

TeMPO: Efficient Time-Multiplexed Dynamic Photonic Tensor Core for Edge AI with Compact Slow-Light Electro-Optic Modulator

Meng Zhang,^{1, a)} Dennis Yin,^{2, a)} Nicholas Gangi,¹ Amir Begović,¹ Alexander Chen,¹ Zhaoran Rena Huang*,¹ and Jiaqi Gu*²

¹⁾*Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA*

²⁾*School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287, USA*

(*Electronic mail: jiaqigu@asu.edu.)

(*Electronic mail: huangz3@rpi.edu)

(Dated: 13 February 2024)

Electronic-photonic computing systems offer immense potential in energy-efficient artificial intelligence (AI) acceleration tasks due to the superior computing speed and efficiency of optics, especially for real-time, low-energy deep neural network (DNN) inference tasks on resource-restricted edge platforms. However, current optical neural accelerators based on foundry-available devices and conventional system architecture still encounter a performance gap compared to highly customized electronic counterparts. To bridge the performance gap due to lack of domain specialization, we present a time-multiplexed dynamic photonic tensor accelerator, dubbed TeMPO, with cross-layer device/circuit/architecture customization. At the device level, we present foundry-compatible, customized photonic devices, including a slow-light electro-optic modulator with experimental demonstration, optical splitters, and phase shifters that significantly reduce the footprint and power in input encoding and dot-product calculation. At the circuit level, partial products are hierarchically accumulated via parallel photocurrent aggregation, lightweight capacitive temporal integration, and sequential digital summation, considerably relieving the analog-to-digital conversion bottleneck. We also employ a multi-tile, multi-core architecture to maximize hardware sharing for higher efficiency. Across diverse edge AI workloads, TeMPO delivers digital-comparable task accuracy with superior quantization/noise tolerance. We achieve a 368.6 TOPS peak performance, 22.3 TOPS/W energy efficiency, and 1.2 TOPS/mm² compute density, pushing the Pareto frontier in edge AI hardware. This work signifies the power of cross-layer co-design and domain-specific customization, paving the way for future electronic-photonic accelerators with even greater performance and efficiency.

I. INTRODUCTION

Photonic computing has emerged as a promising technology for high-performance and energy-efficient computing, particularly in computation-intensive artificial intelligence (AI) tasks. Various integrated photonic tensor core (PTC) designs have been introduced and demonstrated for ultra-fast photonic analog linear operation acceleration. Coherent PTCs that leverage interference and diffraction include MZI arrays¹, butterfly-style meshes^{2,3}, auto-designed photonic circuits⁴, coupler-crossbar array⁵, star-coupler-based design⁶, and metalens-based diffractive PTCs⁷, etc. Besides, to leverage the wavelength-division multiplexing (WDM) technique, there are incoherent multi-wavelength PTCs, e.g., MRR weight bank^{8–11}, PCM crossbar arrays¹², micro-comb-based computing engine^{13,14}. We emphasize three key features of efficient PTCs required by general edge AI from the perspective of versatility, dynamic reprogrammability, and domain-specific customization, respectively, shown in Fig. 1.

Versatility, or universality, is one of the important features of photonic AI hardware to accelerate a variety of DNN workloads. A versatile/generic photonic accelerator based on universal optical linear units is capable of realizing general ma-

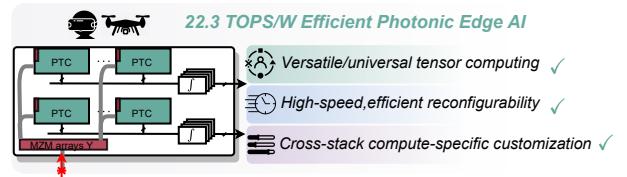


FIG. 1: Our versatile, reconfigurable, cross-stack customized photonic accelerator TeMPO achieves digital-comparable accuracy with 22.3 TOPS/W efficiency on edge AI.

trix multiplication (GEMM) and thus directly implementing a wide spectrum of pre-trained digital DNNs. Many specialized linear units are not applicable to generic tensor computation since they restrict their matrix expressivity to a subspace of specialized matrices for higher hardware efficiency, e.g., butterfly meshes³ and tensorized MZI arrays¹⁵.

Besides versatility, photonic computing requires real-time, efficient input tensor encoding with low reconfiguration costs. One example is the MZI arrays, which support arbitrary weight matrices but suffer from high weight encoding costs due to the high complexity of matrix decomposition required to encode weights. Similarly, many subspace linear unit designs can approximate GEMM operations by cascading more programmable devices but require an even more costly optimization-based approach to map the weight matrix^{3,6}. Such a property restricts those designs to only support weight-

^{a)}Meng Zhang and Dennis Yin are equal contributors to this work and designated as co-first authors.

static linear operations, e.g., fully-connected (FC) layers and convolutional (CONV) layers, where weights are pretrained and pre-encoded into the device/circuit transmissions. However, advanced AI models, e.g., Transformer¹⁶⁻²¹ based on attention operations where both matrix multiplication operands are dynamic, full-range, and general tensors, cannot be efficiently mapped to those weight-static PTCs.

The third critical feature to enable efficient, scalable PTCs is domain-specific hardware customization. ① At the device level, many optical computing hardware demonstrations are based on standard foundry PDK elements, which are designed for optical communications and not optimized for analog neuromorphic computing. For example, bulky electro-optic (E-O) modulators (~mm-level in length)²² can be used as the transmitter module for high-speed communication but are not suitable for analog computing as the footprint is intractable with quadratically many such modulators for input encoding. On the other hand, thermo-optic MZI modulators are usually compact but can only be modulated at KHz frequency due to the $\sim 10 \mu\text{s}$ thermal constant and are usually power-consuming. Plasmonic devices²³ are compact and high-speed but show high insertion loss ($> 10 \text{ dB}$), leading to significant laser power consumption. Hence, compact, low-power, low-loss, and high-speed modulators are in high demand for efficient optical computing. MRRs are compact and low-loss; however, their high locking power and high sensitivity to thermal variations limit their efficiency and robustness²⁴. To bridge the gap at the device level, it is necessary to customize computing-specific optical components, e.g., multi-operand devices for compact neural computing^{25,26}, diffractive meta-computing systems⁷. ② At the circuit level, customization is critical to reducing the long-lasting analog-to-digital and optical-to-electrical conversion bottlenecks. ③ At the architecture level, due to the lack of optical memory, the large spatial footprint of photonic circuits, and the high digital memory access cost, the architecture topology and dataflow also need to be customized to fully leverage the temporal locality to reduce data movement cost and maximize hardware sharing. Only with device-circuit-architecture cross-layer co-design and customization can we realize photonic computing's advantages compared to its electronic counterparts.

In this work, we present a time-multiplexed dynamic photonic tensor accelerator design, dubbed TeMPO, for efficient edge AI acceleration, featuring ultra-compact slow-light electro-optic modulators for input operand encoding, hierarchical partial product accumulation with lightweight capacitive temporal integration modules and multi-core architecture to maximize sharing of data input/readout circuitry. One key innovation of this work is the utilization of custom-designed, foundry-fabricated slow-light MZI modulators (SL-MZM) with enhanced light-matter interaction for size and power reduction. It has a phase shifter length of $150\text{--}200 \mu\text{m}$ and a footprint about $10\times$ greater than Si MRR while an order of magnitude smaller than the typical foundry offered Si Mach-Zehnder modulator (MZM) PDK elements. This SL-MZM is thermally robust, with no thermal tuning/locking circuit needed, and can also tolerate large manufacturing variations. Different from a multi-wavelength dynamic PTC de-

signs⁵, TeMPO simplifies the spectral multi-wavelength encoding to high-speed temporal encoding, eliminating the need for complex dispersion-engineered broadband device designs such as Si modulators, optical power splitters and directional couplers as well as remove WDM MUX/DEMUX overhead.

The major contributions of this paper are as follows:

- We present a compact and energy-efficient multi-core photonic edge AI accelerator, TeMPO, with device and architecture co-optimization and customization.
- **Compact & Efficient Photonic Components** – To enable ultra-fast, compact, low-power input operand encoding and dot-product computing, we adopt a customized slow-light MZM device with orders-of-magnitude smaller footprint and switching energy than the PDK MZM. We also customize optical power splitters with varying splitting ratios and an ultra-low power $\pi/2$ phase shifter. With customized devices, TeMPO is $6.8\times$ more compact and $9.1\times$ more power efficient than the foundry counterparts.
- **Hierarchical Product Accumulation** – TeMPO leverages photocurrent aggregation and temporal integration for partial product accumulation in the analog domain, significantly reducing the laser power and analog-to-digital conversion cost. We also enable input modulator sharing and output readout circuitry sharing to minimize the E-O/O-E cost.
- **Versatile and Robust Edge AI Evaluation** – We evaluate TeMPO on both convolutional NNs and Vision Transformers on AR/VR speech recognition, image classification, and advanced semantic segmentation tasks and show comparable accuracy and superior robustness to low-bit quantization and hardware noises from experimental measurement.
- **New Area-Energy Efficiency Pareto Frontier** – We comprehensively evaluate the scalability and efficiency of our proposed TeMPO architecture and show 368.6 TOPS peak performance, 22.3 TOPS/W energy efficiency, and 1.2 TOPS/mm² compute density, outperforming state-of-the-art electronic counterparts.

II. OVERVIEW OF TIME-MULTIPLEXED DYNAMIC PTC ARCHITECTURE DESIGN OF TEMPO

Matrix multiplication is the key linear operation for various information processing workloads. The proposed dynamic photonic tensor core will perform matrix-matrix multiplication. For generality, we consider two input matrices, matrix X with $M \times N$ dimension and matrix Y with $N \times Q$ dimension:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & \ddots & \vdots \\ x_{M1} & \cdots & x_{MN} \end{bmatrix}, \quad Y = \begin{bmatrix} y_{11} & \cdots & y_{1Q} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_{NQ} \end{bmatrix}. \quad (1)$$

The matrix multiplication $Z = X \cdot Y$ is

$$Z = \begin{bmatrix} z_{11} & \cdots & z_{1Q} \\ \vdots & \ddots & \vdots \\ z_{M1} & \cdots & z_{MQ} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & \ddots & \vdots \\ x_{M1} & \cdots & x_{MN} \end{bmatrix} \cdot \begin{bmatrix} y_{11} & \cdots & y_{1Q} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_{NQ} \end{bmatrix}. \quad (2)$$

The resulting Z is an $M \times Q$ matrix; and its a -th row, b -th column element z_{ab} is obtained by calculating the dot-product of a -th row vector of X and b -th column vector of Y , i.e.,

$$z_{ab} = X_a \cdot Y_b = [x_{a1} \cdots x_{aN}] \cdot \begin{bmatrix} y_{1b} \\ \vdots \\ y_{Nb} \end{bmatrix}. \quad (3)$$

Each vector dot-product operation can be mapped to a dynamic dot-product engine. Multiple dot-product engines can form an array structure, i.e., a tensor core, to realize parallel matrix-matrix multiplication. The design of the dot product engine will be discussed in Section II A, and the proposed PTC architecture will be explained in Section II B.

A. Dynamic Photonic Dot-Product Engine

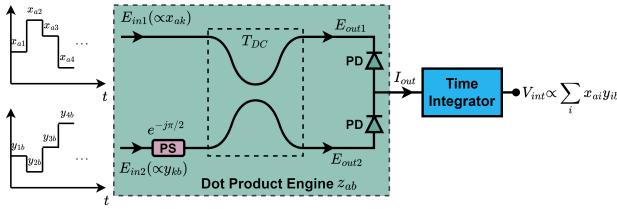


FIG. 2: Schematic of a dynamic optical dot-product engine.

