2023 IEEE International Parallel and Distributed Processing Symposium (IPDPS) | 979-8-3503-3766-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/IPDPS54959.2023.00061

SCONNA: A Stochastic Computing Based Optical Accelerator for Ultra-Fast, Energy-Efficient Inference of Integer-Quantized CNNs

Sairam Sri Vatsavai, Venkata Sai Praneeth Karempudi, Ishan Thakkar, Ahmad Salehi, and Todd Hastings
Department of Electrical and Computer Engineering
University of Kentucky, Lexington, KY 40506, USA
{ssr226, kvspraneeth, igthakkar, sayedsalehi, and todd.hastings}@uky.edu

***Abstract*—Convolutional Neural Networks (CNNs) are used extensively for artificial intelligence applications due to their record-breaking accuracy. For efficient and swift hardware-based acceleration, CNNs are typically quantized to have integer input/weight parameters. The acceleration of a CNN inference task uses convolution operations that are typically transformed into vector-dot-product (VDP) operations. Several photonic microring resonators (MRRs) based hardware architectures have been proposed to accelerate integer-quantized CNNs with remarkably higher throughput and energy efficiency compared to their electronic counterparts. However, the existing photonic MRR-based analog accelerators exhibit a very strong trade-off between the achievable input/weight precision and VDP operation size, which severely restricts their achievable VDP operation size for the quantized input/weight precision of 4 bits and higher. The restricted VDP operation size ultimately suppresses computing throughput to severely diminish the achievable performance benefits. To address this shortcoming, we for the first time present a merger of stochastic computing and MRR-based CNN accelerators. To leverage the innate precision flexibility of stochastic computing, we invent an MRR-based optical stochastic multiplier (OSM). We employ multiple OSMs in a cascaded manner using dense wavelength division multiplexing, to forge a novel Stochastic Computing based Optical Neural Network Accelerator (SCONNA). SCONNA achieves significantly high throughput and energy efficiency for accelerating inferences of high-precision quantized CNNs. Our evaluation for the inference of four modern CNNs at 8-bit input/weight precision indicates that SCONNA provides improvements of up to 66.5×, 90×, and 91× in frames-per-second (FPS), FPS/W and FPS/W/mm^2, respectively, on average over two photonic MRR-based analog CNN accelerators from prior work, with Top-1 accuracy drop of only up to 0.4% for large CNNs and up to 1.5% for small CNNs. We developed a transaction-level, event-driven python-based simulator for the evaluation of SCONNA and other accelerators (https://github.com/uky-UCAT/SC_ONN_SIM.git).**

I. Introduction

Deep Neural Networks (DNNs) have revolutionized the implementation of various artificial intelligence tasks, such as image recognition, language translation, autonomous driving [1], [2], due to their high inference accuracy. Convolutional Neural Networks (CNNs) are specific types of DNNs [3]. CNNs are computationally intensive, and hence, require a long inference time. In CNNs, around 80% of the total processing time is taken by convolution operations that can be decomposed into vector dot product (VDP) operations [4]. The ever-increasing complexity of CNNs has pushed for highly customized CNN hardware accelerators [5]. Often, for efficient and swift hardware-based acceleration, CNNs are typically quantized to have integer input/weight parameters [6]. Among CNN hardware accelerators, silicon-photonic accelerators have shown great promise to provide unparalleled parallelism, ultra-low latency, and high energy efficiency [7]–[12]. Typically, a silicon-photonic CNN accelerator consists of multiple Vector Dot Product Cores (VDPCs) that perform multiple VDP operations in parallel. Several VDPC-based optical CNN accelerators have been proposed in prior works based on various silicon-photonic devices, such as Mach Zehnder Interferometer (MZI) (e.g., [13], [14], [15]) and Microring Resonator (MRR) (e.g., [9], [12], [16], [17]).

Among these optical VDPC-based CNN accelerators from prior work, the MRR-enabled VDPC-based accelerators (e.g., [7]–[9], [12], [17], [18]) have shown disruptive performance and energy efficiencies, due to the MRRs' compact footprint, low dynamic power consumption, and compatibility with cascaded dense-wavelength-division-multiplexing (DWDM). Among these MRR-enabled accelerators, some accelerators utilize digital VDPCs (e.g., [18]), whereas some others employ analog VDPCs (e.g., [9], [12], [17]). In general, a VDPC (analog or digital) transforms convolution operations into vector dot product (VDP) operations by decomposing the input tensors into vectors (1D tensors). In an analog VDPC, such VDP operations are also analog in nature, and they are performed on the individual VDP elements (VDPEs), which are the main MRR-enabled hardware components in the VDPCs. Multiple VDPEs in an analog VDPC can perform multiple analog VDP operations in parallel. The results of these analog VDP operations are converted into the digital format using analog-to-digital converters (ADCs). These results can be summed together (if and when needed) using a partial-sum (*psum*) reduction network, which can be employed outside of the VDPCs as part of the post-processing components of the CNN accelerator. The functioning of the analog VDPCs and their constituent VDPEs in the ultra-high-speed, analog-optical domain results in disruptive throughput for performing analog VDP operations.

We observe that two factors govern the performance of such analog optical VDPCs: (1) the achievable bit-precision (B) and (2) the achievable scalability of the VDPCs, i.e., the achievable count of the individual VDPEs per VDPC (M) and

1530-2075/23/$31.00 ©2023 IEEE
DOI 10.1109/IPDPS54959.2023.00061

the individual VDPE size (the number of multiplications that can be generated and summed up per VDPE) (N). In an analog VDPC, the achievable B affects the inference accuracy of the processed CNNs, whereas the achievable VDPC scalability (i.e., N and M) directly defines the throughput of the VDPC for processing CNNs. Prior works [19] and [20] studied various factors such as optical power budget in waveguides, inter-channel spacing of wavelengths, crosstalk at cascaded MRRs, resolution of ADCs, and photodetector responsivity, to determine the bounds of the achievable B and scalability in analog optical VDPCs. Furthermore, prior work [21] characterized the very strong trade-off between the maximum achievable VDPC size N and B in analog optical VDPCs. From [21], the analog optical VDPCs from prior works cannot support N greater than 44 for B>=4-bit [21]. Achieving such low N can seriously hurt the performance for processing modern CNNs. This is because modern CNNs employ tensors with as high as 4608 points (parameters) per tensor [22]. Processing such large tensors on a VDPC with $N \leq 44$ results in a large number of *psums*, resulting in a very high latency overhead in the *psum* reduction network.

To avoid this undesired outcome, we advocate for such an architecture of MRRs-based CNN accelerator that achieves significantly larger VDPC size N along with weakened interdependence between N and B. To that end, for the first time, we leveraged the inherent precision flexibility of stochastic computing to come up with a novel, MMRs-enabled **S**tochastic **C**omputing based **O**ptical **N**eural **N**etwork **A**ccelerator (SCONNA). SCONNA employs our invented MRR-based Optical Stochastic Multipliers (OSMs) to realize manifold improvements in the throughput and energy efficiency of processing integer-quantized CNNs.

Our key contributions in this paper are summarized below:

- To enable stochastic computing in the optical domain, we present (i) a novel design of optical stochastic multiplier (OSM), and (ii) a novel photo-charge accumulator (PCA) circuit (Section IV);
- We present detailed modeling and characterization of our invented OSM and PCA using foundry-validated, commercial-grade, photonic-electronic design automation tools (Section IV);
- We employ our designed OSMs and PCAs to forge a highly scalable CNN accelerator named SCONNA, which employs OSM and PCA-based scalable VDPCs (Section IV);
- We perform a comprehensive scalability analysis for our SCONNA VDPCs, to determine their achievable maximum size **N**, operating speed, and error susceptibility (Section V);
- We implement and evaluate SCONNA at the system-level using our in-house simulator (https://github.com/uky-UCAT/SC_ONN_SIM.git), and compare its performance and inference accuracy for processing 8-bit integer-quantized CNNs with two widely-known MRR-based analog CNN accelerators from prior works (Section VI).

II. Preliminaries

A. Convolutional Neural Networks (CNNs)

CNNs are specific types of DNNs that have shown remarkable accuracy for image classification. In general, a CNN consists of multiple convolutional layers, pooling layers, and fully connected layers. As shown in Fig. 1, a typical convolutional layer consists of one input tensor $\mathcal{I}(H,W,D)$ and L kernel tensors $\mathcal{F}(K,K,D)$. All of the L kernel tensors convolve over the input tensor using stride (ψ) to produce the output tensor $\mathcal{O}(H^{Out},W^{Out},L)$.

