INVITED PAPER

Computer Systems Based on Silicon Photonic Interconnects

A proposed supercomputer-on-a-chip with optical interconnections between processing elements will require development of new lower-energy optical components and new circuit architectures that match electrical datapaths to complementary optical interfaces.

By Ashok V. Krishnamoorthy, *Member IEEE*, Ron Ho, *Senior Member IEEE*, Xuezhe Zheng, *Senior Member IEEE*, Herb Schwetman, *Member IEEE*, Jon Lexau, *Member IEEE*, Pranay Koka, GuoLiang Li, *Senior Member IEEE*, Ivan Shubin, *Member IEEE*, and John E. Cunningham

ABSTRACT | We present a computing microsystem that uniquely leverages the bandwidth, density, and latency advantages of silicon photonic interconnect to enable highly compact supercomputer-scale systems. We describe and justify single-node and multinode systems interconnected with wavelength-routed optical links, quantify their benefits vis-à-vis electrically connected systems, analyze the constituent optical component and system requirements, and provide an overview of the critical technologies needed to fulfill this system vision. This vision calls for more than a hundredfold reduction in energy to communicate an optical bit of information. We explore the power dissipation of a photonic link, suggest a roadmap to lower the energy-per-bit of silicon photonic interconnects, and identify the challenges that will be faced by device and circuit designers towards this goal.

KEYWORDS | Integrated optics; modulation; optical communication; photodetectors; supercomputers; wavelength-division multiplexing

Manuscript received February 6, 2009; revised March 15, 2009. First published June 10, 2009; current version published June 12, 2009. This work was supported in part by the Defense Advanced Research Projects Agency under HR0011-08-09-0001 and W911NF-07-1-0529.

A. V. Krishnamoorthy, **X. Zheng**, **G. Li**, **I. Shubin**, and **J. E. Cunningham** are with Sun Microsystems, San Diego, CA 92121 USA (e-mail: ashok.k@sun.com; xuezhe.zheng@sun.com; guoliang.li@sun.com; ivan.shubin@sun.com; john.cunningham@sun.com).

R. Ho and **J. Lexau** are with Sun Microsystems, Menlo Park, CA 94025 USA (e-mail: ron.ho@sun.com; jon.lexau@sun.com).

H. Schwetman and **P. Koka** are with Sun Microsystems, Austin, TX 78727 USA (e-mail: herb.schewetman@sun.com; pranay.koka@sun.com).

Digital Object Identifier: 10.1109/JPROC.2009.2020712

I. BACKGROUND

It has been 25 years since the concept for optical interconnections for very-large scale integration (VLSI) systems was proposed [1]. During this period, researchers made considerable progress understanding the benefits of optical interconnects [2], the photonics components themselves, and the integration of the photonic components with VLSI circuits [3]. Much of the initial work focused on switching and routing systems [4] and motivated the hybrid integration of VLSI with surface-normal detectors, modulators, and vertical cavity lasers. Early this century saw the first commercial-grade transceiver products and system insertions to exploit this integration for switching and computing systems [5], [6]. But in spite of this progress, the full inclusion of photonic component manufacturing into mainstream complementary metal–oxide–semiconductor (CMOS) foundries remained elusive.

The key to unlocking this potential began with Soref *et al.* in their seminal investigation of waveguides and silicon's electro-optic effect [7], [8]. Subsequent work at the University of Surrey [9] and at Bookham [10] successfully used silicon microelectronics manufacturing technologies to create a family of active and passive optical structures that could be integrated with high-speed laser sources and detectors. Work continued in parallel to improve the performance and to reduce the size of silicon-based wavelength filter devices [11]. The original goals of these efforts were to reduce the fabrication cost of the optoelectronic components via integration and not necessarily to produce wavelength-division multiplexed (WDM) optical interconnections for the silicon industry.

Nevertheless, the merger of these two efforts was inevitable, as suggested by Soref [12], and the new century has seen a significant effort to create active low-loss high-speed devices that can be manufactured in silicon and other group IV materials [13]–[15].

Another key breakthrough was the creation of a fully CMOS-compatible silicon photonics process. This required a mainstream foundry, capable of building a modern processor chip, to produce photonic components colocated with silicon transistors in a unified process flow. After many years of development, this was demonstrated by the Luxtera-Freescale partnership [16]. Concurrently, the development of high-speed modulators proved that silicon was capable of 10 Gbps and higher modulation speeds [17]–[19] and heralded the era of silicon photonic interconnects.

With such demonstrations of CMOS/photonic co-integration, researchers have now focused on the inclusion of silicon photonic interconnects into multicore computing systems, particularly at the intrachip scale [20]–[25]. These ideas exploited the large aggregate bandwidth and the high density of WDM optical interconnects available with the latest silicon photonic components. In this paper, we will describe another novel concept for a computing microsystem based on silicon photonic interconnect: the macrochip. It uniquely leverages the bandwidth, density, and latency advantages of silicon photonic interconnects to enable highly compact supercomputer-scale systems; like other system concepts, it requires ultra-low-energy photonic interconnects. In this paper, we will explore the power dissipation of a photonic link and suggest a ten-year roadmap for the achievable energy-per-bit of silicon photonic interconnects. We will discuss the challenges that will be faced by device and circuit designers along the way.

II. OPTICS IN COMPUTING SYSTEMS

"Moore's law" is what Carver Mead dubbed Gordon Moore's now-famous 1965 extrapolation of transistor density scaling [26]. Moore's idea was simple: when accounting for device yield, the number of integrated devices on a chip that minimizes system cost will grow geometrically over time. Designers have very effectively converted this increase in transistors into increased performance; for instance, integer benchmarks show an astonishing 35% cumulative annual growth rate for the past 20 years [27].

The close relationship between die transistor count and system performance arises because improving system performance depends on either raising clock frequencies or on increasing instruction, thread, and program parallelism. While clock rate speedup has reached power and complexity limits, increases in parallelism require increases in the number of transistors to enable heavily cached, speculative, multicore and/or multithreaded architectures. However, increasing transistor counts by simply assembling multiple chips together on a printed circuit board does not efficiently improve performance because off-chip wires between chips present fundamental performance bottlenecks, even when used with high-speed and high-power serializer-deserializer (SerDes) circuits. Because on-chip wires offer unmatched bandwidth density, what designers really want is a large continuous piece of silicon on which to place execution engines, memories, program sequencers, and the wires to connect them together.

Unfortunately, design complexity, time-to-market, mask costs, and degraded yield all dramatically worsen with large monolithic chips. These largely economic factors constrain the ability of designers to create single large pieces of silicon. Efforts to break through these Moore's law limits include using capacitive coupling to assemble multichip systems in grid-like aggregations with overlapping "bridge" chips. In such systems, chip-to-chip I/O is mediated by tiny chipface-to-chipface channels, effectively extending highly dense on-chip wires across a chip-to-chip gap, and obviating the need for large-pitch solder and printed circuit board wires. This kind of system provides a logically continuous piece of silicon. Most importantly, because chip-to-chip connections are formed through physical proximity, defective chips can be reworked during system assembly, maintaining a very low system cost. Simple defect and yield models show a 10× cost benefit of rematable smaller chips over single monolithic chips at large die sizes [28].

Using capacitive solderless interconnects to build systems of moderate complexity offers many benefits: high performance from aggregated chip area and enormous on-chip wire bisection bandwidth, low energy from efficient capacitive I/O, and a low total system cost from reworkability. However, by running all chip-to-chip I/O over on-chip wires, a large multichip system would suffer from high message latency: standard on-chip wires propagate at only 5% or 10% the speed of light [29]. This would limit the data to scalability of such a chip grid to a small number of chips, and hence will also limit its performance benefits.

To recap, then: system performance, on critical code benchmarks such as gigaupdates per second (GUPS) and global fast Fourier transform (FFT), comes from the large-scale integration of memory with multiple threads and cores (see Section VI-B). Placing everything on a single monolithic piece of silicon leads to unacceptable design and yield risks, while increasing mask and complexity costs to prohibitive levels. Soldering multiple chips on a printed circuit board (PCB) leads to reduced performance due to chip-to-chip I/O bottlenecks, and soldering them in a multichip package also leads to issues of "known-good die" and dramatically lowered yield. Placing together multiple chips in a reworkable proximity communications grid provides high performance, high yield, and low system cost but at the cost of longer latency across the system.

Optical interconnect provides a potential solution. It enables large aggregations of chips, giving high system performance. True speed-of-light communication offers

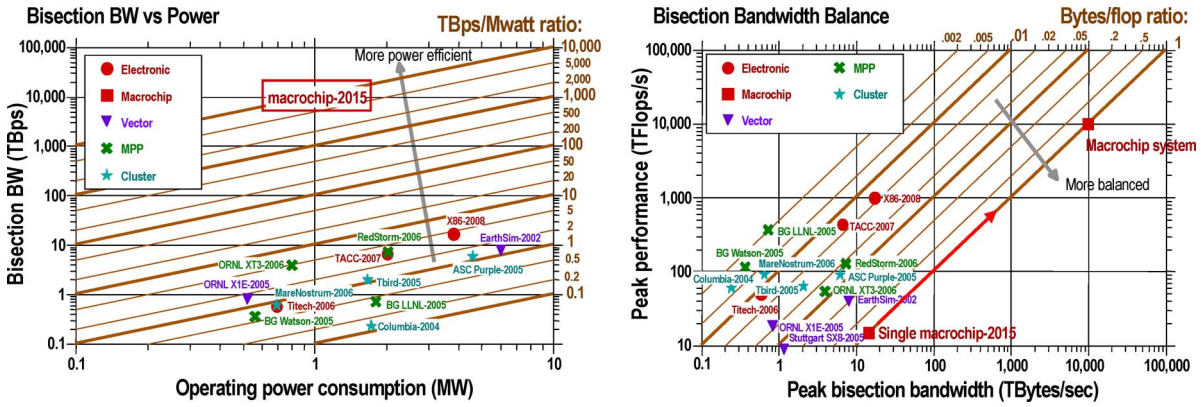


Fig. 1. ***Bisection BW of representative supercomputing systems versus system power and performance. Deployed bisection bandwidth is typically under 10 GBps per kilowatt of system power and must be improved by several orders of magnitude to obtain "balanced" systems with O(1) bytes of bisection bandwidth per flop. Macrochip-based systems with ultra-low-power interconnects offer the promise of breakthrough system bandwidth per watt and per flop.***

reduced latency across the multiple-chip system. Using optical proximity communication [30] to allow chips to communicate without a soldered connection enables seamless reworkability for high system yield. With WDM optics, it offers unmatched bandwidth density. The purpose of this paper is to motivate and introduce microsystem architectures that take advantage of new interconnect technologies to create a collection of tightly connected chips that can exhibit large bisection bandwidths, balance communication and computation, and produce orders of magnitude improvement in performance per watt metrics over existing systems (see Fig. 1).

A. The "Macrochip"

The "macrochip" is a logically contiguous piece of photonically interconnected silicon integrating multicore and multithreaded processors, a system-wide interconnect, and dense memories; it offers unprecedented computational density, energy efficiency, bisection bandwidth, and reduced message latencies. Optical proximity communication connects the processor cores on different sites to the optical layers that route the optical links; these links enable low-latency WDM optical links between sites. This interconnection network uses silicon waveguides and vertical silicon couplers to achieve point-to-point nonblocking links for every site on the macrochip. This network exploits the best features of silicon photonic technology (low latency, high density, long reach), while avoiding its weaknesses, and allows system designers to mix, match, and replace processor or memory die in a modular fashion. This paper will review the physical architecture of the macrochip and its routing network and validate the benefits that derive from the intrachip and interchip optical interconnects. It will also define the energy, density, and performance requirements for the macrochip's silicon photonic components, along with the issues and challenges faced by the silicon circuit and photonic device designers to reach these goals.

Section III will describe the macrochip architecture, specifying its constituent parts. Section IV will present the WDM point-to-point optical network and routing framework that provides low power, high bandwidth, and high-density communication between processor cores. This section will review how the WDM network provides the highest bisection bandwidth for a fixed number of transmitters and receivers, yet is transparent to data rate and communication protocols. Section V will review the physical structure and packaging of the macrochip and show how the various chips within the macrochip are connected and aligned.

Link-level, network-level, and system-level benefits will be analyzed in Section VI, comparing the macrochip with electronic implementations and showing up to a 40× advantage in performance and power efficiency at the system level. For the comparison, we will assume aggressive electronic implementations using capacitive proximity communication [31] to achieve a similar bandwidth in a similar form factor.

Section VII will examine the energy-per-bit of a silicon photonic link and detail its power budget. It will define an aggressive power dissipation roadmap for the link over the next decade and discuss the challenges related to drivers, receivers, modulators, detectors, WDM components, and tuning that will be faced by silicon circuit and photonic component designers. A brief summary will follow in Section VIII.

III. ARCHITECTURE OF THE SILICON PHOTONIC MICROSYSTEM

A macrochip enables vast amounts of processing and system interconnect to be integrated into a single node,

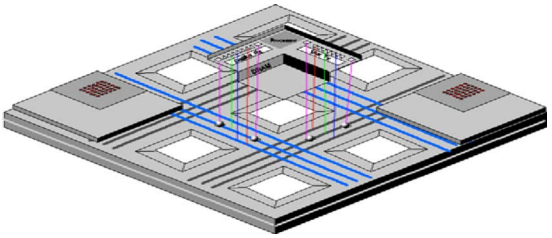


Fig. 2. *A 3 × 3 macrochip. Each site contains DRAM, processors, and photonic bridge chips.*

offering breakthrough performance for a given power and floor space. The macrochip, as described in the following, uses lateral (unfolded) packaging (Fig. 2). This lateral topology exploits the long "reach" of optics by amortizing the signaling power over longer distances and allows "fat" compute nodes that enable rich, high-degree interconnected topologies (such as all-to-all connections) even when scaling up to a multinode supercomputer. A nonblocking, point-to-point WDM routing network used in the macrochip has superior performance and no setup delays when compared to an electrically controlled packet switched network of the same bandwidth. This improvement is particularly evident as the loading of the network goes up and also as the message size goes down. It further simplifies the control of the network and eliminates the resulting power required for network resource arbitration. Further, the static WDM nonblocking network topology described in this paper provides efficient transport for small messages (64 B or less), an important characteristic for supporting shared memory machines, and also favors embedded machines where performance-per-watt on specific high-performance computing (HPC) challenge metrics must be maximized.

A macrochip-based system, optimized for GUPS/watt and global FFT/watt, interconnects message-passing multiprocessors and custom high-bandwidth DRAM chips with optics. It is based on a silicon-on-insulator (SOI) platform that packages and aggregates CPUs, memory, silicon photonics, and fiber interfaces. The requirement for an SOI platform derives from the need for a buried oxide for light confinement in the silicon optical waveguides, although photonics-capable bulk silicon processes that may relieve this requirement are under investigation [32]. A canonical system can be as small as a single macrochip or combine more than a thousand macrochips tied together with a dense fully connected fiber network. In this work we describe a particular configuration of a large-scale system enabled by advances in silicon photonics and by optimistic projections of memories and processors. One can imagine many other possible configurations; we use this one as an "attention-focuser" to explore architectural and system implications of optical and electrical technology choices.

