Electronic and Photonic Integrated Circuits for Fast Data Center Optical Circuit Switches

Luis Chen, Avantika Sohdi, John E. Bowers, and Luke Theogarajan, University of California at Santa Barbara
Jon Roth and Greg Fish, Aurrion

ABSTRACT

This article surveys recent work in optical switching technologies and presents a fast optical circuit switch that intimately integrates the control electrics at the chip level using a novel wafer-scale heterogeneous integration technique.

INTRODUCTION

The growth in data center traffic is demanding higher capacity and more energy-efficient ways of computing and moving data across servers. Recent advances in silicon photonics open up exciting opportunities to solve the bandwidth limitation of both on-chip and off-chip wire interconnects. On-chip photonic interconnects have been proposed to link multi-core systems to create a single logical compute node with terascale processing capability in an energy-efficient manner. Within the data center, packet switching is commonly used to connect compute servers because of low latency and high network utilization. To achieve higher capacity, higher energy efficiency, and lower cost, circuit switching becomes attractive [1]. Micro electromechanical systems (MEMS) optical network switches can be large (nonblocking 320 × 320 port switches), but they are limited to switching times of a few milliseconds [2]. The key for faster switching is to integrate both active and passive photonic devices with complementary metal oxide semiconductors (CMOSs) and take advantage of the economy of scale of mature silicon manufacturing processes, thus lowering the cost while simultaneously enhancing the reliability of optoelectronic systems. The rest of this article is organized as follows. We first survey the most recent work in the literature on implementing optical switching alongside existing data center network infrastructures. We then present a nanosecond-scale 2 × 2 switch design integrated with semiconductor optical amplifiers (SOAs) and power detectors, capable of realizing the building block to construct larger port count switch fabric. To control the switch, a custom designed CMOS integrated circuit (IC) was designed and tightly integrated with the photonic IC via a versatile integration process. Lastly, we discuss the scalability of the switch design to high port counts in terms of power consumption and device area, followed by conclusions.

OPTICAL SWITCHING IN DATA CENTERS

Data centers today employ electronic packet switches for their networks. In data-intensive applications, optical interconnects offer significant advantages over their electrical counterparts in terms of power consumption and latency. For example, optical interconnects have been used extensively in recent record-breaking higher-performance supercomputers. Despite the apparent advantages, the deployment of optical interconnects in data centers has been hampered by the relatively high cost [3]. In the research community, there are many teams pursuing the ultimate goal of optical packet switching; however, the goal of all-optical packet switching in data centers remains elusive[4]. One crucial issue is the lack of power-efficient optical packet buffers needed to store payloads when there is contention in the network. More recently, researchers have been studying the deployment of optical circuit switches alongside traditional electronic packet switches in a hybrid network configuration [1], and they have found that in certain applications, the addition of the optical circuit network can reduce network latency significantly. The c-Through hybrid network architecture presented in [5] represents a new paradigm in deploying optical switching in data centers. c-Through is a hybrid packet- and circuit-switched data network architecture that augments a traditional electrical packet switch hierarchy with a second high-speed rack-to-rack circuit-switched optical

0163-6804/13/$25.00

The gain characteristics of SOAs can vary due to process variation, temperature, and aging effects. Therefore, it is desirable to have the ability to tune each SOA independently through either one-time or periodic offline calibration, or continuously controlled via closed loop feedback.

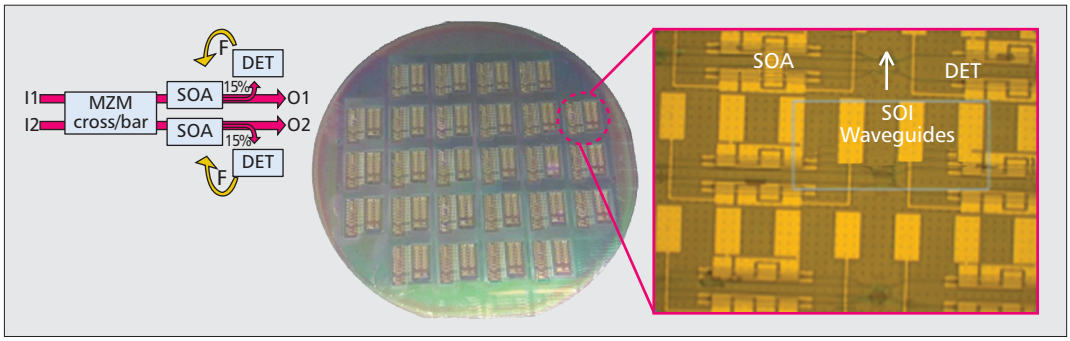

Figure 1. *3-in diameter wafer-scale hybrid silicon process fabricated by Aurrion.*

network. The optical switch cannot switch at packet granularity. c-Through takes advantage of relatively stable rack-to-rack traffic, and using servers to buffer traffic so that the full capacity of the optical link can be utilized. Using servers to buffer packet traffic requires modification of the server's operating system kernel, but the advantage is that it does not require substantial modification to the applications running on the network in order to take advantage of the capacity added by the optical network.

