2023 Design, Automation & Test in Europe Conference (DATE 2023)

Scalable Coherent Optical Crossbar Architecture using PCM for AI Acceleration

Dan Sturm
Electrical and Computer Engineering
University of Washington
Seattle, USA
dansturm@uw.edu

Sajjad Moazeni
Electrical and Computer Engineering
University of Washington
Seattle, USA
smoazeni@uw.edu

***Abstract*—Optical computing has recently been proposed as a new compute paradigm to meet the demands of future AI/ML workloads in datacenters and supercomputers. However, proposed implementations so far suffer from lack of scalability, large footprints and high power consumption, and incomplete system-level architectures inhibit integration within existing datacenter systems for real-world applications. In this work, we present a truly scalable optical AI accelerator based on a crossbar architecture. We have considered all major roadblocks and address them in this design. Weights will be stored on-chip using phase change material (PCM) that can be monolithically integrated in silicon photonic processes. All electro-optical components and circuit blocks are modeled based on measured performance metrics in a 45nm monolithic silicon photonic process, which can be co-packaged with advanced CPU/GPUs and HBM memories. We also present a system-level modeling and analysis of our chip's performance for the Resnet-50V1.5, considering all critical parameters, including memory size, array size, photonic losses, and energy consumption of peripheral electronics. Both on-chip SRAM and off-chip DRAM energy overheads have been considered in this modeling. We additionally address how using a dual-core crossbar design can eliminate programming time overhead at practical SRAM block sizes and batch sizes. Our results show that a 128×128 proposed architecture can achieve inference per second (IPS) similar to Nvidia A100 GPU at $15.4\times$ lower power and $7.24\times$ lower area.**

***Index Terms*—Optical Neural Networks, AI Accelerator, Crossbar, Phase Change Material, System-level Optimization**

I. Introduction

Recent advancements in artificial intelligence (AI) and machine learning (ML) have been challenging our conventional computing paradigms by demanding enormous computing power at a dramatically faster pace than Moore's law [1]. We can compare the performance of today's AI/ML processors from two key aspects of compute power in terms of Tera operations per second (TOPS) and energy-efficiency (TOPS/W) as illustrated in Fig. 1. Despite the promising success of neuromorphic and analog-based computing in electrical domains for low TOPS applications (edge-computing), these approaches can not satisfy the requirements of datacenters and super computers. Due fundamental bandwidth limitations, they cannot achieve high throughput. Optical neural networks (ONNs) can potentially overcome this barrier by providing tens of GHz bandwidths and ultra-low losses of photonic integrated circuits [2], [3]. However, realizing a practical ONN-based AI accelerator requires a holistic system design that considers devices, circuits, chip architectures, and algorithms.

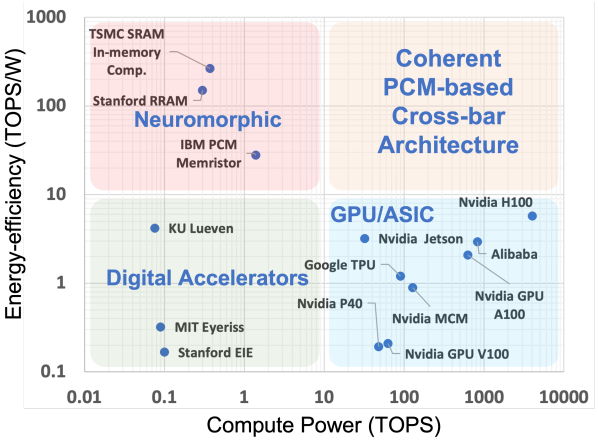

Fig. 1: Comparison of state-of-the-art AI/ML processors.

In this paper, we present a novel architecture for an ONN accelerator in which the multiply-and-accumulate (MAC) operation is performed on a coherent photonic crossbar with programmable phase-change materials (PCM). PCM enables low power photonic computing by storing the weights on-chip in a nonvolatile fashion. This design provides a compact and scalable solution for the first time, which relies minimally on thermo-optic phase-shifters. We have considered all critical circuit blocks and parameters, including memory size, array size, photonic losses, and energy consumption of peripheral electronics including analog-to-digital converters (ADCs), digital-to-analog converters (DACs), serializers, and clocking. We develop a custom simulation framework based on existing cycle-accurate simulation tools to model the compute cycles, programming cycles, and DRAM accesses for a given neural network running on a specific set of parameters (including size of SRAM, array, and batch).