The matrix dot product operation that can be realized in photonic/electronic hardware is shown in Fig. 2. Matrix dot product calculates the element z_{ab} while the data pairs (x_{ak}, y_{kb}) ($k = 1, 2, \dots, N$) are encoded to the phase and amplitude of input light to the directional coupler. A phase shifter (PS) is implemented in one input arm of the directional coupler to generate a $-\pi/2$ phase shift. The core of the dot product engine consists of a 2×2 directional coupler connecting followed by a pair of balanced photodetectors. The 2×2 directional coupler provides interference between coherent light inputs of two arms. The transfer matrix for this structure with an ideal, lossless directional coupler can be expressed as

$$T_{DC} \cdot T_{PS} = \begin{bmatrix} t & j\kappa \\ j\kappa & t \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & e^{-j\frac{\pi}{2}} \end{bmatrix}, \quad (4)$$

where t is the through-coupling coefficient, κ is the cross-coupling coefficient and j is the imaginary unit. For dot product computing, 50:50-splitting is used, so $t = \kappa = \sqrt{2}/2$. Consider the electric fields of input signals to the directional coupler $[E_1, E_2]^T$ encoding a data pair $[x_{ak}, y_{kb}]^T$, the output of the

directional coupler $[E_{out1}, E_{out2}]^T$ can be expressed as

$$\begin{bmatrix} E_{out1} \\ E_{out2} \end{bmatrix} = T_{DC} \cdot T_{PS} \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} = \frac{\sqrt{2}}{2} \begin{bmatrix} 1 & j \\ j & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & -j \end{bmatrix} \cdot \begin{bmatrix} x_{ak} \\ y_{kb} \end{bmatrix} = \frac{\sqrt{2}}{2} \begin{bmatrix} x_{ak} + y_{kb} \\ j(x_{ak} - y_{kb}) \end{bmatrix}. \quad (5)$$

The photocurrent of the PDs connected to the directional coupler is proportional to the received optical power. Assume identical responsivity of two cascaded PDs, the output current I_{out} can be calculated by

$$I_{out} \propto |E_{out1}|^2 - |E_{out2}|^2 \propto |x_{ak} + y_{kb}|^2 - |x_{ak} - y_{kb}|^2 \propto x_{ak} y_{kb}. \quad (6)$$

This is the product between two elements. To accomplish the dot-product operation between vector X_a and Y_b , $x_{ak} y_{kb}$ needs to be summed up over all the k labels from 1 to N . The electrical modulated signal to the slow-light MZM follows the sample-and-hold operation to inject the vector elements $x_{a1}, x_{a2}, \dots, x_{aN}$ and $y_{1b}, y_{2b}, \dots, y_{Nb}$ through two slow-light MZMs sequentially. A time integrator is connected right after the dot product engine to operate the summation $\sum_{k=1}^N x_{ak} y_{kb}$ in the time domain so that the integrator readout voltage V_{int} will represent the dot product between vector X_a and Y_b ,

$$V_{int} \propto \sum_{k=1}^N x_{ak} y_{kb} \propto X_a \cdot Y_b. \quad (7)$$

The detailed physical realization of the dot-product engine and time integrator will be discussed in Section III.

B. TeMPO Architecture Overview

We have introduced one dynamic dot-product engine to realize vector dot-product. Now, we introduce a multi-core time-multiplexed photonic tensor accelerator TeMPO for parallel dot-product, shown in Fig. 3. We have R tiles in the architecture, and each tile contains C PTCs. Each PTC is a cross-bar of $K \times K$ dynamic dot-product engines, which can finish a $K \times 1$ times $1 \times K$ vector outer product at each timestep.

Given an $M \times N$ times $N \times Q$ GEMM workload, we first partition the matrix X into M/K horizontal strips, each with a size of $K \times N$, and matrix Y into Q/K vertical strips, each with a size of $N \times K$. One $K \times K$ block in the result matrix $Z_{1:K,1:K}$ can be computed by accumulating N vector outer product, i.e., $Z_{1:K,1:K} = \sum_{t=1}^N X_{1:K,t} \cdot Y_{t,1:K}$. This length- N reduction can be mapped to C PTCs in a tile in parallel, and each PTC is responsible for computing $P = \frac{N}{C}$ vector outer products, which is formally rewritten as $Z_{1:K,1:K} = \sum_{p=1}^P (\sum_{c=1}^C X_{1:K,(c-1)P+p} \cdot Y_{(c-1)P+p,1:K})$. Therefore, the total cycles consumed to compute $Z_{1:K,1:K}$ is $P = \frac{N}{C}$. There are $\frac{M}{K} \times \frac{Q}{K}$ of such matrix blocks in Z , and we mapped them to R tiles in parallel. This entire matrix multiplication requires in total $\frac{MQN}{RCK^2}$ cycles.

Each cycle is defined as (1) feeding one vector into our PTC, (2) reading out the outer product results as photocurrent, (3) converting it to the electronic domain, and (4) accumulating partial product. As we mentioned above, each PTC consumes $P = \frac{N}{C}$ cycles to finish one $K \times K$ block in the

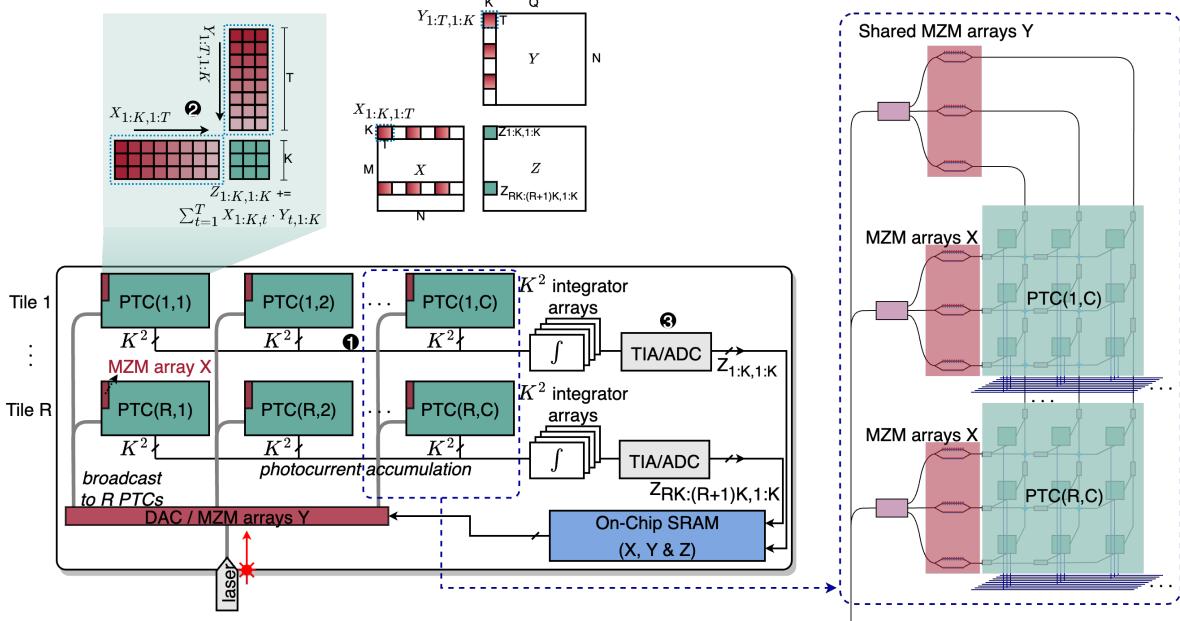


FIG. 3: Our designed multi-core time-multiplexed dynamic photonic tensor accelerator TeMPO. ①-③ correspond to the hierarchical partial product accumulation in Eq. (8). All R PTCs in a column share the same Y matrix MZMs. All C PTCs in a row share the same readout circuitry.

Z matrix, which means a conventional architecture needs to convert the photocurrent as electronic digital signals through trans-impedance amplifier (TIA) and analog-to-digital converter (ADC) *at every cycle for each PTC* and accumulate the result digitally with adders and registers. With a high data rate, e.g., 5-10 GHz, the AD conversion and digital accumulation cost is non-trivial, becoming a bottleneck of the performance and efficiency as the ADC power is proportional to its sampling frequency.

Hierarchical Product Accumulation – To resolve the AD conversion efficiency bottleneck, we adopt hierarchical accumulation both spatially and temporally in the analog domain. The dot-product result is rewritten as

$$Z_{1:K,1:K} = \sum_{p=1}^{P/T} \sum_{t=1}^T \sum_{c=1}^C X_{1:K,(c-1)P+(p-1)T+t} Y_{(c-1)P+(p-1)T+t,1:K} \quad (8)$$

① At each timestep t , the photocurrent carrying the partial product results will first be aggregated from all C PTCs in parallel within the same tile via analog current summation, corresponding to the most-inner summation in Eq. (8). ② Then, the aggregated photocurrents will be further accumulated over T timesteps at the temporal integrator but still in the analog domain. ③ After every T timesteps, the partial sum will be converted to the digital domain via the analog-to-digital converters (ADCs), and the integrators will be reset and prepared for the following T cycles. With this hierarchical accumulation mechanism, the ADC conversion is minimized to merely P/T times per matrix block, leading to T times lower AD conversion frequency and, thus, power consumption.

Input/Output Hardware Sharing – To maximize the hardware sharing of the multi-core accelerator, we explore both input and output sharing. For input sharing, R PTCs across different tiles within the same column will share the same Y vectors. Thus, the input vectors Y can be modulated in the shared MZM arrays and broadcast to them via on-chip waveguide interconnects. For output sharing, the partial products from C PTCs within a tile are aggregated by summing up their photocurrent. For each tile, all C PTCs share the same group of integrators, TIAs, and ADCs. The total cost of those readout circuitry can be reduced by C times with output sharing.

Next, we focus on the detailed design of a $K \times K$ time-multiplexed PTC to explain how our architecture performs dynamic matrix-matrix multiplication. For illustration simplicity, we set the matrix with an equal number of rows and columns, i.e., K , while the architecture can be applied to a matrix with arbitrary dimensions. A coherent monochromatic light source is used as the input to the photonic tensor core units. The input light is first fanned out to $2K$ waveguides via a $1 \times 2K$ splitter. Next, a slow-light Mach-Zehnder modulator (SL-MZM) is connected in each waveguide arm, serving as the input operand modulator of the PTC. Digital electrical signals carrying the matrix information are converted to analog optical signals represented by the amplitude and phase before optical signals reach the dot-product engine for computing. Let E_{in} be the electric field of the input light to the SL-MZM, and the electric field of MZM output can be expressed as $E_{in} \cos \theta$, allowing broadband mapping of both positive and negative values. We consider two optical routing schemes for the PTC architecture in this work, namely a double-layer-splitters scheme TeMPO-D and an embedded-uneven-splitters

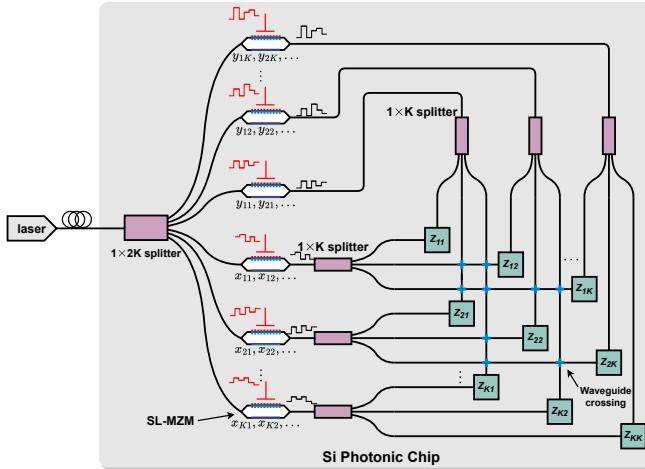


FIG. 4: Schematic of our proposed time-multiplexed double-layer-splitter tensor core TeMPO-D. $K = 3$ is sketched here as an example for illustration.

scheme TeMPO-E to guide the encoding optical signals to the targeting dot product engines. Schematics of the proposed PTC architecture are shown in Fig. 4 and Fig. 5.

