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

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 — Francesco Testa
Trento, Italy — Lorenzo Pavesi
April 2017

Contents

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

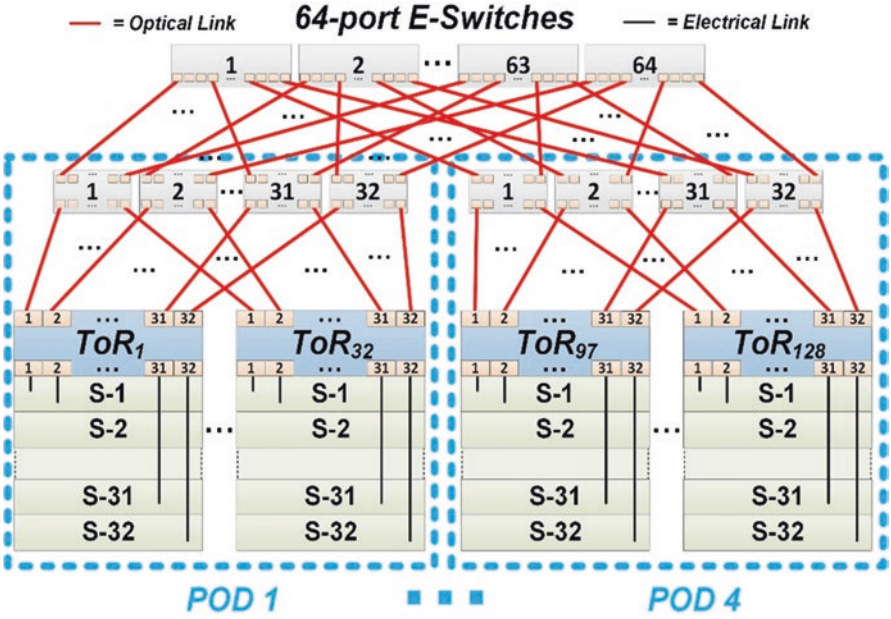

Fig. 1.1 128 rack data center using electrical switches

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.

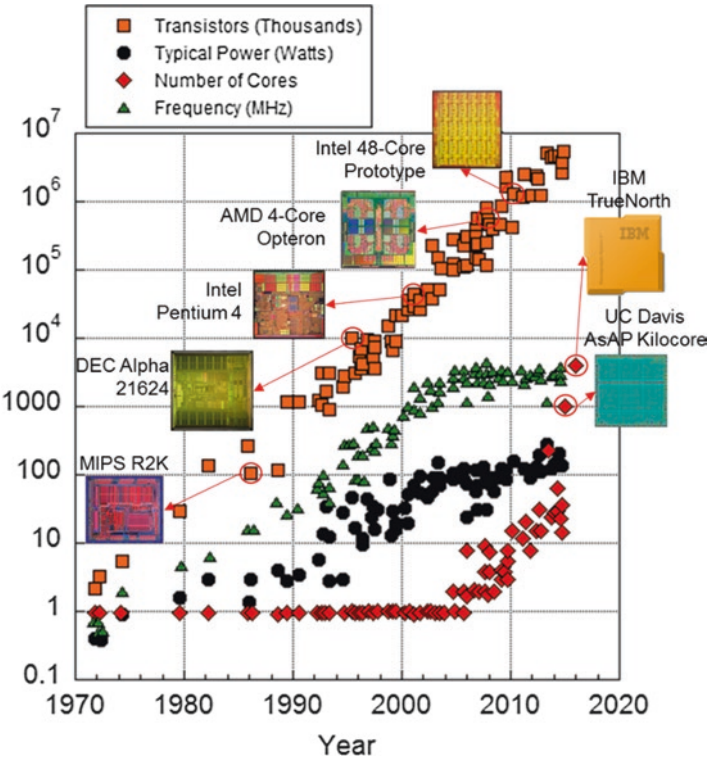

Fig. 1.2 A 45-year trend of a number of transistors per integrated circuit, clock speed (MHz), power (W), performance per clock (ILP), and a number of core per processor die (Figure created based on data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, M. Horowitz, F. Labonte, O. Shacham,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

