

PICoSNN: Partially Incoherent Configurable Optical Computing Architecture for SNN Acceleration

Bowen Duan¹, Zhenhua Zhu^{1†}, Zhengyang Duan¹, Huazhong Yang¹, Yuan Xie², and Yu Wang^{1†}

¹Tsinghua University, Beijing, China ²HKUST, Hong Kong SAR, China

†Corresponding authors: {zhuzhenhua, yu-wang}@tsinghua.edu.cn

Abstract—Optical computing is becoming a promising solution to meet the growing computational demands of increasingly large-scale deep neural networks (DNNs). However, high power consumption from analog-to-digital (ADC) and digital-to-analog (DAC) conversions poses significant challenges for optical computing. Spiking Neural Networks (SNNs), with their binary spike-based input and output, show the potential to address this issue by reducing the need for high-precision DAC/ADC.

In order to exploit the complementary nature of optical computing and spike-based processing, this paper proposes the Partially Incoherent Configurable Optical Computing Architecture for SNN Acceleration (PICoSNN). We address three critical challenges: phase errors in coherent optical computing, limited configurability in weight-stationary architectures, and inefficient mapping of general SNNs to optical computing hardware. We integrate partially incoherent tensor cores with optical leaky integrate-and-fire neurons, minimizing ADC/DAC overhead while supporting dynamic weight mapping. Further, we propose KV Spiking Self-Attention to enable efficient attention with 1-bit multiplications. Experimental results show that PICoSNN achieves up to 70.54× higher throughput and 8.13× lower energy consumption compared to ASIC implementations, while delivering 15.46× better throughput per area and 17.67× better energy efficiency per area than state-of-the-art photonic accelerators.

I. INTRODUCTION

The rapid growth of Deep Neural Networks (DNNs) [1] requires more efficient hardware than existing digital accelerators [2]. Digital accelerators face challenges related to the physical properties of transistors, which may limit further enhancements in computational speed and energy efficiency [3]. Additionally, the bandwidth and computing parallelism of digital circuits can become bottlenecks when performing specific tasks, such as dense matrix operations [4].

Numerous innovative techniques, beyond traditional digital logic computing, have been explored to overcome existing limitations. Optical computing, with its unique advantages of ultra-high bandwidth and low latency, has emerged as a compelling candidate to surpass the constraints of conventional electronic architectures [3], [4]. Recent advances in silicon photonics have enabled optical implementations of essential linear operations for neural networks, such as matrix multiplications via Mach-Zehnder interferometers (MZIs) [5] and microring resonators (MRRs) [6], [7]. In addition, the Lightning-Transformer (LT) [8] introduces enhanced support for transformer-based networks [9]. However, these analog-domain optical computing systems suffer from a critical bottleneck: the power-hungry analog-to-digital (ADC) & digital-to-analog (DAC) converters required for data conversion, which consume up to 50% of the system’s power and area [8], [10].

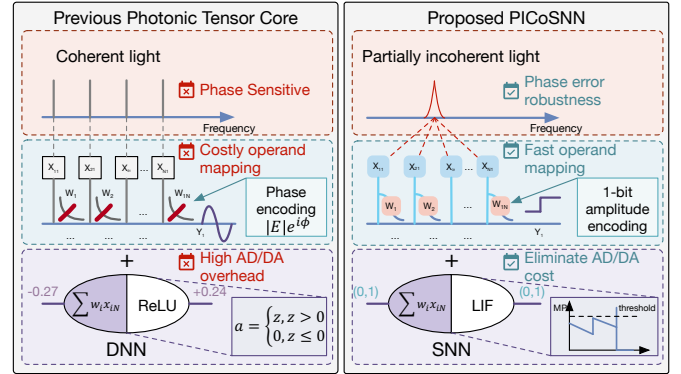


Fig. 1. Comparison between previous photonic accelerators and PICoSNN.

Motivated by the converter’s overhead, we observe that spiking neural networks (SNNs) [11] show promising potential to alleviate the demand for high-precision DACs and ADCs. SNNs use multiple 1-bit spiking activations instead of the continuous real-valued activations used in traditional DNNs, and the outputs of SNNs are binary after being encoded by spiking neurons. This characteristic eliminates the costs associated with ADC/DAC. However, there remain several obstacles to combining these two research fields:

Challenge 1: Phase errors limit the ability of optical computing accelerators to support larger networks. In mainstream coherent optical computing, the phase difference between two input beams of light determines the amplitude and phase of the resulting superimposed output beam. However, optical devices’ phase stability is sensitive to temperature fluctuations, device variations, and waveguide mismatches, inevitably affecting computing accuracy. For example, when the phase error is 0.05 radians, the precision of a 32×32 single-layer neural network can drop by 30~45% [12].

Challenge 2: Limited configurability in optical computing. Most photonic accelerators use the weight-stationary (WS) method for NN tasks [5]–[7]. The optical-based data transfer speed can be ultra-fast, but the speed to update the weights is relatively slow and even non-configurable in some designs [5], which significantly limits further improvements in computational efficiency. Moreover, attention mechanisms, which require dynamic inputs (e.g., Q , K , V), exacerbate these limitations when supporting transformer-based architectures.

