Photonic-Electronic Integrated Circuits for High-Performance Computing and AI Accelerators

Shupeng Ning, Hanqing Zhu, Chenghao Feng, *Member, IEEE*, Jiaqi Gu, *Member, IEEE*, Zhixing Jiang, Zhoufeng Ying, Jason Midkiff, Sourabh Jain, May H. Hlaing, David Z. Pan, *Fellow, IEEE*, and Ray T. Chen, *Fellow, IEEE*

(Invited Paper)

***Abstract*—In recent decades, the demand for computational power has surged, particularly with the rapid expansion of artificial intelligence (AI). As we navigate the post-Moore's law era, the limitations of traditional electrical digital computing, including process bottlenecks and power consumption issues, are propelling the search for alternative computing paradigms. Among various emerging technologies, integrated photonics stands out as a promising solution for next-generation high-performance computing, thanks to the inherent advantages of light, such as low latency, high bandwidth, and unique multiplexing techniques. Furthermore, the progress in photonic integrated circuits (PICs), which are equipped with abundant photoelectronic components, positions photonic-electronic integrated circuits as a viable solution for high-performance computing and hardware AI accelerators. In this review, we survey recent advancements in both PIC-based digital and analog computing for AI, exploring the principal benefits and obstacles of implementation. Additionally, we propose a comprehensive analysis of photonic AI from the perspectives of hardware implementation, accelerator architecture, and software-hardware co-design. In the end, acknowledging the existing challenges, we underscore potential strategies for overcoming these issues and offer insights into the future drivers for optical computing.**

***Index Terms*—AI accelerator, optical computing, optical neural network, photonic integrated circuit, silicon photonics.**

I. Introduction

As THE semiconductor industry advances to process nodes below 3 nanometers, it increasingly encounters inherent physical limitations of both devices and materials [1], [2]. A primary concern is the surge in power consumption as clock frequencies reach gigahertz levels, leading to overwhelming heat generation [2], [3]. Furthermore, at these diminutive scales, quantum uncertainties begin to dominate electron behavior, resulting in increased transistor errors and reduced reliability. Additionally, AI has made remarkable strides in recent years, exerting a growing influence on various aspects of our lives, such as image recognition [4], [5], [6], [7], natural language processing [8], [9], autonomous driving [10], and medical diagnosis [11], [12], which have further increased societal demand for computational power. One notable example is the emergence of large language models (LLMs) such as GPT (Generative Pre-trained Transformer). These models exhibit human-level intelligence and have revolutionized a wide range of applications, from sophisticated chatbots to advanced text analysis tools. However, the advancements of deep neural networks (DNNs) are driven by rapidly increasing model sizes and data volumes, which necessitate significantly expanding computational demands. For instance, the GPT-3 model developed by OpenAI, which contains around 175 billion parameters, requires 14.8 days for training using a cluster of around 10,000 NVIDIA V100 GPUs, with an estimated energy consumption of 1287 MWh [13], [14], [15]. Hence, in the post-Moore's Law era, traditional electronic computing architectures, designed to execute sequential, digital programs, are inadequate to meet the surging demand for high-performance computing and AI tasks. There is a pressing need to develop processing units capable of performing high-speed, energy-efficient computing. In response, both industry and academia are actively exploring alternative avenues from novel materials [16], [17], architectures [4], [18], to the investigation of new computational paradigms.

Among the emerging technologies, integrated photonics is a promising candidate for next-generation computation that can overcome the bottlenecks of their electrical counterparts. First, the speed of optical signals travel within optical waveguides surpasses that of electron-based transit through transistors with multiple fanouts by 1-2 orders of magnitude [19]. The delay and loss in waveguides are primarily determined by the optical path length. Additionally, a series of high-speed and energy-efficient

Manuscript received 15 March 2024; revised 21 June 2024; accepted 10 July 2024. Date of publication 15 July 2024; date of current version 16 November 2024. This work was supported by AFOSR under Grant FA9550-23-1-0452 and Grant FA9550-17-1-0071. *(Shupeng Ning and Hanqing Zhu are co-first authors.) (Corresponding author: Ray T. Chen.)*

Shupeng Ning, Hanqing Zhu, Chenghao Feng, Zhixing Jiang, Zhoufeng Ying, Jason Midkiff, Sourabh Jain, May H. Hlaing, David Z. Pan, and Ray T. Chen are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: chenrt@austin.utexas.edu).

Jiaqi Gu is with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA, and also with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281 USA.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JLT.2024.3427716.

Digital Object Identifier 10.1109/JLT.2024.3427716

(operating on the order of sub-picojoule per bit) devices for optical computing have been developed [20], [21], [22]. Notably, the power consumption of transistor-based electrical circuits exhibits a cubic relationship with the clock frequency f [23], whereas photonics-electronic platforms scale only linearly with f [24], effectively relaxing the frequency constraints associated with the power wall issue. Furthermore, as bosons, photons do not conform to the Pauli exclusion principle, allowing for the utilization of unique multiplexing techniques such as wavelength division multiplexing (WDM), which further increases the overall bandwidth. Moreover, compared to another photonic computing scheme in the form of free-space diffraction [25], [26], integrated photonics offers superior compactness for higher-level integration. The advancement of silicon photonics has enabled the implementation of optical computing on low-cost PICs with high integration density, leveraging CMOS-compatible silicon manufacturing techniques. As an increasing number of foundries develop their validated process design kits (PDKs), the integrated photonics industry is progressively moving towards standardization similar to that of the fabless semiconductor industry [27]. This trend not only improves accessibility for designers and users but also offers more reliable performance.

Integrated photonics has emerged as a promising platform for AI accelerators, benefiting from its inherent attributes of high parallelism, low latency, and low power consumption. In the last decade, a diverse range of PIC-based optical neural networks (ONNs) that implement multilayer perceptrons (MLPs) [28], convolutional neural networks (CNNs) [26], [29], [30], spiking neural networks (SNNs) [31], [32], etc., have been reported, demonstrating remarkable performance on machine learning tasks. The fundamental operations of neural networks, involving data transfers and tensor operations, are achieved through the combination of passive optical devices and high-performance active photonic-electronic components [33], [34]. Specifically, optical signals can be modulated by electrical signals and "multiplied" in accordance with the transmission function of the PIC. The hybrid photonic-electronic platform combines the adaptability of electronic control with the high-speed capabilities of optical computing. Recently, cutting-edge optical processing units have been reported with a matrix processing speed of 3.8 trillion operations per second (TOPS) via time-wavelength multiplexing [30], while other works demonstrated ultra-low power consumption on the order of sub-femtojoules per bit [22].

While integrated photonics offers new opportunities, existing photonic-electronic computing systems still encounter several practical challenges, such as:

- The typical micron-scale dimensions of optical elements in PICs are significantly larger than the transistors in cutting-edge VLSI technologies. Besides, a range of practical issues, such as footprint, control complexity, and accumulated loss, etc., limit the functionality and scalability of PICs for advanced computing applications.
- The widespread reliance on electrical components for electro-optical (E-O) modulation, parameter updates, data transfer, and analog-to-digital/digital-to-analog (A/D, D/A) conversion in photonic-electronic platforms leads to considerable energy consumption.
- The inherent challenges in PICs, such as training algorithms, on-chip implementation of nonlinearity for ONNs and system robustness against noise and crosstalk, require careful consideration in both hardware design and software coordination.

This review focuses on recent progress in photonic-electronic integrated circuits for computing. Spanning from digital computing to analog AI accelerators, this paper is structured as follows. Section II begins with an overview of the fundamental blocks in PIC-based digital computing, followed by a survey of recent highlights ranging from the implementation of logic gates to fully functional photonic processing units. Section III focuses on the implementation of ONNs, covering aspects of photonic tensor cores, nonlinearity, and hardware-aware training strategies. Beyond a review from the device and circuit level, Sections IV and V provide a comprehensive analysis of recent photonic AI efforts from the perspectives of accelerator architectures and software-hardware co-design, respectively. The review culminates with Section VI, which offers an outlook on PIC-based optical computing and provides a summarizing conclusion.

II. Survey of Optical Digital Computing With PICs

PICs are comprised of a range of optical components, both passive and active, featuring various hardware implementations and circuit topologies to fulfill distinct functionalities. This section will focus on E-O digital logic, and provides a concise overview of recent progress in PIC-based digital computing, while highlighting these implementation techniques and associated challenges.

A. Optical Logic Gates

In the digital domain, both input and output are binary, and the resolution is defined by the number of bits and remains unaffected by the circuit size. A range of building blocks for optical digital computing on integrated photonic platforms, such as optical switches [35], modulators [22], [36], interconnects and photodetectors [37], [38], [39], [40], have been experimentally demonstrated. Basic logic operations (NOT, AND, OR, XOR, etc.), which are fundamental elements of digital systems, have been implemented by diverse PICs. Among these, electro-optic logic, also known as optical-directed logic, has been widely investigated by many research groups as well as foundries. As shown in Fig. 1(a), all E-O devices in the functional block, such as Mach-Zehnder Interferometers (MZIs), microring resonators (MRRs), microdisks, etc., are simultaneously configured by electrical signals. When light traverses the block, optical signals are modulated to execute logic operations in accordance with the PIC design and then propagated downstream or detected by monitors to read out the results. An important feature of the E-O digital logic is that each device is controlled by independent electrical input simultaneously and that signals are transmitted via light without the limitation of RC time constant and delay accumulations inherent in electrical systems. In other words, E-O logic merges the convenience and flexibility of electrical control with the high-speed capabilities of optical

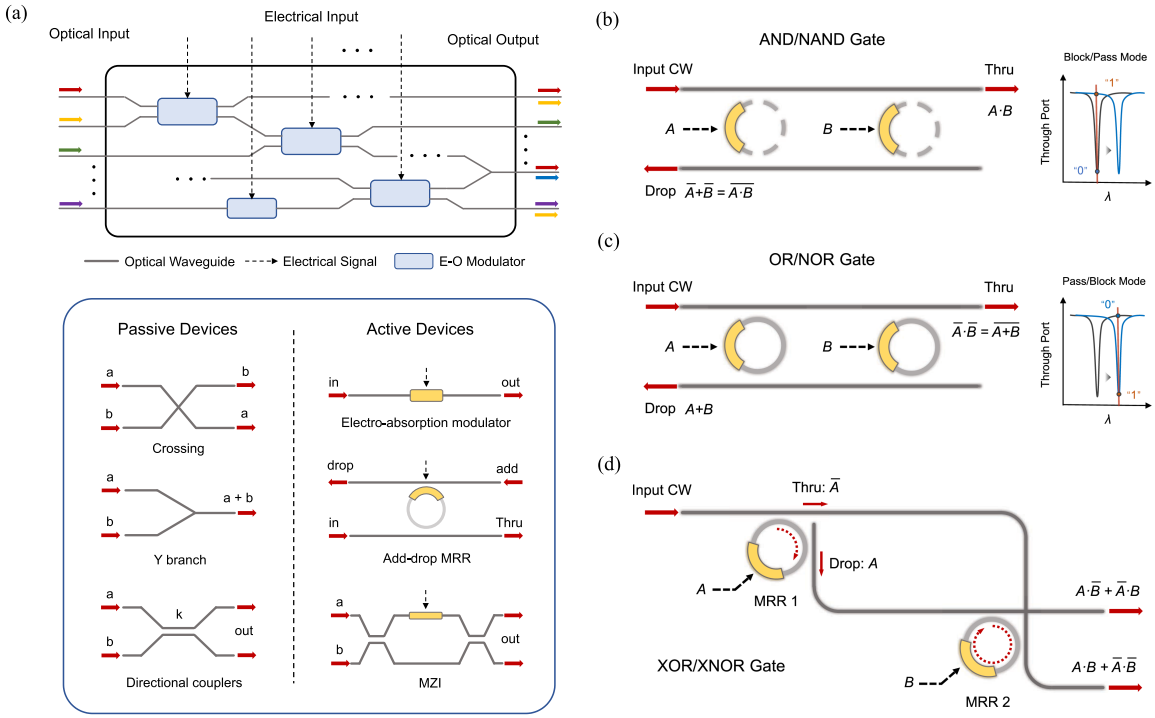

Fig. 1. Implementations of electro-optic logic gates. (a) Schematic diagram of E-O logic composed of passive and active optical components. During each clock cycle, electrical inputs are used to configure the logic circuit, while light carries out logic operations based on the transmission characteristic of the logic block. (b)–(d) Schematic of MRR-based AND/NAND, OR/NOR, and XOR/XNOR gates, as proposed in Ref [41], [42]. In the diagram, the dotted line and the solid line represent MRRs functioning in the "block/pass" and "pass/block" modes, respectively. These configurations correspond to the outputs of "0" and "1" at the through port, given a logic "0" as the electrical input.

computing. Fig. 1(b)-(c) show examples of E-O logic gates using two cascaded MRRs to perform 2-input AND/NAND and OR/NOR operations [41]. The proposed optical logic gate leverages the transmission characteristic of add-drop MRRs, which work as optical switches to implement logic operations and generate complementary outputs at the through and drop ports. Additionally, by configuring the resonant states of MRRs with "0" operands via selecting continuous wave (CW) inputs at either on-resonance or off-resonance wavelengths, a single photonic circuit can execute a variety of logic functions. With consistent logic but varying PIC topologies, XOR and XNOR gates featuring a crossbar structure also have been demonstrated [42] (Fig. 1(d)). Operating with similar mechanisms, MZI-based optical gates have also been extensively developed [43], [44].

In addition to E-O modulation, all-optical logic gate devices have also attracted attention. These devices have been experimentally demonstrated using various structures and phenomena, including photonic crystals [45], [46], surface plasmon polaritons (SPPs) [47], nanowire networks [48], and slot waveguides [49], [50]. A comprehensive classification, explanation of mechanisms, and comparative analysis of all-optical logic gates has been detailed in the prior review [51], [52], [53]. Compared to E-O logic, all-optical logic offers the potential for higher operation speed and bandwidth without extra energy consumption associated with O-E-O conversion. However, the implementation of all-optical logic confronts several practical challenges. Firstly, the complexity and requirements of design and fabrication (e.g., the transmission characteristic of photonic crystal is highly sensitive to its lattice constant) lead to inherent instability and a low contrast ratio between logical states [50]. Secondly, while all-optical gates can be more energy-efficient in signal processing, they often require higher optical signal power (mW-level) to compensate for higher losses or to induce the necessary nonlinear effects for switching. Furthermore, the limited functionality and scalability of all-optical logic, coupled with its higher cost, restrict its widespread application compared to E-O logic.

In recent years, integrated photonics have expanded beyond traditional classical optics, emerging as a compelling platform for quantum information science. Quantum logic gates based on the aforementioned active/passive devices have been widely reported [54], [55], [56]. Quantum PICs offer several significant advantages over bulk optics in the realm of quantum computing. First, PICs enable precise control of phase, polarization, and spatial mode with higher stability, which are essential for manipulating quantum states. Second, silicon exhibits a high third-order nonlinear coefficient $\chi^{(3)}$, facilitates the effective implementation of on-chip single/entangled photon sources through optical processes such as four-wave mixing (FWM) [57]. Third, PICs can integrate the fundamental building blocks of quantum computing–such as photon source, modulators, and single-photon detectors, etc., –in monolithic, hybrid, or heterogeneous configurations [58]. This integration yields scalable, robust, and reconfigurable circuits capable of handling complex quantum computing tasks [57], [59], [60].

B. Combinational Logic and Reconfigurable PICs

In digital circuits, the output of a combinational logic unit is determined solely by the current input combination, without dependence on previous states. Similarly, the implementation of optical combinational logic could begin with extracting the logical expression from its truth table, followed by designing the corresponding PIC based on the simplified expression. The implementation can rely on assembling fundamental logic gates or leveraging the unique characteristics of optical components or multiplexing techniques for fewer devices and compact layouts. Employing these strategies, diverse optical combinational logic units, including but not limited to adders [61], comparators [62], encoders [63], [64] and decoders [65], have been reported.

The aforementioned optical logic gates and units, tailored for specific tasks, are constrained by a fixed or limited logic representation. The inherent limitation not only complicates the development process but also increases the cost. To address this challenge, reconfigurable PICs offer a promising solution by programming the operational states of optical switches within a pre-designed framework using additional signals. The input operands and the reconfiguration signal can be independently managed by two separate control units within the modulator, such as the dual arms of MZIs. Generally, reconfiguration signals do not require high-speed modulation to the same extent as the input signals used as logic operands. Qiu et al. proposed a reconfigurable logic unit based on MRRs embedded with two modulation mechanisms [66]. As is shown in Fig. 2(a), the logic operand is modulated at high speeds via the *p-i-n* junction operating in the carrier injection mode, while the resonant state of MRR can be reconfigured using the microheater. Additionally, multi-operand modulators and non-volatile devices are also promising candidates for reconfigurable PICs [67], [68], [69].

An arbitrary combinational logic expression Y with n inputs $X_1, X_2, \ldots, X_n$ can be represented as a sum of products derived from these inputs, which can be expressed as:

$$Y = y_1 + y_2 + \ldots + y_m, \text{ where } y_i = \prod_{k=1}^{n} X_k \text{ or } \overline{X_k} \quad (1)$$

In this expression, $\overline{X_k}$ denotes the complement logic of input X_k. Using reconfigurable optical switches, the architecture illustrated in Fig. 2(b)-(c) theoretically can implement arbitrary logic functions conforming to the expression format presented in (1). The product term y_i, i.e., the logic AND operation, can be implemented by n serially connected reconfigurable optical switches along single bus waveguides. This block yields logic "1" output only when all switches are in the "pass" state. Reconfiguration signals R_i are used to determine whether the corresponding operand contributes complement logic to y_i. In contrast to electrical digital computing, the OR operation for optical signals can be directly implemented using a combiner and detected by a photodetector. It is important to recognize that when product terms are represented by the same wavelength, the amalgamation may result in logical errors due to coherent interference. This issue can be avoided either by using distinct wavelengths for each branch [66], or equipping each branch with photodetectors individually [70]. Besides, given that photodetectors operate as

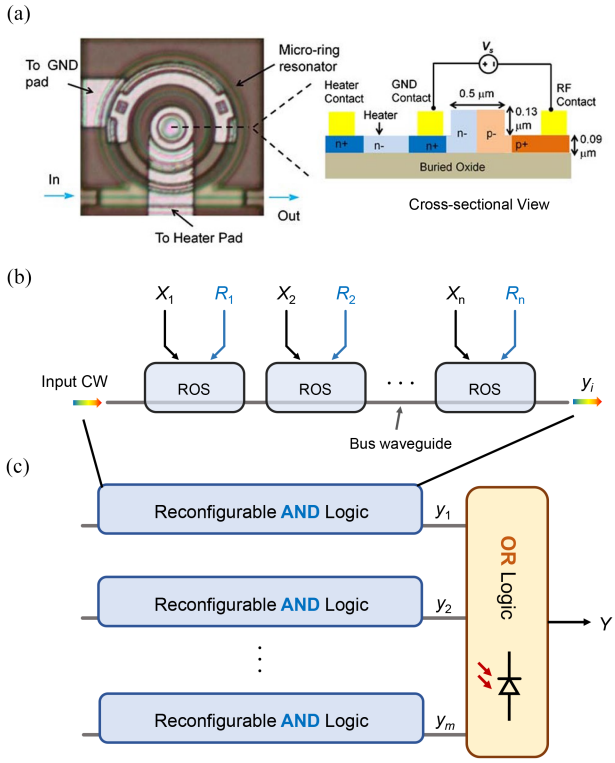

Fig. 2. Reconfigurable PICs for arbitrary combinational logic. (a) Optical micrograph and cross-sectional diagram of a reconfigurable MRR featuring two modulation mechanisms. The *p-i-n* junction is applied by RF signal for input encoding, while the microheater is connected to the low-speed DC signal for resonance mode reconfiguration [66]. (b) and (c) Schematic of a PIC architecture that enables the implementation of arbitrary combinational logic expressions based on reconfigurable optical switchs (ROSs).

