Accelerating artificial intelligence with silicon photonics

Nicholas C. Harris, Ryan Braid, Darius Bunandar, Jim Carr, Brad Dobbie, Carlos Dorta-Quinones, Jon Elmhurst, Martin Forsythe, Michael Gould, Shashank Gupta, Sukeshwar Kannan, Tyler Kenney, Gary Kong, Tomo Lazovich, Scott Mckenzie, Carl Ramey, Chithira Ravi, Michael Scott, John Sweeney, Ozgur Yildirim, Katrina Zhang

Lightmatter, Inc., 61 Chatham St 5th floor, Boston, Massachusetts 02109, USA

Authors listed in alphabetical order, except for the presenter: N. C. Harris.
nick@lightmatter.co

Abstract: As Moore's law and Dennard scaling come to an end, new devices and computing architectures are being explored. The development of computing hardware designed to address the rapidly growing need for computational power to accelerate artificial intelligence applications has prompted investigations into both. While silicon photonics is typically viewed as a communications platform, we discuss its application to artificial intelligence and some outstanding challenges to be addressed.

1. Introduction

Over the past few years, deep neural networks have been widely deployed to address a variety of practical problems ranging from object recognition to natural language processing [1]. Much of this progress has been underpinned by rapid advancement in computational throughput driven by Moore's Law and Dennard scaling–both of which are significantly slowing as they reach fundamental bounds [2]. At the same time that progress in the underlying hardware that powers these computations is slowing, the need for increased computational power is growing: the amount of compute required to train state-of-the-art deep neural networks has been doubling every 3-4 months [3]. To answer this challenge, special purpose hardware accelerators targeted at artificial intelligence are being developed, including digital systolic arrays [4]. While new hardware architectures can provide a boost in compute performance, continued performance advancements for a given hardware architecture are bound by the underlying device physics.

The development of alternative computing technologies, which are not bound by Dennard scaling and Moore's law, are therefore required for continued advancement of artificial intelligence applications. Towards this end, new categories of compute systems including in-memory processing circuits [5], memristor arrays [6], and photonic matrix processors [7] are being explored. Compared to electronic processors (both digital and analog), photonic matrix processors have unique properties including the potential for significantly lower compute latency, the absence of dissipative parasitic resistance and capacitance, and high clock frequency operation [7, 8].

For photonic matrix processors, there is currently a large gap between academic work and the requirements of full-scale industrial applications. Some of the most pressing issues include: integration of high-speed high-accuracy data converters for control and read out, interfacing to standard high speed electronics communications protocols, scaling the number of photonic compute units to be commensurate with that of digital architectures for deep learning, chip packaging, and the development of new algorithms and software for running neural networks on this novel compute platform. Here, we discuss these challenges and our initial work towards overcoming them as we develop high speed, high efficiency photonic artificial intelligence accelerators.

2. Photonic AI acceleration

A number of integrated, photonics-based matrix processing architectures have been proposed [7,9]. Here, we focus on programmable nanophotonic processors (PNPs) [7]. A PNP is composed of an array of Mach-Zehnder interferometers (MZIs)–each of which contains programmable phase shifters. By setting the analog voltage or current on each MZI phase shifter, the 2-dimensional unitary transformation (valid in the absence of loss) applied by the MZI can be controlled. The MZIs are arranged into a reprogrammable network to form an arbitrary linear transformation on an input vector, which is encoded in the intensity and the phase of the optical signals incident on the input of the PNP [10–12]; N^2 MZIs are required to implement an arbitrary N-dimensional linear transformation.

State-of-the-art digital compute systems for deep learning often use 2-dimensional arrays of multiply accumulate units (MACs) to implement N-dimensional linear transformations. For digital compute systems, N is not

