



SEVENTH FRAMEWORK PROGRAMME

D25.2: “Report on Y1 and updated plan for activities”

FP7-ICT-216863/PoliMI/R/PU/D25.2

Project Number: FP7-ICT-2007-1 216863

Project Title: Building the Future Optical Network in Europe (BONE)

Contractual Date of Deliverable: 30/11/09

Actual Date of Delivery: 31/12/09

Workpackage contributing to the Deliverable: WP 25 TP: Optical Interconnects

Nature of the Deliverable R

Dissemination level of Deliverable PU

Editors: Partner / names
PoliMI / Guido Maier, Achille Pattavina

Abstract:

After one-year activity, WP25 has reached the middle of its planned duration. This is the first deliverable of WP25 reporting actual technical content.

The first four general sections compiled by the WP leader are followed by section 5 reporting about each Joint Activity (JA) of the workpackage. This section has been created by merging the contributions received by the JA leaders.

Keyword list:

Optical interconnect (OI), board-to-board OI, rack-to-rack OI, shelf-to-shelf OI, on-chip OI, energy efficiency, power saving, optical backbone



Disclaimer

The information, documentation and figures available in this deliverable, is written by the BONE (“Building the Future Optical Network in Europe) – project consortium under EC co-financing contract FP7-ICT-216863 and does not necessarily reflect the views of the European Commission



Table of Contents

DISCLAIMER.....	2
TABLE OF CONTENTS.....	3
1. EXECUTIVE SUMMARY	4
2. INTRODUCTION	5
3. PARTICIPANTS	5
4. GENERAL STATUS OF THE WORKPACKAGE	6
5. JOINT-ACTIVITY DETAILED REPORTS	8
5.1 JA1: IDENTIFICATION OF REQUIREMENTS OF THE OPTICAL BACKPLANE	8
5.1.1 Description of the work carried out	8
5.1.2 Publications	8
5.1.3 Mobility actions.....	8
5.1.4 Y3 planned activities	8
5.2 JA2: SURVEY ON PHOTONIC TECHNOLOGIES FOR OPTICAL INTERCONNECTIONS.....	9
5.2.1 Description of the work carried out	9
5.2.2 Publications	9
5.2.3 Mobility actions.....	9
5.2.4 Y3 planned activities	9
5.3 JA3: SURVEY ON OPTICAL INTERCONNECTION ARCHITECTURES AND SOLUTIONS.....	10
5.3.1 Description of the work carried out	10
5.3.2 Publications	10
5.3.3 Mobility actions.....	10
5.3.4 Y3 planned activities	10
5.4 JA4: THE "BONE SWITCH": A REFERENCE ARCHITECTURE FOR THE OPTICAL BACKPLANE.....	11
5.5 JA5: PERFORMANCE AND COMPLEXITY ANALYSIS OF OPTICAL SWITCHING FABRICS	12
5.5.1 Description of the work carried out	12
5.5.2 Publications	24
5.5.3 Y3 planned activities	25
5.6 JA6: OPTICAL BACKPLANES UTILIZING MICRORING RESONATORS	26
5.6.1 Description of the work carried out	26
5.6.2 Publications	36
5.6.3 Mobility actions.....	36
5.6.4 Y3 planned activities	36
5.7 JA7: HARDWARE EFFICIENT OPTOELECTRONIC SWITCH FABRIC	38
5.7.1 Description of the work carried out	38
5.7.2 Publications	42
5.7.3 Mobility actions.....	42
5.8 JA8: ON-CHIP OPTICAL SWITCHING NETWORKS	43
5.8.1 Description of the work carried out	43
5.8.2 Y3 planned activities	45



1. Executive Summary

The general purpose of WP25 Topical Project “Optical Interconnects” is to study the optical implementation of the interconnection systems inside a high-performance switching/routing system. The target is to develop the capability to design and optimize an optical backplane, finding the technology and architecture best-fitting system requirements

After one-year activity, WP25 has reached the middle of its planned duration. This is the first deliverable of WP25 reporting actual technical content.

Eleven partners have actively collaborated to the Joint Activities (JAs) of this workpackage during the first year. This document contains the detailed description of the JAs at their current status of development.

At the beginning of the first year of WP25 (BONE Y2), nine JAs were agreed with partners who declared their intention to be involved in the workpackage. Four of these JAs (namely JA5, JA6, JA7 and JA9) were already active in WP14 and were transferred to WP25 (with agreement of BONE scientific coordinator) because better suited to the newly-opened Topical Project.

The first three JAs, dedicated to survey and requirement definition, were planned to be in operation in the first year only of the TP. Thus D25.2 is the final report for JA1, JA2 and JA3. Each of these JAs has worked quite efficiently and has produced an extensive report on time.

The JAs from JA5 to JA9 were planned to span over the two years of the TP. All these activities, except JA9, have started and the preliminary results achieved are reported in this deliverable. The reports show a summary of technical content as well as joint and individual publications of the involved partners.

All activities appear to be in quite advanced state of development. JA8 is a little less advanced in terms of results: this is understandable if we consider that it is a brand new activity, while the others have been active also in Y1 under WP14.

Unfortunately, no activity at all has been carried out in the initially-planned JA9 – perhaps this is justified by the activation in BONE of a new WP whose research topic (power consumption and energy efficiency of switching systems) is partially overlapping with JA9. Therefore the WP25 leader decided to close JA9.

No activity has to be reported on JA4, that was planned to start in month 25 and will be active during Y3.



2. Introduction

After one-year activity, WP25 has reached the middle of its planned duration. This is the first deliverable of WP25 reporting actual technical content.

The next two general sections compiled by the WP leader are followed by the section reporting about each Joint Activity (JA) of the workpackage. This section has been created by merging the contributions received by the JA leaders.

The WP leader would like to thank all the BONE partners contributing to this deliverable for the highly-valued technical content they have provided to the workpackage.

3. Participants

Eleven partners have actively collaborated to the JAs of this workpackage during the first year. Table 1 shows the list of the active partners.

A detailed description of the joint activities is provided in Section 5.

Partner #	Member	Country
29	PoliMI	IT
2	TUW	AU
12	UC3M	ES
14	UPCT ¹	ES
22	UoA	GR
23	UoP	GR
30	PoliTO	IT
32	UniBO	IT
36	TUE	ND
39	PUT	PL
45	UCAM	UK

Table 1: Partners actively contributing to the joint activities of WP25 during BONE Y2

¹ It should be noted that UPCT will provide his active contribution in Y3 because involved in JA4, planned for months 24-36.

4. General status of the workpackage

As well described in the first WP25 deliverable D25.1, the main goal of WP25 Topical Project on “Optical Interconnects” can be summarized by the following statement:

- This project studies the **optical implementation of the interconnection systems inside a high-performance switching/routing system**. The target is to develop the capability to design and optimize an **optical backplane**, finding the technology and architecture best-fitting system requirements

The accomplishment of project goals implies sharing expertise and previous work among the involved partners, merging the complementary visions of the participants on the optical implementation of interconnections within the systems.

With deliverable D25.1 this Topical Project has been organized in a structure represented in Table 2. The five Tasks initially proposed by the WP leader group the JAs of the partners according to general topics and provide specific timing to the JAs.

JA #	JA title	JA leader	Timing	Task
1	Identification of requirements of the optical backplane	TUW	13-24	T1: Survey and requirements
2	Survey on photonic technologies for optical interconnections	PoliMI		
3	Survey on optical interconnection architectures and solutions	PoliMI		
4	The “BONE switch”: a reference architecture for the optical backplane	UPCT	25-36	T2: “BONE switch”
5	Performance and complexity analysis of optical switching fabrics	PoliTO	13-36	T3-5: OI-solution proposal and validation at different interconnection layers (rack-to-rack / shelf-to-shelf, board-to-board, chip-to-chip / on-chip)
6	Optical backplanes utilizing microring resonators	UoA		
7	Hardware efficient optoelectronic switch fabric	UCAM		
8	On-chip optical switching networks	PoliTO		
9	Reliability and power consumption of optical interconnects	ERICSSON		

Table 2. WP25 Topical Project structure, as planned at the beginning of BONE Y2

At the beginning of the first year of WP25 (BONE Y2), nine JAs were agreed with partners who declared their intention to be involved in the workpackage. Four of these JAs (namely JA5, JA6, JA7 and JA9) were already active in WP14 and were transferred to WP25 (with agreement of BONE scientific coordinator) because better suited to the newly-opened Topical Project.

The first three JAs, dedicated to survey and requirement definition, were planned to be in operation in the first year only of the TP. Thus D25.2 is the final report for JA1, JA2 and JA3. Each of these JAs has worked quite efficiently and has produced an extensive report on time. Given the size of these documents, they have not been entirely included in the main document of this deliverable: only the abstracts are included and the full-text reports will be delivered as separate documents attached to the main D25.2 file. With the work carried out in JA1, JA2



and JA3 the first task (T1) of the Topical Project is accomplished: the complete overview on Optical Interconnects prepared by these JAs will constitute a solid reference framework for the development of OI solution proposals for the other JAs. The output of T1 will also be disseminated outside BONE by submitting joint papers based on the three reports.

The JAs of the workpackage corresponding to Tasks T2 to T4 are planned to span over the two years of the TP. All these activities, except JA9, have started and the preliminary results achieved are reported in this deliverable. The reports show a summary of technical content and joint/individual publications of the involved partners. In some cases, mobility actions carried out in Y2 under the umbrella of WP25 are reported. Some JA leaders also provided an explicit research planning for Y3.

All activities appear to be in quite advanced state of development. JA8 is a little less advanced in terms of results: this is understandable if we consider that it is a brand new activity, while the others have been active also in Y1 under WP14.

Unfortunately, no activity at all has been carried out in the initially-planned JA9 – perhaps this is justified by the activation in BONE of a new WP whose research topic (power consumption and energy efficiency of switching systems) is partially overlapping with JA9. Therefore the WP25 leader decided to close JA9.

No activity has to be reported on JA4 (alias Task T5), that, as planned, will start in month 25 and will be active during Y3.

5. Joint-activity detailed reports

In the following the content of the JAs of WP25 is described in details.

5.1 *JA1: Identification of requirements of the optical backplane*

Participants: TUW, PoliTO, UniBO

Responsible person: Slavisa Aleksic (TUW)

Deadline: Month 24

5.1.1 *Description of the work carried out*

This JA achieved the completion of the target survey. D25.2 is the final deliverable for this JA. Given the large extension of the survey, it is reported as a separate document (“Limitations and requirements of electrical and optical technologies for high-capacity interconnects”) attached to this document. We include hereafter only a short abstract.

5.1.1.1 *Abstract*

We show a detailed analysis of systems where optical interconnects can be applied. The application targets are the high-performance network nodes that can more benefit from an optical implementation of the backbone subsystem. Such nodes may comprise switches, routers, digital or optical cross-connects, and add-drop multiplexers with different number of ports and different data rates. In light of these requirements, we will identify potential bottlenecks of “traditional” electronic interconnection systems, in terms of scalability towards the upgraded performance required by the future networks. This document will provide an overview on the state-of-the-art structures and technologies in electronic switching systems to implement interconnections. It will then analyze possible limitations of these solutions in front of new requirements and show in which cases and how optics can offer a real competitive advantage to equipment providers.

5.1.2 *Publications*

The content of the survey presented by this joint activity will be submitted for publication.

5.1.3 *Mobility actions*

No mobility action reported.

5.1.4 *Y3 planned activities*

This JA ends in Y2. However some further work is expected in Y3 to publish the obtained results.



5.2 JA2: Survey on photonic technologies for optical interconnections

Participants: PoliMI, PoliTO, TUE, TUW

Responsible person: Guido Maier (PoliMI)

Deadline: Month 24

5.2.1 Description of the work carried out

This JA achieved the completion of the target survey. D25.2 is the final deliverable for this JA. Given the large extension of the survey, it is reported as a separate document (“State-of-the-art technologies for optical interconnection: optical vs. electrical wiring”) attached to this document. We include hereafter only a short abstract.

5.2.1.1 Abstract

For transmitting very high data rates inside systems, photonics allows to overcome several limitations of electrical wiring. We show how this can occur on the typical interconnection distances rack-to-rack, board-to-board and even on-chip. The comparison between optical and electronic transmission is carried out both for serial and parallel links. Then in this survey we report about the state-of-the-art photonic technologies and devices to implement all the functions needed in an interconnection subsystem or backplane, including: signal generation, signal detection, amplification and switching. Latest advances of research in this field are reviewed. Finally, a comparative analysis of power consumption in optical and electronic interconnection systems is reported.

5.2.2 Publications

The content of the survey presented by this joint activity will be submitted for publication.

5.2.3 Mobility actions

No mobility action reported.

5.2.4 Y3 planned activities

This JA ends in Y2. However some further work is expected in Y3 to publish the obtained results.



5.3 JA3: Survey on optical interconnection architectures and solutions

Participants: PoliMI, PoliTO, PUT, TUW, UoP, UNIBO

Responsible person: Guido Maier (PoliMI)

Deadline: Month 24

5.3.1 Description of the work carried out

This JA achieved the completion of the target survey. D25.2 is the final deliverable for this JA. Given the large extension of the survey, it is reported as a separate document (“Optical interconnection in high-speed switching and processing systems”) attached to this document. We include hereafter only a short abstract.

5.3.1.1 Abstract

Large-capacity interconnection systems only based on electronic technology suffer technical bottlenecks that bound their scalability. Optical interconnection has proved to be a valid alternative solution in many applications, ranging from circuit switching to packet switching, to super-computing, to systems embedded on-chip. In this survey we analyze the main applications of optical interconnection. First we describe the main type of system architectures classified according to the application, analyzing their most relevant requirements of interconnection. Then, for each application class, we present the main architectures that have been proposed so far in literature and prototyped in labs or, in some cases, developed as products available on the market. Our overview aims at testifying that optical interconnection is not only a subject for theoretical studies; it is instead an important resource to increase the capability of actual high-speed systems in telecommunications and computing.

5.3.2 Publications

The content of the survey presented by this joint activity will be submitted for publication.

5.3.3 Mobility actions

No mobility action reported.

5.3.4 Y3 planned activities

This JA ends in Y2. However some further work is expected in Y3 to publish the obtained results.



5.4 JA4: The "BONE switch": a reference architecture for the optical backplane

Participants: UPCT, PoliMI, PUT

Responsible person: Pablo Pavon (UPCT)

Deadline:

Month 36 (starting from Month 24)

This JA will start on Year 3 of the BONE project.

5.5 JA5: Performance and complexity analysis of optical switching fabrics

Participants: PoliTO, PoliMI, UniBo, TUW, UPCT, UoP

Responsible person: Fabio Neri (PoliTO)

Deadline: Month 36

5.5.1 Description of the work carried out

The work on the performance and complexity analysis of optical switching fabrics continued in the second project year with contributions from a number of partners. While several metrics have been used to analyze fabric performance, complexity and cost is a common metric used in most contributions to optimize the designs. The first contribution below is due to TUW, and proposes the implementation of OBS switching fabrics, which use slow and cheap fabrics to implement fast parallel fabrics. The second contribution from UniBO studies the share-per-wavelength architectures as a suitable compromise from the complexity and cost point of view, since the fixed wavelength converters used are more feasible than the tunable counterparts, but still contribute to share the switching resources available for several interconnection applications. The inclusion of fixed microring devices, proposed by PoliMI, constitutes a step ahead in using innovative devices to overcome the passivity of the fabrics (through wavelength selective components) and also looks ahead to their inclusion as simple optical fabrics in a near future. For what regards PoliTO, work was carried on on the simple optical switching fabric architectures studied in the first year, performing a complexity analysis based upon the characterization of commercial optical devices.

5.5.1.1 Optical switch fabric for OBS with channel bonding

Currently, TUW is working on the analysis of an optical switching fabric, optimized to be cascaded and using existing slow optical fabrics (and low cost-complexity). TUW analysed cascability and performance of optical switching fabrics for optical burst-switched (OBS) networks with channel bonding (CB). In OBS-CB networks, bursts are aggregated and transmitted on multiple wavelength channels in parallel, and therefore, the complexity and cost of core nodes can be reduced. Additionally, the number of wavelength converters can be reduced in OBS-CB systems. Due to the fact that both the port count of core nodes and the temporal size of large data bursts are reduced, the scheduling algorithm applied is expected to be simpler than that used in conventional unbonded OBS systems with wavelength switching (OBS-WS).

The proposed OBS-CB system utilizes the wavelength domain to increase the transport rate by transmitting data bursts in a parallel manner [1]. Figure 1 shows the proposed architecture of an OBS-CB core node.

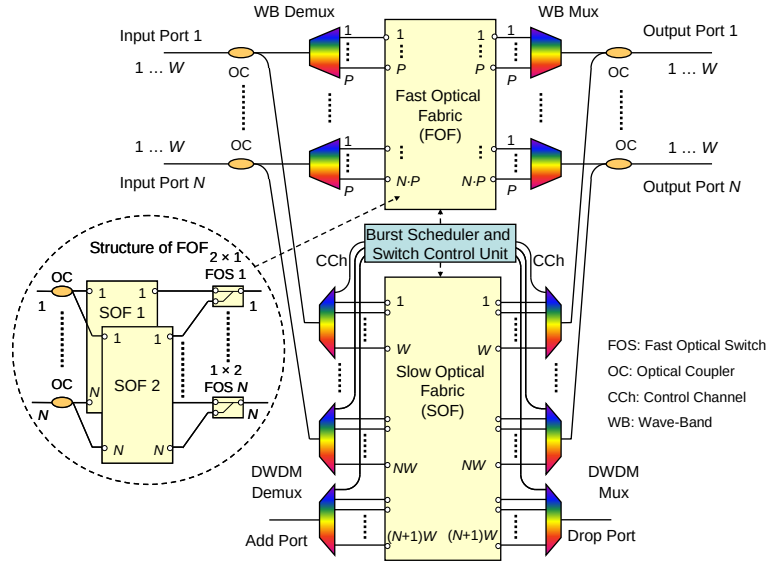


Figure 1 Proposed architecture of an OBS-CB core node.

Let us assume that the number of wavelength channels per fiber port is W . They are grouped into $P = W/K$ wavebands. Each waveband consists of K bonded channels that are used to transmit data bursts in parallel. The role of the fast optical fabric (FOF) is to switch the wavebands from inputs to output ports on a burst-by-burst basis. The FOF can be realized as an arrangement of 2 slow optical fabrics (SOFs) that operate in parallel. When using such an arrangement, active paths through the slow switching fabrics are selected at output ports by means of N fast 2×1 optical switches (FOSs). Within the time period between the arrivals of a burst control packet (BCP) and the corresponding data burst, i.e., within the offset time, a new path can be “offline” prepared through reconfiguring one of the slow optical fabrics. Once the new path is configured, the 2×1 FOS at the respective output port quickly selects the new active path within the short guard time between two bursts. Thus, fast optical burst switching is implemented by using two large low-speed fabrics and N small and fast switches instead of using a large fast optical switching fabric, thereby reducing the overall cost of the core node. The 2×1 fast optical switches can be realized using different components. In this study we consider using optical gates based on semiconductor optical amplifiers (SOAs).

Optionally, an additional slow optical fabric (SOF) can be used to establish paths at the wavelength level as well as to add and drop some unbonded wavelength channels if needed. We assume that the control information is transmitted over a dedicated control channel (CCh) that is separated at the input of the node and processed in the Burst Scheduler and Switch Control Unit.

Switch Cascadability. There are two main limiting factors on cascadability of an OBS-CB switch. The first one is the timing skew that occurs between burst slices transmitted on different wavelength channels due to the chromatic dispersion in fiber transmission links. The second limiting factor is degradation of optical signals during propagation through network nodes. Concerning the first limiting factor, there are several methods to compensate for chromatic dispersion and to eliminate limitations caused by the timing skew. These techniques include the use of dispersion-shifted fiber (DSF), a shepherd pulse technique, electronic bit realignment at the receiver, dispersion compensating fiber (DCF), and dispersion managed fiber (DMF). The influence of the timing skew can be further reduced by increasing the guard band between two consecutive bursts. The semiconductor optical amplifiers, which are considered for use in the fast optical switches, are main sources of the

amplified spontaneous emission (ASE) noise and may cause various undesirable effects when operated in the saturation regime. The accumulation of ASE noise leads to a degradation of the optical signal-to-noise ratio (OSNR), while saturation leads to gain compression, intersymbol interference, and crosstalk between signals on different wavelength channels. Therefore, we analyzed cascability of OBS-CB switches considering two parameters: post-detection signal-to-noise ratio (γ) and saturation power (P_{sat}), similar to the approach presented in [2].

Figure 2 shows the signal path through a cascade of Q OBS-CB switches. The noise in a receiver placed after the Q -th stage would be dominated by the signal-spontaneous and spontaneous-spontaneous beat noise [3]. To achieve a bit error rate of 10^{-9} , the value of the post-detection signal-to-noise ratio (SNR), γ , has to be higher than 144. We determined the maximum number of hops achievable for OBS-CB systems by considering 40 Gbit/s non-return-to-zero (NRZ) signals transmitted over sixteen wavelength channels, where four, eight, or even all sixteen channels are bonded, i.e., for $W = 16$ and $K = 4, 8$, and 16 . The internal gain of SOAs was chosen such that all losses within the switches (L_{in} , $L_{SOA,in}$, $L_{SOA,out}$, and L_{out}) are effectively compensated. The maximum number of hops is limited by the both above mentioned constraints, namely γ has to be higher than 144 and the signal at the output of the cascade should be lower or equal to P_{sat} . The SOA saturation power we set to 13.6 dBm in our calculations.

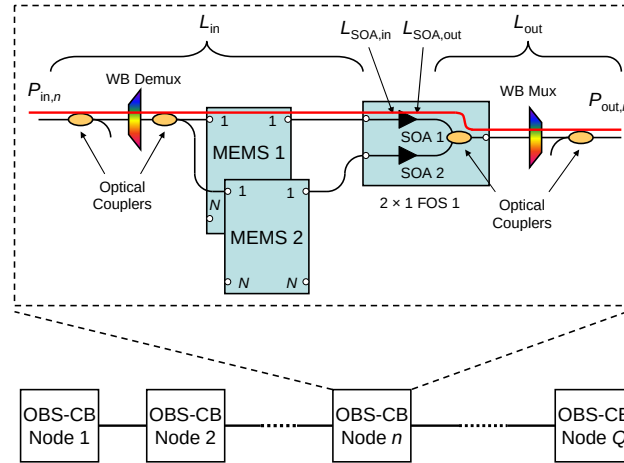


Figure 2 Signal path through OBS-CB switches.

Figure 3 shows the calculated values of γ and optical signal-to-noise ratio (OSNR) in dependence on the number of cascaded switches. It can be observed from the Figure 3 that cascability of OBS-CB switches is mainly limited by the post-detection SNR when considering 40 Gbit/s signals. In this case, a maximum of 7 nodes can be cascaded when assuming a BER of 10^{-9} . If we would only take into consideration the limitation regarding saturation power, then up to 12 hops for $K = 8$ and up to 23 hops for $K = 4$ would be possible. When many channels are bonded, i.e., when more than a half of the total number of channels in the system is used to transmit bursts in a parallel manner, the SOA goes into saturation, which causes a reduction of the maximum achievable hop number.

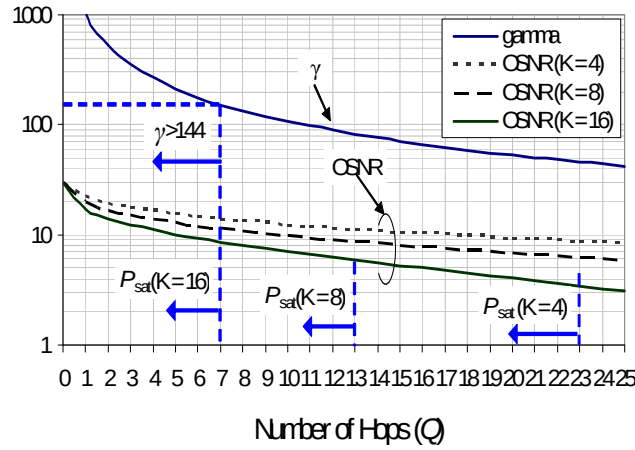


Figure 3 Cascadability of OBS-CB switches considering 40 Gbit/s non-return-to-zero (NRZ) WDM signals.

5.5.1.2 Share-per-wavelength architectures using fixed wavelength converters

During the e-Photon/ONe project the shared-per-input-wavelength (SPIW) switch has been designed as core node for optical packed switched networks [4]. SPIW switch solves contention in the wavelength domain and exploits simple fixed-input/tunable-output wavelength converters (WCs).

To further reduce the complexity of these complex devices, UNIBO propose here an architecture, which exploits fixed wavelength converters (FWCs). A FWC converts any wavelength on input to a fixed wavelength on output, so FWCs do not require a tunability range; this is very important given that WCs with large tuning range and small tuning time are not available yet and are very complex devices. As a matter of fact, the tuning time is related to the tuning range required, the higher the range the higher the tuning time. So FWCs seem to be a good solution especially when the number of wavelengths in the system increases.

The SPOW architecture is sketched in Figure 4 as in the case N Input/Output Fibers (IF/OFs) with M wavelengths each. The switch is equipped with M pools of r_w Fixed-output Wavelength Converters (FWCs), which are shared among all Input Wavelength Channels (IWCs). Each pool is dedicated to a specific wavelength, so each of the r_w FWCs in the k -th pool can convert any input wavelength to a fixed output wavelength k . The SPOW switch employs FWCs instead of tunable WCs, thus allowing easier implementation and cost saving. The total amount of FWCs for the whole architecture is $C_{SPOW} = Mr_w$.

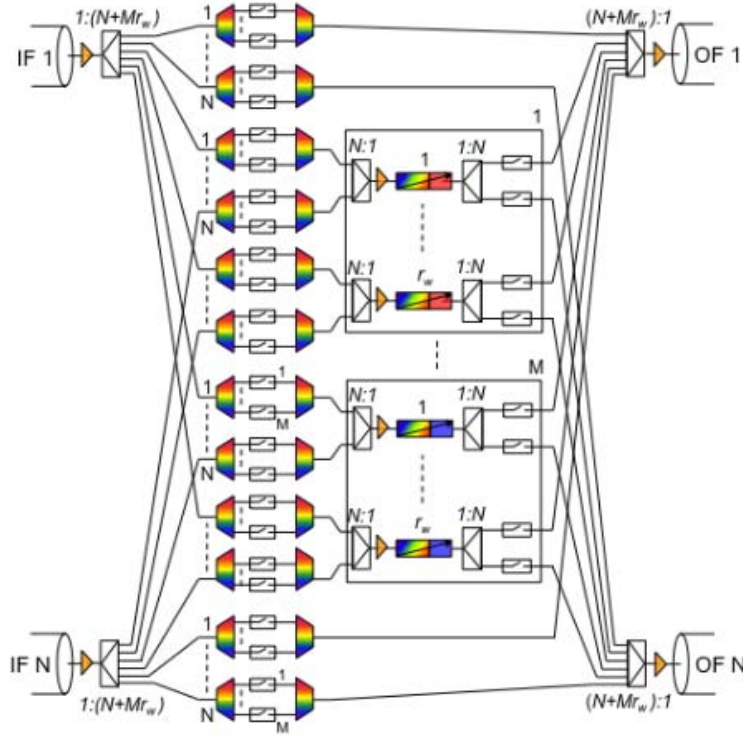


Figure 4 SPOW architecture with N IF/OFs with M wavelengths. The switch is equipped with M pools of r_w WCs dedicated per output wavelength.

This architecture has been proposed as optical switch for optical networks, but it can also be exploited as optical interconnection fabric to connect line cards in a small distance.

The architecture here proposed operates on the Broadcast-&-Select (B&S) principle. Similar B-&S architectures have been proposed to implement the well-known shared-per-node (SPN) and the SPIW schemes in [5]. A packet coming on an IWC is broadcasted to all the OFs, and Optical Gates (OGs) allow to select the desired OF. In particular Wavelength Selectors (WSs), consisting of MUX/DEMUX pairs connected by an array of M OGs, implement the switching matrix. To let a signal on the j -th wavelength go through a WS, the corresponding j -th OG must be set to the ON state. OGs set to the OFF state will block further propagation of the signal. The WDM signal coming on an IF is split into $N+Mr_w$ copies; N of these copies are connected to the N OFs through WSs, to allow the direct connection between the input and output channels, for those packets which can be forwarded without wavelength conversion. The other copies of the input signal are connected to the $C_{SPOW} = Mr_w$ FWCs, again through WSs. In that way, each IWC may exploit, in principle, any FWC. The output of a FWC must be connected to all OFs, so the signal is split into N copies, each one connected to an OF through an OG.

In SPOW, a packet can be forwarded in a given wavelength λ_k if and only if that wavelength is free on the destination OF and there is at least one free FWC in the corresponding pool k . This architecture provides good flexibility in terms of conversion capability. As a matter of fact, when a packet cannot be forwarded on a given wavelength due to FWC unavailability, another wavelength can be checked, until a free wavelength on the OF with an available WC in the corresponding pool is found. A packet may exploit a number of FWCs much larger than r_w .

To correctly manage packet forwarding in SPOW switch, a scheduling algorithm (SA) is defined related to the specific wavelength converters sharing scheme adopted. SAs typically aim at keeping the number of wavelength conversions as low as possible, thus minimizing the

number of wavelength converters needed. The scheduling problem is especially critical in the synchronous case where a decision for a number of packets arriving at the same time slot needs to be made within the slot time, so a SA with limited computational complexity is needed. In this paper a heuristic SA, similar to that presented in [5] for a different sharing scheme, is proposed. This kind of SA can be executed within a reference time of 1 μ s, as discussed in [5].

The SA for the SPOW works in this way: when a packet needs conversion, the SA randomly selects a free wavelength λ_k on the targeted OF and checks whether there is a free FWC among those able to convert the packet on that wavelength (one of the rw on pool k). In case this FWC is not found, the SA selects another free output wavelength and tries again, until a FWC to convert the packet is found.

The packet is lost due to conversion unavailability only in the case all free wavelengths on the destination OF are checked without finding a FWC available to convert the packet.

In this way, the number of times the SA must check the free output channel is not known a priori, and this would result in higher complexity, which is also not easy to be evaluated. To avoid this, the proposed SA uses the intersection among the set of the free wavelengths on a given OF and the wavelengths where the corresponding pool has at least one FWC still available for wavelength conversion. The wavelengths contained in this intersection can be effectively exploited to convert the packet, so it is forwarded to one of those wavelengths, randomly selected.

The computational complexity of this scheduling algorithm is $O(NM)$, as the others proposed for the SPN and SPIW switches.

The performance of the SPOW switch in terms of packet loss probability (PLP) is now presented. Synchronous slotted environment, where packet length is equal to time slot duration is assumed. Bernoulli arrivals are considered due to the buffer-less nature of the switches. Results have been obtained via simulation, with confidence interval less than or equal to 5% of the mean with 95% probability. SPOW is compared to the SPIW and SPN switches, presented in [5] and [6], respectively. Figure 5 and Figure 6 plot the PLP for the SPOW, SPIW and SPN switches. PLP is plotted as a function of the total number of WCs normalized by the total number of channels, $\alpha = \#WCs/NM$, varying load as in the case $N=16$, $M=8$ and $N=8$, $M=48$, respectively. It is worth to remind that α is a discrete value, given by rw/N ($rw = 1, \dots, N$) for SPIW and SPOW.

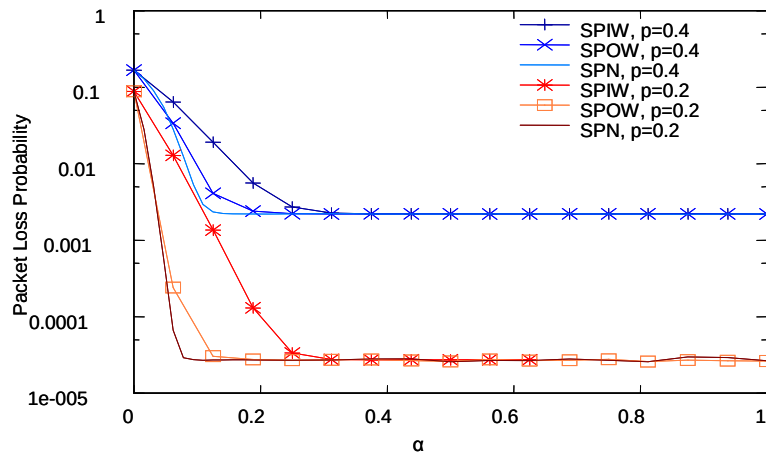


Figure 5 PLP for SPOW, SPIW and SPN as a function of the total number of WCs normalized by the number of channels NM , varying load (p), as in the case $N=16$, $M=8$.

The figures show that the SPOW switch performs very close to the SPN in the region where the WCs are limited, even if it uses FWCs, while it largely improves the performance obtained by SPIW. Figure 6 shows how the SPOW is able to perform closely to the SPN switch even when M increases, thus facilitating the implementation of DWDM high capacity switches (tunable WCs with high tuning range not needed).

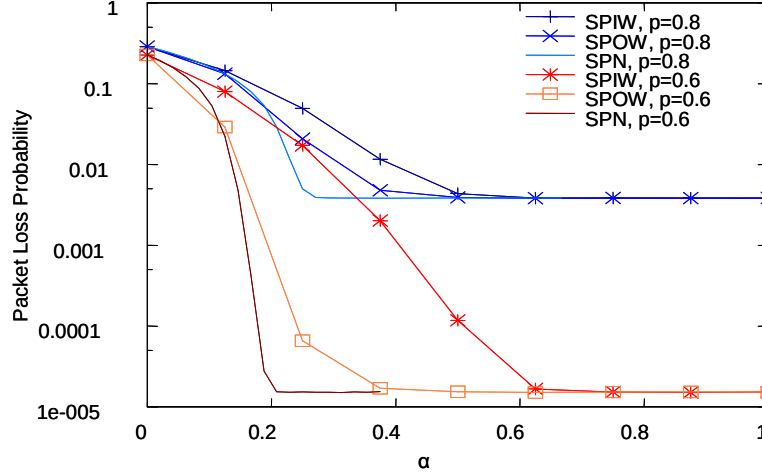


Figure 6 PLP for SPOW, SPIW and SPN as a function of the total number of WCs normalized by the number of channels NM , varying load (p), as in the case $N=16$, $M=8$.

In conclusion, the SPOW provides good PLP (close to SPN, which represents the perfect sharing scheme) for every switch configuration while employing simple FWCs.

The SPIW provides good performance when N is high and M is low, while it performs worse when N is low and M high (see Figure 6). Instead, the SPIW seems to provide some advantage in terms of scalability and power consumption. The analysis of these aspects for the proposed architectures represents the future work UNIBO is planning to do within WP25.

5.5.1.3 Optical switching fabrics using fixed microring filtering structures

The contribution of POLIMI to is based on a feasibility and scalability analysis of switching architectures with overall capacity in the order of Tbit/s [7]. The work has been focused on the issue of interconnecting line cards of routers/switches, exploiting wavelength agility at transmitters to control signal routing across a micro-ring-resonator based structure.

Optical ring resonators can be adopted to create switching matrices working as $N \times N$ wavelength routers or cross-connects. In such a case, N is the number of routed channels that equal the matrix size. The underlying idea of the adopted solution is to create a multi-stage structure in which stages consist of resonators that separate ports the input channels routing them on their “drop” and “through” ports. Routing occurs according to the channel wavelengths and to the Free Spectral Range (FSR) of the ring. Rings of different stages have different FSRs, suitably designed so to minimize crosstalk and maximize scalability. A direct effect of the FSR-design criteria is on the spacing between adjacent channels that will increase stage-by-stage. Hence, FSR values are not directly proportional to N , with a great advantage in terms of scalability.

A 3-stages example of this architecture is shown in Figure 7. Rings are arranged in a tree-like topology, in which the j -th stage ($1 \leq j \leq \log_2 N$) contains 2^{j-1} rings. Rings belonging to same stage have equal radius (and thus equal FSR). N (eight) WDM signals, evenly spaced in wavelength by $\Delta\lambda$, enter the structure at the IN port, and they have to be demultiplexed and transmitted at the outlets (the output ports of all the rings of the last stage). In order to obtain

one wavelength at each output port, the FSR parameter of the j -th stage should be set as it follows: $FSR_j = 2^j \cdot \Delta\lambda$ ($FSR_1 = 2\Delta\lambda$, $FSR_2 = 4\Delta\lambda$ and $FSR_3 = 8\Delta\lambda$). By adopting this approach, the FSR of the rings of the last stage (the maximum FSR in the network) is necessarily directly proportional to the number of managed wavelengths N . Since the FSR is inversely proportional to the ring radius, this implies that, as N increases, rings must be designed smaller and smaller. But given a certain technology, there is always a lower bound on the ring radius (below which waveguides are no longer able to confine optical beams and loss starts to rise very quickly). Thus that imposes a scalability bound on N .

To overcome this limitation the simple solution we are proposing is to set the FSR to a different value, that is: $FSR_j = (2^j \cdot \Delta\lambda) / N$. Exploiting integer sub-dividers of the channel separation $\Delta\lambda$ the ring architecture is still able to perform channel demultiplexing. At the same time, we ensure by this approach that the worst-case FSR is independent of N (i.e. $FSR_n = \Delta\lambda$ with $n = \log_2 N$). This implies that the architecture is fully scalable with N (obviously choosing a suitable channel wavelength-separation).

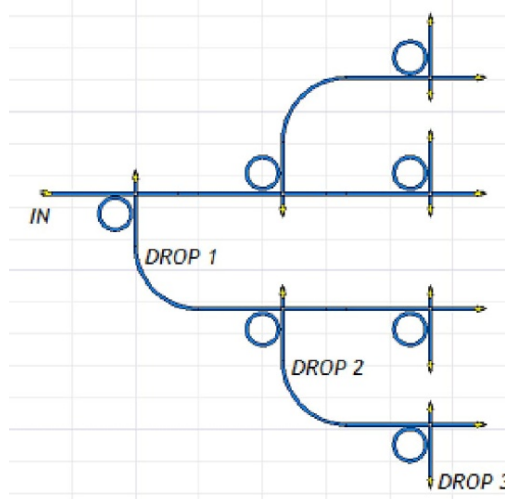


Figure 7 Ring-resonator-based demuxing structure

The aforementioned structure can be used as the central stage of the architecture represented in Figure 8, in which N transmitters are divided in S switching planes, each one comprising N/S transmitters. The multiplane architecture [8] allows to further improve scalability by overcoming limitations due to differential loss between the drop and through ports of the rings. Starting from the left side, a stage of $N/S:1$ couplers collects signals coming from each single plane. Beside the ring structure, there is a second stage of couplers that are needed to achieve connectivity between input/output ports belonging to different switching planes. S EDFA amplifiers are then placed after the couplers and finally S demultiplexers route wavelengths to different output ports within each plane (these latter demultiplexers may be obtained by exploiting again the resonator-ring-based demultiplexing architecture of Figure 8).

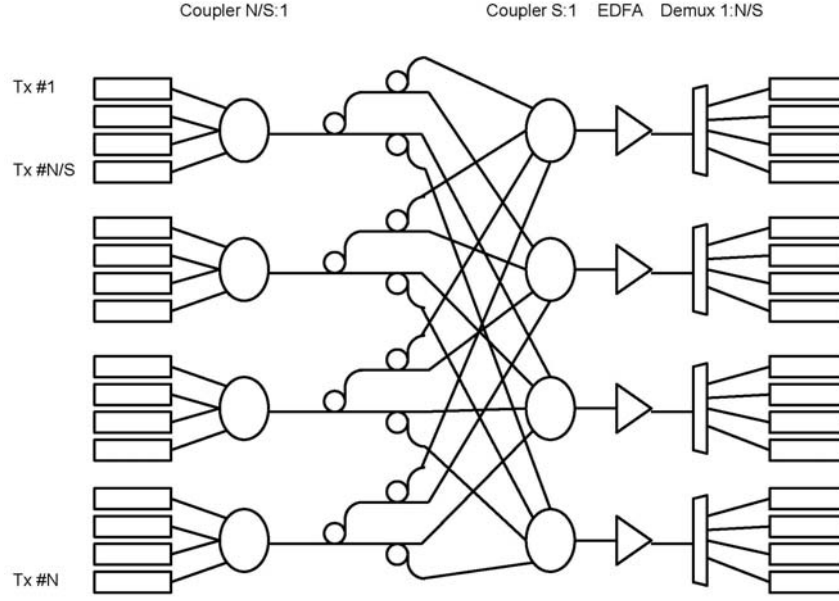


Figure 8 Ring-resonator-based architecture

Once the architecture has been designed PoliMI evaluated the number of rings (K) crossed by a channel. Then the insertion losses were computed for all the rings, which for each ring depends on the ring radius and on the waveguide attenuation. The latter depends on the optical material chosen to fabricate the passive interconnection architecture, while the former is determined by the FSR of the ring. We have assumed equal coupling coefficients between all the waveguides composing each ring, setting the coupling coefficient to the minimal value that guarantees a per-channel passing-bandwidth amplitude equal to the channel bit rate. This choice effectively improves rejection ratio and through-port insertion loss, though it implies an increase in the drop-port insertion loss.

Scalability is then evaluated in terms of total aggregated bandwidth by exploiting the model proposed in [8]. The aggregated bandwidth is calculated as the maximum number of allowed transmitters for the architecture multiplied by the bit rate of a single channel. The characterization of the different components taking into account physical impairments and power penalties have been defined in terms of: (i) average transmitted power for each transmitter, (ii) losses due to modulation and other effects, (iii) losses due to couplers, ring structures and demultiplexers. Two further constraints are considered: (a) signal power at the receiver must be greater than receiver sensitivity and (b) calculated Optical Signal-to-Noise Ratio (OSNR) at the photodiode must be greater than a target OSNR value that guarantees an acceptable BER performance².

Total aggregate bandwidth evaluation has been carried out considering OSNR_{TX} values at the transmitter ranging from 30 dB to 60 dB. Ring resonators operating with spacing of 25, 50, 100 and 200 GHz between the WDM channels have been designed, and transmission rates of 2.5, 10 and 40 Gbit/s are considered. Aggregate bandwidth is directly influenced by the bit rate and by the value of N/S and S . The upper limit for the number of planes is set to $N^{1/2}$. However, if both constraints on signal power at the receiver and on calculated OSNR at the photodiode must be satisfied, optimal S never reaches its maximum theoretical value.

² Defined over a bandwidth equal to the bit rate

With reference to the number of transmitters/receivers in each plane, scalability with $S \leq 50$ ($\Rightarrow N \leq 2500$) and $S \leq 100$ ($\Rightarrow N \leq 10000$) has been investigated. Nominal transmitter OSNR allows to calculate the first noise contribution to the architecture that deeply impacts on the scalability analysis. As the OSNR_{TX} changes, the two constraints differently affect throughput: for values lower than 50 dB, OSNR at receivers is the limiting factor for the scalability; instead, for higher OSNR_{TX} values, power constraint becomes more stringent.

Figure 9 depicts the maximum aggregate bandwidth as a function of OSNR_{TX} for all the considered bit rate/channel spacing combinations and with $S \leq 50$. It can be noticed that it is possible to reach very high capacities, in the order of tens of Tbit/s; higher bit rates clearly allow higher bandwidths, and different bit rates show different trends as OSNR_{TX} increases. This can be traced back to the presence of a throughput dependency on channel spacing that is proportional to the transmission rate. As far as sources at 2.5 Gbit/s are concerned and the previously described constraints are taken into account, it seems that channel spacing does not greatly impacts on maximum capacity. Furthermore, the dispersion of results increases with bit rate: at 10 Gbit/s, the difference of aggregate bandwidth between worst case (i.e. $\Delta\lambda = 25$ GHz) and best case (i.e. $\Delta\lambda = 200$ GHz) is in the order of some Tbit/s, while at 40 Gbit/s it is in the order of tens of Tbit/s.

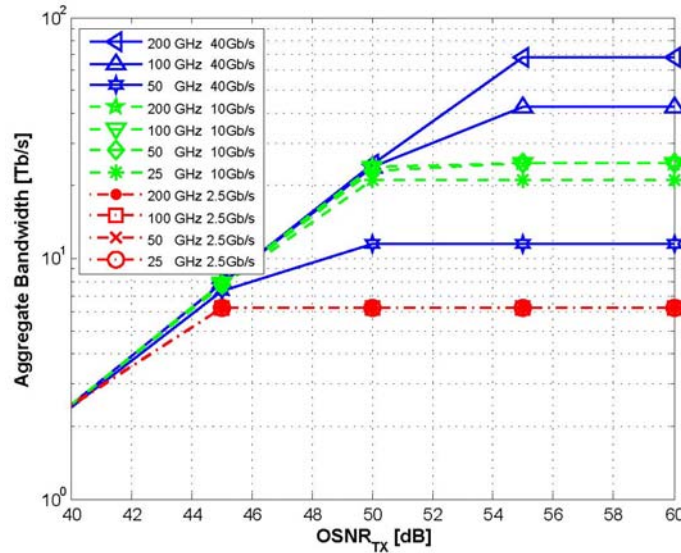


Figure 9 Maximum throughput vs. OSNR_{TX} for $S \leq 50$ and different channel spacing/bit rate

Figure 10 shows the results of a second scalability analysis which differs from the previous one for the limit $S = 100$. As expected, this higher limit for the maximum number of transmitters in each plane produces higher throughputs. Different results with respect to the previous analysis are obtained in the case in which low bit rates are taken into account. For instance, at 2.5 Gbit/s with $S = 50$, channel spacing does not significantly affect throughput because OSNR at receivers is generally high³ and the considered constraints can allow for having a number of transmitters per plane higher than 50 for all channel spacing values. When S_{MAX} is set to 100, throughput becomes sensitive to channel spacing. This observation is generally true also for higher bit rates.

³ Noise power is calculated integrating noise spectral density on a narrower bandwidth

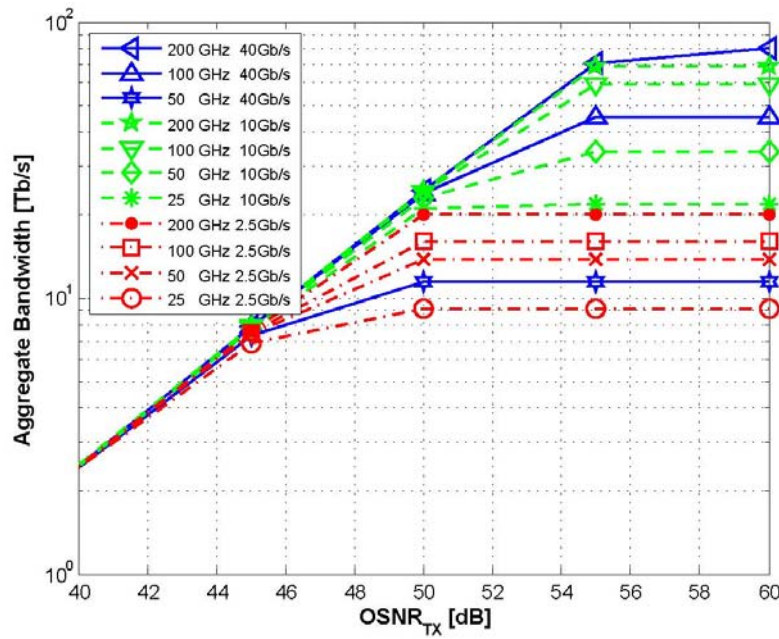


Figure 10 Maximum throughput vs. $OSNR_{TX}$ for $S \leq 100$ and different channel spacing/bit rate

5.5.1.4 Complexity analysis of optical switching fabrics

PoliTO analysed the optical switching fabrics proposed jointly with PoliMI and UniBO in the previous year. Basically three architectures were analysed, which are depicted in Figure 11 , Figure 12 and Figure 13 . The first architecture, called multiplane couple and amplify – MCAD, uses S groups of transmitters and receivers, called planes, in order to balance transmitter tunability with space multiplicity. It is based on space switches (using SOAs as gates) to perform plane selection.

The second architecture, called Wavelength Selective - WSV, uses SOA gates to implement wavelength *and* space selection. The third architecture, called Wavelength Routing Space - WRS, uses wavelength routing with full tunability (transmitters with N wavelengths) to perform all input-output combinations; it exploits the AWG periodicity properties to avoid coherent crosstalk due to wavelength reuse.

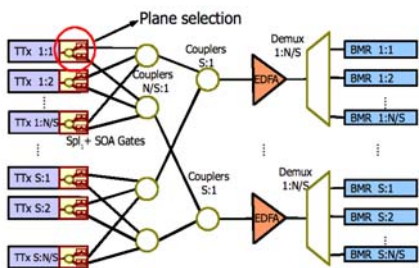


Figure 11 Multiplane Couple-Amplify-Demultiplex (MCAD) architecture

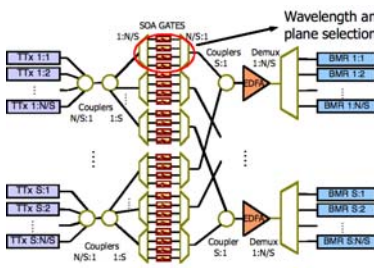


Figure 12 Wavelength Selective (WSV) architecture

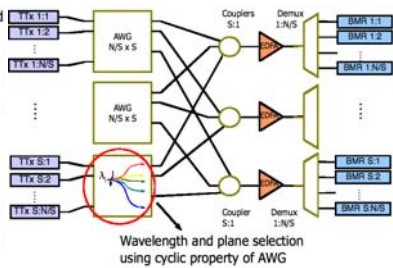


Figure 13 Wavelength Routing Space (WRS) architecture

PoliTO's results show that the scalability (in terms of total switching capacity) is strongly dependent on the transmitter OSNR; noise is accumulated through the fabric from N independent sources and the result is a very low scalability in presence of noisy transmitters (see [9]). When a reasonable level of transmitter noise is assumed (around 50 dB of OSNR), the scalability of all architectures (in terms of aggregate capacity) is very similar, and the complexity of the considered switching fabrics can be fairly evaluated.

PoliTO performed a survey of several optical devices available on the market [10] and proposed a cost model for the different types of devices: couplers, demultiplexers, arrayed waveguide gratings (AWG), transmitters, receivers and amplifiers. An example of this analysis is the coupler cost dependence with its size (number of ports) depicted in Figure 14 , which resulted in a sublinear dependence. Another example are the AWGs, whose cost is determined by their size, and by the amount of wavelength periods upon which they are required to operate (in the case of the WRS architecture). By using these cost models, the previous architectures were characterized in terms of costs when they are configured in such a way that they optimize the optical resources available, i.e., when they scale as much as possible (by varying N and S) while both power budget and OSNR constraints remain just above the feasibility limits to guarantee the target bit error rate (BER). The maximum aggregate bitrates and their associated complexities are summarized in Figure 15 . More details are available in [8].

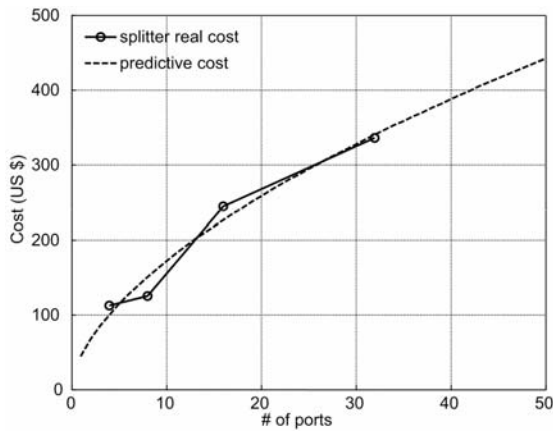


Figure 14 Cost regression taken from commercial splitters

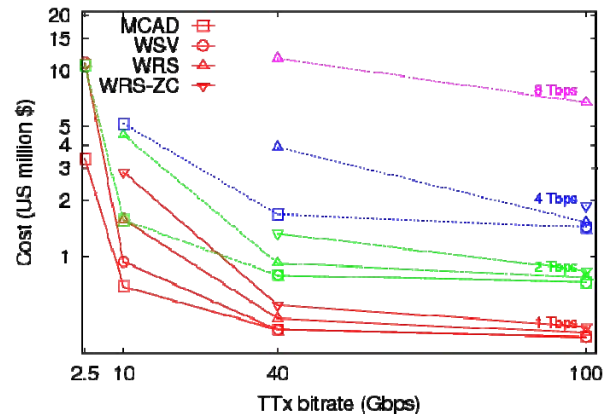


Figure 15 Scalability and Complexity dependence from bitrate

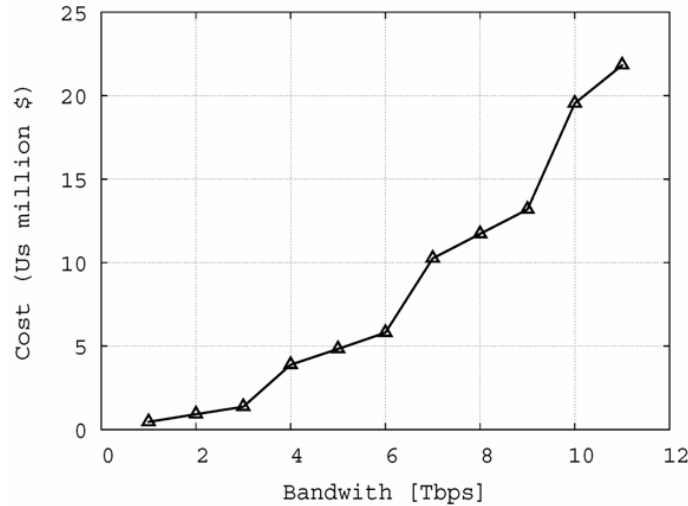


Figure 16 Aggregate bitrate cost trend for WRS at 40 Gbps

The estimated costs are plotted according to the feasible aggregate capacity offered by fabrics (1, 2, 4, 8 Tbps). The limit in source noise considered in this work leads to an advantage to the WRS architecture, since it uses mostly passive devices, and therefore it reaches higher aggregate bitrates; but this increases the cost of the transmitters needed by WRS, since it requires wider transmitter tunability, and transmitter costs were assumed to depend from their

tunability range. On the other hand, several Tbps are achieved when using higher bitrates on the fabric ports, but this means that only few ports are possible (see 4 Tbps, and 8 Tbps curves). The MCAD architecture costs less when feasible due to the optical devices it uses (mainly couplers, which are wideband and cheap). SOA costs influence highly the cost of the WSV architecture, which is feasible also for 2.5 Gbps port bitrates.

These costs, while they still appear high, set an upper bound to the real cost of a final product. Figure 16 shows the scalability-cost relationship for a particular bitrate, which appears to be more than linear. Higher integration levels, and volume discounts for optical components, and the economies of scale can introduce significant cost reductions, which would make optical technologies play a role in future generations of high performance packet switches and routers.

5.5.1.5 References

- [1] S. Aleksic, "Optical Burst-Switched WDM Networks with Channel Bonding", *6th Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP 2008)*, Graz, Austria, July 2008, pp. 405 - 409.
- [2] R. F. Kalman, L. G. Kazovsky, and J. W. Goodman, "Space Division Switches Based on Semiconductor Optical Amplifiers", *IEEE Phot. Techn. Lett.*, Vol. 4, No. 9, 1992, pp. 1040 – 1051.
- [3] N. A. Olsson, "Lightwave Systems with Optical Amplifiers", *IEEE Journal of Lightwave Technology*, Vol. 7, No. 7, 1989, pp. 1071 – 1082.
- [4] C. Raffaelli, M. Savi, A. Stavdas, "Multistage Shared-per-Wavelength Optical Packet Switch: Heuristic Scheduling Algorithm and Performance", *IEEE Journal of Lightwave Technology*, Vol. 27, No. 5, March 2009, pp. 538-551.
- [5] V. Eramo, A. Germoni, C. Raffaelli, M. Savi, "Multifiber Shared-Per-Wavelength All-Optical Switching: Architecture, Control and Performance", *IEEE Journal of Lightwave Technology*, Vol. 26, No. 5, March 2008, pp. 537-551.
- [6] V. Eramo, M. Listanti, P. Pacifici, "A Comparison Study on the Number of Wavelength Converters Needed in Synchronous and Asynchronous All-Optical Switching Architectures", *Journal of Lightwave Technology*, Vol. 21, No. 2, Feb. 2003, pp. 340 - 355.
- [7] R. Gaudino, G. Gavilanes Castillo, F. Neri, and J. Finochietto, "Simple optical fabrics for scalable terabit packet switches", *IEEE International Conference on Communications 2008 (ICC '08)*, May 2008, pp. 5331-5337.
- [8] D. Cuda, R. Gaudino, G. Gavilanes Castillo, F. Neri, G. Maier, C. Raffaelli and M. Savi, "Capacity/cost tradeoffs in optical switching fabrics for terabit packet switches", *International Conference on Optical Network Design and Modeling 2009 (ONDM 2009)*, Feb. 2009, pp. 1-6.
- [9] J. M. Finochietto, R. Gaudino, G. Gavilanes, F. Neri, "Can Simple Optical Switching Fabrics Scale to Terabit/s Switch Capacities?", *IEEE Journal on Optical Communications and Networking*, Vol. 1, No. 3, August 2009, pp. B56 - B69.
- [10] Taken from www.go4fiber.com Product Datasheets

5.5.2 Publications

- V. Eramo (UniRoma1), A. Germoni (CORITEL), A. Cianfrani (UniRoma1), C. Raffaelli (UNIBO), M. Savi (UNIBO), "Evaluation of QoS Differentiation Mechanism in Shared-Per-Wavelength Optical Packet Switches", *ONDM 2009*, Braunschweig, Germany, February 2009
- C. Raffaelli (UNIBO), M. Savi (UNIBO), A. Stavdas (UoP), "Multistage Shared-Per-Wavelength Optical Packet Switch: Heuristic Scheduling Algorithm and Performance", *IEEE Journal of Lightwave Technology*, Vol. 27, No. 5, March 2009, pp. 538-551.

- A. Cianfrani (UniRoma1), V. Eramo (UniRoma1), A. Germoni (CORITEL), C. Raffaelli (UNIBO), M. Savi (UNIBO), “Loss Analysis of Multiple Service Classes in Shared-Per-Wavelength Optical Packet Switches”, *IEEE/OSA Journal of Optical Communications and Networking*, Vol. 1, No. 2, July 2009, pp. A69-A80.
- J. M. Finochietto (PoliTO), R. Gaudino (PoliTO), G. Gavilanes (PoliTO), F. Neri (PoliTO), “Can Simple Optical Switching Fabrics Scale to Terabit/s Switch Capacities?”, *IEEE/OSA Journal on Optical Communications and Networking*, Vol. 1, No. 3, August 2009.
- D. Cuda (PoliTO), R. Gaudino (PoliTO), G.A. Gavilanes Castillo (PoliTO), F. Neri (PoliTO), “Optical Switching Fabrics for Terabit Packet Switches”, *Next-Generation Internet Architectures and Protocols*, B. Ramamurthy, G. Rouskas, K. Sivalingam, Cambridge University Press

5.5.3 Y3 planned activities

Starting from the preliminary results obtained, TUW is planning to continue in the analysis of cascability of nodes with channel bonding technique.

UNIBO is planning to study the advantage in terms of scalability and power consumption of share per wavelength architectures presented here.

PoliMI aims at continuing the study of the proposed architecture and the model adopted for the scalability analysis, comparing performance with other alternative optical-interconnection architectures. PoliMI will also investigate possible improved FSR-dimensioning criteria and techniques to compensate drop-port / through-port differential insertion loss of the resonant rings.

PoliTO is exploring the possibility to use electro-optical devices (i.e. microrings) to move towards a more passive behavior of the previously proposed architectures, in order to increase their performance in terms of two metrics: power efficiency and complexity. At this point PoliMI and PoliTO have been exchanging ideas around this as an initial step of future joint work. The research work of a PhD student at PoliMI will be co-advised by Prof. Fabio Neri of PoliTO.

5.6 JA6: Optical backplanes utilizing microring resonators

Participants: UoA, UC3M, PoliTO, PoliMI

Responsible person: Antonis Bogris (UoA)

Deadline: Month 36

5.6.1 Description of the work carried out

5.6.1.1 A WDM switch based on active InP MRs

Switching with MRs can be achieved by changing the spectral characteristics of the device. The control of the transmission spectrum of the MR cavities has been achieved with several technologies, for example the thermo-optic effect, the free-carrier generation, etc. However, for several reasons they are proven to be inefficient for use in practical systems (slow response time, use of external optical pumping, low switching ratio).

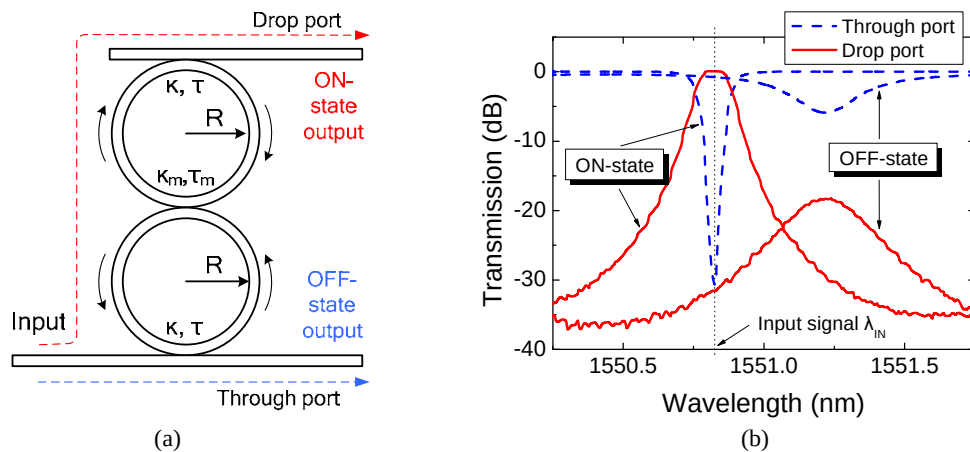


Figure 1 (a) Schematic of the double-ring resonator based switch. (b) Spectral response of both output ports of the switch at ON- and OFF-state.

