CHAPTER 18

Optical interconnection networks for high-performance systems

Qixiang Cheng, Madeleine Glick and Keren Bergman
Columbia University in the city of New York, New York, NY, United States

18.1 Introduction

Large-scale high performance computing (HPC) systems in the form of supercomputers and warehouse scale data centers permeate nearly every corner of modern life from applications in scientific research, medical diagnostics, and national security to film and fashion recommendations. Vast volumes of data are being processed at the same time that the relatively long-term progress of Moore's law is slowing advances in transistor density. Data-intensive computations are putting more stress on the interconnection network, especially those feeding massive data sets into machine learning algorithms. High-bandwidth interconnects, essential for maintaining computation performance, are representing an increasing portion of the total energy and cost budgets.

It is widely accepted that new approaches are required to meet these new challenges [1,2]. Photonic interconnection networks are often cited as ways to break through the energy-bandwidth limitations of conventional electrical wires to solve bottlenecks and improve interconnect performance. In this chapter we begin with an overview of the recent trends in HPC and warehouse scale data centers. We briefly review the challenges due to the slowing of Moore's law and the emergence of machine learning, which are both strongly affecting all aspects of HPC. We then focus on the more immediate supercomputer challenges in the race toward exascale of the CPU-to-memory bottleneck and potential architectural solutions using photonics for bandwidth steering. Turning to the data center, we review the requirements for scaling and improved resource utilization and the need for high-bandwidth intradata center links and the move toward disaggregation. We see that there is a current need for high bandwidth density links in both systems into the server and compute node down to the board and chip module level. At the same time to improve energy efficiency and resource utilization, both supercomputers and data centers are exploring new architectures at all levels of the network from the full system to chip modules. Energy efficient, flexible, adaptable networks involve switched fabrics for which silicon photonics is

Optical Fiber Telecommunications V11
DOI: https://doi.org/10.1016/B978-0-12-816502-7.00020-8

ideally suited. We therefore follow with a review of advances in integrated photonics at the device and system level with specific examples of design explorations.

18.2 Trends and challenges in computing architecture

18.2.1 Overview

We are living in an era of major advances in supercomputing and massive accumulation and analysis of "Big Data." The performance of an Apple iPhone 6 or Samsung Galaxy S5 on standard linear algebra benchmarks exceeds that of a Cray-1, which was considered the first successful supercomputer, and has storage capacity rivaling the text-based content of a major research library [3]. Going forward, both next-generation supercomputers and warehouse scale data centers are facing challenges of scale to transmit, store, and compute. The research and development costs to create an exascale computing system have been estimated to exceed one billion US dollars. At the same time, warehouse scale cloud data centers cost more than $500 million to construct [3].

The cost of electricity to power supercomputing systems and large data centers is a substantial portion of the total cost of ownership. This is a significant part of the motivation for the Department of Energy's (DOE's) Exascale Initiative Steering Committee adopting 20 MW as the upper limit for the system design [4,5].

The supercomputer and data center ecosystems, although not identical in structure or purpose share many similar scaling challenges (see Fig. 18.1). Although in many

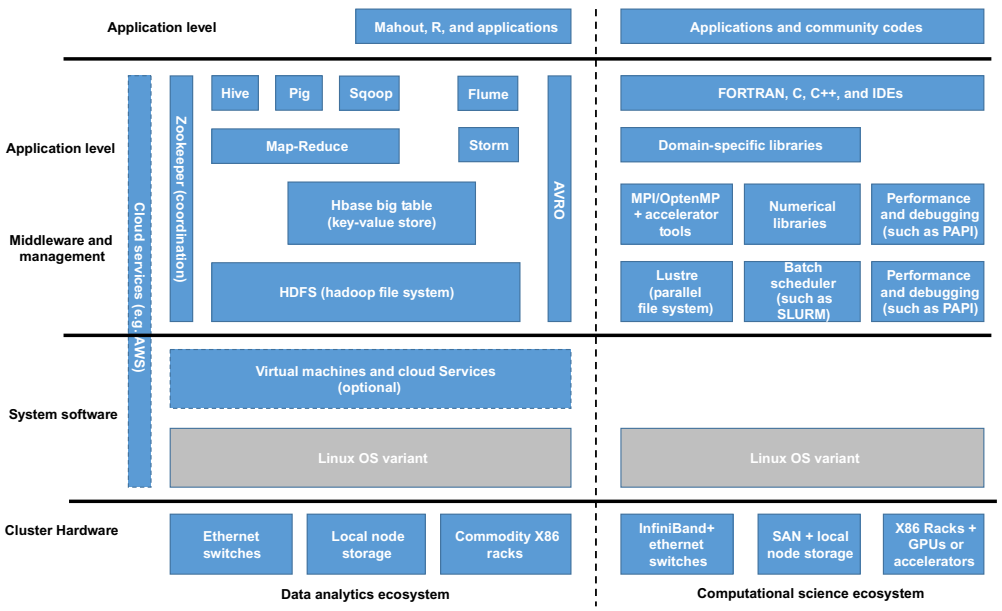

Figure 18.1 High-performance data and computing systems comparison. Source: *From D.A. Reed, J. Dongarra, Exascale computing and big data, Commun. ACM 58 (7) (2015) 56–68 [3].*

ways they are becoming more similar, advanced computing or supercomputing can be defined as those systems computing with multiple petaflops (10^{15} floating operations/second), while cloud data centers can be defined as those with many petabytes of secondary storage. A Cisco study estimates that total data storage capacity will grow from 663 EB in 2016 to 2.6 ZB by 2021 [6].

Supercomputer clusters often make considerable use of accelerators, such as graphical processing units (GPUs) and coprocessors. They use low-latency interconnects (e.g., InfiniBand) and storage area networks. The supercomputer priority is performance rather than minimal cost, while more recently energy efficiency has become a significant metric. Data centers, in contrast to supercomputers, are primarily based on commodity Ethernet servers, often stripped to minimum necessary capabilities, to reduce cost and power consumption.

In the data center, cost and capacity have, until recently, been the primary metrics; however, with new applications, improved performance is also being optimized and accelerators are being used. With increases in scaling, robustness and reliability are becoming higher priorities. Although there are notable differences between the supercomputer platforms and the data/cloud computing centers, significant challenges regarding scaling requirements for power and cost reduction are similar.

Challenges for both the supercomputer and warehouse scale data center are arising from physical hardware limits and burgeoning new applications: from the slowing or ending of Moore's law and the new and almost ubiquitous use of machine learning and data analytics.

18.2.1.1 The end of Moore's law

It has been well known for the last few years that traditional Complementary metal–oxide-semiconductor (CMOS) technology scaling, Moore's law, with the doubling of transistor density every 2 years is ending [7,8]. There is the obvious hard lower limit of the size of molecules in addition to the increasing costs and extreme manufacturing challenges. On-chip power density has reached limitations, and there has been a leveling off of clock frequencies, inhibiting performance increases, which has led to an increased focus on energy efficiency. An exascale supercomputer built with current semiconductor technologies would consume 100s of megawatts of power, an order of magnitude higher than the 20 MW target. This scaling constraint has been met by a turn towards on-chip parallelism and increased use of accelerators. Photonics technology, particularly in the form of wavelength division multiplexing (WDM), is a natural fit for the trends towards parallelism. In addition, CMOS-compatible integrated silicon photonics is enabling high-bandwidth, low-energy interconnects at ever smaller distances, at the board level and possibly eventually on-chip, through monolithic integration and the use of multichip modules (MCMs). In Ref. [10], the authors demonstrate the possibilities of this advanced integration technology

with a report on an electronic–photonic system on a single chip integrating over 70 million transistors and 850 photonic components that work together to provide logic, memory, and interconnect functions.

It is worth pointing out that the end of Moore's law or lithographic scaling does not mean the end of performance scaling. There is considerable research being carried out into post-Moore's law technologies, including novel electronic transistors and spintronic technologies, among others [11]. For any of these new technologies, including photonics, to be incorporated into computing systems, metrics, for lowcost and manufacturability, often high volume manufacturability, must be met before the new technology is widely accepted [1,9].

18.2.1.2 Machine learning and data analytics

Recently, data analytics and machine learning applications are driving increasing amounts of both computation and traffic in the network. Applications include voice assistants such as Apple Siri, Google Voice Search, and Amazon Alexa, facial recognition, spam filters, medical imaging, energy efficiency [12], self-driving vehicles, recommendation services such as on Netflix and Amazon, and many, many more.

Machine learning is a form of computational intelligence that provides computers with the ability to learn and adapt without being explicitly programmed. Neural networks have existed as a form of machine learning since the late 1950s. However, neural networks have become truly useful for perceptual problems over the past 10 years. The change is due to three main factors: (1) availability of large data sets (big data), (2) increased computational capabilities of specialized processors including GPUs and TPUs [13], and (3) advances in machine learning algorithms including parallelization. These all allow predictions to be made in more reasonable time scales [14].

The machine learning process is made up of two main steps: the training stage and inference. Very briefly, in the training stage, the neural network learns to set weights or parameters of the model it is training on [15]. During this stage, training sets of sample data are presented to the network in batches and the weights are adjusted step-by-step (often using stochastic gradient descent) until an acceptable, predefined confidence level is achieved. Training is a compute- and data-intensive process. Often the computation does not fit on a single server or GPU. This is partly due to the large sizes of the data sets [16,17]. The training can take hours [18], days [19], or weeks [20] depending on the number of GPUs available. One Baidu Chinese speech recognition model required 4 terabytes of training data, and 20 exaflops of compute across the entire training cycle [21]. High-bandwidth interconnects are required to maintain performance when the computation must be spread over multiple processing units.

The second main step in machine learning is the inference, or predicting, step, which can be done quickly on a single GPU or processing unit. Here the processor is

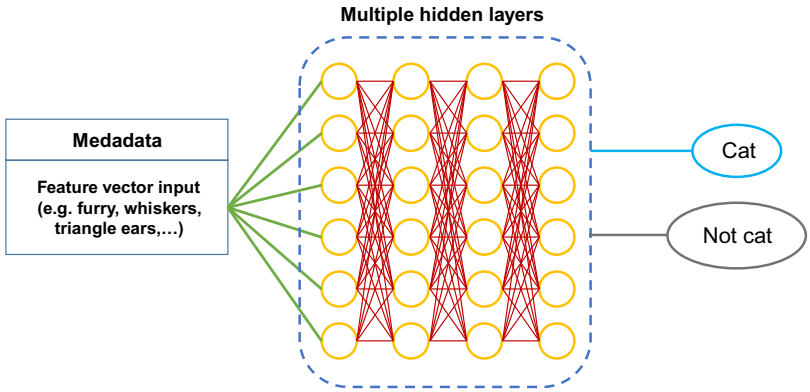

Figure 18.2 Schematic of a machine learning neural network. The input information, the feature vector, gives input values for an appropriate set of features for the problem being studied. The neural network algorithm evaluates the input based on weights on the links defined in the training cycle to determine the prediction. For example, in an image recognition problem, is the picture of a cat?

presented with new information and based on the previously determined weights makes a predictive decision (see Fig. 18.2) [22].

Photonic interconnects have a promising role to play in machine learning in several functions. In particular, as the processing units have significantly differing requirements in the two stages, the architecture may profit from the use of high-bandwidth reconfigurable interconnects. In Ref. [23], the authors use an optical neural network based on a silicon photonic Mach-Zehnder fabric to enhance runtime and energy efficiency.

18.2.2 High performance computing—toward exascale

The next grand challenge for HPC is to reach EFLOPs (10^{18} operations per second), the exascale computer [24,25]. To achieve this in a relatively economical and manufacturably viable manner, the main goal is to design a machine that consumes approximately 20 MW or 50 GFLOPs/W. This goal has been recently made more achievable with major shifts in design that place the memory closer to the GPU [26,27]. Power efficiency has improved in the most recent machines by 2.5 × through the introduction of the new architectures of the Nvidia Tesla P100/Volta V100 and the Zettascaler 2.0 and 2.2. These new architectures including innovative data movement solutions have vastly improved the GFlops/Watt metric [26].

In Figs. 18.3 and 18.4 we can see the trends in the TOP 500 since 2010 [28–30]. The FLOPs/node metric has improved by greater than 50 ×. The byte/FLOP ratio, however, has declined significantly from 0.09 to 0.001. A byte/FLOP ratio below the 0.001 level will result in limitations for programmers, requiring interconnects to scale to far larger bandwidths [27].

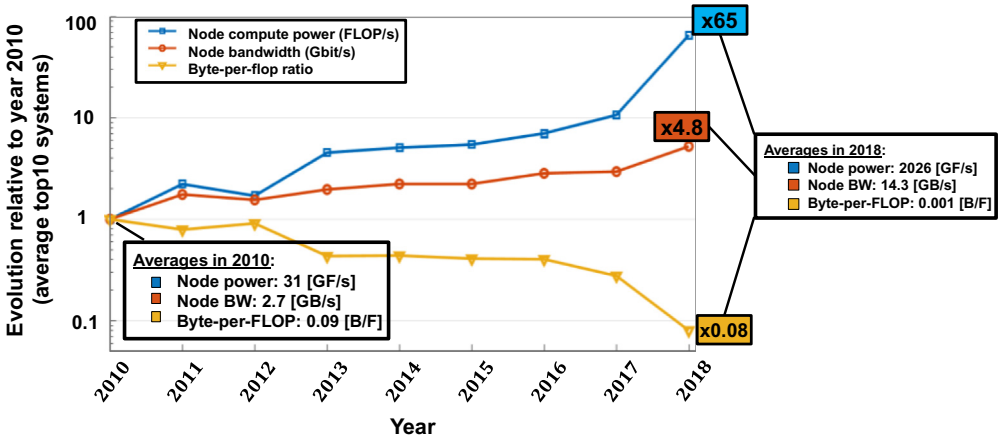

Figure 18.3 Evolution of the average top 10 supercomputers normalized to year 2010.

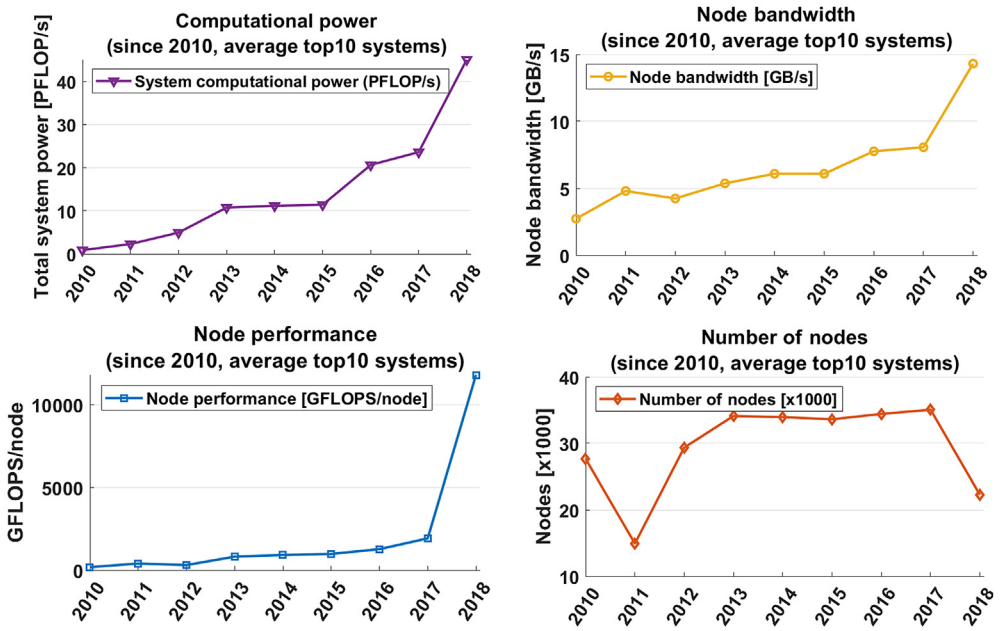

Figure 18.4 Evolution of the average of the top 10 supercomputer systems with respect to computational power, node bandwidth, node performance, and number of nodes.

Photonics is becoming more accepted by the computing community as a technology to provide the required performance. Optical interconnects for supercomputers have been studied since the mid-1980s [31–34]. The use of active optical cables (AOCs) in supercomputers has increased significantly since they were introduced in

2005 [35]. The AOC, however, is a fairly straightforward substitution of an optical link for an electrical cable where longer distance, smaller volume (and larger bend radius), lower weight, and sometimes, even secondarily, higher bandwidth is required. However, photonics can enable further advances in the interconnect, especially with the advent of silicon photonics and photonic integrated circuits (PICs) [36–39].

Current HPC interconnects rely totally on electronics for switching, and still partially for transmission. However, photonics research and development are progressing rapidly in the critical metrics of cost and energy consumption, especially through recent advances of silicon photonics design and manufacturing as described in the next section.

18.2.2.1 The memory bottleneck

For HPC the interconnect has become the bottleneck between CPU and memory. Data movement to other cores is dominating compute power even for short on-chip distances [8,40]. The performance of HPC systems relies heavily on the interconnection network as parallelism increases, resulting in massive data exchange between network endpoints [41–44].

