Research Article

Chenghao Feng, Jiaqi Gu, Hanqing Zhu, Shupeng Ning, Rongxing Tang, May Hlaing, Jason Midkiff, Sourabh Jain, David Z. Pan and Ray T. Chen*

Integrated multi-operand optical neurons for scalable and hardware-efficient deep learning

https://doi.org/10.1515/nanoph-2023-0554
Received August 30, 2023; accepted December 6, 2023; published online January 8, 2024

Abstract: Optical neural networks (ONNs) are promising hardware platforms for next-generation neuromorphic computing due to their high parallelism, low latency, and low energy consumption. However, previous integrated photonic tensor cores (PTCs) consume numerous single-operand optical modulators for signal and weight encoding, leading to large area costs and high propagation loss to implement large tensor operations. This work proposes a scalable and efficient optical dot-product engine based on customized multi-operand photonic devices, namely multi-operand optical neuron (MOON). We experimentally demonstrate the utility of a MOON using a multi-operand-Mach–Zehnder-interferometer (MOMZI) in image recognition tasks. Specifically, our MOMZI-based ONN achieves a measured accuracy of 85.89 % in the street view house number (SVHN) recognition dataset with 4-bit voltage control precision. Furthermore, our performance analysis reveals that a 128 × 128 MOMZI-based PTCs outperform their counterparts based on single-operand MZIs by one to two order-of-magnitudes in propagation loss, optical delay, and total device footprint, with comparable matrix expressivity.

Keywords: multi-operand optical neuron; hardware efficiency; deep learning; photonic tensor core

1 Introduction

Optical neural network (ONN) is an emerging analog artificial intelligence (AI) accelerator that leverages properties of photons, including low latency, wide bandwidth, and high parallelism [1]–[3], to address the growing demand for computing power required to implement deep neural network (DNN) models. Once weight parameters are set, photonic integrated circuits (PICs) can perform tensor operations with near-zero energy consumption at the speed of light [4], [5], making them an ideal platform for accelerating multiply-accumulate (MAC) operations [6]. However, the potential massive parallelism and ultra-high computing speed of ONNs are not fully unleashed with small-size photonic tensor cores (PTCs). To maximize the performance benefit of photonic computing in DNN acceleration, scalable and efficient photonic tensor core designs are in high demand.

The scalability of previous photonic tensor core designs is bottlenecked by the large spatial footprint and insertion loss [7]. For instance, an MZI-based coherent PTC [8] require $O(m^2 + n^2)$ single-operand MZI modulators to construct an $m \times n$ matrix, consuming huge area cost to implement large tensor operations (e.g., 128×128). Moreover, the large number ($\sim 2n$) of cascaded optical devices in the critical path of the circuit leads to unacceptable insertion loss. Even with low-loss MZIs such as thermo-optic MZIs (0.5–1 dB) [9], cascading 128 such devices will result in 64–128 dB propagation loss. In addition, single-operand-device-based PTCs suffer from nontrivial dynamic energy consumption to reconfigure weight parameters. Given the limited chip area and link budget, we have to serialize the matrix multiplication by repeatedly reusing small-size photonic tensor cores, which incurs much longer latency to implement one matrix-vector

***Corresponding author: Ray T. Chen**, Microelectronics Research Center, The University of Texas at Austin, Austin, TX 78758, USA; Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78705, USA; and Omega Optics, Inc., 8500 Shoal Creek Blvd., Bldg. 4, Suite 200, Austin, TX 78757, USA, E-mail: chenrt@austin.utexas.edu
Chenghao Feng, Microelectronics Research Center, The University of Texas at Austin, Austin, TX 78758, USA; and Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78705, USA. https://orcid.org/0000-0002-0751-7681 (C. Feng)
Jiaqi Gu, Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78705, USA; and School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287, USA
Hanqing Zhu, **May Hlaing**, **Jason Midkiff and David Z. Pan**, Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78705, USA
Shupeng Ning, **Rongxing Tang and Sourabh Jain**, Microelectronics Research Center, The University of Texas at Austin, Austin, TX 78758, USA

multiplication, potentially negating the speed advantage of ONNs over electronic analog AI accelerators [10].

Both circuit- and device-level optimizations have been explored to enhance the scalability of ONNs. Circuit-level approaches, such as the butterfly-style circuit mesh [11], have been explored to reduce hardware usage [12], [13]. Moreover, compact device-level photonic tensor cores, such as star couplers and metasurfaces [14], [15], have been proposed to significantly reduce the device footprint and improve the hardware efficiency of tensor operations. However, one major challenge with compact photonic circuit mesh or passive device-level tensor cores is their limited matrix representability, which usually results in accuracy degradation when implementing complicated AI tasks. To address this challenge, we suggest using active device-level photonic tensor cores, which offer the potential to achieve both high representability and high hardware efficiency. Recently there has been a trend to use multi-operand devices for vector operations, which shows great potential to achieve efficiency and scalable breakthroughs [16]. In multi-operand devices, we partition the phase shifter into multiple small segments, each being independently controlled. By leveraging the underlying device transfer function, we can then realize vector operations with nearly the same device footprint and tuning range as the single-operand one. In this work, for the first time, we officially name this photonic structure a multi-operand optical neuron (MOON). Prior work has proposed a microring-based MOON and showed its advantages over standard single-operand micro-ring in neuromorphic computing through simulation [16]. In this work, we introduce a new broadband device in this MOON-family, a multi-operand MZI (MOMZI), and experimentally demonstrate its superior efficiency and scalability for next-generation photonic neuromorphic computing.

In this work, we customize a MOMZI, whose modulation arm is controlled by multiple independent signals, and leverage its transmission to realize vector-vector dot-product. A k-operand (k-op) MOMZI can be used as a length-k vector dot-product engine, directly saving the MZI device usage by a factor of k compared to single-operand MZI arrays [8]. Note that the MZI device footprint and tuning range keep constant and will not scale with k. By combing the result from multiple k-op MOMZIs, we can efficiently scale up to operations with a large vector length with near-constant insertion loss. Using devices from foundry process design kits (PDKs) [17], 128 × 128 photonic tensor cores based on our MOMZIs show a 6.2× smaller total device footprint, 49× lower optical delay, and >256 dB lower propagation loss than previous single-operand MZI arrays [8]. We experimentally demonstrated the representability and trainability of an ONN constructed by 4-op MOMZIs on the street view house number (SVHN) recognition task [18], achieving a measured accuracy of 85.89 % with 4-bit voltage control precision. Our proposed MOMZI-based photonic tensor core enables the implementation of high-performance and energy-efficient neuromorphic computing with a small device footprint, low propagation delay, and low energy consumption.

2 Multi-operand optical neurons

A typical photonic tensor core to implement MAC operation is in Figure 1(a), which contains photonic components to generate input signals, the weight matrix, and the outputs. n high-speed modulators are needed in an n-input, m-output layer. Depending on the weight mapping approach, one needs $\frac{m(m-1)+n(n-1)}{2} + \max(m, n)$ [8] or $m \times n$ active photonic components [19] to implement a $m \times n$ weight matrix. Furthermore, $\sim 2n$ active devices are cascaded in one optical path, resulting in non-negligible propagation loss and requiring more laser power to drive the photonic neural chip.

In this study, we propose a novel approach to reduce the optical component usage by implementing the multiply-accumulate (MAC) operation using an array of multi-operand-modulator-based optical neurons (MOONs), as shown in Figure 1(b). Depending on the area and reliability concerns, one MOON can be a multi-operand active photonic device of any waveguide structure, such as MZI modulators and microring modulators. As illustrated in Figure 1(c) and (d), each row of the layer is divided into $n^* = \frac{n}{k}$ k-operand modulators, and the output of each k-operand modulator is accumulated using on-chip combiners or multiplexers to compute the final output of each row. Consequently, the total number of MOONs required for an n-input, m-output layer is $\frac{mn}{k}$, significantly reducing the number of active optical components.

