Sub-pJ/MAC Silicon Photonic GeMM for Optical Neural Networks using a Time-Space Multiplexed Coherent Xbar

S. Kovaios[1], I. Roumpos[2], A. Tsakyridis[1], G. Giamougiannis[1], M. Moralis- Pegios[1], M. Berciano[3], F. Ferraro[3], D. Bode[3], A. Srinivasan[3,4], M. Pantouvaki[3,5], N. Pleros[1]

[1]Center for Interdisciplinary research and Innovation, School of Informatics, Aristotle University of Thessaloniki, Greece
[2]Center for Interdisciplinary research and Innovation, School of Physics, Aristotle University of Thessaloniki, Greece
[3]Imec, Kapeldreef 75, 3001 Leuven, Belgium., [4]Xanadu Quantum Technologies, Toronto, Canada, 5Microsoft Research Center, Cambridge, UK
Author e-mail address: sdkovaios@csd.auth.gr

Abstract: We present a time-space multiplexed Silicon Photonic Neural Network that acts as a General Matrix Multiply (GeMM) engine, using a 2×2 photonic Xbar prototype for demonstrating experimental results at 20GBd and an accuracy of 93.3% at an energy efficiency of 0.2pJ/MAC.

1. Introduction

Harnessing the speed and energy advantages of Photonic Neural Networks (PNNs) [1] in practical systems can only materialize after bridging the discrepancy between the number of trainable parameters supported by integrated PNNs, which currently range in the low hundreds [2], with the billion-scale number of trainable parameters required by landmark NNs [3]. In this context, recent research has focused on leveraging a plethora of optical multiplexing techniques [4]; the baseline approach employs Space Division Multiplexing (SDM), typically implemented via unitary-based coherent linear optical circuits where the computational throughput increases with the PNN dimensions [5,6]. However, higher scales are naturally associated with high optical losses and low accuracy in matrix representations, supporting only low-bandwidth weight matrix implementations. The scaling perspectives of Wavelength Division Multiplexing (WDM) PNN approaches [7] depend on the number of available wavelength resources, which can naturally not extend to the level of large NNs. Combining multiple dimensions like space, wavelength and radio-frequency can definitely allow for increased NN dimensions [8], but again all these dimensions extend along a finite range of values within practical integrated photonic circuit designs, which can hardly approach the metrics required by large NNs. The only ONN approach that can sustain even large numbers of trainable parameters relies on Time Division Multiplexing (TDM), which can virtually increase the PNN dimensions through unrolling the calculations in the time domain [9], relaxing at the same time the speed and power requirements at the egress-Analog-to-Digital Converter (ADC) stage. Having, however, to unroll the complete NN over time increases latency [10]. Adding the space dimension in hybrid TDM/SDM multiplexed layouts can only conclude to accelerated performance and execution time speedup if the SDM design can sustain fast weight reconfigurability, yielding in this way a time-space multiplexed layout that can effectively act as a GeMM unit similar to the GeMM architectures utilized in conventional electronic NN chipsets.

In this paper, we introduce a silicon photonic GeMM architecture that uses a time-space multiplexed scheme over the recently proposed MxN Coherent Xbar architecture [11], effectively accelerating the workload execution by a factor of MxN compared to TDM-based NN designs while supporting NN dimensions that are higher than the PNN circuit scale. Its proof-of-principle experimental demonstration is presented for a two-layer NN with a total number of 70 trainable parameters that gets executed over a 2x2 silicon photonic Xbar prototype with a total number of only 6 SiGe Electro-absorption modulators (EAMs), i.e. 2 EAMs at its input and 4 EAMs at its weighting matrix. Experimental accuracies of 93.3 % are reported for 5, 10 and 20 Gbaud/axon compute rates, with an energy efficiency of only 0.2pJ/MAC that can scale down to a few fJ/MAC with increasing Xbar dimensions.

2. Time and Space Division Photonic Computing Architecture

Fig. 1(a) describes a typical TDM NN scheme when acting as a perceptron, where the dot-product between an X(t) input vector and a W(t) weight vector requires the use of two cascaded amplitude modulators (AM) that are followed by a photodiode (PD) for electro-optical conversion. The K-element input and weight signal vectors form two respective time series that get modulated via the respective AM modulators, so that the output signal carries the element-wise multiplication of the two vectors when entering the PD. The resulting electrical time multiplexed product is then forwarded to an integrator, a device that performs the accumulation of the TDM signal through time integration and provides the weighted sum $Y_1(t)$, and then to the activation function (A.F.) unit that yields the perceptron output. In this scheme, the linear operations of a single *K*-input neuron will span along *K* time slots. Scaling this towards supporting K neurons instead of a single perceptron has to implement a Matrix-Vector-Multiply operation (MVM) between a KxK weighting matrix and a K-element input signal, as shown on the right side of Fig. 1(a), requiring a total duration of K^2 time slots.

