Challenges and opportunities in system-level evaluation of photonics

George Michelogiannakis, Jeremiah Wilke, Min Yee Teh, Madeleine Glick, John Shalf, et al.

George Michelogiannakis, Jeremiah Wilke, Min Yee Teh, Madeleine Glick, John Shalf, Keren Bergman, "Challenges and opportunities in system-level evaluation of photonics," Proc. SPIE 10946, Metro and Data Center Optical Networks and Short-Reach Links II, 1094607 (1 February 2019); doi: 10.1117/12.2510443

SPIE.

Event: SPIE OPTO, 2019, San Francisco, California, United States

Invited Paper

Challenges and Opportunities in System-Level Evaluation of Photonics

George Michelogiannakis[a], Jeremiah Wilke[b], Min Yee Teh[c], Madeleine Glick[c], John Shalf[a], and Keren Bergman[c]

[a]Lawrence Berkeley National Laboratory
[b]Sandia National Laboratory
[c]Columbia University

ABSTRACT

The cost and complexity of future interconnects create a significant opportunity for emerging photonic technologies such as fibers and switches. These technologies should be evaluated at the system level in order to determine the most efficient way they can be used, as well as provide feedback to photonic developers to better optimize for high-level impact. In this paper, we argue for the need for a systematic methodology to extract system-level models for any emerging photonic component. We then outline our past experience with extracting architectural-level metrics from device demonstrations and conducting architectural-level evaluations. Finally, we discuss qualities for a desirable solution to this problem that requires cross-community collaboration.

1. INTRODUCTION

Large-scale high performance computing (HPC) systems and warehouse scale data centers are increasingly pervasive in different areas of modern life, with applications from climate modeling to consumer analytics. These large investments create a significant drive to preserve performance scaling for these systems but with fixed procurement and power costs. Many such applications are data intensive and thus place significant stress on the interconnection network. This is expected to intensify. For instance, compared to 2010, today's top 10 HPC systems have 65× higher compute throughput in terms of FLOPS per second but only 4.8× the communication bandwidth, thus 0.08× the memory bandwidth per FLOP.[1] In addition, specialization increases computational throughput and consequently stresses communication, which may prove to be the bottleneck. As an example, Google's tensor processing unit (TPU) has a compute capacity of 90 TOPS/s (with 8 bit precision) and a peak off-board communication bandwidth of 34 GB/s.[2] This means that a byte arriving to the board must be re-used 2706 times in order to saturate computation before communication.[1] Most applications in[2] did not have the computational intensity to achieve this. Future deployment of aggressive accelerators in large-scale systems will create similar challenges and make it likely that many applications will saturate communication before computation, at an era where networks will no longer be cheap enough to overdesign in order to mitigate this problem.

Multiple recent advancements in photonics provide promising solutions for future communication challenges and can break through the energy bandwidth limitations of conventional electronic interconnects. At a high level, photonics provide high transmission capacity through optical fibers and can provide low-energy switching. In particular, optical 100 Gb/s fibers used as network channels are experiencing rapid adoption[3] while 200 Gb/s and 400 Gb/s optical link demonstrations are currently in focus.[4] Optical switches have also made significant advancements whether they are based on micro-electro-mechanical systems (MEMS) technology,[5] array waveguide grating routers (AWGRs),[6] or ring oscillators.[7] For both channels and switches, photonics can require less than 1 pJ/bit independent of distance, up to a maximum. These are particularly attractive qualities for future networks.

However, making optimal use of future photonic technologies requires re-thinking the entire system architecture including the software stack. This is in part due to the reconfiguration delay of optical switches and the

George Michelogiannakis E-mail: mihelog@lbl.gov

Metro and Data Center Optical Networks and Short-Reach Links II, edited by Atul K. Srivastava, Madeleine Glick, Youichi Akasaka, Proc. of SPIE Vol. 10946, 1094607 · © 2019 SPIE
CCC code: 0277-786X/19/$18 · doi: 10.1117/12.2510443

inability to perform computation in the optical domain, which push designers towards reconfigurable circuit-switched flow control.[8] In addition, we are at a point where simply replacing current network components with emerging photonics cannot provide the brave advancements necessary for future systems. That is because recent datacenter-scale networks consume only 4% to 11% of overall system power.[9] Even though these numbers may grow in the future, even an unrealistic zero-energy network cannot increase system energy efficiency by 2× if the network consumes half of overall power, which is too high for even far-reaching projections.