Helios [6] is another work that studies the applicability of optical circuit switches in data centers. Like c-Through, Helios augments electrical packet switches with a MEMS-based optical switch to alleviate heavy congestion when there are hot spots in the network. The key difference between c-Through and Helios is that Helios does not require individual servers to buffer the network traffic. Instead, Helios is designed to route a large volume of data between server container pods, and it monitors the packet switches' built-in traffic flow counter to determine when to switch the flow to the optical circuit switch. The advantage is that Helios does not require modification of the server's OS kernel and can readily be adopted within the existing data center infrastructure.

The researchers behind Optical Switching Architecture (OSA) [7] noted that the network topology in c-Through encounters bottleneck when two top-of-rack (ToR) switches want to use the full bandwidth to simultaneously communicate with a third ToR switch because the link bandwidth to each ToR is fixed. OSA solves the bottleneck by using wavelength-division multiplexing (WDM) to dynamically allocate more bandwidth to a particular switch when the demand arises. Essentially, OSA extends the connectivity and available bandwidth by using WDM.

A limitation of research such as c-Through, Helios, and OSA is that they use MEMS optical switches, which have configuration time on the order of milliseconds, so in actual network traffic they are only used when there are large flows of traffic that tend to be stable for a relatively long time. To overcome this limitation, the researchers of Helios presented a follow-up project, Mordia [8], where microsecond switching-time switches are implemented. The idea is to route as much as possible of the traffic in the data center through the optical paths, and with a fast switch, smaller packets can be routed if the switching time overhead is small. Time-division multiplexing (TDM) is a way to share the available network bandwidth among multiple hosts to realize all-to-all connectivity without having to connect them physically with cables. With slow MEMS switches, TDM is not efficient as the switching time creates large overhead. In contrast, a fast switch reduces that overhead and makes TDM feasible. Mordia demonstrated the application of TDM with an optical switch. Another limitation with slow switches is the amount of buffer memory required. For instance, at millisecond switching time and 10 Gb/s Ethernet, the buffer size is on the order of a few gigabytes, but with a microsecond scale switch, the buffer reduces to a few megabytes, which is much more practical to implement.

Integrated Optical Switch

Traditional silicon-on-insulator (SOI) photonic integrated circuits (PICs) with active silicon switches [9] have high fabrication yield, low propagation loss, and relatively low switch insertion loss compared to III-V switches. However, the number of switches that can be cascaded to build a larger switch fabric is limited by total insertion loss. SOI PICs do not have an efficient gain element to compensate for the loss. Therefore, to make large port count switch fabric possible, it is advantageous to use a process that contains passive SOI waveguides and interferometers, active silicon PIN injection modulators, and heterogeneously integrated III-V SOAs. The SOI wafer can be bonded to the III-V wafer using the process described in [10]. The gain characteristics of SOAs can vary due to process variation, temperature, and aging effects. Therefore, it is desirable to have the ability to tune each SOA independently through either one-time calibration or periodic offline calibration, or continuously controlled via closed loop feedback. Figure 1 shows the proposed 2 × 2 Mach-Zehnder interferometer (MZI)-based switch element (SE) with active feedback power equalization fabricated in a wafer-scale hybrid silicon process. The integrated SOAs can provide up to 15 dB of optical gain. A reverse biased SOA is used as a power detector, which has dark current of 6.5 nA [11]. The 2 × 2 MZI can be switched on a nanosecond timescale, which is many orders of magni-

