Silicon Photonics for Extreme Scale Systems

Yiwen Shen, *Student Member, IEEE*, Xiang Meng, *Member, IEEE*, Qixiang Cheng, *Member, IEEE*, Sébastien Rumley, Nathan Abrams, Alexander Gazman, Evgeny Manzhosov, Madeleine Strom Glick, *Senior Member, IEEE, Senior Member, OSA*, and Keren Bergman, *Fellow, IEEE, Fellow, OSA*

(Invited Tutorial)

***Abstract*—High-performance systems are increasingly bottlenecked by the growing energy and communications costs of interconnecting numerous compute and memory resources. Recent advances in integrated silicon photonics offer the opportunity of embedding optical connectivity that directly delivers high off-chip communication bandwidth densities with low power consumption. This paper reviews the design and integration of silicon photonic interconnection networks that address the data-movement challenges in high-performance systems. Beyond alleviating the bandwidth/energy bottlenecks, embedded photonics can enable new disaggregated architectures that leverage the distance independence of optical transmission. This review paper presents some of the key interconnect requirements to create a new generation of photonic architectures for extreme scale systems, including aggregate bandwidth, bandwidth density, energy efficiency, and network resources utilization.**

***Index Terms*—Disaggregation, extreme scale systems, interconnection networks, optical switching, silicon photonics.**

I. Introduction

THE demand for greater computational performance is increasingly reliant on the capability of the interconnection network to provide ultra-high bandwidth connections with efficient power dissipation. Over the last decade, compute power per individual CPU has stalled and increases in performance have since relied on parallelizing computation resources of multiple CPU cores and other discrete components to form powerful microprocessors [1]. Large-scale integration of these microprocessors with memory, accelerators, storage, and network components form the basis for realizing future massive parallel supercomputers. While developments in parallelism have so far produced performance growth, the scalability of the system depends on the underlying interconnection technology to provide adequate bandwidth resources and minimal latency to memory and storage nodes to achieve the full computation potential of each processor.

Over the last few years however, the growth of interconnect bandwidth capacity has not been able to match the pace of the increase in raw processing power gained through parallelism. Despite the rising peak computing performance, the ratio of bandwidth between each node to intrinsic processing power of per node (byte/FLOP ratio) of the Top500 [2] top performing high-performance computing (HPC) systems have decreased by a factor of 6× since June 2010 [3]. As of November 2018, the new *Summit* supercomputer achieves 200.8 PFlops using 4608 servers, which means each server performs 43 TFlops. However, each node only receives 25 GB/s bandwidth with dual NVLINK bricks resulting in a 0.0006 byte/FLOP ratio [4]. Another example is the *Sunway TaihuLight* which has a 0.004 byte/FLOP ratio [3]. The effect of such low byte/FLOP ratios in HPC systems results in data-starved processors whose full utilization are not realized, resulting in suboptimal system performance and energy efficiency.

This effect can be quantitatively observed when the same HPC systems operate the High Performance Gradients (HPCG) benchmark instead of the traditionally used High Performance Linpack (HPL) benchmark: *Summit* performed only 2.9 PFlops (compared to 143.5 PFlops with HPL), while *Sunway TaihuLight* showed a performance of 0.48 PFlops (compared to 93 PFlops with HPL). This is because the Linpack program performs compute-rich algorithms to multiply dense matrices, called Type 1 patterns, favoring systems with high computational capabilities [5]. However, most current applications require lower computation-to-data-access ratios, referred to as Type 2 patterns, which require higher bandwidth communications with the memory subsystem. The HPL program focuses only on Type 1 patterns, while the HPCG benchmark has emerged as a new metric for HPC systems that is designed to evaluate systems with both Type 1 and Type 2 patterns. As such, performance of the top ranked supercomputers operating HPCG show much more modest results compared to their HPL

Manuscript received November 30, 2018; revised January 28, 2019 and January 31, 2019; accepted January 31, 2019. Date of publication February 4, 2019; date of current version February 20, 2019. This work was supported in part by the U.S. Department of Energy Sandia National Laboratories under Contract PO 1426332 and Contract PO 1319001, in part by the Advanced Research Projects Agency Energy (ARPA-E) under the Enlightened Project under Contract DE-AR0000843, and in part by the National Security Agency (NSA) Laboratory for Physical Sciences (LPS) Research Initiative (R3/NSA) under Contract FA8075-14-D-0002-0007, TAT 15-1158. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000. *(Corresponding author: Yiwen Shen.)*

The authors are with the Lightwave Research Laboratory, Columbia University, New York, NY 10027 USA (e-mail: ys2799@columbia.edu; xm2137@columbia.edu; qc2228@columbia.edu; sebastien.rumley@olympe.ch; nca2123@columbia.edu; ag3529@columbia.edu; em3282@columbia.edu; msg144@columbia.edu; bergman@ee.columbia.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JLT.2019.2897365

0733-8724 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

counterparts. The decrease in nearly two orders of magnitude in *Summit*'s and *Sunway TaihuLight*'s performance highlights the crucial role that the interconnection network plays in supplying a system's computational resources with sufficient traffic to memory and other computing nodes in order to bring forth the system's full performance potential.

A suboptimal usage of computation resources due to network congestion also directly translates to unnecessary power dissipation. Currently (as of January 2019), *Summit* ranks 3rd among the Green500 [6] that lists the world's most energy efficient supercomputers, obtained by dividing its HPL performance of 143.5 PFlops with its power usage of 9.8 MW resulting in approximately 14.6 GFlops/W. However, if the performance operating HPCG is used instead, the energy efficiency is only 0.3 GFlops/W. The relationship between low processor utilization and high power dissipation can be understood qualitatively as the low processor utilization leads to a given job/task taking longer to complete, thereby requiring more power to be consumed.

The growing challenge in communications necessitates the development of energy proportional, high bandwidth capacity and low latency interconnects. A possible candidate technology is silicon photonics (SiP), which offers the means to design highly-integrated structures that are fabricated using mature CMOS processes to interconnect electronic processing and memory nodes. The ultra-high bandwidth capacity of SiP interconnects at distances from the order of millimeters to meters can be leveraged within computing systems, shifting the current design paradigm from off-chip data constraints, to flexible, disaggregated systems with greater freedom in network architecture design.

The optimal integration of SiP interconnects is more complex than simply a one-to-one replacement of electronic links. It requires the development of modeling tools for circuit and chip design, methodologies for their integration between electronic peripherals, and network architectures and system control plane that incorporates the optical physical layer characteristics of the entire system [8]. The importance of the network control plane cannot be underestimated - even with ideal ultra-high bandwidth capacity optical links, mismatch between the application traffic characteristics and network resources leads to both congested and under-utilized links resulting in communication bottlenecks with far-reaching effects that severely impede overall system performance [9]–[11].

In this review paper, we focus on the vision of designing an energy efficient, high-bandwidth optically interconnected network fabric. This vision is supported by various pieces: energy efficient link design and techniques for embedding them with electronic components, as well as a flexible network architecture that optimizes bandwidth resource utilization through efficient optical switching to reduce computation time and energy consumption. This paper will present both our works as well as review other state-of-the-art efforts that fall within these pieces. In Section II we summarize the challenges faced by electrical links in terms of maintaining efficient power dissipation levels over a range of distances, and how silicon photonic interconnects can be a potential solution. In Section III, we describe how energy efficient silicon photonic links are designed and how they can be integrated into multi-chip modules (MCMs)

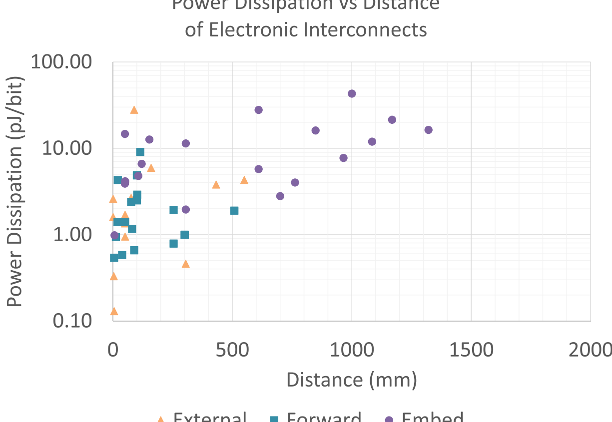

Fig. 1. Power dissipation plotted over distance for electronic interconnects [7].

that contain various computation and memory resources. In Section IV we show how novel network architectures that leverage silicon photonic switches can enable reconfigurable network topologies that use resources efficiently to provide significant performance improvements. Lastly in Section V we describe our future plans of a disaggregated network architecture that support components such as memory, CPUs, GPUs and FPGAs interconnected with a unified silicon photonic fabric.

II. Energy Challenges and Opportunities for Data Movement

There is a considerable body of work on low power electronic interconnects with sub-pJ/bit energy efficiencies over gigabit data rates [7]. Figure 1 shows the energy efficiency over distance for state-of-the-art electronic interconnects, categorized by clocking architecture (external clock, forward, or embedded)[1] and minimum feature size. It can be observed that the lowest power links with sub-pJ/bit energy efficiency are within a distance under 300 mm. As distances increase, energy increases to tens of pJ/bit or greater, meaning that while electronic CMOS techniques are capable of realizing high-bandwidth off-chip communications, pJ/bit efficiency can be realized only at a distance confined within the immediate chip neighborhood. Beyond this range, conventional electronic links operating at high frequency have difficulty maintaining sub pJ/bit efficiency due to various effects that lead to higher losses and the excitation of higher-order parasitic modes. For example, the skin effect problem results in higher resistive losses that scales with the square-root of the frequency, due to the tendency for alternating current to flow near the surface of the conductor. Second, the dielectric loss scales linearly with frequency. Together, these effects constitute the total insertion loss of a printed circuit board (PCB) trace. Lastly, electrical interconnects are limited by wiring density, which exacerbates loss at smaller cross-sectional sizes due to both the resistance and

[1] Clocking architectures - External: same crystal is providing clock to both TX and RX chips; Forward: multiple lanes are used but one lane is dedicated for clock transmission; Embedded: no clock is transmitted from TX and RX recovers the clock from received data

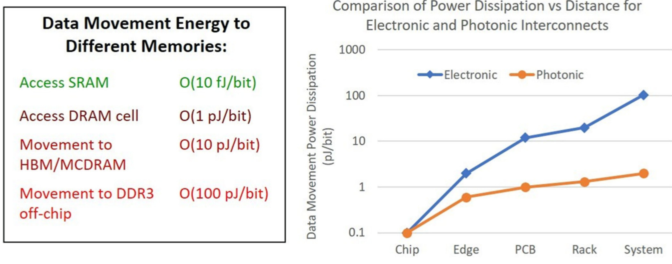

Fig. 2. Power dissipation of data movement to different memory types (left) and different distances (right) [11], [12].

capacitance of the wires. These physical effects cause chip architectures to be increasingly constrained by power dissipation, and limits data throughput achievable by off-chip interconnection networks.

These physical limitations have given rise to the "bandwidth taper" phenomenon, which refers to the orders of magnitude decrease in bandwidth provided by electronic interconnects as data travels over distances ranging beyond on-chip scales, to off-chip, inter-node and inter-rack [11], [12]. This translates to a proportional increase in energy consumption, as displayed in Figure 2 which lists the energy requirements for data movement to different types of memory modules and to varying distances. The pJ/bit unit is used to indicate a power consumption budget for transmission. Access to SRAM occurs on-chip and is on the order of tens of fJ/bit [13]. However, as we move further from the CPU/GPU to access off-chip memory, increases of an order of magnitude in pJ/bit are required.

