Energy-Efficient Photonics in Future High-Connectivity Computing Systems

A. V. Krishnamoorthy, *Fellow, IEEE, Fellow, OSA*, H. Schwetman, X. Zheng, *Senior Member, IEEE*, and R. Ho, *Senior Member, IEEE*

(Invited Paper)

***Abstract*—Energy-efficiency has become a critical parameter in the design of high-performance computing systems. Typically, compute elements consume the most energy in current systems, withmemories and interconnect networks close behind. This paper proposes an energy-efficient, high-connectivity, petascale computing system for the year 2020 timeframe by addressing the energy requirements of these three components. We start with projections based on purely evolutionary computer system design trends, then include the impact of breakthroughs in processor design, memory packaging and optical interconnect technologies. Based on these projections, we motivate the development for a 1-pJ/b optical intrasystem interconnect technology that significantly increases system interconnect bandwidth and relieves the distance-based energy dependence of electrical alternatives. We show that improvements in compute, memory, and IO, when simultaneously applied, become a vision for many-chip photonically-interconnected modules that could lead to an order of magnitude improvement in energy efficiency in the 2020 timeframe. The vision hinges on a high-density, energy-efficient optical link that can connect electronic compute and memory elements across short chip-to-chip distances while also capable of kilometer or longer spans across data centers. We discuss the power budget to enable such a link and review experimental progress toward creating an ultra-dense, hybrid-integrated low-power silicon photonic link that will enable this vision.**

***Index Terms*—Optical communications, optical interconnections, optical receivers, optical transmitters, silicon photonics, transceiver array, VCSELs, WDM.**

I. Introduction

WITH the growth of warehouse-scale applications that must address compute-intensive and/or data-intensive workloads, we see a growing need for energy-efficient datacenter systems. Almost all such applications have the over-arching need for extremely large amounts of main memory - more than can be accommodated on a traditional single-socket blade or a card. As a result, the systems hosting these applications consist of hundreds or thousands of nodes, each with a number of sockets (themselves with multiple computing cores and/or application-specific co-processors) and several giga-bytes of memory; the nodes are linked with an interconnection network. The scale of these systems aggregates sufficient memory so that applications can retain the required data in main memory and execute in an effective manner. A major factor in the construction and operation of such a massively parallel system is the electrical energy and power required to operate and cool the total system. An unfortunate consequence of the scale of such systems and the energy management mandate is that intra-system connectivity is often compromised. In many cases, the time-independent ratio of bandwidth to execution rate, or *Bytes/s* to *Flop/s* dwindles to a few percent of a Byte/Flop.

In this paper, we investigate a number of technology advances that could enable such high-performance and high-capacity systems without requiring excessive operational power while maintaining a high Byte/Flop computing system. We show that future systems will need dramatic improvements in processors, main memory *and* interconnection networks. Related studies have come to similar conclusions regarding the need for more efficient computing hardware and memory [1], [2], and even specifically targeted the interconnect energy as a key metric [2], [3], but did not analyze the system improvements that could be achieved using high-bandwidth, low-energy optical interconnects across the system. In [3], optical interconnects were briefly considered but were dismissed due to inefficient lasers that resulted in large optical link energies. Earlier studies suggested a path to an optically-interconnected machine with high system bisection bandwidth, but did not consider link energy or efficiency requirements that would be necessary to accomplish the system [4]. In this paper, we provide projections for required compute, memory, and interconnect technologies, and emphasize the system impact of an optical interconnect technology capable of high-bandwidth intra-system communication on the order of ∼0.5 Byte/Flop at an energy of ∼1 pJ/bit. We motivate a design for a compact, macrochip that would enable a rack-based petascale system. We review recent progress made towards achieving a pJ/bit silicon photonic data-center link and provide experimental evidence that such transmitters can be used across extended data centers spanning many kilometers. Finally, we develop a detailed link budget based on demonstrated silicon photonic components and hybrid silicon-assisted laser sources.

Manuscript received July 28, 2014; revised December 30, 2014; accepted January 15, 2015. Date of current version February 20, 2015. This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions, and/or findings contained in this article/presentation are those of the author(s)/presenter(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Approved for public release; distribution unlimited.

A. V. Krishnamoorthy and X. Zheng are with Oracle Systems, San Diego, CA 92121 USA (e-mail: ashok.krishnamoorthy@oracle.com; xuezhe.zheng@oracle.com).

H. Schwetman is with Oracle Labs, Austin, TX 78727 USA (e-mail: herb.schwetman@oracle.com).

R. Ho was with Oracle Labs, Redwood Shores, CA 94065 USA, and is now with the Altera Corp., San Jose, CA 95134 USA (e-mail: ronho@ieee.org).

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.2015.2395453

TABLE I
THREE ~1 PETAFLOPS SYSTEMS FROM THE TOP 500 LIST (NOV. 2014)

System	Magnus Cray XC40 Xeon 2690v3	QB-2 Dell C8220X Xeon 2680v2/nVidia K20	Big Red 2 Cray XE6 AMD 6276/nVidia K20
Peak Flops (PFlops)	1.486	1.473	1.000
General purpose cores	35712	9600	21824
Floating point (GPU) cores	0	13440	9464
Memory (TB)	93	30	43
System Power (MW)	0.697	0.500	~0.500
Mem BW (TB/s)	98.812	28	34
I/O BW (TB/s)	3.0	10	35.19
Mem Cap (B/Flops)	0.064	0.021	0.043
Mem BW (Bps/Flops)	0.67	0.019	0.034
I/O BW (Bps/Flops)	0.002	0.007	0.035
Gflops/Watt (energy)	2.129	2.947	2.001

TABLE II
GRAPH500 LIST - GRAPH SCALE 32 – NOV. 2014

System	Peak GFlops/node	Peak network BW/node (GBps)	Bps/Flops
Appro HA-PACS	166.4	4	0.024
SGI Statiscal Science	259.2	6.8	0.026
Intel Endeavor	218	6.8	0.031
NEC TSUBAME-KFC	100.8	6.8	0.067
Cray Big Red 2	981	80	0.082

This paper focuses on a petascale system capable of 10^{15} floating point (or equivalent integer) operations per second. While such systems already exist today, we envision one that is far smaller and more energy efficient than today's systems. Whereas the "Flops" metric is used for compute-intensive systems, we will motivate both compute-intensive and data-intensive systems of roughly equivalent sizes and capabilities.

This paper proceedsas follows: Section II provides an overview of contemporary high-performance computing systems highlighting power consumption and network Byte/Flop ratios; Section III shows a plan for a computing system based on evolutionary technology projections; Section IV shows an alternate plan based on drastic improvements in the computing components; Section V shows a vision for a optically-interconnected computing system in which all components are improved in a super-evolutionary manner and provides basic performance and efficiency projections; Also discussed is the optical network connecting the components; Section VI reviews progress made towards creating dense optical interconnections capable of these performance targets; Section VII presents a link budget representative of a canonical multi-wavelength silicon photonic link Section VIII provides concluding remarks.

II. PETASCALE SYSTEMS TODAY

Modern large-scale datacenters aggregate a collection of processing nodes, where each node has a compute component, typically with multiple sockets and several cores per socket, and a main memory component. The cores can be multi-threaded, thus increasing the number of parallel computations in progress at the same time. All of these nodes are attached to one or more interconnection networks. In some cases, the nodes have privately owned memory components and do not explicitly allow shared memory access by other nodes, but there are exceptions to this arrangement that result in "shared-memory" systems. In all cases the memory on a single node is shared between all processors on the node.

Table I above shows key attributes of three systems from the Top500 Supercomputer List [5]; each provides ~1 petaFlops compute capability. The "QB-2" system and the "Big Red 2" systems include both standard compute processors and GPU processors. All three of these systems are also on the Green500 list [6], a list of the top energy-efficient systems in terms of GFlops/Watt.

The examples in Table I present performance on a compute-intensive application. Such systems are specifically targeted to compute intensive tasks (GFlops) and energy efficiency (GFlops/Watt) and, in the case of the first system, also a high memory bandwidth. However, none of these systems specifically provide large network communication bandwidth, as evidenced by the I/O bandwidth (Byte/Flop) ratios. The Graph500 list, Table II, shows results for a data-intensive application, the Graph500 benchmark [7] which focuses on "Big Data" computing. This benchmark generates a random graph data structure, and then traverses all of the edges in this graph in breadth-first search (BFS) order. The figure of merit for the benchmark is TEPS–traversed edges per second—which is a measure of the edge processing rate of the system.

