

# High-density Integrated Photonic Tensor Processing Unit with a Matrix Multiply Compile

1<sup>st</sup> Hamed Dalir  
Optelligence LLC, Upper  
Marlboro, MD, 20772, USA  
hdalir@optelligence.co

2<sup>nd</sup> Behrouz Movahhed Nouri  
Optelligence LLC, Upper  
Marlboro, MD, 20772, USA  
movahhed@optelligence.co

3<sup>rd</sup> Xiaoxuan Ma  
Optelligence LLC, Upper  
Marlboro, MD, 20772,  
USA  
xma@optelligence.co

4<sup>th</sup> Peserico Nicola  
Department of Electrical and  
Computer Eng George  
Washington University  
npeserico@gwu.edu

5<sup>th</sup> Bhavin J. Shastri  
Department of Physics  
Queen's University  
Kingston, Canada  
bhavin.shastri@queensu.ca

6<sup>th</sup> Volker J. Sorger\*  
Department of Electrical and Computer Eng  
George Washington University  
Washington DC, USA  
sorger@email.gwu.edu

**Abstract**—With the increasing demanding for data processing and application-specific hardware, photonics can out-perform electrical IC in terms of speed and energy saving. Here, we present our Photonic Tensor Core architecture, showing high components density, with error rate lower than 4% on image edge detection.

**Index Terms**—Photonics, Silicon Photonics, Tensor Core

## I. INTRODUCTION

Data centers are the driven point for the new data-based era we are experiencing. Considering the storage capacity of all the data centers, the total amount of data stored has reached 2,300 Exabytes in 2021, with an average year-to-year grow of 22% since 2016 [1]. In this framework, hardware and architectural limits have started to appear, as data centers can't keep the growing pace by simply expanding their infrastructure. Moreover, novel algorithms that highly rely on Machine Learning approaches are becoming more and more diffuse at every levels, adding pressure on the data center hardware, but also on the edge computing units, exploiting the Internet of Thing paradigm. To address these limitations, specific hardware and IC have been designed, such as GPU and TPU, in particular for Deep Neural Network





(DNN) algorithms, that permit to perform Matrix-Vector Multiplication (MVM) at higher speed than traditional CPU. The MVM task is crucial for most of the DNN algorithms, as it is the mathematical way to connect and weight all the neurons of one layer to all the neurons of the following layer. However, electronic hardware still presents some limits, in terms of throughput and energy consumption.

Photonics field have presented several novel architectures to address those limitation, relying on the virtually infinite bandwidth and energy-free interference of the light beams. In table I, we present a comparison between the major integrated architecture, either based on matrix decomposition [2], or on direct matrix multiplication [3], [4].

In this work, we present our approach that links together the major strength points from those architectures. Using coupled add-drop microrings resonators, we can perform MVM task, showing good performance on experimental image processing. Moreover, our architecture shows the possibility to integrated directly P-RAM element, such as GSSe [5], allowing its use also on edge computing devices.

TABLE I

SCALING COMPARISON OF VARIOUS APPROACHES ON PERFORMING MVM OPERATIONS USING PHOTONIC CHIP-BASED COMPONENTS. N = SIZE OF INPUT VECTOR; M = SIZE OUTPUT VECTOR; P-RAM = PHOTONIC RANDOM ACCESS MEMORY, ALLOWING FOR ZERO-STATIC POWER CONSUMPTION, ONCE THE WEIGHTS ARE SET.

2*Type of Operation	$Y = V^T \Sigma U X$ 		$Y = M X$ 	 [This work]
Input	1 Laser, N Modulators	1 Comb Laser, N Modulators	N Lasers, N Modulators	N Lasers, N Modulators
Outputs	M(=N) Photodiodes	M Photodiodes	M or 2*M Photodiodes	M Photodiodes
Area/Basic Element Area	$N^2 + N$	$N \times M$	$N \times M$	$N \times M$
Controllers	$2N^2 + N$	$N \times M$	$2(N \times M)$	$N \times M$
Parallelization	No	WDM Off Chip	WDM On Chip	WDM On Chip
Weight Bit Resolution	8/10	5	>5	5
P-RAM	No	Yes	No	Yes

