Photonic Neuromorphic Computing: Architectures, Technologies, and Training Models

Miltiadis Moralis-Pegios, Angelina Totovic, George Dabos, Apostolos Tsakyridis, George Giamougiannis, George Mourgias-Alexandris, Nikolaos Passalis, Manos Kirtas, Anastasios Tefas, Nikos Pleros

Department of Informatics, Aristotle University of Thessaloniki, 54124, Thessaloniki, Greece

mmoralis@csd.auth.gr, angelina@auth.gr, {ntamposg, atsakyrid, giamouge, mourgias, passalis, eakirtas, tefas, npleros}@csd.auth.gr

Abstract: We summarize recent developments in neuromorphic photonics, including our work and the advances it brings beyond the state-of-the-art demonstrators in terms of architectures, technologies, and training models for a synergistic hardware/software codesign approach.

1. Introduction

In a rush to harness the potential of Artificial Intelligence (AI) and Deep Learning (DL), optics became a widely studied platform for computational tasks, especially as photonic integration is entering its mature stage [1]. The low-latency, -energy, and -footprint, together with tunability, allow the photonic devices to carry out Multiply-Accumulate (MAC) operations, with a predicted computational energy and area efficiency of a few fJ/MAC and >TMAC/sec/mm^2, respectively [2, 3]. To meet these goals, all constituent scientific and technological fields – from device design and architectures to DL training models – need to be considered in a synergistic co-design and co-development roadmap.

In this article we present a summary of the progress made in neuromorphic photonic building blocks, towards sustaining higher on-chip compute rates within low-power and small-size envelope, along with the associated challenges that emerge. Motivated by advances in the field of analog electronic in-memory computing, we present neuromorphic optical architectures, able to operate as universal linear operators [4], that outperform the state-of-the-art neuromorphic photonic engines in terms of compute rate efficiency and accuracy performance. We also share our recent experimental demonstrations of feed-forward coherent photonic neural networks (NNs) employed in Modified National Institute of Standards and Technology (MNIST) handwritten digit recognition dataset [5-8], unveiling how hardware-aware training models can sustain near-to-software accuracy even at high compute line-rates [9, 10].

2. Neuromorphic photonics state-of-the-art review

To be able to challenge the electronic solutions in terms of computational power and area efficiency, neuromorphic photonic engines need to, simultaneously, sustain >10 Gb/sec line-rates [1-3] and offer up to 8-bit-resolution DL environment. Figure 1(a) shows compute rate performance in MAC/sec/axon of Wavelength Division Multiplexed (WDM) and coherent architectures experimentally demonstrated within the last five years [6, 7], [11-21]. Regardless of different architectural schemes and photonic technologies, a clear discrepancy in performance can be seen between WDM architectures [11-17], which consistently operate in GHz regime, and interferometric coherent layouts [6, 7], [18-21], typically clustered around sub-MHz operating regime. However, WDM technology typically necessitates a high amount of wavelength resources for increasing fan-in and computational power [17] unlike single-channel coherent layouts which allow for more resourceful scaling. Thus far, coherent linear optics has been almost exclusively relying on multiple cascaded stages of 2x2 Mach-Zehnder Interferometric (MZI) meshes [19], a design highly

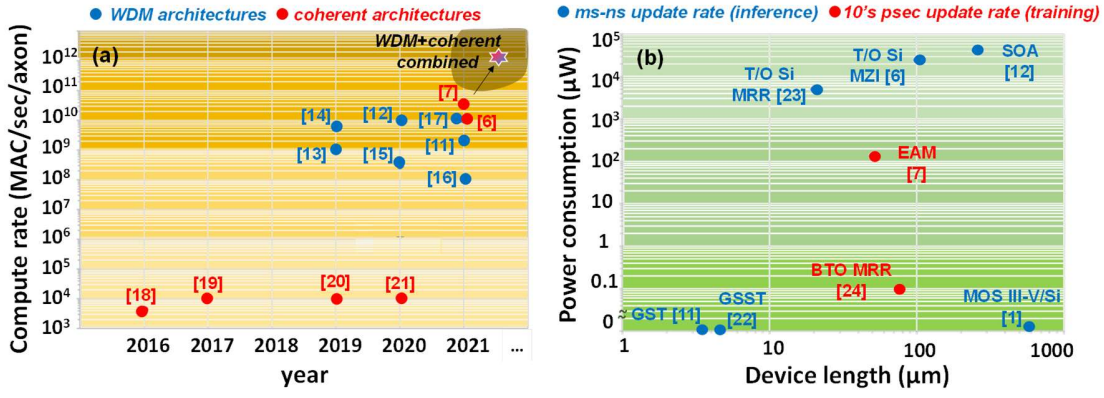

Fig. 1. (a) Compute rate per axon performance of WDM and coherent neuromorphic architectures demonstrated experimentally within the last 5 years; (b) power consumption vs. device length for the different optical weight enabling technologies, with red dots designating the potential for operating also in training applications.

