

Optical and Electrical Memories for Optical Computing

Sadra Rahimi Kari, *Student Member, IEEE*, Carlos Ríos, *Member, IEEE*, Lei Jiang, *Member, IEEE*, Jiawei Meng, *Student Member, IEEE*, Nicola Peserico, *Member, IEEE*, Volker J. Sorger, *Senior Member, IEEE*, Juejun Hu, *Member, IEEE*, Nathan Youngblood, *Member, IEEE*

Abstract—Key to recent successes in the field of artificial intelligence (AI) has been the ability to train a growing number of parameters which form fixed connectivity matrices between layers of nonlinear nodes. This “deep learning” approach to AI has historically required an exponential growth in processing power which far exceeds the growth in computational throughput of digital hardware as well as trends in processing efficiency. New computing paradigms are therefore required to enable efficient processing of information while drastically improving computational throughput. Emerging strategies for analog computing in the photonic domain have the potential to drastically reduce latency but require the ability to modify optical processing elements according to the learned parameters of the neural network. In this point-of-view article, here we provide a forward-looking perspectives on both optical and electrical memories coupled to integrated photonic hardware in the context of AI. We show that for programmed memories the READ energy-latency-product of photonic random access memory (PRAM) technology can be orders of magnitude lower as compared to electronic SRAMs. However, current PRAM-based devices are bulk compared to electronics and we comment on the need for further material and device-design optimizations all together leading to a PRAM technology roadmap. It is our intent to share a path that PRAMs become an integral part of future foundry processes give these promising initial device performance and relevance for emerging AI hardware and machine learning accelerators, but also for future network edge modules for the looming industry-4.0 era.

Index Terms—Artificial intelligence, neural network hardware, analog computers, optical computing, analog processing circuits

I. INTRODUCTION

RECENT progress in the field of AI has been fueled by two major research thrusts: 1) finding ways to train increasingly large deep neural networks (DNNs) and 2) applying new insights from neuroscience to computing algorithms and hardware, commonly known as “neuromorphic computing.” These approaches to AI make the shift from specialized “expert models” which rely on a human

understanding of the data to generalized “neural networks” which typically use a very large number of free parameters to statistically fit the data [1]. In fact, the performance of a DNN has been shown to improve when the number of free parameters exceeds that of the available training data [2]. The vast and tunable 3D connectivity of billions of neurons in the brain is similarly considered a key contributor to intelligence in humans and other animals. Thus, the immense number of trainable parameters in biological and deep neural networks leads to both its generality as well as computational complexity [3].

In both deep learning and neuromorphic computing, the compute operations needed varies drastically from the precise, sequential arithmetic operations that have driven digital hardware design for the past half century. Instead, computation is limited by memory access bottlenecks rather than processor speed, leading to memory-centric design approaches (e.g., weight stationary systolic arrays [4], in-memory computation [5], etc.). These approaches typically minimize the movement of fixed parameters to improve latency and energy efficiency. However, since all electrical processors are fundamentally limited by an energy-bandwidth tradeoff stemming from the capacitance of their interconnects [6], this ultimately limits the maximum compute efficiency achievable (typically measured in operations per watt, “OPS/W”).

Analog computation in the optical domain is an exciting alternative to electrical processors which side-steps this energy-bandwidth tradeoff [7]–[9]. The bandwidth of an optical channel (waveguide, fiber, or even free space) is independent of modulation frequency and therefore extremely high data throughput can be achieved in the optical domain. Additionally, the wave nature of optical signals allows passive elements to achieve unitary linear transformations with no power penalty in lossless materials [10]. These properties make optical analog computing highly attractive for ultrafast, low-power linear operations—the major computational bottleneck in today’s neural networks. However, for these optical computations to be

This work was supported in part by the U.S. National Science Foundation under Grants ECCS-2028624, DMR-2003325, and CISE-2105972. V.J.S. is supported by the Air Force Office for Scientific Research (AFOSR) PECASE, FA9550-1-20-0193. **Others?** (*Corresponding author: N. Youngblood*)

S. Rahimi Kari and N. Youngblood are with the Department of Electrical and Computer Engineering, Swanson School of Engineering, University of Pittsburgh, Pittsburgh, PA 15261 USA.

C. Ríos is with the Department of Materials Science and the Institute for Research in Electronics and Applied Physics, University of Maryland, College Park, MD 20742, USA.

J. Meng N and Peserico are with the Department of Electrical and Computer Engineering, School of Engineering & Applied Science, George Washington University, Washington, DC 20052, USA.

V. J. Sorger is with the Department of Electrical and Computer Engineering, School of Engineering & Applied Science, George Washington University, Washington, DC 20052, USA, and with Optelligence LLC, Upper Marlboro, MD 20772, USA.

J. Hu is with the Massachusetts Institute of Technology, Department of Materials Science and Engineering, Cambridge, MA 02139-4307, USA.

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

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

useful, they must be coupled to the trained parameters of the neural network through analog optical memories.

In this point of view article, we first identify important features needed for analog optical memory memories and their respective challenges. We then discuss different approaches used by the community to implement optical memory for processing information in the optical domain. Next, we perform an energy-latency analysis to identify the applications where these various approaches have a distinct advantage. Finally, we end our discussion with an outlook of the current state of optical memory technologies and present a roadmap identifying the key technological challenges where continued innovation is most needed.

II. KEY REQUIREMENTS OF PHOTONIC MEMORY TECHNOLOGY

At the highest level, photonic computing strategies can be most generally divided into two main categories—coherent or incoherent. These distinct strategies place important physical constraints on the optical memory cells used since in the case of coherent photonic architectures, both the amplitude and phase of the optical signals are used to perform computation [11]–[13]. Incoherent architectures instead use only the amplitude of the optical signal to perform computation, but require sources with many different optical frequencies to prevent unwanted interference effects [14]–[16]. Therefore, for coherent architectures, the insertion loss (IL), amplitude-independent phase control, and fabrication variability of the memory cell directly impact the compute accuracy [17]. These strict requirements are largely reduced for incoherent architectures, but extinction ratio (ER), crosstalk, and precision of the memory cell still limit the ultimate accuracy that can be achieved [18]. Despite these architecture-specific requirements, several key metrics of the memory cell have similar impact on the performance of the photonic processor regardless of the computing strategy. Here, we summarize these metrics and their importance for photonic computing.

Insertion loss (IL). IL of the memory cell impacts the maximum optical power that can be transmitted and read out by detection circuitry when the memory is in the fully “on” state. Since computation occurs in the analog domain, the precision of the optical readout is fundamentally limited by photon shot noise. Improving the IL, therefore reduces the optical power required to perform computation. For coherent architectures, if the IL differs between two interfering optical paths, the interference contrast will be reduced and limit compute accuracy (sometimes also referred to as “fidelity” [11]).

Precision. Optical memory cells are typically tuned with a continuous parameter since they are analog in nature. Therefore, the maximum achievable precision is typically limited by either the stability of the memory cell itself, the noise of the control circuitry, or the optoelectronic noise at detection. Fortunately, many studies have shown that neural networks require relatively low precision memory (even as low as 1 or 2 bits [19]–[21]) and that uncorrelated noise can serve as a method for regularization and improved resilience [22]–[24].