Challenge 3: Lack of flexible and generalizable SNN mapping in photonic implementations. Despite advances in photonic neuromorphic computing [13]–[16], current implementations remain limited to specific NN models. Furthermore, the

unsupported operators in the optical domain and data storage necessitate costly optical-electrical conversions. Existing approaches for fixed network architectures [13] lack support for newer paradigms such as transformer-based SNNs, and face challenges in mapping neural weights to optical components (e.g., MZI meshes) [14]. Even with configurable architectures, implementing trained SNNs all-optically requires hardware-algorithm co-design to minimize conversions while supporting diverse and evolving SNN architectures [17], [18].

To overcome these challenges, we propose the Partially Incoherent Configurable Optical Computing Architecture for SNN Acceleration (PICoSNN), a co-design approach spanning the algorithm, core, and architecture levels, with an overall comparison to previous optical accelerators shown in Figure 1. Our contributions include:

(1) At the **algorithm level**, we propose KV Spiking Self-Attention (KV_SSA) to make attention mechanisms optical computing-friendly while maintaining the accuracy of the algorithm. In addition, we propose an equivalent SNN leaky integrate-and-fire (LIF) neuron and prove that it can support linear transformations commonly employed in state-of-the-art (SOTA) SNNs without extra hardware overhead.

(2) At the **core level**, we design the partially incoherent tensor core (PITC) using the partially incoherent (PI) mechanism for increased robustness. PITC addresses the issue of slow weight updates in photonic accelerators. Furthermore, we develop an optical-based LIF to minimize ADC/DAC costs.

(3) At the **architecture level**, we design a reconfigurable hardware architecture to support different bit-width multiplications for both linear layers and attention computation. On the dataflow side, we split outputs into multiple optical channels to enable efficient tiling of large weight matrices.

The experimental results show that PICoSNN outperforms SOTA ASIC and photonic accelerators in terms of throughput, energy efficiency, and silicon footprint. Specifically, it delivers $15.46\times$ better throughput per area and $17.67\times$ better energy efficiency per area than the previous SOTA photonic design.

II. PRELIMINARIES AND BACKGROUND

A. SNNs Background

SNNs, inspired by biological neurons, offer improved energy efficiency over DNNs [17]–[24]. Unlike DNNs that use multi-bit activations, SNNs process information temporally using single-bit spikes generated by spiking neurons over time [11], as shown in Figure 1. When the neuron receives spikes, the corresponding input weight is added to the membrane potential (MP). Upon the potential exceeding a certain threshold, a spike is fired at that time step, and the potential is reset. The Leaky-Integrate-and-Fire (LIF) neuron model [25] is used by most SNNs [17], [18]. A LIF neuron in layer l can be described as:

$$H_l(t) = \alpha V_l(t) + \beta (X_l(t) + V_{reset}),$$

$$V_l(t+1) = \begin{cases} V_{reset}, & \text{if } H_l(t) > V_{th}, \\ H_l(t), & \text{else;} \end{cases} \quad S_l(t) = \begin{cases} 1, & \text{if } H_l(t) > V_{th}, \\ 0, & \text{else.} \end{cases} \quad (1)$$

where $H_l(t)$ is the membrane potential (MP) of the LIF neuron at time t , $V_l(t)$ represents the residual MP from the previous time ($t - 1$), $\tau \in (0, 1)$ serves as the leakage parameter indicating potential decay, $X_l(t)$ denotes the input computed via convolution, linear, or attention layers, $S_l(t)$ is the output spike, and V_{reset} along with V_{th} is the reset and threshold voltages, respectively. The LIF neuron provides non-linearities for SNNs, and its properties make the combination of SNNs and optical computing possible.

B. Optical Computing Basics

Basic Photonic Devices. Key photonic devices for optical computing include: the Mach-Zehnder Interferometer (MZI), which acts as a fundamental building block for switches and transformations by controlling phase shifts; the Mach-Zehnder Modulator (MZM), which uses the MZI principle to encode data onto light; the Electro-Absorption Modulator (EAM), which controls light intensity via an electric field; and the Microring Resonator (MRR), a compact structure that functions as a wavelength-selective filter or modulator.

Prior Photonic Architecture. Optical computing has traditionally relied on coherent light sources [26]. Key implementations include MZI-based circuits [5], MRR arrays [6], [7], and diffractive networks [27]. Recent photonic SNN advances feature all-optical implementations with phase-change materials [13] and nanophotonic-electronic hybrids [14]. These architectures demonstrate various spiking mechanisms, but face challenges in mapping complex SNN topologies to optical hardware [17]–[19].

III. OPTIMIZED SNN FOR OPTICAL COMPUTING

SNNs exhibit binary inputs and outputs that align well with optical computing and promise to reduce the significant overhead of ADC and DAC. However, deploying SNNs on photonic hardware faces three main challenges: being unable to store the partial sum results before passing through the spiking neuron, the influence of device noise on weight accuracy, and the energy loss of the device to light. To address these challenges, we enhance the spiking attention mechanism for optical computing and quantize model weights. We also demonstrate that scaling factors and biases can be integrated into neurons. This integration supports quantization and helps mitigate the effects of light decay in the devices.

A. KV Spiking Self Attention

The standard spiking self-attention (SSA) mechanism [17], [18], defined as $SSA = \mathcal{SN}((QK^T)V) * s_a = \mathcal{SN}(Q(K^T V) * s_a)^1$, is difficult to implement efficiently in photonic hardware. The primary issue is that the intermediate matrix product, such as $K^T V$, is an analog value that does not pass through a spiking neuron. Storing or pipelining such analog results between optical computing stages is inefficient and costly.