Memory interfaces and communication links on modern computing systems are currently dominated by electrical/copper technology. However, copper wires are reaching limits of bit rate scaling as wire lengths decrease [45,46]. In Refs. [45,46] the author notes that natural bit rate capacity of the wire depends on the aspect ratio, the ratio of the length to the cross-sectional area for a constant input voltage, and does not improve as we shrink the wires down with smaller lithographic processes. As a consequence, power consumption increases proportionally to the bit rate and is highly distance dependent. Photonics technologies have the advantage of having minimal distance dependence and are "transparent" to the signaling rate. Short electronic interconnects are reaching the 1 pJ/bit mark [47–50] but face steep physical limits to get much lower [27,45,46]. Based on these considerations and derived in detail in [4,26,27], energy consumption below 1 pJ/bit has become the target metric for off-chip photonic links.

The memory bandwidth increase is also stressing the pin count limit of the processor package. The pin density of standard chip package cannot scale indefinitely. With pinout for advanced packages already up to 6,000 pins, no opportunity remains to scale performance simply by increasing the pinout [51]. Each SiP waveguide can support terabit/s bandwidth, orders of magnitude higher than what can be achieved with conventional electrical I/O. For example, while an 8-channel (4-layer) high bandwidth memory cube requires a 1024-bit bus for 100 Gb/s, a single SiP waveguide can provide the same bandwidth with 32 wavelengths each at 25 Gb/s [52].

Silicon photonics offers the promise of breaking through the limited bandwidth and packaging constraints of organic carriers using electrical pins, thus solving the challenge of pin-limited bandwidth [8].

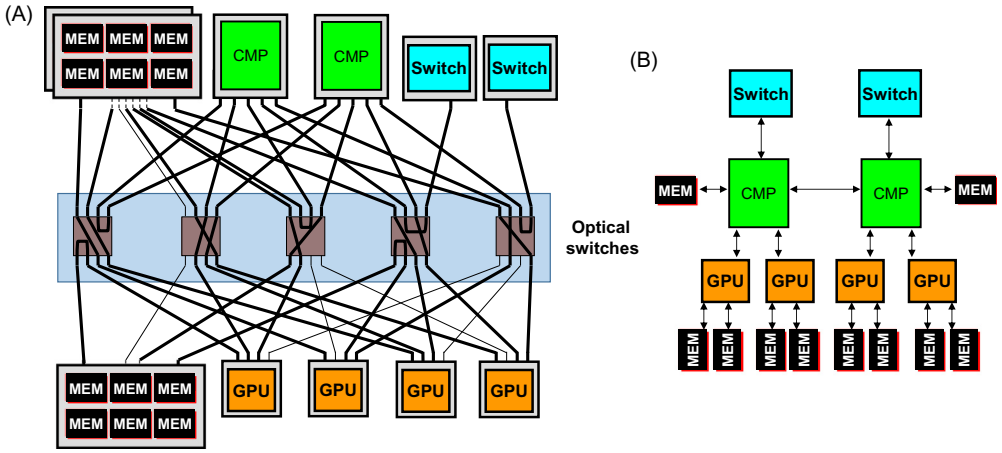

Figure 18.5 An example of beam steering. Photonic switches may be used in (A) to assemble optimized nodes (B) by configuration of the optical switches (shown within the *light blue box*).

18.2.2.2 Bandwidth steering

Applications and network architectures drive the traffic patterns in the computer network. It would enhance the performance of the network architecture if it could match the traffic pattern under consideration. The traffic pattern contains the information of the communication between the nodes in the network. Knowledge of the traffic patterns are therefore critical for optimizing the performance of the architecture. Traffic patterns are usually proprietary. In addition, there can be variations depending on the specific network and the applications running on it [53–55]. Often there is insufficient information on the current and future traffic patterns the architecture should support. A solution to this challenge is the development of flexible, adaptive networks that can take advantage of network resources efficiently and at low cost while meeting bandwidth and latency requirements. The Flexfly network proposed in Ref. [40] uses low to medium radix switches to rewire the interconnect, as required by the application, bandwidth steering, in order to achieve a high-bandwidth low-energy interconnection network with improved resource utilization. More generally, bandwidth steering can be used to change the network configuration dynamically to match the application, as shown schematically in Fig. 18.5 [56]. The original Flexfly network was proposed to modify the Dragonfly architecture, however, the concept of bandwidth steering using silicon photonic switches to improve resource utilization can also be applied to data center networks.

18.2.3 Data centers—scaling and resource utilization

Traffic increases inside the data center are staggering. A Cisco study estimates that the amount of annual global data center traffic in 2016 was 6.8 ZB and will triple

to 20.6 ZB per year by 2021 (see Figs. 18.6 and 18.7). This includes traffic within the data center [6]. Total intradata center traffic does not include traffic local to the rack level, which according to the study is approximately twice the size of the within data center volumes shown in the forecast. The inclusion of rack-local traffic would change our traffic distribution to show more than 90% of traffic remaining local to the data center.

With the growing traffic, there are increasing stresses on the network and the hardware. Autonomous vehicles can produce over 500 GB data per vehicle per day [57].

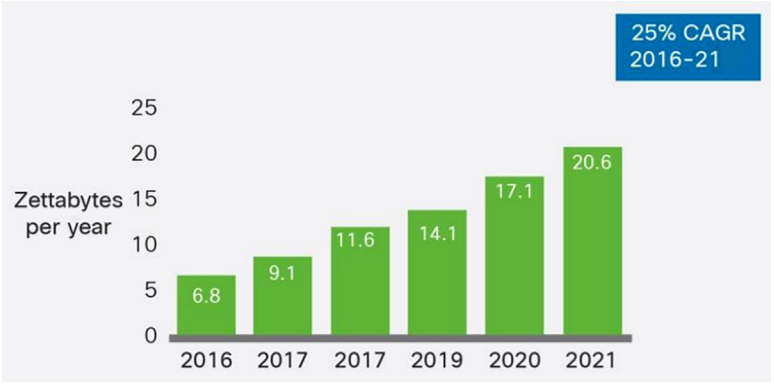

Figure 18.6 Global data center IP traffic growth. Source: *From Cisco Global Cloud Index: Forecast and Methodology, 2016–2021 White Paper. Available from: <https://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.html> [6].*

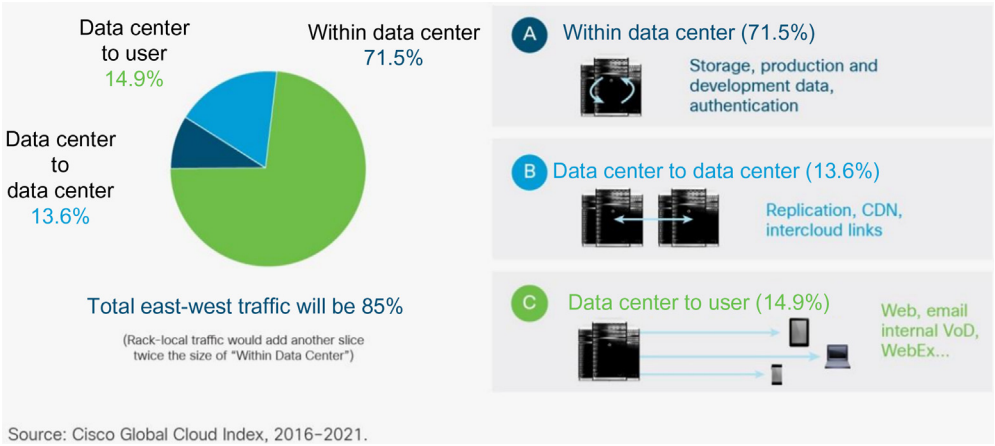

Figure 18.7 Global data center traffic by destination in 2021. Source: *From Cisco Global Cloud Index: Forecast and Methodology, 2016–2021 White Paper. Available from: <https://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.html> [6].*

Some machine learning applications, for example, the training of self-driving vehicles, can use 100 TBps TBps of data and are bounded by available resources. Given these constraints on growth, many research and development programs are seeking ways to enhance performance through improvements at all levels of the architecture, software and hardware. Software improvements are often more easily adopted, as they are usually less risky and less costly. However, despite initial increased hardware costs, Facebook, Google, and Microsoft have found it economically justified to move toward multiwavelength links to achieve higher bandwidth transmission, starting with coarse WDM [58–60]. New architectures have been proposed to improve data center performance, many taking advantage of the high bandwidth density of optics and using optical switches [61,62]. The evaluation of the data center network at the system level depends on several metrics beyond those of cost and power consumption of the hardware. Data throughput and job completion time are also prime metrics. These depend on several factors including scheduling packet transmission and congestion control. In this chapter we focus on the performance of the interconnect level hardware as a basis toward improved performance.

Two current trends for improving data center performance are (1) high bandwidth density communication links and (2) improved resource utilization through disaggregation. In both these areas the advantages of photonic interconnects makes photonics an enabling technology.

18.2.3.1 High-bandwidth links in the data center

There have already been considerable advances in high-bandwidth pluggable optical interconnects for the data center. Large-scale data centers adopted optical transmission technologies during the transition from the 1 to 10 Gbps link data rate between 2007 and 2010.

In 2007, Google began using optical interconnects in its data centers with the introduction of 10 Gbps vertical cavity surface-emitting laser (VCSEL) and multimode fiber-based SFP transceivers for link lengths up to 200 m [63]. With the massive increase of traffic from data center servers over the last several years, it was obvious that the transceiver data rate would be increasing as it has from 10 to 40 Gbps, then 40 to 100 Gbps [64,65]. 100 Gbps links have been commercially available since 2014 and are currently installed in production data centers. Increases to higher rates of 400 Gbps are planned [63]. 400 Gbps transceivers are being standardized by the efforts of IEEE 802.3bs 400 Gb/s Task Force on standardizing short-range (500 m to 10 km) intradata center interconnects over standard single-mode fiber [66,67]. Even higher data rates are being studied with the exception of the eventual need for Tpbs transceivers in the near future [27]. Applications involving machine learning are driving a good portion of this increased need. For example, the DGX-1 station from Nvidia, optimized for machine learning, uses 400 Gbps of network bandwidth [58]. In

addition to expanded bandwidths, optical equipment with improved energy efficiency [10] is also required. It is widely accepted that to achieve the required bandwidth density for the data center, onboard silicon photonics will be used. This can be accomplished either with 2.5D integration on a MCM (Fig. 18.8B and C) or with more advanced 3D integration using through silicon vias (Fig. 18.8D). 2.5D integration is defined by packages in which chips are placed side by side and interconnected through an interposer or substrate.

Advances in silicon photonics manufacturing capabilities are expected to lead to higher bandwidth and considerable energy savings compared to pluggable optics [68]. QSFP56 based on 50 Gbps signaling should increase the front panel BW to 7.2 Tbps;

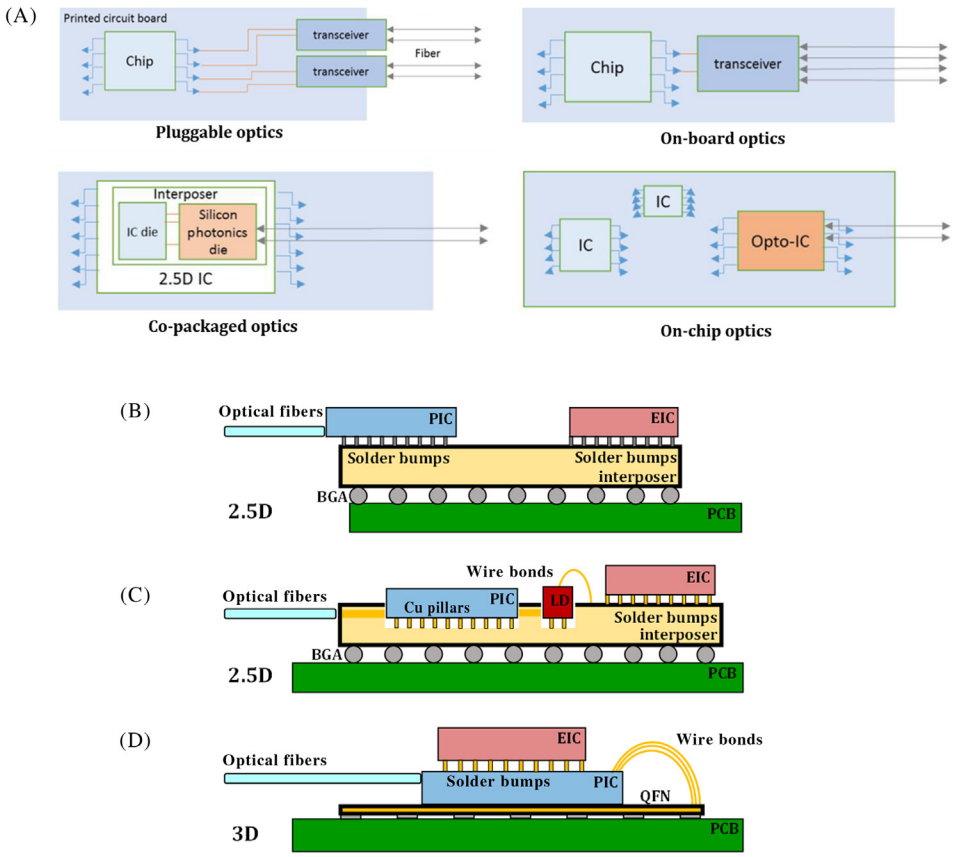

Figure 18.8 (A) Optical interface for pluggable optics, for onboard optics, for copackaged optics and on-chip optics. (B) Schematic of a 2.5D multichip module cointegrating electronics and photonics via an interposer. The interposer only serves as an electrical redistribution layer. (C) Schematic of a 2.5D multichip module. The interposer has both electrical traces and optical waveguides. (D) Schematic of a 3D integrated module.

however, there will eventually be hard limitations to increased bandwidth due to limited area at the front panel and channel impairments on higher data rates [69].

Although the concept of on board optical transceivers is not new, the nearer term data center requirements have provoked vendors to push the technology forward to reduce cost. The Consortium for OnBoard Optics (COBO), led by Microsoft, is defining the standard for optical modules that can be mounted or socketed on a network switch or adapter motherboard. Their initial focus has been on high-density 400 GbE applications [70] with large cloud providers as the early adopters.

Given the requirement for high bandwidth density at low cost and low power consumption, it is not surprising that silicon photonics, fabricated in high volume CMOS-compatible foundries [71,72], is a prime candidate for the interconnection network. Photonic roadmap predictions expect [61] early deployment of 2.5-D integrated-photonic technologies by 2020, and pervasive deployment of WDM interconnects and the beginnings of commercial chip-to-chip intrapackage photonic interconnects by 2025. Roadmapping [73] also sees demand for links to 1 Tbps on boards and 1–4 Tbps within a module by 2020. For very short mm's to cm's distance links on these modules, the energy target is on the order of 0.1 pJ/bit. In the near term the aim is to achieve manufacturable results below 1 pJ/bit.

18.2.3.2 Resource utilization and disaggregation

The traditional data center is built around servers as building blocks. Each server is composed of CPU, memory, one or more network interfaces, specialized hardware such as GPUs, and possibly some storage systems (hard disks or solid state disks). This manner of organizing the hardware is now hitting cost and utilization challenges. Each server element each has its own trends of cost and performance. Upgrading the server to incorporate more recent versions of the CPU or memory requires an entirely new server with new motherboard design [74]. Traditional data centers also suffer from resource fragmentation. Data gathered from data centers show that server memory is unused by as much as 50% or higher [75,76]. This occurs in situations where resources (CPU, memory, storage IO, network IO) are mismatched with workload requirements. For example, a compute-intensive task may not use the full memory capacity or a communication intensive task may not fully use the CPU. These challenges become motivations for disaggregation of the server.

Disaggregation is a concept in which similar resources are pooled and used as required for the application. This enables both the possibility of the resources being independently upgraded and also adaptively configuring the system for optimized performance. The network can be disaggregated at different levels, for example, at the rack or server scale [75,77], as illustrated in Fig. 18.9.

