Scaling photonic neural networks: A silicon photonic GeMM leveraging a Time-Space multiplexed Xbar

S. Kovaios, I. Roumpos, M. Moralis-Pegios, G. Giamougiannis, M. Berciano, F. Ferraro, D. Bode, S. A. Srinivasan, M. Pantouvaki, N. Pleros and A. Tsakyridis

***Abstract*— We demonstrate a silicon photonic Xbar-based general matrix multiplier (Xbar GeMM) for optical neural network (NN) applications, utilizing a hybrid time-space multiplexing scheme for supporting matrix dimensions far beyond the dimensions of the Xbar circuit. We present the operational principle of the silicon photonic accelerator that is capable of merging space and time division multiplexing techniques through the use of high-speed input and weighting nodes within a coherent *M*×*N* Xbar. The proposed scheme was demonstrated experimentally using a 2×2 Xbar that employs electro-absorption modulators (EAM) with 56 GHz bandwidth both at its input signal vector and its weight matrix modulation stages. Its experimental validation as a photonic GeMM engine was performed for 5, 10, 20, 30 and 50 GBd compute rates and was benchmarked as a NN classifier for the IRIS dataset, successfully executing a total number of 2100 products over a 2×2 matrix hardware with an accuracy up to 93.3%. All SiGe EAMs were driven by high-speed electrical signals with a peak-to-peak voltage ranging between 0.9-1.2 V, suggesting a strong potential for a photonic engine that will be capable to perform with CMOS-compatible driving voltages. Finally, we discuss the pros and cons of the proposed hybrid multiplexing scheme, concluding to a thorough system performance and energy efficiency analysis.**

***Index Terms*—coherent photonic Xbar, photonic neural networks, space division multiplexing, time division multiplexing**

I. Introduction

The development of a revolutionary computing technology that will be capable of keeping pace with the exponential rise of artificial intelligence (AI) and deep neural networks (DNNs), appears currently as an imperative solution for sustaining the advances of next generation computing systems [1]. Within this frame, the emersion of optical computing and its projection onto modern photonic integration technologies seem to hold a strong potential for shaping a highly prominent computational fabric that can harmonically combine bandwidth, energy and footprint merits [2], [3]. This builds upon the significant progress that has been witnessed in the field of integrated optical Matrix-Vector-Multiplication (MVM) layouts and linear optical circuits [4], while utilizing the unique native properties of light and photonics: their broadband analog waveform processing credentials together with numerous physical degrees of freedom that are exploitable in different multiplexing schemes.

Several demonstrations of photonic systems incorporating different multiplexing paradigms exist in the literature [5]-[25], reporting on the computation of large matrix-vector (MV) products through limited photonic hardware. Prominently, space (SDM), wavelength (WDM) and time (TDM) division multiplexing exploit spatial, spectral and time dimensions respectively, for extending the available computational space offered by photonic hardware. In more detail, SDM-based layouts consider primarily photonic meshes, in which computations are performed by synergizing cascaded stages of Mach-Zehnder interferometer (MZI) nodes [5]-[8]. This approach directly associates physical with computational space, imposing in this way scalability challenges, since larger network sizes would inevitably lead to higher insertion losses and lower fidelity performance [9],[10], thereby limiting the potential of SDM layouts to match the large neural network (NN) sizes. Similarly, WDM layouts rely on the deployment and manipulation of multiple wavelengths from the photonic hardware [11]-[14], with existing technologies failing to provide a sufficient number of optical channels compliant with the size of NN parameters. Finally, TDM schemes unfold the NN operations over the time domain by serializing multidimensional data [15]-[17], potentially enabling the calculation of arbitrarily large MV products through any dimensioned photonic NN architecture. However, this comes at the cost of increased latency [18]- an effect that becomes even

Manuscript received XX XX, 2024; revised Month XX, 2024; accepted Month XX, 2024. Date of publication Month XX, 2024; date of current version Month XX, 2024. This work was supported by the European Commission through the HORIZON Projects SIPHO-G (101017194), PARALIA (101093013) and GATEPOST (101120938). (*corresponding author:* sdkovaios@csd.auth.gr)

S. Kovaios, M. Moralis-Pegios, G. Giamougiannis¸ N. Pleros and A. Tsakyridis are with the School of Informatics, Aristotle University of Thessaloniki, 54124, Thessaloniki, Greece and the Center for Interdisciplinary Research & Innovation (CIRI-AUTH), Balkan Center, 57001, Greece (e-mail: sdkovaios@csd.auth.gr; mmoralis@csd.auth.gr; giamouge@csd.auth.gr; npleros@csd.auth.gr; atsakyrid@csd.auth.gr).

I. Roumpos is with the School of Physics, Aristotle university of Thessaloniki and the Center of Interdisciplinary Research and Innovation (CIRI-AUTH), Balkan Center, 57001, Greece (e-mail: ioroumpo@auth.gr).

M. Berciano, F. Ferraro, D. Bode, are with IMEC, Kapeldreef 75, 3001, Leuven, Belgium (email: mathias.berciano@imec.be; filippo.ferraro@imec.be; dieter.bode@imec.be).

A. Srinivasan was with IMEC, now with Xanadu Quantum Technologies, Toronto, Canada (email: ashwynsrinivasan@icloud.com)

M. Pantouvaki was with IMEC, now with Microsoft Research Center, Microsoft, Cambridge, UK (email: mpantouvaki@microsoft.com)

