



Lightning-Transformer: A Dynamically-operated Optically-interconnected Photonic Transformer Accelerator

Hanqing Zhu¹, Jiaqi Gu^{1,3}, Hanrui Wang², Zixuan Jiang¹,
Zhekai Zhang², Rongxing Tang¹, Chenghao Feng¹,
Song Han², Ray T. Chen¹, David Z. Pan¹

¹The University of Texas at Austin, ²Massachusetts Institute of Technology, ³Arizona State University
{hqzhu, jqgu, zixuan, rt970117, fengchenghao1996}@utexas.edu, {hanrui, zhangzk, songhan}@mit.edu,
{chen, dpan}@ece.utexas.edu

Abstract— The wide adoption and significant computing resource cost of attention-based transformers, e.g., Vision Transformers and large language models, have driven the demand for efficient hardware accelerators. While electronic accelerators have been commonly used, there is a growing interest in exploring photonics as an alternative technology due to its high energy efficiency and ultra-fast processing speed. Photonic accelerators have demonstrated promising results for convolutional neural networks (CNNs) workloads, which predominantly rely on weight-static linear operations. However, they encounter challenges when it comes to efficiently supporting attention-based Transformer architectures, raising questions about the applicability of photonics to advanced machine-learning tasks. The primary hurdle lies in their inefficiency in handling the unique workloads inherent to Transformers, i.e., dynamic and full-range tensor multiplication.

In this work, we propose **Lightning-Transformer**, the *first* light-empowered, high-performance, and energy-efficient photonic Transformer accelerator. To overcome the fundamental limitation of existing photonic tensor core designs, we introduce a novel dynamically-operated photonic tensor core, **DPTC**, consisting of a crossbar array of interference-based optical vector dot-product engines, supporting highly parallel, dynamic, and full-range matrix multiplication. Furthermore, we design a dedicated accelerator that integrates our novel photonic computing cores with photonic interconnects for inter-core data broadcast, fully unleashing the power of optics. The comprehensive evaluation demonstrates that **Lightning-Transformer** achieves $>2.6\times$ energy and $>12\times$ latency reductions compared to prior photonic accelerators and delivers the lowest energy cost and 2 to 3 orders of magnitude lower energy-delay product compared to the electronic Transformer accelerator, all while maintaining digital-comparable accuracy. Our work highlights the immense potential of photonics for efficient hardware accelerators, particularly for advanced machine-learning workloads, such as Transformer-backed large language models (LLM). Our implementation is available at <https://github.com/zhuhanqing/Lightning-Transformer>.

I. INTRODUCTION

Recently, attention-based Transformers have gained immense popularity and demonstrated remarkable success in various domains, e.g., natural language processing (NLP) [5], [25], [38], [60] and computer vision (CV) [8], [14], [54]. The attention mechanism enables dynamic feature aggregation, long-distance modeling, and global context extraction, contributing significantly to the impressive performance [14], [55]. However,

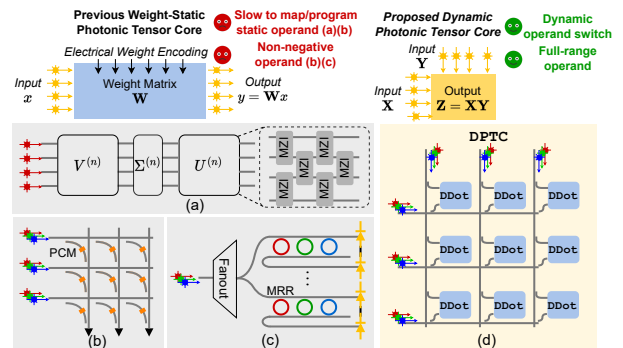


Fig. 1: (a), (b), (c) Prior weight-static photonic tensor core designs [16], [47], [52]. (d) Our proposed dynamic photonic tensor core design without static weight constraints.

this exceptional performance comes at a considerable computational cost. The quadratic complexity of attention, in terms of computation and memory, combined with a large number of parameters, demands substantial computational resources. This poses a challenge for deploying Transformers, particularly in resource-constrained systems where such computational demands are prohibitive. Hence, there is a pressing need to develop domain-specific hardware accelerators for the efficient deployment of Transformers in real-world applications.

