Silicon-integrated coherent neurons with 32GMAC/sec/axon compute line-rates using EAM-based input and weighting cells

George Giamougiannis[(1)], Apostolos Tsakyridis[(1)], George Mourgias-Alexandris[(1)], Miltiadis Moralis-Pegios[(1)], Angelina Totovic[(1)], George Dabos[(1)], Nikos Passalis[(1)], Manos Kirtas[(1)], Nikos Bamiedakis[(2),] Anastasios Tefas[(1)],David Lazovsky[(2)] and Nikos Pleros[(1)]

[(1)] Department of Informatics, Aristotle University of Thessaloniki, 54124, Thessaloniki, Greece, giamouge@csd.auth.gr

[(2)] Celestial AI, 100 Mathilda Place, Suite 170, Campbell, CA 95008, United States

Abstract: We experimentally demonstrate a coherent SiPho neuron that relies on EAM for both on-chip data generation and weighting. A record-high 32GMAC/s/axon compute rate and an accuracy of 95.91% is reported, when the neuron is deployed as a hidden layer of a MNIST classifier neural network.

Introduction

Recent advances in photonic integration technologies[1] and theoretical progress in optical computing architectures[2],[3], have fueled research interest in photonic neuromorphic computing, aiming to transfer the low-power and high-speed and density credentials of light manipulating components in parallel computing layouts[4]-[6]. Previous implementations of photonic neuromorphic hardware can be classified in two broad categories: i) in multi-wavelength resonant layouts, where every axon is realized via a different wavelength[7]-[11], ii) in coherent layouts that use a single laser source and utilize the phase of transmitted light both for sign representation as well as for matrix manipulation in interferometric layouts[12]-[14].

Previous demonstrations of WDM architectures achieved up to 10 Gbaud and 91.7% accuracy for classifying the IRIS dataset[8]. However, scaling of these schemes requires a larger number of wavelengths and accurate resonant control[10], limiting their potential in high-radix layouts. Coherent neuromorphic demonstrations are usually based on the Reck linear optical circuit architecture[2], that requires N-1 cascaded stages of Mach-Zehnder interferometers (MZI) for an N-input linear neuron. Given the direct trade-off between insertion loss and modulation bandwidth, phase and amplitude modulation photonic devices[1], such layouts must balance between high-speed operation and achievable laser powers. The majority of prior demonstrations is opting for the latter, implementing MHz scale thermo-optic (TO) photonic devices[12],[13]. As such, operating speeds have been restricted so far to 10 kHz with reported accuracy values that can reach 90% when performing MNIST classification[12] and vowel recognition[13]. We have recently validated the potential for increasing on-chip compute rates, to up to 10 GMAC/sec/axon, in coherent neuromorphic photonics by adopting a novel architectural scheme[14] and by deploying electro-optic and TO Si-based modulators for data generation and weighting, respectively[15].

In this paper, we report for the first time, to the best of our knowledge, a silicon-integrated coherent linear neuron (COLN) that relies on SiGe electro-absorption modulators (EAM)[16] both for its on-chip data generation and weighting stage and demonstrate a record-high 32 GMAC/sec/axon compute linerate. The performance of the photonic neuron was assessed through the on-chip implementation of the functionality of the penultimate hidden layer of a neural network (NN) that classified the handwritten digits of the MNIST dataset. We report experimentally obtained accuracy of 98.09% and 95.91% at 16 and 32 Gbaud line rates, respectively. Additionally, the energy efficiency of the constituent weighting element of our architecture was measured to be 0.083 pJ/MAC, while the computational density of the basic X×W computational cell of our COLN architecture was measured to be 0.32 TMAC/s/mm^2.