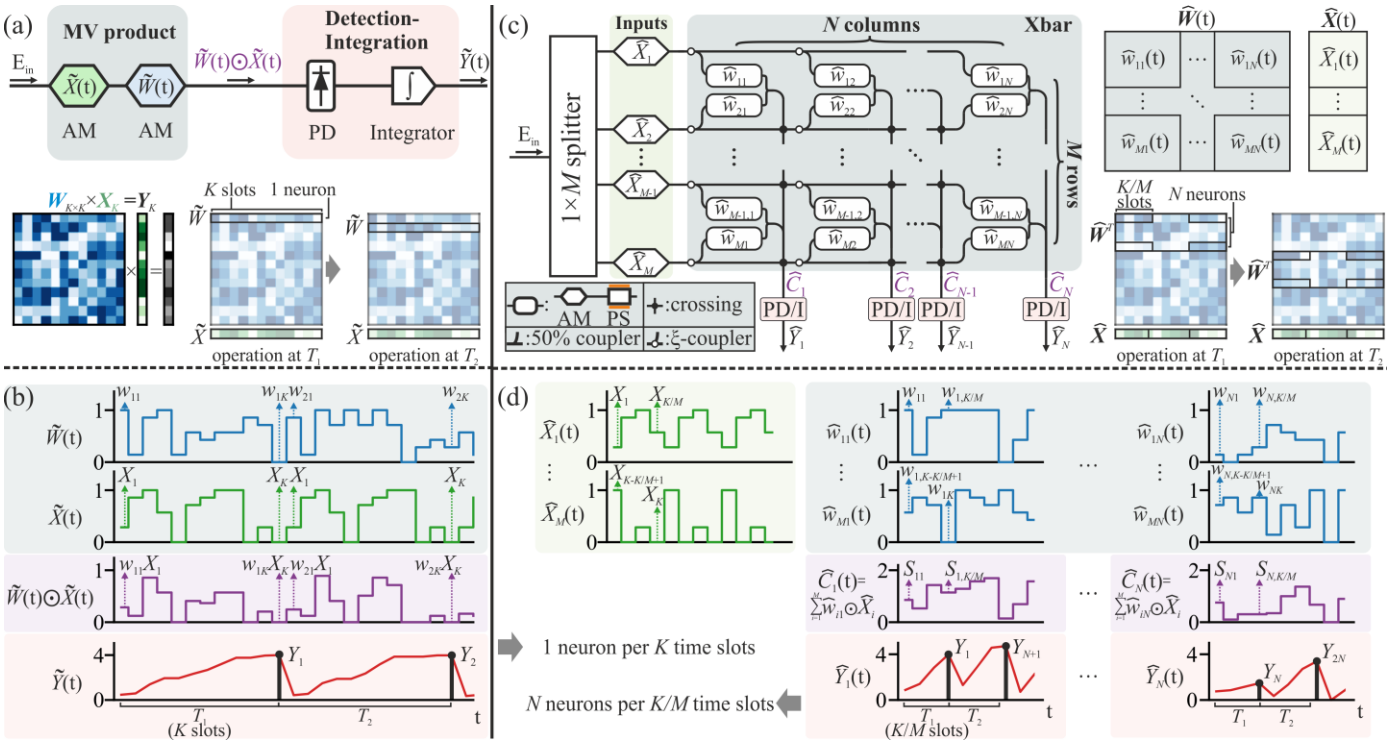

Fig. 1. (a) A fundamental TDM architecture for photonic neural networks, incorporating the MV product unit, a photodetector and a temporal integrator. (b) TDM workflow, resulting in the computation of one neuron per K time slots. (c) The $M\times N$ Xbar GeMM layout, embracing time and space division multiplexing. (d) Hybrid space-time division multiplexing workflow of the Xbar GeMM, allowing the computation of N neurons per K/M time slots.

more pronounced when slow weight update rate is utilized [21].

Reducing this latency, while maintaining the TDM credentials for supporting arbitrarily large matrix dimensions, can yield the analogous of a typical General Matrix Multiplier (GeMM) circuit in the photonic domain. This requires, however, the use of an additional multiplexing dimension, such as space and/or wavelength, to be enforced simultaneously with the TDM scheme for parallelizing computations. Although several hybrid multiplexing schemes have been reported so far for optical NN demonstrations [22], most of them are tailored only along inference operations without supporting fast weight reconfigurability and GeMM functionality.

In this paper, we present a photonic GeMM for operation up to 50 GHz by combining for the first time space and time division multiplexing over a coherent linear optical circuit. Its architecture relies on the use of the recently introduced photonic Crossbar (Xbar) architecture [10] that supports high-speed modulation both at its input and weighting stage. This allows to support two additional spatial dimensions for parallelizing operations and reducing latency, i.e. the number of Xbar rows and the number of Xbar columns. The photonic GeMM architecture was validated experimentally via a silicon photonic (SiPho) 2×2 Xbar circuit that incorporates high-speed silicon germanium (GeSi) electro-absorption modulators (EAMs) to imprint both the input vector and the weight matrix of an NN. The performance of the 2×2 Xbar GeMM is benchmarked in a NN classification task using the IRIS dataset, for compute rates ranging from 5 to 50 GBd, demonstrating a successful execution of 2100 products over a 2×2 matrix hardware with accuracies up to 93.3%. Finally, a discussion on the advantages and disadvantages of the proposed Xbar GeMM is conducted, along with a comprehensive energy efficiency analysis that reveals sub-pJ/MAC all-optical performance.

The rest of the paper is organized as follows. Section II presents the operational principles of TDM and SDM in the Xbar GeMM along with its computational workflow. Section III reports on the experimental study of the 2×2 Xbar, including DC characterization and the execution of an NN classification task. Section IV discusses the operation of the Xbar GeMM and provides a projected energy efficiency analysis for different compute rates and circuit sizes. Finally, Section V outlines the main conclusions of our work.

II. TIME AND SPACE DIVISION MULTIPLEXING

A. *Time division multiplexing*

An elementary photonic architecture supporting TDM and its computational workflow to compute large MV products of the form $\boldsymbol{Y}=\boldsymbol{W}\times\boldsymbol{X}$ are presented in Fig. 1(a) and 1(b), respectively. Two cascaded amplitude modulators, operating at the same compute rate, are imprinting the elements of the weight matrix $\boldsymbol{W}$ (of size $K\times K$) and the input vector $\boldsymbol{X}$ (of size K) of a specific neural network layer on the optical domain. The input time vector $\tilde{X}(t)$, consisting of sequential copies of the NN input vector, and the serialized weight time vector $\tilde{W}(t)$, generated by flattening the weight matrix, are fed to two modulators, producing a time vector $\tilde{W}(t)\odot\tilde{X}(t)$, where $\odot$ denotes the element-wise product between two time vectors. The generated waveform contains K^2 products $W_{ij}X_j$, each assigned to a specific time slot. After photodetection, the signal is forwarded to an electronic integrator, designed to perform the necessary accumulation operations for a single neuron, over a time span of K slots. At the output of the TDM system, the serialized MV

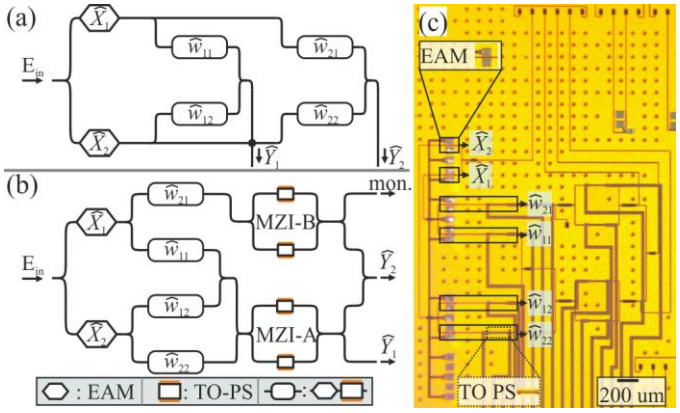

Fig. 2. (a) The 2×2 Xbar. (b) Equivalent design of 2×2 Xbar, incorporating MZIs instead of waveguide crossings. (c) Fabricated SiPho chip.

product $\tilde{Y}(t)$ is obtained, with each neuron spanning an interval of K time slots. Even though the TDM scheme allows for large MV products computations through limited photonic hardware, it inevitably increases the operation time of a single NN layer, imposing additional challenges when dealing with contemporary vast-dimensional NNs.