To provide perspective on the I/O energy budget, the general accepted energy constraint to realize an Exascale compute system is 50 GFlops/W, which stems from dividing 1 Exaflop with the energy constraint of 20 MW generally given to a large-scale supercomputer. The target of 50 GFlops/W is equivalent to 20 pJ/FLOP. With a (double-precision) FLOP taking two operands in (2×64 bits) and one out (64 bits), or roughly 200 bits for each I/O operation, this means that no more than 0.1 pJ can be dissipated per bit. Comparing 0.1 pJ/bit to the values listed in Figure 2 shows the extremely tight budget for I/O. For example, the energy required to access a bit in a modern memory chip (e.g. HBM2) is approximately 3–5 pJ/bit [14]. Of course, the existence of high-capacity on-chip caches means that not every bit involved in a FLOP must be taken out of off-chip memory. Yet with 4 pJ/bit, a budget of 20 pJ/FLOP, and 4 pJ for the arithmetic, no more than half a byte is available for each FLOP. This is insufficient, especially for data intensive applications such as deep learning. To ensure comfortable memory bandwidth and thus an effective utilization of the computing resources, off-chip IO operations should consume 1 pJ/bit or less.

In general, the system requirements for optical interconnections are <1 pJ/bit for rack-to-rack distances, and <0.1 pJ/bit for chip-to-chip distances. This target may allow electronic links to remain viable for small distance connections up to the chip-to-chip range, but the fundamental properties of electrical wires limits the energy efficiency and bandwidth density at longer distances and higher data rates. Meanwhile, energy consumption of an optical link depends on the laser source, data modulation and its associated electronic drivers. Losses in an optical link are distance dependent, but are much lower than electrical interconnects, and they are not data rate dependent. [12] states that there is a break-even distance where optical interconnects become more power efficient than their electrical counterparts, and this distance depends on the data rate. Above 10 Gbps, energy required for modulation becomes dominant, and when the energy efficiency for modulation is <1 pJ/bit, this break-even length is 1 mm. Currently, conventional electronic technologies provides approximately on the order of 10 pJ/bit for rack scale data movement, and on the order of 100 pJ/bit for system scale data movement, as shown in Figure 2 [11].

The need for an energy efficient interconnect paves the way for SiP to transform existing link and network design paradigms. With recent developments that enable photonic components to be tightly integrated with electronic processing and memory peripherals [15], [16], SiP interconnects can potentially enable truly scalable extreme-scale computating platforms and opportunities for ultra-low energy transmission of close-proximity electronic driver circuitry integrated on-chip. As of now, rack-to-rack interconnects used by *Summit* are 100G InfiniBand, which use QSFP28 transceivers with <3.5 W of power, or 35 pJ/bit. In the next section, we will show our path to reach a lower power consumption of 1 pJ/bit through in-depth optimization of a microring-based transceiver architecture with a comb laser source that offers dense wavelength division multiplexing (WDM). We also present the methodologies in fabrication for embedding these links with electronic peripherals.

III. Silicon Photonic Interconnects at the System Level

The design of high bandwidth and efficient silicon photonic links involves finding the optimal combination between number of wavelengths, N_λ, and data rate for each channel, r_b, so that minimal energy consumption is achieved for a given optical aggregation rate. Our link design presented in the following section primarily focuses on minimizing energy consumption. The results presented are obtained using PhoenixSim [17], an open-source software platform developed at the Lightwave Research Laboratory, used for efficient design with electric and photonic co-optimization and analysis of physical layer, link layer and system-level silicon photonic interconnects. We also discuss current efforts in the integration of silicon photonic links to enable photonic interconnection between CPUs, GPUs, FPGAs, and memory components.

A. Energy Efficient Photonic Link Design

In this section we present a design exploration for obtaining maximum achievable aggregate bandwidth using silicon photonic microring modulators and filters. The microring structure is chosen because it has a small footprint and high performance. Silicon microrings have demonstrated significant reduction in footprint due to very high refractive index contrast, which yields much lower propagation loss compare to other on-chip structures. Extremely compact rings supports an FSR over 20 nm at

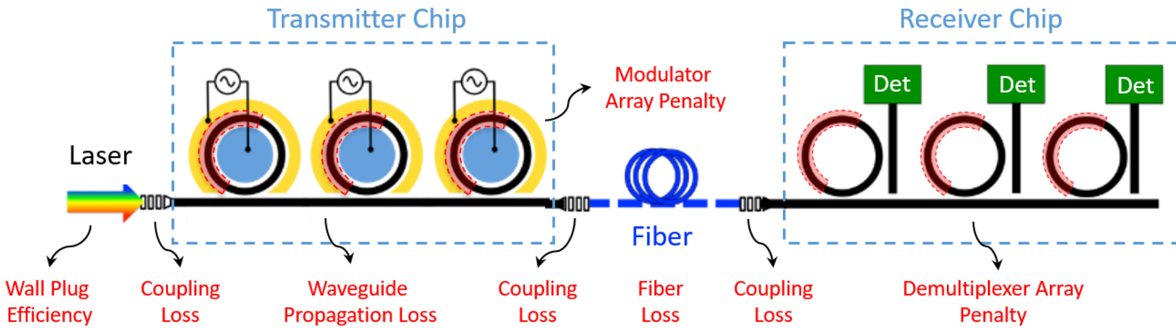

Fig. 3. Chip-to-Chip silicon photonic interconnect with an MRR-based WDM link. The optical interface of the transmitter chip includes MRR modulators that use the carrier dispersion principle for high speed modulation. The optical interface at the receiver includes demux filters, photodetectors, and electronic decision circuitry (Det: detector). Wall-plug efficiency corresponds to the electrical to optical power conversion of the laser [18].

communication wavelengths, which offers great scalability to achieve high-speed data communication through dense WDM [19]. The small footprint of silicon microring also yields low power consumption with high modulation depth [20]. We analyze the impacts of impairments and translate these results into power penalties. The power penalty is the extra optical power required to compensate for the effects of impairments on bit-error-ratio (BER) [21]–[25]. We also describe the changes in spectral properties of modulated light as it travels through a ring demultiplexer and how these changes translate to power penalty.

Microring-resonator (MRR)-based links offer the highest bandwidth density and most energy efficient performance for a high bandwidth above 100 Gbps among current silicon photonic interconnect devices [27]–[29]. MRRs can be designed to perform a number of interconnect functions and represent the key building blocks of silicon photonic systems [19], [30] such as active modulators and optical switches, as well as passive wavelength-selective filters. When acting as modulators [20], [31]–[38], MRRs leverage the plasma dispersion effect of silicon [39]–[41] which provides an efficient way to change the effective refractive index of the ring by injecting or depleting charge carriers. This modulation scheme can deliver high-speed modulations 10 Gb/s or higher with low energy consumption [42]–[44]. The photonic link is based on an external frequency comb laser, and optical couplers with losses lower than 1 dB. EIC components are optimized for minimum the optical power penalty on a 65 nm CMOS technology. We use 65 nm drive electronics as it provides the best cost-performance balance for a channel rate of 10-15 Gb/s due to its energy efficiency. However, if we focus on maximizing data rate, better CMOS process nodes will be more suitable for driver electronics. The optical power is modulated with non-return-to-zero on-off keying (NRZ-OOK) modulation format for minimal power consumption. The receiver of the link has a sensitivity of –17 dBm with a fixed BER of 10^{-12}.

When used as demultiplexers, a ring operates as a passive drop filter with its resonance tuned to a specific channel wavelength [28], [45], [46]. In both cases, fine-tuning is carried out via integrated heaters implemented above the silicon microring [47], [48]. Due to their small size, multiple microrings can be placed along a single waveguide on chip, resulting in a dense WDM design [49]–[51]. However, WDM links may suffer from spectral degradation of channels and inter-channel crosstalk [52]–[58], which is treated as an optical penalty in our model. These impairments eventually set an upper limit on both the number of channels and modulation speed of each channel, resulting in an upper bound on the aggregate rate to the link [59], [60]. To minimize optical crosstalk, the main factor to design the proper channel spacing which is dictated by the linewidth of the MRR and its modulation rate.

Consider a chip-to-chip silicon photonic link as shown in Figure 3. The off-chip frequency comb laser source is coupled into the MRRs chip through the fiber-to-chip edge coupler [61]. The incoming wavelengths, once imprinted with data, are then transmitted through an optical fiber to a receiver chip. The receiver chip consists of multiple passive MRRs 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, as well as to choose higher bit rates for each channel to allow 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 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 optical impairments through the entire link. All the power penalties of the link, PP^{dB}, for a single channel, are shown in the following inequality [62]:

$$[P_{\mathrm{laser}}^{\mathrm{dBm}} - 10\log_{10}(N_\lambda)] - P_{\mathrm{sensitivity}}^{\mathrm{dBm}} \geq PP^{\mathrm{dB}} \quad (1)$$

In general, aggregated optical power P_{laser} (summed over all wavelengths) must stay below the nonlinear threshold of the silicon waveguides at any point of the link [56], [60], [63], [64], as higher-order effects, such as two-photon absorption, free-carrier effects and optical Kerr effect, would significantly affect the signal quality when the optical power is above the threshold. 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 receiver side. A typical receiver may have a sensitivity of −12.7 dBm at 8 Gb/s operation [23], while a better receiver may have a sensitivity of approximately −21 dBm at 10 Gb/s

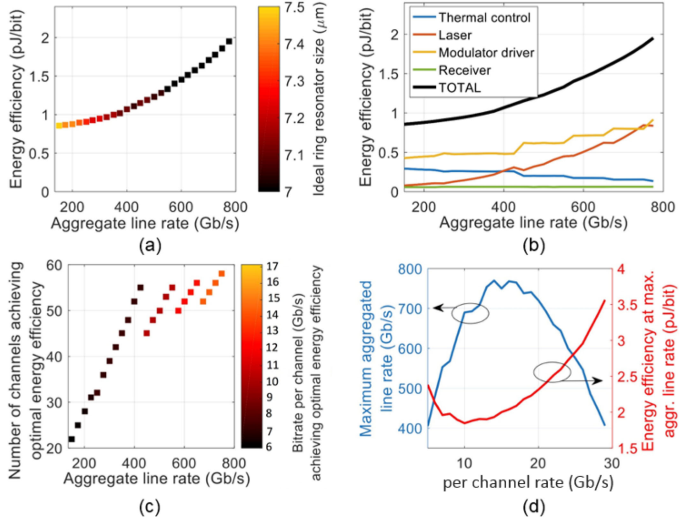

Fig. 4. (a) Minimum energy consumption of the link for given aggregations rates based on optimum values for the ring radius. (b) Breakdown of energy consumption. (c) Breakdown of the number of channels and the required data rate per channel for minimum energy consumption. (d) Evaluation of the maximum supported aggregation and the associated energy consumption for various channel rates [26].

[21], [65]–[68]. The difference between these thresholds can be exploited to find the maximum power budget. This maximum budget accounts for the power penalty, PP^{dB}, per channel over N_{λ} channels. Power impairments induced by the MRRs depend on the channel spacing, which is inversely proportional to the number of channels, and on the modulation rate. Note that if the data rate of the individual modulator from the WDM link is too low, the link does not offer energy efficient performance due to its non-ideal utilization of optical power. On the other hand, if the data rate of the individual modulator is too high, it would lead to higher power consumption from the electronic components, such as modulator drivers and trans-impedance amplifiers. Therefore, the trade-off and balance between electronic and photonic components leads to a favorable data rate range. The detailed link-level impairments and power penalties is provided in [69].

Figure 4(a) presents results on the minimal energy for 200 Gbps to 800 Gbps aggregation based on optimization of the physical parameters of the MRR. The rings are configured to operate at their critical coupling point [70], [71]. This chart shows that a ring radius of $\sim$7 μm leads to the best energy performance of the link. Figure 4(b) illustrates the breakdown of the factors on energy consumption. At higher data rates, we see that laser power consumption becomes critical, but energy consumed by the static thermal tuning declines. Figure 4(c) provides details on the number of channels and the required data rate per channel that lead to minimal energy consumption for the target aggregation rate. Note that the energy efficient link with high data rate in ranges of Tb/s can be achieved through dense WDM with 10-15 Gb/s per channel.