UoA has studied a new topology of optical switching relying on free-carrier injection in active MRs [P1]. The static and dynamic properties of a wavelength selective switching module based on double series-coupled MRs has been analyzed numerically. The switching is achieved by the tuning of the cavity resonances with the variation of the refractive index by proper current injection. The switch provides high contrast between the switching states with low index change requirements. The switching speed is limited by the carrier lifetime of the gain medium, hence switching times in the ns scale can be straightforwardly attained. The wavelength selectivity of the switching module makes it an appropriate building block in integrated metro network nodes with add-drop functionalities. Except from this type of applications, the specific device can be alternatively utilized as a space switch. Possible scenarios of metro network node architectures employing the specific device are the subject of future work.

The ON-OFF contrast ratio (CR) is defined as the ratio between the transmission levels of the ON-state and the OFF-state at the input wavelength λ_{in} at the drop port of the switch. **Errore. L'origine riferimento non è stata trovata.** a shows the CR against the refractive index change Δn ($= n_{ON} - n_{OFF}$) applied to the rings for changing state, and its dependence on the bus-to-ring field coupling coefficient (κ). High CR (> 25 dB) is observed for all κ values for a Δn of 1×10^{-3} .

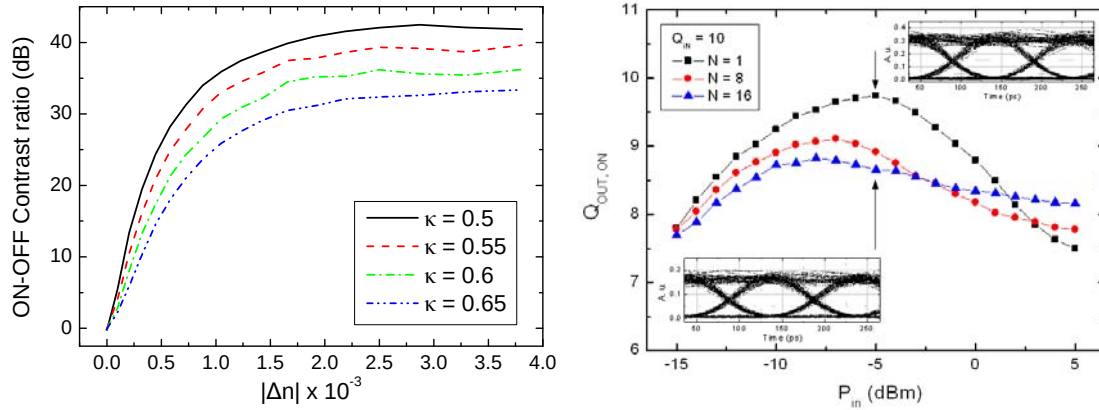


Figure 2 (a) The switching contrast ratio as a function of the change in the refractive index (Δn) of the second ring, for various field coupling coefficients (κ). The ring radius $R = 68.2 \mu m$ ($n = 3.5$, FSR = 200 GHz). (b) Q-factor of the 10 Gbps NRZ signal at the drop output (ON-state) as a function of input power, assuming a switch with $R = 68.2 \mu m$, $\kappa = 0.55$, and a multi-wavelength input signal with one, eight and sixteen wavelengths ($N = 1, 8$ and 16 , respectively). Insets show the corresponding eye diagrams for input power $P_{in} = -5$ dBm and for $N = 1$ (top) and $N = 16$ (bottom).

The switching speed is determined by the carrier lifetime of the InGaAsP medium, which is expected to depend on the current density and the input power. The switching time is calculated as the time required for the output power to change from 10% to 90% of the total change between the two states (low current – high current). Our calculations showed that the switching is limited in the ns scale by τ_{ON} and this behaviour is related to the effective carrier lifetime at the ON- and OFF-state which indicates how fast the carrier density can be modulated by the current.

The operation of the switch under modulated inputs is evaluated with calculations of the output signal Q-factor, assuming operation at ON-state. As input, a 100 GHz - spaced multi-wavelength signal is considered, thus simulating a comb switch. Each wavelength is 10 Gbps non return-to-zero (NRZ) intensity modulated. **Errore. L'origine riferimento non è stata trovata.** b shows the Q-factor of one output wavelength out of the N as a function of the input power of a single wavelength (P_{in}). A wide dynamic range of at least 15 dB for the input power levels is observed for all cases. With the increase of P_{in} , and the number of wavelengths leads to the degradation of the output signal, due to speed limitations of the carrier dynamics at high gain saturation conditions. However a total capacitance of 160 Gbps is exhibited.

5.6.1.2 Scalability of Optical Interconnects based on Microring Resonators

The requirements of high-performance computing applications are making interconnections in electronic technology infeasible. The predictions outlined in the International Roadmap for semiconductors [1] show for example that the most critical performance limitations of on-chip interconnections depend on the latency and power requirements of metal-dielectric wiring,

especially for lengths above the millimeter. Optical interconnection technologies and architectures are therefore receiving increasing attention.

Microring resonators are an interesting optical technology that has been successfully employed to build photonic switching elements, which can be used to design compound interconnection networks. In this research work, building up on experimental results reported in [2], PoliTO proposed a physical layer model of different microring-based Switching Elements (SE), and assessed the scalability of several microring-based interconnection systems. In addition, a new switching element was proposed to achieve better performance and larger scalability.

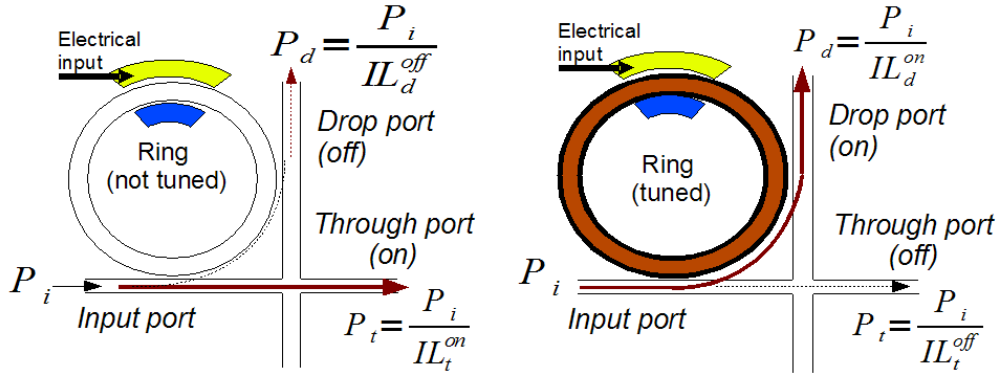


Figure 3 1x2 Microring-based SE in its 2 possible states

A basic structure of a 1x2 SE is depicted in Figure 3. It is composed by two perpendicular waveguides and another one bent into itself, forming the true ring. Both waveguides are near enough to couple light to and from the ring. The ring waveguide can be tuned in such a way that its effective refractive index is changed electrically by carrier injection using p-i-n junctions [4] and consequently the amount of power coupled to the ring (and to the waveguides) depends on an electrical control of the junction. An optical signal entering the input port can be deflected either to the drop port (when the ring is properly tuned to the signal wavelength), or to the through port (if the ring is normally detuned). The structure has an asymmetric switching behavior, because the input signal suffers different penalties depending on the output port it is directed to. Experimental measurements [2] show that entering the ring and exiting by the drop port costs a non negligible power penalty, due to ring coupling and decoupling to the drop port. In Figure 3, the losses perceived by the input signal are denoted as IL_d and IL_t for the drop and through ports respectively; the former represents the coupling loss and propagation through the tuned ring and the drop waveguide; the latter represents losses that generate signal leakage (the dotted path) in the through port.

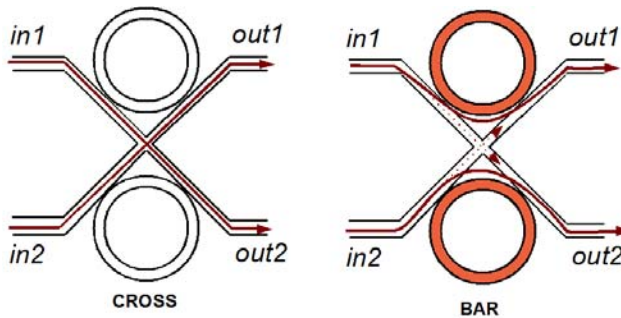


Figure 4 2x2 basic microring switching block in its CROSS and BAR states

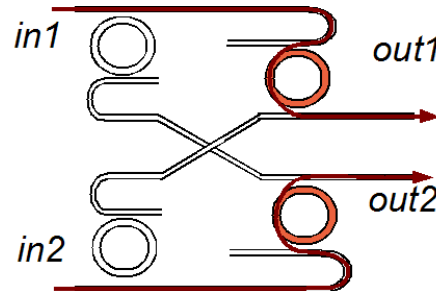


Figure 5 2x2 dilated microring switching block in BAR state; all internal waveguide terminations are anti-reflecting

The Extinction Ratio (ER) for each port can be defined as the ratio between the power measured in the on and off states of the port. Thus, for the drop and through ports: $ER_d = IL_d^{off} / IL_d^{on}$ and $ER_t = IL_t^{off} / IL_t^{on}$. The measured ERs for the two ports are different and this is also evident in the insertion loss asymmetry (given that $IL_t^{on} < IL_d^{on}$).

Other microring structures can be devised to build scalable interconnects. This is the case, e.g., of the 2x2 Basic Switching Block (BSB) depicted in Figure 4, which is based on 2 SEs jointly controlled to provide two possible states, bar (in1 to out1, in2 to out2) and cross (in1 to out2, in2 to out1). Its performance measurements were reported in [5]. The BSB has an asymmetric behavior in the signal impairment introduced in the cross and bar states. Indeed, due to the above mentioned asymmetry of 1x2 SEs, the cross state of BSBs introduces high signal leakage (crosstalk), while the bar state has higher losses. Due to this asymmetry, in larger interconnection networks, the total effect in signal propagation is the result of the state of all SBs in the lightpath.

In Figure 5 we propose a new version of a 2x2 SB, named Dilated Switching Block (DSB), whose main advantage is the symmetry of its penalties regarding the switch state; the incoming signal is always deflected exactly once regardless of the path followed in the DSB. In the cross (bar) state, signal at in1 enters (goes straight) the first ring and goes straight (enters) the second ring. The same holds for the signal entering at in2. The rightmost rings have also an isolating purpose, i.e., in the cross state, the residual signal from in1 (in the through port in the left ring), is not directly coupled to out1 (instead it is passed to an anti-reflective waveguide termination in rightmost ring) and contributes weakly to crosstalk in out1; crosstalk power is constrained to pass by 2 rings with penalty IL_t^{off} and IL_d^{off} respectively.

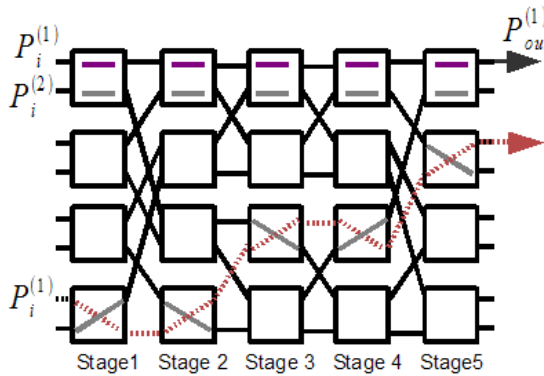


Figure 6 Benes network (8x8)

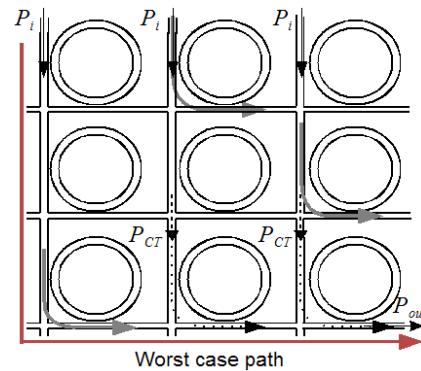


Figure 7 Crossbar (3x3)

PolITO analyzed three compound interconnection architectures: a crossbar interconnect, based on simple 1x2 SEs (see Figure 7), and two Benes networks (based on 2x2 BSBs and DSBs respectively, see Figure 6). These two architectures represent respectively two design targets for the interconnection architectures: the good physical-layer scalability and the minimum complexity in terms of number of rings.

Let define $X^{crossbar}$ and X^{benes} as the normalized crosstalk power, relative to the total output power at any output port for the respective architecture (in superscript), that is, $X^{crossbar, benes} = P_{xtalk} / P_{useful}$. This characterization was derived for each architecture from the basic SE model. With this scope, the scalability of these basic architectures was calculated assuming a transmitter, losses and receiver model (bitrate-dependent sensitivity model, and maximum noise target to guarantee a target BER of 10^{-12}).

After the assumption of single-wavelength transmissions and the non ideal switching states in all the switching blocks described before, the power leakage accumulated at the receiver can

be a limiting factor, since in-band signals from all inputs can affect the performance. The impact on the receiver was properly modeled in [3]; we used those results in our analysis, applying the relative crosstalk contribution $X^{\text{crossbar,benes}}$ to calculate the power penalty due to the presence of crosstalk at the receiver. In order to set an upper bound to the impact, the worst case of signal propagation was considered for Benes and crossbar architectures; examples are illustrated in Figure 6 and Figure 7 with dotted lines.

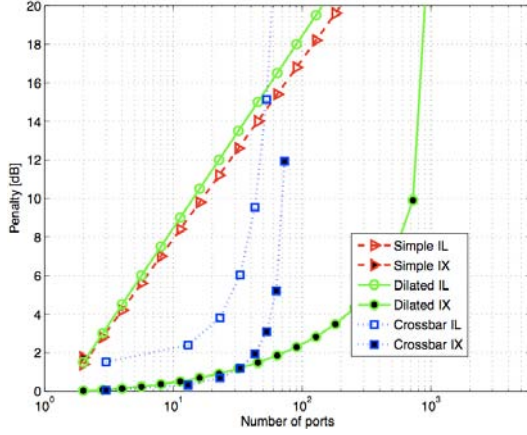


Figure 8 Penalties for Benes (BSB and DSB) and crossbar architectures as N increases; IL are insertion losses; IX are penalties due to coherent crosstalk

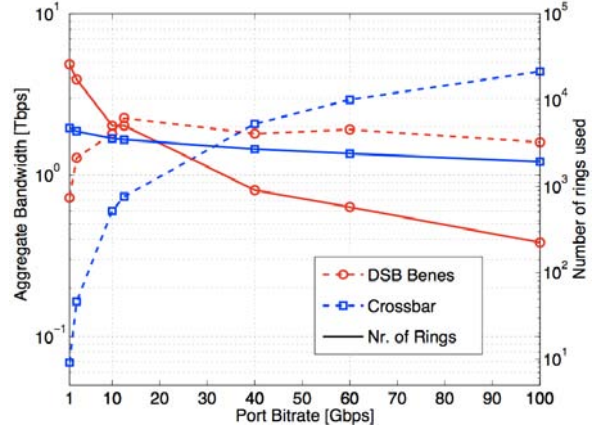


Figure 9 Aggregate bitrate and complexity scalability vs transmitter bitrate

The physical scalability allowed to obtain the maximum-capacity of feasible Benes (both for BSB and for DSB) and crossbar configurations was limited by the amount of crosstalk accumulated through all the fabric. Figure 8 shows the upper bound set by the crosstalk, when it becomes critical (crosstalk power becomes too high, thus a lot of signal power is required to distinguish ones from zeros). The Benes architecture made with BSBs corrupts the signal after the second BSB, so only a single point at $N=2$ is observed. There is also an advantage of using DSBs in Benes architectures because it weakly impacts the signal. Scalability is bounded to around 600 ports in presence of crosstalk for DSBs and around 70 ports for the crossbar. It is also important to note the advantage found at 40 Gbps (see Figure 9), where both crossbar and DSB-based Benes have very similar aggregate capacity (constrained by the bitrate-dependent sensitivity model assumed), but DSB-Benes employs less rings than the crossbar.

With this research work, which was published in [P1], PoliTO has proposed a model to evaluate feasibility of optical switching fabrics based on microring resonators switching elements; additional to the model, a dilated architecture was proposed to improve scalability and cost with respect to the crossbar switching fabric.

5.6.1.3 Optical interconnection networks based on microring resonators

From the architectural point of view, in [P2] PoliTO faced the scalability of optical switching fabrics with microring resonators by considering the bar and cross states of 2×2 SEs as “turn” (or High Loss State - HLS) and “straight” (or Low Loss State - LLS) states respectively. The target here is to find an architecture that minimizes the maximum number of blocks configured in HLS for all input-output permutations. This number, selected among all possible matchings in the fabric, is called X and characterizes the “impairment” introduced by the fabric in the worst case.

At this point, well-known switching theory results are recalled in order to classify the interconnection architectures according to their characteristics when using microring SEs as building element. An exhaustive procedure is applied to obtain the performance of small

classical Crossbar and Benes structures. Two performance metrics are considered: Scalability (in terms of the maximum number of ports) and complexity (in terms of the number of rings employed).

Thereafter, the scalability response of larger switching fabrics is explored. Each architecture is characterised by X depending on its size N (number of ports). The scalability limit is imposed by the target X_T , which means the maximum X allowed in the fabric (which can be determined a priori for a target BER at the receiver).

Based on these feasibility constraints, PoliTO defined construction rules based on several techniques: i) hybrid networks and ii) mirroring. The hybrid networks try to combine the low complexity of Benes architecture and the low signal impairment of crossbars; hybrid Benes crossbar (HBC) and the hybrid crossbar Benes (HCB), depicted in Figure 10 and Figure 11, use Benes architecture in outer and inner blocks respectively. They are characterized by the size of the inner blocks n (several outer blocks are required for a given N), i.e. $n=4$ and $N=16$ in Figure 10, so 2 additional stages of SEs are required; another possible configuration would be to use $n=8$ and only one stage of 2×2 SEs, but in the calculations, only the optimum n was used, which achieves minimum complexity and satisfies target X_T .

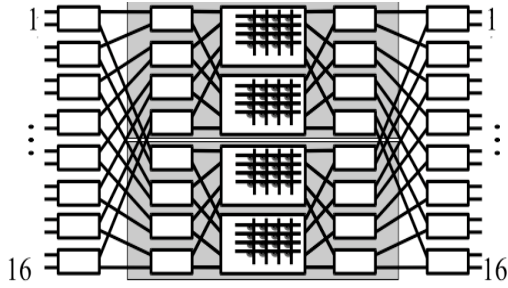


Figure 10 Hybrid Benes-Crossbar network

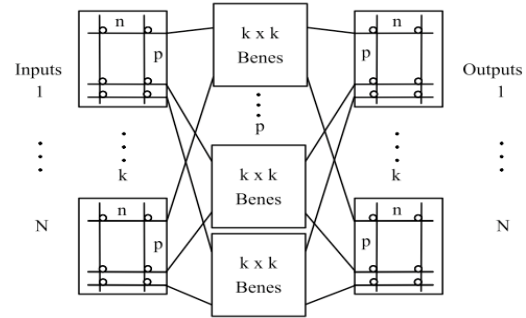


Figure 11 Hybrid Crossbar-Benes network

On the other hand, the mirroring technique defines a mirrored switching element (2M-SE) depicted in Figure 13, which exchanges HLS and LLS states. It is possible to build the whole switching planes with 2M-SEs (mirrored planes) configured in parallel with (normal) planes built with 2B-SEs (Figure 12). In this way, there is always an alternate route through any of the planes with a lower X than a normal plane alone. The idea behind this construction is depicted in Figure 14. Lets take X_{T-SP} as the target for a single-plane architecture; since 2M-SEs swap HLS with LLS and vice versa, all input/output connections that would cross $K \geq X_{T-SP}/2+1$ HLS 2B-SEs in the normal plane can clearly be routed on the mirrored plane, crossing $(X_{T-SP}-K)$ HLS 2M-SEs. Thus, by using the mirroring technique, the maximum number of HLS SEs which optical signal crosses reduces to $X_{T-SP}/2+1$ instead of X_{T-SP} as in the original single-plane architecture. Regarding complexity, the mirroring increases the complexity, since it duplicates the number of microrings used in single-plane and adds N rings for plane selection.

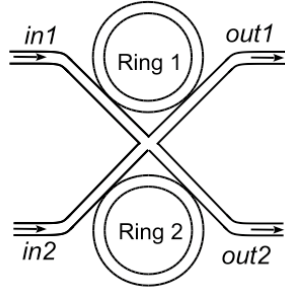


Figure 12 Basic 2x2 switching element (2B-SE)

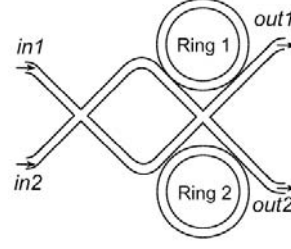


Figure 13 Mirrored switching element (2M-SE)

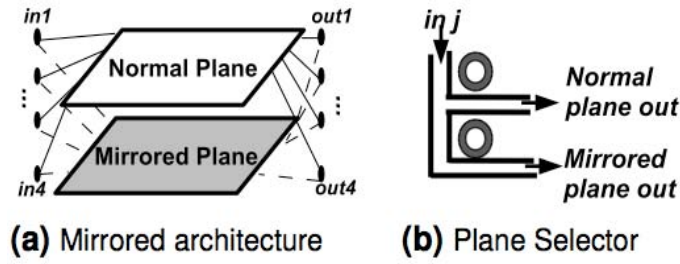
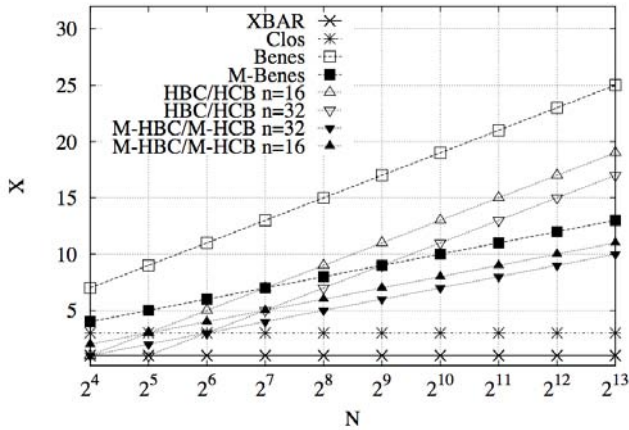
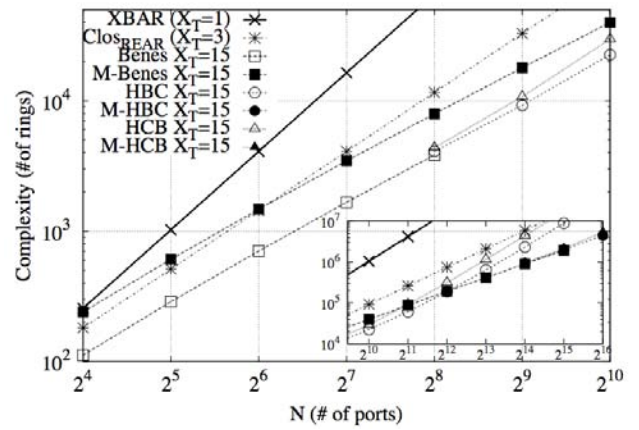


Figure 14 Mirroring technique

The mirroring technique can also be applied to the HCB and the HBC networks. In both cases, the mirroring is applied only to the Benes-like part of the networks, because mirroring the crossbar SEs would not provide any advantage.

PoliTO considered also Clos solutions in non blocking configurations composed entirely by crossbar blocks with size $n = p$ (use Figure 11 for reference), which is the optimal choice from the cost perspective [6].


Figure 15 X as a function of N for different networks. For the HBC/HCB, $n=[16 \ 32]$ ports.

Figure 16 Port scalability for crossbar, Benes, HBC and HCB networks for a target $X_T=15$ turns.

In Figure 15, the evolution of the physical impairment is characterized for the different architectures. Single plane architectures are identified by white markers while mirrored architectures are identified by black markers and the “M-” prefix.

Crossbar and Clos networks show a constant X equal to 1 and 3, respectively. On the contrary, the Benes network shows the worst scalability performance, because X scales

logarithmically with respect to the number of inputs. The HCB and the HBC solution show the same X scalability law. Results for the feasible configurations $N \times n$ are presented for the two different cases $n = 16$ and $n = 32$. The mirroring technique roughly reduces X by a factor of two. Thus, for a fixed X , it is possible to build interconnection networks with a number of ports equal to $N^2/2$ instead of N .

From the complexity point of view, it is possible to observe in Figure 1 the Benes and crossbar are lower and upper bounds respectively. While Benes can scale only to 2^8 ports due to the X_T constraint (see Figure 15), the mirrored Benes with the same target X_T can reach larger number of ports with a reasonable complexity. Hybrid solutions offer a better compromise since they scale to larger ports given that the target $X_T = 15$ allows them to make the internal Benes larger and benefit from low complexity.

In conclusion, this work describes, defines, analyses and presents several alternative solutions to the trade-off between complexity and scalability of interconnections using microring resonators. The new proposed techniques (mirroring, hybrid) are effective enough to optimize scalability and complexity. PoliTO set a starting point to the analysis and the design of this kind of microring-based networks. Finally, as future research lines PoliTO is thinking on the study of a microring model, physical impairments, integration with other passive devices and scheduling algorithms.

5.6.1.4 Simple and compound μ RR as Wavelength Selective Switches

Different optical configurations, simple and compound, based on ring resonators and using liquid crystals as the tuning or controlling block, have been designed [7] at UC3M. Specific attention is given to their application as tunable filters and wavelength switches, in this last case as part of complex matrix fabrics for all optical switching. Simulations of integrated optics ring resonators (see Figure 17) are developed using *FullWave*. In the compound configurations that use serial micro ring resonators it is implemented a 3 port reconfigurable demultiplexer in a compact cross-grid configuration. Theoretical 10 nm tuning is reported on SOI substrate with nematic liquid crystals. Optimization of the critical coupling condition by changing the evanescent coupling length is also reported. Theoretical analysis is presented to identify and emphasize the design parameters of each configuration as regards its application of interest. The reported structures can be developed in integrated optic technologies; all designs correspond to state of the art integrated optics technology.

But simple configuration has a very sharp wavelength response that can affect switching matrix performance due to their accumulated effect. More complex structure can be used to reshape filter response, so improving the overall switch performance. In order to solve this problem new optical filter synthesis methods can help for designing flatter responses, as the one reported in [8] using optical mirror ring resonator (OMRR)-based structures, as building blocks of infinite impulse response (IIR) filters. The method is based on cascading all-pole and all-zero structures, but using an all-pole second-order structure, made of a Sagnac interferometer in a ring resonator, named SIRR (see Figure 19). The main advantages of using the SIRR are that there is no need of phase shifters to adjust single μ RR in cascade and its immunity to variations in the ring lengths. This is a key factor to produce the same time delay in a second-order transfer function realized with cascaded first-order transfer functions. Finally, it has design flexibility, because this structure permits the synthesis of complex conjugated poles, located in any place, by varying either the optical path or only a coupling coefficient. An example of a filter with a full-width at half-maximum of 0.13 times the free-spectral range (FSR), with less than 2 dB of ripple and a crosstalk of 70 dB is shown in Figure

19.b. This is a fourth-order AR filter having real coefficients, so two SIRR stages are needed for implementation.

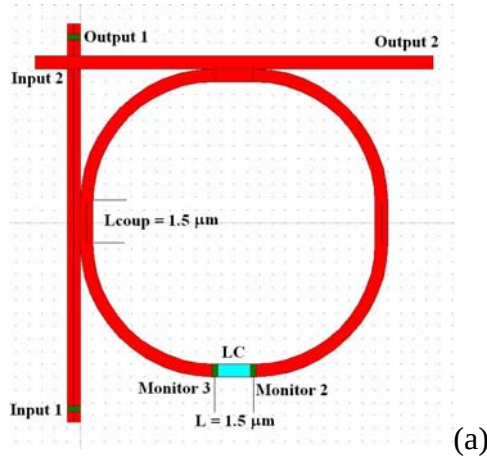


Figure 17 (a) Schematic of mRR with LC

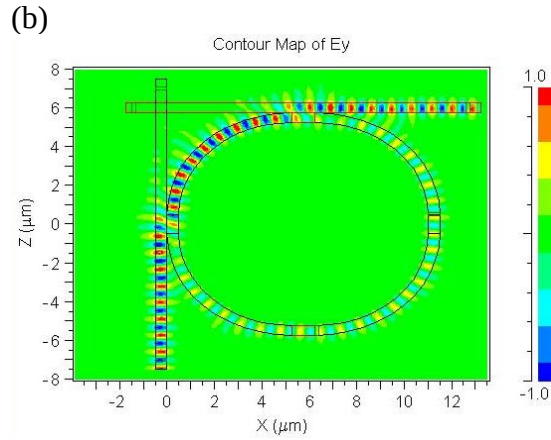


Figure 18 (b) Numerical simulations of the instantaneous electric field intensity pattern with LC ON at $\lambda = 1.564\text{mm}$ in cross-grid configuration

This structure along is a resonant cavity with a very sharp response in comparison with conventional μRR cavities so been proposed to be used as a very selective reflector [9].

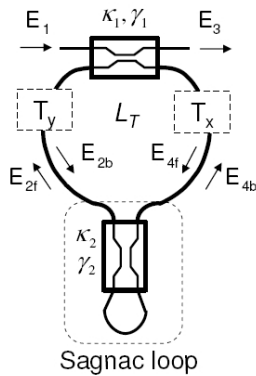


Figure 19 (a) SIRR Sagnac interferometer in a ring resonator

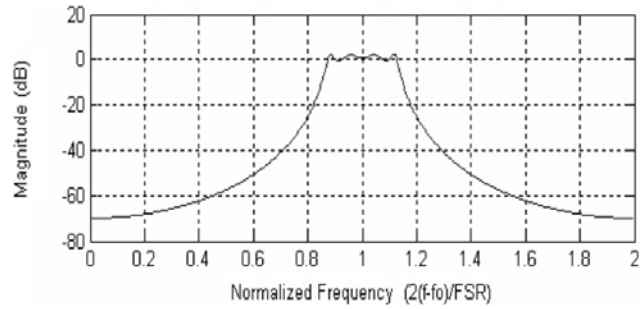
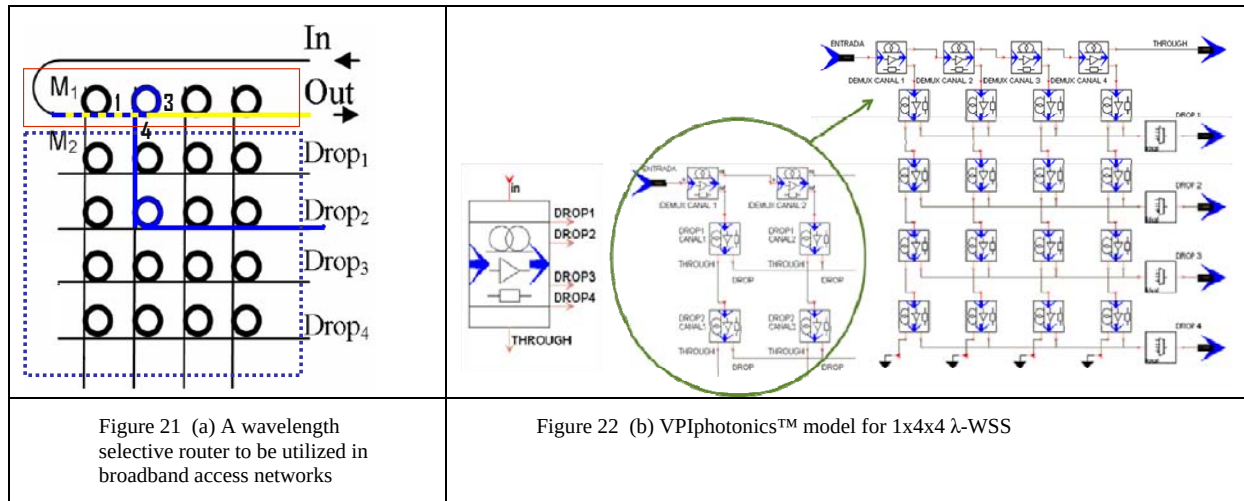


Figure 20 (b) Frequency response of fourth-order filter with SIRR. f_0 is the central frequency.

On the other hand, improving broadband access implies using optical switching elements able to operate at high bit rates and distribution of multicasting and selective signals with low cost. A specific configuration in a reconfigurable ring topology is reported in [10]. The channels in the network, each with a capacity of 1.25 Gbit/s, can be allocated to a single or a group of users via reconfigurable nodes that are controlled from the central node where the various up- and downstream signals are handled. For approaching this, the $1 \times 4 \times 4$ wavelength selective dynamic reconfigurable λ router based on μRR shown in Figure 21 is reported. The λ -router consists of a single 4- μRR de-multiplexer (the upper row) and a 4×4 μRR routing matrix. The de-multiplexer consists of four μRR s that drop up to four channels from the input port into the columns of the routing matrix. From here, by tuning the resonance wavelength of the μRR s in the routing matrix, the channels can be dropped onto the drop ports in any given combination.



A model in order to numerically evaluate the performance of the all-optical switch in such a wavelength router configuration using VPIphotonics™ has been developed at UC3M; a schematic of the different blocks can be seen in Figure 21 (b). First evaluations of improving performance using simple and compound structures are under development.

5.6.1.5 References

- [1] S. I. Association, International technology roadmap for semiconductors, [http://www.itrs.net/Links/2007ITRS/2007 Chapters/](http://www.itrs.net/Links/2007ITRS/2007%20Chapters/)
- [2] B. G. Lee, A. Biberman, P. Dong, M. Lipson, and K. Bergman, "All optical comb switch for multiwavelength message routing in silicon photonic networks", *IEEE Photonic Technology Letters*, Vol. 20, No. 10, May 2008, pp. 767–769.
- [3] H. Takahashi, K. Oda, H. Toba, "Impact of crosstalk in an arrayed-waveguide multiplexer on N×N optical interconnection", *J. of Lightwave Technology*, Vol. 14, No. 6, June 1996, pp. 1097–1105.
- [4] C. Li and A. Poon, "Silicon electro-optic switching based on coupled microring resonators", *Conference on Lasers and Electro-Optics (CLEO)*, May 2007.
- [5] B. G. Lee, A. Biberman, N. Sherwood-Droz, C. B. Poitras, M. Lipson, and K. Bergman, "High-speed 2x2 switch for multi-wavelength message routing in on-chip silicon photonic networks", *European Conference on Optical Communication (ECOC)*, May 2008.
- [6] A. Pattavina, *Switching Theory, Architectures and Performance in Broadband ATM Networks*, Wiley, 1998.
- [7] C. Vázquez, P. Contreras, J. Montalvo, J. M. Sánchez-Pena, Antonio d'Alessandro and D. Donisi, "Switches and Tunable Filters based on Ring Resonators and Liquid Crystals", *Proc. SPIE 6593: Photonic Materials, Devices and Applications II*, paper 65931F, U.S.A., 2007.
- [8] Salvador Vargas, and Carmen Vázquez, "Synthesis of Optical Filters Using Sagnac Interferometer in Ring Resonator" *IEEE Photonic Technology Letters*, Vol. 19, No. 23, December 2007, pp. 1877–1879
- [9] C. Vázquez, O. Schwelb, "Tunable, narrow-band, grating-assisted microring reflectors" *Optics Communications*, Vol. 281, No. 19, October 2008, pp. 4910–4916.
- [10] E.J. Klein, P. Urban, G. Sengo, L.T.H. Hilderink, M. Hoekman, R. Pellens, P. van Dijk, A. Driessen, "Densely integrated microring resonator based photonic devices for use in access networks", *Optics Express*, Vol. 15, No. 16, 2007, pp. 10346–10355.

5.6.2 Publications

- [P1] Hercules Simos (UoA), Adonis Bogris (UoA), Nikos Raptis (UoA), and Dimitris Syvridis (UoA), “Dynamic properties of a WDM switching module based on active microring resonators”, *IEEE Photonics Technology Letters*, accepted for publication.
- [P2] A. Bianco (PoliTO), D. Cuda (PoliTO), P. Giaccone (PoliTO), M. Garrich (UPC), R. Gaudino (PoliTO), G. Gavilanes Castillo (PoliTO), F. Neri (PoliTO), “Optical Interconnection Architectures based on Microring Resonators”, *Photonics in Switching 2009*, Pisa, Italy, September 2009.
- [P3] A. Bianco (PoliTO), D. Cuda (PoliTO), P. Giaccone (PoliTO), M. Garrich (UPC), R. Gaudino (PoliTO), G. Gavilanes Castillo (PoliTO), F. Neri (PoliTO), “Optical Interconnection Networks based on Microring Resonators”, submitted paper, *IEEE ICC 2010 – Optical Networks and Systems Symposium*, Cape Town, South Africa, May 2010.
- [P4] C. Vázquez (UC3M), I. Pérez (UC3M), P. Contreras (UC3M), B. Fracasso (GET), B. Vinouze (GET), chapter 9 “Liquid Crystal Optical Switches” in book entitled *Optical switches: materials and design*. Book editors: Soo Jin Chua and Baojun Li, U. Singapore © Woodhead Publishing Ltd to be published around May 2010

5.6.3 Mobility actions

Miquel Garrich (UPC) moved to PoliTO and collaborated in the study of microring-based architectures of interconnection networks, working towards his MS thesis; Feb 2009 to Oct 2009.

5.6.4 Y3 planned activities

PoliTO and UoA are planning to perform a joint research activity that consists in considering by simulation most of the transmission effects limiting optical interconnects using microring resonators and the evaluation of the system performance under dynamic operation. The target is to merge the complementary expertise and develop common simulation tools for the investigation of optical interconnects based on MRs. They have made a work plan which includes the architectures to be studied and the corresponding switching technologies.

PoliMI is also planning to contribute to JA6 in the next year, working jointly with PoliTO and other possible partners. PoliMI work is currently in its preliminary stages of investigation. The general reference scenario will be optical interconnection “inside the box”, possibly exploring over different interconnection distances ranging from on-chip to board-to-board.

As previously discussed, micro-ring resonators (in silicon or in other optical technologies) can be adopted as building blocks for optical switching matrices, starting from 1 x 2 and 2 x 2 MicroRing-based switching elements (MRSE). Nonetheless, various constraints and issues must be taken into account in order to create efficient structures. PoliMI activity will focus in particular the waveguides of the interconnection stages of MRSE switching matrices, which may cross over several times on the same substrate. Crossover points (i) represent practical realization difficulties, since techniques for their minimization add complexity to the final layout and (ii) introduce crosstalk between the optical signals that pass through the crossing guides. Although the magnitude of power penalty introduced by this waveguide discontinuity



is less than the one caused by the crosstalk within the MRSEs, passing through a high number of waveguide crossovers may contribute to a high level of crosstalk due to scattering. For these reasons, the crossover of the guides should be avoided (or their effects mitigated) when possible. Every feasible solution to this issue has a direct impact on the layout design of the switching structure. PoliMI will investigate different architectures that minimize the number of crossovers while taking into account physical fabrication constraints.

5.7 JA7: Hardware efficient optoelectronic switch fabric

Participants: UCAM, PUT, TUE

Responsible person: Jose Bernardo Rosas Fernandez (UCAM)

Deadline: Month 36

5.7.1 Description of the work carried out

During our studies we designed a new optical switching network architecture (called the $\log_2 N - 1$ switching fabric) which is build from asymmetrical optical switching elements. Each of these optical switching elements (OSE) consists of Semiconductor Optical Amplifiers (SOAs), splitters, and combiners (s/c). Each SOA amplify the optical signal which goes through the OSE. Each splitter splits optical signal from one its input to the proper number of its outputs. Each combiner combines all optical signals which appear at its inputs and send them as one signal to the its output. Thanks to different sizes of switching elements it is possible to construct a switching network of capacity N (N being the power of 2) but composed of fewer number optical elements than their counterparts in the baseline switching fabrics.

The new structure is cheaper than the traditional baseline networks. It has always fewer number of s/c and has equal (only for capacity 8) or fewer (for other capacities) optical amplifiers. It is especially important since the cost of SOAs is very high.

The new structure gives the same signal-to-noise ratio (SNR) for capacity 16 and better SNR for other capacities. Better SNR guarantees better quality of the optical signal at the switching network's output. Moreover, in the new switching architecture the optical signal goes through one stage less than in the typical baseline network of the same capacity. It is especially important because the number of SOAs an optical signal can goes through before regeneration is limited. Furthermore, regeneration is usually more expensive than amplification. Therefore, smaller number of stages has direct influence on the signal quality and the cost of the whole optical switch.

The new structure can be also nonblocking in strict-sense or rearrangeable when the proper number of planes is connected in parallel. It is called the multi- $\log_2 N - 1$ optical switching fabric. The new multi-plane structure consists of more planes than the traditional baseline network of the same capacity. But one plane in the $\log_2 N - 1$ is cheaper than one plane in the baseline fabric. So, totally the new multi-plane architecture is cheaper than traditional baseline networks and guarantee better quality of the optical signal at the switching network's output.

UCAM has been working in high speed interconnections using semiconductor optical amplifiers integrated switches. The first monolithically integrated 16x16 SOA crosspoint switch was presented at ECOC 2009 post-deadline session. The switch has a positive facet-to-facet gain, and a 1dB output saturation power of -11.5dBm. Routing of a 10 Gbit/s signal is achieved with a minimum power penalty of 2.5dB, showing its promising ability for high speed switching applications such a packet switching.

The monolithic integrated 16x16 switch used is based on a 3-stage hybrid Clos-Tree structure, shown in Figure 1. Three columns of four 4x4 port switching elements are connected by two

shuffle networks to create the 16x16 port switch. Each of the 4x4 switching elements comprises its own input and output shuffle networks and 16 SOA gates.

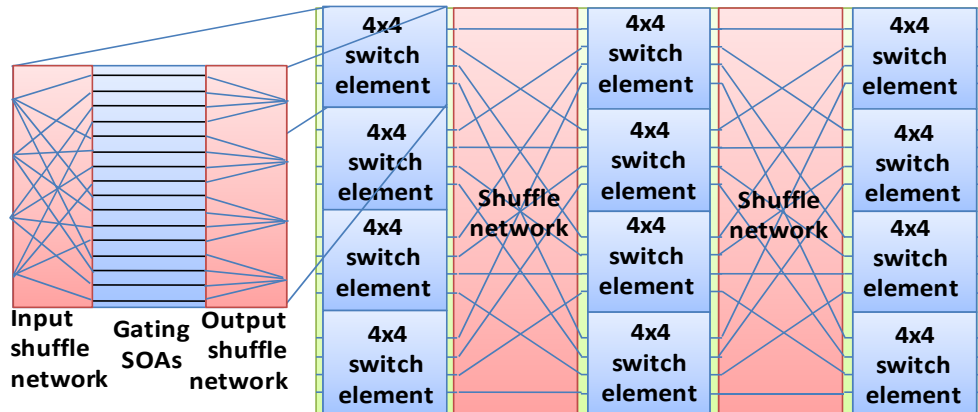


Figure 1 Schematic of 3-stage 16x16 switch architecture based on 4x4 switch elements

The switch is fabricated from all active material with a 6 QW AlGaInAs active region, grown on an InP substrate. The paths through the switch are constructed from 2 μ m ridge waveguides, whereas the SOA gates are tapered to 4 μ m wide ridges. The waveguide interconnections comprising the shuffle networks are constructed from deep etched total internal reflection (TIR) mirrors and beam splitters formed with tapered waveguides.

The waveguides in each shuffle network are biased to provide a small loss between successive SOA gates, to minimize the buildup of amplified spontaneous emission. The SOA gates provide routing functionality and sufficient gain to allow lossless operation of the entire switch.

It should be stressed that the current chip is re-growth free i.e. without active-passive interfaces. It is believed that the inclusion of, for example, passive shuffle networks would reduce ASE, nonlinear distortion and power consumption.

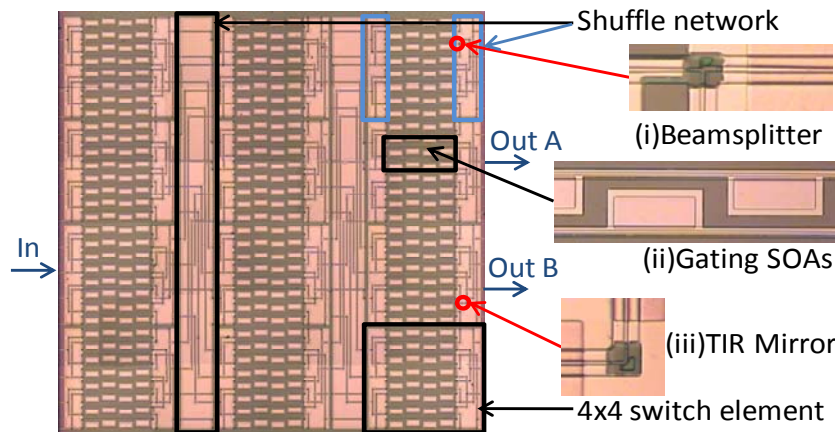


Figure 2 Schematic of the 16x16 switch showing the 4x4 switch elements and shuffle network sections (left). Insets show (i) waveguide beam splitter, (ii) gating SOA waveguides and (iii) TIR turning mirror.

Figure 2 shows the layout of the switch. Each path has three gating SOAs and goes through eight shuffle sections with a mean path length of 9mm. The switch contains 192 SOAs, 210 waveguide crossings, 288 splitters, 424 etched corner mirrors and has dimensions of 6.3mm x 6.5mm.

UCAM and Tue also worked in In this work therefore we report the first QD 1x8 port integrated switch, suitable for a variety of reconfigurable star, PON and switch applications. We demonstrate the potential for cascading this device to enable the construction of larger bespoke switch architectures.

The 1x8 switch, which is made of a five stack QD active layer embedded in a Q1.15µm InGaAsP separate confinement heterostructure⁶, comprises an input waveguide with a tree of 3dB couplers providing 8 output branches. Each of the output waveguides has separate electrical contacts, allowing independent control of each of the outputs. The input and output waveguides are angled to minimise back reflections

Experiment

A chip containing two 1x8 integrated switches, shown in figure 1A, is used as this allows cascaded performance of multiple devices to be investigated. 4 controllable lensed fibres are used, to allow simultaneous access of an input and output waveguide of both devices and to test different paths.

Figure 3 b shows the experimental setup in which a 10 Gbit/s 231-1 pseudo random bit sequence from a bit error rate test set (BERT) is used to modulate a Mach-Zehnder modulator whose optical input is from a tunable laser. The output from the modulator is then amplified with an EDFA, optically filtered and attenuated to provide a variable input power to the switch. The output is then filtered, and fed via a variable attenuator to an optically pre-amplified receiver allowing the performance of a 1x8 switch to be investigated using the error analyser on the BERT.

The output from the first optical switch can be looped back to the second 1x8 switch, allowing cascaded performance to be measured.

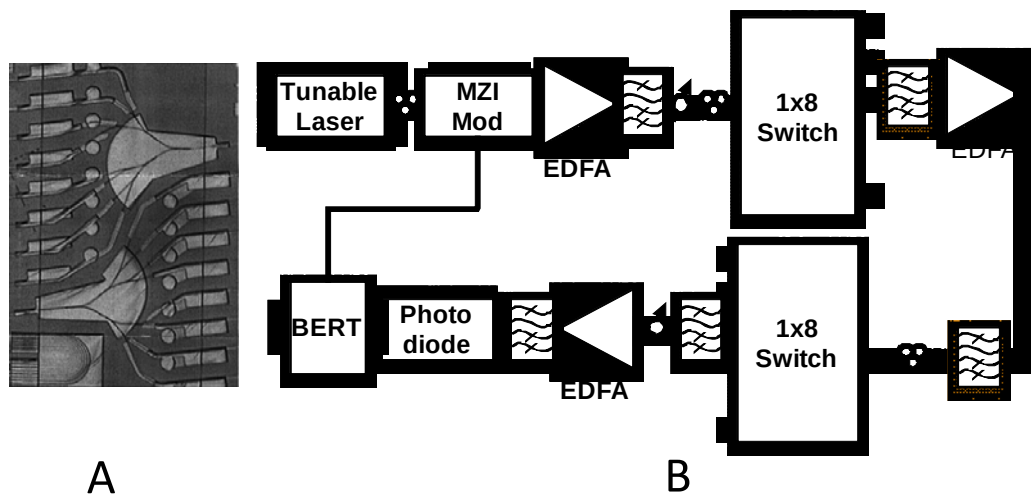


Figure 3 A: Chip schematic. B: Experimental schematic.

The electrical contacts of the optical switches are driven by a switchable power supply so that the temporal performance of the switch can be studied. In the case of a cascaded switch an EDFA and optical filter are used to compensate for the fibre coupling loss at the angled facets. The on chip performance of the switch has been characterised after removing fibre coupling losses, which are estimated from photocurrent measurements to be 12dB in this

device. The 1x8 switch has 3 cascaded 3dB splitters in each path, introducing 9dB of loss, and, together with some excess loss within the switch, this yields a net face.

UCAM also has been working in a A 4x4 port integrated SOA switch is used for space switching and wavelength conversion. This novel design requires no additional active components for the wavelength conversion. Low power penalties of 0.45dB are achieved for wavelength conversion at 5 Gbit/s, with error free operation at 7 Gbit/s in this un-optimised device.

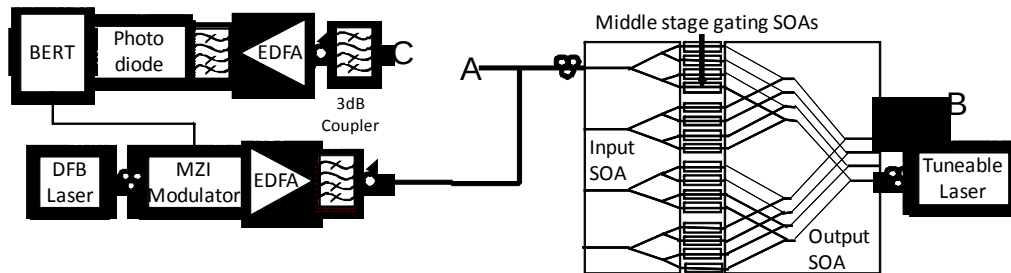


Figure 4 Experimental setup

The experimental setup is shown in Figure 4.

The switch is operated optimally with 200mA to the input waveguides, 400mA to the output waveguides, 25mA to a through-switching SOA and 70mA to a wavelength converting SOA.

The static transfer function of the switch is shown in Figure 5 with an on-chip input power of -5dBm and an optimal SOA current of 70mA. It can be seen that a 12dB extinction of the converted wavelength is easily achieved. A variation of less than 0.5dB is achieved over a 10nm converted wavelength range.

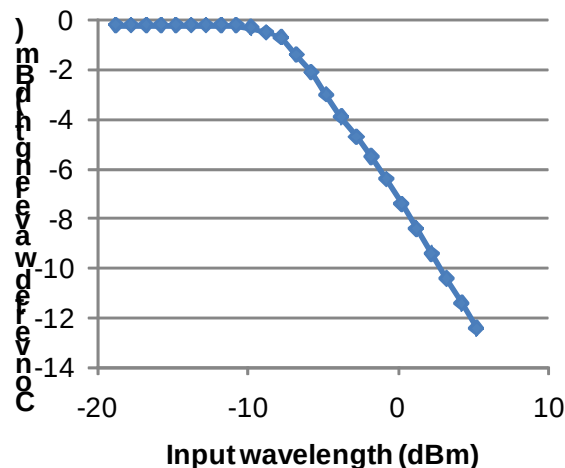


Figure 5 Static transfer function of the switch.

Very high line rate serial 160 Gbit/s data transmission has been demonstrated at TU/e through a semiconductor optical amplifier based multistage switching matrix. Active passive monolithic integration is implemented to ensure the optimum performance. This represents both the leading edge in monolithic switching circuit complexity with four cascaded switching stages and the highest reported line rates through monolithically cascaded switching networks. Bit error rate studies are performed to show error free operation and low

signal degradation. Power penalties of order 0.6dB and 1.2dB are observed for two stages and four stages respectively in the monolithic circuits at 160 Gbit/s per path.

5.7.2 Publications

- G. Danilewicz (PUT), W. Kabaciński (PUT), R. Rajewski (PUT), “The $\log_2 N - 1$ Optical Switching Fabric”, submitted to *IEEE Transactions on Communications*.
- H. Wang (UCAM), A. Wonfor (UCAM), K. Williams (TUE), R. Pentty (UCAM), I. White (UCAM), “Demonstration of a Lossless Monolithic 16x16 QW SOA Switch”, *ECOC 2009*, Vol. Postdeadline paper, September 2009. (JOINT).
- H. Wang (UCAM), K. Williams (TUE), A. Wonfor (UCAM), T. de Vries (TUE), E. Smalbrugge (TUE), Y. Oei (TUE), M. Smit (TUE), R. Noetzel (TUE), S. Liu (UCAM), R. Pentty (UCAM), I. White (UCAM), “Low Penalty Cascaded Operation of a Monolithically Integrated Quantum Dot 1x8 Port Optical Switch”, *ECOC 2009*, September 2009.
- H. Wang (UCAM), A. Wonfor (UCAM), S. Liu (UCAM), R. Pentty (UCAM), I. White (UCAM), “Novel integrated SOA space switch and wavelength converter”, *Photonic in Switching 2009*, September 2009.
- A. Albores-Mejia, K.A. Williams, F. Gomez-Agisz, S. Zhang, H.J.S. Dorren, X.J.M. Leijtens, T. de Vries, Y.S. Oei, M.J.R. Heck, L.M. Augustin, R. Nötzel, D.J. Robbins and M.K. Smit, “160Gb/s serial line rates in a monolithic optoelectronic multistage interconnection network”, *IEEE Symposium on High Performance Interconnects (HOTI)*, New York, 2009.

5.7.3 Mobility actions

- PhD student R. Rajewski from Poznan University of Technology visit to University of Cambridge – three or four weeks – for studying high-bandwidth and low-latency optical system architectures and their simulation in commercial software.
- Dr M. Żal (Poznan University of Technology) periodical visit to University of Cambridge – one week
- Dr R. Stabile: 19-23 October 2009. Financial support from the British Council and Beta Techniek. Joint experiments on switching network control
- Dr R. Stabile: Planned visit 8 to 26 March 2010



5.8 JA8: On-chip optical switching networks

Participants: PoliTO, PoliMI

Responsible person: Fabio Neri (PoliTO)

Deadline: Month 36

5.8.1 Description of the work carried out

Given the trend to increase the number of computing cores in future microprocessor chips, proper interconnections among these cores must be realized, leading to the so-called Networks on Chip (NoC).

Power constraints due to packaging and power budget limitations are among the most limiting factors in current electronic NoC (E-NoC) systems. In current prototypes with tens of cores the power dissipated by the E-NoC's accounts for over 25% of the overall power, and the power needed by a NoC implemented with current electronic circuit techniques is estimated to be too high (by a factor of 10) to meet the expected needs of future multicore systems.

Optical technologies can play interesting roles and complement low-power electronics in the realization of NoC's, for which they promise reduced energy consumption and smaller line delays. Given the above, this JA focuses on issues related to optical implementations of NoC's, also called O-NoC's (Optical Networks on Chip).

During the initial months of the BONE project, a collaboration was established between Politecnico di Torino and the Computer Science Department of Columbia University in New York, USA, on issues related to O-NoC's. The PoliTO's PhD student Michele Petracca moved to Columbia University for a thesis on O-NOC's, under the supervision of Luca Carloni and Keren Bergman. O-NoC architectures exploiting resonant micro-rings were carefully analyzed [1] in collaboration with important industrial partners, proposing new topologies and information routing schemes, and showing that a tenfold reduction of the power dissipated by data transfers in a multi-core chip when a large FFT is computed can be obtained in comparison with electronic technologies.

In the follow-up of that activity, the research efforts at PoliTO were centered on optical interconnection architectures for on-chip applications, always based upon microring resonators as switching technology, and designed according to crossbars and Benes interconnection architectures, similarly to what was reported in this document for BONE WP25 Joint Activity 6 ("Optical backplanes using microring resonators"). Complexity and physical scalability analysis have received most of the attention, but many research issues are arising due to this approach, ranging from the identification of high performance computing applications and their specific traffic patterns, to new discrete devices for off-chip operation in support of the O-NoC's. PoliTO believes that this initial investigation has been a useful step to keep on developing other structures for O-NoC's.

While several research topics are new and stimulating, the suitability of optical technologies for on-chip applications is still under ample debate, and it is not clear if O-NoC's will ever reach commercial maturity. Even the above mentioned researchers at Columbia have mitigated the initial enthusiasms on the effectiveness of optical technologies for in-chip applications, and consider the photonic domain advantageous only if jointly exploited for in-chip and chip-to-chip communications.

During the first year of WP25, PoliTO has been revising extensively the literature around NoC's, since transmission systems at the NoC scales require many specific considerations. Some of the findings are listed below.

Power consumption: Assessment of power consumption in O-NoC's is ongoing; previous similar analyses were done for optical switching fabrics, and PoliTO is considering how to apply them to NoC's in order to quantitatively assess the power consumption of on-chip networks in electronic and optical technologies, since power consumption is one of most frequently claimed advantages of O-NoC's over E-NoC's.

Application and topologies for O-NoC's: Several studies found that fully connected network topologies are largely over-provisioned for most applications [3]; mapping application-level communication patterns and logical topologies onto simpler interconnection topologies such as meshes, tora or stars, leads to non-trivial topology mapping and scheduling challenges [4].

Control Requirements: Until now, delays inside electronic NoC's have been identified as a major performance bottleneck. Hence, one of the possible drivers for O-NoC's is reducing delays in information delivery among cores; in particular, it was found out that:

- There is a reference distance below which the use of electronics is still convenient in terms of delay (since no E-O-E conversion is required); this distance is around tens of mm [4], which is still compatible with intra-chip distances. This sets an upper bound for Electronic-Optics-Electronic conversion delays for O-NoC's.
- Delay optimization in electronic interconnects has been achieved by improving the propagation delay in conductors (by reducing RC impedances); this however correspondingly reduces the bandwidth density, since smaller delays imply wider conductors [2]. At the moment, given the ongoing evolution of electronic interconnects with node technologies, the information density in electrical interconnects is still a moving target; as new node technologies are available, delay-optimized interconnects become more compact and keep competing with optical solutions, since they require less conductor spacing with respect to optical waveguides. WDM solutions are the most viable solution to this limitation, as they provide much larger information densities with low signal penalties [6].

Optical components: Not all kinds of optical components required for NoC's are suitable for integration; this poses limitations and trade-offs in their performance (losses, delay, and space occupied in silicon):

- The integration of optical sources is still under research, and several issues arise with the possibility of integration. The direct modulation (by modulating the polarization current) of the laser decrements its lifetime and reliability and makes its design more complex. In general, several studies [1],[9] consider laser integration still impractical for fabrication and operation, so off-chip reliable Continuous Wave (CW) sources are usually considered.
- Mach Zehnder interferometers [1],[7], ring resonator modulators [8], and p-i-n detector receivers have instead less limitations in integration, both for what regards their required size and polarization-related issues.

Resource utilization: The time required to establish an optical path, or in general to allocate a physical resource, is a critical performance parameter; since this time is not easy to reduce, optical interconnects need to maximize the duration of the use of optical resources, so as to emerge with higher performance than the electronic counterparts.

5.8.1.1 References

- [1] M. Petracca, B. G. Lee, K. Bergman, and L.P. Carloni, “Design Exploration of Optical Interconnection Networks for Chip Multiprocessors”, *16th Annual IEEE Symposium on High-Performance Interconnects (HotI)*, Stanford, Palo Alto, CA, August, 27th-28th, 2008, Vol. 16, pp. 31-40, DOI: 10.1109/HOTI.2008.20.
- [2] M. Haurylau, H. Chen, J. Zhang, G. Chen, N. A. Nelson, D. H. Albonesi, E. G. Friedman, P. M. Fauchet, “On-chip Optical interconnect roadmap: challenges and critical directions”, *IEEE Journal of selected topics in quantum electronics*, Vol. 12, No. 6, part 2, December 2006, pp. 1699-1705.
- [3] S. Kamil, A. Pinar, D. Gunter, M. Lijewski, L. Oliker, and J. Shalf, “Reconfigurable hybrid interconnection for static and dynamic applications”, *ACM International Conference on Computing Frontiers*, Ischia, Italy, May 7-9 2007.
- [4] G. Hendry, S. Kamil, A. Biberman, J. Chan, B. Lee, M. Mohiyuddin, A. Jain, K. Bergman, L.P. Carloni, J. Kubiawicz, L. Oliker and J. Shalf, “Analysis of Photonic Networks for a Chip Multi-Processor Using Scientific Applications”, *Third International Symposium on Networks-on-Chip*, May 10-13, 2009, San Diego, CA
- [5] The international technology roadmap for semiconductors (ITRS): <http://www.itrs.net>
- [6] Guoqing Chen, Hui Chen, Mikhail Haurylau, Nicholas Nelson Philippe M. Fauchet, and Eby G. Friedman, D. Albonesi, “Predictions of CMOS Compatible On-Chip Optical Interconnect”, *2005 International workshop on System level interconnect prediction*, San Francisco, California, USA
- [7] C.Gunn, “CMOS photonics for high speed interconnects”, *IEEE Micro*, Vol. 26, No. 2, March/April 2006, pp. 58-66.
- [8] Q. Xu, B. Schmidt, S. Pradhan, M. Lipson, “Micrometre-scale silicon electro-optic modulator”, *Nature*, Vol. 435, No. 7040, May 2005, pp. 325-327.
- [9] N. Kirman, M. Kirman, R. Dokania, J.F. Martinez, A. Apsel, M. Watkins, D. Albonesi, “Leveraging optical technology in future bus-based chip multiprocessors”, *IEEE Micro*, Vol. 27, No. 1, January/February 2007.
- [10] A. Bianco (PoliTO), D. Cuda (PoliTO), P. Giaccone (PoliTO), M. Garrich (UPC), R. Gaudino (PoliTO), G. Gavilanes Castillo (PoliTO), F. Neri (PoliTO), “Optical Interconnection Architectures based on Microring Resonators”, *Photonics in Switching 2009*, Pisa, Italy, September 2009.
- [11] A. Bianco (PoliTO), D. Cuda (PoliTO), P. Giaccone (PoliTO), M. Garrich (UPC), R. Gaudino (PoliTO), G. Gavilanes Castillo (PoliTO), F. Neri (PoliTO), “Optical Interconnection Networks based on Microring Resonators”, paper submitted to *IEEE ICC 2010 – Optical Networks and Systems Symposium*, Cape Town, South Africa, May 2010.

