Optical Interconnects for High-Performance Computing

Marc A. Taubenblatt, *Member, IEEE*

(Invited Tutorial)

***Abstract*—High-performance computing systems are of steadily growing interest to provide new levels of computational capability for an increasing range of applications. The growing use of and dependence on optical interconnects to meet these system's scaling bandwidth demands has given rise to "computercom" as a distinct market segment, alongside the traditional datacom and telecom markets. This paper discusses the trends, requirements, tradeoffs, and potential technologies for this market.**

***Index Terms*—Computer networks, high-performance computing (HPC), optical interconnections, supercomputing.**

I. Introduction

HIGH-PERFORMANCE computing (HPC) systems are of steadily growing interest, not just for "one of" government systems, but to provide commercial industry with new levels of computational capability. These could range from geophysical data processing to drug discovery to multiscale modeling to environment and climate modeling to the analysis of huge datasets made available through our increasingly connected world. The growing use of and dependence on optical interconnects to meet these system's scaling bandwidth (BW) demands has given rise to "computercom" as a distinct market segment, alongside the traditional datacom and telecom markets. The very high aggregate BW demands of these systems have opened up opportunities for optics to compete with electrical interconnects at shorter and shorter distances.

What distinguishes the computercom market is not only the need for very short, less than 10 m interconnect lengths, but also an enormous pressure for reduced cost, power, and reliability to meet the demands of future systems. Power considerations require placement of optics very close to the signal source, which in turn is driving the need for very high density interconnects and highly integrated packaging as well.

Manuscript received July 11, 2011; revised September 30, 2011; accepted October 07, 2011. Date of publication October 20, 2011; date of current version February 02, 2012. This work was supported in part by the Defense Advanced Research Projects Agency under Contract HR0011-07-9-0002, Contract HR0011-08-C-0102, and Contract MDA972-03-3-0004A.

The author is with IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: tauben@us.ibm.com).

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

Digital Object Identifier 10.1109/JLT.2011.2172989

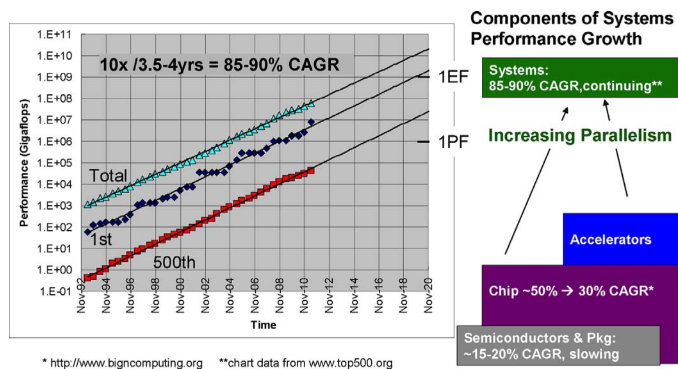


Fig. 1. HPC performance trends. (Left side) Historical and future trend based on Top500 data. (Right side) Additive components needed to make up the overall growth at the system level.

In order to achieve performance for future systems approaching the exaflop scale, a more holistic view of optimization across the system will be required, as well as new technologies and approaches. The system interconnects are no exception, and therefore, the optical interconnects for these systems must be engineered by considering a much larger scope of issues that arise in HPC systems in order to make the appropriate decisions and tradeoffs. With respect to optical interconnects, this will certainly require a much closer level of cooperation between system providers and optics suppliers.

II. HPC Trends

As shown in Fig. 1, over the last two decades, HPC has maintained a performance improvement trend of 85–90% compound annual growth rate, almost doubling performance every year. Historically, growth in transistor speed and improved CPU design (taking advantage of ever cheaper and smaller devices) provided a large part of this growth. As chip speeds have leveled off more recently, multicore architectures have helped fill the gap. Moving forward, new innovations will be required, such as the increased use of specialized computing elements or accelerators (e.g., graphics processing units), as well as a further increase in the reliance on continued growth in large scale system parallelism to make up the difference. Thus, interconnect BW requirements continue to scale at all physical levels of the system.

Typical server interconnects include a core-to-core bus on a single chip, chip to chip (for CPU-to-cache- or CPU-to-CPU communications), chip-to-memory on card, CPU-to-CPU cluster fabric between cards and racks, CPU node to storage

0733-8724/$26.00 © 2011 IEEE

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on September 11,2020 at 08:01:03 UTC from IEEE Xplore. Restrictions apply.

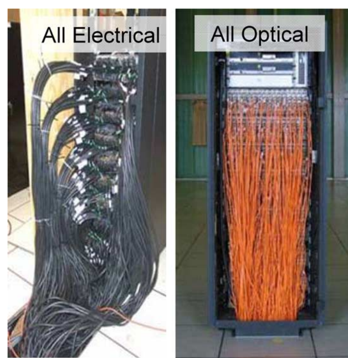


Fig. 2. IBM Federation switch rack for ASCI Purple. (Left side) Preproduction system with all electrical cabling. (Right side) Rack with all optical cabling.

(typically rack to rack), and LAN/WAN links (local/wide area networks) which go beyond the immediate computing or data center building [1]. Typically, CPU-to-memory links and symmetric multiprocessor links, a cache-coherent architecture, are the most latency sensitive, though much of that latency is due to the memory controller, with a smaller dependence on time of flight delays (about 5 ns/m for optical fiber).

Historically, storage links were the first to utilize optical interconnects (e.g., IBM's ESCON technology since 1990 [2]), in large part based on the longer distances from compute to storage racks and the relatively modest, and therefore low total cost, BW required (the need is limited by disk access speed). In 2005, IBM introduced the ASCI Purple system, which was one of the first systems to utilize optics for rack-to-rack cluster links. Initially, the system deployed all electrical links for rack-to-rack cluster interconnects; however, as the price of optics continued to drop during this period, the longer links were replaced with optics. This change relieved significant cable congestion in these systems, and is shown in Fig. 2.

More recently, the IBM Power 775 [3] has made use of a fiber cable optical backplane within the rack as well as for the rack-to-rack cluster fabric. In order to reap the further benefits of an all optical backplane, the optics modules are located on the same first level package as the router chip, in this case on a glass–ceramic multichip module (MCM). Fig. 3 shows one side of the large system card, with eight router chip MCMs and their associated optics. A single MCM contains 28 transmit and 28 receive modules, each of which have 12 channels running at 10 Gbps per channel. In the upper right inset, the underlying glass–ceramic substrate is shown with mounted router chip and with the land grid array pads for the (unmounted) optics modules. Each of the router MCMs is connected electrically to the microprocessor MCMs on the same card (not shown). By mounting the optics on the router MCM, interconnection BW is provided from both the bottom (electrical) and top (optical) of the MCM. As well, the signal path for the electrical link to the optical interconnects is improved.

For cluster fabric in particular, clever networking topologies can mitigate BW costs while maintaining reasonable performance, although not without tradeoffs. For example, mesh or torus networks can require many fewer interconnects, but will need more hops and, thus, have longer latency to pass data between more distant nodes. The IBM Power 775 supercomputer

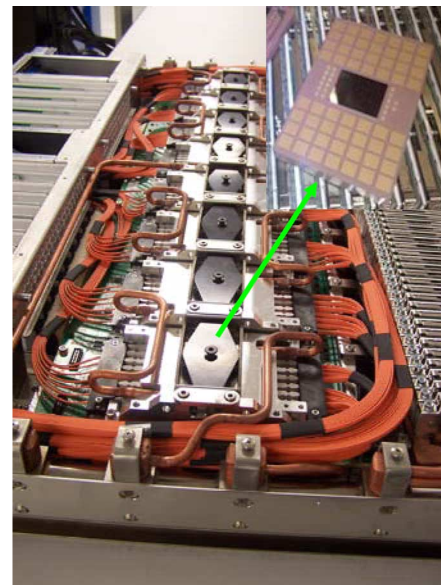


Fig. 3. IBM Power 775 system drawer showing eight router MCMs with integrated optical interconnects. Inset shows ceramic MCM with router chip and contact pads for optical modules.

is an example of a two stage all-to-all network (also known as a dragonfly) [4], [5], which is able to provide a low-latency, high-BW connectivity between random nodes in the system. This provides a high-performance network suitable for a large range of workloads and is shown in Fig. 4. Each node consists of four Power7 chips (eight cores each) on a quad-chip module. Then, 32 nodes are connected with an all-to-all network to form a supernode. Then each supernode in the system is further connected by a second level of all-to-all network.