1. Double-Layer-Splitter PTC Design TeMPO-D

A double-layer-splitter PTC design consists of two layers of optical splitters to route the encoded optical signals to the targeted dot product engines for matrix calculation, and a schematic of the architecture is shown in Fig. 4. After the 1st fan-out $1 \times 2K$ splitter, half of the optical paths (bottom K paths) are used to encode matrix X via an SL-MZM array, mapping to a row vector of matrix X : $X_a = [x_{a1}, \dots, x_{aK}]$, ($a = 1, 2, \dots$). SL-MZMs on the top K arms of the $1 \times 2K$ splitter couple data of N column vectors of matrix Y : $Y_b = [y_{1b}, \dots, y_{Kb}]^T$, ($b = 1, 2, \dots$). The second layer consists of $2K$ $1 \times K$ optical splitters, each of which evenly splits the optical power with encoded information into K secondary output arms so that dot products between any pair of X_a and Y_b can be calculated simultaneously at K^2 dot product engines. Waveguide crossings are needed for this architecture. The coded optical signals may pass up to $(K - 1)^2$ crossings to reach the dot product engine.

2. Embedded-Uneven-Splitters PTC Design TeMPO-E

A schematic of the embedded-uneven-splitters PTC design, TeMPO-E is illustrated in Fig. 5. Different from the TeMPO-D design, this architecture adopts a series of uneven splitters to eliminate waveguide crossings. The splitting ratios are set at $1 : (K - 1)$, $1 : (K - 2)$, \dots , and $1 : 1$. For a PTC with K^2 dot product engines, the splitting ratios of the two optical splitters that guide light into the dot product engine z_{ab} are $1 : (K - a)$ and $1 : (K - b)$, respectively, to ensure identical input power to

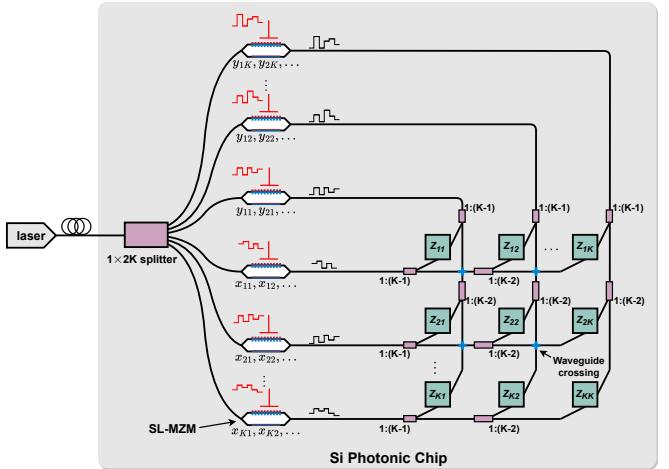


FIG. 5: Schematic of our proposed time-multiplexed embedded-uneven-splitter tensor core TeMPO-E. $K = 3$ is sketched here as an example for illustration.

each dot product engine. The maximum number of crossings on the optical path is $K - 1$.

Comparing TeMPO-D with TeMPO-E, TeMPO-D design only requires 1 optical splitter before reaching the DOT engine with the cost of the increased number of waveguide crossings in some waveguide paths. For the TeMPO-E design, the number of uneven power splitters and waveguide crossings needed in each path are both $K - 1$, while TeMPO-D design requires $(K - 1)^2$ waveguide crossings. We anticipate lower accumulated device loss in the TeMPO-E design when K is large. In the following discussion, we only focus on the embedded-uneven-splitters design TeMPO-E and simplify it as TeMPO.

III. PHOTONIC COMPONENTS FOR PTCs

A. Laser Source

A PTC utilizing optical wave phase and amplitude in time-domain processing only requires a monochromatic light source for optical signal processing. In the realm of integrated photonic computing chip design, o-band operation, in comparison with c-band components, offers several distinct advantages such as a smaller optical mode volume in Si/SiO₂ waveguide structure, higher mode confinement with tighter bending radius and $> 1.5 \times$ higher in Ge PD responsivity^{27,28}.

The second consideration pertains to the choice between an on-chip III-V integrated laser diode and an off-chip laser module. While the heterogeneously bonded laser to Si holds the promise of the miniaturized, photolithographically defined coherent on-chip light source, it has yet to mature for mass production. The long-term reliability of on-chip lasers remains undetermined. Laser cavities are highly sensitive to temperature variations, thus heterogeneously intergated on-chip laser, being in the close vicinity of other electronics that generate considerable heat would demand more complex electronics circuits in thermal management to maintain on-chip

laser diode emission stability in optical mode/wavelength, polarization, and optical power. Integrated optical isolators on Si platform are not yet available from SiPho foundry; while an optical isolator is critical in minimizing reflections that could disturb laser operation if the reflection is not addressed. Varying laser operation will also, in turn, degrade the PTC performance. In this work, we advocate a technological path that utilizes a separate, off-chip laser module that takes advantage of the latest advancement in optical packaging to achieve low insertion loss at the fiber-to-chip interface.

High-power monolithic o-band lasers, capable of producing output powers as high as 150 mW²⁹, are commercially available now. In this work, we utilize a moderate laser power of 100 mW for system power-related analysis and evaluation. Utilizing index-matched epoxy and emerging packaging technology, such as photonic wires^{30–32}, one can expect 0.5dB - 2dB insertion loss at the fiber to chip facet.

B. Slow-Light Mach-Zehnder Modulator

Mach-Zehnder Modulators (MZMs) play a crucial role in the conversion of electrical signals to the optical domain in chip-scale PTC. Si modulators, utilizing the carrier plasma effect, offer a cost-effective and high-density integration solution for on-chip PTC. Achieving a dot product operation for matrices of the size of $K \times K$ requires $2K$ modulators for signal conversion. The physical dimension of these Si modulators serves as a critical design parameter, impacting the scalability of matrix operation. In this study, our approach involves the adoption of a 1D dielectric photonic crystal waveguide, specifically a rectangular-shaped Bragg grating³³, as a slow-light-enabled compact modulator to significantly reduce the footprint of the modulator array^{34,35}. Lately, we have experimentally demonstrated a Si slow-light MZM (SL-MZM) with a phase shifter length (L_{PS}) of $150 \mu\text{m}$ for optical compute application³⁶. The SL-MZM reported in this work was fabricated at AIM Photonics under a multi-project wafer (MPW) run, ensuring complete foundry compatibility. The modulator output is routed to an on-chip Ge photodetector (PD), a standard AIM PDK component with a tested bit rate of 15 Gbps. The SL-MZM, operating under maximum V_{pp} signals of 3.5V, was characterized with up to 6-bit of resolution using both staircase and random data inputs. The readout signals from the PD are displayed on a real-time oscilloscope, shown in Fig. 6. The averaged variance during bit-holding time is reported as 9.72×10^{-7} and 6.59×10^{-5} for the staircase and random signal input cases, respectively.

Reflection occurring at different junctions within the modulator device, optical absorption due to carriers in waveguides, propagation loss in the Bragg grating phase shifter due to increased group indices, and mode mismatch at the Bragg grating waveguide interfaces are the primary factors contributing to the modulator insertion loss. The measured total modulator insertion loss is ~ 6.4 dB for $L_{PS} = 150 \mu\text{m}$ and is utilized as the loss figure in the system evaluation.

To achieve high-bit resolution at a high computing clock frequency, it is imperative to optimize both the electrical band-

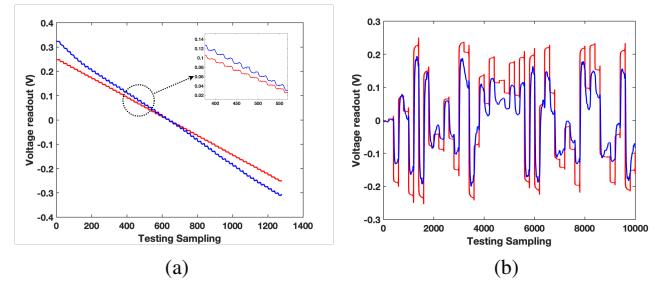


FIG. 6: Bit resolution testing of SL-MZM at 100 MHz clock frequency with (a) 6-bits staircase signal and (b) 6-bits random signal. The red curves show direct driving signals from the arbitrary waveform generator (AWG), while the blue curves represent the SL-MZM response readout by the on-chip PD.

width and linearity of a Si modulator. Operating under reverse bias, the speed of a Si SL-MZM is limited by its RC time constant and photon lifetime. Typically, the PN junctions are doped at an elevated level (ranging from $10^{18}/\text{cm}^3$ to $10^{19}/\text{cm}^3$) to enhance the carrier plasma effect. As the phase shifter length is reduced in an SL-MZM, the total capacitance decreases. In this work, the measured SL-MZM junction capacitance C_j was approximately $\sim 0.75 \text{ pF}$. Depending on the doping level in the connecting Si bar from the ridge waveguide to the via contacts, the intrinsic resistance of a SL-MZM ranges from 5 to 10 ohms. The estimated RC time-limited electrical bandwidth of a SL-MZM is thus in the hundreds of GHz. The slow-light effect can be viewed as a traveling wave resonant in its propagation direction, with the optical bandwidth determined by the Q-factor of the resonator. For the rectangular Bragg grating-shaped slow-light, an optical bandwidth of approximately $\sim 26 \text{ GHz}$ is estimated³⁷. However, the SL-MZM of this work didn't reach its maximum bandwidth potential due to impedance mismatch of the electrodes³⁴, mismatch of the RF signals speed with the optical wave with high group index³⁸ and waveguide dispersion in the slow light spectrum. Dispersion engineering techniques such as phase-shifted Bragg grating, dispersion compensation³⁹, and line-shift photonic crystal waveguide are all effective approaches in reducing the dispersion-induced bandwidth penalty. With careful device design and optimization, a SL-MZM operating at a 5GHz clock frequency is feasible, as assumed for system-level performance evaluation in this study.

C. Optical Power Splitter

The optical splitter is a crucial passive photonic component in integrated photonic systems for splitting optical power. Various types of structures such as Y-junction splitters⁴⁰, multimode interferometers (MMIs)^{41,42} and directional couplers⁴³ have been demonstrated to achieve power splitting with varying splitting ratios. Y-junction splitters are usually compact

and broadband, but the sharp corners can lead to increased reflection, resulting in unwanted FR resonance in a photonic system⁴⁰. The MMI-based power splitter is suitable for $1 \times K$ uniform power splitting, while the shape of tapered input and output waveguides needs to be carefully designed⁴⁴. By adjusting the coupling length, a directional coupler can also be used to obtain varying optical power splitting.

1. $1 \times 2K$ Optical Splitter

In our proposed PTC, the first layer $1 \times 2K$ splitter adopts the design of MMI to fan out the CW laser light to $2K$ slow-light MZMs. For a center-excited $1 \times K$ MMI splitter, K -folded self-imaging can be reproduced at MMI output when the length of the multimode waveguide section L_{MMI} satisfies

$$L_{MMI} = \frac{3iL_\pi}{4K}, i = 1, 2, 3, \dots, \quad (9)$$

where L_π represents the beating length given by

$$L_\pi = \frac{\pi}{\beta_0 - \beta_1} \approx \frac{4n_{eff}w_e^2}{3\lambda_0}. \quad (10)$$

Here β_0 , β_1 represent the propagation constants of the fundamental mode and first-order mode, n_{eff} is the effective index of the multimode waveguide section, λ_0 is the operated wavelength and w_e is the effective width and can be approximated as the multimode waveguide section width W_{MMI} in silicon photonics⁴⁵. We use the 1×8 MMI design in⁴⁴ for $K = 4$ case and develop the 1×10 and 1×12 MMI designs based on Eq. (9). Consider a silicon waveguide layer thickness of 220 nm. The waveguide width is 450 nm and tapered to 1.2 μm at the multimode waveguide segment. The simulated electric field profiles of 1×8 , 1×10 , and 1×12 MMIs are shown in Fig. 7, corresponding to $N = 4$, $N = 5$, and $N = 6$ scenarios. The dimensions ($L_{MMI} \times W_{MMI}$) of 1×8 , 1×10 and 1×12 MMIs are $27.8 \mu m \times 11.3 \mu m$, $34.6 \mu m \times 14.1 \mu m$, $41.4 \mu m \times 16.9 \mu m$, respectively. The insertion loss (IL) is calculated to be 0.14dB, 0.20dB, and 0.21dB for 1×8 , 1×10 , and 1×12 MMIs, respectively. We adopt the 1×10 MMI as a base design and assume a linear scaling law in MMI's length/width for a generic $1 \times 2K$ MMI and a near-constant insertion loss regardless of fanout in later discussion.

