Photonic Analog-to-Digital Architecture for Accelerating Multiply-Accumulate Operations

Nathaniel Nauman,[1,*] James Robinson,[1] Yuyang Wang,[1] Kaylx Jang,[1] Xiang Meng,[1] and Keren Bergman[1]

[1] *Department of Electrical Engineering, Columbia University, New York, NY, 10027, USA*

[*] *n.nauman@columbia.edu*

Abstract: We demonstrate a 3-bit photonic analog-to-digital architecture to accelerate multiply-accumulate operations, achieving a ±300mV buffer for 50mV steps within a 350mV range. This architecture enhances energy efficiency for near-memory computation while maintaining full digital precision.

1. Introduction

Memory transfer is becoming the dominant bottleneck in serving large AI models, where the speed of computation often outpaces the ability to fetch data from memory [1]. This disparity, known as the "memory wall," has grown increasingly severe. Peak hardware FLOPS have improved by 60,000x over the last two decades while DRAM and interconnect bandwidth have only scaled by factors of 100x and 30x, respectively [1]. In matrix multiplication-heavy applications, such as AI workloads, this bottleneck is particularly pronounced during multiply-accumulate (MAC) operations, where off-chip memory transfer incurs significant energy and latency overhead [2].

Recent efforts to alleviate this bottleneck have explored in-memory computing (IMC) approaches, such as SRAM-based IMC, which integrate analog computation near memory cells [2]. However, these solutions often compromise precision, particularly when relying on flash analog-to-digital converters (ADCs) [2]. Similarly, prior photonic MAC accelerators sacrifice precision for power savings [3]. Architectures like microring resonator (MRR)-based weight banks encode entire numbers through analog light intensities which results in low-precision multipliers [4], and Mach-Zehnder interferometer (MZI) meshes are susceptible to cascaded, asymmetric losses which are exacerbated as the matrix becomes larger [5].

Prior photonic ADC architectures have used the input signal to modulate a wavelength multiplexed optical pulse train and then measure the amplitude of the demultiplexed pulses with photodetectors (PDs) [6]. However, this requires ultra-stable timing intervals between the pulses and multiple demultiplexing steps [6].

In this work, we present and demonstrate a novel photonic ADC architecture tailored for MAC acceleration which maintains full digital precision throughout computation. By processing bit place columns with high-precision ADC steps of 50mV within a 350mV range, and generating a buffer of 600mV around the comparator threshold voltage, our system preserves accuracy while minimizing energy costs. The proposed architecture encodes partial products as analog light intensities, which are then converted into digital outputs by the photonic ADC. This approach avoids the accuracy-efficiency trade-offs that limited prior implementations, making it well-suited for near-memory MAC operations in AI and high-performance computing systems.

2. Algorithm and Architecture

In Fig. 1, we illustrate the overall algorithm and photonic integrated circuit (PIC). The multipliers and multiplicands, labeled as 4-bit registers $a-d$, are stored in memory, labeled DRAM (Fig. 1(a)). Although this block could be any form of memory, one of the purposes of IMC is to minimize the shuttling of data from off-chip DRAM to the processor's on-chip SRAM. Therefore, our proposed architecture keeps data near DRAM instead of the CPU/GPU. Matrix multiplication is a series of simultaneous dot product operations between the rows and columns of two matrices, and a dot product between vectors of length N entails the calculation of N products, all of which are added together. Instead of using half and full adders to compute the product of two inputs at a time, we propose filling register arrays with the partial products of many inputs and then performing addition optically.

First, the multipliers and multiplicands are accessed from memory and their partial products fill a set of shift registers (Fig. 1(a)). Each register array corresponds to a particular bit place of the final MAC value. The bits in these arrays drive the photonic digital-to-analog converter (DAC) by shifting modulators on or off-resonance, causing the light intensity at the output of the bus to change by a discrete amount proportional to the modulator's extinction ratio. The output light intensity is sent to a PD and a transimpedance amplifier (TIA) to drive the photonic ADC (Fig. 1(a)).

The ADC converts the analog voltage, which represents the sum of logic high bits in one bit place, into a binary count. The photonic ADC is composed of add-drop microresonator modulators with PDs and TIAs at their drop ports (Fig. 1(b)). Each microresonator is incrementally shifted in the frequency domain further away from its target wavelength channel. An analog voltage is applied to all microresonators at the same time and shifts a subset of PDs such that their current exceeds the threshold for logic high (Fig. 1(b)). Since the extent of detuning increases monotonically for each subsequent microresonator along one direction of the shared bus, the resulting drop port digital values are thermometer encoded, like an electronic flash ADC. Therefore, the comparators are followed by a thermometer-to-binary converter.

