

Integrated Photonic Tensor Processing Unit for a Matrix Multiply: A Review

Nicola Peserico , Bhavin J. Shastri, *Senior Member, IEEE*, and Volker J. Sorger , *Senior Member, IEEE*

(Invited Paper)

Abstract—The explosion of artificial intelligence and machine-learning algorithms, connected to the exponential growth of the exchanged data, is driving a search for novel application-specific hardware accelerators. Among the many, the photonics field appears to be in the perfect spotlight for this global data explosion, thanks to its almost infinite bandwidth capacity associated with limited energy consumption. In this review, we will overview the major advantages that photonics has over electronics for hardware accelerators, followed by a comparison between the major architectures implemented on Photonics Integrated Circuits (PIC) for both the linear and nonlinear parts of Neural Networks. By the end, we will highlight the main driving forces for the next generation of photonic accelerators, as well as the main limits that must be overcome.

Index Terms—Matrix-vector multiplication, photonics, PICs, silicon photonics, tensor core.

I. INTRODUCTION

THE latest decade has seen the exponential growth of Machine Learning (ML) as one of the main branches of the Artificial Intelligence field [1], [2]. At the core of this branch, there is the assumption that a machine can learn to perform any task if a training algorithm is applied. While historically the concept of ML can be tracked back from the '50s [3], [4], just in recent decades the concept has started to attract more and more interest [5], thanks to the improvement of the mathematical approaches (such as back-propagation [6]), and computation capabilities, that allowed to run complex ML algorithms.

To implement ML applications, several algorithms and circuits have been proposed [7]. One approach relies on mimicking the human brain structure, which has led to several implementations, where Neural Networks (NNs) have become the most

popular (Fig. 1), thanks to its flexibility and scalability [8], [9]. A NN is formed by a sequence of interconnected layers of neurons, whose inputs are the output of all the neurons of the previous layer (Fig. 1(a)). The output of a single neuron is the result of the scaled linear summation of the input passed by an activation (nonlinear) function (Fig. 1(b)). In this framework, a whole layer can be seen as matrix multiplication, followed by the activation function, allowing for a more straightforward implementation on hardware. The values used to scale the inputs (the W matrix) are the learning parameters that the NN needs to compute using the selected method (i.e. back-propagation). By so, for each NN, we can see two separate steps: the training one, where all the parameters are computed using training algorithms and dataset, and the second one, called inference or classification, where the NN is used over a novel set of data input. Research on NN has brought other implementations for each layer, based on the application and/or input. For example, convolution layers are widely used in the image and video context, where a certain trainable filter is applied to a portion of a 2D input [10]. More and more complex tasks can be performed by NN by adding more and more layers implementing Deep Neural Networks (DNNs) for Deep Learning.

After the initial creation of the ML concept, followed by a winter phase due to the lack of hardware [4], ML has raised again following the exponential increase of computer performance, creating an environment where DNN can have tens of layers and millions of parameters. One example that has shown all the potential of this approach is called DALL · E2, one of the most advanced text-to-image DNN, with over 3.5 billion parameters [11].

Such large and extended networks raise an enormous demand in terms of computational power [12], challenging current hardware technologies in terms of operation per second, latency, and power consumption. The flexibility and scalability of digital electronics have allowed the creation of a framework where NNs can be coded, tested, and used [13]. As the NN became larger and larger, the digital approach started to look for novel solutions to keep pace and deliver enough performance levels to run the NN [14]. Those solutions are based on scaling, by using interconnected hardware in data centers, or by architecture changes, for example moving from generic CPU to application- or numerical- specific ones, such as FPGA, GPU, or ASIC, called Tensor Core [15], [16], [17]. However, some of the limitations still exist, due to more physical reasons, such as energy

Manuscript received 29 November 2022; revised 6 April 2023; accepted 13 April 2023. Date of publication 2 May 2023; date of current version 27 June 2023. This work was supported by AFOSR under Grant FA9550-22-1-0100. The work of Volker J. Sorger was supported by the PECASE Award through the AFOSR under Grant FAA9550-20-1-0193. (Corresponding author: Volker J. Sorger.)

Nicola Peserico is with the Department of Electrical and Computer Engineering, George Washington University, Washington, DC 20052 USA (e-mail: npeserico@gwu.edu).

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

Volker J. Sorger is with the Department of Electrical and Computer Engineering, George Washington University, Washington, DC 20052 USA, and also with the Optelligence LLC, Wilmington, DE 19801 USA (e-mail: sorger@gwu.edu).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/JLT.2023.3269957>.

Digital Object Identifier 10.1109/JLT.2023.3269957

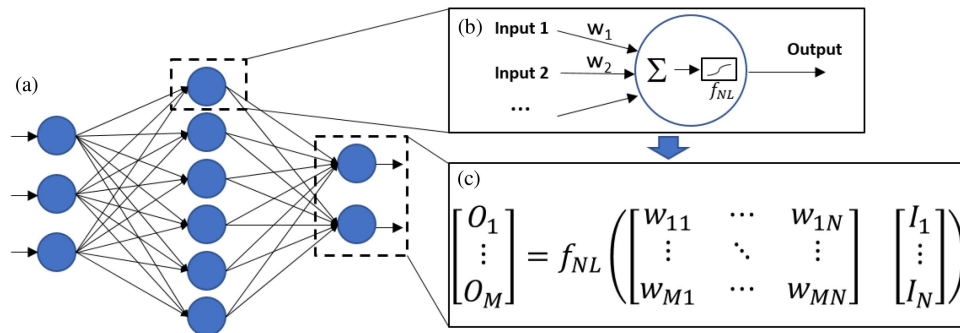


Fig. 1. Breaking down of a Fully-Connected Neural Network. (a) Example of a NN having one hidden layer. (b) Every single neuron receives the input signals from all the previous layer neurons, scaled by a factor w , performing their summation and passing through an activation function. (c) This single neuron can be generalized by including the whole layer, employing a matrix representation.

consumption and latency [18]. For these reasons, research has started to look for novel technologies that can provide a better hardware accelerator for NNs. Optics (and photonics) have been raised as an alternative approach for hardware implementation of NN, thanks to its speed-of-light latency and low energy consumption [19], [20], [21], [22], [23]. Moreover, Silicon Photonics has started to become a reliable and diffuse technology, allowing the implementation of Photonic Neural Network (PNN) hardware accelerator at the chip scale, to better fit the needs of final users [24], [25], [26].

In this paper, we will review why and how silicon photonics chips have addressed the challenge of providing a hardware accelerator for PNN. After an initial part on electronics limitations and photons potential in this field, we will look into the main implementations of Photonic Tensor Core (PTC), either based on coherent interference or WDM/MDM approaches. We will address the limitations and scalability of such solutions, focusing on the most challenging part related to the activation function. We will conclude with a discussion of what the near and long-term future look like for such PNNs.

II. ELECTRONICS VS. PHOTONICS

Digital electronics has been the hardware foundation that allowed the growth of NNs since it can provide flexibility, scalability, and fast delivery times. Even if the Von Neumann architecture is not the best one for NN applications [27], it has provided the right framework to develop NNs in their early stage. Moreover, the diffusion of programming languages for software development, and the following NN-specific libraries, has permitted the spread of NN applications since the '90 [5]. The continuous improvement in computer performances (in terms of processors, memory, and network) thanks to the development of smaller and more dense CMOS transistors [28], has permitted to keep pace with the increasing complexity of NNs.

However, in the last decade, the complexity, layer density, and datasets size have evolved to a scale that a single CPU cannot handle, neither for inference nor training [29]. The main limitations come from the size of the NN, which could require millions of parameters, and so the memory size and throughput become important bottlenecks, as well as the limited capability of CPU to perform float multiplication and summation, that are required for every neural layer, as shown

before. All these aspects have pushed also the energy consumption related to the NN [30], for both training and inference, posing an additional challenge from the hardware perspective.

To overcome such limitations, several paths and solutions have been explored and adopted, from both software and hardware sides. From the software and theoretical side, several strategies and optimizations have been proposed. For example, model compression allows the reduction of the number and size of weights, and by so reducing the need to transfer them from the processing unit to the memory and vice versa [31], [32], [33]. Many studies have shown how the whole system's power consumption can be easily dominated by the access cost per bit to off-chip DRAM memory [34]. Some of these strategies include weight quantization [35], connection pruning [36], low rank approximation [37], and low bit weights [38]. From the hardware side, there have been two main shifts: the first takes advantage of the computation parallelization, and the second push for more application-specific hardware, in particular on the math unit. By using multiple systems in a balanced scheme, it is possible to parallelize the layer computation over different systems, and so assure a more high throughput, even for DNN [39]. Today's market presents many data centers and cloud services that provide these types of schemes, from Google Cloud to Amazon Web Service [40]. The diffusion and expansion of those data centers have reached a threshold regarding their power consumption pace rate [41], [42]. The second approach works directly on the hardware optimization connected to the computation part of the NN [43]. Since CPUs provide a limited amount of resources for math computation, NNs have moved toward GPUs, which provide faster and more specific hardware to perform float multiplication and accumulation (MAC), as a key task for each NN. The main acceleration of GPUs over CPUs is an increased number of ALU (Arithmetic Logic Unit) cores to parallelize MAC operations, roughly 1000 vs. 10, respectively. Following this trend, the use of ASIC and Tensor Processing Units (TPUs) has grown in recent years, where the actual hardware can implement the required tasks in a heavily optimized fashion as they are written in the electronic architecture [15], [16], [17]. TPUs continued the GPU push, reaching about 32,000 cores, but also added reduced memory access by deploying an systolic array, which uses an approach of featuring an array thus processing once input vector at the same time [44]. Examples of

ASIC can be found in many companies, such as Nvidia, Intel, and Tesla [45], [46], [47].

Even with those optimizations, digital electronic presents important limitations for NN implementations. For example, speed is always limited by the clock cycles and transistors' energy consumption, as it has been for CPUs, capping the clock to a few GHz (1–3 GHz). Moreover, the latency in the computation can be dominant, since float MAC operations still require several cycles to be performed. For applications where timing and energy consumption are a concern, such as autonomous driving for small drones, those limitations pose complex challenges to the NN engineers.