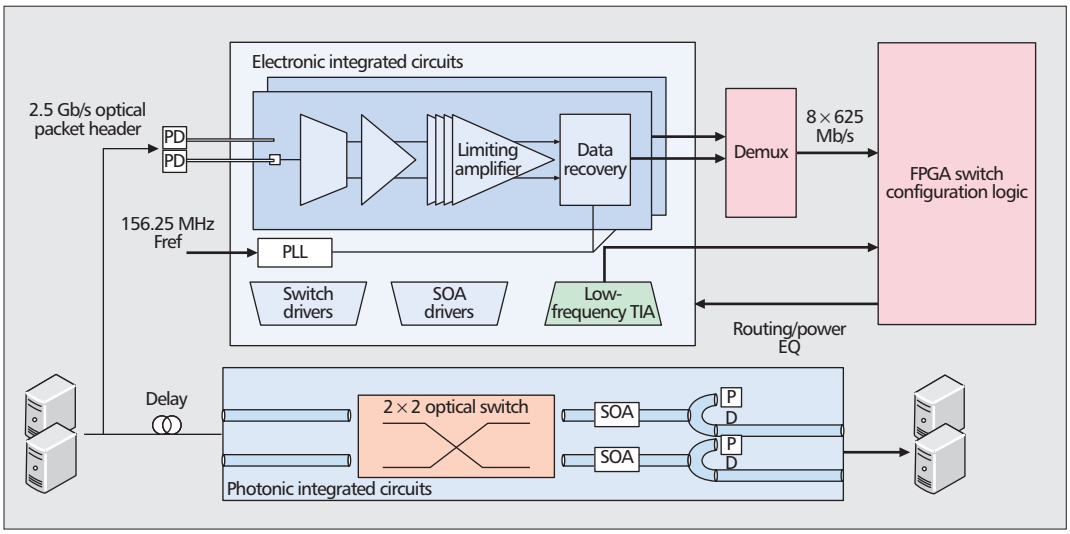

Figure 2. *Block diagram of the electronic controller.*

Recent developments in the field of silicon photonics have enabled development of a highly integrated electronic-photonic platform. However, it is very difficult to realize active optical components, such as lasers and photodetectors, in silicon photonics.

tude faster than MEMS counterparts, while consuming milliwatt-scale power. Since the switch itself performs fast circuit switching, it is independent of packet formats. Packet buffering, routing, and contention resolution can be processed on the server side using the system architectures presented in [5–8].

CMOS Controller IC

A custom IC was designed in a 130 nm CMOS process. The IC is designed to perform optical to electrical conversion of the header information, automatic optical power equalization and it also integrates drivers that changes the state of the 2 × 2 MZI switches. Figure 2 shows a functional block diagram of the IC interfacing with the 2 × 2 switch as well as the external FPGA for signal processing.

A network session begins with an optical header requesting a destination port. The optical signal is converted to an electrical current by a photodiode. The small current is amplified by a transimpedance amplifier (TIA); further amplification is performed by a set of limiting amplifiers (LAs) to overcome the switching threshold of the clock and data recovery (CDR) unit. The CDR recovers the timing information in relation to the bits received so that proper decoding can be performed. The receiver is especially designed to process bursts of data as fast as possible without the need for a long preamble time. The header information is routed to an external field programmable gate array (FPGA) to determine whether the requested port is available. If the port is not available, the request is dropped and retransmission is needed. If the request is granted, the FPGA then instructs the switch driver on the IC to change the state of the 2 × 2 switch. The switching needs to be completed within a few nanoseconds to take full advantage of the fast optical switch.

The IC also supports power equalization by using a built-in low-frequency TIA and an SOA driver. The low-frequency TIA is designed to convert slow-changing variation of optical power detected by a photodiode to a frequency variation, essentially converting amplitude information to time variation, or current to frequency conversion. The advantage is that it can cover a broad range of input amplitudes without saturating, and the output signal level can readily interface with CMOS logic without using an analog-to-digital converter. The SOA driver is a digitally controlled switch that drives a programmable amount of current to the SOA to change the gain level. Closed loop feedback enables the automatic power control that compensates the losses in the 2 × 2 switch due to variations in process, runtime temperature changes, and aging effects. The power gain also enables the ability to cascade many of these 2 × 2 switches to build larger switch fabric, which is one of the main goals of this work. Table 1 summarizes the key performance characteristics of each functional block.

Heterogeneous Electronic and Photonic Integration

Recent developments in the field of silicon photonics have enabled development of a highly integrated electronic-photonic platform. However, it is very difficult to realize active optical components, such as lasers and photodetectors, in silicon photonics. We aim to design a process of integration that is robust and can enable integration of III-V semiconductor photonics devices with silicon CMOS and silicon photonics to take advantage of the high-speed operation of optical devices and the economy of scale of the mature silicon manufacturing process. Traditional approaches to integration can be summarized as follows:

- Embedding the photonic layer into the metal interconnect layers [12]
- Combined electronic and photonic front-end [13]
- Backside fabrication of photonic components [14]

In the first method, the photonic layer is connected to the metal layer of the electronics, and the substrate of the photonic layer is then

