Optical Neural Networks with Tensor Compression and Photonic Memory

Xian Xiao,* Stanley Cheung, Bassem Tossoun, Thomas Van Vaerenbergh, Geza Kurczveil, and Raymond G. Beausoleil

Hewlett Packard Labs, Hewlett Packard Enterprise, 820 N. McCarthy Blvd., Milpitas, California 95305, USA

**xian.xiao@hpe.com*

Abstract: This paper introduces our recent efforts on scalable, energy-efficient, and low-latency tensorized optical neural networks, including design considerations, options for wavelength-parallel photonic tensor cores, and photonic memory for non-volatile tuning.

1. Introduction

Optics plays a crucial role in interconnects in modern data centers by providing high-bandwidth, low-power, low-latency, and reconfigurable data transmission. However, not all of these advantages naturally carry over to optical computing, especially when compared with the state-of-the-art electronic processors [1]. One major issue for optical computing is that there is no practical optical memory. That means E/O and O/E conversions and DAC/ADCs are involved during memory access. As a result, even though optical computation can happen with extremely low latency and energy consumption, the interfaces between optics and electronics can cause significant bottlenecks. In a system-level study of a photonic artificial intelligence (AI) accelerator, only ~10% of the overall power is consumed in the optical devices [2]. The latency and energy consumption bottleneck caused by memory access and transduction is even more severe when running large-scale models. Because of the constraint of the wavelength of light, the footprint of optical multiply-accumulate (MAC) units (e.g., Mach-Zehnder interferometers (MZIs) or microring resonators (MRRs)) are orders of magnitude larger than electronic transistors. Additionally, due to the $O(N^2)$ scaling rule, large-scale optical weight matrices (e.g., 1024×1024 and beyond) don't fit in a single die and exhibit insurmountable insertion loss. As a result, large-scale optical matrices are computed by tiles or blocks with time multiplexing, demanding intensive memory access to store the intermediate data.

Here, we address the issue mentioned above from two aspects. First, we introduce the tensorized optical neural network (TONN) architecture by leveraging the tensor-train (TT) decomposition algorithm (as shown in Fig. 1) to compress the less important parameters in deep neural networks (DNNs) [3]. Such architecture significantly reduces the footprint, control complexity, and insertion loss of large-scale ONNs. It enables large-scale matrix multiplication in a single clock cycle and eliminates undesired memory access. Second, we developed the photonic memory for non-volatile tuning of the phase shifters based on charge-trap flash memory [4] and memristor [5] mechanisms, demonstrating the in-memory computing concepts.

2. Tensorized Optical Neural Networks

2.1. Architecture and Design Considerations

The idea of TONN originates from the model compression with pruning technique. The pruning technique, leveraging weight sparsity, has been widely applied in efficient AI computing products [6]. Power-gating the less

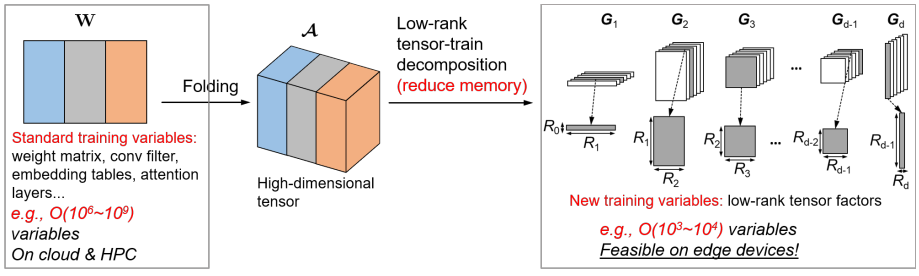

Fig. 1. Tensor-train decomposition.