2. Optical Power Splitter Guiding to the Dot-Product Engine

The TeMPO adopts directional couplers with varying splitting ratios to guide the coded optical signals to each DOT engine for matrix computing. A directional coupler with even splitting is often offered as a standard PDK component from SiPho foundries. Keeping the waveguide gap constant, one only needs to change the coupling length to adjust the splitting ratio. With 480 nm waveguide width and 200 nm gap between two parallel waveguides in the coupling region, our simulation shows that the coupling length is 14.6 μm , 11.2 μm , 9.2 μm , 8 μm , and 7 μm to achieve splitting ratios of 1:1, 1:2, 1:3, 1:4, and 1:5, respectively.

D. Dot Product Engine Design

The dot product engine to realize vector-vector dot product is the key computation unit in our proposed photonic tensor core. A dot product engine consists of a 2×2 optical power splitter, a $\pi/2$ phase shifter, a pair of balanced PD, and a time integrator. They will be discussed separately in this section.

1. 2×2 Optical Power Splitter

A 2×2 50:50 optical power splitter is needed to generate interference between the optical signals from 2 input arms. both directional couplers and MMIs can be used to generate 50:50 power splitting. The directional coupler consists of two closely placed parallel waveguides, and the splitting ratio is wavelength-dependent, thus sensitive to the fabrication accuracy. The 2×2 MMI power splitting is less wavelength sensitive than the directional coupler, while it is challenging to achieve an exact 50:50 splitting ratio, and insertion loss is usually higher than the directional coupler. Two interference mechanisms, namely paired interference and general interference, can be applied to MMI design. The paired interference mechanism is generally used for designing $2 \times K$ MMIs, where the modes contributing to the imaging in the multimode section are paired⁴⁵. The length of the multimode waveguide section L_{MMI} satisfies

$$L_{MMI} = \frac{iL_\pi}{K}, i = 1, 2, 3, \dots. \quad (11)$$

The two input waveguides have to be placed at $+\frac{W_{MMI}}{6}$ and $-\frac{W_{MMI}}{6}$ vertically from the center. For 2×2 MMI based on a general $K \times K$ interference mechanism, there is no restriction on the location of the input waveguides⁴⁵. The length of the multimode waveguide section L_{MMI} can be expressed as

$$L_{MMI} = \frac{3iL_\pi}{K}, i = 1, 2, 3, \dots. \quad (12)$$

L_π in Eq. (11) and Eq. (12) follows the same definition as Eq. (10). Three 2×2 optical power splitter designs are developed, and the results are summarized in Table I. The simulated electric field profiles are illustrated in Fig. 8, where the optical power is coupled in through one input arm, and output power is measured through both output arms. Overall, the directional coupler features lower insertion loss and smaller size, while the two MMI designs have larger bandwidth near the targeting 50:50 splitting ratio. Taking the dimension, splitting ratio, and insertion loss into consideration, the directional coupler-based 2×2 optical power splitter design will be utilized in the following system-level simulation study.

2. $\pi/2$ Phase Shifter

Maintaining a consistent $\pi/2$ phase difference between two optical paths can be realized through the utilization of either

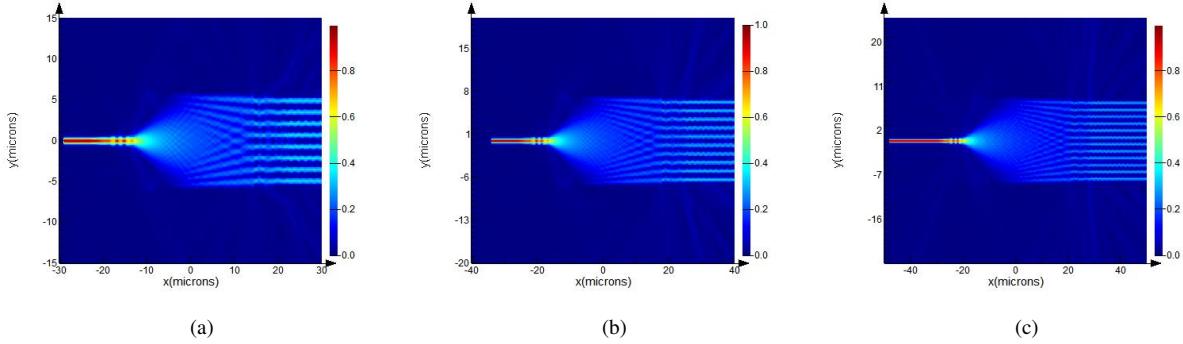


FIG. 7: Simulated electric field profiles of (a) 1×8 , (b) 1×10 , and (c) 1×12 MMIs.

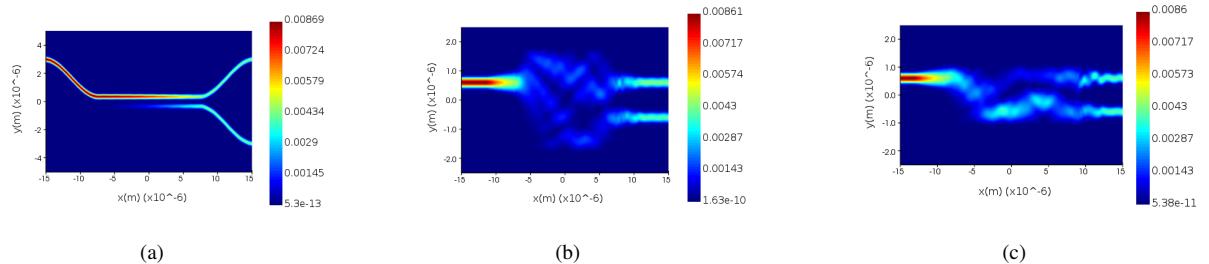


FIG. 8: Simulated electric power profiles of (a) directional coupler, (b) MMI with paired interference mechanism, and (c) MMI with general interference mechanism.

TABLE I: Simulation results of three 2×2 optical splitter designs.

Splitter Design	Directional Coupler	MMI (Paired Interference)	MMI (General Interference)
Optical Coupling Region Dimension ($W \times L$)	$1.2 \mu m \times 14.6 \mu m$	$3.6 \mu m \times 14.5 \mu m$	$2.2 \mu m \times 18 \mu m$
Block Dimension ($W \times L$)	$6.5 \mu m \times 31 \mu m$	$7 \mu m \times 40.5 \mu m$	$7 \mu m \times 44 \mu m$
Splitting Ratio	50:50	52.5:47.5	50:50
Bandwidth at Targeting Splitting Ratio	1550 nm	1500 nm \sim 1600 nm	1530 nm \sim 1570 nm
Insertion Loss at 1550 nm	0.05dB	0.18dB	0.37dB

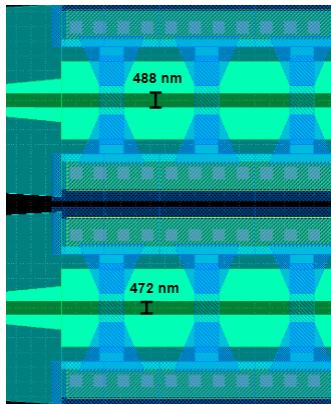


FIG. 9: Side-by-side comparison of the thermo-optic phase shifters with a waveguide width difference.

a path length difference or a waveguide effective index differ-

ence. In practice, there will be deviations from the targeted phase shifter (PS) owing to variations in waveguide dimensions induced during the manufacturing process. Thermal tuning is an effective method to adjust the offset to reach a precise $\pi/2$ phase difference. Optimized for the lowest static thermal tuning power, we adopt the design of n_{eff} difference to achieve a $\pi/2$ phase shifter. The difference in phase ϕ between the two arms with identical lengths is

$$\beta_1 L - \beta_2 L = \phi, \quad (13)$$

where β_1 and β_2 represent the propagation constants of the two arms and L is arm length. We set the global waveguide width to 480 nm while the two arms are set at 488 nm and 472 nm. A 5 μm taper is connected to the PS region. For a PS length of 30 μm , it will produce a $\sim \pi/2$ phase difference. A resistive heater is placed in the optical path of both arms following the design in⁴⁶. As those two arms are placed in close vicinity, we anticipate minimum width difference variation, though their actual dimensions can deviate substantially from

targeted values. In an extreme fabrication variation scenario of 488+2 nm and 472-2 nm, the phase difference is 0.6345π , corresponding to an estimated heater tuning power of 5 mW to reach $\pi/2$. When the fabrication variation is relatively small with advanced fabrication technology, we only need negligible active tuning power to compensate for the phase errors.

3. Photodetector Responsivity and Sensitivity

The sensitivity and responsivity of photodetectors are closely related to the laser power requirement and integrator designs. Sensitivity SPD defines the minimum gap between two levels of optical power received by photodetectors given a certain bit error rate. The loss of the circuit, including power splitting loss and insertion loss, is as follows

$$IL = IL_{couple} + 10\log_{10} K^2 + IL_{MZM} + (K-1)IL_{cross} + K \cdot IL_{splitter} + IL_{PS} + IL_{DC} \quad (14)$$

Given the circuit insertion loss IL and PD sensitivity, we can derive the laser power (mW) requirement for each PTC to obtain b -bit output resolution,

$$\frac{P_{laser} \cdot (1 - 10^{-ER/10})}{10^{IL/10}} \geq I_{noise}/R_{PD} + 2^b \cdot 10^{SPD/10}, \quad (15)$$

where I_{noise} is the dark current noise floor of the PD, R_{PD} is the PD responsivity, and ER is the modulator extinction ratio. $(1 - 10^{-ER/10})$ is the power penalty to compensate for the range reduction due to the non-ideal ER . For example, with 20 dB insertion loss, 1 A/W responsivity, 20 nA dark current noise floor, 10 dB extinction ratio, and -27 dBm PD sensitivity, the minimum optical power from the laser to obtain 6-bit output is 14.2 mW.

Meanwhile, the balanced PD's current range determines the integrator's design. Given the principle of time integration, i.e., $V_{out} \propto \int \frac{I_{PD}}{C_{int}} dt$, the maximum voltage with T -timestep integration of f -frequency datarate is $V_{max} \propto \int_0^{T/f} \frac{I_{PD,max}}{C_{int}} dt = T \cdot I_{PD,max} / (C_{int} f)$. To avoid saturation-induced integration error, i.e., $V_{max} \leq V_{DD}$, we must carefully design the integration timestep T and the capacitance C_{int} given the maximum photocurrent generated by the balanced PD. Detailed integrator design specifications are introduced in the following section.

4. Temporal Integrator

The proposed time-multiplexed approach requires integration of photodetector output current for the accumulation operation as in Eq. (7). This is one of the key mechanisms in TeMPO to significantly relieve the ADC power bottleneck.

Integrator Design and Optimization – Our integrator design objective is to support a target maximum integration timestep T with good linearity in the voltage response and fast reset speed. We adopt a simple, compact, and foundry-compatible means of time integration using a capacitor. Capacitive elements are well suited for analog integration of

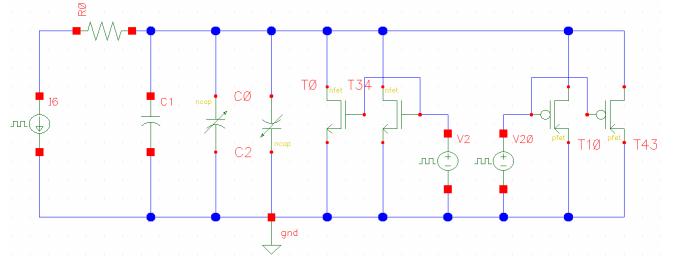


FIG. 10: Schematic of the capacitive temporal integrator.

current-based signals. The voltage across the terminals is proportional to the time-integral of the current from the photodiode. After each multiply-accumulate operation is complete, the capacitor integrator will need to be discharged (reset) before the next operation. By turning on field-effect transistors (FETs) in parallel to the capacitor, the charge across the capacitor can be rapidly dissipated for reset.

