A TeraFLOP Photonic Matrix Multiplier using Time-Space-Wavelength Multiplexed AWGR-based Architectures

Christos Pappas[1], Theodoros Moschos[1], Miltiadis Moralis-Pegios[1], George Giamougiannis[1], Apostolos Tsakyridis[1], Manos Kirtas[2], Nikolaos Passalis[2], Anastasios Tefas[2] and Nikos Pleros[1]
[1]Centre for Interdisciplinary Research and Innovation, Informatics Dept. Aristotle University of Thessaloniki, Greece
[2]Computational Intelligence and Deep Learning Group, Dept. of Informatics, Aristotle University of Thessaloniki, Greece
Author e-mail address: chripapp@csd.auth.gr

Abstract: We demonstrate experimentally a novel 8×8 AWGR-based photonic matrix multiplier that enables simultaneously time-, wavelength- and space- division multiplexed computing with a computational power of 1.28 TeraFLOP.

1. Introduction

The continuously growing compute data volume emanating from the burst of Artificial Intelligence (AI) and Large Language Models (LLMs) [1], alongside the fundamental plateaus of digital electronic AI computing systems [2] has stimulated extensive research in the pursuit of lower energy and higher speed AI chipsets. Photonic platforms have been among the most favored candidates, since they can exploit the unparalleled speed of light in the context of the massive compute parallelism offered by the physical dimensions of time, wavelength, and space [3].

Numerous optical neural network (ONN) architectures have been developed based on these multiple degrees of freedom, targeting large-scale interconnectivity and high computational capacity [4]-[10]. Multiwavelength approaches mainly attempted to scale both in space and wavelength domain, yet with the one division overlapping the other, since every spatially separated computing-cell encodes the information on a different wavelength [4]-[6]. On the other hand, single-λ ONN schemes have mainly expanded either along space-multiplexed designs in the form of coherent optical linear circuits [7]-[10] or along time-division multiplexed (TDM) schemes [11]-[13] yet in both cases neglecting the wavelength parallelism that can be offered by photonic platforms.

In this paper, we introduce a novel arrayed waveguide grating router (AWGR)-based ONN architecture that can simultaneously support time-, wavelength- and space- division multiplexed (T-W-SDM) computing and demonstrate experimentally matrix-by-matrix multiplication (MbMM) operations at 10 GHz line-rates with a total computational power of 1.28 TeraFLOP. The proposed architecture exploits high-speed and broadband optical modulators together with the cyclic wavelength routing and the port- and wavelength-scalability of an AWGR module [14] towards supporting high computational powers and a large number of trainable parameters. The proposed ONN architecture is demonstrated experimentally using an 8×8 AWGR for the classification of the Fashion MNIST dataset, yielding a mean classification accuracy of 87.1% of the software-acquired accuracy obtained via a 10:8:5 NN at 10 Gbaud.

2. The AWGR-based TWSDM Architecture and Experimental Setup

Figure 1 (a) illustrates the Time-Wavelength multiplexing approach obtained via an AWGR-based layout, prior introducing also the space dimension. It comprises an N×N AWGR-based scheme where a set of N modulators is connected to its #N input ports and only output port #1 is connected to an output modulator. Feeding the i-th (i∈[1, N]) input modulator with a different wavelength λ_i and driving it with a respective $\overrightarrow{W_i}(t)$ electrical time series, every λ_i beam carries a corresponding optical weight time vector prior entering the AWGR input port. The AWGR multiplexing properties allow then all N optical signals to exit through the same AWGR output port and enter the output modulator, which is driven by an electrical $\overrightarrow{X_1}(t)$ time series that has the same length L with the weighting vectors. In this way, the multi-λ signal that exits the output modulator carries every Hadamard product $\overrightarrow{W_i}(t)\circ\overrightarrow{X_1}(t)$ as a time series at a different wavelength λ_i. This signal is then demultiplexed and every wavelength enters a different integrator that integrates the time series over a number of L symbols, as proposed in [12], effectively providing in this way the dot-product between the vector $\overrightarrow{W_i}(t)$ and $\overrightarrow{X_1}(t)$ at its output port $Y_{i,1}$. In this way, the Y_N DEMUX output ports provide the Matrix-by-Vector multiplication (MbVM) shown at the right side of Fig. 1(a).

