Fully Integrated Photonic Tensor Core for Neural Network Applications

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

2023 IEEE Photonics Conference (IPC) | 979-8-3503-4722-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/IPC57732.2023.10360525

***Abstract*— Neural Networks applications have exploded in recent times. Here, we present the first fully-integrated Photonic Tensor Core, capable of accelerating the computation of Neural Networks by exploiting a multi-wavelength light approach. In a single chip, we integrated all the optical components, from laser to photodetectors.**

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

I. Introduction

Artificial Intelligence applications are the driving force of the new industrial revolution, with Machine Learning (ML) applications based on Neural Networks that have been showing their initial full potential in the last years [1]. Applications for text generation and complex image production have shown how far ML can achieve for the general public. However, those networks are now working with billions of parameters for every single inference, while training could take weeks for a data center to run, with the related high energy consumption. With the expansion of ML applications, the need for a faster and more efficient way to compute matrix multiplication for Neural Networks is essential.

Optics (and photonics) have proven to be able to accelerate Neural Networks computation by leveraging light propagation properties, such as low energy consumption, high bandwidth, and low latency [2,3]. Integrated photonics solutions have been proposed over the last years, showing the possibility to compute matrix-vector operations at high speed and low latency. However, up to now most of the solutions lack either full integration or a high-speed weight update for addressing online training (fig. 1a) [4].

Here, we present our Fully Integrated Photonic Tensor Core (PTC), based on Photonic Wire Bonding laser integration and Silicon Photonics technology for high-speed input rate, weights update, and photodetectors. This solution has been packed into a QFN carrier, for easier employment into high-speed electronic PCB.

II. Design and Result

A. Photonic Tensor Core Design

The proposed PTC encodes the input vector among 3 DFB lasers utilizing high-speed modulators. The lasers are integrated and connected to the chip by using Photonic Wire Bonding [5,6].

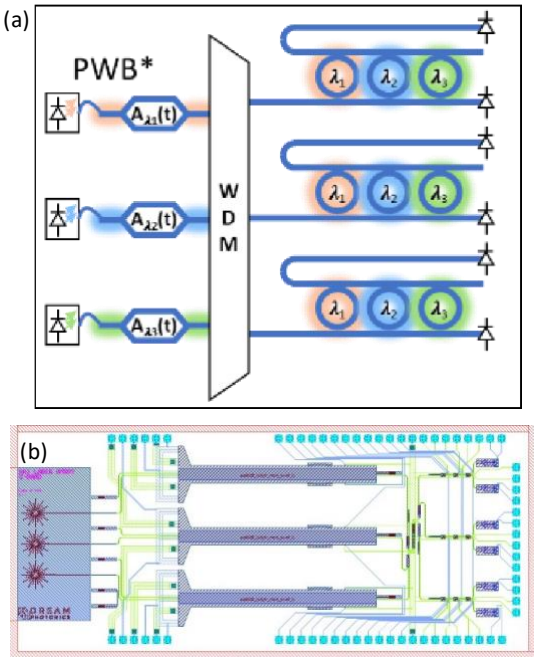

Fig. 1. (a) Scheme of the Photonic Tensor Core for a 3x3 matrix -vector multiplications. (b) GDS layout of the Silicon Photonic Chip.

The signals are then multiplexed and fed into a 3x3 matrix of Microring Modulators, which perform the multiplication by redistributing the optical power between the two output branches. Two SiGe photodetectors are collecting the output power and return it to the analog electronics domain (Fig. 1a).

The chip has been realized by AIM Photonics, employing one of its Silicon Photonic active MPW shuttle runs. The lasers are formed by a 4-fold InP DFB laser bar, integrated into a deeply etched cavity of the Silicon Photonic chip. The components for encoding the vector and matrix, as well as the photodetectors, have been selected to reach the maximum performances in terms of speed and Signal-to-Noise Ratio (fig. 1b).

The overall heterogeneous integrated chip has been packed into a QFN carrier, realizing the first "photonic black box", where all the optics is placed inside the carrier, providing just electrical I/O (fig. 2). This type of packaging allows to integrate one or multiple photonic chips into complex electronic circuits, without the need for fiber arrays or external laser sources.

