INVITED PAPER Joint Special Section on Opto-electronics and Communications for Future Optical Network
The Role of Photonics in Future Computing and Data Centers
S. J. Ben Yoo†, Nonmember

SUMMARY This paper covers new architectures, technologies, and performance benchmarking together with prospects for high productivity and high performance computing enabled by photonics. The exponential and sustained increases in computing and data center needs are driving the demands for exascale computing in the future. Power-efficient and parallel computing with balanced system design is essential for reaching that goal as should support ~billon total concurrencies and ~billon core interconnections with ~exabyte/bisection bandwidth. Photonic interconnects offer a disruptive technology solution that fundamentally changes the computing architectural design considerations. Optics provide high throughput, massive parallelism, minimal access latencies, and low power dissipation that remains independent of capacity and distance. In addition to the energy efficiency and many of the fundamental physical problems, optics will bring high productivity computing where programmers can ignore locality between billions of processors and memory where data resides. Repeaterless interconnection links across the entire computing system and all-to-all massively parallel interconnection switch will significantly transform not only the hardware aspects of computing but the way people program and harness the computing capability. This impacts programmability and productivity of computing. Benchmarking and optimization of the configuration of the computing system is very important. Practical and scalable deployment of photonic interconnected computing systems are likely to be aided by emergence of athermal silicon photonics and hybrid integration technologies.

key words: computing, data centers, optical interconnects, silicon photonics

1. Introduction

Ubiquitous computing and cloud computing have become an essential part of our daily lives, and we rely increasingly on these computers for everything from healthcare and climate predictions to entertainment and shopping. In the healthcare sector alone, we are seeing rapid transitions in data processing from two-dimensional images to three-dimensional or hyper-spectral real-time 3D images. However, today’s data centers and computing systems have reached scalability power limitations to scale further. Typical data centers are consuming megawatts of power, and the desire to realize exascale computing is seriously challenged by the power limitations projected to approach 0.5 GW power consumption for Exascale computing.

Figure 1 shows the rapid exponential growth of supercomputer performance based on Top500.org where LINPACK benchmarking was used.

Table 1 summarizes the top 10 on the TOP500 list, and

Table 1 List of Top 10 supercomputers [1].

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Manufacturer</th>
<th>Computer</th>
<th>Country</th>
<th>Linpack (TFLOPS)</th>
<th>Speed (TFLOPS)</th>
<th>Power (KW)</th>
<th>Energy/ Rmax (GFLOPS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>National University of Defense Technology</td>
<td>Intel</td>
<td>Tianhe-2</td>
<td>China</td>
<td>55.9</td>
<td>54.9</td>
<td>17.8</td>
<td>0.33</td>
</tr>
<tr>
<td>2</td>
<td>Oak Ridge National Laboratory</td>
<td>Cray</td>
<td>Titan</td>
<td>USA</td>
<td>51.4</td>
<td>27.1</td>
<td>8.21</td>
<td>0.67</td>
</tr>
<tr>
<td>3</td>
<td>Lawrence Livermore National Laboratory</td>
<td>IBM</td>
<td>Sequoia</td>
<td>USA</td>
<td>17.1</td>
<td>24.1</td>
<td>7.89</td>
<td>0.46</td>
</tr>
<tr>
<td>4</td>
<td>KISTI Advanced Institute for Computational</td>
<td>Fujitsu</td>
<td>K Computer</td>
<td>Japan</td>
<td>16.5</td>
<td>12.3</td>
<td>12.7</td>
<td>1.21</td>
</tr>
<tr>
<td>5</td>
<td>Argonne National Laboratory</td>
<td>IBM</td>
<td>BlueGene/Q</td>
<td>USA</td>
<td>14.0</td>
<td>14.0</td>
<td>3.15</td>
<td>0.46</td>
</tr>
<tr>
<td>6</td>
<td>Swiss National Supercomputing Center (CSCS)</td>
<td>Cray Inc.</td>
<td>Piz Daint Cray XC30</td>
<td>Switzerland</td>
<td>9.5</td>
<td>7.8</td>
<td>2.13</td>
<td>0.57</td>
</tr>
<tr>
<td>7</td>
<td>Texas Advanced Computing Center</td>
<td>Dell</td>
<td>Stampede PowerEdge C6220</td>
<td>USA</td>
<td>5.2</td>
<td>8.5</td>
<td>4.3</td>
<td>0.87</td>
</tr>
<tr>
<td>8</td>
<td>Forschungszentrum Juelich (FZI)</td>
<td>IBM</td>
<td>Juqueen BlueGene/Q</td>
<td>Germany</td>
<td>5.0</td>
<td>5.9</td>
<td>2.3</td>
<td>0.46</td>
</tr>
<tr>
<td>9</td>
<td>Lawrence Livermore National Laboratory</td>
<td>IBM</td>
<td>BlueGene/Q</td>
<td>USA</td>
<td>4.3</td>
<td>5.0</td>
<td>1.97</td>
<td>0.46</td>
</tr>
<tr>
<td>10</td>
<td>Leibniz Rechenzentrum</td>
<td>IBM</td>
<td>SuperMUC-5/6</td>
<td>Germany</td>
<td>3.9</td>
<td>3.2</td>
<td>3.32</td>
<td>1.21</td>
</tr>
</tbody>
</table>