Optics and photonics have been raised as one of the possibilities to overcome these limitations [19], [48]. The use of photons instead of electrons allows a virtually infinite bandwidth (> 100 GHz), speed-of-light propagation latency (\sim ns), and almost zero power consumption, thanks to the lack of RC wire charge connected to the propagation of electrons [22], [49]. Silicon Photonics, in particular, is in the right spot to provide the next generation of hardware accelerators for PNNs [23], thanks to the important progress that happened in the last decade [50], [51], [52], such as component density, laser integration, high-speed (> 100 GHz) modulators and photodetectors, and low propagation losses. Other benefits that photonics has over digital MAC accelerators include 1) the ability to perform summation in the analog domain at full bit precision before ADC quantization happens; 2) temporal pooling of data such as for convolution operations by increasing the integration time of the receiver, which also lower ADC requirements; 3) high-level of fan-out via copying data passively; 4) energy-free Fourier transformation via the Fourier Theorem performing a passive FFT by an optical lens [53] (i.e. also on-chip [54]); 5) the possibility to process image or lidar input directly as light signals. As we will see in the next section, several Photonic Integrated Circuits (PICs) have been presented in this field, showing the potential of such Photonic Tensor Cores (PTCs) in real applications.

It has to make clear that photonics brings its challenges too, from the energy cost of moving back and forth from the digital domain (from where data come from) to the analog (the optical) one [55], to the noise management for high bandwidth that limits the bit resolution at the output. Other aspects are related to the architecture implementations, as photons require an electrical system to be controlled and keep operational, making each PIC strongly related to an FPGA/ASIC that must assure its working operations [56], [57].

III. PHOTONICS INTEGRATED CIRCUIT FOR NN: ARCHITECTURES

Several PIC architectures have been proposed over the last years to perform the Tensor Core tasks for PNNs [58], [59], [60]. Considering the main PTC task, the MAC operation benefits from the coherent electromagnetic nature of the light, implying the possibility to perform multiplication by lossless interference, while the accumulation is performed directly on the photodetector once light signals are collected. Moreover, by allowing

manipulation of light employing nanoscale waveguides, PIC can integrate a large number of MAC operations on small scale, employing a high number of inputs, high-speed modulators, and photodetectors.

To perform the MAC function, several different approaches have been proposed during the latest years, varying the basic components elements, as well as the input, the weights, and the output configurations. Those different architectures show different performances, in terms of actual speed (measured as MAC operations per second), footprint, energy consumption, etc [23].

Here, we will review these approaches integrated into PICs, as we focus on the main differences among the architectures. Several figure-of-merits are commonly used to compare different PTC, such as MAC operation per second, or footprint [58], [59]. They come from a system-level perspective, and are easily comparable among different architectures, even across different domains. However, for the photonics field, they mainly depend on both the technologies used for modulators (for input vectors) or the photodiodes (for output vectors) used in each implementation, which follow the possibilities given by the foundries and rarely are due to architecture choices [61], [62], [63]. Following that, it is more interesting to focus on common limitations, such as the number of controllers that each circuit requires, the footprint scaling, and the possibility to implement nonvolatile memory elements, such as Photonic RAM (P-RAM) components using Phase Change Materials (PCMs) [64], [65], to further reduce energy consumption. Those figure-of-merits better describe the differences between different circuits, showing that trade-offs must be addressed to evolve into this field.

To start the review, we first divided the PIC into two main categories, based on the mathematical approaches for the MAC operation: the first one relies on the singularization, where the main matrix is divided by the meaning of singular value decomposition into 3 matrices; the second approach avoids this decomposition, by implementing schemes that directly reflect the main matrix.

A. $Y = (V^t \Sigma U) X$

One type of PICs exploits the single value decomposition (SVD) of matrices where the main weight matrix is divided into 3 matrices, that can be directly implemented by using cascaded Mach-Zehnder Interferometers (MZIs). This approach has its root in a work by Reck et al. in 1994 [66], where they describe a simple algorithm for the realization of any $N \times N$ unitary matrix multiplication. By using the SVD, the external matrices V and U are unitary matrices, so the implementation can be straightforward by using interconnected MZIs, while the diagonal matrix Σ can be implemented by a series of attenuators, usually implemented by MZIs too. A more complete description and discussion were later provided by Miller et al. in 2013 [67]. Some examples of this architecture are shown in Fig. 2

The first experimental implementations were presented for quantum optics, by Carolan et al. [74], where 15 MZI were integrated into one single silicon photonic chip. The work was followed by Riberio et al., demonstrating a 4×4 -port universal

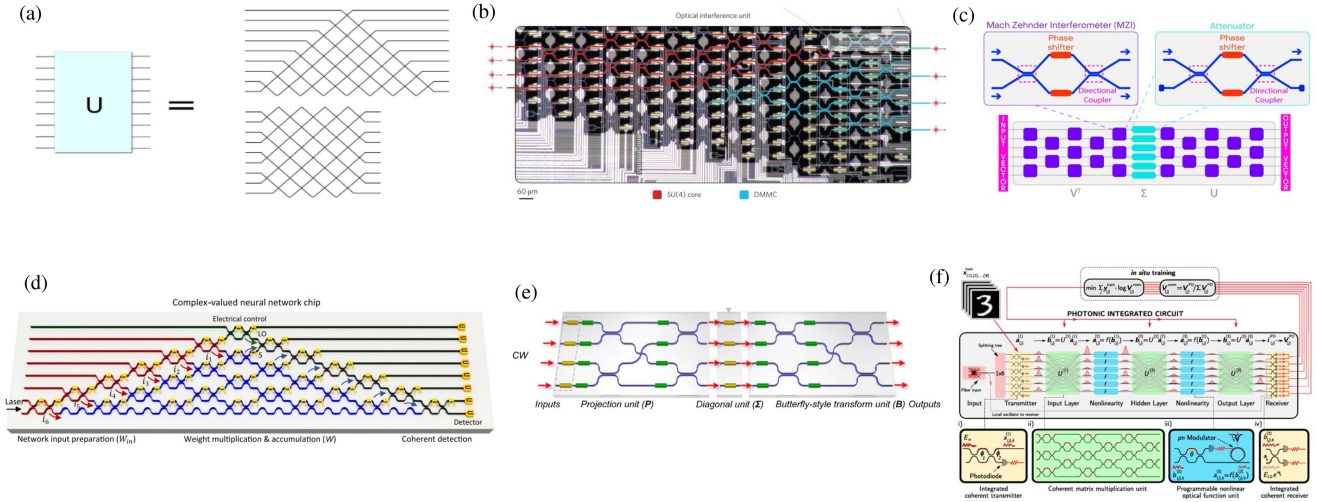


Fig. 2. Examples of Photonic Integrated Circuit for Neural Networks, using Mach-Zehnder Interferometers mesh. (a) Comparison between the original Reck mesh and the improvement proposed by Clements et al. [68]. (b) Photo of the Optical Interference Unit proposed by Shen et al. [69]. (c) Similar architecture, showing the actual SVD with central attenuators line, proposed by Demirkiran et al. [70]. (d) Reck mesh implementing complex values for Neural Networks, done by Zhang et al. [71]. (e) Butterfly solution, exploiting pruning, presented by Feng et al. [72]. (f) Full integrated Neural Network, using MZIs mesh and integrated activation functions, from Bandyopadhyay et al. [73].

linear circuit [75], and by Annoni et al., presenting a mode demultiplexer with the same MZIs architecture [76].

A theoretical discussion was presented by Clements et al. in 2016 on the MZI layout [68], shown in Fig. 2(a). The paper shows a way to implement the same MZI mesh network more compactly, allowing to reduce of the insertion loss due to the shorter path length, without any mathematical limitation in the starting unitary U matrix. To be noticed, this novel approach reduces the length of the device but does not reduce the number of components required.

The first implementation of the MZM mesh as a PTC device for NN comes from Shen et al. in 2017 [69] (Fig. 2(b)). The MZI mesh was used as part of an Optical Neural Network (ONN) in a Deep Learning scheme performing vowel recognition. The chip integrated 56 MZIs, showing good accuracy data and opening the path for more ONN as a way to improve energy efficiency and computational speed. From part of this work, a spin-off company was created and recently started to show its architecture [70] (Fig. 2(c)). In this case the silicon photonic chip has the same MZI mesh approach, but it integrates directly all the 3 matrices of the SVD, together with integrated photodetectors. While the first work was based on heaters to control the MZIs, this later one leveraged MEMS, providing a $\times 100$ higher speed. This latter work shows 8-bit precision and the clear leverage that photonics can provide to AI accelerators in terms of energy efficiency per operation (up to 7,500 Inference-per-Second IPS/W). A step forward was been done by Zhang et al. as they implemented a PNN with complex values, using the original Reck MZI scheme [71] (Fig. 2(d)).

While all these implementations allow having a full matrix, and so to implement a fully connected neural layer, a recent trend following the electronic approach is exploring pruning as a technique to reduce the number of connections between layers. One example in the photonic field has been presented by Feng et al. [72], shown in Fig. 2(e). In this case, the matrices V and

U are substituted with projection and transform units, that have a large reduction of the number of MZI [33]. The authors show that, despite the reduction in the number of MZI, the PNN was capable to perform digit recognition over MNIST dataset with an accuracy of over 94%.

The latest implementation that we present in this overview comes from Bandyopadhyay et al. where they present a full Neural Network chip [73] (Fig. 2(f)). The chip presents input modulators to encode the input, 3 matrix multiplication unit using the MZI mesh, interleaved by 2 nonlinear layers. The nonlinear function will be discussed in a later section. Even in this complex chip, it is possible to perform in-situ training, showing how a silicon photonic chip can cover all the tasks required by a NN. The authors use heaters to control the MZIs but provide alternatives for future higher-speed implementations.

The use of MZI mesh comes with several advantages, like the ideality of the MZI response (even without perfect components [77]), the coherent scheme that requires just one single laser, and the speed of reconfigurability allowed by the pull-down p-n junction configuration of the MZI (reaching GHz bandwidth), or MEMS (reaching sub- μs speed [78]). Thanks to the reliability of the configuration and the single laser source, this approach already showed promising results and startups hit the market with solutions based on it. Moreover, even the bit resolution achieved takes advantage of this advanced state-of-the-art, reaching a high bit resolution, up to 10 b.

On the other side, this configuration comes with some limits, mainly due to the higher complexity behind SVD and the footprint required to fulfill this operation. Dividing the matrix requires a pre-computational step, as well as more components integrated into the PIC, increasing the complexity of the whole architecture.