From Tables I and II, note that the Big Red 2 system is on the Top500 list, the Green500 list and the Graph500 list representing strong performance in raw compute power, compute efficiency as well as graph processing. This system is also of interest because it is equipped with two kinds of processing nodes: 676 nodes, each having an nVidia GPU, and an AMD Opteron CPU, and 344 nodes, each having two AMD Opteron CPUs. The Top500 result (one petaFlops) was achieved using the nVidia GPUs, while the Graph500 result used the AMD CPUs [8].

Despite such innovations, all of the systems display relatively modest communication/compute ratios resulting from network bandwidth limitations – even the Big Red 2 system which is among the highest (Table II). This tradeoff of compute efficiency versus network I/O bandwidth is systemic of large datacenter deployments and represents a key interconnect efficiency versus performance bottleneck. This can also be seen with new benchmarks such as the Green Graph500 [9], aimed at capturing the most energy-efficient graph processing computing systems in terms of MTEPS/W. The top 20 GreenGraph performers are smaller, single-node machines - highlighting the difficulty in

TABLE III
TARGET PARAMETERS FOR PROPOSED DATACENTER

Peak effective Flops	1 PetaFlops
Memory capacity	256 TB
Memory b/w per flop (Bps/Flops)	1
Network b/w per flop (Bps/Flops)	0.5
Total memory B/W	1000 TBps
Total network B/W	512 TBps
Last level cache per Flop (MB/GFlops)	0.0625
Total LLC capacity	64 MB
Disk capacity per flop (B/Flops)	3.5
Total disk capacity	3.5 PB

efficiently scaling-up interconnect-intensive big-data tasks to large numbers of nodes.

III. Evolutionary Datacenters

How can we understand future highly-connected petascale systems that could display far more energy efficiency than Table I machines? One method is to optimistically extrapolate the power, energy, and performance of existing designs to a petascale system of the future. We call this the evolutionary approach, where evolutionary refers to a system that undergoes changes that evolve from predictable trends in current components and technology. This is similar to the methodology used in [2], [3]. Table III gives the requirements of the system used in this paper. The parameters in Table III were chosen to be representative of an aggressive petascale system of the future that simultaneously provides a system-wide memory bandwidth of 1 Byte/Flop and a high intra-system network bandwidth of 0.5 Byte/Flop. Such a machine would be valuable not only for graph processing applications, but also for scaling-up well-known challenge benchmarks including GFFT, STREAM, and GUPS, by removing memory-bandwidth and interconnectivity bottlenecks [10]. Hence, in Table III, the target memory bandwidth is specified at 1 byte/s per flops (0.5 bytes in and 0.5 bytes out), and the target network bandwidth is balanced with this at 0.5 byte/s per flops.

A modern CPU, for instance the Intel Haswell, achieves ~500 peak GFlops with a thermal design power of 135 watts. This includes the on-chip caches and the power required to drive the memory interconnect. We assume that this memory interconnect energy use is 10 pJ/bit (representing just the interconnect and the timing circuits that drive data to and from the DRAM), leading to approximately 10 watts for 1 Byte/Flop memory usage, and hence a compute, cache, and memory interconnect energy of approximately 1.5 GFlops/watt. The Haswell processor is built in a 22 nm technology; in the 2020 time frame, smaller feature sizes will lead to approximately a factor of 2 improvement in energy to 3 Gflops/watt. This compute factor (Table IV) is calculated as the power required for the computation (FPU) plus processor overhead plus the power required for the on-chip cache. We also expect that the power factor for main memory will decline from 500 mW/GB to 200 mW/GB and the energy to drive the memory interconnection network will decline from 10 pJ/bit to 5 pJ/bit.

TABLE IV
TECHNOLOGY ASSUMPTIONS FOR MACROCHIP-BASED DATA CENTER

Technology assumptions	Today (2014)	Evolutionary (2018)	Hybrid (2018)	Macrochip (2020)
FPU power (Gflops/W)	37	6.1	50	50
Processor overhead factor	2.0	2.0	2.0	2.0
Cache power (W/MB)	0.1	0.1	0.1	0.1
Compute power (Gflops/W)	1.5	3.0	21.6	21.6
Memory power (mW/GB)	500	200	200	50
Cache - memcontroller power (pJ/bit)	0.1	0.1	0.1	0.1
Memory interconnect power ((pJ/bit)	10	5.0	5.0	1
Network /on-macrochip power (pJ/bit)	5.0	5.0	5.0	1.0
Switch/ off-macrochip power (pJ/bit)	7.0	7.0	7.0	3.0
Cooling overhead	0.1	0.1	0.1	0.1
Disk power (W/GB)	1.7E-3	1.7E-3	1.7E-3	1.73E-3

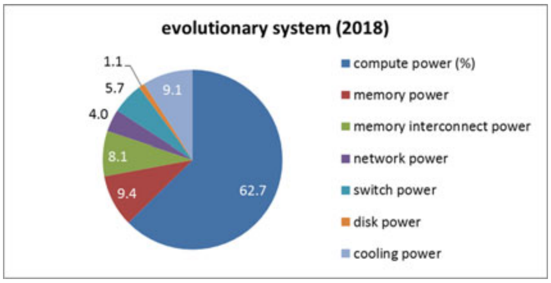


Fig. 1. System power allocation (%) for an evolutionary system in 2018.

Using these assumptions leads to the power projections summarized in Table IV, Section V. The allocation of these power categories in terms of percentages for the evolutionary system is given below (Fig. 1) showing that the compute section consumes the greatest proportion of the power.

IV. Improving Computing Efficiency

Increasingly, computing systems are turning towards highly specialized compute units tailored towards specific workloads. Incorporating such customized blocks, for example as heterogeneous co-processors in a multi-chip CPU, appears to be an effective way to improve the energy efficiency of compute units. The nVidia K20 GPU has an energy efficiency of approximately 5 GFlops/W; by contrast, the Intel Sandy Bridge processor has an energy efficiency of approximately 1.5 GFlops/W.

What is the limit for such energy efficiency? We start with a basic arithmetic function [3]. In an advanced technology a 64-bit floating point ADD consumes 25 pJ, which corresponds to an efficiency of 50 GFlops/W if the entire application could

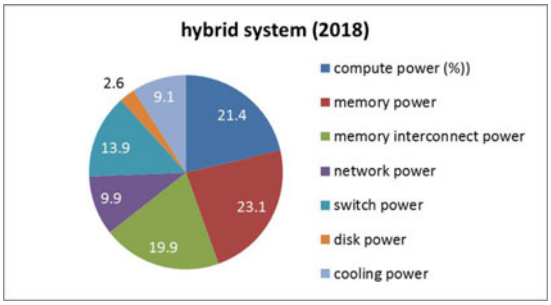


Fig. 2. System power allocation (%) for a hybrid system with improved compute efficiency.

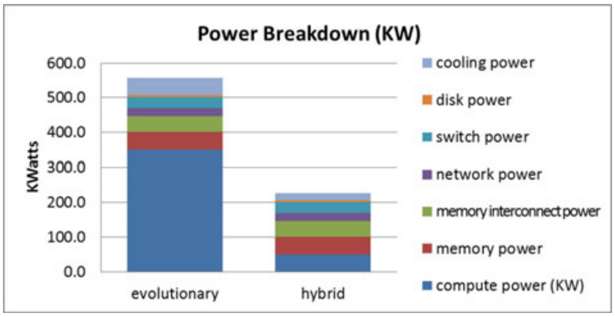
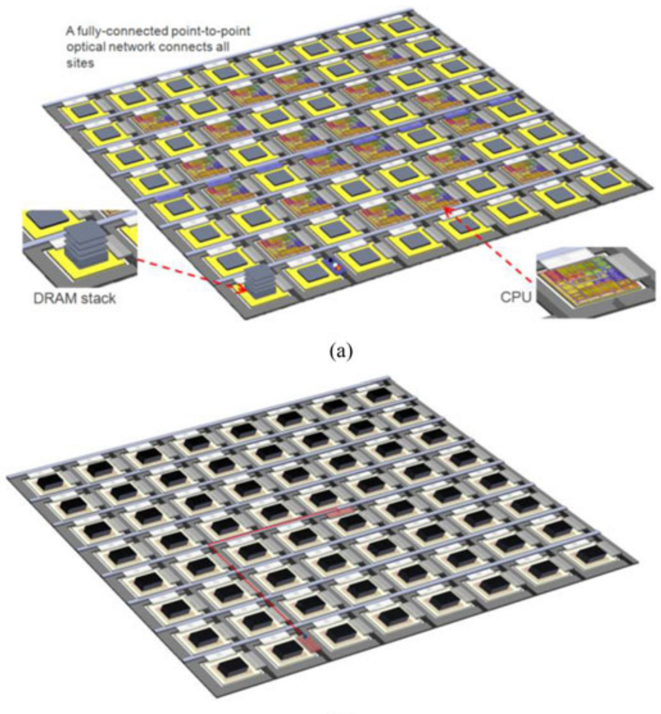


Fig. 3. Comparison of an “evolutionary” system to a “hybrid” system.