The computation required to produce each point $O(i, j, l)$ in the output tensor $\mathcal{O}(H^{out},W^{out},L)$ can be given as Eq. 1.

$$O(i,j,l) = \sum_{d=1}^{D} \sum_{q=1}^{K} \sum_{r=1}^{K} F(r,q,d) I(i \times \psi + r, j \times \psi + q, d) \quad (1)$$

Here, d=[1,D], q=[1,K], r=[1,K], i=[1,H^{Out}], l=[1,L], and j=[1,W^{Out}] are various indices and their value ranges for the kernel and output tensors. $O(i,j,l)$ in Eq. 1 is the sum of a total of $K \times K \times D$ products (products of the individual points of tensors $\mathcal{F}$ and $\mathcal{I}(K,K,D)$; $\mathcal{I}(K,K,D)$ is the gray-highlighted part of $\mathcal{I}(H,W,D)$ in Fig. 1). Thus, producing $O(i,j,l)$ requires $K \times K \times D$ point-wise multiplications (to produce $K \times K \times D$ point-wise products) and one sum-of-products operation. The combination of these point-wise multiplications and the corresponding sum-of-products operation is mathematically equivalent to a Vector Dot Product (VDP) operation. A VDP operation typically occurs between two vectors. This implies that I and F in Eq. 1 are vectors, which are basically flattened (in 1D) versions of tensors $\mathcal{I}(K,K,D)$ and $\mathcal{F}(K,K,D)$ respectively. Note that vectors I and K have a total of $S = K \times K \times D$ points each. Henceforth, We refer to I and K as input vector and kernel vector, respectively.

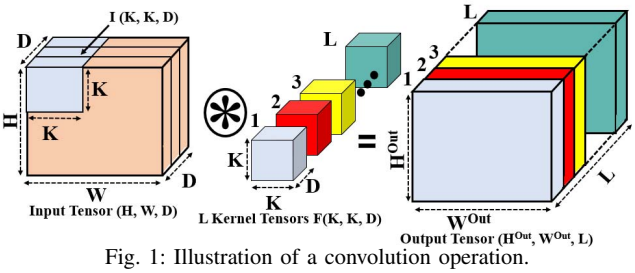

Fig. 1: Illustration of a convolution operation.

B. Processing Convolutions on VDPCs

Producing the output tensor $\mathcal{O}(H^{Out},W^{Out},L)$ (Fig. 1) requires that the VDP operation shown in Eq. 1 is implemented multiple times, i.e., a total of $H^{Out} \times W^{Out} \times L$ times. In Eq. 1, the output $O(i,j,l)$ is the result of the VDP operation between the corresponding input vector and kernel vector, each of size $S = K \times K \times D$ (Section II.A). Typically, for a CNN, the values K and D vary dramatically across different kernel tensors of the CNN. Therefore, $S = K \times K \times D$ also varies dramatically. The value S for CNNs can be as large as 4608 (e.g., ResNet50 [21]). Because of such large S, to accelerate VDP operations on a VDPC, it is intuitive to have the size N of

the constituent VDPEs of the VDPC (defined as the number of point-wise multiplications a VDPE can concurrently perform) to be as large as *S*. However, it is hardly possible to have *N* to be equal to *S* in optical MRR-based analog VDPCs. Therefore, input vector and kernel vector are generally divided into multiple decomposed input vectors (*DIV*s) (this and other abbreviations are defined in Table III) and decomposed kernel vectors (*DKV*s) first, and then these *DIV*s and *DKV*s are processed on the VDPEs (Section III.A). Having to decompose the input vector and kernel vector into multiple *DIV*s and *DKV*s raises several challenges as discussed in Section III.A.

C. Optical Analog VDPC-Based CNN Accelerators

Most of the optical MRR-enabled analog, incoherent CNN accelerators from prior work employ multiple optical analog VDPCs that work in parallel. A brief review of prior works on optical accelerators is provided in Section VII. Typically, an analog VDPC implements the decomposed VDP operations of a convolution operation using DKVs and DIVs (Section II.A). In general, a VDPC consists of five blocks (Fig. 2(a)): *(i)* a laser block that consists of *N* laser diodes (LDs) to generate *N* optical wavelength channels; *(ii)* an aggregation block that aggregates the generated optical wavelength channels into a single photonic waveguide through dense wavelength division multiplexing (DWDM) (using an $N\times1$ multiplexer) and then splits the optical power of these *N* wavelength channels equally into *M* separate waveguides (using a $1\times M$ splitter); *(iii)* a modulation block, also referred to as *DIV* block, that employs *M* arrays of MRRs (one array per waveguide, with each array having *N* MRRs; each array referred to as *DIV* element) to imprint *M* DIVs of *N* points each onto the $N\times M$ wavelength channels by modulating the analog power amplitudes of the wavelength channels; *(iv)* another modulation block, referred to as *DKV* block, that employs another *M* arrays of MRRs (one array per waveguide, with each array having *N* MRRs; each array referred to as *DKV* element) to further modulate the $N\times M$ wavelength channels with *DKV*s, so that the analog power amplitudes of the individual wavelength channels then represent the point-wise products of the utilized *DKV*s and *DIV*s; and *(v)* a summation block (SB) that employs a total of *M* summation elements (SEs), with each SE having two balanced photodiodes (PDs) upon which the point-wise-product-modulated *N* wavelength channels are incident to produce the output current that is proportional to the result of the VDP operation between the corresponding *DKV* and *DIV*. The laser block and SB are typically positioned at the two ends of the VDPC, with the aggregation, modulation (*DIV*), and modulation (*DKV*) blocks placed in between them.

Based on the order in which these intermediate blocks (aggregation, modulation (*DIV*), modulation (*DKV*) blocks) are positioned between the laser block and SB, we classify the MRR-based VDPC organizations from prior work as MAM (Modulation, Aggregation, Modulation) (e.g., [7], [19]) or AMM (Aggregation, Modulation, Modulation) (e.g., [11], [8], [9]). Fig. 2 illustrates MAM and AMM VDPC organizations. From Fig. 2(a), the AMM VDPC organization positions the aggregation block first after the laser block, and then the *DIV* modulation block followed by the *DKV* modulation block. In contrast, the MAM VDPC in Fig. 2(b) positions the *DIV* modulation block first after the laser block, and then positions the aggregation block followed by the DKV modulation block. Note that the MAM *DIV* block is structurally different from the AMM *DIV* block. The MAM *DIV* block employs only one MRR per waveguide, and as a result, it can imprint only one *DIV* with *N* points onto the *N* wavelength channels. This one *DIV* is shared among all DKVs in the MAM VDPC, whereas each *DKV* can have a different *DIV* corresponding to it in the AMM VDPC. Most MAM and AMM VDPCs from prior works have $M{=}N$.

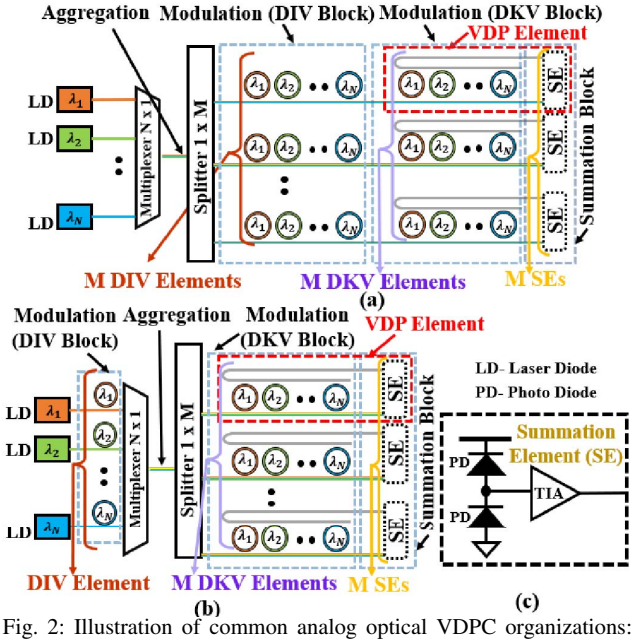

Fig. 2: Illustration of common analog optical VDPC organizations: (a) AMM VDPC, (b) MAM VDPC. (c) Summation Element.

In both the AMM and MAM VDPC organizations, we refer to the combination of a *DKV* element and the corresponding SE as VDP element (VDPE). However, the size and point-wise product precision of MRR-based VDPEs have certain limitations (discussed in Section III). These limitations demand exploration of new computing options to improve MRR-based VDPCs, and stochastic computing is an attractive option.

