Photonic tensor cores for machine learning

Mario Miscuglio[1], Volker J. Sorger[1]

[1]*Department of Electrical and Computer Engineering, George Washington University, Washington, DC 20052, USA*

Abstract: With an ongoing trend in computing hardware towards increased heterogeneity, domain-specific co-processors are emerging as alternatives to centralized paradigms. The tensor core unit (TPU) has shown to outperform graphic process units by almost 3-orders of magnitude enabled by higher signal throughout and energy efficiency. In this context, photons bear a number of synergistic physical properties while phase-change materials allow for local nonvolatile mnemonic functionality in these emerging distributed non van-Neumann architectures. While several photonic neural network designs have been explored, a photonic TPU to perform matrix vector multiplication and summation is yet outstanding. Here we introduced an integrated photonics-based TPU by strategically utilizing a) photonic parallelism via wavelength division multiplexing, b) high 2 Peta-operations-per second throughputs enabled by 10's of picosecond-short delays from optoelectronics and compact photonic integrated circuitry, and c) zero power-consuming novel photonic multi-state memories based on phase-change materials featuring vanishing losses in the amorphous state. Combining these physical synergies of material, function, and system, we show that the performance of this 8-bit photonic TPU can be 2-3 orders higher compared to an electrical TPU whilst featuring similar chip areas. This work shows that photonic specialized processors have the potential to augment electronic systems and may perform exceptionally well in network-edge devices in the looming 5G networks and beyond.

I. Introduction

Aiming to replicate brain functionalities remains a captivating challenge, which does not only aspire and intrigue human feats, but also has shown to provide technological usefulness for modern societies. Indeed, Machine Learning (ML), performance by neural networks (NN), has become a popular approach to Artificial Intelligence (AI), and consists of training a system to learn how to perform unsupervised decision classifications on unseen data; once a NN is trained, it can be implemented to produce an *inference*, in other words recognizing and classifying objects or patterns.

Most NNs unravel multiple layers of interconnected neurons/nodes. Each neuron and layer, as well as the network interconnectivity, is essential to perform the task which the network has been trained for. In their connected layer, NNs strongly rely on vector matrix math operations, in which large matrices of input data and weights are multiplied, according to the training. Complex multi-layered deep NNs, in fact, require a sizeable amount of bandwidth and low latency for satisfying the vast operation required for performing large matrix multiplication (MM) without sacrificing efficiency and speed.

Since the dawn of the computing era, due to the ubiquity of matrix math, which extends to neuromorphic computing, researchers have been investigating optimized ways to efficiently multiply matrices. To corroborate this statement, engineering a platform which performs energy efficient and faster matrix multiplication enables solving linear algebraic problems, such as inverting matrices, systems of linear equations, and finding determinants. Even some basic graph algorithms are obstructed by the speed at which matrix multiplication is computed.

For a general-purpose processor offering high computational flexibility, these matrix operations take place serially, one-at-a-time, while requiring continuous access to the cache memory, thus generating the so called "von Neumann bottleneck". Specialized architectures for NNs such as Graphic Process Units (GPUs) and Tensor Process Units (TPUs), have been engineered to reduce the effect of the von Neumann bottleneck enabling cutting-edge machine learning models. The paradigm of these architectures is to offer domain-specificity such as being optimized for convolutions or Matrix Vector Multiplications (MVMs) performing operations, unlike CPUs, in parallel deploying a systolic algorithm.

GPUs have thousands of processing cores optimized for matrix math operations, providing tens to hundreds of TFLOPS (Floating point operations) of performance which makes GPUs the obvious computing platform for deep NN-based AI and ML applications. GPUs and TPUs are particularly beneficial with respect to CPUs, but when used to implement deep NN performing inference on large 2-dimensional

data sets such as images, they are rather power-hungry and require longer computation time (> tens of ms). Moreover, smaller matrix multiplication for less complex inference tasks (e.g. MIST[1]) are still challenged by a non-negligible latency predominantly due to the access overhead of the various memory hierarchies and the latency in executing each instruction in the GPU[2].

