# Energy-Efficient Multiply-and-Accumulate using Silicon Photonics for Deep Neural Networks

Kyle Shiflett*, Avinash Karanth*, Ahmed Louri† and Razvan Bunescu*
*School of Electrical Engineering and Computer Science, Ohio University, Athens, Ohio 45701
†Dept. of Electrical and Computer Engineering, George Washington University, Washington, DC 20052
Email: *{ks117713, karanth, bunescu}@ohio.edu, †louri@gwu.edu

***Abstract*—We propose two optical hybrid matrix multipliers for deep neural networks. Our results indicate our all-optical design achieved the best performance in energy efficiency and latency, with an energy-delay product reduction of 33.1% and 76.4% for conservative and aggressive estimates, respectively.**

## I. Introduction

Escalating power densities due to increased transistor counts have created a performance barrier for modern multicore systems. To continue performance scaling, chip architects have shifted their focus towards application-specific accelerator designs that surpass the efficiency of general purpose processors. Deep neural network (DNN) architectures are of particular interest due to their unparalleled accuracy on contemporary applications like speech recognition and image classification.

Each hidden layer of a DNN involves concurrent matrix-vector multiplications (MVMs), consisting of several intermediate inner-product steps realized through the use of multiply-accumulate (MAC) units. Electronic-based accelerators implement large broadcast buses for these MVMs that limited by their clock rates, and are subject to Joule heating. Emerging silicon photonics technology is capable of providing the high bandwidths necessary for DNN computations, and is able to exploit higher levels of parallelism through wavelength-division multiplexing (WDM) while minimizing energy consumption.

## II. Photonic Matrix Multiplication Architecture

In this paper, we describe the design of two hybrid optical-electrical MAC processing elements (PEs). The PEs are divided into 2 sectors. Sectors 1 (S1) and 2 (S2) perform the multiplication, where S1 performs logical $AND$, and S2 performs the accumulation of these $AND$s by shifting and adding each successive $AND$. Sector 3 (S3) is the final summation between the PE multiplications , and contains the nonlinear activation function logic, which will give the output corresponding to a neuron in a DNN. Our two hybrid PE designs are **O-E-E** (optical $AND$ - electrical accumulate - electrical summation) and **O-O-E** (optical $AND$ - optical accumulate - electrical summation), which we compare with an all electrical design **E-E-E**.

O-E-E utilizes microring resonators (MRRs) to perform optical $AND$ in S1, which couples the optical signal when a bias voltage (logical 1) is applied by the weight corresponding to that multiplication. Each optical signal then undergoes

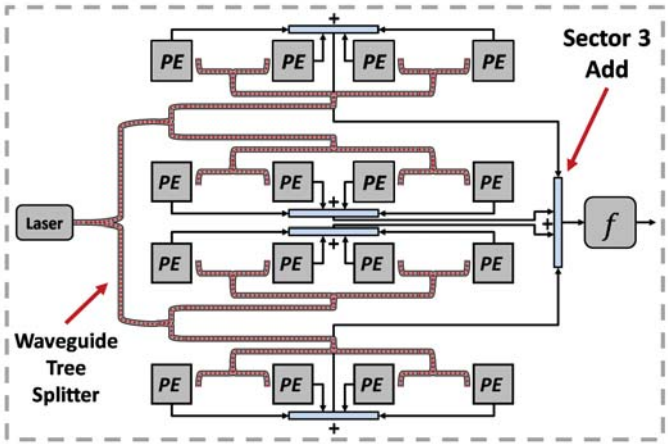


Fig. 1. **16 Processor Element (PE) arrangement with waveguide tree splitter, Sector 3 (S3) summation, and activation function**

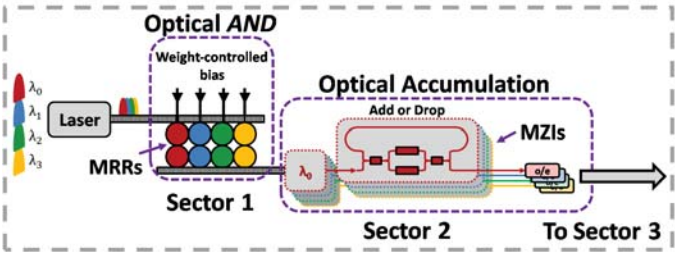


Fig. 2. **O-O-E PE architecture with 4 inputs**

an optical-to-electrical (o/e) conversion in S2, where it is then bit-shifted and added to each successive $AND$ in the multiplication. This multiplication product is then sent to the final adder tree in S3, which performs summation across several PEs, and is then fed to the activation function to give the final output of that neuron.

O-O-E utilizes the same MRR design as O-E-E. S2 of O-O-E now performs the shifting and accumulation in optics. The shifting and addition of optical signals is achieved through the employment of Mach-Zehnder interferometers (MZIs). This 2-input, 2-output devices is able to modulate signals through the use of 2 phase-shifting arms. MZIs can selectively direct an input signal to a desired output port, and has been shown to combine both inputs into a single output port in an additive manner [1]. This will lead to an optical signal with an amplitude that is the summation of the two input

signals. We connect this additive output port of the MZI back to one of its input ports to recursively accumulate signals, and the path length of this connection is designed to apply a delay to the output so it is effectively shifted before adding to the next input signal. O-O-E has the same S3 adder tree as described in O-E-E. The O-O-E design is shown in Fig. 2, and operates on multiple wavelengths by exploiting WDM, essentially increasing parallelism of the MAC.

Our accelerator design is a 4x4 PE arrangement, with each PE operating on 16 inputs (16 wavelengths for optical devices) delivered through a $0.3\,\mathrm{dB}$ waveguide splitter tree shown in Fig. 1. The S3 add with activation function is performed across the entire PE array, which gives the final output of the MVM.