Several hardware accelerators based on digital electronics have been proposed to accelerate the inference of Transformers [12], [50], [57], [64], [69]. However, traditional electrical digital computing platforms face significant challenges as transistor-based chips reach the limits of Dennard scaling, leading to increased power dissipation per unit area and diminishing performance improvements. As a compelling alternative, integrated photonic accelerators [18], [46] have emerged as next-generation computation platforms offering ultra-high speed, high parallelism, and low energy consumption. Various optical systems are actively being studied for accelerating convolutional neural network (CNN) workloads with different photonic tensor core (PTC) designs, e.g., Mach-Zehnder interferometer (MZI) array [47], Micro-ring Resonator (MRR) bank [51], [52], and

non-volatile Phase Change Material (PCM)-based crossbar [16].

However, existing photonic accelerators are mainly designed and optimized for weight-static NNs, e.g., CNNs, where convolution and fully-connected layers only involve matrix multiplications (MM) between the learned static weight matrix and (non-negative) input tensors. They fail to efficiently support attention-based Transformers due to the following *challenges*:

Matrix multiplication with two dynamic input operands.

Unlike digital electronics, prior photonic accelerators typically need to map one operand of MM (usually the weight matrix \mathbf{W}) onto PTC’s circuit transmission by programming its devices, shown in Figure 1. This procedure typically involves costly operand mapping and slow device programming, leading to a preference for keeping the encoded operand *static* and reusing it for many inputs to amortize the cost, i.e., “weight-static”. For example, MZI array [47] requires singular value decomposition (SVD) and phase decomposition for operand mapping. To map a 12×12 matrix, it takes ~ 1.5 ms for SVD and phase decomposition on CPU. Moreover, facing the challenges of bulky area and large insertion loss, PTC designs prefer to use low-loss, compact, and non-volatile devices that usually cost 10 ns-10 μ s to be programmed [3], [16], [42]. However, attention in Transformers is built on dynamic matrix multiplication, with both operands being dynamically generated activations. This dynamic nature necessitates frequent real-time operand mapping and device reprogramming. Real-time mapping and reprogramming are generally not affordable or challenging to amortize given the orders-of-magnitude higher runtime than ultra-fast computing (< 100 ps) and limited reuse opportunity in the dynamic MM scenario, incurring long latency for preparing PTCs and leading to severe system stall.

Matrix multiplication with full-range input operands.

Transformers require full-range matrix multiplications as activations are not constrained to non-negative only. However, prior incoherent PTCs, such as MRR bank, pose range limitations on operands. At least one of the operands is limited to be non-negative as their computation is based on light intensity modulation (non-negative only). The absence of full range support often requires decomposing full-range operands into differences of non-negative operands $(X_+ - X_-)(Y_+ - Y_-)$ and processing X_+Y_+ , X_+Y_- , X_-Y_+ , and X_-Y_- separately [3], [51] with extra accumulation steps, incurring $> 2-4 \times$ hardware cost compared to one with full-range operand support. Full-range operand encoding is a unique feature of coherent PTCs where signs can be encoded to phases and processed via interference. Coherent MZI array indeed can achieve full-range MVM with a single wavelength. In comparison, our PTC design maintains the full-range feature with a novel interference circuit design while utilizing multiple wavelengths for ultra-parallel MM.

Above all, existing photonic accelerators encounter significant difficulties in efficiently accelerating Transformers’ unique dynamic and full-range MM workloads. To address those challenges, in this work, we propose the *first* customized photonic accelerator to support attention-based Transformers, named *Lightening-Transformer*. We first design a *coherent* dot-product unit *DDot* that enables multiplication between two

dynamically encoded full-range optical vectors. *DDot* directly encodes both operands as high-speed *coherent* optical signals, thus can represent signs as signal phases and support real-time operand switching with negligible mapping or programming cost (< 10 ps). The dot-product mechanism is based on coherent light interference with a coupler and balanced photodetection, further enabling to detect full-range outputs as positive or negative photocurrent. Based on *DDot*, we devise a novel crossbar-style photonic tensor core, *DPTC*, for ultra-parallel and energy-efficient dynamic full-range matrix multiplication, which is the key building block of our Transformer accelerator. We fully leverage both spectral and spatial parallelism of optics by exploiting the wavelength-division multiplexing (WDM) capacity, and unleash the natural optical broadcast capability to enable intra-core and inter-core operand sharing. As a result, we can maximize the hardware sharing, trim modulation overhead, and significantly boost processing parallelism and efficiency.

