# Parallel Dot Products Using Silicon Photonics

Andy Wolff, Kyle Shiflett, and Avinash Karanth
*School of Electrical Engineering and Computer Science, Ohio University*
Email: {aw415517, ks117713, karanth}@ohio.edu

***Abstract*—This paper proposes a parallel photonic architecture for computing dense dot products, such as those found during deep neural network (DNN) inference, and quantifies the architecture's computation error induced by crosstalk and noise.**

***Keywords*—optical computing; silicon photonics; analog arithmetic; microring resonator; deep neural networks**

## I. Introduction

The recent proliferation of machine learning (ML) applications, specifically deep neural networks (DNN), is attributed to their high accuracies on classification and regression tasks. As DNN architectures scale in both size and complexity, general purpose processors have failed to exploit sufficient parallelism for energy-efficient and low-latency inference. This has caused a shift towards heterogeneous computation architectures, where domain-specific hardware accelerators are tasked with carrying out computationally demanding operations. Silicon photonics has been proposed as an alternative technology for scaling DNN inference, and new computing architectures have been developed using microring resonators (MRR) [2] and programmable Mach-Zehnder interferometer (MZI) meshes [5]. The inherent parallelism of optics can be utilized through wavelength-division multiplexing (WDM), a technique used in fiber and on-chip interconnects for increasing bandwidth density, which can be taken advantage of to increase computational density in photonic devices. Furthermore, optics are suitable for broadcast and multicast data distributions such as those found in DNNs, because optical signals can be passively split and distributed via waveguide Y-branches, couplers, and free propagation regions. In this paper, we present a photonic building block that leverages these properties for computing highly parallel dot-products, such as those found in convolutional neural networks (CNN).

## II. Photonic Dot Product Architecture

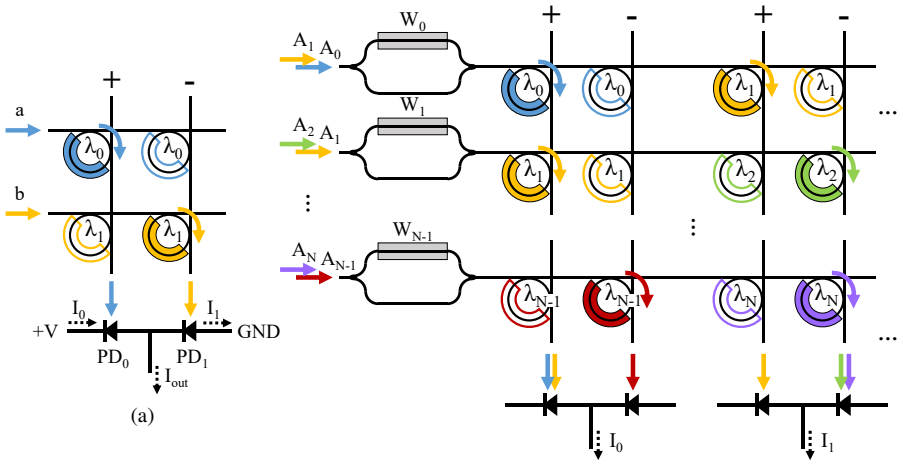
The basic building block for the proposed dot product architecture is the adder-subtractor crossbar (ASC), which performs analog arithmetic on input optical signals. The ASC is comprised of switching MRRs that drop a signal on an accumulation waveguide (WG), and there is a positive accumulation WG and a negative accumulation WG. Input operand values are carried by optical power amplitudes on separate wavelengths, so there is no signed representation of values. A values sign must be explicitly represented by switching into the appropriate accumulation WG. The summation of values is performed by a balanced photodiode (PD) pair, which subtracts the negative accumulation WG's induced current from the positive accumulation WG's current. The ASC is shown in Figure 1(a) for two inputs *a* and *b*.

Photonic dot products are implemented by including a MZI at the input ports of the ASC. The modulating MZI's output is $0 \leq P_{\text{out}} \leq P_{\text{in}}$, which gives the multiplication with some weight $0 \leq W \leq 1$. Assuming positive input signals $A$, which is the case for activations in a CNN layer produced by a nonlinear activation function like the rectified linear unit (ReLU), accumulation WG switching is dependent only on the multiplying weight's sign. The dot products in convolutional layers of CNNs often share weights, which can be leveraged using WDM and using multiple ASCs for a single multiplying MZI. This allows a single MZI to multiply several input signals at once, and the MZI must utilize Y-branches that have a broadband response to keep computation consistent across the different input wavelengths. The parallel dot product architecture is shown in Figure 1(b), which depicts two ASCs used for computing two results of the convolution operation. Note that $W_0$ is shared by inputs $A_0$ and $A_1$, and $W_1$ is shared by inputs $A_1$ and $A_2$, since the convolution window in the input vector $A$ that the weight vector $W$ is applied to is slid by one element, which corresponds to the following dot products being computed in parallel: $O_0 = A_0W_0 + A_1W_1 + ...A_{N-1}W_{N-1}$ and $O_1 = A_1W_0 + A_2W_1 + ...A_NW_{N-1}$. This parallel dot product architecture is the fundamental structure of the Albireo accelerator [7], which improved latency by 4.8 X, reduced energy consumption by 4.9 X, and reduced energy-delay product by 23.9 X when compared to DEAP-CNN [2] for CNN inference.