TABLE V
Power Allocation for Macrochip-based System

Power (KW)	Today (2014)	Evolutionary (2018)	Hybrid (2018)	Macrochip (2020)
Compute power	973.4	349.5	48.5	48.5
Memory power	131.1	52.4	52.4	13.1
Memory interconnect power	90.1	45.0	45.0	8.4
Network power	22.5	22.5	22.5	2.1
Switch/ I/O port power	31.5	31.5	31.5	6.3
Disk power	6.0	6.0	6.0	6.0
Cooling power	98.7	50.7	20.6	7.8
Total system power	940.0	557.7	226.6	92.2

be run within specialized blocks just like that simple adder. Our system also needs a sequencing engine that pipelines and controls these specialized blocks, handles interrupts, and satisfies memory references. After accounting for this control processor, we would be hard-pressed to reach 25 GFlops/W. Caching and off-chip memory, interconnect power might reduce this further to 21 GFlops/W. Without attempting to fully describe a design that meets this efficiency target, we might reasonably claim that a system built from specialized compute blocks and a separate sequencing engine would hit a stretch goal of 21 GFlops/W for compute efficiency. A system using the components in Section III along with this specialized processor (the hybrid system) would have power characteristics as shown in Figs. 2 and 3 and summarized in Table V, Section V.


A fully-connected point-to-point optical network connects all sites
DRAM stack
CPU

(a)

(b)

Fig. 4. Diagram of an 8 × 8 macrochip with: (a) separate CPU & stacked DRAM sites connected with optical links; (b) 64 combination CPU & stacked DRAM sites with electrical CPU-to-DRAM communication within a site and optical communication (e.g. highlighted link) between sites. Not shown are optical waveguides from each site to the edge of the macrochip to an additional I/O port that provides fiber connections to other macrochips.

V. Improving All Major Components of the Datacenter Using a Macrochip

The data in Figs. 2 and 3 shows that even with specialized hardware blocks to improve compute efficiency, significant energy is still consumed by the memory unit, the processor-memory interconnect, and the processor-processor network. Further advances in system efficiency must come from technology advances that broadly address these other components, through packaging and interconnect that enables compact multichip solutions. One way to attack these problems is represented by what we call a “macrochip” [10]-a proposed multichip platform that offers significant energy reductions in these categories of power usage.

The macrochip (visualized in Fig. 4) uses a silicon substrate with embedded silicon photonic (optical) links to interconnect up to 64 sites bonded to the top of the substrate. A local “bridge” converts electrical IO on each site to optical signals for transmission across the built-in optical links. One design employs a single communications layer, configured in a planar point-to-point network implementing all-to-all communications for the sites [11]. Fig. 5 shows a suggested layout for an 8 × 8 macrochip with only two optical hops (bridge-to-substrate at the beginning of the link and substrate-to-bridge at the end). An alternate Manhattan waveguide geometry would reduce the average length of each optical link but would also require four optical hops.

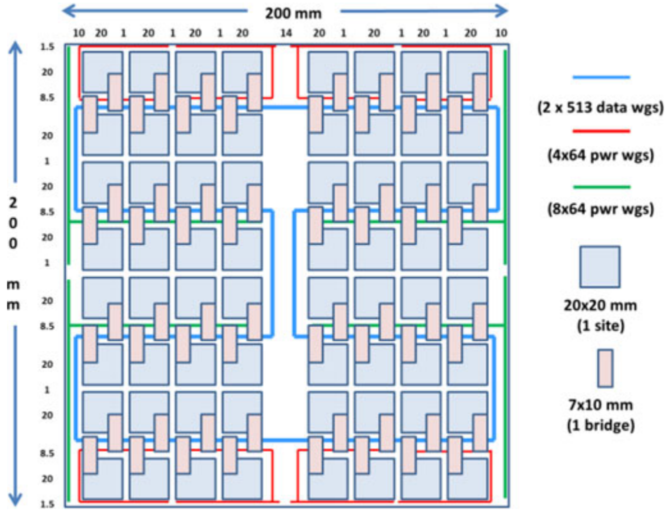


Fig. 5. Layout for 8 × 8 macrochip.

Each site in this version of a macrochip is a "combo" site as depicted in Fig. 4(b): it includes both a compute component and a portion of the main memory. The compute component consists of a CPU with optimized FPU units, and caches. It communicates electrically to its on-site main memory, a 3D stack of memory chips. This particular macrochip form is designed to take advantage of 3-D stacked memory modules that are expected by the end of this decade.

The datacenter will consist of 32 of these macrochips housed in two racks (connected by single-mode optical fibers, with these fibers being connected to the edges of each macrochip.

The system has 1 PFlops of compute composed of 32 macrochip nodes each providing 32 TFlops of compute capacity. Each node thus has 64 sites and each site has 512 GFlops of compute capacity. Each site has memory (with 1B/Flop of electronic memory bandwidth) and optical interconnect bridges that provide 0.5 Byte/Flop of network bandwidth. The optical I/O bandwidth per site therefore is 256 GByte/s = 2048 Gbps of optical transmit/receive bandwidth. Half of this bandwidth exits the macrochip for inter-macrochip communication, and half stays on for inter-site communication within the macrochip. Each wavelength supports 16 Gbps. There are 64 sites on each macrochip. Hence there is exactly one wavelength x 16 Gbps from a site to every other site. Each site on a macrochip also has $64 \times 16 = 1024$ Gbps to another macrochip. Hence there are 8 waveguides with 8 wavelengths each with 16 Gbps/wavelength (=1024 Gbps) exiting each site of the macrochip - terminating into 8 fibers - for a total of 64 × 8 fibers = 512 exiting fiber connections exiting each macrochip (and a corresponding 512 entering). These 512 fibers are partitioned into 2 × 32 macrochip I/O ports with the following rule: 8 fibers from site (2N-1) and 8 fibers from site 2N form the macrochip I/O port N, N = 1..32.

Each macrochip has 32 I/O ports. Assuming a direct point-to-point network between macrochips, the network can be specified as follows: output ports 1 to 31 of macrochip 1 are connected to all the input ports 1 of macrochip 2 to 32; output ports 1 to 31 of macrochip *i* are connected to all the input ports *i* of macrochip 1 to 31; and finally outputs 1 to 31 of macrochip 32 are connected to all the input ports 32 of macrochip 1 to macrochip 31. Each connection is 16 fibers wide per direction (Fig. 6).

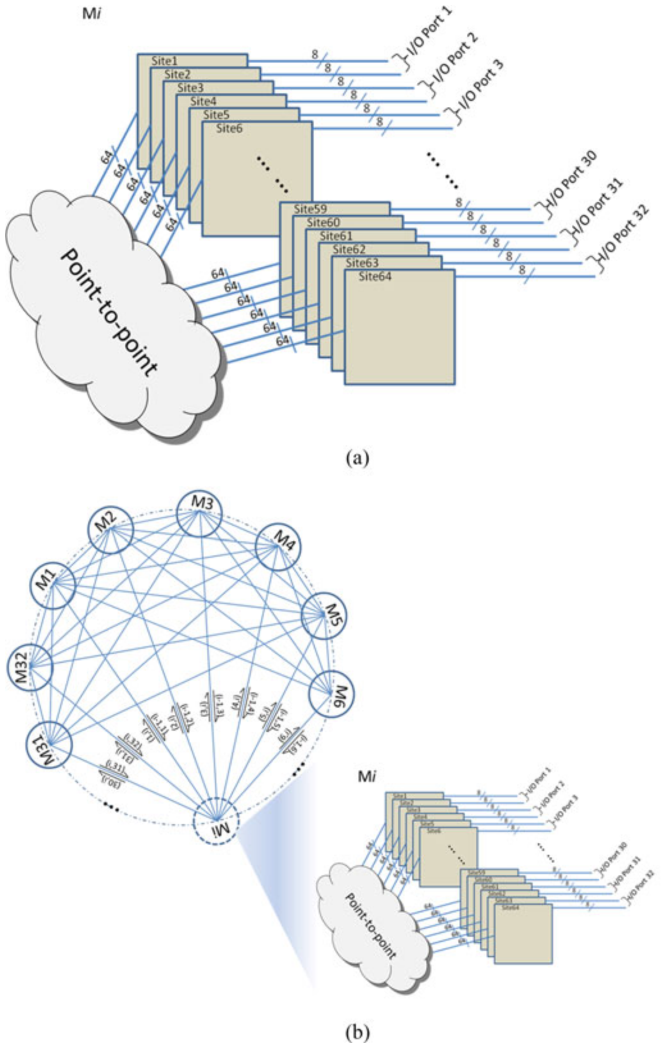


Fig. 6. Connectivity for a 32-macrochip system: (a) the on-macrochip network is a fully-connected point-to-point network. The inter-macrochip network is a fully-connected point-to-point network between 32 macrochips using 32 network ports per macrochip. Each connection provides 2.048 Tbps using 16 fibers with 8 wavelengths and 16 Gbps/wavelength.

