Review article

Thomas Ferreira de Lima, Alexander N. Tait, Armin Mehrabian, Mitchell A. Nahmias, Chaoran Huang, Hsuan-Tung Peng, Bicky A. Marquez, Mario Miscuglio, Tarek El-Ghazawi, Volker J. Sorger, Bhavin J. Shastri* and Paul R. Prucnal

Primer on silicon neuromorphic photonic processors: architecture and compiler

https://doi.org/10.1515/nanoph-2020-0172
Received March 6, 2020; accepted June 10, 2020; published online August 10, 2020

Abstract: Microelectronic computers have encountered challenges in meeting all of today's demands for information processing. Meeting these demands will require the development of unconventional computers employing alternative processing models and new device physics. Neural network models have come to dominate modern machine learning algorithms, and specialized electronic hardware has been developed to implement them more efficiently. A silicon photonic integration industry promises to bring manufacturing ecosystems normally reserved for microelectronics to photonics. Photonic devices have already found simple analog signal processing niches where electronics cannot provide sufficient bandwidth and reconfigurability. In order to solve more complex information processing problems, they will have to adopt a processing model that generalizes and scales. *Neuromorphic photonics* aims to map physical models of optoelectronic systems to abstract models of neural networks. It represents a new opportunity for machine information processing on sub-nanosecond timescales, with application to mathematical programming, intelligent radio frequency signal processing, and real-time control. The strategy of neuromorphic engineering is to externalize the risk of developing computational theory alongside hardware. The strategy of remaining compatible with silicon photonics externalizes the risk of platform development. In this perspective article, we provide a rationale for a neuromorphic photonics processor, envisioning its architecture and a compiler. We also discuss how it can be interfaced with a general purpose computer, i.e. a CPU, as a coprocessor to target specific applications. This paper is intended for a wide audience and provides a roadmap for expanding research in the direction of transforming neuromorphic photonics into a viable and useful candidate for accelerating neuromorphic computing.

Keywords: neuromorphic computing; optical neural networks; photonic integrated circuits; silicon photonics; ultrafast information processing.

***Corresponding author: Bhavin J. Shastri**, Department of Physics, Engineering Physics & Astronomy, Queen's University, Kingston, ON KL7 3N6, Canada; and Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA,
E-mail: bhavin.shastri@queensu.ca. https://orcid.org/0000-0001-5040-8248
Thomas Ferreira de Lima, Alexander N. Tait, Mitchell A. Nahmias, Chaoran Huang, Hsuan-Tung Peng and Paul R. Prucnal: Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA, E-mail: tlima@princeton.edu (T. Ferreira de Lima), atait@ieee.org (A.N. Tait), mitch@luminouscomputing.com (M.A. Nahmias), chaoranh@princeton.edu (C. Huang), hpeng@princeton.edu (H.T. Peng), prucnal@princeton.edu (P.R. Prucnal)
Armin Mehrabian, Mario Miscuglio, Tarek El-Ghazawi and Volker J. Sorger: Department of Electrical and Computer Engineering, George Washington University, Washington, DC 20052, USA,
E-mail: armin@gwmail.gwu.edu (A. Mehrabian), mmiscuglio@email.gwu.edu (M. Miscuglio), tarek@gwu.edu (T. El-Ghazawi), sorger@gwu.edu (V.J. Sorger)
Bicky A. Marquez: Department of Physics, Engineering Physics & Astronomy, Queen's University, Kingston, ON KL7 3N6, Canada,
E-mail: bama@queensu.ca

1 Introduction

Computing today is in many ways the same as it was in the 1960s: digital microelectronics implementing a centralized processing unit (CPU) architecture. Throughout this time, their performance improved exponentially according to what is known as Moore's law. There have always been predictions that Moore's law is ending. Nevertheless, microelectronics have managed to maintain an exponential rate of improvement through the development of new technologies and architectures, such as multi-core architectures, graphical processing units (GPUs), and field-programmable gate arrays (FPGAs). Today, there are a growing number of computational problems that seem well out of reach, even when extrapolating for microelectronic

performance advances. Conventional computers are here to stay; however, recent years have seen a resurgence in unconventional computing approaches, ranging from neuromorphic electronics to radio frequency (RF) photonics.

Photonics does not exhibit the same physical properties of semiconductor electronics. It is unquestionable that photonics, specifically fiber optics, is preferable for high-bandwidth communication over long distances. Motivated by successes in optical communication in the 1960s and 1980s, some began to ask if optics could be used to process information[1], as opposed to only communicating information. These inquiries included fascinating strides in optical neural networks, but they ultimately proved to be mistimed. What is different today includes (1) neural networks models have come to dominate machine learning—we know these models are applicable to modern computing problems, and we know how to program them, and (2) silicon photonic integration provides an unprecedented platform to produce large-scale and low-cost photonic systems. Neuromorphic silicon photonics aims to bring the two together and to understand the impacts that both paradigms can generate on machine information processing.

This paper: (1) provides a rationale for a silicon neuromorphic photonic processor as a complement to digital microelectronic computing (e.g. as an accelerator for high-performance computing) or as an alternate platform to enable new domains of applications (e.g. nonlinear programming or wideband radio signal processing); (2) describes a vision for such a neuromorphic processor and how it can be interfaced with a general purpose computer; and (3) discusses a compiler that can be used to program such a photonic processor. We also summarize recent analysis comparing analog photonics and analog electronics.

2 Neural networks in computing

Computational modeling of neural networks has been motivated by neuroscience—the promise of understanding cognition and pathology; it has been motivated by engineering—the promise of making smarter, more efficient machines. These diverging motivations have yielded diverging research into systems that attempt to retain a faithfulness to biological neural networks and systems that do not. Nevertheless, both schools demand a consciousness of computing hardware performance. Research on the biology side has become limited by simulation tools for very large networks. On the artificial side, machine learning algorithms are taking a growing portion of data center resources.

Training Google's state-of-the-art large-scale language model BERT a single time, for example, takes 5056 GPU-hours, which corresponds to 1438 lbs of CO_2 emissions [1]. Compare that to a one-passenger carbon footprint in a New York to San Francisco flight, 1984 lbs. It is important to note that the developmental cost of these models is not included: tuning a model to a new dataset may cost dozens of times that much, and developing a new model from scratch, thousands more [1]. The short-term trend is even more worrisome. A recent analysis by OpenAI established that the computational resources required to train these large machine learning models have been doubling every 3.4 months since 2012 [2].

Neural network models have solidified as a pillar of modern machine learning under the alias "deep learning" [3]. Machine learning is experiencing rapid progress and skyrocketing demand in applications from autonomous vehicles to consumer devices to medical imaging analysis. Learning algorithms related to neural networks date to the 1950s [4], but have only in the past several years become an indispensable piece of machine learning. This return could be attributed to: (1) decisive algorithmic innovations [5, 6], (2) the Internet: an inexhaustible source of millions of training examples, and (3) new hardware, specifically graphical processing units (GPUs) [7].

It is increasingly apparent that conventional microelectronics are not optimized to implement neural network models with speed and energy efficiency that are sufficient for application demands. Artificial neural network models consist of relatively simple nonlinear nodes interconnected by a configurable linear network. This parallel, distributed structure is dissimilar to the serial nature of conventional computing architectures, and so unconventional hardware has been investigated. Unconventional architectures adopting neural network organizational principles at the hardware level were conceived of in the 1950s [8], and, like neural algorithms, have experienced a rapid and recent revival [9–12]. Modern neuromorphic electronics are used in applications ranging from battery-sipping audio recognition [13] to supercomputing resources for neuroscientists [14].

1 By "information processing," we mean operations whose purpose involves destroying or discarding informational content, for example the two-input, one-output AND gate. Contrast with "communication," where an information carrier is transported, and "signal processing," where the information carrier is enhanced (e.g. amplified, regenerated, etc.). In communication and signal processing, the information content ideally stays the same.

3 Photonic information processing

The question of what an optical computer is or should be has inspired intense debate, one could say an identity crisis. The notion of optical computing is alluring and prone to hype. An experienced retrospective is found in Ref. [15] and a recent survey is found in Ref. [16]. Contrasting philosophical views of the cast of characters in 1980s optical computing are represented in lay form in Ref. [17]. From these sources, we can perceive the tantalizing excitement and multifariousness of ideas in optical computing. Research in past decades (usually referred to as optical computing) has strongly influenced the research that has surged in recent years (usually referred to as photonic information processing). It identified ideas that work well and also pointed out pitfalls from which to learn.

3.1 Previous decades

Optical logic based on nonlinear and bistable devices was proposed in the 1960s [18], but optical devices fail to provide some fundamental properties required to make logic gates [19, 20]. In electronic gates, both input and output are represented by voltage, so the ability of one gate to drive another is taken for granted. In most nonlinear optical devices, on the other hand, the output signal must be a different wavelength than the pump input, meaning that that gate cannot drive a similar gate. Secondly, electronics have a consistent reference potential by which to define logical "0" (0 V) and "1" (1 V), so the ability of one gate to drive multiple others (i.e. fan-out) is taken for granted. When signals are represented by optical energy, they attenuate when splitting, so the definition of logical "1" changes when signals fan-out [21]. These two barriers stem from representing information as potential energy (voltage) versus a conserved energy unit (photons or current). This difference favors a non-digital architecture for light-based processing.

The second choice for optical system reconfiguration such as switching, is then to rely on electronic–photonic interactions, which are inherently inefficient given the three-orders of magnitude dimension difference of the underlying wave functions. The latter has, however, led researchers to explore to the fields of nanophotonics and plasmonics, where the light–matter interactions can be enhanced. Thus, micrometer-small opto-electronic devices have been recently demonstrated, however, often with some tradeoffs in optical signal loss [22].

Many other approaches avoid a digital computing paradigm altogether and instead target specialized tasks. At first, these included Fourier transforms [23] for pattern recognition, matrix-vector multiplication, and radar [24]. Optical technologies for interconnection have long been recognized as potential media for artificial optical neural networks (ONNs), reviewed in Ref. [25]. For the most part, approaches to ONN interconnection have focused on spatial multiplexing techniques, including configurable spatial light modulation [26], matrix grating holograms [27], and volume holograms [28, 29]. Although they are dense techniques for all-to-all interconnection, free-space and holographic devices are difficult to integrate and also require precise alignment.

Attempts to realize optical computers in previous decades encountered three major barriers. Firstly, many approaches could not be integrated, making them expensive and environmentally sensitive. Secondly, the utility of overspecialized optical "co-processors" was overtaken by simultaneous performance leaps in microelectronics. Thirdly, fundamental concepts of what makes a computer were often neglected. Concepts such as cascadability, fan-out, fan-in, and metrics are unchanging and taken for granted in conventional computers; however, they must be completely reconsidered for unconventional physical hardware.

3.2 "Optical silicon"

While optical computing has been explored in many forms, all of these approaches were hindered by a lack of a low-cost platform. In the 1986 Spectrum issue on optical computing [17], Tanguay is attributed with the phrase "optical silicon," in the sense of a low-cost, manufacturable, and versatile platform for optics. This term coined the view of many researchers that without such a platform, optical technologies, especially computers, will always encounter great difficulty competing with electronics. In 1987 coincidentally, the field acquired an idea that "Maybe 'optical silicon' is just that: silicon" [30]. Thirty years later, silicon photonics has arisen as a billion dollar industry that some believe is only just beginning to realize its full potential.

Silicon photonic foundry platforms could bring to photonics the manufacturing economies historically reserved for microelectronics. Immense resources have been poured into silicon photonics for two main reasons. Firstly, silicon photonic device sets can be manufactured with standard silicon foundries and processing

capabilities, albeit with some process modifications [31–33]. Secondly, datacenters now rely on a staggering number of short reach optical interconnects, and silicon transceivers are lower cost than their overperforming III–V counterparts. Silicon photonic platforms are still in a nascent period with many potential advantages yet to be realized [34]; however, it is reasonable to predict robust and vibrant progress in the field. New foundry lines are being created, increasing the supply; new and larger datacenters are being built, increasing the demand for short reach optical interconnects. Silicon photonics is not as simple as supply/demand: the growth of the industry has spawned an entire research ecosystem.

