Beyond CPO: A Motivation and Approach for Bringing Optics Onto the Silicon Interposer

Benjamin G. Lee, *Senior Member, IEEE*, Nikola Nedovic, *Member, IEEE*, Thomas H. Greer III, *Member, IEEE*, and C. Thomas Gray, *Senior Member, IEEE*

(Invited Paper)

***Abstract*—Co-packaged optics (CPO) technology is well positioned to break through the bottlenecks that impede efficient bandwidth scaling in key near-term commercial integrated circuits. We begin by providing some historical context for this important sea change in the optical communications industry. Then, motivated by GPU-based accelerated computing requirements, we investigate the next pain points that are poised to constrain bandwidth and efficiency in future CPO-based systems. We identify 2.5D integrated optics (i.e., bringing optics onto the interposer) as a promising solution that can enable continued scaling for these systems due to the dense wiring available which facilitates more efficient slow-and-wide electrical interfaces. We explore the benefits, challenges, and requirements associated with such tight coupling of the processors and optical engines by considering high-level photonic link design, technology, and packaging. We demonstrate the viability of a control loop which can adequately regulate temperature within the aggressive thermal environment. Then, we introduce a custom simulation framework that allows quantified comparisons of detailed design decisions; the simulations validate the feasibility of the general approach while also providing key guidance to designers on best directions to pursue for efficient optimization.**

***Index Terms*—Chip-scale packaging, photonic integrated circuits, photonics.**

I. Introduction

With the advent of low-loss fiber and fiber amplifiers in the 1970s and 1980s, optical data transmission began to reshape global communication infrastructure. Today, optical data transmission continues to be a disruptive force in communication systems, and not just at telecommunications distance scales. Since the 1980s, fiber-optic transmission has been replacing electrical data transmission at shorter and shorter distances as Table I highlights. A major milestone came in the 2010s when optics began being used copiously inside singular systems, such as datacenter networks [1] or high-performance computing (HPC) systems [2], rather than only within networks that connect disparate machines in separate locations. This paradigm shift was driven by the needs of the new cloud infrastructure, with its rows and rows of compute racks, necessitating cheaper and more scalable optics to support the ever-growing bandwidth demands. The shift furthermore caused an increase in optical transceiver (TRX) volumes that drove an ecosystem supporting generations of pluggable optical modules that have scaled in bandwidth from a few gigabits per second to nearly one terabit per second today. Now, we are on the cusp of another major disruption with the introduction of co-packaged optics (CPO), where optics are moved onto the same package as the host application-specific integrated circuit (ASIC) to facilitate higher bandwidths at reduced energies. Despite early system deployments of CPO a decade ago [3], a host of recent research demonstrations [4], [5], [6], [7] and product-related statements for CPO-integrated switch ASICs [8], [9], [10], [11] point toward broad market adoption within the next few years.

In this work, we endeavor to look beyond today's CPO product development efforts and investigate the demands of the next generation of computing hardware—represented by the light gray shaded region in Table I—that drive optical integration even closer to the ASICs. We motivate a need for further integration of optical engines onto the silicon interposer, next to the ASIC [12], [13], [14] (though we do not see a need to have optics embedded within the silicon interposer in this timeframe [15], [16]). This is driven by a very steep energy penalty associated with continual scaling of baud rates and the unending demand for increasing chip input/output (I/O) bandwidths, resulting in bandwidth bottlenecks at the package electrical interface. In Section II we identify Accelerated Computing as a major application driver and consider I/O requirements of future accelerated computing chips. Section III explores the energies and bandwidth densities needed for the electrical interfaces between the ASICs and the optics, and Section IV motivates design choices for selection of 2.5D integrated optical TRX architectures. Section V presents a link model that validates the approach while exploring various design tradeoffs. We conclude our paper in Section VI.

II. Accelerated Computing

The world's HPC systems are increasingly leveraging graphics processing units (GPU) to accelerate computation. Of the top

Manuscript received 9 July 2022; revised 14 October 2022; accepted 17 October 2022. Date of publication 4 November 2022; date of current version 16 February 2023. *(Corresponding author: Benjamin G. Lee.)*

Benjamin G. Lee is with the NVIDIA, Ridgefield, CT 06877 USA (e-mail: benjlee@nvidia.com).

Nikola Nedovic is with the NVIDIA, Santa Clara, CA 95051 USA (e-mail: nnedovic@nvidia.com).

Thomas H. Greer and C. Thomas Gray are with the NVIDIA, Durham, NC 27713 USA (e-mail: tgreer@nvidia.com; tgray@nvidia.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JLT.2022.3219379.

Digital Object Identifier 10.1109/JLT.2022.3219379

TABLE I
HISTORICAL PERSPECTIVE OF OPTICAL INTEGRATION TRENDS

Market	Optical TRX Form Factor	Optical TRX Location	Decade of Mass Adoption	Distance Scale	Electrical Technologies Displaced	Enabling Optical Technologies
Longhaul	Discrete	Rack	1980s	> 100 km	Twisted pair	Low-loss fiber, fiber amplifiers
Metro area	Discrete	Rack	1990s	10 – 100 km	Twisted pair	WDM
Local area	Pluggable	Edge of card	2000s	100 m – 10 km	Twisted pair, coaxial cable	VCSELs, DMLs
Datacenter, HPC	Active optical cable, pluggable	Edge of card, on board	2010s	10 m – 1 km	Passive copper, direct-attach copper	VCSELs, Si photonics
Datacenter, HPC, AI	CPO	On package	2020s	100 mm – 1 km	PCB traces, flyover cables	Si photonics
	CPO	On interposer		> 10 mm	Package traces	
	3DIC	On die		> 1 mm	Silicon wiring	

3DIC: three-dimensional integrated circuit, AI: artificial intelligence, CPO: co-packaged optics, DML: directly modulated laser, HPC: high-performance computing, PCB: printed circuit board, TRX: transceiver, VCSEL: vertical cavity surface emitting laser, WDM: wavelength-division multiplexing.

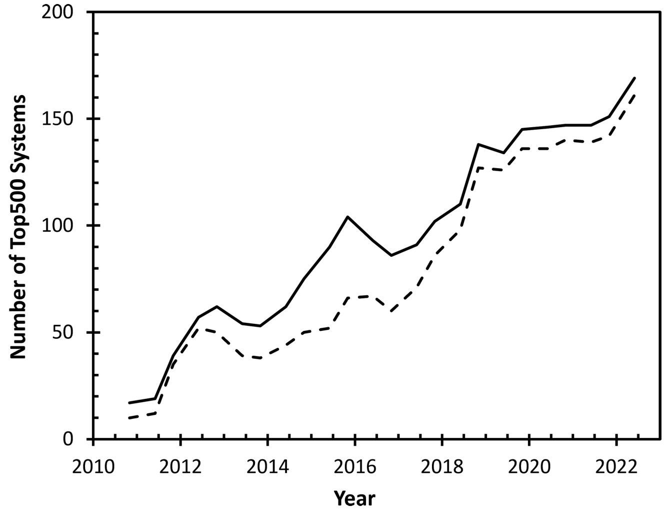

Fig. 1. Data showing, of the 500 highest-ranking computing systems, the number using an accelerated co-processor architecture (solid), and the number of those using NVIDIA GPUs (dashed) over the last 12 years [17].

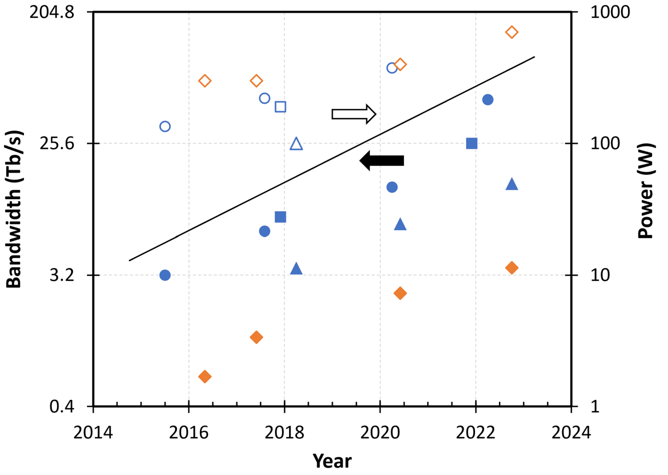

Fig. 2. Published I/O bandwidth (filled symbols) and power (open symbols) for recent NVIDIA switch (blue) and GPU (orange) ASICs versus year of product launch, including Spectrum series Ethernet switches (●) [23], [24], Quantum series Infini-Band switches (■) [25], [26], NVSwitch series NVLink switches (▲) [27], [28], and four GPU generations (♦) [28], [29], [30].