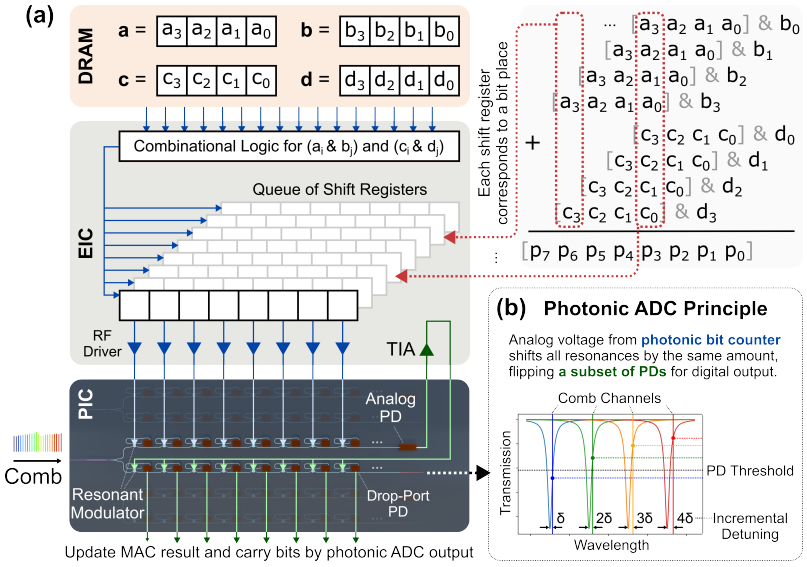

Fig. 1. (a) By accessing partial products from memory cells directly, the photonic DAC and ADC count the number of logic high bits in each bit place of the MAC. (b) The photonic ADC operates by modulating a ladder of incrementally detuned microresonators.

The least significant bit of the ADC's binary count is retained in the same MAC bit place's register array, while the rest of the binary count is passed along the pipeline to consecutive register arrays as carry bits. Unlike prior IMC approaches, this algorithm ensures the full digital precision of the final MAC value without requiring a redesign of the memory cell hardware. Leveraging the integration of complex multilayered systems [7], the photonic circuit could be flip-chipped and positioned beneath off-chip memory to reduce data transfer overhead (Fig. 2(a)).

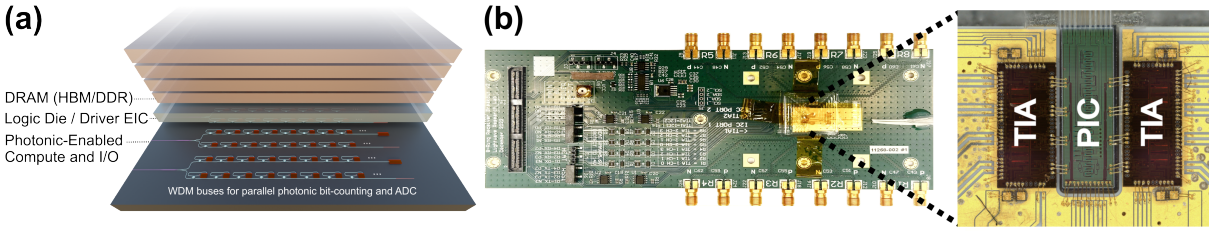

Fig. 2. (a) Proposed heterogeneous integration of the photonic MAC accelerator, driver and decoder EIC, and memory. (b) The experimental setup includes a photonic circuit composed of 8 MRRs with drop port PDs and TIAs wire bonded to an interposer.

3. Results

The same PIC can perform both the DAC and ADC operations, so a photonic circuit comprised of a single bus and eight add-drop wide-FSR microresonators with drop port PDs and TIA channels was wirebonded to an interposer and PCB board (Fig. 2(b)). The microresonators were thermally tuned and the TIA chips were programmed by

an FPGA. The DAC's output light intensity was converted to a voltage through an external Thorlabs PD and subsequently measured by the Agilent 34401A 6½ Digit DMM (Fig. 3(a)). Seven microresonators are sufficient to encode eight possible states of the output's analog light intensity.