The barriers to conducting silicon photonics research are continually lowering with the advent of shared wafers, design tools, prototyping services, open-access libraries, and educational resources [35]. An important development in the field is the fabrication-less or "fabless" research model [36]. In the fabless design process, research groups outsource the fabrication of their designs to silicon photonics foundries. The foundries can then produce the chips at a lower cost by placing multiple designs from different groups on shared wafers. The fabless design model with foundry lines specifically tailored to silicon photonics make state-of-the-art devices and scalability accessible at a low cost and small volume.

The accessibility of silicon photonics opens the door for advanced system-level research that is academic and/or exploratory in nature [37]. The fabless design model crucially relieves the burden of early-stage research to demonstrate concrete market demand. For example, integrated RF photonics are anticipated to impact future wireless operations in the millimeter-wave bands. Because their exact impact is still uncertain, the crucial early-stage research in the field depends on the accessibility of the fabless design model.

Silicon photonics apparently meets the criteria for an imagined "optical silicon," which carries the potential to reinvigorate investigation into advanced photonic information processing.

3.3 Contemporary approaches

A scalable model of computing is necessary to make computers that leverage photonics. A variety of models have recently gained traction in the field. This article will focus on the neuromorphic model, but it is worth noting contemporary approaches to computing with light, namely *quantum silicon photonics*, *photonic reservoir computing*, and *RF photonics*. These approaches take advantage of various properties of lightwaves, respectively: their quantum coherence in low loss waveguides; their high-speed dynamics in nonlinear media; and their linearity across a high dynamic range.

3.3.1 Quantum silicon photonics

Silicon photonic systems have been proposed for quantum processing models, including quantum simulation and quantum computing [38, 39]. A universal quantum computing approach based on implanted donor qubits coupled through configurable silicon photonic circuits was proposed in Ref. [40]. The proposal has already shown favorable quantum computing potential compared to the handful of established quantum computing proposals. Silicon and silicon nitride platforms have attracted interest as high-density platforms for squeezed light sources [41, 42] for continuous-variable quantum computing [43] and boson sampling. In most of these approaches, silicon (nitride) photonics is used as a strategy to externalize platform development to foundries in a way similar to the neuromorphic contemporaries discussed below. In fact, the programmable nanophotonic processor developed for quantum transport simulations in Ref. [44] was the same chip studied in the context of deep learning in Ref. [45] and the basis for a quantum optical neural network [46]. Technological needs also overlap: for example, the need to control resonator wavelengths and large systems of configurable elements.

3.3.2 Photonic reservoir computing

Photonic reservoir computing (PRC) techniques that take inspiration from certain brain properties (e.g. analog, distributed) have received substantial recent attention from the photonics community [47–50]. These techniques are a hardware implementation of the broader *reservoir computing* concept based on recurrent neural networks. Reservoir techniques rely on supervised learning to discern a desired behavior from a large number of complex dynamics, instead of relying on establishing an isomorphism with a model. Neuromorphic and reservoir approaches differ fundamentally and possess complementary advantages. Both derive a broad repertoire of behaviors (often referred to as complexity) from a large number of physical degrees-of-freedom coupled through nonlinear interaction parameters. In a reservoir computer, the interaction

parameters do not need to be configurable, observable, or even repeatable from system-to-system.[2] Reservoirs thus possess a desireable ability to harness complexity from physical processes that are difficult to model and control, such as coupled amplifiers [51], coupled nonlinear MRRs [52], time-delayed dynamics in fibers [49], and fixed interferometric circuits [48]. Furthermore, reservoirs do not need to be fully modeled and do not require significant hardware to control the state of the reservoir. On the other hand, neuromorphic hardware has a burden of establishing an isomorphic mapping to an artificial neural network. In return, the neuromorphic system can leverage existing machine learning algorithms, map training results between simulation and hardware, and guarantee particular behaviors. Photonic reservoir computers can of course be simulated and leverage training algorithms borrowed from the reservoir computing community; however, they have no corresponding a-priori guarantee that a particular hardware instance will reproduce a simulated behavior or that training will be able to converge to this behavior.

3.3.3 RF photonics

Radio frequency (RF) front-ends must improve drastically to handle new millimeter-wave bands and advanced multi-antenna approaches. They have encountered limitations of RF electronics, analog-to-digital converters (ADCs), and digital signal processors (DSP) [53, 54]. RF photonics appear to have potential to rise to these challenges, offering improvements for programmable filtering [55], time delays [56–59], interference suppression [60, 61], and waveform generation [62, 63]. They can also reap considerable performance advantages by moving simple processing tasks from DSP into the analog subsystem [62, 64–66] which is capable of processing information at high speed and low latency. However, the main drawback of analog systems is bit resolution, and work has shown that the break-even bit-density between analog versus digital systems is about 6–7 bits [67] RF photonic circuits that can be transcribed from fiber to silicon are likely to reap the economic benefits of silicon photonic integration. In a distinct vein, large-scale programmable processors are surpassing the complexity and flexibility of processors that could be considered in fiber [68–70]. There are only so many tasks that linear models can perform. By introducing nonlinear elements, a much broader and more complex repertoire of information processing tasks can be considered.

4 Neuromorphic photonics

Neuromorphic photonics is the idea to create photonic devices and circuits whose governing equations are isomorphic to the equations describing a neural model. As mentioned in Section 2, there are two diverging motivations for the study of neuromorphic computing: one, biologically-driven, to understand cognition; and another, engineering-driven, to enhance computing. The prevalent motivation in neuromorphic photonics is the latter. Thus, the degree of neuromorphism present in photonic systems is only useful insofar as it helps with computing tasks. The hardware's adherence to neural models unlocks a wealth of metrics [71], algorithms [72, 73], tools [74, 75], and benchmarks [76] developed specifically for neural networks. A diversity of research in neuromorphic photonics are covered in the textbook "Neuromorphic Photonics" [77], which includes results and discussion on: applications of neuromorphic photonics (Chp. 1); background on silicon photonics for neuromorphic photonics processor (Chp. 3); advanced topics in silicon network implementations, topologies, function, and robustness (Chps. 8 and 11); and learning (Chp. 12). In addition, summaries of recent research in the field are presented in an encyclopedia article [78] and tutorial [79].

Neuromorphic can pertain to individual lasers and a single spiking neuron [80]. Interest in integrated lasers that are isomorphic to spiking neurons has flourished over the past seven years [81]. Experimental work in the field has so far focused on isolated neurons [82–87] and feedforward chains [88–91]. Reference [77] considers how to make cascadable neurons out of some of these non-cascadable or partially cascadable spiking lasers. Spiking neuromorphic light sources have also been implemented with superconducting electronics and all-silicon light emitting diodes [92]. Photonic isomorphism can also refer to photonic circuits and entire neural networks.

Neuromorphic silicon photonics adheres not just to a neural network model, but also to the industrial foundry platforms available today, thereby enabling their near-

2 We acknowledge some divergent opinions in drawing this contrast between neuromorphic photonics and photonic reservoir computing. A reviewer of the manuscript pointed out that some researchers consider PRC as neuromorphic because it is also inspired by the biological nervous system. This article uses neuromorphic in the sense that a mathematical isomorphism between hardware and neural network is necessary in its approach to a computational task at hand. Reservoir computing, in general, does not require a concrete establishment of isomorphism but instead relies on some amount of hardware training. In this way, we have here chosen to base the distinction on the practical implications for computing, rather than relationship to biological principles.

term construction. An architecture for integrated photonic neural networks was first proposed in 2014 [93] and demonstrated in 2017 [94]. In this broadcast-and-weight architecture, each neuron is assigned a wavelength to carry its output signal. These signals are multiplexed and broadcast to all other neurons where each is weighted by the transmission through a tunable microring resonator [95]. The weighted signals are detected, and the resulting electronic signal modulates the amplitude of that neuron's wavelength. The nonlinear transfer function of the neurons is derived from the electro-optic modulation effect. It should be noted that researchers have also showed theoretically [96] and experimentally [97] all-optical nonlinear activation function using electromagnetically induced transparency (EIT).

Since lasers are not yet widely available on silicon, modulator-class neurons were a final step to complete compatibility with silicon foundry platforms. In the authors' recent work in refs. [94, 98–100], neurons are implemented by modulators that exploit the nonlinearity of the electro-optic transfer function. Specifically, Tait et al. [98] demonstrated a modulator neuron with a silicon microring resonator (MRR) with embedded PN modulator. George et al. [100] proposed quantum well absorption modulator-based electro-optic neuron, and Amin et al. [99] proposed an indium tin oxide (ITO)-based electro-absorption modulator for photonic neural activation function. Two important figures of merit for these modulator-class neurons are the energy per bit and the capacitance of the link. In an O–E–O link, the lower the switching voltage, the lower the required transimpedance for it to work. Coupled to a low capacitance, this yields a low CV^2 switching energy and a low RC switching delay. Using photonic crystals in Indium Phosphide, Nozaki et al. [101] neared fundamental limits in classical optics with a 41 aJ/bit & 1.64 fF O–E–O link.

Modulators do not exhibit spiking dynamics, so this move represents a critical departure from the earlier work on spiking, a departure further discussed in [98, Sec. IVB]. The current wave of neuromorphic electronics, in their majority, encapsulate neural spikes in asynchronous, time-multiplexed digital packets, emulating very large spiking neural networks with modest interconnects. In photonics, it is more practical to target modest-sized, but fast networks, with direct spatial or wavelength-multiplexed interconnects. Whether in continuous-time or in spikes, information can be carried in the analog domain with low crosstalk, low dispersion, and without capacitive loading of the interconnect line.

Other neuromorphic photonic networks in silicon were proposed in 2017. A fully integrated superconducting optoelectronic network was proposed in Ref. [102] to offer unmatched energy efficiency. Communication is accomplished with single photons detected by superconducting nanowires. Superconducting electronics offer means to process [103] and amplify [104] the single-photon signal so that it can then drive an all-silicon light emitting diode [105]. While based on an exotic superconducting platform, this approach accomplishes fan-in using incoherent optical power detection in a way reminiscent of and possibly compatible with the broadcast-and-weight protocol. A programmable nanophotonic processor was studied in the context of deep learning in Ref. [45]. Weights are configured by the transmission through a coherent mesh of Mach–Zehnder interferometers. Training might be achieved by sending light through the mesh in the opposite direction [106]. Coherent optical interconnects exhibit a sensitivity to optical phase that must be re-synchronized after each layer. All-optical nonlinear devices for counteracting signal-dependent phase shifts induced by nonlinear materials are yet to be proposed, although a mesh-compatible, partially electronic, neuron was recently proposed in Ref. [107].

Neuromorphic photonics complements neuromorphic electronics in terms of application domains. Its advantages stem from bandwidth and energy efficiency, yet they will likely never compete with the low system power and large scale of neuromorphic electronics. None-withstanding, both offer commensurate power reductions compared to state-of-the-art conventional computers based on multiple cores, GPUs, and FPGAs.

At increased scale, neuromorphic photonic systems could be applied to currently unaddressable computational areas in scientific computing and RF signal processing. A key benefit of neuromorphic engineering is that existing algorithms can be leveraged. A subset of neural networks, Hopfield networks [108], have been used extensively in mathematical programming and optimization problems [72]. The ubiquity of differential equation problems in scientific computing has motivated the development of analog electronic neural emulators [109]. Predictive control algorithms were mapped to neuromorphic photonics in simulation by Ferreira de Lima et al. [79].