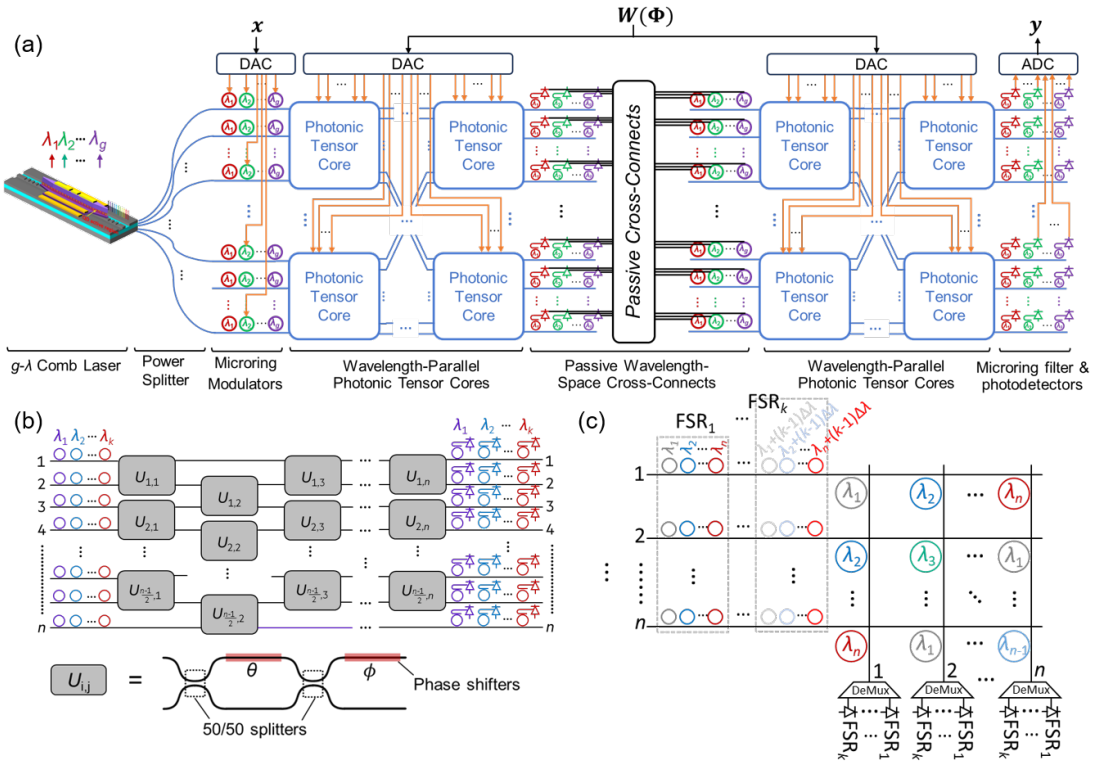

Fig. 2. (a) Tensorized optical neural network architecture with parallelisms in space and wavelength domains. (b) Wavelength-parallel photonic tensor core based on wideband MZI mesh. (c) Wavelength-parallel photonic tensor core based on multi-FSR MRR crossbar array.

important connections in, e.g., MZI meshes, only reduces the power consumption (assuming volatile tuning) but not the footprint and control complexity [7]. TT decomposition can be seen as a regulated pruning method or, specifically, a singular value decomposition for multi-dimensional matrices (tensors).

Since tensor operations don't naturally exist in 2-dimensional photonic chips, we have invented the TONN architecture [3], a photonic implementation of TT-decomposed large-scale weight matrices (Fig. 2(a)). In such architecture, the tensor products are emulated by representing the tensor indices in the wavelength and space domain and multiplying them with an array of wavelength-parallel photonic tensor cores. Several tuning knobs, including the factorization of the scale, TT-ranks, and the number of wavelengths, control the chip layout of TONN. With the folded layout scheme, 2048×2048 and 4096×4096 TONNs can fit in a single DUV stepper die.

2.2. *Wavelength-Parallel Photonic Tensor Cores*

The TONN architecture requires the photonic tensor cores to provide identical weight values among different wavelength channels. The key point to making MZI meshes wavelength-parallel is that the building blocks, 2×2 MZIs, need to be balanced in two arms, as shown in Fig. 2(b). This way, the bandwidth of 2×2 MZIs can be tens of nanometers, and the MZI meshes can be wideband. The second option for wavelength-parallel photonic tensor core is to exploit the multiple free spectral ranges (multi-FSRs) of the MRR crossbar array [8], as shown in Fig. 2(c). Due to the periodicity of the resonances, the lineshapes of the MRR spectra at resonances in different FSRs are similar, providing nearly the same weight values. The multi-FSR MRR crossbar option has a smaller footprint and easier programming but requires more wavelengths than the wideband MZI mesh option.

3. Photonic Memory for Non-Volatile Tuning

We have developed two photonic memory mechanisms on HPE's densely integrated III-V-on-Si metal–oxide–semiconductor capacitor (MOSCAP) platform (Fig. 3(a-c)). The first mechanism, charge-trap flash memory (CTM) [4], uses a layer of insulating material to trap and store electrons and holes and alter the effective index of the optical mode by plasma dispersion effect. Fig. 3(d) shows the energy-band diagram during the writing process. Fig. 3(e) and (f) shows the measured optical spectra of CTM MZI and two cascaded double ring structures, demonstrating tuning between the initial, volatile, non-volatile, reset, and final states. The second mechanism heterogeneously integrates a memristor with III-V-on-Si phase shifters [5]. The device switches its