At the other end of the spectrum, the IBM Blue Gene machine utilizes a torus network [6], shown in Fig. 5. Simply explained, a torus network consists of "nearest neighbor" interconnects that wrap around at the edges. To get to further away nodes requires multiple hops and, therefore, greater latency to go from node to node, but requires fewer interconnects and typically shorter interconnect distances. By further exploiting a 6-D torus network, the number one machine in June 2011, the Fujitsu K Computer, is able to satisfy interconnect requirements without using optics at all [7].

However, for a given topology, further scaling to higher performance will require BW scaling at each level of the packaging hierarchy causing bottlenecks: off chip, off module, off card, and rack to rack. For example, the first two Blue Gene machines (BG/L $\sim$0.6 TF machine and BGF/P $\sim$1 PF machine) both used electrical interconnect for the torus; however, the BG/Q machine (10's of PF) will use optical links for the torus to accommodate higher data rates.

Channel data rates for these off-chip interconnects have, therefore, been steadily increasing in response to system scaling needs, as this has been the best way to improve cost per transported bit, power per transported bit, and to meet BW density (wiring and area density) requirements. However, electrical interconnects become much more difficult to successfully design as data rates begin to exceed 10 Gbps, due to frequency dependent losses, crosstalk, and frequency resonance effects

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on September 11,2020 at 08:01:03 UTC from IEEE Xplore. Restrictions apply.

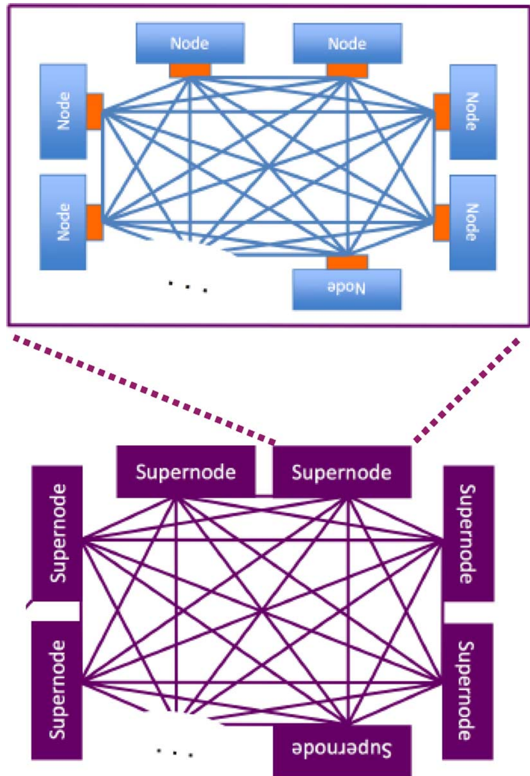


Fig. 4. IBM Power 775 two stage all-to-all network [4].

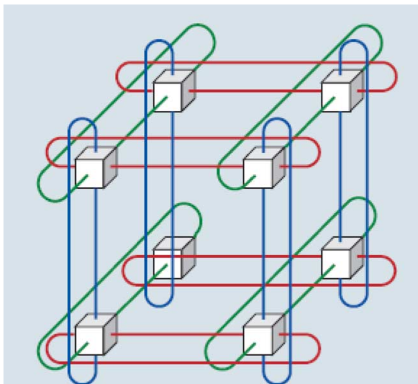


Fig. 5. Example of a 3-D Torus network such as that used in the IBM BlueGene L machine [6].

[8]. Fig. 6 shows a typical electrical backplane consisting of a set of cards plugged into a backplane. Copper traces suffer from increasingly large losses at higher frequencies due to skin effect (and dielectric losses to some degree). This can be mitigated by using fatter wiring, but this only exacerbates wiring density and routing problems. At each packaging or connector juncture, the line impedance must be well matched or signals will be partially reflected causing signal degradation. Via stubs, a consequence of typical drill and fill printed circuit board (PCB) technology also cause reflections, and although these can be eliminated with back drilling or stubless manufacturing methods, these generally increase the costs of the PCB boards. Multilevel signaling has also been considered; however, in our

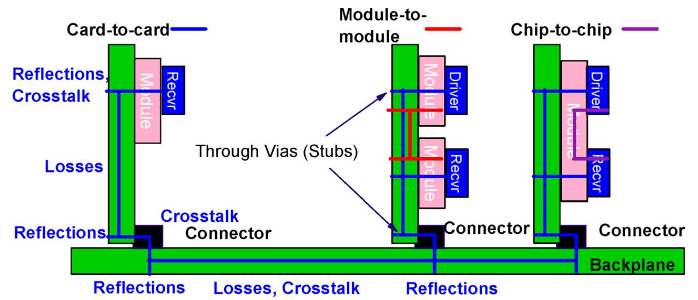


Fig. 6. Electrical links suffering from multiple signal integrity degradations as channel rates increase, including high losses, crosstalk, and reflections.

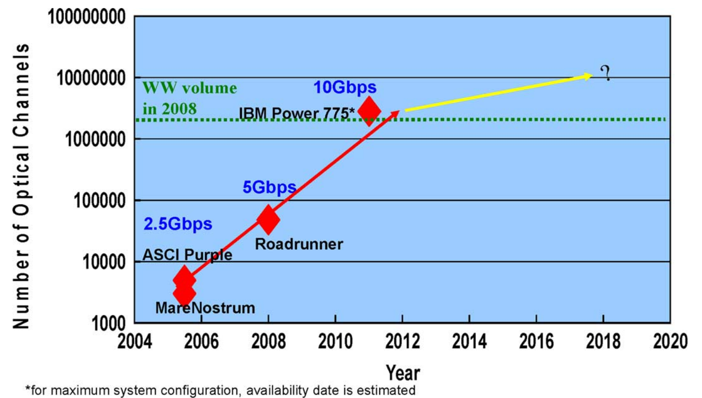


Fig. 7. HPC systems driving optics volumes.

analysis [8], single-bit per baud signaling still performs better even at 25 Gbps data rates.

Optical interconnects do not suffer such strong signal integrity degradations, and provide additional benefits, including reduced cable bulk, smaller connector size, and reduced electromagnetic interference.

Due to the benefits of optical interconnects, there has been a steadily increasing use of optics in these large systems, so much so, that the number of optical channels in a single HPC supercomputing system today can be on par with the worldwide volume in parallel optical interconnects just a few years ago. This trend is shown in Fig. 7. While the upward trend is very clear, it is less clear to what degree the number of optics links will be mitigated by the use of shorter distance topologies (e.g., mesh or torus) or thinner networks. The considerations will be the cost of optical interconnects versus copper interconnects and the resultant performance tradeoffs, i.e., the ability to create topologies coupled with algorithms which maintain performance despite thinner networks and longer latencies to more distant nodes.