D. Stochastic Computing

Stochastic Computing (SC) is an unconventional form of computing that represents and processes data in the form of probabilistic values called stochastic numbers (SNs) [23]–[25]. In SC's unipolar format, an SN *W* is a bit-stream of *N* bits that represents a real-valued variable $v \in [0,1]$ by encoding v through the ratio N_1/N, where N_1 is the number of 1's in *W*. SC offers several advantages over conventional binary computing such as high error tolerance, low power consumption, small circuit area, and low-cost arithmetic operations consisting of standard digital logic components [25]. For example, multiplication can be performed by a stochastic circuit consisting of a single AND gate.

Fig. 3: Multiplication between unipolar stochastic numbers *I* and *W*.

Fig. 3 illustrates a multiplication between two unipolar stochastic bit-streams *I* and *W* using an AND gate. The probabilities of seeing '1's in the bit-streams *I* and *W* are (4/8) and (6/8), respectively. The AND gate performs bit-wise logical AND operation on the bit-streams to produce the output bit-stream *A*. In *A*, the probability of seeing '1's is (3/8), which is equal to (4/8)×(6/8), i.e., the multiplication (or product) of the input probabilities. Note that for the AND gate to produce an error-free multiplication output, the marginal probability of one bit-stream (i.e., *I* or *W*) should be equal to its conditional probability given the other bit-stream (i.e., *I* given *W* or *W* given *I*) [26]. Also note that, because of its advantages, SC has been adopted in stochastic deep CNNs [27]–[29], GEMM computation [26], and image processing [30]. We use stochastic computing in this paper to relax the inherently strong scalability-precision trade-off in the optical VDPCs. This trade-off is explained in the next section.

III. Motivation

A. Scalability Limitations of MRR-Based Analog VDPCs

Prior works [31], [21], and [19] have analysed the scalability (i.e., achievable value of VDPE size *N* under the constraints of bit precision and data rate) of AMM and MAM VDPCs. Table I reproduces the supported values of VDPE size *N* (considering *M*=*N*) for AMM and MAM VDPCs at various data rates (DRs) and bit precision from [21]. From Table 1, the maximum *N*=44 is obtained for MAM VDPC across all tested DR and B values. For MAM VDPC for 1 GS/s, maximum *N* reduces from 44 to 12 as we increase the input/weight precision from 4-bit to 6-bit. The reason for such strong trade-off between *N* and achievable input/weight precision (referred to as *B*, henceforth) in MAM and AMM VDPCs is that both *B* and *N* strongly depend on the number of distinguishable analog optical power levels [21] [31], which is proportional to $N \times 2^B$. Hence, for a fixed number of distinguishable analog optical power levels, which is defined by the analog optical power resolution of the utilized summation elements (SEs) (see SEs in Fig. 2) the supported *N* drastically decreases with an increase in *B*. As a result, *N* decreases all the way to 1 when *B* increases to 8-bit [21].

Due to such strong trade-off between *N* and *B*, the MAM and AMM type of analog VDPCs face two consequences. First, they produce high number of partial sums and incur significantly high latency for partial sum reduction. For example, a VDPE with *N*=44 for *B*=4-bit can only produce a VDP operation between two 44-point vectors. Therefore, producing a VDP operation between an input vector and kernel vector with *S*=4608 (e.g., ResNet50 [22] [21]) requires that the input vector is first decomposed into a total of *C*=Ceil(S/N)=105 *DIV*s of *N*=44 points each. Similarly, the kernel vector also needs to be decomposed into a total of *C*=Ceil(S/N)=105 *DKV*s of *N*=44 points each. Then, a total of 105 VDPEs can be employed to perform 105 VDP operations between 105 pairs of *DKV*s and *DIV*s, to consequently produce a total of 105 intermediate VDP results (i.e., partial sums (*psums*)). Although these 105 VDP operations can be parallelized over 105 VDPEs, producing the final VDP result of *S*=4608 would require the accumulation of the 105 *psums*. Doing so can incur very high latency and energy consumption, which should be avoided using a more efficient VDPC design.

TABLE I: VDPE size *N* for input/weight precision={4,6}-bit at data rates (DRs)={1,3,5,10}GS/s, for AMM and MAM VDPCs.

VDPC	Precision	Datarate(DR)			
		1 GS/s	3 GS/s	5 GS/s	10 GS/s
AMM	4-bit	31	20	16	11
	6-bit	6	3	2	1
MAM	4-bit	44	29	22	16
	6-bit	12	7	5	3

TABLE II: Total number of kernels (T_L) of different DKV sizes (*S*) for various CNNs. The T_L values were extracted for trained CNN models from Keras Applications [32].

Model	T_L	S	Model	T_L	S
ResNet50	1	S≤44	GoogleNet	13	S≤44
	26562	S>44		7554	S>44
VGG16	69	S≤44	DenseNet	1	S≤44
	4168	S>44		10242	S>44

As the second consequence, the throughput of the MAM and AMM VDPCs decreases at higher bit precision (higher value of *B*). This is because to avoid a drastic decrease in *N* as *B* increases beyond 4-bit, the AMM and MAM type of analog VDPCs typically operate at *B*=4-bit [21]. However, using at least 8-bit precision (*B*=8-bit) for the integer-quantized CNN models is recommended to achieve competitive inference accuracy, while also reducing the computational effort, memory requirements, and energy consumption [6]. To meet this requirement, analog VDPCs from [7] (an MAM VDPC) and [9] (an AMM VDPC) employ bit-slicing of input/weight parameters. They slice each 8-bit integer input/weigh parameter into two slices of 4-bit each. Then, they employ two VDPCs in parallel; each VDPC processes one 4-bit slice of the input/weight parameters. The corresponding 4-bit VDP results from these two 4-bit VDPCs are then combined using shifters and adders to produce the final 8-bit results. Thus, performing VDP operations using bit slices reduces the total number of VDP results that can be produced by a fixed number of VDPCs, because multiple VDPCs are needed to produce a single set of VDP results. This can severely degrade the throughput of such VDPCs. Such undesired outcome should be avoided by designing a more efficient VDPC.

B. Need for Stochastic Computing

Table II reports the counts of kernel tensors according to their sizes *S* (*S*<=44 and *S*>44) for four modern CNN models. From Table II, more than 98% of the kernel tensors across all four CNNs have *S*>44, and thus, they require VDPEs with size *N*>44 to process their corresponding VDP operations. But, from Table I, the maximum achievable *N* for analog VDPCs at 4-bit precision (*B*=4-bit) is limited to 44; therefore, processing the VDP operations corresponding to more than 98% of kernel tensors that have *S*>44 would lead to high

psum reduction latency (see Section III-A). However, reducing this *psum* reduction latency in analog VDPCs is challenging, as they have a strong trade-off between *N* and *B*, and this is because the required number of analog optical power levels (i.e., 2^B) to support a specific *B* consumes a large part of the available dynamic range of optical power in analog VDPCs. To this end, the remaining dynamic range of optical power within the total allowable optical power budget restricts the supported *N* in analog VDPCs. This limitation can be addressed by performing VDP operations in the digital domain [18]. There is no need to support any analog optical power levels in the digital domain; therefore, most of the available dynamic range of optical power in a digital VDPC can be used to support a higher *N*. But, the MRR-based binary digital VDPCs (e.g., [18], [33]) suffer from very high hardware complexity, and their multiply-accumulate and bit-shifting circuits consume huge area. These drawbacks motivate the need to examine new options for realizing optical digital VDPCs.

One such option is stochastic computing. In stochastic computing, multiplications can be replaced with simple bitwise AND operations [25]. This can be leveraged to perform point-wise multiplications between *DKV*s and *DIV*s (Section II-C) with less hardware complexity than binary digital VDPCs. In addition, since stochastic computing is also implemented in the digital domain (non-binary), a stochastic computing based optical VDPC can support a large *N* due to the large available dynamic range of optical power, just as discussed above for a binary digital VDPC. Moreover, a stochastic computing based optical VDPC can attain different precision levels by merely changing the number of bits in the stochastic bit-streams, without requiring different analog optical power levels. Therefore, to utilize these advantages of stochastic computing, prior works [34] and [35] proposed stochastic computing based photonic acceleration. [34] reports acceleration of Markov Random Field Inference and [35] employs photonic crystals and MZIs to build an edge detection filter. However, none of the prior works have employed stochastic computing based photonic acceleration for neural network inference. To fill this gap, we invent an MRR-based optical stochastic multiplier (OSM) and employ multiple OSMs to forge a novel Stochastic Computing Optical Neural Network Accelerator (SCONNA). The following section discusses our SCONNA architecture.

