A Codesigned Integrated Photonic Electronic Neuron

Lorenzo De Marinis, Alessandro Catania, *Member, IEEE*, Piero Castoldi, Giampiero Contestabile, Paolo Bruschi, Massimo Piotto, *Member, IEEE*, and Nicola Andriolli, *Senior Member, IEEE*

***Abstract*— In the modern era of artificial intelligence, increasingly sophisticated artificial neural networks (ANNs) are implemented, which pose challenges in terms of execution speed and power consumption. To tackle this problem, recent research on reduced-precision ANNs opened the possibility to exploit analog hardware for neuromorphic acceleration. In this scenario, photonic-electronic engines are emerging as a short-medium term solution to exploit the high speed and inherent parallelism of optics for linear computations needed in ANN, while resorting to electronic circuitry for signal conditioning and memory storage. In this paper we introduce a precision-scalable integrated Photonic-Electronic Multiply-Accumulate Neuron (PEMAN). The proposed device relies on (i) an analog photonic engine to perform reduced-precision multiplications at high speed and low power, and (ii) an electronic front-end for accumulation and application of the nonlinear activation function by means of a nonlinear encoding in the analog-to-digital converter (ADC). The device has been numerically validated through cosimulations to perform multiply-accumulate operations (MAC). Simulations are based on the iSiPP50G SOI process for the photonic engine and a commercial 28 nm CMOS process for the electronic front-end. The PEMAN exhibits a multiplication accuracy of 6.1 ENOB up to 10 GMAC/s, while it can perform computations up to 56 GMAC/s with a reduced accuracy down to 2.1 ENOB. The device can trade off speed and power consumption with resolution, significantly outperforming its analog electronics counterparts both in terms of speed and energy consumption. With respect to other photonic ANNs, the PEMAN has comparable speed and energy consumption with a higher resolution, while outperforming them by a hundredfold in the fan-in, which opens the possibility to accelerate more complex networks.**

***Index Terms*— Photonic-electronic codesign, photonic neural networks, photonic analog computing, neural network accelerator, reduced precision computing.**

I. Introduction

NOWADAYS machine learning technology is pervasively used for a wide range of applications including image classification, speech recognition and language translation, decision making, web searches, content filtering on social networks, recommendations on e-commerce websites [1]. Deep learning is one of the fastest-growing machine learning methods, exploiting multi-layered artificial neural networks (ANNs) implemented in digital electronics for processing large data sets, combining and analysing vast amounts of information quickly without the need of explicit instructions [2]. The spreading of artificial intelligence (AI)-driven systems for an increasing number of applications is testified by the fact that the computing power required to train state-of-the-art AI doubled every 3.4 months since 2012 [3]. Notable deep learning milestones include ResNet winning the ImageNet challenge in 2015 by reaching a super-human level of accuracy in object recognition [4], and GPT-3, the largest AI model up to date, capable of producing high quality human-like writings thanks to an over-100-billion parameter ANN trained over a large part of the internet [5].

These results have been achieved thanks to increasingly sophisticated ANNs and training algorithms, and leveraging a very large amount of computing power. Indeed, general purpose graphical processing units (GPGPUs) have been identified as particularly suitable for implementing the parallel computing tasks typical of ANNs, and contributed significantly to their current success in real application scenarios [6]. Recently, field-programmable gate arrays (FPGAs) and digital or mixed-signal application-specific integrated circuits (ASICs) [7]–[9] have been specifically designed to implement ANN computations, improving both speed and energy efficiency for learning tasks. To this aim, these novel electronic solutions focus on advanced numerical representations and memory architectures suitable for high-speed matrix multiplications, and on a very high bidirectional off-chip bandwidth (exceeding 1 Tb/s) to enable model and data parallelism. Compact and energy-efficient neuromorphic hardware is indeed of paramount importance due to the high power dissipation of large neural network models, reaching several KWh during both training and inference [10]–[12].

Driven by the research on low-precision computing for ANNs, analog engines (e.g., based on memristors [13]) are promising as neuromorphic accelerators. The aim is to avoid the quadratic growth of the linear ANN computations (i.e., vector-matrix multiplications) as a function of the neural network layer size. Indeed, analog hardware, though more expensive than digital solutions, can be used to parallelize the linear computations.

Manuscript received 1 March 2022; revised 14 April 2022; accepted 20 May 2022. Date of publication 25 May 2022; date of current version 3 August 2022. This work was supportedin part by the CrossLab project (Departments of Excellence) funded by the Italian Ministry of Education and Research (MIUR) to the Department of Information Engineering of the University of Pisa. *(Corresponding author: Lorenzo De Marinis.)*

Lorenzo De Marinis, Piero Castoldi, and Giampiero Contestabile are with the Scuola Superiore Sant'Anna, 56124 Pisa, Italy (e-mail: lorenzo.demarinis@santannapisa.it).

Alessandro Catania, Paolo Bruschi, and Massimo Piotto are with the Department of Information Engineering, University of Pisa, 56122 Pisa, Italy.

Nicola Andriolli is with the National Research Council of Italy (CNR-IEIIT), 56122 Pisa, Italy.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JQE.2022.3177793.

Digital Object Identifier 10.1109/JQE.2022.3177793

In this scenario, photonic solutions show a great potential to realize analog low-power high-throughput accelerators for machine learning [14]–[16]. Photonic implementations typically exploit free-space optics, such as the diffractive architectures that make use of micromachined lenses [17], or integrated optics, e.g., the coherent solutions based on Mach-Zehnder interferometer meshes [18], [19]. Despite many research efforts, all-optical approaches must still overcome several challenges before their practical exploitation. The issues concern the large-scale integration and control of many photonic devices (comprising light sources) and the lack of suitable photonic nonlinearities for the activation function. The latter appears to be the main limitation towards truly deep photonic neural networks [20]. While some interesting works on optical nonlinearities are emerging [21]–[24], photonics at this state of development is promising in the short-medium term mainly to implement the linear computations required in ANNs, used in combination with electronic circuitry, thus realizing hybrid photonic-electronic accelerators.

The DEAP (Digital Electronics and Analog Photonics), proposed in [25], is an example of such photonic-electronic neuromorphic cores derived from the broadcast-and-weight architecture [26]. It is a wavelength division multiplexing (WDM)-based optical network that relies on double bus ring resonators connected to a balanced photodetector to perform bipolar multiplications. Another example of these hybrid devices is represented by the photonic tensor core proposed in [27]. This architecture exploits a phase change material to implement photonic memory elements used to record multiplicands. In both solutions, the multiplication results are encoded in the amplitude of a photocurrent after photodetection. Even though these solutions provide a sound system-level photonic-electronic architecture, an in-depth codesign of the photonic and electronic circuits towards the integration of both parts has still to be properly tackled.

Building upon the preliminary results reported in [28], in this paper we present the photonic-electronic multiply-accumulate neuron (PEMAN). It is a reduced-precision integrated photonic-electronic device based on a multiply-accumulate (MAC) processor with an ADC-embedded nonlinearity, suited to accelerate ANNs based on memoryless layers [29]. The PEMAN photonic engine exploits two Mach-Zehnder modulators and a balanced photodetector to perform high-speed bipolar multiplications. The electronic front-end comprises an accumulation capacitance and a loop-unrolled successive approximation register (SAR) ADC. This last element applies the nonlinearity of interest within the analog to digital conversion. This architecture is able to trade off speed with multiplication accuracy. In this work we focus on how the accelerator performance is affected by the non-idealities of the photonic linear engine, which represents the most critical part of the PEMAN. Due to the reduced precision and the relaxation on the operating frequency after the accumulation, the analog front-end is briefly presented with the aim to estimate the power consumption of the overall photonic-electronic device.