Extinction ratio (ER). The ER of the memory cell is linked to the precision and determines the maximum optical contrast between the “on” and “off” states. Improving the ER will help

to distinguish between neighboring analog levels of transmission or phase, increasing the maximum compute precision (and typically accuracy) achievable. Detecting the difference in intensity between the add and drop ports of a microring resonator (MRR) or in the relative transmission of two memory cells are methods for improving ER while also achieving both positive and negative values for weights [16].

Programming latency. While access and read latency can be a bottleneck for electronic memory cells, the write speed of the memory cell is usually the limiting factor for photonics. Reading the state of memory in the optical domain is fundamentally limited by the speed of light traveling through the bus waveguides, but in practice readout is limited by the speed of the detection circuitry at the output. Therefore, in the case of frequent weight updates, the programming latency could dominate (especially in the limit of large matrix operations which exceed the available on-chip photonic memory [25]). Therefore, minimizing the latency for frequent weight updates is crucial for maximizing throughput when faced with realistic constraints on physical optical hardware.

Programming energy and static power. Similar to the case of latency, if the computing application requires frequent updating of the optical weights (e.g., in a photonic tensor core [26]), the optical memory cell programming energy could potentially dominate the power consumption of the chip. Additionally, when using volatile optical responses to store data—such as thermo-optic, electro-absorptive, or plasma-dispersion effects—the static power consumption needed to hold a fixed weight can contribute a significant amount to the overall power budget of the computing system [18].

Cycling endurance. The minimum number of cycles required for an optical memory cell will vary greatly depending on the use case. For example, a fixed-weight architecture that does not require frequent weight updates (e.g., a small convolutional layer implemented optically [14], [15]) will have a much lower cycling requirement compared to a neuromorphic architecture where accumulation of optical pulses occurs in the memory cells themselves [27], [28]. As a point of reference, NAND flash memory used in consumer-grade USB flash drives typically have endurance ranging from 10^4 to 10^6 cycles [29], but these devices are used for storage rather than computation.

Footprint. The footprint of the optical memory cell limits the integration density on chip and can be the limiting factor for scalability. This has important implications on the efficiency and latency of the photonic processor since smaller memory arrays will require more frequent weight updates than large-scale memory arrays for the same matrix operation [25]. While the footprint of photonic memory cells is much larger than that of electronic memory, with the waveguide dimensions and evanescent coupling as the main limiting factors, the compute density can be much greater for optical memory due to high-speed analog operations [18].

III. CURRENT IMPLEMENTATIONS OF PHOTONIC MEMORY

A. Electronic memories coupled to optical components

One common method for implementing optical memory is to use an optical modulator coupled to electrical memory. This first involves digital-to-analog conversion (DAC) of the digital

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

weight, followed by electrical-to-optical conversion (E/O) of the analog electrical signal. E/O conversion is most commonly achieved by modulating the real or imaginary refractive index of a material through different physical effects, such as thermo-optic, electro-absorption, or plasma-dispersion [30]–[35]. This approach to optical memory has the notable benefit of foundry compatibility which has enabled several key proof-of-concept demonstrations of photonic processors [8]. Additionally, by decoupling the device used for optical modulation from that of data storage, both devices can independently optimize important metrics that could be high challenging to optimize in a single material platform (e.g., programming speed and cycling endurance). However, most physical effects used for optical modulation are both volatile and weak (e.g., $\Delta n \sim 10^{-3}$ to 10^{-4} per volt, $^{\circ}\text{C}$, etc.). This translates to constant external biasing (e.g., P-N junction) or power dissipation (e.g., resistive microheater) to maintain the state of an optical weight, as well as large device footprints for non-resonant devices such as MZIs and electro-absorptive modulators. Below, we briefly describe the most common devices used to implement optical memory and their operation.

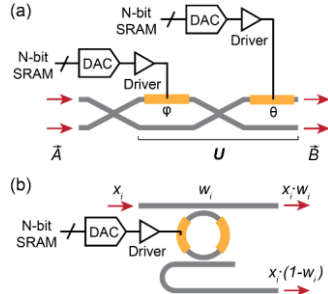


Fig. 1. Electronic memories coupled to optical modulators. (a) Schematic of a reconfigurable MZI implementing the 2×2 unitary matrix U . (b) Schematic of a programmable add-drop MRR using differential weighting to implement positive and negative weights.

A *Mach-Zehnder Interferometer* (MZI) is a reconfigurable 2×2 photonic coupler that uses two pairs of phase shifters and bidirectional couplers to implement a 2×2 unitary weight matrix U . Normalized incident field amplitudes are used to represent the elements of an input vector $\vec{A}$. The optical output vector from the MZI is then equal to $\vec{B} = U\vec{A}$. To reconfigure the weight matrix U , a pair of phase shifters are arranged on any two arms of the MZI to control both the interference and relative phase of the two outputs. Assuming coherent inputs, 50:50 couplers, and two phase shifters ϕ and θ , the output amplitudes can be described as:

$$\vec{B} = \begin{bmatrix} e^{j\phi} \sin(\theta) & \cos(\theta) \\ e^{j\phi} \cos(\theta) & -\sin(\theta) \end{bmatrix} \vec{A}. \quad (1)$$

MZIs can be organized into a mesh to serve as an optical linear unit that performs matrix multiplications [36]. An $N \times N$ arbitrary unitary matrix can be deployed on MZIs connected in

various mesh topologies, e.g., triangular [37], rectangular [38], and binary tree [39]. While mathematically elegant, one drawback of this approach is the requirement of $\sim N^2$ MZIs to implement arbitrary $N \times N$ matrices through the singular value decomposition approach [36] which can lead to large footprints and low compute density [18].

A *Micro-Ring Resonator* (MRR) is a reconfigurable optical device that can be used to tune the relative transmission of its through and drop ports at specific optical frequencies which depend on the radius of the ring [40]. To implement matrix multiplication, an $N \times N$ array of MRRs can be used in a wavelength-division multiplexing (WDM) scheme to form a “broadcast and weight” architecture [16]. Input vectors are encoded as the modulated light intensities of multiple wavelengths, while each MRR acts as a filter to selectively apply attenuation to a specific input wavelength according to a corresponding matrix element [41]. Crosstalk between MRRs of similar optical resonance and free spectral range limit the ultimate size of the $N \times N$ matrix which can be implemented. Moreover, MRRs also suffer from high sensitivity to temperature and fabrication variations.

Resistive heaters and P-N junctions are most commonly used as phase shifters in MZIs and MRRs [31]–[33], [35]. These two modulation approaches have certain advantages and disadvantages for optical memory. For instance, despite having very low insertion losses, resistive heaters suffer from slow switching speeds (hundreds of kHz) and high static power consumption (several mW). On the other hand, P-N junctions offer high switching speeds and typically dissipate very little static power. However, their insertion loss is high due to free-carrier absorption and also dependent on the applied bias, making them unsuitable for photonic processors using the coherent schemes mentioned above.

When using these volatile optical modulators as memory units, each modulator requires designated control circuitry to read digital data from memory and then hold the transmission or phase of the modulator constant. This not only introduces complexity to the integrated system, but it also increases static power dissipation from the DAC and driver blocks needed to hold the state of each modulator. When combined with the energy and latency of high-speed DACs, this can increase the overall power consumption and latency of the photonic processor and is analyzed in more detail in Section IV.