Of these potential application areas, the most immediate needs for real-time bandwidth lie in RF problems, in particular, commercial telecommunications, spectrum monitoring, and microwave metrology operations. Digital signal processors based on high-performance semiconductors are unable to handle increasingly complex access strategies (e.g. opportunistic) in increasingly broad bands (e.g. millimeter wave). Neural algorithms have been developed for real-time RF signal processing, including

spectral mining [110], spread spectrum channel estimation [111], and arrayed antenna control [112]. There is insistent demand to implement these tasks at wider bandwidths using less power than possible with RF electronics. As of now, RF-relevant information processing tasks of principal component analysis [113, 114] and blind source separation [115] have been demonstrated in silicon photonic weight banks.

Beyond merely improving current machine learning calculations, neuromorphic photonics could enable as-of-yet unforeseen applications in sub-nanosecond domains, for example, measurement and control for ultrafast physical phenomena. For example, Gordon [116] investigated potentials of deep neural networks to classify microwave qubit states based on observations of nanosecond transients. Developing that direction will call for the close involvement of experts in experimental physics. Application developers can apply new computer technology only if they can rely on the fact that a propositional computer will adhere to a model. That ability to investigate future applications of scaled-up networks illustrates the power of model adherence – the neuromorphic idea.

5 Neuromorphic processor architecture

Recently, in our tutorial, Ref. [79], we proposed a vision for a neuromorphic processor. We discussed how such a neuromorphic chip could potentially be interfaced with a general-purpose computer, i.e. a CPU, as a coprocessor to target specific applications. Broadly speaking, there are two levels of complexity associated with co-integrating a general-purpose electronic processor with an application-specific optical processor. Firstly, a CPU processes a series of computation instructions in an undecided amount of time and is not guaranteed to be completed. Neural networks, on the other hand, can process data in parallel and in a deterministic amount of time. CPUs have a concept of a 'fixed' instruction set on top of which computer software can be developed. However, a neuromorphic processor would require a hardware description language (HDL) because it describes the intended behavior of a hardware in real-time. Secondly, seamlessly interfacing a photonic integrated circuit with an electronic integrated circuit will take several advances in science and technology including on-chip lasers and amplifiers, co-integration of CMOS with silicon photonics, system packaging, high-bandwidth digital-to-analog converters (DAC) and analog-to-digital converters (ADCs).

We first discuss the need for a high-level processor specification that users can interface with, and then detail the architecture components required to build a neuromorphic photonic processor.

5.1 Processor firmware specifications

As highlighted in Section 4, there has been much work on photonic neural networks with different approaches (all-optical, optoelectronic etc.) in different platforms (silicon, III–V, heterogeneous integration). This is called the physical implementation (or layer) of the neural network. An abstraction above this layer is the behavioral layer which describes how information is encoded, transformed, and decoded as it flows along a network, and how the network should learn new behavior from new information.

New hardware development must be accompanied with software specification. HDLs describe circuits in a way that a computer can understand and simulate. This is useful in making the hardware (i.e. the physical layer) agnostic to the behavioral layer. Consequently, this makes programming or executing a task straightforward. For example, suppose that a particular neural network that executes an inference task can be implemented using an artificial or a spiking neural network. Both of these networks require different coding schemes, but could be used to accomplish the same task with different efficiencies and speed. It is obvious that these coding schemes require different hardware, but they also require different control algorithms and network configuration. That is why it is important to be able to express the function of the neuromorphic circuit without fixing the hardware.

This abstraction is necessary to allow integrated photonics professionals to be able to build neuromorphic processors *to spec*. It also allows them to simulate speed and power consumption before sending a chip layout for manufacture—these metrics depend not only on the performance of individual photonic devices inside a chip, but also more importantly on system tradeoff choices.

5.2 Architecture components

In all of the many existing physical implementations of photonic neural networks, every photonic device must be carefully started up and maintained while the processor is active. In particular, these devices are sensitive to temperature fluctuations and mechanical stress to a much higher degree than electronic devices. As a result, laboratory tests of these devices generally occur in temperature

controlled and stationary environments. As soon as they go out of these, their operation can be disrupted. But the relationship between performance and environmental variables are well understood by the integrated photonics community. Research devoted to this problem aims at either making devices that are more robust to environmental stress, or providing circuits that actively regulate their performance. Consequently, a neuromorphic photonic processor must be organized such that the processor core is accompanied by dedicated hardware responsible for this regulation, as well as other circuits dedicated to interfacing with the real world (Figure 1). Thus, there are two kinds of signals flowing through the processor: high-speed, analog; and low-speed, digital. The analog signals flow through the processor core, which performs a real-time neuromorphic computation, while the digital signals are intended for configuration and control of the core processor, which can happen at lower speeds.

5.2.1 Processor core

The processor core contains a reconfigurable photonic neural network, capable of taking an array of high-bandwidth optical inputs, computing a nonlinear, multivariate transformation, and producing optical outputs in real time. This core is a photonic integrated circuit (PIC) fabricated on a silicon photonics platform, for example.

A "photonic neuron" is by definition a device that can weight multiple optical signals, sum them together, and compute a nonlinear function based on the sum, in addition to having the ability to be networked together [79] in a scalable way. As we discussed in Section 4, a possible interconnection scheme is based on wavelength-division multiplexing (WDM), where all inputs to each neural layers lie in a single waveguide, separated by wavelength. This compensates for the relatively large size of photonic components relative to electronic transistors.

To get a sense of how dense it is, we estimated how many MAC operations[3] a neuromorphic photonic processor core can perform pear area. Suppose that a single broadcast waveguide loop contains 30 independent WDM channels, which can be accessed by any neuron connected to it (Sec. 4 [94, 98]). The number of MACs is proportional to how many total weights there are on chip, divided by its area. Implementing this architecture in silicon photonic foundries yield around 10^4 weights/mm^2, considering a metal routing overhead of 200%. At a conservative 1 GHz signal bandwidth, this corresponds to a 10 TMAC/s/mm$_2$ processing density with ≈30 fJ/MAC efficiency.[4] As a comparison (cf. Section 7), digital electronic neuromorphic architectures are in the 0.5 TMAC/mm^2—pJ/MAC range [86].

A fair comparison between electronic versus photonic performance metrics, however, can only be done after tallying the aggregate size and power consumption of the full processor architecture, not just the MAC computation efficiency in the processor core. This is especially true if the relevant data stream is not analog, because there are involved costs in digital-to-analog conversion. The processor PIC can have internal or external laser sources, depending on the manufacture platform, and electrical or optical I/O, which incur extra costs in generating this continuous wave light source. It also needs thermal management and isolation from other heat-sourcing circuits, which could negatively affect the normal operation of the chip. One way or another it needs to be controlled by an

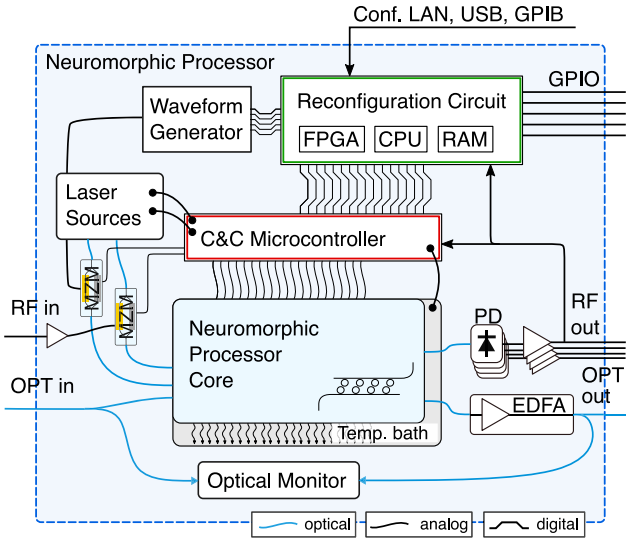

Figure 1: Simplified schematics of a neuromorphic processor. Thanks to integrated laser sources and photodetectors, it can input and output RF signals directly as an option to optically-modulated signals. The waveform generator allows for programming arbitrary stimulus that can be used as part of a machine learning task. The abbreviations are explained in "Reconfiguration Circuit" (Section 5.2). Reproduced from [79].

3 Multiply-accumulate operations. See footnote 9.

4 The efficiency value was estimated by the authors by dividing a typical silicon photonic modulator's switching energy (1 pJ/bit) by the number of maximum weights per neuron (30 MAC/bitperiod) in the latest generation of our silicon photonic neuromorphic chip. This number only takes into account the dynamic power consumption of neuron circuit, ignoring sources of static power, including weight-locking and laser pump inefficiencies. Table 2 compares these metrics for many technologies at their limit.

electronic circuit, which we refer to here as a *Command & Control* circuit.

5.2.2 Command & control circuit

As mentioned, the *Command & Control* circuit, implemented in a standard electronic platform, corrects the fabrication variations, regulates the PIC against thermal fluctuations, and protects it against over-voltage damage. In other words, it ensures that the processor core is well calibrated and run at peak performance at all times.

Its main function is to translate a weight matrix, digitally loaded to its running memory, into a set of analog control signals responsible for locking the microring weight mechanisms in the PIC. It does that by synthesizing information from external laser parameters such as wavelength, and local optical power monitors and temperature sensors embedded on the chip. This control scheme, usually based on locally heating silicon waveguides, has been thoroughly demonstrated in silicon photonics [117, 118]. This technique has been perfected to perform principal component analysis [119] and independent component analysis [114], which rely in precise multivariate weighting of high-bandwidth analog signals. It is worth noting that current research on phase-change materials can enable photonic non-volatile memory, which will simplify these weight setting mechanisms significantly [120]. They would effectively reduce the static power consumption of weight locking to a negligible amount compared to heater approaches.

This micro-controller has a very high analog DC I/O count to control each and every weight unit in the PIC, and a high-throughput digital interface with a *reconfiguration* circuit. Circuits based on this design should be able to reprogram about 10,000 wt per millisecond. The reconfiguration circuit is the highest-level sub-processor of the neuromorphic processor, and it is the low-bandwidth interface to the host computer and the real world.

5.2.3 Reconfiguration circuit – interfacing with the real world

To illustrate the function of each circuit, take for example the Model Predictive Control (MPC) task, introduced in Ref. [79]. In the MPC task, the controller must solve a quadratic optimization problem at each time step. This problem can be mapped to a recurrent neural network, whose steady-state solution corresponds to the solution of the optimization problem, and whose weights correspond to the control law. The neural network can be implemented by a neuromorphic processor. The processor core is doing the real-time computations in the analog domain, without the need for high-speed analog-digital-analog conversion. It acts as the controller in the system. However, its transfer function cannot be insulated from the "plant"[5], because different control rules might be in place depending on the conditions of the plant, such as temperature conditions, time-of-day, humidity or even human-made policy decisions.

The reconfiguration circuit (Figure 1) is used to regulate the high-speed controller based on input data from the plant's sensors. It receives instructions from a CPU, live-data from the environment and the state of the command and control (C&C) circuit and makes decisions about how the network is to be configured in real-time. The entire processor must be designed to account for that, gathering as much information as possible from the environment and from onboard sensors. As a result, it is crucial to maintain a high-bandwidth communication link with a computer motherboard, represented as GPIO in Figure 1. It is best implemented with a combination of interconnected FPGA, CPU, and RAM modules[6].

The reconfiguration circuit marks the boundary between the photonic engineers and the digital hardware programmers. Therefore, it must be the one that receives the instructions (synthesized and assembled from an HDL program) and takes care of not only configuring the core processor but also handling training, online learning, and digital and analog interconnects. It is at this stage that one could envision a neuromorphic compiler for this novel photonic hardware, capable of translating a neural network-compatible task to a program that runs on the processor, breaking down how to configure neuron biases, weight matrices, optical power levels, amplification gain, radio-frequency filtering, learning rules etc.