current sources, a straightforward parallel configuration could achieve electrical OR logic without introducing additional delays. While this strategy offers a general solution for optical combinational logic, the limited scalability restricts its practical applicability in scenarios involving numerous operands. For an n-operands system, the complexity of this processing unit escalates as $O(n \cdot 2^n)$, indicating an exponential increase in the requisite number of switches with n. Furthermore, issues related to power consumption and accumulated losses also need to be considered. The dynamic power of the system P_{dynamic} can be expressed as:

$$P_{\text{dynamic}} = \frac{1}{4}\alpha C V^2 f \quad (2)$$

where, α is the activity factor and C is the total capacitance of the E-O modulators, which increases proportionally with the number of modulators. It is important to note that while (2) has a similar expression to that used for CMOS transistors, the supply voltage V does not necessarily scale with f. Therefore, to reduce P_{dynamic}, two straightforward strategies can be considered: 1) Decrease the number of modulators, for example, by employing multi-operand logic gates to squeeze logic functions into fewer devices [68]; 2) Utilize devices with low capacitance, such as microdisks with capacitance in 10's fF [24]. Additionally, when present, thermal tuning typically dominates the power consumption for modulation. This portion of power consumption can be

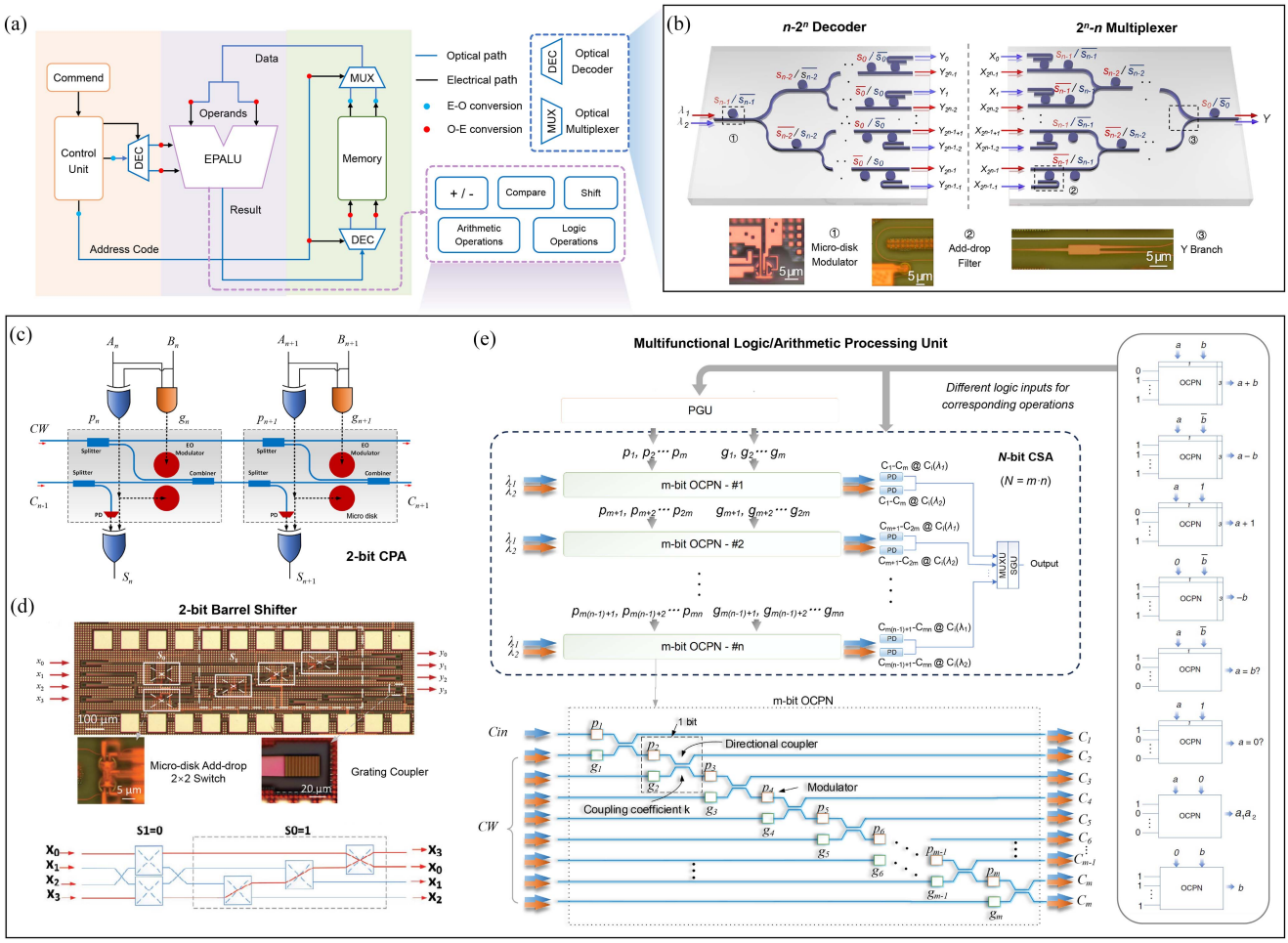

Fig. 3. EPALU architecture for high-performance digital computing. (a) Schematic of electronic–photonic microprocessor with its building blocks and data path. (b) Schematic of an n-2^n E-O decoder and 2^n-n O-E multiplexer [79], where s_i refers to the electrical input signal. (c) Schematic of a 2-bit CPA using E-O logic [61]. (d) Layout of a 2-bit barrel shifter using microdisk add-drop switch array, and the optical datapath with $S =$ '01' [77]. (e) The architecture of the WDM-based N-bit multifunctional processing unit consists of a (p, g) generation unit (PGU), n sets of m-bit optical carry propagation networks (OCPNs), and an array of photodetectors (PDs) along with a network of electronic multiplexer units (MUXU) and an electronic sum generation unit (SGU) [24]. With different input combinations, EPALU can perform various logic/arithmetic functions (right).

reduced or even eliminated through device-level optimization, such as post-fabrication trimming [71], or using energy-efficient tuning mechanisms (discussed in Section III-A). The total power consumption also includes the laser and detection parts, as well as the potential static power consumption. The laser power is primarily determined by the losses, typically at the mW level with dB/cm-level loss, which also correlates with the PIC scale.

In the realm of electrical digital design, electronic design automation (EDA) tools assist designers by automatically generating and optimizing circuits from logic expressions, as well as auto-placing and routing in physical design. With the advancement of PIC-based computing, logic synthesis methods specified for PIC design have been proposed, enabling large-scale design automation and optimization of compact PICs for digital computing [72], [73], [74].

C. Toward a Fully-Functional EPALU

Besides the aforementioned advantages of optical logic units, unique multiplexing techniques play a pivotal role in further improving computing capacity and performance. An example is the WDM-based electronics-photonic arithmetic logic unit (EPALU) (Fig. 3(a)) [24]. The arithmetic logic unit (ALU), which performs arithmetic and bitwise operations, is an essential component of modern computing systems. Within ALUs, the full adder plays a critical role, with its logic being expressed as follows:

$$C_n = p_n \cdot C_{n-1} + g_n \ , \quad S_n = C_{n-1} \oplus p_n$$
$$g_n = A_n \cdot B_n \ , \quad p_n = A_n \oplus B_n \tag{3}$$

where A, B, and C represent the two operands and carry, respectively; p and g denote propagation and generation, and the subscripts indicate the specific bits. Based on these expressions, the scalable electro-optic carry propagation adder (CPA) has been developed [61], with the schematic shown in Fig. 3(c). For a N-bit CPA, the carry output from one stage is the carry input to the next stage, thus the final result cannot be calculated until the carry has rippled through all stages. As an optimized architecture, the carry select adder (CSA) splits N-bit operands into

n m-bit CPA ($N = m \times n$). It speeds up addition by computing two possible outcomes for each m-bit CPA simultaneously–assuming carry-in values of "0" and "1"–and then selects the correct result based on the actual carry-in using multiplexers (MUXs) [75]. As a trade-off, CSAs require two sets of circuits with different carry inputs. However, in the optical domain, this can be efficiently implemented with a single optical path in the PIC, where different carry signals are encoded into two distinct wavelengths using WDM (Fig. 3(e)). Beyond arithmetic addition, Ying et al. developed that the multifunctional architecture could perform addition, subtraction, comparison, and bitwise operations operating at 20 GB/s with various input combinations [24]. Based on time-space multiplexing, Zhang et al. demonstrated a photonic-electronic digital multiplier capable of processing up to 32 × 4-bit binary inputs at 25 Mbit/s [76]. Additionally, the E-O shifter within the EPALU architecture has also been experimentally demonstrated (Fig. 3(d)) [77].

From the perspective of computer architecture, the ALU operates under the directives issued by electronic control units, with its inputs being fetched from memory. Upon completing a designated operation, the ALU's output is then stored back in memory to be accessed for subsequent computations (Fig. 3(a)). Although optical computing has significantly reduced the latency of arithmetic operations, data access and E-O/O-E conversions can create major bandwidth bottlenecks and serve as a significant source of energy consumption in the computing system. Therefore, the exploration of high-speed interconnects between different modules within an electronic–photonic microprocessor is important as well [78], [79]. In [79], a decoder and multiplexer designed for the EPALU architecture have been demonstrated, achieving data transportation and processing at a speed of 20 Gb/s (Fig. 3(b)). Another potential approach to reduce the time and energy costs of E-O/O-E conversion is through the implementation of various forms of optical memory, as discussed in [80], [81].

III. PIC-Based Analog Computing for AI: Fundamentals and Implementation

Undoubtedly, modern AI, functioning on digital computing systems, has achieved significant progress in diverse fields and has even exceeded human performance in specific tasks. With the advancements in integrated photonics, the PIC platform emerges as a compelling candidate for AI accelerators. Leveraging the optical logic gates and computing units detailed in Section II, some studies have illustrated the capability of PICs to perform tensor operations, including accumulation, dot products, and matrix-vector multiplications (MVMs), for diverse machine learning applications within the digital domain [82], [83], [84]. These studies have highlighted the inherent advantages of optical computing in terms of latency, speed, and power efficiency. However, the digital representation can encounter challenges stemming from hardware complexity overhead and speed reduction caused by the sampling and digitization into binary streams processed by logic units. These challenges are especially significant in the context of high-throughput or high-precision tensor operations for various machine-learning tasks.

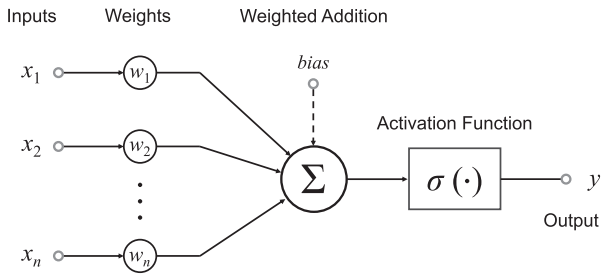

Fig. 4. Schematic of an artificial neuron with simple synaptic model.

On the other hand, the human brain, operating as an analog signal "processor", is estimated to perform at a rate of 10^{18} multiply-accumulate (MAC)/s with a power consumption estimated at just ∼10–20 W [85], [86], demonstrating remarkable efficiency compared to the substantial energy requirements of cutting-edge AI. This efficiency can potentially be attributed to the parallel processing capability and reduced precision requirements inherent in analog computing. While the mechanisms of brain function remain incompletely understood, there is growing interest in incorporating analog computing into machine learning. Before discussing the details of implementing optical analog computing for AI, it is helpful to provide a concise overview of artificial neural networks (ANNs) and the neuron model. The schematic of an artificial neuron with a basic synaptic model is illustrated in Fig. 4, where $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{w}$ represent the inputs from the pre-synaptic neuron, post-synaptic output, and weights of the connection, respectively. The activation functions $\sigma(\cdot)$, such as sigmoid, ReLU, and the leaky integrate-and-fire (LIF) function for SNNs, introduce nonlinearity into the model along with various engineering considerations [87], [88].

A. Programmable Modulation for Optical Analog Computing

In analog AI accelerators, both inputs **x** and weights **w** could correspond to a higher resolution, in contrast to the binary values in digital computing circuits. The hardware implementation of the above process using PICs requires the reconfigurable programming of network parameters, which relies on the modulation of optical components. While the weights in a trained model may remain fixed during the inference process, it is still necessary to calibrate the network through modulation due to the fabrication variations of PICs. A number of modulation mechanisms have been developed, among which tuning the effective refractive index n_{eff} of waveguides is a widely adopted approach in ONNs.

1) Thermal Tuning Mechanism: Based on thermo-optic effects, the transmission characteristics of devices can be modulated by changing the n_{eff} of waveguide through integrated filament microheaters. The heat generated by the ohmic microheater is proportional to the square of the bias voltage. Fig. 5(a) shows the schematic and transmission curve of a thermo-optic MZI fabricated by *Advanced Micro Foundry* [89]. Thermal tuning demonstrates the adaptability to various devices and substrate materials, with minimal constraints imposed by the fabrication process. Furthermore, compared with other mechanisms (especially for silicon), it can induce large changes in n_{eff} within a

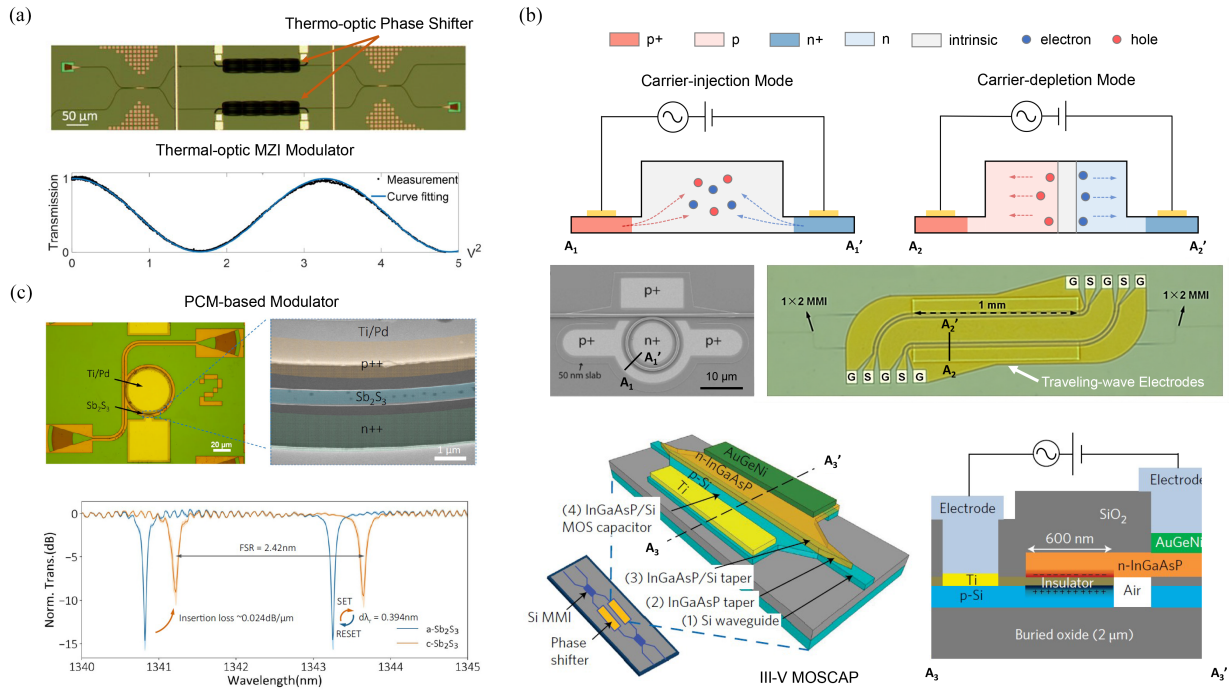

Fig. 5. Modulation techniques of photonic-electronic devices on PICs. (a) Optical micrograph and the normalized transmission curve of a thermo-optic MZI. Each arm of the MZI incorporates a microheater as a phase shifter [89]. While both arms can be used to modulate the output signal, typically, one is used to set the modulator output at the quadrature point for a high extinction ratio in monotonic modulation. (b) Schematics and micrographs of free-carrier effect-based modulators working in carrier-injection mode [110], carrier-depletion mode [95], and by the MOSCAP-driven tuning mechanism [98]. (c) Top, optical micrograph of an MRR modulator loaded with 10 μm long 20-nm-thick Sb_2S_3 and doped silicon PIN heater. Bottom, normalized transmission spectra of MRR when switching between two phases of Sb_2S_3 [109].

small footprint. However, thermo-optic devices encounter difficulties in achieving high-speed modulation, currently limited to a few hundred KHz [90], although sufficient for inference tasks with static weights. Additionally, thermal tuning is a volatile configuration process, requiring continuous external biasing and power supply (typically ∼mW level) to hold its functionality. Another issue arises from thermal crosstalk when heat dispersion cannot be adequately contained without physical constraints, such as trenches, which need appropriate consideration in both schematic and layout design.

2) Field-Effect Tuning Mechanism: Except for the thermo-optic effect, the n_{eff} can also be tuned by electric fields. In silicon photonics, a straightforward approach involves doping the silicon waveguide and applying an external electric field to manipulate the carrier concentration and tune n_{eff}. A variety of modulators operating in carrier-injection (forward-bias *p-i-n*), carrier-depletion (reverse-bias *p-n* junction), and carrier-accumulation (metal-oxide-semiconductor capacitor, MOSCAP) mode have been widely demonstrated (Fig. 5(b)) [91], [92], [93], [94], [95], [96]. These CMOS-compatible mechanisms allow gigahertz-level tuning speeds, making them suitable for high-speed encoding in ONNs. Particularly, in depletion mode, the bandwidth is determined by the majority carriers' dynamics, which are not limited by the slower processes of carrier generation and recombination [94]. Moreover, while the depletion-mode modulator remains a volatile device, it exhibits low static power consumption attributed to the reverse-biased junction. Compared to thermo-optic devices, the free-carrier effect-based modulators typically require longer modulation lengths and larger footprints, primarily due to their lower tunning efficiencies (as evaluated by voltage-length product $V_\pi L$) or the necessity for traveling-wave electrodes for high-speed modulation. Expanding beyond the conventional silicon platform, an alternative approach is the utilization of materials with significant electro-optic effects–such as the free-carrier plasma dispersions, the Pockels effect, the Kerr effect, and the Quantum Confined Stark Effect (QCSE)–for the core or cladding of waveguides. The modulation efficiency of these devices is intrinsically linked to the material properties. Prominent examples include III–V semiconductors [96], [97], [98], lithium niobate [99], and some polymers [100], [101]. A comprehensive discussion of all these mechanisms exceeds the scope of this review. Interested readers are referred to relevant reviews and reports for further information [102], [103], [104].

Most modulators discussed above have nonlinear transmission curves, primarily due to the E-O modulation mechanisms involved. Although the Pockels effect provides a linear change in Δn in response to electric field intensity, the inherent transfer function of the modulator itself can still exhibit a nonlinearity (such as the sinusoidal response of MZI). To reduce undesired nonlinearity in computing architectures, several solutions have been proposed: 1) Introduce a nonlinear mapping during D/A conversion, although this may necessitate additional data processing; 2) Utilize the nearly linear segment of the transmission curve, at the tradeoff of a reduced modulation dynamic range; 3) Design PICs that compensate for nonlinearity by novel devices or modulation mechanisms [100], [105], [106], [107], [108]. Additionally, the impact of nonlinearity on system noise also needs

TABLE I
SUMMARY OF MODULATION TECHNIQUES WITH SILICON WAVEGUIDE

Mechanism	$V_\pi L$ (V·cm)	IL (dB)	f_{3dB}	Data Rate	Energy Efficiency	Active Length	Modulator	Ref
Thermo-optics	∼0.023	<1	∼4 KHz	—	∼1.36 mW*	∼170 μm	MZM	[89]
Carrier Injection	0.036	12	16 GHz	10 Gb/s	51 mW* / 5 pJ/bit	200 μm	MZM	[92]
Carrier Depletion	1.62–2.05	3.9	30 GHz	44 Gb/s	2.84 pJ/bit	1000 μm	MZM	[95]
Si-MOSCAP	∼0.8	9	50 GHz	>100 Gb/s	—	R = 15 μm	MRR	[96]
III-V-MOSCAP	0.09	1	2.2 GHz	32 Gb/s	—	250 μm	MZM	[98]
Graphene-MOSCAP	0.28	∼7	5 GHz	10 Gb/s	1 pJ/bit	300 μm	MZM	[111]
PCM	L_π=30.7μm	<1	∼10 KHz	—	11.6 mJ for SET 197 nJ for RESET	R = 30 μm	MZM/MRR	[109]

* Static power consumption
∼ Measured from figures or calculated based on other parameters in the articles
- Not mentioned or unavailable in the article

consideration. Within modulators exhibiting strong nonlinearity and steep slopes, such as high-Q MRRs, slight noise or offsets in the drive signal can potentially result in significant deviations.

