Francesco Testa · Lorenzo Pavesi *Editors*

Optical Switching in Next Generation Data Centers

Springer

Optical Switching in Next Generation Data Centers

Francesco Testa • Lorenzo Pavesi
Editors

Optical Switching in Next Generation Data Centers

Springer

Editors
Francesco Testa
Ericsson Research
Pisa, Italy

Lorenzo Pavesi
Department of Physics, Nanoscience Lab
University of Trento
Povo, Trento, Italy

ISBN 978-3-319-61051-1 ISBN 978-3-319-61052-8 (eBook)
DOI 10.1007/978-3-319-61052-8

Library of Congress Control Number: 2017949388

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Francesco Testa:
To you, Cinzia

Lorenzo Pavesi:
To Sofia, an unexpected sunbeam and a dawn of hope

Preface

The ever-growing society of information has its backbone on a complex network of optical links among nodes where data are stored and processed. These nodes are mostly constituted by data centers. Here a large amount of traffic is handled which is unequally shared between the inter-data center traffic and out-of-the-data center traffic. This huge amount of data is exchanged by using optical communication technologies, thanks to their unique characteristics of high bandwidth, low power consumption, transparency to signal protocol, and enormous bit rate. Advanced optical network technologies are used in data transmission and are reaching the interior of the data centers to meet the increasing demands of capacity, flexibility, low latency, connectivity, and energy efficiency.

In fact, optical point-to-point interconnects support intra-data center networking of compute nodes, electrical packet switching fabrics, and storage equipments. The evolution is toward the use of optical switching to handle optically the data flow. It is difficult to predict when optical switching will be adopted in data centers, which architectures will be used, and which device technologies will be developed to implement these architectures. It is therefore important to make the point now of the technology to help move this transition. This is the scope of the present book.

Recently, many network architectures have been proposed, and many experiments have been realized to demonstrate intra-data center networking based on optical circuit switching, since it improves the performances of a packet communication network working in the electrical domain. Other networking experiments have been based instead on optical packet and burst switching, in which short data packets or longer data bursts are switched directly in the optical domain to further improve the networking resource utilization and energy efficiency. This reduces the electro-optical conversion to a minimum and improves the flexibility by using sub-wavelength bandwidth granularity and statistical multiplexing. In this framework, software-defined networking (SDN) and network function virtualization (NFV) are widely considered key enablers for flexible, agile, and reconfigurable optical data centers since they will provide the coordinated control of network resources and the capability to allocate dynamically the network capacity.

These demonstrations have been made possible by the development of photonic integrated circuits tailored to high-speed optical interconnects and low-cost integrated switches. Integrated approaches allow lowering the cost, footprint, and power consumption with respect to traditional discrete component-based counterparts. The various optical switching architectures make use of a specific optical platform for both transmission and switching, and some of them are based on transmission and switching of gray optical signals, while others exploit the advantages represented by wavelength switching.

This book introduces the reader to the optical switching technology for its application to data centers. In addition, it takes a picture of the status of the technology evolution and of the research in the area of optical networking in a data center. It is clear to the editors that there is still work to do in both the system architecture (toward a scalable architecture) and the device technology (toward high-performance and large-scale integration optical devices) before the introduction of optical switching in commercial networking equipments for data centers. However, the recent progress in the field make us confident that this technology is going to make a big impact on how the future data centers will be run.

The book is organized in four parts: the first part is focused on the system aspects of optical switching in intra-data center networking, the second part is dedicated to describing the recently demonstrated optical switching networks, the third part deals with the latest technologies developed to enable optical switching, and, finally, the fourth part of the book outlines the future prospects and trends.

In Chap. 1, the challenges in current and future data center architectures in terms of scalability, performances, and power consumption are discussed, and the need to develop new hardware platforms based on a tight integration of photonic ICs with electronic ICs and optoelectronic printed circuit boards is underlined. In this chapter also, a hybrid switch architecture based on small electrical switches interconnected by a wavelength router is presented, and the benefit of software-defined networking (SDN) for switch re-configurability and efficient bandwidth utilization is explained.

Chapter 2 reviews the optical circuit switching networks which have been recently proposed with the following main motivations: (a) improvement of the data center networking performances in terms of latency and power consumption by off-loading long-lived bulky data flow from the electrical switching domain to the optical switching networks, (b) provision of a flexible capacity to the intra-data center networking in order to increase the resource utilization, and (c) build of a high-capacity, future-proof networking infrastructure which is transparent to bit rate and protocol.

While an optical circuit switching layer has to operate in conjunction with a more dynamic electrical packet switching layer, optical packet/burst switching systems improve the bandwidth efficiency with sub-wavelength granularity and have the right dynamicity to handle effectively bursty traffic, eventually replacing completely the electrical packet switching layer. Chapter 3 presents and discusses optical packet/burst switching architectures, defines the challenges, and briefly introduces the enabling technologies.

Chapter 4 begins the second part of the book, which is dedicated to the system demonstrations. The chapter describes the implementation and performances of the OSA system architecture. OSA is an optical circuit switched network, which provides a highly flexible optical communication infrastructure between top-of-the-rack (ToR) switches. A first aggregation layer of wavelength-selective switches and a higher level of optical space switches constitute it. This network is able to adapt both the interconnect topology and the capacity to the changing traffic demand, and it supports on-demand connectivity avoiding or greatly reducing oversubscription.

The Hi-Ring architecture is described in Chap. 5. It is based on a multidimensional all-optical switching network interconnecting top-of-the-rack switches. The multidimensional switch comprises a lower layer of space switches, a medium layer of wavelength-selective switches, and a top layer of time-slot switches. While slower space and wavelength switches handle highly aggregated data flows, fast switches are used for time-slot switching of bursty traffic with sub-wavelength granularity. The use of multiple switching allows to implement an optimized network infrastructure with fewer nodes and links among servers with benefits in terms of power consumption, cost, and latency.

In Chap. 6, the LIONS optical network switch is presented, and its experimental demonstrations are discussed. LIONS is a very low-latency, high-bandwidth, energy-efficient switch that interconnects many servers and is implemented in two versions: passive architecture and active architecture. Both types of systems are based on array waveguide router (AWGR) devices. LIONS exploits the AWGR property of de-multiplexing into different output ports a comb of wavelengths received at each input port and multiplexing in a cyclical manner, at each output port, the wavelengths coming from different input ports. There is no need to use fast optical switching fabric, and the wavelength switching is performed by fast tunable laser diodes. Active LIONS is an all-optical packet switch, while in passive LIONS, packet switching is performed in the electrical domain at the network edge with the AWGR performing wavelength routing.

The torus photonic data center is presented in Chap. 7. The top-of-the-rack switches are connected to a network of hybrid optoelectronic routers (HOPRs) interconnected with a torus topology and controlled by a centralized network controller. Such network architecture is characterized by flexible scalability since it can be expanded by simply adding nodes in a plug-and-play manner. In this way, robust redundancy of the links due to the many alternative routes can be made available. Moreover, it does not require high-radix optical switches. The torus network supports optical packet switching (OPS), optical circuit switching (OCS), and the novel virtual optical circuit switching (VOCS).

LIGHTNESS is another switching network for communication among ToRs that combines OPS and OCS in an interchangeable manner with OPS switching short-lived data flows and OCS handling long-lived data flows, and it is controlled by an SDN-enabled control plane. This network is dealt with in Chap. 8.

In Chap. 9, two network architectures are presented. The first is a hybrid optical/electrical packet switching (OPS/EPS) network in which the data packets are separated in small data packet to be handled in the electrical domain and large data packet to be handled in the optical domain. Short packets are forwarded by using conventional protocol, while long packets are processed in an aggregation node by converting each of them into a photonic frame (adding label, guard gap and scrambling.) before sending them to the optical packet switch. The second network is the pure photonic packet switching network that is a synchronous (time-slotted) OPS, handling all types of packets, and is based on a photonic frame wrapper and on the separation of the control path and the data path.

The last recently demonstrated intra-data center network is the optical pyramid data center (OPMDC), which is discussed in Chap. 10. It is a recursive network, based on a pyramid construct, interconnecting ToR switches. OPMDC comprises three tiers of wavelength-selective optical switching nodes; the first is a reconfigurable optical add/drop multiplexer (ROADM) directly connected to the ToR switches, and the upper tiers are wavelength cross-connects (WXC). This network enables extensive wavelength reuse and efficient allocation of wavelength channels, managed by a centralized SDN controller, in order to support packet-based and circuit-based data transfer with low latency.

The third part of the book, dedicated to the enabling technologies, starts with Chap. 11 that reviews the commercially available optical switch technologies. Microelectromechanical system (MEMS), piezoelectric, liquid crystal, $LiNbO_3$, semiconductor optical amplifier (SOA), and photonic lightwave circuit (PLC)-based switches are presented and discussed. A table is included for comparing the key parameters.

Chapter 12 explains the physical effects and mechanisms for optical switching in silicon and presents the different types of switching cells used in large-scale integration silicon photonic switch matrices. The most used silicon photonic matrix architectures are presented and discussed, and three types of matrices are considered: those with switching speed in the range of microseconds, those with switching speed in the range of nanoseconds, and the wavelength-selective switch matrices. The recently demonstrated matrices are here reviewed and compared.

The other key enabling technology for the introduction of optical switching in data centers is the optical transceiver technology. High-speed, low-cost, short-reach optical interconnects must be deployed with efficient modulation formats and photonic integration. Two chapters are focused on this aspect. Chapter 13 presents the trend in high-speed interconnects reviewing the multidimensional modulation formats that allow increasing the transmission rate with respect to on-off key modulation (OOK) without the need of using costly coherent detection systems. The evolution of the transceiver architecture toward a high-dimensional format from 1D to 4D is discussed, and the digital signal processing functions enabling these types of modulations and their direct detection are briefly described.

Chapter 14 reviews the techniques, capabilities, and future potential of InP monolithic integrated technology for the implementation of optical transceivers and optical switches for data centers.

Finally, the fourth part of the book presents, in Chap. 15, an overview of the recent and future trends in technologies and architectures for high-performance optically switched interconnects. The different aspects are discussed: on-chip, on-board, and rack-to-rack optical interconnects and optical switching. Recent research is addressed on the development of new technologies for increasing capacity and performance of optical networks while providing high flexibility and high energy efficiency to support future cloud applications.

We are grateful to our past and present colleagues, students, and friends at Ericsson and at the Nanoscience Laboratory of the Department of Physics of the University of Trento, for maintaining an environment of scientific excellence and friendship over the years. We owe special thanks to the authors of the various chapters for their excellent work. In addition to thanking the authors, we would like to thank Brinda Megasyamalan and Mary E. James for the help, assistance, and patience.

Pisa, Italy
Trento, Italy
April 2017

Francesco Testa
Lorenzo Pavesi

Contents

Part III Technologies for Optical Switching in Data Centers

Part IV Prospects and Future Trends

Part I
System Aspects of Intra Data Center Networking

Chapter 1
Photonics in Data Centers

S.J. Ben Yoo, Roberto Proietti, and Paolo Grani

1.1 Introduction: Recent Trends and Future Challenges of Data Centers and Cloud Computing

Our everyday lives critically depend on data centers. From healthcare to daily banking and everyday commutes, data centers are constantly working with users around the world. With IPv6, the data center can now address every appliance and sensor on earth. The rich set of data will be networked, processed, and accessed on a virtual platform, often called the cloud, which consists of data systems, networks (optical and electrical/wireless and wireline), and client interfaces (e.g., terminals, handheld devices). Warehouse-scale computing systems or data centers are collections of internetworked servers designed to store, access, and process data for the clients. With the explosive growth of data that need to be stored, accessed, and processed, the current trend of the warehouse-scale computing systems is becoming even larger and deeply networked to become hyper-scale data centers. There are three main challenges for such data centers as we look toward the future. Firstly, the power consumption of the data centers limits the scalability. Secondly, the internal data networking limits its performance. Thirdly, the external data networking limits the performance and utility of the cloud. In particular, the energy efficiency of the cyberinfrastructure links all three issues together. In this chapter, we will discuss how photonics can help to enhance energy efficiency of future data centers.

Today's data centers already consume megawatts of power and require large power distribution and cooling infrastructure. Global data center IP traffic expects to grow threefold over the next 5 years, at a Compound Annual Growth Rate (CAGR) of 25% from 2016 to 2021. At the same time, the energy consumption in US data centers reached 91 TWh in 2013 and is expected to increase at a rate that

S.J. Ben Yoo (✉) • R. Proietti • P. Grani
Department of Electrical and Computer Engineering, University of California,
One Shields Ave, Davis, CA 95616, USA
e-mail: sbyoo@ucdavis.edu

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
DOI 10.1007/978-3-319-61052-8_1

doubles about every 8 years [1]. While the exponential trend of data growth has brought optical communications between racks in data centers and computing centers, energy efficiency remains poor due to several reasons.

First, as Fig. 1.1 example illustrates, typical computing and data centers utilize interconnection of various size electronic switches in many cascaded stages. Due to limitations in radix (port count) and bandwidth of the electronic switches, the inefficiency of the cascaded switch stages compounds, especially in terms of latency, throughput, and power consumption. Second, while the need for high-capacity communications brought photonic technologies to data centers, today's embedded-optics solutions (mostly based on pluggable optical modules) do not offer significant savings in the communication chain. Historically, integrated circuits and systems have improved by collapsing functions into a single integrated circuit and eliminating interfaces. Embedded solutions proposed by COBO [2] fail to eliminate any intermediate electronic interfaces such as equalizers and SERializer/DESerializers (SERDES). The transmission distances on electrical wires without repeaters are severely limited due to losses (skin effects or bulk resistivity) and distortion imposed on the signals due to the impedance of the electrical wires [3]. According to Miller and Ozaktas [3], the transmission distance limit is $l = \sqrt{(B_0 / B)(1/A)}$ where A is the cross-sectional area of the electrical wire, B is the line rate, and B_0 is 10^{15} b/s (LC lines) -10^{16} b/s (RC lines), which indicate <1 cm transmission limit at 25 Gb/s line rates for typical modern on-chip electrical interconnects. On the other hand, optical interconnects is free of such impedance effects and becomes advantageous over electrical interconnects beyond a certain distance at a given line rate. Naeemi et al. [4] defined this distance as a "partition length," and Beausoleil et al. [5] have provided

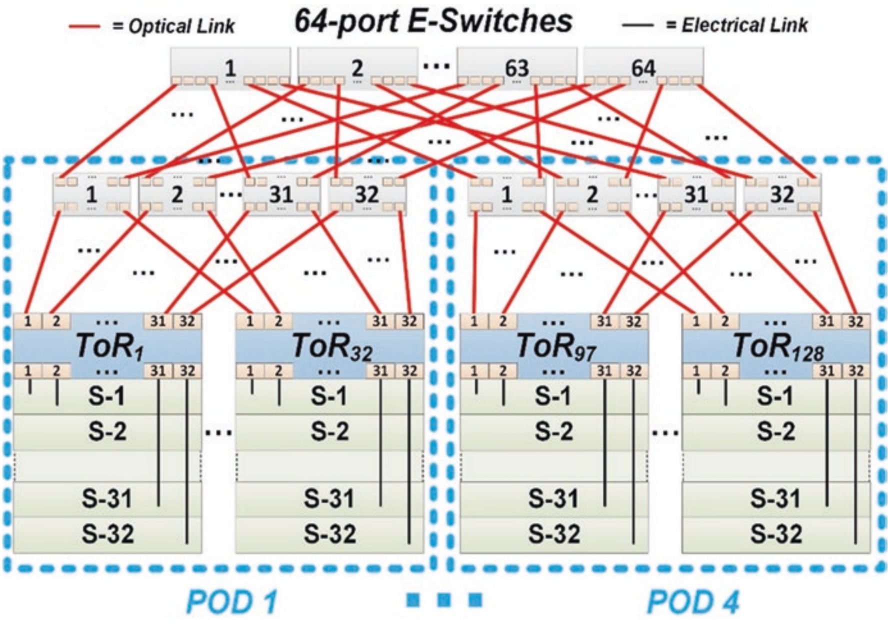


Fig. 1.1 128 rack data center using electrical switches

detailed calculation of these lengths according to ITRS [6], where he found that the partition length for a wire or waveguide width of 1 μm is less than 2 mm.

Thirdly, today's computing systems are designed for a fixed topology with fixed patterns of data movements at fixed data rates, while actual computations have large peak-to-average ratios in processing, bursty data traffic, dynamically changing data movement patterns, and heterogeneous processing threads.

In the following sections, we will discuss the solutions to the three issues after we visit the limitations in the use of electronics for processing and switching the data. Power consumption in electronics is a serious roadblock even if we can keep up with Moore's law (Fig. 1.2) regarding integrating more devices on each chip to meet the demands of future applications. At the device level, Dennard's law [7], which described the simultaneous improvements in transistor density, switching speed, and power dissipation [7] to follow Moore's law [8], has already become obsolete in 2004. As Fig. 1.2 illustrates, while Moore's law continued for more than four decades, the clock speeds and the power efficiencies have flatlined since ~2004. Multi-core solutions emerged shortly after 2004 as a new processor-level solution for the power efficiency and scaling, and it is sometimes called a new Moore's law. While we believe that the multi-core and chip-level parallelism solutions will continue to expand, the communication and data movements will continue to be a challenge. For this reason, for over three successive generations, the performance/watt has improved only marginally. Multicore and GPU-based solutions improved the performance/watt very recently, but these improvements appear to be a one-time reprieve.

Obviously, electronics alone cannot provide solutions to all the challenges to massively parallel data processing. Electronics accompany skin effects, capacitance, electromagnetic interference (EMI), and distortion/dispersion, while photonics support nearly distance-independent parallel transport across the vast optical bandwidth [9]. As the computing nodes are evolving to multi-core and multiprocessor systems with very high bandwidth requirements between processors and memory banks on the same board, inter-chip optical interconnects can also provide significant benefits in terms of energy per bit. Reference [9] shows a comparison between optical interconnects with on-chip and off-chip laser and electrical interconnects, showing significant advantages of the optical solution for distances above few tens of millimeter, and Ref. [10–13] provide more information regarding optical vs. electrical interconnects.

On the other hand, photons cannot be stored easily nor can they interfere easily to be part of three-terminal devices. As Fig. 1.3 illustrates, hybrid solutions exploiting the best of both worlds (photonics and electronics) will be beneficial to future data centers.

Such hybrid solutions should be sought not only between the racks but also between the boards and the chips. Unlike telecommunications where typically ~80% traffic bypasses and ~20% traffic adds/drops locally, data traffic in computing systems is ~80% internal and ~20% external. The statistics from Cisco [14] shown in Fig. 1.4 supports this argument by showing that 77% is internal and 23% is external. Hence the symbiotic integration of photonics and electronics has to happen at every level—between racks, boards, cards, chips, and cores. One of the main emphases we place in this chapter is a computing system architecture based on embedded photonics—photonics will be everywhere in the data system at every hierarchy.

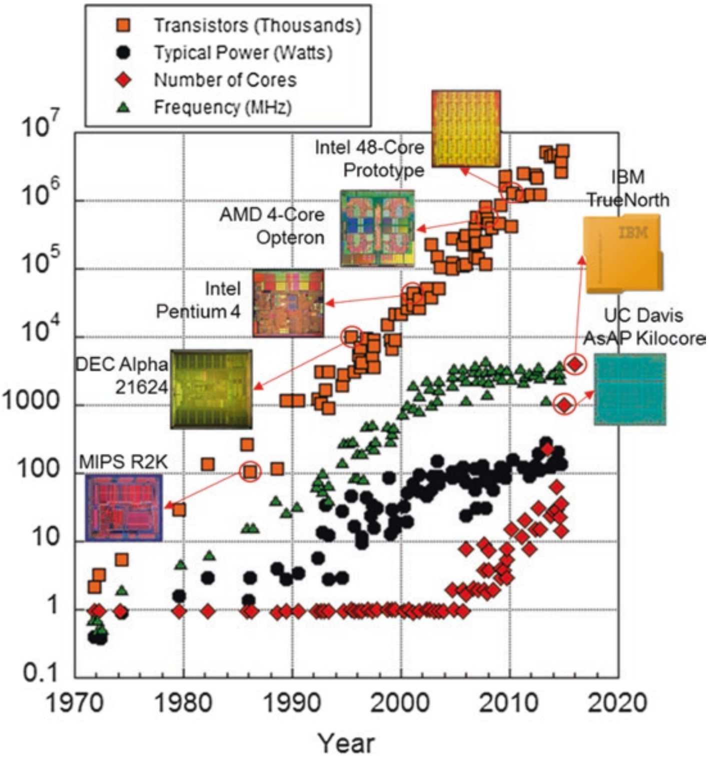
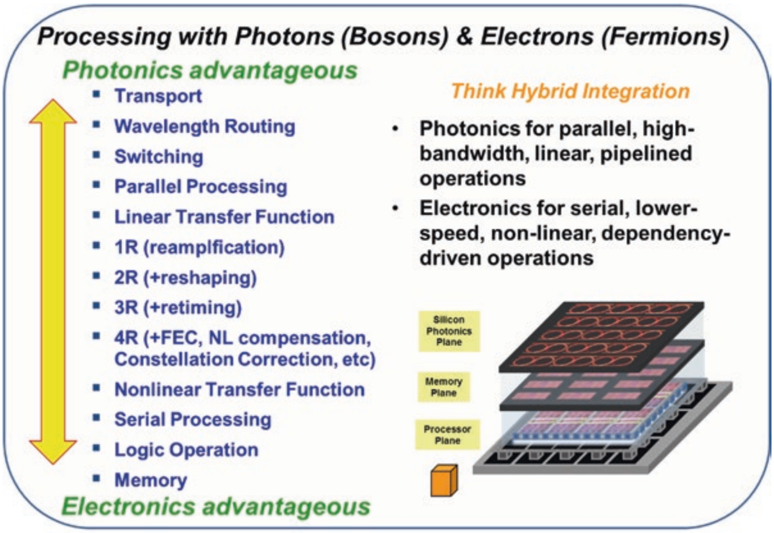


Fig. 1.2 A 45-year trend of a number of transistors per integrated circuit, clock speed (MHz), power (W), performance per clock (ILP), and a number of core per processor die (Figure created based on data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, M. Horowitz, F. Labonte, O. Shacham, and Christopher Batten)



Processing with Photons (Bosons) & Electrons (Fermions)

Photonics advantageous

- Transport
- Wavelength Routing
- Switching
- Parallel Processing
- Linear Transfer Function
- 1R (reamplfication)
- 2R (+reshaping)
- 3R (+retiming)
- 4R (+FEC, NL compensation, Constellation Correction, etc)
- Nonlinear Transfer Function
- Serial Processing
- Logic Operation
- Memory

Electronics advantageous

Think Hybrid Integration

- Photonics for parallel, high-bandwidth, linear, pipelined operations
- Electronics for serial, lower-speed, non-linear, dependency-driven operations

Fig. 1.3 Hybrid photonic-electronic solutions in data systems can offer best of both worlds

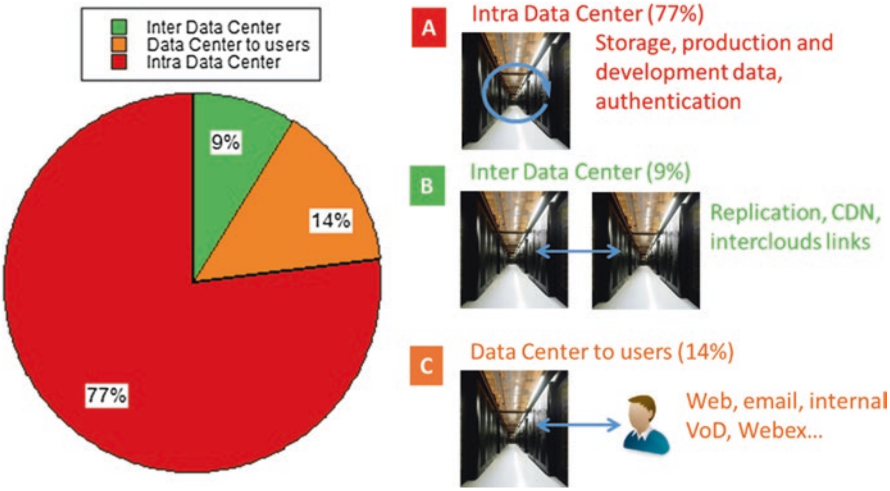


Fig. 1.4 [14] Global Data Center Traffic by Destination in 2020 (Source: Cisco Global Index, 2015–2020; Synergy Research)

1.2 New Directions for Data Centers with Embedded Photonics

It is possible to address the three challenges above as follows. First, we can avoid the typical data center interconnection architectures based on many cascaded stages of electronic switches of limited radix and bandwidth by introducing a flat architecture with an all-to-all interconnection as shown in Fig. 1.5a, b to support contention-free interconnection (no arbitration is required) at full throughput (total of N^2 links). Using a $N \times N$ cyclic arrayed waveguide grating router (AWGR) [16] with its unique wavelength routing property [$N = 5$ example shown in Fig. 1.5d], fully connected all-to-all interconnection (total of N^2 links) without contention becomes possible with N wavelength channels. Then, the all-to-all interconnection topology of Fig. 1.5a is simplified as Fig. 1.5b with each fiber containing N wavelengths. Figure 1.5e shows an implementation [15] of the all-to-all interconnection topology involving N compute nodes with silicon photonic micro-resonator-ring modulators and detectors and an $N \times N$ cyclic AWGR low-latency interconnect optical networks (LIONS) ($k_t = k_r = N$ example) [17]. While a 512 × 512 AWGR [18] has been demonstrated, the requirement for $N \times (N - 1)$ transceivers on N wavelength channels becomes challenging and not scalable for a large number of nodes, N. Fortunately, hierarchical all-to-all interconnection networking called HALL [19] greatly reduces the number of required transceivers and wavelengths while supporting maximum throughput of ~97% and all-to-all connectivity at every hierarchy. Figure 1.5c shows a simpler two-hierarchy topology, called RH-LIONS (reconfigurable hierarchical-LIONS) [20], utilizing smaller reconfigurable all-optical switch at the higher hierarchy. Here, cost reductions can be achieved by introducing partial and reconfigurable all-to-all interconnection at the higher hierarchy while maintaining full all-to-all connectivity at the lowest

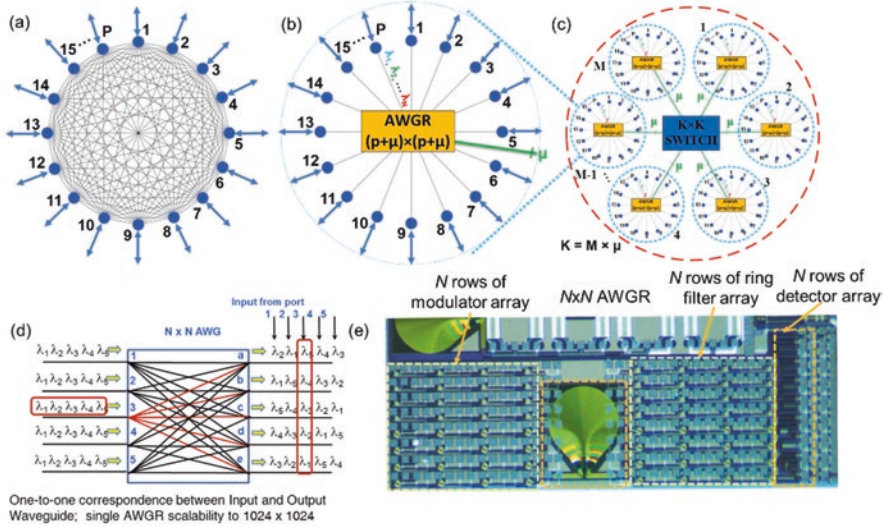

(a)
P
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
(b)
AWGR
(p+μ)×(p+μ)
μ
(c)
K×K SWITCH
M
M-1
K = M × μ
(d)
N x N AWG
Input from port
a
b
c
d
e
One-to-one correspondence between Input and Output Waveguide; single AWGR scalability to 1024 x 1024
(e)
N rows of modulator array
NxN AWGR
N rows of ring filter array
N rows of detector array

Fig. 1.5 (**a**) Fully connected all-to-all interconnection network, (**b**) fully connected all-to-all interconnection network utilizing wavelength routing by an arrayed waveguide grating router (AWGR), (**c**) RH-LIONS with fully connected subnetworks that are interconnected with a reconfigurable optical switch, (**d**) all-to-all wavelength routing interconnection pattern of a $N \times N$ cyclic AWGR using N wavelengths ($N = 5$ example), (**e**) a silicon photonic chip implementation [15] of (**b**) for $N = p = 8$, $\mu = 0$ with silicon photonic microring-resonator modulators and detectors with all-to-all interconnection via a $N \times N$ AWGR (LIONS)

hierarchy. Indeed, while the lowest hierarchy of the interconnection can exploit all-to-all connectivity, the inter-board and the intercluster interconnections already set up for all-to-all connectivity can be reconfigured to more effectively map the network topology to resemble workflow topology. Dynamic assignment of more (less) bandwidths and channels in different routes can remove hotspot congestion and improve energy efficiency (see section below on software-defined elasticity).

Second, although the recent efforts including COBO [2] helped transition from Fig. 1.6a to Fig. 1.6b, this solution fails to eliminate any intermediate electronic interfaces such as equalizers and SERializer/DESerializers (SERDES). A new embedded photonics solution depicted in Fig. 1.6c intimately integrates electronics, silicon photonics [21, 22], and optical interposers [23–25]. Today's embedded optics makes use of standard pluggable optical interfaces connecting to ball grid array (BGA) packaged application-specific integrated circuits (ASICs). This approach typically requires more than 25 mm long electrical interconnections. The embedded photonics with 2.5D and 3D integration using silicon photonic interposers utilizes interconnection lengths below 100 μm between the electronics and silicon photonics. Embedded photonics can significantly impact the energy efficiency and the cost of chip-to-chip, board-to-board, and rack-to-rack data communications. However, today's embedded optics provide limited energy efficiency improvements by requiring SERDES and clock-data recovery (CDR). By reducing its reliance on electrical SERDES, CDR, and equalizers, embedded photonics can greatly reduce the power consumption and operating costs.

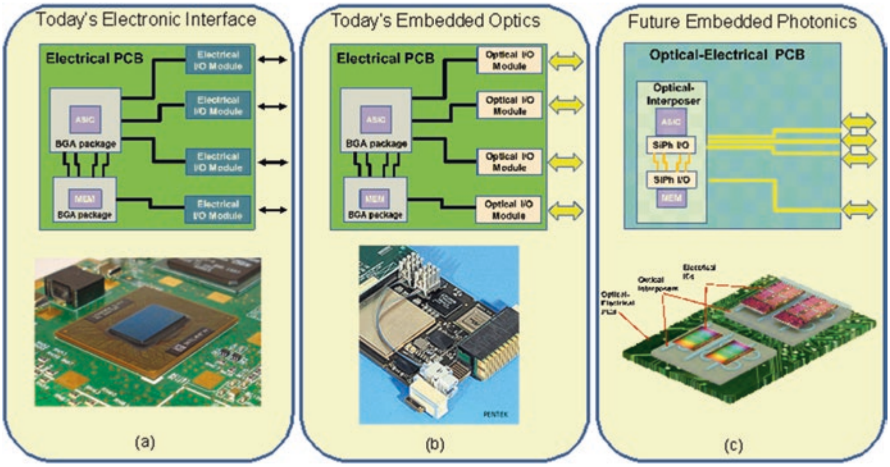


Fig. 1.6 Comparisons of (**a**) today's electronic interfaces with electronic I/Os, (**b**) today's embedded optics with standard pluggable optical interfaces to BGA-packaged ASICs, and (**c**) the proposed embedded photonics with 2.5D and 3D integration using silicon photonic interposers

Thirdly, today's computing systems with fixed topology with fixed patterns of data movements at fixed data rates can be renovated into an optically and electronically reconfigurable system architecture adapting to workload and traffic patterns. In particular, software-defined network solutions with virtualization can involve both photonic and electronic solutions therein. Recent trends are clearly showing a preference for modular scalability and software-defined reconfigurability of data centers. Embedded silicon photonics with ASICs and memories on photonic-electronic interposers can plug into optical-electrical printed circuit boards (OE-PCBs) [25, 26], which will, in turn, plug into OE-backplanes. Optical reconfigurations could exploit miniature optical microelectromechanical switches (MEMS) and wavelength assignments driven by software-defined control planes like in application-driven reconfigurable optical network (ARON) data centers [27].

1.3 Arrival of Embedded Photonics, Silicon Photonics, and Heterogeneous 2.5D and 3D Integration

We envision that future data centers will exploit photonics embedded with electronics through close integration everywhere, in chip-to-chip, board-to-board, and rack-to-rack interconnections. While monolithic co-integration of CMOS and silicon photonic in the same fabrication runs is attractive, the yield and the required technological compatibility challenges make it impractically expensive. Optical interposers and OE-PCBs are practical and effective technologies that enable reduced parasitic, low power consumption, dense optical interconnects, and close integration of photonics and electronics while allowing flexible combinations of heterogeneous technologies with reasonable yield.

Figure 1.7a–d illustrates a method of embedded photonics utilizing active silicon photonic interposers (optical interposer with silicon photonic modulators and detectors) interfacing with electronic ICs and OE-PCBs. (a) and (b) show a side view and a top view schematic of 2.5D integration of the electronic ICs, active silicon photonic interposers, and OE-PCBs achieved by this method, (c) shows an assembly process using evanescent coupling between the silicon photonic waveguides and the OE-PCB waveguides, and (d) illustrates the case with such multiple optical interposers assembled on a larger OE-PCB.

Figure 1.8 show that <0.1 dB optical loss is maintained even for ±1 μm misalignment tolerance between the silicon photonic active optical interposer and the OE-PCB. OE-PCBs and OE-backplanes will exploit low-loss optical waveguide layers laminated on the conventional electrical PCBs.

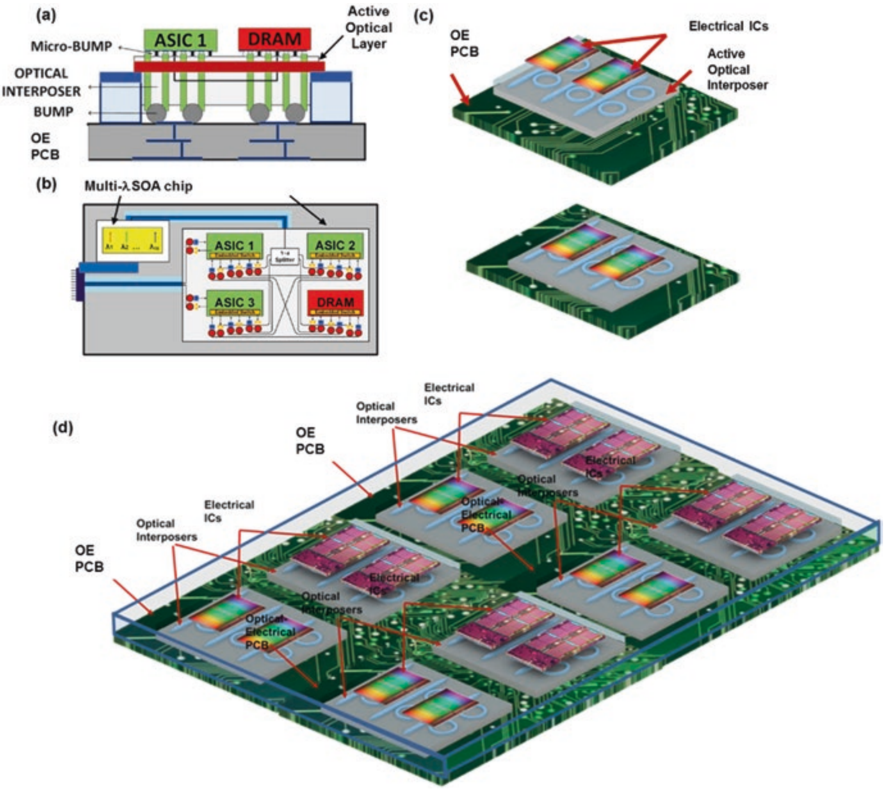


Fig. 1.7 (**a**) and (**b**) show a side view and a top view schematic of 2.5D integration of the electronic ICs, active silicon photonic interposers, and OE-PCBs, (**c**) shows an assembly process using evanescent coupling between the silicon photonic waveguides and the OE-PCB waveguides, and (**d**) illustrates an OE-PCB containing multiple silicon photonic optical interposers and electronic ICs

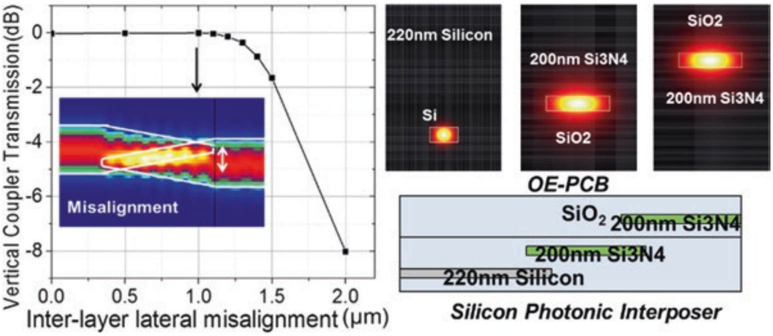


Fig. 1.8 Coupling and misalignment tolerance between the optical interposer and a silicon photonic die consisting of negative tapers indicating +/− 1 μm lateral misalignment tolerance

1.4 OE-PCBs and OE-Backplanes

PCBs with embedded optical layers offer a cost-effective opportunity to reduce energy consumptions and latency induced by electrical wires at high data rates [28–33]. Successful OE-PCBs will eliminate any need for high-speed electrical interconnections on board, and electrical connections will only support power and low-speed control and programming. The majority of the past efforts [28–33] focused on multimode polymer optical waveguides within FR4 PCBs, where multimode dispersions and high losses limited the performance and energy efficiency improvements. There has been recent advances in single-mode optical polymer PCBs [34] and multimode glass waveguide PCBs [35–37] to pursue single-mode glass optical waveguides embedded in electrical PCBs. Initial efforts will utilize the glass lamination technology mentioned in [36] to embed ion-exchanged silica waveguide layer in between two FR4 electrical PCBs [36] as shown in Fig. 1.9a. This method offers relatively sturdy operation which somewhat mitigates the difference in thermal coefficients of expansion (TCEs) between the glass and the FR4 but requires the opening of the FR4 in shape to drop in the optical interposer modules. Similar openings should be made on the FR4 on the other side to balance the stress and TCE difference. Successful progress in developing OE-PCBs with optical and electrical connectors will allow realizing OE-PCBs and building of servers and switches interconnected with optical waveguides as shown in Fig. 1.9c.

1.4.1 High-Radix Optical Switches

As mentioned above, the limited radix and bandwidth of electronic switches severely affect data center scalability regarding latency, throughput, and power consumption. Optical switching can potentially overcome the above limitations, and many optical switch architectures have been investigated and reported in the literature. Table 1.1

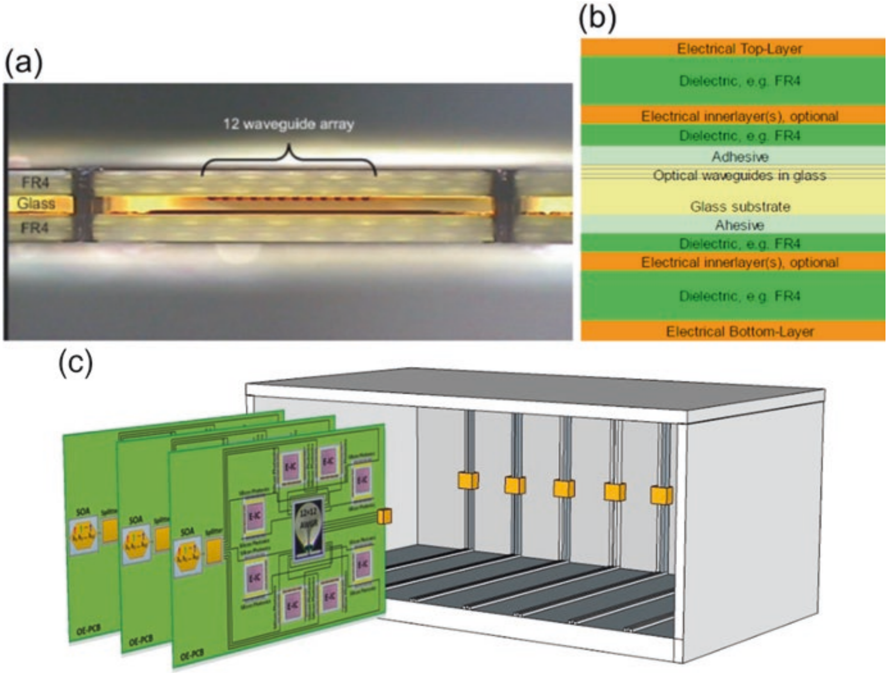


Fig. 1.9 (**a**) A cross-section photograph [36] and (**b**) the composition of a multimode OE-PCB with an ion-exchanged glass waveguide layer sandwiched between two FR4 electrical PCBs [36]. (**c**) Connectorizing the OE-PCBs to realize a chassis with an OE-backplane

summarizes the main all-optical switching technologies highlighting pros and cons of each solution (please also refer to Chap. 8 for more details on AWGR-based switching).

Despite the significant differences highlighted in Table 1.1, all these switches share a common aspect: they are bufferless (no buffering operation at the switch input and output ports) and therefore cannot be cascaded. Therefore, they could be used mainly as core switches in folded-CLOS type of architectures (i.e., Fat Tree [50, 51]) and also in directly connected architectures like torus [52], flattened butterfly [53, 54], or dragonfly [55], where they can interconnect directly computing nodes or top-of-rack (ToR) switches. Also, MEMS switches are the only ones, among the solutions in Table 1.1, currently commercially available. However, due to the slow switching time in the order of milliseconds, MEMS can only switch the so-called elephant flows, and they cannot replace the packet-switching features of electronic switches.

Next-generation high-radix high-bandwidth data center switches will make use of multiple electronic switches optically interconnected on a common silicon interposer (see next section for more details) with 2.5D and 3D hybrid electro-optic integration platforms. The low crosstalk, very low loss, and high energy

Table 1.1 Different all-optical switching technologies

Technology	Pros	Cons
MEMS [38–40]	• Transparent to line rate and modulation format • WDM compatible • High radix (up to 1024)	• ms switching time (only elephant flows)
SOA [41–44]	• Transparent to line rate and modulation format • WDM compatible • *ns* switching time	• Number of SOAs scales nonlinearly with switch radix • High-power consumption • Radix limited to <32
AWGR [45–47]	• ns switching time when used together with fast tunable lasers • WDM implements output queuing • Number of active element scales linearly with switch radix	• Port line rate limited by the AWGR channel bandwidth • Radix limited by in-band crosstalk (Radix <= 128)
MRR [48, 49]	• MRR tuning permits flexible bandwidth allocations • Because of the dense wavelength division multiplexing (DWDM), a high-radix photonic switch will have fewer off-chip fiber connections than pins in a comparable electronic switch	• High switching latency • 1 microsecond → no packet-switching • Too many MRRs for an all-to-all connection → very high power consumption due to thermal tuning • Arbitration might be required • Radix limited by in-band crosstalk (Radix <32)

MEMS microelectromechanical system, *SOA* semiconductor optical amplifier, *AWGR* arrayed waveguide grating router, *MRR* microring resonator

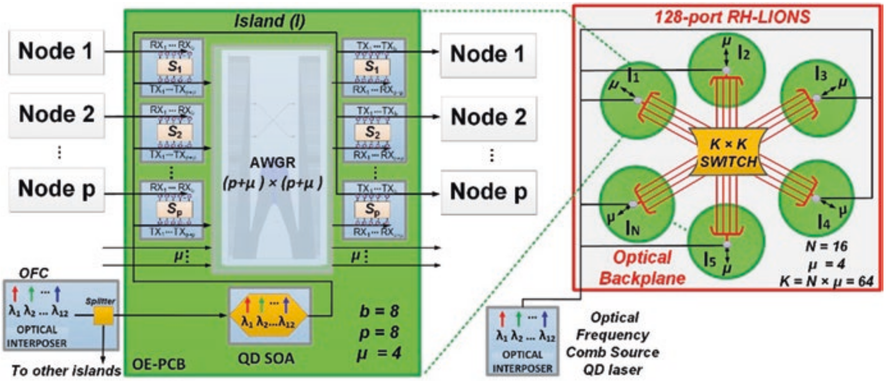


Fig. 1.10 A two-hierarchy switch RH-LIONS switch of size *pN* ×*pN*. Shown is an example of a 128-port switch with $N = 16$, $b = p = 8$, $\mu = 4$, $K = N \times \mu = 64$

efficiency provided by photonic interconnects will potentially enable unprecedented switch bandwidth and radix. Figure 1.10 shows an example of such hybrid approach currently under development at UC Davis NGNS laboratories. This switch is called RH-LIONS (reconfigurable hierarchical low-latency interconnect optical network switch).

RH-LIONS makes use of the electronic-photonic integration technologies to implement a switching architecture with small electrical switches at the edges all-optical interconnected via wavelength routing in AWGR. The proposed solution can scale far beyond 128-port and 50 Tb/s capacity. Figure 1.10 shows a two-hierarchy switch RH-LIONS switch *with p *N ports*. Figure 1.10 shows an example of a 128-port switch with $N = 16$, $b = p = 8$, $\mu = 4$, $K = N \times \mu = 64$. In general, two-hierarchy RH-LIONS switch includes N islands (green boxes in Fig. 1.10 [right]). Each island is composed of p electronic switches (S in the figure) and connects to p nodes. p also represents the number of required wavelengths and AWGR ports for intraisland communication, to let the nodes communicate with an all-to-all scheme through the AWGR. μ is the number of AWGR ports and the number of wavelengths reserved for inter-island communications. Therefore, $p + \mu$ is the total number of AWGR ports and wavelengths required. To reach a very high board-level I/O (i.e., bandwidth per switch port), we are expected to use a WDM optical value of b (number of wavelengths per port), where b = *aggregate-switch-BW / #switch port / line rate / 2*. Finally, a $K \times K$ optical switch (circuit switching) allows to use fewer optical transceivers μ and to reconfigure the topology between the islands. For instance, if $\mu = 4$, we can create a baseline *mesh*, and then the $K \times K$ optical switch can be used to modify the topology according to the traffic patterns. To build a 128-port RH-LIONS switch, we need 16 12-port AWGRs (one AWGR per island, $N = 16$ is the total number of islands), 128 E-switches 20-port ($p + \mu + b$ ports, S in the figure), and one circuit-based optical switch (e.g., MEMS K-port, $K = \mu \times N$, where $\mu = 4$ and $N = 16$). In Fig. 1.10, each port of the switch can support up to 200 Gbps (board-level I/O, $b = 8$ WDM, 25 GHz optical frequency), for a total aggregate bandwidth of 51.2 Tbps. The E-switch can be a commodity switch die and can be very energy efficient (300 mW per port with up to 24 ports at 25Gbps [56] – note that this is the power consumption for the packaged chip).

1.5 Software-Defined Elasticity in Data Centers and Clients

Typical data centers run heterogeneous applications that exhibit various communication patterns among the computing nodes. To optimize the performance of an application, we need to match the communication network to the specific application. However, today's data centers use a single architecture to serve various applications. We believe that flexible physical topology reconfiguration exploiting optical switches as illustrated in Fig. 1.5c and investigated in [27, 39, 57] can play a major role in efficiency and optimization of future data center optical networks.

Figure 1.11 shows an example of what could be achieved with such reconfigurability. As Fig. 1.11a shows, *mesh* is the suitable topology under neighborhood traffic, since it can achieve similar performance with much fewer transceivers than HyperX. When the traffic changes to uniform random, we configured to HyperX to achieve 1.26× higher throughput and 1.97× lower latency. Figure 1.11b

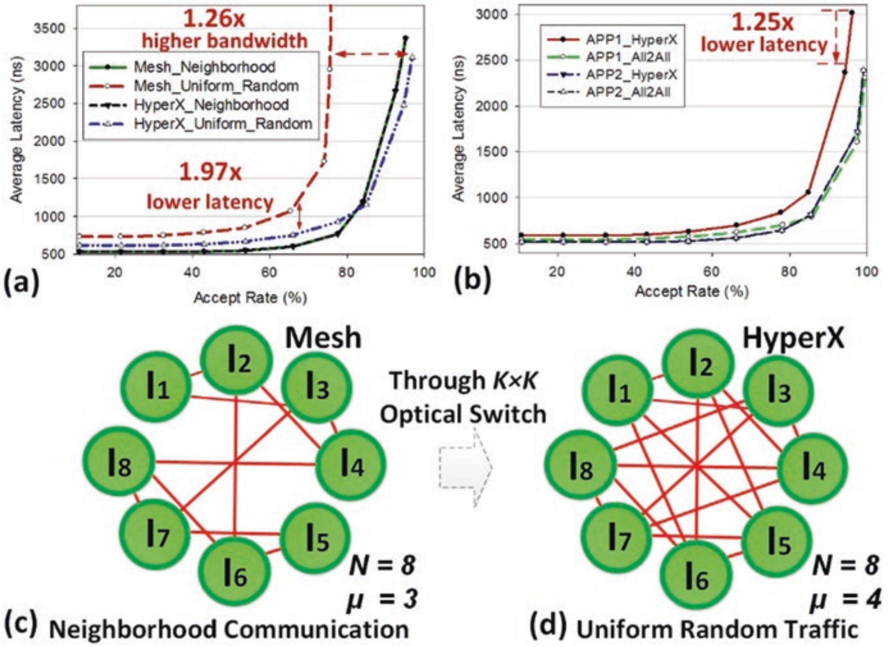


Fig. 1.11 (**a**) Performance of running a single application; (**b**) performance of running two applications; (**c**) full-system mesh with neighborhood communication traffic; (**d**) full-system HyperX with uniform random traffic

shows that, compared with a full-system HyperX, all-to-all achieve similar throughput but with 1.25× lower latency [27]. See also Chapter 6 on LIONS testbed demonstrations to see additional examples of the benefits that can be achieved by reconfigurability.

In addition to network topology reconfiguration between racks and clusters, another level of flexibility and control could be achieved at the link level by adopting dynamic voltage and frequency scaling (DVFS [58]) to adjust dynamically the transceiver bandwidth according to the link utilization. This technology, already applied in the electronic domain inside the processors or computing boards, could be extended to the longer optical interconnects to improve further the energy efficiency of the data links, which are known to be bursty. It is well known that the dynamic power of CMOS transistor scales as $\propto V_{dd}^{2} * f$, where V_{dd} is the driving voltage and f is the clock speed. If V_{dd} can be lowered for circuits with low f, it is then possible to obtain significant improvement in energy efficiency by lowering the clock speed in combination with the driving voltage (nearly 2× improvements in power efficiency for 20% underclocking).

Figure 1.12 shows an example of an optically interconnected multi-socket board (MSB). Figure 1.13 shows some achieved results for an AWGR-based

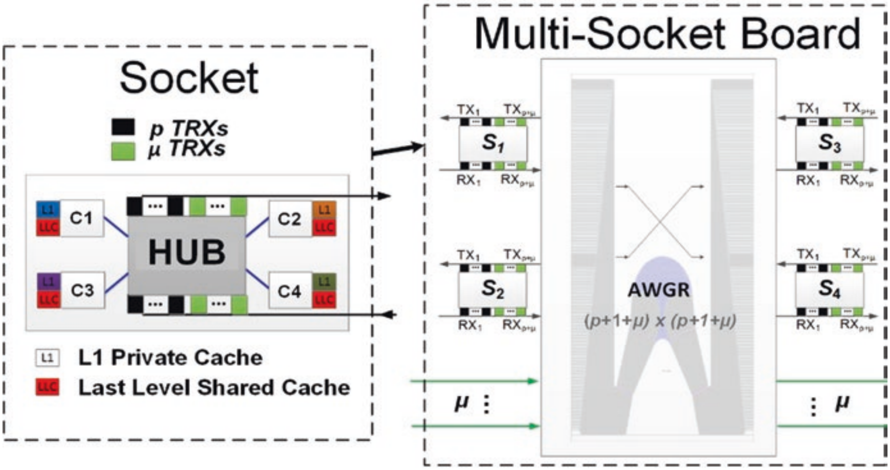


Fig. 1.12 An example of hierarchical optical interconnected architecture for inter-socket communication within a board and between multiple boards. (*Left*): the socket (S) topology with the hub switch connecting the four computing cores with private and shared cache memory. (*Center*): the multi-socket board (MSB) with four sockets based on passive AWGR all-to-all interconnection

board-level architecture [45, 59]. The comparison has been performed with a state-of-the-art electronic board-level topology regarding normalized execution time (top) and normalized *transaction/Joule* (bottom) for each one of the considered applications. In our AWGR solution, we applied a conventional clock and data recovery (CDR) technique, as well as source synchronous [60] and dynamic voltage and frequency scaling (DVFS) [61–64] model. With DVFS the system can dynamically set the transmitter frequency and voltage supply to different values depending on the traffic load. The proposed optical architecture achieved an average execution time improvement of a factor of ~3× when exploiting a CDR-based transmission and of ~2.5×, when exploiting a DVFS transmission scheme with source synchronous technique. Figure 1.13b shows the improvements achieved in terms of *transaction/Joule* exploiting the optical solutions. On average, we were already able to outperform the electronic baseline by a factor of >2.5×. We defined the concept of *transaction* for each one of the considered applications. For instance, *ferret* is an image similarity search benchmark. For this application, we considered a *query* as a transaction.

Another interesting application in the suite is *swaptions* which replicates a financial analysis and in which the transaction can be defined as a *bank atomic operation*. Note that these benchmarks used in this example are only some of the benchmarks that could be used to evaluate the performance of different data center solutions. In fact, choosing the appropriate benchmarks to mimic and provide the best representation of data center traffic is very challenging due to the heterogeneity of the many and concurrent applications running [66–68].

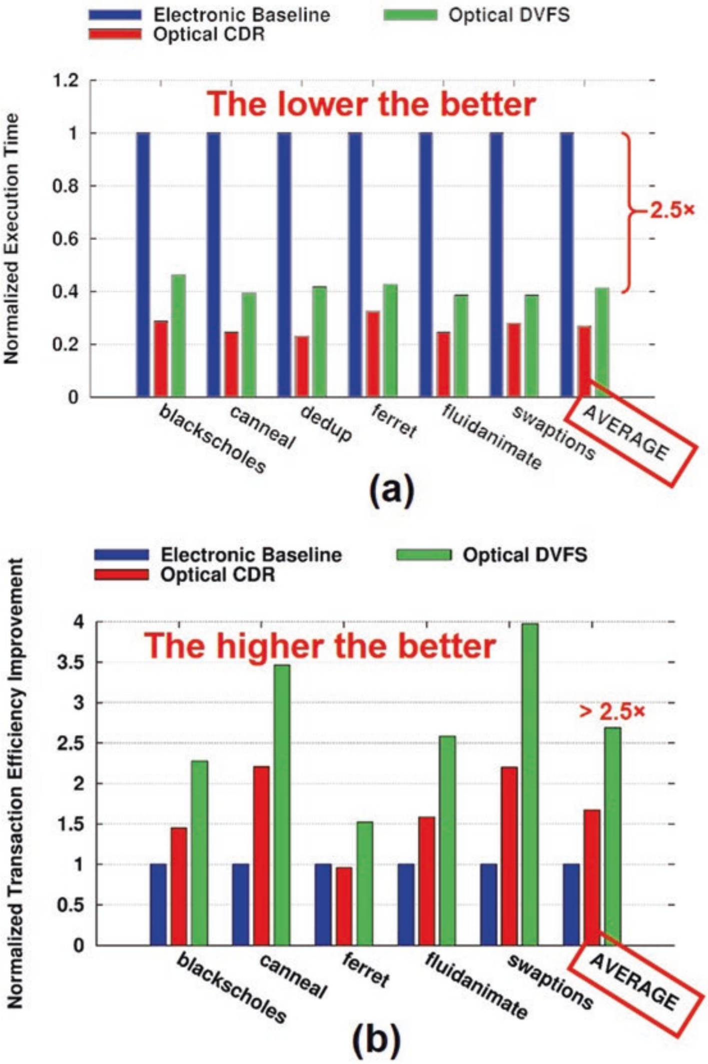


Fig. 1.13 (**a**) Execution time and (**b**) transaction energy efficiency normalized to the electronic baseline in comparison with an optical hierarchical solution with CDR and DVFS [65]

1.6 Summary

This chapter introduced the challenges faced by today's data centers in terms of scalability, power consumption, and performance due to the limitations of electrical interconnects and switches as the bandwidth per port and bisection bandwidth requirements increase. While photonics is already currently used in point-to-point communications between ToR switches, the benefits of today's embedded optics with standard pluggable modules are still quite limited. For photonics to bring transformative changes and benefits for next-generation data centers, it is necessary to develop mature technologies for intimate integration of photonic devices and electronic ICs on a common silicon interposer for on-board chip-to-chip communication. Eventually, multiple interposers will be able to communicate through an optoelectronic PCB for board-to-board communication with superior performance regarding energy efficiency and link density and bandwidth.

While many interesting all-optical switching technologies have been extensively studied and reported in the literature (see Table 1.1), none of them are still technologically and commercially viable to be deployed in the field. So, the core of future data center switching will still be relying on electrical ICs but densely packed on the same interposer and OE-PCB to achieve the required bisection bandwidth with limited power (see above section "High-Radix Optical Switches").

Finally, the chapter introduced the possibility to exploit software-defined optical reconfigurability at the node and network level to adapt the link bandwidth and network topology to the dynamic traffic profile typical of datacom systems. Workload-aware reconfiguration allows optimal data center performance and energy utilization adapting to the workload.

References

1. J. Koomey, Growth in data center electricity use 2005 to 2010 (2011)
2. COBO, *Consortium for On-Board Optics* (2015); Available from: http://cobo.azurewebsites.net
3. D.A.B. Miller, H.M. Ozaktas, Limit to the bit-rate capacity of electrical interconnects from the aspect ratio of the system architecture. J. Parallel Distribut. Comput. **41**(1), 42–52 (1997)
4. A. Naeemi et al., Optical and electrical interconnect partition length based on chip-to-chip bandwidth maximization. IEEE Photon. Technol. Lett. **16**(4), 1221–1223 (2004)
5. R.G. Beausoleil et al., Nanoelectronic and nanophotonic interconnect. Proc. IEEE **96**(2), 230–247 (2008)
6. I.R. Committee, in *International Technology Roadmap for Semiconductors: 2013 Edition Executive Summary* (Semiconductor Industry Association, San Francisco, 2013). Available at: http://www.itrs.net/Links/2013ITRS/2013Chapters/2013ExecutiveSummary.pdf
7. R.H. Dennard et al., Design of ion-implanted mosfets with very small physical dimensions. IEEE J. Solid-State Circuit **9**(5), 256–268 (1974)
8. G.E. Moore, Cramming more components onto integrated circuits. Electronics **38**(8) (1965)
9. M. Stucchi et al., On-chip optical interconnects versus electrical interconnects for high-performance applications. Microelectron. Eng. **112**, 84–91 (2013)
10. H. Cho, P. Kapur, K.C. Saraswat, Power comparison between high-speed electrical and optical interconnects for interchip communication. J. Lightw. Technol. **22**(9), 2021 (2004)
11. C. Guoqing et al., Electrical and optical on-chip interconnects in scaled microprocessors. in *2005 IEEE International Symposium on Circuits and Systems*, 2005
12. D.A.B. Miller, Optical interconnects to electronic chips. Appl. Opt. **49**(25), F59–F70 (2010)
13. S. Rakheja, V. Kumar, Comparison of electrical, optical and plasmonic on-chip interconnects based on delay and energy considerations. in *Thirteenth International Symposium on Quality Electronic Design (ISQED)*, 2012

14. Cisco, Cisco Global Cloud Index: Forecast and Methodology, 2015–2020 (White Paper) (Cisco, 2016)
15. R. Yu et al., A scalable silicon photonic chip-scale optical switch for high performance computing systems. Opt. Express **21**(26), 32655–32667 (2013)
16. I.P. Kaminow et al., A wideband all-optical WDM network. IEEE J. Select. Areas Commun. **14**(5), 780–799 (1996)
17. X. Ye et al., DOS – a scalable optical switch for datacenters. in *ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS),* 2010
18. S. Cheung et al., Ultra-compact silicon photonic 512 x 512 25 GHz arrayed waveguide grating router. IEEE J. Sel. Top. Quantum Electron. **20**(4), 310–316 (2014)
19. Z. Cao, R. Proietti, S.J.B. Yoo, HALL: a hierarchical all-to-all optical interconnect architecture. in *2014 Optical Interconnects Conference*, 2014
20. Z. Cao et al., Experimental Demonstration of flexible bandwidth optical data center Core network with all-to-all interconnectivity. J. Lightwave Technol. **33**(8), 1578–1585 (2015)
21. M. Hochberg et al., Silicon photonics: the next fabless semiconductor industry. IEEE Solid-State Circuit Mag. **5**(1), 48–58 (2013)
22. L. Chrostowski, M. Hochberg, *Silicon Photonics Design: From Devices to Systems* (Cambridge University Press, Cambridge, 2015)
23. Y. Urino et al., *First demonstration of athermal silicon optical interposers with quantum dot lasers operating up to 125 &#x00B0;C*. J. Lightw. Technol. **33**(6), 1223–1229 (2015)
24. K.W. Lee et al., Three-dimensional hybrid integration technology of CMOS, MEMS, and photonics circuits for optoelectronic heterogeneous integrated systems. IEEE Trans. Electron Devices **58**(3), 748–757 (2011)
25. A. Michaels, E. Yablonovitch, Reinventing circuit boards with high density optical interconnects. in *2016 IEEE Photonics Society Summer Topical Meeting Series (SUM)*, 2016
26. L. Brusberg et al., Electro-optical circuit board with single-mode glass waveguide optical interconnects (2016)
27. Y. Guojun et al., ARON: application-driven reconfigurable optical networking for HPC data centers. in *European Conference on Optical Communication, ECOC*, Dusseldorf, 2016
28. T. Ishigure et al., Low-loss design and fabrication of multimode polymer optical waveguide circuit with crossings for high-density optical PCB. in *2013 IEEE 63rd Electronic Components and Technology Conference (ECTC)*, 2013, pp. 297–304
29. R. Kinoshita et al., Polymer optical waveguides with GI and W-shaped cores for high-bandwidth-density on-board interconnects. J. Lightw. Technol. **31**(24), 4004–4015 (2013)
30. R. Pitwon et al., International standards for optical circuit board fabrication, assembly and measurement. Opt. Commun. **362**, 22–32 (2016)
31. R. Pitwon et al., International standardisation of optical circuit board measurement and fabrication procedures. in *Optical Interconnects Xv*, ed. by H. Schroder, R.T. Chen (2015)
32. A.F. Rizky et al., Polymer waveguide-coupled 14-Gb/s x 12-channel parallel-optical modules mounted on optical PCB through Sn-Ag-Cu-solder reflow. in *2013 IEEE 3rd CPMT Symposium Japan*, 2013
33. K. Soma, T. Ishigure, *Fabrication of a graded-index circular-core polymer parallel optical waveguide using a microdispenser for a high-density optical printed circuit oard*. IEEE J. Sel. Top. Quantum Electron. **19**(2) (2013)
34. R. Dangel et al., Polymer waveguides for electro-optical integration in data centers and high-performance computers. Opt. Express **23**(4), 4736–4750 (2015)
35. H. Schröeder et al., Glass panel processing for electrical and optical packaging. in *2011 IEEE 61st Electronic Components and Technology Conference (ECTC)*, 2011
36. H. Schröeder et al., Advanced thin glass Based photonic PCB integration. in *2012 IEEE 62nd Electronic Components and Technology Conference*, 2012
37. L. Brusberg et al., High performance ion-exchanged integrated waveguides in thin glass for Board-level multimode optical interconnects. in *2015 European Conference on Optical Communication (ECOC)*, 2015

38. S. Han et al., Large-scale silicon photonic switches with movable directional couplers. Optica **2**(4), 370–375 (2015)
39. K. Chen et al., OSA: an optical switching architecture for data center networks with unprecedented flexibility. IEEE/ACM Trans. Netw. **22**(2), 498–511 (2014)
40. N. Farrington et al., Helios: a hybrid electrical/optical switch architecture for modular data centers. in *Proceedings of the ACM SIGCOMM 2010 Conference* (ACM, New Delhi, 2010) pp. 339–350
41. R. Hemenway et al., Optical-packet-/switched interconnect for supercomputer applications. J. Opt. Netw. (2004).
42. C. Hawkins et al., The data vortex, an all optical path multicomputer interconnection network. IEEE Trans. Parallel Distrib. Syst. **18**(3), 409–420 (2007)
43. O. Liboiron-Ladouceur et al., *The data vortex optical packet switched interconnection network*. J. Lightw. Technol. **26**(13) (2008)
44. K. Bergman, D. Keezer, S. Wills, Design*,* demonstration and evaluation of an all optical processor memory-interconnection network for petaflop supercomputing. in *ACS Interconnects Workshop* (2010), http://lightwave.ee.columbia.edu/?s=research&p=high-performance_computing_systems#dv
45. P. Grani et al., Flat-topology high-throughput compute node with AWGR-based optical-interconnects. J. Lightw. Technol. **34**(12), 2959–2968 (2016)
46. R. Proietti et al., Scalable optical interconnect architecture using AWGR-based TONAK LION switch with limited number of wavelengths. J. Lightw. Technol. **31**(24), 4087–4097 (2013)
47. R. Proietti et al., All-optical physical layer NACK in aWGR-based optical interconnects. IEEE Photon. Technol. Lett. **24**(5), 410–412 (2012)
48. A.S.P. Khope et al., Elastic WDM crossbar switch for data centers. in *2016 IEEE Optical Interconnects Conference (OI)*, 2016
49. N. Binkert et al., The role of optics in future high radix switch design. in *2011 38th Annual International Symposium on Computer Architecture (ISCA)*,2011. IEEE
50. W. Coomans et al., Solitary and coupled semiconductor ring lasers as optical spiking neurons. Phys. Rev. E **84**(3), 036209 (2011)
51. M. Al-Fares, A. Loukissas, A. Vahdat, A scalable, commodity data center network architecture. SIGCOMM Comput. Commun. Rev. **38**(4), 63–74 (2008)
52. S. Horiguchi, T. Ooki, Hierarchical 3D-torus interconnection network. in *Parallel Architectures, Algorithms and Networks, 2000. I-SPAN 2000. Proceedings. International Symposium on*, 2000. IEEE
53. D. Abts et al., Energy proportional datacenter networks. ACM SIGARCH Comput. Archit. News **38**(3), 338–347 (2010)
54. J. Kim, W.J. Dally, D. Abts, *Flattened butterfly : a cost-efficient topology for high-radix networks*. Comput. Archit. News **35**(2) (2007)
55. J. Kim et al., Technology-driven, highly-scalable dragonfly topology. in *ACM SIGARCH Computer Architecture News* (IEEE Computer Society, 2008)
56. IDT, *RXS RapidIO Switches,* http://www.prnewswire.com/news-releases/idt-launches-next-generation-rapidio-switches-for-5g-mobile-network-development-and-mobile-edge-computing-300221151.html. https://www.idt.com/document/prb/rxs-rapidio-switches-product-brief (Integrated Device Technology, 2016)
57. W.C. Moody et al., Reconfigurable network testbed for evaluation of datacenter topologies. in *Proceedings of the sixth international workshop on Data intensive distributed computing* (ACM, Vancouver, 2014) pp. 11–20
58. C. Xuning, W. Gu-Yeon, P. Li-Shiuan, Design of low-power short-distance opto-electronic transceiver front-ends with scalable supply voltages and frequencies. in *International Symposium on Low Power Electronics and Design*, 2008
59. P. Grani et al., Photonic interconnects for interposer-based 2.5D/3D integrated Systems on a Chip. International Symposium on Memory Systems, MEMSYS16, Washington, DC, USA, 2016

60. C. Gray et al., *Multi-Gigahertz Source Synchronous Testing of an Optical Packet Switching Network*. Mixed-Signals Test Workshop, 2006.
61. A.K. Kodi, A. Louri, Power-aware bandwidth-reconfigurable optical interconnects for high-performance computing (HPC) systems. in *IEEE International on Parallel and Distributed Processing Symposium, 2007. IPDPS 2007*, 2007.
62. P.P. Dash, G. Cowan, O. Liboiron-Ladouceur, A variable-Bandwidth, power-scalable optical receiver front-end in 65 nm. in *2013 IEEE 56th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2013
63. J.E. Proesel et al., Ultra-low-power 10 to 28.5 gb/s CMOS-driven VCSEL-based optical links [invited]. J. Opt. Commun. Netw IEEE/OSA **4**(11), B114–B123 (2012)
64. X. Chen et al., Exploring the design space of power-aware opto-electronic networked systems. in *11th International Symposium on High-Performance Computer Architecture, 2005. HPCA-11*, 2005. IEEE
65. P. Grani, R. Proietti, S.B. Yoo, Benchmark analysis of AWGR-based optical tiled architectures for multi-socket HPC boards. in *2015 International Conference on Photonics in Switching (PS)*, 2015. IEEE
66. T. Benson, A. Akella, D.A. Maltz, Network traffic characteristics of data centers in the wild. in *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, 2010. ACM
67. T. Benson et al., Understanding data center traffic characteristics. ACM SIGCOMM Comput. Commun. Rev. **40**(1), 92–99 (2010)
68. M. Awasthi et al., System-level characterization of datacenter applications. in *Proceedings of the 6th ACM/SPEC International Conference on Performance Engineering* (ACM, Austin, 2015) pp. 27–38

Chapter 2
Optical Switching in Datacenters: Architectures Based on Optical Circuit Switching

Liam P. Barry, Jingyan Wang, Conor McArdle, and Dan Kilper

2.1 Introduction

The proliferation of modern computer applications such as cloud computing, social networking services, multimedia streaming, and the Internet of Things (IoT) creates thousands of terabytes of heterogeneous data every minute [1], creating huge volumes of complex workload in datacenter servers and vast aggregate traffic loads across datacenter networks. Internet content providers such as Google, Facebook, Microsoft, Amazon, and Apple have installed mega datacenters which house hundreds of thousands, even millions, of servers in very large-scale layouts. As well as this scale-out, server network interfaces and top-of-rack (ToR) switches are being scaled up, from 10 to 25 Gbps and 100 Gbps data rates, to keep pace with the required workload rates and increasing service capacity [2]. Big data and cloud computing in particular have been shifting datacenter workloads from north-south to east-west, meaning that more data is flowing within the datacenter than in and out of the datacenter, driving up the internal datacenter network capacities. These trends are presenting substantial challenges to future-proofing of datacenter-scale networking, in terms of the required latency, bandwidth capacity, server connectivity, energy and cost efficiency, network configuration, and control complexity. Furthermore, as datacenters run increasingly mixed applications with more complex workloads, traffic complexity will evolve rapidly and exhibit more diverse and unpredictable communication patterns [3–7]. Thus, dynamic bandwidth-on-demand provisioning needs to be made possible for future datacenter network fabrics. Solutions toward highly scalable, high data rate, low-latency, agile datacenter

L.P. Barry (✉) • J. Wang • C. McArdle
School of Electronic Engineering, Dublin City University, Dublin 9, Ireland
e-mail: liam.barry@dcu.ie

D. Kilper
College of Optical Sciences, The University of Arizona,
1630 E. University Blvd., Tucson, AZ 85721-0094, USA

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
DOI 10.1007/978-3-319-61052-8_2

network infrastructure, with on-demand service provisioning capacity that responds quickly and dynamically to changing application requirements, are urgently required in order to maintain communication performance for building-scale and distributed multisite server installations of the future.

Today's large-scale datacenter networks are constructed by interconnecting a massive number of commodity hardware components, such as low-radix electronic packet switches, transceivers, and optical-electronic-optical (O/E/O) converters and optical fibers, into multi-tier interconnection topologies. Datacenter networks are typically organized as three-layer fat-tree [8] or Clos [9] topologies, where racks of servers are attached directly to a network access layer composed of top-of-rack (ToR) switches. A network aggregation layer collects traffic flows from the access layer and then forwards them to a core layer. Commonly, a datacenter network employs a Clos-based topology with centralized control, where commodity switches based on merchant silicon are arranged into multiple stages to form a large multistage switching fabric [10]. For example, Google has exploited a modular design concept, where a pool of machine servers and network devices are grouped into a cluster networking block to enable fine-grained resource provisioning and boosted operation efficiency [10, 11]. Facebook's current datacenter networks follow a hierarchical treelike networking model composed of three layers of electronic packet switches, but are now migrating toward a three-layer folded-Clos networking model with no oversubscription between racks [12, 13]. This network infrastructure is also designed with a modular structure using clusters or pods. Furthermore, a new two-tier datacenter network platform, the leaf-spine architecture, is emerging as an alternative to the traditional three-tier network topology. The leaf-spine architecture is a two-layer folded-Clos network comprising leaf-and-spine layers with each leaf switch connected to each of the spine switches [14]. Other electronic network architectures such as DCell [15], BCube [16], and FiConn [17] have been proposed to alleviate the bottlenecks of the three-layer fat-tree networks by building a recursive interconnection architecture.

The infrastructure scaling of these hierarchical networks, to accommodate the ever-growing number of servers, is potentially viable by adding more commodity network components and increasing the number of switching stages. For example, a new Facebook datacenter network with 100,000 servers placed in racks of 48 servers requires 2084 rack switches, each configured with 64 × 10 Gbps port capacity and 176 aggregation switches, each with 384 × 10Gbps port capacity and 192 core switches, and each of which has at least 176 × 10Gbps port capacity. Networking such a large number of servers, using electronic packet switches and links, imposes significant challenges on performance scalability, energy consumption cost, flexibility, network configuration, and management complexity. These issues are further exacerbated by the unprecedented datacenter traffic growth and the exploding bandwidth demands, which are rapidly outpacing the capacity of the state-of-the-art electronic switches, thereby necessitating the scaling of data rates, connectivity, and bandwidth capacity in the electronic switches. Electronic switch design currently faces switch port and capacity constraints due to high costs and excessive amounts of power consumption [18]. Ultimately, scaling of electronically switched datacenters calls for a significant upgrading/replacement of the interconnection networks in order to support heterogeneous hardware and speeds. Furthermore, with server capacity scaling and communication speed rising, the higher interconnection level of the multilayer datacenter topology faces substantial resource stranding. That is, the aggregated traffic demands at higher levels of the system hierarchy (aggregation and core layers) potentially require progressively larger bandwidth capacities and consequently put the statically provisioned and oversubscribed upper tiers under strain. These limitations are stressing the existing datacenter networks and motivating the network to evolve toward more advanced networking solutions, augmenting innovative networking technologies, topological structures, and network management techniques.

A constructive approach toward the aforementioned design challenges involves application of optical switching and networking technologies. The most important attribute of optical switching is ultrahigh bandwidth capacity, which would provide unpreceded data carrying capacity for datacenter networking. Additionally, optical switching is potentially a low-latency communication technology, possibly achieving nanosecond-scale switching speed [19]. Further, optical switching can potentially future-proof datacenter network infrastructure, due to transparency to data rates, modulation formats, and protocols, and thus, except at the edge, does not require upgrading/replacement of the underlying physical network components as network link capacities evolve from 10 to 25 Gbps and 100 Gbps and beyond. Low power consumption is another prominent feature of optical switching technology. For example, a commercial 320 × 320 3D MEMS-based optical switch typically consumes 45 watts [20], whereas an Ethernet packet switch with a maximum 224 × 10Gbps ports requires a typical power consumption of 1363 watts [21]. This contrast suggests that, compared to electronic switching, optical switching potentially offers substantial economic and technical benefits. Also at the link level, all-optical switching eliminates the need for expensive, power-consuming optical-electronic-optical (O/E/O) conversions which are necessary to deploy point-to-point optical links in electronically switched networks. The advantages of low loss, low power, and the high capacity of optics, coupled with the rapid development of photonic integration technology, highlight the opportunity to fabricate compact, high-stability, low-cost, and energy-efficient large-radix optical switching fabrics. Deploying high-radix optical switches in massive-scale datacenter networks would significantly flatten the network topology and simplify the network architecture and control complexity. From the network management perspective, software-defined networking (SDN) is becoming progressively common in datacenter networks and in the optical switching domain in general. This emerging trend facilitates the realization of an agile, flexible, and scalable optical datacenter interconnection network, with SDN enabling coordinated control of network resources and support for on-demand capacity provisioning by flexibly allocating network capacities to dynamically changing application demands. The abovementioned benefits indicate that an agile optically switched datacenter network has the potential to address current and future datacenter application demands, making it a suitable substitution for electronic networks. Indeed, research effort in this area has seen considerable growth in recent years.

In this chapter, the main focus is on reviewing the previous approaches to integrating mature optical circuit switching technologies into datacenter networks, to facilitate near- and medium-term expansion of datacenter and cloud computing capacity. In these approaches, high-radix optical circuit switches are interconnected into massive-scale datacenter interconnects using various fundamental networking structures such as Clos [9], Spanke [29], Benes [30], and flattened butterfly [31]. The current proposals are reviewed and discussed in the next section. Following that, a novel, agile datacenter network architecture, which supports dynamic and efficient sharing of network resources and flow-level provisioning, is presented. By exploiting the potential benefits of the passive fast-speed, low-radix flexible arrayed waveguide grating (AWG) switches coupled with high port-count optical circuit switches (OCSs), the proposed dynamic network offers the prospect of building a highly scalable, highly flexible, efficient large-scale datacenter network. We conclude the chapter with a perspective on how the current trends in OCS-based networks are expected to shape the future directions and capabilities of datacenter networking.

2.2 Optical Circuit Switching in Datacenter Networks

Optical switching technologies can be generally classified into two categories, optical circuit switching (OCS) and optical packet (or burst) switching (OPS), depending on the switching granularity. Optical circuit switching (OCS) is a relatively mature commercially available technology. Typically, an OCS is built based on optical 2D or 3D microelectromechanical systems (MEMS) technology. Currently, commercial 3D MEMS OCS supports up to 320 ports [20], and Polatis beam-steering OCS supports 384 ports [22], with high capacity, low loss, and low energy consumption. Port counts of more than 1000 have been offered previously for telecom applications and may become viable again for datacenters. OCS is a coarse-grained switching technique operating at the granularity of a full optical fiber [23]. That is, a dedicated point-to-point optical connection between an input and output fiber pair is set up for the entire data transmission session, providing guaranteed uncontended bandwidth [24] and ensuring quality of service (QoS). Nonetheless, the circuit-oriented configuration of the 3D MEMS-based OCS exhibits slow switch reconfiguration time, in the order of milliseconds. Microsecond switching times have been reported in experimental demonstrations of smaller-port-count optical space switches, such as a 64 × 64 compact silicon photonic MEMS module [25] and a 24 × 24 OCS built of 2D MEMS wavelength selective switch (WSS) modules [34]. Slow OCS potentially creates significant configuration overhead for traffic flows, and, as such, high-capacity optical circuit switching has previously been proposed for transferring only highly aggregated bulk datasets where the data transmission period is significantly longer than the switch setup time overhead. Slow switching time optical circuit switches exhibits relatively low flexibility and potentially low average utilization if employed in lower network tiers, due to the inability to efficiently handle dynamic, bursty traffic streams. In comparison, optical packet switching (OPS) is a fine-grained, flexible switching paradigm which supports fast optical switching at packet (or burst) level. Optical packet switching is expected to be able to fully exploit the advantages and potential of optical switching, and OPS has been extensively studied in a number of large-scale research projects including WASPNET [26], LIGHTNESS [27, 28], and in [19]. However, the required optics and photonic technologies are generally considered not yet mature enough to fabricate a commercial large-scale optical packet switch. The main technical challenges faced by OPS involve high-speed optical packet header processing, lack of efficient optical packet buffering mechanisms, optical wavelength conversion and regeneration, scalability, and reliability. These difficulties would seem to make the wide deployment of OPS in commercial, cost-sensitive networks unachievable in the very near future. The deployment of optical circuit switches as optical cross-connects (OXCs) in datacenter-scale networking has instead been considered the first step to realizing scalable, transparent optical datacenter infrastructure. Extensive research efforts have been devoted to expanding the applicability of OCS from high-speed core telecommunication networks (which is implemented with wavelength switching) to datacenter networking. The major research directions include hybrid electronic/OCS datacenter networks [32–38]; microsecond OCS networks [34–36]; wavelength-, space-, or time-division multiplexing (WDM/SDM/TDM) OCS datacenter networks [39–41]; software-defined networking (SDN)-controlled OCS networks [20, 22, 39–44]; flexible fixed-grid OCS networks [45, 46]; elastic OCS datacenter networks [48, 49]; hybrid OCS/OPS interconnects [27, 28, 51]; hybrid wireless/wired datacenter networks; and various combinations of these emerging trends.

Hybrid electronic/OCS networks [32, 33] integrate advanced OCS technology into existing well-established electronic packet switching (EPS) network architecture, with EPS accommodating short-lived bursty traffic and slow, high-port-count OCS targeting large data transfers. In OCS-based datacenter network design [34–36], the driving goal is to achieve relatively fast microsecond switching speeds. This flexibility allows the OCS to efficiently handle more dynamic traffic patterns and route a larger fraction of the datacenter traffic, in comparison to traditional OCS networks which are expected to switch only in the millisecond range. Advances in WDM/SDM/TDM transmission and switching technologies motivate the exploitation of these technologies in datacenter networking [39–41]. Deploying WDM/SDM/TDM technologies in high-capacity OCS networks has led to significant benefits regarding network capacity, flexibility, and scalability. This flexibility is supported by SDN, a unified network control and management paradigm, which decouples the network control plane from the underlying switching and routing data plane, places the network intelligence into a logically centralized control system, and applies software-programmable functionality in the networking devices to facilitate the deployment of new network applications [42]. SDN promises network configuration optimization, high-level network flexibility, efficient capacity utilization, and guaranteed application performance.

Flexible fixed-grid OCS networks [45, 46] attempt to promote wavelength assignment capacity by augmenting the OCS with flexible fixed-grid optical wavelength switching and scheduling technologies, i.e., wavelength selective switching (WSS), which allow an arbitrary wavelength to be switched to any output port of a WSS device. To further incorporate an even higher level of network elasticity, elastic OCS datacenter networks [48, 49] supporting arbitrary modulation formats and dynamic optical spectrum allocation with channel capacity ranging from subchannels to super-channels have been designed to optimize flexibility and efficiency. Transparent hybrid OCS/OPS datacenter networks [27, 28, 51] rely on both OPS and OCS to realize the central all-optical switching matrix. This scheme aims to combine the merits of OCS and OPS so as to support finer switching granularity. Another emerging trend is the possible deployment of the above-described wired network and the advanced wireless technologies such as free-space optics (FSO) [52] and 60 GHz wireless technology [53–54]. Recently, a MEMS-based approach was used to implement a high-radix OCS in free space across a datacenter [55]. With radix above 10,000, this approach enables a flat architecture with the benefits of OCS switching. In this section, several OCS-based datacenter architectures in the research literature are reviewed, and the opportunities and challenges of these proposals are discussed.

c-Through [32] and Helios [33] are two examples of hybrid EPS/OCS datacenter networks. The system-level structures of the c-Through and Helios networks are illustrated in Figs. 2.1 and 2.2, respectively. These hybrid network configurations combine a traditional hierarchical electronic packet switching network with a complementary optical circuit switching network, with the EPS network mainly managing small traffic flows and the OCS supporting large-volume, aggregated data transfers with guaranteed bandwidth. As demonstrated in Fig. 2.1, the c-Through electronic network follows a three-tier spanning-tree topology, whereas Helios deploys a two-level multi-rooted networking structure (Fig. 2.2). The optical circuit-switched network is built of slow, high-capacity MEMS-based OCS which provides direct interconnections between access switches. Helios also makes use of wavelength-division multiplexing (WDM), which is highly advantageous as it expands the capacity and flexibility of optical circuit provisioning. An important implication of the hybrid designs is that the collaborative operations of two independent switch fabrics to optimize network resources necessitate the need for traffic estimation and demultiplexing, which is conducted in c-Through end hosts and Helios switches, respectively. The hybrids c-Through and Helios offer the prospect of efficient networking by exploiting the best of EPS and OCS technologies, and, to some extent, they alleviate the oversubscription, reliability, scalability, flexibility, and capacity issues in current datacenters without requiring a complete replacement of hardware equipment (Fig. 2.3).

The Helios research group has further designed Mordia [34, 35] and REACToR [36]. Mordia is a prototype demonstration of a hybrid EPS/OCS network. One of the most significant characteristics of Mordia is that it adopts microsecond OCS networking, which is designed based on multiple 2D MEMS wavelength selective switches (WSSs). The microsecond OCS supports substantially higher switching speed, 2–3 orders of magnitude faster than commercial 3D MEMS OCS, and thus can potentially mitigate the slow switching issue and high buffering/aggregation requirement in traditional hybrid EPS/OCS datacenter networks. The topological structure of the Mordia OCS prototype follows a unidirectional ring supporting all-to-all connectivity and arbitrary input/output mapping. Additionally, a novel control algorithm called traffic matrix scheduling (TMS) is proposed, which is compatible with the

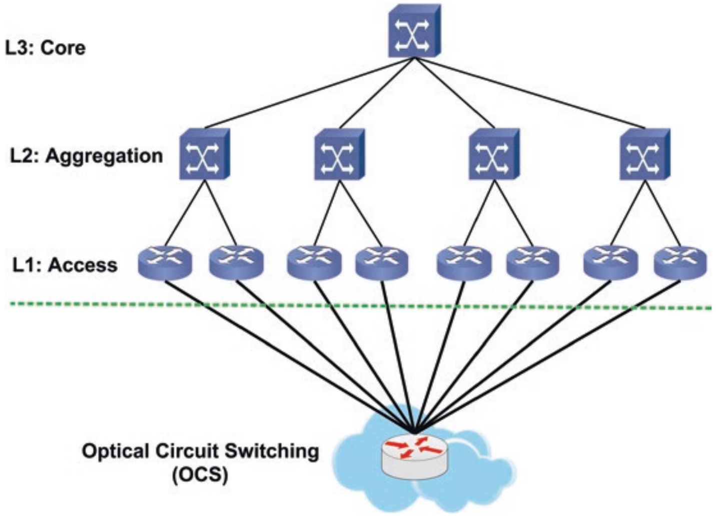


Fig. 2.1 Hybrid c-Through network

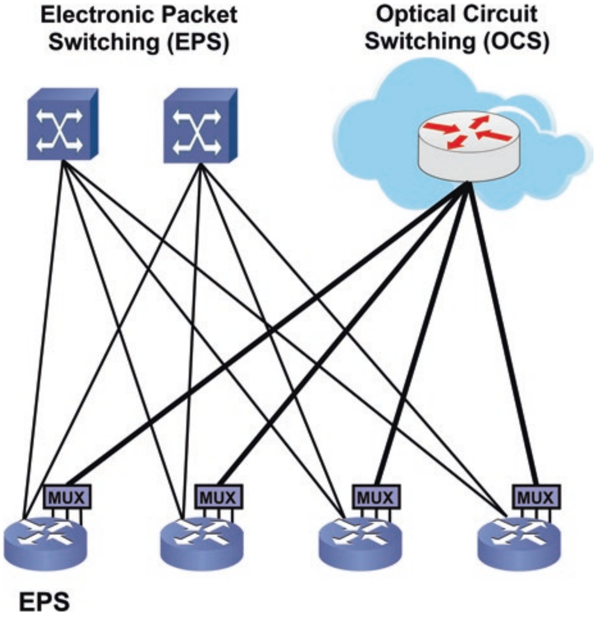


Fig. 2.2 Hybrid Helios network

microsecond-latency OCS. The TMS predicts traffic demands and then schedules short-time circuits to carry these demands. The short-lived traffic scheduling, coupled with microsecond circuit switching, greatly boosts the switching flexibility and capacity, making the OCS react rapidly to changing traffic patterns. The 24-port architectural demonstration is tested on a small-size network comprising 23 nodes.

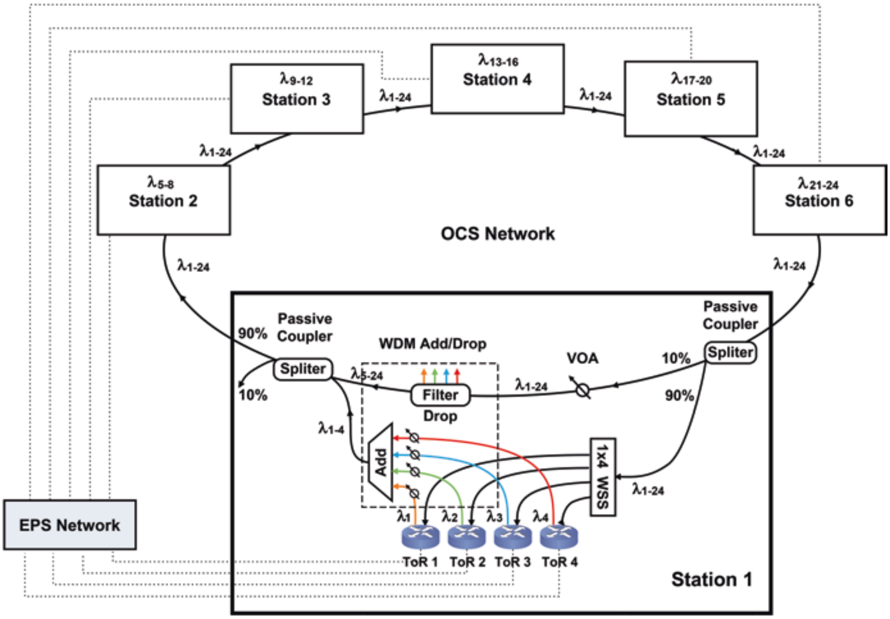


Fig. 2.3 The Mordia network [34, 35]

In large-scale datacenter networks, the implementations of the WSS-based OCS and its corresponding control algorithm may not be sufficiently scalable, due to the fact that the Mordia WSS-based OCS is a DWDM ring assigning one wavelength per port. As a consequence, scaling beyond 88 DWDM channels imposes significant technical and economic challenges.

The ideas behind Mordia, using WSS-based OCS delivering optical circuit switching at microsecond speeds, bring OCS closer to emulating electronic packet switching speeds, capable of supporting more dynamic traffic patterns, compared to the commercial OCS. Motivated by this, the prototype hybrid REACToR architecture [36] has been demonstrated on the basis of the Mordia datacenter network. In REACToR, the buffering and scheduling are mainly performed in the end hosts where the data packets are aggregated, grouped, and stored in transmission queues on a per-destination-ToR basis. Once an optical circuit path has been set up from the REACToR to an end host, the REACToR schedules the appropriate source end host to constantly pump data packets into the network through the established end-to-end optical circuit. In this architecture, direct, high-capacity, flow-level optical connections are provisioned between end hosts to serve high-volume server-to-server traffic, rather than serve rack-to-rack traffic and being shared by many flows as in traditional hybrid networks. The high-speed WSS-based OCS can be reconfigured to adapt to the changing traffic patterns of datacenter applications. During the reconfiguration times of the OCS, the EPS network is used to transmit host-to-host data traffic. This hybrid operation ensures non-degraded performance of the core network.

Thus, by integrating high-speed OCS supporting microsecond switching times with the existing high-speed, flexible EPS network, the REACToR hybrid EPS/OCS network, coupled with an efficient control plane, functions similarly to packet-switched ToR switches and supports fine-grained scheduling and thus can effectively schedule dynamic, rack-level traffic demands and achieve high link bandwidth utilization. This facilitates an upgrade path for existing EPS networks to higher data rates, higher flexibility, and better performance with significant economic benefits. Nevertheless, in addition to the previously described design challenges of a large-scale Mordia OCS network, a major problem that remains is how to efficiently interconnect a large number of the REACToR switches in a massive-scale datacenter, with effective global scheduling, control, and synchronization.

The single-stage shuffle-exchange (SSX) architecture [37, 38] is another hybrid EPS/OCS datacenter architecture. Unlike c-Through and Helios, the SSX scheme augments EPS and OCS in the same interconnection framework where the OCS functions as the central switching fabric that directly interconnects edge EPSs, as shown in Fig. 2.4. The SSX architecture operates on the principle that the *exchange* is performed first and then the *shuffling*. More precisely, an optical signal injected to an EPS is first switched to the appropriate output port of the EPS which directly leads to an input port of the OCS, and then, through the OCS, the optical signal is delivered to the destination EPS.

With the development of high-radix OCS, the SSX architecture now has the port densities required for large-scale datacenters. Nevertheless, the MEM-based OCS, as discussed previously, exhibits coarse-grained switching granularity and long reconfiguration times, which partially offset the benefits of high-capacity optical switching. To avoid constant reconfiguration and to expand the connectivity of the OCS, SSX deploys a hop-by-hop routing strategy which allows communications between two EPSs which are not currently being interconnected by the OCS. However, to do this, the input signal needs to travel through multiple EPSs and undergoes multiple optical-electronic-optical conversions before reaching its destination EPS, which consequently imposes increased communication latency and power consumption. Hence, a trade-off is presented by the SSX architecture.

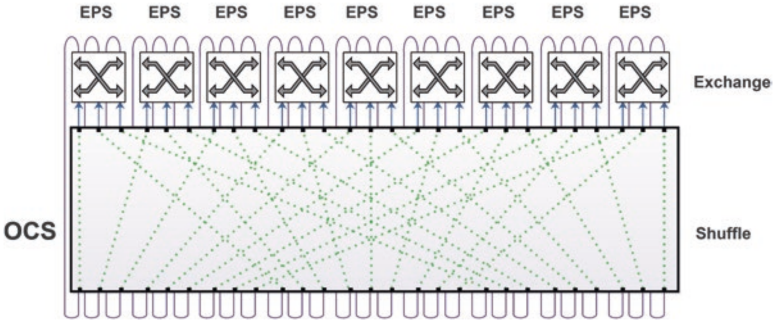


Fig. 2.4 Hybrid SSX architecture [37, 38]

[39, 40] propose a converged EPS/OCS datacenter network with SDN control, where an all-optical circuit-switched core network interconnects ToR electronic switches (EPSs), as illustrated in Fig. 2.5. The OCS network is designed with a flattened butterfly topology [31], where the basic building modules—optical virtual switches (OvS) enabled by optical wavelength switching—are arranged into a 2D array. In the OvS architecture, the optical wavelengths injected from the transceivers in a ToR are combined into a DWDM channel by an optical multiplexer (MUX) and subsequently forwarded to two passive tap coupler tree modules called passive routing fabric (PRF) blocks, which broadcast the DWDM signal in two dimensions. The receiving block of the OvS is mainly based on an optical wavelength switching component, i.e., WSS, which is flexibly configured to allow the desired wavelengths to pass through and block all other wavelengths. The passing wavelengths are then routed through an optical demultiplexer (DEMUX) toward the ToR. Essentially, the OvS is a broadcast-and-select (B&S) building block exploiting the advantages of

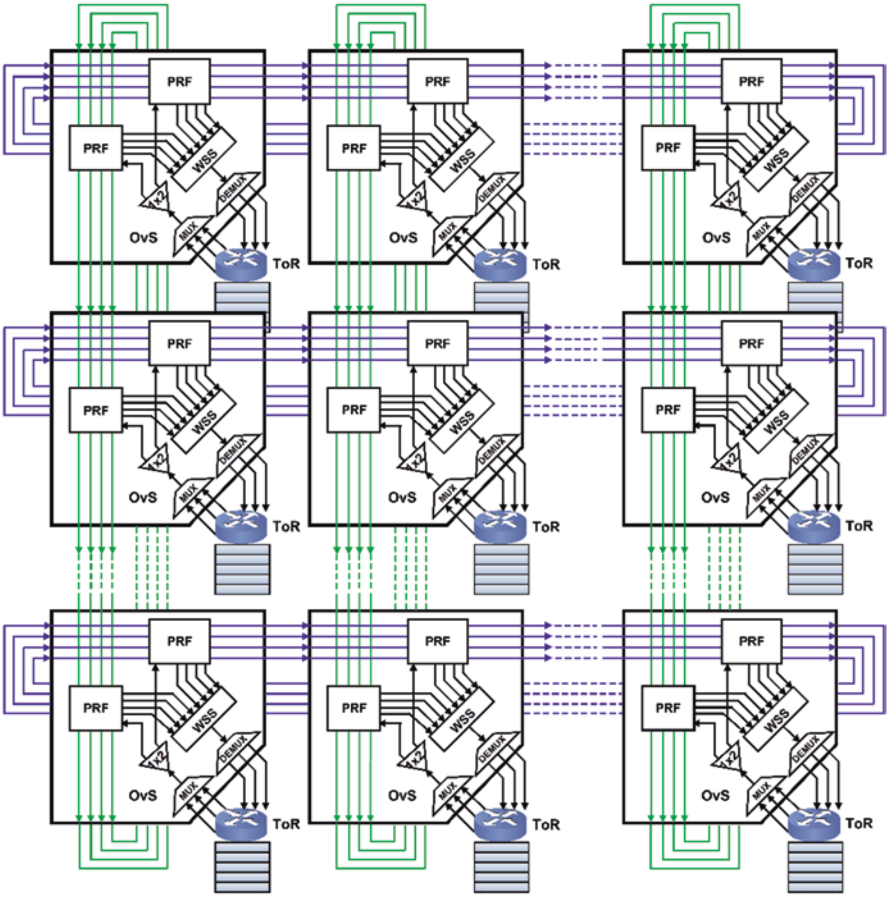


Fig. 2.5 The DWDM/SDM OCS network [39, 40]

DWDM, space-division multiplexing (SDM), and flexible optical wavelength switching technologies. The 2D-positioned OCS network, when utilized in conjunction with the ToR EPS, supports fully meshed interconnectivity among all rack switches, thus facilitating building large-scale networks with good system performance. Particularly, the integration of the B&S architecture and dynamic optical wavelength switching yields high network flexibility, and as such the datacenter network can dynamically support diverse traffic patterns such as multi-cast, in-cast, and all-to-all cast. Nonetheless, the end-to-end reconfiguration speed largely relies on the optical wavelength switching speed, which could be multi-milliseconds with current technologies. In addition, the further scaling of the network is limited by the number of DWDM channels, the port density of the optical wavelength switching component (WSS), and the costs of the DWDM transceivers.

Archon [41] is a transparent optical circuit-switched intra- and inter-datacenter network which adopts solely optical switching for core cross-connection and eliminates electronic switching, as shown in Fig. 2.6. In comparison with the DWDM/SDM datacenter architecture in [39, 40], the Archon network takes advantages of both space- and time-division multiplexing (SDM/TDM) technologies to provide high-capacity interconnection. In Archon, racks of servers are organized into clusters, with each cluster comprising a group of racks. Inside a cluster, all ToR switches are directly connected to a large-scale, high-capacity OCS via high-capacity SDM (multielement fiber (MEF)) optical links. The intra-cluster OCS is mainly used to handle long-lived bulky intra-cluster data flows. To efficiently switch short-lived, bursty intra-cluster traffic streams, an additional high-speed TDM switch with relatively low bandwidth capacity is adopted to support intra-cluster communications with variable link capacity. The TDM switch, as well as amplifiers, optical splitters, and optical couplers, is flexibly connected to a group of ToR switches through the reconfigurable intra-cluster OCS. Alternatively, the intercluster network is constructed by directly interconnecting the intra-cluster OCSs through a large-port-count intercluster OCS via SDM optical links. As for the inter-datacenter connectivity, the SDM signal from the central OCS is routed to the metro/core network through a SDM-to-WDM converter. In the Archon architecture, the unstable, shifting intra-cluster traffic is supported by two complementary switching paradigms, a high-capacity, slow OCS and a low-capacity, flexible TDM switch, while the relatively stable aggregated intercluster traffic is transferred over a centralized flexible OCS. The highly integrated SDM technology promises high capacity, high scalability, and simplified fiber connection complexity. However, as each cluster is equipped with a high-port-count OCS, a large-scale datacenter network composed of many clusters requires a large number of OCSs, which potentially results in high network costs. Another challenge is the practical implementation of the TDM switches and the slow-speed, limited flexibility intra-/intercluster OCS to support excessive traffic demands and heterogeneity in datacenter-scale networking, limiting the network's ability to scale to very large sizes.

Innovative SDN-enabled OCS datacenter architectures have been demonstrated, such as Calient [20], Polatis [22], proposal in [39, 40], Archon [41], SDN implementations in hybrid OCS/EPS datacenters [43], and C-Share [44]. Calient's SDN-based

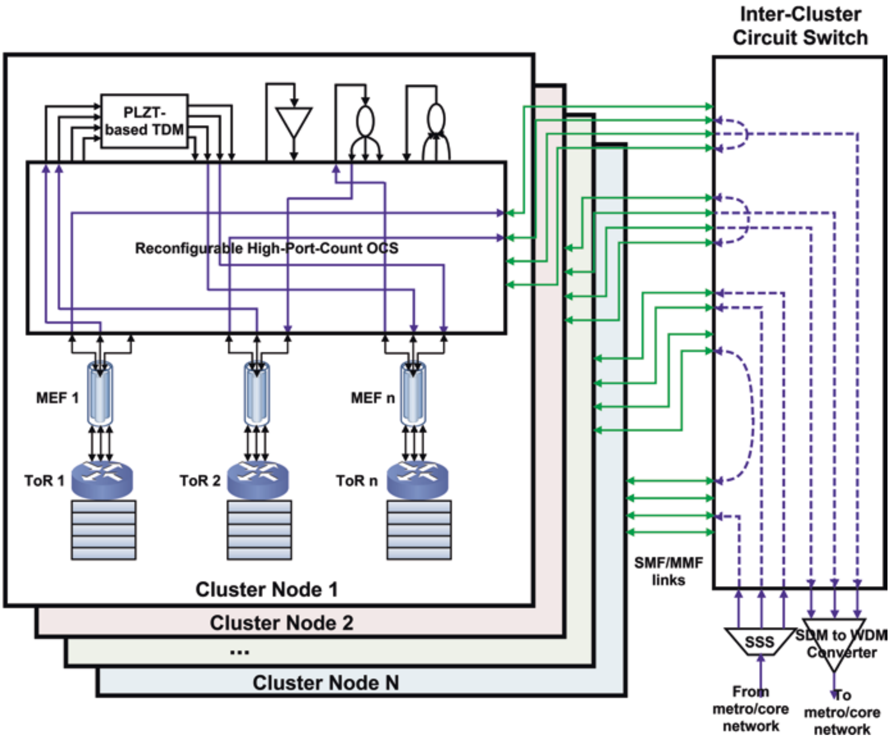


Fig. 2.6 The SDM/TDM/WDM Archon architecture [41]

3D MEMS OCS [20] with up to 320 ports allows dynamic reconfiguration of the optical circuit switching matrix, thereby further accelerating the deployment of the hybrid EPS/OCS datacenter architectures. Polatis has recently launched commercial 384 × 384 software-defined OCS [22] for use in building high-performance optical datacenter networks. The Polatis OCS supports the highest port density commercially available today in all-optical switching fabrics and promotes fully non-blocking all-optical switching operations. The SDN-enabled C-Share network [44] makes efficient use of optical circuits by envisioning an SDN-based circuit sharing scheme in the flattened leaf-and-spine EPS/OCS network topology. If a shared optical circuit is provisioned from a source endpoint to a destination endpoint, the SDN-based traffic rerouting strategy redirects the elephant flows from other endpoints to this source endpoint, such that these large flows can be transmitted through the established optical circuit toward the destination endpoint, enabling optical circuit sharing among multiple elephant flows. This novel integration scheme of OCS into datacenter networks facilitates the small/large flow separation, off-loads the workloads from the EPS network onto optical circuits, and allows small flows to be efficiently transmitted through the core EPS network with alleviated congestion, thereby yielding an improved overall network performance.

As an alternative, fixed-grid OCS datacenter architectures utilizing fixed channel spacing of 50 or 100 GHz have been designed and demonstrated, with the aim to support dynamic topology and link capacities. A prominent example of this category is OSA [45, 46], where the backplane of the network features a high-port-count OCS and reconfigurable wavelength selective switches (WSSs). In OSA, a rack switch can communicate with one or multiple rack switches simultaneously using a number of wavelength channels through the OCS. By exploiting dynamically reconfigurable WSS technology, these wavelength channels are adaptively assigned to the communication requests based on their traffic demands. This flexible bandwidth assignment scheme introduces a high degree of network flexibility and boosts communication performance. Similar to the SSX architecture, OSA also employs a multi-hop routing policy so as to alleviate the inherent connectivity limitations of the OCS, but potentially at the expense of increased routing complexity, communication latency, and optical-electronic-optical conversions.

Elastic OCS-based datacenter networking is a more flexible networking concept where flexible-grid OCS modules are deployed to set up adaptable wavelength channels [47–49]. The network elasticity is mainly reflected in two aspects: mixed line rates and multiple modulation formats. Elastic OCS networks support adaptive optical spectrum assignments with almost an arbitrary spectral width, which allows optical signals to be transmitted with a wide range of line rates depending on traffic requirements. Alternatively, the adoption of adaptive modulation formats gives the network the ability to change the signal's modulation formats according to transmission distances and traffic profiles [47]. The very fine-grid spectrum allocation and variable modulation formats together optimize the spectrum resource utilization and facilitate on-demand datacenter service provisioning. The STRAUSS project [48], the Archon architecture [41], and the elastic ring-based OCS datacenter network proposed in [49] focus on introducing elasticity in optical circuit provisioning to support finer and multiple granularities in large-scale networks. The key enabling technologies include bandwidth-variable transceivers; bandwidth-variable OCS components, i.e., bandwidth-variable WSS; and SDN control technology [50]. Compared to the conventional fixed-grid OCS networks, the flexible-grid OCS designs future-proof the network with respect to channel baud rate and modulation formats, and enable greatly enhanced capacity, spectrum efficiency, and flexibility, which could better serve diverse datacenter applications. However, the flexible-grid software and hardware development, the design of an efficient spectrum allocation algorithm, and the control complexity involved need to be resolved [47]. Advances in integrated photonics are needed to bring the cost and energy of these components down to levels required for datacenters.

The hybrid OCS/OPS datacenter interconnect is another important area of research—replacing the EPS with an optical solution. An example of a hybrid OCS/OPS interconnect is the SDN LIGHTNESS architecture [27, 28] with highly distributed control, which has been proposed by the EC FP7 LIGHTNESS project. In this network configuration, the core layer includes transparent OCS and OPS networks, both providing fully meshed interconnectivity among rack switches. In [51] an energy-efficient torus datacenter structure combining OPS and agile OCS, where a novel flow management scheme is deployed to support optical path on-demand, is introduced. Nevertheless, a cost-effective hybrid OPS/OCS network is still elusive and faces substantial challenges related to optical packet switching technologies, traffic prediction and classification, scalability, complexity, and costs.

An alternative to the hybrid packet and circuit switching is the new “seamless” switching approach used in the ProjecToR architecture [55]. In this case electronic switching and aggregation are performed entirely at the ToR, and all connections above the ToR are implemented using an OCS space switch. With high port counts in the OCS, including multiple ports per ToR, the OCS switch can be thought of as a set of reconfigurable fibers between the ToRs—carrying all traffic elephant and mice. The majority of OCS ports are reconfigured on a slow time scale of hours or even daily, while a small subset is reserved for cases in which the fly reconfiguration is needed—such as long-lived deterministic elephant flows or congestion relief. Multi-hop routing is used in the ToRs when needed as well. This seamless switching approach was adopted in part because analysis of Microsoft production datacenter traffic showed that separating elephant flows through traffic analysis is not feasible based on the measured statistics [55].

It is worth mentioning that development in free-space optics (FSO) [52, 55] and 60 GHz [53–54] wireless technologies enables the possible implementation of wireless links combined with OCS technology to realize a high-capacity, dynamically reconfigurable hybrid wireless/wired datacenter network. The wireless links hold significant advantages with respect to interconnectivity, power consumption, and low cabling complexity. Using MEMS switching through optical diffraction in free space, as opposed to just reflection, has been further shown to enable port counts exceeding 10,000 with switching transition times in the microsecond range. These free-space technologies can be envisioned to provide additional dynamic bandwidth capacity to relieve the congested hotspots in datacenter networks.

2.3 Agile Optical Datacenter Network Architecture

Having considered recent advances in the application of optical circuit switching to future large-scale datacenter networks, in this section we propose a novel architecture for a highly agile optical datacenter network which supports adaptive topologies and efficient sharing of network resources at flow level to provide a dynamic configuration response to changing datacenter application and traffic requirements. By exploiting the potential benefits of the passive, low-radix flexible arrayed waveguide grating (AWG) switch together with high-port-count optical circuit switches (OCSs), the proposed dynamic network offers the prospect of building a highly scalable, highly flexible, efficient large-scale datacenter network. It also supports pulling the metro network into the datacenter network using the OCS fabric for efficient metro implementations.

The proposed network infrastructure follows a flat topology comprising two layers of switch elements (Fig. 2.7). Server racks are directly attached to the top-of-rack (ToR) switches, which provide server-to-server connectivity within racks. ToR switches are connected to the large-scale reconfigurable optical space switch (OSS) which is constructed by interconnecting optical circuit switches (OCSs) with hundreds of ports in a multistage topology such as Clos [9], Spanke [29], Benes [30], or flattened butterfly [31]. On top of the large OSS lies an additional layer of high-speed switching modules including optical wavelength-routed arrayed waveguide grating (AWG) switches, direct fibers (DFs), and small electronic packet switches (EPSs). Metro ROADM switches can also be incorporated directly above the OSS. To make the most of the high-capacity optical circuits and the high-speed interconnect connections such as AWGs, DFs, and EPSs, so as to efficiently serve diverse network-wide demands, a large OSS is appropriately configured to connect the high-speed switching modules, mainly the AWG switches, with a collection of frequently communicating ToR switches, based on traffic demand estimations over relatively long time scales. For that, the designed optical network is given the ability to support circuit-switched provisioning carrying flows of packets, thereby achieving high-capacity, high-speed all-optical switching. Importantly, by flexibly reconfiguring the OSS, various combinations of network connections can be established, thus adapting the embedded network topology to the changing traffic patterns. Strictly speaking, the proposed novel dynamic datacenter network is akin to a flattened two-layer leaf-and-spine architecture consisting of a ToR switch layer and a fast optical switching layer with the layers interconnected via the OSS. This could be implemented using the ProjecToR free-space switching for the OSS function [55]. Because of the high port counts in ProjecToR, the OSS can be flattened and simplified.