Given this context of computational hardware for obtaining architectures that mimic efficiently the biological circuitry of the brain, it is necessary to explore and reinvent the operational paradigms of current logic computing platforms when performing matrix algebra, by replacing sequential and temporized operations, and their associated continuous access to memory, with massively parallelized distributed analog dynamical units, towards delivering efficient post-CMOS devices and systems summarized as non von Neumann architectures. In this paradigm shift the wave nature of light and related inherent operations, such as interference and diffraction, can play a major role in enhancing computational throughput and concurrently reducing the power consumption of neuromorphic platforms. In recent years, the revolutionizing impact of NNs contributed to the development of a plethora of emerging technologies, ranging from free space diffractive optics[3] to nanophotonic processors[4–8] aiming to improve the computational efficiency of specific tasks performed by NN. Integrated photonic platforms can indeed provide parallel, power-efficient and low-latency computing, which is possible because analog wave chips can a) perform the dot product inherently using light matter interactions such as via a phase shifter or modulator, b) enable signal accumulation (summation) by either electromagnetic coherent interference or incoherent accumulation through detectors, and c) enable parallelism strategies and higher throughput using multiplexing schemes.

Additionally, we firmly believe, assisted by state-of-the-art theoretical framework[9], that future technologies should perform computing tasks in the domain in which their time varying input signals lay, exploiting their intrinsic physical operations. In this view, photons are an ideal match for computing node-distributed networks and engines performing intelligent tasks over large data at the edge of a network (e.g. 5G), where the data signals may exist already in the form of photons (e.g. surveillance camera, optical sensor, etc …), thus pre-filtering and intelligently regulating the amount of data traffic that is allowed to proceed downstream towards data centers and cloud systems[10]. Here we explore a photonic tensor core able of performing 4x4 matrix multiplication and accumulation with a trained kernel in one-shot (i.e. non-iterative) and entirely passive; that is, once a NN is trained, the weights are stored in an octet wide (8-bit) multilevel photonic memory directly implemented on-chip, without the need of neither additional electro-optic circuitry nor off-chip DRAM. The photonic memories feature low-losses phase-change nanophotonic circuits based on wires of $G_2Sb_2Se_5$ deposited on a planarized waveguides which can be updated using electrothermal switching and read all-optically. Electrothermal switching is enabled by tungsten heating electrodes which clamps the Phase Change Memory (PCM) wire.

This work represents the first approach towards the realization of a photonic tensor processor which could scale the number of multiply-accumulate (MAC) operations by several orders of magnitude while significantly suppressing power consumption and latency compared to the state-of-the-art hardware accelerators.

II. RESULTS AND DISCCUSION

II.1 Matrix multiplication algorithms

Considering a naïve (‘schoolbook’) algorithm, a multiplication between two square matrices, each with $n \times n$ entries, is characterized by a computational complexity of $O(n^3)$ (**Fig, 1**). Which means that the total operations required for performing this operation scales cubically with the size (n) of the square matrix. Even for optimized algorithms, such as Strassen[11] or Winograd[12], the complexity of the algorithm still requires a $O(n^{2.373})$ ***(Fig. 1a)***. Such ‘flat’ computational complexity scaling requires long latency when computed in regular CPUs since operations are executed sequentially at each clock cycle. A tensor core unit algorithm is an interesting alternative, however not for a reduced computational (operation) complexity, which is indeed still $O(n^3)$, but because of its capability of exploiting parallel architectures and a systolic algorithm[13]; for instance, it can implement a multiplication between two $n \times n$ matrices with an operational time complexity of $O(n^2)$, primarily dominated by reading/writing the input and output matrices,

even if the number of total operations is still $O(n^3)$.[14] In other words, one must distinguish between the complexities of the computational algorithm versus that of the system's execution time. In this regard, the computational complexity scaling results indeed suggest that the main focus should be placed on the fundamental improvements in the time complexity, thus diverting the focus onto hardware such as architectural, circuit, and component-level novelties including new device physics to speed-up the matrix processor's signal 'rate', or in exploiting parallelization strategies (**Fig. 1**). Interestingly, both suggest to consider optics or integrated photonics given a) the various physical multiplexing options (e.g. spectral, mode, polarization, etc...), and b) the short delay in the range of 1-10's of picoseconds from modern optoelectronics devices with 3-dB bandwidths of 10's GHz in photonic foundries [IMEC, AIMs] and even approaching 100's of GHz in laboratory settings.[15]