A. The Macrochip Architecture

The logical architecture of the macrochip is based on an 8 × 8 array of sites, where each site has a four-core processor and 8 GB of DRAM. The ample amount of DRAM per site leads naturally to a wafer-size macrochip implemented on an SOI platform that packages CPUs, memory, silicon photonics, and fiber interfaces (Fig. 3). In more detail, a macrochip contains 64 sites (or super cores) in an 8 × 8 matrix, with each site containing a 400 mm^2 DRAM chip. A bridge chip is mounted face down over the DRAM chip. The bridge chip contains a processor and a system interface and communicates with the DRAM chip using electrical proximity communication and to waveguides using optical proximity communication. The 64 sites are interconnected by a static point-to-point 8 × 8 WDM network described below. The estimated power (assuming a forward-looking implementation on a 22 nm node) is 187 W for processors and DRAM and 26 W for the optical interconnect. Of the 187 W of electrical power, 15 W is required for the electronic proximity interface between the processors and DRAM. This yields a total of 41 W for all interconnect to memory.

Each site is comprised of four 40 Gflop cores, 8 GB of DRAM memory, and a system interface. The system interface connects the four cores to their DRAM slice and has 64 links, 10 GBps each (in + out), connected to the 64 sites in the processor. This yields 640 GBps (in + out) of aggregate network bandwidth per site, and a bisection bandwidth of 10 TBps. To match the 640 GBps of network port bandwidth, each site has 640 GBps of bandwidth (in + out) to the system interface. The DRAM has a 2 ns access time and a 64 B wide interface—this requires 20 banks per slice. External I/O is connected via an additional pair of optical waveguides, yielding 40 GBps (in +

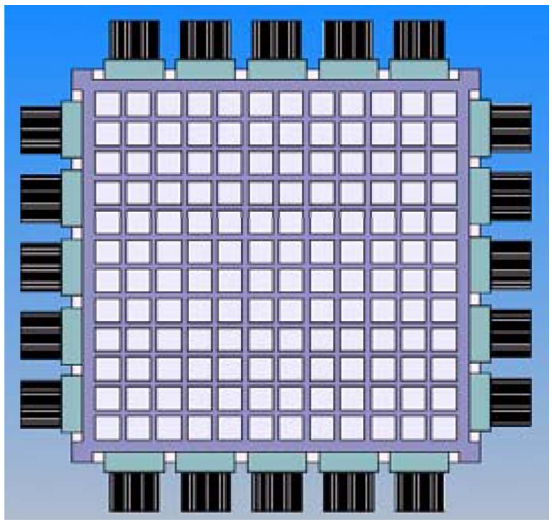


Fig. 3. *A wafer-sized macrochip contains multiple sites and fibers attached along the edges for I/O.*

Table 1 Nominal Compute Configuration for One Site in an 8 × 8 Macrochip in a 22 nm 2015 Technology

Processor chip						DRAM chip				Memory access	On-chip data network		Off-chip IO
Clock rate	Cores	Flops/ core	Total flops	Size	Power	Capacity	IO	Size	Power	On-chip memory BW	Single chip IO	Bisection BW	Inter- face
(GHz)	(#)	(GF)	(TF)	(mm^2)	(W)	(GB)	(GBps)	(mm^2)	(W)	(TBps)	(GBps)	(TBps)	(TBps)
10	4	160	10	8	1.32	8	640	400	1.6	40.96	640	10.24	2.56

out) of I/O bandwidth. These waveguides are routed from the system interfaces at each site or supercore-to-fiber connectors at the edges of the DRAM/processor slices. The cores are relatively simple four-threaded processors supporting two double precision multiply add operations per cycle; they are optimized to run at 10 GHz in 22 nm technology, with an estimated power of 330 mW. We assume an aggressive, custom-designed 22 nm DRAM, optimized for bandwidth and power. An open bitline DRAM array with 50% logic overhead packs 8 GB of capacity in a 400 mm^2 chip. Heavily banking the memory, as done in today's reduced latency DRAM, can potentially offer a total memory bandwidth of 640 GBps per slice and a cycle time and latency of 2 ns [compared to 15 ns in today's reduced-latency dynamic random-access memory (RLDRAM)].

This system configuration assumes the existence of several technologies, including 10 GHz processors and 2 ns access time DRAMs. Large and complex processors today run at speeds up to 5 GHz, but this doubling of clock frequency may not happen as technologies scale, due to tight power constraints and because native transistor speeds (for example, measured in constant fanout inverter delays) are no longer scaling directly with technology. However, a reduced-complexity core, specially designed for a specific workload, such as FFT or cross-memory bit operations (measured in GUPS), might efficiently run at clock rates much higher than those preferred for complex processors. Our macrochip concept is designed to not limit system performance in that case. Similarly, DRAMs with very low access times generally pay a much higher area and energy cost to support low-density bitlines and the wide routing channels required for heavily banked arrays. So, such a fast DRAM, while technologically feasible, will fall outside the economics of volume commercial parts and hence command a higher price.

The system interfaces support both fine grained message passing (for messages as small as 16 B) and shared memory across the entire machine. Message passing support includes hardware facilities for constructing and sending messages in user space using only a few instructions; it also includes efficient polling mechanisms that can be used in user space with very few instructions. Shared memory is supported to improve the programmability (and productivity) of the macrochip. Coding studies by Sun have demonstrated 3× to 10× reductions in code size when a shared memory coding style is used. Similar results were obtained for other codes [33]. A variety of shared memory models are made possible by the stacked macrochip. In the simplest shared memory paradigm, a portion of local DRAM is private and cacheable to a single core, with all other DRAM being noncacheable but directly accessible via load/store instructions. This avoids complex cache coherence hardware support while providing strong support for emerging programming styles [34]. More complex shared memory implementations are possible that enforce cache coherence across all 256 cores or support some form of transactional coherence [35], [36].

Using the technology assumptions discussed above, the macrochip supports 512 GB of memory, sufficient for large FFT problem sets. We assume an extra metal layer in the DRAM for interbank routing and electrical proximity communication for low-energy and low-latency DRAM I/O. For estimating power, we take the average power from industry data sheets, increase it by 40% for maximum power, scale it to 2015, and add extra routing power to connect together all of the banks. Table 1 shows a nominal configuration of the macrochip used for analysis later in the paper.

Each multithreaded core can execute two fused multiply add double-precision floating-point operations per cycle. This leads to 160 GFlops per processor, or 80 GFlops for FFT (which does not use a fused multiply-add operation). The processor supports fine-grained 16 B messages, useful both for updates and for FFT transpose. It employs a 256 KB L1 data cache to support matrix blocking for FFT but dispenses with an L2 cache because the DRAM main memory is very close. We estimate power per core by scaling the maximum power of today's UltraSPARC processor cores from Sun Microsystems to 2015.

This logical structure could also support a smaller macrochip with reduced memory capacity and alternative packaging options. Other processor and memory technologies may also be considered. Embedded DRAM offers very low latency (1.5 ns demonstrated in 2007) but at a 5× density disadvantage relative to DRAMs and with higher power, for lower overall performance per watt. Other technologies, such as SOI-based zero capacitor random access memory (Z-RAM) or thyristor-based random access memory (T-RAM), are speculative alternatives.

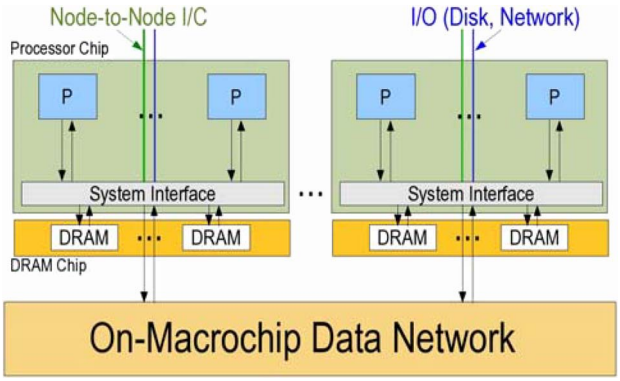


Fig. 4. ***Each site connects to an on-macrochip data network and two off-macrochip data networks.***

B. Optical Proximity Communication

An on-macrochip data network (Fig. 4) connects all of the CPU/DRAM sites by carrying messages passed between processors; the processors can pool their local memory together into a single shared address space and can employ various coherence schemes to maintain consistency in their local caches. This network carries 640 GBps (in plus out) at each processor and is organized as an optical point-to-point connection from each processor to every other processor on the macrochip.

Optical proximity communication (OPxC) couples optical signals between silicon chips placed face-to-face. This may be accomplished by collecting the light from the waveguide in the first chip, bending the light out of the plane of the first chip, coupling across the chip gap, then guiding the light into the waveguide of the second chip. The coupling may be accomplished with waveguide gratings or simple mirrored surfaces. Multiple-hop optical proximity communication between SOI waveguides on cascaded chips [37] has recently been achieved with passive alignment using a ball-in-pit alignment method (see Section V). The chips were passively and rematebly aligned with balls in pits in an experimental package that achieved a 4 dB loss per hop; this was 1 dB larger than the optimized efficiency achieved with active alignment using the identical chips. Bit error rate measurements versus received power of 10 Gbps optical data transmitted across the optical proximity hop showed no measurable penalty and no error floors, confirming that the optical loss associated with packaged OPxC appeared as a static attenuator of the optical signal.

IV. PHOTONIC NETWORK FOR THE MACROCHIP

The objective of the photonic network is to provide low power, high bandwidth, and high-density communication between cores. Although a number of different network topologies may be possible for core-to-core routing on the macrochip, the on-macrochip photonics network described here uses a point-to-point topology with static WDM routing. The point-to-point network offers the lowest optical loss between any two points. It also provides the highest bisection bandwidth for a fixed number of transmitters and receivers. The network is transparent to data rate and communication protocols and exploits the best features of optical technology (low latency, high density, long reach) while avoiding its weaknesses (lack of buffering). This is more efficient than broadcast, arbitrated mesh, or carrier-sense multiple-access/collision detect (i.e., Ethernet) networks because it minimizes optical power loss by avoiding splitters and reducing the number of optical components in a link, and can therefore drastically reduce the optical power requirements. Point-to-point topologies are also well suited to a message passing architecture. They impose serialization delays, but this penalizes neither the GUPS nor the FFT benchmarks significantly.

A site or supercore in the static WDM point-to-point optical network (Fig. 5) uses a combination of wavelength-division multiplexing in a waveguide and space-division multiplexing across multiple waveguides to establish a unique link to every other site in the macrochip [38]. These are shown as 16 waveguides, each carrying one wavelength and each destined for a different target bridge. We draw all 16 for illustrative reasons only; in a real system, we would only need 15 optical transmitters to reach all other bridges. A WDM mux merges groups of four transmit (TX) waveguides into a single four-wavelength WDM waveguide. This runs east–west over to the column containing the destinations of those four TX waveguides. Through an interlayer coupler, this four-wavelength bundle drops to the second SOI routing layer and then runs north–south. At each of the four destination chips, a WDM drop filter pulls off the appropriate wavelength, routes it back through the first SOI wafer, and then onto the target bridge chip. The network has no in-line switching and is nonblocking, yielding lower latency and higher sustained bandwidth than certain hybrid optical networks employing dynamic, electronic switch configuration.

Table 2 shows the nominal configuration of the optical network for an 8×8 macrochip. Signals in this on-macrochip network travel in low-loss waveguides drawn directly on the macrochip SOI wafer itself. Using two such wafer layers with orthogonal routing avoids waveguide crossings; interlayer couplers are used to connect between the wafer layers. Fig. 6 outlines the routes on this network for a sample 4×4 macrochip. Each L-shaped chip is a bridge (or a pair of separate bridges) overlapping a DRAM chip that contains the multicore processors and the photonic devices and circuits. The shape and amount of overlap of the bridge chip will be dictated by alignment and packaging considerations discussed in the next section. The optical signals coming out from each bridge couple into the routing layer 1 and are multiplexed together into row waveguides that run across different columns. At the

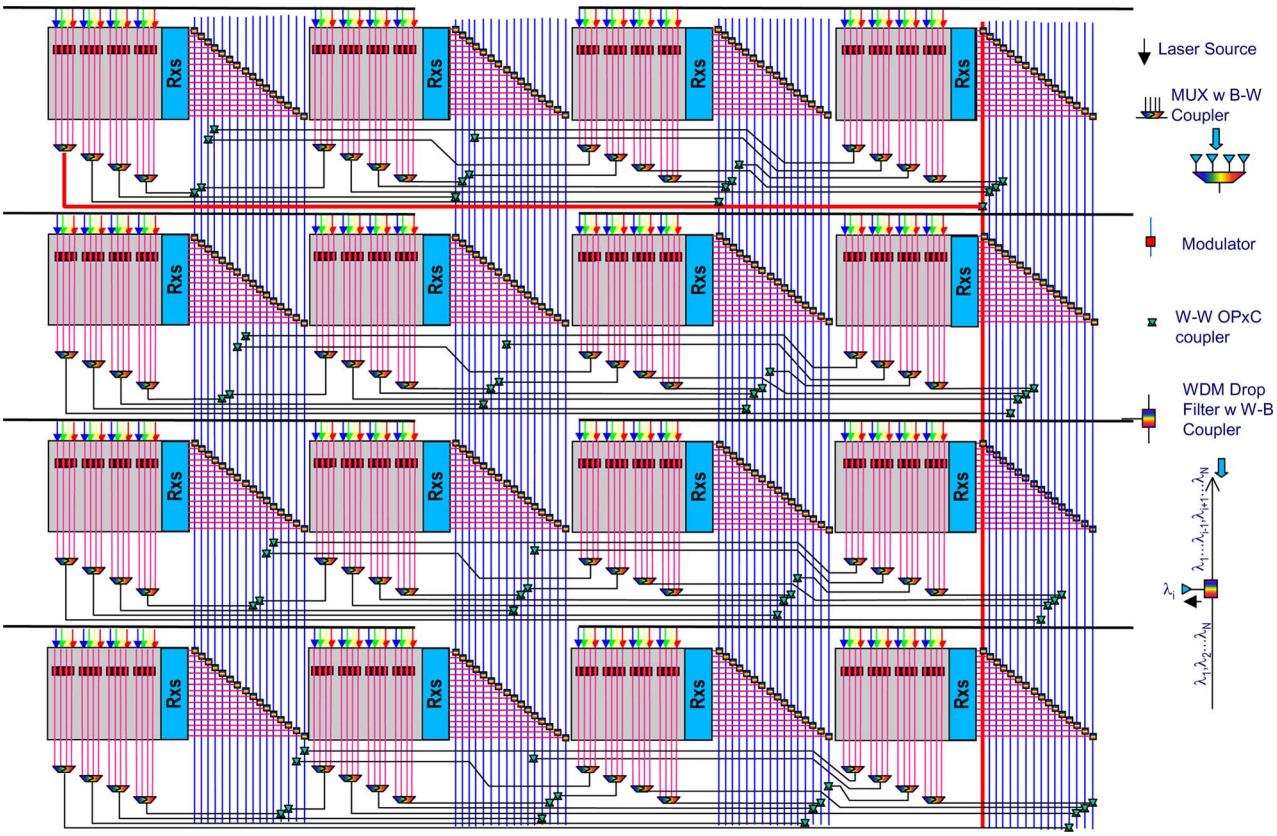


Fig. 5. ***Point-to-point networking on an $N \times N$ macrochip: N waveguides per site and N wavelengths per waveguide are used to create a fully connected network. $N = 4$ in this example.***

destination column, the optical signals couple into column waveguides on the second routing layer via interlayer coupling and are guided to all the chip sites within that column. Different wavelength channels drop at different sites via face-to-face layer coupling. In this way, the network achieves fully nonblocking, source-based routing, without relying on slotting or worrying about collisions. As shown in Figs. 5 and 6, the fully connected point-to-point network can also be used for broadcast from one site to a given column by sending the same message on all four wavelengths on a given waveguide, and similarly to a given row by sending the same message on the same wavelength on all four given waveguides. The topology for a 8×8 macrochip is a simple extension of this arrangement utilizing 8 waveguides and eight wavelengths per waveguide.