¹ $\mathcal{SN}(\cdot)$ is the spiking neuron, and s_a is the scaling factor of attention

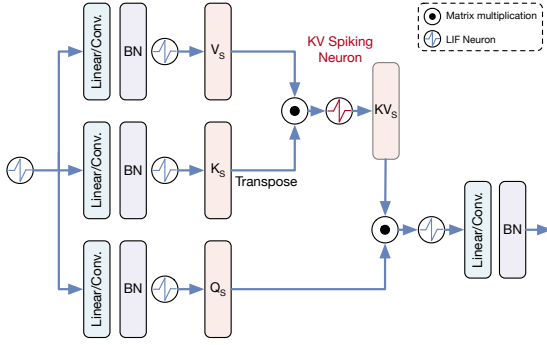


Fig. 2. Spiking self-attention with KV spiking neuron.

To address this, we propose KV Spiking Self-Attention (KV_SSA), which introduces a spiking neuron after the initial matrix multiplication, as shown in Figure 2:

$$\text{KV_SSA}(Q, K, V) = \mathcal{N}(Q \cdot \mathcal{N}(K^T V) * s_a). \quad (2)$$

This modification makes SNN hardware-friendly by converting the intermediate analog result into a binary spike matrix that can be efficiently processed in the next computation. Our experiments show that KV_SSA causes a negligible accuracy drop within 1%. The justification is grounded in two aspects:

(1) $(K^T V)$ generates a feature relevance matrix by weighting the aggregation of features (V) based on their relevance (K). The spiking neuron acts as a filter, suppressing the less important $(K^T V)$ values to zero. This enables denoising and feature extraction while preserving key information for subsequent attention calculations. In addition, $\mathcal{N}(\cdot)$ introduces non-linearity that compensates for the removal of the softmax function in the attention.

(2) Compared to the original SSA, KV_SSA retains sufficient parameters and learnable nonlinearities to approximate a functionally equivalent transformation. The computational complexity also remains linear with respect to the sequence length.

B. Equivalent LIF Neuron

Linear transformations $(X'(t) = (X(t) - b)/scale)$ play an important role in optical-based SNN computing. On the one hand, b and $scale$ can be used to model optical device non-ideal effects like signal decay (insertion loss) and noise. On the other hand, they can also capture algorithmic operators such as batch normalization and data quantization. We propose an equivalent LIF neuron that absorbs the linear transformation into the LIF without extra hardware overhead.

Theorem III.1. Consider the original LIF model defined by Equation 1, after input transformation $X'(t) = \frac{X(t) - b}{scale}$, setting

$$V_{th}^{new} = \frac{V_{th} - (\alpha + \beta)(b + V_{reset})}{scale} \quad (3)$$

ensures equivalent spiking behavior.

Proof. To establish equivalent spiking conditions, we introduce the relationship $V(t) = scale V'(t) + (b + V_{reset})$.

Substituting this into the original integrated potential:

$$\begin{aligned} H(t) &= \alpha \left(scale V'(t) \right. \\ &\quad \left. + (b + V_{reset}) \right) + \beta \left(scale X'(t) + b + V_{reset} \right) \\ &= scale \left(\alpha V'(t) + \beta X'(t) \right) + (\alpha + \beta)(b + V_{reset}) \\ &= scale H'(t) + (\alpha + \beta)(b + V_{reset}) \end{aligned} \quad (4)$$

The original spiking condition $H(t) > V_{th}$ becomes:

$$\begin{aligned} scale H'(t) + (\alpha + \beta)(b + V_{reset}) &> V_{th} \\ H'(t) &> \frac{V_{th} - (\alpha + \beta)(b + V_{reset})}{scale} \end{aligned} \quad (5)$$

Therefore, setting $V_{th}^{new} = (V_{th} - (\alpha + \beta)(b + V_{reset}))/scale$ makes $H'(t) > V_{th}^{new}$ equivalent to $H(t) > V_{th}$. Thus, with this threshold, both models produce identical spiking patterns. \square

Theorem III.1 shows that we can implement more complex computations in optical computing, e.g., scaling factors for quantization, attention mechanisms, and biases in linear layers, by adjusting the threshold of the photo detector (PD) in the optical LIF neuron. Besides, it demonstrates the robustness of the LIF neuron against decay and noise in optical devices, further confirming the suitability of SNNs for optical computing.

IV. PROPOSED PICOSSN ACCELERATOR ARCHITECTURE

To implement the proposed KV_SSA algorithm and fully leverage the advantages of SNN, we design the PICOSSN. We first introduce the overall architecture of PICOSSN, then describe its components, and finally explain the overall data flow.

A. Overall Architecture Design

The PICOSSN architecture, shown in Figure 3(a), contains two partially incoherent tensor cores (PITCs), an optical LIF neuron array, and other supporting electronic modules for data storage and control. The light source is used to drive the tensor core operation. Input activation values and weights are stored in the buffer and mapped to the array via modulators (MZM, EAM), avoiding the use of DACs. The light passes through the modulator and the waveguide to get the result of the matrix multiplication, then is converted into a digital format for storage after passing through the LIF unit.