Unlike conventional PTCs designed for general matrix multiplications (GEMMs), the nonlinear transfer function between the electrical signal and the transmission of the MOON needs to be considered when training DNN models. The input vector $\boldsymbol{x}_{\mathbf{in}}$ is encoded as the amplitude of the optical signals and will also be partitioned into $n^* = \frac{n}{k}$ length-k segments $\boldsymbol{x}_{\mathbf{in}} = \left(\boldsymbol{x}_{\mathbf{in}}^1, \boldsymbol{x}_{\mathbf{in}}^2, \ldots, \boldsymbol{x}_{\mathbf{in}}^{n^*}\right)$. Each segment is encoded on one MOON to implement one k-length vector-vector inner product. Thus, the output signals of one layer can be expressed as follows:

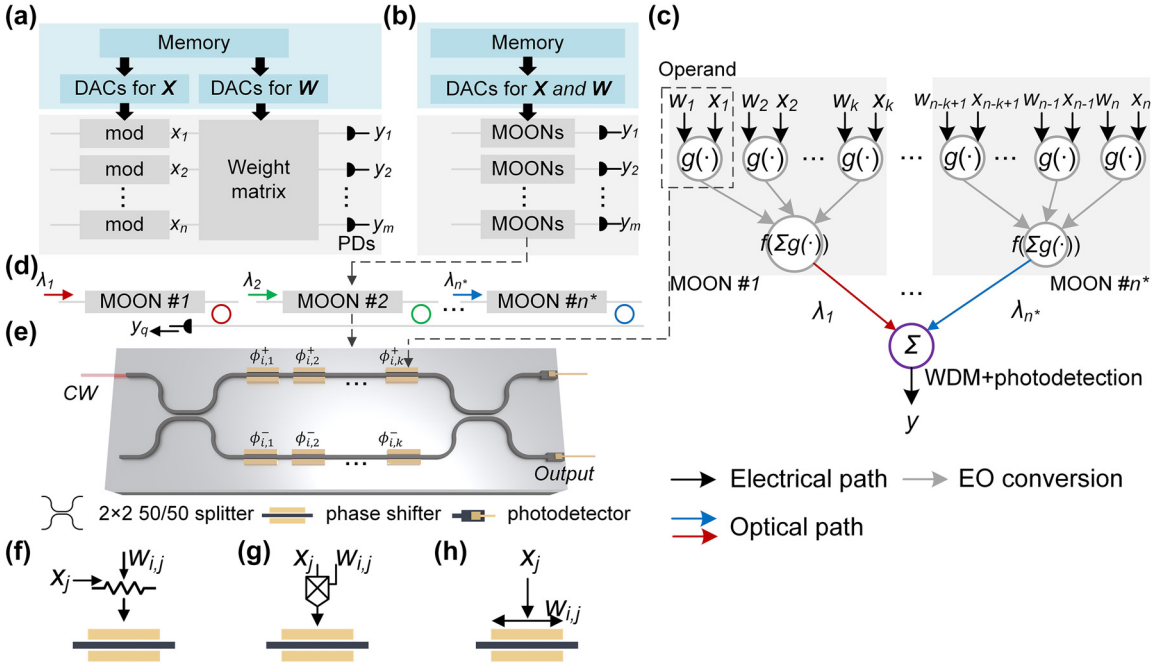

Figure 1: General architecture of the MOON-based photonic tensor core. (a) A conventional photonic tensor core based on single-operand modulators, which has an array of input modulators and $O(mn)$ photonic devices to construct the weight matrix. (b) Schematic of the MOON-based PTC to implement an n-input, m-output layer. (c) Shows the diagram of using $n^* = \frac{n}{k}$ k-operand MOONs to implement a length-n vector operations, and its circuit structure is shown in (d). In each MOON, the weight signals and input signals are operated simultaneously on each operand. The scalar multiplication and partial accumulation are implemented during electrical-to-optical (EO) conversion. The output is obtained by accumulating the output signal of each MOON using multiplexers and photodetectors. (e) Schematic of a k-operand MOMZI-based MOON, which consists of k operands on each arm. There are various approaches to encoding weight signals w_i and input signals x_i on each modulation region (operand). To realize $\phi_i = g(w_i x_i)$ on MOMZI-based MOON, one can use (f) programmable resistances to encode w_i and current signals to encode x_i, or (g) tunable amplifiers/attenuators to encode w_i and voltage signal to encode x_i, or (h) adjust modulation length to encode fixed w_i and voltage signals to encode x_i.

$$\boldsymbol{x}'_{\mathbf{out}} = \mathcal{F}(\boldsymbol{W}, \boldsymbol{x}_{\mathbf{in}}) = \begin{pmatrix} \sum_{i=1}^{n^*} f\left(\sum_{j=1}^{k} g\left(W_{1,j+(i-1)k}, x_{in}^{j+(i-1)k} \right) \right) \\ \sum_{i=1}^{n^*} f\left(\sum_{j=1}^{k} g\left(W_{2,j+(i-1)k}, x_{in}^{j+(i-1)k} \right) \right) \\ \vdots \\ \sum_{i=1}^{n^*} f\left(\sum_{j=1}^{k} g\left(W_{m,j+(i-1)k}, x_{in}^{j+(i-1)k} \right) \right) \end{pmatrix}. \quad (1)$$

where function $f(\cdot)$ represents the relationship between the total phase shift or amplitude response of all the operands and the optical output signal of each MOON, whereas $g(w_i, x_i)$ is determined by the weight/signal encoding way and each operand's phase/amplitude response. In this work, we use $g(w_i, x_i) = g(w_i * x_i)$, where we encode $V_i = w_i \cdot x_i$ as the operating voltage on each operand. $\mathcal{F}(\bullet)$ is the output result of the layer with weight and input signals W and x_{in}. The mechanism of the MOON-based PTC is shown in Figure 1(c). As depicted in Figure 1(f)–(h), w_i can be encoded by programmable resistances (e.g., memristors or phase change materials [20]), tunable electrical amplifiers/attenuators, or the length of modulation arms if the weights are fixed. x_i refers to the input current or voltage signals from input sources or the previous layer. After obtaining the transfer function of MOON (Eq. (1)), one can deploy them in commercial deep learning platforms, e.g., Pytorch, to train MOON-based PTCs.

Our MOON-based PTC significantly improves computational efficiency compared to previous GEMM-based PTCs [8]. A k-operand MOMZI has a similar device footprint and dynamic tuning range to a single-operand MZI, but it can implement k MACs. This outperforms a single-operand MZI in area- and energy- efficiency since it can only perform approximately one MAC operation per MZI device in single-operand MZI-based PTCs. To be more specific, if the total dynamic phase tuning range is $\Sigma\Delta\phi = \pi$, a single-operand MZI with a length-L_0 phase shifter is the same with a k-op MOMZI with k length-$\frac{L_0}{k}$ phase shifters in the phase tuning region's area, which dominates high-speed MZI's footprint. Moreover, the phase-tuning range of each operand is $\frac{\pi}{k}$, hence, the energy consumption of a k-op MOMZI is the same

as that of a single-operand high-speed MZI. One only needs $k-1$ additional waveguides to connect the operands, whose footprints are negligible compared to active phase shifters. The advantage of MOONs lies in their ability to perform multiple MAC operations using a single MZI device, making them more computationally efficient than previous ONNs.

Moreover, as shown in Figure 1(d), only one MOON is cascaded in one optical path of our circuit architecture, resulting in much smaller propagation loss compared to MZI-based or microring-based ONNs, where $2n+1$ MZIs or n microrings are cascaded. As a result, we can deploy compact but lossy optical modulators, e.g., plasmonic-on-silicon modulators [21], as MOONs in our PTC, trading higher insertion loss for a much smaller chip footprint and lower modulation power. Detailed performance evaluations will be provided in our discussions.

3 Multi-operand-MZI-based optical neural network

In this work, we demonstrate the use of k-operand MZI modulators as the fundamental building blocks for constructing our MOMZI-PTC. Figure 1(e) shows the structure of a MOMZI. Unlike the traditional MZI modulators with one or two phase modulators, a k-op MOMZI has k active phase shifters on each modulation arm, and each phase shifter is controlled by an independent signal. This structure is similar to lumped-segment MZIs used in optical communications [22], but the driving signals on each operand are independent and analog. For MZI modulators with dual modulation arms, the total number of operands can increase to $2k$ to enable both positive and negative phase shifts. Suppose each shifter contributes to a phase shift ϕ_i, the output intensity of a MOMZI can be expressed as:

$$\begin{aligned} y_i &= f\left(\sum \phi_i\right) \\ &= f\left(\sum_{i=1}^{k}\phi_i^+ - \sum_{i=1}^{k}\phi_i^-\right) \\ &= \cos^2\left(\frac{\sum_{i=1}^{k}\phi_i^+ - \sum_{i=1}^{k}\phi_i^- + \phi_b}{2}\right) \\ &= \frac{1}{2}\cos\left(\sum_{i=1}^{k}\phi_i^+ - \sum_{i=1}^{k}\phi_i^- + \phi_b\right) + \frac{1}{2} \end{aligned} \tag{2}$$

where $f(\cdot)=\cos^2\left(\frac{\cdot}{2}\right)$, ϕ_i^+ denotes the ith phase shifters on the upper arm of the modulator, and ϕ_i^- denotes that on the lower arm. Consequently, positive weight signals are encoded on upper modulation arms, while negative weight signals are encoded on the lower ones. ϕ_b is the phase bias when no input signals are operated on the modulation arms, which is used to tune the transfer function of the MOMZI.

The modulation mechanism of the MOMZI plays a critical role in determining their transfer function with an input voltage signal. As shown in Figure 2, the transfer function of MZI modulators using the same foundry [17] can exhibit sinusoidal, quadratic (linear field intensity response), or other nonlinear transfer functions with the operating voltage V. The specific shape of the transfer function depends on the modulation mechanism ($\phi_i^{\pm}(V)$) and the modulator's waveguide structure ($f(\cdot)$). By optimizing these parameters, one can customize the transfer function of the MOMZI to realize certain nonlinear activation functions of DNNs. We will discuss this hereinafter.