III. Requirements and Tradeoffs

To respond to the severe requirements of future systems that are approaching one exaflop, optical interconnects will need to continue to make improvements in four major areas: power, cost, density, and reliability. To achieve multi-100 PF and beyond systems, optics power will have to be driven well below 10 mW/Gbps for unidirectional links, perhaps as low as 1 mW/Gbps, with costs below 10's of cents/Gbps or lower. These targets are a consequence of maintaining reasonable costs and

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on September 11,2020 at 08:01:03 UTC from IEEE Xplore. Restrictions apply.

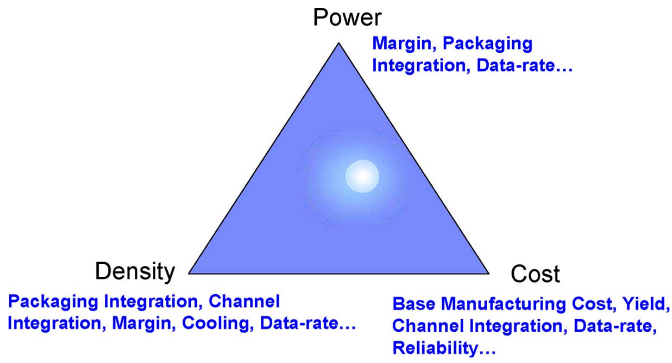


Fig. 8. Optimized solutions will require detailed analysis of tradeoffs.

power targets for these large machines. By assessing the total system BW requirements, one can determine the total optics cost and power in relation to the machine targets. For example, typical system architectures might require a BW of 0.1 to 1 B per flop (unidirectional), depending on the intended application set. Thus, a 100 PF system with a 0.2 B per flop network would require a 200 Pbps network (assuming a 10 bit byte for coding and redundancy). If half the network were optical links at a cost of \$0.50/Gbps and if a complete electro-optical link consumed 20 mW/Gbps, the optics alone would cost \$50 M and consume 2 MW. Numbers this large are not likely tolerable in a 100 PF system and will need to improve by at least an order of magnitude or more for a 1 EF system.

To achieve cost goals, careful rethinking of optical interconnects for low-cost high-volume manufacturing will be required. Furthermore, to achieve the lowest power links, optics modules will need to be situated close to the signal source, requiring very dense modules, of order 1 Tbps/cm^2. Achieving these goals will require careful balancing of tradeoffs, as shown in Fig. 8. For example, trimming power will narrow link margins, lowering yields, and, therefore, increasing costs. By locating optics closer to the signal source, electrical link power can be reduced, but optimization of optics packaging close to the signal source will require much denser and more integrated optical modules which will require greater cooperation between systems providers and optical interconnect suppliers.

A. Density

Fig. 9 shows examples of the SNAP12 form factor, commonly used in the mid 2000s compared with the Avago MicroPOD module, which needed to be accommodated directly on the MCM for the IBM Power 775 supercomputer. Future optical interconnects may require even denser optics, such as might be obtained by flip chip assemblies employing greater numbers of parallel channels, also shown in Fig. 9. Density improvements will also need to be made at the card edge, where the density of optical interconnections required can quickly consume the available back of the drawer area. For example, Fig. 10 shows a partially populated IBM Power 775 rack, with some 60 K fibers per rack needing accommodation.

B. Power Consumption

The power of a link must consider not only the optical link but also the electrical part of the links to get to and from the optics

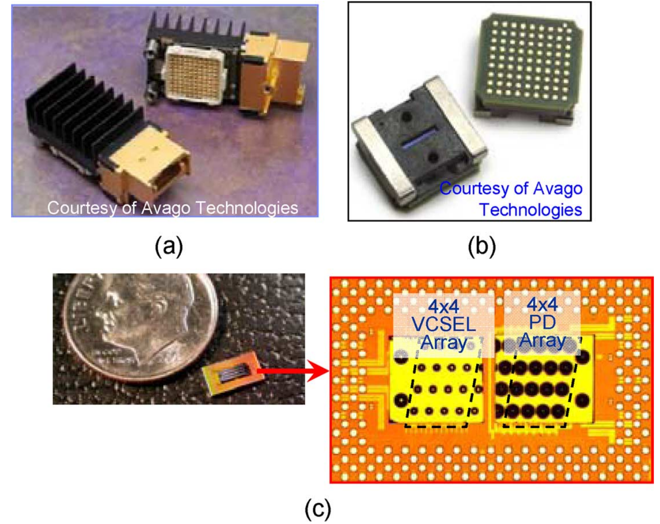


Fig. 9. (a) SNAP12 form factor; $\sim$18 mm $\times$ 41 mm, 1.27 mm pitch. (b) Avago MicroPOD module used in IBM Power 775; $\sim$8 mm $\times$ 8 mm, $\sim$0.75 mm pitch. (c) Prototype optical subassembly; $\sim$5 mm $\times$ 3 mm, 0.2 mm pitch.

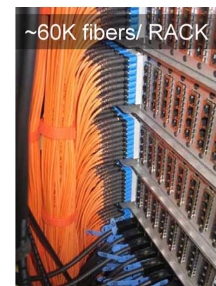


Fig. 10. Partially populated optics cables in an IBM Power 775 supercomputer rack.

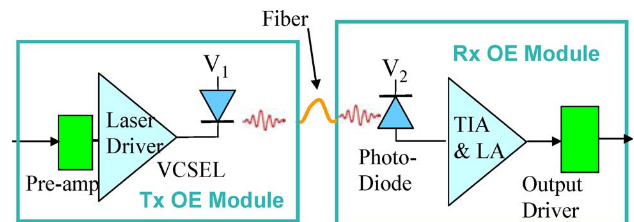


Fig. 11. Power consuming components for a typical VCSEL-based optical interconnect.

TABLE I
POWER CONSUMPTION BREAKOUT OF TYPICAL OPTICAL LINK FUNCTIONAL COMPONENTS FOR A NOMINAL 10 MW/GBPS LINK (TYPICAL OF DESIGNS AT 10–20 GBPS)

Component	Power (mw/Gbps)
Tx Pre-amp	2.0-2.5
Tx Laser Driver	1.0-2.0
Tx VCSEL	1.0
Rx Trans-Impedence Amp (TIA)	1.0-1.5
Rx Limiting Amp (LA)	3.0-3.5
Rx Output Driver	0.5-2.5
Total	**8.5-13.0**

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on September 11,2020 at 08:01:03 UTC from IEEE Xplore. Restrictions apply.

Fig. 12. Example of an active optical cable.

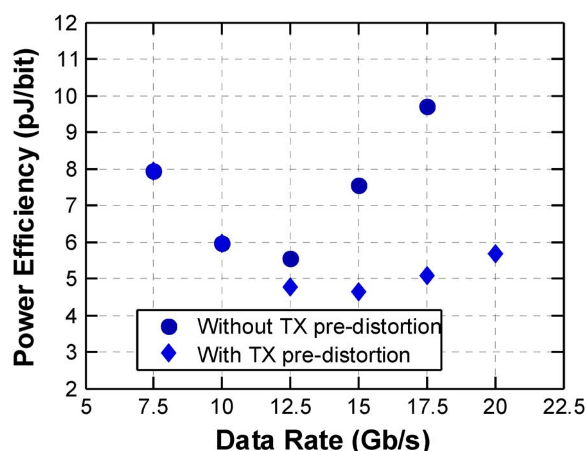


Fig. 13. Power consumption of a 20 Gbps optical link can be greatly improved through the use of predistortion [9].

as well. Fig. 11 along with Table I shows a typical optical link power breakdown for a nominal 10 mW link for typical transmit side (Tx) and receive side (Rx) components. Note that the efficiency of the vertical cavity surface emitting laser (VCSEL) will affect the pre-amp and laser driver component power consumption and Rx output driver power consumption will depend on the electrical channel length and quality. One unidirectional electrical link must also be added to account for the electrical transmit to the optics and the electrical receive from the optics, which, depending on the length and data rate could add an additional 1–10 mW to the total link power depending on the link distance and physical characteristics of the channel. While the active optical cable concept (shown in Fig. 12) is a popular form factor today (using optics housed in an electrical connector to mate directly to the back of rack), the long distance trace along the PCB to get to the optics will necessitate higher electrical link power, making this albeit convenient but less integrated approach a more difficult option in view of the overall power expended. There is still much room for improvement in optical links by employing many of the same signal enhancing techniques which have been employed in electrical links today, such as predistortion, equalization, and decision feedback equalization. Fig. 13 shows the benefits to optical power by the addition of a predistortion stage which compensates for frequency-dependent signal degradation in the VCSEL and initial stages of the receivers [9].