6 Envisioning a photonic compiler

One of the major challenges in neuromorphic photonic design flow is the lack of sufficient tools between the high-level and low-level abstractions. While it took decades for digital electronics to gradually bridge the abstraction gaps, photonics and more specifically neuromorphic photonics can adopt many of such experiences and methodologies. Neuromorphic computers are divergent from their conventional von Neumann counter-parts in three major ways.

5 Plant as defined in control theory.

6 FPGA: Field-programmable gate array. CPU: Central processing unit. RAM: Random-access memory. GPIO: General-Purpose Input/Output. These are common modules in modern digital hardware.

First, the co-localized processor and memory scheme, which brings the memory and processor closer together. A byproduct of this is that the programs move closer to processors. Second, inspired by the human brain, processing elements in neuromorphic computers are inherently heterogeneous. This heterogeneity elevates various functionalities while preserving locality. Third, programming through learning, rather than explicit algorithms, is an important differentiating factor from conventional programming paradigms and lead to robust systems, or possibly even improved performance [121].

Our perspective is that a design flow for neuromorphic processors, and neuromorphic photonics in particular, should be cognizant of hardware, even from early *specification* stages. This can be achieved by including accurate behavioral models into the training and initial inference tests. Figure 2 shows an envisioned design flow for neuromorphic photonics. Particular to neuromorphic systems, top layers accommodate for steps necessary to translate an application statement into a neural network graph with a trained set of weights. In this section, we briefly review some of the reported design and simulation methodologies and tools in photonics and how they fit within the design flow of neuromorphic photonics. Table 1 compares various photonic simulation tools in terms of their level of abstraction and whether they provide neuromorphic design support. Then, we review examples of recent efforts that aim to simulate modern photonic neural networks.

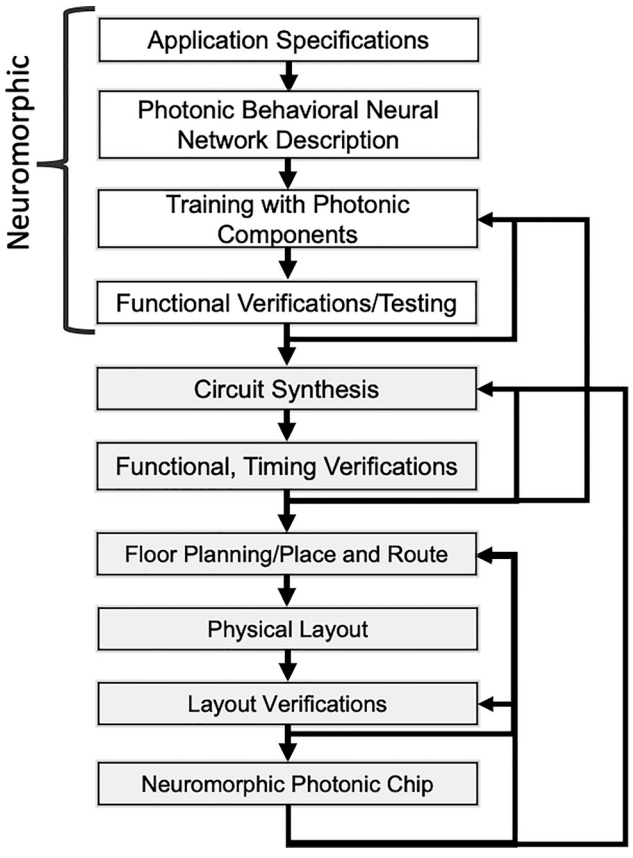

Figure 2: Envisioned design flow for neuromorphic photonics system design. At the top of the design flow stack, neuromorphic-specific steps account for accurate translation of application statement into trained set of weights.

Traditionally, electromagnetic simulation workflows are used to design and optimize nanophotonic devices by methods such as finite difference time domain (FDTD) or beam propagation method (BPM), which simulate the propagation of the electromagnetic wave in a particular geometry. While FDTD and similar methods can be optimized and efficient for simulating single devices or circuits with few components, they are not scalable for large circuits or system-level designs. Some of the more recent efforts are focused on creating a facilitated design environment for photonic components from behavioral device specifications. These specifications are commonly laid out through a scripting language such as SKILLS, AMPLE, TCL, or Python. This helps to not only shorten the design time, but also standardize the design process [122–124].

At the circuit level, photonic platforms allow for more complex functionalities that would not be feasible to directly expand from device level approaches. Hence, some of the more coarse-grain features need to be abstracted away into aggregate and behavioral representations in the form of input-output relations. The majority of the current photonic circuit design tools enable designers to schematically connect photonic building blocks and translate them into photonic circuits (e.g. Lumerical [122]). Caphe [127] creates a closed simulation loop that generates circuit-level models from electromagnetic simulation and refines those models based on actual measurements from taped out circuits. These tools commonly offer parameterized modules with attached layouts and API-like interfaces to automate some parts of the process, without providing full customization capabilities. In neuromorphic computing, processing units are inherently heterogeneous and hierarchical. This highlights the necessity of circuit design, simulation, and layout tools that enable hierarchical flows equipped with a comprehensive library of basic neuromorphic processing circuits.

With the success of artificial neural networks (ANNs) over the past decade, academia and industry have further introduced novel conceptual neural network architectures and applying them to existing problems. However, the bulk of these efforts are undertaken with digital electronic hardware being the primary computing platform with

precise digital number representations and safeguarding mechanisms such as error detection and error correction. Even mixed-signal neuromorphic platforms of interest with analog cores [133, 134] in large part rely on offline training with algorithms and methods with high bit-precision representations and computationally near-ideal transfer functions. It should be noted that the assumption behind using trained weights for inference is that the inference hardware should retain the integrity of connections, weights, and their dynamics in the case of spiking neural networks.

Currently, at the top level neuromorphic processing software stack, neural network platforms such as Tensorflow and Pytorch carry out the translation of application-level specifications to computational graphs and their corresponding trained weights. However, these weights are trained without knowledge of the inference hardware. In digital electronics, this disparity is usually manifested when the bit-resolution of the inference hardware is different from that of the trained weights [135]. In analog neuromorphic circuits and photonics in particular, by mapping arithmetic functions to non-ideal transfer functions of the photonic components, the integrity of trained weights is challenged. In addition, unlike digital circuits, even small amounts of noise can skew numerical representations. Having said that, we envision a high-level hardware-aware layer in the photonic neuromorphic software stack, which can natively simulate and execute an artificial neural network. One effective approach to realize such layer is to conduct training and inference by including behavioral models of photonic components early on in the neural network selection and training process.

The two more established neuromorphic photonic networking architectures (coherent and broadcast-and-weight interconnects, Section 7) have inspired a few photonic-aware simulation tools. In Ref. [107], a machine learning and photonic simulation framework based on the unitary matrix multiplications compatible with coherent Mach–Zehnder meshes [45] (Neurophox[7]). In Ref. [131], the authors introduced a simulator tool to investigate the prediction accuracy of a convolutional neural network (CNN) executed on a broadcast-and-weight architecture. Behavioral models of photonic components were used to perform the CNN operations including the convolution. In Ref. [132] a similar photonic neural network design methodology is proposed (Photonflow[8]). The idea behind this methodology is to extend existing and familiar neural network tools, i.e. Tensorflow, with behavioral and performance models of photonic components.

The advantage of the latter approach, as opposed to building a fully home-grown tool is three fold. First, the familiar interface allows the large group of users to benefit from extended models with minimal effort. Secondly, many of their auxiliary computational modules such as *optimizers* can be adopted to simulate photonic circuits, not limited to neural networks, within the same framework. For instance, in Ref. [129] the neural network framework, PyTorch, is leveraged for time and frequency simulation and optimization of photonic circuits. Lastly, this approach enables designers to preserve many of the low level software capabilities available in commonly adopted neural network tools such as distributed processing and device placement on a variety of backend processors such as CPUs, GPUs, and TPUs. Figure 3 depicts how Tensorflow toolkit hierarchy with libraries is extended by photonic models. As an example, it can be seen that the base multiply operation can be extended to incorporate behavioral models of photodiodes, MRR weights, and modulators. In addition, noise as in the analog circuits can be modeled and simulated in neural networks via this approach.

7 Discussion on photonic and electronic approaches

There are currently two major bottlenecks in the energy efficiency of artificial intelligence accelerators: data movement, and the performance of multiply-accumulate (MAC) operations[9], or matrix multiplications. Light is an established communication medium, and has traditionally

7 Project in development at: https://github.com/solgaardlab/neurophox.

8 Project in development at: https://github.com/openhpclgw/photonflow.

9 The multiply-accumulate (MAC) operation calculates the product of two numbers and adds the result to an accumulator. For a given accumulation variable a and modified state a', the operation takes the following form: $a' \leftarrow a + (w \times x)$. MACs are constituents of a number of linear mathematical operations, including dot products, matrix multiplication, function evaluation, Fourier transforms, and convolutions. MACs have traditionally characterized the performance signal processing (DSP) applications [136, 137], but have become increasingly prominent in modern HPC. MACs constitute the largest bottleneck in a variety of computing problems. For example, linear systems that can be cast in terms of matrix-vector multiplication operations can be represented as MACs. In many benchmarking tables for AI hardware, TOPs (Tera-operations per second) are used instead of MACs. Its precise definition is manufacturer specific, but a rough conversion of 1 TOPs = 2 TMAC/s can be used (since one MAC involves two operations).

Table 1: Comparison of various photonic tools versus their scope.

Tool	Scope			
	Device simulation	**Circuit modelling**	**System simulation**	**Neural network**[a]
Lumerical [122]	✓	✓	✗	✗
OmniSim [125]	✓	✗	✗	✗
Synopsys [126]	✓	✓	✗	✗
Caphe [127]	✓	✓	✗	✗
VPIphotonics [128]	✓	✓	✗	✗
PhotonTorch [129]	✗	✓	✗	✗
QNET [130]	✗	✓	✗	✗
Neurophox[b] [107]	✗	✗	✓	✓
DEAP CNN [131]	✗	✗	✓	✓
PhotonFlow[c] [132]	✗	✗	✓	✓

[a] *Neural Network* column indicates whether the tool can natively design and simulate a modern neural network.
[b] Footnote: neurophox.
[c] Footnote: photonflow.

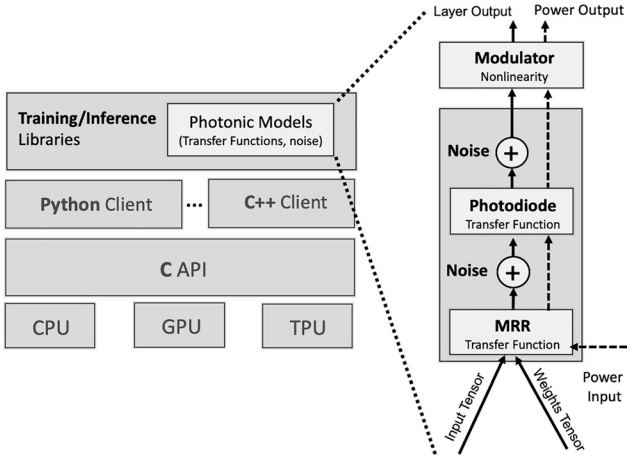

Figure 3: Software illustration of the extended tensorflow toolkit to handle WDM-based photonic systems.

been used to address data movement on a larger scale. As photonic links are scaled smaller and some of their practical problems addressed, photonic devices have the potential to address both of these bottlenecks on-chip simultaneously. Such photonic systems have been proposed in various configurations to accelerate neural network operations (see [45, 94, 138]). However, their main advantage comes from addressing MAC operations directly. Here, we will look at the advantages of a simple matrix vector multiplication (MVM) unit made of integrated photonic components, in which inputs and outputs are encoded as light signals, and analog matrix multiplications are performed using a passive optical array.

