Driving Down Link Energy and Driving Up Link Density in GPU Networks

Benjamin G. Lee
NVIDIA, 2600 Meridian Pkwy, Durham, NC 27713
benjlee@nvidia.com

Abstract: GPU-accelerated computing systems, which power the AI revolution, rely on increasing amounts of off-chip I/O. To continue scaling, very dense integration of ultra-efficient optical transceivers alongside next-generation processor die will be needed. © 2022 The Author(s).

1. GPU Networks

The world's high-performance computing (HPC) systems are increasingly leveraging graphics processing units (GPU) to accelerate computation. Of the top 500 machines currently ranked by top500.org [1], 147 use accelerated co-processor architectures, up from 110 three years ago. Of those 147 systems, 139 use NVIDIA GPUs. These systems provide the backbone for exploring society's most challenging computational problems—such as drug discovery, weather and climate prediction, fuel-cell optimization, and genome sequencing. In addition to traditional scientific applications, areas exploiting data science—including artificial intelligence (AI) and machine learning—leverage advanced computing technology to increase automation and improve efficiencies across both science and business. In fact, the AI-driven semiconductor market is expected to grow five times faster than for non-AI semiconductor applications through 2025 [2]. Due to the growing costs to procure and power HPC machines, many are looking toward computing in the cloud [3], where the volume of computing resources far outweighs today's most advanced HPC machines. Though less customizable, cloud-scale computing offers per-use pricing models that expand access to accelerated computing. Continuing to scale HPC and cloud computing resources in a cost-feasible and energy-efficient manner will be one of the most important technical accomplishments of our time.

A computing system can be scaled by increasing the computing power per node (scale up) or the number of interconnected nodes (scale out). In both cases, a high-performance switched interconnect has become pivotal to maintaining performance at scale. Scale-up switching is illustrated by NVIDIA's NVSwitch, which provides 6.4 Tb/s of bandwidth to a local network of eight to sixteen A100 NVIDIA GPUs in the DGX A100 system [4]. Scale out is embodied by a multitude of HPC and cloud systems, e.g. [5], where switched interconnect fabrics such as InfiniBand or Ethernet use multiple stages of switches and links to connect thousands of GPU-accelerated nodes.

2. Switch ASIC Scaling Trends

To better understand the scaling challenges, consider the bandwidths provided by application-specific integrated circuits (ASIC) built for switching. Fig. 1(a) plots a sampling of data from a variety of vendors of commercial switch ASICs over the past two decades. Electrical switch ASICs have maintained a remarkable scaling trend, roughly doubling bandwidth every two years, bringing the state-of-the-art switch bandwidth from < 100 Gb/s in the early 2000s to > 25 Tb/s today. Switch vendors have leveraged CMOS scaling to continually increase bandwidth while keeping chip area constrained. Fig. 1(b) charts the evolution of CMOS nodes used by the switch ASICs reported in Fig. 1(a) as a function of switching bandwidth, and Fig. 1(c) highlights the gains in energy efficiency by plotting the chip power divided by the bandwidth (i.e. energy per bit). As bandwidths have scaled, energy per bit has decreased in large part due to CMOS scaling. However, energy per bit has not decreased fast enough to keep power envelopes bounded, and power levels will soon be approaching 1 kW. Exacerbating the problem, chip powers will increase more rapidly in the future than they have in the past due to a reduction in the pace—and eventually an end—of CMOS scaling.

Rising power levels are forcing costly changes in system architecture. To control cost, cloud architects have traditionally preferred air-cooling using low-cost fans and heat sinks, while sufficiently spacing out compute infrastructure to keep power densities constrained. Now, the chip-level power dissipations are necessitating more complex and costly liquid-cooling solutions. Furthermore, copper cable power efficiency and reach are diminishing at the increased signal speeds. This places pressure on designers to either construct more integrated systems with shorter link lengths—aggravating the power density challenges, or accelerate the adoption of optics, which relaxes the density requirements but also increases cost. Architects face a challenging optimization problem, balancing the performance-per-cost of compute, cooling, and communication systems.

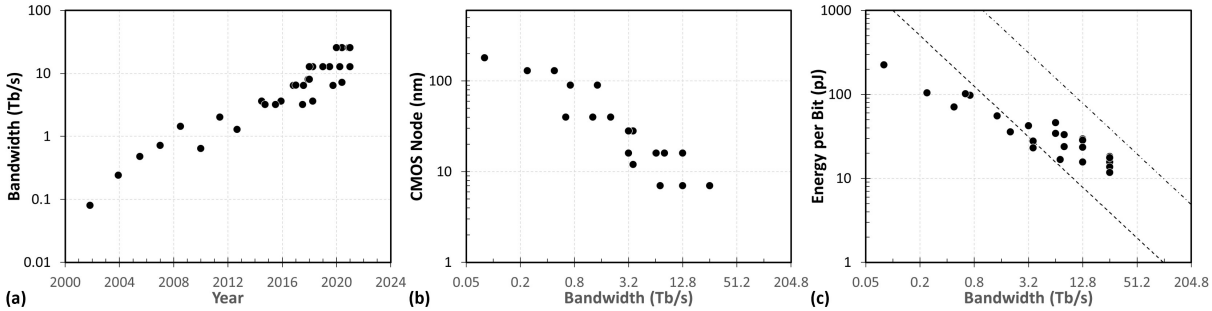

Fig. 1: Sampling of publicly available historical data for commercial Ethernet and InfiniBand switch ASICs produced by a variety of vendors. (a) Switch bandwidth versus date of announcement. (b) CMOS node and (c) energy per bit versus switch chip bandwidth. The dashed and dash-dotted lines demark power contours of 100 W and 1 kW, respectively.

