High-Performance CMOS Photonic Interfaces: From AI to IoT

(Plenary)

Rajeev J. Ram

Massachusetts Institute of Technology, Cambridge, MA, 02139, USA
* rajeev@mit.edu

Computing with reduced power and increased speed is required for the *ubiquitous data* era. While such advancements are increasingly hard to achieve with conventional CMOS logic, the CMOS ecosystem enables a host of devices beyond electronics – including *native* photonic components and heterogeneous integration. Here, we present photonic interfaces from data processing and sensing at the edge to high-performance classical and quantum computing.

Keywords: *CMOS, Integrated Photonics, Silicon Photonics, High Performance Computing*

INTRODUCTION

The last 36 years has seen a steady increase in the deployment of photonic integrated components. Over most of this history, the development of integrated photonic systems in both III-V and Group IV materials has been driven by the needs of fibre optic systems – from telecommunications, to LAN, to FTTH and data centre networks. In these applications, progress has been driven primarily by the properties of the transmission media (single-mode vs multi-mode fibre, fibre gain, etc). Today, photonic integration is increasingly driven by the unique properties of high-performance electronic-photonic interfaces. The low-capacitance, low-energy, high-bandwidth density of photonic integrated systems is now driving optical interconnection into board-scale, chip-scale, and intra-chip photonic systems.

Here, we consider new applications ranging from (1) deep learning systems and other data intensive classical compute applications, (2) optically addressed quantum computing fabrics – with tremendous progress being made today in the area of trapped-ion quantum computing, and (3) next-generation brain-computer interfaces where photonics may play an important role in the massively parallel signal detection.

DATA INTENSIVE COMPUTING

Today, deep learning algorithms are employed in an ever-expanding range of applications. They represent an important class of applications for compute systems from mobile phones to large-scale data centres. Recent, energy analysis on a range of representative deep learning neural networks (DNNs) has demonstrated that data movement represents a larger fraction of energy consumption than computation (multiplication-and-accumulation)[1]. While local (on-chip) memory access can be less expensive than accessing larger off-chip DRAM, together memory access is 2x (local) and 200x (DRAM) more energy intensive than computation resulting in as much as 90% of overall energy consumption devoted to data movement[1]. This memory bottleneck is well established in other data intensive computing applications and has motivated us to consider low-energy photonic interconnections[2]. CMOS photonic integration based on microring resonant modulators and detectors locked to arrays of DWDM light sources have been demonstrated as low-energy, high-bandwidth platforms for optical I/O [3]. As of 2019, CMOS photonic systems showed in-socket energy dissipation (<5pJ/bit) [4] comparable to in-package high bandwidth memory (HBM3 3.9 pJ/bit [5]). Continued innovation in the CMOS photonic platform – minimizing junction capacitance and increasing detector responsivity [6]– promises to further reduce energy consumption per bit by as much as 10x (<0.5 pJ/bit) – a reduction in GPU-to-memory access of this scale could dramatically reduce deep learning energy consumption. Optical I/O not only lower-energy access to larger memory resources but offers new networking possibilities. Recent analysis suggests that such terabit DWDM I/O could further reduce training times for deep learning by 1.9x [7].

OPTICALLY ADDRESSED QUANTUM COMPUTING

Trapped ion qubits enable high-fidelity quantum compute fabrics –these atomic qubits offer low decoherence over typical gate operation times. The performance of these systems is already competitive with superconducting, transmon based quantum compute architectures [8]and has motivated large investments by companies such as IonQ and Quantinuum/Honeywell. We proposed a photonic integrated architecture for scalable trapped-ion quantum computation. A visible waveguide platform employing focusing vertical grating couplers offers a path to

simultaneous addressing of large ensembles of trapped ion qubits[9]. The performance of each qubit can improved because (1) the waveguide devices can be integrated directly into the surface electrode Paul trap hence eliminating beam pointing stability – an important source of technical noise and loss of fidelity and (2) the optical beams can be tightly focused allowing for faster gate operations (Rabi frequency). Since our initial proposal and demonstration of photonically addressed trapped ion qubits and CMOS Paul traps with integrated SPAD readout[10], several groups have made impressive progress. Mehta, et al. at ETH/Cornell have demonstrated multi-ion addressing with improved gate operation [11] and MIT Lincoln Lab has demonstrated fast readout of ion state using integrated SPADs [12]. Future embodiments of this platform can incorporate CMOS Paul traps, SPADs for readout, and visible waveguides in a single platform.

BRAIN-COMPUTER INTERFACES (BCI)

The platforms that have been developed for photonic integration in both high-performance computing and optically addressed quantum computing may also find a home in this most difficult of computing problems – the brain-computer interface. A closed-loop BCI couples automated detection of neural signals for behaviour with therapeutic neuromodulation [13]and has been proposed as a treatment for Parkinson's disease, epilepsy, chronic pain, and depression. The development of in vivo electrical and optical probes [14] for highly multiplexed recording and stimulation of neural circuits has been the focus of several international programs and commercial ventures. Integrated photonics may have an important role to play in both of these sensing and stimulation modalities.