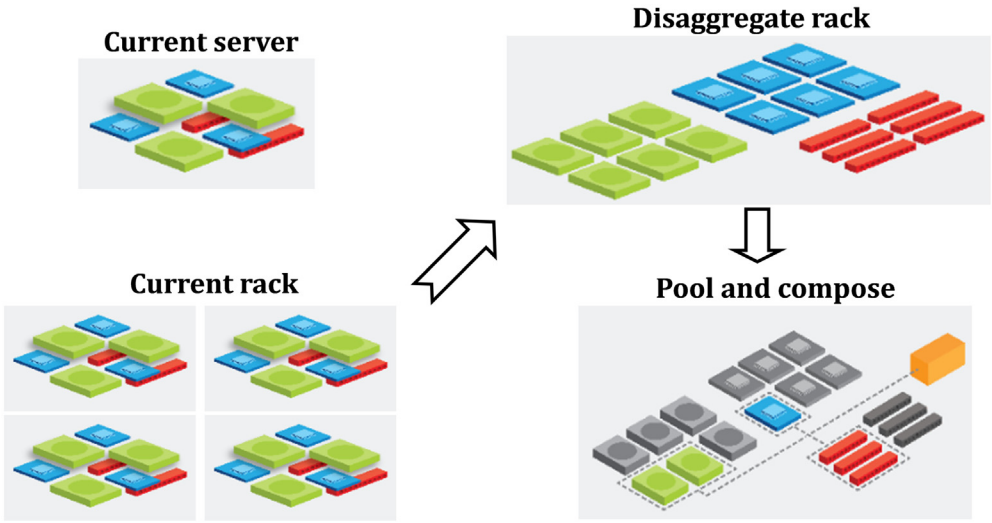

Figure 18.9 A disaggregated rack places resources of different types in different parts of the data center compared to traditional servers and uses networking to pool and compose needed resources together. In the bottom right figure, a logical node is constructed from distant resources.

The disaggregated data center requires a modified interconnection fabric that must carry the additional traffic engendered by the disaggregation and have low latency in order to not only maintain but also improve performance. The network requires a switching fabric to adaptively provision the computing resources. Optical circuit switches are prime candidates for reconfiguration of resources in the disaggregated network. Several reconfigurable data center architectures with optical switch fabrics have been proposed [53,76,78].

In a traditional server, with memory close to the CPU, latency is on the order of 10 seconds of nanoseconds. As a disaggregated network involves additional switched paths, attention must be paid to prevent added latency leading to performance degradation. The cost of the added interconnect components compared to resource savings through improved utilization must also be balanced. Several groups have developed guidelines to achieve these goals [75,77].

Given the requirement for high bandwidth density at low cost and low power consumption, it is not surprising that photonics, and especially silicon photonics, fabricated in high-volume CMOS-compatible foundries [79], is a prime candidate for the disaggregated interconnection network. [75] explores a cost/performance analysis including cost of latency and bandwidth to determine at what point a data center disaggregated memory system would be cost competitive with a conventional direct attached memory system. The authors find that the current cost of an optically switched interconnect should be reduced by approximately a factor of 10 to be an economically viable solution.

18.3 Energy-efficient links

In this section we first review photonic link architectures together with key photonic building blocks. Then we discuss the electrical and optical models used for optical links and present an example of designing the silicon photonic link with performance analysis. The objective of the design is to find the optimal combination of the number of wavelengths, N_λ, and data rate for each channel, r_b, in order to achieve the minimal energy consumption for a given optical aggregation rate. The following work is done through open-source software developed in the Lightwave Research Laboratory, PhoenixSim, which offers a unique and comprehensive modeling platform for efficient design and analysis of the physical layer, link- and system level silicon photonic interconnects.

18.3.1 Anatomy of optical link architectures

As a fundamental building block of optical interconnects, optical transceivers, which consist of the laser light source, modulator, (de)multiplexer and photodetector, are critical for the performance of an optical link. Currently, VCSEL-based transceivers and parallel fibers are the dominant technology in HPC systems. As discussed earlier, the roadmap for ultrahigh-bandwidth, low-energy links requires WDM technology leveraging PICs. Fig. 18.10 schematically shows an anatomy of options for link architectures. Fig. 18.10A shows the transceiver design for a single channel link. Fig. 18.10B shows an approach that is commonly used in telecommunications to combine modulated colored channels using (de)-multiplexers, and in Fig. 18.10C the architecture is equipped with DeMux/Mux stages utilizing broadband modulators, such as electro-absorption modulators and Mach-Zehnder modulators. Another promising architecture, illustrated in Fig. 18.10D, takes advantage of wavelength-selective microring modulators (MRRs) implemented in a cascaded structure, enabling ultrahigh on-chip bandwidth density. To reach the Tbps regime, a large number of wavelengths are required; thus the development of comb lasers with the capability of emitting over 100 individual wavelengths is a promising next step for the progress of transceiver architectures.

18.3.2 Comb laser

The optical frequency comb laser is an appealing alternative to continuous wave (CW) laser arrays as a source for HPC systems in terms of footprint, cost, and energy consumption. A comb laser consists of equally spaced lines in the frequency domain that can be used as separate optical carriers for WDM. Since the comb is generated from a single source and has intrinsically equidistant spacing between its lines, it has the potential to eliminate the energy overhead associated with independently tuning many CW lasers to maintain the desired channel locking. Currently there are two main

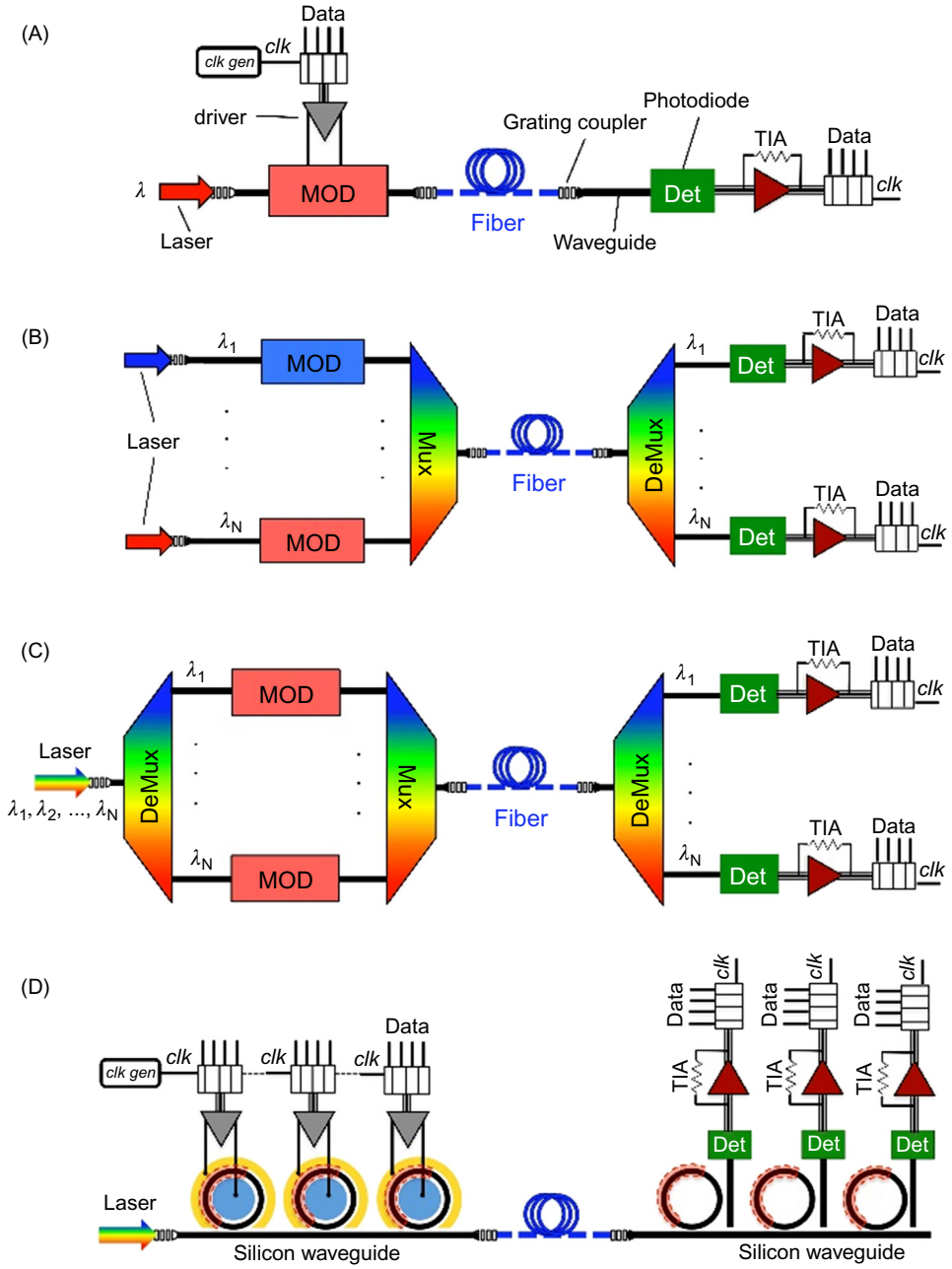

Figure 18.10 Anatomy of various link architectures: (A) single wavelength point to-point photonic link; (B) wavelength division multiplexing (WDM) photonic link based on separate lasers and broadband modulators; (C) photonic link based on parallel broadband modulators and DeMux/Mux. (D) WDM photonic link based on cascaded microring resonators and cascaded drop filters [1].

methods used for generating combs: mode-locking lasers and nonlinear generation using four-wave mixing (FWM) in a microcavity.

Comb generation can occur in a laser by inducing a fixed-phase relation between the longitudinal cavity modes in a Fabry-Perot cavity (mode locking), leading to a stable pulse train in the time domain and therefore a comb with precise spacing in the frequency domain. The channel spacing can be tuned by changing the cavity length. Quantum dot (QD) mode-locked semiconductor lasers, which can be directly grown on silicon [80], are an attractive comb laser source, as the high nonlinear gain saturation of the QD active layer results in low relative intensity noise. Moreover, by controlling the distribution of the sizes of the QDs, intentional inhomogeneous broadening of the gain spectrum, such as a 75 nm broad spectrum of emission, can be achieved [81]. The amplitude and phase noise of such a comb laser source has been greatly reduced through active mode-locking, with reduced optical linewidths of the carriers and increased effective bandwidth that is compatible with coherent systems [82].

Frequency comb generation has also been realized with a CMOS-compatible Si_3N_4 ring resonator through the nonlinear process of FWM in an optical parametric oscillator [83,84], which can be directly integrated in the current silicon photonics platform. In this implementation, numerous equally spaced narrow-linewidth sources can be generated simultaneously using a microresonator with an off-chip CW optical pump, as illustrated in Fig. 18.11A [83]. The pump field undergoes FWM in the resonator and creates signal and idler fields that also satisfy the cavity resonance; these signal and idler fields then seed further FWM, leading to a cascade effect that fills the remaining resonances of the cavity (as illustrated by Fig. 18.11B). This yields many equally spaced optical carriers (with spacing depending on the FSR of the cavity) and has a high pump-to-comb conversion efficiency of up to 31.8% when operating in

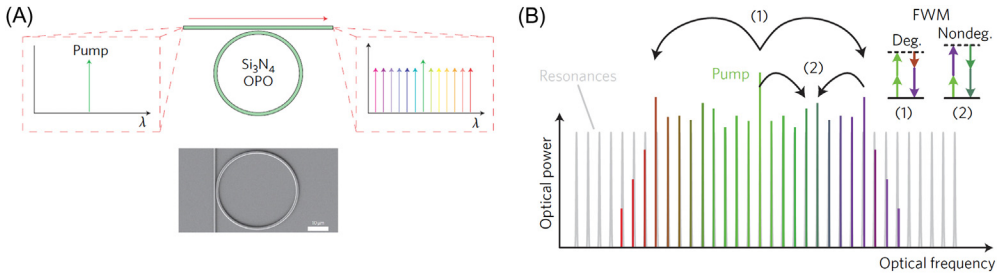

Figure 18.11 (A) On-chip optical comb generator using silicon nitride ring resonator with a single external pump laser. (B) Principle of Kerr comb formation by FWM. Source: *From (A) J.S. Levy, A. Gondarenko, M.A. Foster, A.C. Turner-Foster, A.L. Gaeta, M. Lipson, CMOS-compatible multiple-wavelength oscillator for on-chip optical interconnects, Nat. Photon 4 (2009) 37 [78]; (B) J. Pfeifle, V. Brasch, M. Lauermann, Y. Yu, D. Wegner, T. Herr, et al., Coherent terabit communications with microresonator Kerr frequency combs, Nat. Photon. 8, 375 [84].*

the normal dispersion regime [85]. Recently, a chip-scale comb source was reported [86] using an integrated semiconductor laser pumping an ultrahigh-quality factor (Q) Si_3N_4 ring resonator. This small-footprint device had performance lasting longer than ~200 hours powered from a standard dry cell battery. It is a strong candidate for an energy-efficient optical source for high-performance systems.

Even though the comb laser is a very promising direction of research, in order to be adopted in high-performance systems, several challenges remain. The comb laser must demonstrate advantages over CW laser arrays in terms of energy efficiency, cost, and footprint. To achieve this, the comb lines need to be fully utilized and the comb laser should demonstrate a relatively flat power profile. The optical power per channel needs to be greater than that needed to meet the power budget of the link. Currently most demonstrations have insufficient optical power per comb line and amplification is required to overcome the link power budget. The poor conversion efficiency in the anomalous dispersion regime (~2% pump-to-comb conversion efficiency) poses a major challenge for the wall plug efficiency of the comb source including pump laser.

18.3.3 Microring-based modulators

With its small footprint and wavelength-selective nature, the MRR is a highly promising candidate for realizing high-throughput optical interconnects compatible with comb lasers [87]. Since its introduction in 2005 [88], tremendous improvements have been demonstrated, such as modulation with high speed [89], ZigZag [90], and interdigitated [91] junctions. Advance modulation is also achieved by cascading MRRs along a single bus waveguide, as demonstrated in Refs. [92–94]. Recently the MRR has also been used for higher order amplitude modulation formats such as four-level pulse amplitude modulation (PAM4) at 128 Gb/s [95] and PAM8 at 45 Gb/s [96], achieving an energy consumption of as low as 1fJ/bit [96].

To quantify the performance of an MRR modulator in a cascaded architecture, the power penalty metric associated with the bit error rate (BER) is typically used [97]. The modulated light has a certain optical modulation amplitude (OMA) based on the spectral shift of the resonator. As shown in Fig. 18.12, the spectral response of a PIN-based ring modulator experiences a blue shift with the addition of some excess cavity loss. Such changes are due to the cavity phase shift and round-trip loss, which can be controlled through the driving voltage/current of the modulator [98,99].

Intermodulation crosstalk [97,100] can impact the overall power penalty of modulators in such a cascaded arrangement. A trade-off exists between the spacing of the channels and the shift of the resonance. A larger shift of resonance results in an improved OMA and lower modulator insertion loss but leads to a higher average loss of optical power due to on-off keying (OOK) and higher intermodulation crosstalk. In addition, even though ideally the operation point for the shift of resonance

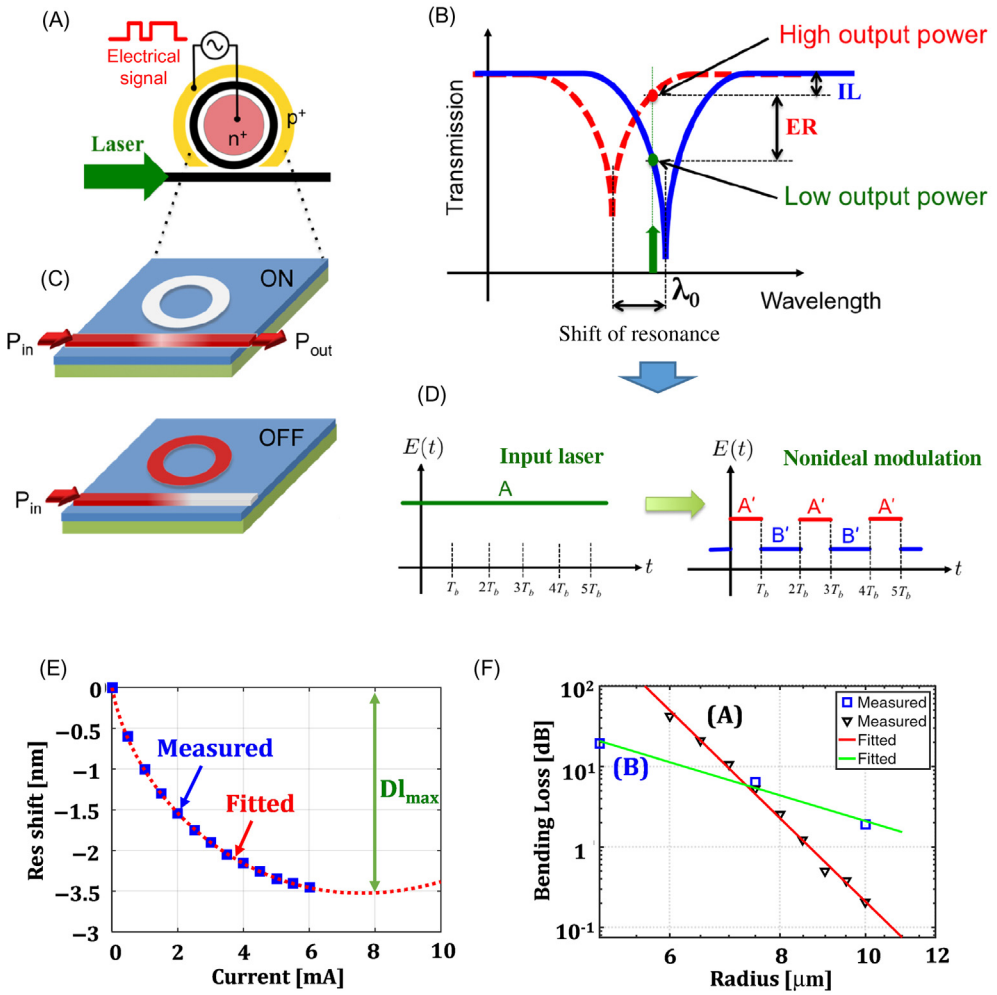

Figure 18.12 Schematic view of a microring modulator (MRR). (A) The high-speed electrical signal is applied to the pn-junction embedded inside the silicon ring. (B) Modulation of the input laser by shifting the resonance of the ring to create high and low levels of optical power at the output. (C) Graphical view of the ON and OFF states of light at the output. (D) Time-domain presentation of a nonideal NRZ OOK modulation. (E) Spectral shift of a PIN-based ring modulator as a function of injected current [94]. (F) Measured bending loss of ring resonators as a function of radius reported in Refs. [103,104] (both horizontal and vertical axes are in log scale). Source: *(A–D) From Q. Cheng, M. Bahadori, M. Glick, S. Rumley, and K. Bergman, "*Recent advances in optical technologies for data centers: a review,*" Optica 5, 1354–1370 (2018) [1].*

should be close to half of the spacing between the channels, PIN-based modulators suffer from Ohmic heating due to the injection of current inside the waveguide. The Ohmic heating limits the blue shift of the spectrum to about 2.5 nm, as shown in Fig. 18.12B. This situation is even worse for PN-based ring modulators due to

their relatively low electro-optic modulation efficiency [101]. The choice of PN or PIN design for ring modulators is therefore twofold based on the desired optical penalty and the operation speed. PN-based modulators exhibit higher optical penalty compared to their PIN-based counterparts but benefit from operating at higher speeds [102].