Therefore, we need to change system architecture to better make use of emerging photonics and thus increase system-wide improvements. However, doing so requires accurate and scalable high-level system evaluations which in turn require accurate system-level models of emerging photonic technologies. Otherwise, evaluating optical fibers or switches in isolation does not provide accurate information for the system-wide impact of each technology. This makes it challenging for system architects to adjust their designs to better make use of photonics. In addition, system-level evaluations are also valuable to photonics developers because simulations give developers an understanding of how to best change their device for best system-level impact. For example, the tradeoff between error rate and energy is better evaluated at the system scale where the impact of increased errors is better quantified, and in fact can be significant.[10] This can drive photonics designers to prioritize reducing errors which are not obvious at the device scale.

Numerous system or architecture simulators exist in literature such as SST[11] and Gem5,[12] which are two widely used examples. These simulators are capable of executing application code and include behavioral descriptions of important components from the system network, on-chip network, processors, accelerators (including GPUs), memories, caches, and others. They report high-level metrics such as application execution time, network load, memory usage, and other relevant information to system architects. In addition, they can be paired with energy reporting tools that understand the underlying technology library and hardware implementation to convert usage statistics to power consumption and area.[13,14]

Past work has addressed this problem but does not provide a complete solution even though one is necessary for the future. Notably, PhoenixSim[15] provides a tool to connect the physical layer with optic devices to the network and then the application layer, such that figures of merit from any of the three layers can be extracted for photonics devices. This is an example of a photonic design tool and is significant step forward and demonstrates the need for such tools. However, it is not meant to perform detailed and cycle-accurate simulators or to design circuits and networks. Other past work developed physical-level descriptions of photonic devices in Verilog-A[16] or other SPICE formats.[17] However, these works focus on on-chip photonics and do not provide complete coverage of emerging technologies.

In this paper, we argue for the need for a systematic methodology to extract system-level models for any emerging photonic technology, whether that is a channel, switch, off chip, or on chip. In particular, we begin by outlining our past experience with extracting architectural-level metrics from device demonstrations and conducting architectural-level simulations. Then we discuss challenges of the current state of the art towards wider adoption of photonics, and then discuss qualities for a solution to this problem. Finally, we discuss desired functionality, pitfalls, and challenges. This is an endeavor that requires cross-community collaboration to produce lasting results.

2. EXAMPLE TESTCASE USING DEVICE DEMONSTRATIONS

In this section, we describe one recent example of how to extract high-level performance and energy numbers such as energy per bit from photonic device demonstrations. The optical link, in our case a transceiver plus laser source, has power consumption contributions from the laser source, the modulator driver and the receiver TIA. In addition, if we are using micro ring based transceivers we must include wavelength stabilization and wavelength locking of the rings in order to compensate for fabrication tolerances and environmental changes. Note that in this case we are not considering SERDES, or "gear" changes needed if there is different data rate transmission between the electronics and the optics.

The laser optical power should be sufficient to close the required power budget, which is a function of the receiver sensitivity, bit rate and optical losses across the link. A margin may be added (often 3dB). For the total power consumption one must take into account a factor for the wall plug efficiency of the laser. One then

multiplies by the number of lasers. Based on our previous link design and analysis (100Gb/s link with receiver sensitivity of -17dBm), we consider a wall plug efficiency of 10%, with a laser power of 1mW. These numbers depend on the specific components chosen and critically on the laser chosen and its actual wall plug efficiency. A discussion of the optical link trade offs is presented in.[18]

The microring power consumption has 3 components: 1) tuning to the desired wavelength, 2) locking on the desired wavelength, 3) active power for driving modulation or switching. The power consumed for tuning to the desired wavelength depends on fabrication variations. In some cases, this may not be required at all. However, in the typical case some tuning is required. The resonance of a typical silicon microring resonator is shifted by 10 Ghz (0.07 nm) for each degree Kelvin change in the temperature of the microring resonator. This shift is utilized to tune the ring to the desired wavelength and also in thermo-optic modulation and switching.[19] Therefore, in this example we estimate that for a microring resonator with a typical free spectral range (FSR) of approximately 25nm, tuning the ring to half FSR, the worst case, consumes roughly 12.5 mW. For the optical switch, in which the light passes through two rings we estimate a power consumption of 25 mW per pass. Once the ring is tuned to the correct wavelength it has to be maintained or locked at that wavelength consuming approximately 385 μW.[20] The power consumption of the link also includes the modulator drivers and the detector TIA for each ring, which we estimate to be 60 mW and 50 mW respectively.