In recent years several methods have been used to eliminate the need for DACs and directly use binary data with E/O modulators. Examples include directly modulating light with binary inputs using segmented MZIs [34] and MRRs [31] with up to 4 bits of resolution. This is a promising approach for optical memory as such schemes can even improve the DAC linearity [31]. We compare the various modulation schemes described above in **Table 1**.

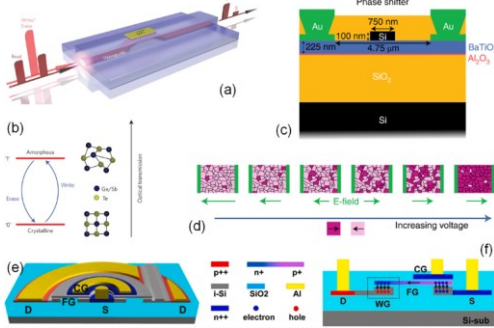
Technology	Speed	Energy/Power	IL (ER)	DAC?
Segmented P-N MRR [31]	20 Gb/s (NRZ)	155 fJ/bit	5.5 dB (3 dB)	No
	40 Gb/s (PAM-4)	42 fJ/bit		

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

Segmented SIS-CAP MZI [34]	20 Gb/s (NRZ)	4.5 pJ/bit	NA	No
	40 Gb/s (PAM-16)	250 fJ/bit	NA	
Single P-N MRR [33]	44 Gb/s (NRZ)	17.4 fJ/bit	0.9 dB (8 dB)	Yes
Thermal MZI [42]	2.4 μ s	12.7 mW (P_{π})	0.5 dB (20 dB)	Yes
Thermal MRR [43]	1.3 μ s	1.47 nm/mW	NA (15 dB)	Yes

Table 1: Comparison of metrics for various optical modulators.**B. On-chip memories based on nonvolatile photonics**

A second approach for implementing on-chip photonic memories involves nonvolatile optical materials or phenomena, where the stored weights are recorded in the form of erasable refractive index and/or optical absorption changes. The examples include: 1) phase change materials (PCMs), which exhibit giant optical property change upon undergoing a nonvolatile amorphous-crystalline structural transition [44]; 2) ferroelectric (FE) crystals exemplified by BaTiO₃ (BTO) whose electric polarization can be switched by an external electrical field in a nonvolatile manner [45]; and 3) charge accumulation in a floating gate or charge trapping in a dielectric layer, the mechanism responsible for data storage in electronic flash memories, which modifies the optical attributes in a Si waveguide via free carrier plasma dispersion [46] (**Fig. 2**). All the schemes are amenable to electrical writing and optical reading [47]–[50]. Another key feature of these memories is multi-level operation capacity, where the presence of intermediate states (corresponding to e.g. mixtures of amorphous/crystalline phases in PCMs [51] or partial FE domain switching in FE crystals [52]) can be used to encode multi-bit information in one single memory cell [53]–[55]. In-memory computing based on nonvolatile photonic memories have been demonstrated in single memory cells [56] as well as in large crossbar arrays [57].

**Fig. 2. Nonvolatile optical memory technologies.** (a) Schematic illustration of a PCM-integrated photonic memory; (b) operating mechanism of the PCM-integrated memory: less optical power is transmitted through the waveguide if the PCM is in the crystalline state than when it is in the amorphous state [44]; (c) cross-section structure of a nonvolatile waveguide phase shifter integrated with FE BTO

crystal, which can serve as a basic building block for photonic memory; (d) schematics depicting progressive FE domain switching with increasing the voltage applied between the electrodes [45]; (e) tilted and (f) cross-sectional schematics of a photonic memory device based on charge accumulation in a floating gate. The black arrows indicate the charge carrier flow directions during write and erase operations [58].

Compared to electronic memory driven approaches discussed in the previous section, nonvolatile photonic memories allow fixed weight storage with zero static power dissipation while affording improved long-term data retention. These nonvolatile photonic memory technologies also each boasts unique advantages with respective technical limitations. In addition to using variable attenuation to represent weights as is illustrated in **Fig. 2b**, low-loss PCMs [59] can execute phase-only encoding functions in a coherent network [60]. PCM photonic memory cells are also ultra-compact, only a few microns in length. However, they require relatively large switching power (sub-nJ for all-optical switching [44] and a few nJ's for electrothermal switching [61]). Moreover, their cycling endurance must be further improved [62]. In comparison, FE devices claim considerably reduced switching power consumption down to tens of pJ's [45] as well as enhanced endurance [63], although they require much larger footprint and a constant DC bias to maintain electro-optic index change during readout. Both PCM and FE devices also involve new materials and special processes (backend deposition for PCMs and wafer bonding for FE crystals) for integration with standard Si photonic foundry process. The charge accumulation or trapping devices hold the advantage of full CMOS compatibility, although they suffer from similar limitations as their electronic flash memory counterpart in low write/erase speed and endurance.

C. Passive optical memories

Controlling signal propagation through delay lines is another promising approach to implement optical memory. This approach has been used as volatile optical memory for computing in both recurrent and convolutional photonic neural networks [14], [64]–[66]. When combined with time-multiplexing and wavelength dispersion, optical delay lines have been used to achieve extremely high computational throughput with ultra-low latencies [14]. The fact that they are fully passive and have minimal latency (i.e., time of flight of the optical signal) are two major advantages of using optical delay lines for temporary data storage. However, optical delay lines require significant area on-chip—limited by the bending radius and spacing between neighboring waveguides—which increases with the required delay. Additionally, it is challenging to efficiently tune these delays after fabrication. Heterogeneous approaches which integrate multiple optical degrees of freedom using WDM, optical memories, and delay lines is a promising direction for photonic computing [66].

IV. ENERGY-LATENCY ANALYSIS

In order to establish a comparison between emerging memory technologies in the optical domain (O) with their electronic (E)

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

counterparts, we can utilize the figure of merit defined as the READ-WRITE operations ratio, as well as the overall energy and latency cost when considering E/O and O/E conversions.

A. READ operation

For an ideal photonic memory based on PCMs or other nonvolatile material platform, the READ operation requires the energies for the creation and detection of a single photon to access the stored data [67]. Considering a laser source, a memory insertion loss (0.005 dB/bit), and photodetector readout, the READ (access) energy of a photonic random-access memory (P-RAM) takes <1 fJ/bit for an on-off-keyed signal at 30 GHz data rates, or, about 10 fJ/bit access for a higher bit resolution (e.g. PAM-16 for a 4-bit one) [68]–[71]. State-of-the-art SRAM memory using flip-flops, which can be in one of two bistable states, has an access latency of 0.21 ns and costs about 5 pJ/bit access [72], [73]. Energy and latency penalties increase when accessing data stored in SRAM cache memories, costing around 180 pJ and 1.66 ns per access for FinFET-based technologies [74]. Thus, a generic photonic link offers MAC operations and memory access of $10\text{--}100\times$ higher MAC/s/J/access than SRAM, highlighting how a P-RAM can improve the performance of a computational processor, as compared in the Table 2.

