European Conference on Optical Communication (ECOC) 2022 © Optica Publishing Group 2022

Photonic Circuits for Accelerated Computing Systems

Benjamin G. Lee

NVIDIA, benjlee@nvidia.com

Abstract *GPU-based accelerated computing is powering the AI revolution. These systems include processors and switches which push thermal power density limits while demanding large I/O bandwidths. To continue scaling, very dense integration of ultra-efficient optical transceivers is called for to alleviate current inefficiencies in off-package signalling. ©2022 The Author.*

Accelerated Computing

High-performance computing (HPC) systems are increasingly turning to accelerated co-processor architectures leveraging graphics processing units (GPU) to maintain the pace of performance scaling. Of the world's 500 highest-performing computers, as ranked by top500.org [1], 169 use accelerated co-processors and 161 of those use NVIDIA GPUs. Fig. 1 shows how this trend has been increasing over the past twelve years. These few but ultra-performant systems provide the backbone for exploring society's most computationally challenging scientific investigations—such as new drug discovery, weather and climate prediction, fuel-cell optimization, and genome sequencing. In addition to traditional scientific computing applications, artificial intelligence (AI) and machine learning rely on advanced computing technology to increase automation and improve efficiencies for a broad scope of industries across science and business. This has led to an explosion in demand for compute acceleration that is finding its fulfilment in the cloud, where resource sharing provides cost optimization over a large range of job sizes.

Continuing to scale both HPC and cloud-based accelerated computing resources in a cost-feasible and energy-efficient manner will be an important focus area over the next decade. A high-performance network—including state-of-the-art switches and interconnects—is key to maintaining performance as systems scale, both in single node performance (scale up) and number of connected nodes (scale out). For example, in the recently announced DGX H100 system [2], four NVSwitches provide > 50 Tb/s of aggregate bandwidth to a local network of eight H100 GPUs. DGX boxes can then be linked together through an optically connected NVLink network or attached to an InfiniBand or Ethernet fabric.

Switch ASIC Scaling Trends

Electrical switch application-specific integrated circuits (ASIC) 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 > 50 Tb/s today. Switch vendors have leveraged CMOS scaling to continually increase bandwidth while keeping chip area constrained. As bandwidths have scaled, energy per bit has decreased (approximately 15 pJ/b in many of today's commercial switches), but not fast enough to keep power envelopes bounded. ASIC power is now or will be soon approaching 1 kW. Exacerbating the problem, chip powers are expected to increase more rapidly in the future than they have in the past due to the reduced pace—and eventually end—of CMOS scaling.

Part of the rise in chip power is due to signalling constraints. The overall portion of ASIC power spent on input/output (I/O) signalling has been increasing over the past few generations [3]. One of the primary challenges for switch scaling is getting the I/O signals into and out of the chip with reasonable power and cost.

Electrical Interfaces

The Optical Internetworking Forum's (OIF) Common Electrical I/O (CEI) 112-Gb/s long reach (LR) standard [4] provides for 1-m of reach. Measured energies of demonstrated 112-Gb/s

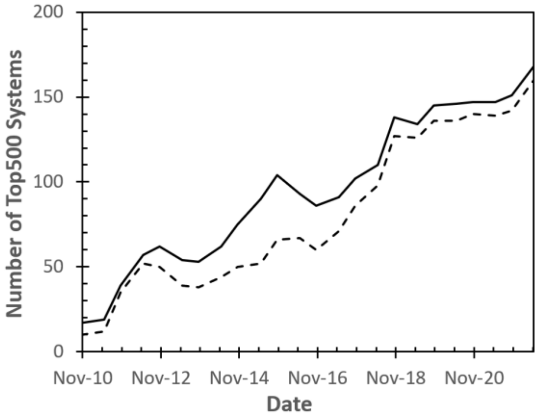

Fig. 1: Data from [1] showing—of the 500 highest-ranking computing systems worldwide—the number using an accelerated co-processor architecture (solid line) and the number of those systems using NVIDIA GPUs (dashed line), as it has progressed over the last twelve years.

LR interfaces are 4.5 to 6.5 pJ/b [5, 6]. LR interfaces are also used to connect to on-board or pluggable edge-of-card optics for extended reach. The optical modules typically add another 10 pJ/b or more to the system power.

Co-packaged optics (CPO) hold promise to reduce chip power while also extending reach compared to purely electrical signalling. By integrating the optics on the multi-chip module (MCM) package along with the ASIC, the electrical interface efficiency can be improved. However, now both ends of the electrical link dissipate heat within the package environment, essentially doubling the impact of each pico-Joule per bit contributed by the electrical portion of the link. The CEI-112G-XSR (extreme short reach) standard provides for up to 100 mm of reach on an organic MCM. Demonstrations have achieved 1.24 to 1.7 pJ/b [7–10]. XSR interfaces have shown electrical edge bandwidth densities ranging from 475 to 870 Gb/s/mm [7–9].