B. Partially Incoherent Tensor Core

Partially incoherent optical computing operates on the principle of using light with reduced coherence to perform calculations by **summing light intensity**, while existing coherent optical computing relies on the controlled interference of **phase-sensitive light waves** and necessitates a unique wavelength for each input channel. Therefore, our solution makes the system robust against phase errors [28], [29], as illustrated in Figure 1.

Building on the concept of partially incoherent optical computing, we propose the PITC, depicted in Figure 3(a). We first enable PITC to compute the 1-bit matrix-vector multiplication between the Q , K , and V matrices in KV_SSA. We modulate the input X using the MZM after splitting the source light. The modulated light is then directed into the computational

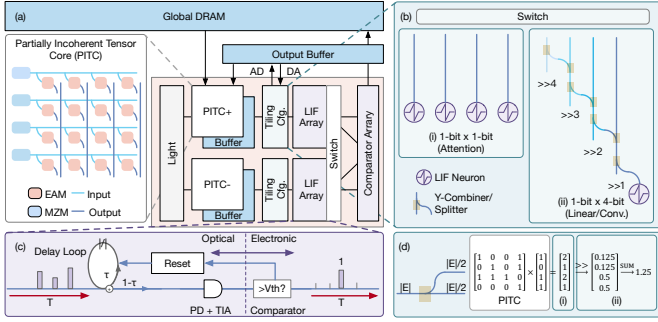


Fig. 3. (a) The proposed PICoSNN architecture. (b) The precision-configurable compute unit. (c) The proposed optical LIF neuron. (d) How PITC supports both 1-bit \times 4-bit and 1-bit \times 1-bit multiplication through the configurable compute architecture.

array, where the weight W is modulated onto the array using the EAM. Thus, the modulator functions like a switch: a ‘1’ in the weights indicates that the light is allowed to pass, while a ‘0’ means that the light is blocked. We employ MZM to modulate X because it supports higher frequency modulation. In contrast, we use EAM to modulate W since W can be reused and updated at a lower frequency, making EAM more energy-efficient.

To handle negative weights in convolutional and linear layers, PICoSNN employs two PITCs (PITC+ for positive weights, PITC- for negative weights) that process inputs in parallel. The results passing through separate LIF neurons before a differential comparator combine them to generate the final output. Two PITCs can simultaneously compute positive Q , K , and V matrices for the proposed KV_SSA mechanisms.

C. Precision-Configurable Compute Unit

The PITC is capable of supporting matrix multiplications used in SNNs. Figure 3(b) illustrates our precision-configurable architecture that accommodates both convolutional and linear layer computations. We implement multi-bit computations by splitting and combining light. The light splitter reduces the light energy by half, akin to a right-shift operation. Then the light combiner adds the attenuated light from the left column to that of the current column. Figure 3(b) shows an example of a 4-bit unsigned integer representation: $2^4 \cdot (a \cdot 2^{-4} + b \cdot 2^{-3} + c \cdot 2^{-2} + d \cdot 2^{-1})$, where a, b, c, d can be either 0 or 1. This representation allows us to obtain the corresponding result by modulating the bits from left to right onto the PITC. The process begins by summing the bits of the weight that correspond to ‘1’s in the input and then performing shifts and summation to produce the output.

The example in Figure 3(d) demonstrates the precision-configurable process. Part (i) illustrates the fundamental PITC operation, which involves a matrix multiplication with 1-bit inputs. The resulting output from this stage is directly applicable for the 1-bit \times 1-bit KV_SSA calculations shown in Figure 2. For the multi-bit operations required in linear and convolutional layers, part (ii) depicts the subsequent processing steps. The analog outputs from the PITC columns in part (i), each corresponding to a different bit-plane of the

4-bit weights, are individually scaled by powers of two and then summed together to produce the final result.

D. Optical LIF Neurons

After obtaining the corresponding output using PITC (where the energy of the light corresponds to the result), we direct the light to the optical LIF neuron. The structure of the optical LIF neuron is shown in Figure 3(c).

At the time step $t=0$, the input $X(0)$ is first combined with V_{reset} , whose energy is provided by the MZM responsible for the reset operation. The light is then split by the waveguide, with a proportion τ entering the delay loop, while the remaining portion $(1-\tau)$ leads to the photo diode (PD). At $t=0$, we have $V(0) = 0$, hence the light portion directed to the PD is exactly $H(0) = (1-\tau)(X(0) + V_{reset})$, as shown in Equation (1). The energy in the delay loop is:

$$V_{delay} = \tau(X(0) + V_{reset}) = \frac{\tau}{1-\tau}H(0) = \frac{\tau}{1-\tau}V(1). \quad (6)$$

At the next time step, $t=1$, the energy in the delay loop is delayed and added to the V_{reset} provided by the modulator, along with the input $X(1)$. This results in the total energy at the next time step before splitting the light: $V_{delay} + X(1) + V_{reset}$. After the light is split, the total energy leading to the PD is given by:

$$\begin{aligned} H(1) &= (1-\tau)(V_{delay} + X(1) + V_{reset}) \\ &= (1-\tau) \left(\frac{\tau}{1-\tau}V(1) + X(1) + V_{reset} \right) \\ &= \tau V(1) + (1-\tau)(X(1) + V_{reset}), \end{aligned} \quad (7)$$

which is consistent with Equation (1). For subsequent time steps, the result follows the same pattern. The PD and the transimpedance amplifier (TIA) convert the light intensity to a voltage that is proportional to it. The comparator compares the voltage with the threshold V_{th} to determine whether to output a digital spike. When the comparator outputs a spike, the modulator receives the spiking signal simultaneously and modulates the energy in the delay loop back to V_{reset} .

