Coherent Photonic Crossbar Arrays for Large-Scale Matrix-Matrix Multiplication

Nathan Youngblood, *Member, IEEE*

***Abstract*—Advances in deep learning research over the past decade have been enabled by an increasingly unsustainable demand for compute power. This trend has dramatically outpaced the slowing growth in the performance and efficiency of electronic computing hardware. Here, we propose a hybrid photonic-electronic computing architecture which leverages a photonic crossbar array and homodyne detection to perform large-scale coherent matrix-matrix multiplication. This approach bypasses the requirements of high-speed electronic readout and frequent reprogramming of photonic weights which significantly reduces energy consumption and latency in the limit of large matrices—two major factors limiting efficiency for many analog computing approaches.**

***Index Terms*—Artificial intelligence, Neural network hardware, Analog computers, Optical computing, Analog processing circuits**

I. Introduction

FEW technological innovations have been as wide-reaching or impactful within the last decade as the field of deep learning. Advances in AI through the development of Deep Neural Networks (DNNs) have transformed a broad range of disciplines such as medical imaging and diagnostics, materials discovery, autonomous navigation, and natural language processing. While the positive societal impact of DNNs have been thrilling to witness, they come with a voracious appetite for computing resources—an increasingly unsustainable paradigm. Thus, the generality and accuracy of DNNs, which fundamentally scales with the amount of training data and available computation, is also their Achilles' heel [1], [2]. While graphics processors (GPUs) have historically enabled continued advances in deep learning, this is due to their suitability for distributed training of DNNs across large clusters of individual nodes, rather than significantly improving computational throughput of a single node. This distributed approach to deep learning development can easily take several months, cost millions of dollars in computing services, and expel hundreds of tons of CO_2 to optimally train a complex DNN [3]. With these current trends, continued progress in the field of deep learning using conventional computing hardware is both economically and environmentally unsustainable.

Computing in the optical domain is one approach to overcome the energy-bandwidth trade-off intrinsic to electronic deep learning hardware [4] and has already shown significant experimental progress in the last few years. Various photonic architectures, such as cascaded Mach-Zehnder interferometers [5], [6], in-memory computing [7], [8], reconfigurable metasurfaces [9], frequency comb shaping [10], and neuromorphic computing [11]–[13] have all demonstrated the feasibility of analog computing in the photonic domain. However, the majority of these approaches rely on fixed photonic weights and high-speed photodetectors and analog-to-digital converters (ADCs) to convert the results of an optical matrix-vector multiplication (MVM) back into the digital domain for further processing. Therefore, the opto-electronic readout circuitry *must operate at the same speed* as the electro-optical modulators at the input, and thus place an upper limit on the overall throughput and energy efficiency of the photonic accelerator. Additionally, unlike digital-to-analog conversion which can be highly efficient [14], conversion from the analog to digital domain is nontrivial and energy consumption scales with the operation frequency of the ADC [15], [16]. Therefore, the overall energy consumption of the readout circuitry—a large fraction of the overall power consumption for many analog computing systems [17], [18]—roughly scales as $\sim N \times f$, where N is the number of optical output channels and f is the ADC operating speed.

To address this challenge, Hamerly et al. [19] recently proposed a novel method for achieving large-scale, multiply-accumulate operations in the optical domain via homodyne detection. This approach has several benefits: **(1)** It decouples the modulation frequency of the optical inputs from the speed of the electrical readout circuitry. **(2)** The differential nature of homodyne detection enables both positive and negative numbers (i.e., $\mathbb{R} \in [-1, 1]$) to be implemented by controlling the phase and amplitude of two coherent optical inputs. **(3)** Homodyne detection removes common-mode noise which allows one to use extremely low optical powers which approach the standard quantum limit determined by the photodetector shot noise. **(4)** Finally, by multiplexing multiply-accumulate operations in space and time, the system is scalable to very large matrix operations. However, in spite these advantages, experimental implementation using free space optics is extremely challenging since the optical path of the two beams must be both *spatially and temporally coherent*. Additionally, the spatial light modulators (SLMs) needed to encode matrix values in this free space architecture are currently limited to modulation speeds of ~1 kHz or less.

Here, we propose an integrated photonic platform to implement large-scale matrix-matrix multiplication (MMM) which overcomes both phase-matching and modulation challenges of a free space approach. Leveraging prior

Manuscript received October 22, 2021. This work was supported in part by the U.S. National Science Foundation under Grants ECCS-2028624, DMR-2003325, and CISE-2105972 as well as support through Pitt Momentum Funds at the University of Pittsburgh.

N. Youngblood is with the Department of Electrical and Computer Engineering, Swanson School of Engineering, University of Pittsburgh, Pittsburgh, PA 15261 USA (e-mail: nathan.youngblood@pitt.edu).

experimental demonstrations of large scale photonic phased-arrays [20], nanophotonic LIDAR [21], [22], and our in-memory photonic computing architecture [7], [8], we use an array of waveguide crossings, directional couplers, and balanced photodetection to achieve fan-out and coherent interference of optical signals on-chip. Our design (illustrated in Fig. 1) uses robust components which are well suited for large scale fabrication in a photonics foundry. In addition to de-coupling the requirement for high-speed electrical read-out from the data modulation rate, we also encode both matrices in the optical input signals, thus removing the costly reprogramming step required by many other photonics approaches. In **Sections II** and **III**, we present an approach for designing an integrated photonic matrix-matrix multiplier and analyze the effects of system noise on computational precision. We then estimate the energy consumption of our platform in **Section IV**. Finally, we compare the overall energy consumption and latency of MMM operations with other computing approaches in the optical and electronic domains and propose a mixed architecture approach to computing (**Sections V** and **VI**).

II. Design of Integrated Photonic Matrix-Matrix Multiplier

A. Background

The multiplication of two matrices $A_{m\times n}$ and $B_{n\times p}$ is simply the result of mp dot-products between the row vectors of matrix A and the column vectors of matrix B. Thus, each element in the resulting matrix of size $m \times p$ can be written as:

$$(AB)_{ij} = \sum_{r=1}^{n} a_{ir} b_{rj} = \vec{a}_i \cdot \vec{b}_j \qquad (1)$$

where $\vec{a}_i$ is the i^{th} row of A and $\vec{b}_j$ is the j^{th} column of B. If the above summation of products between a_{ir} and b_{rj} are multiplexed in time and scaled such that $|a_{ir}|, |b_{rj}| \in [0,1]$, this dot product can be computed optically using a balanced homodyne detection scheme [19] as illustrated in Fig. 1a. In this approach, vectors $\vec{a}_i$ and $\vec{b}_j$ from equation *(1)* are encoded in the time-varying amplitudes of two interfering electric fields $\vec{E}_a(t) = \hat{a}E_a(t)e^{i\varphi_a(t)}$ and $\vec{E}_b(t) = \hat{b}E_b(t)e^{i\varphi_b(t)}$ incident on a 3dB directional coupler (or 50:50 beam splitter [19]). Due to