* Tianhe-2 supercomputer power number did not include the power of the cooling system, and the actual power consumption is estimated to be 30% higher than shown. K-computer as tested in this Top500.org data utilities electrical interconnects only.

Copyright © 2014 The Institute of Electronics, Information and Communication Engineers
rent level at \(\sim 0.5 \text{nJ/FLOP}\) so that sustained exascale computing can be realized at below 10 MW average power consumption. Koomey [2], [3] noted in 2010 that the energy efficiency of computing doubles nearly every 1.5 years, but as Fig. 2 illustrates, the Koomey’s law is starting to break down as energy efficiency starts to flatten. At the device level, the Dennard’s law in 1974 [4] describing MOSFET scaling rules for obtaining simultaneous improvements in transistor density, switching speed and power dissipation [4] to follow the Moore’s Law [5], thus doubling every 20 months, has already become obsolete in 2004. (It is often incorrectly quoted as a doubling of transistors every 18 months, as David House, an Intel Executive, gave that period to chip performance increase. The actual period was about 20 months for doubling the number of transistors).

Figure 3 illustrates the continuation of the Moore’s law while the Dennard’s law has stopped to keep pace with the Moore’s law beyond 2004. As the device dimensions such as gate oxide thickness reduce to several atomic layers, tunneling and leakage current become significant. The International Technology Roadmap for Semiconductors (ITRS) [7] refers to this as a ‘red brick wall’ as there is no known technology solution beyond 2016 when CMOS scaling is expected to stop (note that CMOS power density scaling already stopped in 2004).

Hence, while the number of transistors continue to increase, new solutions for improving energy efficiency is no longer about increasing the clock speed in a large single processor, but rather by introducing many small processors running at moderate clock speeds and interconnecting them in a way to support massive concurrencies in parallel at an optimized system performance design configuration. The Amdahl’s law [8] suggests that a parallel computing system with balanced processing, memory, and communications performs best across most applications. This indicates that for an optimized exascale computing system, we expect to need 1 exabyte/second, or 8 million terabit/second bisection bandwidth. Furthermore, as Fig. 4 and Fig. 5 illustrated, the number of cores and the total concurrency are increasing exponentially, now approaching or exceeding 1 million in a large computing and data systems. Unfortunately, currently most of the computing systems are 10x\(\sim\)100x misbalanced and deprived of sufficient interconnection bandwidths and parallelism.