500 HPC machines currently ranked by top500.org [17], 169 use accelerated co-processor architectures. Fig. 1 illustrates the growth of this trend over time, along with the fraction of accelerated computing systems that use NVIDIA GPUs. Moreover, accelerated computing machines are weighted toward the upper ranks, with 4 of the top 8 machines using NVIDIA GPUs to speed up computation. In addition to traditional scientific applications, areas exploiting data science—including artificial intelligence (AI) and machine learning—leverage GPUs to increase automation and improve efficiencies across science and business. Finally, accelerated computing in the cloud [18] provides broad access to GPU-based accelerators. Thus, a GPU-centric cloud computing infrastructure is emerging that combines the benefits of datacenters, supercomputers, and AI accelerators. For example, NVIDIA's DGX platform [19] can be used as a local AI accelerator, scaled out to realize a top-tier supercomputer [20], or accessed virtually through a variety of cloud platforms.

In these emerging systems, high-performance switched interconnection networks efficiently route communication between a growing number of compute resources. The networks have become pivotal not only to scale out performance across a large system—as in traditional HPC and datacenter networks, e.g., [21]—but increasingly also to scale up the performance of each node. As a case study, four 3rd-generation NVSwitches provide up to 54.4 Tb/s of aggregate bandwidth to a local network of eight H100 GPUs in the recently announced DGX H100 system [22]. The DGX boxes may be scaled into a GPU cluster through an optically connected NVLink network or attached to an existing InfiniBand or Ethernet fabric.

Future system scaling therefore relies on continual improvements in switch bandwidths as well as processor performance. Fig. 2 plots a sampling of power consumption and I/O bandwidth from NVIDIA GPUs and switch ASICs over the past several years. Switches have roughly doubled in bandwidth every two years. While GPU I/O bandwidths lag switch bandwidths by about 10×, the chip power envelopes are comparable. As chip performance has scaled, so has the power, which is approaching 1 kW and requiring complex cooling solutions. A major factor contributing to switch chip power is the off-chip I/O, which has been accounting for an increasing portion of the total [10]. Fundamentally, this is occurring because chip and package pin counts (determined by component size and pin pitch) are scaling slowly compared to steeply rising bandwidth demands. This forces a rapid increase in the signaling rates per pin, which comes at the cost of energy efficiency. Reducing the I/O power by minimizing the length of the signal paths and by moving toward lower speed signaling with more parallelism is critical for increasing chip performance in future generations, both for switches and GPUs. In the next section, we explore the performance and limitations of the current electrical interfaces.

III. Electrical Interfaces

The Optical Internetworking Forum's (OIF) Common Electrical I/O (CEI) 112-Gb/s long reach (LR) standard [31] provides for 1-m of reach over a twin-ax cable or printed circuit board (PCB) trace with 2 connectors. Measured energies of demonstrated 112-Gb/s LR interfaces are 4.5 to 6.5 pJ/b [32], [33], [34]. A 100-Tb/s switch using a purely electrical LR interface would consume at least 450 W, just for off-chip communication. LR interfaces are also used to connect to on-board or pluggable edge-of-card optics to extend reach. In addition to adding cost, the optical modules increase system power (typically by >10 pJ/b) while doing nothing to reduce the chip power. Clearly, more efficient interfaces are needed.

CPO may reduce chip power while also extending reach compared to purely electrical signaling. By integrating the optics on package with the ASIC, the electrical interface efficiency can be improved. The CEI-112G-XSR (extreme short reach) standard provides for up to 100 mm of electrical wiring on an organic multi-chip module (MCM). Demonstrations have achieved 1.24 to 1.7 pJ/b [35], [36], [37], [38]. For CPO, off-chip communication must traverse two electrical links and one optical link, and all three dissipate power within the switch package. Therefore, XSR interfaces to/from CPO in a 100-Tb/s switch package are predicted to consume ∼250 to 350 W (not counting the optical link portion). This may be an improvement over the chip power dissipation of an LR ASIC, but only if the optics are quite efficient. Moreover, XSR interfaces have shown electrical edge bandwidth densities ranging from 475 to 870 Gb/s/mm [35], [36], [37]. Though further improvement is needed, these densities are nearing the 100-Tb/s switch requirements (i.e., ∼2 Tb/s/mm assuming 100-mm chip perimeter with 100 Tb/s ingress and 100 Tb/s egress). Therefore, for the 100-Tb/s generation CPO on MCM has the potential for modest power savings at the switch module, and—with further improvement—potentially enough bandwidth density. Finally, although the module power may not be dramatically reduced with CPO on MCM, the overall system power likely will be, since the power consumed by the pluggable optics can be eliminated from the system.

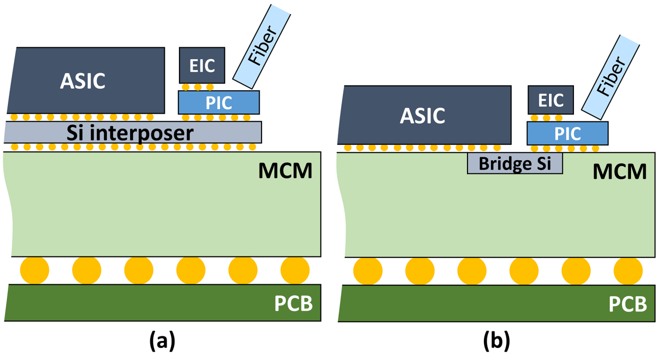

Fig. 3. Two concepts for 2.5D integration of photonics alongside an ASIC. In (a) the PIC is integrated on the interposer, whereas in (b) the PIC is integrated on the organic MCM substrate but connected to the ASIC via a bridging layer of local silicon interconnect.

Scaling beyond 100 Tb/s will require denser integration with electrical edge bandwidth densities of multiple Tb/s/mm and full link (electrical + optical + electrical) energy efficiencies of 1-2 pJ/b. For this, 2.5D integration on silicon interposer or local silicon interconnect will be required [39], [40], [41]; two potential configurations are illustrated in Fig. 3. Closer integration reduces the transmission distance, but more importantly denser wiring allows the per-wire rates to be relaxed, which significantly improves energy efficiency. Recent results highlight the opportunity for a dense and efficient on-interposer interface: a 50-Gb/s link in 5-nm CMOS was demonstrated across a 1.2-mm silicon channel consuming 0.3 pJ/b and achieving an edge bandwidth density of >2 Tb/s/mm with scalability to >10 Tb/s/mm [42].

IV. On-Interposer Optical Engines

Integrating optical engines (OE) on the same interposer as the ASIC creates many challenges. The bandwidth density (both edge and areal) and energy per bit of the electrical integrated circuit (EIC) and photonic integrated circuit (PIC) which comprise the OE becomes even more critical. As Fig. 4 illustrates, the perimeter of a module using CPO on MCM (e.g., ∼400 mm) is most often several times longer than the perimeter of the ASIC (e.g., ∼100 mm). Here, the bottleneck is in the electrical edge bandwidth density attainable on the organic substrate, while the optical edge bandwidth density in the fiber attach region is comparatively relaxed due to beachfront expansion. In contrast, there is very little beachfront expansion on the interposer since trace lengths must be kept short and silicon area is limited. So, while bringing the optics onto the interposer allows significant improvement in the electrical edge bandwidth density between the ASIC and the optics (due to denser wiring), it also requires that the optics achieve similar edge bandwidth density at the fiber interface. The electrical interface energy can be substantially reduced as a consequence of the shorter reach and lower baud rates, as argued above. Our efforts are targeting electrical interface energies of 0.25 pJ/b with an optical link energy of 1 pJ/b and a remote laser source consuming about 2 pJ/b, resulting in 1.5 pJ/b dissipated in the module and 3.5 pJ/b

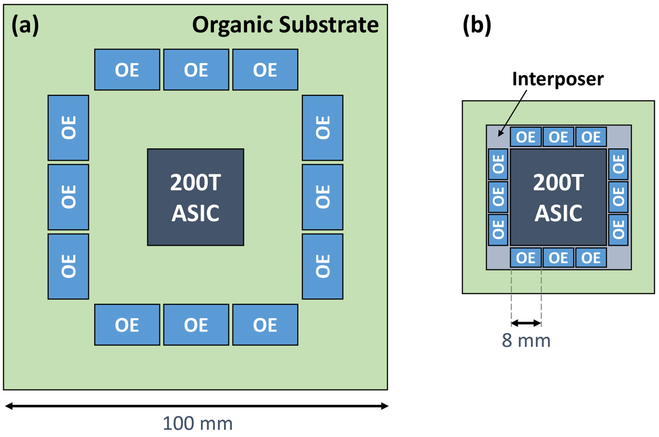

Fig. 4. Top-view package diagrams illustrating the limitations on beachfront expansion for 2.5D integrated optics. The diagrams compare a 200-Tb/s ASIC with a 25-mm edge using (a) CPO-on-MCM and (b) 2.5D integrated optics. In (a) the electrical bandwidth density between the OE and the ASIC is limited by the organic substrate density, while the optical bandwidth density is comparatively relaxed due to the availability of the larger package edge. In (b) the electrical bandwidth density is improved due to Si wiring density, while the optical bandwidth density is constrained by the interposer.

overall [13]. Next, we describe high-level design considerations for the laser source modules, the photonic link architecture, and the package.