The major contributions of this paper are as follows:

- We present, *for the first time*, a light-empowered, high-performance, and energy-efficient photonic Transformer accelerator, dubbed *Lightening-Transformer*, overcoming the efficiency and flexibility limitations of prior photonic accelerators for Transformer acceleration.
- We design a novel dynamically-operated photonic tensor core, *DPTC*, that enables ultra-parallel and energy-efficient dynamic full-range general matrix multiplication in a one-shot way. *DPTC* utilizes the WDM technique to enable spectral parallelism and a crossbar-based circuit topology to explore intra-core operand sharing, offering exceptional processing parallelism and energy efficiency.
- We develop a dedicated dynamically-operated optically-interconnected accelerator that utilizes our superior photonic tensor cores for efficient computing and optical interconnects for efficient inter-core data broadcast to fully unleash the power of optics. To further reduce the signal conversion cost, we employ architecture-level optimization to reduce input electrical-to-optical (E-O) conversion cost via sharing optical signals inter-core and reduce output optical-to-electrical (O-E) conversion cost via analog-domain temporal accumulation.
- We evaluate proposed *Lightening-Transformer* comprehensively compared to photonic and electronic accelerators across different Transformer benchmarks. *Lightening-Transformer* significantly outperforms prior photonic designs, achieving over $2.6 \times$ energy and over $12 \times$ latency reductions. Furthermore, compared to state-of-the-art electronic Transformer accelerators, our accelerator consistently delivers the lowest energy consumption and achieves 2 to 3 orders of magnitude lower energy-delay product while keeping digital-comparable accuracy.

II. PRELIMINARIES AND BACKGROUND

A. Transformer and Self-Attention

Transformer [55] was initially proposed as a sequence transduction model for NLP tasks. Three mainstream Transformer

TABLE I: Comparison of our dynamically-operated photonic tensor core, DPTC, to prior PTC designs. Each PTC takes two operands to perform either matrix-vector multiplication (MVM) or matrix multiplication (MM). Previous PTC designs fail to efficiently support (1) dynamic MM in attention due to the high cost of operand mapping or device programming and (2) full-range MM with no extra overhead, such as duplicated PTCs or multiple inferences.

PTC Designs	MZI array [47]	PCM Crossbar [16]	MRR Bank 1 [52]	MRR Bank 2 [51]	Ours DPTC
Operand 1	Static, Full-range	Static, Positive Only	Dynamic, Full-range	Dynamic, Positive Only	Dynamic, Full-range
Operand 2	Dynamic, Full-range	Dynamic, Positive Only	Dynamic, Positive Only	Dynamic, Positive Only	Dynamic, Full-range
Mapping & Programming Cost	High	Medium	Low	Low	Low
Operation Type	MVM	MM	MVM	MVM	MM
Dynamic MM Support (Attention)	✗	✗	✓	✓	✓
Full-range MM Support (No overhead)	✓	✗	✗	✗	✓

architectures are encoder-decoder (BERT [25], ViT [14]), causal decoder (GPT-series [6]), and the prefix decoder (GLM-130B [66]). Despite different Transformer architectures, they are usually a stack of several identical blocks. Both encoder and decoder blocks comprise a multi-head self-attention (MHA) module, a feed-forward network (FFN), the shortcut connection, and layer normalization (LN) [4]. The decoder additionally adds cross-attention and masked self-attention modules. We use the basic encoder block as an example, defined as

$$\begin{aligned} \mathbf{X}'_{l+1} &= \text{MHA}(\text{LN}(\mathbf{X}_l)) + \mathbf{X}_l; \\ \mathbf{X}_{l+1} &= \text{FFN}(\text{LN}(\mathbf{X}'_{l+1})) + \mathbf{X}'_{l+1}, \end{aligned} \quad (1)$$

where \mathbf{X}_l is input sequences of l -th layer.

Multi-head Self-Attention (MHA). Attention is a novel feature of Transformers where pairwise correlations across the entire input sequence are computed. MHA has H self-attention heads. In each head, the input vector is transformed into the query (\mathbf{Q}), key (\mathbf{K}), and value (\mathbf{V}) vectors by linear projection. Then, the attention function between different input vectors is

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\mathbf{Q}\mathbf{K}^T / \sqrt{d_k})\mathbf{V}, \quad (2)$$

where d_k is \mathbf{Q} and \mathbf{K} 's dimension. Although the attention is still based on matrix multiplication, notably, it is totally different from the linear layer, which poses unique challenges for photonic inference accelerators. In a linear layer, the MM is conducted between the static weight matrix and the dynamic input matrix. In contrast, attention requires the MM between dynamically generated matrices, i.e., \mathbf{Q} , \mathbf{K} , and \mathbf{V} .