IV. Our Proposed SCONNA Architecture

A. Overview of SCONNA VDPC

Fig. 4(a) illustrates the VDPC organization of our SCONNA architecture. Like the VDPCs of analog optical accelerators, a SCONNA VDPC also implements multiple VDP operations in parallel. For that, an array of total *N* single-wavelength laser diodes (LDs) are used, with each LD sourcing optical power of $P_{\lambda_i}^{in}$ amount at a distinct wavelength λ_i. The total power from all *N* LDs (λ_1 to λ_N) multiplexed into a single photonic waveguide through wavelength division multiplexing (WDM). These multiplexed wavelengths split into *M* input waveguide arms (IWAs). Every IWA receives *N*-wavelength optical power and guides it to a VDPE. There are a total of *M* IWAs and *M* VDPEs in the SCONNA VDPC (Fig. 4(a)).

Each VDPE consists of three components: (i) a cascade of *N* Optical Stochastic Multipliers (OSMs); (ii) a bank of filter MRRs; (iii) a Photo-Charge Accumulator (PCA) pair. Each OSM performs stochastic multiplication between an input bit-stream *I* (corresponding to a point in an *N*-point DIV) and weight bit-stream *W* (corresponding to a point in an *N*-point DKV). Each OSM receives its bit-streams *I* and *W* from its corresponding peripherals at a supported bitrate (*BR*). Bit-stream *W* provides a weight value along with a sign bit. Bit-stream *I* provides the RELU-activated output value from the previous CNN layer, without a sign bit as RELU has a non-negative output. The detailed design of *OSM*s and their peripherals is explained in Section IV-B. Each OSM performs a bit-wise logical AND operation between the *I* and *W* bit-streams to produce a resultant optical bit-stream that represents the stochastic multiplication between the *I* and *W* bit-streams. The *N* optical bit-streams from the cascade of *N* OSMs, with each bit-stream carrying a stochastic multiplication result, reach the bank of filter MRRs. In this bank, each filter MRR operates on a distinct optical bit-stream λ_i. Each filter MRR receives the sign bit from the peripheral *W* of its corresponding OSM (Fig. 4(a)). The sign bit operates the filter to steer the incoming optical bit-stream λ_i to the output waveguide arm *OWA* (if the sign bit is '0') or *OWA'* (if the sign bit is '1'). Thus, the OWA and OWA' of a VDPE guide the optical bit-streams, carrying the stochastic multiplication results, to PCAs. A PCA is a circuit that collects all the optical bit-streams (i.e., stochastic multiplication results) from its corresponding OWA (or OWA') and generates the accumulation result in the binary format (details about PCA in Section IV-C). In a VDPE, the OWA-coupled PCA combines with the corresponding OWA'-coupled PCA to generate a signed accumulation result.

B. Optical Stochastic Multiplier

Our Optical Stochastic Multiplier (OSM) consists of peripherals and an Optical 'AND' Gate (OAG) (Fig. 5). The peripherals convert a binary input value I_b and binary weight value W_b into unipolar stochastic bit-streams *I* and *W*, and OAG performs multiplication-equivalent bitwise AND operation between the stochastic bit-streams *I* and *W*.

From Fig. 5, the peripherals of our OSM use a lookup table and serializers to generate a combination of unipolar stochastic bit-streams *I* and *W*. From [26], two unipolar stochastic bit-streams, for their eventual error-free multiplication using an AND gate, should be generated in combination with each other, so that they are uncorrelated, i.e., the marginal probability of one bit-stream (i.e., *I* or *W*) is equal to its conditional probability given the other bit-stream (i.e., *I* given *W* or *W* given *I*). For our OSM, we propose to generate all possible combinations of uncorrelated bit-streams *I* and *W* a priori (offline) using the unipolar circuit from [26], and then store these bit-streams in the bit-vector (bit-parallel) format in the lookup table (Fig. 5). As a result, each entry in the lookup table stores a combination of uncorrelated bit-vectors I_v and W_v. To index into this lookup table, our OSM creates a unique identifier for each combination of binary values I_b and W_b (that are accessed from a buffer (a scratchpad memory); Fig.

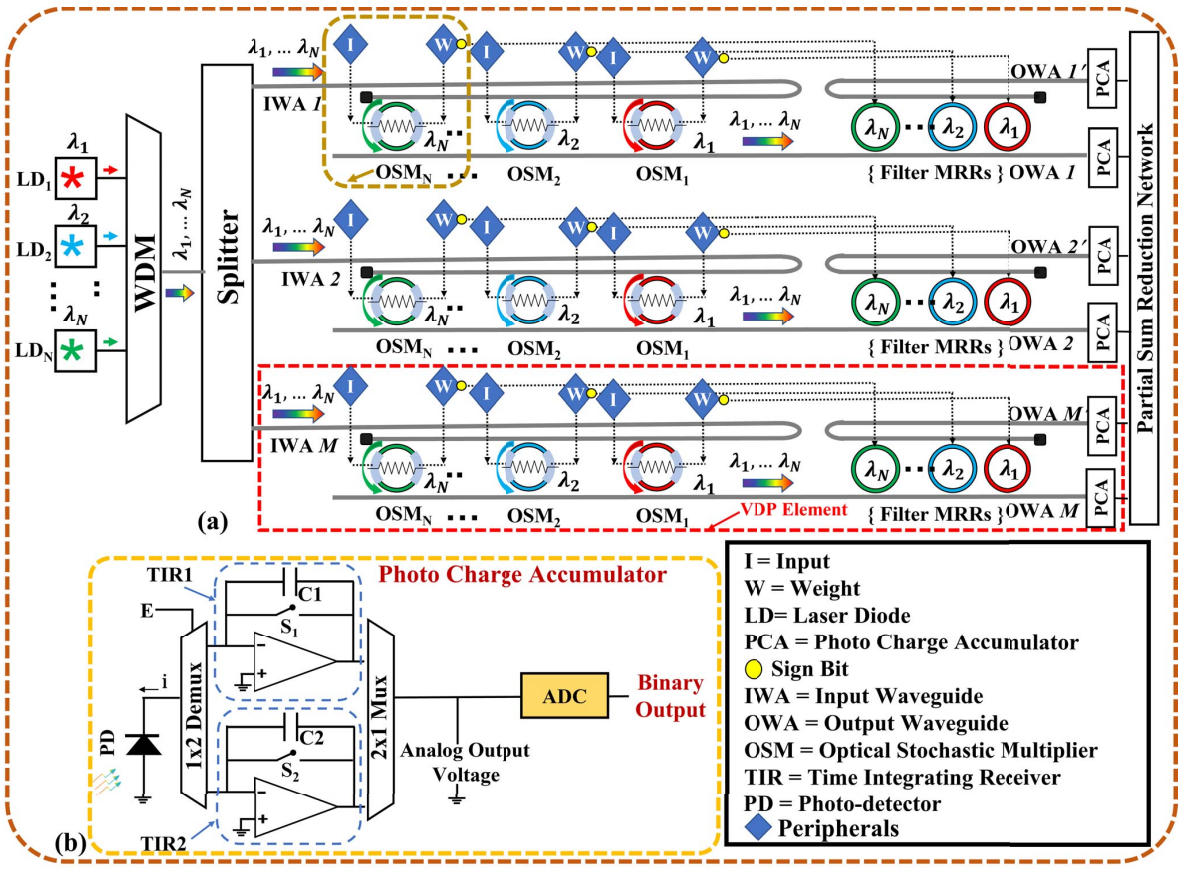

Fig. 4: Schematics of (a) Our SCONNA VDPC (b) Photo-Charge Accumulator (PCA) Circuit.

5) by performing an XOR-based hash function $I_b \oplus W_b$. Thus, our OSM uses a $I_b \oplus W_b$ value to fetch the desired combination of I_v and W_v from the lookup table. Then, it pushes these I_v and W_v through dedicated high-speed serializers, to generate bit-streams I and W. Lookup table size: If precision B=8-bit for binary I_b and W_b, there are 2^B entries in the lookup table, with each entry storing two 2^B-bits long bit-vectors.

The stochastic bit-streams I and W, generated by the peripherals of our OSM, are then fed to the OAG via high-speed drivers for their stochastic multiplication (Fig. 5). The design of our OAG is illustrated in Fig. 6(a). Our OAG is an add-drop microring resonator (MRR), which has two operand terminals (realized as embedded PN-junctions) that can take two stochastic bit-streams I and W (Fig. 6(a)) as inputs at a predefined bitrate (BR). Fig. 6(b) shows the passbands of the MRR for different operand inputs and temperature conditions. The MRR's temperature can be increased using the integrated microheater (Fig. 6(a)), to consequently tune its operand-independent resonance from its fabrication-defined initial position γ to its programmed position η, relative to the input optical wavelength position λ_{in} (Fig. 6(b)). For each bit combination at the operand terminals ((I,W) = (0,1), (1,0), or (1,1)), the MRR's resonance passband electro-refractively moves to an operand-driven position (red and blue passbands in Fig. 6(b)). Based on the MRR resonance passband's programmed position η relative to λ_{in}, the drop-port transmission ($T(\lambda_{in})$) of the MRR provides bit-wise logical AND operation between the inputs I and W.