Now, we show the detailed integrator design with a target maximum integration timestep T and linearity and reset speed considerations. The proposed integration unit is shown in Fig. 10. As indicated by the insertion loss analysis and the PD responsivity, the estimated maximum photocurrent $I_{PD,max}$ is $110 \mu\text{A}$. Given a maximum targeted voltage of $V_{DD} = 240 \text{ mV}$, the signal data rate of 5 GHz, and a target integration timestep $T=60$, we can derive the capacitor $C_{int} = I_{PD,max}T/(fV_{DD}) = 5500 \text{ fF}$. Therefore, two foundry-compatible thin oxide capacitors with a capacitance range of 809 fF to 3.9 nF are connected to the PD's output. Note that besides scaling up capacitors proportionally with T , one can equivalently consider scaling down laser power and thus $I_{PD,max}$ by a factor of T . This can significantly reduce laser power but at the cost of a worse signal-to-noise ratio. In our design, we maintain the same laser power and include the T factor in the capacitance.

For a linear integrator response, multiple flipped capacitor pairs are connected in parallel to achieve a symmetric circuit topology. To enable fast periodic reset, ten 40 nm n-channel and p-channel FETs are connected in parallel with the capacitor to ensure sufficient current driving capability for reset within a single baud time period. This choice accounts for the possibility of both positive and negative source current flow from the balanced photodiode, ensuring effective reset regardless of signal polarity. For simplicity, only two of each type of FET are depicted in Fig. 10.

Note that we prefer this capacitor-based design to an alternative operational amplifier (op-amp) based design due to efficiency considerations. Integrators with an op-amp and a capacitive feedback loop show desired input/output impedance; however, they are more suitable for voltage integration tasks with notably increased chip space usage and power. In contrast, the capacitor-based design has near-zero power and is more suitable for our photocurrent accumulation mechanism in TeMPO.

Integrator SPICE Simulation – The integrator unit's simulation employs flipped capacitor pairs and 40 nm FETs, as previously mentioned. We simulated a maximum current of

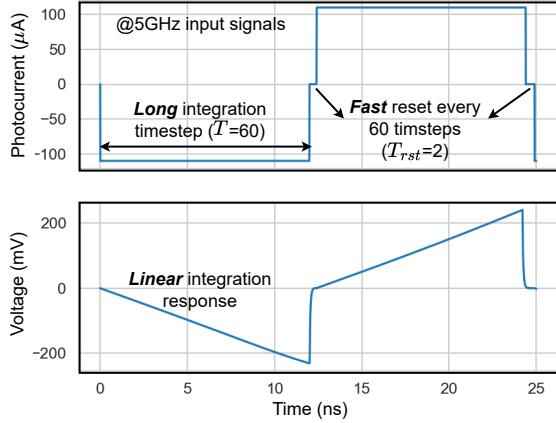


FIG. 11: Simulated waveforms of the integrator unit. Input photocurrent (Top) and integrated voltage signal (Bottom) show linear integration and rapid discharge (reset).

$\pm 110\mu\text{A}$ over the entire integration period ($T=60$) to ensure saturation of the capacitor does not occur. The FET gates received 2.5 V for 120 ps, with additional rise and fall times of 40 ps, ensuring a complete reset within a timestep of $T_{rst}=2$. The waveforms for both the current signal and the integrated voltage signal are illustrated in Fig. 11. Given the maximum anticipated current of $\pm 110\mu\text{A}$, we recorded peak voltages of approximately $\mp 240\text{ mV}$.

Integrator Cost Analysis – Our design shows a compact footprint of $A_{int}=560\text{ }\mu\text{m}^2$, a low power consumption of 0.3 mW, and a long integration timestep $T=60$, with a fast reset time T_{rst} of 2 timesteps. Note that the integrator arrays are shared across C cores in a tile; the integrator area/power cost can be further amortized by a factor of C , leading to marginal hardware overhead at the system level.

Integrator’s Benefits to System Efficiency – To justify the efficiency benefit by setting T to 60, we simulate how timestep T impacts the system power consumption when mapping a large matrix multiplication workload onto our architecture in Fig. 12. The TIA/ADC sampling frequency can be scaled down proportionally by T times, approximately leading to $T \times$ lower power. To keep ADC/TIA power less than 5%, we set T to 60 such that the on-chip power consumption can be drastically reduced from 68 W to 16 W, with the ADC/TIA bottleneck completely eliminated.

IV. EVALUATION RESULTS

In this section, we will analyze the accuracy and hardware cost of our TeMPO architecture. We focus on three variants of our TeMPO with different device configurations listed in Table II. TeMPO-Custom-SL is the fully-customized architecture settings used as our final design. For a comprehensive evaluation of TeMPO-Custom-SL, we also incorporated the analysis of on-chip memory, considering its area and power impact⁵. Similar to⁵, the architecture has a 2MB global on-chip SRAM buffer and 4KB on-chip local SRAM buffer for each tile, de-

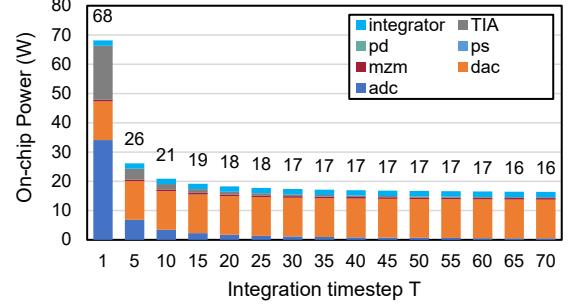


FIG. 12: Impact of temporal integration timestep T to the on-chip system power consumption for TeMPO-Custom-SL. Note that memory and off-chip laser are excluded.

signed to hold two 512×512 matrix multiplication workloads. To summarize, TeMPO-Custom-SL consumes 321 mm^2 area, 17.5 W power at 5 GHz and $T=60$ integration timestep, and realizes 368.6 TOPS peak computing speed with 6-bit precision, 22.3 TOPS/W energy efficiency, and 1.2 TOPS/mm² compute density.

A. Accuracy Evaluation on Various Edge AI Workloads

The performance of the proposed TeMPO is evaluated on real-world edge machine learning tasks, including a Vision Transformer (ViT) DeiT-Tiny¹⁸ on image recognition on ImageNet-1k⁵¹, a convolutional neural network (CNN) on the AR/VR voice keyword spotting task on Google Speech Command dataset⁵², and a FCN-ResNet50⁵³ model on semantic segmentation on PASCAL VOC2012⁵⁴. Our evaluation covers both weight-static CNNs and Transformers with dynamic self-attention operations for both speech and vision tasks to demonstrate our versatility for diverse edge ML. During model training, we adopt a hardware-aware training flow to consider the 6-bit weight/input quantization and hardware-measurement noises to guarantee a robust deployment on our photonic tensor cores.

Noise/Quantization-Aware Training – We adopt learnable step-size per-channel quantization⁵⁵ for both input operands X and Y and the output S . For weight/activation quantization, the i -th channel of the quantized tensors is

$$X_q^i = \mathcal{Q}(X^i) = ([\text{clip}(X^i/\alpha^i + z^i, -2^{b-1}, 2^{b-1} - 1)] - z^i) \cdot \alpha^i, \quad (16)$$

where the scaling factor α^i and the zero point z^i can be trained with gradient descent for the i -th channel/kernel. The gradient of the non-differentiable rounding function can be estimated by using a straight-through estimator (STE). After quantization, we also dynamically inject relative random Gaussian noises with a noise intensity of σ to both input tensors in matrix multiplication, i.e., $\tilde{X}_q = X_q + \Delta X$, where $\Delta X \sim \mathcal{N}(0, (\sigma|X_q|)^2)$.

Figure 13 visualizes our proposed TeMPO on three representative edge AI workloads. Table III shows the task performance on each application with 6-bit weight/activation quantization and noise perturbations. Our 6-bit quantized TeMPO

TABLE II: Component parameters used in three of TeMPO variants. IL represents insertion loss.

Device	Parameter	Value	TeMPO Foundry	TeMPO Foundry-SL	TeMPO Custom-SL
DAC ⁴⁷	Precision Power Area	8-bit 50 mW(@14GSPS) 11,000 μm^2	★	★	★
ADC ⁴⁸	Precision Power Area	8-bit 14.8 mW(@10GSPS) 2,850 μm^2	★	★	★
Foundry Photodetector ^{22,49}	Power Sensitivity Area Bandwidth Responsivity	25 nW at -1 V -27 dBm 16 \times 20 μm^2 27 GHz 1.1 A/W	★	★	★
TIA ⁵⁰	Power Area Bandwidth	3 mW <50 μm^2 40 GHz	★	★	★
Foundry MZM ²²	Static power IL Area EO bandwidth Modulation efficiency Extinction ratio	70 nW 3 dB 1600 \times 460 μm^2 12.5 GHz 450 fJ/bit >15 dB	★		
Customized SL-MZM ³⁴ (Fabricated at AIM)	Static power IL Area EO bandwidth Modulation efficiency Extinction ratio	70 nW at -3.5 V 6.4 dB 250 \times 25 μm^2 10 GHz (foreseeable) 50 fJ/bit 6 dB		★	★
Foundry 2 \times 2 50:50 MMI ⁴⁹	IL Area	0.11 dB 36 \times 10 μm^2	★	★	
Customized 2 \times 2 50:50 Directional coupler	IL Area	0.05 dB 31 \times 6.5 μm^2			★
Foundry TO phase shifter ⁴⁹	IL Area Power	0.03 dB 75 \times 75 μm^2 $P_{\pi}=7$ mW	★	★	
Customized phase shifter	IL Area Power	0.05 dB 0.5 \times 33 μm^2 ~ 0 W			★
Customized 1 \times 10 splitter	IL Area	0.199 dB 34.6 \times 14.1 μm^2	★	★	★
Foundry 1 \times 2 50:50 MMI ⁴⁹	IL Area	0.1 dB 22 \times 10 μm^2	★	★	★
Foundry waveguide crossing ⁴⁹	IL Area	0.23 dB 8 \times 8 μm^2	★	★	★
Fiber/chip coupling	IL	2 dB	★	★	★
Laser	Wavelength	1550 nm	★	★	★

can realize comparable recognition and segmentation performance on edge AI tasks.

Noise Robustness Evaluation – To assess the robustness of our architecture against noise, we tested our speech recognition model with noise-aware training under various noise intensities injected in inference. Figure 14(b) indicates that our architecture demonstrates superior robustness to random noises. Even when increasing the relative noise intensity σ from 0 to 0.08, the accuracy drops by only 1%. Additionally, we measure the real noises in the chip testing in Fig. 14(a), which causes a negligible accuracy drop.

B. System Architecture-Level Performance Analysis

As a case study, we configure our architecture with 6×6 PTCs ($R = C = 6$), and each PTC is of size 32×32 ($K = 32$), working at a clock rate of 5 GHz. We give area and power estimation of our architecture.