Feed-forward Network (FFN). The FFN module usually contains two linear layers with an activation function in between. GELU [22] is a popular option with better performance.

B. Optical Computing Device Basics

Phase shifter (PS): PS is an active device that produces a controlled phase shift ϕ on the light signal x by manipulating the waveguide's effective refractive index. The output is $e^{j\phi}x$.

Directional coupler (DC): DC is a passive device that can produce interference between two coherent light signals. The device consists of two waveguides positioned close to each other such that energy can transfer between them. Its transfer matrix of a 2-by-2 DC is $\begin{bmatrix} t & \sqrt{1-t^2}j \\ \sqrt{1-t^2}j & t \end{bmatrix}$, where t is the transmission coefficient. $t=\sqrt{2}/2$ in a 3 dB 50:50 DC.

Mach-Zehnder modulator (MZM): MZM starts with one splitter that splits the optical input E_{in} into upper and lower modulator arms. After being phase modulated, the signals from

the two arms are recombined as the optical output E_{out} . With equal splitting and differential phase shift, $+\phi$ and $-\phi$ on the two arms [27], respectively, full-range encoding $E_{out} = E_{in} \cos \phi$ can be achieved with MZM by tuning $\phi \in [0, \pi]$.

Microring/microdisk resonator: Micro-ring (MRR) and microdisk (MD) resonators are compact photonic devices that serve as narrowband filters to enable the transmission of a certain wavelength. They can be utilized for constructing optical switches and WDM MUX and DEMUX units.

Mach-Zehnder interferometer (MZI): MZI consists of two cascaded directional couplers and two phase shifters. It can perform arbitrary 2-D unitary matrix operations, thus serving as a basic building block for constructing the MZI array [47].

C. Challenges of Prior Photonic Accelerators for Transformer

Unique features of Transformer workloads. Compared to weight-static NN architectures, the unique execution patterns of Transformers have brought unique workload characteristics.

① Attention modules require **MM with two dynamic input operands**, which requires frequent operand switching and real-time operand mapping and device programming. If operand mapping and device programming are fairly slow compared to ultra-fast computing speed, it will result in severe system stall and significantly increase latency. To clarify, we define the concepts of *dynamic* and *static* operands. In attention, both operands are input-dependent activations generated in real-time, rendering them dynamic. In contrast, weight matrices are fixed (static) during the inference. ② Transformer requires **MM with full-range operands**, as activations are not restricted to non-negative only. This is especially true across Transformer layers with the widespread utilization of GELU and LayerNorm.

Prior photonic accelerators. In this paper, we focus on universal linear units that can potentially support MHA and FFN in Transformers. Specifically, we consider MZI array [47], MRR bank [51], [52], and non-volatile PCM-based crossbar [16]. We omit the discussion of sub-space or convolution-specialized ones [17], [19], [29], [48], [70]. Table I provides a comprehensive comparison of operand characteristics, mapping & programming cost, operation type (MM/MVM) between PTC designs of existing accelerators and our PTC design, DPTC. We highlight whether they can efficiently support unique Transformer workloads, i.e., dynamic MM and full-range MM.

Challenge 1: Prior weight-static PTCs fail to efficiently support dynamic MMs. Before performing optical computing, we need to map operands onto PTC and program devices to get the desired transfer matrix. However, MZI array requires extra

SVD and phase decomposition to obtain phase information needed for device programming. As both operands of attention are activations, this complicated operand mapping step needs to be performed at runtime, which can introduce significant delay. For instance, on a CPU, the required SVD and phase decomposition step takes ~ 1.5 ms for a 12×12 matrix. As a result, severe system stalls occur, making the use of MZI array impractical for dynamic MM scenarios. Besides, current PTC designs face challenges in terms of bulky area and huge insertion loss. Thus, they favor low-loss, compact, and non-volatile devices, as in the MZI array and PCM crossbar. However, these devices suffer from slow programming (10 ns-10 μ s) that cannot be easily amortized, given the orders-of-magnitude higher programming latency than ultra-fast computing.