The result of this example is an optical link model of a power consumption of 2 pJ / bit with different wavelengths at 10 Gpbs each. There are several avenues to reduce the power consumption which, for example, include developing a more efficient laser and reducing insertion losses.

3. NETWORK PLANE SIMULATION CHALLENGES

Here, we emphasize some system-level challenges that have complicated may yet further complicate mass deployment of photonics within the context of computing. Arguably the prime challenge faced in the context of networking is that photonically-interconnected networks require us to challenge the solid fundamentals on which much of today's networking infrastructure has been built. Given the reconfigurable nature of photonically-connected networks, there needs to be a tight-coupling between the routing components and topology components of the system. Up till today however, the networking community has tended to implement compute clusters with decentralized routing to prevent single-point failures. Hinging on the success of Google's Jupiter fabrics,[21] which implements a centralized routing scheme, we argue that there should be a shift of focus towards centralized-routing within the HPC and cloud communities.

To date, there is limited support for simulating all components of a photonically-connected system at the architectural level, and even more limited in the open source domain. This is due to available network simulators such as the ones mentioned in the introduction that generally consider a static physical network layer. This is an assumption fundamentally at odds with photonic networks that derive their advantages from their reconfigurability. The established convention for system-level workflow can be broadly binned into two phases: the first phase is the systems architecture setup phase in which the general machine parameters are set up, and the second phase being the actual simulations phase. The general workflow of simulators is to set up the data structures that represent the physical network layer at setup time. This is shown in the left of Figure 1.

The left side of Figure 1 shows a more photonic-friendly approach by separating the physical topology component from the logical topology, with the physical topology being configured at setup time but the logical topology being configured at run time. In other words, current simulators not only are challenged in deriving energy models, but should also provide a more configurable view of the network that better matches the strengths of photonic interconnects. Solving this challenge will provide the first steps to produce simulations that more accurately reflect the physical reality.

4. NETWORK ENDPOINT AND SOFTWARE STACK SIMULATION CHALLENGES

Also challenging for simulation is how changes to the physical architecture propagate up the network stack. For a fixed application, the network stack details can greatly impact the traffic pattern or traffic model injected into a simulated system architecture. For example, consider the message passing interface (MPI) which is the dominant network middleware for HPC. The popular OpenMPI implementation provides numerous transport

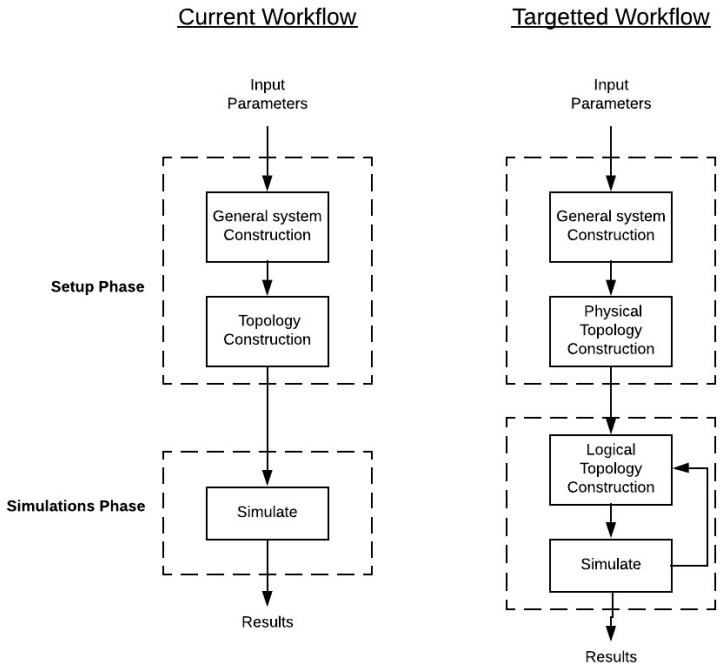

Figure 1. General workflow of network simulators today is presented on the left, with the right figure being a more photonic networks-friendly workflow.

