Integrated Photonics for Computing and Artificial Intelligence

2023 IEEE Photonics Society Summer Topicals Meeting Series (SUM) | 979-8-3503-4720-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/SUM57928.2023.10224461

Chenghao Feng
Microelectronics Research Center
The University of Texas at Austin
Austin, USA

Shupeng Ning
Microelectronics Research Center
The University of Texas at Austin
Austin, USA

Jiaqi Gu
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, USA

Hanqing Zhu
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, USA

David Z. Pan
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, USA

Ray T. Chen
Microelectronics Research Center
The University of Texas at Austin
Austin, USA
chenrt@austin.utexas.edu

***Abstract*— In this paper, we review the progress of integrated photonics in both digital computing and analog neuromorphic computing. We introduce methods to design scalable, area-efficient, and energy-efficient integrated photonic computing chips for computing and artificial intelligence acceleration with experimental demonstrations.**

Keywords—integrated photonics, digital computing, analog neuromorphic computing

I. INTRODUCTION

Integrated photonics is a promising technology for next-generation computing because of the essential characteristics of light, including low latency, high bandwidth, and low power consumption[1,2]. A variety of compact, energy-efficient, and high-speed photonic devices have been demonstrated and can be found in component libraries at foundries. Current foundries also enable the co-integration of silicon-based electrical and photonic circuits on the same interposer or the same substrate to implement complicated computing tasks. In the past decades, numerous integrated photonic chips have been investigated for both digital and analog computing [1-7].

In this presentation, we will guide the interested audience on a journey toward next-generation optical processing platforms with a comprehensive introduction to optical devices, photonic integrated circuits, hardware-software co-design, hardware realization, and experimental demos of optical computing. We will also review efforts to improve the scalability, area efficiency, and energy efficiency of integrated photonic computing chips for next-generation AI accelerators.

II. OPTICAL DIGITAL COMPUTING

In optical digital computing, people use EO modulators and other components as basic logic gates to implement combinational logic functions such as AND, OR, and XOR operations. Both the input and output data are binary in optical digital computing systems. Similar to electrical digital logic circuits, the precision of optical digital computing circuits is determined by the bit number and is independent of the scale of the circuits.

Direct-logic-based electronic-photonic computing architecture, which utilizes the advantages of electronics and photonics, is widely explored in optical digital computing. A typical example is the proposed electronic-photonic arithmetic logic unit (EPALU), which includes the experimental demonstration of an optical full adder at 20 Gb/s [1]. Other logic circuits in EPALU, such as digital comparators [3] and decoders [4], are also designed with high-speed (20 Gb/s) experimental demonstration. These integrated photonic digital computing circuits are scalable and capable of processing larger bit-width inputs, such as 64 or 128-bit data. Additionally, the building blocks of the EPALU incorporate wavelength-division multiplexing (WDM) to improve the area efficiency of optical digital computing circuits. Performance analysis shows the EPALU can be operated at over 20 Gb/s with one to two orders of magnitudes better energy efficiency than transistor-based electrical counterparts.

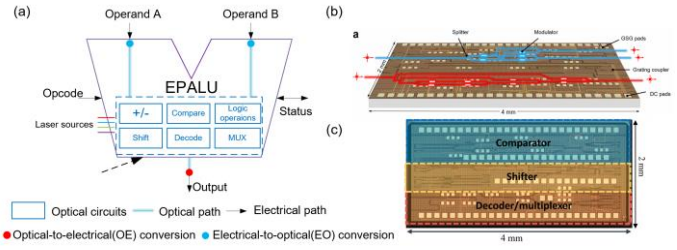

Fig. 1 Schematic of electronic-photonic ALU and its building blocks such as (b) optical full adder [3], (c) comparator [4] and decoder [5].

III. OPTICAL ANALOG NEUROMORPHIC COMPUTING

In analog optical AI accelerators, people manipulate the transfer matrix of photonic integrated circuits (PICs) to implement matrix-vector multiplications (MVMs), which are fundamental operations in artificial intelligence and signal processing. Compared to optical digital computing chips, optical analog computing chips have less precision but can achieve better parallelism and area efficiency than digital ones. As a result, analog optical computing chips are designed for applications that do not require high precision, e.g., neuromorphic computing tasks. Based on the mechanism of integrated photonic tensor cores (PTCs) are generally categorized into coherent and incoherent ONNs. Coherent PTCs, e.g., MZI-based PTCs (Fig. 1(a)), utilize singular-vector-decomposition (SVD) to implement matrix multiplications [1]. Incoherent ONNs such as microring-based PTCs first use EO modulators to implement dot products in parallel and then combine the signals on one optical path and accumulate the dot product results after photodetection [5]. After the weight parameters are mapped on the integrated photonic tensor cores (PTCs) by programming active components such as MZI or microring modulators, an integrated PTC can implement parallel matrix multiplications at the speed of light with near-zero energy consumption. Therefore, PTCs can achieve orders of magnitudes better than their electrical counterparts in both latency and energy efficiency [6].

