High-Speed and Energy-Efficient Non-Binary Computing with Polymorphic Electro-Optic Circuits and Architectures

Ishan Thakkar
Department of ECE
University of Kentucky
Lexington, Kentucky, USA
igthakkar@uky.edu

Sairam Sri Vatsavai
Department of ECE
University of Kentucky
Lexington, Kentucky, USA
ssr226@uky.edu

Venkata Sai Praneeth Karempudi
Department of ECE
University of Kentucky
Lexington, Kentucky, USA
kvspraneeth@uky.edu

ABSTRACT

In this paper, we present microring resonator (MRR) based polymorphic E-O circuits and architectures that can be employed for high-speed and energy-efficient non-binary reconfigurable computing. Our polymorphic E-O circuits can be dynamically programmed to implement different logic and arithmetic functions at different times. They can provide compactness and polymorphism to consequently improve operand handling, reduce idle time, and increase amortization of area and static power overheads. When combined with flexible photodetectors with the innate ability to accumulate a high number of optical pulses in situ, our circuits can support energy-efficient processing of data in non-binary formats such as stochastic/unary and high-dimensional reservoir formats. Furthermore, our polymorphic E-O circuits enable configurable E-O computing accelerator architectures for processing binarized and integer quantized convolutional neural networks (CNNs). We compare our designed polymorphic E-O circuits and architectures to several circuits and architectures from prior works in terms of area, latency, and energy consumption.

CCS CONCEPTS

• **Computer systems organization** → **Neural networks**; **Reconfigurable computing**; **Optical computing**.

KEYWORDS

Electro-Optic Polymorphic Circuits, Non-Binary Computing, Microring Resonators (MRRs)

ACM Reference Format:
Ishan Thakkar, Sairam Sri Vatsavai, and Venkata Sai Praneeth Karempudi. 2023. High-Speed and Energy-Efficient Non-Binary Computing with Polymorphic Electro-Optic Circuits and Architectures. In *Proceedings of the Great Lakes Symposium on VLSI 2023 (GLSVLSI '23), June 5–7, 2023, Knoxville, TN, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3583781.3590258

GLSVLSI '23, June 5–7, 2023, Knoxville, TN, USA

ACM ISBN 979-8-4007-0125-2/23/06…$15.00
https://doi.org/10.1145/3583781.3590258

1 INTRODUCTION

In recent years, Moore's law has faced fatal challenges as the nanofabrication technology is experiencing serious limitations, due to the exceedingly small size of transistors [1]. In the wake of dwindling Moore's law, fortunately, integrated electro-optic (E-O) computing circuits have shown the revolutionary potential to provide progressively faster and more efficient hardware for computing. The E-O circuits for computing, which have been demonstrated in prior works (e.g., [15, 19, 21, 22, 33–35]), are typically used to implement the following four types of logical and arithmetic functions: **(I) Basic logic-gate functions** [21, 22, 35] with two binary input operands that aid the acceleration of neural networks. **(II) Arbitrary combinational logic functions** [19, 34] that can work as the direct optical replacement of field programmable gate arrays (FPGAs). **(III) Two operand arithmetic functions** [15, 33] for accumulation that can support custom precision and full precision arithmetic operations. **(IV) Multi-operand linear arithmetic functions** [7, 17, 21, 35] to implement Multiply-Accumulate (MAC) and Vector Dot Product (VDP) operations for deep learning workloads. However, as elaborated in [16], these E-O circuits face shortcomings due to their *(i)* long idle time and resultant non-amortizable high area and static power overheads and *(ii)* strong trade-off between wavelength parallelism and achievable bit-precision.

To alleviate these shortcomings, our contribution in this paper is two-fold: **(i)** Invention of a polymorphic E-O circuit (PEOC) and **(ii)** Design of a configurable E-O computing accelerator (CEONA). We show that our PEOC can be reconfigured to implement different arithmetic and logic functions at different times. Such PEOCs are employed in our CEONA in a wavelength division multiplexing (WDM) manner to provide flexible support for accelerating convolutional neural networks (CNNs) with various bit-precisions. Furthermore, CEONA enables the acceleration of delayed feedback reservoir computing (DFRC)-based applications. We compare our designed PEOC and CEONA to several circuits and architectures from prior works and show their benefits in terms of area, latency, and energy consumption. To gain preliminary knowledge before digging into the paper, we recommend the reader to go through the tutorials on microring resonators (MRRs) [8] and optical computing architectures [7, 20].