A key step to establishing the design space exploration of MRRs is to relate the spectral parameters (Q-factor, round-trip loss) to the geometrical parameters (radius, coupling gaps) [105]. The bending loss of silicon ring resonators (in dB/cm) is a critical factor. Fig. 18.12F shows two sets of measurements for the bending loss of silicon ring resonators as a function of the radius, as reported in Refs. [103,104]. This leads to a power-law relation between the bending loss α and radius: $\alpha = A^0 \times R^{-B}$. An analytical approach can then be used for estimating the coupling coefficients between the ring and waveguides as a function of radius and coupling gaps [104]. One can then explore the design space of WDM links based on ring parameters [106]. Other design trade-offs also need to be taken into consideration. For instance, a large FSR supports more optical channels in the cascaded WDM configuration but requires a small radius leading to high bending loss [107].

18.3.4 Microring-based drop filters

As a resonance cavity, MRRs can also be employed in the form of add-drop structures. They are capable of performing wavelength de-multiplexing due to their wavelength-selective spectral response. Based on the desired passband and the rejection ratio of the filter, first-order [108,109] or higher order [110] add-drop filters are used. Higher order filters provide a better rejection ratio but suffer from a higher loss or resonance splitting in their passbands.

The power penalty of ring filters can be estimated based on the Lorentzian spectral shape of the filter. As shown in Fig. 18.13C, if the data rate of the OOK channel is much smaller than the 3 dB bandwidth of the ring, the power penalty is simply based on the spectral attenuation of the MRR. However, if the data rate is comparable to the bandwidth of the filter, a correction needs to be introduced to include the data rate impact on the filter power penalty and the crosstalk effects in a cascaded arrangement [109]. An example of the design space of silicon-based add-drop filters under the critical coupling condition is shown in Fig. 18.13D, with preset design metrics of the insertion loss, optical bandwidth, FSR of the filter, and the extinction of resonance.

For each individual channel, an optimization of the add-drop MRR in the cascaded arrangement can be utilized so that the power penalty associated with the entire demultiplexer array is minimized [97]. This optimization depends on the parameters of the ring, as well as data rate, number of channels, and channel spacing. The

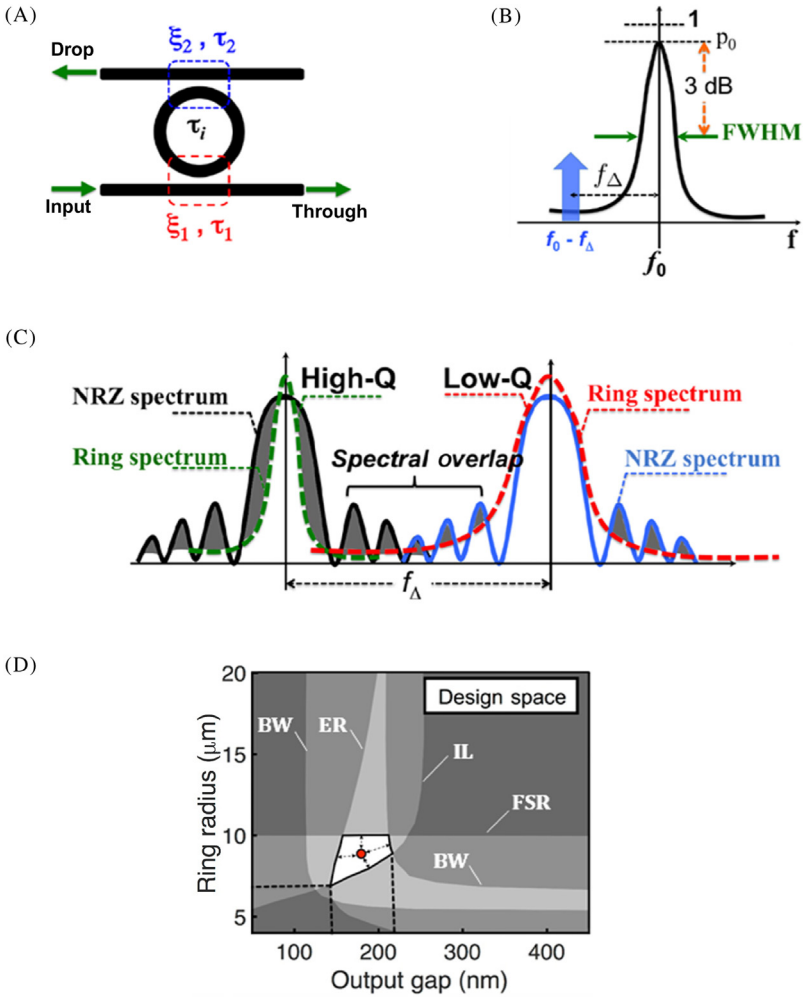

Figure 18.13 (A) Intrinsic and coupling decay rates of a ring add-drop filter. (B) Transmission spectrum of the drop path of a demux ring with the 3-dB bandwidth denoted as FWHM. f_Δ is the possible detuning between the resonance and the channel. (C) Schematic view of two adjacent channels with a fair amount of spectral overlap. A low-Q ring will result in more crosstalk effect but less spectral distortion (highlighted areas on the NRZ spectrum). f_Δ denotes the spacing between channels. (D) Design space of a critically coupled demux add-drop ring. Source: *(A–C) From Q. Cheng, M. Bahadori, M. Glick, S. Rumley, and K. Bergman,* "Recent advances in optical technologies for data centers: a review," *Optica 5, 1354–1370 (2018)* [1].

power penalty imposed on each channel consists of three parts: (1) The insertion loss of the ring—independent of the number of channels and their data rate. (2) The truncation effect—only dependent on the data rate. Strong truncation arises when the 3 dB bandwidth of the MRR is small compared to the signal bandwidth.

(3) Optical crosstalk due to the imperfect suppression of adjacent channels—a function of number of channels and channel spacing. As shown in Fig. 18.13C, the Q-factor of the MRRs is the determining factor in the power penalty space [97]. Increasing the Q will increase the insertion loss of the ring and truncation of the OOK signal, but doing so results in suppression of optical crosstalk. Therefore, an optimized point exists for the minimal penalty.

In addition to the physical properties discussed, other challenges need to be carefully addressed to fully utilize the advantages of MRRs for optical interconnects: (1) Thermal sensitivity: Thermal effects significantly impact the optical response of silicon-based resonant devices due to the strong thermo-optic coefficient of silicon. The resonance of a typical silicon MRR is shifted by $\sim$9 GHz for each degree Kelvin change in the temperature [111]. Such thermal drift in high-Q MRRs can impose more than 1 dB of penalty on high-speed OOK signals [97]. (2) Self-heating: The enhancement of optical power inside the MRR is proportional to the finesse, or Q-factor. Even a slight internal absorption in a high Q MRR can lead to a noticeable thermal drift of resonance. A recent transceiver design has proposed a thermal tuning algorithm based on the statistics of the data stream to counteract this effect [112]. (3) Fabrication variation: The spectral parameters of MRRs such as resonance wavelength, FSR, and the 3 dB optical bandwidth largely depend on their geometrical parameters. It is known that current silicon photonic fabrication imposes variations on the dimensions of the waveguides [113]. This results in deviations of the resonance wavelength from the original design [114] and requires thermal tuning, hence degrading the energy efficiency of the link. Various wavelength locking schemes based on analog and digital feedback [115], bit statistics [112], and pulse width modulation and thermal rectification [111] have been proposed and implemented to overcome the unwanted variations due to the fabrication. (4) Backscattering: In applications where narrow optical linewidths (i.e., $Q > 10{,}000$) are required, even a slight roughness on the sidewalls of the MRRs will cause back reflections inside the ring [116]. The effect of backscattering in MRRs is typically observed in the form of a splitting of the resonance in the spectral response [117]. This spectral distortion adds extra complexity to the design of optical links and further narrows the design space of MRRs [118].

18.3.5 Energy-efficient photonic links

In this section we present a design exploration for short-reach silicon photonic links using MRRs and filters. We seek to obtain the maximum achievable aggregate bandwidth. This is achieved by analyzing the impacts due to the induced impairments and translating them into power penalties, the extra optical power required to compensate for the effects of such impairments on the bit error ratio performance of the system

[98,119–122]. We also show how the spectral statistics of modulated light changes as it travels through a ring demultiplexer and how the changes can be become a power penalty.

We concentrate our efforts on MRR-based links, as they offer the highest bandwidth density and most energy-efficient performance among current silicon photonic interconnect devices [123–125]. Due to their small size, multiple microrings can be placed along a single waveguide on chip, facilitating a dense WDM design [126,127]. However, WDM links may suffer from spectral degradation of channels and interchannel crosstalk [128–131]. These impairments eventually set an upper limit on both the number of channels and the modulation speed of each channel, thus placing an upper bound on the aggregate rate to the link [132,133].

Consider a simple chip-to-chip silicon photonic link as shown in Fig. 18.14. Microrings are placed along an on-chip waveguide to modulate the incoming multiwavelength light generated by a comb laser source [87]. The incoming wavelengths, once imprinted with data, are then transmitted through an optical waveguide to a receiver chip. The receiver chip consists of multiple passive microrings with resonances tuned to the channel wavelengths. The total capacity of this link is obtained by multiplying the number of channels N_λ with the modulation bit rate r_b.

Intuitively, it is tempting to maximize the number of wavelengths and/or to choose higher bit rates for each channel. This allows for higher utilization of the available spectrum in the transmission media. However, as the number of wavelengths and/or the bit rate grows, crosstalk between channels and other undesired impairments emerge, which eventually prevent a reliable transmission through the link. Therefore, the total capacity of the link is closely tied to the optical power losses and other unde-

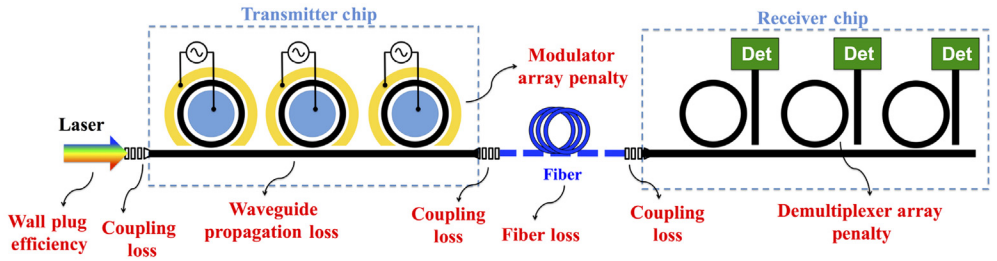

Figure 18.14 Chip-to-chip silicon photonic interconnect with an MRR-based wavelength division multiplexing (WDM) link. The optical interface of the transmitter chip includes MRR modulators that use carrier dispersion for high-speed modulation. The optical interface at the receiver includes demultiplexing filters, photodetectors, and electronic decision circuitry (Det: detector). Wall-plug efficiency corresponds to the electrical to optical power conversion of the laser. Source: *From M. Bahadori, S. Rumley, D. Nikolova, K. Bergman, Comprehensive design space exploration of silicon photonic interconnects, J. Lightw. Technol. 34 (12) (2016) 2975–2987 [97].*

sired optical impairments through the entire link. Summing up all the power penalties of the link, PP^{dB}, for a single channel, the following inequality must hold [134]:

$$\left[P_{\mathrm{laser}}^{\mathrm{dBm}} - 10\log 10(N\lambda)\right] - P_{\mathrm{sensitivity}}^{\mathrm{dBm}} \geq PP^{\mathrm{dB}} \tag{20.1}$$

In general, aggregated optical power P_{laser} (sum over all wavelengths) must stay below the nonlinear threshold of the silicon waveguides at any point of the link [133,135]. On the other hand, the signal powers should stay above the sensitivity of the detectors $P_{\mathrm{sensitivity}}$ (minimum number of photons or equivalently a certain amount of optical power) at the receive side. A typical receiver may have a sensitivity of -12.7 dBm at 8 Gb/s operation [120], while a good receiver may exhibit a sensitivity to -21 dBm at 10 Gb/s [119]. The difference between these higher and lower thresholds can be exploited to find the maximum power budget. This budget accounts for the power penalty, PP^{dB}, per channel over the N_λ channels. We will show that the power impairments induced by the microrings depend on the channel spacing, which is inversely proportional to the number of channels, and on the modulation rate.

Here we present a study on the minimal energy consumption for 200–800 Gbps data rate aggregation on the link, as indicated in Fig. 18.15A, based on the optimization of the physical parameters of the microring. The rings are configured to operate at their critical coupling point. The results in the figure show that the smaller ring radius in the $\sim 7\,\mu$m regime leads to the best energy performance of the link. Fig. 18.15B shows the breakdown of the factors contributing to energy consumption. At higher data rates, the laser power consumption becomes critical, but the energy consumed by the static thermal tuning declines. Finally, Fig. 18.15C provides details of the number of channels and the required data rate per channel that lead to the minimal energy consumption for the target aggregation rate.

When the available optical power budget is fully utilized, the study also investigates the maximum aggregation rate based on the product of the number of channels and the optical data rate of each channel. Fig. 18.15D indicates a maximum possible aggregation rate of ~ 800 Gbps at 15 Gbps data rate per channel. For each data rate, its associated energy efficiency is also plotted. Note that the lowest energy efficiency is not associated with the highest aggregation rate. This further reiterates the fact that designing a silicon photonic link requires a trade-off between energy consumption and high-speed performance. We note that the results depend on the ring resonator parameters and vary if different parameters are used. In addition, the losses in the ring have significant impact on the maximum aggregation rate. The details for the model of the ring resonators associated with these results can be found in Ref. [107].