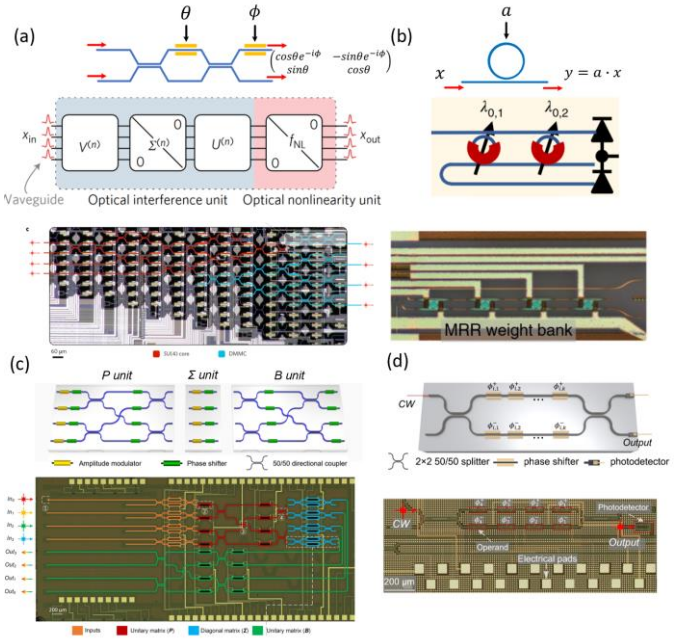

Fig. 2 Integrated photonic chips for analog AI acceleration. (a) and (b) show an MZI-based [1] and a microring-based [5] photonic tensor cores (PTCs) designed for general matrix multiplications (GEMMs). PTCs shown in (c) and (d) are designed to improve optical AI accelerators' area and energy efficiency. (c) uses a compact butterfly-style photonic mesh to reduce the number of optical components from the circuit level [7]. (d) uses a customized MZI-based multi-operand optical neuron to improve the efficiency of implementing tensor operations from the device level [8].

To maximize the performance benefit of photonic computing in AI acceleration, scalable and efficient photonic tensor core designs are in high demand to implement large-size tensor operations (e.g., 128×128). The majority of the analog photonic AI chips are designed to implement general matrix multiplications, leading to unnecessarily large area costs and high control complexity. For instance, MZI-based PTC requires $O(m^2 + n^2)$ MZIs and ~$(m + n)$ cascaded MZIs in one optical path to implementing a n-input, m-output layer, consuming huge area cost and unacceptable high propagation loss to implement large tensor operations (e.g., 128×128). Both circuit- and device-level optimizations have been explored to enhance the scalability of ONNs. Circuit-level approaches, such as the butterfly-style circuit mesh, have been explored to reduce hardware usage [7]. Moreover, compact customized device-level photonic tensor cores, e.g., multi-operand optical neurons, have been proposed to significantly reduce the device footprint and improve the hardware efficiency of tensor operations [8]. Using hardware-software co-design and hardware-aware training approaches, these optimized PTCs can achieve one to orders of magnitudes smaller footprint and lower propagation loss compared to PTCs designed GEMMs with similar task performance. More details about the designs and experimental demonstrations of these scalable hardware-efficient PTCs will be provided in the presentation.

IV. CONCLUSION

In conclusion, we have proposed the progress of integrated photonics in digital computing and analog neuromorphic computing, showing their potential in next-generation computing. We also introduce efforts to improve the scalability, area efficiency, and energy efficiency of integrated photonic computing chips to push the limits of the practical deployment of photonic AI accelerators.

ACKNOWLEDGMENT

The authors acknowledge support from the Multidisciplinary University Research Initiative(MURI) program (Grant No.FA 9550-17-1-0071), monitored by Dr. Gernot S. Pomrenke.

REFERENCES

[1] Z. Ying, et al. "Electronic-photonic arithmetic logic unit for high-speed computing." Nature communications 11.1 (2020): 2154.

[2] Y. Shen, et al. "Deep learning with coherent nanophotonic circuits." Nature photonics 11.7 (2017): 441-446.

[3] C. Feng, et al. "Toward High‐Speed and Energy‐Efficient Computing: A WDM‐Based Scalable On‐Chip Silicon Integrated Optical Comparator." Laser & Photonics Reviews 15.8 (2021): 2000275.

[4] C. Feng, et al. "Wavelength-division-multiplexing (WDM)-based integrated electronic–photonic switching network (EPSN) for high-speed data processing and transportation." Nanophotonics 9.15 (2020): 4579-4588.

[5] C. Huang, et al. "A silicon photonic–electronic neural network for fibre nonlinearity compensation." Nature Electronics 4.11 (2021): 837-844.

[6] H. Zhou, et al. "Photonic matrix multiplication lights up photonic accelerator and beyond." Light: Science & Applications 11.1 (2022): 30.

[7] C. Feng, et al. "A Compact Butterfly-Style Silicon Photonic–Electronic Neural Chip for Hardware-Efficient Deep Learning." ACS Photonics 9.12 (2022): 3906-3916.

[8] C. Feng, et al. "Optically-interconnected, hardware-efficient, electronic-photonic neural network using compact multi-operand photonic devices." Optical Interconnects XXIII. SPIE, 2023.