In terms of component scaling and technologies, the MZM can present limitations and opportunities [60]. In the Reck scheme, the number of MZI needed to implement one of the

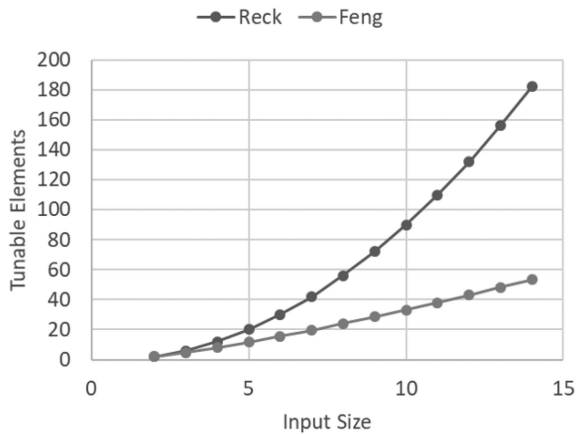


Fig. 3. Scaling comparison between the original Reck scheme [66], and the one proposed by Feng et al. [72], that uses pruning to reduce the connections [33].

two unitary matrices is $N(N-1)$, where N is the number of inputs, resulting in a scaling law of $O(N^2)$. In particular, for each MZI 2 phase controls are needed (one for one input, and one for one of the arms). By pruning, the MZI required can be reduced to $N \log_2(N)$, under certain conditions, resulting in an important reduction of the controllers needed, as shown in Fig. 3. However, to use the scheme proposed by Feng et al., the number of inputs should be a power of 2, or the optical power unbalanced must be addressed with more MZIs.

Some of the downsides of the approach using MZI mesh can be identified in the single MZI element. For example, MZI requires precise control of the phase of each path, making the phase actuator a key element in the performances, as well as being sensitive to the fabrication variability on each waveguide. Several groups analyzed the actual errors and noise due to the phase mismatch to better calibrate the impact on the NN. On the other side, some groups implemented on-chip training, forcing the same NN to calibrate itself on these errors [79], [80], [81].

Other limitations that come from the use of the MZI are the lack of parallelism and P-RAM elements. First, by using MZI, the calibration is wavelength dependent, making more challenging the implementation of a WDM-based scheme on the same MZI mesh. This lack of parallelism could limit the possibility of the architecture, relying just on the speed of the input modulators and output photodetectors. The second element is the complex integration of P-RAM components in the mesh. Those components are one of the keys for an energy-efficient PNN chip, as the PCM material they are based on, can store the weight values in a non-volatile fashion, reducing further the operation-over-energy figure of merit [82]. However, most of the PCM materials have an impact on both amplitude and phase, making the control of one MZI more challenging. Moreover, due to the bi-level nature of the PCM, multiple strips might be required to match the offset due to fabrication phase mismatch.

B. $Y = MX$

Another approach to performing the matrix multiplication is the direct mapping of the M matrix into the PIC, by exploiting one of the degrees of freedom that photonics has,

such as wavelength, modes, or polarization. The most common is Wavelength-division multiplexing (WDM), where different scaled wavelength sources are combined to obtain an equivalent dot product at the photodetectors.

Initial architectures come from the optical computing field, where several researchers were emulating the digital logic functions of electronics [90], [91]. The first implementation in a full WDM scheme was presented by Yang et al. in 2012 [83] (Fig. 4(a)). In this work, the matrix values are mapped one-by-one on the microring resonators grid, exploiting the MUX/DEMUX scheme for WDM lasers, where the input vector is encoded into the amplitude of the same lasers. The photodetectors at each output provide the summation of the different wavelength signals. As most of the schemes in this section, mapping the matrix M , the complexity of the circuit scales with the size of the matrix itself, so $O(N^2)$ for a square matrix of size N , but it can support rectangular matrices, as well as branch pruning to reduce the scaling factor.

A step forward was made by Tait et al. in 2014, describing the “broadcast and weight” approach for the optical neural network [85], later implemented in 2017 [92] (Fig. 4(c)), that achieves an efficiency of $180 fJ$ per Synaptic operation. The architecture shows the broadcast of all the input to all the microring resonator weight banks, whose outputs are fed into the input by an amplitude modulator. The weighting is performed by tuning the microring resonators to the input wavelengths, archiving both positive and negative values thanks to the balanced photodetectors. This approach permits obtaining an optical neural network that has been demonstrated useful for many applications [93]. Other implementations have exploited the tunability of add-drop microring resonators as weights to perform the multiplication as attenuation of the incoming light beam [86] (Fig. 4(d)), reaching up to 9 b resolution [94]. The use of microrings allows for an important footprint reduction (using SiPh, microring could be downsized to a $10 \mu m$ radius) while having high-speed reconfiguration thanks to the internal p-i-n junction, that nowadays could reach a bandwidth of more than 25 GHz. Moreover, thanks to the add-drop configuration, the architecture could have both positive and negative sign weights in the matrix, without the need for post-processing to correct the data. The main disadvantage is coming from the control perspective, as microring tends to be a sensitive element towards noise sources, such as temperature variation, stress, and so on. By so, besides the modulation controlling the p-i-n junction, another signal must be applied to the heater to assure a perfect alignment between the microring’s resonance and the laser’s wavelength, doubling the number of controls, and limiting the speed for a fine reconfiguration to few kHz. Moreover, due to this high integration and need for resonance stabilization, integration of P-RAM elements in the ring itself is challenging due to the double $n-k$ impact of the material and the finite number of states, making this architecture not directly suitable for low-energy applications, such as edge computing.

Another approach exploits tunable couplers between rows and columns of an optical waveguide grid, presented by Feldmann et al. in 2021 [84] (Fig. 4(b)). Each wavelength coming from a Comb laser source is modulated and fed into a certain row. The

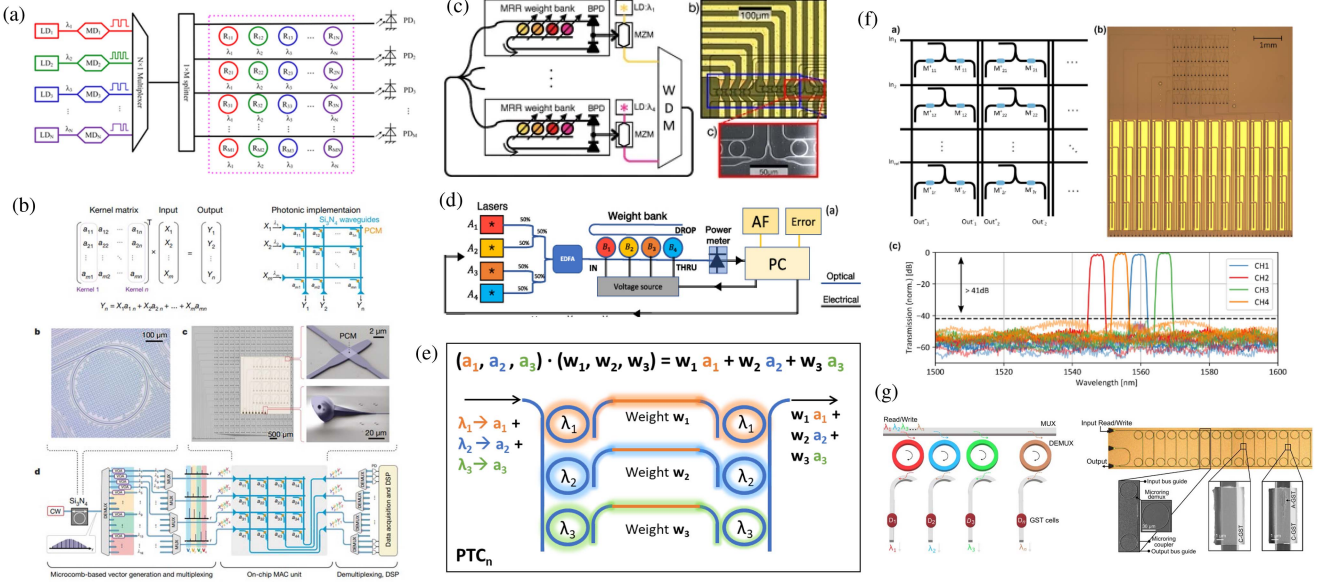


Fig. 4. Architecture and PICs using WDM to implement the Matrix-Vector Multiplication. (a) First architecture proposed by Yang et al. [83], where the one-to-one matrix mapping is clearly visible. (b) Architecture exploiting cross-bar attenuated couplers, presented by Feldmann et al. [84]. (c) First implementation of “broadcast and weight” approach from Tait et al. [85]. (d) Similar “broadcast and weight” approach, that can perform training and testing of a Hopfield network [86]. (e) Implementation of WDM matrix multiplication using add-drop microring resonators, implemented by Ma et al. [87]. (f) Recent implementation of the cross-bar approach, improved by Bragg gratings to reduce the crosstalk [88]. (g) Add-drop PCM microring approach to demonstrate an integrated engine for unsupervised correlation detection [89].

tunable couplers bend a certain amount of the incoming light toward the selected column. The photodiode at the end of the column collects the composition of the different light beams, whose amplitude is determined by the couplers and the P-DRAM element placed after the coupler. This scheme relies on the simplicity of the implementation that reduces the number of controllers to the minimum (equal to the size of the matrix), and implementing them with PCM allows having almost 0 energy cost, but limits the speed of reconfiguration. To improve this figure-of-merit, MEMS actuators can be implemented, at marginal increasing of energy consumption. A further improvement was presented in 2022 [88], where Bragg gratings are used to reduce the crosstalk between channels, and so improving the resolution (Fig. 4(f)).

The last architecture was presented by Miscuglio et al. [21], and later implemented by Ma et al. [87] (Fig. 4(e)). This architecture takes advantage of the add-drop microring as the element to fan-out the WDM inputs and recombines them after attenuation is applied in the waveguide link between them. This approach has the advantage to be able to use PCM (\sim s reconfiguration speed), slow-speed heater-based components (\sim kHz), and high-speed p-i-n junction (\sim GHz) to achieve the required attenuation, by so fulfilling the requirement of both edge computing applications and cloud one. The number of controls could be high in principle (up to 3 controllers for each element of the matrix), however, by relying on the fabrication quality and accepting a reduction of the resolution, the control could be reduced to just an attenuator per element of the matrix. Sarwat et al. used the same approach to demonstrate an integrated engine for unsupervised correlation detection [89], achieving a total energy consumption of 2.50 mJ per 1 Million Inputs (Fig. 4(g)).

