Vector-mode Multiplexing For Photonic Tensor Accelerator

Alireza Fardoost[1], Fatemeh Ghaedi Vanani[1], Zheyuan Zhu[1], Christopher Doerr[2], Shuo Pang[1], Guifang Li[1]
[1]CREOL, The College of Optics and Photonics, University of Central Florida, 4304 Scorpius St., Orlando, FL, USA
[2]Aloe Semiconductor, Inc. 1715 Highway 35, Suite 303, Middletown, NJ, USA
Pang@ucf.edu, Li@ucf.edu

Abstract: We propose a coherent multi-dimensional (wavelength, spatial mode, polarization, etc.) photonic tensor accelerator capable of performing high-speed artificial neural network computation. High-speed matrix-vector and matrix-matrix multiplication were experimentally demonstrated.
Keywords: *Photonic Accelerator, Matrix Multiplication, Optics for AI, Analog Computing*

I. Introduction

Electronic hardware accelerators have played an important role in the rapid advancements of artificial intelligence (AI). An AI accelerator, such as graphic processing unit (GPU) and tensor processing unit (TPU), is a high-performance parallel computation machine that propelled computing performances beyond the scaling limits of Moore's law. However, the cost will eventually limit the parallelization, and physics will limit the chip's efficiency. Therefore, it is necessary to develop new technologies far beyond today's capabilities to support the astonishing growth of AI computation. Passive linear optical circuits have the potentials to greatly reduce both computing and data-movement energy consumption. Low-power optical-to-electrical conversion has also recently been studied and shown promising results [1]. Accordingly, the field of optical artificial neural networks is experiencing a resurgence [2-4]. Here, for the first time, we combined wavelength-division multiplexing (WDM) and mode-division multiplexing (MDM) to encode matrices of unprecedented sizes. The proposed method of multidimensional encoding and coherent mixing supports vector, matrix, and tensor processing within a single clock cycle. Since reliable communication technologies can be deployed in modulation, multiplexing, and detection, the expected computation speed can go up to 10s of GHz and the energy efficiency can be optimized. We envision our photonics platform to be the building block for AI applications, including but not limited to, deep neural networks, real-time image processing, and dynamic control systems.

II. Photonic Tensor Accelerator (PTA)

The number of operations that can be performed in parallel will directly determine the speed of a hardware accelerator. In the proposed PTA, we take advantage of all degrees of freedom of light to maximize the number of parallel operations through one single clock cycle. Matrix multiplication, as the foundational process of any neural network, consists of multiply and accumulate (MAC). In the proposed PTA, scalar multiplication is performed via interference and coherent detection. The weight $\left(E_w = w.\exp\left(j\omega_0 t\right)\right)$ and input $\left(E_x = x.\exp\left(j\omega_0 t\right)\right)$ electric fields are combined on a balanced photodetector (BPD) and the output photocurrent would be their scalar multiplication $w \times x$.

Mapping the weight and vector elements onto different wavelengths can result in an extension of the scalar multiplication to the inner product of two vectors. Therefore, the BPD output will be $\sum_{n=1}^{N} w_n \times x_n$ where N is the length of the vectors. Analogous to the orthogonality of wavelength modes in time, spatial modes are orthogonal in space. Hence, a similar mapping to spatial modes will lead to the same inner product of the vectors.

Combining WDM and MDM with parallelization in space will enable the PTA to perform matrix-matrix multiplication in one clock cycle as shown schematically in Fig. 1. As shown here, the W matrix is projected on the (mode, space)

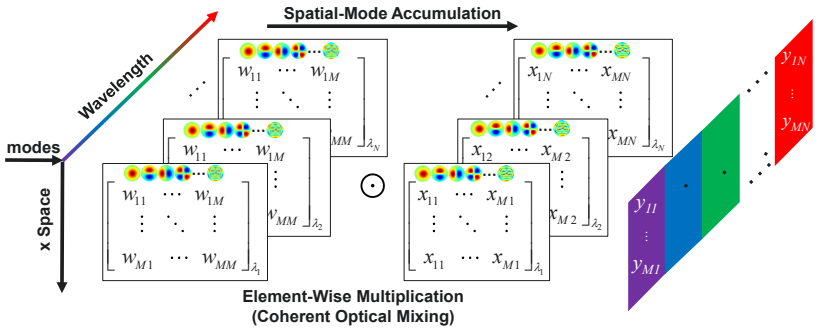

Fig. 1. Photonic Tensor Accelerator (PTA). Schematic mapping of the matrix elements on wavelengths, spatial modes, and space to implement matrix-matrix multiplication in a single clock cycle.

Funding: This work was supported in part by the Office of Naval Research under contract N00014-20-1-2441, NSF under grant number ECCS-1932858, and the Army Research Office under contract W911NF2110321.

dimensions and duplicated in the wavelength dimension. Similarly, the X matrix is projected on the (mode, wavelength) dimensions and duplicated in space. The output matrix elements are mapped to different wavelengths and spatial locations.

