Silicon Photonics for Exascale Systems

Sébastien Rumley, Dessislava Nikolova, Robert Hendry, Qi Li, *Student Member, IEEE*, David Calhoun, *Member, IEEE*, and Keren Bergman, *Fellow, IEEE*

(Invited Tutorial)

***Abstract*—With the extraordinary growth in parallelism at all system scales driven by multicore architectures, computing performance is increasingly determined by how efficiently high-bandwidth data is communicated among the numerous compute resources. High-performance systems are especially challenged by the growing energy costs dominated by data movement. As future computing systems aim to realize the Exascale regime—surpassing 10^{18} operations per second—achieving energy efficient high-bandwidth communication becomes paramount to scaled performance. Silicon photonics offers the possibility of delivering the needed communication bandwidths to match the growing computing powers of these highly parallel architectures with extremely scalable energy efficiency. However, the insertion of photonic interconnects is not a one-for-one replacement. The lack of practical buffering and the fundamental circuit switched nature of optical data communications require a holistic approach to designing system-wide photonic interconnection networks. New network architectures are required and must include arbitration strategies that incorporate the characteristics of the optical physical layer. This paper reviews the recent progresses in silicon photonic based interconnect devices along with the system level requirements for Exascale. We present a co-design approach for building silicon photonic interconnection networks that leverages the unique optical data movement capabilities and offers a path toward realizing future Exascale systems.**

***Index Terms*—Exascale high performance computing, interconnection networks silicon photonics, optical interconnects.**

I. Introduction

THE number of floating point operations realized per second (Flop/s) is a relatively simple way to denote the raw power of a computer. It is a wide spread ranking criteria for Supercomputing systems around the world [1], used to designate the "world's most powerful computer". By following these rankings over time, one can observe that the TeraFlop (10^{12}) mark was attained in June 1997, and PetaFlop (10^{15}) in June 2008. Logically, the next barrier to break is the ExaFlop mark (10^{18}), and if the trend is conserved, it should be attained in 2019.

Manuscript received August 11, 2014; revised October 3, 2014; accepted October 3, 2014. Date of publication October 19, 2014; date of current version February 17, 2015. This work was supported in part by the U.S. Department of Energy Sandia National Laboratories under contracts PO 1426332 and PO 1319001. 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. The authors further acknowledge support by MIT Lincoln Laboratory under contract PO MIT-7000135026.

The authors are with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: sr3061@columbia.edu; dnn2108@columbia.edu; rh2519@columbia.edu; ql2163@columbia.edu; dmc2202@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.2014.2363947

The most recent world leading supercomputer, the Tianhe-2, realized 33 PetaFlops/s (PF) in June 2013. In terms of exponential growth, this marks the half-way point to *Exascale* computing (i.e. realizing one ExaFlop/s - EF). However, it is widely acknowledged that the second half of the way toward Exascale will be significantly more challenging [2]–[4]. Reaching Exascale will require simultaneous solutions to several complex problems that range from the management of hardware failures at runtime [5], to dealing with bit flips [6], to identifying the adequate massively parallel programming approaches [7]. Power is the central constraint, as the overall machine power consumption must be maintained within a "manageable" envelope of 20–50MW [8].

Perhaps the most critical barrier toward realizing Exascale computing is the fundamental challenge of data movement. With the extraordinary growth in parallelism at all system scales, even performance of today's systems is increasingly determined by how data is communicated among the numerous compute resources rather than by the total raw computation resources available. With the progress made on transistor miniaturization and integration, the energy cost associated with a computing operation has been drastically reduced in the last several decades (e.g. $\sim$0.5 pJ for a 32-bit addition [9]). In contrast, the cost associated with moving the operands, even on chip, has not been reduced in the same proportions. Within the upcoming CMOS fabrication platforms, Borkar [10] predicts a 6x improvement factor for compute energy over the transition from 45 to 7 nm fabrication technologies, compared to a mere 1.6x factor for communications. At Exascale, communication challenges are daunting, as energy consumption is completely dominated by costs of data movement and is thereby the main constraint on ultimate performance [11].

This has deep implications in the CPU and CMP (Chip Multi Processor) architectures. Higher transistor counts per chip *will* be attained in the upcoming years, but chip computing capabilities *will not* grow in proportion.

At the single CPU level, boosting compute power mainly means increasing the *instruction level parallelism*, as it is generally recognized that clock speeds are at the maximum; however, this comes at a price. More and more complex logic is required to detect and solve the instruction dependencies (or conversely, to identify the instructions that can be parallelized). At some point, the cost of this extra logic can negate the benefits of the additional parallel instruction execution units. Therefore, most of the 30x improvement factor required to reach Exascale computing will be provided by increased CPU parallelism. In fact, CPU parallelism is already the main path toward scaling computing systems. Fig. 1 illustrates that core parallelism

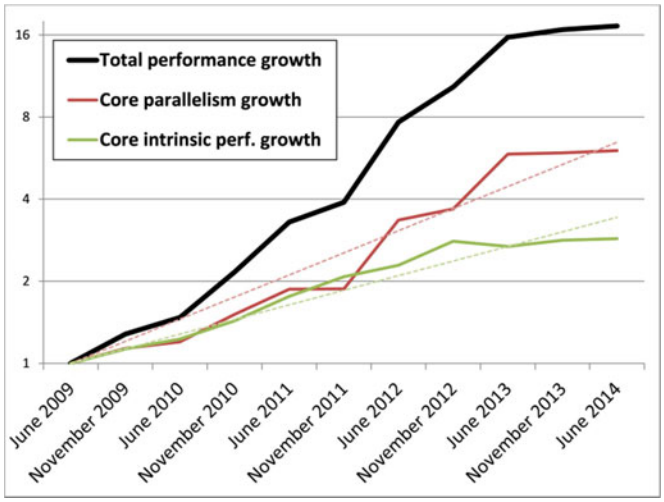


Fig. 1. Average LINPACK performance (black line) and number of cores (red line) of the top-20 supercomputers over the last years [1]. The intrinsic core performance curve (green line) is obtained by dividing the global performance by the core count. Trends are represented in dotted lines.

and core performance growths each have contributed to global performance growth over the last five years. Until 2011, core and parallelism growths progressed similarly; however, parallelism has progressed faster in recent years, while core intrinsic performance has stalled. In 2007 only one supercomputer incorporated more than 100 000 cores; in 2011, sixteen were above this mark. In June 2013, six systems were counting 500 000 cores or more; among them, Tianhe-2 with more than three million. According to current forecasts, Exascale-capable systems will involve nearly twenty million cores. This a factor of 6 compared to Tianhe-2, and more than all other factors combined (these factors are described in section II).

With growing CPUs the performance scalability of multicore chip architectures is becoming increasingly constrained by limitations in power dissipation, chip packaging, and the data throughput achievable by the off-chip interconnection networks. As a result, a significant part of the additional die area, obtained thanks to tighter integration, will be used to expand chip caches, in order to improve the data reuse and thus compensate for the lack of bandwidth and long latencies when going off-chip. The off-chip communication bottleneck, a major challenge due to limited on-chip power budget and pinout, becomes a key scaling barrier to memory bandwidth and system wide communications, themselves exacerbated by the growing CPUs. Furthermore, at the application layer, keeping millions of CPUs busy simultaneously requires a true work of art on the programming side, but above all an extremely agile (and therefore uncongested) communication layer to ensure tight synchronization.

In summary, realizing Exascale performance will essentially require extreme scaling of the communication capabilities at an almost unchanged energy budget (Tianhe-2 already consumes nearly 18 [1] – 2MW below the aforementioned manageable envelope).

Among the technologies emerging toward creating a fundamentally energy efficient interconnect, photonics is perhaps the most promising to enable a transition to a new generation of scaled extreme performance computing systems. Optical links are already present in most recent supercomputers, primarily in the form of active optical cables [12]. They are used to reduce the cabling density and the energy cost of transmission over meters-long distances. However, optical systems have not penetrated deeply within the supercomputer interconnection network yet, mainly due to the bulkiness and high cost of end-point systems.

This is subject to change thanks to the recent achievements in photonic device integration and fabrication, especially on silicon platforms. Unlike prior generations of photonic technologies, recent breakthroughs in *silicon photonics* offers the possibility of creating highly-integrated platforms with dimensions and fabrication processes compatible with electronic logic and memory devices. During the past decade, a series of major breakthroughs in silicon photonic devices have demonstrated that all the components that are necessary to build chip-scale photonic interconnect components (e.g. modulators, filters, switches, detectors) can be fabricated using common CMOS processes.

Turning silicon photonics into the interconnection network technology for next-generation systems requires numerous efforts on various fronts. Progress on the mass fabrication of reliable and low cost integrated silicon photonic subsystems is required. Optical interconnection networks are not a one-to-one replacement with existing electronic-based architectures, thus necessitating the development of methodologies and modeling approaches to effectively combine network architectures, topologies and arbitration strategies that incorporate the optical physical layer characteristics. At the application layer, parallel programs will likely have to be adapted to leverage the unique aspects of data movement in the optical domain.

This paper offers an overview of the research progress with respect to these diverse fronts and provides a roadmap for extending these efforts toward realizing silicon photonic enabled Exascale computing. The paper begins with quantitatively defining the system level requirements for data movement (see Section II). Section III introduces the relevant silicon photonic interconnect device technologies. In Section IV, modeling and design approaches for conceiving the first network blocks are presented. General directions to scale out these blocks to cover the full interconnect are reviewed in Section V, where we further introduce the concept of hardware-application co-design for photonic-enabled systems. Section VI provides initial insights on how futuristic designs can progressively be validated in a hardware/software implementation platform. Section VII summarizes and concludes the paper.

II. Exascale in Numbers

This section aims to provide a quantitative summary of Exascale systems requirements. We note that covering all aspects of supercomputer performance is beyond the scope of this paper and therefore limit our analysis to the main descriptors (number of cores, clocking frequency).

One way of summarizing the capabilities of a supercomputer is to express them as the following product:

$$\mathrm{P_{tot}} = ILP \cdot TLP \cdot f \cdot \mathrm{C_N} \cdot \mathrm{N} \cdot \eta. \quad (1)$$

TABLE I
AGGREGATED COMPUTING POWER OF CMP CHIPS

Name	Frequency f	ILP	TLP	Resulting computing power per chip P_{chip}
Ivy Bridge	4 Ghz	8	6	192 GF
Xeon Phi	1.1 Ghz	16	∼50*	∼880 GF*
Opteron	2.8 Ghz	4	16+	179 GF
Best-of-all	*4 Ghz*	*16*	*50*	*3.2 TF*
Exascale	*4 Ghz*	*20*	*125*	*10 TF*

*The number of cores in Xeon Phi varies across versions. It can be up to 61. The version deployed in Tianhe-2 uses only 50.
+ The Opteron implements an intermediate concept of module which regroups two cores. This blurs the distinction between ILP and TLP.

TABLE II
CHARACTERISTICS OF LEADING SUPERCOMPUTERS

Name	P_{chip}	C_N	N	η	Resulting computing power P_{flop}
Tianhe-2	1.15 TF	3	16 000	61%	33 PF
Sequoia	0.2 TF	1	98 304	85%	17 PF
Titan	0.72 TF*	2	18 688	64%	17 PF
Best-of-all	1.15 TF	3	98,304	85%	288 PF
Exascale 1	*1.15 TF*	*3*	*475 000*	*61%*	*1 EF*
Exascale 2	*10 TF*	*8*	*15 000*	*84%*	*1 EF*
Exascale 3	*10 TF*	*1*	*120 000*	*84%*	*1 EF*

*Titan regroups a CPU chip and an accelerator chip within each node. The value indicated here is the average of the two chip computing power.

The three first elements of this product reflect the computing capabilities of a single CMP chip, and is denoted P_{chip} hereafter. *ILP* stands for *Instruction Level Parallelism* and denotes the highest number of instructions that can be *terminated* in parallel within a single core at each clock cycle (its unit are Flops/cycle). *TLP* stands for *thread level parallelism*. It is a unitless factor that represents the highest number of threads that can be advanced in parallel, which equals the number of cores present on the chip. The clock frequency of the chip is designated as *f* (and given in Hz or cycles/s). The product (*ILP* · *TLP* · *f*) returns the *peak* computing capability of the chip in Flops/s. The notion of *peak* power must be introduced because, in practice, *ILP* instructions do not always leave the CPU pipeline at each cycle. The instruction throughput is maximal only when the CPU pipeline is fully loaded, and when all loaded instructions can be executed independently. Similarly, the thread level parallelism is maximal only if enough threads occupy the cores, and if none of the threads is stalling, for example waiting for a message to arrive.