3) Non-Volatile Modulation: Each aforementioned mechanism is a volatile process and requires continuous power supply even for infrequent or static programming tasks. Non-volatile phase change materials (PCMs) offer a unique opportunity to avoid these scenarios. PCMs have two switchable phases, i.e. amorphous and crystalline states, with drastically different n_{eff}, and can achieve reversible phase transitions within various temperature ranges. For the PIC platform, heating can be achieved by an integrated pulse microheater or alternatively, through light signals themselves, enabling 'all-optical' modulation. Chen et al. developed several non-volatile silicon photonics modulators with Sb_2S_3 PCM cladding and experimentally demonstrated 5-bit multilevel programming with a high cyclability >1600 switching events (Fig. 5(c)) [109]. Besides, all-optical SNNs using chalcogenide PCM GeSbTe (GST) also have been reported [32], and more details will be discussed in Section. III-C2. A brief summary of modulation techniques with Si waveguide is shown in Table I.

B. Implementations of Photonic Tensor Cores

As illustrated in Fig. 4, the interconnections in ANNs can be conceptualized as weights akin to synaptic coupling coefficients in biological systems. This analogy extends to representing these connections through tensor operations, thereby abstracting the complex interactions in a computationally manageable form. Building upon the aforementioned modulation mechanisms, various active devices and encoding mechanisms have been utilized effectively in photonic neurons. This section categorizes ONNs from multiple perspectives, which aims to provide a comprehensive analysis and comparison of the implementations of photonic tensor cores (PTCs) from diverse optical components to the architectural design.

1) Coherent vs Incoherent ONN: From the viewpoint of signal properties, ONNs can be classified into coherent and incoherent systems. Within coherent ONNs, both weights and inputs can be encoded in the complex plane, allowing multiplication through lossless interference. For instance, a pair of beam splitters and phase shifters in the form of MZIs are widely adopted for conducting linear operations in coherent ONNs due to the programmable amplitude and phase response [112]. The schematic of a 2×2 MZI is shown in Fig. 6(a). Assuming that each beam splitter is an ideal 50:50 directional coupler (i.e., with a coupling coefficient $\kappa = 1/\sqrt{2}$), the transition matrix of the MZI can be represented as:

$$T_{\text{MZI}} = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & j\\ j & 1\end{bmatrix}\cdot\begin{bmatrix}e^{-j\phi_1} & 0\\ 0 & e^{-j\phi_2}\end{bmatrix}\cdot\frac{1}{\sqrt{2}}\begin{bmatrix}1 & j\\ j & 1\end{bmatrix} \tag{4}$$

Here, ϕ_1 and ϕ_2 represent the phase changes as signals traverse two arms. In some cases, ϕ_1 and ϕ_2 are denoted as $\Delta\phi = \phi_1 - \phi_2$ and 0, respectively, given that the response is predominantly governed by the phase difference between its two arms. If only one input port receives a signal $\boldsymbol{E}_{\text{in}}$, the field and power intensities at the output port can be expressed as follows:

$$y_1 = -j\cdot\exp\left[-j\left(\phi_2+\frac{\Delta\phi}{2}\right)\right]\cdot\sin\left(\frac{\Delta\phi}{2}\right)\cdot E_{\text{in}}\ ;$$

$$y_2 = \quad j\cdot\exp\left[-j\left(\phi_2+\frac{\Delta\phi}{2}\right)\right]\cdot\cos\left(\frac{\Delta\phi}{2}\right)\cdot E_{\text{in}} \tag{5}$$

$$|y_{1,2}|^2 = |E_{\text{in}}|^2\cdot\frac{1\mp\cos\Delta\phi}{2} \tag{6}$$

This implies that the phase and amplitude of output can be modulated by varying $\Delta\phi$ from 0 to π. The smooth transmission curve, stemming from the cosine term, enables high-resolution analog computing and enhances noise resistance. In a notable case, the two arms are encoded by opposite phase shifts, represented as $\phi \pm \Delta\phi/2$ for ϕ_1 and ϕ_2, respectively. When incorporated into (4), the Mach-Zehnder modulator (MZM) working in "push-pull" mode allows for the independent modulation of field intensity while maintaining a constant phase.

An MZI cascaded with a phase shifter can implement a 2×2 unitary transformation:

$$U(2) = \overbrace{\begin{bmatrix}e^{-j\theta} & 0\\ 0 & 1\end{bmatrix}}^{\text{Phase Shifter}}\cdot\overbrace{\frac{1}{2}\begin{bmatrix}e^{-j\Delta\phi}-1 & j(e^{-j\Delta\phi}+1)\\ j(e^{-j\Delta\phi}+1) & -(e^{-j\Delta\phi}-1)\end{bmatrix}}^{\text{MZI}}$$

$$= \frac{1}{2}\begin{bmatrix}e^{-j\theta}(e^{-j\Delta\phi}-1) & je^{-j\theta}(e^{-j\Delta\phi}+1)\\ j(e^{-j\Delta\phi}+1) & -(e^{-j\Delta\phi}-1)\end{bmatrix} \tag{7}$$

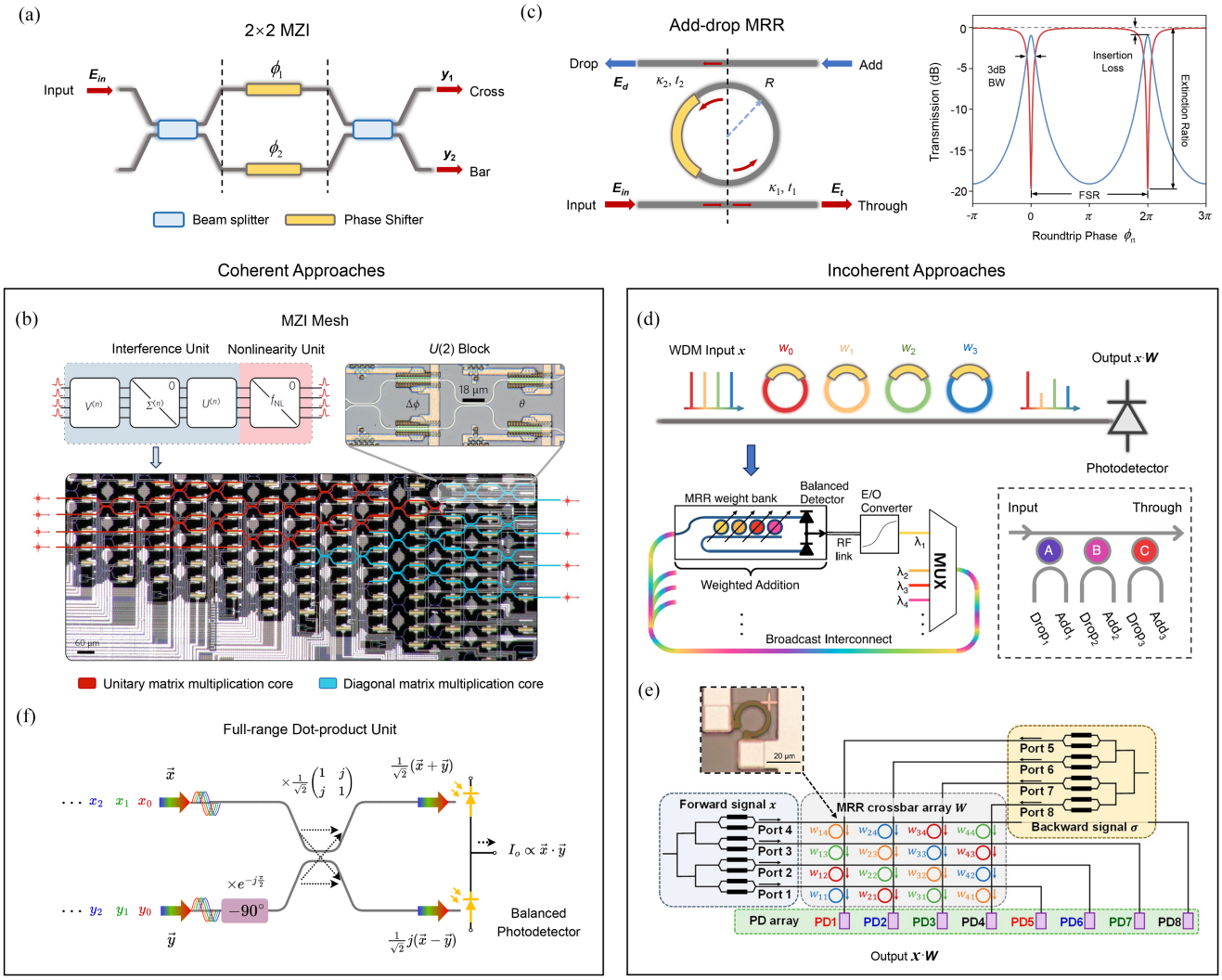

Fig. 6. Various implementations of photonic-electronic tensor cores. (a) Schematic of a 2×2 MZI where the beam splitters can be implemented by multi-mode interferometer (MMIs) or directional couplers. An alternative configuration of MZI features a single input and output, with the splitter designed as a Y branch. (b) An MZI-based coherent photonics tensor core for ONNs [28]. The weight matrix can be represented by SVD in the form $\boldsymbol{U\Sigma V}^{\dagger}$. Due to chip size and complexity considerations, this work only implements $\boldsymbol{U}$ and $\boldsymbol{\Sigma}$ on a single pass through the circuit. (c) Schematic and power spectral responses of an add-drop MRR with power coupling coefficients ${\kappa_1}^2 = {\kappa_2}^2 = 0.2$ (assuming the coupling junctions are lossless, i.e., $\kappa^2 + t^2 = 1$) and 5% round-trip power loss. (d) The weight bank architecture utilizes MRRs as tunable filters for parallel modulation of WDM signals [122]. The series-connected MRRs share a single bus waveguide as a through port, while the other waveguide can be shared or independent, enabling the selective addition or removal of wavelengths. (e) Schematic of 4×4 MRR crossbar array demonstrated in Ref [123]. Input signals are first modulated via MZMs, and then distributed and weighted through the crossbar array, followed by a photodetector array that sums the weighted WDM elements. (f) The Schematic of an optical tensor core supports dynamic full-range general matrix multiplication based on WDM and coherent interference [126].

Unitary matrices of rank N can be decomposed into sets of $U(2)$ blocks, which can be implemented using cascaded 2×2 MZIs to form a mesh structure. Despite the unitary nature of the transition matrix in an MZI-mesh, an arbitrary weight matrix for an ONN can be implemented by singular vector decomposition (SVD). The SVD approach was thoroughly discussed by Miller et al. in 2003 and first experimentally implemented in ONNs by Shen et al. as shown in Fig. 6(b) [28], [113]. Specifically, an arbitrary real-valued matrix $\boldsymbol{M}$ can be decomposed into the form $\boldsymbol{U\Sigma V}^{\dagger}$, where $\boldsymbol{\Sigma}$ is a diagonal matrix and the remaining two are unitary matrices, which can be implemented with a set of tunable attenuators and an MZI mesh, respectively.

Another characteristic of the devices based on phase modulation is their broad spectral bandwidth. During the modulation process, the phase change ψ can be expressed as follows:

$$\psi = \frac{2\pi L_{\mathrm{PS}} \Delta n_{\mathrm{eff}}}{\lambda} \tag{8}$$

where λ and L_{PS} represent operating wavelength and effective modulation length. Within the C-band and O-band commonly used in silicon photonics, the transmission characteristic demonstrates a low sensitivity to λ over a range of several tens of nanometers. Particularly for thermo-optic devices, the thermal

modulation coefficient dn_{eff}/dT shows a slight increase corresponding to the increase in wavelength, which compensates for the decrease in wave number $2\pi/\lambda$ [114]. Therefore, an MZM followed by a phase shifter can achieve parallel modulation of multiple signals with different wavelengths based on WDM. However, it should be noted that waveguides induce different phase changes at various λ, which must be considered when using WDM in coherent ONNs.

Thanks to the phase encoding mechanism, coherent ONNs can perform multiplication with full-range input operands. Besides the SVD approach, a more intuitive example is that the sign and absolute value of the scalar can be encoded by the phase shifter and MZM separately. Following this methodology, [115] presents a photonic neuron architecture enabling full-range dot product operations. However, this architecture also faces limitations in terms of complexity and scalability. Specifically, implementing SVD required a matrix pre-processing step for phase mapping, which consumes extra time and power. In addition, the number of MZIs escalates quadratically with matrix size, leading to scalability issues due to the larger footprint and accumulated losses. Moreover, the phase control of the coherence network presents a challenge as well. Within the MZI mesh, each $U(2)$ operation demands a minimum of two active phase modulators. Without a reference signal, phase calibration becomes more complex than amplitude calibration (which can be straightforwardly measured using photodetectors), especially considering the fabrication deviation across individual devices and waveguides. However, some studies demonstrate the feasibility of addressing the phase calibration issue through on-chip training [116], [117].

Incoherent PICs offer an alternative approach to execute tensor operations. In the absence of coherence requirements, an incoherent ONN enables broader flexibility in employing multiplexing techniques. Because of its wavelength-dependent transmission characteristics, MRRs are widely employed for encoding weights and inputs in WDM-based incoherent ONNs. From the perspective of physical design, the small footprint of MRR (achieving a radius $<10\ \mu$m) affords a compact layout and an enhanced scalability. As shown in Fig. 6(c), the add-drop MRR is a four-port optical device that consists of a microring evanescently coupled to two bus waveguides. The transfer characteristics at the through port and drop port can be expressed as (9) and (10).

$$T_t = \left|\frac{E_t}{E_{in}}\right|^2 = \frac{t_1^2 + t_2^2\alpha_{rt}^2 - 2t_1t_2\alpha_{rt}\cos\phi_{rt}}{1 + t_1^2t_2^2\alpha_{rt}^2 - 2t_1t_2\alpha_{rt}\cos\phi_{rt}} \tag{9}$$

$$T_d = \left|\frac{E_d}{E_{in}}\right|^2 = \frac{\kappa_1^2\kappa_2^2\alpha_{rt}}{1 + t_1^2t_2^2\alpha_{rt}^2 - 2t_1t_2\alpha_{rt}\cos\phi_{rt}} \tag{10}$$

Here, κ and t represent the field coupling factor and transmission factor, respectively. $\alpha_{rt} = exp(-\alpha \cdot 2\pi R)$ and $\phi_{rt} = \beta \cdot 2\pi R$ are the round-trip field attenuation factor and phase, with α and β being the real and imaginary parts of the complex transmission coefficients [118]. The resonance condition is $\phi_{rt} = 2m\pi$, i.e., $\lambda_{\text{res}} = 2\pi n_{\text{eff}}R/m$, where m is a positive integer. A potential issue in achieving full-range modulation (i.e., from 0 to 1) arises, as it can only be attained with symmetric, lossless coupling ($\alpha_{rt} = 1,\ t_1 = t_2$), which is unrealistic for real devices. An all-pass MRR can be regarded as a special case of an add-drop MRR, distinguished by the absence of a drop bus waveguide with $t_2 = 1$ and $\kappa_2 = 0$ for (9). While the phase at the through/drop port can also be derived from (9) and (10), the phase response and tuning range are highly sensitive to coupling conditions and round-trip loss. Hence, MRRs are primarily employed for amplitude modulation in incoherent systems. Another challenge comes from the experimental perspective: MRRs, particularly those with high quality factors, are sensitive to environmental factors such as temperature variations and vibrations. For a more in-depth theoretical analysis of MRRs, readers may refer to relevant literature and books [118], [119].

In tensor operations, MRRs can be modulated by tuning coupling coefficients or the round-trip phase [120]. Unlike MZIs, MRRs demonstrate significant wavelength selectivity. An MRR with a moderate quality factor can achieve a resonance peak with a 3 dB bandwidth of a few hundred picometers. The narrow bandwidth makes MRRs particularly effective as tunable filters in WDM-based PICs, where they are often arranged in series for the selective modulation of signals with different wavelengths. This application is exemplified by the concept of "weight banks" proposed by Tait et al. [121], [122], as shown in Fig. 6(d). In this architecture, photodetectors can spontaneously perform the sum of operands encoded on different wavelengths. Another incoherent architecture is the crossbar array, which implements MVMs using its programmable transfer matrix in the form of a tunable switch array [123], [124], [125]. Ohno, et al. demonstrated a 4×4 add-drop MRR crossbar array for MVMs (Fig. 6(e)) [123]. Specifically, each wavelength coming from the row bus waveguide can be weighted and subsequently directed into a specific column by the MRR array. The weighted elements can be collected and aggregated by a photodetector at the end of the column, thus carrying out MVM.

A notable constraint of incoherent ONNs is the challenge of directly implementing negative operands due to the amplitude modulation mechanism. An intuitive solution is dividing the matrix into positive and negative parts, subsequently mapping into the network up to 4 times to execute $(X_+ - X_-)(Y_+ - Y_-)$, and thereafter deducting the outcomes within the electrical domain. Alternatively, the positive and negative components can be processed simultaneously through two identical incoherent hardware networks. Both strategies, however, necessitate either extended processing time or increased hardware resources. Another strategy leverages the complementary output from the through and drop ports of the add-drop, enabling the full-range output as demonstrated in the work of Tait et al. [121], [122]. For the last two approaches, the post-processing subtraction can be physically implemented using balanced photodetectors, as illustrated in Fig 6(d).

2) Static and Dynamic Weight Encoding: The above-mentioned architectures topically map one operand of MVMs (typically a weight matrix *W*) onto hardware, executing multiplication through the transmission matrix of the PIC. Therefore, the operational speed of an ONN is, in theory, only determined by the modulation rate of input signals x. However, practical implementations face several limitations. Primarily, deploying

MVMs with large dimensions is challenging due to the cost, complexity, loss, and other scalability issues of PICs. While multiplexing techniques can enhance the parallelism of ONNs, time-domain hardware reuse is necessary to fulfill the computational requirements of complex machine learning tasks. This indicates that the "weight-static" architecture constitutes a bottleneck for large networks because of the huge gap between ultra-fast optical computing and slow mapping/reprogramming. To fully unleash the potential of optical analog computing, the critical role of dynamic encoding for specific tasks should be recognized. For dynamically-operated ONNs, two essential requirements must be satisfied. First, the parameters in the network need to be high-speed reprogrammable for efficient hardware reuse. This requires that the modulators operate at gigahertz-level rates, as exemplified by those based on field-effect tuning mechanisms. Secondly, dynamic operation mandates that parameter mapping and output reading be conducted "directly" without additional signal preprocessing or data correction. For instance, to map a 12×12 matrix to an MZI-mesh framework requires ∼1.5 ms to perform SVD and phase decomposition on a CPU [126], which precludes dynamic operation due to the delay.

The importance of dynamic encoding is further highlighted by its compatibility with *Transformer* [9]. Transformer and the unique attention mechanisms have attracted significant research interest in recent years due to their exceptional performance in natural language processing (NLP), machine vision (MV), and large-scale language models (LLM). Unlike weight-static architectures, Transformer employs the multi-head self-attention (MHA) mechanism within its encoder and decoder blocks, necessitating matrix multiplication involving two dynamic, full-range operands. Zhu et al. presented a dynamically-operated PTC for the first photonic Transformer accelerator, leveraging coherent light interference and WDM [126]. As shown in Fig. 6(f), the elements within the input vectors x and y are encoded at distinct wavelengths, and then add a $-90°$phase shift to one vector. Through a directional coupler, two orthogonal signals can be recombined into the form $(x \pm y)$, facilitating the computation of the dot product via a balanced photodetector.

3) Hardware-Efficient PTCs: For the aforementioned structures, the total number of optical components required to implement an $m \times n$ general matrix is similar, e.g., $m(m-1)/2 + n(n-1)/2 + \max(m,n)$ MZIs for an SVD-based MZI mesh, and $m \times n$ MRRs in microring-based ONNs. In addition to the high hardware costs and inherent control complexities, the required electrical components make up a large proportion of energy consumption, particularly for high-speed and high-resolution modulation. To address these challenges, researchers are exploring strategies to enhance hardware efficiency across multiple levels–from the device to the architecture–to improve the scalability of ONNs.

