A scalable integrated photonics tensor core for large vector-matrix multiplications

Emanuel Peinke[a]*, Thomas Lacasse[ab], Najla Najeeb[a], Md Mahadi Masnad[ac], Mohamed Fouda[a], Yves-Alain Peter[b], Odile Liboiron-Ladouceur[c]

[a]3E8 Inc., 780 av. Brewster, H4C 2K1 Montreal QC, Canada;
[b]Polytechnique Montréal, Microphotonics Laboratory, Engineering Physics, 2500 chemin de Polytechnique, H3T 1J4 Montreal QC, Canada;
[c]McGill University, The Photonic DataCom Team, Electrical & Computer Engineering, McConnell Engineering Building, 3480 University Street, H3A 0E9 Montreal QC, Canada

ABSTRACT

Optical computing promises to play a major role in hardware chips dedicated to artificial intelligence (AI). Digital electronics, when employed in computing hardware, face the sunset of Moore's law and the acknowledged end of Dennard Scaling (energy density of shrinking transistors). In response to these limitations, a paradigm shift towards non-digital processing is on the horizon. In optical computing devices for AI, the dominant mathematical operation is vector-matrix multiplication. It is typically limited to very small vector and matrix sizes. Most approaches don't allow for significant scaling. In this context, our work focuses on the development of a silicon photonics tensor core that exhibits a unique scalability feature, enabling effective expansion to accommodate large matrix sizes. This scalability is deemed essential for the realization of meaningful AI accelerator products leveraging photonic hardware.

Keywords: optics computing, silicon photonics, AI

1. THE PROMISE OF OPTICS COMPUTING

The concept of optics computing was introduced decades ago and is expected to revolutionize computing hardware. This dream of optics computing has its origin and justification in the incredible performances such a system might achieve, combining minimal latency, low loss propagation, and high operation frequencies. Photonics systems can operate above 100 GHz without suffering from the frequency scaling limitations prevalent in equivalent electronic systems. In electronics, frequency increase leads to a significant increase in energy consumption. This limits most commercial digital electronics to around 3 GHz so to comply with energy density budgets. Reducing the transistor size did not change this. When considering latency, matrix multiplication in the optical domain typically takes hundreds of picoseconds, while digital electronics latency is in the order of hundreds of nanoseconds. When further considering optical data input and a tensor core made of passive optical elements, optical matrix multiplication can happen at game-changing speed and with negligible (almost zero!) power consumption. For realistic optics computing applications, major challenges remain a full system integration including analog and digital (drive-) electronics, efficient signal conversion from digital memory and/or electronic signals to analog optics and back, and a scalable optical tensor core solution allowing for large and meaningful matrix sizes. This work focuses on the latter aspect of scalability.

2. A SCALABLE SOLUTION

An optical tensor core system, which is scalable and made of components not consuming electrical power, may achieve large vector-matrix multiplications at the highest speed with minimal power consumption, as required for meaningful optics computing applications.