A. Laser Source

Remote lasers eliminate both area and power from the ASIC package, as well as improve laser performance and lifetime, at the expense of more coupling loss and less optical edge bandwidth density (since supply fibers are needed in addition to transmit and receive fibers). The added coupling loss must be compensated by higher laser power, but this increase does not contribute to the in-package power envelope. Remote laser sources capable of producing a comb of dense wavelengths (see Section IV.B) typically fall into one of three categories: (1) component-level assembly of a III-V chip, consisting of an array of DFB lasers emitting at different wavelengths, with passive optical combiners integrated in an oxide or nitride platform [43], [44]; (2) bonding of III-V laser gain material on silicon to produce an array of DFB lasers on the same substrate with fixed or tunable silicon photonic devices used to combine wavelengths [45], [46], [47]; and (3) single cavity mode-locked lasers that produce the set of wavelengths from a single source with no combiner needed [48]. These options provide an array of cost-benefit tradeoffs at various maturity levels.

B. Photonic Link Architecture

The optical edge bandwidth density can be scaled along the dimensions of time, frequency, space, polarization, phase, and amplitude, but there are limitations. Spatial density is constrained by practical aspects of fiber manufacturing, which place useful fiber diameters in the vicinity of 80-125 μm. Overlapping spatial modes (e.g., mode-division multiplexing) generally require digital signal processing (DSP) for reliable demultiplexing. Baud rates are difficult to scale without incurring steep energy penalties [49]. Polarization provides a factor of two at the price of design complexity, area, and cost. Limited signal-to-noise ratios dictate that scaling through the amplitude (e.g., pulse amplitude modulation, PAM) and phase (e.g., quadrature phase-shift keying, QPSK) domains extract a penalty in cost and power for DSP. All of these can provide some scaling; however, the frequency (i.e., wavelength) domain may be the least constrained, although there are yet challenges in realizing low-cost multi-wavelength lasers and in overcoming optical link losses.

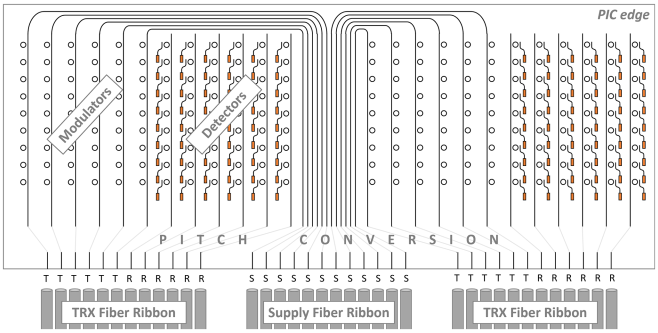

Fig. 5. PIC architecture diagram, for simplicity drawn employing 12 channels and 8 wavelengths per channel with an extra ring on each bus dedicated to clock forwarding. Each PIC connects to two TRX fiber ribbons—each with 6 TX (T) + 6 RX (R) fibers—and one 12-fiber supply (S) ribbon with each fiber carrying 9 continuous-wave wavelength channels. (For consistency with Fig. 4, the PIC would require 21 channels carrying 800+800 Gb/s.).

Coarse wavelength-division multiplexing (CWDM) systems are in use today, but dense WDM (DWDM) will be needed to simultaneously meet tomorrow's bandwidth density and power efficiency targets. At 250-μm fiber pitch, an initial 8-wavelength single-polarization 25-Gbaud PAM-2 (i.e., non-return-to-zero, NRZ) DWDM link can achieve a raw bandwidth density of 0.8 Tb/s/mm. By scaling 2× in each dimension, a 16-wavelength dual-polarization 50-Gbaud PAM-4 link with 127-μm fiber pitch can achieve a raw bandwidth density of 25.6 Tb/s/mm. These raw bandwidth densities must still be allocated between supply, transmit, and receive fibers with some overhead included for fiber attach; however, the example illustrates the scaling potential.

A micro-ring resonator-based link architecture, as illustrated in Fig. 5, provides several benefits. It is DWDM compatible. It does not require grating- or interferometer-based multiplexer/demultiplexers, which occupy significant area. The ring modulators and ring filters are area and energy efficient [43], [50], and capable of scaling well beyond 100 Gb/s [51]. The lumped-element modulators simplify driver circuits and are capable of operating at reasonable voltage swing. Micro-ring resonators require closed-loop control, but the power and area overheads can be relatively small in advanced CMOS nodes.

Such an architecture can deliver the efficiencies needed for 2.5D integrated optical engines, but many other challenges remain. For one, packaging becomes much more constrained. Socketed electrical connectors and pigtailed optical fibers cannot be used. Optical facets must be preserved through a wafer-scale assembly process. Optical connector density should be dramatically increased. These challenges seem plausibly resolvable. However, of paramount concern is the question of the thermal environment. We consider this in the next section.

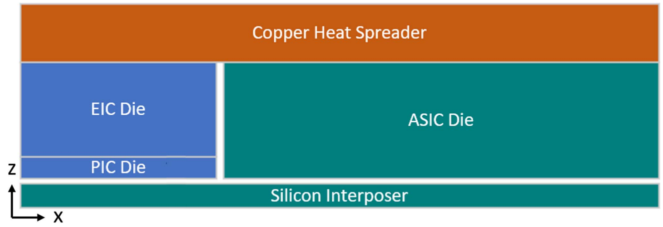

Fig. 6. Cross section of the simulated thermal environment with rings located on the upper surface of the PIC die.

C. Packaging and Thermal Considerations

In this section, we investigate the effects of the thermal environment on link performance, including control of the thermo-optic tuning elements that are used to align resonances. We consider a linear time-invariant thermal system comprised of an OE which is integrated on the silicon interposer next to an ASIC [Fig. 3(a)]. Such a system is completely determined by the geometry in combination with the material properties, such as thermal resistivities, specific heats, and densities (all of which we assume are temperature independent). Fig. 6 shows a cross section of the simulated structure. The interposer and PIC are 100-μm thick to allow through silicon vias. The ASIC is 1-mm thick to support neighboring chip stacks. The EIC/PIC stack is separated from the ASIC by a 70-nm spacer of plastic molding compound. The copper heat spreader is 0.5-mm thick. All simulation boundaries are insulating except for the upper surface of the heat spreader which has a thermal impedance to ground of 3e-5 $m^2 \cdot K/W$.

The optical transceiver depends on thermally tuned ring resonators on the top surface of the PIC. In our analysis, the ring temperatures are maintained by metal heaters located in glass 1 μm above each ring. The rings are insulated from the silicon below by 2 μm of buried oxide, and from the EIC active area above by $\sim$10 μm of back-end-of-line (BEOL) glass. A thermal control loop for each ring maintains the ring temperature in the presence of power transients in the ASIC and in neighboring ring heaters. (The EIC power is assumed to be constant.)

Thermal simulations are performed in Lumerical HEAT. The first, a 2D cross section simulation, determines the temperature response of the ring environment to a maximal 500-W power step distributed evenly across the bottom of the 800 mm^2 ASIC die. The resulting thermal response (Fig. 7) at typical ring positions (the two leftmost designated locations) shows a temperature rise of about 10 K with a maximum slope of about 70 K/s. The contour lines illustrate that some heat is moving laterally from the ASIC to the PIC through the silicon interposer. Heat flow from the ASIC directly to the PIC and EIC is largely blocked by the plastic molding compound between the two dice. Some heat does cross over through the copper heat spreader ($z > 1000$ μm) but does not significantly impact the ring temperatures. The insulating layer of BEOL glass between the PIC and EIC is visible at $z = 100$ μm.

A second thermal simulation, 3D this time, is performed to determine the ring response to a 1-mW step in heater power. This simulation is also used to quantify the effect of the same step in heater power on a neighboring ring 100 μm away. This second simulation employs the same cross section (Fig. 6), but with lateral extent restricted to ± 1 mm from the main ring location. The transient results (Fig. 8) depict the temperature change of the heater, the ring, and the neighboring ring versus time. The ring temperature rises in the steady state by about 11 K with a time constant near 10 μs. (This is in good agreement with another study [52].) The neighboring ring temperature rises by about 0.7 K with maximum slope of $\sim$50 K/s, which is similar to the slope caused by the 500-W step in ASIC power.