V. PHYSICAL STRUCTURE OF THE MACROCHIP

This section describes the physical structure of the hybrid chips comprising the macrochip. These include processor, memory, and photonic bridge chips arranged

Table 2 Nominal WDM Network for an 8×8 Macrochip. Bandwidths Listed are for Payloads, Which do not Include Protocol Overhead.

Network	Architecture	Data rate (Gbps)	Pulse width (ps)	Number of λs (#)	Spectral BW (GHz)	Channel spacing (nm)	Spectral range (nm)
Photonic data network	WDM PtP	20	50	8	50	1.6	12.8
Off-chip interface	Pt-to-pt	20	50	8	50	1.6	12.8

Network	Architecture	Power per λ (dBm)	Number of TX per chip (#)	Total TX BW per chip (GBps)	Number of RX per chip (#)	Total RX BW per chip (GBps)	Macrochip total BW (TBps)
Photonic data network	WDM PtP	0	128	320	128	320	20.48
Off-chip interface	Pt-to-pt	-3	8	20	8	20	2.56

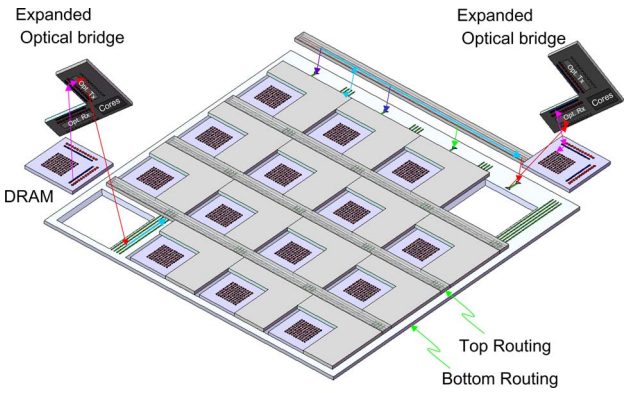


Fig. 6. ***Example routing of a WDM broadcast of a message from the row 1, column 1 chip link on a 4 × 4 macrochip to all rows of column 4, where different wavelengths are dropped onto the different rows of column 4.***

in two-dimensional arrays. For discussion purposes, it is useful to connote the lower chips in Fig. 6 as island chips and the upper chips as bridge chips. Hence, power and ground may be externally provided to the island chips, which may additionally have much greater functionality, processing power, and consequently power consumption. The bridge chips will contain the optical transmitter and receiver circuits and also capacitive proximity communication circuitry. Bridge chips containing processor cores may also be bonded to the memory chips. An alternate arrangement would have both the memory and processor chips face-up, with the bridge chips facing down. The bridge chips can be flip-chip bonded to island chips and derive power from the island. This allows separation of the cooling and power delivery functions on opposite sides of the macrochip. The bridge chip can be thinned and can have an arbitrary shape, defined by both saw-cut edges and etched and lapped features. Two, three, or four wings can be provided which allows compliant overlap with neighboring chips along one or two dimensions.

The photonic bridge chips must be flip-chip attached to the island chips, as shown in Fig. 6. Links are established between chips with OPxC or capacitive proximity communication (PxC). In order to ensure reliable, low power, low bit error rate off-chip proximity communication, neighboring chips need to be positioned to a fraction of the size of the capacitive pad for PxC or a fraction of the optical mode size for the OPxC. This chip-to-chip separation must also be controlled to a few micrometers to ensure the fidelity of the communication channels. Additionally, the macrochip components need to be powered and cooled down while maintaining the chip alignment intact.

We developed a manufacturable process for putting alignment features into silicon containing CMOS integrated circuits. The aligning pit fabrication follows a foundry standard back-end-of-line process and classifies as a post-CMOS process. For packaging a macrochip, the alignment pits are based on silicon micromachining of lithographically defined features in CMOS that are best described as truncated pyramids [37]. The alignment mechanism takes advantage of miniaturized versions of two of nature's perfect shapes: an inverse pyramidal shape with atomically smooth surfaces and a sphere. The inversed pyramids are defined by a self-terminating anisotropic wet etch process in silicon while spheres are a mass-produced commodity made of sapphire or other materials with attainable smoothness of up to a tenth of an optical wavelength. This alignment mechanism enables the two versions of communication between chips in the form of optical proximity communication as well as capacitive proximity communication. Both are depicted in Fig. 7.

A key aspect of the alignment mechanism is the silicon etch-pit. When silicon is etched through a defined (100) surface, four (111) facets appear to form an inverse pyramidal structure. The angle of the etch-pit sidewall, set by the (111) planes when etching a (100) silicon surface, is precisely 54.7°. A set of such etch pits can be fabricated into the corners of each silicon chip containing the proximity communication circuits. As shown in Fig. 7, the silicon chips are then positioned face to face. Etch pit wells capture precision spherical balls that are inserted into the wells before the positioning step. The two chips are then brought in mechanical alignment. As the joining process continues, the balls eventually colocate the two chips as the balls equilibrate into each corner of the chips. Because of photolithographically defined size and the location of the etch pits, and the autocentered mating of the ball in the inverse pyramidal etch pit, the precise relative position between two chips is established (Fig. 8). In addition, with uniform size precision balls, the gap between the two chips can be accurately controlled and maintained in a range less than 1 μm to over 100 μm. The etch-pits can be created

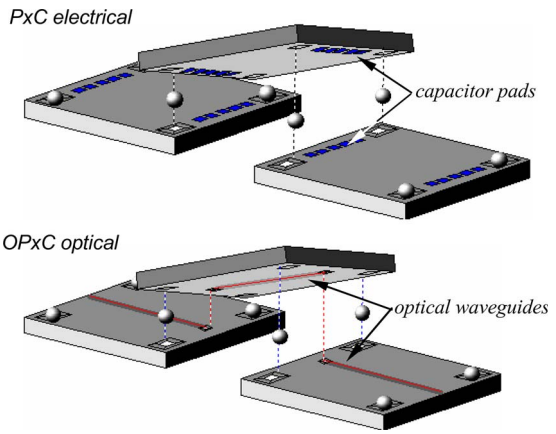


Fig. 7. ***One method of aligning chip for capacitive or optical proximity communication is based on pyramidal etch pits and microspheres.***

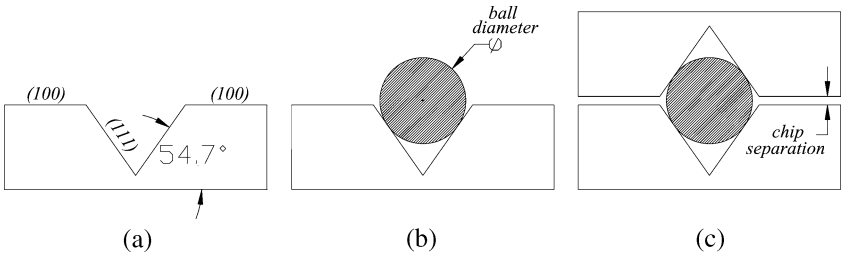


Fig. 8. *(a) Silicon etch along the preferential (111) plane; (b) ball residing on the etched surface; and (c) chip-to-chip alignment with ball in the pit allowing for simultaneous control in x-, y- and z- relative positioning.*

with photolithographic precision before, during, or after circuit fabrication on the silicon chips. This enables the etch-pits to be precisely defined and positioned in relationship to the circuits. The exact depth of the well is unimportant when using the etch chemistry described above. This is a key component to enable a simple (low-cost) manufacturing solution since neither a timed etch nor a special stop-etch layer is necessary.

The ball-in-etch-pit alignment concept was applied to a demonstration of a silicon macrochip package. The macrochip was comprised of silicon bridge and island chips (without active circuitry) that were fabricated at wafer-scale. Each bridge chip was specially shaped with wet etch silicon micromachining and featured with four "wings." Two etch-pits were placed into each "wing" which then housed correspondingly sized sapphire balls (Fig. 9).

The island chips, square in shape, carried two matching etch pits on each edge allowing for a perfect locking to the bridge. The 4 × 4 array composed of bridge and island chips was easily assembled on the copper heat sink as shown in Fig. 10. The chip-to-chip relative distances were measured; each chip was found within 3 μm of its ideal placement.

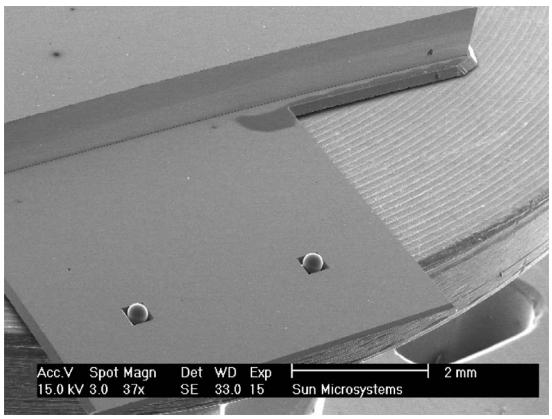


Fig. 9. *One wing of a bridge chip. Installed in the etch pits are sapphire balls ready for macrochip assembly.*

As mentioned earlier, a critical step for packaging PxC and OPxC modules is to find the means to deliver power to the bridge chips which are facing the heat sink in the example shown in Fig. 10. We have developed a high density, low resistance electrical interconnect, microsolder, which powers the bridge chip when it is flip-chip bonded to another island chip. This island chip could be of the same or different functionality from the island chip that was used earlier for alignment with balls in pits. The microsolder is a dense array of specially shaped microbumps. They are designed to have small pitch with low electrical resistance and a high level of compliance after flip-chip bonding to result in extremely small (1 μm or less) chip-to-chip separation. Each microbump is a metal alloy consisting of a square 3 μm tall base and "crown" elevated over the base edges by 4 μm. This special shape ensures high conductivity as the crown is embedded into an opposing pad during flip-chip bonding. Bumps are e-beam deposited onto aluminum pads of an island chip. The bumps could be scaled down to several micrometers in diameter with a comparable separation. Fig. 11 shows square-shaped 18 μm bumps on a 45 μm pitch. The interconnection is completed with flip chip bonding by the means of thermal compression. After alignment of the chips, low-viscosity epoxy is introduced on the chip surface; the chips are brought together and compressed under several pounds of loading pressure at modest temperature. Fig. 12 shows an individual microbump before the flip chip bonding and a cross-section of the compressed bump between the bridge and island chips. The resulting electrical resistance per microbump is under 100 mΩ.

The macrochip package includes a liquid cooled cold plate underneath the multilayer silicon photonic wafer and a power plate atop it; the power plate is then connected to an industry standard electrical interface. Power can be delivered through this plate to the die in the macrochip wafers. Bolster plates surrounding the entire structure provide structural rigidity, and harnesses, voltage regulators, and dc converters provide for power delivery. A more detailed discussion of the packaging of the macrochip is beyond the scope of this paper and will be the focus of a future publication.

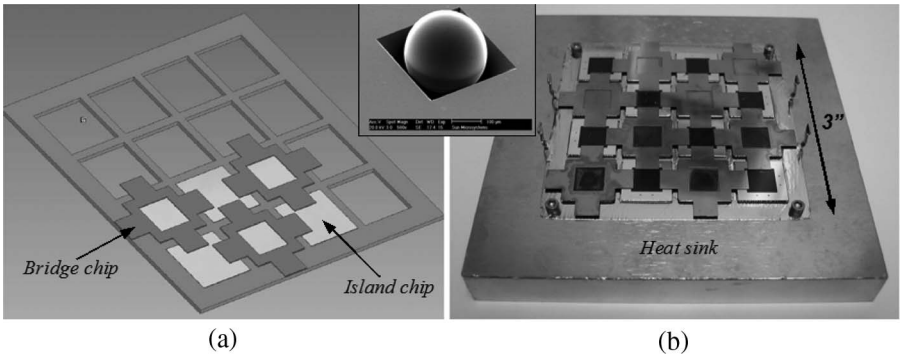


Fig. 10. *Macrochip demonstration in a 4 × 4 array. Schematic assembly is shown in (a) as the array is being populated. (b) Complete array is mounted on the copper heat sink. The inset above displays an actual sapphire ball placed into an etch-pit. Bridge and island chips are self aligned and locked within microns of each other with balls in the pits.*

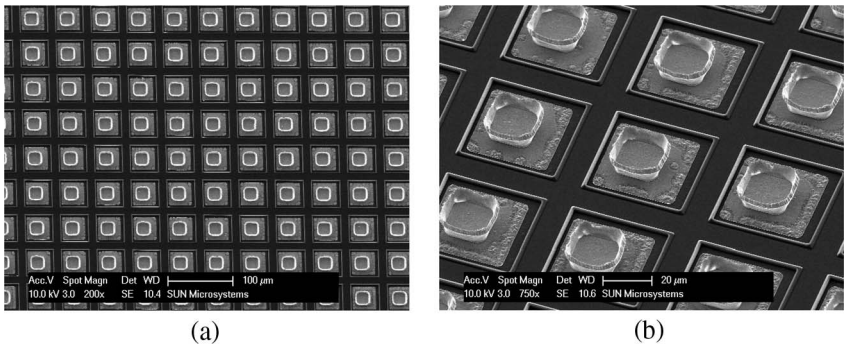


Fig. 11. *Microsolder SEM micrographs. (a) Top view of the high-density array and (b) a closeup of several microbumps.*

VI. BENEFITS OF THE SILICON PHOTONIC MICROSYSTEM

A. Link-Level Comparison

The proposed macrochip design (summarized in Tables 1 and 2) requires a DRAM bandwidth of 640 GBps per site, a system bisection bandwidth of 1 B/flop (or 10 TBps), an aggregated bandwidth of 20 TBps, I/O of 2.5 TBps (to disks or users), and additional scalable off-macrochip bandwidth for node-to-node fiber interconnects. This design uses three types of interconnects: high bandwidth, low latency memory access for the processors; massive, high-density message-passing on the macrochip among processors; and off-macrochip I/O for node-to-node interconnects. We have optimized all three interconnection types not only for high bandwidth and low latency but also for power efficiency.