3. Co-Packaged Optics Opportunities and Constraints

The fraction of switch power devoted to communication in and out of the chip is also increasing [6,7]. The Optical Internetworking Forum (OIF) defines Common Electrical I/O (CEI) standards for off-chip electrical interfaces [8]. The CEI-112G-LR (long reach) standard provides for 1-m of reach over a twin-ax cable with 2 connectors at 112 Gb/s, while CEI-112G-MR (medium reach) allows for 0.5-m across a printed circuit board with 1 connector at the same rate. Energy estimates of 112-Gb/s LR interfaces are > 5 pJ/b, and MR affords only a slight improvement to ~ 4 pJ/b [9]. A 100-Tb/s switch using a purely electrical LR (MR) interface would consume at least 500 W (400 W) in the chip for off-chip communication only. Both interfaces can also be used to connect to on-board or pluggable edge-of-card optics for longer reach. In addition to adding cost, the optical modules increase system power while doing nothing to reduce the power consumed in the ASIC. Clearly, more efficient interfaces are needed.

Co-packaged optics (CPO) provide potentially reduced chip power envelopes while also eliminating the reach limitations imposed by purely electrical signaling. Several CPO prototypes were recently demonstrated [10-12]. By integrating the optics on package with the ASIC, the electrical interface efficiency can be improved. The CEI-112G-XSR standard provides for up to 100 mm of electrical wiring on an organic multi-chip module (MCM). Projections target ~ 1-2 pJ/b per interface [9,13-15]. For CPO, off-chip communication must traverse two electrical links and one optical link; all three dissipate power within the switch package. Therefore, XSR interfaces to CPO in a 100-Tb/s switch may be an improvement over LR, but only with very efficient optics, and even so, the off-chip communication will still account for a considerable amount of in-package power. Moreover, XSR interfaces on MCM achieve electrical edge bandwidth densities ~ 1 Tb/s/mm [15]. This is 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), but needs further improvement. Therefore, for the 100-Tb/s generation CPO on MCM has the potential for modest power savings at the switch module, and—with improvement—potentially enough bandwidth density. (Although the module power may not be dramatically reduced, the overall system power may, since the power consumed by the pluggable optics can be eliminated.) Scaling beyond 100 Tb/s will require denser integration with electrical edge bandwidth densities of multi-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 [16-18].

4. Scaling with Densely Integrated Co-Packaged Optics

Integrating optics this close to the ASIC creates additional challenges. The bandwidth density (both edge and areal) and energy per bit of the optical elements becomes even more critical. For example, where CPO on MCM provides a module edge (~ 100 mm) several times longer than the chip edge (~ 25 mm), 2.5D integration requires optical edge bandwidth densities on par with that of the electrical edge bandwidth densities emerging from the ASIC.

To keep power and footprint as small as possible, one needs to remove unnecessary components. 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 laser 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.

The optical edge bandwidth density must be scaled primarily in the time and frequency domains since spatial density is limited by fiber diameter in practical near-term systems. Faster signaling incurs a premium on energy consumption [19]; thus, wavelength-domain scaling is preferred. Coarse wavelength-division multiplexing (WDM) systems are in use today, but dense WDM will be needed to meet the bandwidth density targets. Cost-effective

dense WDM solutions for the lasers and optical transceivers are not commercially available today, and significant development will be needed. Polarization multiplexing and PAM-4 modulation may each be used to double bandwidth density as well. Altogether, if successful, this approach may lead to edge bandwidth densities approaching 30 Tb/s/mm, which must then be divided across supply, transmit, and receive fibers [20,21].

A micro-ring resonator-based link architecture provides several benefits. It is compatible with dense WDM. It eliminates grating- and interferometer-based wavelength multiplexers/demultiplexers, which occupy significant area. The ring modulators and filters are area and energy efficient. The spectral bandwidth is limited by the free-spectral range, which has been demonstrated out to about 3 THz [22,23]. Micro-ring resonators require integrated control, but the power and area overhead for the control circuits is relatively small in advanced CMOS nodes.

Such an architecture can deliver the power and area efficiencies needed for highly integrated optical engines. However, many other challenges remain for such tight integration of optics. Packaging becomes much more complicated. Socketed solutions are no longer possible. Fiber coupling becomes very constrained. Nevertheless, overcoming these challenges can enable a continuation of bandwidth scaling.