The presented work focuses on inference of convolutional neural networks (CNN) such as ResNet50 v1.5, which is used as a benchmark to compare our proposed accelerator performance with the state-of-the-art. Additionally, while precision and process variation are major factors in all analog-

This work is supported by National Science Foundation (NSF) under Grant No. CCF-2105972, and Clean Energy Institute (CEI) Graduate Fellowship.

978-3-9819263-7-8/DATE23/© 2023 EDAA

based computers, here we assume a INT6 precision for all the components as it has been shown to be sufficient for neural networks with high accuracy [4], [5].

In this paper, we briefly describe related work in Section II. In Section III, we explain the principles of performing the MAC operation in this work. We describe overall chip architecture and CNN operation in Section IV, and section V explains our custom simulation methodology. Finally, we present results from our fully optimized design in Section VI, and compare those with the state-of-the-art in Section VIII.

II. Related Work

Researchers have recently proposed a variety of methods to realize an ONN. Most of these works focus only on the physics and devices rather than providing a system-level solution and analysis. In this presented discussion, we only consider integrated solutions, as free-space solutions [6] lack the reconfigurability that is an essential part of any "computer" and compatibility with mainstream CMOS technology. Furthermore, we note that a suitable application space for ONNs can be datacenters as opposed to edge computing according to Fig. 1. Below we discuss the most promising solutions so far from the perspective of three critical factors that we believe have been addressed in this work:

(1) Scalability: While on-chip photonics provide high bandwidths, their footprints are fundamentally orders of magnitude larger than advanced nm-scale CMOS. With only one or two routing layers, building an ONN processor with big dimensions has remained elusive. This will become even more challenging considering the need for compact ADCs and DACs. Mach-Zehnder Interferometer (MZI)-based coherent architectures such as [2] have large chip areas and large-scale realizations end up exceeding a few cm^2. Non-coherent PCM-based crossbars have been also proposed [7], however they require many wavelengths for large matrix operations and that is impractical. Time-multiplexed coherent arrays [8], [9] also require 2D arrays of free-space detectors yet only perform vector-vector multiplication in one clock cycle.

(2) Monolithic Integration: Since any practical computing system will eventually require high-density CMOS electronics for I/O and memory, we argue that electronics and photonic should be monolithically integrated on a single chip using processes such as GF 45CLO [10]. While 3D integration is typically proposed as an alternative, existing advanced 3D integration technologies (micro-bumps at $55\mu m$ pitch [11]) cannot provide density for optical computing applications.

(3) Full Architecture-level Modeling and Optimization: Practical AI accelerators, including optical processors, should be modeled at the system level. This has been previously discussed in [12], however, accessing DRAM through a PCIe switch will have large energy and latency overheads. We elaborate on this aspect here, and model the system with co-packaged high-bandwith memory (HBM), similar to state-of-the-art AI accelerators. In addition, we discuss impacts and trade-offs between the programming time, batch size, and multiple cores in this work.

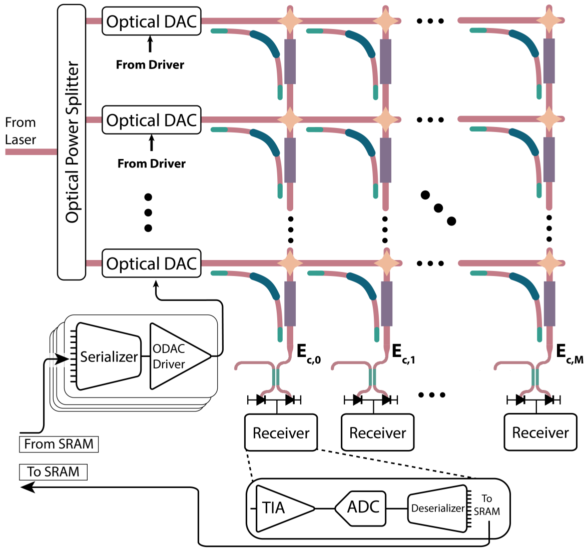

Fig. 2: Photonic crossbar array with peripheral electronics for transmitter and receiver

III. Proposed Crossbar Design for Optical MAC

The overall proposed crossbar ONN is shown in Fig. 2. Below, we describe two key components of this design:

A. Analog-based Optical MAC Core

The cross-bar design is an $N \times M$ array of PCM-based unit cells. Details of these unit cells and how the array performs MAC operations is briefly described below.