Firstly, optical devices or structures featuring compact topology have been proposed to reduce PIC footprint and improve hardware efficiency. Zhu et al. developed an ONN architecture consisting of two diffractive cells and MZIs exhibiting linear scaling in relation to input dimensions [127], as shown in Fig. 7(a). The star-shaped diffractive cell shows the capability of performing on-chip parallel Fourier transform and convolution operations. In a similar vein, an ONN leveraging the combination of WDM and MMIs has been developed and demonstrates an accuracy of 92.17% on the MNIST dataset. Beyond waveguide-based PIC, Wang et al. proposed a metasurface-based processing unit that provides ultra-high throughput for MVM [128] (Fig. 7(b)). Within the subwavelength structure, each pair of slots acts as a weight element and connects to the following layers via in-plane diffraction and interference. For a specific neural network, the width and length of the slot need to be designed to map the corresponding weights.

The passive ONNs not only yield smaller footprints and enhanced hardware efficiency but also lower power consumption. However, a significant challenge associated with the passive task-specific PTCs is their fixed configuration, which can lead to degraded effectiveness when applied to varying tasks. To address this challenge, active multi-operand devices present another opportunity to break the fundamental limitation to achieve high-density tensor operation by squeezing MAC operation into a single device. Specifically, the execution of the length-k dot product between input vector $\boldsymbol{x}_{\text{in}}$ and weight vector $\boldsymbol{w}$ by k-operand devices can be represented as follows:

$$x_{\text{out}} = f(\boldsymbol{w} \cdot \boldsymbol{x}_{\text{in}}) = f\left(\sum_{i=1}^{k} g(w_i, x_i)\right) \tag{11}$$

where the function $f(\cdot)$ represents the E-O transmission function, while $g(w,x)$ is related to the encoding mechanism. Here, $\boldsymbol{w}$ can be encoded using programmable resistances (such as memristors or PCMs), tunable amplifiers/attenuators, or the effective modulation length in cases, while $\boldsymbol{x}_{\text{in}}$ can be encoded by electrical signals. The type of multi-operand devices can either be the modulators mentioned above or based on other passive optical components such as MMI [129]. Feng et al. first demonstrated this method experimentally on a 4-operand MZI, achieving a measured accuracy of 85.9% in SVHN recognition tasks with 4-bit control precision (Fig. 7(c)). Gu et al. proposed a compact ONN architecture based on multi-operand MRRs (MOMRR) [130]. The MOMRR-based ONN supports weight encoding through two sets of rails, with full-range results carried out by balanced photodetectors [131]. Additionally, the architecture can be combined with the structured pruning strategy to further improve network scalability. Theoretically, multi-operand optical synapses could execute vector operations with nearly the same footprint as the single-operand counterparts. However, the intrinsic nonlinearity and possible crosstalk among operands could present challenges for training and calibration.

Beyond improvements at the device level, another promising approach enhances hardware efficiency at the circuit level by software-hardware co-design approaches. Subspace neural networks, for instance, sacrifice a portion of matrix representability in exchange for fewer parameters instead of implementing universal linear operations or general matrix multiplication (GEMM). The effectiveness of this strategy can extend to ONNs by trading the universality of weight representation for higher hardware efficiency. An example is the butterfly-style PTC [89] (Fig. 7(d)). By parameter pruning, this architecture

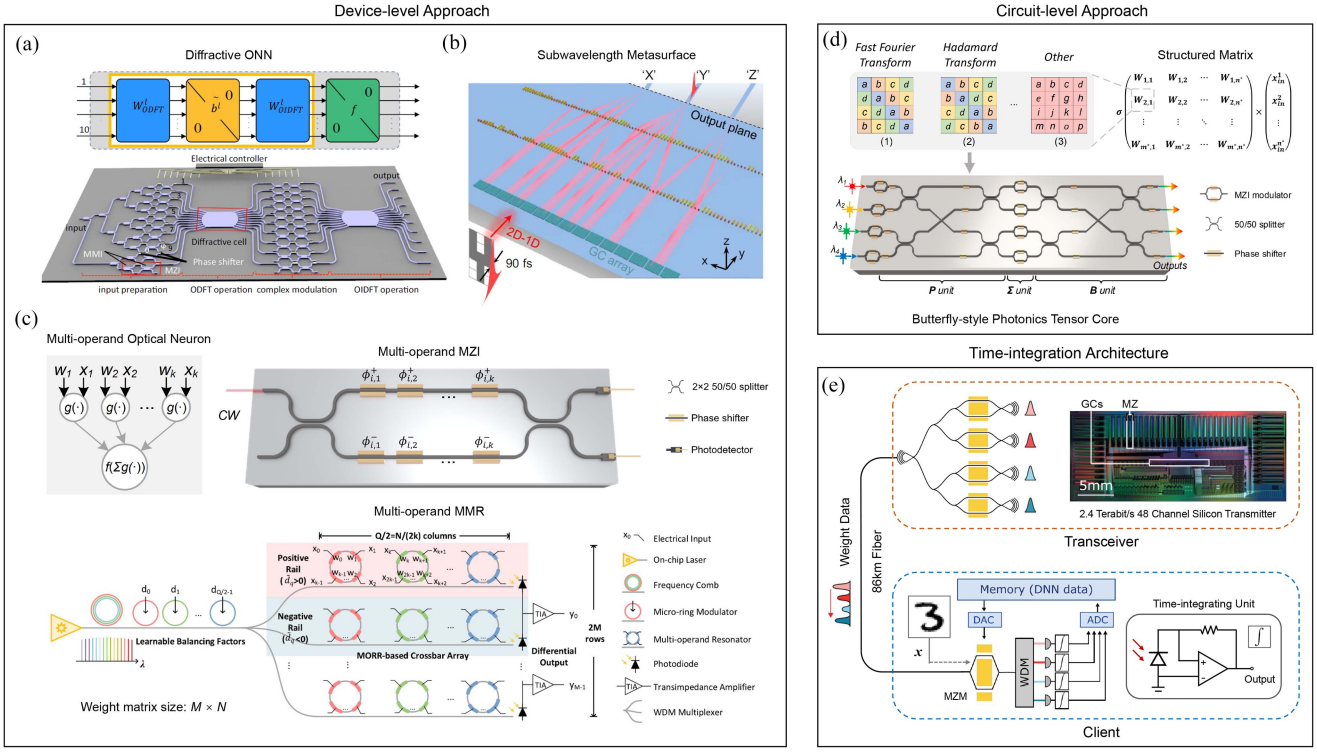

Fig. 7. Implementation of hardware-efficient PTCs. (a) and (b) Schematics of space-efficient ONNs using compact diffractive cell [127] and metasurface structure [128], respectively. O(I)DFT: Optical (inverse) discrete Fourier transform (c) Diagram of a multi-operand optical neuron, along with schematics of muti-operand MZI [130] and muti-operand MRR [131]. (d) The general architecture of the subspace ONNs (top), and schematic of 4×4 butterfly-style tensor core (bottom) [89]. Here, B and P are both unitary matrices and $\mathbf{\Sigma}$ is a diagonal matrix. Here, B and P represent butterfly-style transform unit and projection unit, respectively (e) Architecture diagram of a delocalized time-integration optical computing system consists of smart transceivers as cloud infrastructure and the edge client devices [132]. The parameters of the neural network model are encoded using WDM on the cloud side and transmitted to client nodes via a long-haul optical fiber. The MVMs are carried out in time-integrating of photocurrent generated by the photodetector in the client device.

could be implemented using significantly fewer MZIs with a scale factor of $O(n\log_2^n)$ rather than the $O(n^2)$ required by an MZI mesh for GEMM. This study experimentally showcases a measured accuracy of over 94% on the MNIST hand-written digits classification task, with up to 7× fewer active optical components, a 3.3× smaller footprint, and a 5.5× lower latency compared to conventional MZI mesh. A similar strategy, utilizing an MRR-based crossbar array to implement block-circulant matrices, has been demonstrated in Ref [125]. For subspace networks, the trade-off between hardware efficiency and matrix representability is an important consideration, and more details will be discussed in Section V-A.

Beyond the one-shot broadcast architecture, an alternative method leverages time-domain integration with fewer optical components to perform MACs [132], [133]. Following this approach, Sludds et al. developed a delocalized optical computing system that showcases edge computing capabilities over a span of 86 kilometers [132]. As shown in Fig. 7(e), the modulated WDM signals are separated by a passive demultiplexer and then fed to a set of time-integrating receivers. On the client side, only one MZM, ADC, and DAC are used, which allows an ultra-low power assumption of femtojoules per MAC operation. The photocurrent $I(t)$ generated by the photodetector produces a voltage across the integrating capacitor C by accumulating charge, thereby achieving the summation operation, which can be expressed as (12).

$$V_{\text{out}} = \int \frac{I(t)}{C}\, dt \propto \sum w_i x_i \tag{12}$$

This strategy sacrifices the speed benefit of optical computing to achieve a smaller chip footprint and reduced hardware complexity, which enables the use of milliwatt-class edge devices.

C. Implementations of Nonlinearities

Activation functions introduce the necessary nonlinearity into the network. For PIC-based photonic neurons, the implementations of nonlinearities fall into two major categories: the optical-electrical-optical (O-E-O) and the all-optical approach.

1) O-E-O Nonlinearities: The activation function within O-E-O neurons can be realized most directly by routing the weighted sum through A/D conversion into digital processing units, such as CPUs, GPUs, or FPGAs. This process performs nonlinear operations digitally, and then converts the digital output back into analog signals that are subsequently fed into photonic neurons. The primary advantage of this approach is its extensive transfer ability of existing ANN architectures to the PIC platform, which also facilitates the implementation of arbitrary activation functions. However, the compulsory A/D and D/A conversion, along with digital processing, impose latency

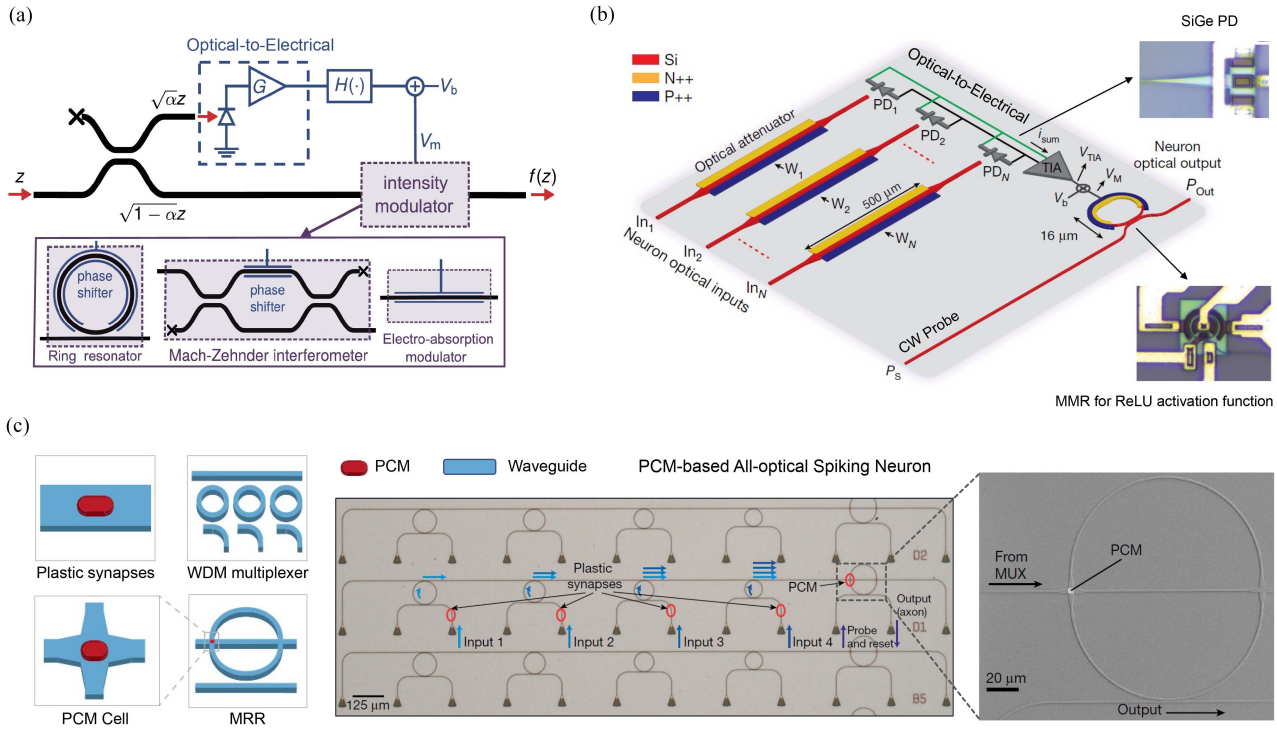

Fig. 8. On-chip implementations of programmable nonlinear activation functions. (a) and (b) Schematics of O-E-O nonlinear units for realization of reconfigurable activation functions [136], [137], [138]. (c) A PCM-based all-optical spiking neuron [32]. The input spikes are weighted via PCM cells, and the PCM cell on the MRR could switch the resonate condition when the accumulated power of the postsynaptic spikes exceeds a predefined threshold, thereby controlling the generation of output spikes.

and power consumption, thereby undermining the efficiency advantage inherent in ONNs. As mentioned in Section II, the performance and speed of an optical computing system are constrained by its weakest link, which in this case is the E-O interface and the digital processing. Alternatively, nonlinearities can also occur in the analog domain by leveraging the inherent nonlinear responses of specific electronic components or analog circuits [134], [135].

Beyond implementing nonlinearity purely in the electronic domain, nonlinearity can also be introduced within the E-O/O-E conversion processes, such as the nonlinear transmission properties of E-O modulators during encoding. For instance, the nonlinear part of the cosine term in the MZM transfer function exhibits similarities to the sigmoid function. This methodology has been validated on the last two layers of an ONN for the MNIST classification task in [115], and a similar approach involving the built-in nonlinearity of MRRs has been proposed by Gu, et al. [131]. The main advantage of introducing nonlinearity in encoding is it does not need extra O-E conversions. However, since the nonlinearity is entirely contingent on the transmission characteristics of E-O modulators, this dependency may pose challenges during the training process. Inappropriate activation functions, particularly in deep networks, can result in gradient vanishing or explosion and low training efficiency. To achieve reconfigurable activation functions, Williamson et al. introduced a nonlinear unit that converts a small portion of the optical output into an electrical signal to modulate the original optical signal (Fig. 8(a)) [136], [137]. This setup offers two approaches for the electrical part. The first approach converts the tapped signal into an electrical signal directly, allowing for moderate adjustments in nonlinearity through varying electrical biases. The second strategy utilizes reconfigurable lookup tables controlled by a microcontroller, which enables the generation of arbitrary nonlinearities and aligns more closely with an all-electrical approach. Likewise, MRRs can be employed to realize this methodology. In [138] and [139], activation functions are implemented by modulating a CW probe with the output from photodetectors (Fig. 8(b)). While this architecture avoids the loss accumulation issue, it requires an additional laser source as the probe. In addition to introducing nonlinearity from E-O modulation, O-E conversion processes, exemplified by the response of photodetectors, can also serve as the source of nonlinearity [140]. Nevertheless, these approaches are also constrained by limited reconfigurability.

2) All-Optical Nonlinearities: Without electrical components, all-optical approaches implement activation functions via the nonlinear response of materials or devices to optical signals. Several ONNs have been demonstrated using semiconductor optical amplifiers (SOAs) [141], saturable absorbers [142], [143], and techniques exhibiting excitable behavior [144]. All-optical nonlinearities can also be achieved by PCMs, such as the all-optical neuron developed by Feldmann et al. [32]. In this work, the PCM, functioning as a waveguide cladding, is used to modulate the pre-synaptic input and govern the resonance state of MRR, thereby controlling the generation of output spikes (Fig. 8(c)). Beyond the issues of activation function applicability

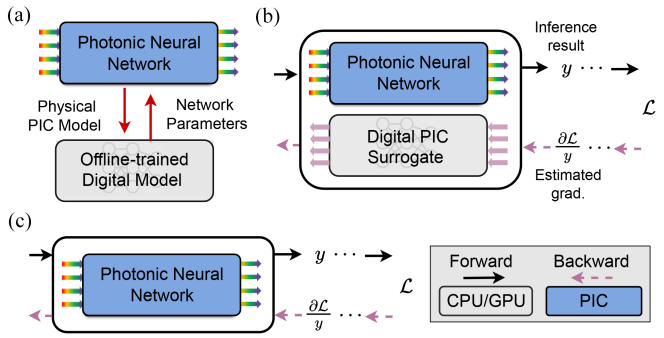

Fig. 9. Optical neural network training methods. (a) Offline training on a digital platform, such as a GPU, mimics the behavior of analog hardware by building a physical model of PICs. (b) Physical neural network training, which utilizes PICs for forward pass and incorporates a differentiable digital model for backpropagation. (c) *In-situ* optical neural network training is performed entirely on photonic hardware.

previously discussed, the implementation of all-optical neurons faces several other challenges. Firstly, even though extra electrical devices are not required, the substantial power required to excite nonlinearities does not offer any power consumption advantages compared to the O-E-O approaches. Additionally, the fabrication and integration of optical components for all-optical nonlinearity, such as PCM, excitable lasers, and optical amplifiers, pose challenges as well. Lastly, since both the control signal and post-synaptic output are optical signals, a potential issue exists in distinguishing the response from the control signals or bias.

D. Optical Neural Network Training

As an analog computing platform, PICs for AI applications are inherently susceptible to various non-ideal conditions, such as environmental changes, process variation, and limited control precision. These factors can decrease inference accuracy and potentially degrade the signal-to-noise ratio (SNR) of computations. To mitigate the decrease in accuracy, it is crucial to train ONNs with careful consideration of non-idealities that may occur during inference, making it a "circuit-aware" approach will enhance its noise resilience. The current ONN training can be classified into two categories: hardware-aware *ex-situ* training, which involves training with the help of digital computers, and on-chip *in-situ* training.

1) Hardware-Aware Ex-Situ Training: Hardware-aware *ex-situ* training offloads the training process to digital computers and utilizes various hardware-aware training techniques to capture PIC behavior as precisely as possible during training. One commonly used technique is noise-aware training [145], [146], as shown in Fig. 9(a). This approach involves modeling the behavior of PICs while considering various non-ideal effects. Subsequently, the PIC model is injected into the training to reduce the gap between training and real inference. Gaussian noise is commonly used to model various noise sources in photonic systems. For instance, shot noise and thermal noise are modeled using a Gaussian distribution by measuring on-chip photonic multiplication results and fitting their distribution [132]. In addition to injecting non-idealities into the training, some work explicitly models the transfer matrix of photonic neurons and embeds them in the forward computation pass during training [130], [131]. This injection is crucial because it introduces a unique nonlinearity term that ONNs need to be aware of in order to accurately capture the behavior of photonic neurons. Moreover, in [126], the noisy transfer matrix is derived and explicitly injected in the training under some noise assumptions. However, this approach encounters two main challenges. Firstly, accurately representing all on-chip non-idealities poses a significant difficulty, and environmental fluctuations can further affect inference accuracy. Second, the computational overhead required to model the physical system accurately can be exceedingly high or even prohibitive, making the training very slow and costly.

Aware of these challenges, some works advocate for training optical neural networks directly with non-ideal physical responses, called physical network training, as shown in Fig. 9(b). The forward pass is executed on PICs and the loss signal $\mathcal{L}$ is obtained by comparing the physical and intended outputs. However, a challenge arises during the backward pass due to the undifferentiable nature of the physical response. To address this challenge, the idea of adopting a differentiable PIC surrogate model in digital domain has been proposed [89], [147], [148]. With the differentiable model, the gradient of the loss can be propagated back with respect to the controllable parameters. This strategy obviates the need for tedious modeling and analysis of on-chip noise sources, incorporating the noisy behavior of photonic chips naturally during training.

2) In-Situ Training: In-situ training aims to perform training directly on-chip, enabling the inherent consideration of all kinds of on-chip non-idealities, as shown in Fig. 9(c). This approach can potentially boost accuracy to the greatest extent by directly incorporating the real-world behavior of photonic hardware into the training process.