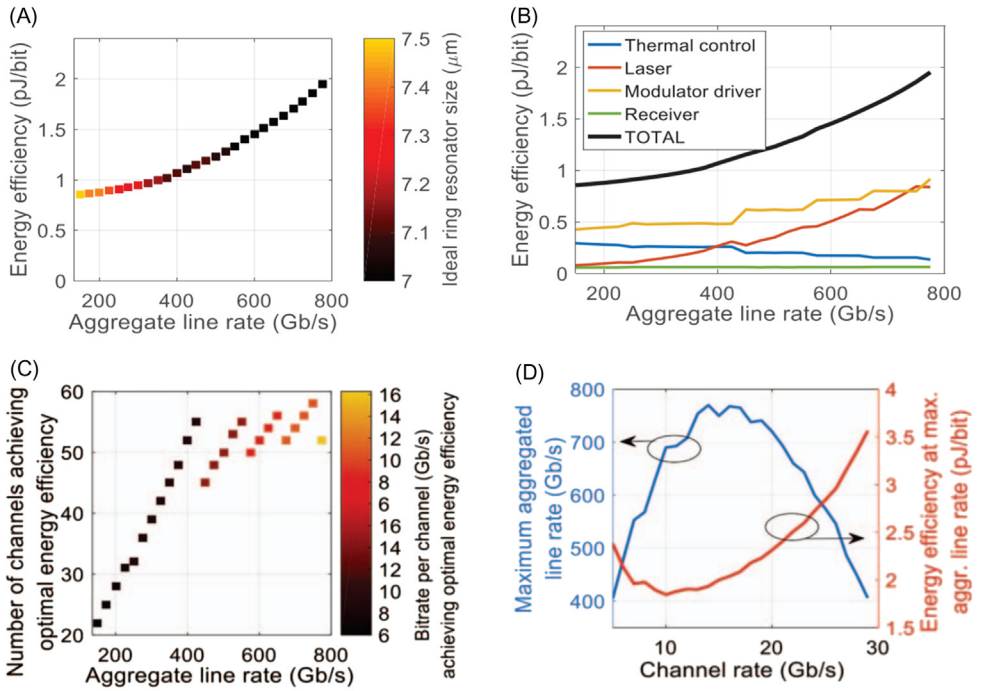

Figure 18.15 (A) Minimum energy consumption of the link for given aggregations based on the optimum value for the ring radius. (B) Breakdown of energy consumption. (C) Breakdown of the number of channels and the required data rate per channel for minimum energy consumption. (D) Evaluation of the maximum supported aggregation and the associated energy consumption for various channel rates Source: *From M. Bahadori, S. Rumley, R. Polster, A. Gazman, M. Traverso, M. Webster, et al., Energy-performance optimized design of silicon photonic interconnection networks for high-performance computing, in: Proceedings of the Conference on Design, Automation & Test in Europe, European Design and Automation Association: Lausanne, Switzerland, 2017, pp. 326–331 [136].*

18.4 Bandwidth steering

In this section, we first briefly survey optical switching technologies, then discuss optimized photonic links with optical switches and introduce the Flexfly architecture—a Dragonfly design that is capable of reconfiguring its bandwidth to match traffic patterns by using low-radix silicon photonics switches [40]. Optical switches are an important component in modern high-speed telecommunications. The advantages they can provide in terms of energy efficiency and high bandwidth density, particularly as cost is reduced through integrated silicon photonics, are an important subject of research and development for computing systems.

18.4.1 Free-space optical switches

Numerous competing optical switching approaches based on free-space technology have been commercially realized, including microelectromechanical systems (MEMS) [137], beam-steering [138], and liquid crystal on silicon [139]. Among these, MEMS-based optical switches are the most common free-space optical switching devices. An electrostatic driver is commonly used because of its low power consumption and ease of control; however, a typical voltage up to 100–150 V is required [61,140]. MEMS spatial switches can be realized in both two-dimensional (2D) and three-dimensional (3D) configurations. The crossbar topology is normally implemented in the 2D configuration with digital operation using a bistable mirror position. 3D MEMS switches, which are assembled using 2D input and output fiber arrays with collimators, have been proposed to support very large-scale optical cross-connect devices [141]. Two stages of independent 2D micromirror arrays with a two-axis tilting structure [137] are used to steer the optical beams in three dimensions.

MEMS switches can support connectivity of hundreds of ports [137,142]; however, the installation and calibration with surface-normal micro optics introduces considerable complexity that is ultimately reflected in the cost per port. This cost remains a challenge for the implementation of MEMS switches in high-performance systems.

18.4.2 Photonic integrated switches

In order to be adopted in high-performance systems, optical switching technologies must demonstrate a path toward high-volume manufacture and ensure low cost per port. This leads to the consideration of lithography-based fabrication and high-level integration. Here we present a brief overview of the switching technologies based on III–V and silicon platforms.

In the integrated devices under consideration, different physical mechanisms have been investigated in order to realize the optical switching process. Physical properties used include phase manipulation through thermal or electrical control in interferometric structures, that is, MZI and MRR, signal amplification/absorption in semiconductor optical amplifiers (SOAs), and MEMS-actuated coupling between different layers of waveguides. In the last decade, photonic integration technologies have quickly matured to realize monolithic integrated circuits of a few thousands of components with increasingly sophisticated functionalities. Notable demonstrations of monolithic switch fabrics are summarized in Fig. 18.16.

InP-based switch fabrics have primarily employed SOA gated elements in the broadcast and select (B&S) topology. B&S networks utilize passive splitters/combiners with each path gated by an SOA element, which can provide chip-level multicast [143]. However, the optical loss due to the various signal splits and recombinations discourage scaling this architecture beyond 4×4 connectivity. As an alternative,

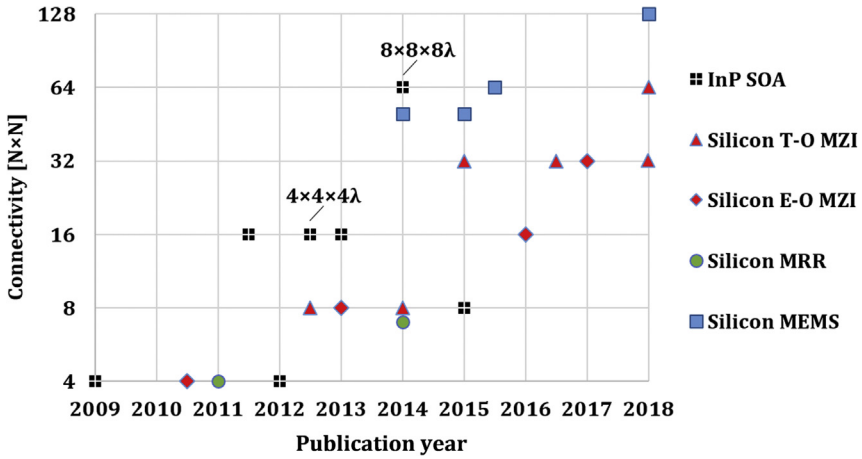

Figure 18.16 High connectivity optical switch matrix technologies shown in terms of input side connectivity. Source: Q. Cheng, M. Bahadori, M. Glick, S. Rumley, and K. Bergman, "*Recent advances in optical technologies for data centers: a review*," Optica 5, 1354–1370 (2018) [1].

multistage architectures using cascaded switching elements have been proposed [144]. 16×16 port count SOA-based switches have been demonstrated using both all-active [145] and passive-active [146] integration schemes. Higher on-chip connectivity, scaling up to 64×64 connections, has been achieved by combining spatial ports with wavelength channels using co-integrated AWGs [147]. Further scaleup would require a large reduction in component-level excess loss, a more careful design of balancing the summed loss and gain per stage, and a close examination of SOA designs for linear operation [148,149].

In addition to InP based switches, the highly advanced CMOS industry with mature fabrication infrastructures and advances in silicon photonics have stimulated the development of silicon-based optical switches. The current record for a monolithic photonic switch radix is 64×64 by a thermo-optic MZI-based Beneš switch [150] and the very recent 240×240 by a MEMS-actuated crossbar switch [151]. Other notable advances include a 32×32 thermally actuated PILOSS MZI switch with $<$13.2 dB insertion loss [152] and a 32×32 electro-optic MZI-based Beneš switch [153].

To support the scaling of integrated photonic switches for high-performance systems, the switch architecture should be reexamined in terms of crosstalk cancellation, the number of cascaded switch stages, and the total number of switch cells, as signal degradation is introduced from accumulated crosstalk and loss exacerbates with increased cascaded stages. We demonstrate an MRR-based modified switch-and-select (S&S) switching circuit with the concept illustrated in Fig. 18.17A and B, where the $1 \times N$ switch unit is built from MRR add-drop cells assembled in a bus coupled

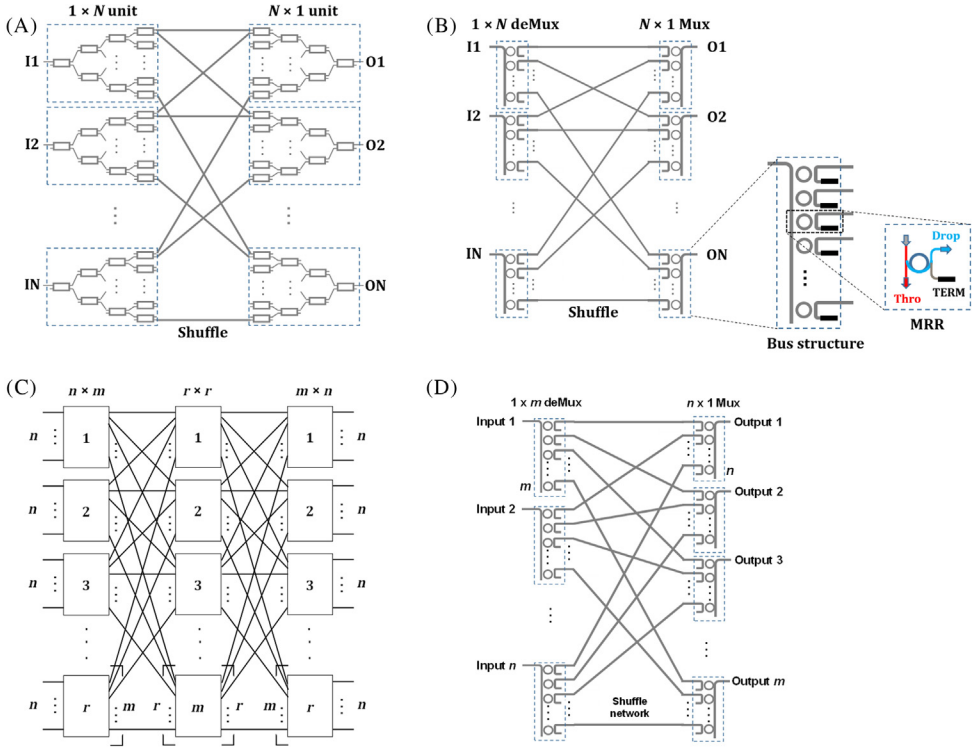

Figure 18.17 (A) Switch-and-select topology with MZI elements arranged in a cascaded structure. (B) Modified switch-and-select topology with MRR based spatial (de)multiplexers. (C) Layout of a generic three-stage Clos network building from r n × m, m r × r, and r m × n blocks. (D) Schematic of an n × m microring-based block in the switch-and-select topology. Source: *From (A-B) Q. Cheng, L. Y. Dai, N. C. Abrams, Y. Hung, P. E. Morrissey, M. Glick, P. O'Brien, and K. Bergman, "*Ultralow-crosstalk, strictly non-blocking microring-based optical switch*," Photon. Res. 7, 155–161 (2019)(C-D) Q. Cheng, M. Bahadori, Y. Hung, Y. Huang, N. Abrams and K. Bergman, "*Scalable Microring-Based Silicon Clos Switch Fabric with Switch-and-Select Stages*," in IEEE Journal of Selected Topics in Quantum Electronics. Doi: 10.1109/JSTQE.2019.2911421 [157]*

structure [154]. Scaling such a structure requires only adding MRRs to the bus waveguide, which effectively reduces the scaling overhead in loss compared to that of the cascaded scheme. The layout of a generic $N \times N$ S&S MRR based switch is depicted in Fig. 18.17B. This configuration has N input spatial $1 \times N$ and N output spatial $N \times 1$ units, maintaining the number of drop microrings at two in any path, therefore the first-order crosstalk is fully blocked. We performed further studies on combining the scalable three-stage Clos network with populated S&S stages [155,156], as shown in Fig. 18.17C and D. The proposed design offers a balance that keeps the number of stages to a modest value while largely reducing the required number of switching elements. The scalability is predicted to be 128×128 [155,156].

Prototyped devices have been fabricated at the American Institute of Manufacturing (AIM photonics) foundry and all designs used the predefined elements in the PDK library to ensure high yield with low cost. The design rules of standard packaging houses, for example, Tyndall National Institute, are employed to achieve low-cost packaging solutions. Fig. 18.18A, B, and C shows the microscope photo of a 4×4 Si/SiN dual-layer S&S switch, a 4×4 silicon S&S switch, and a 12×12 silicon Clos switch, respectively. All devices have been fully packaged with thermo-optical MRRs in use. Small radix switches can be packaged using the QFN-type socket, and directly wire-bonded or flip-chip bonded to a PCB breakout board with a UV-cured fiber array, as shown in Fig. 18.19A, B, and C, respectively. For densely integrated Clos switches, a packaging platform was developed with a silicon interposer (as shown by Fig. 18.18D) as an electrical redistribution layer for an ultra-compact package with low insertion loss. Fig. 18.19D shows the packaged 12×12 Clos switch, which was first flip-chip bonded onto a silicon interposer and then wire-bonded to a PCB breakout board. Excellent testing results were achieved for the fully packaged 4×4 S&S switch with on-chip loss and crosstalk ratio as low as 1.8 and −50 dB, respectively [154,158].

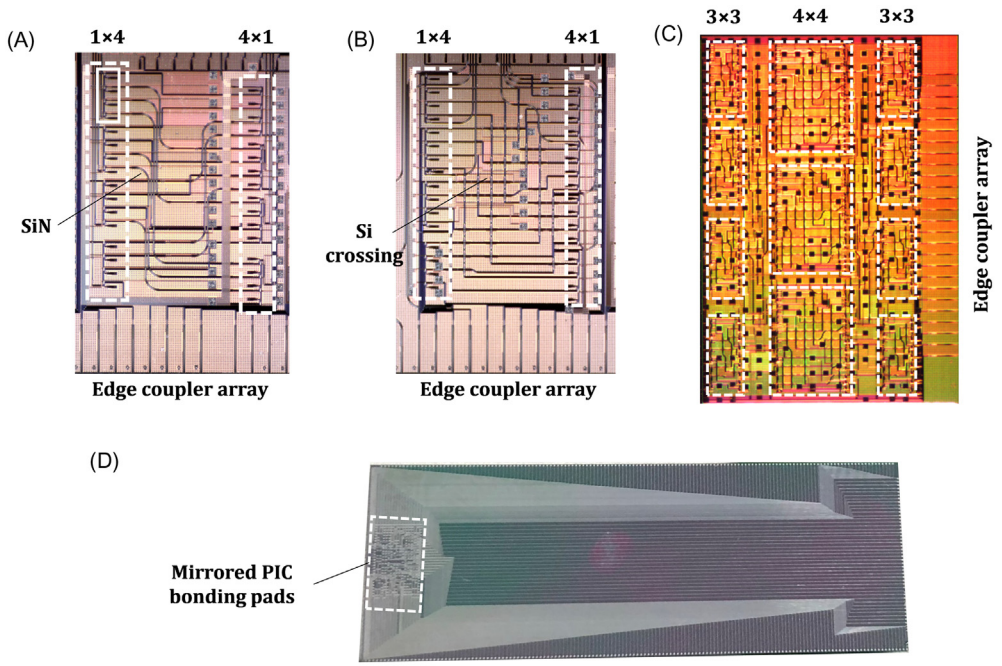

Figure 18.18 Microscope photo of (A) 4×4 Si/SiN dual-layered MRR-based S&S switch, (B) 4×4 Si MRR-based S&S switch, and (C) 12×12 Si MRR-based Clos switch with populated S&S stages. (D) Silicon interposer for the 12×12 Clos switch.

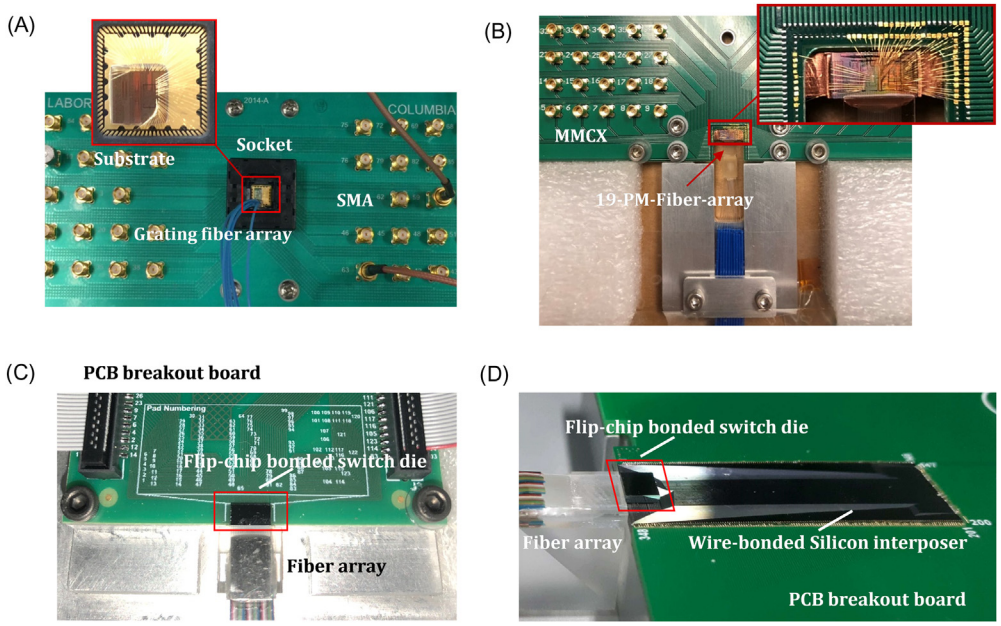

Figure 18.19 (A) QFN type package with socket. (B) Packaged switch device by wire bonding to the PCB breakout board. (C) Packaged switch device by flip-chip bonding on the PCB breakout board. (D) Packaged switch device with silicon interposer. Source: *(B) From Q. Cheng, M. Bahadori, Y. Hung, Y. Huang, N. Abrams and K. Bergman, "*Scalable Microring-Based Silicon Clos Switch Fabric with Switch-and-Select Stages*," in IEEE Journal of Selected Topics in Quantum Electronics. Doi: 10.1109/JSTQE.2019.2911421. (C) From Q. Cheng, L. Y. Dai, N. C. Abrams, Y. Hung, P. E. Morrissey, M. Glick, P. O'Brien, and K. Bergman, "*Ultralow-crosstalk, strictly non-blocking microring-based optical switch", *Photon. Res. 7, 155–161 (2019). (D) From Q. Cheng, M. Bahadori, M. Glick, S. Rumley, and K. Bergman, "*Recent advances in optical technologies for data centers: a review," *Optica 5, 1354–1370 (2018) [158].*

Looking forward, we envision a new class of III–V/Si heterogeneously integrated optical switches leveraging advanced bonding techniques to provide compact, energy-efficient, and low-cost switch fabrics that satisfy the high-performance system metrics [61], where lossless design would be a significant advantage. The implementation can follow the approach demonstrated in the InP MZI-SOA switch fabrics [159–161] or be combined with the S&S topology of MRR add-drop multiplexers. Detailed discussions are found in Ref. [61].