1) PCM-Based Unit Cell: Each unit cell multiplies an input electric field (E-field) by the weight programmed into the PCM section, and adds this product to an externally-inputted electric field through each column. To do so, a portion of light from the row waveguide (E-field into each row is denoted by $|E_{in,i}|$ in Fig. 3) is partially coupled into a bended waveguide via a directional coupler (DC) with a cross-coupling ratio of $k_{in,j}$ (input coupling strength is column-dependent). The portion of the E-field that does not couple into the unit cell passes through a multi-mode interference (MMI) waveguide crossing junction and enters the next column of unit cells to the right. Each bended waveguide has a μm-long section covered with PCM. Individual PCM cells can be programmed with heat (supplied electrically) to be either in the amorphous or crystalline state, or somewhere in between, in a non-volatile fashion [7], [8]. Programming energy is estimated to be around $100pJ$ [7], [8]. The state of PCM changes the absorption coefficient, and hence it can change the E-field amplitude. Consequently, if the PCM's programmed transmission in E-field domain is $w_{i,j}$, the E-field at the end of each bended waveguide will be $|E_{in,i}| \times k_{in,j} \times w_{i,j}$, which we refer to as the product $E_{p,(i,j)}$. Another DC (with $k_{out,i}$ coupling ratio) couples this product into a column waveguide connected to the cells above and below. This E-field in this column waveguide now represents the sum of products computed in the present cell and all above cells, ultimately summing all products in its column. The output coupling strength is row-dependent.

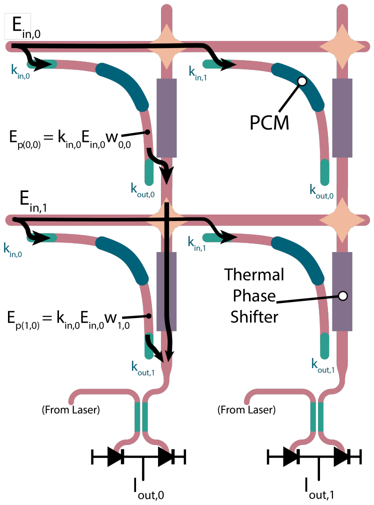

Fig. 3: Principle of optical MAC operation in a coherent crossbar array (an example of 2×2 unit cell array).

2) Crossbar Array of Unit Cells: In order to perform a MAC operation on our $N \times M$ crossbar array (Fig. 3), we need to encode the input data vector (v_{in}) on each row's E-field amplitude. To do so, we use a splitter tree structure and a set of optical DACs (ODACs) to generate $v_{in,i} \times E_{Laser}/\sqrt{N}$ at each row. Correctly designed input coupling coefficients in each unit cell ($k_{in,j}$) distribute this equally throughout the row, so each unit cell receives $v_{in,i} \times E_{Laser}/\sqrt{NM}$ in its bended waveguide.

Similarly, output coupling coefficients ($k_{out,i}$) are designed correctly so that outputted light equally represents the product computed by every unit cell in a given column. Although this adds an electrical field loss of $1/\sqrt{N}$, this enables the entire array to operate on a single frequency of light in a compact footprint (unlike previous works like [2]), which is critical for scaling the array for high-performance compute. We can calculate the resultant E-field at the end of each column as:

$$E_j^c = \frac{E_{Laser}}{N\sqrt{M}} \sum_{i=0}^{N-1} |v_{in,i}| \times w_{i,j} \tag{1}$$

Thus, the overall E^c vector will be equal to the matrix-vector multiplication of input data (v_{in}) by a weights matrix of w. We can convert this back into the electrical domain using coherent detection by coupling each signal with a certain portion of the input laser light in a DC. Coupler outputs enter balanced photodiodes (PD), with $I_{out,j} \propto |E_{Laser}||E_j^c|$. We note that our scheme relies on maintaining coherency across the entire array, which requires precise optical path lengths and phase shift matchings. In order to adjust for potential phase errors due to process variations or random phase variations, etc., we propose adding a small thermal-phase shifter in each unit cell across the column waveguides. The PCM's refractive index may be state-dependent, which can be taken into account when calibrating the appropriate programming weight values.

