OISA: Architecting an Optical In-Sensor Accelerator for Efficient Visual Computing

Mehrdad Morsali[†], Sepehr Tabrizchi[‡], Deniz Najafi[†], Mohsen Imani[§],
Mahdi Nikdast[*], Arman Roohi[‡], and Shaahin Angizi[†]

[†]Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ, USA
[‡]School of Computing, University of Nebraska–Lincoln, Lincoln, NE, USA
[§]Department of Computer Science, University of California Irvine, Irvine, CA, USA
[*]Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO, USA
m.imani@uci.edu, mahdi.nikdast@colostate.edu, aroohi@unl.edu, shaahin.angizi@njit.edu

***Abstract*—Targeting vision applications at the edge, in this work, we systematically explore and propose a high-performance and energy-efficient Optical In-Sensor Accelerator architecture called OISA for the first time. Taking advantage of the promising efficiency of photonic devices, the OISA intrinsically implements a coarse-grained convolution operation on the input frames in an innovative minimum-conversion fashion in low-bit-width neural networks. Such a design remarkably reduces the power consumption of data conversion, transmission, and processing in the conventional cloud-centric architecture as well as recently-presented edge accelerators. Our device-to-architecture simulation results on various image data-sets demonstrate acceptable accuracy while OISA achieves 6.68 TOp/s/W efficiency. OISA reduces power consumption by a factor of 7.9 and 18.4 on average compared with existing electronic in-/near-sensor and ASIC accelerators.**

I. Introduction

While the Internet of Things (IoT) has become ubiquitous, it still lacks inherent intelligence and heavily depends on cloud-based decision-making. In such a cloud-centric scenario, a substantial part of data generated by IoT's sensors is left unprocessed or unanalyzed [1], [2]. Vision sensors typically convert light into electrical signals, which are then saved, processed, transmitted, and utilized. This involves converting all pixels into predetermined digital values with a constant bit depth, such as 8 bits [2], [3]. Reportedly, the majority of power consumption (over 96% [2], [3]) in traditional vision sensors comes from pixel value conversion and storage. This is primarily due to the memory and compute-intensive computing algorithm and the lack of processing capabilities of current IoT devices restricted by power and area factors [4], [5]. To tackle these challenges, a shift from a cloud-centric to a thing-centric (data-centric) approach is required, where the IoT node processes the sensed data locally [6].

Recently, there has been research into developing smarter CMOS image sensors that can accelerate Deep Neural Network (DNN) workloads. One method is to integrate CMOS image sensors and processors on a single chip, referred to as Processing-Near-Sensor (PNS) [7]–[11]. Another approach involves integrating computation units with individual pixels called Processing-In-Sensor (PIS). The PIS platform [2], [3], [12], [13] processes pre-Analog-to-Digital Converter (pre-ADC) data before transmitting it to the on/off-chip processor [1], [14]. However, they still suffer from energy-hungry ADC, DAC, and sense amplifiers [2], [3]. Due to the limited resources of PIS, it has not been feasible to deploy all DNN layers into the pixel array. Therefore, most studies have focused on accelerating the first layer in an analog or digital domain and submitting the remaining layers to a digital accelerator. Nevertheless, three significant challenges have yet to be addressed in current PIS/PNS designs *(i) Current designs still suffer from energy-hungry peripherals and ADC/DAC units even reduced [2], [6], [15] for sensing and computing; (ii) the in-/near-sensor computation imposes a large area overhead and power consumption in more recent PNS/PIS units and typically requires extra memory for intermediate data storage [1], [3], [13], [16]; and (iii) the computation speed has been constrained by the electronic systems (operating at a few GHz) that inherently lack the capability to support both high speeds and the extensive parallelism found in optical systems approaching the photodetection rate (>100GHz) [17]–[19].*

While silicon photonics has already established its efficacy in enabling high-throughput communication and computation across various domains [17], [18], in this work, we systematically explore the potential of deploying it on edge devices. We propose an Optical In-Sensor Accelerator architecture named OISA that leverages the energy efficiency and low latency features of photonic devices and minimizes signal conversion in low-bit-width neural networks to eliminate the need for power-hungry ADC and DACs. Our novel contributions to this work are as follows. (1) For the first time, we develop an optical in-sensor architecture that has been optimized to efficiently process the 1^{st} layer of DNNs with a centralized kernel-based optical processing unit tuned with microring resonator optical devices, resulting in improved energy-efficiency and speed; (2) We design new microarchitectural and circuit-level schemes for OISA supported by novel hardware partitioning and mapping mechanisms; and (3) We create a bottom-up device-to-architecture evaluation framework and extensively analyze and compare the performance of the proposed designs with prior PIS and ASIC designs.