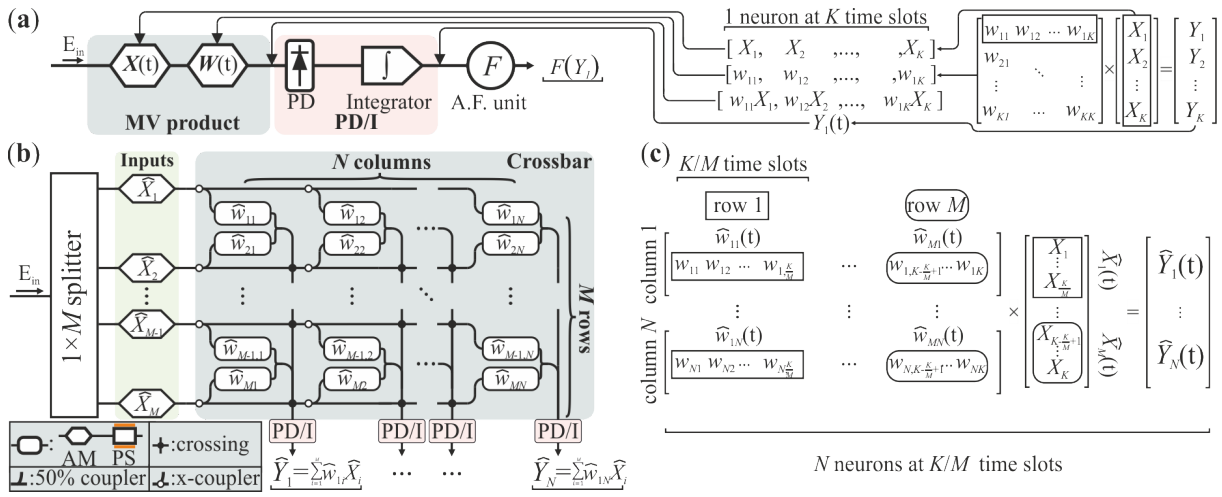

Fig. 1: (a) TDM paradigm for photonic neural networks. The inputs vector and matrix elements are recursively assigned in time slots, with each neuron spannning K time slots, and is derived through an integrator and A.F. unit. (b) STDM paradigm through the $M\times N$ photonic crossbar. Due to the summation performed between different rows of the crossbar in the optical domain, a single neuron spans K/M timeslots, whereas the operation time of a complete NN is accelerated by a factor of MN.

Incorporating the space dimension towards a space-time multiplexed scheme that relies on the photonic crossbar architecture is illustrated in Fig. 1(b). The recently demonstrated $M\times N$ crossbar [11] deploys $M\times N$ computational nodes, each consisting of an AM and a phase modulator (PM) for weighing the respective input signal prior forwarding this to a summation stage, as described in detail in [11]. Assuming the implementation of the dot-product between a K-element input signal and a K-element weight vector over the first column of the Xbar can be performed by distributing the input and weight signal vectors over the M X_i and W_{1i} modulators, so that the K-slot initial time series turns into a K/M-slot sequence. Incorporating also additional weighting vectors and distributing each vector elements over the M weight modulators within an additional Xbar column allows to implement multiple neurons simultaneously, effectively realizing the MVM operation depicted in Fig. 2(c). For a $M\times N$ Xbar size with K-element input and weight vectors, a total number of N neurons is implemented within K/M time slots, finally allowing for a speedup factor equal to MN,

3. Experimental setup and results

The proposed scheme was experimentally tested over a 2×2 silicon photonic crossbar, with the layout and the fabricated chip presented in Fig. 2(a),(b), respectively. The photonic chip was fabricated with IMEC's SiPho 300 mm wafer technology and employed high-bandwidth 50 um Franz-Keldysh electro-absorption modulators (EAMs) for the NN inputs ($\hat{X}_i$) and weights ($\hat{w}_{ij}$) imprinting and thermo-optic phase shifters (TO PS), with P_π=14 mW, for safeguarding coherent constructive inference between its constituent nodes. The insertion loss between a single column output and the common signal input was measured to be 20.2 dB. The deployed experimental setup is