Table I summarizes the *ILP*, *TLP* and *f* values of the CMP chips equipping leading supercomputers. At the chip scale, the typical computing capability is ∼TF (TeraFlop/s). This indicates that the total number of chips required for Exascale is on the order of one million. Table I also displays a "best-of-all" virtual chip that would combine the high clocking frequency of the Ivy Bridge chip with the massive *ILP* and *TLP* of the Xeon Phi. Even so, the chip computing power would not exceed 4 TF. Scaling the chip compute performance will be significantly determined by the off-chip I/O capabilities and particularly the memory bandwidth. Current efforts focus on scaling the memory bandwidth. The recently released hybrid memory cube (HMC) [13] stacks memory in a 3-D structure jointly with a high speed memory controller and interface. This approach permits to reduce the distances and to better exploit the memory space. This novel architecture is expected to allow chip level computing capabilities approaching 10TF (last line of Table I).

Referring back to (1), the two terms C_N and N denote the number of chips assembled inside a *node* and the number of nodes in the system, respectively. The specific definition of a *node* varies across systems and vendors, but generally refers to a compute unit with its own RAM memory, a network interface, and that is capable of running its own instance of the operating system. *Nodes* may also include local long term storage capacities, or dedicated power supplies.

The last parameter, η, represents the efficiency of a system at executing a given task. It expresses the part of the peak computing effectively exploited in practice. The efficiency η is an experimentally measured parameter, which encompasses all uncertainties about the way a workload is executed. It is therefore extremely hard to predict. In the context of the top500 ranking, it is measured while executing the LINPACK workload. Although workload-related, η reflects the parallel agility of a system. A low value of η is synonymous with poor utilization of the available computing power, e.g. due to an under-dimensioned communication layer. Conversely, an η value close to one indicates nearly optimal use of the computing resources.

Table II displays the C_N, N, and (LINPACK measured) η values of the leading supercomputers, together with the peak computing capability of their chip, and their resulting performance P_{flop}. Similar to Table I, it also displays the values of a hypothetical supercomputer combining all "best-of-class" characteristics. If the numerous nodes of Sequoia could be equipped with the same node processing power of Tianhe-2, and the efficiency η kept to around 85%, 288 PF could be achieved, i.e. a 3.5x factor away from the Exascale goal. In fact, to achieve Exascale by scaling Tianhe-2, 475 000 nodes would be required (fifth line of Table II—design option 1 for Exascale). In an alternative design, if the chip could scale to 10 TF, eight such chips aggregated within a node, and the efficiency raised to Sequoia's level, the node count N could be reduced to approximately 15 000 (design option 2).

These projections are quite abstract. In particular, the efficiency should be subject to extreme caution. Even if the η values have been kept at adequate levels in many recent supercomputers *while executing LINPACK* (see Fig. 2), it is unclear how parallel efficiency would evolve for systems involving tens of millions of cores. Nevertheless, these predictions provide an estimate for the size of the network interconnecting the nodes. Unless radical transformations occur in the CPU and/or node architecture, a network interconnecting around 100 000 endpoints will be required. Additionally, an intra-node interconnect will be required if the number of chips per node C_N is pushed beyond 4–8. Under this limit, point-to-point links can be leveraged, as shown in Fig. 3. Note that pertinence of employing such a

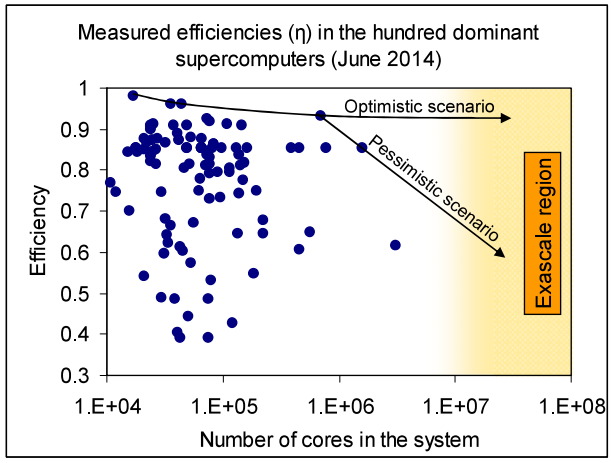


Fig. 2. Efficiencies at executing LINPACK for the 100 dominant supercomputers (as of June 2014 [1]), plotted against the number of cores involved. In the ideal case, efficiencies will be kept high while the number of cores is pushed to the Exascale region.

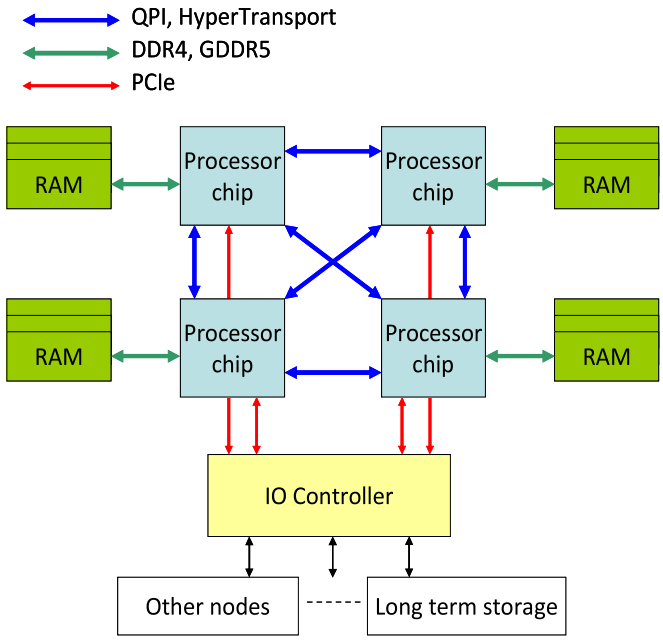


Fig. 3. Typical architecture of a four socket based node, with point-to-point links interconnecting most components [2].

"secondary" interconnect (internal to the node) can be questioned: nodes may also be organized around a single chip, as in the Sequoia Supercomputer today. In that case, the interconnect size will be pushed beyond the 100 000 mark, even with powerful 10 TF chips (last line of Table II).

Besides sizing the interconnection network, predictions in terms of required bandwidth can be derived from existing systems. The per-node available bandwidth, in *Bytes/s* (*B/s*), can be related to the node computing capability, in *Flops/s*, yielding to a *Byte/Flop* (*B/F*) metric. This metric is most often used to characterize the bandwidth connecting each CMP with its RAM

TABLE III
NETWORK CHARACTERISTICS OF LEADING SUPERCOMPUTERS

Name	*Compute power per node*	IO bandwidth per node	Raw data-movement capabilities	Average capabilities
Tianhe-2	3.45 TF	3.5 GB/s	0.001 B/F	0.0005 B/F
Sequoia	0.2 TF	20 GB/s	0.097 B/F	0.009 B/F
Titan	1.44 TF	25 GB/s	0.017 B/F	0.001 B/F
Exascale 1	*3.45 TF*	*350 GB/s*	*0.1 B/F*	–
Exascale 2	*80 TF*	*80 GB/s*	*0.001 B/F*	–

(memory bandwidth), and the one present between a node and the interconnection network (network injection bandwidth).

In the Sequoia supercomputer, the memory bandwidth yields to 0.2 *B/F* and the Xeon Phi architecture equipping Tianhne-2 nodes supports ∼0.25 *B/F*. In the latter case, using six GDDR5, a memory bandwidth of 6×48 GB/s $= 288$ GB/s $= 2.3$ Tb/s is obtained. Hence, to realize similar *B/F* ratios for Exascale systems, bandwidths of tens of Tb/s will be required to interconnect the TF capable chips to memory.

As for the interconnection network, in the Sequoia case, nodes are organized in a 5D-torus topology so each node has ten 2GB/s links available (two per dimension). Since a Sequoia node has a computing capability of 205 GF, the interconnection network offers a raw capacity of ∼0.1 *B/F*. Part of this capacity, however, is taken up by transiting traffic, since the architecture is a direct network (the notion of direct network is described in Section V; readers can also refer to [14]). In the presence of uniform traffic, for instance, transiting traffic represents approximately 91% of the total traffic[1], reducing the useful *B/F* to ∼0.01. In the Tianhne-2, reports indicate a bandwidth of 3.5 GB/s for connecting each 3.45 TF capable node to a first level switch, and 1 GB/s to the fat-tree top. This leads to ∼0.001 *B/F* as the best case ratio, and likely half this value for uniform or adversarial traffic. Finally, for TITAN, each node benefits from approximately 25 GB/s of raw bandwidth or 0.017 *B/F*. In presence of uniform traffic, transiting traffic would account for 94% reducing the available B/F ratio to approximately 0.001*B/F*.

These numbers are summarized in Table III, and compared with the potential Exascale scenarios of Table II. If node computing capabilities cannot be scaled (3.45 TF per node as for Tianhe-2—*Exascale 1* case), more data movement capabilities will likely be required to synchronize the 475,000 nodes corresponding to this scenario. By considering a value of 0.1 *B/F*—similarly to Sequioa, which also involves many nodes—350 GB/s (2.8 Tb/s) are required. If, in contrast, node capabilities are scaled to 80TF (*Exascale 2* case), the number of nodes to synchronize is reduced to 15 000. By considering the *B/F* value of Tianhe-2, which has a similar number of nodes, the per-node

[1]91% is obtained through the following calculation. Sequoia has 98,304 nodes organized as a hypercube of width w $= \sqrt[5]{98,304}$. On average, w nodes share a dimension and thus are connected in a ring topology. Moving in this ring, i.e. along one dimension, takes w/4 hops in average (w/2 in the worst case). A random displacement in the full topology will consists of moving the average distance in each dimension. Therefore, the average hop count in Sequoia is estimated by 5w/4. Resulting traffic is not in transit only for the last hop, so the proportion of transit traffic is (5w/4 - 1)/(5w/4) = 91%.

bandwidth required is 80 GB/s (0.64 Tb/s). In both cases higher underlying parallelism (compared to Sequoia and Tianhne-2) will likely translate in higher *B*/*F* requirements.

This analysis generally indicates that baseline rates of network links in Exascale systems will be on the order of at least Tb/s. This is clearly within the realm of optical systems. Even with extremely high modulation speed, electrical links will have to leverage *spatial parallelism* (i.e. several lanes) to realize such bandwidths. Optical systems, in contrast, can exploit the "additional" WDM dimension and therefore deliver extreme bandwidths densities. Exploiting the CMOS compatibility of silicon photonics, WDM optical links offering bandwidth of 1 Tb/s or more can penetrate directly on chip and the packaging issue. This suggests the possibility to further scale the critical *B*/*F* metric toward (or even beyond) a ratio of 1, and thus directly impacting system performance. The components enabling these concepts are reviewed in the next section.

III. Silicon Photonics Components for Exascale

Unlike prior generations of photonic technologies, the small device footprints, ultra-low capacitances, and tight proximity of electronic drivers in integrated silicon photonics offer fundamentally superior energy efficiency as well as the potential to bypass the electronic pin bandwidth bottleneck entirely. In this section we summarize the key advances in silicon photonic interconnect technologies.

Passive and active silicon photonics devices capable of all of the operations required for transmission and switching have been demonstrated in the past decade. These photonic devices are implemented on a silicon on insulator (SOI) which can be integrated with a computing chip in the CMOS layer, on top of the metal stack or in a different wafer 3-D integrated with CMOS circuitry. The optical wave is guided on the chip within silicon single- or multi-mode waveguides [15], [16]. Different geometries or materials lead to different propagation losses [17]. The fabrication process influences the losses as the roughness of the waveguide's walls can induce undesired scattering [18].