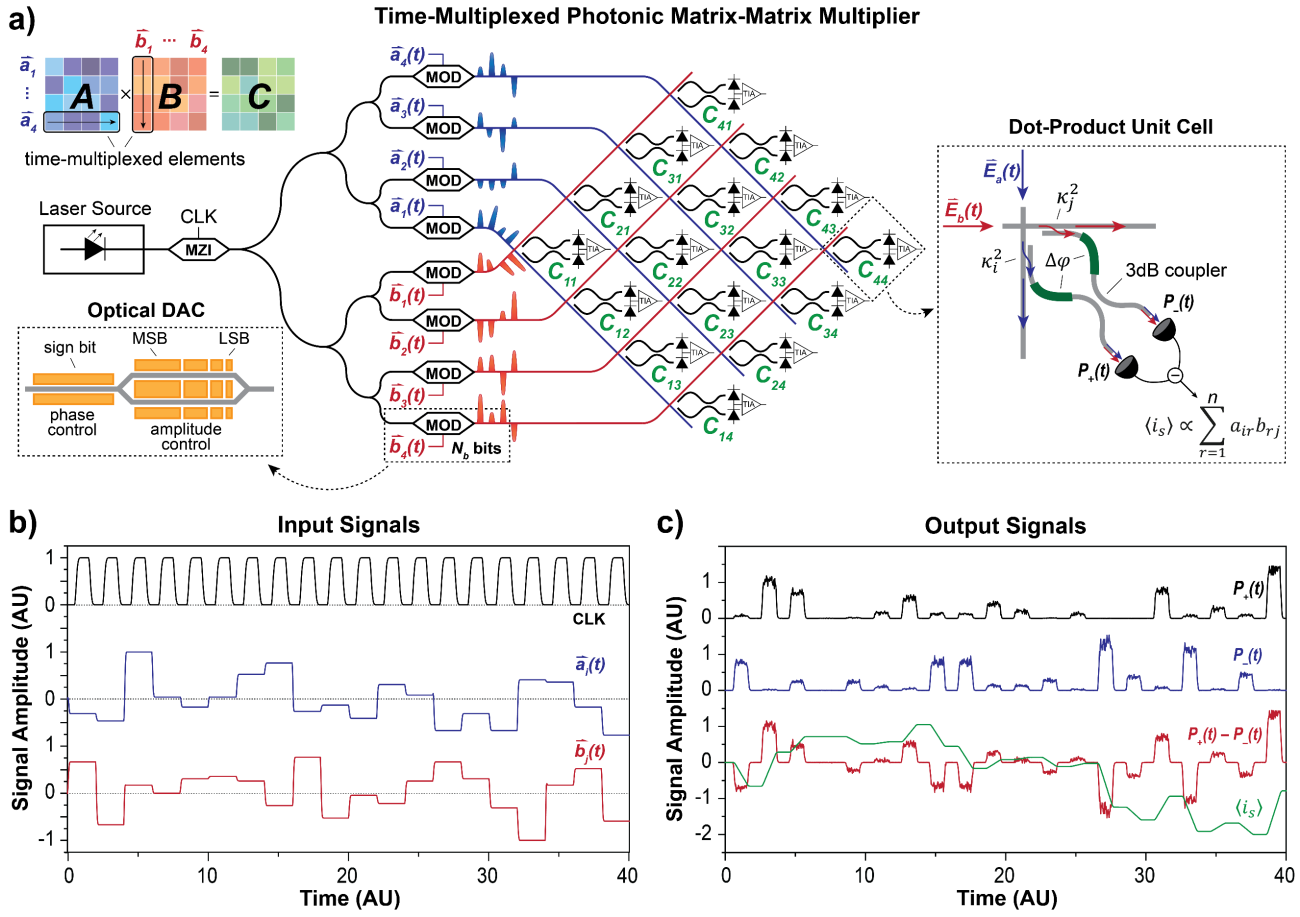


Fig. 1. Time-multiplexed photonic matrix-matrix multiplier (MMM) architecture. **a)** Schematic of a fully integrated photonic MMM platform capable of multiplying two 4×4 matrices. Each input of the coherent crossbar array is modulated in both amplitude and phase. The optical digital-to-analog converter (lower left) uses a segmented modulator to directly encode information from the digital electronic to optical analog domain (i.e., most significant bit "MSB" has longest segment and least significant bit "LSB" has shortest segment for binary code weighted scheme [30]). The intersection of each crossbar contains a photoelectric multiplier unit as proposed by Hamerly et al. [19] to achieve the dot product between two time-multiplexed optical signals (lower right). **b)** Simulated input and **c)** output signals of a single dot-product unit cell using Lumerical INTERCONNECT. The signs of elements in vectors $\vec{a}_i(t)$ and $\vec{b}_j(t)$ are encoded in the optical phase while the magnitudes are encoded in the field amplitudes.

conservation of energy, the cross-coupled (or reflected) beam will experience a π/2-phase shift with respect to the transmitted beam. Assuming $\vec{\boldsymbol{E}}_a(t)$ and $\vec{\boldsymbol{E}}_b(t)$ are temporally and spatially coherent (i.e., phase-matched and single mode) and of the same polarization, the optical signal measured by the photodetectors at the two output ports of the 3-dB coupler can be written as:

$$P_+(t) = \frac{1}{2}(|E_a(t)|^2 + |E_b(t)|^2) + \mathrm{Re}[E_a^*(t)E_b(t)]\sin(\Delta\varphi) \quad (2)$$

$$P_-(t) = \frac{1}{2}(|E_a(t)|^2 + |E_b(t)|^2) - \mathrm{Re}[E_a^*(t)E_b(t)]\sin(\Delta\varphi) \quad (3)$$

where $P_\pm(t)$ is the optical power incident on the two photodetectors and $\Delta\varphi$ is the relative phase difference between $\vec{\boldsymbol{E}}_a(t)$ and $\vec{\boldsymbol{E}}_b(t)$. From equations *(2)* and *(3)* we can see that the first term is simply proportional to the optical power of the two input signals, while the second term contains the product of the field amplitudes which differ by a sign. To convert optical power to photocurrent, we can multiply by the photodetector's responsivity, $R = \frac{\eta e}{h\nu}$, where η is the quantum efficiency of the detector, e is the charge of an electron, and $h\nu$ is the photon energy. Taking the difference of equations *(2)* and *(3)* allows us to cancel the first term and only keep the second using balanced photodetection:

$$\langle i_s\rangle = \frac{1}{n\tau}\frac{\eta e}{h\nu}\int_0^{n\tau}\left(P_+(t) - P_-(t)\right)dt = \frac{2}{n\tau}\frac{\eta e}{h\nu}\int_0^{n\tau} E_a(t)E_b(t)\sin(\Delta\varphi(t) + \Delta\varphi')\,dt \quad (4)$$

$$\langle i_s\rangle \propto \sum_{r=1}^{n} a_{ir}b_{rj} \quad (5)$$

In the above equation, $\langle i_s\rangle$ is the difference signal measured by the homodyne setup, $n\tau$ is the total duration of n pulses of period $\tau = 1/f_{mod}$, and we have assumed the fields $E_a(t)$ and $E_b(t)$ are real. We note that $\Delta\varphi = \Delta\varphi(t) + \Delta\varphi'$ contains both a time-dependent phase difference $\Delta\varphi(t) = \varphi_a(t) - \varphi_b(t)$ and a fixed phase difference ($\Delta\varphi'$) based on the relative optical delay between the source of $\vec{\boldsymbol{E}}_a(t)$ and $\vec{\boldsymbol{E}}_b(t)$ and the two input ports of the 3dB directional coupler. Assuming $\Delta\varphi(t) = q\pi$ (where q is an integer), the difference signal will be maximized by setting $\Delta\varphi' = \pm\pi/2$. This can be accomplished with thermo-optic phase tuning [5], [23], [24], but can also be accomplished using methods which require zero static power, such as laser trimming [25] or low-loss phase change materials [26]–[28]. The phase tuning required to set $\Delta\varphi' = \pi/2$ can be accurately determined experimentally by maximizing $\langle i_s\rangle$ while both $E_a(t)$ and $E_b(t)$ are held constant and the time-dependent phase terms are set to $\varphi_a(t) = \varphi_b(t) = 0$. Once $\Delta\varphi'$ has been trimmed to the correct relative phase difference, the amplitude and phase modulators at each of the inputs can be modulated such that equation *(5)* is satisfied (see the following section).

B. Encoding real numbers in the optical field

To compute the dot-product between two vectors, the vector elements must be encoded in the optical fields. Using a balanced homodyne detection approach as detailed above, it is possible to encode all real-value numbers in the range $[-1, 1]$ by modulating both the phase and amplitude of the optical signals. Amplitude modulation can easily be achieved with integrated high-frequency modulators (such as a silicon plasma-dispersion MZI or microring modulators) which are readily available from most photonics foundries. It is important to note that while microring modulators are desirable for efficient and compact modulation, they also impart a nonlinear phase on the modulated signal which would require special compensation to correct (e.g., two cascaded ring modulators [29]). On the other hand, a balanced MZI modulator based on carrier depletion can be modulated with complementary voltages in both arms and therefore minimize phase modulation of the output optical signal. Additionally, both MZI and ring modulators have been demonstrated with built-in DACs which can efficiently convert a digital input into an amplitude modulated optical output [14], [30], [31]. For example, Moazeni et al. [14] demonstrated a highly linear 4-bit optical DAC capable of 40 Gb/s and with an efficiency of 42 fJ/bit using a segmented silicon microring modulator. This approach allows extremely high-speed electro-optical and digital-to-analog conversion without additional circuitry which would reduce the overall efficiency of our optical computing approach.

From equation *(4)*, we can see that the homodyne signal is proportional to $\sin(\Delta\varphi(t) + \Delta\varphi')$, where $\Delta\varphi' = \pm\pi/2$ does not vary with the optical signal. Therefore, by modulating $\varphi_a(t)$ and $\varphi_b(t)$ to either 0 or π, we can encode both positive and negative numbers. Practically, this can be achieved by cascading an additional phase modulator with each amplitude modulator (see "Optical DAC" in Fig. 1a). Adding this phase term increases the total number of symbols we can encode by 2× without placing additional requirements on the amplitude modulator (e.g., 5-bit signed integers in the case of a PAM-16 modulator in series with a phase modulator). To minimize the effects of the rise and fall times of the amplitude and phase modulators, we add an additional intensity modulator immediately after the optical source to globally gate the optical signal during transitions ("CLK" signal in Fig. 1b).