Insight 1: Mapping and device programming can dominate total latency for weight-static PTCs. Given the huge gap between ultra-fast computing speed and slow mapping/reprogramming speed, the overhead of setting up weight-static cores can hardly be amortized across batch/token dimensions. To eliminate this dominant mapping cost and largely reduce the total latency, we can optically encode both operands for fast dynamic switching. **Challenge 2: Prior incoherent PTCs fail to efficiently support full-range MMs.** As an incoherent design, at least one of the operands of MRR bank is limited to be non-negative as its computation is based on light intensity modulation. The absence of full-range support requires decomposing full-range operands into differences of non-negative ones $(X_+ - X_-)(Y_+ - Y_-)$ and processing X_+Y_+ , X_+Y_- , X_-Y_+ , and X_-Y_- separately by multiple inferences or duplicated PTCs [48], [51], incurring $>2\text{-}4\times$ extra hardware cost compared to one with full-range support. Since the modulation/DAC cost of inputs is doubled, this will eliminate the advantages gained from amortizing DAC and dynamic modulation costs associated with weight-static dataflow.

Insight 2: Phases of light can boost information processing throughput. Instead of performing non-negative matrix multiplication using the weighted sum of light intensities like incoherent PTCs, coherent PTC allows more information to propagate through the circuit by encoding signs to the extra phase dimension and performing signed computation via interference. Hence, we can design a coherent PTC to efficiently support one-shot full-range MM without the above decomposition overhead.

III. PROPOSED PHOTONIC TENSOR CORE DESIGN

To address the above-mentioned challenges of previous designs, we clarify the ultimate goal of this work and introduce our novel PTC design following the above important insights. **Goal: a customized photonic tensor core design for efficient Transformer acceleration** with three essential traits:

- Support for **general matrix multiplications**.
- Support for **full-range inputs/outputs**.
- Efficient handling of **dynamic operands with low encoding and signal modulation cost**.

To meet these traits, we first propose a dynamically-operated dot-product engine, DDot , capable of computing the dot-product of two *dynamically-encoded full-range* optical vectors.

Based on the basic building block DDot , we then introduce a crossbar-style PTC, DPTC , with maximized intra-core operand sharing, enabling ultra-parallel and energy-efficient MM.

A. DDot : Dynamically-Operated Full-range Dot-Product Unit

To perform optical dot-product between vectors \mathbf{x} and \mathbf{y} , we design a DDot based on coherent interference shown in Figure 2(a). We employ the wavelength-division multiplexing (WDM) technique and encode each input pairs (x_i, y_i) in the same wavelength λ_i . We input the WDM light signals carrying \mathbf{x} and \mathbf{y} through the two arms of 50:50 directional coupler (DC) with a -90° phase shifter (PS) on DC's upper left port. Considering each input pair (x_i, y_i) at the same wavelength, the outputs at the right and left output ports of the DC, z_i^0 and z_i^1 can be computed as,

$$\begin{aligned} \begin{pmatrix} z_i^0 \\ z_i^1 \end{pmatrix} &= \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & j \\ j & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & e^{-j\frac{\pi}{2}} \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} \\ &= \frac{1}{\sqrt{2}} \begin{pmatrix} x_i + y_i \\ j(x_i - y_i) \end{pmatrix}. \end{aligned} \quad (3)$$

The two output signals are orthogonal in the complex plane. By adopting broadband devices, we can have the same transfer function for a range of wavelengths. In this way, with WDM signals, each wavelength with (x_i, y_i) encoded interferes in parallel following Eq. (3), while different wavelengths don't interfere. The photo-diode (PD) at the end of each output port of DC converts the incident WDM signals to the photocurrent. The generated photocurrent is proportional to the accumulated intensities of the WDM signals, which is the square of optical magnitudes. Thus, the photocurrents generated at the right and left PD denoted as I^0 and I^1 , can be expressed as,

$$\begin{aligned} \begin{pmatrix} I^0 \\ I^1 \end{pmatrix} &\propto \frac{1}{2} \begin{pmatrix} R^0 \sum_{i=0}^{N-1} \|x_i + y_i\|^2 \\ R^1 \sum_{i=0}^{N-1} \|j(x_i - y_i)\|^2 \end{pmatrix} \\ &\propto \frac{1}{2} \begin{pmatrix} \sum_{i=0}^{N-1} R^0 (x_i + y_i)^2 \\ \sum_{i=0}^{N-1} R^1 (x_i - y_i)^2 \end{pmatrix}. \end{aligned} \quad (4)$$