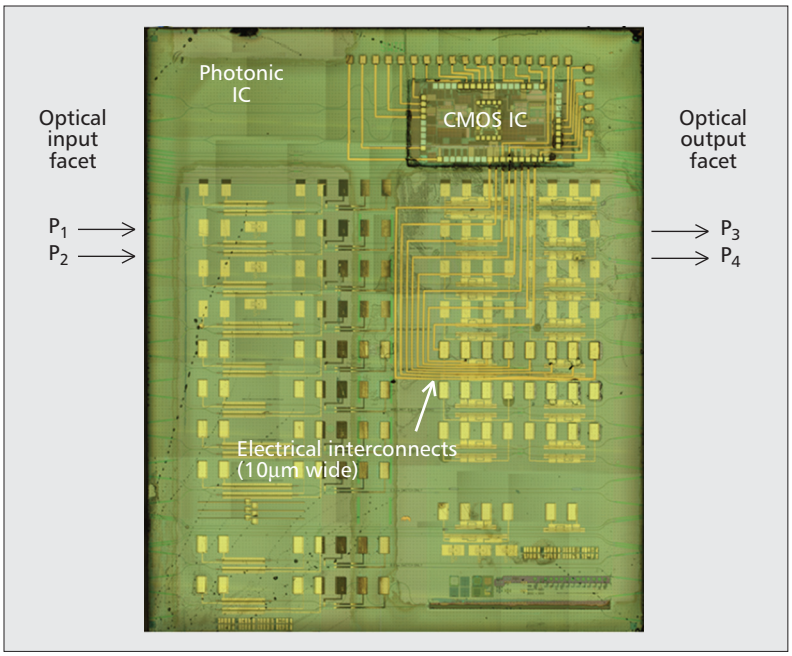

Figure 3. *Integrated electronic and photonic chip with metalized contacts interfacing both chips.*

Burst mode receiver	Performance summary
Bandwidth	2.5 GHz
Preamble time	13 ns
Power consumption	43.6 mW per channel
Area	0.23 mm^2
Switch driver	
Area	0.01 mm^2
Rise time	3.8 ns (20–80%)
Power consumption	150 mW per driver
Adjustable output	Yes
High power supply rejection	Yes
Current-to-frequency converter	
Input current range	10 μA to 1.5 mA
Area	0.005 mm^2
SOA driver	
Control word width	6 bits
Minimum current	3.3 mA
Maximum current	223 mA
Switching speed	10 MHz
Area	0.015 mm^2

Table 1. *Performance summary of the CMOS controller IC.*

removed. However, the most serious drawback of this approach is that length scales between electronics and photonic devices are large, resulting in underutilized silicon area and leading to high cost and low yield.

In the front-end process photonics and electronics structures share the chip footprint, leading to moderate integration density. Some disadvantages of this approach are that the thermal budget of the process needs to be rather high to enable high-temperature processes. However, active devices still need to be external to this system.

In the backside fabrication photonic layers are integrated on the backside of CMOS wafer. The backside approach is developed by Austriamicrosystems within the frame of the Photonics Electronics Functional Integration on CMOS (HELIOS) European Project. In this integration method it is difficult to fabricate on-chip lasers; thus, they are externally bonded onto the chip, making the packaging specific to laser assembly. Also, mismatch in the thermal expansion coefficients of the two materials is a serious concern. One of the key disadvantages of back-end or front-end integration is that one is constrained to use compatible electronic and photonic technologies rather than the best-in-class technologies in both domains.

In this section we describe our initial efforts in integrating both active and passive photonic devices with CMOS by developing a new technique of integration inspired by VCSIT [15]. In this method, PIC fabrication is independent of CMOS chip fabrication, enabling the use of best-in-class processes for both technologies. The CMOS chip is integrated in the PIC by etching a cavity in the PIC, placing the CMOS chip in it, planarizing it, and establishing metallic contacts.

The process begins with forming the cavities in the photonic wafer on the location where the electronic chips are to be integrated. A thin layer of polydimethysiloxane (PDMS) elastomer is spun onto the wafer to serve as the hard mask. The etch area is exposed by imprinting the electronic chip, expulsing the PDMS underneath the chip, and simultaneously heating the chip to induce crosslinking of PDMS surrounding the chip. This is accomplished using a flipchip bonder, with the arm of the bonder holding the chip and heating it. The hard mask is formed by curing all the remaining PDMS. With the sites exposed, the deep reactive ion etch (DRIE) Bosch process is used to define the cavities. Once the cavities are formed, the electronic chip is dropped into the wafer. Alignment between the electronic chip and the photonic wafer is accomplished by aligning the metal contact pads using an infrared optical microscope. The front side of the wafer is aligned with the electronic chip. Since the electronic chip is made thinner,