	Area/bit (μm^2)	Read energy (fJ/bit)	Read Latency (ps)
SRAM cache [74] (64-byte block size)	0.055	350	1,660
SRAM cell [73], [75] (7nm Fin-FET, 6T)	~ 0.01	5,000	210
P-RAM [76]	15	10	< 50

Table 2: Performance table of a photonic random-access memory (P-RAM) as compared to established SRAM shows an several order-of-magnitude higher READ performance. This is particularly relevant for network edge AI with seldomly updated weights (i.e., rare WRITE operations), but frequent READS. Note, this does not include ADC energy or latency for P-RAM READ operations since computation can occur optically across multiple P-RAM memory cells before ADC. Adding the area to the read energy and latency shows an about $5\times$ higher figure-of-merit based on a $(\text{area} \times \text{read energy} \times \text{read latency})^{-1}$.

B. WRITE operation

When writing data to a P-RAM cell, triggering the phase transition of the chalcogenide material, switching ferroelectric domains, etc. is required. This leads to a strong modulation of optical properties (phase for materials such as Sb_2Se_3 and BTO, or amplitude for materials such as GST, GSST, and GSSe). In the case of PCMs, local annealing is used to switch the material—typically either using all-optical heating or an on-chip electro-thermal microheater (e.g., ITO, doped silicon, or metal heaters [68], [70], [77]). This multilevel, ultra-compact approach using PCMs with low IL (such as GSST and GSSe [67], [76]) enables highly efficient fixed weight banks with low power consumption. Compared with writing to SRAM cells, the writing of P-RAM based on (Joule) heating is limited by the behavior of heat propagation and thus requires higher writing energies (few pJ to sub-nJ for all-optical approaches [77] and few nJ for integrated microheaters [78]), as well as higher

latency (sub- μs). In comparison, the SRAM address line, that is operated for opening and closing the switch and to control the certain transistors that permits reading, can experience a writing speed of ~ 1 to 2 ns per access with an associated energy down to <10 pJ/bit. However, unlike the volatile SRAM which needs constant external voltage applied once the information is written to preserve from the current leakage (~ 2 nW/bit [74], [75]), PCM based non-volatile P-RAM does not require continuous external energy after the information is written. Thus, one state of PCM can be maintained passively long term. From an energy perspective, PCM based P-RAM is more suitable for applications which do not require frequent updates and instead require low-cost, long-term data storage which can be rapidly accessed once the information is written. In fact, there is a point beyond which P-RAM becomes more energy efficient compared to the SRAM energy requirements for storing information (Fig. 3). For novel PCM materials, researchers might look for any compounds with lower switching temperatures to further reduce the WRITE energy of the P-RAM, and so reducing the threshold time where P-RAM is more efficient for storing information than SRAM.

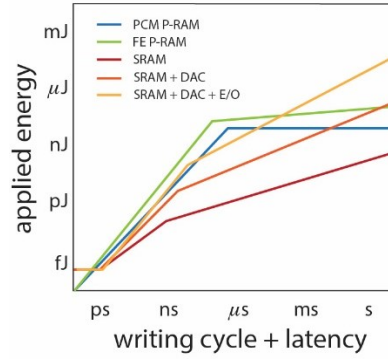


Fig. 3: Trend of total energy consumption for writing over time for P-RAM and SRAM. PCM-based P-RAM does not require additional energy once written, while FE-based P-RAM requires a DC voltage to read the information. SRAM requires a constant power to overcome internal leakage, power that becomes more prominent as DAC and E/O conversion are required to interface the optical waveguides.

C. Electrical-optical conversion

Conversion between the electrical and optical domains is already an overhead cost that many systems pay every day. Assessing the cost in terms of power and latency for these conversions shapes the system design and choice of memory, especially when considering neural networks. Considering electronic memories such as SRAM, the electrical signal needs to go through a DAC (~ 1 nJ and ~ 3 ns [79]), driving amplifier, and electro-optical modulator to convert it into an optical signal. In the same fashion, the detected optical signal requires a trans-impedance amplifier (TIA) and ADC to convert the processed data back to the electronic domain [80]. In this kind of architecture, where each step of the network has to perform a E/O/E conversion, it is straightforward to realize that scaling to multiple processing layers can introduce several problems,

Commented [YN1]: This is for register-based reads, correct?

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

such as the need to buffer intermediate information in an S-RAM cache, as well as limit the latency and efficiency of the network due to the DACs and ADCs. A full optical network, where the weights are stored in a nonvolatile fashion by means of P-RAM elements [26], [81], [82], the signals are converted once to the optical domain, and converted back once at the end of the network, would take full advantage of the wide bandwidth provided by the optical domain and extremely low latency and low energy consumption. However, a lack of efficient, nonlinear optical elements with low optical threshold powers currently limits the practicality of this approach for deep neural networks.

V. OUTLOOK AND ROADMAP FOR DATA STORAGE IN OPTICAL COMPUTING

A. Roadmap for electronic memories for optical computing

Efficient integration of high-density electronic storage with analog optical computing platforms is a challenge that requires alleviating (or removing) the energy-consuming digital-to-analog and electro-optical conversions. The simplest solution is seemingly to adopt a completely analog technology using, for instance, memristors in the electrical domain directly integrated to photonic waveguides [83]–[85]. DACs for data input and ADCs for data output are not needed if the optical processor is communicating with an analog environment and E/O conversion can be realized employing the same memristive element. However, the world runs on digital technology and computing with an analog architecture would certainly require data type conversion. The prospect of E/O conversion of digital signals using optical DACs (see Section IIIA and Table 1), and ideally also ADCs, open the possibility of faster operations with simplified circuitry. The latency can also be further optimized by bringing the electronic memory bank closer to the photonic processor using monolithic co-integration of nanoelectronics and photonics rather than using two separate chiplets [86].

Moreover, novel modulation approaches for electro-optical conversion are necessary to avoid the widespread use of thermo-optical control, which faces serious heating issues when scaling to hundreds of simultaneously operating devices. Similarly, faster carrier-based modulation faces high IL and large form factors—both of which are detrimental to computing tasks since the complexity of the photonic circuitry can afford neither. Optomechanical modulators [87], while still volatile unless using latches or bi-stability [88], [89], are potential CMOS-compatible platforms given their low insertion losses, low powers, and form factors comparable to thermo-optic modulators. Provided CMOS integration in the future, optical modulators based on 2D materials could provide an even closer to optimal platform for energy-efficient modulation [90].

B. Roadmap for photonic memories based on nonvolatile materials

Photonic integrated technologies, as available in current commercial foundries, must deal with large form factors due to waveguide footprints, a fact that could improve in the future by adopting smaller node CMOS fabrication processes to achieve reliable nanophotonic structures [18]. The current form-factor limitation means that electronics' storage densities of 10 Gb/mm² [91] are likely unachievable with photonic memories,

especially those based on material platforms directly embedded into the photonic circuits. Yet, the prospect of a novel optical memory class that, despite the lower storage density, can contribute to and enhance the performance of the memory hierarchy in hybrid optoelectronic architectures—especially photonic computational memory—is enough to motivate the development of an “ideal” photonic memory. The target performance metrics for optical memories (described in detail in Section II) are ultimately determined by the computing task at hand, just like the different electronic technologies in a Von Neumann computer's memory hierarchy. Whether volatile or nonvolatile, written with higher or lower frequency, etc., some features that any ideal photonic memory should have include:

1. CMOS compatibility for guaranteed scalability
2. Low IL comparable to the propagation loss of the platform (<1 dB/cm)
3. READ and WRITE energy consumption of <fJ and fJ-pJ, respectively
4. Large modulation depths >10 dB for amplitude modulation and at least 2π for phase modulation
5. WRITE cyclability > 10^8
6. Precision and stability that are not compromised by environmental effects such as temperature or material degradation