Connectivity between any two sites in the system is achieved through 3 optical links (e.g. a source-macrochip link; an inter-macrochip link; a destination macrochip link). The application performance and specification of this distributed point-to-point interconnection network connecting the macrochips to each other is an area of on-going investigation. For the purposes of this paper, the power requirements for this distributed point-to-point interconnect via the optical links are included in the following analysis, and peak bandwidth numbers are used. The assumptions about power requirements for this system versus the others are given in Table IV, and the power allocation for this system is shown in Table V. Fig. 7 charts corresponding power allocation percentages. Memory power savings come primarily from having compute and memory co-located on each

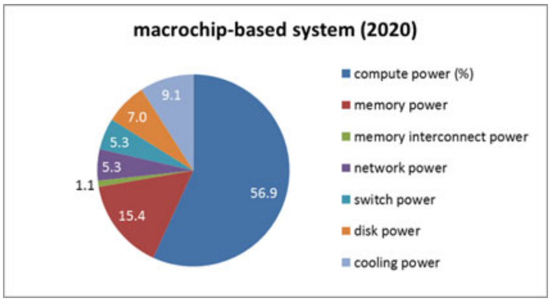


Fig. 7. Power allocation (%) for a macrochip-based system.

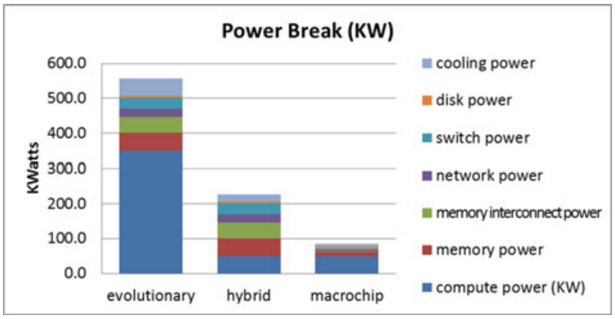


Fig. 8: Comparison of power allocation for three petascale systems (2018+).

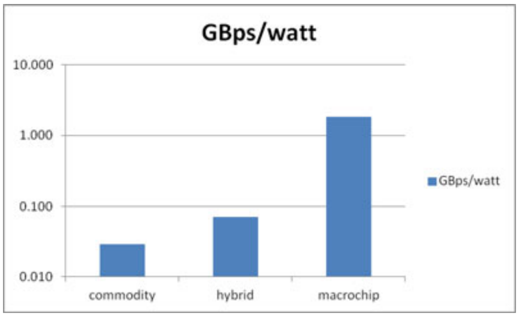


Fig. 9. Bisection bandwidth of evolutionary (commodity), hybrid, and macrochip-based petascale systems.

site. Some of these savings arise from the very short distance between the compute elements and the memory (the memory interconnect power) and some from improvements in memory packaging such as 3-D stacking. We assume that in all cases, commodity memory parts for the specified time frame will be used. The major change will relate to increased bit densities for memory chips, ranging from 0.5 GB/chip today to 4 GB/chip in 2018 (using, for instance, a through-silicon-via-stacked set in a single-die form factor).

The improvement in network and switch power stems from the site-to-site optical interconnection networks on the macrochip and between macrochips. The macrochip has been designed to provide a fully connected, all-to-all network with on-macrochip power of ~1.0 pJ/bit (see Section X). This stems primarily from the enhanced energy efficiency of optical communication. The system will require not only the on-macrochip communications described here but also inter-macrochip communications at ~3 pJ/bit through the macrochip I/O port. As discussed above, we envision fiber optic communications connecting each macrochip pairwise through two intermediate I/O ports. Based on the transmission performance of the silicon photonic links through extended fiber links presented in Section VI, we expect that an extended reach 3-hop network can also be implemented to require ~3 pJ/bit.

As discussed earlier (Table III), the focus of this paper has been a system with a peak compute rate of 1 PFlops, a memory bandwidth of 1 petabyte/s (in + out) and network bandwidth of 0.5 petabytes/s (balanced to the memory bandwidth). The differences in the systems presented for the year 2020 timeframe are in the power required to meet these goals using different implementation technologies and strategies. Fig. 8 shows the power consumption for the three designs. Fig. 9 shows the network bisection bandwidth energy (gigabytes-per-second/watt) for the three designs. A precise prediction of the performance of specific benchmarks or applications relies on achieved compute rates and network usage and is beyond the scope of this paper. Nevertheless, comparing peak performance alone, the 1 PFlops macrochip-based system uses about 6.2 times less power than the equivalent evolutionary system derived from commodity parts projected for the 2018–2020 time frame. It is also interesting to note that the relative proportion of compute, memory, and interconnect power for a macrochip system is similar to that of an evolutionary system – even though the latter has a significantly enhanced peak bisection bandwidth and Byte/Flop ratio.

VI. Dense, Energy-efficient Silicon Photonic Links

One of the central assumptions of the macrochip-based system is the availability of a silicon CMOS photonic foundry that can produce dense, low-power optical link components and optical substrates with high reproducibility and yield [12], [13]. Fortunately, such foundries are being developed around the globe. Using such a foundry, all key technologies on a common platform to support this vision have been demonstrated: ultra-low loss (<2.6 dB/meter) transport waveguides [14]; waveguide grating couplers [15]; silicon-on-insulator optical devices with etched thermal isolation and fine-pitch bumps [16]–[18]; bulk CMOS interface circuits that transmit data from flip-flop to flip-flop with high fidelity and low energy [19]; rematable packaging technologies delivering the required mechanical alignment, hybrid bonding, cooling, and power delivery [20]; high-speed (10–40 Gbps) ring resonator modulators [21]; energy-efficient integrated transmitters (<1 mW) [22], detectors (responsivity > 0.9 A/W across 40 nm) and receivers (<3 mW) [23], 8-wavelength arrays [24]; tunable wavelength multiplexers and demultiplexers [25]; low-power wavelength tracking and locking techniques [26], [27]; preliminary fiber links with an on-chip power of 2 pJ/bit excluding laser [28]; and efficient hybrid external-cavity tunable silicon-assisted lasers [29], [30]. These demonstrated components collectively provide a bandwidth-density of over 500 Gbps/mm^2 [24]. The drivers and receivers made in a bulk CMOS process have been flip-chip integrated to the CMOS photonic chips to create hybrid photonic transmitters and receivers. All photonic components utilized a commercial, foundry-based

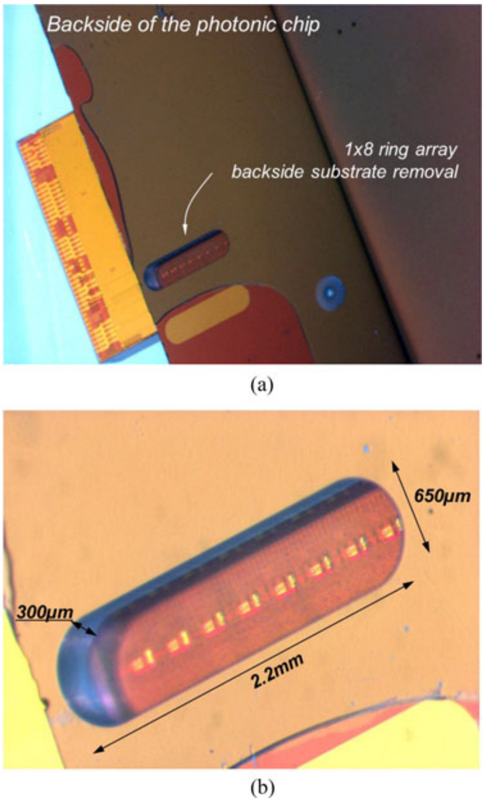


Fig. 10 [37]: (a) Localized silicon substrate removal on the modulators improves thermal tuning efficiency by 10x (b) modulator/mux array seen through buried silicon dioxide layer that serves as an etch stop.