Although *in-situ* training is *straightforward* and *ideal*, it is challenging to implement directly on the optical computing platform. Considering an optical neural network layer l, optical components execute the linear projection $W^l(\mathbf{\Phi}^l)$ complemented by a digital or analog nonlinear transformation f^l. With the input $\mathbf{x}^l$, the forward manner can be defined as,

$$\begin{aligned} \mathbf{y}^l &= W^l(\mathbf{\Phi}^l)\mathbf{x}^l \\ \mathbf{x}^{l+1} &= f^l(\mathbf{y}^l). \end{aligned} \tag{13}$$

Here, $\mathbf{\Phi}^l$ represents the device configurations, which are the device control variables. For the backward pass, assuming we can access the gradient of loss over $\mathbf{x}^{l+1}$, $\partial\mathcal{L}/\partial\mathbf{x}^{l+1}$, we need to obtain the gradient over input $\mathbf{x}^l$ and device configurations $\mathbf{\Phi}^l$. The first step is to obtain the gradient over $\mathbf{y}^l$, which could be prohibitive if the activation function is undifferentiable without analytical formulations, especially for customized analog activations. After determining the gradient $\partial\mathcal{L}/\partial\mathbf{y}^l$, the gradients over inputs and device configurations are defined as $W^l(\mathbf{\Phi}^l)^T\frac{\partial\mathcal{L}}{\partial\mathbf{y}^l}$ and $\frac{\partial\mathcal{L}}{\partial\mathbf{y}^l}\mathbf{x}^T\frac{\partial\mathbf{W}^l(\mathbf{\Phi}^l)}{\mathbf{\Phi}^l}$. The challenges of obtaining the above two items stem from several aspects. First, it requires bidirectional input support to access the transpose of $W^l(\mathbf{\Phi}^l)$. Second, one

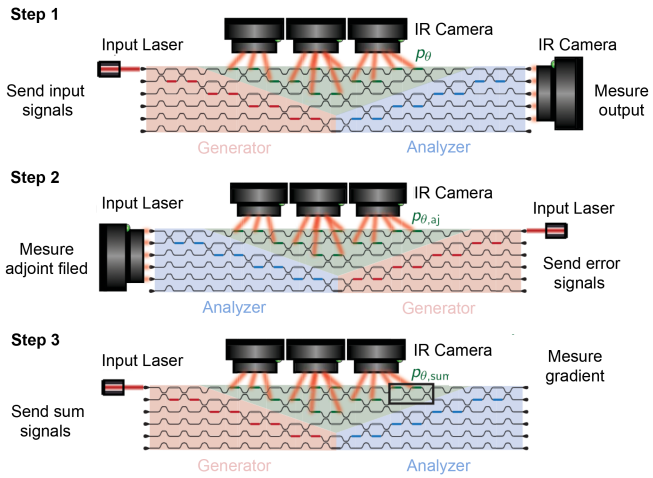

Fig. 10. In-situ backpropagation concept of adjoint-variable method [151]. The forward, backward, and sum steps of their backpropagation concept are shown to derive the gradient with respect to phase parameters θ.

may argue that another optical processor can be used to implement matrix multiplication in the backward pass, thereby avoiding the need for bidirectional support. However, noise and precision limitations cause the gradient computation to deviate from the desired matrix multiplication, thus impeding convergence. Third, obtaining the analytical gradient over the real device configurations, $\partial \mathbf{W}^l(\mathbf{\Phi}^l)/\mathbf{\Phi}^l$, can be very complex and prohibitive, as demonstrated in the case of an MZI-ONN [149].

An adjoint-variable method was proposed theoretically to implement on-chip backpropagation by interfering with adjoint and forward fields [150]. Recent work has implemented the concept of in-situ backpropagation within a triangular MZI mesh [151]. As depicted in Fig. 10, the method involves forward and backward signal propagation, followed by gradient calculation. In step 1 and step 2, the "forward inference" signal x and "backward adjoint" signal x_{aj} are sent forward, respectively. Then the "sum" vector $x - i(x_{aj})^*$ is send forward again. The gradient with respect to the phase variables can be finally obtained. However, this implementation requires additional power/phase monitors, faster microcontrollers, and more precise detectors, which increases hardware control complexity and imposes scalability concerns. Additionally, Ohno et al. attempted to design an MRR crossbar array capable of implementing matrix multiplication between the gradient and the transpose of the weight matrix [123], which is a fundamental operation in conventional backpropagation algorithms. While promising, this method has yet to be demonstrated on-chip. Besides, there is a line of works on on-chip learning protocols evaluated in simulation, which will be discussed in Section V-D.

IV. Photonic-Electronic AI Accelerator: A Glance at the Architecture Level

Photonic AI computing is experiencing rapid advancements in both device and circuit-level innovations. To fully harness the potential of optical computing, it's imperative to develop comprehensive systems that combine PIC-based computing with key auxiliary components, such as memory and datapath elements. This necessitates architecture-level efforts to translate circuit-level innovations into practical frameworks suitable for real-world applications. However, the architecture-level study of photonic AI accelerators is still in its infancy, and limited research has been conducted in this direction. Given the significance of comprehending photonic AI accelerators, this section provides an overview from an architectural perspective, focusing deeply on the system components and covering various design considerations.

A. Fully-Optical Vs. Photonic-Electronic Accelerators

1) Fully-Optical Accelerator: A fully optical accelerator refers to implementing all operations, including both computation and necessary nonlinear activation functions, entirely within the optical domain. Recent works have demonstrated the integration of optical computation and on-chip nonlinearity [32], [152], [153], [154]. Fully optical accelerators, although promising high bandwidth without the power consumption of E-O interference, still confront significant challenges due to scalability and practical implementation concerns. Firstly, the on-chip nonlinearity is not yet a mature scheme with low energy efficiency or flexibility compared to electronics, as discussed in Section III-C2. Secondly, the significant loss imposes a substantial requirement for optical power to meet the detection threshold. Additionally, computation errors will accumulate along the optical computing layers, deviating the final outcome significantly from the expected value.

2) A More Practical Paradigm: Photonic-Electronic Accelerator: Given the challenges associated with practically implementing all-optical accelerators, the photonic-electronic hybrid accelerator emerges as a more feasible and competitive photonic AI solution [155], which is also the key focus in this section. The current hybrid accelerator setup takes advantage of both emerging photonics and mature electronics and builds a system with a tight integration of photonic and electronic integrated circuits. The intensive tensor operations are executed on optical parts in the analog domain, while the electronic segment includes digital memory for data storage and distribution, as well as essential units for data writing/reading, flow control, and minor data processing. Combining digital and analog domains results in a mixed-signal setup, therefore, requiring E-O/O-E, D/A, and A/D conversions. Although the conversion processes incur additional power consumption and delay, they also offer some advantages. For example, the A/D conversion process can be viewed as a denoising step, as it involves discretizing the continuous analog signals into digital representations. This discretization helps filter out noise that is present in the analog signals, avoiding error propagation.

B. Architecture and Workload Mapping

1) Architecture: The photonic-electronic accelerator can be classified into two types based on its application, which range from task-specific accelerators, such as those for CNNs [156], [157], [158], to general-purpose architectures [126], [159]. Despite their varied applications, these accelerators exhibit a similar generic high-level micro-architecture, as shown in Fig. 11.

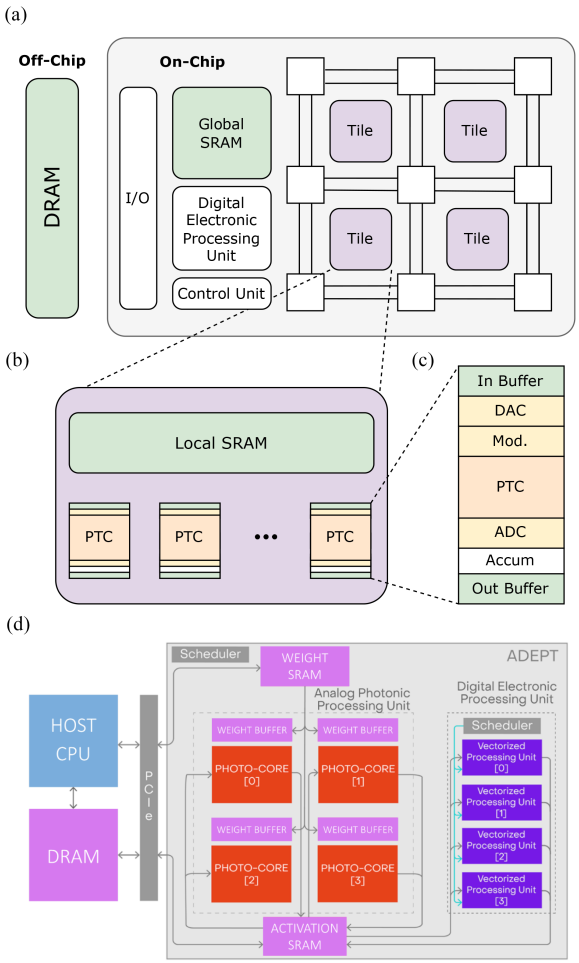

Fig. 11. The generic photonic accelerator architecture. (a) A high-level representation of the full system architecture, including off-chip and on-chip components. (b) The architecture representation of a single tile. (c) The details of the linear computing unit using photonic components. (d) The system architecture of ADEPT interacting with peripheral components [159].

The figure describes the generic architecture in a top-down manner, from the top view of an on-chip accelerator to interaction with peripheral devices, such as off-chip dynamic random-access memory (DRAM). Within the accelerator, the chip node contains both the photonic and electronic circuits. The photonic part handles tensor operations, while the electrical component is responsible for flow management. Each tile is constructed with multiple PTCs and a shared local static random-access memory (SRAM) to store data that can be accessed by different PTCs. The PTC necessitates several digital units to support its operation, such as the in-buffer and out-buffer for data storage, ADC, DAC, and the modulator for signal conversion.

Memory: Memory design is a critical aspect of modern accelerator design, given that data movement frequently constitutes the bottleneck of the entire system. This challenge is observed not only in traditional electronic processors but also in optical processors. [160]. In addition to addressing costly data movement, the ultra-fast characteristic of photonic computing requires specialized memory design to accommodate the requirements for high-speed data access.

Several key design considerations can be taken into account during the memory system design process: 1) *Memory hierarchy:* Adopting a memory system with multiple hierarchical levels is a common strategy to reduce memory access costs and meet the speed requirements. Previous studies have organized memory systems from external off-chip DRAM to internal on-chip SRAM [126], [156], [157], [159], [161]. This progression typically involves using global SRAM to drive multiple local SRAMs that store frequently accessed data. For example, Lightmatter has introduced an architecture denominated ADEPT as shown in Fig. 11(d) [159], featuring separate SRAMs for activations and weights to accommodate their differing access frequencies and patterns. Specifically, weight buffers are utilized to facilitate communication between the large but slow global SRAMs and the ultra-fast PTCs, enabling parallel programming of weight blocks. 2) *Bandwidth matching:* Towards efficient data transfer, bandwidth matching is critical to ensure that the memory bandwidth aligns with the requirements of PTCs. One effective approach is to integrate multiple memory channels, as demonstrated in Ref [126], [159]. To meet latency demands and enable efficient transmission of parameter blocks, the large global SRAM block is partitioned into smaller sub-arrays. This allows data to be read at brief intervals, aligning actual data access latency with specified requirements. 3) *Data prefetching:* Given the regularity of neural network workloads, future memory access can be anticipated, allowing for the use of double buffering to hide data loading latency, as demonstrated in [126]. 4) *Data reuse:* A common approach in photonic accelerators, known as 'weight static', explores weight locality to enhance reuse and minimize data movement.

Optical computation: The fundamental computing primitive is a single PTC designed to accelerate tensor operations. The accelerator could integrate multiple PTCs, allowing for task distribution across multiple cores, thus reducing execution time and increasing throughput. These multiple on-chip PTCs can be further organized into tiles, following a modular design principle where each tile shares associated memory and control logic. Different tiles communicate with each other through router logic, using network-on-chip or optical interconnect, as demonstrated in Fig. 11(a).

Digital electronic processing unit: In neural network acceleration, various non-GEMM operations are essential, such as element-wise non-linear operations (e.g., ReLU, GELU, and sigmoid), reduction operations (e.g., softmax and max-pool), batch and layer normalization, and element-wise operations (e.g., bias). Digital electronic processing units are employed for these tasks due to their higher efficiency. However, to retain the performance benefits of photonics, it is crucial to match the high throughput of photonic part, especially considering the challenges faced by electronic processing units operating beyond 2 GHz. Demirkiran et al. proposed to equip each PTC with vectorized processing units and duplicate logic units internally to enhance processing speed [159].

2) Workload Mapping: This subsection provides an overview of how photonic accelerators perform neural network inference, especially when handling matrix-multiplication-related workloads. The workload needs to be partitioned and

scheduled appropriately to map the matrix multiplication to PICs.

The first mapping method is specific to convolutional layers, where the mapping honors the sliding window property of convolution. The sliding of the input feature map through the kernel can be viewed as a series of dot products between the kernel matrix and the sliding receptive fields. In [30], the sliding receptive fields are sent to the photonic accelerator with a time-wavelength interleaving to enable the multiplication with kernel weights in a temporal manner. [156], [157] further cascade multiple sliding receptive fields with the same kernel to perform an MVM operation. However, this mapping method is specific to convolutional layers, which limits its generalizability across non-convolutional workloads.

The alternative mapping approach involves transforming the tensor multiplication task into a GEMM operation between two large matrices. This conversion is straightforward for linear layer workloads. In the case of convolutional layers, the image-to-column (img2col) technique is commonly employed to convert them into MVM tasks [162], [163]. Matrix tiling is required to partition the input matrices into smaller blocks, and the block-wise multiplication is executed on the photonic accelerator. However, the scheduling of these small blocks, i.e., the order in which these tiles are processed in PTCs, requires careful design, which is sometimes overlooked in recent optical computing studies. Dataflow selection has a significant impact on this scheduling. Currently, many optical computing platforms constrain dataflow selection to weight-stationary dataflow [28], [159]. This is because weight programming is often slow and costly, favoring the weight-static mode. In this scheme, block-wise weight matrices are kept in the PTC to maximize their reuse among different input data sizes. However, recent advancements in dynamically-operated PTC designs, such as those proposed in [126], [164], have eliminated these restrictions due to advancements in weight programmability, enabling flexible dataflow selection based on workload requirements. For example, output-stationary dataflow is employed in [126] to support attention matrix multiplication, as the matrices have limited reuse opportunity in this dynamic matrix multiplication scenario.

V. Toward Efficient Photonic AI Computing: A Software-Hardware Co-Design Perspective

In this review, we have explored the promising potential of photonics for AI acceleration. As a highly interdisciplinary field, photonic AI computing requires contributions from various domains, including device, circuit, architecture, and algorithm levels. Previous sections have examined a range of efforts from academia and industry across the device, circuit, and architecture parts. However, solely focusing on progress at the individual component level is insufficient to fully unlock the vast potential of photonic computing. Instead, a holistic approach across multiple levels is crucial for maximizing the performance of photonic AI systems, necessitating a software-hardware co-design perspective.

In this section, we will delve into the key challenges of current photonic AI systems, specifically in terms of area density, energy efficiency, noise robustness, and trainability. We will feature representative studies that address these challenges through efforts across multiple levels of the system.

A. Area

1) Issue: The large spatial footprint of PICs is a significant concern, as optical devices typically have much larger physical dimensions compared to nanometer-scale transistors, spanning hundreds or thousands of square micrometers. In this case, PICs generally have low packing density and are not competitive in area efficiency. Consequently, it becomes challenging to accommodate a large number of photonic components or a large matrix on a single chip, limiting hardware scalability and compute density.

2) Co-Design Progress: The concern over area cost has prompted the development of compact photonic devices and the advancement of fabrication processes. This, combined with various hardware-efficient PTC designs discussed in Section III-B3, aims to address the challenge of large spatial footprints in PICs. Beyond the methods focusing solely on device or circuit levels, researchers are exploring holistic device-circuit-algorithm co-design efforts in two promising directions: 1) domain-specific photonic computing engines to trade off between matrix expressivity and hardware efficiency; 2) AI-assisted automatic compact PTC design.

The first line of focuses on developing domain-specific photonic computing engines [89], [127], [129], [158], [165], [166], instead of universal linear units as in previous works, such as MZI meshes and MRR banks [28], [121]. The over-parameterization of neural networks has inspired various research efforts exploring the construction of efficient neural networks beyond conventional GEMM within a restricted matrix parameter space, including low-rank neural networks and structured neural networks [167], [168]. We refer to such neural networks with a restricted weight matrix space as subspace neural networks [89], [125]. Subspace neural networks have shown considerable improvements in efficiency while maintaining comparable representability to traditional neural networks. The success of efficient subspace neural networks can be leveraged in ONNs by sacrificing the universality of weight representation in exchange for higher hardware efficiency. For instance, Gu et al. proposed to implement an efficient circulant neural network and devise a novel butterfly photonic architecture with improved area efficiency over previous circuits [165], [168]. It implements the optical fast Fourier transform (OFFT) and its inverse (OIFFT), which are used to efficiently perform circulant matrix multiplication. Gu et al. further extended the proposed architecture to a trainable transform structure to enable the implementation of more matrix transformation [165]. Moreover, to solve the issue of quadratic increase in the number of MZI devices when supporting larger matrices in an MZI mesh [28], Xiao et al. applied tensor-train decomposition first to decompose large over-parameterized weight matrices into

smaller ones, thus substantially reducing the number of MZI devices required [166].

Another noteworthy advancement in this field is the exploration of automatic PTC design, departing from the manual design paradigm. In this approach, the footprint can be incorporated as a constraint to enable the automatic generation of compact PTC. For instance, ADPET introduces the first automatic AI-assisted differentiable search framework for PTC topology design [169]. It first constructs a probabilistic photonic SuperMesh and then employs differentiable optimization in a huge and highly discrete PTC search space. This framework adapts to various circuit footprint constraints and foundry PDKs. The PICs developed using this method demonstrate a substantial increase in footprint compactness, ranging from 2 to 30× compared to both the MZI mesh and the manually designed compact butterfly mesh. This automated approach promises to revolutionize the design process, enabling the creation of more efficient and compact PTCs for AI applications.

B. Energy Efficiency

1) Issue: Energy efficiency is a crucial metric for evaluating computing hardware. When assessing the energy efficiency of photonic AI hardware, especially given its mixed-signal setup, it is essential to consider the energy costs associated with both the digital and optical components of the system. Following previous studies [126], [170], the energy cost of photonic computing can be broadly categorized into optical costs associated with performing computation and electrical costs related to loading operands (X and Y) and detecting output (O), which can be expressed as follows:

$$E = E_{\text{laser}} + \underbrace{E_{\text{comp}}}_{\text{compute}} + \underbrace{E_{\text{load}}}_{\text{load X}} + \underbrace{E_{\text{load}}}_{\text{load Y}} + \underbrace{E_{\text{det}}}_{\text{detect O}}$$

$$E_{\text{load}} = \boldsymbol{E_{\text{read}}} + \boldsymbol{E_{\text{DAC}}} + E_{\text{mod}}$$

$$E_{\text{det}} = E_{\text{PD}} + \boldsymbol{E_{\text{ADC}}} + E_{\text{amp}} + \boldsymbol{E_{\text{write}}}. \tag{14}$$

Here, E_{load} encompasses the energy costs for memory reading (E_{read}), D/A conversion (E_{DAC}), and signal modulation (E_{mod}). E_{det} represents the costs of optical signal detection (E_{PD}), signal amplification (E_{amp}), A/D conversion (E_{ADC}), and the subsequent writing of results back to memory (E_{write}). E_{comp} includes other energy costs associated with performing computation, which could be negligible for a fully passive PIC.

Among all components, the transition between digital and analog signals presents a significant bottleneck in the system energy consumption, particularly evident in data movement (memory-associated cost) and ADC/DAC, as labeled in bold in (14). This transition can occupy more than 80% of the overall energy consumption [160], [171], which is different from the power composition of the optical digital computing outlined in Section II-B.

2) Co-Design Progress: Device-level advancement promises straightforward reduction in energy costs in (14), such as the progress in energy-efficient ADCs [172], efficient optical modulators, and on-chip laser [173]. However, in this section, we explore recent advancements beyond the device level, with a particular focus on addressing the signal conversion and memory aspects.