The photonic elements in our crossbar array present numerous sources of loss, which are critical for system modeling:

- Grating coupler: 2 dB [10], [13]
- Splitting tree: 0.1 dB/splitter [14]
- MMI Crossing Junction: 0.01 dB/junction [15]
- Waveguide loss: 3 dB/cm [10]
- Effective loss due to ODAC optical modulation amplitude (OMA) - 4 dB [16]
- Laser wall-plug efficiency: 15%

B. Peripheral Electronic Circuitry

While the proposed crossbar array can perform optical MAC operations efficiently in a compact footprint with on-chip PCM-based unit cells, the premise of ultra-fast operations relies on energy-efficient and high-speed electro-optical conversions. Here we elaborate on suitable design choices for this aim. We have added area and power estimates of critical peripheral electronics for a 10GHz MAC operation. Note that some assumptions may involve extrapolation from papers reporting work in other processes or operation frequencies.

1) Optical On-chip Transmitter: As mentioned, the amplitude of the input E-field at each row should be modulated with the input vector data. Notice that coherent operation requires that phase of the input E-field remain constant across amplitudes. This has been previously proposed to be done via MZIs, however, the mm-long footprint of MZIs impose unacceptable area and energy overheads for scalable ONN applications. In this design, we propose using micro-ring-assisted MZIs (RAMZIs). RAMZI devices have been previously proposed for high linearity [17], but here we propose using them to perform constant-phase PAM modulation. This can be done by placing one ring-resonator-based ODAC in each arm. Such ODACs have been demonstrated in the 45nm silicon photonics with +20GS/s at low power with up to 6-bit accuracy for data modulation [16]. Each ODAC driver consumes $168fJ$ power and occupies a $0.0012mm^2$ area at 10GS/s [16]. We also add a thermal tuning over head of $0.72mW$ per ring-resonator [16].

2) Optical On-chip Coherent Receiver: Output photocurrents ($I_{out,j}$) should be amplified via a trans-impedance amplifier (TIA) and digitized using an ADC (Fig. 3). Similar coherent receiver designs have been demonstrated in 45nm silicon photonics with $2.25mW$ per TIA [18]. ADC power and area at a 10 GHz sample rate is estimated to be $25mW$ and $0.0475mm^2$, respectively, in 45nm CMOS [19].

3) SerDes and Clocking: Operating the ONN crossbar at high clock rates of +10GHz requires a set of serializer/deserializer (SerDes) to interface the optical transmitter and receiver data with the digital backend (SRAM in this case). The serialization ratio depends on the speed of MAC operation and readout speed from SRAM memory. For a 10GHz ONN operation and a ~1GHz backend clock-rate, we assume a 10:1 ratio. Estimated power is roughly $100fJ$ per bit [16]. The transmit and receive blocks require high speed clock generation and distribution, which we assume will take $200fJ$ power with $0.005mm^2$ per row/column [16].

IV. Overall Accelerator Architecture

In this work we focus on CNNs for our modeling and benchmarking, as they are one of the most popular deep

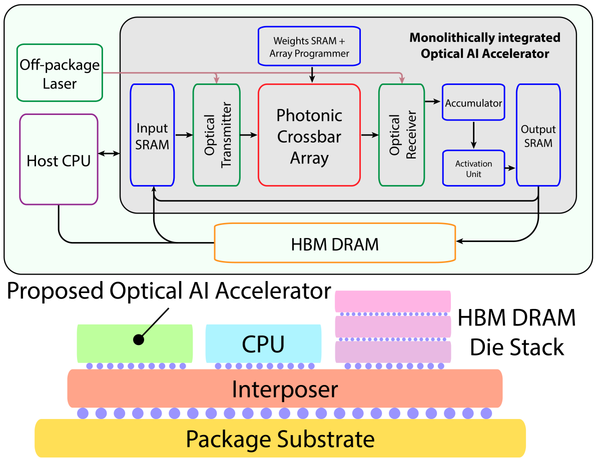

Fig. 4: Architecture of proposed photonic crossbar array AI accelerator.

learning algorithms due to their strong performance in image classification and object detection [20]. Performing inference using a CNN includes multiple convolutional layers using various sets of filters. Our crossbar array capitalizes on the data reuse inherent to the algorithm by minimizing on- and off-chip data access and executes a CNN as follows: The weights in a given 3D filter are flattened and embedded into a column of the crossbar array in the form of PCM programming weights. Since the PCM can only absorb light, all the weights are mapped to a value between 0 and 1 over 64 levels (6 bits). Once the array has been programmed, the optical transmitter iterates through all subsets of input data (across all batches), and the optical receiver records complete MAC operations. For any layer, the array will likely have to be programmed multiple times to fully fit all the filters. An accumulator is used at the output of the ADC and deserializer to hold partial sums and add them to new partial sums from the crossbar array. Once complete MAC operations have been computed, they are passed to an activation unit which performs a non-linear transformation to achieve the final output (Fig. 4).

