Design and testing of a Silicon Photonic Tensor Core with integrated lasers

1st Xiaoxuan Ma
The Department of Electrical and Computer Engineering
The George Washington University
Washington, DC, USA
xxma94@gwu.edu

2nd Nicola Peserico
The Department of Electrical and Computer Engineering
The George Washington University
Washington, DC, USA
npeserico@email.gwu.edu

3rd Bhavin J. Shastri
Department of Physics, Engineering Physics and Astronomy
Queen's University
Kingston, Country
shastri@ieee.org

4th Volker J. Sorger
The Department of Electrical and Computer Engineering
The George Washington University
Washington, DC, USA
sorger@gwu.edu

2023 IEEE Silicon Photonics Conference (SiPhotonics) | 978-1-6654-8655-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/SIPHOTONICS55903.2023.10141963

***Abstract*— Here we present a reliable architecture to perform Matrix-Vector Multiplication exploiting the integration capability of Silicon Photonics, providing lasers and photodetector on-chip. By using the GEMM compiler, we can process images for edge detection, with an error rate lower than 7%.**

Keywords—Photonics, Photonic Tensor Core, Laser Integration

I. Introduction

The exponential increasing generation and elaboration of data and the wide use of artificial intelligence have pushed common computing architectures toward their limits in terms of speed and throughput [1-3]. In particular, Neural Network algorithms are not well suited for common CPUs, as they rely much on the Multiplication and Accumulation (MAC) operation, which CPUs are not designed to perform efficiently [4]. To overcome this limitation novel architectures and paradigms have been developed [5-6]. Among many, optical computing has shown important progress thanks to the almost unlimited bandwidth, the high energy efficiency in performing MAC operations, and the high integration. In particular, in recent years, many architectures have been proposed exploiting the electromagnetic nature of the light propagating in nanoscale structures to perform multiplication and accumulation [6-11]. However, till now, all those approaches have missed the high integration and parallelism that Silicon Photonics can add to the circuits.

Here, we develop and demonstrate a highly integrated Silicon Photonic Tensor Core. By using photonic wire bonding [12], three tunable lasers are integrated into the circuit. Three high-speed Mach-Zehnder modulators (MZMs) are used for encoding the input vector. By using the Y-junctions and three columns of micro-ring resonators with three different resonance frequencies, we achieve on-chip WDM and high-speed weight updates with more than 8 dB extinction ratio. Light signals are then collected by integrated high-speed photodetectors. Based on this PTC architecture, image convolution is achieved in MZM-based weight system, with an error rate of the edge detection kernel is 6.27%.

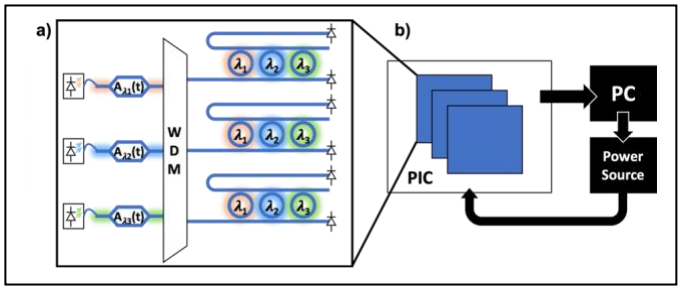

Fig. 1. (a). Schematic of photonic tensor core with laser integration. (b) Schematic of control setup for photonic tensor core.