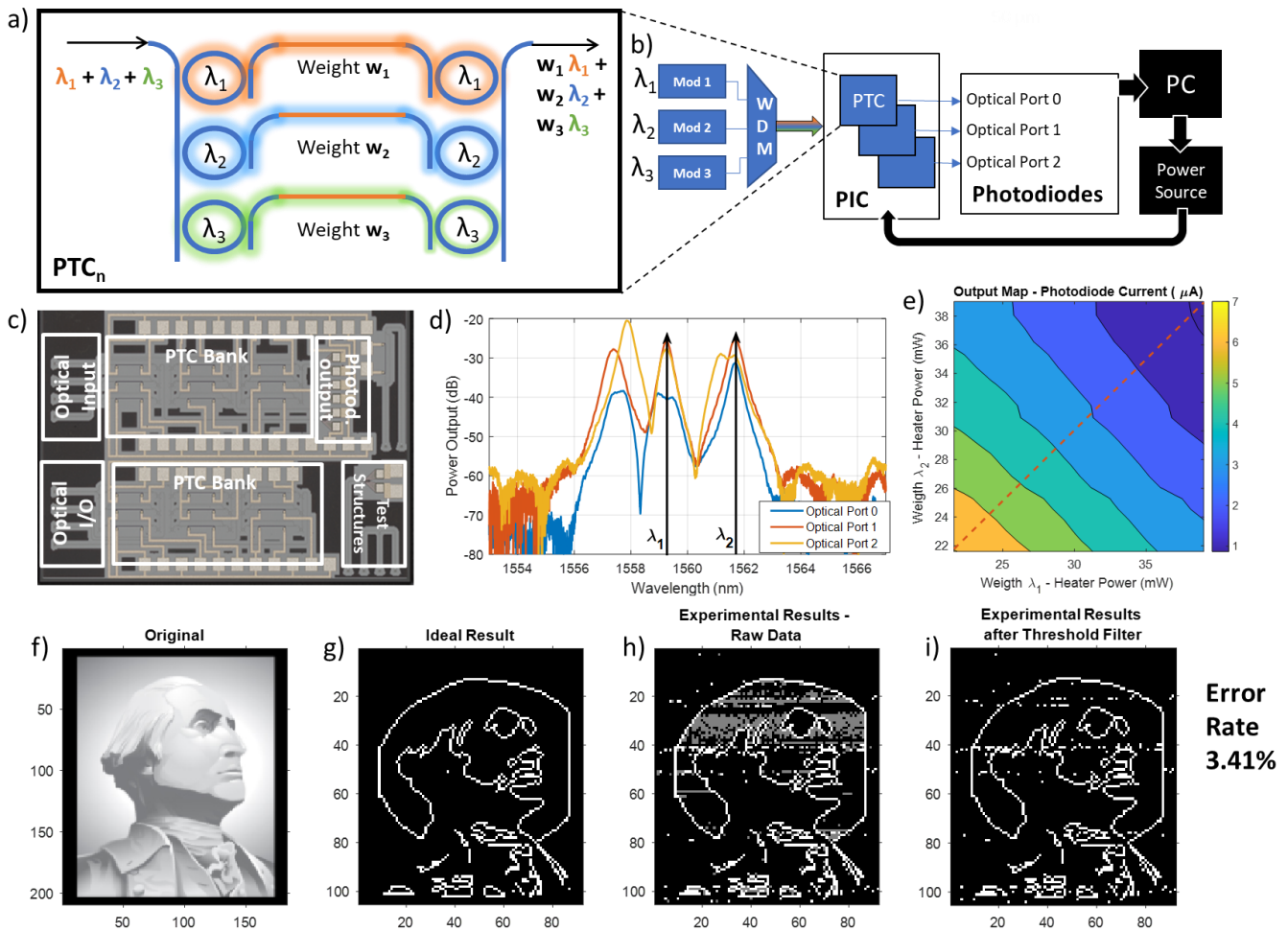


Fig. 1. Architecture and result of our Photonic Integrated Circuit. (a) Structure of a PTC module, (b) its integration on the PIC and setup, and (c) the photo of the PIC. (d) Optical transfer function, with the selected wavelengths. (e) Power map of one module with 2 inputs varying the weights, measure in current of the integrated photodiode. (f-i) Image edge detection: from the original BW figure, we obtain the ideal one (using MATLAB) and the Experimental one, before and after threshold filtering. The result shows an error rate of 3.41%.

## II. RESULT AND DISCUSSION

Our architecture uses coupled add-drop microring resonators linked with attenuator to perform the dot-product for the MVM task. In particular, as shown in figure 1a-c), the first column of microrings acts as WDM de-mux, while the second one acts as WDM combiner. The weights are implemented by attenuator, that can either be high-speed MZI, VOA, or P-RAM element. In this case, we use large ER slow-speed MZI. We realize the Photonic Integrated Circuit (PIC) using active AMF Silicon Photonic platform.

From the measured spectrum in figure 1d, we experimentally obtain the power map from the integrated photodiode using 2 input wavelengths and varying their weights. We then use our PTC to compute the edge detection from the George Washington, as shown in figure 1f-i). The used kernel detects the edge from the four directions, using 3-bit weights. The result shows a good agreement with the computed result from commercial software, indicate the good performance of our

PTC also for large figure. Considering as errors all the PTC pixels that don't match with the ideal result, we can compute an error rate of 3.41%. Future implementation will focused on the fast variation of weights for online training, and edge computing by using P-RAM components integrated in the PIC [6].

## REFERENCES

- [1] Trends Report 2020 Global 2020 Global Networking Trends Report The evolving role of the IT network 7, Cisco, 2021.
- [2] Shen, Yichen, et al. "Deep learning with coherent nanophotonic circuits." *Nature Photonics* 11.7 (2017): 441-446.
- [3] Feldmann, Johannes, et al. "Parallel convolutional processing using an integrated photonic tensor core." *Nature* 589.7840 (2021): 52-58.
- [4] Huang, Chaoran, et al. "A silicon photonic-electronic neural network for fibre nonlinearity compensation." *Nature Electr.* 4.11 (2021): 837-844.
- [5] Meng, Jiawei, et al. "Electrical Pulse Driven Multi-Level Nonvolatile Photonic Memories Using Broadband Transparent Phase Change Materials." *arXiv preprint arXiv:2203.13337* (2022).
- [6] Peserico, Nicola, et al. "Emerging devices and packaging strategies for electronic-photonic AI accelerators: opinion." *Optical Materials Express* 12.4 (2022): 1347-1351.