Area Cost – The total area cost of a $K \times K$ PTC, including photonics and electronics, is estimated as follows

$$A = 2K \cdot A_{DAC} + 2K \cdot A_{MZM} + A_{1 \times 2K \text{ MMI}} + K^2 (A_{node} + A_{int} + A_{TIA} + A_{ADC}), \quad (17)$$

where each node area in the crossbar can be estimated by the bounding box $A_{node} = (L_{\text{splitter}} + 4\text{WBR} + W_{PD} + W_{\text{splitter}} + L_{\text{spacing}})(W_{\text{splitter}} + \text{WBR} + W_{PS} + L_{PD} + W_{\text{spacing}})$,

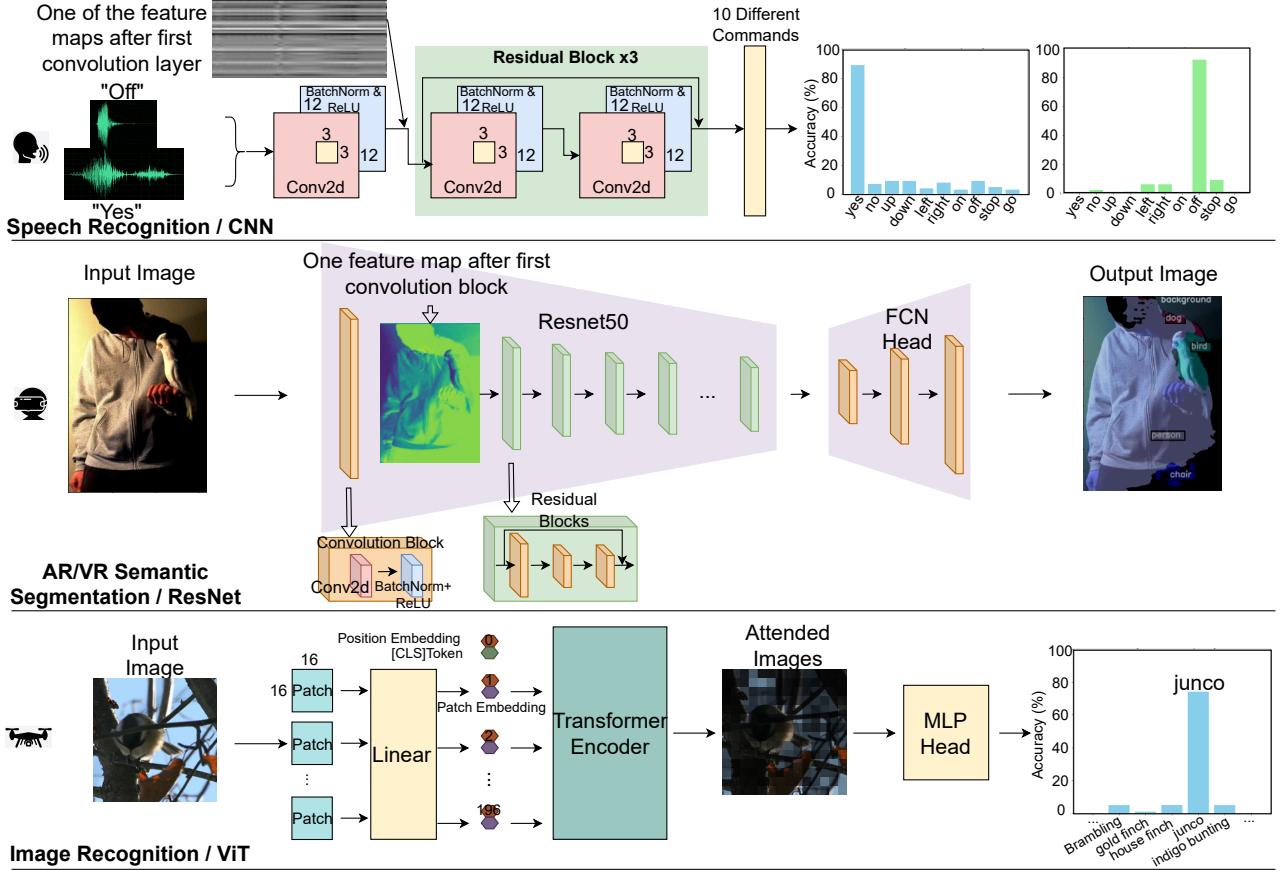


FIG. 13: Evaluation of our TeMPO accelerator on three edge machine learning tasks, including image recognition, voice keyword spotting, and semantic segmentation on CNNs and Vision Transformers (ViT). All optical NNs are trained with 6-bit weight/activation quantization and hardware noise injections.

TABLE III: Accuracy of TeMPO on 4 benchmarks with INT-6 weight/activation quantization and noise perturbation ($\sigma=0.01$).

Task	Dataset	Model	Fp32 Performance	INT6+Noise Acc
Image Recognition	ImageNet-1k ⁵¹	DeiT-Tiny ¹⁸	0.722 (Accuracy)	0.712 (Accuracy)
AR/VR Voice Keyword Spotting	Google Speech Command ⁵²	CNN ⁵²	0.957 (Accuracy)	0.929 (Accuracy)
AR/VR Semantic Segmentation	Pascal VOC2012 ⁵⁴	FCN R-50-D8 ⁵³	52.28 (mIoU)	51.16 (mIoU)

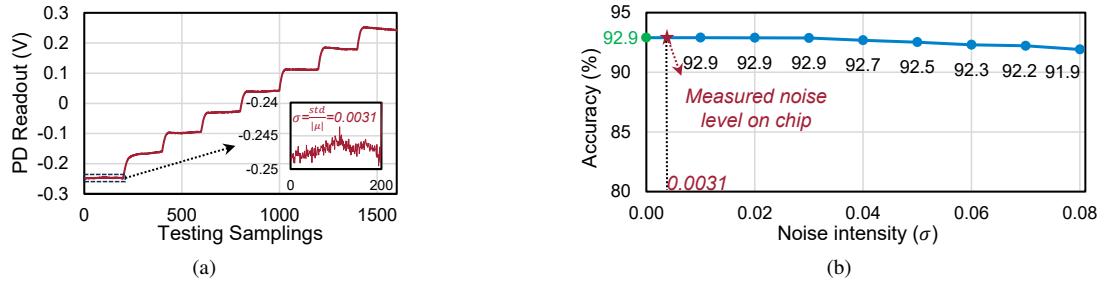


FIG. 14: (a) Noise measurement in experimental chip testing of SL-MZM. (b) Inference accuracy evaluation on the CNN speech command benchmark with various noise intensities (σ) from 0 to 0.08. The model is trained with the noise-aware quantization method. The noise intensity (0.0031) observed in the SL-MZM chip testing shows negligible accuracy impact.

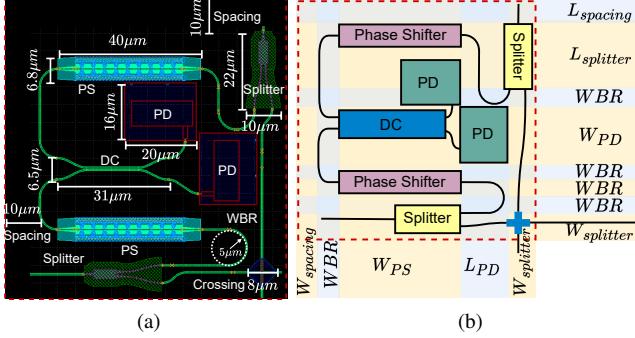


FIG. 15: (a) Layout of one dot-product engine (node). (b) Area breakdown for the node area A_{node} . WBR is denoted as waveguide bending radius, we use 5 μm as the WBR.

where WBR is the waveguide bending radius (set to 5 μm). Figure 15 shows the details of how we derived the node area. We draw the layout in Fig. 15(a) and show the dimension calculation details in Fig. 15(b). Other area terms can be directly obtained from the device area specifications. Note that the $1 \times 2K$ MMI is scaled based on our 1×10 MMI design, assuming length/width is proportional to fanout. Figure 16(a) shows the area comparison among 3 TeMPO variants. With Foundry-based high-speed E-O MZM, the PTC area is bulky, where the MZMs took almost 81% of the total circuit area. With our compact slow-light MZMs, the total area is reduced by 6.8 \times , while the MZMs only take 4.7% of the total area. Figure 17(a) further includes on-chip memory in the breakdown. Our customized architecture's area cost is 321 mm 2 , where 76.3% of the area is from the crossbar structure with minimum peripheral overhead from input encoding and data readout.

Power Consumption – We first give an analysis of the system-level on-chip power

$$P = 2K \cdot (P_{DAC} + P_{MZM}) + K^2 \cdot (2P_{PD} + P_{int} + P_{TIA} + P_{ADC}). \quad (18)$$

The DAC power can be derived by $P_{DAC} = \frac{P_0 b_0 2^b f}{2^b b f_s}$, where P_0 is the DAC power at b_0 -bit precision and f_s sampling rate, and f is the clock frequency. Other power terms can be directly obtained from the device power specification.

We emphasize the benefits of our multi-core architecture and temporal integration mechanism in power efficiency: ① Our multi-tile architecture can reduce the MZM and DAC power by a factor of R for matrix Y since the matrix Y modulation components are shared across R tiles before the on-chip waveguide broadcast, shown in Fig. 3. ② Multiple cores per tile share the same array of integrators, TIAs, and ADCs. Meanwhile, as we analyzed in Section III D 4, temporal integration can further reduce the TIA and ADC working frequency by a factor of T . Hence, the power of TIA/ADC can be overall reduced by CT times.

Figure 16(b) shows the power breakdown of the three variants of TeMPO. Compared to the foundry MZM, which takes 450 fJ to encode each symbol, our designed SL-MZM only takes 50 fJ to encode each symbol, leading to an 89% reduc-

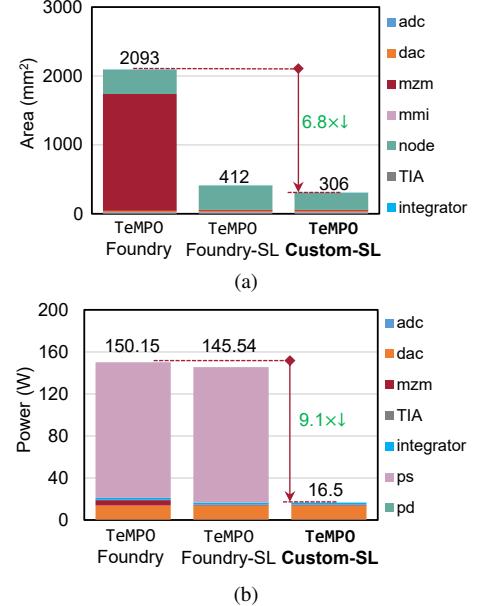


FIG. 16: (a) Area and (b) on-chip power breakdown of our proposed TeMPO across 3 different device configurations (6 \times 6 PTCs, each with a size of 32 \times 32) working at 5 GHz and 1550 nm wavelength. Note that memory is excluded. TeMPO with customized devices achieves 6.8 \times smaller area and 9.1 \times lower power compared to Foundry PDKs.

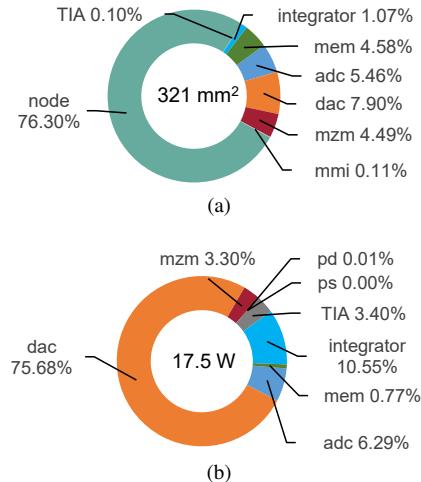


FIG. 17: (a) Area and (b) on-chip power breakdown of our TeMPO-custom-SL architecture (6 \times 6 PTCs, each with a size of 32 \times 32) working at 5 GHz and 1550 nm wavelength. Note that memory is included.

tion in the input tensor modulation power consumption. With time integration ($T = 60$), the ADC/TIA power is reduced by 60 \times , which becomes negligible (<5%) in the system power.