TABLE II
COMPARISON OF RECENT APPROACHES ON SAVING SIGNAL CONVERSION COST, INCLUDING D/A (E_{DAC}), A/D(E_{ADC}), E-O(E_{MOD}) ENERGY COSTS

Method	Input X	Weight Y	Output O
Optical broadcast [121]	$E_{\text{DAC}}, E_{\text{mod}} \downarrow$	-	-
Weight-static dataflow [159]	-	$E_{\text{DAC}}, E_{\text{mod}} \downarrow$	
Spectral parallelism [124], [157]	-	$E_{\text{DAC}}, E_{\text{mod}} \downarrow$	
Time-integration [126], [129], [158]	-	-	$E_{\text{ADC}} \downarrow$
Electro-Optic analog memory [161]	-	$E_{\text{DAC}} \downarrow$	-
Non-volatile device [124], [174]	-	$E_{\text{mod}} \downarrow$	-

At the circuit and architecture levels, several efforts have been made to reduce signal conversion and data movement costs, as summarized in Table II. Optical broadcast is a widely-used technique that enables spatial sharing of encoded signals [121], thereby saving on DAC and E-O modulation costs. This approach has been further extended to a crossbar-style design in [126] to enable both input and weight to be spatially shared. Another strategy involves keeping weights static in photonic devices or employing non-volatile devices to reuse encoded signals across different inputs temporally. Additionally, leveraging spectral parallelism allows for sharing operands among different inputs. Both approaches can lower the encoding costs related to DAC and modulation. Time-integration techniques have been employed to explore analog-domain temporal accumulation [132], significantly reducing the A/D conversion frequency and preserving more computations within the analog domain. Recently, the use of E-O analog memory has gained attention [161], where analog memory is placed near photonic devices, and DACs are reused across rows of analog memory, thereby reducing DAC costs.

At the algorithm level, many studies have adopted low-bit quantization techniques to preserve accuracy while using low-precision weights, inputs, and activations. [128], [146], [165], [175]. This approach can reduce the required DAC and ADC costs with lower resolution, as well as reduce the memory data movement cost. Furthermore, the residue number system has been explored as a method to reduce precision requirements by achieving high-precision computations using low-precision components [176].

C. Noise Robustness

1) Issue: Ensuring functional correctness is a fundamental requirement of computing hardware. However, analog photonic computing inherently faces robustness challenges due to two primary factors, as depicted in Fig. 12: 1) Various non-ideal conditions include process variations, device noises, environmental factors, and limited endurance. 2) Limited precision of inputs and outputs arises from the finite control resolution and the substantial overhead associated with high-precision DACs and ADCs. Unlike low-precision electronics digital computing, as shown in Fig. 12(a), photonic computing uniquely suffers from output precision loss in the conversion of analog outputs back to the digital domain, where the ADC precision is typically unsatisfactory compared to the output precision. To understand

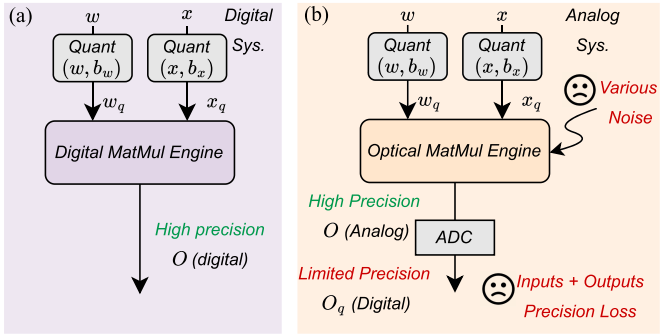

Fig. 12. Comparison between (a) the low-precision digital matrix multiplication (MatMul) engine and (b) the low-precision optical MatMul engine. The optical MatMul engine suffer from various non-idealities during computation and hold precision loss for both inputs and outputs.

the mismatch between an ADC and output precision, consider a straightforward scenario in which a photonic computing engine executes a dot product operation of length N. Note that here we consider an ideal case where the precision constraints are on the weights and activations, and the photonic computing engine can be treated as a dot-product engine based on a scalar product.[1] The multiplication of a B_w-bit signed weight by a B_x-bit signed activation results in a product with $B_w + B_x - 2 + \log_2 N$ magnitude bits and one sign bit. Here, the additional $\log_2 N$ bits of precision stem from the addition of N scalar products. However, achieving this desired precision often exceeds the capabilities of affordable ADCs, typically within 8 bits [177]. For instance, a 4-bit weight and 4-bit activation in a length-32 dot product computation would ideally yield an output precision of 13 bits. The 5-bit overhead underscores a notable difference between photonic computing and traditional low-precision arithmetic units like int8 GPU tensor cores where outputs are preserved in high precision [178], raising a unique precision issue.

2) Co-Design Progress: Here, we explore efforts to mitigate the precision issue and the impact of noise on accuracy. Regarding precision issues, previous efforts have often drawn inspiration from quantization, either employing Quanztiation-aware Training (QAT) [128], [146], [175], [179], [180], or Post-training Quantization (PTQ) [181]. These approaches enable neural networks to inherently tolerate low-precision arithmetic, thereby mitigating the impact of precision limitations. Specifically, Gu et al. first developed the QAT flow by directly optimizing low-precision device control signals in the discretized space [146]. However, the aforementioned quantization techniques are insufficient to compensate for the computation information loss due to the low-precision ADC. This is because errors occur in each partial result, and they finally accumulate when we tile a large matrix workload. Some work in analog computing specifically tune the ADC reference voltage to balance the tradeoff between dynamic quantization range and solution [182]. Furthermore, weight and input slicing are proposed as strategies to manage the trade-off in ADC resolution [183].

For noise mitigation, noise-aware training has emerged as a widely used technique to enhance resilience [132], [145], [146]. This approach involves introducing noises and variations during the training process to enhance the noise resilience of ONNs. Previous studies have focused on modeling various noise sources [132], [184], such as dynamic noise, static manufacturing variation, and thermal crosstalk, then incorporated them into the training process. Furthermore, explicit robustness optimization terms can be integrated into the training process to further enhance robustness against noise. For instance, [146] estimates the noise sensitivity of weights and applies protective regularization terms to sensitive weights during optimization. Similarly, Zhu et al. introduced an additional regularization term on the phase magnitude in an MZI mesh to reduce the crosstalk concerns [185]. Additionally, knowledge distillation strategies, as employed in [186], can be employed to guide the optimization of noisy student ONN models under the guidance of a noise-free teacher model. This approach significantly enhances the robustness of models against static process variation and dynamic input signal noises.

However, noise patterns can be highly intricate, rendering the analysis and modeling of these patterns both challenging and time-consuming. In response to this challenge, some studies have explored training ONNs directly with noisy physical responses from real chips, eliminating the need for explicit noise modeling [89], [147], [148]. This approach, known as physical neural network training, has been introduced in detail in Section III-D. Moreover, on-chip training is another way to recover accuracy by inherently modeling on-chip noise during training, which is further explored in Section V-D.

Besides exploring the inherent noise tolerance of neural networks, another straightforward direction is to compress the on-chip noise levels. Specifically, some work advocate employing device-level and circuit-level design space exploration to mitigate the impact of process variation and crosstalk [187], [188]. [179] suggests avoiding frequent reprogramming on PCM cells by enhancing the similarity of mapping weights to mitigate early wear-out. Additionally, reducing the number of active devices in PICs can also decrease on-chip noise, as noise-induced errors typically correlate positively with the number of noise sources. Therefore, pruning redundant devices [169], or weight blocks [89], [165], brings significant noise robustness improvement.

D. Self-Learnability

1) Issue: The adaptability or trainability of photonic analog computing platforms is another major challenge in their practical application. The significance of self-learnability arises from three main aspects, as shown in Fig. 13. Firstly, on-chip training enhances the adaptability of optical hardware when working conditions drift, or when workloads change. Secondly, the realization of on-chip training can unlock many important edge learning applications such as local online learning, transfer learning, and lifelong learning. Moreover, as discussed

[1]The real case can be more intricate. For example, low precision may be reflected in device control variables, and photonic computing cannot be easily treated based on scalar product.

TABLE III
COMPARISON OF DIFFERENT ON-CHIP TRAINING PROTOCOLS ON GRADIENT ESTIMATION METHODS AND NOTABLE HARDWARE REQUIREMENTS

On-chip training protocol	Backward pass	Hardware requirements	ONN scale
Gradient-free Opt. [28], [116], [189]	No gradient	Precise device control	Low
Direct feedback alignment [190]	Estimated gradient with random projection	Random projection unit	Low
Adjoint-variable method [150], [151]	True gradient with adjoint method	Bidirectional I/O, per device monitor	Low
Zeroth-order Opt. [191]–[193]	Estimated gradient with finite difference method	Forward-only	Medium
First-order Opt. [149]	True first-order gradient	Bidirectional I/O	High

We also show the ONN Scale they can handle in terms of low (~100), medium (~1000), and high (> 10k).

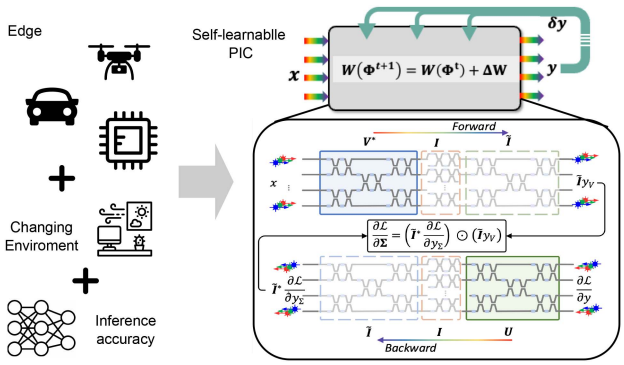

Fig. 13. The self-learnable ONN paradigm exemplified by a single scalable ONN on-chip training framework with *in-situ* gradient calculation [149].

in Section III-D, on-chip training can address robustness issues *in-situ* and bridge the gap between simulation and real implementation.

2) Co-Design Progress: Achieving self-learnability on neuromorphic photonic processors is indeed challenging due to various on-chip restrictions, such as inaccessible gradients for control variables and a lack of full observability for in-situ light fields. Consequently, there is a demand for developing hardware-friendly learning algorithms that can operate within these constraints while maintaining feasibility.

We summarize recent on-chip training protocols, either validated through simulation or real demonstration, in Table III. Generic gradient-free optimization methods, e.g., evolutionary algorithms and brute-force device tuning [28], [116], [189], have been used to optimize on-chip parameters with no gradient information involved. Nevertheless, these approaches encounter difficulties when scaling to large-scale models with slow convergence and poor stability. Besides the gradient-free approach, investigation into the direct feedback alignment (DFA) training algorithm is further explored in [190] for *in-situ* training [194], [195]. DFA propagates errors through fixed random feedback projections from the output layer to each hidden layer in parallel, thereby obviating the necessity for the sequential backpropagation of gradients. Another approach, the adjoint variable method, was introduced to perform on-chip backpropagation by computing the gradients *in-situ* with per-device optical field monitors [150], and recent work has experimentally demonstrated the concept [151]. However, scalability remains a significant concern due to considerable hardware support overhead, such as the need for per-device monitoring.

On-chip training protocols based on forward-only zeroth-order optimization aim to approximate gradients using the finite difference method, perturbing parameters with small random values for gradient estimation. Instead of adjusting individual on-chip parameters separately, Bandyopadhyay et al. proposed a method that samples perturbations for all parameters and shares them among all training instances in a single iteration [191]. Similarly, FLOPS, proposed in [192], shares the sampled perturbations among single mini-batches instead of all mini-batches in the same interaction, showing better convergence than [191]. To further enhance the scalability of zeroth-order training protocols, MixedTrain proposed in Ref [193] partitions PICs into passive and active regions and trains only a small subset of active devices in each iteration, ensuring the capacity to handle a larger scale ONN compared to FLOPS. The zeroth-order-based approach shows the training scale to thousands of MZIs as shown in [192], [193], while it is still not enough when dealing with modern-size machine learning models. L ^{2} ight introduces a subspace optimization algorithm and develops a method for *in-situ* calculation of first-order subspace gradients [149], as shown in Fig. 13. Additionally, a multi-level hardware-aware sparse training method is employed to boost training efficiency. This study demonstrates the first instance of training a million-parameter-level ONN, showcasing exceptional stability.

VI. OUTLOOK AND CONCLUSION

In this review, we have outlined recent advancements in photonic-electronic integrated circuits for computing and AI tasks, covering the progress made at various levels, including the device, circuit, architecture, and system level, as well as progress in cross-layer software-hardware co-design strategies. However, more technical challenges need to be addressed in the future to enable the practical application of photonic computing. Here, we point to possible research directions.

1) Device/new Material Innovation: As highlighted in this review, the characteristics of optical components directly impact the performance of photonic-electronic computing systems. Recent research in emerging materials and novel devices provide new opportunities for enhancing the performance of PICs from the device level. For instance, heterogeneous modulators, leveraging technologies such as III-V MOSCAP and 2-D materials [98], [111], demonstrate high modulation efficiency, gigahertz-level cut-off frequencies, and sub-picojoule-per-bit power consumption, allowing a compact layout with high energy efficiency. Additionally, dataflow management in PIC-based computing, which includes both digital and analog systems,

primarily depends on electrical memory due to the lack of equivalent components in the photonic domain. Previous studies have demonstrated that non-volatile materials and devices have the capability to store data, including pre-trained weight parameters [32], [124]. Significantly, non-volatile photonic memory, in contrast to conventional electronic memory, serves not only as a data storage solution but also functions as a computational unit for 'in-memory' computing. Nevertheless, due to the limited switching frequency of current non-volatile materials, such as PCMs, scenarios requiring dynamic data reading and writing (particularly in high-performance digital processing units) demand high-speed, reconfigurable materials for optical memory.

Beyond the innovation in optical components for computing reviewed in Section III-B3, optimizing other aspects of PICs also presents significant interest. Potential research directions encompass, but are not limited to 1) on-chip light sources to eliminate the complex fiber packaging and alignment processes; 2) integrated frequency combs as WDM sources [124]; 3) customized nonlinear units for on-chip implementation of programmable activation functions; and 4) efficient ADC/DACs and photodetectors, specifically designed for optical computing architectures [196].

2) Advanced Photonics and Electronics Integration: PICs for computing require on-chip interfacing with diverse electronic components, including signal conversion circuits and control systems, leading to substantial data communication between the electronic and photonic parts. Currently, wirebonding is a widely adopted strategy for E-O interconnects in academia. This method is particularly convenient and cost-effective for small-scale networks with few active devices, as it requires only the routing of driving signals to the edges of photonic chips, followed by connections to peripheral electronic units via wire bonds. However, as PICs scale up, managing metal wire routing within constrained chip areas presents a challenge. The beachfront bottleneck of photonic chips and electrical control units restricts the number of connections that can be implemented due to the limited chip perimeter. Furthermore, wirebonding can constrain the bandwidth of high-frequency devices owing to impedance mismatch.

To address this interconnect challenge, advanced integration methods can be employed, aiming to reduce interconnect complexity and improve overall system performance [197]. Flip-chip bonding directly connects two dies by soldering or using conductive bumps to align matching electrical pads on their surfaces. This approach allows flexible floorplanning and pad placement beyond chip edges, enabling higher interconnect density and reducing parasitic impedance. Given the fabrication process and chip sizes, typical integration strategies involve flip-chip bonding single or multiple heterogeneous dies (such as ADC/DAC, memory, microcontroller, FPGA, ASIC, etc.) onto a photonics chip which also serves as an interposer. However, this approach will invariably introduce large temperature excursions on the PIC varying with the workload. To address this issue, one strategy involves developing temperature-resilient photonic platforms [198]. Additionally, the heat spreader and through-silicon via (TSV) technique enables more flexible thermal management within chips [33], [199], [200], [201].

Monolithic fabrication processes that integrate both photonic and electronic components on a single substrate also have shown promising results [202], [203]. While this technology enables higher integration levels, bandwidth, and energy efficiency, significant potential remains for the optimization of fabrication processes, improving yield, and the development of comprehensive PDKs. Besides, further breakthroughs, such as integration with more advanced CMOS nodes and 2.5-D/3-D electro-optical integration, are anticipated to enhance the performance of photonic chips.

3) Scale to Large Models and Advanced Tasks: Photonic computing still faces challenges concerning scalability when moving to support large models. Several directions can be explored. First, the continued exploration of domain-specific PTCs which trade universality for higher scalability, as indicated by previous research efforts [89], [127], [129], [158], [165], [166]. Second, there exists a substantial potential to further improve computational density by engineering tailored, compact photonic devices, such as multi-operand modulators and metasurface-based devices. Third, the exploitation of the unique characteristics of photonics, such as wavelength, time, or mode-division multiplexing, can enable the reuse of hardware for a higher degree of parallel operations. Lastly, the development of a deep understanding of the workload of evolving machine learning models is crucial for designing suitable PICs. For instance, attention-based Transformer models introduce dynamic matrix multiplication, challenging previous PTC designs optimized for CNNs. Recent advancements propose dynamically-operated PTC design to handle dynamic matrix multiplication efficiently, ensuring optimal performance across diverse AI applications.

4) On-Chip ONN Training Protocol: The mainstay of ONN training predominantly relies on simulation, while existing on-chip training experimental demonstrations remain at a small scale with notable overheads. A stable and efficient on-chip training scheme is highly coveted, necessitating breakthroughs in both hardware and training algorithms. These breakthroughs may include innovations such as light-field-driven nonvolatile materials and advancements in training algorithms.

5) Cross-Layer Co-Design and Electronic-Photonic Design Automation (EPDA): Cross-layer efforts open up additional opportunities to optimize the performance of ONNs, as discussed in Section V. With the growing complexity of photonic-electronic hardware platforms, exploring EPDA becomes critical for enhancing productivity and efficiency, such as exploring automatic circuit layout generation and fast photonic circuit simulation [204].

6) System-Level Simulator: As a multidisciplinary emerging area, it is crucial to have a comprehensive simulator framework for evaluating the performance of optical computing systems. Ideal simulators should support seamless integration of new optical hardware, offer automatic algorithms for hardware mapping, and provide evaluation of chip-level performance. Such tools will facilitate fair and straightforward evaluation across different optical circuit designs, helping identify system bottlenecks and guiding further optimizations.

7) Optical AI Software Stack: Developing specialized compilers and instruction sets tailored for photonic computing architectures is necessary to enable the integration of novel optical hardware into the mainstream AI software stack, as envisioned in [205]. While existing deep learning frameworks such as PyTorch and TensorFlow can still serve as the front end, the compiler component must be adapted to generate machine code optimized for the unique photonic accelerators. This ensures efficient utilization of the optical hardware. Moreover, we anticipate open-source efforts in this direction to facilitate the progress of the optical AI software stack.

In conclusion, photonic-electronic integrated circuits stand out for their exceptional advantages in computing–in terms of low latency, high bandwidth, and energy efficiency–and exhibit a high potential to overcome the insurmountable bottlenecks of electronic computing. Nevertheless, whether for digital or analog computing, the PIC scale of reported work remains limited, and further improvements in scalability are essential. Continuous research should aim to enhance throughput via innovative devices, architectural improvements, specialized training algorithms, and hardware-software co-design strategies. Additionally, efforts should focus on reducing power consumption and the costs associated with the E-O interface, in order to rival the performance of state-of-the-art electronic computing systems. This ambition, however, does not imply that the objective of optical computing is to surpass digital-electronic processors in all metrics and replace them. From an industry perspective, optical computing needs to identify niches where it excels over its electronic counterparts. A promising short-term application involves leveraging the high parallelism of optical computing to develop MVM processing units and accelerators for neural networks, an area that has been extensively studied. If the PIC scale can be well aligned with its tasks (including considerations for multiplexing), and use an efficient E-O interfaces, then operation at a rate of one MVM per nanosecond could be achieved and offer significant advantages in high-throughput applications. With sustained innovation and effort, integrated photonics is poised to become a pivotal emerging technology, satisfying the escalating societal demand for high-performance computing and hardware acceleration in AI applications over the long term.

REFERENCES

[1] F. Fang et al., "Towards atomic and close-to-atomic scale manufacturing," *Int. J. Extreme Manuf.*, vol. 1, no. 1, 2019, Art. no. 012001.

[2] R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, Oct. 1974.

[3] M. M. Waldrop, "More than moore," *Nature*, vol. 530, no. 7589, pp. 144–148, 2016.

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

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, Comput. Biol. Learn. Soc., 2015, pp 1–14.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol., Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[9] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.

[10] M. Bojarski et al., "End to end learning for self-driving cars," 2016, *arXiv:1604.07316*.

[11] A. Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Med.*, vol. 25, no. 1, pp. 65–69, 2019.

[12] K. Cao et al., "Large-scale pancreatic cancer detection via non-contrast CT and deep learning," *Nature Med.*, vol. 29, no. 12, pp. 3033–3043, 2023.