B. *Time and space division multiplexing with an Xbar GeMM*

To alleviate the latency-related challenges of the TDM scheme, we have incorporated the space dimension in a space-time multiplexed approach that relies on the recently introduced photonic Xbar architecture [10] shown in Fig. 1(c). The $M\times N$ Xbar, comprises a coherent layout that is capable to directly map the input NN elements (defined as $\hat{X}_i$) on its amplitude modulators (AM) and the NN weight matrix elements (defined as $\widehat{W}_{ij}$) on its computational nodes, consisting of an AM and phase modulator (PM), to imprint the amplitude and sign of the weight, respectively. Each column output is equipped with an optical receiver and an electronic integrator to accumulate the temporal partial linear summations. Xbar architecture has been designed as a loss-balanced layout by properly selecting the splitting ratio of the deployed directional couplers (ξ-couplers) at each column [10]. Therefore, the matrix accuracy representation, also known as fidelity metric, is always unity irrespective of the size and/or the matrix node loss [10], allowing in this way for the deployment of high-speed modulators both for $\hat{X}_i$ and $\widehat{W}_{ij}$.

These benefits facilitate the adoption of time-space multiplexed scheme and enable the computation of large MV products. The Xbar-based GeMM accelerates the calculations required for the computation of a single neuron by distributing the input and weight vector along the different Xbar rows. In parallel, an additional parallelization factor is provided by distributing multiple neurons at different Xbar columns. In more detail, the weights of the m-th neuron are separated into M time vectors $\widehat{w}_{mi}(t)$ and assigned to the corresponding nodes of the m-th column of the Xbar. The same segmentation and mapping is performed for the input vector, leading to the construction of M different input time vectors $\hat{X}_i(t)$, each spanning K/M time slots. At the column's output, the time series $\hat{C}_i(t) = \sum \widehat{w}_{ji} \odot \hat{X}_i$ is calculated in the photonic domain, forming a time series of the neuron's partial sums S_i. Following

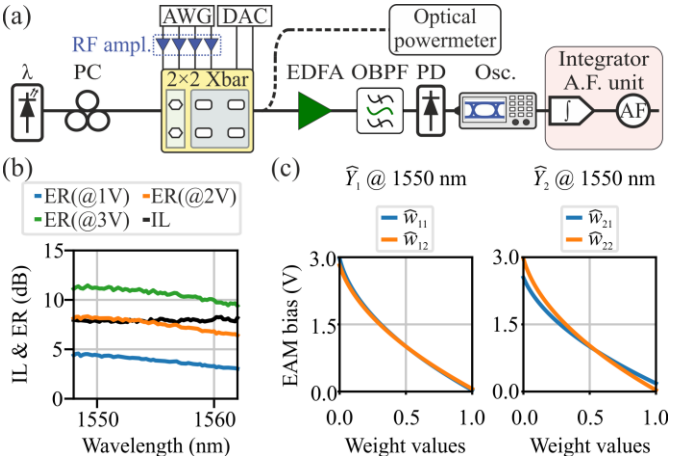

Fig. 3. (a) Experimental setup for DC and RF measurements (b) ER and IL measurements of a standalone EAM, extracted from DC measurements of the 2×2 Xbar. (c) DC weight calibration of the 2×2 Xbar.

photodetection, the remaining accumulations are performed through temporal integration, completing the computation of a single neuron in an interval of K/M time slots at the column output $\hat{Y}_i(t)$, as visually illustrated in Fig. 1(d).

Under this framework, at the first interval of computation T_1 a batch of N neurons will be computed by the Xbar GeMM within a time window of K/M time slots, yielding in this way a time acceleration factor of $M{\cdot}N$ compared to the simple TDM setup of Fig. 1(a). Upon completion, the next batch of N neurons will be imprinted over the same photonic hardware, projecting the calculation of the MV product along a continuous pipeline. As a result, the computation of the MV product involving the W matrix, will be completed over K/N computational intervals, due to the parallelized computations enabled by the N columns of the photonic Xbar. Taking into account that each interval is equal to K/M time slots, it is evident that the output of the MV product will span a total of K^2/MN time slots, significantly improved compared to the case of the TDM scheme.

III. Experimental Demonstration of the 2×2 Xbar GeMM

A. *SiPho chip*

To validate the operation of the Xbar GeMM, we consider the case of a 2×2 Xbar, with its theoretical layout presented in Fig. 2(a). In order to avoid the use of waveguide crossings (WC), that may affect the matrix fidelity performance due to crosstalk, and to allow for the characterization of the different individual components within the Xbar circuit, an alternative yet functionally equivalent design has been finally fabricated, as shown in Fig. 2(b). This design incorporates 1:2 and 2:2 multimode interferometers (MMI) instead of directional couplers, electro-absorption modulators (EAMs) for imprinting the input vector and weight matrix of a NN and thermo-optical phase shifters (TO-PS) for encoding the weight sign information. WCs could be avoided by employing an output stage of two thermo-optic symmetric Mach Zehnder Interferometers (MZI-A and MZI-B) that were both configured to operate at their cross state.

The photonic chip was fabricated in IMEC's SiPho-300 mm wafer technology [24]. The EAMs employed for the input

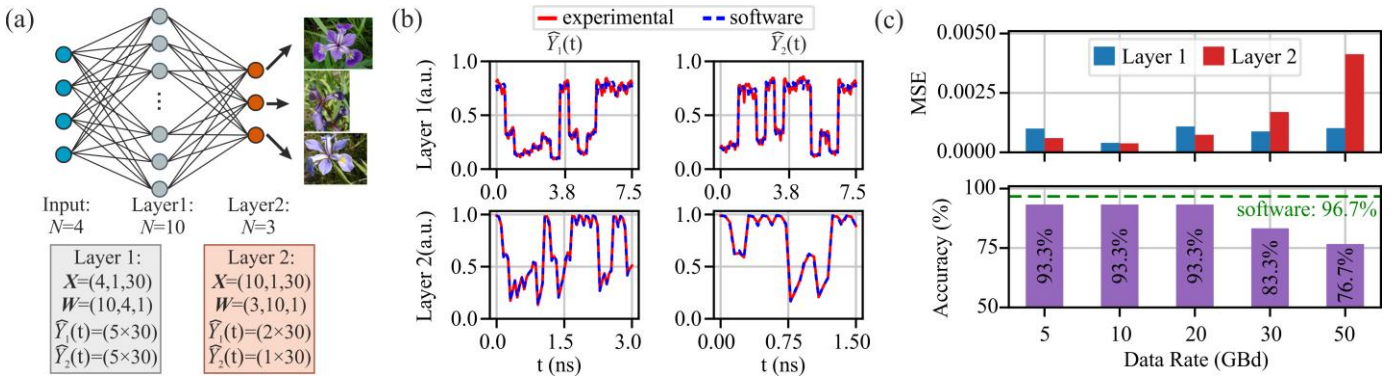

Fig. 4. (a) Fully connected NN, used for the IRIS classification task. (b) Output traces of the different NN's layers at a compute rate of 20 GBd. (c) MSE per layer and total accuracies for 5, 10, 20, 30 and 50 GBd.

$\hat{X}_i$ and weight $\hat{W}_{ij}$ modulators had a bandwidth of 56 GHz and a length of 50 um. The deployed TO PS had a length of 150 um and required a power of ~10 mW for a pi-shift. The photonic integrated circuit (PIC) is presented in Fig. 2(c), with the 2×2 Xbar covering a total area of 5.7 mm^2.