Supposing the dot product information $w_i \cdot x_i$ is directly encoded as the operating voltage V_i on each operand of the MOMZI, we can rewrite Eq. (2) as Eq. (3):

$$\begin{aligned} \boldsymbol{x}'_{\mathbf{out}} &= \mathcal{F}\left(\boldsymbol{W}, \boldsymbol{x}_{\mathbf{in}}\right) \\ &= \frac{1}{2}\begin{pmatrix} \sum_{i=1}^{n^*}\left(\cos\left(\sum_{j=1}^{k}\phi(W_{1,j+(i-1)k}x_{in}^{j+(i-1)k}) + \phi_b^{1,i}\right)\right) \\ \sum_{i=1}^{n^*}\left(\cos\left(\sum_{j=1}^{k}\phi(W_{2,j+(i-1)k}x_{in}^{j+(i-1)k}) + \phi_b^{2,i}\right)\right) \\ \vdots \\ \sum_{i=1}^{n^*}\left(\cos\left(\sum_{j=1}^{k}\phi(W_{m,j+(i-1)k}x_{in}^{j+(i-1)k}) + \phi_b^{m,i}\right)\right) \end{pmatrix} + \boldsymbol{b} \end{aligned} \tag{3}$$

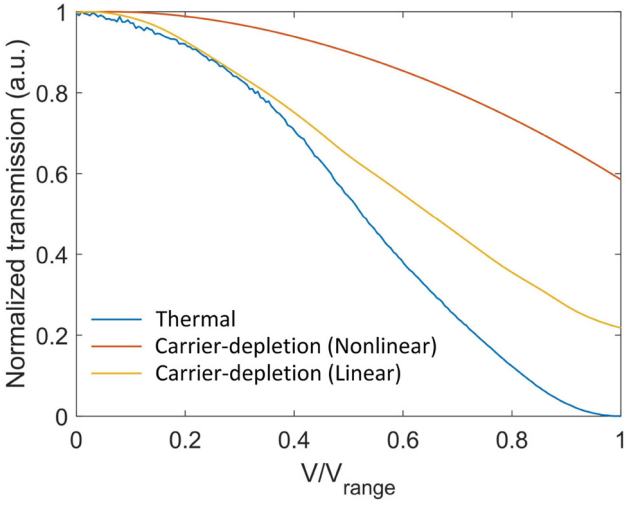

Figure 2: Transfer function of different MZI modulators under different modulation mechanisms. All the data are experimental data from our measurement or the process design kit (PDK) model [17] on Lumerical interconnect. V_{range} is the maximum allowed operating voltage.

In Eq. (3), positive or negative phase shifts are achieved by applying the operating voltages to each phase shifter's upper or lower arm. The phase bias $\phi_b^{p,i}$ of the ith MOMZI on row p of MOMZI-PTC can be adjusted to improve the expressivity of our neural architecture. The constant $\boldsymbol{b} = \frac{n}{2k}$ can be eliminated after photodetection. Using Eq. (3) we can model the MOMZI on commercial deep learning platforms, e.g., PyTorch, making it practical to train and deploy the DNN.

4 Experimental results

In this study, we designed and fabricated a 4-op MOMZI that is capable of implementing a 4×1 vector operation on the silicon photonics platform. This experimental demonstration aims to investigate if the actual performance of MOMZI devices is trainable and learnable to perform deep learning tasks. Additionally, the essential components required for the deployment of MOMZIs in PTCs, including DACs with tunable gains, on-chip combiners, and electrical control circuits, are readily accessible through established foundry services and existing technologies. This paves the way for future large-scale integration of MOMZI-PTCs. The chip layout was drawn and verified using Synopsys OptoDesigner (version 2021) and then fabricated by AIM Photonics. The schematic of the MOMZI is illustrated in Figure 3(b), while Figure 3(a) shows close-up images of its components, including phase shifters, 50–50 directional couplers, and photodetectors.

We use two phase shifters on each modulation arm to enable both positive and negative weights during training. The maximum operating voltage is $V_{\max} \cong \frac{1}{2}V_\pi$ as the tuning range of each phase shifter is $\sim \frac{\pi}{4}$. If adjustable modulation length is allowed in foundries, it is suggested to reduce the length of each operand by 4× to minimize the 4-op MOMZI's device footprint. In experiments, we encode $\phi_i \propto w_i^{\pm} \cdot x_i$, where w_1^+ and w_2^+ are positive weights and w_3^- and w_4^- are absolute values of negative weights, w_1 and w_2 are encoded on the upper arm, while w_3^- and w_4^- are encoded on the downer arm. The transfer function of our modulator can then be written as:

$$\begin{aligned} T &= f\left(\sum_i \phi_i + \phi_b\right) = \cos^2\left[\left(\phi_1\left(w_1^+ \cdot x_1\right) + \phi_2\left(w_2^+ \cdot x_2\right)\right.\right. \\ &\quad \left.\left. - \phi_3\left(w_1^- \cdot x_3\right) - \phi_4\left(w_2^- \cdot x_4\right) + \phi_b\right)/2\right] \\ &= \frac{1}{2}\cos\left[\phi_1\left(w_1^+ \cdot x_1\right) + \phi_2\left(w_2^+ \cdot x_2\right)\right. \\ &\quad \left. - \phi_3\left(w_3^- \cdot x_3\right) - \phi_4\left(w_4^- \cdot x_4\right) + \phi_b\right] + \frac{1}{2} \end{aligned} \tag{4}$$

We tune one additional phase shifter on the upper arm to let $\phi_b \approx \frac{\pi}{2}$ to obtain a relatively linear and balanced output range. The model to implement a length-n vector dot-product with our 4-op MOMZI can be derived from Eq. (3) and Figure 1(c):

$$\begin{aligned} T &= \sum_{\mathrm{j}=1}^{n^*} f\left(\sum_i \phi_{j,i} + \phi_b\right) \\ &= \frac{1}{2}\sum_{\mathrm{j}=1}^{n^*}\left(\cos\left[\phi_{\mathrm{j},1}\left(w_{4\mathrm{j}-3}^+ \cdot x_{4j-3}\right) + \phi_{\mathrm{j},2}\left(w_{4\mathrm{j}-2}^+ \cdot x_{4j-2}\right)\right.\right. \\ &\quad \left.\left. - \phi_{j,3}\left(w_{4j-1}^- \cdot x_{4j-1}\right) - \phi_{j,4}\left(w_{4j}^- \cdot x_{4j}\right)\right]\right) + n/8 \end{aligned} \tag{5}$$

where $n^* = \frac{n}{4}$. The accumulation operation can be realized by on-chip combiners or microring-based multiplexers, which have been widely used in previous PTC works [4], [19].

The schematic of the testing setup is illustrated in Figure 4. Continuous-wave (CW) light is coupled to the chip

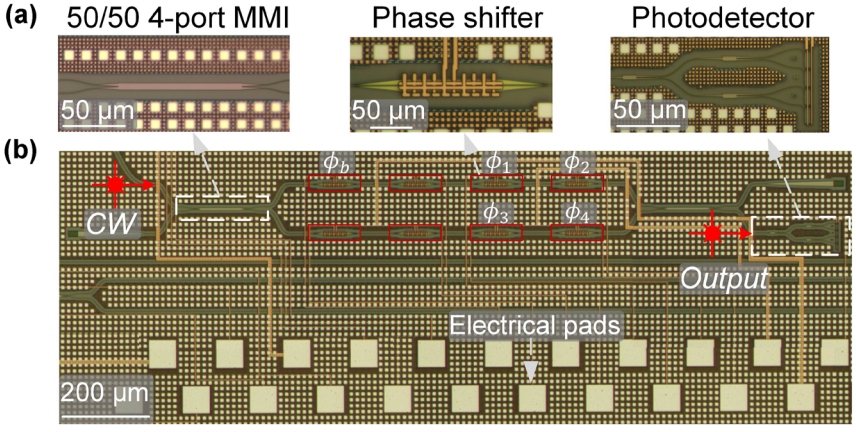

Figure 3: Schematic of the 4-operand MOMZI. The micrographs of necessary optical components are highlighted in (a) and the full schematic of the MOON is shown in (b). The phase shifters we use for training and biasing in this work are marked.