Despite the challenges described in Section III, there is still ample room for improved performance in nonvolatile photonic memory technologies. For instance, even though the PCM photonic memories come with limited endurance today (> 5×10^5 cycles [76]), there does not appear to be any intrinsic limitations that precludes them from reach endurance levels attained in PCM-based RF switches (1.5×10^8 cycles [92]) and electronic memories (> 2×10^{12} cycles [93]). Their energy consumption can also be minimized by searching for new PCM compositions with reduced liquidus temperature and fast crystallization kinetics, and/or further device optimization via engineering the effective device area's thermal capacitance. On the other hand, development of new FE crystals compatible with CMOS backend processing, such as HfO₂-based oxide alloys [94], [95], could potentially facilitate their integration with standard photonic integrated circuits. Finally, other alternative emerging nonvolatile integrated photonics platforms may also prove useful for photonic memory applications [96]–[98]. Whether backend, frontend, or eventually fully integrated into CMOS fabrication processes, the novel active material-based approaches require a scalable fabrication to guarantee high density photonic architectures and mass production.

C. Optical memories in edge/cloud computing

Alleviating the von Neumann bottleneck, especially if using fiber optics to store and fetch data—commonly done in data centers for cloud computing—is the longstanding promise of optical memories in conventional computers. This task is yet to be demonstrated given the complexity of realizing high-density optical storage, mostly due to the lack of fully CMOS compatible platforms and their large footprints. On the other hand, the development of fully integrated optical or electronic memory with a photonic processor either in a von Neumann [99] or brain-inspired architectures [8], [11], [100], together

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

with integrated light sources and photodetectors, can lead to the development of packaged devices with the portability and processing capacity required to enhance edge computing. Inference [11], [27] and high-throughput matrix-vector multiplications [18], [81] have already led to outstanding, high-performance demonstrations using on-chip photonic processors—systems that can be integrated to future edge computing devices.

ACKNOWLEDGMENT

N.Y. acknowledges support from the University of Pittsburgh Momentum Fund. C.R. acknowledges support from the Minta Martin Foundation through the University of Maryland. V.J.S. acknowledges support from the George Washington University Nanofabrication and Imaging Center (GWNIC).

REFERENCES

- [1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
- [2] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, “Theoretical Insights Into the Optimization Landscape of Over-Parameterized Shallow Neural Networks,” *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 742–769, Feb. 2019, doi: 10.1109/TIT.2018.2854560.
- [3] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, “The Computational Limits of Deep Learning,” *MIT INITIATIVE ON THE DIGITAL ECONOMY RESEARCH BRIEF*, vol. 4, Jul. 2020, [Online]. Available: <http://arxiv.org/abs/2007.05558>
- [4] N. P. Jouppi *et al.*, “In-Datacenter Performance Analysis of a Tensor Processing Unit,” in *Proceedings of the 44th Annual International Symposium on Computer Architecture - ISCA '17*, 2017, pp. 1–12, doi: 10.1145/3079856.3080246.
- [5] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, “Memory devices and applications for in-memory computing,” *Nature Nanotechnology*, Mar. 2020, doi: 10.1038/s41565-020-0655-z.
- [6] D. A. B. Miller, “Are optical transistors the logical next step?,” *Nature Photonics*, vol. 4, no. 1, pp. 3–5, Jan. 2010, doi: 10.1038/nphoton.2009.240.
- [7] D. A. B. Miller, “Attojoule Optoelectronics for Low-Energy Information Processing and Communications,” *Journal of Lightwave Technology*, vol. 35, no. 3, pp. 346–396, Feb. 2017, doi: 10.1109/JLT.2017.2647779.
- [8] B. J. Shastri *et al.*, “Photonics for artificial intelligence and neuromorphic computing,” *Nature Photonics*, vol. 15, no. 2, pp. 102–114, Feb. 2021, doi: 10.1038/s41566-020-00754-y.
- [9] G. Wetzstein *et al.*, “Inference in artificial intelligence with deep optics and photonics,” *Nature*, vol. 588, no. 7836, pp. 39–47, Dec. 2020, doi: 10.1038/s41586-020-2973-6.
- [10] N. C. Harris *et al.*, “Linear programmable nanophotonic processors,” *Optica*, vol. 5, no. 12, p. 1623, Dec. 2018, doi: 10.1364/OPTICA.5.001623.
- [11] Y. Shen *et al.*, “Deep learning with coherent nanophotonic circuits,” *Nature Photonics*, vol. 11, no. 7, pp. 441–446, Jun. 2017, doi: 10.1038/nphoton.2017.93.
- [12] T. Zhou *et al.*, “Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit,” *Nature Photonics*, vol. 15, no. 5, pp. 367–373, May 2021, doi: 10.1038/s41566-021-00796-w.
- [13] X. Lin *et al.*, “All-optical machine learning using diffractive deep neural networks,” *Science (1979)*, vol. 361, no. 6406, pp. 1004–1008, Sep. 2018, doi: 10.1126/science.aat8084.
- [14] 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.
- [15] 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.
- [16] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, “Broadcast and Weight: An Integrated Network For Scalable Photonic Spike Processing,” *Journal of Lightwave Technology*, vol. 32, no. 21, pp. 4029–4041, Nov. 2014, doi: 10.1109/JLT.2014.2345652.
- [17] M. Y.-S. Fang, S. Manipatruni, C. Wierzynski, A. Khosrowshahi, and M. R. DeWeese, “Design of optical neural networks with component imprecisions,” *Optics Express*, vol. 27, no. 10, p. 14009, May 2019, doi: 10.1364/OE.27.014009.
- [18] 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 Journal of Selected Topics in Quantum Electronics*, pp. 1–1, 2019, doi: 10.1109/JSTQE.2019.2941485.
- [19] F. Zokaee, Q. Lou, N. Youngblood, W. Liu, Y. Xie, and L. Jiang, “LightBulb: A Photonic-Nonvolatile-Memory-based Accelerator for Binarized Convolutional Neural Networks,” 2020.
- [20] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, “Binarized Neural Networks,” in *Advances in Neural Information Processing Systems 29 (NIPS)*, 2016, pp. 4107–4115.
- [21] Y. Umuroglu *et al.*, “FINN,” in *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, Feb. 2017, pp. 65–74, doi: 10.1145/3020078.3021744.
- [22] H. Noh, T. You, J. Mun, and B. Han, “Regularizing Deep Neural Networks by Noise: Its Interpretation and Optimization,” in *Advances in Neural Information Processing Systems*, 2017, vol. 30. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/file/217e342fc01668b10cb1188d40d3370e-Paper.pdf>