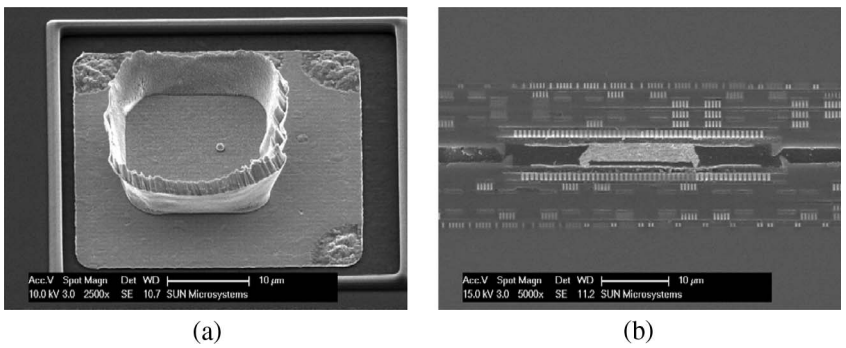


Fig. 12. *Microbump (a) before and (b) its cross-section after flip-chip bonding of two CMOS chips.*

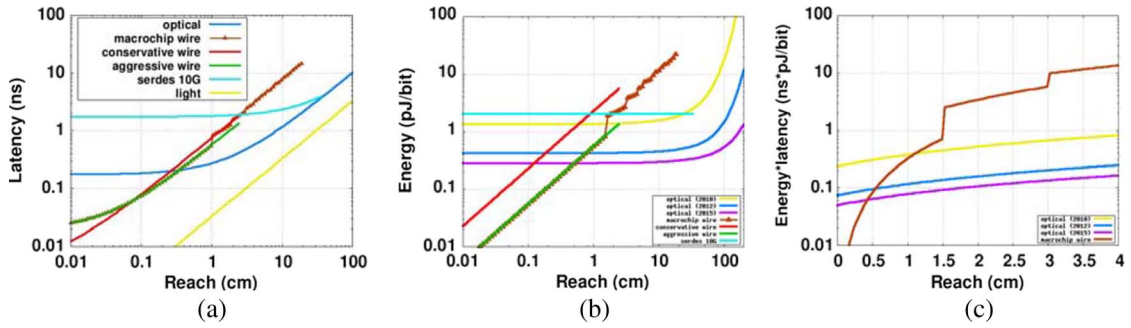


Fig. 13. *Comparison of silicon photonic interconnect signaling across an 8 × 8 macrochip with optical proximity communication versus electrical macrochip wires with capacitive proximity communication and SerDes on a printed circuit board. (a) Latency versus reach; (b) energy versus reach; (c) energy–latency product versus reach. The macrochip wire combination of electrical on-chip wires and capacitive proximity communication shows a staircase-like growth due to the off-chip hops at the edge of each chip in the macrochip.*

Capacitive proximity communication provides low-power short-distance interconnections with very high bandwidth density. Within the macrochip fabric, we bring DRAM chips into close face-to-face alignment with processor chips and use capacitive proximity communication between them. Last-generation proximity communication technology in a 180 nm process is capable of a bandwidth density of 500 Gbps/mm^2 at 3.6 pJ/bit [31]. Assuming a 22 nm process in 2015, we project its capability to be on the order of 1 Tbps/mm^2 at under 1 pJ/bit. To achieve 640 GBps bandwidth, approximately 5 mm^2 of area will be required with a power consumption of about 5 W.

For both latency and energy reasons, silicon photonics-based optical communication appears to be best suited to implement the on-macrochip network. In Fig. 13, a comparison of three candidate interconnect technologies is presented. SerDes links on a conventional PCB [39] are compared with low-power silicon photonic links and also with a combination of capacitive proximity communication and electrical on-chip wires. Fig. 13 shows the latency, energy, and energy-latency product expected of the various technologies. For distances shorter than the size of a single chip in the macrochip, on-chip wires, as expected, have lower latency and use less power than optical links or SerDes links. However, when paired with capacitive proximity communication to hop between chips, their cost grows rapidly with the number of chip hops, while the optical and SerDes links have a lower latency and a constant energy (dominated by the latency and power dissipation of the transmitter and receiver circuits) up to a certain distance beyond which the channel or link loss becomes dominant. Beyond the distance of a few chips, aggressively scaled silicon photonic links become the efficient choice.

Fig. 14 shows histograms of the latency, the energy, and the energy–latency product of the photonic network on a macrochip versus a comparable electronic implementation of macrochip wires based on electrical on-chip routing and capacitive proximity communication. In these histogram plots, messages are assumed to be randomly distributed across the macrochip. Compared to a mixture of on-chip

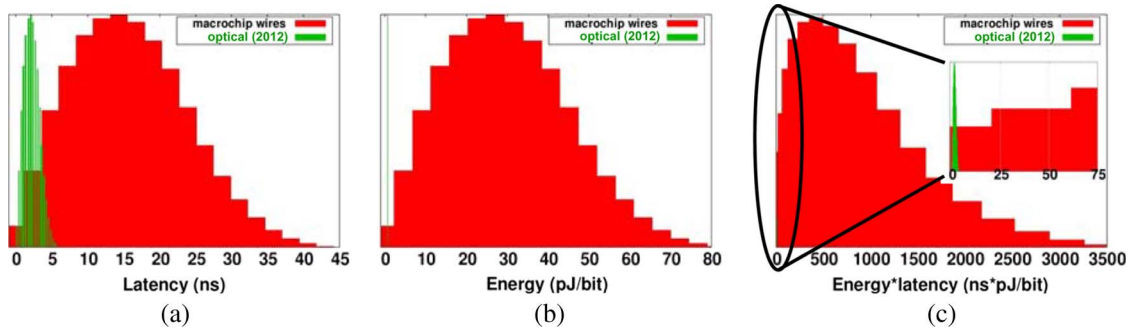


Fig. 14. *Histogram of latencies of electrical and silicon photonic links for a macrochip passing randomly distributed messages using aggressive on-chip wires and capacitive proximity links. (a) Histogram of latency of macrochip links comparing macrochip wires versus silicon photonic interconnects. (b) Histogram of energy of macrochip links. (c) Histogram of energy latency product of macrochip links with inset showing distribution of silicon photonic interconnects.*

wires and capacitive proximity hops, silicon photonic interconnects clearly provide a lower average latency (a 5×–6× improvement) and also a narrower distribution of latencies across the different links in the macrochip. Both of these advantages may be exploited to provide a flatter memory hierarchy across the macrochip [40]. The photonic network can also be 6×–300× more power efficient, assuming that the silicon photonic interconnect roadmap described in Section VII is realized. The results for a comparable electronic implementation assume the use of electronic proximity interconnect to achieve the density required for the same bandwidth as the photonic implementation in a similar form factor; we estimated wiring parameters for both traditional and aggressive electronic interconnect. These include wire delays of either 75 or 55 ps/mm of wire and energy costs of either 1 or 0.1 pJ/bit/mm. Proximity links cost slightly over 200 ps and 0.15 pJ/bit per hop.

B. Macrochip Network-Level Comparison

In order to assess the impact of the silicon photonic interconnect, we compared the performance of the nonblocking point-to-point network with that of a canonical hybrid optical mesh network. The hybrid network is a nonblocking intrachip processor-to-processor mesh network that contains optical switches along the path from source to destination port. The optical switches are controlled via a separate electronic network that has a topology identical to that of the optical network. In order to transmit data optically, an optical path needs to be set up through the optical switches from the source to destination. This is done by explicit path-setup messages forwarded through the electronic network that consists of electronic routers at each mesh intersection. Data can be transmitted optically upon receiving a path-setup acknowledgement from the destination port. Similarly, at the end of data transmission, optical paths are torn down using explicit path-breakdown messages.

We performed a performance comparison using event driven simulation models of both networks for an 8 × 8 macrochip configuration. Each network port in both of the models is assumed to have a peak communication bandwidth of approximately 320 GBps. We assumed an optical latency of 0.1 ns/cm for the point-to-point network; for the hybrid network, we assumed an electronic latency of 0.44 ns for a router hop and a router processing delay of 0.6 ns. We performed simulations for three different packet sizes of 128 Bytes, 1 KB, and 16 KB.

Fig. 15 shows the sustained bandwidth of the two networks as a function of load in the simulation models. Because the electronic paths are slower than the parallel optical paths, there is considerable time required to set up (and tear down) a circuit. This setup time is incorporated into the analysis of the hybrid network: the hybrid network can require several tens of nanoseconds to set up a connection, while the point-to-point WDM network has no

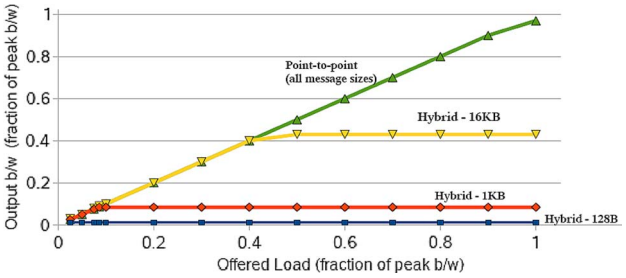


Fig. 15. ***Achieved bandwidth as a function of offered load for the point-to-point network versus a hybrid macrochip network with short, medium, and long message sizes.***

setup time. This means that the effective bandwidth of the hybrid network can be significantly lower than its peak. For even moderately sized data transfers (e.g., 1–2 KB), only a fraction of peak bandwidth would be usable; the network is more efficient when transferring larger data blocks. In the fully connected point-to-point network, there is no bandwidth loss due to path setup and the sole penalty is for an output port at a site, so almost all of the peak bandwidth is usable. As a result, the WDM point-to-point macrochip network has much better latency and bandwidth characteristics than the corresponding hybrid network. Depending upon the communication requirements of the application or benchmark, these characteristics can reduce the energy consumed per useful bit of delivered information.

C. Macrochip System-Level Comparison

We performed an analytical bounds analysis to quantify the performance benefits of a single photonically interconected macrochip to an all-electronic macrochip supporting the same peak bandwidth. Fig. 16 shows the results of this analysis in terms of system performance per watt for four benchmarks: GUPS, FFT, LINPACK (linear algebra package), and STREAM (a test of sustainable memory bandwidth) from the HPC challenge benchmark suite. Results are shown for both the "aggressive" and the "basic" electrical interconnect technology described above.

The system-level power efficiency of photonic interconnects significantly exceeds that of electronic interconnects provisioned for the same bandwidth: for the macrochip, it is 6–40× better. Improved power efficiency returns a greater "win" because it also reduces chip thermal density. The analysis shows that the thermal density of the 2015 macrochip is 2–5× lower than that of recent ×64 processors.

In these optical-to-electrical comparisons we have provisioned equal system bandwidth. However, the latencies of the two systems are dramatically different, by a factor of 5–6×. Architecturally, this is a significant difference that many applications cannot hide. Methods to

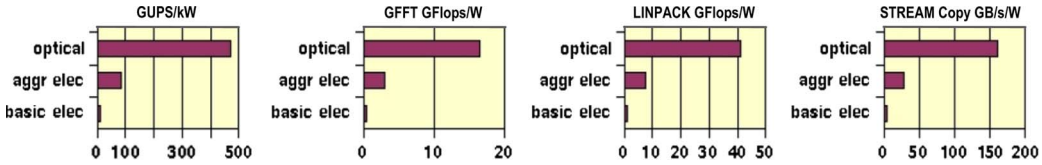


Fig. 16. *Single macrochip node performance per watt on HPC challenge benchmarks. These compare a macrochip implementation with optical links to implementations with basic and aggressive wires using capacitive proximity communication.*

equalize the latency between electronic and optical interconnect are very costly. Making up the entire 5× difference would require impractically expensive on-chip transmission lines, each consuming as much as 40 μm of pitch for differential coplanar microstrips and using several picojoules/bit of energy.

Alternately, we could improve the speed of the traditional repeated on-chip wires. An improvement of 2X in the wires is possible but will consume at least 2X the wire energy (further latency improvements with on-chip wires are far beyond the point of diminishing returns). Thus, closing the latency gap to 2.5X will roughly double the energy difference between optics and electronics. This tradeoff clearly shows and magnifies the benefits of the photonic interconnection network proposed in this paper.

We note that the fully connected point-to-point photonic network allows an extremely high input bandwidth at each network port, so it is important that the processors and memory be able to handle this bandwidth. The macrochip configuration is designed so that the processor and DRAM input bandwidths are matched to the network input bandwidth. Furthermore, the processors are multithreaded with efficient message handling hardware. As mentioned earlier, support of shared memory can yield integer factor reductions in code size and corresponding improvements in productivity. Efficient support of shared memory requires a network that can sustain a high fraction of peak bandwidth at small message sizes.

So far we have discussed the connectivity on a macrochip. A key advantage of the silicon photonic links on the macrochip is that they may be used with single-mode fiber to provide connectivity across a larger system. To facilitate scaling to petascale systems and beyond, additional waveguides in the macrochip wafer lead to two off-macrochip fiber networks: an I/O fiber network uses 128 fibers per macrochip to attach to disk and users, and a node-to-node fiber network can fully connect each of up to 1200 macrochips to every other macrochip, using up to 2400 fibers per macrochip (fewer fibers if wavelength multiplexing and routing is employed). This last network is omitted for single macrochip embedded systems. Together, the two off-macrochip networks combine to maintain 1 Byte/flop of bandwidth across the multinode system, resulting in a fully connected 10 Pflop system with breakthrough bisection bandwidth and bandwidth per watt, as depicted in Fig. 1.

VII. ENERGY-PER-BIT CONSIDERATIONS FOR SILICON PHOTONIC INTERCONNECTS

As discussed earlier, the power consumption of optical communication must be dramatically reduced from the 10–50 pJ/bit levels typical in present-day systems. Optical links will not replace short electrical links in computing systems unless their per-bit energy costs are much lower; at a system level, the benefits of the fully connected network on a macrochip are not appealing if the interconnects consume too much power. Indeed, the simple all-to-all optical network, which provides an overprovisioned total network bandwidth in order to eliminate network switching, is largely predicated on a very low per-link power, which we discuss next.

With the macrochip-on-routing-wafer physical packaging structure, the following components comprise each optical link (Fig. 17): laser source, bridge-to-wafer coupler, modulator, waveguide on bridge, bridge-to-wafer coupler, WDM mux, waveguide on wafer, interlayer coupler, wavelength dropper, interlayer coupler, bridge-to-wafer coupler, and receiver. For a 2015 macrochip, we expect photonics links with a 20 Gbps channel data rate (or a pulse width of 50 ps) and a power efficiency of 160 fJ/bit. We also expect the worst case (or longest) link loss of an 8 × 8 macro chip to be about 17 dB. No optical amplifier is

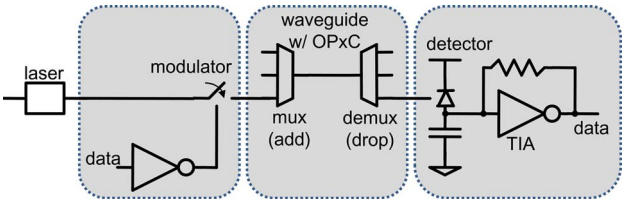


Fig. 17. *A canonical representation of a photonic link. The shaded area represents the on-chip component of the link. Limiting amplifier stages in the receiver often follow the transimpedance amplifier. Not shown are clock and data recovery circuits that may follow the receiver.*

needed (or can be afforded given current amplifier power efficiencies) for a 0 dBm laser source and an optical receiver with −21 dBm sensitivity. A minimum of eight-channel WDM components are assumed at a 200 GHz channel spacing and a 0.2 nm spectral bandwidth to create a 16 nm spectral range.

A. Power Costs and Circuit Topologies

In this section, we discuss the energy of optical links, which is closely related to the electrical circuits that drive them, and how these energy costs can be reduced in a macrochip. We start by enumerating the dissipated power of an on-chip optical link for a given link bit rate

$$\text{Power}_{\text{on-chip}} = P_{\text{RX}} + P_{\text{TX}} + P_{\text{WDM}} + P_{\text{optLoss}}. \quad (1)$$

Here, the total power represents the effects of the receiver, transmitter, WDM mux/demux, and optical (photon) loss on the macrochip. The WDM component can be broken into a mux and a demux portion and primarily accounts for static and dynamic tuning power in the mux and demux. Tuning may also be required for resonant silicon modulator structures, and is discussed later. Tuning power includes not only the actual bias energy used for tuning but also the energy required to control the bias. This has traditionally required a sensor, a feedback loop, and control hardware, all of which cost area and power. However, as demonstrated in clock phase adjusters in electronic chip-to-chip I/O, tuning can be done digitally and with software control in a side or service processor or thread [39]. If dynamic tuning is intended to compensate for relatively slow effects, such as thermal drift, then this control can be made sufficiently fast yet not consume significant on-chip energy

$$P_{\text{RX}} = P^{\text{ckt}}_{\text{static}} + P^{\text{ckt}}_{\text{dynamic}} + P^{\text{detector}}_{\text{bias}} + P^{\text{detector}}_{\text{tuning}}. \quad (2)$$

Receiver power is composed of the receiver circuit bias power plus leakage power, the dynamic switching power of the receiver circuit, the detector bias power (representing the sum of the dark current and the signal current), and any tuning that may be required for the detector, such as if the detector is placed in a resonator. The macrochip architecture is based on optical links running at 20 Gbps and optical-electrical receivers consuming 30 fJ/bit.