As the need arises for further improvement in energy efficiency and bandwidth density, denser integration of the optics with the ASIC will be required. Here, 2.5D integration on silicon interposer or local silicon interconnect can meet the next wave of demand [11–13]. In this approach, the tighter integration not only reduces the transmission distance, but more importantly, the 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 electrical interface where 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 [14].

CPO Integration on Interposer

Integrating optics on the interposer (or attaching them through a bridge layer) with the ASIC creates additional challenges. The edge and areal bandwidth densities and the energy per bit of the optical components become even more constrained. Using a remote laser supply is one way to help keep the in-package power and footprint as small as possible, while additionally improving laser performance and lifetime. Micro-ring resonators can save substantial power and energy compared to Mach Zehnder interferometers for modulating and filtering light.

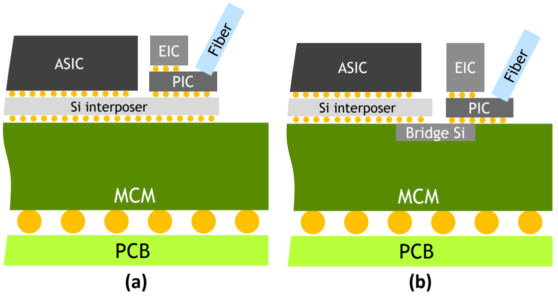

Fig. 2: Two concepts for 2.5D integration of optics alongside an ASIC: (a) integration of optics onto the interposer and (b) integration of optics by way of a bridge layer of local silicon interconnect.

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 even more complicated than CPO on MCM. Fig. 2 illustrates two approaches for realizing 2.5D integrated optics. Of course, many variations are possible, and each of these depend on the availability of a number of foundry-enabled features. In any scheme, socketed solutions will not be feasible, and fiber-last assembly will likely be required, preferably with a removable optical connector that allows replacing broken fibers.

Realizing aggressive optical edge bandwidth densities requires scaling primarily in the frequency (wavelength) domain, since spatial density is limited by fiber diameter in practical near-term systems, and faster signalling incurs a premium on energy consumption [15]. Coarse wavelength-division multiplexing (WDM) systems are in use today, but dense WDM will be needed. Cost-effective dense WDM solutions for the lasers are not yet commercially available, and further development is required. Polarization multiplexing and PAM-4 modulation may each be used to double bandwidth density, but further scaling along these dimensions is constrained. Nevertheless, with reasonable baud rates (e.g. 32 Gbaud), practical fiber pitch (127 μm), and attainable wavelength counts (16), edge bandwidth densities beyond 10 Tb/s/mm may be realizable.

Conclusions

Tighter integration between optical engines and switch silicon may be an attractive way to continue bandwidth scaling beyond the 100 Tb/s switch generation. Integrating the optics onto the interposer alongside the ASIC not only shortens the electrical interface length but also due to higher density allows operating at more energy efficient baud rates. In such a system, optical interfaces to the package will have to support multi-Tb/s/mm bandwidth densities with pico-Joule-per-bit scale energy efficiencies. Once in the optical domain, reaches of 100 m to 1 km can be achieved, decoupling aspects of system design from locality. Dense WDM links employing micro-ring resonators may provide the energy and area efficiency needed to realize 2.5D integrated CPO. 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 processing chips to improve the efficiency of data movement across the entire machine.

Acknowledgements

The author wishes to thank the many NVIDIA colleagues who have contributed to shaping the ideas discussed here.

References

[1] "TOP500 | June 2022 List," Available at https://top500.org/. Accessed May 31, 2022.

[2] "NVIDIA H100 tensor core GPU architecture: Exceptional performance, scalability, and security for the data center," NVIDIA White paper. Available at https://resources.nvidia.com/en-us-tensor-core/. Accessed May 31, 2022.

[3] 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.

[4] 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 May 31, 2022.

[5] P. Mishra, A. Tan, B. Helal, C.R. Ho, C. Loi, J. Riani, J. Sun, K. Mistry, K. Raviprakash, L. Tse, M. Davoodi, M. Takefman, N. Fan, P. Prabha, Q. Liu, Q. Wang, R. Nagulapalli, S. Cyrusian, S. Jantzi, S. Scouten, T. Dusatko, T. Setya, V. Giridharan, V. Gurumoorthy, V. Karam, W. Liew, Y. Liao, Y. Ou, "A 112Gb/s ADC-DSP-based PAM-4 transceiver for long-reach applications with >40dB channel loss in 7nm FinFET," in *Proceedings IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 138-140, Feb. 2021, doi: 10.1109/ISSCC42613.2021.9365929.

[6] Z. Guo, A. Mostafa, A. Elshazly, B. Chen, B. Wang, C. Han, C. Wang, D. Zhou, D. Visani, E. Hsiao, F. Chu, F. Lu, G. Cui, H. Zhang, H. Wang, H. Zhao, J. Lin, J. Gu, L. Luo, L. Jiang, M. Singh, M. Gambhir, M. Hasan, M. Wu, M. J. Yoo, P. Liu, S. Kollu, T. Ye, X. Zhao, X. Yang, X. Han, Y. Huang, Y. Sun, Z. Yu, Z. H. Jiang, Z. Adal, Z. Yan, "A 112.5Gb/s ADC-DSP-based PAM-4 long-reach transceiver with >50dB channel loss in 5nm FinFET," in *Proceedings IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 116-118, Feb. 2022, doi: 10.1109/ISSCC42614.2022.9731650.