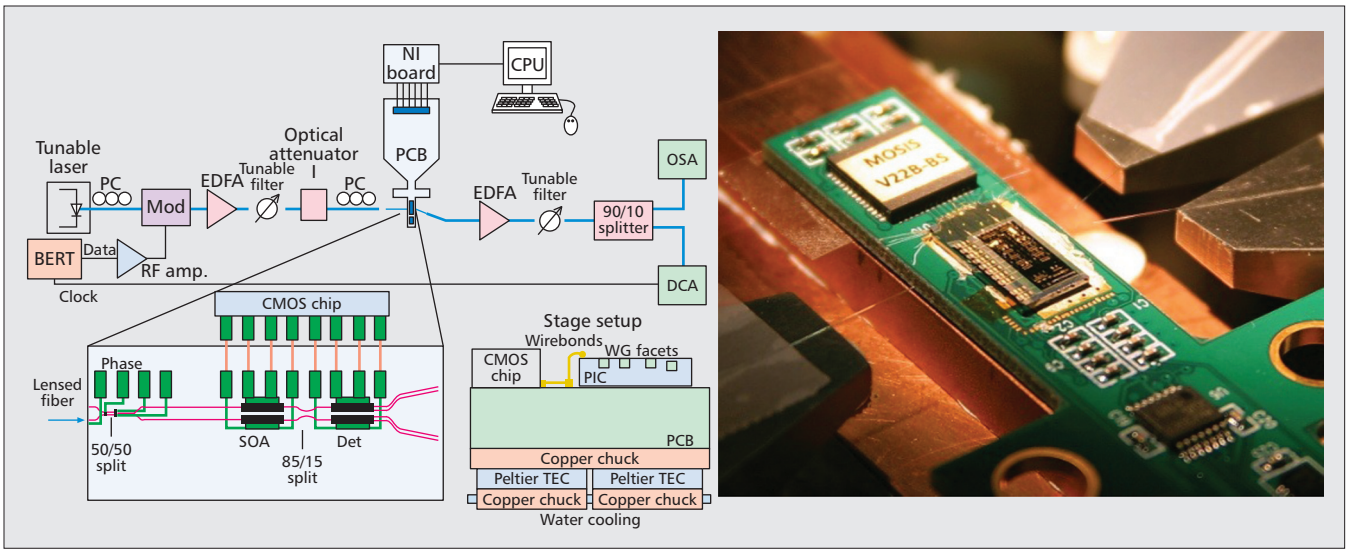

Figure 4. *Test setup for the automatic equalization loop.*

the cavity is filled with spin-on glass (SOG). The front side is passivated with SU8 and planarized using chemical mechanical polishing (CMP). Contacts between the electronic chip and the photonic wafer are lithographically defined; therefore, dense interconnects can be realized. The process ensures that the gap between the electronic chip and the photonic wafer is small so that metal lines can be defined across the chip boundary without breaking. Figure 3 shows a micrograph of the integrated chip. The small CMOS electronic chip can be seen embedded into the photonic chip with gold interconnects contacting the pads on both chips.

The proposed integration technique decouples the electronic and photonic fabrication processing, giving designers of photonic systems the freedom to choose the best technology available in a cost-effective manner. For instance, while CMOS is preferred for digital information processing, gallium arsenide (GaAs) can be used in radio frequency (RF) and high-voltage applications.

Demonstration of the Automatic Equalization Control

A test setup to demonstrate the automatic gain control loop operating on modulated data is shown in Fig. 4. A 1550 nm laser signal is modulated at 10 Gb/s and amplified to 15 dBm. The signal is then wavelength filtered and sent through an optical attenuator to vary the input power at the facet of the PIC. The electronic IC monitors the detector current and adjusts the SOA current to give constant output power. The PIC output signal is amplified, wavelength filtered, and monitored on an optical spectrum analyzer and an oscilloscope. Figure 5 shows the measured output optical power as a function of the input power with the control loop turned on. The system regulates the output to about –6 dBm, excluding coupling losses on the facets (–16 dB). From 0 to 25.4 dBm input, the average output is –6.98 dBm with standard deviation of 0.33 dB.