A multi-wavelength laser source can be off-chip [19] or integrated partially on chip as it has been recently demonstrated [20]. Typical commercial laser efficiencies are on the order of 1% (primarily since they have not been optimized for wall-plug energy efficiency), but significant ongoing research efforts ensue to improve these values. The laser light is coupled to the on-chip waveguides through grating couplers or using tapered fibers. Due to the mode mismatch this coupling can lead to significant loss but recent advances in this technology have shown that 1 dB coupling losses are feasible [21]. Other passive guiding structures like splitters/combiners (Y junction, multimode couplers) and waveguide crossings optimized for low loss and low cross talk have been designed and demonstrated [22]–[24]. The photodetection can be done off chip using Ge photodetectors [25]–[27] or with waveguide integrated receivers [28]. Recently graphene has also been studied for realizing high speed broadband photodetectors [29], [30].

If multi-wavelength signals are utilized, each wavelength needs to be filtered before being coupled out to the photodetector. Silicon microring resonators [31] or arrayed waveguide gratings (AWG) [32] can act as passive filters. The geometry of the device determines which wavelengths will be filtered. For example, in the case of a silicon microring the resonant wavelengths are mainly determined by the ring radius.

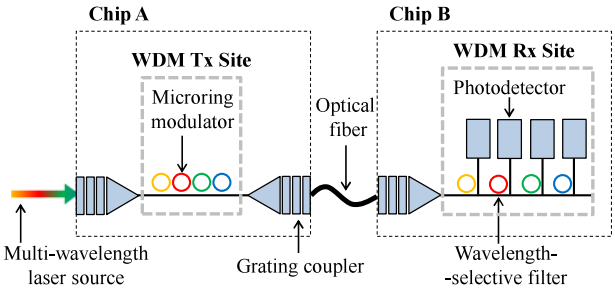


Fig. 4. A silicon photonic transmission link realized with a multi-wavelength laser source coupled onto a chip via a grating coupler and guided via a waveguide to a modulator bank realized by microring resonators tuned on the specific frequencies. The modulated light is coupled out from the chip and guided via an optical fiber to another chip where wavelength filters realized as microring resonators filter the desired wavelengths for detection.

The demonstration of high speed silicon photonic modulators based on Mach-Zehnder interferometers (MZI) [33] and microrings [34] paved the way to apply this technology to transmission links. These devices can perform amplitude and phase modulation, demonstrated at rates reaching 60 Gb/s per wavelength [35]. Other types of silicon integrated modulators have also been demonstrated including microdisc [36] and photonic crystal [37] structures, as well as waveguide-based Ge electro-absorption modulators [38]. Silicon photonic devices like Mach-Zehnder interferometers and optical resonators are uniquely suitable for WDM. These devices have an extremely small footprint (down to several micrometers wide) which results in low power operation and minimal silicon real estate, allowing integration of thousands of such components on a single die. For example, dense WDM modulation can be accomplished by cascading microring modulators on the same waveguide and tuning each ring to resonate on separate wavelengths [39], [40]. Systems with four [41] and eight [42] WDM channels have been demonstrated, the latter achieving 320 Gb/s transmission. These demonstrations indicate that bandwidths up to several Tb/s can be obtained. A schematic view of a transmission link implemented with the silicon photonic components discussed is shown on Fig. 4.

The possibility to spatially switch the signal on chip can significantly increase the range of possible link architectures, system functionalities and throughput [43]. Silicon photonic switches have also been demonstrated [44] including higher radix switches based on microrings or Mach-Zehnder interferometers with 2, 4 and 8 ports [45]–[47]. The scalability of microring [48], [49] and Mach-Zehnder [50], based switches has been investigated, showing the feasibility of implementing high-radix optical switches.

Resonator based devices suffer from thermal fluctuations. Several methods to mitigate and/or control this dependence have been proposed including the use of dithering signal, heterodyne detection and athermal materials [51]–[53]. A survey of the emerging efforts to resolve thermal issues is given in [54].

The insertion of photonic interconnection networks fundamentally changes the energy scaling rules for the system interconnect: once a photonic path is established, the data are transmitted end-to-end without the need for power consuming repeaters, regenerators, or buffers. Hence optics can make the energy cost for off-chip communication nearly the same as for on-chip communication.

IV. Design and Modeling of Silicon Photonics Based Systems

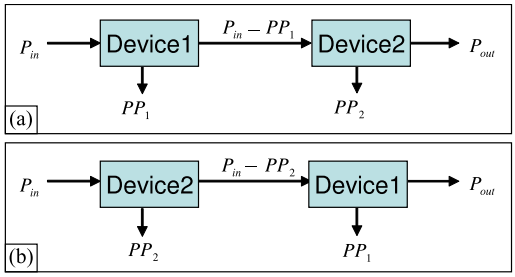
As discussed in the previous section, silicon photonic devices have been shown capable of realizing most of the operations required for creating optical interconnection networks. Most silicon photonic chips fabricated thus far, however, have been produced in academic and research environments and have generally never included more than a few tens of components. In the context of integrating silicon photonics in HPC platforms, the fabricated chips will include a significantly larger number of devices and will require optimization that is driven by system requirements (over even application requirements, as is developed in Section V). Therefore, a methodology that allows highly integrated chips to be optimized at design time is desirable. In this section, we present an approach toward developing such a methodology.

A. Individual Device Modeling

The methodology begins by characterizing each device (modulator, switch, coupler, etc.) along an optical path by its *power penalty*. This is a common approach to modeling optical systems [55]–[57]. As optical signals propagate through optical systems they tend to attenuate, undergo unwanted phase shifts, and accumulate noise. All of these effects degrade the quality of the signal, increasing the probability of transmission errors (i.e. increasing the bit-error-rate, or BER). The *power penalty* denotes the extra optical signal power that would be required to compensate these effects. It is important to note that sometimes additional power *cannot* compensate for distortion. In this case, the power penalty is assumed to be infinite.

Based on this concept, an optical system can be characterized by summing up power penalties of all devices through which a signal propagates until it is finally detected and converted back into the electrical domain. This approach enables a simple initial exploration of the optical system feasibility. It also permits to identify the device(s) that are the dominant contributors to the performance degradation, and hence require major optimization effort.

Device models capturing the main factors that determine power penalty can be based on physical principles, on experimental measurements, or a mix of both. The first group includes numerical methods (e.g. finite element methods) that are computationally intensive and therefore limited to small and specific device networks, but are also able to capture many details with great accuracy. The first group also includes computationally undemanding models abstracting the physical phenomenon via simplifying assumptions [48], [58]. In this approach, devices can be described by their power penalty as a function of a small number of parameters (e.g. loss per centimeter). Finally, empirical models based on experimental measurements provide a first order approximation for devices whose physics aren't well understood yet. They also capture the uncertainties related to fabrication and provide data on the power consumption.



$$PP_{tot} = P_{in} - P_{out} = P_{in} - (P_{in} - PP_1 - PP_2) = PP_1 + PP_2 \quad (1)$$

$$PP_{tot}(P_{in}) = P_{in} - P_{out} = P_{in} - (P_{in} - PP_1(P_{in}) - PP_2(P_{in} - PP_1(P_{in}))) = PP_1(P_{in}) + PP_2(P_{in} - PP_1(P_{in})) \quad (2)$$

$$PP_{tot}(P_{in}) = PP_2(P_{in}) + PP_1(P_{in} - PP_2(P_{in})) \quad (3)$$

Fig. 5. If the power penalty PP_i of a device *i* can be established independently of the input power level, the compound power penalty of a system PP_{tot} can be derived straightforwardly: Equation (1) holds for systems a) and b) (P_{in} and P_{out} are the system input and output optical power respectively). In contrast, components inducing power penalties that depend on the absolute power require more complex calculations as the exact position of each device matters: Equations (2) and (3), corresponding to systems a) and b) respectively, are not equivalent as Dev1 and Dev2 positions are interchanged.

The power penalty approach describes each device through its *relative* power impact. However, this approach hardly supports devices sensitive to the *absolute* optical power level. The main illustration of this concerns silicon photonic waveguides. They are subject to non-linear effects (two photon absorption) above a certain optical power and thus cause power dependent distortions. This limitation is generally circumvented by defining, for each device, a range of power for which the *absolute* power sensitivity can be neglected. Taking waveguides as an example, their power penalty is then only considered as relative below a threshold (generally considered to 20dBm [59]). In return, power levels above this threshold are assumed intolerable.

Maintaining conditions in which the power penalty of a device is independent of the input power greatly simplifies photonic link modeling because links can be defined by accumulating the power penalty from each device regardless of device order [55]–[57]. Fig. 5 illustrates this point with an example.

Laser sources and photodetectors (i.e. link extremities) are not characeterized by power penalty. The laser is simply modeled as a continuous-wave source with its power efficiency as the main characterizing parameter. The photodetector can be reduced to an absolute optical power that guarantees the required transmission quality (e.g. bit-error-rate $< 10^{-12}$). This detector *sensitivity* is symbol-rate and modulation speed dependent.

B. Modeling Devices as Part of a WDM Link

Following the power penalty approach, and provided input power independence, all devices but the initial laser and end

receiver can be aggregated as a compound *link power penalty*. This *link power penalty* must fit within the *link maximal power budget*, which is defined as the difference (in dB) between highest and lowest power levels tolerated by any device along the link. The upper limit on the power level is determined by the aforementioned silicon waveguide threshold, generally fixed to 20dBm, while the lower limit is imposed by the receiver's photodetector sensitivity (typically -20dBm for a 10Gb/s on-off keying modulation).

The margin emerging between the *link power penalty* and the *maximal power budget*, if any, can be used in the link design process. The power budget can be adjusted by reducing the initial link input power (thus laser power). Alternatively, excess power can be used to transmit over multiple wavelength channels (WDM). Expressed in dB, the total optical power required at the WDM link end must be greater than the sum of 10log(N) and P_{det}, where N denotes the number of wavelengths and P_{det} the detector sensitivity[2].

The presence of WDM calls for refinements in the device models and consideration of the channel spacing. As WDM channels get closer, effects that degrade the optical signal arise (e.g. inter-channel crosstalk, channels attenuated by devices meant for other nearby frequencies). These effects have to be taken into account in the power penalty.

We consider microring based modulators as an example. In reference [60] an analytical model is provided for determining the power penalty of intermodulation crosstalk. The practical lower bound on the channel spacing is shown to be 50 GHz for 10 Gb/s modulation. Theoretical models—e.g. the one provided in reference [58], which models the frequency response of a ring resonator—allow the silicon photonic network designer to gain further insight on the trade-offs between WDM channel spacing and the power penalty. In contrast, empirical models taken with a single measurement inevitably lose accuracy as they are extrapolated further from the context of their original measurement.

Fig. 6 illustrates the effects such extrapolations can have on a link level model. A simple, ring resonator based WDM link connecting two chips is considered (10Gb/s per wavelength). If a constant, measurement based link power penalty of 15dB is assumed for this link, the maximum number of supportable wavelengths explodes as the power budget is enlarged (blue curve). Intermediate modeling approaches, assuming both fixed and wavelength dependent power penalties also diverge, as they do not properly capture cross-talk. In contrast, if crosstalk aware models [61], [62] are leveraged to estimate the link power penalty, the supportable number of channels is more modest. Moreover, it eventually saturates, consistent with the spectrum limitation on the link (red curve). When targeting the high bandwidth (dense WDM) requirements for Exascale systems, it is therefore imperative to use models that are sensitive to channel spacing.

[2]Adding 10log(N) to the detector sensitivity—thus multiplying by N in absolute power terms—ensures that each channel conserves enough power to be properly detected.