TABLE II
FAILURE RATE AND TIME TO FIRST FAIL AS A FUNCTION OF NUMBER OF LINKS, AND SPARED/UNSPARED FAILURE RATE [10]

# of Links	No Spare VCSEL + 50 FIT unspared	Spare VCSEL + 50 FIT unspared (time to 1st fail)	Spare VCSEL only (time to 1st fail)
1K	1.5 fails/year	20Khrs	252Khrs
10K	1 fail/month	2Khrs	174kHrs
100K	2.68 fails/week	200hrs	39Khrs
1M	3.8 fails/day	20hrs	12Khrs

C. Costs

The costs associated with an optical link include the bill of materials, assembly and fabrication costs as well as test and adjustment for any yield fall out. The bill of materials may include such items as substrates, lenses, laser and photodiode arrays, microcontroller, driver and receiver chips, connectors (both optical and electrical), fiber cabling, and heat sinks. Assembly and fabrication costs and considerations may include assembly throughput rate and equipment costs, single versus multipart (e.g., panel or wafer level) output, active versus passive alignment, manual versus automated assembly, tolerances versus yield, and final assembly costs for the system. Finally, test and yield cost-related issues include tester time and equipment cost, built in self-test and bit error rate requirements versus test time.

These tradeoffs for cost need to be considered and balanced in each of these categories. For example, the issue of active versus passive alignment has been a long time tradeoff in the industry. From a parts cost and complexity point of view, passive alignment often requires tighter tolerance parts or additional alignment structures, while active alignment parts are simpler but require a longer and more intensive assembly process, with, therefore, higher fabrication costs. One area that clearly needs attention is test costs. As module BW and channel count increase and the level of integration of optics in systems increases, the cost of test is becoming a larger portion of the overall costs. Modules with self-test capability or which can be acceptably tested in phases, partially at the optics supplier and partially as part of system assembly (without incurring high rework costs or yield fallout) will be needed.

D. Reliability

In addition, reliability of optical components will require continued improvement. Due to the shear numbers of optics modules in these large systems, even small failure rates can cause network failure, which if not managed carefully, can cause large jobs to stop, requiring a rerun from the last checkpoint. Reliability can be managed by a combination of low component fail rates and the use of redundant network topologies, providing alternate routing paths in the event of a link fail, and channel sparing, which can be used as a failover to allow the link to continue operation. Table II shows an example of computation [10] based on VCSEL random and wearout fail statistics. With an assumption of a random VCSEL failure in time (FIT) of 10/device, an 11 channel link with 1 potential spare, and an additional

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on September 11,2020 at 08:01:03 UTC from IEEE Xplore. Restrictions apply.

50 FIT of unspared failure in associated components (e.g., packaging, microcontroller), one finds that considering VCSEL only fails, the addition of a spare can make a tremendous difference in overall failure rate (for a 1 M device system, time to first fail of 12 Kh). However, the addition of 50 FIT of unspared failure potential can bring even the time to first fail with spared VCSEL links down to 20 h. Thus, attention will need to be given to balancing not only single optical channel fail rates with sparing but also minimizing the multichannel failure rate and single points of failure scenarios for the entire link. The trend toward water cooling can reduce operating temperatures and, therefore, help to lower laser fail rates.

E. Competing With Copper

For optics to compete with copper at shorter and shorter link distances, there are a number of areas for increased focus. Historically, increasing channel data rates has improved cost, power, and density and that trend will continue with data rates at least to 25 Gbps if not beyond, with the caveat that the power cost of multiplexing slower microprocessor and switch on chip data rates will be begin to rise, limiting the overall benefit of higher data rates. Higher parallelism in optics modules (e.g., $24+24$ channels) will help amortize packaging costs and allow more area efficient packaging, but too high a channel count may outpace the volume market, contributing to higher costs. As discussed previously, integrating optics much closer to the signal source will eliminate excess electrical link power, but can result in too highly customized modules that may not have volume acceptance in the marketplace or require excessive development expense to integrate and test. Development of semicustomizable or standard building blocks for optical links can possibly mitigate the integration issue, however.

IV. Technologies of Interest

This section describes a number of optical technologies which are of high interest to fulfill HPC future requirements. As each of these technologies could merit a separate review paper in their own right, only a brief introduction will be provided along with references for further reading.

A. VCSEL/Fiber Technology

Historically, low-cost optical interconnects for datacom, and now computercom, have been based on multimode fibers and VCSEL technology [11]–[13]. Rack-to-rack cluster fabric in particular has made good use of parallel optical modules employing these technologies. In comparison to single-mode technology, more commonly used in telecom, multimode technology is more alignment tolerant. Multimode VCSELs provide a cheaper and easier to test light source than edge-emitting single-mode lasers, and furthermore, are easy to make in compact arrays, and more power efficient as well. So, although multimode fiber has distance limitations due to path differences between the various modes (e.g., typically in the 100 or 100's of meters range, getting worse with higher data rates), it is still the right choice for these primarily short datacom and computercom links.

As the incumbent, this technology already enjoys a low-cost manufacturing infrastructure, and furthermore one that still has room for improvement. Higher data rate VCSELS are already in development, with many suppliers now focusing on 25 Gbps technology for 100 Gbps Ethernet (4×25 Gbps parallel). Fig. 14 shows a 30 Gbps VCSEL transmit eye shown in a lab demonstration [14]). Mass manufacturing methods are being adopted to further lower cost and meet future high volume demands, as evidenced by the panel-based Avago MicroPOD manufacturing approach [15].

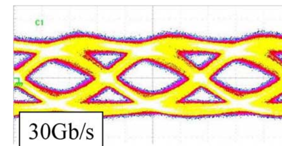
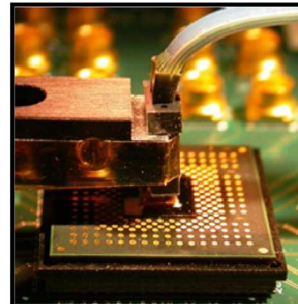


Fig. 14. High-speed prototype VCSEL transmit eye at 30 Gbps [10].



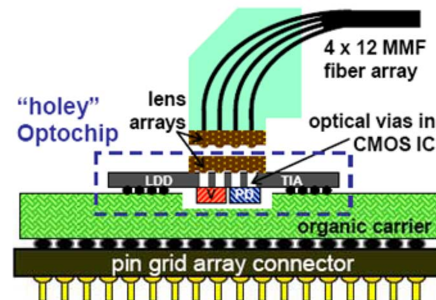


Fig. 15. Prototype high-density optical interconnect transceiver [16].

This technology will continue to improve, with higher, lower cost, lower power, and more compact modules. Fig. 15 shows a prototype highly compact optical module employing flip chip attachment of the VCSELS and photodiode arrays to a CMOS chip with "optical vias" (holes in the Si substrate) to permit coupling to an optical fiber array [16]. This module provides up to 300 Gbps (24×12.5 Gbps in each direction) at 8.2 pJ/bit with a density of 1 Tbps/cm^2.