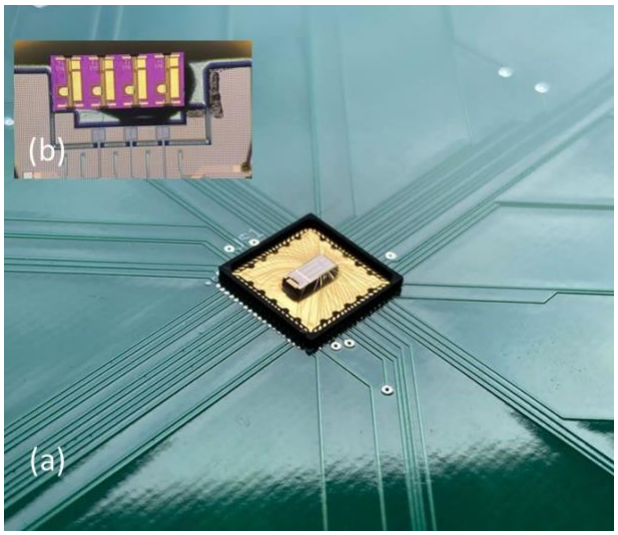

Fig. 3. *(a) Photo of the PTC with integrated laser over a QFN carrier . (b) Zoom of the 4 -fold laser bar placed over the same chip, connected with Photonic Wire Bondings.*

B. Initial Results

As preliminary results, we measured the response of a laser after the integration, utilizing a grating coupler placed after the Mach-Zendher Modulator (MZM) that is encoding the input vector. The expected Insertion Loss of the PWB is estimated at around 1 dB. However, due to complications of the first batch of Silicon Photonic chips, the tapers were several damaged, resulting in a higher than expected IL (>20 dB), on top of the losses due to the components on the circuit. However, we measure the integrated laser spectrum at different current levels (Fig. 3a), as the lasers show good linewidth, having a good agreement with the wafer-level datasheet results.

The spectrum of the MZM bank is shown in good agreement with the design, having over 10 dB Extinction Ratio (ER) per branch. We measure initial temporal results by driving an input modulator and one linked microring modulator with a binary OOK modulation. The result shows a correct inference by the 2 overlapping modulations, as the MZM shows a higher ER than the MRM. Initial emulation of a Neural Network for MNIST digit recognition shows a high accuracy of over 95%.

The overall chip can work to a bandwidth of over 20 GHz, for both input rate and weights update, bringing the performance to 810 GOPS/W. Moreover, the integration allows one to stack multiple chips into the same PCB or the same Si carrier, in a chiplet fashion, allowing one to reach over 20 TOPS per single PCB.

ACKNOWLEDGMENT

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

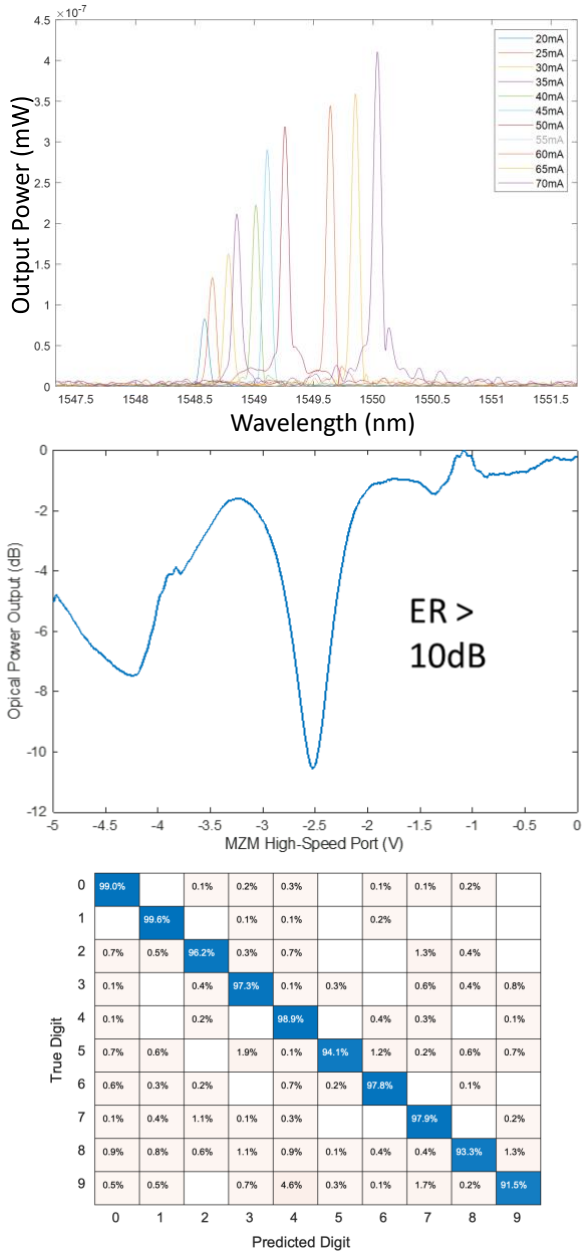

Fig. 2. *(a) Integrated Laser spectrum acquired from the Silicon Photonic chip. (b) Response of the input MZM for encoding the input vector. (c) Emulation results for MNIST digit recognition by Neural Network based on PTC.*

REFERENCES

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

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

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

[4] 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.

[5] Billah, Muhammad Rodlin, et al. "Hybrid integration of silicon photonics circuits and InP lasers by photonic wire bonding." Optica 5.7 (2018): 876883.

[6] 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.