The space dimension towards a TWSDM architecture is achieved by extending this setup towards supporting #K modulators at the respective AWGR output ports (K<=N) and all λ_1-λ_N wavelengths at every AWGR input modulator, as shown in Fig. 1(b). In this way, every i-th input modulator imprints the L-symbols long time-series vector $\overrightarrow{W_i}(t)$ on all #N wavelengths and the cyclic routing properties of the AWGR allow then all #N different time vectors $\overrightarrow{W_1}(t)$, …, $\overrightarrow{W_N}(t)$ to emerge at every AWGR output port and every of the #K AWGR output modulators. By driving every j-th AWGR output modulator with a different $\overrightarrow{X_j}(t)$ time vector with a dimension of L, a different MbVM product is obtained at every AWGR output, similar to the way described in Fig. 1(a). As such, for #K output

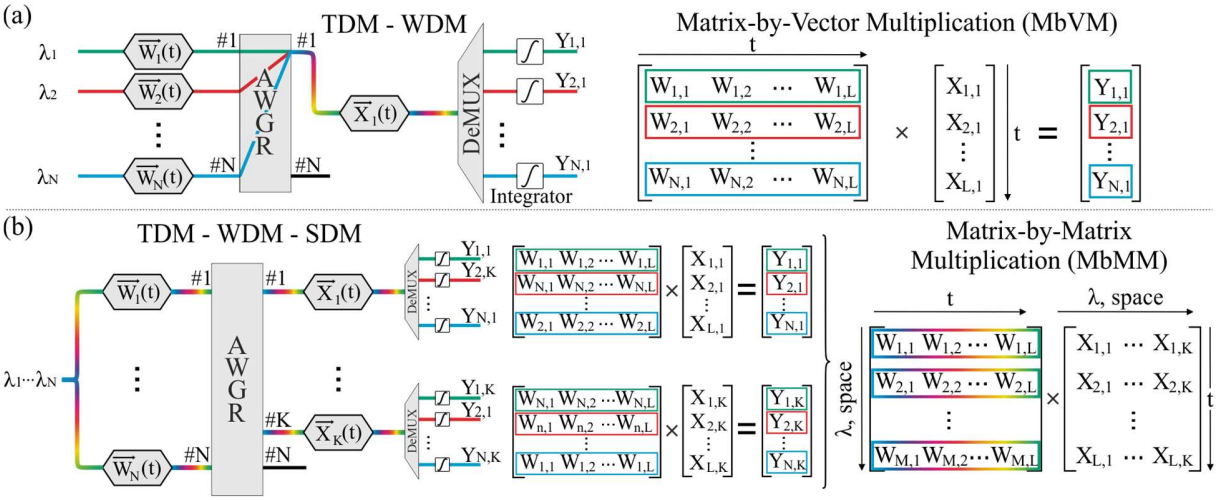

Fig. 1: (a) Conceptual layout of a T-WDM ONN comprised of #N modulators at the input of N×N AWGR interconnected with #1 modulator at its output for the execution of a MbVM operation. (b) T-W-SDM scaling of (a) for the computation of MbMM operations.

modulators that imprint #K different input signal L:1 vectors $\overrightarrow{X_1}(t)$, …,$\overrightarrow{X_K}(t)$, this layout yields successfully an MbMM operation, as shown in Fig. 1(b).

In our experimental demonstration we showcase MbMM operations, by training a 10:8:5 fully-connected NN for classifying 5 classes of the FMINST dataset as illustrated in Fig. 2 (a), achieving a software accuracy of 70.6% and perform the inference operation over the AWGR-based layout by sequentially implementing the two NN layers. The established experimental testbed is depicted in Fig. 2(b). Specifically, eight 30 mW continuous wave signals i.e., λ_1-λ_8 with their wavelengths values ranging from 1547.2 nm to 1558.4 nm and a channel spacing of 1.6 nm were generated from a laser diode (LD) bank and subsequently combined via a multiplexer (MUX). The WDM stream was split by a 90/10 coupler with 90% pointed towards the weight matrix generation and 10% towards a crosstalk emulation stage. The light following the 90% path was split once more by a 3 dB coupler, to create two identical WDM paths. Each path was demultiplexed, to enable the individual control of each wavelength's polarization state and multiplexed again before entering two $LiNbO_3$ Mach- Zehnder modulators (MODs). The latter were driven by an arbitrary waveform generator (AWG) to generate the elements of the NN weight matrices. For an N:L layer, each weight MOD, W_A and W_B imprint: i) W_A: $\overrightarrow{W_1}(t)$ - $\overrightarrow{W_{N/2}}(t)$ and W_B: $\overrightarrow{W_{(N/2)+1}}(t)$ - $\overrightarrow{W_N}(t)$ when N is an even number and ii) W_A: $\overrightarrow{W_1}(t)$ - $\overrightarrow{W_{(N+1)/2}}(t)$ and W_B: $\overrightarrow{W_{((N+1)/2)+1}}(t)$ - $\overrightarrow{W_N}(t)$ with a zero-padding vector $\overrightarrow{W_0}(t)$ when N is an odd number, to balance the pattern lengths withing W_A and W_B. In both cases, the weight vectors are encoded using a TDM scheme. For instance, for the inference of the first neural layer of our model, $\overrightarrow{W_1}(t)$ - $\overrightarrow{W_4}(t)$ = [$W_{1,1}$, …, $W_{1,10}$, …$W_{4,1}$, …, $W_{4,10}$] was encoded to W_A and $\overrightarrow{W_5}(t)$ - $\overrightarrow{W_8}(t)$ = [$W_{5,1}$, …, $W_{5,10}$, …, $W_{8,1}$ …, $W_{8,10}$] to W_B. The elements of the weight matrix of the output layer were encoded accordingly. The WDM weight-modulated signals W_A and W_B were fed to the commercially available 8×8 AWGR through the ports #1 and #2. To emulate the crosstalk of the AWGR that would normally arise from the input ports #3-#8, we loaded the residual AWGR ports with "dummy" 10 Gb/s NRZ signals that were generated by an AWG, as shown in the light blue rectangle of Fig. 2(b). The dummy signal was then amplified via an erbium doped fiber amplifier (EDFA) before being injected to a wave shaper (WS),