It is evident that electronics alone will be unable to meet the requirements of future computing supporting the
massive bandwidth, concurrency, and connectivity requirements scaling to \( \sim \) billion concurrencies and cores for exascale computing. Photonic interconnects offer a disruptive technology solution that fundamentally changes the computing architectural design considerations. Optics provide ultra-high throughput, massive parallelism, minimal access latencies, and low power dissipation that remains independent of capacity and distance. In addition to the energy efficiency, many of the fundamental physical problems of interconnects are directly addressed within the optical technology platform, including precise clock distribution [9], bit rate transparency, and power reduction, without concerns for impedance, crosstalk, voltage isolation, pin inductance, signal distortion, and repeater-induced latency. Exciting opportunities exist in wavelength routing to reconfigure the high-capacity connectivity of multiple wavelengths to reduce contentions and increase system-wide throughput [11],[12]. Recent advances in silicon nanophotonic technologies compatible with nanoelectronics offer new possibilities in realizing future computing systems with a very new architecture.

2. Optical Interconnects Everywhere

The opportunity for optical interconnects [13]–[16] in board-to-board and rack-to-rack communications is already well documented. Table 2 illustrates interconnection hierarchy in the computing system for typical dimensions and targeted energy and cost budget to stay competitive against electronic counterparts. The target energy was calculated for local link at each region to pursue exascale computing at below 20 MW power in the near future. ITRS Longterm Roadmap [17],[18] also indicates similar trend.

In addition to the power requirement, what remains extremely challenging is the cost target of the new photonic technologies for practical implementation of exascale computing and for market acceptance towards ubiquitous computing.

Many literatures [15],[16] already discussed the benefits of optical interconnects in detail. Figure 6 illustrates the advantage of the optical link by reducing the cost, power, and complexity compared to the electronic link. Figure 7 shows a generic optical transceiver which requires additional optical components but reduces the requirement for clock and data recovery circuits. Since electrical communication links suffer from RF skin-effects, RF losses, electromagnetic interference, and limited bandwidths, high-speed parallel data communication need to employ many electronic regenerators including equalization, reshaping, amplification, and reclocking. On the other hand, optical datalink has can achieve multi Tb/s communications on parallel wavelengths with negligible loss, crosstalk, and distortion on a fiber such that repeaters become unnecessary for typical data centers or computing centers, and the equalizers can be simpler. Further, optical clock distribution can potentially eliminate needs for many clocking circuits in each node.

3. Optical Switches and Routing in Computing Systems and LIONS

As Fig. 8 indicates, typical computing and data centers utilize interconnection of various size electronic switches in many cascaded stages. Due to limitations in radix and bandwidth of the electronic switches, the inefficiency of the cascaded switch stages compounds, especially in terms of latency, throughput, and power consumption. This is one of the key reasons behind challenges of parallel pro-
Fig. 8 Typical data center interconnection topology involving various sizes of electronic switches [19].

Fig. 9 Flattened data center interconnection topology involving $N \times N$ AWGR based LION switch.

Fig. 10 $N \times N$ arrayed waveguide grating router’s (a) wavelength routing property ($N = 5$ example), and (b) wavelength assignment table (figures courtesy of NEL).

Fig. 11 The system diagram of the proposed rack to rack LIONS optical switch, LD: Label Detector; OLG: Optical Label Generator; PE: Packet Encapsulation; LE: Label Extractor; FDL: Fiber Delay Line; TX: Transmitter; RX: Receiver; FFC: packet Format Converter; O/E: Optical-to-Electrical converter; E/O: Electrical-to-Optical Converter [22].

programming because the locality constraints for data become very important. On the other hand, optical parallelism and wavelength routing capability of arrayed wavelength grating routers (AWGRs) can collapse the entire network to a single hop and flattened interconnection topology while supporting all-to-all interconnection as shown in Fig. 9.