[13] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[14] D. Patterson et al., "Carbon emissions and large neural network training," 2021, *arXiv:2104.10350*.

[15] E. AI, "Tracking large-scale ai models," 2024. Accessed: Jun. 20, 2024. [Online]. Available: https://epochai.org/data/notable-ai-models

[16] D. Braga and G. Horowitz, "High-performance organic field-effect transistors," *Adv. Mater.*, vol. 21, no. 14/15, pp. 1473–1486, 2009.

[17] V. Podzorov, M. E. Gershenson, C. Kloc, R. Zeis, and E. Bucher, "High-mobility field-effect transistors based on transition metal dichalcogenides," *Appl. Phys. Lett.*, vol. 84, no. 17, pp. 3301–3303, 2004.

[18] Z. Fan, F. Qiu, A. Kaufman, and S. Yoakum-Stover, "GPU cluster for high performance computing," in *Proc. SC'04: Proc. 2004 ACM/IEEE Conf. Supercomputing*, 2004, p. 47.

[19] D. Gostimirovic and W. N. Ye, "Ultracompact CMOS-compatible optical logic using carrier depletion in microdisk resonators," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 12603.

[20] E. Timurdogan et al., "An ultralow power athermal silicon modulator," *Nature Commun.*, vol. 5, no. 1, pp. 1–11, 2014.

[21] K. Xu, "Silicon electro-optic micro-modulator fabricated in standard cmos technology as components for all silicon monolithic integrated optoelectronic systems," *J. Micromechanics Microengineering*, vol. 31, no. 5, 2021, Art. no. 054001.

[22] W. Heni et al., "Plasmonic IQ modulators with attojoule per bit electrical energy consumption," *Nature Commun.*, vol. 10, no. 1, 2019, Art. no. 1694.

[23] S. K. Mathew, M. A. Anders, B. Bloechel, T. Nguyen, R. K. Krishnamurthy, and S. Borkar, "A 4-Ghz 300-mW 64-bit integer execution ALU with dual supply voltages in 90-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp. 44–51, Jan. 2005.

[24] Z. Ying et al., "Electronic-photonic arithmetic logic unit for high-speed computing," *Nature Commun.*, vol. 11, no. 1, 2020, Art. no. 2154.

[25] T. Zhou et al., "Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit," *Nature Photon.*, vol. 15, no. 5, pp. 367–373, 2021.

[26] Y. Chen et al., "All-analog photoelectronic chip for high-speed vision tasks," *Nature*, vol. 623, no. 7985, pp. 48–57, 2023.

[27] A. E.-J. Lim et al., "Review of silicon photonics foundry efforts," *IEEE J. Sel. Topics Quantum Electron.*, vol. 20, no. 4, pp. 405–416, Jul./Aug. 2014.

[28] Y. Shen et al., "Deep learning with coherent nanophotonic circuits," *Nature Photon.*, vol. 11, no. 7, pp. 441–446, 2017.

[29] X. Meng et al., "Compact optical convolution processing unit based on multimode interference," *Nature Commun.*, vol. 14, no. 1, 2023, Art. no. 3000.

[30] X. Xu et al., "11 tops photonic convolutional accelerator for optical neural networks," *Nature*, vol. 589, no. 7840, pp. 44–51, 2021.

[31] I. Chakraborty, G. Saha, and K. Roy, "Photonic in-memory computing primitive for spiking neural networks using phase-change materials," *Phys. Rev. Appl.*, vol. 11, no. 1, 2019, Art. no. 014063.

[32] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," *Nature*, vol. 569, no. 7755, pp. 208–214, 2019.

[33] B. J. Shastri et al., "Photonics for artificial intelligence and neuromorphic computing," *Nature Photon.*, vol. 15, no. 2, pp. 102–114, 2021.

[34] K. Nozaki et al., "Femtofarad optoelectronic integration demonstrating energy-saving signal conversion and nonlinear functions," *Nature Photon.*, vol. 13, no. 7, pp. 454–459, 2019.

[35] M. Jiang et al., "On-chip integrated optical switch based on polymer waveguides," *Opt. Mater.*, vol. 97, 2019, Art. no. 109386.

[36] C. T. Phare, Y.-H. D. Lee, J. Cardenas, and M. Lipson, "Graphene electro-optic modulator with 30 GHz bandwidth," *Nature Photon.*, vol. 9, no. 8, pp. 511–514, 2015.

[37] Z. Ying, G. Wang, X. Zhang, H.-P. Ho, and Y. Huang, "Ultracompact and broadband polarization beam splitter based on polarization-dependent critical guiding condition," *Opt. Lett.*, vol. 40, no. 9, pp. 2134–2137, 2015.

[38] P. Chaisakul et al., "Integrated germanium optical interconnects on silicon substrates," *Nature Photon.*, vol. 8, no. 6, pp. 482–488, 2014.

[39] Y. Ding et al., "Ultra-compact integrated graphene plasmonic photodetector with bandwidth above 110 Ghz," *Nanophotonics*, vol. 9, no. 2, pp. 317–325, 2020.

[40] J. Michel, J. Liu, and L. C. Kimerling, "High-performance Ge-on-Si photodetectors," *Nature Photon.*, vol. 4, no. 8, pp. 527–534, 2010.

[41] Y. Tian et al., "Proof of concept of directed OR/NOR and AND/NAND logic circuit consisting of two parallel microring resonators," *Opt. Lett.*, vol. 36, no. 9, pp. 1650–1652, 2011.

[42] L. Zhang et al., "Demonstration of directed xor/xnor logic gates using two cascaded microring resonators," *Opt. Lett.*, vol. 35, no. 10, pp. 1620–1622, 2010.

[43] A. Kumar et al. "Implementation of xor/xnor and and logic gates by using mach–zehnder interferometers," *Optik*, vol. 125, no. 19, pp. 5764–5767, 2014.

[44] A. Saharia et al., "A comparative study of various all-optical logic gates," in *Optical and Wireless Technologies: Proceedings of OWT 2018*. Berlin, Germany: Springer, 2020, pp. 429–437.

[45] A. Salmanpour, S. Mohammadnejad, and A. Bahrami, "Photonic crystal logic gates: An overview," *Opt. Quantum Electron.*, vol. 47, pp. 2249–2275, 2015.

[46] H. M. Hussein, T. A. Ali, and N. H. Rafat, "A review on the techniques for building all-optical photonic crystal logic gates," *Opt. Laser Technol.*, vol. 106, pp. 385–397, 2018.

[47] M. Ota et al., "Plasmonic-multimode-interference-based logic circuit with simple phase adjustment," *Sci. Rep.*, vol. 6, no. 1, 2016, Art. no. 24546.

[48] H. Wei et al., "Quantum dot-based local field imaging reveals plasmon-based interferometric logic in silver nanowire networks," *Nano Lett.*, vol. 11, no. 2, pp. 471–475, 2011.

[49] D. Pan, H. Wei, and H. Xu, "Optical interferometric logic gates based on metal slot waveguide network realizing whole fundamental logic operations," *Opt. Exp.*, vol. 21, no. 8, pp. 9556–9562, 2013.

[50] Y. Fu et al., "All-optical logic gates based on nanoscale plasmonic slot waveguides," *Nano Lett.*, vol. 12, no. 11, pp. 5784–5790, 2012.

[51] J. J. Singh, D. Dhawan, and N. Gupta, "All-optical photonic crystal logic gates for optical computing: An extensive review," *Opt. Eng.*, vol. 59, no. 11, pp. 110901–110901, 2020.

[52] Y. Fu, X. Hu, and Q. Gong, "Silicon photonic crystal all-optical logic gates," *Phys. Lett. A*, vol. 377, pp. 329–333, 2013.

[53] E. G. Anagha and R. K. Jeyachitra, "Review on all-optical logic gates: Design techniques and classifications–heading toward high-speed optical integrated circuits," *Opt. Eng.*, vol. 61, no. 6, pp. 060902–060902, 2022.

[54] A. Politi, M. J. Cryan, J. G. Rarity, S. Yu, and J. L. O'brien, "Silica-on-silicon waveguide quantum circuits," *Science*, vol. 320, no. 5876, pp. 646–649, 2008.

[55] A. Fedorov, L. Steffen, M. Baur, M. P. da Silva, and A. Wallraff, "Implementation of a Toffoli gate with superconducting circuits," *Nature*, vol. 481, no. 7380, pp. 170–172, 2012.

[56] A. Crespi et al., "Integrated photonic quantum gates for polarization qubits," *Nature Commun.*, vol. 2, no. 1, 2011, Art. no. 566.

[57] N. C. Harris et al., "Large-scale quantum photonic circuits in silicon," *Nanophotonics*, vol. 5, no. 3, pp. 456–468, 2016.

[58] A. W. Elshaari, W. Pernice, K. Srinivasan, O. Benson, and V. Zwiller, "Hybrid integrated quantum photonic circuits," *Nature Photon.*, vol. 14, no. 5, pp. 285–298, 2020.

[59] J. Wang, F. Sciarrino, A. Laing, and M. G. Thompson, "Integrated photonic quantum technologies," *Nature Photon.*, vol. 14, no. 5, pp. 273–284, 2020.

[60] E. Pelucchi et al., "The potential and global outlook of integrated photonics for quantum technologies," *Nature Rev. Phys.*, vol. 4, no. 3, pp. 194–208, 2022.

[61] Z. Ying et al., "Silicon microdisk-based full adders for optical computing," *Opt. Lett.*, vol. 43, no. 5, pp. 983–986, 2018.

[62] L. Yang et al., "Demonstration of a directed optical comparator based on two cascaded microring resonators," *IEEE Photon. Technol. Lett.*, vol. 27, no. 8, pp. 809–812, Apr. 2015.

[63] Y. Tian et al., "Demonstration of a directed optical encoder using microring-resonator-based optical switches," *Opt. Lett.*, vol. 36, no. 19, pp. 3795–3797, 2011.

[64] H. Xiao et al., "Experimental realization of a CMOS-compatible optical directed priority encoder using cascaded micro-ring resonators," *Nanophotonics*, vol. 7, no. 4, pp. 727–733, 2018.

[65] F. Mehdizadeh, M. Soroosh, and H. Alipour-Banaei, "A novel proposal for optical decoder switch based on photonic crystal ring resonators," *Opt. Quantum Electron.*, vol. 48, pp. 1–9, 2016.

[66] C. Qiu, W. Gao, R. Soref, J. T. Robinson, and Q. Xu, "Reconfigurable electro-optical directed-logic circuit using carrier-depletion micro-ring resonators," *Opt. Lett.*, vol. 39, no. 24, pp. 6767–6770, 2014.

[67] M. Ruhul Fatin, D. Gostimirovic, and W. N. Ye, "Reconfigurable optical logic in silicon platform," *Sci. Rep.*, vol. 14, no. 1, 2024, Art. no. 5950.

[68] Z. Ying, C. Feng, Z. Zhao, R. Soref, D. Pan, and R. T. Chen, "Integrated multi-operand electro-optic logic gates for optical computing," *Appl. Phys. Lett.*, vol. 115, no. 17, 2019, Art. no. 171104.

[69] Z. Fang, R. Chen, J. Zheng, and A. Majumdar, "Non-volatile reconfigurable silicon photonics based on phase-change materials," *IEEE J. Sel. Topics Quantum Electron.*, vol. 28, no. 3: Hybrid Integration for Silicon Photonics, May/Jun. 2022, Art. no. 8200317.

[70] Y. Qi, C. Qiu, W. Gao, X. Zhong, and Y. Su, "Silicon reconfigurable electro-optical logic circuit enabled by a single-wavelength light input," *IEEE Photon. Technol. Lett.*, vol. 31, no. 6, pp. 435–438, Mar. 2019.

[71] A. H. Atabaki, A. A. Eftekhar, M. Askari, and A. Adibi, "Accurate post-fabrication trimming of ultra-compact resonators on silicon," *Opt. Exp.*, vol. 21, no. 12, pp. 14139–14145, 2013.

[72] E. Testa, M. Soeken, L. G. Amar, and G. De Micheli, "Logic synthesis for established and emerging computing," *Proc. IEEE*, vol. 107, no. 1, pp. 165–184, Jan. 2019.

[73] Z. Ying et al., "Automated logic synthesis for electro-optic logic-based integrated optical computing," *Opt. Exp.*, vol. 26, no. 21, pp. 28002–28012, 2018.

[74] Z. Zhao et al., "Exploiting wavelength division multiplexing for optical logic synthesis," in *Proc. 2019 IEEE Des., Automat. Test Europe Conf. Exhib. (DATE)*, 2019, pp. 1567–1570.

[75] J. Sklansky, "Conditional-sum addition logic," *IRE Trans. Electron. Comput.*, vol. 2, pp. 226–231, 1960.

[76] W. Zhang et al., "Time-space multiplexed photonic-electronic digital multiplier," *Photon. Res.*, vol. 12, no. 3, pp. 499–504, 2024.

[77] C. Feng, J. Gu, H. Zhu, D. Z. Pan, and R. T. Chen, "Integrated electronic-photonic barrel shifter for high-performance optical computing," in *Proc. 2022 IEEE Conf. Lasers Electro- Opt. (CLEO)*, 2022, pp. 1–2.

[78] C. Sun et al., "Single-chip microprocessor that communicates directly using light," *Nature*, vol. 528, no. 7583, pp. 534–538, 2015.

[79] C. Feng, Z. Ying, Z. Zhao, J. Gu, D. Z. Pan, and R. T. Chen, "Wavelength-division-multiplexing (WDM)-based integrated electronic–photonic switching network (EPSN) for high-speed data processing and transportation: High-speed optical switching network," *Nanophotonics*, vol. 9, no. 15, pp. 4579–4588, 2020.

[80] T. Alexoudi, G. T. Kanellos, and N. Pleros, "Optical ram and integrated optical memories: A survey," *Light: Sci. Appl.*, vol. 9, no. 1, 2020, Art. no. 91.

[81] P. L. McMahon, "The physics of optical computing," *Nature Rev. Phys.*, vol. 5, no. 12, pp. 717–734, 2023.

[82] W. Liu, W. Liu, Y. Ye, Q. Lou, Y. Xie, and L. Jiang, "Holylight: A nanophotonic accelerator for deep learning in data centers," in *Proc. 2019 IEEE Des., Automat. Test Europe Conf. Exhib. (DATE)*, 2019, pp. 1483–1488.

[83] M. Zheng, F. Chen, L. Jiang, and Q. Lou, "Priml: An electro-optical accelerator for private machine learning on encrypted data," in *Proc. 2023 IEEE 24th Int. Symp. Qual. Electron. Des. (ISQED)*, 2023, pp. 1–7.

[84] 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," in *Proc. 2020 IEEE Des., Automat. Test Europe Conf. Exhib. (DATE)*, 2020, pp. 1438–1443.

[85] Y. Chen and J. Zhang, "How energy supports our brain to yield consciousness: Insights from neuroimaging based on the neuroenergetics hypothesis," *Front. Syst. Neurosci.*, vol. 15, 2021, Art. no. 648860.

[86] J. Hsu, "IBM's new brain [news]," *IEEE Spectr.*, vol. 51, no. 10, pp. 17–19, Oct. 2014.

[87] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "Activation functions in deep learning: A comprehensive survey and benchmark," *Neurocomputing*, vol. 503, pp. 92–108, 2022.

[88] W. Teka, T. M. Marinov, and F. Santamaria, "Neuronal spike timing adaptation described with a fractional leaky integrate-and-fire model," *PLoS Comput. Biol.*, vol. 10, no. 3, 2014, Art. no. e1003526.

[89] C. Feng et al., "A compact butterfly-style silicon photonic–electronic neural chip for hardware-efficient deep learning," *ACS Photon.*, vol. 9, no. 12, pp. 3906–3916, 2022.