Figure 2.7 demonstrates the optical schematic design of the modular datacenter network where the large reconfigurable space switch (OSS) logically groups the ToR switches into N clusters, each comprising M ToR switches and a $P \times P$ fast optical AWG switch. In total, this distributed network hosts $M.N$ ToR switches. In a multi-rack cluster, each of the M ToR switches is equipped with one or multiple tunable optical transmitters and then directly connects to an input/output port pair of the $P \times P$ fast AWG switch through the reconfigurable OSS. The tunable optical transmitters and the $P \times P$ AWG switch together feature a reconfigurable, high-speed wavelength router, where the tunable optical transponder tunes the traffic flows originated from the connected ToR switch on the appropriate wavelengths based on the cyclic wavelength routing characteristic of the AWG switch, and subsequently the AWG switch directly routes these traffic flows to their destined AWG output ports. Of particular note is that the wavelength-routed AWG switch allows an optical signal entering from any AWG input port to be switched to any AWG output port by accordingly reconfiguring the output wavelength of the associated tunable transmitter [56]. It also allows multiple optical signals carried on different wavelengths injected from different ingress ToRs to be transmitted concurrently to an egress ToR. In doing so, a fully interconnected cluster network is realized, which moves data across the M ToR switches within the cluster quickly. A local cluster AWG controller is adopted to schedule the traffic flows, configure the tunable transmitters, and manage the network resources. Connections through the AWG are shown as duplex channels with a forward and reverse path through the AWGs. Although many applications will generate highly asymmetric traffic and may not require a high-capacity duplex connection as shown, the reverse channel is included here so that any handshaking or acknowledgments (ACKs) have a low-latency connection available to them so that a traffic flow does not stall due to delays in the

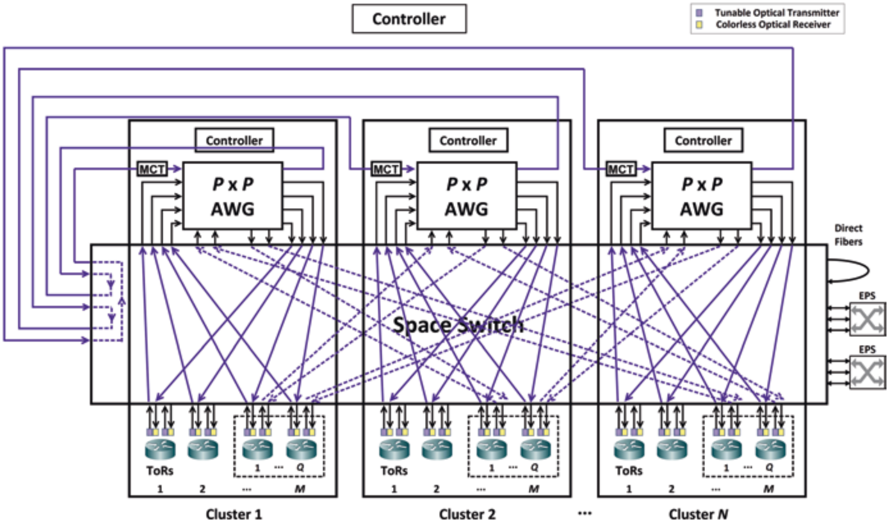


Fig. 2.7 The agile datacenter network architecture. The datacenter network is composed of *N* clusters. Each cluster contains *M* ToR switches and a *P* × *P* fast optical AWG switch. The number of cluster heads in each cluster is configurable and denoted *Q*

reverse channel. The direct fiber (DF) connections shown in the architecture are shown as simplex channels for simplicity and may be used by applications that require long-lived, high-capacity, low-latency links in one direction. The reverse channel may be carried by the electrical network to allow control of this direct fiber connection. Applications that may use such links are typically storage area network applications that may have traditionally been based on Fiber Channel (FC) links. One advantage of this approach is that the AWG switching fabrics can be replaced by metro ROADM switches to connect to a metro network. The wavelength selective switches used in ROADMs can be used to provide the same functionality as the AWG and thereby flexibly set up connections between nearby datacenters.

The *N* cluster AWG switches, in conjunction with the multichannel tuning (MCT) blocks, can be flexibly interconnected into various optically transparent network topologies, such as a linear topology, a ring topology, a mesh topology, and a partially connected mesh topology, depending on the switch configuration of the OSS. Figure 2.7 illustrates an AWG ring topology where the interconnection network of the AWGs is a unidirectional ring. Each AWG is dedicated one or multiple multichannel tuning (MCT) modules. The primary functionality of the MCT is to facilitate high-speed optical wavelength switching in the network of AWGs. That is, the MCT allows a group of desired optical signals to pass through and then tunes these signals on the appropriate carrying wavelengths so that through the cluster AWG switch, they can either be directly dropped to the destination ToRs within the cluster or be transmitted to the next cluster hop. The multichannel tuning (MCT) architecture is demonstrated in Fig. 2.8. The MCT is composed of a wavelength selective switch (WSS), *K*, tunable wavelength converters (TWCs), and optical multiplexers (MUXs).

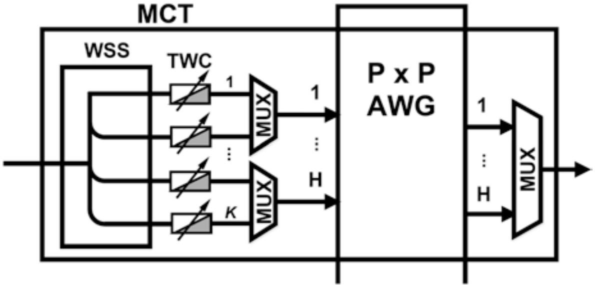


Fig. 2.8 The multichannel tuning (*MCT*) consists of a wavelength selective switch (*WSS*), *K*, tunable wavelength converters (*TWCs*), and optical multiplexers (*MUXs*)

The WSS supports highly flexible wavelength resource allocation, due to its dynamic wavelength route/selection property, which allows each input channel to be individually switched to an arbitrary WSS output port and allows each output signal to include an arbitrary number of wavelengths [57]. In the MCT, the WSS is configured to select the desired wavelengths at desired output ports and then forwards these optical wavelengths to the TWCs for wavelength conversion. The converted optical signals traverse through the core AWG switch and then reach the required AWG ports. The MCT architecture in Fig. 2.8 occupies 1 to *K* AWG ports. Note that there is a tradeoff between the network performance and the number of AWG ports used. The AWG-to-AWG network enables high-speed cluster-to-cluster on-demand connections; thus, the network connectivity is substantially expanded.

One of the most significant aspects of the proposed distributed datacenter network is the design and implementation of the multi-hop routing in managing cluster-to-cluster communications. As illustrated in Fig. 2.7, two distinct fast AWG switches in two separate clusters may be interconnected by one or multiple hops of the ToR switches utilizing bidirectional optical connections through the OSS, so an AWG-to-AWG multi-hop routing path can be established through these intermediate ToR switches. This allows for the exchange of data between two remote clusters that are not directly interconnected. The intermediate ToR switch, referred to as a *cluster head*, is used to aggregate and redirect the received intercluster traffic streams. The main idea behind the multi-hop routing is the deployment of traffic relays, which exploits the routing and data forwarding capabilities of the AWG switches and the ToR switches. That is, a traffic flow containing a sequence of optical packets is propagated along an established path by hopping from one cluster header to another. In doing so, the multi-hop connectivity manages the communications between clusters; thus, it facilitates the realization of network-wide traffic communication and also avoids the constant reconfigurations of the large space switch OSS.

There are two extreme topologies possible with the multi-hop network connection for the proposed datacenter network. In the first network configuration, each cluster contains one cluster head that is directly attached to the logically adjacent cluster. This network design attempts to use only one cluster head per cluster to support intercluster connections. The scaling of the architecture is realized by adding more clusters into the network, without affecting the rest of the network architecture. Nonetheless, the multi-hop routing path is severely unbalanced, as the shortest path includes only two hops, whereas the longest path length grows linearly with the number of clusters in the network. At the other extreme, each cluster has a dedicated cluster head directly connected to each of the other clusters, resulting in balanced multi-hop routing with at most two hops. This network model is highly symmetric and has a low network diameter, but the scalability is limited, as it can accommodate at most $P/2$ clusters, each containing $P/2$ ToR switches, yielding a network totaling at most $P^2/4$ ToR switches. Assuming that the AWG switch has a port count of 64 × 64 ($P = 64$), this topology can interconnect 32 clusters, each of which comprises 32 ToR switches, so a large-scale interconnection network including 1024 ToR switches is then obtained. Although the former network architecture can be scaled out easily, multi-hop routing may include a large number of traffic relays, which leads to performance degradation in terms of loss and communication latency. In contrast, the latter network architecture has limited scalability, but it is symmetric, requires low network diameter, and achieves loss uniformity, due to the fact that there exists a two-hop routing path between any source-destination cluster pair. The comparison indicates that there are obvious trade-offs between the two extreme cases, in terms of scalability, network diameter, loss uniformity, delay, and intra-/intercluster connection capacity. Given the flexibility provided by the physical architecture, arbitrary logical network topologies between these two extremes can be provisioned to balance among these trade-offs to ensure optimal throughput. In practice, the optimized connection structure is selected based on datacenter application requirements over a medium-term time scale.

In the proposed network architecture, the datacenter traffic is classified into two types of communications: intra-cluster communication and intercluster communication. Since the intra-cluster communication is exchanged between two ToRs residing in the same cluster, it can be switched from its source ToR directly to its destination ToR through a high-speed cluster AWG switch which directly interconnects all ToR switches in a cluster. Some heavily communicating ToR switches may be interconnected by the high-speed, high-capacity, low-latency direct fiber connections. A flexible EPS network is also utilized to provide additional intra-cluster connections. Potentially, the intra-cluster connectivity can be supported by three types of connections: fast AWG switches, direct fibers, and EPS modules. Differently, the intercluster connectivity can be provisioned by four types of connections: direct fibers, EPS modules, AWG-to-AWG network, and multi-hop routing. Similarly to the intra-cluster communication, the intercluster communication may be sustained by the direct fibers or the electronic switches, depending on the traffic requirements. In addition to that, an interconnection network of the AWGs is designed to provide high-speed all-optical intercluster connectivity. Furthermore, the multi-hop switching allows allocation of traffic flows on the existing multi-hop paths composed of multiple cluster heads and fast AWG switches. Clearly, the cluster fast AWG switches function as the most essential building blocks in the presented datacenter network and facilitates the routing operations of both intra- and intercluster traffic flows.

To maximize the network performance and resource utilization, an efficient control and scheduling plane is required, which is able to coordinate with the local AWG controllers in managing network resources and performing end-to-end scheduling for flow-level data traffic. Importantly, it needs to compute the most efficient network configuration that guarantees a fully connected graph of the cluster-based network and also fully exploits the capacity of the network connections, in accordance with the predicted traffic requirements. More specifically, the most heavily communicating source/destination ToR switches are provided with direct cross-connections using the fast AWG switches, the direct fibers, and the electronic switches, whereas the remaining communication pairs are interconnected by the AWG-to-AWG connections or the multi-hop paths.

2.4 Conclusions

Optical circuit switching is a promising technology for enabling future expansion of datacenter interconnection networks, in terms of scaling the number of network nodes and scaling link and switch port bandwidths well beyond the levels of today's electronically switched networks. The current research into this area, as reviewed in this chapter, shows a wide range of approaches that generally leverage the potential bandwidths supported by optical circuit switching to off-load aggregated traffic, or longer-lived bulk data flows between servers, from the electronically switched interconnection network. This approach is driving research into faster, submicrosecond optical switching speeds, through integrated photonic optical space switching. As optical switching times reduce, more and more shorter-lived traffic can be switched purely optically, further alleviating traffic overloading and energy consumption and cooling constraints for buffered electronic packet switches. Ultimately, achieving high-port-count nanosecond optical switching fabrics could pave a future path toward elimination of electronic switching altogether above the ToR or server connection level. A further driving factor for high-bandwidth optical switches is to enable future scaling-out of datacenter server numbers. Already, there is a trend in electronically switched datacenter network design toward flattening hierarchical network topologies to two layers, to reduce the requirement for very large core switches and to reduce routing and network configuration and management complexity. As datacenters scale, the aggregation center-stage switching will inevitably come under bandwidth requirement pressures, particularly to maintain high bisection bandwidths needed in cloud and big data applications. With transparency to format and bit-rate, replacing these core switches with fast optically switched alternatives would offer significant future-proofing advantages. Another increasing trend is the requirement for flexible provisioning of inter-datacenter network capacity. As datacenter requirements grow beyond the capacity of a single installation site, seamless distribution of datacenters across multiple sites is becoming more and more desirable. As the required interconnecting metro area and core networks have already been deployed based on optical circuit switching and WDM technologies, a more seamless integration of intra- and inter-datacenter networks is possible if intra-datacenter networks were to be migrated to the same fundamental optically switched technologies. Indeed optical switch vendors are already adapting and targeting these technologies toward datacenter interconnection, which would seem to align well with datacenter operators' future needs.

Acknowledgment The datacenter network project described is a work in progress. The authors would like to acknowledge the support of the Science Foundation of Ireland (SFI) under the IPIC Programme (12/RC/2276), the CONNECT Programme (13/RC/2077), and the US-Ireland R&D Partnership Programme (SFI project code 15/US-C2C/I3132) and the NSF grants #EEC-0812072 and CNS1650669 and DOE grant #DE-SC0015867.

References

1. *Cisco*, *Cisco Global Cloud Index: Forecast and Methodology, 2014–2019*, White paper, October 2015
2. Dell'Oro Group, *Dell'Oro Group: Server Report*. White paper, March 2016
3. S. Kandula et al., Nature of datacenter traffic: measurements and analysis. in *Proceedings of ACM SIGCOMM conference on Internet Measurement*, November 2009
4. T. Benson, A. Akella, D.A. Maltz, Network traffic characteristics of data centers in the wild. in *Proceedings of ACM SIGCOMM conference on Internet measurement*, November 2010
5. P. Bodík, I. Menache, M. Chowdhury, Surviving failures in bandwidth-constrained datacenters. in *Proceedings of ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, August 2012
6. F.P. Tso, D. Pezaros, Improving data center network utilization using near-optimal traffic engineering. IEEE Trans. Parallel Distrib. Syst. **24**(6), 1139–1148 (June 2013)
7. A. Roy, H. Zengy, J. Baggay, G. Porter, A.C. Snoeren, Inside the social network's (Datacenter) network. in *Proceedings of ACM SIGCOMM Conference on Special Interest Group on Data Communication*, August 2015, pp. 123–137
8. M. Al-Fares, A. Loukissas, A. Vahdat, A scalable, commodity data center network architecture. in *Proceedings of ACM SIGCOMM Conference on Data Communication*, August 2008
9. C. Clos, A study of non-blocking switching networks. Bell Syst. Tech. J. **32**(2), 406–424 (1953)
10. A. Singh et al., Jupiter rising: a decade of clos topologies and centralized control in Google's datacenter network. in *Proceedings of ACM SIGCOMM*, August 2015
11. L.A. Barroso, J. Dean, U. Holzle, Web search for a planet: The Google cluster architecture. IEEE Micro **23**(2), 22–28 (2003)
12. N. Farrington, A. Andreyev, "Facebook's data center network architecture. in *Proceedings of IEEE Optical Interconnects Conference (OI)*, 2013
13. A. Andreyev, *Introducing data center fabric, the next-generation Facebook data center network*. White paper, November 2014
14. M. Alizadeh, T. Edsall, *On the data path performance of Leaf-Spine datacenter fabrics*. High-Performance Interconnects (HOTI), August 2013, pp. 71–74
15. C. Guo et al., DCell: a scalable and fault-tolerant network structure for data centers. in *Proceedings of ACM SIGCOMM Conference on Data Communication*, August 2008
16. C. Guo et al., BCube: a high performance, server-centric network architecture for modular data centers. in *Proceedings of ACM SIGCOMM Conference on Data Communication*, August 2009

17. D. Li et al., FiConn: using backup port for server interconnection in data centers. in *Proceedings of IEEE International Conference on Computer Communications*, April 2009, pp. 2276–2285
18. C. Kachris, I. Tomkos, A survey on optical interconnects for data centers. IEEE Commun. Surv. Tut. **14**(4) (2012)
19. D. Nikolova et al., Scaling silicon photonic switch fabrics for data center interconnection networks. Opt. Exp. **23**(2), 1159–1175 (Jan. 2015)
20. CALIENT, *S320 Photonic Switching*. White paper, March 2013
21. ARISTA, *7280R Series Data Center Switch Router*. White paper, July 2016
22. Polatis, *Series 7000 – 384x384 port Software-Defined Optical Circuit Switch*. White paper, March 2016
23. M. Chen, H. Jin, Y. Wen, V.C.M. Leung, Enabling technologies for future data center networking: a primer. IEEE Netw. **27**(4), 8–15 (Aug. 2013)
24. K.J. Barker et al., On the feasibility of optical circuit switching for high performance computing systems. in *Proceedings of ACM/IEEE SC 2005 Conference*, November 2005, pp. 1–22
25. T. Joon Seok et al., Large-scale broadband digital silicon photonic switches with vertical adiabatic couplers. Optica **3**(1) (Jan. 2016)
26. D.K. Hunter et al., WASPNET: a wavelength switched packet network. Commun. Magaz. **37**(3), 120–129 (1999)
27. S.D. Lucente, N. Calabretta, J.A.C. Resing, H.J.S. Dorren, Scaling low-latency optical packet switches to a thousand Ports. J. Opt. Commun. Netw. **4**(9), 17–28 (Sep. 2012)
28. N. Calabretta et al., On the performance of a large-scale optical packet switch under realistic data center traffic. J. Opt. Commun. Netw. **5**(6), 565–573 (June 2013)
29. R.A. Spanke, Architectures for guided-wave optical space switching systems. IEEE Commun. Mag. **25**(5) (1987)
30. V.E. Benes, *Mathematical Theory of Connecting Networks and Telephone Traffic* (Academic Press, New York, 1965)
31. J. Kim, W.J. Dally, D. Abts, Flattened butterfly: a cost-efficient topology for high-radix networks. Int. Symp. Comput. Archit. (ISCA) (2007)
32. G. Wang et al., c-Through: part-time optics in data centers. in *Proceedings of ACM SIGCOMM*, October 2010
33. N. Farrington et al., Helios: a hybrid electrical/optical switch architecture for modular data centers. in *Proceedings of ACM SIGCOMM*, October 2010
34. G. Porter et al., Integrating microsecond circuit switching into the data center. in *Proceedings of ACM SIGCOMM*, 2013, pp. 1–12
35. N. Farrington et al., Multiport microsecond optical circuit switch for data center networking. IEEE Photon. Technol. Lett. **25**(16) (2013)
36. H. Liu et al., Circuit switching under the radar with REACToR. in *Proceedings of USENIX Conference on Networked Systems Design and Implementation*, April 2014
37. D. Lugones, K. Katrinis, M. Collier, A reconfigurable optical/electrical interconnect architecture for large-scale clusters and datacenters. in *Proceedings of ACM International Conference on Computing Frontiers*, May 2012, pp. 13–22
38. D. Lugones et al., A reconfigurable, regular-topology cluster/datacenter network using commodity optical switches. Futur. Gener. Comput. Syst. **30**, 78–89 (2014)
39. Z. Zhu, S. Zhong, Scalable and topology adaptive intra-data center networking enabled by wavelength selective switching. in *Proceedings of OFC*, 2014
40. Z. Zhu, S. Zhong, L. Chen, K. Chen, Fully programmable and scalable optical switching fabric for petabyte data center. Opt. Exp. **23**(3), 3563–3580 (2015)
41. S. Yan et al., Archon: A function programmable optical interconnect architecture for transparent intra and inter data center SDM/TDM/WDM networking. J. Lightw. Technol. **33**(8), 1586–1595 (2015)
42. B.A.A. Nunes et al., A survey of software-defined networking: past, present, and future of programmable networks. IEEE Commun. Surv. Tut. **16**(3), 1617–1634 (2014)

43. K. Katrinis, G. Wang, L. Schares, SDN control for hybrid OCS/electrical datacenter networks: an enabler or just a convenience? IEEE Photon. Soc. Summer Top. Meet. Ser., 242–243 (2013)
44. Y. Ben-Itzhak, C. Caba, L. Schour, S. Vargaftik, *C-Share: Optical Circuits Sharing for Software-Defined Data-Centers*. eprint arXiv, September 2016
45. A. Singla et al., Proteus: a topology malleable data center network. in *Proceedings of ACM SIGCOMM Workshop on Hot Topics in Networks*, October 2010
46. K. Chen et al., OSA: an optical switching architecture for data center networks with unprecedented flexibility. Northwestern University, Technical Report, 2012
47. S.L. Woodward, M.D. Feuer, Benefits and requirements of flexible-grid ROADMs and networks. J. Opt. Commun. Netw. **5**(10), A19–A27 (2013)
48. M. Schlosser et al., Requirements analysis of technology enablers for the flexi-grid optical path-packet infrastructure for Ethernet transport. STRAUSS Project, May 2014
49. Z. Zhang et al., Elastic optical ring with flexible spectrum ROADMs: An optical switching architecture for future data center networks. Opt. Switch. Netw. **19**(part 1), 1–9 (2016)
50. J. Zhang et al., Experimental demonstration of OpenFlow-based control plane for elastic lightpath provisioning in flexi-grid optical networks. Opt. Exp. **21**(2), 1364–1373 (Jan. 2013)
51. K.-I. Kitayama et al., Torus-topology data center network based on optical packet/agile circuit switching with intelligent flow management. J. Lightw. Technol. **33**(5) (Mar. 2015)
52. N. Hamedazimi, et al., "Firefly: a reconfigurable wireless data center fabric using free-space optics," SIGCOMM Comput. Commun. Rev., vol. 44, no. 4, pp. 319–330, Aug. 2014.
53. X. Zhou et al., Mirror mirror on the ceiling: Flexible wireless links for data centers. in *Proceedings of ACM SIGCOMM*, 2012
54. H. Vardhan et al., 60 GHz wireless links in data center networks. Comput. Netw. **58**, 192–205 (2014)
55. M. Ghobadi, R. Mahajan, A. Phanishayee, N. Devanur, J. Kukarni, G. Ranade, P.-A. Blanche, H. Rastegarfar, M. Glick, D. Kilper, ProjecToR: Agile reconfigurable data center interconnect. in *Proceedings of ACM SIGCOMM*, 2016, pp. 216–229
56. S. Cheung, T. Su, K. Okamoto, S.J.B. Yoo, Ultra-compact silicon photonic 512 x 512 25-GHz arrayed waveguide grating router. IEEE J. Select. Top. Quant. Electron. **20**(4), 310–316 (2014)
57. R. Ryf et al., Wavelength-selective switch for few-mode fiber transmission. in *Proceedings of 39th European Conference on Optical Communications (ECOC)*, September 2013

Chapter 3
Optical Switching in Data Centers: Architectures Based on Optical Packet/Burst Switching

Nicola Calabretta and Wang Miao

3.1 Introduction

Data centers (DC) are phenomenally growing in size and complexity, to satisfy the demands for more powerful computational performance driven by the data-intensive applications as well as high-density virtualizations [1]. High-performance and energy-efficient multi-core processors are developed aggressively to provide higher processing capability [2]. Foreseeing the preservation of Moore's law through chip-level parallelism, the multi-core product is expected to scale unabated in computing systems [3], which exposes more pressure on the interconnects and switching elements of the intra-DC network to guarantee a balanced I/O bandwidth performance [4].

Current data center networks (DCNs) are organized in a hierarchical tree-like topology based on the bandwidth-limited electronic switches. A certain degree of oversubscription is commonly enforced [5], resulting in the bandwidth bottleneck and large latency especially for inter-rack/cluster communications. High power consumption related to the O/E/O conversion and format-dependent front end is another issue, limiting the power-efficient and cost-efficient scaling to higher capacity. Therefore, optical switching technology has been considered as a promising candidate for intra-DC networking solutions.

Compared with the optical circuit switching (OCS), the OPS and OBS based on fast optical switches could provide on-demand resource utilization, highly flexible connectivity to effectively cope with bursty traffic features, as well as high fan-in/fan-out hotspot patterns in DCNs. Many techniques have been actively developed, each exhibiting advantages and disadvantages when considered for DCN scenarios. In this chapter, various classes of optical switching technologies in implementing

N. Calabretta (✉) • W. Miao
Electro-Optical Communication Group, Institute for Photonic Integration (IPI), Department of Electrical Engineering, Technical University of Eindhoven, Groene Loper 5, 5612 AE, Flux 9.087, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: n.calabretta@tue.nl

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
DOI 10.1007/978-3-319-61052-8_3

OPS and OBS nodes have been briefly introduced. Then several typical optical DCN architectures based on OPS and OBS are presented, followed by the discussion on the performance focusing on different attributes. In the last section, a novel DCN architecture employing fast optical switches is reported which shows potential settlement to the scalability challenges faced by traditional solutions.

3.2 Data Center Networks: Requirements and Challenges

Data centers consist of a multitude of servers as computing nodes and storage subsystems interconnected with the appropriate networking hardware and accompanied by highly engineered power and cooling subsystems [6]. The DCN is responsible to support the large amounts of workload exchanged among the parallel server machines. The traditional DCN uses a multitier architecture, with tens of servers housed in individual racks, and racks are grouped into clusters. The top-of-the-rack (ToR) switches interconnect the servers via copper or optical links, and the inter-rack communication is handled by layers of electronic switches. Ideally, the DCN should provide a full bisection bandwidth, and thus the oversubscription ratio is 1:1 indicating high server utilization and computation efficiency. However, due to the super linear costs associated with scaling the bandwidth and port density of electronic switches, such a design would be prohibitively expensive for a large-scale DCN. In practice, DCs tend to enforce an oversubscription 1:4 to 1:10 [7]. There is more bandwidth available for intra-rack communication than inter-rack communication, and similar trend would be found at higher switching layers.

A set of stringent requirements are imposed on the DCNs, a few key points of which are listed as follows.

- *Capacity*: An increasing fraction of data centers is migrating to warehouse scales. Although substantial traffic will continue to cross between users and data centers, the vast majority of the data communication is taking place within the data center [8]. Recent studies have shown the continuous increase of the inter-rack traffic with a clear majority of traffic being intra-cluster (> 50%) [9]. Higher bandwidth interconnects in combination with high-capacity switching elements are required especially for inter-rack and inter-cluster communications, to avoid the congestion drops caused by the inherent burstiness of flows and intentionally oversubscribed network [10].
- *Latency*: Packet latency is defined as the time it takes for a packet to traverse the network from the sender to the receiver node (end-to-end latency) which includes both the propagation and switch latency. When it comes to the closed environment like DCs, the latency is dominated by the switch latency mainly contributed by the buffering, routing algorithm, and arbitration. Low latency is a crucial performance requirement especially for mission-critical and latency-sensitive applications where microseconds matter (e.g., financial networking).

- *Interconnectivity*: The servers in data centers have 10s–100s concurrent flows on average [9]. Considering the small fraction of intra-rack traffic, almost all flows will traverse an uplink at the ToR switch as inter-rack communication. Therefore, the degree of the interconnectivity supported by the switching network should be large enough to accommodate the number of concurrent flows. Moreover, considering most flow is short and tends to be internally bursty, fast and dynamic reconfiguration of such interconnectivity (e.g., statistical multiplexing) is also needed to guarantee the efficient bandwidth utilization and timely service delivery.
- *Scalability*: The network architecture should enable scaling to large number of nodes to address future capacity needs in a cost-efficient manner. Extension of an existing network in terms of both node count and bandwidth in an incremental fashion is preferable, i.e., without having to replace a disproportionate amount of the installed hardware.
- *Flexibility*: Data centers are expected to adopt technologies that allow them to flexibly manage the service delivery and adapt to changing needs. To this end, the resources (such as computing, storage and network) are pooled and dynamically optimized by the control plane through software configuration. In addition, open standards, open protocols, and open-source development are more and more involved to facilitate and speed up the deployment operation as well as management in the application- and service-based environment.
- *Power/cost efficiency*: A data center represents a significant investment in which the DCN occupies a significant portion [11]. Besides the costs for hardware and software installation, running a large-scale data center is mainly a power consumption matter. Power efficiency is a key target for reducing the energy-related costs and scaling the bandwidth by improving the power density performance. In this sense, significant efforts have been made toward employment of optical technology and virtualization leading to enhancements in power and cost efficiency [12].

As can be seen from these requirements, high-capacity switching networks with low switching latency and fine switching granularity (e.g., deploying statistical multiplexing) are necessary to effectively improve the bandwidth efficiency and handle the burstiness of the DC traffic flows. The large number of concurrent flows makes large interconnectivity as well as fast reconfiguration a necessity for the switches, in which case the circuit-based approaches may be challenging to employ. The pairwise interconnection and tens of milliseconds reconfiguration time have strictly confined the applications to well-scheduled and long-lived tasks.

With the increasing number of server nodes and rapid upgrade in I/O bandwidth, the abovementioned requirements would be quite challenging for current DCN, in terms of both switching node and network architecture.

First, it is difficult for the electronic switch to satisfy the future bandwidth need. The increasing parallelism in microprocessors has enabled continued advancements in computational density. Despite the continuous efforts from merchant silicon providers toward the development of application-specific integrated circuits (ASICs), the implementation of high-bandwidth electronic switch node is limited by the switch ASIC I/O bandwidth (to multi-Tb/s) due to the scaling issues of the ball grid array (BGA) package [13]. Higher bandwidth is achievable by stacking several ASICs in a multitier structure but at the expense of larger latency and higher cost. Another limiting factor is the power consumption. As electronic switch has to store and transmit each bit of information, it dissipates energy with each bit transition, resulting in power consumption at least proportional to the bit-rate of the information it carries. In addition, the O/E/O conversions and format-dependent interfaces need to be further included as front end, greatly deteriorating the power-efficiency and cost-efficiency performance.

Interconnecting thousands of ToRs, each with large amount of aggregated traffic, would put an enormous pressure on the multitier treelike topology employed by the current DCNs. Due to the limiting performance in terms of bandwidth and port density of conventional electronic switches, network is commonly arranged with oversubscription. Consequently, data-intensive computations become bottlenecked especially for the communication between servers residing in different racks/clusters. The multiple layers of switches also bring large latency when a packet traverses the all DCN to reach its destination, mainly caused by the queueing delay of buffer-related processing. Therefore, to effectively address the bandwidth, latency, scalability, and power requirements imposed by the next-generation DCNs, innovations in the switching technology and network architecture are of paramount significance.

3.3 Optical Data Center Networks

With the prevalence of the high-capacity optical interconnects, optically switched DCNs have been proposed as a solution to overcome the potential scaling issues of the electronic switch and traditional tree-like topology [14, 15]. The switching network handles the data traffic in optical domain thus avoiding the power-consuming O/E/O translations. It also eliminates the dedicated interface for modulation-dependent process, achieving better efficiency and less complexity. Moreover, benefiting from the optical transparency, the switching operation (including the power consumption) is independent of the bit-rate of the information. Scalability to higher bandwidth and employment of WDM technology can be seamlessly supported, enabling superior power-per-unit bandwidth performance.

Various optical switching techniques have been investigated for DC applications, among which the OPS, OBS, and OCS are the most prominent ones. With respect to the requirements for the DCNs, optical switching technologies and the potential of photonic integration can support high-capacity and power-/cost-efficient scaling. Software-defined networking (SDN) is also seeing penetration into the newly proposed optical DCNs to facilitate the flexible provisioning and performance enhancement. However, concerns regarding the limited interconnectivity and handling of applications with fast-changing traffic demands still exist. OCS networks employing slow switches (tens of milliseconds reconfiguration time) have strictly confined the applications to well-scheduled and long-lived tasks. The static-like and pairwise interconnections would only be beneficial as supplementary elements. The OPS and OBS with fast optical switches allowing for on-demand resource utilization and highly flexible connectivity enabled by the statistical multiplexing are becoming the appealing switching schemes for DCNs.

3.4 Optical Packet and Burst Switching Technologies

OPS and OBS technology makes it possible to achieve sub-wavelength bandwidth granularity exploiting statistical multiplexing of bursty flows. The OPS/OBS network consists of a set of electronic edge nodes interconnected by optical switches. At the edge nodes, electrical data packets from the client network with similar attributes are aggregated in an optical packet/burst. It goes through the optical switches transparently without O/E/O conversion. After arriving at the destined edge node, it is disassembled and forwarded to the client network. The switching operation of the packet/burst (usually referred as payload) is determined by a packet header/burst control header (BCH), which is optically encoded but undergoes O/E conversion and electronic processing at the optical switch node. The main differences between OPS and OBS are:

- In OPS networks, the packet durations are in the hundreds of nanoseconds to microseconds range. The packet header is transmitted in the same channel with respect to the payload and either overlaps the payload in time or sits ahead of it. Advance reservation for the connection is not needed, and the bandwidth can be utilized in the most flexible way. These features make OPS a suitable candidate for data center applications which requires transmission of small data sets in an on-demand manner.
- OBS uses more extensive burst aggregation in the order of tens to thousands of microseconds. The BCH is created and sent toward the destination in a separate channel prior to payload transmission. The BCH informs each node of the arrival of the data burst and drives the allocation of an optical end-to-end connection path. OBS enables sub-wavelength granularity by reserving the bandwidth only for the duration of the actual data transfer.

Note that reconfiguration time of the optical switch including the control operation should be much smaller than the duration of the packet/burst, to ensure a low-latency delivery at a fast arrival rate as well as optimized bandwidth utilization. Practical realization of OPS/OBS relies heavily on the implementation of controlling technique and scheme adopted for contention resolution [16]. Table 3.1 summarizes some examples of different classes of optical switching technologies for OPS/OBS node where comparisons of different attributes in terms of switching performance are also reported. A broad range of technologies has been developed for OPS and OBS systems. The space optical switches based on piezoelectric beam

Table 3.1 Optical switching technologies for implementing OPS and OBS node

	Switching time	Transparency	Scale	Loss	Application	Ref.
Piezoelectric beam steering	~10 ms	Good	384 × 384	Low	OBS	[17]
3D optical MEMS	~10 ms	Good	320 × 320	Low	OBS	[18]
2D optical MEMS	~50 µs	Good	50 × 50	Fair	OBS	[19]
LCoS WSS	~10’s ms	Good	1 × 40	Fair	OBS	[21]
Mach-Zehnder (thermo-optic)	~10′ µs	Good	32 × 32	High	OBS	[20]
Micro-ring resonator (thermo-optic)	~10′ µs	Fair	8 × 8	Fair	OBS	[22]
PLZT MZI	~10 ns	Good	8 × 8	High	OPS, OBS	[23]
InP MZI	2.5 ns	Good	8 × 8	High	OPS, OBS	[24]
$LiNbO_3$ MZI	~1 ns	Fair	32 × 32	High	OPS, OBS	[25]
MZI + EAM	< 10 ns	Good	8 × 8	Fair	OPS, OBS	[26]
TWC + AWGR	~10’s ns	Poor	10’s × 10’s	Fair	OPS, OBS	[27]
TL + AWGR	~10’s ns	Good	100′ × 100’s	Low	OPS, OBS	[28]
EAM B&S	~1 ns	Good	8 × 8	High	OPS, OBS	[29]
SOA B&S	~1 ns	Good	64 × 64	Fair	OPS, OBS	[30]
SOA multistage	< 10 ns	Good	16 × 16	Fair	OPS, OBS	[31]
Semiconductor optical phase array	~20 ns	Good	64 × 64	Fair	OPS, OBS	[32]

steering and 3D MEMS as well as wavelength selective switch (WSS) based on Liquid Crystal on Silicon (LCoS) have tens of milliseconds of switching time, which are more suitable for long burst operations. The rest of the techniques are mainly based on interferometric and gating switch elements, holding the potential of photonic integration to further scale the capacity. Large interconnectivity can be enabled by cascading 1 × 2 or 2 × 2 switching elements such as 2 × 2 Mach-Zehnder interferometer (MZI) and micro-ring resonator (MRR). Mach-Zehnder switches with electrooptic switching offer faster reconfiguration than the thermo-optic tuning, and extra optical amplifier is normally needed due to the relatively high insertion loss, and therefore scalability can be compromised by the OSNR degradation. Another category of the fast (nanoseconds) optical switches is implemented by arrayed waveguide grating router (AWGR) along with tunable lasers (TLs) or tunable wavelength converters (TWCs). The interconnection scale and performance is largely dependent on the capability of the TL and TWC. Note that WSS, MRR, and AWGR are all wavelength dependent. For the broadcast-and-select (B&S) architecture, the semiconductor optical amplifier (SOA) and electro-absorption modulator (EAM) are commonly used as the gating elements. The broadcast stage introduces high splitting loss, in which case the SOA can provide loss compensation which is essential in realizing a large connectivity. In the practical implementation of the OPS/OBS network, the techniques (or in combination) listed here can be further included in a network as a basic switching unit [33, 34].

3.4.1 Technical Challenges in OPS/OBS Data Centers

Despite the advantages of increased capacity and power efficiency brought by the optical transparency, employing OPS/OBS is actually facing several challenges which need to be carefully considered for DC networking applications.

- *Lack of optical memory*: As no effective optical memories exist, contention resolution is one of the most critical functionalities that need to be addressed for OPS/OBS node. Several approaches have been proposed to resolve the contention in one or several of the following domains:
  - Time domain: the contending packet/burst can be stored in fixed fiber delay line (FDL) or electronic buffer.
  - Wavelength domain: by means of wavelength conversion, the packet/burst can be forwarded in alternative wavelength channels.
  - Space domain: contending burst is forwarded to another output port (deflection routing).

The techniques based on FDL, wavelength conversion, and deflection routing increase significantly the system complexity in terms of routing control and packet synchronization. Moreover, the power and quality of the signal are degraded which results in limited and fixed buffering time. A promising solution is to exploit the electronic buffer at the edge nodes [35]. To minimize the latency, the optical switch should be as close as possible to the edge nodes and fast decision-making is required. This is feasible in a DC environment with interconnects ranging from few to hundreds of meters.

- *Fast reconfiguration and control mechanism*: To fully benefit from the flexibility enabled by the statistical multiplexing, fast reconfiguration of the optical switch is a key feature. Although OBS is less time demanding, slower OBS can cause inefficiency and unpredictability especially under high network load. Therefore, the optical fabrics with fast-switching time together with fast-controlling mechanism are desired. Regarding the DCN applications, the implementation of the controlling technique should follow the increase of the network scale and optical switch port count and, more importantly, occupy as least resources as possible.
- *Scalability*: Depending on the design and technology employed in optical switches, signal impairment and distortion are observed due to noise and optical nonlinearities. Consequently, the optical switches are realized with limited port count. Scaling the network interconnectivity and maintaining the performance would require the switches to have port count as large as possible and to be intelligently connected to avoid the hierarchical structure. The flat topology also brings the benefits of simplified controlling and buffering which may be problematic for fast optical switches. On the other hand, optical transparency and WDM technology would benefit the DCN in the context of scaling up the bandwidth density. Further improvements could be made by means of photonic integration, which greatly reduces the optical switch footprint and power consumption.

3.5 Optical DCN Architecture Based on OPS/OBS

OPS and OBS technologies providing high bandwidth and power efficiency have been adopted in the recent researches for optical DCNs [36, 37]. This section gives an overview and a general insight on the optical DCN architectures based on OPS/OBS that have been recently proposed, which can be classified into different categories according to the switching technologies used.

3.5.1 Based on OBS

3.5.1.1 OBS with Fast Optical Switches

In [44], a DCN consisting of ToR switches at the edge and an array of fast optical switches at the core to perform optical burst forwarding on the pre-configured light paths has been proposed. It has separate data and control planes as shown in Fig. 3.1. Two-way reservation OBS is implemented, facilitated by the single-hop topology with configuration of only one switch per request. It achieves zero burst loss with slight degradation of the latency owning to the limited round-trip time in DC environment. The centralized control plane is responsible for the routing, scheduling, and switch configuration. It processes the control packets from the ToRs sent through a dedicated optical transceiver, finds appropriate path to the destination ToR, and configures optical switches as allocated in the control packets. Since the fast optical switch connects to every ToRs, scalability is challenging in terms of achievable port count for large number of ToRs. The resulted complexity in the control plane may be another bottleneck in scaling up the network.

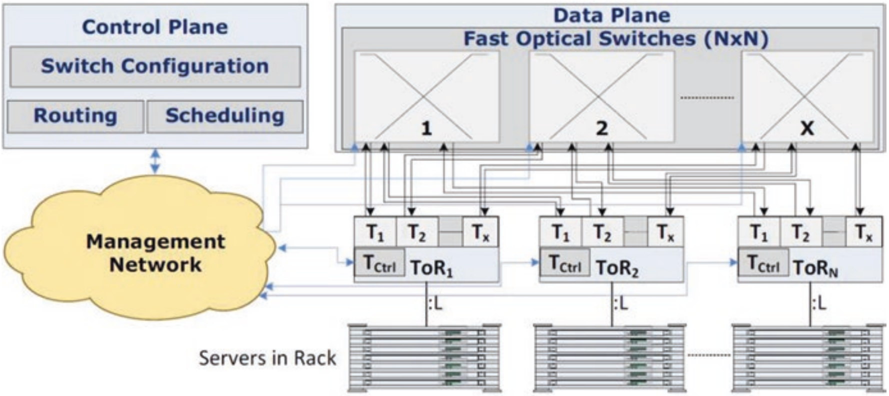


Fig. 3.1 DCN based on OBS with fast optical switches

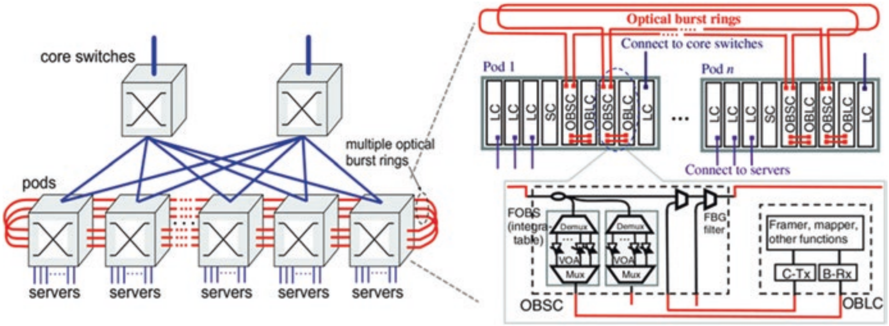


Fig. 3.2 Multiple optical burst rings and internal architecture of the pod

3.5.1.2 Optical Burst Rings

The OBS is utilized in [45] to improve the inter-pod communications in DCNs. The network architecture and the pod are depicted in Fig. 3.2. The pods are connected through the multiple optical burst rings. Bursty and fast-changing inter-pod traffic is handled by the core switches, while the relatively stationary traffic is handled via the optical burst rings. Some line cards (LCs) are configured for connecting the servers, and others are used to access the core switches. The switch cards (SCs) aggregate the traffic and together with the control unit make decision to forward the traffic to the LCs connecting to the core switches or to an optical burst line card (OBLC) which sends the traffic in form of burst to the optical rings. The optical burst switch cards (OBSCs) perform optical burst add/drop to/from the optical burst rings, as shown in Fig. 3.2. The advantages of this architecture are the high inter-pod transmission bandwidth and large number of interconnectivity (>1000 pods). Much shorter connection reconfiguration time is offered compared with OCS-based solutions, achieving better bandwidth utilization.

3.5.1.3 HOS Architecture

An optical switched interconnect based on hybrid optical switching (HOS) has been proposed and investigated in [46]. HOS integrates optical circuit, burst, and packet switching within the same network, so that different DC applications are mapped to the most suitable optical switching mechanisms. As shown in Fig. 3.3, the HOS is arranged in a traditional fat-tree three-tier topology, where the aggregation switches and the core switches are replaced by the HOS edge and core node, respectively. The HOS edge nodes are electronic switches which perform the traffic classification and aggregation. The core node has parallel optical switches composed of switching elements (SEs). A slow optical switch based on 3D MEMS handles circuits and long bursts, and a fast SOA-based optical switch with a three-stage Clos network deals with packets and short bursts. The HOS control plane manages the scheduling

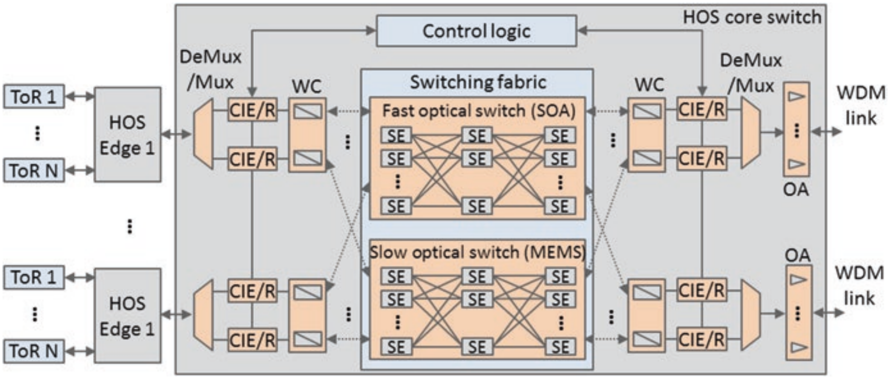


Fig. 3.3 HOS interconnection network

and transmission of the optical circuits, bursts, and packets. Wavelength converters (WCs) are used to solve the possible contentions. Numerical studies show low loss rates and low delays although the practical implementation of a large-scale network remains challenging.

3.5.1.4 HOSA Architecture

HOSA, shown in Fig. 3.4, is another DCN architecture that employs both fast and slow optical switches [47]. Different with the previous work that uses only fast optical switches [44], slow MEMS optical switches are added to exploit the benefits of both types of fabrics. The traffic assembling/disassembling and classification is implemented at the newly designed ToR switch. The control plane still uses a centralized controller which receives connection requests and configures the data plane through a management network. The array of fast optical switches operates in an OBS manner, forwarding the data burst on the predefined connection path. The evaluation results show low-latency and high-throughput performance with low power consumption, assuming large port counts of slow/fast optical switches are deployed in the single-stage network.

3.5.1.5 Torus-Topology DCN

Figure 3.5 shows the Torus DCN [48] based on co-deployment of OPS and OCS. The architecture features with a flat topology where each hybrid optoelectronic router (HOPR), that interconnects a group of ToR switches, is connected to the neighboring HOPRs. The traffic from server is converted into the optical packet and fed into the corresponding HOPR attached with a fixed-length optical label. HOPR uses fast

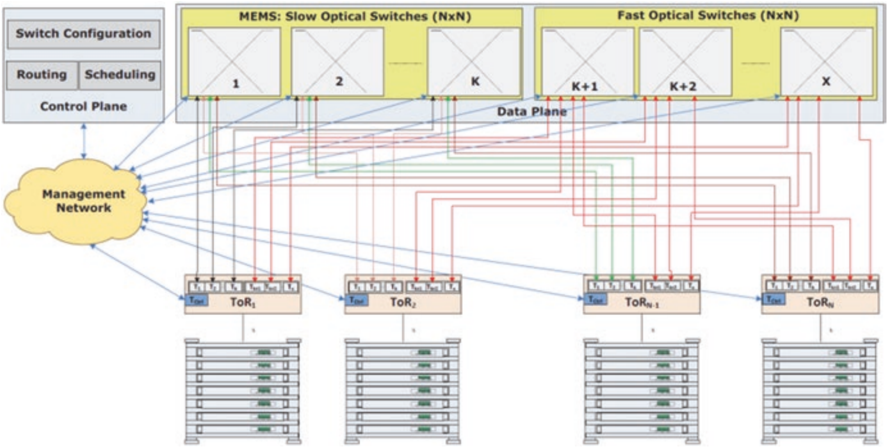


Fig. 3.4 HOSA DCN architecture

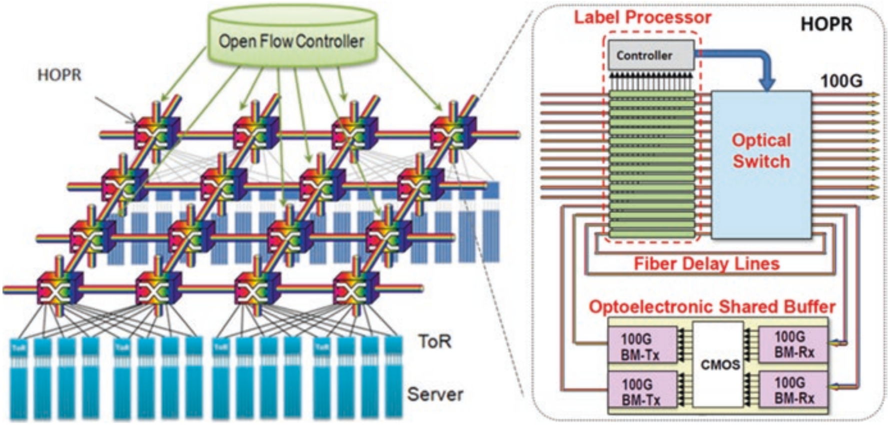


Fig. 3.5 Torus DCN employing hybrid optoelectronic routers (HOPRs)

optical fabric (EAM-based broadcast-and-select structure) which supports both packet operation and circuit operation (express path). The packet contention, which happens when a link is desired by more than one packets or it is reserved by an express path, is solved by different schemes (i.e., deflection routing, FDLs, and optoelectronic shared buffer). The enabling technologies for implementing an HOPR have been detailed, aiming at a high energy efficiency and low latency of 100 ns regime. For an efficient transfer of high-volume traffic, a flow management has been implemented with OpenFlow-controlled express path. Despite the multi-hop transmission may be needed for interconnecting the ToRs, Torus provides the advantages of superior scalability and robust connectivity.

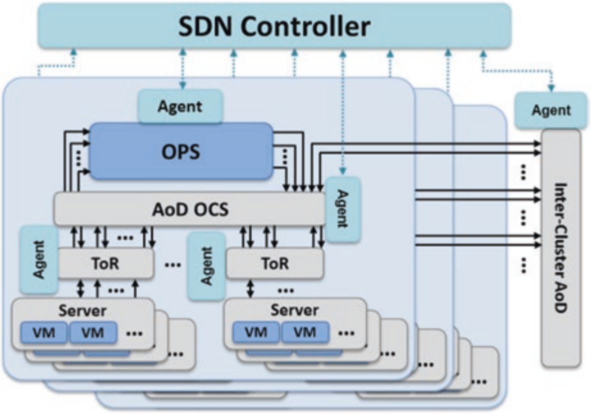


Fig. 3.6 The LIGHTNESS DCN architecture

3.5.1.6 LIGHTNESS DCN Architecture

A flat DCN architecture integrating both OPS and OCS switching technologies to deal with the inconsistent application requirements has been investigated in LIGHTNESS project [49]. The hybrid network interface card (NIC) located in each server supports the switching of the traffic to either OPS or OCS resulting in an efficient utilization of the network bandwidth. As illustrated in Fig. 3.6, the SOA-based OPS which employs broadcast-and-select architecture is plugged into the Architecture on Demand (AoD) backplane as a switching module to handle short-lived data packets. The AoD itself is a large port count fiber switch which can be configured to support OCS function for long-lived data flows. The network can be scaled by interconnecting multiple intra-cluster AoDs with an inter-cluster AoD. Another innovation made by LIGHTNESS is the fully programmable data plane enabled by the unified SDN control plane. It is worth noting that the switching operation of the OPS is controlled by the local switch controller based on the in-band optical labels, which is decoupled from the SDN-based control (e.g., look-up table update and statistic monitoring). Similar schemes have been found in Archon [50] and burst-over-circuit architecture [51] where the OPS is replaced by PLZT-based optical switch and AWGR incorporating with TWC, respectively.

3.5.2 Based on OPS

3.5.2.1 IRIS Project: Photonic Terabit Routers

The IRIS project has developed a photonic packet router that scales to hundreds of terabit/s capacity [38]. As shown in Fig. 3.7, the router employs a load-balanced multistage architecture. Each node (e.g., ToR switch) is connected to an input port

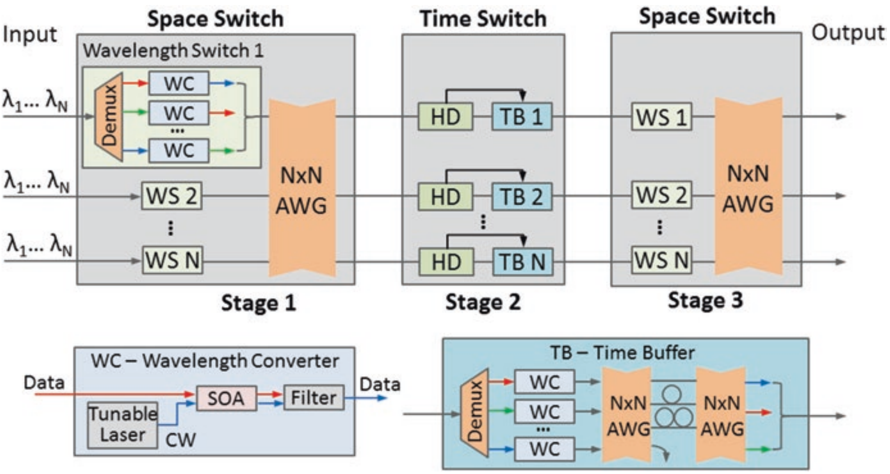


Fig. 3.7 IRIS photonic terabit router

of the first stage using N WDM wavelength channels each carrying synchronous fixed-length data packets. The wavelength switch is based on an array of all-optical SOA-based wavelength converters to set the wavelength routing. The second stage is time switch which contains N time buffers consisting of shared ODLs.

The wavelength is configured in the way that the packet entering on port of the time buffer always exits on the corresponding output port. The third stage then forwards the packet to the desired destination. Due to the periodic operation of the third space switch, the scheduling is local and deterministic to each time buffer which greatly reduces control-plane complexity. The IRIS project has demonstrated the operation of a partially populated router with integrated photonic circuits and developed interoperability card that can connect electronic routers with 10 Gb Ethernet interfaces to the IRIS router. Using 40 Gb/s data packets and 80 × 80 AWGs allows this architecture to scale to 256 Tb/s capacity.

3.5.2.2 Petabit Optical Switch

The petabit optical switch is based on tunable lasers (TLs), tunable wavelength converters (TWCs), and AWGRs, as schematically shown in Fig. 3.8 [39]. The ToR switches are connected to the optical switch, which is three-stage Clos network including input modules (IMs), central modules (CMs), and output modules (OMs). Each module uses an AWGR as the core. The SOA-based TWCs as well as the TLs in the line cards are controlled by the scheduler. A prominent feature of the switch is that packets are buffered only at the line cards, while the IMs, CMs, and OMs do not require buffers. This helps to reduce implementation complexity and to achieve low latency. Performance of the petabit optical switch is evaluated with simulation which shows high throughput benefited from the efficient scheduling algorithm.

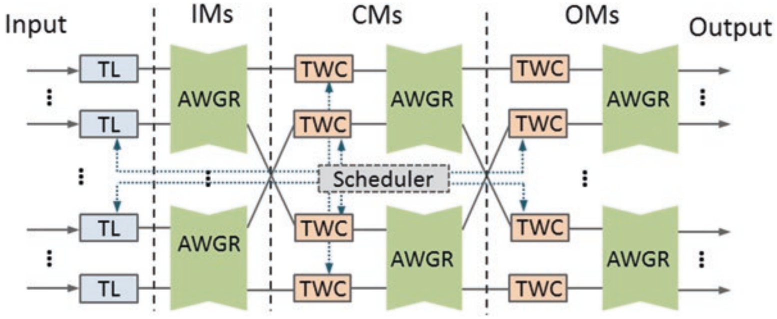


Fig. 3.8 The petabit optical switch architecture

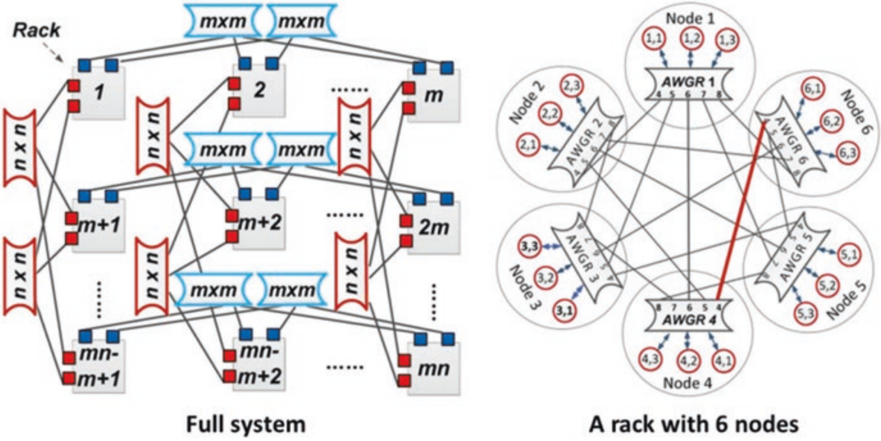


Fig. 3.9 Hi-LION full system with inter-/intra-rack AWGRs communication

3.5.2.3 Hi-LION

A large-scale interconnect optical network Hi-LION has been proposed in [28]. It exploits fast-tunable lasers and high-radix AWGRs in hierarchy to achieve very large-scale and low-latency interconnection of computing nodes. The architecture of the full system and an example of 6-node rack is depicted in Fig. 3.9. The essence is to rely on the unique wavelength routing property assisted by electrical switching embedded in the node to provide all-to-all flat interconnectivity at every level of hierarchy (node-to-node and rack-to-rack). As shown in the Fig. 3.9, the local AWGRs and global AWGRs are used to handle the intra-rack and inter-rack communications, respectively. Single-hop routing in the optical domain also avoids the utilization of optical buffers. However, the maximum hop count for inter-rack

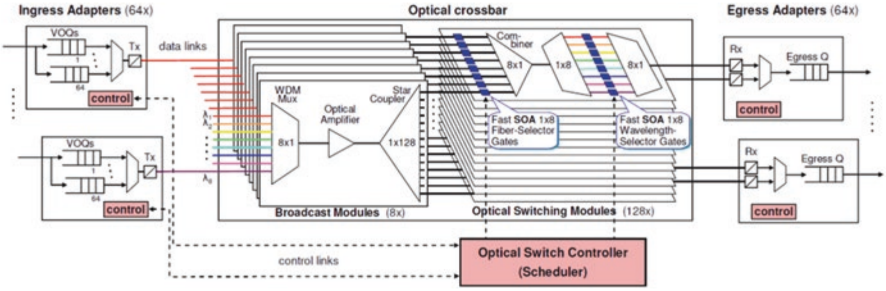


Fig. 3.10 Single-stage OSMOSIS switching system

communication can be seven including the intra-rack forwarding. Compared with previous AWGR-based solutions like DOS (LIONS) [27] and TONAK LION [40], interconnectivity of more than 120,000 nodes can be potentially connected.

3.5.2.4 OSMOSIS Optical Packet Switch

OSMOSIS project targets for accelerating the state of optical switching technology for use in supercomputers [41]. The architecture of the implemented single-stage 64-port optical packet switch is illustrated in Fig. 3.10. It is based on a broadcast-and-select architecture and the switching modules consist of a fiber and a wavelength selection stage, both built with SOAs as the gating elements. The switching of the synchronous fixed-length optical packets is controlled via a separate central scheduler. The performance studies of the OSMOSIS demonstrator confirm high-capacity and low-latency switching capabilities. Two-level fat-tree topology can be potentially built, further scaling the network to 2048 nodes.

3.5.2.5 Data Vortex

The Data Vortex is a distributed interconnection network which is entirely composed of 2 × 2 switching nodes arranged as concentric cylinders [42]. As illustrated in Fig. 3.11, the Data Vortex topology integrates internalized virtual buffering with banyan-style bitwise routing specifically designed for implementation with fiber-optic components. The 2 × 2 node uses SOA as the switching element. The broadband operation of SOA allows for successful routing of multichannel WDM packets. Contentions of packet are resolved through deflection routing. The hierarchical multistage structure is easily scalable to larger network size. However, the practical scale is limited by the increased and nondeterministic latency, as well as deteriorated signal quality.

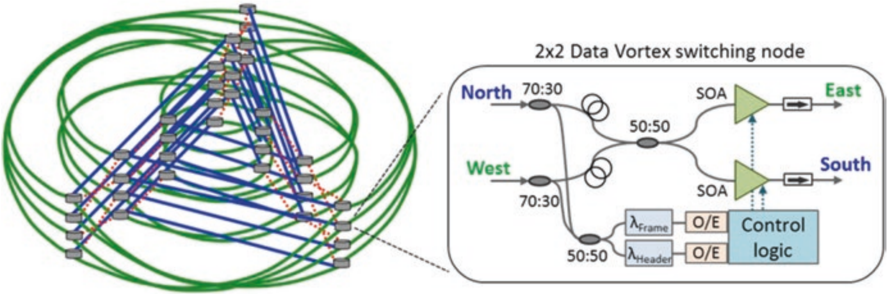


Fig. 3.11 The Data Vortex topology and distributed 2 × 2 nodes

3.6 OPSquare DCN Based on Flow-Controlled Fast Optical Switches

An optical DCN architecture OPSquare has been recently proposed [43]. Fast optical switches, which allow for flexible switching capability in both wavelength and time domains, are employed in two parallel switching networks to properly handle the intra-cluster and inter-cluster communication. Buffer-less operation is enabled by the single-hop optical interconnection and fast optical flow-control mechanism implemented between the optical switches and ToR switches. The parallel switching network also provides path diversity which is improving the resilience of the network. Benefiting from the scalability enabled by the architecture and the transmitter wavelength assignment for the grouped top-of-the-rack (ToR) switches, large interconnectivity can be achieved by utilizing moderate port count optical switches with low broadcasting ratios. The OPSquare also introduces WDM transceiver wavelength assignment for the grouped ToRs, which in combination with the wavelength switching guarantees lower broadcasting ratio for realizing the same port count. The lower splitting losses lead to less OSNR degradation and significant improvement of the scalability and feasibility of the network.

The OPSquare DCN architecture under investigation is shown in Fig. 3.12. It consists of *N* clusters and each cluster groups *M* racks. Each rack contains *K* servers interconnected via an electronic ToR switch. Two WDM bidirectional optical links are equipped at the ToR to access the parallel intra- and inter-cluster switching networks. The N $M \times M$ intra-cluster optical switches (IS) and M $N \times N$ inter-cluster optical switches (ES) are dedicated for intra-cluster and inter-cluster communication, respectively. The *i*-th ES interconnects the *i*-th ToR of each cluster, with $i = 1, 2, \ldots, N$. The number of interconnected ToRs (and servers) scales as $N \times M$, that by using moderate port count 32 × 32 ISs and ESs, up to 1024 ToRs (40,960 servers in case of 40 servers per rack) can be interconnected.

The interface for the intra-/inter-cluster communication consists of *p* WDM transceivers with dedicated electronic buffers to interconnect the ToR to the IS optical switch through the optical flow-controlled link [52], while *q* WDM transceivers interconnect the ToR to the ES optical switch. Optical packet is formed, and a copy

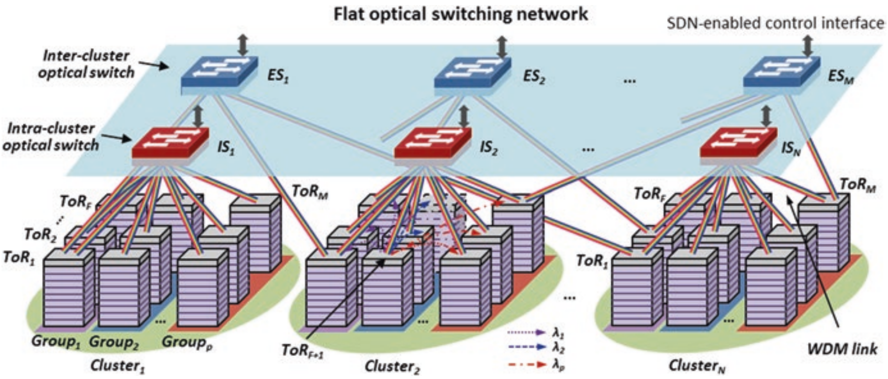


Fig. 3.12 OPSquare DCN architecture built on fast optical switches

is sent to the destination ToR via the fast optical switch. An optical in-band RF tone label is attached to the packet which will be extracted and processed at the fast optical switch node. Multiple (p and q) WDM transceivers allow for scaling the communication bandwidth between the ToRs and the optical network. Moreover, each of the WDM transceivers is dedicated for the communication with a different group of ToRs. For intra-cluster network, the M ToRs are thus divided into p groups, and each group contains $F = M/p$ ToRs. One of the p WDM TXs addresses F (instead of M) possible destination ToRs, in combination with the $1 \times F$ switch at IS. The structure and operation of the inter-cluster interface are similar to the intra-cluster ones.

The schematic of the fast optical switch node acting as IS/ES is shown in Fig. 3.13. The optical switching is realized by SOA-based broadcast-and-select architecture. The fast optical switch node has a modular structure, and each module consists of F units each of which handling the WDM traffic from one of the F ToRs in a single group. The WDM inputs are processed in parallel, and the label extractor (LE) separates the optical label from the payload. The extracted label is processed by the switch controller. The SOA has nanoseconds switching speed and can provide optical amplification to compensate the losses caused by the broadcasting stage. The contention is solved through the optical flow control, according to which ToR releases the packets stored in the buffers (for ACK) or triggers the retransmission (for NACK). Different class of priority can be applied to guarantee traffic with more stringent QoS requirement. The priority is defined by provisioning the look-up table in the switch controller through the SDN control interface [53]. In addition, the SDN control plane can create and manage multiple virtual networks over the same infrastructure by configuring the look-up table and make further optimization through the developed monitoring functions. Note also that benefiting from the WDM TX wavelength assignment for the grouped ToRs and the wavelength switching capability enabled by the optical switch node, the splitting loss of the broadcast-and-select switch is $3 \times \log_2 F$ dB ($F = M/p$ for the IS switch and $F = N/q$ for the ES switch), which is much less than $3 \times \log_2 M$ dB and $3 \times \log_2 N$ dB required in the original configuration. Lower splitting losses lead to less OSNR degradation and

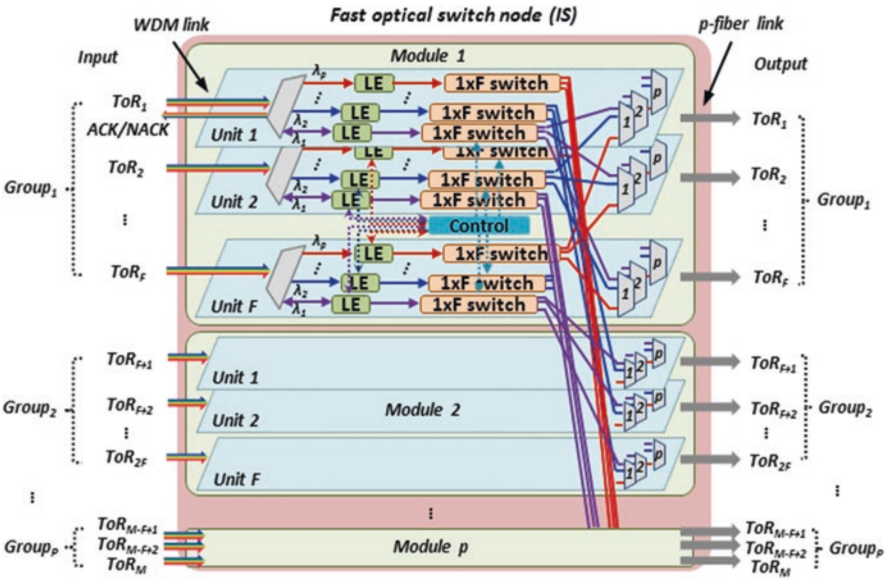


Fig. 3.13 Flow-controlled fast optical switch node

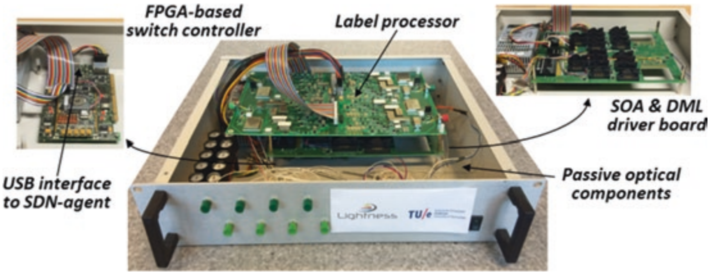

FPGA-based switch controller
Label processor
SOA & DML driver board
USB interface to SDN-agent
Passive optical components
Lightness
TU/e

Fig. 3.14 Flow-controlled fast optical switch node prototype

significant improvement of the scalability and feasibility of the network. The developed 4 × 4 optical switch prototype integrating the FPGA-based switch controller (with interface to SDN-agent), label processor, SOA drivers, and passive optical components (circulators, couplers, etc.) is shown in Fig. 3.14.

3.6.1 Performance Investigation

The performance studies of the OPSquare architecture are reported in Fig. 3.15. The DCN includes 40 servers per rack each with 10 Gb/s uplink programmed to create Ethernet packets with 40–1500 bytes length at a certain load [43]. The round-trip

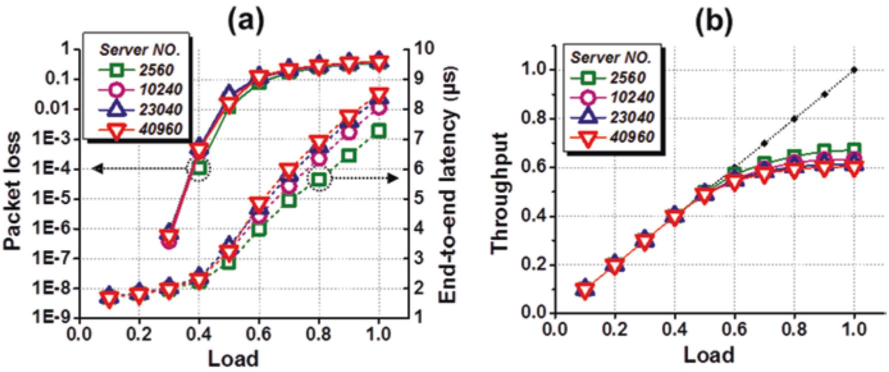


Fig. 3.15 Performance of (**a**) packet loss ratio and server end-to-end latency and (**b**) throughput

time between the ToR and optical switch node is 560 ns (2 × 50 m distance and 60 ns delay caused by the label processor as well as flow control operation). Considering that most of the traffic resides inside the cluster, four transceivers have been assigned to the IS and one for the ES ($p = 4$ and $q = 1$). For the inter-rack communication, the data packets will be forwarded to the ports associated with the intra-/inter-cluster network interface and aggregated to compose a 320 byte optical packet to be transmitted in the fixed 51.2 ns time slot. The delay caused by the head processing and buffering at the ToR input is taken as 80 ns and 51.2 ns, respectively. DCNs with variable amount of servers (from 2560 to 40,960) and racks (from 64 to 1024) with 3:1 intra-/inter-cluster traffic have been investigated. Thus, ES and IS optical switches with port count of 8 × 8 to 32 × 32 are needed to build up the desired DCN size. The buffer is set as 20 KB for each TX. The packet loss ratio and server end-to-end latency reported in Fig. 3.15a as a function of the load indicate almost no performance degradation as the number of servers increases. The packet loss ratio is smaller than 1×10^{-6}, and the server end-to-end latency is lower than 2 μs at load of 0.3 for all scales, which indicates the potential scalability of the OPSquare architecture. Similar results have been achieved for the throughput performance, as clearly shown in Fig. 3.15b.

The performance of the OPSquare architecture in terms of scalability and capacity is also investigated by using different modulation formats. The transparency to the data rate/format enabled by the fast optical switches allows for the immediate capacity upgrade maintaining the same switching infrastructure, without dedicated parallel optics and format-dependent interfaces to be further included as front end. In this respect, three types of directly modulated traffic, namely, 28 Gb/s PAM4, 40 Gb/s DMT, and 4 × 25 Gb/s NRZ OOK featuring with IM/DD, have been investigated. Exploiting the modular structure of the optical switch, the switching performance in OPSquare would mainly depend on the 1 × F broadcast-and-select switch and will be limited by the splitting loss experienced by the payload. Using the prototyped optical switch shown in Fig. 3.14, the switching performance and the port count scalability for realizing a large-scale OPSquare DCN have been assessed. Details on the experimental setup are reported in [43].

The power penalty at BER = 10^{-3} measured at different input optical powers within scale of 32 × 32 and 64 × 64 optical switch for 28 Gb/s PAM4 and 40 Gb/s DMT is depicted in Figs. 3.16a and b, respectively. An example of the optimal bit allocation after bit loading is included in Fig. 3.16b. In 32 × 32 scale, 10 dB dynamic range has been measured with <3 dB penalty, while for 64 × 64, 8 dB dynamic range has been obtained for both traffic with 3 dB penalty. With 4 WDM transceivers ($p = q = 4$) per ToR each operating at 40 Gb/s and 64 × 64 optical switches (1 × 16 broadcasting ratio) used, an OPSquare DCN comprising 4096 ToRs each with 320 Gb/s aggregation bandwidth would have a capacity >1.3 Pb/s. Larger interconnectivity can be achieved either by increasing the broadcasting ratio of the 1 × F switch with limited performance degradation or increasing the number of transceivers per ToR which could also improve the bandwidth performance.

Then the waveband switching of 4 × 25 Gb/s data payload enabled by the broadband operation of the SOA-based switch is analyzed. The power penalty at BER = 10^{-9} with different input optical powers at scales of 32 × 32, 64 × 64, and 128 × 128 optical switch when employing 4 wavebands ($p = q = 4$) is reported in Fig. 3.17c, respectively. The 16 dB input dynamic range is achieved with less than 2 dB power penalty. Here each waveband has a 100 Gb/s capacity which can be increased by inserting more wavelength channels. With four 100 Gb/s wavebands per ToR and 64 × 64 optical switches (1 × 16 broadcasting ratio), an interconnectivity of $64^2 = 4096$ ToRs and a capacity >3.2 Pb/s can be achieved benefiting from the transparency and TX wavelength assignment for the grouped ToRs featured by the OPSquare DCN.

The experimental assessments of the OPSquare DCN reported so far are based on discrete components which would result in power-inefficient bulky systems in practical implementations. Photonic integrated circuits (PICs) can reduce the footprint and the power consumption. In view of this, a 4 × 4 WDM fast optical switch PIC has been designed and fabricated exploiting the modular architecture as shown in Fig. 3.17. The modular photonic chip shown in Fig. 3.17 is a 4x16 port (the combiners shown in the schematic on the left side of Fig. 3.17 were not integrated in this photonic circuit for lack of space) and integrates four optical modules, in which each module includes four WSSs. More than 100 components including the SOAs, AWGs, and couplers are integrated in the same chip. As reported in [54, 55], the compensation of the losses offered by the SOAs allowing for large dynamic range, the low cross talk, and the wavelength, time, and switch nanosecond switch operation indicate the potential scalability to higher data rate and larger number of port counts of the optical switch PIC and potential enhancement of the OPSquare DCN performance.

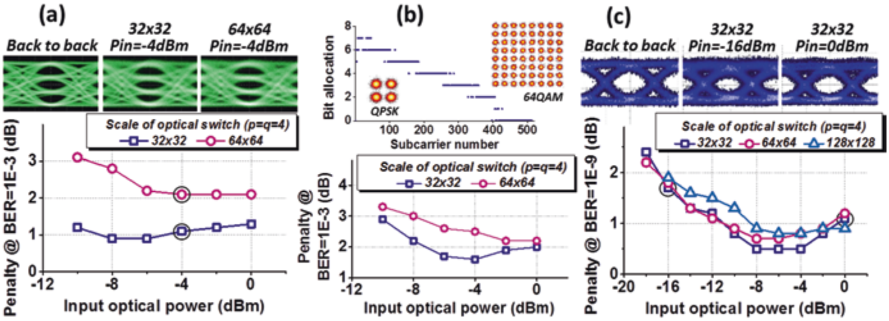


Fig. 3.16 Power penalty vs. input optical power for (**a**) 28 Gb/s PAM4, (**b**) 40 Gb/s DMT, and (**c**) waveband 4 × 25 Gb/s traffic

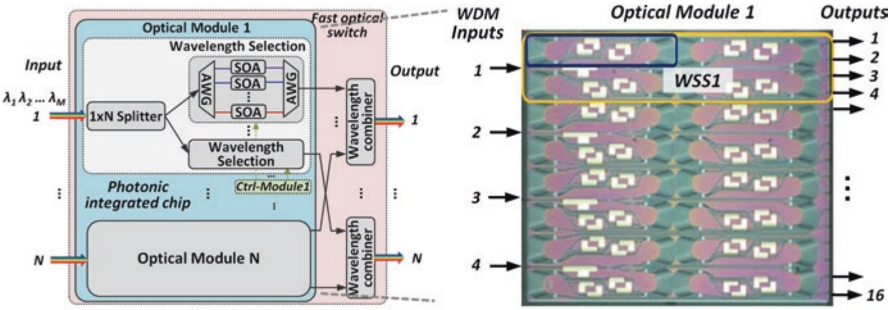


Fig. 3.17 Schematic of the fabricated 4 × 4 fast optical switch PIC

3.7 Conclusions and Discussions

The never-ending growth in the demand for high bandwidth in data centers is accelerating the deployment of more powerful severs and more advanced optical interconnects. To accommodate the increasing volume of traffic with low communication latency and high power efficiency, technological and architectural innovations of the DCNs are required. OPS/OBS based on fast optical switches is an attractive solution, by providing efficient statistical multiplexing and transparent high-capacity operation and eliminating the O/E/O conversions as well as opaque front ends. However, the lack of optical memory, the limited scalability due to the relative low port count of fast (nanoseconds) optical switches, the inefficient and no scalable centralized scheduler/control system capable to fast (tens of nanoseconds) control and configure the overall optical data plane based on fast optical switch, and the compatibility of the OPS technology with commercial Ethernet switches and protocol are some of the practical hurdles to exploit OPS and OBS in DCN. Solving those problems require complete solutions from the DCN architecture down to the devices. Promising results have been shown in the recent investigations to solve those issues. Optical DCN architectures based on OPS and OBS have been presented in this chapter, and the different characteristics in terms of scalability, flexibility, and power/cost efficiency are summarized in Table II. As can be seen in the table, for the contention resolution, most of the schemes use practical electronic buffer (EB) placed at the edge side, either waiting for the command of the scheduler or retransmitting the packet/burst in case of contention. The efficiency of the scheduling, the configuration time of the switch, and the round-trip time would play an important role in reducing the processing latency and the size of the costly buffer. It is difficult for the architectures with a single switching element to scale to large number of interconnections. In this respect, multistage and parallel topologies have been adopted by many solutions. The fast reconfiguration of the optical switches used for OPS and OBS has allowed for flexible interconnectivity which is a desired feature for DC applications. Relatively lower power/cost efficiency is the price need to pay compared with OCS technology, mainly due to the active components and the loss experienced in the switch fabrics. Performance improvement is expected, with the maturing of the fast optical switching technologies in combination with photonic integration (Table 3.2).

Table 3.2 Optical DCN architectures based on OPS and OBS technologies

	Technology	Contention resolution	Scalability	Flexibility	Power/cost efficiency
IRIS	OPS	ODL	Fair	Good	Fair
Petabit	OPS	EB at edge	Fair	Good	Fair
DOS (LIONS)	OPS	EB	Poor	Good	Fair
TONAK-LION	OPS	EB at edge	Poor	Good	Fair
Hi-LION	OPS	EB at edge	Good	Fair	Fair
OSMOSIS	OPS	EB at edge	Fair	Good	Fair
Data Vortex	OPS	Deflection	Fair	Good	Fair
OPSquare	OPS	EB at edge	Good	Good	Fair
OBS with fast optical switch	OBS	EB at edge	Poor	Fair	Good
Optical burst ring	OBS + EPS	EB at edge	Fair	Fair	Good
HOS	OPS + OBS+ OCS	EB at edge	Fair	Good	Fair
HOSA	OBS + OCS	EB at edge	Fair	Fair	Good
Torus	OPS + OCS	Deflection +ODL+ EB	Good	Good	Fair
LIGHTNESS	OPS + OCS	EB at edge	Good	Fair	Fair
Archon	OPS + OCS	EB at edge	Good	Fair	Fair
Burst over circuit	OBS + OCS	EB at edge	Poor	Fair	Good

References

1. International Data Corporation, The new need for speed in the datacenter. in *2015.International Data Corporation, “The New Need for Speed in the Datacenter*, 2015
2. S. Shah, N. Guenov, Multicore Processing: Virtualization and Data Center, Freescale, 2012
3. B. Bland, Titan – early experience with the Titan system at Oak Ridge National Laboratory. in *2012 SC Companion: High Performance Computing, Networking, Storage and Analysis*, 2012
4. G. Bell, J. Gray, A. Szalay, Petascale computational systems. IEEE Comput. **39**(1), 110–112 (2006)
5. T. Benson, A. Akella, D.A. Maltz, Network traffic characteristics of data centers in the wild. in *10th ACM SIGCOMM Conference on Internet Measurement*, 2010
6. L.A. Barroso, J. Clidaras, U. Hölzle, The datacenter as a computer: an introduction to the design of warehouse-scale machines. Synth. Lect. Comput. Archiect. **8**(3), 1–154 (2013)
7. A. Greenberg, J.R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D.A. Maltz, P. Patel, S. Sengupta, VL2: a scalable and flexible data center network. ACM SIGCOMM Comp. Commmun. Rev. **39**(4), 51–62 (2009)
8. Cisco, *Cisco Global Cloud Index: Forecast and Methodology*, 2014–2019, 2015
9. A. Roy, H. Zeng, J. Bagga, G. Porter, A.C. Snoeren, Inside the social network’s (datacenter) network. ACM SIGCOMM Comput. Commun. Rev. SIGCOMM’15 **45**(4), 123–137 (2015)
10. A. Singh, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R. Bannon, S. Boving, G. Desai, B. Felderman, P. Germano, A. Kanagala, J. Provost, J. Simmons, E. Tanda, J. Wanderer, U. Hölzle, S. Stuart, A. Vahdat, Jupiter rising: a decade of clos topologies and centralized control in Google’s datacenter Network. ACM SIGCOMM Comput. Commun. Rev. SIGCOMM’15 **45**(4), 183–197 (2015)
11. A. Greenburg, J. Hamilton, D.A. Maltz, P. Patel, The cost of a cloud: research problems in data center networks. ACM SIGCOMM Comput. Commun. Rev. **391**, 68–73 (2009)
12. A. Hammadi, L. Mhamdi, A survey on architectures and energy efficiency in Data Center Networks. Comput. Commun. **40**, 1–21 (2014)
13. A. Ghiasi, Large data centers interconnect bottlenecks. Opt. Express **23**(3), 2085–2090 (2015)
14. C. Kachris, I. Tomkos, A roadmap on optical interconnects in Data Centre Networks. in *International Conference on Transparent Optical Networks*, 2015.
15. K. Aziz, M. Fayyaz, Optical interconnects for Data Center Networks, in *Handbook on Data Centers*, (Springer, New York, 2015), pp. 449–483
16. S.J. Ben Yoo, Optical packet and burst switching technologies for the future photonic Internet. J. Lightw. Technol. **24**(12), 4468–4492 (2006)
17. Polatis Series 7000, [Online]. Available: http://www.polatis.com/series-7000-384x384-port-software-controlled-optical-circuit-switch-sdn-enabled.asp
18. CALIENT [Online]. Available: http://www.calient.net/products/s-series-photonic-switch/
19. S. Han, T.J. Seok, K. Yu, N. Quack, R.S. Muller, M.C. Wu, 50x50 polarization-insensitive silicon photonic MEMS switches: design and experiment. in *42nd European Conference on Optical Communication*, Paper Th.3.A. 5, Dusseldorf, Germany, 2016
20. K. Tanizawa, K. Suzuki, M. Toyama, M. Ohtsuka, N. Yokoyama, K. Matsumaro, M. Seki, K. Koshino, T. Sugaya, S. Suda, G. Cong, T. Kimura, K. Ikeda, S. Namiki, H. Kawashima, Ultra-compact 32 × 32 strictly-non-blocking Si-wire optical switch with fan-out LGA interposer. Opt. Express **23**, 17599–17606 (2015)
21. M. Iwama, M. Takahashi, M. Kimura, Y. Uchida, J. Hasesawa, R. Kawahara, N. Kagi, LCOS-based flexible grid 1×40 wavelength selective switch using planar lightwave circuit as spot size converter. in *2015 Optical Fiber Communications Conference and Exhibition (OFC)*, Los Angeles, CA, 2015
22. F. Testa et al, Design and implementation of an integrated reconfigurable silicon photonics switch matrix in IRIS project, J. Select. Top. Quant. Electron., vol. 22, n. 6, pp. 155-168, 2016.
23. EpiPhotonics, [Online]. Available: http://epiphotonics.com/products1.htm

24. H. Kouketsu, S. Kawasaki, N. Koyama, A. Takei, T. Taniguchi, Y. Matsushima, K. Utaka, High-speed and Compact Non-blocking 8×8 InAlGaAs/InAlAs Mach-Zehnder-Type Optical Switch Fabric. Opt. Fiber Commun. Conf. **M2K**(3) (2014)
25. H. Okayama, M. Kawahara, Prototype 32×32 optical switch matrix, *Electron. Lett.*, **30**(14), 1128–1129, l 1994
26. Y. Muranaka, T. Segawa, R. Takahashi, Integrated fat-tree optical switch with Cascaded MZIs and EAM-gate array. in *21st OptoElectronics and Communications Conference/International Conference on Photonics in Switching 2016 (OECC/PS 2016)*, paper WF3–2, Niigata, Japan, July 2016
27. Y. Yin, R. Proietti, X. Ye, C.J. Nitta, V. Akella, S.J.B. Yoo, LIONS: an AWGR-based low-latency optical switch for high-performance computing and data centers. IEEE J. Select. Top. Quant. Electron. **19**(2), 3600409 (2012)
28. Z. Cao, R. Proietti, S.J.B. Yoo, Hi-LION: hierarchical large-scale interconnection optical network with AWGRs. J. Opt. Commun. Netw. **7**(1), A97–A105 (2015)
29. T. Segawa, M. Nada, M. Nakamura, Y. Suzaki, R. Takahashi, An 8×8 broadcast-and-select optical switch based on monolithically integrated EAM-gate array. in *European Conf. Exhibition Optical Communication*, Paper TuT4.2, London, UK, 2013
30. R. Luijten, R. Grzybowski, The OSMOSIS optical packet switch for supercomputers. in *Optical Fiber Communication Conference*, 2009
31. H. Wang, A. Wonfor, K.A. Williams, R.V. Penty, I.H. White, Demonstration of a lossless monolithic 16x16 QW SOA switch. in *35th European Conference on Optical Communication*, Vienna, 2009
32. T. Tanemura, I. Soganci, T. Oyama, T. Ohyama, S. Mino, K. Williams, N. Calabretta, H.J.S. Dorren, Y. Nakano, Large-capacity compact optical buffer based on InP integrated phased-array switch and coiled fiber delay lines, *IEEE/OSA*. J. Lightwave Technol. **29**(4), 396–402 (2011)
33. C. Raffaelli, K. Vlachos, N. Andriolli, D. Apostolopoulos, J. Buron, R. van Caenegem, G. Danilewicz, J.M. Finochietto, J. Garcia-Haro, D. Klonidis, M. O'Mahony, G. Maier, A. Pattavina, P. Pavon-Marino, S. Ruepp, M. Savi, M. Scaffardi, I. Tomkos, A. Tzanakaki, L. Wosinska, O. Zouraraki, F. Neri, Photonics in switching: architectures, systems and enabling technologies. Comput. Netw. **52**(10), 1873–1890 (2008)
34. R. Stabile, A. Albores-Mejia, A. Rohit, K.A. Williams, Integrated optical switch matrices for packet data networks. Microsyst. Nanoeng. **2**, 15042 (2016)
35. M. Glick, M. Dales, D. McAuley, T. Lin, K. Williams, R. Penty, I. White, SWIFT: a testbed with optically switched data paths for computing applications. In *Proceedings of 2005 7th International Conference Transparent Optical Networks*, 2005
36. C. Kachris, I. Tomkos, A Survey on Optical Interconnects for Data Centers. IEEE Commun. Surv. Tut. **14**(4), 1021–1036 (2012)
37. C. Kachris, K. Bergman, I. Tomkos, *Optical Interconnects for Future Data Center Networks* (Springer, New York, 2013)
38. J. Gripp, J. E. Simsarian, J.D. LeGrange, P. Bernasconi, D.T. Neilson, Photonic terabit routers: the IRIS project. in *Optical Fiber Communication Conference*, 2012.
39. K. Xi, Y.-H. Kao, H.J. Chao, A Petabit bufferless optical switch for data center networks, Optical Interconnects for Future Data Center Networks, Springer New York, 2013, pp. 135–154.
40. R. Proietti, Y. Yawei, Y. Runxiang, C.J. Nitta, V. Akella, C. Mineo, S.J.B. Yoo, Scalable optical interconnect architecture using AWGR-based TONAK LION switch with limited number of wavelengths. J. Lightw. Technol. **31**, 4087–4097 (2013)
41. R. Luijten, C. Minkenberg, R. Hemenway, M. Sauer, R. Grzybowski, Viable opto-electronic HPC interconnect fabrics. in *Proceedings of Supercomputing 2005*, Seattle, 2005.
42. O. Liboiron-Ladouceur, A. Shacham, B.A. Small, B.G. Lee, H. Wang, C.P. Lai, A. Biberman, K. Bergman, The Data Vortex optical packet switched interconnection network. J. Lightw. Technol. **26**(13), 1777–1789 (2008)

43. W. Miao, F. Yan, N. Calabretta, Towards petabit/s all-optical flat data center networks based on WDM optical cross-connect switches with flow control. J. Lightw. Technol. **34**(17), 4066–4075 (2016)
44. M. Imran, M. Collier, P. Landais, K. Katrinis, Software-defined optical burst switching for HPC and cloud computing data centers. J. Opt. Commun. Netw. **8**(8), 610–620 (2016)
45. C.Y. Li, N. Deng, M. Li, Q. Xue, P.K.A. Wai, Performance analysis and experimental demonstration of a novel network architecture using optical burst rings for interpod communications in data centers. IEEE J. Select. Top. Quant. Electron. **19**(2), 3700508–3700508 (2013)
46. M. Fiorani, S. Aleksic, M. Casoni, Hybrid optical switching for data center networks. J. Electric. Comput. Eng. **2014**, 1–13 (2014)
47. M. Imran, M. Collier, P. Landais, K. Katrinis, HOSA: hybrid optical switch architecture for data center networks. in *12th ACM International Conference on Computing Frontiers*, 2015
48. K.-I. Kitayama, Y.-C. Huang, Y. Yoshida, R. Takahashi, T. Segawa, S. Ibrahim, T. Nakahara, Y. Suzaki, M. Hayashitani, Y. Hasegawa, Y. Mizukoshi, A. Hiramatsu, Torus-Topology Data Center Network Based on Optical Packet/Agile Circuit Switching with Intelligent Flow Management. J. Lightw. Technol. **33**(5), 1063–1071 (2015)
49. LIGHTNESS Project, [Online]. Available: http://www.ict-lightness.eu/
50. S. Yan, E. Hugues-Salas, V.J.F. Rancaňo, Y. Shu, G.M. Saridis, B.R. Rofoee, Y. Yan, A. Peters, S. Jain, T. May-Smith, P. Petropoulos, D.J. Richardson, G. Zervas, D. Simeonidou, Archon: A Function Programmable Optical Interconnect Architecture for Transparent Intra and Inter Data Center SDM/TDM/WDM Networking. J. Lightw. Technol. **33**(8), 1586–1595 (2015)
51. Q. Huang, Y. Yeo, and L. Zhou, Optical burst-over-circuit switching for multi-granularity traffic in data centers. in *Optical Fiber Communication Conference/National Fiber Optic Engineers Conference 2013*, paper OW3H.5, 2013
52. W. Miao, S. Di Lucente, J. Luo, H. Dorren, N. Calabretta, Low latency and efficient optical flow control for intra data center networks. Opt. Express **22**(1), 427–434 (2014)
53. W. Miao, F. Agraz, S. Peng, S. Spadaro, G. Bernini, J. Perello, G. Zervas, R. Nejabati, N. Ciulli, D. Simeonidou, H. Dorren, N. Calabretta, SDN-enabled OPS with QoS guarantee for reconfigurable virtual data center networks, *IEEE/OSA*. J. Opt. Commun. Netw. **7**(7), 634–643 (2015)
54. N. Calabretta, K. Williams, H. Dorren, Monolithically integrated WDM cross-connect switch for nanoseconds wavelength, space, and time switching. in *2015 European Conference on Optical Communication (ECOC)*, 2015
55. N. Calabretta, W. Miao, K. Mekonnen, K. Prifti, K. Williams, Monolithically integrated WDM cross-connect switch for high-performance optical data center networks. in *Optical Fiber Communication Conference 2017 (OFC 2017)*, paper Tu3F.1, 2017

Part II
Demonstrations of Optical Switching in Data Center

Chapter 4
OSA: An Optical Switching Architecture for Data Center Networks with Unprecedented Flexibility

Kai Chen, Ankit Singla, Atul Singh, Kishore Ramachandran, Lei Xu, Yueping Zhang, Xitao Wen, and Yan Chen

4.1 Introduction

Many online services, such as those offered by Amazon, Google, Facebook, and eBay, are powered by massive data centers hosting hundreds of thousands of servers. The network interconnect of the data center plays a key role in the performance and scalability of these services. As the number of hosted applications and the amount of traffic grow, the industry is looking for larger server pools, higher bit-rate network interconnects, and smarter workload placement approaches to satisfy the demand. To meet these goals, a careful examination of traffic characteristics, operator requirements, and network technology trends is critical.

Traffic Characteristics Several recent DCN proposals attempt to provide uniformly high capacity between all servers [1–4]. Given that it is not known a priori which servers will require high-speed connectivity, for a static, electrical network, this appears to be the only way to prevent localized bottlenecks. However, for many real scenarios, such a network may not be fully utilized at all times. For instance, measurement on a 1500-server Microsoft production DCN reveals that only a few ToRs are hot and most of their traffic goes to a few other ToRs [5] . Likewise, an analysis of high-performance computing applications shows that the bulk of interprocessor traffic is degree-bounded and slowly changing [6] . Thus, even for a few

K. Chen (✉)
Hong Kong University of Science and Technology, Kowloon, Hong Kong
e-mail: kaichen@cse.ust.hk

A. Singla
University of Illinois at Urbana–Champaign, Urbana, IL 61801, USA

A. Singh • K. Ramachandran • L. Xu • Y. Zhang
NEC Labs, Princeton, NJ 08540, USA

X. Wen • Y. Chen
Northwestern University, Evanston, IL 60208, USA

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
DOI 10.1007/978-3-319-61052-8_4

thousand servers, uniformly high capacity networks appear to be an overkill. As the size of the network grows, this weighs on the cost, power, and wiring complexity of such networks.

Dealing with the Oversubscribed Networks Achieving high performance for data center services is challenging with oversubscribed networks. One approach is to use intelligent workload placement algorithms to allocate network-bound service components to physical hosts with high bandwidth connectivity [7], e.g., placing these components on the same rack. Such workloads exist in practice: dynamic creation and deletion of VM instances in Amazon's EC2 or periodic backup services running between an EC2 (compute) instance and an S3 (storage) bucket. An alternate approach is to flexibly allocate more network bandwidth to service components with heavy communications. If the network could "shape-shift" in such fashion, this could considerably simplify the workload placement problem.

Higher Bit-Rates There is an increasing trend toward deploying 10 GigE NICs at the end hosts. In fact, Google already has 10 GigE deployments and is pushing the industry for 40/100 GigE [8–10]. Deploying servers with 10 GigE naturally requires much higher capacity at the aggregation layers of the network. Unfortunately, traditional copper wire 10 GigE links are not viable for distances over 10 m [11], due to the high electrical loss at higher data rate, necessitating the need to look for alternative technologies.

The optical networking technology is well suited to meet the above challenges. Optical network elements support on-demand connectivity and capacity where required in the network, thus permitting the construction of thin but flexible interconnects for large server pools. Optical links can support higher bit-rates over longer distances using less power than copper cables. Moreover, optical switches run cooler than electrical ones [12], resulting in lower heat dissipation and cheaper cooling cost. The long-term advantage of optics in DCNs has been noted in the industry [12, 13].

Recent efforts in c-Through [14] and Helios [11] provide a promising direction to exploit the optical networking technology (e.g., one-hop high-capacity optical circuits) for building DCNs. Following this trailblazing research, we present OSA, a novel optical switching architecture for DCNs. OSA achieves high flexibility by leveraging and extending the techniques devised by prior works and further combining them with novel techniques of its own. Similar to prior works, OSA leverages reconfigurability of optical devices to dynamically set up one-hop optical circuits. Then, OSA employs hop-by-hop stitching of multiple optical links to provide overall connectivity for mice flows and bursty communications and to handle workloads involving high fan-in/fan-out hotspots [15] that the existing one-hop electrical/optical architectures cannot address efficiently via their optical interconnects. Furthermore, OSA dynamically adjusts the capacities on the optical links to satisfy changing traffic demand at a finer granularity.

We build a small-scale eight-rack OSA prototype with real optical devices and server-emulated ToRs. Through this test-bed, we evaluate the performance of OSA with all software and hardware overheads. We find that OSA can quickly adapt the topology and link capacities to the changing traffic patterns, and our results show that it achieves nearly 60% of the non-blocking bandwidth in all-to-all communications. We also measure the device characteristics of the optical equipment, evaluate the impact of multi-hop optical-electrical-optical (O-E-O) conversion, and discuss our experience building and evaluating the OSA prototype.

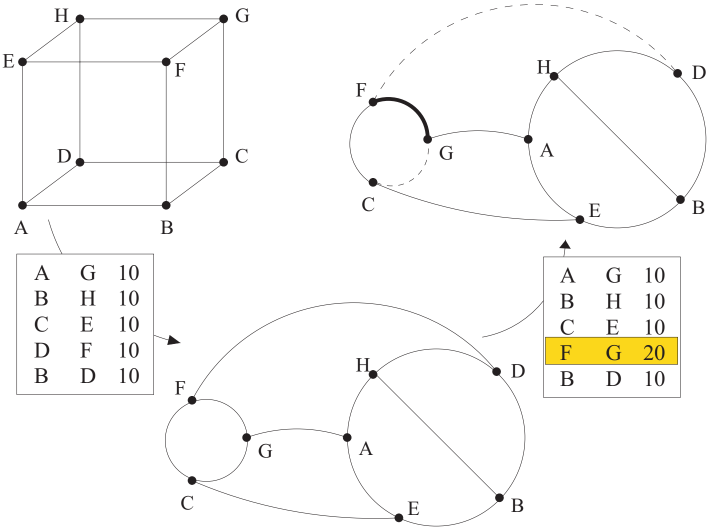
4.2 Motivation and Background

We first use a motivating example to show what kind of flexibility OSA can deliver. Then, we introduce the optical networking technologies that make OSA possible.

4.2.1 A Motivating Example

We discuss the utility of a flexible network using a simple hypothetical example in Fig. 4.1. On the left is a hypercube connecting eight ToRs using 10G links. The traffic demand is shown in the bottom left of Fig. 4.1. For this demand, no matter what routing paths are used on this hypercube, at least one link will be congested. One way to tackle this congestion is to reconnect the ToRs using a different topology (Fig. 4.1, bottom center). In the new topology, all the communicating ToR pairs are directly connected, and their demand can be perfectly satisfied.

Now, suppose the traffic demand changes (Fig. 4.1, bottom right) with a new (highlighted) entry replacing an old one. If no adjustment is made, at least one link



A	G	10
B	H	10
C	E	10
D	F	10
B	D	10

A	G	10
B	H	10
C	E	10
F	G	20
B	D	10

Fig. 4.1 OSA adapts the topology and link capacities to the changing traffic

will face congestion. With the shortest path routing, F↔G will be that link. In this scenario, one way to avoid congestion is to increase the capacity of F↔G to 20G at the expense of decreasing the capacities of F↔D and G↔C to 0. Note that in all these cases, the node degree remains the same (i.e., 3) and the node capacity is no more than 30G.

As above, OSA's flexibility lies in its flexible topology and flexible link capacity. In the absence of such flexibility, the above example would require additional links and capacities to handle both traffic patterns. More generally, a large variety of traffic patterns would necessitate the non-blocking network construction. OSA, with its high flexibility, can avoid such non-blocking construction, while still providing equivalent performance for various traffic patterns.

4.2.2 Optical Networking Technologies

We discuss the optical networking technologies that enable the above flexibility.

1. *Wavelength division multiplexing (WDM)*. Within C-band (conventional band) and with 100 GHz DWDM channel spacing, typically 40 or more wavelength channels can be transmitted over a single optical fiber. For the purpose of our architecture, each wavelength is rate-limited by the electrical port it connects to.
2. *Wavelength selective switch (WSS)*. A WSS is typically a 1 × k switch, consisting of one common port and k wavelength ports. It partitions (runtime-configurable within a few ms) the set of wavelengths coming in through the common port among the N wavelength ports. For example, if the common port receives 80 wavelengths, it can route wavelengths 1–20 on port 1, 30–40 on port 2, etc.
3. *Optical switching matrix (OSM)*. Most OSM modules are bipartite $N \times N$ switching matrix where any input port can connect to any of the output ports. Microelectromechanical system (MEMS) is the most popular OSM technology and achieves reconfigurable (at 10 ms [16]) one-to-one circuit by mechanically adjusting micro mirrors. A few hundred ports are common for commercial products and >1000 for research prototypes [17] . The current commercially available OSM modules are typically oblivious to the wavelengths carried across it. We use MEMS and OSM interchangeably.
4. *Optical circulators*. Circulators enable bidirectional optical transmission over a fiber, allowing more efficient use of the ports of optical switches. An optical circulator is a three-port device: one port is a shared fiber or switching port, and the other two ports serve as send and receive ports.
5. *Optical transceivers*. Optical transceivers can be of two types: coarse WDM (CWDM) and dense WDM (DWDM). We use DWDM-based transceivers in OSA, which support higher bit-rates and more wavelength channels in a single piece of fiber compared to CWDM.

4.3 OSA Network Architecture

We introduce how OSA architecture is built from the above described optical networking technologies. Our current design is intended for container-size DCNs with thousands of servers.

4.3.1 Building Blocks

Flexible Topology OSA achieves the flexible topology by exploiting the reconfigurability of the MEMS. If we start by connecting each of *N* ToRs to one port on an *N*-port MEMS, each ToR can only communicate with one other ToR at any instant given the MEMS's bipartite port-matching. If we connect *N*/*k* ToRs to *k* ports each at the MEMS, each ToR can communicate with *k* other ToRs simultaneously. Here, $k > 1$ is the degree of the ToR, not its port count, in the ToR graph. The MEMS configuration determines which set of ToRs are connected. OSA must ensure that the ToR graph is connected when configuring the MEMS.

Given a ToR topology connected by the MEMS optical circuits, we use hop-by-hop stitching of such circuits to achieve network-wide connectivity. To reach remote ToRs that are not directly connected, a ToR uses one of its *k* connections. This first-hop ToR receives the transmission over fiber, converts it to electrical signals, reads the packet header, and routes it toward the destination. At each hop, the packet experiences the conversion from optics to electronics and then back to optics (O-E-O) and the switching at the ToR. Note that at any port, the aggregate transit, incoming and outgoing traffic, cannot exceed the port's capacity in each direction. So, the high-volume connections must use a minimal number of hops. OSA should manage the topology to adhere to this requirement. Evaluation in Sec. 5 quantifies the overhead (both O-E-O and switching) of the hop-by-hop routing.

Flexible Link Capacity In OSA, each ToR connects to *k* other ToRs. If each link has a fixed capacity, multiple links may be required for this ToR to communicate with another ToR at a rate higher than a single link can support. To overcome this problem, OSA combines the capability of optical fibers to carry multiple wavelengths at the same time (WDM) with the dynamic reconfigurability of the WSS. Consequently, each ToR is connected to the MEMS through a multiplexer (MUX) and a WSS unit (Fig. 4.2).

Specifically, suppose ToR1 wants to communicate with ToR2 using *w* times the line speed of a single port. The ToR will use *w* ports, each associated with a unique wavelength, to serve this request. The WDM enables these *w* wavelengths, together with the rest from this ToR, to be multiplexed into one optical fiber that feeds the WSS. The WSS splits these *w* wavelengths to the appropriate MEMS port which has a circuit to ToR2 (doing likewise for the rest $k - 1$ groups of wavelengths). Thus, a $w \times$ (*line speed*) capacity circuit is set up from ToR1 to ToR2, at runtime. By varying the value of *w* for each MEMS circuit, OSA achieves dynamic link capacity.

We note that a fiber cannot carry two channels over the same wavelength in the same direction. Moreover, to enable a pair of ToRs to communicate using all available wavelengths, we require that each ToR port (facing the optical interconnect) is associated with a wavelength that is unique across the ports of a ToR. The same wavelength is used to receive traffic as well: each port thus sends and receives traffic at one fixed wavelength. This allows all wavelengths at a source ToR to be multiplexed and delivered, after demultiplexing, to individual ports at the destination ToRs.

Efficient Port Usage To make full use of the MEMS ports, we require that each circuit over the MEMS be bidirectional. For this purpose, we use optical circulators between the ToR and the MEMS ports. A circulator connects the send channel of the transceiver from a ToR to the MEMS (after the channel has passed through the MUX and WSS). It simultaneously delivers the incoming traffic toward a ToR from the MEMS (through the coupler and DEMUX) to this ToR. Note that even though the MEMS links are bidirectional, the capacities of the two directions are independent of each other.

4.3.2 Putting It All Together: OSA-2560

Figure 4.2 illustrates the general OSA architecture. We now discuss one specific instantiation – OSA-2560 with $N = 80$ ToRs, $W = 32$ wavelengths, and ToR degree $k = 4$ using a 320-port MEMS to support 2560 servers.

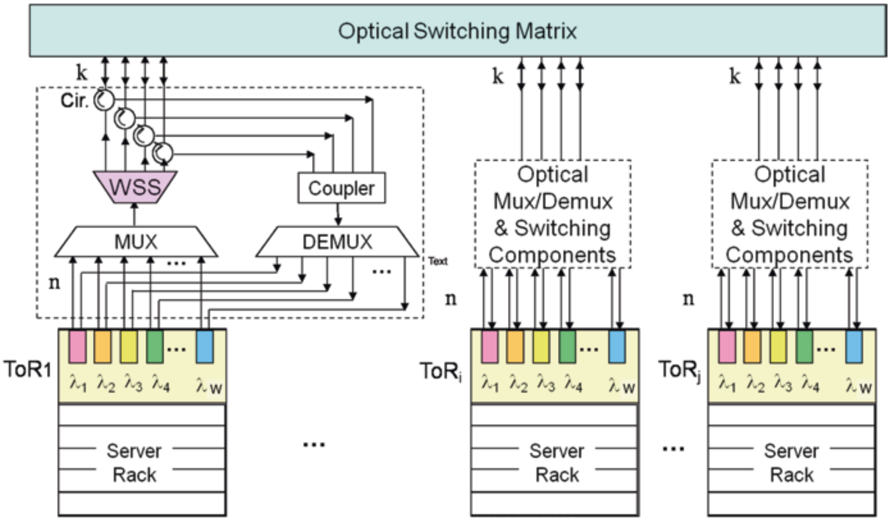

Optical Switching Matrix
k
Cir.
WSS
Coupler
MUX
DEMUX
Text
n
ToR1
λ_1 λ_2 λ_3 λ_4 λ_W
Server Rack
k
Optical Mux/Demux & Switching Components
n
ToR$_i$
λ_1 λ_2 λ_3 λ_4 λ_W
Server Rack
k
Optical Mux/Demux & Switching Components
n
ToR$_j$
λ_1 λ_2 λ_3 λ_4 λ_W
Server Rack

Fig. 4.2 The overall OSA, detailed structure is shown only for ToR1 for clarity

Each ToR is a commodity electrical switch with 64 10-GigE ports [18]: 32 of these ports are connected to servers, while the remaining are connected to the optical interconnect. Each port facing the optical interconnect has a transceiver associated with a fixed and unique wavelength for sending and receiving data. The transceiver uses separate fibers to connect to the send and receive infrastructures.

The send fiber from the transceiver from each of the 32 ports at a ToR is connected to an optical MUX. The MUX feeds a 1 × 4 WSS. The WSS splits the set of 32 wavelengths it sees into four groups, each group being transmitted on its own fiber. These fibers are connected to the MEMS via circulators to enable bidirectional communications. The four receive fibers from four circulators are connected to a power coupler (similar to a multiplexer but simpler), which combines their wavelengths onto one fiber. This fiber feeds a demultiplexer (DEMUX), which assigns each incoming wavelength to its associated port on the ToR.