The ADC uses seven steps of 50mV in the modulation signal for a total range of 350mV and achieves a buffer of at least 600mV around the threshold voltage V_T of 2.3V. Therefore, each channel generates a voltage for logic high (low) $\varepsilon \geq 300$mV above (below) this threshold. The seven microresonators required to achieve 3-bit resolution are labeled $R1 - R7$. The bins in Fig. 3(b) represent the regions of the modulation signal that correspond to all possible 3-bit binary outputs. After the drop port voltage exceeds the threshold voltage, its value is not relevant to the decoder so the corresponding data in Fig. 3(b) is a dashed line. Examining one pair of adjacent microresonators is sufficient to determine the resulting binary output. For example, a binary output of "010" only occurs when microresonator $R2$ produces logic high and $R3$ concurrently produces logic low.

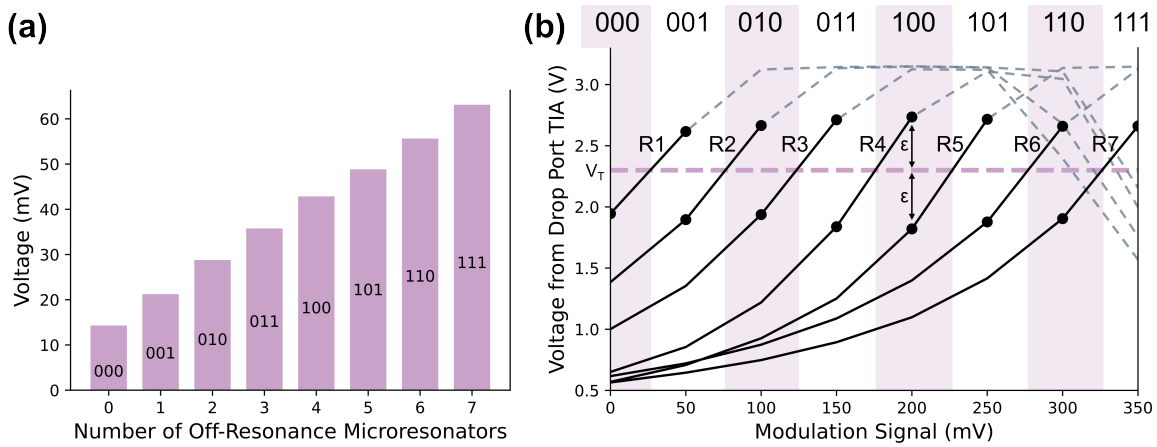

Fig. 3. Experimental results for the (a) photonic DAC and (b) photonic ADC. An external comparator converts the ADC's drop port voltages into logic high or low. There is a buffer $\varepsilon \geq 300$mV above and below the comparator's threshold voltage V_T.

4. Conclusion

In this work, we present and demonstrate a novel photonic ADC architecture for MAC acceleration. By maintaining full digital precision through photonic DAC and ADC circuits, we eliminate the need for costly memory transfers and overcome the accuracy-efficiency trade-offs typically encountered in IMC systems. Our results demonstrate that the photonic ADC can achieve high precision with minimal energy consumption, making it a promising solution for future AI and high-performance computing applications.

References

1. A. Gholami *et al.*, "AI and Memory Wall," IEEE Micro **44**, 3, 33-39 (2024).
2. S. Yin *et al.*, "XNOR-SRAM: in-memory computing SRAM macro for binary/ternary deep neural networks," IEEE **55**, 6, 1733-1743 (2020).
3. H. Zhou *et al.*, "Photonic matrix multiplication lights up photonic accelerator and beyond," Light Sci Appl **11**, 30 (2022).
4. A. Mehrabian, *et al.*, "PCNNA: A Photonic Convolutional Neural Network Accelerator," 2018 31st IEEE International System-on-Chip Conference (SOCC), Arlington, VA, USA, 2018, pp. 169-173.
5. H. Hou, *et al.*, "Hardware Error Correction for MZI-Based Matrix Computation." Micromachines **14**, 955 (2023).
6. K. Qubaisi *et al.*, "Photonic analog-to-digital converters," 2014 XXXIth URSI General Assembly and Scientific Symposium (URSI GASS), Beijing, China, 2014, pp. 1-3.
7. S. Daudlin *et al.*, "3D photonics for ultra-low energy, high bandwidth-density chip data links," arXivpreprint arXiv:2310.01615 (2023).

Acknowledgements: This material is based upon work supported in part by the National Security Agency (NSA) Laboratory for Physical Sciences (LPS) Research Initiative and in part by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-2036197.