II. Background

In-Sensor Accelerators. Boosting throughput and intensifying computation on resource-limited PIS/PNS devices result in elevated temperature, higher power consumption, and increased noise levels. These factors contribute to a decline in accuracy [2], [3], [20]. MACSEN [2] as a PIS platform processes the 1^{st}-convolutional layer of Binary CNN with the correlated double sampling procedure achieving 1000 FPS speed in computation mode. However, it suffers from humongous area-overhead and power consumption mainly due to the SRAM-based PIS

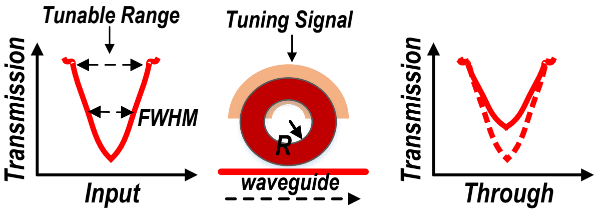


Fig. 1. MR input and through ports' spectra after imprinting a parameter onto the signal. A directional coupler transfers the light from the waveguide into the MR to be recombined. This recombination, influenced by the effective refractive index in the MR which is also impacted by the MR's circumference, induces a phase shift in the combined wave. This phase shift leads to interference with the original light's intensity. Tunable range shows the free resonance spectral range of the MR, where FMHW is the full width at half maximum of the resonance spectrum.

method. In [1], a PIS architecture is designed to support 8-bit activation and weight intended for the first-layer DNN acceleration through pulse modulation. Although the resource utilization of the design improved considerably, the use of power-hungry ADCs in each column, along with capacitors for direct sampling, increased overall energy consumption and area overhead. PISA [3] enables convolutional operations in the first BNN layer by leveraging non-volatile memory to store network weights. Notably, this architecture reduces the energy consumed on data conversion and transmission. However, the power-demanding write operations in non-volatile memories and the use of ADC for data transfer elevate the overall power consumption of the array. In [9], a PNS architecture was introduced, enhancing resolution while minimizing area overhead. However, PNS faces challenges in addressing the under-utilization problem of the first layer, leading to reduced accuracy. Additionally, the use of ADCs further raises power consumption across the entire array. AppCiP [13] implements a folded ADC to decrease comparator count, though the collective power consumption of the ADC units remains an issue. In [21], the PIS architecture leverages a combination of pixel current and charge-sharing events to reduce the power consumption of ADC to enable feature extraction and region-of-interest detection through current-domain MAC operations. However, the design is limited to row-wise computing rather than performing computations across the entire array.

Silicon Photonics Accelerators. Offering notably elevated operational bandwidth compared to electronic accelerators along with addressing fan-in/fan-out problems make silicon-photonic-based accelerators a promising candidate to accelerate DNN and machine vision applications [18], [22]. Such accelerators can be broadly categorized into two primary designs: *coherent* and *non-coherent* architectures. Within the coherent category, a single wavelength is employed for operations, and weight/activation parameters are incorporated into the electrical field amplitude, phase, or polarization of an optical signal [23]. Conversely, the noncoherent designs [17], [18] employ multiple wavelengths each of which capable of conducting computations concurrently. Within coherent architectures, considered in this work, the weight and input parameters are imprinted upon the signal's amplitude. To manipulate individual wavelengths Microring Resonators (i.e., MRs, depicted in Fig. 1) can be employed whose central frequency can be actively adjusted (i.e., through tuning mechanisms using, e.g., microheaters or PIN junctions) to selectively interact with specific wavelengths. By appropriately tuning the MRs, the incoming light intensity of a specific wavelength can be weighted. In the non-coherent designs [17], [18], MRs as a fundamental component store the weight and activation values to be utilized in the MAC operation. During photonic MAC, incoming lights can be multiplied by the value adjusted on the MRs (through applying a tuning signal, see Fig. 1) of the same wavelength.

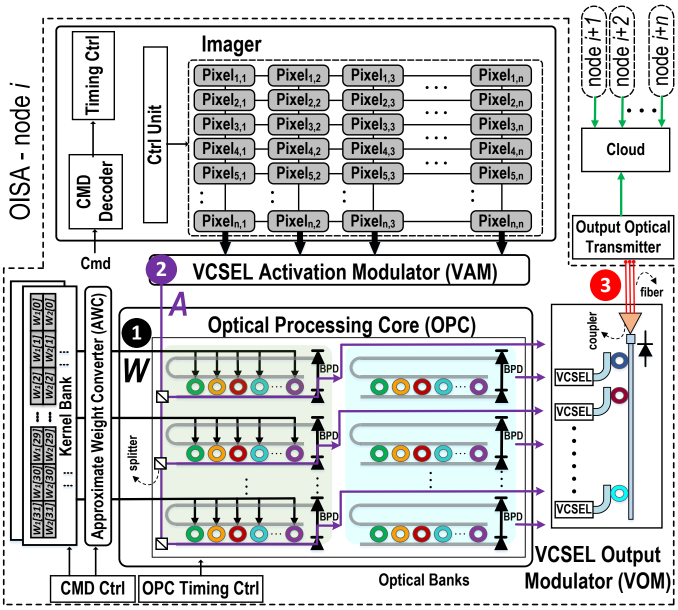


Fig. 2. OISA architecture with Global shutter CMOS imager, VCSEL-based Activation Modulator (VAM), Optical Processing Core (OPC), and VSCEL-based Output Modulator (VOM).

