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

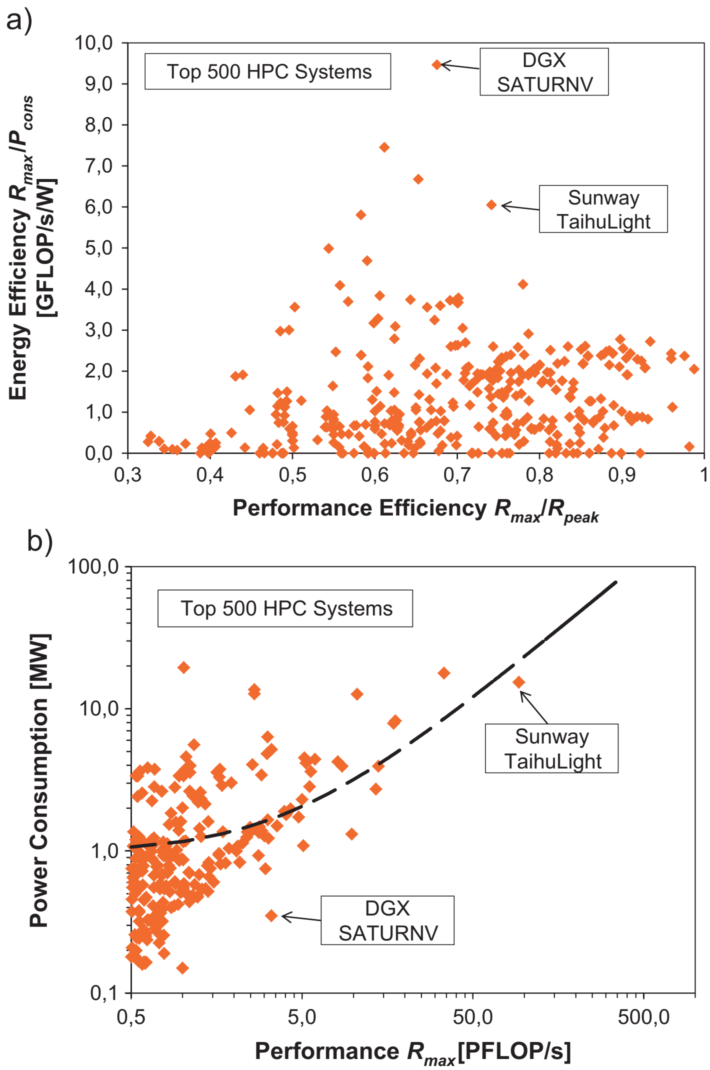

Fig. 15.1 Trends and projections in performance and power consumption of high-performance computers according to the data taken from [2]: (**a**) energy efficiency vs. performance efficiency and (**b**) power consumption vs. computing performance

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.

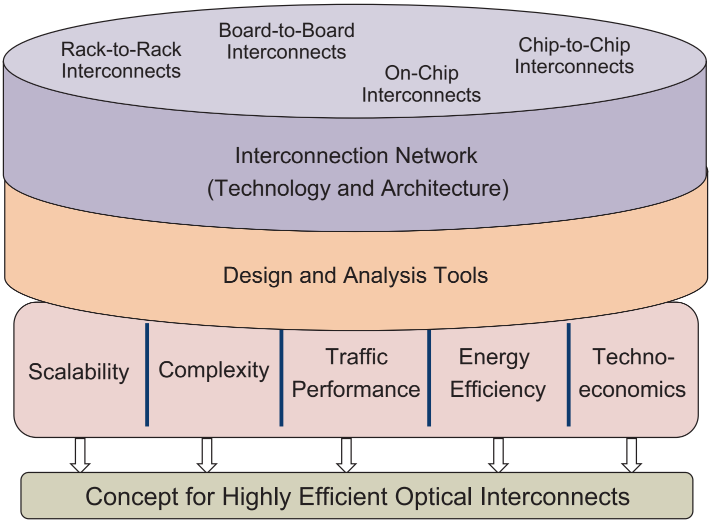
Rack-to-Rack Interconnects
Board-to-Board Interconnects
On-Chip Interconnects
Chip-to-Chip Interconnects
Interconnection Network (Technology and Architecture)
Design and Analysis Tools
Scalability
Complexity
Traffic Performance
Energy Efficiency
Techno-economics
Concept for Highly Efficient Optical Interconnects

