A Codesigned Integrated Photonic Electronic Neuron

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

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, namely 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, based on the iSiPP50G SOI process for the photonic engine and a commercial 28 nm CMOS process for the electronic front end, has been numerically validated through cosimulations to perform multiply-accumulate operations (MAC). 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 with resolution and power consumption, it outperforms its analog electronics counterparts both in terms of speed and power consumption, and brings substantial improvements also compared to a leading GPU.

Index Terms—Photonic-electronic Codesign, Photonic neural networks, Neuromorphic photonics, 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]. Recent 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 MW 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.

In this scenario, photonic solutions show a great potential to realize analog low-power high-throughput accelerators for machine learning [14, 15]. Photonic implementations typically exploit free-space optics, such as the diffractive architectures that make use of micromachined lenses [16], or integrated optics, e.g., the coherent solutions based on Mach-Zehnder interferometer meshes [17, 18]. 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

This paper was carried out within the Department of Excellence in Robotics and Artificial Intelligence funded by MIUR to Scuola Superiore Sant’Anna.

L. De Marinis, P. Castoldi, and G. Contestabile are with Scuola Superiore Sant’Anna, 56124 Pisa, Italy. (email: lorenzo.demarinis@santannapisa.it)

A. Catania, P. Bruschi, and M. Piotto are with the University of Pisa, Dept. of Information Engineering, 56122 Pisa, Italy.

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

to be the main limitation towards truly deep photonic neural networks [19]. While some interesting works on optical nonlinearities are emerging [20]–[22], photonics can be exploited in the short-medium term to implement just the linear computations required in ANNs, thus realizing photonic-electronic accelerators.

The DEAP (Digital Electronics and Analog Photonics), proposed in [23], is an example of such photonic-electronic neuromorphic cores derived from the broadcast-and-weight architecture [24]. 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 [25]. 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 [26], 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 memory-less layers [27]. 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.

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. IV 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. 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 is applied to the accumulation result. This last step is of paramount importance to guarantee the network a proper expressiveness, i.e., the ability to generalize what the ANN has learned to unforeseen data. The computations involved in an ANN layer composed of M neurons fed by a previous layer with N neurons are formalized as follows:

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

where O_i , ($i \in 1, \dots, M$) is the output of the i -th neuron of the layer, x_n , ($n \in 1, \dots, N$) 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 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 $O(M)$ [27]. 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 lower the burden of ANN computations, for instance to exploit hardware-constrained devices and/or to apply ANN in applications with low-latency requirements, such as safety-critical ones [28].

Several hardware and software solutions are emerging in order to meet these low memory and low computing capacity constraints. The main goal of software solutions is to develop ANNs that require less memory, relying on simpler arithmetics, 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 [29]. 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 [30]. These works report ANNs with parameters encoded with ≤ 8 bits [31] down to 1 bit in binary neural networks [32, 33]. 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 [34, 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]–[40]. Analog electronics mainly exploits fundamental circuit laws and device properties (e.g., current sum in a circuit node) to perform MAC operations [41]–[43]. 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 [44, 45]. 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) [45]–[48]. While considerable effort is put to overcome these issues [17, 49]–[52], 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, 23, 53].

B. Resolution in 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. The ENOB depends on the signal over noise and distortion ratio (SINAD), which can be in turn computed from the signal to noise ratio (SNR) and the total harmonic distortion (THD). Equations 2 and 3 state the relations between ENOB, SINAD, SNR and THD, all in dB [54].

$$10^{-\frac{SINAD}{10}} = 10^{-\frac{SNR}{10}} + 10^{+\frac{THD}{10}} \quad (2)$$

$$ENOB = \frac{SINAD - 1.76}{6.02} \quad (3)$$

C. Metrics for analog neuromorphic photonics

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; indeed, for analog systems ENOB (and SINAD) are 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. For the first time in an analog neuromorphic engine, the nonlinearity is computed within the ADC.

A. Working Principle

The photonic engine relies on two cascaded travelling wave (TW) Mach-Zehnder modulators (MZM) to act on the amplitude 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 amplitude of the incoming lightwave in a range from 0 (suppression state)

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

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 properly 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 [55]).

The photonic engine has been emulated using IMEC iSiPP50G platform [56], 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 Figure 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

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.