Possible instantiations of photonic MVMs are shown in Figure 4. Power or phase can be used to encode information, while wavelength or phase selectivity can be used to program the network into a desired configuration. Wavelength division multiplexing (WDM) can further increasing the compute density of the approach. Classic examples include arrays of resonator weight banks [94, 95, 139] or Mach–Zehnder interferometers [45]. The most important metrics are *energy efficiency (energy/MAC), throughput per unit area i.e. compute density (MACs/s/mm^2), speed (MVM/s), and latency (s)*, where both speed and latency are measured across an entire matrix-vector (MVM) operation. In CMOS, MVM operations are typically instantiated using systolic arrays [140] or SIMD units [141], although there are some other architectures that use aspects of both [142]. Digital systems are limited by the use of many transistors to represent simple operations and require machinery to coordinate the data movement involved in both weights and activations. Table 2 provides a comparison between already demonstrated digital and analog electronic implementations, and recently proposed photonic approaches including a possible future platform with subwavelength photonics. Note: the compute densities for the digital system include the overall architecture (core and periphery) while for the rest of the architectures the density is computed with respect to the core(s) only based on available literature. Also, the numbers for the photonic architectures are optimistic and based on extrapolating the observed performance metrics for a handful of neurons. Attaining these values requires solving a number of practical problems which are possible to address in the short term. These are discussed below.

The largest bottleneck in efficient photonic MVM operations is the use of heaters for coarse tuning. Typically, the thermo-optic coefficient (dn/dT) is the strongest effect in most materials of interest (i.e., silicon), leading to heavy use of heaters in almost any tunable passive photonic system. There are several ways these can be eradicated, via the use of post-fabrication trimming [147, 148] or devices with an enhanced electro-optic coefficient or carrier depletion/injection (dn/dT, dα/dT) such that heaters are not as necessary [149, 150]. The second largest problem is fabrication variation, which can result in parameter drifts for devices in an array. Resonators, for example, are highly sensitive to such variation, particularly across a wafer. This can also be remedied by enhancing the electro-optic coefficient of devices and some other tricks (see [151, 152] for resonators). Third, the signal-to-noise ratio of the output

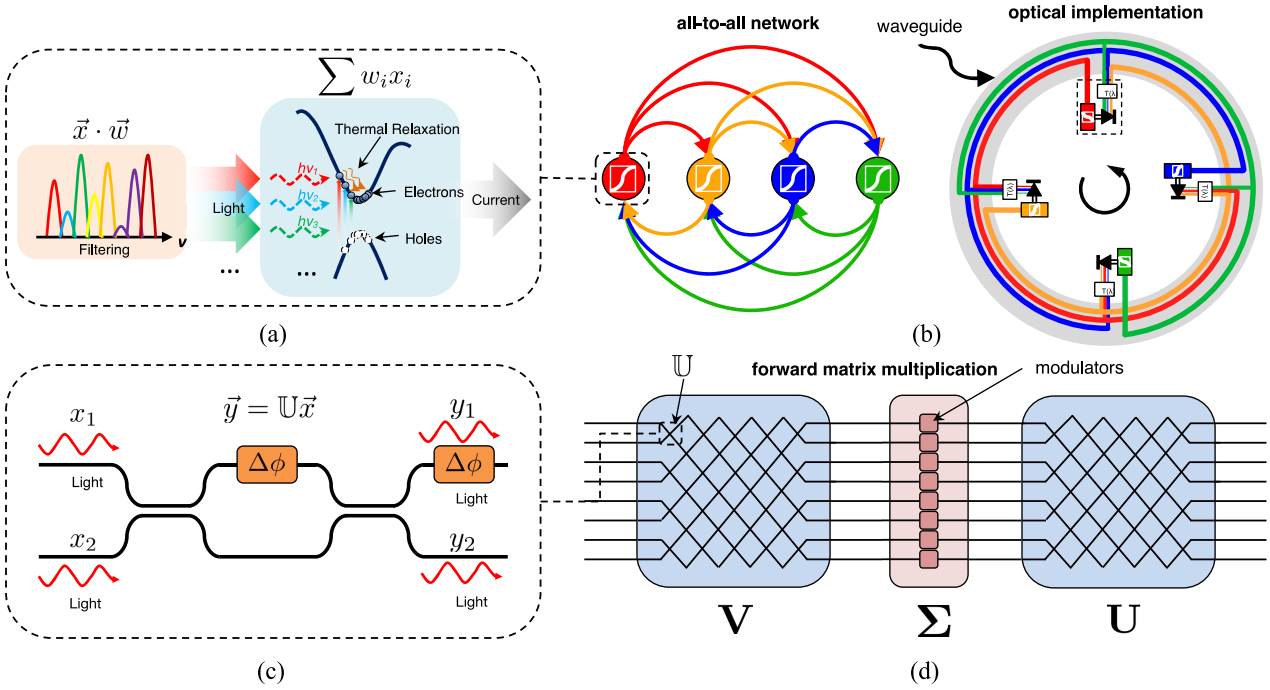

Figure 4: Schematics for incoherent (top) [93, 139] and coherent (bottom) [45] implementations of tunable photonic multiply-accumulate operations. (a) Incoherent approaches can directly perform dot products on optically multiplexed signals. However, they rely on detectors and O/E conversion for summation. (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 $\mathbb{U}$. (d) Example of scaling the system to perform a matrix operation in a feedforward topology, using a $\mathbb{U}$ unit at each crossing together with singular value decomposition. Reproduced from Ref. [143].

Table 2: Comparison of various photonic hardware approaches with a well-known deep learning accelerator during mean operating conditions. Adapted from [143].

Technology	Energy/MAC	Compute density	Vector size	Precision	Latency/speed[a]
Google TPU (digital) [140]	0.43 pJ/MAC[b]	580 GMACs/s/mm^2	256	8 bits	2 µs/1.42 ns
Flash (Analog Sim.) [144, 145]	7 fJ/MAC	18 TMACs/s/mm^2	100	5 bits	15 ns
Coherent Mach–Zehnder interferometer mesh [45]	30 fJ/MAC	0.56 TMACs/s/mm^2	100	8 bits	<100 ps
Hybrid laser NN [80, 86]	0.22 pJ/MAC	4.5 TMACs/s/mm^2	56	5.1+ bits	<100 ps
Silicon photonic broadcast-and-weight NN [98, 146]	2.1 fJ/MAC	50 TMACs/s/mm^2	148	5.1+ bits	<100 ps
Sub-λ nanophotonics (prediction) [143]	30.6 aJ/MAC	5 PMAC/s/mm^2	300	5.1+ bits	<50 ps

[a] Latency is defined as the time between a single matrix multiplication operation at the given vector size. Note: density is computed with respect to the core(s) only (except for the Google TPU).
[b] total power consumption of the chip including the core and periphery. For a fair comparison between the rest of the entries, which only include the core(s), this number would be two orders of magnitude smaller (better) i.e. around 10 fJ/MAC.

must be optimized by reducing the intrinsic loss of photonic components together with the noise on the receiver. There are a variety of technologies that can address this—for example, lasers can be coupled on-chip with <1 dB of loss [153], photonic devices in state-of-the-art silicon foundries can be designed with low scattering [154], while detectors such as avalanche photodiodes [155], can reduce the relative contribution of thermal noise to the signal at the receiver.

Photonic arrays ultimately have very similar limits to analog electronic crossbar arrays, as analyzed in Ref. [143]: single-digit aJ/ MAC efficiencies, and 100 s of PMACs/s/mm^2 compute densities. However, photonic MVMs garner an advantage for larger MVM units, both in the size of the matrix and in the physical footprint of the core. Generally speaking, optimized photonic systems tend to perform worse than their electrical counterparts for smaller arrays (where intra-chip distances are approximately below

100 μm), but perform better for larger arrays (above 100 μm) [143].

In that sense, photonic MVM arrays have a similar profile to photonic communication channels, with better performance over larger distances. However, photonic systems tend to have worse signal-to-noise ratios, as a result of several factors: (1) photonic channels are ultimately shot noise limited, which is more than an order of magnitude greater than the thermal noise limits on resistors [143], and (2) to achieve similar compute densities to electronics, photonic MVMs must run faster to compensate for their larger device sizes, and noise is speed dependent. That being said, there are some architectural options to reduce this issue—for example, optical unitary operations [45] can conserve the variance of the input and output signals. This is in contrast to other approaches such as resistive crossbar arrays where the strength of the signal detected is proportional to $\sqrt{N}$ given by the standard deviation of the signal resulting from an $N \times N$ matrix operation (assuming independent inputs). [156].

Although photonic arrays exhibit some fundamental advantages over analog electronics (particularly for large matrix sizes or large physical sizes), a more important question is whether or not photonics arrays are practical. Thankfully, the transceiver industry has created a silicon photonic ecosystem fully compatible with high volume manufacturing (HVM). Compared to CMOS chips, photonics has costlier packaging, largely because light generation cannot be done easily in silicon—in fact, the cost of a production photonic chip is dominated by packaging. In addition, the tools required for the design and testing of large-scale photonic systems (>10k components) are still in early development—analog photonic systems must grapple with the challenge of addressing yield, variability, precision, and tunability. Nonetheless, the total cost to produce a photonic chip package at high volume is dipping below one hundred dollars, and it is expected that the trend will continue [157]. The orders of magnitude advantages offered by photonics, and its potential for HVM scalability, makes it a viable inroad for the breakneck performance and innovation required by artificial intelligence algorithms in the years to come.

Funding: This article is supported by National Science Foundation (NSF) (Award numbers 1740262, 1740235, 1642991); SRC nCore (Award number 2018-C-A); Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants Program; Canadian Foundation of Innovation (CFI) John R. Evans Fund (JELF); Ontario Research Fund: Small Infrastructure Program.

References

[1] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, vol 1. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 3645–3650 [Online]. Available at: https://www.aclweb.org/anthology/P19-1355.

[2] D. Amodei, D. Hernandez, G. Sastry, J. Clark, G. Brockman, and I. Sutskever, "Ai and compute," 2019 [Online]. Available at: https://openai.com/blog/ai-and-compute/#addendum.

[3] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *J. Big Data*, vol. 2, no. 1, p. 1, 2015. [Online].

[4] J. Von Neumann, "Probabilistic logics and the synthesis of reliable organisms from unreliable components," *Autom. Stud.*, vol. 34, pp. 43–98, 1956 [Online]. Available at: http://fab.cba.mit.edu/classes/862.16/notes/computation/vonNeumann-1956.pdf.

[5] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. [Online].

[7] V. K. Pallipuram, M. Bhuiyan, and M. C. Smith, "A comparative study of GPU programming models and architectures using neural networks," *J. Supercomput.*, vol. 61, no. 3, pp. 673–718, 2012. [Online].

[8] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychol. Rev.*, vol. 65, no. 6, pp. 386–408, 1958 [Online]. Available at: http://dl.acm.org/citation.cfm?id=65669.104386.

[9] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, et al., "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, 2014 [Online]. Available at: http://www.sciencemag.org/content/345/6197/668.full.pdf.

[10] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The SpiNNaker project," *Proc. IEEE*, vol. 102, no. 5, pp. 652–665, 2014.

[11] K. Boahen, "Neurogrid: emulating a million neurons in the cortex," in *IEEE International Conference of the Engineering in Medicine and Biology Society*, 2006.

[12] J. Schemmel, D. Briiderle, A. Griibl, M. Hock, K. Meier, and S. Millner, "A wafer-scale neuromorphic hardware system for large-scale neural modeling," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, 2010, pp. 1947–1950.

[13] W. Y. Tsai, D. Barch, A. Cassidy, et al., "Always-on speech recognition using TrueNorth, a reconfigurable, neurosynaptic processor," *IEEE Trans. Comput.*, vol. 66, no. 6, pp. 996–1007, 2016.

[14] A. Mundy, J. Knight, T. Stewart, and S. Furber, "An efficient SpiNNaker implementation of the neural engineering framework," in *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–8.