Although the 850 nm wavelength for VCSEL links has been the standard for many years (i.e., 1GbE in 1998), the optimal wavelength has been debated for many years [17] as well. Recently, there has been renewed interest in longer wavelengths in the 900–1100 nm range, based on AlGaAs and InGaAs alloys. This interest is spurred by a number of factors, including potential speed, efficiency and reliability improvements, ease in fabricating backside emitting VCSELs (the GaAs substrate is transparent at longer wavelengths, which allows new packaging options), and the potential for low-cost coarse wavelength division multiplexing (CWDM) transceivers [18]. In addition, longer wavelengths enjoy a slight photodetector responsivity

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on September 11,2020 at 08:01:03 UTC from IEEE Xplore. Restrictions apply.

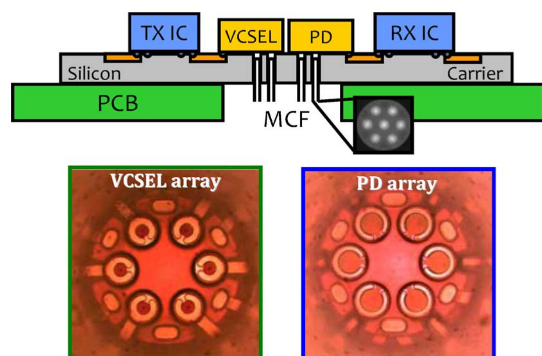


Fig. 16. Prototype optical transceiver utilizing multicore multimode fiber [23].

advantage (producing more current per unit optical power), and more relaxed eye-safety power limits [17]. There is some disadvantage in the use of these longer wavelengths with polymer waveguides (as described in Section IV-B in the following) as the losses tend to be greater for these longer wavelengths.

While there have been significant accomplishments for these longer wavelength VCSELs (e.g., [19], [20]) demonstrating record power conversion efficiencies and good reliability, others have found no significant difference in degradation mechanisms [21] across a range of wavelengths (780–910 nm) for a given set of VCSEL structures. As well, progress continues to be made in 850 nm devices, for example, 100 fJ/bit power dissipation (VCSEL only) at 25 Gbps [22].

To further reduce overall costs of optical links, not only must the cost of the transceivers be reduced but the cost of fiber connectors, cabling, and fiber management must be as well. Higher data rates will help reduce these costs to some degree, but further reduction may be required. One way to accomplish this is to use multiple multimode cores in a single fiber to achieve much higher data rates. Fig. 16 shows such a transceiver which utilizes six cores in a seven core fiber to achieve a 90 Gbps throughput in a single fiber [23]. CWDM transceivers based on multiple wavelengths in the 800–1100 nm range can similarly improve the BW per fiber. Ultimately, this technology may be limited by the heterogeneous nature of packaging integration required and the burgeoning costs of fiber and fiber management, but at present there is still much room for improvement.

B. VCSEL/Optical PCB Technology

To make further gains in cost and level of packaging integration, and compete with copper at on-card distances, optical PCB technology based on polymer waveguide integration with VCSELS may provide the right combination of low-cost manufacturing, module density, and semicustomizable integration [24]–[28]. Fig. 17 shows various elements of this technology, including demonstrations of (a) waveguides fabricated directly on PCB, (b) waveguides on flex and (c) passive shuffle element and (d) connector. Fig. 17(e) shows the construction of the optical module used in 17(b). The Si driver and receiver IC's along with the VCSEL and photodiode arrays are solder attached to a Si carrier. Holes in the Si carrier allow optical access. The optical path is completed through a two lens system to couple into the waveguide mirrors. The two lens system allows a greater misalignment tolerance (>20 μm for 1 dB loss) for the step of attaching the optical module to the PCB, although the individual lens arrays must be attached to their respective sides of the assembly with better tolerance ($< \sim 5$ μm).

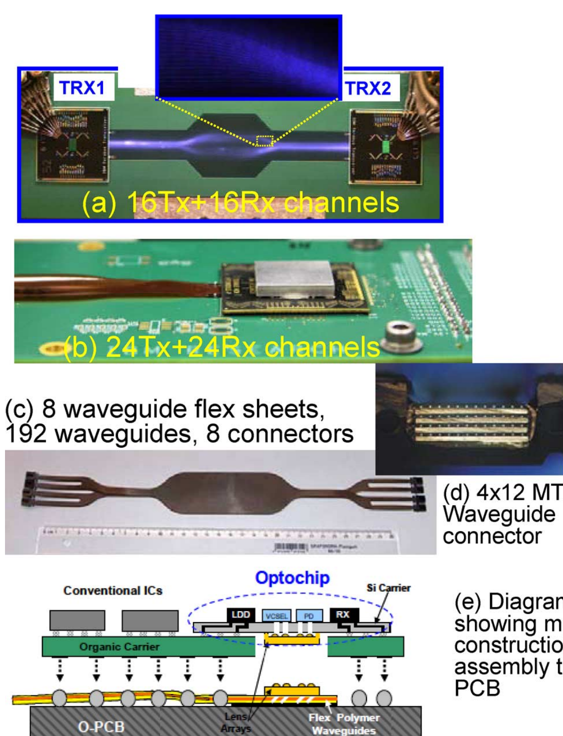


Fig. 17. (a) Polymer waveguides on the PCB substrate demonstrating 16 Tx and 16 Rx channels. (b) Polymer waveguides on a flexible substrate demonstrating 24 Tx + 24 Rx channels. (c) Passive shuffle cable. (d) Four layer waveguide connector. e) Diagram of optical module construction and assembly.

While thermal expansion mismatch is not a problem for attachment of the Si carrier to the underlying organic substrate, some curling of the flex can be observed due to some thermal mismatch between the underlying flex substrate and the polymer waveguide film.

The vision for this technology is to provide an optics technology with the characteristics of electrical PCB technology. PCBs are based on low-cost mass manufacturing methods, yet are customizable for a particular users needs. An optical PCB would mitigate fiber management problems within the card and provide high-density optical transceiver integration close to the processing chips. In order to facilitate a transition to this technology, we anticipate that early offerings would be in the form of an easily replaceable waveguide on flexible substrate assembly, mounted above the board, similar to a fiber ribbon. However, as the technology matures, the polymer waveguides would be incorporated on or within the PCB. While very promising, this technology still has some hurdles to overcome, including improvement in polymer and connector losses and achievement of an infrastructure to allow widespread use of the technology.

C. Silicon Photonics Technology

Finally, silicon photonics is a promising technology which has been studied since the mid 1980s [29], [30], as a platform for optical communications. The technology utilizes single-mode fiber in combination with unmodulated lasers and silicon-based modulators and detectors [31]–[34].

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on September 11,2020 at 08:01:03 UTC from IEEE Xplore. Restrictions apply.

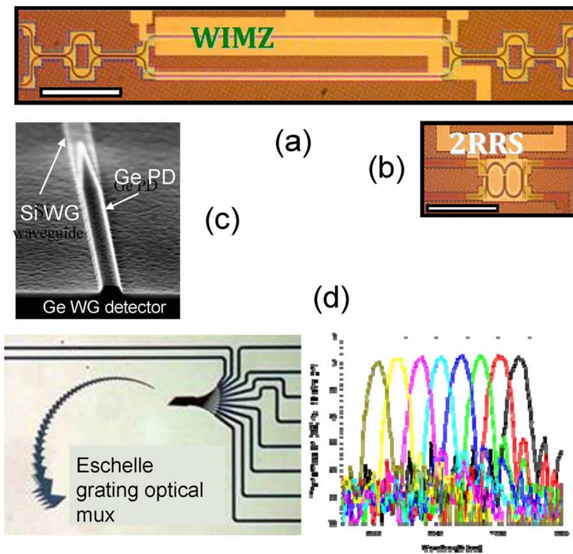


Fig. 18. (a) Wavelength insensitive Mach–Zehnder modulator [37]. (b) Double-ring resonator modulator [38]. (c) Integrated Ge photodetector and Si waveguide [35]. (d) Optical multiplexing/demultiplexing based on an Eschelle grating and associated transmission spectrum [36].