With the condition that available optical power budget is fully utilized, our study also investigates the maximal aggregation rate based on the product of the number of channels and the optical data rate of each channel. Figure 4(d) indicates a maximum possible aggregation rate of $\sim$800 Gbps at 15 Gbps data rate per channel. The energy efficiency for each data rate is also plotted. The energy consumption of our optimized MRR photonic link ranges from 1 pJ/bit to 2 pJ/bit. Note that the optimal energy efficiency is not associated with the highest aggregation rate, a result that highlights the fact that designing a silicon photonic link requires tradeoffs between energy and data rate. We note that results may vary if different parameters for the MRRs are adapted. In addition, losses in the rings have significant impact on the maximum aggregation rate. The details for the model of the MRRs associated with our results can be found in [70].

It is important to note that our link presented is a snapshot of a co-design of electronic and photonic components with a primary focus on minimizing energy consumption. Currently, our link uses NRZ modulation format due to its minimal power consumption. If high data rate was our focus, then we can use more advanced modulation formats such as PAM4, as well as using higher electronic Serializer/Deserializer (SerDes) speeds to increase the data rate. However, the use of PAM4 will require adjusting our link design parameters to optimized with its various additional electronic functions such as (CDR, FEC, FFE, DFE, etc.). Future work will need to be done on the co-design of our photonic link with these options to increase the aggregate data rate while maintaining a sufficiently low energy dissipation to be suitable for next-generation extreme scale computer systems.

B. *Embedding Silicon Photonic Links with Electronic Components*

Leveraging SiP links within extreme scale systems requires tight integration between the optics and the electronic components (CPUs, GPUs, FPGAs, and memory). Integration should focus on maximizing I/O to preserve bandwidth density and minimizing parasitics to preserve the system bandwidth and minimize energy consumption.

Monolithic integration, where the CMOS electrical components are fabricated in the same process as the SiP, is an attractive solution as it allows the driving electronics to be placed as closely as possible to the optics, eliminating the parasitics that result from hybrid integration. While monolithic integration has been demonstrated [72], there is still a need for hybrid integration. The monolithic fabrication process is not as mature as the individual SiP and electronic processes. Additionally, hybrid integration makes it possible for the SiP process to be fully optimized for optical performance and the CMOS process to be fully optimized for electrical performance.

Figure 5 shows different methods of SiP integration. In 2D integration, one of the earliest methods, the SiP chip and the CMOS chip are placed side by side, and wire bonds are used to connect the driving electronics with the optical components [73]. While 2D integration allows connectivity between the photonics and electronics, wirebond connectivity occurs only on a single edge, limiting I/O. Additionally, the wirebond connection between photonics and electronics introduces parasitic inductance. The inductance is directly related to the wirebond length

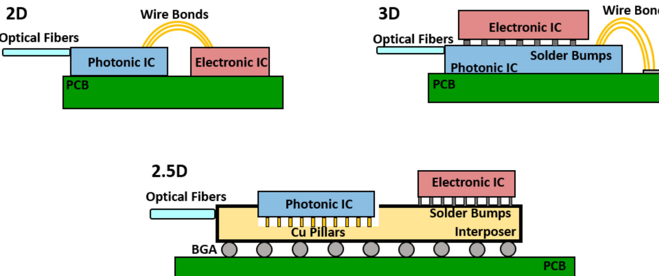

Fig. 5. (a) Schematics of the three integration approaches, 2D, 3D, and 2.5D. In these images the system components that would connect to the electronic IC are omitted. The 2.5D version is shown with the interposer containing an optical redistribution layer. Optical coupling is depicted as edge couplers, but grating couplers could also be used for the 2D and 3D version, as well as the 2.5D version if the photonic ICs substrate was removed.

and diameter, but typically exists in the range of 0.5–1.0 nH/mm, depending on the diameter [74]. Parasitic inductance can be mitigated by keeping the wirebond lengths short and by using larger diameter wirebonds (or ribbon bonds). However, to approach the 0.5 nH/mm requires wirebond diameters to increase, sometimes in excess of 100 μm, placing restrictions on I/O density.

A second SiP integration approach is 3D integration, where one chip is flip chipped on top of the other. Typically the CMOS chip is flipped onto the photonics chip as photonic components are larger than CMOS circuits. 3D integration increases the connectivity as micro solder bumps or copper pillars support pitches well below 100 μm and can be placed throughout the two-dimensional flip chipped area. Flip chip bumps introduce small amounts of parasitic resistance and capacitance. The parasitic resistance is typically below 1 Ω, and can be below 0.1 Ω. The parasitic capacitance has been demonstrated to be on the order of 20–30 fF, though reducing bond size to around 10 μm could further reduce the capacitance [75]–[77].

A third SiP integration approach is 2.5D integration, where an interposer is used to provide connectivity between the photonic integrated circuit (PIC) and the CMOS IC. Both chips are flipped onto the interposer, which serves as a redistribution layer. Similar to the 3D approach, 2.5D integration (sometimes referred to as a multi-chip module, or MCM) allows for high I/O due to micro solder bumps or copper pillars that provide the connection between the chips and the interposer. This approach has more parasitic capacitance than 3D integration, as a driving signal passes through two sets of micro solder bumps or copper pillars and the trace length between the two chips.

The 2.5D and 3D integration approaches allow for increased I/O bandwidth compared to the 2D approach because connections between the driving electronics and the optics occur over a two-dimensional area rather than a one-dimensional edge. Additionally, flip chip bumps allow for a denser pitch than wirebonding. Utilizing 12.5 μm wire with a staggered dual bond pad approach allows for wirebonding to obtain a pitch of 25 μm [78]. High density Cu-Cu bonding has been demonstrated with a 10 μm pitch [79]. The flip chip approach produces a 250-fold increase in the number of I/Os for a square chip. Bandwidth densities of over 1 Tbps/mm^2 have been demonstrated using dense 3D integration with 50 μm pitch microbumps [80].

A main advantage of the MCM approach is that the platform provides further integration for extreme scale systems. To interface into the components for an extreme scale system (CPUs, GPUs, FPGAs, and memory), connectivity must be established with the driving electronics. For 2D integration, this condition would require an additional set of wirebonds between the system components and the driving electronics placing an additional limit on I/O connections. The 3D integration approach requires either wirebonding from the photonic chip to the system components or removing the photonic substrate to provide backside connectivity through through-silicon vias (TSVs). Both approaches place a limit on I/O as wirebonds are limited to edges and solder ball connections from the backside of the photonic chip to the PCB require a pitch of several hundred microns at a minimum. With the MCM integration, bare die system components can be flip chipped to the interposer with micro solder bumps or copper pillars, maximizing I/O. Parasitics can be kept to a minimum because system components can be placed as close as several hundred microns from the driving electronics chip. The different integration approaches introduce different thermal management challenges. The 2D approach keeps the photonics and electronics largely thermally isolated. The 3D approach introduces additional challenges for heat extraction for the photonics. The electronic chip will dissipate a significant amount of heat, which introduces additional challenges for the thermal stabilization of microrings on the photonic chip.

Interposers can also be fabricated with a silicon nitride waveguide to allow optical redistribution, allowing multiple photonic chips to be integrated into the same MCM as well as lasers to be integrated within the same interposer. In this approach, the interposer has a trench etched out to allow for the PIC to be placed within the interposer. The SiN edge couplers in the interposer produce the same spot size as the SiN edge couplers of the PIC, and the trench allows the waveguides in the PIC and interposer to be at the same height, enabling simple butt-coupling. Coupling to the interposer can be achieved with fibers or fiber arrays to the SiN edge couplers in the interposer, in a similar method as coupling to the PIC when it is flipped on top of the interposer (Figure 6).

IV. Silicon Photonic-Switched Network Architectures for Extreme Scale Systems

The interconnection network architecture and the efficient utilization of network resources are another crucial piece for the realization of the energy efficient and high-bandwidth optical network vision. In this section we show how the integration of silicon photonic switches within networks allows for energy-efficient management of network resources through the use of flexible, reconfigurable topologies, allowing bandwidth resources to be utilized in an optimal manner without the need for over-provisioning. This results in a sufficient quantity of data to be supplied to the processors in the system, allowing them to operate at their maximum potential, which lowers application execution time and saves energy consumption as a result.

We begin our discussion by presenting state-of-the-art SiP switches, which are the crucial enabler for resource allocation.

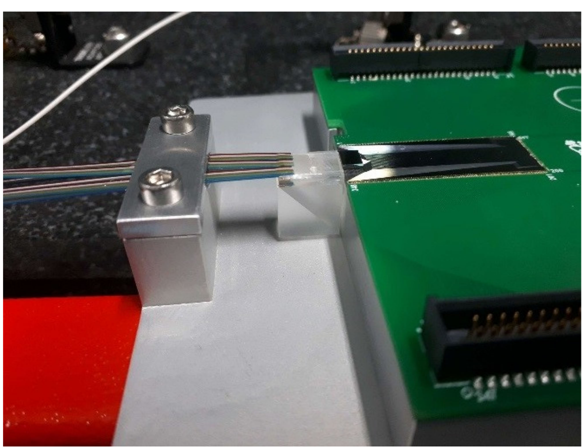

Fig. 6. A DC switch PIC flipped onto an interposer without EICs. Coupling is achieved with a fiber array attached to the SiN edge couplers on the PIC, which is overhung off the interposer. This similar approach can be used for coupling to the SiN edge couplers on the interposer when the PIC is placed within the interposer trench.

Then we will present our network architecture that integrates them into computing environments at the supercomputer or intra-data center level. We show how reconfigurable networks enabled by SiP switches can provide significant performance improvements with a physical testbed demonstration.

A. Silicon Photonic Switches

SiP switches have potential advantages of low power consumption, small area footprint, low fabrication costs at large scales, and the potential for nanosecond range dynamic connectivity [81]. Silicon exhibits a strong thermo-optic (T-O) coefficient (1.8×10^{-4}/K) [82] which can be leveraged to tune the phase of the light passing through the switch in tens of microseconds. SiP switches use the plasma dispersion effect through carrier injection or depletion for nanosecond scale switching times.

One of the first planar integrated Mach-Zehnder Interferometer (MZI) switching circuits of 32×32 connectivity was realized with T-O phase tuners in 2015 [83]. It included 1024 MZI components in a path-independent loss (PILOSS) topology offering strictly non-blocking connections. The current record for thermo-optic silicon switches is a 64×64 implementation of MZI elements in the Bene topology [84]. For fast E-O phase shifting, carrier-injection based PIN junctions are more widely applied than carrier-depletion-type, PN junctions. PIN junctions show higher efficiencies in nm/V terms, leading to a much smaller footprint and/or a lower operating voltage. Monolithic 16×16 E-O Bene switch devices based on MZIs have been demonstrated [85], [86]. More recently, a 32×32 version was proposed [87]. These demonstrations have shown the feasibility of high-level integration in SiPs. However, performance is still limited by high insertion loss and crosstalk. Performance in terms of optical signal quality can be improved by dilating [88] and nesting [89] the 2×2 elementary switch units, by modifying the control plane through introducing push-pull drive [90], or with means of advanced routing algorithm [81], [91].

The largest-scale SiP switch fabric reported to date is the Micro-Electro-Mechanical System (MEMS)-actuated cross-bar switch with 128×128 connectivity [92], constructed with vertical and horizontal waveguide buses with low-loss multi-mode-interference (MMI) crossings. Adiabatic couplers placed at bus intersections enable low-loss and high-extinction coupling; however the actuation voltages on the order of tens of volts are still a challenge.

Switches with the highest number of ports enable the greatest flexibility. With the increase of switch radixes, efficient calibration/testing methods becomes crucial. We proposed algorithm-based methods for device calibration and characterization to eliminate built-in power checkpoints [93]–[95].

The space explored by switching metrics can be ultimately reduced to a cost-per-port/power penalty [96]. For optical switching to be competitive, the cost associated with making a network configurable must be considerably smaller than the cost of adding additional links that lead to similar network performance. Also, the incorporation of optical switches must fit within the optical link power budget, which will be discussed in Section IV-C.