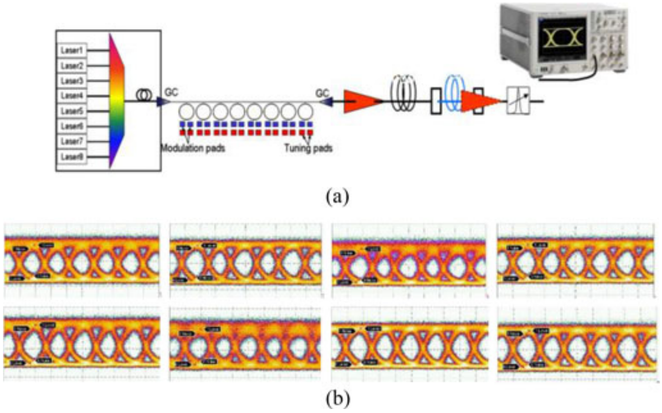
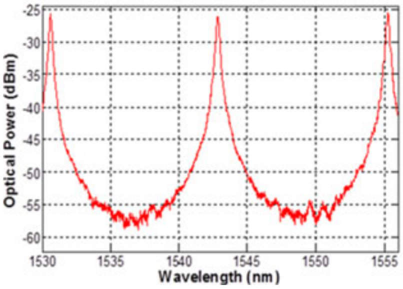


Fig. 11. (a) 100 Gbps (8 channels × 12.5 Gbps/channel) WDM transmission experiment; (b) optical eyes after 40 km of single-mode + dispersion compensated fiber.



	P (mW)
Ch1	0.9
Ch2	0.0
Ch3	0.5
Ch4	0.4
Ch5	0.6
Ch6	1.2
Ch7	1.3
Ch8	0.3
Total	5.1

Fig. 12. (a) Drop port spectrum for one channel; (b) power required per channel to tune to 200 GHz spaced synthetic comb.

digital-CMOS SOI platform co-optimized with a common rib-waveguide that allows all link components to be co-integrated on a photonic wafer with a process that minimizes the number of custom etch steps and variation across wafers [31]. A hybrid flip-chip bonding process was used to integrate the circuits with the photonic components and optimize an energy-performance tradeoff.

As an example, a flip-chip-integrated 100 Gbps ring-modulator-based WDM transmitter using an off-chip laser source is shown in Fig. 10. Multiple rings, each with resonant wavelengths spaced to approximately span the free-spectral range of the rings create a synthetic resonant comb that minimizes tuning range requirements. The link consists of a cascaded array of eight ring modulators that track and modulate the respective input wavelength. The transmitter on-chip power (excluding laser) was 33 mW including integrated on-chip thermal tuners controlled by the CMOS circuit. Also included was a VLSI driver circuit that operated at 12.6 Gbps per channel with a circuit + modulator dynamic power dissipation of less than 1 mW/channel. The thermal tuning efficiency of each ring modulator was improved by an order of magnitude with localized substrate removal. The output waveguide contains the 8-wavelength data (100 Gbps in this case) and is coupled through waveguides (or fiber) to the receiver chip, which contains a ring-resonator-based demultiplexer array that selects the corresponding wavelength and feeds to a high-speed, broadband integrated germanium waveguide detector and CMOS receiver.

The modulation quality of this energy-efficient Si photonic WDM transmitter suggests its use not only for short-reach intra/inter chip interconnects, but also across data-centers and even long-reach WDM link applications. This was tested by measuring eye diagrams back-to-back and after 40 km of standard single-mode + dispersion compensated fiber. Measurements taken after 40 km (Fig. 11) show wide open eyes for all 8 wavelength channels, proving the efficacy of the technology for reliable communication from <1 m to several km –sufficient to support even the largest data-centers.

The demultiplexer is another key component required for a multi-wavelength link. An 8-channel demultiplexer array with silicon-based-resistor heaters built into each ring and localized substrate removal was implemented and tested. The array was tuned to the appropriate 200 GHz spacing ($\sim$1.6 nm) with only $\sim$5 mW of power (or $\sim$0.63 mW per channel) as shown in Fig. 12. This is consistent with statistical data that suggests about 0.52 mW/channel will be needed on average for this particular design and process [31]. The measured tuning required per channel is provided in Fig. 12 and corresponds to $\sim$50 fJ/bit at a channel data rate of 12.5 Gbps.

VII. Link Budget for Silicon Photonic Intra-Module and Inter-Module WDM Connectivity

The silicon photonic intra-module WDM link for a macrochip depicted in Figs. 4 and 5 would employ an array of modulators

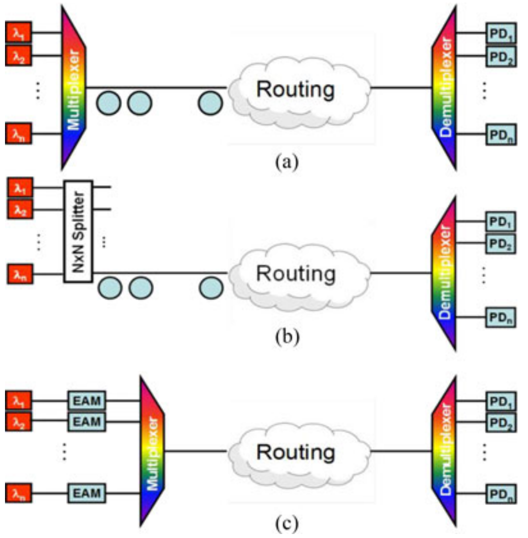


Fig. 13. WDM silicon photonic link configuration using: (a) cascaded ring modulators with source wavelength multiplexer; (b) source power combiner/splitter; and (c) electro-absorption modulators.

multiplexed into one waveguide on the bridge of a source site, coupled to the wafer routing layer (which could be completely passive) through grating coupler-based optical proximity communications (OPxC) [32], and then routed to a destination site using dedicated waveguide. Followed by a second OPxC coupling to the bridge chip of the destination site, the WDM signal would be demultiplexed and received by an array of optical receivers. An inter-module WDM link is slightly different. The WDM signal from/to a source/destination would be coupled to/from the routing layer and routed to/from the edge of the module, and then coupled into a fiber connected to another module.

Depending upon the photonic devices used, the WDM links are configured in slightly different forms. When wavelength selective modulators are used, for example, silicon ring modulators, a comb source is preferred. Unfortunately, it is a challenge to make an efficient comb source good enough to support high speed (e.g. >25 Gbps) optical transmission. One practical implementation of a comb source today would be to use a wavelength multiplexer to combine multiple single wavelength lasers, or an NxN broad band power splitter to mix N laser sources into an N-wavelength comb source with power evenly divided into N waveguides. A complete WDM silicon photonic link based on such integrated comb sources will have the configurations shown respectively in Fig. 13(a) and (b). When broadband modulators are used (e.g. electro-absorption modulators), laser sources with different wavelengths are needed before each modulator, resulting in a WDM link configuration shown in Fig. 13(c). The spatially separated source wavelengths in this configuration can come from multiple single wavelength laser sources, or via demultiplexing of a comb laser source.

Although configured differently, the various types of WDM silicon photonic links largely employ the same set of components including laser sources, modulators, wavelength multiplexers and de-multiplexers, routing, and optical receivers. Depending upon the interconnection network architecture, the routing of the WDM link could involve passive silicon waveguides and multiple interlayer couplers or optical switches. For the scope of this paper, we focus on point-to-point interconnects between a source and a destination site on the same macrochip via waveguides and chip-to-chip couplers, and between macrochips via single mode fibers (SMF), in which case the routing components include waveguides, fiber couplers, SMF, and connectors.

The total power consumption of a WDM optical link includes the electrical power needed for modulator and receiver operation, WDM component tuning and wavelength locking circuitry, as well as the laser source electrical power required to generate sufficient optical power to operate inter-chip links within a macrochip and inter-macrochip links with a bit error rate (BER) no worse than 10^{-12}. To achieve the best energy efficiency necessitates component performance tradeoffs at the link level. Proper optimization is needed between the data rate and the laser power requirement due to the relatively low wall-plug efficiency (<10%), of WDM laser sources. A higher data rate is preferred to amortize the laser power and the WDM component wavelength tuning and control power. However, optical receiver sensitivity can degrade super-linearly with date rate due to structural limits for a given CMOS technology node. Since the loss of the optical channel is largely independent of the data rate, degraded receiver sensitivity directly translates to larger laser optical power, which may result in a reduction of the total link energy efficiency when the low laser WPE is factored in. After analyzing component and system-level tradeoffs [10], we selected a data rate of 16–20 Gbps as the goal for future macrochip links from an energy optimization perspective.