sensitive to individual device loss uniformity and phase control, limiting its operational line-rates. A new paradigm, where bijective mapping between weights and photonic hardware is achieved via dual-IQ-modulator-based computational cells [5], closed the gap between the coherent and WDM solutions, elevating coherent neurons to 10 GMAC/sec/axon compute rates [6]. Migrating to SiGe Electro-Absorption Modulators (EAMs) for both on-chip data generation and weighting purposes, we accomplish the extension of on-chip compute rates to 32 GMAC/sec/axon, leveraging its high-bandwidth capabilities [7]. The employment of the, still, unused wavelength domain in coherent solutions for parallelization of operation paves the way towards >100 GMAC/sec performance per synaptic element.

Driving down the overall power consumption and footprint of the photonic NN brings the focus to weighting technology, since a typical N-input neuromorphic layout requires N^2 weights to perform N^2 MAC operations. Moreover, if in-situ training of the NN is required, the weights need to allow for sub-nsec reconfiguration times, appending an additional load on the already challenging task of simultaneously achieving low power consumption, small footprint, and low insertion losses (ILs) [1-3]. As Fig. 1(b) reveals, Phase-Change-Material (PCM) non-volatile memories are opted in inference engines [11], [22] due to their size and power efficiency; however, on-chip training can be sustained only through electro-optic (E/O) technologies, such as EAM [7] and Barium Titanate (BTO) [24] waveguides.

3. PNN architectures, technologies, and training methods

The utilization of E/O technologies to achieve high input and weight linerates inevitably induces IL penalty, which cannot be tolerated by conventional coherent neuromorphic processors that rely on 2x2 MZI meshes [18-21] employed in a Singular Value Decomposition (SVD) layout, due to the exponential scaling of their IL with the number of inputs. To tackle this challenge, we designed a novel coherent photonic crossbar architecture [4], shown in Fig. 2(a), that can support any linear transformation in the optical domain, while offering significant IL and fidelity benefits compared to SVD-based schemes. Figures 2(b) and (c) depict SiPho single column demonstrators of 4×M and 2×M M-column crossbar versions [6, 7]. More specifically, Fig. 2(b) illustrates a 4-input neuron that employs E/O MZM inputs and thermo-optic (T/O) weights [6], while Fig. 2(c) shows a 2-input neuron that exploits EAMs both for input data and weight imprinting [7].

In both cases, and generally in any analog DL photonic engine, an undesired accumulated noise, that stems from the cascaded photonic elements limits their computational performance in terms of speed and accuracy. In order to implement high-accuracy neural networks through imperfect photonic components, a properly adapted DL training framework is deemed necessary [5-7],[9-10]. We have recently presented a method that allows the inclusion of the noise of photonic components in the training of NNs. This noise-aware training method was validated via the 4-input neuron shown in Fig. 2(b), operating at 10 GMAC/sec/axon, while its accuracy-performance boost is clearly highlighted in Fig. 3(a). The solid lines correspond to the results derived from the simulation model, for different white noise levels with the standard deviation $\sigma \in [0, 0.6]$. The scattered points represent the experimentally recorded data from the baseline and noise-aware training method, respectively, and closely follow the simulation models in both cases, with the noise-aware training apporach offering 2.54% best-case accuracy improvement compared to the baseline scenario. Finally, Fig. 3(b) shows MAC/sec/axon compute rate versus accuracy reported by coherent-based demonstrations so far (black dots), revealing that the achieved experimental accuracy was 72% [20], 76.7% [19] and 90.5% [21], with the compute rate per axon not exceeding 10 kHz. With the employment of single columns of the crossbar (Fig. 2(b),(c)) for the hardware implementation of an NN, we achieved a record-high compute rates at 10 and 32 GMAC/sec/axon with the 4- and 2-input neurons, respectively, while the corresponding experimental accuracy