Similar architectures can be implemented by exploiting mode or polarization multiplexing or mixing different approaches for

more compact and yet performance implementations. The mapping of the weight matrix in a one-to-one fashion allows to have a higher level of flexibility, and requires less pre-computation, as it does not require any matrix decomposition. However, the scaling factor will follow the size of the matrix itself, posing an important limitation due to the high number of components required, and the control electronics circuits they require, such as DACs, current source, and so on.





IV. ARCHITECTURES COMPARISON

As seen, many different architectures could be used to implement MVM for neural networks, as summarized in Table I. In the table, actual Figure-of-Merit MAC/s nor MAC/J is not reported, as for all the architectures, it will mainly rely on the inputs modulator and output photodiodes, whose characteristics are coming from the fabrication process rather than the component used to perform the MAC operation. However, in case where weights must be updated at the same speed of the inputs, the architecture choice will reduce to the ones that allow an high-speed weight updates (for example using p-n junctions), to respect to slow or large footprint ones.

One parameter that influences the choice of architecture is the chip footprint, based on the size of the component used and the scaling compared to the matrix. The basic $Y = MX$ architecture takes advantage of the more direct equation, as scaling is proportional to the matrix size, while the MZM approach suffers from the decomposition matrices. However, for both approaches, the scaling follows $O(N^2)$, except for the butterfly configuration used by Feng et al. This last one exploits pruning as a way to reduce the number of connections, and so the scaling of the circuit. For the number of controls, the best solutions appear to

TABLE I

SCALING COMPARISON OF VARIOUS APPROACHES TO PERFORMING MVM AND MAC OPERATIONS USING PHOTONIC CHIP-BASED COMPONENTS. N = SIZE OF INPUT VECTOR; M = SIZE OUTPUT VECTOR; P-RAM = PHOTONIC RANDOM ACCESS MEMORY, ALLOWING FOR ZERO-STATIC POWER CONSUMPTION, ONCE THE WEIGHTS ARE SET

Type of Operation	$Y = V^T \Sigma U X$ 		$Y = M X$ 	
# Input	1 Laser, N Modulators	1 Comb Laser, N Modulators	N Lasers, N Modulators	N Lasers, N Modulators
# Outputs	N Photodiodes	M Photodiodes	M or 2*M Photodiodes	M Photodiodes
Area/Basic Element Area	$N^2 - N$ or $N \log_2(N)$	$N \times M$	$N \times M$	$N \times M$
Controllers	$2N^2 - N$ or $N \log_2(N)$	$N \times M$	$2(N \times M)$	$N \times M$
Parallelization	No	WDM Off Chip	WDM On Chip	WDM On Chip
Weight Bit Resolution	8/10	5	9	>5
P-RAM	No	Yes	No	Yes

be the one based on couplers and coupled microrings, even if this last one might be affected by the detuning of the microrings that would limit the Extinction Ratio, and so the bit resolution achievable by the NN. The architecture based on single add-drop microrings could either have the same $M \times N$ controllers if just one tuning method is used (for example employing just heaters as tuning weight), but each microring needs to integrate both a trimming method (i.e. heaters) and a high-speed tuning (i.e. the p-i-n junction) to support high-speed reconfiguration, doubling the number of controls. The lack of need for tuning for the coupler architecture comes at the cost of a more complex input that requires a comb laser, and a WDM mux and demux external to the chip for the output, increasing the complexity of the overall system, but allowing for GHz operations exploiting existing modulators and photodetectors.

Bit resolution shows a strong point for the architectures based on MZM, for mainly two reasons: the more straightforward capability of controlling the phase difference in the MZM, resulting in a larger ER, and so larger bit resolution; the advanced stage of the products based on this technology that already reached the market, so having passed the optimization process. Different types of modulations, for example based on Electro-Absorption Modulators [95] or MEMS, can provide a higher ER in a compact way, allowing a high bit resolution also for other architecture with different speeds (sub-MHz to GHz) and limited energy consumption [78]. Moreover, techniques such as coherent detection have been proposed [94], capable to reach 9-bit resolution with WDM MRR architecture.

The last piece of confrontation is regarding the possibility to implement P-RAM on the circuits [64], [65], [82], [96], [97], by using PCMs for example [98]. In a larger view, as more and more MVM circuits will be used to implement NNs, having the possibility to integrate photonic memory elements would have a crucial benefit in terms of energy efficiency, as it reduces the power needed to tune the weight as well as the energy required to access external memory elements in DRAM [29]. That would allow targeting edge computing applications, rather than just cloud applications in data centers, where power consumption is a priority to extend the lifespan of those devices. Up to our knowledge, just two architectures allow the integration of the PCMs, placing those materials either in the couplers or between coupled rings. The architecture based on MZM could benefit in case a phase-only PCM would be presented, as most of the materials are now affecting adsorption too, such as GSST [99], GSSE [98], or GST [100]. Integration of PCMs into microring

resonator might be challenging for the same reason, adding also a problem of cross-heating interference, as the tuning element could affect the phase of the material, resulting in an unwanted switching.

V. NONLINEARITIES

The last piece to turn a PIC circuit performing MVM into a NN is by providing a nonlinear activation function. In many of the NNs we have seen before, this activation function was performed by a CPU or GPU, once the optical signal is transformed into a digital one. This conversion allows several advantages, like performing mathematical complex functions (including calibration), as well as having the flexibility to change the actual activation function based on the goals of the NN. However, it presents several drawbacks: one is the slow speed associated with this procedure, linked with the long latency, that nullify the major advantages of implementing a photonic neural networks. The power involved is also a major drawback: it has been demonstrated that ADCs are the first contributors in the energy cost of the system, especially for high speed ones [55], [105]. Moreover, to perform the following neural layer, all the digitized signals must be converted back into optical analog domain, requiring additional modulation energy.

To overcome these limitations, and so keep the high bandwidth and low latency provided by the optical domain, many researchers have explored different solutions. One of the major paths is the exploiting of material nonlinearities on-chip, which can be exploited by high-power optical signals under certain conditions. While this path comes from a long tradition of exploring nonlinearities in silicon or silicon nitride waveguide (for generating single photons [106], four-wave mixing [107], or comb generation [108]), the cost of dealing with high power signals limit the possibility of implementation into large and deep optical neural network at the moment.

By so, other architectures have been explored to implement such solutions, that use, completely or partially, an electrical-optical domain change, while keeping the signal in the analog domain. Here, we list the major ones, based on the approach used.

A. Full O/E/O Conversion

The first implementation we present is the complete conversion of the optical signal into an electrical one, that would latter pilots a novel optical signal. One implementation presented by

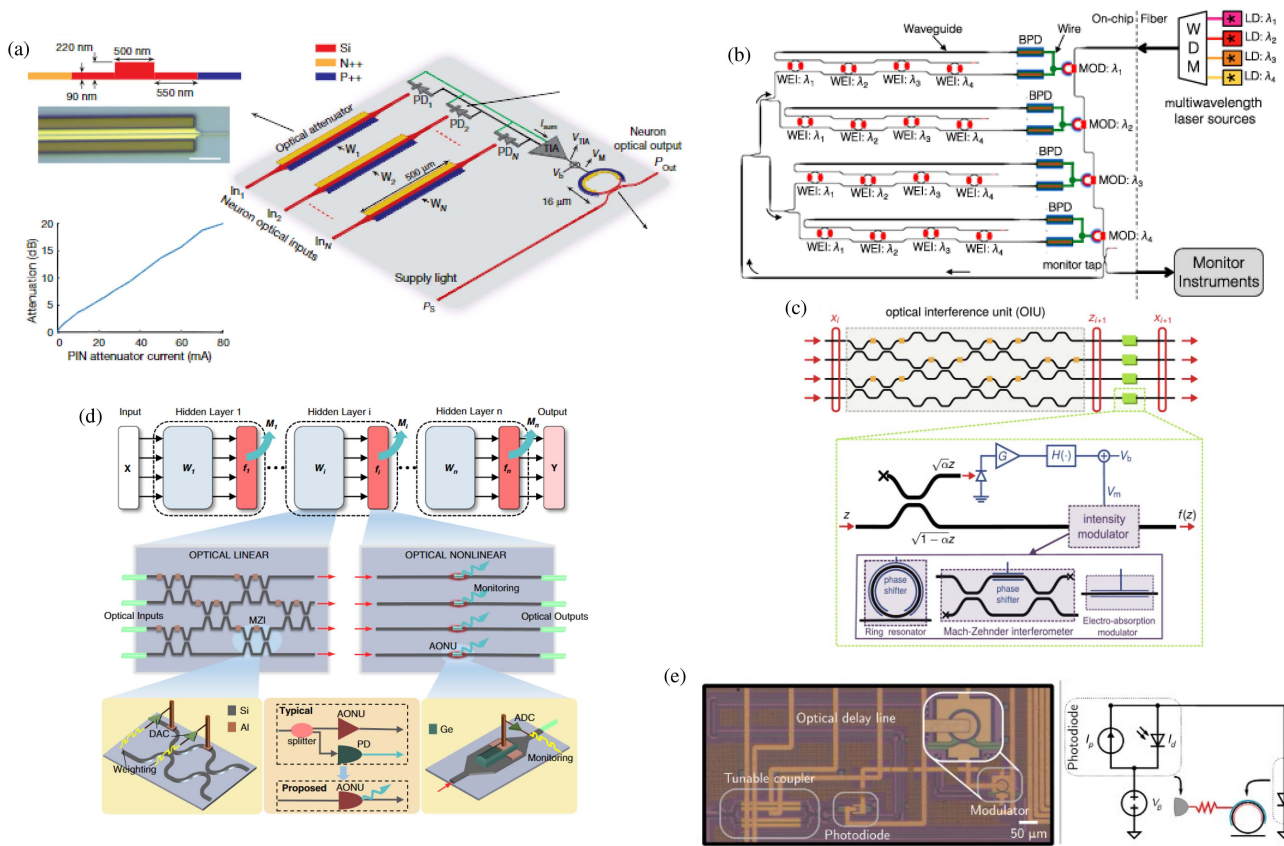


Fig. 5. Circuits and architectures implementing the activation function on-chip. (a) Ashtian et al. show a full O/E/O domain change to pilot a MRR as nonlinear response [101]. (b) Tait et al. use a similar approach to pilot an add-drop microrings bank, using balanced photodetectors [102]. (c) Experimental implementation of arbitrary activation function by tapping part of the optical output power [103]. (d) Shi et al. present the use of a short SiGe photodetectors to implement a nonlinear transfer function between optical power input and output [104]. (e) Implementation used by Bandyopadhyay et al., where just a tap of the optical output is used to modulate itself in nonlinear way by a MRR [73].