## III. Evaluation and Results

The proposed architectures were modeled and evaluated using Synopsys OptSim Circuit, and the photonic device parameters used are tabulated in Figure 1(c). Two parallel dot product architectures were evaluated, a two-input and a four-input variation, and input powers ranged from -3 mW to +3 mW at 1 mW intervals. Figure 2(a) shows the summation current output by the balanced PD pair for the two-input circuit and the absolute error for each result caused by noise, crosstalk, and losses. Each color represents a different summation bin, and the top plot contains a one-dimensional projection of the data to illustrate the separation between sum results. Figure 2(b) shows the results for a four-input circuit, where there is some overlap between

This research was partially supported by NSF grants CCF-1513606, CCF-1703013 and CCF-1901192.




| Device | Parameter | Value |
|---|---|---|
| WG | $n_{\text{eff}}, n_{\text{g}}$ | (2.33, 4.68) @ $\lambda$=1550 nm |
| | prop. loss | 1.5 dB/cm (straight) [3] |
| MRR | radius | 5 µm |
| | coupl. coeff. $k^2$ | 0.03 |
| | WG prop. loss | 3.8 dB/cm (bent) [3] |
| MZI | Y-branch loss | 0.36 dB [4] |
| | mod. loss | 1.2 dB [1] |
| PD | responsivity | 1.1 A/W [6] |
| | dark current | 25 pA @ 1 V [6] |

(c)

Fig. 1. **(a) Two-input adder-subtractor crossbar circuit, (b) Parallel dot product circuit for convolution, and (c) Device parameters.**

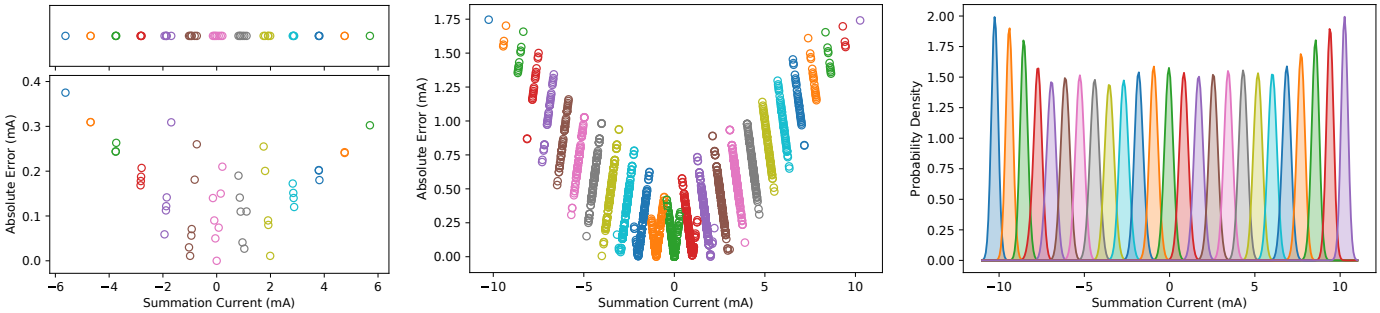



Fig. 2. **(a) Two-input dot product circuit error, (b) Four-input dot product circuit error, and (c) Four-input dot product probability densities. Each color represents a separate summation result.**

summation bins due to MRR crosstalk. The probability density for each summation bin is shown in Figure 2(c), and the overlap between bins indicates there is a small probability of error during operation.

## IV. Conclusions

This paper proposed a new scheme for computing parallel dot products using silicon photonics, motivated by the recent need for fast and efficient DNN inference. A two-input and a four-input photonic dot product architecture was evaluated, and the computation errors were quantified by taking crosstalk and noise into consideration.

## References


[1] S. Akiyama, T. Baba, M. Imai, T. Akagawa, M. Takahashi, N. Hirayama, H. Takahashi, Y. Noguchi, H. Okayama, T. Horikawa, and T. Usuki, "12.5-gb/s operation with 0.29-v·cm v$\pi$l using silicon mach-zehnder modulator based-on forward-biased pin diode," *Opt. Express*, vol. 20, no. 3, pp. 2911–2923, Jan 2012. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-20-3-2911

[2] V. Bangari, B. A. Marquez, H. Miller, A. N. Tait, M. A. Nahmias, T. F. de Lima, H. Peng, P. R. Prucnal, and B. J. Shastri, "Digital electronics and analog photonics for convolutional neural networks (deap-cnns)," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 26, no. 1, pp. 1–13, 2020.

[3] L. Chrostowski, Z. Lu, J. Flueckiger, X. Wang, J. Klein, A. Liu, J. Jhoja, and J. Pond, "Design and simulation of silicon photonic schematics and layouts," in *Silicon Photonics and Photonic Integrated Circuits V*, L. Vivien, L. Pavesi, and S. Pelli, Eds., vol. 9891, International Society for Optics and Photonics. SPIE, 2016, pp. 185 – 195. [Online]. Available: https://doi.org/10.1117/12.2230376

[4] Z. Lin and W. Shi, "Broadband, low-loss silicon photonic y-junction with an arbitrary power splitting ratio," *Opt. Express*, vol. 27, no. 10, pp. 14 338–14 343, May 2019. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-27-10-14338

[5] 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, "Deep learning with coherent nanophotonic circuits," *Nature Photonics*, vol. 11, no. 7, pp. 441–446, 2017.

[6] Z. Sheng, L. Liu, J. Brouckaert, S. He, and D. V. Thourhout, "Ingaas pin photodetectors integrated on silicon-on-insulator waveguides," *Opt. Express*, vol. 18, no. 2, pp. 1756–1761, Jan 2010. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-18-2-1756

[7] K. Shiflett, A. Karanth, R. Bunescu, and A. Louri, "Albireo: Energy-efficient acceleration of convolutional neural networks via silicon photonics," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, in press.