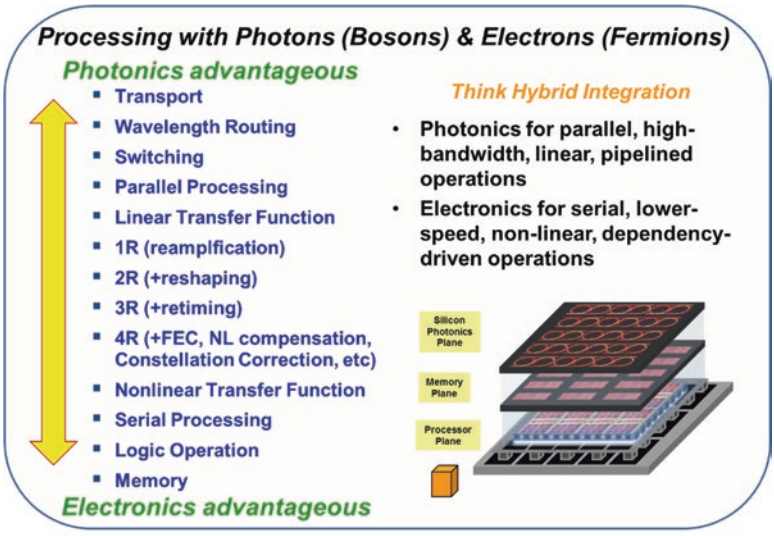

Fig. 1.3 Hybrid photonic-electronic solutions in data systems can offer best of both worlds

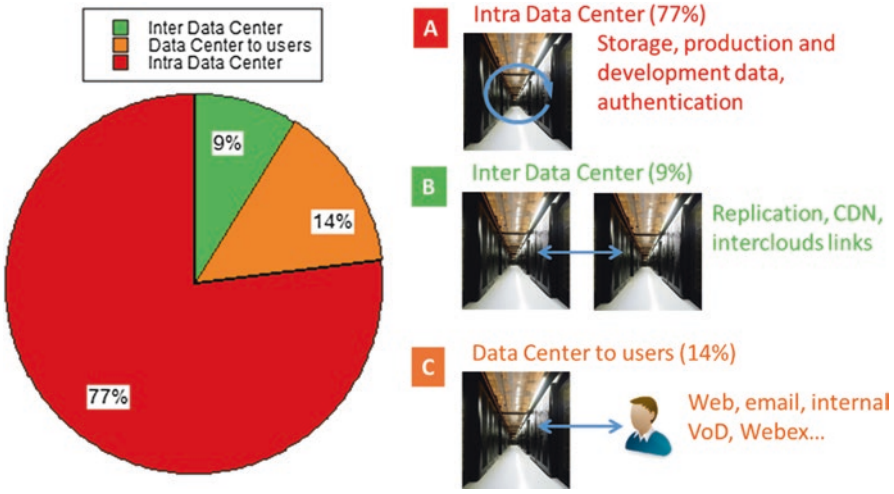

Fig. 1.4 [14] Global Data Center Traffic by Destination in 2020 (Source: Cisco Global Index, 2015–2020; Synergy Research)