In some sense, the ideal receiver is one that simply integrates photodiode current onto a plain capacitor (which could be the receiver capacitance itself), perhaps dumping the integrated charge each cycle to avoid running out of voltage headroom. This idea is only feasible with extremely low detector and parasitic capacitance (under 5 fF); this is set by the relationship of V = current/(capacitance * bit rate/2). Clearly, the additional capacitance of hybrid bonding between VLSI circuits and the photodetectors may exceed this value—even for small flip-chip pads; also, any wiring on a chip may disqualify such "receiver-less" links, as wire imposes a load of a few femtofarads for every 10 μm of length routed. In the remainder of this paper, we will take the view that for these detectors with nontrivial device and parasitic capacitance, some measure of electronic amplification in the receiver will be necessary.

The receiver must translate an input current into an output voltage, i.e., achieve transimpedance. At an input sensitivity of −20 dBm, a contrast ratio of 5×, and a responsivity of 0.75 A/W, the input current has a signal swing of 2.5 to 12.5 μA of input current, overlaid on top of the photodetector's dark current. This input signal swing of 10 μA must be turned into a moderate voltage swing output by the receiver, on the order of 200 mV; a simple and power-efficient clocked sense amplifier can do a subsequent amplification from 200 mV to a full-swing CMOS signal. The first-stage conversion from 10 μA to 200 mV leads to a required transimpedance gain of 20 K. The challenge of the receiver circuit is to provide this transimpedance at a high bit rate, with low power, and with low noise.

One may debate the choice of a 200 mV input voltage to the clocked sense amplifier. This 200 mV swing is intended to overcome six-sigma offsets in the sense amplifier and support circuitry, the residual thermal noise in the sense amplifier, and still provide sufficient signal to maintain fast amplification. This required swing may be reduced somewhat by implementing offset compensation in the sense amplifier; this will, in turn, reduce the required transimpedance and receiver energy costs. However, because clocked sense amplifiers spend half their time precharging, two are required for full data-rate operation, with the data ping-ponging between the two. Offset compensation with two alternating sense amplifiers is moderately complex and involves energy costs that reduce the gains achieved by lowering the required input signal.

Receivers can be as simple as a resistor feeding the photodetector and sense-amp capacitances. Getting a high transimpedance gain simply means using a large resistor. However, the resulting resistor–capacitor time constant is too high for any but the slowest of data rate channels. Classical transimpedance amplifiers (TIAs) with feedback can improve on this situation by providing a low impedance input but a large current-to-voltage gain. However, transistor and voltage scaling reduces device transconductance, reducing the available gain and making traditional shunt–shunt TIA designs increasingly complicated and almost certainly requiring several stages of amplification. From a noise perspective, the feedback TIA closely resembles the plain resistor, as amplifier gain attenuates most of the noise added by transistors. Other common TIA topologies, such as isolation drivers using a regulated cascode, perform slightly more efficiently than the feedback TIA

but at the cost of higher noise, mostly due to the current bias of the regulated cascode.

Two important questions must be addressed by any macrochip optical receiver design. First, how do the receivers know when to clock their data? Full clock-and-data-recovery is typically power-hungry. However, because a macrochip is small, and because every chip in the macrochip shares the same global clock source, clock phase differences between any two chips in the system are bounded; this is known as a "mesochronous" system. In fact, clock phase differences between different chips should change only slowly. Moreover, because optical links are intended to be inexpensive in area and power, the macrochip is expected to run many optical links at double the clock rate instead of a few links at a much higher overclock ratio. This means that data clocking can be based on an already-distributed global digital clock, suitably delayed, for very low incremental cost. Updating this delay can be done slowly and in software running on a service processor or thread.

Secondly, a receiver must compare an input signal (whether current or voltage) to a reference, to determine whether the value was a "0" or a "1." How does it define this reference appropriately? In modern optical links, this is done by comparing a TIA's instantaneous output voltage with a long-running average of the TIA's output, the idea being that the long running average will center halfway between the values corresponding to a digital "0" and "1." This of course requires data that are dc-balanced, or data that has an equal number of "0"s and "1"s. As a result, most modern optical links enforce dc balance through schemes such as 8B10B encoding [41], which maps every byte into a 10-bit space. Because there are only 10-choose-5, or 252, 10-bit words with an even number of 1s and 0s, this mapping is imperfect, and 8B10B may have a "running disparity" of $+1$ or -1 at any given moment. Thus the average is not always a perfect halfway value, but close enough for low bit error rates.

However, such codes impose a cost, such as a 25% overhead for 8B10B. In the macrochip, we can leverage another characteristic, which is that the system is small enough that global events can be reasonably synchronized between any and all of the chips in the system. For instance, a global training session can be established, in which all links are preprogrammed to send a 010101... data pattern, and all receivers take an average of the input data. After sufficient training, the entire macrochip can be synchronously triggered to begin actual operation. In addition, this can be redone periodically, again using a global "refresh" command that reaches all chips in the macrochip at essentially the same time. This shared global sense of time makes such training feasible and inexpensive: a training of 100 cycles done every 100 000 cycles is a much lower overhead than 8B10B-like systems.

For both of these questions, then, we are able to leverage the small physical size of the macrochip to enable energy optimizations that make truly low-power transceivers possible. Traditional long-haul or data-center optical interconnects cannot employ such techniques, and hence pay a higher communication cost. But here, the high density of the macrochip not only improves the communication latency but also provides a framework for power-optimized links

$$P_{\mathrm{TX}} = P^{\mathrm{ckt}}_{\mathrm{static}} + P^{\mathrm{ckt}}_{\mathrm{dynamic}} + P^{\mathrm{mod}}_{\mathrm{bias}} + P^{\mathrm{mod}}_{\mathrm{dynamic}} + P^{\mathrm{mod}}_{\mathrm{tuning}}. \quad (3)$$

For transmitters, the power includes leakage from a reverse-biased modulator and leakage from the circuits, but, depending on the design and implementation of the modulator, power can be dominated by the dynamic switching power of the modulator (in case of larger Mach–Zehnder modulators) or by the tuning power (for highly resonant devices), or even by the absorbed photocurrent (in case of absorption modulators) when the link loss is high. This last term represents photons absorbed in the modulator that convert to electrons and then sunk by the driver circuits

$$P^{\mathrm{mod}}_{\mathrm{bias}} = V^{\mathrm{mod}}_{\mathrm{bias}}\left(I^{\mathrm{leakage}}_{\mathrm{bias}} + I^{\mathrm{photocurrent}}_{\mathrm{bias}}\right). \quad (4)$$

The challenge for transmitter circuits is that modulators often require a modulation voltage of 2 V or higher from technology processes that support 1 V transistors. While most advanced (90, 65, and 45 nm) processes still support 1.8 V and sometimes 2.5 V voltages for legacy bus standards, these signal swings are propagated using thick-oxide and slow devices, unsuitable for high-speed modulation. Swinging high-voltage signals using a lower-voltage technology requires stacking transistors along the signal path, so that a 2 V swing spreads that voltage across two devices, not just one; this minimizes the overstress damage imposed on the transistors. Controlling these transistors properly, under process and environmental variations, requires careful design and margining. Finally, this requirement for high-voltage operation complicates the design of low-energy driver circuits, since a $2\times$ voltage increase represents a $4\times$ energy and power cost.

Finally, we must take into account the modulation and other optical losses on the macrochip with the following terms:

$$P_{\mathrm{optLoss}} = P^{\mathrm{mod}}_{\mathrm{phLoss}} + P^{\mathrm{detector}}_{\mathrm{phLoss}} + P^{\mathrm{waveguide}}_{\mathrm{phLoss}} + P^{\mathrm{WDM}}_{\mathrm{phLoss}}. \quad (5)$$

Here, terms like $P^{\mathrm{mod}}_{\mathrm{phLoss}}$ represent the photon loss in the modulator. Note that the only term we omit is the inefficiency of generating and delivering optical power to

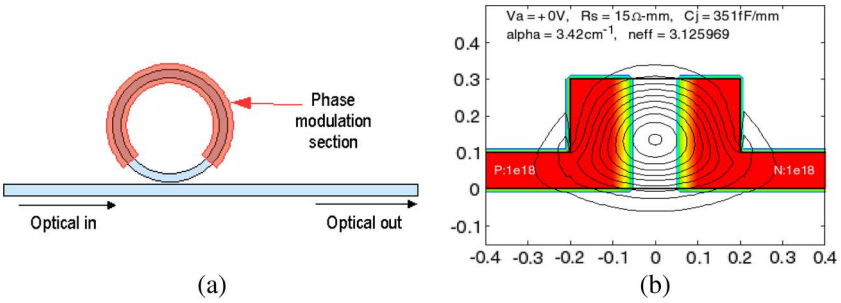


Fig. 18. *(a) Top view of a ring modulator. (b) Cross-section view of the optical waveguide in the phase modulation section, with a two-dimensional color map of the carrier concentrations at 0 V overlapping with a two-dimensional contour plot of the optical mode.*

the macrochip; by analogy, the inefficiency of generating and delivering electrical power to a SerDes is similarly usually omitted. A photon loss of 17 dB can be calculated from a detector with a sensitivity of −21 dBm, a launched optical power of 0 dBm (1 mW), and a margin of 4 dB. At 20 Gbps, this corresponds to 50 fJ/bit (of optical energy) dissipated on the macrochip.

B. Optical Link Power Budget

Every site in a macrochip is interconnected to every other site via static WDM links. The optical signals from (and to) the sites are coupled into (and out of) the photonic routing layer through the face-to-face reflecting pit couplers. The routing layer consists of two layers of wafers with orthogonal waveguides to avoid waveguide crossings. The two wafer layers are interconnected via interlayer couplers.

The expected losses of each component of the optical link are listed in Table 3. The longest optical link in a macrochip needs up to a 40 cm long waveguide on the photonic routing layer. With an expected waveguide loss of 0.05 dB/cm, the total loss of the link is about 17 dB. Assuming 0 dBm optical power launched into the modulator, the optical receiver would receive an optical power of −17 dBm in the worst case. The optical receiver is expected to have a sensitivity of approximately −21 dBm for a bit error rate of 10^{-12} at a data rate of 20 Gbps. This allows sufficient power margin to tolerate link impairments that may include finite modulator extinction ratio, modulation nonlinearity, crosstalk, and so on. Note that in the worst case, an optical route travels through two of each coupler type (bridge-to-wafer, and wafer-to-wafer) and seven through filters before being dropped to its destination site.

Notice that there are potentially significant tuning costs associated with the WDM link when using high-index contrast waveguides with tightly confined optical modes. While the tables show the actual bias energy required for tuning, there is also energy required to control the bias. If relatively slow tuning is sufficient (e.g., to compensate for slow thermal drifts), this control can easily be made sufficiently fast yet not consume significant on-chip energy.

C. Tuning Requirements for Silicon Micro-Resonators

Due to the aggressive energy-per-bit and density specs demanded of the silicon photonic interconnect components, optical microresonator devices appear an attractive choice. These primarily include modulators and mux/demux devices made in microring or microdisk geometries. Among these devices, microring modulators have been investigated by a number of researchers; these efforts are reviewed elsewhere [15]. Fig. 18 shows a schematic top view of a canonical ring modulator, along with a cross-section view of its optical waveguide in the phase modulation section. Here we assume that a reverse-biased PN junction is used for high-speed modulation, and the junction is located at the center of the waveguide, with the same doping density in the P and N regions. This ring modulator structure is fully compatible with standard CMOS fabrication processes. Its modulation bandwidth is determined by its RC limit and its photon lifetime. Assuming a linear-graded doping profile near the junction, we can solve the Poisson equation for the junction and calculate the modulator bandwidth and modulation depth versus voltage swing.

Fig. 19 shows a contour plot of the overall bandwidth (at 0 V) versus the ring radius and doping density for the above ring modulator structure. In the calculation, we assumed a 200 Ω driver impedance and included reasonable parasitic effects. To make the modulator operate at a data rate in excess of 15 Gbps, we require a bandwidth larger than 11 GHz. A ring with a radius of approximately 5 μm and a doping density greater than 10^{18} cm^{-3} can satisfy this requirement, achieving low capacitance ($\approx$15 fF) in a compact footprint ($\approx$100 μm^2). As shown in Fig. 19, the optical loss (2.9 dB) and extinction ratio (7.2 dB) for this design also appear promising.

However, a significant issue for the ring modulator is that it requires accurate tuning to align its resonant spectrum with a fixed laser wavelength. To bias the modulator, a tuning precision within a few tens of picometers may be required, as can be seen in Fig. 19. The problem is exacerbated when the required tuning range increases due to manufacturing imperfections as discussed next. The large

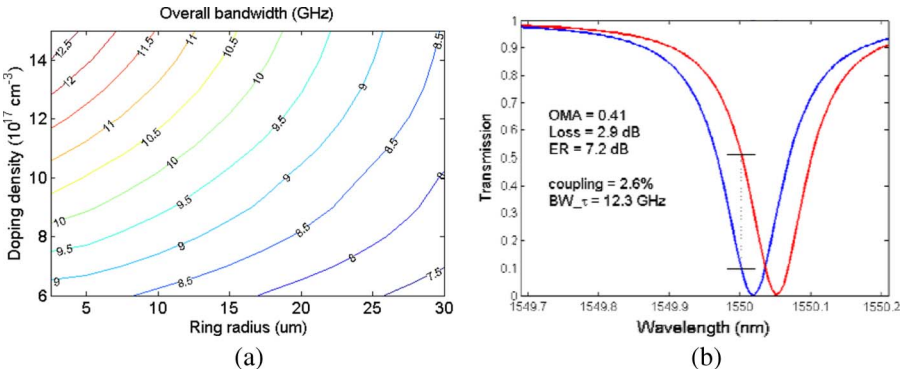


Fig. 19. ***(a) Simulated bandwidth versus ring radius and doping density for depletion-mode ring modulators, including representative parasitic effects and a 200 Ω driver impedance. (b) Transmission spectrum (normalized with respect to input power in linear scale) of a 15 Gbps ring modulator at 0 V (blue curve) and at 2 V (reverse bias, red curve). 60% of the ring perimeter is assumed to constitute the high-speed phase modulation section, and the laser wavelength is fixed at 1550 nm. OMA denotes the normalized optical modulation amplitude. ER is the extinction ratio between 0V and 2V. BW_t represents the photon-lifetime limited bandwidth of the ring. An ON-state loss of 2.9dB and a power coupling loss of 2.6% is assumed for the ring modulator.***

range of tuning may consume a significant amount of power, and the high tuning accuracy may require sophisticated control circuitry. Unless mitigated, these requirements may together annul the benefits of the small capacitance and the low switching energy of the device. Ring mux/demux devices may also require similar tuning, depending on their design.

This component of power consumption is governed by the tuning range for the given resonator design, which is closely related to how much its resonant peak wavelength can vary. Due to the lack of complete published statistical data, the tuning range requirement for silicon photonic microresonator devices has not been conclusively established. In this section, we discuss this question using theoretical analysis and available experimental data.