We point out two key properties of the above interconnect. First, each ToR can communicate simultaneously with any four other ToRs. This implies that the MEMS configuration allows us to construct all possible four regular graphs among ToRs. Second, through WSS configuration, the capacity of each of these four links can be varied in $\{0, 10, 20, \ldots 320\}$ Gbps. The MEMS and WSS configurations are decided by a central OSA manager. The manager estimates the traffic demand, calculates the appropriate configurations, and pushes them to the MEMS, WSS units, and ToRs. This requires direct, out-of-band connections between the OSA manager and those components. Note that our employment of such a central OSA manager is inspired by many recent works [2, 3, 11, 14, 19] in the context of DCNs given the fact that a DCN is usually owned and operated by a single organization.

Furthermore, we choose $k = 4$ for container-size DCNs because it is a trade-off between the network size and performance. A larger k value can enable one ToR to connect to more other ToRs simultaneously, thus achieving higher performance. However, given a 320-port MEMS, it also means that fewer ToRs ($320/k$) can be supported. Our experiments with $k = 1,2,4,8$ indicate that $k = 4$ can deliver considerable bisection bandwidth between thousands of servers.

4.4 Network Optimization

We describe OSA network optimization in detail. Our goal is to compute the optimal topology and link capacities so that the network bisection bandwidth is maximized for a given traffic demand. We need to find 1) a MEMS configuration to adjust the topology to localize high traffic volumes, 2) routes between ToRs to achieve high throughput and low latency or avoid congestion, and 3) a configuration for each WSS to provision the capacities of its outgoing links.

In the following, we first present a mathematical formulation for optimization. Considering its complexity, we then introduce an approximation solution.

4.4.1 Problem Formulation

Given: A traffic demand D between ToRs – d_{ij} is the desired bandwidth from ToR$_i$ to ToR$_j$.

Variables: We use four sets of variables: $l_{ij} = 1$ if ToR$_i$ is connected to ToR$_j$ through MEMS and 0 otherwise; $w_{ijt} = 1$ if l_i carries wavelength λ_t in the $i \rightarrow j$ direction and 0 otherwise; v_{ijt} is the traffic volume carried by wavelength λ_t along $i \rightarrow j$; a traffic-served matrix $S - s_{ij}$ is the bandwidth achieved from ToR$_i$ to ToR$_j$. For the last two sets of variables, s_{ij} have end-to-end meaning, while v_{ijt} have hop-to-hop significance. For all the variables, $t \in \{1,2,\ldots,w\}$; $i,j \in \{1,2,\ldots,N\}$, $i \neq j$; l_{ij} are the only variables for which $l_{ij} = l_{ji}$; and all the other variables are directional.

Objective: To achieve the optimal network bisection bandwidth, we maximize the traffic served:

$$\text{Maximize} \sum_{i,j} s_{ij}. \tag{4.1}$$

Constraints: If the number of outgoing ports of the WSS is k, then ToR$_i$ is connected to exactly k other ToRs:

$$\forall i : \sum_j l_{ij} = k \tag{4.2}$$

A wavelength λ_t can only be used between two ToRs if they are directly connected via MEMS:

$$\forall i, j, t : w_{ijt} \leq l_{ij}. \tag{4.3}$$

To avoid wavelength contention, ToR$_i$ can only receive/send λ_t from/to at most one ToR:

$$\forall i, t : \sum_j w_{jit} \leq 1; \quad \sum_j w_{ijt} \leq 1. \tag{4.4}$$

Traffic carried by λ_t between two ToRs is limited by ToR port capacity (C_{port}) and wavelength capacity (C_λ):

$$\forall i, j, t : v_{ijt} \leq \min\left\{C_{\text{port}}, C_\lambda \times w_{ijt}\right\}. \tag{4.5}$$

The outgoing transit traffic is equal to the incoming transit traffic at ToR$_i$:

$$\forall i : \sum_{j,t} v_{ijt} - \sum_j s_{ij} = \sum_{j,t} v_{jit} - \sum_j s_{ji}. \tag{4.6}$$

Finally, the traffic served is bounded by the demand:

$$\forall i, j : \mathrm{s}_{ij} \leq d_{ij}. \tag{4.7}$$

The above mixed-integer linear program (MILP) can be viewed as a maximum multi-commodity flow problem with degree bounds, further generalized to allow constrained choices in link capacities. While several variants of the degree-bounded subgraph and maximum flow problems have known polynomial time algorithms, the trivial combination of two is NP-hard [20].

4.4.2 Solution

In our approximation solution, we decompose the problem into three sequential subparts as shown in Fig. 4.3, i.e., computing the topology, computing the routing, and computing the wavelength assignment. We adopt the traffic demand estimation introduced in Hedera [21], which is based on the max-min fair bandwidth allocation for TCP flows in an ideal non-blocking network.

1. *Compute the topology*. We localize high-volume communicating ToR pairs over direct MEMS circuit links. This is accomplished by using a weighted b-matching [22], where b represents the number of ToRs that a ToR connects to via MEMS ($b = k = 4$ in OSA-2560). In the ToR graph, we assign the edge weight between two ToRs as the estimated demand between them and then cast the problem of localizing high-volume ToR connections to b-matching. The weighted b-matching is a graph theoretic problem for which polynomial time algorithm exists [22] . We implement it using multiple perfect matchings, for which public library is available [23].

   The b-matching graph above is not necessarily a connected graph. Fortunately, connectivity is easy to achieve via the edge-exchange operation [24] . First, we find all the connected components. If the graph is not connected, we select two edges $a \rightarrow b$ and $c \rightarrow d$ with lowest weights in different connected components and connect them via replacing links $a \rightarrow b$ and $c \rightarrow d$ with links $a \rightarrow c$ and $b \rightarrow d$. We make sure that the links removed are not themselves cuts in the graph. The output of Step 2 is used to tell the MEMS about how to configure the new topology.

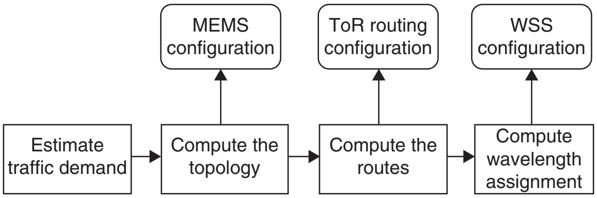


Fig. 4.3 The steps in the OSA control algorithm

2. *Compute the routes*. Once we have connectivity, the MEMS configuration is known. We proceed to compute the routes using any of the standard routing schemes such as the shortest path routing or low congestion routing. Note that some of the routes are single-hop MEMS connections while the others are multi-hop ones. For simplicity, we use the shortest path routing here. However, our framework can be readily applied to other routing schemes. The output of Step 3 is used by the ToRs to configure their routing tables.
3. *Compute the wavelength assignment*. Given the traffic demand and routes among ToRs, we compute the capacity desired on each ToR link in order to serve the traffic demand on this link.

With the desired capacity demand on each link, we need to provision a corresponding amount of wavelengths to serve the demand. However, wavelength assignment is not arbitrary: due to the contention, a wavelength can only be assigned to a ToR at most once. Given this constraint, we reduce the problem to be the edge-coloring problem on a multigraph. We represent our ToR level graph as a multigraph. Multiple edges correspond to the number of wavelengths between two nodes, and we assume each wavelength has a unique color. Thus, a feasible wavelength assignment is equivalent to an assignment from the colors to the edges of the multigraph so that no two adjacent edges have the same color – exactly the edge-coloring problem [25] . The edge coloring is a known problem and fast heuristics are known [26] . Libraries implementing this are publicly available.

We also require at least one wavelength to be assigned to each edge on the physical topology. This guarantees an available path between any pair of ToRs, which may be required for mice/bursty flows. The output of Step 4 is used by the WSS to assign wavelengths.

All the above steps are handled by the OSA manager. Specifically, the OSA manager interacts with the MEMS, WSS units, and ToRs to control the topology, link capacities, and routing, respectively. We note that our decomposition heuristic is not optimal and there is room for further improvement.

4.5 Implementation

We have built a small-scale OSA prototype with real optical devices (Fig. 4.4). We first introduce our test-bed setup and then present our experiments over it.

4.5.1 Test-bed Setup

Our test-bed connects 32 end hosts, uniformly distributed in eight racks. To reduce the cost, we configure eight Dell OptiPlex servers to emulate 32 end hosts. Each server acts as a virtual rack of end hosts (V-Rack), running four virtual machines (VMs) to emulate four end hosts.

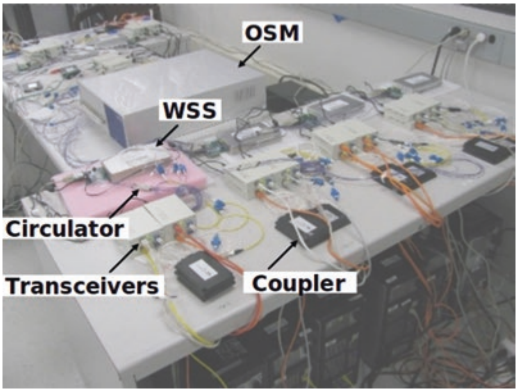


Fig. 4.4 OSA test-bed

We now do not have programmable ToR switches, so we use high-end servers to emulate ToRs. We have four Dell PowerEdge servers, each equipped with an Intel 2.4 GHz quad-core CPU, 8GB DRAM, and 12 × 1 GigE NICs. On each such server, we deploy two VMs, giving us a total of eight virtual ToRs (V-ToRs). Each V-ToR binds to six NICs: one is connected to one V-Rack, one is used for a control connection to the OSA, and the remaining four are used as uplinks to reach other V-ToRs via optical elements.

On top of each V-ToR is a 1 × 4 CoAdna WSS, a coupler, a circulator, a 1 × 4 MUX and DEMUX pair, and four transceivers (which are packaged into a media converted (MC) unit). As in Fig. 4.2, each ToR uplink is connected to a transceiver, with the send fiber of the transceiver connected through the MUX, the WSS and the circulator to the OSM, and the receive fiber connected to the same circulator through the coupler and the DEMUX. We use one Polatis series-1000 OSM (a piezoelectric switch) with 32 ports which allows a 16 × 16 bipartite interconnect. (Each V-ToR has two uplinks connected to each of these two sets of 16 ports.) We use four wavelengths, 1545.32, 1544.53, 1543.73, and 1542.94 nm, corresponding to channel 40, 41, 42, and 43 of ITU grid with 100 GHz channel spacing.

Furthermore, in our test-bed, the OSA manager is a separate Linux server and talks to the OSM and ToRs via Ethernet ports and to the WSS units via RS-232 serial ports.

4.5.2 Understanding the Optical Devices

There are two critical optical devices in OSA: OSM and WSS. A common concern for them is the reconfiguration overhead. To measure the overhead, Fig. 4.5 shows the output power level on two ports of the OSM over time, during a reconfiguration event. We see a clear transition period, i.e., from the high → low output power level shift on one port to the low → high output power level shift on the other port. We observe that the switching delay of our OSM is 9 ms, consistent with [11, 14].

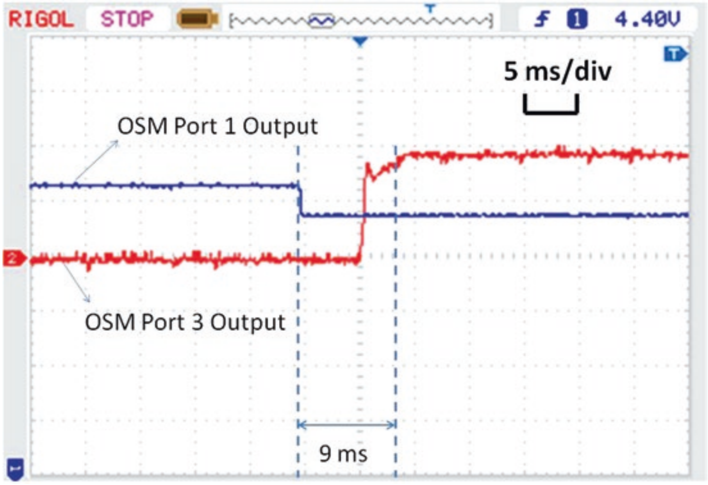


Fig. 4.5 Switching time of our OSM

Next, we measure the reconfiguration time of the WSS by switching a wavelength channel between two output ports. As shown in Fig. 4.6, this transition period is around 14 ms. However, the OSA manager can perform the reconfiguration of OSM and WSS in parallel to reduce the total time of reconfiguration.

4.5.3 *Understanding the O-E-O Conversion*

To measure the impact of O-E-O conversion, we specially connect four servers as in Fig. 4.7 (left). Two servers in the middle are configured as routers and equipped with optical media converters. We create a routing loop by configuring the IP forwarding tables of the routers. In each router, we deploy a `netfilter` kernel module and utilize the `NF_IP_PRE_ROUTING` hook to intercept all IP packets. We record the time lag between the instant when the packets first arrive in the network and when their TTL expires. This way, we are able to measure the multi-hop latency for O-E-O conversion and compare it with the baseline where all the servers are directly connected using only electrical devices. Results in Fig. 4.7 (right) compare the average one-hop switching latency for both the hybrid optical/electrical and pure electrical architectures under different traffic loads. It is evident from the figure that the O-E-O conversion does not incur noticeable (the maximum deviation in the absolute value and standard deviation is 38 and 58 μs, respectively), if any, additional switching latency, demonstrating the feasibility of O-E-O employed by OSA.

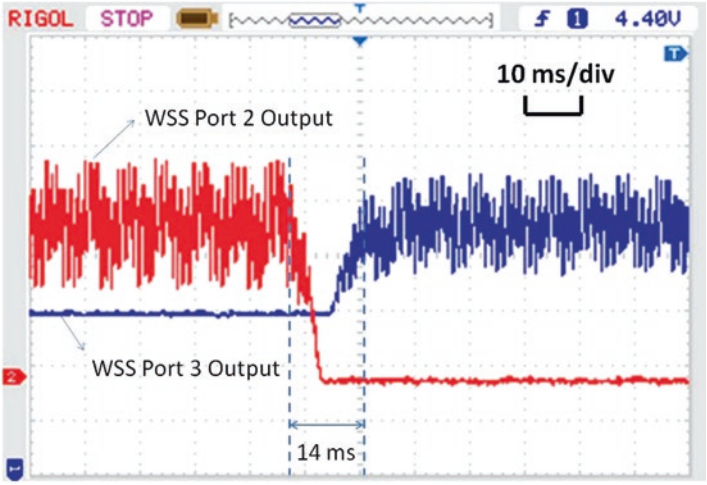


Fig. 4.6 Switching time of our WSS

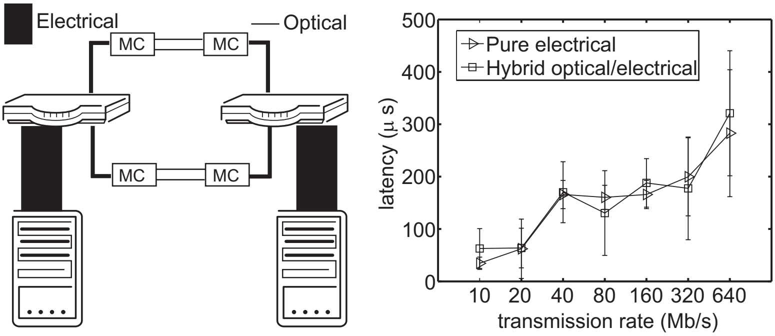


Fig. 4.7 The impact of O-E-O conversion

4.5.4 OSA System Performance

We conduct two sets of experiments: one is for original OSA and the other is for OSA with static topology. We use synthetic traffic pattern which is denoted by parameters (t, r): servers in ToR i ($i = 0\ldots7$) send traffic to servers in t ToRs, i.e., $[i + r, i + r + 1, ..., i + r + (t - 1)]$. We change t from 1 to 7 to generate different traffic loads ($t = 7$ means all-to-all communication). For each t, we vary r from 1 to 7.

Our goal is to compare the achieved bisection bandwidth of OSA against that of a non-blocking network as the traffic spreads out (with increasing t) and to measure the effect of topology reconfiguration. Note that varying r with a fixed t does not produce fundamentally different traffic distributions, as it merely permutes which

ToRs talk with which other ToRs, thus necessitating a change of topology without a change in traffic load or spread.

In our test-bed, the server NICs support 10, 100, and 1000 Mbps full-duplex modes. In all our experiments, we limit the maximum sending rate of each flow to be 100 Mbps. This enables us to emulate a non-blocking network for comparison (Fig. 4.8): for OSA, all the uplink ports of ToRs are set at 100 Mbps, while for the non-blocking, we increase some particular uplink ports to be 1000 Mbps to satisfy the traffic demands we simulate.

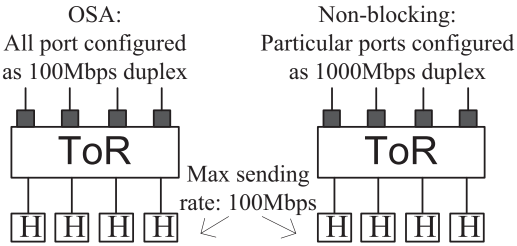
Results of OSA Figure 4.9 shows the average bisection bandwidth of OSA with changing traffic ($t = 1 \cdots 7$). For each t, r steps 1 through 7 every 20 seconds. The network topology is dynamically reconfigured according to the current traffic demand. The results are along expected lines. We observe that the achieved bisection bandwidth of OSA is within 95% of the non-blocking network when t is 1 or 2. This is because when $t = 1$ each ToR talks with two other ToRs and when $t = 2$ each ToR talks with four other ToRs. Given that our topology is a four-regular graph, OSA assigns direct links to each pair of communicating ToRs for efficient communication. For $t > 2$, as expected, the performance of OSA is gradually decreasing because the traffic needs to go through multiple hops before reaching destinations. We find that the performance of OSA under the all-to-all communication is 58% of the non-blocking. We expect that the performance under all-to-all will further degrade in larger networks. However, our intention of the test-bed results is to demonstrate the feasibility of OSA rather than to show the performance achieved in a real deployment.

Next, Fig. 4.10 shows the impact of optical device reconfigurability on the end-to-end throughput between two hosts. We observe that the performance drops during reconfiguration but quickly resumes after it finishes.

We also present the theoretical bisection bandwidth achievable in our test-bed (Fig. 4.9) that ignores the overhead of reconfiguration, software routing, TCP/IP protocol, etc. We observe that the gap between the theoretical values and OSA is within 5%–7%, suggesting that our prototype implementation is efficient.

Results of OSA with a Static Topology We randomly select a topology and run the same experiments as above. We present the results in Fig. 4.11. Given the small diameter of our topology, the static topology OSA still achieves satisfactory performance. For example, in the worst case of all-to-all traffic (i.e., $t = 7$), static OSA achieves more than 40% of the non-blocking network's bisection bandwidth. Since

Fig. 4.8 Make a non-blocking network from OSA



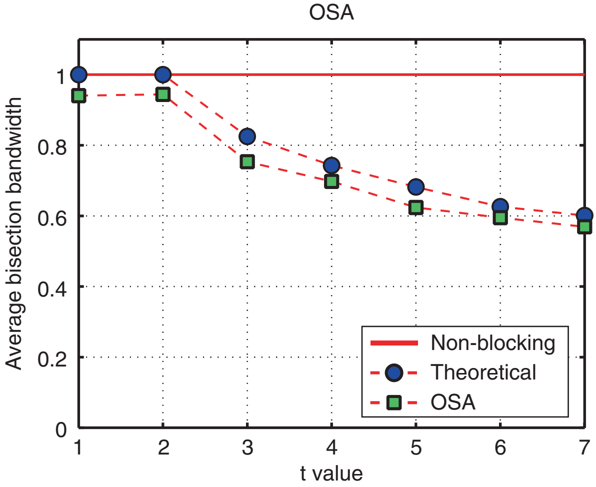


Fig. 4.9 Average bisection bandwidth of OSA

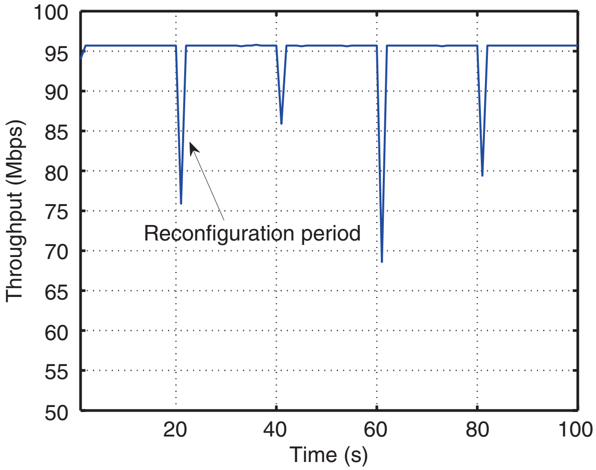


Fig. 4.10 Throughput during reconfigurations

all the paths are one or two hops long, even the randomly selected topology performs satisfactorily.

For different *t* values, we find that the performance of OSA on the static topology is lower than that on the dynamic topology by 10–40%. This is because the topology is not optimized for the current traffic pattern. We expect that on a larger network where OSA topology is sparse, this performance gap will become more pronounced, highlighting the need for a dynamic network for better performance.

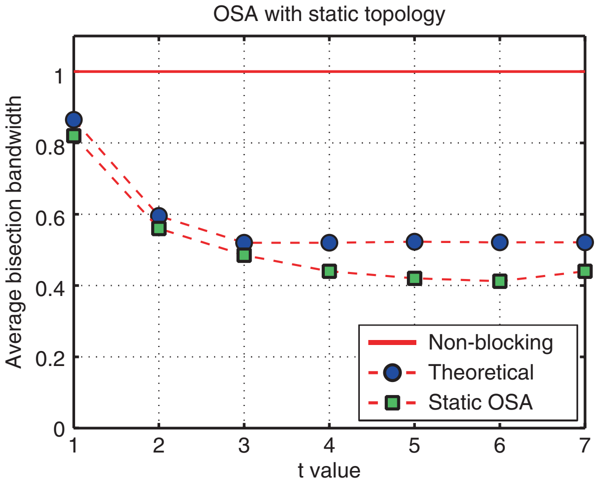


Fig. 4.11 Average bisection bandwidth of OSA with a static topology

4.6 Discussion and Conclusion

Mice Flow During Reconfiguration OSA ensures that all the ToRs are in a connected graph and uses the hop-by-hop stitching of existing circuits to provide overall network connectivity. However, during the network reconfiguration, a pair of ToRs may be temporarily disconnected for around 10 ms. While this can be largely tolerated by latency-insensitive applications such as MapReduce or Dryad, it would affect those operating with latency-sensitive mice flows like Dynamo [28]. We next discuss two possible options to handle this issue.

Our basic idea is to reserve a static, connected channel in OSA network. To do so, we can reserve a small number of wavelengths and MEMS/WSS ports that are never reconfigured and mice flows are sent over them. Such a channel can be simply a spanning tree or other connected topologies. Given the topology of the channel which is controlled by the MEMS, we can arrange it in a low-diameter manner so that the transmission of mice flows is optimized. However, this approach consumes expensive MEMS/WSS ports, which otherwise can be better utilized for other applications or at stable time.

An alternative approach to building the channel without using MEMS/WSS ports is directly connecting all the ToRs together to form a ring or a star network. For the ring, we can reserve two ports on each ToR and directly connect them iteratively. In case of OSA-2560 with 80 ToRs, the diameter is 40 hops. To reduce the path length, it is possible to reserve more ports on each ToR and connect them structurally using DHT techniques [29], e.g., the diameter is expected to be three to four hops with high probability for 80 ToRs if we reserve four ports on each

ToR. Another option is to employ one additional central electrical switch – each ToR uses one port to connect to the central switch. Note that, in Helios or c-Through, the electrical switches (usually forming a tree or even a multi-root tree) are used for overall connectivity among all the Pods/ToRs. In OSA, the all-to-all connectivity is maintained by optical components. A comprehensive evaluation and comparison of these solutions is part of our ongoing work.

OSA Applicability Versus Traffic Properties For the all-to-all traffic, the non-oversubscribed network is indeed more appreciated. However, such workloads are neither reflected in our dataset nor in the measurements elsewhere [2, 15, 27]. Our flexible OSA architecture would work best when the traffic pattern is skewed and stable on the order of seconds. It has been noted in [5] over the measurements of a 1500-server production DCN that "only a few ToRs are hot and most of their traffic goes to a few other ToRs." Another study [2], also on a 1500-server production DCN, shows that more than 90% of bytes are in elephant flows. Regarding the traffic stability, a similarly sized study [30] shows that 60% of ToR pairs see less than 20% change in traffic demands for between 1.6 and 2.2 s on average. Despite these, we expect that OSA may exhibit undesirable performance degradation if the traffic pattern is highly dynamic; in which case any topology adaptation mechanism may be unsuitable as the situation changes instantaneously. In practice, the infrastructure manager should choose the proper sensitivity of OSA according to the operational considerations.

Scalability The current OSA design focuses on the container-size DCNs. To scale, we may confront several challenges. The first one is the MEMS' port density. While the 1000-port MEMS is theoretically feasible, the largest MEMS as of today has 320 ports. One natural way to increase the port density is via interconnecting multiple small MEMS switches. However, this poses additional requirement for fast, coordinated circuit switching. Secondly, larger network size necessitates more control and management overhead. In our OSA-2560 with 80 ToRs, all the intelligences, e.g., the network optimization and routing, are handled by the OSA. How to handle such tasks in a larger DCN with thousands of ToRs is an open question especially when the network environment is dynamic. Further, circuit-switching delay [14] is another issue to notice when scaling. We are considering all these challenges in our continuous effort designing a scalable optical DCN.

Summary In this chapter, we have presented OSA, a novel optical switching architecture for DCNs. OSA is highly flexible because it can adapt its topology as well as link capacities to different traffic patterns. Our implementation and evaluation with the OSA prototype further demonstrate its feasibility. OSA, in its current form, has limitations. Small flows, especially the latency-sensitive ones, may experience nontrivial delay due to the reconfiguration latency (∼10 ms). Another challenge is how to scale OSA from container size to larger DCNs consisting of tens to hundreds of thousands of servers. This requires efforts in both architecture design and management and is left as part of future work.

References

1. M. Al-Fares, A. Loukissas, A. Vahdat, A scalable, commodity data center network architecture. in *SIGCOMM*, 2008
2. A. Greenberg, J.R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, S. Sengupta, VL2: a scalable and flexible data center network. in *ACM SIGCOMM*, 2009
3. R.N. Mysore, A. Pamboris, N. Farrington, N. Huang, P. Miri, S. Radhakrishnan, V. Subramanya, A. Vahdat, Portland: a scalable fault-tolerant layer 2 data center network fabric. in *ACM SIGCOMM*, 2009
4. C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, S. Lu, BCube: a high performance, server-centric network architecture for modular data centers. in *ACM SIGCOMM*, 2009
5. S. Kandula, J. Padhye, P. Bahl, Flyways to de-congest data center networks. in *ACM HotNets*, 2009
6. K. Barker et al., On the feasibility of optical circuit switching for high performance computing systems. in *SC*, 2005
7. J. Hamilton, Data center networks are in my way. http://mvdirona.com/jrh/TalksAndPapers/JamesHamilton_CleanSlateCTO2009.pdf
8. H. Liu, C.F. Lam, C. Johnson, Scaling optical interconnects in datacenter networks opportunities and challenges for wdm. in *IEEE Symposium on High Performance Interconnects*, 2010
9. C. Lam, H. Liu, B. Koley, X. Zhao, V. Kamalov, V. Gill, Fiber optic communication technologies: what's needed for datacenter network operations, 2010
10. J. Rath, Google eyes "optical express" for its network. http://www.datacenterknowledge.com/archives/2010/05/24/google-eyes-optical-express-for-its-network/
11. N. Farrington, G. Porter, S. Radhakrishnan, H.H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, A. Vahdat, Helios: a hybrid electrical/optical switch architecture for modular data centers. in *ACM SIGCOMM*, 2010
12. CIR, 40g ethernet c-closer than ever to an all-optical network. http://cir-inc.com/resources/40-100GigE.pdf
13. ADC, 40 & 100 gigabit ethernet: an imminent reality. http://www.adc.com/Attachment/1270718303886/108956AE,0.pdf
14. G. Wang, D.G. Andersen, M. Kaminsky, K. Papagiannaki, T.S.E. Ng, M. Kozuch, M. Ryan, c-Through: part-time optics in data centers. in *ACM SIGCOMM*, 2010
15. D. Halperin, S. Kandula, J. Padhye, P. Bahl, D. Wetherall, Augmenting data center networks with multi-gigabit wireless links. in *SIGCOMM*, 2011
16. T. Truex, A.A. Bent, N.W. Hagood, Beam steering optical switch fabric utilizing piezoelectric actuation technology. in *NFOEC*, 2003
17. J. Kim et al., 1100×1100 port mems-based optical crossconnect with 4-db maximum loss. IEEE Photon. Technol. Lett. **15**(11), 1537–1539 (2003)
18. Broadcom, Bcm56840 series enables mass deployment of 10gbe in the data center. http://www.broadcom.com/products/features/BCM56840.php
19. K. Chen, C. Guo, H. Wu, J. Yuan, Z. Feng, Y. Chen, S. Lu, W. Wu, Generic and automatic address configuration for data center networks. in *SIGCOMM*, 2010
20. E. Akcali, A. Ungor, Approximation algorithms for degree-constrained bipartite network flow. in *ISCIS*, 2003
21. M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, A. Vahdat, Hedera: Dynamic flow scheduling for data center networks. in *NSDI*, 2010
22. M. Müler-Hannemann, A. Schwartz, Implementing weighted b-matching algorithms: insights from a computational study. J. Exp. Algorithm. **5**, 8 (2000)
23. LEMON-Library, http://lemon.cs.elte.hu
24. K. Obraczka, P. Danzig, Finding low-diameter, low edge-cost, networks. USC, Technical Report, 1997
25. Edge-coloring, http://en.wikipedia.org/wiki/Edge_coloring

26. J. Misra, D. Gries, A constructive proof of vizing's theorem. Inf. Process. Lett. **41**(3), 131–133 (1992)
27. T. Benson, A. Akella, D. Maltz, Network traffic characteristics of data centers in the wild. in *IMC*, 2010
28. G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, Dynamo: Amazon's highly available key-value store. in *SOSP*, 2007
29. A. Rowstron, P. Druschel, Pastry: scalable, decentralized object location and routing for large-scale peer-to-peer systems. in *Middleware*, 2001
30. T. Benson, A. Anand, A. Akella, M. Zhang, The case for fine-grained traffic engineering in data-centers. in *USENIX INM/WREN*, 2010

Chapter 5
The Hi-Ring Architecture for Data Center Networks

Valerija Kamchevska, Yunhong Ding, Michael S. Berger, Lars Dittmann, Leif K. Oxenløwe, and Michael Galili

5.1 Introduction

The explosive growth of Internet-based cloud applications and the rapid growth of data centers in both number and size have attracted vast attention to short-range communications. Data center networks (DCNs) have rapidly evolved into a field of strong commercial and academic interest. The drive toward low-cost, energy-efficient, and low-latency communication has posed the question about the role of optics in DCNs. *Optical switching* is attractive because it enables high bandwidth connectivity and seamless capacity scaling by simultaneous switching of data streams that exploit different multiplexing domains to achieve higher data rates.

In this chapter, we present the Hi-Ring data center network and the concept of multidimensional optical switching. The Hi-Ring exploits photonic technologies offering benefits such as scalability, bit rate-independent switching, and energy-efficient operation. The chapter is subdivided into three sections. The first section deals with the general overview of the proposed architecture and the structure of the multidimensional switching nodes. The second section provides an insight into the requirements behind the proposed optical subwavelength switching technology, the synchronization process, and the control aspects associated with it. The last section is devoted to the on-chip integration of optical devices that hold the promise for realizing future Network-on-Chip (NoC) implementations.

V. Kamchevska (✉) • Y. Ding • M.S. Berger • L. Dittmann • L.K. Oxenløwe • M. Galili
DTU Fotonik, Technical University of Denmark,
Ørsteds Plads 343, 340, Kongens Lyngby, Denmark
e-mail: vaka@fotonik.dtu.dk

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
DOI 10.1007/978-3-319-61052-8_5

5.2 The Hi-Ring Architecture and Multidimensional Switching

Multiplexing technologies such as *time division multiplexing (TDM)*, *wavelength division multiplexing (WDM)*, and *space division multiplexing (SDM)* have been researched considerably over the years. They all provide a means to increased capacity by densely aggregating traffic in each domain, thus saving resources and improving the bandwidth utilization. One of the main advantages of optical technologies is that starting from simple systems exploiting single wavelength communication capacity scaling can be achieved in several different ways by utilizing the various multiplexing domains. This feature is crucial for rapidly expanding DCNs that need to support scaling to immense capacity. However, besides considering scalability as one of the most important issues, DCNs require technologies that are cheap and energy efficient, and hence resource utilization must be maximized. In order to be able to provide support for scaling, while at the same time being able to address individual data flows with fine granularity within the network, an optical switching infrastructure is needed.

This issue has been approached in several different ways in existing proposals. Hybrid approaches that use electrical packet switching combined with optical circuit switching such as c-Through [1] and Helios [2] have been proposed to provide simultaneous support for long-lived and bursty traffic using two different switching technologies. All-optical approaches as in [3, 4] have proposed to replace electrical packet switching with optical packet switching (OPS) for fast routing of small data packets. However, going from electrical to optical packet switching is not straightforward due to the unavailability of mature optical buffering technologies. Several other proposals such as [5, 6] have focused on replacing the functionality provided by electrical packet switching with optical switching technologies combined with enhanced control platforms like software-defined networking (SDN). This allows that buffering in the network is limited to the end points and contention-free scheduled circuit-like transmissions replace OPS.

Another important aspect of designing the DCN is the actual technology used to provide the desired functionalities. Different optical switching technologies are commercially available and have been proposed. Standard optical circuit switching is most commonly done using microelectromechanical switches (MEMS) [7, 8] or piezoelectric beam-steering switches [9]. The main advantages of these switches are the low loss, low cross talk, and high port count. However due to the physical effects behind the switching, the switching time is in the order of microsecond to milliseconds. Fast optical switches such as those relying on modulating gain or refractive index in photonic integrated devices [10–12] provide switching in the nanosecond range. However, they usually suffer from higher insertion loss and cross talk as the port count is scaled up. Wavelength selective switches (WSS) [13] allow for addressing individual wavelength channels in WDM networks. The switching time is in general on the order of microseconds, and high port count WSSs with up to 93 ports have already been demonstrated [14].

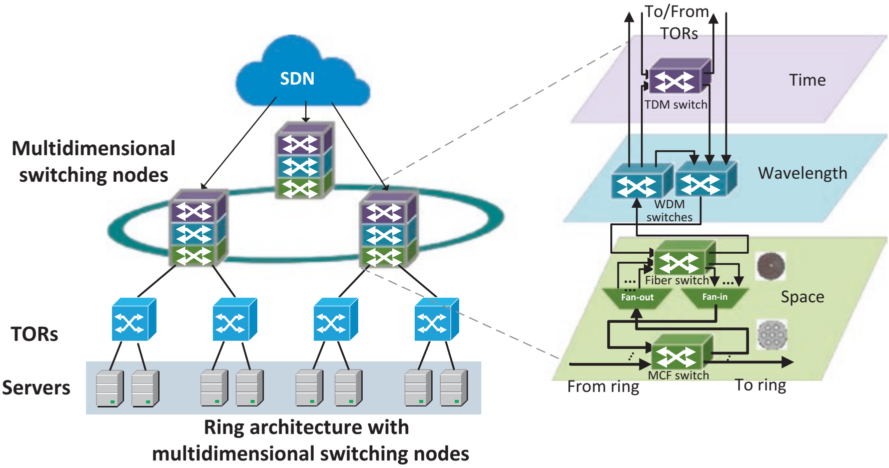


Fig. 5.1 The Hi-Ring architecture with multidimensional switching nodes (Reprinted with permission from Ref. [15]. ©2016 IEEE)

Considering the advantages and disadvantages of each of the aforementioned technologies, we have proposed the *Hi-Ring architecture* for DCNs [15], as shown in Fig. 5.1. This architecture is composed of *multidimensional switching* nodes that allow for all-optical switching in three different dimensions (space, wavelength, and time). The network nodes (NNs) have a hierarchical layout of the switches, meaning that the switch with the coarsest granularity (i.e., allowing for switching of highly aggregated data streams) is positioned at the lowest level of the node, while switches providing fine granularity switching of smaller data streams are positioned higher in the hierarchy. The reasoning behind this decision is to allow that connections are either bypassed at a lower level whenever they are destined to other nodes or processed at a higher level in case switching with finer granularity is required. By bypassing at the lowest level possible, extra processing is removed reducing the use of switch resources. As each node will have to process only a small amount of the incoming traffic and forward most of it, a ring topology for interconnection has initially been chosen.

Switching in the space dimension is performed using two types of switches, one operating at a multicore fiber (MCF) granularity and another one operating at a single fiber core granularity. At both levels, traffic can be added and dropped or can bypass the node, enabling communication with negligible delay. In addition, the fiber switch allows for reshuffling of fiber cores. The WDM switch is a wavelength selective switch that can be reconfigured dynamically. The TDM switch is a fast optical switch that is used for subwavelength switching. The WDM and TDM levels dictate the two different granularities of connections that can be established within the DCN, i.e., long-lived connections that require higher bandwidth are assigned a full wavelength, while bursty connections that require less total bandwidth are assigned a wavelength that is shared among several connections. By sharing a single wavelength, TDM switching allows for increased bandwidth utilization, but unlike

OPS it pushes the buffering toward the network edge, thus eliminating the need for optical buffering within the network. Circuit-like transmission scheduling and reservation of the slotted medium are performed by a centralized SDN controller. The Top-of-Rack (TOR) switches generate traffic with different granularity (wavelength or time slot) toward the NNs, as shown in Fig. 5.1, or in a more long-term view, one could assume that the NNs are access points for the servers themselves, yielding a truly all-optical architecture.

The Hi-Ring architecture has several advantages. Scaling is supported by using several multiplexing dimensions, while still allowing for connections with relatively small granularity, hence maximizing the resource utilization. The hierarchical design allows for traffic aggregation which results in a relatively simplified physical topology. This means that a reasonable number of nodes and physical links can be retained even for a high number of servers. Inventory comparison results in [16] reveal that for the same number of TOR switches, the Hi-Ring architecture would result in 45.8% fewer links and 50% fewer nodes compared to a fat-tree topology. The reduced amount of resources directly implies simplified link management and maintenance. Furthermore, simulation results in [16] indicate 40–99% improvement in connection request blocking and 3–17% improvement in resource utilization compared to a fat-tree topology under different available capacity and varied inter/intracluster traffic ratio.

Considering that node bypass can be performed optically at each level, heavy processing of all incoming traffic at higher levels is circumvented. Moreover, assuming that an SDN control plane exists and facilitates the resource reservation process, connections will experience only the physical propagation delay through the switch, yielding low latency operation. Unlike Ethernet switching, the nodes are fully bit rate independent, i.e., can easily upgrade to a higher bit rate without any additional processing. The use of TDM allows that resources are utilized efficiently by sharing underutilized channels among several connections and hence leading to operational cost reductions. Additionally, by synchronizing the network elements, which can easily be done in a controlled environment such as DCNs, the need for buffering at the network nodes is also eliminated.

There are several approaches regarding the actual physical structure of the node and the level of integration. Modular node design with distinct components has the advantage of simplified upgrade/component repair and seamless and elastic migration toward new technologies. However, although individual switches at each level may provide greater flexibility, integration of all components on a single platform also has several crucial benefits. Not only does this approach hold the promise for developing future NoC with small footprint, but integration is very important when it comes to addressing relevant commercial issues such as fabrication and packaging, cost, power consumption, cooling approach, etc. Furthermore, as in most DCN architectures based on optical switching technologies, optical amplification is a necessity, so integration can facilitate combining the switching devices with proper on-chip optical amplifiers.

In order to investigate the performance of the architecture, we have experimentally implemented a small subset of the network consisting of three nodes. The

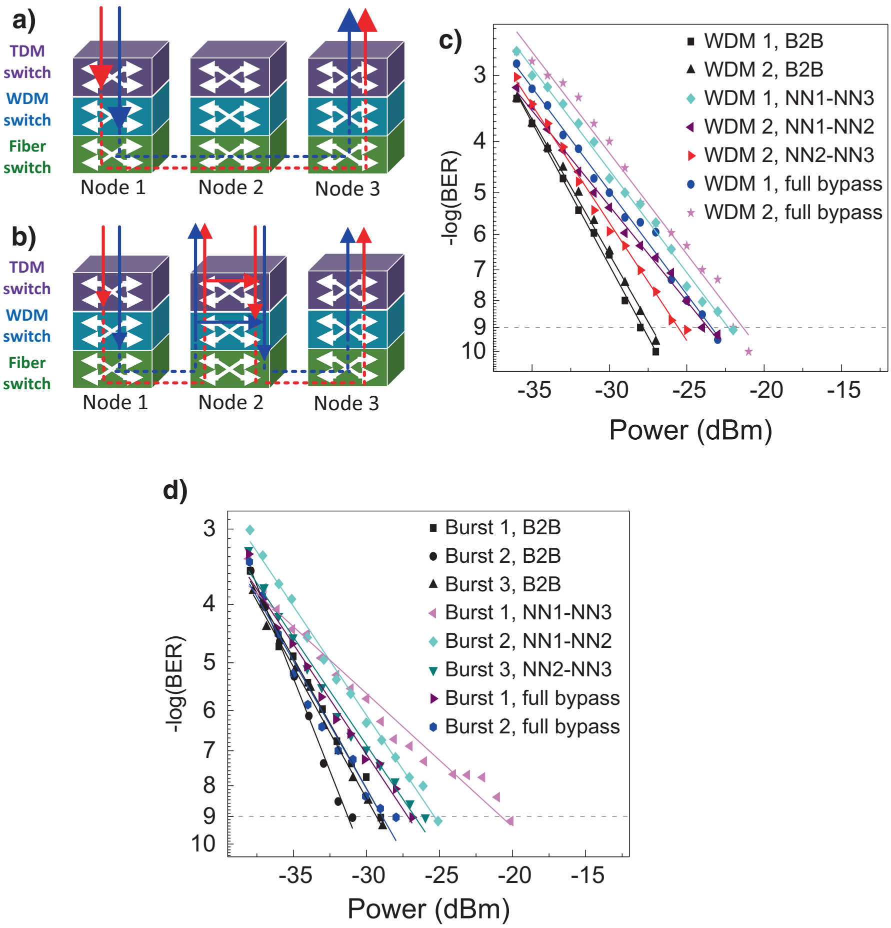


Fig. 5.2 Experimental scenarios: (**a**) full bypass and (**b**) add/drop/bypass (Reprinted with permission from Ref. [15]. ©2016 IEEE). BER performance of the (**c**) WDM connections and (**d**) TDM connections for both scenarios (Replotted using data from Ref. [15]. ©2016 IEEE)

nodes are interconnected with a single seven-core MCF on each link, thus omitting the MCF switch at the nodes. Each node is composed of an optical fiber switch, a WSS used as a WDM switch, and a $LiNbO_3$-based electro-optic switch as a TDM switch. As shown in Fig. 5.2a and b, wavelength connections (blue) and time-slotted connections (red) are established using 40 Gbit/s on-off keying (OOK)-modulated data on 100 GHz spaced channels. Two scenarios are investigated. In the "full bypass" scenario, traffic is generated and multiplexed at the first node, bypassed at the lowest level in the second node, and fully received at the third node. In the "add/drop/bypass" scenario, traffic is generated at the first node, partially dropped, and received at the second node, where terminated connections are replaced with new connections, and finally all remaining traffic is received at the third node.

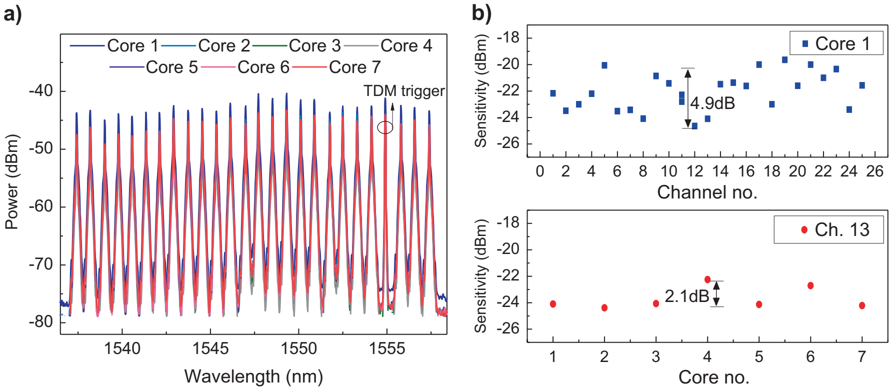


Fig. 5.3 (**a**) Spectra of all received channels and (**b**) receiver sensitivity of all channels in one core (*top*) and a single channel in all cores (*bottom*) (Reprinted with permission from Ref. [15]. ©2016 IEEE)

Figure 5.2c and d illustrates the measured BER results for the WDM and TDM connections in both scenarios. It can be seen that error-free performance (BER < 10^{-9}) is achieved for all connections. The results from both scenarios confirm that it is preferable to perform bypass at lower levels in intermediate nodes. Bypass at higher levels leads to slightly higher penalty due to signal degradation as a result of additional loss and limited suppression ratio of the switches. However, exploiting effective scheduling that considers the power budget and signal degradation can allow for minimizing or completely avoiding this penalty.

Additionally, we have fully loaded the system with 25 wavelength channels yielding a 1 Tbit/s/core capacity and total throughput of 7 Tbit per MCF. Figure 5.3a shows the spectra of all the received channels in all cores. In this case the system performance is evaluated only in the “full bypass” scenario. The measured receiver sensitivities (at BER = 10^{-9}) of all channels in a single core and of a single channel in all cores are shown in Fig. 5.3b. The consistent behavior and error-free performance confirm not only the feasibility of the system but also the ability to provide support for high capacity data networking applications such as data centers.

5.3 Optical Subwavelength Switching and Synchronization

Optical subwavelength switching is an important element of the proposed architecture, mainly because it allows for improved bandwidth utilization by accommodating bursty traffic using time-sharing of a single wavelength among different connections. Unlike OPS and optical burst switching (OBS), *optical TDM switching* is based on circuit transmissions over a time-slotted medium, where the medium access is arbitrated by an SDN controller. Transmissions are permitted during time slots arranged in a periodic frame structure, which allows that sufficient control

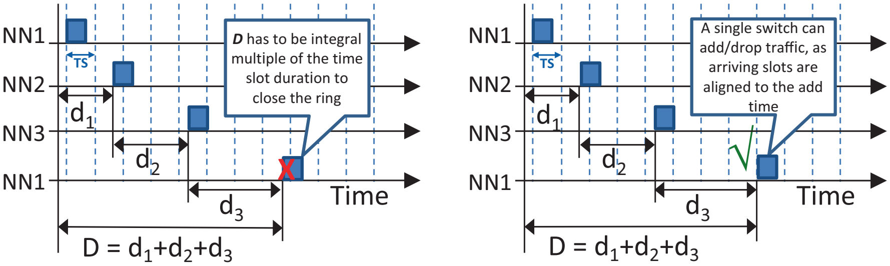

NN1
NN2
NN3
NN1
TS
d_1
d_2
d_3
$D = d_1+d_2+d_3$
Time
D has to be integral multiple of the time slot duration to close the ring
X
NN1
NN2
NN3
NN1
TS
d_1
d_2
d_3
$D = d_1+d_2+d_3$
Time
A single switch can add/drop traffic, as arriving slots are aligned to the add time
√

Fig. 5.4 A ring network with propagation delay, *D*, along the ring that is not (*left*) and that is (*right*) an integer multiple of the time slot duration. Depicted is the time at which a burst of data reaches each node in the ring for a full loop (Reprinted with permission from Ref. [22]. ©2016 OSA)

information is delivered to the switches on a frame basis. The slot allocation is made based on an existing schedule of slot assignments that can be dynamically modified. This allows for a trade-off between inflexible static allocation and complex dynamic assignment. Assuming that all nodes are time aware and considering that no optical buffering is used within the network, a centralized entity such as an SDN controller with global overview of the network can push control plane decisions to SDN agents deployed in the network nodes. As there is no need to transmit control information using packet headers, extra header processing is fully eliminated and better bandwidth utilization can be achieved. Furthermore, the control is completely decoupled from the data plane, allowing for transparent networking where data is processed only at the end sides.

In order to avoid overlaps in transmissions and provide accurate switching, it is necessary to have a mechanism for precise *synchronization* among network elements. In general, any optical subwavelength switching technology that relies on using *NxN* fast optical switches, such as OPS, OBS, or optical circuit-based time slot switching, requires synchronization among the data arriving on different inputs of the switch. Additionally, in a ring network that employs *NxN* fast optical switches, such as the Hi-Ring network, there is a requirement that the ring propagation delay, *D*, is an integer multiple of the slot duration. As shown in Fig. 5.4, if the drop time at each node is dictated by the arrival of the burst from NN1, then the drop time at NN1 has to coincide with the initial add time. Any deviation from this condition will cause improper switching at NN1, and the ring cannot be closed.

There are different ways to provide synchronization in a network composed of fast optical switches that perform optical subwavelength switching. One way that has commonly been associated with OPS is to use synchronizers for each input that can dynamically compensate for delay offsets [17]. However, the cascaded switching in the synchronizers leads to high insertion loss and impaired performance due to cross talk. Other examples include global synchronization mechanism that uses either the global positioning system (GPS) [18] or a master clock reference [19] that is distributed in a network with pre-engineered fiber link lengths. Although theoretically this enables accurate operation, in practice it limits severely the network flexibility.

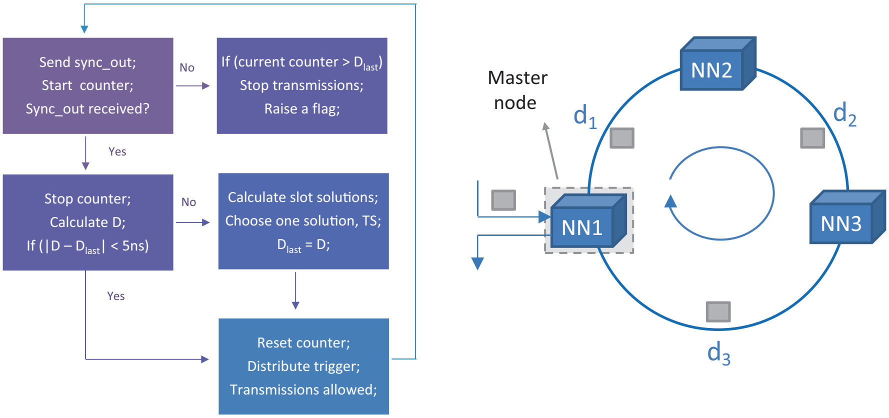


Fig. 5.5 Diagram of the proposed synchronization algorithm (*left*) and an overview of the implemented synchronization plane (*right*) (Reprinted with permission from Ref. [22]. ©2016 OSA)

In [20], an algorithm is presented for synchronization of time slots in multi-ring networks by separating add and drop time at each node. This approach works well when using tunable wavelength converters in combination with arrayed waveguide grating (AWG). However, it cannot be applied if *NxN* optical switches are used to perform simultaneous add/drop of bursts.

Considering the Hi-Ring architecture, we have proposed a synchronization algorithm [21, 22], which works by estimating the propagation delay along the ring and decides on a slot duration such that the propagation delay is an integer multiple of the slot size. The detailed algorithm behavior is illustrated in Fig. 5.5a, and an overview of the implemented synchronization plane is given in Fig. 5.5b. The main idea behind the algorithm is that one node in the ring, denoted as master node, is responsible for running the algorithm and providing synchronization signals to the remaining nodes. The master node sends out a signal, *sync_out* periodically, and estimates the propagation delay, *D*, along the ring based on the time after which the signal is received back. As DCNs are a closed and well controlled environment, this measurement can be done reliably and accurately. Furthermore, by using a clock with high frequency, the elapsed time can be measured with high accuracy. If the signal is not received after a certain time, this can be used as an indication of a link/node failure which can be communicated to a network controller.

Once the propagation delay is estimated, the allowed slot solutions are found as the slot sizes within a predefined range such that *D* is an integer multiple of the slot size. In case there is no slot size satisfying this condition, approximate solutions can be accepted as valid, by allowing some flexibility in the duration of the guard interval between time slots. When the search for slot sizes is concluded, a solution can be chosen for data plane operation. Furthermore, if the solutions are disclosed to an SDN controller, dynamic change within the solution space is also possible, allowing that a better slot size is chosen to match the current traffic profile. At last, once a slot size is chosen, the whole operation is reset and restarted. The synchronization within

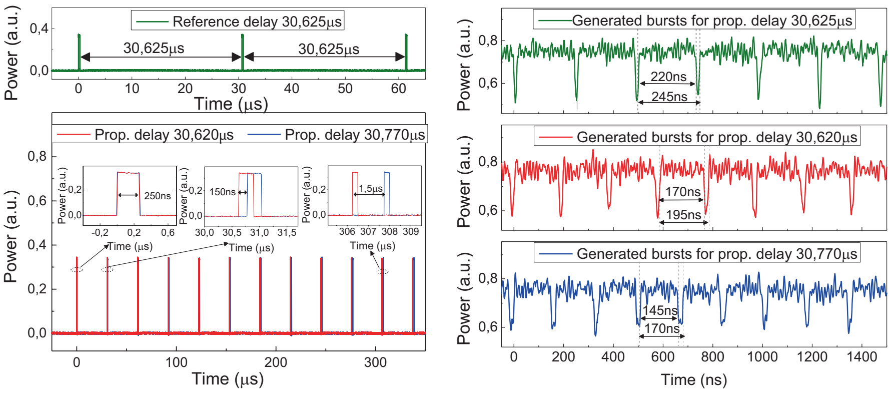


Fig. 5.6 Time domain traces of the received *sync_out* signal after propagation (*left*) and the generated bursts for the different propagation delays (*right*) (Reprinted with permission from Ref. [22]. ©2016 OSA)

the network is provided by distributing a trigger. Each node runs a counter based on the received trigger, hence automatically compensating for the propagation delay.

In order to verify the algorithm behavior experimentally, we use a ring of three nodes interconnected with a set of two cores of a single 2 km-long seven-core multicore fiber. One core is used for transmitting the *sync_out* signal and a second core for transmitting the data and trigger. By varying the propagation delay along the ring, it can be verified that the algorithm can provide automatic synchronization for three different ring lengths. The time domain traces of the received signal after propagation for the three cases are shown in Fig. 5.6a. It can be seen that the periodicity of the received signal corresponds to the propagation delay measured in the three cases. For this implementation, a slot size is randomly selected from the solution space with a fixed 25 ns gap and used for all transmissions within the data plane. To confirm that a correct slot size is chosen, we record the time domain traces of a transmitter that is using the chosen slot size for transmitting data as shown in Fig. 5.6b. For all cases the algorithm choses a correct solution, confirming proper implementation and operation of the scheme.

After verifying that the algorithm is correctly deciding on a slot size, the data plane performance is assessed. Figure 5.7 shows the time domain traces of the generated and switched bursts illustrating the envisioned connectivity among the three nodes. It is important to note that the propagation delay is continuously measured during the data plane operation. The data bursts are generated using an FPGA and carry 10 Gbit/s OOK-modulated data. An OpenDaylight SDN controller is used to control the switches at each node. Similar BER performance is achieved for all bursts dropped at each node and received using a standard optically preamplified receiver as shown in Fig. 5.7. These results verify that correct dynamic network operation can easily be achieved in practical conditions.

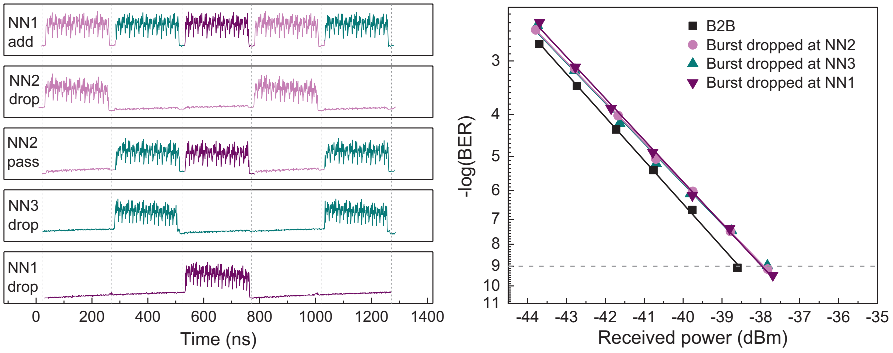


Fig. 5.7 Time domain traces of the generated and switched bursts at the different nodes (*left*) and BER performance of the bursts dropped at each node (*right*) (Reprinted with permission from Ref. [22]. ©2016 OSA)

5.4 On-Chip Integration Using Silicon Photonics

On-chip integration is an important aspect of developing data center technologies. Integration holds the promise to provide low-cost, low footprint, and energy-efficient subsystems fully integrated on a single platform like silicon on insulator (SOI). Integration is important from two standpoints. One aspect is the integration of the actual devices and the driving electronics required to configure them. Another aspect is the ability to integrate as many devices as possible with different functionalities on a single chip, ultimately leading to NoC implementations. Previous demonstrations of monolithically integrated switches [8, 10] confirm that *silicon photonics* is a suitable platform for fabricating devices with the potential for commercial development. Regarding the latter and as previously discussed, the integration of a subset of elements from a single node in the Hi-Ring network is a very important step toward building compact and fully functional NoC. Although possible, full node integration is also challenging, and thus we have initially focused on partial integration of only few components, allowing for a compromise between the modular node design and the integrated approach. In [23–25] we presented a novel PIC composed of MCF coupling devices and a fiber switch; we performed basic characterization and demonstrated error-free system performance. This device allows for integrating node components that operate within the space dimension of the hierarchical structure. Space division multiplexing has attracted increasing interest due to its ability to provide significant increase of the available capacity over a single fiber. However, deployment of SDM technologies such as MCFs depends on the availability of devices such as fan-in/fan-out devices, MCF amplifiers, add/drop modules, switches, etc. Considering that each node of the Hi-Ring architecture deploys SDM components such as the MCF switch and coupling devices to access the fiber switch, integration of any of these entities is extremely valuable. The fabricated PIC as shown in Fig. 5.8 is composed of two grating couplers for spatial multiplexing/demultiplexing of data carried in the different cores of

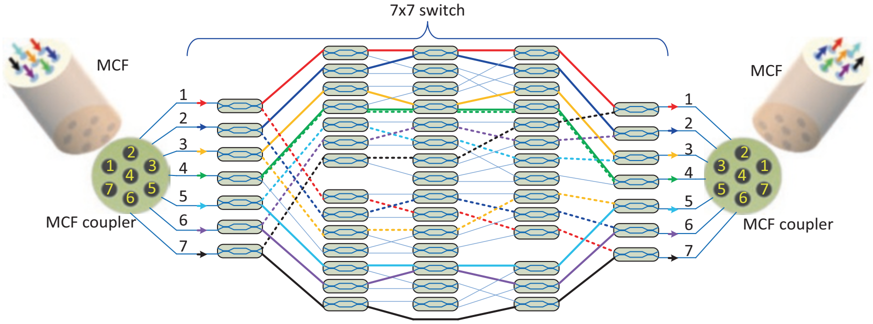


Fig. 5.8 Switch matrix of the fabricated device with the established paths for bar (*solid line*) and cross (*dashed line*) configuration (Reprinted with permission from Ref. [24]. ©2016 OSA)

a seven-core MCF. The grating coupler array fan-in/fan-out [26] consists of seven grating couplers aligned to the arrangement of the cores of the MCF. In addition, Al mirrors are introduced below the grating couplers by bonding in order to achieve low coupling loss [27]. After demultiplexing, each core is connected to the input of the 7 × 7 fiber switch that is able to switch the different cores of the MCF. The switch is built as a five-stage switching matrix composed of 57 Mach-Zehnder interferometer (MZI) elements, each incorporating a heater allowing for thermally controlled switching. After switching, the seven spatial channels are multiplexed and launched to the output MCF. In this scheme, any core at the input can be switched to any core of the output MCF. For example, a bar (solid line) and cross (dash line) switching configuration can be set by configuring the relevant MZIs as shown in Fig. 5.8. In addition, in [28] we have shown that this switch can be used not only for unicast switching but also to provide support for multicast and optical grooming which can be useful for scheduled or backup tasks in DCNs.

The 12 mm × 5 mm silicon PIC as shown in Fig. 5.9a is fabricated on an SOI platform with top silicon thickness of 250 nm. The device performance is characterized for two switching configurations, namely, bar and cross as shown in Fig. 5.8. Figure 5.9b and c illustrates the measured transmission and cross talk for both configurations. It can be seen that the insertion loss and cross talk for C-band are less than 8 dB and −30 dB, respectively. The power consumption per heater is approximately 13 mW, and the average measured switching time is around 30 μs.

In order to investigate the performance of the PIC, we generated 1 Tbit/s/core using 25 channels carrying 40 Gb/s OOK-modulated data on a 100 GHz grid in the C-band. The data is launched into a 2 km MCF, and after propagation the MCF is coupled directly to the silicon PIC. The switch is configured in bar and cross configuration, and the measured receiver sensitivities of all channels in one core and a single channel in all cores are shown in Fig. 5.10. It can be seen that all channels experience error-free performance thus confirming that integration of this kind is a feasible solution.

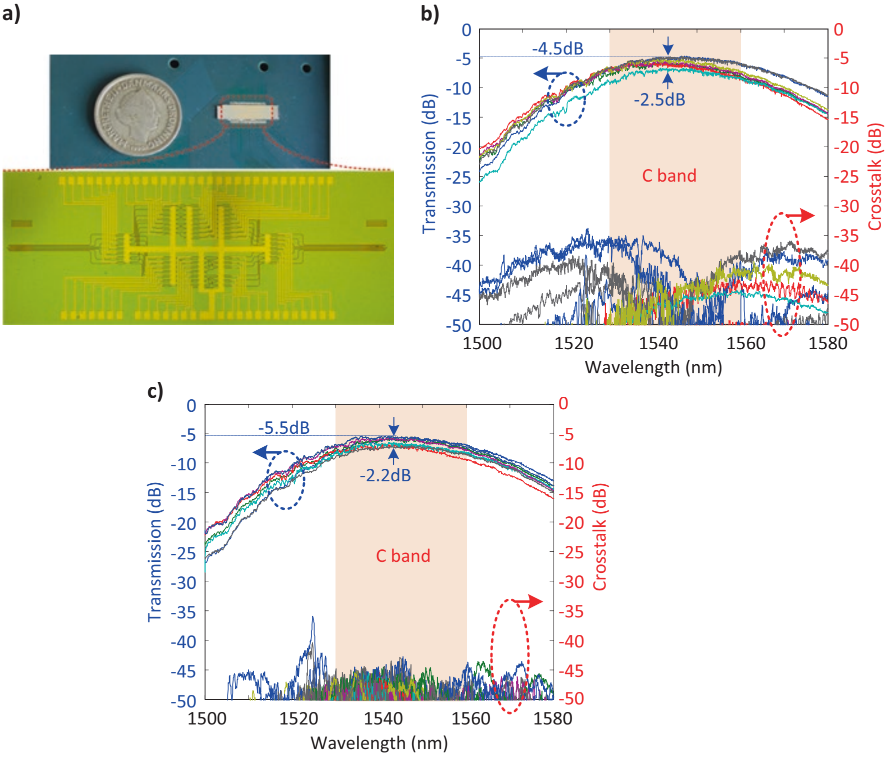


Fig. 5.9 (**a**) Chip layout. Device characterization for configuration (**b**) bar and (**c**) cross (Reprinted with permission from Ref. [24]. ©2016 OSA)

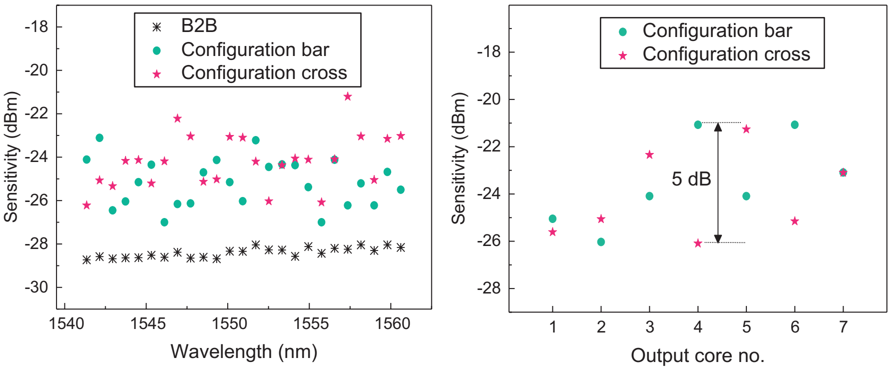


Fig. 5.10 Receiver sensitivity of all channels in a single core (*left*) and one channel in all cores (*right*) (Replotted using data from Ref. [23]. ©2016 OSA)

5.5 Perspectives and Research Directions

In this chapter the Hi-Ring data center architecture has been presented. The architecture is based on the concept of multidimensional switching, where nodes composed of all-optical switches provide switching with different granularity. The fact that photonic technologies are used to perform these functions offers a lot of benefits such as support for scaling, bit rate-independent switching, and energy-efficient operation. Furthermore, by exploiting photonic integration and utilizing platforms such as silicon photonics, the proposed architecture could ultimately be realized as a low-footprint and low-cost NoC.

Although promising, a lot of aspects of this architecture and similar optical architectures proposed still face a lot of challenges. Optical devices have to undergo an additional development phase where properties like insertion loss and cross talk are minimized. Existing implementations would most likely require that optical amplification is used in order to compensate for the loss experienced when going through several optical switches.

Replacing electrical packet switching with an optical subwavelength technology is not a straightforward solution, as optics cannot provide the same functionalities as electronics. However, combining optical TDM switching with an SDN controller that will be responsible for the medium access is a promising way to enable subwavelength connectivity using optical technologies. Additionally, precise operation cannot be achieved without deploying a synchronization mechanism that will be responsible for synchronizing all network elements. As discussed and demonstrated in the previous sections, realization of both of these aspects is not only feasible but also a promising way to introduce optics in the data center as an alternative to electrical packet switching.

Finally, in order to take advantage of the benefits that optical technologies have to offer, it is crucial to consider photonic integration. Photonic integration is a promising solution to incorporate both the actual switches and any additional devices required such as optical amplifiers, space division (de)multiplexers, etc. and reduce coupling losses between individual components. Among different platforms, silicon photonics seems most promising, mainly because of the mature CMOS infrastructure and the ability to integrate photonic components with driving electronics. This yields low production cost and paves the way toward commercialization of optical technologies for data center application and deployment.

References

1. G. Wang *et al.*, C-through: part-time optics in data centers. in *Proceedings of ACM SIGCOMM 2010*, 2010, pp. 327–338
2. N. Farrington et al., Helios: a hybrid electrical/optical switch architecture for modular data centers. in *Proceedings of ACM SIGCOMM 2010*, 2010, pp. 339–350
3. K. Kitayama et al., Torus-topology data center network based on optical packet/agile circuit switching with intelligent flow management. J. Lightw. Technol. **33**(5), 1063–1071 (2015)

4. G.M. Saridis et al., Lightness: a function-virtualizable software defined data center network with all-optical circuit/packet switching. J. Lightw. Technol. **34**(7), 1618–1627 (2016)
5. V. Kamchevska et al., Experimental demonstration of multidimensional switching nodes for all-optical data center networks. in *Proceedings of ECOC 2015*, paper Tu.1.2.2, 2016
6. L. Schares et al., A throughput-optimized optical network for data-intensive computing. IEEE Micro **34**(5), 52–63 (2014)
7. Calient, www.calient.com
8. T.J. Seok et al., 64x64 low-loss and broadband digital silicon photonic MEMS switches. in *Proceedings of ECOC 2015*, paper Tu.1.2.1, 2016
9. Polatis, www.polatis.com
10. B.G. Lee et al., Monolithic silicon integration of scaled photonic switch fabrics, CMOS logic, and device driver circuits. J. Lightw. Technol. **32**(4), 743–751 (2014)
11. L. Qiao, W. Tang, T. Chu, 16×16 non-blocking silicon electro-optic switch based on mach-zehnder interferometers. in *Proceedings of OFC 2016*, paper Th1C.2, 2016
12. X. Chen et al., Monolithic InP-based fast optical switch module for optical networks of the future. in *Proceedings of PS* 2015, 2015
13. Finisar, www.finisar.com
14. M. Iwama et al., Low loss 1x93 wavelength selective switch using PLC-based spot size converter. in *Proceedings of ECOC 2015*, paper Mo.4.2.2, 2015
15. V. Kamchevska et al., Experimental demonstration of multidimensional switching nodes for all-optical data center networks. J. Lightw. Technol. **34**(8), 1837–1843 (2016)
16. A.M. Fagertun et al., Ring-based all-optical datacenter networks. in *Proceedings of ECOC 2015*, paper P.6.9, 2015
17. A. Stavdas, A. Salis, A. Dupas, D. Chiaroni, All-optical packet synchronizer for slotted core/metropolitan networks. J. Opt. Commun. Netw. **7**(1), 88–93 (2008)
18. M. Baldi et al., Scalable fractional lambda switching: a testbed. J. Opt. Commun. Netw. **3**(5), 447–457 (2011)
19. B.R. Rofoee et al., Demonstration of low latency intra/inter data-centre heterogeneous optical sub-wavelength network using extended GMPLS-PCE control plane. Opt. Express **21**(5), 5463–5474 (2013)
20. K. Hattori, M. Nakagawa, M. Katayama, H. Ogawa, Method for synchronizing timeslot of WDM/TDM multi-ring network independent of fiber delay. in *Proceedings of OECC/ACOFT*, Melbourne, Australia, 2014, pp. 227–229
21. V. Kamchevska et al., Synchronization algorithm for SDN-controlled all-optical TDM switching in a random length ring network. in *Proceedings of OFC*, paper Th3I.2, 2016
22. V. Kamchevska et al., Synchronization in a random length ring network for SDN-controlled optical TDM switching. J. Opt. Commun. Netw. **9**(1), A26–A34 (2017)
23. Y. Ding et al., Experimental demonstration of 7 tb/s switching using novel silicon photonic integrated circuit. in *Proceedings of CLEO 2016*, paper Stu1G.3, 2016
24. Y. Ding et al., Silicon photonics for multicore fiber communication. in *Proceedings of ACP 2016*, paper AF1G.1, 2016
25. Y. Ding et al., Reconfigurable SDM switching using novel silicon photonic integrated circuit. Sci. Rep. **6**, 39058 (2016)
26. Y. Ding et al., On-Chip grating coupler Array on the SOI platform for fan-in/fan-out of MCFs with low insertion loss and crosstalk. Opt. Express **23**(3), 3292–3298 (2015)
27. Y. Ding, C. Peucheret, H. Ou, K. Yvind, Fully etched apodized grating coupler on the SOI platform with −0.58 dB coupling efficiency. Opt. Express **39**(18), 5348–5350 (2014)
28. V. Kamchevska et al., On-chip SDM switching for unicast, multicast and traffic grooming in data center networks. Photon. Technol. Letters **29**(2), 231–234 (2016)

Chapter 6
Low-Latency Interconnect Optical Network Switch (LIONS)

Roberto Proietti, Yawei Yin, Zheng Cao, C.J. Nitta, V. Akella, and S.J. Ben Yoo

6.1 Introduction

Optical interconnects can bring transformative changes to cloud computing system architectures. Compared to electrical interconnects, optical interconnects can provide (1) higher transmission bandwidth with lower energy and independently of distance, (2) low interference and crosstalk, (3) inherent parallelism, and (4) low parasitic. In addition, optical domain offers wavelength (frequency)-routing capability not available in electronics. Thus, optically interconnected computing systems could achieve (1) higher scalability, (2) high-density parallel links and buses overcoming input/output pin density limits, and (3) low latency avoiding the need for including repeaters or switches with store-and-forward architectures. Further, optical devices with wavelength routing capability can achieve all-to-all interconnection between computing nodes without contention.

An arrayed waveguide grating router (AWGR) [1, 2] is an example of devices with such wavelength routing capability. As Fig. 6.1 illustrates, the well-known wavelength routing property of an AWGR allows any input port to communicate with any output port simultaneously using different wavelengths without contention. Thus, an $N \times N$ AWGR intrinsically provides all-to-all communication among N compute nodes in a flat topology using N wavelengths. This realization will be called passive AWGR switch or passive low-latency interconnect optical network switch (LIONS) since no optical reconfiguration is necessary. This is true under the assumption that each node connected to the AWGR has N TRXs and an embedded switch forwarding the packets to the proper TRX based on the destination.

If the number of RXs (k_r) per node is $< N$, then contending conditions exist. Ref. [3] demonstrated buffered architecture, while [4] presented bufferless all-optical

R. Proietti (✉) • Y. Yin • Z. Cao • C.J. Nitta • V. Akella • S.J. Ben Yoo
Department of Electrical and Computer Engineering, University of California,
One Shields Ave, Davis, CA 95616, USA
e-mail: rproietti@ucdavis.edu

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
DOI 10.1007/978-3-319-61052-8_6

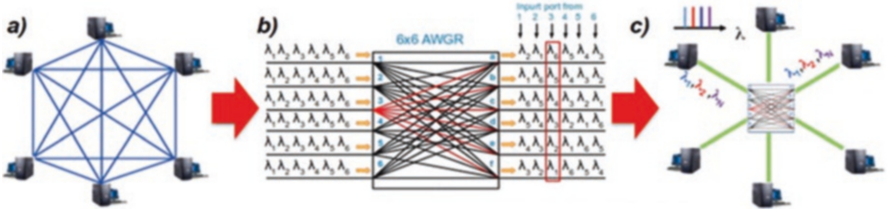


Fig. 6.1 (**a**) All-to-all interconnection with *N*(N-1)* point-to-point links *(N = 6)*. (**b**) Wavelength routing in a six-port AWGR. (**c**) All-to-all interconnection with only *N* links and *N* wavelengths

distributed control plane with performance far greater than typical electronic switches. In this case, it is required to tune the signal wavelength corresponding to the desired destination node according to the AWGR wavelength routing table. This case will be called an active AWGR switch or active LIONS. Ref. [3] investigated cases where the number of TXs (k_t) = 1 (single tunable transmitter) and $k_r < N$ for a large-scale rack-to-rack interconnect network. The active AWGR should deal with contention resolution, even though the contention probability is minimized by the use of k_r RXs per output port and by the unique wavelength routing and multiplexing properties in AWGR. Ref. [3, 4] show the performance of different LIONS architectures interconnecting a number of nodes ≥ 64 with $k_r = 4$. Even with a limited amount of RXs, the switch can sustain a throughput above 80% in case of uniform random traffic.

This chapter summarizes all the active and passive LIONS experimental demonstrations and testbeds carried out at UC Davis Next Generation Networking Systems (NGNS) laboratories. The first part focuses on active LIONS demonstrations with loopback buffer (LB-LIONS) and all-optical TOKEN and NACK techniques (TOKEN-LIONS and TONAK-LIONS). The second part of this chapter summarizes instead the experimental work related to passive LIONS architectures for hierarchical all-to-all interconnection.

6.2 Active LIONS Demonstrations

Figure 6.2a shows the loopback buffer low-latency interconnect optical network switch (LB-LIONS) consisting of an AWGR with $N + R$ ports, N tunable wavelength converters (TWCs) , an FPGA-based electrical control plane and electrical loopback buffers, label extractors (LE), and fiber delay lines (FDLs). The LION switch uses a forward-and-store strategy for packets, as opposed to the store-and-forward strategy employed in an electrical switch. Only the contending packets that fail to get grants are stored in the LB (see [5] for details on the different buffer architectures). Note that, the use of k_r RXs per node, together with the unique wavelength routing property in AWGR, and wavelength division multiplexing (WDM), naturally implements output queuing (OQ). OQ is very challenging in electronics,

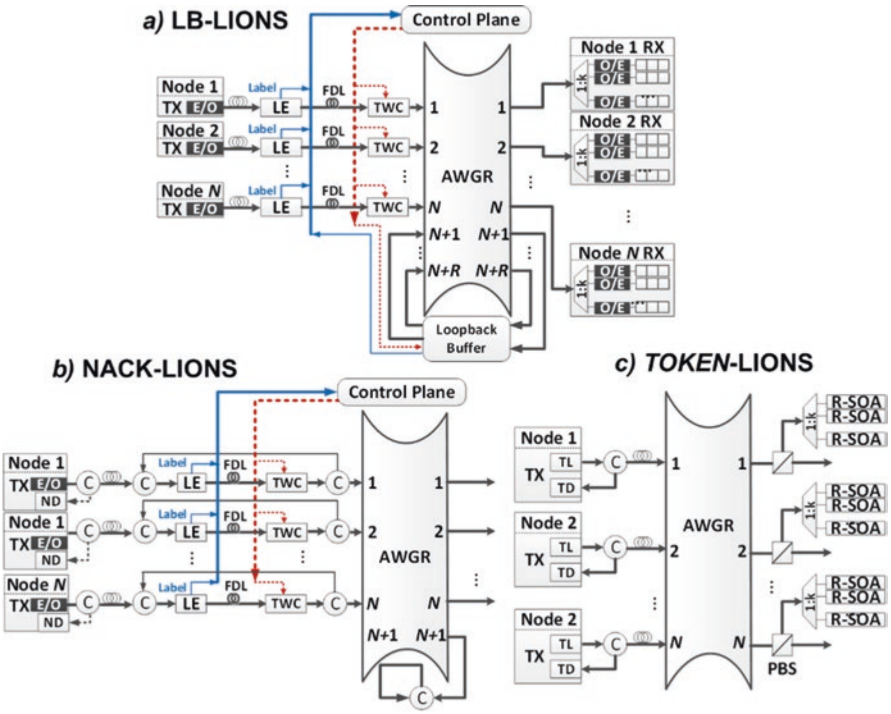


Fig. 6.2 (**a**) Loopback buffer LIONS. (**b**) All-optical NACK-LIONS. (**c**) All-optical TOKEN-LIONS. *LE* label extractor, *FDL* fiber delay line, *C* optical circulator, *TWC* tunable wavelength converter, *ND* NACK detector, *TD* TOKEN detector, *RSOA* reflective semiconductor optical amplifier

especially at high bit rates. However, because the LB still requires to work at line-rate speed (10Gb/s or higher), and it also needs a large amount of O/E/O conversion, in [6] we proposed the all-optical negative acknowledgment (AO-NACK) technique. This technique allows to replace the whole LB with a simple optical circulator (C) acting as reflector. This solution greatly simplifies and improves the line-rate scalability and cost of the LION switch. Exploiting the duplex nature of the AWGR, the “dropped” packets are simultaneously directed to the circulator port and reflected to the TX nodes, where a simple edge detector (ND – NACK detector) senses that a packet has been sent back (NACK) (see Fig. 6.2b). The node receiving the NACK retransmits the packet based on certain retransmission policies. Ref. [7] shows that the LIONS-NACK performs as well as LIONS-LB.

To further simplify the LIONS, we also designed the all-optical token (AO-TOKEN) LIONS (see Fig. 6.2c), which exploits the saturation effect in a reflective semiconductor optical amplifier (RSOA) [4]. AO-TOKEN removes the centralized control plane (now optical and distributed) and the TWCs. The basic idea is the use of one or more RSOAs as the mutual exclusion (mutex) type of arbiter at each output port of the AWGR (up to k_r RSOAs per output port). The TOKEN

Table 6.1 Comparison among the different LIONS architectures

	LB-LIONS	NACK-LIONS	TOKEN-LIONS	TONAK-LIONS
Pros	High throughput and low latency	High throughput and low latency	Distributed CP; no TWCs	Same as NACK and TOKEN
Cons	Power hungry LB and TWCs; centralized CP	Power hungry TWCs; centralized CP	Performance affected by distance and packet size	More complex than TOKEN

signaling and data can share the same physical layer by using polarization diversity, as demonstrated in [8] (there is a polarization beam splitter – PBS – at each AWGR output to separate the TOKEN and data signals). The transmission of a packet happens only after a node applies for and receives a grant from the specific output port arbiters. The major advantage of the TOKEN technique is that it distributes the contention resolution in the control plane without the requirement of a global coordination scheme. However, the delay caused by the wait for the token response can negatively affect the switch performance, as studied in [7]. Overall, the host-switch distance and the packet size play a significant role. It is possible to overcome the above limitation by combining TOKEN and NACK techniques in the same architecture, named TONAK-LIONS [4]. Table 6.1 summarized pros and cons of these different implementations of active LIONS.

6.3 LIONS-LB Testbed Demonstration

Figure 6.3 (Top) shows a testbed for a 4x4 LIONS using a 32 × 32 50 GHz-spacing AWGR. The testbed also includes wavelength converters (WCs) based on cross-phase modulation (XPM) in a semiconductor optical amplifier Mach-Zehnder interferometer (SOA-MZI). Each WC accepts one continuous wave (CW) input signal from a tunable laser diode (TLD) board. The TLD guarantee nanosecond switching time over the entire C band with a wavelength accuracy of 0.02 nm. By reading the 5-bit parallel control signals coming from the FPGA-based control plane, each TLD board tunes its wavelength according to a routing table stored on a complex programmable logic device (CPLD) chip mounted on the board itself. An optical path is established between each AWGR input and output on a packet basis, according to the destination address carried by each packet label.

The control plane and loopback buffers are implemented using a Xilinx Virtex 5 FPGA ML523 characterization platform, which is capable of instantiating eight high-speed RocketIO GTP tile transceivers connected to 16 pairs of differential SMA connectors. Four pairs of the transceivers are used as the four control plane channels, each of which receives the labels from its label extractor (i.e., the 90/10 splitter). The labels are encapsulated in the packet headers in the time domain for simplicity. Note that the labels can be wavelength multiplexed together with the payload to separate control and data planes. The control plane reads the label and

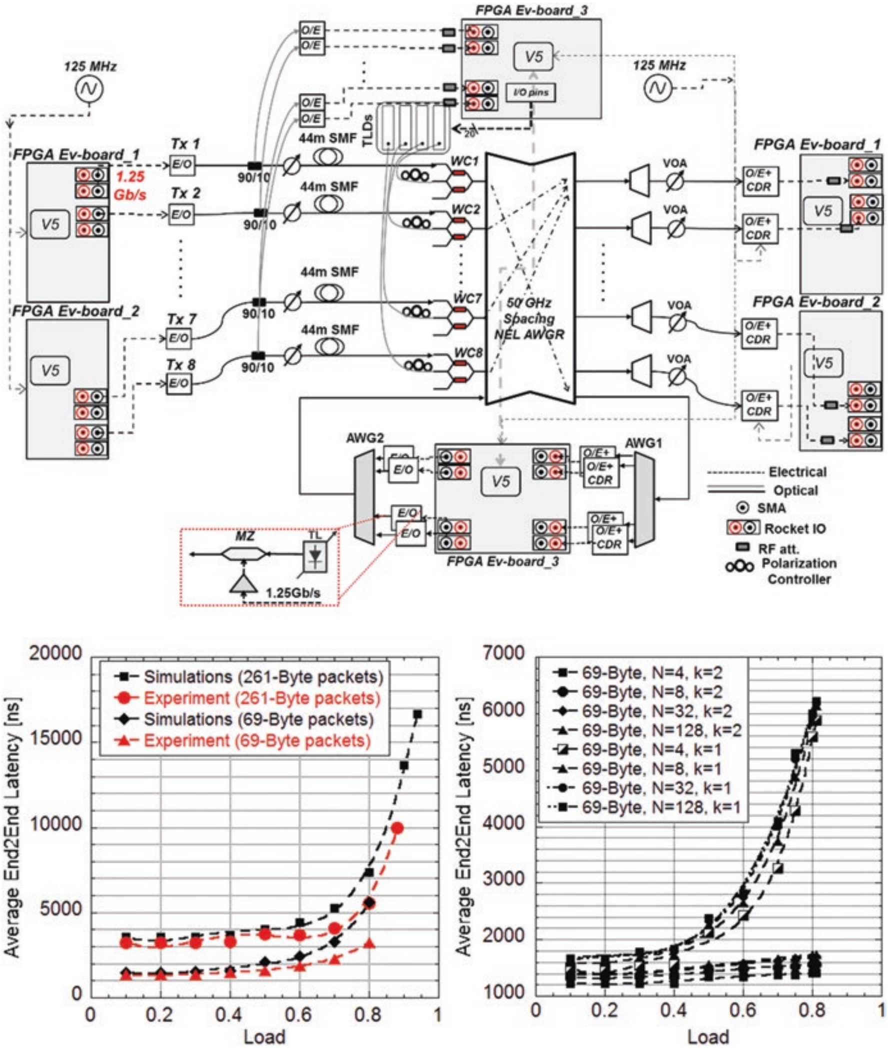


Fig. 6.3 (*Top*) 4 × 4 LIONS testbed with shared loopback buffer. (*Bottom*) Comparison of simulation results with experimental results (*left*); projection of simulation results to high-port count with large packet size and k = 1, 2 (*right*)

generates the 5-bit control signals to TLD boards after arbitration. The contended packets that fail to win arbitrations are directed to the inputs of the loopback buffer.

The I/O ports of the loopback buffer are implemented using another four pairs of RocketIO transceivers. Since the WCs will turn the optical power on and off while switching, burst-mode receivers (BM-Rxs) are used in the testbed beside the RocketIO interfaces of end-hosts and loopback buffers.

Four MicroBlaze soft processor cores [9] were instantiated on two Virtex 5 FPGA boards (same model used for the control plane) with MPI interfaces capable of doing remote direct memory access (RDMA) operations. The data are firstly written to the BRAM block on FPGA and then moved into the RocketIO transmitter output queue using direct memory access (DMA) operation. Then the packets are encapsulated and serialized by RocketIO at the 1.25 Gbps output line rate. On the Rx side, the received data packets are directly moved from the input queue to the DDR2 SDRAM memory on board using DMA operation.

The end-to-end latency is one of the important performance metrics to the switch. A synthetic traffic model is used in the testbed. The data streams at each host are encapsulated into fixed size packets with uniform random destination address. Each packet is with 5-byte header (2-byte preamble, 1-byte destination address, 1-byte source address, and 1-byte packet length). Different offered load values can be achieved by changing the guard time between packets. Note that a minimum guard time of 17 bytes should be guaranteed due to the hardware constraints (i.e., worst case TLD tuning time, burst-mode receiver settling time and comma alignment delay in SERDES, etc.). Since the traffic is uniformly randomly distributed, the end-to-end latency statistics can be collected at any of the output ports. In the experiment, only host 2 was used to collect data.

Figure 6.3 (Bottom, left) shows the comparison of statistic results from both simulations (black lines) and experiments (red lines). The square and circle dots show the 4x4 experimental and simulation data with $k = 1$ and packet size of 256 bytes plus 5 bytes for the header. The diamond and triangle dots show the same results with packet size of 64 bytes plus 5 bytes for the header. Here k means the number of parallel wavelengths can be received by the same host simultaneously from one output port of the switch [10]. As shown, the comparison of the results shows a close match between the experimental data and the simulation data, which verifies the correctness and accuracy of the simulator we developed. The other curves in Fig. 6.3 (bottom, right) show the projection of the results to high-port count and to $k = 2$ case. The increase of the LIONS radix does not significantly affect the end-to-end latency, while $k = 2$ can dramatically reduce it since it reduces the contention probability at each output port.

The line rate of the testbed is simply limited by the commercially available the BM-CDR modules running only at 1.25 Gbps. Despite the low line rate of the demo, the match between the simulation and experiment is still very meaningful to assess the correctness of our simulation framework. In fact, considering that the simulation at 10 Gbps or higher line rate will only change the packet transmission time (the line rate of label is independent to the line rate of the payload when using wavelength multiplexed labels), the arbitration process stays the same no matter what line rate is used in the simulation.

6.4 All-Optical TOKEN/TONAK Demonstrator

It is evident from the experimentation reported above that it would be desirable to reduce the complexity of the LIONS active switch by removing the loopback buffer as well as the centralized controller and the many tunable wavelength converters. It is indeed possible to achieve this with an all-optical solution that (1) does not have the complexity nor the power consumption of the loopback buffer solution and (2) is distributed and yet, at the same time, exhibits low latency and high scalability. To this aim, we introduced the use of a reflective semiconductor optical amplifier (RSOA) [11], a widely used optical active component, as a distributed mutual exclusion element (mutex). We also proposed a simple protocol to detect contention and retransmit packets without incurring a significant latency overhead. Most importantly, we demonstrated that the proposed solution is simple to implement and makes the control plane of an AWGR-based optical switch fully distributed and hence arbitrarily scalable.

The gain saturation effect in a RSOA can be used to realize mutual exclusion behavior. Let us assume that *N* different nodes can make simultaneous requests R_1, $R_2 \ldots R_n$ (using different wavelengths λ_1, λ_2, … λ_n) to the RSOA associated with a given AWGR output port. The first request, say R_i, saturates the RSOA, which results in P_{tot} power reflected back to the sender node *I*. The RSOA stays saturated as long as the request on λ_i is held. A detector that is set to trigger at P_{tot} produces the grant signal. If another request R_j (on λ_j) from node *J* arrives while R_i is still active, the power reflected at λ_j will be $\approx P_{tot}/2$ (because of the saturation effect in the RSOA), which is not enough to set the trigger condition; hence the second request will be excluded.

The experimental demonstration of the RSOA-based contention resolution is shown in Fig. 6.4. Two polarization-diverse TXs (PD-TX) are connected to input ports 1 and 4 of a 200GHz-spaced 8*x*8 AWGR (8 dB uniform insertion loss). Polarization controllers (PCs) at the AWGR inputs align the signal polarization with the PBSs at the AWGR outputs. Alternatively, all polarization maintaining (PM) components could be used. Each PD-TX includes a PBS and polarization beam combiner (PBC) to multiplex in the polarization domain the data and control plane request paths. The mutex arm of the PD-TX includes a Mach-Zehnder (MZ) modulator. Two MZs are used in the data arm as data modulator and gate. The gate is controlled by an FPGA and remains open unless the request is not granted. The FPGA also generates the control plane requests, while the 10 Gb/s 406.9 ns-long packets are generated with a pattern generator, with each packet containing a portion of 2^{31}–1 pseudorandom bit sequence (PRBS). A PBS is placed at AWGR output 3. The PBS extracts the requests, which enter an RSOA implemented here with an optical circulator and a SOA. The PC at the SOA output maximizes the optical power going back through the PBS and reaching the token detectors (TDs), also called grant detectors (GDs). The second PBS output connects to an O/E converter for BER measurements on the data path.

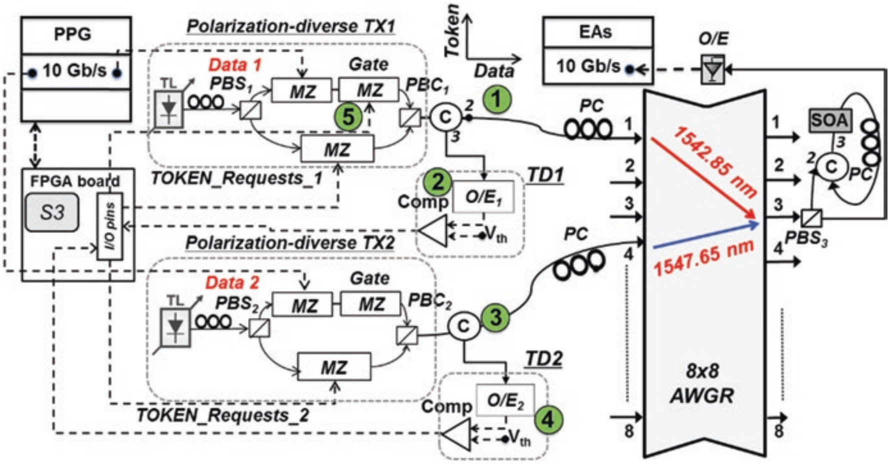


Fig. 6.4 Experimental testbed: *PC* polarization controller, *PBS* polarization beam splitter, *PBC* polarization beam combiner, *MZ* Mach-Zehnder modulator

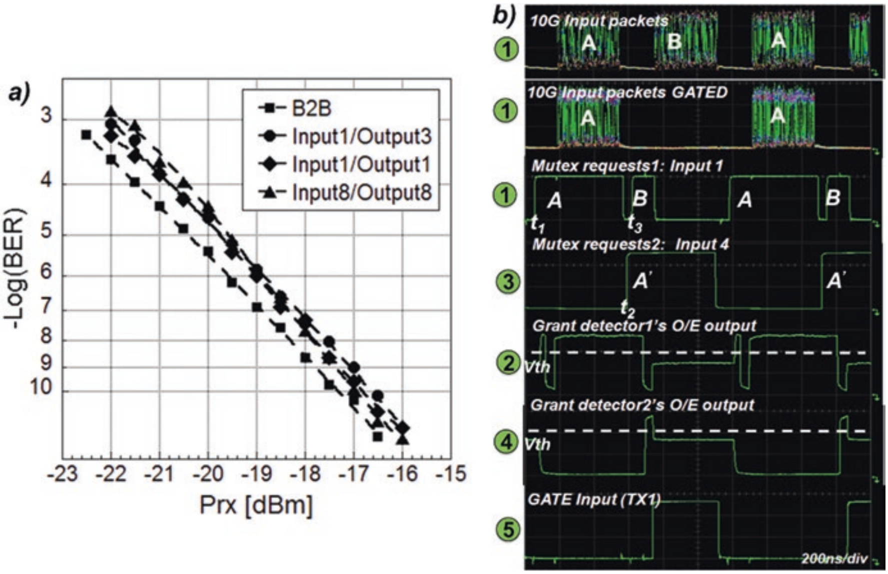


Fig. 6.5 (**a**) BER measurements; (**b**) measured traces

Figure 6.5b shows the measured traces for the packets at AWGR input 1, mutex requests at AWGR inputs 1 and 4, GD1's and GD2's O/Es and comparators outputs, and gate 1 output. Numbered dots refer to the points at which, in the experiment setup of Fig. 6.4, each trace was measured. Note that there is a delay between the leading edges of the mutex requests and the relative GD outputs. This delay is simply due to the propagation delay caused by the fiber pigtails of the bulky components used in the experiment. There is also a small delay between GD1's O/E converter leading edge (related to mutex request B) and the gate input (TX1) signal. This small delay is ≤ two FPGA clock cycles (≤ 12.8 ns in this experiment since the FPGA was running at 156 MHz).

Figure 6.5a shows BER measurements for the data packets at AWGR output 3, with one of every two packets coming from Node1 blocked by the RSOA-based all-optical mutex. The power penalty, compared to the back-to-back (BtB) curve (black squares), is negligible. These results demonstrate that the mutex technique works properly, granting the transmission of A packets only upon successful request, while transmission of B packets is always denied. These results also demonstrate that the coherent crosstalk penalty caused by the mutex requests on the related packets under transmission is not a serious problem for the proposed implementation. In fact, the polarization extinction ratio of the PBS (≈30 dB) and the power values used in the experiment (−13 dBm and −6 dBm are the power values at the PBS3 outputs for data and control plane requests, respectively) guarantee a signal to coherent crosstalk ratio ≈25 dB [12].

As explained above, the RSOA-based mutex exploits the saturation effect in SOAs. Therefore, the technique is subject to the wavelength dependence of the RSOA gain, which can pose a higher and lower bound to the wavelength operating range of the technique, under the assumption that the V_{th} in the GD is kept constant, as in this experiment. In practice, the technique can work over a wider wavelength range, which can be considered approximately equal to the 1 dB bandwidth of the SOA used in this experiment, i.e., ≈ 40 nm. More experimental details and analysis about the wavelength operating range, the crosstalk impairments in this system, and the minimum interval between two successive requests that guarantees the earliest request can be found in [13].

As mentioned above, in TOKEN-LIONS, the host-switch distance and the packet size affect significantly the switch performance (see [7] for more details). Fortunately, it is possible to overcome the above limitation by combining the TOKEN solution with the all-optical NACK technique demonstrated in [6]. We named this new architecture as TONAK-LIONS [4].

Figure 6.6 illustrates the testbed used for the proof-of-concept experimental demonstration of TONAK architecture. A Virtex5 FPGA evaluation board generates token requests and related packets. The packets are generated through a rocket IO GTX interface, which limits the line rate used in this experiment to 6 Gb/s. Standard user IOs pins are used to control and tune two fast tunable lasers (TLs) [13, 27], i.e., TLD 1 and TLD 2. The FPGA tunes TLD 1 (TLD 2) sending two control signals named txa_tld_bit (txb_tld_bit) and txa_tld_en (txb_tld_en). TLD 1 is the laser for transmitter A (TX-A), which connects to the TONAK switch input port 1 through a TONAK linecard. TX-A generates a sequence of token requests and related packets, as explained in detail later. TLD 1 connects to a 3 dB power splitter. One splitter output connects to a Mach-Zehnder (MZ) modulator for data packet modulation (a 10 GHz electrical amplifier drives the modulator with the data generated by rocket IO interface). The modulator connects to the data input of the TONAK linecard through an optical circulator, which extracts the counterpropagating AO-NACK

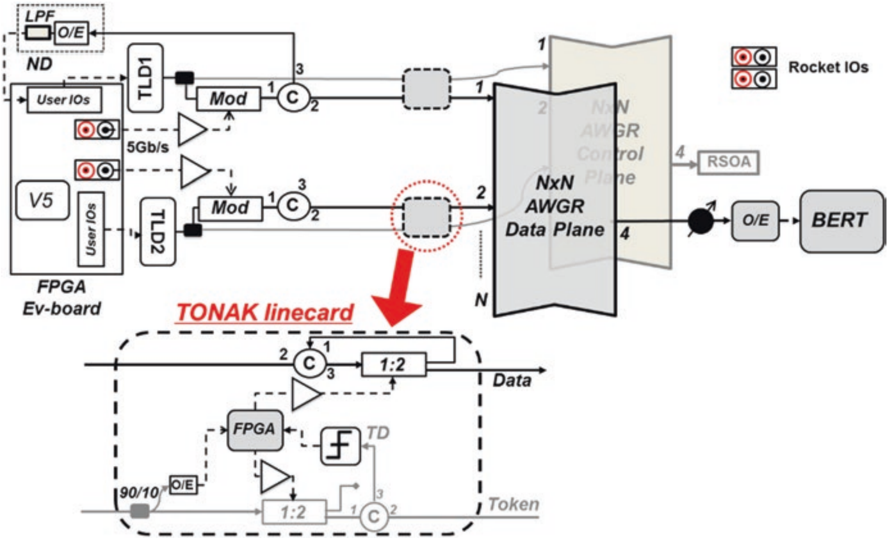


Fig. 6.6 Experimental setup of the TONAK-LION switch testbed. *V5 FPGA* Virtex 5 Field Programmable Gate Array, *ND* NACK detector, *LPF* low-pass filter, *Mod* modulator, *C* optical circulator, *ED* edge detector, *TD* token detector, *AWGR* arrayed waveguide grating router, *SOA* semiconductor optical amplifier, *α* variable optical attenuator, *BERT* bit error rate tester

messages. The AO-NACK messages are then detected by a NACK detector (ND) connected to one FPGA IO pin. The ND in this experiment is implemented with a simple 1.25GHz O/E converter (with limiting amplifier) and a 400 MHz low-pass filter. The second splitter output connects directly to the TOKEN input of the TONAK linecard. The TONAK linecard has two inputs and two outputs. The DATA output (black) connects to the DATA plane AWGR, while the TOKEN output (gray) connects to the TOKEN plane AWGR. The DATA path (black) contains an optical circulator followed by a 1:2 MZ switch. Its default position is in bar state (output connected to optical circulator). If the TOKEN response coming from the distributed control plane is positive (TOKEN detector output is “1”), a V5 FPGA changes the MZ switch to cross state (output connected to AWGR DATA plane) to let the incoming packet going to the AWGR DATA plane input and reach the desired output. In case the response to a TOKEN request is negative (TOKEN detector output is “0”), the incoming packet is reflected to the TX, where the ND detects it.

The TONAK linecard TOKEN path (gray) contains a 90/10 splitter, a 1:2 MZ switch, and a circulator. Default state for the MZ switch is bar (idle output). The power splitter taps 10% of the optical power of an incoming token request to feed a token detector (ED). Since a TOKEN request is initiated with a change of TL wavelength from $\lambda_{\mathrm{DEFAULT}}$ to $\lambda_{\mathrm{SIGNAL}}$ (see Table 6.2), the TD is composed of a passband filter centered at $\lambda_{\mathrm{DEFAULT}}$ and an O/E converter. When the TD senses an incoming TOKEN request, the V5 FPGA, triggered on the falling edge of the TD output, changes the 1:2 MZ switch to cross state to let the request reaching the TOKEN

Table 6.2 Wavelength values used in the experiment

	λ_{DEFAULT} (tx_tld_bit)	λ_{SIGNAL} (tx_tld_bit)
TLD 1	1547.85 nm (0)	1550.1 nm (1)
TLD 2	1547.35 nm (1)	1547.6 nm (0)

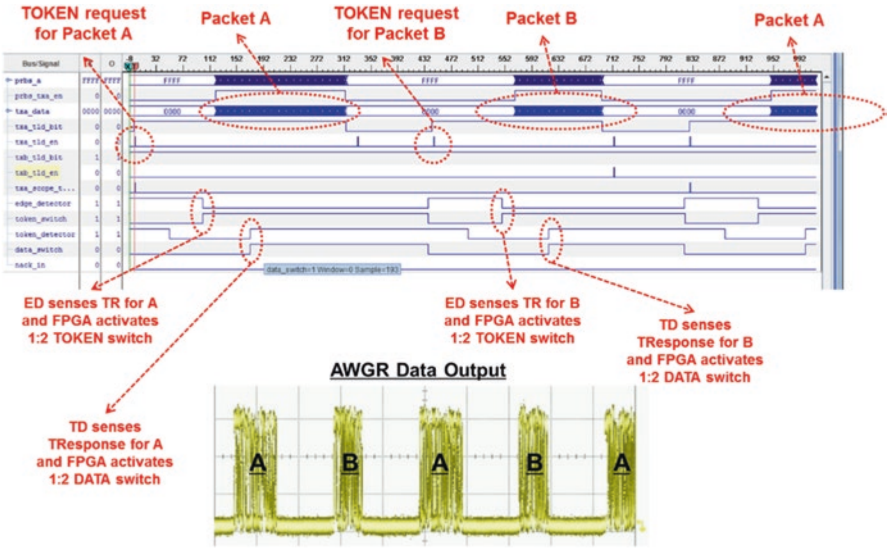

TOKEN request for Packet A
Packet A
TOKEN request for Packet B
Packet B
Packet A
ED senses TR for A and FPGA activates 1:2 TOKEN switch
ED senses TR for B and FPGA activates 1:2 TOKEN switch
TD senses TResponse for B and FPGA activates 1:2 DATA switch
TD senses TResponse for A and FPGA activates 1:2 DATA switch
AWGR Data Output
A
B
A
B
A

Fig. 6.7 ChipScope experimental timing diagram demonstrating the TONAK technique in the case of no contention

plane AWGR input and then the SOA at the desired AWGR output. If the token response is positive, the 1:2 MZ switch stays in cross state for the entire packet duration. In case of a negative token response (desired TONAK switch output is not available), the FPGA change the MZ switch state back to bar state.

Two 50 GHz-spacing 32 × 32 AWGR with uniform insertion loss of 8 dB and cyclic frequency characteristic (ULCF AWGR [28]) represent the core of the TONAK switch architecture. One AWGR (black) acts as the data plane switch fabric, while the other AWGR (gray) implements, together with a SOA, the distributed all-optical TOKEN-based control plane described above. Because an RSOA was not available, we emulated the RSOA function with a SOA and an optical circulator.

In this proof-of-concept demonstration, TX-A generates endlessly two packets, A and B (Fig. 6.7), directed to the TONAK switch output 4. TLD 2, connected to input 2 of the TOKEN plane AWGR, generates only a periodic contending token request that causes contention for packet B. Because of this contention event, packet B is reflected to TX-A and retransmitted at a later time, as shown in Fig. 6.7. The SOA is placed at output 4 of the TOKEN plane AWGR.

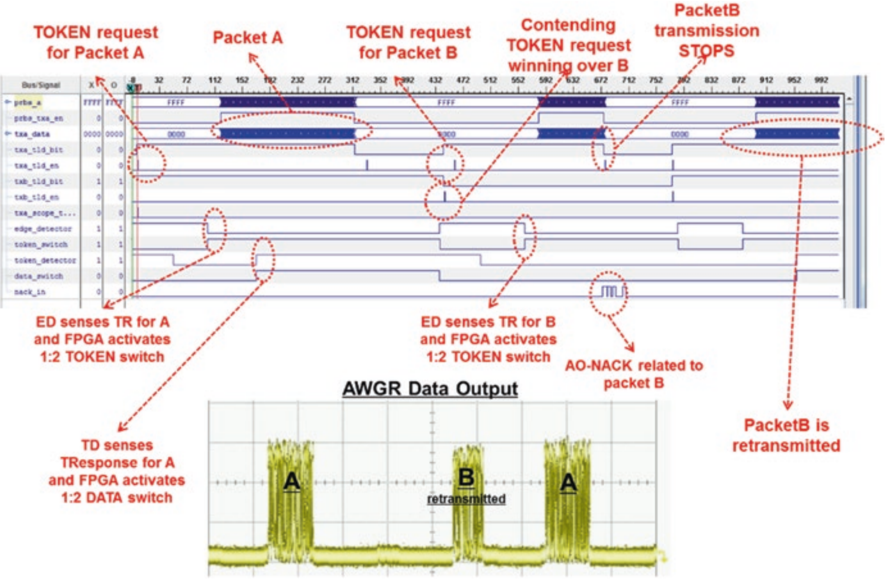


Fig. 6.8 ChipScope experimental timing diagram demonstrating the TONAK retransmission technique in case of contention

Table 6.2 shows the wavelength values used in the experiment and related values for the control bit signals (txa_tld_bit and txb_tld_bit). The wavelength values named as $\lambda_{\mathrm{SIGNAL}}$ are determined by the AWGR routing table. In particular, 1550.1 nm and 1547.6 nm are the wavelength values to reach AWGR output 4 from AWGR inputs 1 and 2, respectively. The values named as $\lambda_{\mathrm{DEFAULT}}$ do not belong to the AWGR grid so that the optical power from TLD 1 and TLD 2 is blocked when no packets have to be transmitted. This is important in an actual implementation to avoid crosstalk at the switch outputs.

Figures 6.7 and 6.8 illustrate two timing diagrams showing traces acquired with Xilinx ISE ChipScope tool, which allows capturing the electrical signals at the different FPGA IOs during the experiment. Figure 6.7 timing diagram is for a case in which no contending token requests are generated by TLD 2. Figure 6.8 timing diagram shows the case when the control plane detects contention for packets B, which are then retransmitted. Note that the numbers at the top of each timing diagram represent the time evolution in clock cycle (2.67 ns/clk in this experiment). Clock cycle “0” corresponds to the trigger event given by the enable pulse for TLD 1, which determines the beginning of a TOKEN request for packet A.

So, when the FPGA generates a pulse enable signal on *txa_tld_en* IO pin and sets the *txa_tld_bit* pin output to “1,” TLD 1 tunes its wavelength from $\lambda_{\mathrm{DEFAULT}}$ to $\lambda_{\mathrm{SIGNAL}}$. After a certain amount of clock cycles, the TOKEN request A reaches the ED that triggers the FPGA (edge_detector signal on Chip_scope goes “low”) in TONAK linecard, which then sets the MZ TOKEN switch in *cross* state (ChipScope *token_* *switch* signal goes "high"). In this way, the TOKEN request can reach the SOA at output 4 of the TOKEN plane AWGR.

After 80 clock cycles (equivalent to the round-trip time for the TOKEN request to reach the SOA, being reflected and reach the TOKEN detector), the TOKEN detector senses the reflected TOKEN request. Since the SOA was not saturated (which means that the target TONAK switch output is available), the optical power is enough to trigger the TOKEN detector (Chip_scope *token_detector* signal is "high"). Then, FPGA sets the MZ DATA switch to *cross* state (*data_switch* signal on Chip_scope goes "high"), allowing the incoming packet *A* to enter the AWGR data plane and being routed to the desired output port 10. Note that packet *A* transmission (*txa_data* on Chip_scope) starts with a certain delay compared to the related TOKEN request. This delay is to account for the latency in the TL board and the ED to TD round-trip time.

When transmission of packet *A* has been completed, FPGA sets *txa_tld_bit* at "0" and generates an enable pulse on *txa_tld_en* to return TLD 1 to $\lambda_{\mathrm{DEFAULT}}$. Transmission of packet B follows the same process described above. The only difference is the length of packet *B*, which is two-thirds of packet *A* length. Both packets *A* and *B* contain a different portion of a PRBS 2^{15}–1.

Figure 6.8 shows the case with contention. The contention happens for packet *B*. Note that, a few clock cycles before the generation of TOKEN request for packet *B*, *txb_tld_bit*, is set at "0," and an enable pulse is generated on *txb_tld_en*. This means that TL-B tunes from its default position to 1547.6 nm, the wavelength values to reach and saturate the SOA. This time, the TOKEN request for packet *B* finds the SOA already saturated and reaches the TD with an optical power value too low to trigger the TD. The FPGA, not seeing the *token_detector* signal going "high" when expected, understands that the TOKEN request for packet B is not successful (desired AWGR output is not available). Then, the MZ data switch is left in default position (*bar* state), and the incoming packet B gets reflected to TX-A, where it gets detected by the NACK detector (see *nack_in* Chipscope signal going "high"). As response to the detection of an AO-NACK message, TX-A stops immediately transmission of packet *B* (which was still under transmission, brings TL-A back to its default wavelength, and schedules retransmission of packet *B* at a later time (around clock cycle number 886)). This time, retransmission of packet *B* is successful.