Neural network architecture and experimental layout

In order to assess the perfomance of the fabricated COLN, we designed and trained a 5-layer neural network, depicted in Fig. 1(a), that can perform image classification of the MNIST dataset. The gray-scale images of the dataset are decomposed to 1902 values that comprise the inputs of the input layer of the designed NN. The implemented NN layout is initiated by two convolutional layers of 32 (L#1) and 64 (L#2) 3×3 filters, respectively, that utilize the ReLU activation function (AF). Three linear layers of 4, 2 and 1 neurons follow, with the ReLU AF used in the first linear layer (L#3) and the $\sin^2(x)$ AF employed in the last two (L#4 and L#5). The Xavier scheme with a gain of 2 was applied for the initialization of the network, while the training

978-1-6654-3868-1/21/$31.00 ©2021 IEEE

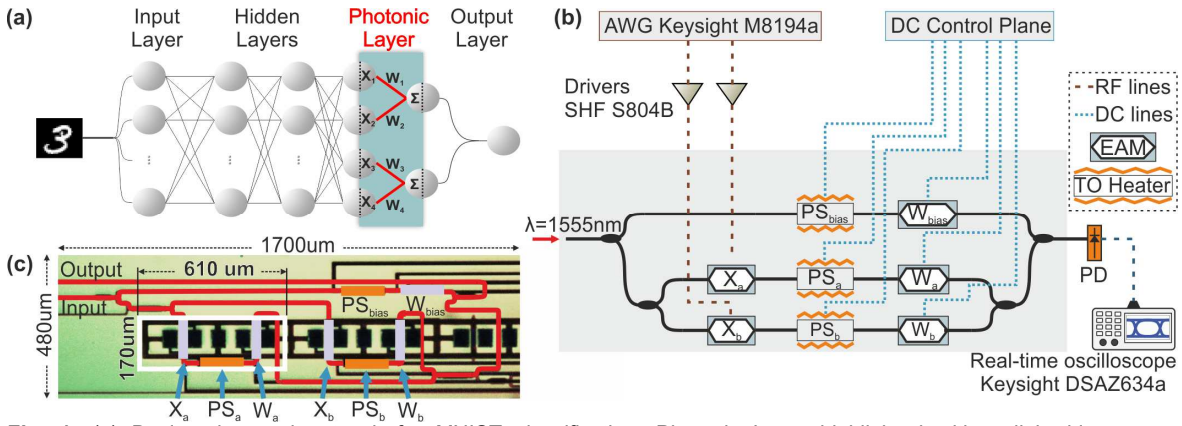

Fig. 1: (a) Designed neural network for MNIST classification. Photonic Layer highlighted with a light blue rectangle (b) Experimental setup for the evaluation of the SiPho COLN (c) Microscope photo of the integrated COLN. The elementary computational cell is encapsulated within a white rectangle.

optimization was achieved via the Adam optimizer. The network was trained for the recognition of the digits 3 and 5 for 20 epochs, with a batch size of 256 samples and a learning rate of 10^{-4}.

The penultimate hidden layer, highlighted in the light blue box of Fig. 1(a), was implemented on the photonic chip and comprises 4 inputs (X_1, X_2, X_3, X_4) with their respective weights (W_1, W_2, W_3, W_4) and two outputs (Σ_1, Σ_2). The 4 inputs X_1, X_2, X_3 and X_4, are individually weighted and summed up in pairs of two, in order to form the two outputs of the neuron i.e $\Sigma_1 = X_1W_1 + X_2W_2$ and $\Sigma_2 = X_3W_3 + X_4W_4$.

The experimental setup, which reflects the SiPho chip's layout, is portrayed in Fig. 1(b). The light, first, splits via a 3dB coupler, into the bias branch (upper branch) and the nested MZI (lower branch). Each branch of the nested MZI comprises an RF driven EAM, utilized for data input imprinting (X_i), a DC driven EAM, for weight amplitude imprinting (W_i) and a TO Phase Shifter (PS) used for sign control. The bias branch consists of a TO PS followed by a DC driven GeSi EAM. The signal exiting the nested MZI recombines in a 50/50 coupler with the signal that emerges from the bias branch prior reaching the receiver site. This nested MZI scheme follows the principles described in[14],[15] with the TO PSs of the inner MZI imprinting the phase of the weight values, with φ=0 and φ=π corresponding to the positive and negative weighting, respectively. Additionally, the constructive or destructive interference of the signal emerging from the inner MZI with the bias signal ensures that, even if the weighted summation has a negative value, the information will be preserved due to the DC offset that the bias signal introduces.