$(N + 1)$ -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} .

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 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 [57], and the switch parasitic capacitances. Due to the non-linearity of these capacitances, a large V_C swing may cause severe harmonic distortions, limiting the SINAD of the analog front-end. 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. 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 100 ps are here used to estimate the maximum number of accumulations $N + 1 \sim 200$. 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 [58] 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

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.

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 a larger number of capacitors. 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 iSiPP50G [56]. This section details the photonic engine and reports the results of numerical simulations performed to validate its performance. The first part focuses on the characterization of the single MZM-based weighting element. The second part presents the validation of the whole photonic engine.

Each MZM has been simulated in Lumerical Interconnect exploiting TW elements with 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. The phase shifters are driven in

Fig. 5: Weighting element performance, reporting the SNR, THD and the corresponding ENOB as a function of frequency.

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.

A. MZM-based Bipolar Multiplication

This section reports the characterization of the bipolar MZM element, i.e., the one used to encode the weight w within the interval $[-1, 1]$. As the model library provides just the 1×1 TW MZM, the 1×2 version has been implemented by adding a 1×2 multi-mode interferometer (MMI) to split the input into two 1×1 MZMs.

As discussed in Sec. II-B, the characterization of an analog processing element is done taking into account both noise (SNR) and distortions (THD), which determine the SINAD, ultimately dictating the element resolution through the ENOB. To this aim, the elements must be driven to provide a sinusoidal signal with full scale i.e., with the largest peak-to-peak voltage (V_{PP}). Given the inherent MZM sinusoidal characteristic, to obtain a sine wave it must be driven with triangular waves in push-pull configuration. 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.

Simulations have been performed at first as a function of the triangular wave 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 relate the influence of the laser power to the ENOB are reported, varying the input power from 0.05 mW (-13 dBm) to 10 mW (10 dBm) at a constant frequency of 10 GHz. 128 wave periods have been simulated with 512 points per period. The

Fig. 6: Weighting element performance, reporting the SNR, THD and the corresponding ENOB as a function of the CW laser power.

output current in the time domain has been exported and FFT-transformed using a Hanning window. SNR and THD values have been then calculated in a bandwidth from 0 to 200 GHz (excluding DC), deriving thus the SINAD and ENOB.

The results as a function of the frequency are summarized in Fig. 5, reporting the SNR, the opposite of THD, and the ENOB. The main limiting factor at low frequencies is the THD (mainly caused by the MZM nonlinear behaviour), that varies in a non-monotone fashion with a peak near 5 GHz, where the distortion is minimum and the obtained ENOB is the largest, about 7.5. Conversely, the SNR exhibits a monotonously decreasing behaviour, becoming the main limiting factor at higher frequencies. Indeed the ENOB follows the THD trend at lower frequencies, and then it is SNR-limited at higher frequencies, being ≤ 5.5 above 50 GHz.

Fig. 6 shows the ENOB as a function of the input laser power. SNR and ENOB grow for increasing power, while THD does not vary significantly. The results show the relevance of the distortion in the MZM-based weighting element: an input power increase of 23 dB causes an ENOB increment of 2, lower than the maximum value of 3.8 obtainable through the SNR increase, according to Eq. 3. This is a common pattern in analog systems, where an increased power increases proportionally both the signal and distortions, thus keeping the THD value constant. This behaviour hinders the effect of the SNR improvement on the overall SINAD and ENOB.

B. PEMAN Validation with Random-valued Multiplications

In this section we discuss the simulations done to evaluate the whole photonic engine composed by the CW laser, a 1×1 MZM, a 1×2 MZM and the balanced PD. In this case the method used for the single MZM cannot be exploited, since SNR and THD of multiple cascaded devices cannot be directly derived. Instead, an equivalent noise interval has been derived by means of random-valued multiplications, aimed to assess the ENOB of the full 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

Fig. 7: PEMAN accuracy in ENOB as a function of the MAC rate.

used to compute an equivalent noise interval (equal to 6σ) taking into account distortions and bandwidth limitations, thus evaluating the system resolution.

The PEMAN has been tested for same frequencies and input powers used for the MZM-based weighting element in Sec. IV-A, 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 static characteristic. 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 used for the weighting MZM characterisation and using 256 points per period.

