Photonic Tensor Core with Photonic Compute-in-Memory

Xiaoxuan Ma[1], Jiawei Meng[1], Nicola Peserico[1], Mario Miscuglio[1], Yifei Zhang[2], Juejun Hu[2], and Volker J. Sorger[1]

[1]*Department of Electrical and Computer Engineering, The George Washington University, Washington, D.C., USA*
[2]*Department of Materials Science & Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA*
E-mail: sorger@gwu.edu

Abstract: Here we demonstrate a photonic tensor core based on a silicon photonics dot-product engine. Utilizing compact electronic phase-change-material based photonic memory and WDM we show the highest throughput density to date of 3.8 MAC/s/mm^2. © 2022 The Author(s)

1. Introduction

With the exponentially increasing amounts of data and the rapid development of artificial intelligence, the high throughput, short latency, and parallelized computing hardware is critical [1]. The critical operation of convolutions (conv) makes for about >90% of all machine learning (ML) compute effort [2-3]. The fundamental reason of why conv processing is demanding is the high mathematical runtime complexity scaling with $(N\text{-}K)^2$, where N is the data input and kernel matrix sizes, respectively (assuming squared matrices). For classification ML tasks the kernel is quasi static requiring rare kernel weight updating. Hence, for improved power efficiency, photonic ASICs should include compute-in-memory functionality to eliminate memory-access bottlenecks known from van-Neuman systems. In fact, a brief analysis shows a ~100x superior potential of photonic memory over state-of-art SRAM with respect to data baud rate (speed) and memory access energy; in brief, an SRAM has an access latency of 0.3ns and costing about 100fJ/access [4]. A photonic memory based on phase-chance-materials (PCM), once WRITTEN, requires only the photon creation and detection energies. The minimum power of foundry-based PIC detectors in the C-band, for example, are about 50nW for signals above 30GHz. Assuming a 1% for the laser wall-plug efficiency and optical losses on the PIC and coupling to the PIC (~2dB per coupler, PWB [5]), a memory READ (access) energy of a PCM-written photonic random-access memory (P-RAM) takes <1fJ/access for OOK signal at 30GHz data rates, or ~10fJ/access for a higher bit resolution (e.g., PAM16 for a 4-bit ML classifier). Thus, a generic photonic link offers MAC operations and memory access of 100x higher MAC/s/J/access than SRAM. Following this potential for PIC-based MAC acceleration, electro-optic reconfigurable photonic integrated circuits (PIC) have been predicted and demonstrated [6-7] to process the repeating conv-underlying MAC operation (multiply-accumulate) in inference tasks on off-chip trained kernels. Specifically, utilizing fiber optical-based discrete components (non-PIC based) [8] and PIC-based [9] demonstrations show the possibility for efficient photonic MAC and hence conv acceleration. Photonic tensor processors are bound by the same system integration-performance scaling laws as electronic ASICs (Fig. 1); demonstrated MAC throughputs are 11.3 TOPS and 2 TOPS for these two prototypes. However, each falls short with respect to a) chip density or system compactness, and b) electronic control of the PIC. In the former case, no PICs were deployed. A system based on discretely packaged components of the shelf (COTS) is not only bulky, but also costly, prone to errors and hence does not offer much to for scaling up the number of matrix-size in the optical neural networks (ONN), as Infinera arguably demonstrated over a decade ago already. The second shortcoming of former work is to rely on an optical signal to program the P-RAMs. That is, requiring multiple lasers and having to ensure on complex waveguide routing is not only cumbersome, but leads to reduced throughput density (TOPS/mm^2) and makes the use of optical MAC processing questionable since electronic signal triggers for these optical WRITE cycles must be used anyhow.

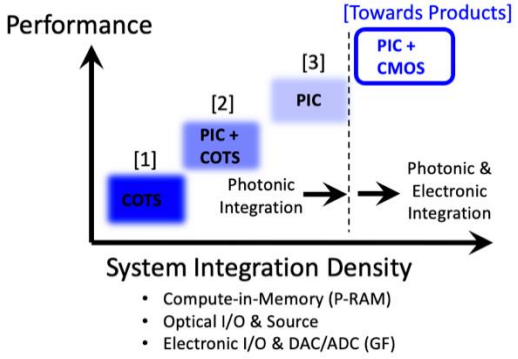

Fig.1. Photonic ASIC system performance improves with degree of integration. Here we introduce a photonic tensor core ASIC processor to accelerate MAC operations featuring electronic-on-chip programmable photonic random access memories (P-RAM).