III. OISA Architecture

We propose OISA as a scalable, high-performance, and low-power solution for real-time image processing at edge devices. OISA integrates sensing and processing phases and intrinsically supports a low bit-width (2-bit (Ternary) activation and up to 4-bit weight) MAC operation of the 1^{st}-layer in Multi-Layer Perceptron (MLPs) and Convolutional Neural Networks (CNNs) while submitting the next layers to an off-chip processor through ultra-fast optical transmitters. The high-level overview of the proposed architecture, denoted by node *i* in a multi-node IoT structure, is shown in Fig. 2 consisting of six key components: *(i)* an ADC-less global shutter CMOS imager comprising a n×n conventional pixel structure to capture frames; *(ii)* a Vertical-Cavity Surface-Emitting Lasers (VCSEL)-based Activation Modulator (VAM) that is dedicated to directly modulate every pixel's voltage drop value after exposure to light (activation) with a predetermined wavelength and intensity proportional to the original light that is absorbed by pixel; *(iii)* an Approximate Weight Converter (AWC) to convert weight values stored on kernel banks to a current driving MRs; *(iv)* an Optical Processing Core (OPC) to execute parallel MAC operation between activation (A) and weight (W) parameters in the photonic domain; *(v)* a VSCEL-based Output Modulator (VOM) which is only used during the MLP processing or large kernel processing to break down the MAC operation when the number of elements in partial sum is

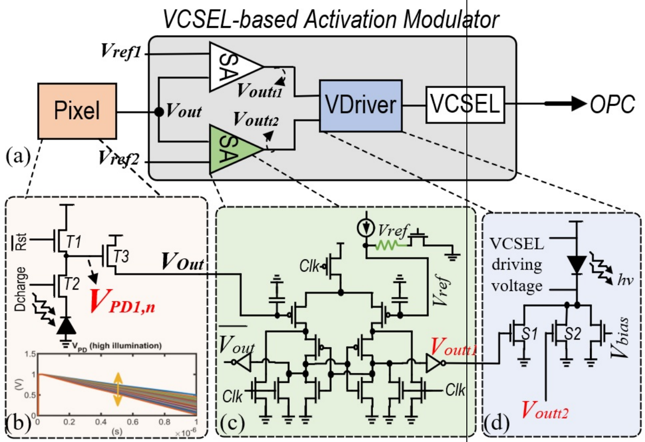


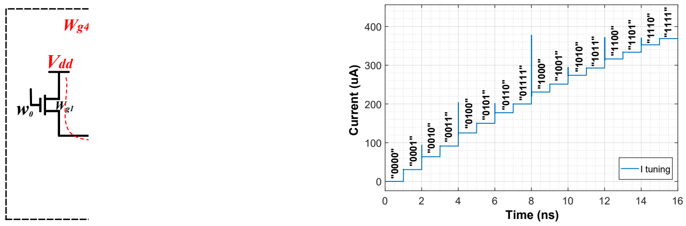
Fig. 3. (a) Proposed VCSEL-based Activation Modulator (VAM) diagram, (b) 3T pixel structure and the voltage across PD plot in high illumination, (c) Sense amplifier for thresholding, and (d) VCSEL Driver.

huge; and *(vi)* a controller to configure the timing and optical banks to perform data-parallel intra-bank computations. In the following, we elaborate on each component.

A. Microarchitectural Design

ADC-Less Imager. The imager consists of pixels with a 3-transistor-1-photodiode structure as shown in Fig. 3(b) with a Photodiode (PD) as the primary sensing component, a reset transistor (T1), discharge transistor (T2), and a source–follower (T3). In the sensing mode, by initially setting Rst='low' and Dcharge='high', the PD connected to the T2 transistor turns into inverse polarization and fully charges the PD capacitor. By turning off T1, PD generates a photo-current with respect to the external light intensity which in turn leads to a voltage drop (V_{PD}) at the gate of T3.

VCSEL-based Activation Modulator. OISA benefits from an optimized VSCEL driver for direct activation modulation that offers power efficiency and low cost, eliminates the need for external modulators, and provides ternary input for OPC. As shown in Fig. 3(a), the VAM consists of two sense amplifiers, a modified VCSEL driver (VDriver), and the VCSEL itself. In the proposed design, the output signal from the pixel is utilized to control the biasing current of the VCSEL. By appropriately realizing one distinct reference voltage for each SA (V_{ref} in Fig. 3(c)), V_{Out1}, and V_{Out2} (SA outputs) generate three states, i.e., both are zero, V_{Out1} is VDD, and V_{Out2} is zero, or both equal VDD. V_{Out1} and V_{Out2} are then used to control S1 and S2 transistors which control the bising current of the VCSEL as shown in Fig. 3(d). Completely turning off the VCSEL and turning it on again to warm up imposes extra energy and delay to the design [24]. To avoid that, another biasing transistor controlled by V_{bias} is added to keep the VCSEL always on. Thus, we have a non-returning-to-zero VCSEL implementation to save time and energy. The output voltages of the SAs determine the biasing current of the VCSEL and the light intensity of the generated light by the VCSEL. Accordingly, the output light intensity of the VCSEL won't be raw light, rather, it will carry ternary encoded data that corresponds to the light intensity absorbed by the pixel. In this way, our activation for the MAC operation is already modulated to the light by controlling electrical signals that run the VCSEL.



(a) (b)

Fig. 4. (a) Approximate weight converter, (b) Transient simulation results.