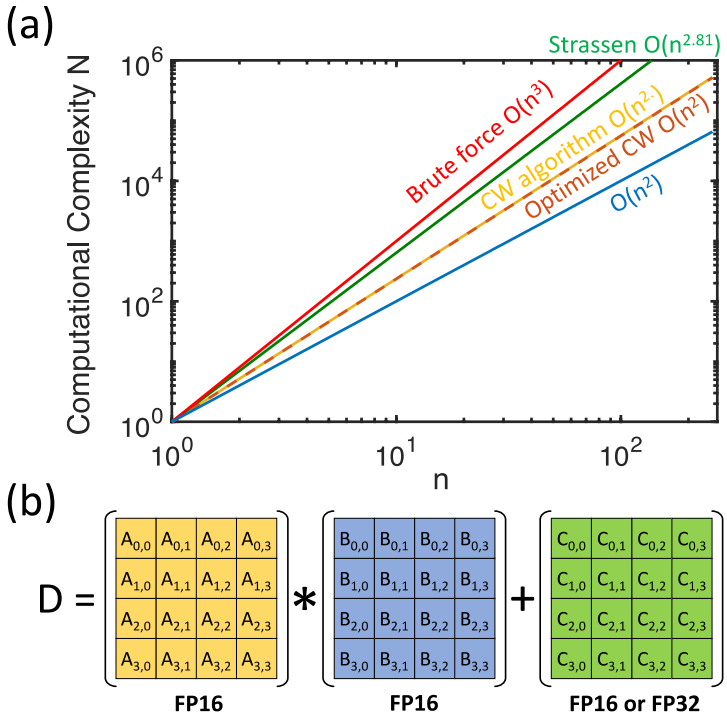

Figure 1. Matrix Multiplication. *(a) Computational complexity of a n × n matrices multiplication for different algorithms: Brute-force (naïve), Strassen, Coppersmith–Winograd, Optimized CW and the longed $O(n^2)$ algorithms when performed on GPUs TCUs. (b) 4x4 Matrix multiplication and accumulation performed by a GPU tensor core.*

For this reason, tensor cores are used to perform large-scale 2-dimensional, or higher dimensional, matrix operations built up from smaller elements, namely tensor core units (TCUs). Each TCU operates on a 4x4 matrices and performs the following operation: D = A×B + C where A, B, C, and D are 4x4 matrices. The matrix multiply inputs A and B are FP16 matrices, while the accumulation matrices C and D may be FP16 or FP32 matrices (**Fig. 1b**).[16]

In order to significantly reduce the execution time of large matrix multiplication or to avoid the non-negligible latency given by the time to collect data from a specific TCU, here we present a design based on silicon photonic TCUs performing matrix multiplication inherently with its latency time-of-flight only limited by the delay of the detection mechanism, which is well below 10's of ps short in modern high-speed photoreceivers[17–19].

II.2 Photonic Tensor Core architecture

The main advantage of performing a matrix multiplication and accumulation operations in the photonic domain is that they can be performed with almost zero power consumption while allowing for low

latency, given only by the time of flight of photon. To build a photonic tensor core, we use 16 fundamental units, namely dot-product engines, which perform an element-wise multiplication whilst featuring a Wavelength Division Multiplexing (WDM) scheme for parallelizing the operation (**Fig. 2a**). The dot product engine (**Fig. 2b**) performs the multiplication between two vectors, namely, between the i_{th} row of the input matrix A and the j_{th} column of the kernel B. In this scheme, the i_{th} row of the input matrix is given by WDM signals which, if not already in the optical domain, are modulated by high-speed (e.g. Mach Zehnder) modulators. The j_{th} column of the kernel matrix is loaded in the photonic memory by properly setting its weight states. Availing light-matter interaction with the phase-change memory, the inputs, opportunely spectrally filtered by microring resonators (MRR), are weighted in a seemingly quantized electro-absorption scheme (i.e. amplitude modulation), thus performing element-wise multiplication. The element-wise multiplications are thence incoherently summed up using a photodetector, which amounts to a MAC operation (D_{ij}). It is worth noticing that contrary to other photonic NN implementations[5,7,20] based on micro-ring modulators, the transmission of the microrings is not actively tuned for performing filtering, but just utilized for passively selecting frequency to be modulated by the photonic memories. This allows to have more control on the inter-channel crosstalk and potentially extending the number of wavelengths in a Dense WDM (DWDM) scheme without being affected by the induced quality factor variation caused by the variation of absorption coefficient.

