Photonic Multiply-Accumulate Operations for Neural Networks

Mitchell A. Nahmias, Thomas Ferreira de Lima, Alexander N. Tait, Hsuan-Tung Peng, Bhavin J. Shastri, *Member, IEEE*, and Paul R. Prucnal, *Fellow, IEEE*

***Abstract*—It has long been known that photonic communication can alleviate the data movement bottlenecks that plague conventional microelectronic processors. More recently, there has also been interest in its capabilities to implement low precision linear operations, such as matrix multiplications, fast and efficiently. We characterize the performance of photonic and electronic hardware underlying neural network models using multiply-accumulate operations. First, we investigate the limits of analog electronic crossbar arrays and on-chip photonic linear computing systems. Photonic processors are shown to have advantages in the limit of large processor sizes ($>$100 μm), large vector sizes ($N > 500$), and low noise precision ($\leq$4 bits). We discuss several proposed tunable photonic MAC systems, and provide a concrete comparison between deep learning and photonic hardware using several empirically-validated device and system models. We show significant potential improvements over digital electronics in energy ($>10^2$), speed ($>10^3$), and compute density ($>10^2$).**

***Index Terms*—Artificial intelligence, neural networks, analog computers, analog processing circuits, optical computing.**

I. Introduction

PHOTONICS has been well studied for its role in communication systems. Fiber optic links currently form the backbone of the world's telecommunications infrastructure, vastly overshadowing the best electronic communication standards in information capacity. Light signals have many advantageous properties for the transfer of information. For one, a photonic waveguide, with diameters ranging from those in fiber ($\sim$80 μm) to those fabricated on-chip ($\sim$500 nm), can carry information at enormous bandwidth densities—i.e., terabits per second—with an energy efficiency that scales nearly independent of distance. This density is possible thanks to signal parallelization in photonic waveguides, in which hundreds of high speed, multiplexed channels can be independently modulated and detected. Photonic channels also experience less distortion, jitter, and crosstalk between one another compared to their electrical counterparts.

Photonic technology has traditionally been used for long distance communication. However, modern bandwidth requirements and the standardization of silicon photonic integrated circuits (PICs) has lead to the proliferation of shorter distance photonic links. For example, silicon photonic transceivers are now a pervasive component in data-centers. In addition, the efficiency of a photonic link, which is dominated by the E/O and O/E conversion costs between the electrical and photonic domains, is rapidly encroaching on the efficiency of electronic links: the cost to move data photonically between nodes at a data-center ($\sim$1 pJ/bit [1]) is now within order unity from a modern DRAM memory stack to a processor [2].

At the same time, there has been a substantial increase in the use of many-core parallel processing systems for a variety of tasks in high performance computing (HPC). Artificial intelligence (AI), in particular, is growing at an alarming pace: deep learning models have been doubling in size every 3.5 months, far outpacing Moore's law [3]. These systems have much greater communication overheads than classical von Neumann architectures such as CPUs, resulting in a dramatic increase of both the area and energy consumption of metal interconnects (see, for example, Ref. [4]). They are also bottlenecked computationally by the ability to perform matrix multiplications efficiently, which represent the most common operations in HPC.

The most computationally expensive task in current AI models is the implementation of neural networks. Current deep learning models require dense, low-precision matrix computations. Digital instantiations of matrix (or tensor) units typically suffer from high communication overheads, expensive digital operations, and high latencies. On the hand, photonic linear operations—such as passive fourier transforms [5] or matrix operations [6]—exhibit stark advantages in bandwidth density, latency, and energy. As mentioned in [7], [8], photonic computations are passive, exhibiting favorable energy scaling costs which are potentially $O(N)$ for $O(N^2)$ fixed point operations. Photonic matrix multiplication occurs in a single step, only bottlenecked by the periphery of modulation and detection. A more surprising observation is the computational density of such an approach: despite the large sizes of photonic devices, such systems can deliver more operations per second in a given area than those in digital electronics.

This manuscript analyzes the merits of using photonics for simulating neural networks. We begin by exploring the

Manuscript received April 17, 2019; revised September 6, 2019; accepted September 7, 2019. Date of publication September 18, 2019; date of current version December 20, 2019. This work was supported by the National Science Foundation (NSF) (ECCS 1247298, DGE 1148900). *(Corresponding author: Mitchell A. Nahmias.)*

M. A. Nahmias, T. F. de Lima, A. N. Tait, H.-T. Peng, and P. R. Prucnal are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: mnahmias@princeton.edu; tlima@princeton.edu; atait@ieee.org; hpeng@princeton.edu; prucnal@princeton.edu).

B. J. Shastri is with the Department of Physics Engineering Physics and Astronomy, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: bhavin.shastri@queensu.ca).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTQE.2019.2941485

implementation of multiply-accumulate operations (which take the form $a' \leftarrow a + w \cdot x$) in various platforms in Section III, discussing the costs and benefits of digital electronics, analog electronics, and photonics. We provide a comparison of the fundamental limits of electronic crossbar arrays and photonic linear computing systems in Section IV, and analyze the performance of these models across of metrics such as energy, speed, and computational density. We consider the general performance of photonic MACs along these metrics based on practical devices that are compatible with large-scale silicon photonic foundries. In the last section, we provide a concrete comparison between fully-tunable neuromorphic photonic networks based on known photonic device models and principles with electronic state-of-the-art deep learning chips.

II. Multiply-Accumulate Operations

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:

$$a' \leftarrow a + (w \times x) \tag{1}$$

MACs are constituents of a number of linear mathematical operations, including dot products, matrix multiplications, Fourier transforms, and convolutions. MACs have traditionally characterized the performance signal processing (DSP) applications [9], [10], but have become increasingly prominent in modern HPC.

We are most interested in a specific use case: the simulation of neural network models. AI applications typically divide into *training*, in which models learn to understand a data set, and *inference*, in which trained models are deployed on new data to draw conclusions or extract information. For a set of input variables x_i and output variables y_j, each node j (or neuron) receives signals from a large number M of other nodes i. The inputs are combined via a *weighted sum* of the form $y_j = \sum_i w_{ij} x_i$. The input to the next layer x'_j sees y_j go through a nonlinear function:

$$x'_j = f\left\{\sum_i w_{ij} x_i\right\} \tag{2}$$

The function $f\{x\}$ can represent any nonlinear operation (i.e., ReLUs, spiking neurons, pooling, etc.), and can be simulated in the analog or digital domains. The weighted sum can be broken down into a series of MAC operations of the form $a_i = a_{i-1} + w_i x_i$ for $i = 1 \ldots M$. Each neuron requires M parallel MAC operations. Therefore, a neural network of size N requires $M \times N$ MAC operations per time step, or one operation per synapse. In a fully interconnected network with N nodes ($M = N$ case), the number of MAC operations required per time step Δt—or characteristic time constant τ in analog hardware—is N^2 per step. The nonlinear function $f\{x\}$ can also consume energy, but since this operation scales with $O(N)$ rather than $O(N^2)$, it does not represent the most costly operation. As the size of the network N grows large, MACs become the most burdensome hardware bottlenecks in neural networks [11]. It is therefore no surprise that MACs are the most ubiquitous computations in deep learning hardware acceleration, both in training and in inference [12], [13].

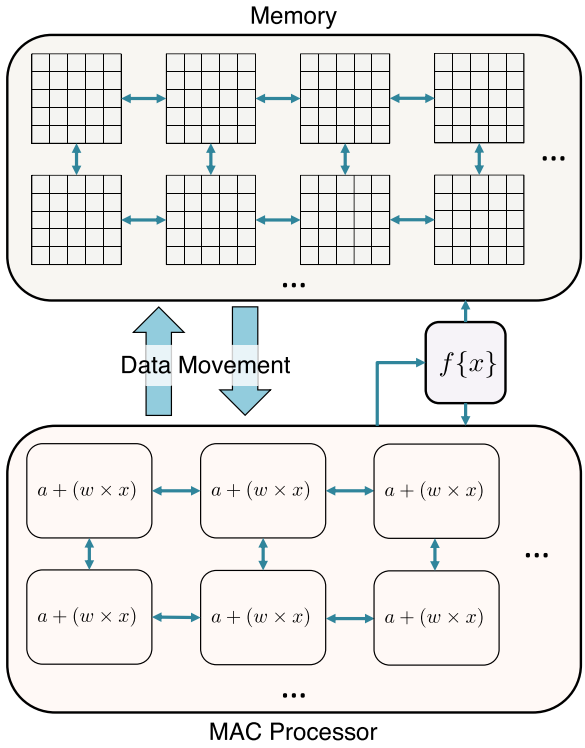


Fig. 1. A typical signal pathway for a modern AI chip. Information is passed between memory chaches, between MAC processors performing $a + (w \times x)$ and nonlinear operations $f\{x\}$. Moving data (blue arrows) consumes the majority of the energy in current systems.

A. Data Movement

Fig. 1 illustrates the signal pathway for a typical AI processor. Tensor or vector data that resides in memory is retrieved and sent to the processor, which performs MAC operations $(a' \leftarrow a + (w \times x))$ and some other nonlinear functions (encompassed in $f\{x\}$) before the result is sent back to memory and stored. Although MACs constitute the majority of operations in AI, in practice, most of the energy is lost data movement [14], [15]. Activations must be shuttled to and from various memory caches and buffers to the matrix multiplication units and back. The cost primarily comes from charging and discharging metal wires, which have a capacitance per unit length of around 100 aF/μm with charging energy proportional to $\sim CV^2$. Since the voltage V is fixed by the fabrication node, conventional digital electronics must necessarily pay this cost [16] (see discussion in Section III-A).

It is well known that photonics has the capability of greatly reducing the data movement problem that currently plagues electronic chips [17]–[20]. Optical loss is nearly negligible for intrachip distances (see Section III-C), so instead of paying an energy cost proportional to the length of each connection, photonic links pay the cost upfront converting from the electrical

domain to the photonic domain and back. Waveguides can thus beat metal wires in efficiency, provided that the cost of $\mathrm{E/O/E}$ conversion is less than that of charging a metal wire over the same distance.

It is not yet clear whether addressing the data movement problem alone is worthwhile—we still pay the $\mathrm{E/O/E}$ cost ($\sim$0.1 pJ to 1 pJ [1]) communicating between cores, which is within order unity of the cost of each MAC operation in state-of-the-art AI chips (see Section VI). Instead, we can garner a larger advantage by using photonics for both data movement and MAC operations simultaneously, interfacing modulators and detectors in close proximity with both local memory banks and a photonic neural network processor. Photonic memory architectures have been studied in depth, having the potential for significant advantages over their electronic counterparts (see for example Ref. [21]–[23]). We focus primarily on the MAC processor in the pages that follow. A key advantage, in this case, is that the memory I/O cost is amortized over the operations performed in the processor. This can lead to significant energy savings, and ultimately, huge performance gains over digital systems.

B. Precision

Analog operations are far more resolution limited than standard floating-point operations. For example, representing a 16-bit value on an optical signal at minimum requires detecting 2^{32} photons per time step to stay above the shot noise limit, which, at typical telecommunications wavelengths ($\lambda \sim 1.55\ \mu$m), puts us above the energy consumption of current digital processors ($\sim$550 pJ per sample, leading to $>$1 pJ/MAC, see Table II). Since analog systems use physical representations of real numbers, they lack the dynamic range to represent different exponents. Their operations are equivalent to fixed point, in which the exponent is fixed during computation.

Thankfully, empirical research has shown that neural networks can operate effectively with both low precision and fixed point operations. Inference models work nearly just as well with 4-8 bits of precision in both activations and weights—sometimes even down to 1–2 bits [24], [25]—and training with nearly 8–16 bits of precision per computation [26], [27]. Training can even work with binary weight evaluations, as long as high resolution stored weights are applied stochastically during training [28]. There is also evidence that fixed point arithmetic within the matrix core is also effective for both inference [27] and training [29]. This puts deep learning in range of analog photonic processing, which has been shown to exhibit a tuning accuracy of more than 4 bits [8], [30], [31]. However, many of these studies have focused on *quantized precision*, in which signals are resolved deterministically via a set of threshold values. Analog systems are for more stochastic, with both unbiased noise from the signal pathway and biased noise from fabrication variation. In the digital domain, there are strict conditions on the number of bit errors that systems can handle (typically, we want $\mathrm{SNR} \sim 10$ dB for a digital channel with forward error correction [32]).

The degree of noise or fault tolerance can vary significantly across different neural network models [33], but interestingly, such models can *be made* robust via proper construction and training [34]–[37]. In some cases, unbiased noise added during training results in a more robust model, effectively acting as a form of regularization [38]. The resulting network becomes more noise-tolerant with an accuracy that is equivalent to a network trained without noise [39]. In practice, noise levels can also approach deterministic precision thresholds: for example, stochastic rounding across signals and weights has many theoretical advantages [40], which is effectively similar to setting the $\mathrm{SNR} \sim 0$ dB relative to the quantization level. In this sense, robustly constructed neural networks can operate with far more noise than standard digital links.

For the remainder of this manuscript, we characterize our precision with respect to the analog noise in each channel. We define a parameter $\mathrm{SNR} \equiv 2^{N_b}$, where N_b represents the number of *bits of noise precision* for a given computation. We will also define a parameter, ρ (see Ref. [41] for an equivalent definition) which represents the loss of precision in the analog domain from the digital domain. For $\rho = N$, we have fixed point arithmetic, in which the precision is only defined with respect to the dynamic range of the output after summation $\sum x_i w_i$. This leads to scaling advantages, as discussed in Sections IV and VI. For $\rho = 1$, we guarantee every output $w_i x_i$ maintains full precision N_b, if even if the weight w_i is small. We can also have $1 < \rho < N$ where the desired precision is in some way dependent to the amplitude of the signal: $\rho = \sqrt{N}$ represents an interesting case, guaranteeing that the precision of a signal in a prior layer maintains the same precision in the next layer after a $1/N$ fan-out loss. Importantly, we will consider only the fixed point case covering the full dynamic range of the output ($\rho = N$), since it leads to great efficiency in the analog domain.