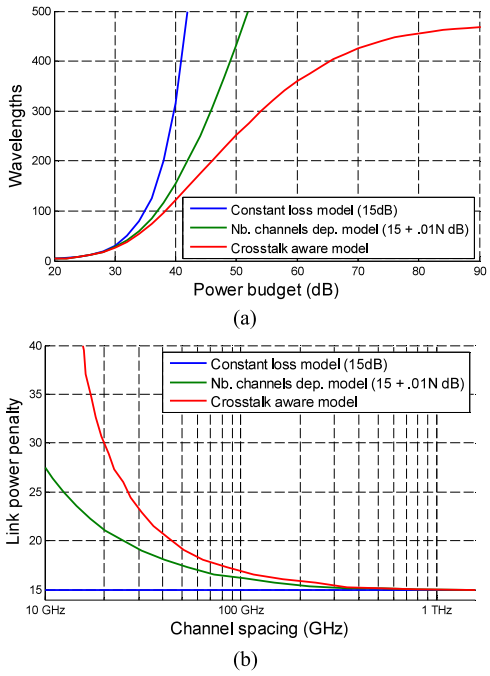


Fig. 6. (a) Maximum number of wavelengths (modulated at 10Gb/s with on-off keying) supported by a link under different device models. Experimentally derived constant power penalty models (blue curve), or constant + wavelength dependent part (green curve), are inaccurate when the number of wavelengths scales as they do not capture appropriately crosstalk related penalties (b) Calculated link power penalties as function of channel spacing. The crosstalk aware model (red curve) is valid for within a range of the link power budgets.

C. Network Transmission Modeling

MZI and microring resonator structures have also been demonstrated as switches. In a similar fashion to the aforementioned link, input power requirements and the general feasibility of a switched network design can be determined by composing device models and calculating the total power penalty experienced by each channel. In this case, the power penalty along the worst-case path through the network bounds the performance characteristics of the network.

When building networks of silicon photonic devices, additional cross-talk effects become important. While a switch is delivering a signal from one input port to an output port, some optical power may "leak" and propagate to other unintended outputs. If other signals are presently using these other outputs, the leaked power becomes cross-talk, for which a power penalty must be accounted. As these switches are cascaded to form more complex networks, this crosstalk is exacerbated. The analysis presented in [48] and [49] illustrates how the power penalty in microring-based silicon photonic networks grows with the network complexity (i.e. number of cascaded switches) and the number of wavelength channels used.

D. Power Consumption

The power consumption associated with active silicon photonic devices is a central consideration of the design. Microring modulators have been demonstrated to operate with

sub-picojoule per bit energy consumption [63]. Recent link analysis work has shown that the foremost contributors to the power consumption are the laser sources, with relatively poor wall-plug efficiencies [64]. The total worst-case power penalty dictates the laser power required at the input of a network, and every wavelength channel must be powered this way. Therefore, the total required laser power scales with the complexity (i.e. number of devices along the path of light) and number of wavelength channels. The total input power to the system, however, is generally 20–100 times greater due to poor electrical-optical conversion during lasing and the need for cooling and stabilization mechanisms in the laser. Laser wall-plug efficiencies (i.e. laser optical output power divided by total laser power consumption, including cooling) range from 1–5% in the state-of-the-art devices today and are projected to reach only 10% in the near future [19]. Laser sources can be located on separate chips from the active optical components and surrounding electronics in order to alleviate chip package power constraints. However, the laser power dissipation remains part of the overall system power budget.

The resonance of a microring is susceptible to local thermal fluctuations that inevitably arise in computing systems. In addition, fabrication variations lead to devices that are not centered directly on the desired wavelength. As a result, mechanisms are required to trim (adjust for fabrication imperfections) and tune (adjust for dynamic fluctuations) silicon photonic devices to maintain operational links, predominately implemented with electronic heaters [54]. Various approaches have been used experimentally to control these heaters and achieve wavelength stabilization [51], [52], [65]–[67]. Athermal devices have also been explored [53], [68]–[70], but there are still many outstanding challenges, such as integration of non-CMOS compatible materials and high losses in these devices. A thorough review of thermal stabilization issues in resonator devices can be found in [54].

These "static" power figures (i.e. lasers and thermal tuning mechanisms) make up the majority of the total power consumption in silicon photonic links and networks [62]. Turning off lasers or heaters while they are not in use may not always be feasible because of relatively long stabilization times [71]. New techniques that allow these components' power to be dynamically scaled without incurring long re-stabilization times would help considerably in reducing the power overhead that comes with optical networking, especially in cases where the optical links are utilized sparingly.

Optical networks should be dimensioned carefully according to the overall system characteristics, including minimization of static power waste and power consumption while fulfilling parallel applications' bandwidth requirements. The dimensions of this design space are numerous. At the link level, they include the number of wavelength channels per link—particularly emphasized in this section—and the modulation rate and type. *Space-division multiplexing (SDM)* [103], i.e. the use of several parallel and independent optical waveguides, is also an option that can be exploited in complement to WDM. By spreading WDM channels over multiple waveguides, channel spacing can be enlarged. This in turn facilitates device tuning, decreases the power penalties and generally reduces the power consumption. However, as explained hereafter, spatial parallelism hardens the problem of building an interconnection topology. Beyond the choice of the link constituents, the manner in which multiple links are interconnected to form a topology is another crucial aspect of the dimensioning process. The following section discusses this design space and the approach toward developing a silicon photonic interconnection network architecture.

V. Toward Full Scale Silicon Photonics Interconnection Networks

In this section, we discuss the methodology and highlight the challenges of realizing designs for full scale optical interconnects.

A. Interconnection Topologies for Exascale

Exascale interconnection networks will be composed of an interconnected set of lower port count switches or routers, as it is unlikely that monolithic switches that are scalable to thousands of ports will emerge. Interconnection network topologies have been extensively investigated in the last decades. Topologies are generally categorized as *direct* or *indirect* [14]. In *direct* topologies (e.g. in Sequoia, as briefly mentioned in Section II), compute nodes are equally distributed among the switches. These switches are in turn interconnected among themselves. As switches are often too numerous to allow a one-to-one (i.e. fully-meshed) interconnection, only given pairs are connected. Different schemes have been proposed to select these pairs, for example, n-dimensional torus [72], Flattened Butterfly [73] or Dragonfly [74]. In *indirect* topologies, switches can be divided in two groups: access, and internal. Compute nodes are connected to the former exclusively, while the internal switches ensure a global connectivity among the access switches. *Indirect* topologies are also often referred as fat-tree based, as the internal switches can be seen as forming a tree-like hierarchical structure [75]. Both *direct* and *indirect* topologies can be found in recent supercomputers. Tianhe-2 leverages a two level fat-tree, while TITAN and Sequoia both implement direct interconnects (3D- and 5D-torus, respectively).

The size of an interconnection network, i.e. principally the number of switches and links it regroups, is mainly determined by the number of ports offered by its switches, i.e. their *radixes* [76]. This is the case for both *direct* and *indirect* topologies. Fig. 7 shows the number of switches required to interconnect 100 000 compute nodes, as a function of the number of ports offered. We take the 100 000 from the analysis made in Section II and suppose here that each compute node is either connected with a *single WDM* link (100 000 endpoints) or with *two spatially parallel WDM* links (200 000 endpoints). As shown, if more ports per switches are available, a lower number of switches are required. However, one notes that the decrease is super-linear. By reducing the number of switches, the number of message stopovers is also reduced, impacting the total network traffic. This translates *in fine* into further reduced requirements in terms of *ports per compute node,* as exhibited in Fig. 8, which shows the impact of radix on the total number of ports. High

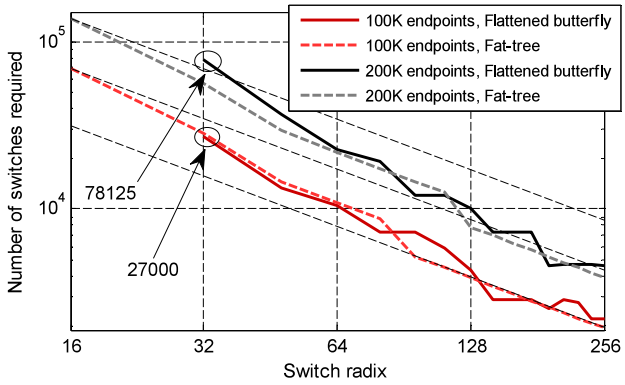


Fig. 7. Number of switches required to interconnect 100 000 or 200 000 endpoints, for different switch radixes, and different topologies. Larger radixes allow a super-linear decrease of the number of required switches (dotted diagonal lines represent linear decrease). In return, doubling the number of endpoints to support spatial parallelism can imply an almost threefold increased number of switches.

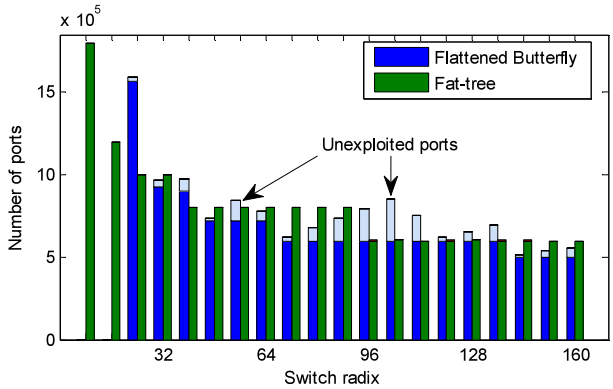


Fig. 8. Total number of ports in an interconnect regrouping 100 000 nodes, for different switch radixes, and different topologies. In the Flattened Butterfly case, some radixes result in unexploited ports as nodes are evenly distributed among *all* switches. In Fat-tree case, nodes are distributed among *entry* switches only which limits the number of unexploited ports.

radix switches are thus highly beneficial. In contrast, *Space Division Multiplexing (SDM)* should be conservatively exploited, as apparent in Fig. 7. If each computer node must be connected with *two* WDM links instead of one, requirements in terms of number of switches (at constant radix) are more than doubled.

Demonstrated radixes in silicon photonics switches are thus far relatively small, mainly limited by the optical loss budget [49]. As shown in Fig. 9, scaling the photonic switches to larger radixes and toward supporting Tb/s links will largely depend on the available power budget. With projected device improvements in terms of reduced insertion losses and enhanced receiver sensitivities, it can be expected that silicon photonic switches will scale to support 128 or even 256 ports. To support the required connectivity among the thousands of nodes in Exascale systems, the photonic interconnection network topology will be of *diameter* 2 or greater[3]. This implies that communication

[3]The *diameter* of a topology denotes the number of hops separating the most distant vertices pair, i.e. the worst case distance between two vertices.

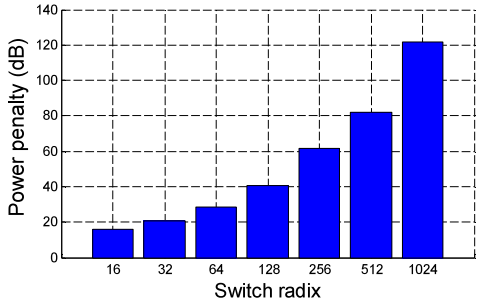


Fig. 9. Power penalty of a ring resonator based optical switch supporting 1 Tb/s per port constructed with DWDM of 100 wavelength channels each at 10Gb/s [49].

messages exchanged by many pairs of compute nodes will have to traverse at least three switches before reaching their final destination (the ingress and egress switches, plus at least one internal "transit switch").

B. Photonic Interconnection Network Design

In such multihop networking scenarios where optical signals traverse multiple photonic switches and links, the following additional design considerations arise:

- **Optical amplification:** amplification will most likely be required to extend the reach of optical signals and to support propagation over the multiple switch stages.
- **Optical paths reconfiguration times:** The interconnection network reconfiguration times will critically impact the latency and throughput performance, and should therefore be minimized.
- **Bufferless arbitration:** the circuit switched nature of photonics requires new designs for the flow control and arbitration.

Each of these key design consideration is addressed in further detail in the following sub-sections.

C. Optical Amplification

Considering the physical size, cost, gain bandwidth and latency requirements, integrated optical amplifier technologies such as semiconductor optical amplifiers (SOAs) are potential candidates for the system interconnection network. SOAs based on the quantum-dot structure have been shown to have gains $>$25 dB, noise figures $<$5 dB, and able to deliver output optical powers $>$20dBm over a 90-nm bandwidth [77]. The advantages of small physical size, wide gain bandwidth and potentially low-cost integration make SOAs attractive for the desired WDM operation. SOAs can further be integrated in the CMOS chip platform via hybrid integration to realize the needed laser sources and have already gained commercial attention for applications in data centers and high performance computing [78].

In the design of system-wide optical interconnection networks that leverage SOAs there are several additional considerations. The non-linear effects and noise power of SOAs will need to be taken into account in the physical link modeling. In addition, SOA transients, especially if the required gain varies

from message to message can adversely impact the dynamic capabilities of the network. Finally, the SOAs will need to be designed to operate at low powers within the system budget.