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

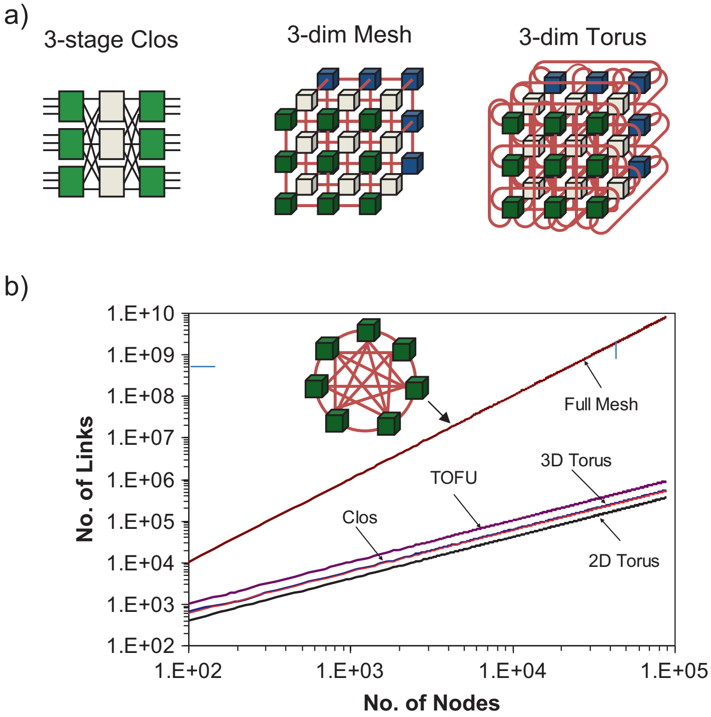

Fig. 15.3 (**a**) Examples of network topologies: three-stage Clos, three-dimensional mesh, and three-dimensional torus. (**b**) Scalability of five exemplary network topologies with respect to the required number of links

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 nonlinear 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].

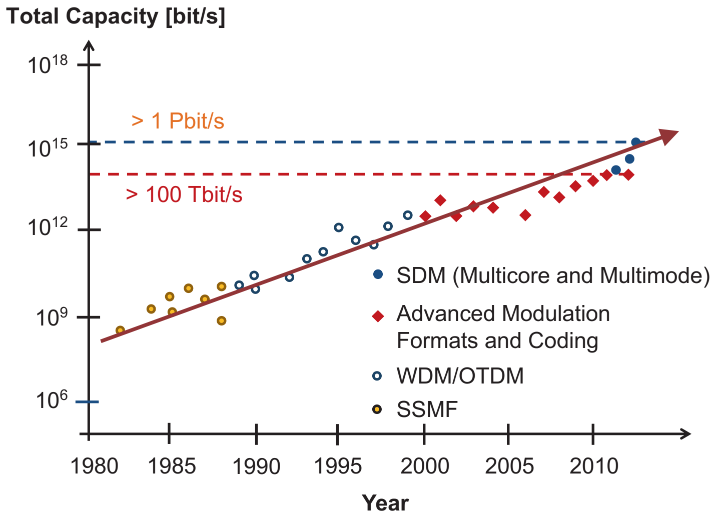

Fig. 15.4 Increase of optical link capacity. SDM, spatial-division multiplexing; *SSMF* standard single-mode fiber, *WDM* wavelength-division multiplexing, *OTDM* optical time-division multiplexing

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.

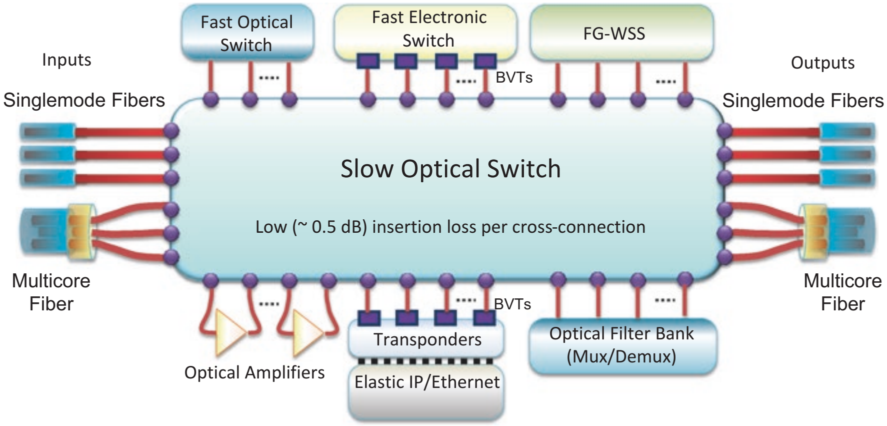

Fig. 15.5 Adaptive node architecture. FG-WSS, flex-grid wavelength selective switch

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

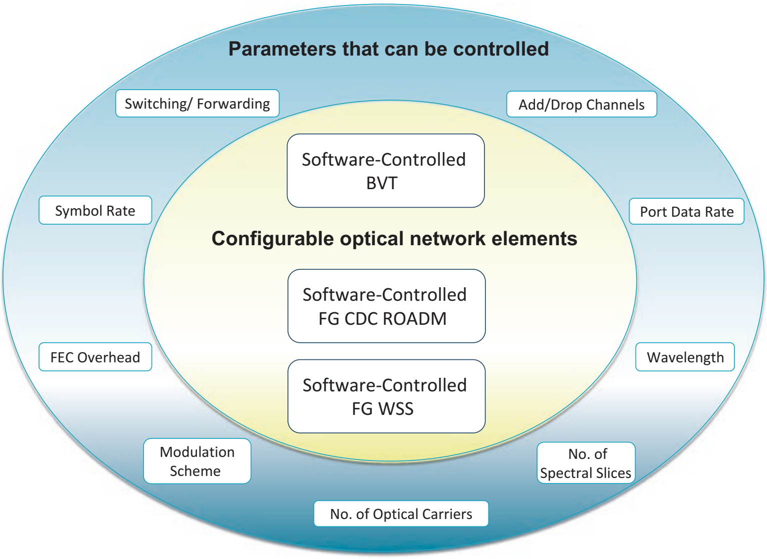

Fig. 15.6 Examples of parameters and network elements that could potentially be software controlled in optical software-defined networking (SDN). *FG* flex-grid, *CDC* colorless, directionless, and contentionless, *ROADM* reconfigurable optical add/drop multiplexer, *BVT* bandwidth-variable transceiver, *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)