C. Compute Density

Throughout this manuscript, we define a figure-of-merit that we can use to compare various architectures with one another. This metric (previously used to benchmark power-performant floating point operations in digital electronics [42]) will be referred to as *compute density*, and is defined as follows:

$$D = \frac{\text{Speed (MACs/s)}}{\text{Area per MAC unit } (\text{mm}^2)} \tag{3}$$

Compute density is related to several other well established metrics. For example, since it is limited by the ability to communicate across each MAC unit, its upper bounded by *bandwidth density* (bits/s/mm^2 in Ref. [18]). It is also affected by energy efficiency, since we must keep our system within a reasonable power density ($<$1 W/mm^2 [42]) to prevent thermal runaway. We analyze these limits in Section IV.

There are a number of reasons why compute density is useful, particularly when we are comparing different kinds of architectures that may multiplex signals differently or run at vastly different clock rates. When we look at crossbar arrays (such as Ref. [43]) or digital matrix configurations such as

systolic arrays [12], [44], there are well defined notions of MAC area, MAC density, memory density, and speed. However, digital architectures that use time multiplexing strategies (i.e., TrueNorth [45]), or photonic strategies that could use either time or wavelength multiplexing (i.e., those described in Sections IV and V) do not necessarily have the same clear definitions, because there are many more MACs being implemented than the number of physical units. It depends on whether we consider these "virtual" MACs as part of the density calculation or not, which can complicate our comparisons.

Defining a compute density metric remedies these ambiguities, providing a grounded way to speak about processing power that is relatively invariant to the multiplexing or channelization strategy. We will also see that, like bandwidth density, it is not necessary to define how we divide the spectrum up into independent channels in order to talk intelligently about the limits of compute density. And ultimately, we are interested in the total amount of computational power (op/s) that a given system can exhibit. Microprocessor areas are fairly invariant—they tend to occupy 100 s of mm^2 because of limits in cost, yield, and reticle size. From this perspective, compute density also acts as a measure of the compute power of a microprocessor that uses a given architecture, since chipsets likely occupy areas that are within order unity.

III. Physical Implementations of Neural Network Hardware

In order to compare electronic and photonic processing with one another, we will use the multiply-accumulate (MAC) operation, defined in Section II. Below, we explore the advantages and disadvantages of implementing these operations in digital microelectronics, analog electronics, and photonics.

A. Electronic Implementations

1) Digital Electronics: Conventional digital computers are based on the von Neumann architecture [46] (also called the Princeton architecture), and are typically implemented in silicon microelectronics. They include a memory bank that stores both data and instructions, and a central processing unit (CPU) that performs nonlinear operations. Instructions and data stored in the memory unit lie behind a shared multiplexed bus which means that both cannot be accessed simultaneously. This leads to the well known von Neumann bottleneck [47] which fundamentally limits the performance of the system—a problem that is aggravated as processors run memory-bound algorithms. Nonetheless, this computing paradigm has dominated for over 60 years, driven in part by the continual progress dictated by Moore's law for CPU scaling—the number of transistors that can be put on a microchip doubles every 18 months to 24months [48]—and Koomey's law—the number of computations per joule of energy dissipated doubles approximately every 1.57 years [49].

These limitations have lead to the massive parallelization and specialization of hardware architectures [50]. CPUs used to be the most common choice for most applications, but in recent years, many-core architectures such as GPUs and FPGAs have expanded to encompass general purpose tasks in the high performance computing arena. Concurrently, specialized ASICs are becoming increasingly popular for the implementation of artificial intelligence algorithms, which require low precision, high density matrix computations, a notable example of which is Google's Tensor Processing Unit (TPU) [12]. Although parallelization can break down tasks that are highly distributable [51], the performance of this operation eventually leads to diminishing returns as a result of Amdahl's law [52]. As a separate issue, I/O latency and sequential processing capabilities cannot exceed the time resolution of the processor itself, which is ultimately bounded by its *clock rate*. Even MAC units need to serialize the summands to perform weighted addition (Fig. 2(a)).

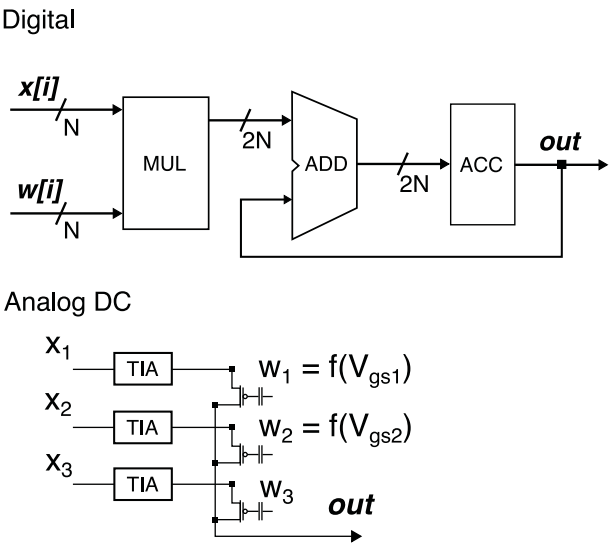


Fig. 2. Parallel MAC operations (i.e. weighted addition $\sum_i w_i x_i$) in different electronic implementations. (a) A typical MAC unit today includes separate multiply and accumulate operations, implemented in digital logic. (b) An analog implementation could use tunable impedance to implement weights, which can be instantiated in dense crossbar arrays.

Although digital microelectronics continue to increase in performance as lower nodes are introduced, an increasing number of practical barriers are inhibiting the scaling of processing and energy densities. As an illustrative example, clock rates have saturated to around 500 MHz to 4 GHz [53], and chip designers have been forced to focus efforts on parallelism instead [54]. Attempting to drive processors faster, or with higher compute density, results in several runaway effects, including:

- **Energy Consumption:** The scalability of modern microprocessors is largely limited by power density, or energy dissipation per unit area (W/mm^2). There is a trade-off between bandwidth and energy consumption in electronic devices. Ideally, the energy lost is almost entirely due to capacitive discharging. Dynamic power scales according to:

$$P = \alpha_{0\to 1}\frac{1}{2}CV^2 f_{\mathrm{s}} \tag{4}$$

for node transition activity factor $\alpha_{0\to 1}$, capacitance C, driving voltage V, and switching frequency f_{s} [55]. However, at higher frequencies, secondary effects such as short circuit current and leakage become more pronounced,

causing $\alpha_{0\to1}$ to decrease and P to hit a floor value. Different material structures, device architectures, or higher driving voltages V can offset these effects, but typically increase energy consumption. This, in turn, produces more heat, which can manifest in runaway thermal effects. These thermal limitations are often the dominant limiting factor for chip scalability [56].

The largest energy contribution originates from communication, which primarily involves charging and discharging many metal wires. Metal lines, like electronic devices, dissipate energy resistively (via Eq. (4)). In many processors with high communication overheads—such as FPGAs or deep learning chips—communication can easily occupy more than half the energy cost [57], [58]. As it stands, digital architectures are far from optimal: the power efficiency of biological systems is estimated to be <1 aJ [11] per MAC operation, which is six orders of magnitude greater than the power efficiency of current state-of-the-art machine learning chips at $\sim$1 pJ (see Fig. 7).

- **Signal Bandwidth:** Since interconnects are restricted by geometric constraints, microelectronic circuits typically rely on some form of temporal multiplexing for widespread, parallel data distribution between processors. For example, many neuromorphic architectures employ a digitization scheme called address event representation (AER) to communicate events between different neural processor cores [59], [60]. Unfortunately, electronic connections experience harsh trade-offs between bandwidth and interconnectivity. Signal bandwidth for both capacitive and inductive lines scale according to

$$B_l \propto \frac{A}{L^2} \tag{5}$$

for bandwidth B_l, cross sectional area A, and [61], [62]. As a result, metal wires are typically limited to signals no faster than several gigahertz in frequency. Temporal multiplexing strategies lead to even harsher trade-offs, since multiplexing N channels each with channel bandwidths B_c requires a total bandwidth of at least $B_l \geq NB_c$ per multiplexed line.

B. Analog Electronics: Spatial Multiplexing

One way to avoid digital bottlenecks is to use an analog networking configuration in which each connection is represented by a physical wire. Dense connections can be instantiated in space-efficient topology such as crossbar arrays [63], [64]. Summation and multiplication can both be performed simultaneously using resistive elements together with Kirchhoff's current law (Fig. 2(b)). However, closely spaced wires also experience bandwith-distance trade-offs. As an illustrative example, for a cluster of adjacent wires with pitch P, width $P/2$, thickness T, length L, RC bandwidth scales according to [61]:

$$B_l \propto \frac{1}{L^2}\left(\frac{1}{P^2}+\frac{1}{T^2}\right)^{-1} \tag{6}$$

This can become particularly problematic for large $L > 1$ mm^2, and is responsible for the enormous energy costs seen for off-chip communication in electronics. That being said, if L is kept small, the bandwidth can actually be quite high and the energetic communication cost low [17]. One must be careful to shrink the cores in a small area to keep the efficiency as high as possible (this point is discussed in more detail in Section IV).

One of the primary difficulties of analog electronic arrays is finding a good linear and tunable resistive element—traditional transistors, optimized for digital operations, do not have the linear transconductance profiles to make this tenable. New materials or fabrication approaches are therefore a necessity in creating efficient analog electronic arrays. To this end, memristive devices have been explored quite extensively (see Ref. [43], [65], [66]) along with phase change memory (see Ref. [67] for a good review), which have yielded a number of interesting approaches for high-density storage and computing. For example, memristive memory now beat traditional flash memory in performance along many metrics, including density, reliability, speed, and endurance (see for example Ref. [68]). Nonetheless, for tunable resistive elements to take full advantage of the possibilities that crossbar arrays have to offer (as discussed in Section IV), we need to see additional performance improvements, and there needs to be a low-cost way to integrate them into standard fabrication processes.

C. Photonic Implementations

Photonic signals can support much greater bandwidth densities and consume less energy for longer distances than the electrical counterparts [16]. This has motivated the development of fiber optic technology in telecommunication networks and now, interconnections in datacenters and processors [62]. The advantages of photonics are especially relevant for systems with high communication or bandwidth overheads. There are several unique physical properties that allow optical signals to manifest these advantages:

- **Bandwidth:** Optical carrier waves possess different orthogonal features, including wavelength, spatial mode and polarization, which do not interact with each other in passive devices. The total complex electric field $\vec{E}(x,y,z,t)$ in a waveguide or fiber optic communication channel can be described as a sum over every optical mode m, polarization p, and wavelength n:

$$\vec{E}(x,y,z,t) = \sum_m \sum_p \sum_n \vec{e_p} A_{mp}(x,y) B_{mnp}(t') \times \cos(\omega_n t - \beta_i z + \phi_n)$$

for unit vector $\vec{e_p}$, mode profile A_{mp}, time-dependent term B_{mnp}, angular frequency ω_n, propagation vector β_n, $t' = t - z/v_g$, and group velocity v_g. Each term can be modulated independently via B_{mnp} and, in the absence of interference, can be separated using linear photonic devices. The optical telecommunication band itself has $\Delta f \sim 5$ THz of spectral bandwidth, which provides approximately $\sim$5 Tb/s of information capacity for every

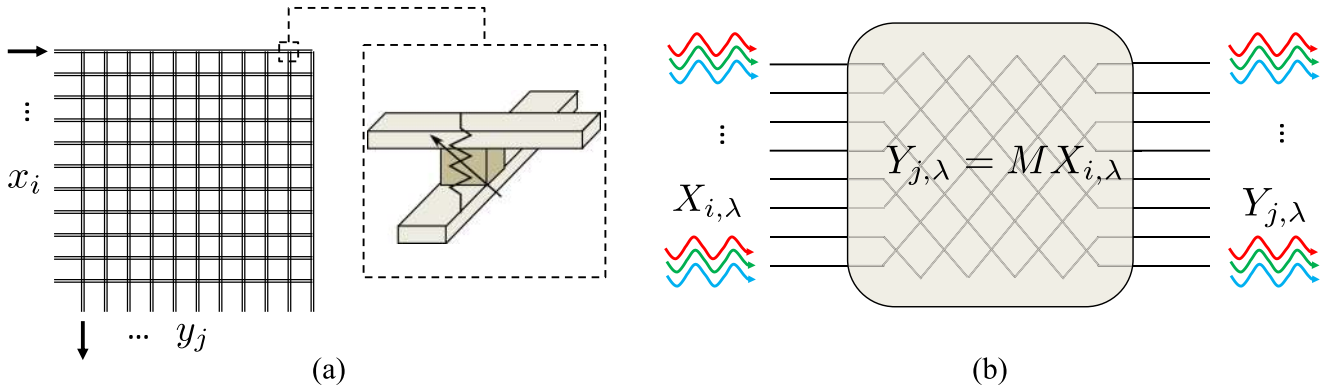


Fig. 3. Schematic of analog matrix cores in electronics and photonics. (a) A schematic of an $N \times N$ resistive crossbar array, with a tunable resistive element at each junction that represents the matrix element being applied for input voltages (or currents) x_i and output currents (or voltages) y_j. The size of the core scales with $(NP)^2$ for wire pitch P. (b) A schematic of a hypothetical $N \times N$ evanescent-field-limited wavelength multiplexed optical matrix core, with wavelength multiplexed inputs $X_{i,\lambda}$ and outputs $Y_{j,\lambda}$ along k wavelengths (labeled by λ) in different waveguides (labeled by i). M applies some linear function to the inputs to create the output vector, using local operations at waveguide junctions. The size of the core scales with $([N/k]P_\lambda)^2$ for waveguide pitch P_λ.

mode m and polarization p. Unlike in electronics, bandwidth and linear separability is an *intrinsic property* of the electromagnetic wave, i.e. it is *independent* of design constraints such as waveguide length or proximity.