Data can be stored on-chip in SRAM blocks (one each for the input, filters, and output) for rapid and low-power access. SRAM area is estimated to be $0.45mm^2$ per $1MB$ in 45nm CMOS [21], with an access energy of $50fJ/bit$ [21]. On-chip SRAM cannot store the entire dataset and all network parameters for practical CNNs. This necessitates an off-chip DRAM memory, which typically consumes up to $15pJ/b$ [22] and can dominate the total accelerator power. Here, similar to state-of-the-art AI accelerators, we assume that our chip-scale system can be co-packaged with HBM DRAM stack using advanced packaging [23] (fig. 4), which only consumes $3.9pJ/bit$ [22]. In order to minimize DRAM access, data can be sent directly from output SRAM to input SRAM at the end of a full layer computation.

Like many non-volatile memories, the PCM in this work has low programming speed, which can impact crossbar array performance. For instance, PCM programming time can be around $100ns$ [7], which is $1000\times$ slower than our target 10GHz MAC operation speed. Since the array cannot perform any computations during PCM programming, this has the potential to significantly slow down the effective inference/MAC speed. We propose a "dual core" design featuring two copies of the crossbar array (and electrical PCM programming circuits) to mitigate this issue. While one core performs computation, the other core can be programmed with the next set of weights. Assuming that programming the array takes less time than running all input data through it, this effectively hides the programming latency, so that computation runs at all times. We note that this solution can be practical only if the crossbar array forms a small portion of total area. A single laser source, optical transmitter, and optical receiver can be shared among two crossbar cores via optical and electrical switches. We discuss this scheme further in Section VI.

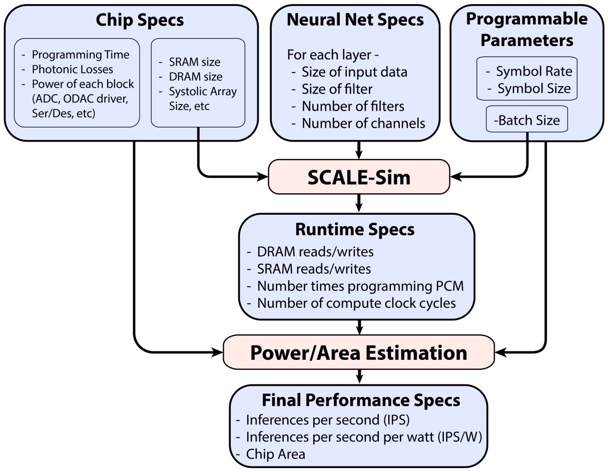

Fig. 5: Simulation framework for estimating IPS and IPS/W.

V. Simulation Framework

We modeled our proposed system architecture comprehensively to measure key performance metrics including inferences per second (IPS), inferences per second per watt (IPS/W), and total chip area based on the 45nm monolithic silicon photonics technology. For this aim, we developed a custom simulation framework based on SCALE-sim [24] as shown in Fig. 5. There are two main steps in estimating the key performance metrics values:

(1) Calculating the runtime specs: This includes the number of SRAM/DRAM accesses, crossbar programming cycles, and number of MAC compute cycles.

(2) Calculating high-level metrics: This is estimated using the runtime specs as well as the specs of all chip components - programming time, photonic losses, ADC power, DRAM access power, etc. (See Sections III and IV)

Calculating runtime specs must be done for a specific neural network and dataset. We achieve this using SCALE-Sim, a cycle-accurate tool for simulating CNN accelerators [24]. We made slight modifications to SCALE-Sim to account for non-unity batch sizes, the presence of an accumulator, and output SRAM data reuse.