II. Result and Discussion

A. Result

For performing matrix-vector multiplication (MVM), the light with different wavelength are injected from the on-chip laser to photonic circuit. The input vector is encoded in different wavelengths by MZMs. Through the on-chip fan-out wavelength-division multiplexing (WDM), different

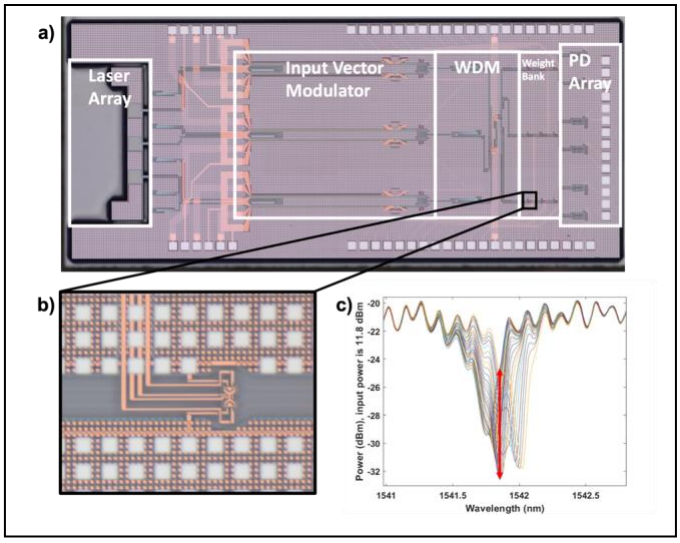

Fig. 2. (a) Microscope image of photonic tensor core with laser integration. (b) Microscope image of weight based on micro-ring modulater. (c) Performance of micro-ring modulator

978-1-6654-8655-2/23/$31.00 ©2023 IEEE

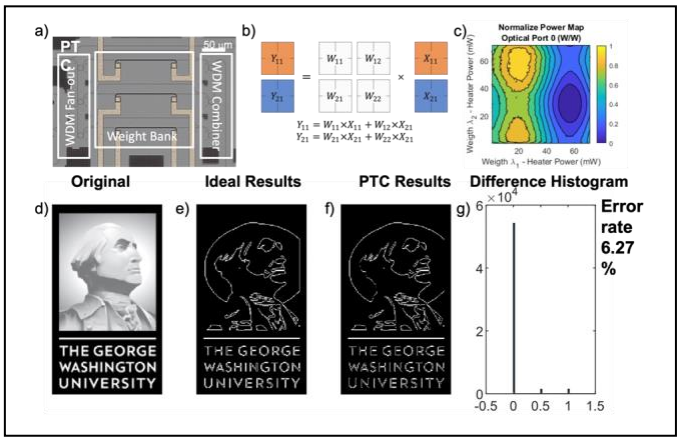

Fig. 3. (a) Microscope image of photonic tensor core with MZM based weight. (b) GeMM Compilers for processing input data matrices that are larger (i.e. more matrix entries) than the physical hardware (i.e. PTC). (c) Power map of one photonic tensor core. (d-g) Result of edge detection

wavelengths are separated and sent to the micro-ring modulator based weight bank. The weighted light signals with different wavelengths are combined in the bus waveguide and detected by the on-chip photodiodes (Fig. 1a).

We designed and taped out 3×3 photonic tensor core (PTC) from AIM Photonics. The single PTC includes a laser array, an input vector modulator array, a WDM combiner, a weight bank, and a photodiode array (Fig. 2a). The light with three different wavelengths is generated by the on-chip tunable lasers. Through the photonic wire bonding, the light is coupled into the photonic integrated circuit. The high-speed MZM array is used to encode the input vector by modulating the optical power coming from lasers. The encoded light is combined to bus exploiting a WDM scheme. The weight bank built by the high-speed micro-ring modulator array (Fig 2b). By tuning the micro-ring modulators, the differential combination of the weighted optical signals at different wavelengths can be achieved using bus and cross bus. Based on the preliminary measurements, the micro-ring modulators in our system have 8 dB extinction radio, which can result in a bit resolution of 4 bits (Fig. 2c). To complete the summation in this architecture, the on-chip photodiode array is used to sum the optical power of different wavelength from the through and cross bus waveguide. The current difference between the through and cross photodetectors, which is the result of MVM, is read by the computer. A power source is used to set the weight bank, controlled by the computer (Fig 1b).