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

- [23] Z. You, J. Ye, K. Li, Z. Xu, and P. Wang, "Adversarial Noise Layer: Regularize Neural Network by Adding Noise," in *2019 IEEE International Conference on Image Processing (ICIP)*, Sep. 2019, pp. 909–913, doi: 10.1109/ICIP.2019.8803055.
- [24] C. Wu *et al.*, "Harnessing optoelectronic noises in a photonic generative network," *Science Advances*, vol. 8, no. 3, Jan. 2022, doi: 10.1126/sciadv.abm2956.
- [25] N. Youngblood, "Coherent Photonic Crossbar Arrays for Large-Scale Matrix-Matrix Multiplication," *IEEE Journal of Selected Topics in Quantum Electronics*, pp. 1–1, 2022, doi: 10.1109/JSTQE.2022.3171167.
- [26] M. Miscuglio and V. J. Sorger, "Photonic tensor cores for machine learning," *Applied Physics Reviews*, vol. 7, no. 3, p. 031404, Sep. 2020, doi: 10.1063/5.0001942.
- [27] 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.
- [28] J. Feldmann *et al.*, "Calculating with light using a chip-scale all-optical abacus," *Nature Communications*, vol. 8, no. 1, p. 1256, Dec. 2017, doi: 10.1038/s41467-017-01506-3.
- [29] A. Spinelli, C. Compagnoni, and A. Lacaita, "Reliability of NAND Flash Memories: Planar Cells and Emerging Issues in 3D Devices," *Computers*, vol. 6, no. 2, p. 16, Apr. 2017, doi: 10.3390/computers6020016.
- [30] R. Amin *et al.*, "ITO-based electro-absorption modulator for photonic neural activation function," *APL Materials*, vol. 7, no. 8, p. 081112, Aug. 2019, doi: 10.1063/1.5109039.
- [31] S. Moazeni *et al.*, "A 40-Gb/s PAM-4 Transmitter Based on a Ring-Resonator Optical DAC in 45-nm SOI CMOS," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 12, pp. 3503–3516, Dec. 2017, doi: 10.1109/JSSC.2017.2748620.
- [32] P. Dong *et al.*, "Thermally tunable silicon racetrack resonators with ultralow tuning power," *Optics Express*, vol. 18, no. 19, p. 20298, Sep. 2010, doi: 10.1364/OE.18.020298.
- [33] E. Timurdogan, C. M. Sorace-Agaskar, J. Sun, E. Shah Hosseini, A. Biberman, and M. R. Watts, "An ultralow power athermal silicon modulator," *Nature Communications*, vol. 5, no. 1, p. 4008, Sep. 2014, doi: 10.1038/ncomms5008.
- [34] X. Wu *et al.*, "A 20Gb/s NRZ/PAM-4 1V transmitter in 40nm CMOS driving a Si-photonic modulator in 0.13 μ m CMOS," in *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, Feb. 2013, pp. 128–129, doi: 10.1109/ISSCC.2013.6487667.
- [35] N. C. Harris *et al.*, "Efficient, compact and low loss thermo-optic phase shifter in silicon," *Optics Express*, vol. 22, no. 9, p. 10487, May 2014, doi: 10.1364/OE.22.010487.
- [36] D. A. B. Miller, "Self-configuring universal linear optical component [Invited]," *Photonics Research*, vol. 1, no. 1, p. 1, Jun. 2013, doi: 10.1364/PRJ.1.000001.
- [37] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, "Experimental realization of any discrete unitary operator," *Physical Review Letters*, vol. 73, no. 1, pp. 58–61, Jul. 1994, doi: 10.1103/PhysRevLett.73.58.
- [38] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walsmley, "Optimal design for universal multiport interferometers," *Optica*, vol. 3, no. 12, p. 1460, Dec. 2016, doi: 10.1364/OPTICA.3.001460.
- [39] D. A. B. Miller, "Self-aligning universal beam coupler," *Optics Express*, vol. 21, no. 5, p. 6360, Mar. 2013, doi: 10.1364/OE.21.006360.
- [40] W. Bogaerts *et al.*, "Silicon microring resonators," *Laser & Photonics Reviews*, vol. 6, no. 1, pp. 47–73, Jan. 2012, doi: 10.1002/lpor.201100017.
- [41] C. Huang *et al.*, "Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits," *APL Photonics*, vol. 5, no. 4, p. 040803, Apr. 2020, doi: 10.1063/1.5144121.
- [42] M. R. Watts, J. Sun, C. DeRose, D. C. Trotter, R. W. Young, and G. N. Nielson, "Adiabatic thermo-optic Mach-Zehnder switch," *Optics Letters*, vol. 38, no. 5, p. 733, Mar. 2013, doi: 10.1364/OL.38.000733.
- [43] A. H. Atabaki, A. A. Eftekhari, S. Yegnanarayanan, and A. Adibi, "Sub-100-nanosecond thermal reconfiguration of silicon photonic devices," *Optics Express*, vol. 21, no. 13, p. 15706, Jul. 2013, doi: 10.1364/OE.21.015706.
- [44] C. Rios *et al.*, "Integrated all-photonic non-volatile multi-level memory," *Nature Photonics*, vol. 9, no. 11, pp. 725–732, Oct. 2015, doi: 10.1038/nphoton.2015.182.
- [45] J. Geler-Kremer *et al.*, "A ferroelectric multilevel non-volatile photonic phase shifter," *Nature Photonics*, May 2022, doi: 10.1038/s41566-022-01003-0.
- [46] C. A. Barrios and M. Lipson, "Silicon photonic read-only memory," *Journal of Lightwave Technology*, vol. 24, no. 7, pp. 2898–2905, Jul. 2006, doi: 10.1109/JLT.2006.875964.
- [47] 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/acsp Photonics.9b00819.
- [48] J. Zheng *et al.*, "Nonvolatile Electrically Reconfigurable Integrated Photonic Switch Enabled by a Silicon PIN Diode Heater," *Advanced Materials*, vol. 32, no. 31, p. 2001218, Jun. 2020, doi: 10.1002/adma.202001218.
- [49] N. Farmakidis *et al.*, "Electronically Reconfigurable Photonic Switches Incorporating Plasmonic Structures and Phase Change Materials," *Advanced Science*, p. 2200383, 2022, doi: 10.1002/ADVS.202200383.
- [50] M. Grajower, N. Mazurski, J. Shappir, and U. Levy, "Non-Volatile Silicon Photonics Using Nanoscale Flash Memory Technology," *Laser & Photonics*