The power consumption of the switch fabrics is also an important factor, which directly relates to the maintaining cost of data centers. The electro-optic switch circuit consumes as low as 7.7 mW per path [96]. Thermo-optic devices are generally more power hungry: the state-of-the-art MZI-based switch features a power consumption of approximately 60 mW [97]. The highly efficient phase tuning of ring resonators can be leveraged to reduce the power consumption. We have recently demonstrated a device with the power consumption of approximately 20 mW per path [89] and this number can be further reduced to approximately 6 mW by redesigning the ring element to have a smaller FSR.

Lastly, we have demonstrated microring based switch-and-select switching circuits (S&S) [96], [98]. This device is designed to for TE mode operation only but a polarization diversity design has been done and currently under evaluation. Scaling this structure requires adding only MRRs to the bus waveguide, which effectively reduces the scaling overhead in loss. Prototyped devices have been fabricated at the foundry of the American Institute of Manufacturing (AIM Photonics). All designs use the pre-defined elements in the PDK library to ensure high yield and low cost, and followed the design rules of standard packaging houses, i.e. Tyndall National Institute, to be compatible with low-cost packaging solutions. Excellent testing results were achieved for the fully packaged 4×4 S&S switch with on-chip loss as low as 1.8 dB and crosstalk ratio as –50 dB [96], [98], [99] This device incorporates an array of standard edge couplers with a 5.5 dB coupling loss, which, however, can be significantly reduced with the novel 3D polymer assisted couplers [100].

B. Integration of Silicon Photonic Switches for Bandwidth Steering

Here, we present a network architecture that integrates both MZI or MRR-based silicon photonic switches within a

packet-switched environment to create a flexible network topology that can adapt to different application traffic characteristics, resulting in improved application execution time [10], [101]–[103]. In [10], [101]–[103], we used the standard Dragonfly network topology as a starting point, a topology commonly considered for HPC systems. The Dragonfly topology [104] is a hierarchical topology that organizes the network into groups of top-of-rack (ToR) packet switches, with each ToR switch connecting to server nodes. Within the group, ToR switches are fully connected with each other. Each group is connected to every other group with at least one link. This topology provides high-connectivity with all-to-all global links at the inter-group level to minimize hop count.

The advantages of high-connectivity, however, are diluted by low per-link bandwidth. In particular, the bandwidth of inter-group (global) links, carrying all the traffic between two large sets of routers groups, becomes the most scarce resource and can create a bottleneck for the entire network. One reason for the bandwidth bottleneck within inter-group links is the highly skewed traffic characteristics present in nearly every HPC application [9]. Skewed traffic characteristics equate to traffic matrices that focus traffic on only a small percentage of inter-group links, so that certain links are highly congested while most others are severely under-utilized. Such traffic characteristics also remain the same for a large portion of the application's runtime, or even for its entirety. Due to the skewed traffic, current best-for-all approach characterized by static, over-provisioned networks will become a bottleneck for the next-generation extreme scale computing platforms [105]–[108].

In response to this observation, there has been various efforts in using optical circuit switching for system level networks connecting racks. They can be categorized into two groups: the first group uses optical links to carry elephant flows in order to reduce the load on the electronic packet switch (EPS), shown in [109]–[111]. The second group of works connects optical circuit switches in order to connect different electronic packet switches, such as [112]–[114]. Many of these works in both categories rely on centralized large port-count switches such as 3D MEMS whose scalability is constrained. The Flexfly architecture shows that low-radix SiP switches distributed across the network can be used to enable significant performance improvements. One recent work that follows a similar concept is [115], which features low-radix switches integrated into data centers in order to reconfigure the network from Fat-Tree to Random Graph topologies based on the application that is being operated.

Achieving a network design that properly balances traffic is a challenging problem. Over-provisioning the network incurs unnecessary energy and infrastructure costs as well as wasted network resources [116], while under-provisioning limits system performance because data-starved processors are unable to leverage their full computing power. We introduce a photonic architecture, Flexfly [117], that allows for flexible global links among groups through the insertion of low-radix silicon photonic switches at the inter-group network layer which enables different groups to be connected. Flexfly does not introduce new bandwidth - it takes under-utilized links and reassigns them to intensively communicating group pairs in a process called bandwidth steering, illustrated in Figure 7. The bandwidth steering concept rests upon the assumption that there are both congested links and under-utilized links at the same time during an application run, which occurs due to a skewed traffic pattern. If an application's traffic is evenly distributed across the entire network, then bandwidth steering and flexible network topologies would not be able to provide any performance benefit.

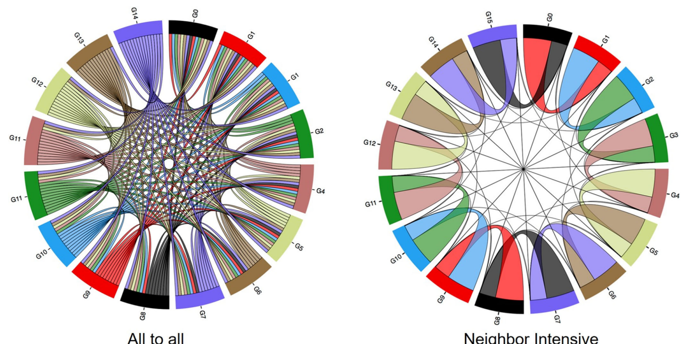

Fig. 7. Standard Dragonfly topology with all-to-all inter-group links (left) and re-configured topology after bandwidth steering focusing on neighbor-intensive traffic (right) [10].

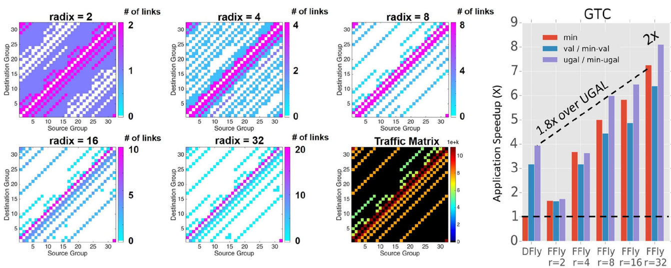

Fig. 8. (Left) The first five traffic matrices shows the physical network topology adapting to the traffic matrix of the GTC application [120] (shown at bottom right) with increasing silicon photonic switch radices [10]. (Right) Performance improvement of various switch radices for different routing mechanisms over standard Dragonfly.

The insertion of SiP switches allows us to take advantage of its near data rate transparent property to perform bandwidth steering on the large quantities of aggregate WDM traffic of entire Dragonfly router groups per each input port. As the optical switch performs its switching functions purely on the physical layer, it is transparent to higher network layers and applications, which reduces the latency, energy consumption, and routing complexity compared to performing the same task in with another electronic packet switch. Ultimately, the SiP switch enables dynamic physical network topology adaptation that matches the application's unique traffic matrix. Additionally, unlike previous optical switching solutions that rely on large port counts [118], [119], Flexfly is designed to support the use of low-radix silicon photonic switches, realizable through low-cost CMOS fabrication technology.

Figure 8 shows how the physical connections of the network topology approaches the traffic matrix of the GTC benchmark application, becoming increasingly optimized for this application's traffic characteristics. With increasing switch radix, more links can be connected to the photonic switch and resulting in

TABLE I
NUMBER OF SWITCHES AND CONNECTORS PER BLADE FOR DIFFERENT G (DRAGONFLY GROUPS) AND r (GROUPS PER CABINET ROW)

G	r	# of supergroups	# of switches	# of connectors
8	4	2	7	56
16	4	4	15	120
16	8	2	15	240
32	4	8	31	248
32	8	4	31	496

the network becoming more flexible as a whole. As can be seen in the bottom right of Figure 8, the traffic matrix for the network at a switch radix of 32 is identical to the traffic matrix of the GTC application [120]. Although it seems that there is little difference in the traffic matrix diagrams of radix = 8 to radix = 32, the colors for each graph are normalized despite different number of links. With a switch of radix = 8 there are a maximum of 8 links available to be steered from under-utilized connections (any white squares) and be given to the most intensive traffic shown at the diagonal. With a switch of radix = 32 however, there are 20 links available to be allocated to the most intensive traffic, which means that not only more bandwidth can be steered, it also allows the network controller to have finer granularity in its resource allocation. The graph on the right shows the speedup of Flexfly with different switch radices as compared to the standard Dragonfly. We use minimal, Valient, and Universal Globally-Adaptive Load-balanced (UGAL) routing for Dragonfly, and minimal, Min-Val, and Min-UGAL routing for Flexfly. The Dragonfly with minimal routing is normalized to 1.0x. As can be seen, the speedup of Flexfly increases with higher switch radix, and Flexfly with minimal routing used achieved 1.8x speedup over Dragonfly with UGAL routing. The hop count and cross-group message latency are also halved compared to the Dragonfly topology.

The scalability of the Flexfly architecture is described as follows: We divide a total of G Dragonfly groups into r groups per cabinet row, resulting in $\frac{G}{r}$ supergroups. A Flexfly switch blade is associated with each supergroup, which contains all the SiP switches associated with the inter-group links of that supergroup. This switch blade will have $G-1$ switches. With r groups per supergroup, each supergroup will have $r(G-1)$ links that fully connect to $G-1$ switches, each with r ports, and $2r(G-1)$ fiber connectors. Table I shows the number of switches and connectors per blade for different G and r values. The compatibility of SiP with CMOS foundries allows for a large number of SiP switches to be placed on a single chip. Looking at the values shown in Table I, the cost and space needed for incorporating $\frac{G}{r}$ switch blades in a Dragonfly topology HPC system are deemed to be scalable [10]."

C. High Performance Computing Physical Testbed Demonstration

A 32-server mini-supercomputer testbed was built with PICA8 electronic packet switches serving as ToR switches interconnected by a silicon photonic MRR-based switch (Figure 10). The testbed features a Dragonfly topology consisting of 4 groups

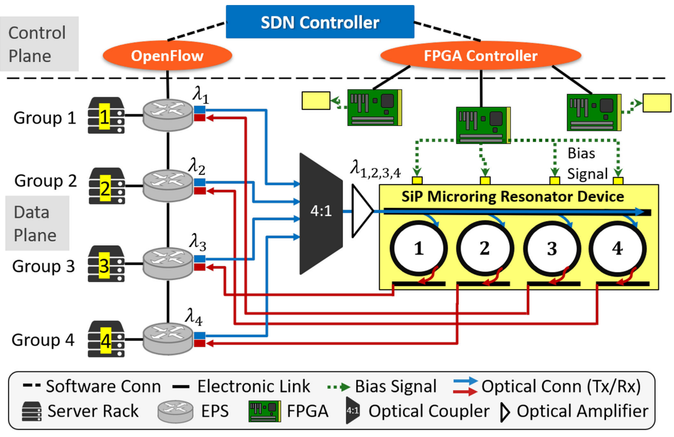

Fig. 9. Network architecture showing the connections of the servers and top-of-rack EPSs to the silicon photonic MRR device [102].

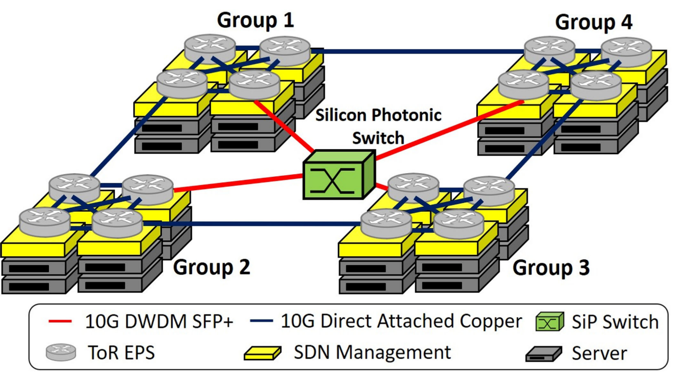

Fig. 10. Testbed consisting of a Dragonfly topology and a silicon photonic switch for inter-group reconfiguration [102].

each having 4 ToR EPSs. Inter-group links have 10 Gbps bandwidth and are a combination of both static (to support basic connectivity) and dynamic links (for bandwidth-steering using the silicon photonic switch). Dynamic links are simply optical links that are connected to the SiP switch, whose different switching configurations allow for different endpoints (Dragonfly groups) to be connected to each other, thereby changing the amount of total bandwidth available between different groups on a link granularity.