C. Directional coupler design

Next, we discuss a method to ensure equal power distribution to each dot product unit cell within the array using a photonic crossbar architecture for fan-out [8]. Fig. 2a illustrates the parameters which define the cross-coupling coefficients (κ_n^2) and the transmission of a single directional coupler (η_{DC}) and waveguide crossing (η_x). For simplicity, we assume the insertion loss for the directional coupler is independent of coupling length (i.e., absorption and scattering in the coupling region are negligible compared to mode-mismatch). In order to have equal power distribution from the input waveguide to each unit cell in a given row, the following must be true:

$$|E_0|^2\eta_x\kappa_1^2 = |E_0|^2\eta_x^2\eta_{DC}(1-\kappa_1^2)\kappa_2^2 = |E_0|^2\eta_x^3\eta_{DC}^2(1-\kappa_1^2)(1-\kappa_2^2)\kappa_3^2 \quad (6)$$

In general, this leads to the following relationship between two neighboring directional couplers:

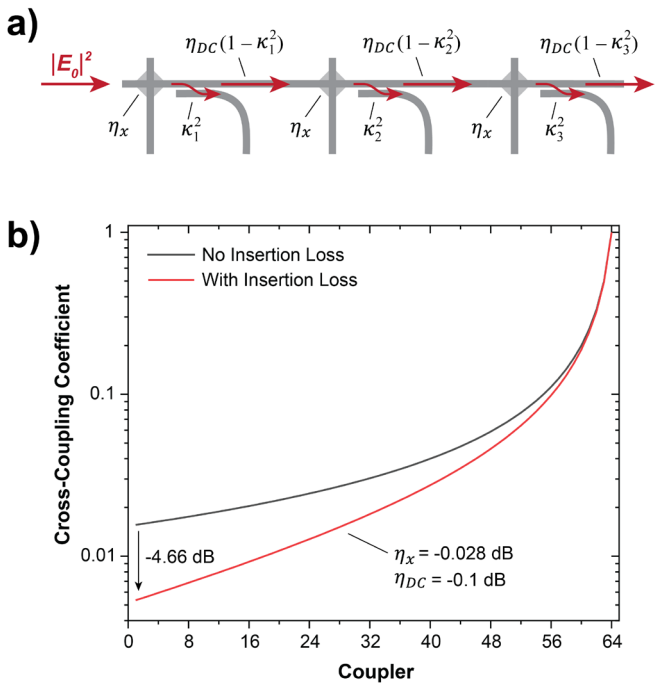


Fig. 2. Loss-compensated fan-out design. **a)** Illustration of first three unit cells of the crossbar array in a given row. Insertion losses of the crossbar and directional couplers (η_x and η_{DC}, respectively) are accounted for in the cross-coupling coefficients (κ_i^2). **b)** Calculated cross-coupling coefficients for a 64×64 crossbar array using experimentally measured insertion losses from [8], [32].

$$\kappa_n^2 = \frac{\kappa_{n+1}^2}{\frac{1}{\eta_{IL}} + \kappa_{n+1}^2} \tag{7}$$

where $\eta_{IL} = \eta_x \eta_{DC}$ is the insertion loss of each unit cell and can be modified to include the waveguide loss as well (e.g., $\eta_{IL} = \eta_x \eta_{DC} e^{-\alpha_{loss} L}$). The above equation also holds true for equal power distribution along a column. If the total number of unit cells in a given row or column is N, then we can set the final cross coupling term to $\kappa_N^2 = 1$ and solve for all previous coupling coefficients recursively. We can also choose $\kappa_N^2 = 0.5$ if we wish to use the output of the final coupler through port to calibrate the average insertion loss of a given row or column. This design choice is useful to experimentally determine the average unit cell transmission $\overline{\eta_{IL}}$ after fabrication. The coupling coefficients for an ideal array ($\eta_{IL} = 1$) and an array with realistic loss [8], [32] are shown in Fig. 2b for an array with 64 unit cells in a row.

D. Matrix-matrix multiplication

Assume we have a photonic crossbar array as described above with $k \times k$ unit cells and wish to perform a matrix-matrix multiplication between $A_{m\times n}$ and $B_{n\times p}$. Using our proposed crossbar architecture, each unit cell of the crossbar performs the dot product $(AB)_{ij} = \vec{a}_i \cdot \vec{b}_j$, where i and j are the row and column index of the unit cell (Fig. 1a). From equation *(4)*, we can see that the time required to perform each dot product will be dependent on the modulation speed and the number of elements in the vectors $\vec{a}_i$ and $\vec{b}_j$. However, the strength of our approach lies in the fact that we can perform k^2 dot products in parallel. Thus, if $k \geq m, p$, the operation $A \times B = C$ has time complexity of $O(n)$. Compared to matrix multiplication in the digital domain which scales between $O(n^3)$ and $O(n^{2.373})$ for two square matrices of size $n \times n$ [33]–[35], the linear scaling of our approach demonstrates the significant speed advantage of computing in the analog domain. For the case of both GPUs and tensor processing units (TPUs), latency is reduced from $O(n^3)$ to $\sim O(n)$ by significantly increasing the parallelism of the hardware and data pipeline. However, these parallel digital approaches do not overcome the $O(n^{2.373})$ lower bound on computational complexity for matrix-matrix multiplication.

It is important to note that while the compute time of matrix-vector operations scale as $O(1)$ for both optical and electrical in-memory computing approaches [8], the output of our crossbar array is a full $k \times k$ matrix rather than a single vector of length k. Thus, the operation $A_{m\times n} \times B_{n\times p}$ scales as $O(p)$ for an in-memory architecture where $A_{m\times n}$ is a memory array of fixed weights. Frequency multiplexing approaches demonstrated in both the optical [8], [10], [36] and electrical domains [37] can reduce this to $O\left(\left\lceil \frac{p}{d} \right\rceil\right)$, where d is the number of frequency channels used simultaneously.

In the more likely scenario that $m, p > k$, the time complexity becomes $\sim O\left(n \left\lceil \frac{m}{k} \right\rceil \left\lceil \frac{p}{k} \right\rceil\right)$ for a single crossbar array. In this case, we have subdivided $A_{m\times n} \times B_{n\times p}$ into $\left\lceil \frac{m}{k} \right\rceil \left\lceil \frac{p}{k} \right\rceil$ sequential operations of size $A_{k\times n} \times B_{n\times k}$ to match the dimensions of our photonic crossbar (illustrated in Fig. 4b). Since these operations are independent of one another, they can be parallelized across multiple crossbar arrays to reduce the time complexity back to $O(n)$. Note that unlike a fixed-matrix approach which places an upper limit of $n \leq k$ for a $k \times k$ array of weights, we are encoding $k \times n$ weights in the time-domain such that n is no longer limited by physical hardware (i.e., $n \gg k$). This has significant implications on both the compute efficiency and latency which we explore in more detail in **Section V**.

III. Noise Analysis:

The computational precision of any analog computing system is fundamentally limited by the signal-to-noise ratio (SNR). The minimum acceptable SNR is highly dependent on the application, though neural networks in general seem to be relatively robust to unstructured noise [38] (and can even benefit from added noise in the case of limited precision [39]). In the case of analog computing systems that are applied to machine learning problems, using fixed precision arithmetic is a logical choice [39], [40]. Therefore, if we require an output precision of N_b bits, we can define the minimum SNR of our system to be:

$$\text{SNR}^2 = 2^{2N_b} = \frac{\langle i_s^2 \rangle}{2e(\langle i_{SN} \rangle + \langle i_D \rangle)\Delta f + \langle i_{RN}^2 \rangle} \tag{8}$$

where $\langle i_s^2 \rangle$ is the mean square value of the measured homodyne photocurrent, $\langle i_{SN} \rangle$ is the photocurrent due to photon shot noise, $\langle i_D \rangle$ is the dark current of the photodetector, Δf is the

bandwidth of the read-out circuitry, and $\langle i_{RN}^2 \rangle$ is the noise of the read-out circuitry (including Johnson noise, $1/f$ noise from amplifier, etc.). If we assume that the measurement is limited by shot noise, then $\langle i_{SN} \rangle \gg \langle i_D \rangle$ and $\frac{\langle i_{RN}^2 \rangle}{2e\Delta f}$. This is reasonable in the case of well-designed read-out circuitry and for $i_D \ll \frac{\eta e}{h\nu}\bar{P}_{\pm}$, where $\bar{P}_{\pm}$ is the average power incident on each photodetector. The photocurrent due to shot noise can be written as:

$$\langle i_{SN} \rangle = \frac{1}{n\tau}\frac{\eta e}{h\nu}\int_0^{n\tau} \big(P_+(t) + P_-(t)\big)\, dt$$
$$= \frac{1}{n\tau}\frac{\eta e}{h\nu}\int_0^{n\tau} (|E_a(t)|^2 + |E_b(t)|^2)\, dt$$

$$\langle i_{SN} \rangle = \frac{\eta e}{h\nu}(\bar{P}_a + \bar{P}_b) \qquad (9)$$

where $\bar{P}_a$ and $\bar{P}_b$ are the time-averaged optical powers of the two input signals. The photocurrent due to optical shot noise is therefore dependent on the total optical power used to compute the dot product. Combining equations *(4)*, *(8)*, and *(9)*, we have the following expression:

$$\frac{h\nu}{2\eta}(\bar{P}_a + \bar{P}_b)\Delta f \cdot 2^{2N_b} = \left[\int_0^{n\tau} \frac{E_a(t)E_b(t)}{n\tau} dt\right]^2 \qquad (10)$$

where we have removed the term $\sin(\Delta\varphi(t) + \Delta\varphi')$ by setting $\Delta\varphi' = \pm\pi/2$ and requiring $\Delta\varphi(t) = 0$ or π. This is equivalent to restricting the normalized electric field amplitude to $\mathbb{R}[-1, 1]$ which is the real number encoding system we have defined in **Section II**. We can also assume that by modulating the intensity of the optical source using a clock signal, we can mitigate any transition effects due to modulating $E_a(t)$ and $E_b(t)$ such that their values are constant over the duration of a single pulse (see simulation results of Fig. 1b-c). Thus, $E_a(t)$ and $E_b(t)$ can be represented by the discrete variables a_i and b_i normalized by the maximum field amplitude such that the integral in equation *(10)* becomes a summation:

$$\left[\int_0^{n\tau} \frac{E_a(t)E_b(t)}{n\tau} dt\right]^2 = \max(|E_a|^2|E_b|^2)\left[\frac{1}{n}\sum_{i=1}^{n} a_i b_i\right]^2 \qquad (11)$$