The remainder of this paper is structured as follows: after a background on ANN, precision-scalable and analog computing reported in Sec. II, in Sec. III we present the integrated photonic-electronic neuron. Sec. III-B analyzes the performance of the components and of the full photonic engine through circuit-based simulations, while Sec. V discusses speed, resolution, and energy consumption of the designed photonic-electronic device and compares it with analog electronic neuromorphic engines. Sec. VI concludes the paper.

II. Background

After recalling the main operations involved in ANN computation, this section focuses on the rationale behind reduced-precision computing for neuromorphic applications, and on the problem of interfacing analog computing to digital memories, with an emphasis on the relevant metrics.

ANNs are a class of machine learning methods vaguely inspired by biological neurons. An ANN is a collection of elementary units, called neurons, arranged in layers. Neurons can be connected either to all or to only a part of the neurons in adjacent layers, thus forming either fully-connected or sparsely-connected layers, respectively. In ANN models, the stimulus of a neuron is computed by adding all the input values, each one multiplied by a proper weight, which corresponds to a MAC operation. Finally, a nonlinear function (called activation) is applied to the accumulation result. The computations involved in an ANN layer composed of M neurons fed by a previous layer with N neurons are formalized as follows:

$$O_i = f\left(\sum_{n=1}^{N} w_{i,n} \cdot x_n + \theta_i\right) \tag{1}$$

where O_i, $(i \in 1, \ldots, M)$ is the output of the i-th neuron of the layer, x_n, $(n \in 1, \ldots, N)$ is the output of the n-th neuron of the previous layer feeding the current layer, $w_{i,n}$ is the weight from the n-th neuron of the previous layer and the i-th neuron of the current layer, θ_i is the i-th neuron bias term, and $f(\cdot)$ is the nonlinear activation function. The building blocks of an ANN are therefore three:

1) a linear part, performing the MAC operations;
2) a nonlinear part, which applies the nonlinear function to the result of MAC operations;
3) a memory element, storing the neuron output in order to be utilized in the successive layers.

While weights $w_{i,n}$ are normally bipolar, positive-only inputs x_n are widely used in neuromorphic applications since many nonlinear functions $f(\cdot)$ (i.e., the sigmoid, the softplus, and the rectified linear unit or ReLU) have positive-only outputs [4].

A. Reduced-Precision Computing in ANN

The most computationally intensive and time consuming workload in ANNs is constituted by the linear part, i.e., by MAC operations. This is because MAC operations in an ANN layer, described by Eq. 1, grow as $O(MN)$, while the computations in the nonlinear part grow only as $O(M)$ [29]. For this reason GPGPUs, particularly suited to perform vector-matrix multiplications, have enabled the

effective use of deep ANNs with thousands or even millions of neurons per layer [6].

In recent years, many research activities have been performed to reduce the complexity of ANN computations, for instance to apply ANN in safety-critical applications, where results should be obtained with low latency [30], or to exploit hardware-constrained devices. Several hardware and software solutions are indeed emerging in order to meet these low computing capacity constraints. The main goal of software solutions is to develop ANNs that, relying on simpler arithmetics, require less memory, while exhibiting negligible accuracy losses. For instance, ANNs have been pruned by removing less relevant connections, parameters have been normalized, and optimizations have been performed in dataflows to reduce data movement and storage [31]. Furthermore, works on reduced-precision computing have demonstrated the possibility to avoid the cumbersome floating point (FP) arithmetic by exploiting a small number of bits to represent ANN parameters with nearly negligible accuracy loss in several edge node applications [32]. These works report ANNs with parameters encoded with ≤ 8 bits [33] down to 1 bit in binary neural networks [34]. Based on the research activities in the field of reduced-precision ANNs, state-of-the-art GPGPUs implement dedicated hardware to perform integer operations (down to 1 bit) in order to reduce the power consumption and latency of ANNs [35]. Moreover, a new class of devices has recently emerged: precision-scalable MAC architectures [36]. These digital electronic architectures are designed to accelerate MAC operations in ANNs, making it possible to choose the number of bits used in computations, typically in three configurations: either 8, 4 or 2 bits of resolution. A lower precision results in higher speed and energy efficiency, making it possible to trade off speed and power efficiency with bit resolution.

In this scenario, analog hardware is gaining momentum to implement neuromorphic accelerators exploiting physical properties of circuits [37]. These analog engines aim to circumvent the quadratic growth in computational time associated with the number of neurons per layer, at the expense of more complex hardware [38], [39]. Analog electronics mainly exploits fundamental circuit laws and device properties (e.g., current sum in a circuit node) to perform MAC operations [40], [41]. A remarkable class of electronic neuromorphic devices are memristor crossbar arrays, also known as resistive RAMs (ReRAMs) [13]. However, ReRAM-based engines (being inherently resistive) suffer from high power dissipation issues, and lack reliable process standards and accurate models for simulation frameworks.

In the roadmap towards low power and high density MAC engines, neuromorphic photonics promise to bring sub-fJ per MAC power efficiency with high compactness, while relying on an inherently parallel hardware that reduces the complexity growth [42]. Nevertheless, several challenges must be tackled to enable effective all-optical approaches for neuromorphic hardware, including the efficient large-scale integration of many active and passive devices, and the reduction of losses and impairments, which may cause a significant accuracy drop (up to 70% in Mach-Zehnder-based coherent approaches) [43], [44]. While considerable effort is put to overcome these issues [18], [45], [46], photonic analog processors are also emerging within hybrid photonic-electronic accelerators, being particularly suited to perform high-speed MAC operations for reduced-precision ANNs [14], [25].

B. Resolution and Metrics for Analog Engines

Analog signals can be represented by a set of continuous values, while digital ones can be represented by a set of discrete values. However, analog computing cannot express continuously variable quantities, i.e., with arbitrarily high resolution, because of noise and distortions introduced by the analog hardware. This indeed limits the resolution of the analog system, i.e., the minimum distance between two distinguishable values. For any noise distribution, the standard deviation σ provides an estimate of the noise interval, namely the spreading of the values around the expected value.

As currently there is no established analog memory, information needs to be digitized in order to be stored. For this reason, the use of the "number of bits" is an appropriate metric to define the resolution of a photonic or electronic analog system, as it provides the bits needed to manage and store the information. To this aim the effective number of bits (ENOB) can be estimated, taking into account both noise and distortions.

Digital hardware makes use of floating point operations per second (i.e., FLOPS) to evaluate the computational speed. Systems based on reduced precision, such as analog engines, cannot be directly compared to electronic engines based on a floating point arithmetic. Once a given arithmetic precision has been chosen, an appropriate metric for reduced-precision systems is the MAC/s, which quantifies the speed at which MAC operations are carried out. Another metric that cannot be used for analog computing is the bit error rate (BER), assessing the number of altered bits in digital communications; instead, for analog systems ENOB is relevant. Moreover, the energy efficiency of analog processors has to be properly normalized over the kind of operation performed, i.e., Joule per MAC (J/MAC).