B. *Experimental setup*

The setup used for the experimental study of the Xbar GeMM is presented in Fig. 3(a). A continuous wave (CW) beam, with an output optical power of 6 dBm is injected to the PIC through a fiber array, with a polarization controller (PC) establishing the optimum coupling efficiency of the incident light beam with the integrated grating couplers. The insertion loss (IL) of a single grating coupler was measured to be 2.5 dB at 1550 nm, through reference waveguide measurements. The outputs of the chip were directed to an optical power meter for the DC measurements of the PIC. The EAMs' reverse bias voltage and all the thermo-optical elements were controlled through external digital to analog converters (DAC).

For the RF measurements, and the demonstration of the GeMM operation in NN tasks, the two input EAMs together with two weight EAMs were driven by a 4-channel Keysight M8194a arbitrary waveform generator (AWG), allowing the full dot-product operation of a single column per measurement. Four SHF-S804b electrical amplifiers were equipped at each channel of the AWG, to ensure that the high-speed analog signals were reaching the EAMs with adequate electrical power. The modulated optical output of the Xbar was directed to an erbium-doped amplifier (EDFA), and an optical band pass filter, before captured by a 70 GHz photodiode (PD). The received signal was recorded by a 72 GHz Keysight N1046A sampling scope with both the integrator and the activation function (AF) units implemented in software, to complete the operation of a single NN layer.

C. *DC characterization - calibration*

The characterization of the EAMs included in the 2×2 Xbar was performed through a series of DC measurements, with the indicative IL and static extinction ratio (ER) as a function of wavelength presented in Fig. 3(b). By setting the two MZIs at a bar-state, the products $\hat{w}_{21}\hat{X}_1$ and $\hat{w}_{22}\hat{X}_2$ can be directed to the "monitor" and "$\hat{Y}_1$" ports respectively, permitting the extraction of the IL and ER figures. Both MZIs were tuned through iterative measurements at a wavelength of 1550 nm. Under the assumption that both the input and the weight EAMs present approximately the same behavior, the IL of a standalone EAM was identified to be 8 dB in the range of 1550-1560 nm. The measured ER reached approximately 5 and 11 dB for a reverse bias voltage of 1 and 3 V respectively. Even though standalone EAM measurements could provide more appropriate metrics, our implicit characterization provides an estimation of the effective performance of the computational nodes, in the presence of wavelength dependent variations of the constituent components. Based on our measurements, a wavelength of 1550 nm was selected for the RF measurements.

Furthermore, to account for the different losses of the optical paths from the input EAMs up to the output of a single column, the calibration of the 2×2 Xbar was performed, with the results presented in Fig. 3(c). Following the analysis presented in [26], we record the optical Xbar outputs for all the combinations of applied voltages, with each EAM being reverse biased at 0, 1, 2 or 3 V respectively. Through a linear regression analysis, the differential path losses within a single column can be extracted and consequently compensated by proper selection of the weight EAM driving voltage. The latter is depicted in Fig. 3(c), where a non-linear relationship between the targeted weight values and the EAM bias voltage is observed, with the minor differences occurring between the different Xbar EAM voltage curves being enforced for balancing any intra-column differential path losses.

D. *The Xbar GeMM for the IRIS classification task*

To evaluate and benchmark the performance of the 2×2 Xbar GeMM in NN applications, we consider a fully connected NN, trained for the classification of the IRIS dataset. The deployed NN, illustrated in Fig. 4(a), consists of 2 fully connected layers, with a topology of 4:10:3, where each input has a batch size of 30. Therefore, the full inference of the NN would require the computation of 1200 and 900 products for Layer 1 and Layer 2 respectively, promoting this model for benchmarking the accelerator's performance, due to its relatively large size. The GeMM output signals were captured and subsequently normalized in order to allow for their comparison with their expected software-obtained waveform counterparts. For both layers, an ideal tunable integrator is considered to perform the remaining accumulations, followed by an ideal sigmoid AF [27]. The experimental traces obtained within an indicative time window at both Xbar outputs and for both NN layers at 20 GBd

are presented in Fig. 4(b) together with the corresponding waveforms produced when executing the NN in software. As can be seen, only negligible deviations were observed between the experimental and the software-based traces, validating the operation of the photonic processor. The amplitude of the EAMs' electrical driving signals lied in the range of 0.9 – 1.2 V for both NN layers and for all compute rates used in our experiments, indicating a strong potential to operate under CMOS compatible driving voltages.

To benchmark the performance of the Xbar GeMM, we proceed with a series of measurements for 5, 10, 20, 30 and 50 GBd rates, with the extracted MSE values per layer and NN accuracies summarized in Fig. 4(c). The low MSE values, below 0.0025 for all compute rates and layers, as well as the high experimental accuracy of 93.3%, negligibly lower than the software accuracy of 96.6%, validate the near error free operation of the proposed layout for compute rates up to 20 GBd. Increasing further the compute rate induces a small degradation to the performance of the system, which is imprinted in the increase of the MSE values and the decrease of the experimental accuracy that goes down to 76.7% when performing at 50 GBd. This behavior is expected, as the increased analog noise associated with higher compute rates translates to lower bit resolution performance [28]. Finally, it is worth noting that the SiPho processor was utilized for executing the NN inference, while the training was performed offline. The rationale behind this selection is two-fold: (a) inference process is, generally, more power consuming than training because of the multiplicative factor of using the deployed system many times [29], while training, even if it involves repetitions, is performed only once. For example, in modern large language models, inference workloads are estimated to consume 25×-1386× higher power than training [30]. Therefore, the energetic savings of photonic accelerators would be much more manifest when inference process is targeted. (b) Inference models work quite well with 4–8-bit resolution and sometimes even down to 1-2 bits [31],[32], a performance that PNNs could handle by trading off the computational rate [28]. On the other hand, the precision requirements during the training process could reach up to 8-16 bits [33],[34] which is rather challenging to achieve with current photonic technology. On top of that, during the training process there are additional operations, beyond matrix multiplication, such as loss function and gradient calculation that, are still immature in the optical domain. However, given the continuous advancements in optical technologies and the crossbar's capability to perform NN inference operations with high accuracy in the GHz regime, we anticipate that in the near future, our architecture could even be leveraged to accelerate the training process.

IV. Discussion

We have experimentally demonstrated a robust framework that exploits hybrid time-space division multiplexing for enabling the computation of large MV products within a reduced latency envelope. The proposed scheme simultaneously accelerates and parallelizes the computation of all multiply operations encountered in MV products, requiring just a few accumulations to be performed through temporal integration at the circuit output. The use of the integrator allows for the employment of lower speed and as such lower power analog-to-digital converters (ADCs) at the Xbar outputs [35], contributing to reduced energy consumption.