5. Conclusions

Dense WDM links employing micro-ring resonators may provide the energy and area efficiency needed for CPO beyond 100 Tb/s. In these systems, silicon interconnect between the switch and transceivers will be required to achieve the needed density. Optical interfaces to the package will have to support multi-Tb/s/mm. Electrical interfaces operating at 0.25 pJ/b and 1-pJ/b optical links—including drivers, tuning, and control—can deliver a total off-chip communications power of 300 W for a 200-Tb/s switch. Once in the optical domain, reaches of 100 m to 1 km can easily be achieved, decoupling aspects of system design from locality. Finally, if achieved, the densely integrated solution will not only help computing systems continue to scale up and out via switched interconnect; it can also be replicated within GPUs, CPUs, and other ASICs to improve efficiency across the entire machine.

Acknowledgments

The author would like to thank T. Greer, W. Turner, and T. Gray for helpful discussions.

References

[1] "TOP500 | June 2021 List." Available at https://top500.org/. Accessed Nov. 3, 2021.
[2] G. Batra *et al.*, "Artificial-intelligence hardware: New opportunities for semiconductor companies," *McKinsey & Co. Mag.*, p. 1 (2018).
[3] G. B. Berriman *et al.*, "The application of cloud computing to scientific workflows: a study of cost and performance," *Phil. Trans. R. Soc.*, **371** (1983), p. 1 (2013).
[4] J. Choquette *et al.*, "The A100 datacenter GPU and Ampere architecture," *Int. Solid- State Circuits Conf. (ISSCC)*, p. 48 (2021).
[5] C. B. Stunkel *et al.*, "The high-speed networks of the Summit and Sierra supercomputers," *IBM J. Res. and Dev.*, **64** (3/4), p. 3:1 (2020).
[6] C. Minkenberg *et al.*, "Co-packaged datacenter optics: Opportunities and challenges," *IET Optoelectron.*, **15** (2), p. 77 (2021).
[7] R. Chopra, "Co-packaged optics and an open ecosystem," *Cisco Blogs*, published Jan. 11, 2021. Available at https://blogs.cisco.com/sp/co-packaged-optics-and-an-open-ecosystem.
[8] Optical Internetworking Forum, "Common electrical I/O (CEI)-112G," Available at https://www.oiforum.com/technical-work/hot-topics/common-electrical-interface-cei-112g-2/. Accessed Nov. 3, 2021.
[9] M. Y. Frankel, "Prospects for optical transceivers expanding to access, metro and long-haul," *Opt. Fiber Commun. Conf. (OFC)*, paper Tu5A.2 (2021).
[10] P. Maniotis *et al.*, "Toward lower-diameter large-scale HPC and data center networks with co-packaged optics," *J. Opt. Commun. and Netw.*, **13** (1), p. A67 (2021).
[11] N. Margalit *et al.*, "Perspective on the future of silicon photonics and electronics", *Appl. Phys. Lett.*, **118** (220501), p. 1 (2021).
[12] R. Mahajan *et al.*, "Co-packaged photonics for high performance computing: status, challenges and opportunities," *J. Lightw. Technol.*, pre-print.
[13] R. Ward, "Co-packaged optics, near-packaged optics, or neither?" *IEEE Photon. Conf. (IPC)*, workshop WG3 (2021).
[14] D. Goodwill, "The road to co-packaged optics," *IEEE Photon. Conf. (IPC)*, workshop WG3 (2021).
[15] R. Shivnaraine *et al.*, "A 26.5625-to-106.25Gb/s XSR SerDes with 1.55pJ/b Efficiency in 7nm CMOS," *Int. Solid- State Circuits Conf. (ISSCC)*, p. 181 (2021).
[16] R. Mahajan *et al.*, "Embedded multi-die interconnect bridge (EMIB) -- A high density, high bandwidth packaging interconnect," *Electron. Components Technol. Conf. (ECTC)*, p. 557 (2016).
[17] P. K. Huang *et al.*, "Wafer level system integration of the fifth generation CoWoS®-S with high performance Si interposer at 2500 mm^2," *Electron. Components Technol. Conf. (ECTC)*, p. 101 (2021).
[18] K. Sikka *et al.*, "Direct bonded heterogeneous integration (DBHi) Si bridge," *Electron. Components Technol. Conf. (ECTC)*, p. 136 (2021).
[19] D. Miller, "Getting to femtojoule optics – what physics and what technology?" *Opt. Fiber Commun. Conf. (OFC)*, paper Tu5A.3 (2021).
[20] H. Hou, "Overview of silicon photonics design, process, and applications," *IEEE Photon. Conf. (IPC)*, plenary talk TuA1.2 (2021).
[21] K. Giewont *et al.*, "Path to beyond 1Tb/s optical channels in monolithic SiPh foundry," *IEEE Photon. Conf. (IPC)*, workshop WG3 (2021).
[22] M. Wade *et al.*, "An error-free 1 Tbps WDM optical I/O chiplet and multi-wavelength multi-port laser," *Opt. Fiber Commun. Conf. (OFC)*, paper F3C.6 (2021).
[23] M. Sakib *et al.*, "A high-speed micro-ring modulator for next generation energy-efficient optical networks beyond 100 Gbaud," *Conf. Lasers Electro-Optics*, paper SF1C.3 (2021).