Electrophysiological recording of biopotentials requires readout and processing of 1-10 mV signal levels (requiring < 10μV noise). While these signals are relatively low bandwidth, the low electrical signal levels require amplification before analysis. Analog electrical amplification for each probe signal can require 5.2 μW per pixel (excluding ADC) with a bandwidth of 27 kHz (190 pJ/bit) [15]. These links can be highly asymmetric as the temperature and power dissipation at the probe (brain) are critical whereas the machine terminal has relaxed energy and temperature requirements. Such low-energy, low-bandwidth optical links would not typically be feasible, however, recent improvements in silicon photonic technology has demonstrated low-optical insertion loss modulators that operate at mV-levels [16]. We have demonstrated photovoltaic modulators that harvest some of the energy lost due to background absorption to bootstrap a 10x voltage gain [17]. In this way, 1MHz bandwidth modulators with mV sensitivities have been realized with total (optical + electrical) power dissipation less than 8 μW per channel. Further, improvements in device and system performance may enable 10-70x reductions in total power per channel.

Acknowledgements: The author acknowledges many years of fruitful collaboration with researchers in the Physical Optics and Electronics Group at MIT, Ayar Labs, UC Berkeley and Boston University.

References

[1] Yang, T.J., Chen, Y.H., Emer, J. and Sze, V., *A method to estimate the energy consumption of deep neural networks*, Proceedings of 51st asilomar conference on signals, systems, and computers, pp. 1916-1920, 2017.
[2] Batten, C., *et al*., *Building many-core processor-to-DRAM networks with monolithic CMOS silicon photonics*, IEEE Micro, vol. *29*, no. 4, pp.8-21, 2009.
[3] Sun, C., *et al. Single-chip microprocessor that communicates directly using light,* Nature,vol. *528* no. 7583, pp.534-538, 2015.
[4] Wade, M., *et al., TeraPHY: a chiplet technology for low-power, high-bandwidth in-package optical I/O*. IEEE Micro, *40*(2), pp.63-71, 2020.
[5] O'Connor, M., Chatterjee, N., Lee, D., Wilson, J., Agrawal, A., Keckler, S.W. and Dally, W.J., *Fine-grained DRAM: Energy-efficient DRAM for extreme bandwidth systems*, Proceedings *2017* 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO),pp. 41-54, 2017.
[6] Rakowski, M., *et al*., *45nm CMOS-silicon photonics monolithic technology (45CLO) for next-generation, low power and high speed optical interconnects*. *Proceedings of Optical Fiber Communication Conference* (pp. T3H-3). Optical Society of America, 2020.
[7] Khani, M., Ghobadi, M., Alizadeh, M., Zhu, Z., Glick, M., Bergman, K., Vahdat, A., Klenk, B. and Ebrahimi, E., SiP-ML: high-bandwidth optical network interconnects for machine learning training, *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, pp. 657-675, 2021.
[8] Linke, N.M., Maslov, D., Roetteler, M., Debnath, S., Figgatt, C., Landsman, K.A., Wright, K. and Monroe, C., *Experimental comparison of two quantum computing architectures*, Proceedings of the National Academy of Sciences, vol. *114*, no. 13, pp.3305-3310, 2017.
[9] Mehta, K.K., Bruzewicz, C.D., McConnell, R., Ram, R.J., Sage, J.M. and Chiaverini, J., *Integrated optical addressing of an ion qubit*, Nature nanotechnology, *11*(12), pp.1066-1070, 2016.

[10] R. J. Ram, *CMOS Photonic Circuits for Trapped Ion Quantum Computing and Molecular Sensing*, *Proceedings of Conference on Lasers and Electro-Optics (CLEO)*, 2018, pp. 1-2, 2018.
[11] Mehta, K.K., Zhang, C., Malinowski, M., Nguyen, T.L., Stadler, M. and Home, J.P., *Integrated optical multi-ion quantum logic*, Nature, *586*(7830), pp.533-537, 2020.
[12] Reens, D., *et al*., *High-Fidelity Ion State Detection Using Trap-Integrated Avalanche Photodiodes*, *arXiv:2202.01715, 2022*.
[13] Zhang, Q., *et al. A prototype closed-loop brain–machine interface for the study and treatment of pain*, Nature Biomedical Engineering, pp.1-13, 2021.
[14] Moreaux, L.C., et al., *Integrated neurophotonics: toward dense volumetric interrogation of brain circuit activity—at depth and in real time*, Neuron, vol.*108*, no.1, pp.66-92, 2020.
[15] Musk, E., *An integrated brain-machine interface platform with thousands of channels*, Journal of medical Internet research, vol.*21*, no.10, p.e16194, 2019.
[16] de Cea, M., Wollman, E.E., Atabaki, A.H., Gray, D.J., Shaw, M.D. and Ram, R.J., *Photonic readout of superconducting nanowire single photon counting detectors*, Scientific reports, vol.*10*, no.1, pp.1-8, 2020.
[17] de Cea, M., Atabaki, A.H. and Ram, R.J., *Energy harvesting optical modulators with sub-attojoule per bit electrical energy consumption*, Nature communications, *12*(1), pp.1-9, 2021.