All-function Integrated Silicon Photonic Tensor Core (PTC) AI Accelerator

Nicola Peserico[1], Xiaoxuan Ma[1], Behrouz Movahhed Nouri[1,2], Bhavin J. Shastri[3], Hamed Dalir[1,2] , Volker J. Sorger[1,2*]

[1]Department of Electrical and Computer Engineering, George Washington University, Washington, DC, 20052, USA
[2]Optelligence LLC, Upper Marlboro, MD, 20772, USA.
[3]Department of Physics, Engineering Physics and Astronomy, Queens University, Kingston, Ontario, Canada
Author e-mail address: sorger(@gwu.edu, optelligence.co)

Abstract: Here we present our architecture for Silicon Photonic Tensor Core (PTC) capable of accelerating computational needs of neural networks and augmented/virtual reality applications. We present a novel fully-integrated PTC including chip-based lasers, modulators, and photodetectors.

1. Introduction: With the explosion of data-based applications, where fast elaboration of the incoming data streams from different sources is required, the request for highly specialized hardware accelerators has increased to match these demands. Applications that use Deep Neural Networks, Augmented Reality, and Virtual Reality bring huge demand for hardware capable of performing mathematical operations with low latency and high bandwidth [1]. Moreover, the same hardware must limit its energy impact, so that it can be implemented in network-edge devices, such as smartphones or drones. One of the main tasks that are required by those new applications is performing tensor operations, mainly matrix multiplications. The digital electronics field has specialized in providing Application-Specific Integrated Circuit (ASIC) to better perform those tasks, with GPU or FPGA as solutions. However, those solutions come with high energy consumption, and high latency, due to the speed limit of digital electronics computation. Optics and photonics provide a solution that can overcome the limitation of digital electronics, thanks to the physical properties of light [2]. The virtual-zero losses of optical mode, the possibility to perform interference between different optical beams, and the low latency of light propagation give photonics an advantage over the electronic counterpart, which has to deal with the charging of RC lines and clock cycles. Moreover, thanks to the progress of Silicon Photonics, photonics accelerators can be realized on-chip and large scale, without compromising any performances. Here, we present our Silicon Photonic architecture to perform Matrix-Vector Multiplication (MVM) on-chip. The architecture shows the possibility to implement the weight matrix with different technologies, to better adapt the Photonic Integrated Circuit (PIC) to the target application and requirements. While the implementation using PCM-based P-RAM components was already presented [3], here we show the implementation of the PIC using MZI and MRR, which shows the potential of this approach. Moreover, we will overview the next-generation approach that integrates all the needed components, from the lasers (using Photonic Wire Bonding) to the high-speed photodetectors.

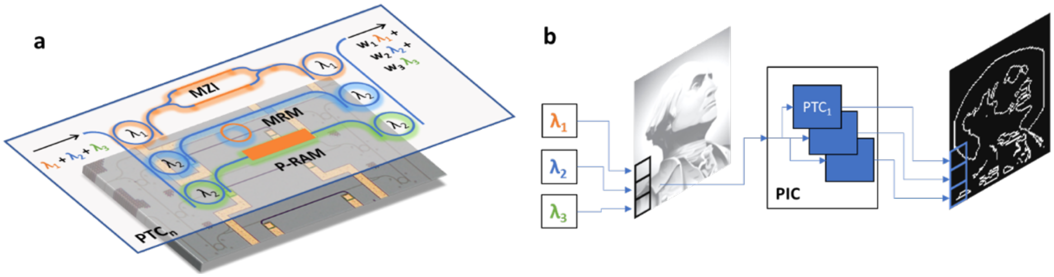

Figure 1: PTC architecture. (a) Design of the single PTC element, performing the scaled combination of the combined WDM inputs. (b) System overview, where the lasers source are modulated based on the input and sent to the PIC, that performs the MVM operation.

2. Results: The PIC exploits a WDM scheme, where the data are encoded by amplitude modulators and coupled into the MVM part of the circuit, where a sequence of WDM demux, amplitude modulators, and WDM mux perform the multiplication part of the computation, and an integrated photodetector combines the beams, returning the final result (fig. 1a). This single vector multiplication can be scaled into a proper MVM directly on-chip (fig. 1b). We showed, in a previous paper [3], the performance of this architecture using PCM-based P-RAM components, in terms of energy efficiency and accuracy, due to the non-volatile properties that do not require fixed voltage or power to maintain the optical properties of the material. In the new implementation, we use MZI to perform the weight of the inputs. This solution allows for a faster reconfiguration of the weight matrix and direct implementation on SiPh commercial foundries. We realized a SiPh chip (fig. 2a) implementing our PTC. The chip has two sections, each performing a 3x3 MVM, with optical input from grating couplers. The two sections differ in the output, as one has grating couplers, while the second one has integrated photodetectors. This last one presents better results in terms of MVM operation, thanks to the more linearity provided by the MZIs. Fig. 2b shows the power map when 2 CW lasers (1554.8 nm and 1550.1 nm) are coupled into the chip, as it performs the summation varying the weights. Since the MZI presents a cycled spectral response, only a subset of the total power map is used to perform the MVM. On this subset, to verify the accuracy, we extract the linear summation (fig. 2c) and compear it with the ideal output. As shown, the experimental results follow the ideal one from 1.6 to 8.8 uA. Considering the noise level of the photodetectors (50 nA), we can compute an equivalent bit resolution of 5.7 bits. The high-speed photodetectors allow reaching a bandwidth of over 10 GHz, while the latency is due to just the propagation of the light into the 3x4 mm^2 chip. We tested our system with different applications to prove its performance [4]. In the first test, we run an edge detection over the GW face from the GWU logo. The weight matrix used has negative values, which we decompose into 2 submatrices, both with just positive numbers, and combined the results from the PTC to return the right values. The error rate over the image was estimated at 3.4%. The second application is shown in fig. 2d-g, and it consists of image elaboration for augmented reality, where a generated mask is placed on top of the GW face. In this case, the result shows good agreement and an error rate of 0.3%. In the last series of tests, we emulate a Convolution Neural Network using the power maps obtained from the PIC, to perform digit recognition over the MNIST dataset. In this case, the accuracy reached over 96% when the PTC was used in the two convolution layers while performing the fully connected one on the computer. This solution is the most efficient one in terms of accuracy and speed over energy, opening the path for hybrid systems to be implemented. Integrated- and Silicon photonics can be the driving force for the next generation of hardware accelerators for many applications that require high speed and low latency, especially when PIC-IC co-designs approaches are utilized [5].