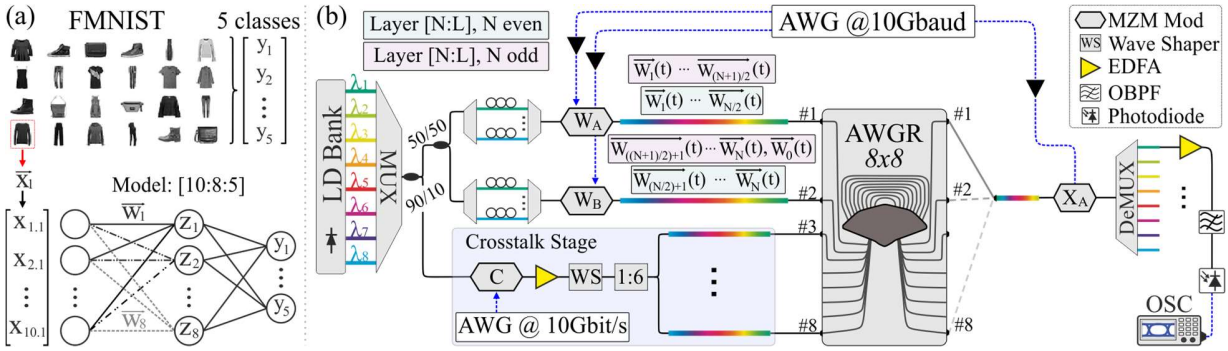

Fig. 2: (a) The NN model sized as [10:8:5] based on the first 5 output classes of the Fashion MNIST dataset, (b) Experimental setup for the 8×8 AWGR-based PNN with 10 Gbaud data generation.

responsible for equalizing the dummy signals with the information carrying signals and split into 6 components to load the #3-#8 inputs of the AWGR. A $LiNbO_3$ MOD, X_A, was driven by the AWG to generate the NN input vectors and was sequentially connected to all eight output ports #1-#8 of the AWGR. For each port i (i∈[1, N]), 32 different NN input vectors $\vec{X_i}(t)$ were imprinted time-wise, equally splitting the #256 samples of the FMNIST dataset to the #8 NN input encoding MODs. Finally, the WDM output of X_A was demultiplexed to λ_1-λ_8. Since two MODs were employed for the NN weights encoding, only 2 wavelengths in each output contained the useful multiplication information, with the latter being collected by a PD and captured by a 50 GHz sample scope, after loss compensation by an EDFA. Time-demultiplexing, integration of the hardware multiplications, as well as application of the non-linear activation function were performed off-line with a software routine.