D. Optical Path Reconfiguration Time

As relevant to the network performance, we define the *optical path Reconfiguration Time* (RT) as the time required to *both* change the spatial switching configuration, *and* re-establish an error-free optical communication link.

In general, long ($>100\mu$s) RTs will prevent the optical network from efficiently carrying traffic with small to medium sized and/or latency sensitive messages. Architectures with long optical RTs have been be used as over-flow networks, devoted to long and non-latency sensitive data-transfer [79], or as ancillary networks, that provide a very specific connectivity (typically matching application requirements) for sufficiently long durations [80], [81].

By driving down the RT the optically switched architecture can accommodate a larger portion of the traffic. The ratio s/RT, where s stands for the average message serialization time, can be used as an indicator of the range of packet sizes that can be supported by the optical network. If s/RT is equal to one, switches are transmitting half of the time, and reconfiguring the other half. This results in a maximal link utilization of 50%. Larger s/RT values (i.e. longer messages or faster switches) improve this utilization. In contrast, the value of the optical network may be questionable for small values of this metric [10]. Assuming a line rate of 1 Tb/s, and targeting $s/RT \geq 1$, the average message size should be $\geq$ 125MB with 1ms switching time, $\geq$ 125KB with 1μs, and $\geq$ 125B with 1ns. In practice, the average message size will range between 125 B and 125KB. Therefore, realizing nanosecond scale reconfiguration times is an important goal for silicon photonic networks to deliver high link utilization while supporting small packets. With microsecond scale RTs, an ancillary electrical network will likely be required to transmit the smallest messages.

With current state-of-the-art thermal stabilization mechanisms, link configuration times of $\sim$100μs have been measured with ring resonators based switches [81], and improved designs, in particular based on athermal devices, are expected to achieve sub-μs times.

E. Bufferless Arbitration

From a utilization standpoint, shorter reconfiguration times allow for an increasing portion of the traffic to be optically supported. However, faster reconfiguration times, allowing smaller messages, also imply an increased number of messages flowing through the switches that require arbitration.

In a networking context, any element (switch, link, etc.) potentially traversed by different traffic flows must be *arbitrated* to prevent these flows from mixing. Arbitration leads to contention when a network element has been assigned to a flow while another flow requires it. In most networks, contentions can be easily solved by delaying the contending flow (i.e. by buffering its content) until the required element can be reassigned. Optical data flows, however, cannot be delayed in a flexible way. Contention must therefore be solved by dropping the contenting flow or by sending it to another output (deflection routing). Alternatively, contentions can be *prevented* by designing the network interfaces such that nodes do not inject contending flows simultaneously.

Arbitration approaches developed for bufferless optical networks will depend on the reconfiguration times. In a central arbitration scheme, a unique arbiter receives requests for circuits from the network clients, schedules the requests (accounting for reconfiguration times), and communicates its arbitration decisions on the optical path availability back to the clients. In this way, the clients know exactly when to send the data, while the arbiter knows when to reassign the network elements. This scheme is well-suited for architectures with relatively long reconfiguration times ($>100\mu$s) to allow for the central processing. Moreover, although circuit switched arbitration requires a round-trip time (RTT) to the arbiter, it is largely dominated by the reconfiguration time. In contrast, when the optical switches RTs are fast (sub- μs), it is advantageous to consider alternatives to central circuit switched arbitration, whose performance is limited by RTTs and the arbiter processing bandwidth.

One alternative approach consists of dividing the interconnection network into separate domains that are independently arbitrated [82]. This approach reduces the distances to the arbiter as well as its computational load. However, in this scenario, communications spanning over multiple domains can potentially lead to contention and electrical buffering is required at the domain boundaries.

Another approach consists of replacing the two-way reservation scheme of centrally arbitrated circuit switching (request - grant) with a packet oriented, one-way reservation scheme: the request for network resources is included with the packet, or precedes the packet payload by a short time interval. In both cases, the payload is sent *opportunistically*, i.e. without any insurance of resource availability. In the bufferless optical network, this inevitably leads to packet dropping and the need for retransmission [83]. Packet dropping can be mitigated (but not suppressed) by means of deflection. Opportunistic approaches provide good performance in presence of sporadic traffic flows which leads to rare contentions. Under highly loaded traffic conditions, in contrast, opportunistic approaches fail to achieve high throughput as frequent retransmissions exhaust the bandwidth.

Importantly, each dropped message represents wasted energy [84]. Under high loads, deflection is of little help as deflected messages do not leave the network. They are thus still subject to dropping and occupy elements normally devoted to other flows [85]. Packet contention also scales with the architecture size. Finally, in architectures where multiple optical switches are cascaded—with optical amplifiers interleaved—messages dropped close to the destination represent an even higher waste of energy. Opportunistic arbitration schemes must therefore be carefully designed to account for both the energy consumption and performance. Performances can vary substantially with different communication patterns and must therefore be tested, and their parameters optimized, in presence of application traffic.

F. Architectural Improvements

Contention within opportunistically arbitrated networks can be reduced by introducing time slot mechanisms [83] which diminish packet overlap. Flow control mechanisms, such as back-pressure, can be leveraged to dynamically adjust the traffic volume entering the network.

The Data Vortex architecture is one example of an optical interconnection network design that combined deflection, flow-control and amplification [86]. A silicon photonics based Data Vortex can be foreseen, with ring or MZI comb switches advantageously replacing the SOA based switch elements used in the prior implementation [87].

The capabilities silicon photonics based interconnection networks can deliver to Exascale systems performance are primarily dependent on the reconfiguration times associated with optical switching. As switching speeds are pushed toward the nanosecond levels, the scope and range of applications benefiting from the photonic data movement will broaden. For these future photonic-enabled systems, the overarching goal is to evaluate the performance impact on the executed Exascale applications, as discussed in the next section.

G. Application Aware Design of Exascale Optical Networks

Exascale systems will ultimately aim to execute relevant applications with optimized time-to-solution and/or energy-to-solution. Therefore, hardware design decisions, including *dimensioning* decisions (*B*/*F* ratios, bisectional bandwidths, link line rates, etc.), as well as operational design choices (such as the arbitration scheme consideration for latency or energy efficiency), must account for the applications' performance. Particularly for the photonic network, different design trade-offs are considered that emphasize performance and/or energy consumption.

In this subsection we provide an overview of the optical network design space considerations and their impact on applications execution. One can frame a lower bound on the application runtime as determined by the Flops it requires divided by the peak system computing capability. If an application requires 10 PetaFlops, it will require at least 10 s of execution time on PF machine. In practice, the execution will be longer and will depend on the computing efficiency (of $1/\eta$, c.f. Section II). A portion of this extra-time can be attributed to the parallel nature of the application structure: no application is perfectly parallelizable over all cores at all times. Another portion of this extra time, however, will be attributed to computing resources idly waiting for the required computation operands, due to limitations in memory access, network latency, or other system limitations. In other words, network latencies resulting from limited bandwidths, arbitration delays, the number of hops in the topology, etc. directly impact the application execution time performance.

The network may be independently designed to maximize bandwidth, minimize arbitration latency, to support small hop count compact topology, or to target a tradeoff among these metrics. Each possible design, however, uniquely impacts the application performance and/or the energy consumption. For example maximizing bandwidth is typically at the expense of increased power consumption. Moreover, minimizing the *average* arbitration latency can be made at the expense of higher *peak* latency. This is typically what happens within single-way arbitration schemes: even if the majority of the packets experience no drop, and thus benefit from the single-way reservation, the minority that is retransmitted is heavily impacted (long latency tails), which can have detrimental effects at the application level [88]. Navigating the design space toward developing a *balanced* system is therefore a complex process.

An empirical notion of a balanced system can be obtained by examining today's supercomputer systems. In terms of interconnects specifically, Table III shows that node injection bandwidths should be maintained above 0.001 *B/F*, while bisectional bandwidths can be one order of magnitude lower. Although these values can be used as guidelines for the initial optical network design, evaluating the needs of applications on Exascale systems with photonic-enabled data movement is extremely challenging.

One approach involved analyzing the detailed communication activity of applications and mapping them onto the interconnect design [80]. However, this approach is limited by the fact that application traffic itself depends on the interconnect capabilities. Indeed, delayed messages might slow down the application proceeding, which in turn will decrease the message injection rate. Therefore, the interconnect *and* the application must be jointly modeled. Since application behaviors cannot typically be described analytically, simulation based modeling must be employed. In the past years, large research efforts have been devoted to developing extensive simulation platform for joint (i.e. hardware and software) modeling of Exascale systems [89], [90]. As photonic interconnect systems are developed for Exascale, it becomes important to incorporate their models in these large scale simulation platforms [91] to enable hardware/software co-design and validation [92], [93].

VI. Experimental Hardware/Software Validation of Silicon Photonic Interconnected Systems

Physical layer demonstrations of individual devices, links or networks, and architectural modeling of potential photonic interconnection networks are crucial stepping stones toward the realization of silicon photonics in next generation Exascale systems. It is clear however that an integrated source of hardware-software validation is required for moving forward with implementing silicon photonics in next-generation computing systems. There is a need to fill the gap between early stage development approaches and the ultimate goal of mass fabrication and insertion into commercial systems. An intermediate validation step, which proves the feasibility and on-the-field utilization of systems aims to address this gap.

In this section we describe an *integrated hardware-software validation platform* with dynamic capabilities that can provide a systems-level perspective of the feasibility of silicon photonics for computing systems.

A. Dynamic Functionality Enabled by FPGAs

Field programmable gate arrays (FPGAs) provide dynamic reconfiguration of electronic logic gates—i.e., logic gates that

can be configured in a variety of ways to create state-based logic, registers, memory, etc. This "logic substrate" can be used to emulate computing components, control systems, and network functionalities that physically implement critical components of an Exascale computing system. Additional programmable hardware—a local processor architecture and associated memory capable of encapsulating and executing a complete software stack—enable software programmability of the custom hardware, ranging from control of individual registers to full-scale multi-processor operating systems [94].

In addition to reconfigurable logic gates, the FPGA chips within the test platform embed special-purpose high-speed data transceivers (also known as serial-deserializers, or SerDes). The essential functionality of these transceivers is to convert parallel data—inherent to computing elements—to serial I/O that is compatible with silicon-photonic transmit and receive devices. The physical serialization realized by these embedded transceivers cannot be modified (it is often constrained to low-voltage differential signaling); however, higher-layer protocols and data formatting are easily controlled and configured according to custom logic.

Silicon photonic devices also require a certain degree of control and maintenance [54], according to the aforementioned thermal characteristics in Section III, which can be implemented with state-based logic [67]. This logic drives additional analog-to-digital (AD) and digital-to-analog (DA) hardware—coupled with FPGAs through a daughtercard interface such as FPGA mezzanine card or high-speed mezzanine card—that provides complete sampling and multi-level control of silicon photonic devices. The coupling of data transmission and control provides an implementation of a complete silicon photonic link layer protocol, thereby validating the capabilities of silicon photonics as a data transport medium.

FPGAs can also be leveraged to control silicon photonic devices arranged in a network—i.e., to implement the logic to arbitrate fundamental networking functions, such as switching and routing, and has been successfully demonstrated [95], [96]. Controlling networked silicon photonic devices using FPGAs provides not only necessary device stabilization and control, but an aggregated method for actuating computing/networking primitives in a software-programmable fashion.

B. Demonstration Through Testing and Validation

The FPGA-based hardware-software integration platform provides a test vehicle for 1) discovering implementation-dependent effects that cannot be found with initial simulation/modeling, and 2) demonstrating the functionality of silicon photonic devices in an interconnection network and implementing valid execution models for system runtime 3) paving the way toward all-integrated systems (devices and control logic, all inclusive).

Implementation-dependent effects are any physical, architectural, or programming peculiarities that are not initially part of device- and architectural-level simulation/modeling. One such example is pattern-dependent effects of data transmission on an injection-mode microring modulator, which exacerbates thermal effects and results in significant power penalty [97]. Forward current effects of an integrated *p-i-n* diode (used for modulation) might not be initially transparent; the aforementioned study highlights reverse-biased depletion mode device operation, which is insensitive to pattern-dependent effects. On the contrary, error-prone systems—such as memory transactions in optically-connected memory—adhere to well-known error-correction techniques [98], demonstrating the feasibility of silicon photonic enabled large-scale systems. Furthermore, validation of advanced modulation techniques using particular silicon photonic implementation effects shows further functionality and efficiency for optics in high performance and potentially Exascale systems [99].