In step two of our framework, the number of array programming cycles and MAC compute operation cycles are

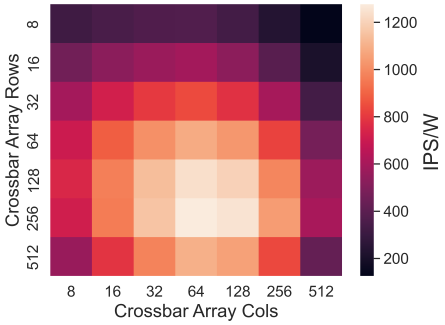

Fig. 6: IPS/W as a function of crossbar rows and columns.

combined with MAC operation speed and programming time to calculate inference time and corresponding IPS. Total chip power is calculated based on clock speed, number of SRAM and DRAM accesses (from step one), plus the energy of optical and electrical components. Chip area is estimated similarly from SRAM sizes, digital accumulator and activation block areas, photonic device footprints, and areas of peripheral electronics, including ADCs, DACs, and clocking.

VI. System-Level Performance Trends and Optimization

Our holistic simulation framework enables us to search the many-dimensional chip design space to find the optimal configuration for ResNet50 inference. We do so by studying the trends in terms of IPS and IPS/W for various design parameters. Here we present observed critical trends and our approach to choosing the optimal design. While we investigate the effect of array size, batch size, dual core scheme, and SRAM size, we hold a constant MAC operation speed of $10GHz$, since the power of many of the peripheral electronics (specifically ADCs) may rise steeply at higher speeds.

A. Trends in Array Performance

Major trends are described below. All figures and data referenced in this section assume the following default chip parameters except when clearly being swept: Array size: 32 x 32, SRAM sizing: 26.3 MB (input), 0.75 MB (output), 0.75 MB (filter), and 0.75 MB (accumulator), dual-core architecture, and batch size: 32.

1) Dual vs. Single Core: Dual core design increases the IPS, but power consumption is also consistently higher since computing and programming happen simultaneously. As a result, IPS/W is the same regardless of the core count.

2) Array Size: IPS always increases approximately linearly with the array size ($N \times M$). This is no surprise because a larger array can process larger matrices/vectors, so less programming and compute cycles are required to reach the final sum. However, the effect of array size on IPS/W is more complicated. As the array size increases, the power of peripheral electronics (such as ODAC drivers or ADCs) grows. Also, higher laser power is required for larger arrays because the power input to each row must spread out to more columns and rows. Both of these effects increase the power less than linearly since many components are unaffected (such as DRAM accesses). However, photonic losses in the crossbar array, such as those due to MMI waveguide crossings or waveguides themselves, grow exponentially. Ultimately this causes power to grow more than linearly with array size, so IPS/W drops. We observe a peak IPS/W at array size of 128-256 rows and 64-128 columns (Fig. 6).

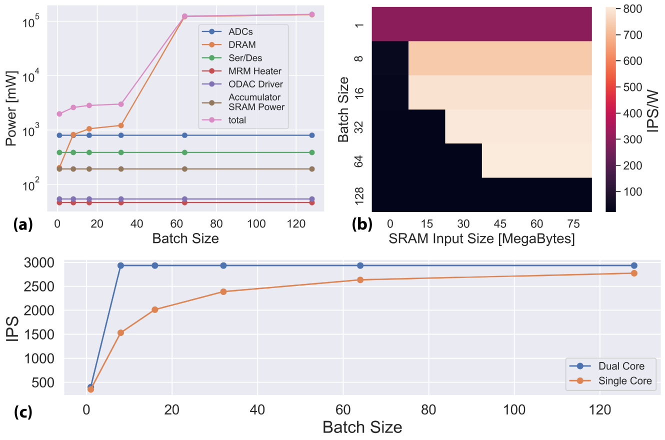

Fig. 7: (a, b) Effect of the SRAM size and batch size on chip power and IPS/W, (c) Effect of dual-core architecture on relationship between batch size and IPS.

3) Batch Size and SRAM Size: Running CNNs with large batch sizes offers the possibility of amortizing the cost of array programming time over a larger set of input data, increasing IPS. However, if not all of the input data across all batches can fit on the input SRAM, data will routinely have to be loaded on and off the chip. This effect can be seen in 7a, where DRAM access energy rises steeply between batch sizes of 32 and 64. For any batch size there is a critical input SRAM size that must be met for optimal IPS/W, and any larger input SRAM size does not significantly increase the performance further (Fig. 7b).