TABLE III
COMPARISON OF TWO DIFFERENT MODULATOR DESIGNS BUILT IN SI PHOTONICS AND SHOWN IN FIG. 18(A) AND (B) [38]

	Wavelength-Insensitive Mach Zehnder (WIMZ)	Ring Resonator (2RRS)
no. stages	1	2
diode length (μm)	200	90
series resistance (Ω)	10	18
footprint (mm2)	0.02	0.001
optical bandwidth (nm)	> 100	< 1
ON power (mW)	4	< 3
digital response	no	yes
port configuration	2×2	1×2

This technology may offer the ultimate in integration capability as well as low cost by utilizing mature CMOS fabrication to produce highly integrated assemblies with most elements fabricated directly in CMOS. In addition, by greatly reducing the costs of introducing wavelength division multiplexing (WDM) capability, which allows multiple wavelengths on the same fiber, the costs of fiber cabling and connectors can be amortized over much greater BW per fiber. Fig. 18 shows some of the technology elements which are required [35]–[38]: Si waveguides, integrated Ge detector, modulator based on either a Mach–Zehnder interferometer or ring resonator and WDM multiplexing elements.

The technology has matured significantly over the years with commercial products available today in an active cable form [31]. Since the technology requires use of single-mode fiber and longer wavelengths [typically $\sim$1300 nm or $\sim$1500 nm, to take advantage of already developed continuous wave (CW) telecommunications single-mode lasers which operate in a region where Si is transparent], Si-photonics-based transceivers are incompatible with the shorter wavelength and multimode technology based on VCSELs. Active cables offer a good initial entry into the market, since interoperability is not an issue. In addition, the lower signal distortion and losses experienced in single-mode fiber at these longer wavelengths allows longer lengths (e.g., 2 km) at high data rates which are of potential interest in very large installations, and not easily attainable with multimode links.

The much lower parasitics of highly integrated electrical and optical devices will be of great help in improving power consumption at high data rates. In order to design modulators, Si photonics will, however, have to contend with the nature of Si as an indirect bandgap material which must be coaxed into performing optically. Modulator design is an exercise in balancing optical BW, temperature sensitivity and control, power consumption, and optical losses. For example, Table III [38] shows a comparison of the two different optical modulator designs shown in Fig. 18. The Mach–Zehnder design has a larger optical BW and less temperature sensitivity, at a cost of larger area and more power over the ring resonator design. Fortunately, large-scale HPC machines in the future are likely to be water cooled, which can help to reduce the required temperature range of operation to a few tens of °C. However, to address a larger market, Si photonics will have to address a more typical temperature range, e.g., 0–70 °C.

Laser light sources can be packaged onto the chip [31] or located at a convenient location off-chip and coupled to the chip via fiber. The on-chip location offers the convenience of more integrated and potentially lower cost packaging but will be exposed to a more challenging thermal environment. The off-chip location offers a separated environment for the laser, such that the temperature (and therefore wavelength) can be more accurately controlled. A lower temperature environment may help improve laser reliability as well. In addition, one might conceive of high-power off-chip lasers which could be split among many transceivers thereby amortizing the cost of the laser and the laser packaging and cooling across many more optical channels.

Packaging is another area that is often overlooked in discussions on Si photonics. While the Si photonics chip itself may be of relatively low cost, the coupling of the chip to fiber and the addition of a CW laser can add substantial cost. Packaging which meets single-mode tolerances (typically <1 μm) can be considerably more expensive than packaging which meets multimode tolerances ($\sim$10 μm). In addition to the cost of the laser, an optical isolator may be required, as single-mode edge-emitting lasers are more sensitive to reflected light than multimode VCSELs, and the amount of optical feedback must generally be quite low ($\sim -$ 30–40 dB).

Finally, one must consider total power consumption for the Si photonics links. While there is significant potential for very low power optical links (e.g., modulator performance in the $\sim$100 fJ/bit range [39]), designs with a practical balance of performance and temperature tolerance and a proper accounting for all power sources (including temperature control, CW laser, and any control or clocking logic) may find that this potential advantage is significantly eroded.

The "killer app" for Si photonics may ultimately be to integrate these optical transceivers directly into a 3-D chip stack,

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on September 11,2020 at 08:01:03 UTC from IEEE Xplore. Restrictions apply.

and although, this will require significant development and maturation, will be difficult for any other technology to match.

D. Passive Connectors and Cabling

In addition to the active transceiver technologies discussed previously, linking these transceivers across cards, boards, or racks will require passive connectors and cabling. In the case of fiber-based VCSEL links, these rely on well-developed parallel fiber ribbon and connectors, e.g., MPO (Multi-fiber Push On), which have been used for many years for these multimode links. Connector losses are typically no more than 0.5 dB, including misalignment tolerances.

For polymer-waveguide-based links, longer connections (e.g., 1 m between boards) will require improved losses if polymer waveguides are to be used; however, the use of a waveguide to fiber connector also allows use of lower loss fiber for these links [e.g., as shown in Fig. 17(d)]. These connectors can have additional losses of ~ 0.5 dB due to the geometric mismatch between the round fiber core and the square waveguide core (may be asymmetric depending on choice of dimensions).

Si photonics technology will require single-mode fiber and connectors. These connectors can typically have 0.25 dB additional loss (over multimode) due to the tighter alignment tolerances required, although lower loss components are available at higher costs. In addition, greater care must be paid during assembly to prevent contamination from ambient particulates, which can more easily degrade these single-mode fiber connections than multimode connections as dust can more easily occlude the smaller single-mode core size (9 μm versus 50 μm for multimode).

V. New Architectures

In these large computing systems, the switching functionality of the network is a major power consumer and cost. Over the years, there have been many demonstrations of optical circuit switching as a more efficient alternative to electrical packet-switched networks. Optical circuit-switched networks may have a role to play in large-scale computing systems, as a more efficient means to move large amounts of data between compute nodes or allow reconfigurability in the network topology. One issue that arises with these new architectures is that it is difficult to simultaneously introduce a new architecture and a new technology; there is simply too much risk. Thus, an early introduction of optics into the switching network might utilize existing microelectromechanical systems technology [40], [41], which provides circuit switching capability with an order of $\sim$10 ms switching time. Such networks might allow reconfigurability of networks to optimize application performance as long as application run time is long enough to overcome the reconfiguration time scales. A joint IBM/Corning project, Osmosis [42] broadcast the optical signal over multiple parallel channels and wavelengths to all receivers, and then utilized semiconductor optical amplifier (SOAs) elements as gates to allow only the intended nodes to receive the signal. Although a relatively complex electrical "central scheduler" is required to set the gates, the chip only needs to manage the routing information, while the payloads are sent over an entirely optical channel. Another optical switch approach (data vortex) taken by Columbia University [43] utilizes wavelength-based coding to achieve optical packet switching, where only the wavelength-based codes need to be converted to electrical signals to properly route the all optical payloads. One issue with these latter two approaches is the reliance on still relatively expensive elements such as SOAs. Si photonics may ultimately provide a much lower cost optical switch [44] by integrating all optical switching elements with the electronic control and scheduling functions. One caveat is that today Si photonics is better suited to very low radix switch elements (e.g., 2×2 or 4×4) since larger radix switch functionality will need to be stitched together from many smaller elements and will, therefore, be limited due to losses and the lack of a cost efficient optical gain element. Future advances in Si photonics will hopefully ameliorate these issues. Finally, while all optical switching remains a worthy goal, the lack of optical memory as a means to buffer contentions remains an issue that must be overcome.