layers for message matching (MTL) or raw data movement (BTL) ranging from ethernet (TCP/IP) to RDMA (e.g. libfabrics).[22] On top of this, each transport layer may have different predictive or reservation protocols for exchanging data in the point-to-point message layer (PML). Even the same MPI application can therefore have quite distinct traffic characteristics depending on the low-level network "provider".

Within the OSI[23] or TCP/IP stack, an example would be shifting from packet switching to circuit switching. For TCP/IP to handle photonic network components it would need to differentiate between packets travelling through electrical switches that therefore may face congestion, and packets traveling through optical switches that require advance reservations made in a circuit-switched manner. In that case, traffic through the optical switches would be injected as TCP "optical packets", only replacing the underlying network and physical layers. Bandwidth steering with photonics has generally preserved the network interfaces used with electrical packet switching. For example, wavelength-multiplexing within ethernet simply provides extra bandwidth density and the endpoint model (injected traffic) is unchanged. However, entire flows can be delivered over an optical circuit without packetization. Therefore, for flows traversing through optical circuits rely on resource scheduling instead of congestion control. Software-defined networks (SDN),[24] source-based routing,[25] congestion control,[26] software error correction,[27] packetization, and source or destination message ordering[28] are some of the implementation details handled in the endpoint software stack which could greatly impact traffic the network experience and therefore network requirements because they change the times that data packets are admitted into the network as well as the control packets necessary.

Simulation is perhaps most useful as a co-design tool, allowing both the physical architecture and supporting software to be simultaneously developed. This model, successful in embedded systems,[29] should ideally be extended to the system-level challenges discussed here. As photonic devices are incorporated, how can the software itself adapt to best use the hardware? System-level simulation tools vary in the level of detail given to the different layers of the network stack. SST,[30] SMPI,[31] and CODES[32] provide their own simulator-specific MPI implementations (and therefor simulator specific backend "providers"). SST and CODES have packet-level routing and contention modeled, but some details of the physical or data link layers are approximated for

Table 1. An example collection of some metrics at the device level and at the architectural level. The challenge is to systematically and rapidly translate between the two domains for a variety of existing and emerging photonic components to make architectural design space exploration more efficient.

Device level	Architectural level
Optical (db) loss	Error rate per bit
Crosstalk	Bandwidth density
Wavelengths (comb diagram)	Bandwidth and number of channels
Eye diagram	Energy per bit
Conversion energy and latency	Energy and latency per bit
Mechanically directing laser beams	MEMS switch reconfiguration delay

simulation efficiency. Booksim generally provides more detail for the network components, but is generally used in conjunction with tools like SST/macro to provide the actual traffic injection.[33] The LogP simulator has been used to explore numerous software-level network strategies, but estimates only flow-level network delays without any packet modeling.[34] PhoenixSim[15] provides numerous physical models, but relies on simulation-specific traffic generators via its LWSim component. Different simulation tools are generally suited to particular problems, but none provides a "holistic" simulation framework for tuning all levels of the network stack together. A major simulation challenge going forward is therefore solving the Venn diagram of network simulators for true co-design of network hardware and network stack.

Photonics may naturally fit into existing network interfaces, but also has the potential to be highly disruptive. Network middleware cannot be agnostic to features like broadband or wavelength-selective switching, packet or flow-level completion, packet switching or flow-level switch reconfiguration or job-level/SDN switch reconfiguration. Simulation tools must therefore become more flexible, moving beyond assumptions of network stacks for electric packet switching that dominate current simulator designs. Some progress has been made in this direction, for example in providing compiler support to adapt network middleware source code to simulator backends.[35] Significant tool development remains, though, for simulation to support the full software/hardware co-design space with photonics.