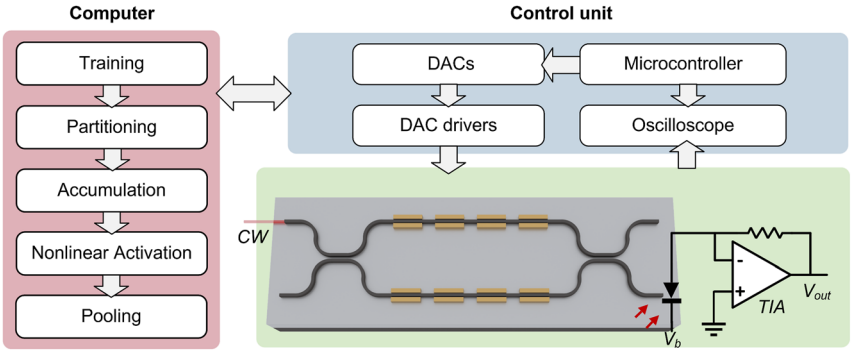

Figure 4: Experimental setup of MOMZI-ONN. Schematic of our MOMZI-ONN test flow. The entire tensor operation is first partitioned into multiple 4×1 blocks, and each block is implemented optically on a 4-op MOMZI. The weight parameters and the input signals are programmed by a multi-channel digital-to-analog converter (DAC). Then the output optical signals are converted to photocurrents using on-chip photodetectors. We use an off-chip TIA to convert the output photocurrent to electrical signals, which are then read by the oscilloscope. Both the oscilloscope and the DAC are controlled by a microcontroller. The tensor operation results are provided to the computer for data processing in order to train and deploy the DNN.

through an edge coupler. The MOMZI's phase shifters are programmed using a high-precision multi-channel digital-to-analog converter (DAC). The on-chip photodetector, along with an off-chip trans-impedance amplifier (TIA) converts the output optical signal to electrical voltage outputs. These converted electrical outputs will subsequently be read using oscilloscopes. A microcontroller is used to program the electrical signals that represent $w_i \cdot x_i$ to the DAC and read the output signals in this work. We use computers to process the measurement data, train the DNN parameters, and implement the DNN model. This work uses the microcontroller to program the DAC to emulate the dot-product operation directly. In real applications, high-speed DACs with programmable swings implement the multiplication as shown in Figure 1(f) and (g). High-speed (up to 224 Gbps) DACs with tunable gain have been demonstrated and are available in industry [23], [24]. Meanwhile, the energy efficiency of tunable DACs can be further improved with programmable memristors or phase-change devices. In addition, current fabrication and co-packaging technologies enable the integration of electrical control circuits and the laser on a single substrate [25] or a single chip [26], resulting in much higher compactness, shorter interconnect paths, and higher efficiency.

In this work, we construct a CNN with our MOMZI and benchmark its performance on a street view house number (SVHN) dataset. It is more complicated than the MNIST dataset [18] since each image contains color information and various natural backgrounds. To perform convolutional operations with our PTCs, we employ the widely-used tensor unrolling method (im2col) [27]. Large-size tensor operations are partitioned into 4×1 blocks and mapped onto our MOMZI. We first calibrate the behavior of each phase shifter for training and model it using Eq. (4), as shown in Figure 5(a). Based on the chip measurement data, our proposed hardware-aware training framework can efficiently train the ONN weights while being fully aware of all the physical non-idealities during optimization, e.g., process variations, thermal crosstalk, and signal quantization [14]. The dynamic noises are also measured (shown in Figure 5(b)) and added to the training framework to improve the robustness of ONNs (see Supplementary Materials Note 1). Additional power monitors can be added to the output port or the drop port of the MOMZI to realize *in-situ* training [28]–[30], which can continuously monitor the MOMZI's performance and update our training framework to improve the training accuracy. *In-situ* training can also potentially improve the training speed of MOMZI-PTCs by training multiple MOMZIs in parallel. After modeling the chip's actual response, we map the trained weights to our MOMZI to implement tensor operations. Figure 5(c) shows some normalized measured output results. Finally, we evaluate the task performance of our photonic neural chip on different ML tasks, where partial accumulation, nonlinearity, and other post-processing operations are offloaded to the digital computer. Figure 5(d) illustrates the network structure for training our MOMZI-ONN as well as the flow to implement im2col method with MOMZIs.

Our experiments show that under 4-bit voltage control resolution (16 phase shift levels for each operand), the inference accuracy of the CNN reaches ~85.89 % in our experimental demonstration. The confusion matrix depicting the prediction results is shown in Figure 5(e). Figure 5(f) and (g) shows the tested probability distribution of different street-view numbers. As a reference, we can achieve 91.8 % accuracy using an ideal CNN model with the same network structure on 64-bit computers. One can improve the task

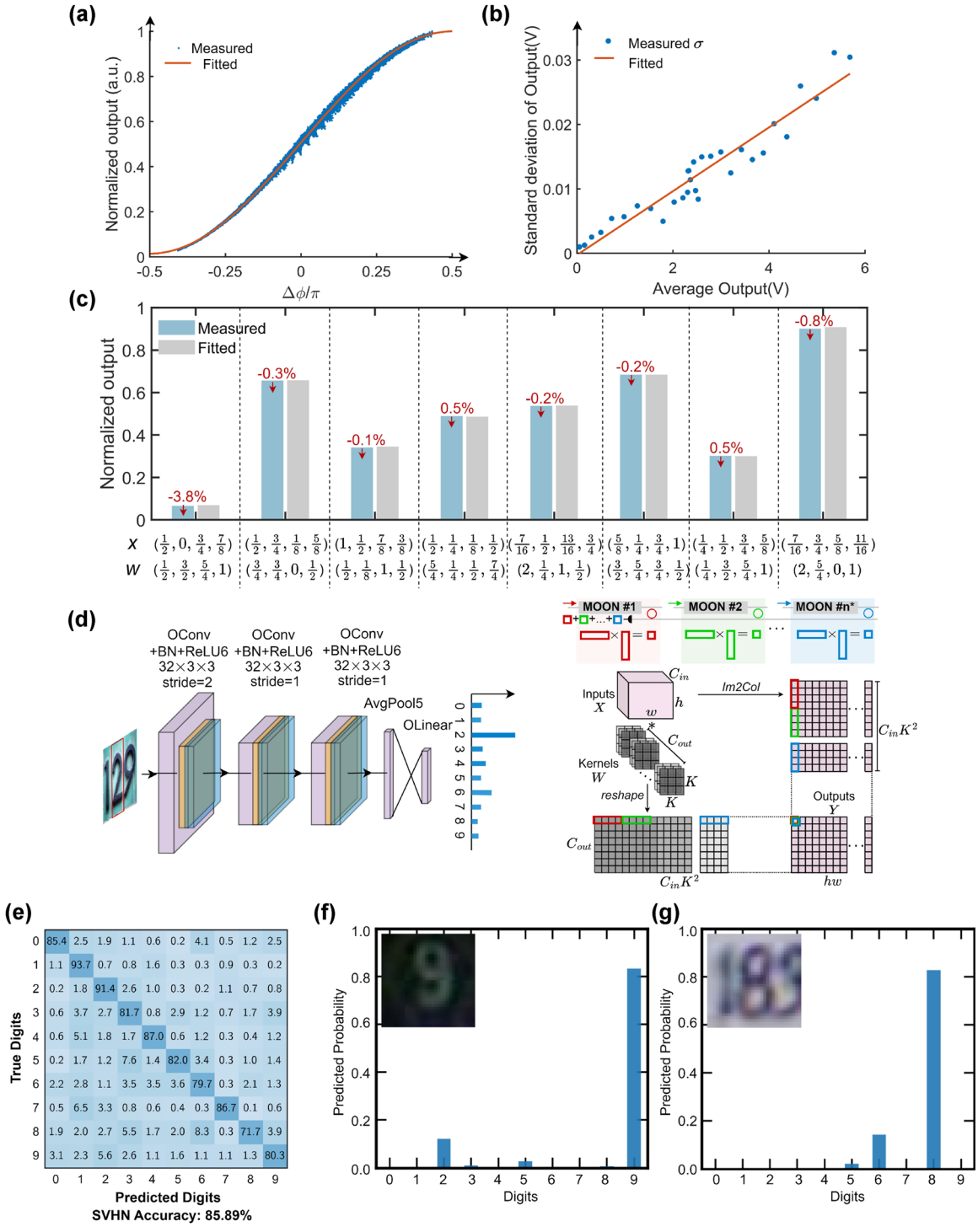

Figure 5: Experimental result of street view house number (SVHN) recognition with the MOMZI-ONN. (a) Our measured output data and curve fitting for training the MOMZI. The tuning range of the total phase shift of four operands is $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. (b) Dynamic noise analysis of output signal of MOMZI, the measured standard deviation of the dynamic noise is ∼0.5 %. (c) Comparison between experimentally measured output and fitted output. The deviation is marked in red. (d) Structure of the CNN. The first convolutional layer has three input channels and 32 output channels with a stride of 2. The subsequent two convolutional layers have 32 input/output channels with a stride of 1. After adaptive average pooling, we use a linear classifier with 10 outputs for final recognition. The convolution is realized by MOMZIs with im2col approach (shown on the right). When convolution is mapped to a matrix multiplication, each length-k vector dot-product is mapped to one k-op MOMZI. (e) Our measured output data and curve fitting for training the MOMZI-ONN. The tuning range of the total phase shift of four operands is $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. (e) The confusion matrix of the trained MOMZI-ONN on the SVHN dataset shows a measured accuracy of 85.89 %. (f) and (g) Show the predicted probability distribution of our MOMZI-ONN on two selected test digits in the SVHN dataset.

performance of MOMZIs using operands with more linear phase responses and higher control precision, which will be shown hereinafter.