[15] H. J. Caulfield, "Perspectives in optical computing," *Computer*, vol. 31, no. 2, pp. 22–25, 1998.

[16] P. Ambs, "Optical computing: a 60-year adventure," *Adv. Opt. Technol.*, vol. 2010, pp. 22–25, 2010 [Online]. https://doi.org/10.1155/2010/372652.
[17] T. E. Bell, "Optical computing: a field in flux," *IEEE Spectr.*, vol. 23, no. 8, pp. 34–57, 1986.
[18] O. A. Reimann and W. F. Kosonocky, "Progress in optical computer research," *IEEE Spectr.*, vol. 2, no. 3, pp. 181–195, 1965.
[19] R. W. Keyes, "What makes a good computer device?," *Science*, vol. 230, no. 4722, pp. 138–144, 1985 [Online]. Available at: http://science.sciencemag.org/content/230/4722/138.
[20] R. W. Keyes, "Optical logic-in the light of computer technology," *Opt. Acta Int. J. Opt.*, vol. 32, no. 5, pp. 525–535, 1985.
[21] J. W. Goodman, "Fan-in and fan-out with optical interconnections," *Opt. Acta Int. J. Opt.*, vol. 32, no. 12, pp. 1489–1496, 1985 [Online].
[22] R. Amin, R. Maiti, Y. Gui, et al., *Broadband sub-λ ghz ito plasmonic Mach–Zehnder modulator on silicon photonics*, 2019.
[23] J. W. Goodman, *Introduction to Fourier optics*. San Francisco, USA: McGraw-Hill, 1968.
[24] L. J. Cutrona, E. N. Leith, L. J. Porcello, and W. E. Vivian, "On the application of coherent optical processing techniques to synthetic-aperture radar," *Proc. IEEE*, vol. 54, no. 8, pp. 1026–1032, 1966.
[25] J. Misra and I. Saha, "Artificial neural networks in hardware: a survey of two decades of progress," *Neurocomputing*, vol. 74, no. 1–3, pp. 239–255, 2010 [Online]. Available at: http://www.sciencedirect.com/science/article/pii/S092523121000216X.
[26] E. C. Mos, J. J. H. B. Schleipen, H. de Waardt, and D. G. D. Khoe, "Loop mirror laser neural network with a fast liquid-crystal display," *Appl. Opt.*, vol. 38, no. 20, pp. 4359–4368, 1999 [Online]. Available at: http://ao.osa.org/abstract.cfm?URI=ao-38-20-4359.
[27] S. L. Yeh, R. C. Lo, and C. Y. Shi, "Optical implementation of the Hopfield neural network with matrix gratings," *Appl. Opt.*, vol. 43, no. 4, pp. 858–865, 2004 [Online]. Available at: http://ao.osa.org/abstract.cfm?URI=ao-43-4-858.
[28] P. Asthana, G. P. Nordin, J. Armand, R. Tanguay, and B. K. Jenkins, "Analysis of weighted fan-out/fan-in volume holographic optical interconnections," *Appl. Opt.*, vol. 32, no. 8, pp. 1441–1469, 1993 [Online]. Available at: http://ao.osa.org/abstract.cfm?URI=ao-32-8-1441.
[29] J. Shamir, H. J. Caulfield, and R. B. Johnson, "Massive holographic interconnection networks and their limitations," *Appl. Opt.*, vol. 28, no. 2, pp. 311–324, 1989 [Online]. Available at: http://ao.osa.org/abstract.cfm?URI=ao-28-2-311.
[30] R. Soref and B. Bennett, "Electrooptical effects in silicon," *IEEE J. Quantum Electron.*, vol. 23, no. 1, pp. 123–129, 1987.
[31] W. Bogaerts, R. Baets, P. Dumon, et al., "Nanophotonic waveguides in silicon-on-insulator fabricated with cmos technology," *J. Lightwave Technol.*, vol. 23, no. 1, pp. 401–412, 2005.
[32] R. G. Beausoleil, "Large-scale integrated photonics for high-performance interconnects," *J. Emerg. Technol. Comput. Syst.*, vol. 7, no. 2, pp. 6:1–6:54, Jul. 2011 [Online].
[33] W. Bogaerts, M. Fiers, and P. Dumon, "Design challenges in silicon photonics," *IEEE J. Sel. Top. Quantum Electron.*, vol. 20, no. 4, pp. 1–8, 2014.
[34] T. Baehr-Jones, T. Pinguet, P. Lo Guo-Qiang, S. Danziger, D. Prather, and M. Hochberg, "Myths and rumours of silicon photonics," *Nat. Photon.*, vol. 6, no. 4, pp. 206–208, 2012 [Online].
[35] L. Chrostowski and M. Hochberg, *Silicon Photonics Design: From Devices to Systems*. Cambridge, UK, Cambridge University Press, 2015.
[36] A.-J. Lim, J. Song, Q. Fang, et al., "Review of silicon photonics foundry efforts," *IEEE J. Sel. Top. Quantum Electron.*, vol. 20, no. 4, pp. 405–416, 2014.
[37] M. Hochberg, N. C. Harris, R. Ding, et al., "Silicon photonics: the next fabless semiconductor industry," *IEEE Solid State Circuits Mag.*, vol. 5, no. 1, pp. 48–58, 2013.
[38] T. Rudolph, *Why I am optimistic about the silicon-photonic route to quantum computing*, 2016 [Online]. Available at: http://arxiv.org/abs/1607.08535.
[39] J. W. Silverstone, D. Bonneau, J. L. O'Brien, and M. G. Thompson, "Silicon quantum photonics," *IEEE J. Sel. Top. Quantum Electron.*, vol. 22, no. 6, pp. 390–402, 2016.
[40] K. J. Morse, R. J. S. Abraham, A. DeAbreu, et al., "A photonic platform for donor spin qubits in silicon," *Sci. Adv.*, vol. 3, no. 7, 2017 [Online]. Available at: http://advances.sciencemag.org/content/3/7/e1700930, https://doi.org/10.1126/sciadv.1700930.
[41] Z. Vernon, N. Quesada, M. Liscidini, et al., "Scalable squeezed-light source for continuous-variable quantum sampling," *Phys. Rev. Appl.*, vol. 12, p. 064024, 2019.
[42] V. D. Vaidya, B. Morrison, L. G. Helt, et al., *Broadband Quadrature-Squeezed Vacuum and Nonclassical Photon Number Correlations from a Nanophotonic Device*, arXiv eprint, arXiv: 1904.07833, 2020.
[43] C. Weedbrook, S. Pirandola, R. García-Patrón, et al., "Gaussian quantum information," *Rev. Mod. Phys.*, vol. 84, pp. 621–669, 2012 [Online].
[44] N. C. Harris, G. R. Steinbrecher, M. Prabhu, et al., "Quantum transport simulations in a programmable nanophotonic processor," *Nat. Photon.*, vol. 11, p. 447 EP –, 06 2017 [Online].
[45] Y. Shen, N. C. Harris, S. Skirlo, et al., "Deep learning with coherent nanophotonic circuits," *Nat. Photon.*, vol. 11, no. 7, pp. 441–446, 2017 [Online].
[46] G. R. Steinbrecher, J. P. Olson, D. Englund, and J. Carolan, "Quantum optical neural networks," *NPJ Quantum Inf.*, vol. 5, no. 1, p. 60, 2019 [Online].
[47] D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer, "Parallel photonic information processing at gigabyte per second data rates using transient states," *Nat. Commun.*, vol. 4, p. 1364, 2013 [Online].
[48] K. Vandoorne, P. Mechet, T. Van Vaerenbergh, et al., "Experimental demonstration of reservoir computing on a silicon photonics chip," *Nat. Commun.*, vol. 5, p. 3541, 2014 [Online].
[49] M. C. Soriano, D. Brunner, M. Escalona-Morán, C. R. Mirasso, and I. Fischer, "Minimal approach to neuro-inspired information processing," *Front. Comput. Neurosci.*, vol. 9, p. 68, 2015 [Online].
[50] F. Duport, A. Smerieri, A. Akrout, M. Haelterman, and S. Massar, "Fully analogue photonic reservoir computer," *Sci. Rep.*, vol. 6, pp. 22–381 EP –, 03 2016 [Online].
[51] K. Vandoorne, W. Dierckx, B. Schrauwen, et al., "Toward optical signal processing using photonic reservoir computing," *Opt. Express*, vol. 16, no. 15, pp. 11182–11192, 2008 [Online]. Available at: http://www.opticsexpress.org/abstract.cfm?URI=oe-16-15-11182.

[52] C. Mesaritakis, V. Papataxiarhis, and D. Syvridis, "Micro ring resonators as building blocks for an all-optical high-speed reservoir-computing bit-pattern-recognition system," *J. Opt. Soc. Am. B*, vol. 30, no. 11, pp. 3048–3055, 2013 [Online]. Available at: http://josab.osa.org/abstract.cfm?URI=josab-30-11-3048.

[53] J. Capmany, J. Mora, I. Gasulla, J. Sancho, J. Lloret, and S. Sales, "Microwave photonic signal processing," *J. Lightwave Technol.*, vol. 31, no. 4, pp. 571–586, 2013.

[54] A. Farsaei, Y. Wang, R. Molavi, et al., "A review of wireless-photonic systems: Design methodologies and topologies, constraints, challenges, and innovations in electronics and photonics," *Opt. Commun.*, 2016 [Online]. Available at: http://www.sciencedirect.com/science/article/pii/S0030401816300748, https://doi.org/10.1364/iprsn.2016.jtu4a.14.

[55] W. Liu, M. Li, R. S. Guzzon, et al., "A fully reconfigurable photonic integrated signal processor," *Nat. Photon.*, vol. 10, no. 3, pp. 190–195, 2016.

[56] Y. Liu, A. Choudhary, D. Marpaung, and B. J. Eggleton, "Gigahertz optical tuning of an on-chip radio frequency photonic delay line," *Optica*, vol. 4, no. 4, pp. 418–423, 2017 [Online]. Available at: http://www.osapublishing.org/optica/abstract.cfm?URI=optica-4-4-418.

[57] M. Burla, L. R. Cortés, M. Li, X. Wang, L. Chrostowski, and J. Azaña, "On-chip ultra-wideband microwave photonic phase shifter and true time delay line based on a single phase-shifted waveguide Bragg grating," in *2013 International Topical Meeting on Microwave Photonics (MWP)*, 2013, pp. 92–95.

[58] S. Khan and S. Fathpour, "Demonstration of complementary apodized cascaded grating waveguides for tunable optical delay lines," *Opt. Lett.*, vol. 38, no. 19, pp. 3914–3917, 2013 [Online]. Available at: http://ol.osa.org/abstract.cfm?URI=ol-38-19-3914.

[59] J. Cardenas, M. A. Foster, N. Sherwood-Droz, et al., "Wide-bandwidth continuously tunable optical delay line using silicon microring resonators," *Opt. Express*, vol. 18, no. 25, pp. 26525–26534, 2010 [Online]. Available at: http://www.opticsexpress.org/abstract.cfm?URI=oe-18-25-26525.

[60] Y. Liu, D. Marpaung, A. Choudhary, and B. J. Eggleton, "Lossless and high-resolution RF photonic notch filter," *Opt. Lett.*, vol. 41, no. 22, pp. 5306–5309, 2016 [Online]. Available at: http://ol.osa.org/abstract.cfm?URI=ol-41-22-5306.

[61] M. P. Chang, E. C. Blow, J. J. Sun, M. Z. Lu, and P. R. Prucnal, "Integrated microwave photonic circuit for self-interference cancellation," *IEEE Trans. Microw. Theory Tech.*, vol. 65, no. 11, pp. 1–9, 2017.

[62] M. H. Khan, H. Shen, Y. Xuan, et al., "Ultrabroad-bandwidth arbitrary radiofrequency waveform generation with a silicon photonic chip-based spectral shaper," *Nat. Photon.*, vol. 4, no. 2, pp. 117–122, 2010.