The optical LIF neuron only outputs a binary sequence of 0s and 1s, taking the place of traditional ADC devices for optical-to-electrical conversion. Thus, it leads to reduced power consumption and enhanced robustness and also showcases the adaptability of SNN to optical computing architectures.

E. Data flow and Tiling

For the matrix multiplications shown in Figure 4(a), we adopt the weight-stationary (WS) data flow with tiling across the M , N , and K dimensions and prioritize the tiling in the K dimension for large-scale SNNs. For matrices exceeding the dimensions of the EAM array, partial sums must be stored and accumulated, typically requiring ADC/DAC conversions. PICoSNN minimizes this overhead by meticulously designing tiling and pipeline strategies.

Figure 4(b) shows how PICoSNN supports matrix multiplication with a K dimension four times larger than the array size. With a $k \times 4$ array and a weight matrix $K \times N$ where $K = 4k$, we process the calculation in four cycles. In the first cycle, the array uses the first column to compute the multiplication result of the first k numbers from the first row of the input matrix

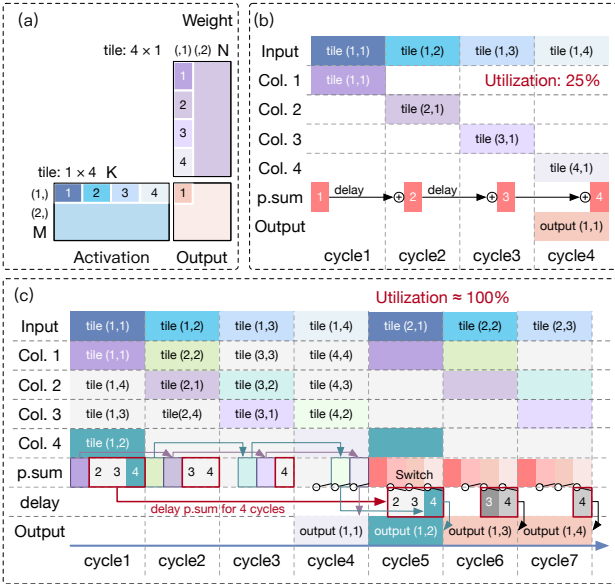


Fig. 4. Data flow and tiling strategy. (a) Tiling the input and weight matrix in the K dimension. (b) The proposed tiling strategy supports larger matrices. (c) The proposed pipeline design for improving utilization.

and the first column of the weight matrix (referred to as tile 1). In cycle 2, the result of tile 1 is delayed by one cycle due to the delay loop and is then added to the result of tile 2 computed by column 2. In cycle 4, the results of all four tiles are accumulated through delay loops. So the array utilization is only 25%. If we want to use all columns simultaneously, we need to store the results from the other columns.

We improve the array utilization with a pipelining scheme shown in Figure 4(c), allowing all four columns to compute in parallel. After cycle 1, the results from the last three columns are delayed by 4 cycles. In cycle 5, the result of tile (1,2) of the second weight column (green one) is selected and added through a switch from the last column of the array to obtain the second output. Simultaneously, results completed in column 4 during cycle 5 are delayed by four cycles, preventing conflicts at the partial summation unit with tile (1,2). The optimized pipelining scheme increases the utilization to nearly 100%.

V. EVALUATION

A. Methodology

Evaluation models and datasets. We assess our PICoSNN across different SNNs, such as spiking convolutional neural networks (SCNN) ResNet [30], along with spiking transformers like Spikformer [17] and SDT-V2 [18]. All models follow the settings from their original papers regarding the number of layers, dimensions, and time steps. We run these models on tasks including CIFAR10, CIFAR100 [31], CIFAR10-DVS [32], and ImageNet [33]. We sourced all models from their original open-source repositories, trained them with PyTorch to replicate the reported accuracies, and then fine-tuned them with Learned Step Size Quantization (LSQ) [34] and our proposed KV Spiking Self-Attention.

System setup. We developed a simulator based on validated open-source optical computing simulation code [8] to evaluate

TABLE I
ADOPTED DEVICES IN OUR PICoSNN. MODIFIED FROM [8]

Device	Specification
DAC [8]	Precision: 8-bit; Area: $11,000 \mu\text{m}^2$ Power: 50 mW (@14 GSPS)
ADC [8]	Precision: 8-bit; Area: $2,850 \mu\text{m}^2$ Power: 14.8 mW (@10 GSPS)
TIA [8]	Power: 3 mW; Area: $50 \mu\text{m}^2$
MZM [8]	Tuning Power: 2.25 mW IL: 1.2 dB; Area: $260 \times 20 \mu\text{m}^2$
EAM [36]	Tuning Power: 0.58 mW IL: 5 dB; Area: $0.6 \times 50 \mu\text{m}^2$
Photodetector [8]	Sensitivity: -25 dBm Power: 1.1 mW; Area: $4 \times 10 \mu\text{m}^2$
Y-branch [8]	IL: 0.3 dB; Area: $1.8 \times 1.3 \mu\text{m}^2$
On-chip Laser [8]	Wall-plug Efficiency: 0.2 Area: $400 \times 300 \mu\text{m}^2$
Comparator [37]	Power: 2.2 mW; Area: $78 \mu\text{m}^2$
Optical Switch [38]	Tuning Power: 0.1 mW IL: 1.04 dB; Area: $18 \times 200 \mu\text{m}^2$

TABLE II
PICoSNN ARCHITECTURE SETUP.