- **Impedance:** In optical systems, one only needs to match the refractive index to prevent reflections. In addition, since electric/optical (E/O) and optical/electrical (O/E) conversion is an inherently quantum process, electric nodes which communicate using photonic edges need not be electrically impedance matched with one another [69]. This reduces many of the design constraints that typically limit microwave electronic circuits.
- **Energy:** Since photonic signals are not subject to Joule heating, waveguides and fibers can be designed with very low signal attenuation (i.e. $<$.1 dB/cm [70] and $<$.1 dB/m in some cases [71], [72]), allowing for communication costs that scale independently of distance. This allows for the propagation of higher power signals without the associated contribution to thermal runaway. In addition, communication or computations in the optical domain could be performed with minimal or theoretically even *zero* energy consumption—especially for linear or unitary operations.

In addition to these physical benefits, there are also practical ones. While there has been research on photonic integration for some time, in the past five years, there has there been a paradigm-shift in photonic integration that could garner the manufacturing benefits enjoyed by digital microelectronics [73], [74], namely:

- **Performance:** Shrinking devices reduces their energy requirements, and allows for continuous performance scaling. Furthermore, the high yields attainable only in foundries enable the fabrication of complex photonic systems.
- **Economics:** The presence of large markets driving silicon photonics (i.e., data-center transcievers) enables economies of scale in production, amortizing the cost of fabrication and packaging.
- **Standardization:** Every foundry line has a standard library of heavily optimized device designs through which, smaller enterprises can effectively utilize the fruits of millions of dollars worth of industrial research.

Silicon photonics offers a combination of foundry compatibility, device compactness, and cost that enables the creation of scalable photonic systems on chip. Its heavy use for data-center transcievers have lead to a decrease in overall packaging costs. Of course, the industry is still new, so photonic chips are not without their challenges. A prime example is that tunable photonic devices are currently energetically expensive: microring resonators and phase shifters currently use heaters for coarse tuning, which can consume significant energy. This point is discussed more in Section V-A.

IV. Analog Matrix Multiplication: A Comparison Between Photonics and Electronics

It's clear that analog computing in both the electronic and photonic domains offer many advantages over digital microelectronics. So which one will win? To get a better sense of their performance bounds, we will compare an electronic crossbar array (the most common architecture for devices in Secton III-B) with a hypothetical dense photonic matrix core in which MACs are performed using a resistive approach in electronics and passive linear approach in photonics. Inputs for the electronic core are analog voltages and currents, whereas the inputs and outputs for the photonic core are optically multiplexed signals with analog light intensities.

We use an example of performing a single, square matrix-vector operation, consisting of N input channels and N output channels (N^2 MAC operations) with a fixed preconfigured matrix. We implicitly assume that there is a set of devices that can fully tune resistance or optical loss locally and selectively without a significant quiescent power overhead. A schematic of these models is shown in Fig. 3.

A. Bandwidth Density

We first consider how our bandwidth density limits the overall *compute density* (see Section II-C) of each approach. A given

compute core must simultaneously address both processing within the core (i.e., an efficient implementation of a MAC operation $a = a + w \times x$) and data movement *across* the core (i.e., each MAC operation requires a result from a prior MAC unit in order to perform a full dot product $\sum w_i x_i$ at the end of each row). As we will see, the data movement constraint can bound the performance of each of the cores.

We assume that there is a tunable, resistive element at the interface between metal crossbars, and each tunable element emulates a simple resistor associated with a fixed weight w. Kirchhoff's current law performs the summation $\sum w_i x_i$ with the weights within each matrix, determined by the relative resistance values along each wire. A standard formula for assessing the bandwidth of on-chip metal interconnects is $B_E \leq B_{RC} A/L^2$ per wire, for constant bit rate B_E, architecture-dependent constant B_{RC} (typically $B_{RC} \sim 10^{16}$ for on-chip RC interconnects [62]), cross sectional area A, and length L of the wire. Extending this analysis to crossbars, we make the simple observation that the area occupied by each resistive element is approximately equivalent to the cross-sectional wiring area A in two dimensions. Computing over a $N \times N$ matrix multiply array with $L = NP_l$ for crossbar line pitch P_l, this gives us our bandwidth-limited electronic compute density D_E:

$$D_E \leq \frac{B_{RC}}{L^2} \tag{7}$$

in units of 1/s/mm^2.

In the optical domain, each waveguide has an intrinsic bandwidth B_O upper bounded by the speed of the wave itself—for standard telecommunications wavelengths (1550 nm), this upper bound is in the range of $B_O \sim 100 \times 10^{12}\text{s}^{-1}$ for multiplexed signals (from $f = 193$ THz), but more realistically $\sim$5 THz for WDM-multiplexed systems in the 1.3 μm or 1.55 μm wavelength bands. Photonic waveguides are limited by the evanescent field coupling overlap between adjacent modes, which is a function of the wavelength of light. We can thus derive a minimum pitch P_λ between waveguides. This leads to a maximum bandwidth-limited photonic compute density D_O of:

$$D_O \leq \frac{B_O}{P_\lambda^2} \tag{8}$$

There is a critical difference here: electronic crossbars decrease in bandwidth density as the size of the crossbar (L^2) grows larger, whereas photonic systems maintain their density, independent of size. For fairly reasonable values based on the gain bandwidths in typical III-V devices and preventing crosstalk between waveguides ($B_O \sim 3 \times 10^{12}$ bits/s, $P_\lambda^2 \sim 2$ um), the crossover point at which $D_O > D_E$ occurs near $L > 100$ μm. Put another way, photonics is expected to exhibit a greater on-chip bandwidth density limit than electronics for cores that occupy more than $L^2 > 0.01$ mm^2 of area.

There are a number of factors that this analysis did not take into account. Channel crosstalk becomes a bigger problem for electronic systems, but this can be greatly reduced placing an isolating ground wire between each signal wire, keeping the bandwidth density still within order unity. Also, both optical and metal crossbar arrays can be scaled vertically with using 3D stacking technology (see [75] for the optical case), and optical waveguides can also include mode multiplexing, which may shrink the effective pitch P_λ. Nonetheless, the analysis above provides a good first principles look at the bandwidth density, and shows that they are both capable of enormous compute densities, with photonics winning in the large L limit.

B. Switching & Driving Energy

Here, we consider contributions from the *driving energy*—i.e., the amplitude of the signals required to drive any output circuitry—and the capacitive switching energy for both analog electronic and photonic cores. We will assume that the input and output voltages are compatible with transistors, restricting values to $\sim$0.5 V or larger to prevent thermal leakage (see discussion in Ref. [16]).

Given this voltage condition, the main way through which electronic crossbar arrays lose energy is capacitance discharge across the array. Note that it may also be possible to make the array appear purely resistive in dissipation–using, for example, inductors to cancel out the capacitance at a given frequency. This case is not considered here. The energy lost per cycle is $\frac{1}{2}CV_l^2$, where C is the capacitance of the array and V_l is the line voltage. To arrive at a per-operation metric, we consider the contribution of charging each group of metal wires surrounding each resistive element: for a wire pitch P_l, this is $L = 2P_l$. Given standard capacitances of about $c_l = 200$ aF/μm [62] a discharging according to $\frac{1}{2}C_l V_r^2$ for total capacitance $C_l = 2c_l P_l$, our energy consumption becomes:

$$E_{\text{MAC(E)}} = c_l P_l V_r^2 \tag{9}$$

per operation. For a standard line pitch $P_l \sim$ 80 nm and $V_r \sim$0.5 V, we arrive at $E_{\text{MAC(E)}} \sim 4$ aJ. This is quite low, and may be brought lower if advanced techniques are employed to reduce this pitch (i.e., $P_l \sim$12 nm in Ref. [76]).

The optical case has a potential scaling advantage, because metal wires need not be charged at each junction. In particular, photonics only requires charging N detectors for N^2 operations. However, we must generate enough light to *drive* the detector with sufficient charge, which can be significantly limiting [62]. This depends on the amount of light that each detector receives, which can be affected by the precision loss ρ. For example, in a conservative estimate, a given signal in an $N \times N$ matrix is split to $1/N$, and we must multiply our light power by N to make up for the loss if we are to maintain the same input precision ($\rho = \sqrt{N}$). In a better case (i.e., fixed point arithmetic with $\rho = N$), we care less about the signal and more about the full dynamic range of the output. For some power P_L driving a laser with efficiency η_L, some loss through the optical system η_{wg}, and detection efficiency η_d, the current we see at the detector is $I_d = \eta_L \eta_{wg} \eta_d P_L / E_{ph}$ for photon energy $E_{ph} = h\nu$. Lumping the efficiencies into a single quantum efficiency $\eta = \eta_L \eta_{wg} \eta_d$, this gives us a minimum energy of:

$$E_{\text{MAC(O)}} \geq \frac{N}{\rho^2} \cdot C_d V_r \cdot \frac{h\nu}{e\eta} \tag{10}$$

for photon energy $h\nu$ and elementary charge e. Note that we also have capacitive discharge from the detector (scaling according

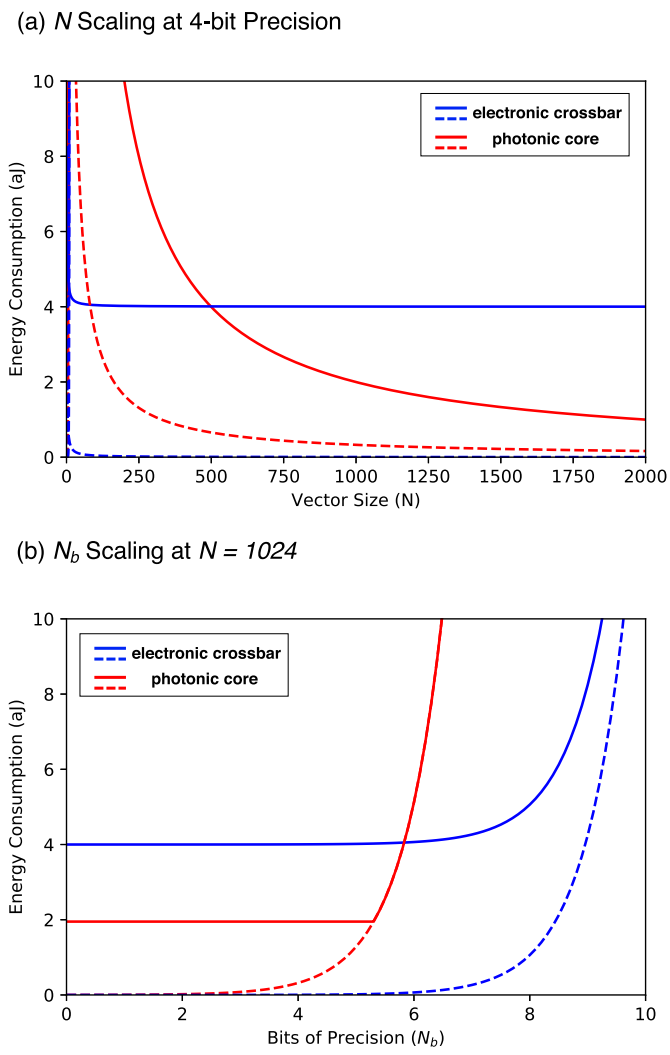


Fig. 4. Various scaling laws near the limits for photonic (red) and electronic (blue) fixed point compute cores ($\eta = 0.2, \mathrm{V_r} = 0.5$ V, $\rho = N$, $T = 300$ K, $C_d = 1$ fF, $\nu = 193$ THz). We neglect periphery costs, including the capacitive charging and discharging of drivers and receivers. Solid lines represent total energy/MAC, while dotted lines represent the noise power contribution to this energy.

to $(1/N) \cdot (1/2)CV_r^2$ per operation), although it typically has a smaller effect on the energy consumption than the driving condition above.

If we consider deep learning framework compatible with fixed point arithmetic ($\rho = N$), we see that, unlike in the electronic case, the capacitive charging scales with N rather than N^2. Choosing a high performance detector with $C_d \sim 1$ fF [77], $V_r = 0.5$ V (bringing the optical link energy to <500 aJ, see Ref. [16] for further discussion), and assuming a fairly efficient laser source ($\eta = 0.2$), we start to see a difference around $N > 500$ as shown in Fig. 4. We once again observe optical matrix multiplication cores gaining an advantage as the matrix becomes larger—in this case, we have a direct dependence on the $N \times N$ matrix size. Note that the single digit aJ/MAC bound is still a factor of 1×10^5 out of range relative to current state-of-the-art technologies (which are >100 fJ/MAC, see Section VI), so it is a far cry from limits we are seeing in the near term. Nonetheless, it is clear that both approaches have the potential for very low energy operations, with photonics exhibiting a greater overall advantage in the large N limit for capacitively-limited arrays and fixed point operations.

C. Noise

Noise affects analog precision during computations and has a strong effect on the energy consumption of each analog core. Reading values from a resistive crossbar with some SNR is fundamentally limited by thermal noise [41]. Using ρ and N_b as defined in Section II, this gives us the following expression for the energy per MAC operation:

$$E_{\mathrm{MAC(E)}} \geq \frac{N}{\rho^2} \cdot 4k_B T \cdot 2^{2N_b} \tag{11}$$

We again consider the case of full fixed point precision, where we define the precision with respect to the total output dynamic range (as discussed in Section II) and set $\rho = N$. Our MAC energy numbers become $E_{\mathrm{MAC(E)}} \sim 4$ aJ$/N$ for 4-bit operations, and $E_{\mathrm{MAC(E)}} \sim 1$ fJ$/N$ for 8-bit.

In the case of the optical matrix multiplier, we need to consider the noise on the E/O and O/E interfaces to and from the input and output. At the detector, the fundamental limit is shot noise, resulting from photon fluctuations from the incoming wave. Considering the total quantum efficiency η, we arrive at an analogous expression as above, but for shot noise:

$$E_{\mathrm{MAC(O)}} \geq \frac{N}{\rho^2} \cdot \frac{2h\nu}{\eta} \cdot 2^{2N_b} \tag{12}$$

Using a fixed point representation ($\rho = N$) with an efficient laser ($\eta \sim 0.2$) in the C-band, this gives us .33 fJ/N at 4-bit and 84 fJ/N for 8-bit.

