Silicon Photonics for AI Hardware

B. J. Shastri[1,2], B. A. Marquez[1], M. Filipovich[1], Z. Guo[1], E. R. Howard[1], H. Morison[1], A. N. Tait[2], T. Ferreira de Lima[2], C. Huang[2], and P. R. Prucnal[2]
[1]*Department of Physics, Engineering Physics & Astronomy, Queen's University, Kingston, ON K7L 3N6, Canada*
[2]*Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA*
shastri@ieee.org

Abstract: Neuromorphic photonic processors promise orders of magnitude improvements in both speed and energy efficiency over purely digital electronic approaches. We will provide an overview of silicon photonic systems for deep learning inference and in situ training. © 2020 The Author(s)

OCIS codes: (200.3050) Information processing; (200.4700) Optical neural systems; (250.5300) Photonic integrated circuits.

There has been a substantial increase in the use of many-core parallel processing systems for a variety of tasks in high performance computing (HPC). Artificial intelligence (AI), in particular, is growing at an alarming pace: deep learning models have been doubling in size every 3.5 months, far outpacing Moore's law [1]. These systems have much greater communication overheads than classical von Neumann architectures such as CPUs, resulting in a dramatic increase of both the area and energy consumption of metal interconnects (see, for example, Ref. [2]). They are also bottlenecked computationally by the ability to perform matrix multiplications efficiently, which represent the most common operations in HPC. The most computationally expensive task in current AI models is the implementation of neural networks. Current deep learning models require dense, low-precision matrix computations [3,4]. Digital instantiations of matrix (or tensor) units typically suffer from high communication overheads, expensive digital operations, and high latencies. On the other hand, photonic linear operations—such as passive fourier transforms [5] or matrix operations [6]—exhibit stark advantages in bandwidth density, latency, and energy.

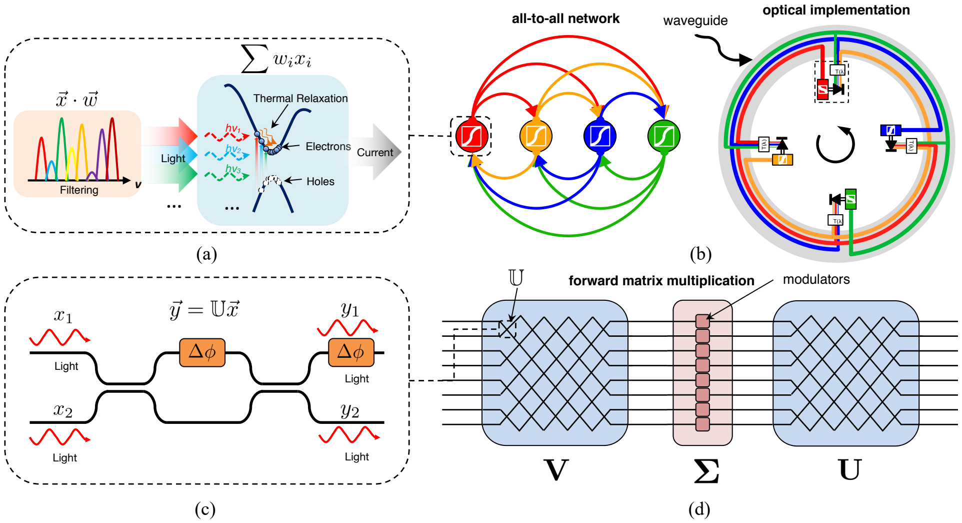

Fig. 1. Schematics for incoherent (top) [7], [8] and coherent (bottom) [9] implementations of tunable photonic multiply-accumulate operations. (a) Incoherent approaches can directly perform dot products on optically multiplexed signals. However, they rely on detectors and O/E conversion for summation. (b) The ability to multiplex allows for network flexibility, which can enable larger-scale networks with minimal waveguide usage. (c) Coherent approaches can apply a unitary rotation to incoming lightwaves. This unit can perform a tunable 2 × 2 unitary rotation denoted by U. (d) Example of scaling the system to perform a matrix operation in a feedforward topology, using a U unit at each crossing together with singular value decomposition.

Photonic technology has traditionally been used for long distance communication. However, modern bandwidth requirements and the standardization of silicon photonic integrated circuits (PICs) has led to the proliferation of shorter distance photonic links. For example, silicon photonic transceivers are now a pervasive component in datacenters. There has been growing interest in using silicon photonics as an enabling platform for unconventional computing [7], [9-13]. This could result in a new class of ultrafast information processors for neuromorphic information and signal processing, machine learning, and HPC. Neuromorphic photonics [14] can enable new class of applications where low latency, high bandwidth, and low switching energies are paramount. These applications could include nonlinear programming (e.g. solving optimization problems and partial differential equations), scientific (e.g. protein folding simulations) and high-performance computing (e.g. vector-matrix multiplications), machine learning acceleration (e.g. deep learning inference, and ultrafast and online learning), and intelligent signal processing (e.g. wideband RF signals, fiber transmission equalization, spectral mining).

In this talk, we will provide an overview of current neuromorphic silicon photonics architectures (se Fig. 1) including coherent approaches based on Mach-Zehnder interferometers [9], [13] and incoherent (multiwavelength) optoeletronic approaches based on microring resonators [7], [11], [12]. We also describe several real-world applications for control and deep learning inference. Lastly, we will discuss scalability in the context of designing a full-scale neuromorphic photonic processing system, considering aspects such as signal integrity, noise, and hardware fabrication platforms [3].

References

[1] D. Amodei and D. Hernandez, "AI and compute," Available: https:// blog.openai.com/ ai- and- compute/

[2] F. Akopyan, et al. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **34**, 1537 (2015).

[3] M. A. Nahmias, et al. *IEEE J. Sel. Top. Quantum Electron.* **26**, 7701518 (2020).

[4] V. Bangari, et al. *IEEE J. Sel. Top. Quantum Electron.* **26**, 7701213 (2020).

[5] C. Dragone, *J. Lightw. Technol.* **7**, 479 (1989).

[6] R. A. Athale and W. C. Collins, *Appl. Opt.* **21**, 2089 (1982).

[7] A. N. Tait et al. *J. Light. Technol.* **32**, 4029 (2014).

[8] L. Yang et al. *Opt. Express* **20**, 13560 (2012).

[9] Y. Shen et al. *Nat. Photonics* **11**, 441 (2017).

[10] T. Ferreira de Lima et al. *J. Light. Technol.* **37**, 1515 (2019).

[11] A. N. Tait, et al. *Sci. Rep.* **7**, 7430 (2017).

[12] A. N. Tait, et al. *Phys. Rev. Appl.* **11**, 064043 (2019).

[13] T. W. Hughes, et al. *Optica* **5**, 864 (2018).

[14] P. R. Prucnal and B. J. Shastri. *Neuromorphic Photonics* (CRC Press, 2017).