5.8.2 Y3 planned activities

Partners are working towards the identification of scenarios (applications, or systems-on-chip) in which O-NoC's are most suitable (or even mandatory), and propose several candidate architectures and/or reference applications to complement the proposals currently available in the literature. This identification is linked to the characterization of the proposed physical architectures and their feasibility assessment.

Furthermore, partners are working on the following targets:

- Architectures for Optical Networks On-Chip (O-NOC's)
- Protocols and access schemes suited for these O-NOC's



JA1: Identification of requirements of the optical backplane

Limitations and Requirements of Electrical and Optical Technologies for High-Capacity Interconnects

Document attached to D25.2 of BONE WP25

Abstract – We show a detailed analysis of systems where optical interconnects can be applied. The application targets are the high-performance network nodes that can more benefit from an optical implementation of the backbone subsystem. Such nodes may comprise switches, routers, digital or optical cross-connects, and add-drop multiplexers with different number of ports and different data rates. In light of these requirements, we will identify potential bottlenecks of “traditional” electronic interconnection systems, in terms of scalability towards the upgraded performance required by the future networks. This document will provide an overview on the state-of-the-art structures and technologies in electronic switching systems to implement interconnections. It will then analyze possible limitations of these solutions in front of new requirements and show in which cases and how optics can offer a real competitive advantage to equipment providers.

Index Terms – Optical interconnects, electrical interconnects, scalability limitations,

I. Introduction

Historically, printed circuit board (PCB) assemblies were dominated by parallel communication technologies for many years. This type of interface requires large banks of parallel line transmitters and receivers which cause alignment problems for outside communication. Clock skew, data skew, rise and fall times as well as jitter limit the introduction of higher data rates and speeds. As a result, serial point-to-point technology was introduced, initially for box-to-box communication and afterwards for backplane communication. The SERDES (“serializer/deserializer”) receives parallel data from PCB parallel busses and allows transmission outside electronic boards of high data rates over a serial link. Smaller pin counts, less expensive connectors and better signal integrity at larger data rates are also ensured.

SERDES chips have become over time a very important building block for high-performance computing systems, including switches and routers. The need for driving long lines brings significant power consumption, and SERDESes often dominate the power consumption of switch boards.

II. Overview of Electrical Interconnection Technologies

The continuous exponential reduction in feature size of ever faster silicon devices has led to an increase in interconnection requirements within large-scale systems. More functionality on boards needs more interconnects between boards that are able to operate at higher data rates. Due to this denser packing, electrical interconnects are more vulnerable to cross-talk and the need for higher bandwidth is obstructed by increased dielectric losses. Consequently, the overall system performance scalability depends on how well the internal interconnection system scales.

In order to overcome this electronic bottleneck and meet growing demand for both high speed and bandwidth, standardization bodies and forums such as IEEE Ethernet Working Group, OIF, Rapid TA, PCI-SIG, OBSAI and CPRI have developed various electrical interconnecting standards and protocols with improved signal integrity specifications, line coding formats as well as design techniques such as pre-emphasis and equalization. High-speed electrical signaling has also experienced considerable enhancements. These technologies are in the process of being adopted by some system and chip vendors.

The following table briefly outlines most typical electronic technologies and summarizes some important physical differences.

Name	Nr. of diff. pairs		Line Rate [Gb/s]				Total Rate [Gb/s]				Payload Rate [Gb/s]				Maximum Transmit Distance					
	TX	RX													PCB [cm]	Optic [m]	Copper [m]			
XGMII	32	32	0.313				0.3125x32 =				10				<7					
XAUI	4	4	3.125				12.5				(3,125x4)x(8/10) = 10				< 50	<300	<15			
XLGMII	32	32	1.25				1.25x32=				40									
XLAUI	4	4	10.313				41.25				(10.3125x4)x(64/66)= 40				25					
CGMII	32	32	3.125				3,125x32=				100									
CAUI	10	10	10.313				10.313				(10.3125x10)x(64/66)= 100				25					
			Lane Rate [Gbaud/s]																	
CEI-6G-LR	Duplex connection not defined		4.976 - 6.375				No specific data pattern (coding scheme) defined Total Rate= Payload Rate				4.976 - 6.375				0 - 100					
9.95 - 11.1				9.95 - 11.1							0 - 100									
28				28							30									
CEI-28G-SR	FFE DFE		25								25				68.58					
Interlaken	Any Nr. of Lanes		3.125, 6.25 and 10.3125				Depends on PHY specif.													
--> Exp:	4	4	6.25				6.25x4=				25				0 - 100					
FC-XAUI	4	4	3.1875				12.75				(3.1875x4)x(8/10)=10.2				< 50	<300	<15			
SPI-4.1	64	64	0.15625				0.15625x64=				10									
SPI-4.2	16	16	0.622				0.622x16=				9,952									
SPI-5	16	16	2.488 - 3.125								40									
SFI-4.1	16	16	0.622				0.622x16=				9,952									
SFI-4.2	4	4	2.488				2.488x4=				9,952									
SFI-5	16	16	2.488 - 3.125				39.8 - 50				39.8				20.32					
TFI-5	unidirectional		2.488 or 3.1104												76.2		<100			
InfiniBand			SDR	DDR	QDR	EDR*	SDR	DDR	QDR	EDR*	SDR	DDR	QDR	EDR*		SDR	DDR	SDR	DDR	
	1	1	2.5	5	10	20	2.5	5	10	20	2	4	8	16	73.5			20		
	4	4	2.5	5	10	20	10	20	40	80	8	16	32	64		200	200	15	10	
	8	8	2.5	5	10	20	20	40	80	160	16	32	64	128						
	12	12	2.5	5	10	20	30	60	120	240	24	48	96	192		200		10	7	
sRIO	1x, 2x, 4x, 8x, 16x Lane Payload Rate -->		1.25	2.5	3.125	5	6.3									<50				
			1	2	2.5	4	5									<60				
RIO	8 or 16	8 or 16																		
PCIe (PHY)	1x, 2x, 4x, 8x, 16x, 32x		0.25; 0.5 and 1				max 16				12.8				75 - 100		<7			
SAS	4	4	3 (2x1.5) and 6		1.5x(8/10)=1.2		12 or 24								100		8			
SATA	4	4	3 (2x1.5) and 6		1.5x(8/10)=1.2		12 or 24								200		1			

* future standard

Table 1: Overview of electronic interconnecting technologies

A. *State-of-the-art of electrical interconnection technologies*

In what follows, we present a short overview of some physical and logical specifications of the state-of-the-art electrical interconnection technologies. Each technology is unique but they all have a layered architecture. The implementation of physical layer has more or less fixed functionality where upper layers are more customizable and can vary greatly from one technology to another.

II-A-1 10GbE XAUI

10 Gigabit Ethernet 0 family of standards was first published in 2002 as IEEE Std 802.3ae-2002 and is the fastest of currently deployed Ethernet Working Group standards, continuing the natural evolution of Ethernet in speed and transmission distance.

Structure

10GbE only operates in full-duplex mode, meaning that collision detection is not necessary. In other aspects, 10GbE remains loyal to the original Ethernet model. As a Layer 2 protocol, it uses the IEEE 802.3 Ethernet Media Access Control (MAC) protocol, the IEEE 802.3 Ethernet frame format, and the minimum and maximum IEEE 802.3 frame size 0.

The Ethernet architecture consists of PHYSical Layer, which corresponds to Layer 1 of the OSI model and MAC Layer. An Ethernet physical layer device connects through optic or copper to the MAC Layer (OSI Layer 2). Physical layer is further divided into PCS (Physical Coding Sublayer), PMA (Physical Medium Attachment) and PMD (Physical Medium Dependent) sublayers. Optical transceivers, for example, are PMDs.

The 802.3ae specification defines two PHY types: the LAN PHY with payload rate of 10 Gb/s and the WAN PHY with payload rate of 9.29 Gb/s. Physical specifications cover various media types for single-mode fiber (long haul), multi-mode fiber (up to 300 m), copper backplane (up to 1 m) and copper twisted pair up to 100 m) as shown in Table 2 0. The options highlighted in yellow can be used for inter-module interconnects.

Device	Standard	Media Type	Distance [m]
10GBASE-SR	802.3ae	MMF	26-82
10GBASE-LX4	802.3ae	MMF with WDM	240-300
10GBASE-LX4	802.3ae	SMF	10 000
10GBASE-LR	802.3ae	SMF	10 000
10GBASE-ER	802.3ae	SMF	40 000
10GBASE-CX4	802.3ak	Copper cable	<15
10GBASE-KX4	802.3ap	Copper backplane over 4 lanes	1
10GBASE-KR	802.3ap	Copper backplane over 1 lane	1
10GBASE-SW	802.3ae	MMF	26-821
10GBASE-LW	802.3ae	SMF	10 000
10GBASE-EW	802.3ae	SMF	40 000

Table 2: Overview of 10GbE standard options, media types and supported distances

There are several interfaces defined in the 10GbE standard in order to interconnect different parts of a 10 Gb/s Ethernet system. For example, the XGMII interface is located between MAC and PHYsical Layer providing full duplex operation at 10 Gb/s with each line operating at 0.3125 Gb/s. It is 74 signal wide interface and consists of 32-bit data path in each direction as well as control and clock signals. Because of its broad interface and short recommended distance of 7 cm, chip-to-chip, board-to board and chip-to-module applications are not feasible with this interface.

10 Gigabit Attachment Unit Interface, XAUI is designed to resolve these issues. Conversion between the XGMII and XAUI interfaces takes place at XGXS (XGMII Extender Sublayer).

The major functions of the XGXS are the following 0:

- The optional XGXS can be inserted between the Reconciliation Sublayer and the PHY in order to extend the physical reach i.e. operational distance of the XGMII and reduce the interface pin count.
- The XGMII is organized into four lanes with each lane conveying a data octet (8) or control character on each edge of the associated clock. The XGXS converts bytes on an XGMII lane into a self clocked, serial, 8B/10B encoded data stream. Each of the four XGMII lanes is transmitted across one of the four XAUI lanes.
- The source XGXS converts XGMII Idle control characters (interframe) into an 8B/10B code sequence. The destination XGXS recovers clock and data from each XAUI lane and deskews the four XAUI lanes into the single-clock XGMII.
- The destination XGXS adds to or deletes from the interframe as needed for clock rate disparity compensation prior to converting the interframe code sequence back into XGMII Idle control characters.
- The XGXS uses the same code and coding rules as the 10GBASE-X PCS and PMA.

Physical Specification

XAUI is full duplex interface that uses 4 self-clocked serial differential links in each direction, i.e. 8B/10B encoded data streams consisting of data and timing clock run over four serial lanes each operating in continuous mode at 3.125 Gb/s. Total aggregate bandwidth of 12.5 Gb/s (10 Gb/s useful payload with 8B/10B encoding overhead) is achieved. Hence, XAUI reduces 10 GbE's 72-pin XGMII interface to 16 pins. Due to self-clocked nature of the interface, skew problems between clock and data are eliminated and transmit distance of 7 cm is extended for another 50 cm over standard PCBs. The actual performance of an XAUI interconnect is highly dependent on the implementation. Loss and jitter budgets are standardized and listed in Table 3 [2].

	Loss (dB)	Differential skew (ps _{p-p})	Total jitter (UI _{p-p})	Deterministic jitter (UI _{p-p})
Driver	0	15	0.35	0.17
Interconnect	7.5	60	0.20	0.20
Crosstalk, Noise, etc.	4.5		0.10	0.10
Total	12	75	0.65	0.47

Table 3: Loss and jitter budgets of the XAUI interface

Logical Specification

As mentioned above, XAUI employs the 8B/10B transmission code so providing a high-level of signal integrity. Each serial transmission lane operates independently which leads to appearance of lane-to-lane skew. In order to achieve line alignment, an extra control word defined for alignment is used. Frame synchronization is done with the aid of some extra code groups for control signaling. XAUI functions as a self-managed interface maintaining frame synchronization and lane alignment without support of upper layers [3].

Moreover, XAUI compensates clock differences that often appear between transmitter and receiver by adding or deleting specific control words. The XAUI interface is universal interface introduced in many vendor-specific products. Through extended physical separation between MAC and PHY, the interface supports connection between line cards within the system chassis.

II-A-2 FC XAUI

The Fibre Channel (FC) [4] is logically a bi-directional point-to-point serial data channel and physically an interconnection of one or more point-to-point links. In the case of FC, fiber is a general term for all supported media including optical fiber, twisted pair and coaxial cable.

Structure

10GFC has the same layered protocol structure as original FC with some additional functionality within each layer. Fibre Channel structure is depicted in Figure 1.

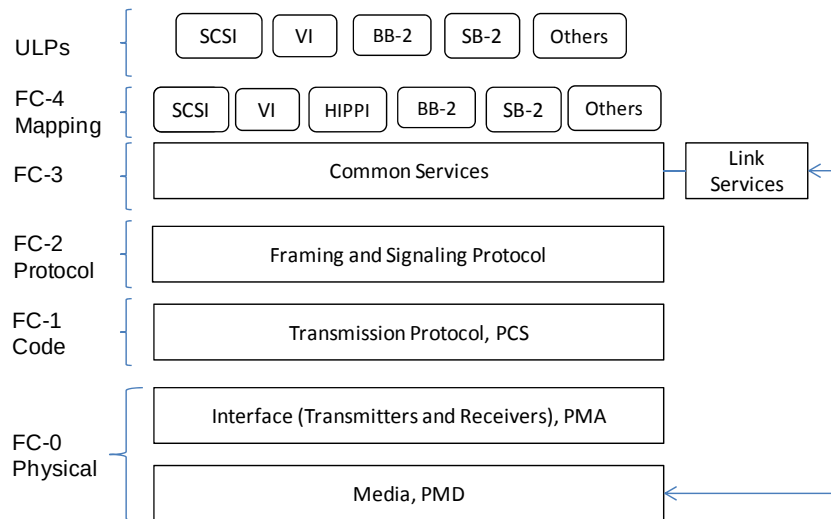


Figure 1 Structure of Fibre Channel

In general, Fibre Channel consists of five layers. The three lower layers are layers of interest for us. FC-3 level of 10GFC is extended by port management interface and register set and low level signalling protocol. In order to support a data rate of 10.2 Gb/s over physical variants of 10GFC, FC-2 level of 10GFC extends the FC-2 levels of FC-FS and FC-AL-2. Transport at 10.2 Gb/s is provided by the FC-1 level of 10GFC which adopts FC-2 level functions to the XGMII interface as defined by

10GbE. 10GFC FC-1 level specifies optional electrical interfaces XGMII and XAUI which may be used to interconnect the physical embodiment of various sublayers. 8B/10B and 64B/66B PCS sublayers are also functions of 10GFC FC-1 level. The FC-0 Level of 10 GFC describes Fibre Channel link, i.e. different media types and associated transmitter and receiver. Serial physical full-duplex variants specified to support the transport of encoded FC-1 data over fiber-optic medium include [4]:

- four quarter speed lanes over individual fibers
- four quarter speed lanes optically multiplexed over a single fiber and
- one full speed lane over one fiber.

The optical interfaces (PMDs) for all physical variants are included in the 10GFC specification [4]:

- 850 nm Parallel (four lane) PMD, type 1200-MX-SN4P-I, specified in 10GFC;
- 850 nm Serial PMD, type 1200-MX-SN-I, specified in 10GbE;
- 850 nm CWDM (four wavelength) PMD, type 1200-MX-SN4-I, specified in 10GFC;
- 1310 nm Serial PMD, type 1200-SM-LL-L, specified in 10GbE;
- 1310 nm CWDM (four wavelengths) PMD, type 1200-X-LC4-L, specified in 10GbE.

The 8B/10B PCS and PMA sublayers are intended for transport over PMD mentioned above at cumulative data rate at 10.2 Gb/s. The 10GFC PCS and PMA sublayers are generally specified in FC-FS and 10GbE with some extensions. They operate at signalling rate of 3.1875 Gbaud +/- 100 ppm, exactly 2% higher than the PCS and PMA specified in 10GbE.

The 64B/66B transmission code PCS sublayer is intended for 10.2 Gb/s data transport across a single mode fiber. The 10GFC 64B/66B transmission code operates at a signalling rate of exactly 2% higher than the 64B/66B transmission code specified in 10GbE.

Physical specification

The purpose of the XAUI interface is to offer an optional electrical interconnect. XGXS and XAUI can be used to physically partition FC-1 and FC-0 elements and are fully specified in 10GbE.

10GFC XAUI interface is a four serial lane interface where 8B/10B coding and operation at 3.1875 Gb/s +/-100 ppm, exactly 2% higher than the XAUI specified in 10GbE are performed on a per lane basis.

The implementation of the interface is an extended chip-to-chip interface implemented with traces on a printed circuit board. XAUI may be used as the electrical interface to optical PCS devices.

II-A-3 CEI-6G

The Common Electrical I/O CEI-6G [5] was developed by OIF (Optical Internetworking Forum) with aim to double the capacity of conventional backplanes from 2.5-3.125 Gb/s to 4.976 to 6.375 Gsym/s with low BER < 10^{-15} .

The Common Electrical I/O implementation agreement specifies the transmitter, receiver and interconnect channel for 6 Gb/s interfaces as well as jitter and measurement methodologies for application in high-speed backplanes, chip-to-chip interconnects and optical modules. It defines applicable data characteristics (e.g. DC balance, transition density, and maximum run length), channel

models and compliance points/parameters supporting the physical reach and conditions. Pin-out, management interface, power-supply specification, connector or higher-level activity such as addressing or error control is not specified. Furthermore, no particular data protocol is favoured or specified.

The CEI Implementation Agreement includes [5]:

- Electrical and jitter methodologies for new high speed interfaces and including the following older OIF interfaces: SxI-5, SFI-4.2, SFI-5.1, SPI-5.1 and TFI-5.
- A CEI-6G-SR specification for: Data lane(s) that support bit rates from 4.976 to 6.375Gsym/s over Printed Circuit Boards. Physical reach from 0 to 200mm and up to 1 connector.
- A CEI-6G-LR specification for Data lane(s) that support bit rates from 4.976 to 6.375Gsym/s over Printed Circuit Boards. Physical reach from 0 to 1m and up to 2 connectors.

The signal paths or CEI lanes are unidirectional point-to-point connections. Each CEI Lane is made up of a balanced differential pair. A CEI Link can be comprised of a unidirectional single lane or parallel lanes in either transmit or receive direction. A CEI Link does not imply duplex operation.

Rather than specifying materials, channel components, or configurations, the CEI focuses on effective channel characteristics. Hence, a short length of poorer material should be equivalent to a longer length of premium material. A ‘length’ is effectively defined in terms of its attenuation rather than physical length. The transmitter is specified in terms of its ability to pre-equalize the transmit signal and the receiver must work to the given BER using a compliant transmitter and channel.

In order to achieve the bit rate of up to 6 Gb/s, electrical parameters and attributes common to all links are defined. These are data pattern, signal levels, BER, ground differences, cross talk, transmitter lane-to-lane skew and so on. Although IA (Implementation Agreement) does not have any specific data patterns (i.e. 8B/10B, 64/66B, SONET scrambling, stream cipher, raw data, etc.) there are some necessary requirements regarding this issue in order to insure proper operation. If all conditions are not met, then the link may not work to the full distance, or meet the BER, or in fact work at all. If a fixed block coding scheme is used (e.g. 8B/10B), the input data must be either scrambled before coding or the coded data must be scrambled prior to transmission.

CEI-6G-LR

CEI-6G-LR long-reach high speed electrical interface between nominal baud rates of 4.976Gsym/s to 6.375Gsym/s using NRZ coding (hence 1 bit per symbol at the electrical level) with distance of up to 1 m is solution of interest for backplane applications. The electrical interface is based on high speed, low voltage logic. Connections are point-to-point balanced differential pair and signalling is unidirectional. The electrical IA is based on loss & jitter budgets and defines the characteristics required to communicate between a CEI-6G-LR transmitter and a CEI-6G-LR receiver using copper signal traces on a printed circuit board.

A compliant device must meet the following requirements [5]:

1. Support serial baud rate from 4.976Gsym/s to 6.375Gsym/s.
2. Capable of low bit error rate (required BER of 10⁻¹⁵).
3. Capable of driving 0 – 1m of PCB (such as IEEE 802.3 XAUI/TFI-5 compliant backplane) and up to 2 connectors.
4. Shall support AC-coupled operation and optionally DC-coupled operation.

5. Shall allow multi-lanes (1:N).
6. Shall support hot plug.

II-A-4 CEI-11G

CEI-11G is further specification developed by OIF and also specified in the IA.

The CEI Implementation Agreement includes [5]:

- Electrical and jitter methodologies for new high speed interfaces and including the following older OIF interfaces: SxI-5, SFI-4.2, SFI-5.1, SPI-5.1 and TFI-5.
- A CEI-11G-SR specification for: Data lane(s) that support bit rates from 9.95 to 11.1Gsym/s over Printed Circuit Boards. Physical reach from 0 to 200mm and up to 1 connector.
- A CEI-11G-LR(MR) specification for Data lane(s) that support bit rates from 9.95 to 11.1 Gsym/s over Printed Circuit Boards. Physical reach from 0 to 1m and up to 2 connectors.

CEI-11G-LR/MR

CEI-11G-LR/MR is a long-reach high-speed electrical interface between nominal baud rates of 9.95 Gsym/s to 11.1 Gsym/s. It uses NRZ coding on distances of up to 1 m making it suitable for use in backplane applications. The electrical interface is based on high speed, low voltage logic. Connections are point-to-point balanced differential pair and signalling is unidirectional. The electrical IA is based on loss and jitter budgets and defines the characteristics required to communicate between a CEI-11G-LR transmitter and a CEI-11G-LR receiver and between a CEI-11G-MR transmitter and a CEI-11G-MR receiver, using copper signal traces on a printed circuit board.

CEI-11G-LR as well as CEI-11G-MR devices from different manufacturers shall be inter-operable. The transmitter is specified in terms of its ability to pre-equalize the transmit signal and the receiver must work to the given BER using a compliant transmitter and channel. The CEI-11G-LR implementation agreement is optimized for overall cost-effective system performance including total power dissipation. Low power option is CEI-11G-MR which is based upon following [5]:

- A channel compliance specification is defined stricter than that of CEI-11G-LR.
- CEI-11G-MR uses the same Transmitter device as is specified for CEI-11G-LR with certain functions defined otherwise as optional.
- CEI-11G-MR uses a Receiver device that is similar to the device specified for CEI- 11G-SR but with extended Output Differential Voltage range.

A compliant device must meet the following requirements [5]:

1. Support NRZ coded serial data rate from 9.95 Gsym/s to 11.1 Gsym/s.
2. Capable of low bit error rate (required BER < 10⁻¹⁵).
3. Capable of driving 0 — 1 meter (39 inches) of PCB and up to 2 connectors.
4. Capable of driving 0 — 600 mm of PCB and up to 2 connectors for low-power applications.
5. Shall support AC-coupled and optionally DC-coupled operation.
6. Shall allow multi-lanes (1 to n).
7. Shall support hot plug.

II-A-5 Interlaken

Interlaken is a scalable, chip-to-chip interconnect protocol designed to enable transmission speeds from 10 to 100 Gbps and beyond. Interlaken can run on multiple media such as FR4 (PCB), backplanes, or over cable and it can be used in a variety of applications:

- Framer/MAC to NPU or L2/L3 switch interface
- Line Card to Switch Fabric Interface.

Interlaken is characterized by the following features [6]:

- Support for 256 logical communications channels, or up to 64K with channel extension
- A simple control word structure to delineate packets, similar in function to SPI 4.2 (see subsection II-A-9)
- A continuous Meta Frame of programmable frequency to guarantee lane alignment, synchronize the scrambler, perform clock compensation, and indicate lane health
- Protocol independence from the number of serdes lanes and serdes rates
- Both out-of-band and in-band per-channel flow control options, with a simple Xon/Xoff semantic
- 64B/67B data encoding and scrambling
- Performance that scales with the number of lanes

Since Interlaken has no upper limit on the bandwidth range although it is primarily targeted for 10 Gb/s to 100 Gb/s connections, it is adequate for wide range of applications and allows backward compatibility of existing devices. It is highly scalable and flexible protocol due to its ability to run over varying number of lanes, hence increasing connection capacity simply by adding more lanes or by increasing data rate per lane. Consequently, scalability is determined by number of lanes and frequency of each lane. Interlaken combines the channelization and flow control attributes of SPI 4.2 with the long reach and low pin count requirements of XAUI.

Impact of bit errors is minimized by using strong cyclic redundancy check (CRC) protection for each transfer. A CRC24 polynomial was chosen to ensure reliable communication [6].

Interlaken Protocol is defined by two fundamental structures: the Data Transmission Format and the Meta Frame [7]. The data transmission format relies significantly on the concepts of SPI 4.2. Data is segmented into bursts representing subsets of the original packet data. Each burst is bounded by two control words and associated with a logical channel, which can represent a physical networking port in the system or some other logically connected stream of data. Packet data is transmitted sequentially via one or more bursts whose size is a configurable parameter. By segmenting the data into bursts, the interface allows the interleaving of data transmissions from different channels for low-latency operation. The Meta Frame is defined to support the transmission of the data over a SerDes infrastructure. It contains a set of four unique control words, which are defined to provide lane alignment, scrambler initialization, clock compensation, and diagnostic functions. The Meta Frame uses the specific formatting of the control words to distinguish it from the data.

Physical Specification

Interlaken is specified as a multi-lane full-duplex interface, using differential pairs connected to Serdes. As mentioned above, the protocol scales from one to any number of lanes that are practical to allocate on a single IC. Current serdes technology allows for operation from 1-6.375 Gbps, including the widely-deployed XAUI rate at 3.125 Gbps, 4G Fibre Channel at 4.25 Gbps, next generation PCI Express at 5 Gbps, and the OIF's CEI-6 specification of 4.976 - 6.375 Gbps. In order to allow interoperability, the base electrical specification for Interlaken is the OIF's 2003.104.10 [5].

Recently, the Interlaken Alliance has defined a set of recommended baseline configurations which target typical applications at data transfer rates of 10Gb/s, 25Gb/s, 50Gb/s and 100Gb/s. Physical interoperability recommendations are also included. Summary of the recommendations is given in the underlying table (Table 4) [8].

Transfer Rate	Recommendation #1 Wide Options			Recommendation #2 Narrow Options		
	Lanes	Rate [Gb/s]	Electrical	Lanes	Rate [Gb/s]	Electrical
10 Gb/s	5	3.125		3	6.25	CEI-6G-SR
25 Gb/s	5	6.25	CEI-6G-SR			
50 Gb/s	10	6.25	CEI-6G-SR	6	10.3125	CEI-11G-SR
100 Gb/s	20	6.25	CEI-6G-SR	12	10.3125	CEI-11G-SR

Table 4: Interlaken Interoperability Recommendations

II-A-6 PCIe

PCI Express (PCIe) was introduced by Intel and others as the successor to PCI. Unlike earlier buses, PCIe is based upon point-to-point serial links rather than a shared parallel bus structure where all devices connected to the bus shared the same bandwidth. Timing skew typical for traditional parallel format in the underlying physical interface limits bandwidth scalability. PCIe avoids this issue by using serial channel (bonding) where physical link is composed of multiple lanes that do not require the bits to arrive simultaneously so achieving better scalability. This multichannel serial design also offers more flexibility.

Physical Specification

PCIe has layered protocol architecture, consisting of physical layer, data link layer and transaction layer. Functions of the each layer will not be treated here any further, but some interesting specifications of the physical layer.

The physical layer of the PCIe is divided into logical and electrical sub-blocks that coordinate the state of each transceiver.

The electrical sub-block contains a Transmitter and a Receiver. The Transmitter serializes the Logical sub-block Symbols and transmits onto a Lane. The Receiver becomes serialized Symbols from the Lane. It transforms the electrical signals into a bit stream which is de-serialized and transmitted to the Logical sub-block together with a clock recovered from the incoming serial sequence.

The logical sub-block of the physical layer is responsible for following functions [9]:

- Symbol Encoding (8B/10B)
- Framing and Application of Symbols to Lanes
- Data Scrambling
- Link Initialization and Training with *Lane-to-Lane De-skew*
- Clock Tolerance Compensation

The logical sublayer is sometimes further subdivided in MAC and PCS although it is not part of the PCIe specification. In the specification published by Intel, called PIPE (The PHY Interface for the PCI Express Architecture) [10], MAC/PCS functional partitioning is defined as well as the interface between these two layers. It provides the specification of functionally equivalent PCIe PHY as “one way to implement” base specification requirements in order to speed up PCIe endpoint device development. Physical Media Attachment Layer (PMA) with SerDes and other functionality is identified as shown in the Figure 2 below, but PMA/PCS interface is not defined since implementation of SERDES devices vary among IC (integrated circuits) vendors [10].

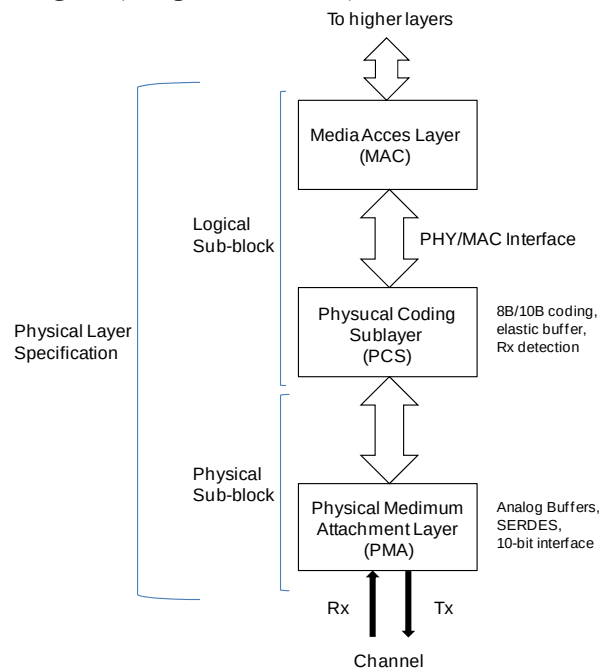


Figure 2: Partition of the PHY Layer in PCIe

As mentioned above, the fundamental PCI Express serial link is the lane, full-duplex low-voltage differential pair of signals. A data clock is embedded using 8B/10B encoding in order to achieve high data rates. There is from one to thirty-two in powers of two (1x, 2x, 4x, 8x, 16x and 32x) multiple lanes that comprise physical PCIe channel. In PCIe 1.x, each lane carries 250 MB/s. For example, an 8x PCIe channel (8 lanes in each direction) can achieve 2 GB/s data rate in each direction, resulting in an aggregate bandwidth of 4 GB/s. Since PCIe technology is under constant development and progress, there are different generations of the standards with ever-growing data rates as outlined in the Table 5 below. PCIe 3.0 is the newest generation and early PCIe 3.0 products will emerge starting in the 2010 timeframe [11].

	PCIe 1.0	PCIe 2.0	PCIe 3.0
1x	0.25 GB/s	0.5 GB/s	1 GB/s
2x	0.5 GB/s	1 GB/s	2 GB/s
4x	1 GB/s	2 GB/s	4 GB/s
8x	2 GB/s	4 GB/s	8 GB/s
16x	4 GB/s	8 GB/s	16 GB/s
32x	8 GB/s	16 GB/s	32 GB/s

Table 5: PCIe Standard Overview

In year 2007, PCI-SIG® released a PCI Express External Cabling 1.0 specification [12], which extends the PCI Express interconnecting technology to “outside the box” applications. It defines attributes and electrical characteristics of copper cables and connectors for x1, x2, x8 and x16 link widths, with a transfer rate of 250 MB/s per lane as found in PCI 1.0. A maximum cable distance is still unspecified.

II-A-7 InfiniBand

InfiniBand [13] is a scalable, high-speed industry-standard architecture developed by InfiniBand Trade Association (IBTA). It was primary designed as a general purpose I/O interconnect but it has found its most effective deployment in storage applications and high-performance computing. This true fabric architecture with switched point-to-point serial links can achieve a transfer rate of up to 120 Gb/s, both in chassis backplane applications and external through copper or optic cable. It supports several basic data rates that can be bonded together to achieve higher bandwidths. These benefits made InfiniBand become a high-speed connectivity solution between servers and switches.

The smallest complete InfiniBand unit is a subnet (fabric). In order to create a large InfiniBand network, multiple subnets can be joined by routers. The key elements are: collection of links, HCA (Host Channel Adapter), TCA (Target Channel Adapter), SM (Subnet manager) and InfiniBand Switch. HCA and TCA are adapters needed for communication across InfiniBand infrastructure and while HCA resides on the host node or server, TCA is used to connect an external storage system or I/O modules to the InfiniBand system area network. SM is used to manage InfiniBand Fabric and to monitor interconnect performance. A fabric can be simply point-to-point connection or multiple connections through one or more switches. Scalability is provided through a number of HCAs, TCAs and other InfiniBand switches that can connect to the InfiniBand infrastructure

As all other interconnecting technologies, InfiniBand uses a multi-layer architecture [13]. On the physical layer, the basic path between any two fabric elements is full-duplex transmission link. The InfiniBand base serial connection data rate is 2.5 Gb/s in each direction which is termed Single Data Rate (SDR). InfiniBand further supports double and quadruple data rates, termed DDR (Double Data Rate) and QDR (Quadruple Data Rate) respectively. Summarized, lane data rates of SDR, DDR and QDR are 2.5 Gb/s, 5Gb/s and 10Gb/s, respectively. Beside these lane speeds, bandwidth is further increased by adding more lanes to the link. InfiniBand defines following full-duplex link widths: 1x, 4x, 8x or 12x as one can see in the Table 6 below.

Link Width	SDR		DDR		QDR	
	Link Rate	Payload Rate	Link Rate	Payload Rate	Link Rate	Payload Rate
1x	2.5	2	5	4	10	8
4x	5	4	20	16	40	32
8x	20	16	40	32	80	64
12x	30	24	60	48	120	96

Table 6: An Overview of InfiniBand Data Rates

The 20 % difference between link rate and payload rate is due to the 8B/10B channel coding.

However, actual utilization of the available bandwidth may be limited mainly by the efficiency of the server BUS architecture, namely PCIe, etc. With continuous development of the PCIe and recent availability of the PCIe 2.0 specification, higher data rates of external InfiniBand interconnects are realizable. In the reference paper [14], an evaluation of PCIe 2.0 and QDR InfiniBand performance was carried out and it was shown that PCIe 2.0 significantly improves overall system performance by lowering latency and increasing bandwidth.

Although InfiniBand standard does not define maximum cable length that guarantees these data rates, signal integrity requirements limit the transmission between end-points using copper cabling. As the transmission data rate increases, maximum achievable copper cable distance decreases due to data skew and well-known signal attenuation at higher frequencies caused by physical effects such as crosstalk.

The following table (Table 7) shows possible transmission distances, achieved for different InfiniBand data rates and link widths by cable manufacturers [16].

Line Rate		Max Distance (copper)
SDR	1x	20m
	4x	15m
	12x	10m
DDR	1x	
	4x	10m
	12x	7m

Table 7: Cable Lengths for SDR and DDR

Better transmission distances are attained using optical cabling even though it causes higher costs. For example, the maximum achievable distance for 4x-DDR over 50 micron optical fibres is 200 m.

II-A-8 RapidIO

Rapid IO (RIO) [18] is an open-standard, high-performance packet-switch interconnection technology designed by leaders in embedded computing for interconnecting chips on circuit boards and printed circuit boards for backplanes.

The RIO standard is specified in three layers: logical, transport and physical layer. The logical layer defines overall protocol and packet formats necessary for end points to initiate and complete a transaction. The logical specifications are defined in a bit stream format because they do not imply a

specific transport or physical interface [18]. Rapid IO is not dependent on the bandwidth or latency of the physical fabric and works well whether the physical fabric is a point-to-point ring, a switched multi-dimensional network or a duplex serial connection. The transport layer provides routing information in order to transport packets from one end to another end in the system. Physical layer defines how adjacent processing elements electrically connect to each other and describes physical interface properties such as transport mechanisms; interconnect topology, flow control, electrical characteristics needed to move packets between two physical devices. This layered architecture makes RIO flexible, so that new transaction types can be added to the logical specification without need for modification in lower layers.

Rapid IO is a standard in constant development. The current available standard is RapidIO 2.0 [18]. It builds on revision 1.3 capabilities and introduces new 5.0 and 6.25 Gbaud PHY. Among other enhancements it also defines how other protocols such as PCIe and Ethernet can be encapsulated and transported seamlessly through the RIO fabric.

Physical specification

In general, there are two different types of physical channel defined by RIO, namely parallel RIO and serial RIO (sRIO). The main difference is that parallel RIO has separate clock signal while sRIO uses 8B/10B encoding to transmit clock together with data serially.

The RapidIO 2.0 Part 4: 8/16 LP-LVDS Physical Layer Specification [18] defines 8-bit and 16-bit parallel (8/16 LP-LVDS), point-to-point unidirectional interfaces. An 8/16 LP-LVDS device interface contains a dedicated 8- or 16-bit port with a source-synchronous-clock signal and frame signals, and a 8- or 16-bit output port with clock and frame signals. Differential signalling is used. A RapidIO channel completely constructed onto a single board offers the highest performance in terms of clock rate and signal integrity.

Serial RIO [18] (XAUI compatible) addresses the physical layer requirements for devices utilizing an electrical serial connection medium. This specification defines a full duplex serial physical layer interface (link) between devices. The links are comprised of one or more lanes and allows bonding of up to sixteen (x16) serial lanes for applications requiring higher link performance.

The general serial RIO physical layer properties are [18]:

- Embeds the transmission clock with data using an 8B/10B encoding
- Supports links from one lane up to sixteen bonded lanes (1, 2, 4, 8 and 16 lanes per direction) where each lane is a pair of unidirectional serial paths with one path in each direction
- Allows switching packets between RapidIO LP-Serial ports and 8/16 LP-LVDS parallel ports without requiring packet manipulation.
- Supports transmission rates of 1.25, 2.5, 3.125, 5 and 6.25Gbaud (data rates of 1, 2, 2.5, 4 and 5 Gbps) per lane.
- Supports division of the physical layer bandwidth into up to 9 virtual channels with independent flow control.

The sRIO physical layer comprises Physical Coding Sublayer (PCS) and Physical Media Attachment (PMA) sublayer (terminology is adopted from IEEE 802.3).

The PCS function is responsible for idle sequence generation, lane striping, scrambling and encoding for transmission and decoding, lane alignment, descrambling and destripping on reception. The PCS uses the 8B/10B encoding for embedding clock with data and transmission over the link.

The PMA is responsible for serializing 10-bit parallel code-groups to/from a serial bit stream on a lane-by-lane basis. Upon receiving data, the PMA function provides alignment of the received bit stream to 10-bit code-group boundaries, independently on a lane-by-lane basis.

Indeed, it is important to mention that serial RIO 1.25Gbaud, 2.5Gbaud, and 3.125Gbaud electrical specifications are guided by the XAUI electrical interface. The Electrical specification for 5Gbaud and 6.25Gbaud are based upon the Optical Internetworking Forum's Common Electrical Interface [5], with the difference that OIF CEI documentation defines support for 4.976 to 6.375Gbaud while RapidIO only supports 5.0 and 6.25 Gbaud data rates.

Distances supported by this physical links that use NRZ coding (hence 1 bit per symbol at the electrical level) with data rates of 1.25Gbaud, 2.5Gbaud, and 3.125Gbaud are:

- Short run: at least 20cm of PCB material with up to a single connector
- Long run: at least 50cm of PCB material with up to two connectors

Physical links with NRZ coding and data rates of 5 and 6.25 Gbaud have following distances:

- Short run: at least 20cm of PCB
- Medium run: at least 60cm of PCB material with up to two connectors
- Long run: at least 100cm of PCB material with up to two connectors

Because it is XAUI and CEI compatible and offers scalability in bandwidth through different lane versions, serial RapidIO technology is gaining acceptance as a backplane interconnect.

II-A-9 SFI and SPI

System Packet Interface (SPI) and System Framer Interface (SFI) specify interfaces commonly used in synchronous optical networking (SONET/SDH) and other applications. As all Implementation Agreements (IAs) developed by OIF which standardize interfaces within telecom systems, Serdes Framer Interface (SFI) and System Packet Interface (SPI) are also based upon OIF System Reference Model as shown in Figure 3 [30].

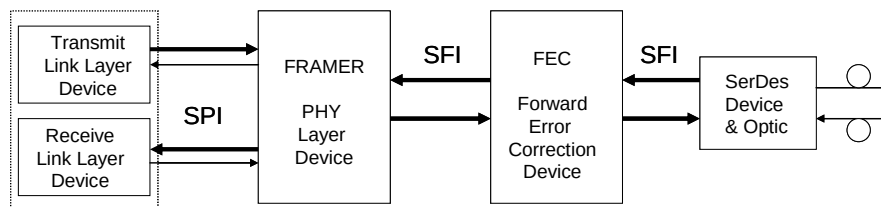


Figure 3: The OIF System Reference Model

The typical interface of a communication system with optical links consists of three devices interconnected electronically: an optical module comprising Serdes component, FEC (Forward Error Correction) device and Framer, as can be seen on the Figure above.

SPI typically connects Link Layer Devices to Physical Layer Devices (i.e. Framer for optical network or MAC for IP network) while SFIs connect Physical Layer Devices to each other. SFI is relatively short chip-to-chip or chip-to-module interconnect and SPI supports larger distances over backplane since Link Layer devices are frequently not located on the same card as Physical Layer devices. There are several agreements for SPI and SFI (as can be seen in Table 1) that defer, among others, in line rate, supported total data rate, number of control and clock signals and channel coding:

- SPI-3 – System Packet Interface for 2.488 Gb/s (OC-48)
- SPI-4.1 – System Packet Interface Level 4, Phase 1 for 10 Gb/s (OC-192)
- SPI-4.2 – System Packet Interface Level 4, Phase 2 for 10 Gb/s (OC-192)
- SPI-5 – Packet Interface Level 5 for 40 Gb/s (OC-768)
- SFI-4.1 – System Framer Interface Level 4, Phase 1 for 10 Gb/s (OC-192)
- SFI-4.2 – System Framer Interface Level 4, Phase 2 for 10 Gb/s
- SFI-5 – Packet Interface level 5 for 40 Gb/s
- SPI-5 [31] is an interface for aggregate bandwidths of OC-768 ATM and Packet over SONET/SDH.

Although it is designed to meet requirements of this specific application, it may also be used in other applications at the data rate of 40 Gb/s.

The general properties of SPI-5 are [31]:

- Point-to-point connection
- Support for 256 ports with address extension to 2144 ports.
- Transmit / Receive Data Path
- Data transfer segmented in bursts that are multiples of 16 words (32 bytes) in length.
- 2.488 Gb/s minimum data rate per line
- Independent X11 + X9 + 1 stream cipher scrambling on each bit lane
- Source synchronous clock at one quarter the data bit rate

Serdes Framer Interface SFI-5 has following characteristics [32]:

- Point-to-point connections of the Serdes component to the FEC processor, the FEC processor to the SONET/SDH framer or the Serdes component directly to the SONET/SDH framer.
- 16-bit wide data bus with each channel operating at up to 3.125 Gb/s.
- Supports SONET OC-768, SDH STM-256, OTN OTU-3 and other systems operating at the payload data rate in the 40Gb/s range, with up to 25% FEC overhead
- Deskew Channel included in both RX and TX directions

The electrical specification for both SPI-5 and SFI-5 is defined in System Interface Level 5 (SxI-5): Common Electrical Characteristics for 2.488 – 3.125Gbps Parallel Interfaces [33]. Some of the fundamental characteristics are [33]:

- Serial data rate support from 2.488 Gb/s to 3.125 Gb/s
- Well defined voltage levels and timing budgets
- Low bit error rate (BER)
- Capable of driving at least 8 inches of FR4 with 1 or 2 connectors
- Compatible with other high-speed I/O families

II-A-10 Serial Attached SCSI (SAS)

Serial Attached SCSI (SAS), first introduced in 2004, is a point-to-point interconnecting technology developed by the International Committee for Information Technology Standards (INCITS) that replaces the parallel SCSI bus structure. It is intended for use in storage networks for communication between host and disk drive devices. SAS-1.0 protocol currently supports an interconnect speed of 1.5 Gb/s and 3 Gb/s whereas near future standard (SAS-2.0) will provide data rates of 6 Gb/s. 12 Gb/s SAS-3.0 is planned for the Year 2015 [34]. Beside this better utilization of the bandwidth, the 6 Gb/s SAS-2.0 standard will also provide easier management mechanisms and network robustness. The Serial Attached SCSI standard defines five layers: Application, Transport, Link, PHY and Physical. The physical layer defines electrical and physical characteristics and three types of connectors:

SFF 8482 – SATA compatible (i.e. physical specifications of SAS are compatible with that of SATA)

SFF 8484 – up to four devices

SFF 8470 – external InfiniBand connector, up to four devices

PHY Layer comprises 8B/10B block coding, link initialization and speed negotiation functionalities.

B. New proposals and future trends

As the need for higher system capacities becomes ever greater, a lot of efforts are made by electronic standardization bodies such as IEEE 802.3ba Ethernet Task Force, OIF, Fibre Channel, InfiniBand, Rapid IO and PCI Express in order to achieve an improvement in existing interconnecting technologies regarding data rate and efficiency. IEEE 802.3ba Ethernet Task Force is currently working on 40 Gigabit Ethernet and 100 Gigabit Ethernet. The Physical and Link Layer (PLL) Working Group of the OIF develops physical specification CEI-25G in order to achieve lane signalling rate of up to 28 Gbaud/s. InfiniBand has announced future plans and their next step (as taken from their homepage) is EDR (Eight x Data Rate) 1x 20 Gb/s per lane that is planned for the year 2011. As already mentioned above, beside the lane speed InfiniBand also supports full-duplex link widths of 4x, 8x or 12x which are also planned to be achieved one after another. Further, PCIe is also about to release the final specifications for PCI Express 3.0 with improved data rates. Rapid IO future work is focused on higher data rates which will come with their serial Rapid IO 3.0 standardization. In what follows, we briefly present two of the most mature technologies currently under development.

II-B-1 CEI-25G

Common Electrical I/O 25 Gb/s (CEI-25) [20] is the project currently under development by the Physical and Link Layer (PLL) Working Group of the Optical Internetworking Forum (OIF). This project is based on improvements in both signalling technology and channel technology [20] and it defines electrical specifications for up to 28 Gbaud/s signalling. It is intended for next generation chip-to-chip and chip-to-module as well as for backplane applications that support transmission of 100 Gb/s, such as OTU4 or 100GbE (100 Gigabit Ethernet). Therefore, there are two types of definition: CEI-28G-SR for short reach applications (chip-to-chip and chip-to-module) with 28 Gbaud/s and CEI-25-LR for long reach applications (backplane) with signalling rate of 25 Gbaud/s.

Equalized NRZ (Non-Return-to-Zero) is the signalling technology of choice for this physical specification. In addition to standard NRZ signalling, equalizer circuits in both transmitter and receiver are used in order to combat ISI (Inter Symbol Interference). DFE (Decision Feedback Equalizer) is typically used in the receiver since it can cancel the precursor ISI (interference from future symbols) without noise enhancement, while a FFE (Feed Forward Equalizer) is typically included in receiver. One possible solution that shows good performance at desired data rates is receiver that consists of cascaded CTE (Continuous Time Equalizer) with the DFE [20].

As defined in [37] the basic requirements on CEI-28G-SR are:

1. Support serial data rate from 19.90 Gsym/s to 28.00 Gsym/s.
2. Capable of low bit error rate (10^{-12} , with an objective of 10^{-15}).
3. Capable of driving 0 – 300 mm of PCB and up to 1 connector.
4. Shall support AC-coupled operation and optionally DC-coupled operation.
5. Shall allow multi-lanes (1 to n).

Requirements on CEI-25G-LR are [21]:

1. Support serial data rate from 19.90 Gsym/s to 25.00 Gsym/s.
2. Capable of low bit error rate (10^{-12} , with an objective of 10^{-15}).
3. Capable of driving 0 – 686 mm of PCB and up to 2 connectors.
4. Shall support AC-coupled operation.
5. Shall allow multi-lanes (1 to n).

Development of 100 GbE with narrower interfaces would be possible with these signalling rates. This electrical interface would be four lanes wide ($4 \times 25 \text{ GB/s} = 100 \text{ GB/s}$) in each direction. Such interfaces would bring a lot of advantages such as lower pin count and smaller package sizes of electrical devices, connectors and optical modules as well as lower power dissipation.

II-B-2 IEEE 802.3ba (40GbE and 100GbE)

The IEEE 802.3ba Ethernet Task Force has been working on the development of the 40 Gigabit Ethernet (40GbE) for server and computing application and 100 Gigabit Ethernet (100GbE) for network aggregation application.

The objectives of this project are [23]:

- Support full-duplex operation only
- Preserve the 802.3 / Ethernet frame format utilizing the 802.3 media access controller (MAC)
- Preserve minimum and maximum frame size of current 802.3 standard
- Support a bit error rate (BER) better than or equal to 10^{-12} at the MAC/ physical layer service interface
- Provide appropriate support for optical transport network (OTN)
- Support a MAC data rate of 40 gigabit per second
- Provide physical layer specifications which support 40 gigabit per second operation over:
 - at least 10km on single mode fibre (SMF)
 - at least 100m on OM3 multi-mode fibre (MMF)
 - at least 10m over a copper cable assembly
 - at least 1m over a backplane
- Support a MAC data rate of 100 gigabit per second
- Provide physical layer specifications which support 100 gigabit per second operation over:
 - at least 40km on SMF
 - at least 10km on SMF
 - at least 100m on OM3 MMF
 - at least 10m over a copper cable assembly

The IEEE 802.3ba amendment defines a single architecture for 40GbE and 100GbE and all physical layer specifications mentioned above [24]. This architecture consists of a MAC Layer (corresponds to Layer 2 of the OSI model) and Ethernet physical device (corresponds to Layer 1 of the OSI model). The physical device consists of PMD, PMA and PCS sublayer as it is the case by 10GbE. For backplane and copper applications, there are two another sublayers, AN (Autonegotiation) and FEC (Forward Error Correction) sublayers.

PCS is responsible for encoding and uses 64B/66B coding scheme. In order to decouple the evolution of the optical interface widths from the evolution of the electrical interface widths a Multilane Distribution (MLD) scheme was introduced by the Task Force [23] as the basis for PCS for 40GbE and 100GbE. It is a low overhead scheme that is based upon striping of the 66-bit code blocks over multiple lanes. The number of the needed PCS lanes is the least common multiple of the expected widths of optical and electrical interfaces [23]. For 100GbE 20 PCS lanes has been chosen thus supporting interconnect widths of 1,2,4,5,10 and 20 channels and 4 PCS lanes for 40GbE with interface widths of 1,2 and 4.

PMA sublayer is responsible for transmission, reception, collision detection, clock recovery and skew alignment. It connects PCS and PMD sublayers. Certain flexibility is needed by the PMA sublayer since its actual implementation depends on specific PMDs used.

Different PMDs are under development for 40GbE and 100GbE and they all are briefly outlined in the Table 8 [23].

PMD	Media Type	Distance [m]
100GBASE-SR10	OM3 MMF	100
100GBASE-LR4	SMF	10 000
100GBASE-ER4	SMF	40 000
100GBASE-CR10	Copper cable	10
40GBASE-KR4	Copper backplane	1
40GBASE-CR4	Copper cable	10
40GBASE-SR4	OM3 MMF	100
40GBASE-LR4	SMF	10 000

Table 8: An Overview of different PMD types

40GbE and 100GbE Interfaces

In the IEEE 802.3ba specification two types of interfaces will be defined, the logical and the physical interface (as in the past). While logical specification only defines logical signal and their behaviour, physical interface specification defines electrical and timing parameters of physical signals.

Logical interfaces for 40GbE and 100GbE are XGMII and CGMII, respectively. They are defined as interfaces between MAC and PCS and the only difference between is the data rate (clock frequency). This type of interface is 64 bits wide that are grouped in 8 lanes of 8 bits.

The low pin count physical interfaces are XLAUI and CAUI for 40GbE and 100GbE, respectively, that enables partitioning in the similar way as XAUI. Each lane has an effective 64B/66B encoded data rate of 10 Gb/s.

In the case of XLAUI, there are 4 transmitting and 4 receiving lanes, i.e., 8 pairs or 16 signals while CAUI utilizes 10 transmit lanes and 10 receive lanes, i.e., 20 pairs or 40 signals. These interfaces are primarily chip-to-chip interfaces but they could also be used as interfaces for pluggable form factor modules. Transmit reach is based on a channel of a 25 cm length over FR4 PCB with one connector.

III. Usage Models for Electrical High-Speed Interconnects

In this section we provide an overview of the various usage models for high-speed, point-to-point serial interconnecting technologies. It provides its basic categorization and description.

A. Backplane (Single-Shelf) Applications

A basic backplane application includes transmitter, transmitter package, at least one electrical connector, PCB assemblies (traces, wires, etc.), receiver package, and receiver. Interconnect between transmitter and receiver can be configured as a single channel, multiple channels (independent) or as multiple channels (bonded). Some applications include daughter cards, chassis and switch fabrics. A generic model for point-to-point backplane medium is illustrated in Figure 4.

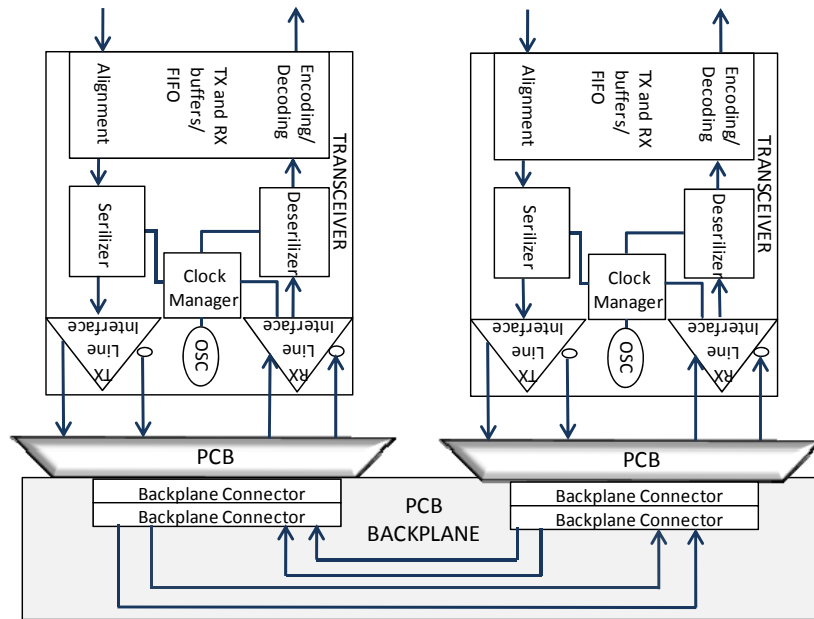


Figure 4: Backplane Interconnect Generic Block Diagram

Technologies described in Section II, such as XAUI, serial Rapid IO, InfiniBand and so on, are interconnect standards supporting backplane interfacing. Proprietary interfaces designed by chip vendors are also common.

A typical system consists of multiple line cards located in a rack that are electrically connected by a backplane. In order to interconnect a large number of line cards, the designer has to choose between distributed and centralized switch fabric architecture.

In the distributed architecture, switching is implemented on the line cards before data is sent out onto the backplane. Each line card is connected to every other line card in the system as shown in Figure 5 below. In the case of N line cards, each line card comprises $N-1$ transceivers. Total number of such point-to-point channels as depicted in Figure above (Figure 4) is $(N-1) \times N$.

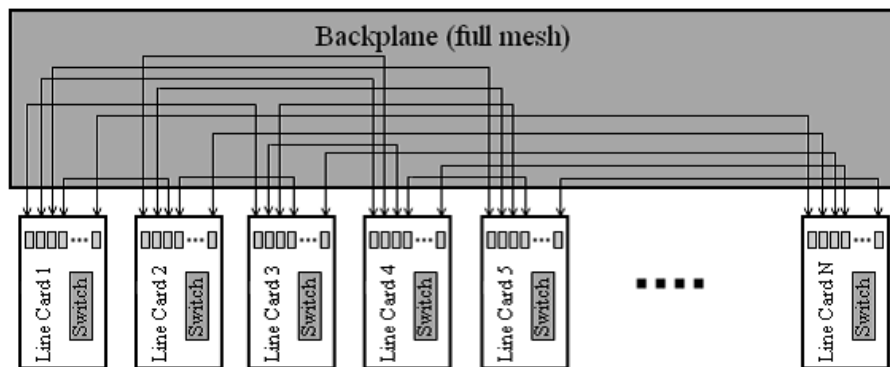


Figure 5: Distributed Architecture

The centralized architecture (Figure 6) uses a switch fabric card that is responsible for interconnecting all ports on all the line cards in the system. For the system comprising of N line cards and one switch fabric, there is one transceiver per line card and N transceivers on the switching card so building the total number of N point-to-point interconnects.

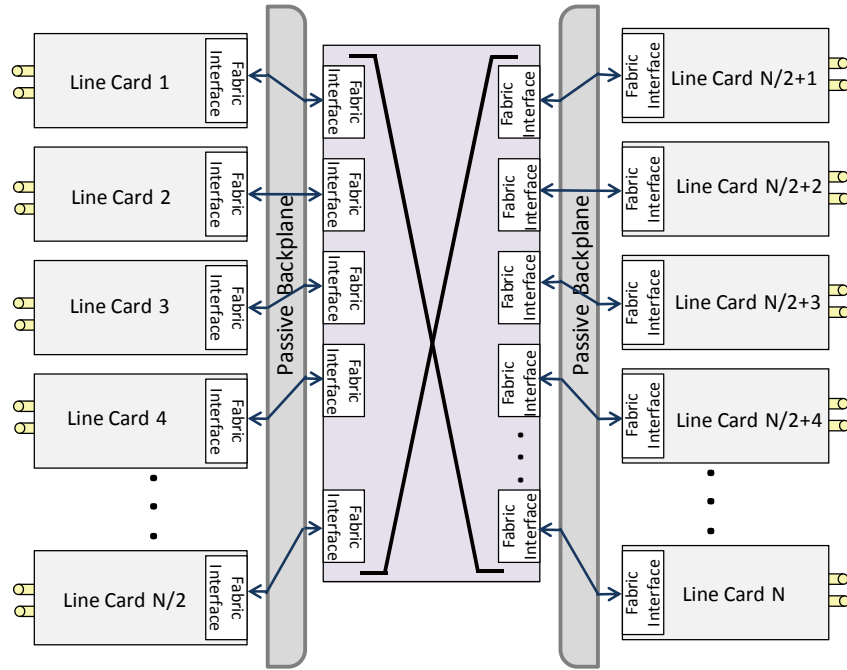


Figure 6: Centralized Architecture

B. Multi-Shelf Applications

The most widely used interconnection medium in multi-shelf applications is copper cable. Copper cabling can be used to interconnect line cards over short to medium distances that reside in different racks of equipment that are parts of the large high-capacity network node. These cables are generally based upon coaxial, shielded twisted pair (STP) or similar technologies. Single channel transceivers (single differential pair in each direction) are mostly used, although configurations with multiple channels can also be deployed. The interconnecting standards such as Fibre Channel, 10GbE and InfinBand support this art of connection. As pointed out in Figure 7, a generic model for this type of medium include transmitter, transmitter package, PCB assemblies (traces, wires, etc.), electrical connectors, cable, receiver package, and receiver.