Figure 6.9 shows BER measurements at AWGR output 4 for the contention-less (squares) and contention (triangles) scenarios described above. In both cases the BER reaches error-free condition. Note that there is some penalty associated with the switched packets in case of contention and retransmission of packets B. This penalty is caused by the limited extinction ratio (< 20 dB) of the 1:2 MZ switch used in the experiment. Note that devices with higher extinction ratio (> 30 dB) are available (http://www.eospace.com/switches.htm). The finite ER of the 1:2 switches can cause out-of-band crosstalk. It is then important to maximize the ER at the 1:2 MZ switch output. In fact, the worst case for this type of out-of-band crosstalk is when *N*-1 inputs are trying to send data to the same output simultaneously.

Assuming *k* RSOAs per output port, there would be *k* requests granted and *N*-1-*k* requests rejected, causing up to *N*-1-*k* sources of out-of-band crosstalk. Since each

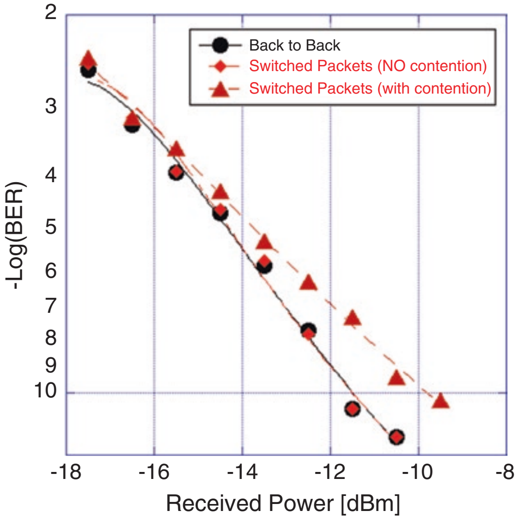


Fig. 6.9 BER measurements: Back to Back (*circle*); Switched packets at AWGR output 4 without contention (*squares*) and with contention for packets B (*triangles*)

node has one 1:*k* demux and *k* receivers, each receiver would only see (*N*-1-*k)/k* crosstalk terms. For *N* = 128 and *k* = 4, this would be equivalent to 30 crosstalk terms (a factor of 14.7 dB). So, in the worst case, the signal to crosstalk ratio would be 30–14.7 = 15.3 dB. This value, according to Ref. [14], gives negligible power penalty.

6.5 Passive LIONS Demonstrations

In the sections above, we discussed the experiment testbeds and demonstrations for different AWGR-based active LIONS architectures. Being these switches, as any other optical switch architecture, bufferless, they cannot be cascaded. Therefore, they can be used mainly as core switches in folded CLOS type of architectures (i.e., fat tree) or in directly connected architectures like torus, flattened butterfly, or Hyper X where these optical switches can interconnect directly computing nodes or top-of-rack (ToR) switches.

As part of the research efforts carried out at UC Davis NGNS laboratories, this section reports experimental demonstration of passive LIONS hierarchical all-to-all architectures. The AWGR is still the core wavelength routing component, but the switching of the packets happens at the edges and in the electrical domain. This approach can still support packet switching and circuit switching but represents more of a custom solution for data center or HPC architectures that requires high-bandwidth all-to-all interconnection at the rack or cluster level (see [15] for more details). This architectures could be very suitable also for on-board or on-chip interconnects as discussed in [16, 17], but we will not address this aspect in this chapter.

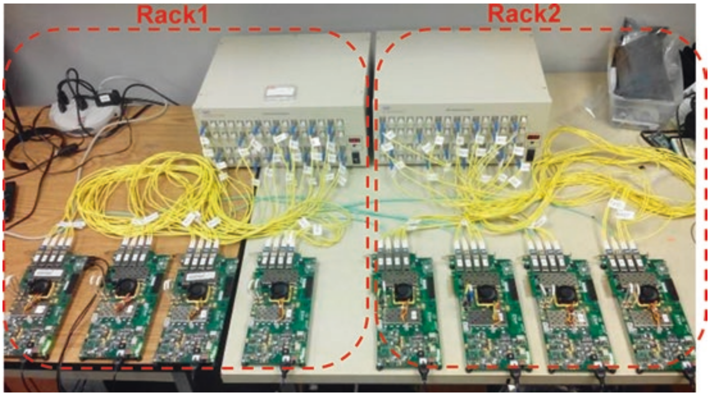


Fig. 6.10 Picture of the hardware demo using eight VIRTEX 7 FPGA boards and two 32-port AWGRs

6.6 Hierarchical All-to-All Eight-Node Demo

As part of the emerging technologies demo session at the 2014 Super Computing Conference, we build the demo shown in Fig. 6.10. This is a hardware testbed for a hierarchical all-to-all cluster composed of two racks and a total of eight computing nodes. Each node is implemented with one Xilinx VC709 FPGA board equipped with a Virtex 7 chip. Each FPGA implements traffic generation function and embedded switch functions with four 10Gbps WDM SFP$^+$ transceivers (TRXs): three for all-to-all intra-rack communication and one for inter-rack communication (within the same cluster). The wavelengths used for the experiment are in the range 1546.04–1561.04 nm, with a 0.4 nm (50 GHz) frequency grid. The network traffic generated by the FPGAs is converted in the optical domain, wavelength-routed by two 32-port AWGRs (one per rack), and received by the destination servers (FPGAs). The AWGRs' insertion loss is 8 dB. The TRXs' output power is ~3 dBm and the RX sensitivity is −24 dBm at BER = 10^{-12}.

Figure 6.11 shows the architecture of the emulated embedded switch and traffic generator inside the FPGA VIRTEX 7 chip. The core switching fabric is an 11 × 5 crossbar (including the virtual channels). A matrix arbiter implements one-cycle fair arbitration in the 11 × 5 crossbar. The inter-rack port (the green port in Fig. 6.11) has one virtual channel (VC) for the packets directed to the server and three virtual channels for redirecting in the incoming packets to the other three nodes in the same rack. The traffic generator has also four VCs: one for each of the intra-rack ports and inter-rack port.

Under uniform random traffic, the normalized system-wide network throughput of the cluster testbed is higher than 97% with latency below 364 ns, only limited by the FPGA speed (see Fig. 6.12). To compare the testbed results with simulations, we used the parameters measured from the testbed (see Table 6.3) in the simulator. The results shown in Fig. 6.12 evidence a very similar behavior between experiment and

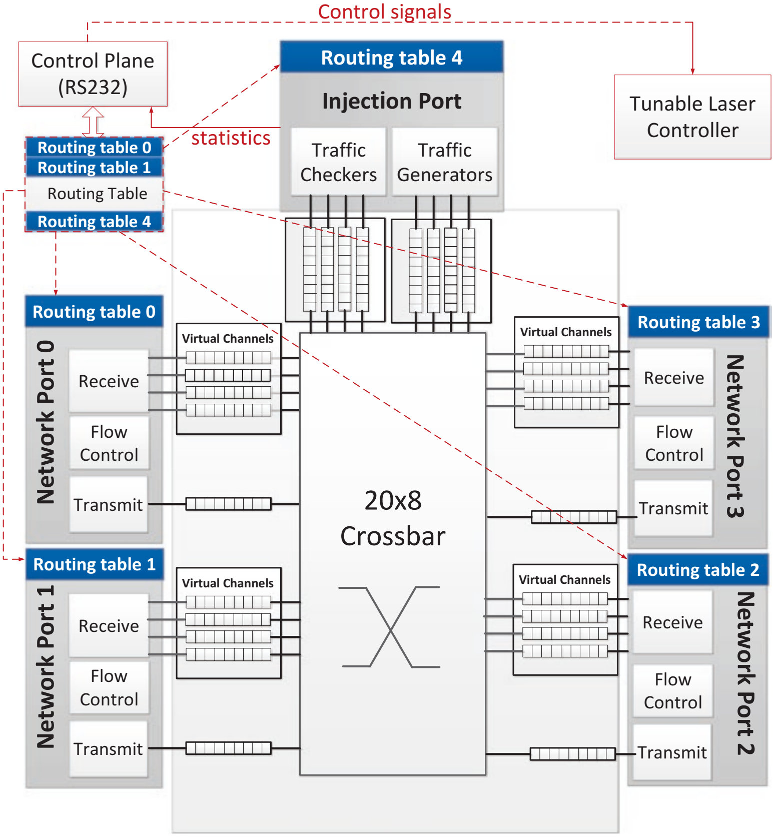


Fig. 6.11 Emulated intra-cluster top-level switch

simulated scenarios, validating the accuracy of the simulator. The packet size used for the simulation and experiment is 256 B. The slight difference between the simulated and experimental curve is most likely due to some differences in the characteristics of the random traffic generation in the hardware and simulator.

6.7 Flexible Bandwidth Optical Data Center Core Network with All-to-All Interconnectivity

Building upon the architecture and demo results presented above, we performed an experiment demonstration of flexible bandwidth allocation exploiting wavelength routing and multiplexing property of AWGR. Figure 6.13 shows the experimental setup. Eight Xilinx VC709 boards with high-speed Rocket I/O TRXs at 10 Gb/s emulate eight nodes (these could be eight servers, ToRs, or clusters). The TRXs are

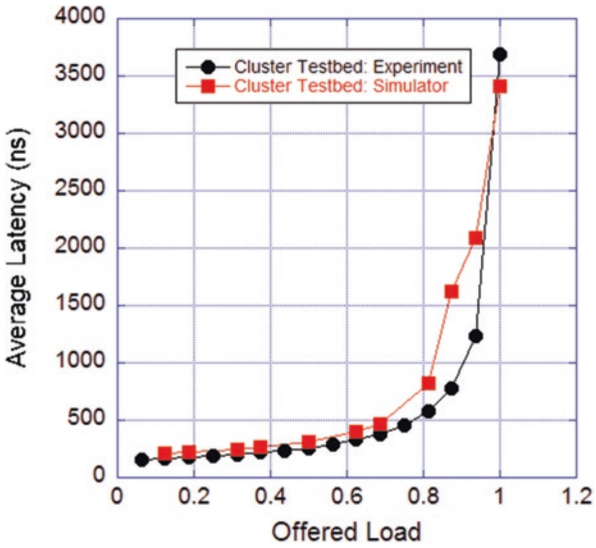


Fig. 6.12 Latency vs. throughput performance comparison between hardware and simulations

Table 6.3 Simulation parameters

Parameter	Value	Parameter	Value
VC's buffer size	4Kbyte	SerDes RX&TX Delay (GTH)	113 ns ≈ 14 (cycles)
Delay in receiver module(RX)	4 (cycles)	Wire delay per meter	5 ns
Delay in transmit module(TX)	4 (cycles)	Delay in passive AWGR	1 ns
Switching delay(crossbar)	5 (cycles)	Active AWGR switching	12 ns

connected to two 32-port AWGRs. The two AWGRs communicate by a single fiber (carrying WDM signals). The AWGRs' channel spacing is 50 GHz, and their insertion loss is 8 dB. The wavelengths used in the experiment are in the range 1546.04 ~ 1561.04 nm on a 0.4 nm (50 GHz) grid.

Each FPGA board makes use of three TRXs for all-to-all intra-region communication and one TRX for inter-region communication. TX1 and TX3 in FPGA1 and TX3 in FPGA4 and FPGA5 are implemented with fast tunable lasers with tuning time as short as few nanoseconds for fast flexible bandwidth adjustment between hot spots. All the other TXs and RXs are commercial small form pluggable (SFP) TRXs at 10 Gbps with a RX sensitivity of −24 dBm.

Each FPGA works as intra-cluster top-level switch (TLS) and traffic generator, as already discussed above and shown in Fig. 6.11. Each TLS has four network ports (each one connected with one of the four TRXs), one injection port with four independent 10Gb/s traffic generators, and one 20 × 8 crossbar switch. Each net-

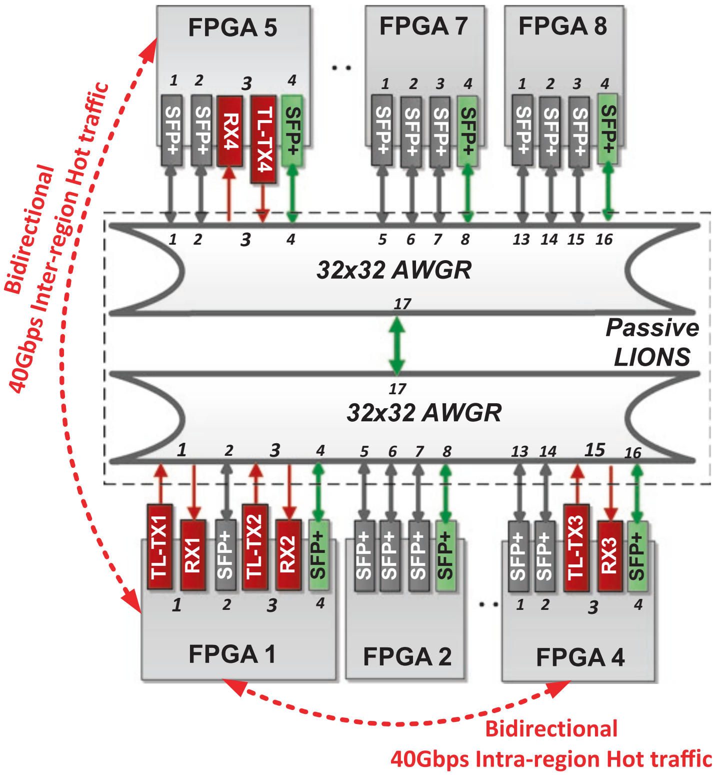


Fig. 6.13 Experiment setup of the full system interconnection network. *TL-TX* tunable transmitter (10 Gb/s), *RX* receiver (10 Gb/s), *SFP+* small form pluggable transceiver (10 Gb/s), *AWGR* arrayed waveguide grating router, *FPGA* field programmable gate array

work port implements a virtual output queuing architecture (four virtual channels) to avoid the head-of-line blocking issue. Each injection port can generate up to 40 Gbps traffic. The crossbar switch performs switching among 20 input channels and 8 output channels. To perform seamless network reconfiguration, the TLS makes use of two routing tables: a working table and a look-ahead table. The working table is used for forwarding the packets in the default hierarchical all-to-all scenario, while the look-ahead table is used for accepting the new table content containing the new routing information required for the reconfigured network. All the updated routing information is written into the look-ahead table during the network reconfiguration. After the look-ahead table has been updated, the controller tunes the tunable lasers and uses the look-ahead table as the new working table.

To measure the performance under hot-spot traffic, each FPGA board generates up to 10 Gbps of background (cold) traffic with uniform random distribution and 40 Gbps bidirectional hot-spot traffic between two hot-spot points in the network. As shown in Fig. 6.13, we demonstrated the four scenarios: 40 Gbps hot-spot traffic between FPGA 1 and 4 (intra-region) with and without channel bonding and 40

Table 6.4 Wavelength allocation of the TL-TXs

	Without flexibility	With intra-region flexibility	With inter-region flexibility
TL-TX1	1561.41 nm	1552.50 nm	1561.41 nm
TL-TX2	1559.79 nm	1559.79 nm	1548.08 nm
TL-TX3	1546.04 nm	1552.50 nm	1559.79 nm
TL-TX4	1559.79 nm	1559.79 nm	1548.08 nm

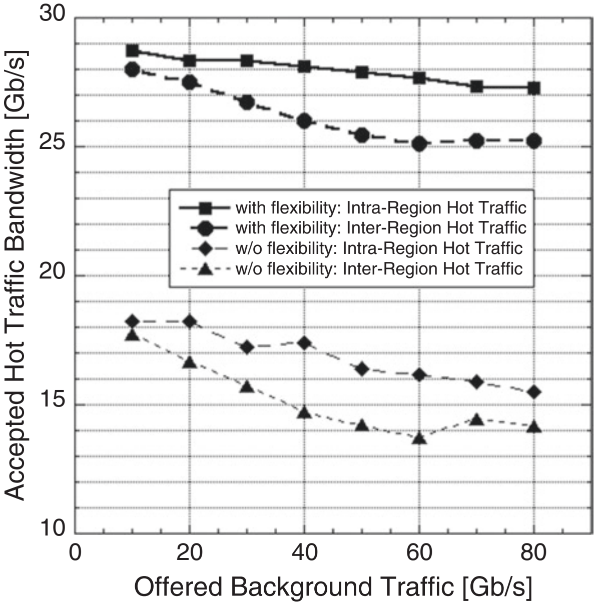


Fig. 6.14 Measured accepted hot-spot traffic bandwidth

Gbps hot-spot traffic between FPGA 1 and 5 (inter-region) with and without channel bonding.

The intra-region experiment makes use of TL-TX1 and TL-TX3 in FPGA1 and FPGA4, respectively. The inter-region experiment makes use ofTL-TX2 in FPGA1 and TL-TX4 in FPGA5. Table 6.4 shows the wavelengths used.

Figure 6.14 shows that for the hierarchical all-to-all network, the accepted hot-spot traffic will keep decreasing as the bandwidth of background traffic increases due to the limited bandwidth of the single link between the hot spots. On the contrary, by using channel bonding for flexible bandwidth adjustment, it is possible to achieve up to ~1.77× improvement in accepted hot-spot traffic. Figure 6.15 shows that the reconfiguration of the links dedicated to certain clusters does not reduce but can increase the accepted background traffic by releasing the congestion caused by the hot-spot traffic.

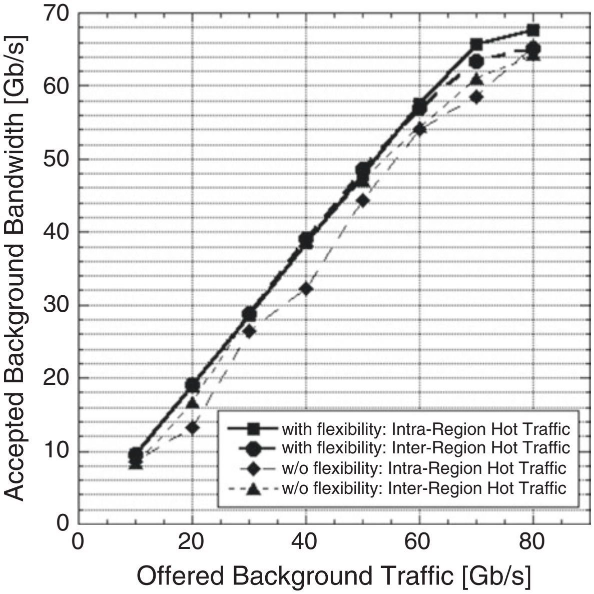


Fig. 6.15 Measured accepted background traffic bandwidth

6.8 Conclusions

The experiment testbeds and demonstrations carried out at UC Davis NGNS laboratories clearly demonstrate the validity and effectiveness of AWGR as an optical switching and routing fabric for optical circuit and packet switching. Certainly, the active LIONS solutions, as any other optical switch proposed in literature, cannot be cascaded since there are no buffers at the switch ports. This limit the use of these optical switches either in the core stage of folded Clos-like architectures or in direct connected topologies like flattened butterfly, torus, and Hyper X.

Among the proposed active LIONS solutions, TOKEN-LIONS and TONAK-LIONS are the one more effective since they eliminate the need of the expensive loopback buffer and wavelength converters, and they decentralize the control plane and tunable lasers (which are now located at the nodes' site) improving the overall scalability of the switch architecture. The key technologies required in the passive and active LIONS architectures (AWGRs, RSOAs, and tunable lasers) are commercially available and could be readily used, with the exception for fast tunable lasers which are still available only as research prototypes. However, whether optical switches can become techno-economically viable is still an open question that goes beyond the scope of this chapter.

We believe that next-generation data center and HPC switches will likely adopt optical technologies in a hybrid integration platform, with several electronic switches interconnected through very high-bandwidth and high-energy efficient optical interconnect architectures (like the ones we demonstrated in the above section on passive LIONS) to create very high-port count and multi-Tbps bandwidth switches.

References

1. S. Kamei et al., 64 × 64-channel uniform-loss and cyclic-frequency arrayed-waveguide grating router module. Electron. Lett. **39**(1), 83–84 (2003)
2. B. Glance, I.P. Kaminow, R.W. Wilson, Applications of the integrated waveguide grating router. J. Lightwave Technol. **12**(6), 957–962 (1994)
3. Ye, X., et al., *DOS – A scalable Optical Switch for Datacenters.* ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS), (2010)
4. R. Proietti et al., Scalable optical interconnect architecture using AWGR-based TONAK LION switch with limited number of wavelengths. J. Lightwave Technol. **31**(24), 4087–4097 (2013)
5. X. Ye et al., Buffering and flow control in optical switches for high performance computing. J. Opt. Commun. Netw. IEEE/OSA **3**(8), A59–A72 (2011)
6. R. Proietti et al., All-optical physical layer NACK in AWGR-based optical interconnects. IEEE Photon. Technol. Lett. **24**(5), 410–412 (2012)
7. R. Proietti et al., Scalable and distributed contention resolution in AWGR-based data center switches using RSOA-based optical mutual exclusion. IEEE J. Sel. Top. Quantum Electron. **19**(2), 3600111–3600111 (2013)
8. R. Proietti et al., An all-optical token technique enabling a fully-distributed control plane in AWGR-based optical interconnects. J. Lightwave Technol. **31**(3), 414–422 (2013)
9. Hubner, M., K. Paulsson, and J. Becker. *Parallel and Flexible Multiprocessor System-On-Chip for Adaptive Automotive Applications based on Xilinx MicroBlaze Soft-Cores*. in *Parallel and Distributed Processing Symposium, 2005. 19th IEEE International*. (2005)
10. Ye, X., et al. *DOS – A scalable Optical Switch for Datacenters*. in ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS). (2010)
11. Shin, H.C., et al., *Reflective semiconductor optical amplifier*. Google Patents, (2006)
12. Kim, H. and S. Chandrasekhar, *Dependence of coherent crosstalk penalty on the OSNR of the signal.* Optical Fiber Communication Conference (OFC). (2000)
13. Proietti, R., et al., *Fully-distributed Control Plane by All-Optical Token Technique in AWGR-based Optical Interconnects.* J. Lightwave Technol. (2012)
14. L. Buckman, L. Chen, K. Lau, Crosstalk penalty in all-optical distributed switching networks. IEEE Photon. Technol. Lett. **9**(2), 250–252 (1997)
15. R. Proietti et al., A scalable, low-latency, high-throughput, optical interconnect architecture based on arrayed waveguide grating routers. J. Lightwave Technol. **33**(4), 911–920 (2015)
16. R. Yu et al., A scalable silicon photonic chip-scale optical switch for high performance computing systems. Opt. Express **21**(26), 32655–32667 (2013)
17. P. Grani et al., Photonic interconnects for interposer-based 2.5D/3D integrated systems on a Chip, in *Proceedings of the Second International Symposium on Memory Systems*, (ACM, Alexandria, 2016), pp. 377–386

Chapter 7
Torus-Topology Data Center Networks with Hybrid Optoelectronic Routers

Ryo Takahashi and Ken-ichi Kitayama

7.1 Introduction

Current data center (DC) networks [1–3] rely on electrical packet switching (EPS), and so far their evolution has been supported by the advancement of CMOS ASIC and optical transceiver module technologies. However, these technologies have already become exhaustively advanced awaiting various difficulties for further improvement in the future [4]. Releasing the DCs from such traditional OEO-based paradigm would drastically improve the network performance. Thus, there is a need for new photonic networks with new topology and networking model [5–15], where best-effort services and highly reliable services can, respectively, be supported by optical packet switching (OPS) on a packet basis and optical circuit switching (OCS) on a flow basis.

However, the realization of the photonic network still requires great research efforts on various aspects from devices to networking with an overall consideration for the following issues.

R. Takahashi (✉)
Nippon Telegraph and Telephone Corporation (NTT), Device Technology Laboratories, 3-1, Morinosato Wakamiya, Atsugi-shi, Kanagawa 234-0198, Japan

National Institute of Information & Communications Technology (NICT), Network System Research Institute, 4-2-1, Nukui-Kitamachi, Koganei, Tokyo 184-8795, Japan
e-mail: t.ryo@nict.go.jp

K.-i. Kitayama
The Graduate School for the Creation of New Photonics Industries (GPI),
1955-1 Kurematsu-cho, Nishiku, Hamamatsu, Shizuoka 431-1202, Japan

National Institute of Information & Communications Technology (NICT), Network System Research Institute, 4-2-1, Nukui-Kitamachi, Koganei, Tokyo 184-8795, Japan

Professor Emeritus, Osaka University, Osaka, Japan
e-mail: kitayama@gpi.ac.jp

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
DOI 10.1007/978-3-319-61052-8_7

- *Networking with OPS and OCS*: These schemes possess complementary features and would best fit for different kinds of services. To cope with versatile services in DCs in a cost-effective way, both schemes should be efficiently unified on a single platform without relying on different types of switches and/or allocation of separate wavelengths.
- *Network topology*: It should be selected while considering OPS-related demands such as the difficulty in realizing high-speed optical switches with large port counts and the need for fast deterministic forwarding algorithm and effective contention resolution strategy in the absence of a viable optical random access memory, in addition to other general requirements of flexible scalability, robust redundancy, connectivity between any-to-any server, and easy controllability.
- *Node technology*: It requires high capabilities for both the optical and electrical processing of high-speed burst-mode optical packets. Making optimal use of both optics and electronics would enable these processes to be done with low power and low latency while keeping a high signal quality.

In this chapter, we present our approach [14, 15] for meeting the above requirements with a torus-topology photonic DC network, where the OPS, OCS, and novel virtual OCS (VOCS) schemes are all simultaneously supported on a unified platform and at the same wavelength. The node equipment called the hybrid optoelectronic router (HOPR) is currently being developed for the torus network to support the OPS/OCS/VOCS schemes and to handle 100-Gbps burst-mode optical packets with very low power consumption and latency. This chapter is subdivided into three sections. The first section explains what the torus topology is and discusses about why it has been selected. The second section presents the data/control planes of the torus DC network and explains the operation mechanisms of the hybrid OPS/OCS/VOCS transmission schemes, followed by an evaluation for their performance based on numerical simulation. The third section reviews HOPR's architecture, targeted specs, and the key enabling technologies.

7.2 Torus Topology

In an N-dimensional torus network $T(N,M)$ composed of M nodes, the address of each node can be simply represented by its coordinates $(x_1, x_2, .., x_i, \ldots, x_N)$, where $x_i \in \{0, 1, 2, \ldots, K_i - 1\}$ and $M = \prod_{i=1}^{N} K_i$, where K_i is the number of nodes for each dimension and a group of K_i nodes along each coordinate axis forms an individual ring network as shown in Fig. 7.1, and as a result each node connects 2 N neighboring nodes for bidirectional transmission. In the torus network, it is straightforward to place the nodes at the network coordinates.

The performance of a torus network strongly depends on its dimension N and node count M. For simplification, assuming the radix $K_i = K$ for all dimensions, then the total number of nodes and total number of links are given by K^N and $2NK^N$,

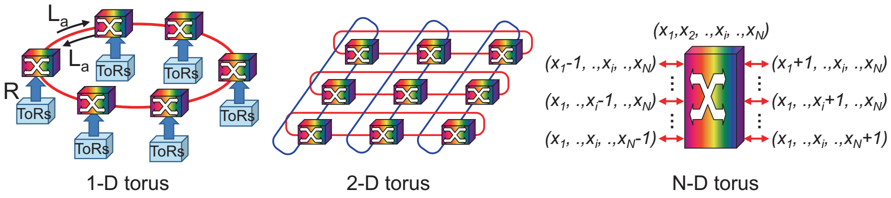


Fig. 7.1 Torus topology

Table 7.1 Comparisons of different dimensions

Dimension, N	1		2		3		4		6
Radix, K	4096		64		16		8		4
# total nodes, K^N	4096	=	64^2	=	16^3	=	8^4	=	4^6
# total links, $2NK^N$	8192		16,384		24,576		32,768		49,152
Average hop count, KN/4–1	1023		31		11		7		5
Critical input load, R_C [Gbps] (L_N = 100 Gbps)	0.2		12.5		50		100		200

respectively. The average hop count, which is defined by the number of nodes between source and destination nodes, is given by NK/4–1 and $N(K^2–1)/4$ K-1 for even and odd values of K, respectively. Assuming a constant data load R [Gbps] sent from a group of top-of-rack (ToR) switches to each node and uniform traffic distribution without packet contention, then the average traffic L_a present at each link between neighboring nodes is given by RK/8 and $R(K^2–1)/8$ K for even and odd values of K, respectively. And as it is limited by the link capacity L_N of the network ($L_a < L_N$), then we get $R < R_C = 8L_N/K$ and $8KL_N/(K^2–1)$ for even and odd values of K, respectively, where R_C is the critical input load that causes network overflow. Table 7.1 shows the basic network parameters that depend on its dimension while keeping the same sufficiently large number of nodes (4096).

To realize an OPS-based large-scale DC network, a higher-dimension torus network is promising for the reasons discussed below,

- *Redundancy*: Higher dimension torus NW provides a large number of alternative shortest routes with the same latency, enabling robust redundancy for supporting the continuous DC operation even when some node failures occur.
- *Scalability*: Current spine-leaf-type fabric networks need to deploy switches with large port counts in advance, whereas the installation of fiber links gets more complicated when adding new nodes. The torus NW is expandable simply by adding the new node in a plug-and-play manner without interrupting the already running NW, allowing effective small start and huge expandability.
- *Hop count*: Looking inside each electrical switch in the current DC, a close architecture with multiple switch ASICs is typically adopted [3], implying that the packet needs to go through at least seven switch ASICs between leaf switches. On the other hand, the average hop count gets reduced to as low as five hops in a

6-D torus NW composed of as many as 4096 nodes. This is advantageous for the OPS packets to keep high signal quality and low latency.

- *Optical switch*: The feasible optical switches for OPS use that can handle high-speed WDM optical packets with a switching time of less than 10 ns are currently limited to around 16 or 32 ports. Only low-radix switches are demanded with the torus topology, and thus a torus-based OPS network is currently realizable.
- *Forwarding and contention resolution*: To enable fast packet forwarding, the label processor requires a simple deterministic forwarding algorithm for deciding the output port. In torus networks, since all propagation delays between adjacent nodes are equal, the label processor can quickly decide the shortest route and the corresponding output port by only considering the destination address. Moreover, when contention occurs, multiple alternative routes are available for deflection routing. Thus, utilizing the link fibers as if they are optical buffers enables a large reduction of the packet loss probability without changing the end-to-end latency. In addition, to resolve contention incidents, the time needed for arbitration increases exponentially with the input port count, and thus it is very desirable to realize OPS-based networks with low-radix switching nodes.
- *Network scale and throughput*: With an input load R from ToR switches, Fig. 7.2a shows the average traffic L_a for torus networks of different scales and dimensions. Considering a given input load R, a torus network with higher dimension enables more reduction of the average traffic L_a of each link and thus enables the deployment of more nodes with a given link capacity. Figure 7.2b shows the critical input load R_C for a fixed link capacity L_N of 100 Gbps. Similarly with increasing the NW dimension, the max acceptable data rate from ToR switches increases and thus enhances the network throughput. In other words, for a given server count, each node can aggregate more servers, and the node count would accordingly be reduced.

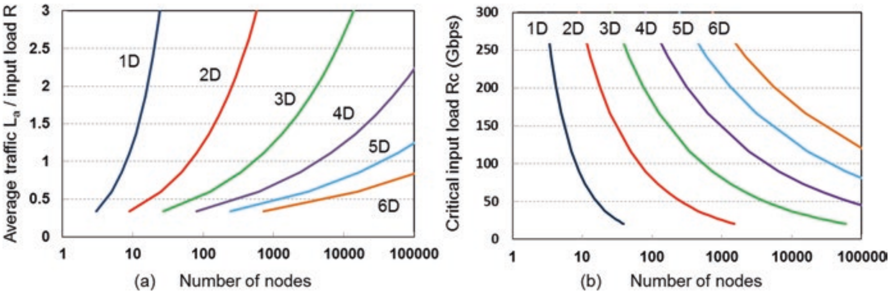


Fig. 7.2 (**a**) Ratios of the average traffic L_a and input load R and (**b**) critical input load R_C assuming the L_N = 100 Gbps versus DC scale for different torus dimensions

7.3 Torus Data Center Networks

Figure 7.3 illustrates the architecture of the torus DC network. Each node is composed of an aggregation switch part and optical packet switch part. Here, we refer to the combination of the optical packet switch and aggregation switch all together as the hybrid optoelectronic router (HOPR), which can be seen as a sort of label-switched router equipped with an Ethernet/OPS network interface. The torus network is controlled by a SDN (OpenFlow)-based control plane, which enables the OPS, OCS, and VOCS schemes to be all simultaneously supported on a unified data plane. The control and data planes are separately operated with Ethernet packets at 1.3-µm band and burst-mode optical packets at 1.5-µm band, respectively, while sharing the same fiber links as explained below.

- *Control plane (1.3-µm band)*: The network controller plays a key role in managing the entire DC network as if it is a very large L2 switch. To implement the MAC learning process, the controller gathers the ARP (address resolution protocol) packets from servers, updates the MAC address table that includes the port number and address of the HOPR unit to which each server is connected, and then distributes the table to all HOPRs. The controller's other vital role is controlling the OCS/VOCS flows and establishing their necessary paths. In preparation for a new path request, the controller gathers the traffic information of each HOPR unit at constant time intervals and correspondingly estimates the best path among HOPRs that would less affect the data traffic. Once an OCS path is requested, the controller quickly distributes the already prepared path setting table that includes the information of the input/output HOPRs' ports along the selected path.

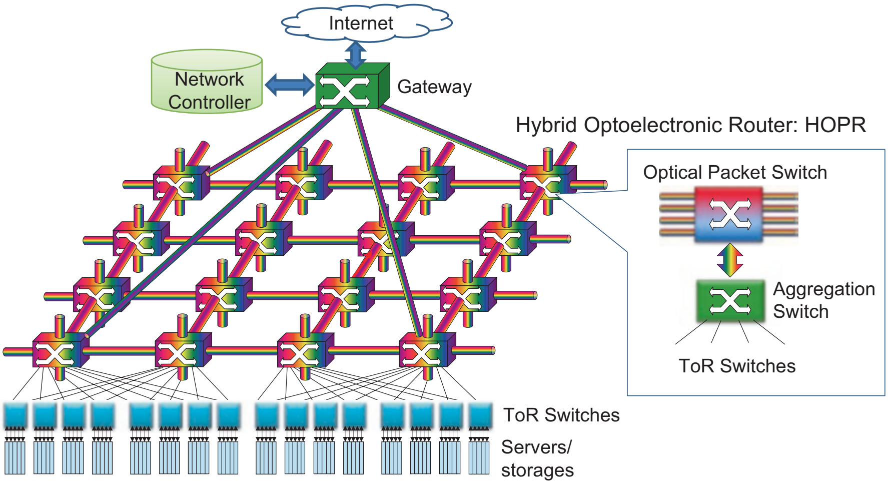


Fig. 7.3 Illustration of the torus-topology data center network

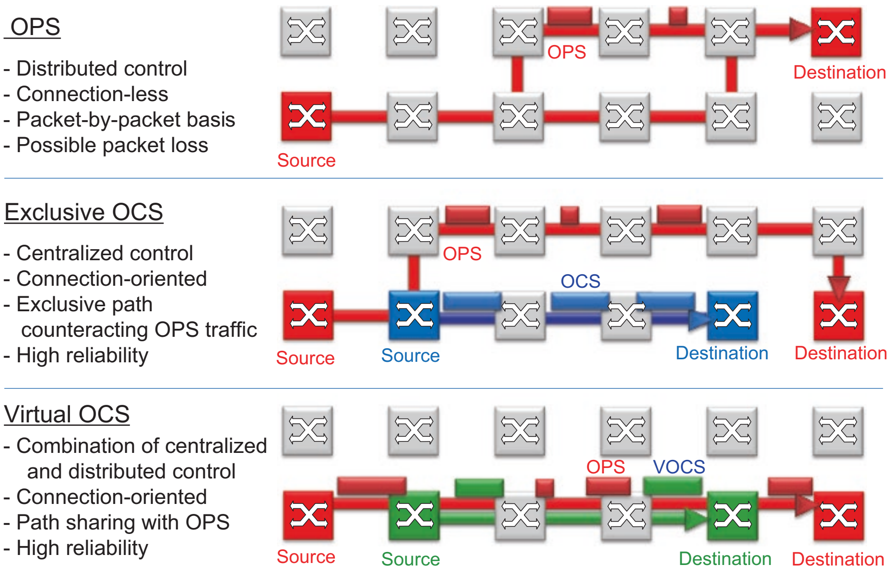


Fig. 7.4 OPS, OCS, and VOCS transmission schemes

- *Data plane (1.5-μm band)*: Ethernet signals from a group of ToR switches are sent to the attached aggregation switch, where the MAC frame is encapsulated with a fixed length label to form a high-speed burst-mode optical packet (100–400 Gbps). The optical packets are then forwarded up to their destinations via the optical packet switches. As illustrated in Fig. 7.4, the data format of the burst-mode optical packet, i.e., OPS packet, is also used for the OCS and VOCS cases. It is noteworthy that the three transmission schemes can be operated on a single data plane and at the same wavelength, instead of relying on separate switches/platforms and/or allocating them separate wavelengths. Wavelength resources are thus used for increasing the link capacity by WDM technology (see Fig. 7.8c). This enables us to cost-effectively construct and manage the DC for coping with a diversity of services. The details of the OPS, OCS, and VOCS schemes are described below. Hereafter, we refer to the transmitted packets by the name of their scheme, i.e., OPS/OCS/VOCS packets.

7.3.1 Optical Packet Switching

The OPS scheme relies on distributed control with label processors where packets are transferred on a packet-by-packet basis. The packet latency can be reduced with a connection-less protocol like the User Datagram Protocol (UDP), but the possibility of packet contention is unavoidable. The OPS mode is thus suitable for best-effort services or latency-sensitive applications.

Consider a packet arriving at the node $P(c_1, c_2, .., c_N)$ and by denoting the source and destination node of the packet by $S(s_1, s_2, .., s_N)$ and $D(d_1, d_2,.., d_N)$, respectively. Then, the packet should be forwarded to the destination along the vector $\overrightarrow{PD} = (\delta_1, \delta_2, \ldots, \delta_N)$, where by considering the bidirectional ring topology,

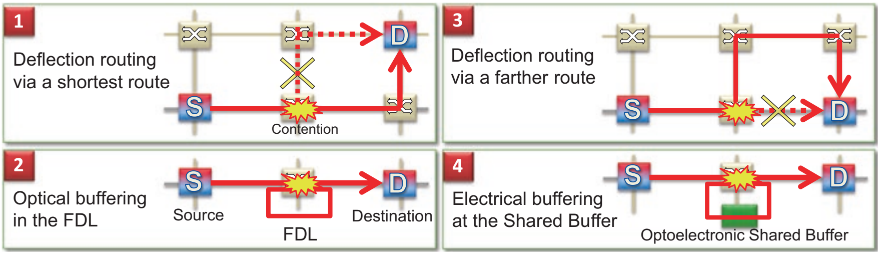
$$\delta_i = \min\{|d_i - c_i|, K_i - |d_i - c_i|\} \text{ for } i = 1, 2, \ldots, N.$$

In case of no contention, each hop decrements one of the elements δ_i of $\overrightarrow{PD}$ by 1, and finally after $\sum_{i=1}^{N} \delta_i$ hops, the packet arrives at the destination. It is noted that the packet latency is the same even when the vector elements are decremented in any order. Here, the label processor follows a deterministic rule in which the vector element with maximum value is selected by default, and just in case of packet contention, the nonzero elements are used in turn to provide an alternative output port for deflection routing. If some elements have the same maximum value, the element of the lowest dimension is selected.

Figure 7.5 shows the contention resolution strategy based on the maximum utilization of network links as optical buffers.

- The first choice: Deflection routing via the shortest alternative route in the case that more than two elements exist in the forwarding vector
- The second choice: Optical buffering through a fiber delay line in the case that the output ports for the shortest deflection routing are all occupied or the forwarding vector has only one nonzero element
- The third choice: Deflection routing via a farther route, i.e., not the shortest one, in case that the FDL is also occupied
- A final choice: Electrical buffering at the optoelectronic shared buffer

Following the mentioned forwarding algorithm and contention resolution strategy, all OPS packets are automatically forwarded to their destinations, theoretically speaking without packet loss. However packets might change their routes according to traffic conditions. In fact, packet loss occurs when the signal quality gets so degraded by undergoing many hops, such that the error correction procedure cannot manage to recover the error bits (label recognition error), or when the electrical buf-



1
Deflection routing via a shortest route
S
D
Contention
3
Deflection routing via a farther route
S
D
2
Optical buffering in the FDL
S
Source
D
Destination
FDL
4
Electrical buffering at the Shared Buffer
S
D
Optoelectronic Shared Buffer

Fig. 7.5 Contention resolution strategy

fer memory overflows. As explained in Fig. 7.2, a higher network dimension increases the critical input load and drastically decreases packet loss.

7.3.2 Optical Circuit Switching

The OCS scheme is connection oriented, relying on a centralized network controller. When an OCS path is requested, the controller quickly reserves the optical switches at the HOPR units located along the path to establish the connection. The resulting path setting time is just several tens of microseconds, which is quite faster than a few tens of milliseconds in case of MEMS switches. The OCS mode thus provides an exclusive path that guarantees constant latency and high reliability without packet loss. However, as shown in Fig. 7.4, the ordinary OPS packets cannot go through the exclusive OCS path and should be deflected via a farther route instead. Thus increasing the number of OCS paths strongly counteracts the ordinary OPS packets and degrades the packet loss probability of the OPS packets. In addition, different from the collision of OPS-OPS packets, the OPS-OCS collision sometimes cannot be resolved by the abovementioned strategy. This is because as long as the OCS path remains established, the contented OPS packet repeats the same action for contention resolution, such as circulating in an FDL, and therefore the hop count keeps increasing and results in a higher packet loss probability. A different strategy called "back-last deflection routing" is introduced to solve this issue, where the output port of the router from which the packet came at the last hop is given the lowest priority [17]. Figure 7.6 shows simulation results for the packet loss probability of ordinary OPS packets in a 6-D torus network comprising 4096 HOPRs with a configuration shown in Fig. 7.8b. For a pure OPS mode without any OCS paths, even when the input is 160 Gbps which is close to the critical input load, the packet loss rate remains well below 10^{-3}. Whereas by increasing the number of OCS paths, the OPS packet loss probability gradually degrades. For 885 OCS paths, the OPS traffic is severely counteracted where the packet loss probability cannot anymore go

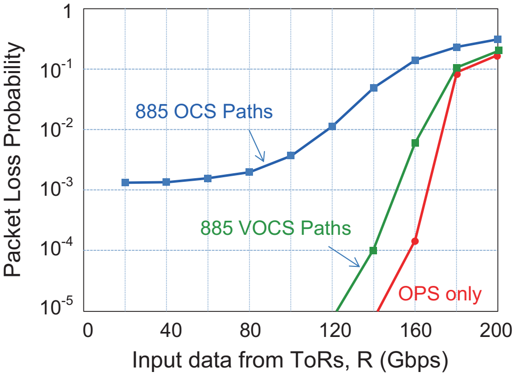


Fig. 7.6 Packet loss probability of OPS packets for OPS, OCS, and VOCS schemes

less than 10^{-3}. However, with a low number of paths, the OCS scheme is powerful for transmitting extremely large-capacity data at a high data rate allowing for full exploitation of the link capacity as in case of VMs cluster migration or fast backup of large-capacity storage data. The details of calculation conditions and other results can be found in Ref. [17].

7.3.3 *Virtual Optical Circuit Switching*

When data is transmitted in cases similar to individual VM migration or single video/file transfer, the HOPR-HOPR link capacity is not usually fully exploited because the link capacity is much higher than the HOPR-ToR link capacity. It is then very inefficient to have sparse transmission of OCS packets in an exclusively reserved link. Thus, a novel transmission scheme is highly demanded to establish a dedicated path that can still be shared with OPS packets without a counter effect.

In conventional EPS networks that utilize electrical switches, a scheme that allows transmission sharing over dedicated paths has been already realized. Such scheme is called the *virtual circuit switching (VCS)* scheme, in which electrical switches operate in a store-and-forward manner relying on electrical buffering. Therefore, even when contention occurs, packets can keep going through the dedicated path by being stored at each node. On the other hand, the optical packets have to be basically forwarded in the optical domain in a cut-through manner, because a viable optical random access memory is still missing. Here, by using an FDL for optical buffering instead of electrical buffering, we realize such a dedicated path and refer to it as *virtual optical circuit switching (VOCS)* [15, 16].

Figure 7.7 illustrates the operation of VOCS. When a VOCS path is requested, the network controller distributes the path setting table in the same way as for the OCS mode. In conventional OCS, the optical switches are exclusively reserved, whereas

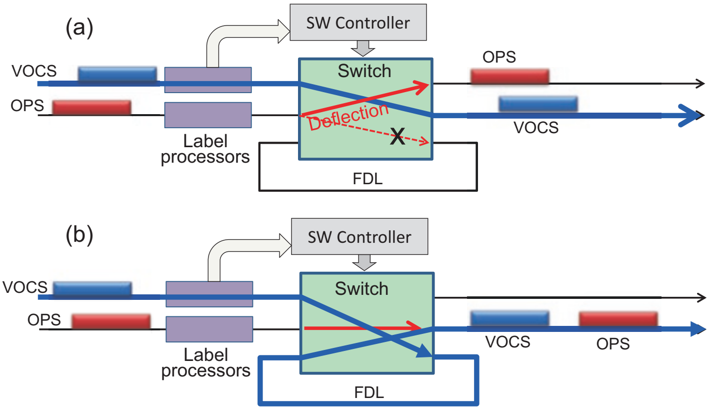


Fig. 7.7 Illustration for the VOCS operation (**a**) without and (**b**) with contention

in case of VOCS, the FDLs and not the optical switches are exclusively reserved. When a VOCS packet comes in, the label processor recognizes the virtual path identifier, not the destination address. Then according to the path setting table already present at that node, the packet is forwarded to the desired output port along the intended path. However, if the VOCS packet arrives late and collides with an OPS packet, the OPS packet is allowed to keep going through the desired output port, whereas the label processor reserves that output port for the VOCS packet from that time on. Simultaneously, the VOCS packet is forwarded to the already-reserved FDL, and after the OPS packet passes away, the use of the output port is permitted only for the VOCS packet. Therefore, the packet can be always forwarded to the desired output port after returning to the optical switch coming out of the FDL.

As shown in Fig. 7.6, unlike the conventional OCS scheme, the novel VOCS hardly affects the packet loss of ordinary OPS packets even with as many as 885 paths being established [16]. Relying on a combination of centralized control with the network controller and distributed control based on label processors, the VOCS can provide a reliable dedicated path without packet deflection or loss for VOCS packets and very importantly without affecting ordinary OPS packets. Thus, many VOCS paths can be simultaneously established; however, the number of VOCS paths crossing a HOPR unit is limited by the number of FDLs at that unit, as one FDL is demanded for each path. The VOCS scheme is very beneficial for handling large-capacity data without preventing the OPS packets from sharing the same network resources.

7.4 Hybrid Optoelectronic Router (HOPR)

The performance of the torus DC network depends on HOPR’s specifications, and hence a large reduction of latency, power consumption, and signal quality degradation is strongly demanded in a highly functional HOPR unit. Compared to the previous HOPR prototype completed in 2009 [18], an upgraded HOPR prototype is being developed with highly enhanced specifications. In this section, we review HOPR’s architecture and enabling technologies and reveal its new specs.

Figure 7.8(a) shows HOPR’s basic architecture that employs an 8x8 optical switch for a 3-D torus topology. An incoming optical packet is basically forwarded in a cut-through manner through the optical switch to the output port decided by the label processor or network controller without OE conversion, thus enabling very low latency. An FDL and electrical shared buffer are available for contention resolution to reduce the packet loss probability as explained above. In this case, since a 3-D torus network requires a switching port count of 6, the shared buffer has only one set of 100-Gbps (25 Gbps × 4 λs) burst-mode Tx/Rx. Thus the maximum input load sent from ToRs to the buffer and the link capacity between adjacent HOPRs are also both limited to 100 Gbps. To enable a 6-D torus NW, a 16x16 optical switch is required as shown in Fig. 7.8(b), where the link capacity is still 100 Gbps, but the maximum input load is enhanced up to 200 Gbps, and therefore more server traffic

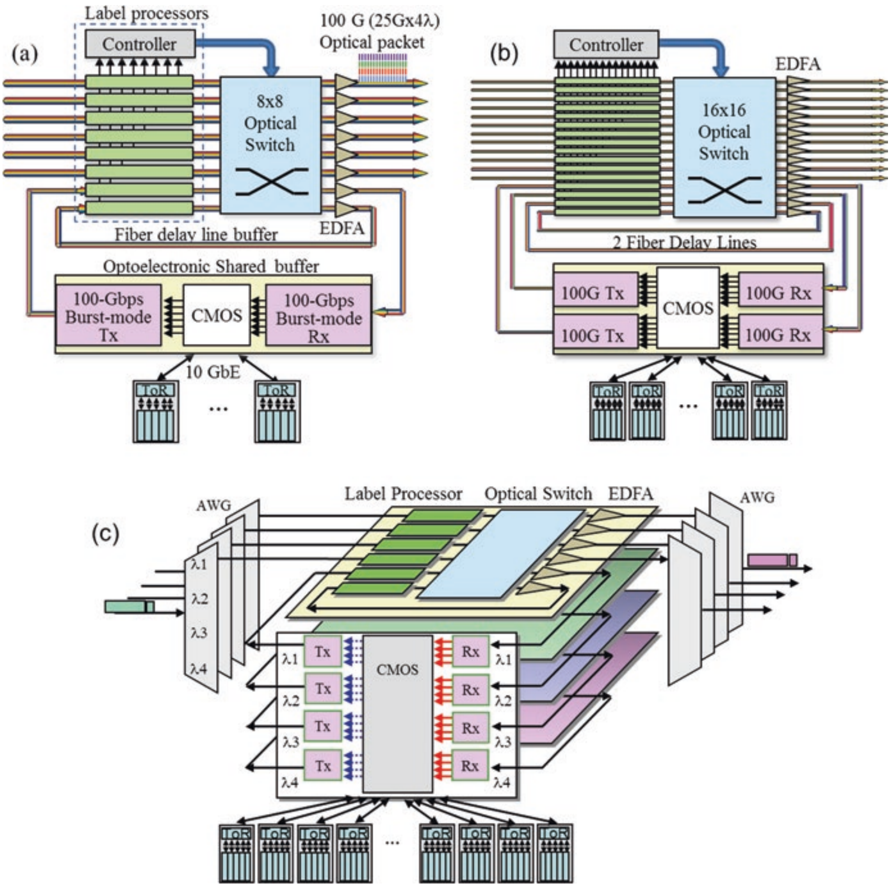


Fig. 7.8 Different HOPR configurations for (**a**) 3-D torus, (**b**) 6-D torus, and (**c**) scaling up the link capacity by using WDM

can be aggregated by the shared buffer. Moreover, as shown in Fig. 7.8(c), with the WDM configuration, where each layer handles waveband packets, e. g., 25 Gbps × 4 λs, both the link capacity and the maximum input load can be easily increased as the product of the number of wavelength layers and the data rate of packets generated by each transmitter to allow achieving 400 ~ 1000 Gbps.

The upgraded HOPR prototype is still under construction, where new devices and subsystems are being particularly developed to meet HOPR's new target specs that are listed below.

- *Throughput*: 1280 Gbps. Six input/output ports for 100-Gbps optical packets and four input/output ports for 10-GbE connection.
- *Power consumption*: 110 W. The breakdown is ~40 W for the optical packet switch part including eight label processors and an 8x8 optical switch and ~70 W for the aggregation switch part (optoelectronic shared buffer). The power con-

sumed by control plane, cooling, GUI for external computer control, 10-GbE optical transceiver modules, and DC voltage conversions is all included.

- *Latency*: 140 ns. The breakdown is ~60 ns for transmission delay via optical components including the optical switch and EDFA and ~80 ns for the label processor which is dominated by arbitration time for contention resolution.

On the other hand, current electrical switches have been significantly improved where the state-of-the-art CMOS ASIC technology allows the monolithic integration of all necessary components on a single chip. However, electrical switches require a NIC card and optical transceiver module at each switching port. Considering that a CFP 100-GbE optical transceiver module consumes 6–20 W for different transmission distances, the value of 5 W/100-G port for the optical packet switch part is obviously lower than the power of just a single CFP module. It is also worth mentioning that the power of the optical packet switch almost remains constant even if the packet data rate is increased to 400 Gbps in the future. Moreover, the power consumed by the shared buffer can be further reduced if instead of combining an FPGA and discrete electronic components, the state-of-the-art integrated ASIC technologies are introduced in a way similar to the electrical switches.

To achieve such large reduction of power consumption and latency, a set of novel optical devices and subsystems have been developed as summarized in Fig. 7.9. Here we quickly review HOPR's enabling technologies, and for more details please check the corresponding references.

- *Packet format*: EPS handles stream format signals in which scramble bits fill in the inter-frame gaps between packets to keep the clock recovery going on. Differently for the OPS case, processing burst-mode optical packets demands individual fast clock recovery for each arriving packet. Realizing burst-mode clock recovery is usually done by adding long preamble bits in front of the packets [19], or otherwise by considering an analog phase picking method [20]. However, long preambles increase the overhead and reduce the throughput. Alternatively, the novel mechanisms of trigger pulse generation [22] in label processors and optical clock generation [27] in a shared buffer (Fig. 7.9) allow handling burst-mode optical packets without any preambles.

  In the prototype, we adopt an in-band 32-bit label (header) modulated at 25 Gbps similar to the packet payload, where the first 16 bits include the destination address, broadcast flag, virtual path identifier, checksum bits, etc., and the label processor recognizes them to decide on packet forwarding, whereas the other 16 bits that include the source address/port and packet ID are used at the shared buffer to check the received packets.
- *Optical label processor*: The label processor plays many important roles including arbitration in case of contention, deciding of the output port based on the forwarding table, setting the OCS/VOCS paths, and controlling the optical switch. To enable easy label recognition by a CMOS processor (FPGA), an optoelectronic integrated circuit (OEIC) called the *optically clocked transistor array (OCTA)* [21] performs a bit-by-bit serial-to-parallel conversion (SPC) only for

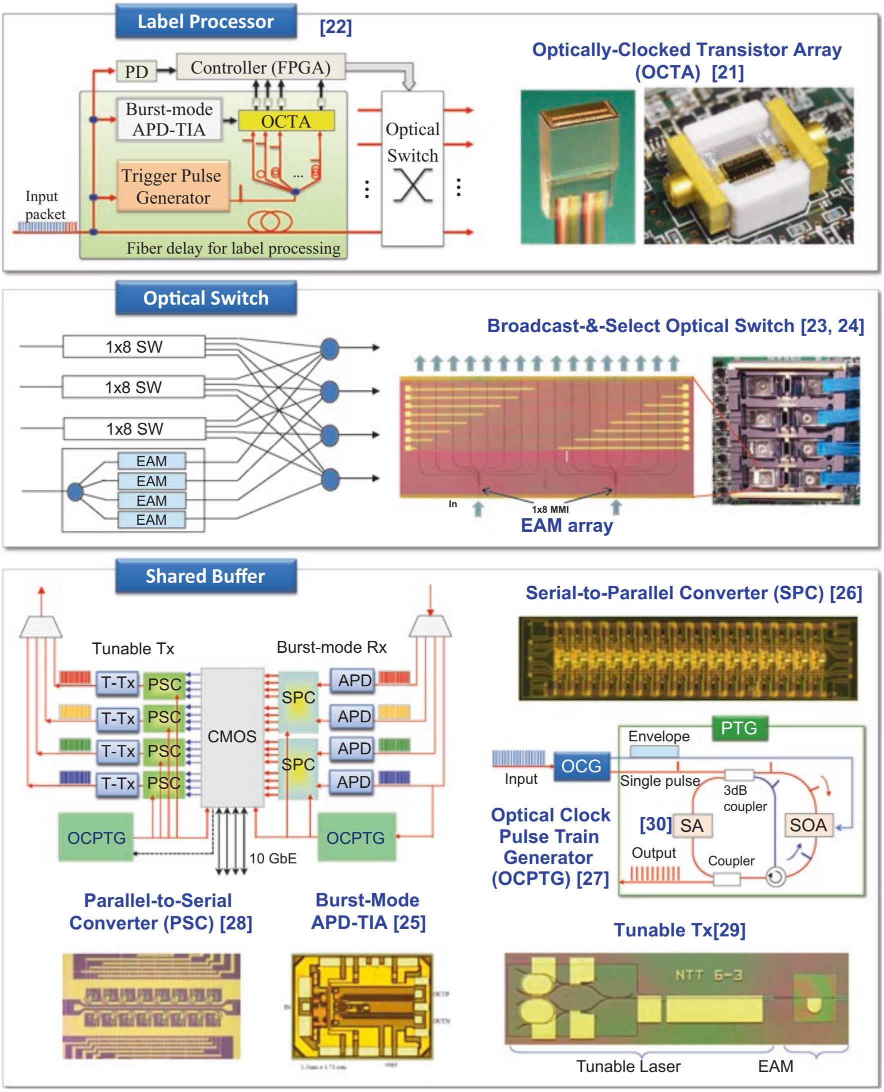


Fig. 7.9 Summary for the enabling technologies of the hybrid optoelectronic router (HOPR)

the label's bits with a very low power consumption of about 5 mW/ch and high-speed operation of up to 65 Gbps. The OEIC chip is directly attached to a glass block of fiber array without lenses. This enables a drastic reduction in OEIC packaging cost, besides resulting in a very compact package with easy and quick alignment. Moreover, the OE conversion efficiency is almost doubled with the backside irradiation.

The key point for further reduction of power consumption is limiting power dissipation only for the short label duration when a packet is received and not otherwise. High-speed electronic devices and EDFAs demand continuous power supply even during packet absence, resulting in a wasteful high power consumption. Here, an optical trigger pulse generator (TPG) is driven by a narrow current pulse (~1 ns) from a novel OEIC driver only when a packet comes in and thus enables a drastic reduction of power consumption from 3 W previously down to as low as 120 mW [22].

- *Optical switch:* The optical switch needs to be transparent with respect to bit-rate and packet format, besides fulfilling other basic requirements such as fast switching, low power, wavelength/polarization/temperature insensitivity, high extinction ratio, low cross talk, ease of controllability, and compactness.

  A variety of optical switches have been demonstrated so far, including the matrix switch, phased array switch, wavelength routing switch, and broadcast-and-select (B&S) switch. Among them, the B&S switch is promising because it can meet most of the requirements to a good extent, in addition to possessing the capability of unicast/multicast/broadcast forwarding. However owing to its inherently large splitting loss, its port count scalability is limited. To compensate for this loss, usually semiconductor optical amplifiers (SOAs) are used as optical gates. However, the SOA severely degrades the signal quality due to its large ASE noise, pattern dependence, and nonlinear effects. Moreover, it requires a high-speed and high-current driver that is power hungry. To solve these issues, we have developed a compact 8 × 8 B&S optical switch based on monolithically integrated electro-absorption modulator (EAM) gate arrays [23, 24] with a switching time of less than 10 ns, an extinction ratio of more than 40 dB, together with extremely low power consumption of less than 3 W, data format transparency, and polarization/wavelength/temperature insensitivity.
- *Optoelectronic shared buffer*: The shared buffer relies on a CMOS core (FPGA) for performing the necessary complex signal processing functions such as data format conversion between 10-GbEther signals and 100-Gbps burst-mode optical packets, insertion/erasure of labels, performing L2 switching among the buffer-attached ToRs, buffering for contention resolution, queuing for QoS, and forwarding error correction. It also includes a burst-mode receiver/transmitter (Rx/Tx). The Rx comprises four 25-Gbps burst-mode APD-TIAs [25] and two OEIC-based SPCs [26] driven by an optical clock pulse train generator (OCPTG) [27], where 25 Gbps × 4 λs are converted to 1.56 Gbps × 64. Conversely, when an optical packet is retrieved, the 64 parallel electrical signals are output from the FPGA, converted to 25 Gbps × 4 with similar OEIC-based parallel-to-serial converters (PSCs) [28], and finally transmitted as a 100-Gbps optical packet by four tunable transmitters (T-Tx) [29].

  The OCPTG is a critical device for generating optical clock pulses without demanding preamble bits. When an optical packet comes in, the optical clock pulse generator (OCG) generates a single short optical pulse and an electrical signal taking the form of the packet envelope. Both signals are fed into the optical pulse train generator (PTG), and as long as the SOA remains activated by the packet envelope signal, optical clock pulses are periodically output from the PTG throughout the whole packet duration. To enable energy-efficient and stable generation of the clock pulses, a combination of an SOA and spin-polarized saturable absorber [30] is used. With such combination, the SOA's low-power ASE noise is cut off, and when the input power undergoes some fluctuation, the output power automatically converges to a constant value by a negative feedback mechanism. This self-stabilization mechanism completely eliminates the demand for external stabilizing equipment, enabling a drastic reduction of OCPTG's size and power consumption.

7.5 Perspectives and Research Directions

In this chapter, the hybrid OPS/OCS/VOCS torus-topology DC network based on HOPR has been presented. Unlike conventional EPS that adopts a store-and-forward mechanism, the absence of a viable optical RAM makes the cut-through mechanism in the optical domain the only feasible option for the OPS case and implies the necessity of a fast forwarding algorithm for label processing. In addition, a reliable strategy for contention resolution is pivotal for reducing the packet loss probability, where it is useful to consider all the network links as available optical buffers. A high-dimensional torus topology meets these OPS requirements due to the regularity of its node sequence.

To cope with the wide diversity of services demanded by the DC, other than the OPS-based services, the OCS scheme is indispensable as it enables a dedicated path for highly reliable transmission of large-capacity data flows without packet loss. The important point is how both OPS and OCS are unified on a single platform and at the same wavelength to construct a cost-efficient DC. The VOCS scheme enables the coexistence of OPS and OCS and further enhances the bandwidth utilization.

The control plane is essential for managing the whole network, and it should be operated separately from the data plane. However sharing the same platform between the data and control planes is desirable. Other important aspects still need to be solved such as issues related to the OAM functions and an effective broadcast algorithm in the optical domain. Moreover to achieve higher reliability and low latency for OPS packets, a new communication protocol that is a midway approach between the TCP and UDP may also be necessary.

On the other hand, innovative node hardware that can handle burst-mode optical packets is indispensable to meet a long set of demands including high-speed operation, high throughput, low power consumption, low latency, and maintained high signal quality, in addition to providing high functionalities such as QoS, FEC, data format conversion, contention resolution, etc.

To meet the demand for reducing the power consumption, a novel approach is required, where power is dissipated only when a packet is received and not otherwise. Dissipation time is even further reduced to a minimum as, for example, in the label processor case where it takes place only for the label duration. The latency of the node is dominated by the arbitration time for contention resolution, and thus a fast arbitration algorithm is required. To maintain high signal quality over multiple hops, EDFAs and not SOAs should be used along the main data stream. This minimizes unwanted signal degradation due to ASE accumulation and nonlinear effects. Whereas the SOA attractive features of compactness and fast operation make them suitable for optical signal processing such as in label processors or optical clock generators.

Photonic networks based on OPS are promising candidates for future DC networks, and despite the need for further research efforts from the device level to the network level in order to increase their maturity, recent developments have been revealing the high potential of OPS networks that allows them to significantly surpass the traditional EPS-based networks.

Acknowledgment This work has been funded by the National Institute of Information and Communications Technology (NICT) R&D program, Basic Technologies for High-Performance Hybrid Optoelectronic Router (2011–2016).

References

1. Cisco Global Cloud Index: Forecast and Methodology, 2013–2018, Cisco White Paper
2. A data center fabric is critical to a next-generation unified data center, in Cisco White Paper, (2011)
3. Arjun Singh et al., Jupiter rising: A decade of clos topologies and centralized control in google's datacenter network, in proc. of ACM SIGCOMM2015, (2015), pp.183–197
4. A. Ghiasi, Large data centers interconnect bottlenecks. Opt. Express **23**(3), 2085–2090 (2015)
5. D.T. Neilson, Photonics for switching and routing. IEEE J. Sel. Top. Quantum Electron. **12**(4), 669–678 (2006)
6. W. Zhang, H. Wang, K. Bergman, Next-generation optically-interconnected high-performance data centers. J. Lightwave Technol. **30**(24), 3836–3844 (2012)
7. C. Kachris, K. Kanonakis, I. Tomkos, Optical interconnection networks in data centers: Recent trends and future challenges. IEEE Commun. Mag. **51**(9), 39–45 (2013)
8. J. Gripp, J. E. Simsarian, J. D. LeGrange, P. Bernasconi, and D. T. Neilson, Photonic tera-bit routers: The IRIS project, in Optical Fiber Communication Conf. (OFC), paper OThP3 (2010)
9. N. Farrington, G. Porter, S. Radhakrishnan, H. Bazzaz, V.Subramanya, Y. Fainman, G. Pa-pen, and A. Vahdat, Helios: A hybrid electrical/optical switch architecture for modular data centers, in Proc. ACM SIGCOMM, (2010), pp. 339–350
10. Y. Yin, R. Proietti, X. Ye, C.J. Nitta, V. Akella, S.J.B. Yoo, LIONS: An AWGR-based low latency optical switch for high performance computing and data centers. IEEE J. Sel. Top. Quantum Electron. **19**(2), 3600409 (2013)
11. COSIGN: Combining Optics and SDN In next Generation data center Networks, Tech. Rep. [Online]. Available: http://www.fp7-cosign.eu/
12. H. Furukawa et al., Development of optical packet and circuit integrated ring network testbed. Opt. Express **19**(26), B242–B250 (2011)
13. M. A. Mestre, G. de Valicourt, P. Jennevé, H. Mardoyan, S. Bigo, Y. Pointurier, Opti-cal Slot Switching-Based Datacenters With Elastic Burst-Mode Coherent Transponders, in European Conf. on Optical Communication (ECOC), paper Th.2.2.3 (2014)
14. K. Kitayama, Y. Huang, Y. Yoshida, R. Takahashi, T. Segawa, S. Ibrahim, T. Nakahara, Y. Suzaki, M. Hayashitani, Y. Hasegawa, Y. Mizukoshi, A. Hiramatsu, Torus-topology data center network based on optical packet/agile circuit switching with intelligent flow management. J. Lightwave Technol. **33**(5), 1063–1071 (2015)

15. R. Takahashi, T. Segawa, S. Iblahim, T. Nakahara, H. Ishikawa, A. Hiramatsu, Y. Huang, K. Kitayama, Torus data center network with smart flow control enabled by hybrid optoelectronic routers. OSA/IEEE J. of Optical Communications and Networking **7**(12), 141–152 (2015)
16. Y. Huang, Y. Yoshida, S. Ibrahim, R. Takahashi, A. Hiramatsu, K. Kitayama, Novel virtual OCS in OPS data center networks. Photon Netw. Commun. **31**(3), 448–456 (2016)
17. Y. Huang, Y. Yoshida, K. Kitayama, S. Ibrahim, R. Takahashi, A. Hiramatsu, OPS/agile OCS data center network with flow management. OSA/IEEE J. of Optical Communications and Networking **7**(12), 1109–1119 (2015)
18. R. Takahashi, T. Nakahara, Y. Suzaki, T. Segawa, H. Ishikawa, and S. Ibrahim, Recent progress on hybrid optoelectronic router for future energy-efficient optical packet switched networks, in Photonics in Switching, (2012)
19. A. Rylyakov, J. Proesel, S. Rylov, B. G. Lee, J. Bulzacchelli, A. Ardey, C. Schow, and M. Meghelli, A 25 Gb/s burst-mode receiver for low latency photonic switch network, in Optical Fiber Communication Conf. (OFC), paper W3D.2 (2015)
20. R. Yu, R. Proietti, S. Yin, J. Kurumida, S.J.B. Yoo, A 10-Gbps BM-CDR circuit with synchronous data output for optical networks. IEEE Photon. Technol. Lett. **25**(5), 508–511 (2013)
21. S. Ibrahim, H. Ishikawa, T. Nakahara, Y. Suzaki, R. Takahashi, A novel optoelectronic 32-bit serial-to-parallel converter for 25-Gbps optical label processing. IEICE Trans. Electron. **E97.C**(7), 773–780 (July 2014)
22. S. Ibrahim, T. Nakahara, H. Ishikawa, R. Takahashi, Burst-mode optical label processor with ultralow power consumption. OSA Optics Exp. **24**(7), 6985–6995 (2016)
23. T. Segawa, S. Ibrahim, T. Nakahara, Y. Muranaka, R. Takahashi, Low-power optical packet switching for 100-Gbps burst optical packets with a label processor and 8×8 optical switch. IEEE J. Lightwave Technol. **34**(8), 1844–1850 (2016)
24. Y. Muranaka, T. Segawa, Y. Ogiso, T. Fujii, R. Takahashi, Performance im-provement of an EAM-based broadcast-and-select optical switch. IEEE Photonics J. **8**(2) (2016)
25. M. Nada, M. Nakamura, H. Matsuzaki, 25-Gbit/s burst-mode optical receiver using high-speed avalanche photodiode for 100-Gbit/s optical packet switching. Opt. Express **22**(1), 443–449 (2014)
26. S. Ibrahim, H. Ishikawa, T. Nakahara, R. Takahashi, A novel optoelectronic serial-to-parallel converter for 25-Gbps optical packets. Opt. Exp. **22**(1), 157–165 (2014)
27. T. Nakahara, R. Takahashi, Self-stabilizing optical clock pulse-train generator using SOA and saturable absorber for asynchronous optical packet processing. Opt. Express **21**(9), 10712–10719 (2013)
28. H. Ishikawa, T. Nakahara, H. Sugiyama, R. Takahashi, A parallel-to-serial converter based on a differentially operated optically clocked transistor array. IEICE Electron. Ex-press **10**(20), 20130709 (2013)
29. T. Segawa, W. Kobayashi, T. Sato, S. Matsuo, R. Iga, R. Takahashi, A flat-output widely tunable laser based on parallel-ring resonator integrated with EA modulator. Opt. Express **20**(26), B485–B492 (2012)
30. R. Takahashi, T. Yasui, J.-K. Seo, H. Suzuki, Ultrafast all-optical serial-to-parallel converters based on spin-polarized surface-normal optical switches. IEEE J. Sel. Top. Quantum Electron. **13**(1), 92–103 (2007)

Chapter 8
LIGHTNESS: All-Optical SDN-enabled Intra-DCN with Optical Circuit and Packet Switching

George M. Saridis, Alejandro Aguado, Yan Yan, Wang Miao, Nicola Calabretta, Georgios Zervas, and Dimitra Simeonidou

8.1 Introduction

In this book chapter, a flat all-optical intra-data center network architecture is introduced and validated throughout various experimental demonstrations. The architecture, which combines novel photonic switching technologies with a fully SDN-enabled control plane, aims at delivering scalable, flexible, low-latency, and high-capacity interconnection on demand. The architecture and technology described in this chapter is built under the EU FP7 project LIGHTNESS [1–3], containing both data and control plane state-of-the-art features.

The chapter begins with the "LIGHTNESS" architecture, which provides an analytical view of the proposed data plane design while also offering an insight into the control plane architecture. The overall architecture targets for utilizing hybrid OCS/OPS principals interchangeably in all parts and levels of the DCN, from the server interface cards to the all-optical top of the rack (ToR) switch and top of the cluster (ToC) switches. On top of it, a unified SDN controller is in charge of network resource allocation, decision making, and command forwarding, making the data plane fully programmable.

The following section is about the LIGHTNESS technology enablers and the hybrid OPS/OCS interconnect. It explains the technologies that were developed and

G.M. Saridis (✉) • A. Aguado • Y. Yan • D. Simeonidou
High Performance Networks Group, Department of Electrical & Electronic Engineering,
University of Bristol, Woodland Road, Bristol BS8 1UB, UK
e-mail: george.saridis@bristol.ac.uk

W. Miao • N. Calabretta
Electro-Optical Communication Systems, Institute for Photonic Integration (IPI), Department of Electrical Engineering, Technical University of Eindhoven, Eindhoven, The Netherlands

G. Zervas
Department of Electronic & Electrical Engineering, University College London,
Torrington Place, London WC1E 7JE, UK

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
DOI 10.1007/978-3-319-61052-8_8

used in the LIGHTNESS framework. It includes the design and functionality of the FPGA-based optoelectrical network interfaces, the optical ToR switch, as well as the OCS and OPS switching technologies.

Later, “experimental demonstration and evaluation” section reports the demonstration of SDN and virtualization-based capabilities (such as monitoring and file transfer or database migration) entirely integrated with an advanced all-optical physical layer. The performance of the above network is experimentally evaluated in terms of BER, end-to-end latency, and control plane functionality.

In the end, in the “Discussions” section, an overview of the proposed architecture is provided, including the pros and cons of such a design, as well as further suggestions for future work.

8.2 LIGHTNESS Data Plane Architecture

As shown in Fig. 8.1, servers at each rack are interconnected to the hybrid OCS/OPS DCN data plane via an optical ToR switch. The optical ToR switch can be implemented in many ways, as a passive optical element, such as an arrayed waveguide grating (AWG), a routing-AWG (R-AWG), or an active optical switch, such as a wavelength/spectrum selective switch (WSS/SSS) or a fiber/space switch. The functionality of the classical electrical ToR switch is moved by a large extent toward the advanced network interface cards (NICs), interfacing each server. Each NIC interface performs traffic aggregation and application-aware classification of data flows to either short- or long-lived ones. This offers an increased degree of programmability and dynamicity that is essential for modern intra-DCN capabilities.

Scalability is yet another objective of LIGHTNESS, and it is realized through the deployment of architecture-on-demand (AoD)-based [4] large-port count fiber switches (OCS) and numerous nanosecond fast switches (OPS) interconnecting racks in the DC, which allows for extending the number of input/output wavelengths between ToRs (WDM scaling) when the port count of OPS or OCS switches becomes a limitation. For higher scalability, the creation of clusters made of fixed number of racks could be an alternative, e.g., the top-of-the-cluster AoD-based switch of each cluster can be interconnected with other clusters through a large port-count fiber switch.

Based on the AoD design shown in Fig. 8.1, OCS switches support the interconnection of both short-lived and long-lived traffic flows within and between clusters of the DCN. In addition, for the case of the short-lived traffic flows, OPS subsystems can efficiently fill in the gaps of the OCS by loosening interconnection demands for the OCS with an optical packet-based approach, which offers finer network granularity while allowing the sharing of the same WDM channels among different servers. Additional links directly interconnecting servers between them can be an alternative approach (as shown in Fig. 8.1); even if it is limited by the port number, the NIC is able to support.

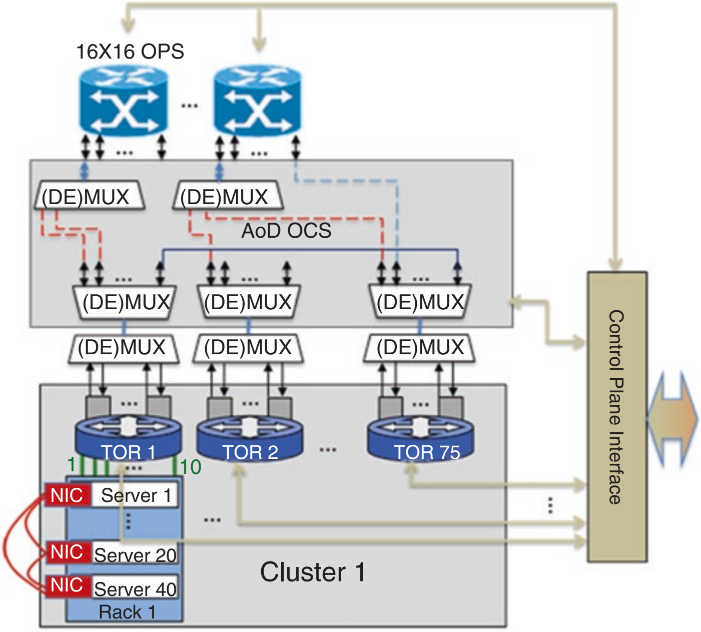


Fig. 8.1 Cluster-level LIGHTNESS AoD-based architecture

The combination of OPS and AoD-enabled OCS modules makes it possible to switch traffic in all three optical dimensions, namely, space, frequency, and time, as well as to provide a range of additional capabilities on demand. The reconfigurable optical backplane provides flexible connectivity for the AoD node. Based on the AoD-based OCS concept, the LIGHTNESS network shows the following advantages:

- Flattened network infrastructure providing the on-demand bandwidth allocation and low latency which are desirable for next generation DCN.
- Multi-granularity configuration: fiber switching, spectrum (or a single wavelength) switching, and subwavelength switching (optical packets).
- Fully reconfigurable and programmable connectivity: each NIC transponder can be directed to either OCS or OPS for transmitting information among servers.
- According to the traffic demand, OCS/OPS or both can be selected.

The proposed design for intra-cluster DCN is illustrated in Fig. 8.1. In this case, one cluster consists of several racks of servers. A ToR switch is used for interconnecting the servers in one rack to the AoD cluster backplane. Each ToR switch provides ten channels with 10Gbps/channel capacity, and each channel can support either OPS or OCS transmission, which is again directed by the ToR switch. Those ten channels are combined by a mux or a demux, creating the optical fiber input and output ports of one ToR. As shown in Fig. 8.2, an AoD-enabled OCS interconnects all the input and output ports of different ToRs, OPS modules, and traffic from/to other clusters as well.

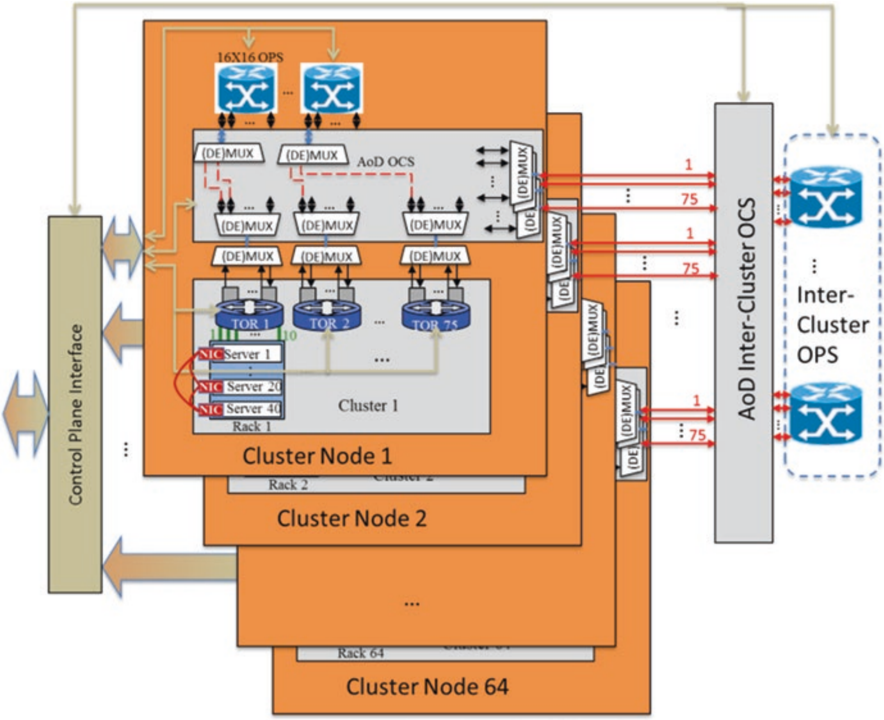


Fig. 8.2 High-level LIGHTNESS AoD-based architecture including inter-cluster connectivity

The use of the AoD structure for the OCS node makes it capable of supporting multi-granularity switching (from 10Gbps to 100Gbps). For example, by cross-connecting two ToRs directly, e.g., ToR to ToR, coarse granularity switching is achieved, whereas lower bandwidth granularities can be accomplished by separating channels from each ToR and assembling them again when required. After demultiplexing, channels from each ToR can be connected to the OPS module or interconnected with each other as OCS connections, depending on the requirement.

For this design, the OPS module could be either 4 × 4 or 16 × 16, depending on the port count and bandwidth required for the OPS. The number of links between each ToR and OPS module can be flexible, while it is also possible to scale the OPS switch by cascading several small switch modules in a Clos topology. The WDM mux/demux (e.g., AWG) interface between ToRs and AoD OCS backplane can potentially be replaced by the space division multiplexing (SDM) technologies, like multiple fibers (or ribbon fibers) or multicore fiber for further cost savings [5]. However, this might compromise the switching granularity at the OCS level.

8.3 LIGHTNESS Control Plane Architecture

The proposed LIGHTNESS control plane allows the provisioning of the hybrid optical data center network resources through a set of integrated and innovative functionalities and procedures: implementing connectivity services setup and teardown, dynamic service modification, path and flow computation, data center network resiliency and monitoring, and dynamic and automated resource optimization. The LIGHNTESS control plane is also equipped with an open, flexible, and extensible interface at the northbound for cooperation with management, VDC planning, and orchestration entities or user applications. In addition, an open, standard, vendor-independent and technology-agnostic southbound interface allows to configure and monitor the underlying physical devices composing the hybrid data center network. It is important to note that the term northbound interface (NBI) refers to any interface in a system used to communicate with higher layer systems and applications, as shown in Fig. 8.3. Similarly, southbound interfaces are defined to communicate a system with lower level systems and devices.

The finalized LIGHTNESS SDN control plane architecture is shown in Fig. 8.3. It is composed by an SDN controller natively implementing a set of basic network control functions and protocols, which are crucial to meet the requirements for data center services and applications.

A component that differentiates the LIGHTNESS SDN control plane from other SDN approaches is the virtualization manager (highlighted in red in Fig. 8.3). This module enables multi-tenancy within the hybrid data center by provisioning virtual DCNs. It is introduced as a new base service in the SDN controller, sitting on top of the resource manager and directly using the abstracted view of the hybrid optical data center network the resource manager exposes. In particular, the virtualization manager allows users to create their own virtual topologies on demand based on their specific QoS requirements.

Also, by taking advantage of the SDN approach, any features or functionalities supported by the LIGHTNESS control plane can be easily extended and implemented as network applications running on top of the SDN controller. As an example, a VDC composition application is proposed to enhance the aforementioned LIGHTNESS architecture to enable users to specify and dynamically modify their virtual DCNs (network application highlighted in red, top right of Fig. 8.3).

The VDC composition application is the component that implements the logic and the intelligence to interface with core virtualization manager for DCN virtualization. This means that any virtualization algorithm can run inside this VDC component, while the QoS provisioning and guarantee are provided by the virtualization manager. In particular, various algorithms and procedures for the VDC allocation are designed and evaluated in LIGHTNESS, e.g., a static VDC alloca-

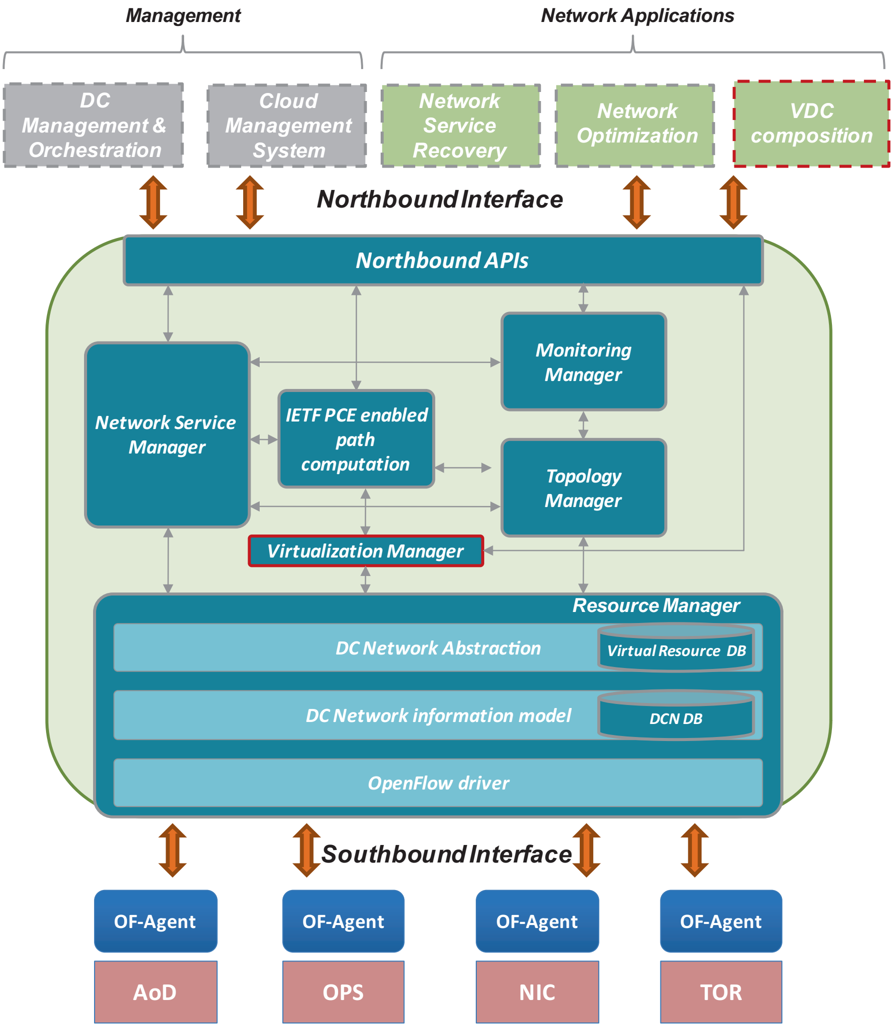


Fig. 8.3 High-level LIGHTNESS SDN-based control plane architecture

tion and a dynamic virtual slice allocation algorithm. In the first case, the minimization of the number of optical transponders is attained (DCN's CAPEX optimization), while in the second case, the number of blocked VDC allocation requests is evaluated. The main rationale behind such algorithm designs is the evaluation of the performance (in terms of allocated VDCs) of the LIGHTNESS architecture based on the hybrid data plane.

8.4 Technology Enablers for Flexible OCS/OPS

8.4.1 FPGA-Based Network Interface Card (NIC)

For DCNs, such as cloud data centers, special focus is needed on improving the intra-rack communication performance. The programmable NIC is designed and implemented to enable high-bandwidth and low-latency intra-rack communication and further empowers a flat and scalable all-optical data center inter- and intra- cluster architecture. The programmable NIC plugged directly to the server replaces the commodity NIC and enables intra-rack server-to-server full-mesh interconnection. The NIC is designed and implemented based on high-speed FPGA platforms [6] and optoelectronic transceivers. The 10x10G transceiver interfaces can be anything from ready off-the-shelf components (SFP+, CFP2, CFP4, etc.) or custom-made integrated PICs based on silicon photonics or III–V materials. In particular, hardware programming, framing, and processing methods are adopted for ultra-low latency processing and switching, as well as traffic aggregation techniques for maximum capacity handling. The traffic generated from servers is dissected over standardized network layer protocols (e.g., Ethernet) and then allocated into optical packets with the lowest possible processing delay to preserve ultra-low latency communication. It is capable of switching among multiple technologies, such as OCS and OPS, in a hitless manner, to achieve discrete bandwidth granularities. Most importantly the proposed NIC is also able to change between many-to-many aggregation mode and high-throughput point-to-point mode with little or no disruption to running applications. By using the programmable hybrid OCS/OPS NIC, the traffic features related to the OPS operation become flexible and programmable, allowing repurposing of the synchronous time-slotted mode OPS function to request and impose different levels of quality of service (QoS). These features include (a) variable optical packet size with (b) variable payload and overhead and are implemented as programmable network

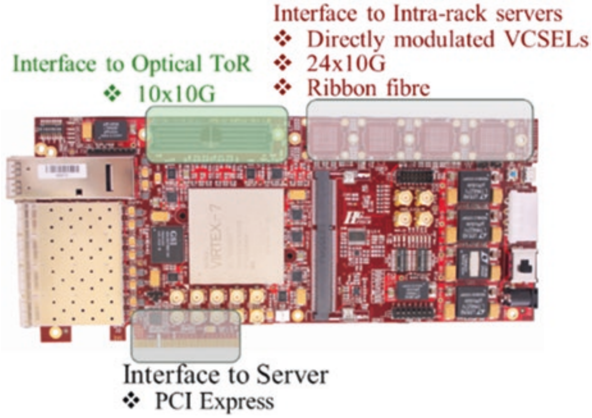


Fig. 8.4 FPGA-based network function programmable interface card (NIC)

functions rather than rigid hardware. All the above programmable features and FPGA hardware implementation are effectively exposed to the SDN-based control plane (Fig. 8.4).

8.4.2 OPS Module

A well-equipped 4 × 4 prototype integrating optical label processor (LP), optical switching fabric, and switch controller has been completed in the LIGHTNESS framework. As depicted in Fig. 8.5a, the proposed modular architecture allows the 4 × 4 OPS prototype to logically perform as two 2 × 2 OPS switches. The electrical label bits generated by each NIC are encoded in an in-band optical RF tone label [7, 8] by a prototyped label generator. The in-band optical labels are then coupled to each of the optical packets. Due to the lack of an optical buffer, a copy of the transmitted packet is stored, and a fast optical flow control between the OPS and the NIC is implemented for packet retransmission in case of contention. Figure 8.5b shows the photos of the 4 × 4 OPS prototype. The optical label of each packet is filtered out by a fiber Bragg grating (FBG) with a narrow passband. It is then detected and processed by the LP, and the recovered label bits are sent to the switch controller. The payload is fed into a 1 × N SOA-based broadcast and select stage. The switch controller checks the possible contention and, according to the lookup table (LUT), configures the 1 × N switch to forward the packets to the destination. The LUT can be remotely configured through the SDN-based control interface and can be seamlessly updated when necessary. In case of contention, the low-priority packet will be blocked, and a fast optical flow control signal (negative ACK) generated by using a low-speed

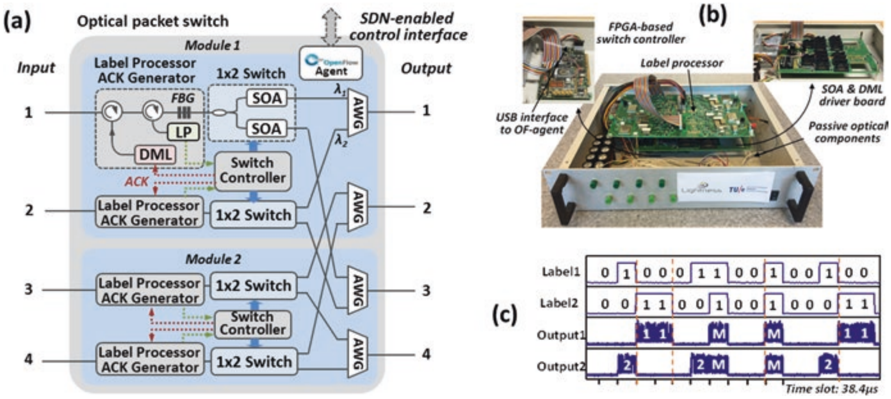


Fig. 8.5 (**a**) Schematic of the SDN-enabled OPS, (**b**) 4 × 4 OPS prototype, and (**c**) time traces of labels and OPS switch outputs for normal switching operation and multicasting. "*1*" and "*2*" indicate whether the present optical packet (*blue*) is switched in port No. 1 or 2. "*M*" indicates "multicasting," and thus the packet is present in both output ports of the switch

directly modulated laser (DML) is sent back to the NIC to request for the retransmission. If no contention occurs, a positive ACK informs the NIC to remove the packet from the buffer. As shown in the time traces of Fig. 8.5c, depending on the combination of the values of the labels, different (or no) outputs are activated for each 38.4 μsec timeslot. Multicasting is triggered when two label bits have been set as "11."

The fast response of the SOA in combination with the parallel processing of the label bits allows 20 nanoseconds switch reconfiguration time regardless of the port count. Besides the re-configuration of the LUT, the SDN-enabled control interface also enables programmable and flexible access from the SDN controller including the monitoring of the statistics.

8.5 Experimental Demonstration and Evaluation

The proposed fully dynamic all-optical circuit and packet-switched experimental data plane is able to carry out unicast/multicast switchover on request, while the powerful control plane enables the abstraction and virtualization of the networking resources. Thus, virtual data centers (VDCs) and virtual network functions (VNFs) are created on top of the data plane infrastructure. We have experimentally demonstrated practical intra-DCN interconnection use cases with deterministic latencies for both unicast and multicast, exhibiting monitoring and database transfer scenarios, each of which is facilitated by a joint software element based on the NFV and SDN principles. The outcomes validate a fully working thorough unification of the advanced optical data plane with the SDN-based control plane, committing to more efficient management of the forthcoming data center's compute and network resources.

8.5.1 Overall Experimental Architecture

The introduced high-level architecture, previously shown in Figs. 8.2 and 8.3, displays a next-generation fully reconfigurable DCN relied upon both optical circuit and optical packet switching technologies. Now for the overall experimental DCN design, as illustrated in Fig. 8.6, server blades within individual racks are connected via dedicated optoelectronic interfaces and all-optical ToR switches to the rest of the programmable DCN. The FPGA-based NICs operate as SDN-enabled hybrid OCS/OPS interfaces that support dynamic composition and transmission of Ethernet frames and/or optical packets with associated labels [9]. In order to provide direct OCS multicasting capabilities in each rack separately, optical power splitters are applied at each of the optical ToRs.

We propose a large port-count space/fiber switch as an optical ToR, because when combined with the advanced NICs on each server, they prevent the use of power-hungry electronic packet ToR switches (EPS) with unregulated latency values.

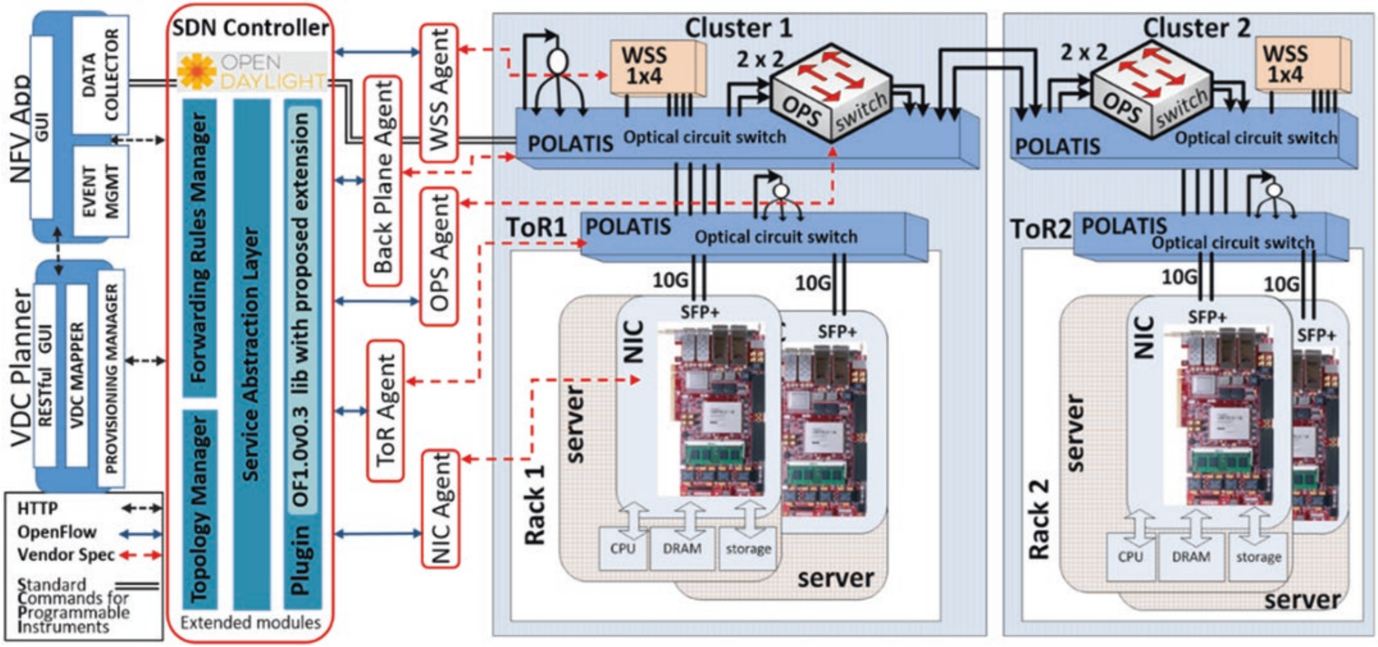


Fig. 8.6 Overall data center network architecture; experimental data plane (*center-right*), control plane (*left*), and virtualization schemes (*far left*)

Therefore, this flat design offers fixed low interconnection latency and potential cutback in power consumption of the overall network, due to the eventual lack of repeated O/E/O conversions. It also establishes full bandwidth transparency, since optical switches are entirely agnostic of the actual operated channel data rate, network protocol, or optical modulation format.

Within each cluster, the optical ToRs are linked to a top-of-the-cluster (ToC) optical switch, as shown in Fig. 8.6. The ToC incorporates a modular flexible optical network, containing a high-radix optical switch, serving as the optical backplane, optical power splitters, a wavelength/spectrum selective switch (WSS/SSS), and a 2 × 2 optical packet switch. The optical splitter at this network position enables OCS multicasting among various servers of separate racks within the same cluster and between remote clusters. The WSS's role is to groom/distribute multiple channels and traffic for inter-cluster communication in an elastic way. A passive optical filtering device, such as an AWG, could operate comparable jobs and be deployed instead of the WSS. However, a WSS is preferred due to its increased reconfigurability, flexibility, and bandwidth adaptability. A WSS/SSS could easily adapt to future possible bandwidth allocation needs with higher spectrum efficiency and narrower channel spacing than the present standardized ones.

The optical backplane also encompasses SDN-enabled OPS switching nodes that can perform nanosecond-fast packet switching, multicasting, as well as supporting of monitoring capabilities for optical packet reception and contention. To accomplish forwarding operation in the nanosecond scale, the OPS deals with the optical packets according to the optical label sent by the NIC [10], while the SDN controller has the role to arrange and supply the lookup tables to the OPS and NIC modules. This favors the decoupling of the fast (nanoseconds) forwarding operation of the optical data plane (to support time domain fast statistical multiplexing capability), from the slower SDN control plane (milliseconds), DCN virtualization, and VDC planner application.

Exploiting statistical multiplexing, OPS can also provide efficient and flexible bandwidth utilization therefore lowering the required optical port count at the backplane while guaranteeing the appropriate connectivity. In conjunction with the nanoseconds-fast label detection and switching control, bursty traffic demands are handled with higher degree of bandwidth granularity, lower latency, and adjustable per-packet processing agility.

Each of the NICs, optical ToRs, OCS and OPS switching nodes, and WSS switches is utterly controllable by the centralized SDN controller through a consistent control software interface, featured by each device's dedicated agent, as shown in Fig. 8.6 left. The agents abstract essential information from the hardware network modules, keeping the SDN controller up to date with every instance of the network, while they also translate and forward the management commands to the physical layer devices. Furthermore, logically on top of the SDN controller (and his left in Fig. 8.6), the VDC planner and the NFV applications provide an extra layer of abstraction and virtualization of the deployed physical computing and network infrastructure.

In summary, the programmable data plane facilitates the SDN-based DCN control plane to form and amend the physical layer topology, by flexibly arranging the relevant cross-connections in the optical backplane to suit the various applications' demands. Additionally, based on the DCN's specific requirements and data flow durations each time, the FPGA-based hybrid OCS/OPS NIC can be set up by the SDN controller on request along with the optical ToRs, ToCs, and OPS switches, achieving also unicast and/or multicast communication among servers.

8.5.2 All-Optical Experimental Data Plane

The data plane test-bed used for the experimental studies includes four rack-mounted Dell PowerEdge T630 servers, each supplied with a state-of-the-art FPGA-based NIC board with 10G SFP+ transceivers, serving as the reconfigurable interface of the computer blades to the optical network [9]. These servers were populated by miscellaneous virtual machines (VMs), one of which also hosts the SDN controller. All servers are joined to the 192 × 192 port optical circuit switch (supplied by Polatis), which in our experiment acts both as a ToR switch and as the OCS backplane on top of each cluster. Polatis beam-steering switching technology adds around 1 dB of loss per optical cross-connection (OXC), so multiple OXCs and hops are sustainable without extensive power and signal integrity penalties. As mentioned above in the overall architecture description, a 1 × 4 optical power splitter, two 1 × 4 WSS, and one SOA-based 4 × 4 OPS [10] are also attached to the optical backplane.

The novel NIC's range of capabilities combines network interface functions, programmable aggregation and segregation duties, OCS/OPS switching, and layer 2 switching services. The FPGA-based hybrid OCS/OPS NIC has been implemented on top of the NetFPGA-SUME development board [11, 12] and has been constructed to fit directly into a server's motherboard by replacing the conventional NIC. In the original design, it has an eight-lane Gen3 PCI Express interface for DRAM communication, one 10 Gb/s dual-line optical interface for receiving instructions from the SDN control agent and forwarding any feedback, two OCS/OPS hybrid 10 Gb/s SFP+ ports for inter-server communication, and an OPS label pin interface connected to the OPS label generator. The SFP+ transceivers' frequencies are selected in the C-band and are ITU compatible, in order to be consistent with the LCoS-based WSS and SOA-based OPS modules, which both normally operate in the 1550 nm frequency region.

The 1 × 4 optical power splitter supports any OCS one-to-four multicasting schedules. The WSSs are mostly operated for merging inter-cluster (or even inter-DC) traffic carried by channels from different servers or racks into a WDM superchannel. In the destination cluster, the local WSS separates the original WDM channels and properly routes them to the receiving racks and servers.

8.5.3 SDN-Enabled Experimental Control Plane

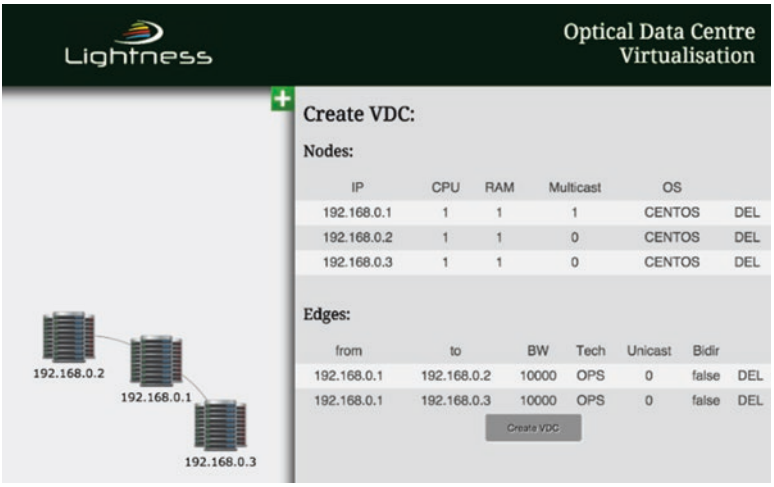
For these experiments, OpenFlow (OF) was selected as the standard control plane protocol to communicate the network devices with the controller. OpenDaylight (ODL) was operating as the SDN controller, and OF agents for Polatis, WSS, OPS switch, and hybrid OCS/OPS NIC were implemented to enable SDN-based programmability, as shown in Fig. 8.6. Further to the OF extensions, as previously reported in [6], the NIC OF agent is additionally enhanced to allow rearrangement of the generated OPS packets' duration. Moreover, ODL internal software elements were extended to provide some new network device-specific characteristics. For instance, in relation to the OPS and WSS ports, the switch manager and service abstract layer (SAL) were further developed to capture the supported wavelength and supported spectrum range, respectively, both of which were then used to validate the current configuration. Furthermore, the statistics of transmitted optical packet were collected and preserved by the statistics manager. In order to program accordingly the deployed optical devices, the forwarding rules manager has been also extended to build the appropriate set of configuration details. For example, for the OPS switch, label and output port details were included. For the WSS, central frequency, bandwidth, and output port information were contained in the forwarding rules manager extensions, whereas traffic matching to OCS or OPS, label, and output port information were included regarding the NIC. More specifically, the FPGA-based hybrid OCS/OPS NIC communicates with the OF agent through a bidirectional 10Gbps SFP+ Ethernet interface. The commands and information are encapsulated in a predefined 1504-byte Ethernet frame. Furthermore, through an extended ODL northbound interface, various applications can communicate directly with the hardware using the widely supported representational state transfer application programming interface (RESTful API).

8.5.4 Optical Data Center Virtualization Demonstration

As experimental objectives, two control plane operations have been produced and positioned on top of the ODL: a virtual data center planner (VDC planner) and a virtual network monitoring function (monitoring VNF).

First, the VDC planner allows the composition and arrangement of virtual network slices within the DCN, thus empowering multi-tenancy characteristics in data centers. In LIGHTNESS demonstration, the VDC planner consists of a graphical user interface (GUI) implemented in HTML/JavaScript that interfaces with a backend application developed in Python 2.7, which is capable of direct interaction with the ODL controller. The potential DC manager or user has access to the GUI with any existing browser, while he can create a VDC request, as shown in an example of Fig. 8.7. The specifiable VDC creation parameters are (i) number of servers to be allocated, (ii) optical links to be established, (iii) the preferred optical interconnection technology (OCS/OPS) for each link, and finally (iv) advice if there are any multicast properties for the allocated linked servers. Other parameters that are available (not mandatorily applied for the present VDC creation algorithm) are the required interconnection bandwidth and the bidirectionality of a given link.

Once the application has received the set of the abovementioned parameters for the VDC and has generated a group of static control flows to be forwarded to the



Create VDC:

Nodes:

IP	CPU	RAM	Multicast	OS	
192.168.0.1	1	1	1	CENTOS	DEL
192.168.0.2	1	1	0	CENTOS	DEL
192.168.0.3	1	1	0	CENTOS	DEL

Edges:

from	to	BW	Tech	Unicast	Bidir	
192.168.0.1	192.168.0.2	10000	OPS	0	false	DEL
192.168.0.1	192.168.0.3	10000	OPS	0	false	DEL

Fig. 8.7 VDC planner application with OPS or OCS multicasting options

DCN devices by ODL, it distributes them among the various data plane modules (NIC cards, OPS switches, WSS, and OCS backplane). This set of flows is produced in JavaScript Object Notation (JSON) format and sent to the ODL controller via a RESTful API.

Figure 8.7 shows an example of a VDC request using the aforementioned VDC planner. The user defines, in this example, three server hosts, one of them chosen as a multicast node, which will duplicate the content to the other two. In this case, the request also specifies OPS as the multicast technology for the VDC interconnection.

8.5.5 *Experimental Results and Evaluation*

For the experimental demonstration, we incorporated all the available data plane and control plane resources, as presented in the previous sections, and evaluated several intra-DCN interconnection schemes, based on VDC applications' and VNF's requests and commands.

First, for the physical layer the DCN was comprehensively evaluated for intra-rack, inter-rack, and inter-cluster unicast and multicast communication by measuring the bit error rate (BER) for both OCS and OPS switching technologies with realistic traffic (scrambled PRBS payload from the traffic analyzer). The results are presented in Fig. 8.8. The traffic analyzer provides the FPGA-based NIC with 10 Gb/s Ethernet traffic, and then the NIC forwards the data to one of its hybrid OCS/OPS ports. When OPS mode is selected, the NIC, relying on the configuration instructed by the SDN controller, sets the optical packet duration, encapsu-

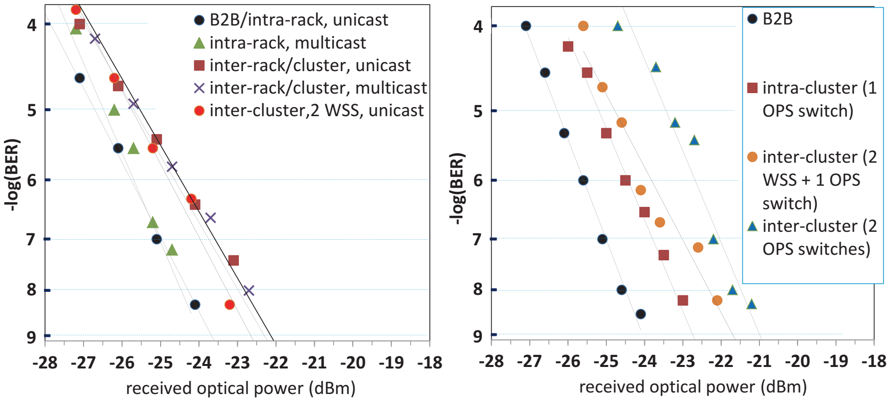


Fig. 8.8 (*Left*) OCS BER curves for intra-/inter-rack unicast/multicast and inter-cluster through 2 WSSs with 10-Gigabit Ethernet (GbE) traffic. (*Right*) OPS BER curves for intra-cluster and inter-cluster through WSS and 1 or 2 OPS switches with 10GbE traffic

lates a certain number of Ethernet frames, and issues the optical packet, while the packet label is generated and integrated in parallel. On the receiver end, in the lack of a burst mode receiver and in order to properly recover the clock and lock the data, a preamble between 10us and 30us was used, depending on the desired quality of transport.

Intra-rack communication is achieved by pushing optical flows from source to destination server via the optical ToR switch for unicast and through an optical splitter for multicast operation. Inter-rack and inter-cluster communications are identically accomplished by going through multiple Polatis OXCs and/or optical power splitters. For inter-cluster groomed interconnection, optical signals have to propagate through two extra WSSs for WDM mux/demux and switching functions. Small penalties of <2 dB are observed for all OCS interconnection scenarios (unicast/multicast), as shown in the BER curves of Fig. 8.8 (left).