Note that although our testbed currently uses a SiP switch based on MRRs, the Flexfly architecture is not limited to any specific type of switching technology. For a practical installation of optical switches in the link, the key challenge would be the induced excess optical power penalty on the link budget. Ideally the maximum link budget available depends on the input laser power, to keep lower than a threshold to avoid the nonlinear effects, and the receiver sensitivity, the minimum optical power necessary in order to recover the signals. This ideal optical budget can be about 25 dB based on our link design presented in Section III-A, where the link components such as modulators and filters would introduce around 15 dB power penalty with 1 dB insertion loss per optical coupler, and leave less than 10 dB budget to optical switches. Currently, scalable SiP switch designs with a 32 × 32 port count have been demonstrated with

Fig. 11. Throughput over time of inter-group links over the run-time of GTC application for an all-to-all Dragonfly and a bandwidth-steered topology [103].

$<$13.2 dB of insertion loss [97], which is slightly larger than the available margin. However, with future low-loss architecture due to advanced fabrication techniques, the insertion of SiP switches within photonic link become viable.

The testbed setup consisting of the control and data planes is shown in Figure 9. The control plane consists of the software-defined networking (SDN) controller, which acts as the top-level management for both the electronic packet switches (EPSs) through the OpenFlow protocol, and the silicon photonic switches through an FPGA interface. During a switching operation, different server groups are chosen to be bi-directionally connected to each other by tuning the resonances of the microrings to the input wavelengths in different configurations, which redistributes the input signals to different output ports, allowing the device to act as a wavelength and spatial switch [102].

Figure 11(a) shows experiments where a skeletonized version of the GTC benchmark application [120] was operated over a baseline all-to-all Dragonfly topology versus a bandwidth-steered topology. The plots shows the traffic throughput during the time the application ran for each of the inter-group links in the testbed network (traditional Dragonfly topology at the top, and bandwidth optimized topology matching the application traffic at the bottom). For the purpose of demonstrating the effect of bandwidth steering on a physical testbed to its maximum which is limited in its capabilities compared to a real HPC system, the rank assignment of the GTC application to the physical machines was done to concentrate traffic and force congestion in the links between Groups 1 and 2, and Groups 3 and 4. Then, by configuring the silicon photonic switch to move under-utilized links between Groups 1 and 3 and Groups 2 and 4 to the heavily communicating Groups 1 and 2 and Groups 3 and 4, a 40% reduction in the application execution time was observed. Intuitively, a more severely congested link that is subsequently relieved by an additional link to share its load will have a higher performance improvement. Overall, our results show that our control plane can leverage silicon photonic switches to manage existing link resources effectively to enable performance benefits without adding bandwidth capacity. Further details can be found in [103].

The reconfiguration time of the control plane is on the order of hundreds of microseconds, while the SiP switch, configured with thermal tuning, has a switching latency on the order of single-digit microseconds [101]. It is possible to have used electro-optic tuning of the SiP switch instead, which would have a switching time of tens to hundreds of nanoseconds. However, HPC application jobs have typical runtimes on the order of hours, and their traffic characteristics are very stable, meaning that the links that are either congested or under-utilized remain so for large portions of the runtime of the application [10]. This allows a switching time of microseconds for traffic patterns that change on an hourly timeframe to be acceptable. Additionally, the stability of the traffic allows for the traffic pattern to be known a-priori if the rank assignment is known. The runtime middleware can also easily characterize the network traffic based on the first few iterations and subsequently perform the network topology reconfiguration [10]. However for traffic such as from a cloud data center which is more dynamic due to many users with different types of applications being operated, traffic matrices remain skewed but are no longer as stable as HPC traffic. This means that the optical circuit switch will need to reconfigure more frequently, so that the end-to-end switching latency due to the control plane and related hardware becomes more crucial.

Lastly, polarization diversity is also a potential roadblock to the practical installation of SiP switches in data center networks, since polarization-maintaining fibers (PMF) are too costly for data center applications. While currently, most of the demonstrations in SiP switch only support one (TE) polarization merely for the purpose of exploiting the feasibility of building

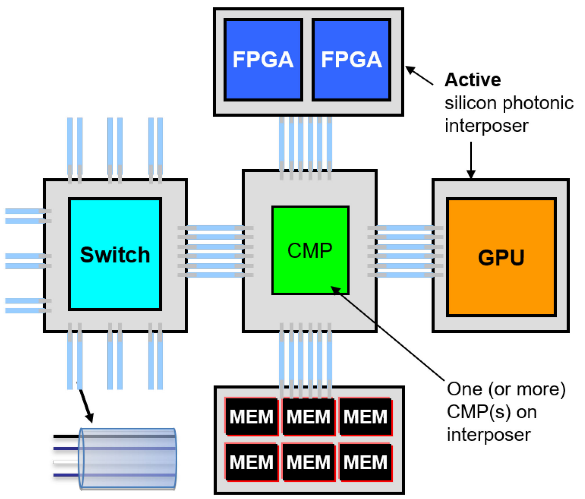

Fig. 12. Conceptual diagram of the optically connected multi-chip module (OC-MCM).

large-scale, high-performance devices, studies on the polarization diversity has been carried out by either co-integrated polarization splitter-rotators [121], [122] or eliminating polarization sensitive elements [123].

V. Disaggregated Server Concept and Future Work

In this section we present our concept of the Optically Connected Multi-Chip Module, or OC-MCM (Figure 12) intra-server architecture, where disaggregated compute resources are envisioned supported by interposers that are interconnected by high-speed optical links, and managed using a control plane leveraging a silicon photonic switching platform. There have been some efforts in a disaggregated network for the intra-data center level already [124], [125]. We focus mainly on the hardware infrastructure within an individual server node, although there is a great deal of challenges on other software aspects such as the operating system, drivers, and common protocols for data accessing [126].

In Section IV-B and IV-C we showed how silicon photonic switches within the packet switched network environment of a high performance system can enable significant performance improvements. This was partially enabled by the packet switches which served as the endpoints for the silicon photonic switch as well as the mature Layer 3 communication protocols that supports all desired topologies. Within the server node however, links between compute peripherals such as processor, memory and GPUs require tight integration with unique protocols designed for different types of components. The architecture of these networks within a server are therefore restricted by the silicon interposer technology which must support Tb/s interfaces in a millimeter scale and limited by printed-circuit-board (PCB) design constraints for Gb/s interfaces within few tens of centimeters. Electrical integration challenges also limit the amount of resources per node and network reconfiguration capabilities.

The ideal situation would be a merged interconnection network that supports any combination of a variable number of different components in a disaggregated manner. To enable such a system requires significant changes to both the operating system and drivers to support global addressing, common protocols for data accessing, and an open switching technology. The interconnect system must also allow for flexibility in the number of connection ports and data rates. Many organizations are currently working together to formulate an agreement on a common semantic/protocol, as well as an open switching platform, such as the Gen-Z consortium [127] and AMD's Infinity Fabric [128]. These efforts aim to commoditize computing components, blurring the limit between processors, memories, accelerators and network peripherals by creating a interconnection fabric that allows these devices to communicate with each other with simple commands. The landscape of computing devices needs to adapt to this as well - technology providers must allow for electronic components to be utilized in this new disaggregated system in their design considerations. Last but not least, the merged interconnection network must rest on a physical layer foundation that is capable of supporting a range of distances while remaining cost and power efficient. A promising enabler for such a technology is embedded silicon photonics.

Similar to the adaptive capabilities of the network architecture shown in the Flexfly concept, the ideal intra-node network should allow for support of dynamic reconfiguration of computing resources that is optimized for the application currently operating over the system. For example, if a compute system is engaged in a deep neural network (DNN) training phase, the topology can be reconfigured to focus on high connectivity between GPUs. Subsequently during the inferencing phase, the topology would reorganize to ensure high bandwidth between each GPU and memory. Additionally, the links can also be reconfigured to interconnect different types of resources. For example, during a DNN training phase, a network consisting of a large percentage of GPUs, an appropriate amount of memory, and a single CPU for orchestration can be allocated. Then, during the inferencing phase, computing resources switch over to a large percentage of FPGAs or ASICs supported by an appropriate amount of HBM and SSDs. If each resource can be provisioned dynamically, then the system can avoid wasting unused resources. We have demonstrated in a fully connected scenario that the optical switching capability of the silicon photonic platform has the potential to reduce the peer-peer latency by almost 25% [129].

However, many challenges remain - as was mentioned previously in Section IV-C, current SiP switches have insertion losses that are higher than the 10 dB optical margin shown by our link design. State of the art reports on the total insertion loss associated with SiP switches under both thermal- and electro-optic control show many with insertion loss values less than 20 dB [16]. Currently, this loss is too high to be integrated with our photonic links without amplification and still have a viable energy efficiency, but future work focusing on low-loss couplers and switch architectures can pave the way for a reconfigurable network that do not require an optical amplification. Recently, we have developed an accurate tapless and photodetector-less testing procedure that is based on the photoconductance effect of the SiP control elements which can further reduce the insertion loss and design complexity of high radix switches [130].

The integration of SiP switches with photonic links also has many fabrication challenges. Two approaches exist for integrating silicon photonic transceivers and switches. One option is for the photonic elements of the transceiver and the switch to share the same photonic integrated circuit (PIC). This approach can be seen in network-on-chip structures. A second option is for the transceiver and switch to be on different PICs but share the same package, with connectivity achieved using an interposer. One of the challenges for the first approach is the electrical signal routing. The number of electrical signals for a modest radix switch can easily be in the hundreds. Routing these signals as well as the DC electrical signals and RF signals due to the transceiver photonic elements is often a difficult hurdle. One of the obstacles for the second approach is the optical connectivity between the transceiver and the switch. Some foundries, such as AIM Photonics, use an interposer with a SiN waveguide to provide the optical connection between the two elements, and etch trenches within the interposer to allow for the PICs edge couplers to sit at the same height as the interposer waveguide. The challenge associated with this approach is the reflow of multiple PICs while maintaining optical alignment. Overall, many challenges will need to be addressed to realize a truly photonic interconnected disaggregated server architecture.

VI. Conclusion

The realization of next-generation extreme scale computing systems requires important developments in the system's ability to sustain its computational power by efficiently moving large amounts of data through low power, ultra high bandwidth connections. Silicon photonic interconnects are a candidate technology to deliver the necessary communication bandwidths with scalable energy efficiencies. The integration of silicon photonics requires extensive development of techniques to control and embed them within different electronic environments of the compute system. First we must develop methodologies to maximize the bandwidth capacity and energy efficiency of silicon photonic links, and develop methods to integrate the links to support communication between electronic endpoints at various network layers. We must also develop silicon photonic switching platforms and control planes to enable flexible network architectures that can optimize the utilization of network resources thereby supplying processors with sufficient data to maximize computational capability, resulting in minimal execution time and saving energy. These combined technologies are important and significant steps in the development of increased bandwidth and energy efficient interconnects for next-generation extreme scale computing platforms.

References

[1] P. M. Kogge and J. Shalf, "Exascale computing trends: Adjusting to the "new normal" for computer architecture," *Comput. Sci. Eng.*, vol. 15, pp. 16–26, 2013.

[2] Top500, 2018. [Online]. Available: https://www.top500.org. Accessed on: Sep. 10, 2018.