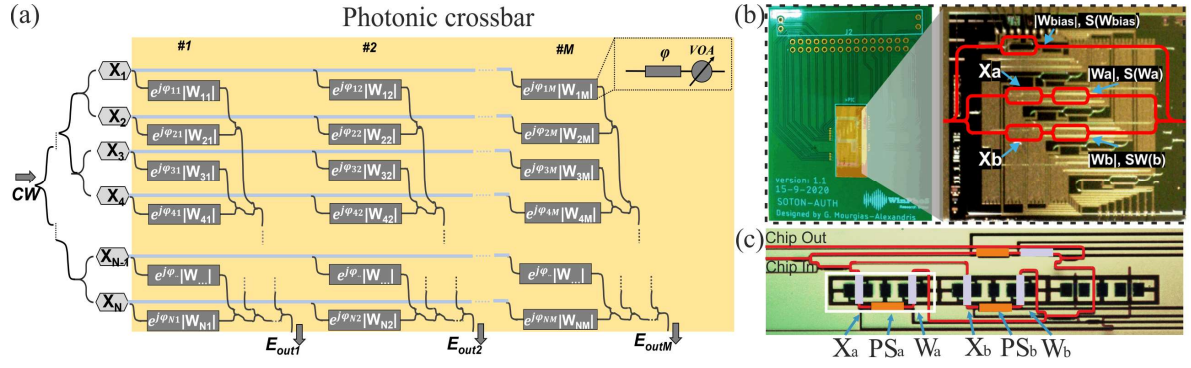

Fig. 2. (a) The photonic crossbar architecture, with the inset illustrating a weighting module that is composed of an optical phase shifter followed by a Variable Optical Attenuator (VOA). (b) 4:1 single column crossbar layout in SiPho chip with MZI-based T/O weighting modules, (c) 2:1 single column crossbar layout in SiPho chip with EAM-based E/O weighting structures.

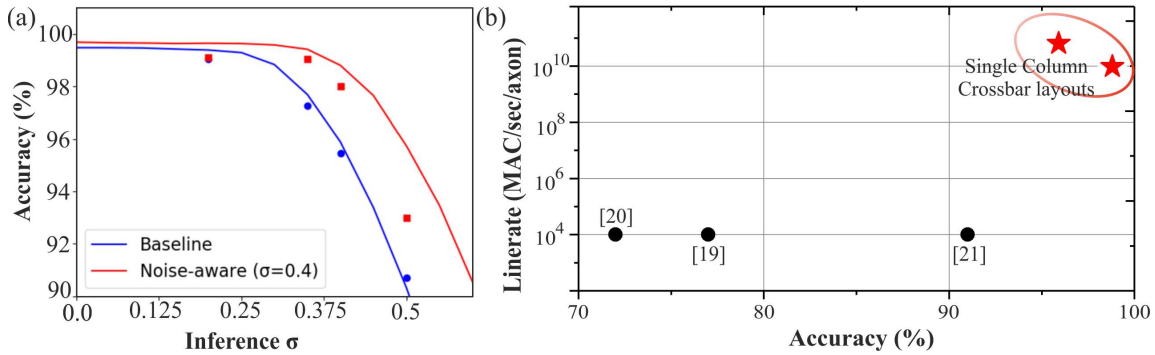

Fig. 3. (a) Accuracy performance on MNIST classification task versus the standard deviation of the noise σ at 10 GMAC/sec/axon, when using the 4:1 SiPho single column crossbar layout. Solid lines illustrate the simulated results, while the scatter points represent the experimentally recorded data. (b) Linerate in MAC/sec/axon versus classification accuracy of state-of-the-art coherent linear demonstrators.

values remain as high as ~98% and ~96%, outperforming all state-of-the-art coherent layouts by ~6 orders of magnitude in terms of compute rate and 7% in terms of classification accuracy performance.

4. Conclusion

We reviewed the neuromorphic photonic architectures and technologies, highlighting the key challenges and limitations of the state-of-the-art demonstrators. Additionally, we present our recent work on photonic NNs, demonstrating a linear crossbar architecture, along with a noise-aware training method to improve NN accuracy.

Acknowledgments

This work was supported by the European Commission (EC) through H2020 Projects PLASMONIAC (871391) and SIPHO-G (101017194) and by the Hellenic Foundation for Research and Innovation (H.F.R.I.) through project DeepLight (4233).