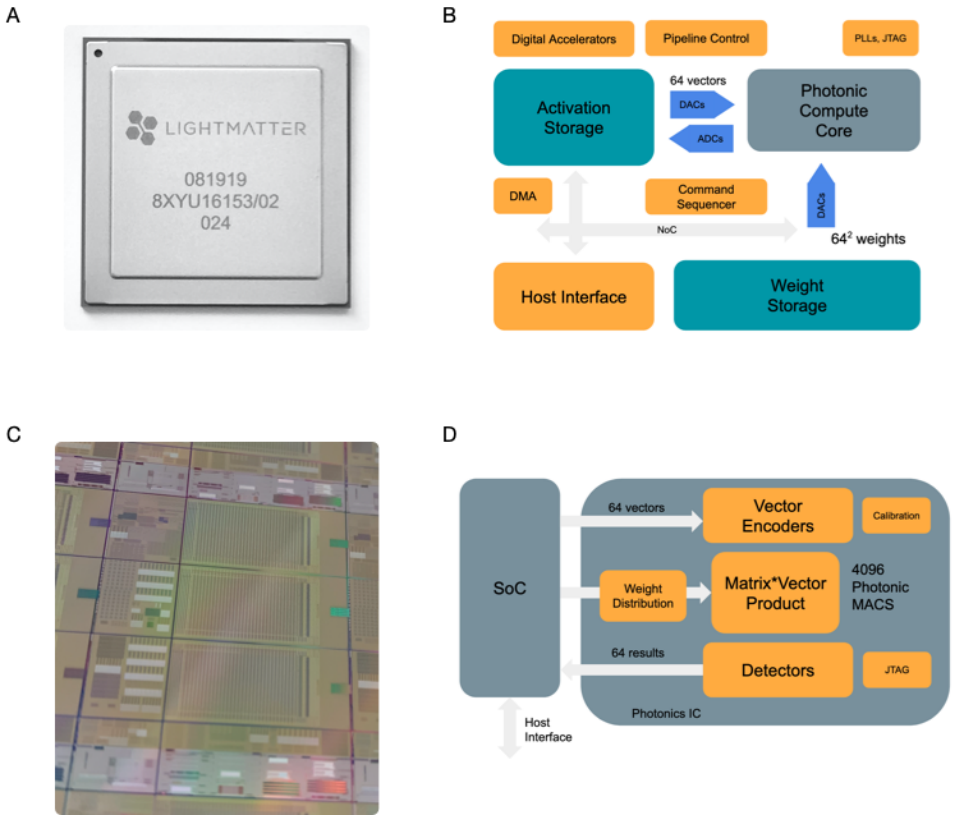

Fig. 1. (A) Photograph of an advanced 12 nm digital-analog electronic control processor. (B) Block diagram of the electronic control processor in operation with the photonic accelerator including digital compute units, pipeline control, phase locked loops (PLLs), JTAG debug interfaces, a host interface, a large static random access memory cache for neural network weights and activations, and arrays of digital-to-analog and analog-to-digital converters. (C) Micrograph of a 4096 compute element photonic accelerator. (D) Block diagram of the photonic accelerator including vector encoding modulators, a 4096 element PNP, and detectors.

principally bound by the number of MACs that can fit on a chip, but by the power budget for the chip–$N = 256$ systems have been shown [4]. For photonic system to reach this scale, significant advances in phase shifter loss and compactness are necessary (ideally while maintaining large analog bandwidths). Potential technology candidates include barium titanate [13] and nonlinear polymers [14].

In the context of computing, PNPs implement general matrix multiply (GEMM) operations, (i.e. $A \times B = C$), that are ubiquitous in deep learning algorithms by (1) applying phase settings to the MZI phase shifters that correspond to the entries of A, (2) exciting the input modes of the PNP with light of intensities and phases corresponding to the entries of a column vector from the matrix B, and (3) detecting the output light that corresponds to a column vector of the output matrix C. The steps are repeated until all vectors of the matrix B have been propagated through the PNP. Since the computation occurs entirely in the optical domain, the system can be operated at clock frequencies exceeding those typically used in large-scale electronic computing engines for machine learning [4, 15]. Machine learning algorithms are known to be robust to statistical errors and are therefore an ideal target application for analog computing technologies [16–18].

In this conference, we will discuss our work towards the development of integrated photonic processors for artificial intelligence. Fig. 1 (A) shows a photograph of a system on chip (SoC) fabricated in a 12 nanometer feature-size complimentary metal-oxide semiconductor process. As shown in Fig. 1 (B), the SoC provides control signals to each of the 64^2 photonic compute elements. In addition the SoC contains standard electronic communications interfaces to external systems through the host interface and debugging JTAG port as well as a large static random access memory cache. Data converter performance generally improves with decreasing process node [19], and so there is a general advantage to using so-called 'advanced-node' CMOS processes for implementing these circuits. Fig. 1 (C) shows a photograph of a 4096 photonic compute element PNP which is controlled by the SoC shown in Fig. 1 (A). Fig. 1 (D) shows a block-level representation of the photonic chip which consists largely of

the compute elements, vector encoder intensity modulators, and photodetectors.

3. Conclusion

We have discussed some of the challenges associated with bridging the gap between academic research into photonic matrix processors for artificial intelligence and industrial requirements, and we briefly presented our work towards addressing these issues. Continued scaling in the dimension of these systems will require the development of compact, low loss, and fast phase shifters as well as tight integration with the control electronics. The realization of such systems could enable high-performance, high-efficiency analog optical computing with programmable nanophotonic processors.