Comparing these two quantities directly, the optical shot noise factor $\frac{2h\nu}{\eta}$ is about an order of magnitude off from the thermal noise factor $4k_BT$. If we let our E/O/E efficiency $\eta \to 1$ in the best case, the ratio between the energies is still $E_{\mathrm{MAC(O)}}/E_{\mathrm{MAC(E)}} \sim 15$, which is larger than order unity. So we see that in the limit of noise power limited operation at high precision, electrical crossbars have an advantage over photonics.

D. Discussion

We have considered the bandwidth density, switching energy, and noise at the physical limits of both electronic and optical matrix multiplier cores. We see that photonic cores exhibit scaling advantages over electronics for large core areas ($L > 100$ μm) or large channel counts ($N > 500$, see Fig. 4), but perform worse, in the limit, if the system is noise-power limited.

To illustrate performance differences between the two approaches, let's set a vector size of $N = 1024$, which is within an order of magnitude of current conventions [12]. We calculate the maximum compute density with both 4 bit and 8 bit operations. For a given energy E_{MAC}, our power density is $D_P = E_{\mathrm{MAC}} \Delta f / P^2$ and our computational density (ops/s/mm^2) is $D = \Delta f / P^2$ for pitch P between MAC elements and signal bandwidth Δf. We restrict the power density below a critical threshold $D_P < 1$ W/mm^2 [42] to prevent anticipated thermal issues that would otherwise result. We use the following parameters: a pitch of $P_E = 80$ nm

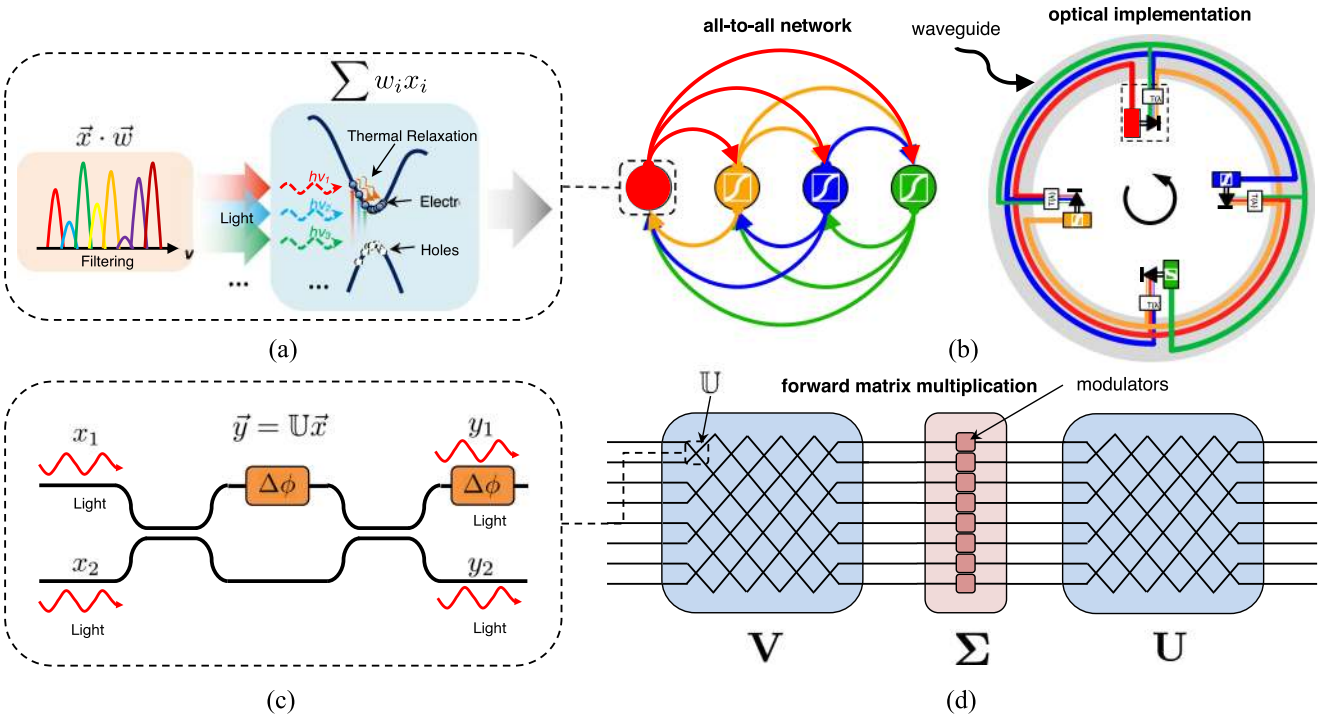


Fig. 5. Schematics for incoherent (top) [78], [79] and coherent (bottom) [8] implementations of tunable photonic multiply-accumulate operations. (a) Incoherent approaches can directly perform dot products on optically multiplexed signals. However, they rely on detectors and O/E conversion for summation. (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.

TABLE I
COMPUTE DENSITY PERFORMANCE FOR IDEALIZED ELECTRONIC AND PHOTONIC MATRIX CORES WITH $N = 1024$, SUBJECT TO POWER DENSITY <1 W/mm^2

Technology	**Noise Precision**	**Energy** (aJ/MAC)	**Compute Density** (PMACs/s/mm^2)
Electronic	4 bit	4.0	250
Crossbar	8 bit	5.0	198
Photonic	4 bit	2.0	513
Core	8 bit	81.9	12.2

for electronic crossbars, a driving voltage of $V_l = 0.5$ V, $B_O = 5$ THz, $P_\lambda = 2$ μm, $C_d = 1$ fF and $\eta = 0.2$ (assuming a fairly efficient laser source). The results are shown in Table I.

For 4-bit operations, switching energy largely dominates over noise energy for both photonics and electronics. Optical systems exhibit an advantage here: electronic cores hit the thermal density limit, but photonic cores are able to saturate their full bandwidth density limit before that point. In the 8-bit case, we see noise energy becoming significantly larger. There is a large jump in the photonic energy consumption as we move to higher precision, thanks to a quadratic dependence on the relative noise power of each signal. In cases in which high precision is necessary, operating in a noise power limited regime results in electronics crossbars performing better.

Note that although electrical crossbars are less noise-bound than photonic cores, it is unclear if this increased precision capacity is important for artificial intelligence. Ref. [27], [28] have shown that the forward compute step does not need high precision even during training, as long as the underlying weight storage and gradient rules maintain granularity. Also, since shot and thermal noise are unbiased, batching can be used to average the noise over a given set of training data (where the effective precision over the batch with M samples is equal to $N_b' = N_b + \log_2 \sqrt{M}$).

The limits discussed here are a far cry away from current technology—compute densities in the range of 100 s of PMACs/s/mm^2 are a factor of $>1 \times 10^5$ from the current state-of-the-art as discussed in Section VI. This shows that both electronic and photonic arrays have immense computational capacity, and what may ultimately differentiate them may be short term technological developments, i.e., cheap, high endurance, and tunable weight elements, or the efficiency of the nonlinear periphery surrounding each matrix core.

An interesting note is that optical systems are optically limited by P_λ, and electrical crossbars can have much smaller pitches ($<$100 nm). This means that, in the limit, photonic devices will be much larger but run at much higher speeds. This can actually a significant practical advantage: larger photonic devices may not be as sensitive to device variations or yield in a given fabrication process. We shall see that this size difference also occurs in nearer term systems, explored more closely in the sections that follow.

V. Photonic Multiply-Accumulate Operations

Here, we consider the practical performance of photonic MACs based on existing photonic devices. There are a variety of methods for implementing photonic multiply-accumulate operations using tunable photonic elements [8], [78], [80], [81] and also in many fixed network implementations in reservoir computing approaches [82]–[85]. We will distinguish between two primary mechanisms for implementing linear summations: *coherent* or *incoherent*, as defined in Ref. [7]. The former uses interferometry to implement linear operations via constructive and destructive interference, changing the relative power levels of a coherent beam. The second utilizes excited carriers to perform summations or nonlinear operations, and can potentially accept multiple wavelengths or modes.

Coherent approaches can implement linear, unitary operations while only consuming energy resulting from passive loss. However, operations must be performed within a single wavelength and mode for a given matrix—or else constructive and destructive interference would not occur between interacting lightwaves—and all-optical nonlinearities are generally challenging to implement at low optical signal intensities. Systems that fall under the interference-mediated approach include the passive reservoir [85] and the interference-based processor described in Ref. [8].

Incoherent photonic MAC units are capable of operating across different wavelengths, modes, or polarizations. For dot product functionality, filter banks (described in [78], [79]), can apply weights via the partial transmission of signals to one (or more) detectors. This can greatly increase the information density on-chip, since many independent channels can coexist in a single waveguide. Performing a MAC is also passive in the incoherent approach: for a fixed filter topology, the computations are performed as lightwaves flow to their respective destinations. Unlike in the coherent approach, semiconductor devices (and therefore, O/E conversions) are required at each nonlinear processing stage. Systems that occupy this category include those described in Ref. [78], [82], [86], [87]. A more detailed discussion of these relationship is also provided in Ref. [7].

For both approaches, we will speak broadly about photonic MAC operations in the context of an $N \times N$ matrix operation. We consider the energy per MAC, speed (signal bandwidth and latency), and computation density (i.e., MACs/s/mm^2).

A. Energy

Photonic devices, much like their resistive electronic counterparts, implement matrix operations passively and linearly. This leads to a number of advantages—in particular, for an $N \times N$ matrix, many of the most expensive energy costs scale with the size of the vector $O(N)$ rather than the size of the matrix $O(N^2)$. Below, we outline a general framework for understanding energy consumption in passive $N \times N$ photonic arrays, and provide some analysis on the trade-offs between various tunable devices.

First, we consider the cost of driving the system with a light source. An unavoidable, fundamental contribution is from shot noise, as explored in Section IV-C. We can also have relative intensity noise (RIN) on each laser input, which can affect our precision N_b. However, this is typically close to the shot noise level for sufficiently high modulation frequencies. Secondly, we must drive the capacitor of the detector with enough light to switch it (see Section IV-B). The main point to consider is whether these energies scale with $O(N)$, $O(N^2)$ or something worse, which depends on the precision loss ρ. As mentioned in Section II, it is likely that deep learning algorithms work well in fixed point arithmetic, allowing us to recover an $O(N)$ scaling law for our light input with $\rho = N$. Therefore, we potentially have a favorable scaling law for our light source, depending on the nature of the computations being performed.

Secondly, we consider costs that scale only with $O(N)$ rather than $O(N^2)$, which are those involving the periphery around the $N \times N$ matrix. Since we must first retrieve data from memory, modulate N signals on the input and detect N such signals at the output to place back into memory, we must consider the intrinsic costs associated with the driving and receiving circuitry, the modulators, detectors, and memory I/O. These energies are similar to those in digital photonic links (see Ref. [23], [88]), which include both driving and tuning the modulating device and the amplification and the recovery circuitry in the electronic receiver. Energy per sample can reach in the hundreds of fJ for co-optimized photonic platforms [88], [89].

Lastly, we consider what can be the largest contribution to energy: costs that scale with $O(N^2)$ with every photonic device. Although fixed systems can implement a pre-defined weight matrix W passively with low loss, tunable systems require a way of modifying the weight w. Photonic devices currently use heaters for coarse tuning, which consume a significant amount of power. Phase shifters in coherent approaches typically consume 10 mW to 20 mW per unit for thermal shifting [90], while microring heaters can consume $\sim$1 mW [91]. However, given the nature of passive photonic systems, these limits are not inherent. There are a variety of device modifications that promise to alleviate these problems that could see integration into foundries very soon. For example, phase shifters can be greatly enhanced with slow light cavities [92], and microresonators can be trimmed to the desired value using foundry-compatible techniques, negating the need for a heater [93]–[95].

Considering all these factors, our full energy per MAC equation is as follows:

$$E_{\mathrm{MAC}} = \frac{N}{\rho^2} \cdot \frac{h\nu}{\eta} \cdot \max\left[2^{2N_b+1}, \frac{C_d V_r}{e}\right] + \frac{1}{N} \cdot (E_{\mathrm{mod}} + E_{\mathrm{rec}} + E_{\mathrm{mem}}/M) + \frac{P_q}{\Delta\tau} \tag{13}$$

The first term accounts for the optical power supplied to the system, which may either be noise limited (left) or swing-limited (right). The second term accounts for the capacitive switching and driving circuitry for the modulators (E_{mod}), detectors (E_{rec}) and the memory retrieval cost (E_{mem}) (which includes DAC/ADC conversion if digital memory is used, which can be made quite efficient [97]). M refers to the number of compute

cycles that occur before data is passed back to memory: for example, in a hardware neural network processor with a feed-forward topology, M is equivalent to the number of network layers that have been fabricated onchip. The last term is the quiescent power use P_q for each element, which includes the power of coarse tuned heaters and the leakage power across diode junctions. We operate our system over some characteristic sampling time window $\Delta\tau$ with some effective sampling rate $1/\Delta\tau$.

In practice, for heater-tuned resonators and phase shifters, the primary source of energy consumption is from tuning each element. If we operate the system at 10 GS/s (see Ref. [88] for various photonic link speeds), this puts the energy squarely in the range of 150 fJ/MAC to 1.5 pJ/MAC for rese shifters (on the high end). If we use techniques to remedy this cost as discussed above, our next primary contributions are the link energy $E_L = E_{\text{mod}} + E_{\text{rec}} + E_{\text{mem}}/M$—which is typically in the 100 s of fJ range—and the capacitive charge of the detector, which consumes several fJ, even with conservative assumptions on precision ($\rho = \sqrt{(N)}$). The former quantity divides by N, so with channel counts in the hundreds, we are quickly brought to the low fJ/MAC range. This means that with $N > 100$ and the eradication of power hungry heaters, the single digit fJ/MAC range becomes tenable, a $>10^2$ improvement over the current state-of-the-art in energy efficiency.