The mth resonant wavelength λ of a resonator is determined by the following equation:

$$n_{\text{eff}} \cdot L = m \cdot \lambda \tag{6}$$

where n_{eff} is the effective index of the optical mode and L is the resonator perimeter length. A simple derivative of the above equation can lead to

$$\frac{\delta\lambda}{\lambda} = \frac{\delta L}{L} + \frac{\delta n_{\text{eff}}}{n_{\text{eff}}}. \tag{7}$$

The above equation indicates that the resonant wavelength variation $\delta\lambda$ can arise from the optical mode effective index variation δn_{eff} and the resonator perimeter length variation δL. Here we focus primarily on δn_{eff}, which can be caused by various manufacturing variations including waveguide width variation, etch depth variation, silicon layer thickness variation, material stress, and interface roughness. The $\delta\lambda$ caused by these manufacturing variations may be relatively large (up to $\pm$10 nm, as discussed below), and requires static tuning. In addition, temperature variation can also change n_{eff} and cause wavelength drift, although this effect is relatively small (under $\pm$1 nm with $\pm$10 °C temperature variation). This temperature effect, as well as the laser wavelength drift (typically less than 0.1 nm), may necessitate dynamic tuning of resonant wavelength.

Silicon photonic devices are usually fabricated on SOI wafers, having a typical silicon layer thickness of 200–300 nm with a variation on the order of $\pm$10 nm. Etch depth variation can also be up to $\pm$10 nm for an etch depth of $\approx$200 nm. The variation of waveguide width is controlled by multiple manufacturing steps (lithography, etching and oxidation, etc). It may again be on the order of $\pm$10 nm when using a 193 nm lithography tool. As shown in Fig. 20, etch depth (in rib waveguides) and waveguide width variations can cause resonant wavelength shifts up to $\pm$5 nm whereas silicon thickness variations can result in shifts up to $\pm$14 nm. While we expect that all these variations can be improved with process technology advancement, some may be difficult to improve with acceptable yield and cost. After statistical summation, the overall resonant wavelength variation $\delta\lambda$ might easily approach $\pm$10 nm.

The above estimated $\delta\lambda$ is for wafer-to-wafer and lot-to-lot variations for resonators using rib waveguides after undergoing a full CMOS process flow. The use of an abbreviated CMOS process flow (without the backend metal process) or with a non-CMOS process (such as a research facility process or an e-beam process) can reduce $\delta\lambda$. Employing channel waveguides can eliminate etch-depth variations. Restricting fabrication to a smaller area of one wafer, although detrimental to yield and cost, can further improve the wavelength shift. For instance, researchers

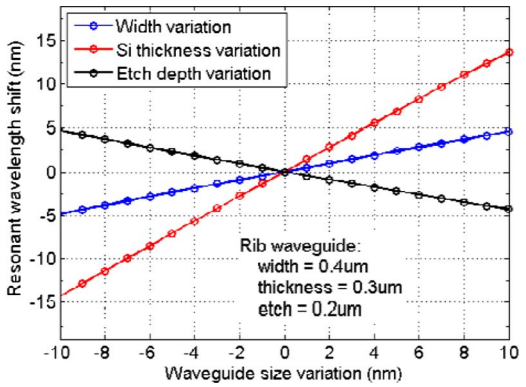


Fig. 20. *Simulated resonant wavelength shift due to variations of waveguide width, silicon thickness, and etch depth for a ring resonator using the waveguide geometry of Fig. 18. The rib waveguide is sandwiched by SiO_2 above and below.*

have built an optical delay line made of 56 closely located identical rings (with a 5 μm radius) fabricated in a research facility [42]. The authors used data fitting to indicate that the resonant wavelengths of the 56 rings could be represented by a Gaussian distribution with $\sigma = 0.4$ nm. Other research results for an array of four closely located ring modulators fabricated with e-beam lithography showed a distribution of channel spacing from 1.3 to 4.0 nm [43]. The test data in both of the above examples indicated that the resonant wavelength can vary more than 1 nm, even though the tested devices were within a small area of one wafer fabricated without the full complexity of a CMOS process. More recent work using a 193 nm lithography tool achieved 7.5 nm of 3σ variation in waveguide width for 175 samples across one wafer (measured after etching, without a full CMOS process) [44]. It also showed that resonant wavelength variation could be improved with finer CMOS fabrication tools, which is encouraging.

The next question is: how much do we need to tune given a specific $\delta\lambda$? Because we may align the laser wavelength to one of many degenerate resonant peaks, we need not tune across the entire $\delta\lambda$ range when the resonant wavelength variation $\delta\lambda$ exceeds the free spectral range (FSR) of the resonator. The worst case tuning range is one FSR (i.e., 2π phase) for unidirectional tuning, or half the FSR (i.e., π phase) for bidirectional tuning. The FSR is inversely proportional to ring radius, and for a small ring with 5 μm radius is $\approx$19 nm. Therefore, based upon the above $\delta\lambda$ estimate and FSR value, we may need to tune as much as π phase (statistical mean value for unidirectional tuning) for many resonator devices. An interesting consequence is that when the wavelength shift cannot be accurately controlled, it may be prudent to use as large a ring diameter as possible and hence reduce the worst case tuning range, trading off against a smaller ring size to improve the switching energy and speed of the ring (19) while accommodating the required number of wavelength channels.

It is clear from this discussion that manufacturing variations must be controlled in order to achieve uniformity in the resonant wavelength of small microresonator devices, and thus simultaneously achieve low tuning and low switching energies. On the positive side, it is quite likely that these manufacturing variations will continue to reduce as lithography improves and CMOS feature sizes shrink. In spite of a number of interesting early experiments, the availability and understanding of resonant wavelength variation data is limited, from which it is hard to derive meaningful information about the wafer-to-wafer and lot-to-lot variations. In addition, these data cannot represent the true variation that one might expect when integrating the microring or microdisc devices into a full CMOS fabrication process and the increased resonant wavelength variation resulting from the wafer-to-wafer and lot-to-lot variations. Ideally, one would be able to measure the resonant spectrum of a large number of small rings (e.g., with 5 μm radius to ensure a large FSR) across each wafer and on multiple wafer lots.

Given the large tuning requirement (potentially up to π phase), a challenge is to implement low power tuning without significantly degrading the resonator performance. One tuning method is to use a forward-biased diode to do carrier injection tuning. It can tune π phase with only $\approx$1 mA injection current and consumes only $\approx$1 mW power. However, a fatal problem with this tuning method is that optical loss also increases with carrier injection. In fact, the changes in optical refractive index and optical absorption coefficient are both roughly proportional to the carrier density change [8]. From these relationships, one can easily calculate the relationship between the phase tuning and the optical loss in the resonator, as shown in Fig. 21. In order to tune π phase, an extra optical loss of $\approx$2 dB will be introduced. Because the total loss is only 0.05 dB for a 5 μm ring with $Q = 20\,000$, this large excess loss will destroy the resonator performance.

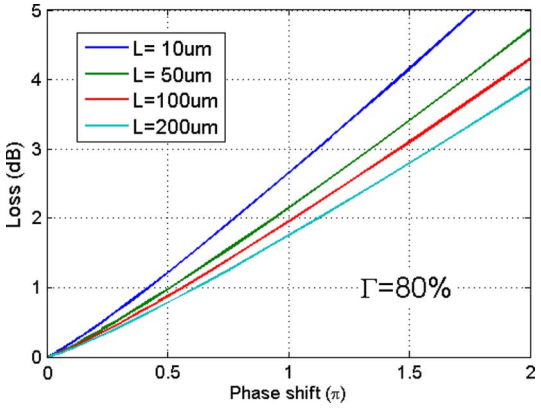


Fig. 21. *Optical loss versus phase shift with carrier injection tuning using forward-biased PIN diodes with different diode length L.*

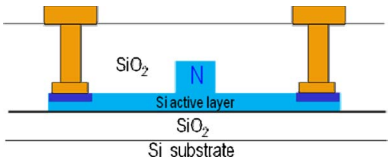


Fig. 22. *A thermal tuning structure by doping the Si waveguide as a resistor.*

Another popular method is thermal tuning. One approach is to put a metal heater on top of the optical waveguide [45]. With this method, an SiO_2 layer is needed between the metal heater and the waveguide for optical isolation, which makes the heat transfer inefficient since SiO_2 is a poor heat conductor. This results in a relatively large dissipation (a power of over 100 mW reported to shift the resonant wavelength by 6.4 nm or $2/3\pi$ phase for a ring with an FSR of $\approx$19 nm). A more efficient thermal tuning method might be to directly heat up the Si waveguide by doping the waveguide as a resistor, as illustrated in Fig. 22. By applying voltage across the resistor, this structure could generate heat directly in the optical waveguide. Its efficiency depends directly on the thermal resistance. A simple thermal analysis for such a silicon resistor structure indicates that the thermal resistance would be on the order of 100 °C/mW for a 1 μm long waveguide; it would take $\approx$50 mW to tune π phase, regardless of waveguide length. This power consumption is still unacceptably high. A large improvement (100×) is needed in order to make it useful for the silicon photonic interconnects considered here.

The above discussion shows that tuning can be a challenging problem for photonic microresonator devices. Several solutions are possible. One approach might be to reduce manufacturing variations and tuning range requirements using more advanced fabrication tools; a second could be to use photonic device and network designs with flexible wavelength registration. A third solution would be to develop novel tuning structures with significantly lower power consumption and low optical loss. This might be done either with dramatic improvements over existing structures or by employing novel tuning mechanisms. Another approach, discussed next, could be to develop broadband optical devices that may not need tuning.

D. Group IV Electroabsorption Devices

Another class of optoelectronic devices is based on the electroabsorption effects in group IV materials. In this section, we consider their performance as detectors and modulators for the macrochip. The primary device geometry chosen for analysis is waveguide-based although other geometries may prove feasible, such as surface normal device operation combined with OPxC. To gauge the overall strength of the electro-optic coefficient for group IV materials, one can examine the absorption coefficient of various materials. Most notable is the strong direct band edge of germanium that lies at 1530 nm with a strength of 6000 cm^{-1}. This strength is comparable to III–V materials. Further, it has more than three orders of magnitude stronger absorption than the plasma dispersion effect used in Mach–Zehnder or microring modulators. This suggests that group IV materials, when used in electroabsorption geometries, can have significantly large electro-optic coefficients that can enable compact low-voltage devices. We note that silicon has its corresponding direct edge well into the deep ultraviolet and hence silicon-compatible electroabsorption devices will most likely require significant germanium content to make C-band wavelength operation feasible.

A number of efforts have led to successful germanium waveguide detectors [47]–[50], and even to compact Ge detectors fully integrated into a CMOS process flow [51]. Many of these were integrated into waveguides on SOI enabling compact footprints and low capacitance, with responsivities that ranged 0.5 to 1.0 A/W and speeds up to 40 Gbps. One issue is the considerable variance in the observed dark currents ranging from 100 nA to 10 μA. This appears to be related to the quality of the germanium epitaxy on SOI. In macrochip applications, the germanium detector quality must insure a sufficiently low dark current detector under bias to enable the CMOS receiver to reach sensitivities as low as −21 dBm. A second issue is that the variation in the responsivity appears to be associated with insertion loss related to the integration into waveguides. This can likely be solved by appropriate design of the optical mode coupling into the detector taking into account the mode profile and the refractive index mismatch.

A concern for germanium-based modulators arises from the indirect absorption tail of the material that causes unwanted insertion loss and complicates the design of waveguide-based modulator geometries. Nevertheless, these devices have the potential for broadband operation and several noteworthy device concepts should be recognized. One particularly interesting device is based on the Franz–Keldysh (FK) effect [52]. These devices have been studied as modulators in bulk germanium and early results on integrated devices on SOI as thin layers in a waveguide geometry have been recently reported [53]. An optimized FK device could use an electric field of approximately 5 V/μm to achieve an on–off absorption difference $\delta\alpha$ on the order of 400 cm^{-1} over a 20 nm or greater wavelength range. A 40 μm long, 0.5 μm wide modulator could then achieve an extinction ratio of about 7 dB and an insertion loss of 3.5 dB not including coupling-related penalties. Based on these characteristics, the capacitance of the device could be 30 fF and its resistance 200 Ω, giving it ample bandwidth and a switching energy on the order of 50 fJ/bit. In this case, another factor in the transmitter power arises from the absorbed photocurrent in the biased modulator off-state [the last term in (4)]. Such a device might consume an additional 45 fJ/bit due to this term. The germanium FK modulator operates red-shifted from the germanium detector where its responsivity is lower.

Table 3 Expected Link Budget for Optical Components of the WDM Link

Component	Loss (dB)
Modulator	4
Waveguide on source site	1
Face-to-face coupler	1
Mux	2.5
Waveguide on SiP layers	2
Inter-layer coupler	1.2
Waveguide on destination	1
Drop filter, pass-through λs	0.1
Drop filter, dropped λ	1.5

Table 4 Energy per Bit for Optical Links in 2012–2015 and 2015–2018 Timeframes, Listed in fJ/bit

Component	2012-2015	2015-2018
Modulators and CMOS drivers	80	35
Detectors and CMOS receivers	120	65
Photon loss	70	50
WDM mux/demux and tuning	30	10

Another interesting electroabsorption device is based on the quantum-confined stark effect in a multiple quantum-well device based on Ge quantum wells separated by SiGe barriers [54]. Strong exitonic peaks with sharp absorption spectra have been observed, which may lead to devices that can achieve an on–off absorption difference $\delta\alpha$ on the order of 1000 cm^{-1} over a 20 nm wavelength range. Although the integration of multiple quantum well materials into a silicon photonics foundry is an open issue, such a device can potentially reduce both the switching and the tuning energies to the levels necessary to meet the long-range macrochip targets.

E. Roadmap for Silicon Photonic Interconnects

Current best-in-class SerDes transceivers are expected to yield signaling densities between 120–200 Gbps/mm^2. In the capacitive case, the electrical pad pitch may be on the order of 20 μm. Each pad can drive signals at line rates of 5 Gbps. This provides a potential communication density in excess of 5 Tbps/mm^2. Experimental capacitive proximity communication circuits have yielded areal densities up to 430 Gbps/mm^2 to date. In the optical case, an optical coupler can be as small as 20 μm on a side. The optical coupler may communicate many wavelength-multiplexed channels (e.g., as few as eight or as many as 64 may be envisaged with current technologies), with each channel operating at line rates of 10 Gbps and larger. The assumption of eight wavelength channels at 20 Gbps per channel with an optical coupler pitch of 35 μm results in a potential communication density of 128 Tbps/mm^2. In both the capacitive coupling and the optical proximity case, the engineering limits to signal density will result from the area and power of the transmitter and receiver circuits.