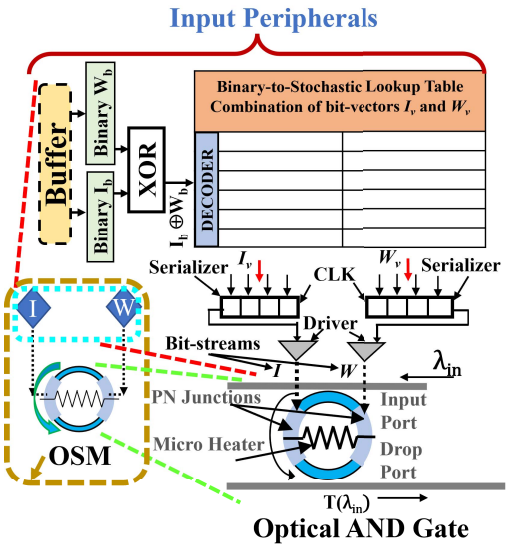

Fig. 5: Schematic of our Optical Stochastic Multiplier (OSM).

To validate our OAG, we performed transient analysis with two pseudo-random numbers as shown in Fig. 6(c). For that, we modelled and simulated our OAG using the

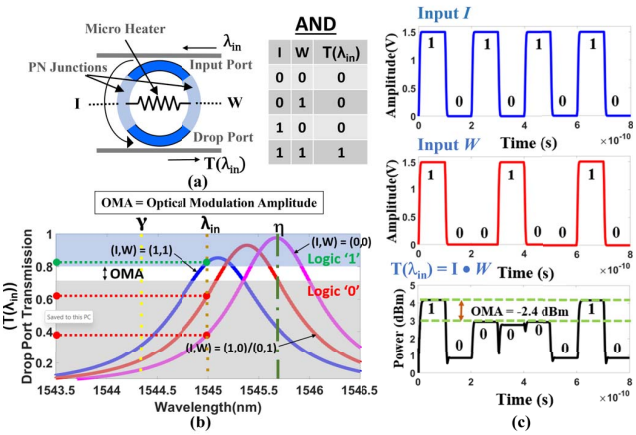

Fig. 6: (a) Schematic of our Optical AND Gate (OAG), (b) operation of OAG, (c) results of OAG's transient analysis.

foundry-validated tools from the Ansys/Lumerical's DEVICE, CHARGE, and INTERCONNECT suites [36]. Fig. 6(c) shows two input bit-streams (*I*, *W*) applied to the two PN junctions of our OAG at BR=10 Gbps. By looking at the output optical bit-stream T(λ_{in}) in Fig. 6 (c), we can say T(λ_{in}) = *I* AND *W*, which validates the functionality of our OAG as a logical AND gate. Thus, since the input bit-streams *I* and *W* are in the unipolar stochastic format, the output optical bit-stream at the drop port of the OAG provides the unipolar stochastic multiplication between *I* and *W*.

C. Photo Charge Accumulator (PCA)

From Section IV-A, the stochastic multiplication bit-streams generated by OSMs are guided to a PCA, where they are accumulated to generate a binary output value equivalent to the VDP result. Our PCA is inspired from the time integrating receiver (TIR) design from [37] and the photodetector-based optical-pulse accumulator design from [38]. A PCA circuit, shown in Fig 4(c), has two stages: (i) a stochastic-to-analog conversion stage; (ii) an analog-to-binary conversion stage. The stochastic-to-analog stage employs a photodetector and two TIR circuits (one TIR circuit remains redundant, enabled by the demux and mux; Fig 4(b)). The photodetector generates a current pulse for each optical logic '1' incident upon it. This current pulse accumulates a certain amount of charge on the capacitor of the active TIR circuit (e.g., the circuit with C1 capacitor); as a result, the capacitor accrues an analog voltage level. Hence, when one or more output optical bit-streams are incident upon the photodetector, the total accumulated charge (and thus, the accrued analog voltage level) on the active capacitor (e.g., C1) is proportional to the total number of '1's in the incident bit-streams. The number of 1's that can be accumulated in such manner might be limited, as the charge across the capacitor of TIR circuit (Fig. 4(b)) might saturate (this is further analysed in section V-C). Once the TIR output saturates, a discharge of the active capacitor (e.g., C1) is needed to prepare the circuit for the next accumulation phase. While capacitor C1 is discharging, capacitor C2 of the redundant TIR circuit mitigates the discharge latency by allowing a continuation of a concurrent accumulation phase. The output analog voltage computed by the stochastic-to-analog conversion stage represents the unipolar unscaled addition [26] of the stochastic bit-streams. To convert this analog voltage into a binary value, the analog-to-binary stage of the PCA circuit employs an analog-to-digital converter (ADC). This binary value is the VDP result.

V. Scalability Analysis of SCONNA Architecture

To understand the scalability of our SCONNA architecture, in this section, we analyze the achievable operating speed of the OSMs, achievable size *N* of the SCONNA VDPC, and the accumulation capacity of the PCA circuits.

A. Operating Speed and Latency Overhead of OSM

The peripherals of an OSM can incur some latency for accessing the scratchpad buffer and eDRAM-based lookup table. We consider 2ns latency each for the scratchpad buffer [39] and eDRAM-based lookup table [40]. Beyond this latency, the speed of an OSM depends on the achievable operating speed (bit-rate (BR)) of the constituent OAG. Analysis of OAG's BR: For the output optical bit-stream T(λ_{in}) in Fig. 6(c), the optical modulation amplitude (OMA) is the output power difference between the highest logic '0' power level and the lowest logic '1' power level. OMA should be at least equal to or greater than the sensitivity of the photodetector in the PCA circuit, to ensure that the photodetector in the PCA circuit can produce a distinguishably higher-amplitude current pulse for an optical logic '1' bit compared to an optical logic '0' bit. Keeping the OMA to be greater than or equal to the given photodetector sensitivity (P_{PD-opt}=-28dBm; Section V-B) depends on the OAG's BR and FWHM (full passband width at half maximum). Therefore, to analyze this dependency, we simulated BR and FWHM duplets for which OMA = -28 dBm, as shown in Fig. 7(a). As evident, supported BR increases as FWHM increases. However, at (FWHM≈0.8nm), BR saturates at 40 Gbps. Therefore, we aim to operate our OAG at BR<=40Gbps for FWHM<=0.8nm.

B. Achievable Size of SCONNA VDPC

We consider optimistic free-spectral range (FSR) of 50 nm [19] for the constituent MRR-based OAGs of our SCONNA VDPC. In addition, we consider the inter-wavelength gap of 0.25 nm. This allows the *N* for our SCONNA VDPC to be 200 (=FSR/0.25nm), theoretically. However, even if we consider FSR=50nm to be practically achievable for our OAGs, achieving *N*=200 for our SCONNA VDPC might not be possible in practice. This is because when we aim to operate our OAGs at a high BR of <= 40 Gbps, for FWHM <= 0.8 nm, the total power penalty for our SCONNA VDPC might increase significantly owing to the increased impacts of optical crosstalk effects at OSMs, signal truncation at MRR filters, and BR-dependent increase in the photodetector sensitivity [41]–[43]. This increase in power penalty can reduce *N* to be less than 200. Therefore, to determine the achievable *N* for our SCONNA VDPC at *B*=8-bit precision, we adopt the scalability

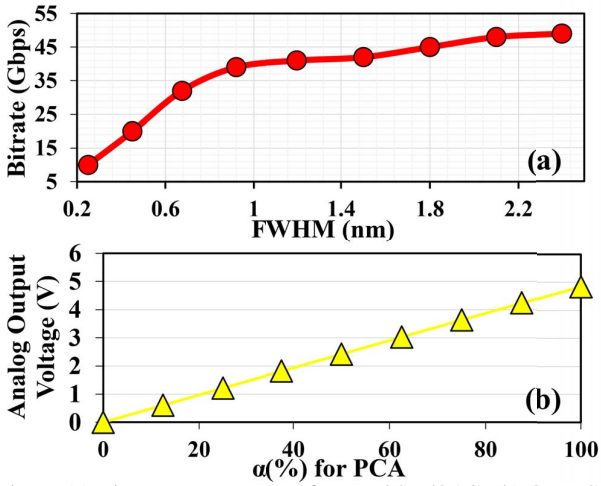

Fig. 7: (a) Bitrate versus FWHM for our OSM/OAG, (b) Our PCA's analog output voltage versus α

.

analysis equations (Eq. 2, Eq. 3, and Eq. 4) from [19], [21]. Table III reports the definitions of the parameters and their values used in these equations. Since our SCONNA VDPC processes stochastic bit-streams, which are digital in format, it requires the bit resolution of B_{Res} = 1-bit in the equations. Moreover, we conservatively choose to operate OSMs/OAGs at BR=30Gbps. We consider *M*=*N*. We first solve Eq. 2 and Eq. 3 for datarate (DR)=BR$*2^B$, to find P_{PD-opt} to be -28 dBm. Then, we solve Eq. 4 for *N*, to find *N*=*M*=176, which is a very large *N* compared to analog VDPCs that have *N*<=44. Such large *N* significantly improves the overall throughput and energy efficiency (Section VI).