In order to experimentally validate the performance of the integrated neuron we, sequentially, interfaced the two pairs of the linear layer L#4 of the NN into the SiPho chip. Initially, the first pair of waveforms emerging from the 3rd Layer of our NN (X_1, X_2) were upsampled from 1 sample per symbol (sps) to 6 and 3 sps for the 16 and 32 Gbaud cases, respectively. Subsequently, they were filtered through a finite impulse response (FIR) filter, that counterbalanced the non-ideal response of the photonic devices and quantized with 8-bit resolution. The resulting signals were generated with a Keysight's M8194a arbitrary waveform generator (AWG) operating at 96 GSa/s and fed via two linear RF amplifiers (SHF S804B) into the input X_i EAMs of the nested MZI, with a driving voltage of approximately 3 Vpp. At the same time, a light beam at λ=1555 nm was injected into the SiPho chip input via a TE grating coupler. Finally, the signal was retrieved via a 66GHz bandwidth real-time oscilloscope (RTO) (Keysight DSAZ634a) at 160 GSa/s, after being captured by a 70GHz PIN photodiode. The received signal was, then, filtered with a Gaussian filter and downsampled to 1sps before being fed again to the NN. The same procedure was followed for the experimental evaluation of the X_3, X_4 input pair of L#4.

The SiPho neuromorphic chip, was fabricated in IMECs ISIPP50G platform, and is depicted in Fig. 1(c). The white rectangle of Fig. 1(c), highlights one basic X×W computational cell, comprising two EAMs[16] and a PS, for input data generation, weight amplitude and sign control, respectively. Its dimensions were approximately 609×170 um^2, while the 2-fan in COLN occupied a total area of 1700×480 um^2.

Experimental results and scaling analysis

Figure 2 (a)-(d) depict the experimentally obtained (rendered with blue lines) X_1, X_2 and X_3, X_4 time traces, respectively, along with the corresponding expected ones (orange lines) at

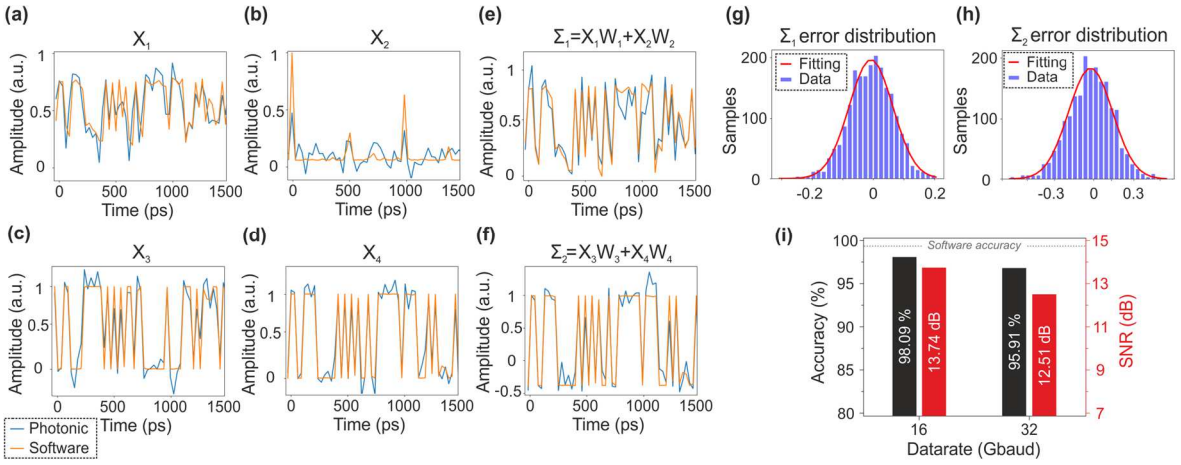

Fig. 2: Time traces of the obtained (blue) and expected (orange) signals at 32 Gbaud: (a)-(d) when the inputs X_1, X_2, X_3 and X_4 of Layer #4 are interfaced to the SiPho neuron, (e),(f) after the weighted summation of the two input pairs X_1, X_2 and X_3, X_4. (g),(h) Error distributions of the weighted sums of the two input pairs. (i) Classification accuracy and SNR measurements at 16 and 32 Gbaud.