[3] S. Rumley *et al.*, "Optical interconnects for extreme scale computing systems," *Parallel Comput.*, vol. 64, no. C, pp. 65–80, May 2017. [Online]. Available: https://doi.org/10.1016/j.parco.2017.02.001

[4] System Overview Oak Ridge Leadership Computing Facility, 2019. [Online]. Available: https://www.olcf.ornl.gov/for-users/system-user-guides/summit/system-overview/. Accessed on: Oct. 23, 2018.

[5] J. J. Dongarra, M. A. Heroux, and P. Luszczek, "HPCG benchmark a new metric for ranking high performance computing systems," *Int. J. High Performance Comput. Appl.*, vol. 30, Aug. 2015. doi: 10.1177/1094342015593158.

[6] Green500 List – June 2018 – Top500 Supercomputer Sites, 2019. [Online]. Available: https://www.top500.org/green500/list/2018/06/. Accessed on: Aug. 25, 2018.

[7] Wireline Link Survey, 2016. [Online]. Available: https://web.engr.oregonstate.edu/anandt/linksurvey/. Accessed on: Aug. 26, 2018.

[8] A. Gazman *et al.*, "Software-defined control-plane for wavelength selective unicast and multicast of optical data in a silicon photonic platform," *Opt. Express*, vol. 25, no. 1, pp. 232–242, Jan. 2017. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-25-1-232

[9] S. Kamil, A. Pinar, D. Gunter, M. Lijewski, L. Oliker, and J. Shalf, "Reconfigurable hybrid interconnection for static and dynamic scientific applications," in *Proc. 4th Int. Conf. Comput. Frontiers*, 2007, pp. 183–194. [Online]. Available: http://doi.acm.org/10.1145/1242531.1242559

[10] K. Wen *et al.*, "Flexfly: Enabling a reconfigurable dragonfly through silicon photonics," in *Proc. Int. Conf. High Perform. Comput. Netw. Storage Anal.*, Nov. 2016, pp. 166–177.

[11] L. Robert *et al.*, "Top ten exascale research challenges," U.S. Department of Energy, Washington, DC, USA, DOE ASCAC Subcommitte Rep., Feb. 2014.

[12] C. A. Thraskias *et al.*, "Survey of photonic and plasmonic interconnect technologies for intra-datacenter and high-performance computing communications," *IEEE Commun. Surveys Tut.*, vol. 20, no. 4, pp. 2758–2783, Oct.–Dec. 2018.

[13] J. Lee, D. Shin, Y. Kim, and H. Yoo, "A 17.5 fJ/bit energy-efficient analog SRAM for mixed-signal processing," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2016, pp. 1010–1013.

[14] M. O'Connor *et al.*, "Fine-grained DRAM: Energy-efficient DRAM for extreme bandwidth systems," in *Proc. 50th Annu. IEEE/ACM Int. Symp. Microarchitecture*, 2017, pp. 41–54.

[15] Y. Arakawa, T. Nakamura, Y. Urino, and T. Fujita, "Silicon photonics for next generation system integration platform," *IEEE Commun. Mag.*, vol. 51, no. 3, pp. 72–77, Mar. 2013.

[16] Q. Cheng, M. Bahadori, M. Glick, S. Rumley, and K. Bergman, "Recent advances in optical technologies for data centers: A review," *Optica*, vol. 5, no. 11, pp. 1354–1370, Nov. 2018.

[17] Lightwave Research Laboratory. (2018). PhoenixSim. [Online]. Available: https://lightwave2.ee.columbia.edu/svn/columbia_research/

[18] M. Bahadori, S. Rumley, D. Nikolova, and K. Bergman, "Comprehensive design space exploration of silicon photonic interconnects," *J. Lightw. Technol.*, vol. 34, no. 12, pp. 2975–2987, Jun. 2016.

[19] W. Bogaerts *et al.*, "Silicon microring resonators," *Laser Photon. Rev.*, vol. 6, no. 1, pp. 47–73, 2012.

[20] G. T. Reed, G. Mashanovich, F. Y. Gardes, and D. Thomson, "Silicon optical modulators," *Nature Photon.*, vol. 4, no. 8, pp. 518–526, 2010.

[21] G. Masini *et al.*, "CMOS photonics for optical engines and interconnects," in *Proc. Opt. Fiber Commun. Conf*, 2012, Paper OTu2I-1.

[22] R. Wu, C.-H. Chen, J.-M. Fedeli, M. Fournier, R. G. Beausoleil, and K.-T. Cheng, "Compact modeling and system implications of microring modulators in nanophotonic interconnects," in *Proc. ACM/IEEE Int. Workshop Syst. Level Interconnect Prediction*, 2015, pp. 1–6.

[23] C.-H. Chen *et al.*, "DWDM silicon photonic transceivers for optical interconnect," *IEEE Opt. Interconnects Conf. (OI)*, San Diego, CA, Apr. 2015, pp. 52–53, doi: 10.1109/OIC.2015.7115682.

[24] A. Biberman, J. Chan, and K. Bergman, "On-chip optical interconnection network performance evaluation using power penalty metrics from silicon photonic modulators," in *Proc. Int. Interconnect Technol. Conf.*, 2010, pp. 1–3.

[25] Q. Li *et al.*, "Single microring-based 2 × 2 silicon photonic crossbar switches," *IEEE Photon. Technol. Lett.*, vol. 27, no. 18, pp. 1981–1984, Sep. 2015.

[26] M. Bahadori *et al.*, "Energy-performance optimized design of silicon photonic interconnection networks for high-performance computing," in *Proc. Conf. Des., Autom. Test Europe*, 2017, pp. 326–331. [Online]. Available: http://dl.acm.org/citation.cfm?id=3130379.3130456

[27] R. Ding *et al.*, "A compact low-power 320-Gb/s WDM transmitter based on silicon microrings," *IEEE Photon. J.*, vol. 6, no. 3, 2014, Art. 6600608.

[28] P. Dong *et al.*, "Low power and compact reconfigurable multiplexing devices based on silicon microring resonators," *Opt. Express*, vol. 18, no. 10, pp. 9852–9858, 2010.

[29] P. Dong *et al.*, "Low Vpp, ultralow-energy, compact, high-speed silicon electro-optic modulator," *Opt. Express*, vol. 17, no. 25, pp. 22 484–22 490, 2009.

[30] W. Bogaerts *et al.*, "Nanophotonic waveguides in silicon-on-insulator fabricated with CMOS technology," *J. Lightw. Technol.*, vol. 23, no. 1, pp. 401–412, Jan. 2005.

[31] B. G. Lee, A. Biberman, J. Chan, and K. Bergman, "High-performance modulators and switches for silicon photonic networks-on-chip," *IEEE J. Sel. Top. Quantum Electron.*, vol. 16, no. 1, pp. 6–22, Jan./Feb. 2010.

[32] Q. Xu, B. Schmidt, S. Pradhan, and M. Lipson, "Micrometre-scale silicon electro-optic modulator," *Nature*, vol. 435, no. 7040, pp. 325–327, 2005.

[33] S. Lin, E. Schonbrun, and K. Crozier, "Optical manipulation with planar silicon microring resonators," *Nano Lett.*, vol. 10, no. 7, pp. 2408–2411, 2010.

[34] G. Li *et al.*, "Ring resonator modulators in silicon for interchip photonic links," *IEEE J. Sel. Top. Quantum Electron.*, vol. 19, no. 6, pp. 95–113, Nov./Dec. 2013.

[35] L. Zhang, Y. Li, M. Song, J.-Y. Yang, R. G. Beausoleil, and A. E. Willner, "Silicon microring-based signal modulation for chip-scale optical interconnection," *Appl. Phys. A*, vol. 95, no. 4, 2009, Art. 1089.

[36] P. Dong *et al.*, "Wavelength-tunable silicon microring modulator," *Opt. Express*, vol. 18, no. 11, pp. 10 941–10 946, 2010.

[37] R. Osgood Jr., J. Dadap, A. Ahmed, and X. Meng, "New advances in nanophotonic device physics," in *Proc. 3rd ACM Int. Conf. Nanoscale Comput. Commun.*, 2016, pp. 22:1–22:2.

[38] A. Ahmed, X. Meng, J. I. Dadap, R. El-Ganainy, and R. M. Osgood, "Non-Hermitian parametric amplification via four wave mixing," in *Proc. Frontiers Opt.*, 2016, Paper FTu2D-4.

[39] C. Manolatou and M. Lipson, "All-optical silicon modulators based on carrier injection by two-photon absorption," *J. Lightw. Technol.*, vol. 24, no. 3, pp. 1433–1439, Mar. 2006.

[40] R. Soref and B. Bennett, "Electrooptical effects in silicon," *IEEE J. Quantum Electron.*, vol. 23, no. 1, pp. 123–129, 1987.

[41] S. Stepanov and S. Ruschin, "Modulation of light by light in silicon-on-insulator waveguides," *Appl. Phys. Lett.*, vol. 83, no. 25, pp. 5151–5153, Mar. 2003.

[42] Y. Liu *et al.*, "40-Gb/s silicon modulators for mid-reach applications at 1550 nm," in *Proc. IEEE Opt. Interconnects Conf.*, 2014, pp. 19–20.

[43] T. Baba *et al.*, "50-Gb/s ring-resonator-based silicon modulator," *Opt. Express*, vol. 21, no. 10, pp. 11 869–11 876, 2013.

[44] J. Sun *et al.*, "A 128 Gb/s PAM4 silicon microring modulator," in *Proc. Opt. Fiber Commun. Conf. Expo.*, 2018, pp. 1–3.

[45] X. Zheng *et al.*, "A tunable 1 × 4 silicon CMOS photonic wavelength multiplexer/demultiplexer for dense optical interconnects," *Opt. Express*, vol. 18, no. 5, pp. 5151–5160, 2010.

[46] J. Wang, "Recent progress in on-chip multiplexing/demultiplexing silicon photonic devices and technologies," in *Proc. Prog. Electromagn. Res. Symp.*, 2014, pp. 28–36.

[47] J. E. Cunningham *et al.*, "Highly-efficient thermally-tuned resonant optical filters," *Opt. Express*, vol. 18, no. 18, pp. 19055–19063, 2010.

[48] M. S. Dahlem, C. W. Holzwarth, A. Khilo, F. X. Kärtner, H. I. Smith, and E. P. Ippen, "Reconfigurable multi-channel second-order silicon microring-resonator filterbanks for on-chip WDM systems," *Opt. Express*, vol. 19, no. 1, pp. 306–316, 2011.

[49] A. W. Poon, F. Xu, and X. Luo, "Cascaded active silicon microresonator array cross-connect circuits for WDM networks-on-chip," *Proc. SPIE*, vol. 6898, 2008, Art. 689812.

[50] A. W. Poon, X. Luo, F. Xu, and H. Chen, "Cascaded microresonator-based matrix switch for silicon on-chip optical interconnection," *Proc. IEEE*, vol. 97, no. 7, pp. 1216–1238, Jul. 2009.

[51] A. Bianco, D. Cuda, R. Gaudino, G. Gavilanes, F. Neri, and M. Petracca, "Scalability of optical interconnects based on microring resonators," *IEEE Photon. Technol. Lett.*, vol. 22, no. 15, pp. 1081–1083, Aug. 2010.

[52] K. Yu *et al.*, "25 Gb/s hybrid-integrated silicon photonic receiver with microring wavelength stabilization," in *Proc. Opt. Fiber Commun. Conf.*, 2015, Paper W3A-6.

[53] Q. Xu, B. Schmidt, J. Shakya, and M. Lipson, "Cascaded silicon micro-ring modulators for WDM optical interconnection," *Opt. Express*, vol. 14, no. 20, pp. 9431–9436, 2006.

[54] K. Preston, N. Sherwood-Droz, J. S. Levy, and M. Lipson, "Performance guidelines for WDM interconnects based on silicon microring resonators," *CLEO: 2011 - Laser Sci. Photon. Appli.*, Baltimore, MD, May 2011, pp. 1–2, doi: 10.1364/CLEO_SI.2011.CThP4.