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

$$\beta = \sqrt{2q(RP_{PD-opt} + I_d) + \frac{4kT}{R_L} + R^2P^2_{PD-opt}RIN} \quad (3)$$

$$P_{Laser} = \frac{10^{\frac{\eta_{WG(dB)}[N(d_{OSM})]}{10}}M}{\eta_{SMF}\eta_{EC}IL_{i/p-OSM}} \times \frac{P_{PD-opt}}{\eta_{WPE}IL_{MRR}} \times \frac{1}{(OBL_{OSM})^{N-1}(EL_{splitter})^{log_2M}} \times \frac{1}{(OBL_{MRR})^{N-1}(IL_{penalty})} \quad (4)$$

C. Accumulation Capacity and Error Susceptibility of PCA

From Section V-B, our SCONNA VDPC has *N*=176. For precision *B*=8, each optical bit-stream in a SCONNA VDPC has 2^B=256 bits. Therefore, each PCA in a SCONNA VDPC needs to be able to accumulate a total $N \times 2^B$=176×256 optical '1' bits, at the least. We modeled the photodetector of our PCA circuit using the INTERCONNECT tool from Ansys/Lumerical [36] for R_{PD}=1.2 A/W and P_{PD-opt}=-28 dBm, and extracted the current pulse values corresponding optical '1's and '0's that are consumed by the photodetector. We then imported these values in our MultiSim [44] based model of the TIR circuit of the PCA, to find out that our PCA should have R=50Ω, C=250pF, and voltage amplifier gain=80. For these parameters, we simulated to the analog output voltage at the PCA using MultiSim [44] for different valus of α=(actual # of '1's in incident bit-streams/176×256)×100%. The results are shown in Fig. 7(b). As evident, the analog output voltage increases linearly with α without saturating at α=100%. This outcome shows that our PCA can efficiently support the accumulation of *N*=176 bit-streams. Note that the analog output voltage from the amplifier of the PCA circuit does not incur any errors in computation. But, the ADC introduces errors in the generated binary result (we evaluate mean absolute percentage error to be 1.3% for the ADC), and we later evaluate the impact of these errors on the CNN inference accuracy (Section VI).

VI. System-Level Implementation and Evaluation

A. System-Level Implementation of SCONNA

Fig. 8 illustrates the system-level implementation of our SCONNA accelerator. It consists of global memory for storing CNN parameters, and a preprocessing and mapping unit for decomposing the tensors into DIVs/DKVs and mapping them onto VDPEs. It has a mesh of tiles connected to routers, and this mesh network facilitates parameter communication among tiles. Each tile consists of 4 SCONNA VDPCs interconnected (via H-tree network) with output buffer, activation, and pooling units. In addition, each tile also contains a *psum* reduction network.

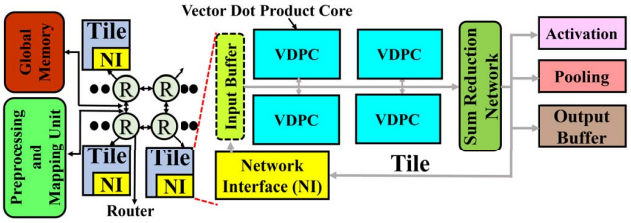

Fig. 8: System-level overview of our SCONNA CNN accelerator.

B. Simulation Setup

For evaluation, we model our SCONNA accelerator from Fig. 8 using our developed custom, transaction-level, event-driven python-based simulator (https://github.com/uky-UCAT/SC_ONN_SIM.git). Using the simulator, we simulated the inference four CNN models (with batch size of 1): GoogleNet [50], ResNet50 [22], MobileNet_V2 [51], and ShuffleNet_V2 [52]. We evaluate the metrics such as Frames per second (FPS), FPS/W (energy efficiency) and FPS/W/mm^2 (area efficiency). We also evaluate the impact of PCA error on Top-1 and Top-5 inference accuracy of the CNN models for ImageNet validation dataset [53].

We compared our accelerator with the analog optical accelerators AMM (DEAPCNN [9]) and MAM (HOLYLIGHT [7]) at 8-bits integer quantization for CNN inference. We omitted comparison with CMOS-based digital CNN accelerators as

TABLE III: List of abbreviations and their full forms used in this paper. Definition and values of various parameters (obtained from [19]) used in Eq. 2, Eq. 3, and Eq. 4 for the scalability analysis of our SCONNA VDPCs.

Abbreviations	Full form	Parameter	Definition	Value
VDPC	Vector Dot Processing Core	P_{Laser}	Laser Power Intensity	10 dBm
PCA	Photo Charge Accumulator	R_{PD}	PD Responsivity	1.2 A/W
OAG	Optical AND Gate	R_L	Load Resistance	50 Ω
SE	Summation Element	I_d	Dark Current	35 nA
SC	Stochastic Computing	T	Absolute Temperature	300 K
DKV	Decomposed Kernel Vector	BR	Bitrate	30 Gbps
DIV	Decomposed Input Vector	RIN	Relative Intensity Noise	-140 dB/Hz
VDP	Vector Dot Product	η_{WPE}	Wall Plug Efficiency	0.1
S	Size of DKV	IL_{SMF}(dB)	Single Mode Fiber Insertion Loss	0
psum	Partial Sum	IL_{EC}(dB)	Fiber to Chip Coupling Insertion Loss	1.6
OSM	Optical Stochastic Multiplier	IL_{WG}(dB/mm)	Silicon Waveguide Insertion Loss	0.3
DR	Data rate	$EL_{Splitter}$(dB)	Splitter Insertion Loss	0.01
VDPE	Vector Dot Product Element	IL_{OSM}(dB)	Optical Stochastic Multiplier (OSM) Insertion Loss	4
N	Size of VDPE	OBL_{OSM}(dB)	Out of Band Loss Optical Stochastic Multiplier	0.01
M	Number of VDPEs per VDPC Unit	IL_{MRR}(dB)	Microring Resonator(MRR) Insertion Loss	0.01
OMA	Optical Modulation Amplitude	$IL_{penalty}$(dB)	Network Penalty	7.3
B	Binary Bit Precision	d_{OSM}	Gap between two adjacent OSMs	20 μm
B_{Res}	Bit Resolution	P_{PD-opt}	Output Photodetector Sensitivity	-

TABLE IV: Peripherals Parameters for Accelerators [6].

	Power (mW)	Area (mm^2)	Latency
Reduction Network	0.05	3.00E-05	3.125ns
Activation Unit	0.52	6.00E-04	0.78ns
IO Interface	140.18	2.44E-02	0.78ns
Pooling Unit	0.4	2.40E-04	3.125ns
eDRAM	41.1	1.66E-01	1.56ns
Bus	7	9.00E-03	5 cycles
Router	42	0.151	2 cycles
AMM/MAM			
DAC [45]	30	0.034	0.78ns
ADC [46]	29	0.103	0.78ns
SCONNA			
ADC [47]	2.55	0.002	0.78ns
Serializer per OSM [48]	5	5.9	0.03ns
LUT per OSM [49]	0.06	0.09	2ns
PCA [44]	0.02	0.28	-

prior analog optical photonic CNN accelerators have outperformed them [9], [12]. We simulate analog optical accelerators for 5 GS/s [31] and from Section III-A, at *B*=4-bit precision, we set *N=16* for AMM (DEAPCNN), and *N=22* for MAM (HOLYLIGHT). Prior works, AMM (DEAPCNN) and MAM (HOLYLIGHT) employ weight stationary dataflow, therefore our evaluation is based on weight stationary dataflow. For fair comparison, we perform area proportionate analysis. In the area proportionate analysis, we altered the VDPE count of each analog optical accelerator, across all of the accelerator's VDPCs, to match with the area of the SCONNA accelerator having 1024 VDPEs. The scaled VDPE count of MAM (HOLYLIGHT) and AMM (DEAPCNN) are 3971 and 3172, respectively.