5. A SYSTEMATIC AND RECONFIGURABLE APPROACH

As shown above, current approaches take considerable effort but also do not readily transfer to new photonic components because the entire process needs to be repeated. Also, current simulators present a too static view of the network. Tackling this challenge in a systematic and reconfigurable manner requires bridging the gap between photonic device designers and system architects. This means that low-level metrics need to be consistently translated to high-level metrics for the purpose of being added to simulation tools. To illustrate the design space of performance and energy, Table 1 shows an example collection of some metrics that are relevant at the device level and at the architecture level. In addition, the functionality at a clock cycle level of photonic devices should be described to better configure simulators.

To this end, we argue that the architecture and photonics communities should develop a design automation tool that facilitates this translation. Essentially, this means that observations from device-level experiments are scaled up to block (e.g., entire switch) level metrics, and metrics are converted appropriately. For channels this process is more straightforward, but larger and more complicated blocks such as switches require knowledge of the internal architecture. This can be handled by providing the means to describe new architectures as well as having a library of state-of-the-art designs for common building blocks such as multiplexers. In addition, technology libraries for electronics (e.g., CMOS transistors) should be included where appropriate, such as optical to electrical conversions. Figure 2 provides a high-level overview.

To make this vision a reality, we need to work towards solving some related challenges. For instance, the second step that composes devices to create an architecture needs to consider how these components interact, whether that is crosstalk, conversion latency, or wavelengths. This requires accurate models but also a synthesis step where a high-level metric such as energy is optimized while preserving constraints. In addition, translating low- to high-level metrics requires assumptions such as operating temperatures that this flow must make available

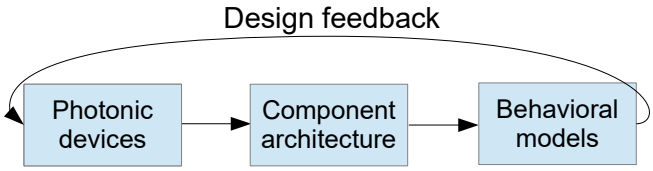

Figure 2. At a high level, the desired process will take device-level descriptions of photonic devices. Based on a described architecture that describes how each device is connected to each other to compose a photonic component, behavioral performance and power models are created for each component (e.g., switch or fiber). There is also a feedback loop such that system-level experiments can affect device design to optimize system impact. For instance, if the observed error rate at the system scale causes too much adversity, devices can reduce their error rate at the expense of performance or energy.

to the designer. Even so, that information may not be available in advance which risks affecting the accuracy of the process. Also, there needs to be a method to encode hard constraints such as maximum switch radices.

Further considerations include validation because combining models of different devices in a high-level simulation does not necessarily preserve correct behaviorial or energy models. Therefore, verification should be hierarchical, where small-scale models are compared against physical-level simulations or actual device demonstrations. From that, larger simulations should be built with consideration for design limitations such as maximum number of hops in the optical domain before re-amplification, assumptions on operating conditions, and others.

Still, the value in this approach is to provide a rapid and easy way to gage system-level impact for any emerging photonic device such that first order approximations can be made before more effort is spent to increase accuracy. If this process maintains sufficient reconfigurability to increase accuracy as well as extendability to support future device types and new description formats, this flow will provide a basis for future design methodologies. Verification becomes a grand challenge for future-looking devices where manufactures devices do not yet exist. Still, early simulations of future devices can provide invaluable insight as long as reasonably accurate models can be derived.

Moreover, while our current description simplistically assumes that the metrics at the system level are fixed numerical values (e.g., error rate 1% and latency of 2ns), this need not be the case. Ultimately, the relationship between variables is better described with functions. This will show the impact in error rate (as an example) if we chose to increase bandwidth of a fiber. This is useful information to system architects so they can rapidly examine tradeoffs at the system scale and thus be in a better position to pick design points based on a cost/benefit analysis. However, this requires exploring the design space at the device level to generate these functions and therefore may be impractical for some designs.

Interestingly, a systematic methodology for high-level modeling based on device models is also a challenge in computer architecture where the plethora of devices, memories, specialized architectures, 3D integration techniques, and other emerging technologies to preserve digital computing performance scaling.[36] This motivated the development of PARADISE[37] that takes as input physical level descriptions of emerging technologies and extracts chip-level power and performance metrics. This approach has been proven successful, which shows the promise when applying a similar approach in the photonics domain. In addition, PARADISE is open source to aid adoption. Realizing this goal required adding design automation functionality where not available from the open source domain, but also connecting existing software tools and providing a user interface.