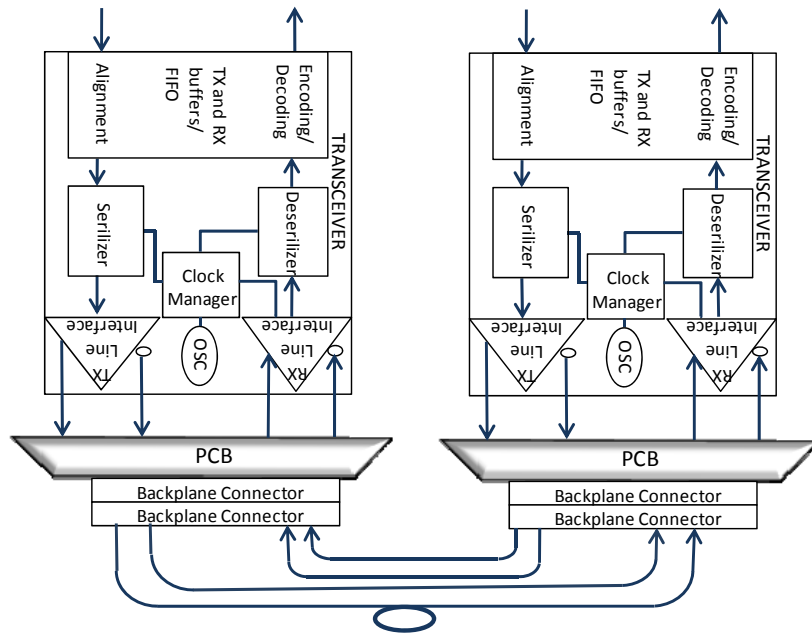


Figure 7: Copper Cable Interconnect Generic Block Diagram

Interconnects that use optical fibres instead of copper-based cables are capable of providing both larger transmission distance and higher data rate. It is the first generation of optical intra-system interconnects introduced into traditionally electronic equipment in order to meet high throughput requirements. There are several solutions offered, from short reach systems using multimode fibres to long reach options using single mode fiber and supporting various distances. Interconnects between transmitter and receiver can be configured as a serial single channel, or multiple channels achieved through parallel optical ribbon cable or WDM. The interconnecting standards such as Fibre Channel, 10GbE and InfinBand support optical cabling.

A generic model of the first generation optical interfaces is shown in Figure 8 and includes electrical transceiver, transceiver package, PCB assemblies (traces, wires, etc.), electro-optical converters (optical transceiver), optical media (fibre and connectors) and receiver package and receiver.

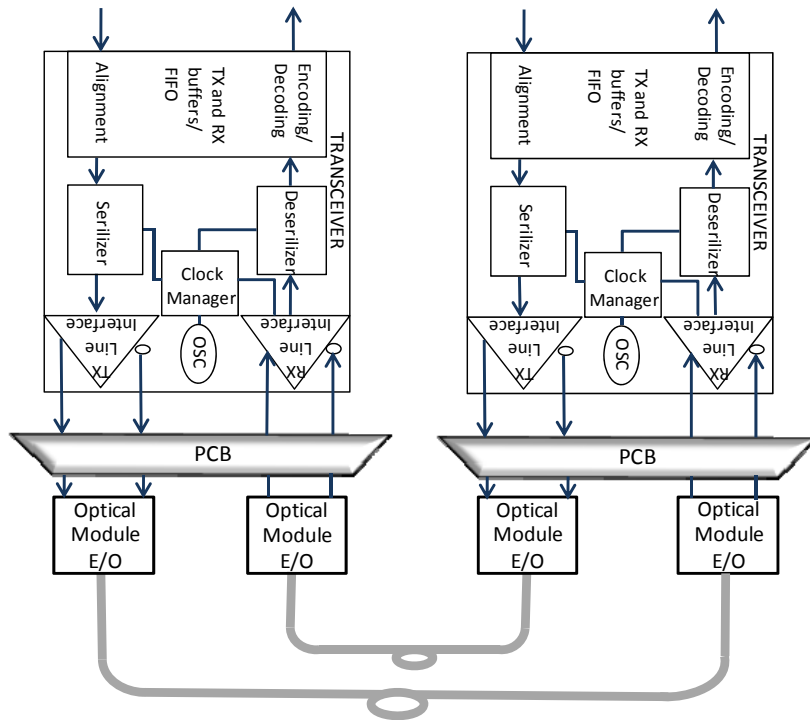


Figure 8: Optical Cable Interconnect Block Diagram

Following Figures (Figure 9 and Figure 10) show our developed high-capacity multi-shelf system models for calculating required number of optical modules, fibre connections, copper transceivers and copper connections for different aggregate capacities and maximum achievable distance. Figure 9 shows centralized architecture. Each line card (LC), located in a line card rack (LCR), is directly connected via point-to-point optical or copper links with a central switch that comprises a number of switching cards (SCs). The partitioned architecture shown in Figure 10 uses a three-stage switch architecture, where first and third stages are located in LCRs and connected to the middle stage by optical interconnects. LCs are connected electronically over backplane with the first and third switch.

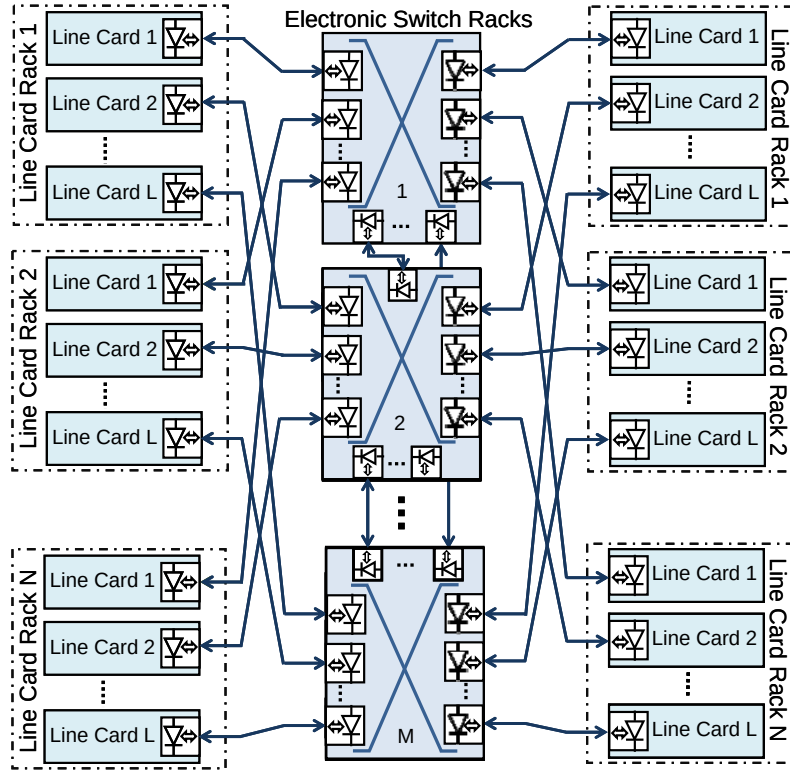


Figure 9: Generic multi-shelf centralized model

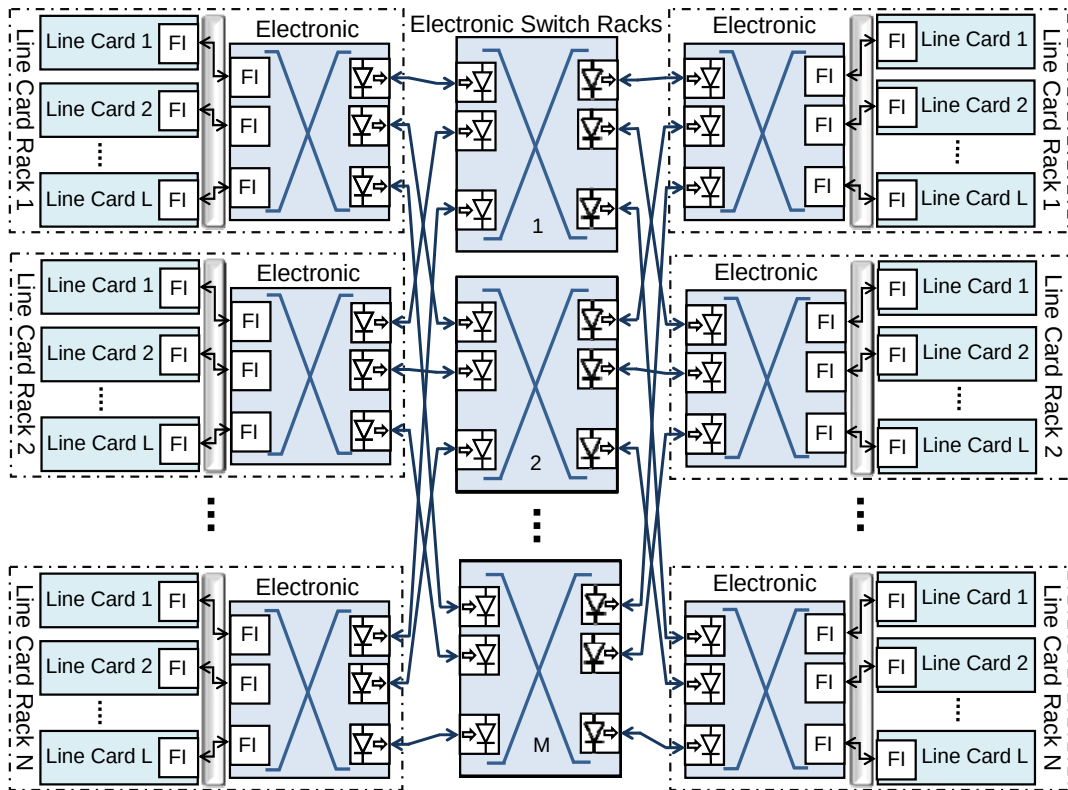


Figure 10: Generic multi-shelf partitioned model

IV. Interconnects in Current High-Performance Switches and Routers

The Cisco CRS-1 Carrier Routing System [27] is the system with highest aggregate capacity currently available on the market. Its main components are: 4-Slot Single-Shelf System, 8-Slot Single-Shelf System, 16-Slot Single-Shelf System and 24-Slot Fabric Card Rack. It offers two possible configurations: single-shelf system and multi-shelf system. Single-shelf system consists of just one rack comprising line cards and switch cards with total throughput of up to 1.2 Tb/s (for 16-Slot Single-Shelf configuration) whereas multi-shelf system can have up to 72 line card racks (8-Slot and 16-Slot Racks only) and up to 8 fabric card racks with up to 92 Tb/s total capacity when full configuration is used. A switching fabric deployed in the CRS-1 Routing System consists of eight planes over which packets are distributed. Within these planes, three fabric stages – S1, S2 and S3 in a Benes interconnection setup, are responsible for routing. In the single-shelf configuration, the fabric cards allocated in the shelf contain all three Benes stages. In the multi-shelf configuration S2 stage is realized using one to eight fabric card racks and S2 and S3 stages are switching cards that reside in the line card racks. In partitioned model in Figure 10, the Cisco CRS-1 full multi-shelf configuration would be for L=16 line cards in each line card rack, M=8 switch fabric racks and N=72 line card racks. Each of the M=8 connections consists of three fibre array cables and hence there are $8 \times 3 = 24$ fibre array cables that connect each line card rack with fabric card racks.

The following table (Table 9) summarises Cisco CRS-1 shelf dimensions.

Cisco CRS-1 Shelves	Physical Dimensions		
	Height [in(cm)]	Width [in(cm)]	Depth [in(cm)]
8-slot Line Card Shelf	38.5 (97.79)	17.5 (44.45)	36.6 (92.964)
16-slot Line Card Shelf	84 (213.36)	23.6 (59.944)	36-39.72 (91.44-100.84)
24-slot Line Card Shelf	84 (213.36)	23.6 (59.944)	36-41.5 (91.44-105.41)

Table 9: Cisco CRS-1 shelf dimensions

V. Limitations of Current Electrical Approaches

As presented in Table 2 each interconnecting technology has certain channel bandwidth defined. Further, it is upper bounded regarding its transmission distance which is clear because of the fundamental physical limitations of electronic interconnects. A question that arises is when actually this plays a decisive role.

In order to answer to this question, we developed the models presented in Chapter III.B for calculating required number of optical transceivers and fibre connections for different aggregate capacities in order to estimate overall system scalability.

We evaluated single-shelf distributed and centralized architecture as well as multishelf centralized and partitioned architecture

In addition, our analysis was based upon physical dimensions of the 23" ETSI (European Telecommunications Standard Institute) cabinet standard [28]. In terms of comparison, typical physical dimensions of the 23" ETSI and 19" IEC Racks are listed in the table (Table 10) below.

Components	Typical Physical Dimensions			Card Pitch
	Height [in(cm)]	Width [in(cm)]	Depth [in(cm)]	
23" ETSI Rack	86.614 (220)	23.622 (60)		
19" IEC60297 Rack	Up to 78.74 (200)	19 (48.26)	optional	1 HP= 5.08 mm (2/10"); nxHP

Table 10: 23" ETSI and 19" IEC typical physical dimensions

These standards does not define number of cards that fit into single shelf but the distance between them should be a multiple of the modular pitch of 1 HP = 5.08 mm = 0.2" (n x 1 HP) [29]. Therefore, we assumed that the cards are spaced at a 1.6" (8HP pitch). Using this cards distance, 12 cards can fit into single 23" (600 mm) shelf. Thus, one rack consists of two such shelves.

Further, for the backplane analysis, we assumed that a line cards operate at 40 Gb/s and a switching capacity of each switch card was supposed to be 160 Gb/s.

Line cards considered in multi-shelf applications were supposed to operate at 100 Gb/s. Switching card capacity for centralized architecture was supposed to be 200 Gb/s. In partitioned architecture, the mentioned line cards were used in LCRs and switching cards with capacity of 600 Gb/s was assumed within middle switch.

In our calculations we assumed that packaging density will increase in the near future, so that 100 Gb/s line cards can also be plugged into the same 23" ETSI cabinet with a card spacing of 1.6" (8HP pitch).

A. Single-Shelf System

First we examined different interconnecting technologies in terms of their usability in the Single-Shelf and Single-Rack configuration.

SFI-5 was not considered here since it cannot be used as backplane interconnect in high-capacity network nodes as its maximum achievable distance is 20.32 cm. All other current technologies are well suited for single-shelf applications regarding supported transmission distance. Using line cards with data rates of 40 Gb/s each, the maximum single-shelf configuration has aggregate capacity of 640 Gb/s, while the maximum single-rack configuration reaches 1280 Gb/s. Here we assume that two shelves can be placed in a single rack, while each shelf carries 16 line cards.

XAUI backplane interconnection solution is limited by its maximum transmission distance of less than 50 cm over PCB. In the full constellation of 12 cards, the distributed architecture is possible with the XAUI if transceivers on the cards are placed so that an end-to-end connection is less than 50 cm. However, in the centralized architecture of essential importance is location of the switching cards within one shelf. They should be placed in the middle of the shelf for XAUI to be used since the distance between switching cards and line cards is halved.

In the case of a single shelf configuration a PCB backplane can be used. However, required length of PCB traces exceeds the maximum allowed distance of 50 cm if more than 12 cards are used, i.e., if more than one shelf in a 32" ETSI rack is needed (see Figure 11). This conclusion can be applied to all backplane technologies with respect to maximum achievable distance presented in Table 1. It can be seen that for smaller capacities of up to 1.2 Tbit/s, the distributed architecture allows shorter interconnection distances than the centralized one. For higher capacities, i.e. larger number of line

cards, the centralized architecture is clearly beneficial. Regarding the total number of 10G XAUI transceivers, the centralized architecture is superior already for capacities of more than 320 Gb/s as it can be seen from Figure 12.

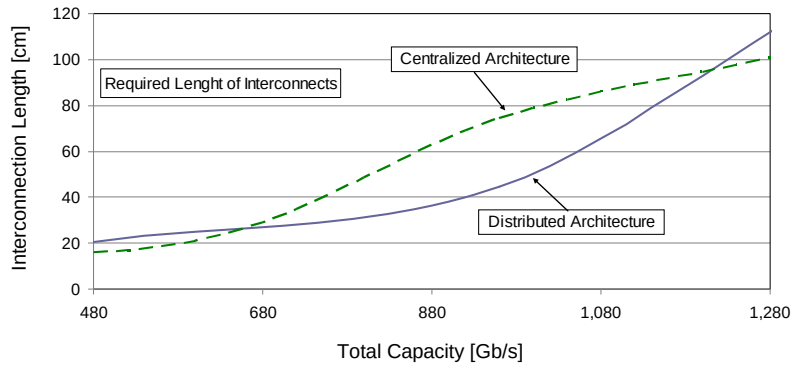


Figure 11: Interconnection length for a single-rack system when assuming either distributed or centralized architecture.

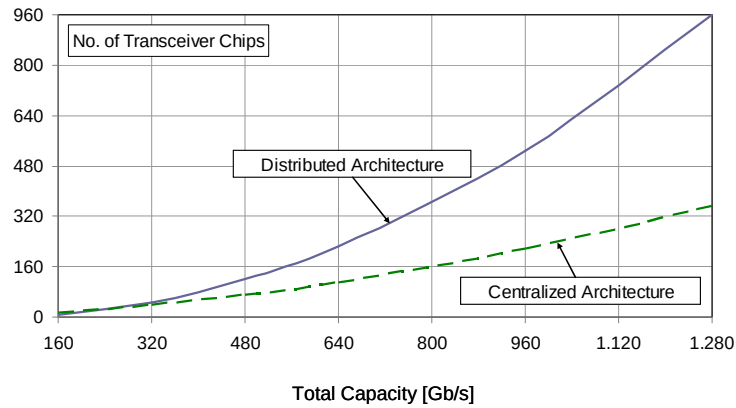


Figure 12: Comparison between distributed and centralized architectures for a single rack system regarding the total number of 10G XAUI transceivers required.

Figure 13 shows comparison between distributed and centralized architectures for the backplane interconnects based on CEI-6G-LR, CEI-11G-LR and CEI-25G-LR. Again, centralized architecture performs better. Clearly, the increase in data rate per channel leads to less number of transceivers and thus fewer backplane connections. CEI-25G-SR with only 56 transceivers for throughput of 640 Gb/s is a future OIF standard and is considered here only for the purpose of comparison.

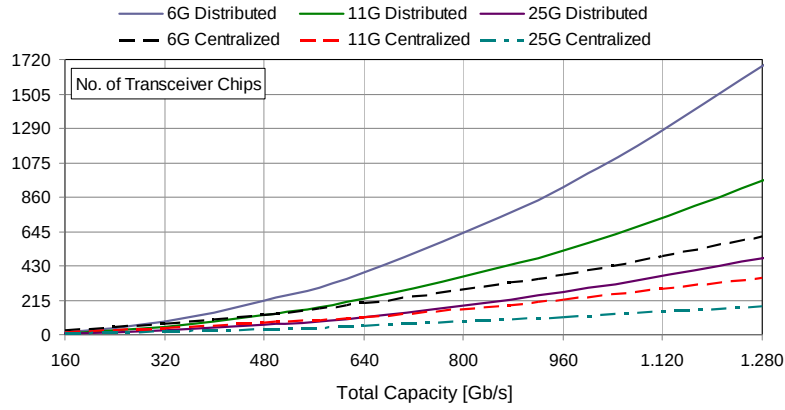


Figure 13: Comparison between distributed and centralized architectures for a single rack system regarding the total number of different OIF CEI transceivers required.

Number of transceiver chips for different sRIO channel data rates is presented in Figure 14. Compared to OIF CEI transceivers, a larger number of PCB traces is required for the same aggregate capacity. Since the payload data rate of sRIO is 5 Gb/s, it performs better than 10G XAUI when considering number of transceivers and PCB traces. Figure 15 shows further data rate enhancement by means of the link aggregation using 1x, 2x, 8x, 16x and 32x links for centralized architecture. This causes reduction in number of transceiver chips needed, but the number of connections remains the same as before since the data rate per channel is the same.

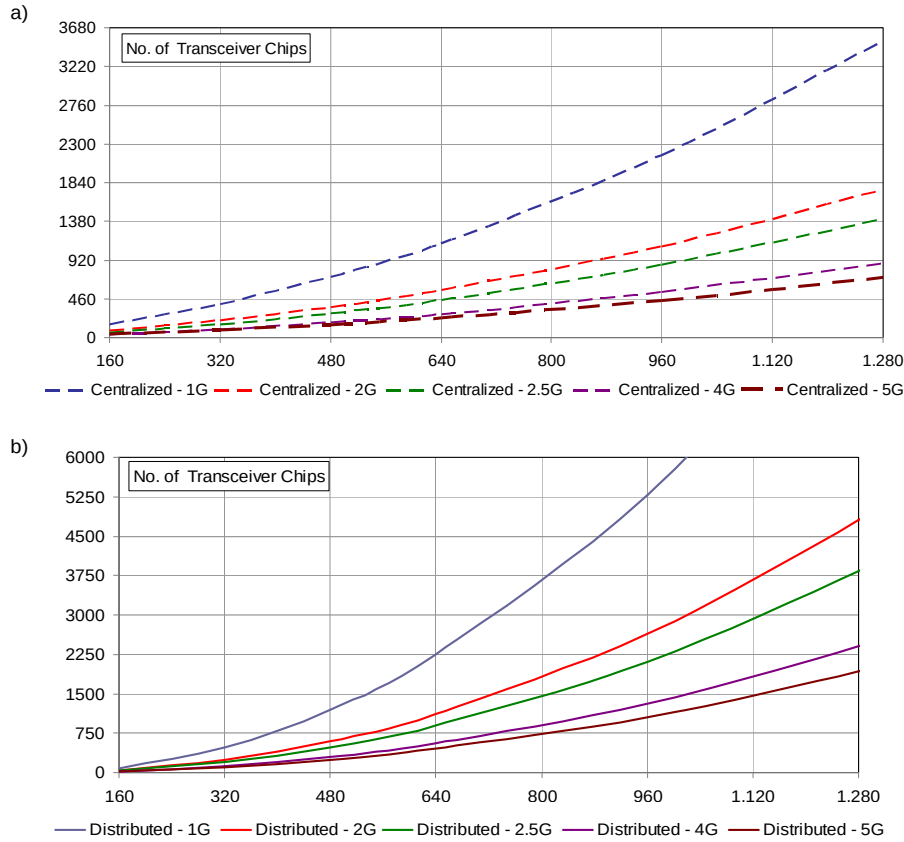


Figure 14: Comparison between distributed and centralized architectures for different sRIO channel data rates regarding the total number of transceivers required.

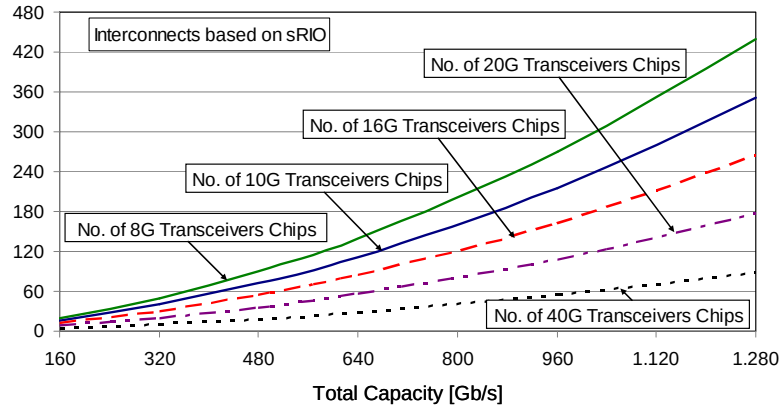


Figure 15: Number of sRIO transceiver chips with higher data rates achieved through link aggregation

InfiniBand performance with respect to payload channel data rates of 2Gb/s, 4Gb/s and 8 Gb/s is shown in Figure 16. As it was the case for other technologies, also InfiniBand used in the centralized architecture has advantage over the distributed one. Channel bundling reduces the number of transceiver chips as in case of sRIO interconnects, which can be seen in the Figure 17.

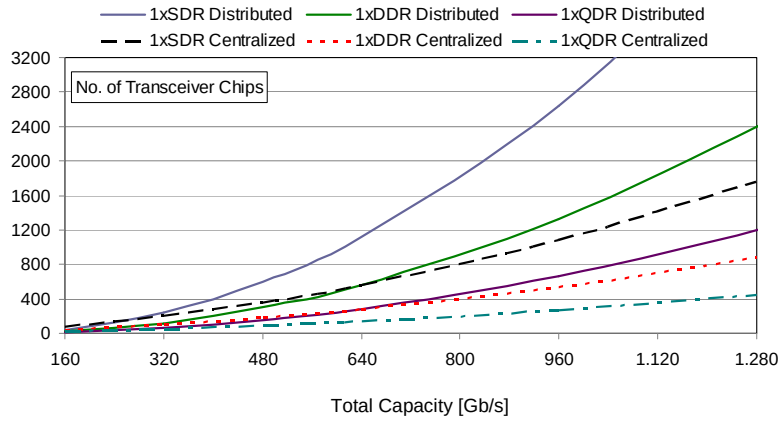


Figure 16: Number of transceiver InfiniBand chips

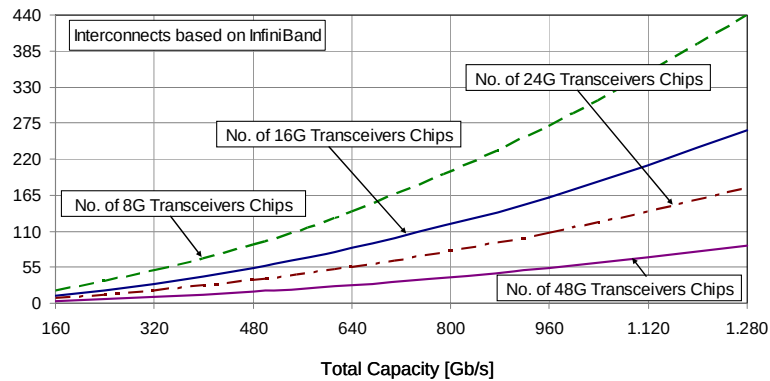


Figure 17: Number of InfiniBand transceiver chips with higher data rates achieved through link aggregation

The following two Figures (Figure 18 and Figure 19) show the number of PCIe and SAS & SATA transceiver chips for the centralized architecture.

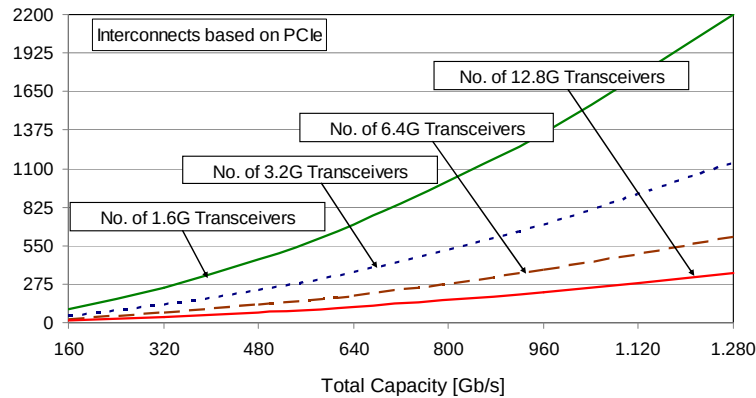


Figure 18: Number of PCIe transceiver chips with link aggregation for the centralized architecture

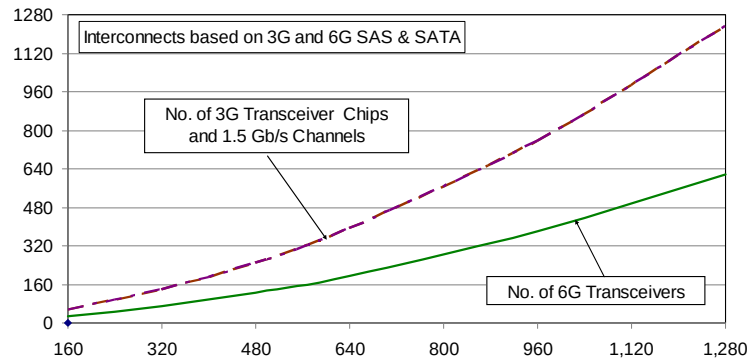


Figure 19: SAS & SATA backplane transceivers for the centralized architecture

Finally, the upper limits for backplane interconnects using above mentioned technologies are summarized in Figure 20. The use of PCB backplanes is possible only in small systems having total capacity of no more than 1280 Gb/s. The capacity of 1280 Gb/s can be achieved using SAS & SATA interconnects as well as PCIe, CEI-6GLR and CEI-11G-LR. Although CEI-25G-LR has the highest line rate, it is limited by its maximum transmit distance so achieving maximum capacity of 960 Gb/s. The smallest aggregate capacity of 640 Gb/s is attained using 10G XAUI and sRIO interconnecting technologies.

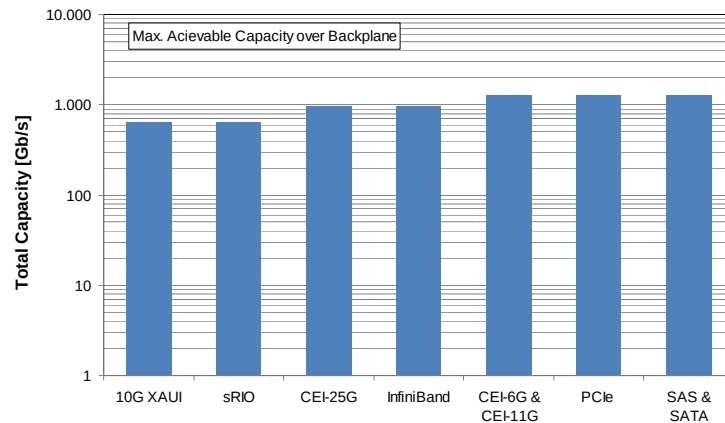


Figure 20: Limitations of interconnecting technologies for PCB backplane

B. Multi-Shelf System

In order to identify potential bottlenecks of multi-rack solution we analysed it regarding number of interconnections and transceivers needed as well as regarding scalability with respect to achievable capacity and footprint. As mentioned above, our multi-shelf system analysis was based on developed generic models presented in Figure 9 and Figure 10 and line cards with 100 Gb/s capacity. For the purpose of maximum distance calculation between racks, we used two basic multi-rack interconnects configurations depicted in the following Figures (Figure 21 and Figure 22). In both, centralized and partitioned architectures, switch card racks were supposed to be placed in the middle of the constellation.

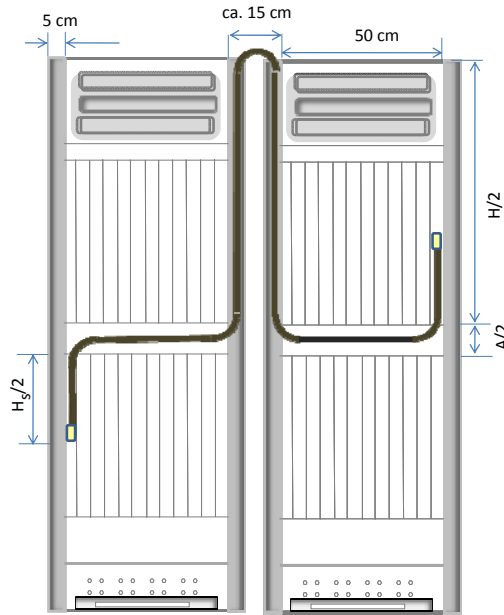


Figure 21: Rack-to-Rack Connection Model 1



Figure 22: Rack-to-Rack Connection Model 2

Regarding XAUI interface, there are three possible media that can be used. The initial specification of XAUI envisages the use of XAUI over PCB traces of length smaller than 50 cm. Other options are to use a copper or an optical cable. Using XAUI copper and optical cables enables larger transmission distances, and hence, implementation of interconnects in multi-rack systems.

Figure 23 shows upper limits of interconnects using 10G XAUI. Using first generation optical interconnects where optical transceivers are deployed on line cards in order to transmit optical signals using fibres over longer distances in a point-to-point manner, multi-shelf systems having capacities of more than 100 Tb/s are possible. However, several hundreds of Tb/s or even 1 Pb/s total capacities can not be reached.

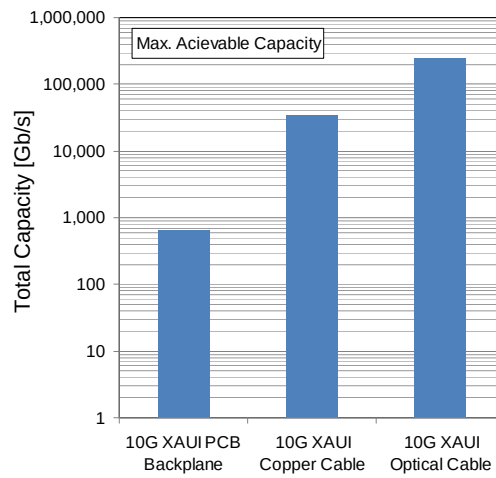


Figure 23: Limitations of 10G XAUI interconnects when using PCB backplane, copper cable or optical cable (first generation optical interconnects).

Another technology that also supports copper- and fiber-based connection in large-scale switches and routers is InfiniBand. The maximum achievable distances are not standardized, but commercial products have reaches of up to 200 m over optical cable and about 15 m over copper cable.

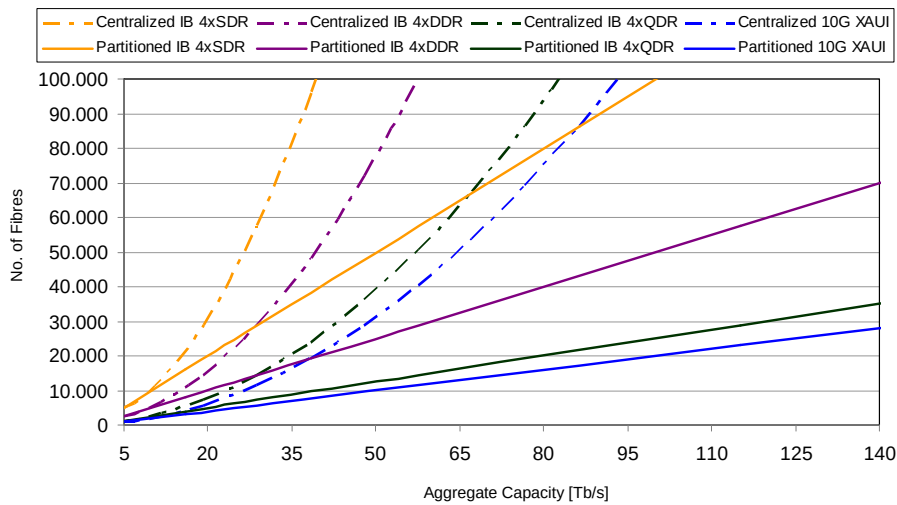


Figure 24: Number of fibers for different optical intermodule point-to-point interconnections

Regarding number of fibers needed for a particular aggregate capacity, partitioned architecture performs better as depicted in Figure 24. The interconnecting technology with the smallest number of fiber connections is 10G XAUI since it supports the highest line rate. Further technologies that only support copper connections are PCIe and SAS with distances of 7 m and 8 m over copper cable, respectively.

Figure 25 shows limitation of each considered interconnecting technology. With respect to other technologies considered in this document (InfiniBand options summarized in Table 1 only considered here), 10G XAUI performs the best.

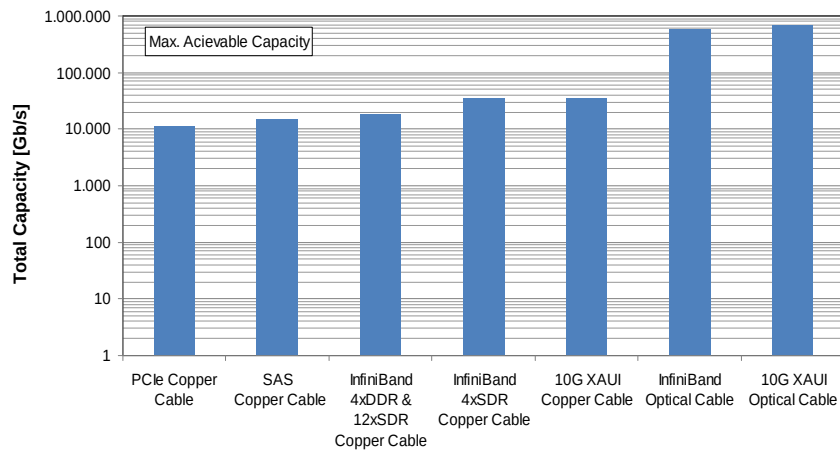


Figure 25: Limitation of optical and copper cable intramodule interconnects

Cable lengths needed to achieve a specific aggregate capacity when using parallel 100 Gb/s interconnects are shown in Figure 26. Since the maximum transmission length of the Ethernet future standard 100GBASE-SR10 is limited to 100 m, the achievable capacity is below 720 Tb/s.

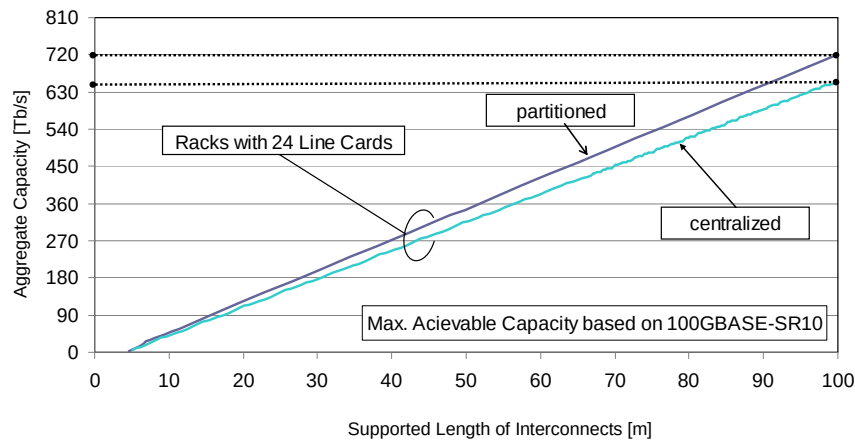


Figure 26: Maximum achievable capacity

VI. Requirements for Optical Interconnects

From the section II.B one can see that electronic interconnecting technologies are not in a stalemate, but the essential concerns that remain are cost, energy dissipation and how long can they keep up with ever increasing capacity needs because of the well-known physical electronic constraints. In this section the main topics related to optical technology, considered as a challenging alternative for interconnection, will be reviewed.

A. Requirements for optical interconnection networks in future high capacity routers

Support of future Internet services will require dynamic optical interconnections where resources are reconfigured at different time scales in relation to traffic granularities and quality of service needs. The emerging opinion is that multi-granular interconnections relying on hybrid technology can efficiently meet this task [35]. The resulting network should be able to adapt itself to application requirements on one side and benefits of different kinds of available technology on the implementation side. This network paradigm will employ Optical Circuit Switching (OCS) at the wavelength (and, eventually, waveband) level, Optical Burst Switching and Optical Packet Switching at subwavelength level. Multi-granular optical switches are then required with adequate re-configuration capability.

Optical technology has been applied to connect high capacity routers and offer huge transport capacity among them. The still open problem is how to employ optical technology at shorter distances, that is in board-to-board, chip-to-chip or even on-chip interconnections, or, in other words, ‘inside the box’. Many solutions of optical cross-connects and switch architectures have been proposed in the last decade which can be considered as a basis for this kind of investigations, at least in the board-to-board and chip-to-chip scenarios. In any case the short range characteristic of this new area of research will pose new challenges and opportunities in the design of optical interconnections.

Different aspects in optical interconnects need to be studied and analyzed, to provide a characterization in terms of performance and feasibility

B. Contention resolution in optical interconnection fabrics

The main problem to solve when interconnecting resources is to optimize resource design and sharing by adopting techniques to solve contention for resource usage. Contention arises in circuit, burst or packet switching at different time scales and influence call blocking or burst/packet blocking, depending on the transfer mode. This may result in delay or loss in information transfer, in relation to the particular contention resolution technique applied.

The techniques to solve contention are defined in relation to the applied domain. Options are space, time, wavelength or, even if still less investigated, code domains.

Contention resolution in space domain exploits physically separated resources to solve contention. An example can be the introduction of parallel alternative interconnection paths, typically available in short range interconnections. Contention resolution in the time domain relies on the availability of

devices which act as queues to delay the use of a contended resource. Queues are implemented by memory cards in electronics, while optical RAMs are not still available. Typical queueing solutions in optics are based on fibre delay lines (FDLs), which delay the information transfer of multiples of a basic delay unit. Contention resolution in the wavelength domain exploits the property of wavelength separation (typical in WDM) in the usage of an optical resource. When contention occurs on a wavelength channel, contending information are shifted to different wavelengths. This operation is performed by wavelength converters. Exploitation of the code domain for contention resolution in optical interconnects has been still less studied. It relies on the availability of optical coders/decoders which are able to change the code associated to the information to solve contention. The recent demonstration of the availability of these devices poses new challenges in the design of optical interconnects.

In any case contention resolution is performed by the electronic forwarding control which has the task to optimize information transfer in relation to contention occurrence and interconnection network structure. Forwarding algorithms and related software implementation are key components of the interconnection implementation. This software interacts with the interconnection network topology and implementation on one side and with routing control with the other side. A recent view of the organization of all these functionalities is proposed in [29].

C. Roles of wavelength converters in contention resolution

Output contention on wavelength channels can be solved by means of wavelength converters (WCs) which are able to shift the information from the encoding wavelength to a different one. Feasibility of these components has been widely demonstrated (electro-optical WCs are nowadays feasible, while all-optical WCs technology need further improvement). In any case they are active complex components whose number must be limited in simple interconnection matrices, both for cost and power consumption. Sharing schemes have been proposed in literature to design architectures where the WCs are a shared optimized resource.

Wavelength converters functionalities are of different types depending on the tuneability of input and output lines:

- FIFO (fixed input - fixed output)
- TIFO (tuneable input - fixed output)
- TITO (tuneable input – tuneable output)
- FITO (fixed input- tuneable output)

In addition the range of tuneability can be limited with respect to the full range of wavelength channels available on fiber interfaces.

In relation to these functional characteristics, different sharing schemes and the related interconnection architectures can be designed. The main architectural models are represented by:

- Shared per output link
- Shared per node
- Shared per (input or output) wavelength

These models allow different implementations based on single or multi-stage interconnection network organizations. Moreover they can be jointly applied with space domain contention resolutions as happens with multi-fiber interfaces. The space isolation operated by separated fibers on the same forwarding interface allows to separate the wavelength channels and to reuse the same wavelength on different fibers of the same interface with consequent reduction of the wavelength converters range.

D. How optics can support Quality of Service?

The QoS mechanism adopted in electronic domain cannot be inherited for optical interconnects due to the lack of random access memories. In the optical domain some basic QoS mechanisms must be provided to overcome the limitations in traffic performance consequent to contention and serve traffic with particular restrictions with some privileges.

The domains available to support quality of service in optics are the same which are applied for contention resolution. QoS differentiation is achieved by applying differentiated resolution schemes which gives priority to some flows instead of others. Few service classes can be adopted with the aim to maintain processing simple.

Examples of these strategies have been presented in literature and applies the following schemes:

- Wavelength channels reservation: wavelength channels are reserved for high priority flows, typically with some dynamicity.
- Queue scheduling: information in queues is served according to scheduling algorithm which resembles electronic solutions.

QoS differentiation schemes for asynchronous bufferless OPS, namely, the wavelength allocation algorithm (WA), the preemptive drop policy (PDP), and intentional packet dropping (IPD), have been analyzed and compared in [36]. The WA isolates the different service classes by reserving a certain number of wavelengths exclusively for highpriority traffic. In the PDP, QoS differentiation is obtained by letting high-priority traffic preempt low priority traffic currently in transmission in the case of contention. With IPD, low-priority packets are intentionally dropped with a certain probability, which results in higher packet loss probability (PLP) for low priority traffic compared with high-priority traffic. The results reported in show that in terms of throughput, the PDP has much better performance than the WA and IPD. As a matter of fact, in the PDP, packets are dropped only when contention occurs, while, in the WA and IPD, packets are dropped although resources are available. A QoS differentiation scheme for slotted OPS is proposed in [37]. That scheme provides differentiation thanks to wavelength allocation to the various traffic classes. A greater number of wavelengths is allocated for higher-priority packets. The scheme is highly efficient, as there is no penalty in the switch throughput when migrating from a best effort scenario to one with several classes of service (CoSs).

Algorithms which combines time and wavelength domain were presented in [38] and [39]. They consider both threshold based techniques applied to FDL buffers and wavelength reservation applied to the wavelength pool available to provide simple two-level quality of service differentiation with respect to packet loss. Some architectures allow to by-pass the FIFO queuing discipline by means of suitable scheduling applied to buffer in order to serve packets which need time transparency [40].

In order to manage quality of service within interconnection networks, forwarding control must implement the differentiation strategy, which turns into increased complexity in information scheduling.

E. Performance comparison criteria for optical interconnects

To identify the most promising solution for the mid-to-long term application in next-generation routers, optical interconnection fabrics must be compared under several aspects. The most important ones are here reported:

- *feasibility* in terms of optical devices employed to implement it and *scalability* as the number of fibers/wavelengths to be connected increases;
- *scheduling algorithms* applied to manage contention resolution and packet forwarding in a given architecture
- *performance* (in terms of loss and delay) in relation to the traffic supported;
- *power consumption* of the optical switching fabrics.

These aspects are also related each other (as an example, performance are related to the scheduling algorithm applied, power consumption is related to the electro/optical devices employed and so on).

As a matter of fact optical interconnects need to improve features of the electronic interconnects at least in terms of power consumption, scalability and footprint, otherwise their deployment is useless.

For this reason, a reasonable target for an optical interconnects is to reach an aggregated throughput equal to or higher than the one provided by electronic interconnects, while reducing power consumption.

The rest of this section gives some guidelines and criteria to evaluate and compare optical switching fabrics taking into account the above mentioned aspects.

VI-A-1 Feasibility/scalability

To ensure feasibility on large scale and not only test-bed implementations, an optical interconnection architecture should be designed taking already available or near-to-mature optical devices into account [41]. For this reason, the interconnection matrices should be designed taking the maximum number of ports of the devices into account, and the signal degradation and power loss introduced in the optical signal by any kind of optical devices should also be accounted. Example of works in this sense can be found in [42], [43], [44]. In [42], power loss and optical signal-to-noise ratio (OSNR) of many optical devices that can be employed to build an optical switching fabric are described, as well as different kind of simple fabrics. Also, an evaluation of the scalability (in maximum number of ports) and maximum aggregate throughput achievable is presented, taking the OSNR and power sensitivity of the receiver into account. The evaluation is done by analyzing the path of the optical signal along the fabric, taking the losses introduced by the optical devices the signal crosses into account. This paper provides a sort of reference, given that similar considerations have to be done for every kind of fabrics, in order to evaluate their scalability. In [43], more advanced multi-plane fabric solutions to interconnect different line cards are introduced, with an evaluation of the power budget and maximum aggregate throughput achievable tasking the power loss into account. In [44] other architectures are considered in terms of both scalability/power aspects and cost.

In the design phase a careful evaluation of the components that can be exploited is needed, to take into account the most advanced optical and electronic devices, in terms of switching speed, signal degradation, power loss, maximum size (in number of ports) and cost (per port) [45]. Furthermore, complex (but needed) devices such as WCs and optical gates must be kept as simple as possible and employed in a limited number.

Hybrid multi-granular switching fabrics can be considered to ensure easy implementation with available technology, while meeting the requirements of different traffic flows. An example of this kind of architecture, designed for optical switching nodes for core optical networks, can be found in [46], [47], [48]. Similar hybrid architectures can be employed as interconnection matrices, to reduce the number of expensive and power consuming optical devices. In this case, the scalability of these fabrics needs to be evaluated considering the power loss and OSNR introduced, taking the different available paths through the fabric into account.

Also, the connection of these components to implement the fabric must be (possibly) modular [43] to ensure scalability as the number of inputs/outputs and/or wavelengths increases. The target is to obtain an architecture where inputs and/or outputs can be easily added/removed (as well as the number of wavelengths) to increase the capacity of the whole interconnection when needed. In this sense, a modular organization of the fabric is preferable with respect to another one, composed by a single fabric.

VI-A-2 Control/scheduling algorithms

The scheduling algorithm adopted for contention resolution and packet forwarding plays an important role in interconnection fabrics. Indeed, some architecture needs complex and time-consuming scheduling algorithms (to properly control the available hardware in relation to the features of the fabric), which can limit the performance themselves. Other architectures can be managed with simple heuristic algorithms, such as i-SLIP algorithm proposed for electronic routers. In [49], a simple protocol, named multi-FASNET, was proposed to manage packet forwarding in a slotted WDM metropolitan area network (MAN), ensuring the fairness among transmitting nodes. Starting from this MAN architecture, a switching fabric to interconnect electronic line cards was proposed, and a distributed scheduling to manage packet forwarding in a slotted environment was presented in [50]. This scheduling algorithm is distributed among the line cards, so that each line card takes its own decision about the packets to be transmitted in a time slot.

Furthermore, in [51], the authors presented a test-bed implementation of optically switched data paths for computing applications. Basically, computers are connected through a switching fabric built-up with Semiconductor optical amplifiers (SOAs), and a central arbiter decides which hosts are allowed to transmit in a frame-by-frame basis. Each frame is composed by a fixed number of time slots, and the hosts, which have packets to transmit, send the requests to send to the central arbiter at the beginning of each frame. This is an example of centralized scheduling algorithm. A detailed description of the scheduling algorithm is proposed in [52].

The definition of the scheduling algorithm is really important when an interconnection fabric is proposed. Different options can be considered, and as a consequence the scheduling algorithms can be divided in different categories, here reported:

- synchronous/asynchronous: in the first case, the packets arrive at the inputs of the fabric at the same time, or are aligned at the fabric's inputs; packets have in general the same size equal to a time slot duration and, in this case, the scheduling algorithm decides in each time slot, or in a frame of more than one time slots, which packets can be forwarded and which are blocked. The scheduling algorithms proposed in [50], [52] are time slot based. In the second case, packets of variable size arrive at the inputs asynchronously, and the scheduling algorithm takes a decision just when a packet arrives.

- centralized/distributed: in a centralized scheduling algorithm [52], an arbiter collects all the requests to send packets and decides which packets can be forwarded and which not; in the distributed algorithms [50], each input line card is able to take its own decisions, avoiding conflicts with other line cards. Centralized algorithms provide best performance in terms of loss but require complex and slow procedures to communicate with all line cards, thus limiting the speed when the number of line cards increases. On the other hand, distributed algorithms are less efficient in solving contention, but the algorithm scale very well with the number of line cards.
- optimum/heuristic: optimum scheduling algorithm are able to maximize the throughput, but they are often complex and their computation complexity could be too high for an implementation in a real system; heuristic scheduling algorithms are less complex and less time consuming but provide (in some cases) lower throughput performance.

VI-A-3 Traffic performance

One of the most important parameter to evaluate an interconnection fabric is the performance it is able to provide, in terms of information loss and delay. Both parameters are related to the structure of the fabric and of course also to the scheduling algorithm applied to manage it, so performance evaluation is carried out once defined the architecture of the fabric and the related algorithm applied to manage the switching resources. It is very important to obtain a balance between loss/delay, especially in relation to the requirements of the input traffic (real-time applications are delay-sensitive while file transfer is loss-sensitive). When the same fabric is exploited for different kind of traffic, multi-granular fabrics can be defined to treat different traffic flows in different ways according to their needs. Performance evaluation can be carried out through analytical models, simulation or experimental results, and different architectures can be compared in terms of loss and delay to understand which is the best choice for a given scenario.

VI-A-4 Power consumption

Another important aspect for optical interconnects is the power consumption. As a matter of fact, one of the reasons why optics should be convenient with respect to electronics is that optical technology potentially provides lower power consumption in a smaller footprint than electronics. So, it is very important to evaluate the power consumption of the proposed fabrics, and to see whether or not they provide advantages with respect to electronic interconnections regarding power consumed. This aspect is really relevant as the ICT consumes a relevant fraction of the electricity (4% today), and this value is expected to increase rapidly in the future in relation to the capacity (and average per-user bit-rate) of the network. Power consumption have to take into account different aspects, in fact the power consumed increases as the bit-rate per channel increases, as the dimension of the fabric in terms of fibers/wavelengths increases (higher number of optical devices and so on).

Anyway, the fact that optics will provide power saving with respect to electronics is still a controversial, with pro and cons opinions. In [53], the authors state that optical switching routers, especially when based on SOA technology, will not provide benefits in consumed power with respect to electronics, when considering the optical network equipments. Instead, in [48] the authors show how a bufferless all-optical switch will provide power reduction with respect to electronic routers. For this

reason, a careful evaluation of the power consumption of different optical interconnection fabrics is necessary, taking different aspects into account.

References

- [1] IEEE Standard for Information technology—Telecommunications and information exchange between systems, IEEE Std 802.3™-2008 (Revision of IEEE Std 802.3-2005), 2008
- [2] IEEE Std 802.3™-2008 (Revision of IEEE Std 802.3-2005), Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) access method and Physical Layer specifications, Section 4, 2008
- [3] John D'Ambrosia, Shawn Rogers, Jim Quilici "XAUI: An Overview, Version 1.0", March 2002
- [4] American National Standard for information Technology - Fibre Channel - 10 Gigabit (10GFC), Project 1413-D, 10 Gigabit Fibre Channel, Rev 4.0, April 1, 2004
- [5] I/O IA # OIF-CEI-02.0 Common Electrical I/O (CEI) - Electrical and Jitter Interoperability agreements for 6G+ bps and 11G+ bps I/O, 28th February 2005
- [6] Interlaken Protocol Definition, A Joint Specification of Cortina Systems and Cisco Systems Revision: 1.1, July 25, 2006
- [7] Mark Gustlin, Fred Olsson, Matt Weber, "Interlaken Technology: New-Generation Packet Interconnect Protocol", White Paper, 08 March 2007
- [8] Interlaken Alliance, "Interlaken Interoperability Recommendations", Revision 1.3, 5 January 2009
- [9] PCI-SIG, "PCI Express Base Specification", Revision 1.0, April 29, 2002
- [10] Intel Corporation, "PHY Interface for the PCI Express™ Architecture", Version 2.00, 2007
- [11] PCI-SIG, PCI Express 3.0 Frequently Asked Questions, <http://www.pcisig.com>
- [12] Lisa Sherwin, "PCI Express external cabling specification completed by PCI-SIG", February 7, 2007
- [13] InfiniBand™ Trade Association, InfiniBand™ Architecture Specification Volume 2 Release 1.2, October 2004
- [14] Matthew J. Koop, Wei Huang Karthik, Gopalakrishnan Dhabaleswar, K. Panda, "Performance Analysis and Evaluation of PCIe 2.0 and Quad-Data Rate InfiniBand", 16th IEEE Symposium on High Performance Interconnects, 2008
- [15] Jiuxing Liu, Amith Mamidala, Abhinav Vishnu, Dhabaleswar K. Panda, "Evaluating InfiniBand Performance with PCI Express", IEEE 2005
- [16] Cisco Systems, "InfiniBand SDR, DDR, and QDR Technology Guide", white paper, 2006
- [17] Makia Minich, "Infiniband Based Cable Comparison", June 25, 2007
- [18] RapidIO Trade Association, RapidIO™ Interconnect Specification Rev. 2.0, 03/2008
- [19] RapidIO Trade Association, RapidIO™ Interconnect Specification Rev. 2.0, Part 10: Data Streaming Logical Specification, 03/2008
- [20] Optical Internetworking Forum (OIF), Ethernet Alliance, "CEI-28G: Paving the Way for 100 Gigabit", white paper, www.oiforum.com
- [21] Elizabeth A Donnay, David R Stauffer, "CEI-28G-SR Working Clause Proposal", Physical Link Layer (PLL) Working Group, July 23, 2008
- [22] Elizabeth A Donnay, Joel R Goergen, 2CEI-25G-LR Baseline Clause Proposal", Physical Link Layer (PLL) Working Group, July 23, 2008
- [23] John D'Ambrosia, David Law, Mark Nowell, "40 Gigabit Ethernet and 100 Gigabit Ethernet Technology Overview", November 2008
- [24] IXIA, "An Overview of Next-Generation 100 and 40 Gigabit Ethernet Technologies", May, 2009, <http://www.ixiacom.com/>
- [25] Abhijit Athavale, Carl Christensen, "High-Speed Serial I/O Made Simple", XILINX Connectivity Solutions Edition 1.0
- [26] Scott Irwin, "Usage Models for Multi-Gigabit Serial Transceivers", White Paper: XILINX Virtex-II Series, 2002
- [27] Cisco CRS-1 Carrier Routing System 16-Slot Line Card Chassis System Description, CISCO, March 2008
- [28] European Telecommunications Standards Institute (ETSI), ETSI Cabinets, <http://www.etsi.org/WebSite/homepage.aspx>
- [29] International Electrotechnical Commission (IEC), International Standard IEC 60297-3-101, 08-2004
- [30] D. R. Stauffer, J. T. Mechler, M. A. Sorna, K. Dramstad, C. R. Ogilvie "High Speed Serdes Devices and Applications", Springer, 2008
- [31] Karl Gass, Russ Tuck, Jeffrey Lynch "System Packet Interface Level 5 (SPI-5): OC-768 System Interface for Physical and Link Layer Devices", OIF-SPI5-01.1, September 2002
- [32] OIF-SFI-5-01.02, "Serdes Framer Interface Level 5 (SFI-5): Implementation Agreement for 40Gb/s Interface for Physical Layer Devices", January 2002
- [33] System Interface Level 5 (SxI-5): Common Electrical Characteristics for 2.488 – 3.125Gbps Parallel Interfaces, OIF-SxI-5-01.0, October 2002
- [34] Chris Hoffman, "6Gb/s SAS — An Evolutionary Step for SAS Technology", White Paper, 2008
- [35] G. Zervas, R. Nejabati, D. Simeonidou, C. Raffaelli, M. Savi, C. Devellder, M. De Leenheer, Dider Colle, M. Schiano, Programmable Multi-Granular Optical Networks: Requirements and Architecture, Broadnets 2009, Madrid, Spain, *Invited paper*, IEEE catalog number CFP09356, ISBN: 978-963-9799-49-3.
- [36] H. Overby, N. Stol, and M. Nord, "Evaluation of QoS differentiation mechanism in asynchronous bufferless optical packetswitched networks," *IEEE Commun. Mag.*, vol. 44, no. 8, pp. 52–57, Aug. 2006.

- [37] H. Overby, "Packet loss rate differentiation in slotted optical packet switched networks," *IEEE Photon. Technol. Lett.*, vol. 17, no. 11, pp. 2469-2471, Nov. 2005.
- [38] F. Callegati, W. Cerroni, C. Raffaelli, P. Zaffoni, "Wavelength and time domain exploitation for QoS management in optical packet switches", *Computer Networks*, Elsevier B. V., Vol. 44, N. 4, 15 Marzo 2004, pagine 569-582.
- [39] F. Callegati, W. Cerroni, C. Raffaelli, M. Savi, "QoS Differentiation in Optical Packet-Switched Networks", *Special Issue of Elsevier Computer Communications Journal*, Vol. 29, N. 7, 2006, p. 855-864.
- [40] C. Raffaelli, M. Savi, "QoS Aware Optical Packet Switch with Shared Electronic Buffers" *Broadnets/GOSP 2008, invited paper*, Londra, Settembre 2008.
- [41] S. J. B. Yoo, "Optical Packet and Burst Switching Technologies for the Future Photonic Internet," *Journal of Lightwave Technology*, 24(12):4468-4492, Dec 2006.
- [42] J. M. Finochietto, R. Gaudino, G. A. Gavilanes Castillo, F. Neri, "Simple Optical Fabrics for Terabit Packet Switches", 2008 IEEE International Communications Conference (ICC 2008) (Beijing (China) 19-23 May 2008 pp.5331-5337.
- [43] J. M. Finochietto, R. Gaudino, G. A. Gavilanes Castillo, F. Neri, "Multiplane Optical Fabrics for Terabit Packet Switches", 2008 IEEE/IFIP Conference on Optical Network Design and Modelling (ONDM 2008), Vilanova i la Geltru, Catalonia, Spain, 12-14 March 2008.
- [44] D. Cuda, R. Gaudino, G. A. Gavilanes Castillo, F. Neri, G. Maier, C. Raffaelli, M. Savi, "Capacity/Cost Tradeoffs in Optical Switching Fabrics for Terabit Packet Switches", *ONDM 2009: Conference on Optical Networking Design and Modeling*, Braunschweig, Germany, 18-20 February 2009.
- [45] R. Gaudino, G. A. Gavilanes Castillo, F. Neri, J. M. Finochietto, "Can Simple Optical Switching Fabrics Scale to Terabit per Second Switch Capacities?", *Journal of Optical Communications and Networking (JOCN)*, Vol. 1, Issue 3, pp. B56-B69.
- [46] M. Savi, G. Zervas, Y. Qin, V. Martini, C. Raffaelli, F. Baroncelli, B. Martini, P. Castoldi, R. Nejabati, D. Simeonidou, "Data-Plane Architectures for Multi-Granular OBS Network", *Proceedings of IEEE/OSA OFC 2009, Optical Fiber Communication Conference*, San Diego, USA, 22-26 March, 2009.
- [47] C.T. Politi, C. Matrakidis, A. Stavdas, D. Gavalas, M.J. O'Mahony, "Single Layer Multigranular OXCs architecture with conversion capability and enhanced flexibility", *Journal of Optical Networking*, 5(12):1002-1012, Dec 2006.
- [48] G. Zervas, M. De Leenheer, L. Sadeghioon, D. Klonidis, R. Nejabati, D. Simeonidou, C. Develder, B. Dhoedt, P. Demeester, "Multi-Granular Optical Cross-Connect: Design, Analysis and Demonstration", *IEEE journal of selected area in communications (JSAC)*, vol. 27, no. 4, April 2009.
- [49] A. Bianco, D. Cuda, J. M. Finochietto, F. Neri, C. Piglione, "Multi-Fasnet Protocol: Short-Term Fairness Control in WDM Slotted MANs", *IEEE GLOBECOM 2006*, San Francisco, CA, USA, 27-30 November 2006.
- [50] A. Bianco; E. Carta; D. Cuda, J. Finochietto, F. Neri, "A distributed scheduling algorithm for an optical switching fabric", in *proceedings of IEEE International Communication Conference (ICC 2008)*, Beijing, China, May 2008.
- [51] Glick, M.; Dales, M.; McAuley, D.; Tao Lin; Williams, K.; Penty, R.; White, I.; SWIFT: a testbed with optically switched data paths for computing applications, (invited) *Transparent Optical Networks*, 2005, *Proceedings of 2005 7th International Conference*, Volume 2, 3-7 July 2005 Page(s):29 - 32
- [52] Michael Dales, Madeleine Glick, "SWIFT: A high capacity wavelength-striped optically switched network with electronic control", *IEEE INFOCOM 2005 Poster/Demo session*.
- [53] R. S. Tucker, "Optical packet switching: A reality check", *Optical Switching and Networking* 5 (2008) 2-9.
- [54] V. Eramo, M. Listanti, "Power Consumption in Bufferless Optical Packet Switches in SOA Technology", *Journal of Optical Communications and Networking*, Vol. 1, Issue 3, pp. B15-B29.