OPS BER plots in Fig. 8.8 (right) indicate 1 and <3 dB penalties when signals pass through one (for intra-cluster) and two (for inter-cluster) switches, respectively.

Following the BER test of the physical links, we collected network layer 2 results regarding the chip-to-chip interconnection latency. This is the access latency between one NIC's direct memory access (DMA) and the destination NIC's DMA. DMA driver's actual delays are excluded and are separately measured later for different DMA lengths and fixed Ethernet frames. In addition, interconnection throughput is monitored and plotted, exhibiting OCS-to-OPS switchover and vice versa.

We measured the DMA-to-DMA access latency with Ethernet traffic generated from the traffic analyzer with various PRBS payloads. The traffic analyzer firstly feeds the transmitting FPGA-based NIC with the traffic. Then, the NIC pushes the data flows to the all-optical network, employing either OCS or OPS, toward the destination NIC. Optical signals traverse through the DC network and the appropriately established cross-connections. Finally, the destination NIC forwards the received

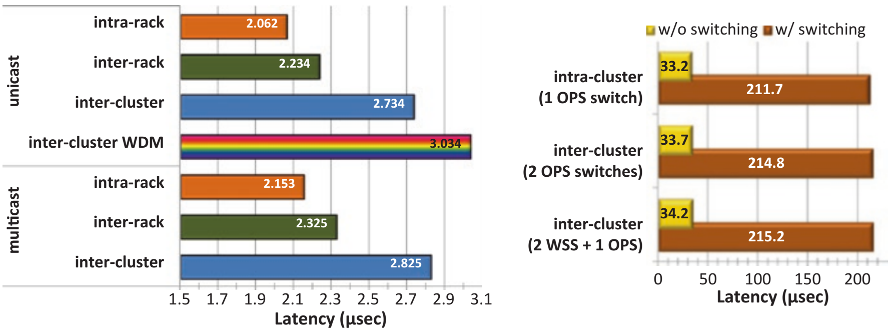


Fig. 8.9 DMA-to-DMA OCS (*left*) and OPS (*right*) latency for various intra-DC scenarios

traffic flows back to the traffic analyzer. The accumulated chip-to-chip latency was estimated by subtracting the traffic analyzer-to-NIC (and vice versa) delays and the traffic analyzers' processing delays (a few hundreds of microseconds).

The displayed investigation and measurements are based on the best possible latency with maximum bitrate, so, for OPS with switching, actual throughput bitrate is around 3 Gb/s, and for OCS it is around 8 Gb/s. All quantified latency values include FPGA physical and logic delays, which can strongly vary depending on the used frame length, selected transmission/switching scheme (OCS or OPS), and the FPGA design.

Figure 8.9 (left) shows unicast and multicast OCS access latencies for all the studied interconnection scenarios. Predictable latency values are exhibited between 2 and 3 μsec for most communication scenarios. The majority of the latency is contributed by the electronic processing (PHY and logic) of the Ethernet traffic in the source and destination FPGA-based nodes.

Figure 8.9 (right) shows intra-/inter-cluster OPS access latencies with and without switching. When no switching is performed, the clock of the receiving end of the transceiver is continuous, so there is no need for recovering it with extra payload (e.g., preamble dummy key characters).

Figure 8.10 (right) depicts the variations of interconnection throughput when a change from normal operation OCS to OPS and vice versa is initiated by the NFV application and executed in the data plane. Protocol overheads are restricting throughput in OCS whereas in OPS the dummy key characters used for receiver side and clock recovery are confining the maximum throughput, plus the fact that OPS traffic is transmitted and switched in a 50% manner for this experiment.

8.6 Conclusion: Discussions

Regarding the energy efficiency of the proposed architecture, it is widely known and proven that optical network modules deliver higher port count with fixed power consumption and non-limiting switching capacity. On the contrary,

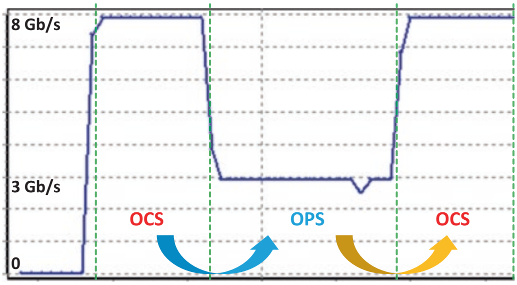


Fig. 8.10 Throughput plot over time, illustrating the OCS-to-OPS switchovers and vice versa

electrical network elements usually provide fewer ports with restricted switching capacity and significantly higher power consumption values, which can also vary counting on the switching traffic load. The predominant reason why all-optical switching is more energy efficient than electrical switching is due to the lack of optical transceivers, which count for more than 50% of the total power expenditure. Optical switching devices used in this experiment show much more modest power consumption (some tens of Watts) than regular electrical switches (several hundreds of Watts). More specifically, the 192 × 192 OCS switch consumes 75 Watts in regular operation, the 1 × 4 SSS consumes less than 10 Watts, while the total power consumption of the 4 × 4 SOA-based OPS prototype is around 50 Watts. OPS's breakdown energy contributions are FPGA controller 15 Watts, label processor 20 Watts, and SOA driver 15 Watts. Those values are based on off-the-shelf components, and further power reduction is possible by dedicated hardware photonic integration. Furthermore, recent experimental and simulation research [13, 14] have shown considerable distinction in terms of energy efficiency between architectures using all-optical switching and in others using conventional electrical switching equipment.

Lastly, with regards to the performance of the demonstrated VDC and NFV, not only the total (re)configuration times but also the contribution of each individual element was calculated. The total OCS/OPS channel configuration span includes (i) the ODL SDN controller processing time, (ii) control message transmission time (which strongly depends on the actual experiment setup), and (iii) the device reconfiguration time. Particularly in this experiment, the SDN controller needs around 210 msec to process requests arriving from the RESTful API in order to push the matching OF configuration commands to the network devices' OF agents. It approximately lasts a further 200 msec for those commands to reach the OF agents, to be processed and forwarded. At last, Polatis OCS switch, OPS switch, WSS, and NIC require approximately 16, 10, 300, and 18 msec, respectively, to properly configure themselves. The above device configurations of course can be carried out in parallel. So, assuming that in order to establish an end-to-end OCS channel we need to successfully configure the optical ToR before configuring NIC, establishing an OCS channel will need 970 msec (also using the WSS) or 690 msec without WSS, while it takes around to 420 msec for OPS connection establishment.

This chapter demonstrates an all-optical programmable DCN architecture enabling OCS/OPS multicasting for realistic monitoring, migration, and transferring scenarios. The novel networking schemes demonstrated in this chapter include an SDN-enabled, virtualize-able, and reconfigurable optical data plane integrated and supported by an extended control plane. In this work, the SDN controller and NFV server are able to offer data plane monitoring and database migration function virtualization, on top of a virtual data center environment implemented and managed by a VDC planner application.

It is apparent that there is a trend for all-optical switching in DCNs in order to tackle the disadvantages of current architectures, exactly as it was done with metro and core networks a couple of decades ago. However, the requirements of those two categories of networks are very different. Hence, WDM technologies commercially available and suitable for metro core and regional networks cannot be introduced in the intra-DCNs without further modification. The introduction of space division multiplexing (SDM) [5] and the latest advances in photonic integration will play a critical role in the future development of intra-DCN architectures and designs, by attempting to exploit the space dimension, in addition to frequency and time, and by bringing massive manufacturing costs of optical components down, while further improving their energy efficiency.

References

1. 'Lightness EU FP7 project'. [Online]. Available: http://www.ict-lightness.eu/
2. G.M. Saridis et al., Lightness: A Function-Virtualizable Software Defined Data Center Network With All-Optical Circuit/Packet Switching. J. Lightwave Technol. **34**(7), 1618–1627 (Apr. 2016)
3. G. M. Saridis et al., 'LIGHTNESS: A Deeply-Programmable SDN-enabled Data Centre Network with OCS/OPS Multicast/Unicast Switch-over', in *European Conference on Optical Communication (ECOC)*, Valencia, 2015, p. PDP 4.2
4. N. Amaya, G. S. Zervas, D. Simeonidou, 'Architecture on demand for transparent optical networks', in *2011 13th International Conference on Transparent Optical Networks*, 2011, pp. 1–4
5. G. M. Saridis, D. Alexandropoulos, G. Zervas, D. Simeonidou, 'Survey and Evaluation of Space Division Multiplexing: From Technologies to Optical Networks', *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 2136–2156, Fourthquarter 2015
6. B. Guo et al., 'SDN-enabled programmable optical packet/circuit switched intra data centre network', in *Optical Fiber Communications Conference and Exhibition (OFC), 2015*, 2015, pp. 1–3
7. W. Miao, F. Yan, H. Dorren, N. Calabretta, 'Petabit/s Data Center Network Architecture with Sub-microseconds Latency Based on Fast Optical Switches', in *European Conference on Optical Communication (ECOC)*, Valencia, 2015
8. W. Miao, X. Yin, J. Bauwelinck, H. Dorren, N. Calabretta, 'Performance assessment of optical packet switching system with burst-mode receivers for intra-data center networks', in *2014 European Conference on Optical Communication (ECOC)*, 2014, pp. 1–3
9. Y. Yan, Y. Shu, G. M. Saridis, B. R. Rofoee, G. Zervas, D. Simeonidou, 'FPGA-based Optical Programmable Switch and Interface Card for Disaggregated OPS/OCS Data Centre Networks', in *European Conference on Optical Communication (ECOC)*, Valencia, 2015
10. W. Miao et al., SDN-enabled OPS with QoS guarantee for reconfigurable virtual data center networks. IEEE/OSA Journal of Optical Communications and Networking **7**(7), 634–643 (Jul. 2015)

11. ‘NetFPGA-SUME Virtex-7 FPGA Development Board’, *Digilent*. [Online]. Available: http://store.digilentinc.com/netfpga-sume-virtex-7-fpga-development-board/
12. ‘NetFPGA’. [Online]. Available: https://netfpga.org/site/#/systems/1netfpga-sume/details/
13. M. Imran, P. Landais, M. Collier, K. Katrinis, ‘A data center network featuring low latency and energy efficiency based on all optical core interconnect’, in *2015 17th International Conference on Transparent Optical Networks (ICTON)*, 2015, pp. 1–4
14. Y. Ji et al., All Optical Switching Networks With Energy-Efficient Technologies From Components Level to Network Level. IEEE Journal on Selected Areas in Communications **32**(8), 1600–1614 (Aug. 2014)

Chapter 9
Hybrid OPS/EPS Photonic Ethernet Switch and Pure Photonic Packet Switch

Hamid Mehrvar, Huixiao Ma, Xiaoling Yang, Yan Wang, Dominic Goodwill, and Eric Bernier

9.1 Introduction

There have been many efforts to address the limitation of electronic packet switches by utilizing photonic switches that are high capacity and energy efficient for the datacenter core switches. Examples include: arrayed waveguide grating (AWG) and fast-tuning lasers [1] and a hybrid of MEMS optical circuit switching (OCS) and electrical packet switching (EPS) to share the traffic load [2, 3]. Given that AWG systems only work with single-DWDM wavelength per signal, it will be a challenge to support low-cost 40GE/100GE transceivers. On the other hand, the slow switching speed of MEMS limits this scheme to steady traffic sessions that last for seconds. Such traffic sessions are small fraction of the traffic. As a result, this chapter focuses on solutions where the core switches are wavelength-agnostic photonic packet switches, which may be implemented in silicon photonics, as the scale and performance of silicon photonics improve.

Optical packet switching (OPS) [4] has been proposed as a solution to replace the OCS [5]. However, there are several challenges in implementing silicon photonic packet switches that are buffer-less and operate at 40GE/100GE or higher rates. One, probably insurmountable, challenge is to connect an optical signal from an input port to an output port within the nanosecond(s) time frame allowed by the

H. Mehrvar (✉) • D. Goodwill • E. Bernier
Huawei Technologies Canada Co. Ltd, 303 Terry Fox Drive, Ottawa, ON K2V 3J1, Canada
e-mail: Hamid.Mehrvar@huawei.com

H. Ma • X. Yang • Y. Wang
Huawei Technology Co., LTD, Longgang, Shenzhen 518129, China

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
DOI 10.1007/978-3-319-61052-8_9

inter-packet gap (IPG) of native Ethernet packets. It is extremely difficult to design a photonic switch with response time of less than the inter-frame gap. Another problem is that photonic switches are buffer-less space switches that disconnect the physical layer as packets are switched. As a result, proper considerations need to be given to photonic framing to address clock recovery due to light discontinuity and the buffer-less behavior of the photonic switches.

Section 9.2 demonstrates a hybrid OPS/EPS that off-loads large packets for photonic switching, whereas control packets, broadcast, and small data packets are switched by an electronic packet switch. The hybrid OPS/EPS is our first step toward photonic packet switching [6]. It discusses how a hybrid OPS/EPS can be equipped with packet compression to increase the inter-frame gap for photonic switching as well as using a framing scheme with proper frame scrambler to solve the receiver decode synchronization. When a hybrid scheme is equipped with proper control architecture, it permits the datacenter capacity to grow as over 90% of traffic can be off-loaded to photonic packet switches.

In Sect. 9.3, we take a step beyond hybrid OPS/EPS architecture. We discuss the functions of a pure photonic packet switch in terms of photonic frame wrapper, control signaling, and synchronization. Photonic frame wrapper/un-wrapper allows switching of all packets through the photonic fabric. The material in this section has been originally published in [7]. In Sect. 9.4, we briefly address the scalability scenario of a photonic switch fabric, followed by conclusion and discussion on research direction in Sect. 9.5.

9.2 Hybrid OPS/EPS

A hybrid of OPS/EPS is required to address the switching during IPG as well as a proper photonic framing for receiver clock recovery. As we discuss below, these challenges can be addressed by accelerating the photonic line rate by at least 10% faster than the native rate and equipping photonic frames with a long preamble sequence for frame recovery.

An example of photonic-based hybrid OPS/EPS datacenter architecture is illustrated in Fig. 9.1. It features N servers, M top-of-rack (TOR) switches, P aggregation nodes each enhanced by new functions, a PxP switch OPS, and a tandem EPS of the same size but lower capacity. The connectivity architecture shown is a three-tier hierarchy with N servers connected to a TOR, M TORs interconnected by an aggregation node, and P aggregation nodes interconnected by the core hybrid OPS/EPS. Assuming $N = 48$, $M = 32$, and $P = 32$, the total number of servers is approximately 50,000. A two-tier architecture is possible by including the enhanced functions of aggregation nodes into the TORs or edge switches.

The use of a hybrid switch core is feasible because the traffic can be segmented based on packet length. Traffic inside general-purpose datacenters is typically

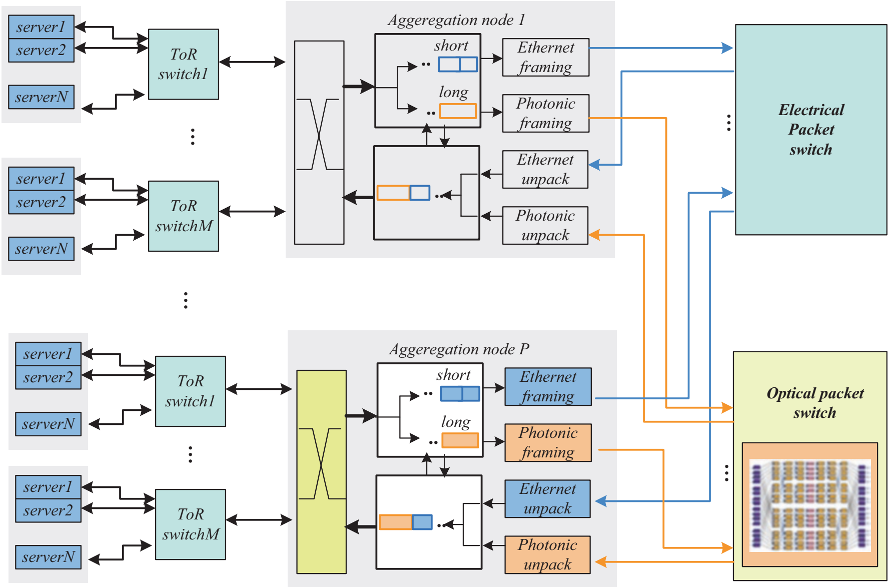


Fig. 9.1 Structure of a hybrid OPS/EPS datacenter network

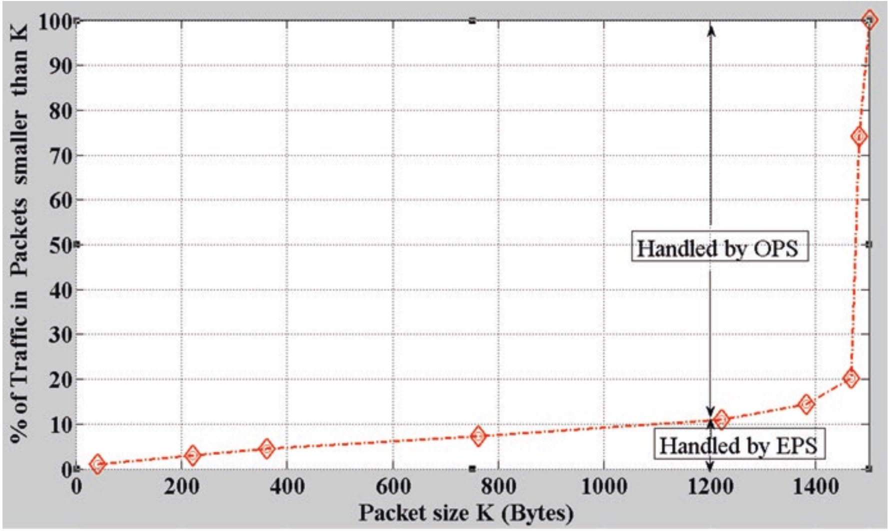


Fig. 9.2. Percentage of traffic in packets in a datacenter (calculated using data from [8])

bimodal, with 45% of packets less than 500B (bytes), 5% between 500 and 1400B, and 50% more than 1400B [8]. By defining a packet length splitting threshold *K*, in bytes, one can switch long packets using the OPS and the short packets using the EPS. Figure 9.2. shows the percentage of the total bandwidth as a function of the packet length splitting threshold K. Clearly, the large majority of the bandwidth uses long packets. For a threshold of $K = 1200$B, 10% of the bandwidth is switched by the EPS, and 90% by the OPS, which is expected to deliver a huge saving in power and cost compared to using a very large EPS for all datacenter traffic. We observe that while the bandwidth of the EPS is capped, the number of packets handled by the EPS does increase. Thus, attention must be paid to the packet-rate handling capacity of the EPS.

9.2.1 *Photonic Functions of Hybrid OPS/EPS*

Aside from packet contention scheduling, the photonic protocol functions, as shown in Fig. 9.3, are packet separation, photonic framing to allow label insertion and receiver clock synchronization as the packets are switched and light is discontinued. On the ingress side, the long packets are separated from the short packets. The short packets are forwarded to the EPS using conventional protocols. Long packets are prepared for photonic switching by resolving packet contention, then compressing, bitwise scrambling, adding a photonic destination label, and sending them out through an optical transmitter. By compressing the long packets, there is enough inter-packet gap for the insertion of the photonic label, but more importantly there is more time for photonic switch connection setup and for receiver synchronization at the destination aggregation switch. The packet compression is achieved by raising the clocking rate on the output physical layer. For egress, away from the core, the aggregation switch performs the inverse operations, receiving photonic long

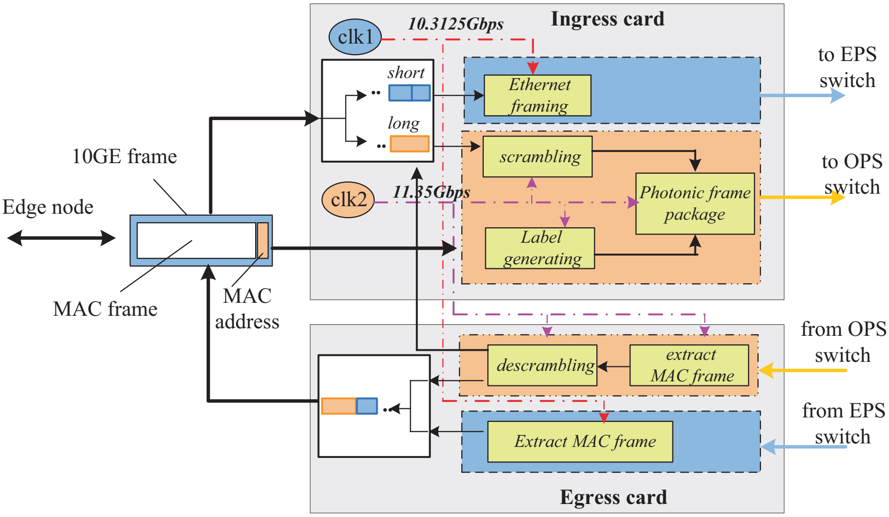
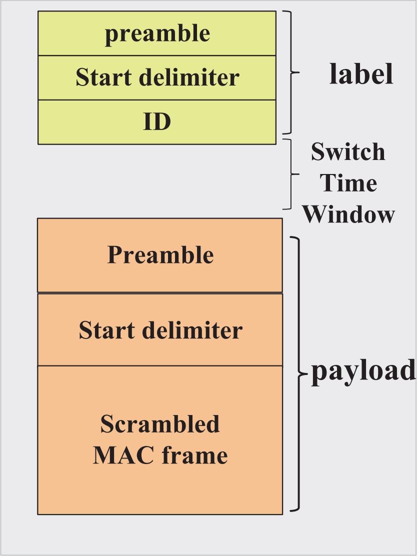


Fig. 9.3 Block diagram of core-facing egress/ingress of aggregation switch

Fig. 9.4 Photonic frame format



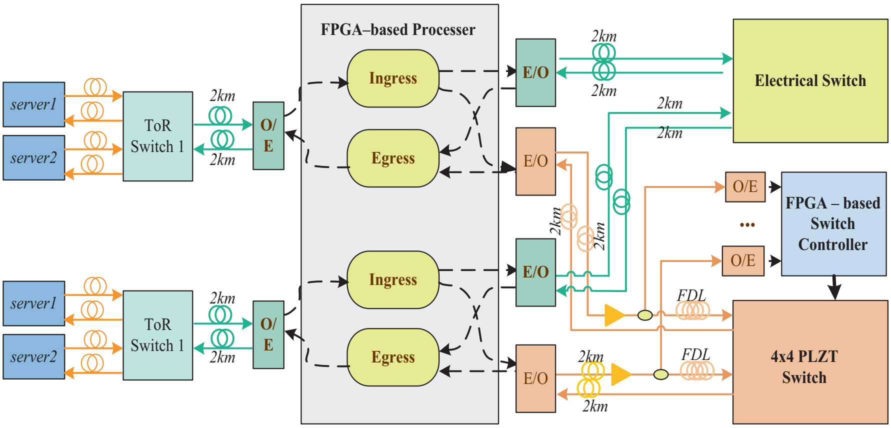


Fig. 9.5 Test-bed configuration

packets from the OPS and conventional short packets from the EPS, reordering the packets, and forwarding all traffic as conventional IP/Ethernet packets toward destination TORs.

Figure 9.4 shows a photonic frame structure containing an in-band inserted label, switch time window, and payload. The “preamble” section in both label and payload is used for clock recovery in the receiver. The “start delimiter” is a fixed pattern used for frame alignment and to locate the label and payload. The duration of the time window is set to allow enough time for photonic switch setup. The ID field in the photonic label refers to the output physical port number of the OPS, and it is obtained from a mapping table between the destination MAC addresses that are reachable via a given OPS port. Another enhancement is a scrambling suitable for photonic frames (e.g., an OTN frame scrambling algorithm) that scramble each MAC frame independently before encapsulating it in a photonic frame.

9.2.2 Experimental Setup of Hybrid OPS/EPS

An experimental setup of an OPS/EPS test-bed, using a 4 × 4 PLZT [9], also called lead lanthanum zirconate titanate, electro-optic switch with switching time of 10–20 ns is shown in Fig. 9.5. The servers transmit packets with lengths varying between 64B and 1510B. The links between TOR switches and ingress/egress card are 10 Gigabit Ethernet LAN PHY, although the architecture can work for higher rates with proper compression scheme. The ingress card receives server packets and determines their length. It retransmits each short packet as an Ethernet packet on a 1310 nm SFP+. It converts each long packet into a photonic frame of Fig. 9.4 by adding scrambling, label, and guard gap and sends it out the photonic interface using the transmit side of a commercial 1550 nm XFP, running 10% over-clocked (11.35Gb/s). Each ingress card concatenates a label followed by a scrambled MAC frame into a photonic frame. The scrambler is x^{16} type. The label is a short non-scrambled pattern at 11.35Gb/s. The photonic packet is sent through a fiber to the PLZT switch, which has 11–13 dB loss depending on the path, with an EDFA as preamplifier.

At the OPS, a fiber splitter taps 10% power to the OPS label reader, which sets the connectivity of the switch for each packet. Since the test-bed is a single-hop network, the label detector has zero lock time for each label. In the absence of packet labels, the photonic transmitter sends an idle pattern to maintain continuity at the label detector. All the electrical functions of the aggregation node and OPS are implemented in Xilinx Virtex-7 FPGAs [10]. The servers, TORs, and EPS are standard off-the-shelf products. A fiber delay line (FDL) between the tap and the PLZT switch delays the signal long enough for the switch controller to read the photonic label and set the switch con-

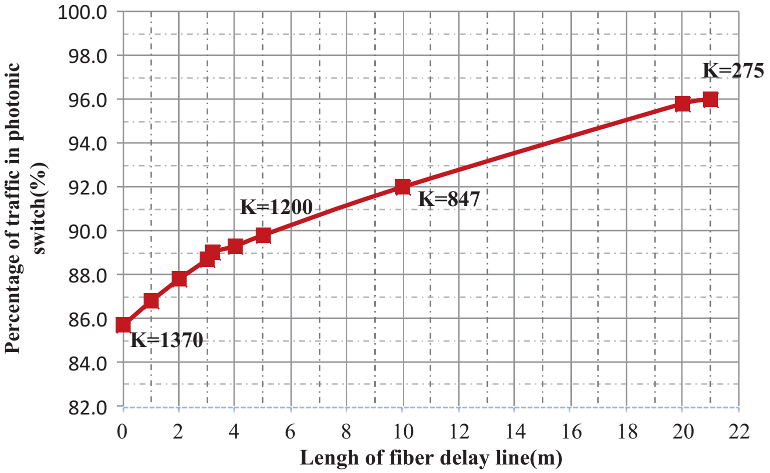
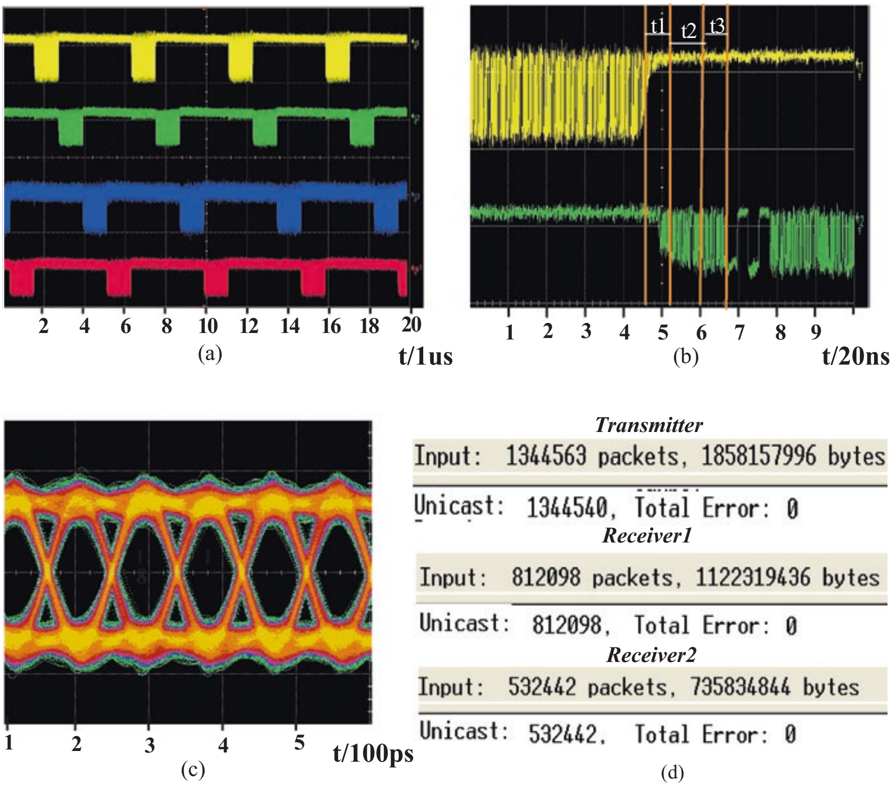


Fig. 9.6 Percentage of traffic handled by the photonic switch vs. the length of fiber delay line



Transmitter

```
Input:  1344563 packets, 1858157996 bytes
Unicast:  1344540,  Total Error: 0
```

Receiver1

```
Input:  812098 packets, 1122319436 bytes
Unicast:  812098,  Total Error: 0
```

Receiver2

```
Input:  532442 packets, 735834844 bytes
Unicast:  532442,  Total Error: 0
```

Fig. 9.7 Experimental results for the test-bed: (**a**) Packet waveform on all four output ports of the OPS. (**b**) Detailed output packet waveform. (**c**) Eye diagram of the switched signal. (**d**) Packet statistics of two video sessions through the test-bed

nection. The relationship between threshold, K; processing time, p; compression rate, r_1/r_2; and propagation delay in the fiber delay line, D_{FDL}, is governed by:

$$K > \frac{p - D_{FDL}}{r_1^{-1} - r_2^{-1}}, \tag{9.1}$$

where r_1 = 10Gb/s and r_2 = 11.35Gb/s are the pre- and post-encapsulation rates. With total processing time, p, measured to be 130 ns, the smallest possible packet splitting threshold K without fiber delay line is about 1370B, which results in the OPS handling of 86% of the bandwidth. Figure 9.6 shows the combined results obtained from (1), and Fig. 9.2., plots the percentage of bandwidth handled by the OPS as a function of packet splitting threshold K and the length of fiber delay line.

Figure 9.7 shows some of the experimental results for the test-bed. For simplicity, in this implementation, contending long packets are silently dropped as there is no contention resolution. In Sect. 9.3, we discuss an upgrade for contention handling and scheduling. One server, e.g., server 1, sends Ethernet packets with four different destination MAC addresses, each destined to a different photonic output port of the OPS. Figure 9.7a shows the waveforms captured at the four output ports of the PLZT switch. The photo-receiver voltage polarity is inverted, with horizontal lines as no light and the waveforms as the switched packets. The details of outputs 1 and 2 are shown in Fig. 9.7b, where output 1 completes a photonic frame transmission and output 2 starts sending preamble and photonic label. The switch response time is $t_1 = 12\ ns$, residual preamble for receiver synchronization $t_2 = 15\ ns$, and start-frame-delimiter time $t_3 = 12\ ns$. The eye diagram at the receiver is shown in Fig. 9.7c. Since total processing time is measured to be 130 ns, the latency for control processing is approximately $130\text{-}t_1\text{-}t_2 = 103\ ns$. This delay can be compensated by the FDL with length of 21 *m*, resulting in the OPS handling a maximum of 96% of datacenter-type traffic with threshold setting of 275B (Fig. 9.6).

The test-bed was used to transmit UDP real-time traffic. Videos play smoothly at the receiver with no packet loss. Figure 9.7d shows statistics of the packets sent by the TOR that is attached to the server and the packets received by each TOR that is attached to a client. The data shows successful transmission of all packets through the OPS.

9.3 Pure Photonic Packet Switching

In Sect. 9.2, we discussed off-load of large packets in a hybrid system comprising an electronic core switch which handled control, broadcast, and small data packets and a photonic core switch which handled large packets and photonic label recognition at an accelerated photonic line rate.

The main issue associated with the hybrid OPS/EPS solution is its limited application due to the following reasons. First, splitting a packet stream to long and short packets is troublesome for TCP connections; second, there may be MAC learning confusion as there are two receive interfaces; third, packet reordering is required as there are different path delays for OPS and EPS. A pure photonic switch could address these issues.

In this section, we discuss a synchronous photonic packet switch system that handles all packet types and offers two distinctive features. One is the separation of control path and data path, and the other is a photonic framing architecture that allows carrying of all packet sizes while addressing the challenges of switching speed and signal recovery. Separation of data path and control not only allows conventional low-cost 40GE/100GE optics but also lays the groundwork for scalable datacenter architectures with the possibility of integrating control with software-defined networking (SDN). By wrapping many packets destined to the same output into a single photonic frame, the efficiency of the photonic packet switch is improved. With out-of-band signaling for photonic labels and system synchronization, the physical layer is simplified.

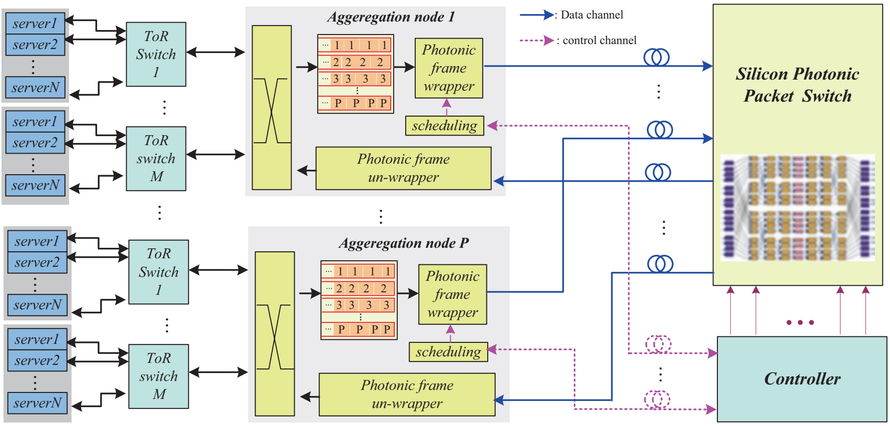


Fig. 9.8 The structure of a pure photonic switch for datacenter

Figure 9.8 shows a three-tier synchronous photonic packet switch datacenter architecture, a modified structure of Fig. 9.1. The aggregation switches are modified, to interface with the photonic switch. As mentioned previously, a two-tier architecture can be implemented by moving the modified functions inside the TORs and not using aggregation switches. The core switch comprises a buffer-less *PxP* silicon photonic time-slot space switch for packet switching and a controller for connection management, scheduling, and system synchronization. The control data for synchronization, scheduling, and routing uses a different wavelength than the data path. Using out-of-band labels simplifies the controller and simplifies the physical layer of the photonic switch.

9.3.1 Photonic Functions of Pure Photonic Packet Switch

The network of Fig. 9.8 is a time-slot-based synchronous network, in which the length of a wrapped photonic frame is equal to the slot time, e.g., 1 microsecond. Each aggregation node adjusts its frame transmit boundary clock to ensure that the photonic frames arrive at the inputs of the photonic packet switch at the same time, despite diverse fiber lengths, using frame synchronization.

Figure 9.9a, b shows the flow of photonic frames at the output of the aggregation node and the input of the photonic packet switch, respectively. If the network is small, each aggregation node may have a fixed allocation of time slots to allow simpler scheduling and contention control, e.g., round robin. In a large network, a request-grant approach with efficient scheduling scheme can be applied [11].

Two functions are introduced to achieve the goal of handling unicast and broadcast packets of all lengths, for both data and control, at line rate using only the photonic switch. One is a photonic frame wrapping/unwrapping function, and the other is a

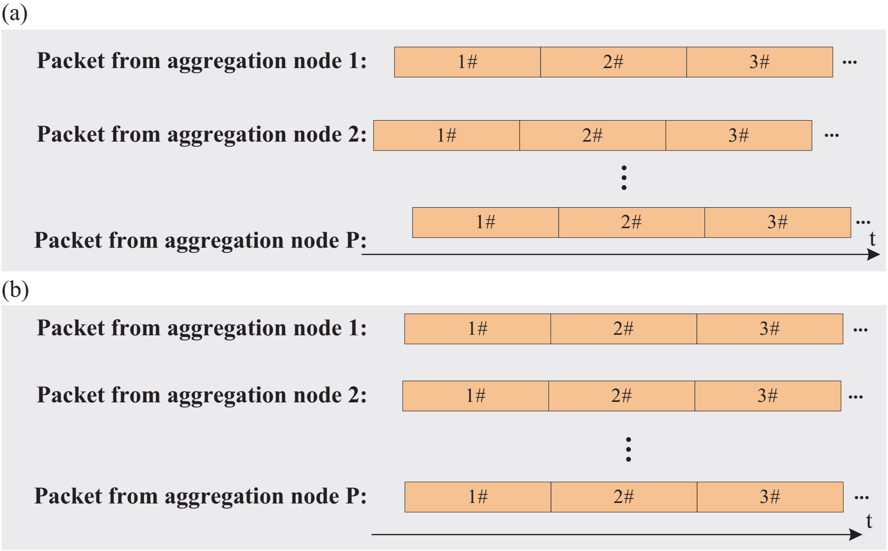

(a)
Packet from aggregation node 1:
1#
2#
3#
...
Packet from aggregation node 2:
1#
2#
3#
...
Packet from aggregation node P:
1#
2#
3#
...
t
(b)
Packet from aggregation node 1:
1#
2#
3#
...
Packet from aggregation node 2:
1#
2#
3#
...
Packet from aggregation node P:
1#
2#
3#
...
t

Fig. 9.9 Flow of packets (frames): (**a**) at the output of the aggregation node and (**b**) at the input of the photonic switch

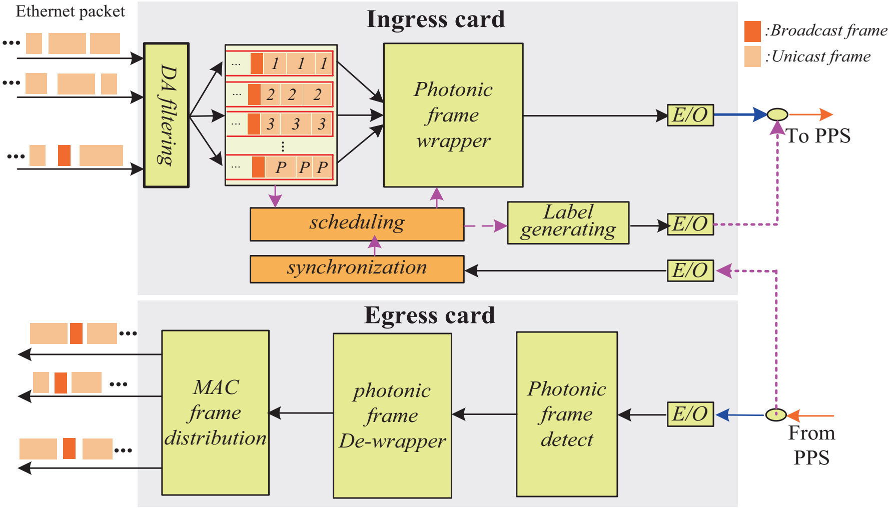


Fig. 9.10 Egress/ingress of aggregation switch, facing toward core photonic packet switch

synchronization scheme. Figure 9.10 is a block diagram of the wrapping/unwrapping element, respectively, in the ingress/egress of the aggregation node.

In the ingress, native Ethernet/IP packets are placed into one of *P* virtual output queues based on their destination MAC address (or IP address). If the frame is a broadcast frame, it is copied into all virtual queues.

The switch controller and scheduler determine which queue can wrap its packets into a photonic frame in the next time slot. The wrapping function includes de-queueing multiple MAC (or IP) packets from a given virtual queue, packaging them into one large photonic frame, followed by inserting the label, the preamble, and the start delimiter sequence. The inserted label is transmitted onto the control channel, e.g., at 1310 nm wavelength, while the wrapped payload is transmitted though the data channel with single or multiple wavelengths in the 1550 nm band in which the photonic switch operates.

The egress performs inverse operations. It receives photonic packets from the photonic packet switch (PPS), disassembles the packets, and forwards each conventional Ethernet packet to the destination TOR using conventional routing or switching protocols. If the frame is a broadcast frame, it is sent to every attached TOR.

Figure 9.11 shows photonic frame wrapping structure and its out-of-band label. Packets for the same destination are concatenated in a wrapper with no inter-packet gap (IPG). The *preamble* is used for receiver recovery at the egress cards, and the *start delimiter* is a fixed pattern used for frame alignment.

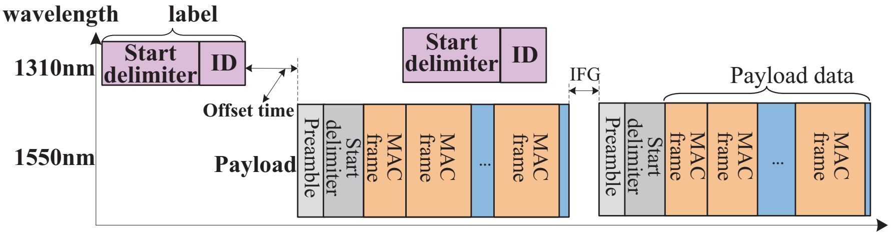


Fig. 9.11 Photonic frame format

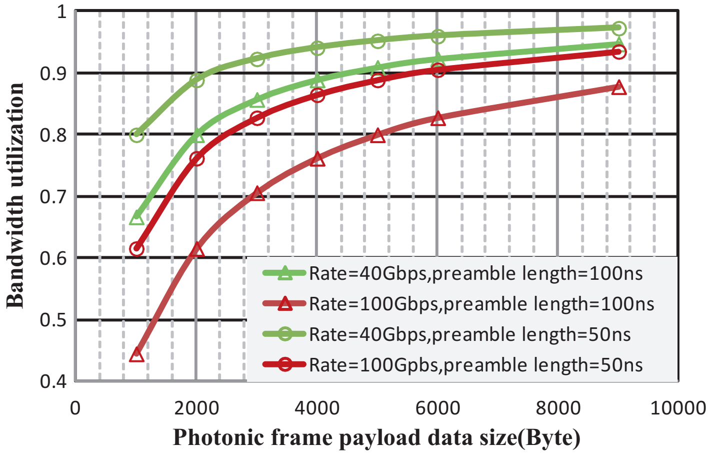


Fig. 9.12 Bandwidth utilization ratio versus photonic frame wrapper size

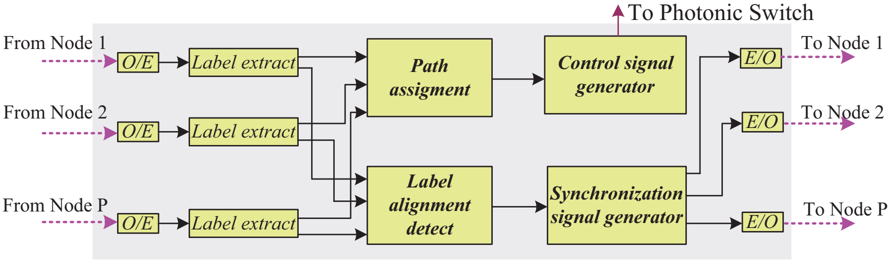


Fig. 9.13 Photonic packet switch controller

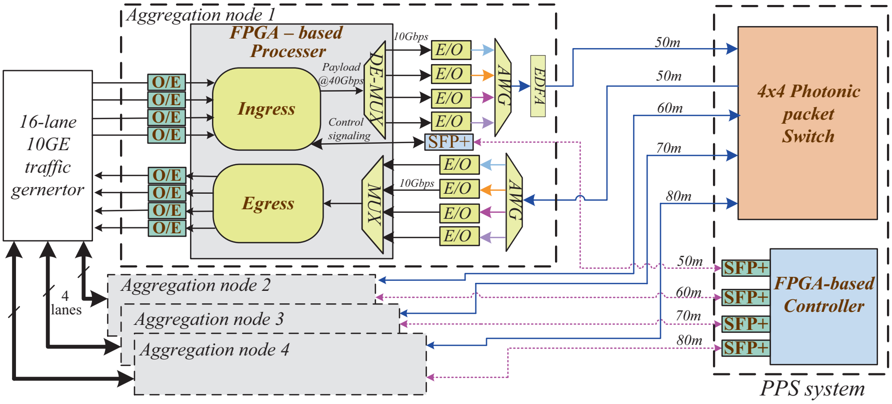


Fig. 9.14 Test-bed configuration

The offset time between label and payload allows label decoding. Switch setup is performed during the inter-frame gap (IFG). The ID field of the photonic label identifies the destination aggregation node. Bandwidth utilization is impacted by the number of packets in a wrapper. To achieve full utilization, it is required that

$$\text{packets-per-wrapper}^{*}\text{inter-packet-gap} > \text{preamble} + \text{start delimiter} + \text{inter-frame-gap}$$

Figure 9.12 shows bandwidth utilization versus wrapper size, for different line rates and preamble sizes. Assuming constant switching setup time and start delimiter duration, the two factors that impact bandwidth utilization are wrapper size and burst recovery time at the receiver.

To achieve time-slot synchronization among aggregation nodes, the time and frequency offset between the aggregation switches is adjusted continually. This function is performed by the photonic packet switch controller shown in Fig. 9.13. The controller extracts the labels to generate the signals that set up the photonic switch fabric and determines the time offset values between the incoming labels. This offset is used to send synchronization signals to all aggregation switches through the control channel. Once an aggregation switch receives the synchronization signal, it either delays or advances the photonic wrapper frame clock accordingly.

9.3.2 Experimental Example and Results

Figure 9.14 shows an experimental setup of a pure photonic packet switch test-bed. It uses a 4 x 4 photonic switch that operates in the 1550 nm band and connects four aggregation nodes with 40Gb/s interfaces. Data path packets (unicast 64–1518B) and control/comms packets (both unicast and broadcast) are generated by four 10GE test generators. When an aggregation node receives a packet from any of the four 10GE client lanes, it extracts the MAC destination address and enqueues it to the virtual queue of the destination egress card. It then wraps many MAC frames of the same queue into a fixed-length photonic frame and adds a label, preamble, and start of the payload sequence. The corresponding label is sent on a 1310 nm SFP+ with an offset time before the photonic data packet, to allow processing time in the photonic core node.

The wrapped photonic frame is sent at 40Gb/s. To mimic standard four-wavelength 40GE and to overcome I/O rate limitations of the FPGA, the 40G signal in the demonstration is implemented using four commercial XFPs at four different 10Gbp/s wavelengths. The 1550 nm light is preamplified by an EDFA before being sent to the photonic packet switch. Each aggregation node is connected to the photonic switch by a different length of fiber. The contention scheme is implemented by a simple round robin scheduler, as the test-bed switch size is small.

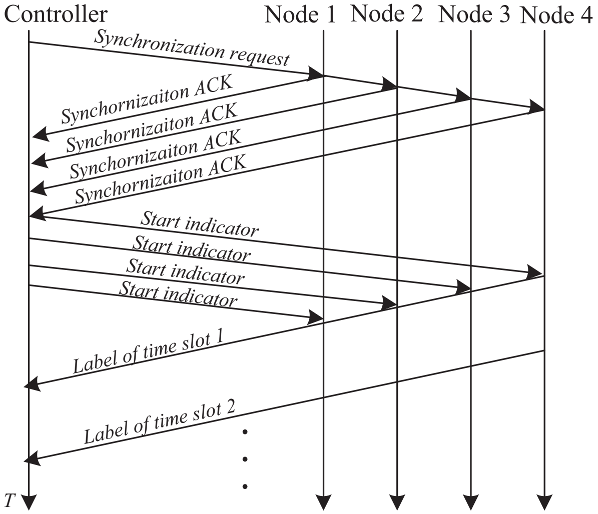


Fig. 9.15 Flowchart of the synchronization process

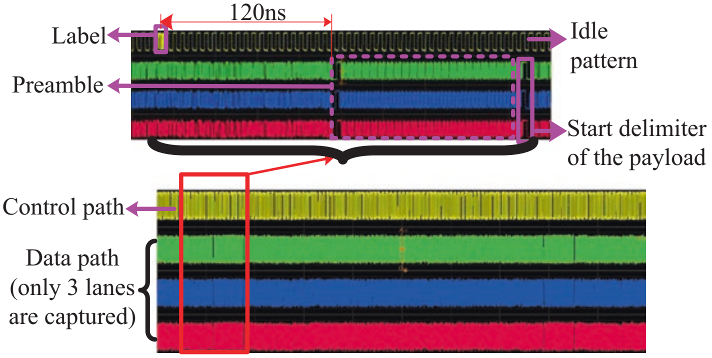


Fig. 9.16 Photonic frame output by an ingress node

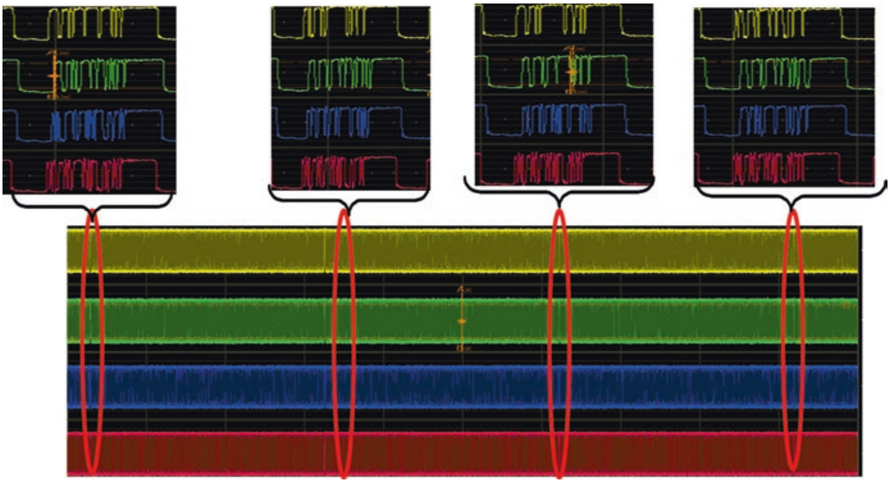


Fig. 9.17 Control signals (four lanes label and idle pattern) at the input of the controller

When the photonic packet switch controller receives a label on the 1310 nm channel, it decodes the label and sets the connectivity of the switch for the photonic packet. To avoid label lock time at the controller, each ingress card sends an idle pattern between the labels. The modifications of the aggregation node are implemented in external Xilinx Virtex-7 FPGAs [10] that share a common clock. This allowed simplification of clock synchronization as there is no frequency offset among the four aggregation nodes. As shown in Fig. 9.15, the controller sends a synchronization message to the ingress card first. Upon receipt, the ingress returns ACK to the controller. From the ACK arrival time, the controller determines the time that "start indicator" signal should be sent. When the egress card receives the start indicator signal, it wraps the phonic frame and sends it out. If there is no MAC frame to send, an idle photonic frame is sent.

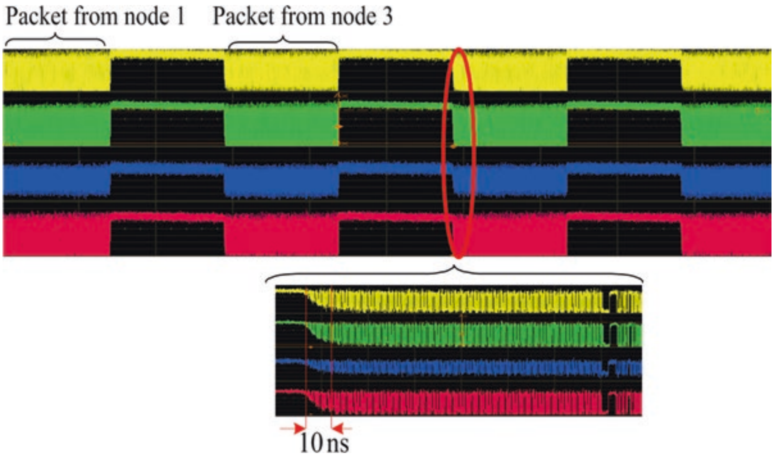


Fig. 9.18 Signal at the output 1 of the switch

Figure 9.16 shows the waveforms captured at the output port of the ingress cards, using a four-channel real-time oscilloscope that captures three lanes of data path and the control path. As seen in the magnified part, the payload follows the label with a 120 ns offset because the total controller processing time was previously measured to be 120 ns. Also, an idle pattern is sent between labels to ensure that the label detector sees a continuous signal at the physical layer. The magnified part of the data path shows the preamble for receiver recovery, followed by the start delimiter of the payload. Figure 9.17 shows the four captured labels at the input of the controller have been well aligned by the synchronization process.

Figure 9.18 shows the switched signal at output 1 of the photonic packet switch (four wavelengths) with received packets from aggregations nodes 1 and 3. The photo-receiver voltage polarity is inverted, with horizontal lines representing no light and the waveforms indicating the switched packets. For better illustration, nodes 2 and 4 are setup to have no packets to send. As seen in Fig. 9.18, the photonic switch physical response time is about 10 ns.

9.4 Scalability of Photonic Packet Switch

In the previous section, we demonstrated how a single photonic packet switch can be implemented. One of the challenges for deployment of photonic switching is how to scale photonic switches. Scaling photonic packet switches can be achieved by using a set of fast photonic switch matrices, with less than 10 nanosecond switch time, implemented as carrier injection optical silicon photonic Mach-Zehnder matrices (e.g., [12]) with a time-slot synchronous centralized control to perform scheduling of photonic frames through the stack of photonic switches, with no reordering of frames. With the use of M small $N \times N$ photonic switch matrices, all controlled by a time-synchronous controller, the stack of switches can act as one switch with capacity of $(M \times N) \times (M \times N)$. In [13], it has been demonstrated that two 4 × 4 photonic switches in parallel can be aggregated by such a control scheme to deliver eight-channel capacity. This stacking principle can be extended to larger networks.

9.5 Discussion and Conclusion

In this chapter, we presented a single-hop photonic switching node for interconnecting aggregation nodes or edge switches of datacenters. In summary, while photonic switching of native Ethernet packets can be applied to any packet size, it is more efficient to be applied to large packets to allow enough time for photonic switching time and clock recovery due to light discontinuity. A hybrid photonic switch can use packet length discrimination, compression, and scrambling to off-load electrical packet switches. Results indicate that a hybrid photonic switch core can handle 96% of the native Ethernet traffic in a datacenter. Yet, the hybrid approach is deemed to be impractical due to requirement of packet re-ordering and the challenges posed for handling TCP traffic as a result of splitting the traffic flow stream. On the other hand, a pure photonic switching fabric based on fast carrier injection optical switches [12] that is buffer-less, wavelength agnostic and operates as time-space switch can address these challenges. The presented pure photonic switch fabric is characterized by: packets wrapped into a photonic frame, signaling waveband separated from data path waveband, and a synchronization scheme for alignment of photonic frames from all aggregation switches at the input of the photonic core switch. An example of a 40GE system connected by a 4 × 4 photonic switch and synchronized by a photonic controller was presented.

The advantage of the proposed time-slot synchronous pure photonic space switch based on carrier injection is that it is rate agnostic, buffer-less, and scalable. It allows the buffering of packets to be performed in the electronic domain at the aggregation node or edge switches. Upon receipt of a grant by scheduler, packets are wrapped into photonic frames and sent through photonic fabric to their destinations.

As for the future research direction, special focus is required in two main areas. One is in the design of fast and scalable switch fabric building blocks with low insertion loss. Lowering the insertion loss can be addressed by improving the design, manufacturing, and packaging process and by integrating on-chip semiconductor optical amplifier (SOA) to compensate for the optical loss. The other area is related to the control of a large-scale photonic switch fabric that interconnects many aggregation nodes within 2 km in a next-generation datacenter. While this chapter discussed a centralized synchronous system for a single-hop core photonic switch, scaling the control scheme to a cluster of photonic switches, or multi-hop photonic switching, remains a challenge. In other words, while scaling the data path of a photonic fabric can be achieved by synchronizing an array of many smaller photonic switches, e.g., [13], scaling the control architecture in terms of synchronization and scheduling scheme of photonic packet fabric constitute future areas of research. The research question that needs to be addressed is that what type of control architecture, i.e., centralized or distributed control architecture, would allow operation of a petabit photonic fabric. It is important to note that while a petabit photonic fabric offers energy efficiency (power and cooling) and small physical footprint, it is required that they offer, at a minimum, same quality of service (QoS) offered by today's electronic packet switches. Given the desire to perform buffering in the electronic domain, such QoS performance metric includes packet end-to-end delay and jitters from one aggregation node to another one through the photonic fabric.

References

1. K. Xi et al., A petabit bufferless optical switch for data center networks, in *Optical networks*, (Springer, New York, 2013), pp. 135–154
2. G. Wang, D.G. Anderson, c-Through: part-time optics in data center, SiGCOMM'10, August 30, 2010, New Delhi, India
3. The software defined hybrid packet optical datacenter network. Calient Technology white paper, 2013
4. Martin Maier, *Optical switching networks*. (Cambridge University Press, Cambridge (Feb. 2008)
5. H.T. Mouftah, P.-H. Ho, *Optical networks: architectures and survivability* (Springer, New York, 2002)
6. H. Ma et al, *Hybrid photonic ethernet switch for datacenters.* OFC (2014)
7. X. Yang et al, 40Gb/s pure photonic packet switch for datacenters OFC (2015)
8. T. Benson et al, Understanding data center traffic characteristics, WREN'09, August 21, 2009, Barcelona, Spain
9. EpiPhotonics Corp. http://epiphotonics.com/
10. Xilinx Inc. https://www.xilinx.com/products/silicon-devices/fpga/virtex-7.html
11. M. Kiaei et al, Scalable architecture and low-latency scheduling schemes for next generation photonic datacenters. ICC (2016)
12. N. Dupuis et al, Nanosecond-scale Mach-Zehnder-based CMOS photonic switch fabrics. J. Lightwave Technol. **35**, 615–623 (August, 2016)
13. H. Mehrvar et al, Scalable Photonic packet switch test-bed for datacenters. OFC (2016)

Chapter 10
OPMDC: Optical Pyramid Data Center Network

Maria Yuang and Po-Lung Tien

10.1 Introduction

Data center networks (DCNs) [1–3] have been designed and deployed to provide a reliable and efficient infrastructure for supporting a wide variety of emerging cloud and enterprise applications and services. Evidence shows that these applications and services not only involve much client-server (north-south) traffic flowing in and out of DCNs but also spawn a massive amount of east-west server-to-server traffic within DCNs. These applications and services are data rich by nature and demand high-bandwidth and low-latency transport of data. Besides, recent studies have further shown an ever-growing trend toward the variety and complexity of new cloud and enterprise applications and services. Such a trend places a higher demand for large-scale DCNs [4–6] that can deliver substantially high bandwidth, low latency, and reduced power consumption. These facts altogether bring about an urgent need for the design and implementation of next-generation DCN architectures and technology that can meet the demand.

There has been an increasing tendency toward a *modular* [3] and *incremental* [4] design for large-scale DCNs. A modular data center is constructed from purpose-engineered modules (e.g., pods, containers) that are flexibly expanded to the original data center infrastructure in an architecture compliant manner. The incremental design allows small rollouts and seamless expansion, resulting in agile and economical deployment and delivering resources on fully as-needed basis.

M. Yuang (✉)
Department of Computer Science and Information Engineering, National Chiao Tung University, 1001 University Road, Hsinchu 30050, Taiwan
e-mail: mariayuang@gmail.com; mcyuang@cs.nctu.edu.tw

P.-L. Tien
Department of Electrical and Computer Engineering, National Chiao Tung University, 1001 University Road, Hsinchu 30050, Taiwan
e-mail: polungtien@gmail.com

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
DOI 10.1007/978-3-319-61052-8_10

Current state-of-the-art DCNs [2, 3] embrace optical transmissions but electrical switching of packets via electrical switches, such as top of rack (ToR), aggregation, and core switches. The electrical switches are interconnected based on two architecture designs: scale-up and scale-out. The scale-up approach uses a hierarchical tree structure in which the switches toward higher level of the hierarchy demand higher capacity and port count. On the other hand, the scale-out approach, aka the leaf-spine architecture, uses a large number of identical low-cost tier-1 ToR and tier-2 aggregation switches to deliver full bisectional bandwidth with extensive path diversity between servers. Both approaches have different pros and cons, but result in high power consumption [7] due to using power-hungry electrical-to-optical (E/O) and optical-to-electrical (O/E) transceivers. By and large, the electrical switching-based approaches have been deemed to be incapable of meeting the aforementioned DCN demands. This fact, coupled with recent advances in semiconductors and silicon photonics, becomes key driving forces for developing new optical architectures and technologies for next-generation DCNs.

Thanks to advances in silicon photonics and wavelength division multiplexing (WDM) technologies, optical WDM switching networks and systems have been proposed and widely deployed in long-haul and metro networks. Examples are optical wavelength cross-connects (OXCs/WXCs) [8] and reconfigurable optical add-drop multiplexers (ROADMs) [9, 10]. Optical WDM switching possesses some attributes, such as high bandwidth, low latency, and low power consumption, which are proved advantageous to future DCNs. A number of optical WDM DCN architectures that have been proposed [11–13] are based on various types of optical switching devices. Of these devices, the wavelength selective switch (WSS) has been considered the most promising candidate for building next-generation DCNs due to its flexible per-wavelength switching capability, besides being technologically mature and commercially available.

Being a key enabler for ROADMs, WSS is tailored to flexible per-wavelength provisioning. It is typical a $1 \times N$ optical switch that flexibly routes each wavelength from the input port to any of the N multiwavelength output ports, independent of how other wavelength channels are routed. WSS features [14] simple electronic control, low cost, high reliability (low FIT rate), and low power consumption (e.g., <2 W for a typical 1x9 WSS), but at the expense of a reconfiguration delay of a few milliseconds. Such a delay poses a challenge of supporting dynamic packet-based transport that is of crucial importance for future DCNs. The major contribution of our work lies in the design of a unique DCN architecture that operates in conjunction with SDN-based resource management, with the result that, despite the high reconfiguration delay limitation, the DCN efficiently achieves ultra-low-latency packet-based communications.

10.2 OPMDC Architecture

The architecture of a full-scale optical pyramid modular data center network (OPMDC) [4] is shown in Fig. 10.1. It consists of three types of WSS-based optical switching nodes in three tiers: (tier-1) ROADM, tier-2 WXC, and tier-3 WXC.

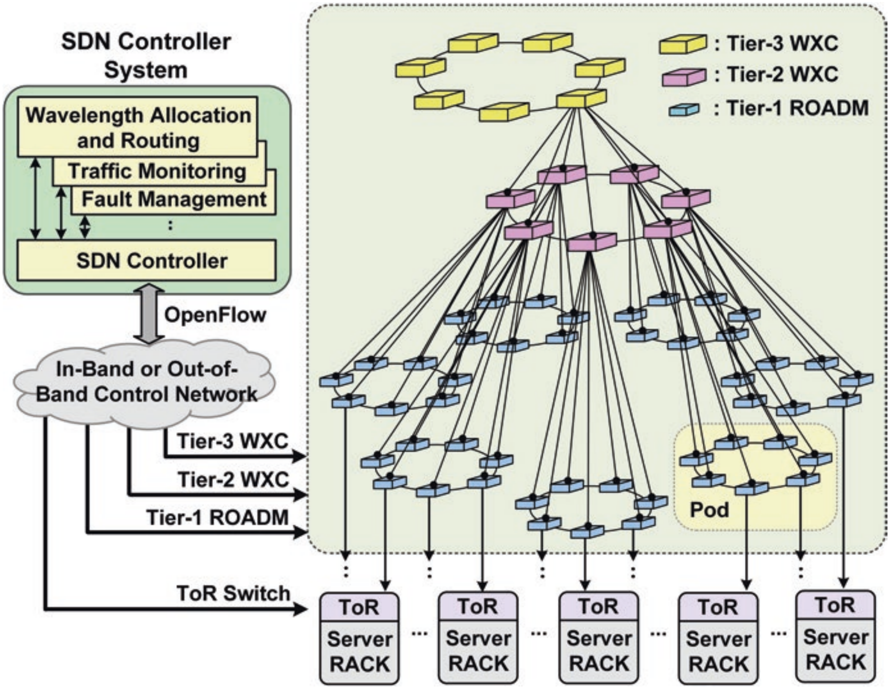


Fig. 10.1 The OPMDC architecture ($B = 7$)

While each ROADM node is directly connected to a ToR switch in tier 1, WXC nodes perform high-bandwidth optical per-wavelength switching in tiers 2 and 3. Further, OPMDC is controlled and managed by a software-defined networking (SDN) controller system in a centralized manner. The system consists of wavelength allocation and traffic engineering modules as well as an SDN controller. The controller governs the operation/configuration of OPMDC switching nodes (optical nodes and ToR switches) based on the OpenFlow protocol via an in-band or out-of-band control network. More details about the implementation of the SDN controller system are given in Sect. 10.4.

OPMDC is recursively built based on a pyramid construct that contains a polygonal base with an odd number (B) of nodes that are mesh connected (not a ring). In the example shown in Fig. 10.1, $B = 7$. The mesh connection is made via ribbon fiber cables, as will be described in detail later. Accordingly, two types of building blocks in OPMDC can be constructed incrementally: *pod* and *macro-pod*. A pod is the basic building block that spans tiers 1 and 2. It consists of B ROADM nodes at the base of its pyramid, each of which is down connected to a ToR switch and up connected to the apex (a tier-2 WXC node) of its pyramid.

A macro-pod is the larger building block that spans three tiers. It consists of B tier-2 WXC nodes (that are mesh connected), each of which is down connected to a pod and up connected to the apex of its pyramid in tier 3. For example, the OPMDC shown in Fig. 10.1 delineates a complete macro-pod that contains B (=7) pods, or B^2

(=49) ROADM nodes. Further, a full-scale OPMDC contains *B* macro-pods that are connected through *B* tier-3 WXC nodes that are also mesh connected. As described, these building blocks can be deployed on an incremental basis. For example, to interconnect only a total of 3*B* server racks, OPMDC will contain three pyramids each of which has *B* ROADMs at the base (tier 1) and one tier-2 WXC at the apex, while the three tier-2 WXCs are mesh connected.

OPMDC boasts four unique features that are deemed crucially significant to support emerging cloud applications. *First*, due to the pyramid topology and horizontal mesh interconnection, OPMDC offers powerful broadcast capability without overloading the network. *Second*, OPMDC allows extensive wavelength reuse. For example, the same set of wavelengths can be reused for transporting traffic within different pods. Such a feature enables OPMDC to employ highly efficient static preallocation of wavelengths, thereby accomplishing ultra-low-latency packet-based transport under a substantial portion of traffic patterns. This will further be described in Sect. 10.3. *Third*, OPMDC is highly fault tolerant [15]. Due to short distances within data centers, failures in optical links are generally disregarded. In addition, as was mentioned earlier, the key device in ROADM and WXC is WSS, which possesses an exceedingly low failure-in-time (FIT) rate [14]. Unlike E-switch-based nodes, each optical ROADM or WXC node contains individual active and passive devices that collaboratively support a number of parallel light paths. Any failure in a node occurs only on the basis of an individual device (also with low probability) rather than the entire node. Additionally, with the rich horizontal mesh connectivity, the occurrence of a few failures results in only minor throughput degradation instead of node disconnections from the rest of the network.

Finally, the pyramid topology allows OPMDC to adopt fairly simple routing under both normal and fault conditions. Under the normal condition, traffic from one ROADM to another ROADM within the same pod is routed through the mesh connection of the pod. For the traffic within a macro-pod but without a pod, packets are passed from the source ROADM to its tier-2 WXC, then through the mesh connection to the destination pod's tier-2 WXC, and finally down to the destination ROADM. By the same token, inter-macro-pod traffic is routed through two tier-3 WXC nodes. Under the condition of a fault in a pyramid, alternative routes can simply be taken through other available horizontal mesh connections or the apex of the pyramid.

10.2.1 Internal Design of ROADM and WXC Nodes

The tier-1 ROADM and tier-2/3 WXC nodes have been designed in such a way that they can be implemented using commercially available components that have already been widely deployed. Their key component is the $N \times 1$ WSS module. Its distinctive features, including low cost, high port count, low power consumption, and high reliability, are ideally suited for data center switching.

10.2.1.1 Tier-1 ROADM Node

The tier-1 ROADM node was originally developed by CoAdna Photonics [16] and then revised to tailor for the OPMDC project. As shown in Fig. 10.2a for $B = 7$, each ROADM node contains an optical multiplexer (MUX) and demultiplexer (DEMUX), a $B \times 1$ WSS, an erbium-doped fiber amplifier (EDFA), ribbon cables, and a series of passive splitters (a 3-way splitter and a number of tap couplers). Each ROADM is horizontally connected to $(B-1)/2$ peer ROADM nodes in the east via $(B-1)/2$ pairs of fibers, and likewise for the west.

For $B = 7$ as shown in Fig. 10.2, among the 3 pairs of fibers for either direction, 2 fiber pairs are for pass-through traffic and 1 fiber pair for add and drop of traffic to the local node. While travelling along the fibers, packets are tapped into the WSS of each of the three ROADM nodes through tap couplers for broadcasting.

Each ROADM node is directly connected to a ToR switch. There are W uplink ports in the ToR switch that are populated with W wavelength-specific DWDM transceivers, respectively, where W is the total number of wavelength channels. For transmissions, the W channels of optical signals are combined via MUX and passed to EDFA that boosts the peak signal power to ensure sufficient power budget. With a 1×3 splitter, the multiplexed traffic is 3-way broadcast to east and west ROADM nodes and the tier-2 WXC node. For receiving, a $B \times 1$ WSS is used to select W signals from the B input ports of WSS (i.e., $(B-1)/2$ ports from the east, $(B-1)/2$ ports from the west, and 1 from the tier-2 WXC). After DEMUX, W channel signals are passed to the corresponding ports of the ToR switch.

Figure 10.2b depicts how the B (=7) nodes of a pod are horizontally interconnected. Notice that, for any traffic within a pod, packets from the same source node to different destination nodes share the same fiber link. Thus, it requires distinct wavelengths to carry traffic to different destination nodes. On the other hand,

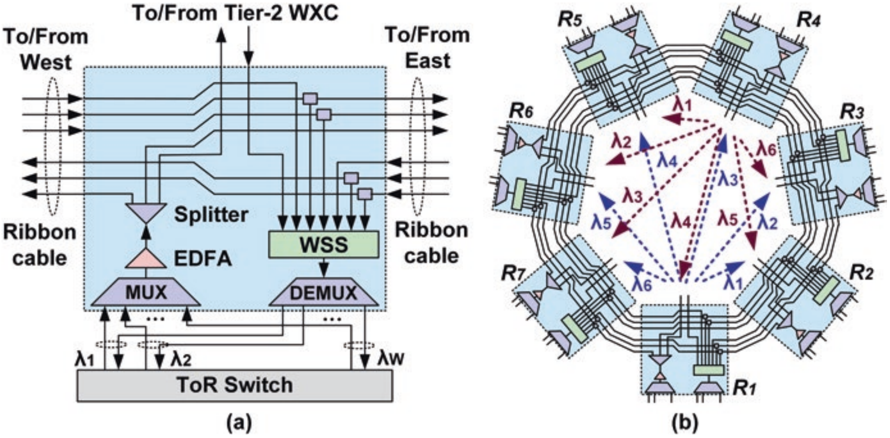


Fig. 10.2 (**a**) Design block diagram of tier-1 ROADM node ($B = 7$), (**b**) Horizontal interconnection of ROADM nodes in a pod

packets from different source nodes are carried by different fiber links, thus causing no wavelength contention. For example, node R_1 can send packets to node R_3 via a wavelength, while R_2 can send packets to node R_4 via the same wavelength without contention. As a result, as shown in Fig. 10.2b, both nodes R_1 and R_4 can send packets independently to six other nodes via λ_1 to λ_6. Thus, it takes a total of six (i.e., $B-1$) wavelengths to facilitate all-to-all independent communications within a pod. Importantly, such wavelength reuse can be applied to all other pods. Namely, the same six wavelengths can be fully and independently reused within the B^2 pods to provide parallel intra-pod transport.

10.2.1.2 Tier-2 WXC Node

The tier-2 switching is performed via a 4-way WXC node, as shown in Fig. 10.3, for the case of $B = 7$. As shown in Fig. 10.3a, the WXC node is south connected to seven ROADM nodes of its pod and north connected to a tier-3 WXC node. The node is also connected to 3 peer east/west WXC nodes via 3 pairs of fibers, in which 2 pairs are for pass-through traffic and 1 pair for traffic being switched to other ports. The key switching element of the WXC node is the 17 × 13 WSS module that can be implemented via commercially available WSS devices, ranging from size 10 × 1 to 17 × 1 (see Fig. 10.3b).

Notice that there are four pairs of parallel fibers for the northbound transport. This is because each edge between the tier-2 and tier-3 WXC node requires a capacity of nearly 4 W to assure the DCN of being congestion free (described in Sect. 10.2.2). Accordingly, for $B = 7$, a WXC node is equipped with a WSS that has 17 input ports (3 + 3 + 7 + 4) and 13 output ports (two east and two west ports are for pass-through traffic only). For the ease of illustration, we delineate in the figure

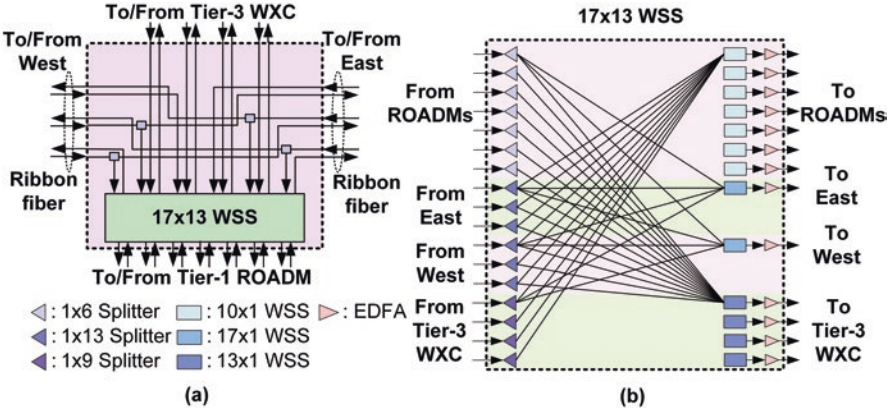


Fig. 10.3 (a) Design block diagram of tier-2 WXC node, (b) Design block diagram of WSS in tier-2 WXC

the exact number of ports of each splitter and WSS while omitting several direct connections between some splitters to WSSs. For example, it requires a 17 × 1 WSS for switching traffic "to east" from any of the 17 input ports.

10.2.1.3 Tier-3 WXC Node

The tier-3 switching for traffic that crosses macro-pods is performed via a 3-way (east, west, and south) WXC node. Since its overall structure is similar to that of a tier-2 WXC node, the designed block diagram is omitted here, and interested readers can refer to [4]. It is worth mentioning that there is a new feature that has been designed at tier 3 for achieving better scalability and fault tolerance purposes. Specifically, recall that each of 7 tier-2 WXCs is connected to a tier-3 WXC via 4 pairs of fiber links. Rather than feeding all 7 × 4 = 28 links from the tier-2 pyramid base into one tier-3 WXC node, the tier-3 WXC node is functionally divided into four smaller-size independent WXC nodes, to which each pair of fiber links is connected. As a result of the division, there are four switching planes at tier 3 (each of which consists of 7 smaller-size WXC nodes) that operate in parallel, thereby offering higher fault tolerance. Further, each main WSS module at each WXC node is reduced in port size to 25 × 13, achieving better scalability.

10.2.2 Edge Capacity and Structure

In this subsection, we are to answer the next design question: how many fiber links are required between any two adjacent optical switching nodes? First, let the *edge* between two adjacent nodes be defined as the inclusion of all parallel fiber links connecting the two nodes. Let W denote the total number of wavelength channels on each fiber link. The *edge capacity* of an edge between two adjacent nodes is defined as the total number of required wavelengths, satisfying an oversubscription ratio of one (i.e., the total output link rate of the first switching node is equal to its total input link rate). Here, we first derive the edge capacities, followed by determining the edge structure, i.e., the number of parallel fiber links on each edge.

It is clear that the determination of all edge capacities depends on the traffic distribution within and outside of pods and macro-pods. Let P_{TL} denote the *traffic locality probability*, which defines the traffic distribution within and outside of a module- pod and macro-pod alike. Specifically, given a source (ROADM) node of a flow that belongs to a macro-pod, P_{TL} is defined as the probability that its destination node falls within the same macro-pod, and $1 - P_{TL}$ the probability outside of the macro-pod. Further, conditional to a given macro-pod, given a source node in a pod within the macro-pod, P_{TL} is defined as the probability that the destination node falls within the same pod, and $1 - P_{TL}$ the probability outside of the pod but within the same macro-pod. Accordingly, a flow is destined to a node within the same pod with probability, $(P_{TL})^2$; is destined to a node in a different pod but within the same macro-pod with probability, $(1 - P_{TL})P_{TL}$; and is destined to a node in a different macro-pod with probability, $1 - P_{TL}$. Once the locality is determined, the destinations are assumed uniformly distributed. Notice that it is highly expected that the normalized traffic destined to any node within the pod/macro-pod is greater than any node outside of it. For the OPMDC prototyping system, we use $B = 7$, and its edge capacity and structure design is based on a modest locality probability, $P_{TL} = 0.5$.

Let $C_H(B, T)$ denote the edge capacity between two adjacent horizontal nodes in tier T, where B is the number of base nodes in a pyramid. Let $C_V(B, T)$ denote the edge capacity between two adjacent vertical nodes at tiers T and $T + 1$, respectively. First, we are to compute $C_H(B, T)$, where $T = 1$, 2, and 3, and $C_V(B, T)$, where $T = 1$ and 2.

Assume that each traffic flow requires a bandwidth of one wavelength channel, and the total number of flows emitting from any ToR switch (or ROADM) is W. Let $F(B, k)$ denote the mean total number of flows from source ROADM node s to destination node d, where $k = 1, 2, 3$ correspond to three cases for the locality of nodes s and d, as stated in Eq. (10.1). For any given source node (s) in a pod, for case I ($k = 1$), there are $B - 1$ possible destination nodes (d) in the same pod; for case II ($k = 2$), there are $B(B - 1)$ destination nodes in different pods of the same macro-pod; and for case III ($k = 3$), there are $B^2(B - 1)$ nodes in different macro-pods. So, $F(B, k)$ can be given as

$$F(B,k) = \begin{cases} \dfrac{W \cdot (P_{TL})^2}{B-1}, & k = 1, \mathit{if}\ \ s, d \in \mathit{same}\,\text{pod}; \\ \dfrac{W \cdot P_{TL} \cdot (1 - P_{TL})}{B(B-1)}, & k = 2, \mathit{if}\ \ s, d \notin \mathit{same}\,\text{pod, and } s, \\ & \quad d \in \mathit{same\,macro} - \text{pod}; \\ \dfrac{W \cdot (1 - P_{TL})}{B^2(B-1)}, & k = 3, \mathit{if}\ \ s, d \notin \mathit{same\,macro} - \text{pod}. \end{cases} \tag{10.1}$$

Recall that all traffic emitted from a ROADM node is broadcast via a 1 × 3 splitter (see Fig. 10.2a) to horizontal ROADM nodes and tier-2 WXC node. Thus, we can directly get $C_H(B, 1) = W$ and $C_V(B, 1) = W$. Next, the capacity $C_H(B, 2)$ needs to accommodate all the traffic from one whole pod to its adjacent pod. Moreover, the capacity is shared for transporting traffic from a tier-2 WXC to $(B-1)/2$ peer WXC nodes. Accordingly, we have

$$C_H(B,2) = F(B,2) \cdot B^2 \cdot \frac{B-1}{2} = \frac{B \cdot W \cdot P_{TL} \cdot (1 - P_{TL})}{2} \tag{10.2}$$

By the same token, $C_H(B, 3)$ at tier 3 needs to accommodate all the traffic from one whole macro-pod to its adjacent macro-pod, and the capacity is also shared for transmitting $(B-1)/2$ sets of traffic. We thus get

$$C_H\left(B,3\right) = F\left(B,3\right)\cdot B^2 \cdot B^2 \cdot \frac{B-1}{2} = \frac{B^2 \cdot W \cdot \left(1-P_{TL}\right)}{2} \tag{10.3}$$

Finally, $C_V(B, 2)$ for a tier-2 WXC is to interconnect its entire pod to any node outside of the macro-pod it belongs to, with a probability of $1 - P_{TL}$. This implies

$$C_V\left(B,2\right) = B \cdot W \cdot (1-P_{TL}) \tag{10.4}$$

Given $P_{TL} = 0.5$ and $B = 7$, we get from Eqs. (10.2, 10.3, and 10.4) that $C_H(7, 2) = 0.875W$, $C_V(7, 2) = 3.5W$, and $C_H(7, 3) = 12.25W$ in OPMDC. This explains the requirements of 4 pairs of parallel fiber links to connect a tier-2 WXC to its tier-3 WXC and 12 pairs of parallel fiber connections to connect two adjacent tier-3 WXC nodes before the division into four switching planes is applied.

10.3 Wavelength Allocation Strategies

OPMDC strives for flexible optical packet-based and circuit-based transport based on three innovative wavelength allocation strategies. They are (1) static pre-allocation, (2) relay-based allocation, and (3) dynamic allocation. While strategies 1 and 2 require no further WSS reconfiguration and thus are best suited for supporting packet-based transport, strategy 3 provides efficient dynamic establishment of new optical paths for circuit-based transport. They are described in the following.

10.3.1 Strategy 1: Static Wavelength Pre-allocation

The first strategy caters to all intra-pod and intra-macro-pod transport, based on static wavelength pre-allocation. Significantly, as will be shown, the static pre-allocation allows all intra-pod and intra-macro-pod transport to be facilitated fully in parallel using only a total of $2(B-1)$ wavelengths. Due to the avoidance of WSS reconfiguration, this strategy is capable of meeting the demand of ultra-low latency for supporting packet-based transport.

In the case of intra-pod communications, as was illustrated in Fig. 10.2b, a small fixed number ($= B-1$) of wavelengths can be pre-allocated and reused to simultaneously support all intra-pod transport for all B^2 pods in OPMDC. Due to incurring no WSS reconfiguration and any other delays, this class of the packet-based transport receives near-zero latency.

For intra-macro-pod communications, since the horizontal mesh connection at tier 2 is the same as that at tier 1, the same wavelength reuse principle (illustrated in Fig. 10.2b) can be applied to tier 2. That is, for OPMDC with $B = 7$, it takes only a total of six (i.e., $B-1$) wavelengths to facilitate all pod-to-pod communications independently within any macro-pod in OPMDC. For example, assume a wavelength is designated for the communications from s-pod to d-pod. Since there are seven ROADM nodes in any pod, the designated wavelength can be used to connect one random ROADM node (say R_1) in s-pod to another random ROADM node (say R_3) in d-pod. Then, all other communications between two different pairs of ROADM nodes from s-pod to d-pod can be established by means of packet relays at nodes R_1 and R_3 (and their ToR switches). Such a relay operation is employed in Strategy 2, which is described in detail in the next subsection.

Further, such wavelength reuse can be applied to all other macro-pods. Therefore, it takes only a total of $B-1$ wavelengths to establish all intra-macro-pod packet-based transport fully in parallel for all B macro-pods in OPMDC. Without any WSS reconfiguration, this class of packet-based transport experiences an ultra-low delay resulting from two (or less) relays at the ToR E-switches (described next).

10.3.2 Strategy 2: Relay-Based Wavelength Allocation

The relay-based wavelength allocation aims to facilitate intra-macro-pod and inter-macro-pod transport using existing optical paths between the source and destination pods located at different pods and macro-pods, respectively. Notice that such existing optical paths for intra-macro-pod transport are to be statically pre-allocated based on Strategy 1. Basically, this relay-based strategy employs a combined *relay* and *aggregation* operation at the source and/or destination pods, referred to as SDRA. The SDRA mechanism allows new flows to be transited using existing optical paths by means of flow relay and aggregation through the horizontal mesh connections in tier 1 at the source and/or destination pods.

The SDRA mechanism can be best explained via an example illustrated in Fig. 10.4. Suppose there is a new traffic flow from node S in s-pod to node D in

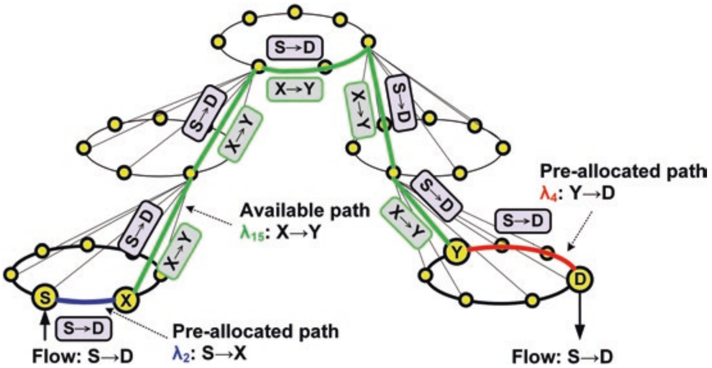


Fig. 10.4 Relay-based wavelength allocation for low-latency packet-based transport

d-pod, and there exists an available optical path, λ_{15}, for the traffic flow between nodes X and Y in s-pod and d-pod, respectively. S-to-D packets are first sent to X through the horizontal pre-allocated path (λ_2) and then relayed at X's ToR switch. These packets are optical-to-electrical (OE) converted, aggregated with X-to-Y packets, and together delivered through available optical path λ_{15}. Upon having arrived at node Y in d-pod, S-to-D packets are OE-converted again and transported together with Y-to-D packets through horizontal pre-allocated path λ_4. The S-to-D packets are finally dropped after reaching node D. Notice that if node S = X, the relay/aggregation operation is not invoked in the source pod. Likewise, if node D = Y, the operation is avoided in the destination pod.

The SDRA mechanism results in high utilization due to packet aggregation. In addition, the mechanism offers low-latency transport owing to the avoidance of WSS reconfiguration. The price paid is only no more than two additional hops of E-switch processing delay. Simulation results [4] show that employing the SDRA mechanism yields a substantial throughput improvement from 42.5 to 87.9%, due to taking advantage of available optical paths.

10.3.3 Strategy 3: Dynamic Wavelength Allocation

The third strategy aims to establish new optical paths for inter-macro-pod transport, based on dynamic wavelength allocation. Due to the need for WSS reconfiguration that causes a few millisecond delay, this strategy is best suited for supporting circuit-based transport.

The dynamic wavelength allocation problem can be formally defined as: given a set of available wavelengths and circuit flows to be served, the problem is to assign the wavelengths to a maximum number of circuit flows, subject to being free from wavelength contention at any fiber link. To maximize the throughput, the wavelength allocation problem boils down to the proper determination of the order by which the circuit flows are assigned wavelengths. Specifically, the flow that contends with a higher number of flows should be served first in order to reduce the contention probability. Accordingly, we have proposed a heuristic algorithm, called the most contentious first (MCF) [4]. The algorithm first ranks each flow according to the total number of all other flows in its *most congested* link of the path the flow travels through. The MCF algorithm then assigns wavelengths to the flows sequentially in descending order of the flow ranks.

The performance of the MCF algorithm was evaluated via experimental testbed results (see Sect. 10.4) as well as simulation results [4]. In particular, the simulation results show that OPMDC achieves 95.8% throughput under $P_{TL} = 0.5$ and 80% throughput even under poor traffic locality, $P_{TL} = 0.3$.

10.4 Prototype and Performance Assessment

In this section, we present our OPMDC prototype system and give an assessment of its performance with respect to scalability, power consumption, wiring complexity, fault tolerance, and mean packet latency.

10.4.1 OPMDC Prototyping System

We have built a prototyping system of OPMDC with $B = 7$ using seven ROADM nodes, in which the WSS module is based on the CoAdna's LightFlow™ digital LC platform [14]. Each ROADM node is directly connected to a Pica8 P-3295 ToR switch that is compliant with the OpenFlow software-defined networking (SDN) [17] standard interface. Each ToR switch provides 48 10-Gb/s ports. Among them, 16 ports are populated with 16 10-Gb/s DWDM transceivers associated with 16 wavelength channels, respectively. We have implemented an OpenFlow-1.3 [18] SDN controller based on a Ryu 3.9 open source system. The SDN controller system runs the wavelength allocation algorithms and, in turn, performs the reconfiguration and control of ROADM nodes and ToR SDN switches.

Each ROADM node is equipped with a Raspberry Pi [19] embedded firmware system that operates with a ROADM controller, developed under a Debian 7.2 Linux kernel-based operating system. The controller is primarily responsible for the real-time reconfiguring and monitoring and periodic reporting of the health of the hardware devices, e.g., WSS and EDFA. Specifically, the controller performs the configuration of its optical WSS whenever it receives a control message from the SDN controller after making new wavelength allocation decisions. The control message includes information such as wavelength channel IDs and WSS ports. For the time being, the communication between the SDN controller and ROADM controller is facilitated through a TCP socket interface.

10.4.2 Performance Assessment

An overall performance assessment of OPMDC is made in the following:

Scalability: OPMDC supports up to B^3 ROADM nodes. Each ROADM provides dynamic switching of up to 96 (W) wavelengths under the 50-GHz channel spacing. Each wavelength channel can currently accommodate a capacity (C) of 100 Gb/s. Thus, OPMDC can support a maximal capacity of B^3WC, which is 3.3 Pb/s with $B = 7$, or 7 Pb/s with $B = 9$. As such, OPMDC provides high and scalable bandwidth for both cloud and enterprise DCNs.

Power Consumption: Power consumption is often estimated on a per port basis. For E-switch-based DCNs, ARISTA has claimed to provide industry leading power efficiency, achieving a typical power consumption of 5 watts per 40-GbE port in its 7050X Series of products. For the OPMDC that supports 48 wavelengths under the 100-GHz channel spacing, the typical power consumptions of a ROADM and a tier-2/tier-3 WXC node, dominated by the EDFA module, are around 15 watts and 12 watts, respectively. With $B = 7$, there are $7^3 = 343$ ROADM nodes, 49 tier-2 WXC, and 7 × 4 = 28 tier-3 WXC nodes; and each WXC node is populated with 13 pairs of WSS/EDFA modules. The total power consumption becomes (343 × 15w) + (49 × 13 × 12w) + (28 × 13 × 12w) = 17, 157 watts. For OPMDC with 343 48-port ROADM nodes, i.e., 16,464 ports altogether, the typical power consumption is around 1 watt per port. Thus, OPMDC currently achieves five times as power efficient as E-switch DCNs. In our future work, we aim at designing an optical DCN architecture using fewer EDFAs, for achieving more power saving.

Wiring Complexity: The wiring complexity and overall cost of DCNs are directly proportional to the total number of distinctive optical fibers interconnecting all switching nodes. We now draw a comparison of the total number of fibers required between OPMDC and a typical DCN, populated with the same port number. We consider $B = 7$ in this comparison. Recall that OPMDC supports $7^3 = 343$ ROADM nodes, and each ROADM connects to the 6 40-Gb/s ports of a ToR switch. To support the same number of 40-Gb/s ports, i.e., 343 × 6 = 2058 ports, a typical leaf-and-spine DCN demands 2058 fibers altogether. Due to the use of ribbon fibers for all the horizontal connections, for example, each tier-1 pyramid pod needs 7 horizontal and 7 vertical fibers. Therefore, OPMDC requires a total of 959 fibers, including (7 + 7) × 49 fibers for 49 tier-1 pods, [7 + (4 × 7)] × 7 fibers for 7 tier-2 macro-pods, and 7 × 4 fibers for 4 planes of tier-3 nodes. Compared to a typical spine-and-leaf DCN, OPMDC achieves a reduction of more than half in wiring complexity.

Fault Tolerance: Existing DCNs achieve fault tolerance and high availability by means of switch redundancy and path diversity. However, the price paid is high wiring complexity and poor resource efficiencies. For designing large-scale but manageable cost DCNs, we incorporate additional resiliency means. OPMDC achieves fault tolerance especially for two lower tiers solely by using highly reliable WSS-based switching modules and periodic monitoring and fast recovery of hardware devices. For tier 3, as explained in Sect. 10.2.1, OPMDC additionally employs the segregation of the tier-3 WXC backbone into four identical but scale-down switching planes. The division results in path diversity, thus offering an additional level of reliability.

Cost: Due to not yet being prevalent on the market, the costs of the major components (purchased in small quantity) are currently high. For example, a 1 × 4 WSS costs around US$3000, and a DWDM 10G transceiver with a channel spacing of 100 GHz costs around a range of US$300–$400. Nevertheless, the costs can be greatly reduced if the optical components are produced in large quantity. Furthermore, with the advances in emerging photonic integrated circuit (PIC) technology, these optical switching nodes can be PIC-based designed and implemented, resulting in significant cost reduction.

10.4.3 Packet Latency Performance

We take the direct measurements of packet latency from experiments on the prototype system. In the experiments, we send a large video file via python-based TCP socket, from the source server (S) (socket client) to the destination server (D) (socket server). The buffer size of the socket is set to be 60 Kbytes. To measure the latency for both circuit- and packet-based transport, we adopt two types of flows: c-flow (2000 S-to-D transmissions of 60-Kbyte packets) and p-flow (50 S-to-D transmissions of 60-Kbyte packets). In particular, the c-flow requires the setup of a new optical path based on the MCF algorithm (allocation Strategy 3). The p-flow is transported via existing optical paths based on the SDRA algorithm (Strategy 2). Moreover, to measure the one-way S-to-D packet delay, we implement a simple ACK program at the D-node, sending an acknowledgment packet to the S-node as soon as a 60-Kbyte packet has been received from its TCP socket. The round-trip time is measured at the S-node upon having received the ACK packet. Since the ACK packet undertakes the same overhead as that of its video packet, the S-to-D packet latency is calculated as half of the round-trip time. The measurement of mean latency performance is summarized in Table 10.1.

For c-flow, the ROADM reconfiguration delay consists of two parts: firmware response time and WSS response time. The firmware response time, including Ethernet delay, is 10 ~ 12 ms. The single-channel (sc) and multichannel (mc) WSS response time, including WSS switching time and universal asynchronous receiver/transmitter (UART) delay, are 3 ms and 35 ms, respectively. Therefore, the total reconfiguration delay is 13 ~ 15 ms for a single wavelength and 45 ~ 47 ms for

Table 10.1 Mean packet latency performance of the OPMDC prototyping system

Task	Description	Time
c-flow: Circuit-based transport (requiring new optical paths using Strategy 3)		
A. ROADM reconfiguration	Firmware response time (12 ms)	15 ms(sc); 47 ms(mc)
	WSS response time (single-channel/multichannel) (3 ms/35 ms)	
B. Run time	MCF algorithm (48 × 7 flows)	0.184 ms
C. ToR switch configuration	Packet-in/packet-out (max. two times)	6.47 ms
	Flow entry setup in 2 switches	
D. Packet transmission	Packet size = 60 Kbytes	2410 ms
	Flow duration = 2000 packets	
Mean S-to-D packet latency = (A + B + C + D)/2000		1.232 ms
p-flow: Packet-based transport (using existing optical paths based on Strategy 2)		
E. Run time	SDRA algorithm (48 × 7 flows)	0.07 ms
F. ToR switch configuration	Packet-in/packet-out (max. two times)	6.54 ms
	Flow entry setup in 4 switches	
G. Packet transmission	Packet size = 60 Kbytes	79.2 ms
	Flow duration = 50 packets	
Mean S-to-D packet latency = (E + F + G)/50		1.716 ms

multiple wavelengths. For p-flow, the configuration of four SDN-enabled ToR switches takes place in parallel, with the result that it takes no more than two times of packet-in/packet-out as that for transporting c-flow. With the SDRA mechanism employed, the bottleneck of latency lies in populating the flow tables in the ToR switches.

10.5 Conclusions and Research Directions

In this chapter, we have presented a novel optical pyramid DCN architecture and its prototype, OPMDC, including the design and implementation of its three types of WSS-based optical switching nodes. After introducing a traffic locality parameter, P_{TL}, we derived the edge capacities and determined the edge structure that satisfies the need of OPMDC for being bottleneck free. Owing to the pyramid architecture, OPMDC enables powerful wavelength reuse and broadcast capability. Specifically, we proposed three wavelength allocation strategies, boasting flexible ultra-low-latency optical packet-based transport and high-throughput circuit-based transport. We demonstrated experimental testbed results to justify that OPMDC achieves high and scalable bandwidth, low latency, high fault tolerance, and reduced power consumption and wiring complexity.

We are currently undertaking a number of research work related to OPMDC. First, recall that one of the distinguishing features of OPMDC is its modular and incremental design of the architecture. As a result, such a design makes OPMDC highly flexible to serve the needs for providing different-scale data centers at different locations of networks. Thus, one of our current research tasks is to design micro/mini-data centers that are constructed based on a few pods or macro-pods of OPMDC. These smaller-scale data centers are targeted at facilitating near-zero-latency mobile computing at the edge of 5G mobile networks. The second research task is to design broadcast-based traffic control mechanisms that cater for supporting parallel-processing cloud applications, such as Big Data computing. Last but not least, we have continually been refining the OPMDC architecture and the internal designs of optical switching nodes in an effort to reduce the number of power-hungry EDFA devices for achieving greater power efficiency.

References

1. N. Bitar, S. Gringeri, and T. Xia, “Technologies and Protocols for Data Center and Cloud Networking,” *IEEE Commun. Mag.*, vol. 51, no. 9, Sep. 2013, pp. 24–31
2. Y. Liu, J. Muppala, M. Veeraraghavan, and D. Lin, *Data Center Networks: Topologies* (Springer, Architectures and Fault-Tolerance Characteristics, 2013)
3. C. Kachris, K. Bergman, and I. Tomkos, *Optical Interconnects for Future Data Center Networks* (Springer, 2013)

4. M. Yuang, P. Tien, H. Chen, W. Ruan, S. Zhong, J. Zhu, Y. Chen, and J. Chen, "OPMDC: Architecture Design and Implementation of a New Optical Pyramid Data Center Network," *IEEE/OSA Journal of Lightwave Technology*, vol. 33, no. 10, May 2015, pp. 2019–2031
5. Z. Li, Z. Guo, and Y. Yang, "BCCC: An Expandable Network for Data Centers," *IEEE/ACM Trans. Networking*, vol. 24, no. 6, Dec. 2016, pp. 3740–3755
6. N. Han, Y. Chung, and M. Jo, "Green Data Centers for Cloud-Assisted Mobile Ad Hoc Networks in 5G," *IEEE Network*, vol. 29, no. 2, March/April 2015, pp. 70–76
7. C. Kachris, K. Kanonakis, and I. Tomkos, "Optical Interconnection Networks in Data Centers: Recent Trends and Future Challenges," *IEEE Commun. Mag.* **51**(9), 39–45 (Sep. 2013)
8. T. Ban, H. Hasegawa, K. Sato, T. Watanabe, and H. Takahashi, "A Novel Large-scale OXC Architecture and an Experimental System that Utilizes Wavelength Path Switching and Fiber Selection," *Opt. Express* **21**(1), 469–477 (Jan. 2013)
9. J. Homa and K. Bala, "ROADM Architectures and Their Enabling WSS Technology," *IEEE Commun. Mag*. **46**(7), 150–154 (June 2008)
10. Y. Li, Li Gao, G. Shen, and L. Peng, "Impact of ROADM Colorless, Directionless, and Contentionless (CDC) Features on Optical Network Performance," *IEEE/OSA J. Optical Communications and Networking*, vol. 4, no. 11, Nov. 2012, pp. 58–67
11. C. Kachris and I. Tomkos, "A Survey on Optical Interconnects for Data Centers," *IEEE Communications Surveys and Tutorials* **14**(4), 1021–1036 (2012)
12. Jordi Perello, et al., "All-Optical Packet/Circuit Switching-Based Data Center Network for Enhanced Scalability, Latency, and Throughput," *IEEE Network*, vol. 27, no. 6, Nov./Dec. 2013, pp. 14–22
13. K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, X. Wen, and Y. Chen, "OSA: An Optical Switching Architecture for Data Center Networks With Unprecedented Flexibility," *IEEE/ACM Trans. Networking* **22**(2), 498–511 (April 2014)
14. CoAdna, "50GHz Wavelength Selective Switch- High performance with integrated functionalities in a small footprint," http://www.coadna.com/2/products.html#_top
15. M. Yuang, J. Yang, H. Chen, and P. Tien, "Fault-Tolerance Enhanced Design and Analyses for Optical Pyramid Data Center Network (OPMDC)," *Optical Fiber Communication (OFC) Conference*, 2016
16. S. Zhong and Z. Zhu, "Distributed Optical Switching Architecture for Internal Data Center Networking," USA Provisional Patent, OMB 0651–0032, Jan. 2014
17. P. Goransson and C. Black, *Software Defined Networks: A Comprehensive Approach* (Morgan Kaufmann, San Francisco, 2014)
18. OpenFlow Switch Consortium and Others. OpenFlow Switch Specification Version 1.4.0. 2013. https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/openflow/openflow-spec-v1.4.0.pdf, Jan. 2014
19. Raspberry Pi. Available online: http://www.raspberrypi.org

Part III
Technologies for Optical Switching in Data Centers

Chapter 11
Commercial Optical Switches

Qirui Huang

11.1 Introduction

The consistent increase in data traffic of various services such as social networking, online gaming, e-commerce, and CDN (content delivery network) in the Internet demands even higher interconnect bandwidth in telecommunication network and higher switching capacities for intra-datacenter network. This trend brings new challenges to legacy switching and routing devices of the networks. Although optical fiber links have been widely deployed to increase the transmission capacity between switching nodes in datacenters, the switching/routing functions are performed by electrical packet switches that require a large number of optical/electronic/optical (O/E/O) conversions. Another major concern is the power consumption of these devices. For instance, a 25-Tbit/s core router (like Juniper TX Matrix Plus) with 128 40-Gb/s I/O ports needs 768 O/E/O converters and consumes over 12,000 Watts power [1]. Using a large number of electronic packet switches not only causes severe power consumption issues but also increases cabling complexity and operating cost for a modern datacenter. In order to mitigate the pressure brought by electrical switching, many optical switching techniques have been proposed, developed, and gradually deployed in telecommunication network and datacenters [2–4].

Optical switching exhibits many potential advantages in terms of throughput capacity, power consumption, flexibility, and so on. With optical switch devices, a source signal can be maintained in the optical domain for high-speed transmission and routing until it reaches destination. This process does not need any expensive and energy-intensive O/E/O convertors. Although a wide range deployment of optical switches is still in immaturity, with the progress in material science and fabrication, more and more advanced optical switch devices will become commercially available in future.

Q. Huang (✉)
Temasek Lab, Singapore University of Technology and Design, Singapore, Singapore
e-mail: qirui_huang@sutd.edu.sg

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
DOI 10.1007/978-3-319-61052-8_11

Optical switch technologies can be categorized in different ways [5]. Based on switching dimension, there are optical space, time, and wavelength domain switches; according to switching granularity, there are optical circuit, burst, and packet switches; depending on the fabrication material and physical effect, there are polymer-based, semiconductor-based, fiber-based, liquid crystal-based, electro-optic, thermo-optic, and opto-mechanical switch devices. In this chapter, we cover the following optical switches that have been widely developed for commercialization and deployment in telecom and datacom networks, with focuses on their principle, structure, performance, and applications:

- Microelectromechanical system optical switch
- Beam-steering-based optical switch
- Liquid crystal optical switch
- Electro-optic optical switch
- Semiconductor optical amplifier-based switch
- Thermo-optic optical switch

11.2 Microelectromechanical System-Based Optical Switch

Microelectromechanical system (MEMS) technology has been widely used in various areas including biochemistry, aeronautics, manufacturing, display, sensors, and telecommunications. Even though used for different applications, MEMS devices have common features such as small volume, easy integration, and low power consumption. MEMS-based optical switches are the most common and mature commercial opto-mechanical free-space devices [6].

There are many designs of MEMS-based optical switches, which can mainly be classified into two categories according to their working principles. The first is based on the manipulation of the propagation direction of the light by reflecting structures. Typically, such MEMS optical switches consist of an array of tiny micro-mirrors and moving parts arranged on special substrate material with CMOS-integrated circuits. Figure 11.1 (left) illustrates the structure of a micro-mirror fabricated by MEMS technology, which uses a two-axis tilting structure [7]. The micro-mirror is usually fabricated by depositing dielectric layers on a substrate, which is typically a single crystal or polycrystalline silicon substrate, and then etching selected material using photolithographic technology. The reflector surface of the mirror is often made of thin metal coating such as aluminum or gold. By applying current on the moving parts, the mirror and the surrounding ring can rotate by a certain angle and reflect the beam of light from an input port to a specified output port. This structure can easily be batch-fabricated in the form of arrays of hundreds of mirror elements in a single chip of a few centimeter square as shown in Fig. 11.1 (right) [8]. The second category is based on the adjusting of the phase of light through mechanical motion in MEMS to achieve switching function due to diffractive or interference effects. Examples of this are the mechanical anti-reflection

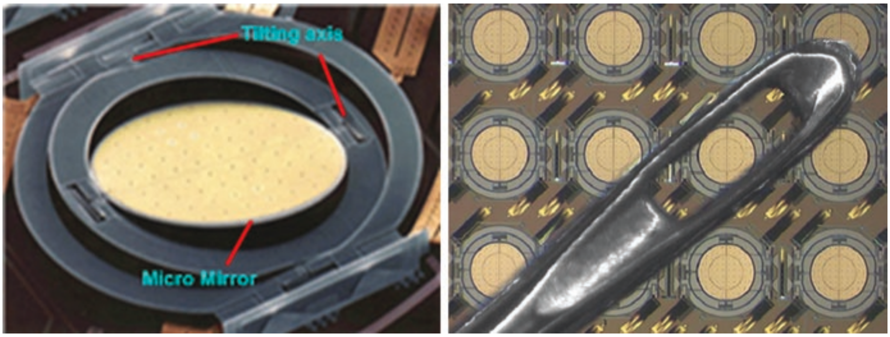
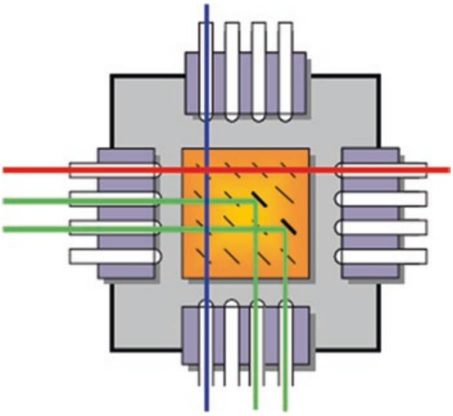


Fig. 11.1 Micro-mirror structure (*left*) and mirror array (*right*) of MEMS

Fig. 11.2 2D MEMS switch used for OADM



switch [9] and the tunable grating structure-based switch [10]. Both categories make use of motion of micro-mirrors to realize switching and share the common features of MEMS devices, while the first category is more widespread for commercialization due to its advantage in batch fabrication.

A commercial MEMS optical switch consists of MEMS arrays integrated on a single chip, driving circuits, control software, and input/output ports. MEMS arrays are usually built in two-dimensional (2D) and three-dimensional (3D) configurations [5]. In 2D MEMS, shown in Fig. 11.2 [11], light beams from the input ports are steered to the desired output ports by the corresponding switching elements within a single plane. Although a MEMS switch with a 2D structure is easy to control, a $N \times N$ configuration requires N^2 micro-mirrors, and the switch size is limited by the number of light paths and achievable area of micro-mirrors. 256 × 256 2D MEMS-based switch fabric has been reported in research [8]. For commercialization, 16 × 16 2D MEMS has early been built and used as optical add/drop multiplexer (OADM) for optical transmission networks [12].

In order to support large-scale switching, MEMS optical switches with 3D structure have been proposed [13, 14]. The architecture of a $N \times N$ 3D MEMS optical switch, which consists of a pair of two-axis (2D) tilting mirror arrays and optical fiber collimator arrays, is illustrated in Fig. 11.3 [13]. The light beams from the input ports are collimated by input collimators onto the first tilting micro-mirror array, which then reflects the light beams to the second micro-mirror array that redirects the beams to the corresponding output ports through the output collimators. Since the tilting range of each micro-mirror in the first array can continuously cover all the micro-mirrors of the second array, the light beams from each input port can reach any output port by precisely adjusting the tilting angles of micro-mirrors in two arrays, yielding a non-blocking switching in this structure. Compared with 2D structure, the 3D structure allows the switch to achieve even larger scalability. One of the challenges of 3D MEMS optical switch is the fact that the system requires complex control software to coordinate operations of thousands of micro-mirrors.

Three-dimensional MEMS optical switches have been commercialized by several vendors and deployed for circuit-switched networks in datacenters by service providers. An example of 320 × 320 optical circuit switch is shown in Fig. 11.4 [15]. The switch includes eight mirror drivers controlled by FPGA processors to provide appropriate voltages to each MEMS mirror. It is capable of monitoring status of existing input-output connections by tapped light from input and output ports and

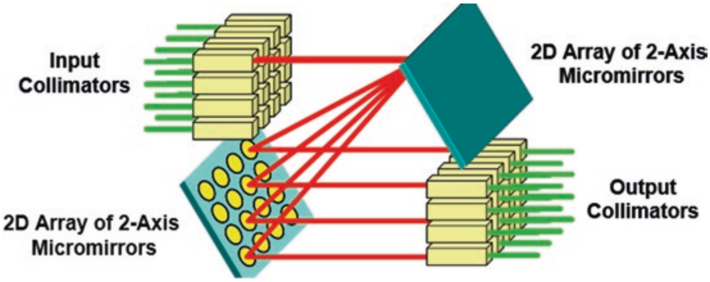
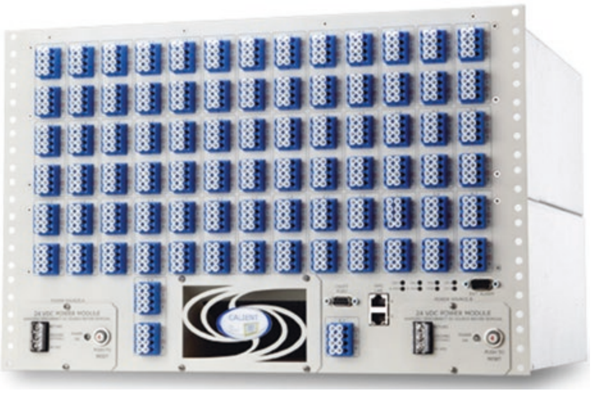

Input Collimators
2D Array of 2-Axis Micromirrors
Output Collimators
2D Array of 2-Axis Micromirrors

Fig. 11.3 3D MEMS optical switch structure

Fig. 11.4 320 × 320 optical circuit switch by Calient



optimizing the connections by adjusting the voltage to each MEMS mirror. The switch can achieve less than 3.5 dB (typical) insertion loss and 50 milliseconds (ms) switching speed in a wavelength range of 1260–1630 nm with only 45 Watts power consumption.