III. 4×4 MATRIX-VECTOR MULTIPLICATION DEMONSTRATION

The free-space implementation of a 4×4 matrix-vector multiplier is illustrated in Fig. 2(a). A CW laser at $\lambda_1 = 1545nm$ was modulated with a 16Gbps PRBS (2^{15}-1) using an EO modulator. To alleviate the requirements for a large number of high-speed equipment, we use only one modulator and appropriate delays in fiber and free space so that all 20 matrix and vector elements are decorrelated with each other. Three fiber delays are added between the four inputs of the all-fiber mode-selective photonic lantern (PL). Consequently, the output of the PL in four spatial modes consists of four vector elements of X ($x_{11}(\lambda_1, LP_{01})$, $x_{21}(\lambda_1, LP_{11a})$, $x_{31}(\lambda_1, \mathrm{LP}_{11b})$, and $x_{41}(\lambda_1, LP_{21a})$). The interference and detection are performed in free space where rows of the weight matrix $W_{4\times4}$ are delayed and thus decorrelated with the input vector $X_{4\times1}$ and each other. As a result, all four elements of the output matrix $Y_{4\times1}$ are obtained in one single clock cycle.

Measured intensity waveforms of the elements of Y are shown in Fig. 2(b) and are compared with their corresponding expected true output (Y_T) to show the well-matched condition of the results. The signal-to-noise ratio (SNR) of each output element can be defined as $SNR = \left(V_{rms} / error_{rms} \right)^2$ where $error = Y_T - Y$ and V_{rms} is the root mean square (RMS) voltage of the measured signal. The SNR is found to be 19-20dB for the elements of Y. Additionally, a more generalized parameter, the normalized mean square error (NMSE), $\left\| \vec{Y}_T - \vec{Y} \right\|^2 / \left\| \vec{Y}_T \right\|^2$, which is related to SNRs of all vector elements, can be defined to characterize the accuracy of the multiplication [5].

The maximum number of detectable levels is equal to $1/\sqrt{NMSE}$. Finding NMSE for more than 5000 different symbols of the signal in time, the overall bit precision for the matrix multiplication can be denoted as $-\log_2 \left\langle \sqrt{NMSE} \right\rangle$ where $\langle . \rangle$ is the ensemble average over NMSEs. The experimental results show 9 detectable levels and a bit precision of 3.2 bits.

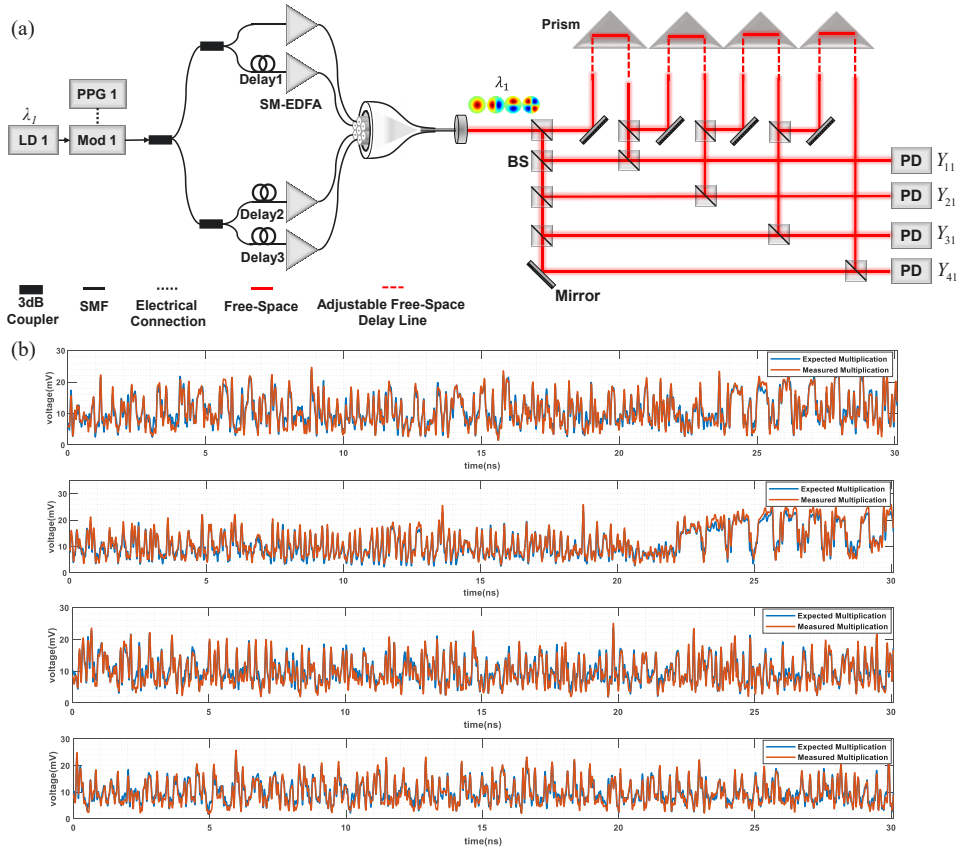

Fig. 2. Experiment setup and results for 4×4 matrix-vector multiplication in free space PTA. (a) Schematic of the experimental setup demonstration. LD: Laser Diode, PPG: Pulse Pattern Generator, BS: Beam Splitter, CF: Color Filter, PD: Photodetector. (b) Measured multiplication results, in comparison with the expected result in blue for all 4 output elements of $Y_{4\times1}$.