[55] M. Bahadori, D. Nikolova, S. Rumley, C. P. Chen, and K. Bergman, "Optimization of microring-based filters for dense WDM silicon photonic interconnects," in *Proc. IEEE Opt. Interconnects Conf.*, 2015, pp. 84–85.

[56] R. Hendry, D. Nikolova, S. Rumley, N. Ophir, and K. Bergman, "Physical layer analysis and modeling of silicon photonic WDM bus architectures," in *Proc. High Perform. Embedded Architecture Compil. Workshop*, 2014, pp. 20–22.

[57] A. Biberman, P. Dong, B. G. Lee, J. D. Foster, M. Lipson, and K. Bergman, "Silicon microring resonator-based broadband comb switch for wavelength-parallel message routing," in *Proc. IEEE 20th Annu. Meeting Lasers Electro-Opt. Soc.*, 2007, pp. 474–475.

[58] B. G. Lee, B. A. Small, K. Bergman, Q. Xu, and M. Lipson, "Transmission of high-data-rate optical signals through a micrometer-scale silicon ring resonator," *Opt. Lett.*, vol. 31, no. 18, pp. 2701–2703, 2006.

[59] M. Georgas, J. Leu, B. Moss, C. Sun, and V. Stojanović, "Addressing link-level design tradeoffs for integrated photonic interconnects," in *Proc. IEEE Custom Integrat. Circuits Conf.*, 2011, pp. 1–8.

[60] N. Ophir, C. Mineo, D. Mountain, and K. Bergman, "Silicon photonic microring links for high-bandwidth-density, low-power chip I/O," *IEEE Micro*, vol. 33, no. 1, pp. 54–67, Jan./Feb. 2013.

[61] C.-H. Chen *et al.*, "A comb laser-driven DWDM silicon photonic transmitter based on microring modulators," *Opt. Express*, vol. 23, no. 16, pp. 21 541–21 548, 2015.

[62] R. Hendry, D. Nikolova, S. Rumley, and K. Bergman, "Modeling and evaluation of chip-to-chip scale silicon photonic networks," in *Proc. IEEE 22nd Annu. Symp. High-Perform. Interconnects*, 2014, pp. 1–8.

[63] X. Meng, R. R. Grote, W. Jin, J. I. Dadap, N. C. Panoiu, and R. M. Osgood, "Rigorous theoretical analysis of a surface-plasmon nanolaser with monolayer MoS_2 gain medium," *Opt. Lett.*, vol. 41, no. 11, pp. 2636–2639, 2016.

[64] X. Meng, R. R. Grote, J. I. Dadap, and R. M. Osgood, "Threshold analysis in plasmonic nanolaser with monolayer semiconductor as gain medium," in *Proc. Joint Symp. Jpn. Soc. Appl. Phys. Opt. Soc. Amer.*, 2015, Paper 15p_2D_3.

[65] A. Ahmed *et al.*, "Differential phase-shift-keying demodulation by coherent perfect absorption in silicon photonics," *Opt. Lett.*, vol. 43, no. 16, pp. 4061–4064, 2018.

[66] B. Souhan, R. R. Grote, J. B. Driscoll, X. Meng, and R. M. Osgood, "Metal-semiconductor-metal monolithic silicon-waveguide photodiode design and analysis," in *Proc. Frontiers Opt.*, 2013, pp. FM3E-3.

[67] X. Meng *et al.*, "Emerging plasmonic applications explored with cluster parallel computing," in *Proc. Photon. North*, 2015, p. 1.

[68] X. Meng, A. Ahmed, J. Dadap, K. Bergman, and R. Osgood, "Dispersion engineering of silicon/plasmonics hybrid optical interconnections," in *Proc. Integrat. Photon. Res. Silicon Nanophoton.*, 2015, Paper IW2A-2.

[69] M. Bahadori, R. Polster, S. Rumley, Y. Thonnart, J.-L. Gonzalez-Jimenez, and K. Bergman, "Energy-bandwidth design exploration of silicon photonic interconnects in 65 nm CMOS," in *Proc. IEEE Opt. Interconnects Conf.*, 2016, pp. 2–3.

[70] S. Rumley, M. Bahadori, D. Nikolova, and K. Bergman, "Physical layer compact models for ring resonators based dense WDM optical interconnects," in *Proc. 42nd Eur. Conf. Opt. Commun.; Proc.* VDE, 2016, pp. 1–3.

[71] A. Yariv, "Universal relations for coupling of optical power between microresonators and dielectric waveguides," *Electron. Lett.*, vol. 36, no. 4, pp. 321–322, 2000.

[72] V. Stojanović *et al.*, "Monolithic silicon-photonic platforms in state-of-the-art CMOS SOI processes [invited]," *Opt. Express*, vol. 26, no. 10, pp. 13106–13121, May 2018. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-26-10-13106

[73] H. Li *et al.*, "A 25 Gb/s, 4.4 V-swing, ac-coupled ring modulator-based WDM transmitter with wavelength stabilization in 65 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 50, no. 12, pp. 3145–3159, Dec. 2015.

[74] I. Ndip, A. Öz, S. Guttowski, H. Reichl, K. Lang, and H. Henke, "Modeling and minimizing the inductance of bond wire interconnects," in *Proc. IEEE 17th Workshop Signal Power Integrity*, May 2013, pp. 1–4.

[75] S. Saeedi, S. Menezo, G. Pares, and A. Emami, "A 25 Gb/s 3D-integrated CMOS/silicon-photonic receiver for low-power high-sensitivity optical communication," *J. Lightw. Technol.*, vol. 34, no. 12, pp. 2924–2933, Jun. 2016.

[76] X. Zheng *et al.*, "Ultra-efficient 10Gb/s hybrid integrated silicon photonic transmitter and receiver," *Opt. Express*, vol. 19, no. 6, pp. 5172–5186, Mar. 2011. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-19-6-5172

[77] X. Zheng *et al.*, "Ultra-low power silicon photonic transceivers for inter/intra-chip interconnects," *Proc SPIE*, vol. 7797, 2010, Art. 779702.

[78] 2014. [Online]. Available: https://www.hesse-mechatronics.com/wp-content/uploads/bros-mk-bj653-en- 20181029.pdf. Accessed on: Jan. 20, 2019.

[79] J. D. Reed, M. Lueck, C. Gregory, A. Huffman, J. M. Lannon, and D. Temple, "High density interconnect at 10 m pitch with mechanically keyed Cu/Sn-Cu and Cu-Cu bonding for 3-D integration," in *Proc. 60th Electron. Compon. Technol. Conf.*, Jun. 2010, pp. 846–852.

[80] M. Rakowski *et al.*, "Hybrid 14 nm FinFET - silicon photonics technology for low-power Tb/s/mm^2 optical I/O," in *Proc. IEEE Symp. VLSI Technol.*, 2018, pp. 221–222.

[81] Q. Cheng, M. Bahadori, and K. Bergman, "Advanced path mapping for silicon photonic switch fabrics," in *Proc. Conf. Lasers Electro-Opt.*, 2017, Paper SW1O.5.

[82] M. Bahadori *et al.*, "Thermal rectification of integrated microheaters for microring resonators in silicon photonics platform," *J. Lightw. Technol.*, vol. 36, no. 3, pp. 773–788, Feb. 2018.

[83] K. Tanizawa *et al.*, "Ultra-compact 32 × 32 strictly-non-blocking Si-wire optical switch with fan-out LGA interposer," *Opt. Express*, vol. 23, no. 13, pp. 17599–17606, Jun. 2015. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-23-13-17599

[84] T. Chu, L. Qiao, W. Tang, D. Guo, and W. Wu, "Fast, high-radix silicon photonic switches," in *Proc. Opt. Fiber Commun. Conf*, 2018, Paper Th1J.4. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=OFC-2018-Th1J.4

[85] L. Qiao, W. Tang, and T. Chu, "16 × 16 Non-blocking silicon electro-optic switch based on Mach-Zehnder interferometers," in *Proc. Opt. Fiber Commun. Conf. Exhib.*, Mar. 2016, pp. 1–3.

[86] L. Lu *et al.*, "16 × 16 Non-blocking silicon optical switch based on electro-optic Mach-Zehnder interferometers," *Opt. Express*, vol. 24, no. 9, pp. 9295–9307, May 2016. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-24-9-9295

[87] L. Qiao, W. Tang, and T. Chu, "32 × 32 Silicon electro-optic switch with built-in monitors and balanced-status units," *Sci. Rep.*, vol. 7, Feb. 2017, Art. 42306.

[88] J. Xing, Z. Li, Y. Yu, and J. Yu, "Low cross-talk 2 × 2 silicon electro-optic switch matrix with a double-gate configuration," *Opt. Lett.*, vol. 38, no. 22, pp. 4774–4776, Nov. 2013.

[89] N. Dupuis *et al.*, "Ultralow crosstalk nanosecond-scale nested 2 × 2 Mach–Zehnder silicon photonic switch," *Opt. Lett.*, vol. 41, no. 13, pp. 3002–3005, Jul. 2016.

[90] J. Xing, Z. Li, P. Zhou, X. Xiao, J. Yu, and Y. Yu, "Nonblocking 4 × 4 silicon electro-optic switch matrix with push–pull drive," *Opt. Lett.*, vol. 38, no. 19, pp. 3926–3929, Oct. 2013. [Online]. Available: http://ol.osa.org/abstract.cfm?URI=ol-38-19-3926

[91] Q. Cheng, M. Bahadori, Y. Huang, S. Rumley, and K. Bergman, "Smart routing tables for integrated photonic switch fabrics," in *Proc. Eur. Conf. Opt. Commun.*, Sep. 2017, pp. 1–3.

[92] K. Kwon *et al.*, "128 × 128 Silicon photonic MEMS switch with scalable row/column addressing," in *Proc. Conf. Lasers Electro-Opt.*, 2018, Paper SF1A.4.

[93] Y. Huang, Q. Cheng, N. C. Abrams, J. Zhou, S. Rumley, and K. Bergman, "Automated calibration and characterization for scalable integrated optical switch fabrics without built-in power monitors," in *Proc. 43th Eur. Conf. Opt. Commun.*, 2017, pp. 1–2.

[94] Y. Huang, Q. Cheng, and K. Bergman, "Crosstalk-aware calibration for fast and automated functionalization of photonic integrated switch fabrics," in *Proc. Conf. Lasers Electro-Opt.*, 2018, Paper STh3B.6. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=CLEO_SI-2018-STh3B.6

[95] Y. Huang, Q. Cheng, and K. Bergman, "Automated calibration of balanced control to optimize performance of silicon photonic switch fabrics," in *Proc. Opt. Fiber Commun. Conf.*, 2018, Paper Th1G.2. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=OFC-2018-Th1G.2

[96] Q. Cheng, S. Rumley, M. Bahadori, and K. Bergman, "Photonic switching in high performance datacenters [invited]," *Opt. Express*, vol. 26, no. 12, pp. 16022–16043, Jun. 2018. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-26-12-16022

[97] K. Suzuki *et al.*, "Low insertion loss and power efficient 32 × 32 silicon photonics switch with extremely-high-δ PLC connector," in *Proc. Opt. Fiber Commun. Conf. Postdeadline Papers*, 2018, Paper Th4B.5. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=OFC-2018-Th4B.5

[98] Q. Cheng *et al.*, "Si/SiN microring-based optical router in switch-and-select topology," in *Proc. Eur. Conf. Opt. Commun.*, 2018, pp. 1–3.

[99] Q. Cheng *et al.*, "Ultralow-crosstalk, strictly non-blocking microring-based optical switch," *Photon. Res.*, vol. 7, no. 2, pp. 155–161, Feb. 2019.