In order for us to go beyond into the ~aJ range that we have explored at the analog limits (Section IV), we rely on the (I) creation of very low energy optoelectronic devices to reduce E_L significantly as discussed in Ref. [16], and (II) fixed point operations with $\rho = N$ to reduce the energy cost of the light source, which reduces both the shot noise contribution and the light required to drive each detector, and (III) a reduction of the memory I/O cost via either efficient photonic links or many-layer physical neural networks. We explore an architecture aimed at bringing aJ/MAC efficiencies in Section VI.

B. Speed

Photonic MACs can be done at very high speeds, limited only by the optoelectronic devices that encode and decode the signals on the input and output. An $N \times N$ matrix only requires one time step to compute the result. We can divide speed into two primary components: signal bandwidth and latency. If the system is bandwidth-limited by multiple parts of the signal pathway with time constants $\tau_1, \tau_2, \tau_3 \ldots$, we can approximate the total bandwidth as

$$\tau^2 \sim \tau_1^2 + \tau_2^2 \cdots$$

The delay for each component is about half the bandwidth, i.e., $\tau_1/2, \tau_2/2 \ldots$ and the total latency is the addition of all the delays s.t.

$$T \sim \tau_1 + \tau_2 \cdots$$

Several properties of photonic devices lead to their operation at much higher speeds compared to digital and analog electronic devices: (1) they do not suffer from data movement and clock distribution costs along metal wires, reducing the thermal barrier and allowing for higher clock rates, (2) a small number of photonic devices are required to perform the same MAC operations, greatly reducing latency, (3) photonic devices have a larger footprint than analog electrical devices and thus run faster to saturate the available bandwidth density, and (4) photonic arrays do not suffer from the clock jitter problems that plague metal wires and cause inconsistent signal arrival times. With typical bandwidths of >20 GHz per photonic device and only several photonic devices in a signal pathway for a given $N \times N$ matrix operation, the signal bandwidth of each input can readily exceed 10 GS/s. Similarly, a <50 ps delay time for most photonic components and only several devices per pathway (see, for example Fig. 6) results in a delay that is <100 ps. In other words, the entire matrix is effectively computed in less than a *single digital electronic clock cycle*. This contrasts quite sharply with the ~μs latencies and > 1 ns speeds seen in current electronic approaches [12]. We thus see a stark $>10^3$ decrease in

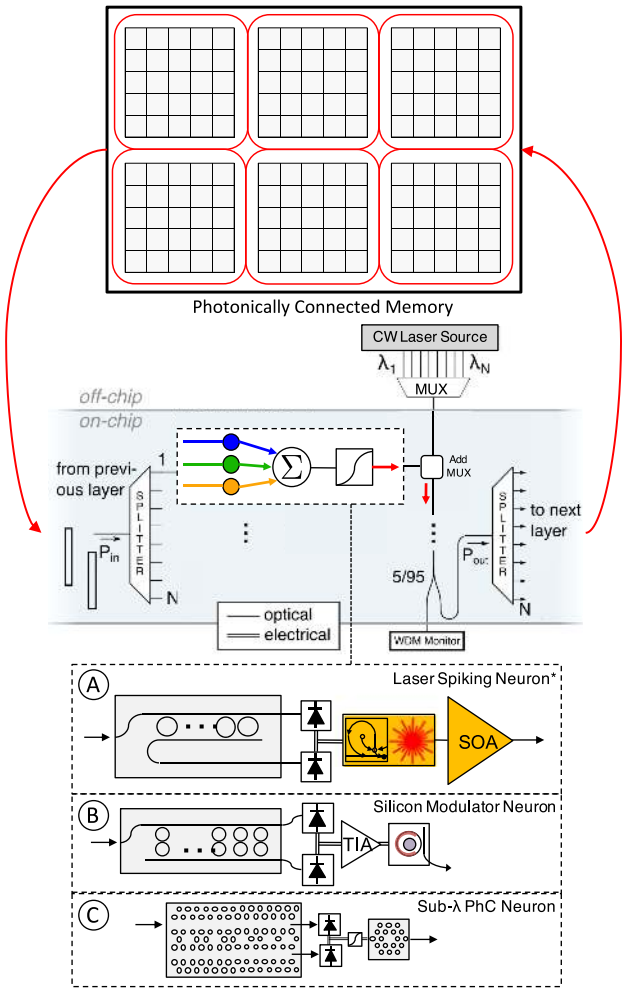


Fig. 6. Schematic of the neuromorphic photonic models under comparison with photonically-connected memory [21]–[23]. The abstract neuron model (above) can be represented using: (A) A hybrid spiking laser neuron, investigate in Ref. [7], [103]. (B) A co-integrated silicon modulator neuron, based on the system in Ref. [80], [104]. (C) A sub-λ photonic crystal neuron, running close to fundamental photonic limits. Photonically connected memory refers to models such as [21]–[23]. *Note that A does not require the off-chip laser source since it generates its own light.

TABLE II
COMPARISON OF ELECTRONIC ARCHITECTURES (TOP) WITH ESTIMATES FOR VARIOUS PHOTONIC NEURAL NETWORK (NN) APPROACHES (BOTTOM). DENSITY IS COMPUTED WITH RESPECT TO THE CORE(S) ONLY.

Technology	Energy/MAC	Compute Density	Vector Size	Precision	Latency/Speed*
Google TPU (Digital) [12]	0.43 pJ/MAC	580 GMACs/s/mm^2	256	8 bits	2 μs/1.42 ns
Flash (Analog Sim.) [96], [110]	7 fJ/MAC	18 TMACs/s/mm^2	100	5 bits	15 ns
Hybrid Laser NN [7], [109]	0.22 pJ/MAC	4.5 TMACs/s/mm^2	56	5.1+ bits	<100 ps
Co-Integrated Silicon NN [80], [104]	2.7 fJ/MAC	50 TMACs/s/mm^2	148	5.1+ bits	<100 ps
Sub-λ Nanophotonics	30.6 aJ/MAC	5 PMAC/s/mm^2	300	5.1+ bits	<50 ps

*Latency is Defined as the Time it Takes for a Single Matrix Multiplication Operation to Compute at the Given Vector Size. Speed Is Defined as the Time Between Subsequent Matrix Multiplies.

latency, meaning that any practical system will be limited more by the periphery circuitry than the neural network core itself.

C. Compute Density

We use the same compute density metric defined in Section II-C: the number of operations (MACs) performed in a given area (mm^2) per unit of time (seconds). The underlying density of a photonic compute core can be quite high using standard photonic components, which we will illustrate with a simple example: suppose we took the 512×512 AWG prototyped in Ref. [96] and used it to apply N^2 linear operations over a vectorized set of input light intensities. Suppose that there were multiple sets of these signals at different wavelengths s.t. they were multiplexed across the entire $\sim$5 THz wavelength band. If we took the number of operations and divided by the area of the chip, we get the rather large compute density of 6.8 PMAC/s/mm^2, exceeding the state-of-the-art in digital electronics by $>10^4$. This gives a picture for the capacity of photonics—the large value stems largely from the ability to multiplex both signals and connections, a technique exploited quite often by optical reservoir computing approaches (see for example Ref. [82], [85]).

However, making matrices with adjustable weight values w_{ij} can be more challenging—tunable photonic systems typically require N^2 photonic devices, since there must be a device for every weight w_{ij}. As discussed in Section V-A, there are a couple tunable approaches that have received significant attention: the coherent and incoherent approaches, which require $2N^2$ Mach-Zehnder interferometers (MZIs) or N^2 resonators, respectively. The former currently loses on compute density, since each MZI requires significantly more area ($\sim$10000 μm^2 in Ref. [8]) compared to microresonators ($\sim$250 μm^2 or much smaller). Miniaturizing each MZI relies on some complex modifications, such slow-light enhanced structures [92] or perhaps inverse design [98], [99]. whereas resonators can increase in performance as they are shrunken in size [100].

To get a better sense of what N^2 photonic devices can achieve, we can look towards prototyped devices that are compatible with silicon photonic foundry models. Standard microrings of size 50 μm $\times$ 50 μm operating at a sampling speed of 10 GS/s results in a computational density of 10 TMACs/s/mm^2. This is a major improvement over current digital electronic densities, which are around 580 GMACs/mm^2 (see Table II). A key point is that even though photonic devices are much larger than individual transistors, a single MAC unit in digital electronics is actually composed of many hundreds or thousands of transistors, occupying >100 μm in area [101], comparable to the one (or several) elements that can accomplish the same operation in analog photonics. With a higher energy efficiency, photonic elements can be clocked much faster without hitting energy density limits, leading to the overall larger compute density seen here.

What compute densities will photonics be able to attain in the near future? This is considered in the last part of Section VI, in which photonic crystal defect states [102] that occupy close to 2 μm^2 per resonator are closely packed together. As shown in Table II, this can lead to an enormous photonic compute density (5 PMACs/s/mm^2). In conclusion, we can expect photonic devices will exceed current digital electronic systems by $>10^2$ in compute density with miniaturized resonator components. In the future, more exotic structures (such as PhCs) could reach $>10^3$ as photonic devices reach their fundamental limits in size.

VI. NEURAL NETWORK HARDWARE COMPARISON

This section provides comparisons between neuromorphic photonic processing models and digital electronic processing systems. For concreteness, we focus specifically on Broadcast-and-Weight architectures [78], [103], which have been developed enough for a comparison to be possible—in particular, the empirical validation of both tunable weight systems [30], [105]–[107] and nonlinear processors that have a direct functional correspondence with neuron models [7], [104]. Nonetheless, given that photonic architectures are bound by the same physical constraints and underlying devices, this comparison provides some insights for the performance of neuromorphic photonic systems in the more general case. For the photonic platforms, we choose three models with distinct characteristics: 1) a laser neural network based on an instantiation in a hybrid spiking III-V/silicon platform [108], [109], 2) a silicon photonics platform with tight co-integration with digital electronic drivers, controllers, and amplifiers [80], and 3) a nanophotonic platform operating close to fundamental noise limits. These hardware platforms are depicted in Fig. 6. A list of computed values is included in Table II, and a graph depicting the compute density and energy efficiency—along with some of the the analog limits discussed in Section IV—is shown in Fig. 7.

1) Hybrid Laser Neural Network: This model, which is largely the focus of Ref. [103], uses currently available silicon

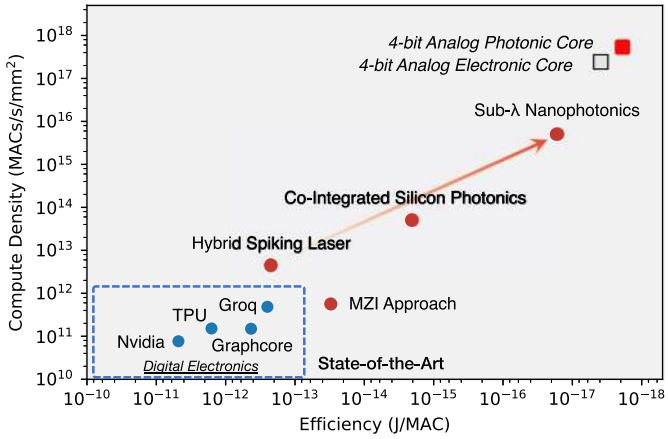


Fig. 7. Comparison of deep learning hardware accelerators with photonic platforms discussed in Section VI, modified from Ref. [7]. Photonic systems can support high bandwidth densities on-chip while consuming minimal energy both transporting data and performing computations. Metrics for digital electronic architectures taken from various sources [12], [124]–[127]. Also included are the analog limits for photonic and electronic matrix cores with $N = 1024$ and 4 bits of precision, from Table I.

photonic technology together with integrated III-V lasers to emulate biological spiking behavior. It has been proposed together with the Broadcast-and-Weight networking framework [78], and has also received considerable experimental validation, both in the tunable weight units [105] and the nonlinear processors that communicate using such units [7], [111], [112]. These systems are limited by two primary sources of energy consumption: the quiescent power of the laser and amplifier units (which can be as large as 200 mW), and the static power consumption of the heaters used within each filter bank (which can be as large as 2 mW each). For the comparison, we assume an all-to-all network with a channel number of $N = 56$, based on limits discussed in Ref. [105]. The precision is based on experimentally-validated measurements of microring filters [107]. We assume that, for excitable operation, lasers are biased close to that threshold. We also consider a semiconductor optical amplifier on the output port to generate enough output power for the next stage. For an $N \times N$ fully-connected network, the energy consumption per MAC operation can be expressed as:

$$E_{\mathrm{MAC}} = \frac{1}{N} \cdot \underbrace{\frac{P_{\lambda(\mathrm{th})} + P_{\mathrm{SOA}}}{\tau_s}}_{\text{node energy}} + \underbrace{\frac{P_{\mathrm{h}} + P_{\mathrm{l}}}{\tau_s}}_{\text{edge energy}} \tag{14}$$

where $P_{\lambda(\mathrm{th})} = I_{\mathrm{th}} V_L$ represents the laser power consumption at threshold current I_{th}, $P_{\mathrm{SOA}} = I_{\mathrm{SOA}} V_{\mathrm{SOA}}$ is the power consumption of each output SOA, $P_{\mathrm{h}} = I_{\mathrm{h}}^2 R_{\mathrm{h}}$ is the average power dissipation of each microring heater, and $P_{\mathrm{l}} = I_{\mathrm{l}} V_{\mathrm{MRR}}$ is the current across the junction biased at V_{MRR}. τ_s represents the effective sampling rate, determined by the bandwidth of the real-time signal pathway and I/O (i.e., $\sim$10 GHz [109]) during operation. We distinguish between power use at each node (which scales with $O(N)$ for N^2 operations) and power use at each edge (which scales with $O(N^2)$ for a MAC performed at each network edge), and omit the memory I/O cost, since it is not dominant here.

In this system, energy efficiency is primarily bottlenecked on the quiescent power consumption of the optical amplifier and that of the heaters. In practice, the remaining contributions—the laser threshold power and leakage terms—are negligible in comparison. In particular, the amplifier must provide enough energy to drive the next stage, meeting *cascadability* conditions as discussed in Ref [7]. With our assumed channel density $N = 56$, and other parameters based on current photonic devices ($\tau_s \sim 100$ ps), we arrive at 0.22 pJ shown in Table II. This system is comparable to deep learning chips and neuromorphic electronic systems in energy consumption, fan-in, and compute density. In the following section, we will explore the improvements that can manifest in systems better optimized for higher energy efficiency.