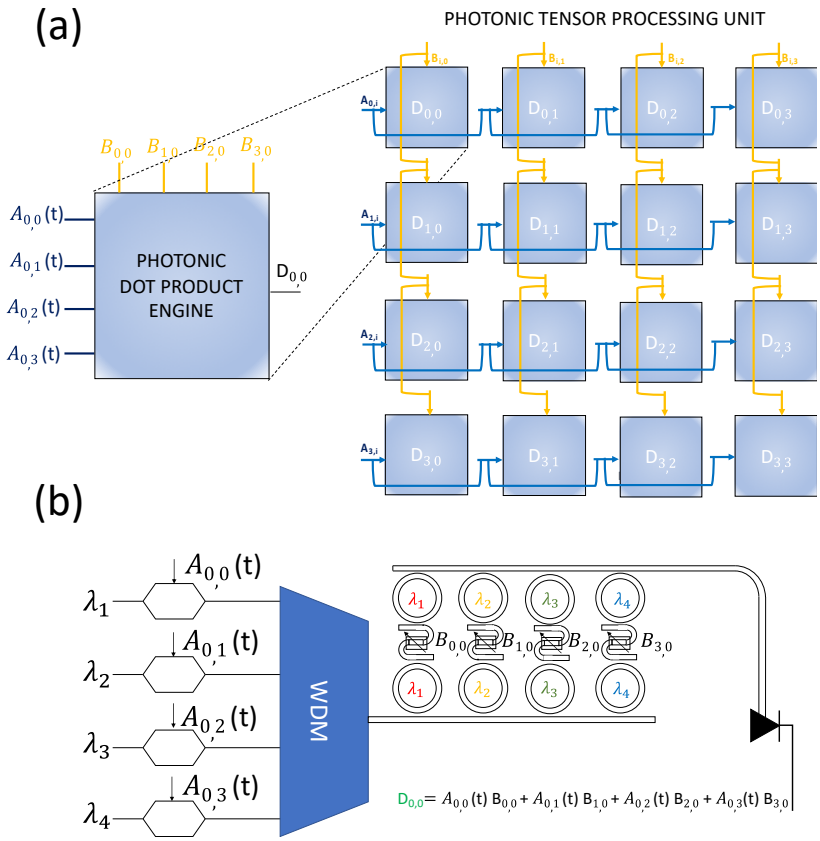

Figure 2. Photonic Tensor Core and Dot-product Engine. *(a) The photonic tensor core is constituted by 16 dot product engine which inherently and independently performs row by column pointwise multiplication and accumulation. (b) Photonics dot product engine is fed with WDM (λi) modulated inputs ($A_{0,i}$), which after opportune filtering, are weighted through photonic memory ($B_{i,0}$). Uncoherent summation is performed by photodetection mechanism ($D_{0,0}$).*

II.3 Photonic Memories

The benefits given by the intrinsic electromagnetic nature of the signals can potentially be hindered by the optoelectrical and electrooptical transductions, as well as by the repeated access to a digital and

nonvolatile memory, which impacts on the overall operation speed, while producing considerable additional energy loss. For this reason, having a heterogeneously integrated optimized photonic memory which retains information in a non-volatile fashion, poses a great advantage, especially when implementing NN performing inference, where the trained weights are only rarely update not very often (i.e. depending on the application daily, monthly, yearly, if ever). To provide this functionality, a multi-state photonic memory device, which comprises of multiple $Ge_2Sb_2Se_5$ photonic memory wires, is placed in between two resonant rings (**Fig. 1b**) to selects the opportune wavelength front and back, respectively (**Fig. 3a**). That is, once the PCM memory states are set (WRITE operation) in this photonic kernel (i.e. matrix *B*), this architecture allows performing the weighting functionality entirely passive. Selective 'writing' is achieved by changing the phase of the corresponding number of PCM wires that we deposited on the waveguides, by local electrostatic heating, which promotes crystallization or amorphization, and consequently modifies the waveguide modal refractive index in a reversible process.