[63] A. M. Weiner, "Ultrafast optical pulse shaping: a tutorial review," *Opt. Commun.*, vol. 284, no. 15, pp. 3669–3692, 2011 [Online]. Available at: http://www.sciencedirect.com/science/article/pii/S0030401811003750.

[64] J. Chang, J. Meister, and P. R. Prucnal, "Implementing a novel highly scalable adaptive photonic beamformer using "blind" guided accelerated random search," *J. Lightwave Technol.*, vol. 32, no. 20, pp. 3623–3629, 2014.

[65] T. Ferreira de Lima, A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Scalable wideband principal component analysis via microwave photonics," *IEEE Photon. J.*, vol. 8, no. 2, pp. 1–9, 2016.

[66] A. N. Tait and P. R. Prucnal, "Applications of wavelength-fan-in for high-performance distributed processing systems," in *2014 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, 2014, pp. 177–178.

[67] Y. Choukroun, E. Kravchik, F. Yang, and P. Kisilev, "Low-bit quantization of neural networks for efficient inference," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3009–3018.

[68] D. Pérez, I. Gasulla, L. Crudgington, et al., "Multipurpose silicon photonics signal processor core," *Nat. Commun.*, vol. 8, no. 1, p. 636, 2017.

[69] D. Pérez, I. Gasulla, and J. Capmany, "Field-programmable photonic arrays," *Opt. Express*, vol. 26, no. 21, pp. 27265–27278, 2018 [Online]. Available at: http://www.opticsexpress.org/abstract.cfm?URI=oe-26-21-27265.

[70] L. Zhuang, C. G. H. Roeloffzen, M. Hoekman, K.-J. Boller, and A. J. Lowery, "Programmable photonic signal processor chip for radiofrequency applications," *Optica*, vol. 2, no. 10, pp. 854–859, 2015 [Online]. Available at: http://www.osapublishing.org/optica/abstract.cfm?URI=optica-2-10-854.

[71] J. Hasler and H. B. Marr, "Finding a roadmap to achieve large neuromorphic hardware systems," *Front. Neurosci.*, vol. 7, no. 118, 2013, https://doi.org/10.3389/fnins.2013.00118.

[72] U.-P. Wen, K.-M. Lan, and H.-S. Shih, "A review of Hopfield neural networks for solving mathematical programming problems," *Eur. J. Oper. Res.*, vol. 198, no. 3, pp. 675–687, 2009 [Online]. Available at: http://www.sciencedirect.com/science/article/pii/S0377221708009788.

[73] T. Lee and F. Theunissen, "A single microphone noise reduction algorithm based on the detection and reconstruction of spectro-temporal features," *Proc. R. Soc. Lond. A Math. Phys. Eng. Sci.*, vol. 471, no. 2184, p. 20150309, 2015 [Online]. Available at: http://rspa.royalsocietypublishing.org/content/471/2184/20150309, https://doi.org/10.1098/rspa.2015.0309.

[74] C. Eliasmith and C. H. Anderson, *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*. Cambridge, USA, MIT Press, 2004.

[75] F. Donnarumma, R. Prevete, A. de Giorgio, G. Montone, and G. Pezzulo, "Learning programs is better than learning dynamics: a programmable neural network hierarchical architecture in a multi-task scenario," *Adapt. Behav.*, vol. 24, no. 1, pp. 27–51, 2016 [Online]. Available at: http://adb.sagepub.com/content/24/1/27.abstract.

[76] A. Diamond, T. Nowotny, and M. Schmuker, "Comparing neuromorphic solutions in action: implementing a bio-inspired solution to a benchmark classification task on three parallel-computing platforms," *Front. Neurosci.*, vol. 9, no. 491, p. 118, 2016 [Online]. Available at: http://www.frontiersin.org/neuromorphic_engineering/10.3389/fnins.2015.00491/abstract.

[77] P. R. Prucnal and B. J. Shastri, *Neuromorphic Photonics*. Boca Raton, FL: CRC Press, 2017.

[78] B. J. Shastri, A. N. Tait, T. F. de Lima, M. A. Nahmias, H.-T. Peng, and P. R. Prucnal, *Principles of Neuromorphic Photonics*, 2018 [Online]. Available at: http://arxiv.org/abs/1801.00016.

[79] T. Ferreira de Lima, H. Peng, A. N. Tait, et al., "Machine learning with neuromorphic photonics," *J. Lightwave Technol.*, vol. 37, no. 5, pp. 1515–1534, 2019.

[80] M. A. Nahmias, B. J. Shastri, A. N. Tait, and P. R. Prucnal, "A leaky integrate-and-fire laser neuron for ultrafast cognitive

computing," *IEEE J. Sel. Top. Quantum Electron.*, vol. 19, no. 5, pp. 1–12, 2013.

[81] P. R. Prucnal, B. J. Shastri, T. Ferreira de Lima, M. A. Nahmias, and A. N. Tait, "Recent progress in semiconductor excitable lasers for photonic spike processing," *Adv. Opt. Photon.*, vol. 8, no. 2, pp. 228–299, 2016 [Online]. Available at: http://aop.osa.org/abstract.cfm?URI=aop-8-2-228.

[82] A. Hurtado, K. Schires, I. Henning, and M. Adams, "Investigation of vertical cavity surface emitting laser dynamics for neuromorphic photonic systems," *Appl. Phys. Lett.*, vol. 100, no. 10, p. 103703, 2012.

[83] F. Selmi, R. Braive, G. Beaudoin, I. Sagnes, R. Kuszelewicz, and S. Barbay, "Relative refractory period in an excitable semiconductor laser," *Phys. Rev. Lett.*, vol. 112, p. 183902, 2014 [Online].

[84] B. Romeira, R. Avó, J. M. L. Figueiredo, S. Barland, and J. Javaloyes, "Regenerative memory in time-delayed neuromorphic photonic resonators," *Sci. Rep.*, vol. 6, p. 19510, 2016 [Online].

[85] M. A. Nahmias, A. N. Tait, L. Tolias, et al., "An integrated analog O/E/O link for multi-channel laser neurons," *Appl. Phys. Lett.*, vol. 108, no. 15, p. 151106, 2016 [Online]. Available at: http://scitation.aip.org/content/aip/journal/apl/108/15/10.1063/1.4945368.

[86] H. T. Peng, M. A. Nahmias, T. F. de Lima, A. N. Tait, B. J. Shastri, and P. Prucnal, "Neuromorphic photonic integrated circuits," *IEEE J. Sel. Top. Quantum Electron.*, vol. 24, no. 6, pp. 1–15, 2018.

[87] H. Peng, G. Angelatos, T. F. de Lima, et al., "Temporal information processing with an integrated laser neuron," *IEEE J. Sel. Top. Quantum Electron.*, vol. 26, no. 1, pp. 1–9, 2020.

[88] T. V. Vaerenbergh, M. Fiers, P. Mechet, et al., "Cascadable excitability in microrings," *Opt. Express*, vol. 20, no. 18, pp. 20292–20308, 2012 [Online]. Available at: http://www.opticsexpress.org/abstract.cfm?URI=oe-20-18-20292.

[89] B. J. Shastri, M. A. Nahmias, A. N. Tait, A. W. Rodriguez, B. Wu, and P. R. Prucnal, "Spike processing with a graphene excitable laser," *Sci. Rep.*, vol. 6, p. 19126, Jan. 2016 [Online].

[90] T. Deng, J. Robertson, and A. Hurtado, "Controlled propagation of spiking dynamics in vertical-cavity surface-emitting lasers: towards neuromorphic photonic networks," *IEEE J. Sel. Top. Quantum Electron.*, vol. 23, no. 6, pp. 1–8, 2017.

[91] S. Y. Xiang, H. Zhang, X. X. Guo, et al., "Cascadable neuron-like spiking dynamics in coupled vcsels subject to orthogonally polarized optical pulse injection," *IEEE J. Sel. Top. Quantum Electron.*, vol. 23, no. 6, pp. 1–7, 2017.

[92] J. M. Shainline, S. M. Buckley, A. N. McCaughan, et al., "Superconducting optoelectronic loop neurons," *J. Appl. Phys.*, vol. 126, no. 4, p. 044902, 2019 [Online].

[93] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Broadcast and weight: an integrated network for scalable photonic spike processing," *J. Lightwave Technol.*, vol. 32, no. 21, pp. 4029–4041, 2014.

[94] A. N. Tait, T. F. de Lima, E. Zhou, et al., "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, vol. 7, no. 1, p. 7430, 2017 [Online].

[95] A. N. Tait, A. X. Wu, T. Ferreira de Lima, et al., "Microring weight banks," *IEEE J. Sel. Top. Quantum Electron.*, vol. 22, no. 6, pp. 2276–2279, 2016.

[96] M. Miscuglio, A. Mehrabian, Z. Hu, et al., "All-optical nonlinear activation function for photonic neural networks invited," *Opt. Mater. Express*, vol. 8, no. 12, pp. 3851–3863, 2018 [Online]. Available at: http://www.osapublishing.org/ome/abstract.cfm?URI=ome-8-12-3851.

[97] Y. Zuo, B. Li, Y. Zhao, et al., "All-optical neural network with nonlinear activation functions," *Optica*, vol. 6, no. 9, pp. 1132–1137, 2019 [Online]. Available at: http://www.osapublishing.org/optica/abstract.cfm?URI=optica-6-9-1132.

[98] A. N. Tait, T. Ferreira de Lima, M. A. Nahmias, et al., "Silicon photonic modulator neuron," *Phys. Rev. Appl.*, vol. 11, no. 6, p. 064043, 2019 [Online]. https://doi.org/10.1103/PhysRevApplied.11.064043.

[99] R. Amin, J. K. George, S. Sun, et al., "Ito-based electro-absorption modulator for photonic neural activation function," *APL Mater.*, vol. 7, no. 8, p. 081112, 2019 [Online]. Available at:

[100] J. K. George, A. Mehrabian, R. Amin, et al., "Neuromorphic photonics with electro-absorption modulators," *Opt. Express*, vol. 27, no. 4, pp. 5181–5191, 2019 [Online]. Available at: http://www.opticsexpress.org/abstract.cfm?URI=oe-27-4-5181.

[101] K. Nozaki, S. Matsuo, T. Fujii, et al., "Femtofarad optoelectronic integration demonstrating energy-saving signal conversion and nonlinear functions," *Nat. Photon.*, vol. 13, no. 7, pp. 454–459, 2019 [Online].

[102] J. M. Shainline, S. M. Buckley, R. P. Mirin, and S. W. Nam, "Superconducting optoelectronic circuits for neuromorphic computing," *Phys. Rev. Appl.*, vol. 7, p. 034013, 2017 [Online].

[103] J. M. Shainline, "Fluxonic processing of photonic synapse events," *IEEE J. Sel. Top. Quantum Electron.*, vol. 26, no. 1, pp. 1–15, 2020.

[104] A. N. McCaughan, V. B. Verma, S. M. Buckley, et al., "A superconducting thermal switch with ultrahigh impedance for interfacing superconductors to semiconductors," *Nat. Electron.*, vol. 2, no. 10, pp. 451–456, 2019 [Online].

[105] S. Buckley, J. Chiles, A. N. McCaughan, et al., "All-silicon light-emitting diodes waveguide-integrated with superconducting single-photon detectors," *Appl. Phys. Lett.*, vol. 111, no. 14, p. 141101, 2017 [Online].

[106] T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, "Training of photonic neural networks through in situ backpropagation and gradient measurement," *Optica*, vol. 5, no. 7, pp. 864–871, 2018 [Online]. Available at: http://www.osapublishing.org/optica/abstract.cfm?URI=optica-5-7-864.