The distribution of the discrete variables a_i and b_i will have a significant impact on the SNR we measure at the output. If we restrict $a_i, b_i \in \mathbb{R}[0, 1]$, the product of $a_i b_i$ will always be a positive value. Thus, assuming a_i and b_i are independent random variables with a mean value of $\bar{a}_i = \bar{b}_i = 0.5$, the expected value of equation *(11)* is:

$$\mathrm{E}\left(\max(|E_a|^2|E_b|^2)\left[\frac{1}{N'}\sum_{i=1}^{N'} a_i b_i\right]^2\right) = \bar{P}_a\bar{P}_b \qquad (12)$$

where we have replaced $\max\left(|E_{a,b}|^2\right) = 4\bar{P}_{a,b}$, which is the average optical power in each signal if $\bar{a}_i = \bar{b}_i = 0.5$. The SNR is maximized when $\bar{P}_a = \bar{P}_b$. Therefore, the minimum average optical power required to resolve the dot product of two vectors with positive, random inputs will be:

$$\bar{P}_{min} = \frac{h\nu}{\eta}\cdot\frac{f_{mod}}{n}\cdot 2^{2N_b} \quad (0 \le a_i b_i \le 1) \qquad (13)$$

An important observation of equation *(13)* is that the minimum optical power is proportional to the measurement bandwidth $\Delta f = f_{mod}/n$. Therefore, a longer integration time (longer input vector) will require less optical power per multiply-accumulate (MAC) operation. If we solve for the average optical energy per MAC operation, we find:

$$\bar{E}_{MAC} = \frac{\bar{P}_{min}}{f_{mod}} = \frac{h\nu}{\eta}\cdot\frac{2^{2N_b}}{n} \quad (0 \le a_i b_i \le 1) \qquad (14)$$

Similar to the case of electronic crossbar arrays [40], the total noise limited optical energy required to compute the dot product $\vec{a}_i \cdot \vec{b}_j$ *does not depend on the input vector size* for fixed precision arithmetic. It is helpful to compare the derived minimum optical power in equation *(14)* to that of n incoherent MAC operations using a single photodetector. Assuming input vector $\vec{a}$ is encoded on the optical power and $\vec{b}$ on the optical transmission of the network (e.g., microring resonators or optical phase-change memory [8], [41]) and $\bar{a}_i = \bar{b}_i = 0.5$, equations *(13)* and *(14)* become:

$$\bar{P}_{min} = \frac{4h\nu}{\eta}\frac{f_{mod}}{n}\cdot 2^{2N_b},$$
$$\bar{E}_{MAC} = \frac{4h\nu}{\eta}\cdot\frac{2^{2N_b}}{n} \quad (0 \le a_i b_i \le 1) \qquad (15)$$

which is 4× larger than the coherent case. The reason for this is twofold. First, we have a 2× advantage in SNR using homodyne detection [42] and secondly, we are performing multiplication using the optical field rather than the optical intensity resulting in an average 2× greater contribution to the signal photocurrent compared to the shot noise. However, for analog computing approaches the optical power is typically dwarfed by the power consumption of the readout electronics (especially the ADC) which scales approximately linearly with the sampling rate [15]. Thus, reducing the ADC operation frequency by $1/n$ is likely to result in the largest energy savings of our proposed approach. We note that equation *(15)* is a factor of 4× larger than the lower bound for an incoherent photonic MAC architecture as derived by Nahmias et al. [43] (note that $2\bar{E}_{MAC} = E_{MAC(O)}$ in *Eq. (12)* of [43]). This is because we wish to resolve the expected value of two random input vectors to N_b bits of precision, rather than the maximum signal possible (i.e., $a_i = b_i = 1$ for all i) which is of trivial interest computationally in most cases (for a more detailed analysis of the impact of random variable distributions on SNR and power consumption, see [44]).

If we make full use of both phase and intensity modulation, a_i and b_i can be both positive and negative such that the product $a_i b_i \in \mathbb{R}[-1, 1]$. For the case of deep neural networks, we can assume that the data passing between layers is positive after the activation function (e.g., ReLU, softmax, etc.), while the connectivity matrix is normally distributed within $\mathbb{R}[-1, 1]$ with a mean of zero ($b_i \sim N(0, \sigma_b)$). From the law of expectations, the average product of $\overline{a_i b_i} = \bar{a}_i\bar{b}_i = 0$ and our signal $\langle i_s^2 \rangle$ will sum to zero on average. In this case, we wish to resolve the variance (rather than the mean) of $\sum a_i b_i$ to N_b bits of resolution [40]. If a_i and b_i are independent random variables, we have:

$$\begin{aligned}&\mathrm{Var}\left(\frac{\max(|E_a||E_b|)}{n}\sum_{i=1}^{n}a_ib_i\right)\\&=\frac{\max(|E_a|^2|E_b|^2)}{n^2}\sum_{i=1}^{n}\mathrm{Var}(a_ib_i)\\&=16\frac{\bar{P}_a\bar{P}_b}{n}(\bar{a}_i^2+\sigma_a^2)\sigma_b^2\quad(\bar{b}_i=0)\end{aligned}\tag{16}$$

where σ_a^2 and σ_b^2 are the variance of a_i and b_i, respectively. If we let $\sigma_b = 0.5$ and a_i is uniformly distributed on the interval $[0,1]$, $\bar{a}_i = 0.5$ and $\sigma_a^2 = 1/12$ so equation *(16)* becomes:

$$\mathrm{Var}\left(\frac{\max(|E_a||E_b|)}{n}\sum_{i=1}^{n}a_ib_i\right)=\frac{4\bar{P}_a\bar{P}_b}{3n}\tag{17}$$

Again setting $\bar{P}_a = \bar{P}_b$ to maximize SNR, our expressions for the minimum average optical power and average optical energy per MAC operation become:

$$\begin{aligned}\bar{P}_{min}&=\frac{4h\nu}{3\eta}n\Delta f\cdot 2^{2N_b}=\frac{4h\nu}{3\eta}f_{mod}\cdot 2^{2N_b}\\\bar{E}_{MAC}&=\frac{\bar{P}_{min}}{f_{mod}}=\frac{4h\nu}{3\eta}\cdot 2^{2N_b}\quad(-1\le a_ib_i\le 1)\end{aligned}\tag{18}$$

Unlike the case for $a_ib_i \in [0,1]$, the average optical energy per MAC operation does not depend on the length of the input vectors a_i and b_i. Thus, the optical energy required to compute the dot product between two vectors within the range of $[-1,1]$ scales linearly with the input vector size, n. One approach to overcome this issue (assuming $\vec{a}_i$ is positive) is to perform two dot products instead of one such that the input vectors $\vec{a}_i, \vec{b}_j^+, \vec{b}_j^- \in [0,1]$ are all positive numbers: $\vec{a}_i \cdot \vec{b}_j = \vec{a}_i \cdot \vec{b}_j^+ - \vec{a}_i \cdot \vec{b}_j^-$. Using this strategy significantly reduces the energy consumption for large input vectors at the cost of doubling either the computation time or hardware footprint. We note that a recent work on quantifying power consumption in photonic neural network accelerators has suggested that the optical power and energy of analog photonic processors actually scales as 2^{3N_b} rather than 2^{2N_b} in the shot-noise-limited regime [44]. This would place an even stricter upper limit on the maximum computational precision that can be practically achieved in photonic neural networks, thus we have limited our analysis in **Sections IV** and **V** to $N_b = 5$-bits.

IV. Energy and Compute Density Analysis:

We now estimate the total energy consumption and compute efficiency of our photonic crossbar array. Using an externally modulated continuous-wave laser source, the minimum total optical power needed to overcome the quantum limited shot noise for positive valued inputs is:

$$\begin{aligned}P_{min}^{optical}&\ge\frac{4h\nu f_{mod}}{\eta_{mod}^2\eta_{PD}\eta_x\kappa_1^2}\left(\frac{k}{n}\right)\cdot 2^{2N_b}\\&\approx\frac{4h\nu f_{mod}}{\eta_{mod}^2\eta_{PD}}\left(\frac{k^2}{n}\right)\cdot 2^{2N_b}\end{aligned}\tag{19}$$

where η_{mod} is the transmission of the clock and input optical modulators, η_{PD} is the quantum efficiency of the photodetectors, $\eta_x\kappa_1^2$ is the fraction of power coupled into the first unit cell (defined in equations *(6)* and *(7)*), and $k \times k$ is the size of the crossbar array. The extra factor of 4× arises from the fact that while the average power is $|E_{a,b}/2|^2$, the maximum power required to cover the full range $[0,1]$ is $|E_{a,b}|^2 = 4\bar{P}_{min}$.