III. The Integrated Photonic Electronic Neuron

This section introduces the PEMAN, an integrated photonic-electronic precision-scalable MAC architecture with ADC embedded nonlinearity. The device has been codesigned in order to exploit the strengths of both photonic and electronic domains to perform the computations needed in neuromorphic applications. In particular, as depicted in Fig. 1, the PEMAN leverages: (i) an analog photonic engine to carry out reduced-precision multiplications at high speed and low power, and (ii) an electronic front-end to accumulate the multiplication results and compute the nonlinear function. The nonlinearity is inherently computed within the ADC, removing the need of a shared Look Up Table (LUT) as in [47]. The PEMAN performs all the operations required in a neuron. Its output is stored in the digital domain by design and can be exploited by the same device to implement neurons in the same or successive layers without scalability issues.

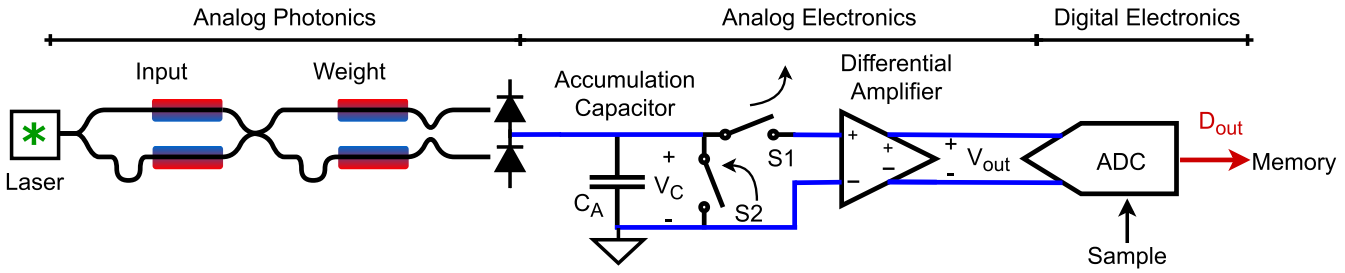

Fig. 1. The PEMAN schematic composed of: (i) an analog photonic engine that performs high speed multiplications, (ii) an analog electronic front-end to perform signal conditioning and accumulation, and (iii) an analog-to-digital converter with embedded nonlinearity.

A. Working Principle

The photonic engine relies on two cascaded travelling wave (TW) Mach-Zehnder modulators (MZM) to act on the intensity of an incoming lightwave and perform dot product multiplications. The first one is a 1×1 MZM able to impress input values in the range [0, 1], the second one is a 1×2 MZM connected to a balanced photodetector (PD) able to encode both positive and negative weights within the range $[-1, 1]$. The choice of encoding input x within [0, 1] fits well with the operation of the first MZM, which modulates the intensity of the incoming lightwave in a range from 0 (suppression state) to 1 (all-pass state); the unity-limited range can be overcome by a simple scaling of inputs and output.

The photocurrent generated by the balanced PD represents the multiplication result. The accumulation is then carried out after the opto-electronic conversion by charging (or discharging) a capacitor, thus implementing the MAC operation. The capacitor voltage is reset every $N+1$ accumulations of the results of the N input-weight multiplications and of the bias term θ_i, as shown in Eq. 1. The capacitor is connected to a differential amplifier, needed to drive the subsequent ADC. During the reset phase, the amplifier input is disconnected, the ADC samples the capacitor voltage, and subsequently the capacitor is reset to zero. The ADC has been designed with a nonlinear coding that allows inherently applying the neuron nonlinearity within the sampling operation, as detailed in Sec. III-B.

Differently from a transimpedance amplifier (TIA)-based photoreceiver, the integrating front-end accumulates in the analog domain the results of several operations before sampling, hence relaxing the ADC bandwidth specifications, In particular, sampling every $N+1$ operations allows the ADC rate to be $N+1$ times lower than the MAC rate. This is a critical aspect to reduce the ADC power consumption and to increase the achievable ENOB (typically quite low for high-speed ADC, e.g., ~ 2 for ADC operating at ≥ 5 GSa/s [48]).

The photonic engine has been emulated using imec iSiPP50G platform [49], while the electronic front-end has been designed using a commercial 28 nm CMOS process. The entire PEMAN system has been validated through cosimulations using Lumerical Interconnect and Cadence Spectre for the photonic and electrical domain, respectively.

B. Electronic Analog Front-End

The electronic analog front-end implements the second part of the MAC operations, i.e., the accumulation, followed by the analog-to-digital conversion embedding the activation function. A single frame of the MAC operations performed by the PEMAN is depicted in Fig. 2(a). The results of N consecutive multiplications ($w_{i,n} \cdot x_n$) plus the bias term θ_i, each of them associated to a time slot of length T_{MAC}, are accumulated on the Accumulation Capacitor C_A shown in Fig. 1. Index n represents the multiplication step during the accumulation, while index i represents the computed output, i.e., the i-th overall PEMAN operation. After the $(N+1)$-th T_{MAC}, i.e., after T_{ACC}, a transition of a digital signal commands the sampling of the amplifier output voltage V_{out} by the ADC. To finalize these operations, a time T_S is needed for the amplifier to reach a stable state after the last accumulation ($MAC_{i,N+1}$ in Fig. 2). Finally, the ADC, within a conversion time T_C, converts the result of the sampling operation into the digital code $D_{out,i}$, then stored in a memory. During this conversion phase NLF_i, the nonlinear function is also applied. The sum of the accumulation time, the sampling period and the conversion time determines the time needed for a whole PEMAN operation, T_{PEMAN}.

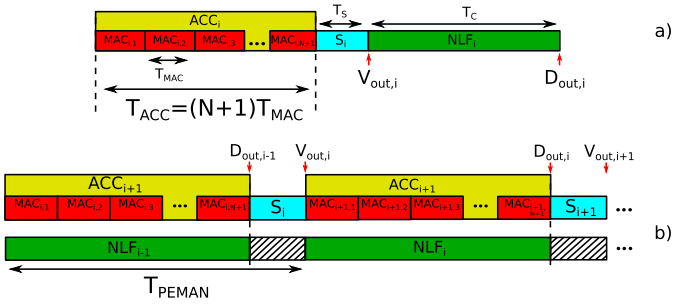

Fig. 2. Timing diagram of the PEMAN. a) A single frame of the MAC operation and the following analog-to-digital conversion. b) Time-interleaved operations of the proposed architecture.

The proposed architecture offers the possibility to time-interleave part of the operations of the analog front-end, thus allowing a lower T_{PEMAN} and a higher computational speed, without penalizing the electronic performance. In particular, once the correct reset of the accumulation capacitor and the proper sampling of the amplifier output voltage $V_{out,i}$ are guaranteed, the following accumulation ACC_{i+1} can start, while the ADC is still converting the result of the previous sampling phase S_i. With this approach, the conversion time of the ADC could be as long as the accumulation time T_{ACC} without introducing penalties on the PEMAN speed. The overall T_{PEMAN} period is then equal to the sum of

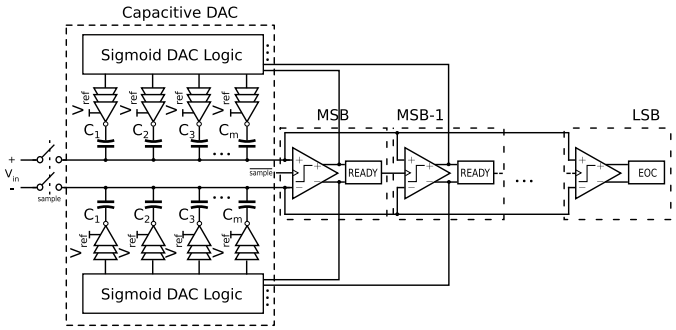

Fig. 3. Schematic view of the SAR ADC implementing the sigmoid function of the PEMAN.