6. CONCLUSION

In this paper, we discuss the need to quickly integrate photonics in future systems, the current challenges in extracting architectural-level metrics from device demonstration as well as conducting large-scale experiments with photonics. We then make the case for a systematic approach to bridge the gap between photonic device developers and system architects. We believe that work in this area promises to significantly accelerate adoption of photonic technologies at the system scale, as well as change system architectures to better match the strengths that photonic technologies provide. In addition, photonic developers will have a tool to show them how to maximize their impact at the system and application levels.

ACKNOWLEDGMENTS

This work was supported by ARPA-E ENLITENED Program (project award DE-AR00000843) and the National Security Agency (NSA) Laboratory for Physical Sciences (LPS) Research Initiative (project award FA8075-14-D-0002-0007).

REFERENCES

[1] Bergman, K., "Empowering flexible and scalable high performance architectures with embedded photonics," in [*2018 IEEE International Parallel and Distributed Processing Symposium, IPDPS 2018, Vancouver, BC, Canada, May 21-25, 2018*], 378 (2018).

[2] Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., Bates, S., Bhatia, S., Boden, N., Borchers, A., Boyle, R., Cantin, P.-l., Chao, C., Clark, C., Coriell, J., Daley, M., Dau, M., Dean, J., Gelb, B., Ghaemmaghami, T. V., Gottipati, R., Gulland, W., Hagmann, R., Ho, C. R., Hogberg, D., Hu, J., Hundt, R., Hurt, D., Ibarz, J., Jaffey, A., Jaworski, A., Kaplan, A., Khaitan, H., Killebrew, D., Koch, A., Kumar, N., Lacy, S., Laudon, J., Law, J., Le, D., Leary, C., Liu, Z., Lucke, K., Lundin, A., MacKean, G., Maggiore, A., Mahony, M., Miller, K., Nagarajan, R., Narayanaswami, R., Ni, R., Nix, K., Norrie, T., Omernick, M., Penukonda, N., Phelps, A., Ross, J., Ross, M., Salek, A., Samadiani, E., Severn, C., Sizikov, G., Snelham, M., Souter, J., Steinberg, D., Swing, A., Tan, M., Thorson, G., Tian, B., Toma, H., Tuttle, E., Vasudevan, V., Walter, R., Wang, W., Wilcox, E., and Yoon, D. H., "In-datacenter performance analysis of a tensor processing unit," in [*Proceedings of the 44th Annual International Symposium on Computer Architecture*], *ISCA '17*, 1–12 (2017).

[3] Yin, X., Verplaetse, M., Breyne, L., Kerrebrouck, J. V., Keulenaer, T. D., Vyncke, A., Pierco, R., Vaernewyck, R., Spiga, S., Amann, M. ., Chen, J., Steenberge, G. V., Torfs, G., and Bauwelinck, J., "Towards efficient 100 gb/s serial rate optical interconnects: A duobinary way," in [*2017 IEEE Optical Interconnects Conference (OI)*], 33–34 (June 2017).

[4] Sun, Y. and Lingle, R., "Technical feasibility of new 200 gb/s and 400 gb/s links for data centers," in [*2018 IEEE Optical Interconnects Conference (OI)*], 37–38 (June 2018).

[5] Plander, I. and Stepanovsky, M., "MEMS technology in optical switching," in [*2017 IEEE 14th International Scientific Conference on Informatics*], 299–305 (Nov 2017).

[6] Lea, C., "A scalable AWGR-based optical switch," *Journal of Lightwave Technology* **33**, 4612–4621 (Nov 2015).

[7] Descos, A., Seyedi, M. A., Chen, C., Fiorentino, M., Vincent, F., Penkler, D., Szelag, B., and Beausoleil, R. G., "Silicon photonics optical switch based on ring resonator," in [*2016 21st OptoElectronics and Communications Conference (OECC) held jointly with 2016 International Conference on Photonics in Switching (PS)*], 1–3 (July 2016).