Arrayed waveguide grating router (AWGR) [20], [21] is a passive wavelength routing component that enables all-optical switching. The main difference between the electronic crossbar switch and the AWGR switch is that (a) the $N \times N$ AWGR switch provides simultaneously fully connected all-to-all interconnection thus providing $N^2$ simultaneous connection links, (b) there is no switching component within the core of the AWGR switch fabric (AWGR is a passive element), (c) the switching component scales linearly as the number of connection links supported by the switch, and (d) the AWGR supports parallel wavelength interconnection covering huge (> 20 THz) optical bandwidths. Figure 10 illustrates the (a) wavelength routing property ($N = 5$ example) and (b) wavelength assignment table for switching of an AWGR. As Fig. 10 demonstrates, the well-known wavelength routing property of the $N \times N$ AWGR supports simultaneous and non-blocking interconnections of each of $N$ input ports with all of $N$ output ports by using $N$ wavelengths. For the $N = 5$ example, there are $N^2 = 25$ simultaneous interconnections. Each input port can utilize a wavelength (instead of electronic headers) to address the output port. For this reason, any input port can use a tunable transmitter to address each output port by tuning the transmitter to the corresponding wavelength without contention if each output port is equipped with a wavelength demultiplexers and $N$ receivers. Alternatively, the $N \times N$ AWGR can be used in all-to-all interconnection without contention by employing $N$ transmitters and $N$ receivers at each port [22].

Figure 11 shows an AWGR based hybrid interconnecting architecture, referred to as lightwave interconnect optical network switch (LION switch, or LIONS), which also includes optical channel adapters (OCA), tunable wavelength converters (TWC), loopback buffers and the control plane for the switching fabric. LIONS uses label switching with the optical label transmitted on a different wavelength.

We compared the performance of rack to rack LIONS with other state-of-art rack to rack switches including the electrical switching network architecture employing flattened butterfly topology, the IBM-Corning’s Optical Packet Switch for Supercomputers (OSMOSIS) and the Data Vortex optical switch [23]–[26]. Figures 12(a) and (b) show the effective bandwidth and the end-to-end latency of LIONS comparing with a flattened butterfly network under a 32- and 128-node for a message size of 128 bytes. The average end-to-end latency of the flattened butterfly network increases much faster than that of the LIONS system under moderate network load. The flattened butterfly network saturates more easily with increasing network size, while the latency of LIONS is almost independent of the size of the network. The effective bandwidth comparison shows that LIONS can support heavy network load of up to 90% and beyond. We simulated the case where the LIONS has a radix of 64 and the message size is 256 bytes (the packet size is 324 bytes) so that the setting is comparable with that of OSMOSIS demonstrator. At the load of 90%, the latency of L-
ONS is still less than 190 ns. In comparison, the minimum achievable latency of OSMOSIS is above 700 nanoseconds by only considering data path delay as well as STX arbitration [27]. Figure 12(d) shows the effective bandwidth comparison. LIONS achieves a little higher effective bandwidth than OSMOSIS, because the InfiniBand label is smaller than the total overhead of the OSMOSIS. Figure 12(d) shows the effective bandwidth comparison. Notice that Data Vortex saturates before the uniform network load reaches 0.5 [27]. Clearly LIONS outperforms both OSMOSIS and Data Vortex in terms of effective bandwidth. The results also indicate label rate actually affects latency performance of LIONS, higher label rate can reduce the end-to-end latency.

The control plane in Fig. 11 can be distributed by all-optical token and accelerated by all-optical negative acknowledgment mechanisms, combined into all-optical TONAK described in Refs. [28], [29]. Figure 13 illustrates the working principle of all-optical token where saturable reflector (reflective-SOA) reflects the first packet with high gain while others with lower saturated gain. By placing the R-SOAs at each receiver, there can be as many parallel tokens available in the network as the number of receivers in the network. Thus arbitration can be achieved in parallel without requiring the centralized control plane.

The LIONS described in Fig. 11 can be a top-of-the-rack switch interconnecting many racks or supernodes at the top hierarchy mentioned in Table 2. As Fig. 14 [30] illustrates, a passive \( N \times N \) AWGR without tunable lasers can interface \( N \) compute nodes or processors to support many parallel interconnections up to \( N^2 \). Figure 14 includes a centralized control plane for illustration, but it will be replaced with a distributed control plane using the all-optical TOKEN technique [28]. Each node has a transmitter array that uses \( k_l \) ring modulators to generate the data packets and \( k_r \) receivers with ring filter to receive packets. An off-chip comb generator provides the \( N \) wavelengths required by the cyclic frequency AWGR for wavelength routing. We restrict the number of rings on each bus waveguide to a fix number \( k_l \) and \( k_r < N \). This passive LIONS can be in multiple hierarchy with interfaces between the same and higher hierarchy as Fig. 15 illustrates for chip-to-chip hierarchy and board-to-board hierarchy.