It is important to note that SRAM size cannot be increased indefinitely due to chip area limitations. While smaller batch size is favorable from this perspective because it reduces the need for large SRAM, it can limit IPS. Our dual-core scheme alleviates this tradeoff by hiding the array's programming latency, so IPS becomes only a function of computation time. Fig. 7c shows the large increase in IPS a dual-core architecture can have for small-batch systems.

B. Proposed Approach for Optimization

Building on these observed trends, we develop the following flow for optimizing system performance: Firstly, we find the smallest batch size that is large enough to avoid a low IPS caused by a large programming time that cannot be hidden by our dual core architecture. We do not need to be overly concerned with increasing the batch size beyond this minimal amount because it will have minimal effect on IPS/W, even for infinitely large SRAM (Fig. 7b). Next, we maximize the SRAM size without exceeding a practical chip size ($\sim 1cm^2$ in here), assuming that SRAM dominates chip area, validated in Fig. 8. Finally, we find the the array size that maximizes

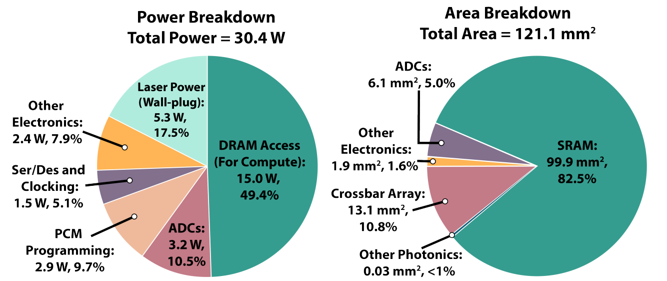

Fig. 8: Power and area breakdown of our proposed accelerator.

IPS/W. When there are multiple array sizes with similar IPS/W, we pick that largest array as it yields a higher IPS, even though exact optimal size may depend on the CNN.

VII. Results and Comparison to State-of-the-Art

Using our optimization framework, we arrive at the following system parameters for optimal performance of our ONN, with dual-core architecture and a $10GHz$ MAC operation speed.

- Array size: 128 rows × 128 columns
- SRAM sizes: 26.3 MB (input), 0.75 MB (output), 0.75 MB (filter), and 0.75 MB (accumulator)
- Batch Size: 32

The performance of this system is compared with the state-of-the-art NVIDIA A100 GPU (INT8 mode with a batch of 128) for ResNet50 in the table below [25]. Our proposed system achieves similar IPS with 15.4× lower power and 7.24× lower area. Overall power and area breakdowns are illustrated in Fig. 8. While the power consumption is dominated by DRAM accesses, the area is mainly dominated by the SRAM blocks. Our system can achieve 327 TOPS and 10.9 TOPS/W.

System	**IPS**	**IPS/W**	**Power**	**Area**
This work	36382	1196	30W	121 mm^2
Nvidia A100	29,733	75	396W	826 mm^2

VIII. Conclusion

We have presented an optical AI accelerator using the coherent crossbar design with on-chip PCM weight storage. The key performance metrics of this design are holistically modeled and estimated using practical and realistic reported results based on 45nm monolithic silicon photonic technology. We have discussed our modeling and system optimization framework. Our approach can be used for optimizing other types of emerging AI accelerators as well. Furthermore, we proposed a dual core system design that can hide programming latency. Our results show that system performances greater than the state-of-the-art can be achieved. This work can be a first step towards designing and optimizing large-scale ONN accelerators for real-world applications by considering all the crucial aspects including SRAM/DRAM memory, electronic-photonic integration, packaging, and all peripheral electronics.

References

[1] D. Amodei and D. Hernandez, "AI and Compute. Blog post, OpenAI," 2018.

[2] Y. Shen, N. C. Harris *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature Photonics*, vol. 11, no. 7, pp. 441–446, 2017.

[3] X. Xu, M. Tan *et al.*, "11 TOPS photonic convolutional accelerator for optical neural networks," *Nature*, vol. 589, no. 7840, pp. 44–51, 2021.

[4] N. Wang, J. Choi *et al.*, "Training Deep Neural Networks with 8-bit Floating Point Numbers," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach *et al.*, Eds., vol. 31. Curran Associates, Inc., 2018.