Moreover, the proposed Xbar GeMM has the potential to form a CMOS compatible solution that can be directly adapted to an integrated opto-electronic system, thanks to the low driving voltage requirements of the incorporated EAMs. In particular, the EAMs provide a static ER of 5 dB at 1550 nm at 1 V reverse bias voltage. This feature is also supported by the supplemented RF electrical power measurements since the amplitude of the required EAM electrical driving signals ranged between 0.9 and 1.2 V when measured just before being applied to the photonic chip. These voltage requirements assert a strong potential for operating with CMOS driving voltages, eliminating the need for on-chip RF amplifiers. This is expected to provide an additional boost to energy efficiency and to reduce the system complexity of scaled up architectures. In addition, EAMs serve as an optimal modulator choice for scalable Xbar GeMM layouts compared to other technologies [36],[37]. Their small footprint (in contrast with Mach Zehnder modulators), combined with the high bandwidth credentials, offer the potential of realizing scaled up Xbar layouts, whereas their relatively low IL (compared to other high-speed modulators such as plasmonic organic modulators) and low complexity configuration (compared to exhaustive external thermal control of micro-ring modulators), currently place EAMs as the preferred technology for amplitude modulation in Xbar GeMMs. An illustration of the envisioned Xbar GeMM is presented in Fig. 5(a), including all the necessary optical and electronic components.

The energy efficiency of the Xbar GeMM can be theoretically calculated both for different Xbar dimensions as well as for different computing rates, following the theoretical analysis presented in [37]. We consider square Xbar layouts, of size $N\times N$, operating at 1550 nm, comprising directional couplers with splitting ratio ξ_i for the inter-column splitting, a binary tree of 1:2 MMIs for the initial splitting stage and 2:1 MMIs at the recombination stages of each Xbar column. A requirement of 4-bit resolution for the peripheral electronics is assumed, whereas the integrator and the ADCs are considered to operate at a rate of a few hundred MHz, implying that the targeted MV products exceed the size of the Xbar architecture by at least an order of magnitude.

The energy efficiency e of the Xbar GeMM can be decomposed into two contributions: the optical energy efficiency, accounting for the power consumption of the laser and the opto-electronic components, and the electrical efficiency, accounting for the power consumption of the peripheral electronics.

$$e = \frac{total\ power\ consumption}{\#\ MAC/s} = e_{opt} + e_{el} \qquad (1)$$

The optical efficiency e_{opt} is defined as:

$$e_{opt} = \frac{P_{laser} + NP_{EAM-in} + N^2(P_{EAM-weight} + P_{PS})}{CR \cdot N^2} \qquad (2)$$

where P_{laser}, P_{EAM-in}, P_{PS} and $P_{EAM-weight}$ are the laser, input EAM, weight PS and weight EAM electrical power

consumption values, respectively, with CR denoting the deployed compute rate. The power consumption of the laser can be directly extracted from the required output optical power I_{laser} as P_{laser}=I_{laser}/a, where a is the wall plug efficiency of the laser. The optical laser power required for the circuit operation is equal to:

$$I_{laser}[dBm] = s_{R_x} + IL_{Xbar} \tag{3}$$

with s_{R_x} being the receiver sensitivity and IL_{Xbar} being the Xbar insertion loss [10], which is defined as:

$$IL_{Xbar}[dB] = 2IL_{EAM} + IL_{PS} + IL_{\xi} - 10\log_{10}\xi_1^2 + \left(\frac{N}{2} - 1\right)IL_X + (2\log_2 N)IL_{MMI} \tag{4}$$

where IL_{EAM}, IL_{PS}, IL_{ξ}, IL_{MMI} and IL_X are the insertion loss of a single EAM, a single PS, a directional coupler, an MMI and a waveguide crossing respectively, and ξ_1 is the splitting ratio of the implemented directional couplers that ensures a loss-balanced Xbar layout as thoroughly analyzed in [10].

Finally, the power consumption of a standalone EAM can be divided in a dynamic and a static part [38],[39] and set equal to:

$$P_{EAM} = CR \cdot \frac{CV_{dyn}^2}{4} + \frac{1}{2}R \cdot V_{stat} \cdot I_{EAM} \tag{4}$$

where C is the capacitance of the EAM, V_{dyn} and V_{stat} are the dynamic and the static bias voltage of the EAM respectively, R is the responsivity of the EAM and I_{EAM} is the optical power reaching the EAM. Since both input and weight EAMs are operating at the same high compute rate, the only factor that differs is the input optical power, which can be defined for each case as:

$$I_{EAM-in} = I_{laser}[dBm] - 10\log_{10} N - (\log_2 N)IL_{MMI} \tag{5}$$

$$I_{EAM-weight} = I_{EAM-in} - IL_{EAM} - IL_{\xi} - 10\log_{10}\xi_1^2 \tag{6}$$

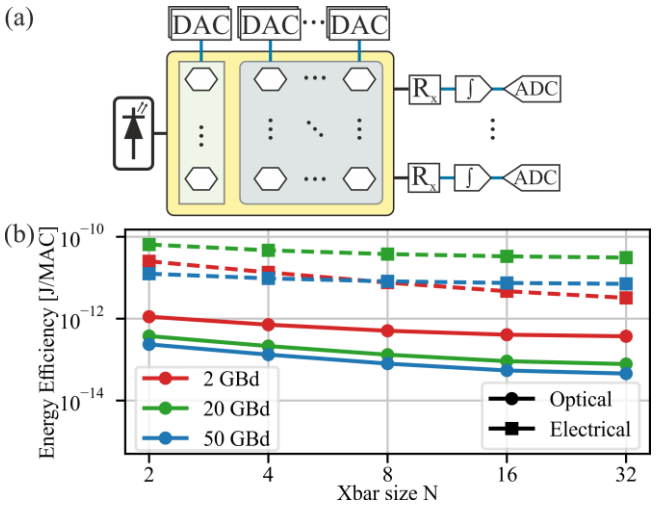

Fig. 5. (a) Perspective of the system layout of the Xbar GeMM (b) Energy efficiency of the N×N Xbar GeMM as a function of Xbar size N for 2, 20 and 50 GBd compute rates.

Finally, the electrical energy efficiency for arbitrary bit resolutions can be calculated as:

$$e_{el} = \frac{(N^2 + N)P_{el,EAM} + NP_{el,PD/I}}{CR \cdot N^2} \tag{7}$$

where $P_{el,EAM}$ is the power consumption of the electronic hardware driving the EAMs and $P_{el,PD/I}$ is the power consumption of the electronic components following photo-detection. Considering that the optical receiver's power consumption is dictated by the transimpedance amplifier (TIA), which is followed by the low bandwidth integrator and ADC, the electrical energy efficiency is deduced to:

$$e_{el} = 2^n \frac{(N^2 + N)P_{DAC} + N(P_{TIA})}{CR \cdot N^2} + \frac{N(P_{int} + P_{ADC})}{CR \cdot N^2} \tag{8}$$

with P_{DAC} and P_{TIA} being the power consumption of the digital to analog converter (DAC) and receiver's TIA circuitry respectively, scaled to 1-bit resolution [40], while P_{ADC} and P_{int} correspond to the absolute power consumption of the ADC and integrator circuitries respectively. Equations (2) and (7) respectively allow the calculation of the energy efficiency for arbitrary compute rates and Xbar designs, with the complete list of parameters adapted in our analysis summarized in Table 1.