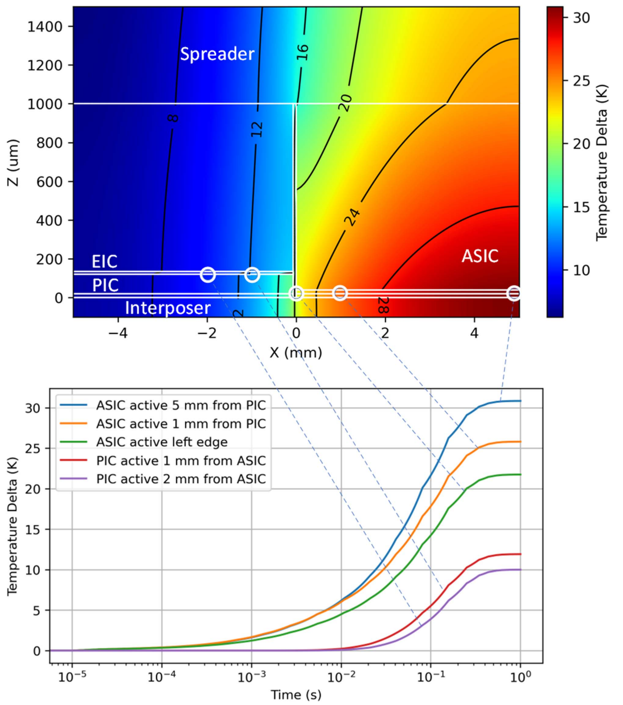

Fig. 7. (bottom) Transient temperature response of the system to a 500-W step in ASIC power at various locations along the PIC and ASIC. (top) Steady-state temperature profile of the simulated cross section at a time of 1 s. Note the different x and z scales in the upper figure.

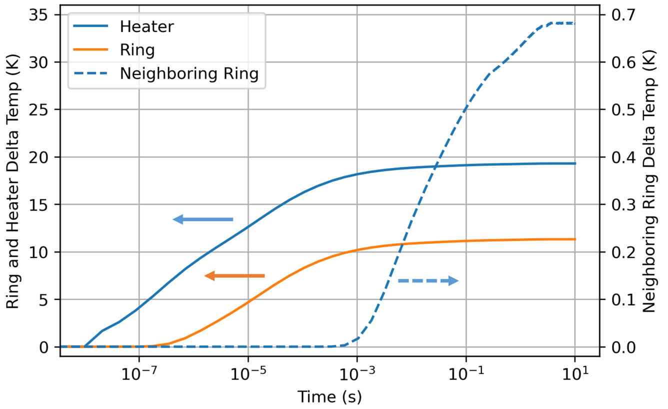

Fig. 8. Transient temperature response of the heater (blue solid), ring waveguide (orange solid), and neighboring ring (blue dashed) to a 1-mW step in heater electrical power dissipation.

Next, we must determine how precisely the control loop needs to maintain the ring temperature in the presence of the external

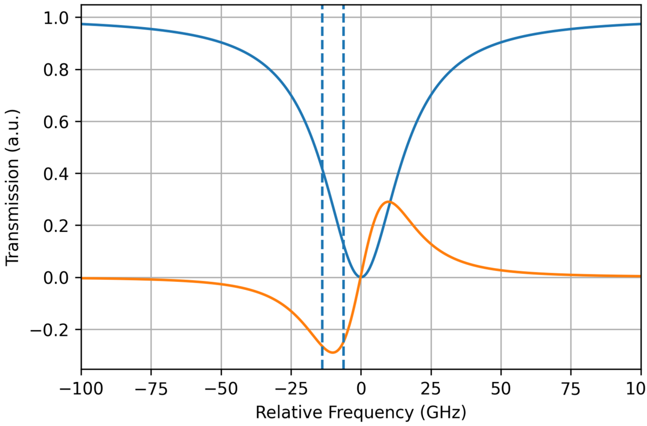

Fig. 9. TX ring optical response. The ring spectrum (blue solid) has a Q of 7000. A modulation swing of 8 GHz peak-to-peak is denoted (vertical dashed). The relative OMA, defined as the modulator's output OMA per input continuous-wave power, is plotted for the stated modulation swing versus TX frequency offset (orange solid). The control loop must maintain temperature with ample precision to remain in the maximum or minimum region of the relative OMA (about ± 5 GHz).

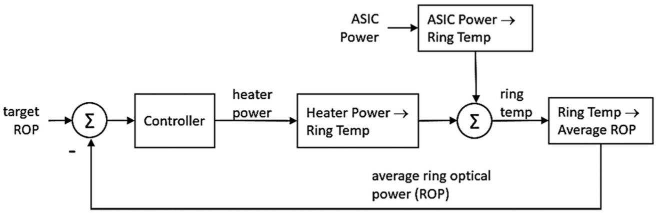

Fig. 10. Control loop block diagram.

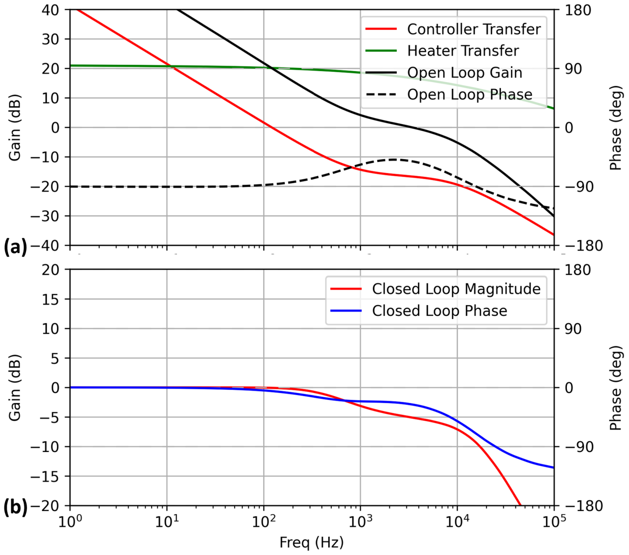

Fig. 11. (a) Open-loop frequency-domain characteristics of controller. (b) Closed-loop frequency-domain characteristics of controller.

temperature aggressors (i.e., the ASIC power swings and the neighboring ring heaters). For this, we need to explore in more detail the optical performance of the rings (Fig. 9). Ultimately, the control systems for transmitter (TX) and receiver (RX) rings will differ substantially. TX rings are tuned so that the laser is on the flank of the resonance to maximize optical modulation amplitude (OMA), while RX rings are tuned so that the laser is centered on the resonance to maximize drop-port power. For NRZ modulation at 32 Gbaud (see Section V), we assume a TX ring with quality factor (Q) of 7000 and a peak-to-peak frequency shift of 8 GHz. As indicated by Fig. 9, this requires a frequency control of ± 5 GHz (± 30 pm), which corresponds to a temperature stabilization of ± 0.4 K. For the RX side, we employed a frequency-domain analysis to verify that there is negligible eye closure for a ring with Q of 5000 and the same ± 0.4 K temperature stabilization as needed for the TX.

TX Loop: The TX temperature control loop block diagram is illustrated in Fig. 10. For the controller we use a conventional proportional-integral (PI) thermal controller [53], which transforms error temperature into heater power. We implement the controller with a proportional constant K_p of 0.15 and an integrating time constant T_i of 200 μs. The blocks labeled "Heater Power → Ring Temp" and "ASIC Power → Ring Temp" are derived from the thermal simulations described above. The outputs of the two thermal transfer functions, denoting the impacts of ASIC and heater powers on ring temperature, are summed to obtain the ring temperature. The ring temperature is then transformed by the ring's Lorentzian response curve to the average ring optical power (ROP), which is then compared with the target ROP, and the error is fed back to the controller. The transformation from temperature to ROP is a linear approximation; a crude estimate of the gain suffices to keep the ring sufficiently fast and stable. The optimum target ROP depends on the laser power level. The target value may be updated in a slow dithering loop to obtain maximum OMA based on feedback from the RX. A weakly coupled drop port and photodiode measure the TX ring's average optical power, with 8b/10b data encoding ensuring no data dependance.

RX Loop: In the RX we want to hold the optical power at a maximum (at resonance). We do this by dithering the heater power, then estimating the temperature error using the phase and magnitude of the fundamental frequency response of the optical power to the dither. The dither frequency is chosen sufficiently above the control loop bandwidth, but at a low enough frequency that the thermal mass of the ring allows an adequate response to the dither. We will see below that a trivially stable control loop bandwidth of 1 kHz is sufficient to keep the temperature error within the target bounds. The ring temperature response to heater modulation at 10 kHz is ∼5 dB down from DC, making 10 kHz an acceptable dither frequency.