Ashtian et al. [101] (Fig. 5(a)), and similarly by Tait et al. [102] (Fig. 5(b)), implements the activation function by modulating the wavelength resonance of a microring resonator, that is fed from an external CW laser source that can be directly sent into the following neural layer. The 2 implementations have some differences: Ashtian et al. implement the summation by combining the current of the photodetectors, directly connected to the modulator and inputs. Moreover, a stage of amplification is placed to match the voltage levels between the sum of the photocurrents and the p-n junction of the microring resonator. Using this scheme, there is no need for WDM multiplexing, and low latency is assured (down to 570 ps per classification). On the other hand, Tait et al. use a WDM scheme in loop-back with differential photodetectors to tune directly the microring resonators. This scheme presents some limitations and opportunities, in particular, can be sensitive to fabrication differences between photodetectors pairs, unbalancing the actual response of the microring and adding parasitic capacitance. This fully O/E/O approach, where a complete domain change, from optical to electrical and back to optical is used, takes advantage of the full bandwidth of the components as the signals stay in the analog domain. However, the O/E/O approach adds noise sources, in particular due to the photodetectors and amplification stage [109]. To reduce the actual noise, one proposed

solution is using modulators that require a lower $V_{\pi}L$, to reduce or completely avoid the amplification stage. Heterogeneously integrated devices, such as ITO-based modulators [110], [111], or ITO-graphene device [112], can reduce the $V_{\pi}L$ by orders of magnitude, reducing drastically the need for amplification stages, and so the noise introduced by them. With the implementation of these novel materials, this approach can be scaled several times, as the optical propagation losses are not a limitation.

B. Nonlinear Adsorption Devices

Another architecture to implement a nonlinearity is by design a custom nonlinear device, so providing a nonlinear function between the optical input and output. A solution based on SiGe photodetector has been proposed by Shi et al. [104], leveraging the short structure of the SiGe to limit the maximum optical power output from the component, by so implementing a nonlinear transfer function (Fig. 5(d)). This solution has the advantages to permit the monitoring of the power while preserving the latency of the optical circuit, maintaining the large optical bandwidth. Similar approaches have been explored by other groups, to find the best material to perform this function both on the detection and modulation side, aiming for a better energy-efficient way [110], [113]. This type of approach has

the potential to leverage on the different types of material that can be used, limited by the compatibility with the SiPh CMOS process. The main drawback can be identified in the scaling limitations: since the input power must meet a certain criteria to activate the nonlinear function (mW-level) and the propagation is still affected by dB/cm losses, a large NN with several consecutive layers, or a high loss PIC would not be suitable for this approach, unless other adjustments (like on-chip amplification stage) would be adopted.

C. Light Splitting and Detection

Another approach has been used by Moayed et al. [103], and Bandyopadhyay et al. [73], shown respectively in Fig. 5(c)–(e). In this case, the linear part of the NN is based on the MZI mesh, and the nonlinear function is activated by just part of the light power of the output waveguide (splitting), which is detected and the signal used to modulate the amplitude of the remaining part of the optical signal. This allows feeding the whole network with the same optical input signals, reducing the need to have more lasers or input couplers. However, the network will add layers of modulation on top of each other, making the same scheme more sensitive to noise and not directly suitable for WDM expansions. Following the same approach, Xu et al. [114] propose a similar scheme. In this case the NL part is implemented using a MZI, where one of the arm is controlled by a optical memory-based feedback circuit, using a PCM material as nonvolatile element. The light-splitting-and-detection has some clear advantages as the modulation is directly on the same optical signal, with a clear advantage in terms of speed and latency. However, the tap requires an electrical circuit capable of reading low currents from the photodetectors and translate into proper signals, limiting the actual bandwidth to few GHz, and posing limitation in the energy consumption as well, determine by the Transimpedance amplifiers (TIA). Moreover, the continuous splitting layer after layer increases the insertion loss of the overall photonic circuit, putting more pressure on the performances of the latter activation functions in terms of minimum optical power detection.

In all cases, the activation function is encoded at the hardware level, resulting in a fixed size of the number of inputs, layers, and outputs, and so fixing the scalability and limiting parallelization. Schemes that can be used to subdivide the matrix into smaller ones to fit large MVM on smaller hardware cannot be used with hardware activation functions, as the nonlinear function is applied *a-priori*, and so it does not allow for a temporal multiplexing summation. By so, the research may investigate schemes that allow an actual flexible implementation of the nonlinear function, by exploiting more programmable photonic circuits or optical buffer for example, at a cost of increasing the latency, that has to count for all the signals to be accumulated before performing the activation function. Energy consumption is also a concern for the activation function. The All-Optical scheme must provide enough optical power for each layer, creating the need for optical on-chip amplifiers, while the other schemes must take into account the consumption of high-speed TIAs (tens of mW per each TIA [115]), but with the possibility to tune the amplification stage to reach different types of activation

functions. Mixing the approaches (all-optical for first layers when the optical power is not a concern, using the others for the following layers) might result in a good trade-off between energy efficiency and complexity.

VI. DISCUSSION

In all this review, it has been clear that photonics has a great opportunity to be the hardware accelerator for NN applications, as the increasing number of machine learning applications is driving the actual hardware to its limit. Integrated photonics, and Silicon Photonics (SiPh) as the main actor, have several advantages and directions that could drive the implementation of fast and reliable Photonic Neural Networks (PNNs). The research and progress that have been done in the last decades for mainly telecom purposes have now a new shine in another field. Among them, we can see the main driving forces:

- *Components*: SiPh can now show several components that are over the possibilities of any electrical counterpart in terms of speed and energy efficiency. Modulators up to 100 GHz [51], and photodetectors that can reach over 200 GHz have been presented [116], while CMOS foundries are more and more implementing SiPh lines, with state-of-the-art components in their PDKs. Note that all these components come from another field (telecom mainly), but their impact can be further beyond the initial field, resulting in a reduced initial cost.
- *Emerging Materials and Devices*: The research over new materials and new devices has brought several innovations in recent years [97], [117]. Among many, ITO has shown the most potential, especially in terms of energy efficiency and footprint, being 1000x smaller than Si EOM, and 10,000x smaller than Lithium Niobate [110], [111], [112], [118], [119], [120] with similar performances. Beyond ITO, two-dimensional material-based solutions may yet play a role in future semiconductor chips and tensor core processors; for instance, the accumulation operation can be performed simply and incoherently using a photodetector as discussed above. The detector's figure of merit, the gain-bandwidth-product, falls into either sensitive-but-slow or into fast-but-non-sensitive quadrants [121]. Recent developments on slot-based 2D material detectors show to overcome the transient-time and RC-delay time limitations offering sensitive and fast detectors while offering a minuscule footprint. Furthermore, PN-junction-based 2D detectors have demonstrated promising performance while not requiring a bias, thus saving power and wire-routing complexity [122], [123].
- *I/O*: One of the limitations that slowed down the expansion of SiPh was the actual coupling in/out of the chip, an essential piece considering also the lack of integrated light sources. SiPh has now advanced packaging tools to provide small form factor chips, with laser sources on-chip [124]. Moreover, the expansion of the materials used has brought new devices, such as P-RAM [82], [98], to be integrated, reducing the dependency on external digital electronic memories, one of the bottlenecks of electronics. The next

steps will focus on inter-chip communications, as well as intra-chip ones.

- *Domain Crossings*: Photonic-based tensor core processors are analog in nature and hence may require digital to analog and vice versa domain crossings. Above 5 GHz baud rates and 8-bit resolution DACs and ADCs become quite expensive to operate [55]. If the PTC application allows processing data in the optical domain (from an optical input, such as for intra data-center, for example), then a photonic PIC-based DAC would be beneficial [125]. This could include also energy harvesting, such as recapturing optical nonlinearities [126], nanoscale RF antennas or solar cells [127].
- *Architectures*: As seen, several architectures have been proposed and demonstrated. While a clear winner is still to be found, all of them can push towards several improvements to further expand their performances [23]. On one side, parallelization exploiting other degrees of freedom can further push the performances. On the other side, techniques such as pruning or others can be implemented on-chip as well, making room for improvements in the overall system. There may also be options to learn from emerging architectures such as hybrid (electronic-photonic) network-on-chip approaches that allocate interconnect technology between local (electronic) and distant (optical) requirements, which may also allow for some degree of network reconfiguration for demand optimization [128], [129].

However, photonics has still to improve some aspects to become a viable solution for deep learning and machine intelligence. Adding a nonlinear activation function in the optical domain is challenging, and more efficient all-optical nonlinearities need to be explored, yet, electro-optical nonlinearity devices are promising, as shown, despite some architectural overhead, such footprint, accumulation detectors, and ADCs. Analog-to-digital and digital-to-analog conversions must be taken into account too: domain crossings (i.e. DAC & ADCs) constitute the majority power consumption for heterogeneous photonic-electronic machine intelligence accelerators [55]. However, emerging monolithic integration solutions (e.g. Global Foundry 45 nm, GF45SPLCO [130]) hold great promise to minimized communication overhead. Furthermore, emerging packaging solutions including stacking multiple BEOLs [131], [132], integrating plurality of chiplets onto the same interposer, with world-record pin pitches of 10 μm [133], will enable extremely tight integrated heterogeneous PNN-CMOS ASIC solution with unprecedented performance. The upcoming SRC JUMP2.0 Center for Heterogeneous Integration of Micro Electronic Systems (CHIMES) will explore the latter in detail. By last, laser integration must become a standard in the SiPh process, allowing high energy-efficient ($>5\%$) lasers to be implemented on-chip, exploring integration [124], or Photonic Wire Bonding [134].