Table IV gives the parameters used for evaluating the overhead of the peripherals in our evaluated accelerators. We consider each laser diode to emit input optical power of 10 mW (10 dBm) (Table III) [9], multiplexer and splitter parameters are taken from [7].

C. Evaluation Results

Fig. 9(a) compares the FPS values (log scale) achieved by each accelerator across various CNNs. SCONNA significantly outperforms the analog optical accelerators MAM (HOLYLIGHT) and AMM (DEAPCNN) by 66.5× and 146.4×, respectively, on gmean across the CNNs. These benefits are mainly associated with the superior *N* and higher BR of SCONNA compared to the analog optical accelerators. Because of the high *N*, SCONNA requires less number of *psums* for DKVs with *S>44* (refer Table II), while generating the final VDP result. The reduced *psum*s drastically reduces the *psum* reduction latency. The higher operating BR=30Gbps compensates for the lengthy stochastic bit-streams of 2^B=256 bits used by SCONNA. The improvements for SCONNA are more evident for large CNNs such as GoogleNet [50] and ResNet50 [22] compared to smaller CNNs such as MobileNet_V2 [51] and ShuffleNet_V2 [52]. This is due to the fact that MobileNet_V2 [51] and ShuffleNet_V2 [52] employ depthwise separable convolutions which use DKVs with S<44 more frequently than large CNNs. Overall, SCONNA gives exceedingly better FPS compared to the analog optical accelerators.

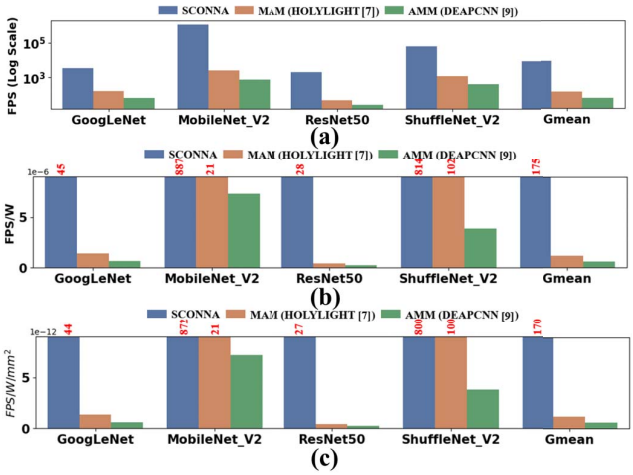

Fig. 9: (a) FPS (Log Scale) (b) FPS/W (c) FPS/W/mm^2 for SCONNA versus MAM and AMM accelerators for *B*=8-bits.

Fig. 9(b) gives the energy efficiency (FPS/W) values for each accelerator across various CNNs. It is evident that SCONNA attains substantially better energy efficiency than the analog optical accelerators. Our SCONNA gains 90× and 183× better FPS/W against analog MAM (HOLY-

TABLE V: Top-1 and Top-5 inference accuracy comparison of SCONNA versus MAM for 8-bit quantized CNNs {GoogleNet (GNet), ResNet50 (RNet50), MobileNet_V2 (MNet_V2), ShuffleNet_V2 (SNet_V2)} and ImageNet dataset [53].

SCONNA ACCURACY DROP (%)	GNet [50]	RNet [22]	MNet_V2 [51]	SNet_V2 [52]	Gmean
TOP-1	0.1	0.4	1.5	0.5	0.4
TOP-5	0.1	0.3	0.7	0.4	0.3

LIGHT) and AMM (DEAPCNN), respectively, on gmean across the CNNs. These energy efficiency benefits are due to the improved throughput and flexible precision support of SCONNA VDPCs. The analog MAM (HOLYLIGHT) and AMM (DEAPCNN), due to their limited 4-bit precision, employ two VDPEs to attain an 8-bit precision using bit-slicing. This decreases the throughput and also increases the energy consumption compared to SCONNA VDPCs. In addition, during area proportionate analysis, MAM (HOLYLIGHT) and AMM (DEAPCNN) get scaled to large VDPE counts (3971 and 3172), leading to overall higher static power consumption compared to SCONNA. Therefore, SCONNA achieves better energy efficiency compared to all the other tested accelerators.

Fig. 9(c) shows the area efficiency values (FPS/W/mm^2) for each accelerator across various CNNs. The area efficiency results look similar to energy efficiency as we match the area of all the accelerators to SCONNA (for the area proportionate analysis). SCONNA gains 91× and 184× better FPS/W/mm^2 against analog MAM (HOLYLIGHT) and AMM (DEAPCNN), respectively, on gmean across the CNNs. Overall, SCONNA significantly improves the throughput, energy efficiency and area efficiency compared to the tested analog optical accelerators.

D. Inference Accuracy Results

As discussed in Section IV-C, the ADC in the PCA circuits of our SCONNA VDPCs incurs the mean absolute percentage error of 1.3% on the computed binary results. To evaluate the impact of these errors on the CNN inference accuracy, we simulated the inference of four CNNs on SCONNA and analog optical accelerator MAM (HOLYLIGHT). We integrated our custom simulator with ML-framework PyTorch [54] and performed the inference using ImageNet validation dataset [55] (50k images and 1k classes). Table V reports the Top-1 and Top-5 inference accuracies obtained for our SCONNA and MAM for four 8-bit integer-quantized CNNs. As evident, SCONNA yields Top-1 and Top-5 accuracy drop of only 0.4% and 0.3%, respectively, on gmean across the tested CNNs. The large CNN models ResNet50 [22] and GoogelNet [50] have more tolerance to the errors, and hence, they show minimal to zero drop in accuracy for SCONNA. Furthermore, SCONNA's accuracy drop can be improved by performing stochastic computing aware training of the CNN models on SCONNA [56]. Our SCONNA accelerator's significant gains in the FPS, FPS/W, and FPS/W/mm^2, overshadows the minor drop in the CNN inference accuracy.

VII. Related Work on Optical CNN Accelerators

To accelerate CNN inferences with low latency and low energy consumption, prior works proposed various accelerators based on photonic integrated circuits (PICs) (e.g., [7], [11]–[14]). These accelerators employ PIC-based Vector Dot Product Cores (VDPCs) to perform multiple parallel VDP operations. Some accelerators implement digital VDPCs (e.g., [18], [31]), whereas some others employ analog VDPCs (e.g., [7], [9], [12], [17]). Different VDPC implementations employ MRRs (e.g., [7], [9], [12], [57], [58]) or MZIs (e.g., [13]–[15]) or both (e.g., [18], [31]). The analog VDPCs can be further classified as incoherent (e.g., [7], [9], [12]) or coherent (e.g., [59]–[64]). To set and update the values of the individual input and kernel tensors used for vector dot product operations, the incoherent VDPCs utilize the analog optical signal power, whereas the coherent VDPCs utilize the electrical field amplitude and phase. The coherent VDPCs achieve low inference latency, but they suffer from control complexity, high area overhead, low scalability, low flexibility, high encoding noise, and phase error accumulation issues [65]. In contrast, the MRRs-enabled incoherent VDPCs based accelerators achieve better scalability and lower footprint, because they use PICs that are based on compact MRRs [9], unlike the coherent VDPCs that use PICs based on bulky MZIs. Various state-of-the-art PIC-based optical CNN accelerators are well discussed in a survey paper [66]. Because of the inherent advantages of MRR-enabled incoherent VDPCs, there is impetus to design more energy-efficient and scalable CNN accelerators employing MRR-enabled incohorent VDPCs.

VIII. Conclusions

To mitigate the very strong scalability versus bit-precision trade-off innately present in analog optical CNN accelerators, we demonstrated a merger of stochastic computing and MRR-based CNN accelerators for the first time in this paper. We invented an MRR-based optical stochastic multiplier (OSM) and employed multiple OSMs to forge a novel stochastic computing based CNN accelerator called SCONNA. Our evaluation results for four CNN models show that SCONNA provides improvements of up to 66.5×, 90×, and 91× in throughput, energy efficiency, and area efficiency, respectively, compared to two analog optical accelerators AMM and MAM, with Top-1 accuracy drop of only up to 0.4% for large CNNs and up to 1.5% for small CNNs.

Acknowledgments

We thank the anonymous reviewers whose valuable feedback helped us improve this paper. We would also like to acknowledge the National Science Foundation (NSF) as this research was supported by NSF under grant CNS-2139167.