Software-dependent communications refer to execution of physical layer data flows that are controlled by a higher-level arbitration layer. A complex silicon photonic network could consist of various wavelength division multiplexed data paths, originating from/going to emulated CPU cores or physical memory banks, and silicon photonic circuit-switched and wavelength-selective components arbitrated by high-speed state-based logic on FPGAs. Previous sections of this document contemplate the execution and delivery of such flows in tandem with arbitration of complex network architectures, but without actual implementation of such a system, its runtime complexities cannot be entirely characterized. The interaction between physical layer startup sequencing, arbitration of silicon photonic devices, and data delivery must be orchestrated with an overall control interface to ensure successful end-to-end operation. One such method that has been successfully shown using FPGAs is a message-passing interface (MPI) used for direct memory access [100]. Considering the inclusion of complex computing components necessary for Exascale systems—such as the Hybrid Memory Cube [13] and multicore processing architectures—further exacerbates the need for complete system implementation.

A hardware-software integration platform as described here provides a mechanism for demonstrating data that is both driven by system software and compatible with optical interconnection network protocols. The relationship between data programmability and protocol compatibility is not necessarily clear during the design process; aforementioned implementation-dependent and runtime execution effects could hinder system functionality. Integrating the physical layer devices and link layer control mechanisms provides critical feedback on device start-up sequencing, control loop timing, design of silicon photonic transmission protocols, and software execution models for system runtime [67], [101], [102]. The criticality of this type of design and validation is further accentuated as hardware becomes more complicated in an implemented design, due to scalability of aggregated control and software models.

VII. Conclusion

Data movement is perhaps the dominating challenge on the way to realizing Exascale systems. Breakthrough technological advances are most likely required to address these challenges in a scalable manner. With its CMOS-compatible fabrication

and compact integration within the computing/memory chips, silicon photonics has emerged as one of the promising communications technologies for Exascale data movement. This paper provided an overview of the different research efforts conducted to allow further scaling of supercomputer capabilities by means of ultra-high bandwidth, energy efficient silicon photonics interconnection networks.

The design of an Exascale grade interconnection system is a multi-faceted problem which requires an integrated effort that cross-cuts several system planes. Nanophotonic devices able to perform all the operations required for transmission and switching have been demonstrated over the past decade, but these designs have not been optimized for system level considerations. The specific system implementation in turn impacts the device characteristics and performance. Therefore, a holistic approach achieving individual device parameters optimization jointly with the architecture and networking mechanisms design must be targeted. Models of photonic networks must also be incorporated in simulation platforms to allow co-design of the executed applications and of the interconnection systems. Finally, early hardware/software joint testing and validation of new devices, systems and architectures is crucial to reach the goal of overall design of next-generation Exascale computing platforms.

REFERENCES

[1] (2014, Oct.). [Online]. Available: www.top500.org

[2] P. M. Kogge and D. R. Resnick, "Yearly update: Exascale projections for 2013," Sandia Report SAND, pp. 2013–9229, Oct. 2013.

[3] J. Shalf, S. Dosanjh, and J. Morrison, "Exascale computing technology challenges," *VECPAR*, vol. 6449, pp. 1–25, 2011.

[4] V. Getov, A. Hoisie, and H. J. Wasserman, "Codesign for systems and applications: Charting the path to exascale computing," *Computer*, vol. 44, no. 11, pp. 19–21, Nov. 2011.

[5] F. Cappello, A. Geist, B. Gropp, L. Kale, B. Kramer, and M. Snir, "Toward exascale resilience," *Int. J. High Perform. Comput. Appl.*, vol. 23, no. 4, pp. 374–388, Nov. 2009.

[6] M. Snir, R. W. Wisniewski, J. A. Abraham, S. V. Adve, S. Bagchi, P. Balaji, J. Belak, P. Bose, F. Cappello, B. Carlson, A. A. Chien, P. Coteus, N. A Debardeleben, P. Diniz, C. Engelmann, M. Erez, S. Fazzari, A. Geist, R. Gupta, F. Johnson, S. Krishnamoorthy, S. Leyffer, D. Liberty, S. Mitra, T. Munson, R. Schreiber, J. Stearley, and E. V. Hensbergen, "Addressing failures in exascale computing," *Int. J. High Performance Comput. Appl.*, vol. 28, no. 2, pp. 129–173, 2014.

[7] A. Geist and R. Lucas, Major computer science challenges at exascale, *Int. J. High Performance Comput. Appl.*, vol. 23, no. 4, pp. 427–436, 2009.

[8] J. Dongarra, P. Beckman,T. Moore, P. Aerts, G. Aloisio, J.-C. Andre, D. Barkai, J.-Y. Berthou, T. Boku, B. Braunschweig, F. Cappello, B. Chapman, X. Chi, A. Choudhary, S. Dosanjh, T. Dunning, S. Fiore, A. Geist, B. Gropp, R. Harrison, M. Hereld, M. Heroux, A. Hoisie, K. Hotta, Y. Ishikawa, Z. Jin, F. Johnson, S. Kale, R. Kenway, D. Keyes, B. Kramer, J. Labarta, A. Lichnewsky, T. Lippert, B. Lucas, B. Maccabe, S. Matsuoka, P. Messina, P. Michielse, B. Mohr, M. Mueller, W. Nagel, H. Nakashima, M. E. Papka, D. Reed, M. Sato, E. Seidel, J. Shalf, D. Skinner, M. Snir, T. Sterling, R. Stevens, F. Streitz, B. Sugar, S. Sumimoto, W. Tang, J. Taylor, R. Thakur, A. Trefethen, M. Valero, A. van der Steen, J. Vetter, P. Williams, R. Wisniewski, and K. Yelick, "The international exascale software project roadmap," *Int. J. High Perform. Comput. Appl.*, 2011.

[9] J. D. Balfour. "Efficient embedded computing," PhD thesis, Stanford University, Stanford, CA, USA, 2010.

[10] S. Borkar, "Role of interconnects in the future of computing," *J. Lightw. Technol.*, vol. 31, no. 24, pp. 3927–3933, Dec. 2013.

[11] R. Lucas, J. Ang, K. Bergman, S. Borkar, W. Carlson, L. Carrington, G. Chiu, R. Colwell, W. Dally, J. Dongarra, A. Geist, G. Grider, R. Haring, J. Hittinger, A. Hoisie, D. Klein, P. Kogge, R. Lethin, V. Sarkar, R. Schreiber, J. Shalf, T. Sterling, and R. Stevens, "Top ten exascale research challenges," DOE ASCAC Subcommitte Report, Feb. 2014.

[12] T. Pinguet, B. Analui, E. Balmater, D. Guckenberger, M. Harrison, R. Koumans, D. Kucharski, Y. Liang, G. Masini, A. Mekis, S. Mirsaidi, A. Narasimha, M. Peterson, D. Rines, V. Sadagopan, S. Sahni, T. J. Sleboda, D. Song, Y. Wang, B. Welch, J. Witzens, J. Yao, S. Abdalla, S. Gloeckner, P. De Dobbelaere, and G. Capellini, "Monolithically integrated high-speed CMOS photonic transceivers," in *Proc. 5th IEEE Int. Conf Group IV Photon.*, Sep. 2008, pp. 362–364.

[13] J. Jeddeloh and B. Keeth, "Hybrid memory cube new DRAM architecture increases density and performance," in *Proc. Symp. VLSI Technol.*, Jun. 2012, pp. 87–88.

[14] A. Agarwal, "Limits on interconnection network performance," *IEEE Trans. Parallel Distrib. Syst.*, vol. 2, no. 4, pp. 398–412, Oct. 1991.

[15] Y. Vlasov and S. McNab, "Losses in single-mode silicon-on-insulator strip waveguides and bends," *Opt. Exp.*, vol. 12, pp. 1622–1631, 2004.

[16] L. W. Luo, N. Ophir, C. P. Chen, L. H. Gabrieli, C. B. Poitras, K. Bergman, and M. Lipson, "WDM-compatible mode-division multiplexing on a silicon chip," *Nature Commun.*, vol. 5, p. 3069, Jan. 2014.

[17] G. Li, J. Yao, H. Thacker, A. Mekis, X. Zheng, I. Shubin, Y. Luo, J.-H. Lee, K. Raj, J. E. Cunningham, and A. V. Krishnamoorthy, "Ultralow-loss, high-density SOI optical waveguide routing for macrochip interconnects," *Opt. Exp.*, vol. 20, no. 11, pp. 12035–12039, 2012.

[18] J. Cardenas, C. B. Poitras, J. T. Robinson, K. Preston, L. Chen, and M. Lipson, "Low loss etchless silicon photonic waveguides," *Opt. Exp.*, vol. 17, no. 6, pp. 1–2, 2009.

[19] X. Zheng, S. Lin, Y. Luo, J. Yao, G. Li, S. S. Djordjevic, J.-H. Lee, H. D. Thacker, I. Shubin, K. Raj, J. E Cunningham, and A. V. Krishnamoorthy, "Efficient WDM laser sources towards terabyte/s silicon photonic interconnects," *J. Lightw. Technol.*, vol. 31, no. 15, pp. 4142–4154, Dec. 2013.

[20] B. Koch, E. J. Norberg, B. Kim, J. Hutchinson, J.-H. Shin, G. Fish, and A. Fang, "Integrated silicon photonic laser sources for telecom and datacom," presented at the Opt. Fiber Commun., Washington, D.C, USA, 2013.

[21] M. Pu, L. Liu, H. Ou, K. Yvind, and J. Hvam, "Ultra-low-loss inverted taper coupler for silicon-on-insulator ridge waveguide," *Opt. Commun.*, vol. 283, no. 19, pp. 3678–3682, Oct. 2010.

[22] Y. Zhang, S. Yang, A. E.-J. Lim, G.-Q. Lo, C. Galland, T. Baehr-Jones, and M. Hochberg, "A compact and low loss Y-junction for submicron silicon waveguide," *Optics Exp.*, vol. 21, no. 1, pp. 1310–1316, 2013.

[23] D. Kwong, Y. Zhang, A. Hosseini, Y. Liu, and R. T. Chen, "Demonstration of Rib waveguide based 1×12 multimode interference optical beam splitter on silicon-on-insulator," in *Proc. Photon. Soc. Summer Top. Meeting Series*, July 19–21, 2010, pp. 221–222.

[24] Y. Ma, Y. Zhang, S. Yang, A. Novack, R. Ding, A. E.-J. Lim, G.-Q. Lo, T. Baehr-Jones, and M. Hochberg, "Ultralow loss single layer submicron silicon waveguide crossing for SOI optical interconnect," *Opt. Exp.*, vol. 21, no. 24, pp. 29374–29382, Nov. 2013.

[25] R. H. Derksen, G. Lehmann, C.-J. Weiske, C. Schubert, R. Ludwig, S. Ferber, C. Schmidt-Langhorst, M. Moller, and J. Lutz, "Integrated 100 Gbit/s ETDM receiver in a transmission experiment over 480 km DMF," presented at the Opt. Fiber Commun. Opt. Soc. Am., Washington, DC, USA, 2006, PDP37.

[26] J. H. Sinsky, A. Adamiecki, L. Buhl, G. Raybon, P. Winzer, O. Wohlgemuth, M. Duelk, C. R. Doerr, A. Umbach, H. G. Bach, and D. Schmidt, "A 107-Gbit/s optoelectronic receiver utilizing hybrid integration of a photodetector and electronic demultiplexer," *J. Lightw. Technol.*, vol. 26, no. 1, pp. 114–120, Jan. 2008.

[27] D. Kucharski, D. Guckenberger, G. Masini, S. Abdalla, J. Witzens, and S. Sahni,, "10-Gb/s 15-mW optical receiver with integrated germanium photo-detector and hybrid inductor peaking in 0.13-μm SOI CMOS technology," *Proc. IEEE Int. Solid-State Circuits Conf.*, Feb. 7–11, 2010, pp. 360–361.