[5] B. Darvish Rouhani, D. Lo *et al.*, "Pushing the Limits of Narrow Precision Inferencing at Cloud Scale with Microsoft Floating Point," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato *et al.*, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 10 271–10 281.

[6] X. Lin, Y. Rivenson *et al.*, "All-optical machine learning using diffractive deep neural networks," *Science*, vol. 361, no. 6406, pp. 1004–8, 2018.

[7] J. Feldmann, N. Youngblood *et al.*, "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, no. 7840, pp. 52–58, 2021.

[8] N. Youngblood, "Coherent Photonic Crossbar Arrays for Large-Scale Matrix-Matrix Multiplication," *IEEE Journal of Selected Topics in Quantum Electronics*, pp. 1–1, 2022.

[9] R. Hamerly, A. Sludds *et al.*, "Large-Scale Optical Neural Networks based on Photoelectric Multiplication," *ArXiv*, vol. abs/1812.07614, 2019.

[10] M. Rakowski, C. Meagher *et al.*, "45nm CMOS — Silicon Photonics Monolithic Technology (45CLO) for Next-Generation, Low Power and High Speed Optical Interconnects," in *2020 Optical Fiber Communications Conference and Exhibition (OFC)*, 2020, pp. 1–3.

[11] R. Mahajan, R. Sankman *et al.*, "Embedded multi-die interconnect bridge (emib) – a high density, high bandwidth packaging interconnect," in *2016 IEEE 66th Electronic Components and Technology Conference (ECTC)*, 2016, pp. 557–565.

[12] C. Demirkiran, F. Eris *et al.*, "An electro-photonic system for accelerating deep neural networks," *CoRR*, vol. abs/2109.01126, 2021.

[13] Y. Luo, Z. Nong *et al.*, "Low-loss two-dimensional silicon photonic grating coupler with a backside metal mirror," *Opt. Lett.*, vol. 43, no. 3, pp. 474–477, Feb 2018.

[14] A. Zanzi, A. Brimont *et al.*, "Compact and low-loss asymmetrical multimode interference splitter for power monitoring applications," *Opt. Lett.*, vol. 41, no. 2, pp. 227–229, Jan 2016.

[15] Y. Ma, Y. Zhang *et al.*, "Ultralow loss single layer submicron silicon waveguide crossing for soi optical interconnect," *Opt. Express*, vol. 21, no. 24, pp. 29 374–29 382, Dec 2013.

[16] S. Moazeni, S. Lin *et al.*, "A 40-Gb/s PAM-4 Transmitter Based on a Ring-Resonator Optical DAC in 45-nm SOI CMOS," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 12, pp. 3503–3516, 2017.

[17] J. Cardenas, h. A. Morton *et al.*, "Linearized silicon modulator based on a ring assisted Mach Zehnder inteferometer," *Optics Express*, vol. 21, no. 19, p. 22549, 2013.

[18] N. Mehta, S. Lin *et al.*, "A laser-forwarded coherent 10gb/s bpsk transceiver using monolithic microring resonators in 45nm soi cmos," in *2019 Symposium on VLSI Circuits*, 2019, pp. C192–C193.

[19] M. Zhang, Y. Zhu *et al.*, "16.2 a 4× interleaved 10GS/s 8b time-domain ADC with 16× interpolation-based inter-stage gain achieving 37.5dB SNDR at 18GHz input," in *2020 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE, 2020, pp. 252–254.

[20] J. Gu, Z. Wang *et al.*, "Recent advances in convolutional neural networks," *Pattern recognition*, vol. 77, pp. 354–377, 2018.

[21] G. K. Chen, "Power management and sram for energy-autonomous and low-power systems," Ph.D. dissertation, University of Michigan, 2011.

[22] M. O'Connor, N. Chatterjee *et al.*, "Fine-grained DRAM: Energy-efficient DRAM for extreme bandwidth systems," in *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture*, 2017, pp. 41–54.

[23] K. Hosseini, E. Kok *et al.*, "5.12 Tbps Co-Packaged FPGA and Silicon Photonics Interconnect I/O," in *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, 2022, pp. 260–261.

[24] A. Samajdar, Y. Zhu *et al.*, "Scale-sim: Systolic cnn accelerator simulator," *arXiv preprint arXiv:1811.02883*, 2018.

[25] "Nvidia data center deep learning product performance," Sep 2022.