Now we show the loop dynamics of a generic control loop assuming the thermal responses from simulation, the PI controller parameters mentioned earlier, and perfect temperature estimation. Fig. 11 shows the open- and closed-loop frequency-domain characteristics of the loop. The closed-loop transient response in Fig. 12(a) shows that the heater power swing required to handle 500-W transients in the ASIC is about 1 mW. The error temperature extreme of about 8 mK is well within the 400 mK target, even with the low loop bandwidth of 1 kHz. Fig. 12(b) shows that even a 1-mW step on a neighboring ring heater results in only a 7 mK closed-loop error temperature on the ring, again

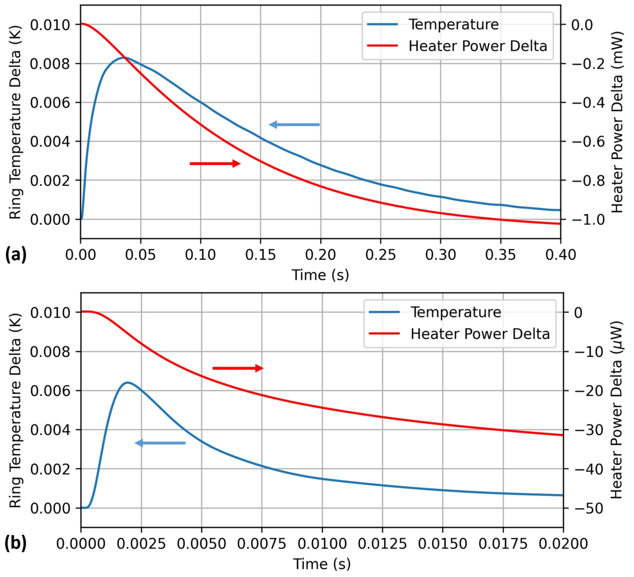

Fig. 12. The change in temperature (blue) and change in heater power (red) under closed-loop control in response to a (a) 500-W step in ASIC power and (b) 1-mW step in neighboring ring power versus time.

well within bounds. Note that while this temperature excursion is about the same magnitude as the 500-W ASIC step excursion, the timescale is much shorter. These results indicate that even within the aggressive thermal environment that 2.5D integrated optics will endure, it is feasible to develop controllers that lock the rings and stabilize to external temperature variations.

V. Optical Link Analysis

In this section, we explore ring-based DWDM link architectures for 200-Tb/s switches, packaged as in Fig. 4(b) and having 100 mm of perimeter photonic I/O. This generates a target of 800 Gb/s per fiber per direction, assuming one supply fiber for each TX/RX pair with 127-μm fiber pitch. The photonic link architecture design space is immense. For this study, we restrict the analysis to NRZ format using a single polarization with clock forwarding on a dedicated wavelength [42], [54]. (We assume each clock lane can service up to 16 data lanes.) We use 8b/10b data encoding (25% bandwidth overhead) and lightweight forward error correction, requiring a raw bit error rate (BER) $< 10^{-8}$. We then focus the analysis on the exploration of the optimal number, spacing, and data rate of DWDM channels which provide either the most open eye for a given energy, or the lowest energy for a given eye opening. This optimization is performed for various properties of the electrical and optical devices and circuits, such as modulation efficiency, bandwidth, equalization, demultiplexer filter order, etc. For each case, we optimize the rings for the best tradeoff between channel bandwidth and inter-channel crosstalk.

A. Statistical Link Tool

For the exploration, we use an in-house statistical DWDM link tool that provides very fast and reasonably accurate estimations of multi-channel link performance and energy efficiency. The tool (1) calculates the step-response transients of each of the components in the DWDM link, (2) convolves the responses encountered by a signal along any given path, (3) builds a statistical eye diagram for a user-defined pattern probability density function (PDF), (4) adds relevant random noise parameters, and (5) plots the statistical eye diagram. From the statistical eye, the user can monitor the vertical and horizontal eye opening at a given BER and repeat the measurement across various link-, circuit-, and device-level settings. To compare energy between different architectures at the same eye quality, the laser power can be scaled up or down to match the eye-opening target.

For the purpose of this study, we consider only the following random noise sources: output-referred noise at each TX, output-referred noise at each RX, and the laser relative intensity noise (RIN). Each are treated as white Gaussian noise, independent of any signal and of each other. Furthermore, we do not calculate sensitivity as conventionally defined at the input of the RX, but instead observe the statistics of the eye at the output of the TIA/limiting amplifier. We require that this opening, after all ISI, crosstalk, and noise, is larger than the output referred sensitivity defined at the target BER. This sensitivity is assumed to depend linearly on the data rate in our model.

B. Architecture Exploration

With the scope of the analysis and the tool described above, we begin an architectural exploration to determine the feasibility of the 800-Gb/s/fiber target using DWDM with access to EIC and PIC technologies available today or in the near future. The most critical decision is the number of wavelength channels and data rate per channel, such that the total throughput remains constant. This choice is affected by many device- and circuit-level assumptions such as the process node(s), photonic device characteristics, laser wall plug efficiency (WPE), ring tuning efficiency and range, etc. We estimate the energy efficiency of the entire macro required to transfer 800-Gb/s payload over a single fiber while meeting sensitivity plus margin, as well as the worst-case eye height and eye width at a fixed laser output power of 3 dBm per line (chosen to ensure open eyes for most links studied). We vary the raw data rate per channel simultaneously with the number of channels so that total throughput stays the same, including the coding and clocking overheads. This choice explores rich tradeoffs between the bandwidth of a single channel and crosstalk, which are of primary concern in DWDM links designed for high capacity. For the same free spectral range (FSR), links with a larger number of slower channels will suffer fewer bandwidth limitations, both electrical and optical, but more inter-channel crosstalk due to spectral proximity of neighboring channels and poor ring selectivity. As a result, they will operate at narrower spectral bandwidth, or equivalently higher Q, compared to the links with fewer faster channels. Each link architecture will therefore need to optimize ring Q, which will also depend on many low-level device and circuit parameters. We explore the effect of some of the most important of these parameters, as follows:

Circuits: We assume the ring driver and the TIA consist of three identical single-pole stages for simplicity. Although this choice excludes more complex circuits with peaking stages, complex poles, and analog equalization, it still allows us to explore the effects of circuit bandwidth and EIC process node.

Equalization: Equalization is studied by enabling a one-tap decision feedback equalizer (DFE), which we use to provide insight into how DFE and equalization in general affect the link architecture and shift the balance between the bandwidth and crosstalk by allowing high data rates that would have otherwise been unattainable. The DFE coefficient is adjusted for each data rate to maximize the eye opening for the given BER.

Ring modulator: The effect of the ring modulator is explored by varying its modulation efficiency, which is the wavelength shift per applied voltage. Different types of ring modulators (e.g., lateral or vertical junction, interdigitated, etc. [55], [56]) have different modulation efficiencies, but these also depend on doping characteristics. Normally, higher doping gives larger modulation efficiency, but also higher loss in the ring waveguide, which limits the achievable Q. For this reason, as we vary modulation efficiency between 22 pm/V (∼3.8 GHz/V), typical for lateral junctions with moderate doping concentrations, to 50 pm/V (∼8.7 GHz/V), typical for vertical junctions with higher doping concentrations (both defined for 0 V to –2 V levels), we simultaneously vary the ring waveguide loss between 60 dB/cm and 80 dB/cm, respectively.

Ring filter: Use of a higher order RX ring filter affects the bandwidth-crosstalk tradeoff by suppressing the filter stop-band more and therefore making the entire link less sensitive to crosstalk. In addition, flatter passband characteristics can be engineered to reduce ISI. We study the effect of the RX ring order by comparing the links with conventional single rings to those that use double rings designed to fit the maximally flat Butterworth characteristics. We ignore the effects of potentially more complicated thermal tuning of such structures. Other link parameters are fixed at reasonable values for the purpose of containing the scope of the analysis. A few of these are listed here. The laser provides N equidistant lines with equal power, WPE of 8%, and RIN of -145 dBc/Hz. All rings have a radius of 5 μm and a thermal tuning efficiency of 200 nm/W (∼35 GHz/mW). Loss between rings is 0.1 dB per ring, which includes the losses of the off-resonance ring coupler and the inter-ring waveguide segment. All optical components apart from the laser, the rings, the waveguide between the rings, and the photodiode are modeled as a lumped broadband loss of 10 dB. Ring spatial arrangement on the TX and RX buses is such that the k-th ring modulator on the TX bus (counting from the laser) is locked to the same laser channel as the k-th ring on the RX bus (counting from the input coupler). TX driver power is CV^2f-dominated, while TIA power is constant with data rate. Finally, we assume that the power consumption of the electrical digital back-end depends mostly on the total throughput and not on the data rate of a single channel.