VI. Conclusion

Computercom is an emerging market with a need for very short links, many less than 10 m, with very high aggregate BW and very high demands for low cost, low power, high reliability, and high density. Tightly integrated optics packaging will be required to achieve these goals along with a broad view to optimize technology across system requirements and close cooperation between the system providers and component suppliers as well.

Acknowledgment

The author would like to thank C. Schow, J. Kash, P. Pepeljugoski, D. Kuchta, L. Schares, F. Doany, C. Baks, M. Ritter, L. Shan, K. Gu, D. Kam, Y. Kwark, R. Budd, F. Libsch, C. Tsang, J. Knickerbocker, P. Coteus, A. Gara, Y. Vlasov, S. Assefa, W. Green, B. Offrein, R. Dangel, F. Horst, Y. Taira, Y. Katayama, B. Lee, J. Van Campenhout, A. V. Rylyakov, M. Yang, J. Rosenberg, S. Nakagawa, A. Benner, D. Stigliani, C. DeCusatis, H. Bagheri, K. Akasofu and many others at IBM for their technical work and insights contributing to and underlying the content of this paper, and M. Soyuer for his management support.

References

[1] A. Benner, M. Ignatowski, J. Kash, D. Kuchta, and M. Ritter, "Exploitation of optical interconnects in future server architectures," *IBM J. Res. Develop.*, vol. 49, pp. 755–775, 2005.

[2] C. DeCusatis and C. J. S. DeCusatis, *Fiber Optic Essentials*. New York: Academic, 2006, pp. 154–155.

[3] [Online]. Available: http://www-03.ibm.com/systems/power/hardware/775/

[4] B. Arimilli, R. Arimilli, V. Chung, S. Clark, W. Denzel, B. Drerup, T. Hoefler, J. Joyner, J. Lewis, L. Jian, N. Nan, and R. Rajamony, "The PERCS high-performance interconnect," in *18th IEEE Annu. Symp. High Perform. Interconnects*, 2010, pp. 75–82.

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

[6] A. Gara, M. A. Blumrich, D. Chen, G. L.-T. Chiu, P. Coteus, M. E. Giampapa, R. A. Haring, P. Heidelberger, D. Hoenicke, G. V. Kopcsay, T. A. Liebsch, M. Ohmacht, B. D. Steinmacher-Burow, T. Takken, and P. Vranas, "Overview of the blue Gene/L system architecture," *IBM J. Res. Develop.*, vol. 49, no. 2--3, pp. 195–212, 2005.

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on September 11,2020 at 08:01:03 UTC from IEEE Xplore. Restrictions apply.