2 POLYMORPHIC ELECTRO-OPTIC CIRCUIT

Fig. 1(a) illustrates the structure of our polymorphic electro-optic (E-O) circuit (PEOC). It consists of an active MRR that can be utilized as either a microring modulator (MRM) or a polymorphic E-O logic gate (MRR-PEOLG) [16]. Using the active MRR as an MRM enables

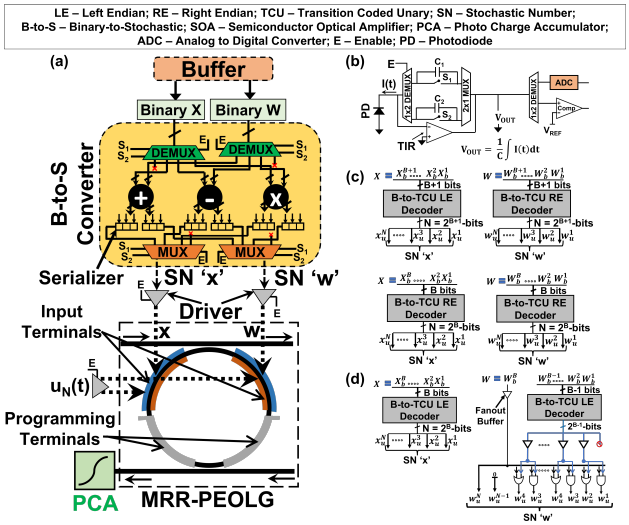

Figure 1: (a) Structure of our MRR-based polymorphic electro-optic circuit (PEOC), (b) photo-charge accumulator (PCA). B-to-S conversion circuits for ADD ((c)-top), SUB ((c)-bottom), and MUL (d) operations.

modulation of an incoming, electrical time-series signal $u_N(t)$ onto an output optical signal [32]. Moreover, when the active MRR is used as an MRR-PEOLG, it can be reconfigured to implement different logic functions at different times [16]. All of the AND, OR, XOR, NAND, NOR, and XNOR logic functions have been demonstrated [16]. In addition, in Fig. 1, a binary-to-stochastic (B-to-S) conversion circuit and a photo-charge accumulator (PCA) are integrated with the MRR-PEOLG to transform the PEOC into a polymorphic binary arithmetic unit (PBAU). PBAU leverages the OR, AND, and XOR functions of the MRR-PEOLG to implement stochastic computing (SC) based approximate addition (ADD), multiplication (MUL), and subtraction (SUB) operations [3]. More details on the structure and operation of our MRR-PEOLG, PCA, and PBAU are provided in the upcoming subsections.

Table 1: Performance comparison of E-O circuits.

Metrics	XNOR-POPCOUNT		Bit-serial Multiplier	
	[35]	MRR-PEOLG	[22]	MRR-PEOLG
A (mm^2)	0.013	0.011 (1.16×)	0.023	0.011 (2.08×)
E (nJ)	0.05	0.032 (1.53×)	0.327	0.033 (9.89×)
L (ns)	0.02	0.025 (0.8×)	0.1	0.025 (4×)
A*E*L	1.3e-5	0.9e-5 (1.44×)	75.2e-5	0.91e-5 (82.6×)

2.1 MRR-Based Polymorphic E-O Logic Gate

Our invented MRR-PEOLG is described in [16]. From [16], to program MRR-PEOLG to implement a specific logic-gate function, the MRR's operand-independent resonance position 'κ' (magenta-colored passband in Fig. 2) is adjusted to a specific spectral position with respect to the input wavelength 'λ_{in}' and the MRR's initial resonance position 'η', by applying a voltage to the programming terminals of the MRR-PEOLG (see the terminals in Fig. 1(a)). Then, the electrical input operands are applied to the PN junction-based input terminals of the MRR ((x,w) in Figs. 1 and 2). Upon doing so, the resonance of the MRR shifts towards shorter wavelengths depending on the combination of applied input operands. Applying the input operand bits to the input terminals makes the drop-port and through-port optical responses of our MRR-PEOLG follow the truth table of logic gate functions for which the MRR-PEOLG is programmed. In this manner, our MRR-PEOLG can perform different logic functions at different times (Fig. 2). To validate this polymorphic functionality of our MRR-PEOLG, we also performed a time-domain (transient) analysis using the INTERCONNECT tool of Ansys/Lumerical suite [4]. For that, we provided two electrical pulses (Figs. 3(a) and 3(b)) to the input terminals of our MRR-PEOLG and collected the output pulse patterns corresponding to different logic functions at the drop-port (Figs. 3(c), 3(e) and 3(g)) and through-port (Figs. 3(d), 3(f) and 3(h)) of our MRR-PEOLG. As evident, the output signals follow the pulse-wise truth tables of the respective logic functions, which demonstrates the capability of our MRR-PEOLG to implement different logic functions. In addition, we evaluated how the use of our MRR-PEOLG improves the area, latency, and energy consumption of two E-O circuits from prior works [35] and [22]. The results are provided in Table 1.