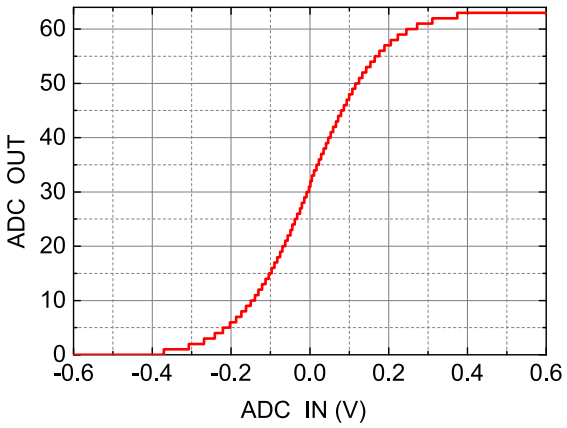

Fig. 4. DC input-output characteristic of the 6 bit, 1.4 GS/s, SAR ADC resembling the sigmoid function, obtained by means of Cadence Spectre electrical simulations.

the accumulation time $T_{ACC} = (N + 1)T_{MAC}$ and the sampling period T_S. It is worth noting that a larger number of accumulations relaxes the design of both the differential amplifier and the ADC. Compared to the minimum achievable $T_{PEMAN} \sim T_{ACC}$, the only overhead introduced by the proposed solution is T_S, which cannot be avoided in any electronic front-end to allow the correct settling before the analog-to-digital conversion. Nevertheless, the larger is N, the lower is the impact of the sampling period on the PEMAN speed. Similarly, a large number of accumulations allows an extended conversion time for the ADC.

The maximum number of accumulations allowed by the PEMAN is related to the maximum photodiode current, the accumulation capacitor C_A, its maximum voltage swing V_C, and the integration time. For the sake of simplicity, let us consider an integration time equal to T_{MAC} and a photodiode current constant during the whole period T_{ACC}. Considering the maximum photodiode current, as in the case of maximum input x_n and maximum weight $w_{i,n}$, we will obtain the worst-case estimate of the maximum number of accumulations. The minimum value of the accumulation capacitor C_A is given by the parallel of the photodiodes' parasitic capacitances, which could be as low as few hundreds of femtofarads as in monolithically-integrated solutions [50], and the switch parasitic capacitances. Due to the non-linearity of these capacitances, a large V_C swing may cause severe harmonic distortions. For this reason, additional linear capacitors (Metal-Insulator-Metal or Metal-Oxide-Metal capacitors, typically present in current commercial CMOS processes) with capacitances of the order of picofarads or tens of picofarads can be added in parallel, increasing the overall value of C_A and the linearity of the analog front-end, with relatively small impact on the area occupation. Consequently, the voltage headroom of V_C is mainly limited by the supply voltage of the front-end, which is close to 1 V in modern CMOS processes. Therefore, we considered a value of $C_A = 20$ pF and $V_{C,max} = 0.5$ V (the maximum voltage swing of V_C is half of the supply voltage). Anticipating some values obtained from the numerical simulations described in Sec. IV-B, a maximum photodiode current of 1 mA and a T_{MAC} of 50 ps, corresponding to half period at 10 GHz, are here used to estimate the maximum number of accumulations $N \sim (V_{C,max}C_A)/(I_{max}T_{MAC}) \sim 200$. Notably, with other parameters unaltered, this value increases linearly with the MAC rate. Considering the high speed of the photonic operations (tens of GHz) and a flexible range of N from few tens to hundreds, the sampling frequency of the ADC is on the order of 100 MS/s - 1 GS/s. Given the constraints of speed and resolution, a feasible and energy efficient solution is represented by the Successive Approximation Register (SAR) converter, depicted in Fig. 3. In particular, a loop unrolled topology [51] has been chosen due to the improved feedback delay, which guarantees a higher sampling rate compared to the conventional SAR topology. The fully-differential architecture brings several advantages in terms of improved linearity, common-mode noise rejection, and SAR algorithm efficiency. For this reason, the ADC is preceded by the differential amplifier that converts the unipolar voltage V_C into the differential voltage V_{out}. The presence of N_0 different comparators, where N_0 is the nominal resolution, allows the intrinsic speed-up of the chosen topology by removing the digital delay to store each comparison result, as well as the comparator reset time. At the same time, it has some unavoidable drawbacks, namely an increase of area consumption, and the need of additional hardware overhead for the offset calibration.

The nonlinear function of the neuron is embedded inside the capacitive DAC of the SAR converter. Instead of employing the typical binary weighting, an ad-hoc weighting strategy has been developed, obtaining a sigmoid transfer function, as shown in Fig. 4. The system has been designed with a standard 28 nm bulk CMOS process, with a nominal resolution of 6 bit and a max sampling frequency of 1.4 GS/s, and simulated by means of Cadence Spectre. The non-linear encoding requires additional logical circuitry, which slightly increases the delay time of the critical path, and the number of capacitors and switches, if the weighting and thus the nonlinear function needs to be flexibly reprogrammed. Nevertheless, it is possible to keep the total capacitance of the DAC (and consequently the area and the power consumption of the ADC) sufficiently low by exploiting parallel connections of several capacitors at the same depth level of the algorithm.

IV. The Photonic Linear Engine

The PEMAN optical engine has been simulated on a commercial silicon photonic platform, namely Imec