[8] Wang, C. and Javidi, T., "Adaptive policies for scheduling with reconfiguration delay: An end-to-end solution for all-optical data centers," *IEEE/ACM Transactions on Networking* **25**, 1555–1568 (June 2017).

[9] Pries, R., Jarschel, M., Schlosser, D., Klopf, M., and Tran-Gia, P., "Power consumption analysis of data center architectures," in [*Green Communications and Networking*], Rodrigues, J. J. P. C., Zhou, L., Chen, M., and Kailas, A., eds., 114–124, Springer Berlin Heidelberg, Berlin, Heidelberg (2012).

[10] Li, L., Xiao, L., Cao, X., Qi, C., and Mao, Z., "Impact of BTI aging effect on soft error rate of combination circuit," in [*2017 Prognostics and System Health Management Conference (PHM-Harbin)*], 1–5 (July 2017).

[11] Rodrigues, A. F., Hemmert, K. S., Barrett, B. W., Kersey, C., Oldfield, R., Weston, M., Risen, R., Cook, J., Rosenfeld, P., Cooper-Balis, E., and Jacob, B., "The structural simulation toolkit," *SIGMETRICS Perform. Eval. Rev.* **38**, 37–42 (Mar. 2011).

[12] Binkert, N., Beckmann, B., Black, G., Reinhardt, S. K., Saidi, A., Basu, A., Hestness, J., Hower, D. R., Krishna, T., Sardashti, S., Sen, R., Sewell, K., Shoaib, M., Vaish, N., Hill, M. D., and Wood, D. A., "The gem5 simulator," *SIGARCH Comput. Archit. News* **39**, 1–7 (Aug. 2011).

[13] Sun, C., Chen, C. O., Kurian, G., Wei, L., Miller, J., Agarwal, A., Peh, L., and Stojanovic, V., "Dsent - a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling," in [*2012 IEEE/ACM Sixth International Symposium on Networks-on-Chip*], 201–210 (May 2012).

[14] Wilton, S. J. E. and Jouppi, N. P., "Cacti: an enhanced cache access and cycle time model," *IEEE Journal of Solid-State Circuits* **31**, 677–688 (May 1996).

[15] Rumley, S., Bahadori, M., Wen, K., Nikolova, D., and Bergman, K., "Phoenixsim: Crosslayer design and modeling of silicon photonic interconnects," in [*Proceedings of the 1st International Workshop on Advanced Interconnect Solutions and Technologies for Emerging Computing Systems*], *AISTECS '16*, 7:1–7:6 (2016).

[16] Martin, P., Gays, F., Grellier, E., Myko, A., and Menezo, S., "Modeling of silicon photonics devices with verilog-a," in [*2014 29th International Conference on Microelectronics Proceedings - MIEL 2014*], 209–212 (May 2014).

[17] Zhang, Z., Wu, R., Wang, Y., Zhang, C., Stanton, E. J., Schow, C. L., Cheng, K., and Bowers, J. E., "Compact modeling for silicon photonic heterogeneously integrated circuits," *Journal of Lightwave Technology* **35**, 2973–2980 (July 2017).

[18] Bahadori, M., Rumley, S., Polster, R., Gazman, A., Traverso, M., Webster, M., Patel, K., and Bergman, K., "Energy-performance optimized design of silicon photonic interconnection networks for high-performance computing," in [*Design, Automation Test in Europe Conference Exhibition (DATE), 2017*], 326–331 (March 2017).

[19] Bahadori, M. and Bergman, K., "Low-power optical interconnects based on resonant silicon photonic devices: Recent advances and challenges," in [*Proceedings of the 2018 on Great Lakes Symposium on VLSI*], *GLSVLSI '18*, 305–310 (2018).

[20] Padmaraju, K., Logan, D. F., Shiraishi, T., Ackert, J. J., Knights, A. P., and Bergman, K., "Wavelength locking and thermally stabilizing microring resonators using dithering signals," *Journal of Lightwave Technology* **32**, 505–512 (Feb 2014).

