Frontiers in Optics + Laser Science 2023 (FiO/LS) © Optica Publishing Group 2023

Fully Integrated Photonic Tensor Core Accelerator for Neural Network Applications

N. Peserico[1,2], X. Ma[1], B. Movahhed Nouri[1], H. Dalir[2], B. J. Shastri[3], V. J. Sorger[1,2]

1. Department of Electronics and Computer Engineering, George Washington University, 800 22nd St, 20052, Washington, DC, USA
2. Department of Electrical and Computer Engineering, University of Florida, Gainesville, Florida 32611, USA
3. Department of Electronics and Computer Science, Queens University, Kingston, Canada

volker.sorger@ufl.edu

***Abstract*— Machine Learning applications have exploded in recent times. Here, we present the first fully-integrated Photonic Tensor Core accelerator, capable of computing Neural Networks by integrating all the optical components, from laser to photodetectors.**

Keywords*— *Silicon Photonics, Packaging, Neural Networks, Laser, Heterogeneous Integration

I. Introduction

The new industrial revolution is being driven by applications of artificial intelligence, namely Machine Learning (ML) applications based on Neural Networks, which have recently begun to reach their full potential [1]. How far ML can advance for the general audience has been demonstrated by applications for text generation and complicated image synthesis. However, since those networks are now working with billions of parameters for every single inference, training could take weeks for a data center to run, with the related high energy consumption of up to millions of dollars [2]. The need for a quicker and more effective method to compute matrix multiplication for neural networks is critical given the growth of ML applications.

By utilizing the characteristics of light propagation, such as low energy consumption, high bandwidth, and low latency, optics (and photonics) have demonstrated their ability to accelerate Neural Network computing [3,4]. Over the past few years, integrated photonics systems have been put out, demonstrating the ability to calculate matrix-vector operations quickly and with minimal latency. However, as of right now, most of the options for dealing with online training either lack complete integration or a high-speed weight update [5], two important components for the full-scale deployment of photonic computing solutions.

Here, we present our Fully Integrated Photonic Tensor Core (PTC), based on Photonic Wire Bonding InP laser integration and Silicon Photonics technology for high-speed input rate, weights update, and fast read-out. This solution has been packed into a QFN carrier without any need for fiber coupling, for easier employment on high-speed electronic PCB boards.

II. Design and Result

A. Photonic Tensor Core Design

The proposed PTC is formed by a main Silicon Photonic chip, where an array of 3 DFB lasers is placed on a trench (Fig. 1a-b). Photonic Wire Bonding is used to integrate and link the lasers to the silicon waveguides [6,7] with low losses. High-speed Mach-Zhender modulators are used to encode the input vector over the 3 wavelengths. After being multiplexed, the signals are routed into a 3x3 matrix of add-drop Microring Modulators, which multiply the signals by dividing the optical power between the two output

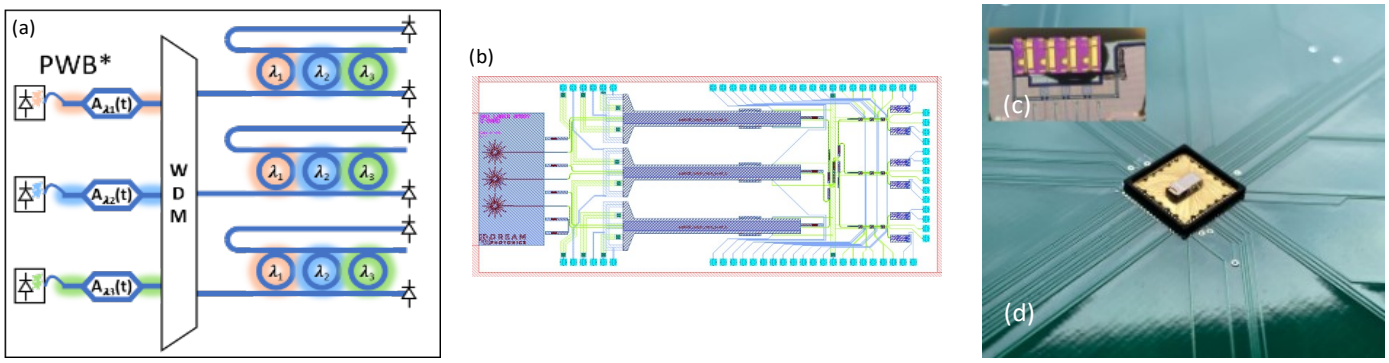

Figure 1: Scheme and photo of the realized Photonic Tensor Core. (a) Scheme of the PTC, with integrated laser sources, input vector encoding modulators, microring modulators weight matrix, and balanced photodetectors. (b) GDS design of the PTC. (c) DFB Laser array mounted on the Silicon Photonic chip. (d) Photo of the realized fully integrated PTC on a QFN package. No optical fiber is needed.