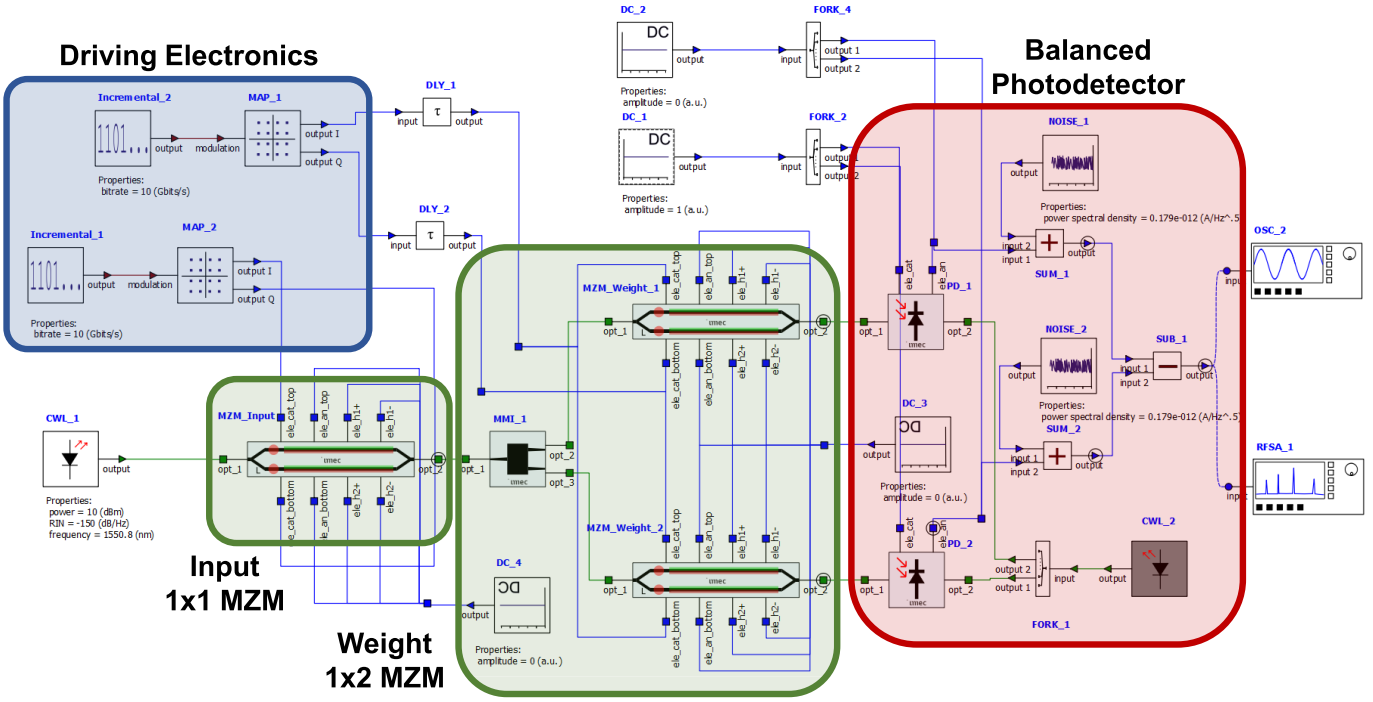

Fig. 5. The Lumerical INTERCONNECT simulation setup encompassing: input and weight MZMs (Green), the driving electronics (Blue) and balanced photodetector (Red).

iSiPP50G [49]. This section details the photonic engine and reports the results of numerical simulations performed to validate its performance. The first part focuses on the simulation setup, the second part presents the validation of the whole photonic engine.

A. Analog Photonic Engine Simulation Setting

The simulations for the analog photonic engine were performed in the Lumerical INTERCONNECT environment. The Lumerical simulative setup is depicted in Fig. 5, consisting of a CW laser, waveguides, TW MZMs and two PDs. All elements are compact models from the imec iSiPP50G library with the exception of the CW laser, which is a customized model.

The MZMs have the following characteristics: 2.5 mm-long electro-optic phase shifters, $V_\pi = 3.6$ V, and a free spectral range of 14.5 nm. In all simulations, the phase shifters are driven in a push-pull configuration within the range $[0, V_\pi]$. The MZMs are unbalanced in order to match the driving voltage ends with the representation ends: the minimum (maximum) value is encoded by applying a null (V_π) voltage to the upper arm and a V_π (null) voltage in the lower arm of the MZM. Regarding the 1×2 MZM, to obtain 0 both arms are driven with $V_\pi/2$, thus resulting in a theoretically null current at the balanced PD. To simulate the 1×2 weighting TW MZM, two 1×1 were used instead, identical to the input one, connected with a 1×2 multimode interferometer (MMI) as no built-in TW MZM with two outputs is present in the library. This is depicted by the right-most green rectangle in Fig. 5.

The iSiPP50G does not provide a built-in model for a balanced photodetector. As represented in the red rectangle of Fig. 5. To simulate the behaviour of a balanced PD, two PDs were instantiated and their output photocurrent subtracted. The following noise sources have been taken into account in the simulations: -150 dB/Hz laser relative intensity noise (RIN), 1 MHz laser linewidth, thermal noise (Temperature 300 K), dark current and shot noise in the PD. TW phase shifters allow taking into account the related delays and distortions.

B. PEMAN Validation

In this section we discuss the simulations carried out to evaluate the photonic engine composed by the CW laser, a 1×1 MZM, a 1×2 MZM and the balanced PD. An equivalent noise interval has been derived by means of random-valued multiplications, aimed to assess the ENOB of the photonic circuit. Sec. II-B discussed the derivation of ENOB from the noise standard deviation. Using the same rationale the multiplication error standard deviation has been used to compute an equivalent noise interval (equal to 6σ) taking into account distortions and bandwidth limitations, thus evaluating the system resolution.

Simulations have been performed at first as a function of multiplication frequency (from 1 GHz to 56 GHz) to evaluate the impact of the MZM finite bandwidth on the ENOB. These simulations have been carried out at a fixed laser power of 10 dBm. Subsequently, simulations aimed to assess the influence of the laser power on the ENOB are reported, varying the input power from 0.05 mW (-13 dBm) to

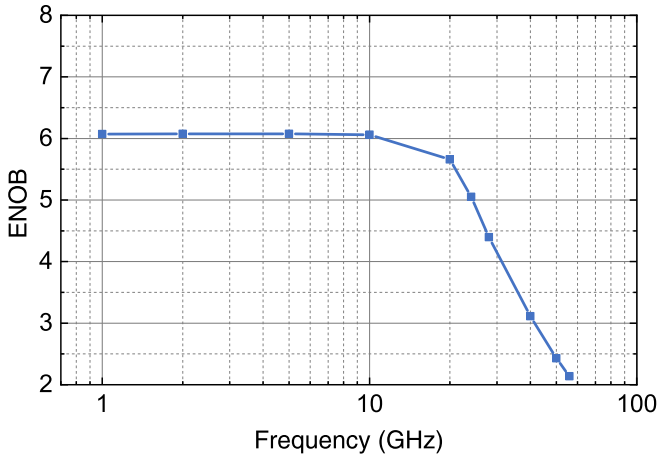

Fig. 6. PEMAN accuracy in ENOB as a function of the MAC rate.

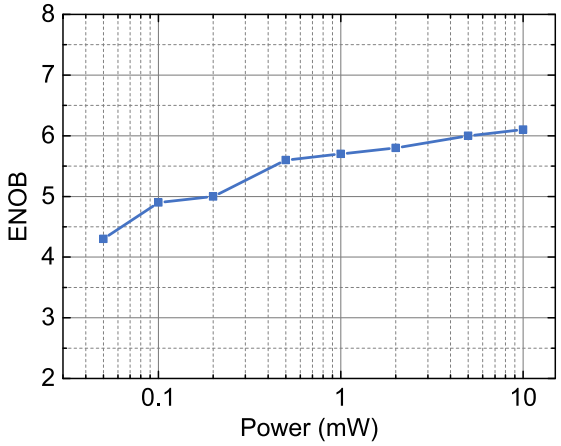

Fig. 7. PEMAN accuracy in ENOB at as a function of the CW laser power.

10 mW (10 dBm) at a constant multiplication rate of 10 GHz. The PEMAN resolution has been thus evaluated performing dot product multiplications on a dataset of random-valued input-weight pairs. The dataset has been produced using the Python library NumPy, thus generating values in the range $[0, 1]$ for x and values in the range $[-1, 1]$ for w, both rounded at the third decimal. The obtained values have been translated into the corresponding MZM voltage values by means of a nonlinear coding based on the modulator non-linear characteristic as resulting from static simulations. The simulation output returns the multiplication results as time-dependent photocurrent waveforms. The waveforms have been analyzed to extract the standard deviation σ relative to the multiplication error. These simulations have been performed with the same settings of the above, using a sampling rate equals to 256 points per period.