2.2 Photo-Charge Accumulator

Fig. 1(b) illustrates our Photo-Charge Accumulator (PCA), which is collectively inspired by the time integrating receiver (TIR) from [24] and the photodetector (PD)-based optical-pulse accumulator from [9]. Hence, our PCA employs a PD, a TIR, and two capacitors. The ADC or comparator is connected to the TIR based on the intended use case. The PD has the ability of dual coherent-incoherent superposition and photo-charge accumulation [9]. If the PD has inverse bandwidth of t=(1/symbol rate), the number of free carriers generated in the PD during every interval of t is proportional to the number of photons absorbed, and hence, the output photocurrent in the interval is proportional to the sum of average optical powers of all the coherent and incoherent optical pulses that are incident on the PD during the interval [9]. This output photocurrent is collected by the TIR of our PCA to generate a proportional voltage on one of the capacitors. This output photocurrent, and hence the voltage accrued on the capacitor, provides the accumulation result of all the optical pulses that are incident on the PD during the interval of time t [24]. The TIR circuit can allow the accumulation of a total of γ such intervals of time t each before the PCA circuit saturates. γ is known as PCA's accumulation capacity. After the optical pulse accumulation over γ intervals, a discharge of the active capacitor (e.g., C1) is needed to prepare the circuit for the next accumulation. While capacitor C1 is discharging, capacitor C2 mitigates the discharge latency by allowing a continuation of another concurrent accumulation. Table 2 reports our PCA's accumulation capacity γ at different symbol rates from [30]. At 50 GS/s, our PCA has γ=8503 (Table 2), which is greater than the required accumulation count per neuron for most modern CNNs. Such large γ for our PCA eliminates the need to decompose the required accumulations per output neuron into multiple partial sums [30].

2.3 Polymorphic Binary Arithmetic Unit

From Section 2 and Fig. 1, integrating our MRR-PEOLG with a B-to-S conversion circuit and a PCA transforms the PEOC into

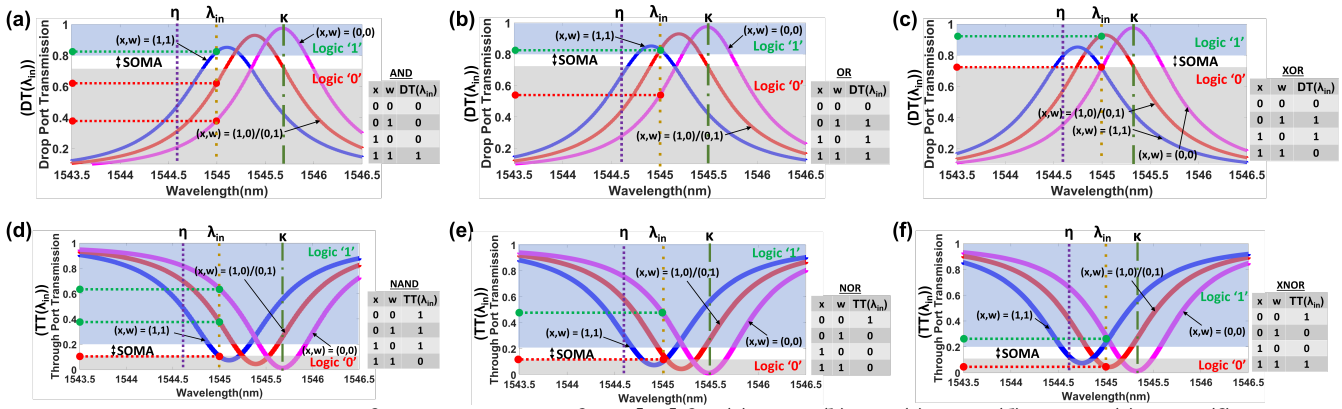

Figure 2: Transmission spectra of our MRR-PEOLG from [16] for (a) AND, (b) OR, (c) XOR, (d) NAND, (e) NOR, (f) XNOR.

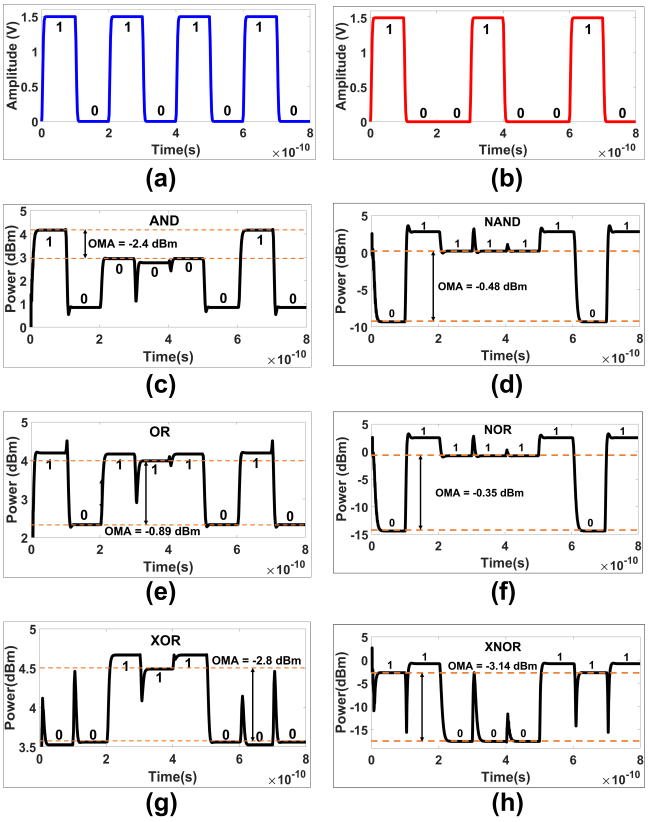

Figure 3: Transient analysis results for our MRR-PEOLG from [16]. (a),(b) Input electrical pulses. Output optical pulses for (c) AND, (d) NAND, (e) OR, (f) NOR, (g) XOR, (h) XNOR.