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

- Reviews*, vol. 12, no. 4, p. 1700190, Apr. 2018, doi: 10.1002/LPOR.201700190.
- [51] Y. Zhang *et al.*, "Electrically reconfigurable non-volatile metasurface using low-loss optical phase-change material," *Nature Nanotechnology*, vol. 16, pp. 661–666, 2021, doi: 10.1038/s41565-021-00881-9.
- [52] M. Mishra, N. R. Das, A. Melloni, and F. Morichetti, "Modelling domain switching of ferroelectric BaTiO₃ integrated in silicon photonic waveguides," *Optics Communications*, vol. 448, pp. 19–25, Oct. 2019, doi: 10.1016/J.OPTCOM.2019.05.001.
- [53] X. Li *et al.*, "Fast and reliable storage using a 5 bit, nonvolatile photonic memory cell," *Optica*, vol. 6, no. 1, p. 1, Jan. 2019, doi: 10.1364/OPTICA.6.000001.
- [54] D. Yao *et al.*, "Energy-efficient non-volatile ferroelectric based electrostatic doping multilevel optical readout memory," *Optics Express*, vol. 30, no. 8, pp. 13572–13582, Apr. 2022, doi: 10.1364/OE.456048.
- [55] 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," *Nature Communications*, vol. 12, p. 96, Dec. 2021, doi: 10.1038/s41467-020-20365-z.
- [56] C. Ríos *et al.*, "In-memory computing on a photonic platform," *Science Advances*, vol. 5, no. 2, p. eaau5759, Feb. 2019, doi: 10.1126/sciadv.aau5759.
- [57] 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.
- [58] J. F. Song *et al.*, "Integrated photonics with programmable non-volatile memory," *Scientific Reports*, vol. 6, no. 1, p. 22616, Mar. 2016, doi: 10.1038/srep22616.
- [59] C. Ríos *et al.*, "Ultra-compact nonvolatile photonics based on electrically reprogrammable transparent phase change materials," *arXiv:2105.06010*, 2021.
- [60] N. Youngblood, "Coherent Photonic Crossbar Arrays for Large-Scale Matrix-Matrix Multiplication," *IEEE Journal of Selected Topics in Quantum Electronics*, 2022, doi: 10.1109/JSTQE.2022.3171167.
- [61] Y. Zhang *et al.*, "Myths and truths about optical phase change materials: A perspective," *Applied Physics Letters*, vol. 118, no. 21, p. 210501, May 2021, doi: 10.1063/5.0054114.
- [62] L. Martin-Monier *et al.*, "Endurance of Chalcogenide Optical Phase Change Materials: a Review," *Optical Materials Express*, vol. 12, no. 6, pp. 2145–2167, Jun. 2022, doi: 10.1364/ome.456428.
- [63] R. Cao *et al.*, "Improvement of Endurance in HZO-Based Ferroelectric Capacitor Using Ru Electrode," *IEEE Electron Device Letters*, vol. 40, no. 11, pp. 1744–1747, Nov. 2019, doi: 10.1109/LED.2019.2944960.
- [64] K. Vandoorne *et al.*, "Experimental demonstration of reservoir computing on a silicon photonics chip," *Nature Communications*, vol. 5, pp. 1–6, 2014, doi: 10.1038/ncomms4541.
- [65] D. Brunner, B. Penkovsky, B. A. Marquez, M. Jacquot, I. Fischer, and L. Larger, "Tutorial: Photonic neural networks in delay systems," *Journal of Applied Physics*, vol. 124, no. 15, p. 152004, Oct. 2018, doi: 10.1063/1.5042342.
- [66] S. Xu, J. Wang, S. Yi, and W. Zou, "High-order tensor flow processing using integrated photonic circuits," Dec. 2021.
- [67] Y. Zhang *et al.*, "Broadband transparent optical phase change materials for high-performance nonvolatile photonics," *Nature Communications*, vol. 10, no. 1, p. 4279, Dec. 2019, doi: 10.1038/s41467-019-12196-4.
- [68] R. Chen, Z. Fang, J. E. Fröch, P. Xu, J. Zheng, and A. Majumdar, "Broadband Nonvolatile Electrically Controlled Programmable Units in Silicon Photonics," *ACS Photonics*, vol. 9, no. 6, pp. 2142–2150, Jun. 2022, doi: 10.1021/acsp Photonics.2c00452.
- [69] J. Meng, M. Miscuglio, and V. J. Sorger, "Multi-level Nonvolatile Photonic Memories Using Broadband Transparent Phase change materials," in *OSA Advanced Photonics Congress 2021*, 2021, p. IF3A.2, doi: 10.1364/IPRSN.2021.IF3A.2.
- [70] K. Kato, M. Kuwahara, H. Kawashima, T. Tsuruoka, and H. Tsuda, "Current-driven phase-change optical gate switch using indium-tin-oxide heater," *Applied Physics Express*, vol. 10, no. 7, 2017, doi: 10.7567/APEX.10.072201.
- [71] X. Li *et al.*, "Fast and reliable storage using a 5 bit, nonvolatile photonic memory cell," *Optica*, vol. 6, no. 1, p. 1, Jan. 2019, doi: 10.1364/OPTICA.6.000001.
- [72] Xiaoyao Liang, Kerem Turgay, and D. Brooks, "Architectural power models for sram and cam structures based on hybrid analytical/empirical techniques," in *2007 IEEE/ACM International Conference on Computer-Aided Design*, Nov. 2007, pp. 824–830. doi: 10.1109/ICCAD.2007.4397367.
- [73] A. Biswas and A. P. Chandrakasan, "CONV-SRAM: An Energy-Efficient SRAM With In-Memory Dot-Product Computation for Low-Power Convolutional Neural Networks," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 217–230, Jan. 2019, doi: 10.1109/JSSC.2018.2880918.
- [74] D. P. Ravipati, R. Kedia, V. M. van Santen, J. Henkel, P. R. Panda, and H. Amrouch, "FN-CACTI: Advanced CACTI for FinFET and NC-FinFET Technologies," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 30, no. 3, pp. 339–352, Mar. 2022, doi: 10.1109/TVLSI.2021.3123112.
- [75] A. Shafaei, Y. Wang, X. Lin, and M. Pedram, "FinCACTI: Architectural Analysis and Modeling of Caches with Deeply-Scaled FinFET Devices," in *2014 IEEE Computer Society Annual Symposium on VLSI*, Jul. 2014, pp. 290–295. doi: 10.1109/ISVLSI.2014.94.
- [76] J. Meng *et al.*, "Electrical Programmable Low-loss high cyclable Nonvolatile Photonic Random-Access Memory," *arXiv:2203.13337*, Mar. 2022, Accessed: Jul. 05, 2022. [Online]. Available: <https://arxiv.org/abs/2203.13337v4>