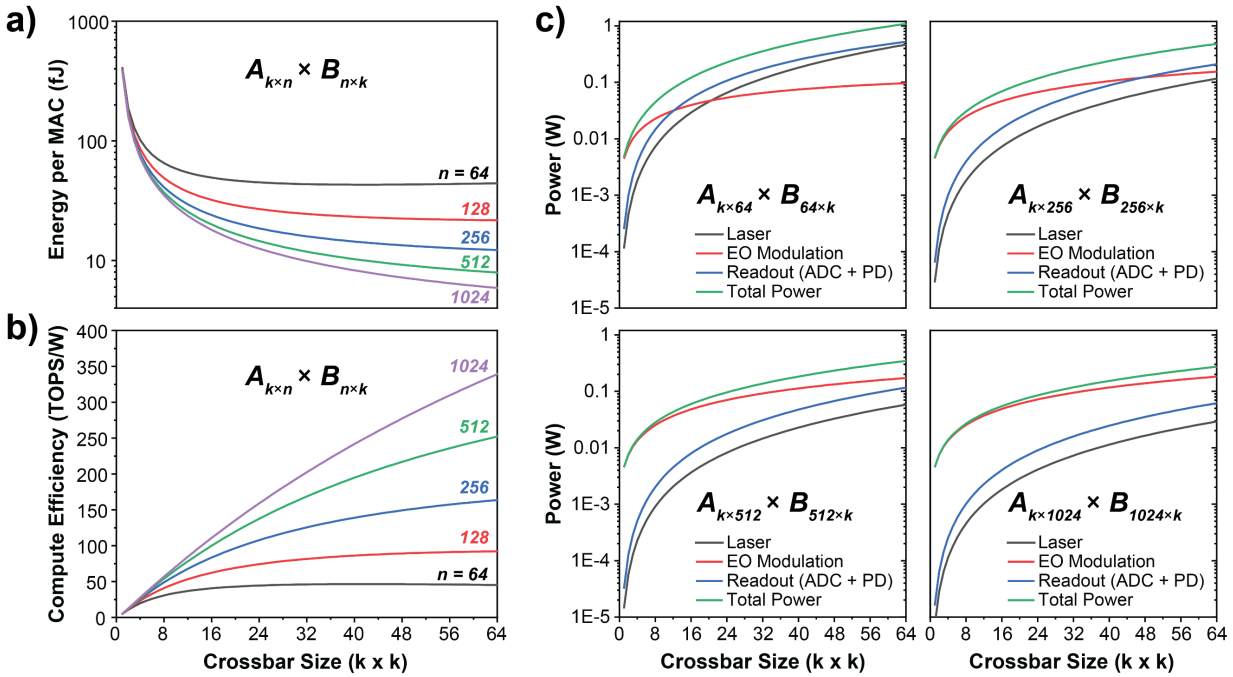


Fig. 3. Influence of matrix and crossbar dimensions on computing efficiency ($f_{mod} = 12$ GHz, $N_b = 5$-bits). **a)** Compute efficiency in energy per multiply-accumulate (MAC) operation and **b)** Tera-operations per Watt (TOPS/W) as a function of crossbar size and input matrix dimension. **c)** Breakdown of power consumption as a function of crossbar size for four different matrix dimensions. As the row/column dimension (n) increases, the power of the optical source and readout circuitry decreases, causing the E/O modulation power to dominate.

In the ideal case of lossless passive components, $\eta_x \kappa_1^2 \approx 1/k$ to account for fan-out. The total power required to operate the crossbar array will be:

$$P_{total} \approx \left(\frac{k^2}{n}\right) \cdot \frac{4h\nu f_{mod}}{\eta_{total}} \cdot 2^{2N_b} + (2k+1) \cdot P_{mod}^{E/O} + k^2 \cdot P_{read}^{O/E} \quad (20)$$

where $\eta_{total} = \eta_{mod}^2 \eta_{PD} \eta_{laser}$ includes the laser wall plug efficiency (typically assumed to be ~20%), $P_{mod}^{E/O}$ is the power consumption of each modulator, and $P_{read}^{O/E}$ is the electrical power necessary to read-out a single dot-product unit cell including analog to digital conversion. We note that while our proposed architecture requires k^2 balanced photodetector units with accompanying readout circuitry, the readout rate is f_{mod}/n and therefore the readout power scales linearly with crossbar dimension k if $k \approx n$. This is more obvious if we calculate the energy consumption per MAC operation for the entire crossbar array:

$$E_{MAC} = \frac{P_{total}}{\#\,\mathrm{MAC/s}} \approx \frac{1}{n} \cdot \frac{4h\nu}{\eta_{total}} \cdot 2^{2N_b} + \frac{(2k+1)}{k^2} \cdot \beta_{mod} N_b + \frac{1}{n} \cdot E_{read}^{O/E} \quad (21)$$

where $\beta_{mod} N_b f_{mod} = P_{mod}^{E/O}$, $E_{read}^{O/E} \cdot \frac{f_{mod}}{n} = P_{read}^{O/E}$, and β_{mod} is the modulation efficiency in J/bit. This result has a similar form to *Eq. (13)* of [43]. We can thus conclude that since E_{MAC} is inversely proportional to both k and n, larger matrix operations will result in the larger energy savings due to the advantages of fan-out and our choice of fixed-precision operations. Using the values in Table 4 of the Supplementary, Fig. 3 plots the energy consumption of our coherent matrix-multiplier as a function of photonic crossbar and input matrix dimensions k and n.

Fig. 3a-b plot the total energy per MAC and overall compute efficiency (in Tera-Ops/W or "TOPS/W") of our photonic matrix-matrix multiplier. For $n \approx k$, the compute efficiency saturates at relatively small crossbar sizes since the laser and electrical readout energies dominate equation *(21)*. However, when $n \gg k$, the $1/n$ term causes the laser and readout energies to become negligible, and the E/O modulation energy becomes dominant. This is more clearly evident in Fig. 3c where we break down the total power consumption into the power used by the optical source, E/O modulation, and O/E conversion. The $1/n$ term in equation *(21)* leads to very favorable energy scaling for large matrix operations as observed by Hamerly et al. [19] since both the minimum optical power and relative number of O/E conversions decrease significantly. Recent work by Wang et al. [45] has experimentally demonstrated that less than 1 photon per MAC is possible for optical dot products with large vector sizes (>10^3 elements). Thus, for large scale matrix operations, optical energy is unlikely to be dominant in the overall power consumption of the system as seen in Fig. 3c.

V. Comparison with other computing architectures:

We now compare our proposed photonic matrix-matrix multiplier against several integrated photonic computing architectures that have been previously demonstrated experimentally. While these demonstrations have been limited to relatively small weight matrices (a maximum weight matrix of 4×4 and 9×4 was demonstrated by Shen et al. [5] and Feldmann et al. [8], respectively) we have used idealized scaling to project the best-case performance and have limited $N_b = 5$-bits throughout. For all fixed-weight architectures, we have assumed a single photonic core that requires reprogramming if the dimensions of the input matrix $A_{m\times n}$ exceeds that of the available photonic weights ($m, n > k$). For simplicity, we have also assumed square matrices in our simulations ($m = n = p$). Note, for the broadcast-and-weight architecture using microring resonators [46], the number of wavelength channels on a single bus waveguide is limited to $k \leq 56$ based on crosstalk between nearest neighbors [43].

The fundamental difference between a fixed-weight photonic architecture and our time-multiplexed architecture is highlighted in Fig. 4a-b. In the case when $m, n > k$ (very likely for practical machine learning tasks with many millions of trained weights as illustrated in **Section VI**), the matrix $A_{m\times n}$ must be split into MN sub-matrices of dimension $k \times k$ (for example, see A_{11} in Fig. 4a). To compute the sub-matrix C_{11}, requires N matrix-matrix MAC operations with N reprogramming steps of the photonic array between [36]. Additionally, the results of each sub-matrix operation require the O/E conversion and digital storage of $(N-1)k^2$ intermediate results which can cause additional latency and energy consumption that greatly outweighs the advantages of computing in the photonic domain. By contrast, our time-multiplexed architecture allows the entire row and column of the input matrix to be processed sequentially with a single readout of the final result (C_{11}). This approach is much more efficient and does not require any additional O/E conversions or digital storage operations. Additionally, the energy savings improves with matrix dimension for positive valued inputs as highlighted in the previous section.

A. Computational efficiency and latency:

To estimate the computational efficiency of various fixed-weight photonic platforms, we use the following equations to account for the total energy consumption:

$$E_{MAC} = \frac{1}{mnp}\left(E_{laser} + E_{mod}^{E/O} + E_{weights} + E_{update} + E_{read}^{O/E} + E_{mem} + E_{digital}\right) \quad (22)$$

where we have defined the various computing energies in Supplementary Table 1.