VII. CONCLUSION

In this paper, we review the main aspects that enable integrated photonic technologies to become a key resource for the current and next generation of hardware accelerators for Neural

Networks. We review the main advantages that photonics has compared to electronics, in terms of power efficiency, latency, and bandwidth. The main architectures that have been used so far to implement linear and nonlinear operations on PIC have been shown, highlighting the pros and cons of each one of them, and outlining a comparison among them. We finally discuss among the main drive forces that will boost the photonic approach in the next years.

Considering all those aspects, photonics will still play an important role in the research for the next generation of hardware accelerators. As more and more computational power is required and considering energy efficiency a key factor, photonics will be in the spotlight in the near and long-term future.

REFERENCES

- [1] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, 2021.
- [2] B. J. Shastri, A. N. Tait, C. Huang, V. J. Sorger, and P. R. Prucnal, "Prospects and applications of silicon photonic neural networks," in *Proc. Eur. Conf. Opt. Commun.*, vol. 12019, pp. 135–144, 2022.
- [3] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM J. Res. Dev.*, vol. 3, no. 3, pp. 210–229, 1959.
- [4] A. L. Fradkov, "Early history of machine learning," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 1385–1390, 2020.
- [5] N. Yadav, A. Yadav, and M. Kumar, "History of neural networks," in *An Introduction to Neural Network Methods for Differential Equations*. Berlin, Germany: Springer, 2015, pp. 13–15.
- [6] J. Leonard and M. Kramer, "Improvement of the backpropagation algorithm for training neural networks," *Comput. Chem. Eng.*, vol. 14, pp. 337–341, 1990.
- [7] B. Mahesh, "Machine learning algorithms—A review," *Int. J. Sci. Res.*, vol. 9, pp. 381–386, 2020.
- [8] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, 2015.
- [9] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, 2018, Art. no. e00938.
- [10] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, 2017.
- [11] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," 2022, *arXiv:2204.06125*.
- [12] T. Hwang, "Computational power and the social impact of artificial intelligence," 2018, *arXiv:1803.08971*.
- [13] B. J. Erickson, P. Korfiatis, Z. Akkus, T. Kline, and K. Philbrick, "Toolkits and libraries for deep learning," *J. Digit. Imag.*, vol. 30, no. 4, pp. 400–405, 2017.
- [14] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "Survey and benchmarking of machine learning accelerators," in *Proc. IEEE High Perform. Extreme Comput. Conf.*, 2019, pp. 1–9.
- [15] N. Suda et al., "Throughput-optimized openCL-based FPGA accelerator for large-scale convolutional neural networks," in *Proc. ACM/SIGDA Int. Symp. Field-Programmable Gate Arrays*, 2016, pp. 16–25.
- [16] J. Zhou et al., "Tunao: A high-performance and energy-efficient reconfigurable accelerator for graph processing," in *Proc. IEEE/ACM 17th Int. Symp. Cluster, Cloud, Grid Comput.*, 2017, pp. 731–734.
- [17] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "Survey of machine learning accelerators," in *Proc. IEEE High Perform. Extreme Comput. Conf.*, 2020, pp. 1–12.
- [18] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2015, vol. 1, pp. 1135–1143.
- [19] B. J. Shastri et al., "Photonics for artificial intelligence and neuromorphic computing," *Nature Photon.*, vol. 15, pp. 102–114, 2021.
- [20] W. Ma, Z. Liu, Z. A. Kudyshev, A. Boltasseva, W. Cai, and Y. Liu, "Deep learning for the design of photonic structures," *Nature Photon.*, vol. 15, no. 2, pp. 77–90, 2021.

- [21] M. Miscuglio and V. J. Sorger, "Photonic tensor cores for machine learning," *Appl. Phys. Rev.*, vol. 7, 2020, Art. no. 031404, doi: [10.1063/5.0001942](https://doi.org/10.1063/5.0001942).
- [22] D. A. Miller, "Attojoule optoelectronics for low-energy information processing and communications," *J. Lightw. Technol.*, vol. 35, no. 3, pp. 346–396, Feb. 2017.
- [23] F. P. Sunny, E. Taheri, M. Nikdast, and S. Pasricha, "A survey on silicon photonics for deep learning," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 17, no. 4, pp. 1–57, 2021.
- [24] R. Soref and B. Bennett, "Electrooptical effects in silicon," *IEEE J. Quantum Electron.*, vol. 23, no. 1, pp. 123–129, Jan. 1987.
- [25] L. Chrostowski and M. Hochberg, *Silicon Photonics Design: From Devices to Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2015.
- [26] S. Y. Siew et al., "Review of silicon photonics technology and platform development," *J. Lightw. Technol.*, vol. 39, no. 13, pp. 4374–4389, Jul. 2021.
- [27] W. Haensch, "Scaling is over-What now?," in *Proc. IEEE 75th Annu. Device Res. Conf.*, 2017, pp. 1–2.
- [28] D. Etiemble, "45-year CPU evolution: One law and two equations," 2018, *arXiv:1803.00254*.
- [29] S. Petrenko, "Limitations of Von Neumann architecture," in *Big Data Technologies for Monitoring of Computer Security: A Case Study of the Russian Federation*. Berlin, Germany: Springer, 2018, pp. 115–173.
- [30] A. Ganguly, R. Muralidhar, and V. Singh, "Towards energy efficient non-von Neumann architectures for deep learning," in *Proc. IEEE 20th Int. Symp. Qual. Electron. Des.*, 2019, pp. 335–342.
- [31] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "A survey of model compression and acceleration for deep neural networks," 2017, *arXiv:1710.09282*.
- [32] L. Deng, G. Li, S. Han, L. Shi, and Y. Xie, "Model compression and hardware acceleration for neural networks: A comprehensive survey," *Proc. IEEE*, vol. 108, no. 4, pp. 485–532, Apr. 2020.
- [33] C. Ding et al., "C_{IR} CNN: Accelerating and compressing deep neural networks using block-circulant weight matrices; circnn: Accelerating and compressing deep neural networks using block-circulant weight matrices," in *Proc. IEEE/ACM 50th Annu. Int. Symp. Microarchit.*, 2017, vol. 14, pp. 13–17, doi: [10.1145/3123939.3124552](https://doi.org/10.1145/3123939.3124552).
- [34] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in *Proc. Int. Conf. Learn. Representations*, 2016.
- [35] W. Sung, S. Shin, and K. Hwang, "Resiliency of deep neural networks under quantization," 2015, *arXiv:1511.06488*.
- [36] D. Blalock, J. J. G. Ortiz, J. Frankle, and J. Guttat, "What is the state of neural network pruning?," in *Proc. Conf. Mach. Learn. Syst.*, 2020, vol. 2, pp. 129–146.
- [37] M. Astrid and S.-I. Lee, "Deep compression of convolutional neural networks with low-rank approximation," *ETRI J.*, vol. 40, no. 4, pp. 421–434, 2018.
- [38] B. Zhuang, C. Shen, M. Tan, L. Liu, and I. Reid, "Towards effective low-bitwidth convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7920–7928.
- [39] M. Pethick, M. Liddle, P. Werstein, and Z. Huang, "Parallelization of a backpropagation neural network on a cluster computer," in *Proc. Int. Conf. Parallel Distrib. Comput. Syst.*, 2003, vol. 392, pp. 165–208.
- [40] A. Saiyeda and M. A. Mir, "Cloud computing for deep learning analytics: A survey of current trends and challenges," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 2, pp. 68–72, 2017.
- [41] T.-J. Yang, Y.-H. Chen, J. Emer, and V. Sze, "A method to estimate the energy consumption of deep neural networks," in *Proc. IEEE 51st Asilomar Conf. Signals, Syst., Comput.*, 2017, pp. 1916–1920.
- [42] R. Rani and R. Garg, "A survey of thermal management in cloud data centre: Techniques and open issues," *Wireless Pers. Commun.*, vol. 118, no. 1, pp. 679–713, 2021.
- [43] R. Machupalli, M. Hossain, and M. Mandal, "Review of ASIC accelerators for deep neural network," *Microprocess. Microsyst.*, vol. 89, 2022, Art. no. 104441.
- [44] J. J. Zhang, T. Gu, K. Basu, and S. Garg, "Analyzing and mitigating the impact of permanent faults on a systolic array based neural network accelerator," in *Proc. IEEE 36th VLSI Test Symp. (VTS)*, 2018, pp. 1–6.
- [45] J. Burgess, "RTX on-the NVIDIA Turing GPU," in *Proc. IEEE Hot Chips 31 Symp.*, Cupertino, CA, USA, 2019, pp. 1–27, doi: [10.1109/HOTCHIPS.2019.8875651](https://doi.org/10.1109/HOTCHIPS.2019.8875651).
- [46] A. Yang, "Deep learning training at scale spring crest deep learning accelerator (intel Nervana NNP-T)," in *Proc. IEEE Hot Chips 31st Symp.*, Cupertino, CA, USA, 2019, pp. 1–20, doi: [10.1109/HOTCHIPS.2019.8875643](https://doi.org/10.1109/HOTCHIPS.2019.8875643).
- [47] P. Bannon, G. Venkataraman, D. D. Sarma, and E. Talpes, "Computer and redundancy solution for the full self-driving computer," in *Proc. IEEE Hot Chips 31st Symp.*, Cupertino, CA, USA, 2019, pp. 1–22, doi: [10.1109/HOTCHIPS.2019.8875645](https://doi.org/10.1109/HOTCHIPS.2019.8875645).
- [48] C. Huang et al., "Prospects and applications of photonic neural networks," 2022. [Online]. Available: <https://www.tandfonline.com/action/journalInformation?journalCode=tapx20>
- [49] A. N. Tait, "Quantifying power in silicon photonic neural networks," *Phys. Rev. Appl.*, vol. 17, 2022, Art. no. 054029.
- [50] M. Hochberg and T. Baehr-Jones, "Towards fabless silicon photonics," *Nature Photon.*, vol. 4, no. 8, pp. 492–494, 2010.
- [51] A. Rahim et al., "Taking silicon photonics modulators to a higher performance level: State-of-the-art and a review of new technologies," *Adv. Photon.*, vol. 3, no. 2, 2021, Art. no. 024003.
- [52] S. Lischke et al., "Ultra-fast germanium photodiode with 3-dB bandwidth of 265 GHz," to be published, doi: [10.1038/s41566-021-00893-w](https://doi.org/10.1038/s41566-021-00893-w).
- [53] M. Miscuglio et al., "Massively parallel amplitude-only Fourier neural network," *Optica*, vol. 7, no. 12, pp. 1812–1819, 2020.
- [54] N. Peserico et al., "Design and testing of integrated photonic chip for convolution neural network," in *Proc. Int. Conf. Imag. Syst. Appl.*, 2022, paper no. ITH3D-7.
- [55] T. Drenski and J. C. Rasmussen, "ADC/DAC and ASIC technology trends," in *Proc. IEEE 24th OptoElectron. Commun. Conf., Int. Conf. Photon. Switching Comput.*, 2019, pp. 1–3.
- [56] Q. Cheng, J. Kwon, M. Glick, M. Bahadori, L. P. Carloni, and K. Bergman, "Silicon photonics codesign for deep learning," *Proc. IEEE*, vol. 108, no. 8, pp. 1261–1282, Aug. 2020.
- [57] F. Morichetti, S. Grillanda, and A. Melloni, "Breakthroughs in photonics 2013: Toward feedback-controlled integrated photonics," *IEEE Photon. J.*, vol. 6, no. 2, Apr. 2014, Art. no. 0701306.
- [58] H. Zhou et al., "Photonic matrix multiplication lights up photonic accelerator and beyond," *Light: Sci. Appl.*, vol. 11, no. 1, pp. 1–21, 2022.
- [59] B. Yunping et al., "Photonic multiplexing techniques for neuromorphic computing," *Nanophotonics*, vol. 12, no. 5, pp. 795–817, 2023, doi: [10.1515/nanoph-2022-0485](https://doi.org/10.1515/nanoph-2022-0485).
- [60] M. A. Al-Qadasi, L. Chrostowski, and B. J. Shastri, "Scaling up silicon photonic-based accelerators: Challenges and opportunities collections articles you may be interested in," *APL Photon.*, vol. 7, 2022, Art. no. 20902, doi: [10.1063/5.0070992](https://doi.org/10.1063/5.0070992).
- [61] N. Margalit, C. Xiang, S. M. Bowers, A. Bjorlin, R. Blum, and J. E. Bowers, "Perspective on the future of silicon photonics and electronics," *Appl. Phys. Lett.*, vol. 118, no. 22, 2021, Art. no. 220501.
- [62] W. Bogaerts and L. Chrostowski, "Silicon photonics circuit design: Methods, tools and challenges," *Laser Photon. Rev.*, vol. 12, 2018, Art. no. 1700237.
- [63] L. Chrostowski et al., "Silicon photonic circuit design using rapid prototyping foundry process design kits," *IEEE J. Sel. Topics Quantum Electron.*, vol. 25, no. 5, Sep./Oct. 2019, Art. no. 8201326.
- [64] C. Ríos et al., "Integrated all-photonic non-volatile multi-level memory," *Nature Photon.*, vol. 9, no. 11, pp. 725–732, 2015.
- [65] Z. Cheng, C. Ríos, N. Youngblood, C. D. Wright, W. H. Pernice, and H. Bhaskaran, "Device-level photonic memories and logic applications using phase-change materials," *Adv. Mater.*, vol. 30, no. 32, 2018, Art. no. 1802435.
- [66] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, "Experimental realization of any discrete unitary operator," *Phys. Rev. Lett.*, vol. 73, no. 7, 1994, Art. no. 58. [Online]. Available: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.73.58>
- [67] D. A. Miller, "Self-configuring universal linear optical component," *Photon. Res.*, vol. 1, no. 1, pp. 1–15, 2013.
- [68] B. J. Metcalf, I. A. Walmsley, P. C. Humphreys, W. S. Kolthammer, and W. R. Clements, "Optimal design for universal multipoint interferometers," *Optica*, vol. 3, no. 12, pp. 1460–1465, Dec. 2016.
- [69] Y. Shen et al., "Deep learning with coherent nanophotonic circuits," *Nature Photon.*, vol. 11, pp. 441–446, Jun. 2017.
- [70] C. Demirkiran et al., "An electro-photonic system for accelerating deep neural networks," 2021. [Online]. Available: <http://arxiv.org/abs/2109.01126>
- [71] H. Zhang et al., "An optical neural chip for implementing complex-valued neural network," *Nature Commun.*, vol. 12, no. 1, 2021, Art. no. 457, doi: [10.1038/s41467-020-20719-7](https://doi.org/10.1038/s41467-020-20719-7).
- [72] 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, 2021.