5 Discussion

5.1 Expressivity evaluation

Our MOMZI-ONN exhibits comparable trainability and expressivity with ONNs designed for GEMMs with k times fewer optical component (MZI) usage (k is the number of operands). By explicitly modeling the transfer function of the MOMZI during ONN training, we can efficiently learn the mapping from the software model to the MZI devices. Here, we simulate the task performance of our MOMZI-ONN with different numbers of operands on the SVHN dataset using the same NN model and control precision. An ideal CNN model with the same model architecture is also trained as a reference. In the evaluation, the phase response of each operand is $\Delta\phi = \gamma\Delta V$ (γ is the modulation coefficient), which can be realized on linear phase shifters such as lithium niobate EO phase shifters [31]. In simulations, we add a phase bias $\phi_b = \frac{\pi}{2}$ to enable a balanced output range. The evaluation results are shown in Figure 6, showing that our MOMZI-ONNs can achieve >91 % accuracy on the SVHN dataset, which has <0.6 % accuracy difference compared to the ideal CNN model. It should be noted that the task performance of MOMZI-ONN is insensitive to the number of operands once we properly normalize the operands. Moreover, the number of active MZI devices to implement an n-input, m-output linear layer are $\frac{mn}{k}$ k-operand MOMZIs. Therefore, ONNs based on MOMZIs with a large number of operands will significantly reduce the hardware cost without accuracy loss.

5.2 Propagation delay and loss

By minimizing the number of cascaded MZI devices in the critical path of PTCs, MOMZI-PTC outperforms single-operand MZI-PTC in both propagation loss and optical delay by one to two orders of magnitude. In this work, we evaluate the propagation delay and loss of MOMZI-PTC using the foundry's process-design-kit (PDK) libraries. The parameters of optical devices are given in Table S1 (see Supplementary Materials Note 2). As shown in Figure 1(d), the MOMZIs in one optical path are placed parallelly in our PTC, so the insertion loss and propagation loss contributed by lossy MZIs will not accumulate when the size of the DNN model increases. As a result, the optical delay and the propagation loss of a MOMZI-PTC with n-inputs and m- outputs can be calculated as follows:

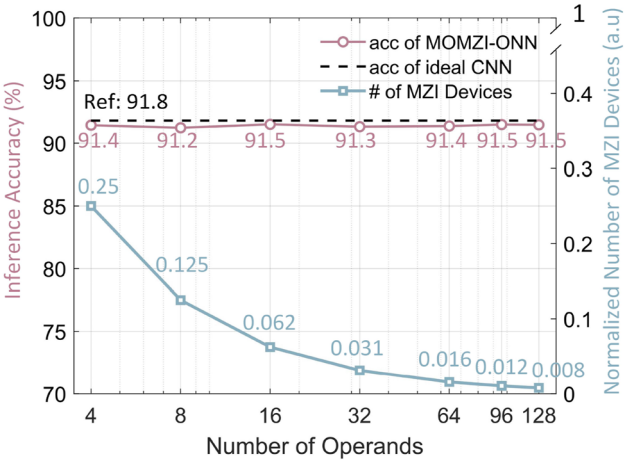

Figure 6: Task performance and hardware cost of MOMZI-ONN on SVHN dataset. Inference accuracies of MOMZI-ONNs with different operand numbers are shown. Using the same neural network structure, the accuracy of an ideal CNN model is 91.8 %. The normalized total number of MZI devices with different operand numbers of MOMZI-ONNs compared to ideal CNN models is shown. Suppose the matrix size is $n \times n$, and the number of microring- and MZI-PTCs is normalized to 1.

$$\tau_{\mathrm{MOMZI-PTC}} = \frac{n_g}{c}\left(L_{\mathrm{MOMZI}} + L_{\mathrm{combiner}}\right) \tag{6}$$

$$IL_{\mathrm{MOMZI-PTC}} = IL_{\mathrm{MOMZI}} + IL_{\mathrm{combiner}} \tag{7}$$

In Eq. (6), $n_g = 4.3$ is the group index of silicon waveguides. L_{MOMZI} is the length of the MOMZI, which depends on the operands and the waveguides used to connect these operands. Since the tuning ranges of a MOMZI and a single-operand MZI are the same, the total length of the operands of MOMZI should also be the same as the length of a high-speed electro-optic (EO) modulator. Here we assume the distance between each operand to be $d = 10$ µm based on the device layout of a recently-published two-operand 10-µm-radius microring modulator [32]. L_{combiner} is the length of the on-chip combiners/multiplexers, for microring-filter-based multiplexers, $L_{\mathrm{combiner}} = \frac{n}{k}L_{\mathrm{ring}}$. IL_{MOON} is the insertion loss of one multi-operand modulator, IL_{combiner} is the total insertion loss of the combiner, which is $\frac{n}{k}IL_{\mathrm{ring}}$ with add-drop microrings as multiplexers. Increasing the operand number k can potentially reduce both the IL and propagation delay.

On the other hand, the propagation loss of single-operand MZI-PTC can be estimated as $(n+m+1)IL_{\mathrm{MZI(ls)}} + IL_{\mathrm{MZI(hs)}}$, while the total device length can be expressed as $(n+m+1)L_{\mathrm{MZI(ls)}} + L_{\mathrm{MZI(hs)}}$. Here, MZI(hs) denotes the high-speed EO modulators for input signal encoding, and MZI(ls) is the TO switch for weight encoding. Because the tuning range and the modulation mechanism of the MOMZI should be the same as that of the input EO modulator, we let

$IL_{\mathrm{MOMZI}} = IL_{\mathrm{MZI(hs)}}$. The model parameters are available in Table S1.

The results presented in Figure 7(a) and (b) demonstrate that using the same component library [17], our MOMZI-PTC can achieve an optical delay that is approximately 49 times lower than that of a single-operand MZI-PTC. Furthermore, the propagation loss of our MOMZI-PTC is ∼257 dB lower than that of the single-operand MZI-PTC, which results in lower laser power requirements to drive the ONN and a lower response time.

5.3 Computational speed

MOMZI-PTC outperforms single-operand MZI-PTC in computational speed by minimizing optical propagation delay in the critical path. The total delay of MOMZI-PTC is determined by factors such as the response time of tunable DACs, EO response time, optical propagation delay, photodetection time, and other electrical processing circuits. With the availability of high-speed (224 GBaud) DACs featuring tunable gains [24], and considering that other electrical components remain the same between MOMZI-PTCs and single-operand MZI-PTCs, MOMZI-PTC's total delay is significantly lower than that of single-operand MZI-PTC, thanks to a substantial reduction in optical propagation delay. Moreover, because the optical propagation loss of MOMZI-PTC is significantly lower than single-operand MZI-PTCs, it is possible to reduce the driving ability and optimize the photodetection circuits for higher bandwidth, potentially enhancing the computational speed even further.

Like single-operand MZI-PTCs, MOMZI-PTC exhibits a relatively lower weight programming speed than computational speed, which is primarily constrained by the programming speed of the gain control circuit or the response time of memristors. However, it's feasible to tailor the transfer function of MOONs and modify signal encoding techniques for each operand to facilitate high-speed programming of both weights and signals, which will be explored further in the later discussions.