JA2: Survey on photonic technologies for optical interconnections

State-of-the-art technologies for optical interconnection: optical vs. electrical wiring

Document attached to D25.2 of BONE WP25

Abstract— For transmitting very high data rates inside systems, photonics allows to overcome several limitations of electrical wiring. We show how this can occur on the typical interconnection distances rack-to-rack, board-to-board and even on-chip. The comparison between optical and electronic transmission is carried out both for serial and parallel links. Then in this survey we report about the state-of-the-art photonic technologies and devices to implement all the functions needed in an interconnection subsystem or backplane, including: signal generation, signal detection, amplification and switching. Latest advances of research in this field are reviewed. Finally, a comparative analysis of power consumption in optical and electronic interconnection systems is reported

Index Terms—Optical interconnects, photonic technology, transmission lines, integrated optics, optical switching, optical backplane, rack-to-rack, shelf-to-shelf, board-to-board, on-chip

1. INTRODUCTION	2
2. PASSIVE TRANSMISSION LINES.....	3
2.1 SINGLE-LINK TRANSMISSION.....	3
2.1.1 Backplane Level.....	5
2.1.2 Rack-to-rack level	8
2.1.3 On-chip level	10
2.2 PARALLEL-LINK TRANSMISSION.....	15
2.2.1 Backplane Level.....	15
2.2.2 Rack-to-rack Level	16
2.2.3 On-chip level	18
3. DEVICES FOR INTERCONNECTION.....	19
3.1 TRANSMISSION	19
3.2 DETECTION.....	21
3.3 AMPLIFICATION/REGENERATION.....	22
3.4 SWITCHING	23
4. POWER CONSUMPTION COMPARISON	28
4.1 SINGLE-SHELF SYSTEM	29
4.2 MULTI-SHELF SYSTEM	29
5. REFERENCES	30

1. Introduction

As increasing voice, audio, video, TV, and gaming traffic is carried over IP, Internet traffic continues to grow rapidly. Many network-related applications are emerging for portable devices. As smart cellular phone technology advances, the price decreases, and the infrastructure to support wireless applications (voice, data, video) is being deployed ubiquitously to meet unprecedented demands from users. All of these fast growing services translate into the high volume of Internet traffic, stringent quality of service (QoS) requirements, large number of hosts/devices to be supported, high speed packet processing, and large storage capability. Network operators need not only to expand their access networks with broadband technology, but also have to adapt the capacities of their optical core networks accordingly to the need for high-speed and high-capacity transmission/switching equipment. When designing/operating next generation switches and routers, these factors create new specifications and new challenges for equipment vendors and network providers.

High-performance router architectures are typically very large and highly distributed, with diameter spanning up to tens of meters, so the capability of covering such distances with high bandwidth and low latency is a significant advantage. As a matter of fact, electrical wiring and electronic technology only may potentially lead to system bottlenecks when interconnecting many node-interfaces or network-processing modules. This will constraint scalability of large-capacity networking systems. To address such expected hard-wired bottleneck problem, the development of an alternative interconnection technology that can replace the conventional method is urgently needed.

Optical interconnection is the most promising candidate for providing a solution for hard-wired deadlock. It is well known that optics is a valuable resource for transmission on long-distance link between the nodes of the network. The conjecture is that it can play an important role also “inside the box”, to implement short-distance interconnection between different modules of high-performance router/switches architectures. For transmitting high data rates over typical interconnection distances, fiber is far better suited than copper. Also, optical interconnects neither generate nor are sensitive to electromagnetic interference a clear benefit at multigigabit rates.

Advances in optical technologies, coupled with enormous demands for network bandwidth, are fuelling an increasing interest and amount of research projects in the field of optical interconnections. We intend to address here one key question: has optical-interconnection technology already the capability of scaling-up the switching capacity of routers and switches beyond the frontier set by the electronic interconnection systems?

This question does not have a simple yes/no answer: it requires analyzing systems and technology in order to identify the working conditions under which optical interconnects becomes an advantage.

We would like in this report to present an attempt of such analysis, based on an ample and updated survey on the state-of-the-art technology. The analysis is divided in two main sections, the first dealing with passive interconnection lines and the second dedicated to active optical devices. A final section provides a comparison between electronic and optical interconnections based on power consumption and energetic issues.

2. Passive transmission lines

In this section we present the passive transmission media typically used for optical interconnection and we compare them to transmission lines used for electrical wiring. We consider the different layers of interconnection with progressively smaller distances that can occur in a system, i.e.: rack-to-rack (1 - 100 m), board-to-board (backplane level) (25 cm - 1 m), and on-chip (< 2 cm). For each layer we find in which conditions optics becomes more convenient than electronic (from a technical point of view). We apply the analysis first to serial links, extending it then to parallel links.

2.1 Single-link transmission

On the side of electronic interconnections, high-speed systems adopt various types of microwave-guiding structures as transmission media. Popular configurations for rack-to-rack and board-to-board digital communications include various types of coaxial cables and twisted-pair cables with and without shielding. On smaller distances (backplane layer), Printed Circuit Board (PCB) either micro-strips or strip-line traces are often used today. Transmission-line interconnections built on-chip typically use structures that resemble the micro-strip, with either polysilicon or metal layers used for the conductors and a doped semiconductor substrate used as the underlying solid return path.

Surprisingly, one can model all these media with the same equations [1]. Provided there exists a well-defined uniform path for the flow of both signal and return current, and further provided that the conductors are closely spaced in comparison to the wavelength of the signals conveyed, the telegrapher's equations accurately model the propagation of electrical currents and voltages along the structure. The telegrapher's equations were developed in the late 19th century to explain the behavior of high-speed signals on long telegraph lines and they still apply today. In general the telegrapher's equations may be successfully applied to any transmission media having the following characteristics:

- at least two conductors, insulated from each other
- uniform cross section along the entire length of the structure
- small cross-sectional geometry relatively to the wavelength of the conveyed signals
- transmission distance much greater than the spacing between the conductors

For electrical interconnects, the distance that a signal can travel is limited by the distortions due to frequency-dependent loss [1]. The attenuation of the transmission line as a function of the carrier frequency is given by

$$A(f)[\text{dB}] = 8.686 \left(\frac{R_{DC}}{2Z_0(f)} \right) + A_s(f)[\text{dB}] + A_D(f)[\text{dB}] \quad (1)$$

where Z_0 is the characteristic impedance, R_{DC} is the Direct Current (DC) resistance, A_s is loss due to skin effect, and A_D is the dielectric loss.

The transmission loss associated with any conductive transmission media increases monotonically with frequency. From low frequencies to high, the slope of the loss curve changes in a predictable way, passing through various regions of operation. Figure 1 displays the attenuation functions of six distinct types of coaxial cabling, plus one typical PCB trace. The horizontal axis shows the operating frequency in Hz. The vertical axis shows cable attenuation in dB. Both axes use logarithmic scales.

Each curve may be divided into distinct regions, each one with a characteristic shape of the loss function. The sequence of regions, in order of increasing frequency, is as follows:

- RC region: at low frequencies (close to DC) the magnitude of the inductive reactance, ωL dwindles to negligibility in comparison to the DC resistance. All that matters below this point is the relation between the DC resistance of the line and its capacitance. Lines at such low frequencies are said to operate in the RC region.
- LC region: at higher frequencies the inductive reactance grows, eventually exceeding the magnitude of the DC resistance, forcing the line into the LC region.
- Skin-effect region: beyond the LC transition the internal inductance of the conductors (a mere fraction of the total inductance) becomes significant compared to the DC resistance. This development forces a redistribution of current within the bodies of the conductors. The redistribution of current heralds the arrival of the skin-effect region.
- Dielectric loss region: dielectric losses are present at all frequencies, growing progressively more severe at higher frequencies. These losses become noticeable only when they rise to a level comparable with the resistive losses, a point after which the line is said to operate in the dielectric-loss-limited region.

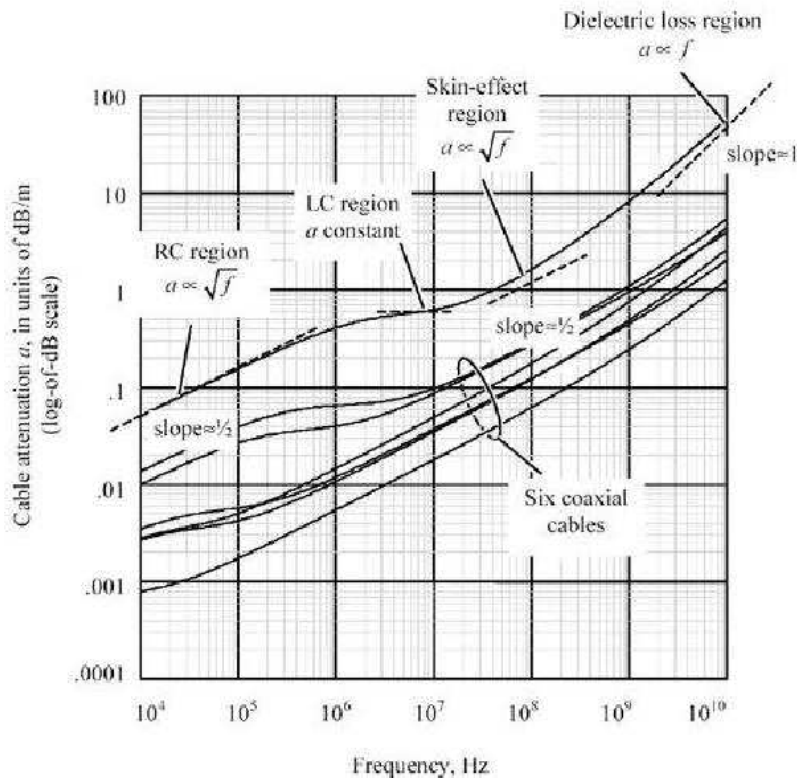


Figure 1 Attenuation curve of copper interconnection lines [2]

At high frequencies, dielectric loss dominates and the loss becomes directly proportional to frequency. This implies that, if the transmitted-signal bandwidth is large enough, contributes to not only power dissipation but also signal distortion.

Beside attenuation, also noise imposes limits on the communication channel. Because the exact calculation of the Bit Error Rate (BER) requires that we know detailed information about the system, (such as receiver technology, modulation format, line coding, etc.) it would not be suitable for a general scalability study. Thus, in this analysis we will use the Signal to Noise Ratio (SNR) as noise-related performance parameter, since loss and noise of channel are fundamental characteristics not related to any specific technology. The SNR is generally

calculated as the ratio between signal (average) power and noise power spectral density; the latter is integrated over the detection bandwidth.

The fundamental noise power spectral density limit for an ideal detector is given by [3]:

$$\psi_d = \frac{h\nu}{e^{h\nu/kT} - 1} + \frac{h\nu}{2} \quad (2)$$

where: ν is the frequency of operation, T is the equivalent noise temperature of the receiver (in Kelvin degrees), h and k are the Plank and Boltzman constants respectively.

Detectors in photonics ($h\nu \gg kT$) are typically shot-noise limited while electronic interconnects ($h\nu \ll kT$) are generally thermal noise limited. The difference between the quantum mechanical nature of light and the classical nature of electronics has lead to some debate over the limit on the power dissipation required for an optical link versus an electrical link.

Thermal noise is a function of temperature and quantum noise is a function of photon energy and thus of frequency. In the electrical channel, photon energy is significantly lower than the thermal energy and therefore, the overall noise power spectral density of the electrical link is determined by the thermal noise. In contrast, optical link performance is limited by the statistical nature of the photo-counting process performed by the detector, i.e. quantum noise. This is the dominant contribution, as the photon energy is well above kT .

Figure 2 indicates that the noise power spectral density of an optical channel operating at 850 nm ($\nu = 353$ THz) is 18 dB, almost two orders of magnitude higher than an electrical channel operating at 290 K°. Consequently, in order to achieve the same BER performance the optical receivers need 18 dB more signal power than the electrical ones.

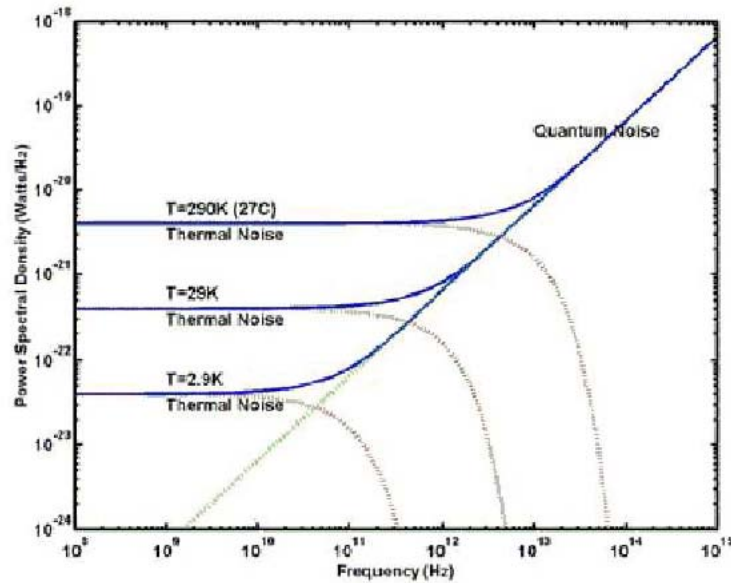


Figure 2 Power spectral densities of thermal noise and quantum noise vs. frequency.

2.1.1 Backplane Level

Let us consider more in details the interconnection at the board-to-board level, i.e. at distances typical of the backplane of switches and routers.

Electrical backplane implementations are typically based on the same microwave-guiding structures introduced above, i.e. coaxial cables and PCB strips.

Many different optical backplane constructions have been suggested and demonstrated, including glass fibers embedded in polymer sheets, polymer light guides [4], lens and mirror system [5],[6] diffractive optical elements [7], and active pointing with micro-electro-mechanical systems devices.

It is important to note that in addition to being low, the loss at optical frequencies does not increase with the modulated data rate but remains essentially flat over several THz. This makes the use of photonic devices much more attractive where the data rates are higher.

Electrical backplanes based on FR4 dielectric substrate have about 4.5 dB/m/GHz of dielectric loss [8], while PCBs with GETEK polymer as dielectric substrate [9] have 2 dB/m/GHz of dielectric loss. The use of low-loss dielectric, such as in the RO4000 series, can decrease the dielectric loss down to 0.6 dB/m/GHz: this allows for an increase the data rates of the backplane.

On the side of photonics, silica on silicon waveguides and polymeric backplanes can achieve a 2-5 dB/m loss.

In our study we have evaluated the loss of optical waveguides and electrical strips at two different propagation distances, representing typical intra-board (25 cm) and backplane (1 m) distances. We have assumed 2 dB/m material loss for waveguide links and an extra 1.5 dB loss at the two terminations of the link as coupling loss for connectors.

The comparison between the losses of electrical and optical backplane for 1 m distance shows that, as expected, electrical links are outperformed by photonics for transmission bit-rates higher than a certain threshold speed, while for lower bitrates electrical solutions do better. The crossover bitrate value is at about 1GHz, 2.5GHz, 8GHz for standard FR4, GETEK PCB and GETEK with Roger dielectric, respectively (Figure 3). However, when also the noise is taken into account, the 18dB difference between the noise power spectral density of the two channels shifts the crossover points at 5GHz, 10GHz and 40GHz (Figure 4).

We have carried out similar analysis for shorter links, 25 cm, without (Figure 5) and with (Figure 6) taking into account the channel noise. With channel noise included, the penetration of optics into-the-box is effective at data rates beyond 20GHz, 40GHz, 150GHz for standard FR4, GETEK PCB and GETEK with Roger dielectric, respectively.

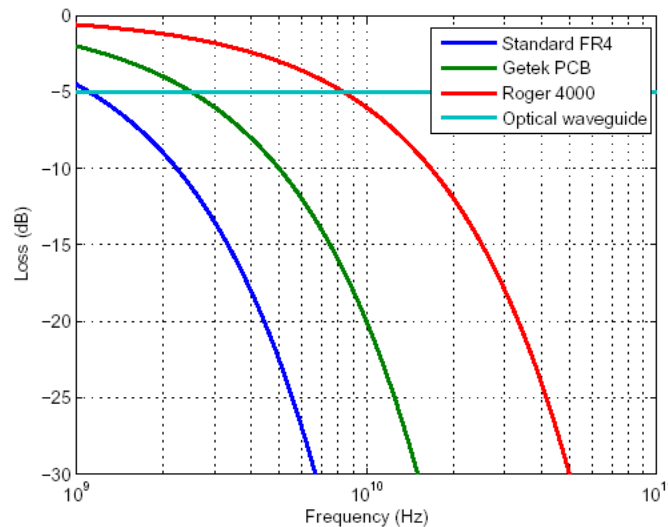


Figure 3 Crossover point backplane, channel noise not included (1 m).

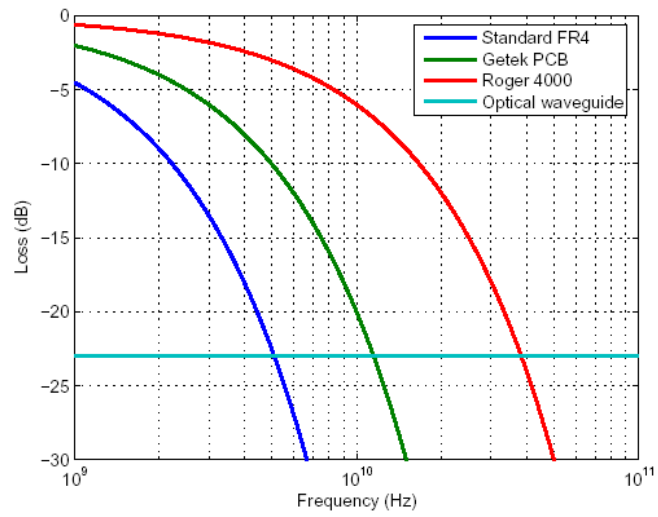


Figure 4 Crossover point backplane, channel noise included (1 m).

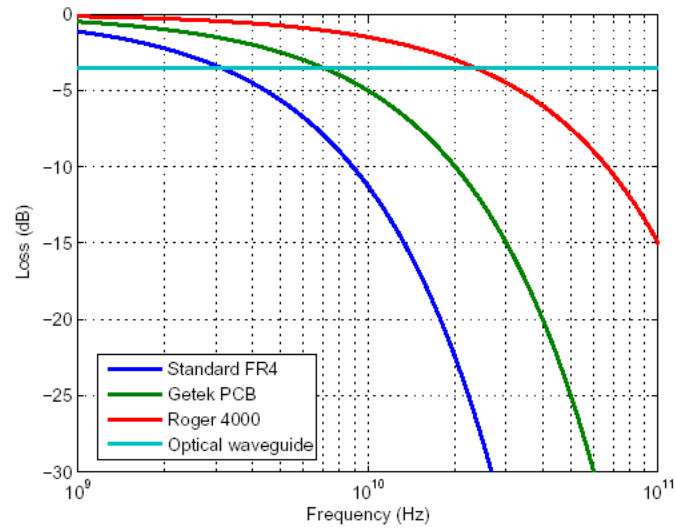


Figure 5 Crossover point backplane, channel noise not included (25 cm).

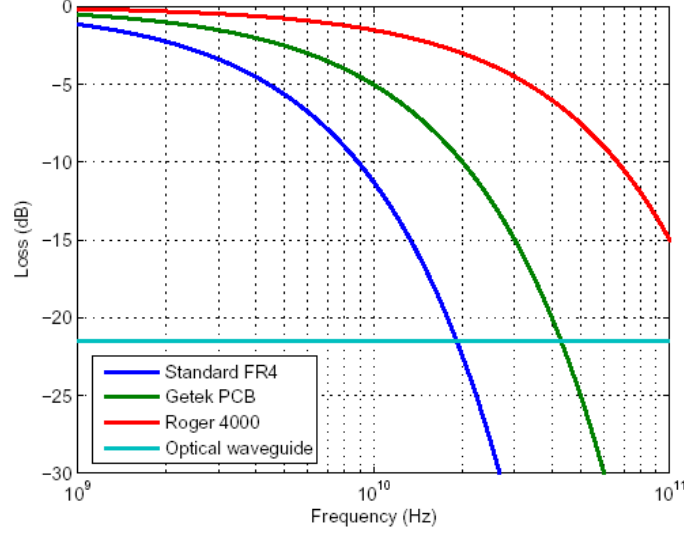


Figure 6 Crossover point backplane, channel noise included (25 cm).

2.1.2 Rack-to-rack level

For cabling between racks, low loss coax cable using twinaxial pairs (MT125C) [10] can achieve 0.3 dB/m/GHz; low-loss coax cable using ultra low density (UFA125A) [11] can achieve 0.07 dB/m/GHz. Both these solutions allow data transmission at rates up to 50 Gbit/s over 20 m distance. Greatly improved over traditional technologies, cables twinaxial pairs employs a unique homogeneous construction with 100% shielding that enables high data rates over a single line. Crosstalk, delay skew and amplitude skew are minimized because the one integral dielectric eliminates material variations and forces a nearly identical distance between conductors to be maintained (this is more important for parallel than single links – see next sections). The corresponding reduction in mux/de-mux requirements as well as a reduction in both size and quantity of carriers can greatly simplify high speed systems.

Coax cable with ultra low density dielectric is constructed using PTFE material, offering excellent loss characteristics, outstanding phase stability, and unsurpassed flexibility compared to standard flexible cables, all without sacrificing mechanical integrity.

The photonic counterpart to electrical rack-to-rack connection can benefit from the very low-loss propagation media provided by the single-mode silica optical fiber with its 0.2 dB/Km (at 1.55 μm).

Similarly to what was done for the board-to-board interconnections, we have compared the loss of optical and electrical links at three transmission distances typical of rack-to-rack, i.e. 10 m, 50 m and 100 m. As usual, electrical links are better than optical for low transmission bitrates, up to a crossover bitrate value after which optical fiber becomes more convenient. If the noise is included, over a distance of 10m the crossover bitrates are at 7 GHz and 30 GHz, for MT125C and UFA125A, respectively (Figure 7). However, increasing of the length to 50 m (Figure 8) the crossover points shift at about 1 GHz and 6 GHz. Finally, for longer links, 100m (Figure 9), the only solution comparable with optics is the MT125C with a crossover point at 3 GHz.

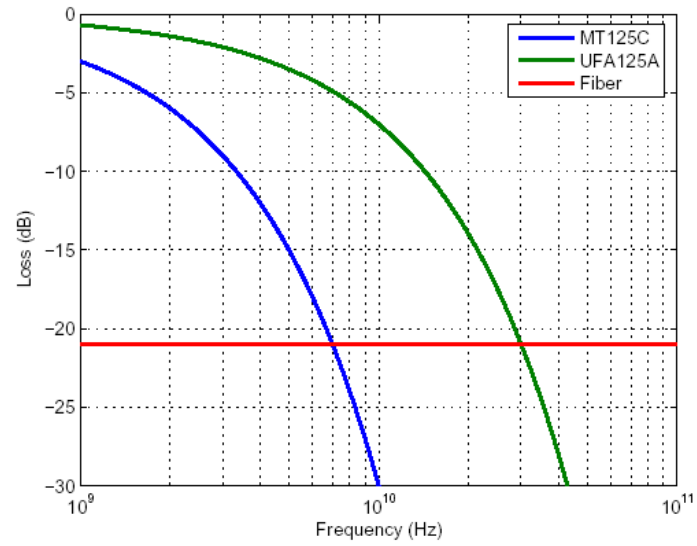


Figure 7 Crossover point rack-to-rack, channel noise included (10m).

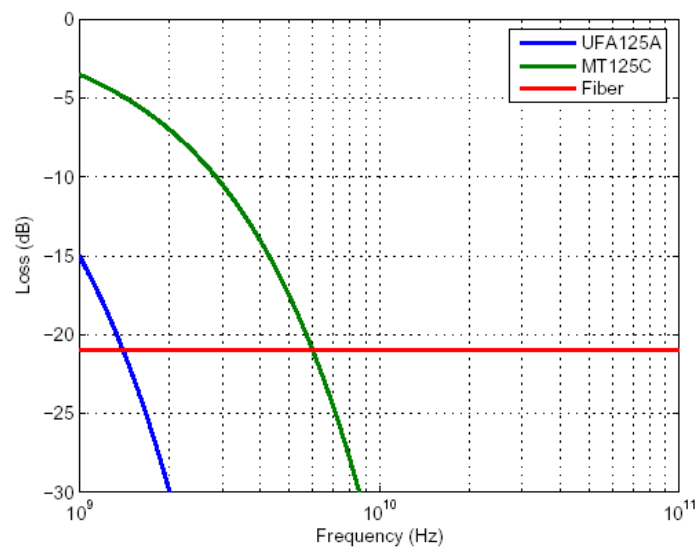


Figure 8 Crossover point rack-to-rack, channel noise included (50m).

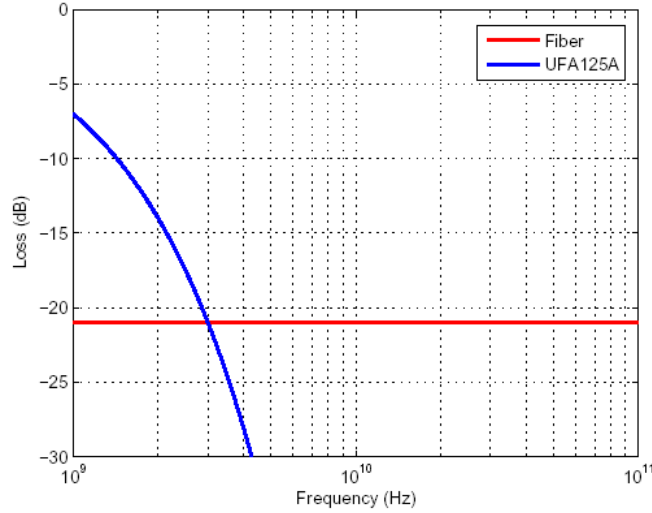


Figure 9 Crossover point rack-to-rack, channel noise included (100m).

2.1.3 On-chip level

The current trend of integrating tens or even hundreds of processing elements into a single die of semiconductor raises the importance of suitable interconnections among them in order to exploit the advantages of multi processor architectures at a low cost in performance. Different issues must be taken into account in order to provide point to point communication solutions between on-chip processing cores. Considering this on-chip scenario, the links used for interconnections are required to provide an overall system demand of low latency, bandwidth and compactness.

From the optical point of view, recent breakthroughs in CMOS-compatible silicon photonics suggest that the integration of optical components is likely to become feasible in the near future [12]. The presented comparison is performed between electrical and optical global interconnects for an on-chip scenario.

The absence of efficient monolithic on-chip light sources constitutes an important challenge of component integration in silicon. The most suitable model for optical communication in this scenario is made with an on-chip modulator, a waveguide for the optical link and a photodetector at the receiver (see Figure 10). The work done in [13] (or the theoretical considerations in [14]) highlights the importance to consider an off-chip laser source in order to save on-chip power, area and cost.

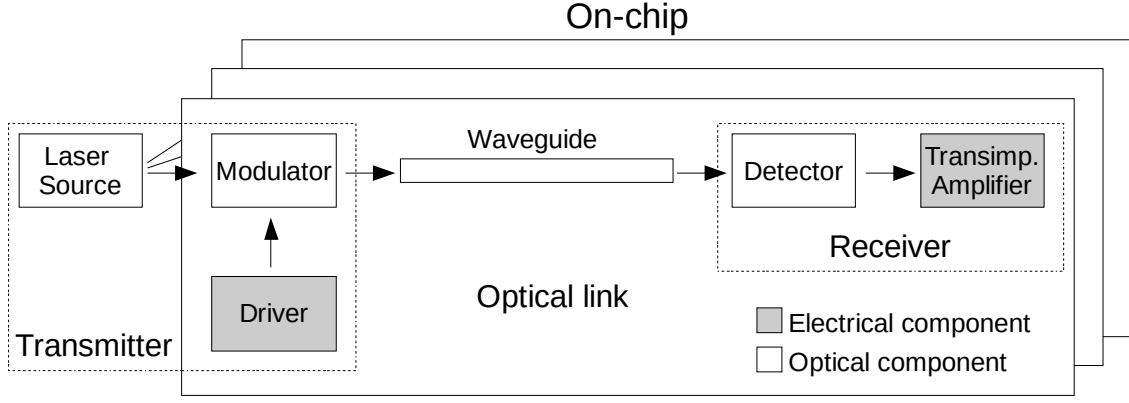


Figure 10 Main components in an on-chip optical transmission

It is important to highlight that the system performance will depend on the minimum optical power required by the receiver and on the efficiency of passive optical devices used in the system. The total loss in any optical link (as presented in Figure 10) is the contribution of all optical components. Different sources of loss can be identified from the laser source to the detector. These contributions can be separated into several losses through the optical link as depicted in Figure 12.

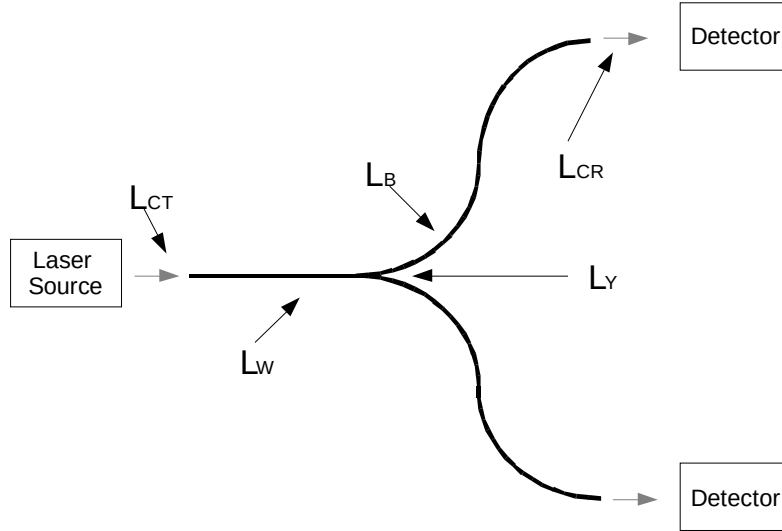


Figure 11 Main losses in an on-chip optical transmission

Where the different losses are:

- L_{CT} : Coupling coefficient between the photonic source and the optical waveguide. Given that this coupling involves off chip coupling, generally at most 50% coupling efficiency [15] can be achieved.
- L_W : Waveguide transmission loss is defined as the attenuation per unit distance of the optical power. It is important to mention that the waveguide dimensions and the contrast between core/cladding refractive indexes determine the loss due to the side-wall scattering. It is briefly discussed in the following as a function of the core material.
- L_B : Bending loss is generally expressed as the attenuation due to the curvature of the waveguide. L_B is highly dependent on the refractive index difference $\Delta = (n_1^2 - n_2^2)/2n_1^2$ between the core and the cladding: the larger is Δ , the smaller are bending

losses. When considering a silicon waveguide (core Si, $n_1 = 3.5$ and cladding SiO_2 , $n_2 \approx 1.5$), Δ is high enough to allow a bend radius of a few μm . It can be assumed that bending losses are negligible if the radius of curvature is greater than $2 \mu\text{m}$ for a Si/SiO₂ waveguide [15].

- L_Y : Is the loss due to the Y-branch or splitters through which light is routed in order to reach several receivers. For high Δ , L_Y is significantly lower than for the low Δ waveguides. Additional to power splitting, Excess losses associated to splitting, reported in [15], are less than 0.2 dB per 50/50 split.
- L_{CR} : Is the coupling loss from the waveguide to the optical receiver. It is possible to achieve an almost 100% coupling efficiency from the waveguide to the optical receiver. In [15] L_{CR} is assumed conservatively to be 0.6dB.
- $L_{\text{total}} = L_{CT} + L_W + L_B + L_Y + L_{CR}$
- Hence, the sum of all these source of losses can be expressed as a function of the interconnect distance, l , and the number of splitters per path, y , without considering stringent bending radii.
- $L_{\text{total}}(l, y) = 3.6 + L_W \cdot l + 3.2 \cdot y$

Since waveguides are the paths through which the light is routed, they play a fundamental role in the optical link. Geometrical considerations and the index contrast between the core and the cladding must be taken into account in order to determine propagation delay and bandwidth density (defined as the number of bits that can be transmitted per unit of surface area).

From the geometrical point of view, Figure 12 depicts the distances involved: waveguide width W , the separation S and pitch P between two waveguides.

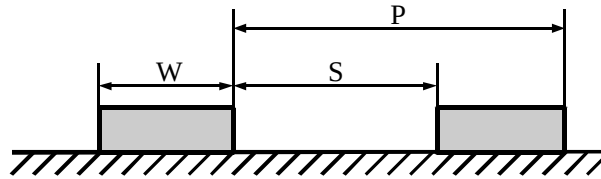


Figure 12 Width, separation and pitch detail

It is important to mention that for a particular waveguide material, an optimum ratio exists between the waveguide width W and pitch P (separation between components in silicon). For a fixed P , if a waveguide is too wide, the crosstalk turns to be relevant, due to the proximity between the sides of adjacent waveguides.

From the index contrast between the core and the cladding perspective, silicon (Si) and polymer seem to be the most promising materials for the waveguides' fabrication. Some of their most relevant features are given in Table 1.

Waveguide	Si	Polymer
Refractive index, n	3.5	1.49 - 1.6 [16]
Width (μm)	0.5	5
Separation (μm)	5	20
Pitch (μm)	5.5	25
Delay (ps/mm)	10.45	4.93

Table 1 Characteristics of silicon and polymer on-chip waveguides [18]

Regarding silicon, we can infer from Table 1 that a high refractive index difference $\Delta = (n_1^2 - n_2^2)/2n_1^2$ between the core (Si, $n_1 = 3.5$) and cladding (SiO_2 , $n_2 \approx 1.5$) allows a small waveguide's width. Therefore, it is possible the realisation of a compact optical circuit with pitch and separation dimensions of the order of a few μm . In addition, the same silicon that forms the transistors can be used for the silicon waveguides. This compatibility with standard CMOS current processes does not require the development of a new CMOS technology.

Despite these advantages, the higher refractive index of the silicon core waveguides compared with the polymer, results in a lower propagation speed, and motivates considering polymer waveguides. Two important characteristics of polymer waveguides are identified in [16]:

- Different materials can be considered for the cladding, from Teflon ($n_2 = 1.3$) to a wide range of polymers' refractive indexes ($n_2 = 1.49$ to 1.61). Nevertheless, the small index contrast Δ between the core (n_1) and cladding (n_2) exhibited by polymer waveguides requires larger dimensions (width, separation and pitch) than for silicon. This reduces the bandwidth density achieved by polymer waveguides.
- The integration of polymer waveguides could become limited because of the absence of suitable polymer-based modulators. Since their refractive index can be weakly modulated, they are bulky and require high driving voltages for high frequency operation, this could limit their applicability to on-chip optical links.

From the loss point of view, Table 2 presents the losses per unit lengths for several silicon and polymer waveguides.

Material	Silicon	Silicon	Polymer NOA 61	Silicon	Polymer UFC 170	Polymer SU-8
Loss [dB/cm] ($\lambda=1.55 \mu\text{m}$)	0.3 [17]	0.8 [15]	1.1 [16]	1.3 [18]	3 [16]	4 [16]

Table 2 Different losses per unit length for several silicon and polymer waveguides

Concretely, [17] presents low loss silicon waveguides at $\lambda=1.55\mu\text{m}$ with propagation losses of 0.3dB/cm for 1 μm width waveguides and 0.5dB/cm for 1.5 μm width waveguides. Moreover, [17] also mentions that the slightly higher losses from the 1.5 μm waveguide are most likely due to its multimode nature.

On-chip electronic interconnects (EI) use copper wires that can be characterized by resistance, capacitance and inductance [19]. Long wires required for global on-chip interconnects tend to exhibit high RC time constants, which increase the delay, raise/fall transition times, and suffer from crosstalk. In order to face these issues, CMOS technologies enable the use of repeaters to separate long distances into smaller segments. In this way (and by properly optimizing the inter-repeater distance) the delay becomes practically linear with the link length, instead of quadratic. In fact, the delay is reduced when the number of repeaters increases, but there is a delay and power penalty associated with each repeater. For interconnects with a fixed geometry, there is an optimal repeater size and spacing in order to achieve a minimum delay [20].

Assuming EI with optimized repeaters insertion, the delay can also be reduced by increasing the interconnect width at the expense of a smaller bandwidth density. The study in [14] showed that technology scaling (that is, the decrease in the element sizes on silicon) has insignificant effect on delay-optimised EIs with optimal number of repeaters. The minimum achievable interconnect delay remains fixed at around 20 ps/mm for all technology nodes. Hence, the bandwidth density can be further increased by changing the wire pitch (as depicted in 0).

Summarizing, EIs can be optimally designed either for bandwidth density but presenting a large RC impedance (which reduces their delay), or for delay, reducing bandwidth density in favour of a higher propagation speed by using wider wires. The following comparison between optical and electrical interconnects is focused on global links, and since EIs are typically delay-limited in long on-chip distances, only delay-optimised EIs are considered.

As previously explained, the propagation speed in the optical waveguide depends on its core material. Hence, assuming a delay of 10.45 ps/mm for the silicon-core and 4.93 ps/mm for the polymer-core waveguides, Figure 13 presents a comparison of both optical interconnects with the EI (fixed at approximately 20 ps/mm for node at different successive technologies).

The discussion presented in [14], based on a projected (by the International Technology Roadmap for Semiconductors – ITRS [21]) chip edge size of around 17.6 mm, highlights the conversion delay required (electronic-optic-electronic) for silicon and polymer waveguides in order to satisfy the delay of an equivalent electronic global interconnect. We can infer from the y axis on Figure 13 that the conversion delay allowed on a silicon link is almost 180 ps while in polymer waveguide links is almost 280 ps.

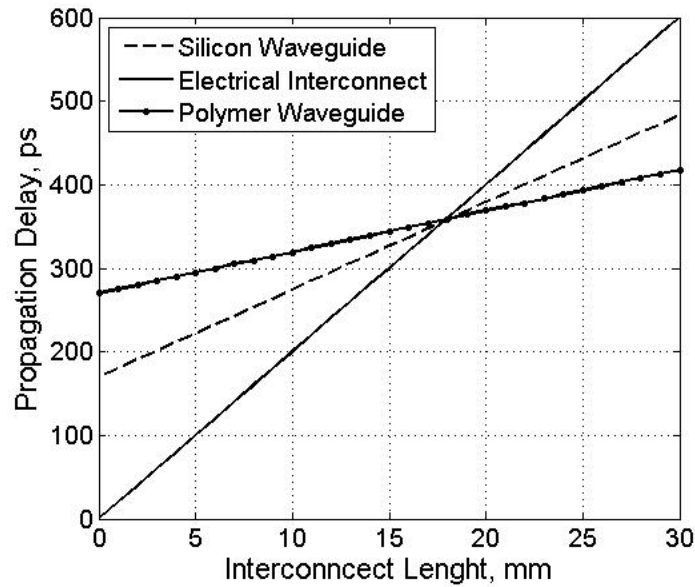


Figure 13 Optical and electrical delays for different interconnection lengths

Bandwidth density is the metric that characterises the number of bits per second that can be transmitted per unit surface area. We have seen that the pitch required in optical links is about 5.5 μm for a silicon-core waveguide and 25 μm for a polymer waveguide. EIs are spaced around 5-7 technology node sizes when a delay-optimised design is considered. Therefore, EIs have a significant advantage in bandwidth density in comparison with single-wavelength optical interconnects (see Figure 14), since more conductors can be concentrated in the same area. Moreover, EIs can also efficiently exploit repeaters at every successive technology node. The increase in optical bandwidth density shown in the graph is due solely to the higher bit rate allowed at each technology node through the waveguides with a fixed pitch.

Consequently, single wavelength optical link is inferior to delay-optimised electrical wire in terms of bandwidth density.

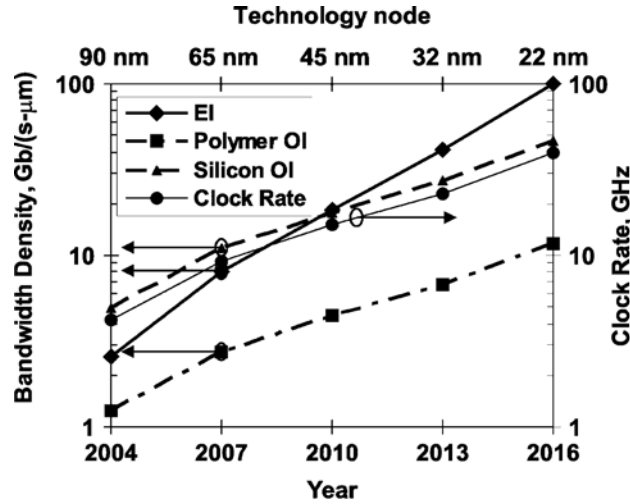


Figure 14 Optical and electrical bandwidth density as a function of technology nodes / year [14]

2.2 Parallel-link transmission

Let us now consider the case of multiple parallel links. Electronic wiring allows implementing bundles of parallel links with a higher density compared to optical interconnections. On the other hand, photonics offers a very important resource to exploit in achieving parallelism, that is wavelength division multiplexing.

2.2.1 Backplane Level

In addition to loss and SNR we have already discussed for single links, in parallel link other constraints must be taken into account while designing the system.

The co-presence of several signals travelling in parallel along very close links makes power leaked by one channel to couple and thus interfere with the adjacent or close ones, thus leading to the crosstalk phenomenon. Moreover, the parallel links must be connected to the sources and terminations of signals on the pieces of equipment in communication. The connector of each single link of the parallel set often has a footprint larger than the cross section of the link itself. When putting together more links in a parallel set (e.g. a ribbon or a 2-D matrix), this is an important aspect to consider.

Both crosstalk and connectorization can limit the scalability of parallel links, as we are going to discuss in the following.

Crosstalk is a major source of noise that can limit wiring density. Generally the total crosstalk component affecting a signal should be limited to 5% of the signal swing (-13 dB), in order to maintain signal integrity.

Low-loss laminate materials like Rogers 4000 have been developed for printed circuit boards. The raw material is five times more expensive than FR-4, and more difficult to process into high layer count boards. However, it does provide higher bandwidth density. With Rogers 4000 more than 1 Gbit/s/mm bandwidth density can be obtained with modest transceiver technology while only 400 Mbit/s/mm is available with FR-4 [22].

Standard printed circuit boards are constructed from etched copper foil traces on FR-4 laminate with plated through holes. Typically, these boards are manufactured with trace width of 100 μm . The edge-to-edge spacing between traces should be greater than 150 μm , to obtain crosstalk limited to 5%, and the center-to-center trace pitch is 250 μm , equal to the center-to-

center fiber spacing in a fiber ribbon cable. In this case, fibers have no advantage in wiring density at the same data rate.

In a study [23] on the maximum achievable speeds in interconnection systems that use traditional printed-wiring boards have been presented. At 2.4 Gbit/s the systems output eye patterns revealed that standard FR4, GETEK PCB and Roger dielectric materials permit satisfactory system operation. At 4.8 Gbit/s, examination of the eye patterns reveals that the system performance of FR4 and GETEK may not be acceptable. In contrast, Rogers provides enough improvement to offer potentially acceptable system performance. At 9.6 Gbit/s, the system yielded unacceptable results using any of the dielectric materials. At speeds approaching 9.6 Gbit/s, Roger dielectric materials enhance trace performance, but system components, such as the connectors and plated through holes, tend to dominate signal losses. At 9.6 Gbit/s, the comparison of system to trace-only results shows that connectorization is responsible for a significant signal degradation. Whereas trace-only performance at 9.6 Gbit/s was reasonably acceptable, the system performance at the same speed was unacceptable, even with Roger dielectric materials. Beyond 9.6 GHz, with the current technology, optics is necessary independently of the loss.

However, signal processing techniques can be used to increase bandwidth on FR4 boards. For example, Lucent Technologies' Bell Labs and FCI announced the successful demonstration of 25 Gbit/s data transmission over an electrical backplane using a Bell Labs-developed new duobinary line-coding technique (that uses three electrical signal levels to represent binary code in a communications transmission) [24]. The team achieved error-free performance at 25 Gbit/s over electrical traces as long as 60 cm using a backplane made with FR4 laminate. These methods will extend the useful lifetime of FR4 technology.

In optics the crosstalk is primarily caused by light coupled to the waveguide from adjacent VCSELs in the array. This kind of crosstalk can be eliminated by using VCSELs that have a low divergence angle. In fact, experiments indicate that polymer multimode waveguides have shown an impressive crosstalk of -30 dB for 10 μm waveguide spacing [25]. This allows high density guide circuits and connectorization schemes. Typical crosstalk value for 250 μm spacing waveguide is lower than -40 dB. Single mode guides have been shown to evanescently couple up to about 20 μm spacing; thus waveguides must be further separated to insure no crosstalk for guides with matched propagation coefficients. However, given that packaging and alignment requirements with single-mode waveguides are intrinsically more demanding, they can be partially mitigated by using multimode waveguides.

With current VCSEL technology, an optical transceiver can transmit up with 12 SERDES (serializer/deserializer) channels, each running at 2.5 or 3.125 Gbit/s (also 10 Gbit/s [26]). A work [27] presented a backplane architecture exploits 850 nm VCSEL and photodiode arrays that can be mounted on each of the line cards and connected to the backplane via ribbon fibres. The waveguides links have a $50 \times 50 \mu\text{m}^2$ cross-section to optimize coupling with standard 50/125 μm multimode ribbon fibre, while the waveguide center-to-center separation is 250 μm to match the pitch of the ribbon fibre. The reported backplane demonstrator uses a matrix of 100 waveguides each capable of 10 Gbit/s operation to interconnect 10 cards for a total capacity of a Tbit/s aggregate data rate.

2.2.2 Rack-to-rack Level

For cabling between racks, low loss twinax cable MT125C can achieve more than 40 Gbit/s, while UFA125A can allow up to 50 Gbit/s over 20 m². Both of these solutions have an outer diameter of 3.17 mm. The connectors have been optimized to achieve the lowest possible Standing Wave Ratio (SWR) across the bandwidth. In addition, the connector attachment has

been designed to provide high reliability and withstand heavy stress. The connector body, dielectric, and center contact are completely captivated guaranteeing the cable assembly will keep its excellent properties even after hard use. These solutions support plugs and jacks that are 2.4 mm wide.

In optics, the transmitter may contain a linear array of 12 VCSELs operating at 850 nm; the receiver version of the module consists of an array of surface-illuminated photodiodes, with amplifiers to drive the output electrical signal. Although VCSELs can be produced to operate at wavelengths from approximately 650 nm to 1550 nm, the 850 nm wavelength is chosen primarily because it is the wavelength for which the lowest-cost and highest-reliability devices can be found from multiple vendors. Commercial modules commonly operate at up to approximately 3.3 Gbit/s per channel, for an aggregate throughput of approximately 40 Gbit/s for a module pair. The fiber that connect the receiver and the transmitter modules consists of 1×12 linear array of fibers on a 250 μm . The module connector has a wide approximately of 17.5 mm.

Reusing the 12 VCSELs at 10 Gbit/s per channel is possible [28], achieving a 120 Gbit/s link. Since the power dissipation of the 10 Gbit/s is nearly the same as at 3 Gbit/s, the thermal properties of the module package are sufficient for heat-sinking. Experiments [29] were conducted on a range of different fiber lengths from 10 m to 300 m. For the 300 m experiments, next generation multimode 50 μm fiber (OM-3) was used. During one test in which the link was operating at 120 Gbit/s, nine errors were observed in 68 hours for a BER better than 10^{-14} .

To get denser optical links, an array with more than two rows of VCSELs is needed. However it is very difficult to assemble an array at the $250 \times 250 \mu\text{m}$ pitch of typical multimode fiber ribbons and connector using wire bonds. This problem has been solved by flip-chip bonding between the driver electronics and the O/E array [30]. A virtue of the flip-chip attach is that it introduces less parasitic series inductance than wire bonding, which is helpful for high frequency interconnects.

48 (or 4×12) VCSELs on a 48-channel driver integrated circuit is presented with a compact width of 4.6 mm. One feature of this module is that it reduces the fiber count by a factor of 4 by using CWDM. The overall bandwidth is 10 Gbit/s in each of the 48 channels.

In Figure 15 we report the total output bandwidth from a board of (theoretical) width from 25 to 100 cm comparing the electrical to the optical solutions. In our analysis we have considered, for each technological solution, the maximum bandwidth supported and the typical width of the connector more suitable for that specific solution. Results show that electrical connectorization offers a bandwidth density greater than the 1×12 traditional VCSEL, due to the dimension of the optical module. However, in 4×12 VCSEL modules flip-chip bonding allows a denser connectorization which leads optics to be more attractive than electronic.

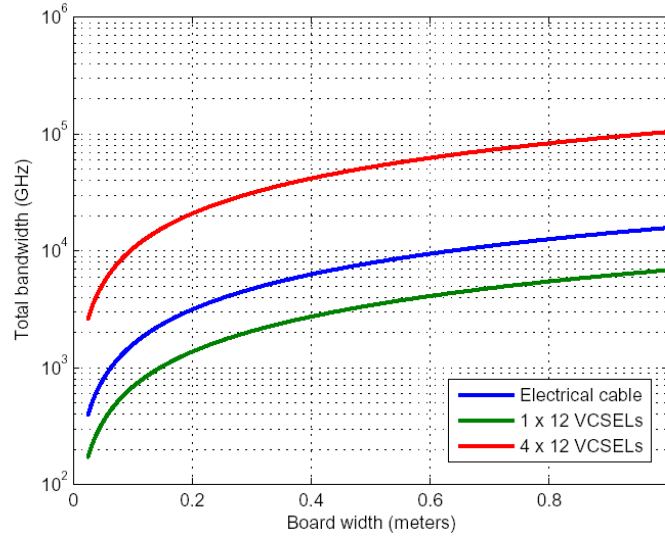


Figure 15 Total bandwidth as a function of the board width.

2.2.3 On-chip level

Wavelength division multiplexing (WDM) seems the proper interconnect evolution in order to increase the bandwidth density in optical links. The number of WDM channels required for silicon and polymer waveguides to reach the bandwidth density of the EI is shown in Figure 16.

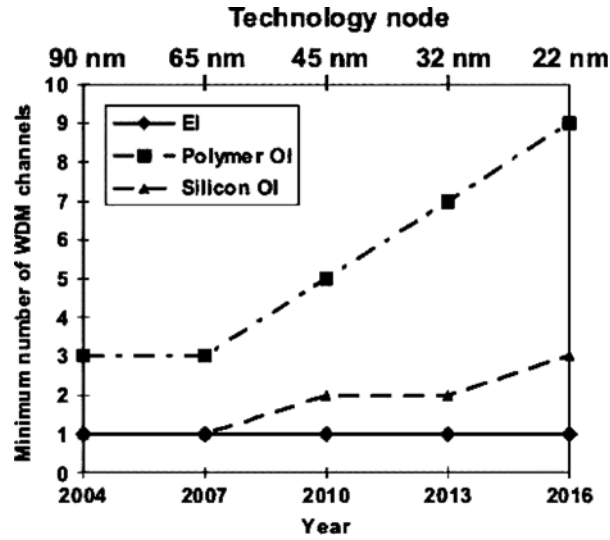


Figure 16 Number of WDM channels required to reach EI bandwidth density as a function of technology nodes / year in [14]

We can infer from Figure 16 that polymer waveguides require more WDM channels to match the bandwidth density due to their higher pitch size. However, because of their higher propagation speed they allow a considerable conversion delay overhead. On the other hand, silicon waveguides do not require a high number of WDM channels, but their higher delay compared with polymer waveguides implies faster (E-O-E conversion) transmitters and receivers.

Finally, it is important to mention that for every additional WDM channel considered, several parameters, such as the insertion of different modulators for each channel or wavelength selectors at the receiver (and their required area) must also be taken into proper account.

3. Devices for interconnection

In this section we present a set of active components that the state-of-the-art photonic technology makes available today for optical interconnection applications. We have classified the components in six categories, according to the function they would be dedicated to in an interconnection system. Thus, we have technologies for: transmission, detection, re-amplification and regeneration, switching. For each class we will discuss the best conditions of usage, state of development and performance limitations.

3.1 Transmission

Interfacing moderate-line-rate, widely-parallel, electronic busses to optical transmission media requires unprecedented data rates, stringent packaging and energy consumption requirements. Key areas of current research include:

- Parallel optical links
- Advanced multiplexing formats
- Energy efficiency
- Photonic integration
- Nanoscale miniaturisation

Parallel fiber-optics have already enabled impressive performance metrics in terms of energy consumption at tens of milliwatts per Gbit/s. Vertically emitting laser arrays have been efficiently used in combination with parallel multimode fiber transmission. Wafer scale test and ease of assembly for broadband and low threshold VCSELs has been an important driver. The low-power solutions scale well in 2-D arrays enabling Terabit class interconnection rates [31]. Integration of optical transceivers on a single CMOS chip has similarly enables multi-fibre cables, although so far just to 40 Gbit/s with operating powers of 20mW/Gbps [32]. Parallel optical outputs may however present packaging and assembly challenges and potentially constrain bandwidth density.

Serial multiplexing approaches include electronic time division multiplexing, optical time, code or wavelength multiplexing, optical polarisation diversity, phase- and amplitude-keying and the relative merits for many of these schemes remains an active areas of study [33]. Direct laser modulation continues to be of interest when optical losses cannot be tolerated. Recent reports of error free transmission over multimode fibre at data rates up to 32 Gbit/s at 25C and 25 Gbit/s at 85C using an oxide confined 850 nm VCSELs offer some promise [34],[35].

Energy efficiency is intrinsically linked to the confinement of charge carriers in the active region and non-radiative recombination of charge carriers. While Phosphorus containing alloys have provided the bedrock for telecommunications lasers, epitaxial engineering using Aluminium containing alloys enable high speed operation of lasers at very operating temperature [36],[37]. This being particularly important when the energy and packaging overheads for thermoelectric cooling cannot be tolerated. Quantum dot epitaxies further confine current through the limited number of energy states participating in laser oscillation, thus providing a route to high speed and high temperature [38]. However low optical overlap

can impede very high speed modulation and research into performance limits at elevated temperatures has so far been inconclusive. Despite this, ultracompact microring quantum dot lasers have been reported in the important 1550nm band [39]. The growth of long wavelength quantum dots on GaAs substrates is a promising avenue of research for low fabrication complexity long wavelength VCSELs [40].

Integration technology is enabling very high line rates per optical connections and enables the removal of energy overhead and packaging costs for planar optoelectronic circuits. The hybrid integration of de/multiplexers and uncooled transmitter/receiver arrays has allowed coarse wavelength multiplexed transceivers [41]. Planar waveguide circuits incorporating distributed feedback laser arrays, modulator arrays, and arrayed waveguide gratings have been pioneered by Infinera [42]. The incorporation of phase diversity modulation with wideband tunable lasers offers further scope for energy reduction and packaging simplification [43]. Microring laser arrays vertically coupled to a single passive bus waveguide have also highlighted a route to both low-power operation and monolithic wavelength division multiplexed sources [44] with 7-Gbit/s data modulation being demonstrated [45]. Membrane technologies have been further developed to circumvent the absence of electro-optic gain in Silicon circuits. Most recently, electrically pumped hybrid silicon-InGaAlAs microrings have been demonstrated with 5.4mA threshold currents and operation to 65C [46]. While further work is required to meet the performance levels achieved in monolithic circuits, the prospect of CMOS compliant optical transmitters is attractive. Parallel advances on CMOS compliant Mach-Zehnder modulators are also being made with submillimetre Silicon p+-i-n+ devices. The V L figure of merit is measured at 0.36V-mm and optical modulation rates of 10 Gbit/s have been achieved for 5pJ/bit radio frequency power consumption, leading to 6dB extinction ratio when used with electronic preemphasis [47]

Nanoscale miniaturisation has become a reality in the last two years through with metal-insulator-metal sub-wavelength plasmonic waveguides at wavelengths of 1500nm [48]. The electrically pumped semiconductor cores have widths below the diffraction limit. Oscillation in gap plasmon modes show that losses in sub-wavelength MIM waveguides can be overcome to create small plasmon mode lasers at wavelengths near 1500nm at room temperature.

As discussed in section 2.1.3, in optical interconnection systems for on-chip applications the typical configuration is with an external cw-laser source coupled with an on-chip modulator. Let us focus, then, here the modulation devices. The two main types of integrated modulators, suited for on-chip applications, are [14]: i) microresonator-based modulators [49], [50] and ii) Mach-Zehnder interferometer-based modulators [51].

The performance of the modulator depends on its extinction ratio (defined as the relation between the power measured in the on and off state). A high extinction ratio enables suitable signal detection which improves the system performance.

i) Microring resonator-based modulators

The work done in [49] demonstrates silicon resonator-based modulators that can switch 20 wavelengths simultaneously with a low-penalty active operation and broadband uniformity. Also [50] reports the practical implementation of these modulators, and foresees the possibility to be used as WDM modulators. In order to modulate light, a microring is placed so as to be coupled to a waveguide. The resonance condition (for those wavelengths such that the length of the ring circumference is equal to an integer number of wavelengths) the waveguide's light is coupled to the ring and it is scattered in the ring. The wavelength (or index of refraction) can be further fine-tuned electrically by injecting charge into the ring (by means of p-i-n junctions) or changing the temperature of the ring. This brings the ring in and

out of resonance. The first timing method seems more suitable for fast modulation (less than 1 ns) while the second can be used for slow tuning.

ii) MZ-based modulators

In [51] it is described an approach based on a metal–oxide–semiconductor (MOS) capacitor structure embedded in a silicon waveguide that can produce high-speed optical phase modulation. It is also demonstrated that this MZ optical modulator can exceed a modulation bandwidth of nearly 1GHz. When a Mach–Zehnder interferometer structure is considered, Δn_{eff} determines the length of the phase shifters. As this technology is compatible with conventional complementary MOS (CMOS) logic, monolithic integration of the silicon modulator with advanced electronics on a single silicon substrate becomes possible.

3.2 Detection

The receiver comprises the optoelectronic detector, transimpedance amplifier and voltage amplifier. At increasing line-rates, intimate connection between elements become critical in removing bandwidth limiting parasitics. III-V technologies continue to provide the bedrock for commercial telecom deployments but the attraction of a monolithic solution is driving research into Silicon- and CMOS-compliant detection even at long wavelengths. Current research highlights include

- Intimate Silicon integration
- Gain-bandwidth product enhancement
- Miniaturisation
- Functional enhancement through photonic integration

Silicon compliant detection at long wavelength has evolved in the last few years. Normal incidence p-i-n photodiodes have been realised for the monolithic integration. An enhanced efficiency in the near-infrared regime and the absorption edge shifting to longer wavelength is achieved with responsivities at 1.48, 1.525, and 1.55 μm are 0.8, 0.7, and 0.64 A/W, respectively [52].

Avalanche detection offers a route to increased sensitivity, but the fabrication of reliable high-gain, high-bandwidth devices has been restricted to a small number of groups. Epitaxially grown Si/Ge APDs have been demonstrated for the 1.3 μm wavelength by Intel [53]. The collaborating team at Santa Barbara has been able to experimentally and theoretically investigate the frequency response enhancement in Ge/Si APDs to demonstrate a gain bandwidth product of 845GHz for a 30 μm -diameter device [54]. Side-illuminated waveguide structures are anticipated to enable even higher gain-bandwidth products.

A route to nanoscale photodetection with sub-wavelength structures is currently under study. The first plasmonic device integrated in Si CMOS, a C-shaped nanoaperture photodetector, has been realised with a feature sizes of only 40 nm [55]. The technology has been proven in the blue wavelength regime showing polarisation dependence attributable to the test structures used, and with impressive quantum efficiency. Multiple electrode designs are biased using algorithms implemented by nanosecond reconfigurable logic to allow the synthesis of optical filter transfer functions at the detector [56].

Coherent receiver technologies required for emerging phase and polarisation keyed modulation and multiplexing formats have developed through increasingly complex integration technologies. Bell Labs have most recently realized a monolithic polarization-diversity dual-quadrature coherent receiver with on-chip balanced detection in a silicon photonic integrated circuit with integrated germanium detectors. We tested it using a polarization-division multiplexed quadrature phase-shift keyed signal at 112-Gbit/s [57].

The receiver is made with a photodetector that converts light into electrical signals and receiver circuitry that converts the analogical signal into a digital electrical signal. Metal-semiconductor-metal photodetector receivers already satisfy the requirements of on-chip global interconnections since they can be monolithically integrated, and they exhibit their fast response and roughly good quantum efficiency.