R^0 and R^1 are the responsivities of right and left PD. Balanced photodetection ($R^0 = R^1 = R$) is employed for subtraction between I^0 and I^1 , producing the final output current as,

$$\begin{aligned} I_o &\propto R^0 \sum_{i=0}^{N-1} (x_i + y_i)^2 - R^1 \sum_{i=0}^{N-1} (x_i - y_i)^2 \\ &\propto R \left(\sum_{i=0}^{N-1} (x_i + y_i)^2 - \sum_{i=0}^{N-1} (x_i - y_i)^2 \right) \propto R \left(\sum_{i=0}^{N-1} x_i y_i \right) \propto \mathbf{x} \cdot \mathbf{y}. \end{aligned} \quad (5)$$

The differential photocurrent naturally cancels out the quadratic terms and carries the dot-product of \mathbf{x} and \mathbf{y} . Note that \mathbf{x} and \mathbf{y} for DDot can be arbitrary vectors without any restriction.

Our DDot leverages coherent light interference and WDM technique to enable general full-range vector dot-product. Unlike prior PTC designs [47], [52], our DDot can *simultaneously* support the following three critical features: **1 Support full-range inputs and outputs**. Given the coherent interference mechanism of DDot , the signs of inputs can be encoded to phases using MZM to achieve full-range encoding. As shown

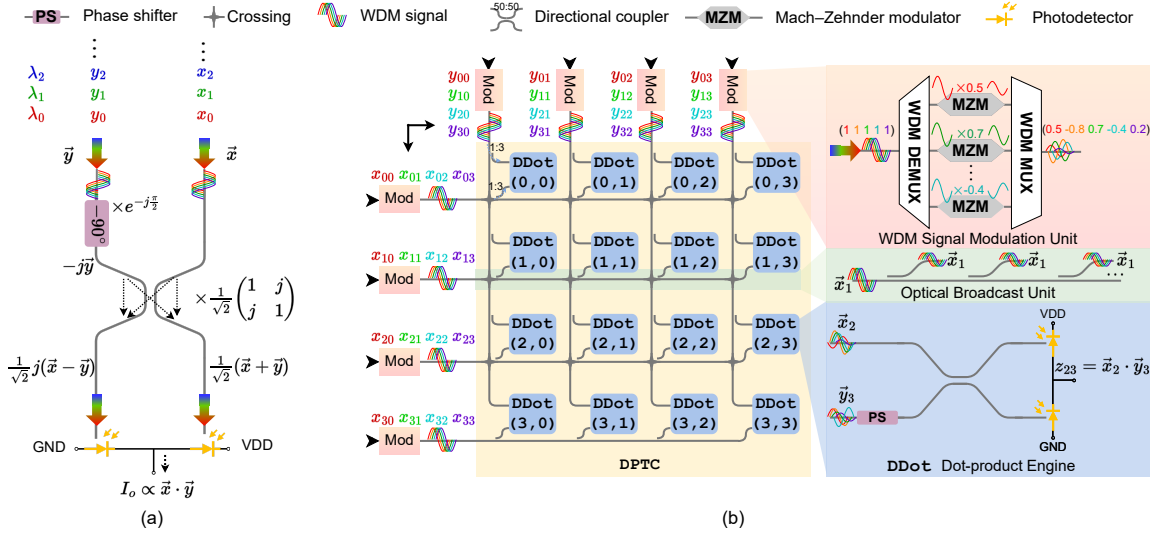


Fig. 2: (a) The proposed DDot dot-product engine. Multi-wavelength signals propagate concurrently on the waveguide. (b) The proposed DPTC matrix-matrix multiplication unit with input WDM signals broadcasting.

in Section II-B, the electric field at the output port of MZM $E_{out} = E_{in} \cos \phi$ that allows for a full encoding range of $[-1, 1]$ by tuning $\theta \in [0, \pi]$. Besides, balanced photodetection enables the detection of full-range outputs as positive or negative photocurrent. Different from incoherent designs [48], [51], [52] that require decomposing full-range operands into two non-negative ones and processing separately, we inherently support full-range operands at one shot with no extra overhead. **⊕ Support for dynamic operand.** Both operands are optically encoded at high speed (~ 10 ps), thus, can be flexibly switched/reprogrammed without causing any latency bottleneck, unlike previous weight-static PTCs [16], [47], a crucial feature to support attention. **⊕ Superior computing parallelism and fully-passive computing core.** DDot leverages WDM to enable spectral parallelism such that different wavelengths can share the same DDot unit, achieving superior computing density and parallelism. The PS and DC in DDot are entirely passive/fixed, resulting in zero energy consumption, no external control overhead, and no thermal crosstalk concern.