Approximate Weight Converter. In the OISA architecture, the weight parameters are initially held in on-chip kernel banks to be mapped to the MR elements in the OPC core. For this purpose, an approximate converter as shown in Fig. 4(a) is utilized. AWC is responsible for tuning the MRs to the desired weight values. Unlike prior optical accelerators that use area-consuming and power-hungry DACs to convert digital weight values to analog MRs' tuning signals [17], [18], [25], we propose an n-bit approximate weight converter ($n \leq 4$). As shown in Fig. 4(a), weight bits denoted by w_0 to w_3 are connected to the gate of T1 to T4. The key idea of AWC is to realize multi-level weighted current w.r.t. various weight values to mimic DAC behavior. Based on our circuit-level analysis, we have determined that increasing the width of transistors T1 to T4 results in a reliably enhanced current doubling effect where in the source node, all of these currents are summed. Thus, according to the spatial value of the weight bits, as depicted in Fig. 4(b), the AWC generates up to 16 levels ($n = 4$) of current to be used for MR's tuning purposes. We elaborate on AWC latency and power consumption in Section IV.

Optical Processing Core & VCSEL Output Modulator. The proposed OPC as shown in Fig. 2 is a non-coherent photonic computing core composed of MRs that are arrayed in arms. Each arm is comprised of 10 MRs and two waveguides used for positive and negative weights. As shown in Fig. 2 ❶, digital weights (W) of the DNN or MLP are converted to the analog tuning control values by the AWC unit. Utilizing these values, the weights are mapped to the MRs. This step is crucial only when OISA takes a new set of weight kernels into the processing core; once the weights are mapped, it can bypass this step. In ❷, activation values coming from the pixel plane are modulated to their respective wavelength using VCSELs in the VAM unit. The resultant lights whose intensity corresponds to the pixel's output values are then applied to the MR banks in the OPC. Inside the arms, each MR affects a specific wavelength of the applied light and weights the intensity of that particular wavelength. Thus, the multiplication operation of the weight values stored in the MRs and the light coming from VCSELs is conducted. At the end of each arm, two Balanced PhotoDiodes (BPD), shown in Fig. 2, perform the summation operation of both positive and negative lights resulting in an electrical output voltage that represents the result of MAC operation between the stored weights and the incoming light. In other words, BPDs convert the optical values to the electrical signals representing the sum of the dot products. Depending on the weight kernel size, these values in ❸ are either summed up using extra optical summation arms or transmitted to the output of our

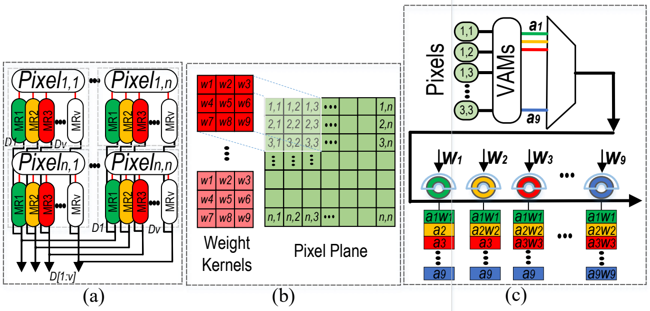


Fig. 5. (a) High-level hardware mapping to OISA, (b) Stride of a 3×3 Kernel in convolution operation, (c) Implementing a stride in an arm.

chip directly for further processing. In the case of the MLP, the number of dot products is enormous. To reduce the complexity of the calculations, the VOM unit is added to the architecture that enables OISA to break the intensive MAC operations into smaller parts and perform the calculations. It is noteworthy that OISA uses off-chip resources to perform the non-linearity (activation function).

MR Device Engineering. Increasing the MR resolution ideally leads to higher weight/activation precision and accuracy. However, it demands a reduction in the Quality Factor (Q-factor), another pivotal characteristic of the MRs. The Q-factor describes the resonance's sharpness relative to the MR's central frequency. A higher Q-factor results in a sharper resonance which can make the system more sensitive to noise, as even a slight deviation in the central frequency can result in significant losses. Here, a smaller Q-factor is preferred over a large one [19]. Yet, achieving smaller Q-factors often entails enlarging the dimensions of the MRs, which can, in turn, introduce substantial optical crosstalk noise and energy demands for tuning. In this work, we tune and leverage the effective bit resolution of 4-bit for MRs [18], [19] to make a balance among the above-mentioned parameters. Using Lumerical tools, we designed an MR with a radius of $5\mu m$ and a ring waveguide width of 760 nm. These dimensions resulted in a relatively small Q-factor ($\approx$ 5000) which provides sufficient differentiation levels to carry out our intended multi-bit design. To tune MRs during weight mapping, Thermo-Optic (TO) or Electro-Optic (EO) methods are widely utilized. EO tuning is faster than TO but it can only create a slight change in MR's resonant wavelength. On the other hand, TO tuning has the capability to largely shift the MR's resonant wavelength but at a cost of larger delay and more power consumption. Similar to [18], a hybrid TO-EO tuning method is utilized to leverage both methods' advantages.