18.4.3 Network performance

In today's computing systems, the high bandwidth densities that optics is capable of are not fully leveraged. Implementing optical interconnects for a system with 10 K computing nodes entirely with optical cables and optical switches remains cost-

prohibitive. Currently, network topologies implemented in computing systems use electrical packet switches and optical cables for longer reach links (>1 m). Supercomputers are often perceived as the first adopters in the market for innovative optical technologies (as occurred, for instance, with AOCs). In the supercomputer, where torus topologies previously dominated, high-radix packet switches have made hierarchical topologies such as the Dragonfly and the fat tree more popular. The Dragonfly topology [41,162] provides high connectivity with all-to-all global links at the intergroup level, and aims at minimizing the number of long distance links to reduce the cost. However, the advantages of high connectivity are diluted by low per link bandwidth. The bandwidth of intergroup (global) links, carrying the traffic between Dragonfly groups, can become the bottleneck for an entire network.

A major reason for this bandwidth bottleneck is due to the highly skewed traffic characteristics of HPC applications. These traffic patterns concentrate traffic on only a small percentage of links, so that only a few links are congested while most others are severely underutilized. Thus, the current, best-for-all approach using static, over-provisioned networks have topologies that are mismatched with the applications that operate over them, which will likely become a bottleneck for the next-generation Exaflop platforms [42–44,163].

Designing networks that properly balance traffic is challenging: over-provisioning the network incurs unnecessary cost and energy [164], while under-provisioning leads to limitations on system performance due to data-starved processors. In this section we present results from a study on Flexfly [40] a photonic architecture that trades global links among dragonfly groups using low-radix silicon photonic switches, allowing the network topology to be dynamically reconfigured to match HPC application traffic. In Flexfly the global links initially defining the all-to-all topology can be taken from their original destination groups and reassigned to traffic-intensive ones. By trading the global links in this way, Flexfly creates additional direct bandwidth for intensively communicating group pairs where and when it is needed. This is illustrated in Fig. 18.20, showing an all-to-all Dragonfly topology being reconfigured to a bandwidth-steered topology that focuses on maximizing bandwidth between neighboring groups. It achieves such reconfigurability through the use of transparent silicon photonic circuit switching. Flexfly is designed to support the use of low-radix optical switches, realizable through low-cost fabrication technologies. Simulations on the Flexfly architecture with applications such as GTC, Nekbone, and LULESH show up to $1.8\times$ speedup over the Dragonfly topology paired with UGAL routing. The hop count and cross-group message latency are also halved compare to the Dragonfly topology.

An experimental 32-node Flexfly prototype was built with four groups connected through a silicon photonic switch [56]. The interconnect reconfiguration time was 820 ns. The network architecture consisting of the control and data planes is shown in Fig. 18.21. The control plane is an Ryu-based SDN controller, and each rack sends a

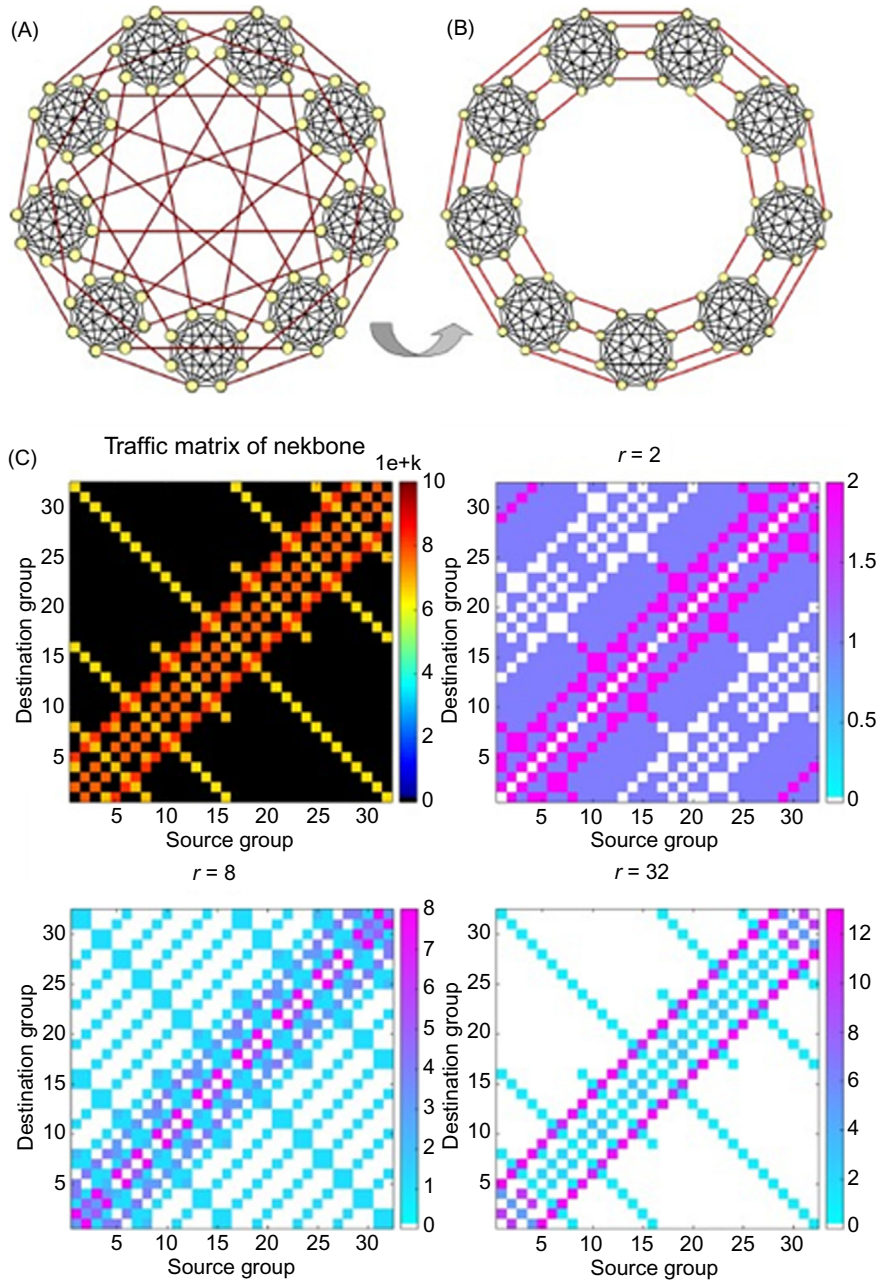

Figure 18.20 (A) Regular Dragonfly topology with all-to-all intergroup links. (B) Reorganized topology after bandwidth steering using optical switching. (C) Top-left matrix shows traffic distribution across pairs of Dragonfly groups during execution of Nekbone workload; other matrices show the network topology for different silicon photonic switch radices. Source: *From J. Wilke, Bringing minimal routing back to HPC through silicon photonics: a study of "flexfly" architectures with the structural simulation toolkit (SST), in: Proceedings of the 2nd International Workshop on Advanced Interconnect Solutions and Technologies for Emerging Computing Systems, ACM, Stockholm, Sweden, 2017, p. 5-5 [165].*

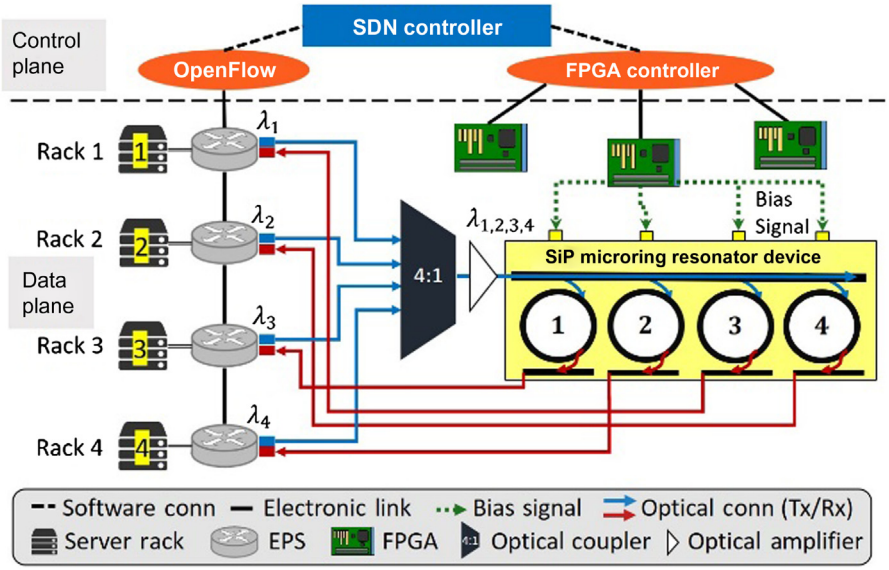

Figure 18.21 Network architecture showing the connections of the servers and top-of-rack electronic packet switches (EPSs) to the SiP MRR device. Source: *From Y. Shen, A. Gazman, Z. Zhu, M.Y. Teh, M. Hattink, S. Rumley, et al., Autonomous dynamic bandwidth steering with silicon photonic-based wavelength and spatial switching for Datacom networks, in: Optical Fiber Communication Conference, Optical Society of America, 2018 [56].*

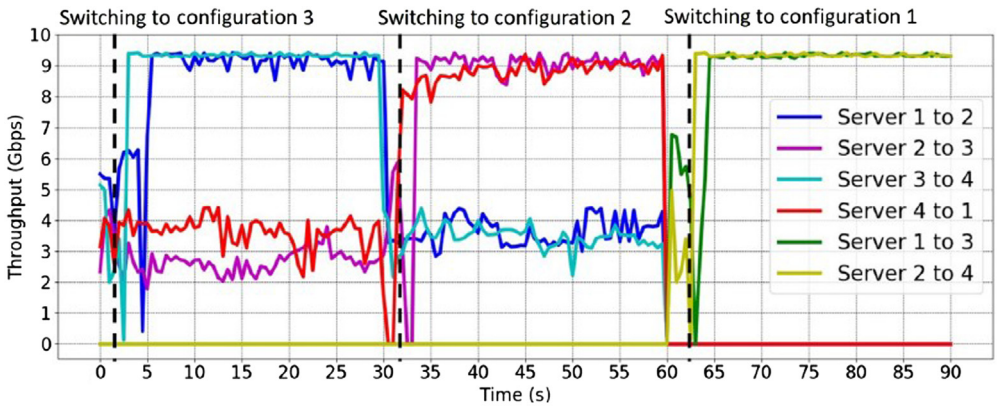

Figure 18.22 Throughput of various intergroup flows over time demonstrating control plane bandwidth steering capabilities. Source: *From Y. Shen, A. Gazman, Z. Zhu, M.Y. Teh, M. Hattink, S. Rumley, et al., Autonomous dynamic bandwidth steering with silicon photonic-based wavelength and spatial switching for Datacom networks, in: Optical Fiber Communication Conference, Optical Society of America, 2018 [56].*

unique wavelength through the electronic packet switches (EPSs). By tuning the rings in different ways, the device allows different servers to be connected bidirectionally and thus act as a wavelength and spatial optical circuit switch.

The results of the bandwidth steering are shown in Fig. 18.22. When traffic is added between servers, the control plane detects the changes and initiates the configuration updates. The configurations provide a direct connection for traffic between particular servers, so that they are able to reach near full link capacity, while the other flows must compete with the background traffic and are therefore limited in throughput. The results demonstrate that Flexfly-based optically switched networks are a promising solution for improved network efficiency and resource allocation.

18.5 Conclusions

Looking forward we see several factors pointing toward excellent opportunities for optical interconnection networks to be increasingly deployed in high-performance systems. Although the details of the architectures of the supercomputer and warehouse scale data center are different, currents trends are leading to similar scaling challenges. The slowing of Moore's law leads to more parallelism and a greater focus on energy efficiency. Photonic solutions are a natural fit for increased parallelism and as data rates get higher and are at an advantage in the energy/bit metric. The need for networks with higher and higher bandwidth is exacerbated by new applications using machine learning algorithms and data analytics. By their nature, these applications require much greater storage capabilities and are communication intensive. Communication-intensive calculations using large amounts of data also favor the photonic interconnects' high bandwidth capabilities. With an increased emphasis on cost savings and energy efficiency, network architects are looking to get the highest utilization out of the equipment in the network. This is leading to a re-architecting of the topologies toward adaptability and reconfiguration to more efficiently match traffic and application requirements. Techniques such as bandwidth steering using low-radix optical switches can efficiently reconfigure the topology to place the bandwidth where required to meet the application. At the same time as the needs are expanding, there are major advances being made to lower the cost of photonic-based circuits through the use of CMOS-compatible silicon photonics and leveraging manufacturing technology used in the electronics industry to reduce the cost of fabrication and packaging. This combination of an avenue toward reduced cost and greater needs suited to photonics capabilities are expected to lead to major advances in research and deployment of photonic interconnection networks in the near future.

References

[1] Q. Cheng, M. Bahadori, M. Glick, S. Rumley, K. Bergman, Recent advances in optical technologies for data centers: a review, Optica 5 (2018) 1354–1370.
[2] Christos A. Thraskias, et al., Survey of photonic and plasmonic interconnect technologies for intra-datacenter and high-performance computing communications, IEEE Communications Surveys & Tutorials 20 (4) (2018) 2758–2783.
[3] D.A. Reed, J. Dongarra, Exascale computing and big data, Commun. ACM 58 (7) (2015) 56–68.
[4] J. Shalf, S. Dosanjh, J. Morrison. Exascale computing technology challenges, in: International Conference on High Performance Computing for Computational Science, Springer, 2010.
[5] R. Meisner, A Platform Strategy for the Advanced Simulation and Computing Program. NA-ASC-113R-07-Vol. 1-Rev. 0, 2007.
[6] Cisco Global Cloud Index: Forecast and Methodology, 2016–2021 White Paper. Available from: <https://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.html>.
[7] M.M. Waldrop, The chips are down for Moore's law, Nat. News 530 (7589) (2016) 144.
[8] G. Michelogiannakis, J. Shalf, D. Donofrio, J. Bachan, Continuing the Scaling of Digital Computing Post Moore's Law. 2016.
[9] Ken Giewont, et al., 300mm Monolithic Silicon Photonics Foundry Technology, IEEE Journal of Selected Topics in Quantum Electronics (2019).
[10] C. Sun, M.T. Wade, Y. Lee, J.S. Orcutt, L. Alloatti, M.S. Georgas, et al., Single-chip microprocessor that communicates directly using light, Nature 528 (2015) 534.
[11] D.E. Nikonov, I.A. Young, Benchmarking of beyond-CMOS exploratory devices for logic integrated circuits, IEEE J. Explorat. Solid-State Comput. Dev. Circuits 1 (2015) 3–11.
[12] DeepMind AI Reduces Google Data Centre Cooling Bill by 40%. Available from: <https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/>.
[13] N.P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, et al., In-datacenter performance analysis of a tensor processing unit, in: Computer Architecture (ISCA), 2017 ACM/IEEE 44th Annual International Symposium on, EEE, 2017.
[14] K. Hazelwood, S. Bird, D. Brooks, S. Chintala, U. Diril, D. Dzhulgakov, et al., Applied machine learning at facebook: a datacenter infrastructure perspective, in: High Performance Computer Architecture (HPCA), 2018 IEEE International Symposium on, IEEE, 2018.
[15] N.M. Nasrabadi, Pattern recognition and machine learning, J. Electr. imaging 16 (4) (2007) 049901.
[16] N. Farrington, A. Andreyev, Facebook's data center network architecture, in: Optical Interconnects Conference, 2013 IEEE. Citeseer, 2013.
[17] S. Kanev, J.P. Darago, K. Hazelwood, P. Ranganathan, T. Moseley, G.-Y. Wei, et al., Profiling a warehouse-scale computer, in: ACM SIGARCH Computer Architecture News, ACM, 2015.
[18] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, et al., Accurate, large minibatch SGD: training imagenet in 1 hour. arXiv preprint arXiv:1706.02677, 2017.
[19] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems. 2012, pp. 1097–1105.
[20] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
[21] Training Recurrent Neural Networks at Scale. Available from: <https://www.slideshare.net/SessionsEvents/erich-elsen-research-scientist-baidu-research-at-mlconf-nyc-41516>.
[22] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436.
[23] Y. Shen, N.C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, et al., Deep learning with coherent nanophotonic circuits, Nat. Photon. 11 (2017) 441.
[24] J. Hines, Stepping up to Summit, Comput. Sci. Eng. 20 (2) (2018) 78–82.
[25] D. Schneider, US supercomputing strikes back, IEEE Spectrum 55 (1) (2018) 52–53.
[26] K. Bergman, Silicon photonics for high performance interconnection networks, in: Optical Fiber Communication Conference, Optical Society of America, 2018.