The optical and electrical energy efficiencies of the Xbar GeMM as a function of its size for compute rates equal to 2, 20 and 50 GBd are presented in Fig. 5(b). The Xbar GeMM demonstrates sub-pJ/MAC optical energy efficiencies for almost all examined scenarios, that can be improved by increasing Xbar size and compute rate. Specifically, the 2×2 Xbar presents optical energy efficiencies of 1.11, 0.36 and 0.23 pJ/MAC at 2, 20 and 50 GBd respectively, which scale down to 0.38, 0.08 and 0.05 pJ/MAC for the 32×32 design. The situation is different when incorporating the electrical energy efficiency: the calculated electrical energy efficiency of the 2×2 Xbar exceeds 10 pJ/MAC for all three compute rates, improving with the circuit dimensions but still being above the pJ/MAC regime even for the 32×32 case. This is mainly the result of requiring a discrete DAC for driving each input and weighting node EAM, so that the electrical energy efficiency is mainly dictated by the energy efficiency of the DAC module and converges to this value as the circuit scales to higher dimensions. This also explains why the energy efficiency at 2 GBd gets finally lower than the respective values at higher rates, since the DAC energy efficiency tends to improve with decreasing compute rates. To this end, the roadmap towards enabling sub-pJ/MAC energy efficiency in this time-space multiplexed photonic GeMM layout seems to extend along two main pillars: i) it closely

Table 1. Energy efficiency parameters.

Constant parameters		
a = 0.2	C =20 fF [38]	P_{ADC} = 8.4 mW [45]
IL_{EAM} = 8 dB	IL_{ξ} = 0.1 dB [41]	P_{int} = 2 mW [46]
V_{dyn}= 1 V	IL_{MMI} = 0.06 dB [42]	n = 4
V_{stat} = −0.5 V	IL_X = 0.02 dB [43]	
IL_{PS} = 0.3 dB [44]	P_{PS} = 0.49 mW [44]	

Compute rate dependent parameters			
CR (GBd)	s_{R_x} (dBm)	P_{TIA} (mW)	P_{DAC} (mW)
2	-18 [47]	5 @1-bit res [47]	14 @6-bit res. [49]
20	-12 [47]	5 @1-bit res [47]	144 @2-bit res. [50]
50	-10 [48]	59 @2-bit res [48]	168 @3-bit res. [51]

follows the roadmap for reducing energy consumption of high-speed DACs, but also ii) can support an additional energy saving mechanism by extending the hybrid multiplexing scheme into the 3D domain, introducing also wavelength simultaneously with the time-space multiplexing scheme [52], [53].

V. Conclusion

We have presented a photonic GeMM accelerator that relies on the Xbar architecture and operates under a hybrid space and time division multiplexing scheme. Following the description of the Xbar GeMM, its evaluation within a 2×2 experimental matrix layout was performed in a series of experiments, serving as the neural layers required for executing the classification of the IRIS dataset. The experimental evaluation of the 2×2 Xbar GeMM revealed constant accuracies of 93.3% for compute rates below 20 GBd, with 83.3% and 76.7% accuracies achieved at 30 and 50 GBd respectively. Finally, a thorough analysis of the performance and energy efficiency of the proposed photonic accelerator was performed, revealing the energy consumption of the electrical DAC as the dominant power contributing factor and highlighting the extension towards a 3D multiplexing scheme as the way for sub-pJ/MAC energy efficiencies.

References

[1]. V. Sze, Y.-H. Chen, J. Emer, A. Suleiman and Z. Zhang, "Hardware for machine learning: Challenges and opportunities," in *2017 IEEE Custom Integrated Circuits Conference (CICC)*, 2017.

[2]. Apostolos Tsakyridis, Miltiadis Moralis-Pegios, George Giamougiannis, Manos Kirtas, Nikolaos Passalis, Anastasios Tefas, Nikos Pleros; Photonic neural networks and optics-informed deep learning fundamentals. APL Photonics 1 January 2024; 9 (1): 011102.

[3]. B. J. Shastri, A. N. Tait, T. Ferreira de Lima et al., "Photonics for artificial intelligence and neuromorphic computing," Nat. Photonics 15, 102–114 (2021).

[4]. M. Moralis-Pegios, G. Mourgias-Alexandris, A. Tsakyridis, G. Giamougiannis, A. Totovic, G. Dabos, N. Passalis, M. Kirtas, T. Rutirawut, F. Y. Gardes, A. Tefas and N. Pleros, "Neuromorphic Silicon Photonics and Hardware-Aware Deep Learning for High-Speed Inference," *Journal of Lightwave Technology*, vol. 40, pp. 3243-3254, 2022.

[5]. Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund and M. Soljačić, "Deep learning with coherent nanophotonic circuits," *Nature Photonics*, vol. 11, p. 441–446, 2017.

[6]. H. H. Zhu, J. Zou, H. Zhang, Y. Z. Shi, S. B. Luo, N. Wang, H. Cai, L. X. Wan, B. Wang, X. D. Jiang, J. Thompson, X. S. Luo, X. H. Zhou, L. M. Xiao, W. Huang, L. Patrick, M. Gu, L. C. Kwek and A. Q. Liu, "Space-efficient optical computing with an integrated chip diffractive neural network," Nature Communications, vol. 13, p. 1044, 2022.

[7]. N. Youngblood, "Coherent Photonic Crossbar Arrays for Large-Scale Matrix-Matrix Multiplication," IEEE Journal of Selected Topics in Quantum Electronics, vol. 29, pp. 1-11, 2023.

[8]. H. Zhang, M. Gu, X. D. Jiang, J. Thompson, H. Cai, S. Paesani, R. Santagati, A. Laing, Y. Zhang, M. H. Yung, Y. Z. Shi, F. K. Muhammad, G. Q. Lo, X. S. Luo, B. Dong, D. L. Kwong, L. C. Kwek and A. Q. Liu, "An optical neural chip for implementing complex-valued neural network," Nature Communications, vol. 12, p. 457, 2021.

[9]. G. Giamougiannis, A. Tsakyridis, M. Moralis-Pegios, A. R. Totovic, M. Kirtas, N. Passalis, A. Tefas, D. Lazovsky and N. Pleros, "Universal Linear Optics Revisited: New Perspectives for Neuromorphic Computing With Silicon Photonics," IEEE Journal of Selected Topics in Quantum Electronics, vol. 29, pp. 1-16, 2023.

[10]. G. Giamougiannis, A. Tsakyridis, Y. Ma, A. Totović, M. Moralis-Pegios, D. Lazovsky and N. Pleros, "A Coherent Photonic Crossbar for Scalable Universal Linear Optics," Journal of Lightwave Technology, vol. 41, pp. 2425-2442, 2023.

[11]. L. Yang, R. Ji, L. Zhang, J. Ding and Q. Xu, "On-chip CMOS-compatible optical signal processor," Opt. Express, vol. 20, p. 13560–13565, June 2012.

[12]. A. Totovic, G. Giamougiannis, A. Tsakyridis, D. Lazovsky and N. Pleros, "Programmable photonic neural networks combining WDM with coherent linear optics," Scientific Reports, vol. 12, p. 5605, 2022.

[13]. X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, R. Morandotti, A. Mitchell and D. J. Moss, "11 TOPS photonic convolutional accelerator for optical neural networks," Nature, vol. 589, p. 44–51, 2021.