Parameter	Value
PITC Array Size	$k = 144, n = 32$
On-chip Buffer Size	18KB Activation; 144KB Weight; 128 KB Partial Sum
HBM DRAM	1TB/s
Clock Frequency	5GHz

latency, power, and area. We used CACTI 7.0 (scaled to 14 nm [35]) for on-chip buffers and DRAMsim3 for off-chip DRAM. ADC/DAC parameters were sourced from comparable 14/16 nm technology nodes. Table I lists the photonic device parameters used in the simulation, and Table II details our architecture's configuration.

Baselines. We compare PICoSNN against state-of-the-art ASIC SNN accelerators (Spiking Eyeriss [2], PTB [21], SATO [22], Prosperity [23]) and photonic accelerators (Lightning-Transformer (LT) [8], MRR bank [6], MZI array [5]). We simulated the ASIC accelerators at 500 MHz and all photonic systems, including our own, at a conservative 5 GHz to highlight the throughput advantage of optical computing. To ensure a fair comparison, all photonic baselines were simulated using the same device parameters (Table I), SRAM, and DRAM technology as PICoSNN.

B. PICoSNN Evaluation Results

Figure 5 shows PICoSNN's efficiency compared to baselines in throughput and energy efficiency.

Performance. Photonic accelerators generally outperform ASICs due to higher clock frequencies and parallelism. Among photonic baselines, LT is the strongest due to its support for higher parallelism and dynamic inputs. PICoSNN achieves

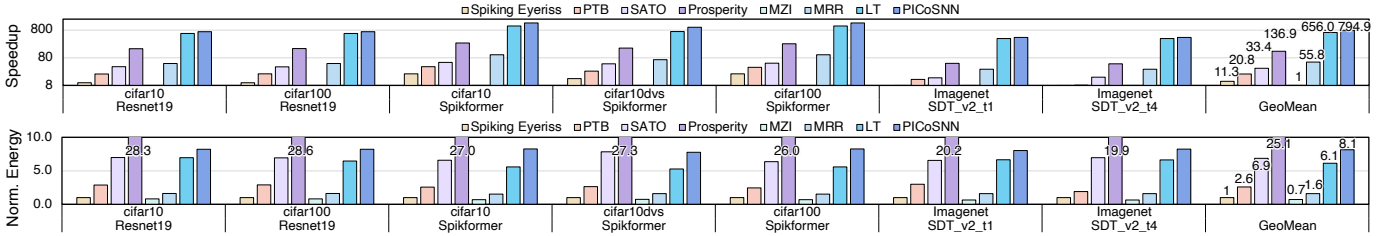


Fig. 5. Speedup (normalized by MZI) and Energy Efficiency (normalized by Spiking Eyeriss) of PICoSNN and baselines.

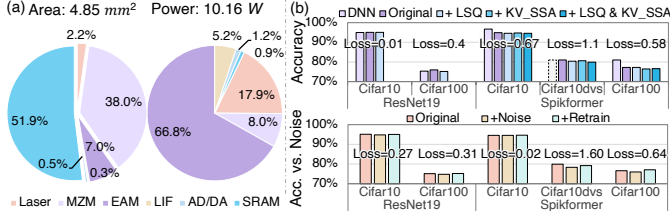


Fig. 6. (a) PICoSNN's power and area breakdown. (b) PICoSNN's accuracy (Acc.) results.

TABLE III
AREA AND EFFICIENCY COMPARISON FOR PICoSNN AND PHOTONIC BASELINES.

Area (mm^2) (%)	MZI	MRR	LT	PICoSNN
Core	20.55 (33.65%)	11.09 (17.98%)	28.12 (46.64%)	1.90 (40.22%)
Memory	14.90 (24.40%)	26.08 (42.29%)	14.70 (24.37%)	2.80 (59.20%)
AD/DA	25.62 (41.95%)	24.50 (39.73%)	17.48 (28.99%)	0.03 (0.59%)
Total	61.07	61.67	60.30	4.73
Normalized Throughput/Area Efficiency	1.00 \times	55.24 \times	664.34 \times	10268.72 \times
Normalized Energy/Area Efficiency	1.00 \times	2.18 \times	8.60 \times	151.98 \times

speedups ranging from $1.21\times$ to $794.95\times$ against all baselines. This significant performance gain stems from our partially incoherent design, which further enhances parallelism and supports dynamic inputs on both sides, and our integrated optical LIF unit, which eliminates the ADC/DAC latency that bottlenecks other photonic systems.

Energy. PICoSNN is more energy-efficient than most baselines, achieving up to an $8.13\times$ reduction compared to Spiking Eyeriss and a $1.33\times$ reduction against LT. This is primarily due to our architecture minimizing ADC/DAC overhead. However, Prosperity remains $3.1\times$ more efficient because its design is highly optimized for extreme activation sparsity, a domain where optical computing still has room for future exploration.