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

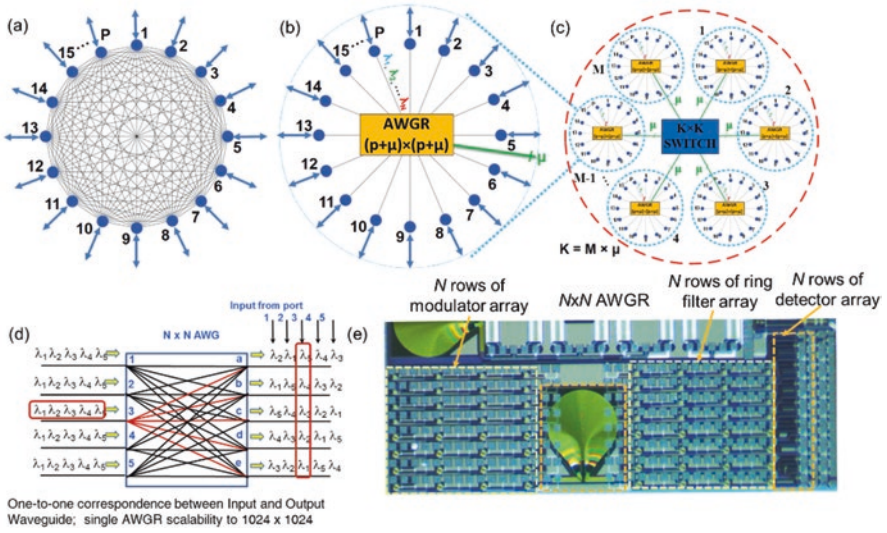

Fig. 1.5 (**a**) Fully connected all-to-all interconnection network, (**b**) fully connected all-to-all interconnection network utilizing wavelength routing by an arrayed waveguide grating router (AWGR), (**c**) RH-LIONS with fully connected subnetworks that are interconnected with a reconfigurable optical 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.

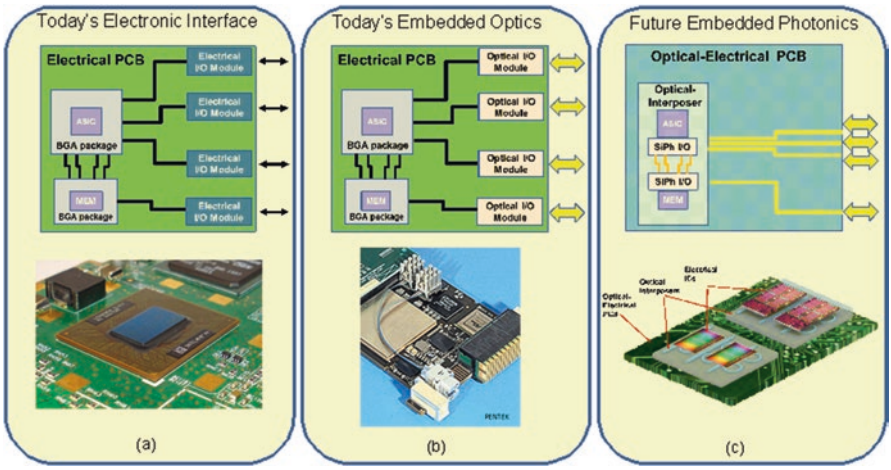

Fig. 1.6 Comparisons of (**a**) today's electronic interfaces with electronic I/Os, (**b**) today's embedded optics with standard pluggable optical interfaces to BGA-packaged ASICs, and (**c**) the proposed embedded photonics with 2.5D and 3D integration using silicon photonic interposers

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.