In the case of our time-multiplexed architecture, there are no weight components or intermediate sub-matrix products to be stored/processed ($E_{weights}, E_{update}, E_{mem}, E_{digital} = 0$). This significantly reduces the overall energy per MAC by approximately four orders of magnitude compared to the most efficient fixed-weight architecture (broadcast-and-weight microring resonators [43], [46] or "MRR") as shown in Fig. 4c. We also see more than 100× greater compute efficiency than state-of-the-art commercial GPUs/TPUs [47], [48] in the limit of large n. While this is highly promising, full system modeling of data transfer between digital memory and the photonic chip

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTQE.2022.3171167, IEEE Journal of Selected Topics in Quantum Electronics

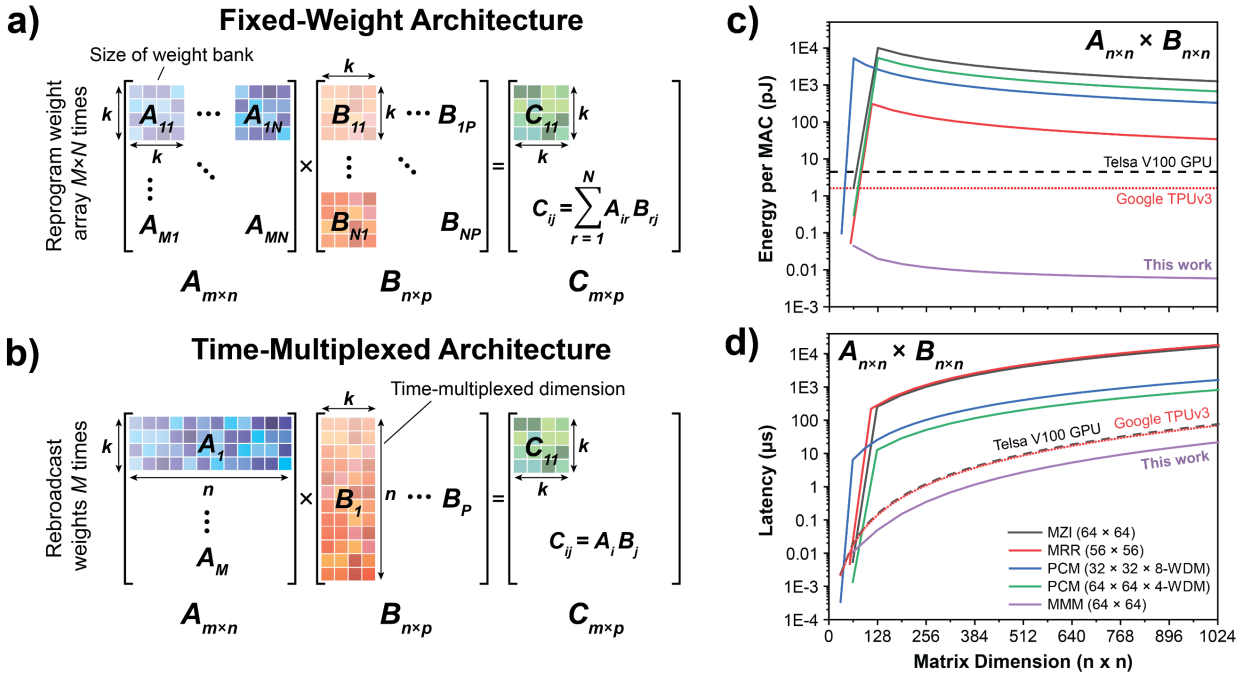


Fig. 4. Comparison of fixed-weight versus time-multiplexed architectures for matrix-matrix multiplication. **a)** For a fixed-weight architecture where $m, n > k$, the array will have to be reprogramed a minimum of MN times which leads to significant latency and energy costs. **b)** For our time-multiplexed architecture, the entire sub-matrix C_{11} is computed in n time steps without requiring any reprogramming or additional matrix-matrix operations. **c)** Energy per MAC and **d)** latency versus matrix dimension for various photonic architectures ($f_{mod} = 12$ GHz, $N_b = 5$-bits). The large discontinuity for $m, n > k$ in the MZI [5], MRR [43], [46], and PCM [8] architectures is caused by the slow and power-hungry reprogramming operations of the weight array. Energy per MAC and latency for GPU and TPU architectures estimated from reported FLOPS and wall plug power [47], [48]. Parameters used in calculations are listed in Supplementary Table 4. Note: Transfer of digital data between memory and the photonic chip has not been considered in the photonic architectures compared in **c)** and **d)**.

is needed for a more accurate comparison which is outside the scope of this paper.

The dramatic increase in energy consumption for the fixed-weight architectures is due to the need for multiple sub-matrix operations which requires reprogramming of the photonic weight array. In the case of the MZI [5] and MRR [46] architectures, we have assumed that reprogramming a column-addressed array of thermal phase shifters requires a settling time of at least ~10 µs per column, which significantly increases the overall energy consumption. While MEMS and electro-optic modulators have been proposed [44], [49] to overcome the static power consumption and slow update speed of thermal phase shifters, these approaches have their own challenges (i.e., optical insertion loss, footprint, leakage current, limited multi-bit resolution, etc.) and have yet to be experimentally demonstrated for scalable photonic computing. Electronic switching speeds as fast as ~10 to 20 ns has been demonstrated for phase-change photonic memory cells [50], [51], but the switching energy is still on the order of ~1 nJ to 10 nJ per switching event. In our architecture, the energy per weight is approximately $\sim\beta_{mod}N_b/k$ in the limit of large n and k. Since E/O modulator efficiencies can be on the order of ~fJ/bit or less [4], [52] the cost per weight is on the order of a few femto-joules or less. Any fixed-weight architecture that requires frequent weight updates during computation will likely perform much worse than our approach.

Fig. 4d compares the latency of various computing architectures as a function of matrix dimension. Similar to the case of computing efficiency, the latency of fixed-weight architectures increases dramatically once weight updates are considered. Again, we have assumed that columns are written in parallel, but rows are written sequentially for the MZI, MRR, and PCM architectures. Additionally, we have added a digital processing time to account for the N additional sub-matrix accumulate operations. The total processing time can be expressed as follows:

$$\tau_{total} = \tau_{mod} + \tau_{update} + \tau_{digital} \qquad (23)$$

where τ_{mod}, τ_{update}, and $\tau_{digital}$ are the times required for modulation, weight updates, and digital processing of sub-matrix results. Since these time delays are quite dependent on the specific architecture in question, we summarize the various sources of latency in Supplementary Table 2.

B. Fabrication variability:

Unlike the majority of other photonic computing approaches, our architecture is highly robust to fabrication variability across the crossbar array. To highlight this advantage, consider the effect of random variation in the coupling efficiency of one of the row or column directional couplers comprising a unit cell ($\widetilde{\kappa_i} = \kappa_i + \Delta\kappa_i$, $\widetilde{\kappa_j} = \kappa_j + \Delta\kappa_j$). This is likely to be the source of the greatest fabrication error in our proposed architecture. The non-ideal directional coupler will scale $\langle i_s \rangle$ by $\widetilde{\kappa_i}\widetilde{\kappa_j}$ which can be factored outside of the integral in equation *(4)* and thus simply scales the dot-product $\vec{a}_i \cdot \vec{b}_j$ by a constant. This can be compensated across the entire crossbar array by performing a

single Hadamard product between the computed output matrix and a calibrated $k \times k$ look-up table. Alternatively, one could conceivably reduce the computational burden even further by adjusting the relative gain of each unit cell's differential amplifier at the hardware level.

By a similar analysis, variations in the fan-out distribution network before the row and column modulators will also simply introduce a scaling term for each unit cell. In fact, the most significant impact of fabrication variability in the passive photonic crossbar is the requirement to increase the total input power of the optical source such that the minimum optical power derived in equation *(15)* (or equation *(18)* for negative inputs) is satisfied across all unit cells.

VI. Mixed Architectures for Efficient Data Processing

In this final section, we propose a mixed architecture approach which combines the relative strengths of fixed-weight and time-multiplexed architectures to achieve efficient