Area and Power Breakdown. Figure 6(a) illustrates the breakdown of PICoSNN's area and power consumption. The ADC/DAC has a relatively small area and power usage, and the LIF, which serves as a substitute for the ADC, also occupies minimal space and power. Table III provides a

comparison of the area breakdown between PICoSNN and the photonic accelerator baselines, which include core, memory, and AD/DA converter components. PICoSNN attains a $15.46\times$ improvement in normalized throughput/area efficiency and an impressive $17.67\times$ improvement in normalized energy and area efficiency compared to the SOTA baseline LT.

Model Accuracy. Figure 6(b) illustrates the accuracy results of our algorithm design, which includes KV_SSA and LSQ quantization. It shows that our algorithm modifications maintain the accuracy loss within 1%, while ensuring that the SNN model is applicable to PICoSNN. We also tested the impact of device noise on model accuracy, employing the same methodology as LT [8]. The results indicate that device noise affects accuracy by approximately 1%, but this can be recovered through further training.

VI. CONCLUSION

This paper presented PICoSNN, a novel photonic accelerator specifically designed for spiking neural networks. By leveraging partially incoherent light for reduced phase sensitivity and developing optical LIF neurons to minimize ADC/DAC overhead, our architecture achieves significant advantages in throughput, energy efficiency, and area utilization compared to existing ASIC and photonic implementations. Our experiments show that PICoSNN achieves up to $70.54\times$ higher throughput and $8.13\times$ lower energy consumption compared to conventional ASIC designs. These results highlight the promising potential of combining optical computing with spiking neural networks for next-generation AI hardware, particularly for transformer-based architectures requiring high-throughput, energy-efficient inference capabilities. Future research will incorporate more refined physical noise models and hardware-in-the-loop calibration to further validate its robustness.

VII. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (62504139, 62325405, U24B6015), Beijing Natural Science Foundation (L257010), the National Key R&D Program of China (2023YFB4502200), Beijing National Research Center for Information Science, Technology (No. BNR2024TD03001), Beijing Innovation Center for Future Chips, and State Key Laboratory of Space Network and Communications. This research was partially supported by ACCESS – AI Chip Center for Emerging Smart Systems, sponsored by InnoHK funding, Hong Kong SAR (HKSAR) and Research Grants Council of HKSAR (16213824).