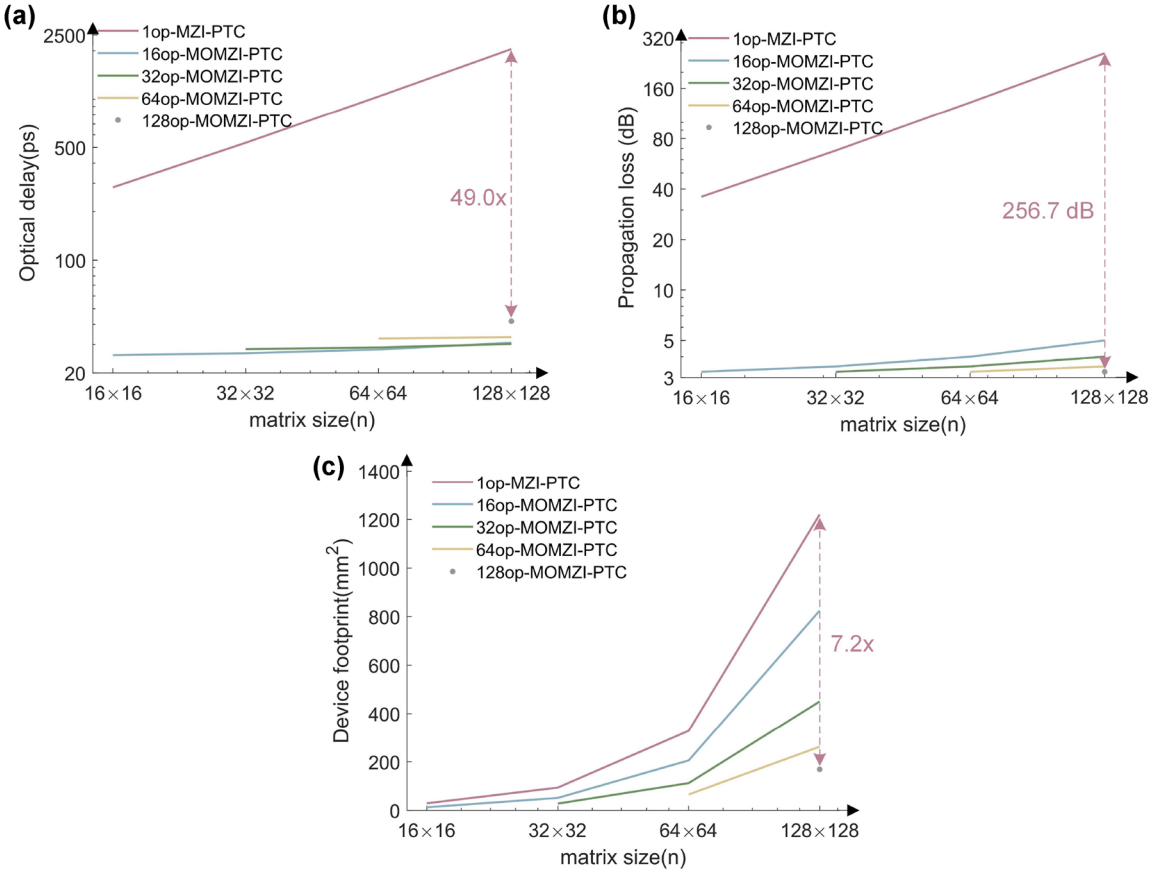

Figure 7: Performance analysis of MOMZI-PTC and comparison with single-operand (1op) MZI-PTC [8] using foundry PDKs [17]. MOMZIs with different operand numbers are shown. Here we suppose the circuit structure of the MZI-based PTC is Clement-style [33]. (a) Optical propagation delay in log scale. (b) Optical propagation loss. (c) Device footprint.

5.4 Footprint

Our MOMZI-PTC significantly improves the area efficiency and reduces the number of MZI devices compared to a single-operand MZI-PTC [8]. Unlike single-operand active devices such as single-operand MZI, our k-op MOMZI is capable of implementing length-k vector-vector inner products, which results in a much higher hardware efficiency in terms of #MAC/MZI. The total device footprint of k-op MOMZI-PTCs can be estimated using Eq. (8):

$$S_{\text{MOMZI}-\text{PTC}} = \frac{m \times n}{k} S_{\text{MOMZI}} + S_{\text{combiner}} \tag{8}$$

where we assume a distance of $d = 10\,\mu\text{m}$ between neighboring operands. Suppose the device footprint of a high-speed MZI modulator is $S_{\text{MZI(hs)}} = L_{\text{MZI(hs)}} W_{\text{MZI(hs)}}$. The device footprint of one k-op MOMZI can then be estimated as $S_{\text{MOMZI}} = \left(L_{\text{MZI(hs)}} + (k-1)d\right) \cdot W_{\text{MZI(hs)}}$. Figure 7(c) shows the estimated device footprint of MOMZI-based PTC and single-operand MZI-PTC based on our assumptions. The estimated device footprint of MOMZI-PTC and MZI-PTC is shown in Figure 7(c). When the matrix size is 128×128, our 128-op MOMZI-PTC consumes ∼127× fewer MZI modulators, leading to ∼6.2× footprint reduction compared to single-operand MZI-PTC [8] with the same matrix size and optical component selection.

From Eq. (8) and Figure 7(c), MOMZI-based PTC will be more area efficient with a larger number of operands k on each MOMZI. The foundry's fabrication process precision, which determines the shortest operand one can design, restricts the maximum number of operands. Moreover, the area for metal routing and placement of electrical tunable DACs also limits the size of each operand. Previous has shown that a 10-µm-radius silicon-based microring modulator can be divided into 32 independent active segments using a 45-nm technology node [34], where each operand only consumes 2 µm in length. This means that an MZI-modulator with a 1.6 mm-length modulation arm has the potential to support up to 800 operands using current layout technology, which should be comparable with other analog electronic tensor cores in scalability, e.g., 256×256 memristor-based crossbar arrays [35].

Another big advantage of the proposed MOON-based PTC is its superior compatibility with compact, high-speed optical modulators, even with high insertion loss, e.g., plasmonic-on-silicon modulators, which have only 15 µm modulation length and 11.2 dB IL [5]. The fundamental reason is the small number of cascaded devices in the critical path. Figure 8 shows the normalized device footprint compared with silicon-based MZI-PTCs, which shows the plasmonic-on-silicon-MOMZI-PTC can reduce the footprint

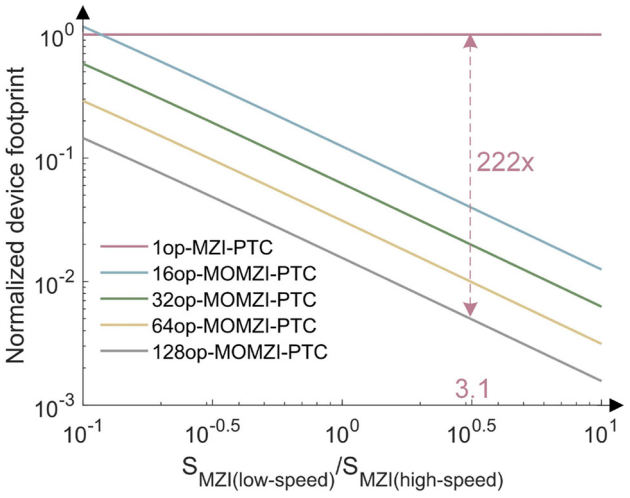

Figure 8: Normalized footprint of MOMZI-based PTC using scaling technologies. The x-axis is the ratio between the area of low-speed silicon-based TO MZI ($550 \times 125\,\mu\text{m}^2$) and high-speed MZI. Using compact plasmonic-on-silicon high-speed modulators ($\sim 220 \times 100\,\mu\text{m}^2$) [21], $S_{\text{MZI(low-speed)}}/S_{\text{MZI(high-speed)}} \cong 3.12$, and a 128-op MOMZI-PTC consumes a 222× smaller footprint than single-operand MZI-PTCs using silicon-based MZI modulators. For simplicity, we assume the entire waveguide length for connecting the operands is the same as the total length of the operands of MOMZIs, so $S_{\text{MOMZI}} = 2S_{\text{MZI(high-speed)}}$.

by 177× compared to single-operand silicon-MZI-based PTC. Single-operand MZI-PTCs are not compatible with these compact high-loss modulators because there are $2n + 1$ MZIs in the critical path. Using compact high-loss modulators for weight configuration will lead to significant propagation loss and require high laser power to drive the neural chip.

Finally, the hardware cost of MOMZI-PTC can be further optimized with operand pruning strategies. To implement an FC layer in DNN models, especially with sparse matrices [38], we only need to encode non-zero weights on the MOMZIs. The operands of MOMZIs with zero weight values can be either removed from the device to save footprint or power-gated to reduce energy consumption. Sparsity-aware training [39] can be applied to prune redundant MOMZI operands while maintaining task accuracy.

5.5 Energy efficiency

MOMZI-based PTC is a more energy-efficient alternative to single-operand MZI-PTCs for implementing large-tensor-size operations due to its lower propagation loss, which allows it to consume over 256 dB less laser power. The total power consumption of MOMZI-PTC for computing comprises the power required to drive the lasers, modulators, and photodetectors and for biasing the MOMZI, as well as the power needed to drive the digital-to-analog converters

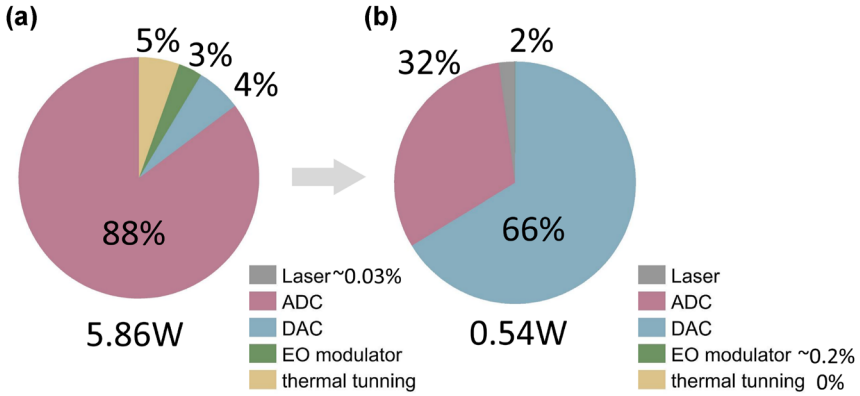

Figure 9: Power breakdown of a 128 × 128 photonic tensor core implemented by 128 128-op MOMZIs using existing technology (a) and emerging technology (b). (a) The total power of the MOMZI-ONN is 5.7 W at 10 GHz clock rate (56 TOPS/W). (b) Using emerging technologies, we use ADC-less designs (e.g., magnetic-tunnel-junction (MTJ)-based analog content-addressable memory (ACAM) [36], [37]) to boost the energy efficiency to ∼604 TOPS/W.