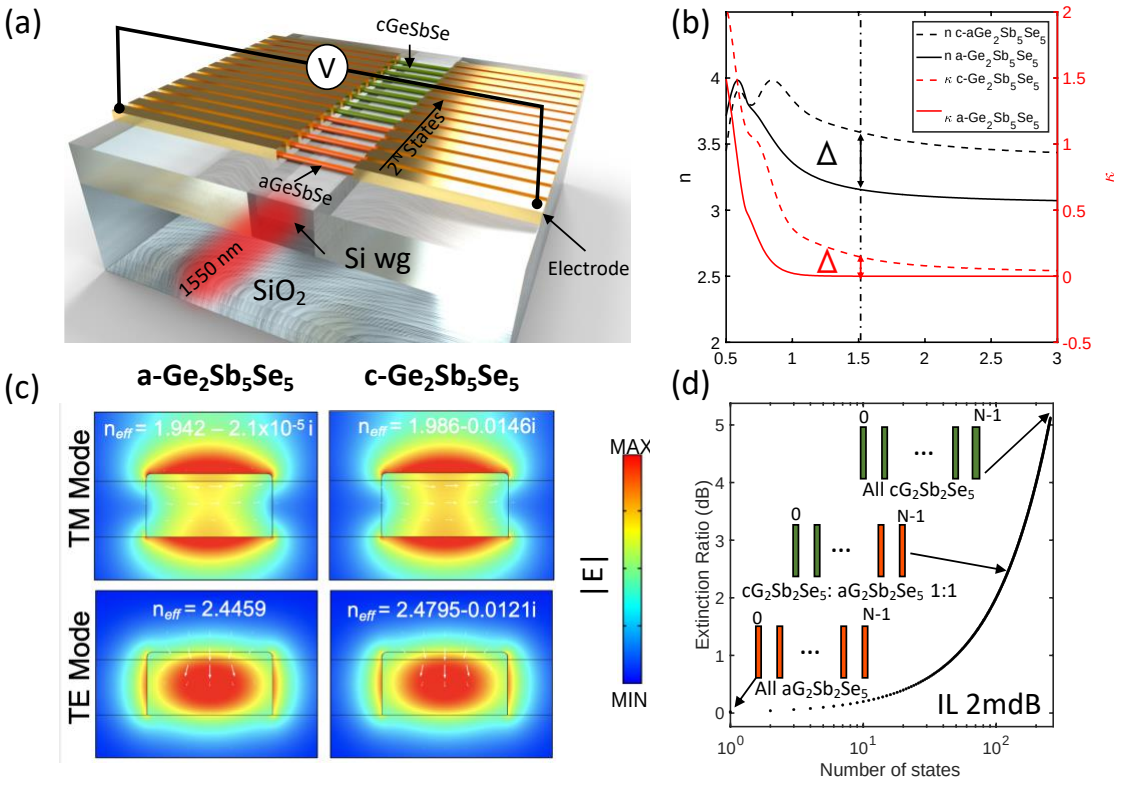

Figure 3. Multistate on-chip photonic memory embedded in the photonic tensor core. *(a) Schematic representation of the multistate reprogrammable photonic memory. 100 nm thick $Ge_2Sb_2Se_5$ wires can be patterned using a combination of lithographic process, sputtering and liftoff. The resulting wire can be contacted with tungsten electrodes which represents the heating elements. (b) Experimentally obtained (ellipsometry) optical properties of phase change material ($Ge_2Sb_2Se_5$) film. Real (n, left y-axis) and imaginary (κ, right y-axis) parts of the refractive indices of the amorphous (solid line) and crystalline alloys. (dashed line). The $Ge_2Sb_2Se_5$ shows a sensitive variation of the absorption coefficient, while simultaneously showing small induced loss. (c) Simulated interaction between the $Ge_2Sb_2Se_5$ nanowire and the fundamental optical mode for both the amorphous (left panel) and crystalline (right panel) state. Fundamental transversal electric (TE) and Transverse magnetic mode profiles (normalized electric field) of the $Ge_2Sb_2Se_5$-Silicon hybrid waveguide at 1550 nm for amorphous and crystalline state show a strong index (imaginary-part) difference ~ 0.01, while incurring a relatively low insertion loss. White arrows represent the direction and intensity of the magnetic field (Hx,Hy). (d) Extinction ratio as function of the number of nanowires (i.e. photonic multi-cell memory) that has changed states from amorphous (orange) to crystalline (green). The total extinction ratio provided by 256 nanowires 100 nm thick is 5 dB with negligible insertion losses (IL < 0.01 dB).*

We decide to implement the photonic memory kernels based on $Ge_2Sb_2Se_5$, since this material presents broadband transparent region for telecommunication wavelengths in its amorphous state and can be used to implement high-performance nonvolatile multistate photonic memories.[21] $Ge_2Sb_2Se_5$ exhibits 3-orders of magnitude lower absorption coefficient with respect to regularly employed GST at 1550 nm, and features still a high optical (real part) index contrast Δn of 0.5 across the near- to mid-IR bands and around 0.2 $\Delta\kappa$ in the C-band (**Fig. 3b**). Remarkably, the optical absorption in the amorphous state is vanishingly small and non-measurable when heterogeneously integrated in silicon photonics of ~100 micrometer long lengths. Moreover, the relatively lower variation of the absorption coefficient, indeed, makes it a promising material for multistate devices, avoiding the utilization of high laser power and extremely low noise equivalent power detectors.