[21] Singh, A., Ong, J., Agarwal, A., Anderson, G., Armistead, A., Bannon, R., Boving, S., Desai, G., Felderman, B., Germano, P., et al., "Jupiter rising: a decade of clos topologies and centralized control in google's datacenter network," *Communications of the ACM* **59**(9), 88–97 (2016).

[22] Graham, R. L., Shipman, G. M., Barrett, B. W., Castain, R. H., Bosilca, G., and Lumsdaine, A., "Open MPI: A High-Performance, Heterogeneous MPI," in [*2006 IEEE International Conference on Cluster Computing*], 1–9 (2006).

[23] "Information Technology – Open Systems Interconnection – Basic Reference Model: The Basic Model," standard, International Organization for Standardization, Geneva, CH (Mar. 1996).

[24] Shen, Y., Hattink, M. H. N., Samadi, P., Cheng, Q., Hu, Z., Gazman, A., and Bergman, K., "Software-defined networking control plane for seamless integration of multiple silicon photonic switches in Datacom networks," *Opt. Exp.* **26**, 10914–10929 (2018).

[25] Jiang, N., Kim, J., and Dally, W. J., "Indirect Adaptive Routing on Large Scale Interconnection Networks," in [*ISCA 2009: 36th Annual International Symposium on Computer Architecture*], 220–231 (2009).

[26] Salim, J. H. and Ahmed, U., "Performance Evaluation of Explicit Congestion Notification (ECN) in IP Networks," (2000).

[27] Fiala, D., Mueller, F., Engelmann, C., Riesen, R., Ferreira, K., and Brightwell, R., "Detection and correction of silent data corruption for large-scale high-performance computing," in [*SC '12: International Conference on High Performance Computing, Networking, Storage and Analysis*], 1–12 (2012).

[28] Eberle, H. and Dennison, L., "Light-weight protocols for wire-speed ordering," in [*SC '18: International Conference for High Performance Computing, Networking, Storage, and Analysis*], 1–12 (2018).

[29] Wolf, W. H., "Hardware-software co-design of embedded systems," *Proc. IEEE* **82**, 967–989 (1994).

[30] Groves, T., Grant, R. E., Hemmer, S., Hammond, S., Levenhagen, M., and Arnold, D. C., "(SAI) Stalled, Active and Idle: Characterizing Power and Performance of Large-Scale Dragonfly Networks," in [*2016 IEEE International Conference on Cluster Computing (CLUSTER)*], 50–59 (2016).

[31] Degomme, A., Legrand, A., Markomanolis, G. S., Quinson, M., Stillwell, M., and Suter, F., "Simulating MPI Applications: The SMPI Approach," *IEEE Transactions on Parallel Distrib. Syst.* **28**, 2387–2400 (2017).

[32] Jain, N., Bhatele, A., White, S., Gamblin, T., and Kale, L. V., "Evaluating HPC Networks via Simulation of Parallel Workloads," in [*SC16: International Conference for High Performance Computing, Networking, Storage and Analysis*], 154–165 (2016).

[33] Blumrich, M. A., Jiang, N., and Dennison, L. R., "Exploiting idle resources in a high-radix switch for supplemental storage," in [*SC '18: International Conference for High Performance Computing, Networking, Storage, and Analysis*], 1–13 (2018).

[34] Hoefler, T., Schneider, T., and Lumsdaine, A., "LogGOPSim: simulating large-scale applications in the LogGOPS model," in [*HPDC: International Symposium on High Performance Distributed Computing*], 597–604 (2010).

[35] Wilke, J. J., Kenny, J., and Knight, S., "Supercomputer in a laptop: Distributed application and runtime development via architecture simulation," in [*(To appear) International Workshop on Communication Architectures for HPC, Big Data, Deep Learning and Clouds at Extreme Scale (Exacomm)*], (2018).

[36] Shalf, J. M. and Leland, R., "Computing beyond moore's law," *Computer* **48**, 14–23 (Dec 2015).

[37] Vasudevan, D., Butko, A., Michelogiannakis, G., Donofrio, D., and Shalf, J., "Towards an integrated strategy to preserve digital computing performance scaling using emerging technologies," in [*High Performance Computing*], Kunkel, J. M., Yokota, R., Taufer, M., and Shalf, J., eds., 115–123, Springer International Publishing, Cham (2017).