The study is organized as a comparison between a baseline choice of parameters and variations to that baseline. Those baseline parameters are: non-equalized electrical circuits with bandwidths of 40 GHz per stage, modulation efficiency of 22 pm/V with 60 dB/cm ring loss, and single-ring filters. Each of these parameter settings is studied for the set of {data rate, channel count} pairs. The goal of this approach is to identify the parameters that have the most impact on the link performance, which can serve as input to the architecture, circuit design, or EIC/PIC foundry.

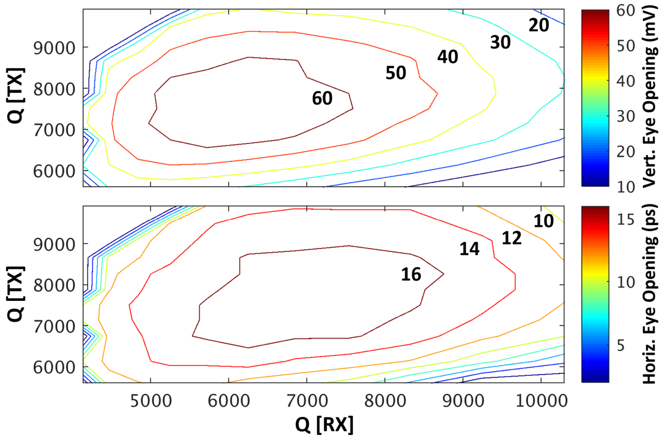

Fig. 13. Vertical (top) and horizontal (bottom) eye opening at a BER of 10^{-8} as a function of the TX and RX ring Qs for the baseline design with 34 32-Gb/s channels and a laser output power of 3 dBm.

The TX and RX ring optimizations for each architecture are performed by varying the coupling coefficients, while maintaining critical coupling, and minimizing the drop-port loss (effectively forcing drop and bus coupling coefficients of the RX ring to be equal). There is a 1:1 correspondence between the optimal coupling coefficients and the TX and RX ring Qs, which are plotted in Fig. 13 versus the eye height and eye width at a fixed laser output power of 3 dBm for a single architecture – 34 channels, 32 Gb/s/channel with a fixed set of parameters that correspond to our baseline configuration. Fig. 14 shows the statistical eye diagrams and eye contours of two illustrative channels (#1 and #34 for the same architecture with the Qs that maximize the eye opening from Fig. 13). We see that channel #1 (the channel locked to the first RX ring) experiences the worst-case crosstalk. Note that the first RX ring experiences all the aggressor crosstalk channels, on both the blue and the red sides, since the laser spectrum is spread equidistantly over the entire FSR. The crosstalk to all other rings is attenuated due to the filtering action of the RX rings that remove most of the signal from all potential aggressor channels that are spatially preceding the main channel. The very last ring sees almost no crosstalk because all potential aggressor channels have already been removed from the bus.

Figs. 15 and 16 summarize the performance of various link architectures, where we vary the channel count and data rate simultaneously while keeping the total throughput at 800 Gb/s. For both figures, we compare the baseline with the following cases: (a) 1-tap DFE where the coefficients are set for each data rate to maximize the eye opening, (b) modulator ring with 50 pm/V modulation efficiency, and (c) RX with a double ring. Fig. 15 reports vertical and horizontal eye openings at fixed laser power, while Fig. 16 shows energy efficiency (by scaling laser power to achieve the eye opening needed to produce the raw BER of 10^{-8} at the given data rate) for two thermal tuning scenarios.

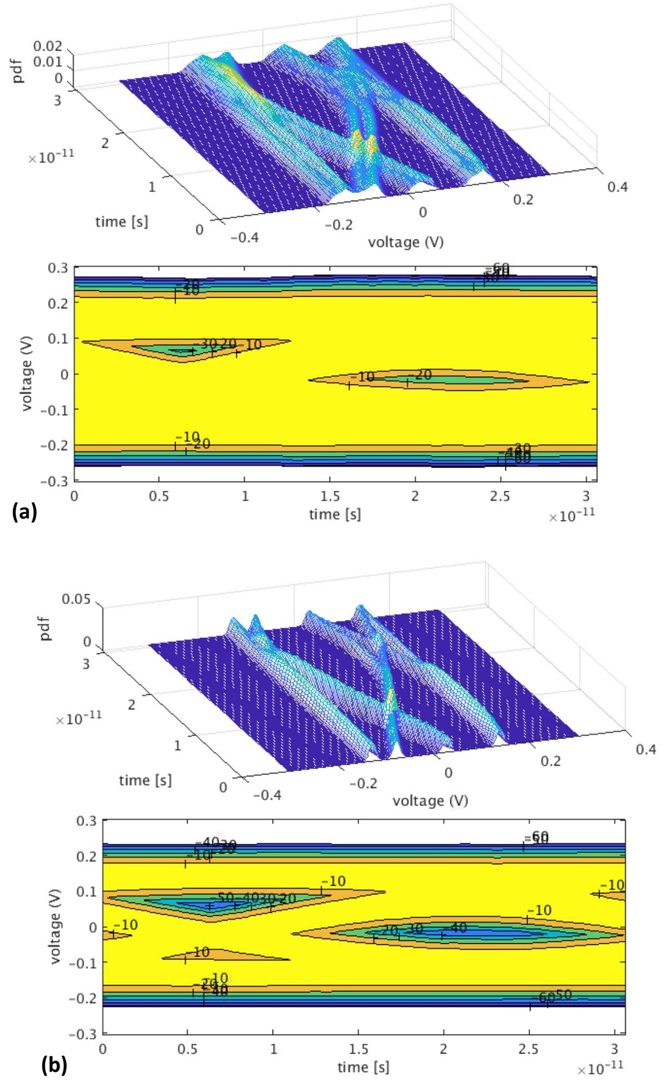

Fig. 14. Statistical eye diagrams (top) and eye contours (bottom) of two channels in the baseline architecture with 34 32-Gb/s channels: (a) channel #1 (closest to the RX input coupler), (b) channel #34 (last channel on the bus).

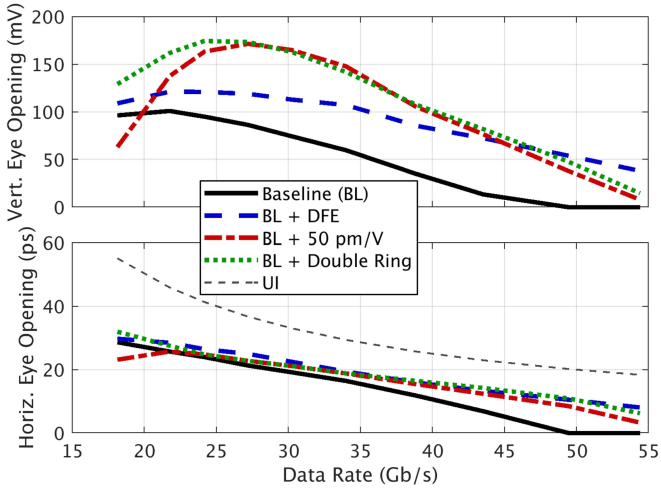

Fig. 15. Vertical (top) and horizontal (bottom) eye opening at a BER of 10^{-8} with a laser output power of 3 dBm versus raw data rate for the set of design parameters at 800-Gb/s total throughput.

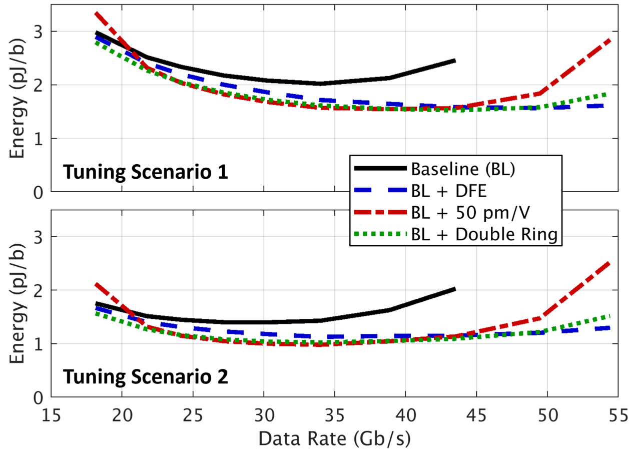

Fig. 16. Energy per bit for wide (top) and narrow (bottom) ring heater tuning range versus raw data rate for the set of design parameters at 800-Gb/s total throughput.

In the first scenario, corresponding to the case where process variations are large, the TX ring modulators and RX ring filter must be tuned by FSR/5 and FSR/8, respectively. In the second low-process-variation scenario, the rings only need to lock to their nearest laser line; thus, the required tuning range is FSR/*N* where *N* is the number of channels.