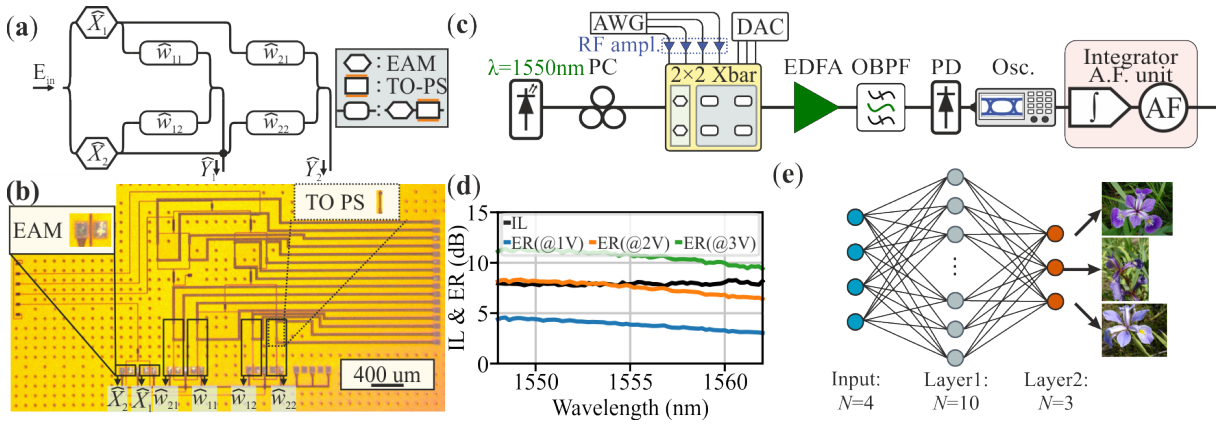

Fig. 2: (a) The 2×2 crossbar architecture (b) The fabricated 2×2 crossbar chip (c) Experimental setup (d) IL and ER of a standalone 50um FK-EAM (e) Deployed NN topology, for the IRIS classification task.

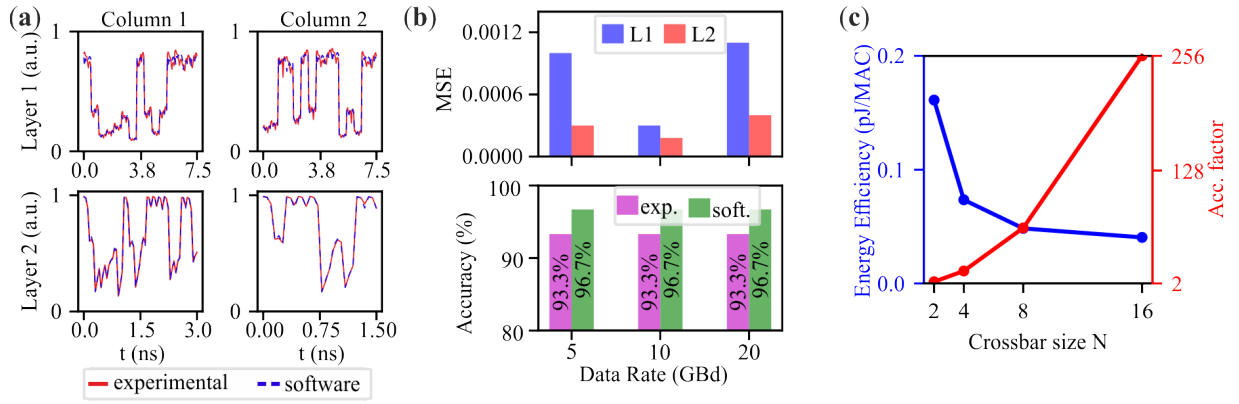

Fig. 3: (a) Experimental and theoretical output traces of L1, and L2 at 20 GBd, for the columns 1 and 2 of the 2×2 crossbar (b) MSE and accuracy for 5, 10 and 20 GBd. (c) Energy efficiency and accelerator factor of an TDM/SDM $N\times N$ Xbar, for different circuit sizes M.

presented in Fig 2(c). An optical continuous wave (CW) signal at 1550 nm is injected to the chip by a TE-grating coupler and modulated through the input EAMs. The EAMs were driven by a 38 GHz Keysight 8194a arbitrary waveform generator (AWG) with an amplitude of 100 mV_{pp} that was amplified by 67 GHz bandwidth (BW) RF amplifiers to ~0.9 - 1.0 V_{pp} which lies in the range of the CMOS technology voltage levels. A low-speed multi-DAC module was employed to tune the TO PSs. The output optical signal was amplified by an EDFA, to compensate for the acquired optical losses and filtered by an optical bandpass filter to remove the Amplified Spontaneous Emission (ASE) noise. The signal was subsequently injected in a 70 GHz photodiode, with its output recorded by a digital sampling oscilloscope. Finally, the integration of the multiplexed waveforms and the non-linear activation function was applied digitally via a software implemented routine.