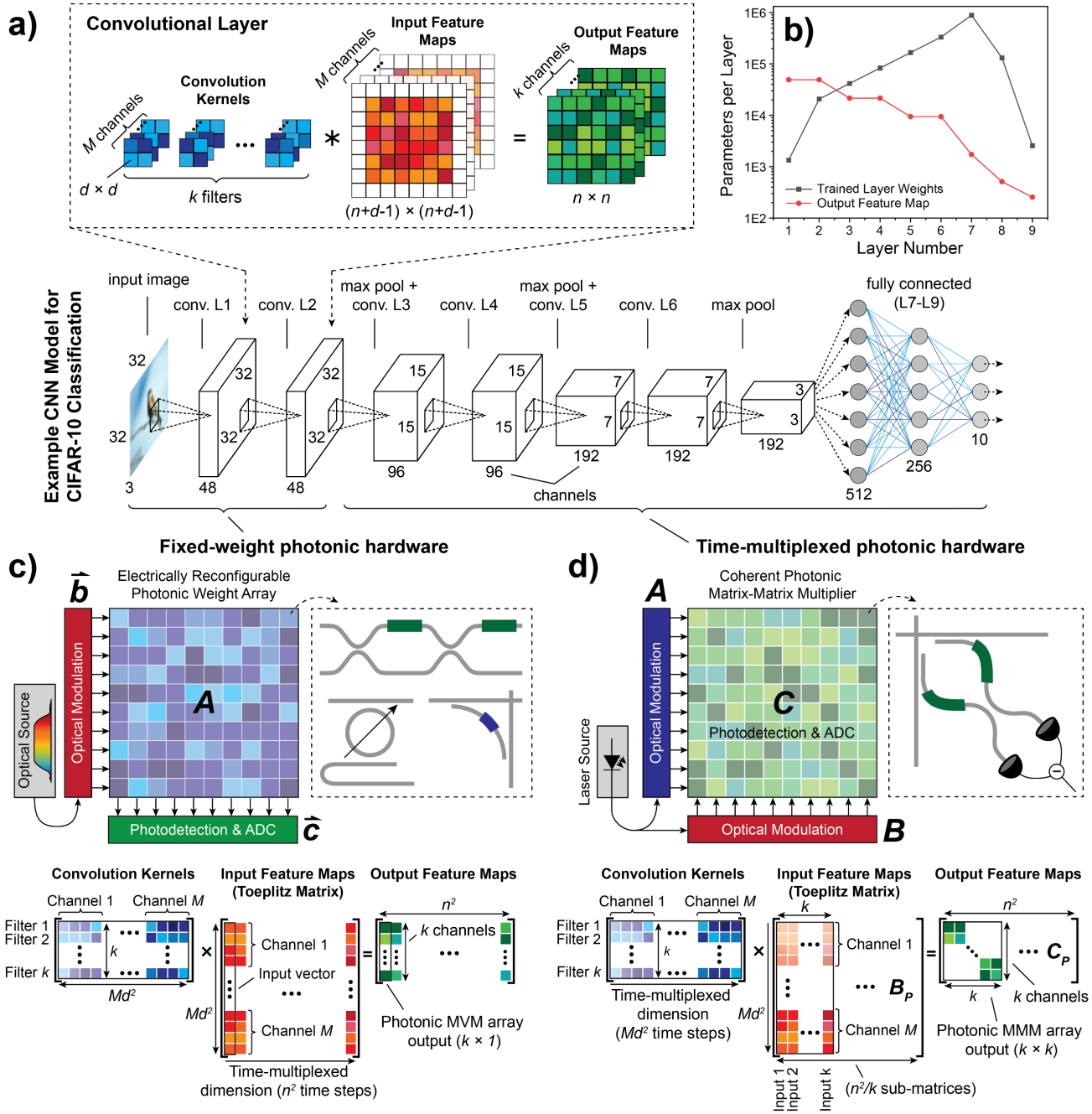


Fig. 5. Overview of mixed-architecture implementation for an example CNN. **a)** Data flow for a 9-layer CNN used to classify images from the CIFAR-10 dataset [53], [54]. Input, output, and kernel data dimensions for a convolutional layer (top inset). **b)** Total count of parameters stored and computed for a given layer in the network. **c)** Architecture overview and convolutional layer implementation for **c)** fixed-weight and **d)** time-multiplexed photonic hardware. The fixed-weight approach in **c)** has lower latency and is more efficient when the entire convolutional layer can be stored in photonic weights ($Md^2 \ll n^2$). However, as the number of parameters within a layer grows ($Md^2 \gg n^2$), a time-multiplexed approach will scale more efficiently.

photonic computing in large-scale neural networks. We illustrate this concept through a small, yet practical convolutional neural network (CNN) model used for image classification on the CIFAR-10 dataset [53], [54] shown in Fig. 5a. This CNN model has 6 convolutional layers and 3 fully connected layers for a total of ~1.7 million parameters as detailed in Table 3 of the Supplementary.

To store the entire model simultaneously in photonic hardware would require more than 400 separate photonic weight banks of size 64×64. This corresponds to a total footprint of >10 cm^2 when assuming an ambitious 25×25 μm^2 unit cell. Rather than storing the entire model in photonic memory or exclusively using a time-multiplexed approach, it could be advantageous to use a fixed-weight photonic computing architecture for the first few convolutional layers. This takes advantage of the high speed MVM operations possible with a fixed-weight approach when the output feature maps are at their largest, while the number of stored weights is smallest (i.e., $Md^2 \ll n^2$ in Fig. 5a). As data flows through the convolutional layers, the number of parameters in each layer grows, while the output feature maps are reduced in size due to repeated max pooling as plotted in Fig. 5b. This is ideal for our proposed architecture since the time-multiplexed dimension can easily accommodate the growing number of parameters. Additionally, since the time-multiplexed dimension is larger than the input feature map ($Md^2 \gg n^2$), we sum along the growing number of channels in the time domain which minimizes the number of opto-electronic conversions. Finally, since we are using a time-multiplexed approach for the layers deeper in the network, we minimize the number of costly weight updates in physical hardware during training, thus further improving the efficiency of the photonic network. While we have presented an intuitive approach for mixed-architecture design in the limit of $Md^2 \ll n^2$ or $\gg n^2$, intermediate cases will depend on the specifics of the photonic architecture and neural network being implemented. This requires full system modeling [49] which includes data movement, memory access, and other factors outside the scope of this work.

VII. Conclusions:

We have presented a photonic approach to large-scale matrix-matrix multiplication using standard components commonly available at PIC foundries. Our approach significantly reduces the ADC energy consumption and high-speed electronic design requirements of prior photonic matrix-vector multiplier strategies, while addressing the challenge of maintaining both spatial and temporal coherence between optical fields—a major difficulty in free space approaches to photonic computing. Additionally, our approach is easily scalable to large matrix-matrix operations without introducing the additional latency and energy needed to reconfigure fixed photonic weights. We have also shown that a computational efficiency of ~340 TeraOPs/W (~5.8 fJ/MAC) and peak computational speed of ~98 TeraOPs (64×64 array at 12 GHz modulation speed) are possible using experimentally demonstrated components. Finally, we have proposed a mixed architecture approach to photonic AI hardware design, providing a route toward ultrafast and efficient machine learning.

References

[1] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, "The Computational Limits of Deep Learning," *MIT Initiat. Digit. Econ. Res. Br.*, vol. 4, Jul. 2020, [Online]. Available: http://arxiv.org/abs/2007.05558.

[2] D. Amodei and D. Hernandez, "AI and Compute," 2018. https://openai.com/blog/ai-and-compute/ (accessed Jan. 11, 2019).

[3] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3645–3650, doi: 10.18653/v1/P19-1355.

[4] D. A. B. Miller, "Attojoule Optoelectronics for Low-Energy Information Processing and Communications," *J. Light. Technol.*, vol. 35, no. 3, pp. 346–396, Feb. 2017, doi: 10.1109/JLT.2017.2647779.

[5] Y. Shen *et al.*, "Deep learning with coherent nanophotonic circuits," *Nat. Photonics*, vol. 11, no. 7, pp. 441–446, Jun. 2017, doi: 10.1038/nphoton.2017.93.

[6] H. Zhang *et al.*, "An optical neural chip for implementing complex-valued neural network," *Nat. Commun.*, vol. 12, no. 1, p. 457, Dec. 2021, doi: 10.1038/s41467-020-20719-7.

[7] C. Ríos *et al.*, "In-memory computing on a photonic platform," *Sci. Adv.*, vol. 5, no. 2, p. eaau5759, Feb. 2019, doi: 10.1126/sciadv.aau5759.

[8] J. Feldmann *et al.*, "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, no. 7840, pp. 52–58, Jan. 2021, doi: 10.1038/s41586-020-03070-1.

[9] C. Wu, H. Yu, S. Lee, R. Peng, I. Takeuchi, and M. Li, "Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network," *Nat. Commun.*, vol. 12, no. 1, p. 96, Dec. 2021, doi: 10.1038/s41467-020-20365-z.

[10] X. Xu *et al.*, "11 TOPS photonic convolutional accelerator for optical neural networks," *Nature*, vol. 589, no. 7840, pp. 44–51, Jan. 2021, doi: 10.1038/s41586-020-03063-0.

[11] A. N. Tait, J. Chang, B. J. Shastri, M. A. Nahmias, and P. R. Prucnal, "Demonstration of WDM weighted addition for principal component analysis," *Opt. Express*, vol. 23, no. 10, p. 12758, May 2015, doi: 10.1364/OE.23.012758.

[12] H. T. Peng, M. A. Nahmias, T. F. De Lima, A. N. Tait, B. J. Shastri, and P. R. Prucnal, "Neuromorphic Photonic Integrated Circuits," *IEEE J. Sel. Top. Quantum Electron.*, vol. 24, no. 6, pp. 1–16, 2018, doi: 10.1109/JSTQE.2018.2840448.

[13] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," *Nature*, vol. 569, no. 7755, pp. 208–214, May 2019, doi: 10.1038/s41586-019-1157-8.

[14] S. Moazeni *et al.*, "A 40-Gb/s PAM-4 Transmitter Based on a Ring-Resonator Optical DAC in 45-nm SOI CMOS," *IEEE J. Solid-State Circuits*, vol. 52, no. 12, pp. 3503–3516, Dec. 2017, doi: 10.1109/JSSC.2017.2748620.

[15] K. Uyttenhove and M. S. J. Steyaert, "Speed-power-accuracy tradeoff in high-speed CMOS ADCs," *IEEE Trans. Circuits Syst. II Analog Digit. Signal Process.*, vol. 49, no. 4, pp. 280–287, Apr. 2002, doi: 10.1109/TCSII.2002.801191.

[16] B. Murmann, "ADC Performance Survey 1997-2021," 2021. http://web.stanford.edu/~murmann/adcsurvey.html.

[17] W. Liu, W. Liu, Y. Ye, Q. Lou, Y. Xie, and L. Jiang, "HolyLight: A Nanophotonic Accelerator for Deep Learning in Data Centers," in *Design, Automation Test in Europe Conference Exhibition*, 2019.

[18] S. Ambrogio *et al.*, "Equivalent-accuracy accelerated neural-network training using analogue memory," *Nature*, vol. 558, no. 7708, pp. 60–67, Jun. 2018, doi: 10.1038/s41586-018-0180-5.

[19] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, "Large-Scale Optical Neural Networks Based on Photoelectric Multiplication," *Phys. Rev. X*, vol. 9, no. 2, p. 021032, May 2019, doi: 10.1103/PhysRevX.9.021032.