We developed a tensor core architecture made of only passive integrated silicon photonics elements, namely silicon (Si) and silicon nitride (SiN) waveguides, Si-to-SiN couplers and multimode interferometers (MMIs). Photonic memory (PMEM) cells [1], electronically programmable and optically readable, store the matrix weights. The tensor core multiplies an input vector of length N with a stored matrix of size MxN, to obtain a result vector of length M. Every element of the input vector is encoded in the intensity of an optical signal, typically of the C- or O-band. Here, all of the *N* optical signals have a slightly different wavelength, so to be carried by one only waveguide, which is the physical input to the tensor core. On-chip, the *N* signals are separated by a DEMUX into *N* waveguides. (Alternatively, less DEMUX but more waveguides entering the chip could be used, reducing DEMUX but adding I/O constraints.) Using MMIs, each waveguide is split into *M* waveguides carrying identical intensity (typically $1/M^{\text{th}}$ of the respective input power), resulting in *M*x*N* waveguides. On each waveguide is a PMEM cell for the *M*x*N* dot-products. One 90-degree bend and a Si-to-SiN coupler per waveguide allow sorting the *M*x*N* waveguides in *M* bundles of *N* waveguides. Together, the *N* waveguides of each bundle carry the intensity of one element of the result vector. On every waveguide, after a Si-to-SiN coupler, a high-speed photodetector converts the optical signal into an electrical current. The currents of each bundle's photodectors are summed in the analog electrical domain, resulting in the *M* currents of the result vector. Summation in the electronic domain allows for linear addition of current intensities, instead of more complex optical interference of electromagnetic fields typically requiring active control for correct summation. The transition to the SiN layer allows for very low loss crossings of waveguides when building the matrix, as only waveguides of different layers (Si and SiN) cross. The architecture is particularly scalable as the number of building blocks (except the MMIs) scales linearly with the matrix size. By design, insertion losses are identical for every single dot-product and total losses only differ in negligible different propagation losses for slightly different lengths of straight waveguide sections. No photonic element needs active control, simplifying strongly drive electronics and reducing power consumption to its minimum. Only reprogramming of the PMEM requires electronic control and energy [1], not the matrix multiplication. Computing latency is given by the limited physical size of the tensor core, as the propagation of the signals happens at the speed of light. Together, these characteristics result in strongly improved scalability and simplicity of the presented architecture, useful for successful application. Compared to existing silicon photonics tensor core architectures that use micro-rings, Mach-Zehnder interferometers, or other structures [2-5], the presented architecture differentiates by its scalability, homogeneous losses and extensive use of fully passive elements, simplifying integration of larger-matrix tensor cores.

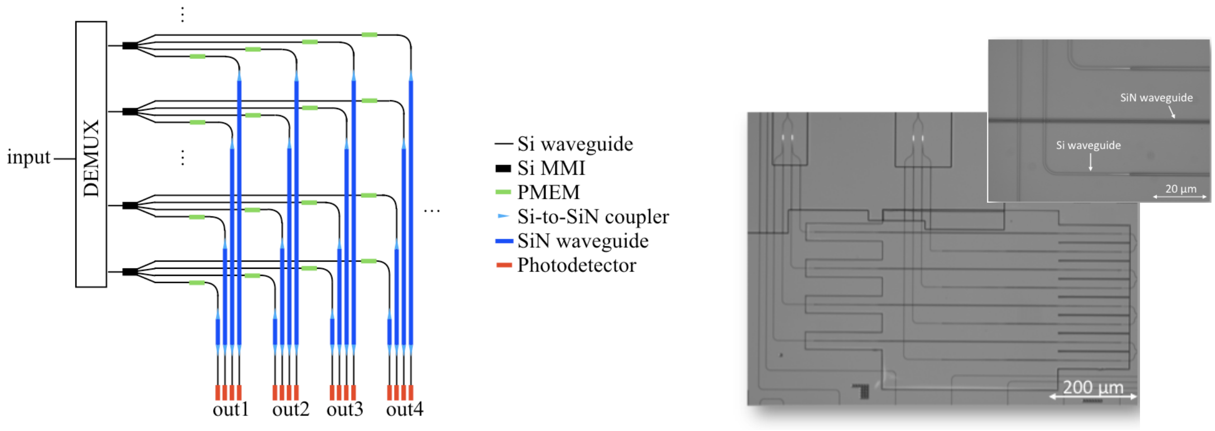

Figure 1. Sketch of a 4x4 tensor core that can be scaled to larger matrix size by adding more columns or rows, as indicated by the dots in the drawing. The photo and inset show a part of the test chip.

References

[1] J. Meng *et al.*, Electrical programmable multilevel nonvolatile photonic random-access memory, *Light Sci Appl* 12, 189 (2023).

[2] M. Miscuglio *et al.*, Photonic tensor cores for machine learning, *Applied Physics Reviews*, *7*(3), 031404 (2020).

[3] Y. Shen *et al.*, Deep learning with coherent nanophotonic circuits, *Nature Photon.*, vol. 11, pp. 441–446 (2017).

[4] F. Shokraneh *et al.*, Theoretical and Experimental Analysis of a 4×4 Reconfigurable MZI-Based Linear Optical Processor, *Journal of Lightwave Technology,* vol. 38, pp. 1258-1267 (2020).

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

We would like to acknowledge the Mitacs Accelerate Program for funding and CMC Microsystems for the provision of products and services that facilitated this research.

*emanuel.peinke@icloud.com