[100] O. A. J. Gordillo, M. A. Tadayon, Y.-C. Chang, and M. Lipson, "3D photonic structure for plug-and-play fiber to waveguide coupling," in *Proc. Conf. Lasers Electro-Opt.*, 2018, Paper STh4B.7. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=CLEO_SI-2018-STh4B.7

[101] Y. Shen *et al.*, "Software-defined networking control plane for seamless integration of multiple silicon photonic switches in datacom networks," *Opt. Express*, vol. 26, no. 8, pp. 10 914–10 929, Apr. 2018. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-26-8-10914

[102] Y. Shen *et al.*, "Autonomous dynamic bandwidth steering with silicon photonic-based wavelength and spatial switching for datacom networks," in *Proc. Opt. Fiber Commun. Conf.*, 2018, Paper Tu3F.2. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=OFC-2018-Tu3F.2

[103] Y. Shen, S. Rumley, K. Wen, Z. Zhu, A. Gazman, and K. Bergman, "Acceleration of high performance data centers using silicon photonic switch-enabled bandwidth steering," in *Proc. 44th Eur. Conf. Opt. Commun.*, 2018, pp. 1–2.

[104] J. Kim, W. J. Dally, S. Scott, and D. Abts, "Technology-driven, highly-scalable dragonfly topology," in *Proc. 35th Int. Symp. Comput. Architecture*, 2008, pp. 77–88.

[105] J. H. Ahn, N. Binkert, A. Davis, M. McLaren, and R. S. Schreiber, "HyperX: Topology, routing, and packaging of efficient large-scale networks," in *Proc. Conf. High Performance Comput. Netw., Storage Anal.*, 2009, pp. 1–11.

[106] B. Arimilli *et al.*, "The PERCS high-performance interconnect," in *Proc. IEEE 18th Symp. High Performance Interconnects*, 2010, pp. 75–82.

[107] M. Besta and T. Hoefler, "Slim fly: A cost effective low-diameter network topology," in *Proc. Int. Conf. High Performance Comput., Netw., Storage Anal.*, 2014, pp. 348–359.

[108] S. Rumley, D. Nikolova, R. Hendry, Q. Li, D. Calhoun, and K. Bergman, "Silicon photonics for exascale systems," *J. Lightw. Technol.*, vol. 33, no. 3, pp. 547–562, Feb. 2015.

[109] N. Farrington *et al.*, "Helios: A hybrid electrical/optical switch architecture for modular data centers," in *Proc. ACM SIGCOMM Conf.*, 2010, pp. 339–350. [Online]. Available: http://doi.acm.org/10.1145/1851182.1851223

[110] N. Farrington *et al.*, "A 10 s hybrid optical-circuit/electrical-packet network for datacenters," in *Proc. Opt. Fiber Commun. Conf. Expo. Nat. Fiber Opt. Engineers Conf.*, Mar. 2013, pp. 1–3.

[111] G. Wang *et al.*, "c-Through: Part-time optics in data centers," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, Aug. 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=2043164.1851222

[112] Y. Xia, M. Schlansker, T. S. E. Ng, and J. Tourrilhes, "Enabling topological flexibility for data centers using omniswitch," in *Proc. 7th USENIX Workshop Hot Topics Cloud Comput.*, 2015. [Online]. Available: https://www.usenix.org/conference/hotcloud15/workshop-program/presentat ion/xia

[113] C. Minkenberg *et al.*, "Performance benefits of optical circuit switches for large-scale dragonfly networks," in *Proc. Opt. Fiber Commun. Conf. Exhib.*, Mar. 2016, pp. 1–3.

[114] J. Shalf, S. Kamil, L. Oliker, and D. Skinner, "Analyzing ultra-scale application communication requirements for a reconfigurable hybrid interconnect," in *Proc. ACM/IEEE Conf. Supercomputing*, Nov. 2005, p. 17.

[115] Y. Xia, X. S. Sun, S. Dzinamarira, D. Wu, X. S. Huang, and T. S. E. Ng, "A tale of two topologies: Exploring convertible data center network architectures with flat-tree," in *Proc. Conf. ACM Special Interest Group Data Commun.*, 2017, pp. 295–308. [Online]. Available: http://doi.acm.org/10.1145/3098822.3098837

[116] K. Wen *et al.*, "Reuse distance based circuit replacement in silicon photonic interconnection networks for HPC," in *Proc. IEEE 22nd Annu. Symp. High-Perform. Interconnects*, 2014, pp. 49–56.

[117] K. Wen *et al.*, "Flexfly: Enabling a reconfigurable dragonfly through silicon photonics," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, 2016, pp. 166–177.

[118] Y. Xia, M. Schlansker, T. E. Ng, and J. Tourrilhes, "Enabling topological flexibility for data centers using omniswitch," in *Proc. USENIX Workshop Hot Topics Cloud Comput.*, 2015.

[119] C. Minkenberg *et al.*, "Performance benefits of optical circuit switches for large-scale dragonfly networks," in *Proc. Opt. Fiber Commun. Conf.*, 2016, Paper W3J-3.

[120] GTC, 2013. [Online]. Available: http://www.nersc.gov/users/computational-systems/cori/nersc-8-procurement/trinity-nersc-8-rfp/nersc-8-trinity-benchmarks/gtc/. Accessed on: Sep. 27, 2018.

[121] K. Suzuki *et al.*, “Polarization-diversity 4 × 4 Si-wire optical switch,” in *Proc. Int. Conf. Photon. Switching*, 2015, pp. 121–123.
[122] K. Tanizawa, K. Suzuki, K. Ikeda, S. Namiki, and H. Kawashima, “Non-duplicate polarization-diversity 8 × 8 Si-wire PILOSS switch integrated with polarization splitter-rotators,” *Opt. Express*, vol. 25, no. 10, pp. 10885–10892, May 2017. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-25-10-10885
[123] S. Han, T. J. Seok, K. Yu, N. Quack, R. S. Muller, and M. C. Wu, “Large-scale polarization-insensitive silicon photonic MEMS switches,” *J. Lightw. Technol.*, vol. 36, no. 10, pp. 1824–1830, May 2018.
[124] K. Katrinis *et al.*, “Rack-scale disaggregated cloud data centers: The dredbox project vision,” in *Proc. Des., Autom. Test Eur. Conf. Exhib.*, Mar. 2016, pp. 690–695.
[125] Y. Yan *et al.*, “All-optical programmable disaggregated data centre network realized by FPGA-based switch and interface card,” *J. Lightw. Technol.*, vol. 34, no. 8, pp. 1925–1932, Apr. 2016.
[126] B. Abali, R. J. Eickemeyer, H. Franke, C. S. Li, and M. Taubenblatt, “Disaggregated and optically interconnected memory: When will it be cost effective?” *CoRR*, vol. abs/1503.01416, 2015, [Online]. Available: http://arxiv.org/abs/1503.01416
[127] Gen-Z Technology | Gen-Z Consortium, 2019. [Online]. Available: https://genzconsortium.org/about-us/gen-z-technology/. Accessed on: Aug. 29, 2018.
[128] AMD’s Infinity Fabric Detailed - The Innovative, Real-World Implementation of the Company’s ‘Perfect Lego’ Philosophy, 2019. [Online]. Available: https://wccftech.com/amds-infinity-fabric-detailed/. Accessed on: Aug. 29, 2018.
[129] E. Anderson, J. Gonzalez, A. Gazman, R. Azevedo, and K. Bergman, “Optically connected and reconfigurable GPU architecture for optimized peer-to-peer access,” in *Proc. Int. Symp. Memory Syst.*, 2018, p. W3J.3.
[130] A. Gazman *et al.*, “Tapless and topology agnostic calibration solution for silicon photonic switches,” *Opt. Express*, vol. 26, no. 25, pp. 32 662–32 674, Dec. 2018.

Yiwen Shen (SM’18) received the B.S. degree from the University of Toronto, Toronto, ON, Canada, in 2014, and the M.S. degree from Columbia University, New York, NY, USA, in 2015, both in electrical engineering. He is currently working toward the Ph.D. degree at Lightwave Research Lab, Columbia University. His current research interests involve the development of network architectures focused software-defined networking (SDN), silicon photonic interconnects and their integration within high-performance systems and data center environments.

Xiang Meng (M’19) received the B.Sc. degree in computer science and the B.Eng. degree in electrical engineering from the University of Saskatchewan, Saskatoon, SK, Canada, in 2011, and the M.S. and Ph.D. degrees in electrical engineering from Columbia University, New York, NY, USA, in 2012 and 2017, respectively. His research interests include scientific parallel computing and numerical analysis on emerging nanophotonic devices, ranging from nanolasers, nano-sensors, to high-speed optical transceivers and energy efficient photonic interconnects mainly for applications in high-performance computing and data center platform.

Qixiang Cheng (M’18) received the B.S degree from the Huazhong University of Science and Technology, Hubei, China, in 2010 and the Ph.D. degree from the University of Cambridge, Cambridge, U.K., in 2014, in the field of III/V integrated optical switches. He was a Research Assistant at the University of Cambridge. In March 2015, he joined Shannon Laboratory, Huawei, Beijing, China, researching future optical computing systems. He is currently a Research Scientist at the Lightwave Research Laboratory, Columbia University, New York, NY, USA. He has authored and co-authored more than 60 papers. His current research interests include design, simulation, and characterization of large-scale optical integrated devices for data center and optical computing applications.

Sébastien Rumley studied in Lausanne, Zurich (ETHZ) and Santiago de Chile (PUC). He received the M.S. degree in communication systems and the Ph.D. degree in computer and communication sciences, both from the Ecole Polytechnique Fdrale de Lausanne (EPFL), Switzerland, in 2005 and 2011, respectively. He is currently an Associate Research Scientist in the Department of Electrical Engineering, Columbia University, New York, NY, USA. His research focuses on optical interconnect.

Nathan Abrams received the B.S. degree in electrical engineering from Columbia University, New York, NY, USA, and the B.A. degree in natural mathematics and sciences from Whitman College, Walla Walla, WA, USA, both in 2014. He is currently working toward the M.S. and Ph.D. degrees at Columbia University. His research interests relate to photonic devices.

Alexander Gazman received the B.S. degree in electrical engineering from Boston University, Boston, MA, USA, in 2013, and the M.S. degree in electrical engineering from Columbia University, New York, NY, USA, in 2015, where he is currently working toward the Ph.D. degree at the Lightwave Research Laboratory. His research interests include developing silicon photonic subsystems and integration of optical interconnects in high-performance computing systems.

Evgeny Manzhosov received the B.Sc. degree in physics and the B.Sc. degree in electrical engineering from the Technion – Israel Institute of Technology, Haifa, Israel, in 2014, majoring in VLSI and circuits design. He is currently working toward the M.Sc./Ph.D. degree at Columbia University, New York, NY, USA, with focus on photonic devices and systems. Since 2011, he worked at Intel, Apple, and Cisco as a Physical Design Engineer.

Madeleine Strom Glick (M’99–SM’16) received the Ph.D. degree in physics from Columbia University, New York, NY, USA, for research on electro-optic effects of GaAs/AlGaAs quantum wells. After receiving the degree, she joined the Department of Physics, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland, where she continued her research in electro-optic effects in GaAs and InP-based materials. From 1992 to 1996, she was a Research Associate with CERN, Geneva, Switzerland, as part of the Lightwave Links for Analogue Signal Transfer Project for the Large Hadron Collider. From 2002 to 2011, she was a Principal Engineer at Intel (Intel Research Cambridge UK, Intel Research Pittsburgh) leading research on optical interconnects for computer systems. Her research interests are in applying photonic devices and interconnects to computing systems.

Dr. Strom Glick is a Fellow of the Institute of Physics and a Senior Member of OSA.

Keren Bergman (S’87–M’93–SM’07–F’09) received the B.S. degree from Bucknell University, Lewisburg, PA, USA, in 1988, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1991 and 1994, respectively, all in electrical engineering. She is currently a Charles Batchelor Professor at Columbia University, New York, NY, USA, where she also directs the Lightwave Research Laboratory. She leads multiple research programs on optical interconnection networks for advanced computing systems, data centers, optical packet switched routers, and chip multiprocessor nanophotonic networks-on-chip.

Dr. Bergman is a Fellow of OSA.