4. References

References

1. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436–444 (2015).
2. T. N. Theis and H.-S. P. Wong, "The end of moore's law: A new beginning for information technology," Comput. Sci. & Eng. **19**, 41 (2017).
3. D. Amodei, D. Hernandez, G. Sastry, J. Clark, G. Brockman, and I. Sutskever, "Ai and compute," (2019).
4. N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P.-l. Cantin, C. Chao, C. Clark, J. Coriell, M. Daley, M. Dau, J. Dean, B. Gelb, T. V. Ghaemmaghami, R. Gottipati, W. Gulland, R. Hagmann, C. R. Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, A. Jaworski, A. Kaplan, H. Khaitan, D. Killebrew, A. Koch, N. Kumar, S. Lacy, J. Laudon, J. Law, D. Le, C. Leary, Z. Liu, K. Lucke, A. Lundin, G. MacKean, A. Maggiore, M. Mahony, K. Miller, R. Nagarajan, R. Narayanaswami, R. Ni, K. Nix, T. Norrie, M. Omernick, N. Penukonda, A. Phelps, J. Ross, M. Ross, A. Salek, E. Samadiani, C. Severn, G. Sizikov, M. Snelham, J. Souter, D. Steinberg, A. Swing, M. Tan, G. Thorson, B. Tian, H. Toma, E. Tuttle, V. Vasudevan, R. Walter, W. Wang, E. Wilcox, and D. H. Yoon, "In-datacenter performance analysis of a tensor processing unit," SIGARCH Comput. Archit. News **45**, 1–12 (2017).
5. P. Chi, S. Li, C. Xu, T. Zhang, J. Zhao, Y. Liu, Y. Wang, and Y. Xie, "Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory," in *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA),* (2016), pp. 27–39.
6. M. Hu, C. E. Graves, C. Li, Y. Li, N. Ge, E. Montgomery, N. Davila, H. Jiang, R. S. Williams, J. J. Yang, Q. Xia, and J. P. Strachan, "Memristor-based analog computation and neural network classification with a dot product engine," Adv. Mater. **30**, 1705914 (2018).
7. N. C. Harris, J. Carolan, D. Bunandar, M. Prabhu, M. Hochberg, T. Baehr-Jones, M. L. Fanto, A. M. Smith, C. C. Tison, P. M. Alsing *et al.*, "Linear programmable nanophotonic processors," Optica **5**, 1623–1631 (2018).
8. M. A. Nahmias, T. F. De Lima, A. N. Tait, H. Peng, B. J. Shastri, and P. R. Prucnal, "Photonic multiply-accumulate operations for neural networks," IEEE J. Sel. Top. Quantum Electron. pp. 1–1 (2019).
9. A. N. Tait, T. F. De Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," Sci. reports **7**, 7430 (2017).
10. M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, "Experimental realization of any discrete unitary operator," Phys. Rev. Lett. **73**, 58–61 (1994).
11. W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley, "Optimal design for universal multiport interferometers," Optica **3**, 1460–1465 (2016).
12. D. A. B. Miller, "Self-configuring universal linear optical component [invited]," Photon. Res. **1**, 1–15 (2013).
13. S. Abel, F. Eltes, J. E. Ortmann, A. Messner, P. Castera, T. Wagner, D. Urbonas, A. Rosa, A. M. Gutierrez, D. Tulli *et al.*, "Large pockels effect in micro-and nanostructured barium titanate integrated on silicon," Nat. materials **18**, 42 (2019).
14. L. Alloatti, R. Palmer, S. Diebold, K. P. Pahl, B. Chen, R. Dinu, M. Fournier, J.-M. Fedeli, T. Zwick, W. Freude *et al.*, "100 ghz silicon–organic hybrid modulator," Light. Sci. & Appl. **3**, e173 (2014).
15. Graphcore, "Scalable silicon compute," in *NIPS 2017 Workshop on Deep Learning at Supercomputer Scale,* (2017).
16. R. Krishnamoorthi, "Quantizing deep convolutional networks for efficient inference: A whitepaper," arXiv preprint arXiv:1806.08342 (2018).

17. E. Fiesler, A. Choudry, and H. J. Caulfield, "Weight discretization paradigm for optical neural networks," in *Optical interconnections and networks,* , vol. 1281 (International Society for Optics and Photonics, 1990), pp. 164–173.
18. I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," The J. Mach. Learn. Res. **18**, 6869–6898 (2017).
19. B. Murmann, "A/d converter trends: Power dissipation, scaling and digitally assisted architectures," in *2008 IEEE Custom Integrated Circuits Conference,* (IEEE, 2008), pp. 105–112.