Table 2: PCA accumulation capacity (γ) for different symbol rates (SRs) (GS/s) [30].

SR	3	5	10	20	30	40	50
γ	39682	29761	19841	14880	10822	9920	8503

a polymorphic binary arithmetic unit (PBAU). The structure of our PBAU can be segmented into three stages. First, the B-to-S peripheral circuit aids in converting the input binary operands (N-bit) into stochastic bit-streams (2^N-bits). The B-to-S conversion, in a nutshell, is implemented through bit-parallel binary-to-transition coded unary (B-to-TCU) decoders [26]. The bit-parallel outputs of B-to-TCU decoders are then converted into bit-streams using high-speed serializers. The B-to-TCU decoders are custom designed for ADD, MUL, and SUB functions, to ensure that the input stochastic bit-streams ('x' and 'w') have an appropriate correlation to minimize the errors in their results [3]. To achieve appropriate correlation among the stochastic bit-streams, we learned from [3] to endow the function-specific B-to-S circuits (and the constituent B-to-TCU decoders) with appropriate endianness and bit-stream sizes. For instance, for ADD functions, bit-streams x and w have opposite endianness, whereas for SUB and MUL they both have the same endianness (right endianness) (Figs. 1(c) and 1(d)). Moreover, the generated stochastic bit-streams from the B-to-S conversion circuits have 2^N bits for MUL and SUB functions, whereas they have 2^{N+1} bit for ADD function [3]. Further, our B-to-S conversion circuit for MUL function (Fig. 1(d)) minimizes correlation-related errors in the results by ensuring that the conditional probability P(w/x) is equal to the marginal probability P(x) [26].

The stochastic bit-streams generated from the B-to-S conversion stage are given as input to the second stage, which consists of our MRR-PEOLG [16]. Our MRR-PEOLG as described in Section 2.1 implements AND, OR, and XOR logical functions, which are applied to the stochastic bit-streams in a bit-wise manner to implement the target MUL, ADD, and SUB functions respectively. The third stage of PBAU consists of our PCA, which converts the resultant stochastic bit-stream from the MRR-PEOLG into the binary format.

Table 3: Performance of our PBAU for ADD, SUB, and MUL. MAE=Mean Absolute Error.

Bit Precision	6-bit			8-bit		
	ADD	SUB	MUL	ADD	SUB	MUL
Latency (ns)	5.32	2.74	2.76	20.51	10.27	10.29
Energy (pJ)	16.1	6.8	10.2	60.1	23.6	36.2
MAE	0	0	0.03	0	0	0.04

In Table 3, we have evaluated the per-operation latency, energy, and mean absolute error (MAE) values of our PBAU for MUL, SUB, and ADD functions across the binary (integer) operand precision of 6-bit and 8-bit. As evident, our PBAU incurs no errors for SUB

Table 4: Comparison of latency, energy and area of PBAU with E-O arithmetic circuits from prior work.

	Area	Energy	Area*Latency
8-bit PBAU	0.0012mm^2	36.2pJ	3.312mm^2.ps
8-bit PoNALU [15]	0.6 mm^2	31.25nJ	201.3mm^2.ps
8-bit EPALU [33]	1.4 mm^2	37.5nJ	523.5 mm^2.ps
8-bit PIXEL [21]	0.00359 mm^2	51.2pJ	36.9mm^2.ps

and ADD functions, and the MAE values for MUL function are also negligibly low. Similarly, Table 4 provides a comparison of latency, energy, and area of our PBAU with E-O arithmetic circuits from prior works. From Table 4, our PBAU consumes substantially less energy and occupies less area.

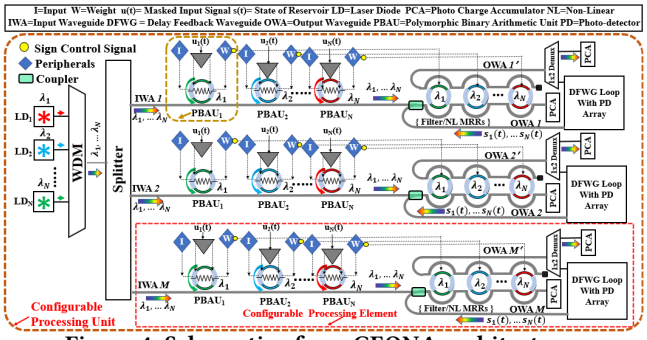

Figure 4: Schematic of our CEONA architecture.

3 CONFIGURABLE E-O COMPUTING ACCELERATOR

3.1 Overview

Multiple PBAUs are organized in an array employing wavelength division multiplexing (WDM) to constitute the main processing unit of our Configurable E-O computing Accelerator (CEONA) architecture. This unit is called a configurable processing unit (CoPU), as illustrated in Fig. 4. A CoPU consists of a comb laser source [14] that emits optical power at N distinct wavelengths (i.e., from λ_1 to λ_N). The optical power at each of these wavelengths is split into M input waveguides (IWA), each of which is connected to a configurable processing element (CoPE). Each CoPE consists of an array of N PBAUs arranged in a WDM manner aside a bank of MRRs that act as filters or non-linear active MRRs (Fig. 4). Each MRR bank connects to PCAs or a delay feedback loop waveguide (DFWG). Based on the configuration of the constituent PBAUs and MRR banks, and CoPE's connection to PCAs or DFWG, our CEONA accelerator can be employed in two use cases.