3.3 Amplification/regeneration

Regeneration in terms of retiming, reamplification, and reshaping of optical data sequences may be performed by a number of methods [59]. Studies into the relative merits of all-optical, optoelectronic and optical-electronic-optical conversion are ongoing. Power consumption considerations have so far ensured the dominance of OEO techniques. With increased photonic integration lowering the cost in optoelectronic conversion, some question the role of all-optical techniques [60]. None-the-less, data transparency becomes a real concern when seeking to future-proof for as yet unspecified formats and the massive linerates for tomorrow's technologies. Hot topics within the photonics research community have therefore included:

- Nonlinear photonics in nanostructures
- Ultrahigh speed and multisignal regeneration
- Photonic integration

Nonlinear processes have been pursued using nanostructured wiring. A four wave mixing conversion efficiency of -5.5 dB has been demonstrated in a low-loss 2.5 cm long sub-micron silicon-on-insulator rib waveguide [61]. Wavelength conversion, inherent in a number of optical regeneration schemes, has been shown at 40 Gbit/s in such silicon waveguides [62]. Such techniques enable conjugate phase outputs which may be exploited for include dispersion compensation [63]. While extensive multiband wavelength conversion ranges have been reported, power penalties have been poor at 2.9-4.7dB so far [64]. Pump powers of order 160mW are still required, enabling a power density in the waveguide's core of 430 MW cm², and thus work still needs to be done to improve energy efficiency. Internal conversion efficiencies of 10.6 dB have been obtained [65]. Routes to reducing the power have been identified with hybridly integration InGaAsP microdisc lasers [66]. Wavelength conversion is demonstrated with under 4mA of bias current, and estimate on-chip powers of 100uW and 10uW for the signal and pump powers respectively. While measurements are frustrated by the very low optical powers in early prototypes, error rates below 10⁻⁴ have now been shown at 2.5 Gbit/s.

Multichannel systems have been highlighted as a route to infrastructure reduction [67]. Field experiments for multichannel wavelength-path switching has been successfully demonstrated using polarization-independent multichannel wavelength conversion in quasi-phase-matched lithium niobate waveguide with an 8×8 matrix switch [68]. Terabit rate conversion has also been shown, although with a 10dB degradation in Q factor [69]. Proposed exploitation with integration compliant quantum dot amplifier systems also offers an interesting avenue for further study [70].

Integration of mode-locked lasers has provided a route to clock extraction from high line rate data. Monolithic techniques enabling 160 Gbit/s have been pioneered at Oki Electric [71]. More recent work has explored potential with hybrid-Silicon integrated mode locked lasers. Operation at 30.4 Gbit/s has enabled the extraction of a clock signal with 10dB extinction ratio and 1.7ps jitter derived from 3.8dB with 14ps jitter input data stream [72].

3.4 Switching

Optical switching technologies may be classified based on the underlying physical effect that is responsible for the switching process. The main optical switching technologies available today are: Electro-Optic (EO) Switching, Acousto-Optic (AO) Switching, Thermo-Optic (TO) Switching, Opto-Mechanical (OM) Switching. Each of these categories comprises a number of technology types depending on the manner in which the physical effect is exploited, device design, material used, and on other considerations:

Electro-Optic (EO) Switching

This category utilizes electro-optic effects. The most well-known switching technology in this category is based on optical waveguides and directional couplers which are implemented on Lithium Niobate substrates. In the EO effect, an electric field causes electron or crystal-lattice displacements, which result in refractive index change. Very fast response is expected in devices based on this effect. LiNbO₃ is a unique EO crystal in that large wafer substrates can be grown and can exhibit small dielectric constant, fast device responses, and low power consumption. Other ferroelectric materials that are used in optical switches include Lead Lanthanum Zirconate Titanate (PLZT). These materials exhibit electro-optic coefficients that are one order of magnitude larger than LiNbO₃. However, difficulty of fabricating low-loss waveguides hinders their use in optical switching. Recent improvements in fabrication methods can lead to lower waveguide loss [73]. Moreover fast electro-optical amplified switch with high contrast and large bandwidth based on Semiconductor Optical Amplifiers (SOA) have been developed. Finally, silicon microring resonator-based electro-optic switch array coupled with waveguide crossings can constitute a potential architecture for on-chip photonic crossbar interconnection. However, one key challenge in realizing such design in silicon is the considerable insertion loss and crosstalk imposed at the high-index-contrast submicrometer-sized waveguide intersections.

Acousto-Optic (AO) Switching

This guided-wave category exploits the acousto-optic effect and can also be implemented on Lithium Niobate. In this case, the refractive index of an optical medium is modulated by acoustic waves. An acoustic or elastic wave travelling in a medium induces periodical deformation in the form of alternating compressions and rarefactions. This results in a periodical strain in the medium. The vibrations of the molecules due to this strain alter the optical polarizability of the material and, consequently, its refractive index. An acoustic wave can be generated for this purpose by the piezoelectric effect either in the bulk of the material (leading to bulk acoustic waves) or on its surface (leading to surface acoustic wave, or SAW). This is done by applying a radio-frequency (RF) electric field. The electrodes across which the field is applied to induce the AO effect are called acoustic transducers. LiNbO₃ has been widely used to make acoustooptic waveguide devices since it is characterized by large acousto-optic effect and because low-loss waveguides can be fabricated on it using well established fabrication technologies. Furthermore, lithium niobate is a good piezoelectric material which means that RF energy can be efficiently converted into acoustic energy. Due to their capability to perform wavelength-selective optical switching, LiNbO₃ AO switches were thoroughly studied in the past decade for wavelength division multiplexing (WDM) applications and networks. Acousto-optic devices have been fabricated also using III-V semiconductors such as Gallium Arsenide (GaAs) and Indium Gallium Arsenide Phosphide (InGaAsP), where monolithic integration may be obtained easily. While these compound semiconductors are used to fabricate excellent lasers, photo-detectors and optical waveguides, AO devices made from them require higher driving power and yield much lower diffraction efficiencies than their LiNbO₃ counterparts. Furthermore since III-V semiconductors are not good piezoelectric materials, a thin piezoelectric film of zinc oxide (ZnO) must be deposited

between electrodes and waveguides to get efficient excitation of SAW's. Low loss and low driving power routing switches are also demonstrated by exploiting the acousto-optic effect in all fiber directional couplers.

Thermo-Optic (TO) Switching

This category utilizes the thermo-optic effect in optical waveguides. Two main types of materials are deployed to implement these switches, namely, silica and polymers. Early thermo-optic devices were fabricated using ion exchanged glass and Titanium diffused Lithium Niobate (Ti:LiNbO₃). However, since the TO effect in these materials was relatively weak, devices based on them were not efficient. Polymers became candidates to fabricate thermo-optic devices due to their large thermo-optic coefficients and low thermal conductivity. A lot of research was devoted to thermo-optic switches especially in polymers. Thermally-optimized silica based interferometric TO switches became also popular. Virtually, all thermo-optic switching principles and device concepts, including directional couplers, Mach-Zehnder interferometers (MZI) and digital switches, had been used before in EO switching and were adapted with slight modifications for TO switching. Thermo-optic waveguide switches have the advantages of relying on guided lightwaves (no free space optics) and not having moving parts. They are robust, enjoy long-term stability and their electrical driving circuitry can be very simple. Interaction of electro-magnetic lightwaves with the core and cladding of a dielectric waveguide is governed by the geometry of the waveguide and its refractive indices. Thermo-optic switching utilizes the temperature dependence of the refractive index dn/dT to realize switching functionality. Small thermally-induced changes of the refractive index can significantly alter the light-intensity distribution within waveguides, especially in coupled systems. By using micro heaters, temperature gradients can be induced within these waveguide structures leading to changes in the refractive index profile, which, in turn, is used to switch lightpaths.

Opto-Mechanical (OM) Switching

These devices are mostly based on free-space optics. It comprises classical technologies based on moving fiber, and/or moving macroscopic optical components as well as modern technologies such as those based on optical Micro Electro Mechanical Systems (MEMS). Using MEMS technology, optical elements and functions have been demonstrated, including mirrors, shutters, filters, lenses, and polarizers. They maintain the excellent optical performance characteristics of their macroscopic counterparts, but are much smaller, faster and can be fabricated using very large scale integration (VLSI) fabrication processes, similar to those of the electronics industry. Additionally, such optical MEMS can easily be batch-fabricated in arrays of hundreds or thousands of elements, which can be put in a single compact package. This makes them suitable for the demanding high-complexity optical networking applications, which need high levels of functionality per module. Integration is a key enabler of high complexity subsystems, allowing a new application to be addressed by providing novel functions, impractical to achieve otherwise. All micro mechanical optical switching devices may be roughly classified into two distinct classes. Devices in the first class achieve the switching function by manipulating the intensity or the direction of propagation of a light beam via reflective or refractive structures. In the second class the steering or switching function is achieved by diffractive or interference effects, where the mechanical motion adjusts the phase of light. Examples of the first class are: a mirror of a fixed orientation inserted into the path of a light beam to redirect all or part of the beam into a different output; a mirror permanently located in the path of the beam and can change orientation to steer the beam to the desired output. Some examples from the second class are: a tunable cavity formed by a membrane such as the mechanical antireflection switch (MARS); switches based on tunable gratings, which are revealed by moving a grating structure out of plane. The systems in the second class rely on the wave properties of the

electromagnetic field and are fundamentally wavelength dependent. Achieving high contrast switching relies on the precision of the phase control to cancel out the unwanted signal(s) via destructive interference. Consequently achieving high contrast over a wide wavelengths range can be difficult. However, one benefit of this kind of devices is that the small range of motion required, of the order of half a wavelength, which can allow for low actuation forces, or high modulation and switching rates. The first class has the advantage of wavelength independent operation over a broad spectrum. This can be achieved by appropriate mirror coatings. This approach is therefore much more suitable for building large scale, low loss, wavelength independent optical switching fabrics. The requirements in terms of size and motion of the switching elements will depend upon the type of optical system in which they are to be used. For a large, low-loss free-space switch fabric the beam is typically several hundreds of microns in diameter, dictating large mirror size and mirror stroke. Alternatively, for a waveguide based system, switching may be accomplished by inserting a mirror in a narrow gap between fixed guides, where the beam size is comparable to the mode size. In this case, strokes and mirror sizes of a couple of tens of microns may be sufficient. A final consideration is the number and type of degrees of freedom the device will have. Single axis motion of a rotational type is usually the simplest to implement since the motion in this case is around a fixed point and generally the support structure and the actuators can remain fixed. Pure translation is typically more challenging since the actuation and support may be required to move with the element. Motion about two axes adds significant complexity. Implementation of robust two-axis translation still poses significant challenges.

It must be noted that regardless of the physical effect responsible for the switching process, which is used as basis for this categorization, external control of the switching device is electrical in the vast majority of cases. Electrical control is used to trigger electro-optic effects in EO switching, to generate acoustic surface waves in AO switching, to supply heating energy in TO switching, to actuate OM switching systems, to control SOA gates, and so on.

Any utilization of optical switching technology should lead to an exhaustive analysis of the following performance parameters:

- *Insertion loss*: the insertion loss of optical switching technology is defined as the optical power loss when optical signals pass through the optical switch, consisting of coupling loss, waveguide propagation loss, and excess loss. When designing a network according to optical power budget, optical switches and their cascading impact network performance greatly. Furthermore, insertion loss limits the scalability of optical switches and increases system cost.
- *Switching speed*: the switching speed is defined as the time period from the moment the command is given to the switch to change state to the moment the insertion loss of the switched path achieves more than 90 percent of its final value. According to switching applications, the switching time is divided into three levels: multimillisecond order for protection application, nanosecond order for packet switching application, and picosecond order for bit-level optical time-division multiplexing (OTDM) application.
- *Crosstalk*: crosstalk is the ratio of the power leaked to the wrong output and the power correct output, used to measure the signal interference between channels. Low crosstalk and high extinction ratio (difference between two optical power levels) indicate small signal interference or high signal quality. Typically, the crosstalk value is around 40 or 50 dB.

- *Polarization-dependent loss*: polarization sensitivity is used to measure polarization dependence. When it is very high, it harms transmission reliability, and increases monitoring and dynamic compensation requirements.
- *Power supplied / power dissipated*: high power consumption increases a system's cost, and associated heat dissipation increases a system's ambient temperature.

Table 3 reports performance data of some optical switches.

Switch Effect	Switch type	Insertion Loss	Switching speed	Crosstalk	PDL	Power supplied
Electro-optic switch	LiNbO ₃ (1 x 2) [1]	4 dB	5 ns	45 dB		18 V
	LiNbO ₃ (2 x 2) [2]	20 dB	0.1 ns	55 dB		20 V
	PLZT (1 x 2 - 8 x 8) [3]	5 dB	20 ns	20 dB	0	10 V
	PLZT (1 x 2 - 1 x 8) [4]	10 dB	2.5 ns	20 dB	0	10 V
	SOA (1 x 2) [5]	0 dB	650 ps		<1 dB	200 mA
	SOA (2 x 2) [6]	1 dB	100 ps			15.6 fJ
	Ring resonator (4 x 4) [7]	Low	< 1 ns	Low		20 μ W
Acousto-optic switch	Acoustooptic (2 x 2) [8]	15 dB	155 ns	40 dB		200 mW
Thermal-optic switch	Thermaloptic (4 x 4) [9]	10 dB	4.6 μ s		0.8 dB	200 mW
	Ring resonator (8 x 4) [10]	4 dB	300 μ s	30 dB		6 W
Electro-Mechanical switch	MEMS (1 x 2) [11]		310 ns	?		22 V

Table 3 Optical switching technology comparison

References within Table 3:

- [1] R. Krhenbhl, M. M. Howerton, J. Dubinger, and A. S. Greenblatt, "Performance and modeling of advanced ti : Linbo3 digital optical switches", *IEEE Journal of Lightwave Technology*, Vol. 20, No. 1, Jan 2002, pp. 92–99.
- [2] A. Chiba, T. Kawanishi, T. Sakamoto, K. Higuma and M. Izutsu, "High extinction-ratio (> 55 dB) port selection by using a high-speed LiNbO₃ optical switch with intensity trimmers", *Photonics in Switching*, 2007, Aug. 2007, pp. 51–52.
- [3] K. Nashimoto, H. Moriyama, S. Nakamura, M. Watanabe, T. Morikawa, E. Osakabe, and K. Haga, "PLZT electro-optic waveguides and switches", *OFC 2001*.
- [4] H. Furukawa, N. Wada, N. Takezawa, K. Nashimoto, and T. Miyazaki, "640 (2 x 32 x 10) Gbit/s polarization-multiplexed, wide-colored optical packet switching achieved by polarization-independent high-speed plzt switch", *OFC 2008*, pp. 1–3.
- [5] C. M. Galleg and E. Conforti, "A fast electro-optical amplified switch using a resistive combiner for multi-pulse injection", *Microwave Symposium Digest 2006*, pp. 1935–1938.
- [6] A. Malacarne, J. Wang, Y. Zhang, A. Das Barman, G. Berrettini, L. Pot and A. Bogoni, "20 ps transition time all-optical SOA-based flip-flop used for photonic 10 Gbit/s switching operation without any bit loss", *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 14, No. 3, June 2008, pp. 808–815.
- [7] F. Xu and A. W. Poon, "Multimode-interference waveguide crossing coupled microring-resonator-based switch nodes for photonic networks-on-chip", *CLEO/QELS 2008*, May 2008, pp. 1–2.
- [8] J. Aboujeib, A. Prennou, V. Quintard, J. L. Bihan, "Evaluation of the crosstalk and losses in a multitransducer acousto-optic switch", *ICTTA 2008*, pp. 1–6.
- [9] Y. Li, J. Yu, and S. Chen, "Rearrangeable nonblocking soi waveguide thermooptic 4x4 switch matrix with low insertion loss and fast response", *IEEE Photonics Technology Letters*, Vol. 17, No. 8, Aug. 2005, pp. 1641–1643.
- [10] S.-J. Chang, C.-Y. Ni, Z. Wang and Y.-J. Chen, "A compact and low power consumption optical switch based on microrings", *IEEE Photonics Technology Letters*, Vol. 20, No. 12, June 2008 pp. 1021–1023.
- [11] G. N. Nielson, R. H. Olsson III, P. R. Resnick, O. B. Spahn, "High-speed mems micromirror switching", *CLEO 2007*, pp. 1–2.

The diversity in opto-electronic switching technology is extensive. Techniques exploit a bewildering variety of physical effects. The most recent research has been at the integrated circuit level for fast programmable switch interconnects. Areas of particular focus include:

- Wavelength conversion
- Fast programmable filters
- Monolithic integration

Wavelength conversion, summarised in the section on regeneration, becomes relevant to optical switching when augmented by a means to switch wavelength quickly in a broadly tunable laser without incurring thermal and electro-optic transients. A number of groups are now demonstrating nanosecond switching speeds, with varying degrees of device performance and complexity [74]-[78]. The first fully integrated 8x8 tunable router based on wavelength convertors has been realised by the group at Santa Barbara. They demonstrate the first InP monolithic tunable optical router with error-free 40 Gbps operation per port. The device has eight wavelength converters and an 8x8 arrayed-waveguide grating router, yielding more than 200 on-chip functional elements [79].

Nanosecond-timescale programmable filtering offers efficient reuse of fixed broadband infrastructure for bursty traffic environments. Integration-compliant filtering has been proposed with Bragg grating filters [80],[81], ring resonators [82]-[88] and tunable arrayed waveguide gratings [89],[90]. The Silicon CMOS integration environment provides an important level of process stability which has enabled research into microring based circuits. Researchers at IBM have devised filter networks with 1.8dB loss for 1dB flat top bandwidth of 310GHz and out of band rejection of 40dB [91]. Such approaches will be attractive for use in combination with a low power wavelength conversion technology or robust electro-optic tuning. The robust pass-band specification for arrayed waveguide gratings [92] already enables broadband multistage networks. Polarization and temperature insensitive design is routinely achieved and loss independent scaling to hundreds of connections is feasible in single stage circuits [93]. Research into wavelength blockers and gated cyclic routers may therefore offer a route to direct reconfigurability without stringent fabrication tolerances [94],[95].

Electro-optic switching between arbitrary combinations of waveguides has lead to renewed interest into amplifier based switching [96]. Semiconductor optical amplifier based photonic achieve lossless switching of high capacity data with nanosecond switching timescales [97]. Such switches have recently undergone considerable development, with integrated 2x2 [98], 4x4 [99] and 8x8 [100] port switches being reported, the latter being achieved with multiple integrated 1x8 switching elements. Scaling further has required unprecedented levels of component integration [101]. However, for practical applications, devices with port counts of at least 16x16 are needed for systems applications [102]. Recently, feasibility studies have shown that 16x16 port counts can be achieved using cascaded 4x4 devices with fibre interconnections [103]. This lead to the first demonstration of lossless routing with a minimum power penalty of 2.5dB at 10Gbit/s in an integrated 16x16 fabric [104]. Larger scale fabrics have also been predicted from simulations [105]. Channel line rates have increased markedly in the last year. Serial 160Gbit/s data transmission through a semiconductor optical amplifier based multistage switching matrix has been achieved in monolithically cascaded switching networks. Bit error rate studies are performed to show only modest levels of signal degradation. Power penalties of order 0.6dB and 1.2dB are observed for two stages and four stages respectively in the monolithic circuits at 160Gbit/s per path [106].

4. Power Consumption Comparison

Power consumption per rack of equipment of high-capacity network elements has already reached more than 10 kW, which represents a serious problem for a further increase in capacity. Due to the high power consumption and the need for more aggregate capacity, single-rack systems have been replaced through multi-rack systems where the system power is uniformly distributed over multiple racks. However, this extension implies larger system with more interconnections. Due to high signal loss and intersymbol interference (ISI) in copper cable interconnects at high data rates and large distances, optical point-to-point interconnects with low signal attenuation and crosstalk are used for rack-to-rack interconnections in order to achieve higher overall throughput and better transmission conditions with longer reaches. These interconnects are E/O/E devices and comprise optical transmitter, optical fiber and optical receiver. Currently deployed optical interconnects (OIs) provide channel data rates of up to 10 Gbit/s and have reaches of up to 300 m. A further increase in data rate per channel is, however, limited due to the transmission impairments in optical fiber such as chromatic dispersion and polarization mode dispersion. Thus, maximum data rate per channel and transmission distance are both limiting factors when an extension to larger systems, and thus greater aggregate capacities, is needed. The further limiting factor is, however, power consumption of the internal interconnects, since they dissipate considerable amount of power and occupy significant physical area.

In case of optical interconnects, the most critical parameters that influence power consumption are detector capacitance, coupling efficiency and the type of modulator used while for electrical intermodule interconnects, besides above mentioned limits, there are also receiver sensitivity/offset and impedance mismatch [107].

The IEEE 802.3ba Ethernet Task Force has been working on further transmission rate enhancement with development of the 40 Gigabit Ethernet (40GbE) and 100 Gigabit Ethernet (100GbE) options. The future standard envisages the use of either four or ten parallel lines at 10Gbit/s each over up to 100 m of multi mode fiber for 40GbE (40GBASE-SR4) and 100GbE (100GBASE-SR10), respectively. Further considered options for the implementation are wavelength-division-multiplexed channels.

For the purpose of estimation of the overall system scalability in terms of power consumption, we developed a multi-shelf model for calculating total power consumption for different aggregate capacities. This model considers two interconnection arrangements, namely a centralized one and a partitioned one. A single-shelf implementation, which is typical for backplane applications, was also considered and includes the two above mentioned interconnection arrangements. Our analysis was based upon determining total number of transceivers required for different aggregate capacities according to allocated power per transceiver.

In order to determinate realistic values for power consumption of electrical (backplane and copper cable transceivers) and optical modules, a database of large numbers of currently (or in the near future) available transceivers from different manufacturers was created with specifications and product briefs for all interconnecting technologies aimed for single-shelf and multi-shelf applications. Power consumption per transceiver block for one specific data rate and one specific target application was simply calculated as a mean value of all such components consisting of the same implemented blocks, but from different vendors. Thus, realistic values for power consumption were obtained.

4.1 Single-Shelf System

Figure 17 gives an overview on power consumption comparison of two considered architectures for 10G XAUI backplane transceivers. The centralized architecture achieves better performance. For a total capacity of 640 Gbit/s, the power consumption of the centralized architecture is about 220 W, which is two times lower than that of the distributed architecture assuming the same capacity.

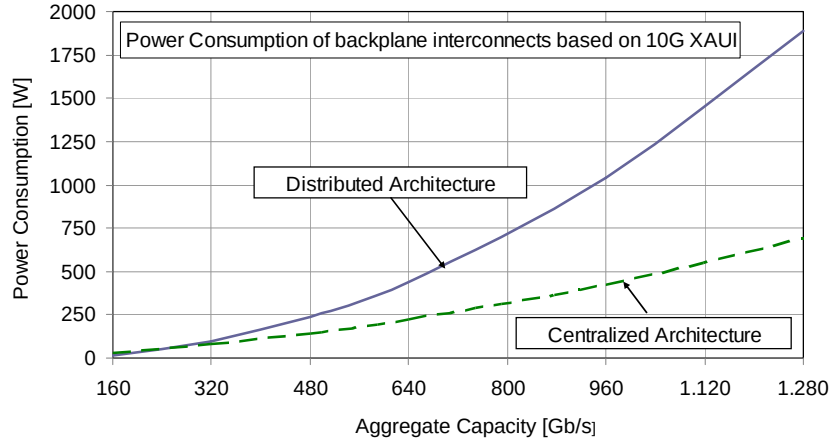


Figure 17 Power consumption of XAUI backplane transceivers

4.2 Multi-Shelf System

Figure 18 shows the results for power consumption of optical modules used for multi-shelf interconnection of systems with aggregate capacity of 100 Tbit/s. The 40GbE and 100GbE options, namely 40GBASE-SR4 and 100GBASE-SR10, were also considered. It can be seen that 12xQDR IB consumes the lowest power, while the most consuming technology option is 10G XAUI. 40G and 100G WDM interconnects also consume considerable amount of power, but this could rely on the fact that the components we found for WDM interconnects target 10 km applications. Even if very low-power interconnects would become broadly available, the management of large number of interconnecting cables remains very difficult [108]. When using WDM technology, the number of fibers needed to transmit 100 Gbit/s payload data rate can be reduced from 10 to 1. Hence, WDM interconnects provide an additional benefit of reducing the total number of fibers, which consequently make the whole system simpler and cheaper.

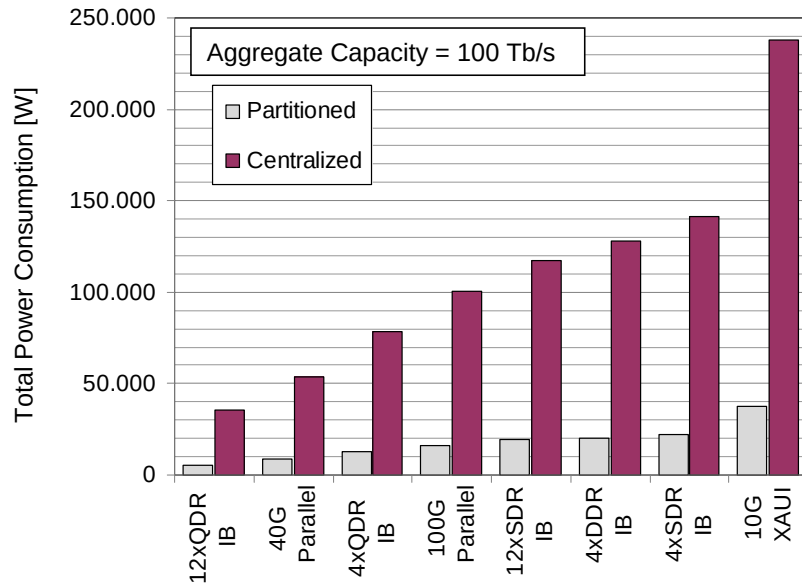


Figure 18 Power consumption of different interconnecting technologies

5. References

- [1] H. Johnson and M. Graham, *High-speed signal propagation*, Prentice Hall, February 2003.
- [2] D. T. Neilson, "Photonics for switching and routing", *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 12, No. 4, July-Aug. 2006, pp. 669–678.
- [3] B. M. Oliver, "Thermal and quantum noise", *Optics in Computing*, Vol. 53, No. 5, May 1965, pp. 436–454.
- [4] R. T. Chen, L. Lei, C. Chulchae, Y. J. Liu, B. Bihari, L. Wu, S. Tang, R. Wickman, B. Picor, M. K. Hibb-Brenner, J. Bristow, and Y. S. Liu, "Fully embedded board-level guided-wave optoelectronic interconnects", *Proc. of IEEE*, Vol. 88, June 2000, pp. 780–793.
- [5] X. Zheng, P. Marchand, D. Huang, and S. Esener, "Free-space parallel multichip interconnection system", *Appl. Opt.*, vol. 39, July 2000, pp. 3516–3524.
- [6] P. S. Guilfoyle, J. M. Hessenbruch, and R. V. Stone, "Free-space optical interconnects for high performance optoelectronic switching", *IEEE Trans. Comput.*, Vol. 31, Feb. 1998, pp. 69–75.
- [7] D. W. Prather, S. Venkataraman, M. Lecompte, F. Kiamilev, J. N. Mait, and G. J. Simonis, "Optoelectronic multichip module integration for chip level optical interconnects", *IEEE Photonics Technology Letters*, vol. 13, Oct. 2001, pp. 1112–1114.
- [8] S. C. Esener, E. Yuceturk, D. Huang, and T. C. Sze, "Comparative study of very short distance electrical and optical interconnects based on channel characteristics", *Optics in Computing*, Vol. 90, 2003, pp. 7–9.
- [9] H. Cho, P. Kapur, and K. C. Saraswat, "Power comparison between high-speed electrical and optical interconnects for interchip communication", *IEEE Journal of Lightwave Technology*, Vol. 22, No. 9, Sep. 2004, pp. 2021–2033.
- [10] "Data sheet MT 125C", www.micro-coax.com.
- [11] "Data sheet UFA 125A", www.micro-coax.com.
- [12] C. Gunn, "CMOS photonics for high speed interconnects", *IEEE Micro*, Vol. 26, No. 2, March/April 2006, pp. 58–66.

- [13] N. Kirman, M. Kirman, R. K. Dokania, J. F. Martínez, A.B. Apsel, M. A. Watkins, D. H. Albonesi, "On-chip optical technology in future bus-based multicore designs", *IEEE Micro*, Vol. 27, No. 1, Jan.-Feb. 2007, pp.56–66.
- [14] M. Haurylau, H. Chen, J. Zhang, G. Chen, N. A. Nelson, D. H. Albonesi, E. G. Friedman, P. M. Fauchet, "On-chip Optical interconnect roadmap: challenges and critical directions", *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 12, No. 6, part 2, Dec. 2006, pp. 1699-1705.
- [15] Ian O'Connor, "Optical solutions for system-level interconnect", *2004 International workshop on System level interconnect prediction*, ACM Press, Newport Beach, California, USA, Dec. 2004, pp.79–88.
- [16] P. Rabiei, W. H. Steier, C. Zhang, L. R. Dalton, "Polymer Micro-Ring Filters and Modulators" *IEEE Journal of Lightwave Technology*, Vol. 20, No. 11, Nov. 2002, pp. 1968-1975.
- [17] J. Cardenas, C.B. Poitras, J.T. Robinson, K. Preston, L. Chen, M. Lipson, "Low Loss Etchless Silicon Photonic Waveguides", *Conference on Quantum electronics and Laser Science Conference, CLEO/QELS 2009*, Munich, Germany, June 2009, pp.1 – 2.
- [18] N. Kirman, M. Kirman, R. K. Dokania, J. F. Martínez, A.B. Apsel, M. A. Watkins, D. H. Albonesi, "Leveraging Optical Technology in Future Bus-Based Chip Multiprocessors", *39th Annual IEEE/ACM International Symposium on Microarchitecture*, Washington, DC, USA, Dec. 2006, pp. 492-503,
- [19] R. Ho, K. W. Mai, M. A. Horowitz, "The future of wires", *Proc. of IEEE*, Vol. 89, No. 4, New York, USA, Apr. 2001, pp. 490–504.
- [20] B. S. Cherkauer, E. G. Friedman, "A unified design methodology for CMOS tapered buffers" *IEEE Trans. Very Large Scale (VLSI) Integr. Syst.*, Vol. 3, No. 1, Mar. 1995, pp. 99–111.
- [21] The ITRS Technology Working Groups, <http://public.itrs.net>. International Technology Roadmap for Semiconductors (ITRS).
- [22] A. L. R. L. Dawei Huang, T. Sze and H. Davidson, "Optical interconnects: out of the box forever?", *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 9, No. 2, Apr. 2003, pp. 614–623.
- [23] C. Morgan and D. Helster, "New printed-wiring-board materials guard against garbled gigabits", www.ednmag.com, Nov. 1999.
- [24] PHYSorg.com, "Bell labs and fci demonstrate 25gb/s data transmission over electrical backplane connectors", www.physorg.com/news2899.html, Feb. 2005.
- [25] B. L. Booth , J. E. Marchegiano , C. T. Chang, R. Furmanak , D. M. Graham, R. G. Wagner, "Polyguide polymeric technology for optical interconnect circuits and components", DuPont Company, 1997.
- [26] D. M. Kuchta, Y. H. Kwark, C. Schuster, C. Baks, C. Haymes, J. Schaub, P. Pepeljugoski, L. Shan, R. John, D. Kucharski, D. Rogers, M. Ritter, J. Jewell, L. A. Graham, K. Schrödinger, A. Schild, and H. Rein, "120-gb/s vcsel-based parallel-optical interconnect and custom 120-gb/s testing station", *IEEE Journal of Lightwave Technology*, Vol. 22, No. 9, Sep. 2004, pp. 2200–2212.
- [27] J. Beals IV, N. Bamiedakis, A. Wonfor, R.V. Penty, I.H. White, J.V. DeGroot Jr., K. Hueston, T.V. Clapp, M. Glick, "A terabit capacity passive polymer optical backplane based on a novel meshed waveguide architecture", *Applied Physics A: Materials Science and Processing*, Feb. 2009.
- [28] D. Kuchta, Y. Kwark, C. Schuster, C. Baks, C. Haymes, J. Schaub, P. Pepeljugoski, J. Jewell, L. Graham, D. Kucharski, "120 gb/s vcsel based parallel optical transmitter and custom 120 gb/s testing station", *Lasers and Electro-Optics Society. LEOS 2003*, vol. 1, Oct. 2003.

- [29] A. F. Benner, M. Ignatowski, J. A. Kash, D. M. Kuchta, M. B. Ritter, "Exploitation of optical interconnects in future server architectures", *IBM Journal of Research and Development*, July-Sep. 2005.
- [30] A.V. Krishnamoorthy, L.M.F. Chirovsky, W.S. Hobson, R.E. Leibengath, S.P. Hui, G.J. Zydzik, K.W. Goossen, J.D. Wynn, B.J. Tseng, J. Lopata, J.A. Walker, J.E. Cunningham, L.A. D'Asaro, "Vertical-cavity surface-emitting lasers flip-chip bonded to gigabit-per-second cmos circuits", *IEEE Photonics Technology Letters*, Vol. 11, No. 1, Jan. 1999, pp. 128–130.
- [31] L. Schares et al., "Terabus: Terabit/second class card-level optical interconnect technologies", *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 12, No. 5, Sep.-Oct. 2006, pp. 1032-1044.
- [32] www.luxtera.com.
- [33] P. J. Winzer, "Modulation and multiplexing in optical communications systems", *IEEE LEOS Newsletter*, 4-10, Feb. 2009
- [34] A. Mutig et al., "Temperature-dependent small-signal analysis of high-speed high-temperature stable 980-nm VCSELs", *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 15, No. 3, 2009, pp. 679 – 686.
- [35] P. Westbergh et al., "32 Gbit/s multimode fibre transmission using high-speed, low current density 850 nm VCSEL", *Electronics Letters*, Vol. 45, No. 7, 2009, pp. 366 – 368.
- [36] R. Paoletti et al., "Highly reliable and high yield 1300nm InGaAlAs directly modulated ridge Fabry-Perot lasers operating with 10-Gb/s up to 110 C with constant current swing", *IEEE Journal of Lightwave Technology*, Vol. 24, No. 1, Jan. 2006, pp. 143- 149.
- [37] T. Nakamura et al., "1.3- μ m AlGaInAs strain compensated MQW-buried-heterostructure lasers for uncooled 10-Gb/s operation", *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 11, No. 1, 2005, pp. 141-148.
- [38] K.T. Tan et al., "High bit rate and elevated temperature data transmission using InGaAs quantum-dot lasers", *IEEE Photonics Technology Letters*, Vol. 16, No. 5, May 2004, pp. 1415-1417.
- [39] M.T. Hill et al., "InAs-InP (1.55- μ m region) quantum-dot microring lasers", *IEEE Photonics Technology Letters*, Vol. 20, No. 6, 2008, pp. 446-448.
- [40] R. Suzuki, "Multilayer 1.4 μ m InAs quantum dots with thin spacer using GaNAs strain compensation layer", *20th International Conference on Indium Phosphide and Related Materials*, 2008.
- [41] L. A. Buckman et al., "Demonstration of a small-form-factor WDM transceiver module for 10-Gb/s local area networks", *IEEE Photonics Technology Letters*, Vol. 14, No. 5, 2002, pp. 702–705.
- [42] D. F. Welch et al., "The realization of large-scale photonic integrated circuits and the associated impact on fiber-optic communication systems", *IEEE Journal of Lightwave Technology*, Vol. 24, No. 12, 2006, pp. 4674–4683.
- [43] M.J. Wale, "Technology options for future WDM-PON access systems", *OECC 2009*.
- [44] A. Bennecer et al., "Directly modulated wavelength-multiplexed integrated microring laser array", *IEEE Photonics Technology Letters*, Vol. 20, No. 16, 2008, pp. 1411-1413.
- [45] A. Kapsalis et al., "7 Gb/s direct modulation of vertically coupled microring lasers", *OFC/NFOEC 2008*, San Jose, CA, Feb. 2008.
- [46] D. Liang et al., "Electrically-pumped compact hybrid silicon microring lasers for optical interconnects", *Optics Express*, Vol. 17, No. 22, 2009, pp. 20355-20364.
- [47] W.M.J. Green et al., "Ultra-compact low RF power 10Gb/s Silicon Mach-Zehnder modulator", *Optics Express*, Vol. 15, No. 25, 2007, pp. 17106.
- [48] M.T. Hill et al., "Lasing in metal-insulator-metal sub-wavelength plasmonic waveguides", *Optics Express*, Vol. 17, No. 13, 2009, pp. 11107.

- [49] B. G. Lee, A. Biberman, P. Dong, M. Lipson, K. Bergman, "All-Optical Comb Switch for Multiwavelength Message Routing in Silicon Photonic Networks", *IEEE Photonics Technology Letters*, Vol. 20, No. 10, 15 May 2008, pp. 767-769.
- [50] Q. Xu, B. Schmidt, S. Pradhan, M. Lipson, "Micrometre-scale silicon electro-optic modulator", *Nature* 435, May 2005, pp. 325-327.
- [51] A. Liu, R. Jones, L. Liao, D. Samara-Rubio, D. Rubin, O. Cohen, R. Nicolaescu, M. Paniccia, "A high-speed silicon optical modulator based on a metal-oxide-semiconductor capacitor", *Nature* 427, Feb. 2004, pp. 615-618.
- [52] H. Yu et al., "High-efficiency p-i-n photodetectors on selective-area-grown Ge for monolithic integration", *IEEE Electron Device Letters*, Vol. 30, No. 11, 2009, pp. 1161.
- [53] Y. Kang et al., "Epitaxially-grown Ge/Si avalanche photodiodes for 1.3 μ m light detection", *Optics Express*, Vol. 16, No. 13, 2008, pp. 9365.
- [54] W.S. Zaoui et al., "Frequency response and bandwidth enhancement in Ge/Si avalanche photodiodes with over 840GHz gain-bandwidth-product", *Optics Express*, Vol. 17, No. 15, 2009, 12641-12649.
- [55] L. Tang et al., "Plasmonic device in silicon CMOS", L. Tang, S. Latif and D.A.B. Millier, *Electronics Letters*, Vol. 45, No. 13, 2009, pp. 706-708.
- [56] R. Chen et al., "Design and analysis of CMOS-controlled tunable photodetectors for multiwavelength discrimination", *IEEE Journal of Lightwave Technology*, Vol. 27, No. 23, 2009, pp. 5451-5460.
- [57] C.R. Doerr et al., "Monolithic Silicon coherent receiver", *OFC 2009*, post-deadline paper PDPB2.
- [58] J. Koester, C.L. Schow, L. Schares, G. Dehlinger, J.D. Schaub, F.E. Doany, R.A. John, "Ge-on-SOI-Detector/Si-CMOS-Amplifier Receivers for High-Performance Optical-Communication Applications", *IEEE Journal of Lightwave Technology*, Vol. 25, No. 1, Jan 2007, pp. 46-57.
- [59] O. Leclerc et al, "Optical regeneration at 40Gb/s and beyond", *IEEE Journal of Lightwave Technology*, Vol. 21, 2003, pp. 2779-2790.
- [60] D.F. Welch, "The advantages of PIC-based digital optical networks", *Photonics in Switching 2007*, paper WB1.3.
- [61] H. Rong et al., "Efficient wavelength conversion via four-wave mixing in sub-micron silicon rib waveguides", *IEEE/LEOS Winter Topicals Meeting Series*, 2009, paper MA4.1.
- [62] Y.H. Kuo et al., "Demonstration of wavelength conversion at 40 Gb/s data rate in silicon waveguides", *Optics Express*, Vol. 14, 2006, pp. 11721-11726.
- [63] S. Ayotte et al., "Dispersion compensation by optical phase conjugation in silicon waveguide", *Electronics Letters*, Vol. 43, 2007, pp. 1037-1039.
- [64] B.G. Lee et al., "Demonstration of broadband wavelength conversion at 40 Gb/s in Silicon waveguides", *IEEE Photonics Technology Letters*, Vol. 21, No. 3, 2009, pp. 182-184.
- [65] K. Yamada et al., "All-optical efficient wavelength conversion using Silicon photonic wire waveguide", *IEEE Photonics Technology Letters*, Vol. 18, No. 9, 2006, pp. 1046-1048.
- [66] O. Raz et al., "High speed wavelength conversion in a heterogeneously integrated disc laser over Silicon on insulator for network on a chip applications", *ECOC 2009*.
- [67] C. Okonkwo, "Performance analysis of an optical packet switch with shared parametric wavelength converters", *IEEE Communications Letters*, Vol. 12, No. 8, 2008, pp. 596-598.
- [68] J. Yamawaku et al., "Field demonstration of up to 640 Gb/s (64 ch $\times$ 10 Gb/s) GWP switching networks with a QPM-LN wavelength converter in JGN II test bed", *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 12, No. 4, 2006, pp. 529-535.
- [69] J. Yamawaku et al., "Low-crosstalk 103 channel $\times$ 10 Gb/s (1.03 Tb/s) wavelength conversion with a quasi-phase-matched LiNbO₃ waveguide", *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 12, No. 4, 2006, pp. 521-528.

- [70] D. Nielsen, "High-speed wavelength conversion in quantum dot and quantum well semiconductor optical amplifiers", *Applied Physics Letters*, Vol. 92, No. 21, 2008, id. 21110.
- [71] S. Arahira et al., "Reshaping and retiming function of all-optical clock extraction at 160Gb/s in a monolithic mode-locked laser diode", *Journal of Quantum Electronics*, Vol. 41, No. 7, July 2005, pp. 937-944.
- [72] B.R. Koch et al., "All-optical clock recovery with retiming and reshaping using a Silicon evanescent mode-locked ring laser", *OFC 2008*.
- [73] K. Nashimoto, H. Moriyama, S. Nakamura, M. Watanabe, T. Morikawa, E. Osakabe, and K. Haga, "PLZT electro-optic waveguide and switches", *Optical Fiber Communication Conference 2001*, Vol. 4, Paper PD10.
- [74] M.J.R. Heck et al., "Monolithic AWG-based discretely tunable laser diode with nanosecond switching speed", *IEEE Photonics Technology Letters*, Vol. 21, No. 13, 2009, pp. 905-907.
- [75] M. Kauer et al., "16-channel digitally tunable external-cavity laser with nanosecond switching time", *IEEE Photonics Technology Letters*, Vol. 15, No. 3, Mar. 2003, pp. 371-373.
- [76] B. Puttnam, "Nanosecond tuning of a DS-DBR laser for dynamic optical networks", *CLEO Europe*, 2009.
- [77] B. Roycroft et al., "Fast switching tunable laser sources for wavelength division multiplexing in passive optical access networks", *IEEE 19th International Conference on Indium Phosphide & Related Materials*, 2007.
- [78] O.A. Lavrova et al., "10-Gb/s agile wavelength conversion with nanosecond tuning times using a multisection widely tunable laser", *IEEE Journal of Lightwave Technology*, Vol. 20, No. 4, Apr. 2002, pp. 712-717.
- [79] S.C. Nicholes et al., "The world's first InP 8x8 monolithic tunable optical router (MOTOR) operating at 40 Gbps line rate per port", *OFC 2009*, post-deadline paper PDPB1.
- [80] R.C. Alferness, et al., "Grating-assisted InGaAsP/InP vertical co-directional coupler filter", *Applied Physics Letters*, Vol. 55, No. 19, 1989, pp. 2011-2013.
- [81] J.P. Weber et al., "An integratable polarization-independent tunable filter for WDM systems: the multigrating filter", *IEEE Journal of Lightwave Technology*, Vol. 14, No. 12, 1996, pp. 2719-2735.
- [82] S. J. Choi et al., "An eight-channel demultiplexing switch array using vertically coupled active semiconductor microdisk resonators", *IEEE Photonics Technology Letters*, Vol. 16, No. 11, Nov. 2004, pp. 2517-2519.
- [83] A.W. Poon et al., "Cascaded microresonator-based matrix switch for Silicon on-chip optical interconnection", *Proceedings of the IEEE*, Vol. 97, No. 7, 2009, pp. 1216-1238.
- [84] N. Sherwood-Droz, et al., "Optical 4x4 hitless Silicon router for optical Networks-on-Chip (NoC)", *Optics Express*, Vol. 16, No. 20, 2008, 15915-15922.
- [85] B.G. Lee et al., "All-optical comb switch for multi-wavelength message routing in Silicon photonic networks", *IEEE Photonics Technology Letters*, Vol. 20, No. 10, 2008, pp. 767-769.
- [86] A. Bennecer et al., "Ultracompact microring laser-based optical-add multiplexer", *IET Electronics Letters*, Vol. 44, No. 9, 2008, pp. 593-594.
- [87] D. G. Rabus et al., "Optical filters based on ring resonators with integrated semiconductor optical amplifiers in GaInAsP-InP", *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 8, No. 6, 2002, pp. 1405-1411.
- [88] B. E. Little, S. T. Chu, P. P. Absil, J. V. Hryniewicz, F. G. Johnson, F. Seiferth, D. Gill, V. Van, O. King, and M. Trakalo, "Very high-order microring resonator filters for WDM applications", *IEEE Photonics Technology Letters*, Vol. 16, No. 10, 2004, pp. 2263-2265.

- [89] T. Tanemura, "InP/InGaAsP integrated 1x5 optical switch using arrayed phase shifters", *IEEE Photonics Technology Letters*, Vol. 20, No. 12, 2008, pp. 1063-1065.
- [90] I.M. Soganci et al., "Integrated 1x8 optical phased-array Switch with low polarisation sensitivity for broadband optical packet switching", *IEEE Photonics Journal*, Vol. 1, No. 2, 2009, pp. 80-87.
- [91] F. Xia et al., "Ultra-compact high order ring resonator filters using submicron photonic wires for on-chip optical interconnects", *Optics Express*, Vol. 15, No. 19, 2007, pp. 11934.
- [92] N. Yurt et al., "Design and fabrication of a broadband polarization and temperature insensitive arrayed waveguide grating on InP", *Optics Express*, Vol. 13, 2005, pp. 5535-5541.
- [93] S. Araki et al., "Highly scalable optoelectronic packet-switching fabric based on wavelength-division and space-division optical switch architecture for use in the photonic core node [Invited]", *Journal of Optical Networking*, Vol. 2, 2003, pp. 213-228.
- [94] J. Gripp et al., "Optical switch fabrics for ultra-high-capacity IP routers", *IEEE Journal of Lightwave Technology*, Vol. 21, No. 11, 2003, pp. 2839-2850.
- [95] A. Rohit et al., "Monolithic multiband wavelength router for fast reconfigurable data networking", *Photonics in Switching*, 2009.
- [96] O. Liboiron-Ladouceur et al., "The data vortex optical packet switched interconnection network", *IEEE Journal of Lightwave Technology*, Vol. 26, No. 13, 2008, pp. 1777-1789.
- [97] R. Luijters and R. Grzybowski, "The OSMOSIS optical packet switch for supercomputers", *Optical Fiber Communications Conference 2009*, San Diego, paper OtuF3.
- [98] E.T. Aw et al. CLEO'08 (2008).
- [99] H. Wang. et al., "Temperature independent optical amplification in uncooled quantum dot optical amplifiers", *Optical Fiber Communications conference*, 2008.
- [100] Y. Kai et al. Proc ECOC '08 (2008).
- [101] Nagarajan, R. and Smit, M. "Photonic integration", *IEEE LEOS News*, Vol. 21, No. 3, June 2007, pp. 4-10.
- [102] L. Dittmann et al., "The European IST project DAVID: a viable approach toward optical packet switching", *IEEE Journal of Selected Areas in Communications*, Vol. 21, No. 7, 2003, pp. 1026-1040.
- [103] H. Wang et al., "Lossless multistage SOA switch fabrics using high capacity monolithic 4x4 SOA circuits", *Optical Fiber Communications Conference 2009*, paper OWQ2.
- [104] H. Wang et al., "Demonstration of a lossless monolithic 16x16 QW SOA switch", *ECOC 2009*, postdeadline paper.
- [105] I.H. White et al., "Scalable optical switches for computing applications", *OSA Journal of Optical Networking*, Vol. 8, No. 2, 2009, pp. 215-224.
- [106] A. Albores-Mejia et al., "160Gb/s Serial Line Rates in a Monolithic Optoelectronic Multistage Interconnection Network", *17th IEEE Symposium on High Performance Interconnects*, New York, 2009.
- [107] Hyeol Cho, "Power Comparison Between High-Speed Electrical and Optical Interconnects for Interchip Communication", *IEEE Journal of Lightwave Technology*, Vol. 22, No. 9, Sep. 2004, pp. 2021- 2033.
- [108] Rodney S. Tucker, "The Role of Optic and Electronic in High-Capacity Routers", *IEEE Journal of Lightwave Technology*, Vol. 24, No. 12, Dec. 2006, pp. 4655-4673.



JA3: Survey on optical interconnection architectures and solutions

Optical interconnection in high-speed switching and processing systems

Document attached to D25.2 of BONE WP25

Abstract—Large-capacity interconnection systems only based on electronic technology suffer technical bottlenecks that bound their scalability. Optical interconnection has proved to be a valid alternative solution in many applications, ranging from circuit switching to packet switching, to super-computing, to systems embedded on-chip. In this survey we analyze the main applications of optical interconnection. First we describe the main type of system architectures classified according to the application, analyzing their most relevant requirements of interconnection. Then, for each application class, we present the main architectures that have been proposed so far in literature and prototyped in labs or, in some cases, developed as products available on the market. Our overview aims at testifying that optical interconnection is not only a subject for theoretical studies; it is instead an important resource to increase the capability of actual high-speed systems in telecommunications and computing.

Index Terms – Switching systems, processing systems, system architecture, optical add-drop multiplexer, optical cross connect, optical packet switch, super-computer, on-chip network, optical interconnects, optical backplane

1. INTRODUCTION	2
2. FUNCTIONAL ARCHITECTURES.....	3
2.1 ADD-DROP MULTIPLEXERS AND CROSS-CONNECTS	3
2.2 FRAME/PACKET SWITCHES.....	5
2.3 SUPERCOMPUTING MACHINES.....	8
2.4 ON-CHIP NETWORKS	9
3. OPTICAL INTERCONNECT ARCHITECTURES.....	9
3.1 ADD-DROP MULTIPLEXERS.....	9
3.1.1 ROADMs based on wavelength blocker systems.....	10
3.1.2 ROADM systems based on PLC.....	12
3.1.3 The ROADM systems based on WSS.....	13
3.2 OPTICAL CROSS-CONNECTS.....	17
3.3 FRAME/PACKET SWITCHES.....	23
3.3.1 HYPASS (Bell Communications Research).....	24
3.3.2 STAR-TRACK (Bell Communications Research).....	25
3.3.3 BNR (Bell North Research).....	26
3.3.4 WAVE-MUX SWITCH (Oki Electric Industry Corporation).....	27
3.3.5 VLSI PRM (Washington University).....	28
3.3.6 BROADCAST-AND-SELECT OPTICAL INTERCONNECT (e-photon/ONe)	29
3.4 SUPERCOMPUTING MACHINES.....	31
3.4.1 INFINIBAND (Infiniband Trade association).....	32
3.4.2 MYRINET (Myricom)	32
3.4.3 RAPID (University of Arizona)	33
3.4.4 DATA VORTEX (IBM Research)	34
3.4.5 OSMOSIS (IBM Research).....	35
3.5 ON-CHIP ARCHITECTURES IN MULTICORE APPLICATIONS	36
3.5.1 Manycore (Massachusetts Institute of Technology, University of Carolina).....	37
3.5.2 CORONA (University of Wisconsin, Hewlett Packard, University of Utah).....	38
3.5.3 Photonic NoC (Columbia University).....	39
3.5.4 HyperMesh (Shanghai Jiao Tong University).....	40
3.5.5 Summary of on-chip architectures	42
4. REFERENCES	43

1. Introduction

Modern information processing systems are composed of several subsystems physically implemented as separate modules which need to communicate each other in order to carry out their duties. Let us think very simply to our lap-top computer: there are at least half a dozen devices, namely CPU, RAM, hard-disk, keyboard, screen, Ethernet card, which exchange data through the mother-board.

In a high-performance large system, being it a core IP router, a large network switch or a super-computer, the number of hardware modules to be interconnected increases enormously and the interconnection data-speed scales-up as well. Not only size and bandwidth increase, but also interconnection topology becomes more complex, passing from the simple linear parallel bus of the pc to rings, grids, full-mesh or complicated multistage architectures. Designing the interconnections inside a system becomes a task *per-se* complicated almost as much as designing the main modules.

The vast majority of systems on the market today adopt electrical wiring and electronics as interconnection technology. Conventional hard-wired electronics has however a fundamental drawback in broadband transmission: the high-frequency components of the signal are attenuated due to increased conductor resistance in the high frequency region, and also cable

material degrades the signal intensity as frequency increases. The transmission distance has to be shortened for higher transmission rates to occur. This limitation has a crucial effect on the system design. We have then to take into account that electrical transmission lines are highly lossy and highly sensitive to electro-magnetic interference: these effects pose further constraints in terms of power dissipation and interconnection-line density.

In this perspective, Optical Interconnections (OI) based on optical fibers, waveguides or free-space beam propagation (according to the specific system size and interconnection distance) are powerful alternatives to electrical hard-wiring. Though many studies have been already carried out, OI is still an open and promising research field. But with this survey report we would like to show that OI is not only a matter for theoretical speculation. Even if commercial products based on OI are still a minority, there is a vast set of prototypes reported in literature in which this technology has been successfully experimented. In this report we would like to focus the architectural aspects, so to provide a reference basis for designers who attempt to conceive new architectures for photonic interconnection systems.

The interconnection-design problem is common to systems for quite different applications in the area both of telecommunications and information processing. It should be noted that the interest to optics for interconnection has developed first in the area of super-computing and several new optical components have been developed for that application. Therefore our study will consider the following application scenarios: switches for optical-circuit networks (optical add-drop multiplexers and cross-connects), packet-switching nodes, super-computing systems. We will consider also a final very particular application, which is network on-chip. As a matter of fact, also integrated circuits have become very-complex “multi-module” systems, despite their small foot-print, and they require an internal interconnection network to transfer data between the different zones of the chip: OI has shown to be a helpful technology also for this applications.

The report is organized as follows. In the first section we will characterize the application scenarios listed above, trying to identify for each the main modules and subsystems that need to be interconnected and the requirements. In the second section we present, for each application, the state-of-the-art of optical interconnection and the best known OI solutions that have been reported in literature.

2. Functional architectures

Requirements and topology of interconnection systems depend on the number and types of modules which compose a system and their mutual communication relationships. These features usually depend on the functions the system performs and how its tasks are distributed between its subsystems: in a word, they depend on the functional architecture of the system.

This section tries to summarize the functional architectures of the systems we will consider, explaining for each the main features and operations. The functional architecture is in most cases independent of the interconnection technology adopted for implementation, i.e. it is the same regardless whether optics or electrical wires are used for interconnection. This section is intended to be an introduction to better understand Section 3, specifically dedicated to OI.

2.1 Add-Drop Multiplexers and Cross-Connects

In the first generation optical networks only point-to-point transmission systems were used. Since wavelengths division multiplexing (WDM) was unknown, only one wavelength went through an optical fiber. To increase the transmission rate between two adjacent nodes either additional fiber had to be applied, or faster transmitters and receivers had to be used. But in many practical cases these solutions were too expensive or just impossible. The WDM

technology allowed telecommunication operators to use one optical fiber in a better way. Several or even dozens of wavelengths could be sent in the same optical fiber at the same time.

Optical Add/Drop Multiplexers (OADM) were used to add and/or remove a wavelength (or wavelengths) in a node of a WDM network. They allow network operators to more easily manage network traffic; to introduce new wavelength services and faster react to users' needs. When the number of wavelengths grew and more nodes were installed in a network, traditional OADM systems became only a temporary solution. Reconfiguration required long time and a very precise work of specialists. This was of course a very expensive solution. Therefore, introduction of Reconfigurable OADMs (ROADMs) was quite natural and only a matter of time. It gives a possibility to reconfigure connections remotely in quite a short time (few minutes) and using very friendly graphical interfaces. ROADMs are equipped with some kind of switching elements.

The standard and simple ROADM system works in a WDM ring in two directions. Such node is called two-degree or two-dimension. Two directions in which the wavelengths are directed are called east and west, respectively. In case when an ROADM combines few rings together or combines a ring (or rings) with a mesh network, it can be of 3-, 4- or even 8-degree. An example of a 2-degree ROADM is shown in Figure 1. All modern ROADM systems have automatic power equalizer at inputs, outputs, as well as add and drop channels. Most of them have also Optical Power Monitoring (OPM).

The ROADM systems were firstly implemented in WAN networks, nowadays they are also used in Metro networks. According to questionnaire done by the Infonetics Research in 2005 among the world biggest telecommunication operators at least 62% of them wanted to use ROADMs to build their Metro networks [1].

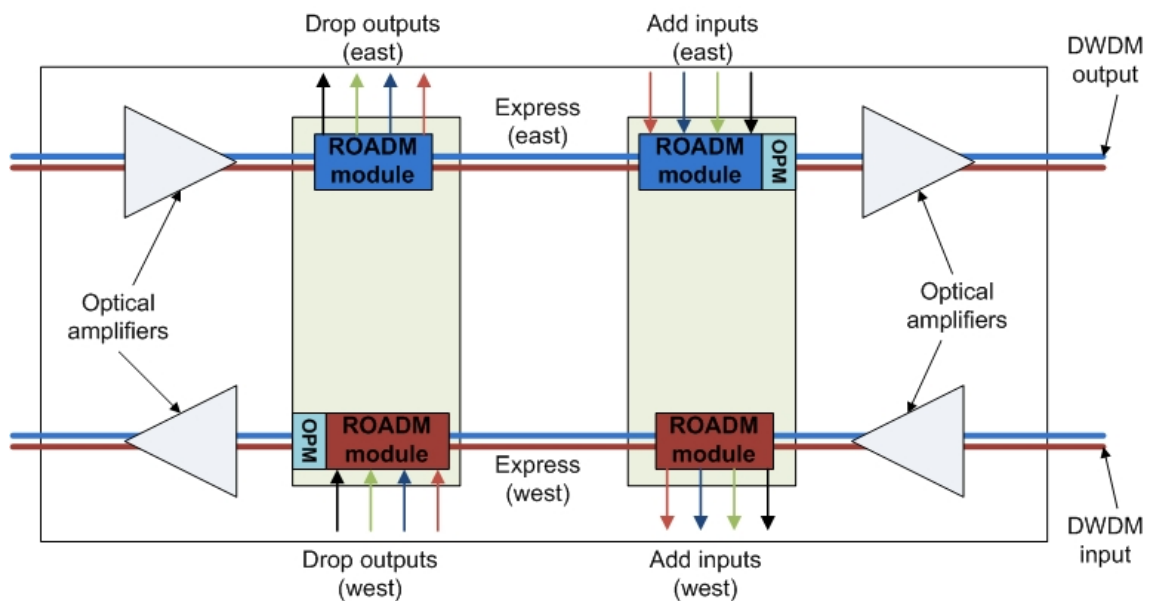


Figure 1 The 2-degree ROADM system

A further evolution of optical-switching equipment for WDM networks led from the OADM systems to the optical cross connects. Optical Cross Connects (OXC) (see scheme in Figure 2) are located at nodes cross-connecting a number of fiber pairs and support add and drop of local traffic. To offer pairs and support add and drop of local traffic. To offer flexible path provisioning and network resilience, they normally utilise a switching fabric for routing of any incoming channels to the appropriate output ports. In terms of technology, OXCs can be

opaque or transparent. Opaque OXCs are either based on electrical or optical switch fabrics surrounded by OEO interfaces. In OXCs using electrical switching, depending on the technology and architecture, sub-wavelength switching granularities can be supported providing edge and intermediate grooming capabilities for more efficient bandwidth utilisation. Opaque OXCs also offer inherent regeneration, wavelength conversion and bit-level monitoring. In transparent OXCs the incoming signals are routed transparently through the optical switch without optoelectronic conversions. The switching granularity may vary between fibre, wavelength band or wavelength channel level. Transparent nodes do not support regeneration capabilities and may significantly impact the scalability of the overall solution as present optical networks are analogue in nature and both transmission and switching introduce impairments. To overcome this, partially regenerating architectures have been proposed. These provide the ability to selectively regenerate individual channels of degraded signal quality through a set of regenerators that can be selectively accessed. Full and partial wavelength conversion can also be applied to reduce wavelength blocking, offer improved bandwidth utilisation. All-optical solutions offer transparency to the signal bit-rate and format.

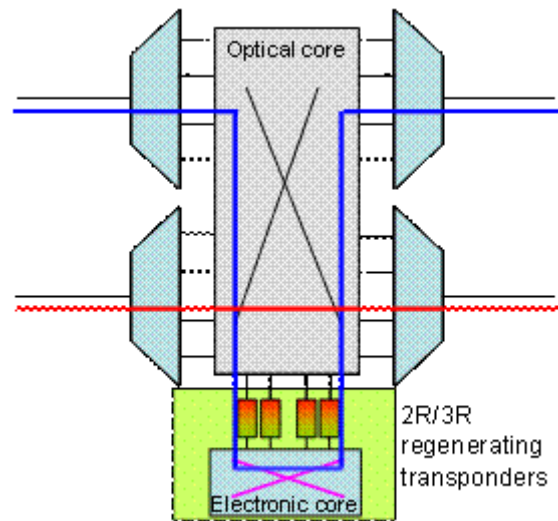


Figure 2 Architecture of a translucent optical cross-connect node

OADM, R-OADM and OXC are all suitable for nodes of optical circuit-switched networks. Circuit switching is a relatively simple networking function, as control of the node is fully separated from data flow and the node does not take configuration decisions inferring information from transit data. Switching-time constraints are not critical as there is a setup phase for each lightpath prior that data are transferred.