References

[1] Y. LeCun *et al.*, "Deep learning," *Nature*, 2015.
[2] W. Liu *et al.*, "A survey of deep neural network architectures and their applications," *Neurocomputing*, 2017.
[3] Z. Li *et al.*, "A survey of convolutional neural networks: Analysis, applications, and prospects," *CoRR*, 2020. [Online]. Available: https://arxiv.org/abs/2004.02806

[4] X. Xu *et al.*, "Scaling for edge inference of deep neural networks," *Nature Electronics*, 2018.
[5] L. Baischer *et al.*, "Learning on hardware: A tutorial on neural network accelerators and co-processors," 2021.
[6] R. Krishnamoorthi, "Quantizing deep convolutional networks for efficient inference: A whitepaper," *arXiv preprint arXiv:1806.08342*, 2018.
[7] W. Liu *et al.*, "Holylight: A nanophotonic accelerator for deep learning in data centers," in *DATE*, 2019.
[8] J. Gu *et al.*, "Squeezelight: Towards scalable optical neural networks with multi-operand ring resonators," in *DATE*, 2021.
[9] V. Bangari *et al.*, "Digital electronics and analog photonics for convolutional neural networks (DEAP-CNNs)," *JSTQE*, 2020.
[10] Q. Cheng *et al.*, "Silicon photonics codesign for deep learning," *Proceedings of the IEEE*, 2020.
[11] L. Yang *et al.*, "On-chip optical matrix-vector multiplier," in *Optics and Photonics for Information Processing*. SPIE, 2013.
[12] F. Sunny *et al.*, "Crosslight: A cross-layer optimized silicon photonic neural network accelerator," *CoRR*, 2021.
[13] H. Bagherian *et al.*, "On-chip optical convolutional neural networks," *CoRR*, vol. abs/1808.03303, 2018.
[14] H. Zhang *et al.*, "An optical neural chip for implementing complex-valued neural network," *Nature Communications*, 2021.
[15] C. Demirkiran *et al.*, "An electro-photonic system for accelerating deep neural networks," *CoRR*, 2021.
[16] A. N. Tait *et al.*, "Broadcast and weight: An integrated network for scalable photonic spike processing," *J. Lightwave Technol.*, 2014.
[17] P. R. Prucnal *et al.*, *Neuromorphic Photonics*. CRC Press, 2017.
[18] K. Shiflett *et al.*, "Pixel: Photonic neural network accelerator," in *HPCA*, 2020.
[19] M. A. Al-Qadasi *et al.*, "Scaling up silicon photonic based accelerators: Challenges and opportunities," *APL Photonics*, vol. 7, no. 2, p. 020902, 2022. [Online]. Available: https://doi.org/10.1063/5.0070992
[20] C. Huang *et al.*, "Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits," *APL Photonics*, Apr. 2020.
[21] V. S. Sri and T. I. G, "Photonic reconfigurable accelerators for efficient inference of cnns with mixed-sized tensors," *TCAD*, 2022. [Online]. Available: https://arxiv.org/abs/2207.05278
[22] H. Kaiming *et al.*, "Deep residual learning for image recognition," *CoRR*, 2015.
[23] B. R. Gaines, *Stochastic Computing Systems*, 1969.
[24] A. Alaghi *et al.*, "Survey of stochastic computing," *ACM Trans. Embed. Comput. Syst.*, 2013.
[25] A. Armin *et al.*, "The promise and challenge of stochastic computing," *TCAD*, 2018.
[26] D. Wu *et al.*, "Ugemm: Unary computing architecture for gemm applications," in *ISCA*, 2020.
[27] A. Ren *et al.*, "Sc-dcnn: Highly-scalable deep convolutional neural network using stochastic computing," ser. ASPLOS '17. New York, NY, USA: ACM, 2017.
[28] J. Li *et al.*, "Towards acceleration of deep convolutional neural networks using stochastic computing," in *ASP-DAC*, 2017.
[29] Z. Li *et al.*, "Dscnn: Hardware-oriented optimization for stochastic computing based deep convolutional neural networks," in *ICCD*, 2016.
[30] P. Li *et al.*, "Computation on stochastic bit streams digital image processing case studies," *TVLSI*, 2014.
[31] S. Kyle *et al.*, "Albireo: Energy-efficient acceleration of convolutional neural networks via silicon photonics," in *ISCA*, 2021.
[32] F. Chollet *et al.*, "Keras," https://keras.io/api/applications/, 2015.
[33] V. S. P. Karempudi *et al.*, "Design exploration and scalability analysis of a cmos-integrated, polymorphic, nanophotonic arithmetic-logic unit," in *CENSS*, 2021.
[34] X. Zhang *et al.*, "Architecting a stochastic computing unit with molecular optical devices," in *ISCA*, 2018.
[35] H. El-Derhalli *et al.*, "Towards all-optical stochastic computing using photonic crystal nanocavities," *JETC*, 2021.
[36] "Pic design and simulation software - lumerical interconnect," Apr 2021. [Online]. Available: https://www.lumerical.com/products/interconnect/
[37] A. Sludds *et al.*, "Delocalized photonic deep learning on the internet's edge," *Science*, 2022.
[38] F. Brückerhoff-Plückelmann *et al.*, "A large scale photonic matrix processor enabled by charge accumulation," *Nanophotonics*, 2022.
[39] R. Balasubramonian, A. B. Kahng, N. Muralimanohar, A. Shafiee, and V. Srinivas, "Cacti 7: New tools for interconnect exploration in innovative off-chip memories," *ACM Trans. Archit. Code Optim.*, 2017.
[40] J. Lira *et al.*, "Implementing a hybrid sram / edram nuca architecture," in *ICHPC*, 2011.
[41] B. Meisam *et al.*, "Crosstalk penalty in microring-based silicon photonic interconnect systems," *JLT*, 2016.
[42] M. Bahadori *et al.*, "Comprehensive design space exploration of silicon photonic interconnects," *JLT*, 2016.
[43] B. Meisam *et al.*, "Energy-performance optimized design of silicon photonic interconnection networks for high-performance computing," in *DATE*, 2017.
[44] "Multisim." [Online]. Available: https://www.ni.com/en-us/shop/software/products/multisim.html
[45] F. N. U. Juanda *et al.*, "A 10-gs/s 4-bit single-core digital-to-analog converter for cognitive ultrawidebands," *TCS*, 2017.
[46] M. Guo *et al.*, "A 29mw 5gs/s time-interleaved sar adc achieving 48.5db sndr with fully-digital timing-skew calibration based on digital-mixing," in *VLSI Circuits*, 2019.
[47] D.-R. Oh *et al.*, "An 8b 1gs/s 2.55mw sar-flash adc with complementary dynamic amplifiers," in *IVLSIC*, 2020.
[48] C. Sun *et al.*, "A 45 nm cmos-soi monolithic photonics platform with bit-statistics-based resonant microring thermal tuning," *IJSSC*, 2016.
[49] H. Ye *et al.*, "Double-gate w-doped amorphous indium oxide transistors for monolithic 3d capacitorless gain cell edram," in *IEDM*, 2020.
[50] C. Szegedy *et al.*, "Going deeper with convolutions," *CoRR*, 2014.
[51] M. Sandler *et al.*, "Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation," *CoRR*, 2018.
[52] X. Zhang *et al.*, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," *CoRR*, 2017.
[53] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.
[54] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, 2019.
[55] A. D. Team, "Hub: A dataset format for ai. a simple api for creating, storing, collaborating on ai datasets of any size and streaming them to ml frameworks at scale." *GitHub. Note: https://github.com/activeloopai/Hub*, 2022.
[56] H. Benmeziane *et al.*, "A comprehensive survey on hardware-aware neural architecture search," *arXiv preprint arXiv:2101.09336*, 2021.
[57] P. Y. Ma *et al.*, "Photonic independent component analysis using an on-chip microring weight bank," *Optics Express*, 2020.
[58] J. Feldmann *et al.*, "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, 2021.
[59] R. Hamerly *et al.*, "Large-scale optical neural networks based on photoelectric multiplication," *Phys. Rev. X*, 2019.
[60] Z. Zhao *et al.*, "Hardware-software co-design of slimmed optical neural networks," ser. ASPDAC '19, 2019.
[61] X. Xu *et al.*, "11 TOPS photonic convolutional accelerator for optical neural networks," *Nature*, 2021.
[62] X. Zhao *et al.*, "An integrated optical neural network chip based on mach-zehnder interferometers," 2018.
[63] X. Lin and Others, "All-optical machine learning using diffractive deep neural networks," *Science*, 2018.
[64] T. Zhou *et al.*, "Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit," *Nature Photonics*, 2021.
[65] Mourgias-Alexandris *et al.*, "Neuromorphic photonics with coherent linear neurons using dual-iq modulation cells," *JLT*, 2020.
[66] L. De Marinis *et al.*, "Photonic neural networks: A survey," *IEEE Access*, 2019.