For each defined link configuration, we estimate the corresponding link power consumption based on demonstrated component performance results. These are presented in Table VI. On the transmitter side, several silicon modulator types are possible. However, the choice for ultra-efficient silicon photonic links is limited because the silicon modulator not only has to be compact, high speed, and low total capacitance, but also has to operate with low driver voltage-swings due to limits from transistor breakdown, interconnect dielectric breakdown, and compatibility with advanced CMOS technology nodes. Promising candidates include reverse-biased depletion ring modulators [33], [34], and Ge modulators using Franz-Keldysh effect [35], [36].

Using a cascode driver, a modulation energy efficiency of 80 fJ/bit including the driving circuits was demonstrated at a data rate of 12.5 Gbps for silicon ring modulator-based transmitters [37]. We believe comparable energy efficiencies can be achieved at 16–25 Gbps since a reverse biased depletion ring modulator behaves essentially as a capacitive load to its VLSI driver. For Ge FK absorption-modulator transmitters, a higher power consumption of ∼6 mW was demonstrated at a similar data rate [38]. The increased power was due to its leakage current and power consumption associated with the absorbed photocurrent. Upon improving fabrication processes, material quality and device structure, we expect the leakage current can be reduced to a negligible level and a modulation efficiency comparable to ring modulators can be achieved at 16–25 Gbps.

based WDM links, the laser source wavelength multiplexing loss can be reduced to ~0.5 dB based on numerical design simulation. It is very attractive from a link loss perspective, but it requires each laser source to have N times of the optical power sufficient to support one link, where N is the splitting ratio. It becomes impractical quickly when N scales up; thus for the future 8-wavelength 20 Gbps WDM link, we assume regular wavelength multiplexers will be used.

Another loss source is the optical modulator. The total power penalty of an optical modulator includes it's "ON" state loss, the modulation loss and eye closure penalty due to imperfect modulation. We demonstrated a ring modulator with ~8 dB total power penalty for 10 Gbps [24] and later 20 Gbps operation [49]. Ge FK modulators, on the other hand, showed slightly a higher total power penalty of ~10 dB [47] due to their relatively low extinction ratio. Although ring/disc modulators can potentially be improved with substantially lower total power penalty of <5 dB [50] employing a vertical junction design on a thinner SOI platform, we use a conservative 8 dB based on our measured device for future goal for consistency with other components.

Signal routing loss is slightly different for intra- and inter-module links. As described earlier, the intra-module routing includes OPxC coupling from/to the bridge to/from the routing wafer and the transport waveguide on the routing wafer, while the inter-module link would have additional fiber couplings (in and out) and a fiber transport. Using post-process techniques including top dielectric layer removal and back-side etch-pit mirror, a surface-normal 2 dB OPxC was demonstrated using grating couplers [32]. We expect such OPxC loss to be further improved to 1.5 dB or better when an advanced CMOS process with finer etch resolution is used to achieve better GC mode matching. Huge improvements in the manufacture of low-loss silicon waveguides have also been achieved. Using advanced photolithography and waveguide side wall smoothing techniques, submicron single mode silicon waveguides achieved a low loss of around 2~3 dB/cm [51]. By reducing the mode overlap with the side walls using wider waveguides with shallow etching, an ultra-low loss of 0.027 dB/cm (2.7 dB/m) has been achieved with deep-etching compatible with compact micro-rings for up to 1 m long passive waveguides [14]. An intra-module link may require a routing waveguide 10s of centimeters long corresponding to <3 dB waveguide loss. An inter-module link, on the other hand, could use shorter transport waveguides to route the signal from the bridge to the edge of the routing wafer, but it has to go through another fiber transport interface. The loss of SMF is on the order of 0.3 dB/km for the wavelength band of interest, hence negligible compared to other losses. Using double SOI substrates and an advanced CMOS process, a GC fiber coupler can be fabricated with loss as low as ~1 dB [52]. Considering additional transport waveguide loss on the routing wafer, we use an aggressive routing loss budget of 3 dB for the inter-module links.

Knowing the loss of the components and the receiver sensitivity, it is straightforward to calculate the laser optical power needed to support an optical link. The laser optical power present in the waveguide (after accounting for coupling losses) divided by its wall-plug efficiency (WPE) is the total laser electrical power for a wavelength channel. The details of such link power budget for all three link configurations shown in Fig. 13 are summarized in Table VI. A tunable or wavelength-locked laser WPE of 10% is assumed for total laser power calculation. Such lasers have recently been developed by several groups – with waveguide-coupled WPEs approaching 10% [53]–[56] and promising up to 20% in the near future.

VIII. Conclusion

An analysis of petascale systems suggests that relying on evolutionary trends in processor, memory and interconnect components will not lead to efficient systems with high connectivity and good Byte/Flop ratios. In this paper, we have analyzed a futuristic design based on highly optimized processor designs with specialized floating-point units that will lead to significant improvements in performance-power characteristics. By addressing other major users of power, namely memory and interconnection networks, even more power-efficient systems with reduced form factors can emerge. Furthermore, such systems *need not compromise* intra-system bandwidth as a default consequence of meeting energy efficiency targets, but in fact, if designed around the interconnect - can lead to more compact systems with lower end-to-end latency. We presented and motivated a system design with nodes consisting of multi-chip substrates with silicon photonic intra-node communications (the macrochip) and hybrid inter-node links (optical fibers with electrical switches) as candidates for power-efficient petascale systems for the 2020+ timeframe. Power efficiency and peak bandwidth estimates were provided showing an energy improvement larger than 6x over evolutionary systems.

A key requirement for such systems is a 1 pJ/bit optical link that can support distances as short as a few cm in silicon waveguides all the way to several km in single-mode optical fiber, while providing high bandwidth-density and compact waveguide-to-fiber connectivity. In this paper, we have shown that wavelength-division multiplexed single-mode silicon photonic links can meet this need. We have provided examples of each link component required for the high-density silicon photonic link demonstrated in a commercial, foundry-based 130 nm digital-CMOS SOI platform flip-chip integrated to circuits implemented in a 40 nm bulk CMOS process. We reviewed some of the recent link component demonstrations based on this technology including an 8-channel 100 Gbps tunable silicon photonic WDM ring-resonator-based multiplexed transmitter array and showed that high-fidelity transmission was possible for distances up to 40 km. We summarized recent progress in efficient, tunable silicon-assisted lasers as well as fiber-based link results and all-solid-state link experiments to support the system vision. Based on these demonstrated components, we presented a detailed link budget analysis establishing that a ~1 pJ/bit link supportive of the vision for photonically-interconnected many-chip modules was feasible. Significant system packaging, alignment, and assembly challenges still remain. Architectural investigations that prove application-level benefits are also needed. Nevertheless, the technology transition and commercialization prospects are promising.

ACKNOWLEDGMENT

The author would like to thank many current and former collaborators at Oracle Systems and Oracle Labs who have contributed to the work reviewed in this paper. They would also like to thank S. Radic and J. Ford of UCSD for preparation of the dispersion-compensated single-mode fiber spools. They would also like to thank the anonymous reviewers for their constructive comments on the paper and S. Djordjevic for editorial assistance.

REFERENCES

[1] P. Kogge, K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, K. Hill, J. Hiller, S. Karp, S. Keckler, D. Klein, R. Lucas, M. Richards, A. Scarpelli, S. Scott, A. Snarvely, T. Sterling, R. S. William, and K. Yelick, "Exascale computing study: Technology challenges in achieving exascale systems," Dept. Defense, United States Government, Washington, DC, USA, 2008.

[2] S. Borkar and A. Chien, "The future of microprocessors," *Commun. ACM*, vol. 54, no. 5, pp. 67–77, May 2011.

[3] S. Borkar, "How to stop interconnects from hindering the future of computing," in *Proc. IEEE Opt. Interconnects Conf.*, 2013, pp. 96–97.

[4] R. Drost, C. Forrest, B. Guenin, R. Ho, A. Krishnamoorthy, D. Cohen, J. Cunningham, B. Tourancheau, A. Zingher, A. Chow, G. Lauterbach, and I. Sutherland, " Challenges in building a flat-bandwidth memory hierarchy for a large-scale computer with proximity communication," in *Proc. 13th Symp. High Perform. Interconnects*, 2005, pp. 13–22.

[5] (2014). Top 500 supercomputer sites—Presented at the SC14 [Online]. Available: http://www.top500.org/lists/2013/11/

[6] (2014). The green 500 list—Benchmark results [Online]. Available: http://www.green500.org

[7] (2013). The graph 500 list—Benchmark results [Online]. Available: http://www.graph500.org/

[8] IU Computing Center, Private Communication, 2014.