When the network is trained, the extracted weights are set by changing through electrothermal switching individual states of photonic memories, instead of the previously used optical pulses[22]. Each state of the memory can reversibly be written, by selective transitioning between amorphous (*a*) and crystalline (*c*) phase using electrothermal switching induced by Joule heating, as previously demonstrated[23]. In our scheme heat is applied to the material externally via joule heating of a tungsten metal layer in contact with the wire, as shown in Figure 3a. Different pulse train profiles according to the type of transition (*a*-*c* or *c*-*a*) are applied to the wire via the connected in series to the device.[23]

Light signals that couple with this phase-change memory probes the variation of the absorption coefficient over phase transition (READ operation). For the fundamental TM mode of the waveguide the phase transition produces a variation of the effective absorption coefficient $\Delta\kappa \sim 0.01$ to which corresponds 0.21dB/μm (**Fig. 3c**). Considering a total of 256 states, the 100 nm thin programable PCM-wires arranged in a grating fashion (duty cycle 50%), it is possible to obtain an 8-bit memory for each element of the kernel (B_{ij}), with an total length of just ~50 μm, excluding electrical circuitry. The maximum extinction ratio (when all the wires are in the crystalline state) provides about 5 dB of modulation depth with a quantization step of 0.02 dB/bit (**Fig. 3d**).

The periodicity and intensity of the electric pulses applied to the tungsten electrodes, used for writing the memory, has to be adjusted for providing sufficient thermal energy to $Ge_2Sb_2Se_5$ wire according to in which phase has to be switched. The voltage, number of pulses and periodicity can be regulated to a) heat up the PCM wire up to 150°C and anneal for few tens of micro seconds to crystallize it b) melt it, increasing the temperature to over 600 °C, for the amorphization.[22,23]

II.4 Performances

The photonic tensor core implemented according the proposed scheme can perform matrix multiplication with 8-bit precision, completely passively once the weights are stored in the photonic network, which is a one-time operation. Considering the use of a photonic foundry Germanium photodetector and being inputs in the optical domain, a tensor core requires just 0.02 ns for processing a simple 4x4 matrix multiplication with a footprint which scales primarily with the number of wavelengths/MRR $NxNx2xA_{MRR}$. This translates into a chip of 1.6 mm^2 assuming 640 tensor core which performs a 4x4 matrix multiplication and accumulation. The electrical circuitry for read/out can be vertically interconnected through-silicon vias. Table I shows the comparison between a NVIDIA Tesla Core V100 and the photonic tensor core for the same number of cores.