11.3 Beam-Steering Optical Switch

Unlike 3D MEMS-based optical switch, the DirectLight beam-steering technology patented by Polatis is another 3D-based solution for scalable optical switches. The principle of beam-steering technology is illustrated in Fig. 11.5 [16], where two 2D standard fiber collimator arrays are faced. The light beams from the left fiber collimator array are directly coupled into the fiber collimators of the right array. The left fiber collimators are controlled by individual 2D piezoelectric actuators such that the beams can be steered in two angular dimensions for covering all the fiber collimator on the right array. Position sensors are used to monitor the pointing angles of the actuator and to provide feedbacks to the digital position control loop to obtain optimum target positions.

Figure 11.6 shows a 384 × 384 optical circuit switch based on beam-steering technology by Polatis [17]. The switch features 25 ms switching time with 2.5 dB

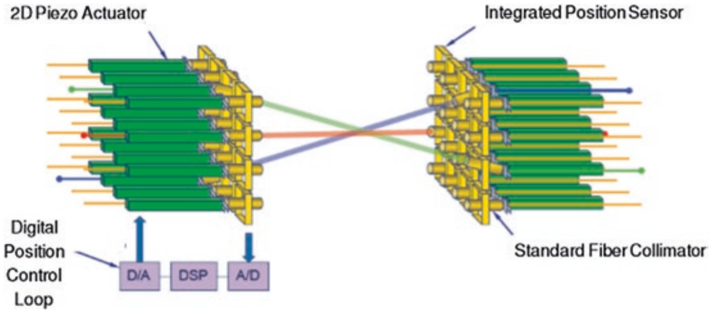
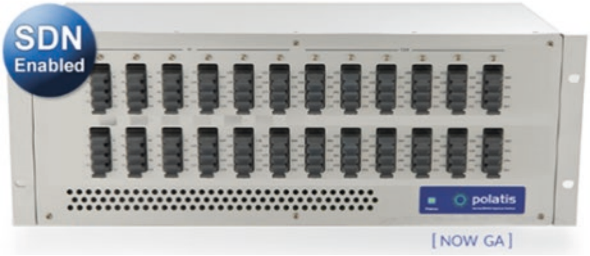

2D Piezo Actuator
Integrated Position Sensor
Digital Position Control Loop
D/A
DSP
A/D
Standard Fiber Collimator

Fig. 11.5 DirectLight beam-steering switching technology by Polatis

Fig. 11.6 384 × 384 SDN-enabled optical circuit switch based on beam-steering technology



insertion loss and 70 dB optical crosstalk isolation between channels. In the control plane, it provides SDN (software-defined networking)-based management interfaces that can work with any SDN controllers to enable flexible service provision and fiber layer connectivity.

11.4 Liquid Crystal Optical Switch

Liquid crystal (LC), which has been widely used in displays, is another promising technique of optical switching [18–21]. In LC materials, molecules have a certain average relative orientation. By applying modest voltage across the LC material, the molecular alignment of the LC will be changed accordingly. This leads to variation in the optical properties such as polarization or refractive index of the LC material.

There are many designs of LC optical switches, which can mainly be classified into two types: polarization based and refractive index based [5]. An example of 1 × 2 polarization-based switch is shown in Fig. 11.7, where a light beam at the input is divided into two polarization beams by a polarization beam splitter and each beam is then redirected to a LC-based polarization rotator. When there is no voltage applied on the LC cells, these two components will recombine at one output via a second polarization beam splitter; when the LC cells are biased with appropriate voltages, the polarizations of the two beams will be rotated orthogonally and coupled to another output via a second polarization beam splitter. Although this kind of switch is polarization independent, since the two beams of incoming light may experience different paths through the switch, there will be difference in optical loss for the two polarization components. This can result in polarization-dependent loss (PDL). The crosstalk is another challenge of this kind of switch. Thus polarization-based LC switches usually have small sizes. Although a large switching size could be implemented by interconnecting a number of switch units, this is only available at the research level.

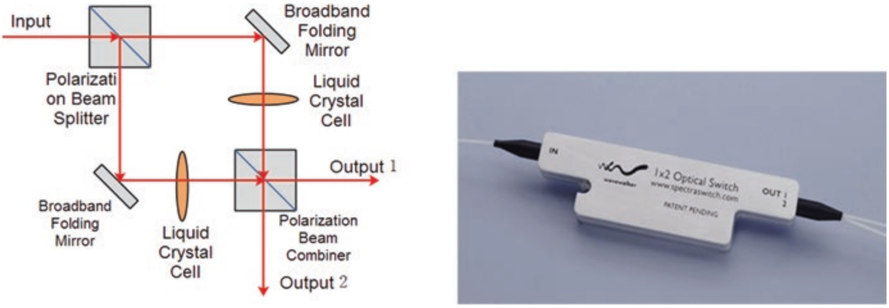


Fig. 11.7 Structure (*left*) and sample (*right*) of 1 × 2 LC optical switch

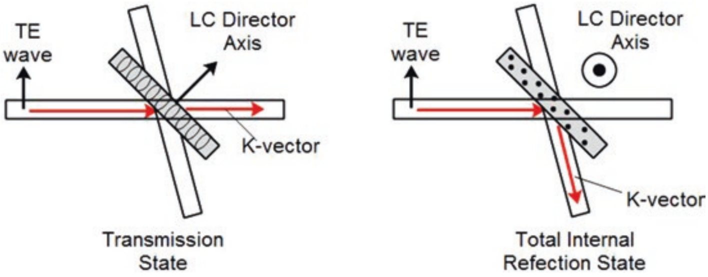


Fig. 11.8 Principle of refractive index-based LC switch

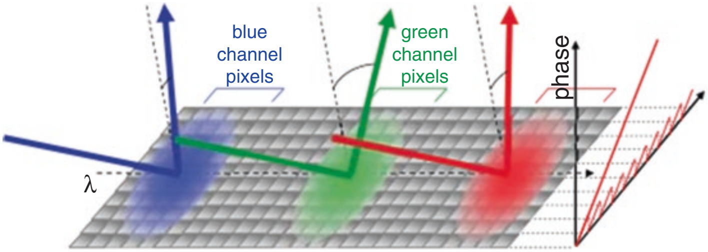


Fig. 11.9 Principle of LCoS technology [23]

The birefringence of LC crystal can also be used to realize another switching function, which is referred to as refractive index-type LC switch. The refractive index of an LC crystal can be varied by controlling the alignment of the LC molecules relative to the propagating direction of the light [18]. As shown in Fig. 11.8, by a careful design, the change of refractive index could facilitate transmission or total internal reflection (TIR) of the light on the LC crystal surface to achieve a switching function in free space [5].

The optical properties in reflection of LC can be further explored with silicon technology to implement wavelength selective switches (WSS) for larger scalability. Liquid crystal on silicon (LCoS) technology is another application of LC in optical switching [22]. As shown in Fig. 11.9, this technology makes use of 2D pixel array on LCoS plane to control the reflection of light beams by producing a linear optical phase retardation in the intended deflection direction [23].

A schematic design of LCoS-based WSS is shown in Fig. 11.10. Input WDM light of random polarizations is first converted into linearly polarized light by polarization diversity optical components and then reflected to conventional grating for wavelength diffraction by a cylindrical mirror. The beams of different wavelengths are fed to corresponding pixel areas on the LCoS plane that will reflect the beams back to the intended output ports through the cylindrical mirror and imaging optics. Since each beam is steered independently, the switch can obtain any combination of these beams at any output. Based on the LCoS switching technology, several wavelength selective switches have been proposed [24–27]. An example of

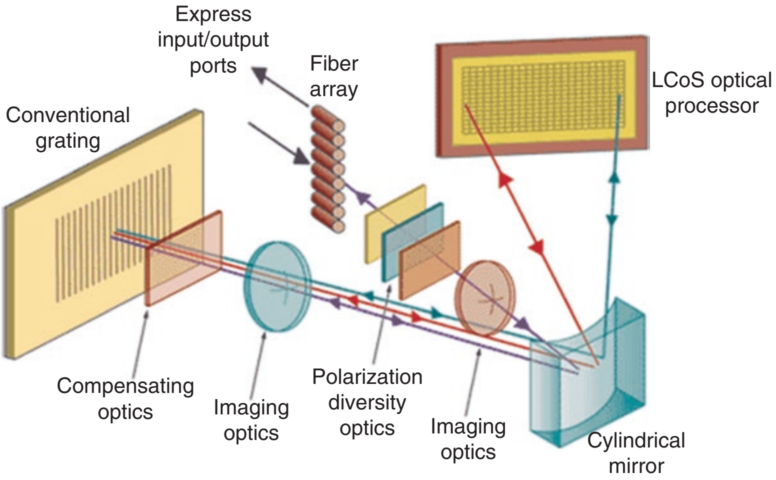


Fig. 11.10 Configuration of LCoS-based WSS [23]

Fig. 11.11 1 × 20 wavelength selective switch based on LCoS by Finisar

LCoS-based 1 × 20 WSS from Finisar is show in Fig. 11.11, which features dynamic channel width control with 6.25 GHz center resolution and 12.5 GHz width resolution [28].

11.5 Electro-optic Switch

Electro-optic (EO) switch is based on the electro-optic effect in which physical properties such as the refractive index of the waveguide material are changed when an electric field is applied. These switches can be implemented using liquid crystals, semiconductor optical amplifiers, waveguide Bragg grating, and lithium niobate ($LiNbO_3$) [29]. We focus on the $LiNbO_3$ optical switches in that they are widely commercialized, thanks to its ultrafast electro-optic effect response time (a few nanoseconds), good optical performance, and low power consumption. Besides

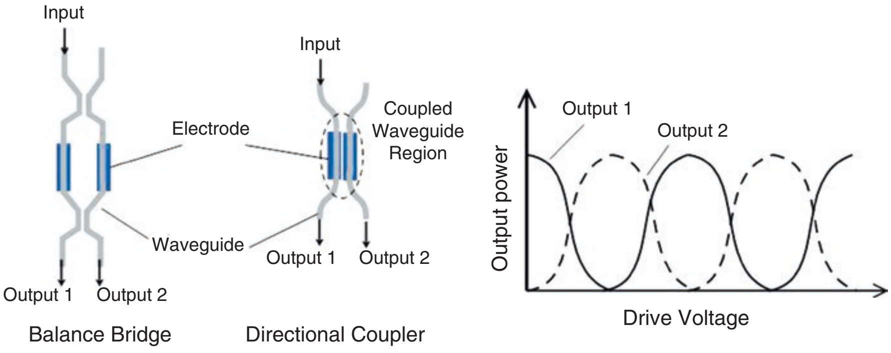


Fig. 11.12 Structures of interferometric $LiNbO_3$ switches

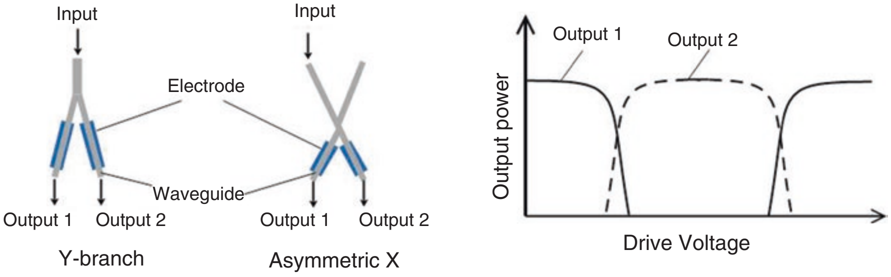


Fig. 11.13 Structures of digital $LiNbO_3$ switches

optical switches, $LiNbO_3$ has been used for implementing various optical modulators for high-speed long-haul transmission [30].

$LiNbO_3$ optical switches are based on the phase modulation and interference phenomena in waveguide structure when external electric field is applied. There are many structural designs of $LiNbO_3$ optical switches [31–33], which can be mainly categorized into interferometric and digital types based on their switching characteristics. Figure 11.12 depicts some examples of interferometric structures, where the curve of their output optical power versus voltage applied is like a periodic sinusoid. The switch operates at optimum state when the output optical power reaches a peak point by a certain voltage value. For digital type, as shown in Fig. 11.13, the output optical power can be maintained at the peak point in a certain voltage range. This characteristic is desirable as the switches can operate at the optimum state without being affected by drive voltage variations.

EO-based $LiNbO_3$ optical switches work on the basis of phase modulation of the light in waveguides. They are capable of 100 ns switching speed. The extinction ratio (ER) of the output signal is not quite good, and it can become even worse as the switch size increases. In addition, the insertion loss is another factor limiting the scalability. As shown in Fig. 11.14, the insertion loss of a 2 × 2 $LiNbO_3$ optical switch is about 3 dB, while the insertion loss of the 8 × 8 configuration increases to

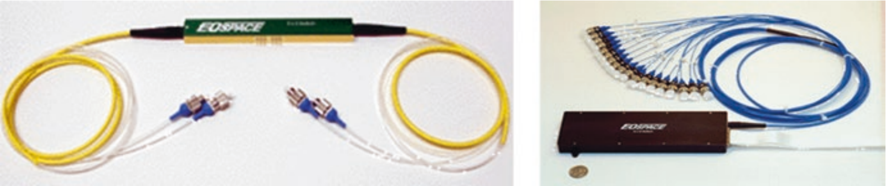


Fig. 11.14 2 × 2 (*left*) and 8 × 8 (*right*) non-blocking $LiNbO_3$ optical switch by EOSPACE

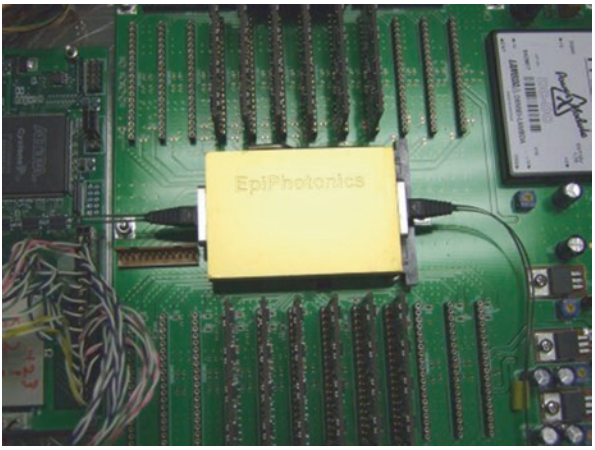


Fig. 11.15 1 × 16 PLZT optical switch by EpiPhotonics

10 dB. To date, commercial EO-based $LiNbO_3$ switches can reach a maximum size of 16 outputs for the 1 × N configuration and 8 outputs for the $N \times N$ configuration, respectively. Another material such as lead lanthanum zirconate titanate (PLZT) can be used for fast optical switching due to its higher EO effect compared to $LiNbO_3$. Figure 11.15 shows a 1 × 16 PLZT EO-based switch by EpiPhotonics. The switch can achieve 10 ns switching speed driven by FPGA controller.

11.6 Semiconductor Optical Amplifier-Based Switch

Research in semiconductor optical amplifier (SOA) has been active for a long time. SOA-based devices have a wide range of applications such as signal amplification/ regeneration, all-optical wavelength conversion, and switching in optical networks [34–37].

A conventional SOA has a double hetero-structure design where an active region is sandwiched by an n-type and p-type cladding layers. An SOA works based on the stimulated emission of the incident light when it passes through the active region which is driven into the positive gain regime by the injected current. As discrete components, SOAs are often used to amplify optical signal in their linear operation

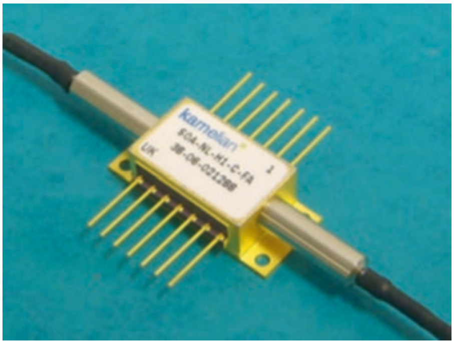


Fig. 11.16 An SOA component for *on*/*off* optical switching by Kamelian

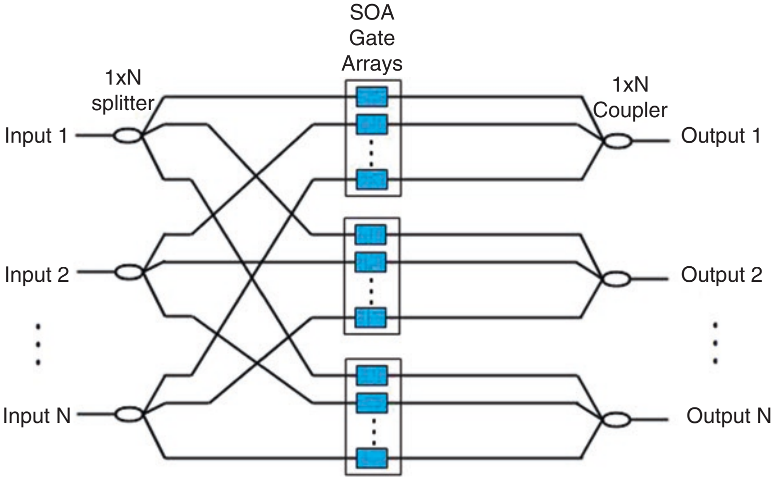


Fig. 11.17 Broadcast-and-select optical switching matrix based on SOA gates

mode. Figure 11.16 shows an SOA component manufactured by Kamelian, which has 10 dB fiber-to-fiber gain and 25 ps (picoseconds) gain recovery time. When no electrical current is present, incident light is blocked by the SOA; when sufficient electrical current is applied, incident light passes through the SOA with gain. Thus, SOAs are also good *on*/*off* gates and have been extensively researched for fabricating many large-size matrices for optical switching and routing [38–40].

The broadcast-and-select architecture is a typical optical switching matrix based on SOA gates shown in Fig. 11.17. The switch architecture consists of N inputs and N outputs. Each input and output is connected to a 1 × N splitter and N × 1 combiner, respectively. The input optical signals are first split into N copies, each of which is fed to an array of SOA-based gates, but only one copy from different inputs will be selected by an array and transmitted to the desired output. This architecture is easy to be implemented with discrete optical components, and several variants have been

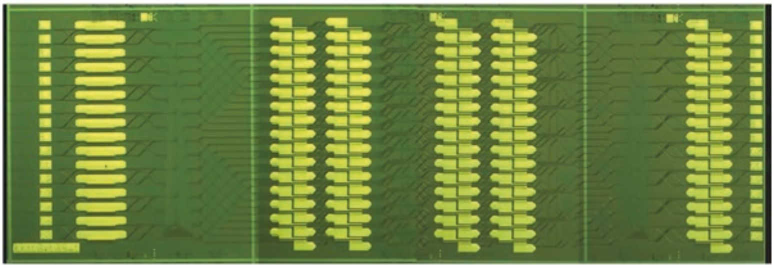


Fig. 11.18 16 × 16 monolithically integrated switch matrix-based cascaded SOA arrays

proposed to address issues such as insertion loss, OSNR, and scalability for large-scale optical packet switching [41–44]. Figure 11.18 shows a 16 × 16 monolithically integrated switch matrix constructed by multistage 2 × 2 SOA-based switching units [45]. However, the main concern for commercialization is the cost due to the large number (N^2) of SOA required.

11.7 Thermo-optic Switch

Thermo-optic (TO) switches are based on the thermo-optic effect in optical waveguide material that utilizes the temperature dependence of the refractive index to achieve switching functionality [46]. These switches are generally fabricated on silicon and polymer materials where a waveguide structure is produced by either chemical vapor deposition (CVD) or flame hydrolysis deposition (FHD) technique and, then, is cladded by metal films that are used as heating electrodes. By applying suitable current to the heating electrodes, the refractive index of the waveguide is varied, leading to the coupling of light from one waveguide branch to the other.

Similar to electro-optic switches, thermo-optics switches can be also classified into interferometric and digital devices according to their implementation principles [47]. The interferometric thermo-optic switches usually adopt interferometric structures such as directional coupler and Mach-Zehnder interferometer [48–50]. Figure 11.19 shows a 2 × 2 interferometric switch based on Mach-Zehnder interferometer. Input optical signal is split by a 3 dB coupler into two components that are coupled into two parallel waveguide arms. A heating electrode is mounted in one of the arms to introduce a phase difference between the two components by thermo-optic effect. When phase conditions are matched, either constructive or destructive interference will occur, and the optical signal will be directed to either of the outputs. As such, thermo-optic devices are based on interference of the light; precise phase control and thermal tolerance are essential in fabrication. Interferometric structure is usually applied to build small scale switch, typically 1 × 2, 1 × 4, or 2 × 2. Large-size switch matrix can be implemented by planar lightwave circuit (PLC) technology to integrate small switching units in a single chip. To date, 16 × 16 switch fabric has been built by NTT Electronics Corporation [51].

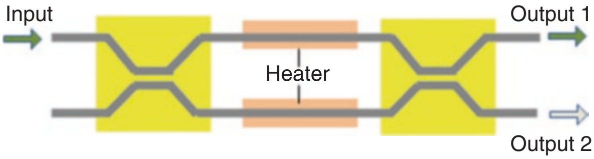


Fig. 11.19 1 × 2 interferometric TO switch based on Mach-Zehnder interferometer

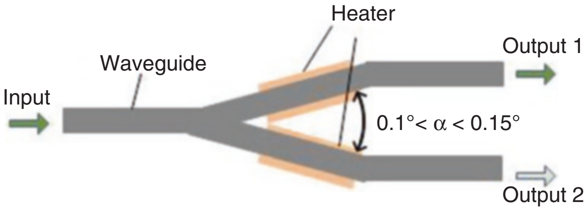


Fig. 11.20 1 × 2 digital TO switch

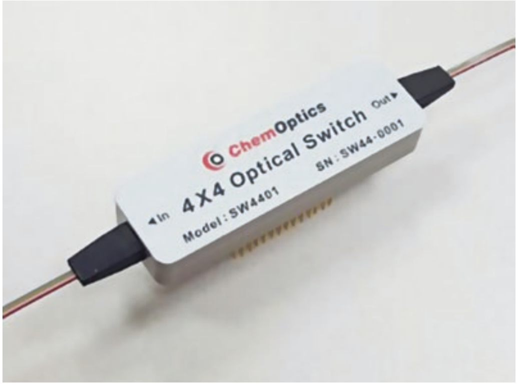


Fig. 11.21 4 × 4 thermo-optic switch by ChemOptics

Unlike interferometric thermo-optic switch, digital thermo-optic switches use mode coupling in waveguides. An example of 1 × 2 Y-branch switch is shown in Fig. 11.20, where the structure has a very small angle α (usually 0.1° < α < 0.15°) between the two waveguide branches. Both branches are equipped with heaters to change their refractive index. When one waveguide branch is heated, its effective refractive index becomes lower than that of the other waveguide branch that is unheated. This will result in coupling of fundamental mode from the branch of lower index to the one of higher index [52]. Although digital thermo-optic switches require more thermal power than interferometric switching to achieve switching function, they can maintain switching status within a larger temperature window without precise thermal power control, and they are wavelength and polarization independent. Figure 11.21 shows a 4 × 4 thermo-optic switch by ChemOptics [53]. The switch can achieve around 5 ms response time and 40 dB crosstalk.

Thermo-optic switches have good stability and robustness as they have no moving parts and their electrical driving circuitry for heating the waveguides is simple. They are suitable for small size switching and routing requirements such as ROADM and optical network protection. Due to high thermal power required, thermo-optic switches are usually built on materials like polymers that have high thermo-optic coefficients and low thermal conductivity.

11.8 Comparisons and Discussions

As discussed above, each switch technology has its own characteristics and application scenarios. In summary, Table 11.1 compares these switch technologies from the perspectives of performance, power efficiency, cost, and application. Specifically, MEMS- and beam-steering-based optical switches have large scalability, but relatively slow switching speed (10–20 ms) and low reliability; thus, they usually require complex driving circuitry to control their moving parts, which could increase the implementation cost. These optical switches are suitable for circuit switching of "elephant" traffic that includes long-lived and stable data flows in hybrid electronic/optical switch networks. Liquid crystal-based optical switches have high reliability, low power consumption, good wavelength selectivity and low packing cost, but medium scalability. These features make liquid crystal-based switches desirable for ROADM, WSS (for LCoS), and failure protection switching in WDM transmission networks or interconnection of inter-datacenters. Electro-optic and SOA-based optical switches are capable of fast switching speeds in nanoseconds, which make them suitable for OPS, various high-speed optical modulators as well as waveguide converters, but due to the complex manufacturing processes, they have much high implementation cost that limits their deployment in small- or medium-scale system. Compared with other types of optical switch technologies, thermo-optic-based switches require more power in operation, but they have good reliability and low

Table 11.1 Comparison of different commercial optical switch technologies

	Technologies					
Characteristics	MEMS based	Beam-Steering	Liquid crystal (LCoS)	Electro-optic	SOA based	Thermo-optic
Switching speed	10–20 ms	25 ms	100 ms	~ ns	~ ns	5–10 ms
Insertion loss	Medium	Low	High	High	Low	Low
Power efficiency	Medium	Low	High	High	Low	Low
Scalability	Large	Large	Medium	Medium	Medium	Small
	320 × 320	384 × 384	1 × 20	32 × 32	16 × 16	8 × 8
Reliability	Low	Low	High	High	Medium	High
Implementation cost	Medium	Medium	Low	High	High	Low
Application	OCS, WSS	OCS	ROADM, protection switching, and WSS	OPS, optical modulators	OPS, WSS	Protection switching

implementation cost, which are suitable for batch production. Thermo-optic-based switches are usually used for protection switching in optical transport networks.

References

1. D. J. Blumenthal, et al., “Integrated photonics for low-power packet networking,” IEEE Journal of Selected Topics in Quantum Electronics, vol. 17, no. 2, pp. 458–471, (March/April, 2011).
2. M. Glick, “Optical switching for next generation data centers,” in Proc. International Conference on Photonics in Switching 2009 (PS’09), Pisa, Italy, 2009, pp. 1–4.
3. A. Vahdat, H. Liu, X. Zhao, C. Johnson, “The emerging optical data center,” in Proc. OFC/NFOEC 2011, Los Angeles, CA, 2011, paper OTuH2.
4. A. Wonfor, H. Wang, R.V. Penty, I.H. White, Large port count high-speed optical switch fabric for use within datacenters [invited]. IEEE/OSA Journal of Optical Communication Network **3**(8), A32–A39 (Aug. 2011)
5. T.S. EI-Bawab, *Optical Switching* (Springer, New York, 2006)
6. K. W. Markus, “Commercialization of optical MEMS-volume manufacturing approaches,” the 2000 IEEE/LEOS International Conference on Optical MEMS, pp. 7–8, (2000).
7. V.A. Aksyuk, F. Pardo, D. Carr, D. Greywall, H.B. Chan, M.E. Simon, A. Gasparyan, H. Shea, V. Lifton, C. Bolle, S. Arney, R. Frahm, M. Paczkowski, M. Haueis, R. Ryf, D.T. Neilson, J. Kim, C.R. Giles, D. Bishop, Beam-steering micromirrors for large optical crossconnects. IEEE J. Lightwave Technol **21**(3), 634–642 (March 2003)
8. D.J. Bishop, C.R. Giles, G.P. Austin, The lucent lambda router: MEMS technology of the future here today. IEEE Comm. Mag. **40**(3), 75–79 (2002)
9. J. A. Walker, K. W. Goossen, S. C. Arney, “Mechanical anti-reflection switch (MARS) device for fiber-in-the-loop applications,“Advanced Applications of Lasers in Materials Processing/Broadband Optical Networks/Smart Pixels/Optical MEMs and Their Applications, IEEE/LEOS 1996 Summer Topical Meetings, pp. 59–60, (1996).
10. D.E. Sene, J. W. Grantham, V. M. Bright, J. H. Comtois, “Development and characterization of micro-mechanical gratings for optical modulation,” An Investigation of Micro Structures, Sensors, Actuators, Machines and Systems, IEEE MEMS ‘96, pp. 222–227, Feb 1996.
11. L. Fan, S. Gloeckner, P. D. Dobblelaere, S. Patra, D. Reiley, C. King, T. Yeh, J. Gritters, S. Gutierrez, Y. Loke, M. Harburn, R. Chen, E. Kruglick, M. Wu, A. Husain, “Digital MEMS switch for planar photonic crossconnects,” in OFC’02, Washington, DC, vol. 1, pp. 93–94, 2002.
12. P. De Dobbelaere, K. Falta, S. Gloeckner, S. Patra, Digital MEMS for optical switching. IEEE Commun. Mag. **40**(3), 88–95 (Mar. 2002)
13. M.C. Wu, O. Solgaard, J.E. Ford, Optical MEMS for lightwave communication. IEEE J. Lightwave Technol **24**(12), 4433–4454 (Dec. 2006)
14. W. C. Dickson, B. P. Staker, Gene Campbell, and William C. Banyai, “64 × 64 3D–MEMS switch control system with robustness to MEMS resonant frequency variation and pointing drift,” in OFC’04, Los Angeles, CA, 2004, pp. ThQ5.
15. Calient, Product user guide – “S320 Photonic Switch Getting Started Guide”
16. “Polatis technology – Directlight® Beam-Steering All-Optical Switch” [Online]. Available: http://www.polatis.com/polatis-all-optical-switch-technology-lowest-loss-highest-performance-directlight-beam-steering.asp
17. “Polatis - series 7000 384 × 384 switch - cross-connect- compact single mode all-optical low loss switch up to 384 × 384 ports.” [Online]. Available: http://www.polatis.com/series-7000-384x384-port-software-controlled-optical-circuit-switch-sdn-enabled.asp
18. R. E. Wagner, J. Cheng, “Electrically Controlled Optical switch for Multimode Fiber Applications,” Applied Optics, Vol. 19, No. 17, pp. 2921–2925, (September 1, 1980).
19. P. Yeh, C. Gu, *Optics of Liquid Crystal Displays* (Wiley, New York, 1999)

20. J-C. Chiao, K-Y. Wu, and J-Y. Liu, "Liquid-Crystal WDM Optical Signal Processors," Broad Band Communications for The Internet Era Symposium Digest, 2001 IEEE Emerging Technologies Symposium, pp. 53–57, (2001).
21. K. Noguchi, Optical free-space multichannel switches composed of liquid-crystal light-modulator arrays and Birefringent crystals. IEEE J. Lightwave Technol. **16**(8), 1473–1481 (August, 1998)
22. G. Baxter, S. Frisken, D. Abakoumov, H. Zhou, I. Clarke, A. Bartos, S. Poole, "Highly Programmable Wavelength Selective Switch Based on Liquid Crystal on Silicon Switching Elements," in OFC'06, Anaheim, CA, 2006, pp. OTuF2.
23. JDSU, White Paper "A performance comparison of WSS switch engine technologies".
24. S. Frisken, "Advances in liquid crystal on silicon wavelength selective switching," in OFC'07, Anaheim, CA, 2007, pp. OWV4.
25. P. Wall, P. Colbourne, C. Reimer, S. McLaughlin, "WSS switching engine technologies," in OFC'08, San Diego, CA, 2008, pp. OWC1.
26. P. Colbourne, B. Collings, "ROADM switching technologies," in OFC'11, Los Angeles, CA, 2011, pp. OTuD1.
27. Steve Frisken, Glenn Baxter, Dmitri Abakoumov, Hao Zhou, Ian Clarke, Simon Poole, "Flexible and Grid-less Wavelength Selective Switch using LCOS Technology," in OFC'11, Los Angeles, CA, 2011, pp. OTuM3.
28. Finisar, Product data sheet "1 × 9/1 × 20 Flexgrid Wavelength Selective Switch (WSS)".
29. E.L. Wooten, K.M. Kissa, A. Yi-Yan, et al., A review of lithium niobate modulators for fiber-optic communications systems. IEEE Journal of Selected Topics in Quantum Electronics **6**(1), 69–82 (2000)
30. T. Volk, M. Wöhlecke, *Lithium Niobate: Defects, Photorefraction and Ferroelectric Switching* (Springer Science & Business Media, Berlin, 2008)
31. Y. Silberberg, P. Perlmutter, J.E. Baran, Digital optical switch. Appl. Phys. Lett. **51**(16), 1230–1232 (1987)
32. K. Suzuki, T. Yamada, M. Ishii, et al., High-speed optical 1× 4 switch based on generalized Mach–Zehnder interferometer with hybrid configuration of silica-based PLC and lithium Niobate phase-shifter Array. IEEE Photon. Technol. Lett. **19**(9), 674–676 (2007)
33. G.L. Li, P.K.L. Yu, I. Tutorial, Optical intensity modulators for digital and analog applications. IEEE J. Lightwave Technol **21**(9), 2010–2030 (2003)
34. M.J. Connelly, *Semiconductor Optical Amplifiers* (Springer Science & Business Media, New Year, 2007)
35. K.E. Stubkjaer, Semiconductor optical amplifier-based all-optical gates for high-speed optical processing. IEEE Journal of Selected Topics in Quantum Electronics **6**(6), 1428–1435 (2000)
36. J.H. Kim, Y.M. Jhon, Y.T. Byun, et al., All-optical XOR gate using semiconductor optical amplifiers without additional input beam. IEEE Photon. Technol. Lett. **14**(10), 1436–1438 (2002)
37. Z. Li, G. Li, Ultrahigh-speed reconfigurable logic gates based on four-wave mixing in a semiconductor optical amplifier. IEEE Photon. Technol. Lett. **18**(12), 1341–1343 (June 2006)
38. R. Hemenway, R. Grzybowski, C. Minkenberg, R. Luijten, Optical-packet-switched interconnect for supercomputer applications. J. Opt. Netw. **3**(12), 900–913 (Dec. 2004)
39. O. Liboiron-Ladouceur et al., The data vortex optical packet switched interconnection network. J. Lightwave Technol. **26**(13), 1777–1789 (July 2008)
40. Y.K. Yeo, Q. Huang, L. Zhou, Large port-count optical crossconnects for data centers, in *Proc. International Conference on Photonics in Switching (PS)*, (2012)
41. P. Gambini, M. Renaud, C. Guillemot, et al., Transparent optical packet switching: Network architecture and demonstrators in the KEOPS project. IEEE Journal on Selected Areas in Communications **16**(7), 1245–1259 (1998)
42. R. Luijten, W. E. Denzel, R. R. Grzybowski, et al, "Optical interconnection networks: The OSMOSIS project," in Proc. The 17th Annual Meeting of the IEEE Lasers and Electro-Optics Society. 2004.

43. G. Nakagawa, Y. Kai, S. Yoshida, et al, “High-speed and high-reliability optical selector for 256 × 256 large-scale, nanosecond-order optical switching,” in OFC’08, San Diego, CA, 2008, pp. OWI5.
44. C. Develder, J. Cheyns, M. Pickavet, et al, “Multistage architectures for optical packet switching using SOA-based broadcast-and-select switches”, in OFC’03, Atlanta, GA, 2003, pp. FS3.
45. R. Stabile, A. Albores-Mejia, K. A. Williams, “Monolithic active-passive 16× 16 optoelectronic switch,” Optics letters, vol. **37**, no. 22, 2012, pp. 4666–4668, 2012.
46. M.B.J. Diemeer, J.J. Brons, E.S. Trommel, Polymeric optical waveguide switch using the thermooptic effect. J. Lightwave Technol. **7**(3), 449–453 (March 1989)
47. G. Coppola, L. Sirleto, I. Rendina, et al, “Advance in thermo-optical switches: principles, materials, design, and device structure,” Optical Engineering, vol. 50, no. 7, pp. 071112–071112-14, 2011.
48. R.C. Alferness, Guided-wave devices for optical communication. IEEE J. Quantum Electron. **17**(6), 946–959 (June 1981)
49. T. Goh, M. Yasu, K. Hattori, A. Himeno, M. Okuno, Y. Ohmori, “Low loss and high extinction ratio strictly non blocking 16 × 16 thermo-optic matrix switch on 6-in wafer using silica-based planar lightwave circuit technology,” IEEE J. Lightwave Technol., Vol. 19, No. 3, pp. 371–379, March 2001.
50. T. Chu, H. Yamada, S. Ishida, et al., Compact 1× N thermo-optic switches based on silicon photonic wire waveguides. Opt. Express **13**(25), 10109–10114 (2005)
51. T. Goh, “Recent advances in large-scale silica-based thermo-optic switches,” in Proc. Asia-Pacific Optical and Wireless Communications Conference and Exhibit, pp. 49–56, 2001
52. W.K. Burns, A.F. Milton, Mode conversion in planar-dielectric separating waveguides. IEEE J. Quantum Electron. **11**(1), 32–39 (January 1975)
53. “Chem Optics Product Digest” [Online]. Available: http://www.chemoptics.co.kr/main/chemoptics_2015brochure.pdf.

Chapter 12
Silicon Photonics Switch Matrices: Technologies and Architectures

Francesco Testa, Alberto Bianchi, and Marco Romagnoli

12.1 Introduction

Optical switching in *intra-data center networking* is gaining momentum, thanks to the unique advantages that it offers in terms of energy efficiency, low latency, cost reduction, wide bandwidth and transparency to bit rate and protocol with respect to electronic switching [1–3].

Key parameters for the introduction of optical switching in *data centers* are cost, integration level, port count, and footprint of the optical switching matrices. Recent achievements in silicon *photonics integrated technologies* for high-speed optical interconnects boost the advancement in the realization of new highly *integrated optical switching devices*.

The *integrated optical switch matrices* considered here are monolithically integrated photonic chips in which the signals to be processed propagate through optical waveguides.

Silicon integration will allow to exploit the highly developed silicon manufacturing infrastructure and the materials used for electronic integrated circuits. This ensures low cost and mass manufacturability, while the high refractive index contrast of silicon photonics allows to implement highly miniaturized optical circuits. Low power consumption is ensured by the energy-efficient physical mechanisms

The original version of this chapter was revised. The correction to this chapter can be found at https://doi.org/10.1007/978-3-319-61052-8_16

F. Testa (✉) • A. Bianchi
Ericsson Research, Via G. Moruzzi, 1, 56124 Pisa, Italy
e-mail: francesco.testa@ericsson.com; alberto.bianchi@ericsson.com

M. Romagnoli
Consorzio Nazionale Inter-Universitario per le Telecomunicazioni (CNIT),
Via G. Moruzzi, 1, 56124 Pisa, Italy
e-mail: mromagnoli@cnit.it

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
DOI 10.1007/978-3-319-61052-8_12

used to switch the light, by the small dimensions of the photonic integrated circuits, and by the tight integration of photonic and control electronic circuits made possible by the use of the same production technologies. There are also some limitations to cope with using silicon photonics such as the need to generate light by heterogeneously integrating blocks taken from InP-based technology, a more complex handling of signal polarization and coupling with single-mode fibers with respect to conventional optical technologies.

Optical switch matrices for data centers should allow to connect any idle input port to an arbitrary idle output port (non-blocking characteristic), in order to fully utilize the switch internal bandwidth. Moreover, they should have port count as large as possible (larger than 64 × 64) to allow building up of a larger multistage optical switching fabric ensuring networking of thousands of servers. They should have small size to be included in a module of few squared centimeters in order to utilize effectively the data center footprint and low loss, low differential loss among the different paths, and low crosstalk in order to realize scalable matrices. The characteristics of low polarization sensitivity are also crucial due the use of a single-mode fiber as interconnect infrastructure.

However, integrated optical switches, different from integrated electrical switch counterparts, are purely analog devices. While in digital electrical switches signal regeneration occurs in the internal CMOS gates, inside the optical switches, loss and crosstalk are experienced during processing through the many traversed switching cells. Loss and crosstalk increase with the switch matrix port count and complexity and eventually limit the scalability. Different prototypes of silicon-integrated optical switching matrices are presented, explained, and discussed in this chapter, with non-blocking characteristics, having potential high port count, low loss, and low crosstalk.

The chapter is organized as follows: the physical effects used to switch optical signals in silicon-integrated matrices are presented in Sect. 12.2, and the relevant switching cell types are discussed in Sect. 12.3. In Sect. 12.4, the optical switch matrix architectures for optical circuit switching are explained, and the relative integrated devices proposed and demonstrated so far are reported. Section 12.5 is dedicated to the presentation of optical switching matrices used in optical packet switching. Section 12.6 reports on a silicon photonics wavelength selective switch implementation, and in Sect. 12.7, a comparison table among the switching fabrics is presented. Finally, in Sect. 12.8, the perspectives and future research directions are briefly discussed.

12.2 Physical Effects and Mechanisms for Optical Switching in Silicon

In a first category of photonics integrated switch matrices, the mechanisms used for optical switching are associated to a change in the index of refraction of optical waveguides in interferometric- or resonant circuit-based switching cells. In such integrated switching cells, this index variation causes a change of the signal path from one input port to another output port.

In such silicon photonics switches, two physical effects are used: the plasma dispersion effect and the thermo-optic effect with quite different response times. With the plasma dispersion effect, the switch can be reconfigured in a nanosecond scale, while with the thermo-optic effect, the reconfiguration time is in the order of microseconds.

A completely different mechanism used for switching in silicon monolithic integrated optical devices is the physical movement of optical waveguides driven by electrostatic actuators. They have become attractive due to the widely developed micro-electromechanical system (MEMS) process technology in which micrometric-scale switching cells are created combining mechanical and electrical components. This technology leverages on silicon wafer-scale fabrication processing to achieve low cost and high repeatability, and it is already widely adopted in consumer electronics accelerometers and gyroscopes for mobile phones. These types of optical switches have reconfiguration times very much dependent on the mechanical design of the MEMS switching cell; typically, they are spanning from 1 to 100 μs.

12.2.1 Plasma Dispersion Effect

Plasma dispersion effect, also named free carrier dispersion effect, has been investigated and numerically analyzed for the first time in 30 years [4]. From that time, many progresses have been made, and today optical modulator in silicon photonics using this physical effect has become commercially available in optical interconnects [5]. The optical switches utilize the same effect of modulators consisting of the variation of the refractive index in an optical waveguide induced by the variation in the concentration of free carriers.

Unfortunately, it is not possible to change the waveguide refractive index without affecting the optical loss generated by the free carrier absorption since the two effects are coupled by the Kramers-Kronig dispersion equation:

$$\Delta n(\omega) = \left(\frac{c}{\pi}\right) P \int_{\infty}^{0} \frac{\Delta\alpha(\omega')\, d\omega'}{\omega'^2 - \omega^2} \tag{12.1}$$

in which P indicates the principal part of the integral due to the singularity at $\omega = \omega'$, Δn is the refractive index change, Δα is the absorption coefficient change, c is the light velocity, and ω is the angular frequency of the light.

The refractive index and absorption coefficient change with respect to the change in the carrier density have been quantified in [6]:

$$-\Delta n(\lambda) = p(\lambda)\Delta N_e^{q(\lambda)} + r(\lambda)\Delta N_h^{s(\lambda)} \tag{12.2}$$

$$\Delta\alpha(\lambda) = a(\lambda)\Delta N_e^{b(\lambda)} + c(\lambda)\Delta N_h^{d(\lambda)} \tag{12.3}$$

in which ΔN_e and ΔN_h are relevant with the concentration of electrons and holes, respectively.

The coefficients have been empirically found in [6] for crystalline silicon for wavelength from 1 up to 14 μm.

At the wavelengths of interest in optical fiber communication systems at 1.3 μm and 1.5 μm, the coefficients are:

for $\lambda = 1.3$ μm

$$\begin{aligned} &a\left(\mathrm{cm}^2\right) = 3.48\cdot 10^{-22} \\ &b = 1.229 \\ &c\left(\mathrm{cm}^2\right) = 1.02\cdot 10^{-19} \\ &d = 1.089 \\ &p\left(\mathrm{cm}^3\right) = 2.98\cdot 10^{-22} \\ &q = 1.016 \\ &r\left(\mathrm{cm}^3\right) = 1.25\cdot 10^{-18} \\ &s = 0.835 \end{aligned} \tag{12.4}$$

and for $\lambda = 1.55$ μm

$$\begin{aligned} &a\left(\mathrm{cm}^2\right) = 8.88\cdot 10^{-21} \\ &b = 1.167 \\ &c\left(\mathrm{cm}^2\right) = 5.84\cdot 10^{-20} \\ &d = 1.109 \\ &p\left(\mathrm{cm}^3\right) = 5.40\cdot 10^{-22} \\ &q = 1.011 \\ &r\left(\mathrm{cm}^3\right) = 1.53\cdot 10^{-18} \\ &s = 0.838 \end{aligned} \tag{12.5}$$

From the above values, it results that at 1.3 and 1.55 μm, it is possible to find a region of the carrier density around 10^{18} cm^{-3} in which the optical phase shift component in the waveguide is dominant with respect to the optical absorption component. It is therefore possible to obtain an optical phase shift of about π radians, as needed for optical switching in a Mach-Zehnder interferometer-based switching cell (see Sect. 12.3.1), at low loss and with a reasonably short length.

As an example, the characteristics of a phase shifter working at 1.3 μm wavelength with a carrier concentration (both holes and electrons) of 10^{18} cm^{-3} are calculated.

From the coefficients reported above, it is obtained for the change of the refractive index:

$$-\Delta n\left(1.3\ \mu m\right) = 2.98\ 10^{-22}\left(10^{18}\right)^{1.016} + 1.25\ 10^{-18}\left(10^{18}\right)^{0.835} = 1.9\ 10^{-3} \tag{12.6}$$

The waveguide length for a π radian phase shift, L_π, can be obtained by the following formula:

$$L_\pi = \frac{\Delta\theta\lambda}{2\pi\Delta n} \tag{12.7}$$

and by imposing $\Delta\theta = -\pi$, it is found: $L_\pi = 340$ μm.

The absorption coefficient variation is given by:

$$\Delta\alpha\left(1.3\ \mu m\right) = 3.48\ 10^{-22}\left(10^{18}\right)^{1.229} + 1.02\ 10^{-19}\left(10^{18}\right)^{1.089} = 8.65\ \text{cm}^{-1} \tag{12.8}$$

Optical loss for a phase shifter with 340 μm length is given by:

$$\Delta\alpha L_\pi = 0.29 \tag{12.9}$$

and it corresponds to an optical loss <1.5 dB.

It has to be noticed that holes have a much stronger effect (about 10×) than electrons on the change of refractive index.

In some optical switching cells based on Mach-Zehnder interferometers (see Sect. 12.3.1) electrically driven in push-pull configuration, the phase shift required on both arms is $\pi/2$, and the phase shifter length can be halved [7].

Plasma dispersion-based switches are implemented as p-i-n junctions operating in forward mode. In such structures, higly doped p- and n-regions are separated by an intrinsic region where the waveguide is formed [8]. By forward-biasing the p-i-n junction, free carriers are injected in the intrinsic waveguide region, as depicted in Fig. 12.1, changing its refractive index. The doped region of the silicon waveguide constitutes therefore a phase shifter, and the phase change depends on both the injection current and the length of the phase shifter. This phase shifter can be inserted in a waveguide-based switching cell, for instance, Mach-Zehnder interferometers or micro-ring resonators, for tuning or switching. The time response of the plasma dispersion effect in a forward biased p-i-n junction is limited by the carrier diffusion time in the injection region and is in the order of nanoseconds: this value complies with most optical packet switching requirements [9, 10]. Alternatively, the forward bias can be substituted by a carrier accumulation mechanism. Carrier accumulation occurs around a thin oxide region in the p-oxide-n structure. The thickness of the insulator region is a trade-off between bandwidth and driving voltage. The thinner the oxide region, the higher is the value of the capacitance and consequently the lower is the driving voltage but also the bandwidth would be limited. This solution is quite interesting because based on charge and discharge of the capacitor formed by the insulator and its interfaces. The index change depends therefore on the charge accumulation only, i.e., the charge of the capacitor, and may be of the

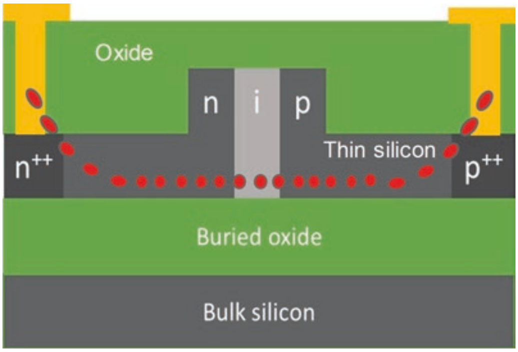


Fig. 12.1 Plasma dispersion effect in a silicon-on-insulator (SOI) rib waveguide

order of 10^{-3}. In this condition, there is no current flow, and the power consumption can be extremely reduced. In the forward biased p-i-n junction instead, the current flow is continuous. Presently this solution has been adopted for modulators only [11], and it is affected by a nonnegligible loss that in optical switching matrices eventually limits the scalability.

12.2.2 Thermo-optic Effect

A second physical effect used in silicon photonics optical switching devices is the thermo-optic effect.

Bulk silicon has a quite large thermo-optic coefficient defined as the change in refractive index *n* with temperature *T*:

$$\frac{\partial n}{\partial T} = 1.86 \times 10^{-4} / K \text{ at } 300 \text{ K}, \tag{12.10}$$

and thermal conductivity:

$$k_{\mathrm{Si}} = 148 \text{ W}/(\text{mK}) \tag{12.11}$$

These parameters, together with the high index contrast (between waveguide core and cladding) of silicon photonics, allow the realization of highly miniaturized optical waveguide phase shifters with micrometric size and low switching power (in the mW range) made by placing micro-heaters in the vicinity of the optical waveguide in order to change its local temperature and consequently the refractive index. These thermally actuated phase shifters, placed in one or both arms of a Mach-Zehnder interferometer (MZI) or along a micro-ring resonator, could act as active elements in a switching cell of a big integrated optical switch matrix.

Two types of micro-heaters have been developed so far: silicon doped [12] and metallic [13].

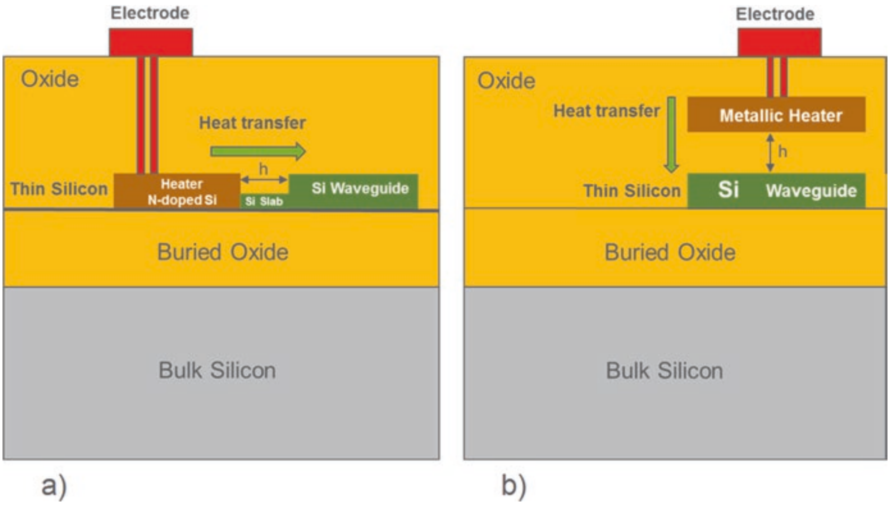


Fig. 12.2 Thermo-optic effect in a silicon-on-insulator (SOI) optical waveguide. (**a**) Silicon doped micro-heaters, (**b**) metallic micro-heater

A phase shifter with silicon-doped micro-heater is depicted in Fig. 12.2a. At one side of the silicon optical waveguide, surrounded by the buried oxide (BOX) layer and cladding material, a micro-heater made by N-doped silicon is implemented. A silicon slab of width h could be inserted, in some designs, between the waveguide and the doped heater to increase the thermal conductivity, the slab being low enough to avoid optical leakage through it. To activate the phase shifter, a current is injected into the heater that works as a resistor, increasing the temperature of the adjacent optical waveguide and changing its refractive index.

In Fig. 12.2b, a phase shifter is made of a metal line with a suitable width and thickness to ensure the expected value of resistivity, and it is placed at a certain distance h (less than 1 μm) above the silicon optical waveguide. The metal line is, for instance, made by Ti/TiN, and the current flows through the metal resistor causing the increase of temperature and the generation of the heat necessary to change the refractive index of the silicon waveguide core. In both heater configurations, the heat transport by conduction is defined by the following equation [13]:

$$\nabla\left(-k\nabla T\right)+\rho c\frac{\partial T}{\partial t}=q_s \tag{12.12}$$

where k is the thermal conductivity of the used materials, c is the heat capacity, ρ is the material density, q_s is the heat flux density, and T is the temperature. Since ρ is fixed and it depends on the materials, the design parameters that have to be optimized to achieve a fast time response, a high power efficiency, and a low optical loss of a waveguide phase shifter are mainly k and c. They depend mostly on the thickness of the BOX and the cladding layers. Thinner BOX and cladding layers give a faster response time and a better power efficiency, but a minimum thickness of about 1 μm has to be set for both BOX layer and cladding to avoid loss increase due to leakages of the optical field into the underneath silicon substrate and optical absorption by the metal line of the heaters.

The thermo-optic effect, differently from the plasma dispersion effect, is inherently low loss and can be utilized in optical waveguide-based switching cells for tuning and switching photonic devices with negligible insertion loss. However, the response time of such thermally actuated switches is slower than the plasma dispersion-based counterparts, and it is usually of few microseconds limited by the heat diffusion process.

Power consumption of thermal heaters is a function of the current needed to reach the wanted temperature, and it depends on the thermal impedance of the insulator between the heater and the waveguide and on the distance between the silicon core and the heat sink that usually is the silicon substrate. The insulator thickness on top of the silicon core is designed to avoid loss due to optical mode overlap with the heater. Typically, metal heaters are placed at 0.7–1 μm above the silicon core. When the time response of the heater is not critical, an underetch of the buried oxide below the silicon core can help in reducing power consumption, i.e., the underetch serves to thermally insulate the waveguide, and the reduced heat exchange serves to maintain the steady-state temperature of the waveguide core. A comparison between thermo-optic- and p-i-n junction-based phase shifters used for a $N = 64$ port switch is reported in [14]. The time response is in favor of p-i-n junctions, but the insertion loss is much lower in thermo-optic-based switches. For $N = 64$, 8.5 dB of insertion loss for the current-driven thermo-optic switch and 17 dB [15] or 29 dB [16] for the carrier-injection switches have been reported.

12.3 Integrated Switching Cell Technologies

The key building block in an integrated optical switching matrix is the switching cell, that is, the single *switching element* (normally a 2 × 2 switch or in some cases a 1 × 2 switch) interconnected to many other switching elements of the same type in a two-dimensional matrix. The switching cell characteristics of bandwidth, loss, and extinction ratio are crucial to realize a scalable, high-performance, integrated switching matrix. The switching cell states can be identified either with *bar* and *cross* (if related to the optical path through the switching cell) or alternatively with *on* and *off* (if related to the state of the electrical control signal), as depicted in Fig. 12.3. The cell characteristics in the two alternative states are generally different.

Loss of a switching cell is defined as the reduction in the signal power from input (P_i) to output (P_o) port of the wanted channel (see Fig. 12.3).

Extinction ratio of a switching cell is the ratio of the signal power level at the output port of interest (P_o) to the signal power level at the unwanted port (P_x) (see Fig. 12.4).

Crosstalk is the ratio of total power in the disturbing signals to that in the wanted signal.

While the extinction ratio is a property of the switching cells, “crosstalk” is reserved for the description of system effects of a complete switching matrix and may comprise many unwanted signals interfering with the wanted signal.

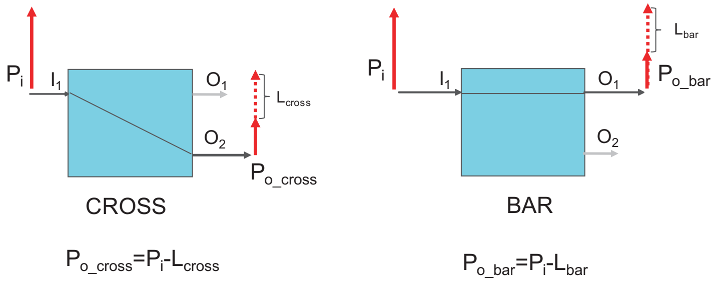


Fig. 12.3 Loss of a switching cell in *cross* and *bar* state

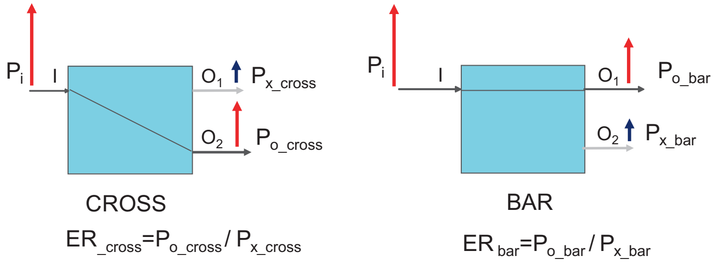


Fig. 12.4 Extinction ratio of a switching cell in *cross* and *bar* state

Crosstalk is the ratio of total power in the disturbing signals to that in the wanted signal.

In the following, the most relevant switching cells are presented and discussed.

12.3.1 Mach-Zehnder Interferometer (MZI) Switching Cell

The switching cell based on *Mach-Zehnder interferometers* (MZI) is shown in Fig. 12.5. It consists of an optical interferometer with an input 3 dB coupler, two arms equipped with phase shifter sections, and an output 3 dB coupler. The cell works with the following principle: the optical power of the input signal is split by the first coupler into two equal portions that interfere at the output coupler depending on the relative phase shift. If the phase difference among the two arms is 0 radians, as depicted in Fig. 12.6, a constructive interference occurs at the crossover port, and the signal entering at I_1 goes out at O_2, and similarly the signal at I_2 port goes out at O_1 port: in this condition, the switch is defined to be set on *cross* state. When the phase difference among the arms is changed to π radians, activated by an electrical control signal, the constructive interference occurs at the through port, and the signal entering at I_1 goes out at O_1 (see Fig. 12.6), while the signal entering at port I_2 goes out at O_2: in this condition, the switch is defined to be in *bar* state. This switch cell can be activated by using either the thermo-optic effect or the

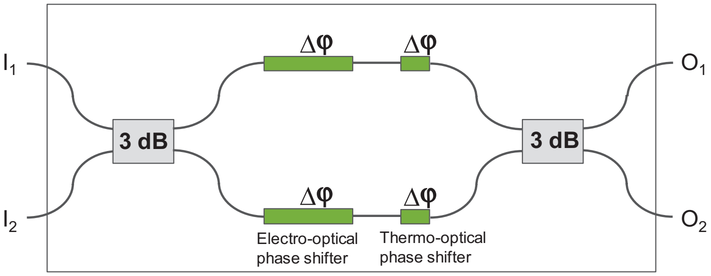


Fig. 12.5 Mach-Zehnder interferometer switching cell

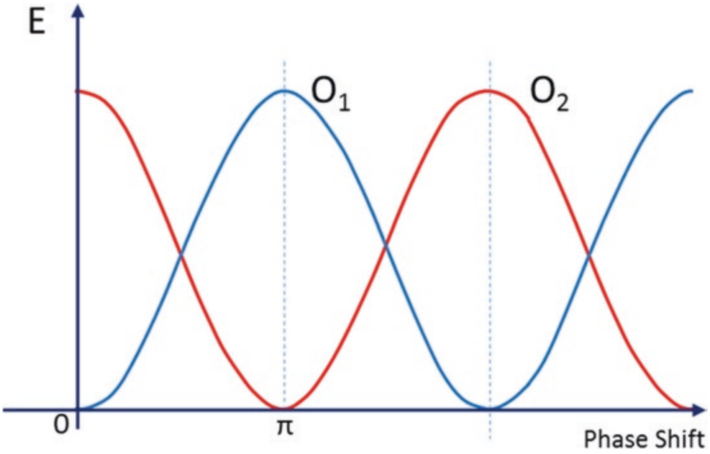


Fig. 12.6 Electrical field amplitude vs phase shift between the two arms at output ports O_1 and O_2 of the MZI when the input signal is injected at port I_1

plasma dispersion effect (see Sect. 12.2) or both to change the refractive index in the waveguides of the phase shifter sections: the selection of the switching mechanism depends on the switching speed requirements that should be in the microsecond range for the thermo-optic effect and nanosecond range for the plasma dispersion effect. In some high-speed switch designs, both types of phase shifter can be used to improve the extinction ratio performances with the plasma dispersion-based phase shifters activated for fast switching and the thermo-optic phase shifters tuned to compensate interferometer unbalances due to production imperfections without introducing significant excess losses.

The switching cells working with plasma dispersion effect are normally designed to be in *cross* state with no activation signal (*off* state), and for that reason, they have, in general, better loss and extinction ratio performance in *cross* than in *bar* state that needs current injection with the relative increase of loss and decrease of extinction ratio.

In most of the integrated switch design, the π relative phase shift between the two arms of the MZI, needed to set the switch in *bar* state, is actuated by two phase shifters, one per each arm driven in push-pull manner. In this condition, only $\pi/2$ phase shift is needed per phase shifter, and this results in a lower loss with respect to the MZI configuration with only one shifter in one arm. In this configuration, the carrier density can be halved with consequent improvement of loss and extinction ratio due to a better power balance between the two interfering signals at the MZI output.

12.3.2 Resonant Switching Cell

A second type of monolithically integrated silicon photonics switching cell is the one based on waveguide-coupled *micro-ring resonators* (MRR) [17]. This is a resonant structure consisting of two bus waveguides (W_1 and W_2) and one or more equal MRRs coupled to each other (see Fig. 12.7). The MRR has periodic frequency response repeating itself each free spectral range (FSR). If the wavelenght of the signal going into port I_1 is aligned with the resonance of the MRR (*on-resonance*) (see Fig. 12.7a), then that signal from W_1 couples into the first ring, couples subsequently to the other intermediate rings and finally couples out from the last ring into W_2 and exits the device at the drop port O_1. If the input signal instead is *off-resonance* with the rings (see Fig. 12.7b), it remains in W_1 and exits at port O_2 (through port) nearly unperturbed. At the same time, in *off-resonance* state, a signal going into port I_2 remains in W_2 and exits port O_1 unperturbed.

The resonance condition is according with the following equation:

$$2\pi R n_{\text{eff}} = m\lambda \tag{12.13}$$

in which R is the ring radius, n_{eff} is the effective refractive index of the waveguide, λ is the signal wavelength, and m is an integer number.

The switching principle of the MRR is shown in Fig. 12.8a.

Micro-ring resonator switching cell is narrowband, and if a single micro-ring is used, it has Lorentzian shape with slow roll-off of the resonance passband resulting in a relatively low switch extinction ratio.

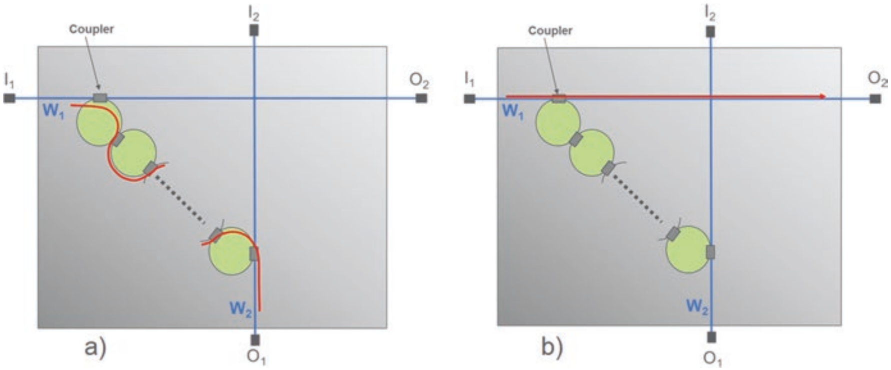


Fig. 12.7 Resonant switching cell states: cell set (**a**) *on-resonance* with the signal, (**b**) *off-resonance* with the signal

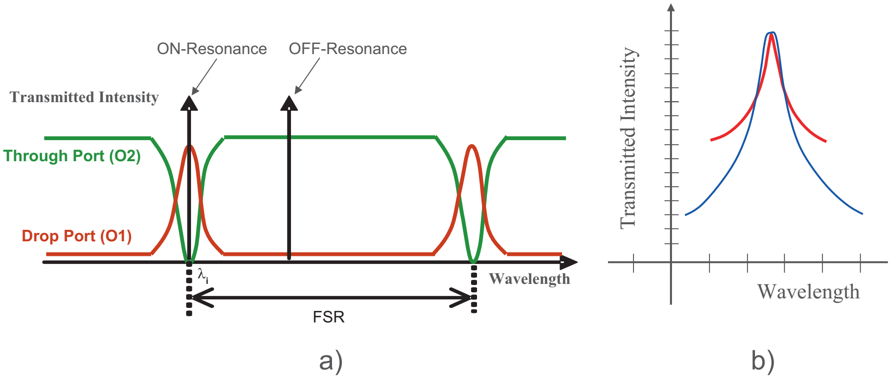


Fig. 12.8 (**a**) Transmitted intensity at ports O1 and O2 with the MRR *on-resonance* and *off-resonance*; (**b**) MRR frequency response of a single micro-ring (*red curve*) and two-coupled micro-rings (*blue curve*)

Higher-order MRR-based filters are required to improve both the passband width and the extinction ratio. In this case, a filter with high-order-coupled micro-rings can be suitably synthesized [18]. The frequency response of a two-coupled micro-ring switching cell is shown in Fig. 12.8b, and it is compared with that of a single micro-ring.

This switching cell can be activated by an electrical control signal to change the effective refractive index n_{eff} of the ring waveguides in order to match the resonance condition (see formula 12.13). This is achievable using either the plasma dispersion effect or the thermal effect, as discussed in Sect. 12.2. With the former, electrical carriers are injected into the ring waveguides, and the switch time response is very fast in the nanosecond range, while with the latter, the local temperature of the ring waveguide is changed using a micro-heater, and in this case, the response time is in the microsecond range. Generally, this type of switching cells is set as *on-resonance* when activated (*on-state*) and *off-resonance* when deactivated (*off-state*), and they have better loss and extinction ratio in this switch's last state since the optical path is shorter.

12.3.3 Micro-electromechanical System (MEMS) Switching Cell

The key building block in silicon *MEMS*-based optical switching cells is the comb electrostatic actuator. It has been widely used in MEMS devices because of easy fabrication and mass manufacturability, and it has the function of translating movable micrometric parts with high precision level [19].

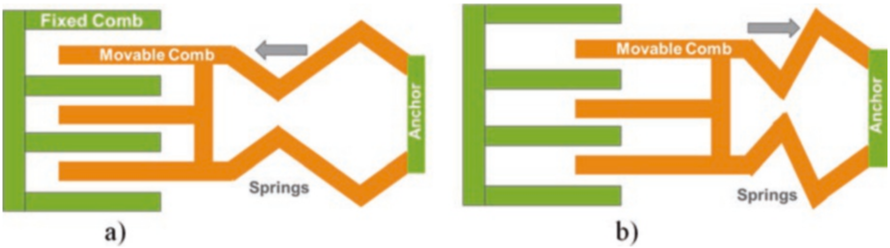


Fig. 12.9 Comb electrostatic MEMS actuator: (**a**) actuator in *on*, (**b**) actuator in *off*

It consists of two plates each with a number of fine interdigitated fingers, the fix comb and the movable comb (see Fig. 12.9)—by applying a voltage between them, the fixed comb attracts the movable comb with a force given by:

$$F = \frac{1}{2}\frac{\partial C}{\partial x}V^2 \tag{12.14}$$

where C is the capacitance between the plates and V is the applied voltage and x is the axis of translation.

Since

$$C = \epsilon_0 \epsilon_r \frac{A}{d} \tag{12.15}$$

where ϵ_0 is the permittivity of the free space, ϵ_r is the relative permittivity of the material between plates, A is the overlapping plates area, and d is the plate separation, the more numerous are the fingers, the larger is the capacitance surface and hence the higher is the force. An interesting advantage of this type of actuator is the fact that the electric circuit is capacitive, and hence the driving current needed to load the capacity is very low even if the voltage is high.

In the electrostatic actuator shown in Fig. 12.9, the fixed comb that is totally anchored to the substrate is indicated in green, while the movable comb, supported by a spring, is indicated in orange, and it translates to left when a driving voltage is applied between the conductive plates. The movable comb is free to move except at the right end where it is also anchored to the substrate, so that the actuator can come back to its initial position with no applied voltage, thanks to the spring force.

Optical switching cells based on silicon MEMS consist of a fixed waveguide section and a movable waveguide section, and the light coming from the fixed waveguide is coupled to two alternative movable optical waveguides under the comb actuator control. This type of switching cell based on moving waveguides can easily achieve high extintion ratio by providing enough separation between the alternative waveguides to which the output signals are to be coupled.

12.4 Integrated Matrices with μs Response Time for Optical Circuit Switching

Optical circuit switching has been researched as an attractive technology to increase the bandwidth and reduce latency and power consumption in intra-data center networking. Optical switching matrices have unsurpassed capability in routing high-capacity dataflows of 100 Gbps, 400 Gbps, or more with high energy efficiency and bit rate and protocol transparency, but, differently from electrical switch matrix counterparts, they cannot process a data stream bit by bit in the optical domain. Hence, hybrid switching architectures have been proposed in [1, 20] in which bursty and latency-sensitive traffic (mice flows) is switched by electrical packet switches, while long-lasting dataflows (elephant flows) are routed by optical circuit switches (OCS). It has been found that the addition of an optical switching layer working in cooperation with the electrical packet switching (EPS) layer improves the data center networking performances. However, in such hybrid networks, the data throughput is affected by the optical switch response time, and optical switches with speed in the order of microseconds are needed. If the switch response time is fast enough, they could also be used in optical burst switching (OBS) [21]: according to this technique, optical packets with the same destination, before switching, are wrapped into longer burst frame lasting for many microseconds to reduce the impact of the switch reconfiguration time on the bandwidth utilization efficiency and latency. In the following, the architectures and technologies for integrated circuit switches are presented and discussed.

12.4.1 Crossbar Switch Architecture

An N × M integrated optical *crossbar switch matrix* is a photonic network with N horizontal optical waveguides and M vertical optical waveguides containing N × M optical switching cells (crosspoint switch elements), each one placed at the crossing point between a vertical waveguide and a horizontal waveguide, as depicted in Fig. 12.10. The input signals traverse a number of cascaded switching cells until they arrive at the output ports. Along this path, only one crosspoint cell is activated, and it is the one deviating the signals from the horizontal waveguide to the vertical waveguide. In Fig. 12.10, the internal switch connection from I_2 to O_3 is indicated by the orange line, and it is obtained by activating the crosspoint cell S_{23}.

The schematic block diagram of the 2 × 2 crosspoint switching cell is shown in Fig. 12.11. It has two input ports I_1 and I_2 and two output ports O_1 and O_2, and if the switch cell state is in *off* (switching cell deactivated), the input signals entering from I_1 and/or I_2 proceed straight on its direction toward O_1 and O_2, respectively. When the switch cell is set *on*, the signal from I_1 is deviated to the output port O_2, while the ports I_2 and O_1 remain unconnected.

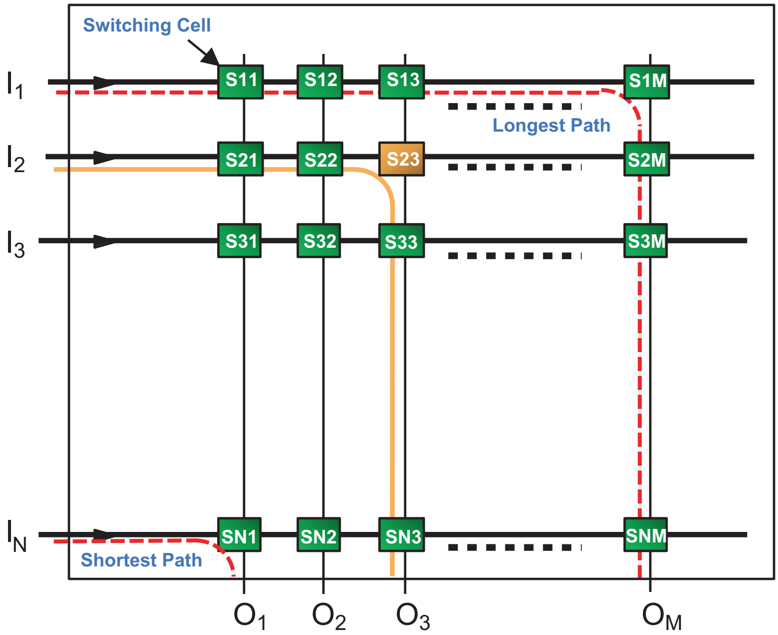


Fig. 12.10 N × M optical crossbar switch matrix

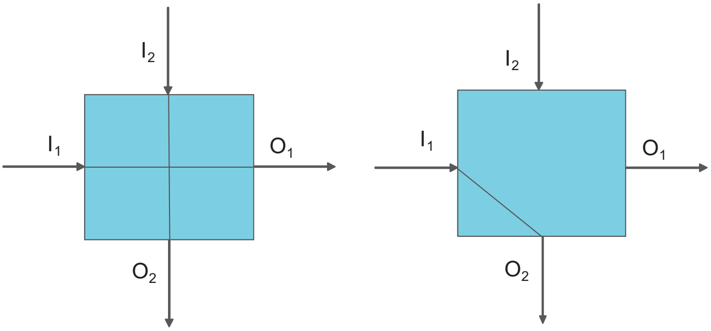


Fig. 12.11 2 × 2 crosspoint switching cell

Optical crossbar switches have been realized with the same architecture of electrical CMOS counterparts that are commercially available with 160 × 160 ports at 12 Gbps gigabit rates [22] and 16 × 16 ports at 28 Gbps rate [23]. In these electrical crossbar matrices, CMOS gates are placed at the crossing points to interconnect two electrical lines.

In a crossbar matrix architecture, each switching cell is dedicated to a specific input to output connection, and hence it is always possible to connect idle inputs to arbitrary idle outputs independently of the already established input to output connections: this results in a strictly non-blocking (SNB) characteristic and low power consumption since only one switching cell is activated for each input signal.

Crossbar switching matrices have different insertion loss experienced by the input signals travelling through the matrix due to different optical paths and different number of switching cells traversed. In Fig. 12.10, the lowest loss is for the signal that enters at I_N and gets out at port O_1, while the highest loss is for the signals I_1-O_M (see red dashed lines). To develop scalable crossbar matrices, the loss of the waveguide crossing circuits and that of the switching cell when set in *off* must be very low (few hundredths of a dB for waveguide crossing and few tenths of a dB for the cell set in *off* in a 64 × 64 matrix); this is because all the cells traversed by the signals are set in *off* but one.

In crossbar switch, the extinction ratio requirement of the switching cell must also be high to keep the crosstalk low. This is because the various signals travelling through the horizontal lines of the matrix leak a portion of power at each of the switched *off* traversed cells, due to non-ideal extinction ratio. These unwanted leakages propagate along the vertical lines together with the wanted signal and disturb it as unwanted crosstalk.

In a N × M switch matrix, the number of switching cells is equal to N × M, while the maximum number of activated switching cells is equal to N.

Two types of silicon-integrated optical crossbar switch matrices have been investigated and demonstrated recently: the 64 × 64 digital silicon photonics MEMS switch and the 8 × 7 switch matrix based on resonant switching cells. They are presented and discussed in the following.

12.4.1.1 64 × 64 Digital Silicon Photonics MEMS Switch Matrix

A new type of monolithically integrated MEMS switch matrix in *silicon photonics* has been demonstrated and reported in [24–27]. The switching cell is based on vertical adiabatic couplers actuated by MEMS (see Fig. 12.12a).

The 64 × 64 optical crossbar switch matrix [27] was fabricated on two stacked silicon optical layers: a first 220 nm thick crystalline silicon layer on top of a 3 μm thick buried oxide (BOX) layer in a SOI wafer and a second deposited polysilicon layer.

The matrix of passive rib optical waveguides with 60 nm ridge height and 600 nm width is patterned in the crystalline silicon layer, and a multimode interference (MMI) waveguide crossing is placed at each matrix crossing point in order to achieve low loss and high isolation.

In the top polysilicon layer, a MEMS electrostatic actuator is patterned together with a vertical adiabatic coupler made by a ridge waveguide with core thickness of 300 nm and ridge height of 200 nm.

When the switching cell is in *off* (see Fig. 12.12b), the adiabatic coupler is positioned 1 μm above the passive waveguide crossing, and no coupling occurs in the top polysilicon circuit layer: the light entering in the switch cell at the “input” port

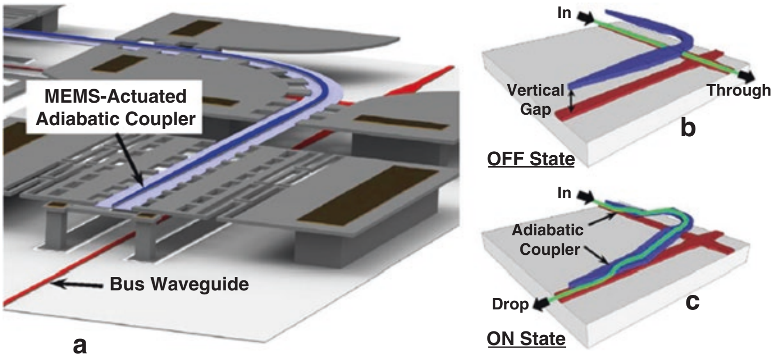


Fig. 12.12 Silicon photonics MEMS switching cell [27]: (**a**) overview of the cell, (**b**) cell set in *off*, (**c**) cell set in *on*

proceeds propagating in the passive waveguide layer toward the "through" port with very low disturbances. When the switch cell is set in *on* (see Fig. 12.12c) by imposing a control voltage at the electrodes of the electrostatic actuator, the adiabatic coupler is moved by the MEMS actuator toward the waveguide crossing, and the light in the bottom layer entering at the "input" port is coupled into the waveguide in the top layer by a first adiabatic coupler. After a rotation of 90°, obtained inside a curved waveguide, the light is coupled back in the bottom layer by a second adiabatic coupler and goes out at the "drop" port.

To digitally control the switch operations, mechanical stoppers are used to fix the vertical gap between the top waveguide and the bottom waveguide (at about 125 nm) when the switch is in *on*, while micro-springs are used to move waveguides away when the switch is in *off*.

Off-chip connection to optical fiber arrays is provided by integrated grating coupler at input and output ports.

This matrix comprises of 4096 MEMS-actuated switching cells with 110 × 110 μm^2 size. The switching cells have high extinction ratio of > 60 dB in *off* state.

The loss of the switch cell is 0.47 and 0.026 dB in *on* and *off*, respectively, while the loss of the waveguide crossing is 0.017 dB, and the chip propagation loss is 1.1 dB/cm. With these values, the on-chip loss, not including input and output fiber coupling losses, is path dependent spanning from about 0.47 dB for the shortest path (corresponding to one switch cell set in *on*) to 3.7 dB for the longest path (corresponding to 63 switch cell set in *off* with 0.026 dB/cell loss plus 1 cell set in *on*). The *on* and *off* switching time response is 0.91 and 0.28 μs, respectively. The total 64 × 64 matrix chip area, including grating couplers, is 8.6 mm × 8.6 mm. The switch cells are voltage controlled with a driving voltage of about 42 V, and the driving current needed to load the electrostatic MEMS actuator is very small due to the fact that electric circuit is purely capacitive. The operating wavelength range is as broad as 300 nm spanning from 1400 to 1700 nm.

This matrix is polarization dependent since it works correctly only with input signals having TE polarization, while for a practical system utilization, the improvement to a polarization-insensitive device is essential.

12.4.1.2 8 × 7 Switch Matrix Based on Resonant Switching Cells

An 8 × 7 crossbar switch matrix based on resonant switching cells has been demonstrated in [28]. The chip was fabricated in a SOI wafer having a 220 thick silicon layer over a 2 μm thick BOX layer. The waveguide dimensions are 220 nm × 500 nm, and the cladding layer on top of it is 1.2 μm thick.

The switching cell consists of a fifth-order ring resonator as shown in Fig. 12.13a. The five rings have a racetrack shape (to increase the coupling length and ensure a phase shift of π) with 5 μm bend radius and a free spectral range (FSR) of about 350 GHz. In Fig. 12.13b, the layout of the switching cell is shown including the optical waveguides and waveguide crossing (indicated in black) and the heaters (indicated in orange). Each ring is coupled with its neighbor by directional couplers. The crossing elements needed in each switching cell of the crossbar matrix are made by MMI waveguide crossing. The size of each switching cell is 100 × 115 μm^2.

Grating couplers are used for off-chip connection to optical fibers at input and output ports.

The physical effect used for switching is the thermo-optic effect with the rings heated by metal Ti/Pt micro-heaters placed on top of the ring waveguides and separated by a 1.2 μm silica cladding layer.

The switching cells have a passband of 100 GHz centered at wavelength 1551.35 nm, the loss is about 0.9 dB in *off* state and about 2 dB in *on* state, and each switching cell needs to be calibrated both for *on* and *off* operation by setting the bias voltage. The switch matrix extinction ratio is >19 dB, and the total loss is path

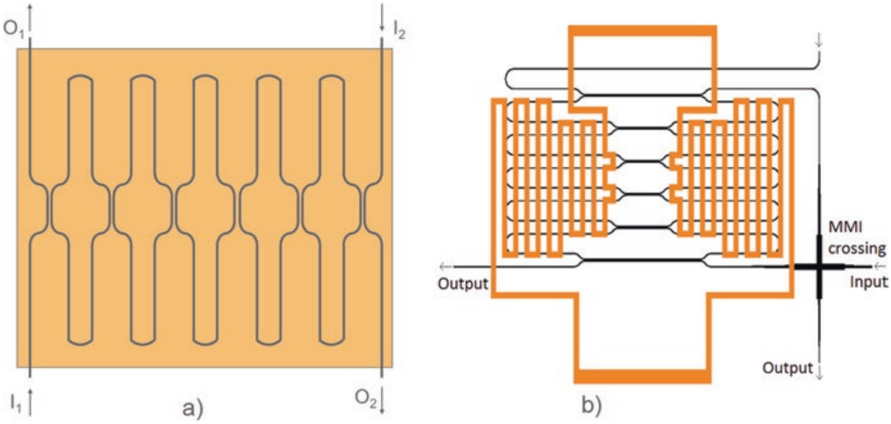


Fig. 12.13 Switching cell based on five-coupled micro-ring-resonators: (**a**) coupled rings, (**b**) layout of the complete switching cell (Courtesy of COBRA Research Institute)

dependent spanning from 14 dB for the shortest path to 25 dB in the longest path (including input and output coupling losses that count 12 dB). The *on* and *off* response time is 17 μs and 4 μs, respectively.

The drawback of this switch matrix comes from the fact that the switching cell has a limited bandwidth corresponding to the bandwidth of the micro-ring resonance. To work properly this matrix requires either a complex resonance calibration procedure at the start-up, a tuning range wide enough to comprise the system wavelength range and resonance locking to the channel wavelength during operation, or the use of fixed wavelength transmitter. This makes the matrix less attractive unless its characteristic is advantageously exploited for wavelength selective switching. The demonstrated matrix must operate with TE polarized input signals while polarization independent characteristics are necessary since the interconnect infrastructure in data center is based on single mode fibers.

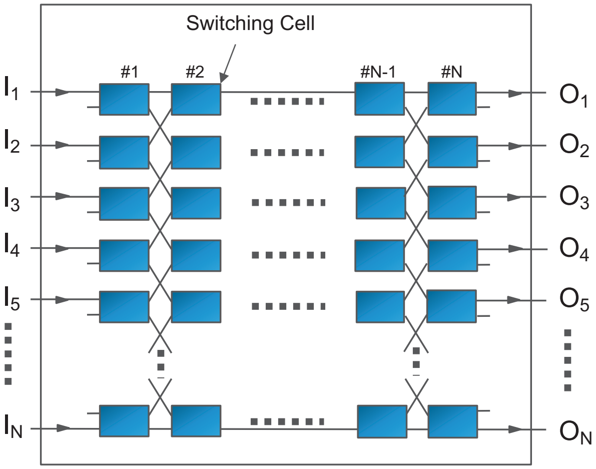
12.4.2 PILOSS Switch Architecture

In crossbar switch matrices, as described above, the different signals experience different insertion losses depending on their internal path through the matrix. This aspect limits the scalability unless very low loss of few hundredths of a dB is achieved for the switching cells set in *off*. That is because a high differential loss between the signals at the matrix output results in a crosstalk increase especially in multistage optical switching fabric systems. To avoid this detrimental effect, channel power equalization has to be implemented with consequent increase in loss and device complexity. Alternatively, a *path-independent insertion loss (PILOSS) matrix* architecture may be adopted. It was proposed originally in [29] and is schematically presented in Fig. 12.14. Each cell in the matrix is connected to the four similar cells at the corners of the square surrounding that cell, and only the cells in the first and last rows are connected to the adjacent cells in the same row. The switching cell (see Fig. 12.15) is a 2 × 2 switch in which, in *off* state, signals at inputs I_1 and I_2 follow the diagonal paths to O_2 and O_1, respectively, while with the switch cell set in *on*, the signal at I_1 is transferred to O_1 and that at I_2 to O_2.

In a N × N PILOSS switch matrix, the number of cells the signals traverse is always constant, is independent to the path, and is equal to N. Similar to the crossbar architecture, for each path, only one switching cell is set in *on*. As an example, three path setups, namely, I_1–O_3, I_3–O_7, and I_5–O_4, are shown in Fig. 12.16 for an 8 × 8 PILOSS matrix.

PILOSS matrix, like the crossbar counterpart, is strictly non-blocking, and the number of switching cells in a N × N matrix is equal to N^2. The same considerations made above for the scalability of crossbar matrices are valid here, concerning the requirements of very low loss and high extinction ratio of the switching cells and waveguide crossing.

Fig. 12.14 PILOSS switch matrix architecture



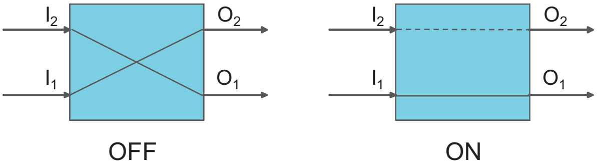


Fig. 12.15 PILOSS 2 × 2 switching cell

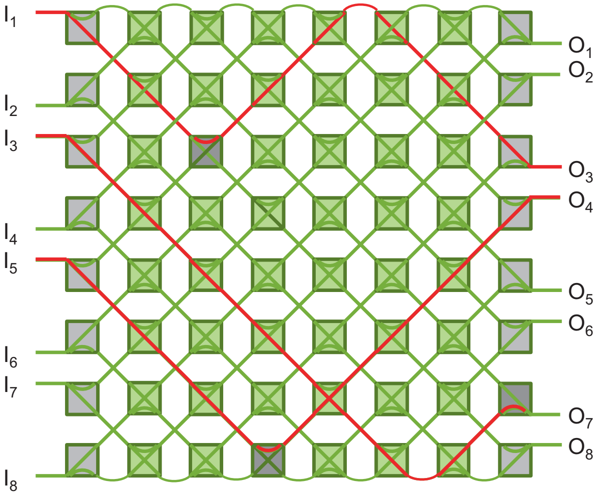


Fig. 12.16 Example of three optical path setups (*red lines*), I_1–O_3, I_3–O_7, and I_5–O_4, in an 8 × 8 PILOSS matrix: switching cells in *on* are indicated with *dark boxes*

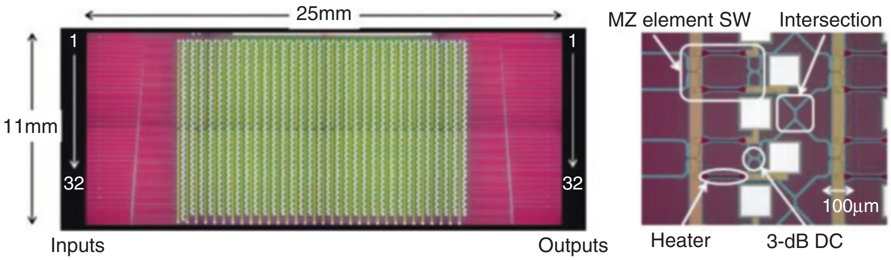


Fig. 12.17 32 × 32 Si-wire switch chip images [31]

Two types of silicon-integrated PILOSS switch matrices have been investigated and proposed recently: the 32 × 32 Si-wire switch [30, 31] and the micro-opto-electro-mechanical system (MOEMS) [32].

12.4.2.1 32 × 32 Si-Wire Switch Matrix

The 32 × 32 PILOSS matrix [30, 31] is based on the 2 × 2 MZI switching cell type described above. It was fabricated on 300 mm SOI wafer using a CMOS-compatible process. The chip integrates 1024 MZI cells (see Sect. 12.3.1), each thermally controlled by a TiN micro-heater on both arms of the interferometer and 961 waveguide crossings. The chip area is 11 × 25 mm^2 (see Fig. 12.17). In the MZI cells, the 3 dB couplers are directional couplers working with the TM mode. The waveguides are 430 nm wide and 220 nm high, buried in a 1.5 µm thick SiO_2 cladding. The waveguide crossings are made by 0 dB directional couplers (in which all the power is transferred to the adjacent waveguide) with simulated loss and isolation of 0.1 dB and 45 dB, respectively. The estimated loss of the switching cell (including one waveguide crossing) is 0.44 dB, and the extinction ratio is <35 dB but in a bandwidth of only 2.3 nm centered at wavelength 1545 nm, due to the use of wavelength-sensitive directional couplers. The complete device has an on-chip insertion loss of about 15 dB (not including fiber coupling loss), and the worst case crosstalk is estimated to be about –20 dB in a band of 1.8 nm. The switch response time is of about 30 µs, and the electrical power to set the switch cell in *on* is 45 mW, while the power to keep it in *off* is 1.5 mW (that is the power to trim the MZI).

This matrix is polarization sensitive since the directional couplers have been designed for TM polarization. Future efforts should be dedicated to widen the operating bandwidth of the device (due to the directional couplers), to reduce the on-chip loss, and to provide polarization insensitivity. A very recent work has been published in [60] reporting on a smaller size 8×8 Si-wire switch matrix with improved bandwidth and polarization diversity architecture.

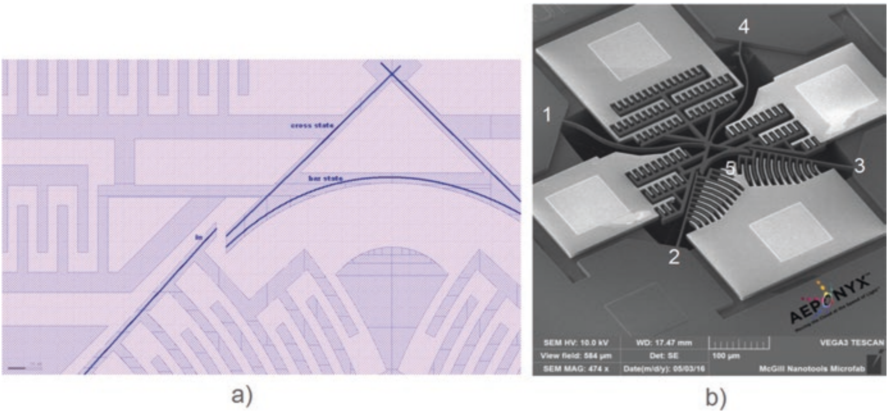


Fig. 12.18 MOEMS switching cell, (**a**) details of the waveguide coupling, (**b**) SEM picture of the switching cell (Courtesy of Aeponyx)

12.4.2.2 Micro-Opto-Electro-Mechanical Switch (MOEMS) Matrix

A new type of silicon-integrated switching cell for use in a PILOSS matrix has been proposed in [32]. It is based on a micro-opto-electro-mechanical system (MOEMS) in which comb electrostatic actuators are used to move optical waveguide ends causing light coupling into two alternative paths: the one with the straight waveguide or the one with the curved waveguide (see Fig. 12.18a). The cell consists of a suspended element including the crossing waveguide segments 1 and 4 together with a curved waveguide segment 5 and a fixed part that includes the waveguide segments 2 and 3 (see Fig. 12.18b). When the switch cell is in *off* (*cross* state), the ends of waveguides 1 and 4 in the suspended part are aligned to the ends of waveguides 2 and 3 of the fixed part, respectively, and the two input signals (e.g., at waveguides 2 and 3) cross the cell with very low coupling loss (going out through waveguides 1 and 4). With the switch cell in *on* (*bar* state), the ends of the waveguides 2 and 3 in the fixed parts are aligned to the left and right ends of the curved waveguide 5 in the suspended part, and the input signal at waveguide 2 is redirected to output waveguide 3 with low coupling loss. The waveguide alignment occurs through a thin air gap of about 50 nm or less.

The alignment procedure works as follows, and it applies for setting either *on* or *off* state: firstly, the left and right actuators at the bottom of Fig. 12.16a rotate the waveguides in the fixed part to open the two air gaps. Secondly, the suspended part with the crossing and curved waveguides translates up or down. Finally, the waveguides in the fixed part rotate back to close the air gaps.

The waveguides are composed of a Si_3N_4 core and a SiO_2 cladding to ensure low propagation losses and polarization-independent operation. The expected cell loss is about 0.2 dB (including the cell internal waveguide crossing loss and the loss in the thin air gap between fixed and movable waveguides) in *off* (corresponding to *cross* state) and 0.5 dB in *on* (corresponding to *bar* state) with a slight increase with respect to the *cross* state due to the propagation in the curved waveguide [58].

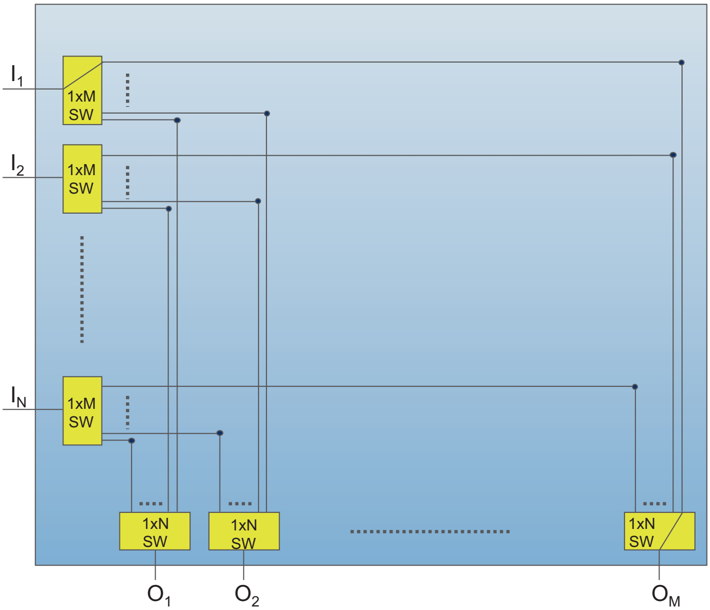


Fig. 12.19 N × M switch and select architecture

The waveguide crossing loss is estimated to be about 0.01 dB, and the propagation loss in the waveguide is estimated to be 0.6 dB/cm.

A high cell extinction ratio (>60 dB) can be easily achieved by separating the curved and straight waveguides in the suspended part by few μm.

The cell dimensions are 350 × 350 μm^2, and the cell response time is <200 μs.

The switch is controlled by analog voltage signals of about 150 V for the gap closer actuator (once optimized could be reduced to 100V) and about 200 V for the actuator that translates the suspended part. A 48×48 switch matrix based on the MOEMS switch elements described above is under development and the total estimated insertion loss is about 12 dB including 0.6 dB of input and output edge coupling loss. The operating bandwidth is in principle very wide, limited only by the Si_3N_4 transparency.

This matrix is slightly larger and has a longer response time with respect to the matrices presented above based on Si nanowires, but it has the relevant advantage of being polarization insensitive.

12.4.3 Switch and Select Architecture

The N × M *switch and select matrix architecture* is depicted in Fig. 12.19. It comprises of an array of N 1 × M input switches, a large passive interconnect network, and an array of M 1 × N output switches.

The input to output connections are established by appropriate selection of the paths at both the input switch array and the output switch array as indicated in the Fig. 12.19 in which the connection between I_1 and O_M is shown.

The switching elements here, instead of being distributed across the two-dimensional matrix, like in crossbar and PILOSS architectures, are concentrated in two linear switch arrays. In this switch and select architecture, each couple of switches (in the input and output arrays) is dedicated to a specific input to output connection, and hence, like the crossbar and PILOSS, it is strictly non-blocking (SNB) as it is always possible to connect idle inputs to arbitrary idle outputs independently of the already established input to output connections.

Considering a N × N switch matrix, the 1 × N switches in the two arrays could be made by cascading a number of 1 × 2 MZI-based switching cells, similar to the ones shown in Fig. 12.5 but with only one input port. The number of 1 × 2 cells in the matrix scales with the number of ports as 2 N*(N − 1). All input to output connections traverse a constant number of 1 × 2 switching stages (switching cells), like in PILOSS architecture, but here this number is lower being equal to $2^*\log_2 N$ instead of N. This is obtained by increasing the complexity of the passive interconnect network between the two switch arrays that must interconnect N^2 input waveguides to N^2 output waveguides. This represents the fundamental limitation to the scalability of this architecture as long as it is implemented in single optical layer photonic chips. In such an architecture, some signals, in the worst case, must traverse up to $(N - 1)^2$ waveguide crossing points resulting in a big insertion loss and in high differential loss between those unfortunate signals and the lucky signals that do not traverse any crossing. In a 64 × 64 switch matrix, the number of waveguide crossing traversed by the signals in the worst case is up to 3969, and even using waveguide crossings with very low loss of about 0.01 dB, the insertion loss and differential loss experienced by the signals could easily reach up to 40 dB that is an unacceptable value for an optical switch.

However, this architecture could be attractive in the future if multilayer photonic chip will be developed. 3D heterogeneous photonic integrated circuits have been presented in [33], and multilayer photonic chips compatible with CMOS fabrication processes and based on a Si_3N_4-SOI platform have been investigated in [34]. With a two-photonic-layer chip, the large passive interconnect network of Fig. 12.19 could be realized with the vertical connection laid out in one layer and the horizontal connections on the other layer in such a way that waveguide crossings are completely avoided and the signals could exchange layer by means of low-loss optical vias.

12.4.3.1 8 × 8 Switch and Select Optical Matrix

An 8 × 8 switch and select matrix has been implemented and reported in [35]. The 1 × 8 switch in the arrays is implemented with three stages of 1 × 2 MZI-based switching cells, each one using thermally tuned phase shifters on both arms and two adiabatic power splitters having a bandwidth wider than directional coupler-based 3 dB power splitters and a loss lower than MMI-based 3 dB power splitters. The phase shifters have etched trench and partially removed silicon to reduce power consumption and thermal crosstalk, and they are controlled by 224 heaters made by a thin TiN layer. The passive interconnecting network is implemented on a single layer, and the waveguides are joined by the use of low-loss tight bends. The waveguide crossings are the most critical circuit in this architecture, and they have a simulated loss of 0.015 dB and a crosstalk of about −40 dB at 1550 nm wavelength. The on-chip loss, not including the fiber coupling loss, was estimated as 4 dB for the longest path, and it includes losses of 3 + 3 stages of MZI, waveguide bend, and propagation in a 1.6 cm long waveguide.

The extinction ratio of the switching cell is >16 dB in an 80 nm bandwidth from 1500 to 1580 nm. Due to the fact that for establishing an input to output connection, the path selection is made by cascading an input and an output switch, the port to port isolation of this switch matrix is increased to >30 db. The chip footprint is 8 mm × 8 mm, and the power needed by the 1 × 2 MZI switching cell is as low as 1.5 mW, while the power consumption of the matrix when all the eight connections are established is about 70 mW. The high energy efficiency is due to the isolation trenches and substrate removal in the phase shifters, but the thermal isolation affects also the response time that increased to 250 μs.

This switch, as most of the other types of demonstrated silicon photonics switching matrices presented above, works with the fundamental TE mode only, and for a practical use, it needs to be enhanced with polarization-independent characteristics. High port count in this matrix can be achieved with a photonic chip having two layers of waveguide (one for vertical and one for horizontal waveguides) in order to avoid waveguide crossings with the related losses.

12.5 Integrated Matrices with ns Response Time for Optical Packet Switching

Integrated switching matrices with very fast response time of the order of nanoseconds have recently captured high interest since, by leveraging on silicon photonics integration, they allow the implementation of optical packet switching (OPS). In OPS short data packets are exchanged between servers in a performance optimized data center (POD) entirely in the optical domain. For this application, a short reconfiguration time of the switch is needed [36, 37]. OPS enables a high energy efficiency intra-data center networking since there is no need of electro-optical and optical-electrical conversion prior to an electrical packet switch (EPS), and it also permits to obtain a lower latency. With respect to optical circuit switching (OCS), OPS has a more efficient bandwidth utilization, and it could become, in the future, the ultimate and sole switching technology replacing the electrical packet switching for intra-data center networking.

In the following, the architectures and technologies for silicon-integrated optical packet switch matrices are presented and discussed.

12.5.1 Benes Switch Architecture

The optical switch matrix architectures presented in Sect. 12.4, used for optical circuit switching, are all strictly non-blocking. In these architectures, the number of traversed switching cells is equal to N for PILOSS, 2N − 1 for the worst case in crossbar. To obtain a good scalability in such architectures characterized by a large number of traversed cells, the cell loss must be very low, as already discussed in the above sections. Switching cells based on MEMS and micro-ring resonators have loss as low as few tenths of a dB in through port (*off* state), while the loss in *cross* is less critical due to the fact that only one cell is activated for each input to output connection.

Switching cells based on MZI have been proposed in PILOSS architectures in conjunction with micro-heaters as active element since the loss introduced by the presence of the heaters on top of the waveguides is quite low. But in high-speed switching matrices, the plasma dispersion effect has to be exploited to activate the switching cells, and the total losses become too high and limit the matrix scalability to few ports (8 × 8 or 16 × 16). This is due to the fact that by injecting carriers in the phase shifters of the MZI switching cells, not only the refractive index is changed but, as non-wanted effect, the optical loss is increased due to free carrier absorption (see Sect. 12.2).

For this reason, the fast optical switching matrices proposed and experimentally demonstrated so far, instead of being SNB, adopt rearrangeable non-blocking architectures (RNB) because of a reduced number of traversed switching cells and waveguide crossings, resulting in a lower loss and lower crosstalk. However, the drawback is that, even if it is always possible to connect an idle input port to an arbitrary output port, in some cases, it may be necessary to apply an internal rearrangement of the connections already established. This drawback implies the implementation of complex switching control algorithms and it does not affect the switch performances only in switching matrices operating in synchronous time slotted mode. In such matrices all the connections are set up and torn down periodically at the same time as in the system architecture presented in [37]: only in this condition an RNB matrix becomes SNB [40].

RNB matrix architectures adopted in silicon photonics switches are the Benes architecture [38–40]. They are attractive for the low-cost index (number of switching cells in a matrix) that increases asymptotically as $O[N \log_2 N]$, to be compared with the cost index of crossbar and PILOSS that increases as $O[N^2]$.

The *Benes silicon photonics switch matrix* is constructed starting from the 2 × 2 MZI switching cell, already presented in Sect. 12.3.1 and shown in Fig. 12.5. A Benes matrix with an arbitrary size of $N \times N$ is recursively constructed starting from a top Benes sub-matrix of size $N/2 \times N/2$ and an equal bottom Benes sub-matrix of equal size. The two sub-matrices are interconnected by N/2 2 × 2 MZI switching cells included in two extra-symmetrical stages, named Banyan stages. In the Banyan stages, one port of each 2 × 2 switching cell is connected to the top sub-matrix and the other to the bottom sub-matrix, as depicted in Fig. 12.20.

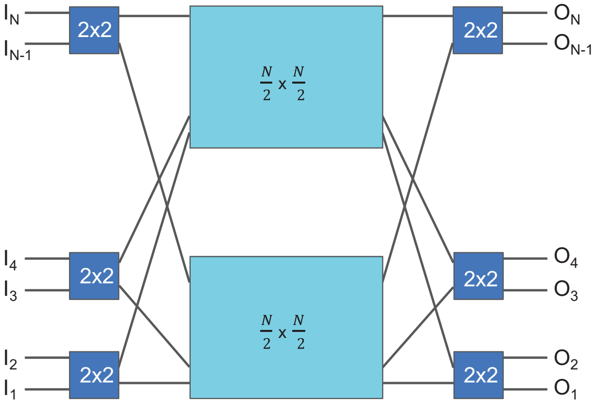


Fig. 12.20 Construction of an arbitrary size Benes switching matrix

A N × N Benes switch matrix has a number of stages equal to:

$$S_N = \left(2\log_2 N - 1\right) \tag{12.16}$$

The number of switching cells in each stage is:

$$C_S = 2^{\left(\log_2 N - 1\right)} \tag{12.17}$$

An example of the Benes matrix scaling is depicted in Fig. 12.21 starting with the 2 × 2 MZI-based switching cell and then showing the 4 × 4 matrix and the 8 × 8 matrix.

Two Benes switch matrices have been demonstrated recently and are reported in the following.

12.5.1.1 16 × 16 Benes Switching Matrix

A 16×16 integrated silicon photonics switching matrix based on MZI cells with Benes architecture has been presented in [41]. It comprises seven stages and eight switching cells on each stage and it was fabricated on a SOI wafer with a thin 220 nm silicon layer on top of a 2 μm thick BOX layer. The MZI switching cells were similar to the type shown in Fig. 12.5 with two broadband MMI-based 3 dB couplers and two arms of equal length: one arm is equipped with both a high-speed p-i-n phase shifter and a thermo-optic phase tuner, while the other arm is equipped with the thermo-optic phase tuner only. The thermo-optic tuner sections are used for compensating phase error due to fabrication and to set *cross* state at the initialization. They are made by TiN micro-heaters with 300 μm length placed on top of the waveguides. Air trenches surrounding the waveguides are used to improve

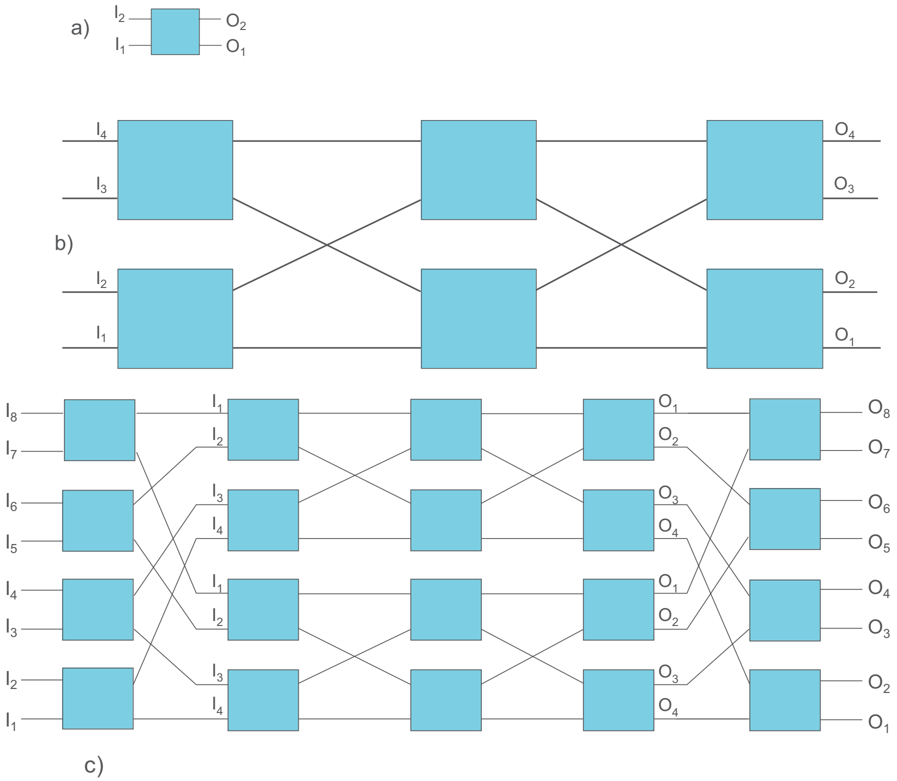


Fig. 12.21 Scaling a Benes matrix: (**a**) 2 × 2 MZI switching cell, (**b**) 4 × 4 Benes matrix, (**c**) 8 × 8 Benes matrix

energy efficiency of the phase tuner. The p-i-n diode phase shifter, used for fast switching, has a length of 380 μm and heavily doped *p*+ and *n*+ regions with carrier concentration of about 10^{20} cm^{-3}.