Commercial photonic links based on vertical-cavity surface-emitting lasers (VCSELs) as well as those based on silicon photonics currently operate at tens of picojoules per bit and higher. As discussed in the previous sections and summarized in Tables 3 and 4, the energy-per-bit of a silicon photonic link is a key metric that must be reduced by over two orders of magnitude to achieve the many benefits of the macrochip. Fig. 23 depicts a potential roadmap for the evolution of the energy per bit of a silicon photonic interconnect over the next decade and also shows, as a baseline for comparison, state-of-the-art link energies for SerDes [39] and VCSELs [46] (excluding clock and data recovery). From Fig. 23 and the discussion in Section VII-C, it is clear that the tuning energy for the silicon photonic devices such as the

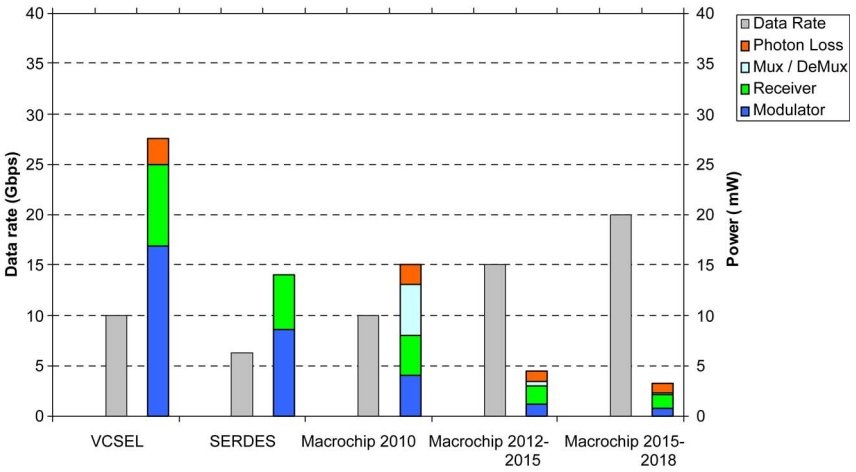


Fig. 23. *Roadmap for power and data-rate in the 2010–2018 timeframe for on-macrochip optical links versus best-in-class research results to date for SerDes [39] and VCSEL [46] links. The ratio of the second column to the first is the energy-per-bit of the interconnect.*

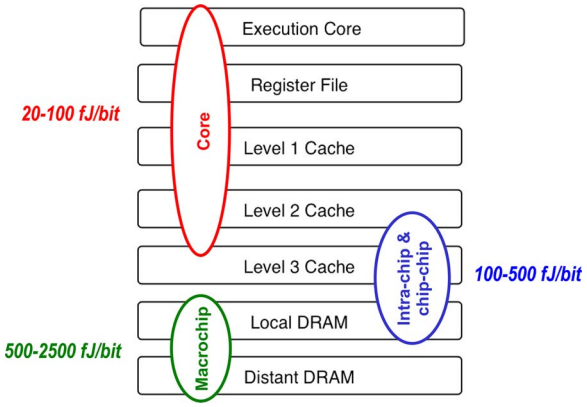


Fig. 24. ***Processor-to-memory interconnect hierarchy on the macrochip and the energy-per-bit requirements for silicon photonic interconnect for consideration as a replacement for wires in the hierarchy.***

modulator, the mux, and the demux can be expected to constitute a significant fraction of the energy per bit in the near term. However, these can be expected to reduce with improved device design, better manufacturing processes, and perhaps with the broadband materials reviewed in Section VII-D. As the switching, tuning, and circuit energies diminish, the photon loss becomes significant, and a low-loss link will ultimately be needed to hit the final targets.

The penetration of silicon photonic interconnects into the computing systems hierarchy as a replacement for wires will, in our opinion, be directly related to the energy of the optical link versus the distance of the interconnect. In Fig. 24, we show the interconnect hierarchy with the corresponding target energy-per-bit requirements for silicon photonic interconnects. In this paper, we considered the use of optical links commensurate with the macrochip and chip-to-chip levels of this hierarchy. Consideration for even deeper penetration of optical links will likely have technology, integration, and packaging consequences beyond the scope of this paper.

VIII. CONCLUSION

In this paper, we presented a canonical macrochip computing system, described its benefits, analyzed the constituent optical component and system requirements, and provided an overview of the requirements for the critical technologies needed to fulfill this system vision. We described the macrochip from an architecture and system viewpoint and quantified single-node and multinode macrochip performance in absolute terms and in relation to electrically interconnected systems. The macrochip utilizes optical proximity communication and the energy, density, and latency advantages of wavelength-division multiplexed optical links to allow "fat" compute nodes that enable rich, highly interconnected topologies (such as all-to-all connections) even when scaling up to a multinode supercomputer. We described a nonblocking, point-to-point WDM routing network that has superior performance and no setup delays when compared to an electrically controlled, packet-switched network of the same bandwidth. This improvement is particularly evident as the loading of the network goes up and also as the message size goes down. It further simplifies the control of the network and eliminates the resulting power required for network resource arbitration. The static WDM nonblocking network topology provides efficient transport for small messages (64 B or less), an important characteristic for supporting shared memory machines or HPC challenge benchmarks that make use of small messages. This network topology also favors embedded machines where the requirement is to maximize performance-per-watt on specific HPC metrics such as GUPS/watt and FFT/watt. We note that such small message support can also simplify programming although the exposition of this point is beyond the scope of this paper.

The development of the macrochip calls for a hundredfold to a thousandfold reduction in energy to communicate an optical bit of information, thereby enabling silicon photonic interconnects to transition to the intrachip stage and provide systems level benefits exceeding those offered by scaled electrical technologies. To meet this aggressive challenge, it is likely that a new class of photonic components as well as a comprehensive design toolkit for photonics will be needed. It is also crucial that traditional power-hungry analog optoelectronic interfaces be replaced by low-voltage low-energy circuit architectures and families that match electrical datapaths to complementary optical interfaces and take advantage of the small physical extent of the macrochip. Although we stress the performance aspects of optical links vis-à-vis density and energy requirements, we note finally that the ultimate choice of silicon photonic interconnects versus wires on the macrochip will also depend on other factors including manufacturability, reliability, and cost. ■

Acknowledgment

The authors thank Dr. J. Shah of Darpa MTO for his inspiration and support for this program. The authors gratefully acknowledge many valuable suggestions, technical guidance, and support from Dr. J. Mitchell, Dr. R. Drost, Dr. D. Cohen, Dr. M. McCracken, and Dr. K. Raj. Special thanks to Dr. B. O'Krafka for his contributions to the macrochip architecture and to A. Charlesworth for providing data on HPC systems. The authors also acknowledge valuable discussions and insights provided by Dr. C. Gunn and Dr. T. Pinguet of Luxtera, Dr. M. Asghari and Dr. D. Feng of Kotura, Prof. D. A. B. Miller of Stanford, and Prof. Y. Fainman of the University of California, San Diego.

REFERENCES

[1] J. W. Goodman, F. J. Leonberger, S.-Y. Kung, and R. A. Athale, "Optical interconnections for VLSI systems," *Proc. IEEE*, vol. 72, pp. 850–866, Jul. 1984.

[2] D. A. B. Miller, "Rationale and challenges for optical interconnects to electronic chips," *Proc. IEEE*, vol. 88, pp. 728–749, Jun. 2000.

[3] A. V. Krishnamoorthy, "The intimate integration of photonics and electronics," in *Advances in Information Optics and Photonics*, A. T. Friberg and R. Dändliker, Eds., SPIE Press, 2008, pp. 581–598.

[4] H. S. Hinton, T. J. Cloonan, J. McCormick, F. B., A. L. Lentine, and F. A. P. Tooley, "Free-space digital optical systems," *Proc. IEEE*, vol. 82, no. 11, pp. 1632–1649, Nov. 1994.

[5] C. Cook, J. E. Cunningham, A. Hargrove, G. G. Ger, K. W. Goossen, W. Y. Jan, H. H. Kim, R. Krause, M. Manges, M. Morrissey, M. Perinpanayagam, A. Persaud, G. J. Shevchuk, V. Sinyansky, and A. V. Krishnamoorthy, "A 36-channel parallel optical interconnect module based on optoelectronics-on-VLSI technology," *IEEE J. Sel. Topics Quantum Electron.*, vol. 9, pp. 387–399, Mar./Apr. 2003.

[6] L. Schares, J. A. Kash, F. E. Doany, C. L. Schow, C. Schuster, D. M. Kuchta, P. K. Pepeljugoski, J. M. Trewhella, C. W. Baks, R. A. John, L. Shan, Y. H. Kwark, R. A. Budd, P. Chiniwalla, F. R. Libsch, J. Rosner, C. K. Tsang, C. S. Patel, J. D. Schaub, R. Dangel, F. Horst, B. J. Offrein, D. Kucharski, D. Guckenberger, S. Hegde, H. Nyikal, C. K. Lin, A. Tandon, G. R. Trott, M. Nystrom, D. P. Bour, M. R. T. Tan, and D. W. Dolfi, "Terabus: Terabit/second-class card-level optical interconnect technologies," *IEEE J. Sel. Topics Quantum Electron.*, vol. 12, pp. 1032–1044, Sep./Oct. 2006.

[7] R. Soref and J. Larenzo, "All-silicon active and passive guided-wave components for $\lambda = 1.3$ and 1.6 μm," *IEEE J. Quantum Electron.*, vol. QE-22, pp. 873–879, Jun. 1986.

[8] R. Soref and B. Bennett, "Electrooptical effects in silicon," *IEEE J. Quantum Electron.*, vol. QE-23, pp. 123–129, Jan. 1987.

[9] C. K. Tang, A. K. Kewell, G. T. Reed, A. G. Rickman, and F. Namavar, "Development of a library of low-loss silicon-on-insulator optoelectronic devices," *Proc. Inst. Elect. Eng. Optoelectron.*, vol. 143, pp. 312–315, Oct. 1996.

[10] T. Bestwick, "ASOC—A silicon-based integrated optical manufacturing technology," in *Proc. 48th IEEE Electron. Compon. Technol. Conf.*, Seattle, WA, May 1998, pp. 566–571.

[11] B. E. Little, J. S. Foresi, G. Steinmeyer, E. R. Thoen, S. T. Chu, H. A. Haus, E. P. Ippen, L. C. Kimerling, and W. Greene, "Ultra-compact si-sio2 microring resonator optical channel dropping filters," *IEEE Photon. Technol. Lett.*, vol. 10, pp. 549–551, Apr. 1998.

[12] R. A. Soref, "Silicon-based optoelectronics," *Proc. IEEE*, vol. 81, pp. 1687–1706, Dec. 1993.

[13] L. Kimerling, "Silicon microphotonics: The next killer technology," in *Towards the First Silicon Lasers*, Pavesi, Ed. Norwell, MA: Kluwer, 2003, pp. 465–476.

[14] B. Jalali and S. Fathpour, "Silicon photonics," *J. Lightw. Technol.*, vol. 24, pp. 4600–4615, Dec. 2006.

[15] R. Soref, "The past, present, and future of silicon photonics," *IEEE J. Sel. Topics Quantum Electron.*, vol. 12, pp. 1678–1687, Nov./Dec. 2006.

[16] C. Gunn, "CMOS photonics for high-speed interconnects," *IEEE Micro*, vol. 26, pp. 58–66, Mar./Apr. 2006.

[17] L. Liao, D. Samara-Rubio, M. Morse, A. Liu, D. Hodge, D. Rubin, U. Keil, and T. Franck, "High speed silicon Mach-Zehnder modulator," *Opt. Express*, vol. 13, no. 8, pp. 3129–3135, Apr. 2005.

[18] Q. Xu, S. Manipatruni, B. Schmidt, J. Shakya, and M. Lipson, "12.5 gbit/s carrier-injection-based silicon micro-ring silicon modulators," *Opt. Express*, vol. 15, no. 2, pp. 430–436, Jan. 2007.

[19] A. Narasimha, B. Analui, Y. Liang, T. J. Sleboda, S. Abdalla, E. Balmater, S. Gloeckner, D. Guckenberger, M. Harrison, R. G. M. P. Koumans, D. Kucharski, A. Mekis, S. Mirsaidi, D. Song, and T. Pinguet, "A fully integrated 4 × 10-gb/s DWDM optoelectronic transceiver implemented in a standard 0.13 micron CMOS SOI technology," *IEEE J. Solid-State Circuits*, vol. 42, pp. 2736–2744, Dec. 2007.

[20] B. G. Lee, X. Chen, A. Biberman, X. Liu, I.-W. Hsieh, C.-Y. Chou, J. I. Dadap, F. Xia, W. M. J. Green, L. Sekaric, Y. A. Vlasov, R. M. Osgood, and K. Bergman, "Ultrahigh-bandwidth silicon photonic nanowire waveguides for on-chip networks," *IEEE Photon. Technol. Lett.*, vol. 20, pp. 398–400, Mar. 2008.

[21] R. G. Beausoleil, P. J. Kuekes, G. S. Snider, S.-Y. Wang, and R. S. Williams, "Nanoelectronic and nanophotonic interconnect," *Proc. IEEE*, vol. 96, pp. 230–247, Feb. 2008.

[22] M. Petracca, B. G. Lee, K. Bergman, and L. P. Carloni, "Design exploration of optical interconnection networks for chip multiprocessors," in *Proc. 16th IEEE Symp. High Perform. Interconnects*, Stanford, CA, Aug. 2008, pp. 31–40.

[23] A. Hadke, T. Benavides, S. J. B. Yoo, R. Amirtharajah, and V. Akella, "OCDIMM: Scaling the DRAM memory wall using WDM based optical interconnects," in *Proc. 16th IEEE Symp. High Perform. Interconnects*, Stanford, CA, Aug. 2008, pp. 57–63.

[24] R. G. Beausoleil, J. Ahn, N. Binkert, A. Davis, D. Fattal, M. Fiorentino, N. P. Jouppi, M. McLaren, C. M. Santori, R. S. Schreiber, S. M. Spillane, D. Vantrease, and Q. Xu, "A nanophotonic interconnect for high-performance many-core computation," in *Proc. 16th IEEE Symp. High Perform. Interconnects*, Stanford, CA, Aug. 2008, pp. 182–189.

[25] C. Batten, A. Joshi, J. Orcutt, A. Khilo, B. Moss, C. Holzwarth, M. Popovic, H. Li, H. Smith, J. Hoyt, F. Kartner, R. Ram, V. Stojanovic, and K. Asanovic, "Building manycore processor-to-DRAM networks with monolithic silicon photonics," in *Proc. 16th IEEE Symp. High Perform. Interconnects*, Stanford, CA, Aug. 2008, pp. 21–30.

[26] G. Moore, "Cramming more components onto integrated circuits," *Electron.*, vol. 38, no. 8, pp. 114–117, Apr. 1965.

[27] M. Horowitz, E. Alon, D. Patil, S. Naffziger, R. Kumar, and K. Bernstein, "Scaling, power, and the future of CMOS," in *Proc. IEEE Int. Electron Devices Meeting*, Dec. 2005.

[28] A. Chow, D. Hopkins, R. Drost, and R. Ho, "Exploiting capacitance in high-performance computer systems," in *Proc. IEEE Int. Symp. VLSI Design, Automat. Test*, Hsinchu, Taiwan, Apr. 2008, pp. 55–58.

[29] R. Ho, K. W. Mai, and M. A. Horowitz, "The future of wires," *Proc. IEEE*, vol. 89, pp. 490–504, Apr. 2001.

[30] X. Zheng, J. Lexau, J. Bergey, J. Cunningham, R. Ho, R. Drost, and A. V. Krishnamoorthy, "Optical proximity communication using reflective mirrors," *Opt. Express*, vol. 16, no. 19, pp. 15 052–15 058, Sep. 2008.

[31] D. Hopkins, A. Chow, R. Bosnyak, B. Coates, J. Ebergen, S. Fairbanks, J. Gainsley, R. Ho, J. Lexau, F. Liu, T. Ono, J. Schauer, I. Sutherland, and R. Drost, "Circuit techniques to enable 430 gb/s/mm^2 proximity communication," in *Dig. Tech. Papers IEEE Int. Solid-State Circuits Conf.*, San Francisco, CA, Feb. 2007, pp. 368–609.

[32] J. S. Orcutt, K. Anatol, M. A. Popovic, C. W. Holzwarth, B. Moss, H. Li, M. S. Dahlem, T. D. Bonifield, F. X. Kartner, E. P. Ippen, J. L. Hoyt, R. J. Ram, and V. Stojanovic, "Demostration of an electronic photonic integrated circuit in a commercial scaled bulk cmos process," in *Proc. CLEO/IQEC Conf. Lasers Electro Opt./Int. Quantum Electron. Conf.*, May 2008.