[14]. W. Zhou, B. Dong, N. Farmakidis, X. Li, N. Youngblood, K. Huang, Y. He, C. David Wright, W. H. P. Pernice and H. Bhaskaran, "In-memory photonic dot-product engine with electrically programmable weight banks," Nature Communications, vol. 14, p. 2887, 2023.

[15]. R. Hamerly et. al., "Large-Scale Optical Neural Networks Based on Photoelectric Multiplication," Phys. Rev. X, vol. 9, no. 2, p. 021032, May 2019.

[16]. F. Duport, B. Schneider, A. Smerieri, M. Haelterman and S. Massar, "All-optical reservoir computing," Opt. Express, vol. 20, p. 22783–22795, September 2012.

[17]. M. Nakajima, K. Tanaka and T. Hashimoto, "Scalable reservoir computing on coherent linear photonic processor," Communications Physics, vol. 4, p. 20, 2021.

[18]. M. A. Nahmias, T. F. de Lima, A. N. Tait, H.-T. Peng, B. J. Shastri and P. R. Prucnal, "Photonic Multiply-Accumulate Operations for Neural Networks," IEEE Journal of Selected Topics in Quantum Electronics, vol. 26, pp. 1-18, 2020.

[19]. A. Tait, T. Ferreira de Lima, M. Nahmias, H. Miller, H. Peng, B. Shastri, and P. Prucnal, "Silicon photonic modulator neuron," Phys. Rev. Appl. 11, 064043 (2019).

[20]. A. Tsakyridis, G. Giamougiannis, A. Totovic, and N. Pleros, "Fidelity restorable universal linear optics," Adv. Photonics Res., 3, 2200001 (2022).

[21]. G. Giamougiannis, et. al., "Neuromorphic silicon photonics with 50 GHz tiled matrix multiplication for deep-learning applications," Advanced Photonics, vol. 5, p. 016004, 2023.

[22]. Y. Bai, X. Xu, M. Tan, Y. Sun, Y. Li, J. Wu, R. Morandotti, A. Mitchell, K. Xu and D. J. Moss, "Photonic multiplexing techniques for neuromorphic computing," Nanophotonics, vol. 12, p. 795–817, 2023

[23]. G. Mourgias-Alexandris et al., "Neuromorphic photonics with coherent linear neurons using dual-IQ modulation cells," J. Lightwave Technol. 38(4), 811–819 (2020).

[24]. F. J. Ferraro, et. al., "Imec silicon photonics platforms: performance overview and roadmap," in Next-Generation Optical Communication: Components, Sub-Systems, and Systems XII, 2023.

[25]. G. Mourgias-Alexandris, M. Moralis-Pegios, A. Tsakyridis et al., "Noise resilient and high-speed deep learning with coherent silicon photonics," Nat. Commun. 13, 5572 (2022).

[26]. M. Moralis-Pegios, G. Giamougiannis, A. Tsakyridis, D. Lazovsky and N. Pleros, "Perfect Linear Optics using Silicon Photonics," arXiv:2306.17728, 2023.

[27]. G. Mourgias-Alexandris, A. Tsakyridis, N. Passalis, A. Tefas, K. Vyrsokinos and N. Pleros, "An all-optical neuron with sigmoid activation function," Opt. Express, vol. 27, p. 9620–9630, April 2019.

[28]. G. Giamougiannis, A. Tsakyridis, M. Moralis-Pegios, C. Pappas, M. Kirtas, N. Passalis, D. Lazovsky, A. Tefas and N. Pleros, "Analog nanophotonic computing going practical: silicon photonic deep learning engines for tiled optical matrix multiplication with dynamic precision," Nanophotonics, vol. 12, p. 963–973, 2023.

[29]. Desislavov, R., Martínez-Plumed, F. and Hernández-Orallo, J. (2023) 'Trends in AI Inference Energy Consumption: Beyond the performance-vs-parameter laws of Deep Learning', Sustainable Computing: Informatics and Systems, 38, p. 100857. doi:10.1016/j.suscom.2023.100857

[30]. A. A. Chien, L. Lin, H. Nguyen, V. Rao, T. Sharma, and R. Wijayawardana, "Reducing the carbon impact of generative AI inference (today and in 2035)," Proceedings of the 2nd Workshop on Sustainable Computer Systems, 2023.

[31]. I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," in Advances in neural information processing systems, 2016, pp. 4107–4115.

[32]. Y. Umuroglu, N. J. Fraser, G. Gambardella, M. Blott, P. Leong, M. Jahre, and K. Vissers, "Finn: A framework for fast, scalable binarized neural network inference," in Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays. ACM, 2017, pp. 65–74.