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

- [77] C. Ríos *et al.*, “Integrated all-photonic non-volatile multi-level memory,” *Nature Photonics*, vol. 9, no. 11, pp. 725–732, Sep. 2015, doi: 10.1038/nphoton.2015.182.
- [78] J. Zheng *et al.*, “Nonvolatile Electrically Reconfigurable Integrated Photonic Switch Enabled by a Silicon PIN Diode Heater,” *Advanced Materials*, vol. 32, no. 31, p. 2001218, Aug. 2020, doi: 10.1002/adma.202001218.
- [79] Juanda, W. Shu, and J. S. Chang, “A Calibration-Free/DEM-Free 8-bit 2.4-GS/s Single-Core Digital-to-Analog Converter With a Distributed Biasing Scheme,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 11, pp. 2299–2309, Nov. 2018, doi: 10.1109/TVLSI.2018.2850919.
- [80] F. Ashtiani, A. J. Geers, and F. Aflatouni, “An on-chip photonic deep neural network for image classification,” *Nature*, vol. 606, no. 7914, pp. 501–506, Jun. 2022, doi: 10.1038/s41586-022-04714-0.
- [81] 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.
- [82] 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,” *Nature Communications*, vol. 12, no. 1, p. 96, Dec. 2021, doi: 10.1038/s41467-020-20365-z.
- [83] B. M. Tossoun, X. Sheng, J. P. Strachan, D. Liang, and R. G. Beausoleil, “Hybrid memristor optoelectronic integrated circuits for optical computing,” in *Smart Photonic and Optoelectronic Integrated Circuits 2022*, Mar. 2022, p. 11. doi: 10.1117/12.2614073.
- [84] B. Tossoun, X. Sheng, J. P. Strachan, D. Liang, and R. G. Beausoleil, “Memristor Photonics,” in *Photonics in Switching and Computing 2021*, 2021, p. Tu5B.3. doi: 10.1364/PSC.2021.Tu5B.3.
- [85] K. Portner *et al.*, “Analog Nanoscale Electro-Optical Synapses for Neuromorphic Computing Applications,” *ACS Nano*, vol. 15, no. 9, pp. 14776–14785, Sep. 2021, doi: 10.1021/acsnano.1c04654.
- [86] 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, doi: 10.1038/s41586-018-0028-z.
- [87] C. Errando-Herranz, A. Y. Takabayashi, P. Edinger, H. Sattari, K. B. Gylfason, and N. Quack, “MEMS for Photonic Integrated Circuits,” *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 26, no. 2, 2020, doi: 10.1109/JSTQE.2019.2943384.
- [88] H. Sattari, A. Toros, T. Graziosi, and N. Quack, “Bistable silicon photonic MEMS switches,” in *MOEMS and Miniaturized Systems XVIII*, Mar. 2019, p. 13. doi: 10.1117/12.2507192.
- [89] P. Edinger *et al.*, “A Bistable Silicon Photonic Mems Phase Switch For Nonvolatile Photonic Circuits,” in *2022 IEEE 35th International Conference on Micro Electro Mechanical Systems Conference (MEMS)*, Jan. 2022, pp. 995–997. doi: 10.1109/MEMS51670.2022.9699739.
- [90] I. Datta *et al.*, “Low-loss composite photonic platform based on 2D semiconductor monolayers,” *Nature Photonics*, vol. 14, no. 4, pp. 256–262, Apr. 2020, doi: 10.1038/s41566-020-0590-4.
- [91] A. Goda, “Recent Progress on 3D NAND Flash Technologies,” *Electronics (Basel)*, vol. 10, no. 24, p. 3156, Dec. 2021, doi: 10.3390/electronics10243156.
- [92] J.-S. Moon *et al.*, “Reconfigurable infrared spectral imaging with phase change materials,” in *Micro- and Nanotechnology Sensors, Systems, and Applications XI*, May 2019, p. 32. doi: 10.1117/12.2519492.
- [93] W. Kim *et al.*, “ALD-based confined PCM with a metallic liner toward unlimited endurance,” in *2016 IEEE International Electron Devices Meeting (IEDM)*, Dec. 2016, pp. 4.2.1–4.2.4. doi: 10.1109/IEDM.2016.7838343.
- [94] M. Halter *et al.*, “Back-End, CMOS-Compatible Ferroelectric Field-Effect Transistor for Synaptic Weights,” *ACS Applied Materials and Interfaces*, vol. 12, no. 15, pp. 17725–17732, Apr. 2020, doi: 10.1021/ACSAMI.0C00877/SUPPL_FILE/AM0C00877_SI_001.PDF.
- [95] J. Qin *et al.*, “Enhanced Second Harmonic Generation from Ferroelectric HfO₂-Based Hybrid Metasurfaces,” *ACS Nano*, vol. 13, no. 2, pp. 1213–1222, Feb. 2019, doi: 10.1021/ACS.NANO.8B06308/ASSET/IMAGES/LARGE/NN-2018-06308K_0006.JPEG.
- [96] C. Lian, C. Vagionas, T. Alexoudi, N. Pleros, N. Youngblood, and C. Ríos, “Photonic (computational) memories: tunable nanophotonics for data storage and computing,” *Nanophotonics*, 2022, doi: 10.1515/NANOPH-2022-0089.
- [97] J. Parra, I. Olivares, A. Brimont, and P. Sanchis, “Toward Nonvolatile Switching in Silicon Photonic Devices,” *Laser & Photonics Reviews*, vol. 15, no. 6, p. 2000501, Jun. 2021, doi: 10.1002/LPOR.202000501.
- [98] Y. Zhai *et al.*, “Toward non-volatile photonic memory: concept, material and design,” *Materials Horizons*, vol. 5, no. 4, pp. 641–654, 2018, doi: 10.1039/C8MH00110C.
- [99] A. Narayan, Y. Thonnart, P. Vivet, A. K. Coskun, and A. Joshi, “Architecting Optically-Controlled Phase Change Memory,” Jul. 2021, Accessed: Jul. 05, 2022. [Online]. Available: <http://arxiv.org/abs/2107.11516>
- [100] V. Bangari *et al.*, “Digital Electronics and Analog Photonics for Convolutional Neural Networks (DEAP-CNNs),” *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 26, no. 1, pp. 1–13, Jan. 2020, doi: 10.1109/JSTQE.2019.2945540.

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

Sadra Rahimi Kari, (Student Member, IEEE)



Carlos A. Ríos Ocampo (Carlos Ríos) (Member, IEEE) is an Assistant Professor at the University of Maryland, College Park. He received a Ph.D. from the University of Oxford, UK (2017), a M.Sc. from Karlsruhe Institute of Technology, Germany (2014), and a B.S. in Physics from the University of Antioquia, Colombia (2010). Prior to joining UMD, Carlos was a Postdoctoral Associate at MIT between 2018 and 2021. Carlos's scientific interests focus on studying and developing new on-chip technologies driven by optical nanomaterials and nanophotonics.

Fellow of Optica (former OSA), a Fellow of SPIE, a Fellow of the German National Academic Foundation, and a Senior Member of IEEE. He is a co-founder of Optelligence Company.

Lei Jiang, (Member, IEEE)



Jiawei Meng (Student member, IEEE) received his B.S. (2017) in computer engineering from Miami Univ. in Oxford, OH and the M.S. (2019) in electrical engineering from The George Washington Univ., Washington, DC. He is currently working on his Ph.D. in electrical engineering at The George Washington Univ., Washington, DC. His research

interests include photonic integrated circuit analysis and design, phase change material (PCM) on integrated circuits, specifically designing for photonic random-access memory.



Dr. Nicola Peserico (Member, IEEE) received his PhD at Politecnico di Milano (Italy) in 2018. In 2019, he joined Femtorays (Italy), a silicon photonics startup for biosensing. He is now a Post-doc Researcher in the Department of Electrical and Computer Engineering at the

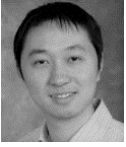
George Washington University, Washington, DC. His research area includes silicon photonics, AI/ML accelerators, optoelectronics devices and components, and bio-sensing with photonic integrated circuits.



Dr. Volker J. Sorger (Senior member, IEEE) is an Associate Professor in the Department of Electrical and Computer Engineering and the Director of the Institute on AI & Photonics, the Head of the Devices & Intelligent Systems Laboratory at the George Washington University. His research

areas include devices & optoelectronics, AI/ML accelerators, mixed-signal ASICs, quantum matter & quantum processors, cryptography. For his work, Dr. Sorger received multiple awards including the Presidential PECASE Award, the AFOSR YIP Award, the Emil Wolf Prize, and the National Academy of Sciences award of the year. Dr. Sorger is an Associate editor for OPTICA, serves on the board of Chip, and was the former editor-in-chief of Nanophotonics. He is a

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <



Juejun (JJ) Hu (Member, IEEE), is currently a Professor of Materials Science and Engineering at MIT. He holds a Ph.D. degree (2009) from MIT and a B.S. degree (2004) from Tsinghua University, China, both in Materials Science and Engineering. Prior to joining MIT, Hu was an Assistant Professor at the University of Delaware from 2010 to 2014. His primary research interest covers new optical materials exemplified by chalcogenide compounds, as well as enhanced photon-matter interactions in nanophotonic structures.



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.