[7] R. Shivnaraine, M. van Ierssel, K. Farzan, D. Diclemente, G. Ng, N. Wang, J. Musayev, G. Dutta, M. Shibata, A. Moradi, H. Vahedi, M. Farzad, P. Kainth, M. Yu, N. Nguyen, J. Pham, A. McLaren, "A 26.5625-to-106.25Gb/s XSR SerDes with 1.55pJ/b efficiency in 7nm CMOS," in *Proceedings International Solid- State Circuits Conference (ISSCC)*, p. 181, Feb. 2021, doi: 10.1109/ISSCC42613.2021.9365975.

[8] G. Gangasani, D. Hanson, D. Storaska, H. H. Xu, M. Kelly, M. Shannon, M. Sorna, M. Wielgos, P. B. Ramakrishna, S. Shi, S. Parker, U. K. Shukla, W. Kelly, W. Su, Z. Yu, "A 1.6Tb/s chiplet over XSR-MCM channels using 113Gb/s PAM-4 transceiver with dynamic receiver-driven adaptation of TX-FFE and programmable roaming taps in 5nm CMOS," in Proceedings *International Solid-State Circuits Conference (ISSCC)*, pp. 122-124, Feb. 2022, doi: 10.1109/ISSCC42614.2022.9731636.

[9] C. F. Poon, W. Zhang, J. Cho, S. Ma, Y. Wang, Y. Cao, A. Laraba, E. Ho, W. Lin, D. Z. Wu, K. H. Tan, P. Upadhyaya, Y. Frans, "A 1.24-pJ/b 112-Gb/s (870 Gb/s/mm) transceiver for in-package links in 7-nm FinFET," *IEEE Journal of Solid-State Circuits*, vol. 57, no. 4, pp. 1199-1210, Apr. 2022, doi: 10.1109/JSSC.2022.3141802.

[10] R. Yousry, E. Chen, Y.-M. Ying, M. Abdullatif, M. Elbadry, A. ElShater, T.-B. Liu, J. Lee, D. Ramachandran, K. Wang, C.-H. Weng, M.-L. Wu, T. Ali, "A 1.7pJ/b 112Gb/s XSR transceiver for intra-package communication in 7nm FinFET technology," in *Proceedings IEEE International Solid- State Circuits Conference (ISSCC)*, pp. 180-182, Feb. 2021, doi: 10.1109/ISSCC42613.2021.9365752.

[11] R. Mahajan, R. Sankman, N. Patel, D.-W. Kim, K. Aygun, Z. Qian, Y. Mekonnen, I. Salama, S. Sharan, D. Iyengar, D. Mallik, "Embedded multi-die interconnect bridge (EMIB) -- A high density, high bandwidth packaging interconnect," in *Proceedings Electronics Components Technology Conference (ECTC)*, p. 557-565, May 2016, doi: 10.1109/ECTC.2016.201.

[12] P. K. Huang, C. Y. Lu, W. H. Wei, C. Chiu, K. C. Ting, C. Hu, C. H. Tsai, S. Y. Hou, W. C. Chiou, C. T. Wang, D. Yu, "Wafer level system integration of the fifth generation CoWoS®-S with high performance Si interposer at 2500 mm2," in *Proceedings Electronics Components Technology Conference (ECTC)*, p. 101-104, Jun. 2021, doi: 10.1109/ECTC32696.2021.00028.

[13] K. Sikka, R. Bonam, Y. Liu, P. Andry, D. Parekh, A. Jain, M. Bergendahl, R. Divakaruni, M. Cournoyer, P. Gagnon, C. Dufort, I. de Sousa, H. Zhang, E. Cropp, T. Wassick, H. Mori, S. Kohara, "Direct bonded heterogeneous integration (DBHi) Si bridge," in *Proceedings Electronics Components Technology Conference (ECTC)*, p. 136-147, Jun. 2021, doi: 10.1109/ECTC32696.2021.00034.

[14] Y. Nishi, J. W. Poulton, X. Chen, S. Song, B. Zimmer, W. J. Turner, S. G. Tell, N. Nedovic, J. M. Wilson, W. J. Dally, C. T. Gray, "A 0.297-pJ/bit 50.4-Gb/s/wire inverter-based short-reach simultaneous bidirectional transceiver for die-to-die interface in 5nm CMOS," in *Proceedings Symposium on VLSI Technology and Circuits (VLSI)*, paper C17-4, Jun. 2022.

[15] D. A. B. Miller, "Getting to femtojoule optics – what physics and what technology?" in *Proceedings Optical Fiber Communications Conference (OFC)*, paper Tu5A.3, San Diego, California, Mar. 2021.