From Fig. 15, for data rates lower than what the baseline design can comfortably support, there is no benefit from lowering data rate further due to the crosstalk from the decreasing channel spacing. This is most pronounced for the link with the high modulation efficiency ring modulator as its high ring loss imposes higher penalty for the high-Q modulators needed in this regime. At the other end of the spectrum, the bandwidth enhancing DFE configuration tends to scale the best with the data rate. The baseline configuration, and to a lesser degree the other configurations as well, cannot support very high data rates. The optimum eye height at the fixed laser power is found at about 25-30 Gb/s/channel.

Similar trends can be seen in Fig. 16. The comparison between the two tuning scenarios shows that the narrower tuning range results in better energy efficiency, as expected. However, it also shows that with our assumptions, the ring heaters consume a sizable portion of the total link energy. The implication is that if the untuned ring resonant wavelengths have large variation, other architecture and circuit choices will matter less, and the architectures with few faster channels will be more beneficial simply because they use fewer rings. As a result, the energy-optimized data rate is higher than the eye-opening optimized data rate. The amount of shift depends on what portion of the link energy is proportional to the number of rings: tuning scenario 1 shifts the energy efficiency optimal point towards higher data rates and fewer rings; tuning scenario 2 tends to favor comparatively lower data rates as it imposes less energy penalty for adding more channels to the link, given that the total heater power does not depend on data rate and number of channels. As an example, for 800 Gb/s/channel and the set of assumptions outlined above, the baseline design should target about 28-33 Gb/s/channel to achieve 1.5-2 pJ/b (not surprisingly, limited by the raw bandwidth of the electrical front-end), while any of the device and circuit techniques we study would improve the energy efficiency to 1-1.5 pJ/b at data rates in the range of 30-45 Gb/s/channel.

C. Practical Considerations and Limitations

Our analysis indicates that 800 Gb/s/fiber is feasible with the technology available today and with reasonable energy efficiency of <2 pJ/b. The choice of the architecture depends on many assumptions and design choices regarding process technology, electrical circuits, and photonic devices, some of which were included in our study. There are many other factors that we have not considered here that may incur additional power penalty. One of these is the impact of non-ideal laser sources, including non-uniform channel spacing. Another is the implication of the architecture to the optical packaging and assembly. Yet another is the effect of the detailed thermal control implementation.

Some other limitations of our analysis are:

-To speed up the simulation, our approach calculates contributions only from a user-defined number of spectral neighbors to each channel. Throughout the simulations we used four spectral neighbors to each side of the main channel. This parameter allows for an indication of how many neighbors are important (more with narrower channel spacing). We found that including four neighbors accounts for virtually all crosstalk in practical cases.

-The statistical tool performs calculations in baseband, rather than at optical frequencies. The assumption is that the data stream at each carrier is independent of the others. As such, the nonlinear effects from inter-channel interactions such as cross-modulation at the photodiode are not included. These effects are generally negligible for practical choices of channel spacing and RX front-end bandwidth.

-Thermal tuning is assumed to be ideal; no power penalty is added due to imperfect tuning. Tuning penalty is generally small, although not negligible, and we expect it will not significantly change the results and conclusion.

-Effects of process variations are not considered (e.g., double rings are assumed identical, laser sources are assumed uniform in power and spacing, etc.). These effects will be a focus of a future study.

VI. Conclusion

Accelerated computing systems, whether on-premise or cloud-based, rely increasingly on high-bandwidth switched interconnects to scale performance through parallelism. For switch ASICs beyond the 100-Tb/s generation, CPO on MCM will face efficiency and density challenges at the electrical interface between the ASIC and the optics. In these systems, a silicon interconnect (either on interposer or a bridge layer) that joins the ASIC with the optical TRXs can overcome these challenges. Electrical interfaces and optical links—including drivers, tuning, and control—that operate at 0.25 pJ/b and 1 pJ/b, respectively, can deliver a total off-chip communications power of 300 W in the package for a 200-Tb/s switch. Once in the optical domain, reaches of 100 m to 1 km are achievable. Fiber interfaces to the 2.5D integrated optics will have to support multi-Tb/s/mm edge bandwidth densities. DWDM links employing micro-ring resonators, which can scale in bandwidth through wavelength-parallelism, may provide the required energy and area efficiencies. We have established the feasibility of operating the DWDM links in the aggressive thermal environment while maintaining adequate signal integrity at the required energy values. If realized, the densely integrated solution will not only help computing systems continue to scale up and out via switch scaling, it can also be replicated within GPUs, CPUs, and other processing ASICs to improve efficiency across the entire machine

References

[1] A. Singh et al., "Jupiter rising: A decade of clos topologies and centralized control in google's datacenter network," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 4, pp. 183–197, Oct. 2015.

[2] D. Chen et al., "The IBM blue gene/Q interconnection fabric," *IEEE Micro*, vol. 32, no. 1, pp. 32–43, Jan./Feb. 2012.

[3] B. Arimilli et al., "The PERCS high-performance interconnect," in *Proc. IEEE 18th Symp. High Perform. Interconnects*, 2010, pp. 75–82.

[4] S. Fathololoumi et al., "1.6 Tbps silicon photonics integrated circuit and 800 Gbps photonic engine for switch co-packaging demonstration," *J. Lightw. Technol.*, vol. 39, no. 4, pp. 1155–1161, Feb. 2021.

[5] S. Fathololoumi et al., "Highly integrated 4 Tbps silicon photonic IC for compute fabric connectivity," in *Proc. IEEE Symp. High-Perform. Interconnects*, 2022, pp. 1–4.

[6] R. Mahajan et al., "Co-packaged photonics for high performance computing: Status, challenges and opportunities," *J. Lightw. Technol.*, vol. 40, no. 2, pp. 379–392, Jan. 2022.

[7] Ranovus, "Ranovus demonstrates industry's first adaptive compute acceleration co-packaged optics platform with xilinx versal and ranovus odin$^{\text{TM}}$ 800 Gbps CPO 2.0," *Press Release*. Accessed: Mar. 3, 2022. [Online]. Available: https://ranovus.com/ranovus-with-amd-cpo/

[8] N. Margalit et al., "Perspective on the future of silicon photonics and electronics," *Appl. Phys. Lett.*, vol. 118, no. 22, 2021, Art. no. 220501.

[9] H. Tan, R. Velaga, and A. Bjorlin, "Networking overview," J.P. Morgan 19th Annual Tech/Auto Forum. Accessed: Jan. 12, 2021. [Online]. Available: https://investors.broadcom.com/static-files/c5414657-07e3-423f-afec-315135a9cb53

[10] R. Chopra, "Co-packaged optics and an open ecosystem," *Cisco Blogs*. Accessed: Jan. 11, 2021. [Online]. Available: https://blogs.cisco.com/sp/co-packaged-optics-and-an-open-ecosystem

[11] Marvell, "Marvell unveils co-packaged optics technology platform at OFC 2022," *Press Release*. Accessed: Mar. 7, 2022. [Online]. Available: https://www.marvell.com/company/newsroom/marvell-unveils-co-packaged-optics-technology-platform-at-ofc-2022.html

[12] B. G. Lee, "Driving down link energy and driving up link density in GPU networks," in *Proc. IEEE Opt. Fiber Commun. Conf.*, 2022, pp. 1–3.

[13] B. Dally, "Optical communication for data center GPUs," in *Proc. Opt. Fiber Commun. Conf.*, Mar. 2022, Paper Tu2A.1.

[14] M. Wade et al., "TeraPHY: A chiplet technology for low-power, high-bandwidth in-package optical I/O," *IEEE Micro*, vol. 40, no. 2, pp. 63–71, Mar./Apr. 2020.

[15] K. Takemura et al., "Silicon-photonics-embedded interposers as co-packaged optics platform," *Trans. Jpn. Inst. Electron. Packag.*, vol. 15, pp. E21-012-1–E21-012-13, 2022.

[16] N. C. Harris, D. Bunandar, A. Joshi, A. Basumallik, and R. Turner, "Passage: A wafer-scale, programmable photonic communication substrate," in *Proc. IEEE Hot Chips 34 Symp.*, 2022, pp. 1–26, doi: 10.1109/HCS55958.2022.9895610.

[17] "TOP500 | June 2022 list." Accessed: Jun. 14, 2022. [Online]. Available: https://top500.org/

[18] G. B. Berriman et al., "The application of cloud computing to scientific workflows: A study of cost and performance," *Philos. Trans. Roy. Soc. A: Math., Phys. Eng. Sci.*, vol. 371, no. 1983, 2013, Art. no. 20120066.

[19] NVIDIA, "NVIDIA DGX systems: Purpose built for the unique demands of AI," Accessed: Jun. 14, 2022. [Online]. Available: https://www.nvidia.com/en-us/data-center/dgx-systems/

[20] C. Boyle, “Meta works with NVIDIA to build massive AI research supercomputer,” *NVIDIA Blogs*. Accessed: Jan. 24, 2022. [Online]. Available: https://blogs.nvidia.com/blog/2022/01/24/meta-ai-supercomputer-dgx/

[21] C. B. Stunkel et al., “The high-speed networks of the Summit and Sierra supercomputers,” *IBM J. Res. Develop.*, vol. 64, no. 3/4, pp. 3:1–3:10, May–Jul. 2020.