The switching cell average loss when the state is in *cross* is about 0.4 dB in the wavelength range of 1530–1590 nm, and the extinction ratio is >30 dB in a 30 nm wavelength band. When the switching cell is in *bar* state, the loss increases to 1 dB, and the extinction ratio decreases to 18 dB due to free carrier absorption in the p-i-n phase shifter. The on-chip insertion loss (not including input and output coupling loss) is 6.7 dB for a path of all *cross* states and 14 dB for a path of all *bar* state. The chip crosstalk is −20 dB for all *cross* states and −10 dB for all *bar* states. The response time is about 3 ns and the chip area is 10.7 x 4.4 mm^2. The worst case power consumption of the thermo-optic tuners is 881 mW and that of the plasma dispersion-based phase shifters is 289 mW. This device works with TE polarized input signals.

12.5.1.2 32 × 32 Benes Switching Matrix

A higher-radix, high-speed monolithically integrated silicon photonics matrix has been demonstrated in [42]. Similar to the previous matrix, it also has Benes architecture and MZI-based switching cells. The matrix is composed of 9 stages with 16 switching cells at each stage. The chip was fabricated with a 180 nm CMOS production process and comprised of 144 switching cells and 416 waveguide crossing with a chip area as small as 12.1 × 5.2 mm^2. The cells are made by two MMI 3 dB couplers and two p-i-n-based phase shifter at both arms of the interferometer having 200 μm length and 450 nm width, driven in push-pull configuration.

The matrix insertion loss was 12.8 dB, and the crosstalk was in the range from −19.2 to −25.1 dB in a 50 nm wavelength bandwidth from 1525 to 1575 nm.

This device to operate correctly needs TE polarized input signals and the response time of the MZI switching cells is 1 ns.

In [42] together with this 32 × 32 fast switching matrix, a 64 × 64 Benes matrix based on MZI cells is presented using the thermo-optic effect as switching mechanism. It has lower loss (12.8 dB) and lower crosstalk (<30 dB) in the 50 nm wavelength bandwidth from 1525 to 1575 nm and a slower response time.

12.6 Silicon Photonics Wavelength-Selective Switch Matrix

In many of the optical switching network architectures proposed so far for data center networking [43–46], wavelength division multiplexing has been introduced in order to increase the link capacity of each single fiber connection. In such architectures, *wavelength selective switches* (*WSS*) work in conjunction with optical space switches to provide connectivity among data center equipments (compute and storage nodes).

Similarly, to the space switch matrices presented and discussed in the above sections, also for wavelength selective switching, high-capacity, highly integrated, and low-cost devices are needed.

A promising technology for the realization of large-scale integration silicon photonics WSS is the one based on micro-ring resonator (MRR) switch elements and researched in the EU FP7 project IRIS[1] (Integrated Reconfigurable silicon photonic-based optical Switch) [47]. The IRIS switch device has been implemented as 4 × 8 transponder aggregator (TPA) block of a colorless-directionless-contentionless reconfigurable optical add drop multiplexer (*CDC ROADM*), but a similar switch architecture can be used for a NxN WSS device for data center applications.

The schematic block diagram of the IRIS WSS switch is depicted in Fig. 12.22.

[1] The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007–2013) under grant agreement n° 619194.

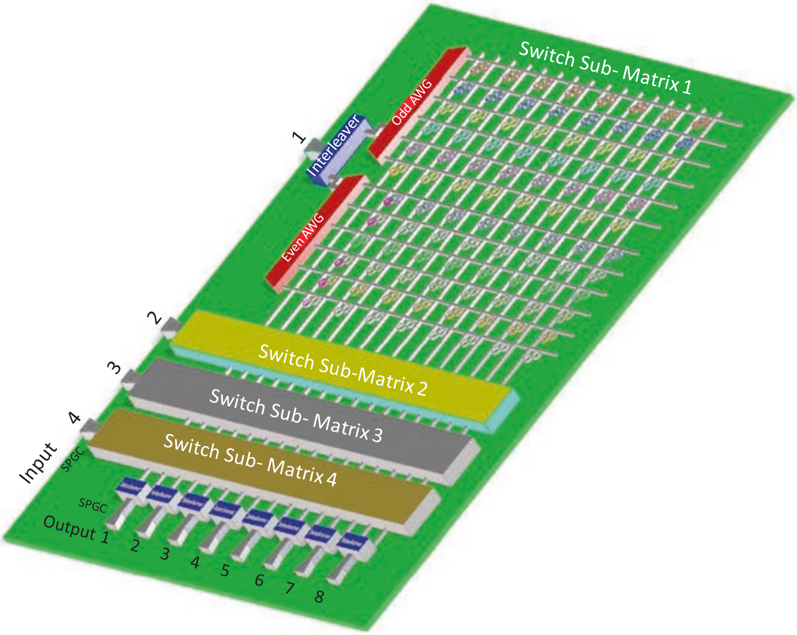


Fig. 12.22 IRIS switch architecture

It has four input line ports and eight output local ports. At each input line port, it receives a comb of 12 multiplexed wavelengths, 200 GHz spaced. Single polarization grating couplers (SPGC) are used to couple the optical signals from the input fiber array to the chip silicon waveguides. Interleaver circuits separate the input channels into odd and even in order to double the channel spacing and relax the channel isolation requirement of the following arrayed waveguide grating (AWG)-based demuxes and that of the switch matrix. Two AWG demux circuits, the odd one and the even one, are used to separate the various channels in the wavelength combs received at the device input. At the demux output, the individual signals are sent to the switching matrix constituted by a crossbar of optical waveguides equipped with micro-ring resonator (MRR) switch elements at each crosspoint. The matrix has 4 × 12 rows (each row dedicated to a certain wavelength) and 8 × 2 columns (one column for odd wavelengths and one for even wavelengths). In the switching matrix, the channels travel along the row corresponding to the wavelength until they intersect the column connected to the wanted local output port to which the channel has to be dropped. At this crossing point, by thermally tuning the corresponding MRR in resonance with the signal wavelength, the signal is transferred to the corresponding drop column (see Fig. 12.23). Note that an interleaver recombines the odd and even wavelengths coming from two distinct columns before the drop port. SPGCs are used at the drop ports coupling the output signals to the optical fibers. The same device described above can be used in the reverse direction to add wavelengths from local ports to line ports.

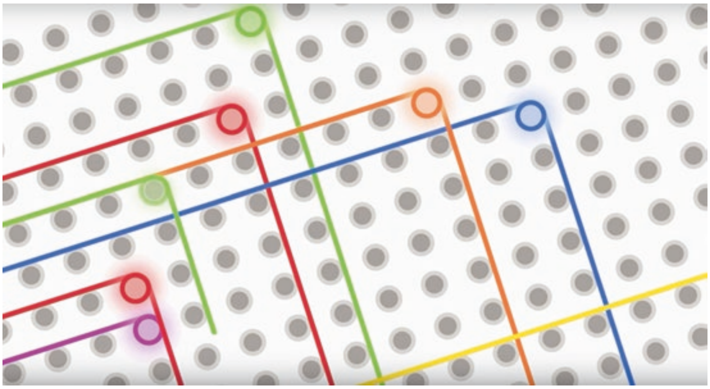


Fig. 12.23 Working principle of the IRIS wavelength selective switch matrix

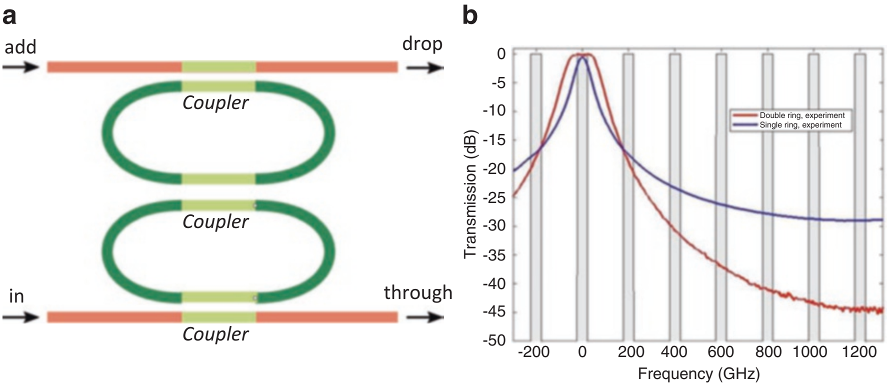


Fig. 12.24 Double ring circuit (**a**), single and double ring frequency response (**b**). *Gray bars* indicate 50 GHz bandwidth

In a TPA device, only one wavelength is dropped on each column due to the fact that a TPA connects the line switch to the transponders, while in a NxN wavelength selective switch for data center, more than one wavelength can be dropped on each column.

The IRIS switch has been implemented on a 8” SOI wafer with 220 nm thin silicon layer on a 2 µm BOX layer. The main waveguides of the circuits are 480 nm width strip waveguides.

High ER of the switching cell in *off* state is achieved by shifting the MRR far enough from the channel wavelength. In order to achieve a wide-enough 1-dB bandwidth of about 50 GHz and a high-enough ER > 35 dB, the MRR-based switch elements are made with two-coupled micro-rings as shown in Fig. 12.24a. In Fig. 12.24b, the spectral response of the MRR switching cell is shown for both single ring (blue curve) and two-coupled rings (red curve) .

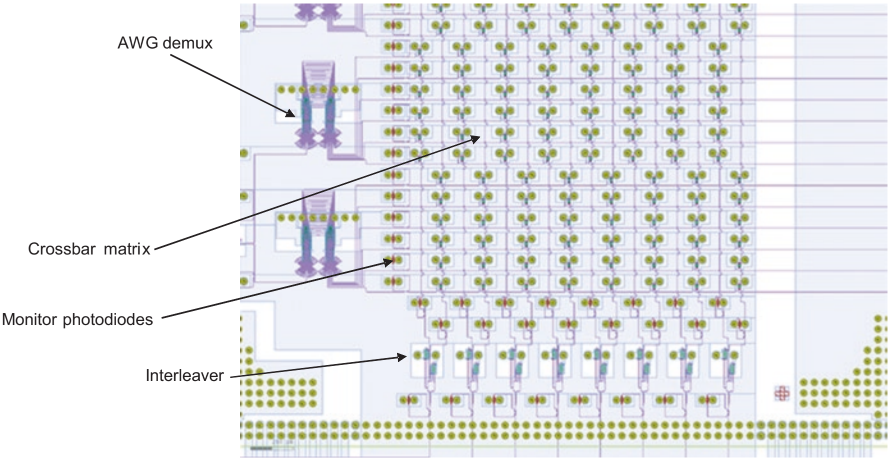


Fig. 12.25 IRIS circuit layout details

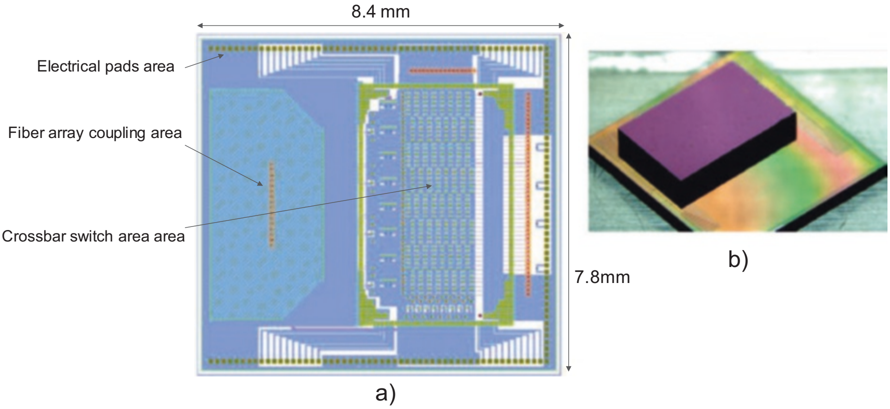


Fig. 12.26 (**a**) IRIS chip layout, (**b**) 3D integration with the control electronic chip (Courtesy of LETI fab)

Details of one section of the chip including the optical circuits are shown in Fig. 12.25.

The chip size is 8.4 mm × 7.8 mm (see Fig. 12.26a). A large portion of the chip area is dedicated to fiber coupling and electrical pads, while a smaller portion comprises photonic circuits.

The two-coupled ring loss with the switch element in *off* is <0.02 dB, while the loss when the switching cell is in *on* is <1 dB. MRR ER at 1000 GHz shift is >40 dB, and the tuning efficiency is 23 mW/FSR (with an FSR of about 20 nm). Crossing loss is <0.03 dB.

The interleavers are infinite impulse response (IIR) filters consisting of a MZI and a ring resonator on one arm [48]. Interleaver channel isolation is >20 dB, while the loss is <1 dB [49]. The AWG channel separation is 400 GHz, the insertion loss is <3 dB, and the channel isolation is >25 dB.

3D integration of the photonic chip with its control electronic chip is shown in Fig. 12.26b: it has been realized by the use of micro-pillar interconnect arrays.

Like most of the silicon photonics switching matrices demonstrated so far, also IRIS switch works with input signals having TE polarization, and further developments are needed for a polarization-insensitive device.

12.7 Switch Matrices Comparison Table

The characteristics of the various switch matrices discussed in the above sections are summarized in Table 12.1.

All matrices are polarization sensitive but one, and they need a further development step to implement polarization-insensitive devices.

For optical circuit switching, the 64 × 64 digital MEMS switch has the lowest on-chip loss and fastest switching time with good performances in terms of extinction ratio and crosstalk but it is polarization sensitive. The MOEMS switch has promising performances in terms of loss, extintion ratio and polarization insensitivity even if it has a larger size and slower response time but a high radix switch matrix is still under development. The 8 × 7 resonant cell matrix is fundamentally narrowband, and it is more suitable for wavelength selective switching, while the 32 × 32 Si-wire switch has a limited spectral response and it is polarization sensitive but very promising improvements in terms of bandwidth and polarization insensitivity have been achieved very recently on a 8 × 8 matrix. The 8 × 8 switch and select matrix has scalability issue, but it could be an interesting option in case of implementation in a multilayer photonic chip. Fast 16 × 16 and 32 × 32 Benes matrices are interesting candidates for optical packet switching applications provided that they become polarization insensitive. The same is valid for IRIS switch that is an attractive technology for integrated wavelength selective switching devices.

12.8 Perspectives and Research Directions

The advancements reported above in the implementation of scalable, highly integrated optical switching matrices in silicon photonics are encouraging for the future realization of commercial products that could enable a pervasive use of optical switching in data centers. These optical matrices will have a high port count (64x64 or more), low power consumption in the order of few watts, high speed either in the order of microseconds for optical circuit switching or of nanoseconds for optical packet switching, low chip area of about tens of squared mm, low cost, wide

Table 12.1 Integrated switch matrices comparison

	Parameter							
Switch type	Cell loss (dB)	Pol. independent	Cell size (μm^2)	ER (dB)	Blocking perf.	Speed (μs)	Bandwidth (nm)	Refs.
64 × 64 Digital MEMS	0.026/0.47 (*off/on*)	No	110 × 110	>60 (in *off*)	SNB	0.28/0.91 (*off/on*)	300	[27]
8 × 7 Resonant cells	0.9/2 (*off/on*)	No	100 × 115	>19.5 (all matrix)	SNB	4/17 (*off/on*)	0,8	[28]
32 × 32 Si-wire	0.44	No		>35	SNB	30	2.3	[30, 31]
MOEMS	0.2/0.5 (*off/on*)	Yes	350 × 350	>60	SNB	200	Very wide (Si_3N_4 transparency)	[32, 58]
8 × 8 Switch and select	<4 (all matrix)	No	8 × 8 mm^2 (all matrix)	>30	SNB	250	80	[35]
16 × 16 Benes	0.4/1 (*cross/bar*)	No	10.7 × 4.4 mm^2 (all matrix)	>30/18 (*cross/bar*)	RNB	0.003	60	[41]
32 × 32 Benes	12.8 (all matrix)	No	12.1 × 5.2 mm^2 (all matrix)	>−19.2 (crosstalk of all matrix)	RNB	0.001	50	[42]
IRIS switch	0.02/1 (*off/on*)	No	100 × 100	>40 (in *off*)	SNB	<4	0.4	[47]

transmission band supporting switching of high-speed signals of >100 Gbps, and wide wavelength operating range (>30 nm).

However, some technical challenges have to be met for the realization of commercial devices in the near future.

The switch matrix loss should be further reduced by improving the loss characteristics of the switching cells and reducing the fiber coupling loss. Heterogeneous integration of optical gain blocks in the switch matrix to compensate for internal losses has been proposed. Demonstration of semiconductor optical amplifier (SOA) array integration and packaging for a scalable silicon photonics switching matrix has been presented in [50]. The integration of SOAs in a high radix switching matrix has the drawback of increasing significantly the power dissipation in the chip and it requires an appropriate packaging for the heat removal that impacts the device form factor. However, it could allow not only the use of simpler optical transceivers with relaxed power budget but also the realization of higher radix strictly non-blocking matrix architectures since a higher number of cascaded switching cells are permitted by the loss compensation. The impacts on the optical signal integrity caused by ASE noise accumulation in the cascaded SOA, especially in a large multi-stage optical switch fabric, should be investigated.

In some of the presented switching cell types, the extinction ratio, the bandwidth, and the operating wavelength range must be improved to become really attractive. Since the switch matrix performances are determined not only by the characteristics of the switching cells but also by the matrix topology, new matrix architectures able to reduce crosstalk without increasing too much the loss and complexity needs to be investigated. In [59] a 32 × 32 silicon photonics matrix of thermally controlled MZI switching cells has been realized by modifying a dilated Benes architecture. It has low crosstalk but it is reconfigurable non blocking and for use in optical circuit switching blocking could occur and has to be minimized to avoid limitations in the full utilization of the switch capacity.

Most of the integrated switching matrices demonstrated so far are polarization sensitive, while polarization insensitivity is an absolute characteristic for use in data centers since the optical fiber infrastructure is made by single-mode fibers: future efforts should be dedicated to realize large-scale integration devices with polarization-insensitive characteristics. This can be achieved either by implementing polarization diversity schemes [51, 52] or, when possible, by designing the waveguide circuits and the switching cell structures [53] with low polarization sensitivity.[2]

An effective integration between the optical chip and its control electronic complement is fundamental. Due to the very high interconnection density required, two different approaches have been investigated, the monolithic integration of control electronics with the photonics [54–56] and the 3D hybrid integration with the electronic chip on top of the optical chip, interconnected to each other by micro-solder bumps [47]. Both approaches need further development to reach the maturity required for commercial deployment.

[2] After the preparation of the manuscript an interesting paper has been published presenting a work on a polarization insensitive silicon photonics 50 × 50 switch matrix [53].

The manufacturing processes should be improved to reduce the fabrication imperfections of the integrated optical circuits that require fine tuning to compensate with the consequent increase in power consumption and calibration complexity. Moving the silicon photonics fabrication to more advanced CMOS nodes and the use of silicon wafer with a better thickness uniformity are strongly recommended.

Last but not the least is the packaging issue. Due to the high impact of packaging on the device cost and since the number of input and output fiber interfaces is high, new technological advances are definitely needed. Two packaging aspects should be researched: the realization of a high-density I/O connection of silicon photonic chip with optical fibers and the realization of low-cost, removable, low-loss, and mass-manufacturable optical connectors. Experiments on a high-dense optical packaging for a high-radix silicon photonic switch are reported in [57] using a hexagonal pitch-reducing optical fiber array (PROFA).

The implementation of silicon photonics high-radix switching matrices is a research area in a rapid evolution, and it may be envisaged that completely new types of switching cells based on new technologies and with low loss and low crosstalk, suitable for large-scale integration, will be invented in the near future.

Acknowledgments The authors are grateful to the partners of FP7-IRIS project and they wish to thank Philippe Babin and Francois Menard from Aeponyx Inc. for sharing the information on the MOEMS switch matrix.

References

1. K.J. Barker, A. Brenner, R. Hoare, A. Hoisie, A.K. Jones, D.J. Kerbyson, D. Li, R. Melhem, R. Rajamony, E. Schenfeld, S. Shao, C. Stunken, P. Walker, On the feasibility of optical circuit switching for high performances computing systems. in *SC '05: Proceedings of the 2005 ACM/IEEE Conference on Supercomputing*
2. A. Vahdat, H. Liu, X. Zhao, C. Johnson, The emerging optical data center. in *Proceedings of OFC*, 2011
3. Y. Ohsita, M. Murata, Data center network topologies using optical packet switches. in *32nd International Conference on Distributed Computing Systems Workshops*, 2012
4. R.A. Soref, B.R. Bennet, Electrooptical effects in silicon. IEEE J. Quant. Electron. **QE-23**(1) (1987)
5. T. Pinguet, G. Armijo, J. Balardeta, S. Barabas, B. Chase, Y. Chi, A. Dahl, P. De Dobbelaere, Y. De Koninck, S. Denton, M. Eker, S. Fathpour, D. Foltz, F. Gholami, S. Gloeckner, K.Y. Hon, S. Hovey, S. Jackson, W. Li, Y. Liang, M. Mack, G. Masini, G. McGee, A. Mekis, S. Pang, M. Peterson, L. Planchon, K. Roberson, S. Sahni, J. Schramm, M. Sharp, C. Sohn, K. Stechschulte, P. Sun, G. Vastola, S. Wang, G. Wong, Xu, K. Yokoyama, S. Yu, R. Zhou, Advanced silicon photonic transceivers. in *IEEE 12th International Conference on Group IV Photonics (GFP)*, Vancouver, 2015
6. M. Nedeljkovic, R. Soref, G.Z. Mashanovic, Free-carrier electrorefraction and electroabsorbtion modulation predictions for silicon over the 1–14-um infrared wavelength range. IEEE Photon. J. **3**(6) (2011)
7. N. Dupuis, Technologies for Fast, Scalable Silicon Photonics Switches, Photonics in Switching (2015)

8. G.T. Reed, G. Mashanovihc, F.Y. Gardes, D.J. Thomson, Silicon optical modulators. Nat. Photon. **4** (2010)
9. B.G. Lee, N. Dupuis, P. Pepeljugoski, L. Schares, R. Budd, J.R. Bickford, C.L. Schow, Silicon photonic switch fabrics in computer communications systems. J. Lightw. Technol. **33**(4) (2015)
10. L. Chen, Y. Tang, J.E. Bowers, L. Theogarajan, CMOS Enabled Silicon Photonics for Data Center Packet Switching. in *IEEE/MTT-S International Microwave Symposium Digest*, 2012
11. M. Webster, K. Lakshmikumar, C. Appel, C. Muzio, B. Dama, K. Shastri, Lowpower MOSCapacitor based silicon photonic modulators and CMOS drivers. in *Proceedings of OFC*, 2015
12. P. Pintus, C. Manganelli, F. Gambini, F. Di Pasquale, M. Fournier, O. Lemonnier, C. Kopp, C.J. Oton, Optimization of integrated silicon doped heaters for optical microring resonators. in *Proceedings of ECOC*, 2016
13. A.H. Atabaki, E. Shah Hosseini, A.A. Eftekhar, S. Yegnanarayanan, A. Adibi, Optimization of metallic microheaters for high-speed reconfigurable silicon photonics. Opt. Exp. **18**(17/16) (2010)
14. Ryan Aguinaldo, Alex Forencich, Christopher DeRose, Anthony Lentine, Douglas C. Trotter, Yeshaiahu Fainman, George Porter, George Papen, and Shayan Mookherjea, Wideband silicon-photonic thermo-optic switch in a wavelength division multiplexed ring network, Opt. Exp., vol. 22, N. 7, 8205 (2014)
15. Joris Van Campenhout, William M. J. Green, Solomon Assefa, and Yurii A. Vlasov, Low-power, 2x2 silicon electro-optic switch with 110-nm bandwidth for broadband reconfigurable optical networks, Opt. Exp. Vol. 17, N. 26, 24020–24029 (2009)
16. P. Dong, S. Liao, H. Liang, R. Shafiiha, D. Feng, G. Li, X. Zheng, A. V. Krishnamoorthy, and M. Asghari, Submilliwatt, ultrafast and broadband electro-optic silicon switches, Opt. Express, Vol. 18, N. 24, 25225–25231 (2010)
17. M.S. Nawrocka, T. Liu, X. Wang, R.R. Panepucci, Tunable silicon microring resonator with wide free spectral range. Appl. Phys. Lett. **89**(071110) (2006)
18. Xianshu Luo, Junfeng Song, Shaoqi Feng, Andrew W. Poon, Tsung-Yang Liow, Mingbin Yu, Guo-Qiang Lo and Dim-Lee Kwong, Electro-optically tunable switches with 100GHz flat-top passband and 45dB extinction ratio using silicon high-order coupled-microring resonators for optical interconnects. in *Proceedings of OFC*, 2012
19. An Introduction to MEMS, Loughborough University, in http://www.lboro.ac.uk/microsites/mechman/research/ipm-ktn/pdf/Technology_review/an-introduction-to-mems.pdf
20. N. Farrington, G. Porter, S. Radhakrishnan, H.H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, A. Vahdat, Helios: a hybrid electrical/optical switch architecture for modular data centers. in *SIGCOMM'10*, New Delhi, India, 30 August–3 September 2010
21. L. Peng, Chan-Hyun Youn, Member, Wan Tang, Chunming Qiao, A novel approach to optical switching for intradatacenter networking. IEEE J. Lightw. Technol. **30**(2) (2012)
22. https://www.macom.com/products/product-detail/M21605
23. http://www.macom.com/products/product-detail/MAXP-37161
24. T.J. Seok, N. Quack, S. Han, R.S. Muller, M.C. Wu, Highly scalable digital silicon photonic MEMS switches. IEEE J. Lightw. Technol. **34**(2) (2016)
25. T.J. Seok, N. Quack, S. Han, W. Zhang, R.S. Muller, M.C. Wu, 64x64 low-loss and broadband digital silicon photonic MEMS switches. in *Proceedings of ECOC*, 2015
26. N. Quack, T.J. Seok, S. Han, W. Zhang, R.S. Muller, M.C. Wu, Row/column addressing of scalable silicon photonic MEMS switches. in *International Conference on Optical MEMS and Nanophotonics*, 2015
27. T.J. Seok, N. Quack, S. Han, R.S. Muller, M.C. Wu, Large-scale broadband digital silicon photonic switches with vertical adiabatic couplers. Optica **3**(1) (2016)
28. P. DasMahapatra, R. Stabile, A. Rohit, K.A. Williams, Optical crosspoint matrix using broad-band resonant switches. IEEE J. Select. Topics Quant. Electron **20**(4) (2014)
29. T. Goh, A. Himeno, M. Okuno, H. Takahashi and K. Hattori, High-extinction ratio and low-loss silica-based 8x8 strictly nonblocking thermooptic matrix switch, IEEE JLT, Vol. 17, N. 7, 1999

30. K. Tanizawa, K. Suzuki, M. Toyama, M. Ohtsuka, N. Yokoyama, K. Matsumaro, M. Seki, K. Koshino, T. Sugaya, S. Suda, G. Cong, T. Kimura, K. Ikeda, S. Namiki, H. Kawashima, Ultra-compact 32x32 strictly-non-blocking Si-wire optical switch with fan-out LGA interposer. Opt. Exp. **23**(13), 17599–17606 (2015)
31. K. Tanizawa, K. Suzuki, M. Toyama, M. Ohtsuka, N. Yokoyama, K. Matsumaro, M. Seki, K. Koshino, T. Sugaya, S. Suda, G. Cong, T. Kimura, K. Ikeda, S. Namiki, H. Kawashima, 32×32 strictly non-blocking Si-wire optical switch on ultra-small die of 11×25 mm^2. in *Proceedings of OFC*, 2015
32. Photonic Switches, *Photonic Switching Fabrics and Methods For Data Centers*, Patent WIPO-PCT WO 2016/149797 A1
33. S.J. Ben Yoo, 2D and 3D heterogeneous photonic integrated circuits. in *Proceedings of SPIE*, vol. 8989 89890A-1
34. Y. Huang, J. Song, X. Luo, T.-Y. Liow, G.-Q. Lo, CMOS compatible monolithic multi-layer Si3N4-on-SOI platform for low-loss high performance silicon photonics dense integration. Opt. Express **22**(18), 21859–21865 (2014)
35. L. Chen, Y.-k. Chen, Compact, low-loss and low-power 8×8 broadband silicon optical switch. Opt. Express **20**(17), 18977–18985 (2012)
36. K.-I. Kitayama, Y.-C. Huang, Y. Yoshida, R. Takahashi, T. Segawa, S. Ibrahim, T. Nakahara, Y. Suzaki, M. Hayashitani, Y. Hasegawa, Y. Mizukoshi, A. Hiramatsu, Torus-topology data center network based on optical packet/agile circuit switching with intelligent flow management. J. Lightw. Technol. **33**(5), 1063–1071 (2015)
37. H. Mehrvar, Y. Wang, X. Yang, M. Kiaei, H. Ma, J. Cao, D. Geng, D. Goodwill, E. Bernier, Scalable photonic packet switch test-bed for datacenters. in *Proceedings of OFC*, 2016
38. V. Benes, Permutation groups, complexes, and rearrangeable multistage connecting networks. Bell Syst. Tech. J. **43**, 1619–1640 (1964)
39. C. Chang, R. Melhem, Arbitrary size benes networks. Parallel Proc Lett (7, 3) (1997)
40. A. Pattavina, *Switching Theory Architecture and Performance in Broadband ATM Networks* (Wiley, Hoboken, 1998)
41. L. Lu, S. Zhao, L. Zhou, D. Li, Z. Li, M. Wang, X. Li, J. Chen, 16 × 16 non-blocking silicon optical switch based on electro-optic Mach-Zehnder interferometers. Opt. Express **24**(9) (2016)
42. L. Qiao, W. Tang, T. Chu, Ultra-large-scale silicon optical switches, Group IV Photonics (GFP). in *2016 IEEE 13th International Conference*
43. M. Yuang, P. Tien, H. Chen, W. Ruan, S. Zhong, J. Zhu, Y. Chen, J. Chen, OPMDC: Architecture design and implementation of a new optical pyramid data center network. IEEE/OSA J. Lightw. Technol. **33**(10), 2019–2031 (2015)
44. V. Kamchevska, A.K. Medhin, F. Da Ros, F. Ye, R. Asif, A.M. Fagertun, S. Ruepp, M. Berger, L. Dittmann, T. Morioka, L.K. Oxenløwe, M. Galili, Experimental demonstration of multidimensional switching nodes for all-optical data center networks. J. Lightw. Technol **34**(8), 1837–1843 (2016)
45. K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, X. Wen and Y. Chen, OSA: an optical switching architecture for data center networks with unprecedented flexibility, IEEE/ACM Trans. Netw., Vol. 22, N. 2, 2014
46. M. George, S.P. Saridis, Y. Yan, A. Aguado, B. Guo, M. Arslan, C. Jackson, W. Miao, N. Calabretta, F. Agraz, S. Spadaro, G. Bernini, N. Ciulli, G. Zervas, R. Nejabati, D. Simeonidou, Lightness: A function-Virtualizable software defined data center network with all-optical circuit/packet switching. J. Lightw. Technol. **34**(7), 1618–1627 (2016)
47. F. Testa, C.J. Oton, C. Kopp, J.-M. Lee, R. Ortuno, R. Enne, S. Tondini, G. Chiaretti, A. Bianchi, P. Pintus, M.-S. Kim, D. Fowler, J.'.A. Ayucar, M. Hofbauer, M. Mancinelli, M. Fournier, G.B. Preve, N.Z.C.L. Manganelli, C. Castellan, G. Pares, O. Lemonnier, F. Gambini, P. Labeye, M. Romagnoli, L. Pavesi, H. Zimmermann, F. Di Pasquale, S. Stracca, Design and implementation of an integrated reconfigurable silicon photonics switch matrix in IRIS project. IEEE J. Select. Topics Quant. Electron. **22**(6) (2016)
48. J. Song, Q. Fang, S.H. Tao, M.B. Yu, G.Q. Lo, D.L. Kwong, Passive ring-assisted Mach-Zehnder interleaver on silicon-on-insulator. Opt. Express **16**(12), 8359–8365 (2008)

49. P. Pintus, C. Manganelli, S. Tondini, M. Mancinelli, F. Gambini, C. Castellan, F. Di Pasquale, L. Pavesi, F. Testa, C. J. Oton, Silicon photonics toolkit for integrated switching matrices, Photonic Technologies (Fotonica 2016). in *18th Italian National Conference*, 2016
50. R.A. Budd, L. Schares, B.G. Lee, F.E. Doany, C. Baks, D.M. Kuchta, C.L. Schow, F. Libsch, Semiconductor optical amplifier (SOA) packaging for scalable and gain-integrated silicon photonic switching platforms. in *IEEE 65th Electronic Components & Technology Conference (ECTC)*, 2015
51. K. Tanizawa, K. Suzuki, S. Suda, G. Cong, K. Ishii, J. Kurumida, K. Ikeda, S. Namiki, H. Kawashima, 4x4 Si-wire optical path switch with off-chip polarization diversity. JMoD **42**, OECC (2015)
52. K. Suzuki, K. Tanizawa, S.-H. Kim, S. Suda, G. Cong, S. Namiki, K. Ikeda, H. Kawashima, Polarization Diversity 4x4 Si-Wire Optical Switch, Photonics in Switching (2015)
53. S. Han, T.J. Seok, K. Yu, N. Quack, R.S. Muller, M.C. Wu, 50x50 Polarization-Insensitive Silicon Photonics MEMS Switches: Design and Experiment, Post Deadline Paper, ECOC (2016)
54. N. Dupuis, B.G. Lee, A.V. Rylyakov, D.M. Kuchta, C.W. Baks, J.S. Orcutt, D.M. Gill, W.M.J. Green, C.L. Schow, Modeling and characterization of a nonblocking 4 × 4 Mach–Zehnder silicon photonic switch fabric. J. Lightw. Technol. **33**(20) (2015)
55. B.G. Lee, A.V. Rylyakov, W.M.J. Green, S. Assefa, C.W. Baks, R. Rimolo-Donadio, D.M. Kuchta, M.H. Khater, T. Barwicz, C. Reinholm, E. Kiewra, S.M. Shank, C.L. Schow, Y.A. Vlasov, Monolithic silicon integration of scaled photonic switch fabrics, CMOS logic, and device driver circuits. J. Lightw. Technol. **32**(4) (2014)
56. N. Dupuis, A.V. Rylyakov, C.L. Schow, D.M. Kuchta, C.W. Baks, J.S. Orcutt, D.M. Gill, W.M.J. Green, B.G. Lee, Nano-second-scale Mach-Zehnder-based CMOS photonic switch fabrics. J. Lightw. Technol. **35**(4) (2016)
57. T.J. Seok, V. Kopp, D. Neugrosch, J. Henriksson, J. Luo, M.C. Wu, High density optical packaging of high radix silicon photonic switches. in *Proceedings of OFC 2017*, paper Th5D.7
58. F. Menard, P. Babin, Aeponyx Inc., private communications, 2017
59. D. Celo, D.J. Goodwill, J. Jiang, P. Dumais, C. Zhang, F. Zhao, T. Xin, C. Zhang, S. Yan, J. He, M. Li, W. Liu, Y. Wei, D. Geng, H. Mehrvar, E. Bernier, 32x32 Silicon Photonic Switch, OECC 2016
60. K. Tanizawa, K. Suzuki, K. Ikeda, S. Namiki, H. Kawashima, Non-duplicate polarization-diversity 8 x 8 Si-wire PILOSS switch integrated with polarization splitter-rotators. Opt. Express **25**(10), 10885–10892 (2017)

Chapter 13
Trends in High-Speed Interconnects for Datacenter Networking: Multidimensional Formats and Their Enabling DSP

David V. Plant, Mohamed H. Morsy-Osman, Mathieu Chagnon, and Stephane Lessard

13.1 Introduction

The role of Internet in every facet of nowadays society is indisputable. Growing Internet traffic due to fast-expanding bandwidth intensive applications across all sectors (e.g., social networking, cloud computing and storage, e-commerce, etc.) is creating a spurring capacity demand in DCs. Driven by the persistent growth of cloud-based business and consumer services, the datacenter (DC) traffic is growing incessantly. Its global estimate is forecasted to reach more than 15 zettabytes (1 zetta = 10^{21}) by 2020 representing more than threefold increase from its 2015 value [1]. Nowadays, the majority of Internet traffic, which used to be dominated by peer-to-peer traffic, originates or terminates at a DC. Approximately 77% of traffic originating from a DC is predicted to stay within the DC in 2020 [1]. Also, *hyperscale data centers*, which already account for 34% of total DC traffic, will account for 53% by 2020 [1]. This relentless growth of intra-DC traffic and DC sizes is driving the need for fast inexpensive short-reach optics that provides the desired capacity over growing intra-DC reaches (<10 km). Inside these mega warehouse-sized DCs (see Fig. 13.1), several hundreds of thousands of servers that store and process massive amounts of cloud data reside in large racks (see Fig. 13.2). Typically, servers within each rack are interconnected using a *top-of-rack (ToR) switch*, and several ToRs are connected through *aggregate switches* [2]. According to [3], current generation top-of-rack (ToR) switches have a switching capacity of 3.2 Tbps that is

D.V. Plant (✉) • M.H. Morsy-Osman • M. Chagnon
Department of Electrical & Computer Engineering, McGill University,
Montréal, QC H3A 2A7, Canada
e-mail: david.plant@mcgill.ca; mohamed.osman2@mcgill.ca; mathieu.chagnon@mail.mcgill.ca

S. Lessard
Ericsson Research, 8275 Route Transcanadienne, Saint-Laurent, QC H4S 0B6, Canada
e-mail: stephane.lessard@ericsson.com

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
DOI 10.1007/978-3-319-61052-8_13

Fig. 13.1 Google's large DC facility in Council Bluffs, Iowa

Fig. 13.2 Racks containing hundreds of thousands of servers inside a Google DC

usually provided across 32 ports, each port hosting an optical transceiver that is typically a *Quad Small Form-Factor Pluggable 28* (QSFP28) attached to the switch faceplate. Each QSFP28 provides a net throughput of 100 Gb/s using 4 × 25 Gb/s lanes. Next-generation ToR switch capacities are expected to be 6.4 and 12.8 Tb/s. In order for switches to provide the desired throughput while maintaining a reasonable port density in 1 rack unit (RU) form factor, the number of lanes per transceiver needs to increase by utilizing more wavelengths or fiber lanes in a *wavelength division multiplexing* (WDM) or *space division multiplexing* (SDM) scheme, respectively. However, this is not a scalable solution due to packaging difficulties that exacerbate upgrade to higher capacities using multiplexing approaches. A more promising approach is to employ advanced modulation formats to increase the bit rate per wavelength (per carrier) and hence reduce the required wavelengths or fiber lanes to achieve the target aggregate throughput. Modulating multiple dimensions of a lightwave over numerous levels allows increasing binary throughputs of single-channel short-reach transceivers while maintaining a cost-effective *self-beating* DD scheme. The abovementioned trend can be also observed by the adoption of *pulse amplitude modulation* over four levels (PAM4) as the modulation format to use in the upcoming 400G Ethernet standard over intra-DC reach replacing legacy on-off keying (OOK) that operated over two levels in past standards [4]. Compared to *coherent detection*, direct detection (1) employs a simpler front-end that does not have a local oscillator (LO), (2) requires simpler DSP since there is no need for laser phase noise and frequency offset compensation, and (3) can operate without thermoelectric coolers (TECs), all of which resulting in a receiver that is inexpensive, compact, and less power hungry.

In this chapter, we present a review of the recent literature of multidimensional modulation and direct detection systems that deliver large bit rates (>100 Gb/s) per wavelength over *short reach* for intra-DC interconnects. We begin by reviewing the mathematical representation of a lightwave and polarization rotation in Stokes space. Then, we present the evolution of the multidimensional modulation formats depicting the architectures of their respective transmitters and receivers. Next, digital signal processing (DSP) functions that enable using direct-detection receivers for such formats are presented. Then, the key results that have been recently reported are summarized. Finally, we conclude and give potential future research directions.

13.2 Representation of a Lightwave and Polarization Rotation in Jones and Stokes Spaces

In this section we introduce different notations to represent an optical field in *single mode fiber* (SMF). SMF are cylindrical waveguides that can support the propagation of two orthogonal complex fields. There are two principal ways to vectorially represent the whole optical field in SMF: the Jones and Stokes representations [5]. Both are based on a reference polarization coordinate system. The *Jones space* is a two-dimensional complex space where each dimension represents an orthogonal state of polarization (SOP) of the light and where the argument of each dimension is the complex optical field in said dimension, providing in total 4 degrees of freedom (DOF). These complex two-dimensional Jones vectors are more concisely expressed using "bra-ket notation," where the "ket" is written as $|E\rangle = [E_x, \mathrm{E}_y]^T$ and the "bra" is $\langle E| = |E\rangle^{\dagger} = [E_x^*, E_y^*]$, where * represents complex conjugation, † the conjugate transpose, and $^{\mathrm{T}}$ the transpose. The fields E_x and E_y are the complex fields on the $\hat{x}$

and $\hat{y}$ polarizations, respectively. The same field can be equivalently represented in Stokes space using a three-dimensional (3D) real-valued vector $\boldsymbol{S} = [S_1, S_2, S_3]^{\mathrm{T}}$, where $S_1 = E_x^* E_x - E_y^* E_y$, $S_2 = 2\,\mathrm{Re}\left\{E_x E_y^*\right\}$, and $S_3 = -2Im\left\{E_x E_y^*\right\}$. By definition, Stokes vectors bear the property that $\left(S_1^2 + S_2^2 + S_3^2\right)^{1/2} = E_x^* E_x + E_y^* E_y$, being the total power. Stokes vectors have only 3 degrees of freedom, being the power on $\hat{x}$, the power on $\hat{y}$, and phase difference between the two polarizations E_x and E_y. In order to have a direct access to the total power in the vector representation, the latter can be added to the Stokes vector $\boldsymbol{S}$ as $\boldsymbol{S}' = [S_0, S_1, S_2, S_3]^{\mathrm{T}}$, where S_0 is the total power. The vector $\boldsymbol{S}'$ is called the *Mueller vector* [6]. For direct access to the first two degrees of freedom, it is also common to cast the four-component Mueller vector in the form $\boldsymbol{V} = [\,E_x^* E_x\,,\ E_y^* E_y\,, S_2, S_3]^{\mathrm{T}}$, where the first two components of $\boldsymbol{S}'$ and $\boldsymbol{V}$ are related as $E_x^* E_x = \left(S_0 + S_1\right)/2$ and $E_y^* E_y = \left(S_0 - S_1\right)/2$. From the representation of the first two components, we will call this modified Mueller vector $\boldsymbol{V}$ the *power vector*. We will use this vector representation in the remaining sections of this manuscript as the reference field representation. It is noteworthy to mention that the presence of complex conjugation in all four Stokes components S_0 to S_3 cancels any absolute phase information in Stokes space, hence the loss of a DOF from four in Jones to three in all variants of Stokes representations.

In single-mode fibers being non-polarization-maintaining fibers, the field inside the latter undergoes polarization rotations due to inherent random *birefringence* [5]. The following Eqs. (13.1–13.2) present the impact of polarization rotation in SMF on both Jones and modified Mueller representations. In complex Jones space, a rotation is a complex unitary matrix $\mathbf{U}$ applied to the Jones vector $|E\rangle$ as

$$\left|E_{\mathrm{rot}}\right\rangle = \mathbf{U}\left|E\right\rangle = \begin{bmatrix} a & -b \\ b^* & a^* \end{bmatrix}\left|E\right\rangle \tag{13.1}$$

where unitary matrices have the property that $\mathbf{U}^{\dagger}\mathbf{U} = \mathbf{I}$, the identity matrix. In the power vector representation, the equivalent rotation on $\boldsymbol{V}$ is cast as $\boldsymbol{V}_{\mathbf{rot}} = \mathbf{M}\boldsymbol{V}$ [6] where the rotation matrix $\mathbf{M}$ is given by

$$\mathbf{M} = \begin{bmatrix} |a|^2 & |b|^2 & -\mathrm{Re}\left\{ab^*\right\} & -Im\left\{ab^*\right\} \\ |b|^2 & |a|^2 & \mathrm{Re}\left\{ab^*\right\} & -Im\left\{ab^*\right\} \\ 2\,\mathrm{Re}\left\{ab\right\} & -2\,\mathrm{Re}\left\{ab\right\} & \mathrm{Re}\left\{a^2\right\} - \mathrm{Re}\left\{b^2\right\} & Im\left\{a^2\right\} + Im\left\{b^2\right\} \\ 2Im\left\{ab\right\} & -2Im\left\{ab\right\} & Im\left\{b^2\right\} - Im\left\{a^2\right\} & \mathrm{Re}\left\{a^2\right\} + \mathrm{Re}\left\{b^2\right\} \end{bmatrix} \tag{13.2}$$

Because power vectors *V* are always real valued, 4-by-4 rotation matrix M is also real valued.

13.3 Evolution of Multidimensional Modulation Formats and Their Transceiver Architectures

Figure 13.3 summarizes the evolution of the dimensionality (N_{dim}) of modulation formats and their underlying transmitter and receiver architectures for self-beating direct detect optical communication systems for 1D, 2D, 3D, and 4D modulation, i.e., for $N_{\text{dim}} = \{1, 2, 3, 4\}$. The figure is complemented with example formats and their corresponding field representation in the optical domain.

Increasing the bit rate of single carriers employing 1D modulation formats ($N_{\text{dim}} = 1$) requires either (1) increasing the symbol rate or (2) increasing the modulation complexity imprinted on this dimension. Faster signaling requires electronics of larger bandwidth, while increased modulation complexity improves the spectral efficiency at the expense of requiring a more complex electrical drive signal. The laser drive signal, and therefore the signal obtained after direct detection, can be simple binary signals or more complex waveforms. The most flexible way to generate complex waveforms is to employ a digital-to-analog converter (DAC) of large bit resolution. Several different modulation formats have recently been demonstrated to increase the spectral efficiency, such as discrete multitone modulation (DMT) [7], half-cycle 16-QAM Nyquist-subcarrier-modulation [8], multiband carrier-less amplitude phase modulation (CAP) [9], and multilevel intensity modulation, also known as pulse amplitude modulation (PAM) [10]. All these formats are greatly leveraged by digital signal processing (DSP) from a DAC, while some (e.g., DMT) inherently require a DAC and DSP. Electrical multilevel signaling, on the other hand, can be generated in a simpler way without a DAC by power combining NRZ two-level tributaries of different amplitudes.

Intensity modulation of a laser can be done using direct modulated lasers (DML), electroabsorption-integrated modulated lasers (EML), or via a Mach-Zehnder modulator (MZM) [11]. The receiver for 1D formats is a simple photodiode, converting the total optical power $\langle E|E\rangle$ into photocurrent. For some 1D IM formats employing higher-order modulation like DMT, the receiver also has to comprise an ADC and a DSP block [7], while any format can benefit from digital signal equalization. The "1D" line of Fig. 13.3 depicts the transmitter and receiver architectures for 1D formats, where dashed boxes represent a DAC- or ADC-based transceiver with DSP.

The current technology for optical Ethernet pluggables in datacenters relies on 1D OOK intensity modulation on a single polarization (SP-OOK). Such format provides 1 bit per symbol. To increase the binary throughput of these pluggables, multiple carriers are independently modulated and multiplexed on either different wavelengths in a fiber or on different fibers representing either wavelength division multiplexing (WDM) or space division multiplexing (SDM), respectively. The transceiver requires as many parallel transmitters (RF signals, amplifiers, electro-optic transducers) and parallel receivers (optoelectronic transducers, amplifiers, detectors) as the multiplexing order. Though these multiplexing solutions are currently adopted, they are not scalable toward future larger aggregate bit rates.

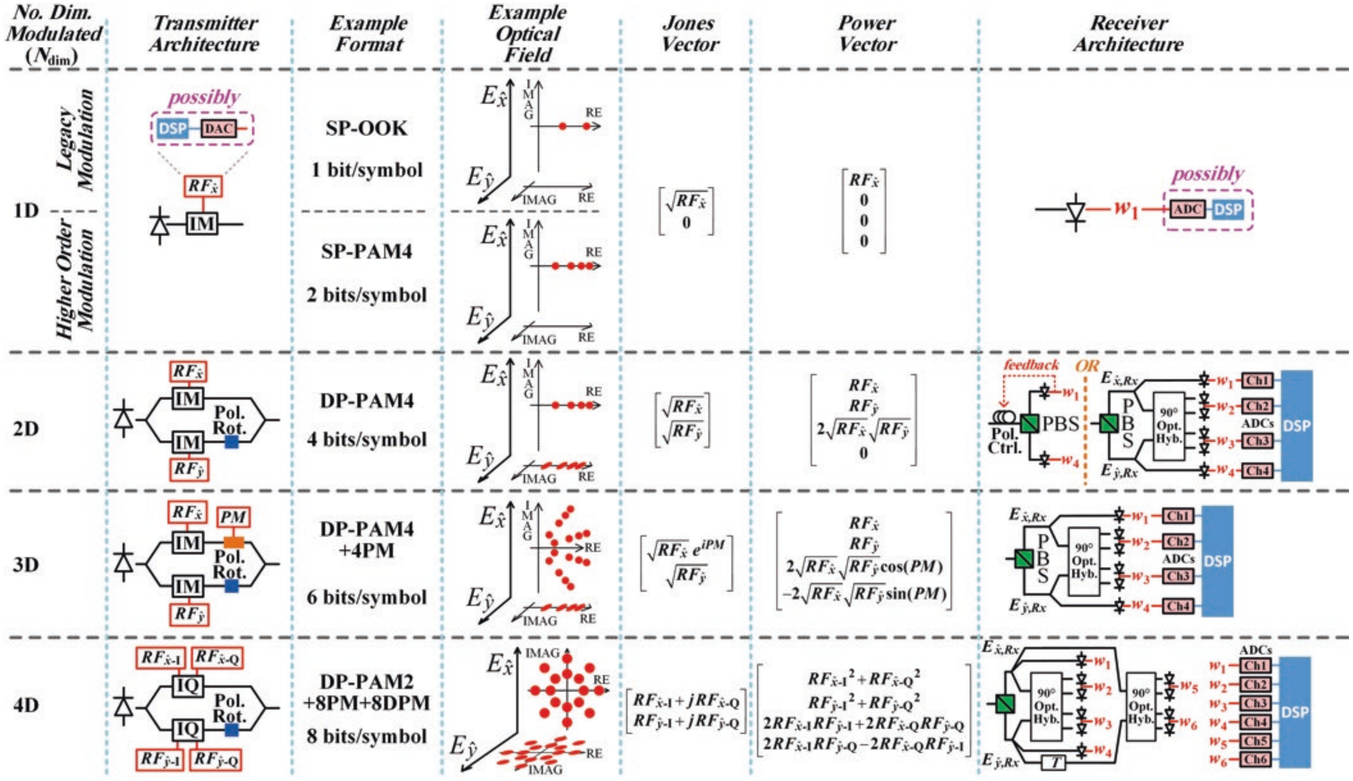

No. Dim. Modulated (N_{dim})
Transmitter Architecture
Example Format
Example Optical Field
Jones Vector
Power Vector
Receiver Architecture
1D
Legacy Modulation
Higher Order Modulation
possibly
DSP
DAC
$RF_{\hat{x}}$
IM
SP-OOK
1 bit/symbol
SP-PAM4
2 bits/symbol
$E_{\hat{x}}$
$E_{\hat{y}}$
IMAG
RE
$\begin{bmatrix} \sqrt{RF_{\hat{x}}} \\ 0 \end{bmatrix}$
$\begin{bmatrix} RF_{\hat{x}} \\ 0 \\ 0 \\ 0 \end{bmatrix}$
w_1
ADC
2D
$RF_{\hat{y}}$
Pol. Rot.
DP-PAM4
4 bits/symbol
$\begin{bmatrix} \sqrt{RF_{\hat{x}}} \\ \sqrt{RF_{\hat{y}}} \end{bmatrix}$
$\begin{bmatrix} RF_{\hat{x}} \\ RF_{\hat{y}} \\ 2\sqrt{RF_{\hat{x}}}\sqrt{RF_{\hat{y}}} \\ 0 \end{bmatrix}$
feedback
OR
Pol. Ctrl.
PBS
$E_{\hat{x},Rx}$
$E_{\hat{y},Rx}$
90° Opt. Hyb.
w_2
w_3
w_4
Ch1
Ch2
Ch3
Ch4
ADCs
3D
PM
DP-PAM4 +4PM
6 bits/symbol
$\begin{bmatrix} \sqrt{RF_{\hat{x}}}\,e^{iPM} \\ \sqrt{RF_{\hat{y}}} \end{bmatrix}$
$\begin{bmatrix} RF_{\hat{x}} \\ RF_{\hat{y}} \\ 2\sqrt{RF_{\hat{x}}}\sqrt{RF_{\hat{y}}}\cos(PM) \\ -2\sqrt{RF_{\hat{x}}}\sqrt{RF_{\hat{y}}}\sin(PM) \end{bmatrix}$
4D
$RF_{\hat{x}\text{-I}}$
$RF_{\hat{x}\text{-Q}}$
$RF_{\hat{y}\text{-I}}$
$RF_{\hat{y}\text{-Q}}$
IQ
DP-PAM2 +8PM+8DPM
8 bits/symbol
$\begin{bmatrix} RF_{\hat{x}\text{-I}}+jRF_{\hat{x}\text{-Q}} \\ RF_{\hat{y}\text{-I}}+jRF_{\hat{y}\text{-Q}} \end{bmatrix}$
$\begin{bmatrix} RF_{\hat{x}\text{-I}}^2+RF_{\hat{x}\text{-Q}}^2 \\ RF_{\hat{y}\text{-I}}^2+RF_{\hat{y}\text{-Q}}^2 \\ 2RF_{\hat{x}\text{-I}}RF_{\hat{y}\text{-I}}+2RF_{\hat{x}\text{-Q}}RF_{\hat{y}\text{-Q}} \\ 2RF_{\hat{x}\text{-I}}RF_{\hat{y}\text{-Q}}-2RF_{\hat{x}\text{-Q}}RF_{\hat{y}\text{-I}} \end{bmatrix}$
T
w_5
w_6
Ch5
Ch6

Fig. 13.3 Evolution of transceiver architectures for 1D, 2D, 3D, and 4D formats for self-beating direct detection systems

Although increasing the modulation order of electrical 1D formats improves the bit rate delivered by improving the spectral efficiency, it comes with an increase of the BER because the SNR remains unchanged [10]. Higher order intensity modulation provides a good example to express this phenomenon. The BER of multilevel formats is function of the SNR and the number of levels M as expressed in Eq. (13.3):

$$\underset{\text{PAM M}}{\text{BER}}\left(M,\text{SNR}\right)=\frac{M-1}{M\log_2\left(M\right)}erfc\left(\sqrt{\frac{3}{2\left(M^2-1\right)}\text{SNR}}\right) \tag{13.3}$$

where *erfc* is the complementary error function.

If a two-level intensity modulated signal providing 1 bit/symbol has a SNR of 16.94 dB, increasing the number of levels to four or eight to double or triple the bit rate worsens the BER from 1×10^{-12} to 6.2×10^{-4} or 3.6×10^{-2}, respectively. Conversely, to maintain the BER at four- and eight-level intensity signaling, the SNR has to increase to 23.88 and 30.07 dB, respectively. In the "1D" line of Fig. 13.3, we depict the symbols in the optical field of both two- and four-level 1D intensity modulation, where both formats have the same mean "DC" and "AC" signal power. By visual inspection, one can observe that the symbol spacings are much closer at four levels, explaining the increase of the BER. Symbols in the optical field are spaced in a square-root fashion such to generate equally spaced levels after square-law direct detection.

Binary throughputs from a single carrier direct detect systems can be further increased by modulating more dimensions of a lightwave, i.e., $N_{\text{dim}} > 1$. In the following, with visual inspections from Fig. 13.3, we present transceiver architectures and corresponding optical fields of 2D, 3D, and 4D format compliant to self-beating direct detection.

One variant of 2D formats is to modulate independently the intensity on each orthogonal polarization. Such format is called *dual-polarization intensity modulation* (DP-IM) and allows to double the bitrate of 1D IM format while maintaining the signaling rate and without affecting the SNR per orthogonal polarization [12]. The transmitter architecture for such 2D format requires two driving signals and intensity modulators, branched off of the same laser source and combined on orthogonal polarizations, thanks to an inline polarization rotator (Pol. Rot.), as depicted in the "2D" line of Fig. 13.3. Other variants of 2D transmitters employ complex IQ modulation on one polarization while sending a copy of the CW transmit laser on the orthogonal polarization [13–15]. For these 2D formats, two different types of receiver can recover the two intensities which are differentiated by the method employed to cancel polarization rotation: optically or digitally.

The first receiver performs polarization derotation in the optical domain and consists of an active polarization controller followed by a polarization beam splitter (PBS) serving as a polarization demultiplexer, where the two outputs are subsequently detected by a photodiode. The two resulting RF photocurrents can be input to ADCs for subsequent DSP for IM formats having complex modulation schemes, like DMT, or where digital filtering is applied to enhance BER performance, like in the case of 1D formats. Other simpler IM formats relying on multilevel modulation can be detected without DSP, although digital filtering tends to improve the BER performance. As polarization slowly wanders inside single-mode fiber (see Eqs. (13.1) and (13.2)), it needs to be tracked such to maintain the two optical outputs of the PBS using a control signal from one of the two photocurrents. The control signal minimizes the inter-polarization crosstalk on said RF output. Such receiver architecture is depicted at the end of the "2D" line of Fig. 13.3.

The other type of receiver for 2D formats is called a *Stokes vector receiver* (SVR) and uses digital signal processing to perform *polarization derotation* [12]. The receiver architecture, also depicted at the end of the "2D" line of Fig. 13.3, is comprised of a PBS where each output is further split in two using couplers. The couplers' outputs from orthogonal polarizations are combined together on the same polarization through a 90° optical hybrid, and the four hybrid outputs are detected in pairs using balanced photodetectors. The other couplers' outputs of each polarization are directed to photodetectors. The four photocurrents generated are real time sampled using four analog-to-digital converters (ADC) and are input to subsequent DSP. The four waveforms w_1 to w_4 are $\left|E_{\hat{x},Rx}\right|^2$, $2\,\mathrm{Re}\left\{E_{\hat{x},Rx}\,E^*_{\hat{y},Rx}\right\}$, $2Im\left\{E_{\hat{x},Rx}\,E^*_{\hat{y},Rx}\right\}$, and $\left|E_{\hat{y},Rx}\right|^2$, respectively, and represent the four components of the rotated modified Mueller power vector introduced in Sect. 2. The DSP algorithm employed to recover the two intensity tributaries is that for 2D formats presented in Sect. 4 and performs polarization derotation and allows for mitigation of *intersymbol interference* (ISI). It is noteworthy to mention that the active optical polarization tracking of the first 2D receiver is now replaced by active digital polarization tracking and the optical receiver front-end of SVR is a passive structure serving to beat the incoming signal with itself.

The "3D" line of Fig. 13.3 presents the transmitter architecture allowing to modulate a third orthogonal degree of freedom while maintaining a direct detection scheme, therefore permitting to triple the bit rate of 1D formats, while the signaling rate stays the same [16]. The transmitter architecture is similar to that of 2D formats, with an additional phase modulator in one of the two IM branches. This phase modulator allows modulating the *inter-polarization phase*: the phase of one polarization with respect to the other orthogonal polarization. When all three driving signals are of four-level modulation, the optical signal carries 6 bits per symbol. The optical field of such format, called dual polarization PAM4 with four inter-polarization phases (DP-PAM4 + 4 PM), is also depicted in Fig. 13.3. The same SVR receiver for 2D formats can be used to recover all three degrees of freedom via the received modified Mueller power vector using the DSP presented in Sect. 4 for "3D" formats, where now all four components of $V_{4\mathrm{D}}$ need to be derotated and recovered digitally.

Finally, it is possible to modulate a fourth dimension of the optical field and recover the latter in a direct detect fashion [17]. Similar to differential phase shift keying (DPSK) modulation, the phase jumps from one Jones vector symbol to the next can be modulated and demodulated using self-beating. Unlike DPSK, there are three possible intersymbol phase jumps that can be modulated: phase jumps from $E_{\hat{x}}\left[\mathrm{n}\right]$ to $E_{\hat{x}}\left[\mathrm{n}+1\right]$, from $E_{\hat{y}}\left[\mathrm{n}\right]$ to $E_{\hat{y}}\left[\mathrm{n}+1\right]$, or from $E_{\hat{x}}\left[\mathrm{n}\right]$ to $E_{\hat{y}}\left[\mathrm{n}+1\right]$, where n is the discrete time index at time $t = \mathrm{n}T$ and where T is the symbol duration. The transmitter architecture to generate the format differs from that of the previous 2D and 3D formats. Instead of employing intensity and phase modulators, the transmitter uses in-phase-quad-phase (IQ) modulators, also known as dual-parallel Mach-Zehnder modulator (MZM), where each MZM is biased at null. The transmitter comprises two IQs modulating independently, in a complex fashion, the same laser source where the polarization of one IQ output is rotated such to combine the two tributaries on orthogonal polarizations. This transmitter architecture is the same as that for polarization-multiplexed coherent transmission, serving as a full four-dimensional field modulator, as depicted in “4D” line of Fig. 13.3. Only specific complex modulation formats relying on multi-ring/multiphase can be employed in order to be recoverable by a self-beating receiver [17]. To demodulate the intersymbol, inter-polarization phase jumps in a self-beating fashion, the SVR used for 2D and 3D formats has to be modified as depicted in the receiver architecture column of the “4D” line of Fig. 13.3, where the received field $E_{\hat{x},Rx}$ beats with a one symbol delayed version of $E_{\hat{\mathrm{y}},Rx}$ through a second 90° optical hybrid. Waveforms w_1 to w_4 are the same as that of the SVR for 2D and 3D formats, and w_5 and w_6 are $2\,\mathrm{Re}\left\{E_{\hat{x},Rx}\left[n\right]E_{\hat{y},Rx}\left[n-1\right]\right\}$ and $-2Im\left\{E_{\hat{x},Rx}\left[n\right]E_{\hat{y},Rx}\left[n-1\right]\right\}$, respectively.

13.4 Enabling DSP for Multidimensional Formats

All modulation formats outlined in the previous section benefit from applying DSP at the transmitter and receiver sides; some of these DSP tasks are mandatory since they, for example, enable polarization demultiplexing in a multidimensional format ($N_{\mathrm{dim}} > 1$) or compensate for a significant amount of ISI due to limited bandwidth of electronics. Figure 13.4a shows a high-level block diagram of the DSP stack of possible functions carried out at the transmitter for a N_{dim}-D format. We assume each dimension of the lightwave carries one symbol that is drawn from a symbol alphabet with size K. This leads to having $N_{\mathrm{dim}}\log_2 K$ bits that get modulated per symbol across all dimensions of the lightwave. Subsequent DSP blocks at the transmitter include: (1) resampling to the sampling rate of the digital-to-analog converter (DAC) if used, (2) pulse shaping using a band-limiting spectral shaping filter, (3) compensation of the nonlinear (NL) transfer characteristics of a Mach-Zehnder modulator (MZM) if used, (4) pre-compensation of ISI using a finite impulse response (FIR) filter, and (5) clipping and quantizing the resulting samples in preparation of feeding them to a DAC having a finite number of bits to represent each quantized level. The chain of the abovementioned DSP functions is applied to all N_{dim} signals that will modulate each of the N_{dim} dimensions of the lightwave. However, some the above functions may not be used for all dimensions. For example, the driving signal that gets imprinted onto the inter-polarization phase does not require any compensation of the NL characteristics of a MZM since it drives a linear phase modulator (PM), not a MZM (refer to the 3D transmitter architecture in Fig. 13.3). Next, Fig. 13.4b shows a high-level block diagram of the DSP stack of possible functions carried out at the receiver for a N_{dim}-D format. Depending on N_{dim}, we assume the input signal to the receiver DSP chain comprises N_{ch} waveforms provided by the ADCs that sample each of the output photocurrents from one the corresponding receiver front-ends in Fig. 13.3. After resampling from the sampling rate of the ADC to twice the symbol rate, the key DSP block at the receiver is a *multiple-input*, *multiple-output* (MIMO) bank of FIR filters each having a temporal length N_{taps}. The task of the bank of FIR filters is twofold: (1) it achieves polarization demultiplexing by inverting the polarization rotation that occurs along the fiber (refer to Eqs. 13.1 and 13.2 in Sect. 2) which results in a misalignment between the transmitter's and receiver's principle axes, denoted by $(\hat{x}, \hat{y})_{Tx}$ and $(\hat{x}, \hat{y})_{Rx}$, and (2) it post-compensates any residual ISI that may have resulted from bandwidth-limited electronic components in the system (e.g., electric amplifiers, ADCs, DACs, etc.). The final two blocks in the receiver DSP chain include clock recovery and hard decision of the received symbols on each of the N_{dim} dimensions. Clearly, the key enabling DSP block in the chains of Fig. 13.4 is the MIMO filtering. This block accepts N_{ch} incoming signals from the receiver front-end and provides N_{dim} signals ready for the retiming and final decision. Henceforth, we explain in detail the contents of the MIMO filtering block for each of the multidimensional formats, i.e., for $N_{\text{dim}} = \{2, 3, 4\}$.

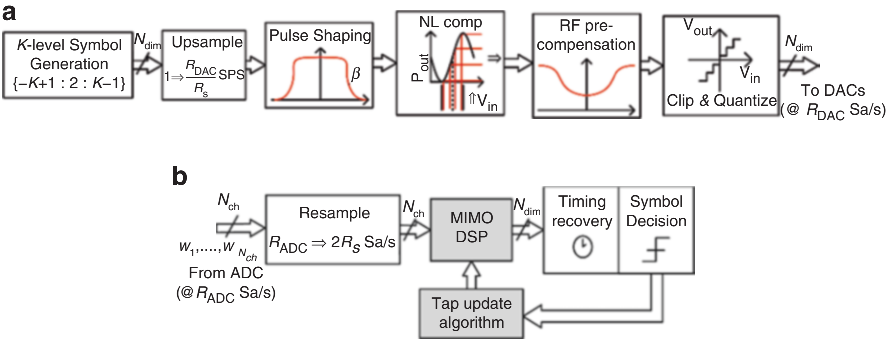


Fig. 13.4 High-level block diagrams of possible DSP functions at (**a**) transmitter and (**b**) receiver, for a N_{dim}-D modulation format. The transmitter DSP stack generates N_{dim} signal streams that drive N_{dim} DAC channels each running at R_{DAC} Sa/s. The receiver DSP chain accepts N_{ch} waveforms from N_{ch} ADC channels each running at R_{ADC} Sa/s from the Stokes receiver front-end depicted earlier in Fig. 13.3

Since 1D formats are detected via a single photodetector which is agnostic to the received state of polarization (SOP), the DSP at the receiver side (if used at all) is useful to post-compensate any residual ISI. Hence the MIMO block in Fig. 13.4b reduces to a single-input, single-output (SISO) FIR filter that post-compensates ISI [18]. On the other hand, 2D, 3D, and 4D formats are severely impacted by the cross-talk resulting from the misalignment of the transmitter's and receiver's principle axes due to random polarization rotation. Hence, MIMO DSP is mandatory to undo the polarization rotation and recover the transmitted data. Figure 13.5a shows a block diagram of the MIMO DSP required for the 2D DP-IM format as was detailed

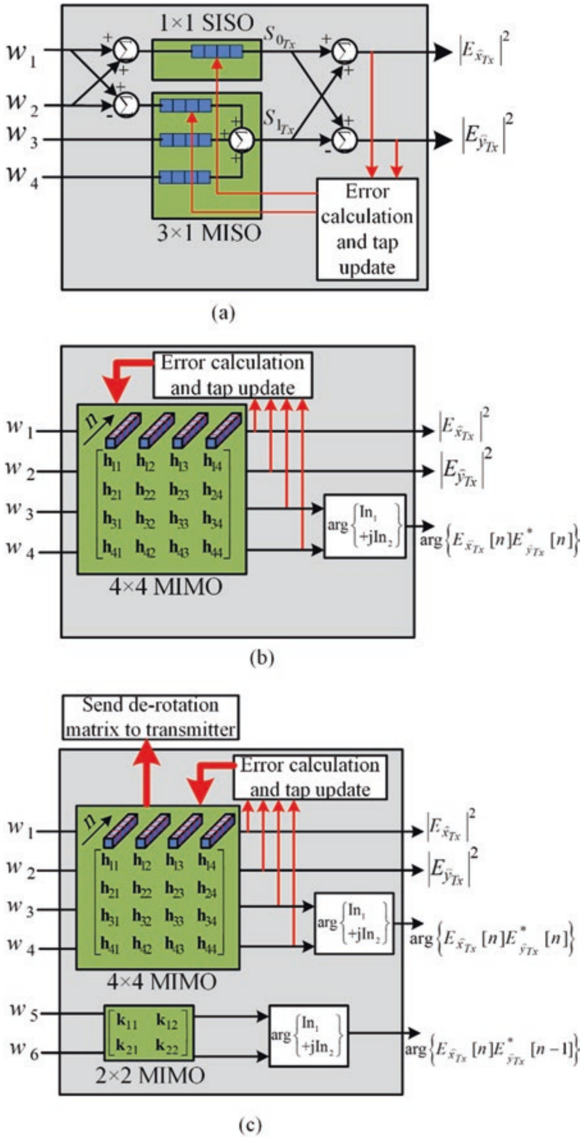

1×1 SISO
3×1 MISO
Error calculation and tap update
w_1
w_2
w_3
w_4
$S_{0_{Tx}}$
$S_{1_{Tx}}$
$|E_{\hat{x}_{Tx}}|^2$
$|E_{\hat{y}_{Tx}}|^2$
(a)
Error calculation and tap update
4×4 MIMO
$\mathbf{h}_{11}$ $\mathbf{h}_{12}$ $\mathbf{h}_{13}$ $\mathbf{h}_{14}$
$\mathbf{h}_{21}$ $\mathbf{h}_{22}$ $\mathbf{h}_{23}$ $\mathbf{h}_{24}$
$\mathbf{h}_{31}$ $\mathbf{h}_{32}$ $\mathbf{h}_{33}$ $\mathbf{h}_{34}$
$\mathbf{h}_{41}$ $\mathbf{h}_{42}$ $\mathbf{h}_{43}$ $\mathbf{h}_{44}$
$\arg\{\mathrm{In}_1 + j\mathrm{In}_2\}$
$\arg\{E_{\hat{x}_{Tx}}[n]E^*_{\hat{y}_{Tx}}[n]\}$
(b)
Send de-rotation matrix to transmitter
Error calculation and tap update
4×4 MIMO
w_5
w_6
$\mathbf{k}_{11}$ $\mathbf{k}_{12}$
$\mathbf{k}_{21}$ $\mathbf{k}_{22}$
2×2 MIMO
$\arg\{E_{\hat{x}_{Tx}}[n]E^*_{\hat{y}_{Tx}}[n-1]\}$
(c)

Fig. 13.5 Structural block diagrams of MIMO DSP that follows the Stokes receiver front-end for (**a**) 2D DP-IM, (**b**) 3D DP-IM-PM, and (**c**) 4D DP-IM-PM-DPM modulation formats

in [19]. The block comprises a combination of one SISO FIR filter and a 3 × 1 multiple-input, single-output (MISO) bank of FIR filters that accepts four waveforms w_1 to w_4 from the front-end in the 2D line of Fig. 13.3 giving the transmitted intensities $\left|E_{\hat{x},Tx}\right|^2$ and $\left|E_{\hat{y},Tx}\right|^2$. In such case, the 3 × 1 MISO block (a) provides the inversion of the polarization rotation, and the temporal length of all the filters is to mitigate the ISI. In case of 3D DP-IM-PM format, the MIMO block required becomes the one in Fig. 13.5b which accepts the same four waveforms w_1 to w_4 from the respective front-end in Fig. 13.3 and recovers $\left|E_{\hat{x},Tx}\right|^2$ and $\left|E_{\hat{y},Tx}\right|^2$ as well as the inter-polarization phase $\arg\left\{E_{\hat{x},Tx}\,E^*_{\hat{y},Tx}\right\}$ [6] − arg $\{S_{2\mathrm{Tx}}+\boldsymbol{i}S_{3\mathrm{Tx}}\}$. In such case, the 4 × 4 MIMO is adapted so that it inverts the polarization rotation besides the temporal length of each FIR filter to remove residual ISI. Finally, the 4D DP-IM-PM-DPM format in [17] requires processing the two additional waveforms w_5 and w_6 from the front-end in the last line of Fig. 13.3. For such 4D format, the MIMO block becomes the one in Fig. 13.5c where the first four waveforms w_1 to w_4 are processed in exactly the same manner as in Fig. 13.5b for the 3D format. As clearly explained in [20], recovering the fourth dimension from w_5 and w_6 is only possible if $\left(\hat{x},\hat{y}\right)_{Tx}$ and $\left(\hat{x},\hat{y}\right)_{Rx}$ are aligned. In order to align the transmitter's and receiver's axes, we make use of the knowledge of the derotation matrix that is obtained from 4 × 4 MIMO operating on the first three dimensions. This matrix is sent back to the transmitter where the DP-IQ modulator can pre-rotate the transmitted waveforms, ensuring aligned axes at the receiver. In such case, w_5 and w_6 can be processed separately by a 2 × 2 MIMO block that recovers the fourth dimension: the inter-polarization differential phase $\arg\left\{E_{\hat{x},Tx}\left[\mathrm{n}\right]E^*_{\hat{y},Tx}\left[\mathrm{n}-1\right]\right\}$.

13.5 State-of-the-Art Experimental Results Using Transceivers Realized by Discrete Components

This section summarizes the key results achieved using the modulation formats whose details and transceiver architectures were presented in Sect. 3 and are enabled using the DSP presented in Sect. 4. Table 13.1 presents the summary of the key results achieved using 1D, 2D, 3D, and 4D formats; each is listed in a separate row. Specifics of the experimental testbeds used for each experiment can be found in the publications cited in the table. All these record-breaking results were generated at the labs of the Photonics Systems Group (PSG) at McGill University. Figure 13.6 shows photos of the testbed used to demonstrate the 2D DP-PAM4 format that generated the results in [19]. Key modulation parameters for each experiment are listed including the symbol rate (or baud rate), the resulting number of bits per symbol, the maximum transmission distance or reach, and the operating wavelength.

Table 13.1 Summary of key achievements using multidimensional formats relying on direct-detection receivers with polarization demultiplexing using DSP (HD-FEC: Hard decision forward error correction)

Number of dimensioNumber of dimensions	Modulation format	Number of bits per symbol	Baud rate (Gsym/s)	Maximum bit rate (Gb/s) achieved at HD-FEC	Transmission distance (km)	Operating wavelength (nm)
1D [18]	PAM4 or PAM8	2 or 3	56 or 37.4	112	10	1310
2D [19]	DP-PAM4	4	56	224	10	1310
3D [6]	DP-PAM4-4 PM	6	50	300	b2b	1550
4D [20]	DP-PAM2-8 PM-8DPM	8	40	320	10	1550

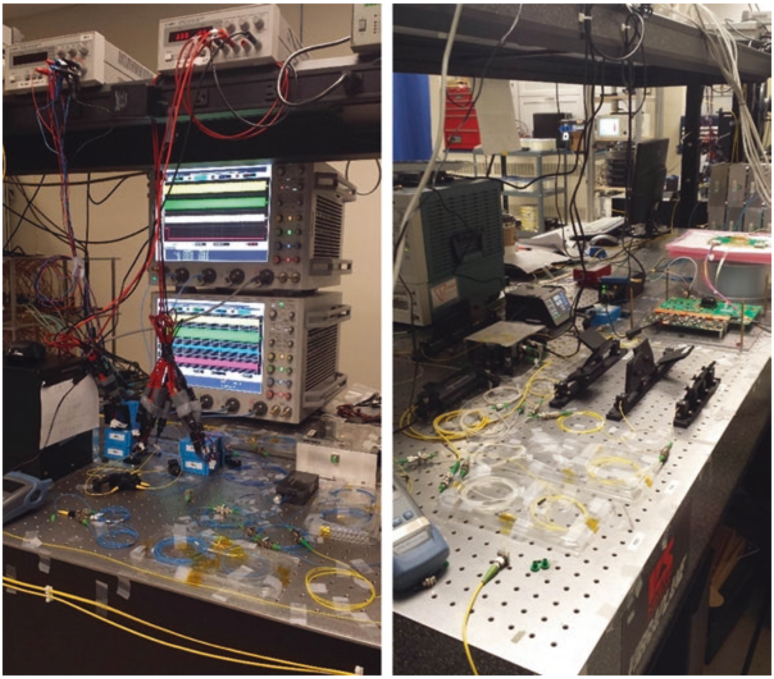


Fig. 13.6 Photos captured at the Photonics Systems Group Lab at McGill University for the experimental setup to demonstrate the 2D DP-PAM4 format in [19]. *Left* photo is the Stokes vector receiver front-end realized using discrete components followed by real-time scopes, and *right* photo is the DP-PAM4 transmitter

Also, Table 13.1 lists the maximum bit rate we were able to achieve using each format, while the bit error rate (BER) stays below the hard decision forward error correction (HD-FEC) threshold of 3.8×10^{-3}. If the BER stays below HD-FEC, the FEC decoder that follows the receiver DSP chain in Fig. 13.4b is guaranteed to produce error-free decoded bits with post-FEC BER in the range of 10^{-12} to 10^{-15} depending on the type of FEC used.

Examining Table 13.1 carefully, we notice that going from one row to the next, the number of modulated dimensions increases meaning more bits per symbol are modulated; hence, the resulting capacity (bit rate) per carrier goes up. Also, we notice that the first two experiments in [18, 19] for 1D and 2D formats were performed in the O band (1310 nm) near the zero dispersion wavelength of standard single-mode fiber (SMF), and hence the only factor that limited the maximum transmission distance is the received power budget. For the following experiment in [6] for the 3D format which was done in the C band (1550 nm) due to availability of parts, the interplay between dispersion and phase modulation limited the transmission distance severely, and we were only able to demonstrate the feasibility of the format in back to back. However, the same experiment could be repeated in the O band where the interplay between dispersion and phase modulation is minimized and reach could be extended. The last experiment in [20] for the 4D format was also performed in the C band, but dispersion pre-compensation using DSP was used at the transmitter prior to driving the DP-IQ modulator in the last line of Fig. 13.3. One final comment on Table 13.1, we should notice that the increase in bit rate from 300 to 320 Gb/s by going from 3D to 4D modulation format seems marginal; however, this increase was realized at a reduced symbol rate (40 instead of 50 Gsym/s) due to the more spectrally efficient 4D format. This means all the electronic, the electro-optic, and the optoelectronic components required can have smaller bandwidths since they will run at a slower signaling rate.

Finally, we should also note that the results in Table 13.1 were achieved using transceivers realized from discrete optical components (e.g., notice the fiber-based couplers, polarization controllers, and optical delay lines in Fig. 13.6). However, all these transceivers depicted in Fig. 13.3 can be potentially integrated on photonic ICs. Refs. [21–23] give examples from the literature of Stokes transmitters and receivers demonstrated experimentally as silicon photonic integrated circuits.

13.6 Conclusion and Future Research Avenues

In this chapter, we presented a review of multidimensional modulation formats which can be direct-detected without a local oscillator laser. These formats are capable of modulation up to four dimensions of a single laser being the intensity on orthogonal polarization states, the inter-polarization phase, and the inter-polarization differential phase. Due to their ability to deliver much larger bit rates at similar symbol rates compared to legacy one-dimensional formats such as OOK, these multidimensional formats are candidates to be deployed in next-generation intra-DC optical interconnects. Using these formats, the number of wavelengths or fiber lanes required to be multiplexed to achieve a target aggregate capacity can be greatly reduced by pushing the per-channel bit rate. The transceiver architectures are depicted together with the enabling digital signal processing required to invert the polarization rotation along the fiber and detect all dimensions carrying modulated data. Experimental results obtained using benchtop discrete components to realize the transceivers as well as offline DSP were also highlighted.

Moving forward, potential research avenues include exploring alternative implementations of multidimensional transceivers that are less complex than the ones already presented. In particular, improvements in receiver sensitivity are highly desirable which can be achieved by using different optical front-ends. Also, fully integrated transceivers should be realized which requires, in addition to PIC fabrication, extensive optical and electrical packaging endeavors. Finally, a more in-depth study of the power consumption of the transceivers including the DACs, ADCs, drivers, and real-time DSPs implemented as application-specific ICs (ASICs) should be performed. Such power consumption study will solidify the likelihood of implementing multidimensional formats in future optical interconnects over intra-DC reaches (<10 km).

References

1. Cisco Corporation, Cisco global cloud index: forecast and methodology, 2015–2020 (2016). Cisco white paper available at http://www.cisco.com/c/dam/en/us/solutions/collateral/ service-provider/global-cloud-index-gci/white-paper-c11–738085.pdf
2. E. Agrell, M. Karlsson, A. Chraplyvy, et al., Roadmap of optical communications. J. Opt. **18**, 063002 (2016)
3. B. Booth, in *Lighting up the cloud. presented at photonic integrated circuits (PIC) international conference*, Brussels, Belgium, 1–2 March 2016
4. IEEE, P802.3bs 200 Gb/s and 400 Gb/s Ethernet Task Force [Online]. Available at http://www.ieee802.org/3/bs/index.html
5. G. Agrawal, *Lightwave Technology Telecommunication Systems* (Wiley-Interscience, Hoboken, 2005)
6. M. Chagnon, M. Morsy-Osman, D. Patel, et al., Digital signal processing for dual-polarization intensity and inter-polarization phase modulation formats using stokes detection. J. Lightw. Technol. **34**, 188–195 (2016)
7. Y. Kai, M. Nishihara, T. Tanaka et al., Experimental comparison of pulse amplitude modulation (PAM) and discrete multi-tone (DMT) for short-reach 400-Gbps data communication. Paper presented at 39th European Conference on Optical Communication (ECOC), London, UK, 22–26 September 2013
8. A. Karar, J. Cartledge, Generation and detection of a 112-Gb/s dual-polarization signal using a directly modulated laser and half-cycle 16-QAM Nyquist-subcarrier-modulation. Paper presented at Post-deadline session of 38th European Conference on Optical Communication (ECOC), Amsterdam, Netherlands, 16–20 September 2012
9. M. Olmedo, T. Zuo, J. Jensen, et al., Multiband carrierless amplitude phase modulation for high capacity optical data links. J. Lightw. Technol. **32**, 798–804 (2014)
10. M. Chagnon, M. Morsy-Osman, M. Poulin, et al., Experimental study of 112 Gb/s short reach transmission employing PAM formats and SiP intensity modulator at 1.3 μm. Opt. Express **22**, 21018–21036 (2014)
11. P. Winzer, R. Essiambre, Advanced optical modulation formats. Proc. IEEE **94**, 952–985 (2006)
12. M. Morsy-Osman, M. Chagnon, M. Poulin et al., 1λ × 224 Gb/s 10 km transmission of polarization division multiplexed PAM-4 signals using 1.3 μm SiP intensity modulator and a direct-detection MIMO-based receiver. Paper presented at Post-deadline session of 40th European Conference on Optical Communication (ECOC), Cannes, France, 21–25 September 2014
13. D. Che, A. Li, X. Chen et al., 160-Gb/s stokes vector direct detection for short reach optical communication. Paper presented at Post-deadline session of 37th Optical Fiber Communication (OFC) Conference, San Francisco, California, 9–14 March 2014
14. D. Che, A. Li, X. Chen, et al., Stokes vector direct detection for linear complex optical channels. J. Lightw. Technol. **33**, 678–684 (2015)
15. M. Sowailem, T. Hoang, M. Chagnon, et al., 100G and 200G single carrier transmission over 2880 and 320 km using an InP IQ modulator and stokes vector receiver. Opt. Express **24**, 30485–30493 (2016)
16. M. Chagnon, M. Osman, D. Patel et al., 1 λ, 6 bits/symbol, 280 and 350 Gb/s direct detection transceiver using intensity modulation, polarization multiplexing, and inter-polarization phase modulation. Paper presented at Post-deadline session of 38th Optical Fiber Communication (OFC) Conference, Los Angeles, California, 22–26 March 2015
17. M. Morsy-Osman, M. Chagnon, D. Plant, Polarization division multiplexed intensity, inter polarization phase and inter polarization differential phase modulation with stokes space direct detection for 1λ×320 Gb/s 10 km transmission at 8 bits/symbol. Paper presented at Post-deadline session of 41st European Conference on Optical Communication (ECOC), Valencia, Spain, 27 September – 1 October 2015

18. M. Chagnon, M. Morsy-Osman, M. Poulin, et al., Experimental parametric study of a silicon photonic modulator enabled 112-Gb/s PAM transmission system with a DAC and ADC. J. Lightw. Technol. **33**, 1380–1387 (2015)
19. M. Morsy-Osman, M. Chagnon, M. Poulin, et al., 224 Gb/s 10 km transmission of PDM PAM-4 at 1.3 μm using a single intensity-modulated laser and a direct-detection MIMO DSP-based receiver. J. Lightw. Technol. **33**, 1417–1424 (2015)
20. M. Morsy-Osman, M. Chagnon, D. Plant, Four dimensional modulation and stokes direct detection of polarization division multiplexed intensities, inter polarization phase and inter polarization differential phase. J. Lightw. Technol. **34**, 1585–1592 (2016)
21. P. Dong, X. Chen, K. Kwangwoong, et al., 128-Gb/s 100-km transmission with direct detection using silicon photonic Stokes vector receiver and I/Q modulator. Opt. Express **24**, 14208–14214 (2016)
22. P. Dong, X. Chen, Stokes vector communications using silicon photonic integrated circuits. Presented at OSA Asia Communications and Photonics Conference (ACP), Wuhan, China, 2–5 November 2016
23. E. Elfiky, M. Sowailem, A. Samani et al., Dual polarization O-Band silicon photonic intensity modulator for Stokes vector direct detection systems," to appear at 40th Optical Fiber Communication (OFC) Conference, Los Angeles, California, 19–23 March 2017

Chapter 14
Trends in High Speed Interconnects: InP Monolithic Integration

Kevin Williams and Boudewijn Docter

14.1 Introduction

Performance has proved to be the primary driver for integrated circuit technology. Wafer scale integration is instrumental in removing assembly complexity and variability and enables sustained increases in the functionality, performance and circuit reliability while at the same time reducing size, power and cost [1, 2]. PIC-enabled products now outperform equivalent combinations of discrete components at the functional level for telecom systems. Digital coherent transceivers are a striking example of circuits which could not be produced cost-effectively in any other way [3]. Wavelength-multiplexed transceivers [4] and potentially optical packet switching circuits [5] may be expected to follow.

Yield has been improved through sustained technology development for InP-based epitaxial growth, fabrication processes and device design innovations. Killer defect densities have been driven down to levels comparable with silicon CMOS in the early 1990s [4] with Infinera reporting random killer defect densities in the range 0.5–1.25 cm^{-2} and functional yields as high as 70% for 440 element circuits [2]. Improved reliability becomes a key advantage for InP-integrated photonics technology as the performance-yield envelope is dominated by packaging and assembly. Strict design methodologies and tightly controlled, standardized processes have been essential. Vertically integrated corporations such as Infinera [4] and open access platforms such as JePPIX [6] have adopted methodologies for InP-based PICs which exhibit similarities to the CMOS electronic IC approach.

K. Williams (✉)
Institute for Photonic Integration, (COBRA), Technical University Eindhoven,
Flux Building 9.067, 5612 AJ Eindhoven, The Netherlands
e-mail: k.a.williams@tue.nl

B. Docter
EFFECT Photonics BV, Torenallee 20, 5617 BC Eindhoven, The Netherlands
e-mail: boudewijndocter@effectphotonics.nl

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
DOI 10.1007/978-3-319-61052-8_14

Scaling laws for InP-integrated photonics are similar to other thin film fabrication technologies like CMOS electronics and silicon photonics. Cost is defined by the market volumes, so an increased wafer size and fab throughput leads to a decreasing price-per-square-millimetre price [7]. Current volumes of photonic product are addressed with wafer sizes of 3″ and 4″, but as volumes increase, the cost advantages of moving to larger wafer sizes, increasing wafer batch sizes and levels of automation will appear. Integration becomes increasingly important as the volumes scale. Much as in electronics, costs will be dominated by test and assembly rather than wafer production providing the driver for further integration.

In this chapter, we review the InP integration techniques most relevant to current and future data interconnection. The InP building blocks are first reviewed, before integration methods, and compared. State-of-the-art devices and circuits are overviewed for direct detection and coherent communications. Components for transparent optical networking are addressed before considering trends in technology integration and wafer-scale assembly.

14.2 The Building Blocks

The rich diversity of InP components can be concisely described in terms of basic building blocks [6] which may be visualized through waveguide cross-sections. Figure 14.1 shows examples of basic building blocks from an open access InP-integrated photonics platform. The five different types of waveguide cross-sections are shown from left to right: (i) optical amplification and absorption, (ii) electro-optic phase modulation, (iii) shallow-ridge optical waveguiding, (iv) deep-ridge optical waveguiding and (v) polarization rotation. Enhanced performance is enabled through adjustments to the precise cross-section and process flow used. The power of monolithic integration is to create these building blocks in the same plane, on the same wafer and in the same process flow, making optical losses and reflections between components negligible.

Amplifiers are implemented through the simultaneous confinement of optical field and charge carriers within a separate confinement heterostructure. Figure 14.1 shows confinement of the optical field (white) in the SCH region (green) and with an additional active layer (red). Here a shallow-ridge waveguide is shown, but deep-ridge and buried heterostructures are also feasible. The function is implemented in a p-i-n diode structure with electronic contacting layers optimized for low electrical resistance and low optical losses. Operating in forward bias enables amplification, while reverse bias enables detector, electro-absorption modulators or a saturable absorber functions. The precise active layer structure can be optimized depending on the precise performance requirements and function. Both multi-quantum well (MQW) and bulk quaternary layer stacks consisting of either InGaAsP or InAlGaAs materials are possible, where the InAlGaAs materials generally provide better high-temperature operation properties.

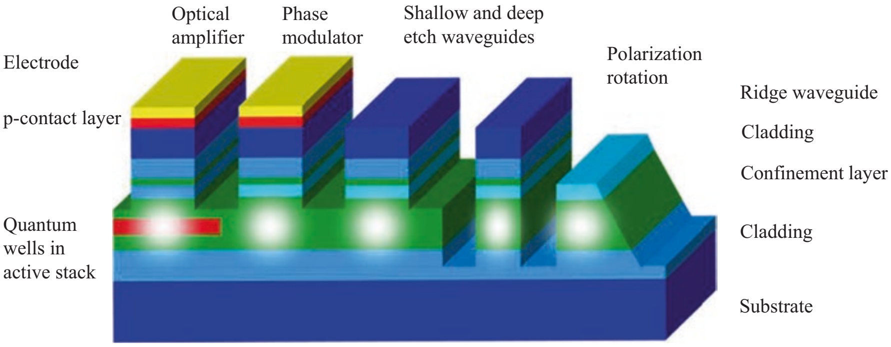


Fig. 14.1 Basic building blocks in InP photonic integration. From *left to right*, an amplifier or absorber, phase modulator, shallow-etch waveguide, deep-etch waveguides and polarization rotator

Phase modulators also use a p-i-n diode waveguide structure, but now the band gap is detuned from the signal wavelength. An increasing reverse bias applied across the junction induces a refractive index perturbation and an increasing phase shift. While bulk waveguide layers can be used, MQW InP modulators offer the most energy-efficient electro-optic conversion. Here the quantum-confined Stark effect (QCSE) is exploited. MQW layers within a PIN structure are designed with exciton absorption at a wavelength well below the desired operation wavelength. InP phase modulators require a larger detuning between signal and band gap wavelength than used for electro-absorption [8], and therefore Mach-Zehnder modulators that are built from these phase modulators have lower insertion loss. Phase modulators can also be operated in forward bias current injection mode. In this case the detuning needs to be smaller, and therefore the insertion loss is higher, but the phase tuning efficiency is very much increased. Typical values are 5 mA to tune 180 degrees in a 100 μm phase section. These phase shifters are commonly used inside laser cavities and distributed Bragg reflector (DBR) gratings to tune the lasing wavelength.

Shallow-ridge optical waveguides provide a means to connect active devices and create interferometric devices without the need to etch through the waveguide core. The removal of the contacting layers provides isolation between active components. Shallow etches may offer advantages in terms of reduced optical losses which arise from finite levels of surface roughness. As processing is increasingly well controlled, it is the dopants which dominate losses. In the absence of p-dopants, losses of below 0.5 dB/cm become feasible [9].

Deep-ridge optical waveguides provide a higher optical confinement and more compact optical bends. Bends with radii of order 20 μm have been reported [10, 11]. Deep-ridge structures are commonly used for multimode couplers, filters and arrayed waveguide grating de/multiplexers [12]. Distributed Bragg reflectors can also be implemented within the waveguides through a variation in the waveguide width, although it is more common to implement these in a separate epitaxial growth step.

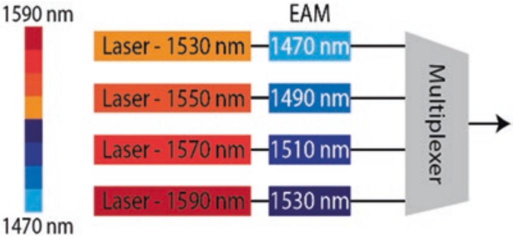


Fig. 14.2 The ability to create devices with different band edges (shown here in terms of wavelength) within the same high-density PIC. Concept for a wavelength-multiplexed transmitter [19]

Polarization rotators may be anticipated to provide the workhorse for a broad range of polarization-processing devices. Controlling the polarization within integrated circuits both accommodates variations in the input state of polarization and also allows an extra dimension for capacity scaling. Slanted side-wall waveguides have been long proposed to rotate the polarization [13]. So far, however, integrated polarization rotation devices have suffered from tight fabrication tolerances [14]. Recently a double-section polarization converter structure has been proposed and validated which has been able to correct for critical dimension variation [15, 16]. Combinations of phase modulators and polarization rotators may be implemented for conversion, splitting and combining, through to modulation [17] and analysis [18].

14.3 Monolithic Technology

Photonic-integrated circuits can be considered as multiple, interconnected, basic building blocks with varying epitaxial layers and waveguide cross-sections. Figure 14.2 shows an example schematic view for a WDM transmitter circuit implemented with a high number of different devices and band edges. Four methods stand out in the development of integrate devices with different band edges: vertically coupled epitaxial layers, quantum well intermixing, selective area growth and butt-joint integration. These methods are described below.

Vertically coupled layers have been used to enable the separate growth of active and passive integrated circuits, with notable examples in optical packet switch matrices [20], transceivers [21] and mode-matching elements [22]. The techniques have been extended most recently to enable the inclusion of InP devices on silicon photonic waveguides and facilitate efficient end-fire fibre coupling from PICs in general. The reliance on precision epitaxy and lithography impacts the yield-performance envelope, but this may be expected to improve with new generations of tooling in photonic circuit production. The physical size of tapers between elements will impact footprint and density. For some variants of silicon-based photonics where some of the active photonic building blocks originate from InP wafers, laser and amplifier integration becomes part of the assembly methodology.

Quantum well intermixing is typical of a broad range of post-growth techniques which allows the post-growth re-engineering of the band gap at the building block level across the wafer [23–25]. The use of masked capping layers in combination with an annealing process allows a change in the composition and definition of the quantum wells and an associated change in emission wavelength. This enables the wavelength of the laser to be tuned relative to passive and absorber structures. Tunable lasers with electro-absorption modulators have been created with such an approach [26].

Selective area growth engineers the band edge during the epitaxial growth. Masking is used to locally enhance the growth rate. The dimensions of the mask define the thickness of the grown quantum wells and therefore the emission wavelength for quantum well lasers and band edge for modulators and detectors [27, 28]. Multiple band gaps can be defined across the same wafer to create combinations of building blocks with a higher level of functionality within one circuit. Lasers are readily incorporated with electro-absorption modulators [29, 30], and lasers can be integrated with passive components [31, 32]. Most recently, eight-channel distributed feedback laser arrays have been created with wavelength coverage from 1447 nm to 1602 nm. Threshold currents of 20 mA and output powers of up to 18 mW at 150 mA indicate efficient operation [33].

Butt-joint integration involves the etching and regrowth of different layer stacks across the wafer. The process can be repeated multiple times with the advantage that the epitaxial layer stack can be optimized at the building block level. For instance, the laser active regions require relatively few QWs for low threshold and high efficiency compared to the modulator where a high optical confinement factor is paramount [3]. This is also the most compact scheme with no inherent distance between the active and passive elements in the circuit [34] and very low reflections from the interfaces [35].

14.4 Transmitters

The techniques and building blocks used for converting electronic signals into the optical domain are increasingly diverse, reflecting a broad range of needs from today's links and networks. The most high-profile techniques in terms of research, development and deployment currently include:

- Directly modulated lasers (DML)
- Integrated laser modulators (ILM)
- Reflective semiconductor optical amplifiers (RSOA)
- Mach-Zehnder modulators (MZMs) and vector modulators

The functionality can be derived from a relatively small set of high-performance building blocks to enable optical gain, wavelength selection and phase modulation. The sophistication in integration technology scales as the module level performance increases to meet continued scaling demands in bandwidth, footprint and power consumption.

Directly modulated laser (DML) has been demonstrated for short, high grating strength distributed feedback lasers with high confinement factor InGaAlAs quantum wells. The cavity length directly influences modulation speed, with 25Gb/s demonstrated for 200 μm length devices and 40 Gb/s for 100 μm length devices [36]. High numbers of high-confinement wells ensure the high differential gain required for the modulation bandwidth, and a high grating strength ensures the low-threshold operation necessary. A current practical consideration for die handling necessitates the butt-joint connection of a passive waveguide for very short devices. Direct modulation bandwidths of 29GHz (3 dB) have been achieved, with a modulation efficiency of 4.85 GHz/mA$^{1/2}$, enabling 40-Gb/s transmission over 40-km-long single-mode fibre. The approach has most recently been extended to incorporate high electronic confinement 1.3-μm InGaAlAs active layers and the low-leakage ridge-shaped-buried heterostructures for a low threshold current of 5.6 mA at 85 °C. Clear eye openings and 10-km signal transmission have been claimed with 50Gb/s data transfer [37]. This already indicates suitability as a cost-effective light source in 400 GbE and OTU5 applications, and the integration with a passive section opens up possibilities for further functional integration. Passive feedback lasers offer an enhanced modulation bandwidth through the interaction with the feedback field. Integration is key in enabling precise phase control in a stable cavity. 40 Gb/s operation was demonstrated [38, 39] with a two-section device comprising a DFB laser and an integrated passive cavity.

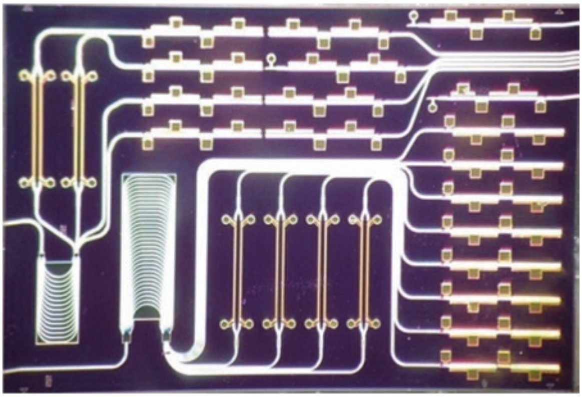
Integrated laser modulators are most compactly implemented with the combination of DFB lasers and electro-absorption modulators. The initial challenge had been to control chirp and laser frequency in the presence of spurious back reflections [40], but significant progress has been made in terms of reflection suppression and modulation bandwidth. An InGaAlAs-based electro-absorption-modulated DFB laser has been operated with up to 56 Gb/s with low chirp and high output power. Lanes operate with extinction ratios of >9 dB at peak-to-peak voltages below 2 V and milliWatt mean modulated power levels in fibre. Such technology can be readily applied for O-band, C-band and L-band transmitters [41–43]. PAM4 coding has also been demonstrated with 56 GBaud/s symbol rates [44]. Beyond 100Gb/s becomes feasible with PAM8 coding at the 28 GBaud/s symbol rate [45] on a single wavelength.

Reflective SOAs have received renewed attention for access networking where the prospect of colourless operation enables enhanced network flexibility with simplified hardware. The amplified spontaneous emission (ASE) may be spectrum sliced, self-seeded or externally seeded by a remote transmitter. The limited modulation bandwidth can be accommodated with bit- and power-loaded discrete multitone (DMT) modulation to achieve data rates of order 30.7 Gbit/s with an externally seeded scheme [46]. Data rates of 28 Gb/s are independently achieved over 20 km with a directly modulated RSOA with the polar return-to-zero (RZ) N-ary pulse-amplitude modulation (PAM) format [47]. Leveraging advanced digital signal processing and digital to analogue conversion has enabled up to 40 Gb/s transmission [48].

Dense wavelength division multiplexing (DWDM) is exploited in the highest capacity PICs. Considerable progress has been made at Infinera, with 1700 components per chip and chip capacity exceeding 2 Tb/s in 2014 [49]. A high proportion of these devices are either energy-efficient InP modulators or lasers, ensuring an inherently scalable and complete chip-scale solution to ever-rising data rate requirements. Figure 14.3 shows the constituent parts in a WDM chip created using open access generic technology (jeppix.eu). From right to left, for the lower part of the die, there is a column of single-frequency distributed Bragg reflector lasers which connect to Mach-Zehnder modulators, an arrayed waveguide grating multiplexer, and then one single fibre-optic connection at the bottom left of the image. In this case, the device is designed according to a generic integration concept, using standardized building blocks, and is fabricated in a multiproject wafer run. The device delivers up to 4 dBm of optical power into the fibre with a modulation data rate of 12.5 Gbps per transmission channel [50].

Mach-Zehnder modulators (MZMs) in their simplest implementation enable on-off-keyed amplitude modulation. Lithium niobate modulators continue to set the standard in terms of modulation linearity and bandwidth, but the efficient quantum-confined Stark effect in InP MQW MZMs provides significantly smaller, submillimetre, interaction lengths compared to devices based on the Pockels effect [3]. InP offers the smallest and most energy-efficient, commercially produced modulators and already enables direct integration with the laser source [51, 52]. By detuning of the material band edge energy by 120 nm from the laser wavelength, InP Mach-Zehnder modulators are sufficiently broadband to allow operation over ninety 50-GHz-spaced DWDM channels [8]. The $V_{2\pi}$ modulation efficiencies are less than 5 V with an extinction ratio of more than 25 dB [53]. Enhanced optical confinement is able to improve efficiency further with the very high optical overlap achieved in substrate-removed devices: Values as low as 0.6 Vπ mm have now been reported under push-pull drive [54, 55], and the longer devices have enabled bandwidth demonstrations to 67 GHz.

Fig. 14.3 Wavelength division multiplexed transmitter [50]



Vector modulators, also known as I&Q (in-phase and quadrature) modulators and nested MZMs, enable the simultaneous modulation of phase and amplitude. For higher data rates, the adoption of more complex modulation formats has contributed to the industry achievement of increasing capacity while satisfying the same link design rules as developed for 10 Gb/s transmission. The adoption of coherent technology, incorporating a local oscillator laser with a polarization and phase-diversity receiver, in combination with powerful CMOS digital signal processing (DSP), has further extended the achievable line rate and spectral efficiency. A dual-polarization (DP) I&Q modulator consisting of four MZMs can be employed to transmit 16-QAM modulation at 32 GBaud/s for a data rate of 200 Gb/s using a single wavelength [13]. Increased optical modulation complexity and loss can be compensated by the integration of optimally located SOA elements to achieve efficient optical gain, low noise figure and high saturation power. This is confirmed by operation within a CFP2 pluggable module [56]. An integrated tunable laser and vector modulator used with external polarization division multiplexing (PDM) emulation and digital coherent detection enabled data rates of 256 Gb/s per wavelength with PDM-16-QAM [57] and power dissipation as low as 3.2 W achievable [58]. The need for even greater capacity and spectral efficiency drives higher cardinality modulation formats such as PM-64QAM [59]. Further energy and footprint reductions are conceivable through, e.g. the integration of polarization optics, and innovative photonic design enables DAC-free operation, integrating traditionally electronic functionality within the PIC itself [60].

Integrated tuneable lasers leveraging sampled grating technologies have been an important enabler for telecom networking. The primary motivation was for inventory reduction, but considerable research interest has explored the potential for fast wavelength reconfigurability [61, 62]. For network reconfiguration, tunable laser products have been wavelength retuned within several nanoseconds with appropriate electronic control planes for the compensation of thermal crosstalk [63]. A programmable wavelength is an enabler for flex-grid and grid-less architectures, and the use of lasers with coherent systems enables optical filter-free detection. Integrating the lasers and modulators has the important manufacturability advantage of enabling full photonic testing at the wafer scale, to ensure the packaging of known good die. Performance parameters such as laser power, threshold and wavelength tuning characteristics as well as modulator switching voltage can be determined with wafer probe testing, before single dies are separated and packaged. Test components that have no function in the final application, such as test laser sources on WDM receiver chips, are now being designed into devices to get as much test and yield data before chips are committed to the packaging and assembly process.

Scaling to higher information densities requires a mitigation and suppression of thermal, optical and electronic crosstalk. The thermal crosstalk between active and passive components can limit the performance of integrated Mach-Zehnder (MZ) modulators operating at high radio frequencies and has been quantified by measuring the effects on the electro-optical response of neighbouring MZ modulators [64]. The role of substrate thickness is similarly important as this defines the relative proximity of the heat sink [65]. Thermal crosstalk can be reduced through the incorporation of deep trenches [66]. Electrical crosstalk is also observed between interconnect lines and electro-optical phase shifters in photonic-integrated circuits. Crosstalk originates from radiative and substrate coupling between lines and from shared ground connections [67].

14.5 Receivers

The photodiodes at the heart of an optical receiver convert optical amplitude-modulated signals to analogue electronic waveform. The range of detectors which have been developed reflects the optimizations which can be made in terms of sensitivity, power handling and electrical energy use. The early emphasis in direct-detection, long-haul fibre optic links has driven ever more sensitive receivers, although more recently, short-reach and digital coherent links have required high operating optical powers. The shorter-reach links have focussed on energy reduction at the link level, and higher received optical powers can lead to energy savings through reduced electronic amplification in the receiver. The main classes of detector include:

- Waveguide P-I-N detectors
- Avalanche photodiodes (APD)
- Optically amplified receivers
- Unitravelling carrier detectors (UTC)
- Coherent detection

Waveguide P-I-N detectors provide a highly efficient, wide-band means to convert amplitude-modulated data into the electronic domain, but the sensitivity is ultimately limited by the thermal noise in the receiver and any dark current. Early experimental evidence indicated that high-speed response with bandwidths of 110 GHz [68] was feasible with waveguide integration, and theoretical estimates indicated that 200 GHz should be feasible with appropriate levels of fabrication control. The waveguide approach enables the simultaneous optimization of detection efficiency and bandwidth efficiency [69]. A 40Gbit/s PIN diode integrated with a transimpedance amplifier can achieve a sensitivity of −10.5 dBm at a BER of 10^{-9} [70]. Side-illuminated photodetectors show an improved high-power behaviour, as the absorption is distributed laterally into a larger length of a thinner absorption layer in a controlled manner, compared to perpendicular illuminated detectors. Line rates of 85Gb/s and bandwidths of 110GHz have been achieved on semi-insulating substrates [71]. Ultimately PIN photodiodes will show saturation due to large densities of carriers generated in the depletion region [72].

Avalanche photodiodes are implemented with waveguides to achieve both high-speed performance and high responsivity through the exploitation of internal gain [72, 73]. A typical improvement of the photoreceiver sensitivity would be 5–10 dB compared with PIN photodiodes [71, 74]. Waveguide-integrated devices with sensitivity of -19 dBm at 40 Gb/s have been demonstrated [75, 76].

Waveguide integration also enables the further integration with other photonic components. Progress in epitaxial growth and specifically techniques with low residual doping and precise control of heterostructure interfaces have been instrumental in optimizing detectors, and this is particularly true for APDs where particularly high electric fields are required for avalanche gain. APD performances are related to electron and hole ionization coefficients in the multiplication region. Very low-noise APDs were demonstrated using a wide range of very thin avalanche layers, including InP, Al_x $In_{1-x}As$, GaAs and $Al_xGa_{1-x}As$.

Optically preamplified receivers with an SOA-PIN combination can enhance the sensitivity of a PIN detector to −17.5 dBm at 40 Gb/s [77]. The combination of optimized SOA preamplification and TIA design enables a large responsivity of 44 A/W, a polarization dependence below 2 dB, a low noise figure of 8.5 dB and a 3-dB bandwidth of 35 GHz [78]. Recently a 40-Gb/s photoreceiver module with differential outputs designed for short-reach applications such as access network and data centre interconnect was reported [79]. It consists of an InP semiconductor optical amplifier monolithically integrated with a p-i-n (SOA-PIN) photodiode, co-packaged with an InP linear transimpedance amplifier (TIA) and a matching circuit between the SOA-PIN and the TIA in order to increase the cutoff frequency. The module exhibits a − 3 dB bandwidth of 43 GHz, a single-ended optoelectrical conversion gain of 10,000 V/W for an optical input power of −25 dBm and a record sensitivity of −22.5 dBm at a bit error rate of 10^{-9} at 40 Gb/s in non-return-to-zero on/off -keying operation (Fig. 14.4).