## III. Results

We take our 2 hybrid PE designs and compare with an all electrical implementation. The bit precision is varied to see how the accelerators scale in terms of energy, latency, and area to perform a MVM.

Our design evaluation was performed using DSENT simulator since it offers a flexible platform for integration of both electronic and photonic components. Our electrical components were evaluated using the Bulk22LVT (22nm CMOS) tech model with $2\,\mathrm{GHz}$ clock. We performed two evaluations on photonic designs, one conservative (C) estimate and one aggressive (A) estimate. For conservative optical designs, our MRRs operate at $100\,\mathrm{fJ/bit}$ (including ring heating), and the MZIs operate at $450\,\mathrm{fJ/bit}$, all at 10 Gbps modulation. For aggressive optical designs, our MRRs operate at $50\,\mathrm{fJ/bit}$ (including ring heating), and the MZIs operate at $100\,\mathrm{fJ/bit}$, all at 12 Gbps modulation. We believe that agressive devices parameters are easily achievable in the near future since MRRs have been shown to have modulation energy as low as $7\,\mathrm{fJ/bit}$ [2] and MZIs have been demonstrated at $32.4\,\mathrm{fJ/bit}$ [3]. Our optical signals are centered around $1550\,\mathrm{nm}$ wavelength in a dense WDM (DWDM) scheme.

For a 16-input, 16-bit design, our aggressive O-O-E (A) required the least amount of energy to perform a full MVM, which consumed only 53.8% of the energy that E-E-E did. Our O-E-E (A) consumed 1.04X more energy than E-E-E because it was not able to fully take advantage of the DWDM system.

Fig. 3 shows the latency of each design and their scaling. At 16-inputs and 16-bits, O-O-E (C) is 77.8% faster than E-E-E, and O-O-E (A) is 80.8% faster than E-E-E. If we scale the design up to 64 bits, O-O-E (C) is 79.4% faster than E-E-E, and O-O-E (A) is 82.7% faster than E-E-E.

Our accelerator was also evaluated for several convolutional neural network (CNN) architectures, which were broken down by layer and analyzed to give the number of MVMs and MACs required for a single inference. We give results from this evaluation as the energy-delay product (EDP) to show the trade-off between energy-efficiency and latency. On average when compared to E-E-E, the EDP of O-O-E (C) is 33.1% lower and O-O-E (A) is 76.4% lower. The EDP comparison for each CNN architecture is shown in Fig. 4.

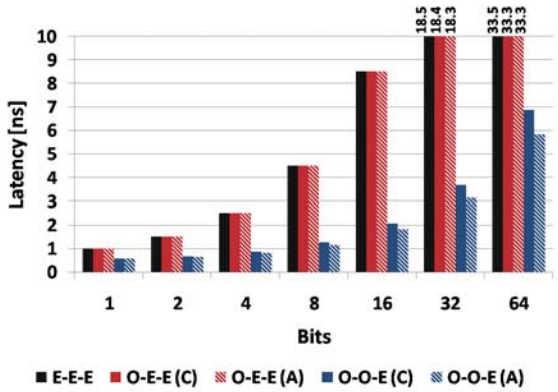


Fig. 3. **Latency for E-E-E, O-E-E (C) and (A), and O-O-E (C) and (A) designs for a 16 input MVM and varying bits**

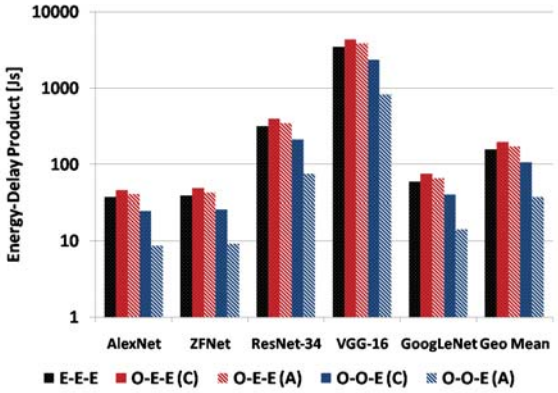


Fig. 4. **Energy-delay product (EDP) for E-E-E, O-E-E (C) and (A), and O-O-E (C) and (A) designs for a 16 PE design with 16 bit precision**

## IV. Conclusions

In this paper, we have proposed two electrical-optical hybrid MVM accelerators for use with DNNs. Our proposed architectures utilize emerging photonic devices in a manner that leverages their low-energy low-latency properties for MVM computation on the bit-level. We found that our O-O-E design gave the best performance in both energy efficiency and latency, with a reduction of 33.1% for EDP with conservative estimates and a 76.4% reduction for EDP with aggressive estimates.


## Acknowledgment

This research was partially supported by NSF grants CCF-1513606, CCF-1703013, CCF-1901192, CCF-1513923, CCF-1547034, CCF-1547035, CCF-1547036, CCF-1702980, and CCF-1901165.


## References


[1] D. A. B. Miller, "Self-aligning universal beam coupler," *Opt. Express*, vol. 21, pp. 6360–6370, Mar 2013.

[2] G. Li, X. Zheng, J. Yao, H. Thacker, I. Shubin, Y. Luo, K. Raj, J. E. Cunningham, and A. V. Krishnamoorthy, "25gb/s 1v-driving CMOS ring modulator with integrated thermal tuning," *Opt. Express*, vol. 19, pp. 20435–20443, Oct 2011.

[3] J. Ding, R. Ji, L. Zhang, and L. Yang, "Electro-optical response analysis of a 40 Gb/s silicon Mach-Zehnder optical modulator," *Journal of Lightwave Technology*, vol. 31, pp. 2434–2440, July 2013.