[28] B. G. Lee, A. V. Rylyakov, J. E. Proesel, C. W. Baks, R. Rimolo-Donadio, C. L. Schow, A. Ramaswamy, J. E. Roth, M. Jacob-Mitos, and G. Fish, "60-Gb/s receiver employing heterogeneously integrated silicon waveguide coupled photodetector," in *Proc. Conf. Lasers Electro. Opt.*, CLEO, San Jose, CA, USA, Jun. 9–14, 2013.

[29] F. Xia, T. Mueller, Y.-m. Lin, A. Valdes-Garcia, and P. Avouris, "Ultrafast graphene photoodetector," *Nature Nanotechnol.*, vol. 4, pp. 839–842, 2009.

[30] C. Liu, Y.-C. Chang, T- B. Norris, and Z. Zhong, "Graphene photodetectors with ultra-broadband and high responsivity at room temperature," *Nature Nanotechnol.*, vol. 9, pp. 273–278, 2014.

[31] S. Xiao, H. Shen, M. H. Khan, and M. Qi, "Silicon microring filters," in *Proc. Conf. Lasers Electro-Opt./Quantum Electron. Laser Sci. Photonic Appl. Syst. Technol., OSA Tech. Dig. Opt. Soc. Am.*, 2008, paper JWA84.

[32] S. T. S. Cheung, B. Guan, S. S. Djordjevic, K. Okamoto, and S. J. B. Yoo, "Low-loss and high contrast silicon-on-insulator (SOI) arrayed waveguide grating," presented at the Conf. Lasers Electro-Opt., San Jose, CA, USA, 2012.

[33] R. Ding, Y. Liu, Y. Ma, Y. Yang, Q. Li, A. E. Lim, G.-Q. Lo, K. Bergman, T. Baehr-Jones, and M. Hochberg, "High-speed silicon modulator with slow-wave electrodes and fully independent differential drive," *IEEE/OSA J. Lightw. Technol.*, vol. 32, no. 12, pp. 2240–2247, Jun. 2014.

[34] Q. Xu, B. Schmidt, S. Pradhan, and M. Lipson, Micrometre-scale silicon electro-optic modulator, *Nature*, vol. 435, pp. 325–327, May 19, 2005.

[35] X. Xiao, H. Xu, X. Li, Z. Li, T. Chu, J. Yu, and Y. Yu, "60 Gbit/s silicon modulators with enhanced electro-optical efficiency," presented at the Opt. Fiber Commun., Anaheim, CA, USA, Mar. 2013.

[36] M. R. Watts, D. C. Trotter, R. W. Young, and A. L. Lentine, "Ultralow power silicon microdisk modulators and switches," in *Proc. 5th IEEE Int. Conf. Group IV Photon.*, pp. 4–6, Sep. 2008.

[37] P. Dong, S. Liao, D. Feng, H. Liang, D. Zheng, R. Shafiiha, C.-C. Kung, W. Qian, G. Li, X. Zheng, A. V. Krishnamoorthy, and M. Asghari, "Low V_{pp} ultralow-energy, compact, high-speed silicon electro-optic modulator," *Opt. Exp.*, vol. 17, no. 25, pp. 1–2, Nov. 2009.

[38] A. Krishnamoorthy, X. Zheng, D. Feng, J. Lexau, J. F. Buckwalter, H. D. Thacker, F. Liu, Y. Luo, E. Chang, P. Amberg, I. Shubin, S. S. Djordjevic, J. H. Lee, S. Lin, H. Liang, A. Abed, R. Shafiiha, K. Raj, R. Ho, M. Asghari, and J. E. Cunningham, "A low-power, high-speed, 9-channel germanium-silicon electro-absorption modulator array integrated with digital CMOS driver and wavelength multiplexer," *Opt. Exp.*, vol. 22, no. 10, pp. 12289–12295, 2014.

[39] W. Bogaerts, P. De Heyn, T. Van Vaerenbergh, K. De Vos, S. Kumar Selvaraja, T. Claes, P. Dumon, P. Bienstman, D. Van Thourhout, and R. Baets, "Silicon microring resonators," *Laser Photon. Rev.*, vol. 6, no. 1, pp. 47–73, 2012.

[40] G. Li, A. V. Krishnamoorthy, I. Shubin, J. Yao, Y. Luo, H. D. Thacker, X. Zheng, K. Raj, and J. E. Cunningham, "Ring resonator modulators in silicon for interchip photonic links," *IEEE J. Sel. Topics Quantum Electron.*, vol. 19, no. 6, art. no. 3401819, Nov./Dec. 2013.

[41] Y. Liu, R. Ding, Y. Ma, Y. Yang, Z. Xuan, Q. Li, A. E.-J. Lim, G.-Q. Lo, K. Bergman, T. Baehr-Jones, and M. Hochberg, "Silicon Mod-MUX-ring transmitter with 4 channels at 40 Gb/s," *Opt. Exp.*, vol. 22, no. 13, pp. 16431–16438, 2014.

[42] R. Ding, Y. Liu, Q. Li, Z. Xuan, Y. Ma, Y. Yang, A. Eu-Jin Lim, G.-Q. Lo. K. Bergman, T. Baehr-Jones, and M. Hochberg, "A compact low-power 320-Gb/s WDM transmitter based on silicon microrings," *IEEE Photon. J.*, vol. 6, no. 3, Article 6600608, Jun. 2014.

[43] L. Chen, E. Hall, L. Theogarajan, and J. Bowers, "Photonic switching for data center applications," *IEEE Photon. J.*, vol. 3, no. 5, pp. 834–844, Oct. 2011.

[44] Y. Vlassov, W. Green, and F. Xia, "High-throughput silicon nanophotonic wavelength-insensitive switch for on-chip optical networks," *Nature Photon.*, vol. 2, pp. 242–246, 2008.

[45] X. Zhu, Q. Li, J. Chan, and A. Ahsan, "4×44 Gb/s packet-level switching in a second-order microring switch," *IEEE Photon. Technol. Lett.*, vol. 24, no. 17, pp. 1555–1557, Aug. 2012.

[46] N. Sherwood-Droz, H. Wang, L. Chen, B. G. Lee, A. Biberman, K. Bergman, and M. Lipson, "Optical 4×4 hitless silicon router for optical networks-on-chip (NoC)," *Opt. Exp.*, vol. 16, no. 20, pp. 15915–15922, 2008.

[47] K. Suzuki, K. Tanizawa, T. Matsukawa, G. Cong, S.-H. Kim, S. Suda, M. Ohno, T. Chiba, H. Tadokoro, M. Yanagihara, Y. Igarashi, M. Masahara, S. Namiki, and H. Kawashima, "Ultra-compact 8×8 strictly-non-blocking Si-wire PILOSS switch," *Opt. Exp.*, vol. 22, no. 4, pp. 3887–3894, 2014.

[48] D. Nikolova and K. Bergman, "Analysis of silicon photonic microring-based multistage switches," *Adv. Photon. Commun.*, vol. 3, Photonics in Switching, San Diego, San Diego, CA, USA, pp. 13–17, Jul. 2014.

[49] D. Nikolova, R. Hendry, S. Rumley, and K. Bergman, "Scalability of silicon photonic microring based switch," presented at the Int. Conf. Transparent Opt. Netw., Graz, Austria, Jul. 2014.

[50] Q. Cheng, A. Wonfor, Richard V. Penty, and I. H. White, "Scalable, low-energy hybrid photonic space switch," *J. Lightw. Technol.*, vol. 31, no. 18, pp. 3077–3084, Sep. 2013.

[51] C. Qiu, J. Shu, Z. Li, X. Zhang, and Q. Xu, "Wavelength tracking with thermally controlled silicon resonators," *Opt. Exp.*, vol. 19, no. 6, pp. 5143–5148, Mar. 2011.

[52] K. Padmaraju, D. Logan, X. Zhu, J. J. Ackert, A. P. Knights, and K. Bergman, "Integrated thermal stablization of a microring modulator," *Opt. Exp.*, vol. 20, no. 27, pp. 14342–14350, 2012.

[53] B. Guha, K. Preston, and M. Lipson, "Athermal silicon microring electro-optic modulator," *Opt. Lett.*, vol. 37, no. 12, pp. 2253–2255, Jun. 2012.

[54] K. Padmaraju and K. Bergman, "Resolving the thermal challenges for silicon microring resonator devices," *Nanophotonics*, vol. 2, no. 4, pp. 1–14, Sep. 2013.

[55] J. Chan, G. Hendry, K. Bergman, and L. Carloni, "Physical-layer modeling and system-level design of chip-scale photonic interconnection networks," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 30, no. 10, pp. 1507–1520, Oct. 2011.

[56] C. Batten, A. Joshi, V. Stojanović, and K. Asanović. "Designing Chip-Level Nanophotonic Interconnection Networks." in Gabriela Nicolescu and Ian O'Connor Eds., *Integrated Optical Interconnect Architectures and Applications in Embedded Systems*. New York, NY, USA: Springer, 2013.

[57] N. Ophir and K. Bergman, "Analysis of high-bandwidth low-power microring links for off-chip interconnects," in *Proc. SPIE. Photonics West,* Feb. , 2013, vol. 8628.

[58] A. Yariv and P. Yeh, *Photonics: Optical Electronics in Modern Communications*. Oxford, U.K.: Oxford University Press 2007.

[59] N. Ophir, *Silicon Photonics for All-Optical Processing and High Bandwidth-Density Interconnects.* Ph.D. Thesis. Columbia Univ., New York, NY, USA.

[60] K. Padmaraju, X. Zhu, L. Chen, M. Lipson, and K. Bergman, "Intermodulation crosstalk characteristics of WDM silicon microring modulators," *IEEE Photon. Technol. Lett.*, vol. 26, no. 14, pp. 1478—1481, Jul. 2014.

[61] R. Hendry, D. Nikolova, S. Rumley, N. Ophir, and K. Bergman, "Physical layer analysis and modeling of silicon photonic WDM bus architectures," in *Proc. HiPEAC Workshop*, HiPEAC 2014, Vienna Austria, Jan. 20–22, 2014.

[62] R. Hendry, D. Nikolova, S. Rumley, and K. Bergman, "Modeling and evaluation of chip-to-chip scale silicon photonic networks," in *Proc. IEEE Symp. High Perform. Interconnects (Hot)*, 2014, Mountain View, Aug. 26–27, 2014.

[63] J. C. Rosenberg, W. M. J. Green, S. Assefa, D. M. Gill, T. Barwicz, M. Yang, S. M. Shank, and Y. A. Vlasov, "A 25 Gbps silicon microring modulator based on an interleaved junction," *Opt. Exp.*, vol. 20, no. 24, pp. 26411–26423, Nov. 2012.

[64] 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.

[65] H. Yu, M. Pantouvaki, S. Dwivedi, and P. Verheyen, "Compact thermally tunable silicon racetrack modulators based on an asymmetric waveguide," *IEEE Photon. Technol. Lett.*, vol. 25, no. 2, pp. 159–162, Jan. 2013.

[66] C. T. DeRose, M. R. Watts, D. C. Trotter, D. L. Luck, G. N. Nielson, and R. W. Young. "Silicon microring modulator with integrated heater and temperature sensor for thermal control," in *Proc Conf. Lasers Electro-Opt.*, 2010, pp. 1–2.

[67] K. Padmaraju, L.-W. Luo, X. Zhu, M. Glick, R. Dutt, M. Lipson, and K. Bergman, "Wavelength locking of a WDM silicon microring demultiplexer using dithering signals," in *Proc. Opt. Fiber Conf.*, vol. 4, Mar. 2014 , pp. 1–3.

[68] J. Teng, P. Dumon, W. Bogaerts, H. Zhang, X. Jian, X. Han, M. Zhao, G. Morthier, and R. Baets, "Athermal silicon-on-insulator ring resonators by overlaying a polymer cladding on narrowed waveguides," *Opt. Exp.*, vol. 17, no. 17, pp. 14627–14633, 2009.

[69] P. Alipour, E. S. Hosseini, A. A. Eftekhar, B. Momen, and A. Adibi, "Athermal performance in high-Q polymer-clad silicon microdisk resonators," *Opt. Lett.*, vol. 35, no. 20, pp. 3462–3464, 2010.