Another key aspect to consider is the overall power consumption of the tensor core, which considering the completely passive operation of the dot-product and accumulation (and zero bias in detection with high responsivity photoreceivers), accounts only for the laser power which is below 5mW. The main limitation of the proposed system is related to the relatively small bit resolution compared to the GPUs tensor cores, since it is constrained by the footprint of the multistate photonic memories.

TABLE I. Comparisons of the figures of merit of the Photonic tensor core and Nvidia Tesla Tensor Core V100 [16]. MRR radius = 10 μm.

	Photonic Tensor Core	**Tesla V100**
Number of tensor cores	640	640
Clock (GHz)	50GHz	1.5GHz
Bit resolution (bit)	8bit	16bit-32bit
Throughput (Flops)	2 Peta operations O(n_3) at 8bit	125 TFlops/s
Power	5mW x Core + EO conversion <10 W	500 W
Footprint	1.6 mm_2 (only photonics components)	815 mm_2

III. CONCLUSION

In summary, we demonstrate a tensor core unit implemented in photonics. The system relies on photonic multiplexed (WDM) signals, weighted, after filtering, using engineered multistate photonic memories based on $Ge_2Sb_2Se_5$ wires patterned on the waveguide. The photonic memories can be reprogrammed by selectively changing phase (amorphous/crystalline) of the wires, using electrothermal switching through Joule heating induced by tungsten electrodes. The photonic memory programming can happen in parallel (few ms). An additional key feature of this design is that no additional losses are introduced by the photonic memories, avoiding repeaters, amplifiers (EDFAs) or cumbersome EO/OE conversions. The architecture shows execution time limited only by the time of flight of the photon in the chip, which is function of the ring size/selectivity (number of wavelengths), and the latency of the photodetector $\boldsymbol{O}(<10_{-1}$ns). The concurrent development of new PCM material and the advancement of the integration of photonic memories can enable the realization of engines based on the proposed scheme able to inherently perform full precision floating point matrix multiplication and accumulation, and consequently opening a pathway towards the realization of all-optical photonic tensor units which can significantly speed up intelligent tasks at the edge of the network without requiring EO conversions and access to external memories.

ACKNOWLEDGMENTS

We acknowledge the support from the Presidential Early Career Award for Scientist and Engineers (PECASE) nominated by the Department of Defense (AFOSR).

Bibliography

[1] Y. Kochura, Y. Gordienko, V. Taran, N. Gordienko, A. Rokovyi, O. Alienin, and S. Stirenko, ArXiv:1812.11731 [Cs, Stat] **938**, 658 (2020).

[2] V. Volkov, 128 (n.d.).

[3] X. Lin, Y. Rivenson, N.T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, Science **361**, 1004 (2018).

[4] M. Miscuglio, A. Mehrabian, Z. Hu, S.I. Azzam, J. George, A.V. Kildishev, M. Pelton, and V.J. Sorger, Opt. Mater. Express, OME **8**, 3851 (2018).

[5] A. Mehrabian, M. Miscuglio, Y. Alkabani, V.J. Sorger, and T. El-Ghazawi, ArXiv:1906.10487 [Cs, Eess] (2019).

[6] Y. Shen, N.C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, Nature Photon **11**, 441 (2017).

[7] A.N. Tait, T.F. de Lima, E. Zhou, A.X. Wu, M.A. Nahmias, B.J. Shastri, and P.R. Prucnal, Scientific Reports **7**, 1 (2017).

[8] R. Amin, J.K. George, S. Sun, T. Ferreira de Lima, A.N. Tait, J.B. Khurgin, M. Miscuglio, B.J. Shastri, P.R. Prucnal, T. El-Ghazawi, and V.J. Sorger, APL Materials **7**, 081112 (2019).