[27] S. Rumley, M. Bahadori, R. Polster, S.D. Hammond, D.M. Calhoun, K. Wen, et al., Optical interconnects for extreme scale computing systems, Parallel Comput. 64 (2017) 65–80.
[28] Top 500. Available from: <https://www.top500.org/.>.
[29] The Green 500. Available from: <https://www.top500.org/green500/>.
[30] Graph 500. Available from: <https://graph500.org/>.
[31] B.E. Floren, Optical interconnects in the Touchstone supercomputer program, in: Integrated Optoelectronics for Communication and Processing. International Society for Optics and Photonics, 1992.
[32] D.H. Hartman, Digital high speed interconnects: a study of the optical alternative, Opt. Eng. 25 (10) (1986) 251086.
[33] J.W. Goodman, F.J. Leonberger, S.-Y. Kung, R.A. Athale, Optical interconnections for VLSI systems, Proc. IEEE 72 (7) (1984) 850–866.
[34] A. Louri, H. Sung, An optical multi-mesh hypercube: a scalable optical interconnection network for massively parallel computing, J. Lightw. Technol. 12 (4) (1994) 704–716.
[35] J.A. Kash, Leveraging optical interconnects in future supercomputers and servers, in: High Performance Interconnects, 2008. HOTI'08. 16th IEEE Symposium on, IEEE, 2008.
[36] A.F. Benner, M. Ignatowski, J.A. Kash, D.M. Kuchta, M.B. Ritter, Exploitation of optical interconnects in future server architectures, IBM J. Res. Dev. 49 (4.5) (2005) 755–775.
[37] M.J. Kobrinsky, B.A. Block, J.-F. Zheng, B.C. Barnett, E. Mohammed, M. Reshotko, et al., On-chip optical interconnects, Intel Technol. J. 8 (2) (2004).
[38] C. Gunn, CMOS photonics for high-speed interconnects, IEEE Micro 26 (2) (2006) 58–66.
[39] A. Shacham, K. Bergman, L.P. Carloni, Maximizing GFLOPS-per-Watt: high-bandwidth, low power photonic on-chip networks, in: P = ac2 Conference, Citeseer, 2006.
[40] K. Wen, P. Samadi, S. Rumley, C.P. Chen, Y. Shen, M. Bahadori, et al. Flexfly: enabling a reconfigurable dragonfly through silicon photonics, in: High Performance Computing, Networking, Storage and Analysis, SC16: International Conference for, IEEE, 2016.
[41] J.H. Ahn, N. Binkert, A. Davis, M. Mclaren, R.S. Schreiber, HyperX: topology, routing, and packaging of efficient large-scale networks, in: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, ACM, 2009.
[42] B. Arimilli, R. Arimilli, V. Chung, S. Clark, W. Denzel, B. Drerup, et al., The PERCS high-performance interconnect, in: High Performance Interconnects (HOTI), 2010 IEEE 18th Annual Symposium on, IEEE, 2010.
[43] M. Besta, T. Hoefler, Slim fly: a cost effective low-diameter network topology, in: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, IEEE, 2014.
[44] S. Rumley, D. Nikolova, R. Hendry, Q. Li, D. Calhoun, K. Bergman, Silicon photonics for exascale systems, J. Lightw. Technol. 33 (3) (2015) 547–562.
[45] D. a B. Miller, H. Ozaktas, Limit to the bit-rate capacity of electrical interconnects from the aspect ratio of the system architecture, J. Parallel Distrib. Comput. 41 (1) (1997) 42–52.
[46] D.A. Miller, Rationale and challenges for optical interconnects to electronic chips, Proc. IEEE 88 (6) (2000) 728–749.
[47] T.O. Dickson, Y. Liu, A. Agrawal, J.F. Bulzacchelli, H.A. Ainspan, Z. Toprak-Deniz, et al., A 1.8 pJ/bit $16\times 16\;\text {Gb/s} $ Source-synchronous parallel interface in 32 nm SOI CMOS with receiver redundancy for link recalibration, IEEE J. Solid-State Circ. 51 (8) (2016) 1744–1755.
[48] T.O. Dickson, Y. Liu, S.V. Rylov, A. Agrawal, S. Kim, P.-H. Hsieh, et al., A 1.4 pJ/bit, power-scalable 16 × 12 Gb/s source-synchronous I/O with DFE receiver in 32 nm SOI CMOS technology, IEEE J. Solid-State Circ. 50 (8) (2015) 1917–1931.
[49] J.W. Poulton, W.J. Dally, X. Chen, J.G. Eyles, T.H. Greer, S.G. Tell, et al., A 0.54 pJ/b 20 Gb/s ground-referenced single-ended short-reach serial link in 28 nm CMOS for advanced packaging applications, J. Solid-State Circ. 48 (12) (2013) 3206–3218.
[50] A. Shokrollahi, D. Carnelli, J. Fox, K. Hofstra, B. Holden, A. Hormati, et al. 10.1 A pin-efficient 20.83 Gb/s/wire 0.94 pJ/bit forwarded clock CNRZ-5-coded SerDes up to 12mm for MCM packages in 28nm CMOS, in: Solid-State Circuits Conference (ISSCC), 2016 IEEE International, IEEE, 2016.

[51] K. Bergman, J. Shalf, T. Hausken, Optical interconnects and extreme computing, Optics and Photonics News 27 (4) (2016) 32–39.
[52] K. Wen, S. Rumley, P. Samadi, C.P. Chen, K. Bergman, Silicon photonics in post Moore's Law era: technological and architectural implications, in: Post-Moore's Era Supercomputing (PMES) Workshop, Salt Lake City, IEEE, 2016.
[53] M. Ghobadi, R. Mahajan, A. Phanishayee, N. Devanur, J. Kulkarni, G. Ranade, et al., ProjecToR: agile reconfigurable data center interconnect, in: Proceedings of the 2016 ACM SIGCOMM Conference, ACM, Florianopolis, Brazil, 2016, pp. 216–229.
[54] T. Benson, A. Akella, D.A. Maltz, Network traffic characteristics of data centers in the wild, in: Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, ACM, 2010.
[55] A. Roy, H. Zeng, J. Bagga, G. Porter, A.C. Snoeren, Inside the social network's (datacenter) network, in: ACM SIGCOMM Computer Communication Review, ACM, 2015.
[56] Y. Shen, A. Gazman, Z. Zhu, M.Y. Teh, M. Hattink, S. Rumley, et al., Autonomous dynamic bandwidth steering with silicon photonic-based wavelength and spatial switching for Datacom networks, in: Optical Fiber Communication Conference, Optical Society of America, 2018.
[57] https://cloud.google.com/solutions/designing-connected-vehicle-platform accessed April 16, 2019.
[58] R. Urata, H. Liu, L. Verslegers, C. Johnson, Silicon photonics technologies: Gaps analysis for data-center interconnects, Silicon Photonics III., Springer, 2016, pp. 473–488.
[59] R. Urata, H. Liu, X. Zhou, A. Vahdat, Datacenter interconnect and networking: from evolution to holistic revolution, in: Optical Fiber Communications Conference and Exhibition (OFC), 2017, IEEE, 2017.
[60] A. Chakravarty, K. Schmidtke, V. Zeng, S. Giridharan, C. Deal, R. Niazmand, 100Gb/s CWDM4 optical interconnect at facebook data centers for bandwidth enhancement, in: Laser Science, Optical Society of America, 2017.
[61] Q. Cheng, S. Rumley, M. Bahadori, K. Bergman, Photonic switching in high performance data-centers [invited], Opt. Express 26 (12) (2018) 16022–16043.
[62] N. Farrington, G. Porter, S. Radhakrishnan, H.H. Bazzaz, V. Subramanya, Y. Fainman, et al., Helios: a hybrid electrical/optical switch architecture for modular data centers, ACM SIGCOMM Comput. Commun. Rev. 40 (4) (2010) 339–350.
[63] X. Zhou, H. Liu, R. Urata, Datacenter optics: requirements, technologies, and trends (Invited Paper), Chin. Opt. Lett. 15 (5) (2017) 120008.
[64] A. Singh, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R. Bannon, et al., Jupiter rising: a decade of clos topologies and centralized control in Google's datacenter network, Commun. ACM 59 (9) (2016) 88–97.
[65] C.F. Lam, Optical network technologies for datacenter networks (invited paper), in: 2010 Conference on Optical Fiber Communication (OFC/NFOEC), collocated National Fiber Optic Engineers Conference, 2010.
[66] Moving data with light. Available from: <https://www.intel.com/content/www/us/en/architecture-and-technology/silicon-photonics/silicon-photonics-overview.html>.
[67] T. Rokkas, I. Neokosmidis, B. Shariati, I. Tomkos, Techno-economic evaluations of 400G optical interconnect implementations for datacenter networks, in: Optical Fiber Communication Conference, Optical Society of America, 2018.
[68] A. Ghiasi, Large data centers interconnect bottlenecks, Opt. Express 23 (3) (2015) 2085–2090.
[69] R.H. Johnson, D.M. Kuchta, 30 Gb/s directly modulated 850 nm datacom VCSELs, in: Conference on Lasers and Electro-Optics/Quantum Electronics and Laser Science Conference and Photonic Applications Systems Technologies, Optical Society of America, San Jose, CA, 2008.
[70] M. Filer, B. Booth, D. Bragg, The role of standards for cloud-scale data centers, in: Optical Fiber Communication Conference, Optical Society of America, San Diego, CA, 2018.
[71] D. Thomson, A. Zilkie, J.E. Bowers, T. Komljenovic, G.T. Reed, L. Vivien, et al., Roadmap on silicon photonics, J. Opt. 18 (7) (2016) 073003.
[72] E.R.H. Fuchs, R.E. Kirchain, S. Liu, The future of silicon photonics: not so fast? Insights from 100G ethernet LAN transceivers, J. Lightw. Technol. 29 (15) (2011) 2319–2326.
[73] M. Glick, L.C. Kimmerling, R.C. Pfahl, A roadmap for integrated photonics, Opt. Photon. News 29 (3) (2018) 36–41.

[74] Disaggregated Servers Drive Data Center Efficiency and Innovation. Available from: <https://www.intel.com/content/www/us/en/it-management/intel-it-best-practices/disaggregated-server-architecture-drives-data-center-efficiency-paper.html>.
[75] B. Abali, R.J. Eickemeyer, H. Franke, C.-S. Li, M.A. Taubenblatt, Disaggregated and optically interconnected memory: when will it be cost effective? arXiv preprint arXiv:1503.01416, 2015.
[76] G. Zervas, H. Yuan, A. Saljoghei, Q. Chen, V. Mishra, Optically disaggregated data centers with minimal remote memory latency: technologies, architectures, and resource allocation, J. Opt. Commun. Netw. 10 (2) (2018) A270–A285.
[77] P.X. Gao, A. Narayan, S. Karandikar, J. Carreira, S. Han, R. Agarwal, et al., Network requirements for resource disaggregation, in: OSDI, 2016.
[78] W.M. Mellette, R. Mcguinness, A. Roy, A. Forencich, G. Papen, A.C. Snoeren, et al., RotorNet: a scalable, low-complexity, optical datacenter network, in: Proceedings of the Conference of the ACM Special Interest Group on Data Communication, ACM, 2017.
[79] K. Hazelwood, S. Bird, D. Brooks, S. Chintala, U. Diril, D. Dzhulgakov, et al., Applied machine learning at facebook: a datacenter infrastructure perspective, in: 2018 IEEE International Symposium on High Performance Computer Architecture (HPCA), 2018.
[80] S. Liu, J.C. Norman, D. Jung, M.J. Kennedy, A.C. Gossard, J.E. Bowers, Monolithic 9 GHz passively mode locked quantum dot lasers directly grown on on-axis (001) Si, Appl. Phys. Lett. 113 (4) (2018) 041108.
[81] A. Kovsh, I. Krestnikov, D. Livshits, S. Mikhrin, J. Weimert, A. Zhukov, Quantum dot laser with 75nm broad spectrum of emission, Opt. Lett. 32 (7) (2007) 793–795.
[82] V. Panapakkam, A.P. Anthur, V. Vujicic, R. Zhou, Q. Gaimard, K. Merghem, et al., Amplitude and phase noise of frequency combs generated by single-section InAs/InP quantum-dash-based passively and actively mode-locked lasers, IEEE J. Quant. Electr. 52 (11) (2016) 1–7.
[83] J.S. Levy, A. Gondarenko, M.A. Foster, A.C. Turner-Foster, A.L. Gaeta, M. Lipson, CMOS-compatible multiple-wavelength oscillator for on-chip optical interconnects, Nat. Photon. 4 (2009) 37.
[84] J. Pfeifle, V. Brasch, M. Lauermann, Y. Yu, D. Wegner, T. Herr, et al., Coherent terabit communications with microresonator Kerr frequency combs, Nat. Photon. 8 (2014) 375.
[85] X. Xue, P.H. Wang, Y. Xuan, M. Qi, A.M. Weiner, High-efficiency WDM sources based on microresonator Kerr frequency combs, in: 2017 Optical Fiber Communications Conference and Exhibition (OFC), 2017.
[86] B. Stern, X. Ji, Y. Okawachi, A.L. Gaeta, M. Lipson, Fully integrated chip platform for electrically pumped frequency comb generation, in: Conference on Lasers and Electro-Optics, Optical Society of America, San Jose, CA, 2018.
[87] C.-H. Chen, M.A. Seyedi, M. Fiorentino, D. Livshits, A. Gubenko, S. Mikhrin, et al., A comb laser-driven DWDM silicon photonic transmitter based on microring modulators, Opt. Express 23 (16) (2015) 21541–21548.
[88] Q. Xu, B. Schmidt, S. Pradhan, M. Lipson, Micrometre-scale silicon electro-optic modulator, Nature 435 (2005) 325.
[89] T. Baba, S. Akiyama, M. Imai, N. Hirayama, H. Takahashi, Y. Noguchi, et al., 50-Gb/s ring-resonator-based silicon modulator, Opt. Express 21 (10) (2013) 11869–11876.
[90] X. Xiao, X. Li, H. Xu, Y. Hu, K. Xiong, Z. Li, et al., 44-Gb/s silicon microring modulators based on zigzag pn junctions, IEEE Photon. Technol. Lett. 24 (19) (2012) 1712–1714.
[91] M. Pantouvaki, H. Yu, M. Rakowski, P. Christie, P. Verheyen, G. Lepage, et al., Comparison of silicon ring modulators with interdigitated and lateral p-n junctions, IEEE J. Select. Topics Quant. Electr. 19 (2) (2013), pp. 7900308–7900308.
[92] Q. Xu, B. Schmidt, J. Shakya, M. Lipson, Cascaded silicon micro-ring modulators for WDM optical interconnection, Opt. Express 14 (20) (2006) 9431–9436.
[93] D. Brunina, X. Zhu, K. Padmaraju, L. Chen, M. Lipson, K. Bergman, 10-Gb/s WDM optically-connected memory system using silicon microring modulators, in: European Conference and Exhibition on Optical Communication, Optical Society of America, Amsterdam, 2012.