[70] Q. Li, S. Yegnanarayanan, M. Soltani, P. Alipour, and A. Adibi, "A temperature-insensitive third-order coupled-resonator filter for on-chip terabit/s optical interconnects," *IEEE Photonic Technol. Lett.*, vol. 22, no. 23, pp. 1768–1770, 2010.

[71] X. Zhu, K. Padmaraju, L. W. Luo, M. Glick, R. Dutt, M. Lipson, and K. Bergman, "Fast wavelength locking of a microring resonator," in *Proc IEEE Opt. Interconnects Conf.*, MB4, p. 1, May 2014.
[72] W. J. Dally, "Performance analysis of k-ary n-cube interconnection networks," *IEEE Trans. Comput.*, vol. 39, no. 6, pp. 775–785, Jun. 1990.
[73] J. Kim, W. J. Dally, and D. Abts. "Flattened butterfly: A cost-efficient topology for high-radix networks," in *Proc. 34th Annu. Int. Symp. Comput. Architect.*, 2007, pp. 126–137.
[74] J. Kim, W. J. Dally, S. Scott, and D. Abts, "Technology-driven, highly-scalable dragonfly topology," in *Proc. 35th Int. Symp. Comput. Architecture.*, Jun. 2008, vol. 77, no. 88, pp. 21–25, .
[75] F. Petrini and M. Vanneschi, "k-ary n-trees: High performance networks for massively parallel architectures," in *Proc. Parallel Process. Symp. 11th Int.*, Apr. 1997, pp. 87–93.
[76] S. Rumley, M. Glick, R. Dutt, and K Bergman, "Impact of photonic switch radix on realizing optical interconnection networks for exascale systems," in *Proc. IEEE Opt. Interconnects Conf.*, May 2014, pp. 88–89.
[77] A. Tomoyuki, M. Sugawara, and Y. Arakawa, "Quantum-dot semiconductor optical amplifiers," *Proc. IEEE*, vol. 95, no. 9, pp. 1757–1766, Sep. 2007.
[78] M. J. Heck, J. F Bauters, M. L. Davenport, J. K. Doylend, S. Jain, G. Kurczveil, S. Srinivasan, Y. Tang, and J. E. Bowers, "Hybrid silicon photonic integrated circuit technology," *IEEE J. Sel. Topics Quantum Electron.*, vol. 19, no. 4, Article 6100117, May 2013.
[79] G. Porter, R. Strong, N. Farrington, A. Forencich, P. Chen-Sun, T. Rosing, Y. Fainman, G. Papen, and A. Vahdat, "Integrating microsecond circuit switching into the data center," *SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 4, pp. 447–458, Aug. 2013.
[80] S. Kamil, L. Oliker, A. Pinar, and J. Shalf, "Communication requirements and interconnect optimization for high-end scientific applications," *IEEE Trans. Parallel Distrib. Syst.*, vol. 21, no. 2, pp. 188–202, Feb. 2010.
[81] K. Wen, D. Calhoun, S. Rumley, X. Zhu, Y. Liu, L. Luo, R. Ding, T. Baehr-Jones, M. Hochberg, M. Lipson, and K. Bergman, "Reuse distance based circuit replacement in silicon photonic interconnection networks for HPC," in *Proc. IEEE Symp. High Perform. Interconnects (Hot)*, Mountain View, Aug. 26–27, 2014.
[82] Q. Li, S. Rumley, M. Glick, J. Chan, H. Wang, K. Bergman, and R. Dutt, "Scaling star-coupler-based optical networks for avionics applications," *J. Opt. Commun. Netw.*, vol. 5, no. 9, pp. 945–956, Sep. 2013.
[83] S. Rumley, M. Glick, G. Dongaonkar, R. Hendry, K. Bergman, and R. Dutt, "Low latency, rack scale optical interconnection network for data center applications," in *Proc. Eur. Conf. Opt. Commun.*, Sep. 2013, pp. 1–3.
[84] K. Wen, S. Rumley, and K. Bergman, "Reducing energy per delivered bit in silicon photonic interconnection networks," in *Proc. Opt. Interconnects Conf.*, May 2014, pp. 123–134.
[85] S. Rumley, C. Gaumier, O. Pedrola, and J. Solé-Pareta, "Feedback based load balancing, deflection routing and admission control in OBS Networks," *J. Netw.*, vol. 5, no. 11, pp. 1290–1299, Nov. 2010.
[86] Q. Yang and K. Bergman, "Performances of the data vortex switch architecture under nonuniform and bursty traffic," *J. Lightw. Technol.*, vol. 20, no. 8, pp. 1242–1247, Aug. 2002.
[87] A. Shacham, B. A. Small, O. Liboiron-Ladouceur, and K. Bergman, "A fully implemented 12x12 data vortex optical packet switching interconnection network," *J. Lightw. Technol.*, vol. 23, no. 10, pp. 3066–3075, Oct. 2005.
[88] J. Dean and L. A. Barroso, "The tail at scale," *Commun. ACM*, vol. 56, no. 2, pp. 74–80, Feb. 2013.
[89] S. Hammond, K. S. Hemmert, S. Kelly, A. Rodrigues, S. Yalmanchili, and J. Wang, "Towards a standard architectural simulation framework," in *Proc. Workshop Model. Simul. Exascale Syst. Appl. Seattle, WA*, Sep. 18–19, 2013.
[90] A. Hoisie, D. Kerbyson, R. Lucas, A. Rodrigues, J. Shalf, J. Vetter, K. Barker, J. Belak, G. Bronevetsky, C. Carothers, B. Norris, and S. Yalamanchili. (2012, Nov.), "Report on the ASCR workshop on modeling and simulation of exascale systems and applications," University of Washington, Seattle, WA, [online]. Available: http://science.energy.gov/˜/media/ascr/pdf/program-documents/docs/ModSim_Report-2012_AH_5.pdf
[91] S. Rumley, R. Hendry, and K. Bergman, "Fast exploration of silicon photonic network designs for exascale systems," in *Proc. ASCR Workshop Model. Simul.*, Seattle, WA, Sep. 18–19, 2013.
[92] C. Chan, D. Unat, M. Lijewski, W. Zhang, J. Bell, and J. Shalf, "Software design space exploration for exascale combustion co-design," *Lecture Notes Comput. Sci.*, vol. 7905, pp. 196–212, 2013.
[93] S. Rumley, L. Pinals, G. Hendry, and K. Bergman, "A synthetic task model for HPC-grade optical network performance evaluation," in *Proc. Workshop Irregular Appl.: Architectures Algorithms*, Denver, CO, Nov. 17, 2013.
[94] P. Huerta, J. Castillo, C. Sanchez, and J. I. Martinez, "Operating system for symmetric multiprocessors on FPGA," in *Proc. Int. Conf. Reconfigurable Comput. FPGAs (ReConFig)*, Dec. 3–5, 2008, pp. 157–162.
[95] D. Calhoun, K. Wen, X. Zhu, S. Rumley, L. Luo, Y. Liu, R. Ding, T. Baehr-Jones, M. Hochberg, M. Lipson, and K. Bergman, "Dynamic reconfiguration of silicon photonic circuit switched interconnection networks," in *Proc. IEEE High Perform. Extreme Comput. Conf.*, Sep. 2014.
[96] C. P. Chen, X. Zhu, Y. Liu, T. Shiraishi, T. Baehr-Jones, M. Hochberg, and K. Bergman, "Multicasting using a high-radix silicon photonic switch," in *Proc. TECHCON*, Sep. 2014.
[97] X. Zhu, K. Padmaraju, D. Logan, L. Chen, J. Ackert, A. Knights, M. Lipson, and K. Bergman, "Pattern-dependent performance of microring modulators," in *Proc. Opt. Fiber Conf.*, Anaheim, CA, Mar. 17–21, 2013.
[98] D. Brunina, C. P. Lai, D. Liu, A. S. Garg, and K. Bergman, "Optically-connected memory with error correction for increased reliability in large-scale computing systems," in *Proc. Opt. Fiber Commun.*, Mar. 2012, pp. 1–3.
[99] Q. Li, Y. Liu, K. Padmaraju, R. Ding, D. F. Logan, J. J. Ackert, A. P. Knights, T. Baehr-Jones, M. Hochberg, and K. Bergman, "A 10-Gb/s silicon microring resonator-based BPSK link," *IEEE Photon. Technol. Lett.*, vol. 26, no. 18, pp. 1805–1808, Sep. 2014.
[100] D. L. Ly, M. Saldana, and P. Chow, "The challenges of using an embedded MPI for hardware-based processing nodes," in *Proc. Int. Conf. Field-Programmable Technol.*, Dec. 9–11, 2009, pp. 120–127.
[101] D. Brunina, X. Zhu, K. Padmaraju, L. Chen, M. Lipson, and K. Bergman, "10-Gb/s WDM optically-connected memory system using silicon microring modulators," in *Proc. Eur. Conf. Opt. Commun.*, Sep. 17, 2012, pp. 1–3.
[102] T. Shiraishi, Q. Li, Y. Liu, X. Zhu, K. Padmaraju, R. Ding, M. Hochberg, and K. Bergman, "A reconfigurable and redundant optically-connected memory system using a silicon photonic switch," *Opt. Fiber Conf.*, San Francisco, CA, Th2A.10, Mar. 2014, pp. 9–13.
[103] D. J. Richardson, J. M. Fini, and L. E. Nelson, "Space-division multiplexing in optical fibres," *Nat. Photon.*, vol. 7, pp. 354–362, 2013.

Sébastien Rumley received the M.S. degree in communication systems and the PhD degree in computer and communication sciences, both from Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, in 2005 and 2011, respectively, after studying in Lausanne, Zurich (ETHZ) and Santiago de Chile (PUC), Santiago, Chile.

He is currently an Associate Research Scientist in the Department of Electrical Engineering, Columbia University in the City of New York, NY, USA. His research focuses on optical interconnect modeling, mainly for applications in High-Performance Computing and Distributed Computing platforms.

Dessislava Nikolova received the M.Sc. degree in solid state physics from Sofia University, Sofia, Bulgaria and the Ph.D. degree in computer science from University of Antwerp, Antwerp, Belgium.

She spent two years as a Research Engineer in optical access networks with Alcatel working on scheduling algorithms for passive optical networks. After receiving the Ph.D. degree, she was awarded a Marie-Curie Fellowship to study the interaction between magnetism and plasmonics on the nanoscale at the London Center for Nanotechnology at University College London. Her current research is focused on the design and analysis of on-chip silicon photonic networks combining high level network design principles with fundamental electromagnetic theory.

Robert Hendry received the B.S. degree in computer science from Hobart College in Geneva, NY, USA, in 2010. He received the M.S. degree in electrical engineering from Columbia University, New York, NY, in 2012. He is currently working toward the Ph.D. degree at Columbia University under the supervision of Professor Keren Bergman, pursuing research in optical interconnection networks for parallel computer architectures and distributed systems.

Qi Li (S'10) received the B.Eng. degree in electrical and computer engineering with a minor in mathematics from the Hong Kong University of Science and Technology, Hong Kong, in 2010. He received the M.S. degree from Columbia University, New York, NY, USA, in 2012, where he is currently working toward the Ph.D. degree in the Department of Electrical Engineering.

He spent the Spring of 2009 as an exchange student at Cornell University, Ithaca, NY, USA. His research interests include silicon photonics and optical interconnection networks. He is a Student Member of the IEEE.

David Calhoun received the B.S. degree in electronics and communication engineering from Rowan University, Glassboro, NJ, USA, in 2012, and the M.S. degree in electrical engineering from Columbia University in the City of New York, NY, USA, in 2013.

He is currently a Graduate Research Assistant with the Lightwave Research Laboratory at Columbia University, New York, NY, USA. His current research focus is at the systems level of silicon photonic networks, developing optical network interface logic and protocols for link layer data delivery methods, physical layer control systems, and application layer runtime execution methods using high-speed FPGAs. He is a Member of the IEEE, OSA, and SPIE, and is currently an NSF-funded Ph.D. Research Fellow through the Columbia University Optics and Quantum Electronics IGERT.

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.Dr. Bergman is currently the Charles Batchelor Professor and Chair of Electrical Engineering 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 the IEEE and OSA.