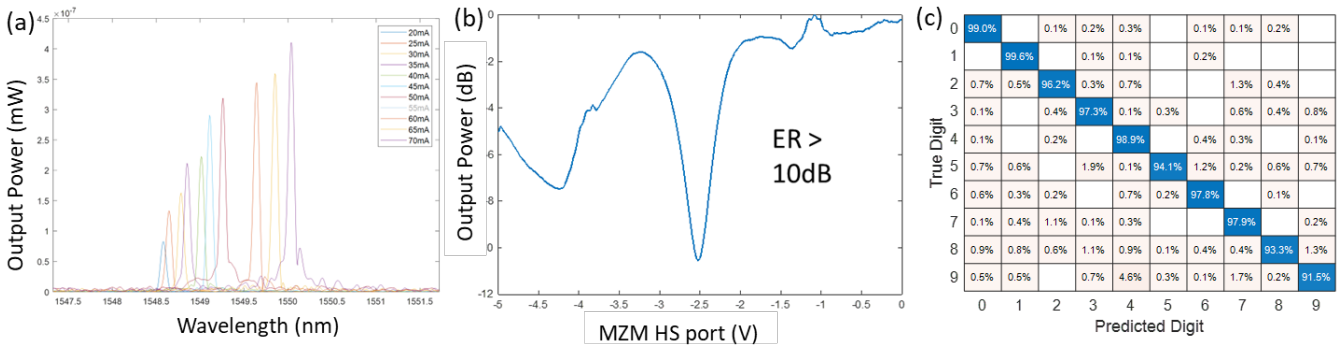

Figure 2: Measures of the different PTC components and results. (a) Experimental spectra of one integrated laser, obtained from a testing coupler on the Silicon Photonic chip. (b) Measure of the optical response of the MZM varying the p-n junction port. (c) Results of an emulation of a Neural Network for MNIST digit recognition.

branches. The output power is collected by two SiGe photodetectors. The silicon photonic chip has been realized by AIM Photonics using one of its Silicon Photonic active MPW shuttle runs. A 4-fold InP DFB laser bar is used to create the lasers, and it is embedded inside a silicon photonic chip with a deeply etched cavity (Fig. 1c). The photodetectors, as well as the vector and matrix encoding components, have been chosen to achieve the highest speeds and Signal-to-Noise Ratios possible (Fig. 1b).

The entire heterogeneous integrated chip has been packed inside a QFN carrier, creating the first "photonic black box" in which all of the optics are contained within the carrier and only electrical I/O is provided (Fig. 1d). Without the use of fiber arrays or external laser sources, this sort of packaging enables the integration of one or more photonic chips into intricate electrical circuits.

B. Initial Results

As a first step to test the chip, we used a grating coupler placed after the Mach-Zehnder Modulator (MZM), which encodes the input vector, to evaluate the laser's reaction following integration. Around 1 dB is thought to be the PWB's predicted Insertion Loss. However, issues with the first batch of Silicon Photonic chips caused a number of the tapers that have to couple the PWB with the silicon waveguide, to be damaged, leading to a greater than anticipated IL (>20 dB), on top of the losses from the circuit's components. However, as the lasers exhibit good linewidth and are in good accord with the findings of the wafer-level datasheet, we examine the integrated laser spectrum at various current levels (Fig. 2a).

With an Extinction Ratio (ER) of more than 10 dB per branch, the spectrum of the MZM bank is demonstrated to be in good accord with the design (Fig. 2b). By using a binary OOK modulation to drive one input modulator and one linked microring modulator, we have measured the initial temporal outcomes. As a result, we have observed the overlapping of the two modulations, with an unbalanced due to the greater ER that MZM has than the microring weight. The first neural network simulation for MNIST digit recognition demonstrates a high accuracy of over 95% (Fig. 2c).

Since we select high-speed input rate and weights update modulators, the whole chip can operate at a bandwidth of over 20 GHz, increasing performance to 810 GOPS/W. Additionally, the integration enables chiplet-style stacking of numerous chips onto a single PCB or Si carrier, enabling the achievement of over 20 TOPS per single chiplet.

ACKNOWLEDGMENT

V.J.S. is supported by the PECASE Award under the AFOSR grant (FAA9550-20-1-0193).

REFERENCES

[1] Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." Science 349.6245 (2015): 255-260.

[2] Batra, Gaurav, Zach Jacobson, Siddarth Madhav, Andrea Queirolo, and Nick Santhanam. "Artificial-intelligence hardware: New opportunities for semiconductor companies." McKinsey and Company, January 2 (2019).

[3] Miscuglio, Mario, and Volker J. Sorger. "Photonic tensor cores for machine learning." Applied Physics Reviews 7.3 (2020): 031404.

[4] Shastri, Bhavin J., et al. "Photonics for artificial intelligence and neuromorphic computing." Nature Photonics 15.2 (2021): 102-114.

[5] N. Peserico, B. J. Shastri and V. J. Sorger, "Integrated Photonic Tensor Processing Unit for a Matrix Multiply: A Review," in Journal of Lightwave Technology, doi: 10.1109/JLT.2023.3269957.

[6] Billah, Muhammad Rodlin, et al. "Hybrid integration of silicon photonics circuits and InP lasers by photonic wire bonding." Optica 5.7 (2018): 876-883.

[7] M. Mitchell et al., "Photonic Wire Bonding for Silicon Photonics III-V Laser Integration," 2021 IEEE 17th International Conference on Group IV Photonics (GFP), Malaga, Spain, 2021, pp. 1-2, doi: 10.1109/GFP51802.2021.9673842.