3.2 Case I: Neural Network Accelerator

CEONA can be configured to perform inference of binary neural networks (BNNs) and integer-quantized CNNs. During inference, CEONA receives weight and input operands that are 1-bit (binarized) for BNNs and 8-bit for integer-quantized CNNs.

CEONA with Binarized Operands: CEONA with binarized operands is referred to as CEONA-B. The inference of BNNs requires XNOR-Bitcount operations [30]. Therefore, with binarized operands, CEONA dynamically configures each CoPE's PBAUs as XNOR gates as discussed in Section 2. The PBAUs perform the XNOR operation between binarized I and W (Fig. 4). The optical outputs of XNOR gates are sent to the bottom OWA (Fig. 4). The MRR banks are turned off to allow all XNOR output bits to reach the bottom PCA (on OWA). The PCA counts the incoming optical bits coming from PBAUs to do bitcount operations to generate final accumulation results. Thus, each CoPE can generate one value of the output BNN tensor without requiring partial sum storage or reduction, as detailed in [30]. Our utilized evaluation setup is reported in [30].

Figs. 5(a) and 5(b) compare FPS (Frames-Per-Second)(throughput) values and FPS/W (energy-efficiency) values achieved by CEONA-B and prior accelerators across various BNNs. Overall, both CEONA-B_5 (SR=5 GS/s) and CEONA-B_50 (SR=50 GS/s) achieve better throughput and energy efficiency than other accelerators. CEONA-B_50 achieves 52×, 7×, and 7× better FPS than ROBIN_EO [28], ROBIN_PO [28], and LIGHTBULB [35], respectively, on gmean across the BNNs. Our CEONA-B_5 gains 2.6×, 3.3×, and 1.7× better FPS/W than ROBIN_EO, ROBIN_PO, and LIGHTBULB, respectively, on gmean across the BNNs. From [30], CEONA-B improves throughput by achieving larger N resulting in higher parallelism. The energy benefits come mainly from the elimination of the need to store and reduce partial sums due to the PCAs' innate capability of performing in-situ accumulations.

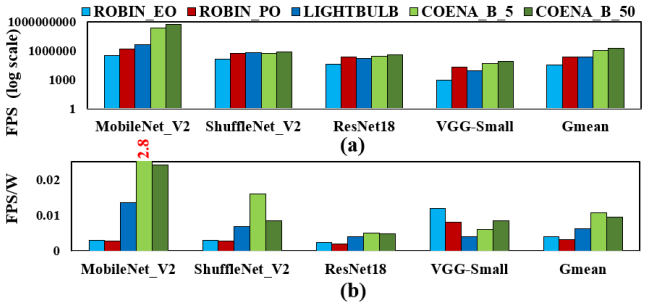

Figure 5: (a) FPS (log scale), (b) FPS/W for CEONA-B versus ROBIN [28] and LIGHTBULB [35] accelerators.

CEONA with Integer Operands: CEONA with integer operands is referred to as CEONA-I. The inference of CNNs requires dot product operations [27]. As discussed in Section 2, with stochastic computing, multiplication can be performed with AND gates. Therefore, for inference of CNNs, CEONA-I configures PBAUs to work as AND gates and they perform pointwise multiplication of I and W [31] (Fig. 4). The MRR banks are operated as filter banks, and sign control signals from corresponding PBAUs turn on/off the filters to enable signed accumulation at the PCAs. A comprehensive explanation of the CEONA-I architecture and employed evaluation setup is provided in [31].

Figs. 6(a), 6(b), and 6(c) compare the FPS (throughput) values, FPS/W (energy efficiency), and FPS/W/mm^2 (area efficiency) achieved by CEONA-I and prior accelerators across various CNNs. From Fig. 6(a), CEONA-I significantly outperforms the analog optical accelerators MAW (HOLYLIGHT) [17] and AMW (DEAPCNN) [7] by 66.5× and 146.4×, respectively, on gmean across the CNNs. From Fig. 6(b), CEONA-I gains 90× and 183× better FPS/W than analog MAW (HOLYLIGHT) and AMW (DEAPCNN), respectively, on gmean across the CNNs. From Fig. 6(c), CEONA-I gains 91× and

184× better FPS/W/mm^2 than analog MAW (HOLYLIGHT) and AMW (DEAPCNN), respectively, on gmean across the CNNs. The throughput benefits are mainly associated with the superior N of CEONA-I compared to the analog optical accelerators [31]. Moreover, the use of PCAs eliminates the need to store and reduce partial sums, providing throughput, energy, and area benefits. Overall, CEONA-I significantly improves throughput, energy efficiency, and area efficiency compared to prior analog accelerators.