[20] J. Sun, E. Timurdogan, A. Yaacobi, E. S. Hosseini, and M. R. Watts, "Large-scale nanophotonic phased array," *Nature*, vol. 493, no. 7431, pp. 195–199, Jan. 2013, doi: 10.1038/nature11727.

[21] C. Rogers *et al.*, "A universal 3D imaging sensor on a silicon photonics platform," *Nature*, vol. 590, no. 7845, pp. 256–261, Feb.

2021, doi: 10.1038/s41586-021-03259-y.

[22] X. Zhang, K. Kwon, J. Henriksson, J. Luo, and M. C. Wu, "A large-scale microelectromechanical-systems-based silicon photonics LiDAR," *Nature*, vol. 603, no. 7900, pp. 253–258, Mar. 2022, doi: 10.1038/s41586-022-04415-8.

[23] A. Annoni *et al.*, "Unscrambling light - Automatically undoing strong mixing between modes," *Light Sci. Appl.*, vol. 6, no. 12, pp. e17110–e17110, Dec. 2017, doi: 10.1038/lsa.2017.110.

[24] A. N. Tait *et al.*, "Feedback control for microring weight banks," *Opt. Express*, vol. 26, no. 20, p. 26422, Oct. 2018, doi: 10.1364/OE.26.026422.

[25] B. Chen *et al.*, "Real-time monitoring and gradient feedback enable accurate trimming of ion-implanted silicon photonic devices," *Opt. Express*, vol. 26, no. 19, p. 24953, Sep. 2018, doi: 10.1364/OE.26.024953.

[26] C. Ríos *et al.*, "Electrically-switchable foundry-processed phase change photonic devices," in *Active Photonic Platforms XIII*, Aug. 2021, p. 66, doi: 10.1117/12.2592021.

[27] M. Delaney *et al.*, "Nonvolatile programmable silicon photonics using an ultralow-loss Sb2Se3 phase change material," *Sci. Adv.*, vol. 7, no. 25, p. eabg3500, Jun. 2021, doi: 10.1126/sciadv.abg3500.

[28] Y. Zhang *et al.*, "Broadband transparent optical phase change materials for high-performance nonvolatile photonics," *Nat. Commun.*, vol. 10, no. 1, p. 4279, Dec. 2019, doi: 10.1038/s41467-019-12196-4.

[29] Y. Ehrlichman, O. Amrani, and S. Ruschin, "Generating arbitrary optical signal constellations using microring resonators," *Opt. Express*, vol. 21, no. 3, p. 3793, Feb. 2013, doi: 10.1364/OE.21.003793.

[30] X. Wu *et al.*, "A 20Gb/s NRZ/PAM-4 1V transmitter in 40nm CMOS driving a Si-photonic modulator in 0.13μm CMOS," in *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, Feb. 2013, pp. 128–129, doi: 10.1109/ISSCC.2013.6487667.

[31] Chi Xiong, D. Gill, J. Proesel, J. Orcutt, W. Haensch, and W. M. J. Green, "A monolithic 56 Gb/s CMOS integrated nanophotonic PAM-4 transmitter," in *2015 IEEE Optical Interconnects Conference (OI)*, Apr. 2015, pp. 16–17, doi: 10.1109/OIC.2015.7115665.

[32] Y. Ma *et al.*, "Ultralow loss single layer submicron silicon waveguide crossing for SOI optical interconnect," *Opt. Express*, vol. 21, no. 24, p. 29374, 2013, doi: 10.1364/oe.21.029374.

[33] D. Coppersmith and S. Winograd, "Matrix multiplication via arithmetic progressions," in *Proceedings of the nineteenth annual ACM conference on Theory of computing - STOC '87*, 1987, pp. 1–6, doi: 10.1145/28395.28396.

[34] V. V. Williams, "Multiplying matrices faster than coppersmith-winograd," in *Proceedings of the 44th symposium on Theory of Computing - STOC '12*, 2012, p. 887, doi: 10.1145/2213977.2214056.

[35] F. Le Gall, "Powers of Tensors and Fast Matrix Multiplication," Jan. 2014, [Online]. Available: http://arxiv.org/abs/1401.7714.

[36] M. Miscuglio and V. J. Sorger, "Photonic tensor cores for machine learning," *Appl. Phys. Rev.*, vol. 7, no. 3, p. 031404, Sep. 2020, doi: 10.1063/5.0001942.

[37] C. Wang *et al.*, "Scalable massively parallel computing using continuous-time data representation in nanoscale crossbar array," *Nat. Nanotechnol.*, Jul. 2021, doi: 10.1038/s41565-021-00943-y.

[38] N. Semenova, X. Porte, L. Andreoli, M. Jacquot, L. Larger, and D. Brunner, "Fundamental aspects of noise in analog-hardware neural networks," *Chaos An Interdiscip. J. Nonlinear Sci.*, vol. 29, no. 10, p. 103128, Oct. 2019, doi: 10.1063/1.5120824.

[39] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep Learning with Limited Numerical Precision," *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 3, pp. 1737–1746, Feb. 2015, Accessed: Aug. 24, 2021. [Online]. Available: https://proceedings.mlr.press/v37/gupta15.html.

[40] S. Agarwal *et al.*, "Energy Scaling Advantages of Resistive Memory Crossbar Based Computation and Its Application to Sparse Coding," *Front. Neurosci.*, vol. 9, Jan. 2016, doi: 10.3389/fnins.2015.00484.

[41] J. Feldmann, N. Youngblood, X. Li, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, "Integrated 256 Cell Photonic Phase-Change Memory With 512-Bit Capacity," *IEEE J. Sel. Top. Quantum Electron.*, vol. 26, no. 2, pp. 1–7, Mar. 2020, doi: 10.1109/JSTQE.2019.2956871.

[42] B. M. Oliver, "Signal-to-noise ratios in photoelectric mixing," *Proc. IRE*, vol. 49, pp. 1960–1961, 1961.

[43] M. A. Nahmias, T. F. De Lima, A. N. Tait, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, "Photonic Multiply-Accumulate Operations for Neural Networks," *IEEE J. Sel. Top. Quantum Electron.*, pp. 1–1, 2019, doi: 10.1109/JSTQE.2019.2941485.

[44] A. N. Tait, "Quantifying power use in silicon photonic neural networks," Aug. 2021, [Online]. Available: http://arxiv.org/abs/2108.04819.

[45] T. Wang, S.-Y. Ma, L. G. Wright, T. Onodera, B. C. Richard, and P. L. McMahon, "An optical neural network using less than 1 photon per multiplication," *Nat. Commun.*, vol. 13, no. 1, p. 123, Dec. 2022, doi: 10.1038/s41467-021-27774-8.

[46] 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. Light. Technol.*, vol. 32, no. 21, pp. 4029–4041, Nov. 2014, doi: 10.1109/JLT.2014.2345652.

[47] N. P. Jouppi *et al.*, "A domain-specific supercomputer for training deep neural networks," *Commun. ACM*, vol. 63, no. 7, pp. 67–78, Jun. 2020, doi: 10.1145/3360307.

[48] Nvidia, "Nvidia V100 Tensor Core GPU," 2020. https://images.nvidia.com/content/technologies/volta/pdf/volta-v100-datasheet-update-us-1165301-r5.pdf.

[49] C. Demirkiran *et al.*, "An Electro-Photonic System for Accelerating Deep Neural Networks," Sep. 2021, [Online]. Available: http://arxiv.org/abs/2109.01126.

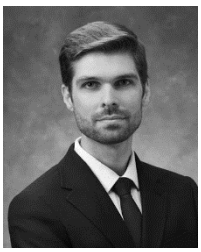
[50] J. Zheng *et al.*, "Nonvolatile Electrically Reconfigurable Integrated Photonic Switch Enabled by a Silicon PIN Diode Heater," *Adv. Mater.*, vol. 32, no. 31, p. 2001218, Aug. 2020, doi: 10.1002/adma.202001218.

[51] H. Zhang *et al.*, "Miniature Multilevel Optical Memristive Switch Using Phase Change Material," *ACS Photonics*, vol. 6, no. 9, pp. 2205–2212, Sep. 2019, doi: 10.1021/acsphotonics.9b00819.

[52] W. Heni *et al.*, "Plasmonic IQ modulators with attojoule per bit electrical energy consumption," *Nat. Commun.*, vol. 10, no. 1, p. 1694, Dec. 2019, doi: 10.1038/s41467-019-09724-7.

[53] S. R. Nandakumar *et al.*, "Mixed-Precision Deep Learning Based on Computational Memory," *Front. Neurosci.*, vol. 14, May 2020, doi: 10.3389/fnins.2020.00406.

[54] P. Kaur, "Convolutional Neural Networks (CNN) for CIFAR-10 Dataset." http://parneetk.github.io/blog/cnn-cifar10/.



Nathan Youngblood (Member, IEEE) received the B.S. degree in physics from Bethel University, St. Paul, MN, USA in 2011 and the Ph.D. degree in electrical engineering from the University of Minnesota, Minneapolis, MN, USA in 2016 where he was involved in the integration of 2-D materials with silicon photonics for optoelectronic applications. After postdoctoral training at the University of Oxford, Oxford, UK, he joined the Department of Electrical and Computer Engineering at the University of Pittsburgh, Pittsburgh, PA, USA in 2019. His research interests include integrated photonics, high-speed optoelectronics, artificial intelligence, and novel computing methods with light.