IV. 2×2 MATRIX-MATRIX MULTIPLICATION DEMONSTRATION

The same concept of multiplexing spatial modes and wavelengths can be extended to realize matrix-matrix multiplier units. A 2×2 matrix-matrix multiplier employing 2 wavelengths (1545 and 1555 nm) and 2 spatial modes (LP_{01}, LP_{11a}) is illustrated in Fig. 3(a). Here, the output of the PL in two wavelengths and two spatial modes consists of four matrix elements of X ($x_{11}(\lambda_1, LP_{01})$, $x_{12}(\lambda_1, LP_{11a})$, $x_{21}(\lambda_2, LP_{01})$, and $x_{22}(\lambda_2, LP_{11a})$). Additionally, 2 free space delays are deployed to generate the 4 decorrelated elements of the weight matrix $W_{2\times 2}$.

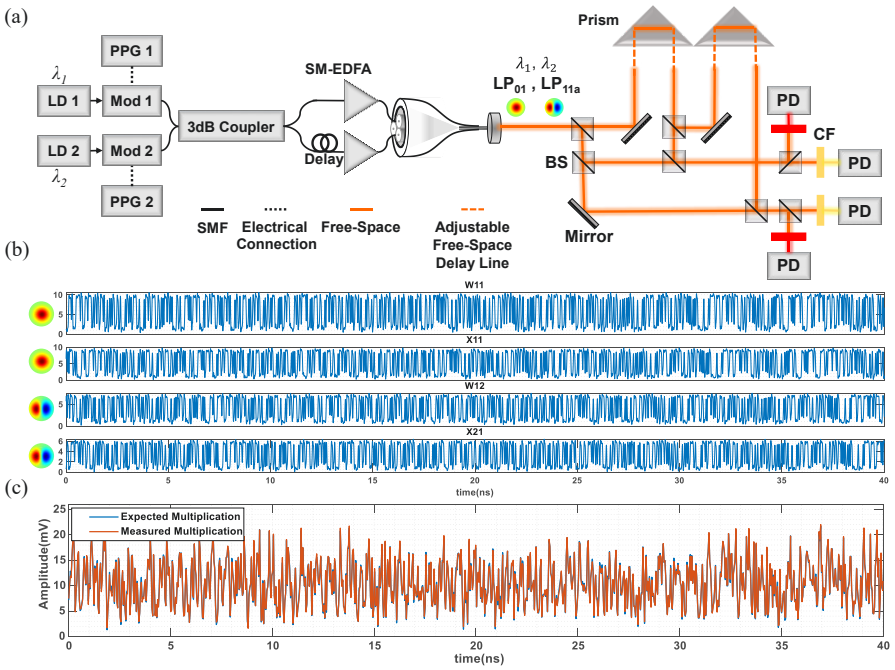

Fig. 3. Experiment setup and results for PTA in free space. (a) Schematic of the experimental setup for the matrix-matrix multiplication demonstration. LD: Laser Diode, PPG: Pulse Pattern Generator, BS: Beam Splitter, CF: Color Filter, PD: Photodetector. (b) Elements of the X and W matrices (c) measured multiplication results ($\tilde{Y}_{11}$ in red), in comparison with the expected result (Y_{T11} in blue).

A 25Gbps PRBS (2^{15}-1) signal is used to emulate the random elements of both matrices. Measured intensity waveforms for X and W , and one element $\left(\tilde{Y}_{11}\right)$ of the matrix multiplication $\left(\tilde{Y}\right)$ are shown for 1000 symbols $(40ns)$ in duration in Fig. 3(b) and (c), respectively. Using a similar method as described in section III, SNR is found to be between 21-22 dB for different elements of the output matrix. Additionally, the output matrix is first vectorized which results in precision of 4.67 bits obtained based on the NMSE definition which equivalently denotes more than 25 detectable levels at the output.

V. CONCLUSION

The key innovation for the photonic tensor accelerator (PTA) lies in exploiting all dimensions of light, each containing many degrees of freedom. Because accumulation is multi-dimensional, the scalability of the proposed PTAs is multiplicative since these dimensions are orthogonal. The PTA can be implemented on chip by taking advantage of mature and reliable photonic integration and optical communications technologies.

REFERENCES

[1] D. A. Miller, "Attojoule optoelectronics for low-energy information processing and communications," *Journal of Lightwave Technology,* vol. 35, no. 3, pp. 346-396, 2017.
[2] X. Lin *et al.*, "All-optical machine learning using diffractive deep neural networks," *Science,* vol. 361, no. 6406, pp. 1004-1008, 2018.
[3] Y. Shen *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature Photonics,* vol. 11, no. 7, pp. 441-446, 2017.
[4] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, "Large-scale optical neural networks based on photoelectric multiplication," *Physical Review X,* vol. 9, no. 2, p. 021032, 2019.
[5] D. M. Allen, "Mean square error of prediction as a criterion for selecting variables," *Technometrics,* vol. 13, no. 3, pp. 469-475, 1971.