- [73] S. Bandyopadhyay et al., "Single chip photonic deep neural network with accelerated training," 2022, *arXiv:2208.01623*.
- [74] J. Carolan et al., "Universal linear optics," *Science*, vol. 349, pp. 711–716, 2015.
- [75] A. Ruocco, A. Ribeiro, L. Vanacker, and W. Bogaerts, "Demonstration of a 4x4-port universal linear circuit," *Optica*, Vol. 3, no. 12, pp. 1348–1357, 2016.
- [76] A. Annoni et al., "Unscrambling light—automatically undoing strong mixing between modes," *Light: Sci. Appl.*, vol. 6, no. 12, pp. e17110–e17110, 2017.
- [77] D. A. Miller, "Perfect optics with imperfect components," *Optica*, vol. 2, no. 8, pp. 747–750, 2015.
- [78] R. Baghdadi et al., "Dual slot-mode NOEM phase shifter," *Opt. Exp.*, vol. 29, no. 12, pp. 19113–19119, 2021.
- [79] A. Cem, S. Yan, Y. Ding, D. Zibar, and F. D. Ros, "Data-driven modeling of Mach-Zehnder interferometer-based optical matrix multipliers," *J. Lightw. Technol.*, to be published, doi: [10.1109/JLT.2023.3263235](https://doi.org/10.1109/JLT.2023.3263235).
- [80] F. Shokraneh, S. Geoffroy-Gagnon, and O. Liboiron-Ladouceur, "Towards phase-error- and loss-tolerant programmable Mach-Zehnder interferometer-based optical processors for optical neural networks," in *Proc. IEEE Conf. Opt. Commun.*, 2020, pp. 1–2.
- [81] R. Hamerly, S. Bandyopadhyay, and D. Englund, "Robust zero-change self-configuration of the rectangular mesh," in *Proc. IEEE Opt. Fiber Commun. Conf. Exhib.* 2021, pp. 1–3.
- [82] S. R. Kari et al., "Optical and electrical memories for analog optical computing," *IEEE J. Sel. Topics Quantum Electron.*, vol. 29, no. 2, Mar./Apr. 2023, Art. no. 6100812.
- [83] R. Ji, J. Ding, L. Yang, L. Zhang, and Q. Xu, "On-chip CMOS-compatible optical signal processor," *Opt. Exp.*, Vol. 20, no. 12, pp. 13560–13565, Jun. 2012.
- [84] J. Feldmann et al., "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, pp. 52–58, 2021, doi: [10.1038/s41586-020-03070-1](https://doi.org/10.1038/s41586-020-03070-1).
- [85] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Broadcast and weight: An integrated network for scalable photonic spike processing," *J. Lightw. Technol.*, vol. 32, no. 21, pp. 4029–4041, Nov. 2014.
- [86] G. Liu et al., "Photonic pattern reconstruction enabled by on-chip online learning and inference," *J. Phys.: Photon.*, vol. 3, 2021, Art. no. 024006. [Online]. Available: <https://iopscience.iop.org/article/10.1088/2515-7647/abe3d9https://iopscience.iop.org/article/10.1088/2515-7647/abe3d9/meta>
- [87] X. Ma et al., "High-density integrated photonic tensor processing unit with a matrix multiply compiler," 2022. [Online]. Available: <https://www.researchsquare.com/article/rs-1833027/latest.pdf>
- [88] F. Brücknerhoff-Plückelmann et al., "Broadband photonic tensor core with integrated ultra-low crosstalkwavelength multiplexers," *Nanophotonics*, vol. 11, pp. 4063–4072, 2022, doi: [10.1515/nanoph-2021-0752/html](https://doi.org/10.1515/nanoph-2021-0752/html).
- [89] S. G. Sarwat et al., "An integrated photonics engine for unsupervised correlation detection," *Sci. Adv.*, vol. 8, 2022, Art. no. 3243, doi: [10.1126/sci-adv.abn3243](https://doi.org/10.1126/sci-adv.abn3243).
- [90] Q. Xu and R. Soref, "Reconfigurable optical directed-logic circuits using microresonator-based optical switches," *Opt. Exp.*, vol. 19, no. 6, pp. 5244–5259, Mar. 2011. [Online]. Available: <https://opg.optica.org/oe/abstract.cfm?URI=oe-19-6-5244>
- [91] T. A. Ibrahim, K. Amarnath, L. C. Kuo, R. Grover, V. Van, and P.-T. Ho, "Photonic logic nor gate based on two symmetric microring resonators," *Opt. Lett.*, vol. 29, 2004, Art. no. 2779.
- [92] A. N. Tait et al., "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, Aug. 2017. [Online]. Available: <https://www.nature.com/articles/s41598-017-07754-z>
- [93] C. Huang et al., "A silicon photonic–electronic neural network for fibre nonlinearity compensation," *Nature Electron.*, vol. 4, no. 11, pp. 837–844, 2021.
- [94] W. Zhang et al., "Silicon microring synapses enable photonic deep learning beyond 9-bit precision," *Optica*, vol. 9, no. 5, pp. 579–584, 2022.
- [95] R. Amin, J. B. Khurgin, and V. J. Sorger, "Waveguide-based electro-absorption modulator performance: Comparative analysis," *Opt. Exp.*, vol. 26, no. 12, pp. 15445–15470, 2018.
- [96] T. Alexoudi, G. T. Kanellos, and N. Pleros, "Optical RAM and integrated optical memories: A survey," *Light: Sci. Appl.*, vol. 9, no. 1, pp. 1–16, 2020.
- [97] N. Peserico, T. F. d. Lima, T. F. d. Lima, P. Prucnal, and V. J. Sorger, "Emerging devices and packaging strategies for electronic-photonic AI accelerators: Opinion," *Opt. Mater. Exp.*, vol. 12, no. 4, pp. 1347–1351, 2022.
- [98] J. Meng et al., "Electrical pulse driven multi-level nonvolatile photonic memories using broadband transparent phase change materials," 2022, *arXiv:2203.13337*.
- [99] D. Sahoo and R. Naik, "GSST phase change materials and its utilization in optoelectronic devices: A review," *Mater. Res. Bull.*, vol. 148, 2022, Art. no. 111679.
- [100] A. Redaelli, E. Petroni, and R. Annunziata, "Material and process engineering challenges in Ge-rich GST for embedded PCM," *Mater. Sci. Semicond. Process.*, vol. 137, 2022, Art. no. 106184.
- [101] 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, doi: [10.1038/s41586-022-04714-0](https://doi.org/10.1038/s41586-022-04714-0).
- [102] A. N. Tait et al., "Silicon photonic modulator neuron," *Phys. Rev. Appl.*, vol. 11, no. 6, 2019, Art. no. 064043.
- [103] M. Moayedi et al., "Experimental realization of arbitrary activation functions for optical neural networks," *Opt. Exp.*, Vol. 28, no. 8, pp. 12138–12148, 2020.
- [104] Y. Shi et al., "Nonlinear germanium-silicon photodiode for activation and monitoring in photonic neuromorphic networks," *Nature Commun.*, vol. 13, 2022, Art. no. 6048, doi: [10.1038/s41467-022-33877-7](https://doi.org/10.1038/s41467-022-33877-7).
- [105] P. Gupta and S. Li, "4F optical neural network acceleration: An architecture perspective," *Proc. SPIE*, vol. 12019, pp. 77–84, 2022.
- [106] J. Leuthold, C. Koos, and W. Freude, "Nonlinear silicon photonics," *Nature Photon.*, vol. 4, no. 8, pp. 535–544, 2010.
- [107] F. Morichetti, A. Canciamilla, C. Ferrari, A. Samarelli, M. Sorel, and A. Melloni, "Travelling-wave resonant four-wave mixing breaks the limits of cavity-enhanced all-optical wavelength conversion," *Nature Commun.*, vol. 2, no. 1, pp. 1–8, 2011.
- [108] L. Chang, S. Liu, and J. E. Bowers, "Integrated optical frequency comb technologies," *Nature Photon.*, vol. 16, no. 2, pp. 95–108, 2022.
- [109] T. F. d. Lima et al., "Noise analysis of photonic modulator neurons," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 1, Jan./Feb. 2020, Art. no. 7600109.
- [110] R. Amin et al., "ITO-based electro-absorption modulator for photonic neural activation function," *APL Mater.*, vol. 7, no. 8, 2019, Art. no. 081112.
- [111] Y. Gui et al., "100 GHz micrometer-compact broadband monolithic ITO Mach-Zehnder interferometer modulator enabling 3500 times higher packing density," *Nanophotonics*, vol. 11, no. 17, pp. 4001–4009, 2022.
- [112] R. Amin et al., "An ITO–graphene heterojunction integrated absorption modulator on Si-photonics for neuromorphic nonlinear activation," *APL Photon.*, vol. 6, no. 12, 2021, Art. no. 120801.
- [113] 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.
- [114] Z. Xu et al., "Reconfigurable nonlinear photonic activation function for photonic neural network based on non-volatile opto-resistive RAM switch," *Light: Sci. Appl.*, vol. 11, no. 1, 2022, Art. no. 288.
- [115] R. Costanzo and S. M. Bowers, "A 10-GHz bandwidth transimpedance amplifier with input DC photocurrent compensation loop," *IEEE Microw. Wireless Compon. Lett.*, vol. 30, no. 7, pp. 673–676, Jul. 2020.
- [116] S. Lischke et al., "Ultra-fast germanium photodiode with 3-dB bandwidth of 265 GHz," *Nature Photon.*, vol. 15, no. 12, pp. 925–931, 2021.
- [117] A. Fratallocchi et al., "Nano-optics gets practical," *Nature Nanotechnol.*, vol. 10, pp. 11–15, 2015.
- [118] H. Wang et al., "High-performance opto-electronics with emerging materials," *Proc. SPIE*, vol. 12200, 2022, Art. no. 1220002.
- [119] R. Amin et al., "Sub-wavelength GHz-fast broadband ito Mach-Zehnder modulator on silicon photonics," *Optica*, vol. 7, no. 4, pp. 333–335, 2020.
- [120] S. K. Pickus, S. Khan, C. Ye, Z. Li, and V. J. Sorger, "Silicon plasmon modulators: Breaking photonic limits," *IEEE Photon. Soc.*, vol. 27, no. 6, 2013.
- [121] V. J. Sorger and R. Maiti, "Roadmap for gain-bandwidth-product enhanced photodetectors: Opinion," *Opt. Mater. Exp.*, vol. 10, no. 9, pp. 2192–2200, 2020.
- [122] C. Patil, H. Dalir, J. H. Kang, A. Davydov, C. W. Wong, and V. J. Sorger, "Highly accurate, reliable, and non-contaminating two-dimensional material transfer system," *Appl. Phys. Rev.*, vol. 9, no. 1, 2022, Art. no. 011419.
- [123] H. Wang et al., "Self-powered broadband photodetector based on MoS₂/Sb₂Te₃ heterojunctions: A promising approach for highly sensitive detection," *Nanophotonics*, vol. 11, no. 22, pp. 5113–5119, 2022.
- [124] Z. Wang et al., "Novel light source integration approaches for silicon photonics," *Laser Photon. Rev.*, vol. 11, no. 4, 2017, Art. no. 1700063.