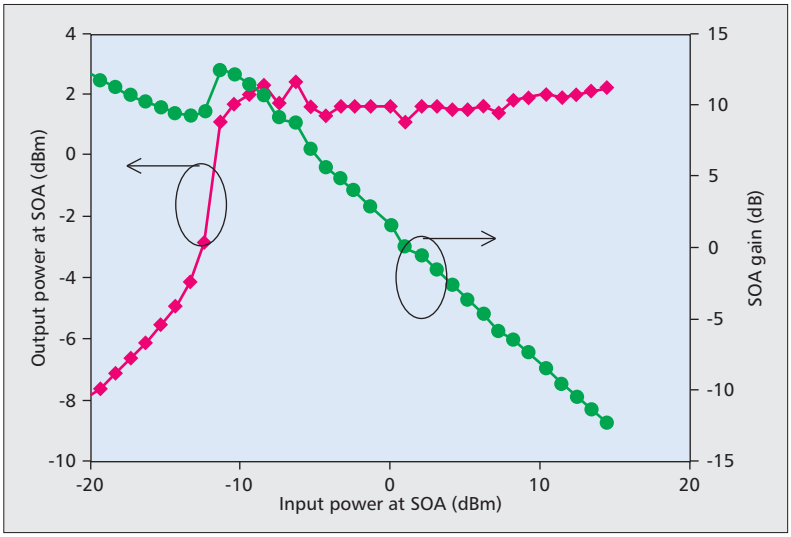

Figure 5. *Result with automatic power equalization.*

Scaling to Higher-Port-Count Switches

To support data-center-scale networks, the switch port count must be able to scale. Figure 6 shows the switch loss as a function of port count, assuming Benes and dilated Benes network configurations. To make the switch cascadable, the switch losses must be compensated using SOAs. As shown in this work, the SOA can compensate for up to 12 dB of loss. For example, to build a 32 × 32 port switch, two SOAs are necessary per port. As the number of ports increases, the number of SOAs required also increases. Scaling beyond 32 ports remains challenging. For instance, at 1024 ports, a Benes arrangement requires up to 19 stages of switches. New design

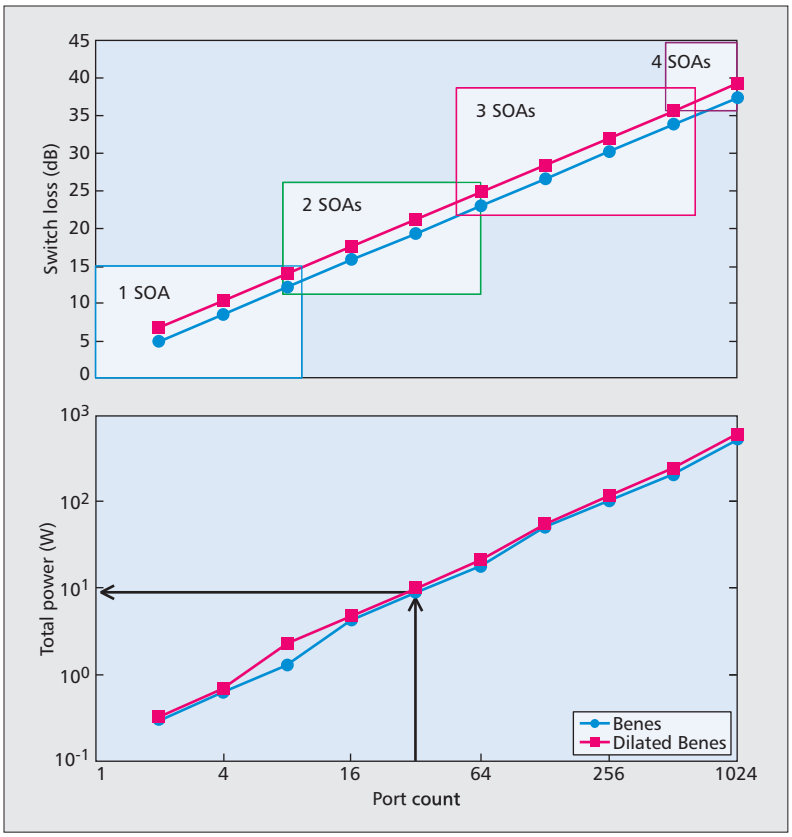

Figure 6. *Estimated switch loss and total power versus port count.*

techniques to improve crosstalk from the SEs and the amplified spontaneous emission (ASE) from the SOAs will be needed.