References

[1] B.J. Shastri *et al.*, “Photonics for artificial intelligence and neuromorphic computing”, *Nat. Photonics* 15, 102–114, 2021.
[2] A. R. Totović *et al.*, “Femtojoule per MAC Neuromorphic Photonics: An Energy and Technology Roadmap,” *IEEE J. Sel. Top. Quantum Electron.*, vol. 26, no. 5, pp. 1-15, 2020.
[3] M. Nahmias, *et. al.*, “Photonic Multiply-Accumulate Operations for Neural Networks,” *IEEE J. Sel. Top. Quantum Electron*, 26 (1), 2020
[4] G. Giamougiannis *et. al.*, “Coherent photonic crossbar as a universal linear operator”, submitted to *Lasers and Photonics Review*, 2021.
[5] G. Mourgias-Alexandris *et al.*, "Neuromorphic Photonics With Coherent Linear Neurons Using Dual-IQ Modulation Cells," *J. of Lightw. Technol.*, vol. 38, no. 4, pp. 811-819, 2020.
[6] G. Mourgias-Alexandris *et al.*, "A Silicon Photonic Coherent Neuron with 10GMAC/sec processing line-rate ", Optical Fiber Comm. Conf., Tu5H.1, 2021.
[7] G. Giamougiannis *et al*, “Silicon-integrated coherent neurons with 32GMAC/sec/axon compute line-rates using EAM-based input and weighting cells”, Eur. Conf. on Optical Comm. 2021.
[8] G. Mourgias-Alexandris *et al*, “An all-optical neuron with sigmoid activation function”, *Opt. Exp.*, Vol. 27, No. 7, pp. 9620-9630, 2019.
[9] N. Passalis *et al*, “Initializing Photonic Feed-forward Neural Networks using Auxiliary Tasks”, *Neural Networks*, vol. 129, pp. 103-108, 2020.
[10] N. Passalis *et al*, “Training Deep Photonic Convolutional Neural Networks with Sinusoidal Activations”, *IEEE Trans. on Emerging Topics in Comp. Intel.*, Vol. 5, No. 3, pp 384-393, 2021.
[11] J. Feldmann *et al.*, “Parallel convolutional processing using an integrated photonic tensor core”, *Nature* 589, 52–58, 2021.
[12] B. Shi *et al*, “Deep Neural Network Through an InP SOA-Based Photonic Integrated Cross-Connect,” *IEEE J. Sel. Top. Quantum Electron.*, vol. 26, no. 1, pp. 1–11, 2020.
[13] A. Tait *et al.*, "Silicon Photonic Modulator Neuron", *Physical Review Applied*, vol. 11, no. 6, 2019.
[14] Y. Huang *et al*, "Programmable matrix operation with reconfigurable time-wavelength plane manipulation and dispersed time delay," *Opt. Express* 27, 20456-20467, 2019.
[15] T. F. de Lima *et al*, "Real-time Operation of Silicon Photonic Neurons," Optical Fiber Communication Conference (OFC) 2020, M2K.4., 2020.
[16] W. Zhang *et al*, “Microring Weight Banks Control beyond 8.5-bits Accuracy”, arXiv preprint arXiv:2104.01164, 2021
[17] X. Xu *et al.*, “11 TOPS photonic convolutional accelerator for optical neural networks”, *Nature* 589, 44–51, 2021.
[18] A. Ribeiro *et al*, "Demonstration of a 4 × 4-port universal linear circuit," *Optica* 3, 1348-1357, 2016.
[19] Y. Shen *et al.*, “Deep learning with coherent nanophotonic circuits,” *Nat. Photonics*, vol. 11, no. 7, pp. 441–446, 2017.
[20] F. Shokraneh *et al*, "A Single Layer Neural Network Implemented by a 4x4 MZI-Based Optical Processor," *IEEE Photon. J.*, vol. 11, no. 6, pp. 1-12, 2019.
[21] H. Zhang *et al*. “An optical neural chip for implementing complex-valued neural network”, *Nat Commun* 12, 457, 2021.
[22] M. Miscuglio *et al*, “Artificial Synapse with Mnemonic Functionality using GSST-based Photonic Integrated Memory”, *Intern. App. Comp. Electromagnetics Society Symp*. (ACES), pp. 1-3, 2020.
[23] A. Tait *et al.*, "Neuromorphic photonic networks using silicon photonic weight banks", *Sci. Rep.*, 7 (1), 2017.
[24] J. Elliott Ortmann *et al*, ”Ultra-Low-Power Tuning in Hybrid Barium Titanate–Silicon Nitride Electro-optic Devices on Silicon”, *ACS Phot*, 6, 11, 2677–2684, 2019.