[22] NVIDIA, “NVIDIA H100 tensor core GPU architecture: Exceptional performance, scalability, and security for the data center,” 2022. [Online]. Available: https://resources.nvidia.com/en-us-tensor-core

[23] T. P. Morgan, “Mellanox doubles up ethernet bandwidth with Spectrum-3,” *The Next Platform*, Accessed: Mar. 26, 2020. [Online]. Available: https://www.nextplatform.com/2020/03/26/mellanox-doubles-up-ethernet-bandwidth-with-spectrum-3/

[24] NVIDIA, “NVIDIA announces spectrum high-performance data center networking infrastructure platform,” *Press Release*. Accessed: Mar. 22, 2022. [Online]. Available: https://nvidianews.nvidia.com/news/nvidia-announces-spectrum-high-performance-data-center-networking-infrastructure-platform

[25] NVIDIA, “NVIDIA Mellanox Quantum HDR 200G infiniband switch silicon,” *NVIDIA Mellanox Product Brief*, Sep. 2020. [Online]. Available: https://www.mellanox.com/sites/default/files/doc-2020/pb-quantum-hdr-switch-silicon.pdf

[26] NVIDIA, “NVIDIA quantum-2 takes supercomputing to new heights, into the cloud,” *Press Release.* Accessed: Nov. 9, 2021. [Online]. Available: https://nvidianews.nvidia.com/news/nvidia-quantum-2-takes-supercomputing-to-new-heights-into-the-cloud

[27] P. Teich, “Building bigger, faster GPU clusters using NVSwitches,” *The Next Platform*. Accessed: Apr. 13, 2018. [Online]. Available: https://www.nextplatform.com/2018/04/13/building-bigger-faster-gpu-clusters-using-nvswitches/

[28] R. Smith, “NVIDIA volta unveiled: GV100 GPU and tesla V100 accelerator announced,” *AnandTech*. Accessed: May 10, 2017. [Online]. Available: https://www.anandtech.com/show/11367/nvidia-volta-unveiled-gv100-gpu-and-tesla-v100-accelerator-announced

[29] J. Choquette, E. Lee, R. Krashinsky, V. Balan, and B. Khailany, “The A100 datacenter GPU and Ampere architecture,” in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2021, pp. 48–50.

[30] L. Gwennap, “NVIDIA Hopper leaps ahead: New 700W AI chip triples performance, adds FP8 support,” *The Linley Group, Mountain View, CA, USA, Microprocessor Report*, Apr. 11, 2022, pp. 1–6.

[31] Optical Internetworking Forum, “Common electrical I/O (CEI)-112G,” Accessed: May 3, 2022. [Online]. Available: https://www.oiforum.com/technical-work/hot-topics/common-electrical-interface-cei-112g-2/

[32] P. Mishra et al., “A 112 Gb/s ADC-DSP-based PAM-4 transceiver for long-reach applications with >40 dB channel loss in 7 nm FinFET,” in *Proc. IEEE Int. Solid- State Circuits Conf. (ISSCC)*, 2021, pp. 138–140.

[33] Z. Guo et al., “A 112.5 Gb/s ADC-DSP-based PAM-4 long-reach transceiver with >50 dB channel loss in 5 nm FinFET,” in *Proc. IEEE Int. Solid- State Circuits Conf. (ISSCC)*, 2022, pp. 116–118.

[34] A. Varzaghani et al., “A 1-to-112 Gb/s DSP-based wireline transceiver with a flexible clocking scheme in 5 nm FinFET,” in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, 2022, pp. 26–27.

[35] R. Shivnaraine et al., “A 26.5625-to-106.25 Gb/s XSR SerDes with 1.55 pJ/b efficiency in 7 nm CMOS,” in *Proc. IEEE Int. Solid- State Circuits Conf.*, 2021, pp. 181–183.

[36] G. Gangasani et al., “A 1.6 Tb/s chiplet over XSR-MCM channels using 113 Gb/s PAM-4 transceiver with dynamic receiver-driven adaptation of TX-FFE and programmable roaming taps in 5 nm CMOS,” in *Proc. IEEE Int. Solid- State Circuits Conf.*, 2022, pp. 122–124.

[37] C. F. Poon et al., “A 1.24-pJ/b 112-Gb/s (870 Gb/s/mm) transceiver for in-package links in 7-nm FinFET,” *IEEE J. Solid-State Circuits*, vol. 57, no. 4, pp. 1199–1210, Apr. 2022.

[38] R. Yousry et al., “A 1.7 pJ/b 112 Gb/s XSR transceiver for intra-package communication in 7 nm FinFET technology,” in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2021, pp. 180–182.

[39] R. Mahajan et al., “Embedded multi-die interconnect bridge (EMIB)—A high density, high bandwidth packaging interconnect,” in *Proc. IEEE 66th Electron. Compon. Technol. Conf.*, 2016, pp. 557–565.

[40] P. K. Huang et al., “Wafer level system integration of the fifth generation CoWoS-S with high performance Si interposer at 2500 mm^2,” in *Proc. IEEE 71st Electron. Compon. Technol. Conf.*, 2021, pp. 101–104.

[41] K. Sikka et al., “Direct bonded heterogeneous integration (DBHi) Si bridge,” in *Proc. IEEE 71st Electron. Compon. Technol. Conf.*, 2021, pp. 136–147.

[42] Y. Nishi et al., “A 0.297-pJ/bit 50.4-Gb/s/wire inverter-based short-reach simultaneous bidirectional transceiver for die-to-die interface in 5 nm CMOS,” in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, 2022, pp. 154–155.

[43] M. Wade et al., “An error-free 1 Tbps WDM optical I/O chiplet and multi-wavelength multi-port laser,” in *Proc. Opt. Fiber Commun. Conf.*, 2021, pp. 1–3.

[44] B. B. Buckley et al., “WDM source based on high-power, efficient 1280-nm DFB lasers for terabit interconnect technologies,” *IEEE Photon. Technol. Lett.*, vol. 30, no. 22, pp. 1929–1932, Nov. 2018.

[45] T. Thiessen et al., “Back-side-on-BOX heterogeneously integrated III-V-on-silicon O-band distributed feedback lasers,” *J. Lightw. Technol.*, vol. 38, no. 11, pp. 3000–3006, Jun. 2020.

[46] J. B. Driscoll et al., “First 400G 8-channel CWDM silicon photonic integrated transmitter,” in *Proc. IEEE 15th Int. Conf. Group IV Photon.*, 2018, pp. 1–2.

[47] A. W. Fang, G. Fish, and E. Hall, “Heterogeneous photonic integrated circuits,” in *Proc. IEEE Photon. Conf.*, 2012, pp. 354–355.

[48] G. Kurczveil et al., “Robust hybrid quantum dot laser for integrated silicon photonics,” *Opt. Exp.*, vol. 24, no. 14, pp. 16167–16174, 2016.

[49] D. Miller, “Getting to femtojoule optics–what physics and what technology?,” in *Proc. IEEE Opt. Fiber Commun. Conf.*, 2021, pp. 1–2.

[50] X. Zheng and A. V. Krishnamoorthy, “Si photonics technology for future optical interconnection,” in *Proc. IEEE Asia Commun. Photon. Conf.*, 2011, pp. 1–11.

[51] M. Sakib et al., “A 240 Gb/s PAM4 silicon micro-ring optical modulator,” in *Proc. IEEE Opt. Fiber Commun. Conf.*, 2022, pp. 1–3.

[52] D. Coenen et al., “Thermal modelling of silicon photonic ring modulator with substrate undercut,” *J. Lightw. Technol.*, vol. 40, no. 13, pp. 4357–4363, Jul. 2022.

[53] N. S. Nise, *Control Systems Engineering*, 8th ed. Hoboken, NJ, USA: Wiley, 2019, pp. 358–419.

[54] J. W. Poulton et al., “A 0.54 pJ/b 20 Gb/s ground-referenced single-ended short-reach serial link in 28 nm CMOS for advanced packaging applications,” *IEEE J. Solid-State Circuits*, vol. 48, no. 12, pp. 3206–3218, Dec. 2013.

[55] M. Pantouvaki et al., “Comparison of silicon ring modulators with interdigitated and lateral p-n junctions,” *IEEE J. Sel. Topics Quantum Electron.*, vol. 19, no. 2, pp. 7900308–7900308, Mar./Apr. 2013.

[56] E. Timurdogan et al., “An ultralow power athermal silicon modulator,” *Nature Commun.*, vol. 5, 2014, Art. no. 4008.