2) Co-Integrated Neuromorphic Silicon Photonic Network: This platform (first discussed in Ref. [80], [104]) uses continuous models and can vastly reduce the energy consumption via a close interface between digital electronic and photonic systems. This interface allows easy E/O and O/E conversions between electrical nonlinearities and photonic linear computation elements. This system also uses silicon photonic technologies that are currently available in foundries, but its performance depends critically on several new developments and insights, including: (I) the use of active electronic amplification to sidestep the gain-bandwidth trade-off in each nonlinear processing unit, and (II) the reduction of static power in microring filters by minimizing the use of heaters. For the remainder of this analysis, we also assume a close proximity, low capacitance interface between electronics and photonics (i.e., TOVs with <50 fF [113]), and low-node electronics (i.e., FinFETs [114]).

One of the first challenges is minimizing the quiescent power usage that results from each filter (scaling with $O(N^2)$) requiring a power hungry heater. Note that this is not a problem inherent in photonic elements, since a pre-fabricated fixed photonic network performs the same computations without consuming power. To avoid the immense cost of tuning across the fabrication variation that occurs across microresonators, we assume that each element is trimmed to avoid the use of heaters, as discussed in Section V-A. Integrating these approaches into the fabrication process would allow for an tremendous reduction ($P_h \to 0$) in energy consumption.

Next, we consider the limitations imposed by amplitude cascadability. In a *passive* neuron configuration in which a detector directly drives a modulator with no intermediate circuitry (i.e., Ref. [104], each nonlinear element must replenish the energy lost from the previous layer. In an all-to-all N-node network with N^2 connections, we must assure that the small-signal gain from layer to layer allows is greater than unity (i.e., $g = dP_{\mathrm{out}}/dP_{\mathrm{in}} > 1$). This puts the following lower bound on the energy consumption per MAC operation [115]:

$$E_{\mathrm{MAC}} \geq \frac{N}{\rho^2} \cdot \underbrace{\frac{h\nu}{\eta}}_{\substack{\text{quantum} \\ \text{efficiency}}} \cdot \frac{1}{e} \underbrace{[2\pi V_s (C_{\mathrm{mod}} + C_{\mathrm{PD}})]}_{\substack{\text{switching charge} \\ \text{for gain cascadability}}} \tag{15}$$

where $\eta = \eta_L \eta_{wg} \eta_d$ is laser efficiency, photonic link efficiency, and photodetector efficiency, respectively; V_s is the inverse slope of the modulator's voltage-to-transmission curve $T(V)$; and $C_{\rm mod}, C_{\rm PD}$ are the joint capacitances of the photodetetor and modulator. In a typical foundry-model where $V_s(C_{\rm mod} + C_{\rm PD}) \sim 70$ fC and $\eta \sim .06$ (which includes the passive losses through the weight banks, which can be made quite small [116]), even with $\rho = N$ in fixed point systems, we arrive at a floor of approximately $E_{\rm MAC} \geq 30$ fJ/MAC.

Going beyond this barrier requires the use of an active trans-impedance amplifier (TIA) placed between the detector and modulator, which can be instantiated using digitally-compatible circuitry in a number of different configurations. This serves several functions: it can separate capacitive contributions of the photodetector and modulator, and it also reduces the impedances associated with each stage. In a low-node electronics platform with TOVs, the energy consumption per sample can be quite low ($<$100 fJ) for a good TIA, see for example analysis in Ref. [88] or Ref. [89], [117]. Given that the signal-to-noise ratio must exceed the given bit precision N_b (i.e., ${\rm SNR} = I_p/\sigma_i > 2^{N_b}$ for received current I_p and RMS shot noise current σ_i at each detector), we arrive at a new energy-per-MAC metric:

$$E_{\rm MAC} = \frac{N}{\rho^2} \cdot \underbrace{\frac{2h\nu}{\eta}}_{\substack{\text{quantum}\\ \text{efficiency}}} \cdot \underbrace{2^{2N_b}}_{\substack{\text{noise and}\\ \text{resolution}}} + \underbrace{\frac{E_{\rm link}}{N}}_{\substack{\text{switching}\\ \text{energy/MAC}}} + \underbrace{\frac{P_l}{\tau_s}}_{\text{leakage}} \tag{16}$$

Here, $E_{\rm link}$ includes contributions from the active TIA, the modulator switching energy per unit of time τ_s, which is typically expressed in J/bit, and the energy associated with memory I/O. We conservatively assume just one neural network layer [$M = 1$ from Eq. (13)]. Note that we neglect the effect of nonlinearity on noise reduction, which can have positive effects on the resulting precision, as discussed in Ref. [115]. With fixed point like precision ($\rho = 1$), power dissipation is dominated by E/O and O/E interfaces together with digital circuitry. Given the similarity between each modulator neuron and the E/O/E interface in a standard photonic link—requiring the same electrical interfaces, amplification, and driver circuitry—we can use $E_{\rm link}$ estimates from digital links [88], [118] and those from photonically connected DRAM memory [21] to arrive at 400 fJ/sample as a relatively accurate proxy for the link energy.

We arrive at our energy consumption of 2.7 fJ as shown in Table II. We assume an improvement in areal density and channel density by shrinking the resonators to $\sim$10 μm in diameter [100] and high fidelity photonic two-pole filters as described in [107].

3) Sub-λ Nanophotonics: Here, we consider the performance of photonic devices as they begin to hit their physical limits in the B&W architecture. The basic principle of operation of each unit is similar to the co-integrated silicon photonic network. The platform is assumed to include both low node electronics and photonics on the same platform (i.e., a variant of [119]) to avoid additional capacitances at the interfaces. Additionally, we assume that there are a significant number of layers in the network ($M \sim 100$) before the information is passed back to memory, further amortizing the energy cost (400 fJ/sample) of the photonic memory link by the factor $1/{\rm M}$ [as described in Eq. (13)]. Each sub-λ neuron uses (I) a nanophotonic photodetector such as [77] with $<$1fF of capacitance, (II) operate in the "near-receiverless" regime discussed in [16], i.e., a minimal gain stage, if any, between the detector and modulator such as a single inverter amplifier (see Ref. [117], [120]), and (III) the filters and modulators are instantiated efficiently using more exotic enhancement techniques [121], [122]. We utilize devices that have been empirically prototyped, but not yet scaled in foundries. Our metrics are based on several insights:

- **Compute Density:** Photonic devices can be shrunk significantly in size compared to where they are now. The smallest known resonators are photonic crystal defect states, [102] which can occupy small footprints—if we pack them very tightly, they can be as small as $\sim$2 μm^2. A single defect state can potentially perform a weight multiplication. This has significant ramifications for compute density ($\sim 10^3$) compared to microring filter banks, even if the effective sampling rate is kept constant.
- **Channel Number:** The number of channels is limited by the total bandwidth available in the optical spectrum. At 10 GS/s, we can fit about fit about $\sim$300 channels in a 30 nm spectral gain curve. Although channel number can be extended further through the use of heteregeneous laser sources or frequency combs, this goes beyond the scope of this work. We also assume low precision, fixed point operations ($\rho = N$).
- **Energy Consumption:** There are many vectors for improvement in Eq. (16). We will assume the reverse-biased filter leakage can be brought down from microamperes [100] to nanoamperes with better manufacturing. The O/E/O switching energy $E_{\rm samp}$—which shares many properties with digital links—can be improved significantly using a variety of techniques to reach the $\sim$1 fJ level [16]. Modulators, for example, can reach in the $\sim$100 aJ per bit range [123]. We also assume that optical losses through the system are small, which can be optimized via passive device engineering. With this in place, the system is now bottlenecked by shot noise at the detector and the cost of the I/O to memory. limiting precision for a given input power. With more efficient laser sources, the total quantum efficiency to as high as $\eta \sim 20\%$. All together, this leads to $E_{\rm MAC} = 17.3$ aJ + 13.3 aJ (memory) = 30.6 aJ.

VII. Summary and Concluding Remarks

Historically, both electronic neuromorphic systems and electronic emulations of neural networks have been constrained by the inherent scaling laws of digital systems and metal interconnects. In particular, energy scales with $O(N^2)$, where N is the number of neurons, and for systems of large numbers of neurons, this becomes untenable for modern applications. Photonics provides a solution, alleviating the energy consumption of both data movement across metal wires and of multiply-accumulate (MAC) computation itself, both of which are major bottlenecks in neural computing.

We have extensively compared the limits of electronic crossbar arrays with photonics linear compute cores, and have

shown that photonics exhibits advantages for large processor sizes ($>100\,\mu$m), large vector sizes ($N > 500$), and low precision ($\leq$4 bits). We have discussed the myriad advantages that photonic multiply-accumulate (MAC) operations possess over their digital electronic counterparts in energy ($>10^2$), speed ($>10^3$), and compute density ($>10^2$). We have analyzed how they can manifest in practical models, based on empirically validated, foundry compatible photonic devices. Although we considered resonator-based methods for networking and linear operations, the advantages of photonic MACs remain relevant for many architectures beyond those presented in this work.

Although photonics has traditionally been studied for its role in communication, there is great potential to address new and emerging bottlenecks in computing. Artificial intelligence has brought unique challenges to processor architectures: modern GPUs and machine learning ASICs now implement high volume, high density, low precision matrix operations with specialized compute cores. These processors are subject to trade-offs that are significantly more communication bottlenecked than traditional von Neumann architectures. There are still many challenges towards seeing functional analog computing systems: for example, one must consider the cost of the periphery, the cost of reprogramming weights during training, the cost of A/D and D/A conversion, and the higher-level communication protocols between multiple neuromorphic cores. Nonetheless, photonics has the potential to address the major bottlenecks present in AI hardware, providing a means to simultaneously move data across a chip and perform matrix multiplication with little cost.

ACKNOWLEDGMENT

I would like to thank both Michael Gao and Marcus Gomez for their stimulating discussions regarding artificial intelligence algorithms and hardware.

REFERENCES

[1] B. R. Moss *et al.*, "A 1.23pj/b 2.5gb/s monolithically integrated optical carrier-injection ring modulator and all-digital driver circuit in commercial 45 nm SOI," in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2013, pp. 126–127.

[2] J. Jeddeloh and B. Keeth, "Hybrid memory cube new dram architecture increases density and performance," in *Proc. Symp. VLSI Technol.*, 2012, pp. 87–88.

[3] D. Amodei and D. Hernandez, "Ai and compute," Accessed: May 16, 2018. [Online]. Available: https://blog.openai.com/ai-and-compute/

[4] F. Akopyan *et al.*, "Truenorth: 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, Oct. 2015. [Online]. Available: http://dx.doi.org/10.1109/TCAD.2015.2474396

[5] C. Dragone, "Efficient n*n star couplers using fourier optics," *J. Lightw. Technol.*, vol. 7, no. 3, pp. 479–489, 1989.

[6] R. A. Athale and W. C. Collins, "Optical matrix–matrix multiplier based on outer productdecomposition," *Appl. Opt.*, vol. 21, no. 12, pp. 2089–2090, 1982. [Online]. Available: http://ao.osa.org/abstract.cfm?URI=ao-21-12-2089

[7] H. Peng, M. A. Nahmias, T. F. de Lima, A. N. Tait, and B. J. Shastri, "Neuromorphic photonic integrated circuits," *IEEE J. Sel. Topics Quantum Electron.*, vol. 24, no. 6, Nov./Dec. 2018, Art. no. 6101715.

[8] Y. Shen *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature Photon.*, vol. 11, pp. 441–446 2017. [Online]. Available: http://dx.doi.org/10.1038/nphoton.2017.93

[9] S. W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing*. San Diego, CA, USA: California Technical Publishing, 1997.

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

[11] J. Hasler and B. Marr, "Finding a roadmap to achieve large neuromorphic hardware systems," *Frontiers Neurosci.*, vol. 7, no. 118, 2013. [Online]. Available: http://dx.doi.org/10.3389/fnins.2013.00118

[12] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proc. 44th Annu. Int. Symp. Comput. Architecture*, 2017, pp. 1–12.

[13] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, 2015.

[14] S. R. Agrawal *et al.*, "A many-core architecture for in-memory data processing," in *Proc. 50th Annu. IEEE/ACM Int. Symp. Microarchitecture*, 2017, pp. 245–258. [Online]. Available: http://doi.acm.org/10.1145/3123939.3123985

[15] P. Jawandhiya, "Hardware design for machine learning," *Int. J. Artif. Intell. Appl.*, vol. 9, no. 1, pp. 63–84, 2018.

[16] D. A. B. Miller, "Attojoule optoelectronics for low-energy information processing and communications," *J. Lightw. Technol.*, vol. 35, no. 3, pp. 346–396, Feb. 2017. [Online]. Available: http://jlt.osa.org/abstract.cfm?URI=jlt-35-3-346

[17] D. A. B. Miller, "Device requirements for optical interconnects to silicon chips," *Proc. IEEE*, vol. 97, no. 7, pp. 1166–1185, 2009.

[18] D. A. B. Miller, "Rationale and challenges for optical interconnects to electronic chips," *Proc. IEEE*, vol. 88, no. 6, pp. 728–749, Jun. 2000. [Online]. Available: http://dx.doi.org/10.1109/5.867687

[19] C. Gunn, "CMOS photonics for high-speed interconnects," *IEEE Micro*, vol. 26, no. 2, pp. 58–66, Mar. 2006.

[20] K. Preston, N. Sherwood-Droz, J. S. Levy, and M. Lipson, "Performance guidelines for wdm interconnects based on silicon microring resonators," in *Proc. CLEO-Laser Sci. Photonic Appl*, 2011.

[21] C. Batten *et al.*, "Building many-core processor-to-DRAM networks with monolithic CMOS silicon photonics," *IEEE Micro*, vol. 29, no. 4, pp. 8–21, Aug. 21, 2009.