Overall, our optimized TeMPO-Custom-SL architecture equipped with energy-efficient SL-MZMs, customized splitters, phase shifters, and temporal integrators can reduce the on-chip system-level power by 9.1 \times compared to foundry PDK variants. Figure 17(b) indicates TeMPO-Custom-SL con-

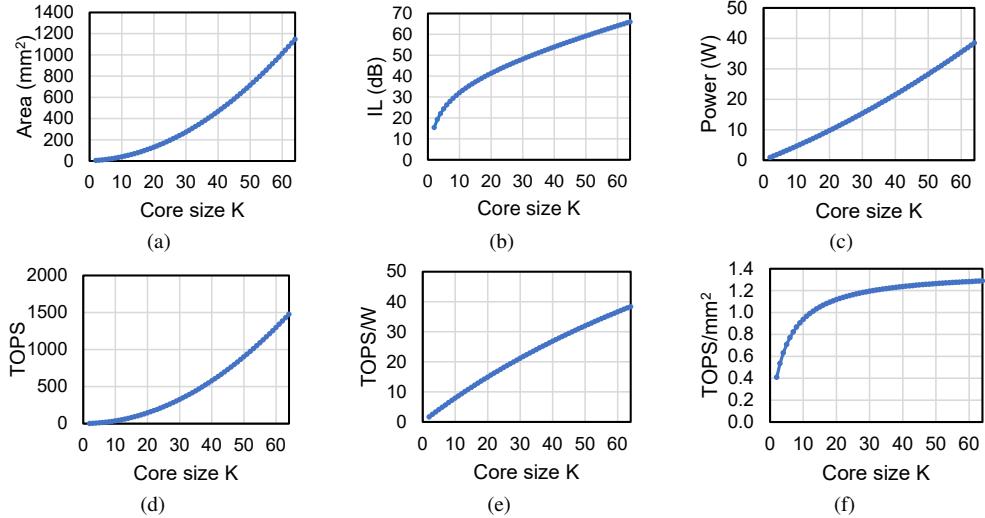


FIG. 18: (a) Area, (b) IL, (c) power, (d) computing speed (TOPS), (e) energy efficiency (TOPS/W) and (f) compute density (TOPS/mm²) with different PTC core size K of our TeMPO-custom-SL (6×6 cores) working at 5 GHz.

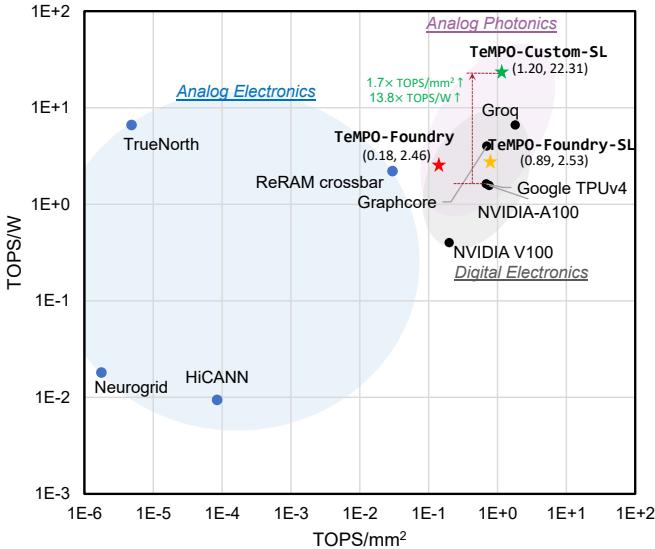


FIG. 19: Compare digital electronics (NVIDIA V100⁵⁶, A100 GPU⁵⁷, Google TPUv4⁵⁸, Groq⁵⁹, and Graphcore⁶⁰), analog electronics (IBM TrueNorth⁶¹, Neurogrid⁶², HiCANN⁶³, and ReRAM crossbar⁶⁴), and our photonic AI hardware TeMPO in compute density (TOPS/mm²) and energy efficiency (TOPS/W). Our TeMPO-Custom-SL with customized devices is at the Pareto front.

sumes 17.5 W power while 76% of power is from DACs. As technology continues advancing, power-efficient DACs are expected to significantly boost the efficiency of TeMPO further.

C. Tensor Core Efficiency and Scalability Analysis

In this section, we show a thorough analysis of the scalability of one PTC with different core sizes K . Besides area, insertion loss (IL), and power, we further define computing speed, energy efficiency, and compute density. To estimate the peak performance, we define the computing speed for each core as $2K^2ft/(T + T_{rst})$. Note that the reset overhead is considered as a scaling factor $T/(T + T_{rst})$. To evaluate the area efficiency, we adopt the metric of peak compute density, which measures how fast the hardware can compute per unit circuit area. For a TeMPO architecture ($R \times C$ cores) with $K \times K$ PTC, the peak compute density is evaluated as $\frac{2K^2RC}{fA(T+T_{rst})}$, where f is the clock frequency (no higher than the maximum ADC sampling rate, i.e., $f \leq f_{ADC,max}$). The energy efficiency of the hardware is defined as $\frac{2K^2RC}{fP}$ if we ignore energy cost during reset as the accelerator is idle, which measures how much energy it consumes to finish one operation.

Our TeMPO architecture has 6×6 PTCs, and each PTC core size varies from 2×2 to 64×64 . Figure 18(a) shows a nearly quadratic area scaling since most of the area is attributed to the crossbar structure with quadratically many dot-product engines. Figure 18(b) shows almost linear insertion loss scaling as the number of crossings and splitters linearly increases with the core size K . Hence, it is not efficient to use an overly large core size due to intractable insertion loss and laser power. In Fig. 18(c), we observe that power linearly scales with core size. Since the hardware power is dominated by DAC and we have a linear number of DAC to encode input vectors. Compared to quadratic power scaling in electronic circuits (as the transistor count quadratically increases with a larger K), this linear power scaling shows the advantage of photonic computing cores. Figure 18(d) shows the superior peak performance of our multi-core photonic accelerator. With 5 GHz

computing frequency and a core size of 30-40, TeMPO can potentially realize Peta operations per second (POPS)-level computing speed. Thanks to the quadratically increasing computing speed and the linear power scaling, TeMPO shows a consistent efficiency boost with a larger core size in Fig. 18(e). In terms of compute density, we can obtain a higher density with a larger core size, as indicated by Fig. 18(f). We expect a higher compute density in the future with more compact coupler and photodetector designs as technology advances. Overall, TeMPO shows good scalability to a larger core size. The ultimate upper bound of core size is from the insertion loss, which can be largely relaxed with customized low-loss optical components.

D. Efficiency Comparison with SoTA Accelerator Designs

We compare our designs with state-of-the-art (SoTA) electronic digital computers, including GPU, TPU, ASIC, and analog neuromorphic processors, e.g., IBM TrueNorth. We observe that our architecture TeMPO can realize competitive energy efficiency and compute density compared to state-of-the-art digital computers. However, standard foundry PDK devices are not the most efficient designs for photonic computing. By replacing the foundry MZM with our SL-MZM alone, we can boost the compute density from 0.18 (TeMPO-Foundry) to 0.89 (TeMPO-Foundry-SL) TOPS/mm². With customized SL-MZM, splitters, and phase shifters, our fully customized TeMPO-Custom-SL pushes the Pareto frontier to a record high level. It achieves 22.3 TOPS/W and 1.2 TOPS/mm², outperforming the foundry PDK variant by 9.1× higher energy efficiency and 6.8× higher compute density, respectively. Compared to NVIDIA A100 GPU and Google TPUs, TeMPO-Custom-SL shows 13.8× higher TOPS/W and 1.7× higher compute density, respectively.

V. CONCLUSION

In this work, we present TeMPO, a time-multiplexed dynamic photonic tensor accelerator designed for energy-efficient edge AI applications. Through careful co-design across device, circuit, and architecture layers, TeMPO achieves significant performance improvements compared to state-of-the-art electronic accelerators. Key innovations include customized slow-light Mach-Zehnder modulator, optical splitter, and phase shifters for low-power dynamic tensor computation, analog domain accumulation via capacitive temporal integration to eliminate analog-to-digital conversion bottleneck, and a multi-core architecture for efficient hardware sharing. TeMPO demonstrates comparable task accuracy with 6-bit quantization to digital counterparts, superior noise tolerance, and a peak performance of 368.6 TOPS, energy efficiency of 22.3 TOPS/W, and compute density of 1.2 TOPS/mm², pushing the Pareto frontier for edge AI hardware. This work establishes a new frontier in energy-efficient analog AI hardware, paving the path for future electronic-photonic accelerators in ubiquitous edge AI applications.

VI. DATA AVAILABILITY

The data that support the findings of this study are available within the article.

REFERENCES

- ¹Y. Shen, N. C. Harris, S. Skirlo, *et al.*, “Deep learning with coherent nanophotonic circuits,” *Nature Photonics* (2017).
- ²J. Gu, Z. Zhao, C. Feng, *et al.*, “Towards area-efficient optical neural networks: an FFT-based architecture,” in *IEEE/ACM Asia and South Pacific Design Automation Conference (ASPDAC)* (2020).
- ³C. Feng, J. Gu, H. Zhu, Z. Ying, Z. Zhao, D. Z. Pan, and R. T. Chen, “A compact butterfly-style silicon photonic-electronic neural chip for hardware-efficient deep learning,” *ACS Photonics* **9**, 3906–3916 (2022).
- ⁴J. Gu, H. Zhu, C. Feng, Z. Jiang, M. Liu, S. Zhang, R. T. Chen, and D. Z. Pan, “ADEPT: Automatic Differentiable DEsign of Photonic Tensor Cores,” in *ACM/IEEE Design Automation Conference (DAC)* (2022).
- ⁵H. Zhu, J. Gu, H. Wang, R. Tang, Z. Zhang, C. Feng, S. Han, R. T. Chen, and D. Z. Pan, “Lightening-Transformer: A Dynamically-operated Optically-interconnected Photonic Transformer Accelerator,” *arXiv preprint arXiv:2305.19533* (2023).
- ⁶H. H. Zhu, J. Zou, H. Zhang, Y. Z. Shi, S. B. Luo, *et al.*, “Space-efficient optical computing with an integrated chip diffractive neural network,” *Nature Communications* (2022).
- ⁷Z. Wang, L. Chang, F. Wang, T. Li, and T. Gu, “Integrated photonic meta-system for image classifications at telecommunication wavelength,” *Nat Commun* **13**, 2131 (2022).
- ⁸A. N. Tait, T. F. de Lima, E. Zhou, *et al.*, “Neuromorphic photonic networks using silicon photonic weight banks,” *Sci. Rep.* (2017).
- ⁹W. Liu, W. Liu, Y. Ye, Q. Lou, Y. Xie, and L. Jiang, “Holylight: A nanophotonic accelerator for deep learning in data centers,” in *IEEE/ACM Proceedings Design, Automation and Test in Europe (DATE)* (2019).
- ¹⁰F. Zokaee, Q. Lou, N. Youngblood, *et al.*, “LightBulb: A Photonic-Nonvolatile-Memory-based Accelerator for Binarized Convolutional Neural Networks,” in *IEEE/ACM Proceedings Design, Automation and Test in Europe (DATE)* (2020).
- ¹¹J. Gu, Z. Zhao, C. Feng, *et al.*, “Towards Hardware-Efficient Optical Neural Networks: Beyond FFT Architecture via Joint Learnability,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)* (2020).
- ¹²X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, R. Morandotti, A. Mitchell, and D. J. Moss, “11 TOPS photonic convolutional accelerator for optical neural networks,” *Nature* (2021).
- ¹³J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. L. Gallo, X. Fu, A. Lukashchuk, A. Raja, J. Liu, D. Wright, A. Sebastian, T. Kippenberg, W. Pernice, and H. Bhaskaran, “Parallel convolutional processing using an integrated photonic tensor core,” *Nature* (2021).
- ¹⁴B. Bai, Q. Yang, H. Shu, L. Chang, *et al.*, “Microcomb-based integrated photonic processing unit,” *Nature Communications* (2023).
- ¹⁵X. Xiao, M. B. On, T. Van Vaerenbergh, D. Liang, R. G. Beausoleil, and S. J. B. Yoo, “Large-scale and energy-efficient tensorized optical neural networks on III-V-on-silicon MOSCAP platform,” *APL Photonics* **6**, 126107 (2021).
- ¹⁶A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations* (2021).
- ¹⁷N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I* **16** (Springer, 2020) pp. 213–229.
- ¹⁸H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in *International conference on machine learning (PMLR, 2021)* pp. 10347–10357.