Figure 16 shows the performance study of the proposed architecture based on a \( 64 \times 64 \) AWGR under various configurations [30]. Even for moderately small \( k_l \) and \( k_r \) at values 2 and 4, there are substantial improvements over the case of 1, equivalent to the case of electronic counterpart. For single-transmitter/receiver-pair configuration \((k_l = 1 \text{ and } k_r = 1)\), the presence of the virtual output queues (VOQ) greatly improves the system performance in all three aspects, but system equipped with multiple transmitter/receiver pairs \((k_l \geq 2 \text{ and } k_r \geq 2)\) still performs better especially in terms of end-to-end latency at high (>90%) input load. Without VOQ, multiple-transmitter/receiver pairs provide significant boost in performance due to the increased statistical multiplexing and the enhanced instantaneous rate at each AWGR inputs and outputs. We observe zero packet loss and 100% throughput for all the cases where \( k_l \geq 2 \) and \( k_r \geq 2 \).

Figure 17 shows a photograph of a fabricated silicon
YOO: THE ROLE OF PHOTONICS IN FUTURE COMPUTING AND DATA CENTERS

Fig. 15 Passive LIONS interfacing other passive LIONS at the same hierarchy and at higher hierarchy.

Fig. 16 Performance study on (a) end-to-end latency, (b) throughput and (c) packet loss rate as functions of the offered load for uniform random traffic distribution on proposed architecture with 64 nodes. Left: no Virtual Output Queue (VOQ), right with VOQ [30].

Fig. 17 Fabricated silicon photonic \( N \times N \) LIONS chip for \( N = 4 \), \( k_x = 4 \) and \( k_y = 4 \) [30].

Fig. 18 Experimental bit-error-rate measurements for (a) 4-by-1 routing and for (b) 1-by-4 routing demonstration [30].

all connectivity without contention.

4. Silicon Photonic Integrated Circuits and Athermalization

Silicon photonics exploits the CMOS infrastructure built up on many billions of dollars of investments in the past. While the photonic devices are quite large in physical size and fabrication resolution compared to the electronic devices, thus the number of devices to be integrated in the die can be orders of magnitude lower, the uniform and well-established CMOS fabrication platform can contributed greatly to manufacturability of photonic components for future computing. Initially, silicon photonics stirred speculations that the future CMOS will combine photonic integration and electronic integration, which expect to bring significant impact by (a) offering intelligent data processing and storage capabilities of electronics together with high capacity and parallelism of photonics, or by (b) realizing signal processing in the photonic domain. However, challenges of photonic-electronic integration lie in the process compatibility between photonic and electronic ICs, isolation, crosstalk, yield, and heat density. The groups at Luxtera, IBM, and MIT/Micron have recently and independently demonstrated electronic-photonic integration on a silicon CMOS platform. Figure 19 shows silicon photonic transmitters and receivers realized on the same electronic platform as silicon CMOS [33].

More practical photonic-electronic integration would be to pursue hybrid integration by die bonding or wafer bonding. As Fig. 20 illustrates, 3D photonic-electronic in-
integration can build up on the already active 3D electronic integration with through-silicon-vias (TSVs).

However, silicon photonics must maintain nearly constant optical response across the operating temperature range. The high temperature sensitivity of silicon material (thermo-optical coefficient of silicon is $1.81 \times 10^{-4} \, \text{K}^{-1}$ which is 20 times greater compared to that of silica) makes this challenging.

Recently, we have realized [34] athermalization of silicon photonic waveguides by using a CMOS compatible process exploiting titanium oxide overcladding. Figure 21 illustrate fabricated athermal micro-resonator ring modulators, and Fig. 22 show experimental testing of the fabricated devices with various waveguide widths showing nearly perfect athermalization as well as blue shift and red shift depending on the waveguide width.