- [125] 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.
- [126] B. Jalali, V. Raghunathan, D. Dimitropoulos, and O. Boyraz, "Raman-based silicon photonics," *IEEE J. Sel. Topics Quantum Electron.*, vol. 12, no. 3, pp. 412–421, May/Jun. 2006.
- [127] M. H. Tahersima and V. J. Sorger, "Enhanced photon absorption in spiral nanostructured solar cells using layered 2D materials," *Nanotechnol.*, vol. 26, no. 34, 2015, Art. no. 344005.
- [128] V. K. Narayana, S. Sun, A.-H. A. Badawy, V. J. Sorger, and T. El-Ghazawi, "Morphonoc: Exploring the design space of a configurable hybrid NoC using nanophotonics," *Microprocess. Microsyst.*, vol. 50, pp. 113–126, 2017.
- [129] C. Shen et al., "Reconfigurable application-specific photonic integrated circuit for solving partial differential equations," 2022, *arXiv:2208.03588*.
- [130] M. Rakowski et al., "45 nm CMOS - silicon photonics monolithic technology (45CLO) for next-generation, low power and high speed optical interconnects," in *Proc. Opt. Fiber Commun. Conf. Exhib.*, 2020, pp. 1–3.
- [131] W.-T. J. Chan, Y. Du, A. B. Kahng, S. Nath, and K. Samadi, "BEOL stack-aware routability prediction from placement using data mining techniques," in *Proc. IEEE 34th Int. Conf. Comput. Des.*, 2016, pp. 41–48.
- [132] T. Standaert et al., "BEOL process integration for the 7 nm technology node," in *Proc. IEEE Int. Interconnect Technol. Conf./Adv. Metallization Conf.*, 2016, pp. 2–4.
- [133] V. Vega-Gonzalez et al., "Three-layer BEOL process integration with supervia and self-aligned-block options for the 3 nm node," in *Proc. IEEE Int. Electron Devices Meeting*, 2019, pp. 19–3.
- [134] M. R. Billah et al., "Hybrid integration of silicon photonics circuits and InP lasers by photonic wire bonding," *Optica*, vol. 5, no. 7, pp. 876–883, 2018.

Nicola Peserico received the Ph.D. degree from the Politecnico di Milano, Milan, Italy, in 2018. In 2019, he joined Femtorays, Italy, a silicon photonics startup for biosensing. He is currently a Postdoc Researcher with the Department of Electrical and Computer Engineering, George Washington University, Washington, DC, USA. His research interests include silicon photonics, AI/ML accelerators, optoelectronics devices and components, and bio-sensing with photonic integrated circuits.

Bhavin J. Shastri (Senior Member, IEEE) received the Ph.D. degree in electrical engineering (photonics) from McGill University, Montreal, QC, Canada, in 2012. He is currently an Assistant Professor of engineering physics with Queen's University, Kingston, ON, Canada, and a Faculty Affiliate with the Vector Institute for Artificial Intelligence, Canada. He was an Associate Research Scholar during 2016–2018 and Banting/NSERC Postdoctoral Fellow during 2012–2016 with Princeton University, Princeton, NJ, USA. He has authored or coauthored more than 70 journal articles and 100 conference proceedings, seven book chapters, and given more than 65 invited talks and lectures, five keynotes and four tutorials. He is a coauthor of the book (CRC Press, 2017) *Neuromorphic Photonics*, a term he helped coin. His research interests include silicon photonics, photonic integrated circuits, neuromorphic computing, and machine learning. Dr. Shastri was the recipient of the 2022 SPIE Early Career Achievement Award and the 2020 IUPAP Young Scientist Prize in Optics for his pioneering contributions to neuromorphic photonics from ICO. He is a Senior Member of Optica (formerly OSA). He was the recipient of the 2014 Banting Postdoctoral Fellowship from the Government of Canada, the 2012 D. W. Ambridge Prize for the top graduating Ph.D. student at McGill, and an IEEE Photonics Society 2011 Graduate Student Fellowship amongst others awards.

Volker J. Sorger (Senior Member, IEEE) is currently a Full Professor with the Department of Electrical and Computer Engineering and the Director of the Institute on AI & Photonics, the Head of the Devices & Intelligent Systems Laboratory, George Washington University, Washington, DC, USA. His research interests include devices & optoelectronics, AI/ML accelerators, mixed-signal ASICs, quantum matter & quantum processors, cryptography. For his work, Dr. Sorger was the recipient of multiple awards including the Presidential PECASE Award, the AFOSR YIP, the Emil Wolf Prize, and the National Academy of Sciences Award of the year. Dr. Sorger is the Editor of *Optica*, *Nanophotonics*, *Applied Physics Reviews*, *eLight*, *Chips*, and was the former Editor-In-Chief of *Nanophotonics*. He is a Fellow of Optica (former OSA), a Fellow of SPIE, a Fellow of the German National Academic Foundation. He is a Founder of Optelligence.