Fig. 7 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. The plateau, not present in the case of the single MZM element, is due to the nonlinear coding used to drive the MZMs, which mitigates the effects of distortions.

Fig. 8 shows the ENOB as a function of the input laser power with a dataset of 256 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. Similarly to the case of the single weighting element, 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 SNR increase, according to Eq. 3. This is due to the fact that an increased power causes higher distortions.

The results obtained on the overall PEMAN architecture are

Fig. 8: PEMAN Accuracy in ENOB at as a function of the CW laser power.

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 [59, 60].

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 1.1 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 digital and analog electronics.

Tab. I reports a comparison in terms of speed, resolution, and power efficiency concerning PEMAN (operating at different MAC rates), four analog electronic solutions (HICANN [61], NeuroGrid [62], SpiNNaker [63], and TrueNorth [64]), and a leading GPGPU (NVIDIA Tesla v100 [65]). To derive the power budget, 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 [44], < 1 mW for the balanced PD, and 13 mW for the electronics (amplifier and ADC). Concerning MZMs, besides a static power consumption < 1 mW, the phase shifter energy consumption per MAC is fixed at 6.9 pJ/MAC, according to its equivalent capacity and neglecting the driving circuitry. The power consumption for the analog electronics has been evaluated by dividing the dissipated power by the number of basic elements (i.e.,

neurons which corresponds to the number of MAC operations) and by the MAC speed per processing core.

The codesigned photonic-electronic neuromorphic device has the potential to outperform its analog electronics counterparts by several orders of magnitude in terms of speed per core. Moreover, it outperforms all the electronic engines apart from the TrueNorth also in terms of power consumption. The adopted opto-electronic approach overcomes the main electronic bottleneck caused by the dynamic power exponentially growing with 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. The drawback for higher MAC rates is a reduced resolution in terms of ENOB due to the finite bandwidth of the MZM elements. As reported in Sec. IV, 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].

TABLE I: Comparison of PEMAN with analog and digital electronics solutions in terms of speed, resolution, and energy consumption.

Architecture	Speed per core GMAC/s	Resolution ENOB or bits	Energy pJ/MAC
PEMAN 10G	10	6.1	16.3
PEMAN 28G	28	4.4	10.2
PEMAN 56G	56	2.1	8.6
HICANN [61]	0.0224	4	198
NeuroGrid [62]	40.1×10^{-6}	13	119
SpiNNaker [63]	3.2×10^{-6}	16	6×10^5
TrueNorth [64]	2.5×10^{-6}	5	0.27
Tesla v100 [65]	2.93	32	20

VI. CONCLUSION

With this paper, we propose, reporting also a first performance evaluation through numerical simulations, a precision-scalable integrated photonic-electronic multiply-accumulate neuron (PEMAN) intended for neuromorphic acceleration at low power and with the ability 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 and digital 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 an extremely low power consumption of 8.6 pJ/MAC is achieved. In addition, the PEMAN can flexibly adapt its operation balancing/trading speed and accuracy needs.

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, also using a common platform for the photonic and electronic parts. 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 work will tackle the design also of 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, May 2015.
- [2] D. Silver *et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, Jan. 2016.
- [3] D. Amodei *et al.*, "AI and Compute," Sep 2020. [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, 2017.
- [5] T. B. Brown *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.
- [6] Y. E. Wang, G.-Y. Wei, and D. Brooks, "Benchmarking TPU, GPU, and CPU platforms for deep learning," *arXiv preprint arXiv:1907.10701*, 2019.
- [7] J. Misra and I. Saha, "Artificial neural networks in hardware: A survey of two decades of progress," *Neurocomputing*, vol. 74, no. 1, pp. 239–255, 2010, artificial Brains.
- [8] Intel delivers 'Real Time AI' in Microsoft's accelerated deep learning platform. [Online]. Available: <https://newsroom.intel.com/news/intel-delivers-real-time-ai-microsofts-accelerated-deep-learning-platform/>
- [9] B. Rajendran *et al.*, "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, 2019.
- [10] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *arXiv preprint arXiv:1907.10597*, 2019.
- [11] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," *arXiv preprint arXiv:1906.02243*, 2019.
- [12] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the carbon emissions of machine learning," *arXiv preprint arXiv:1910.09700*, 2019.
- [13] S. Mittal, "A survey of ReRAM-based architectures for processing-in-memory and neural networks," *Machine learning and knowledge extraction*, vol. 1, no. 1, pp. 75–114, 2019.
- [14] M. A. Nahmias *et al.*, "Photonic multiply-accumulate operations for neural networks," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 1, pp. 1–18, 2019.
- [15] R. Stabile *et al.*, "Neuromorphic photonics: 2D or not 2D?" *Journal of Applied Physics*, vol. 129, no. 20, p. 200901, 2021.
- [16] X. Lin *et al.*, "All-optical machine learning using diffractive deep neural networks," *Science*, vol. 361, no. 6406, pp. 1004–1008, 2018.