Fig. 6 reports the ENOB as a function of the MAC rate with a dataset of 1024 multiplications at a constant input laser power of 10 dBm. It shows a constant ENOB of 6.1 up to 10 GHz, while it decreases down to 2.1 at 56 GHz. The low-frequency plateau and the subsequent decay in the ENOB reflect the fact that the system resolution is noise-limited up to 10 GHz, while at higher frequencies it is bandwidth-limited.

Fig. 7 shows the ENOB as a function of the input laser power. These simulations were carried on a dataset of 256 random multiplications, a fixed MAC rate of 10 GHz, and all other parameters unchanged. The ENOB grows from 4.3 to 6.1 for increasing input laser power. An input power increase of 23 dB causes an ENOB increment of 1.8, lower than the maximum value of 3.8 achievable through the signal-to-noise ratio increase (SNR), according to the SNR-ENOB relation in noise limited scenarios [52]. This is due to the fact that an increased power causes higher distortions.

The results obtained on the overall PEMAN architecture are consistent with the performance of its basic MZM elements, developed for digital communications. In particular, they show that the PEMAN can trade off not only speed with resolution, but also power consumption with resolution. Moreover, the found ENOB are in line with the performance of similar devices found in the recent literature [53], [54].

Concerning the maximum number of MAC operations that can be accumulated, this number can be derived from the maximum photocurrent produced by the device, found to be 1 mA. This value is achieved when both input x_n and weight $w_{i,n}$ are equal to 1, the MAC rate is 10 GMAC/s, and the laser power is set at 10 dBm. In order to have a voltage variation ≤ 0.1 V on the accumulation capacitor V_C, so that the bias point of the photodiodes is not significantly altered, the PEMAN can accumulate ~ 200 multiplications.

V. Discussion

In this section we discuss the obtained results, focusing on the PEMAN performance and physical implementation. We aim to position the proposed photonic-electronic neuron among the current solutions based on analog electronics. An in-depth discussion on the design and benchmarking of neural network models compliant with photonic accelerators, including the PEMAN, can be found in [55].

Table I reports a comparison in terms of speed, resolution, energy consumption, footprint efficiency, and neuron fan-in of PEMAN (operating at different MAC rates) with four analog photonic and four analog electronic neuromorphic engines (HICANN [58], NeuroGrid [59], SpiNNaker [60], and TrueNorth [61]). Among the photonic neuromorphic engines reported in literature, we have chosen two architectures implementing fully-connected layers, and two spiking photonic neural networks (SPNN). The first one is the electro-absorption modulator (EAM) coherent linear neuron (COLN) [16], implementing a high-speed and compact linear neuron by means of EAM, while the second one is the semiconductor optical amplifiers (SOA) PNN [24], that exploits SOAs and a WDM encoding to perform synaptic operations (weighting and nonlinearity). The latter architectures are the resonant tunneling diode (RTD) SPNN [56], based on RTD excitable lasers to emulate spiking neurons, and the Izhikevich (IZK) inspired SPNN [57], built upon optoelectronic spiking neurons consisting of transistors and vertical cavity lasers. In some cases, photonic solutions in Table I show additional values between brackets, referring to projected estimates.

TABLE I

COMPARISON OF PEMAN WITH ANALOG PHOTONIC AND ELECTRONIC SOLUTIONS IN TERMS OF SPEED, RESOLUTION, ENERGY CONSUMPTION, FOOTPRINT EFFICIENCY, AND NEURON FAN-IN. NUMBER BETWEEN BRACKETS ARE PROJECTED VALUES, DERIVED VALUES ARE PRECEDED BY A $\approx$

Architecture	**Speed per core** GMAC/s	**Resolution** ENOB	**Energy consumption** pJ/MAC	**Photonic section energy consumption** pJ/MAC	**Footprint efficiency** GMAC/s/mm^2	**Fan-in** Inputs per neuron
PEMAN 10G	10	6.1	119	15.0	3.3	200
PEMAN 24G	24	5.1	53	11.7	8	500
PEMAN 56G	56	2.1	22	8.4	18.6	1200
EAM COLN [16]	32	≈ 2	-	0.2 (0.09)	320 (3.2×10^5)	2 (64)
SOA PNN [24]	40	≈ 2	-	≈ 30 (12)	-	4 (64)
RTD SPNN [56]	11.9	1	-	0.1	-	5
IZK SPNN [57]	$\approx 20\times10^{-6}$ (10)	1	-	10^6(0.21)	10^{-15} (2×10^3)	10
HICANN [58]	0.0224	4	198	-	51.4	64
NeuroGrid [59]	40.1×10^{-6}	13	119	-	0.9	256
SpiNNaker [60]	3.2×10^{-6}	16	6×10^5	-	-	≈ 1000
TrueNorth [61]	2.5×10^{-6}	5	26	-	1.2	256

The PEMAN has the potential to outperform its analog electronics counterparts by several orders of magnitude in terms of speed per core, while being competitive with the photonic implementations both in speed and resolution. The remarked difference in speed between the photonic and the electronic solutions is due the fact that electronic chips have privileged a distributed computation strategy. In the latter, an high number of operations is reached through the deployment of a large number processors, in the thousands range, characterized by high resolution (ENOB) and a very low speed, ranging in the hundreds or tens of Hz.

To derive the energy consumption, the PEMAN is considered with an input laser power of 10 dBm and the electronic ADC working at its maximum speed of 1.4 GS/s. In these conditions, the power consumption of every element is as follows: 81 mW for the laser source [42], $<$ 1 mW for the balanced PD, and 13 mW for the front-end electronics (amplifier and ADC). The major contribution to the power budget comes from the MZMs driving circuitry, accounting for 180 mW per high speed DAC [62], and 400 mW per RF amplifier [63]–[65]. The energy consumption for the analog electronic solutions has been evaluated by dividing the dissipated power by the number of basic elements (i.e., neurons) and by the MAC speed per processing core. To deal with photonic solutions, Table I provides an additional column for the photonic section energy consumption, which considers just the energy to run the optical elements and to charge and discharge the equivalent capacity of the MZM EO phase shifters, and excludes the energy dissipated by the analog front-end, ADC, RF drivers, and DAC, which is not typically reported. Moreover, the power dissipated by the neuron synaptic weights is not considered in the two SPNNs. The PEMAN outperforms all engines apart from the TrueNorth, EAM COLN, and RTD SPNN, the latter two being significantly constrained in terms of resolution and fan-in, as detailed in the following.

The footprint efficiency is a metric introduced in [42] and is evaluated as MAC speed per wafer area usage (GMAC/s/mm^2). The PEMAN achieves a median value for this metric, which changes significantly among the considered architectures, as it strongly depends on the integration platform (CMOS, InP, SOI) and on the basic element (EAMs, MZMs, transistors).

The last metric assessed in this comparison is the fan-in, i.e., the maximum number of inputs that a neuron can elaborate. This is an important aspect, representing the ability to implement large neural networks. The PEMAN outperforms all the photonic engines by two order of magnitudes, reaching values similar to the electronic ones.