To address both high MAC operation photonic ASPIC efficiency and compact system design, here we demonstrate the first photonic tensor core processor based on our previous design [10], which features electronically written P-RAM and compute-in-memory on-chip elements (Fig. 2). This photonic tensor core ASIC utilizes a one-dimensional degree-of-freedom multiplexer, namely WDM, to enable a N runtime complexity vector-matrix multiplier (VMM).

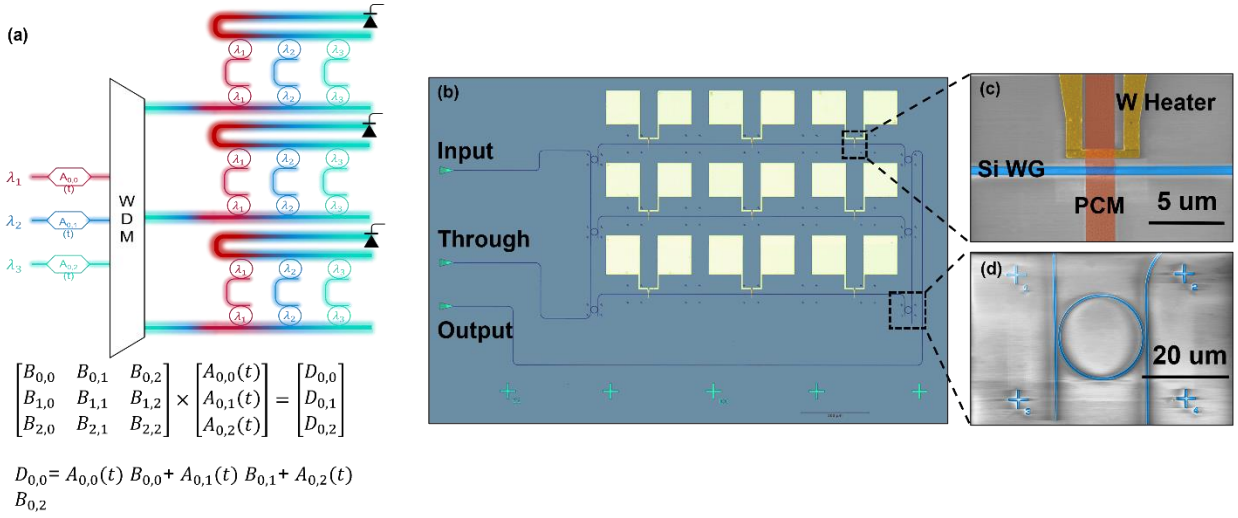

Fig.2. (a) Photonic tensor core architecture. (b). Optical micrograph of a fabricated 1×3 matrix. (c) The SEM image of PCM and heater. (d) The SEM image of micro-ring resonator (MMR)

Note, the matrix is the kernel stored in an array of P-RAMs and the *N* steps are determined by N wavelengths (hence VMM and not MM, matrix-matrix multiplier). The P-RAM are electro-thermally programed and feature not GST ($Ge2Sb2Te_5$), which is too lossy at the C-band, as non-volatile kernel weights, but GeSbSe (GSSe) instead. At 1550 nm wavelength, the GSSE has a significantly lower absorption coefficient (2.0×10-4) compared to GST (0.19) [11], which can significantly reduce the power consumption of our PTC system. Overall, this photonic tensor processor based on low-loss P-RAMs offers higher highest TOP/J/mm^2 as compared to previous demonstrations.

Here, we develop and demonstrate a highly integrated photonic tensor core based on non-volatile photonic memory by using phase change material (GeSbSe, GSSE). Our weight is built by depositing 3 stripes of GSSE with the same dimension on the top of the waveguide. By using the tungsten heater, the GSSE is switched between the amorphous and crystalline states. The absorption coefficient difference between the two states leads to the intensity change of light. By independently controlling the state of the 3 GSSE strip, the 2-bits weight with a total 1 dB extinction ratio (ER) is achieved.

2. Results and Discussion

2.1. Results

For performing matrix-vector multiplication (MVM), each element of a resulting matrix D is obtained through dot product (elementwise multiplication and summation) of each row of B by the vector A. We map this operation directly into the photonic hardware. The input vector A is encoded onto different wavelengths which are modulated by high-speed EOMs. Each wavelength is successively isolated by ring resonator (Fig.2d) and effortlessly weighted according to the tunable absorption coefficient of a multistate photonic memories (Fig.2c), which stores the element of the rows of the kernel B. The elements of B can be written via Joule heating and stored with no further power consumption. The P-RAM comprise an array of phase change material (GSSe) whose states are individually controlled by on-chip tungsten heaters. The number of GSSe stripes represent the number of bit-1 used for representing the kernel elements. After multiplications, light is detected by a high-speed photodetector, which performs the summation across wavelengths in the linear domain (Fig.2a).