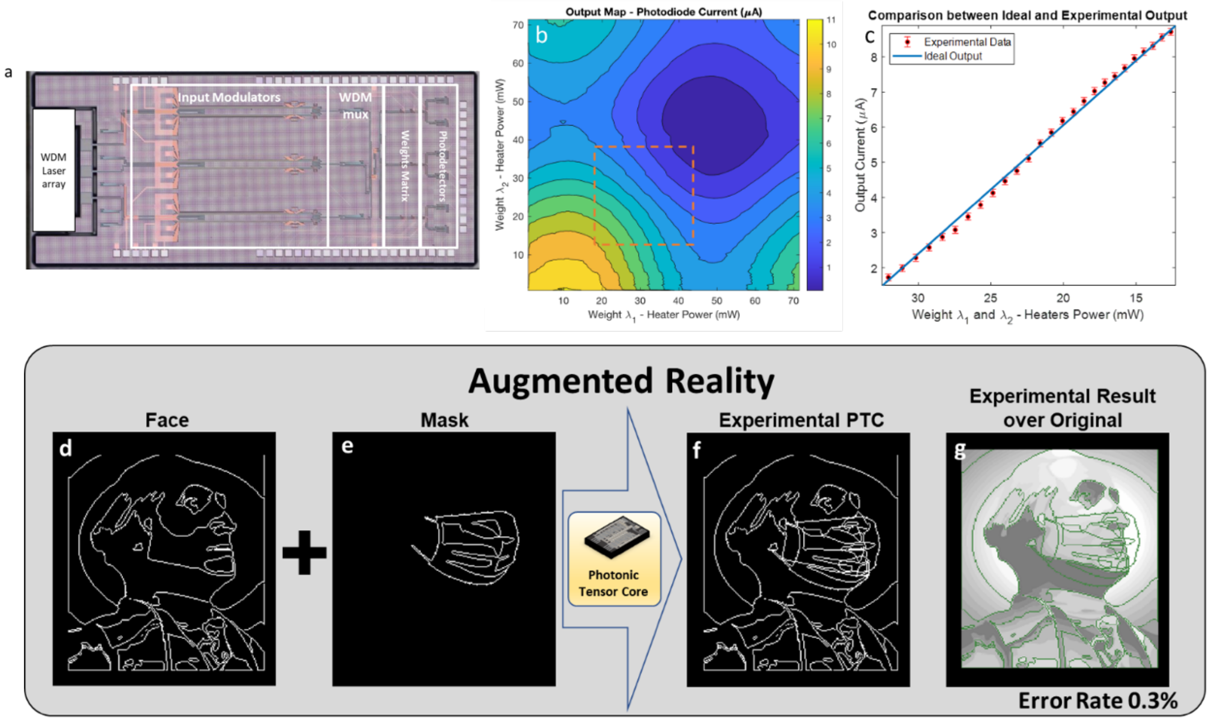

Figure 2: PTC performance and applications. (a) Photo of the SiPh PTC with integrated photodetector. (b) Power map obtained varyng two weights over the full MZI span. (c) Extraction of the linear summation of the 2 weights, and the error scale. (d-g) Example of Augmented Reality application, where a Mask image is placed on top of GW face for GWU logo.

3. References

[1] M. I. Jordan,T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol. 349, no. 6245, pp. 255–260, Jul. 2015.
[2] B. J. Shastri, et al., "Photonics for artificial intelligence and neuromorphic computing," Nature Phot., vol. 15, no. 2, pp. 102–114, Jan. 2021.
[3] Ma, Xiaoxuan, Jiawei Meng, Nicola Peserico, Mario Miscuglio, Yifei Zhang, Juejun Hu, and Volker J. Sorger. "Photonic Tensor Core with Photonic Compute-in-Memory." In Optical Fiber Communication Conference, pp. M2E-4. Optica Publishing Group, 2022.
[4] Ma, Xiaoxuan, Nicola Peserico, Ahmed Khaled, Zhimo Guo, Behrouz Nouri, Hamed Dalir, Bhavin Shastri, and Volker Sorger. "High-density integrated photonic tensor processing unit with a matrix multiply compiler." (2022).
[5] Peserico, Nicola, Thomas Ferreira de Lima, Paul Prucnal, and Volker J. Sorger. "Emerging devices and packaging strategies for electronic-photonic AI accelerators: opinion." Optical Materials Express 12, no. 4 (2022): 1347-1351.
[6] De Lima, Thomas, et al. "Primer on silicon neuromorphic photonic processors: architecture & compiler." *Nanophot.* 9, 4055-4073 (2020).