Figure 6 shows the total power consumption for Benes and dilated Benes networks, including the electronic power consumption. If a 32 × 32 switch were to be built today, it would consume about 8.7 W of power. At 100 Gb operation, it would consume 273 mW per port, or about 2.7 pJ/b/port. In contrast, the state-of-the-art electronic crossbar switch, represented by Vitesse VCS-3144-12, consumes 1.3 W/port at 100 Gb, or 13.1 pJ/b/port. In comparison, the photonic switch consumes 4.8× lower power. Clearly, scaling the photonic switch to larger port counts is advantageous for data centers. The estimated area for the entire photonic integrated circuit for a 32 × 32 switch would be able to fit in a footprint of 32 × 12 mm^2 of chip area, while the electronics would fit in a silicon area of 34 × 27 mm^2. Furthermore, the combination of space switching with TDM and WDM can enable multiplicative scaling of the number of ports, thus enabling data center scaling networks that otherwise might not be feasible to build with a single technology.

Conclusions

Scaling data center network capacity to meet the ever increasing data demands in a sustainable fashion will require technologies that can switch and route data efficiently. Optical networks have remarkable properties, such as low loss and high bandwidth, that make them suitable for long-haul communication infrastructures and are increasingly being deployed in data centers. Using electronics closely integrated with the photonic integrated circuit is important to scaling to large switch sizes because:

- The integration of electronics and photonics allows low-power high-impendence drives in high-speed switches.
- The integration of electronics for a "smart" photonics chip allows drive and bias levels of individual 2 × 2 switches to be optimized along with maintaining appropriate optical power levels throughout the chip.

Acknowledgment

The authors thank Aurrion for fabricating the PIC and E. Kintzer for useful suggestions. This work was supported by the Defense Advanced Research Projects Agency (DARPA) MTO EPHI project and by the Intelligence Advanced Research Project Agency under contract W911QX-11-C-0112.

References

[1] A. Vahdat *et al.*, "The Emerging Optical Data Center," *OFC 2011*.
[2] X. Zheng *et al.*, "Three-Dimensional MEMS Photonic Cross-Connect Switch Design and Performance," *IEE J. Sel. Topics Quantum Electronics*, vol. 9, no. 2, Mar. 2003, pp. 571–79.
[3] M. Glick, "Optical Switching for Next Generation Data Centers," *Photonics in Switching*, Sept. 2009.
[4] E. F. Burmeister *et al.*, "Photonic Integrated Circuit Optical Buffer for Packet-Switched Networks," *Optics Express*, vol. 17, no. 8, Apr. 7, 2009, pp. 6629–35.
[5] G. Wang *et al.*, "c-Through: Part-time Optics in Data Centers," *SIGCOMM 2010*.
[6] N. Farrington *et al.*, "Helios, a Hybrid Electrical/Optical Switch Architecture for Modular Data Centers," *SIGCOMM 2010*.
[7] K. Chen *et al.*, "OSA: An Optical Switching Architecture for Data Center Networks with Unprecedented Flexibility," *Symp. Networked Sys. Design and Implementation*, Apr. 2012.
[8] N. Farrington *et al.*, "Hunting Mice with Microsecond Circuit Switches," *ACM Wksp. Hot Topics in Networks*, Oct. 2012.
[9] Yang *et al.*, "Non-Blocking 4×4 Electro-Optic Silicon Switch for On-Chip Photonic Networks," *Opt. Express*, vol. 19, no. 1, 2011, pp. 47–53.
[10] D. Liang *et al.*, "Low-Temperature, Strong SiO2-SiO2 Covalent Wafer Bonding for III-V Compound Semiconductors-to-Silicon Photonic Integrated Circuits," *J. Electronic Materials*, vol. 37, no. 8, June 1, 2008.
[11] L. Chen *et al.*, "Digitally Tunable Optical Power Qualization for Large Port Count Optical Switches," *OFC/NFOEC*, 17–21 Mar. 2013, pp. 1, 3.
[12] A. V. Krishnamoorthy *et al.*, "3-D Integration of MQW Modulators Over Active Submicron CMOS Circuits: 375 Mb/s Transimpedance Receiver-Transmitter Circuit," *IEEE Photonics Tech. Letters*, vol. 7, no. 11, Nov. 1995, pp. 1288–90.
[13] T. Pinguet *et al.*, "Monolithically Integrated High-Speed CMOS Photonic Transceivers," *IEEE Int'l. Conf. Group IV Photonics*, Sept. 2008, pp. 362–64.
[14] C.Kopp *et al.*, "Silicon Photonic Circuits: On-CMOS Integration, Fiber Optical Coupling, and Packaging," *IEEE J. Sel. Topics in Quantum Electronics*, vol. 17, no. 3, May–June, 2011, pp. 498–509.
[15] A. Uddin *et al.*, "Wafer Scale Integration of CMOS Chips for Biomedical Applications via Self-Aligned Masking," *IEEE Trans. Components Packaging Manufacturing Tech.*, vol.1, no.12, Dec. 2011, pp. 1996–2004.