To prove the function of this architecture, a photonic tensor core with MZM-based weight is tested (Fig 3a). The input is a one-bit GW face logo (Fig. 3c) which is binarized by a computer, while the edge filter is a 3-bit matrix with negative values, which is [0 -1 0; -1 4 -1; 0 -1 0]. To perform the computation, for each pixel the PTC computed the Hadamard product between the 3×3 filter and a 3×3 subset of the input image, using the GeMM compiler to fit the PTC matrix (Fig 3b). Since the filter has negative values, we divide the computation into 2 parts by adding an offset to the filter matrix and subtracting the same offset afterward. For performing dot product, the power map with a fixed input vector is tested. With different voltages applied to different MZMs, the output power is modulated at different wavelengths. The light power accumulates in photodetectors. The gradient change of total power is shown in the power map (Fig 3c). By using a power map as a look-up table (LUT), edge detection based on our photonic tensor core is achieved. The 3 bits kernel [0 -1 0; -1 4 -1; 0 -1 0] is encoded by voltage signals and sent to weight bank. The result of MVM is selected from the power map. Figure 3 d-e are the original picture and ideal result of edge detection, the PTC result of edge detection, and the difference pixel value between the ideal result and PTC result, with an error rate of our PTC of 6.27%.

B. Discussion

We have shown an architecture to achieve a photonic tensor core with on-chip lasers, WDM, and photodetector. By testing the micro-ring based weight, at least 4-bit resolution can be achieved. Also, a photonic tensor core with the same architecture based on MZM weights has been tested and used to prove the functionality of our PTC. With the MZM-based photonic tensor core, the edge detection of the image is achieved. Future steps will include proving the function in our full integration photonic tensor core.

ACKNOWLEDGMENT

This work was supported by the PECASE Award under the AFOSR grant (FAA9550-20-1-0193).

REFERENCES

[1] D. Amodei, “AI and Compute,” https://openai.com/blog/ai-and-compute 2020.

[2] Z. Yang, L. Zhang, K. Aras, I. R. Efimov, and G. C. Adam, “Hardware-Mappable Cellular Neural Networks for Distributed Wavefront Detection in Next-Generation Cardiac Implants,” Advanced Intelligent Systems, p. 2200032, 2022.

[3] B.Dong, M. Liu, and Y. Zhao, “Simulation of a Nano Plasmonic Pillar-Based Optical Sensor with AI-Assisted Signal Processing,” in ECS Meeting Abstracts, 2021, no. 61, p. 1641.

[4] Cong and B. Xiao, "Minimizing Computation in Convolutional Neural Networks", *Artificial Neural Networks and Machine Learning – ICANN 2014*, pp. 281-290, 2014.

[5] J. Peng et al., "DNNARA: A Deep Neural Network Accelerator using Residue Arithmetic and Integrated Photonics", *49th International Conference on Parallel Processing - ICPP*, 2020.

[6] Y. Shen et al., "Deep learning with coherent nanophotonic circuits", *Nature Photonics*, vol. 11, no. 7, pp. 441-446, 2017.

[7] N. Peserico et al., "Emerging devices and packaging strategies for electronic-photonic AI accelerators: opinion", *Opt. Mater. Express*, 12, 1347-1351 2022.

[8] J. Feldmann et al., "Parallel convolutional processing using an integrated photonic tensor core", *Nature*, vol. 589, pp. 52-58, 2021.

[9] X. Ma et al., " Photonic Tensor Core With Photonic Compute-in-Memory ", *Optical Fiber Communication Conference,* 2022.

[10] M. Miscuglio et al., "Photonic tensor cores for machine learning", *Applied Physics Reviews* 7, 031404, 2020.

[11] Shastri, B.J et al., "Photonics for artificial intelligence and neuromorphic computing", *Nat. Photonics* 15, 102–114, 2021.

[12] N. Lindenmann et al., "Photonic wire bonding: a novel concept for chip-scale interconnects", Optics Express, vol. 20, no. 16, p. 17667, 2012.