[9] T.W. Hughes, I.A.D. Williamson, M. Minkov, and S. Fan, Science Advances **5**, eaay6946 (2019).

[10] J. Meng, M. Miscuglio, J.K. George, A. Babakhani, and V.J. Sorger, ArXiv:1911.02511 [Physics] (2019).

[11] J. Li, S. Ranka, and S. Sahni, in *2011 IEEE 17th International Conference on Parallel and Distributed Systems* (IEEE, Tainan, Taiwan, 2011), pp. 157–164.

[12] A. Lavin and S. Gray, ArXiv:1509.09308 [Cs] (2015).

[13] N.P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P. Cantin, C. Chao, C. Clark, J. Coriell, M. Daley, M. Dau, J. Dean, B. Gelb, T.V. Ghaemmaghami, R. Gottipati, W. Gulland, R. Hagmann, C.R. Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, A. Jaworski, A. Kaplan, H. Khaitan, A. Koch, N. Kumar,

S. Lacy, J. Laudon, J. Law, D. Le, C. Leary, Z. Liu, K. Lucke, A. Lundin, G. MacKean, A. Maggiore, M. Mahony, K. Miller, R. Nagarajan, R. Narayanaswami, R. Ni, K. Nix, T. Norrie, M. Omernick, N. Penukonda, A. Phelps, J. Ross, M. Ross, A. Salek, E. Samadiani, C. Severn, G. Sizikov, M. Snelham, J. Souter, D. Steinberg, A. Swing, M. Tan, G. Thorson, B. Tian, H. Toma, E. Tuttle, V. Vasudevan, R. Walter, W. Wang, E. Wilcox, and D.H. Yoon, ArXiv:1704.04760 [Cs] (2017).
[14] F. Silvestri and F. Vella, ArXiv:1908.06649 [Cs] (2019).
[15] L. Alloatti, R. Palmer, S. Diebold, K.P. Pahl, B. Chen, R. Dinu, M. Fournier, J.-M. Fedeli, T. Zwick, W. Freude, C. Koos, and J. Leuthold, Light: Science & Applications **3**, e173 (2014).
[16] NVIDIA TESLA V100 GPU Architecture, https://www.nvidia.com/en-us/data-center/volta-gpu-architecture.
[17] Y. Salamin, P. Ma, B. Baeuerle, A. Emboras, Y. Fedoryshyn, W. Heni, B. Cheng, A. Josten, and J. Leuthold, ACS Photonics **5**, 3291 (2018).
[18] L. Vivien, J. Osmond, J.-M. Fédéli, D. Marris-Morini, P. Crozat, J.-F. Damlencourt, E. Cassan, Y. Lecunff, and S. Laval, Opt. Express, OE **17**, 6252 (2009).
[19] H.T. Chen, P. Verheyen, P. De Heyn, G. Lepage, J. De Coster, P. Absil, G. Roelkens, and J. Van Campenhout, J. Lightwave Technol. **33**, 820 (2015).
[20] Y. Alkabani, M. Miscuglio, V.J. Sorger, and T. El-Ghazawi, IEEE Photonics Journal 1 (2020).
[21] Y. Zhang, J.B. Chou, J. Li, H. Li, Q. Du, A. Yadav, S. Zhou, M.Y. Shalaginov, Z. Fang, H. Zhong, C. Roberts, P. Robinson, B. Bohlin, C. Ríos, H. Lin, M. Kang, T. Gu, J. Warner, V. Liberman, K. Richardson, and J. Hu, Nature Communications **10**, 1 (2019).
[22] C. Ríos, N. Youngblood, Z. Cheng, M.L. Gallo, W.H.P. Pernice, C.D. Wright, A. Sebastian, and H. Bhaskaran, Science Advances **5**, eaau5759 (2019).
[23] M. Miscuglio, J. Meng, O. Yesiliurt, Y. Zhang, L.J. Prokopeva, A. Mehrabian, J. Hu, A.V. Kildishev, and V.J. Sorger, ArXiv:1912.02221 [Physics] (2019).