[22] S. Beamer *et al.*, "Re-architecting DRAM memory systems with monolithically integrated silicon photonics," in *Proc. 37th Annu. Int. Symp. Comput. Architecture*, vol. 38, no. 3, pp. 129–140, Jun. 2010. [Online]. Available: http://doi.acm.org/10.1145/1816038.1815978

[23] A. V. Krishnamoorthy *et al.*, "Computer systems based on silicon photonic interconnects," *Proc. IEEE*, vol. 97, no. 7, pp. 1337–1361, Jul. 2009.

[24] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4107–4115.

[25] Y. Umuroglu *et al.*, "Finn: A framework for fast, scalable binarized neural network inference," in *Proc. ACM/SIGDA Int. Symp. Field-Programmable Gate Arrays*, 2017, pp. 65–74.

[26] R. Banner, I. Hubara, E. Hoffer, and D. Soudry, "Scalable methods for 8-bit training of neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5145–5153.

[27] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1737–1746.

[28] M. Courbariaux, Y. Bengio, and J.-P. David, "Binaryconnect: Training deep neural networks with binary weights during propagations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3123–3131.

[29] U. Köster *et al.*, "Flexpoint: An adaptive numerical format for efficient training of deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1742–1752.

[30] A. N. Tait, T. Ferreira de Lima, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Multi-channel control for microring weight banks," *Optics Express*, vol. 24, no. 8, pp. 8895–8906, Apr. 2016. [Online]. Available: http://dx.doi.org/10.1364/OE.23.012758

[31] A. N. Tait *et al.*, "Neuromorphic photonic networks using silicon photonic weight banks," *Scientific Rep.*, vol. 7, no. 1, p. 7430, 2017. [Online]. Available: https://doi.org/10.1038/s41598-017-07754-z

[32] F. Chang, K. Onohara, and T. Mizuochi, "Forward error correction for 100g transport networks," *IEEE Commun. Mag.*, vol. 48, no. 3, pp. S48–S55, Mar. 2010.

[33] B. Reagen *et al.*, "Ares: A framework for quantifying the resilience of deep neural networks," in *Proc. 55th ACM/ESDA/IEEE Des. Autom. Conf.*, 2018, pp. 1–6.

[34] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," 2017, *arXiv:1705.10694*.

[35] S. Sukhbaatar and R. Fergus, "Learning from noisy labels with deep neural networks," 2014, *arxiv preprint arXiv:1406.20802.3*.

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[37] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus, "Regularization of neural networks using dropconnect," in *Proc. 30th Int. Conf. Mach. Learn.*, Jun. 17–19, 2013, pp. 1058–1066. [Online]. Available: http://proceedings.mlr.press/v28/wan13.html

[38] C. M. Bishop, "Training with noise is equivalent to tikhonov regularization," *Neural Comput.*, vol. 7, no. 1, pp. 108–116, Jan. 1995. [Online]. Available: https://doi.org/10.1162/neco.1995.7.1.108

[39] A. Neelakantan *et al.*, "Adding gradient noise improves learning for very deep networks," 2015, *arXiv:1511.06807*.

[40] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6869–6898, Jan. 2017.

[41] S. Agarwal *et al.*, "Energy scaling advantages of resistive memory crossbar based computation and its application to sparse coding," *Frontiers Neurosci.*, vol. 9, pp. 484–493, 2016. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2015.00484

[42] S. Galal and M. Horowitz, "Energy-efficient floating-point unit design," *IEEE Trans. Comput.*, vol. 60, no. 7, pp. 913–922, Jul. 2011.

[43] J. J. Yang, D. B. Strukov, and D. R. Stewart, "Memristive devices for computing," *Nature Nanotechnol.*, vol. 8, pp. 13–24 2012. [Online]. Available: https://doi.org/10.1038/nnano.2012.240

[44] H.-T. Kung, "Why systolic architectures?" *IEEE Comput.*, vol. 15, no. 1, pp. 37–46, Jan. 1982.

[45] F. Akopyan *et al.*, "Truenorth: 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, Oct. 2015.

[46] J. von Neumann, "First draft of a report on the EDVAC," *IEEE Ann. Hist. Comput.*, vol. 15, no. 4, pp. 27–75, Oct. 1993.

[47] J. Backus, "Can programming be liberated from the von neumann style?: A functional style and its algebra of programs," *Commun. ACM*, vol. 21, no. 8, pp. 613–641, Aug. 1978.

[48] G. E. Moore, "ch. Cramming More Components Onto Integrated Circuits," in *Readings in Computer Architecture*, M. D. Hill, N. P. Jouppi, and G. S. Sohi, Eds. San Francisco, CA, USA: Morgan Kaufmann, 2000, pp. 56–59. [Online]. Available: http://dl.acm.org/citation.cfm?id=333067.333074

[49] J. Koomey, S. Berard, M. Sanchez, and H. Wong, "Implications of historical trends in the electrical efficiency of computing," *IEEE Ann. Hist. Comput.*, vol. 33, no. 3, pp. 46–54, Mar. 2011. [Online]. Available: http://dx.doi.org/10.1109/MAHC.2010.28

[50] S. W. Keckler, W. J. Dally, B. Khailany, M. Garland, and D. Glasco, "GPUs and the future of parallel computing," *IEEE Micro*, vol. 31, no. 5, pp. 7–17, Sep. 2011.

[51] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[52] P. J. Denning and T. G. Lewis, "Exponential laws of computing growth," *Commun. ACM*, vol. 60, pp. 54–65, 2017.

[53] V. Agarwal, M. Hrishikesh, S. W. Keckler, and D. Burger, "Clock rate versus IPC: The end of the road for conventional microarchitectures," *ACM SIGARCH Comput. Architecture News*, vol. 28, no. 2, pp. 248–259, 2000.

[54] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," *IEEE Micro*, vol. 32, no. 3, pp. 122–134, May 2012.

[55] A. P. Chandrakasan and R. W. Brodersen, "Minimizing power consumption in digital CMOS circuits," *Proc. IEEE*, vol. 83, no. 4, pp. 498–523, Apr. 1995.

[56] I. L. Markov, "Limits on fundamental limits to computation," *Nature*, vol. 512, no. 7513, pp. 147–154, 2014.

[57] E. Kadric, D. Lakata, and A. Dehon, "Impact of parallelism and memory architecture on FPGA communication energy," *ACM Trans. Reconfigurable Technol. Syst.*, vol. 9, no. 4, pp. 30-1–30-23, Aug. 2016. [Online]. Available: http://doi.acm.org/10.1145/2857057

[58] M. Horowitz, "1.1 computing's energy problem (and what we can do about it)," in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2014, pp. 10–14.

[59] V. Chan, S.-C. Liu, and A. van Schaik, "AER EAR: A matched silicon cochlea pair with address event representation interface," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 54, no. 1, pp. 48–59, Jan. 2007.

[60] J. Park, T. Yu, S. Joshi, C. Maier, and G. Cauwenberghs, "Hierarchical address event routing for reconfigurable large-scale neuromorphic systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2408–2422, Oct. 2017.

[61] M. T. Bohr, "Interconnect scaling-the real limiter to high performance ULSI," in *Proc. Int. Electron Devices Meet.*, 1995, pp. 241–244.

[62] D. A. B. Miller, "Device requirements for optical interconnects to silicon chips," *Proc. IEEE*, vol. 97, no. 7, pp. 1166–1185, Jul. 2009.

[63] D. B. Strukov and K. K. Likharev, "CMOL FPGA: A reconfigurable architecture for hybrid digital circuits with two-terminal nanodevices," *Nanotechnol.*, vol. 16, no. 6, pp. 888–900, 2005.

[64] J. Borghetti *et al.*, "'Memristive'switches enable 'stateful'logic operations via material implication," *Nature*, vol. 464, no. 7290, pp. 873–876, 2010.

[65] H. Akinaga and H. Shima, "Resistive random access memory (RERAM) based on metal oxides," *Proc. IEEE*, vol. 98, no. 12, pp. 2237–2251, Dec. 2010.

[66] C. Li *et al.*, "Analogue signal and image processing with large memristor crossbars," *Nature Electron.*, vol. 1, no. 1, pp. 52–59, 2018. [Online]. Available: https://doi.org/10.1038/s41928-017-0002-z

[67] G. W. Burr *et al.*, "Recent progress in phase-change memory technology," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 6, no. 2, pp. 146–162, Jun. 2016.

[68] B. Govoreanu *et al.*, "10 × 10 nm2 Hf/Hfox crossbar resistive RAM with excellent performance, reliability and low-energy operation," in *Proc. Int. Electron Devices Meet.*, 2011, pp. 3161–3164.

[69] D. A. Miller, "Optics for low-energy communication inside digital processors: quantum detectors, sources, and modulators as efficient impedance converters," *Opt. Lett.*, vol. 14, no. 2, pp. 146–148, 1989.

[70] A. Gondarenko, J. S. Levy, and M. Lipson, "High confinement micron-scale silicon nitride high q ring resonator," *Opt. Exp.*, vol. 17, no. 14, pp. 11366–11370, Jul. 2009. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-17-14-11366

[71] J. F. Bauters *et al.*, "Planar waveguides with less than 0.1 db/m propagation loss fabricated with wafer bonding," *Opt. Exp.*, vol. 19, no. 24, pp. 24090–24101, Nov. 2011. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-19-24-24090

[72] H. Lee, T. Chen, J. Li, O. Painter, and K. J. Vahala, "Ultra-low-loss optical delay line on a silicon chip," *Nature Commun.*, vol. 3, 2012, Art. no. 867. [Online]. Available: https://doi.org/10.1038/ncomms1876

[73] R. Soref, "The past, present, and future of silicon photonics," *IEEE J. Sel. Topics Quantum Electron.*, vol. 12, no. 6, pp. 1678–1687, Nov./Dec. 2006.

[74] W. Bogaerts, M. Fiers, and P. Dumon, "Design challenges in silicon photonics," *IEEE J. Sel. Topics Quantum Electron.*, vol. 20, no. 4, Jul./Aug. 2014, Art. no. 8202008.

[75] J. Chiles, S. M. Buckley, S. W. Nam, R. P. Mirin, and J. M. Shainline, "Multiplanar dielectric waveguides for neural communication," in *Proc. IEEE 15th Int. Conf. Group IV Photon.*, 2018, pp. 1–2.

[76] S. Pi *et al.*, "Memristor crossbar arrays with 6-nm half-pitch and 2-nm critical dimension," *Nature Nanotechnol.*, vol. 14, no. 1, pp. 35–39, 2019. [Online]. Available: https://doi.org/10.1038/s41565-018-0302-0

[77] K. Nozaki *et al.*, "Photonic-crystal nano-photodetector with ultrasmall capacitance for on-chip light-to-voltage conversion without an amplifier," *Optica*, vol. 3, no. 5, pp. 483–492, May 2016. [Online]. Available: http://www.osapublishing.org/optica/abstract.cfm?URI=optica-3-5-483

[78] 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. Lightw. Technol.*, vol. 32, no. 21, pp. 3427–3439, Nov. 2014. [Online]. Available: http://dx.doi.org/10.1109/JLT.2014.2345652

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

[80] A. N. Tait, "Silicon photonic neural networks," Ph.D. dissertation, Dept. Elect. Eng., Princeton University, Princeton, NJ, USA, Apr. 2018.

[81] J. M. Shainline, S. M. Buckley, R. P. Mirin, and S. W. Nam, "Superconducting optoelectronic circuits for neuromorphic computing," *Phys. Rev. Appl.*, vol. 7, Mar. 2017, Art. no. 034013. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevApplied.7.034013

[82] L. Appeltant *et al.*, "Information processing using a single dynamical node as complex system," *Nature Commun.*, vol. 2, 2011 Art. no. 468. [Online]. Available: http://dx.doi.org/10.1038/ncomms1476

[83] L. Larger *et al.*, "Photonic information processing beyond turing: An optoelectronic implementation of reservoir computing," *Opt. Exp.*, vol. 20, no. 3, pp. 3241–3249, Jan. 2012. [Online]. Available: http://dx.doi.org/10.1364/OE.20.003241

[84] M. C. Soriano *et al.*, "Optoelectronic reservoir computing: tackling noise-induced performance degradation," *Opt. Exp.*, vol. 21, no. 1, pp. 12–20, Jan. 2013.

[85] K. Vandoorne *et al.*, "Experimental demonstration of reservoir computing on a silicon photonics chip," *Nature Commun*, vol. 5, 2014, Art. no. 3541.

[86] K. Vandoorne, J. Dambre, D. Verstraeten, B. Schrauwen, and P. Bienstman, "Parallel reservoir computing using optical amplifiers," *IEEE Trans. Neural Netw.*, vol. 22, no. 9, pp. 1469–1481, Sep. 2011.

[87] Y. Paquot *et al.*, "Optoelectronic reservoir computing," *Scientific Rep.*, vol. 2, 2012, Art. no. 287. [Online]. Available: http://dx.doi.org/10.1038/srep00287

[88] M. Georgas, J. Leu, B. Moss, C. Sun, and V. Stojanović, "Addressing link-level design tradeoffs for integrated photonic interconnects," in *Proc. IEEE Custom Integr. Circuits Conf.*, 2011, pp. 1–8.

[89] L. Szilagyi *et al.*, "A 53-gbit/s optical receiver frontend with 0.65 pj/bit in 28-nm bulk-CMOS," *IEEE J. Solid-State Circuits*, vol. 54, no. 3, pp. 845–855, Mar. 2019.

[90] N. C. Harris *et al.*, "Efficient, compact and low loss thermo-optic phase shifter in silicon," *Opt. Exp.*, vol. 22, no. 9, pp. 10487–10493, May 2014.

[91] C. Sun *et al.*, "A 45 nm CMOS-SOI monolithic photonics platform with bit-statistics-based resonant microring thermal tuning," *IEEE J. Solid-State Circuits*, vol. 51, no. 4, pp. 893–907, Apr. 2016.