[7] Y. Ajima, Y. Takagi, T. Inoue, S. Hiramoto, and T. Shimizu, "The tofu interconnect," in *Proc. 19th IEEE Annu. Symp. High Perform. Interconnects*, Aug. 24–26, 2011, pp. 87–94.
[8] D. G. Kam, M. B. Ritter, T. J. Beukema, J. F. Bulzachelli, P. K. Pepeljugoski, Y. H. Kwark, L. Shan, X. Gu, C. W. Baks, R. A. John, G. Hougham, C. Schuster, R. Rimolo-Donadio, and B. Wu, "Is 25 Gb/s on-board signaling viable?," *IEEE Trans. Adv. Packag.*, vol. 32, no. 2, pp. 328–344, May 2009.
[9] C. Schow *et al.*, "Transmitter pre-distortion for simultaneous improvements in bit-rate, sensitivity, jitter, and power efficiency in 20 Gb/s CMOS-driven VCSEL links," in *Nat. Fiber Opt. Eng. Conf./ Opt. Fiber Conf. Expo.*, 2011, pp. 1–3.
[10] D. Kuchta, "Advances in high speed parallel links for computational systems," in *Proc. IEEE Syst. Packag. Japan Workshop*, Jan. 2010.
[11] P. Pepeljugoski, S. E. Golowich, A. J. Ritger, P. Kolesar, and A. Risteski, "Modeling and simulation of the next generation multimode fiber," *J. Lightw. Technol.*, vol. 21, no. 5, pp. 1242–1255, May 2003.
[12] P. Pepeljugoski, M. J. Hackert, J. S. Abbott, S. E. Swanson, S. E. Golowich, A. J. Ritger, P. Kolesar, Y. C. Chen, and P. Pleunis, "Development of system specification for laser optimized 50-um multimode fiber for multigigabit short-wavelength LANs," *J. Lightw. Technol.*, vol. 21, no. 5, pp. 1256–1275, May 2003.
[13] D. Kuchta, , R. Michalzik and F. Koyama, Eds., "Progress in VCSEL Based Parallel Links," in *VCSELs—Fundamentals, Technology and Applications of Vertical-Cavity Surface-Emitting Lasers*. New York: Springer-Verlag, 2011, .
[14] R. Johnson and D. Kuchta, "30 Gb/s directly modulated 850 nm datacom VCSELs," presented at the presented at the Conf. Lasers ElectroOpt. (CLEO), San Francisco, CA, May 2008, post deadline paper.
[15] M. Fields, "Transceivers and optical engines for computer and datacenter interconnects," in *Proc. Opt. Fiber Commun. Conf.*, 2010, pp. 1–2.
[16] F. Doany, C. L. Schow, B. G. Lee, A. V. Rylakov, C. Jahnes, Y. Kwark, C. Baks, D. M. Kuchta, and J. A. KAsh, "Dense 24 TX + 24 RX fiber-coupled optical module based on a holey CMOS transceiver IC," in *Proc. 60th Electron. Compon. Technol. Conf. (ECTC)*, 2010, pp. 247–255.
[17] D. Hanson, "Case For Using 980 nm (Rather Than 850 nm) VCSELs For Serial 10 Gb/s Links With New Higher-Bandwidth 50 MMF," 1999 [Online]. Available: http://www.ieee802.org/3/10G_study/public/july99/hanson_1_0799.pdf
[18] B. E. Lemoff, M. E. Ali, G. Panotopoulos;, E. de Groot, G. M. Flower, G. H. Rankin, A. J. Schmit, K. D. Djordjev, M. R. T. Tan, A. Tandon, W. Gong, R. P. Tella, B. Law, L.-K. Chia, and D. W. Dolfi, "Demonstration of a compact low-power 250-Gb/s parallel-WDM optical interconnect," *IEEE Photon. Technol. Lett.*, vol. 17, no. 1, pp. 220–222, Jan. 2005.
[19] K. Takaki, S. Imaia, S. Kamivab, H. Shimizua, Y. Kawakita, K. Hiraiwa, T. Takagia, H. Shimizua, J. Yoshida, T. Ishikawab, N. Tsukijia, and A. Kasukawa, "1060 nm VCSEL for inter-chip optical interconnection," in *Proc. Int. Soc. Opt. Eng.*, 2011, vol. 7952, pp. 1–6.
[20] A. Mutig and D. Bimberg, "Progress on high-speed 980 nm VCSELs for short-reach optical interconnects," *Adv. Opt. Technol.*, vol. 2011, pp. 1–15, 2011, Article ID 290508.
[21] J. K. Guenter, B. Hawkins, and R. A. Hawthorne, "Phenomenological study of VCSEL wearout reliability," in *Proc. Int. Soc. Opt. Eng.*, 2011, vol. 7952, pp. 795209-1–795209-8.
[22] P. Moser, W. Hofmann, P. Wolf, J. A. Lott, G. Larisch, A. Payusov, N. N. Ledentsov, and D. Bimberg, "81 fJ/bit energy-to-data ratio of 850 nm vertical-cavity surface-emitting lasers for optical interconnects," *Appl. Phys. Lett.*, vol. 98, pp. 1–3, 2011.
[23] B. G. Lee, D. M. Kuchta, F. E. Doany, C. L. Schow, C. Baks, R. John, P. Pepeljugoski, T. F. Taunay, B. Zhu, M. F. Yan, G. E. Oulundsen, D. S. Vaidva, W. Luo, and N. Li, "Multimode transceiver for interfacing to multicore graded-index fiber capable of carrying 120-Gb/s over 100-m lengths," in *Proc. 23rd Annu. Meet. IEEE Photon. Soc.*, 2010, pp. 564–565.
[24] F. E. Doany, B. G. Lee, C. L. Schow, C. K. Tsang, C. Baks, Y. Kwark, R. John, J. U. Knickerbocker, and J. A. Kash, "Terabit/s-class 24-channel bidirectional optical transceiver module based on TSV Si carrier for board-level interconnects," in *Proc. Electron. Compon. Technol. Conf.*, Jun. 2010, pp. 58–65.
[25] D. Jubin, R. Dangel, N. Meier, F. Horst, T. Lamprecht, J. Weiss, R. Beyeler, and B. J. Offrein, "Polymer waveguide-based multilayer optical connector," in *Proc. Int. Soc. Opt. Eng.*, 2010, vol. 7607, pp. 76070K-1–76070K-9.
[26] F. D. Doany, "160 Gb/s bidirectional polymer-waveguide board-level optical interconnects using CMOS-based transceivers," *IEEE Trans. Adv. Packag.*, vol. 32, no. 2, pp. 345–359, May 2009.
[27] R. Dangel, C. Berger, R. Beveler, L. Dellmann, M. Gmur, R. Hamelin, F. Horst, T. Lamprecht, T. Morf, S. Oggioni, M. Spreafico, and B. J. Offrein, "Polymer-waveguide-based board-level optical interconnect technology for datacom applications," *IEEE Trans. Adv. Packag.*, vol. 31, no. 4, pp. 759–767, Nov. 2008.
[28] F. Doany, C. L. Schow, B. G. Lee, R. Budd, C. Baks, R. Dngel, R. John, F. Libsch, J. A. Kash, B. Chan, H. Lin, C. Carver, J. Huang, J. Berry, and D. Bajkowski, "Terabit/sec-class board-level optical interconnects through polymer waveguides using 24-channel bidirectional transceiver modules," in *Proc. IEEE 61st Electron. Compon. Technol. Conf.*, 2011, pp. 790–797.
[29] R. A. Soref, "Silicon-based optoelectronics," *Proc. IEEE*, vol. 81, no. 12, pp. 1687–1706, Dec. 1993.
[30] R. Soref, "The past, present, and future of silicon photonics," *IEEE J. Sel. Topics Quantum Electron.*, vol. 12, no. 6, pp. 1678–1687, Nov./Dec. 2006.
[31] D. Guckenberger, S. Abdalla, C. Bradbury, J. Clymore, P. De Dobbelaere, D. Folz, S. Gloeckner, M. Harrison, S. Jackson, D. Kucharski, Y. Liang, C. Lo, M. Mack, G. Masini, A. Mekis, A. Narasimha, M. Peterson, T. Pinguet, J. Redman, S. Sahni, B. Welch, K. Yokoyama, and S. Yu, "Advantages of CMOS photonics for future transceiver applications," in *Proc. 36th Eur. Conf. Exhib. Opt. Commun.*, 2010, pp. 1–6.
[32] Y. A. Vlasov, S. Assefa, W. M. J. Green, M. Yang, C. L. Schow, and A. Rylyakov, "CMOS integrated nanophotonics: Enabling technology for exascale computer systems," presented at the SEMICON, SEMI Technol. Symp., Tokyo, Japan, Dec. 2010.
[33] D. Van Thourhout, "Si photonics," in *Proc. Opt. Fiber Commun. Conf.*, 2010 [Online]. Available: http://photonics.intec.ugent.be/download/
[34] D. A. B. Miller, "Optical interconnects," in *Proc. Opt. Fiber Commun. Conf.*, 2010 [Online]. Available: http://ee.stanford.edu/~dabm
[35] S. Assefa, F. Xia, and Y. Vlasov, "Reinventing germanium avalanche photodetector for nanophotonic on-chip optical interconnects," *Nature*, vol. 464, pp. 80–84, Mar. 2010.
[36] F. Horst, "Silicon integrated waveguide devices for filtering and wavelength demultiplexing," in *Proc. Opt. Fiber Commun. Conf.*, 2010, pp. 1–3.
[37] J. Van Campenhout, W. Green, S. Assefa, and Y. Vlasov, "Low-power, 2×2 silicon electro-optic switch with 110-nm bandwidth for broadband reconfigurable optical networks," *Opt. Exp.*, vol. 17, no. 26, pp. 24020–24029, 2009.
[38] B. G. Lee, W. M. J. Green, J. Van Campenhout, C. L. Schow, A. V. Rylakov, S. Assefa, M. Yang, J. Rosenberg, J. A. Kash, and Y. A. Vlasov, "Comparison of ring resonator and Mach–Zehnder photonic switches integrated with digital CMOS drivers," in *Proc. 23rd Annu. Meet. IEEE Photon. Soc.*, 2010, pp. 327–328.
[39] H. Thacker, I. Shubin, Y. Luo, J. Costa, J. Lexau, X. Zheng, L. Guoliang, J. Yao, J. Li, D. Patil, F. Liu, R. Ho, D. Feng, M. Asghari, T. Pinguet, T. K. Rai, J. G. Mitchell, A. V. Krishnamoorthy, and J. E. Cunningham, "Hybrid integration of silicon nanophotonics with 40 nm-CMOS VLSI drivers and receivers," in *Proc. IEEE 61st Electron. Compon. Technol. Conf.*, 2011, pp. 829–835.
[40] L. Schares, X. J. Zhang, R. Wagle, D. Rajan, P. Selo, S. P. Chang, J. Giles, K. Hildrum, D. Kuchta, J. Wolf, and E. Schenfeld, "A reconfigurable interconnect fabric with optical circuit switch and software optimizer for stream computing system," in *Proc. Opt. Fiber Commun. Conf.*, 2009, pp. 1–3.
[41] A. Vahdat, H. Liu, X. Zhao, and C. Johnson, "The emerging optical data center," in *Proc. Opt. Fiber Commun. Conf. Expo.*, 2011, pp. 1–3.
[42] R. Luijten *et al.*, "The OSMOSIS optical packet switch for supercomputers," in *Proc. Opt. Fiber Commun. Conf.*, 2009, pp. 1–3.
[43] O. Liboiron-Ladouceur, "The data vortex optical packet switched interconnection network," *J. Lightw. Technol.*, vol. 26, no. 13, pp. 1777–1789, Jul. 1, 2008.
[44] B. Lee, "Demonstration of a digital CMOS driver co-designed and integrated with a broadband silicon photonic switch," *J. Lightw. Technol.*, vol. 29, no. 8, pp. 1136–1142, Apr. 2011.

Marc A. Taubenblatt (M'87) received the B.S.E.E degree from Princeton University, Princeton, NJ, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA.

He is currently a Senior Manager at Optical Communications and High Speed Test, IBM T. J. Watson Research Center, Yorktown Heights, NY, where his

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on September 11,2020 at 08:01:03 UTC from IEEE Xplore. Restrictions apply.

research is focused on optical interconnects and high-speed electrical packaging for computer systems and test and innovative diagnostic techniques for high-performance computer chips. He has been at IBM Research Center for 26 years. He has had responsibility for the IBM Research WW optical interconnect strategy for the past ten years. He also manages a research program on advanced computing technology and is involved in commercialization of IBM Research technology.

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on September 11,2020 at 08:01:03 UTC from IEEE Xplore. Restrictions apply.