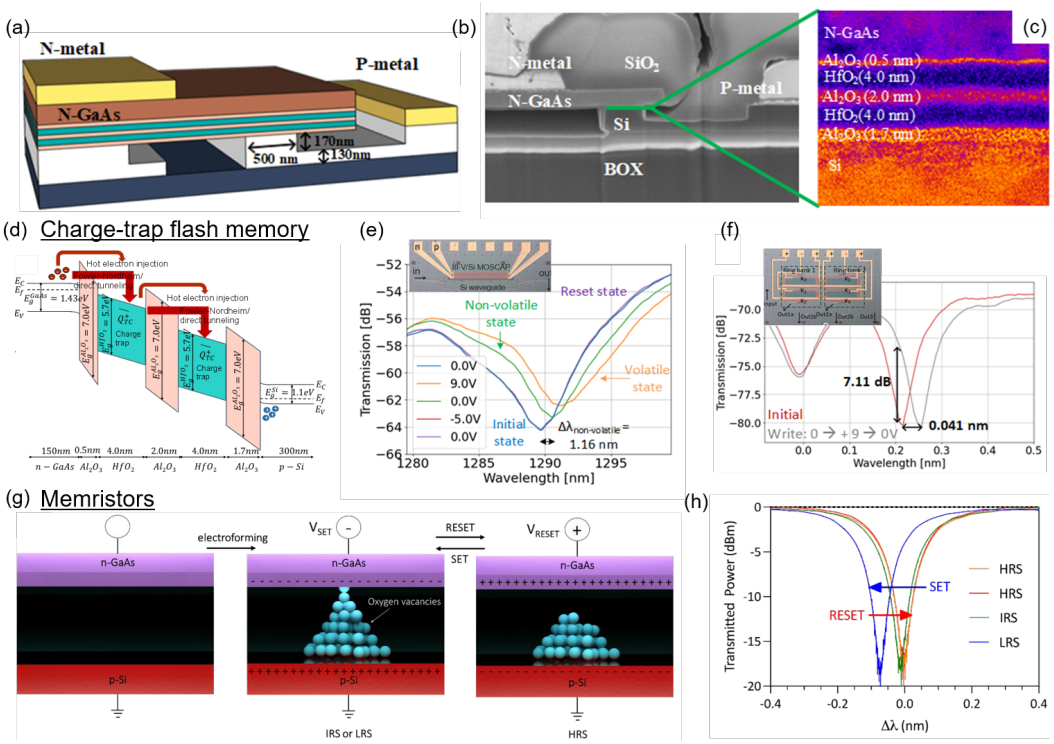

Fig. 3. (a) 3-D schematic of the III-V-on-Si MOSCAP phase shifter. (b) SEM cross section. (c) HRTEM image of the layer stack. (d) Schematic of energy-band diagram for CTM with positive bias. (e) Spectra for different states of CTM MZI. (f) Spectra for two cascaded double ring structures. (g) Schematic diagram of the forming and rupturing conductive filaments within the memristor. (h) Measured spectrum of the memresonator while a 2 V read voltage is applied in different states.

resistance by applying set/reset switching voltages by creating conductive filaments within the oxide material, as shown in Fig. 3(g). The change of resistance leads to an increase in the current and, subsequently, the carrier density within the optical waveguide, causing an enhanced plasma dispersion effect, tuning the resonant wavelength of the memristive MRR, as shown in Fig. 3(h). These photonic memory enables nearly-zero power consumption for DNN inference and can seamlessly integrate with other essential components for photonic AI accelerators, such as quantum dot comb lasers, III-V/Si MOSCAP ring modulators, Si-Ge avalanche photodetectors, and in-situ III-V/Si light monitors. We believe TONN with photonic memory will be a significant step towards breaking the latency and energy consumption bottleneck caused by intensive memory access for the photonic AI accelerator.

References

1. Peter L McMahon. The physics of optical computing. *Nature Reviews Physics*, 2023.
2. Cansu Demirkiran, Furkan Eris, Gongyu Wang, Jonathan Elmhurst, Nick Moore, Nicholas C Harris, Ayon Basumallik, Vijay Janapa Reddi, Ajay Joshi, and Darius Bunandar. An Electro-Photonic System for Accelerating Deep Neural Networks. *J. Emerg. Technol. Comput. Syst.*, 19(4), sep 2023.
3. Xian Xiao, Mehmet Berkay On, Thomas Van Vaerenbergh, Di Liang, Raymond G Beausoleil, and SJ Ben Yoo. Large-scale and energy-efficient tensorized optical neural networks on iii–v-on-silicon moscap platform. *APL Photonics*, 6(12):126107, 2021.
4. Stanley Cheung, Di Liang, Yuan Yuan, Yiwei Peng, Yingtao Hu, Geza Kurczveil, and Raymond G Beausoleil. Non-volatile heterogeneous iii-v/si photonics via optical charge-trap memory. *arXiv preprint arXiv:2305.17578*, 2023.
5. Bassem Tossoun, Di Liang, Stanley Cheung, Zhuoran Fang, Xia Sheng, John Paul Strachan, and Raymond G Beausoleil. High-speed and energy-efficient non-volatile silicon photonic memory based on heterogeneously integrated memresonator. *arXiv preprint arXiv:2303.05644*, 2023.
6. J Choquette. NVIDIA Hopper H100 GPU: Scaling Performance. *IEEE Micro*, 43(3):9–17, 2023.
7. Sanmitra Banerjee, Mahdi Nikdast, Sudeep Pasricha, and Krishnendu Chakrabarty. Pruning coherent integrated photonic neural networks. *IEEE Journal of Selected Topics in Quantum Electronics*, 29(2: Optical Computing):1–13, 2023.
8. Xian Xiao, Stanley Cheung, Sean Hooten, Yiwei Peng, Bassem Tossoun, Thomas Van Vaerenbergh, Geza Kurczveil, and Raymond G Beausoleil. Wavelength-Parallel Photonic Tensor Core Based on Multi-FSR Microring Resonator Crossbar Array. In *Optical Fiber Communication Conference*, page W3G.4, San Diego, CA, 2023.