Asia Communications and Photonics Conference 2021 © Optica Publishing Group 2021

Silicon Photonics for Neuromorphic Computing and Artificial Intelligence

B. J. Shastri[1,2], C. Huang[2], A. N. Tait[1,2], 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 photonics exploit optical device physics for neuron models, and optical interconnects for distributed, parallel, and analog processing for high-bandwidth, low-latency and low switching energy applications in artificial intelligence and neuromorphic computing.** © 2021 The Author(s)

Neuromorphic (i.e., neuron-isomorphic) computing aims to bridge the gap between the energy efficiency of von Neumann computers and the human brain [1], [2]. The rise of neuromorphic computing can be attributed the widening gap between current computing capabilities and current computing needs [3], [4]. Consequently, this has spawned research into novel brain-inspired algorithms and applications uniquely suited to neuromorphic processors. These algorithms attempt to solve artificial intelligence (AI) tasks in real-time while using less energy. We posit that we can make use of the high parallelism and speed of photonics to bring the same neuromorphic algorithms to applications requiring multiple channels of multi-gigahertz analog signals, which digital processing struggles to process in real-time.

By combining the high bandwidth and parallelism of photonic devices with the adaptability and complexity attained by methods similar to those seen in the brain, photonic neural networks (PNNs) have the potential to be orders of magnitude faster than state-of-the-art electronic processors while consuming less energy per computation [5]. The goal of neuromorphic photonic processors is not to replace conventional computers, but to enable applications that are unreachable at present by conventional computing technology—those requiring low latency, high bandwidth and low energies [6], [7]. As shown in Figure 1, examples of applications for ultrafast neural networks include: 1) Enabling fundamental physics breakthroughs: qubit read-out classification, high-energy-particle collision classification, fusion reactor plasma control; 2) Nonlinear programming: solving nonlinear optimization problems (robotics, autonomous vehicles, predictive control) and partial differential equations; 3) Machine learning acceleration: vector–matrix multiplications, deep learning inference, ultrafast or online learning; 4) Intelligent signal processing: wideband radio-frequency signal processing, fibre-optic communication.

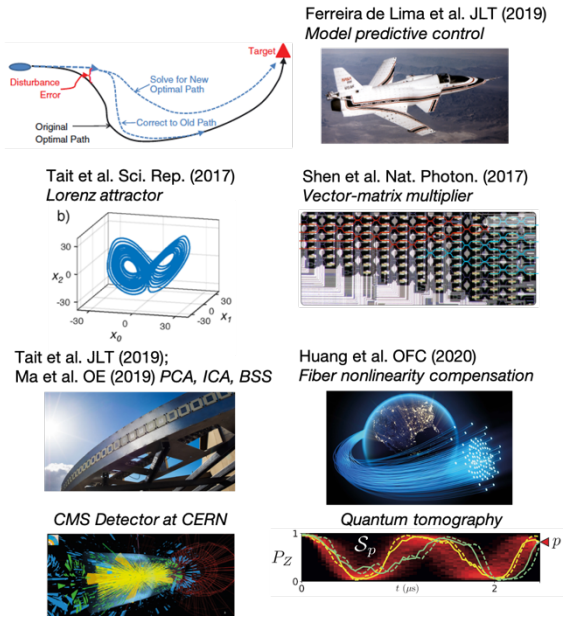

Figure 1. Applications for photonic neural networks that require low latency, high-bandwidth, and real-time processing.

Neuromorphic photonic [6], [8] approaches can be divided into two main categories (Figure 2): coherent (single wavelength) and incoherent (multiwavelength) approaches. Neuromorphic systems based on reservoir computing [9]–[11] and Mach-Zehnder interferometers [12], [13] are example of coherent approaches. In reservoir computing the predefined random weights of their hidden layers cannot be modified. An alternative approach uses silicon photonics to design fully programmable neural networks [14], [15], with a so-called broadcast-and-weight protocol [16]. In this architecture, photonic neurons output optical signals with unique wavelengths. These are multiplexed into a single waveguide and broadcast to all others, weighted, and photodetected. Each connection between a pair of neurons is configured independently by one microring resonator (MRR) weight, and the wavelength division multiplexed (WDM) carriers do not mutually interfere when detected by a single photodetector. Consequently, the physics governing the neural computation is fully analog and does not require any logic operation or sampling, which would involve serialization and sampling. Thus, they exhibit distinct, favorable trends in terms of energy dissipation, latency, crosstalk, and bandwidth when compared to electronic neuromorphic circuits [5]. The advantage of this approach over the aforementioned approaches is that it has already demonstrated fan-in, inhibition, time-resolved processing, and autaptic cascadability [15], [16].

However, the same physics also introduce new challenges, especially reconfigurability, integration, and scalability. Information carried by photons is harder to manipulate compared to electronic signals, especially nonlinear operations and memory storage [6]. Photonic neurons described here solve that problem by using optoelectronic components (O/E/O), which can be mated with standard

electronics providing reconfigurability. However, neuromorphic photonic circuits are challenging to scale up because they do not benefit from digital information, memory units and a serial processor, and therefore requires a physical unit for each element in a neural network, increasing size, area and power consumption. Here, integration costs must also be considered since the advantages of using analog photonics (high parallelism and high bandwidth) must outweigh the costs of interfacing it with digital electronics (requiring both O/E and analog/digital conversion).