[107] I. A. D. Williamson, T. W. Hughes, M. Minkov, B. Bartlett, S. Pai, and S. Fan, "Reprogrammable electro-optic nonlinear activation functions for optical neural networks," *IEEE J. Sel. Top. Quantum Electron.*, vol. 26, no. 1, pp. 1–12, 2020.

[108] J. J. Hopfield and D. W. Tank, ""Neural" computation of decisions in optimization problems," *Biol. Cybern.*, vol. 52, no. 3, pp. 141–152, 1985 [Online].

[109] T. Roska, L. Chua, D. Wolf, T. Kozek, R. Tetzlaff, and F. Puffer, "Simulating nonlinear waves and partial differential equations via CNN. I. Basic techniques," *IEEE Trans. Circuits Syst. I Fundam. Theory Appl.*, vol. 42, no. 10, pp. 807–815, 1995.

[110] V. K. Tumuluru, P. Wang, and D. Niyato, "A neural network based spectrum prediction scheme for cognitive radio," in *2010 IEEE International Conference on Communications (ICC)*, 2010, pp. 1–5.

[111] U. Mitra and H. V. Poor, "Neural network techniques for adaptive multiuser demodulation," *IEEE J. Sel. Areas Commun.*, vol. 12, no. 9, pp. 1460–1470, 1994.

[112] K.-L. Du, A. Lai, K. Cheng, and M. Swamy, "Neural methods for antenna array signal processing: a review," *Sig. Process.*, vol. 82, no. 4, pp. 547–561, 2002 [Online]. Available at: http://www.sciencedirect.com/science/article/pii/S0165168401001852.

[113] A. N. Tait, P. Y. Ma, T. F. de Lima, et al., "Demonstration of multivariate photonics: blind dimensionality reduction with integrated photonics," *J. Lightwave Technol.*, vol. 37, no. 24, pp. 5996–6006, 2019.

[114] P. Y. Ma, A. N. Tait, T. F. de Lima, C. Huang, B. J. Shastri, and P. R. Prucnal, "Photonic independent component analysis using an on-chip microring weight bank," *Opt. Express*, vol. 28, no. 2, pp. 1827–1844, 2020 [Online]. Available at: http://www.opticsexpress.org/abstract.cfm?URI=oe-28-2-1827.

[115] A. N. Tait, T. F. de Lima, P. Y. Ma, M. P. Chang, M. A. Nahmias, B. J. Shastri, P. Mittal, and P. R. Prucnal, "Blind source separation in the physical layer," in *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, 2018, pp. 1–6.

[116] E. Gordon, "Design and control of a photonic neural network applied to high-bandwidth classification," Undergraduate Thesis, Princeton University, 2017.

[117] A. N. Tait, T. Ferreira de Lima, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Multi-channel control for microring weight banks," *Opt. Express*, vol. 24, no. 8, pp. 8895–8906, 2016 [Online]. Available at: http://www.opticsexpress.org/abstract.cfm?URI=oe-24-8-8895.

[118] A. N. Tait, H. Jayatilleka, T. F. D. Lima, P. Y. Ma, M. A. Nahmias, B. J. Shastri, et al., "Feedback control for microring weight banks," *Opt. Express*, vol. 26, no. 20, pp. 26422–26443, 2018 [Online]. Available at: http://www.opticsexpress.org/abstract.cfm?URI=oe-26-20-26422.

[119] P. Y. Ma, A. N. Tait, T. F. de Lima, S. Abbaslou, B. J. Shastri, and P. R. Prucnal, "Photonic principal component analysis using an on-chip microring weight bank," *Opt. Express*, vol. 27, no. 13, pp. 18329–18342, 2019 [Online]. Available at: http://www.opticsexpress.org/abstract.cfm?URI=oe-27-13-18329.

[120] C. Ríos, N. Youngblood, Z. Cheng, et al., "In-memory computing on a photonic platform," *Sci. Adv.*, vol. 5, no. 2, pp. 1–10, 2019.

[121] G. Indiveri and S.-C. Liu, "Memory and information processing in neuromorphic systems," *Proc. IEEE*, vol. 103, no. 8, pp. 1379–1397, 2015.

[122] Lumerical solutions." [Online]. Available at: https://www.lumerical.com/.

[123] L. Alloatti, M. Wade, V. Stojanovic, M. Popovic, and R. J. Ram, "Photonics design tool for advanced cmos nodes," *IET Optoelectron.*, vol. 9, no. 4, pp. 163–167, 2015.

[124] W. Bogaerts and L. Chrostowski, "Silicon photonics circuit design: methods, tools and challenges," *Laser Photon. Rev.*, vol. 12, no. 4, p. 1700237, 2018.

[125] Omnisim omni-directional photonic simulations." [Online]. Available at: https://www.photond.com/products/omnisim.htm.

[126] E. Ghillino, E. Virgillito, P. V. Mena, et al., "The synopsys software environment to design and simulate photonic integrated circuits: a case study for 400g transmission," in *2018 20th International Conference on Transparent Optical Networks (ICTON)*. IEEE, 2018, pp. 1–4.

[127] M. Fiers, T. Van Vaerenbergh, J. Dambre, and P. Bienstman, "Caphe: time-domain and frequency-domain modeling of nonlinear optical components," in *Integrated Photonics Research, Silicon and Nanophotonics*. Optical Society of America, 2012, p. IM2B-3.

[128] VPI photonics." [Online]. Available at: https://www.vpiphotonics.com.

[129] F. Laporte, J. Dambre, and P. Bienstman, "Highly parallel simulation and optimization of photonic circuits in time and frequency domain based on the deep-learning framework pytorch," *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, 2019.

[130] N. Tezak, A. Niederberger, D. S. Pavlichin, G. Sarma, and H. Mabuchi, "Specification of photonic circuits using quantum hardware description language," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 370, no. 1979, pp. 5270–5290, 2012.

[131] V. Bangari, B. A. Marquez, H. Miller, et al., "Digital electronics and analog photonics for convolutional neural networks (DEAP-CNNS)," *IEEE J. Sel. Top. Quantum Electron.*, vol. 26, no. 1, pp. 1–13, 2019.

[132] A. Mehrabian, M. Miscuglio, Y. Alkabani, V. J. Sorger, and T. El-Ghazawi, "A winograd-based integrated photonics accelerator for convolutional neural networks," *IEEE J. Sel. Top. Quantum Electron.*, vol. 26, no. 1, pp. 1–12, 2019 [Online]. Available at: https://github.com/openhpclgw/photonflow.git.

[133] B. V. Benjamin, P. Gao, E. McQuinn, et al., "A mixed-analog-digital multichip system for large-scale neural simulations," *Proc. IEEE*, vol. 102, no. 5, pp. 699–716, 2014.

[134] F. Akopyan, J. Sawada, A. Cassidy, et al., "Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip," *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.*, vol. 34, no. 10, pp. 1537–1557, 2015.

[135] A. Rodriguez, E. Segal, E. Meiri, et al., "Lower numerical precision deep learning inference and training," *Intel White Paper*, vol. 3, 2018.

[136] S. W. Smith, et al., *The scientist and engineer's guide to digital signal processing*, 1997.

[137] G. Frantz, "Digital signal processor trends," *IEEE Micro*, vol. 20, no. 6, pp. 52–59, 2000.

[138] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, "Large-scale optical neural networks based on photoelectric multiplication," *Phys. Rev. X*, vol. 9, p. 021032, 2019 [Online].

[139] L. Yang, R. Ji, L. Zhang, J. Ding, and Q. Xu, "On-chip CMOS-compatible optical signal processor," *Opt. Express*, vol. 20, no. 12, pp. 13560–13565, 2012 [Online]. Available at: http://www.opticsexpress.org/abstract.cfm?URI=oe-20-12-13560.

[140] N. P. Jouppi, C. Young, N. Patil, et al., In-datacenter performance analysis of a tensor processing unit, arXiv:1704.04760, 2017 [Online]. Available at: http://arxiv.org/abs/1704.04760.

[141] J. Fowers, K. Ovtcharov, M. Papamichael, et al., "A configurable cloud-scale dnn processor for real-time AI," in *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, 2018, pp. 1–14.

[142] C. Nicol, "A coarse grain reconfigurable array (CGRA) for statically scheduled data flow computing," Wave Comput. White Paper, 2017.

[143] M. A. Nahmias, T. F. D. 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.*, vol. 26, no. 1, pp. 1–18, 2019.

[144] M. Bavandpour, M. R. Mahmoodi, and D. B. Strukov, "Energy-efficient time-domain vector-by-matrix multiplier for neurocomputing and beyond," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 66, no. 9, pp. 1512–1516, 2019.

[145] M. R. Mahmoodi and D. Strukov, "An ultra-low energy internally analog, externally digital vector-matrix multiplier based on nor flash memory technology," in *2018 55th ACM/ESDA/IEEE design automation conference (DAC)*, 2018, pp. 1–6.

[146] A. N. Tait, "Silicon photonic neural networks," Ph.D. dissertation, Princeton University, 2018 [Online]. Available at: https://dataspace.princeton.edu/jspui/handle/88435/dsp01vh53wz43k.

[147] J. Schrauwen, D. V. Thourhout, and R. Baets, "Trimming of silicon ring resonator by electron beam induced compaction and strain," *Opt. Express*, vol. 16, no. 6, pp. 3738–3743, 2008 [Online]. Available at: http://www.opticsexpress.org/abstract.cfm?URI=oe-16-6-3738.

[148] A. H. Atabaki, A. A. Eftekhar, M. Askari, and A. Adibi, "Accurate post-fabrication trimming of ultra-compact resonators on silicon," *Opt. Express*, vol. 21, no. 12, pp. 14139–14145, 2013 [Online]. Available at: http://www.opticsexpress.org/abstract.cfm?URI=oe-21-12-14139.

[149] S. S. Djordjevic, K. Shang, B. Guan, et al., "CMOS-compatible, athermal silicon ring modulators clad with titanium dioxide," *Opt. Express*, vol. 21, no. 12, pp. 13958–13968, 2013.

[150] K. Padmaraju and K. Bergman, "Resolving the thermal challenges for silicon microring resonator devices," *Nanophotonics*, vol. 3, no. 4–5, pp. 269–281, 2014.

[151] A. V. Krishnamoorthy, X. Zheng, G. Li, et al., "Exploiting CMOS manufacturing to reduce tuning requirements for resonant optical devices," *IEEE Photon. J.*, vol. 3, no. 3, pp. 567–579, 2011.

[152] Z. Su, E. S. Hosseini, E. Timurdogan, et al., "Reduced wafer-scale frequency variation in adiabatic microring resonators," in *OFC 2014*. IEEE, 2014, pp. 1–3.

[153] A. Mekis, S. Gloeckner, G. Masini, et al., "A grating-coupler-enabled cmos photonics platform," *IEEE J. Sel. Top. Quantum Electron.*, vol. 17, no. 3, pp. 597–608, 2011.

[154] W. Bogaerts, S. K. Selvaraja, P. Dumon, et al., "Silicon-on-insulator spectral filters fabricated with cmos technology," *IEEE J. Sel. Top. Quantum Electron.*, vol. 16, no. 1, pp. 33–44, 2010.

[155] S. Assefa, F. Xia, and Y. A. Vlasov, "Reinventing germanium avalanche photodetector for nanophotonic on-chip optical interconnects," *Nat. Lett.*, vol. 464, pp. 80–84, 2010.

[156] S. Agarwal, T.-T. Quach, O. Parekh, et al., "Energy scaling advantages of resistive memory crossbar based computation and its application to sparse coding," *Front. Neurosci.*, vol. 9, p. 484, 2016 [Online]. Available at: https://www.frontiersin.org/article/10.3389/fnins.2015.00484.

[157] M. Glick, L. C. Kimmerling, and R. C. Pfahl, "A roadmap for integrated photonics," *Opt. Photon. News*, vol. 29, no. 3, pp. 36–41, 2018.