B. Hardware Mapping & Bank Allocation

Aiming to process the first layer of DNNs, OISA is equipped with a correlated hardware mapping method to balance the workload and increase the throughput. A high-level hardware mapping is shown in Fig. 5(a), where the inputs from the pixel plane are connected to their respective weight values that are mapped to the MRs. A 3×3 kernel stride over a pixel plane is shown in Fig. 5(b). To localize the kernel stride computation in one arm, OISA supports a kernel size of 3×3 as seen by several well-known DNN models. Fig. 5(c) shows the detailed implementation of a stride where weight values related to a 3×3 kernel are mapped on the MRs in an arm. Input activation values then are modulated on a light with a wavelength according to the resonance wavelength of their corresponding weight. These input-modulated lights are then multiplexed to a single light and are passed through the arm. MRs affect the intensity of the light passing the arm and weight the specific wavelength providing multiplication between 9 activation values and 9 weight values. Later a BPD at the end of the arm will provide the optical summation of multiplication results to conclude the MAC operation. We develop OISA with 10 MRs in each arm that enables us to perform 9×9 (3×3 kernel values by 9 activation values) MAC operations. However, in order to enhance the core's efficiency in handling 5×5 and 7×7 kernels (25 and 49 weight values), we partition the cores in the OPC hierarchy. As shown in Fig. 6, in our core, each bank comprises 5 arms, each equipped with 10 MRs, resulting in a total of 50 MRs per bank. With 80 banks in OPC, OISA consists of 4000 MRs in total. Banks are grouped in the 4 columns. Thus, each row has 40 MRs, and 40 AWC units are assigned to map the weights to MRs. It is worth pointing out that to completely map all the weight values to the OPC, 100 iterations are required. In the case of 3×3 kernels, the MAC result of each arm will represent a stride of the convolution operation and can be directly transferred to the output. Supporting 5×5 and 7×7 weight kernels, the output of each bank will be further processed in the VOM unit to obtain the final MAC results. According to this configuration, the total number of MAC operations that can be processed in one cycle ($N_{\frac{m}{cycle}}$) can be formulated as $f \times (nK^2)$. Where f is the number of banks, $K \in \{3,5,7\}$ is the kernel size. We consider $n = 5$ when $K = 3$, as 5 kernels with the size of 3×3 can be mapped to each bank. Else $n = 1$ as only one 5×5 or 7×7 can be mapped to each bank. Thus, for $K = 3, 5, 7$, in each cycle, OISA conducts 3600, 2000, and, 3920 MAC operations. The total required cycles for performing convolution operation depend on the number of weight kernels and their sizes.

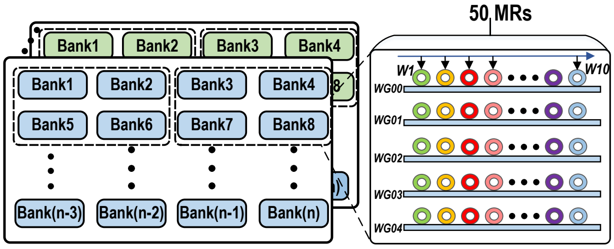


Fig. 6. Optical array partitioning and allocation of OISA.

IV. Performance Evaluation

Framework. The evaluation framework is created through a bottom-up methodology as shown in Fig. 7. At the device level, we fabricated and optimized the MR device and extracted the circuit parameters to co-simulate with interface CMOS circuits in Cadence Spectre and SPICE. At the circuit level, we first implemented the OISA pixel's array and peripheral circuitry using the 45nm NCSU Product Development Kit (PDK) library [26] in Cadence and extracted the output voltages and currents. We then developed all OISA's components except the kernel

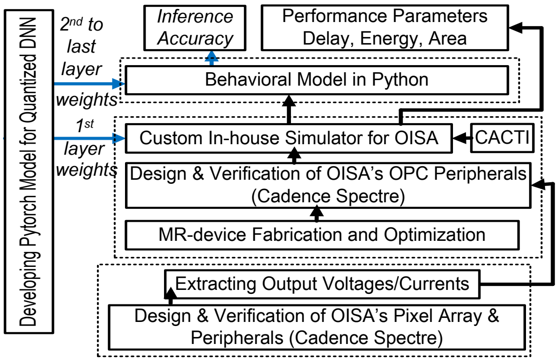

Developing Pytorch Model for Quantized DNN
2nd to last layer weights
1st layer weights
Inference Accuracy
Performance Parameters Delay, Energy, Area
Behavioral Model in Python
Custom In-house Simulator for OISA
CACTI
Design & Verification of OISA's OPC Peripherals (Cadence Spectre)
MR-device Fabrication and Optimization
Extracting Output Voltages/Currents
Design & Verification of OISA's Pixel Array & Peripherals (Cadence Spectre)

Fig. 7. Proposed bottom-up evaluation framework.

banks (implemented in Cacti [27]) in Cadence Spectre. The DNN weight parameters associated with the 1^{st} layer need to be quantized and mapped into the OPC, while the remaining layers are processed with the off-chip processors. We trained a PyTorch model w.r.t. the under-test datasets and extracted the 1^{st}-layer weights. OISA's MR elements are then adjusted with these weights. For computation purposes, we developed a custom in-house simulator for OISA. This simulator computes the overall latency and power consumption required for the network execution. It also offers flexibility in array configuration and peripheral design selection. The results are captured after processing the 1^{st} convolution layer. We then developed a Python-based behavioral DNN model to utilize the computation outcomes and processes the 2^{nd}-to-last layer to calculate inference accuracy.

Functionality. Figure 8 shows the transient simulation waveforms of VAM's thresholding to drive VCSEL depicted in Fig. 3(a). In this figure, the t_1 and t_2 represent two outputs corresponding to three distinct pixels denoted as $Out1$, $Out2$, and $Out3$. Each pixel exhibits a unique voltage implying its absorption of varying light intensities. Specifically, the voltage levels for $Out1 - 3$ are ascertained whenever the Clk signal is low, observable within the time frame of 16 to 17 ns. Within this interval, the voltage for $Out1$ surpasses both sense amplifier thresholds, resulting in both t_1 and t_2 being set to 1. Conversely, the voltage for $Out2$ resides between 0.16V and 0.32V, leading to t_1 equating to 1 and t_2 being 0. As for $Out3$, its voltage is less than 0.16V, thereby setting both t_1 and t_2 to 0.