(DACs) and analog-to-digital converters (ADCs). The silicon-based carrier-depletion MZI's modulation energy consumption in previous work can achieve ∼146 fJ/bit [40]. Furthermore, the power to bias the MOMZI is ∼2.5 mW per phase shifter if we use thermal phase shifters from foundry PDKs [41].

Using the parameters of existing technology provided in Table S3 (see Supplementary Materials Note 3), the optical part of MOMZI-PTC, accounts for <9 % of total power consumption when the tensor size is 128 × 128. The power breakdown analysis shown in Figure 9(a) indicates that our 128-op MOMZI-PTC can achieve ∼56 TOPS/W at a 10 GHz clock rate, 100 % higher than existing analog electronic tensor cores [42] with 100× faster operating speed. Currently, the energy efficiency of MOMZI-PTC is dominated by data converters such as ADCs. This work employs an 8-bit, 10 GSPS ADC that consumes 39 mW per channel [43].

The energy efficiency can be further improved to ∼604 TOPS/W using emerging high-speed and energy-efficient data converters and EO modulators. Recent advances in energy-efficient active optical components, such as the plasmonic-on-silicon modulator that consumes approximately 0.1 fJ per bit modulation energy at 50 GHz operating frequency, have made it possible to reduce the power consumption of MOMZI further [21]. The power to bias the MOMZI can be decreased to zero with phase change materials or nano-opto-electro-mechanical devices [44], [45]. Using energy-efficient modulators, the energy consumption of the optical computing part only accounts for <3 % of the total power consumption, showing that large-size MOMZI-PTC will not bring scalability issues due to excessive laser power. Moreover, we can use energy-efficient analog content-addressable memory (ACM) to replace the ADCs [36], reducing the power consumption of ADCs by ∼33×. As shown in Figure 9(b), the final power breakdown of MOMZI-PTC for computing shows our MOMZI-PTC can achieve a competitive energy efficiency of ∼604 TOPS/W, 20× higher than existing memristor-based analog electronic tensor cores [42]. More details of our power analysis are provided in Supplementary Materials Note 3.

In addition, our k-op MOMZI-PTC can reduce the weight reconfiguration energy by k times compared to single-operand-device-based PTCs, which will bring considerable energy efficiency improvement, especially when the photonic tensor cores need to be frequently reconfigured to map a large number of matrix blocks in DNNs. The number of MZI devices in our MOMZI-PTC is only $O\left(\frac{mn}{k}\right)$, which is k times fewer than that of PTCs with single-operand devices ($\mathcal{O}(mn)$ [19] or $\mathcal{O}\left(\max\left(m^2, n^2\right)\right)$ [46]) This feature of MOMZI-PTC is essential to implement modern DNNs, where weight loading takes nontrivial hardware costs [22].

5.6 Nonlinearity engineering

The nonlinearity of MOONs can be customized in various dimensions to achieve a desired activation function, potentially saving power for doing activation functions electronically. The built-in nonlinearity of MOON is contributed by the weight/signal encoding way and the nonlinear transfer function of the optical modulator with the input voltage. To customize such built-in nonlinearity, one can add electrical or optical components before or after photodetection to alter the optical outputs to implement the activation function. Previous work has widely investigated this approach [47]–[49]. Typically, one can add saturable absorbers before photodetection with a linear optical modulator [50] to

construct a ReLU-like MOON, reducing the hardware cost to realize activation functions electronically.

Depending on the transfer function of the MOON, the weight encoding approach can be designed to enable high-speed dynamic tensor operations beyond ones with stationary weights. Dynamic tensor operations mean both the inputs and the weights can be updated at high speed, which is crucial in emerging applications, such as the self-attention operation in transformer [51] and on-chip training tasks for intelligent edge learning. A specific example of an optical modulator with a linear field response region with voltage ($|\Delta E_{\text{out}}| \propto \Delta V$) is provided here. Suppose the electrical modulation signal of the modulator is bidirectional; then, one can use two MOONs and one differential photodetector to implement high-speed vector-to-vector operations. As shown in Figure 10, the weight and input voltage signal w_i and x_i are encoded with the same phases on operand i of the upper modulator, and high-speed signals w_i and x_i with opposite phases encoded on operand i of the downer modulator. After differential photodetection, one can obtain the output current signal as:

$$\begin{aligned} I_- &= I_0 + \alpha\left(\Sigma\left(w_i - x_i\right)^2\right) \\ I_+ &= I_0 + \alpha\left(\Sigma\left(w_i + x_i\right)^2\right) \\ I_{\text{out}} &= I_+ - I_- = 2\alpha\Sigma\left(w_i \cdot x_i\right) \end{aligned} \tag{9}$$

where α is the modulation efficiency of each operand. I_0 is the output intensity of the modulator at the biased point. Compared to MOONs that use memristors to encode stationary weights, the dual-linear-modulator-based MOON shown in Figure 10 can enable high-speed weight reprogramming/updates to implement high-speed dynamic tensor operations. One can investigate more efficient signal encoding approaches of MOONs to support more types of tensor operations in state-of-the-art DNNs.

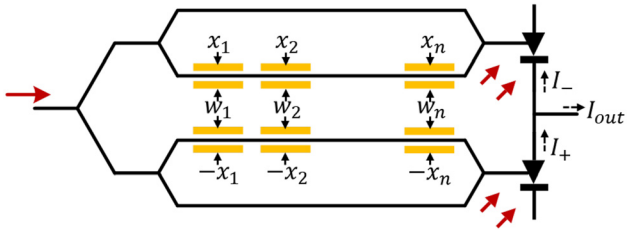

Figure 10: A MOON with two linear modulators for matrix-matrix multiplications. The optical output power of each modulator is proportional to the electrical input power or V^2. Here we apply differential input signals $\pm x_i s$ on upper/lower modulators, and put weight signals $w_i s$ on both upper/lower modulators. The output power after differential photodetection is then proportional to $\sum_1^n w_i \cdot x_i$.

6 Conclusions

We have presented a scalable, energy-efficient optical neural network with customized multi-operand optical neurons (MOONs). We have experimentally demonstrated a 4-operand silicon-photonic MOMZI on practical image recognition tasks. Compared to prior single-operand-MZI-based photonic tensor cores (PTCs), our MOMZI-based PTC design can achieve one to two orders-of-magnitude reduction in MZI device usage, footprint, latency, and propagation loss. The speed, footprint, and energy efficiency of our MOON-based PTC can benefit from more advanced technologies, e.g., faster and more efficient data converters, optical devices, and nonlinearity engineering. Our customized MOON design provides a scalable solution for the next-generation photonic AI accelerators with extreme compute density and energy efficiency.

Acknowledgments: The authors acknowledge support from the Multidisciplinary University Research Initiative (MURI) program through the Air Force Office of Scientific Research (AFOSR), monitored by Dr. Gernot S. Pomrenke.
Research funding: Air Force Office of Scientific Research (AFOSR) (FA 9550-17-1-0071).
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Conflict of interest: Authors state no conflicts of interest.
Informed consent: Informed consent was obtained from all individuals included in this study.
Ethical approval: The conducted research is not related to either human or animals use.
Data availability: The data and codes that support the findings of this study are available from the corresponding author upon reasonable request.