[9] (2014). The green graph 500 list—Benchmark results [Online]. Available: http://www.green.graph500.org

[10] A. Krishnamoorthy, R. Ho, X. Zheng, H. Schwetman, J. Lexau, P. Koka, G. Li, I. Shubin, and J. Cunningham, "Computer systems based on silicon photonic interconnects," *Proc. IEEE*, vol. 97, no. 7, pp. 1337–1361, Jul. 2009.

[11] P. Koka, M. McCracken, H. Schwetman, C.-H. Chen, X. Zheng, R. Ho, K. Raj, and A. Krishnamoorthy, "A micro-architectural analysis of switched photonic multi-chip interconnects," in *Proc. 39th Annu. Int. Symp. Comput. Archit.*, 2012, pp. 153–164.

[12] C. Gunn, "CMOS photonics for high speed interconnects," *IEEE Micro*, vol. 26, no. 2, pp. 58–66, Mar./Apr. 2006.

[13] A. Mekis, S. Gloeckner, G. Masini, A. Narasimha, T. Pinguet, S. Sahni, and P. De Dobbelaere, "A grating-coupler-enabled CMOS photonics platform," *IEEE J. Sel. Topics Quantum Electron.*, vol. 17, no. 3, pp. 597–608, May/Jun. 2011.

[14] 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. 12035–12309, 2012.

[15] J. Yao, X. Zheng, G. Li, I. Shubin, H. Thacker, Y. Luo, K. Raj, J. Cunningham, and A. Krishnamoorthy, "Grating-coupler based low-loss optical interlayer coupling," in *Proc. 8th IEEE Int. Conf. Group IV Photon.*, 2011, pp. 383–385.

[16] J. E. Cunningham, I. Shubin, X. Zheng, T. Pinguet, A. Mekis, Y. Luo, H. Thacker, G. Li, J. Yao, K. Raj, and A. V. Krishnamoorthy, "Highly-efficient thermally-tuned resonant optical filters," *Opt. Exp.*, vol. 18, no. 18, pp. 19055–19063, 2010.

[17] I. Shubin, G. Li, X. Zheng, Y. Luo, H. Thacker, J. Yao, N. Park, A. V. Krishnamoorthy, and J. E. Cunningham, "Integration, processing and performance of low power thermally tunable CMOS-SOI WDM resonators," *Opt. Quantum Electron.*, vol. 44, pp. 1–16, 2012.

[18] H. Thacker, Y. Luo, J. Shi, I. Shubin, J. Lexau, X. Zheng, G. Li, J. Yao, J. Costa, T. Pinguet, A. Mekis, P. Dong, S. Liao, D. Feng, M. Asghari, R. Ho, K. Raj, J. Mitchell, A. Krishnamoorthy, and J. E. Cunningham, "Flip-chip integrated silicon photonic bridge chips for sub-picojoule per bit optical links," in *Proc. 60th Electron. Compon. Technol. Conf.*, 2010, pp. 240–246.

[19] F. Liu, D. Patil, J. Lexau, P. Amberg, M. Dayringer, J. Gainsley, H. Moghadam, X. Zheng, J. Cunningham, A. Krishnamoorthy, E. Alon, and R. Ho, "10-Gbps, 5.3-mW optical transmitter and receiver circuits in 40-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 47, no. 9, pp. 2049–2067, Sep. 2012.

[20] J. Cunningham, A. Krishnamoorthy, R. Ho, I. Shubin, H. Thacker, J. Lexau, D. Lee, D. Feng, E. Chow, Y. Luo, X. Zheng, G. Li, J. Yao, T. Pinguet, K. Raj, M. Asghari, and J. Mitchell, "Integration and packaging of a macrochip with silicon nanophotonic links," *IEEE J. Sel. Topics Quantum Electron.*, vol. 17, no. 3, pp. 546–558, May/Jun. 2011.

[21] G. Li, A. Krishnamoorthy, I. Shubin, J. Yao, Y. Luo, H. Thacker, X. Zheng, K. Raj, and J. Cunningham, "Ring resonator modulators in silicon for interchip photonic links," *IEEE J. Sel. Topics Quantum Electron.*, vol. 19, no. 6, pp. 1–19, Nov./Dec. 2013.

[22] G. Li, X. Zheng, J. Lexau, Y. Luo, H. Thacker, P. Dong, S. Liao, D. Feng, D. Zheng, R. Shafiiha, M. Asghari, J. Yao, J. Shi, P. Amberg, N. Pinckney, K. Raj, R. Ho, J. Cunningham, and A. Krishnamoorthy, "Ultralow-power high-performance Si photonic transmitter," in *Proc. Opt. Fiber Commun.*, 2010, pp. 1–3.

[23] X. Zheng, D. Patil, J. Lexau, F. Liu, G. Li, H. Thacker, Y. Luo, I. Shubin, J. Li, J. Yao, P. Dong, D. Feng, M. Asghari, T. Pinguet, A. Mekis, P. Amberg, M. Dayringer, J. Gainsley, H. F. Moghadam, E. Alon, K. Raj, R. Ho, J. E. Cunningham, and A. V. Krishnamoorthy, "Ultra-efficient 10 Gb/s hybrid integrated silicon photonic transmitter and receiver," *Opt. Exp.*, vol. 19, no. 6, pp. 5172–5186, 2011.

[24] X. Zheng, F. Liu, J. Lexau, D. Patil, G. Li, Y. Luo, H. Thacker, I. Shubin, J. Yao, K. Raj, R. Ho, J. Cunningham, and A. Krishnamoorthy, "Ultralow power 80 Gb/s arrayed CMOS silicon photonic transceivers for WDM optical links," *J. Lightw. Technol.*, vol. 30, no. 4, pp. 641–650, Feb. 2012.

[25] X. Zheng, I. Shubin, G. Li, T. Pinguet, A. Mekis, J. Yao, H. Thacker, Y. Luo, J. Costa, K. Raj, J. E. Cunningham, and A. V. Krishnamoorthy, "A tunable 1 × 4 silicon CMOS photonic wavelength multiplexer/demultiplexer for dense optical interconnects," *Opt. Exp.*, vol. 18, no. 5, pp. 5151–5160, 2010.

[26] P. Amberg, E. Chang, F. Liu, J. Lexau, X. Zheng, G. Li, I. Shubin, J. Cunningham, A. Krishnamoorthy, and R. Ho, "A sub-400 fJ/bit thermal tuner for optical resonant ring modulators in 40 nm CMOS," in *Proc. IEEE Asian Solid State Circuits Conf.*, 2012, pp. 29–32.

[27] X. Zheng, E. Chang, P. Amberg, I. Shubin, J. Lexau, F. Liu, H. Thacker, S. S. Djordjevic, S. Lin, Y. Luo, J. Yao, J.-H. Lee, K. Raj, R. Ho, J. E. Cunningham, and A. V. Krishnamoorthy, "A high-speed, tunable silicon photonic ring modulator integrated with ultra-efficient active wavelength control," *Opt. Exp.*, vol. 22, no. 10, pp. 12628–12633, 2014.

[28] X. Zheng, Y. Luo, J. Lexau, F. Liu, G. Li, H. Thacker, I. Shubin, J. Yao, R. Ho, J. Cunningham, and A. Krishnamoorthy, "2-pJ/bit (on-chip) 10-Gb/s digital CMOS silicon photonic link," *IEEE Photon. Technol. Lett.*, vol. 24, no. 14, pp. 1260–1262, Jul. 2012.

[29] X. Zheng, S. Lin, Y. Luo, J. Yao, G. Li, S. Djordjevic, J.-H. Lee, H. Thacker, I. Shubin, K. Raj, J. Cunningham, and A. Krishnamoorthy, "Efficient WDM laser sources towards terabyte/s silicon photonic interconnects," *J. Lightw. Technol.*, vol. 31, no. 24, pp. 4142–4154, Dec. 2013.

[30] S. Lin, S. S. Djordjevic, J. E. Cunningham, I. Shubin, Y. Luo, J. Yao, G. Li, H. Thacker, J.-H. Lee, K. Raj, X. Zheng, and A. V. Krishnamoorthy, "Vertical-coupled high-efficiency tunable III-V- CMOS SOI hybrid external-cavity laser," *Opt. Exp.*, vol. 21, no. 26, pp. 32425–32431, 2013.

[31] A. Krishnamoorthy, X. Zheng, G. Li, J. Yao, T. Pinguet, A. Mekis, H. Thacker, I. Shubin, Y. Luo, K. Raj, and J. Cunningham, "Exploiting CMOS manufacturing to reduce tuning requirements for resonant optical devices," *IEEE Photon. J.*, vol. 3, no. 3, pp. 567–579, Jun. 2011.