[90] N. C. Harris et al., "Efficient, compact and low loss thermo-optic phase shifter in silicon," *Opt. Exp.*, vol. 22, no. 9, pp. 10487–10493, 2014.
[91] S. Manipatruni, K. Preston, L. Chen, and M. Lipson, "Ultra-low voltage, ultra-small mode volume silicon microring modulator," *Opt. Exp.*, vol. 18, no. 17, pp. 18235–18242, 2010.
[92] W. M. Green, M. J. Rooks, L. Sekaric, and Y. A. Vlasov, "Ultra-compact, low RF power, 10 gb/s silicon mach-zehnder modulator," *Opt. Exp.*, vol. 15, no. 25, pp. 17106–17113, 2007.
[93] B. R. Moss et al., "A 1.23 pj/b 2.5 gb/s monolithically integrated optical carrier-injection ring modulator and all-digital driver circuit in commercial 45 nm SOI," in *Proc. 2013 IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2013, pp. 126–127.
[94] H. Xu, X. Li, X. Xiao, Z. Li, Y. Yu, and J. Yu, "Demonstration and characterization of high-speed silicon depletion-mode Mach–Zehnder modulators," *IEEE J. Sel. Topics Quantum Electron.*, vol. 20, no. 4, Jul./Aug. 2014, Art. no. 3400110.
[95] J. Wang et al., "Optimization and demonstration of a large-bandwidth carrier-depletion silicon optical modulator," *J. Lightw. Technol.*, vol. 31, no. 24, pp. 4119–4125, 2013.
[96] W. Zhang et al., "Harnessing plasma absorption in silicon mos ring modulators," *Nature Photon.*, vol. 17, no. 3, pp. 273–279, 2023.
[97] J.-K. Park, S. Takagi, and M. Takenaka, "Ingaasp mach–zehnder interferometer optical modulator monolithically integrated with ingaas driver MOSFET on a III-V CMOS photonics platform," *Opt. Exp.*, vol. 26, no. 4, pp. 4842–4852, 2018.
[98] T. Hiraki et al., "Heterogeneously integrated III–V/SI MOS capacitor Mach–Zehnder modulator," *Nature Photon.*, vol. 11, no. 8, pp. 482–485, 2017.
[99] C. Wang et al., "Integrated lithium niobate electro-optic modulators operating at CMOS-compatible voltages," *Nature*, vol. 562, no. 7725, pp. 101–104, 2018.
[100] X. Zhang, B. Lee, C.-Y. Lin, A. X. Wang, A. Hosseini, and R. T. Chen, "Highly linear broadband optical modulator based on electro-optic polymer," *IEEE Photon. J.*, vol. 4, no. 6, pp. 2214–2228, Dec. 2012.
[101] X. Zhang et al., "High performance optical modulator based on electro-optic polymer filled silicon slot photonic crystal waveguide," *J. Lightw. Technol.*, vol. 34, no. 12, pp. 2941–2951, Jun. 2016.
[102] K. Liu, C. R. Ye, S. Khan, and V. J. Sorger, "Review and perspective on ultrafast wavelength-size electro-optic modulators," *Laser Photon. Rev.*, vol. 9, no. 2, pp. 172–194, 2015.
[103] G. Sinatkas, T. Christopoulos, O. Tsilipakos, and E. E. Kriezis, "Electro-optic modulation in integrated photonics," *J. Appl. Phys.*, vol. 130, no. 1, 2021, Art. no. 010901.
[104] W.-C. Hsu, B. Zhou, and A. X. Wang, "MOS capacitor-driven silicon modulators: A mini review and comparative analysis of modulation efficiency and optical loss," *IEEE J. Sel. Topics Quantum Electron.*, vol. 28, no. 3: Hybrid Integration for Silicon Photonics, May/Jun., 2021, Art. no. 3400211.
[105] C. Xiong et al., "A linear push-pull silicon optical modulator," in *Frontiers in Optics*, Tucson, AZ, USA: Optica Publishing Group, Oct. 2014, pp. FM3A–4.
[106] H. Yamazaki et al., "Optical modulator with a near-linear field response," *J. Lightw. Technol.*, vol. 34, no. 16, pp. 3796–3802, Aug. 2016.
[107] B. Lee, C. Lin, X. Wang, R. T. Chen, J. Luo, and A. K. Jen, "Bias-free electro-optic polymer-based two-section y-branch waveguide modulator with 22 db linearity enhancement," *Opt. Lett.*, vol. 34, no. 21, pp. 3277–3279, 2009.
[108] H. Feng et al., "Ultra-high-linearity integrated lithium niobate electro-optic modulators," *Photon. Res.*, vol. 10, no. 10, pp. 2366–2373, 2022.
[109] R. Chen et al., "Non-volatile electrically programmable integrated photonics with a 5-bit operation," *Nature Commun.*, vol. 14, no. 1, 2023, Art. no. 3465.
[110] L. Chen, K. Preston, S. Manipatruni, and M. Lipson, "Integrated ghz silicon photonic interconnect with micrometer-scale modulators and detectors," *Opt. Exp.*, vol. 17, no. 17, pp. 15248–15256, 2009.
[111] V. Sorianello et al., "Graphene–silicon phase modulators with gigahertz bandwidth," *Nature Photon.*, vol. 12, no. 1, pp. 40–44, 2018.
[112] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, "Experimental realization of any discrete unitary operator," *Phys. Rev. Lett.*, vol. 73, no. 1, 1994, Art. no. 58.
[113] D. A. Miller, "Self-configuring universal linear optical component," *Photon. Res.*, vol. 1, no. 1, pp. 1–15, 2013.
[114] S. Dwivedi et al., "Experimental extraction of effective refractive index and thermo-optic coefficients of silicon-on-insulator waveguides using interferometers," *J. Lightw. Technol.*, vol. 33, no. 21, pp. 4471–4477, Nov. 2015.
[115] G. Mourgias-Alexandris et al., "Noise-resilient and high-speed deep learning with coherent silicon photonics," *Nature Commun.*, vol. 13, no. 1, 2022, Art. no. 5572.
[116] H. Zhang et al., "Efficient on-chip training of optical neural networks using genetic algorithm," *ACS Photon.*, vol. 8, no. 6, pp. 1662–1672, 2021.
[117] A. Cem, S. Yan, Y. Ding, D. Zibar, and F. Da Ros, "Data-driven modeling of Mach-Zehnder interferometer-based optical matrix multipliers," *J. Lightw. Technol.*, vol. 41, no. 16, pp. 5425–5436, Aug. 2023.
[118] V. Van, *Optical Microring Resonators: Theory, Techniques, and Applications*. Boca Raton, FL, USA: CRC Press, 2016.
[119] W. Bogaerts et al., "Silicon microring resonators," *Laser Photon. Rev.*, vol. 6, no. 1, pp. 47–73, 2012.
[120] P. Dong et al., "Wavelength-tunable silicon microring modulator," *Opt. Exp.*, vol. 18, no. 11, pp. 10941–10946, 2010.
[121] A. N. Tait et al., "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 7430.
[122] A. N. Tait et al., "Microring weight banks," *IEEE J. Sel. Topics Quantum Electron.*, vol. 22, no. 6, Nov./Dec., 2016, Art. no. 5900214.
[123] S. Ohno, R. Tang, K. Toprasertpong, S. Takagi, and M. Takenaka, "Si microring resonator crossbar array for on-chip inference and training of the optical neural network," *ACS Photon.*, vol. 9, no. 8, pp. 2614–2622, 2022.
[124] J. Feldmann et al., "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, no. 7840, pp. 52–58, 2021.
[125] S. Ning et al., "A hardware-efficient silicon electronic-photonic chip for optical structured neural networks," *Proc. SPIE*, vol. 12892, pp. 17–20, 2024.
[126] H. Zhu et al., "Lightening-transformer: A dynamically-operated optically-interconnected photonic transformer accelerator," in *Proc. IEEE Int. Symp. High-Perform. Comput. Architecture (HPCA)*, Mar. 2024, pp. 686–703. [Online]. Available: https://arxiv.org/abs/2305.19533
[127] H. Zhu et al., "Space-efficient optical computing with an integrated chip diffractive neural network," *Nature Commun.*, vol. 13, no. 1, 2022, Art. no. 1044.
[128] Z. Wang, L. Chang, F. Wang, T. Li, and T. Gu, "Integrated photonic metasystem for image classifications at telecommunication wavelength," *Nature Commun.*, vol. 13, no. 1, 2022, Art. no. 2131.
[129] J. Gu, H. Zhu, C. Feng, Z. Jiang, R. T. Chen, and D. Z. Pan, "M3ICRO: Machine learning-enabled compact photonic tensor core based on programmable multi-operand multimode interference," *APL Mach. Learn.*, vol. 2, no. 1, 01 2024, Art. no. 016106, doi: 10.1063/5.0170965.
[130] C. Feng et al., "Integrated multi-operand optical neurons for scalable and hardware-efficient deep learning," *Nanophotonics*, vol. 13, no. 12, pp. 2193–2206, 2024.
[131] J. Gu et al., "SqueezeLight: A multi-operand ring-based optical neural network with cross-layer scalability," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 42, no. 3, pp. 807–819, Mar. 2023.
[132] A. Sludds et al., "Delocalized photonic deep learning on the internet's edge," *Science*, vol. 378, no. 6617, pp. 270–276, 2022.
[133] M. Zhang et al., "Tempo: Efficient time-multiplexed dynamic photonic tensor core for edge AI with compact slow-light electro-optic modulator," *J. Appl. Phys.*, vol. 135, no. 22, 2024.
[134] Z. Chen, X. Li, X. Zhu, H. Liu, H. Tong, and X. Miao, "Full-analog implementation of activation function based on phase-change memory for artificial neural networks," *IEEE Trans. Ind. Electron.*, vol. 71, no. 8, pp. 9914–9922, Aug. 2024.
[135] M. Vatalaro et al., "A low-voltage, low-power reconfigurable current-mode softmax circuit for analog neural networks," *Electron.*, vol. 10, no. 9, 2021, Art. no. 1004.
[136] I. A. Williamson, T. W. Hughes, M. Minkov, B. Bartlett, S. Pai, and S. Fan, "Reprogrammable electro-optic nonlinear activation functions for optical neural networks," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 1, Jan./Feb. 2020, Art no. 7700412.
[137] M. M. P. Fard et al., "Experimental realization of arbitrary activation functions for optical neural networks," *Opt. Exp.*, vol. 28, no. 8, pp. 12138–12148, 2020.
[138] 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, 2022.
[139] A. N. Tait et al., "Silicon photonic modulator neuron," *Phys. Rev. Appl.*, vol. 11, no. 6, 2019, Art. no. 064043.

[140] Y. Shi et al., "Nonlinear germanium-silicon photodiode for activation and monitoring in photonic neuromorphic networks," *Nature Commun.*, vol. 13, no. 1, 2022, Art. no. 6048.

[141] K. Kravtsov, M. P. Fok, D. Rosenbluth, and P. R. Prucnal, "Ultrafast all-optical implementation of a leaky integrate-and-fire neuron," *Opt. Exp.*, vol. 19, no. 3, pp. 2133–2147, 2011.

[142] M. A. Nahmias, B. J. Shastri, A. N. Tait, and P. R. Prucnal, "A leaky integrate-and-fire laser neuron for ultrafast cognitive computing," *IEEE J. Sel. Topics Quantum Electron.*, vol. 19, no. 5, Sep./Oct. 2013, Art. no. 1800212.

[143] Y. Shi et al., "Photonic integrated spiking neuron chip based on a self-pulsating DFB laser with a saturable absorber," *Photon. Res.*, vol. 11, no. 8, pp. 1382–1389, 2023.

[144] B. Romeira, J. Javaloyes, C. N. Ironside, J. M. Figueiredo, S. Balle, and O. Piro, "Excitability and optical pulse generation in semiconductor lasers driven by resonant tunneling diode photo-detectors," *Opt. Exp.*, vol. 21, no. 18, pp. 20931–20940, 2013.

[145] Z. Zhao, J. Gu, Z. Ying, C. Feng, R. T. Chen, and D. Z. Pan, "Design technology for scalable and robust photonic integrated circuits," in *Proc. 2019 IEEE/ACM Int. Conf. Comput.-Aided Des. (ICCAD)*, 2019, pp. 1–7.

[146] J. Gu, Z. Zhao, C. Feng, H. Zhu, R. T. Chen, and D. Z. Pan, "ROQ: A noise-aware quantization scheme towards robust optical neural networks with low-bit controls," in *Proc. 2020 IEEE Des., Automat. Test Europe Conf. Exhib. (DATE)*, 2020, pp. 1586–1589.

[147] L. G. Wright et al., "Deep physical neural networks trained with backpropagation," *Nature*, vol. 601, no. 7894, pp. 549–555, 2022.

[148] Y. Zhan et al., "Physics-aware analytic-gradient training of photonic neural networks," *Laser Photon. Rev.*, vol. 18, no. 4, 2024, Art. no. 2300445.

[149] J. Gu, H. Zhu, C. Feng, Z. Jiang, R. Chen, and D. Pan, "L2IGHT: Enabling on-chip learning for optical neural networks via efficient in-situ subspace optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 8649–8661.

[150] T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, "Training of photonic neural networks through in situ backpropagation and gradient measurement," *Optica*, vol. 5, no. 7, pp. 864–871, 2018.

[151] S. Pai et al., "Experimentally realized in situ backpropagation for deep learning in photonic neural networks," *Science*, vol. 380, no. 6643, pp. 398–404, 2023.

[152] G. Mourgias-Alexandris, A. Tsakyridis, N. Passalis, A. Tefas, K. Vyrsokinos, and N. Pleros, "An all-optical neuron with sigmoid activation function," *Opt. Exp.*, vol. 27, no. 7, pp. 9620–9630, 2019.

[153] J. Crnjanski, M. Krstić, A. Totović, N. Pleros, and D. Gvozdić, "Adaptive sigmoid-like and prelu activation functions for all-optical perceptron," *Opt. Lett.*, vol. 46, no. 9, pp. 2003–2006, 2021.

[154] J. Xiang, A. Torchy, X. Guo, and Y. Su, "All-optical spiking neuron based on passive microresonator," *J. Lightw. Technol.*, vol. 38, no. 15, pp. 4019–4029, Aug. 2020.

[155] H. Zhou et al., "Photonic matrix multiplication lights up photonic accelerator and beyond," *Light: Sci. Appl.*, vol. 11, no. 1, 2022, Art. no. 30.

[156] K. Shiflett, D. Wright, A. Karanth, and A. Louri, "PIXEL: Photonic neural network accelerator," in *Proc. 2020 IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, 2020, pp. 474–487.

[157] K. Shiflett, A. Karanth, R. Bunescu, and A. Louri, "Albireo: Energy-efficient acceleration of convolutional neural networks via silicon photonics," in *Proc. 2021 ACM/IEEE 48th Annu. Int. Symp. Comput. Architecture (ISCA)*, 2021, pp. 860–873.

[158] S. Li, H. Yang, C. W. Wong, V. J. Sorger, and P. Gupta, "PhotoFourier: A photonic joint transform correlator-based neural network accelerator," in *Proc. 2023 IEEE Int. Symp. High- Perform. Comput. Archit. (HPCA)*, 2023, pp. 15–28.

[159] C. Demirkiran et al., "An electro-photonic system for accelerating deep neural networks," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 19, no. 4, pp. 1–31, 2023.

[160] C. Ramey, "Silicon photonics for artificial intelligence acceleration: Hotchips 32," in *Proc. 2020 IEEE Hot Chips 32 Symp. (HCS)*, IEEE Computer Society, 2020, pp. 1–26.

[161] S. Lam et al., "Dynamic electro-optic analog memory for neuromorphic photonic computing," 2024, *arXiv:2401.16515*.

[162] S. Chetlur et al., "cuDNN: Efficient primitives for deep learning," 2014, *arXiv:1410.0759*.

[163] H. Bagherian, S. Skirlo, Y. Shen, H. Meng, V. Ceperic, and M. Soljacic, "On-chip optical convolutional neural networks," 2018, *arXiv:1808.03303*.

[164] Z. Zhu, A. Fardoost, F. G. Vanani, A. B. Klein, G. Li, and S. S. Pang, "Coherent general-purpose photonic matrix processor," *ACS Photon.*, vol. 11, no. 3, pp. 1189–1196, 2024.

[165] J. Gu, Z. Zhao, C. Feng, M. Liu, R. T. Chen, and D. Z. Pan, "Towards area-efficient optical neural networks: An fft-based architecture," in *Proc. 2020 IEEE 25th Asia South Pacific Des. Automat. Conf. (ASP-DAC)*, 2020, pp. 476–481.

[166] X. Xiao, M. B. On, T. Van Vaerenbergh, D. Liang, R. G. Beausoleil, and S. Yoo, "Large-scale and energy-efficient tensorized optical neural networks on III–V-on-silicon MOSCAP platform," *Apl Photon.*, vol. 6, no. 12, 2021, Art. no. 126107.

[167] I. V. Oseledets, "Tensor-train decomposition," *SIAM J. Sci. Comput.*, vol. 33, no. 5, pp. 2295–2317, 2011.

[168] C. Ding et al., "CirCNN: Accelerating and compressing deep neural networks using block-circulant weight matrices," in *Proc. 50th Annu. IEEE/ACM Int. Symp. Microarchitecture*, 2017, pp. 395–408.

[169] J. Gu et al., "ADEPT: Automatic differentiable design of photonic tensor cores," in *Proc. 59th ACM/IEEE Des. Automat. Conf.*, 2022, pp. 937–942.

[170] M. Anderson, S. Ma, T. Wang, L. Wright, and P. McMahon, "Optical transformers," *Trans. Mach. Learn. Res.*, 2024.

[171] H. Zhu et al., "Fuse and mix: Macam-enabled analog activation for energy-efficient neural acceleration," in *Proc. 41st IEEE/ACM Int. Conf. Comput.-Aided Des.*, 2022, pp. 1–9.

[172] B. Murmann, "ADC performance survey 1997-2024," 2024. [Online]. Available: https://github.com/bmurmann/ADC-survey

[173] Z. Zhou et al., "Prospects and applications of on-chip lasers," *Elight*, vol. 3, no. 1, pp. 1–25, 2023.

[174] M. Miscuglio and V. J. Sorger, "Photonic tensor cores for machine learning," *Appl. Phys. Rev.*, vol. 7, no. 3, 2020, Art. no. 031404.

[175] M. Kirtas et al., "Mixed-precision quantization-aware training for photonic neural networks," *Neural Comput. Appl.*, vol. 35, no. 29, pp. 21361–21379, 2023.

[176] C. Demirkiran, G. Yang, D. Bunandar, and A. Joshi, "Accelerating DNN training with photonics: A residue number system-based design," 2023, *arXiv:2311.17323*.

[177] 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. Topics Quantum Electron.*, vol. 26, no. 1, Jan./Feb. 2020, Art. no. 7701518.

[178] J. Choquette, W. Gandhi, O. Giroux, N. Stam, and R. Krashinsky, "NVIDIA A100 tensor core GPU: Performance and innovation," *IEEE Micro*, vol. 41, no. 2, pp. 29–35, Mar./Apr. 2021.

[179] H. Zhu et al., "Elight: Enabling efficient photonic in-memory neurocomputing with life enhancement," in *Proc. 2022 IEEE 27th Asia South Pacific Des. Automat. Conf. (ASP-DAC)*, 2022, pp. 332–338.

[180] Y. Zhu, M. Liu, L. Xu, L. Wang, X. Xiao, and S. Yu, "Multi-wavelength parallel training and quantization-aware tuning for WDM-based optical convolutional neural networks considering wavelength-relative deviations," in *Proc. 28th Asia South Pacific Des. Automat. Conf.*, 2023, pp. 384–389.

[181] D. Zhang et al., "Training and inference of optical neural networks with noise and low-bits control," *Appl. Sci.*, vol. 11, no. 8, 2021, Art. no. 3692.

[182] A. S. Rekhi et al., "Analog/mixed-signal hardware error modeling for deep learning inference," in *Proc. 56th Annu. Des. Automat. Conf.* 2019, 2019, pp. 1–6.

[183] T. Andrulis, J. S. Emer, and V. Sze, "RAELLA: Reforming the arithmetic for efficient, low-resolution, and low-loss analog PIM: No retraining required!," in *Proc. 50th Annu. Int. Symp. Comput. Architecture*, 2023, pp. 1–16.

[184] M. Y.-S. Fang, S. Manipatruni, C. Wierzynski, A. Khosrowshahi, and M. R. DeWeese, "Design of optical neural networks with component imprecisions," *Opt. Exp.*, vol. 27, no. 10, pp. 14009–14029, 2019.

[185] Y. Zhu et al., "Countering variations and thermal effects for accurate optical neural networks," in *Proc. 39th Int. Conf. Comput.-Aided Des.*, 2020, pp. 1–7.

[186] J. Gu, Z. Zhao, C. Feng, Z. Ying, R. T. Chen, and D. Z. Pan, "O2NN: Optical neural networks with differential detection-enabled optical operands," in *Proc. 2021 IEEE Des., Automat. Test Europe Conf. Exhib.*, 2021, pp. 1062–1067.

[187] F. Sunny, A. Mirza, M. Nikdast, and S. Pasricha, "CrossLight: A cross-layer optimized silicon photonic neural network accelerator," in *Proc. 2021 IEEE 58th ACM/IEEE Des. Automat. Conf. (DAC)*, 2021, pp. 1069–1074.

[188] A. Mirza, F. Sunny, P. Walsh, K. Hassan, S. Pasricha, and M. Nikdast, "Silicon photonic microring resonators: A comprehensive design-space exploration and optimization under fabrication-process variations," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 41, no. 10, pp. 3359–3372, Oct. 2022.

[189] H. Zhou et al., "Chip-scale optical matrix computation for PageRank algorithm," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 2, Mar./Apr. 2020, Art. no. 8300910.

[190] M. J. Filipovich et al., "Silicon photonic architecture for training deep neural networks with direct feedback alignment," *Optica*, vol. 9, no. 12, pp. 1323–1332, 2022.

[191] S. Bandyopadhyay et al., "Single chip photonic deep neural network with accelerated training," 2022, *arXiv:2208.01623*.

[192] J. Gu, Z. Zhao, C. Feng, W. Li, R. T. Chen, and D. Z. Pan, "Flops: Efficient on-chip learning for optical neural networks through stochastic zeroth-order optimization," in *Proc. 2020 57th ACM/IEEE Des. Automat. Conf. (DAC)*, 2020, pp. 1–6.

[193] J. Gu, C. Feng, Z. Zhao, Z. Ying, R. T. Chen, and D. Z. Pan, "Efficient on-chip learning for optical neural networks through power-aware sparse zeroth-order optimization," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 7583–7591.

[194] A. Nøkland, "Direct feedback alignment provides learning in deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016.

[195] J. Launay, I. Poli, F. Boniface, and F. Krzakala, "Direct feedback alignment scales to modern deep learning tasks and architectures," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 9346–9360.

[196] J. Meng, M. Miscuglio, J. K. George, A. Babakhani, and V. J. Sorger, "Electronic bottleneck suppression in next-generation networks with integrated photonic digital-to-analog converters," *Adv. Photon. Res.*, vol. 2, no. 2, 2021, Art. no. 2000033.

[197] Lightmatter, "Passage: A wafer-scale, programmable photonic interconnect," 2024. Accessed: Jun. 02, 2024. [Online]. Available: https://lightmatter.co/products/passage/

[198] S. Sahni et al., "Beyond the beachfront: Integration of silicon photonic I/Os under a high-power ASIC," in *Proc. 2024 IEEE Opt. Fiber Commun. Conf. Exhib. (OFC)*, 2024, pp. 1–3.

[199] G. Tang, C. Li, X. Zhang, and D. M. Rhee, "Thermal management solutions and design guidelines for silicon based photonic integrated modules," in *Proc. 2015 IEEE 17th Electron. Packag. Technol. Conf. (EPTC)*, 2015, pp. 1–6.

[200] G. Refai-Ahmed et al., "Lidless and lidded flip chip packages for advanced applications," in *Proc. 2020 IEEE 22nd Electron. Packag. Technol. Conf. (EPTC)*, 2020, pp. 104–111.

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

[202] V. Stojanović et al., "Monolithic silicon-photonic platforms in state-of-the-art CMOS SOI processes," *Opt. Exp.*, vol. 26, no. 10, pp. 13106–13121, 2018.

[203] K. Giewont et al., "300-mm monolithic silicon photonics foundry technology," *IEEE J. Sel. Topics Quantum Electron.*, vol. 25, no. 5, pp. 1–11, 2019.

[204] J. Gu et al., "Neurolight: A physics-agnostic neural operator enabling parametric photonic device simulation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 14623–14636.

[205] T. Ferreira De Lima et al., "Primer on silicon neuromorphic photonic processors: Architecture and compiler," *Nanophotonics*, vol. 9, no. 13, pp. 4055–4073, 2020.