Biographies

Luis Chen received his B.S., M.S., and Ph.D. in electrical engineering in 2006, 2007, and 2013, respectively, from the University of California, Santa Barbara (UCSB). His areas of interest are low-power and high-speed analog and

mixed-signal integrated circuits. He was a recipient of UCSB's Institute for Energy Efficiency Holbrook Foundation fellowship in 2012–2013.

AVANTIKA SOHDI received her B.Tech. in instrumentation and control from Netaji Subhas Institute of Technology, Delhi, India, in 2009. She joined the Biomimetic Circuits and Nanosystems Group at UCSB in fall 2010 and is currently a graduate student researcher in the Department of Electrical Engineering at UCSB under the supervision of Prof. Luke Theogarajan. Her current research interests include optoelectronic integration for the design of ultra-fast low-power optical packet switching.

JOHN E. BOWERS [F] (bowers@ece.ucsb.edu) holds the Fred Kavli Chair in Nanotechnology, and is the director of the Institute for Energy Efficiency and a professor in the Departments of Electrical and Computer Engineering and Materials at UCSB. He is a cofounder of Aurrion, Aerius Photonics, and Calient Networks. He received his M.S. and Ph.D. degrees from Stanford University, and worked for AT&T Bell Laboratories and Honeywell before joining UCSB. He is a member of the National Academy of Engineering, and a Fellow of the OSA and American Physical Society. He is a recipient of the OSA/IEEE Tyndall Award, the OSA Holonyak Prize, the IEEE LEOS William Streifer Award, and the South Coast Business and Technology Entrepreneur of the Year Award. He and coworkers received the *EE Times* Annual Creativity in Electronics (ACE) Award for Most Promising Technology for the hybrid silicon laser in 2007. His research is primarily in optoelectronics and photonic integrated circuits. He has published 10 book chapters, 600 journal papers, and 900 conference papers, and has received 54 patents. He has publised 180 invited and conference papers, and given 16 plenary talks at conferences.

LUKE THEOGARAJAN is currently an associate professor at UCSB in electrical engineering and computer science from the Massachusetts Institute of Technology in 2007. Before starting his Ph.D., he worked for Intel for five years, where he was part of the Pentium 4 design team. He was a 2010 NIH New Innovator Award recipient and a 2011 NSF Career Award recipient. He was awarded the Northrupp Grumman Excellence in Teaching award in 2011, and outstanding faculty member in electrical engieering for 2009, 2010, 2011, and 2012. His research interests are low-power circuits for biological systems, high-speed circuits for optoelectronics, advanced packagint for both biosensors and electronic-photonic integration, nanoscale sensors, and supramolecular chemistry.

JONATHAN E. ROTH received his B.S. degree in biomedical engineering from Case Western Reserve University in Cleveland, Ohio, in 2000, and his Ph.D. degree in electrical engineering from Stanford University, California, in 2007. His dissertation work was on electroabsorption modulators in indium phosphide and silicon germanium. He is employed by Aurrion Inc. as an optoelectronic device engineer, where he designs heterogeneous III-V on silicon devices and photonic integrated circuits.

GREG FISH is the chief technical officer at Aurrion and a leading expert in the field of photonic integration with over 15 years of experience in the field of InP-based photonic integrated circuits (PICs). He began his work in this area while obtaining a B.S. in electrical engineering from the University of Wisconsin at Madison in 1994, and later his M.S. and Ph.D. in electrical engineering from the University of California in Santa Barbara in 1999, where his research focused on optical crossbar switches on InP. He is an author/coauthor of over 40 papers in the field and has 10 patents. He was a cofounder and key leader at Agility Communications from 1999 to 2005, which was formed to commercialize tunable lasers based on PICs for telecommunications applications using technology pioneered at UCSB. This business was acquired by JDSU in late 2005, where it subsequently grew to nearly $100 million in revenue per year. While at Agility and JDSU, he applied innovation and leadership to bring the first generation of products to market, helped demonstrate the future potential of tunable PICs and provided the basis to launch the next generation of tunable transceivers at JDSU. He has been a key member of several multimillion-dollar DARPA projects (DODN, CS-WDM) and performed as the principal investigator for equally large projects at NIST ATP and ONR. At present, he is the principal investigator for many of Aurrion's projects, the largest of which is the DARPA DAHI E-PHI project where he is responsible for organizing the development of leading-edge heterogeneous silicon photonics technology.