5. Benchmarking Evaluations

New computing paradigms and architectures must be put to test under actual workloads. Figure 23 shows the comparison between the performance of an AWGR based all-to-all fully connected network, that of the AWGR based LIONS switch with three variations, and that of an electronic flattened-butterfly-network (with a large number of switches). The giga-updates per second (GUPS) benchmarking shows the AWGR based all-to-all interconnection results in a 16-fold increase and approaches the theoretical maximum GUPS. GUPS benchmarking is well suited for applications that require low latency and processing on many small data packets such as sorting. TOP500 used LINPACK benchmarking that utilized numerical linear algebra, which is not very memory or communication (interconnect) intensive. GRAPH500 utilized Graphs which are a core part of most analytics workloads that better represent data-intensive applications. Since such workloads require much more communication and access with memory, the GRAPH500 ranking in Table 3 differ significantly from Table 1. The computing systems with better balance and higher memory bandwidth emerged (e.g. IBM BlueGene/Q) compared to the others (e.g. Tianhe-2).

Future computing systems must be compared in the context of expected workloads and optimize based on the workloads towards most efficient design, configuration, and
The exponential and sustained increases in computing and data center needs are driving the demands for exascale computing in the future. Power-efficient and parallel computing with balanced system design is essential for reaching that goal as should support in part from DoD ACS grants. Researchers at UC Davis and support in part from DoD ACS grants acknowledge contributions from colleagues and support in part from DoD ACS grants.

### Acknowledgments

The author acknowledges contributions from colleagues and researchers at UC Davis and support in part from DoD ACS project W911NF-13-1-0090.

### References


24. C. Minkenberg, F. Abe1, P. Muller, R. Krishnamurthy, M. Gusat, P. Dill, I. Iliadis, R. Luijten, B.R. Hemenway, R. Grzybowski, and...


S. J. Ben Yoo is Professor of Electrical Engineering at University of California at Davis (UC Davis). His research at UC Davis includes high-performance optical switching systems, optics in computing, elastic optical networking, terahertz bandwidth RF-photonic systems, and software defined heterogeneous networks. Prior to joining UC Davis in 1999, he was a Senior Research Scientist at Bellcore, leading technical efforts in optical networking research and systems integration. His research activities at Bellcore included demonstration of the first reconfigurable multiwavelength optical networks (MONET), the next-generation Internet using optical-label switching, the advanced technology demonstration multiwavelength optical network (ATD/MONET) systems integration, parametric wavelength interchanging crosconnects, and the OC-192 SONET rings. He also participated in a number of standardization activities, which led to documentations of Generic Requirements, GR-2918-CORE (1999), GR-2918-ILR (1999), GR-1377-CORE (1995), and GR-1377-ILR (1995) on dense WDM and OC-192 systems. Prior to joining Bellcore in 1991, he conducted research at Stanford University on nonlinear optical processes in quantum wells, four-wave-mixing study of molecular relaxation mechanisms, and ultrafast diffusion-driven photodetectors. During this period, he also conducted research on intersubband transitions and on nonlinear optical storage mechanisms at Bell Laboratories and IBM Research Laboratories, respectively. Prof. Yoo is a Fellow of IEEE Photonics Society, a Fellow of the Optical Society of America (OSA), a Fellow of NASA NIAc, and a Member of Tau Beta Pi. He is a recipient of the DARPA Award for Sustained Excellence in 1997, the Bellcore CEO Award in 1998, the Outstanding Mid-Career Research Faculty Award (UC Davis) in 2004, and the Outstanding Senior Research Faculty Award (UC Davis) in 2011. Prof. Yoo also served as Associate Editor for IEEE Photonics Technology Letters, Guest Editor for IEEE/OSA, IEEE Networks, and IEEE Journal of Special Topics in Quantum Electronics, and General Co-Chair for Photonics in Switching conference 2007, 2010, and 2012.