[33] K. Datta, D. Bonachea, and K. Yelick, "Titanium performance and potential: An NPB experimental study," in *Proc. 18th Int. Workshop Lang. Compilers Parallel Comput.*, Oct. 2005, pp. 200–214.

[34] J. Dunigan, T. H., J. S. Vetter, I. White, J. B., and P. H. Worley, "Performance evaluation of the cray x1 distributed shared-memory architecture," *IEEE Micro*, vol. 25, pp. 30–40, Jan./Feb. 2005.

[35] D. Lilya, "Cache coherence in large-scale shared-memory multiprocessors: Issues and comparisons," *ACM Comput. Surv.*, vol. 25, no. 3, pp. 303–338, Sep. 1993.

[36] L. Hammond, V. Wong, M. Chen, B. D. Carlstrom, J. D. Davis, B. Hertzberg, M. K. Prabhu, H. Wijaya, C. Kozyrakis, and K. Olukotun, "Transactional memory coherence and consistency," in *Proc. 31st Annu. Int. Symp. Comput. Architect.*, Jun. 2004, pp. 102–113.

[37] A. V. Krishnamoorthy, J. E. Cunningham, X. Zheng, I. Shubin, J. Simons, D. Feng, H. Linag, C. C. Kung, and M. Asghari, "Optical proximity communication with passively aligned silicon photonic chips," *IEEE J. Quantum Electron.*, vol. 45, pp. 409–414, Apr. 2009.

[38] X. Zheng, P. Koka, H. Schwetman, J. Lexau, R. Ho, J. E. Cunningham, and A. V. Krishnamoorthy, "Silicon photonic WDM point-to-point network for multi-chip processor interconnects," in *Proc. 5th IEEE Int. Conf. Group IV Photon.*, Sep. 2008, pp. 380–382.

[39] J. Poulton, R. Palmer, A. M. Fuller, T. Greer, J. Eyles, W. J. Dally, and M. Horowitz, "A 14-mw 6.25-gb/s transceiver in 90-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 42, pp. 2745–2757, Dec. 2007.

[40] R. Drost, C. Forrest, B. Guenin, R. Ho, A. V. Krishnamoorthy, D. Cohen, J. E. 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*, Aug. 2005, pp. 13–22.

[41] A. Widmer and P. Franaszek, "A DC-balanced, partioned-block 8B/100B

transmission code," *IBM J. Res. Develop.*, vol. 27, no. 5, pp. 441–451, Aug. 1983.

[42] F. Xia, L. Sekaric, and Y. Vlasov, "Ultra-compact optical buffers on a silicon chip," *Nature Photon.*, vol. 1, no. 1, p. 65, Jan. 2007.

[43] Q. Xu, B. Schmidt, J. Shakya, and M. Lipson, "Cascaded silicon microring modulators for WDM optical interconnection," *Opt. Express*, vol. 14, no. 20, pp. 9430–9435, Oct. 2006.

[44] S. Selvaraja, W. Bogaerts, D. Van Thourhout, and R. Baets, "Fabrication of uniform photonic devices using 193 nm optical lithography in silicon-on-insulator," in *Proc. 14th Eur. Conf. Integr. Opt.*, Eindhoven, The Netherlands, Jun. 2008, pp. 359–362.

[45] X. Wang, J. A. Martinez, M. S. Nawrocka, and R. R. Panepucci, "Compact thermally tunable silicon wavelength switch: Modeling and characterization," *IEEE Photon. Technol. Lett.*, vol. 20, pp. 936–938, Jun. 2008.

[46] C. Kromer, G. Sialm, C. Berger, T. Morf, M. Schmatz, F. Ellinger, D. Erni, G. Bona, and H. Jackel, "A 100-mW 410 Gb/s transceiver in 80-nm CMOS for high-density optical interconnects," *IEEE J. Solid-State Circuits*, vol. 40, pp. 2667–2679, Dec. 2005.

[47] J. Liu, J. Michel, W. Giziewicz, D. Pan, K. Wada, D. Cannon, S. Jongthammanurak, T. Danielson, L. Kimerling, J. Michel, J. Chen, and F. Kärtner, "High-performance, tensile-strained Ge p-i-n photodetectors on a Si platform," *Appl. Phys. Lett.*, vol. 87, p. 103501, Aug. 2005.

[48] L. Colace, M. Balbi, G. Masini, G. Assanto, H.-C. Luan, and L. Kimerling, "Ge on Si p-i-n photodiodes operating at 10 Gbit/s," *Appl. Phys. Lett.*, vol. 88, p. 101111, Mar. 2006.

[49] T. Yin, R. Cohen, M. Morse, G. Sarid, Y. Chetrit, D. Rubin, and M. Paniccia, "31 GHz Ge n-i-p waveguide photodetectors on silicon-on-insulator substrate," *Opt. Express*, vol. 15, no. 21, pp. 13 965–13 971, Oct. 2007.

[50] D. Suh, S. Kim, J. Joo, G. Kim, and I. Kim, "35 GHz Ge p-i-n photodetectors implemented using RPCVD," in *Proc. 5th IEEE Int. Conf. Group IV Photon.*, Sep. 2008, pp. 191–193.

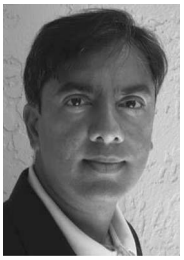
[51] T. Pinguet, B. Analui, E. Balmater, D. Guckenberger, M. Harrison, R. Koumans, D. Kucharski, Y. Liang, G. Masini, A. Mekis, S. Mirsaidi, A. Narasimha, M. Peterson, D. Rines, V. Sadagopan, S. Sahni, T. Sleboda, D. Song, Y. Wang, B. Welch, J. Witzens, J. Yao, S. Abdalla, S. Gloeckner, P. De Dobbelaere,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

[52] A. Frova and P. Handler, "Franz-Keldysh effect in the space-charge region of a germanium p-n junction," *Phys. Rev.*, vol. 137, no. 6A, pp. A1857–A1861, Mar. 1965.

[53] J. Liu, M. Beals, A. Pomerene, S. Bernardis, R. Sun, J. Cheng, L. Kimerling, and J. Michel, "Ultralow energy, integrated GeSi electroabsorption modulators on SOI," in *Proc. 5th IEEE Int. Conf. Group IV Photon.*, Sep. 2008, pp. 10–12.

[54] Y.-H. Kuo, Y. Lee, Y. Ge, S. Ren, J. Roth, T. Kamins, D. A. B. Miller, and J. Harris, "Quantum-confined stark effect in Ge/SiGe quantum wells on Si for optical modulators," *IEEE J. Sel. Topics Quantum Electron.*, vol. 12, pp. 1503–1513, Nov./Dec. 2006.

ABOUT THE AUTHORS



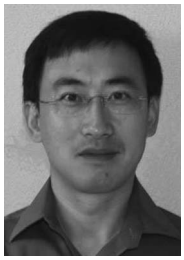
Ashok V. Krishnamoorthy (Member, IEEE) is a Distinguished Engineer and Director with the Sun Microsystems Physical Sciences Center, San Diego, CA, and Principal Investigator for Sun's photonics R&D. Previously, he was with AraLight as its President and CTO as part of one of Lucent's first optical communications spinouts. Prior to that, he was a member of Technical Staff with the Advanced Photonics Research Department, Bell Labs. He has authored six book chapters and more than 160 technical publications. He has received 40 patents and has presented 60 invited conference talks.

Dr. Krishnamoorthy is a member of Tau Beta Pi, Eta Kappa Nu, and Sigma Xi. He has received several individual and team awards, including the Eta Kappa Nu young electrical engineer award, the IEEE LEOS Distinguished Lecturer award, the 2004 ICO International Prize in Optics, and the Chairman's Award for Innovation from Sun Microsystems. He serves on the Technical Advisory Board for several optical technology startups.

Ron Ho (Senior Member, IEEE) received the Ph.D. degree from Stanford University, Stanford, CA, in 2002.

He is a Distinguished Engineer and Director with Sun Microsystems, working in the Sun Labs VLSI Research Group. From 1993 to 2003, he was with Intel Corporation, Santa Clara, CA, working on the Pentium II and later the third-generation Itanium processors. In 2003, he joined Sun Labs, Menlo Park, CA, where he has been working on chip-to-chip and on-chip communication technologies, memory design, and asynchronous circuits. He has authored or coauthored more than 40 technical papers in peer-reviewed conferences and journals and has received 20 U.S. patents. He is a consulting Assistant Professor at Stanford University, Stanford, CA.

Dr. Ho is a member of Tau Beta Pi and Phi Beta Kappa. He received the Chairman's Award for Innovation from Sun in 2004. He is a former member of the Technical Program Committee of the IEEE International Solid-State Circuits Conference and is currently on program committees for the IEEE Asian Solid-State Circuits Conference, the IEEE Hot Interconnects Conference, the IEEE Symposium on Asynchronous Circuits and Systems, and the IEEE VLSI-DAT Conference. He was General Chair of the 2008 IEEE/LEOS Workshop on Interconnections Within High-Speed Digital Systems. He has served as Guest Associate Editor for the IEEE JOURNAL OF QUANTUM ELECTRONICS and the IEEE JOURNAL OF SOLID-STATE CIRCUITS.



Xuezhe Zheng (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in optical instruments from Tsinghua University, Beijing, China, in 1993 and 1997, respectively.

Prior to joining Sun Microsystems Inc., San Diego, CA, as a Senior Staff Engineer, he was with Calient Networks Incorporated, San Jose, CA, where he was a Manager of Optical Engineering working on three-dimensional MEM-based photonic switching and its application in wavelength-division multiplexing (WDM) networks. From 1997 to 1999, he was a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, University of California, San Diego, investigating high-speed, high-density free-space optical interconnects. He has extensive experiences in photonic switching and optical cross-connect, fiber-optic components, dense WDM optical networks, and optical interconnections. His current research interests are in WDM networking and Si photonics for advanced inter/intrachip interconnects. He has published more than 20 papers in technical journals and has received seven U.S. patents.

Dr. Zheng received the Science and Technology Development Award from the National Education Committee of China.

Herb Schwetman (Member, IEEE) received the Ph.D. degree in computer sciences from The University of Texas (UT) at Austin in 1970.

He has been a Senior Staff Engineer with the Computer Architecture and Performance group, Sun Labs, a division of Sun Microsystems, Inc., since 2004. He joined Sun in 2001. In 1994, he founded Mesquite Software to develop, market and support CSIM, a toolkit for building system simulation models. From 1984 to 1994, he was a Senior Member of Technical Staff at MCC, an R&D consortium in Austin, TX. While at MCC, he worked in the parallel processing and parallel database programs. He was a Professor in the Computer Sciences Department, Purdue University, West Lafayette, IN, from 1972 to 1984. He has served as an Adjunct Professor in the Computer Science Department, UT Austin since 1984, teaching courses in computer architecture and systems modeling. He was Chairman of ACM/SIGMETRICS (1980–1984), Department Editor of *Communications of ACM* (1979–1982), Cochair of the International Conference on Parallel Processing (1991), and a Member of the Board of Directors of the Winter Simulation Conference (1990–2001). In summer 1975, he was a Visiting Scientist at CERN, Geneva, Switzerland. During fall 1979, was a Fullbright/Hayes Lecturer in the Department of Computer Science, University of Helsinki, Finland.

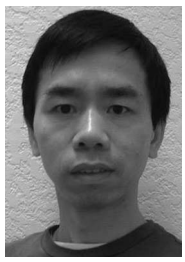
Dr. Schwetman is a member of the IEEE Computer Society and ACM.

Jon Lexau (Member, IEEE) received the B.S. degree in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, NY, in 1989 and the M.S. degree in electrical engineering from Stanford University, Stanford, CA, in 1994.

From 1989 to 1993, he was with Amdahl Corporation designing high-performance mainframe CPUs, working on chips implementing the timer register facility, instruction unit microcode, and high-level architectural specification. He joined Sun Labs in 1994 with the Asynchronous Systems Group and followed its evolution into the current VLSI Research Group. His responsibilities ranged from circuit design to chip layout, PC board design, and chip testing. His most recent work includes contributing to a series of capacitively coupled proximity communications experiments with the VLSI group and electrical transceivers for optical interconnect.

Pranay Koka received the M.S. degree in electrical and computer engineering from the University of Wisconsin, Madison, in 2005 and the M.S. degree in electrical engineering from Southern Illinois University, Edwardsville, in 2002.

He has been with the Computer Architecture and Performance group, Sun Labs, a division of Sun Microsystems, Inc., since 2005.



GuoLiang Li (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of California, San Diego, in 2002.

He is a Senior Staff Engineer with the Physical Sciences Research Center, Sun Microsystems, San Diego, working on the Ultrascale Nanophotonic Intrachip Communication project. In January 2001, he joined the Semiconductor Research Department, Bell Labs, Breinigsville, PA, which was subsequently spun off with Agere Systems, working as Lead Designer for 40 Gb/s InP Mach–Zehnder modulators and 40 Gb/s electroabsorption modulated lasers. In 2003, he joined Luxtera, Carlsbad, CA, where he worked on the development of electronic and photonic integrated circuits and successfully developed the world's first 10 Gb/s Si optical modulator on CMOS fabrication platform. In 2006 and 2007, he was with the Optical Platform Division, Intel Corporation, Fremont, CA, working on SFP+ and X2 transceiver development for 10GBASE-LRM and 10GBASE-LR applications. He has coauthored two book chapters and more than 30 technical papers published in peer-reviewed journals and conferences and has received ten U.S. patents. He is an External Reviewer for the Research Grants Council of Hong Kong.

Dr. Li has served on the Technical Program Committee of the IEEE Microwave Photonics Conference.

Ivan Shubin (Member, IEEE) received the M.S. degree in solid-state and semiconductor physics from the Moscow Institute of Physics and Technology, Moscow, Russia, in 1988 and the Ph.D. degree in electrical engineering from the Center for Research and Education in Optics and Lasers, School of Optics, University of Central Florida, Orlando, in 2001.

From February 2001 to August 2002, he was a Senior Engineer with Optium Corporation. From 2002 until 2007, he was a Project Scientist with the Electrical and Computer Engineering Department, University of California, San Diego. Since 2007, he has been a Senior Member of Technical Staff with Sun Microsystems, San Diego. His research interests include advanced packaging solutions and platforms for electronic, optoelectronic and MEMS applications, wafer-scale packaging, three-dimensional integration, and novel photonic components. He has authored or coauthored more than 30 technical papers and conference presentations.

John E. Cunningham received the B.S. degree from the University of Tennessee at Knoxville and the M.S. and Ph.D. degrees from the University of Illinois at Urbana-Champaign, all in physics.

He is a Distinguished Engineer and Director with Sun Microsystems and a veteran research scientist with more than 25 years of university, Bell Labs, and startup experience in the area of optoelectronic and semiconductor materials and devices used within optical networks. Since joining Sun, he has led packaging initiatives to develop interchip proximity communication and worked on Si nanophotonics solutions for data communications within computers. Before joining Sun Microsystems, he was Chief Scientist with Aralight, where he developed products based on the integration of vertical cavity surface-emitting lasers and photodetectors with CMOS, a technology spun out of Bell Laboratories. While with Bell Laboratories, he also pioneered eight world records on various types of quantum mechanically engineered nanodevices and materials and coauthored more than 360 journal papers as well received 30 U.S. patents. Before joining Bell Laboratories, has was a Member of the Research Faculty in the Physics Department, University of Illinois, where he pioneered metals molecular beam epitaxy.

Dr. Cunningham won the Chairman's Award at Sun Microsystems.