We exemplary fabricated a 1×3 MVM that can be achieved by 3×3 photonic random-access memories (P-RAM) (Fig.2b). Here, we utilize a compact 2 micrometer PCM pads design resulting in an ER of ~1.0 dB or 0.5dB/μm (Fig.3b), which is one of the most area-efficient designs to date[4]. Extending the PCM strip size allows for a) higher ER such as 10dB for 20 micrometers, and b) also enabling multi-state programmability with a single control voltage pulse. We initialize the E-PTC by tuning the MRRs to achieve an equal output power on each wavelength channel. This accounts for the resonance frequency mismatch between MRR pairs to account for fabrication variances (Fig. 3a).

To compensate for the difference insertion loss of MMR, the 8 dB insertion loss as the initial power level is selected and adjusted by tuning the wavelength of injecting laser. The result of multi-time tuning does not affect the target power level at a significant level. To analyze the computational accuracy of the silicon photonics-based MVM

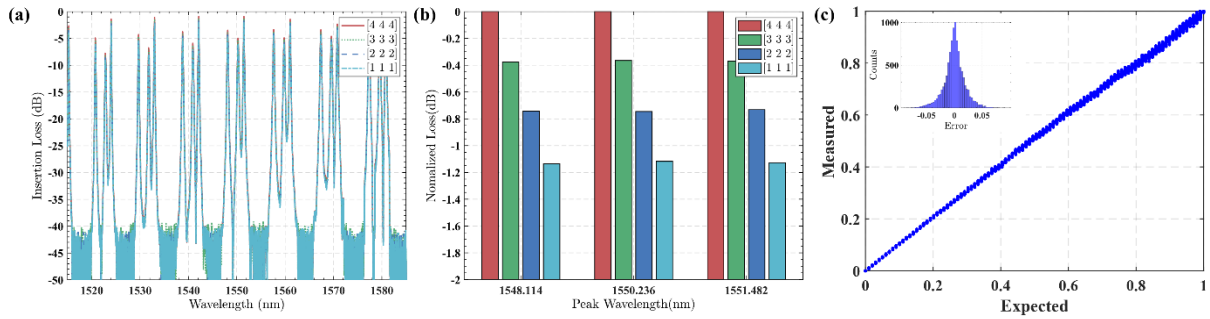

Fig.3. Measured (a) Spectrum, (b) bit resolution, and (c) accuracy of the 1×3 MVM kernal

operation, a randomly chosen 1×3 vector is processed using the different configuration kernels, and compared with the expected analytically calculated multiplication result. The normalized results from the E-PTC for 10000 MAC operations show the standard deviation of 0.019 and the mean value of 0.0016 (Fig. 3c). The fitted line shows the measured results well correspond with expected.

2.2. Discussion

We have shown a photonic tensor core based on photonic dot-product engine and demonstrated the photonic dot-production engine with P-RAM enables parallelizing MAC operations with 2 bits resolution of weight. The standard deviation of 0.019 and mean of 0.0016 of 10000 MAC operation test shows low error rate and stability of our PTC.

3. References

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

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

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

[4] T. Alexoudi et al., "Optical RAM and integrated optical memories: a survey", *Light: Science & Applications*, vol. 9, (2020).

[5] M. Billah et al., "Hybrid integration of silicon photonics circuits and InP lasers by photonic wire bonding", *Optica*, vol. 5, no. 7, p. 876, 2018.

[6] N. Harris et al., "Quantum transport simulations in a programmable nanophotonic processor", *Nature Photonics*, vol. 11, no. 7, pp. 447-452, 2017. Available: 10.1038/nphoton.2017.95.

[7] Y. Shen et al., "Deep learning with coherent nanophotonic circuits", *Nature Photonics*, vol. 11, no. 7, pp. 441-446, 2017. Available: 10.1038/nphoton.2017.93.

[8] X. Xu et al., "11 TOPS photonic convolutional accelerator for optical neural networks", Nature, vol. 589, (2021) , pp. 44-51.

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

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

[11] Y. Zhang et al., "Broadband transparent optical phase change materials for high-performance nonvolatile photonics", Nat Comms, vol. 10, (2019).