Figure 2 (d) presents the experimentally measured insertion loss (IL) and extinction ratio (ER) of a stand-alone EAM, when reverse biased in the range of [-1,3] V. The reported IL in the range of 1550 nm is approximately ~8 dB while the ER equals 5dB @1V and ramps up to 11 dB for 3V reverse bias voltages. The STDM 2×2 Xbar layout was configured to implement a 4:10:3 fully connected NN that was trained to classify the IRIS data set, as depicted in Fig 2(e), with the NN inputs and weight values being imprinted onto the Xbar nodes after following the STDM multiplexing scheme explained in Fig. 1.

The experimental results are summarized in Fig. 3. The experimental waveforms collected at the output of every Xbar column for both NN layers at 20 GBd are presented in Fig. 3(a), showing an excellent qualitative agreement with the respected waveforms expected from the software-deployed NN. Fig. 3(b) presents the extracted mean squared errors (MSE) and NN accuracies for different data rates at both NN layers, where MSE values below 0.001 and experimental accuracies of 93.3 % are obtained in all cases, degraded only by 3.4% compared to the respective digital NN accuracy. Taking into account that the EAM capacitance equals 20 fF and that the use of only 1Vpp driving voltage can allow for driver-less EAM operation, the energy efficiency of a single EAM node has been calculated equal to 4.5fJ/symbol. With a laser power of 6mW at the 2×2 Xbar input and wall-plug efficiency of 20%, the overall energy efficiency of the 2×2 STDM Xbar architecture was calculated following the analysis presented in [5] and was found to be only 0.2 pJ/MAC. This can reduce dramatically as the circuit size increases, as shown in Fig. 3(c), enabling energy efficiencies much lower than 50fJ/MAC even for moderate 16×16 Xbar sizes. The same graph illustrates the speedup or acceleration factor obtained with circuit scaling, revealing that a speedup of up to 256 is obtained for a 16×16 Xbar size.

Acknowledgements:
This work was supported by the European Commission through the HORIZON projects SIPHO-G (101017194) and PARALIA (101093013).

References

[1] A. R. Totović et al, "Femtojoule per MAC Neuromorphic Photonics: An Energy and Technology Roadmap," *IEEE JSTQE,* 2020.
[2] G. Giamougiannis et al, "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.
[3] M. G. Anderson et al, "Optical Transformers," *arXiv:2302.10360,* 2023.
[4] Y. Bai et al, "Photonic multiplexing techniques for neuromorphic computing," *Nanophotonics,* vol. 12, p. 795–817, 2023.
[5] A. Tsakyridis et al, "Universal Linear Optics for Ultra-Fast Neuromorphic Silicon Photonics Towards Fj/MAC and TMAC/sec/mm2 Engines," *IEEE JSTQE,* vol. 28, pp. 1-15, 2022.
[6] W. R. Clements et al, "Optimal design for universal multiport interferometers," *Optica,* vol. 3, p. 1460–1465, 2016.
[7] A. N. Tait et al, "Demonstration of WDM weighted addition for principal component analysis," *Opt. Express,* vol. 23, p. 12758–12765, 2015.
[8] B. Dong et al, "Higher-dimensional processing using a photonic tensor core with continuous-time data," *Nature Photonics,* 2023.
[9] R. Hamerly et al, "Large-Scale Optical Neural Networks Based on Photoelectric Multiplication," *Phys.Rev.X,* vol. 9, p. 021032, May 2019.
[10] N. Youngblood, "Coherent Photonic Crossbar Arrays for Large-Scale Matrix-Matrix Multiplication," *IEEE JSTQE,* vol. 29, pp. 1-11, 2023.
[11] G. Giamougiannis et al, "A Coherent Photonic Crossbar for Scalable Universal Linear Optics," *JLT,* vol. 41, pp. 2425-2442, 2023.