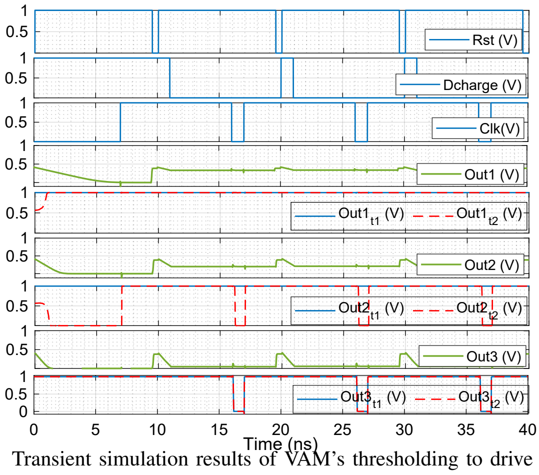
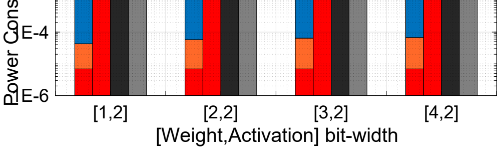
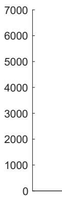


Fig. 8. Transient simulation results of VAM's thresholding to drive VCSEL.





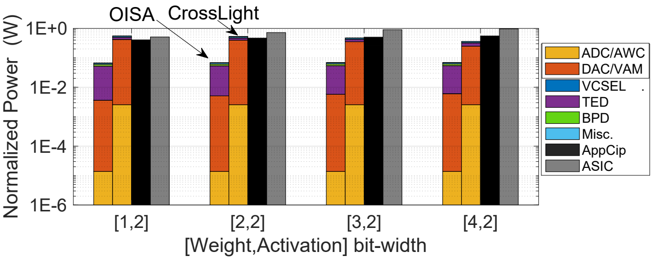


Fig. 9. Normalized log-scaled power consumption of various accelerators. From left to right: OISA, Crosslight, AppCip, and ASIC design.

Power Consumption & Performance. Here, we compare the OISA with three DNN accelerators based on PIS and ASIC as follows assuming a 1-to-4-bit width for the weight parameter. For evaluation, we assume that all platforms process the 1^{st} layer of the ResNet18 model. ***Optical PIS:*** We designed a Crosslight-like [18] platform with 80 banks, each consisting of 5 arms with 10 MRs. For a fair evaluation, we developed the design from scratch using the proposed evaluation framework and the in-house simulator to extract numbers. Note that the Crosslight uses separate banks for weight and activation. ***Electronic PIS:*** We developed an AppCip-like [13] accelerator with non-volatile memory in HSPICE and NVSIM [28] from scratch and extracted the performance parameters. ***ASIC:*** We developed a DaDianNao-like [29] accelerator with 8×8 tile version connected to a conventional 128×128 image sensor. We synthesized the designs with the Design Compiler under the 45 nm process node. The eDRAM and SRAM performance was estimated using CACTI [27]. Figure 9 shows the normalized power consumption of the under-test platforms in various bit-width configurations. From this figure, we observe the superiority of OISA over various under-test platforms where it achieves 8.3×, 7.9×, and 18.4× reduction in power consumption on average compared with Crosslight [18], AppCip [13], and ASIC platforms, respectively. We report the breakdown of power consumption for OISA and Crosslight platforms as well, where we observe a remarkable reduction in power consumption mainly due to ADC/DAC elimination in OISA as compared with Crosslight. As for execution time, considering that the activation and weight values are already mapped to the core, in the OISA, and Crosslight-like designs, the utilized VCSEL and BPD technologies have critical effects on the execution time. To have a fair comparison same VCSEL [30] and BPD [17] technologies have been used for OISA and corsslight-like designs. The total execution time for performing one architecture-wide MAC operation is 55.8 ps which results in 7.1 TOp/s. However, the most important difference between these two designs is that in the OISA, all of the MRs in OPC are allocated to weight values, while in the Crosslight-like design, half of the MRs are considered to be mapped by activation which cuts the total number of operations to half.

Table I presents a comparison of the structural and performance characteristics of selective PIS implementations in the electronic domain and OISA. Since these implementations are tailored for specific domains, we conducted a fair assessment by estimating and normalizing the power consumption considering a scenario where all PIS units process the first layer of a CNN. The OISA reaches the frame rate of 1000 and the

TABLE I
PERFORMANCE COMPARISON OF VARIOUS PIS/PNS/PIP UNITS. NOTE: OISA IS THE ONLY HYBRID IN-SENSOR ACCELERATOR.