[94] J. Li, X. Zheng, A.V. Krishnamoorthy, J.F. Buckwalter, Scaling trends for picojoule-per-bit WDM photonic interconnects in CMOS SOI and FinFET processes, J. Lightw. Technol. 34 (11) (2016) 2730–2742.
[95] J. Sun, M. Sakib, J. Driscoll, R. Kumar, H. Jayatilleka, Y. Chetrit, et al., A 128 Gb/s PAM4 silicon microring modulator, in: 2018 Optical Fiber Communications Conference and Exposition (OFC), 2018.
[96] R. Dubé-Demers, S. Larochelle, W. Shi, Ultrafast pulse-amplitude modulation with a femtojoule silicon photonic modulator, Optica 3 (6) (2016) 622–627.
[97] M. Bahadori, S. Rumley, D. Nikolova, K. Bergman, Comprehensive design space exploration of silicon photonic interconnects, J. Lightw. Technol. 34 (12) (2016) 2975–2987.
[98] R. Wu, C.H. Chen, J.M. Fedeli, M. Fournier, R.G. Beausoleil, K.T. Cheng, Compact modeling and system implications of microring modulators in nanophotonic interconnects, in: 2015 ACM/IEEE International Workshop on System Level Interconnect Prediction (SLIP), 2015.
[99] R. Wu, C.-H. Chen, J.-M. Fedeli, M. Fournier, K.-T. Cheng, R.G. Beausoleil, Compact models for carrier-injection silicon microring modulators, Opt. Express 23 (12) (2015) 15545–15554.
[100] K. Padmaraju, X. Zhu, L. Chen, M. Lipson, K. Bergman, Intermodulation crosstalk characteristics of WDM silicon microring modulators, IEEE Photon. Technol. Lett. 26 (14) (2014) 1478–1481.
[101] O. Dubray, A. Abraham, K. Hassan, S. Olivier, D. Marris-Morini, L. Vivien, et al., Electro-optical ring modulator: an ultracompact model for the comparison and optimization of p-n, p-i-n, and capacitive junction, IEEE J. Select. Topics Quant. Electr. 22 (6) (2016) 89–98.
[102] J.B. Quélène, J.F. Carpentier, Y.L. Guennec, P.L. Maître, Optimization of power coupling coefficient of a carrier depletion silicon ring modulator for WDM optical transmissions, in: 2016 IEEE Optical Interconnects Conference (OI), 2016.
[103] H. Jayatilleka, K. Murray, M. Caverley, N.A.F. Jaeger, L. Chrostowski, S. Shekhar, Crosstalk in SOI microring resonator-based filters, J. Lightw. Technol. 34 (12) (2016) 2886–2896.
[104] M. Bahadori, M. Nikdast, S. Rumley, L.Y. Dai, N. Janosik, T. Van Vaerenbergh, et al., Design space exploration of microring resonators in silicon photonicinterconnects: impact of the ring curvature, J. Lightw. Technol. 36 (13) (2018) 2767–2782.
[105] G. Li, A.V. Krishnamoorthy, I. Shubin, J. Yao, Y. Luo, H. Thacker, et al., Ring resonator modulators in silicon for interchip photonic links, IEEE J. Select. Topics Quant. Electr. 19 (6) (2013) 95–113.
[106] S. Rumley, M. Bahadori, D. Nikolova, K. Bergman, Physical layer compact models for ring resonators based dense WDM optical interconnects, in: ECOC 2016; 42nd European Conference on Optical Communication, 2016.
[107] M.A. Seyedi, R. Wu, C.-H. Chen, M. Fiorentino, R. Beausoleil, 15 Gb/s transmission with wide-FSR carrier injection ring modulator for Tb/s optical links, in: Conference on Lasers and Electro-Optics, Optical Society of America, San Jose, CA, 2016.
[108] M. Bahadori, S. Rumley, H. Jayatilleka, K. Murray, N. a F. Jaeger, L. Chrostowski, et al., Crosstalk penalty in microring-based silicon photonic interconnect systems, J. Lightw. Technol. 34 (17) (2016) 4043–4052.
[109] L. Chen, N. Sherwood-Droz, M. Lipson, Compact bandwidth-tunable microring resonators, Opt. Lett. 32 (22) (2007) 3361–3363.
[110] C.L. Manganelli, P. Pintus, F. Gambini, D. Fowler, M. Fournier, S. Faralli, et al., Large-FSR thermally tunable double-ring filters for WDM applications in silicon photonics, IEEE Photon. J. 9 (1) (2017) 1–10.
[111] M. Bahadori, A. Gazman, N. Janosik, S. Rumley, Z. Zhu, R. Polster, et al., Thermal rectification of integrated microheaters for microring resonators in silicon photonics platform, J. Lightw. Technol. 36 (3) (2018) 773–788.
[112] C. Sun, M. Wade, M. Georgas, S. Lin, L. Alloatti, B. Moss, et al., A 45 nm CMOS-SOI monolithic photonics platform with bit-statistics-based resonant microring thermal tuning, IEEE J. Solid-State Circ. 51 (4) (2016) 893–907.
[113] P.L. Maître, J.F. Carpentier, C. Baudot, N. Vulliet, A. Souhaité, J.B. Quélène, et al., Impact of process variability of active ring resonators in a 300mm silicon photonic platform, in: 2015 European Conference on Optical Communication (ECOC), 2015.

[114] M. Nikdast, G. Nicolescu, J. Trajkovic, O. Liboiron-Ladouceur, Chip-scale silicon photonic interconnects: a formal study on fabrication non-uniformity, J. Lightw. Technol. 34 (16) (2016) 3682–3695.
[115] K. Padmaraju, D.F. Logan, T. Shiraishi, J.J. Ackert, A.P. Knights, K. Bergman, Wavelength locking and thermally stabilizing microring resonators using dithering signals, J. Lightw. Technol. 32 (3) (2014) 505–512.
[116] F. Morichetti, A. Canciamilla, C. Ferrari, M. Torregiani, A. Melloni, M. Martinelli, Roughness induced backscattering in optical silicon waveguides, Phys. Rev. Lett. 104 (3) (2010) 033902.
[117] B.E. Little, J.-P. Laine, S.T. Chu, Surface-roughness-induced contradirectional coupling in ring and disk resonators, Opt. Lett. 22 (1) (1997) 4–6.
[118] M. Bahadori, S. Rumley, Q. Cheng, K. Bergman, Impact of backscattering on microring-based silicon photonic links, in: Optical Interconnects, 2018.
[119] C. Chen, C. Li, R. Bai, K. Yu, J. Fedeli, S. Meassoudene, et al., DWDM silicon photonic transceivers for optical interconnect, in: 2015 IEEE Optical Interconnects Conference (OI), 2015.
[120] A. Biberman, J. Chan, K. Bergman, On-chip optical interconnection network performance evaluation using power penalty metrics from silicon photonic modulators, in: 2010 IEEE International Interconnect Technology Conference, 2010.
[121] Q. Li, D. Nikolova, D.M. Calhoun, Y. Liu, R. Ding, T. Baehr-Jones, et al., Single microring-based 2×2 silicon photonic crossbar switches, IEEE Photon. Technol. Lett. 27 (18) (2015) 1981–1984.
[122] R. Ding, Y. Liu, Q. Li, Z. Xuan, Y. Ma, Y. Yang, et al., A compact low-power 320-Gb/s WDM transmitter based on silicon microrings, IEEE Photon. J. 6 (3) (2014) 1–8.
[123] P. Dong, W. Qian, H. Liang, R. Shafiiha, N.-N. Feng, D. Feng, et al., Low power and compact reconfigurable multiplexing devices based on silicon microring resonators, Opt. Express 18 (10) (2010) 9852–9858.
[124] P. Dong, S. Liao, D. Feng, H. Liang, D. Zheng, R. Shafiiha, et al., Low Vpp, ultralow-energy, compact, high-speed silicon electro-optic modulator, Opt. Express 17 (25) (2009) 22484–22490.
[125] W. Bogaerts, R. Baets, P. Dumon, V. Wiaux, S. Beckx, D. Taillaert, et al., Nanophotonic waveguides in silicon-on-insulator fabricated with CMOS technology, J. Lightw. Technol. 23 (1) (2005) 401.
[126] A. Bianco, D. Cuda, R. Gaudino, G. Gavilanes, F. Neri, M. Petracca, Scalability of optical interconnects based on microring resonators, IEEE Photon. Technol. Lett. 22 (15) (2010) 1081–1083.
[127] K. Yu, C.-H. Chen, A. Titriku, A. Shafik, M. Fiorentino, P.Y. Chiang, S. Palermo, 25Gb/s hybrid-integrated silicon photonic receiver with microring wavelength stabilization, in: Optical Fiber Communication Conference, Optical Society of America, 2015.
[128] M. Bahadori, D. Nikolova, S. Rumley, C.P. Chen, K. Bergman, Optimization of microring-based filters for dense WDM silicon photonic interconnects, in: Optical Interconnects Conference (OI), 2015 IEEE, IEEE, 2015
[129] R. Hendry, D. Nikolova, S. Rumley, N. Ophir, K. Bergman, Physical layer analysis and modeling of silicon photonic WDM bus architectures, in: Proc. HiPEAC Workshop, 2014.
[130] B.G. Lee, A. Biberman, P. Dong, M. Lipson, K. Bergman, All-optical comb switch for multiwavelength message routing in silicon photonic networks, IEEE Photon. Technol. Lett. 20 (10) (2008) 767–769.
[131] M. Georgas, J. Leu, B. Moss, C. Sun, V. Stojanović, Addressing link-level design tradeoffs for integrated photonic interconnects, in: Custom Integrated Circuits Conference (CICC), 2011 IEEE, IEEE, 2011.
[132] N. Ophir, C. Mineo, D. Mountain, K. Bergman, Silicon photonic microring links for high-bandwidth-density, low-power chip I/O, IEEE Micro 33 (1) (2013) 54–67.
[133] R. Hendry, D. Nikolova, S. Rumley, K. Bergman, Modeling and evaluation of chip-to-chip scale silicon photonic networks, in: 2014 IEEE 22nd Annual Symposium on High-Performance Interconnects, 2014.
[134] A. Yariv, Universal relations for coupling of optical power between microresonators and dielectric waveguides, Electron. Lett. 36 (4) (2000) 321–322.
[135] A. Biberman, P. Dong, B.G. Lee, J.D. Foster, M. Lipson, K. Bergman, Silicon microring resonator-based broadband comb switch for wavelength-parallel message routing, in: LEOS 2007 - IEEE Lasers and Electro-Optics Society Annual Meeting Conference Proceedings, 2007.

[136] M. Bahadori, S. Rumley, R. Polster, A. Gazman, M. Traverso, M. Webster, et al., Energy-performance optimized design of silicon photonic interconnection networks for high-performance computing, in: Proceedings of the Conference on Design, Automation & Test in Europe, European Design and Automation Association: Lausanne, Switzerland, 2017, pp. 326–331.
[137] J. Kim, C.J. Nuzman, B. Kumar, D.F. Lieuwen, J.S. Kraus, A. Weiss, et al., 1100 x 1100 port MEMS-based optical crossconnect with 4-dB maximum loss, IEEE Photon. Technol. Lett. 15 (11) (2003) 1537–1539.
[138] Polatis technology – Directlight® Beam-Steering All-Optical Switch. Available from: <http://www.polatis.com/polatis-all-optical-switch-technology-lowest-loss-highest-performance-directlight-beam-steering.asp>.
[139] Z. Zhang, Z. You, D. Chu, Fundamentals of phase-only liquid crystal on silicon (LCOS) devices, Light Sci. Appl. 3 (2014). p. e213.
[140] 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. Select. Topics Quant. Electr. 11 (2) (2005) 383–394.
[141] R. Ryf, J. Kim, J.P. Hickey, A. Gnauck, D. Carr, F. Pardo, et al. 1296-Port MEMS transparent optical crossconnect with 2.07 petabit/s switch capacity, in: OFC 2001. Optical Fiber Communication Conference and Exhibit. Technical Digest Postconference Edition (IEEE Cat. 01CH37171), 2001.
[142] D.T. Neilson, R. Frahm, P. Kolodner, C.A. Bolle, R. Ryf, J. Kim, et al., 256 × 256 Port optical cross-connect subsystem, J. Lightw. Technol. 22 (6) (2004) 1499.
[143] K. Wang, A. Wonfor, R.V. Penty, I.H. White, Active-passive 4x4 SOA-based switch with integrated power monitoring, in: OFC/NFOEC, 2012.
[144] I. White, E.T. Aw, K. Williams, H. Wang, A. Wonfor, R. Penty, Scalable optical switches for computing applications [Invited], J. Opt. Netw. 8 (2) (2009) 215–224.
[145] A. Wonfor, H. Wang, R.V. Penty, I.H. White, Large port count high-speed optical switch fabric for use within datacenters [invited], J. Opt. Commun. Netw. 3 (8) (2011) A32–A39.
[146] R. Stabile, A. Albores-Mejia, K.A. Williams, Monolithic active-passive 16x16 optoelectronic switch, Opt. Lett. 37 (22) (2012) 4666–4668.
[147] R. Stabile, A. Albores-Mejia, K.A. Williams, Monolithically integrated 8 × 8 space and wavelength selective cross-connect, J. Lightw. Technol. 32 (2) (2014) 201–207.
[148] Q. Cheng, M. Ding, A. Wonfor, J. Wei, R.V. Penty, I.H. White, The feasibility of building a 64x64 port count SOA-based optical switch, in: 2015 International Conference on Photonics in Switching (PS), 2015.
[149] Q. Cheng, A. Wonfor, J.L. Wei, R.V. Penty, I.H. White, Low-energy, high-performance lossless 8 × 8 SOA switch, in: 2015 Optical Fiber Communications Conference and Exhibition (OFC), 2015.
[150] T. Chu, L. Qiao, W. Tang, D. Guo, W. Wu, Fast, high-radix silicon photonic switches, in: 2018 Optical Fiber Communications Conference and Exposition (OFC), 2018.
[151] T.J. Seok, K. Kwon, J. Henriksson, J. Luo, M.C. Wu, "240 × 240 Wafer-Scale Silicon Photonic Switches," in: Optical Fiber Communication Conference (OFC) 2019, OSA Technical Digest (Optical Society of America, 2019), paper Th1E.5.
[152] K. Suzuki, R. Konoike, J. Hasegawa, S. Suda, H. Matsuura, K. Ikeda, et al., Low insertion loss and power efficient 32x32 silicon photonics switch with extremely-high-D PLC connector, in: 2018 Optical Fiber Communications Conference and Exposition (OFC), 2018.
[153] L. Qiao, W. Tang, T. Chu, 32 × 32 silicon electro-optic switch with built-in monitors and balanced-status units, Sci. Rep. 7 (2017) 42306.
[154] Q. Cheng, L.Y. Dai, M. Bahadori, N.C. Abrams, P.E. Morrissey, M. Glick, et al., Si/SiN microring-based optical router in switch-and-select topology, in: European Conference on Optical Communication (ECOC), 2018, p. We1C.3.
[155] Q. Cheng, M. Bahadori, S. Rumley, K. Bergman, Highly-scalable, low-crosstalk architecture for ring-based optical space switch fabrics, in: 2017 IEEE Optical Interconnects Conference (OI), 2017.

[156] Q. Cheng, L.Y. Dai, N.C. Abrams, Y. Hung, P.E. Morrissey, M. Glick, P. O'Brien, K. Bergman, Ultralow-crosstalk, strictly non-blocking microring-based optical switch, Photon. Res 7 (2019) 155–161.
[157] Q. Cheng, M. Bahadori, Y. Hung, Y. Huang, N. Abrams, K. Bergman, Scalable Microring-Based Silicon Clos Switch Fabric with Switch-and-Select Stages, in IEEE Journal of Selected Topics in Quantum Electronics. https://doi.org/10.1109/JSTQE.2019.2911421
[158] Q. Cheng, R. Dai, M. Bahadori, P. Morrissey, R. Polster, S. Rumley, et al., Microring-based Si/SiN dual-layer switch fabric, in: Optical Interconnects, IEEE, Santa Fe, New Mexico, USA, 2018.
[159] M. Ding, A. Wonfor, Q. Cheng, R.V. Penty, I.H. White, Hybrid MZI-SOA InGaAs/InP photonic integrated switches, IEEE J. Select. Topics Quant. Electr. 24 (1) (2018) 1–8.
[160] Q. Cheng, A. Wonfor, J.L. Wei, R.V. Penty, I.H. White, Monolithic MZI-SOA hybrid switch for low-power and low-penalty operation, Opt. Lett. 39 (6) (2014) 1449–1452.
[161] Q. Cheng, A. Wonfor, J.L. Wei, R.V. Penty, I.H. White, Demonstration of the feasibility of large-port-count optical switching using a hybrid Mach-Zehnder interferometer-semiconductor optical amplifier switch module in a recirculating loop, Opt. Lett. 39 (18) (2014) 5244–5247.
[162] A. Bhatele, N. Jain, Y. Livnat, V. Pascucci, P. Bremer, Analyzing network health and congestion in dragonfly-based supercomputers, in: 2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS), 2016.
[163] K. Wen, D. Calhoun, S. Rumley, X. Zhu, Y. Liu, L.W. Luo, et al., Reuse distance based circuit replacement in silicon photonic interconnection networks for HPC, in: 2014 IEEE 22nd Annual Symposium on High-Performance Interconnects, 2014.
[164] J. Kim, W.J. Dally, S. Scott, D. Abts, Cost-efficient dragonfly topology for large-scale systems, in: 2009 Conference on Optical Fiber Communication - Incudes Post Deadline Papers, 2009.
[165] J. Wilke, Bringing minimal routing back to HPC through silicon photonics: a study of "flexfly" architectures with the structural simulation toolkit (SST), in: Proceedings of the 2nd International Workshop on Advanced Interconnect Solutions and Technologies for Emerging Computing Systems, ACM, Stockholm, Sweden, 2017, p. 5-5.