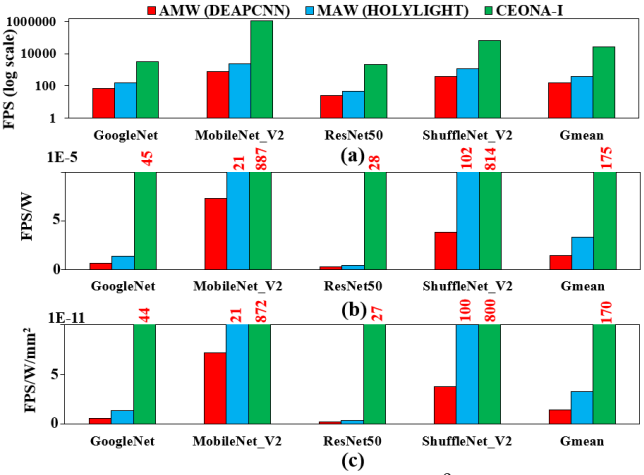

Figure 6: (a) FPS, (b) FPS/W, (c) FPS/W/mm^2 for CEONA-I versus MAW and AMW accelerators for 8-bit integer precision.

Scalability of CEONA-I: To determine the achievable size N for our CEONA-I CoPU across various integer precision levels (B), we adopt scalability analysis equations (Eq. 1, Eq. 2, and Eq. 3) from [2] and [27]. The definitions of the parameters and their values used in these equations are reported in [31]. In Eq. 1, for AMW and MAW architectures, the datarate (DR) is equal to the symbol rate (SR) and $n_{i/p}$=B. But, for CEONA-I architecture, DR=(SR/2^B) and $n_{i/p}$=1 [31]. We consider M=N and solve the equations using the method from [31]. Fig. 7 reports the achievable N of CEONA-I, AMW, and MAW architectures for different B levels across various *SR*s. As evident from Fig. 7, our CEONA-I can support larger N value compared to AMW and MAW at all bit-precision levels across different SRs. For instance, CEONA-I achieves larger *N=192* for 4-bit precision at 1 GS/s, compared to AMW and MAW, which achieve *N=31* and *N=44*, respectively. This is because of CEONA-I's PCAs' in-situ accumulation capacity [9] (see Section 2.2). It allows the PCAs to operate at lower DR=(SR/2^B) which significantly improves N at larger B. In contrast, the support for N decreases for AMW and MAW with an increase in B [27]. Furthermore, the achievable N is also limited by inter-wavelength spacing. We consider optimistic FSR=50nm. For AMW and MAW, the inter-wavelength spacing is set to 0.8nm [27] whereas for CEONA-I it can be set to 0.25nm [31]. Therefore, in Fig. 7, the N values for AMW/MAW and CEONA-I are capped at 62 (=FSR/0.8nm) and 200 (=FSR/0.25nm) respectively.

$$n_{i/p} = \frac{1}{6.02}\left[20log_{10}\left(\frac{R \times P_{PD-opt}}{\beta\sqrt{\frac{DR}{\sqrt{2}}}} - 1.76\right)\right] \quad (1)$$

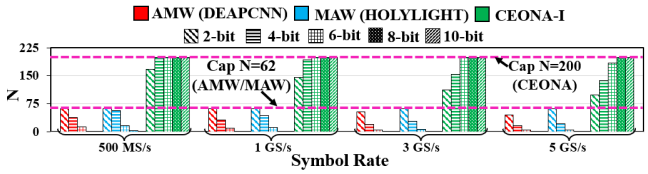

Figure 7: Supported CoPE size N for bit precision ={2, 4, 6, 8, 10}bits at symbol rates (SRs) = {0.5, 1, 3, 5}GS/s, for AMW, MAW, and CEONA-I.

$$\beta = \sqrt{2q(RP_{PD-opt} + I_d) + \frac{4kT}{R_L} + R^2P^2_{PD-opt}RIN} \quad (2)$$

$$P_{Laser} = \frac{10^{\frac{\eta_{WG}(dB)[N(d_{OSM})]}{10}}M}{\eta_{SMF}\eta_{EC}IL_{i/p-OSM}} \times \frac{P_{PD-opt}}{\eta_{WPE}IL_{MRR}} \times \frac{1}{(OBL_{OSM})^{N-1}(EL_{splitter})^{log_2M}} \times \frac{1}{(OBL_{MRR})^{N-1}(IL_{penalty})} \quad (3)$$

3.3 Case II: Reservoir Computing

CEONA can also be configured as a delay feedback reservoir computing accelerator (CEONA-DFRC) for training and inference of time series tasks. For that, CEONA-DFRC configures the PBAUs to work as conventional modulators to modulate input masked signals $u_i(t)$ (Fig. 4). Each MRR in the filter banks is configured to act as a non-linear node of the reservoir [5].

An active MRR shows the rich non-linear response at its drop-port transmission due to Two-Photon Absorption (TPA) [6]. The degree of non-linearity depends on the photon lifetime (τ_{ph}) of the MRR cavity. For an MRR, τ_{ph} depends on the MRR's Q-factor. Therefore, the non-linearity of the MRR can be controlled with the Q-factor (hence, τ_{ph}) of the MRR. To enable control of the MRRs' Q-factor (hence, τ_{ph}), we employ the non-linear MRR design from [18, 23] in Fig. 4.