Unitravelling carrier photodiodes (UTC) offer a means to scale in power and bandwidth. Unipolar photodiodes offer a particularly powerful structure to reduce space charge effect limitations since they use only electrons as active carriers [71]. High-speed unitravelling carrier photodiode has been demonstrated for 100 Gbit/s applications with a 3 dB bandwidth exceeding 110 GHz, a dark current of 1 nA and a peak saturation current of about 30 mA at −2 V [80]. When monolithically integrated with a semiconductor optical amplifier (SOA), a 95 GHz 3 dB bandwidth is still feasible with 8 dB noise figure and a polarization-dependent loss of 1–2 dB. The SOA integration enables a 95 A/W peak responsivity corresponding to record gain-bandwidth product of 6.1 THz [81]. UTC technology has also been extended to

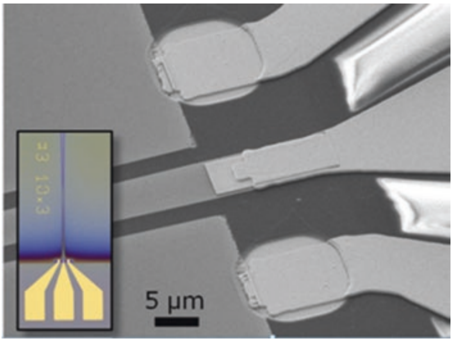


Fig. 14.4 UTC detector with >67 GHz bandwidth using InP membrane on silicon (IMOS) technology [83]

continuous-wave terahertz (THz) signal generation at 1.25 THz [82]. Integration of high-speed UTC devices with the substrate-free IMOS InP-membrane-on-silicon platform also enables integration with surface grating couplers and a route to high-density integration [83]. The higher optical confinement ensures a smaller cross-section, smaller bend radii, the possibility to use high-contrast reflectors for cavities and a route to smaller and denser packed devices. For the case of the photodiodes, a 150 nm thin p-type doped InGaAs layer is used both as the absorption layer and as the p-contact layer. The photogenerated holes are collected directly by the p-contact, while the electrons travel to the non-intentionally doped depletion region. The thickness of the p-layer is chosen as a trade-off between the optical absorption coefficient for efficiency and the electron transit time for bandwidth.

Coherent detection requires further integration, implementing the mixers with near-identical, ground-isolated, photodiodes to enable phase-sensitive detection and a richer range of bandwidth-efficient modulation formats. InP-based MMI (multimode interferometer)-mixer chip has been integrated with photodiodes by the butt-joint process to achieve high responsivity in a chip size of 2.0 mm × 5.1 mm. The 3 dB bandwidth is more than 20 GHz, and uniform characteristics of over four PDs have been achieved to enable 100Gb/s operation [84]. InP coherent receiver chip with the highest reported responsivity (0.15A/W) together with excellent RF bandwidth (32GHz) and 4 × 4 MMI width fabrication control (< ±60 nm 90% population) provides a highly manufacturable receiver for pluggable CFP2 modules [85]. InP offers an attractive and manufacturable platform for size and cost reduction as well as a common platform for full transceiver (laser, transmitter and receiver) integration.

14.6 Optical Switching

The energy and latency cost converting signals back and forth between the electrical and optical domain has long motivated R&D into optical switching technologies. While wavelength-selective switches, photonic switches and reconfigurable optical add-drop multiplexers are now an established part of the network, the technologies so far deployed have been operated as circuit switches. This is primarily a technology limitation. Microelectro-mechanical systems are widely used to provide high connectivity switching fabrics with hundreds of fibre connections. These approaches use free-space imaging of fibre arrays onto two-dimensional arrays of voltage-actuated micromechanical switch elements. This approach can be energy efficient as the low switch actuation power and optical power loss are both loss, and the devices can be transparent to bandwidth, enabling the routing of many tens of high bandwidth channels. However the actuation times, the requirement for power levelling and physical size continue to pose system implementation challenges. A rich vein of InP PIC research has sort to exploit nanosecond switch actuation, on-chip levelling and chip-scale implementation to enable reduced latency networking [86]. Switch architectures are primarily aligned to photonic switches with broadband any-port-to-any-port routing and wavelength-selective switches. Multi-degree ROADMs may be implemented with combinations of the two architectures.

Photonic switches have been implemented at the chip scale using InP PICs with connectivities ranging 1 × 100 and 16 × 16. Single-input-port, integrated, phased-array optical switches offer a high port-count scalability and broad spectral coverage and can be used as building blocks of large-scale optical routers. Single stages of a 1 × 16 switch feature wavelength-independent nanosecond switching characteristics [87]. Scaling to 1 × 100 is demonstrated with an active-passive integration technology and a two-stage phase array interferometric switch. The inclusion of active SOA gates on the output enables an enhanced switch extinction [88] and the possibility for gain compensation. Such circuits have even been used to enable optical buffering experiments [89]. The implementation of multiple arrayed waveguides with shared free-space regions has also been explored with the creation of a strictly nonblocking 8 × 8 switch for high-speed, WDM optical interconnection [90]. The circuit consists of over 200 functional devices such as star couplers, phase shifters and avoided waveguide crossings. C-band operation with extinction ratio performance of more than 20 dB was achieved with nanosecond reconfiguration times. N × N switching matrices have also been implemented with combinations of Mach-Zehnder interferometers and SOA gates for 8 × 8 switch fabrics [91]. Broadcast and select architectures have been implemented in multiple stages to enable higher levels of connectivity in 16 × 16 fabrics [11, 92].

Reconfigurable optical packet switches enable per wavelength routing between multiple ports. InP PICs have been created using filter elements such as chained Mach-Zehnder interferometers [93], cross-point matrix implementations of third-order ring resonators [94] and arrayed waveguide grating-based wavelength selectors [95]. Wavelength selector circuits have also been implemented in parallel on the same die [96]. The demonstrated nanosecond switching of high line-rate data using such switches has provided a powerful means of enabling packet level routing at the chip level. Many-to-many connectivity with wavelength granularity becomes feasible with the combination of shuffle networks and parallel wavelength selectors. An example circuit is shown in Fig. 14.5 with eight-input to

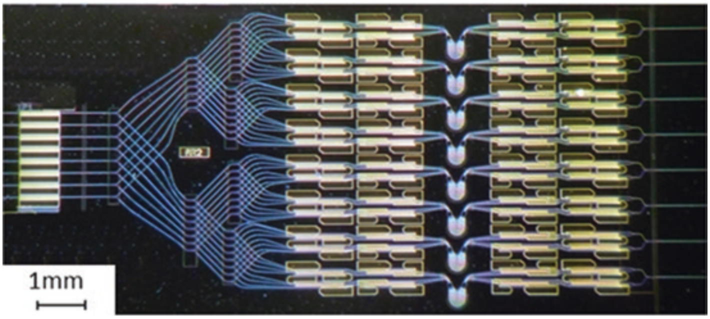


Fig. 14.5 Reconfigurable optical packet switch operating on eight input fibres (*left*), eight wavelength channels and eight output fibres (*right*) [5]

eight-output connectivity with cyclic arrayed waveguide grating routers operating on eight individual wavelengths or combs of wavelengths [5]. Dynamic routing has been demonstrated with real-time path reallocation with 16 channels with microsecond time slots [97].

14.7 Outlook

InP-integrated photonics scales along the same trajectory as other wafer-scale technologies. An increased production volume can be expected to lead to further yield improvements and cost reductions. The low-energy requirement from quantum-well-based building blocks enables shorter components with enhanced packing density for a given thermal load and chip area. This provides an important route to very high-density integration for terabit-class transmitters, receivers and routing circuits. Progress over the last decades has already led to the observation of a Moore's Law in photonic integration [98]. Indium phosphide membrane on silicon technology [99] may be expected to lead to a further step in miniaturization through high-confinement, ultradense InP optoelectronics for further footprint and energy reductions.

The package for PICs currently requires optomechanical connections, electronic connections and thermal management, each of which adds losses and cost. In telecommunications, the achievable data rate for a given link outweighs cost. For the data centre interconnect, data rate, cost and energy use are all critically important. Packages will become smaller with increasing levels of electronics co-packaged with the PIC. New methods to enable relaxed precision assembly will be critical. New techniques for automated optical alignment [100, 101] and reduced complexity electronic connection [100] become areas of active research.

Systems integration will become critical. Matching electronic and photonic design – co-design – and designing for uncooled (high temperature) operation are expected to have a major impact on both energy use and also assembly cost. As with electronic ICs, costs are initially dominated by auxiliary components, package and test, rather than the enabling chip itself. Here InP has a striking advantage. The monolithically integrated light sources already enable wafer-scale self-test. Uniquely, the lasers and amplifiers are created within one chip with high-performance multi-quantum well modulators, detectors and passives without introducing assembly steps between photonic devices. The epitaxial growth of energy-efficient multi-quantum wells and the use of thermally conductive waveguide cladding layers offer important energy efficiency advantages. Butt-joint integration allows component packing at the density limit, without the use of adiabatic tapers or extra-thick cladding layers. Other platforms offer subsets of devices and therefore may be assembled with other platforms in the packaged part, but the monolithic approach feasible for InP PICs ensures a chip-scale solution for sustained year-on-year scaling.

Acknowledgements Past and present members of the Institute for Photonic Integration, formerly known as the COBRA institute, are thanked for their stimulating contributions in the preparation of this chapter. The authors are also grateful to partners in JePPIX – the Joint European Platform for Photonic Integrated Components and Circuits – for their insights. Research projects which have had a particular input include Paradigm, Phastflex, GetPICs, WIPE, Photronics and Smartlight.

References

1. C.R. Doerr, K. Okamoto, Planar lightwave circuits in fiber-optic communications, in *Optical Fiber Telecommunications V A (Fifth Edition) Volume A: Components and Subsystems*, ed. by I. P. Kaminow, T. Li, A. E. Willner (Eds), (Elsevier, Amsterdam, 2008)
2. R. Nagarajan, M. Kato, J. Pleumeekers, P. Evans, S. Corzine, S. Hurtt, A. Dentai, S. Murthy, M. Missey, R. Muthiah, R.A. Salvatore, C. Joyner, R. Schneider, M. Ziari, F. Kish, D. Welch, InP photonic integrated circuits. Invited Paper, IEEE J. Sel. Top. Quantum Electron. **16**(5), 1113 (2010)
3. R.A. Griffin, S.K. Jones, N. Whitbread, S.C. Heck, L.N. Langley, InP Mach–Zehnder modulator platform for 10/40/100/200-Gb/s operation. Invited Paper, IEEE J. Sel. Top. Quantum Electron. **19**(6), 3401209 (2013)
4. F. Kish, R. Nagarajan, D. Welch, et al., From visible light-emitting diodes to large scale III-V photonic integrated circuits. Proc. IEEE **101**(10), 2255–2270 (2013)
5. R. Stabile, A. Rohit, K.A. Williams, Monolithically integrated 8×8 space and wavelength selective cross-connect. J. Lightwave Technol. **32**(2), 201 (2014)
6. M.K. Smit et al., An introduction to InP-based generic integration technology. Semicond. Sci. Technol. **29**(8), 083001–081/41 (2014)
7. JePPIX roadmap http://www.jeppix.eu/document_store/JePPIXRoadmap2015.pdf, 2015
8. R.A. Griffin, B. Pugh, J. Fraser, I.B. Betty, K. Anderson, G. Busico, C. Edge, T. Simmons, Compact, high power, MQW InP Mach-Zehnder transmitters with full-band tunability for 10 Gb/s DWDM. 4, 903–904.proceedings European Conference on Optical Communications (2005)
9. D. d'Agostino, G. Carnicella, C. Ciminelli, H.P.M.M. Ambrosius, M.K. Smit, Design of a compact high-performance InP ring resonator. Proceedings MEPHISTO 2014
10. R. Stabile, K.A. Williams, Relaxed dimensional tolerance whispering gallery microbends. J. Lightwave Technol. **29**(12), 2011 (1892)
11. R. Stabile, A. Albores-Mejia, K.A. Williams, Monolithic active-passive 16 × 16 optoelectronic switch. Opt. Lett. **37**(22), 4666 (2012)
12. K.A. Williams, E.A.J.M. Bente, D. Heiss, Y. Jiao, K. Ławniczuk, X.J.M. Leijtens, J.J.G.M. van der Tol, M.K. Smit, InP photonic circuits using generic integration. Photon. Res. **3**(5), B60–B68 (2015)
13. H. el-Refaei, D. Yevick, T. Jones, Slanted-rib waveguide InGaAsP–InP polarization converters. J. Lightwave Technol. **22**(5), 1352 (2004)
14. M. Zaitsu, T. Tanemura, Y. Nakano, Numerical study on fabrication tolerance of half-ridge InP polarization converters. IEICE Trans. Electron. **97-C**(7), 731 (2014)
15. D.O. Dzibrou, J.J.G.M. van der Tol, M.K. Smit, Improved fabrication process of low-loss and efficient polarization converters in InP-based photonic integrated circuits. Opt. Lett. **38**(7), 1061 (2013)
16. D.O. Dzibrou, J.J.G.M. van der Tol, M.K. Smit, Tolerant polarization converter for InGaAsP-InP photonic integrated circuits. Opt. Lett. **38**(18), 3482 (2013)
17. M.A. Naeem, M. Haji, B.M. Holmes, D.C. Hutchings, J.H. Marsh, A.E. Kelly, Generation of high speed polarization modulated data using a monolithically integrated device. IEEE J. Sel. Top. Quantum Electron. **21**(4), 3400205 (2015)
18. S. Ghosh, Y. Kawabata, T. Tanemura, Y. Nakano, Integrated Stokes vector analyzer on InP. Paper WD4–4, proceedings OECC/PS (2016)

19. M. Trajkovic, High speed electro-absorption modulator: a step towards high performance and high density PICs. Fotonica Magazine (2016)
20. S.C. Lee, R. Varrazza, S. Yu, Advanced optical packet switching functions using active vertical-couplers-based optical switch matrix. J. Sel. Top. Quantum Electron. **12**(4), 817–827 (2006)
21. V. Tolstikhin, Multi-guide vertical integration in InP: PIC technology for cost-sensitive applications. Proceedings Conference on Lasers and Electro-Optics Pacific Rim (2013)
22. I. Moerman, P.P. van Daele, P.M. Demeester, A review on fabrication technologies for the monolithic integration of tapers with III-V semiconductor devices. IEEE J. Sel. Top. Quantum Electron. **3**(6), 1308–1320 (1997)
23. S. McDougall, O. Kowalski, C. Hamilton, F. Camacho, B. Qiu, M. Ke, R. De La Rue, A. Bryce, J. Marsh, Monolithic integration via a universal damage enhanced quantum-well intermixing technique. IEEE J. Sel. Top. Quantum Electron. **4**(4), 636–646 (1998)
24. A. McKee, C.J. McLean, G. Lullo, A.C. Bryce, R.M. Rue, J.H. Marsh, Monolithic integration in InGaAs-InGaAsP multiple-quantum-well structures using laser intermixing. IEEE J. Quantum Electron. **33**, 45–55 (1997)
25. E.J. Skogen, J.W. Raring, G.B. Morrison, C.S. Wang, V. Lal, M.L. Masanovic, L.A. Coldren, Monolithically integrated active components: A quantum-well intermixing approach. J. Sel. Top. Quantum Electron. **11**(2), 343–355 (2005)
26. J.W. Raring, E.J. Skogen, L.A. Johansson, M.N. Sysak, S.P. DenBaars, L.A. Coldren, Widely tunable negative-chirp SG-DBR laser/EA-modulated transmitter. J. Lightwave Technol. **23**(1), 80–86 (2005)
27. R. Bhat, Non-planar and masked-area epitaxy by organometallic chemical vapour deposition. Semicond. Sci. Technol. **8**, 984–993 (1993)
28. M. Gibbon, J.P. Stags, C.G. Cureton, E.J. Thrush, C.J. Jones, Selective-area low-pressure MOCVD of GalnAsP and related materials on planar InP substrates. Semicond. Sci. Technol **8**, 998–1010 (1993)
29. N. Dupuis, J. Décobert, C. Jany, F. Alexandre, A. Garreau, R. Brenot, N. Lagay, F. Martin, D. Carpentier, J. Landreau, F. Pommereau, F. Poingt and C. Kazmierski, Selective area growth engineering for 80 nm spectral range AlGaInAs 10 Gbit/s remote amplified modulator. Proceedings indium phosphide and related materials (2008)
30. H. Debrégeas, J. Decobert, N. Lagay, R. Guillamet, D. Carrara, O. Patard, C. Kazmierski, R. Brenot, Selective-area-growth technology for flexible active building blocks. Proceedings advanced photonics congress, IM2A.3 (2012)
31. J. Décobert, N. Dupuis, P.Y. Lagrée, N. Lagay, A. Ramdane, A. Ougazzaden, F. Poingt, C. Cuisin, C. Kazmierski, Modeling and characterization of AlGaInAs and related materials using selective area growth by metal-organic vapor phase epitaxy. J. Cryst. Growth **298**, 28–31 (2007)
32. J. Decobert, G. Binet, A.D.B. Maia, P.Y. Lagrée, Christophe Kazmierski, "AlGaInAs MOVPE selective area growth for photonic integrated circuits". Adv Opt Technol **4**, 2 (2015)
33. F. Soares, M. F. Baier, Z. Zhang, T. Gaertner, D. Franke, J. Decobert, M. Achouche, D. Schmidt, M. Moehrle, N. Grote, M. Schell, 155 nm-span multi-wavelength DFB laser array fabricated by selective area growth. Paper MoC4–4. Proceedings compound semiconductor week (2016)
34. J. Binsma, P. Thijs, T. van Dongen, E. Jansen, A. Staring, G. van den Hoven, L. Tiemeijer, Characterization of butt-joint InGaAsP waveguides and their application to 1310 nm DBR-type MQW gain-clamped semiconductor optical amplifiers. IEICE Trans. Electron. **E80-C**, 675–681 (1997)
35. Y. Barbarin, E.A.J.M. Bente, C. Marquet, E.J.S. Leclère, J.J.M. Binsma, M.K. Smit, Measurement of reflectivity of butt-joint active–passive interfaces in integrated extended cavity lasers. Photon. Technol. Lett. **17**(11), 2265–2267 (2005)
36. W. Kobayashi, T. Tadokoro, T. Fujisawa, N. Fujiwara, T. Yamanaka and F. Kano, 40-Gbps direct modulation of 1.3-μm InGaAlAs DFB laser in compact TO-CAN package. Paper OWD2, proceedings optical fiber communications conference (2011)

37. K. Nakahara, Y. Wakayama, T. Kitatani, T. Taniguchi, T. Fukamachi, Y. Sakuma, S. Tanaka, Direct modulation at 56 and 50 Gb/s of 1.3-μm InGaAlAs ridge-shaped-BH DFB lasers. IEEE Photon. Technol. Lett. **27**(5), 534–536 (2015)
38. U. Troppenz, J. Kreissl, M. Möhrle, C. Bornholdt, W. Rehbein, B. Sartorius, I. Woods, M. Schell, 40 Gbit/s directly modulated lasers: Physics and application. 7953, 79530F-1–79530F-10, proceedings SPIE (2011)
39. J. Kreissl, V. Vercesi, U. Troppenz, T. Gaertner, W. Wenisch, M. Schell, Up to 40-Gb/s directly modulated laser operating at low driving current: Buried-heterostructure passive feedback laser (BH-PFL). IEEE Photon. Technol. Lett. **24**(5), 362 (2012)
40. J.A.J. Fells, M.A. Gibbon, G.H.B. Thompson, I.H. White, R.V. Penty, A.P. Wright, R.A. Saunders, C.J. Armistead, E.M. Kimber, Chirp and system performance of integrated laser modulators. IEEE Photon. Technol. Lett. **7**(11), 1279 (1995)
41. M. Theurer, Y. Wang, L. Zeng, U. Troppenz, G. Przyrembel, A. Sigmund, M. Moehrle, M. Schell, 2×56 Gb/s from a double side electroabsorption modulated DFB laser. Paper Tu3D.6 OFC (2016)
42. M. Theurer, H. Zhang, Y. Wang, W. Chen, L. Zeng, U. Troppenz, G. Przyrembel, A. Sigmund, M. Moehrle, M. Schell, 2×56 Gb/s from a double side electroabsorption modulated DFB laser and application in novel optical PAM4 generation. J. Lightwave Technol. Accepted for publication 03 August 2016
43. M. Theurer, G. Przyrembel, A. Sigmund, W.D. Molzow, U. Troppenz, M. Mohrle, 56 Gb/s L-band InGaAlAs ridge waveguide electroabsorption modulated laser with integrated SOA. Phys. Status Solidi A **213**(4), 970–974 (2016)
44. M.A. Mestre, H. Mardoyan, C. Caillaud, R. Rios-Muller, J. Renaudier, P. Jenneve, F. Blache, F. Pommereau, J. Decobert, F. Jorge, P. Charbonnier, A. Konczykowska, J.Y. Dupuy, K. Mekhazni, J.F. Paret, M. Faugeron, F. Mallecot, M. Achouche, S. Bigo, Compact InP-based DFB-EAM enabling PAM-4 112 Gb/s transmission over 2 km. J. Lightwave Technol. **34**(7), 1572 (2016)
45. U. Troppenz, M. Narodovitch, C. Kottke, G. Przyrembel, W.D. Molzow, A. Sigmund, H.G. Bach, M. Moehrle, 1.3 μm electroabsorption modulated lasers for PAM4/PAM8 single channel 100 Gb/s. Paper Th-B2–5, Montpelier, international conference on indium phosphide and related materials (2014)
46. S.A. Gebrewold, R. Brenot, R. Bonjour, A. Josten, B. Baeuerle, D. Hillerkuss, C. Hafner, J. Leuthold, Colorless low-cost RSOA based transmitters optimized for highest capacity through bit- and power-loaded DMT. Proceedings optical fiber communications conference, Tu2C.4 (2016)
47. H.K. Shim, H. Kim, Y.C. Chung, Effects of electrical and optical equalizations in 28-Gb/s RSOA-based WDM PON. Photon. Technol. Lett. **28**(22), 2537–2540 (2016)
48. B.Y. Cao, M.L. Deng, R.P. Giddings, X. Duan, Q.W. Zhang, M. Wang, J.M. Tang, RSOA intensity modulator frequency chirp-enabled 40Gb/s over 25km IMDD PON systems. Proceedings optical fiber communications conference, W1J.3 (2015)
49. J. Summers, T. Vallaitis, P. Evans, M. Ziari, P. Studenkov, M. Fisher, J. Sena, A. James, S. Corzine, D. Pavinski, J. Ou-Yang, M. Missey, D. Gold, W. Williams, M. Lai, D. Welch, F. Kish, Monolithic InP-based coherent transmitter photonic integrated circuit with 2.25 Tbit/s capacity. Electron. Lett. **50**(16), 1150 (2014)
50. K. Ławniczuk, C. Kazmierski, J.G. Provost, M.J. Wale, R. Piramidowicz, P. Szczepanski, M.K. Smit, X.J.M. Leijtens, InP-based photonic multiwavelength transmitter with DBR laser array. Photon Technol. Lett. **25**(4), 352 (2013)
51. J.E. Zucker, K.L. Jones, B.I. Miller, U. Koren, Miniature Mach-Zehnder InGaAsP quantum well waveguide interferometers for 1.3 μm. IEEE Photon. Technol. Lett. **2**(1), 32–34 (1990)
52. J.E. Zucker, K.L. Jones, M.A. Newkirk, R.P. Gnall, B.I. Miller, M.G. Young, U. Koren, C.A. Burrus, B. Tell, Quantum well interferometric modulator monolithically integrated with 1.55 μm tunable distributed Bragg reflector laser. Electron. Lett. **28**(20), 1888–1889 (1992)

53. S.C. Heck, S.K. Jones, R.A. Griffin, N. Whitbread, P.A. Bromley, G. Harris, D. Smith, L.N. Langley, T. Goodhall, Miniaturized InP dual I&Q Mach Zehnder modulator with full monitoring functionality for CFP2. Proceedings European conference on optical communications, paper Tu.4.4.2 (2014)
54. S. Dogru, N. Dagli, 0.77-V drive voltage electro-optic modulator with bandwidth exceeding 67 GHz. Opt. Lett. **39**(20), 6074 (2014)
55. S. Dogru, N. Dagli, 0.2V drive voltage substrate removed electro-optic Mach-Zehnder modulators with MQW cores at 1.55 μm. IEEE/OSA J. Lightwave Technol. **32**(3), 435–439 (2014)
56. R.A. Griffin, N. D. Whitbread, S.K. Jones, S.C. Heck, P. Firth, D. Govan, T. Goodall, InP coherent optical modulator with integrated amplification for high capacity transmission. Paper Th4E.2, proceedings optical fiber communications conference (2015)
57. S. Chandrasekhar, X. Liu, P.J. Winzer, J.E. Simsarian, R.A. Griffin, Compact all-InP laser-vector-modulator for generation and transmission of 100-Gb/s PDM-QPSK and 200-Gb/s PDM-16-QAM. J. Lightwave Technol. **32**(4), 736 (2014)
58. T. Tatsumi, N. Itabashi, T. Ikagawa, N. Kono, M. Seki, K. Tanaka, K. Yamaji, Y. Fujimura, K. Uesaka, T. Nakabayashi, H. Shoji, S. Ogita, A compact low-power 224-Gb/s DP-16QAM modulator module with InP-based modulator and linear driver ICs. Paper Tu3H.5, proceedings optical fiber communications conference (2014)
59. W. Forysiak, D.S. Govan, Progress toward 100-G digital coherent pluggables using InP-based photonics. J. Lightwave Technol. **32**(16), 2925 (2014)
60. A. Aimone, I.G. Lopez, S. Alreesh, P. Rito, T. Brast, V. Hohns. G. Fiol, M. Gruner, J. Fischer, J. Honecker, A. Steffan, D. Kissinger, A.C. Ulusoy, M. Schell, DAC-free ultra-low-power dual-polarization 64-QAM transmission with InP IQ segmented MZM module. Postdeadline paper Th5C.6 proceedings optical fiber communications conference (2016)
61. L.A. Coldren, G.A. Fish, Y. Akulova, J.S. Barton, L. Johansson, C.W. Coldren, Tunable semiconductor lasers: A tutorial. J. Lightwave Technol. **22**(1), 193–202 (2004)
62. A.J. Ward, D.J. Robbins, G. Busico, E. Barton, L. Ponnampalam, J.P. Duck, N.D. Whitbread, P.J. Williams, D.C.J. Reid, A.C. Carter, M.J. Wale, Widely tunable DS-DBR laser with monolithically integrated SOA: Design and performance. J. Sel. Top. Quantum Electron. **11**(1), 149 (2005)
63. J.E. Simsarian, M.C. Larson, H.E. Garrett, H. Xu, T.A. Strand, Less than 5-ns wavelength switching with an SG-DBR laser. Photon. Technol. Lett. **18**(4), 565 (2006)
64. G. Gilardi, W. Yao, M.K. Smit, M.J. Wale, Observation of dynamic extinction ratio and bit error rate degradation due to thermal effects in integrated modulators. J. Lightwave Technol. **33**(11), 2199 (2015)
65. G. Gilardi, W. Yao, H.R. Haghighi, M.K. Smit, M.J. Wale, Substrate thickness effects on thermal crosstalk in InP-based photonic integrated circuits. J. Lightwave Technol. **32**(17), 3061 (2014)
66. G. Gilardi, W. Yao, H.R. Haghighi, X.J.M. Leijtens, M.K. Smit, M.J. Wale, Deep trenches for thermal crosstalk reduction in InP-based photonic integrated circuits. J. Lightwave Technol. **32**(24), 4864 (2014)
67. W. Yao, G. Gilardi, N. Calabretta, M.K. Smit, M.J. Wale, Experimental and numerical study of electrical crosstalk in photonic integrated circuits. J. Lightwave Technol. **33**(4), 934 (2015)
68. K. Kato, A. Kozen, Y. Muramoto, Y. Itaya, T. Nagatsuma, M. Yaita, 110 GHz, 50% efficiency mushroom-mesa waveguide p-i-n photodiode for a 1.55 μm wave-length. Photon. Technol. Lett. **6**, 719–721 (1994)
69. J.E. Bowers, C.A. Burrus, Ultrawide-band long-wavelength p-i-n photodetectors. J. Lightwave Technol. **5**(10), 1339–1350 (1987)
70. R. Vetury, I. Gontijo, K. Krishnamurthy, R. Pullela, M.J. Rodwell, High sensitivity and wide-dynamic-range optical receiver for 40 Gbit/s optical communication networks. Electron. Lett. **39**(1), 91–92 (2003)
71. H.G. Bach, Ultra high-speed photodetectors and photoreceivers for telecom and datacom also aiming at THz applications. Proceedings European conference on integrated optics, FB0 (2007)

72. (a) M. Achouche, G. Glastre, C. Caillaud, M. Lahrichi, M. Chtioui, D. Carpentier, InGaAs communication photodiodes: From low- to high-power-level designs. Photon. J. Invited Paper **2**(3), 460 (2010). (b) J. Wei, F. Xia, S.R. Forrest, A high-responsivity high-bandwidth asymmetric twin-waveguide coupled InGaAs-InP-InAlAs avalanche photodiode. IEEE Photon. Technol. Lett. **14**(11), 1590–1592 (2002)
73. K. Shiba, T. Nakata, T. Takeuchi, K. Kasahara, and K. Makita, Theoretical and experimental study on waveguide avalanche photodiodes with an undepleted absorption layer for 25-Gb/s operation. J. Lightwave Technol. **29**(2), 153 (2011)
74. J.C. Campbell, Recent advances in telecommunications avalanche photodiodes. J. Lightwave Technol. **25**(1), 109 (2007)
75. T. Nakata, T. Takeuchi, K. Maliita, Y. Amamiya, T. Kalo, Y. Suzuki, T. Torikai, High-sensitivity 40-Gb/s receiver with a wideband InAlAs waveguide avalanche photodiode. Proceedings European conference on optical commununications, Paper 10.5.1 (2002)
76. K. Makita, T. Nakata, K. Shiba, T. Takeuchi, 40 Gbps waveguide photodiodes. NEC J Adv. Technol. 234–240, Summer (2005)
77. B. Mason, S. Chandrasekhar, A. Ougazzaden, C. Lentz, J.M. Geary, L.L. Buhl, L. Peticolas, K. Glogovsky, J.M. Freund, L. Reynolds, G. Przybylek, F. Walters, A. Sirenko, J. Boardman, T. Kercher, M. Radar, J. Grenko, D. Monroe, L. Ketelsen, Photonic integrated receiver for 40 Gbit/s transmission. Electron. Lett. **38**(20), 1196–1197 (2002)
78. C. Caillaud, P. Chanclou, F. Blache, P. Angelini, B. Duval, P. Charbonnier, D. Lanteri, G. Glastre, M. Achouche, Integrated SOA-PIN detector for high-speed short reach applications. Invited Paper, J. Lightwave Technol. **33**(8), 1596 (2015)
79. P. Angelini, F. Blache, C. Caillaud, P. Chanclou, M. Goix, F. Jorge, K. Mekhazni, J.Y. Dupuy, M. Achouche, Record −22.5 dBm sensitivity SOA-PIN-TIA photoreceiver module for 40 Gb/s applications. IEEE Photon. Technol. Lett. **27**(19), 2027 (2015)
80. M. Anagnosti, C. Caillaud, F. Blache, F. Jorge, P. Angelini, J.F. Paret, M. Achouche, Optimized high speed UTC photodiode for 100 Gbit/s applications. J. Sel. Top. Quantum Electron. **20**(6), 3801107 (2014)
81. M. Anagnosti, C. Caillaud, J.F. Paret, F. Pommereau, G. Glastre, F. Blache, M. Achouche, Record gain × bandwidth (6.1 THz) monolithically integrated SOA-UTC photoreceiver for 100-Gbit/s applications. Invited Paper, J. Lightwave Technol. **33**(6), 1186 (2015)
82. M. Theurer, T. Göbel, D. Stanze, U. Troppenz, F. Soares, N. Grote, M. Schell, Photonic-integrated circuit for continuous-wave THz generation. Opt. Lett. **38**(19), 3724 (2013)
83. L. Shen, Y. Jiao, W. Yao, Z. Cao, J.P. van Engelen, G.C. Roelkens, M.K. Roelkens, M.K. Smit, J.J.G.M. van der Tol, High-bandwidth uni-traveling carrier waveguide photodetector on an InP-membrane-on-silicon platform. Opt. Express **24**(8), 8290–8301 (2016)
84. Y. Tateiwa, M. Takechi, H. Yagi, Y. Yoneda, K. Yamaji, Y. Fujimura, 100 Gbit/s compact digital coherent receiver using InP-based mixer. SEI Tech. Rev. **77**, 59 (2013)
85. S. Farwell, P. Aivaliotis, Y. Qian, P. Bromley, R. Griggs, J.N.Y. Hoe, C. Smith, S. Jones, InP coherent receiver chip with high performance and manufacturability for CFP2 modules. W1I.6, proceedings optical fiber communications conference (2014)
86. R. Stabile, A. Albores-Mejia, A. Rohit, K.A. Williams, Integrated optical switch matrices for packet data networks, Microsys. Nanoeng. Rev. Art. **2**, 15042 (2016)
87. I.M. Soganci, T. Tanemura, K.A. Williams, N. Calabretta, T. de Vries, E. Smalbrugge, M.K. Smit, H.J.S. Dorren, Y. Nakano, Monolithically integrated InP 1 × 16 optical switch with wavelength-insensitive operation. Photon. Technol. Lett. **22**(3), 143–145 (2010)
88. I.M. Soganci, T. Tanemura, Y. Nakano, Integrated phased-array switches for large-scale photonic routing on chip. Laser Photon. Rev. **6**, 549–563 (2012)
89. T. Tanemura, I.M. Soganci, T. Oyama, T. Ohyama, S. Mino, K.A. Williams, N. Calabretta, H.J.S. Dorren, Y. Nakano, Large-capacity compact optical buffer based on InP integrated phased-array switch and coiled fiber delay lines. J. Lightwave Technol. **29**(4), 396–402 (2011)
90. M.J. Kwack, T. Tanemura, A. Higo, Y. Nakano, Monolithic InP strictly non-blocking 8×8 switch for high-speed WDM optical interconnection. Opt. Express **20**(27), 28734 (2012)

91. Q. Cheng, A. Wonfor, R.V. Penty, I.H. White, Scalable, low-energy hybrid photonic space switch. J. Lightwave Technol. **31**(18), 3077–3084 (2013)
92. H. Wang, A. Wonfor, K.A Williams, R.V. Penty and I.H. White, Demonstration of a lossless monolithic 16 × 16 QW SOA switch. Post-deadline paper, proceedings European conference on optical communications (2009)
93. R. Stabile, N. Calabretta, K.A. Williams, Switch-filter wavelength selector: Simulation and experiment. IET Optoelectron. **8**(1), 1–10 (2014)
94. R. Stabile, P. DasMahapatra, K.A. Williams, 4×4 InP switch matrix with electro-optically actuated higher order micro-ring resonators. IEEE Photon. Technol. Lett. Accepted for publication (2016)
95. R. Stabile, N. Calabretta, K.A. Williams, H.J.S. Dorren, Monolithic 16-wavelength selector based on a chain of passband-flattened cyclic AWGs and optical switches. Opt. Lett. **40**(8), 1795–1797 (2015)
96. N. Calabretta, K.A. Williams, H.J.S. Dorren, Monolithically integrated WDM cross-connect switch for nanoseconds wavelength, space, and time switching. Proceedings European conference on optical communications, ID: 0296 (2015)
97. Q. Cheng, R. Stabile, A. Rohit, A. Wonfor, R.V. Penty, I.H. White, K.A. Williams, First demonstration of automated control and assessment of a dynamically reconfigured monolithic 8×8 wavelength-and-space switch. J. Opt. Comm. Networking **7**(3), A388–A395 (2015)
98. M.K. Smit, J. van der Tol, M.T. Hill, Moore's law in photonics. Laser Photon. Rev. **6**, 1–13 (2012)
99. J.J.G.M. van der Tol, R. Zhang, J. Pello, F. Bordas, G.C. Roelkens, H.P.M.M. Ambrosius, P.J.A. Thijs, F. Karouta, M.K. Smit, Photonic integration in indium-phosphide membranes on silicon. IET Optoelectron. **5**(5), 218–225 (2011)
100. wipe. jeppix.eu
101. phastflex. jeppix.eu

Part IV
Prospects and Future Trends

Chapter 15
The Future of Switching in Data Centers

Slavisa Aleksic and Matteo Fiorani

15.1 Introduction

The internal interconnection network of a data center is usually limited by the maximum data rate per link and per cable, the required number of links, and the maximum length of a single interconnection link. This is due to the fact that current intra-data center interconnects mostly use a mix of electronic backplanes, copper cables, and optical fibers, the latter mainly based on multimode fibers interconnecting modules placed in different chassis and racks. In fact, very high data rates over long electronic backplane traces are hardly achievable due to the associated high signal losses and inter-symbol interference (ISI). The maximum switching capacity and the number of switches additionally limit the achievable performance of switched interconnects.

On the other hand, processing electronics show continuous advance in computational bandwidth as well as in reduction of its feature size, thus providing more functionality and higher speed on cards within modules, i.e., system boards. This implies the need for more point-to-point interconnects on boards and between boards, in which denser packing is, however, limited by the crosstalk. Since data centers have been experiencing a heavy increase in the amount of traffic to store and process, optical cables have already found their application in interconnecting racks of equipment within data centers and high-performance computer clusters. Due to the fact that both optical transmission and switching technologies are generally able

S. Aleksic (✉)
Institute of Communications Engineering, Hochschule für Telekommunikation
Leipzig (HfTL), Gustav-Freytag-Str. 43-45, 04277, Leipzig, Germany
e-mail: aleksic@hft.leipzig.de

M. Fiorani
Optical Networks Lab (ONLab), KTH Royal Institute of Technology,
Electrum 229, SE164 40, Kista, Stockholm, Sweden

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
DOI 10.1007/978-3-319-61052-8_15

to provide higher data rates over longer transmission distances and faster switching operation than electrical transmission and switching systems, a natural answer to the scalability problem could be to use optical transmission and switching technologies, i.e., optically switched interconnects, in order to relax the limitations and improve the scalability of internal interconnecting system.

Figure 15.1 represents some recent trends in performance and power consumption of high-performance computers (HPCs) [1]. It is evident that already today, large HPC systems consume almost up to 20 MW of electricity. In the figure, two

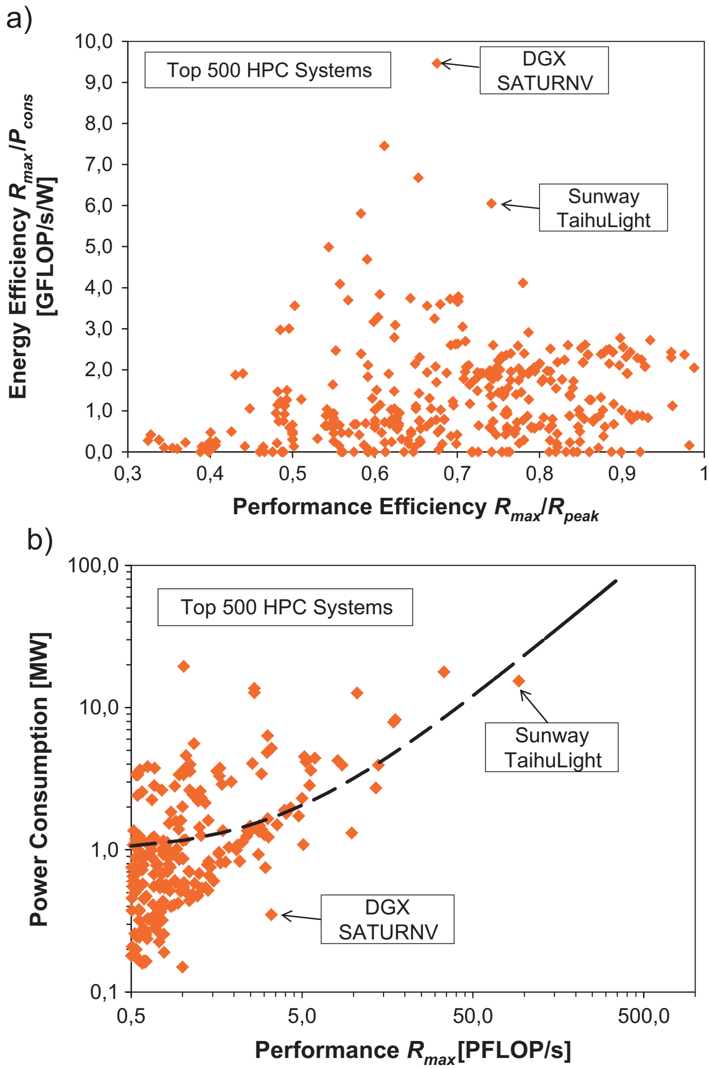


Fig. 15.1 Trends and projections in performance and power consumption of high-performance computers according to the data taken from [2]: (**a**) energy efficiency vs. performance efficiency and (**b**) power consumption vs. computing performance

examples of recent high-performance computers are indicated. One of them is the current most powerful HPC system Sunway TaihuLight, which reaches the maximum computing performance of 93 PFLOPS while consuming more than 15.3 MW of power. This leads to an energy efficiency of 6.05 GFLOPS/s/W. The second example is the most energy-efficient system called DGX SATURNV, providing about 9.46 GFLOPS/s/W and a maximum computing performance of 3.3 PFLOPs.

Even though a lot of effort has been made in the last years to increase the energy efficiency of computing and networking equipment, a further growth of HPC systems and data centers will require additional technological development steps to further increase both the processing speed and the number of cores/servers/nodes, which will unavoidably lead to a higher system complexity and an increased power consumption. Thus, future Exascale computer systems will probably reach a very high level of complexity and are expected to consume even more energy than the current systems, which will set very high requirements on the internal interconnection network as well as power supply and cooling. Therefore, a large attention has to be paid to the research and development of more scalable and energy-efficient structures and technologies to make possible further scaling in both capacity and performance.

15.2 Design Considerations for Advanced Optical Interconnects

Several studies have shown strong potential for relaxing the energy and volume issues through replacing electrical lines by optical interconnects. Above a certain length, the break-even length, optical interconnects consume less energy than electrical ones. The break-even length differs from case to case and has been estimated to be between 43 cm [2] and 50 μm [3]. Other potential benefits of optical interconnections lie in the achievable high interconnection density and signal integrity. Thus, deeper penetration of optics into data centers could provide benefits regarding scalability and power consumption. Additionally, the future viability of optical interconnects also depends on a reduction of costs per input/output port as well as on the achievable performance of the interconnection network. It is thus important to consider all the different factors when examining new concepts for highly scalable and efficient optical interconnects.

Distance- and frequency-dependent attenuation as well as high crosstalk set limits on achievable data rates over copper-based cables and PCB traces. This is the main reason, while recent data centers, supercomputers, and high-capacity routers are increasingly relying on optical point-to-point interconnection links. Standard optical point-to-point links are usually based on directly modulated vertical cavity surface emitting lasers (VCSEL) and multimode fibers (MMFs). However, the capacity of such interconnects is limited by both modulation bandwidth of the laser and the intermodal dispersion in MMFs. Recently, various methods to enable transmission at higher data rates have been proposed and investigated, such as (1) to increase the modulation rate of the laser [4], (2) to use multiple fibers and transceivers in a parallel manner [5], and (3) to employ advanced modulation formats and multiplexing techniques such as optical orthogonal frequency-division multiplexing (OOFDM) and high-order modulation [6].

When considering future requirements, an architecture based on point-to-point, single-channel links will not only lead to poor scalability and large latency but will also cause low power efficiency and high implementation costs. On the other hand, optically switched interconnects that make use of optical switches and wavelength-division multiplexing (WDM) technology can benefit from inherent parallelism and optical transparency. Several realizations of optically switched interconnects based on different interconnecting arrangements and optical switching devices have been already proposed and analyzed in the literature [7–10]. The optically switched interconnects proposed in recent technical literature can be categorized according to the utilized switching technology in hybrid optical/electronic interconnects, optical circuit-switched interconnects, and optical packet-switched interconnects. It has been shown that optically switched interconnects have the potential to achieve high performance and high energy efficiency [11]. However, their future viability also depends on further improvements in scalability and a reduction of the cost per input/output port.

It is crucial to recognize that large internal interconnection systems are much more than point-to-point transmission links. Indeed, additional to the large number of transceivers and links, they also comprise elements that implement different other functions such as synchronization, switching, switch control, scheduling, arbitration, signaling, as well as managing and routing of data units through the internal interconnection network. All these additional elements contribute to an increased complexity and higher energy consumption of the interconnection system. For example, while a single point-to-point, fiber-based interconnection link comprising an optical transmitter and an optical receiver can be realized using the state-of-the-art technology to consume as low as several pJ/bit, the entire internal interconnection network implementing all the above-listed functions usually consumes about two orders of magnitude more energy per bit, thus reaching the level of nJ/bit. Also the topology of the interconnection network influences the achievable performance and efficiency of the entire system. Therefore, new concepts for interconnection systems should be examined by considering, additional to the transmission properties of point-to-point links, also various other technological aspects and interconnection arrangements under all performance metrics, namely, scalability/feasibility, traffic-related performance, power consumption, and techno-economics, as depicted in Fig. 15.2.

The internal interconnections within large data centers, supercomputers, or routers are usually classified into the following four groups, which can be seen as different hierarchical system levels.

- The highest hierarchical level represents the *rack-to-rack* interconnection network, whose link lengths can range from a few meters to several hundreds of meters.

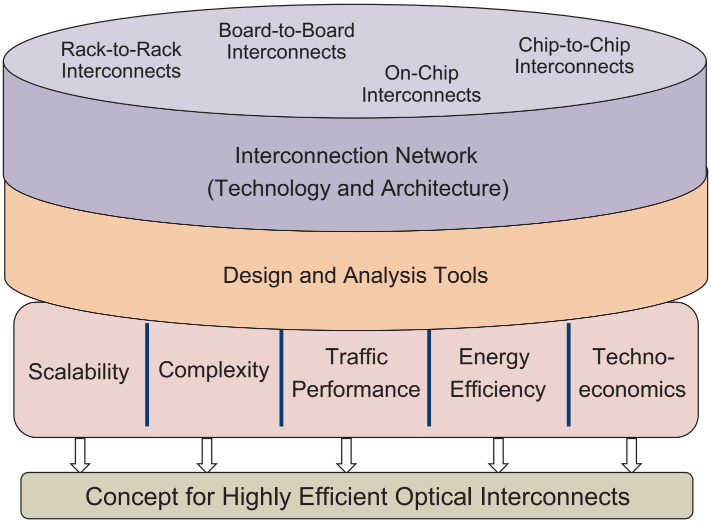

Rack-to-Rack Interconnects
Board-to-Board Interconnects
On-Chip Interconnects
Chip-to-Chip Interconnects
Interconnection Network (Technology and Architecture)
Design and Analysis Tools
Scalability
Complexity
Traffic Performance
Energy Efficiency
Techno-economics
Concept for Highly Efficient Optical Interconnects

Fig. 15.2 Conception and evaluation of highly efficient and scalable optical interconnects for large-scale systems

- Within a rack of equipment, various realizations of backplanes are possible. The length of *intra-rack* links are typically between 15 cm and a few meters.
- The *chip-to-chip* interconnects are providing connections between chips in a single module, i.e., on a board, which are typically shorter than 25 cm.
- Finally, *on-chip* interconnects have usually a length below two centimeters.

Interconnects at various scales can make use of different architectures and technologies since system parameters and design goals can differ significantly. Additional to the transmission distance, also the number of nodes and the number of hops, data rates, and transmission characteristics depend on the hierarchical level and the location within the system. However, even though the hierarchical system levels are often designed and analyzed separately, there is a need for an analysis and optimization at the system level since the overall performance depends not only on the performance level of the individual subsystems but also on the intensity of the interaction between different hierarchical levels.

15.3 Switch Architecture and Network Topology

There are various interconnecting arrangements and thus various architectures of internal switching fabrics that have been used to implement interconnection networks in large-scale systems. Table 15.1 summarizes the most important architecture types and shows their blocking characteristics.

Table 15.1 Some often used architectures for optical switching elements

Switch architectures	Blocking type
Classical logN	Blocking
Benes	Rearrangeably non-blocking
Crossbar	Wide-sense non-blocking
Spanke	Strict-sense non-blocking
Cantor	Strict-sense non-blocking
Banyan	Blocking

Table 15.2 Selected network topologies for internal interconnection networks in large data processing and switching systems

Network topologies	Blocking type
Clos (fat-tree, multistage)	Strict-sense, non-blocking if $p \geq 2n$-1
d-dim symmetric mesh	Rearrangeable, blocking if $p > 2$
d-dim symmetric torus	Rearrangeable, blocking if $p > 2$
d-dim hypercube	Rearrangeable

p is the number of edge switches
n is the number of ports of a single switching element

All architectures have several important characteristics such as the number of stages, the number of feasible connections, and the type of switching elements used to construct a large fabric. Multistage interconnection network is an important class of interconnection arrangements that consists of multiple stages with a number of switching units in each stage. Some selected topologies of multistage interconnection networks that have been often used to implement internal interconnection networks in large data processing and switching systems are listed in Table 15.2. The topologies can be divided into generally blocking, wide-sense non-blocking, rearrangeably non-blocking, and strict-sense non-blocking networks. The interconnecting arrangements can be classified in different ways, e.g., regarding its blocking probability, packet loss probability, the number of stages, or with respect to number of possible paths through the switching fabric. The selection of the topology of the internal interconnection network can have a significant impact on the overall performance of the system. However, it is not possible to design a single interconnection topology that provides best performance for all applications. For example, the fat-tree topology (the Clos network) can be strictly non-blocking if the number of edge switches is about two times larger than the number of ports of a single switching element, i.e., if $p \geq 2n - 1$. Even though this architecture is able to provide a very good performance with respect to bandwidth and latency and has already been used in many internal interconnection networks, the relatively high cost of the entire network due to the large number of high-speed ports limits its scalability [12]. On the other hand, the multidimensional mesh and torus topologies are typically not able to provide strict-sense non-blocking operation, but usually lead to cost-effective implementation at large scales. Especially in applications with locality, which is often the case with applications running on high-performance computers, these topologies can provide better performance than the Clos network. Therefore, d-dimensional mesh and torus networks have been often used in recent implementation of supercomputers. An example is the TOFU interconnect, which has been developed for the K computer [13] and is the Cartesian product of three-dimensional mesh and three-dimensional mesh/torus networks resulting in an overall topology of a six-dimensional mesh/torus.

An important issue that limits both the scalability and the manageability of large-scale systems is the very large number of required interconnection links, which results in a large number of cables. This problem is often referred to as the wiring problem. Thus, the number of required links is an important parameter for the selection of a suitable network topology. Figure 15.3 shows a comparison of several network topologies such as full mesh, Clos, two-dimensional and three-dimensional torus, and TOFU with regard to the required number of interconnection links. As expected, the fully mesh topology is not scalable at all since it would require billions of links to connect several tens of thousands of nodes. The multistage and multidimensional topologies provide much better scalability. Even though the

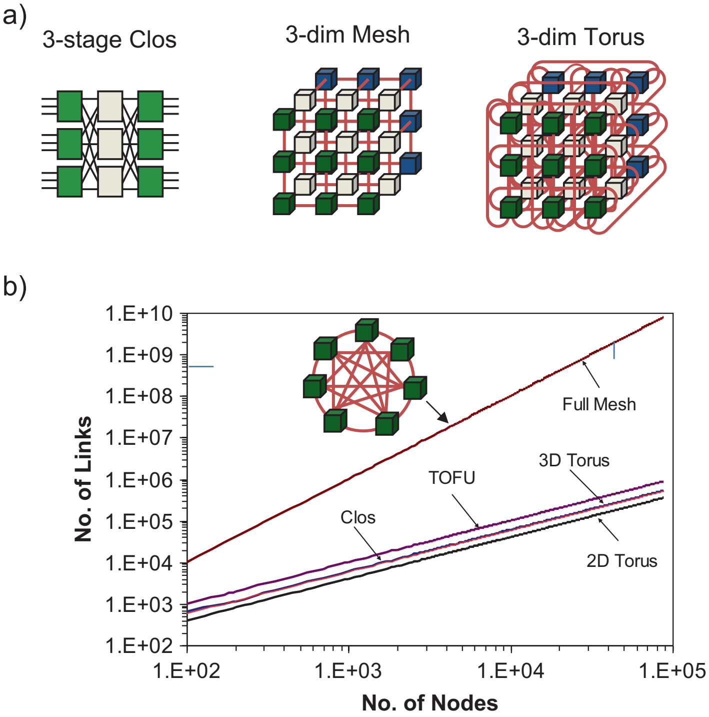


Fig. 15.3 (**a**) Examples of network topologies: three-stage Clos, three-dimensional mesh, and three-dimensional torus. (**b**) Scalability of five exemplary network topologies with respect to the required number of links

required number of links in a full mesh topology can be reduced by several orders of magnitude when using a more scalable network topology, the wiring still remains important. For example, the TOFU interconnect provides a very good scalability, but still requires about 960,000 links in total to support 80,000 nodes. A further reduction of the total number of cables can be achieved through transmitting several wavelength channels over the same optical fiber by utilizing the wavelength-division multiplexing (WDM) technique [10].

15.4 Technology Trends

This section briefly reviews recent trends in research and development of technologies for optical interconnects at different hierarchical system levels.

15.4.1 On-Chip Optical Interconnects

As the number of cores in a single processor chip continuously increases, the requirements on the on-chip interconnection network increase, too. Already today, consumer CPUs comprise up to ten cores, and novel architectures for high-end processors having more than1000 cores have already been designed [14, 15]. It is well known that the transistor performance improves with geometric scaling, which cannot be said for interconnects. This fact will play an important role in the future generations of processors because on-chip interconnects will become an important limiting factor in increasing the overall performance and energy efficiency. The International Technology Roadmap for Semiconductors (ITRS) has indicated in its recent report that this trend may enforce Cu extensions, replacements, and native interconnects [16]. The options for Cu replacements include nanowires, carbon nanotubes, graphene nanoribbons, and optical intra- and inter-chip interconnects.

Most of the recent proposals for optical on-chip communication systems use silicon photonics in combination with silicon nitride and silicon oxide. Silicon photonics has the potential of enabling implementation of cost-effective and high-speed communication links and optical networks on chip (NoC). Different implementation options have been investigated, such as the monolithic integration using either the front-end-of-line (FEOL) or back-end-of-line (BEOL) process [17] or three-dimensional chip stacking [18]. The systems proposed so far include both single and multiwavelength operation and make use of different network topologies. The network topology significantly influences both the performance and reliability of the network-on-chip (NoC) and determines its footprint. The topologies proposed and used so far for optical network-on-chip (NoC) include regular and irregular ones. Regular topologies such as mesh, ring, crossbar, torus, cube, and tree have been extensively investigated for its appropriateness to optical on-chip interconnects [19, 20]. As we have already seen in the previous section, the direct regular topologies such as torus and mesh offer superior scalability. The tree topology, which is an indirect regular network, offers better hardware efficiency than the direct topologies. Irregular topologies are asymmetric and scales nonlinearly, leading to higher requirements on areas and energy budget [21]. The irregularity makes this type of networks better suitable for heterogeneous applications with asymmetric traffic and varying communication requirements between the nodes.

While the advantages of optical interconnects when compared to electrical interconnects are clearly obvious at the system level, i.e., between modules, shelves, and racks, it is not yet sure whether optical technology will be able to provide significant benefits at the chip level. Even though power-hungry trans-impedance amplifiers can be avoided in the 22-nm technology node due to the small enough transistor capacitance and the possibility to drive transistors directly by photodiodes, it is still not obvious whether the advantages of optical interconnects are strong enough to enable a fast penetration of optical technologies on the chip level, especially when considering the additional power needed for thermal tuning and laser drivers. In fact, research on electrical interconnects has also made great progress in the last years, and looking along the technology road map, optical interconnects are expected to outperform their electrical counterparts first at the 8-nm technology node [22]. However, in the long term, it is broadly agreed that nanoscale photonics will play a significant role in enabling further scaling of multiprocessor architectures by offering an improvement of the performance-per-watt metric in next-generation high-performance chips [16, 23].

15.4.2 On-Board Optical Interconnects

The processing power of integrated circuits has been constantly increasing during the last three decades. The requirements on interconnects between chips and modules have increased, too. According to the recent projections of the International Roadmap for Semiconductors (ITRS), single-processor chips with 100 TFLOPS can be expected in 2020. The interconnections between processing nodes should be able to provide capacities of more than 200 Tbit/s and an energy efficiency significantly below 1 pJ per bit. All these requirements can hardly be met by electrical interconnection technologies, while optical interconnects have the potential to provide both high bandwidth density and high energy efficiency. Electrical interconnects suffer from distance- and frequency-dependent attenuation due to two kinds of losses coupled with high frequencies: (i) dielectric loss in PCB substrates and (ii) skin effect in coaxial cables. Optical interconnects exhibit no frequency-dependent loss. Thus, optical interconnects are capable of overcoming most of the physical limitations associated with electronic interconnects regarding interconnection density, timing, signal integrity, crosstalk, and energy consumption. In the following, we present and discuss optical interconnects on PCB boards on the example of an innovative method based on two-photon absorption (TPA). This method can be used to rapidly write multi-core optical waveguides within a polymer material, which can be coated on any standard printed circuit board.

Table 15.3 Fabrication methods for optical interconnects on PCBs

Embossing	Photolithography (UV)
Pros	Pros
Well-known processes	Well-known processes
High refractive index difference realizable	High refractive index difference realizable
Cons	Cons
Require many different cost-intensive steps	Require many different cost-intensive steps
Low precision	Alignment difficult
Expensive	Wet chemical process
	Expensive

Hybrid approach	Two-photon absorption (TPA)
Pros	Pros
Optical fibers (or waveguides) are integrated on PCB	A few production steps
Commercially available	Rapid prototyping possible
High waveguide quality	Optoelectronic components directly mounted on PCB
	Simplified alignment
Cons	Cons
Alignment difficult	Low refractive index difference
Complex procedures	New technology
Expensive	

Optical interconnects on printed circuit boards (PCBs) allow denser waveguides with potentially lower energy consumption per transmitted bit compared to pure electrical interconnects [3]. Further advantages of optical PCBs such as the robustness against electromagnetic interference and the galvanic isolation make them a promising alternative to microstrip lines [24]. In supercomputers and data centers, reliability becomes an important requirement as the number of interconnection links increases. However, environmental stresses on PCB influence the structural integrity and functional parameters of embedded polymer waveguides, which, in turn, impair their reliability. Stress factors influence mostly the refractive index and optical transmissivity. For example, isothermal annealing can reduce the refractive index of optical waveguides [25]. Thus, it is important to develop optical interconnects that satisfy the requirement on high reliability [26].

There have been a number of fabrication processes proposed and used to integrate optical waveguides on printed circuit boards (PCBs). Table 15.3 provides a comparison between four fabrication processes. Most of recently demonstrated optical PCBs are produced using the well-known photolithographic methods for waveguide production [27], but also embossing technologies are used for structuring the optical waveguides [28]. In the hybrid approach, optical waveguides (mostly fibers) are integrated directly on PCB, which allows achieving a high waveguide quality. This fabrication method relies on mature technology and has already reached a certain level of commercialization. However, the process for fabricating the hybrid optical PCBs is relatively complex and the alignment is difficult, which increases the end product price and limits its practicability. Two-photon absorption (TPA) is a relatively new, promising method for rapid prototyping of optical PCBs. It makes possible mounting of optoelectronic components directly on PCB and enables a simplified alignment. Waveguide fabrication using the two-photon absorption (TPA) process allows direct writing of optical waveguides on PCB boards within a few fabrication steps [29–31]. The main advantage of TPA is that it allows both production of waveguides with arbitrary shapes and alignment in one single fabrication step. On the other hand, no industrial TPA waveguide writing site is available so far, which hinders a fast development and fabrication as well as a quick market penetration.

15.4.3 System-Level Interconnection Network

More functionality and processing power on boards lead to a need for high-capacity and high-performance interconnects between boards. Electrical interconnects are more vulnerable to crosstalk, and the need for higher bandwidth is usually obstructed by increased dielectric losses. In order to overcome this electronic bottleneck and meet the growing demand for both low latencies and high bandwidth, various interest groups and standardization bodies have developed enhanced technologies for printed circuit boards as well as standards and protocols with improved signal integrity specifications, line coding formats, and design techniques such as preemphasis and equalization. High-speed electrical signaling has also experienced considerable enhancements. These technologies are either already in use or in the process of being adopted by system and chip vendors.

This section gives an overview on technologies and architectures for system-level interconnects (rack-to-rack) and addresses their scalability limitations and energy consumption.

15.5 Point-to-Point Interconnects

Each interconnection technology is upper bounded regarding its transmission distance, channel data rate, and number of assigned channels [32]. Some of the technologies such as CEI-6G, CEI-11G, and sRIO (Serial Rapid I/O) are only intended for electrical backplane applications, while others such as those based on Ethernet and InfiniBand (IB) also support board-to-board and rack-to-rack applications through using optical point-to-point interconnects. A lot of effort has been made by standardization bodies such as IEEE 802.3ba Ethernet Task Force, Optical Internetworking Forum (OIF), Fiber Channel (FC), InfiniBand (IB), Rapid IO Trade Association, and PCI Express to achieve an improvement of existing interconnecting technologies regarding data rate and efficiency. The IEEE 802.3ba Ethernet Task Force has already standardized 40 Gbit/s (40GbE) and 100 Gbit/s Ethernet (100GbE). The IEEE 802.3bs is also working on the definition of a standard for 200 Gbit/s (200GbE) and 400 Gbit/s Ethernet (400GbE) that is expected to be ready in 2018 [33]. The Physical and Link Layer (PLL) Working Group of the OIF has developed the physical specifications CEI-25G and CEI-28G for achieving lane signaling rate of up to 28 Gbaud/s. They are intended for next-generation chip-to-chip and chip-to-module interconnects as well as for backplane applications that support transmission up to 100 Gbit/s. InfiniBand has introduced additionally to single, double, and quadruple data rate (SDR, DDR, and QDR) and also enhanced data rate (EDR) systems with 20 Gbit/s per lane. In addition, a 200 Gbit/s InfiniBand hardware has been recently introduced [34], which claims to be the first 200 Gbit/s data center interconnect. This solution includes ConnectX-6 adaptors with 200 Gbit/s, Quantum 200 Gbit/s HDR InfiniBand switch, as well as copper and optical cables capable of supporting the data rate up to 200 Gbit/s. The switches support up to 40 or 80 ports with 90 ns of latency. Similarly, PCI Express 3.0 supports row data transfer rates of 8 Gbit/s/lane and up to 32 lanes. The PCI-SIG group is currently working on the specifications for PCI Express 4.0 with improved data rates of up to 16 Gbit/s/lane and a maximum of 265 Gbit/s over 16 lanes. Rapid IO focuses on higher data rates, which reached 25 Gbit/s/lane and 100 Gbit/s per port in the Rapid IO 4.0 specification. Although a lot of progress in signal processing and modulation formats has been made to realize and standardize high-data-rate electrical and optical interconnects, the most of the effort has been put into design and characterization of simple point-to-point links, while switching is done electronically. On the other hand, optically switched interconnects are able to provide both transmission and switching functionalities directly in the optical domain. In the following section, we will discuss the technologies for optically switched interconnects from both device and system perspective.

15.6 Optically Switched Interconnects

Combining optical transmission and optical switching in an optimal way to realize high-capacity optically switched interconnects is a promising approach for high-performance systems. In the following subsection, we very briefly review different optical switching devices that can be used as an alternative to widely used electronic switches.

Optical switching technologies can be classified based upon the underlying physical effect used for the switching process into: electro-optic (EO), acousto-optic (AO), thermo-optic (TO), and opto-mechanical (OM) switching. The EO, AO, and TO effects rely on refractive index changes of the matter through application of an external physical field or action, while in OM switches, optical beams are reflected by electromechanical means.

In the switches utilizing the EO effect, an applied electrical field induces the change in the index of refraction, which then channels the light to the appropriate port.

Lithium niobate ($LiNbO_3$) is a unique crystal that shows large EO effect, AO effect, TO effect, and nonlinear effects. EO devices based on this substrate have very fast response and small dielectric constant. Another EO switch group comprises switches based on liquid crystals, which exhibit high extinction ratio, high reliability, and low power consumption. Also semiconductor optical amplifiers (SOAs) can be used as an ON-OFF switch by varying its bias current. By applying a reduced bias voltage, no population inversion is achieved, and the device rather absorbs the input signal, thereby building the off-state. In contrary, if a sufficiently high bias voltage is applied, the input signal will be amplified, and, thus, the on-state is achieved. In general, EO switches suffer from high insertion loss and polarization-dependent loss (PDL) . PDL can be combat at the cost of higher driving voltage, and consequently lower switching speed, which is not desirable. The switching speed of EO switches usually lies in the order of several nanoseconds or even hundreds of picoseconds, which is sufficiently fast for most applications including optical packet switching.

The changes of refractive index due to the interaction between acoustic and optical waves in the crystal are utilized in the AO switches. The switching speed of AO switches is in the order of hundreds of nanoseconds and is limited by the propagation speed of acoustic waves. AO switches can also be implemented on lithium niobate.

The TO effect utilizes the temperature dependence of the refractive index. The advantage of thermo-optical switches is its generally small size, but the high driving voltage and high power dissipation make such switches highly impractical. Crosstalk and insertion loss values are also not very satisfactory. Mostly used materials for implementing TO switching elements are silica and polymers. Switching time of TO switches lies in the order of milliseconds, thereby making this type of switches less suitable for applications requiring dynamic switching.

Opto-mechanical (OM) switches are based on mechanics and free-space optics. Switching is performed by electromechanical means such as by moving mirrors or directional couplers. Regarding its optical performance parameters, OM switches provide low insertion loss, low polarization-dependent loss (PDL), and low crosstalk. However, drawbacks of this type of switches are their relatively low switching speed in the order of a few milliseconds, which could be unacceptable for some applications requiring fast optical switching. The micro-electromechanical system (MEMS) switches form a subcategory of the OM switches. In particular, 3-D MEMS is the most promising option for applications that do not require fast switching, but instead large port counts. 3-D MEMS switches with more than 1000 ports have already been demonstrated [35]. MEMS devices are scalable and cascadable and consume low power. Challenges regarding MEMS are packaging and time-consuming fabrication.

Arrayed waveguide grating (AWG)-based switches have gained a particular attention for implementation in large-scale switching fabrics. There are several architectures that base upon these particular elements. Since they are passive elements, they can potentially provide low-power operation. However, additional active elements such as wavelength converters (WCs) are needed to implement switching operation. Switching time of AWG-based switches is determined by the tuning speed of the wavelength converter.

According to its switching time, insertion loss, crosstalk, and PDL, a specific device type can be more or less suitable for a particular application. For example, the switching time required for fast packet-switching applications should be small in comparison to the average length of data packets. For a more complete overview of different options for implementing optical switches, the reader is referred to [31, 32].

15.7 Enabling Technologies for Next-Generation System-Level Optical Interconnection Networks

Recent experience and practices in designing and managing large-scale data centers and the growing need for more capacity and higher performance of internal interconnection networks have led to an increased interest in adopting advanced optical technologies to internal data center networks. Similarly, the methods and technologies that have been successfully used in large data centers such as virtualization and consolidation have started to influence the communication network. On the one hand, the optical technologies that have been designed for core and access networks have the potential to significantly improve the performance and scalability of data center interconnects, while on the other, advanced cloud applications and services set high requirements on communication networks, which have to respond with a more dynamic and flexible operation. Some of examples of these trends are the increased use of advanced optical technologies within data centers and recent efforts in improving the flexibility of communication networks by developing, standardizing, and implementing network function virtualization (NFV) and software-defined networking (SDN). In this context, scalability, adaptability, and energy efficiency play an important role for both data centers and communication networks. In the following, we will address the recent research efforts that have been put into the development of new components, methods, and systems for increasing capacity and performance of optical networks, while making the optical infrastructure more flexible and energy efficient [36–46]. These advanced technologies have the potential to revolutionize both the intra- and inter-data center networks and to enable an optimal support of future cloud applications and services.

15.7.1 High-Capacity Optical Links

The capacity limit of optical transmission systems has already approached close to the nonlinear Shannon limit thanks to using and optimally combining different multiplexing formats such as wavelength-division (WDM), optical time-division (OTDM), and polarization-division (PDM) multiplexing together with advanced multilevel modulation formats that exploit intensity and phase modulation of the optical carrier [38, 39]. The spatial-division multiplexing (SDM) technology with optical fibers supporting multiple spatial elements (e.g., multimode, multicore, or multielement fibers) has been recently proposed as a solution to overcome the non-linear Shannon limit and to increase the traffic that can be carried over a single fiber [40]. Recently, transmission of 101.7 Tbit/s with a spectral efficiency as high as 11 bit/sec/Hz over 3x35 km of standard single-mode fiber (SSMF) has been successfully demonstrated [43]. Aggregate data rates in the range of Pbit/s are possible by exploiting SDM in specially designed multicore fibers [44] as shown in Fig. 15.4, which summarizes recent experimental demonstrations of optical high-capacity transmission systems.

Although the presented achievements are really impressive and could theoretically solve some issues regarding the capacity shortage in intra- and inter-data center networks, it is unrealistic to expect that these technologies can soon be adopted for data centers because they are still too complex and expensive. However, multilevel modulation and wavelength-division multiplexing are practical enough to be used for implementing the next-generation data center interconnection networks. Indeed, different options for combining flex-grid WDM and SDM have been recently studied and utilized to design efficient optical switching solutions for large data center interconnects [41, 42].

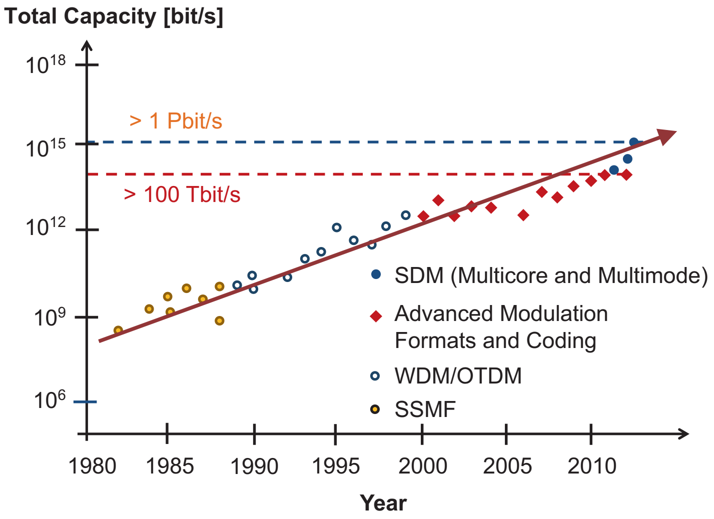


Fig. 15.4 Increase of optical link capacity. SDM, spatial-division multiplexing; *SSMF* standard single-mode fiber, *WDM* wavelength-division multiplexing, *OTDM* optical time-division multiplexing

15.7.2 Bandwidth-Variable and Software-Controllable Optical Transceivers

Bandwidth-variable and software-controllable optical transceivers (BVSCT) are considered to be a key enabling component for future optical transport networks, but can also be used in data center interconnects. It is very probable that BVSCTs will operate on flexible-wavelength grid with 12.5 GHz spectral separation and 6.25 GHz granularity for center frequencies. They will allow optimally accommodating traffic needs by flexibly varying bit rate, reach, and spectral efficiency. New-generation optical coherent transceivers with digital signal processing already provide a high level of adaptability to support trade-offs between bit rate, spectral efficiency, and reach. They can provide different modulation formats such as binary phase-shift keying (BPSK), quadrature phase-shift keying (QPSK), and quadrature amplitude modulation (QAM) together with forward error correction (FEC) [45]. The migration scenarios toward fully flexible and elastic optical data center networks will be influenced by the capabilities and the cost of BVSCTs, which in turn largely depends on the choice of architecture and required features. In addition, there are still challenges for defining efficient techniques for performing traffic grooming and spectral resource allocation in flex-grid WDM optical networks.

15.7.3 Dynamic and Flexible Optical Switching Nodes

It could be very beneficial, if switching nodes for the next-generation optical data center networks would be capable of providing a high level of flexibility in various domains such as wavelength, space, and time as well as to support elastic switching over a flexible wavelength grid. Other important requisites are adaptability, scalability, and resilience. Thus, there is a need for node architectures that allow flexibility and adaptability through a reconfigurable and on-demand structure [46]. Various components such as flex-grid (FG) wavelength selective switches (WSSs) , multiplexers/demultiplexers, optical amplifiers, fast optical and/or electronic switches, and transponders can be a part of such a flexible architecture that allow different configurations to be created by interconnecting functional components to best cope with changing traffic demands (see Fig. 15.5). Additionally, such a modular architecture will enable scalability and an easy extension of the node functionality through adding new functional modules or replacing the old ones in order to optimally support future services. Also redundancy and protection switching can easily be used for critical functions, leading to improved resilience. Inactive modules can be switched off, thereby reducing the energy consumption and increasing energy efficiency.

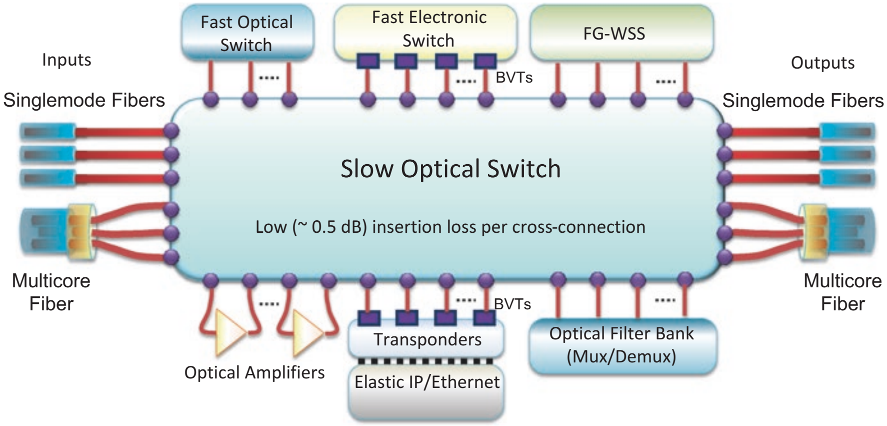


Fig. 15.5 Adaptive node architecture. FG-WSS, flex-grid wavelength selective switch

15.7.4 Energy-Efficient Communication Systems and Networks

Energy efficiency considerations have gained in importance in recent years due to the ever-increasing energy consumption of data centers. Even if the contribution of the networking equipment to the total energy consumption of a data center has been estimated to about 8–10% [47, 48], it is still a large amount of energy consumed, especially when taking into account the growing energy consumption of data centers and communication networks. Therefore, it is important to carefully address technologies and methods for increasing energy efficiency of the interconnection network within the data center.

15.7.5 Multilayer Software-Defined Networking

Software-defined networking (SDN) is a framework to support the programmability and an efficient control of network functions and protocols over several layers as well as to decouple the data plane from the control plane. SDN allows an abstraction of the underlying infrastructure, which could then be used by applications and network services as virtual entities. It makes possible to define and manipulate multiple coexisting virtual network slices in a way that is independent of the underlying transport technology and network protocols. SDN can help in achieving an efficient multilayer and multi-domain transport, reducing provisioning latencies and implementing real-time constraint-based routing. Through SDN and elastic optical networking, application aware and on-demand resource provisioning could become reality, which could help cloud service providers to better utilize their infrastructure and customize it in a dynamic manner and according to the needs of applications and users. Additionally, it could enable a consolidation of different switching technologies in a single dynamic and flexible data center network.

Already for many years now, multilayer integration has been a wish of networking industry. Since multilayer SDN provides centralized network intelligence, it makes possible to inspect all network layers concurrently to determine a path and transport technology best suited to carry traffic. Even on a single path, a data flow can be transported using different technologies and over several layers. Additionally, multilayer SDN could monitor and evaluate the performance at each layer and across several network areas and dynamically reroute traffic or add some bandwidth from a lower layer to avoid congestions and find an optimal solution in milliseconds. This can avoid the need for hold-down timers, which are provisioned waiting periods defined and used by upper layers to provide enough time for lower layers to react to failures. Multilayer SDN can open the way for dynamic network optimization as well as for automated congestion control and cost management. The SDN network control plane can also be connected to a higher layer orchestrator that harmonizes and optimizes the allocation of heterogeneous types of resources in the data center (i.e., network, cloud, and storage). The orchestrator allows to rapidly set up, configure, and manage services that span across different technology domains.

Current SDN implementations focus mainly on Ethernet networks. It is essential to extend and apply the SDN concept to optical interconnection networks on layer 0/1, where there is currently a lack of standards and products providing automated provisioning across these layers. Modern optical network elements are already capable of providing a relatively high level of flexibility and controllable attributes. Some of the attributes can be controlled by software, so a SDN controller can control them. Topology management and virtual routing modules are also available. However, additional standardization is required to allow the SDN controllers to directly manage optical transmission components such as variable bandwidth transceivers (VBTs) and reconfigurable add/drop multiplexers (ROADMs). Within the network, the available spectrum can be flexibly handled by allocating one, two, or more spectral slices to a data flow. Some realizations of ROADMs are very flexible and allow the control of the wavelength (color), ingress/egress direction, and wavelength reuse without restrictions [49–52]. Such ROADMs are called colorless, directionless, and contentionless (CDC) add/drop multiplexers. Figure 15.6 shows a few examples of components that could be used to implement software-defined optical interconnection networks for data centers. Several parameters that could potentially be controlled by software are also indicated in the figure.

15.8 Conclusions and Future Research Directions

Cloud computing is still evolving, and new cloud services with increasing requirements on the data transmission and processing infrastructure are being introduced on a daily basis. This trend generates the need for more capacity, flexibility, and

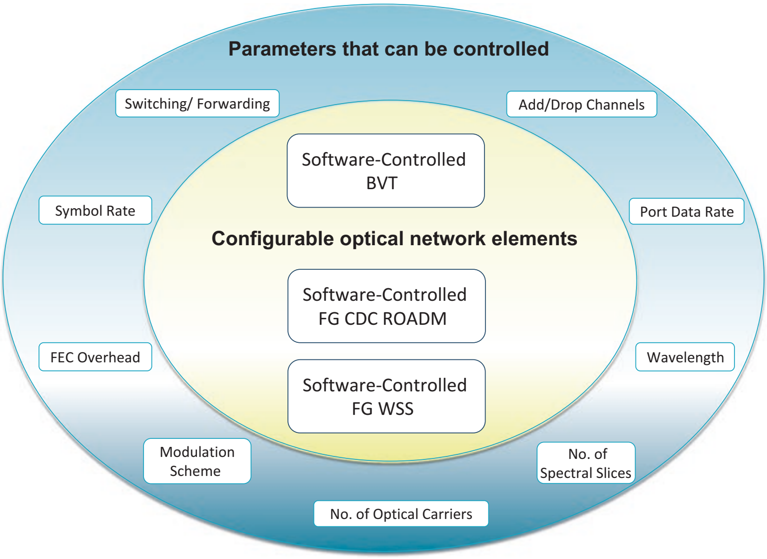


Fig. 15.6 Examples of parameters and network elements that could potentially be software controlled in optical software-defined networking (SDN). *FG* flex-grid, *CDC* colorless, directionless, and contentionless, *ROADM* reconfigurable optical add/drop multiplexer, *BVT* bandwidth-variable transceiver, *WSS* wavelength selective switch

efficiency of data centers and communication networks, in order to provide a seamless end-to-end network infrastructure that is able to optimally support a wide range of different current and future applications and services. To respond to this trend, optical disruptive technologies are expected to penetrate into data centers. Since optical point-to-point interconnects have already been used for years to directly interconnect servers and switches, optically switched interconnects are still in the research phase. Optically switched interconnects basing on either passive [53, 54], active [55, 56], or hybrid [57, 58] architecture and making use of various switching technologies such as optical micro-electromechanical system (MEMS) switches [59], arrayed waveguide routers (AWGRs) [60], electro-optic switches [55, 61], and thermo-optic switches [62] have recently been proposed and demonstrated. The switching time ranges from relatively slow in the order of milliseconds, suitable for a circuit-switched operation, to fast switching in the picosecond range as needed for dynamic optical packet switching. Several data channels can be transmitted over a single fiber by using either time-division multiplexing (TDM) or wavelength-division multiplexing (WDM) or spatial-division multiplexing (SDM), which can contribute to a reduction of the required number of cables. Various multilevel modulation formats can be used to increase the spectral efficiency and, when using bandwidth-variable software-controllable transceivers (BVSCT) and flex-grid optical switches, to increase the flexibility by providing a dynamic routing and spectrum assignment in an elastic manner.

As for on-chip interconnects, various network topologies and switch realizations can be considered along with the possibility to integrate lasers directly on a two-dimensional chip or a three-dimensional chip stack using the hybrid silicon technology. The main target will be to design chip-level interconnects that satisfy the high requirements on bandwidth density, latency, and energy efficiency in order to support future developments in high-performance multicore processors. The chance for optics to become the dominant technology for on-chip interconnects depends also on the future development of other disruptive technologies such as carbon nanotubes [63]. At the board and module level, one of the most promising options from the packaging perspective might be to integrate low-loss optical waveguides on standard PCB materials such as FR4 [29, 30]. Using this method, additional optical layers with polymer waveguides can be embedded into the board that remains compatible with the existing and widely used PCB technology. One of the most important challenges in fabricating board-level optical interconnects is to achieve a precise mounting of active components and an efficient and reliable coupling between the active components and the optical waveguide [31].

Thus, according to the above discussion, one can identify future research directions such as: (*i*) optimizing the internal network architecture, cross-layer, and cross-level performance analysis; (*ii*) introducing elastic and software-defined networking, scalable routing, and optimal load balancing; as well as (*iii*) implementing low-latency and energy- and cost-efficient optical networks on chip (NoC). These research topics are briefly summarized in the following.

15.8.1 Optimizing the Architecture of Optically Switched Interconnects

A lot of recent research work has concentrated on defining and analyzing new and promising architectures for system-level optical interconnects in data centers. Most of the technologies and architectures proposed so far have been initially developed for the application in access and core networks and slightly adapted to match the needs of data centers. The architectures are mostly based on indirect regular networks such as tree, e.g., [57], Clos, e.g., [64], and ring topologies, e.g., [65]. In some recent works, direct regular networks such as n-dimensional cube [66] and n-dimensional torus [67] have been assumed.

Hybrid architectures usually rely on a combination of commercial electronic switches for dynamic packet switching and simple yet energy-efficient optical switches providing circuit switching capabilities. While hybrid architectures are rather flexible and able to adapt to varying traffic situations as well as cost-efficient because they use the commercial state-of-the-art technology, the need for commodity electronic switches makes them less viable long-term solution for future data center networks.

Optically switched interconnects can be seen as a promising candidate for future data centers because they offer the highest capacity and bandwidth density as well as the potential for lowest latency among all interconnection options. When implemented in a pure circuit-switched manner by using large slow optical switches such as optical MEMS switches, the system can be built to provide high scalability, low energy consumption, and a relatively low cost. However, the applications requiring dynamic switching cannot be optimally supported because of the large reconfiguration overhead of circuit switching, which leads to a low transmission efficiency. On the other hand, architectures providing fast all-optical packet switching are usually more complex and expensive and typically less scalable. Additionally, the lack of practical optical buffering technologies limits the achievable performance of large all-optical packet-switched networks. Thus, the architecture of choice needs to provide very good scalability as well as high efficiency and reliability. The term efficiency is to be broadly construed and includes transmission, energy, and cost efficiency. Most probably, there will be not only a single architecture that fits all needs, but rather a number of selected architectures that are designed to best meet the requirements of specific applications.

15.8.2 Cross-Layer and Cross-Level Performance Analysis

In order to identify the most suitable architecture for a specific data center implementation and target applications, one needs first to evaluate its performance. The performance should be evaluated by taking into consideration various technological, architectural, and economic parameters at different hierarchical levels along with realistic traffic scenarios. Since the internal interconnection network in large-scale data centers is usually very complex and there are interdependencies between different hierarchical levels, there is a need for a powerful and efficient holistic toolbox that is able to take into account all the design parameters in order to estimate the most important performance metrics at the level of the entire system.

15.8.3 Elastic and Software-Defined Optical Interconnects

The deployment of elastic network elements such as flex-grid wavelength selective switches (WSS) and reconfigurable add/drop multiplexers (ROADM) in a combination with variable bandwidth transceivers (VBT) can potentially lead to a more flexible operation, a higher utilization of available sources, higher energy efficiency, and better restoration capabilities. Thus, an elastic data center network would make possible to dynamically adapt the allocation of available resources according to application needs. With regard to multiplexing and modulation formats, the elastic optical infrastructure based on optical orthogonal frequency-division multiplexing

(OOFDM) has recently gained particular attention [68]. The advantage of OOFDM is the possibility to achieve an agile optical spectrum management and a seamless integration of the physical transmission layer with upper layers [69].

Various data center applications such as replication, backup, and service migration require guaranteed quality of service (QoS) levels that usually specify short delays, high availability, and high data rate. To be able to provide QoS guaranties in a flexible and efficient manner, software-defined networking (SDN), and particularly the OpenFlow protocol [70], has recently gained popularity [71]. When combining SDN with the network function virtualization (NFV) paradigm and using elastic and software-controlled optical network elements, an efficient, scalable, customizable, and application-aware optical data center network could become reality.

Since flexible and software-controllable optical network technologies have initially been developed for the application in core networks, one could ask whether it makes sense to adopt these technologies in data centers. This question arises especially because core networks have traditionally been designed and operated according to different requirements than data centers. For example, while in core networks, a limited number of fiber cables are used to transmit highly aggregated traffic between several high-capacity nodes, a huge number of transmission links between a large number of servers and switches are typically needed in data centers to transmit individual data flows with a much lower granularity. Additionally, the requirements on cost and energy efficiency of data transmission and processing systems are much more restrictive in data centers than in core networks. However, even though traditional planning and design processes for data centers and core networks follow different goals and the currently relatively high cost of flexible and software-controllable optical systems make their use in data centers less attractive, one can argue that this technology might become one of the most suitable options for future data centers because of its high efficiency, flexibility, and adaptability.

Indeed, both network carriers and cloud infrastructure providers are currently on the verge of a paradigm shift. Current trends in network function virtualization and software-defined networking (SDN) Switching in data center: are significantly changing the core network landscape and require rethinking the traditional approaches for network planning and operation. Actually, virtualization techniques and SDN have primarily been developed for the use in data center environments. On the other hand, optical technologies that have been used for decades in core networks are currently penetrating into data centers. Thus, it seems logical that flexible and efficient optical systems that are capable of providing high data rates in a flexible manner along with software controllability and virtualization capability can be excellent candidates for implementing future high-performance data center interconnection networks, provided the technology becomes less expensive in the future. An additional benefit of using elastic and software-defined optical technologies within data centers is their potential compatibility with future optical core networks, in which the same technology will probably be used to cope with the high requirements set by advanced applications and services in the areas of mobile communications, cloud computing, and cyber-physical systems. The compatibility of inter- and intra-data center network technologies could be proved advantageous in providing, in an energy-efficient way, high data rate and low-latency connections within data centers, between two data centers as well as between data centers and users [72].

15.8.4 Efficient Optical Interconnects

Currently, electronic packet-switched architectures are efficient and relatively cheap when compared with optical WDM solutions. However, this can change in the medium-/long-term future. According to the latest Cisco forecast, the data center traffic is increasing at the very high compound annual growth rate of 25%, and the majority of this traffic is exchanged among servers within the same data center [73]. Moreover, while today most of the servers are equipped with 10 Gbit/s network interface cards (NICs), in the future, more advanced NICs operating at 40 Gbit/s and 100 Gbit/s are expected to be introduced [74]. Consequently, electronic packet switches with huge capacities and equipped with high-speed ports will be needed inside the data center network to keep up with these trends. However, electronic line cards operating at high data rates, e.g., 100 Gbit/s and higher, are still very expensive and consume a large amount of power [75]. In addition, electronic switches are not very scalable because both cost and power consumption increase almost linearly with the aggregate data rate [76]. For this reason, there has been recently significant research effort to define scalable optical switching architectures for data centers [77]. In the short/medium term, hybrid solutions could be adopted, where optical circuit switches are used in parallel to conventional electronic packet switches to transmit elephant flows [78]. However, in the medium-/long-term future, more advanced optical technologies are expected to be gradually introduced in order to meet the very high data center traffic demand [77]. This could be made possible by new, more efficient, and less expensive optical devices, e.g., based on silicon photonics technologies [79].

The comparison between optical and electronic packet-switching technologies for data centers has already been performed several times in recent technical literature, e.g., in [59, 80, 81]. The main conclusion from these studies is that, although electronic packet-switching technologies are still less expensive than their optical counterparts considering current traffic levels, when considering the expected traffic increase in the future, optical switching architectures will become more cost-efficient. This is mostly due to the fact that the cost of optical switches is not very sensitive to an increase in transmission data rate, while the cost of electronic packet switches increases almost linearly with the transmission rate. Based on these results, electronic packet-switching architectures supporting very high-speed connections (e.g., 1 Tb/s) would probably be more expensive and power consuming comparing to optical switching architectures. Another important issue in current data centers is the cabling complexity, which derives from the large number of required fiber links

[82], also referred to as the wiring problem. The wiring problem makes the data center network planning, operation, and maintenance more complex and expensive. To solve this problem, advance optical transmission technologies such flex-grid and spatial-division multiplexing (SDM) can be seen as promising solutions, as they allow transmitting larger amounts of data over a reduced number of optical cables.

15.8.5 Scalable Routing and Load Balancing

It is expected that wavelength-division multiplexing (WDM) will be used at all interconnection levels, i.e., from rack-to-rack to on-chip interconnects, because it can provide high increase in capacity and a reduction of the required number of cables, thereby relaxing the wiring problem. An additional increase of spectral efficiency and better bandwidth granularity can be achieved by combining WDM with other multiplexing and modulation formats and using a flexible wavelength grid. Due to the high diversity of data center traffic and a large number of coexisting connections that need to be set up and maintained in a dynamic manner, it is not a trivial issue to design and implement a dynamic, scalable, and efficient routing and resource provisioning within the internal data center network. Similarly, it is challenging to implement an efficient load balancing method in the optical domain. Therefore, new approaches for efficient and dynamic routing and wavelength assignment as well as for implementing load balancing will be needed.

15.8.6 Low-Latency and Efficient Optical Networks on Chip (NoC)

As with other hierarchical interconnection system levels, future optical networks on chip (NoC) will need to outperform electrical interconnects with respect to all the important metrics such as bandwidth density, latency, and power consumption to become the technology of choice for next-generation processor chips. While the advantages of optical interconnects in comparison with their electrical counterparts are obvious at the rack-to-rack level, it is still not clear, if optical interconnects within the chip are a viable option. The recent developments and the remarkable capabilities of nanoscale silicon photonic technology promise a practical integration of photonic waveguides and components within the commercial CMOS chip manufacturing processes. However, the additional power consumption and latency induced by the electrical-to-optical conversion as well as the losses occurring while coupling the light from external sources into internal waveguides must be further reduced. A possible solution of this problem could be to integrate sources on the processor chip, either by packaging or by bonding, which would eliminate the need for the fiber-to-chip coupling and increase the energy proportionality [83].

References

1. Top 500 Computer Sites Statistics on high-performance computers. http://www.top500.org/
2. H. Cho, P. Kapur, K.C. Saraswat, Power comparison between high-speed electrical and optical interconnects for interchip communication. J. Lightwave Technol. **22**(9), 2021–2033 (2004)
3. D.A.B. Miller, Device requirements for optical interconnects to silicon chips. Proc. IEEE **97**(7), 1166–1185 (2009)
4. P. Westbergh et al., 32 Gbit/s multimode fiber transmission using high-speed, low current density 850 nm VCSEL. Electron. Lett. **45**(7), 366–368 (2009)
5. P. Pepeljugoski et al., Low Power and High Density Optical Interconnects for Future Supercomputers. in *OFC 2010*, San Diego, California, USA, March 2010, paper OThX2
6. Y. Benlachtar et al., Optical OFDM for the Data Center. in *ICTON 2010*, Munich, Germany, June 27–July 1, 2010, paper We.A4.3
7. X. Ye, et al., Assessment of Optical Switching in Data Center Networks. in *OFC 2010*, San Diego, California, USA, March 21–25, 2010, paper JWA63
8. G.I. Papadimitriou, C. Papazoglou, A.S. Pomportsis, Optical switching: switch fabrics, techniques, and architectures. IEEE/OSA JLT **21**(2), 384–405 (2003)
9. K. Vlachos et al., Photonics in switching: Enabling technologies and subsystem design; OSA. J. Opt. Netw. **8**(5), 404–428 (2009)
10. N. Fehratovic and S. Aleksic, Power Consumption and Scalability of Optically Switched Interconnects.in *OFC 2011*, Los Angeles, California, USA, March 2011, paper JWA84
11. C. Kachris, I. Tomkos, K. Bergman, Optical Interconnects for Future Data Center Networks. New York Springer Science & Business Media, 2012, 978-1-4614-4630-9
12. A. D. Hospodor, E. L. Miller, Interconnection Architectures for Petabyte-Scale High-Performance Storage Systems. in *21st IEEE/12th NASA Goddard Conference on Mass Storage Systems and Technologies*, April 2004, pp. 273–281
13. Y. Ajima, T. Inoue, S. Hiramoto, T. Shimizu, Tofu: Interconnect for the K computer. FUJITSU Sci. Tech. J. **48**(3), 280–285 (2012)
14. B. Bohnenstiehl, A. Stillmaker, J. Pimentel, T. Andreas, B. Liu, A. Tran, A. Emmanuel, B. Baas, A 5.8 pJ/Op 115 Billion Ops/sec, to 1.78 Trillion Ops/sec 32 nm 1000-Processor Array. in *IEEE Symposium on VLSI Circuits*, Honolulu, HI, USA, June 2016
15. A. Olofsson, Epiphany-V: A 1024 processor 64-bit RISC System-On-Chip. arXiv:1610.01832v1 [cs.AR], Oct 2016
16. Semiconduction Industry Association International Technology Roadmap for Semiconductors (ITRS) 2,0. *Interconnect*, 2015 Edition
17. J.S. Orcutt, R.J. Ram and V. Stojanović. Integration of silicon photonics into electronic processes. in *SPIE OPTO: Silicon Photonics VIII*, pp. 86290F--86290F, 2013
18. R.W. Morris, A.K. Kodi, A. Louri, R.D. Whaley, Three-dimensional stacked Nanophotonic network-on-Chip architecture with minimal reconfiguration. IEEE Trans on Comput **63**(1), 243–255 (2014)
19. S. Le Beux, H. Li, G. Nicolesu, J. Trajkovic, I. O'Connor, Optical crossbars on chip, a comparative study based on worst-case losses. Concurr and Comput: Pract Exp **26**, 2492–2503 (2014). doi:10.1002/cpe.3336
20. R. Sharma (ed.), *Design of 3D Integrated Circuits and Systems* (CRC Press, Boca Raton, November 2014)
21. P. P. Pande, C. Grecu, M. Jones, A. Ivanov, and R. Saleh, "Effect of traffic localization on energy dissipation in NoC-based interconnect", IEEE International Symposium on Circuits and Systems, vol. 2, Kobe, Japan, 2005, pp. 1774–1777.
22. C. Batten, A. Joshi, V. Stojanović, K. Asanović, Designing chip-level nanophotonic interconnection networks, in *Integrated Optical Interconnect Architectures for Embeded Systems*, (Springer, New York, 2013), pp. 81–135

23. H. Wang, M. Petracca, A. Biberman, B. G. Lee, L. P. Carloni and K. Bergman, Nanophotonic Optical Interconnection Network Architecture for On-Chip and Off-Chip Communications. in *Optical Fiber Communication Conference (OFC/NFOEC)*, San Diego, CA, USA, February 2016, paper JThA92
24. G. Schmid, W.R. Leeb, G. Langer, Experimental demonstration of the robustness against interference of optical interconnects on printed circuit boards, in *IEEE Photonics Society Winter Topicals Meeting*, (Mallorca, Spain, 2010), pp. 93–94
25. M.P. Immonen, M. Karppinen, J.K. Kivilahti, Investigation of environmental reliability of optical polymer waveguides embedded on printed circuit boards. Circuit World **33**(4), 9–19
26. M. A. Taubenblatt, Challenges and opportunities for integrated optics in computing systems. in *SPIE Conference Series Bd. 6124*, 2006, pp. 612406–1 – 612406-11
27. K.K. Tung, W.H. Wong, E.Y.B. Pun, Polymeric Optical Waveguides Using Direct Ultraviolet Photolithography Process. Applied Physics A **80**(3), 621–626 (2005)
28. X. Wang et al., Fully embedded board-level optical interconnects from waveguide fabrication to device integration. IEEE/OSA JLT **26**(2), 243–250 (2008)
29. G. Langer, V. Satzinger, V. Schmid, G. Schmid, W.R. Leeb, PCB with fully integrated optical interconnects. SPIE Photonics West 2011, Optoelectronic Interconnects and Component Integration XI Bd **7944**, 794408-1–794408-15 (2011)
30. R. Houbertz, V. Satzinger, V. Schmid, W. Leeb, G. Langer Optoelectronic printed circuit board: 3D structures written by two-photon absorption. in *Proc. Organic 3D Photonics Materials and Devices II*, pages 70530B, San Diego, August 2008, pp. 1–13
31. S. Aleksic, G. Schmid, N. Fehratovic, Limitations and perspectives of optically switched interconnects for large-scale data processing and storage systems, MRS Proc., Cambridge University Press, Vol. 14382012, 2012, pp. 1–12.
32. S. Aleksic and N. Fehratovic, Scalability analysis of optical intrasystem interconnects", in Journal of Networks, Academy Publisher, Vol. 7, No. 5 2012, pp. 791–799.
33. IEEE, P802.3bs "200 Gbit/s and 400 Gbit/s Ethernet Task Force", http://www.ieee802.org/3/bs/
34. Mellanox Technologies, Introducing 200G HDR InfiniBand Solutions. http://www.mellanox.com/related-docs/whitepapers/WP_Introducing_200G_HDR_ InfiniBand_Solutions.pdf, White Paper, 2016
35. M. Yano, F. Yamagishi, T. Tsuda, Optical MEMS for photonic switching-compact and stable optical crossconnect switches for simple, fast, and flexible wavelength applications in recent photonic networks. IEEE J. Sel. Top. Quantum Electron. **11**, 383–394 (2005)
36. S. Aleksic, Towards the fith-generation (5G) optical transport networks. *Proceedings of the 17th International Conference on Transparent Optical Networks (ICTON 2015)*, Budapest, Hungary, July 2015, pp. 1–4.
37. B. Skubic, G. Bottari, A. Rostami, F. Cavaliere, P. Öhen, Rethinking optical transport to pave the way for 5G and the networked society. J. Lightwave Technol. **33**, 1084–1091 (2015)
38. P.J. Winzer, Scaling optical fiber networks: Challenges and solutions. Opt. Photon. News **26**, 28–35 (2015)
39. I. Djordjevic, M. Cvijetic, C. Lin, Multidimensional signaling and coding enabling multi--Tb/s optical transport and networking: Multidimensional aspects of coded modulation. Signal Process Mag, IEEE **31**, 104–117 (2014)
40. D.J. Richardson, J.M. Fini, L.E. Nelson, Spatial-division multiplexing in optical fibres. Nat. Photonics **7**(2), 354–362 (2013)
41. M. Fiorani, M. Tornatore, J. Chen, L. Wosinska, B. Mukherjee, Optical Spatial Division Multiplexing for Ultra-High-Capacity Modular Data Centers. in *Proc. of IEEE/OSA Optical Fiber Communication Conference and Exposition (OFC)*, March 20–24, Los Angeles, USA 2016
42. M. Fiorani, M. Tornatore, J. Chen, L. Wosinska, B. Mukherjee, Spatial division multiplexing for high capacity optical interconnects in modular data centers. *IEEE/OSA Journal of Opt Commun Networking (JOCN), Special Issue on OFC 2016*, 9, 1, pp. 1–10, 2017
43. D. Qian, M.-F. Huang, E. Ip, Y.-K. Huang, Y. Shao, J. Hu, T. Wang, 101.7-tb/s (370x294-Gbit/s) pdm-128QAM-OFDM transmission over 3x55-km SSMF using pilot-based phase noise mitigation. in *Optical Fiber Communication Conference and Exposition (OFC/NFOEC), 2011 and the National Fiber Optic Engineers Conference*, pp. 1–3, March (2011)

44. D. Qian, E. Ip, M.-F. Huang, M. Jun Li, A. Dogariu, S. Zhang, Y. Shao, Y.-K. Huang, Y. Zhang, X. Cheng, Y. Tian, P. Ji, A. Collier, Y. Geng, J. Linares, C. Montero, V. Moreno, X. Prieto, T. Wang, 1.05 Pb/s transmission with 109 b/s/Hz spectral efficiency using hybrid single- and few-mode cores., in *Frontiers in Optics 2012/Laser Science XXVIII*, p. FW6C.3, Optical Society of America, (2012)
45. J. Fischer, S. Alreesh, R. Elschner, F. Frey, M. Nolle, C. Schmidt-Langhorst, C. Schubert, Bandwidth-variable transceivers based on four-dimensional modulation formats, Lightwave technology. J. Lightwave Technol. **32**, 2886–2895 (2014)
46. E. Hugues-Salas, G. Zervas, D. Simeonidou, E. Kosmatos, T. Orphanoudakis, A. Stavdas, M. Bohn, A. Napoli, T. Rahman, F. Cugini, N. Sambo, S. Frigerio, A. D'Errico, A. Pagano, E. Riccardi, V. Lopez, J. Fernandez-Palacios Gimenez, Next generation optical nodes: The vision of the European research project idealist. Commun Mag, IEEE **53**, 172–181 (2015)
47. Cisco Systems, Power Management in the Cisco Unified Computing System: An Integrated Approach, White Paper, 2011
48. Info-Tech Research Group, Storyboard: Build a Data Center, White Paper, 2009
49. S. Gringeri, B. Basch, V. Shukla, R. Egorov, T. Xia, Flexible architectures for optical transport nodes and networks. Commun Mag, IEEE **48**, 40–50 (2010)
50. M. Xia, M. Shirazipour, Y. Zhang, H. Green, A. Takacs, Optical service chaining for network function virtualization. Commun Mag, IEEE **53**, 152–158 (2015)
51. S. Aleksic, I. Miladinovic, Network virtualization: Paving the way to carrier clouds. in *Proceedings of the 16th International Telecommunications Network Strategy and Planning Symposium (Networks 2014)*, Funchal, Madeira Island, Portugal, pp. 1–6, September 2014
52. S. Peng, R. Nejabati, D. Simeonidou, Role of optical network virtualization in cloud computing [invited], optical communications and networking. IEEE/OSA J **5**, A162–A170 (2013)
53. D. Alistarh, H. Ballani, P. Costa, A. Funnell, J. Benjamin, P. Watts, B. Thomsen, A High-Radix, Low-Latency Optical Switch for Data Centers. *SIGCOMM 15 August 17–21*, London, UK, 367–368 2015
54. J. Chen, Y. Gong, M. Fiorani, S. Aleksic, Optical interconnects at the top of the rack for energy-efficient data centers. In IEEE Communications Magazine **53**(8), 140–148 (2015)
55. R.R. Grzybowski et al., The osmosis optical packet switch for supercomputers: Enabling technologies and measured performance, in *Photonics in Switching*, (IEEE Publications Database, San Francisco, CA, 2007), pp. 21–22
56. N. Calabretta, R.P. Centelles, S. Di Lucente, H.J.S. Dorren, On the performance of a large-scale optical packet switch under realistic data center traffic. J. Opt. Commun. Netw. **5**(6), 565–573 (2013)
57. N. Farrington, Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers. in *ACM SIGCOMM*, New York, NY, USA, 339–350 October 2010
58. M. Fiorani, S. Aleksic, M. Casoni, Hybrid optical switching for data center networks. J Elect Comput Eng **2014**(139213), 1–13 (2014)
59. K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, X. Wen, Y. Chen, OSA: An optical switching architecture for data center networks with unprecedented flexibility. Netw. IEEE/ACM Trans. **22**, 498–511 (2014)
60. K.I. Sato, H. Hasegawa, T. Niwa, T. Watanabe, A large-scale wavelength routing optical switch for data center networks. IEEE Commun. Mag. **51**(9), 46–52 (2013)
61. H. Wang, K. Bergman, A Bidirectional 2x2 Photonic Network Building-Block for High-Performance Data Centers. in *Optical Fiber Communication Conference*, Los Angeles, CA, USA, paper OTuH4 March 2011
62. R. Aguinaldo, A. Forencich, C. DeRose, A. Lentine, D.C. Trotter, Y. Fainman, G. Porter, G. Papen, S. Mookherjea, Wideband silicon-photonic thermo-optic switch in a wavelength-division multiplexed ring network. Opt. Express **22**, 8205–8218 (2014)
63. M.F.L. De Volder, S.H. Tawfick, R.H. Baughmann, A.J. Hart, Carbon nanotubes: Present and future commercial applications. Science **339**(6119), 535–539 (2013)
64. J. Gripp, J.E. Simsarian, J.D. LeGrange, P. Bernasconi and D.T. Neilson, Photonic terabit routers: The IRIS project. *2010 Conference on Optical Fiber Communication (OFC/NFOEC), collocated National Fiber Optic Engineers Conference*, San Diego, CA, paper OThP3 March 2010

65. D. Karthi, G. Das, WMRD net: An optical data center interconnect. *2013 Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC)*, Anaheim, CA, paper OTu3H.3 March 2013
66. K. Chen, X. Wen, X. Ma, Y. Chen, Y. Xia, Q. Dong, WaveCube: A scalable, fault-tolerant, high-performance optical data center architecture. IEEE INFOCOM 2015 **26**, 1903–1911 (2015)
67. K. Kitayama, Y. Huang, Y. Yoshida, R. Takahashi, T. Segawa, S. Ibrahim, T. Nakahara, Y. Suzaki, M. Hayashitani, Y. Hasegawa, Y. Mizukoshi, A. Hiramatsu, Torus-topology data center network based on optical packet/agile circuit switching with intelligent flow management. IEEE Journal of Lightwave Technol **33**(5), 1063–1071., March (2015)
68. P.N. Ji, T. Wang, C. Kachris, I. Tomkos, Energy Efficient Flexible-Bandwidth OFDM-Based Data Center Network 2012. *IEEE 1st International Conference on Cloud Networking*, 119–123, October 2012
69. S. Shen, W. Lu, X. Liu, L. Gong, Z. Zhu, Dynamic advance reservation multicast in data center networks over elastic optical infrastructure. in *39th European Conference and Exhibition on Optical Communication (ECOC 2013)*, London, pp. 1–3, September 2013
70. N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "OpenFlow: enabling innovation in campus networks."ACM SIGCOMM Comp Comm Rev archive. **38**(2), 69–74 April (2008)
71. H. Yang, J. Zhang, Y. Zhao, Y. Ji, Time-aware software defined networking for openflow-based datacenter optical networks", Net Proto Algor, **6**(4), 77–91 December (2014)
72. M. Fiorani, S. Aleksic, P. Monti, J. Chen, M. Casoni, L. Wosinska, Energy efficiency of an integrated intra-data-center and core network with edge caching. in IEEE/OSA Journal of Optical Communications and Networking **6**(4), 421–432 (2014)
73. Cisco White Paper, Global cloud index: forecast and methodology, 2014–2019. August 2016
74. Dell white paper, Data center design considerations with 40 GbE and 100 GbE. August 2013.
75. M.R. Raza, M. Fiorani, B. Skubic, J. Mårtensson, L. Wosinska, P. Monti, Power and cost modeling for 5G transport networks. in *IEEE International Conference on Transparent Optical Networks (ICTON)*, July (2015) pp. 1–7
76. S. Aleksic, Analysis of power consumption in future high-capacity network nodes. J. Opt. Commun. Netw. **1**, 245–258 (2009)
77. C. Kachris, I. Tomkos, K. Bergman, *Optical Interconnects For Future Data Center Networks* (Springer Science & Business Media, NewYork, 2012)
78. M. Fiorani, M. Casoni, S. Aleksic, Performance and power consumption analysis of a hybrid optical Core node. J. Opt. Commun. Netw. **3**, 502–513 (2011)
79. D. Nikolova, S. Rumley, D. Calhoun, Q. Li, R. Hendry, P. Samadi, K. Bergman, Scaling silicon photonic switch fabrics for data center interconnection networks. OSA Opt Express, 1159–1175 (2015)
80. K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, X. Wen, Y. Chen, OSA: An optical switching architecture for data center networks with unprecedented flexibility. in *ACM USENIX Symposium on Networked System Design and Implementation*, April 2012
81. N. Farrington, G. Porter, S. Radhakrishnan, H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, A. Vahdat, Helios: A hybrid electrical/optical switch architecture for modular data centers. in *Proc. ACM SIGCOMM*, September 2010, pp. 339–350
82. S. Aleksic, N. Fehratović, Requirements and limitations of optical interconnects for high-capacity network elements. in *12th International Conference on Transparent Optical Networks*, Munich, 2010, pp. 1–4
83. M.J.R. Heck, J.E. Bowers, Energy efficient and energy proportional optical interconnects for multi-core processors: Driving the need for on-chip sources. IEEE J Sel Top Quantum Electron **20**(4), 1–12 (2014)

Correction to: Silicon Photonics Switch Matrices: Technologies and Architectures

Francesco Testa, Alberto Bianchi, and Marco Romagnoli

Correction to: Chapter 12 in: F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*, https://doi.org/10.1007/978-3-319-61052-8_12

The chapter was inadvertently published with an error. In page 246, line 33, the formula $O[\log_2 N]$ is incorrect. It should be $O[N \log_2 N]$. The same has been updated.

The updated online version of this chapter can be found at
https://doi.org/10.1007/978-3-319-61052-8_12

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
https://doi.org/10.1007/978-3-319-61052-8_16

Index

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
DOI 10.1007/978-3-319-61052-8