- [17] 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," *Optics Express*, vol. 28, no. 16, pp. 23 495–23 508, 2020.
- [18] L. De Marinis, G. Contestabile, P. Castoldi, and N. Andriolli, "A silicon nitride reconfigurable linear optical processor," in *Optical Fiber and Communication Conference, OFC*, 2021, pp. Tu1C–6.
- [19] L. De Marinis, M. Cococcioni, P. Castoldi, and N. Andriolli, "Photonic neural networks: A survey," *IEEE Access*, vol. 7, pp. 175 827–175 841, 2019.
- [20] J. Crnjanski *et al.*, "Adaptive sigmoid-like and prelu activation functions for all-optical perceptron," *Optics Letters*, vol. 46, no. 9, pp. 2003–2006, 2021.
- [21] I. A. Williamson *et al.*, "Reprogrammable electro-optic nonlinear activation functions for optical neural networks," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 1, pp. 1–12, 2019.
- [22] M. Miscuglio *et al.*, "All-optical nonlinear activation function for photonic neural networks," *Optical Materials Express*, vol. 8, no. 12, pp. 3851–3863, 2018.
- [23] 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, 2019.
- [24] 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, 2014.
- [25] M. Miscuglio and V. J. Sorger, "Photonic tensor cores for machine learning," *Applied Physics Reviews*, vol. 7, no. 3, p. 031404, 2020.
- [26] L. De Marinis *et al.*, "A codesigned photonic electronic mac neuron with adc-embedded nonlinearity," in *Conference on Lasers and Electro-Optics*, 2021, pp. AW3E–4.
- [27] T. Bouwmans, S. Javed, M. Sultana, and S. K. Jung, "Deep neural network concepts for background subtraction: A systematic review and comparative evaluation," *Neural Networks*, vol. 117, pp. 8–66, 2019.
- [28] X. Wang *et al.*, "In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning," *IEEE Network*, vol. 33, no. 5, pp. 156–165, 2019.
- [29] P. Molchanov *et al.*, "Importance estimation for neural network pruning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 264–11 272.
- [30] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision," in *International Conference on Machine Learning*, 2015, pp. 1737–1746.
- [31] R. Andri, L. Cavigelli, D. Rossi, and L. Benini, "Yodann: An ultra-low power convolutional neural network accelerator based on binary weights," in *2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2016, pp. 236–241.
- [32] M. Courbariaux *et al.*, "Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1," *arXiv preprint arXiv:1602.02830*, 2016.
- [33] I. Hubara *et al.*, "Quantized neural networks: Training neural networks with low precision weights and activations," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6869–6898, 2017.
- [34] "NVIDIA Tensor Cores: Versatility for HPC & AI." [Online]. Available: <https://www.nvidia.com/en-us/data-center/tensor-cores/>
- [35] J. Choquette *et al.*, "NVIDIA A100 Tensor Core GPU: Performance and innovation," *IEEE Micro*, vol. 41, no. 2, pp. 29–35, 2021.
- [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 Trans. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 4, pp. 697–711, 2019.
- [37] S. Garg, J. Lou, A. Jain, and M. Nahmias, "Dynamic precision analog computing for neural networks," *arXiv preprint arXiv:2102.06365*, 2021.
- [38] K. Berggren *et al.*, "Roadmap on emerging hardware and technology for machine learning," *Nanotechnology*, vol. 32, no. 1, p. 012002, 2020.
- [39] F. Merrikh-Bayat *et al.*, "High-performance mixed-signal neuro-computing with nanoscale floating-gate memory cell arrays," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4782–4790, 2017.
- [40] 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, 2018.