The adopted opto-electronic approach overcomes the main electronic bottleneck caused by the dynamic power exponentially growing with the clock speed [66], trading off speed and resolution. It can reach MAC rates exceeding 50 GMAC/s while reducing the energy per MAC operation, as the static power scales down with speed. At higher MAC rates the drawback is a reduced resolution in terms of ENOB due to the finite bandwidth of the MZM elements. As reported in Sec. III-B, rates below 10 GMAC/s are not convenient, as there is no resolution improvement, while the energy per MAC increases due to the static power consumption.

The computed power consumption accounts also for the nonlinear function computation. By applying the nonlinearity while sampling, the system avoids: (i) an additional DRAM read/write, (ii) the nonlinearity computation (typically 10 arithmetic operations), saving 76 pJ every time the ADC samples, according to the energy cost of DRAM read/write 5 pJ/bit and floating point operation 0.1 pJ/bit [67].

VI. CONCLUSION

In this paper, we propose and numerically assess the performance of a precision-scalable integrated photonic-electronic multiply-accumulate neuron (PEMAN) intended for neuromorphic acceleration at low power, able to trade off speed and accuracy. The hybrid device implements the high speed multiplication stage in the optical domain, while embedding the nonlinear activation function in the analog-to-digital conversion process in the electronic domain. The numerical simulations have been performed considering Imec iSiPP50G silicon photonic platform for the implementation of the optical part and a standard 28 nm CMOS process for the electronic front-end. Photonic-electronic co-simulations show that the PEMAN

has the potential to largely outperform analog electronic equivalent solutions, in particular in terms of power consumption at large operating frequencies in excess of 10 GMAC/s and up to 56 GMAC/s, where a very low power consumption of 22 pJ/MAC is achieved. In addition, the PEMAN can flexibly adapt its operation balancing speed and accuracy needs. This power analysis accounts for all the elements, encomprising the laser, the analog electronic front-end, the DACs and the RF drivers. The PEMAN proves to be competitive also with other photonic solutions, particularly in terms of resolution and fan-in (by a hundredfold on the latter), enabling the possibility to accelerate more complex ANN models. Moreover, with its output in the digital domain by design, neurons in the same or successive layers can be fed without scalability issues.

The choice of a silicon photonics platform for implementing the integrated optical engine has been made envisioning an all-silicon implementation of the PEMAN, ideally using a common platform for the photonic and electronic sections. The technological platforms used for the implementation clearly have deep and complex techno-economic implications in terms of form factors and CAPEX/OPEX of the ANN. Limiting the considerations to the ANN power efficiency, which is the key aspect discussed in the paper, future works will focus on the driving circuitry and consider alternative platforms for implementing the photonic and the electronic parts. For example, the InP monolithic integration platform can be a promising candidate, providing all the required photonic building blocks including the laser source. Also, alternative electronic technologies like the finFET platforms, can be investigated, having the potential to improve speed and power efficiency of the PEMAN.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[2] D. Silver *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[3] D. Amodei, D. Hernandez, G. Sastry, J. Clark, G. Brockman, and I. Sutskever. (Sep. 2020). *AI and Compute*. [Online]. Available: https://openai.com/blog/ai-and-compute/

[4] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.

[5] T. B. Brown *et al.*, "Language models are few-shot learners," 2020, *arXiv:2005.14165*.

[6] Y. Emma Wang, G.-Y. Wei, and D. Brooks, "Benchmarking TPU, GPU, and CPU platforms for deep learning," 2019, *arXiv:1907.10701*.

[7] J. Misra and I. Saha, "Artificial neural networks in hardware: A survey of two decades of progress," *Neurocomputing*, vol. 74, nos. 1–3, pp. 239–255, 2010.

[8] *Intel Delivers 'Real Time AI' in Microsoft's Accelerated Deep Learning Platform*. Accessed: Jul. 24, 2021. [Online]. Available: https://newsroom.intel.com/news/intel-delivers-real-time-ai-microsofts-accelerated-deep-learning-platform/

[9] B. Rajendran, A. Sebastian, M. Schmuker, N. Srinivasa, and E. Eleftheriou, "Low-power neuromorphic hardware for signal processing applications: A review of architectural and system-level design approaches," *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 97–110, Nov. 2019.

[10] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," 2019, *arXiv:1907.10597*.

[11] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," 2019, *arXiv:1906.02243*.

[12] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the carbon emissions of machine learning," 2019, *arXiv:1910.09700*.

[13] S. Mittal, "A survey of ReRAM-based architectures for processing-in-memory and neural networks," *Mach. Learn. Knowl. Extraction*, vol. 1, no. 1, pp. 75–114, 2018.

[14] 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, pp. 1–18, Jan. 2020.

[15] R. Stabile, G. Dabos, C. Vagionas, B. Shi, N. Calabretta, and N. Pleros, "Neuromorphic photonics: 2D or not 2D?" *J. Appl. Phys.*, vol. 129, no. 20, May 2021, Art. no. 200901.

[16] G. Giamougiannis *et al.*, "Silicon-integrated coherent neurons with 32GMAC/sec/axon compute line-rates using EAM-based input and weighting cells," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, Sep. 2021, pp. 1–4.

[17] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, "All-optical machine learning using diffractive deep neural networks," *Science*, vol. 361, no. 6406, pp. 1004–1008, Sep. 2018.

[18] F. Shokraneh, S. Geoffroy-Gagnon, and O. Liboiron-Ladouceur, "The diamond mesh, a phase-error-and loss-tolerant field-programmable MZI-based optical processor for optical neural networks," *Opt. Exp.*, vol. 28, no. 16, pp. 23495–23508, 2020.

[19] L. De Marinis, G. Contestabile, P. Castoldi, and N. Andriolli, "A silicon nitride reconfigurable linear optical processor," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2021, p. Tu1C.6.

[20] L. De Marinis, M. Cococcioni, P. Castoldi, and N. Andriolli, "Photonic neural networks: A survey," *IEEE Access*, vol. 7, pp. 175827–175841, 2019.

[21] 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.

[22] I. A. D. 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, pp. 1–12, Jan. 2020.

[23] M. Miscuglio *et al.*, "All-optical nonlinear activation function for photonic neural networks," *Opt. Mater. Exp.*, vol. 8, no. 12, pp. 3851–3863, 2018.

[24] B. Shi, N. Calabretta, and R. Stabile, "InP photonic integrated multilayer neural networks: Architecture and performance analysis," *APL Photon.*, vol. 7, no. 1, Jan. 2022, Art. no. 010801.

[25] V. Bangari *et al.*, "Digital electronics and analog photonics for convolutional neural networks (DEAP-CNNs)," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 1, pp. 1–13, Jan. 2020.

[26] 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. 1, 2014.

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

[28] L. De Marinis, A. Catania, P. Castoldi, P. Bruschi, M. Piotto, and N. Andriolli, "A codesigned photonic electronic MAC neuron with ADC-embedded nonlinearity," in *Proc. Conf. Lasers Electro-Opt.*, 2021, p. AW3E.

[29] T. Bouwmans, S. Javed, M. Sultana, and S. K. Jung, "Deep neural network concepts for background subtraction: A systematic review and comparative evaluation," *J. Neural Netw.*, vol. 117, pp. 8–66, Sep. 2019.

[30] C.-H. Cheng *et al.*, "Neural networks for safety-critical applications—Challenges, experiments and perspectives," in *Proc. Design, Automat. Test Eur. Conf. Exhib. (DATE)*, Mar. 2018, pp. 1005–1006.