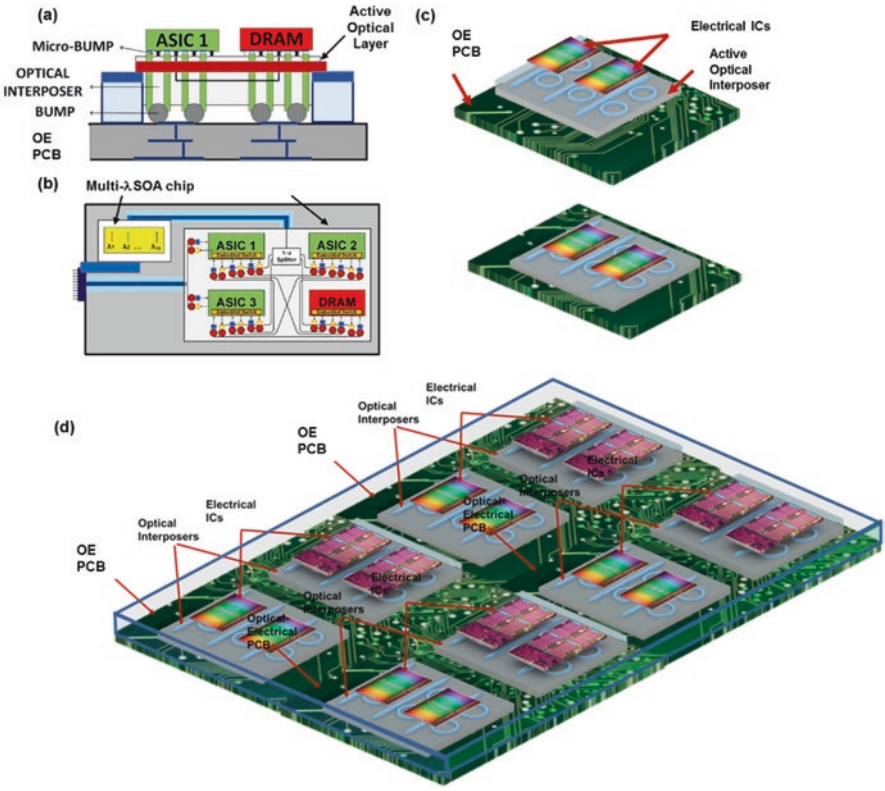

Fig. 1.7 (**a**) and (**b**) show a side view and a top view schematic of 2.5D integration of the electronic ICs, active silicon photonic interposers, and OE-PCBs, (**c**) shows an assembly process using evanescent coupling between the silicon photonic waveguides and the OE-PCB waveguides, and (**d**)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

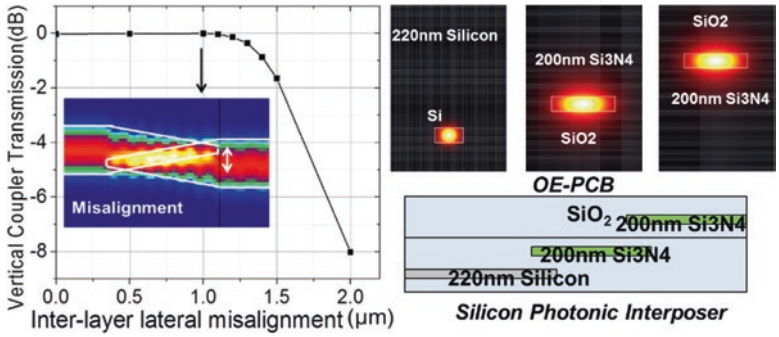

Fig. 1.8 Coupling and misalignment tolerance between the optical interposer and a silicon photonic die consisting of negative tapers indicating +/− 1 μm lateral misalignment tolerance

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

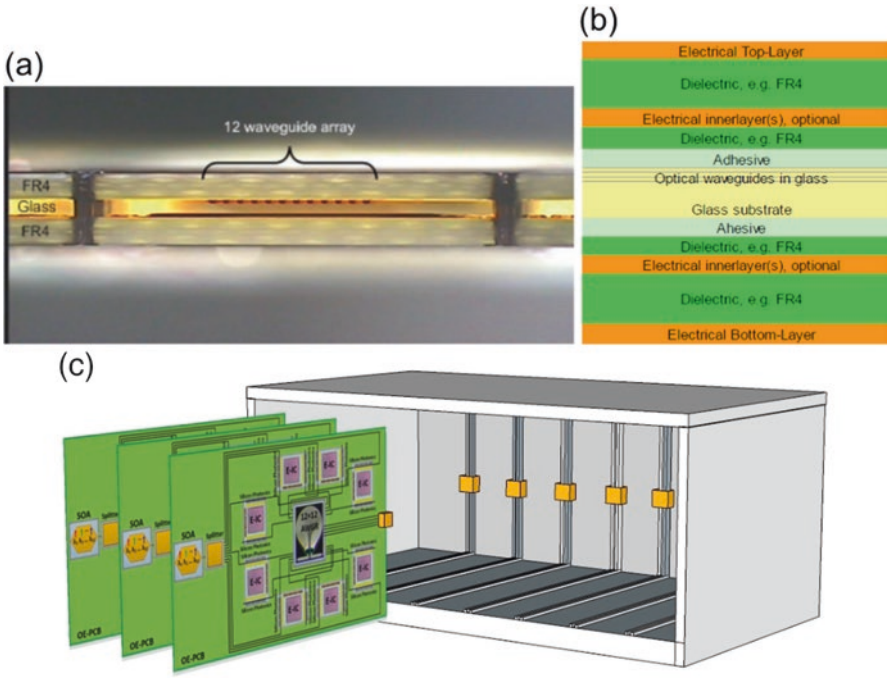

Fig. 1.9 (**a**) A cross-section photograph [36] and (**b**) the composition of a multimode OE-PCB with an ion-exchanged glass waveguide layer sandwiched between two FR4 electrical PCBs [36]. (**c**) Connectorizing the OE-PCBs to realize a chassis with an OE-backplane