[33]. R. Banner, I. Hubara, E. Hoffer, and D. Soudry, “Scalable methods for 8-bit training of neural networks,” in Advances in Neural Information Processing Systems, 2018, pp. 5145–5153.
[34]. 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.
[35]. A. Sludds, S. Bandyopadhyay, Z. Chen, Z. Zhong, J. Cochrane, L. Bernstein, D. Bunandar, P. B. Dixon, S. A. Hamilton, M. Streshinsky, A. Novack, T. Baehr-Jones, M. Hochberg, M. Ghobadi, R. Hamerly and D. Englund, "Delocalized photonic deep learning on the internet’s edge," Science, vol. 378, pp. 270-276, 2022.
[36]. A. Rahim, et al. "Taking silicon photonics modulators to a higher performance level: state-of-the-art and a review of new technologies," in Advanced Photonics, vol. 3, no. 2, pp. 024003, 2021.
[37]. A. Tsakyridis, G. Giamougiannis, M. Moralis-Pegios, G. Mourgias-Alexandris, A. R. Totovic, M. Kirtas, N. Passalis, D. Lazovsky, A. Tefas and N. Pleros, "Universal Linear Optics for Ultra-Fast Neuromorphic Silicon Photonics Towards Fj/MAC and TMAC/sec/mm2 Engines," IEEE Journal of Selected Topics in Quantum Electronics, vol. 28, pp. 1-15, 2022.
[38]. M. Pantouvaki, S. A. Srinivasan, Y. Ban, P. De Heyn, P. Verheyen, G. Lepage, H. Chen, J. De Coster, N. Golshani, S. Balakrishnan, P. Absil and J. Van Campenhout, "Active Components for 50 Gb/s NRZ-OOK Optical Interconnects in a Silicon Photonics Platform," Journal of Lightwave Technology, vol. 35, pp. 631-638, 2017.
[39]. D. A. B. Miller, "Energy consumption in optical modulators for interconnects," Opt. Express, vol. 20, p. A293–A308, March 2012.
[40]. M. A. Al-Qadasi, L. Chrostowski, B. J. Shastri and S. Shekhar, "Scaling up silicon photonic-based accelerators: Challenges and opportunities," APL Photonics, vol. 7, p. 020902, February 2022.
[41]. B. Sharma, K. Kishor, A. Pal, S. Sharma and R. Makkar, "Design and simulation of ultra-low loss triple tapered asymmetric directional coupler at 1330 nm," Microelectronics Journal, vol. 107, p. 104957, 2021.
[42]. Z. Sheng, Z. Wang, C. Qiu, L. Li, A. Pang, A. Wu, X. Wang, S. Zou and F. Gan, "A Compact and Low-Loss MMI Coupler Fabricated With CMOS Technology," IEEE Photonics Journal, vol. 4, pp. 2272-2277, 2012.
[43]. Y. Ma, Y. Zhang, S. Yang, A. Novack, R. Ding, A. E.-J. Lim, G.-Q. Lo, T. Baehr-Jones and M. Hochberg, "Ultralow loss single layer submicron silicon waveguide crossing for SOI optical interconnect," Opt. Express, vol. 21, p. 29374–29382, December 2013.
[44]. Q. Fang et al., "Ultralow Power Silicon Photonics Thermo-Optic Switch with Suspended Phase Arms," in IEEE Photonics Technology Letters, vol. 23, no. 8, pp. 525-527, April15, 2011, doi: 10.1109/LPT.2011.2114336.
[45]. C. Briseno-Vidrios, D. Zhou, S. Prakash, Q. Liu, A. Edward, E. G. Soenen, M. Kinyua and J. Silva-Martinez, "A 44-fJ/Conversion Step 200-MS/s Pipeline ADC Employing Current-Mode MDACs," IEEE Journal of Solid-State Circuits, vol. 53, pp. 3280-3292, 2018.
[46]. B. Pankiewicz and M. Madej, "Design of high frequency OTA in 130nm CMOS technology with single 1.2V power supply," in 2010 2nd International Conference on Information Technology, (2010 ICIT), 2010.
[47]. S. Saeedi, S. Menezo, G. Pares and A. Emami, "A 25 Gb/s 3D-Integrated CMOS/Silicon-Photonic Receiver for Low-Power High-Sensitivity Optical Communication," Journal of Lightwave Technology, vol. 34, pp. 2924-2933, 2016.
[48]. J. Lambrecht, et. al., "90-Gb/s NRZ Optical Receiver in Silicon Using a Fully Differential Transimpedance Amplifier," Journal of Lightwave Technology, vol. 37, pp. 1964-1973, 2019.
[49]. B. Kim, M. Cho, Y. Kim and J. Kwon, "A 1 V 6-bit 2.4 GS/s Nyquist CMOS DAC for UWB systems," in 2010 IEEE MTT-S International Microwave Symposium, 2010.
[50]. A. Nazemi, K. Hu, B. Catli, D. Cui, U. Singh, T. He, Z. Huang, B. Zhang, A. Momtaz and J. Cao, "3.4 A 36Gb/s PAM4 transmitter using an 8b 18GS/S DAC in 28nm CMOS," in 2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers, 2015.
[51]. E. Chong, S. N. Shahi, F. A. Musa, A. N. Mustafa, P. Krotnev, P. Madeira and D. Tonietto, "112G+7-Bit DAC-Based Transmitter in 7-nm FinFET With PAM4/6/8 Modulation," IEEE Solid-State Circuits Letters, vol. 5, pp. 21-24, 2022.
[52]. A. R. Totović, G. Dabos, N. Passalis, A. Tefas and N. Pleros, "Femtojoule per MAC Neuromorphic Photonics: An Energy and Technology Roadmap," IEEE Journal of Selected Topics in Quantum Electronics, vol. 26, pp. 1-15, 2020.
[53]. A. Totovic, C. Pappas, M. Kirtas, A. Tsakyridis, G. Giamougiannis, N. Passalis, M. Moralis-Pegios, A. Tefas and N. Pleros, "WDM equipped universal linear optics for programmable neuromorphic photonic processors," Neuromorphic Computing and Engineering, vol. 2, p. 024010, June 2022.

Stefanos Kovaios received his B.Sc. degree in Physics from the School of Physics of Aristotle University of Thessaloniki in 2016 and his joint M.Sc. degree in Physics from the University of Amsterdam and the Vrije University of Amsterdam in 2020. His M.Sc. thesis, completed in the “Resonant Nanophotonics” group at AMOLF NWO institute, Amsterdam, concerned the spatial programming of gain in plasmonic lattices and metasurfaces. Since 2021 he is a PhD candidate at the School of Informatics of Aristotle University of Thessaloniki and a member of the WinPhos research group. His research interests focus on optical communications and integrated photonic systems.

Ioannis Roumpos received his B.Sc. degree in Physics from University of Ioannina in 2018. At the beginning of 2020 he received his M.Sc. degree on "Materials Physics & Technology" from School of Physics at Aristotle University of Thessaloniki. Since October 2020, he is a PhD candidate and member of WinPhoS research group. His research interests focus on photonic integrated circuits for optical communication and neuromorphic photonics applications

George Giamougiannis has received the Diploma degree in "Electrical and Computer Engineering" from the Aristotle University of Thessaloniki, Greece in 2017. At the beginning of 2020 he received his M.Sc. in "Communication Network and Systems Security" from the same department. Since September 2019, he is a member of the WinPhos research group working as a research assistant. His research interests focus on optical communication and neuromorphic photonics. He currently has 5 issued U.S. patents in the field of linear optics and neuromorphic photonics.

Miltiadis Moralis-Pegios received the B.Sc. degree and the M.Sc. degree in Electrical Engineering from Democritus University of Thrace, Greece, in 2011 and 2013, respectively. In 2020 he received his PhD degree in “Silicon-based Photonic Integrated Circuits and High-Capacity Switching Systems for DataCenters Interconnects” from the Department of Informatics of Aristotle University of Thessaloniki (AUTH). He has been involved on several tasks of the EU-funded projects PhoxTroT, L3MATRIX, ICT-STREAMS and MOICANA, PLASMONIAC while his research interests include large-scale switching architectures for datacenter applications, silicon photonic interconnects for datacenter and high-performance computing systems and neuromorphic photonics. From 2020 he is working as a Postdoctoral Researcher in the Photonic Systems and Networks group in Aristotle University of Thessaloniki.

Nikos Pleros joined the faculty of the Department of Informatics, Aristotle University of Thessaloniki, Greece, in September 2007, where he is currently serving as an Associate Professor. His research interests extend along a broad range of photonic technologies and their use for communications, computing and sensing, including photonic neural networks, optical RAMs, optical interconnects, photonic integrated circuit

technologies, optical switching and fiber-wireless for 5G mobile networks. He has more than 350 archival journal publications and conference presentations including several invited contributions, while his work has been cited >4700 times (GS). He holds 3 US Patents in the fields of photonic biosensing and neuromorphic photonics. He has held positions of responsibility at several major conference committees including ECOC, OFC and SPIE Photonics West and has coordinated several FP7 and H2020 European projects. He has received the 2003 IEEE Photonics Society Graduate Student Fellowship, the 2018 AUTH Excellence Award for his research project funding ID, the 2021 Greek Innovator Award and the 2021 AUTH Excellence Award for Innovation and Research.

Apostolos Tsakyridis has received his university degree from the Department of Electrical and Computer Engineering of University of Thessaly in 2016. At the beginning of 2019 he received M.S. in Computer System and Networking in the Department of Informatics of the Aristotle University of Thessaloniki. He completed his PhD in 2023 in the field of photonic integrated circuits for routing and computing applications and now he is a postdoctoral researcher at the same university. He currently has 7 issued U.S. patents in the field of linear optics and neuromorphic photonics.