Designs	Technology (nm)	Purpose	Comput. Scheme	Memory	NVM	Pixel Size (μm^2)	Array Size	Frame Rate (frame/s)	Power (mW)	Efficiency (TOp/s/W)
[31]	180	2D optic flow est.	row-wise	Yes	No	28.8×28.8	64×64	30	0.029	0.0041
[8]	180	edge*/blur/sharpen/ 1^{st}-layer CNN	row-wise	No	No	7.6×7.6	128×128	480	sensing: 77 processing: 91	0.777
[9]	60/90	STP†	row-wise	Yes	No	3.5×3.5	1296×976	1000	sensing: 230 processing:363	0.386
[2]	180	1^{st}-layer BNN	entire-array	Yes	No	110×110	32×32	1000	0.0121	1.32
[32]	180	edge*/TMF‡	row-wise	Yes	No	32.6×32.6	256×256	100,000	1230	0.535
[3]	65	1^{st}-layer BNN	entire-array	Yes	Yes	55×55	128×128	1000	sensing: 0.025 processing: 0.0088	1.745
[12]	180	1^{st}-layer BNN	entire-array	Yes	No	35×35	32×32	156	0.00014 - 0.00053	9.4-34.6
[21]	65	2 - 64 Conv/ROI**	row-wise	No	No	9×9	160×128	96 - 1072	0.042 - 0.206	0.15 - 3.64
[1]	180	1^{st}-layer CNN	entire-array	No	No	10×10	128×128	3840	0.45 - 1.83	1.41 - 3.37
[13]	45	1^{st}-layer CNN	entire-array	Yes	Yes	38×38	32×32	3000	0.00096 - 0.0028	1.37 - 4.12
OISA	65	1^{st}-layer CNN	entire-array	Yes	No	4.5×4.5	128×128	1000	0.00012-0.00034	6.68

* Edge extraction. †Spatial Temporal Processing. ‡Thresholding Median Filter. **Region Of Interest.

efficiency of ∼6.68 TOp/s/W as one of the most efficient implementations. This comes from the massively-parallel OPC banks and eliminating ADC/DAC for coarse-grained inference. As for the area, according to the MR's dimensions mentioned in section III, and our architecture configurations, the total area of the OISA is 1.92 mm^2. The simulation results reported in Table I demonstrate no modification on the pixel array. Due to our lack of access to the configurations of other layouts, it is challenging to establish a fair comparison with respect to the total area overheads.

Accuracy. We conduct experiments on OISA considering various [Weight: Activation] configurations with several datasets, including MNIST evaluated on LeNet, SVHN on ResNet18, CIFAR-10 on ResNet18, and CIFAR-100 on VGG16 compared with a software baseline, FBNA [33], AppCiP [13], and PISA [3] as recent low bit-with accelerators. The comparison results of classification accuracy are summarized in Table II. We find that 1) OISA shows an acceptable accuracy while providing significant power-delay reduction as discussed earlier compared with other platforms; 2) Generally, our experiments show that weights and inputs are progressively more sensitive to bit-width changes. However, a higher weight bit-width does not necessarily result in a higher accuracy as indicated in the OISA [4:2] configuration. This comes from the fact that AWC may not reliably provide distinct current levels when the number of bits increases; and 3) The accuracy drop of OISA is mainly because of the ADC-DAC-less nature of it that allows processing the 1^{st} convolutional layer with 1 to 4 bits approximated with a converter.

TABLE II
OISA'S ACCURACY (%) ON VARIOUS DATASETS.

Configuration	MNIST	SVHN	CIFAR-10	CIFAR-100
baseline	99.6	97.5	91.37	78.4
FBNA [33]	–	96.9	88.61	71.5
AppCiP [13]	–	96.4	89.51	–
PISA [3]	95.12	90.35	79.80	61.6
OISA [4:2]	95.21	91.74	81.23	61.38
OISA [3:2]	96.18	94.36	84.45	66.89
OISA [2:2]	96.25	93.20	83.85	66.94
OISA [1:2]	95.75	93.16	83.64	66.06

V. CONCLUSION

In this work, we presented OISA as a high-performance and energy-efficient optical in-sensor accelerator architecture. OISA benefits from an innovative design and hardware mapping method to remarkably reduce the power consumption of data conversion, transmission, and processing in the conventional cloud-centric architecture as well as recent edge accelerators. Our results on various image data-sets show acceptable accuracy while OISA achieves 6.68 TOp/s/W efficiency.

VI. ACKNOWLEDGMENT

This work is supported in part by the National Science Foundation (NSF) under grant numbers 2046226, 2006788, 2216772, 2216773, and 2247156.

REFERENCES

[1] R. Song *et al.*, "A reconfigurable convolution-in-pixel cmos image sensor architecture," *IEEE TCSVT*, 2022.

[2] H. Xu *et al.*, "Macsen: A processing-in-sensor architecture integrating mac operations into image sensor for ultra-low-power bnn-based intelligent visual perception," *IEEE TCAS II*, vol. 68, pp. 627–631, 2020.

[3] S. Angizi *et al.*, "Pisa: A non-volatile processing-in-sensor accelerator for imaging systems," *IEEE TETC*, 2023.

[4] K.-T. Tang *et al.*, "Considerations of integrating computing-in-memory and processing-in-sensor into convolutional neural network accelerators for low-power edge devices," in *Symposium on VLSI*. IEEE, 2019.

[5] S. Angizi *et al.*, "A near-sensor processing accelerator for approximate local binary pattern networks," *IEEE Transactions on Emerging Topics in Computing*, 2023.