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 nonlin-early 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

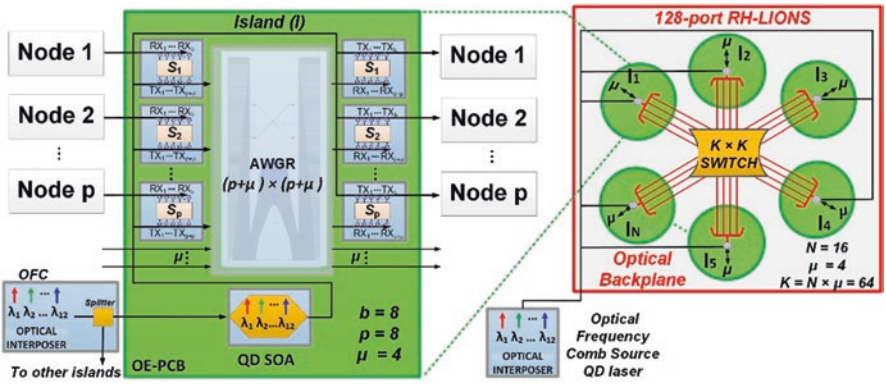

Fig. 1.10 A two-hierarchy switch RH-LIONS switch of size $pN \times pN$. Shown is an example of a 128-port switch with $N = 16$, $b = p = 8$, $\mu = 4$, $K = N \times \mu = 64$

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

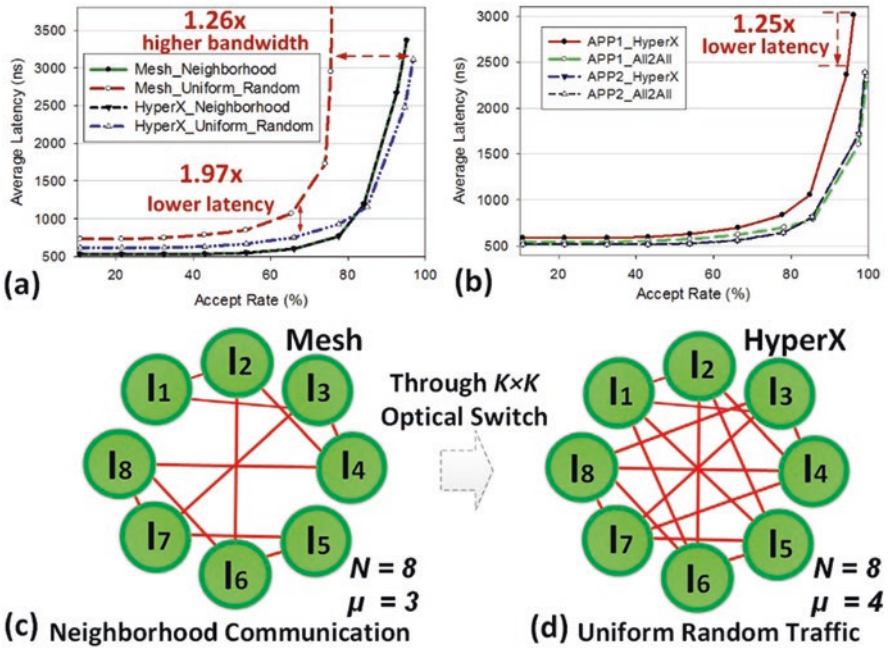

Fig. 1.11 (**a**) Performance of running a single application; (**b**) performance of running two applications; (**c**) full-system mesh with neighborhood communication traffic; (**d**) full-system HyperX with uniform random traffic