2.2 Frame/packet switches

It has been widely accepted that statistical multiplexing at the optical layer exhibits a very promising way of resource utilisation in optical networking, hence switching at the sub-wavelength layer has emerged as important requirement. Optical switching architectures have been proposed for different applications and been implemented at the level of packets, frames or bursts with the architectural details dictated by the network scenario deployed.

Packet switching is much more demanding than circuit switching in terms of node functional architecture. The operation of transferring a packet from an input to an output requires inferring information from the packet, solving possible contentions for the output and

managing a buffer system for packet queuing. The fundamental operations are similar in a frame (layer-2) switch and in a packet (layer-3) router. However, the router is more complicated because of further additional packet-processing operations. Therefore we will refer in the following to a high-performance core-router, to be more comprehensive.

IP-router functions can be classified into two categories: datapath functions and control plane functions [2]

- *The datapath functions* such as forwarding decision, forwarding through the backplane, and output link scheduling are performed on every datagram that passes through the router. When a packet arrives at the forwarding engine, its destination IP address is first masked by the subnet mask (logical AND operation) and the resulting address is used to lookup the forwarding table. A so-called longest prefix matching method is used to find the output port. In some applications, packets are classified based on the IP source/destination addresses, transport layer port numbers (source and destination), and type of protocol, which is generally called 5-tuple. Based on the result of classification, packets may be either discarded or handled at different priority levels.
- *The control plane functions* include the system configuration, management, and exchange of routing table information. These are performed relatively infrequently. The route controller exchanges the topology information with other routers and constructs a routing table based on a routing protocol.

The Router Controller (RC) maintains a routing table and updates it based on the supported routing protocols. The routing table is used to generate a forwarding table that is then downloaded from the RC to the forwarding engines in the interface cards. It is not necessary to download a new forwarding table for every route update. Route updates can be frequent, but routing protocols need time, in the order of minutes, to converge. The RC needs a dynamic routing table designed for fast updates and fast generation of forwarding tables. Forwarding tables, on the other hand, can be optimized for lookup speed.

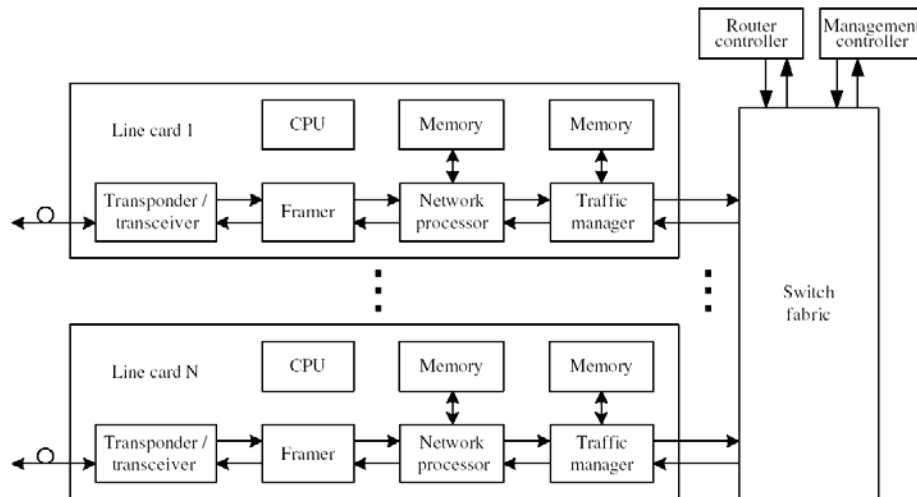


Figure 3 Typical router architecture

Figure 3 shows a typical router architecture, where multiple line cards and an RC are interconnected through a switch fabric. Each line card supports at least one full-duplex fiber connection on the network side, and at least one ingress and one egress connection to the switch fabric backplane. A line card is characterized by:

- *Transponder/Transceiver*: this component performs optical-to-electrical and electrical to optical signal conversions, and serial-to-parallel and parallel-to-serial conversions.

- *Framer*: a framer performs synchronization, frame overhead processing, and cell or packet delineation.
- *Network Processor*: the NP mainly performs table lookup, packet classification, and packet modification.
- *Traffic Manager*: to meet the requirements of each connection and service class, the TM performs various control functions to cell/packet streams, including traffic access control, buffer management, and cell/packet scheduling.
- *Central Processing Unit*: the CPU performs control plane functions including connection set-up/tear-down, table updates, register/buffer management, and exception handling.

Core routers are designed to move traffic as quickly as possible. With the introduction of diverse services at the edges and rapidly increasing bandwidth requirements, core routers now have to be designed to be more flexible and scalable than in the past. To this end, design goals of core routers generally fall into the following categories:

- *Packet forwarding performance*: core routers need to provide packet forwarding performance in the range of hundreds of millions of packets per second. This is required to support existing services at the edges, to grow these services in future, and to facilitate the delivery of new revenue-generating services.
- *Scalability*: as the traffic rate at the edges grows rapidly, service providers are forced to upgrade their equipment every three to five years. Latest core routers are designed to scale well such that subsequent upgrades are cheaper to the providers. To this end, the latest routers are designed as a routing matrix to add future bandwidth while keeping the current infrastructure in place.
- *Power requirements*: another issue with core routers is the amount of real estate and power required to operate them.
- *Service delivery features*: in order to provide end-to-end service guarantees, core routers are also required to provide various services such as aggregate DiffServ classes, packet filtering, policing, rate-limiting, and traffic monitoring at high speeds.
- *Availability*: as core routers form a critical part of the network, any failure of a core router can impact networks dramatically. Therefore, core routers require higher availability during high-traffic conditions and during maintenance. Availability on most core routers is achieved via redundant, hot-swappable hardware components, and modular software design. The latest core routers allow for hardware to be swapped out and permit software upgrades while the system is on-line.
- *Security*: as the backbone of network infrastructure, core routers are required to provide some security related functions as well. Besides a secure design and implementation of their own components against denial of service attacks and other vulnerabilities, the routers also provide rate-limiting, filtering, tracing, and logging to support security services at the edges of networks.
- *Cost*: electronic components, optical components, electrical consumption, build re-engineering, air conditioning
- *Multicast capability*: in order to provide one-to-many interconnections in support to multicast services (e.g. streaming media, internet television).

The complexity of a large-capacity router is much higher than that of an OXC. This is because, for packet switching, there is a requirement to process packets (such as classification, table lookup, and packet header modification), store them, schedule them, and perform buffer management. As the line rate increases, the processing and scheduling time

associated with each packet is proportionally reduced. Also, as the router capacity increases, the time interval for resolving output contentions becomes more constrained. Interconnection and memory technologies are the most demanding when designing a large-capacity packet switch.

We now introduce as benchmark a very popular electrical core router on the market: Cisco System's Carrier Routing System CRS-1 (Figure 4) [3]. Cisco's solution is surely one of the most scalable, ostensibly enabling throughput of up to 92 Tbit/s. CSR-1 design allows the system to combine as many as 72 line card shelves interconnected using eight fabric shelves to operate as a single router or as multiple logical routers. It can be configured to deliver any capacity between 1.2 and 92 Tbit/s and the router as a whole can accommodate 1152 40 Gbit/s interfaces (16 for each line card shelf). Router Engine is implemented using at least two route processors in a line card shelf. Cisco CRS-1 uses a three-stage, dynamically self-routed Benes topology based switching fabric. The first-stage (S1) of the switch is connected to ingress line cards. Stage-2 (S2) fabric cards receive cells from Stage-1 fabric cards and deliver them to Stage-3 fabric cards that are associated with the appropriate egress line cards. Stage-2 fabric cards support speedup and multicast replication. The system has eight such switch fabrics operating in parallel through which cells are transferred evenly. This fabric configuration provides highly scalable, available, and survivable interconnections between the ingress and egress slots. However, it requires 1MW of power supplied and 1MW for cooling.

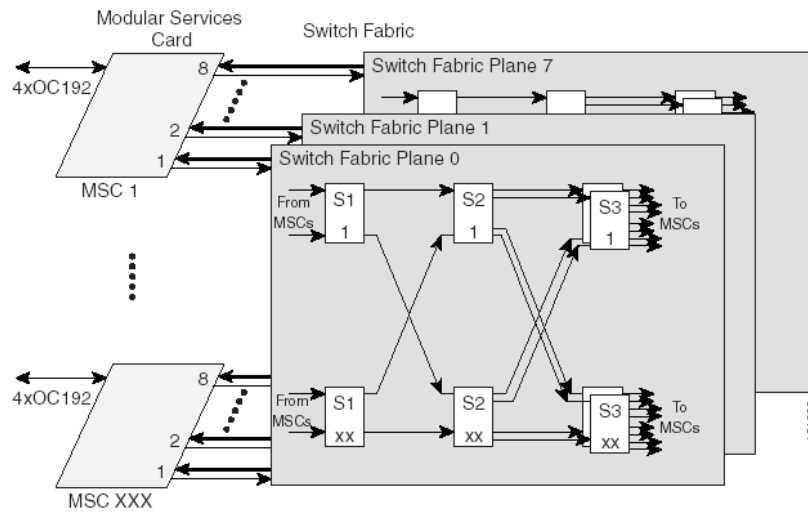


Figure 4 Cisco CRS1 Routing System

2.3 Supercomputing machines

Though advances in processor speed have been outstanding, a single processor is too limited to carry out the very high-demanding tasks required by scientific applications. Usually supercomputing requires sharing the computation load on a plurality of processors working on parallel. Often, also RAM and other functions are distributed over several modules.

Massively parallel high-performance computing (HPC) systems require large-scale, high-bandwidth, low-latency interconnection networks. HPC interconnects must handle the arrival, buffering, routing, forwarding, and error management of high-rate data and control streams in scalable local networks. HPC switches must deliver very low latency, an extremely low bit-

error rate, a high data rate, and extreme scalability. Moreover, they must handle bursty traffic efficiently, with demands ranging from very small messages (such as collectives and syncs) to bulk data set transfers (such as memory pages). The specific HPC-based requirements for recently common implementations are:

- Low latency (Less than 1 μ s)
- Switch ports operating at 40 Gbps
- 10⁻²¹ bit-error rate;
- scalability to at least 2,048 nodes

The system memory model imposes additional requirements. Shared-memory, cache coherent, HPC multiprocessor systems, whether they have uniform or non-uniform memory access, rely on hardware support to enforce memory coherency and cache-line granularity consistency.

2.4 On-chip networks

The current trend of including from tens to hundreds of cores on a single die of silicon is being shaped mainly by what are called the processor-memory network architectures, in which several independent processors and the associated resources (memories, coprocessors, controllers) are interconnected on chip through a high capacity (possibly optical) network.

The advantages of optical links over electronic technologies are under extensive discussion in the scientific community; at the on-chip scale optical interconnections appear as a performance- and power-convenient option, given their high capacity and distance-independent physical impairments.

3. Optical interconnect architectures

In this section we classify optical interconnect systems depending on the field of application: switch/router, high-performance-computing and on-chip networking.

3.1 Add-Drop Multiplexers

Commercial optical switching products have emerged with much attention given to the various ROADMs despite the limited scalability of the corresponding networks. A direct extension of a ROADM to a multi-degree node (number of I/O fibres >2) is the wavelength selection switching (WSS). WSS solutions have been presented from Polatis, CoAdna, JDSU, Metconnex and Capella Photonics, all based on different technologies and varying node degrees. WSS's are designed as 1xK devices, and are able to route any combination of 40 input wavelengths to any of the K outputs (currently K<10). This technology is used in single fibre 40 channel DWDM line systems to provide optical bypass of intermediate nodes and allow remote wavelength circuit provisioning.

We will describe three types of ROADM systems, introduce how they work and what advantages and disadvantages they have.

Nowadays the most popular and very often used technologies to build ROADM systems are:

- liquid crystals – in ROADMs based on Wavelength Blocker (WB) systems,
- Planar Lightwave Circuits (PLC) – in ROADMs designed in one integrated circuit,
- Micro-Electro-Mechanical Systems (MEMS systems) – in ROADMs based on wavelength selection switching (WSS).

Being ROADMs optically-transparent devices, it is clear that OI is a must (it would not have sense to use electronic switching, as these devices are exactly conceived to introduce transparency in networks). We can affirm that the optical interconnection system is in practice the device itself. Therefore in the next we will describe in details the three types of ROADMs mentioned above.

3.1.1 ROADMs based on wavelength blocker systems

The wavelength blocker system was the first system available for commercial use and it was used in circuit switching. An example of the WB system is shown in Figure 5. Wavelength multiplexing and demultiplexing is done using grating of and controlled liquid-crystal elements introduce polarization diversity. This polarization diversity allows to make the proper suppression or even in the extreme case allows to block selected wavelength. The ROADM based on WB are capable of switching high number of channels, for example 128/64 channels with 50/100 GHz channel spacing.

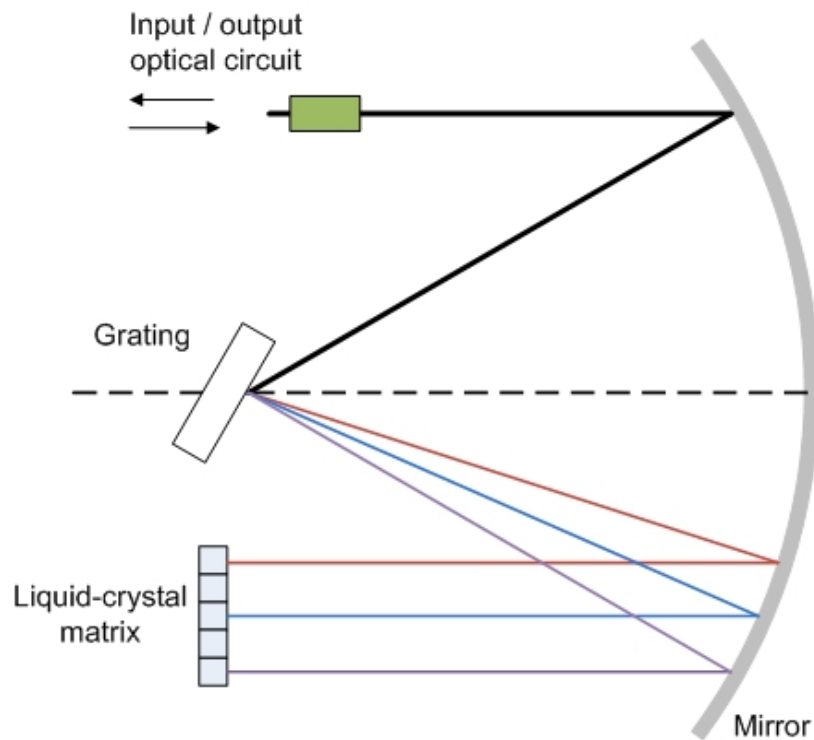


Figure 5 The Wavelength Blocker ROADM system with grating and liquid-crystal elements

These ROADMs have two ports and by using the Dynamic Equalizer Channel (DCE) allows to control a power level for each wavelength separately (see Figure 6). This kind of devices are build from high class optical components with broadband bandwidth, which have low dispersion characteristics and enables transmission rates of 10/40 Gbit/s. This attributes are especially important in ROADM systems based on the liquid-crystal technology connected in cascade.

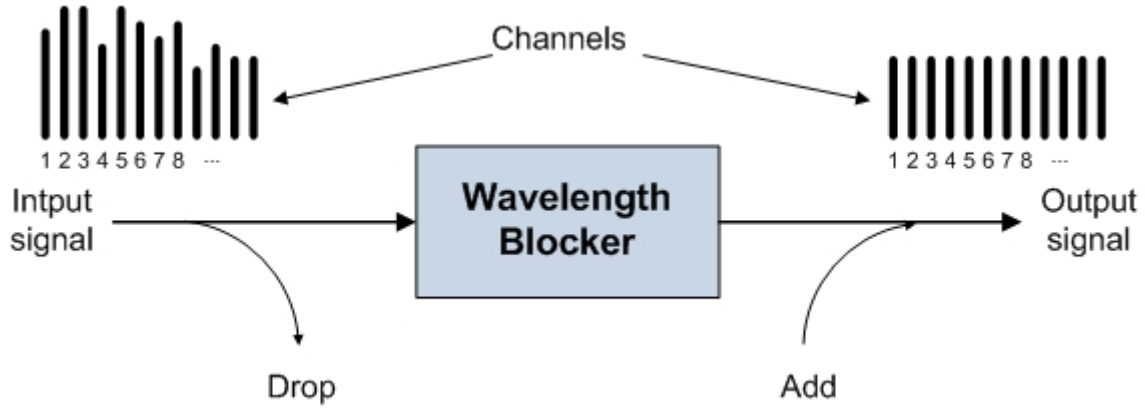


Figure 6 Dynamic Equalizer Channel (DCE)

Table 1 Comparison of the WB-ROADM systems

Company	Model	Band, channel spacing	Attenuation [dB]	Crosstalk [dB]	Dynamic equalization [dB]	Other
Xtellus	Wavelength Blocker 7000	C or L; 100 or 200 GHz spacing	< 5	> 40	< 15	up to 64 wavelengths; no moving elements
	Wavelength Blocker 9000	C or L; 50 or 100 GHz spacing	< 5	> 40	< 15	no moving elements
Santec	WBE 300	C or L; 50 or 100 GHz spacing	n/a	high	n/a	no moving elements
Avanex	Powerblocker	C or L; 50 or 100 GHz spacing	< 60	40	< 20	
CoAdna	Dynamic Channel Equalizer Blocker (DCEB)	n/a	low	n/a	n/a	up to 40 optical channels; no moving elements
Agiltron	Reconfigurable Wavelength Blocker	C or L; 50 or 100 GHz spacing	< 6	40	20	
JDSU	Dynamically Reconfigurable Wavelength Blocker	C or L; 50 or 100 GHz spacing	< 6	about 43	< 20	

The ROADM system based on WB are in the market from many years. They were implemented mostly in long-haul and ultra long-haul networks, they were reliable and the production process is mature. But despite these advantages the WB-ROADM systems have many disadvantages. In the vast majority of the implementation the ports are colored, it means they depends on a wavelength. The colorless ports implementation in the WB-ROADM is possible but tunable filters on Drop ports and tunable lasers on Add ports must be

used. This solution is, however, very expensive. In result, the high cost of the WB modules and limited possibilities (2-degree nodes and colored ports) cause that WB-ROADM are nowadays not so competitive. They are used only in the oldest systems. In Table 1 the comparison of same WB-ROADM models is given.

3.1.2 ROADM systems based on PLC

The Planar Lightwave Circuit technology allows integrating in one chip such components like gratings and switches. This technology seems to be very interesting especially in telecommunication applications. ROADM systems based on PLC can have add/drop ports for all available channels (typically it is about 40 ports in C band with 100 GHz spacing). They are used in nodes build from two modules: an add module and a drop module. About 80% of all nodes in Metro networks are 2-degree, thus the PLC ROADM systems are quite good solution. Typical architecture of 2-degree node is shown in Figure 7.

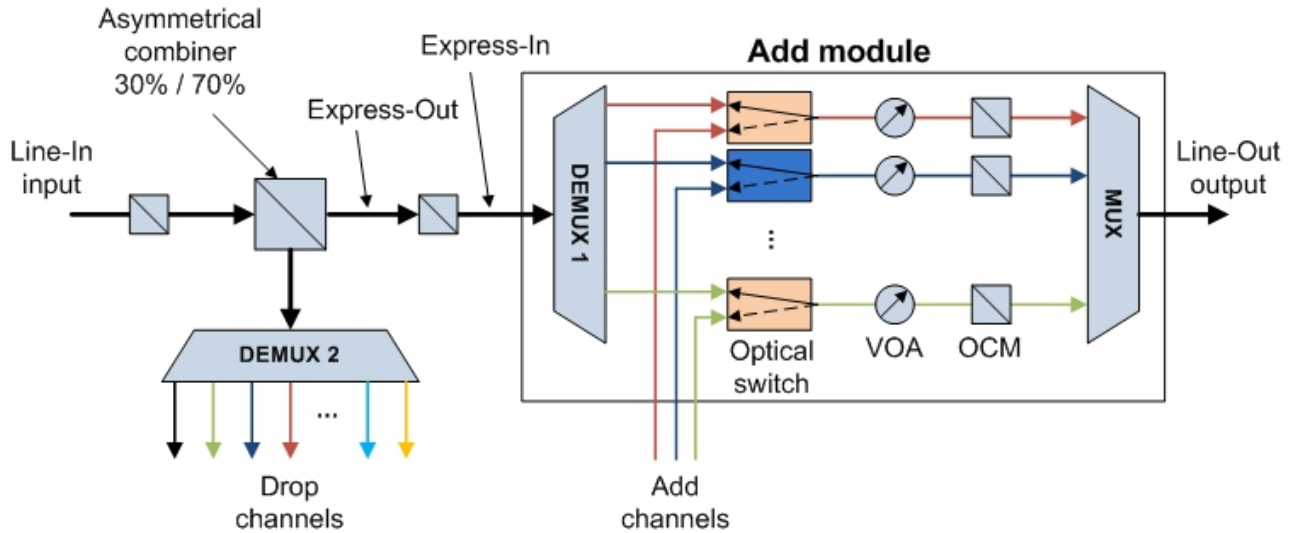


Figure 7 PLC-ROADM system

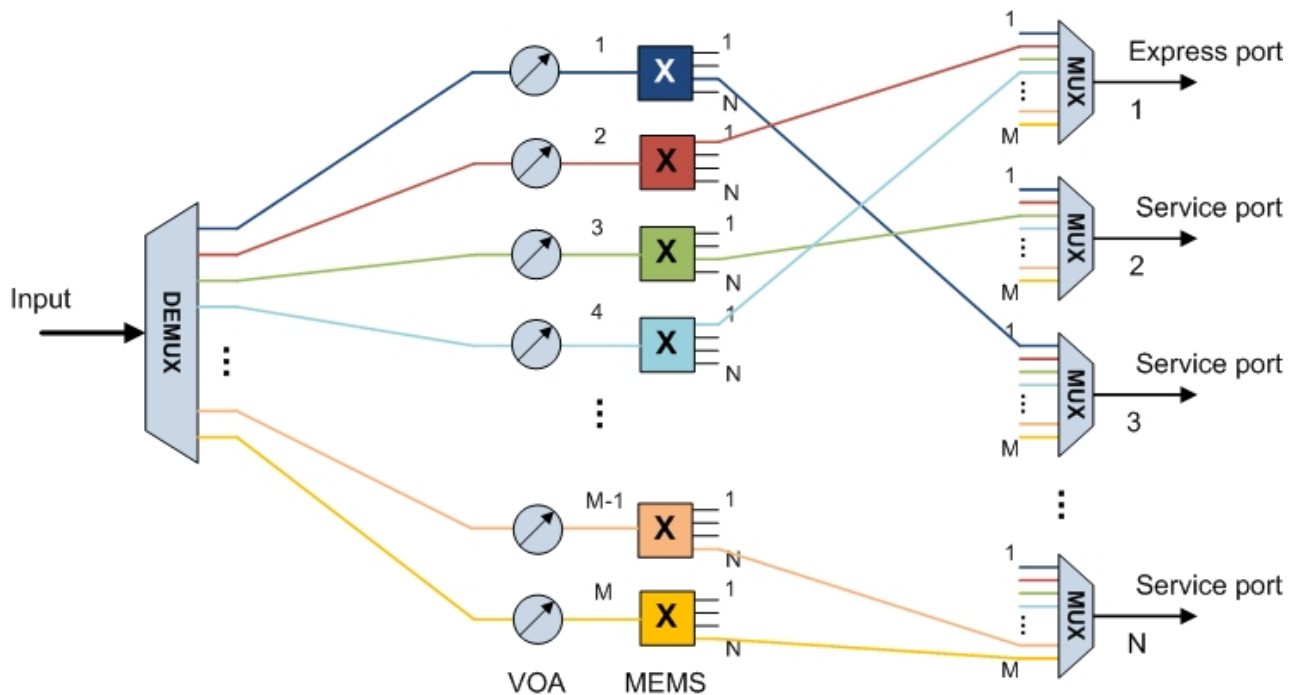
It consists of two pair of modules (the east and the west), the express output and the drop module. The multiplexing is realized by the AWG. Switches, having switching time in milliseconds, pass the proper wavelength through the express channel or just block the input wavelength and run the add function. Additionally, Variable Optical Attenuators (VOA) work only as DCE to align the power of both: added channels and channels from the express part. Optical Channel Monitors (OCM) can be also seen in Figure 7. They are used to control the quality of the optical signal. In Table 2 the comparison of same PLC-ROADM models is shown.

Table 2 Comparison of the PLC-ROADM systems

Company	Model	Band, channel spacing	Attenuation	Crosstalk	Dynamic equalization
JDSU	PLC ROADM	C and L (40 channels)	5,8-7,5 dB	n/a	0-25 dB
Dupont	IRoad 4000	C and L (40 channels); 100 GHz spacing	up to 9,5 dB	50 dB	0-20 dB

3.1.3 The ROADM systems based on WSS

The ROADM node based on wavelength selection switch allows to connect N adjacent network nodes. The WSS-ROADM nodes can be used not only in the simple ring topology. The standard configuration allows to send data in one Express port and in $N-1$ DROP ports. Similarly as WB-ROADM, most of WSS-ROADM systems are based on the free space optics. The difference is that liquid-crystals are now replaced by MEMS devices. The $N-1$ WSS-ROADM system is shown in Figure 8. Application of MEMS gives broad bandwidth, low dispersion and transmission at 10-40 GB/s rates. Furthermore, WSS-ROADMs allow to switch each wavelength to any available port, so we have colorless ports functionality also at ADD/DROP ports.

**Figure 8** The $1 \times N$ WSS-ROADM system

In case when WSS is used in mesh networks, the 1×5 switch can be used in 4-degree node (including possibility for ring-to-ring connections). In this type of usage if one $1 \times N$ WSS is used to serve one degree of the node (for a connection with its neighbour), multiplexers used

to add or remove signals (Add/Drop) are set up permanently. Thus, Add/Drop ports are not colored. In case when WSS is used in the mesh network and its ports should be colored it is necessary to use three 1xN WSS for every node degree.

Table 3 Comparison of the WSS-ROADM systems

Company	Model	Channel Spacing [GHz]	Attenuation [dB]	Crosstalk [dB]	Dynamic equalization [dB]	Number of channels
JDSU	1 × 9	50	< 6,6	40	< 15	96
JDSU	1 × 9	100	< 6,5	45	< 15	46
Optium	DWP 50/100	50	< 6,5	40	< 15	80
		100				40
COADNA	1 × 9	100	low	n/a	n/a	48
Capella	4500	100	low	n/a	n/a	45
Capella	9600	50	low	n/a	n/a	96

Nowadays the WSS platforms use 1x9 devices with 100 GHz spacing and 5x1 devices with 50GHz spacing. Despite the fact that these devices exist not too long on the market, their production process is very mature and they are reliable. To decrease the number of moving elements (MEMS) Liquid Crystal on Silicon (LCoS) technology is used. In Table 3 the comparison of same WSS-ROSDM models is shown.

Table 4 Comparison of the different models of the ROADM systems

ROADM	Advantages	Disadvantages
WB	- first one for commercial usage, - add/drop ports do every $\lambda = N$	- high dimensions, - high cost, - constant λ per port or very expensive tuning laser, - constant degree of the node
PLC	- lowest cost for 2-degree node, - low dimensions, - low power consumption, - scalable, - no mobile elements, - no free-space optic, - shockproof	difficult in extension for > 2-degree nodes and uncolored ports (required big matrixes, for example 40×8)
WSS	- ports can be uncolored (different wavelength between two ports), - can be extended to mesh topology	- high cost, - unprofitable for 2-degree nodes, 3 WSS modules per one degree in node, - use free-space optic, - sensitive for vibrations

In this part of the report we described three different technologies used in the commercial ROADMs systems. They are based on wavelength blocker, on PLC technology and on WSS switches. Advantages and disadvantages of these technologies are compared in Table 4. In Table 5 complete DWDM systems supporting ROADMs are shown. Most of them are based on WSS switches (they use MEMS technology).

Table 5 Comparison of the ROADM systems available in the market

Producer	Model	Number of wavelengths	Topology	Switching	Interfaces
Adfa	FSP 3000 RR	40	add/drop, ring, mesh, star	$1 \times N$ WSS	Ethernet, SONET/SDH, G.709 OTN, Video, SAN from 8Mbit/s to 40Gbit/s, Ethernet 10/100/1000/10000Mbit/s, ESCON, Fibre Channel, FICON, ETR/CLO, OC-3/12/48/192, STM-1/4/16/64, OUT-1/2, ATM 155/622/2488
Adfa	FSP 3000	80 (each 40GB/s)	n/a	n/a	Ethernet, SONET/SDH, G.709 OTN, Video and SAN from 8 Mb/s to 40 Gbit/s Ethernet 10/100/1000/10000 Mbit/s, ESCON, Fibre Channel 1/2/4/(8)/10 Gbit/s, FICON 1/2/4 Gbit/s, Coupling Link 1Gbit/s, 2 Gbit/s, ETR/CLO and STP, OC-3/12/48/192/768, STM-1/4/16/64/256, OTU-1/2/3, ATM 155/622/2488
Alcatel-Lucent	n/a	40	point-point, linear add/drop, ring, mesh	n/a	OC-192 / STM64, 10 GbE, FICON, ESCON, Fibre Channel, Video, Alien Wavelengths
CISCO	ONS 15454	40	n/a	WSS 1×9	GbE, OC-192

Producer	Model	Number of wavelengths	Topology	Switching	Interfaces
Ericsson	Marconi MHL 3000	40 (each 40GB/s), possibility to extend to 80 channels	point-point, ring, mesh	WSS	SDH/SONET/CBR: • STM-1/OC-3/CBR-155Mbit/s, • STM-4/OC-12/CBR-622Mbit/s, • STM-16/OC-48/CBR-2,5Gbit/s, • STM-64/OC-192/CBR-10Gbit/s, • STM-256/OC-768/CBR-40Gbit/s, • Ethernet, • GbE, • 10GbE WAN/LAN-PHY, Video: 270Mbit/s, Storage: FICON, Fiber Channel: 1G/2G/4G/10G, ESCON
Lambda	Lambda Node 5000	n/a	n/a	WSS	n/a
ECI	n/a	n/a	mesh, multi-ring	WSS	n/a
Ciena	CN 4200	n/a	n/a	WSS	n/a
Fujitsu	Flashwave 9500	40	n/a	WSS	OC+3 to OC+192 Gigabit Ethernet
Meriton	Metrocore 7200	n/a	ring, mesh, star, point-point	WSS	DWDM, CWDM, G.709, Ethernet, SONET/SDH
Infinera	n/a	n/a	mesh, multi-ring	WSS	40 Gb/s (OC-768, STM-256), 10 Gb/s (OC-192, STM-64, 10GbE LAN and WAN PHY), 2,5 Gb/s (OC-48, STM-16), GbE, 622 Mb/s (OC-12, STM-4), 155 Mb/s (OC-3, STM-1)

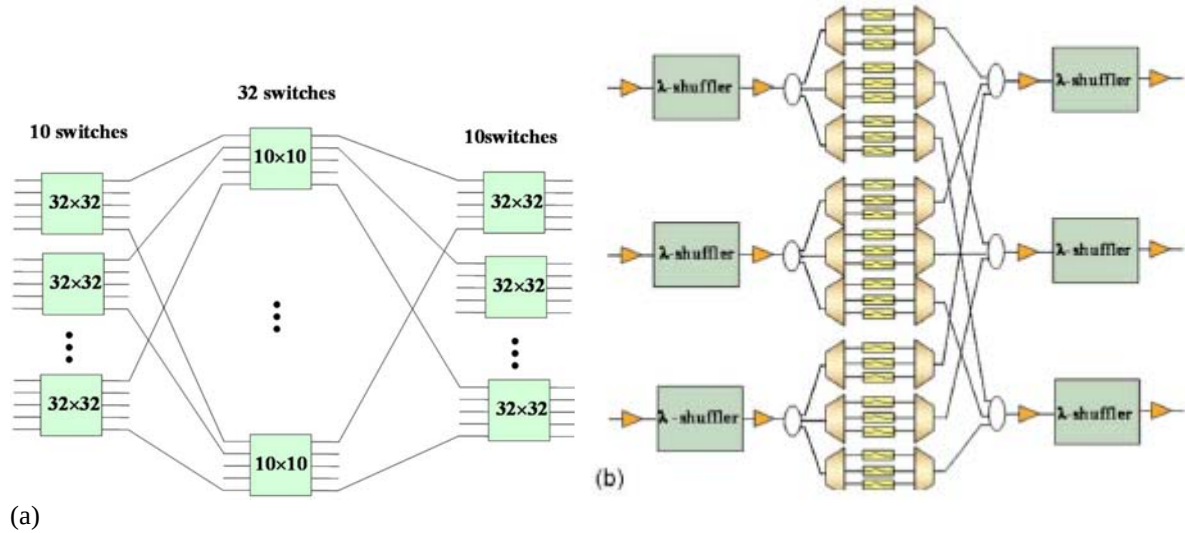
3.2 Optical Cross-Connects

Introduction of optical fibers to communication networks has caused a tremendous increase in the speed of data transmitted. The virtually unlimited bandwidth of optical fibers comes from the carrier frequency of nearly 200 THz [4]. Optical networking technology, such as add-drop multiplexers [5][6], reconfigurable photonic switches [7], and wavelength multiplexing division (WDM), has progressed well and facilitated optical networking in the past few years [8][9]. Especially, recent advances in dense wavelength multiplexing division (DWDM) technology have provided tremendous bandwidth in optical fiber communications [10]. However, the capability of switching and routing information at this high bandwidth (e.g., 1 terabit/s) has lagged far behind the transmission capability. Building a large-capacity switching system using only electronic technology may potentially lead to a system

bottleneck when interconnecting many electronic devices or modules, mainly caused by the enormous interconnection wires and the electromagnetic interference they would generate.

In Optical Circuit Switching (OCS) [11], switching is performed at the granularity of an optical circuit or lightpath. When the lightpath is based on a wavelength, OCS is sometimes referred to as optical wavelength switching. With the advancement of optical devices technology, several switch architectures based on WDM technology have been proposed for large-capacity switches. Optical switches have great potential to scale-up their switching capacity as the technology of some key optical devices matures.

An OXC is networking equipment featuring transparent steering of wavelength channels from input to output, so to setup time-varying circuits. In some cases OXC are also give the wavelength-conversion capability. The very same hardware setup can however also be exploited inside an optical interconnection system in order to switch optical signals “inside the box”. In this section we will present the optical switching architectures as seen in the literature for both telecom applications (OXCs) and optical interconnects.



**Figure 9 a) Schematic representation of a 320 x 320 FXC based on a Clos network arrangement
b) the λ -S- λ configuration**

There three main categories of optical switch architectures that were developed in literature and through R&D projects: multistage switches based on on-off or 2x2 space switches, Broadcast and Select switches and Wavelength routed switches that are usually based on combination of wavelength converter and filtering (typically AWG). A combination of any of the above can be performed and has been investigated in the e-photon/One project [12]. A way to classify the switch architectures refers to the blocking propeties. The switching fabrics are classified as follows:

- Rearrangeable Non Blocking (RNB), where every idle input channel can be connected to any idle output channel, requiring in some cases the internal rearrangement of some of the connections already set-up
- Strictly Non Blocking (SNB), where every idle input channel can be connected to any idle output channel independently from the current network state (so regardless of the connections already set-up and the internal network resources engaged for them).

In order to theoretically investigate and compare the different architectures it is typical to compare physical performance, blocking performance and complexity while investigating its networking performance. Evidently different architectures are implemented in different demonstrations and experiments that will be analysed separately. In Table 6, the complexity of the aforementioned switch architectures is discussed. Here the λ -S- λ (Figure 9(b)), and of the S- λ -S configurations discussed in e-photon/ONE+ are compared against the other known architectures: a single stage crossbar switch, a three-stage minimum-cost SNB Clos (Figure 9(a)) [13], a three-stage minimum-cost RNB Slepian-Duguid [14][15], a single-stage AWG router with tunable wavelength converters [16] (WC+AWG), a three-stage Clos-like configuration with AWG routers and tunable wavelength converters like in (3-stage WC+AWG), the architecture of reference [17] (MWTN) and the architecture of reference (DCS) [18].

Some explanations are necessary to correctly understand the figures shown in Table 6:

Clos, WC+AWG and 3-stage WC+AWG are not modular; hence, the full-size box should be installed in day-one, requiring significant first-day investment. Furthermore, those structures cannot be simply upgraded to accommodate higher capacity requirements. Finally, the DCS solution cannot be wavelength and link modular simultaneously.

In general, only broadcast and select architectures support multicasting and broadcasting, which is an important feature for reducing the amount of traffic spanning a network under dynamic λ -networking.

Tunable wavelength converters are counted as a single element. In reality, they include a number of SOAs and a tunable laser, and despite the possibility of monolithic integration, they are complex sub-systems.

The complexity of 3-stage WC+AWG shown in Table 6 can be lowered and/or the requested wavelength tuneability can be reduced, exactly as in λ -S- λ and S- λ -S architectures. However, since this optimization is beyond the scope of this comparison, the figures shown in Table 6 are obtained under the same operational assumptions used here (i.e. a fabric with M wavelength channels per fiber).

Table 6 Comparison with other architectures

Strictly Non Blocking			
Architecture	Component Count	WC count	WC tunability
Crossbar	$(NM)^2$	MN (fixed)	-
3-stage Clos	$4\sqrt{2} (MN)^{3/2} - 4MN$	MN(fixed)	-
WC+AWG	-	MN (tunable) NM (fixed)	NM -
3 stage AWG+WC	-	MN (tunable) (2M-1)N (tunable) (2M-1)N (tunable) MN (fixed)	2M-1 N M -
MWTN	$(2M-1)N^2$ (2M-1)N tun. filters	(2M-1)N (tunable)	M
DCS	$2MN^2$	MN (tunable)	M

Re-arrangeable Non Blocking			
	Component Count	WC count	WC tunability
Slepian-Duguid	$2\sqrt{2} (MN)^{3/2}$	MN(fixed)	-
3 stage AWG+WC	-	2 MN (tunable) MN (tunable) MN (fixed)	M N -
MWTN	MN^2 MN tun. filters	MN	M
S-λ-S	$2MN^2$	A,B: MN (tunable) C: + MN (fixed)	M -
λ-S-λ	MN^2	A, B: 2MN (tunable) C: + 2MN (fixed)	M -

As far as implementation of switching systems is concerned, the first optical switching concepts, were influenced by the electronic switching paradigm, and were attempting to construct large space switching matrices based on 2x2 elements (e.g. LiNbO₃). The main issue was to reduce the total number of the switching elements exploiting component minimizations concepts like Clos, Banyan etc networks [19]. In the WDM era the evolution of these switching matrix architectures lead to various configurations e.g. [20]. These architectures are still considered of prime interest and are widely deployed using a range of

optical technologies (MEMS, etc). However, these solutions were aiming to allow for optical circuit switching in core networks.

Later, approaches tailored to the properties of photonic technology emerged and were exploited in parallel with the aforementioned. For example, limited scale “single-hop” schemes (e.g. a star configuration or node-clusters interconnected via a central hub) were constructed by means of a central, over-dimensioned, optical node. Such a node can be constructed using either “passive” or “active” technologies or a combination of them. The simplest optical node architecture has the form of a passive star coupler [21] [RAINBOW]. A more advanced configuration, still using passive technology for interconnection, exploited wavelength routers (constructed based either on planar or free-space technology) [22][23] [OPERA]. In both cases, the channel/path selection is made by means of advanced optical technologies like tuneable transmitters and/or receivers. Very recently the [IRIS] project using these concepts is exploring the possibility to form very high capacity switches making use of optoelectronic conversion and electrical buffering. Still, single hop schemes can be implemented using both “passive” and “active” sub-systems like in [24]. The reason for using architectures based on “active” components for the interconnection/hub is either for providing fast reconfiguration that is controlled locally (and hence for reducing the overall control and coordination complexity that otherwise might be significant) or because there is a need to consolidate the expensive components in a single location for a more efficient time sharing of them. Nevertheless, in order to gratify the all-to-all connectivity pattern (i.e. strictly non-blocking operation) essential building blocks of such an architecture are passive splitters/combiners. Thus, what was in [24] the entire switch of the [RAINBOW] architecture, now it is an integral part of a more complex structure.

Recently there have been many projects that are dealing especially with optical interconnects emphasizing on the benefits of optical technology with respect to the electronics. For example ICT BOOM is an integration project that aims to pursue the systematic advancement of Silicon-on-Insulator (SOI) integration technology to develop compact, cost-effective and power efficient silicon photonic components that enable photonic Tb/s capacity systems for current and new generation high speed broadband core networks. As seen in [25] and in Figure 10, BOOM invests in the development of improved CMOS compatible waveguide technologies to fabricate miniaturized, low loss and fully reconfigurable wavelength routing cross-connects based on two dimensional grids of serially interconnected micro-ring resonators. BOOM will perform system level integration of all the components to assemble the first prototype rack-mount, ultra-high capacity router based on silicon photonics that will require minimum board space and power consumption and achieve a total throughput of 640 Gb/s. Equivalent architectures have been proposed in the framework of ICT HISTORIC project (Figure 11) where all-optical packet over electronic solutions are being investigated for several of the building blocks or subfunctions of an optical interconnect/router. To this end, architectures such as the one shown below and fitting into the generic description of an optical packet switched cross-connect will be demonstrated and assessed. However, during the first two years focus will be on the main building blocks such as the optical D-flip-flop and the gate / wavelength converter.

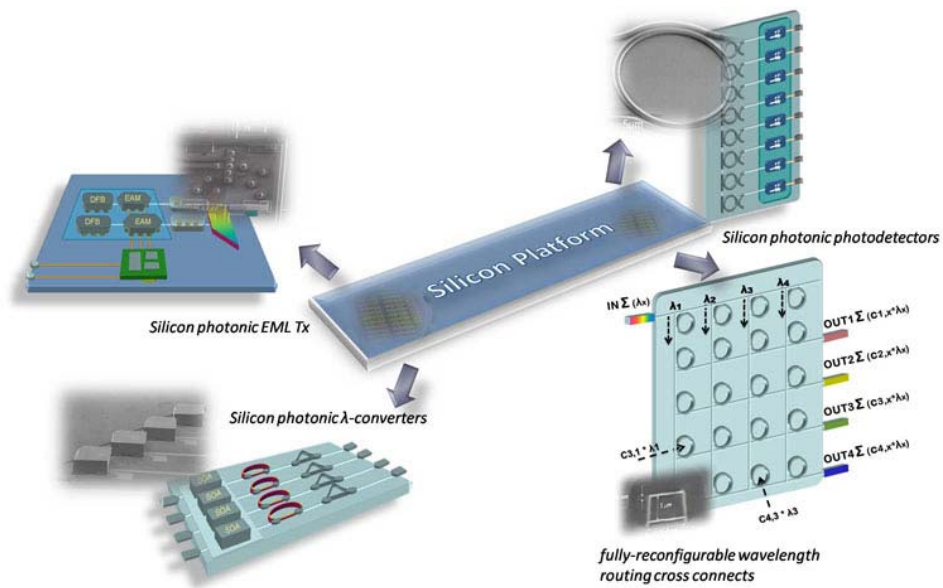


Figure 10 ICT BOOM Optical interconnect architecture [26][27]

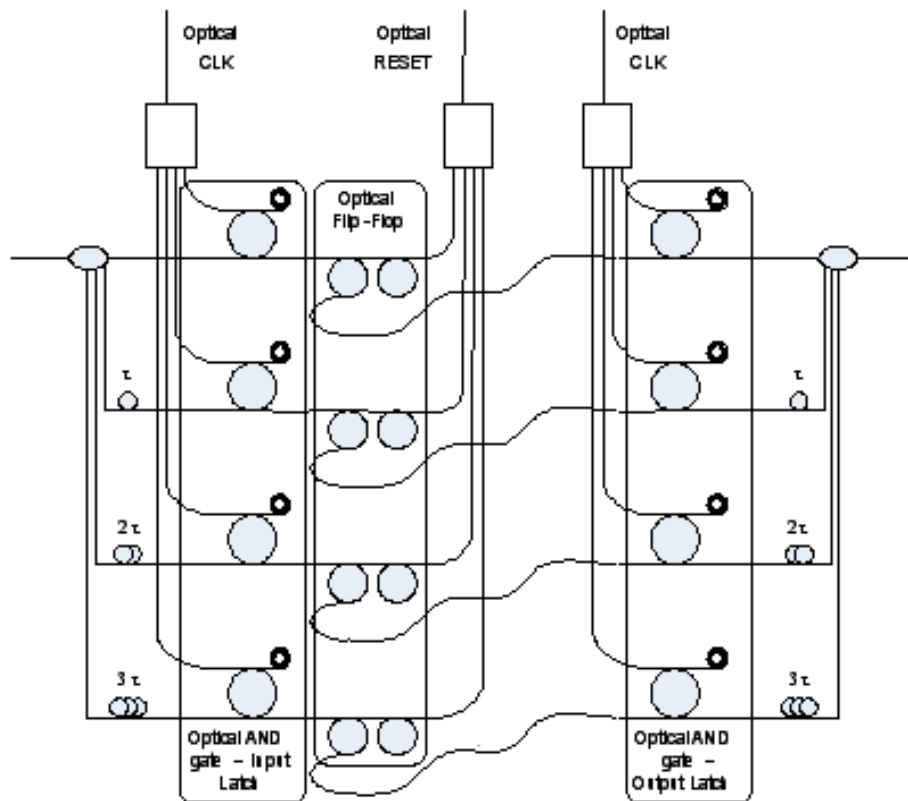


Figure 11 HISTORIC: Illustration of a functional digital photonic circuit based on micro-disk laser diodes; in this case a 4x1 mux/demux with latching memory

3.3 Frame/packet switches

Because data traffic is dominating the network and since circuit switching is not optimized for data traffic, there has been considerable focus by the research community on optical switching performed at the granularity of packets. However, this requires that packet header recognition, header processing, harder switching technology, and control be fully performed in the optical domain on a packet-by-packet basis. The availability of photonic components capable to operate at very high speed compared to electronics gave birth to Optical Packet Switching (OPS) [28] to address the question of multi-granular networking making use of the lowest level of granularity (packet). Thus, circuits would not be needed per se and would be understood as trains of slots/packets.

A photonic switch may require optical devices such as lasers, filters, couplers, memories, multiplexers, demultiplexers, and so on. At the present time, some optical devices are either very power-hungry or too slow in switching compared with electronic devices. However, it is possible to design high-capacity switches by the use of both electronic and optical technologies. In such switches, data transfer can be achieved through optical medium, and complicated functions such as contention resolution and routing control can be performed electronically. These switches are called hybrid switches. In this chapter, we review several approaches to build a large-capacity packet switch and discuss their advantages and disadvantages.

For the opto-electronic packet switches, optical switching networks are used for interconnection and transmission between electronic input and output interface. Logical control, contention resolution, and packet storage are handled electronically. In any adopted switching solution, optical router switches each packet which contains destination information. To minimize delay and maximize throughput, the switching time should be a fraction of the packet duration.

Packet type	byte	2.5Gbps	10Gbps	40Gbps	100Gbps
IPv6 header	40	128 ns	32 ns	8 ns	3,2 ns
ATM cell	53	170 ns	42 ns	11 ns	4,2 ns
MTU IP	512	1638 ns	410 ns	102 ns	41 ns
Max Eth segment	1500	4800 ns	1200 ns	300 ns	120 ns

Two aspects are then relevant for the throughput of the fabric: processing time, T_{proc} , and switch fabric reconfiguration time, T_{switch} . The sum of these must be less than T_{packet} to avoid a data flow bottleneck at the switch input.

Switch resources must be optimized according to target performance and are partially or completely shared on the data plane by all information to be switched. Switching resources operate in the space, time and wavelength domains to perform switch operation. The main aspect to be solved is contention, which arise when more data transfers require the same resource at the same time on the same wavelength channels. To face this task, control forwarding algorithms coordinate the use of switching resources trying to minimize the loss of calls/packets. These scheduling algorithms operate according to the interconnection architecture available, and this latter should be designed so that in the presence of optimal scheduling algorithms the target performance can be achieved. In particular the interconnection network must be equipped in the most general case with a sufficient amount of devices which operate space, wavelength and time domain switching.

A general architecture which performs these functions consists of three main sub-systems: space switching, wavelength conversion and buffer sub-systems. Depending on the interconnection strategy these subsystems can assume different configurations.

A number of very successful projects like ATMOS, KEOPS, DAVID and STOLAS were funded from EU in previous research frameworks in the last 15 years and many others were funded in the US. However it should be clearly mentioned that the notion and the objectives for OPS nodes and networks widely differ in those projects and two distinctive approaches gradually emerged: The first is using optical technology to imitate certain functions of electronic (IP) routers aiming to substitute parts of the electronic circuitry within these systems with advanced optical technology [29]-[42]. This approach is adopting a Router-centric networking paradigm demonstrating a paradigm shift in telecommunication since the term “circuit” is vanishing. The second is trying to identify the optimum balance between optical, electronic technologies and network architectures to achieve the requested functionality [43] [IRIS].

In recent years, OBS [44] was introduced as a compromise between OCS and OPS. Several innovative OBS architectures have been proposed to date. However, they present shortcomings of low efficiency in resource utilization, without significantly reducing the implementation complexity of OPS.

Let us now present in details some of the most relevant packet-switching architectures using optical systems for connecting the different modules, i.e. adopting OI.

3.3.1 HYPASS (Bell Communications Research)

HYPASS [45] in Figure 12 is an opto-electronic hybrid cell switch in which electronic components are used for memory and logic functions and optical components are used for routing and transporting data. In this Figure, bold continuous lines represent optical paths, bold dashed lines represent serial data paths, dotted lines are tuning current paths, and thin continuous lines are control signal paths. The switch is composed of two networks: the transport network and the control network. The architecture is based on the “broadcast and select” approach in both of the networks. There is a unique optical wavelength associated with each of the output ports. As shown in Figure 12, the transport network has tunable wavelength laser transmitters at the input side, fixed-wavelength receivers at the output side, and an $N \times N$ star coupler that transfers the incoming data from inputs to outputs. In order to transfer control information from output ports to the input ports, a similar network is used.

When a cell arrives at an input port of the switch, it is first converted from optical to electronic and its destination address is obtained. Then, the cell is temporarily stored in the corresponding input buffer. The tunable wavelength laser transmitter of the corresponding input port is tuned to the wavelength of the requested output port. When a request-to-send signal (or poll) is received from the corresponding output port via then control network, the cell is transmitted through the transport network.

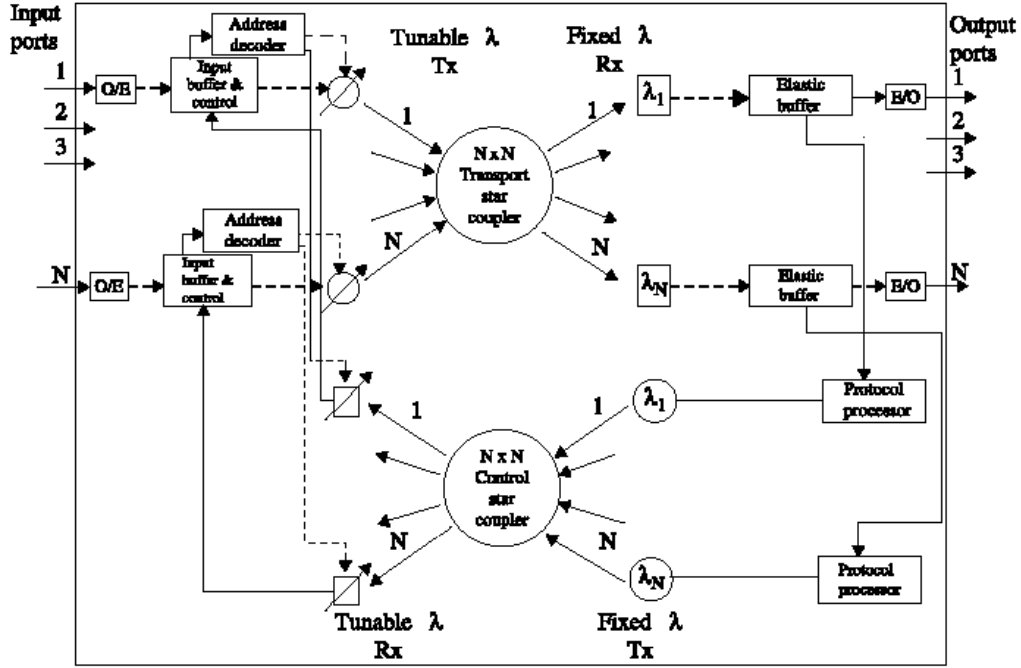


Figure 12 HYPASS implementation

The acknowledgment representing successful delivery of the cell is also transmitted through the control network. If there are multiple cells for the same output port, contention occurs. Power threshold detection or multiple bit detection on the cell preamble could be used to detect collision. The cells that do not get acknowledgments in a slot time are kept to retry later. The HYPASS architecture has advantages due to its parallel structure. However, since a slot time is based on the length of the polling step, transmission of a cell, and receipt of the acknowledgment, the time overhead for the electronic control and optical tuning operations are the factors that limit capacity. The switch does not have a multicasting capability due to the usage of fixed wavelength receivers at the output ports.

3.3.2 STAR-TRACK (Bell Communications Research)

As shown in Figure 13, the switch [46] is composed of two internal networks: an optical star transport network and an electronic control track surrounding the star network. The optical transport network has fixed wavelength optical transmitters at the input-ports side, and wavelength tunable optical receivers at the output-ports side. There is a unique wavelength associated with each input port. Input and output ports are connected through an optical star coupler. Cells arriving at the input ports are stored in the input buffers after optical to electronic conversion. There are two control phases in a cell transmission cycle. In the first phase, input ports write their output port requests into the tokens circulating in the control network. The output ports read the tokens and tune their receivers to the appropriate input port wavelengths in the second phase. Then, the cell transmission starts over the star transport network. The transmission and control cycles are overlapped in time in order to increase throughput. Since each input has a unique wavelength and there is input-output port pair scheduling prior to transmission, cells are transmitted simultaneously without causing contention.

This architecture allows multicasting. However, the throughput of the switch may degrade as the number of multicasting connections increases due to output port collisions in the first phase. It is shown that this problem can be alleviated by call splitting (i.e., allowing a multicast call to be completed in multiple cell slots). This architecture can support different

priority levels for the cell by adding minor tracks into the control network. However, in this case, the token should recirculate among the input ports more than once depending on the number of priority levels. This will increase the length of the write phase and result in longer cell processing time.

The main drawback of the switch is the sequential processing of the token by input and output ports. As a result, the time taken by the token to travel through the entire ring increases with the size of the switch. In case of multiple priority levels, the recirculation period for the token becomes even longer.

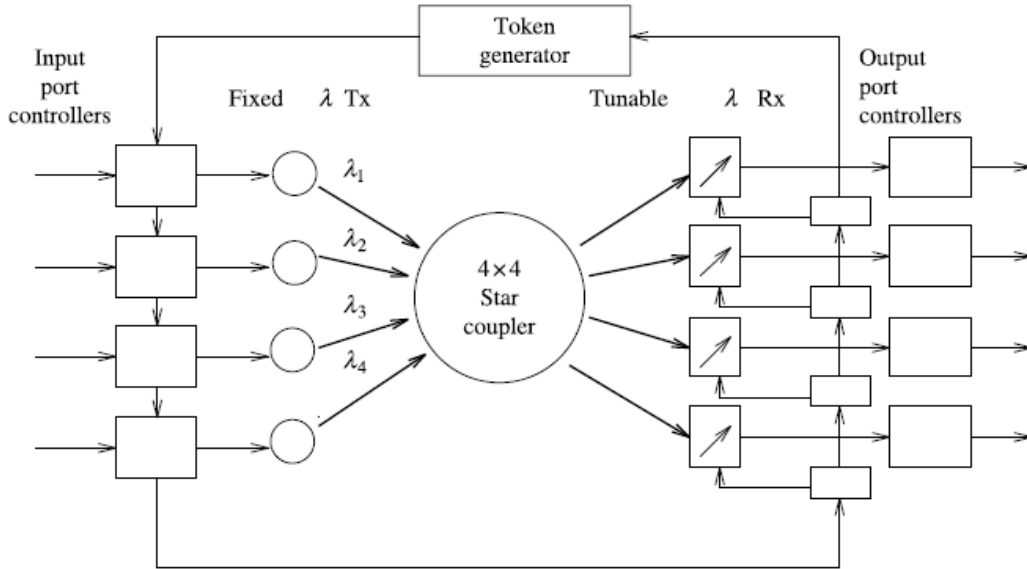


Figure 13 STAR-TRACK implementation

3.3.3 BNR (Bell North Research)

The main components of the switch [47] are input buffer modules, output buffer modules, a high-speed switching core, and a central control unit, as shown in Figure 14. The core switch contains a 16×16 cross-connect network using optical links running at 10 Gbit/s. The central control unit receives requests from input buffer modules and returns grant messages. Each request message indicates the number of queued packets in the input buffer module, which is later used to determine the size of burst allowed to transmit to the switch fabric. A connection can only be made when both input and output ports are free. A control bus is used by the free input ports to broadcast their requests, and by the free output ports to return grant messages.

An arbitration frame consists of 16 packet time slots for a 16×16 core switch. In each slot, the corresponding output port polls all 16 inputs. For example, in time slot 1, output port 1 (if it is idle) will choose the input that has the longest queue destined for output port 1. If the input is busy, another input port that has the second longest queue will be examined. This operation repeats until a free input port is found. If a match is found (free input, free output, and outstanding request), a connection is made for the duration corresponding to the number of packets queued for this connection. So, the switch is a burst switch, not a packet switch. In time slot 2, output port 2 repeats the above operation. The switch capacity is limited by the speed of the central control unit. Packet streams can have a long waiting time in the input buffer modules under a high traffic load.

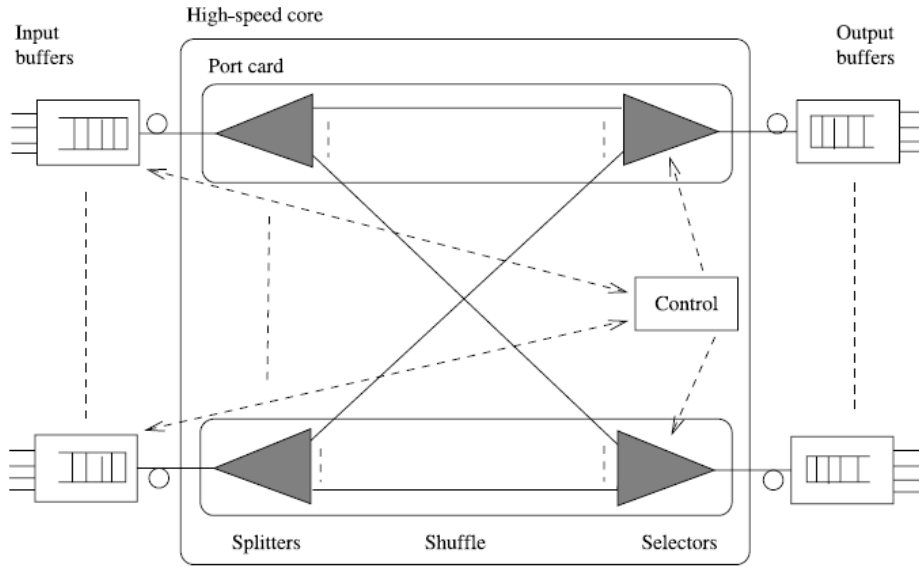


Figure 14 BNR implementation

3.3.4 WAVE-MUX SWITCH (Oki Electric Industry Corporation)

Basically, this switch [48] consists of three kinds of modules: input group module (IGM), switching module (SWM), and output group module (OGM), as shown in Figure 15. They are connected by means of fiber optical lines. The cells arriving through optical lines are first converted to electronic signals by optical-to-electrical (O/E) converters and their header information is electrically processed at the header converter in IGMs. Both the header and payload of the arriving cell are processed and stored in an electronic random access memory (RAM). An optical sorter in each IGM is used to sort the cells with respect to their OGM requests and delivers them to SWM in a cell slot time. In each OGM, it is possible to have the cells with different wavelengths but with the same output port request. This is called output port contention and is solved by the use of an optical buffer. This optical buffer in the OGM is based on the fiber delay line principle. They are distributed to the tunable filters by an optical switch. Each tunable filter is then tuned to the wavelength of the requested output port.

The proposed optical switch architecture needs complex arbitration to solve the contention problem, not only for those cells in same IGM but for the cells in different IGMs as well. This will increase the complexity of control, thus limiting the switch size. There are many O/E and E/O converters required in the switching path, thus increasing implementation costs.