- ¹⁹J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT* (2019) pp. 4171–4186.
- ²⁰T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems* **33**, 1877–1901 (2020).
- ²¹OpenAI, "Gpt-4 technical report," (2023), [arXiv:2303.08774 \[cs.CL\]](https://arxiv.org/abs/2303.08774).
- ²²E. Timurdogan, Z. Su, C. V. Poulton, *et al.*, "AIM Process Design Kit (AIMPDKV2.0): Silicon Photonics Passive and Active Component Libraries on a 300mm Wafer," in *Optical Fiber Communication Conference* (2018).
- ²³R. Amin, R. Maiti, and Y. Gui, "Heterogeneously integrated ito plasmonic mach-zehnder interferometric modulator on soi," *Sci. Rep.* (2021), <https://doi.org/10.1038/s41598-020-80381-3>.
- ²⁴A. Mirza, F. Sunny, P. Walsh, K. Hassan, S. Pasricha, and M. Nikdast, "Silicon Photonic Microring Resonators: A Comprehensive Design-Space Exploration and Optimization under Fabrication-Process Variations," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 1–1 (2021).
- ²⁵J. Gu, C. Feng, H. Zhu, Z. Zhao, Z. Ying, M. Liu, R. T. Chen, and D. Z. Pan, "SqueezeLight: A Multi-Operand Ring-Based Optical Neural Network with Cross-Layer Scalability," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)* (2022).
- ²⁶J. Gu, C. Feng, Z. Zhao, Z. Ying, M. Liu, R. T. Chen, and D. Z. Pan, "SqueezeLight: Towards Scalable Optical Neural Networks with Multi-Operand Ring Resonators," in *IEEE/ACM Proceedings Design, Automation and Test in Europe (DATE)* (2021).
- ²⁷P. C. Eng, S. Song, and B. Ping, "State-of-the-art photodetectors for optoelectronic integration at telecommunication wavelength," *Nanophotonics* **4**, 277–302 (2015).
- ²⁸X. Zhao, G. Wang, H. Lin, Y. Du, X. Luo, Z. Kong, J. Su, J. Li, W. Xiong, Y. Miao, *et al.*, "High performance pin photodetectors on ge-on-insulator platform," *Nanomaterials* **11**, 1125 (2021).
- ²⁹"O-band 4 channel dfb laser source."
- ³⁰T. Barwicz, T. W. Lichoulas, Y. Taira, Y. Martin, S. Takenobu, A. Janta-Polczynski, H. Numata, E. L. Kimbrell, J.-W. Nah, B. Peng, *et al.*, "Automated, high-throughput photonic packaging," *Optical Fiber Technology* **44**, 24–35 (2018).
- ³¹S. Khan, S. M. Buckley, J. Chiles, R. P. Mirin, S. W. Nam, and J. M. Shainline, "Low-loss, high-bandwidth fiber-to-chip coupling using capped adiabatic tapered fibers," *APL Photonics* **5** (2020).
- ³²J. Nauriyal, M. Song, R. Yu, and J. Cardenas, "Fiber-to-chip fusion splicing for low-loss photonic packaging," *Optica* **6**, 549–552 (2019).
- ³³A. Chen, A. Begović, S. Anderson, and Z. R. Huang, "On-chip slow-light sin bragg grating waveguides," *IEEE Photonics Journal* **14**, 1–6 (2022).
- ³⁴S. R. Anderson, A. Begović, H. Jiang, and Z. R. Huang, "Compact slow-light integrated silicon electro-optic modulators with low driving voltage," *IEEE Photonics Technology Letters* **35**, 697–700 (2023).
- ³⁵S. R. Anderson, A. Begović, and Z. R. Huang, "Integrated slow-light enhanced silicon photonic modulators for rf photonic links," *IEEE Photonics Journal* **14**, 1–6 (2022).
- ³⁶M. Zhang, A. Begović, D. Yin, N. Gangi, J. Gu, and Z. R. Huang, "Foundry manufactured 6-bit resolution, 150um long slow-light electro-optic modulator for on-chip photonic tensor computing," in *submitted to 2024 Conference on Lasers and Electro-Optics (CLEO)* (IEEE, 2024).
- ³⁷M. Caverley, X. Wang, K. Murray, N. A. F. Jaeger, and L. Chrostowski, "Silicon-on-insulator modulators using a quarter-wave phase-shifted bragg grating," *IEEE Photonics Technology Letters* **27**, 2331–2334 (2015).
- ³⁸Y. Hinakura, Y. Terada, H. Arai, and T. Baba, "Electro-optic phase matching in a si photonic crystal slow light modulator using meander-line electrodes," *Optics express* **26**, 11538–11545 (2018).
- ³⁹S. Khan and S. Fathpour, "Complementary apodized grating waveguides for tunable optical delay lines," *Optics express* **20**, 19859–19867 (2012).
- ⁴⁰Y. Zhang, S. Yang, A. E.-J. Lim, G.-Q. Lo, C. Galland, T. Baehr-Jones, and M. Hochberg, "A compact and low loss y-junction for submicron silicon waveguide," *Optics express* **21**, 1310–1316 (2013).
- ⁴¹M. van Niekerk, J. A. Steidle, G. A. Howland, M. L. Fanto, N. Soures, F. T. Zohora, D. Kudithipudi, and S. Preble, "Approximating large scale arbitrary unitaries with integrated multimode interferometers," in *Quantum Inf. Sci. Sens. Comput. XI* (2019) p. 20.
- ⁴²Y. Liu, Z. Li, S. Wang, N. Zhang, Y. Yao, J. Du, Z. He, Q. Song, and K. Xu, "Ultra-compact and polarization-insensitive mmi coupler based on inverse design," in *Proc. IEEE OFC* (2019).
- ⁴³H. Sattari, A. Y. Takabayashi, Y. Zhang, P. Verheyen, W. Bogaerts, and N. Quack, "Compact broadband suspended silicon photonic directional coupler," *Optics Letters* **45**, 2997–3000 (2020).
- ⁴⁴R. Yao, H. Li, B. Zhang, W. Chen, P. Wang, S. Dai, Y. Liu, J. Li, Y. Li, Q. Fu, *et al.*, "Compact and low-insertion-loss 1×n power splitter in silicon photonics," *Journal of Lightwave Technology* **39**, 6253–6259 (2021).
- ⁴⁵L. Soldano and E. Pennings, "Optical multi-mode interference devices based on self-imaging: Principles and applications," *J. Lightwave Technol.* **13**, 615–627 (1995).
- ⁴⁶N. C. Harris, Y. Ma, J. Mower, T. Baehr-Jones, D. Englund, M. Hochberg, and C. Galland, "Efficient, compact and low loss thermo-optic phase shifter in silicon," *Opt. Express* **22**, 10487–10493 (2014).
- ⁴⁷P. Caragiulo, O. E. Mattia, A. Arbabian, and B. Murmann, "A compact 14 gs/s 8-bit switched-capacitor dac in 16 nm finfet cmos," *2020 IEEE Symposium on VLSI Circuits*, 1–2 (2020).
- ⁴⁸J. Liu, M. Hassanpourghadi, and M. S.-W. Chen, "A 10gs/s 8b 25fj/c-s 2850um2 two-step time-domain adc using delay-tracking pipelined-sar tdc with 500fs time step in 14nm cmos technology," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, Vol. 65 (2022) pp. 160–162.
- ⁴⁹M. Rakowski, C. Meagher, K. Nummy, A. Aboketaf, J. Ayala, Y. Bian, B. Harris, K. McLean, K. McStay, A. Sahin, L. Medina, B. Peng, Z. Sowinski, A. Stricker, T. Houghton, C. Hedges, K. Giewont, A. Jacob, T. Letavic, D. Riggs, A. Yu, and J. Pellerin, "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.
- ⁵⁰M. Rakowski, Y. Ban, P. De Heyn, N. Pantano, B. Snyder, S. Balakrishnan, S. Van Huylenbroeck, L. Bogaerts, C. Demeurisse, F. Inoue, *et al.*, "Hybrid 14nm finfet-silicon photonics technology for low-power tb/s/mm² optical i/o," in *2018 IEEE Symposium on VLSI Technology* (IEEE, 2018) pp. 221–222.
- ⁵¹J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2009) pp. 248–255.
- ⁵²R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018) pp. 5479–5483.
- ⁵³J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015).
- ⁵⁴M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>.
- ⁵⁵S. K. Esser, J. L. McKinstry, D. Bablani, R. Appuswamy, and D. S. Modha, "Learned step size quantization," in *International Conference on Learning Representations (ICLR)* (2020).
- ⁵⁶J. Choquette, O. Giroux, and D. Foley, "Volta: Performance and programmability," *IEEE Micro* **38**, 42–52 (2018).
- ⁵⁷J. Choquette, W. Gandhi, O. Giroux, N. Stam, and R. Krashinsky, "Nvidia a100 tensor core gpu: Performance and innovation," *IEEE Micro* **41**, 29–35 (2021).
- ⁵⁸N. P. Jouppi, G. Kurian, S. Li, P. Ma, R. Nagarajan, L. Nai, N. Patil, S. Subramanian, A. Swing, B. Towles, C. Young, X. Zhou, Z. Zhou, and D. Patterson, "Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings," (2023), [arXiv:2304.01433 \[cs.AR\]](https://arxiv.org/abs/2304.01433).
- ⁵⁹L. Gwennap, "Groq rocks neural networks, microprocessor report," (2020), <http://groq.com/wp-content/uploads/2020/04/Groq-RocksNNs-Linley-Group-MPR-2020Jan06.pdf>.
- ⁶⁰I. Kacher, M. Portaz, H. Randrianarivo, and S. Peyronnet, "Graphcore c2 card performance for image-based deep learning application: A report," (2020), [arXiv:2002.11670 \[cs.CV\]](https://arxiv.org/abs/2002.11670).
- ⁶¹M. V. DeBole, B. Taba, A. Amir, F. Akopyan, A. Andreopoulos, W. P. Risk, J. Kusnitz, C. Ortega Otero, T. K. Nayak, R. Appuswamy, P. J. Carlson, A. S. Cassidy, P. Datta, S. K. Esser, G. J. Garreau, K. L. Holland,

- S. Lekuch, M. Mastro, J. McKinstry, C. di Nolfo, B. Paulovicks, J. Sawada, K. Schleupen, B. G. Shaw, J. L. Klamo, M. D. Flickner, J. V. Arthur, and D. S. Modha, “Truenorth: Accelerating from zero to 64 million neurons in 10 years,” *Computer* **52**, 20–29 (2019).
- ⁶²B. V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J.-M. Bussat, R. Alvarez-Icaza, J. V. Arthur, P. A. Merolla, and K. Boahen, “Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations,” *Proceedings of the IEEE* **102**, 699–716 (2014).
- ⁶³J. Schemmel, D. Brüderle, A. Grübl, M. Hock, K. Meier, and S. Millner, “A wafer-scale neuromorphic hardware system for large-scale neural modeling,” in *2010 IEEE International Symposium on Circuits and Systems (ISCAS)* (2010) pp. 1947–1950.
- ⁶⁴M. Giordano, K. Prabhu, K. Koul, R. M. Radway, A. Gural, R. Doshi, Z. F. Khan, J. W. Kustin, T. Liu, G. B. Lopes, V. Turbiner, W.-S. Khwa, Y.-D. Chih, M.-F. Chang, G. Lallement, B. Murmann, S. Mitra, and P. Raina, “Chimera: A 0.92 tops, 2.2 tops/w edge ai accelerator with 2 mbyte on-chip foundry resistive ram for efficient training and inference,” in *2021 Symposium on VLSI Circuits* (2021) pp. 1–2.