References

[1] B. J. Shastri, *et al*., "Photonics for artificial intelligence and neuromorphic computing," *Nat. Photonics*, vol. 15, no. 2, pp. 102–114, 2020.
[2] Z. Ying, *et al*., "Electronic-photonic arithmetic logic unit for high-speed computing," *Nat. Commun.*, vol. 11, no. 1, p. 2154, 2020.
[3] H. Zhou, *et al*., "Photonic matrix multiplication lights up photonic accelerator and beyond," *Light Sci. Appl.*, vol. 11, no. 1, p. 30, 2022.
[4] M. Miscuglio and V. J. Sorger, "Photonic tensor cores for machine learning," *Appl. Phys. Rev.*, vol. 7, no. 3, p. 031404, 2020.
[5] J. Feldmann, *et al*., "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, no. 7840, pp. 52–58, 2021.
[6] M. A. Nahmias, T. F. de Lima, A. N. Tait, H. T. Peng, B. J. Shastri, and P. R. Prucnal, "Photonic multiply-accumulate operations for neural

networks," *IEEE J. Sel. Top. Quantum Electron.*, vol. 26, no. 1, pp. 1—18, 2020.
[7] M. A. Al-Qadasi, L. Chrostowski, B. J. Shastri, and S. Shekhar, "Scaling up silicon photonic-based accelerators: challenges and opportunities," *APL Photon.*, vol. 7, no. 2, p. 020902, 2022.
[8] Y. Shen, *et al*., "Deep learning with coherent nanophotonic circuits," *Nat. Photonics*, vol. 11, no. 7, pp. 441—446, 2017.
[9] S. Chen, Y. Shi, and D. Dai, "Low-loss and broadband 2 × 2 silicon thermo-optic Mach—Zehnder switch with bent directional couplers," *Opt. Lett.*, vol. 41, no. 4, pp. 836—839, 2016.
[10] C. Li, *et al*., "Analogue signal and image processing with large memristor crossbars," *Nat. Electron.*, vol. 1, no. 1, pp. 52—59, 2018.
[11] C. Feng, *et al*., "A compact butterfly-style silicon photonic—electronic neural chip for hardware-efficient deep learning," *ACS Photon.*, vol. 9, no. 12, pp. 3906—3916, 2022.
[12] J. Gu, *et al*., "ADEPT: automatic differentiable DEsign of photonic tensor cores," in *Proceedings of the 59th ACM/IEEE Design Automation Conference*, New York, NY, USA, Association for Computing Machinery, 2022, pp. 937—942.
[13] X. Xiao, 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 Photon.*, vol. 6, no. 12, p. 126107, 2021.
[14] H. H. Zhu, *et al*., "Space-efficient optical computing with an integrated chip diffractive neural network," *Nat. Commun.*, vol. 13, no. 1, p. 1044, 2022.
[15] Z. Wang, L. Chang, F. Wang, and T. Gu, "Integrated photonic metasystem for image classifications at telecommunication wavelength," *Nat. Commun.*, vol. 13, no. 1, p. 2131, 2022.
[16] J. Gu, *et al*., "SqueezeLight : towards scalable optical neural networks with multi-operand ring resonators," in *Proceedings -Design, Automation and Test in Europe*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 238—243.
[17] E. Timurdogan, *et al*., "APSUNY process design kit (PDKv3.0): O, C and L band silicon photonics component libraries on 300mm wafers," in *2019 Optical Fiber Communications Conference and Exhibition (OFC)*, OSA, 2019, pp. 1—3.
[18] Y. Lecun, "The MNIST database of handwritten digits," Available at: http://yann.lecun.com/exdb/mnist/.
[19] A. N. Tait, *et al*., "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, vol. 7, no. 1, p. 7430, 2017.
[20] H. Zhang, *et al*., "Miniature multilevel optical memristive switch using phase change material," *ACS Photonics*, vol. 6, no. 9, pp. 2205—2212, 2019.
[21] W. Heni, *et al*., "Plasmonic IQ modulators with attojoule per bit electrical energy consumption," *Nat. Commun.*, vol. 10, no. 1, p. 1694, 2019.
[22] G. L. Li, T. G. B. Mason, and P. K. L. Yu, "Analysis of segmented traveling-wave optical modulators," *J. Lightwave Technol.*, vol. 22, no. 7, p. 1789, 2004.
[23] H. Mardoyan, *et al*., "Single carrier 168-Gb/s line-rate PAM direct detection transmission using high-speed selector power DAC for optical interconnects," *J. Lightwave Technol.*, vol. 34, no. 7, pp. 1593—1598, 2016.
[24] A. Konczykowska, *et al*., "112 GBaud (224 Gb/s) large output swing InP DHBT PAM-4 DAC-driver," in *2022 24th International Microwave and Radar Conference (MIKON)*, 2022, pp. 1—4.
[25] C. Minkenberg, R. Krishnaswamy, A. Zilkie, and D. Nelson, "Co-packaged datacenter optics: opportunities and challenges," *IET Optoelectron.*, vol. 15, no. 2, pp. 77—91, 2021.
[26] A. H. Atabaki, *et al*., "Integrating photonics with silicon nanoelectronics for the next generation of systems on a chip," *Nature*, vol. 556, no. 7701, pp. 349—353, 2018.
[27] S. Chetlur, *et al*., "cuDNN: efficient primitives for deep learning," *arXiv preprint*, vol. arXiv:1410.0759, 2014.
[28] T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, "Training of photonic neural networks through in situ backpropagation and gradient measurement," *Optica*, vol. 5, no. 7, p. 864, 2018.
[29] J. Gu, *et al*., "L2ight: enabling on-chip learning for optical neural networks via efficient in-situ subspace optimization," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 8649—8661.
[30] L. G. Wright, *et al*., "Deep physical neural networks trained with backpropagation," *Nature*, vol. 601, no. 7894, pp. 549—555, 2022.
[31] C. Wang, *et al*., "Integrated lithium niobate electro-optic modulators operating at CMOS-compatible voltages," *Nature*, vol. 562, no. 7725, pp. 101—104, 2018.
[32] J. Sun, R. Kumar, M. Sakib, J. B. Driscoll, H. Jayatilleka, and H. Rong, "A 128 Gb/s PAM4 silicon microring modulator with integrated thermo-optic resonance tuning," *J. Lightwave Technol.*, vol. 37, no. 1, pp. 110—115, 2019.
[33] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walsmley, "Optimal design for universal multiport interferometers," *Optica*, vol. 3, no. 12, pp. 1460—1465, 2016.
[34] S. Moazeni, *et al*., "A 40-Gb/s PAM-4 transmitter based on a ring-resonator optical DAC in 45-nm SOI CMOS," *IEEE J. Solid-State Circ.*, vol. 52, no. 12, pp. 3503—3516, 2017.
[35] J. Yang, *et al*., "Thousands of conductance levels in memristors monolithically integrated on CMOS," preprint, 2022, https://doi.org/10.21203/rs.3.rs-1939455/v1.
[36] C. Li, *et al*., "Analog content-addressable memories with memristors," *Nat. Commun.*, vol. 11, no. 1, p. 1638, 2020.
[37] H. Zhu, *et al*., "Fuse and mix: MACAM-enabled analog activation for energy-efficient neural acceleration," in *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*, New York, NY, USA, Association for Computing Machinery, 2022, pp. 1—9.
[38] W. Wen, *et al*., "Learning structured sparsity in deep neural networks," in *Advances in Neural Information Processing Systems*, vol. 29, 2016.
[39] J. Gu, *et al*., "Towards area-efficient optical neural networks: an FFT-based architecture," in *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*, IEEE, 2020, pp. 476—481.
[40] J. Ding, *et al*., "Ultra-low-power carrier-depletion Mach—Zehnder silicon optical modulator," *Opt. Express*, vol. 20, no. 7, pp. 7081—7087, 2012.
[41] S. Y. Siew, *et al*., "Review of silicon photonics technology and platform development," *J. Lightwave Technol.*, vol. 39, no. 13, pp. 4374—4389, 2021.
[42] S. Ambrogio, *et al*., "Equivalent-accuracy accelerated neural-network training using analogue memory," *Nature*, vol. 558, no. 7708, pp. 60—67, 2018.
[43] ADC (analog-to-digital converters) — alphacore," Available at: https://www.alphacoreinc.com/adc-analog-to-digital-converters/ Accessed: Aug. 25, 2021.

[44] M. Wuttig, H. Bhaskaran, and T. Taubner, "Phase-change materials for non-volatile photonic applications," *Nat. Photonics*, vol. 11, no. 8, pp. 465—476, 2017.
[45] L. Midolo, A. Schliesser, and A. Fiore, "Nano-opto-electro-mechanical systems," *Nat. Nanotechnol.*, vol. 13, no. 1, pp. 11—18, 2018.
[46] Y. Shen, *et al*., "Deep learning with coherent nanophotonic circuits," in *2017 IEEE Photonics Society Summer Topical Meeting Series (SUM)*, vol. 11, IEEE, 2017, pp. 189—190.
[47] I. Williamson, T. W. Hughes, M. Minkov, B. Bartlett, S. Pai, and S. Fan, "Reprogrammable electro-optic nonlinear activation functions for optical neural networks," *IEEE J. Sel. Top. Quantum Electron.*, vol. 26, no. 1, pp. 1—12, 2019.
[48] C. Huang, *et al*., "A silicon photonic—electronic neural network for fibre nonlinearity compensation," *Nat. Electron.*, vol. 4, no. 11, pp. 837—844, 2021.
[49] Z. Xu, *et al*., "Reconfigurable nonlinear photonic activation function for photonic neural network based on non-volatile opto-resistive RAM switch," *Light: Sci. Appl.*, vol. 11, no. 1, pp. 1—11, 2022.
[50] X. Zhang, B. Lee, C.-Y. Lin, A. X. Wang, A. Hosseini, and R. T. Chen, "Highly linear broadband optical modulator based on electro-optic polymer," *IEEE Photonics J.*, vol. 4, no. 6, pp. 2214—2228, 2012.
[51] A. Vaswani, *et al*., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/nanoph-2023-0554).