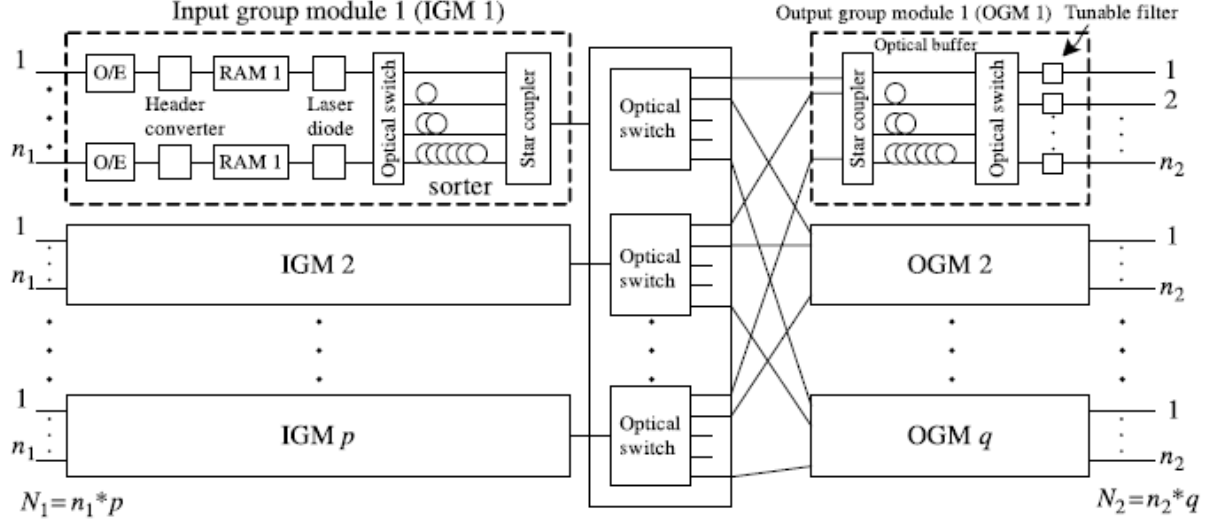


Figure 15 WAVE-MUX implementation

3.3.5 VLSI PRM (Washington University)

At the core of the proposed switch fabric [49] are two dimensional arrays of lasers and detectors. The lasers are Vertical Cavity Surface Emitting Lasers (VCSELs) which emit light perpendicular to the surface of the chip. Chip-to-chip optical interconnections can then be supported through the use of flexible fiber image guides. Figure 16 illustrates the allocation of lasers and detectors for a four channel system utilizing 32x32 arrays of optical elements. If the individual element communicates at 1 Gbit/s, this yields $(32 \times 32)/4 = 256$ Gbit/s per channel. As illustrated in Figure 16 for a node multiring, one subring is associated with each of the ports (or destinations).

The subrings within the multiring share the bandwidth of the entire system where the total bandwidth is determined by the number of laser-detector pairs available and the bandwidth associated with each pair. Bandwidth sharing is done by allocating a portion of the laser-detector pairs to each subring. Within each subring only one source can be active a time. The Deficit Round Robin scheduling algorithm has been selected to arbitrate across competing sources for access to each channel or subring.

The proposed optical switch needs easy arbitration, but it is hardly scalable for the nature of the architecture, which is a ring.

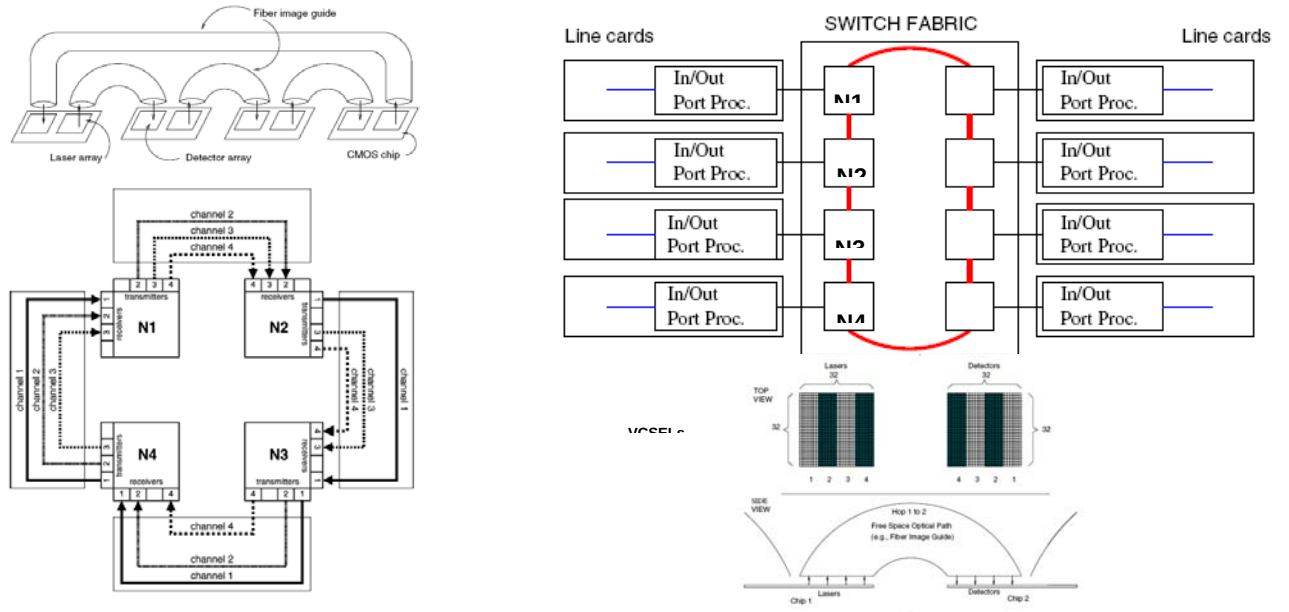


Figure 16 VLSI PRM implementation

3.3.6 BROADCAST-AND-SELECT OPTICAL INTERCONNECT (e-photon/ONe)

One of the most commonly applied switching principles is represented by the broadcast-and-select principle, which is based on the employment of splitters which broadcast the optical signal that is then selected on the basis of the destination address. Different examples of such architecture have been proposed in literature and will be described in the following.

An example of interconnection network based on the broadcast-and-select principle is here given. It was originally proposed in the framework of the e-photon/ONe project and implements the so-called shared-per-wavelength principle. The architecture is hybrid as results by the composition of optical and electronic subsystems. It relies on available optical devices (MUX/DEMUXes, couplers/splitters, optical gates etc...) and electronic buffers. A similar architecture was first presented in [50]. That architecture is re-organized in Figure 17 to better outline its modularity and subsystems which perform contention resolution. The general scheme can be exploited in different scenarios, ranging from optical circuit switching (OCS) to optical burst/packet switching (OBS/OPS). The switch exploits wavelength and time domains to solve contention, through wavelength converters (WCs) and electronic buffers (EBs), respectively, which are shared among the input fibers. The WCs are organized in r_w blocks, and the EBs are organized in B line cards of M first-in-first-out (FIFO) queues each (one per wavelength). The resulting architecture, called H-EOS (Hybrid-Electro/Optical Switch) is shown in Figure 17, as in the case $N=2$ input fibers (IFs)/output fibers (OFs) carrying $M=4$ wavelengths each, $r_w=1$ WC block and $B=1$ EB block.

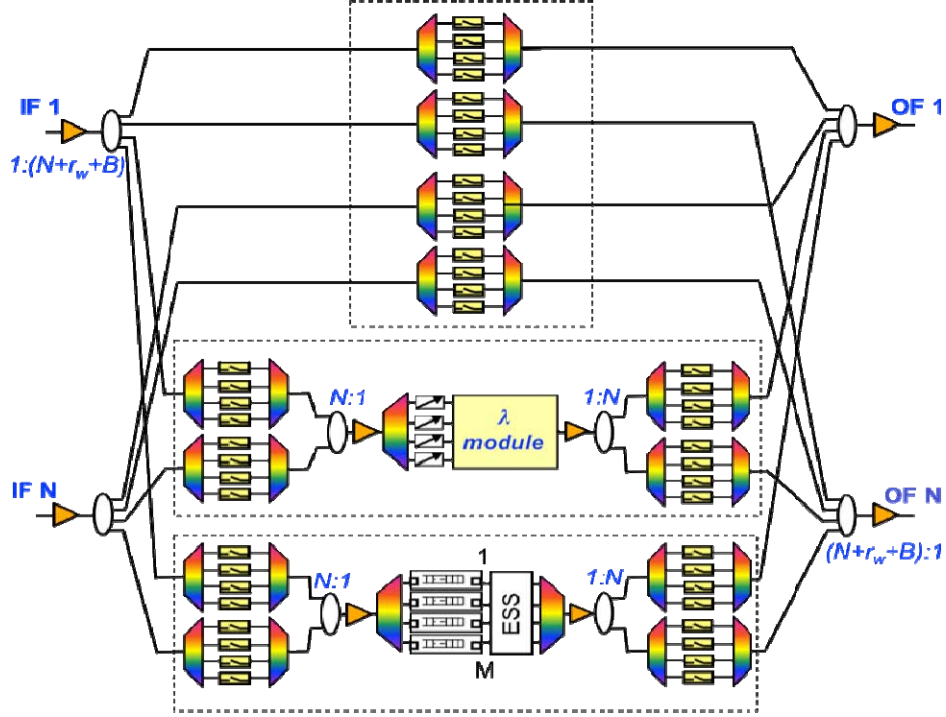


Figure 17 Broadcast-and-select hybrid interconnection.

As can be seen in the figure, a packet can be directly sent to the OFs through the switching fabric, otherwise it can exploit either WCs or EBs. Neither queuing after wavelength conversion nor optical wavelength conversion (through WC blocks) after queuing is allowed. In fact, signals from a WC or EB block are directly sent to the destination OFs. The switching fabric is implemented by means of an array of N^2 Wavelength Selectors (WSs), each consisting of two grating Demux/Mux (or any devices with equivalent functionality) in tandem separated by an array of M optical devices (each dedicated to one wavelength) which are able to operate as ON/OFF gates. This kind of switching stage has been extensively considered in literature and is reported for example in [51]. In this work Semiconductor Optical Amplifiers (SOAs) are considered as optical gates given that they provide high extinction ratio, switching time in the order of few nanoseconds and quite mature technology. This switching stage is wavelength blocking (if two or more packets coming on the same wavelength are directed to the same OF at the same time, one of them is forwarded, the others are blocked), due to wavelength contention at the couplers on each OF. To solve the wavelength blocking of the switching fabric previously described, r_w shared WC blocks are provided. Each WC block is equipped with M Tunable WCs (TWCs), each dedicated to a different wavelength. These WCs should be all-optical, to assure optical transparency, but, due to immaturity of these devices, in a first phase they might be electro-optical devices. The concepts presented in this work are valid for both options, and, if no bit-a-bit operations and 3R regenerations are performed, the optical transparency is maintained also in the latter case. A λ -module combines the M signals in a single WDM multiplex. Each TWC block is equipped with Erbium Doped Fiber Amplifiers (EDFAs) at its ingress and egress, to amplify the WDM signal. A more detailed description of this TWC block and λ -module, together with an analysis of the impact of this particular TWC organization on the scheduling algorithm and performance, can be found in [52]. The TWCs in a block are partitioned among the M wavelengths so that r_w TWCs (placed in different blocks) are shared among the packets coming on the same wavelength. The switch applies the shared-per-wavelength strategy, presented in details in [53]. A packet blocked due to wavelength contention on the destination OF can be forwarded to another free wavelength on the same OF by exploiting one of the

TWCs dedicated to its wavelength. Due to this particular organization, fixed-input/tunable output WCs can be employed instead of the tunable-input/tunable-output ones, that are supposed to be more complex. To connect the IFs to the TWC blocks, the WDM signal coming on an IF is split into further r_w copies, each one connected to a different TWC block by means of a WS. Similarly, to connect the TWC blocks to the OFs, N copies of the signal outgoing from a TWC block are connected to different OFs through WSs. So, $2Nr_w$ additional WSs are needed to connect the TWC blocks to the IFs/OFs. Wavelength conversion reduces the effect of the wavelength blocking, but it is not enough to completely avoid packet blocking due to either output blocking (i.e. lack of free wavelengths on the destination OFs) or lack of TWCs to perform wavelength conversion. For this reason, B EB blocks are available to store blocked packets and solve contention in the time domain. Similarly to TWC blocks, EB blocks are shared among the IFs. Different technological solutions for an EB (line card) can be considered. Demultiplexing, opto/electronic (O/E) conversion, buffering and electronic synchronization stages are in any case performed. The EB blocks are equipped with EDFAs at the ingress and egress, as for TWC blocks. Each EB block is equipped with M electronic queues, one per wavelength. First, the WDM signal entering a block is split by means of a Demux, and the M signals on different wavelengths are interfaced with the M queues. In each of them, a packet is received on the corresponding wavelength and O/E converted. Then it is stored in a simple FIFO electronic queue, with L packet rooms. At the transmission side, an Electronic Slot Synchronizer (ESS) synchronizes the packets to be forwarded in the current time slot, and then they are converted to the optical domain. Queued packets are converted to the optical domain through either Fixed Transmitters (FTs) or Tunable Transmitters (TTs). In the latter case, better performance can be obtained, at expenses of some additional hardware complexity [54].

Contention resolution in the proposed hybrid architecture was discussed in different works for undifferentiated traffic and differentiated (two classes, priority and best effort) traffic. Different scheduling algorithms need to be defined according to the available resources and the requirements of the input traffic. On the other hand, performance in terms of both delay and loss depends on the traffic mix considered and on the dimensioning of the fabric in terms of EBs and WCs. A good trade-off among complexity/cost and performance need to be individuated, taking the available technology into account.

3.4 *Supercomputing machines*

As mentioned before, massively parallel high-performance computing (HPC) in a sense paved the way to OI application also for telecommunication applications.

Optical interconnection is implemented in super-computing systems both at rack-to-rack and board-to-board level and at chip-to-chip level on a board. Chip-to-chip OI is now mainly being pursued using vertically emitting VCSELs + vertically receiving detectors in $<1\mu\text{m}$ wavelength range. However, it is expected that by the required data rate will increase to a level (several 10s of TB) such that wavelength division multiplexing is required to reach the desired data density. ICT VISIT project is dealing with exactly these issues with great success.

Of the several existing example of OI for super-computing, we have chosen to describe here Infiniband and Myrinet systems for their advanced state of development, plus other interesting prototypes.

3.4.1 *INFINIBAND (Infiniband Trade association)*

Infiniband [55] has become rapidly a widely accepted medium for internode networks. The specification was finished in June 2001. From 2002 on a number of vendors has started to offer their products based on the Infiniband standard. Infiniband is employed to connect various system components within a system. The Infiniband fabric can be used for interprocessor networks, attaching I/O subsystems, or to multi-protocol switches like Gbit Ethernet switches, etc. Because of this versatility, the market is not limited just to the interprocessor network segment and so Infiniband is expected to become relatively inexpensive because a higher volume of sellings can be realized. The characteristics of Infiniband are rather nice: there are product definitions both for copper and glass fiber connections, switch and router properties are defined and for high bandwidth multiple connections can be employed. Also the way messages are broken up in packets and reassembled as well as routing, prioritizing, and error handling are all described in the standard. This makes Infiniband independent of a particular technology and it is, because of its completeness, a good basis to implement a communication library on top of it.

Conceptually, Infiniband knows of two types of connectors to the system components, the Host Channel Adapters (HCAs) and Target Channel Adapters (TCAs). The latter are typically used to connect to I/O subsystems while HCAs does more concern us as these are the connectors used in interprocessor communication. Infiniband defines a basic link speed of 2.5 Gb/s (312.5 MB/s) but also a 4× and 12× speed of 1.25 GB/s and 3.75 GB/s, respectively. Also HCAs and TCAs can have multiple ports that are independent and allow for higher reliability and speed.

Messages can be sent from one HCA/TCA to another: a HCA/TCA is permitted to read/write the memory of another HCA/TCA. This enables very fast transfer once permission and a write/read location are given. A port together with its HCA/TCA provide a message with a 128-bit header which is IPv6 compliant and that is used to direct it to its destination via cut-through wormhole routing. In each switching stage the routing to the next stage is decoded and send on. Short messages of 32 B can be embedded in control messages which cuts down on the negotiation time for control messages.

Infiniband switches for HPC are normally offered with 8–288 ports (though Sun is now building a 3456-port switch for its Constellation cluster systems) and presently mostly at a speed of 1.25 GB/s. However, recently the first switches and HCAs accommodating double this speed have become available. Obviously, to take advantage of this speed at least PCI Express must be present at the nodes to which the HCAs are connected. The switches can be configured in any desired topology but in practice a fat tree topology is almost always preferred.

3.4.2 *MYRINET (Myricom)*

Myrinet [56] is a high-speed local area networking system designed by Myricom to be used as an interconnect between multiple machines to form computer clusters. Myrinet has much less protocol overhead than standards such as Ethernet, and therefore provides better throughput, less interference, and less latency while using the host CPU. Although it can be used as a traditional networking system, Myrinet is often used directly by programs that "know" about it, thereby bypassing a call into the operating system.

Myrinet physically consists of two fibre optic cables, upstream and downstream, connected to the host computers with a single connector. Machines are connected via low-overhead routers and switches, as opposed to connecting one machine directly to another. Myrinet includes a number of fault-tolerance features, mostly backed by the switches. These include flow

control, error control, and "heartbeat" monitoring on every link. The newest, "fourth-generation" Myrinet, called Myri-10G, supports a 10 Gbit/s data rate and is inter-operable with 10 Gigabit Ethernet on PHY, the physical layer (cables, connectors, distances, signaling). Myri-10G started shipping at the end of 2006. The latest Myrinet implementation only uses fibers as signal carriers. This gives a high flexibility in the connection and much headroom in the speed of signals but the fiber cables and connectors are rather delicate which can lead to damage when cluster nodes have to be serviced.

Myrinet offers ready-made 8--128 port switches (8--256 for its newest product). The 8 and 16 port switches are full crossbars. In principle all larger networks are build form these using a Clos network topology. An example for a 64-port systems is shown in Figure 18

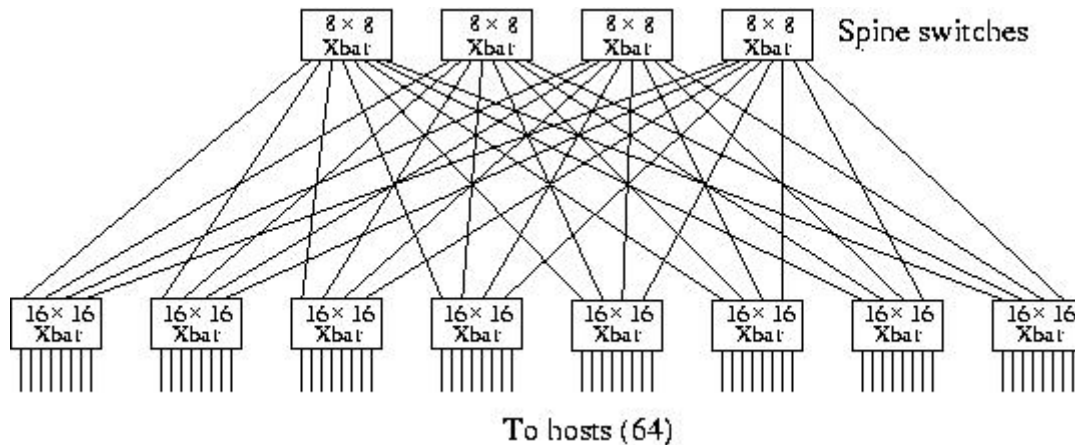


Figure 18 MYRINET implementation

3.4.3 RAPID (University of Arizona)

The RAPID (Reconfigurable and scalable All-Photonic Interconnect for Distributed-shared memory) [57] network is an interconnection network proposed specifically for a large distributed shared memory (DSM) machine. In actuality, the network as proposed is an entire system with a specific interconnect technology (passive/active optics and waveguides on cluster processor boards), cache coherence protocols, power controlled, wavelength assignment schemes, resilience techniques.

RAPID provides fast and efficient unicast, multicast and broadcast capabilities using a combination of aggressively designed wavelength division multiplexing (WDM), time division multiplexing (TDM), and space division multiplexing (SDM) techniques.

Each node in RAPID network, contains the processor and its caches, a portion of the machine's physically distributed main memory, and a node controller (shown as a bus) which manages communication within nodes. Few nodes are connected together to form a group. All nodes are connected to two subnetworks; a scalable intragroup interconnection (IGI) and a scalable intergroup remote interconnection (SIRI). In our remote wavelength assignment scheme shown in Figure 19, all nodes within the source group is assigned a unique wavelength at which it can transmit to communicate with any destination group. RAPID is an OCS techniques.

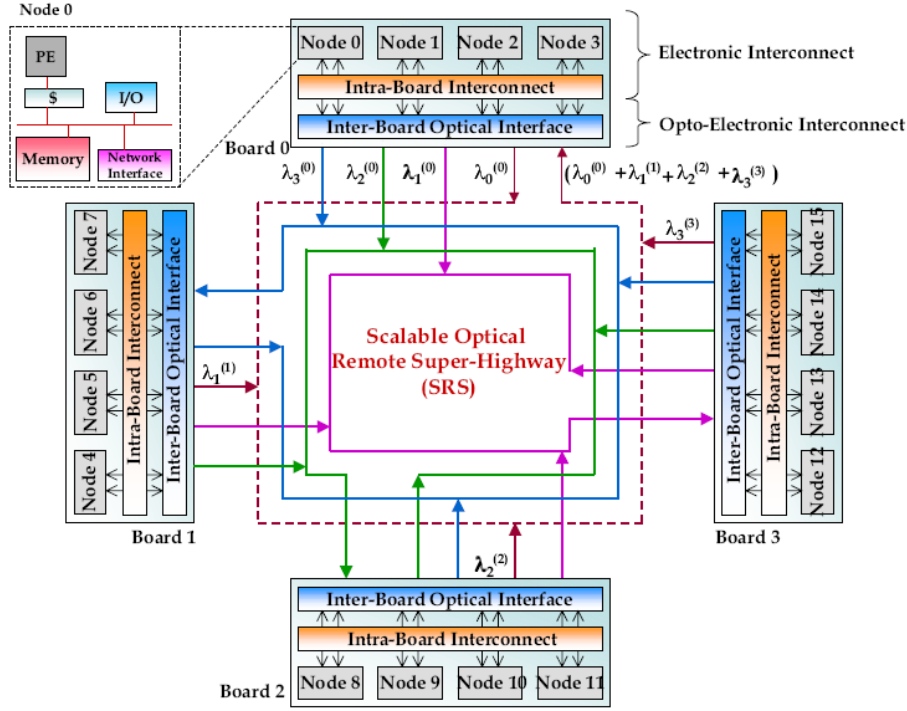


Figure 19 RAPID implementation

3.4.4 DATA VORTEX (IBM Research)

The data vortex [58] is one of the aforementioned photonic networks that are designed to allow always-open deflection paths and thereby avoid the need for optical buffering. It is a highly-scalable photonic packet switching architecture that utilizes self-routing of individual packets and alleviates the need for central scheduling and processing. Deflection routing is used to eliminate internal packet buffering and minimize packet traffic congestion. The data vortex architecture's unique absence of internal optical buffering elements enables the transparent routing of DWDM packet payloads while maintaining flexibility of extending the packet size by simply adding (or removing) wavelength channels. The data vortex optical packet switching network architecture was designed specifically for realization in the optical domain, taking into consideration the difficulty of implementing optical buffering and complex optical logic. Its topology is composed entirely of 2×2 switching elements (also called nodes) arranged in a fully connected, directed graph (Figure 20). The packets are injected at the outer most cylinder from the input, and emerge at the inner most cylinder toward the output. Each packet is selfrouted in the fashion of binary-tree decoding. Without loss of generality, a node A on a generic cylinder has two input ports: one from a node B on the same cylinder and one from a node C on the outer cylinder. A packet passing from B to A causes a control signal to be sent from B to C that blocks data at C from progressing to A. The blocked packet is deflected on its current cylinder level. Header bits are separated from the rest of the packet with conventional fiber optic wavelength filters for detection and conversion to electrical signals which can be used for the routing logic. However, electrical decision circuit to perform header processing may introduce latency.

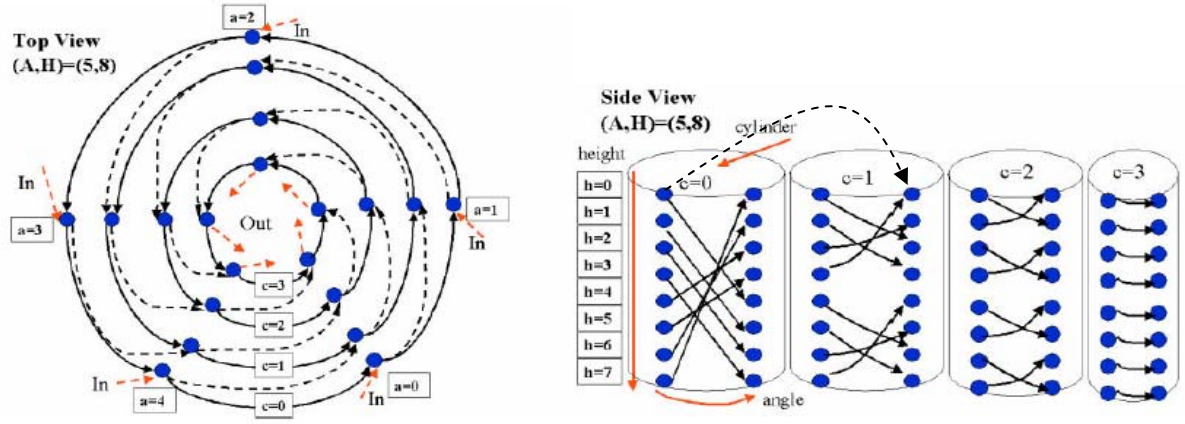


Figure 20 DATA VORTEX implementation

3.4.5 OSMOSIS (IBM Research)

Osmosis [59] is fundamentally different from other optical switches, most of which use techniques such as meshing, circuit switching, provisioning (bandwidth reservation), or aggregation (burst or container switching). These approaches are prohibitively expensive (meshing), are too inefficient (meshing and circuit switching), or introduce too much latency (aggregation). Instead, Osmosis (Figure 21) switches fixed-size cells by reconfiguring Osmosis switches fixed-size cells by reconfiguring The key enabler is the fast switching time (a few nanoseconds) achieved by state-of-the-art semiconductor optical amplifiers (SOAs), which now, for the first time, allow low-overhead optical cell switching (256 byte). SOAs are well suited to implement the optical equivalent of a crossbar using a broadcast-and-select (B&S) data path architecture combining space- and wavelength-division multiplexing. Each optical switching module has fiber and wavelength selection stages.

The fiber selection stage consists of eight SOAs, which select the fiber that carries the cell to be received. A planar light-wave circuit combines the output of all SOAs to a common port, and the resulting stream is then demultiplexed and passed through the wavelength selection stage (also consisting of eight SOAs), which selects the desired wavelength. Synchronously with cell transmission, the centralized scheduler resolves contention on requests it receives via optical control channels and provides control signals that drive the fiber and wavelength selection SOAs just as the cells arrive at each optical switching module. Although SOA technology allows very fast switching operation, a single SOA gate requires high power supply in the order of 100mW. This may reduce the scalability of the overall system. Moreover, also the nature of the B&S data path architecture may lead to an higher power consumption in comparison to other network topologies.

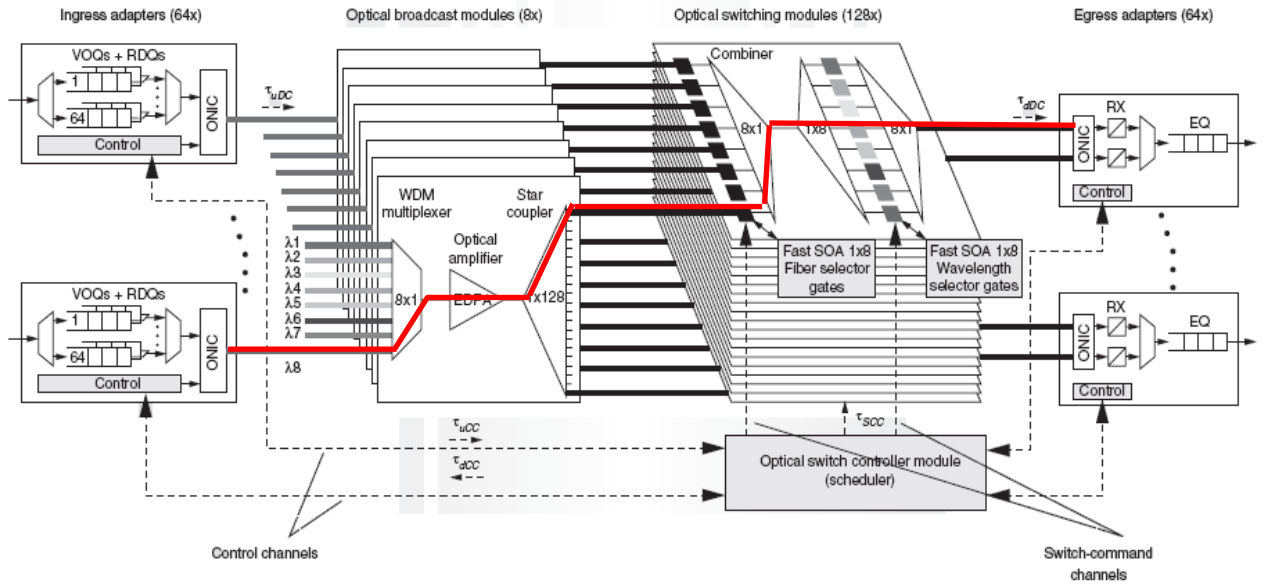


Figure 21 OSMOSIS implementation

3.5 On-chip architectures in multicore applications

There is no commercial deployment of integrated silicon photonics for chip-to-chip (distance < few 10cm) and on-chip communications at the moment. On-chip optical interconnect is not expected before 2018-2022 in commercial applications. However a lot of research effort is currently going on in Europe (ICT WADIMOS project), US (consortia led by IBM, SUN, INTEL, HP) and Japan.

By examining several multicore architectures using optical technologies on-chip, it is possible to identify similar characteristics and trends in their design considerations. Interconnection network topologies (rings, torus, crossbars) are used to transfer data between memories and processors with simple access protocols exploiting wavelength and space diversity. Particularly, the generation of the optical carriers is usually performed by external optical sources, and a network is used to distribute continuous wave (CW) optical signals to the different nodes. The CW optical carriers are used inside the chip by modulating, redirecting or splitting the CW power by means of optical switching structures, such as cross points or resonant modulators.

Modulation, space selection, filtering and even detection operations have found a convergent technical solution in different versions of ring resonator devices, used as add-drops, modulators, channel selectors, and (when rings are properly doped) detectors [60].

Allowing to exploit the physical proximity in different layers (a photonic layer carrying Tbps of data, an electronic layer used to control the photonic one and one or more chip layers for processors and memories), the 3D integration is the technology making these architectures possible optimizing the area occupied by the whole structure.

All architectures shown in the following have been analyzed independently and are assumed to be implemented in technology nodes not currently available, but they constitute a clear step towards replacing electronic interconnects with optical equivalents by strongly stressing the wavelength dimension, an issue that has been pointed out in other WP25 documents. A brief summary of the relevant characteristics of the different architectures is shown in Table 7.

3.5.1 *Manycore (Massachusetts Institute of Technology, University of Carolina)*

The work done in [61] presents a network architecture based on an opto-electronic global crossbar where the photonic links are used to build on-chip and off-chip interconnects.

The architecture is depicted in Figure 22. The system comprises 256 cores, divided in groups of 16 cores each (in Figure 22, the hexadecimal notation indicates the group) and 16 external DRAM modules. All cores have bidirectionally connectivity (using 64 wavelength channels per waveguide) with the external DRAM modules. Each core is connected to one waveguide.

The crossbar is implemented through a waveguide matrix where the switching capability is achieved by means of ring resonant structures. Waveguides are used twice: to transmit i) from cores to memory (or possibly other cores) and ii) from memory to cores. The transmission direction is separated using two different wavelength channel grids; in this way, the waveguide usage is improved (reaching 128 WDM channels) and also memory requests-responses are physically separated to avoid deadlocks. Both cores and memory add or drop optical signals from the waveguides through resonant ring structures. Each modulator operates at 10 Gbps, and each node has access to the waveguide through 16 modulators (4 cores share the same waveguide), achieving a total capacity of 64 bit per core cycle per channel (determined by memory requirements) at a 2.5 Ghz core frequency (since the modulator is 4 times faster than the core). Details about the cores used are not provided, so an estimation in terms of FLOPS cannot be performed accurately. At present, exploiting 65 nm and 45 nm node technologies, several tests around the optical layer, such as detectors, waveguides and cascaded resonant add/drop devices are being reported (note that the architecture is targeted for a 22nm technology nodes).

Simulation tests of 256 cores and 16 external memory show near 8-10 times higher throughput with respect to the electronic equivalent when the energy constraints were maintained constant. This architecture was motivated by the memory bandwidth challenge, derived from the emerging multicore architectures. According to the authors, this would be the future bottleneck of these kind of systems. Moreover, an order of magnitude in the energy-efficiency is predicted respect to the electrical links of the same target technology node. Several power efficiency analyses at this scale are cited to assess power performance.

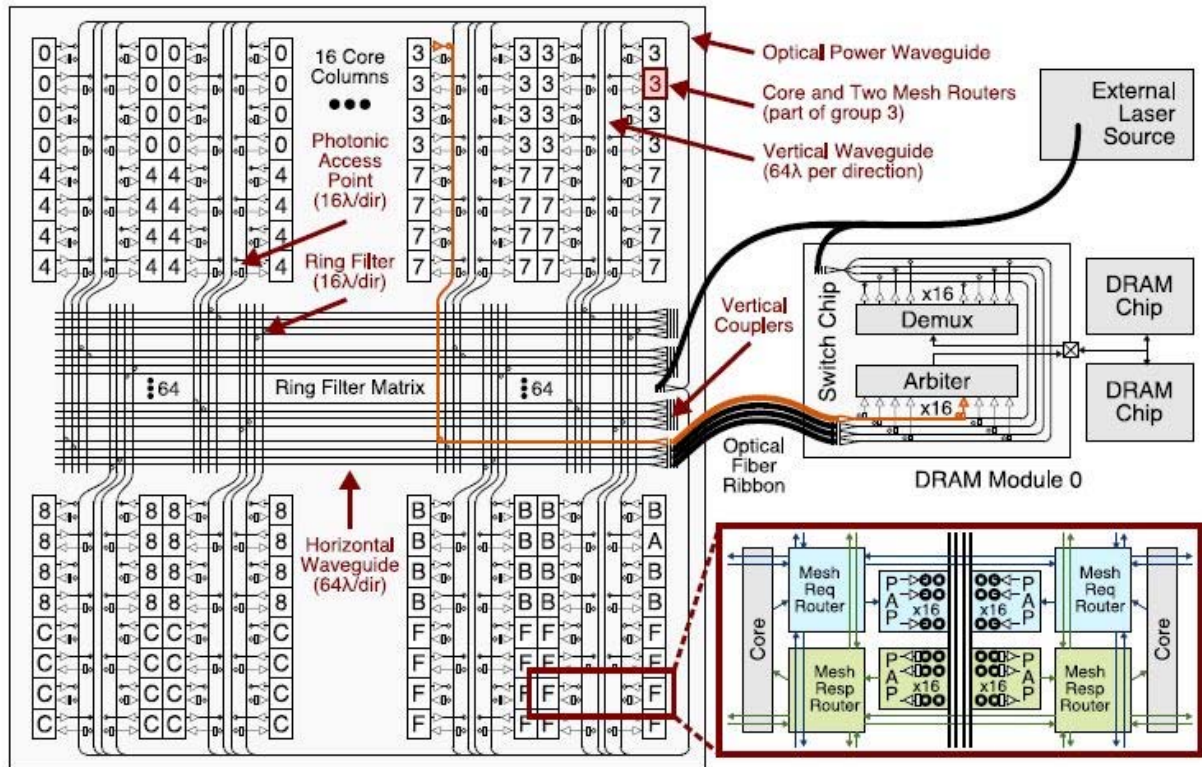


Figure 22 Manycore architecture

3.5.2 CORONA (University of Wisconsin, Hewlett Packard, University of Utah)

CORONA [62] is a 3D based system targeted for 16 nm technology (predicted to be available in 2017), and it is depicted in Figure 23. CORONA is intended to support 256 general purpose cores, grouped in 64 clusters interconnected by a waveguide bundle visible in blue in Figure 23. Clusters are located just above waveguide bundles and are represented by squares and each cluster contains 4 cores, as it is depicted in the top left corner diagram.

The bundle is composed by 64 data waveguides (64 wavelength channels each), plus one for arbitration signalling, one more for clock distribution, and another one for broadcasting. An additional waveguide carries the CW optical signals, which will be used by all nodes' transmissions. Each data waveguide is associated with a cluster (called home cluster for the waveguide), where the ring is interrupted forming an open loop. The home cluster receives on the corresponding waveguide after the waveguide has crossed through all the other clusters, according to a multiple write - single read scheme. Therefore the bundle is a logical crossbar in the sense that there is always a path from a free input to a free output.

All transmissions are controlled by an arbitration scheme, based on the presence or absence of optical signals in the arbitration waveguide, which carries 64 wavelengths, each one associated to a home cluster. A cluster wanting to transmit to a given destination modulates all the wavelengths on the destination cluster's waveguide and turns on the wavelength associated to the destination cluster on the arbitration waveguide. In this scheme, signal retiming is not required, thanks to the clock distribution waveguide, which propagates in parallel to the data waveguides, so the signal's synchronization is recovered directly from the clock.

Another set of waveguides (the green ones in Figure 23) are also used as I/O for optically connected memories; each cluster can access external memories by using a single waveguide (for in-chip memory) or an output fiber (for off-chip memories).

CORONA presents a cluster-to-cluster bandwidth of 2.56 Tbit/s and a total aggregate of 20 TBytes/s; the architecture was compared through simulations to electronically interconnected memories, showing 2-6 times performance gain.

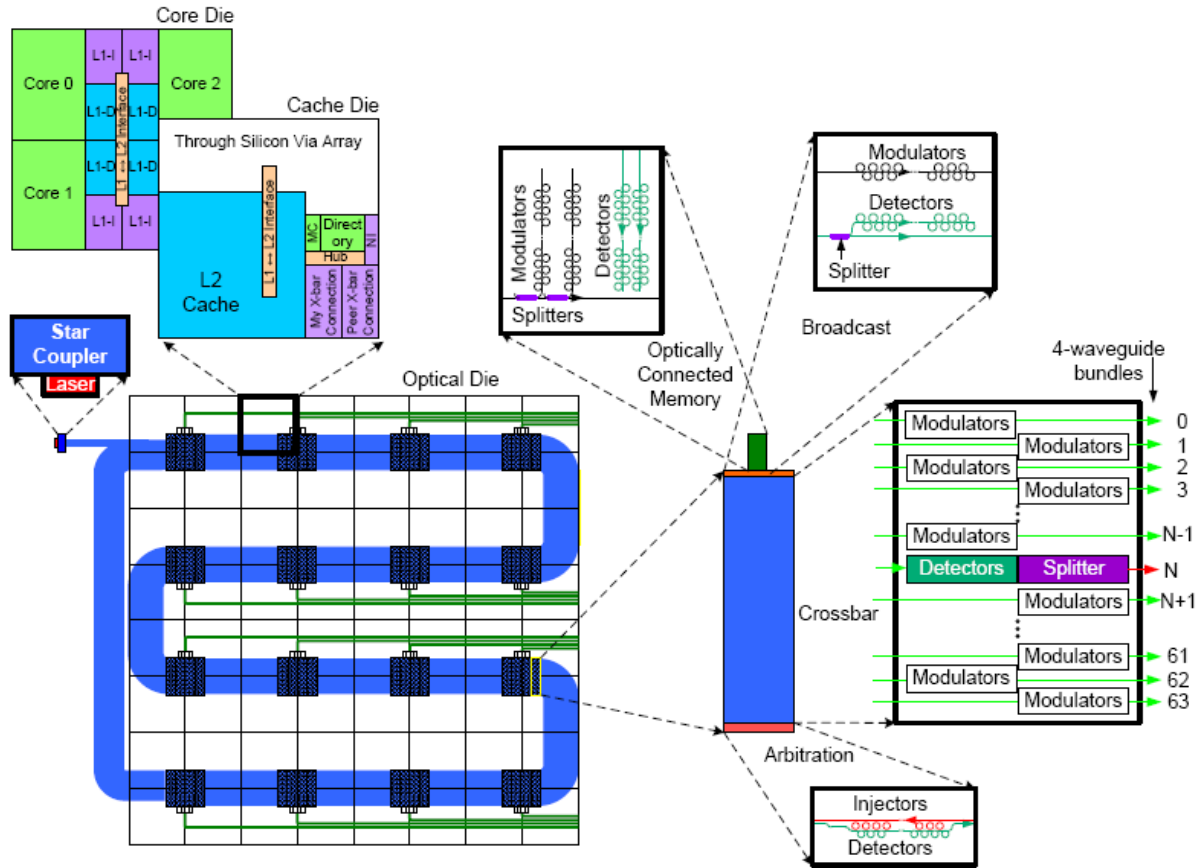


Figure 23 Corona architecture

3.5.3 Photonic NoC (Columbia University)

The work done in [63] combines photonic communication with electronic control to provide an On-chip interconnect solution between independent cores. Based on a hybrid approach, it uses a high-bandwidth circuit-switched optical network (for actual data transfer) and a low-bandwidth packet-switched electronic network (for control purposes). This approach of transferring large amounts of data in a circuit switched paradigm, allows dealing with the lack of optical memories while offering large data messages between core pairs in order to maximize the network usage.

The basic steps of the NoC operation are: first, at the path-setup process a source core reserves the path with the destination core; second, the data from the source core (highlighted as black circles in Figure 24) is transmitted by the photonic circuit (using TDM-WDM resources); finally at the path teardown process, the source finishes the single communication. The basic switching elements are based on 4x4 non-blocking switches built with microring resonators, distributed on an arbitrary grid. It is important to mention that the network

gateways (squares near nodes in Figure 24) include the circuitry necessary for clock synchronization, recovery and serialization/deserialization.

Three different networks are presented (Figure 24(a): blocking torus, Figure 24(b): crossbar and Figure 24(c): non blocking torus). All of them are based on the non blocking 4x4 switch (also proposed by Columbia Univ., and shown in the upper right square in Figure 24(a)) and are shown for the 16-core case.

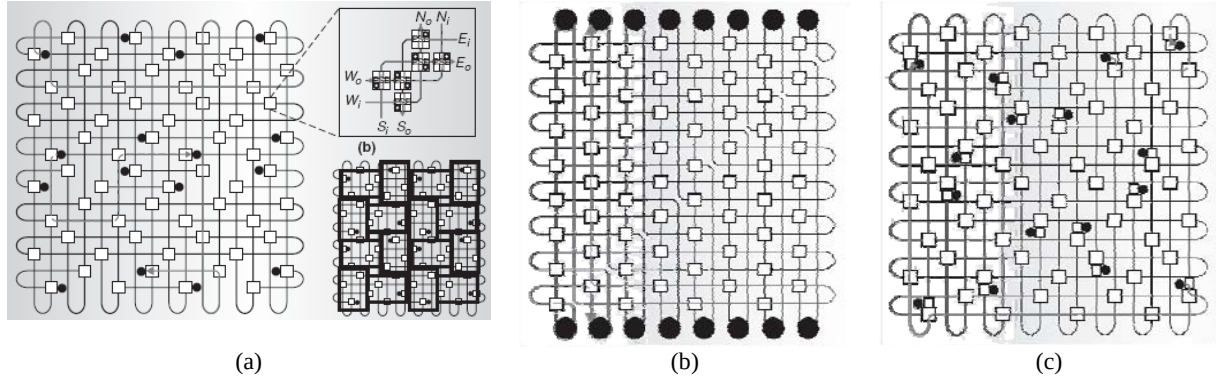


Figure 24 (a) Blocking torus NoC; (b) Crossbar NoC; (c) Non blocking torus NoC

These three different topologies were compared using simulations under different traffic patterns. Under uniform traffic, it was obtained a remarkably similar throughput result for the blocking torus and crossbar networks. This fact is justified with the long electronic paths used by the crossbar, which increase the propagation time of the path-setup packet over the control network. The non blocking torus exhibits a gain about the 13% over blocking torus and crossbar. It is important to mention that the latency results do not diverge considerably between the different network topologies.

For the throughput expected under non uniform traffic, the non blocking torus, together with the crossbar (with very similar performance), offer the best amount of delivered traffic, reaching a gain slightly over 50% larger if compared with the blocking torus topology for a certain specific traffic pattern.

Finally, a simulation case is reported to study the execution of a large FFT (Fast Fourier Transform) in order to compare several performance figures, such as number of parallel operations executed (FLOPS), times required in finishing tasks, and power consumption. For example, for the same execution time, optical interconnects requires ten times less power than an equivalent electrical interconnect.

3.5.4 HyperMesh (Shanghai Jiao Tong University)

Recently presented in [62], this architecture is based on the principle of using global optical interconnections (for long on-chip distances) and local electronic interconnections between near processing cores; this aims at reducing the impact of the power and delay required for OE/EO conversions. It houses 256 cores grouped in clusters with 4 cores (see example in Figure 25) connected by an electronic switch, distributed in a 2D photonic network (Called HyperMesh). Each column and each row of the optical interconnect is served by a set of waveguides, which provide a crossbar-like interconnection among electronic switches. Communication between switches belonging to the same row or to the same column is single-hop, while routing through an intermediate electronic switch is required otherwise.

Row/column waveguides are folded in open loops similarly to the Corona architecture, and switches communicate according to a Multiple-Write-Single-Read scheme. On each cluster, add and drop operations are performed with ring resonant structures, using a total of 34K rings.

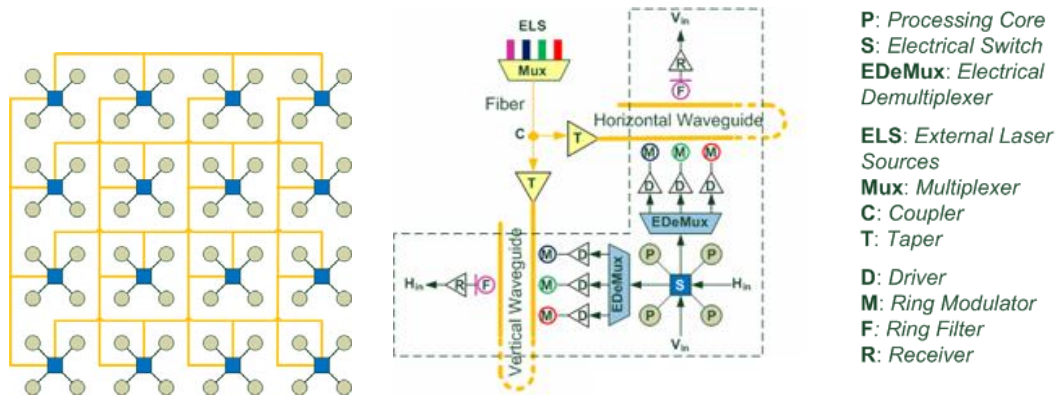


Figure 25 HyperMesh architecture: topology (left) and cluster architecture (right)

3.5.5 Summary of on-chip architectures

Table 7 Summary of on-chip architectures

	P-NoC blocking torus	P-NoC crossbar	P-NoC non blocking torus	Corona	HyperMesh	Manycore
Architecture	Torus	Crossbar	Torus	Mesh / Mesh with crossbar	Mesh/ rossbar	Crossbar ring
Target technology node (nm)	22	22	22	16	----	22
WDM channels	24	24	24	64	----	64 bidirectional (128)
Processors	36	36	36	256	256	256
Grouped in clusters	No	No	No	Yes	Yes	No
3D integration	Yes	Yes	Yes	Yes	Yes	No
Clock freq (GHz)	----	----	----	5	----	2.5
Clock distribution	Optical	Optical	Optical	Optical	----	Optical
Agregate bitrate core-to-core	0.96 TB/s	0.96 TB/s	0.96 TB/s	20 TB/s	----	----
Agregate bitrate core-to-memory	----	----	----	10 TB/s	----	22 TB/s
TFLOPS	0.00812	----	0.00918	10	----	----
Estimated power (W)	24.5	24.5	24.5	40	0.7 (only optical)	20
Estimated Area (mm ²)	625	625	625	423 / 491	----	400
μrings	2592	2592	2880	1056 K	34K	9 K double rings

4. References

- [1] <http://www.lightreading.com/default.asp>
- [2] N. McKeown, 'A fast switched backplane for a gigabit switched router,' *Business Communications Review*, Vol. 27, No. 12, Dec. 1997.
- [3] Cisco CRS-1 Carrier Routing System 16-Slot Line Card Chassis System Description, CISCO, Mar. 2008.
- [4] P.E. Green, *Fiber Optic Communication Networks*, Prentice-Hall, Upper Saddle River, New Jersey, 1992.
- [5] N. V. Srinivasan, "Add-drop multiplexers and cross-connects for multiwavelength optical networking," *OFC'98*, San Jose, California, Feb. 1998, pp. 57–58.
- [6] C. K. Chan, F. Tong, L. K. Chen, and K. W. Cheung, "Demonstration of an add-drop network node with time slot access for high-speed WDMA dual bus/ring packet networks," *OFC'98*, San Jose, California, Feb. 1998, pp. 62–64.
- [7] G. Chang, G. Ellinas, J. K. Gamelin, M. Z. Iqbal, and C. A. Brackett, "Multiwavelength reconfigurable WDM/ATM/SONET network testbed," *IEEE Journal of Lightwave Technology*, Vol. 14, No. 6, June 1996, pp. 1320–1340.
- [8] R. E. Wanger, R. C. Alferness, A. A. M. Saleh, and M. S. Goodman, "MONET: Multiwavelength optical networking," *IEEE/OSA Journal of Lightwave Technology*, Vol. 14, No. 6, June 1996, pp. 1349–1355.
- [9] S. Okamoto and K. Sato, "Optical path cross-connect systems for photonic transport networks," *IEEE Global Telecommun. Conf.*, Houston, Texas, Nov. 1993, pp. 474–480.
- [10] A. K. Srivastava, J. L. Zyskind, Y. Sun, J. W. Sulhoff, C. Wolf, M. Zirngibl, R. Monnard, A. R. Charplyvy, A. A. Abramov, R. P. Espindola, T. A. Strasse, J. R. Pedrazzani, A. M. Vengsarkar, J. Zhou, and D. A. Ferrand, "1 Tbit/s transmission of 100 WDM 10 Gbit/s channels over 400 km of True Wave fiber," *OFC'98*, San Jose, California, Feb. 1998, Postdeadline Papers, pp. PD10-1–PD10-4.
- [11] I. Chlamtac, A. Ganz, and G. Karmi, "Lightpath communications: an approach to high bandwidth optical WAN's", *IEEE Transaction on Communications*, Vol. 40, No. 7, July 1992, pp. 1171–1182.
- [12] E-photon paper-submitted.
- [13] C. Matrakidis, A. Pattavina, S. Sygletos, N. Vogiatzis, F. Neri, A. Stavdas, "New approaches in optical switching in the network of excellence e-photon/One", *Optical Network Design and Modeling 2005*, Feb. 2005, pp. 133 – 139.
- [14] D. Slepian, "Two theorems on a particular crossbar switching network", Unpublished manuscript, 1952.
- [15] A.M. Duguid, "Structural properties of switching networks", Brown Univ. Progr. Rept., BTL-7, 1959.
- [16] D. Hunter et al, "WASPNET: A Wavelength Switched Packet Network," *IEEE Communications Magazine*, Vol. 37, No. 3, Mar. 1999, pp. 120–129.
- [17] G. R. Hill et al.: "A transport network layer based on optical network elements", *IEEE Journal of Lightwave Technology*, Vol. 11, No. 5/6, May/June 1993, pp. 667–679.
- [18] E. Iannone, R. Sabella: "Optical path technologies: A comparison among different crossconnect architectures", *IEEE Journal of Lightwave Technology*, Vol. 14, No. 10, 1996, pp. 2184–2196.
- [19] L. Thylen, et al, "Computer analysis and design of Ti:LiNbO₃ integrated optics devices and comparison with experiments", *ECOC 1983*, pp. 425–428.
- [20] G. Hill et al, "A transport network layer based on optical network elements", *IEEE Journal of Lightwave Technology*, Vol. 11, No. 5, 1993, pp. 667–679.

- [21] E. Hall et al. "The Rainbow-II gigabit optical network", *IEEE Journal on Selected Areas in Communications*, Vol. 14, No. 5, June 1996, pp. 814-823.
- [22] N.P. Caponio et al., "Single-layer optical platform based on WDM/TDM multiple access for large-scale "switchless" networks", *European Transactions on Telecommunications*, Vol. 11, No. 1, Jan.-Feb. 2000, pp. 73-82.
- [23] D. Hunter et al. "WASPNET: A wavelength switched packet network", *IEEE Communications Magazine*, Vol. 37, No. 3, Mar. 1999, pp. 120-129.
- [24] L. Dittman et al "The European IST Project 7: a Viable Approach towards Optical Packet Switching", *IEEE Journal on Selected Areas in Communications*, Vol. 21, No. 7, 2003, pp. 1026-1040.
- [25] <http://www.ict-boom.eu/objectives.html>.
- [26] http://www.ict-boom.eu/Presentations/NetworksandOptical.Communications2008/Development_of_Advanced_Hybrid_Integrated_Photonic_Devices.pdf
- [27] <http://www.ict-historic.eu/index.php?id=16>.
- [28] T. El-Bawab and J.-D. Shin, "Optical Packet Switching in core networks: between vision and reality", *IEEE Communications Magazine*, Vol. 40, No. 9, Sep. 2002, pp. 69-65.
- [29] P.C. Teh et al, "Multi-wavelength (40 WDM X 10 Gbit/s) optical packet router based on superstructure fibre Bragg gratings", *IEICE Transaction: Special Issue on Recent Progress in Optoelect. and Comm.*, Vol.E86B/5, 2003, pp.1487-1492.
- [30] S Yu, et al, "Demonstration of high-speed optical packet routing using vertical coupler crosspoint space switch array", *Electron. Letters.*, Vol. 36, No. 6, 2000, pp. 556-558.
- [31] Y. Yamada et al., "Optical output buffered ATM switch prototype based on FRONTIERNET architecture", *IEEE Journal on Selected Areas in Communications*, Vol. 16, No. 7, 1998, pp. 1298-1308.
- [32] N. Wada, H. Harai, W. Chujo, and F. Kubota, "80G bit/s variable rate photonic packet routing based on multi-wavelength label switch", *ECOC 2002*, Vol. 3, no. We-B-2-3, pp. 308-309.
- [33] N. Wada, W. Chujo, and K. Kitayama., "1.28 Tbit/s (160 Gbit/s x 8 wavelengths) throughput variable length packet switching using optical code based label switch", *ECOC 2001*, Vol.6/ PD-A-1-9, pp. 62-63.
- [34] N. Wada, R. Takemori, N. Kataoka, F. Kubota and K. Kitayama, "200G-Chip/s, 128-Chip Hierarchical Optical BPSK Labels Processing and its Networking Application", *ECOC 2003*, Vol. 2/Tu4.4.2, pp. 304-305.
- [35] N. Wada, H. Harai and F. Kubota, "40Gbit/s, multi-hop optical packet routing using optical code label processing based packet switch prototype", *OFC 2004*, paper FO7, pp. FO-62-FO-64.
- [36] T. Nakahara, et al, "Optical single-clock-pulse generator using a photoconductive sample-and-hold circuit for processing ultrafast asynchronous optical packets," *IEEE Photonic Technology Letters*, Vol. 14, No. 11, 2002, pp.1623 – 1625.
- [37] N. Wada, G. Cincotti, S. Yoshima, N. Kataoka, and K. I. Kitayama, "Characterization of a full encoder/decoder in the AWG configuration for code-based photonic routers—Part II: Experiments and applications", *IEEE Journal of Lightwave Technology.*, Vol. 24, No. 1, Jan. 2006, pp. 113–121.
- [38] B. Yoo "Optical Packet and Burst Switching Technologies for the Future Photonic Internet", *IEEE Journal of Lightwave Technology*, Vol. 24, No. 12, Dec. 2006, pp. 4468-4492.
- [39] Wei Wang, L. Rau and D. J. Blumenthal, "All-Optical Label Switching/Swapping of 160 Gbps Variable Length Packets and 10 Gbps Labels using a WDM Raman

- Enhanced-XPM Fibre Wavelength Converter with Unicast/Multicast Operation”, *OFC 2004*, Post Deadline, paper PDP8.
- [40] S. Rangarajan et al, “All-optical contention resolution with wavelength conversion for asynchronous variable-length 40 Gb/s optical packets”, *IEEE Photonic Technology Letters*, Vol. 16, No. 2, 2004, pp.689 – 691.
 - [41] P. Zhong et al, “Demonstration of variable-length packet contention resolution and packet forwarding in an optical-label switching router”, *IEEE Photonic Technology Letters*, Vol. 16, No. 7, 2004, pp.1772 – 1774.
 - [42] S. J. Ben Yoo, “ High-Performance Optical-Label Switching Packet Routers and Smart Edge Routers for the Next-Generation Internet”, *IEEE Journal on Selected Areas in Communications*, Vol. 21, No. 7, 2003, pp. 1041-1054.
 - [43] D. Wonglumsom, et al, “Experimental demonstration of an access point for HORNET-A packet over WDM multiple-access MAN”, *IEEE Journal of Lightwave Technology*, Vol. 18, No. 12, Dec. 2000, pp. 1709-1717.
 - [44] C. Qiao and M. Yoo, “Optical Burst Switching (OBS) – a new paradigm for an optical Internet”, *Journal on High Speed Networks*, Vol. 8, No. 1, Jan. 1999, pp. 69-84.
 - [45] E. Arthurs, M. S. Goodman, H. Kobrinski, and M. P. Vecchi, “HYPASS: an optoelectronic hybrid packet switching system,” *IEEE Journal on Selected Areas in Communications*, Vol. 6, No. 9, Dec. 1988, pp. 1500–1510.
 - [46] T. T. Lee, M. S. Goodman, and E. Arthurs, “STAR-TRACK: a broadband optical multicast switch,” Bellcore Technical Memorandum Abstract, 1989.
 - [47] E. Munter, L. Parker, and P. Kirkby, “A high-capacity ATM switch based on advanced electronic and optical technologies”, *IEEE Communications Magazine*, Vol. 33, No. 11, Nov. 1995, pp. 64–71.
 - [48] Y. Nakahira, H. Inoue, and Y. Shiraishi, “Evaluation of photonic ATM switch architectureproposal of a new switch architecture,” *International Switching Symposium*, Berlin, Germany, 1995, pp. 128–132.
 - [49] P. Krishnamurthy, M. Franklin and R. Chamberlain, “Dynamic Reconfiguration of an Optical Interconnect”, 36th Annual Symposium on Simulation, Orlando, FL, 2003.
 - [50] C. Raffaelli, M. Savi, "QoS Aware Optical Packet Switch with Shared Electronic Buffers", *3rd International Workshop in Guaranteed Optical Service Provisioning (GOSP 2008)*, London, UK, Sep. 2008.
 - [51] A. Stavdas, C. Matrakidis, C. Politi, "Migration of Broadcast-and-Select Optical Crossconnects from Semi-Static to Dynamic Reconfiguration and their Physical Layer Modeling", *Optics Communications*, Vol. 280, No. 1, Dec. 2007, pp. 49-57.
 - [52] Raffaelli, M. Savi, A. Stavdas, "Sharing Wavelength Converters in Multistage Optical Packet Switches", *IEEE Workshop on High Performance Switching and Routing (HPSR) 2006*, Poznan, Poland, June 2006.
 - [53] C. Raffaelli, M. Savi, A. Stavdas, “Multistage Shared-per-Wavelength Optical Packet Switch: Heuristic Scheduling Algorithm and Performance”, *IEEE Journal of Lightwave Technology*, Vol. 27, No. 5, March 1, 2009, pp. 538 – 551.
 - [54] C. Raffaelli, M.Savi, “Hybrid Contention Resolution in Optical Switching Fabric with QoS traffic”, *WOBS 2009*, Madrid, Spain.
 - [55] <http://www.infiniband.org/>
 - [56] <http://www.myricom.com/myrinet/overview/>
 - [57] Avinash Karanth Kodi and Ahmed Louri, Department of Electrical and Computer Engineering, <http://oucsace.cs.ohiou.edu/~avinashk/pubs.htm>
 - [58] C. Hawkins, B.A. Small, D.S. Wills, K. Bergman, “The Data Vortex, an All Optical Path Multicomputer Interconnection Network”, *IEEE Transactions on Parallel and distributed system*, 2007.

- [59] C. Minkenberg, F. Abel, P. Muller, R. Krishnamurthy, M. Gusat, P. Dill, I. Iliadis, R. Luijten, B.R. Hemenway, R. Grzybowski, E. Schiattarella, “Designing a Crossbar Scheduler for HPC Applications”, *IEEE Micro*, Vol. 26, No. 3, May-June 2006, pp. 58-71.
- [60] M. Petracca, B. G. Lee, K. Bergman, and L. P. Carloni, “Photonic NoCs: System-level Design Exploration”, *IEEE Micro*, Vol.29, No. 4, July-Aug. 2009, pp.74–85.
- [61] C. Batten, A. Joshi, J. Orcutt, A. Khilo, B. Moss, C. Holzwarth, M. Popovic, H. Li, H. Smith, J. Hoyt, F. Kärtner, R. Ram, V. Stojanovic, and K. Asanovic, “Building Manycore Processor-to-DRAM Networks with Monolithic Silicon Photonics”, *Hot Interconnects (HOTI 08)*, Aug. 2008, pp. 21-30.
- [62] Y. Jin, Z. Chang, J. Tang, W. Hu, “Energy-Efficient On-Chip Photonic HyperMesh Network for Multi-Processor Communication”, *4th International Workshop on OPS & OCDMA, (IWOOC 2009)*, Tokyo, Japan, Nov. 2009.
- [63] D. Vantrease, R. Schreiber, M. Monchiero, M. McLaren, N. P. Jouppi, M. Fiorentino, A. Davis, N. Binkert, R. G. Beausoleil, and J. H. Ahn, “Corona: System Implications of Emerging Nanophotonic Technology”, *International Symposium on Computer Architecture (ISCA 08)*, Beijing, China, June 2008, pp. 153-164.