[32] J. Yao, X. Zheng, I. Shubin, G. Li, H. Thacker, Y. Luo, L. Jin-Hyoung, K. Raj, J. E. Cunningham, and A. V. Krishnamoorthy, "Optical interlayer coupling design for optical interconnects based on mirror enhanced grating couplers," in *Proc. IEEE Opt. Interconnects Conf.*, 2012, pp. 27–28.

[33] M. R. Watts, W. A. Zortman, D. C. Trotter, R. W. Young, and A. L. Lentine, "Vertical junction silicon microdisk modulators and switches," *Opt. Exp.*, vol. 19, no. 22, pp. 21989–22003, 2011.

[34] 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," in *Proc. Opt. Fiber Commun. Conf.*, 2013, pp. 1–3.

[35] J. Liu, M. Beals, A. Pomerene, S. Bernardis, S. Sun, J. Cheng, L. C. Kimerling, and J. Michel, "Waveguide-integrated ultra-low-energy GeSi electro-absorption modulators," *Nature Photon.*, vol. 2, pp. 433–437, 2008.

[36] A. E.-J. Lim, T.-Y. Liow, F. Qing, N. Duan, L. Ding, M. Yu, G.-Q. Lo, and D.-L. Kwong, "Novel evanescent-coupled germanium electro-absorption modulator featuring monolithic integration with germanium p-i-n photodetector," *Opt. Exp.*, vol. 19, no. 6, pp. 5040–5046, 2011.

[37] X. Zheng, E. Chang, I. Shubin, G. Li, Y. Luo, J. Yao, H. Thacker, J.-H. Lee, J. Lexau, F. Liu, P. Amberg, K. Raj, Ho, J. Cunningham, and A. Krishnamoorthy, "A 33 mW 100 Gbps CMOS silicon photonic WDM transmitter using off-chip laser sources," in *Proc. Opt. Fiber Commun. Conf. (OFC/NFOEC)*, paper PDP5C.9, pp. 1–3, 2013.

[38] A. V. 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] D. Feng, S. Liao, P. Dong, N. Feng, H. Liang, D. Zheng, C. Kung, J. Fong, R. Shafiiha, J. E. Cunningham, A. V. Krishnamoorthy, and M. Asghari, "High-speed Ge photodetector monolithically integrated with large cross-section silicon-on-insulator waveguide," *Appl. Phys. Lett.*, vol. 95, pp. 261105-1–261105-3, 2009.

[40] C. T. DeRose, D. C. Trotter, W. A. Zortman, A. L. Starbuck, M. Fisher, M. R. Watts, and P. S. Davids, "Ultra compact 45 GHz CMOS compatible germanium waveguide photodiode with low dark current," *Opt. Exp.*, vol. 19, no. 25, pp. 24897–24904, 2011.

[41] L. Vivien, A. Polzer, D. Marris-Morini, J. Osmond, J. Michel Hartmann, P. Crozat, E. Cassan, C. Kopp, H. Zimmermann, and J. Fédéli, "Zero-bias 40 Gbit/s germanium waveguide photodetector on silicon," *Opt. Exp.*, vol. 20, no. 2, pp. 1096–1101, 2012.

[42] W. Zortman, A. Lentine, D. Trotter, and M. Watts, "Bit-error rate monitoring for active wavelength control of resonant modulators," *IEEE Micro*, vol. 33, no. 1, pp. 42–52, Jan-Feb, 2013.

[43] K. Padmaraju, D. Logan, J. Ackert, A. Knights, and K. Bergman, "Microring resonance stabilization using thermal dithering," in *Proc. IEEE Opt. Interconnects Conf.*, 2013, pp. 58–59.

[44] J. Cox, D. Trotter, and A. Starbuck, "Integrated control of silicon-photonic micro-resonator wavelength via balanced homodyne locking," in *Proc. IEEE Opt. Interconnects Conf.*, 2013, pp. 52–53.

[45] P. Dong, W. Qian, H. Liang, R. Shafiiha, D. Feng, G. Li, J. E. Cunningham, A. V. Krishnamoorthy, and M. Asghari, "Thermally tunable silicon racetrack resonators with ultralow tuning power," *Opt. Exp.*, vol. 18, no. 19, pp. 20298–20304, 2010.

[46] D. Feng, S. Liao, H. Liang, J. Fong, B. Bijlani, R. Shafiiha, B. J. Luff, Y. Luo, J. Cunningham, A. V. Krishnamoorthy, and M. Asghari, "High speed GeSi electro-absorption modulator at 1550 nm wavelength on SOI waveguide," *Opt. Exp.*, vol. 20, no. 20, pp. 22224–22232, 2012.

[47] D. Feng, W. Qian, H. Liang, C.-C. Kung, J. Fong, B. Luff, and M. Asghari, "Novel fabrication tolerant flat-top demultiplexers based on etched diffraction gratings in SOI," in *Proc. 5th IEEE Int. Conf. Group IV Photon.*, 2008, pp. 386–388.

[48] B. Little, J. Foresi, G. Steinmeyer, E. Thoen, S. Chu, H. Haus, E. Ippen, L. Kimerling, and W. Greene, "Ultra-compact Si-SiO2 microring resonator optical channel dropping filters," *IEEE Photon. Technol. Lett.*, vol. 10, no. 4, pp. 549–551, Apr. 1998.

[49] G. Li, X. Zheng, J. Yao, H. Thacker, I. Shubin, Y. Luo, K. Raj, J. E. Cunningham, and A. V. Krishnamoorthy, "25 Gb/s 1V-driving CMOS ring modulator with integrated thermal tuning," *Opt. Exp.*, vol. 19, no. 21, pp. 20435–20443, 2011.

[50] E. Timurdogan, C. Sorace-Agaskar, A. Biberman, and M. Watts, "Vertical junction silicon microdisk modulators at 25 Gb/s," in *Proc. Opt. Fiber Commun. Conf. (OFC/NFOEC)*, paper OTh3H.2, pp. 1-3, 2013.

[51] W. Bogaerts, R. Baets, P. Dumon, V. Wiaux, S. Beckx, D. Taillaert, B. Luyssaert, J. Van Campenhout, P. Bienstman, and D. Van Thourhout, "Nanophotonic waveguides in silicon-on-insulator fabricated with CMOS technology," *J. Lightw. Technol.*, vol. 23, no. 1, pp. 401–412, Jan. 2005.

[52] C. Baudot, D. Dutartre, A. Souhaite, N. Vulliet, A. Jones, M. Ries, A. Mekis, L. Verslegers, P. Sun, Y. Chi, S. Cremer, O. Gourhant, D. Benoit, G. Courgoulet, C. Perrot, L. Broussous, T. Pinguet, J. Siniviant, and F. Boeuf, "Low cost 300 mm double-SOI substrate for low insertion loss 1D & 2D grating couplers," in *Proc. IEEE 11th Int. Conf. Group IV Photon.*, 2014, pp. 137–138.

[53] A. J. Zilkie, P. Seddighian, B. J. Bijlani, W. Qian, D. C. Lee, S. Fathololoumi, J. Fong, R. Shafiiha, D. Feng, B. J. Luff, X. Zheng, J. E. Cunningham, A. V. Krishnamoorthy, and M. Asghari, "Power-efficient III-V/silicon external cavity DBR lasers," *Opt. Exp.*, vol. 20, no. 21, pp. 23456–23462, 2012.

[54] K. Nemoto, T. Kita, and H. Yamada, "Narrow spectral linewidth wavelength tunable laser with Si photonic-wire waveguide ring resonators," in *Proc. IEEE 9th Int. Conf. Group IV Photon.*, 2012, pp. 216–218.

[55] B. Koch, E. Norberg, B. Kim, J. Hutchinson, J.-H. Shin, G. Fish, and A. Fang, "Integrated silicon photonic laser sources for telecom and datacom," in *Proc. Opt. Fiber Commun. Conf. (OFC/NFOEC)*, paper PDP5C.8, pp. 1–3, 2013.

[56] J. H. Lee, I. Shubin, J. Yao, J. Bickford, Y. Luo, S. Lin, S. S. Djordjevic, H. D. Thacker, J. E. Cunningham, K. Raj, X. Zheng, and A. V. Krishnamoorthy, "High power and widely tunable Si hybrid external-cavity laser for power efficient Si photonics WDM links," *Opt. Exp.*, vol. 22, no. 7, pp. 7678–7685, 2014.

Authors' biographies not available at the time of publication.