In Fig. 4, the non-linear MRRs along with the DFWG loop form a reservoir [32]. The MRRs' quality factor is adapted [18, 29] to set the degree of nonlinearity depending on the task. The MRRs generate the states of the reservoir by a non-linear transformation of modulated $u_i(t)$. The states $s_i(t)$ of the reservoir are sent to the DFWG loop where they are captured and stored using the PD array, and part of $s_i(t)$ signals are further fed as feedback to the non-linear MRRs with modulated masked inputs $u_i(t+1)$. The operation and architecture of CEONA-DFRC are discussed extensively in [32].

Figs. 8(a), 8(b), and 8(c) compare Symbol Error Rate (SER), Normalized Root Mean Square Error (NRMSE), and training time of various DFRC accelerators across various time series tasks. From Fig. 8(a), on average across various target SNRs, CEONA-DFRC achieves 58.8% lower SER than All_Optical (MZI) [11] on non-linear channel equalization task [13]. Similarly, Fig. 8(b) shows that for NARMA [12] and SantaFe [25], CEONA-DFRC achieves 35% lower NRMSE compared to All_Optical (MZI), and it performs on par with Electronic (MG) [5]. CEONA-DFRC's major benefit can be observed in Fig. 8(c); it significantly speedups training time by 98×

and 93× on average compared to All_Optical (MZI) and Electronic (MG) respectively. CEONA-DFRC leverages the rich non-linearity of the active MRR to realize the non-linear node in the reservoir layer. Moreover, the MRR-based reservoir for CEONA-DFRC takes 168× and $2*10^5$× less time to transform the masked input signal compared to All_Optical MZI and Electronic MG reservoirs. respectively. In addition, CEONA-DFRC uses a photonic waveguide as the delay feedback loop which further reduces the training time [32]. Overall, CEONA-DFRC significantly improves the training time of time series tasks while achieving better or on-par error compared to prior optical and electronic DFRCs.

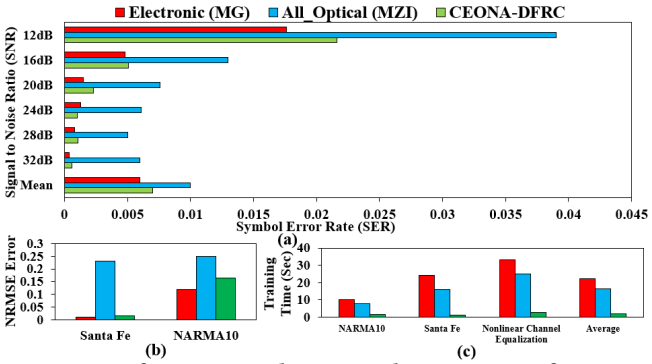

Figure 8: Performance evaluation and comparison of CEONA-DFRC, reproduced from [32]. (a) SER for various target SNRs for channel equalization task, (b) NEMSE for Santa Fe and NARMA series prediction, and (c) training time comparison for considered tasks.

4 SUMMARY AND FUTURE PROSPECTS

In this paper, we demonstrated our invented PEOC and CEONA. We showed that PEOC can be reconfigured to implement different logical and arithmetic functions at different times. Similarly, we showed that CEONA can be used to enable flexible support for accelerating CNNs and DFRC applications. We have shown that our CEONA can be reconfigured to handle binarized CNNs and integer-quantized CNNs. In both cases, CEONA provides significant benefits in throughput, energy efficiency, and area efficiency compared to prior works. In addition, we have also shown that CEONA achieves significant latency benefits for accelerating inference and training of various time series tasks such as NARMA10, SantaFe, and Non-linear Channel Equalization.

We envision that our CEONA can be extended to support the acceleration of mixed-precision CNNs [10]. Furthermore, each CoPE of our CEONA architecture can accelerate multiple time series tasks in a WDM manner to significantly improve the accelerator throughput. Overall, this flexibility can enable CEONA to simultaneously handle multiple workloads from machine learning and artificial intelligence applications.

ACKNOWLEDGMENTS

This research is supported by a grant from NSF (CNS-2139167).