[6] T.-H. Hsu *et al.*, "AI edge devices using computing-in-memory and processing-in-sensor: from system to device," in *IEDM*, 2019.

[7] M. Morsali *et al.*, "Design and evaluation of a near-sensor magneto-electric fet-based event detector," *IEEE Transactions on Electron Devices*, 2023.

[8] T.-H. Hsu, Y.-R. Chen *et al.*, "A 0.5-v real-time computational cmos image sensor with programmable kernel for feature extraction," *IEEE JSSC*, vol. 56, pp. 1588–1596, 2020.

[9] T. Yamazaki *et al.*, "4.9 a 1ms high-speed vision chip with 3d-stacked 140gops column-parallel pes for spatio-temporal image processing," in *ISSCC*. IEEE, 2017, pp. 82–83.

[10] S. Angizi *et al.*, "Cmp-pim: an energy-efficient comparator-based processing-in-memory neural network accelerator," in *Proceedings of the 55th Annual Design Automation Conference*, 2018, pp. 1–6.

[11] S. Angizi, Z. He *et al.*, "Mrima: An mram-based in-memory accelerator," *IEEE TCAD*, vol. 39, no. 5, pp. 1123–1136, 2019.

[12] H. Xu *et al.*, "Senputing: An ultra-low-power always-on vision perception chip featuring the deep fusion of sensing and computing," *IEEE TCASI: Regular Papers*, 2021.

[13] S. Tabrizchi *et al.*, "Appcip: Energy-efficient approximate convolution-in-pixel scheme for neural network acceleration," *IEEE JETCAS*, vol. 13, pp. 225–236, 2023.

[14] A. El Gamal *et al.*, "Pixel-level processing: why, what, and how?" in *Sensors, Cameras, and Applications for Digital Photography*, vol. 3650. SPIE, 1999, pp. 2–13.

[15] J. Choi *et al.*, "An energy/illumination-adaptive cmos image sensor with reconfigurable modes of operations," *IEEE JSCC*, vol. 50, no. 6, pp. 1438–1450, 2015.

[16] A. Roohi *et al.*, "Pipsim: A behavior-level modeling tool for cnn processing-in-pixel accelerators," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2023.

[17] F. P. Sunny *et al.*, "Robin: A robust optical binary neural network accelerator," *ACM TECS*, no. 5s, pp. 1–24, 2021.

[18] F. Sunny *et al.*, "Crosslight: A cross-layer optimized silicon photonic neural network accelerator," in *DAC*. IEEE, 2021, pp. 1069–1074.

[19] Q. Cheng *et al.*, “Silicon photonics codesign for deep learning,” *Proceedings of the IEEE*, vol. 108, pp. 1261–1282, 2020.

[20] H. Xu *et al.*, “Utilizing direct photocurrent computation and 2d kernel scheduling to improve in-sensor-processing efficiency,” in *DAC*, 2020.

[21] M. Lefebvre *et al.*, “7.7 a 0.2-to-3.6 tops/w programmable convolutional imager soc with in-sensor current-domain ternary-weighted mac operations for feature extraction and region-of-interest detection,” in *ISSCC*, vol. 64. IEEE, 2021, pp. 118–120.

[22] F. P. Sunny *et al.*, “Arxon: A framework for approximate communication over photonic networks-on-chip,” *TVLSI*, vol. 29, pp. 1206–1219, 2021.

[23] Z. Zhao *et al.*, “Hardware-software co-design of slimmed optical neural networks,” in *ASP-DAC*, 2019, pp. 705–710.

[24] D. Breuer *et al.*, “Comparison of nrz- and rz-modulation format for 40-gb/s tdm standard-fiber systems,” *IEEE Photonics Technology Letters*, vol. 9, no. 3, pp. 398–400, 1997.

[25] Z. Zhong *et al.*, “Lightning: A reconfigurable photonic-electronic smartnic for fast and energy-efficient inference,” in *ACM SIGCOMM 2023 Conference*, 2023, pp. 452–472.

[26] (2011) Ncsu eda freepdk45. [Online]. Available: http://www.eda.ncsu.edu/wiki/FreePDK45:Contents

[27] S. Thoziyoor, N. Muralimanohar, J. H. Ahn, and N. P. Jouppi, “Cacti 5.1,” Technical Report HPL-2008-20, HP Labs, Tech. Rep., 2008.

[28] X. Dong *et al.*, “Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory,” *IEEE TCAD*, vol. 31, pp. 994–1007, 2012.

[29] Y. Chen *et al.*, “Dadiannao: A machine-learning supercomputer,” in *Micro*. IEEE, 2014, pp. 609–622.

[30] K. S. Kaur *et al.*, “Flip-chip bonding of vcsels to silicon grating couplers via su8 prisms fabricated using laser ablation,” *2015 European Conference on Optical Communication (ECOC)*, pp. 1–3, 2015.

[31] S. Park *et al.*, “7.2 243.3 pj/pixel bio-inspired time-stamp-based 2d optic flow sensor for artificial compound eyes,” in *IEEE ISSCC*. IEEE, 2014.

[32] S. J. Carey *et al.*, “A 100,000 fps vision sensor with embedded 535gops/w 256× 256 simd processor array,” in *Symposium on VLSI*. IEEE, 2013.

[33] P. Guo *et al.*, “Fbna: A fully binarized neural network accelerator,” in *FPL*. IEEE, 2018, pp. 51–513.