[31] P. Molchanov, A. Mallya, S. Tyree, I. Frosio, and J. Kautz, "Importance estimation for neural network pruning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11264–11272.

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

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

[34] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to $+1$ or -1," 2016, *arXiv:1602.02830*.

[35] *NVIDIA Tensor Cores: Versatility for HPC & AI*. Accessed: Jul. 24, 2021. [Online]. Available: https://www.nvidia.com/en-us/data-center/tensor-cores/

[36] V. Camus, L. Mei, C. Enz, and M. Verhelst, "Review and benchmarking of precision-scalable multiply-accumulate unit architectures for embedded neural-network processing," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 4, pp. 697–711, Dec. 2019.

[37] S. Garg, J. Lou, A. Jain, and M. Nahmias, "Dynamic precision analog computing for neural networks," 2021, *arXiv:2102.06365*.

[38] F. Merrikh-Bayat, X. Guo, M. Klachko, M. Prezioso, K. K. Likharev, and D. B. Strukov, "High-performance mixed-signal neurocomputing with nanoscale floating-gate memory cell arrays," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4782–4790, Oct. 2018.

[39] W. Haensch, T. Gokmen, and R. Puri, "The next generation of deep learning hardware: Analog computing," *Proc. IEEE*, vol. 107, no. 1, pp. 108–122, Jan. 2019.

[40] J. Lu, S. Young, I. Arel, and J. Holleman, "A 1 TOPS/W analog deep machine-learning engine with floating-gate storage in 0.13 μm CMOS," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 270–281, Jan. 2015.

[41] M. Paliy, S. Strangio, P. Ruiu, T. Rizzo, and G. Iannaccone, "Analog vector-matrix multiplier based on programmable current mirrors for neural network integrated circuits," *IEEE Access*, vol. 8, pp. 203525–203537, 2020.

[42] A. R. Totovic, G. Dabos, N. Passalis, A. Tefas, and N. Pleros, "Femtojoule per MAC neuromorphic photonics: An energy and technology roadmap," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 5, pp. 1–15, Sep. 2020.

[43] P. Stark, F. Horst, R. Dangel, J. Weiss, and B. J. Offrein, "Opportunities for integrated photonic neural networks," *Nanophotonics*, vol. 9, no. 13, pp. 4221–4232, Aug. 2020.

[44] L. De Marinis, M. Cococcioni, O. Liboiron-Ladouceur, G. Contestabile, P. Castoldi, and N. Andriolli, "Photonic integrated reconfigurable linear processors as neural network accelerators," *Appl. Sci.*, vol. 11, no. 13, p. 6232, Jul. 2021.

[45] J. Binas, D. Neil, G. Indiveri, S.-C. Liu, and M. Pfeiffer, "Precise deep neural network computation on imprecise low-power analog hardware," 2016, *arXiv:1606.07786*.

[46] D. A. Miller, "Perfect optics with imperfect components," *Optica*, vol. 2, no. 8, pp. 747–750, 2015.

[47] M. Giordano *et al.*, "Analog-to-digital conversion with reconfigurable function mapping for neural networks activation function acceleration," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 2, pp. 367–376, Jun. 2019.

[48] S. Tanaka *et al.*, "Ultralow-power (1.59 mW/Gbps), 56-Gbps PAM4 operation of Si photonic transmitter integrating segmented PIN Mach–Zehnder modulator and 28-nm CMOS driver," *J. Lightw. Technol.*, vol. 36, no. 5, pp. 1275–1280, Mar. 1, 2018.

[49] P. Absil *et al.*, "Reliable 50 Gb/s silicon photonics platform for next-generation data center optical interconnects," in *IEDM Tech. Dig.*, Dec. 2017, p. 34.

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

[51] T. Jiang, W. Liu, F. Y. Zhong, C. Zhong, K. Hu, and P. Y. Chiang, "A single-channel, 1.25-GS/s, 6-bit, 6.08-mW asynchronous successive-approximation ADC with improved feedback delay in 40-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 47, no. 10, pp. 2444–2453, Oct. 2012.

[52] J. Millman and C. C. Halkias, *Integrated Electronics: Analog and Digital Circuits and Systems*. New York, NY, USA: McGraw-Hill, 1972.

[53] A. N. Tait *et al.*, "Feedback control for microring weight banks," *Opt. Exp.*, vol. 26, no. 20, pp. 26422–26443, Oct. 2018.

[54] C. Huang *et al.*, "Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits," *APL Photon.*, vol. 5, no. 4, Apr. 2020, Art. no. 040803.

[55] E. Paolini, L. De Marinis, M. Cococcioni, L. Valcarenghi, L. Maggiani, and N. Andriolli, "Photonic-aware neural networks," *Neural Comput. Appl.*, Apr. 2022.

[56] M. Hejda *et al.*, "Resonant tunneling diode nano-optoelectronic excitable nodes for neuromorphic spike-based information processing," *Phys. Rev. A, Gen. Phys.*, vol. 17, no. 2, Feb. 2022, Art. no. 024072.

[57] Y.-J. Lee, M. Berkay On, X. Xiao, R. Proietti, and S. J. Ben Yoo, "Izhikevich-inspired optoelectronic neurons with excitatory and inhibitory inputs for energy-efficient photonic spiking neural networks," 2021, *arXiv:2105.02809*.

[58] J. Schemmel, D. Briiderle, A. Griibl, M. Hock, K. Meier, and S. Millner, "A wafer-scale neuromorphic hardware system for large-scale neural modeling," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2010, pp. 1947–1950.

[59] B. V. Benjamin *et al.*, "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations," *Proc. IEEE*, vol. 102, no. 5, pp. 699–716, May 2014.

[60] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The SpiNNaker project," *Proc. IEEE*, vol. 102, no. 5, pp. 652–665, May 2014.

[61] P. A. Merolla *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, 2014.

[62] B. Moeneclaey *et al.*, "A 6-bit 56-GSa/s DAC in 55 nm SiGe BiCMOS," in *Proc. IEEE BiCMOS Compound Semiconductor Integr. Circuits Technol. Symp. (BCICTS)*, Dec. 2021, pp. 1–4.

[63] (2020). *GaAs, HBT, MMIC, Low Phase Noise Amplifier, 6 GHz to 14 GHz*. Analog Devices. rev. 0. [Online]. Available: https://www.analog.com/media/en/technical-documentation/data-sheets/adl 8150achip.pdf

[64] (2020). *GaAs PHEMT MMIC Medium Power Amplifier, 17.5–25.5 GHz*. Analog Devices. [Online]. Available: https://www.analog.com/media/en/technical-documentation/data-sheets/hmc 442lc3b.pdf

[65] (2020). *GaAs HEMT MMIC Medium Power Amplifier, 55–65 GHz*. Analog Devices. rev. 0. [Online]. Available: https://www.analog.com/media/en/technical-documentation/data-sheets/hmc-abh209.pdf

[66] F. De Lima, T. Ferreira, B. J. Shastri, A. N. Tait, M. A. Nahmias, and P. R. Prucnal, "Progress in neuromorphic photonics," *Nanophotonics*, vol. 6, no. 3, pp. 577–599, 2017.

[67] D. A. Miller, "Attojoule optoelectronics for low-energy information processing and communications," *J. Lightw. Technol.*, vol. 35, no. 3, pp. 346–396, Feb. 1, 2017.