the rate of 32 Gbaud, while Fig. 2 (e) and (f) illustrate the corresponding all-optical weighted sums $\Sigma_1 = X_1W_1 + X_2W_2$ and $\Sigma_2 = X_3W_3 + X_4W_4$, respectively, along with the respective expected NN signals. In order to quantify the divergence of the received weighted sums with respect to the expected ones, we calculated their error distribution shown in Fig. 2 (g) and (h). It can be derived that both error histograms approximately correspond to the Gaussian distribution with $(\mu_{12},\sigma_{12})=(-0.03,0.13)$ and $(\mu_{34},\sigma_{34})=(-0.02,0.16)$. Finally, Fig. 2 (i) depicts the MNIST dataset classification accuracy, for the linerates of 16 and 32 Gbaud being 98.09% and 95.91%, respectively, while the corresponding calculated SNR values were 13.74 dB and 12.51 dB.

The energy efficiency benefits of the EAM-based silicon coherent platform can be outlined by scaling the demonstrated 1×2 vector-vector multiplication to a full N×N vector-matrix configuration following our recently proposed optical crossbar architecture described in [17],[18]. Figure 3 depicts the total insertion loss (IL) and energy efficiency metrics as N scales from 2 to

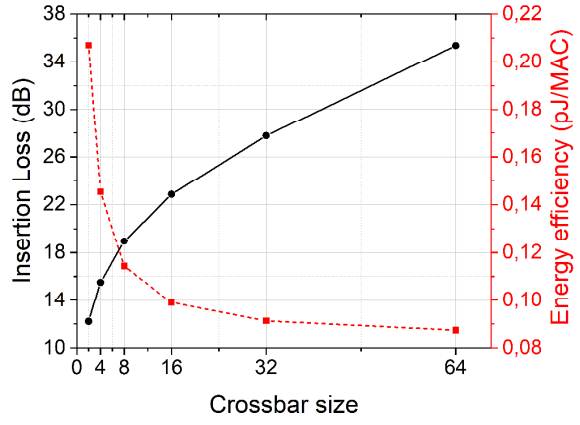

Fig. 3: Insertion Loss (black solid line) and energy efficiency (red dashed line) versus device radix.

64, using the following experimentally verified values for the calculations: X_i EAMs with an IL of 4.5 dB, a responsivity of 0.8 A/W, a capacitance of 20 fF[16] and a 3 Vpp driving voltage, W_i EAMs with an insertion loss of 4.5 dB and an average driving voltage of 1.5 V, TO PS for weight sign imprinting with a 4 mW/π efficiency. The analysis reveals that energy efficiency approaches asymptotically a lower bound as N increases, with the lower bound value dictated by the 0.083 pJ/MAC offered by a single weight EAM. In the case of a 32×32 Crossbar layout operating at 32 GMAC/sec/axon and offering a >32 TMAC/sec computational power, the total IL would remain <30 dB and the electrical energy efficiency would be ~0.09 pJ/MAC.

Conclusions

We demonstrated experimentally a SiPho 2-fan-in COLN, based on EAMs, as an integral part of a NN classifying the MNIST dataset at the compute rates of 16 and 32 GMAC/s/axon, yielding accuracy values of 98.09% and 95.91%, respectively. Additionally, the energy efficiency of the constituent weighting element of our architecture was measured to be 0.083 pJ/MAC, while the computational density of the basic X×W computational cell of our COLN architecture equals 0.32 TMAC/s/mm^2. Finally, an electrical power consumption and loss budget analysis of a scaled-up version of the proposed layout revealed an efficiency of ~0.09 pJ/MAC and an insertion loss of <30 dB at a 32×32 radix.