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

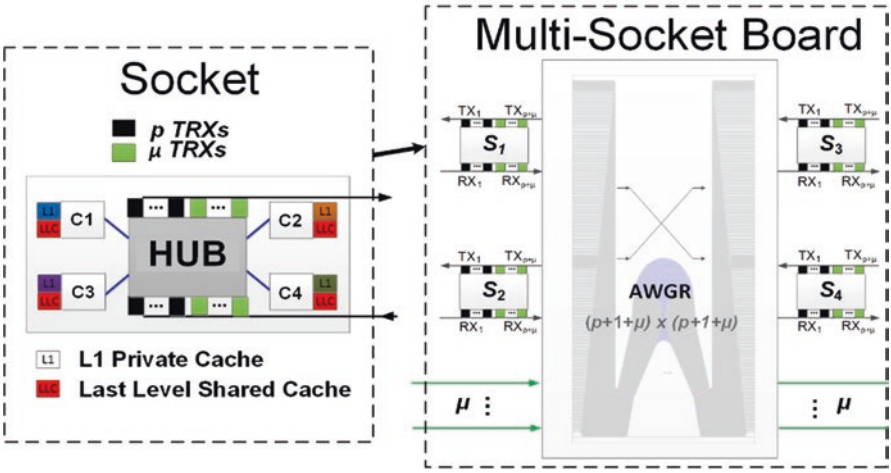

Fig. 1.12 An example of hierarchical optical interconnected architecture for inter-socket communication within a board and between multiple boards. (*Left*): the socket (S) topology with the hub switch connecting the four computing cores with private and shared cache memory. (*Center*): the multi-socket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

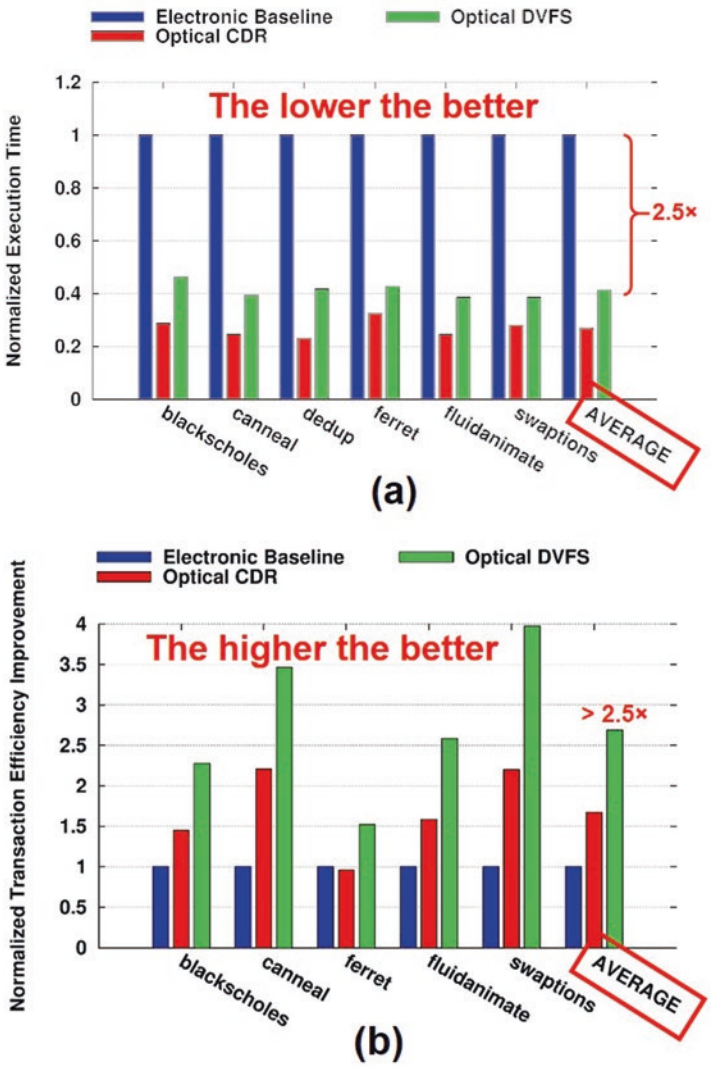

Fig. 1.13 (**a**) Execution time and (**b**) transaction energy efficiency normalized to the electronic baseline in comparison with an optical hierarchical solution with CDR and DVFS [65]

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 °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