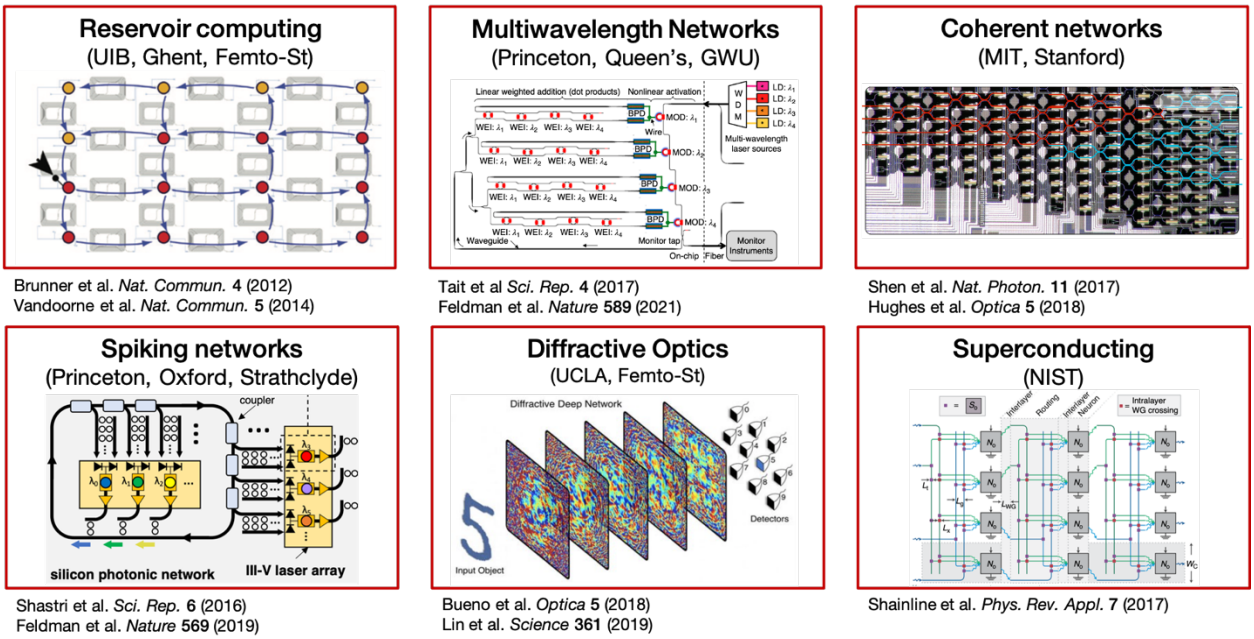

Figure 2. Examples of recently demonstrated photonic neural network architectures.

Neuromorphic photonics has reached an inflection point, benefiting from great opportunities as the world looks for alternative processor architectures. The physical limits of Dennard scaling is galvanizing the community to put forward candidates for next generation computing, from bio- to quantum computers. Photonics and in particular neuromorphic photonics are a formidable candidate for analog reconfigurable processing. We expect the development of this field to accelerate as neuroscience makes further leaps towards our understanding of the nature of cognition and artificial intelligence demands more computational resources for machine learning. As photonics technology matures and becomes more accessible to academic groups and small companies, we expect this acceleration to continue. We describe several real-world applications for control and deep learning inference [17]. 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 [6], [18].

REFERENCES

[1] Schuman, Catherine D., et al. arXiv preprint arXiv:1705.06963 (2017).
[2] Hasler, Jennifer, and Harry Bo Marr. Front. Neurosci. 7 (2013): 118.
[3] Merolla, Paul A., et al. Science 345.6197 (2014): 668-673.
[4] Davies, Mike, et al. IEEE Micro 38.1 (2018): 82-99.
[5] Ferreira De Lima, Thomas, et al. Nanophotonics 6.3 (2017): 577-599.
[6] Shastri, Bhavin J., et al." Nat. Photon. 15.2 (2021): 102-114.
[7] Ferreira De Lima, Thomas, et al. J. Lightwave Technol. 37.5 (2019): 1515-1534.
[8] Prucnal, Paul R., and Bhavin J. Shastri. Neuromorphic photonics. CRC Press, 2017.
[9] Brunner, Daniel, et al. Nat. Commun. 4.1 (2013): 1-7.
[10] Vandoorne, Kristof, et al. Nat. Commun. 5.1 (2014): 1-6.
[11] Larger, Laurent, et al. Opt. Express 20.3 (2012): 3241-3249.
[12] Shen, Yichen, et al. Nat. Photon. 11.7 (2017): 441-446.
[13] Hughes, Tyler W., et al. Optica 5.7 (2018): 864-871.
[14] Tait, Alexander N., et al. Sci. Rep. 7.1 (2017): 1-10.
[15] Tait, Alexander N., et al. Phys. Rev. Appl. 11.6 (2019): 064043.
[16] Tait, Alexander N., et al. J. Lightwave Technol. 32.21 (2014): 4029-4041.
[17] Huang, Chaoran, et al. Opt. Fiber Commun. (OFC) Conf. Postdeadline Papers (2020), paper Th4C.6
[18] Ferreira De Lima, Thomas, et al. Nanophotonics 9.13 (2020): 4055-4073.