REFERENCES

- [1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [2] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE journal of solid-state circuits*, vol. 52, no. 1, pp. 127–138, 2016.
- [3] T. Fu, J. Zhang, R. Sun, Y. Huang, W. Xu, S. Yang, Z. Zhu, and H. Chen, "Optical neural networks: progress and challenges," *Light: Science & Applications*, vol. 13, no. 1, p. 263, 2024.
- [4] B. J. Shastri, A. N. Tait, T. Ferreira de Lima, W. H. Pernice, H. Bhaskaran, C. D. Wright, and P. R. Prucnal, "Photonics for artificial intelligence and neuromorphic computing," *Nature Photonics*, vol. 15, no. 2, pp. 102–114, 2021.
- [5] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature photonics*, vol. 11, no. 7, pp. 441–446, 2017.
- [6] A. N. Tait, T. F. De Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," *Scientific reports*, vol. 7, no. 1, p. 7430, 2017.
- [7] F. Sunny, A. Mirza, M. Nikdast, and S. Pasricha, "Crosslight: A cross-layer optimized silicon photonic neural network accelerator," in *2021 58th ACM/IEEE design automation conference (DAC)*. IEEE, 2021, pp. 1069–1074.
- [8] H. Zhu, J. Gu, H. Wang, Z. Jiang, Z. Zhang, R. Tang, C. Feng, S. Han, R. T. Chen, and D. Z. Pan, "Lightening-transformer: A dynamically-operated optically-interconnected photonic transformer accelerator," in *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2024, pp. 686–703.
- [9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [10] F. P. Sunny, E. Taheri, M. Nikdast, and S. Pasricha, "A survey on silicon photonics for deep learning," *ACM Journal of Emerging Technologies in Computing System*, vol. 17, no. 4, pp. 1–57, 2021.
- [11] A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, and A. Maida, "Deep learning in spiking neural networks," *Neural networks*, vol. 111, pp. 47–63, 2019.
- [12] A. Shafiee, S. Banerjee, K. Chakrabarty, S. Pasricha, and M. Nikdast, "Analysis of optical loss and crosstalk noise in mzi-based coherent photonic neural networks," *Journal of lightwave technology*, 2024.
- [13] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," *Nature*, vol. 569, no. 7755, pp. 208–214, 2019.
- [14] L. El Srouji, Y.-J. Lee, M. B. On, L. Zhang, and S. B. Yoo, "Scalable nanophotonic-electronic spiking neural networks," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 29, no. 2: Optical Computing, pp. 1–13, 2022.
- [15] M. A. Nahmias, B. J. Shastri, A. N. Tait, and P. R. Prucnal, "A leaky integrate-and-fire laser neuron for ultrafast cognitive computing," *IEEE journal of selected topics in quantum electronics*, vol. 19, no. 5, pp. 1–12, 2013.
- [16] J. Robertson, D. Black, G. Donati, Q. R. A. Al-Taai, E. Malysheva, B. Romeira, J. Figueiredo, V. D. Calzadilla, E. Wasige, and A. Hurtado, "Ultrafast and compact photonic-electronic leaky integrate-and-fire circuits based upon resonant tunnelling diodes," *arXiv preprint arXiv:2501.17133*, 2025.
- [17] Z. Zhou, Y. Zhu, C. He, Y. Wang, S. Yan, Y. Tian, and L. Yuan, "Spikformer: When spiking neural network meets transformer," *arXiv preprint arXiv:2209.15425*, 2022.
- [18] M. Yao, J. Hu, T. Hu, Y. Xu, Z. Zhou, Y. Tian, B. Xu, and G. Li, "Spike-driven transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips," *arXiv preprint arXiv:2404.03663*, 2024.
- [19] M. Yao, J. Hu, Z. Zhou, L. Yuan, Y. Tian, B. Xu, and G. Li, "Spike-driven transformer," *Advances in neural information processing systems*, vol. 36, pp. 64 043–64 058, 2023.
- [20] S. Narayanan, K. Taht, R. Balasubramonian, E. Giacomini, and P.-E. Gaillardon, "Spinalflow: An architecture and dataflow tailored for spiking neural networks," in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2020, pp. 349–362.
- [21] J.-J. Lee, W. Zhang, and P. Li, "Parallel time batching: Systolic-array acceleration of sparse spiking neural computation," in *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2022, pp. 317–330.
- [22] F. Liu, W. Zhao, Z. Wang, Y. Chen, T. Yang, Z. He, X. Yang, and L. Jiang, "Sato: spiking neural network acceleration via temporal-oriented dataflow and architecture," in *Proceedings of the 59th ACM/IEEE Design Automation Conference*, 2022, pp. 1105–1110.
- [23] C. Wei, C. Guo, F. Cheng, S. Li, H. F. Yang, H. H. Li, and Y. Chen, "Prosperity: Accelerating spiking neural networks via product sparsity," in *2025 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2025, pp. 806–820.
- [24] C. Wei, B. Duan, C. Guo, J. Zhang, Q. Song, H. Li, and Y. Chen, "Phi: Leveraging pattern-based hierarchical sparsity for high-efficiency spiking neural networks," in *Proceedings of the 52nd Annual International Symposium on Computer Architecture*, 2025, pp. 930–943.
- [25] M. J. Pearson, A. G. Pipe, B. Mitchinson, K. Gurney, C. Melhuish, I. Gilhespy, and M. Nibouche, "Implementing spiking neural networks for real-time signal-processing and control applications: A model-validated fpga approach," *IEEE Transactions on Neural Networks*, vol. 18, no. 5, pp. 1472–1487, 2007.
- [26] T. H. Maiman, "Stimulated optical radiation in ruby," *nature*, vol. 187, no. 4736, pp. 493–494, 1960.
- [27] H. Zhu, J. Zou, H. Zhang, Y. Shi, S. Luo, N. Wang, H. Cai, L. Wan, B. Wang, X. Jiang *et al.*, "Space-efficient optical computing with an integrated chip diffractive neural network," *Nature communications*, vol. 13, no. 1, p. 1044, 2022.
- [28] C. Bourassin-Bouchet and M.-E. Couprie, "Partially coherent ultrafast spectrography," *Nature communications*, vol. 6, no. 1, p. 6465, 2015.
- [29] B. Dong, F. Brücknerhoff-Plückelmann, L. Meyer, J. Dijkstra, I. Bente, D. Wendland, A. Varri, S. Aggarwal, N. Farnakidis, M. Wang *et al.*, "Partial coherence enhances parallelized photonic computing," *Nature*, vol. 632, no. 8023, pp. 55–62, 2024.
- [30] W. Fang, Z. Yu, Y. Chen, T. Huang, T. Masquelier, and Y. Tian, "Deep residual learning in spiking neural networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 056–21 069, 2021.
- [31] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.
- [32] H. Li, H. Liu, X. Ji, G. Li, and L. Shi, "Cifar10-dvs: an event-stream dataset for object classification," *Frontiers in neuroscience*, vol. 11, p. 309, 2017.
- [33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [34] S. K. Esser, J. L. McKinstry, D. Bablani, R. Appuswamy, and D. S. Modha, "Learned step size quantization," *arXiv preprint arXiv:1902.08153*, 2019.
- [35] A. Stillmaker and B. Baas, "Scaling equations for the accurate prediction of cmos device performance from 180 nm to 7 nm," *Integration*, vol. 58, pp. 74–81, 2017.
- [36] J. Liu, D. Pan, S. Jongthammanurak, K. Wada, L. C. Kimerling, and J. Michel, "Design of monolithically integrated gesi electro-absorption modulators and photodetectors on an soi platform," *Optics Express*, vol. 15, no. 2, pp. 623–628, 2007.
- [37] A. T. Ramkaj, M. S. Steyaert, and F. Tavernier, "A 13.5-gb/s 5-mv-sensitivity 26.8-ps-clk-out delay triple-latch feedforward dynamic comparator in 28-nm cmos," in *ESSCIRC 2019-IEEE 45th European Solid State Circuits Conference (ESSCIRC)*. IEEE, 2019, pp. 167–170.
- [38] H. Zhong, J. Li, Y. He, R. Zhang, H. Wang, J. Shen, Y. Zhang, and Y. Su, "Ultra-low-power consumption silicon electro-optic switch based on photonic crystal nanobeam cavity," *npj Nanophotonics*, vol. 1, no. 1, p. 33, 2024.