[92] Y. Lai *et al.*, "Ultra-wide-band structural slow light," *Scientific Rep.*, vol. 8, no. 1, 2018, Art. no. 14811. [Online]. Available: https://doi.org/10.1038/s41598-018-33090-x

[93] G. J. Sharp, C. Klitis, V. Biryukova, B. Holmes, and M. Sorel, "Trimming of silicon-on-insuiator micro-ring resonators by laser irradiation," in *Proc. Conf. Lasers Electro-Opt. Eur. Eur. Quantum Electron. Conf.*, 2017, pp. 1–1.

[94] M. M. Milosevic *et al.*, "Ion implantation in silicon for trimming the operating wavelength of ring resonators," *IEEE J. Sel. Topics Quantum Electron.*, vol. 24, no. 4, Jul./Aug. 2018, Art. no. 8200107.

[95] A. P. Knights, "Device and method for post-fabrication trimming of an optical ring resonator using a dopant-based heater," U.S. Patent 9 946 027, Apr. 17, 2018.

[96] M. R. Mahmoodi and D. Strukov, "An ultra-low energy internally analog, externally digital vector-matrix multiplier based on nor flash memory technology,," in *Proc. 55th Ann. Design Automat. Conf.*, ser. DAC'18. New York, NY, USA, 2018, pp. 22-1–22-6. [Online]. Available: http://doi.acm.org/10.1145/3195970.3195989

[97] S. Cheung, T. Su, K. Okamoto, and S. Yoo, "Ultra-compact silicon photonic 512 × 512 25 GHz arrayed waveguide grating router," *IEEE J. Sel. Topics Quantum Electron.*, vol. 20, no. 4, Jul./Aug. 2014, Art. no. 8202207.

[98] J. Lu and J. Vučković, "Inverse design of nanophotonic structures using complementary convex optimization," *Opt. Exp.*, vol. 18, no. 4, pp. 3793–3804, Feb. 2010.

[99] J. Peurifoy *et al.*, "Nanophotonic particle simulation and inverse design using artificial neural networks," *Sci. Adv.*, vol. 4, no. 6, pp. 4206–4213, 2018. [Online]. Available: https://advances.sciencemag.org/content/4/6/eaar4206

[100] E. Timurdogan *et al.*, "An ultralow power athermal silicon modulator," *Nature Commun.*, vol. 5, 2014, Art. no. 4008. [Online]. Available: https://doi.org/10.1038/ncomms5008

[101] J. Johnson, "Rethinking floating point for deep learning," 2018, *arXiv:1811.01721*.

[102] J. D. Joannopoulos, P. R. Villeneuve, and S. Fan, "Photonic crystals: putting a new twist on light," *Nature*, vol. 386, no. 6621, pp. 143–149, 1997. [Online]. Available: https://doi.org/10.1038/386143a0

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

[104] A. N. Tait *et al.*, "A silicon photonic modulator neuron," *Physical Rev. Appl.*, vol. 11, no. 6, Jun. 18, 2019, Art. no. 064043.

[105] 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. Exp.*, vol. 24, no. 8, pp. 8895–8906, Apr. 2016. [Online]. Available: http://dx.doi.org/10.1364/OE.23.012758

[106] A. N. Tait *et al.*, "Microring weight banks," *IEEE J. Sel. Topics Quantum Electron.*, vol. 22, no. 6, Nov./Dec. 2016, Art. no. 5900214. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7479545

[107] A. N. Tait *et al.*, "Two-pole microring weight banks," *Opt. Lett.*, vol. 43, no. 10, pp. 2276–2279, May 2018. [Online]. Available: https://www.osapublishing.org/ol/abstract.cfm?uri=ol-43-10-2276

[108] A. N. Tait *et al.*, "Feedback control for microring weight banks," *Opt. Exp.*, vol. 26, no. 20, pp. 26422–26443, 2018. [Online]. Available: https://www.osapublishing.org/oe/abstract.cfm?uri=oe-26-20-26422

[109] 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. Topics Quantum Electron.*, vol. 19, no. 5, Sep./Oct. 2013, Art. no. 1800212. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6497478

[110] M. A. Nahmias, A. N. Tait, B. J. Shastri, T. F. de Lima, and P. R. Prucnal, "Excitable laser processing network node in hybrid silicon: Analysis and simulation," *Opt. Exp.*, vol. 23, no. 20, pp. 26800–26813, Oct. 2015. [Online]. Available: http://dx.doi.org/10.1364/OE.23.026800

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

[112] B. J. Shastri *et al.*, "Spike processing with a graphene excitable laser," *Scientific Rep.*, vol. 6, 2016, Art. no. 19126. [Online]. Available: http://dx.doi.org/10.1038/srep19126

[113] G. Katti, M. Stucchi, K. D. Meyer, and W. Dehaene, "Electrical modeling and characterization of through silicon via for three-dimensional ics," *IEEE Trans. Electron Devices*, vol. 57, no. 1, pp. 256–262, Jan. 2010.

[114] M. Jurczak, N. Collaert, A. Veloso, T. Hoffmann, and S. Biesemans, "Review of finfet technology," in *Proc. IEEE Int. SOI Conf.*, 2009, pp. 1–4.

[115] T. F. de Lima *et al.*, "Noise analysis of photonic modulator neurons," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 1, pp. 1–9, 2020.

[116] S. Xiao, M. H. Khan, H. Shen, and M. Qi, "Multiple-channel silicon micro-resonator based filters for WDM applications," *Opt. Exp.*, vol. 15, no. 12, pp. 7489–7498, Jun. 2007. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-15-12-7489

[117] I. Ozkaya *et al.*, "A 64-gb/s 1.4-pj/b NRZ optical receiver data-path in 14-nm CMOS finfet," *IEEE J. Solid-State Circuits*, vol. 52, no. 12, pp. 3458–3473, Dec. 2017.

[118] C. Sun *et al.*, "Dsent—a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling," in *Proc. IEEE/ACM 6th Int. Symp. Netw.-on-Chip*, 2012, pp. 201–210.

[119] A. H. Atabaki *et al.*, "Integrating photonics with silicon nanoelectronics for the next generation of systems on a chip," *Nature*, vol. 556, no. 7701, pp. 349–354, 2018. [Online]. Available: https://doi.org/10.1038/s41586-018-0028-z

[120] G.-S. Jeong, W. Bae, and D.-K. Jeong, "Review of CMOS integrated circuit technologies for high-speed photo-detection," *Sensors*, vol. 17, no. 9, pp. 1962–2002, 2017.

[121] V. J. Sorger, N. D. Lanzillotti-Kimura, R.-M. Ma, and X. Zhang, "Ultra-compact silicon nanophotonic modulator with broadband response," *Nanophotonics*, vol. 1, no. 1, pp. 17–22, 2012.

[122] V. J. Sorger *et al.*, "Scaling vectors of attojoule per bit modulators," *J. Opt.*, vol. 20, no. 1, 2017, Art. no. 014012.

[123] R. Amin *et al.*, "Attojoule-efficient graphene optical modulators," *Appl. Opt.*, vol. 57, no. 18, pp. D130–D140, 2018. [Online]. Available: http://ao.osa.org/abstract.cfm?URI=ao-57-18-D130

[124] "Groq," Nov. 2017. Accessed: May 11, 2018. [Online]. Available: https://groq.com/

[125] R. Smith, "Nvidia volta unveiled: Gv100 GPU and Tesla v100 accelerator announced," May 2017. [Online]. Available: https://www.anandtech.com/show/11367/nvidia-volta-unveiled-gv100-gpu-and-tesla-v100-accelerator-announced

[126] S. Knowles, "Scalable silicon compute," in *Proc. NIPS Workshop Deep Learn. Supercomputer Scale*, Dec. 2017.

[127] P. Wijesinghe, A. Ankit, A. Sengupta, and K. Roy, "An all-memristor deep spiking neural computing system: A step toward realizing the low-power stochastic brain," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 5, pp. 345–358, Oct. 2018.

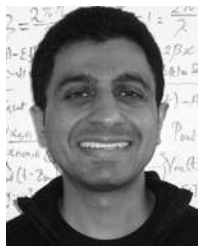
Mitchell A. Nahmias received the B.S. (Hons.) degree in electrical engineering with a Certificate in Engineering Physics and the M.A. degree in electrical engineering in 2012 and 2014, respectively, from Princeton University, Princeton, NJ, USA, where he is currently working toward the Ph.D. degree with the Princeton Lightwave Communications Laboratory. He is a member of the Princeton Lightwave Communications Laboratory. He has authored or coauthored more than 60 journal papers, has been cited more than 1000 times, and is an inventor on several patents. His research interests include photonic integrated circuits, unconventional computing, and neuromorphic photonics. He was the recipient of the Best Engineering Physics Independent Work Award (2012), the National Science Foundation Graduate Research Fellowship, the Best Paper Award at the IEEE Photonics Conference 2014 (third place), and the Best Paper Award at the 2015 IEEE Photonics Society Summer Topicals Meeting Series (first place). He is also a contributing author to the textbook, *Neuromorphic Photonics* (2017).

Thomas Ferreira de Lima received the bachelor's degree and the Ingénieur Polytechnicien master's degree from Ecole Polytechnique, Palaiseau, France, with a focus on Physics for Optics and Nanosciences. He is currently working toward the Ph.D. degree in electrical engineering from the Lightwave Communications Group, Department of Electrical Engineering, Princeton University, Princeton, NJ, USA. He has authored or coauthored more than 40 journal or conference papers, contributes to four major opensource projects, and is a contributing author to the textbook *Neuromorphic Photonics* (2017). His research interests include integrated photonic systems, nonlinear signal processing with photonic devices, spike-timing-based processing, ultrafast cognitive computing, and dynamical lightmatter neuro-inspired learning and computing.

Alexander N. Tait received the B.S.Eng. (Hons.) degree in electrical engineering in 2012 from Princeton University, Princeton, NJ, USA, where he received the Ph.D. degree from the Lightwave Communications Research Laboratory, Department of Electrical Engineering, under the supervision of Prof. P. Prucnal. He has authored nine refereed papers and a book chapter, presented research at 13 technical conferences, and contributed to the textbook *Neuromorphic Photonics* (2017). His research interests include silicon photonics, optical signal processing, optical networks, and neuromorphic engineering.

Dr. Tait is a recipient of the National Science Foundation Graduate Research Fellowship and is a Student Member of the IEEE Photonics Society and the Optical Society of America. He is the recipient of the Award for Excellence from the Princeton School of Engineering and Applied Science, the Optical Engineering Award of Excellence from the Princeton Department of Electrical Engineering, the Best Student Paper Award at the 2016 IEEE Summer Topicals Meeting Series, and the Class of 1883 Writing Prize from the Princeton Department of English.

Hsuan-Tung Peng received the B.S. degree in physics from National Taiwan University, Taipei, Taiwan, in 2015 and the M.A. degree in electrical engineering in 2018 from Princeton University, Princeton, NJ, USA, where he is currently working toward the Ph.D. degree. His current research interests include neuromorphic photonics, photonic integrated circuits, and optical signal processing.



Bhavin J. Shastri received the Ph.D. degree in electrical engineering (photonics) from McGill University, Montreal, QC, Canada, in 2012. He is an Assistant Professor of Engineering Physics with Queen's University, Canada. He is a coauthor of the book *Neuromorphic Photonics* (Taylor & Francis, CRC Press). He was an Associate Research Scholar (2016–2018) and Banting and NSERC Postdoctoral Fellow (2012–2016) with Princeton University. He is a recipient of the 2014 Banting Postdoctoral Fellowship from the Government of Canada, the 2012 D.W. Ambridge Prize for the top graduating Ph.D. student, an IEEE Photonics Society 2011 Graduate Student Fellowship, a 2011 NSERC Postdoctoral Fellowship, a 2011 SPIE Scholarship in Optics and Photonics, a 2008 NSERC Alexander Graham Bell Canada Graduate Scholarship, including the Best Student Paper Awards at the 2014 IEEE Photonics Conference, 2010 IEEE Midwest Symposium on Circuits and Systems, the 2004 IEEE Computer Society Lance Stafford Larson Outstanding Student Award, and the 2003 IEEE Canada Life Member Award.

Paul R. Prucnal received the A.B. *(summa cum laude)* degree in mathematics and physics from Bowdoin College, Brunswick, ME, USA, and the M.S., M.Phil. and Ph.D. degrees in electrical engineering from Columbia University, New York, NY, USA. He was the faculty member with Columbia University, where, as a member of the Columbia Radiation Laboratory, he performed groundbreaking work in OCDMA and self-routed photonic switching. In 1988, he was the faculty member with Princeton University. His research on optical CDMA initiated a new research field in which more than 1000 papers have since been published, exploring applications ranging from information security to communication speed and bandwidth. In 1993, he invented the "Terahertz Optical Asymmetric Demultiplexer," the first optical switch capable of processing terabit per second (Tb/s) pulse trains. He is author of the book *Neuromorphic Photonics* (CRC Press, 2017) and editor of the book *Optical Code Division Multiple Access: Fundamentals and Applications* (CRC Press, 2018). He was an Area Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS. He has authored or coauthored more than 350 journal articles and book chapters and holds 28 U.S. patents. He is a Life Fellow of the Institute of Electrical and Electronics Engineers, the Optical Society of America and the National Academy of Inventors, and a member of honor societies including Phi Beta Kappa and Sigma Xi. He was the recipient of the 1990 Rudolf Kingslake Medal for his paper entitled "Self-routing photonic switching with optically-processed control, received the Gold Medal from the Faculty of Mathematics, Physics and Informatics at the Comenius University, for leadership in the field of Optics 2006 and received the multiple teaching awards with Princeton, including the E-Council Lifetime Achievement Award for Excellence in Teaching, the School of Engineering and Applied Science Distinguished Teacher Award, The President's Award for Distinguished Teaching. He has been instrumental in founding the field of Neuromorphic Photonics and developing the "photonic neuron," a high speed optical computing device modeled on neural networks, as well as integrated optical circuits to improve wireless signal quality by cancelling radio interference.