- [41] 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 Journal of Solid-State Circuits*, vol. 50, no. 1, pp. 270–281, 2014.
- [42] M. Paliy *et al.*, "Analog vector-matrix multiplier based on programmable current mirrors for neural network integrated circuits," *IEEE Access*, vol. 8, pp. 203 525–203 537, 2020.
- [43] D. Marković, A. Mizrahi, D. Querlioz, and J. Grollier, "Physics for neuromorphic computing," *Nature Reviews Physics*, vol. 2, no. 9, pp. 499–510, 2020.
- [44] A. R. Totović *et al.*, "Femtojoule per mac neuromorphic photonics: An energy and technology roadmap," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 5, pp. 1–15, 2020.
- [45] S. Banerjee, M. Nikdast, and K. Chakrabarty, "Modeling silicon-photonic neural networks under uncertainties," *arXiv preprint arXiv:2012.10594*, 2020.
- [46] P. Stark *et al.*, "Opportunities for integrated photonic neural networks," *Nanophotonics*, vol. 9, no. 13, pp. 4221–4232, 2020.
- [47] G. Wetzstein *et al.*, "Inference in artificial intelligence with deep optics and photonics," *Nature*, vol. 588, no. 7836, pp. 39–47, 2020.
- [48] L. De Marinis *et al.*, "Photonic integrated reconfigurable linear processors as neural network accelerators," *Appl. Sci.*, vol. 11, no. 6232, 2021.
- [49] T. F. de Lima *et al.*, "Primer on silicon neuromorphic photonic processors: architecture and compiler," *Nanophotonics*, vol. 9, no. 13, pp. 4055–4073, 2020.
- [50] T. F. De Lima *et al.*, "Machine learning with neuromorphic photonics," *J. Lightw. Technol.*, vol. 37, no. 5, pp. 1515–1534, 2019.
- [51] J. Binas *et al.*, "Precise deep neural network computation on imprecise low-power analog hardware," *arXiv: Computer Science/Neural and Evolutionary Computing*, vol. 1606, no. 1606.07786, pp. 0–0, 2016.
- [52] D. A. Miller, "Perfect optics with imperfect components," *Optica*, vol. 2, no. 8, pp. 747–750, 2015.
- [53] Q. Cheng *et al.*, "Silicon photonics codesign for deep learning," *Proc. IEEE*, 2020.
- [54] J. Millman and C. C. Halkias, "Integrated electronics: analog and digital circuits and systems," *McGraw-Hill*, 1972.
- [55] 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," *Journal of Lightwave Technology*, vol. 36, no. 5, pp. 1275–1280, 2018.
- [56] P. Absil *et al.*, "Reliable 50Gb/s silicon photonics platform for next-generation data center optical interconnects," in *Proc. 63rd Int. Electron. Device Meeting (IEDM)*, 2017, pp. 34.2.1–34.2.4.
- [57] 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, 2020.
- [58] T. Jiang *et al.*, "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 Journal of Solid-State Circuits*, vol. 47, no. 10, pp. 2444–2453, 2012.
- [59] A. N. Tait *et al.*, "Feedback control for microring weight banks," *Optics Express*, vol. 26, no. 20, pp. 26 422–26 443, 2018.
- [60] C. Huang *et al.*, "Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits," *APL Photonics*, vol. 5, no. 4, p. 040803, 2020.
- [61] J. Schemmel *et al.*, "A wafer-scale neuromorphic hardware system for large-scale neural modeling," in *Proc. 2010 IEEE International Symposium on Circuits and Systems*. IEEE, 2010, pp. 1947–1950.
- [62] 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, 2014.
- [63] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The SpiNNaker Project," *Proc. IEEE*, vol. 102, no. 5, pp. 652–665, 2014.
- [64] 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.
- [65] R. Smith, "NVIDIA Volta Unveiled: GV100 GPU and Tesla V100 Accelerator Announced," May 2017. [Online]. Available: <https://www.anandtech.com/show/11367/nvidia-volta-unveiled-gv100-gpu-and-tesla-v100-accelerator-announced>
- [66] T. F. De Lima *et al.*, "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, 2017.