Acknowledgements

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

References

[1] IPSR, “2020 IPSR-International Integrated Photonic Systems Roadmap- Silicon Photonics”, [Online]. Available: https://photonicsmanufacturing.org/sites/default/files/documents/front-end_siph3_0.pdf

[2] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, “Experimental realization of any discrete unitary operator,” Phys. Rev. Lett. 73, 58–61 (1994).

[3] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. Steven Kolthammer and I. A. Walmsley, “Optimal design for universal multiport interferometers,” Optica 3, 1460–1465 (2016).

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

[5] A. R. Totović, G. Dabos, N. Passalis, A. Tefas and N. Pleros, "Femtojoule per MAC Neuromorphic Photonics: An Energy and Technology Roadmap," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 26, no. 5, pp. 1-15, Sept.-Oct. 2020.

[6] H. Peng, M. A. Nahmias, T. F. de Lima, A. N. Tait and B. J. Shastri, "Neuromorphic Photonic Integrated Circuits," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 24, no. 6, pp. 1-15, Nov.-Dec. 2018

[7] A. Tait, T. F. de Lima, M. A. Nahmias, H. B. Miller, H. T. Peng, B. J. Shastri, P. R. Prucnal, "Silicon Photonic Modulator Neuron", Physical Review Applied, vol. 11, no. 6, 2019.

[8] B. Shi, N. Calabretta and R. Stabile, "Deep Neural Network Through an InP SOA-Based Photonic Integrated Cross-Connect," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 26, no. 1, pp. 1-11, Jan.-Feb. 2020.

[9] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, J. Liu, C. D. Wright, A. Sebastian, T. J. Kippenberg, W. H. P. Pernice, H. Bhaskaran, Parallel convolutional processing using an integrated photonic tensor core. Nature 589, 52–58, 2021.

[10] C. Huang, S. Bilodeau, T. Ferreira de Lima, A. N. Tait, Philip Y. Ma, E. C. Blow, A. Jha, H. T. Peng, B J. Shastri, and P. R. Prucnal, "Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits", APL Photonics 5, 040803, 2020.

[11] S. Ohno, K. Toprasertpong, S. Takagi and M. Takenaka, "Demonstration of Classification Task Using Optical Neural Network Based on Si Microring Resonator Crossbar Array," 2020 European Conference on Optical Communications (ECOC), 2020.

[12] 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, A. Q. Liu, An optical neural chip for implementing complex-valued neural network. Nat Commun 12, 457, 2021.

[13] Y. Shen, N. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, M. Soljačić, Deep learning with coherent nanophotonic circuits. Nature Photon 11, 441–446 ,2017.

[14] G. Mourgias-Alexandris, A. Totovic, A. Tsakyridis, N. Passalis, K. Vyrsokinos, A. Tefas, N. Pleros, "Neuromorphic Photonics With Coherent Linear Neurons Using Dual-IQ Modulation Cells," in Journal of Lightwave Technology, vol. 38, no. 4, pp. 811-819, 15 Feb.15, 2020.

[15] G. Mourgias-Alexandris, M. Moralis-Pegios, S. Simos et al., "A Silicon Photonic Coherent Neuron with 10GMAC/sec processing line-rate ", Optical Fiber Communication Conference (OFC), ID : Tu5H.1, 2021.

[16] 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, J. Van Campenhout, "Active Components for 50 Gb/s NRZ-OOK Optical Interconnects in a Silicon Photonics Platform," in Journal of Lightwave Technology, vol. 35, no. 4, pp. 631-638, 15 Feb.15, 2017,

[17] G. Giamougiannis, A. Tsakyridis, Y. Ma, A. Totovic, D. Lazovsky, N. Pleros, “Coherent photonic crossbar as a universal linear operator”, submitted to Physical Review X, 2021.

[18] A. Tsakyridis, G. Giamougiannis, A. Totovic, Y. Ma, D. Lazovsky, N. Pleros, “Universal Linear Optical Operator with Optimal Fidelity Performance”, submitted to Proc. European Conference on Optical Communication (ECOC), 2021, Bordeaux, France