REFERENCES

[1] A.Ganguly *et al.* 2022. Interconnects for DNA, Quantum, In-Memory, and Optical Computing: Insights From a Panel Discussion. *IEEE micro* 42, 3 (2022), 40–49.
[2] M. A. Al-Qadasi et al. 2022. Scaling up silicon photonic based accelerators: Challenges and opportunities. *APL Photonics* (2022).
[3] Armin Alaghi and John P Hayes. 2013. Exploiting correlation in stochastic circuit design. In *2013 IEEE 31st International Conference on Computer Design (ICCD)*. IEEE, 39–46.
[4] ANSYS. 2003. Lumerical. http://www.lumerical.com/products
[5] Lennert Appeltant et al. 2011. Information processing using a single dynamical node as complex system. *Nature Communications* 2 (2011).
[6] Meisam Bahadori et al. 2017. Energy-performance optimized design of silicon photonic interconnection networks for high-performance computing. In *DATE*.
[7] Viraj Bangari et al. 2020. Digital Electronics and Analog Photonics for Convolutional Neural Networks (DEAP-CNNs). *JSTQE* (2020).
[8] W. Bogaerts et al. 2012. Silicon microring resonators. *Laser & Photonics Reviews* 6, 1 (2012), 47–73.
[9] Frank Brückerhoff-Plückelmann et al. 2022. A large scale photonic matrix processor enabled by charge accumulation. *Nanophotonics* (2022).
[10] Weihan Chen et al. 2021. Towards Mixed-Precision Quantization of Neural Networks via Constrained Optimization. In *ICCV*.
[11] François Duport et al. 2016. Fully analogue photonic reservoir computer. *Scientific Reports* (2016).
[12] Herbert Jaeger. 2002. Adaptive Nonlinear System Identification with Echo State Networks. In *NIPS*.
[13] Herbert Jaeger et al. 2004. Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. *Science* (2004).
[14] Venkata Sai Praneeth Karempudi et al. 2022. Photonic Networks-on-Chip Employing Multilevel Signaling: A Cross-Layer Comparative Study. *JETCS* (2022).
[15] Venkata Sai Praneeth Karempudi, Shreyan Datta, and Ishan G Thakkar. 2021. Design Exploration and Scalability Analysis of a CMOS-Integrated, Polymorphic, Nanophotonic Arithmetic-Logic Unit. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 628–634.
[16] V. Karempudi *et al.* 2023. A Polymorphic Electro-Optic Logic Gate for High-Speed Reconfigurable Computing Circuits. *arXiv preprint arXiv:2301.13626* (2023).
[17] Weichen Liu et al. 2019. HolyLight: A Nanophotonic Accelerator for Deep Learning in Data Centers. In *DATE*.
[18] Enxiao Luan et al. 2023. Towards a high-density photonic tensor core enabled by intensity-modulated microrings and photonic wire bonding. *Scientific Reports* (2023).
[19] C. Qiu *et al.* 2012. Demonstration of reconfigurable electro-optical logic with silicon photonic integrated circuits. *Optics letters* 37, 19 (2012), 3942–3944.
[20] B.J. Shastri et al. 2021. Photonics for artificial intelligence and neuromorphic computing. *Nat. Photonics* 15 (2021), 102–114.
[21] Kyle Shiflett et al. 2020. PIXEL: Photonic Neural Network Accelerator. In *HPCA*.
[22] K. Shiflett*et al.* 2021. Bitwise Neural Network Acceleration Using Silicon Photonics. In *GLSVLSI*. 9–14.
[23] Hossam Shoman et al. 2018. Compact Silicon Microring Modulator with Tunable Extinction Ratio and Wide FSR. In *OFC*.
[24] Alexander Sludds et al. 2022. Delocalized photonic deep learning on the internet's edge. *Science* (2022).
[25] Andrew R. Solow. 1994. Forecasting the Future and Understanding the Past. *Science* (1994).
[26] Sairam Sri Vatsavai and Ishan Thakkar. 2023. A Bit-Parallel Deterministic Stochastic Multiplier. *arXiv e-prints* (2023), arXiv–2302.
[27] Sairam Sri Vatsavai and Ishan G. Thakkar. 2022. Photonic Reconfigurable Accelerators for Efficient Inference of CNNs With Mixed-Sized Tensors. *TCAD* (2022).
[28] Febin P. Sunny et al. 2021. ROBIN: A Robust Optical Binary Neural Network Accelerator. *ACM Trans. Embed. Comput. Syst.* (2021).
[29] Sairam Sri Vatsavai et al. 2020. PROTEUS: Rule-Based Self-Adaptation in Photonic NoCs for Loss-Aware Co-Management of Laser Power and Performance. In *NOCS*.
[30] Sairam Sri Vatsavai et al. 2023. An Optical XNOR-Bitcount Based Accelerator for Efficient Inference of Binary Neural Networks. arXiv:2302.06405 [cs.AR]
[31] Sairam Sri Vatsavai et al. 2023. SCONNA: A Stochastic Computing Based Optical Accelerator for Ultra-Fast, Energy-Efficient Inference of Integer-Quantized CNNs. arXiv:2302.07036 [cs.AR]
[32] Sairam Sri Vatsavai and Ishan Thakkar. 2021. Silicon Photonic Microring Based Chip-Scale Accelerator for Delayed Feedback Reservoir Computing. In *VLSID*.
[33] Zhoufeng Ying, Chenghao Feng, Zheng Zhao, Shounak Dhar, Hamed Dalir, Jiaqi Gu, Yue Cheng, Richard Soref, David Z Pan, and Ray T Chen. 2020. Electronic-photonic arithmetic logic unit for high-speed computing. *Nature communications* 11, 1 (2020), 2154.
[34] Z Ying *et al.* 2019. Integrated multi-operand electro-optic logic gates for optical computing. *APL* (2019).
[35] F. Zoakee *et al.* 2020. LightBulb: A photonic-nonvolatile-memory-based accelerator for binarized convolutional neural networks. In *2020 DATE*. IEEE, 1438–1443.