3. Experimental results

The first step in our evaluation of the AWGR-based ONN performance comprised a preliminary characterization of the broadband profile of the employed MODs towards pre-compensating the non-linear response of every RF driver-modulator chain across the wavelength window defined by the 8 employed input signals. Subsequently, the multi-λ multiplication performance was evaluated. A WDM multi-level signal encoded to W_A was injected into the input port #1 of the AWGR, which was then multiplied by a signal encoded in X_A at every AWGR output port, #1-#8. Figure 3 (a) depicts the captured time traces, along with their mean square error (MSE) values, showcasing homogeneous performance across the wavelength channels. Having characterized the WDM performance of the MODs we proceeded by executing the inference of the FMNIST dataset. As detailed in section 2, in each AWGR output port a different set of multiplications between the weight matrix of each layer and 32 samples/port of the FMNIST inputs was calculated. Figure 3(b) presents the experimentally measured accuracy per samples/port as a percentage of the accuracy acquired by the software, revealing a mean value of 87.1% at 10 Gbaud ONN operation. In the same figure we plot with dashed lines the theoretically expected accuracy levels when the operations are quantized to 3-and 4-bits. It can be observed that the experimentally acquired accuracy levels range within these lines, indicating an experimental bit resolution within the [3, 4] range at 10 Gbaud. Finally, a confusion matrix of the 256 inferenced samples is depicted in Fig. 3(c), with each class of the FMNIST dataset defined by their parity bits. As can be observed in the confusion matrix, labels 2 and 3 were the easiest to classify with label 4 being the hardest.

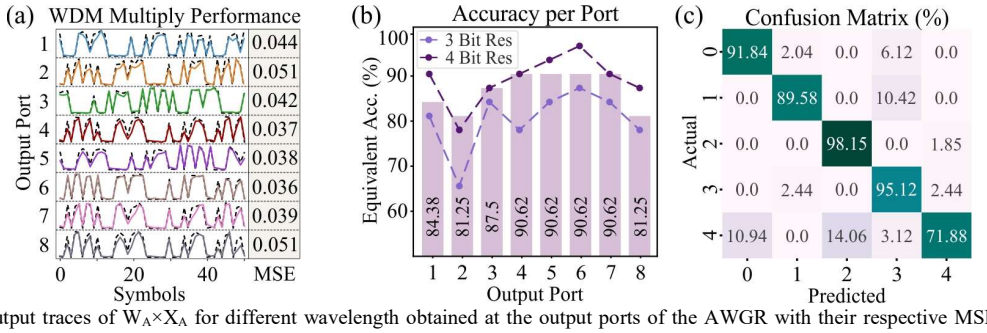

Fig. 3: (a) Output traces of $W_A \times X_A$ for different wavelength obtained at the output ports of the AWGR with their respective MSE values, (b) Achieved accuracy at 10 Gbaud operation for the 8 output ports of the AWGR along with the theoretical inference quantization of 3 and 4 bits resolution and (c) Confusion matrix of the total 256 samples inferenced through the PNN hardware.

Acknowledgements

This work was supported by the European Commission via GATEPOST (101120938).

References

[1] M. U. Hadi et. al., "Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects", TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.23589741.v3 (2023).
[2] Theis T. N. & Wong H.-S. P., "The end of Moore's law: A new beginning for information technology", Comput. Sci. & Eng. 19, 41 (2017).
[3] Y. Bai et.al., "Photonic multiplexing techniques for neuromorphic computing", Nanophotonics, vol. 12, no. 5, 2023, pp. 795-817.
[4] A. Totovic et. al., "WDM equipped universal linear optics for programmable neuromorphic photonic processors," NCE, 2022.
[5] J. Feldmann et. al., "Parallel convolutional processing using an integrated photonic tensor core", Nature 589, 52–58 (2021).
[6] C. Huang et. al., "Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits", APL Photonics 5.
[7] Y. Shen et. al., "Deep learning with coherent nanophotonic circuits", Nat. Photonics 11, 441–446 (2017).
[8] M. Moralis-Pegios et. al., "Perfect Linear Optics using Silicon Photonics", arxiv 2023: arXiv:2306.17728v1
[9] H. Zhang et. al., "An optical neural chip for implementing complex-valued neural network", Nat Commun 12, 457 (2021).
[10] G. Giamougiannis et. al., "Analog nanophotonic computing going practical: silicon photonic deep learning engines for tiled optical matrix multiplication with dynamic precision" Nanophotonics, 2023.
[11] R. Hamerly et.al., "Large-Scale Optical Neural Networks Based on Photoelectric Multiplication", Phys. Rev. X 9, 021032.
[12] L. De Marinis et. al., "A Codesigned Integrated Photonic Electronic Neuron," in IEEE JQE, vol. 58, no. 5, pp. 1-10, Oct. 2022.
[13] N. Youngblood, "Coherent Photonic Crossbar Arrays for Large-Scale Matrix-Matrix Multiplication," in IEEE JSTQE 2023.
[14] S. Cheung et. al., "Ultra-Compact Silicon Photonic 512 × 512 25 GHz Arrayed Waveguide Grating Router," in IEEE JSTQE 2014.