B. DPTC: Dynamically-Operated Photonic Tensor Core

To support MM using optical dot-product engines, some accelerators [1], [35], [51] directly map each dot-product of MM onto the dot-product engine. However, it incurs nontrivial signal modulation overhead as operand sharing is limited.

Therefore, we present a novel photonic tensor core design, named DPTC, by constructing a compact crossbar-style array of DDot units to **maximize the intra-core operand sharing** to largely reduce operand modulation cost, shown in Figure 2(b). This design enables the efficient sharing of photonic waveguide buses across DDot units and facilitates ultra-parallel MM. A $N_v \times N_h$ DPTC consists of $N_v \times N_h$ DDot units, where N_v and N_h represent the numbers of input waveguide along the vertical and horizontal directions, respectively.

WDM signal modulation unit (Figure 2(b) pink region): The optical inputs are driven by coherent sources with phase shifters

to control phases. Each waveguide in DPTC has a WDM signal modulation unit where total N_λ wavelengths are separated by a WDM DEMUX, individually modulated by high-speed MZMs, and merged in one waveguide by a WDM MUX.

Intra-core optical broadcast unit (Figure 2(b) green region): To largely reduce the signal modulation cost, each modulated WDM signal is broadcast to a row or a column of DDot units through the intra-core optical broadcast unit. A copy of the input vectors is coupled out from the optical bus and fed into the DDot, enabling operand sharing and thus largely reducing modulation overhead in our dynamically-operated design. Specifically, for a $[N_h, N_\lambda] \times [N_\lambda, N_v]$ MM workload, the DAC and MZM modulation cost of our DPTC is

$$E_{\text{encode}} \approx (N_h N_\lambda + N_\lambda N_v)(E_{\text{DAC}} + E_{\text{MZM}}). \quad (6)$$

Compared to prior work that simply utilizes separate vector dot-product engines to implement MM without enabling operand sharing [1], [35], [51], whose encoding cost is $(2N_h N_v N_\lambda)(E_{\text{DAC}} + E_{\text{MZM}})$, the intra-core optical broadcast helps DPTC save $(2N_h N_v)/(N_h + N_v) \times$ encoding cost. For instance, when $N_h = N_v = N_\lambda = 12$, DPTC shows $12 \times$ less encoding cost. This is one key reason why our design can still preserve high energy efficiency even if we need dynamically modulate both operands.

To summarize, DPTC inherits the ability to support full-range dynamic operands from DDot and incorporates a more compact and energy-efficient crossbar-style design. Unlike most previous designs supporting only MVM, our DPTC enables one-shot MM with ultra-high computation parallelism. Moreover, leveraging the broadcast ability of light, we maximize the intra-core sharing of modulated signals among multiple DDot units to largely amortize the operand encoding cost.

C. Robustness Analysis of Proposed Photonic Design

Analog optical computing systems are subject to various noises, e.g., encoding noise, WDM dispersion, and device

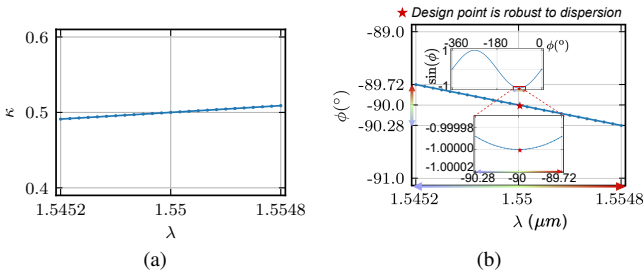


Fig. 3: Our design point is robust to non-ideal dispersion effects. Coupling coefficient κ and phase shift ϕ are not sensitive to wavelength-dependent device responses (i.e., dispersion).

imperfection. Here, we analyze the noise impact on fundamental DDot unit and show its inherent robustness to variations.

Optical encoding noise. In DDot, both operands are encoded as optical signals, which are inevitably susceptible to encoding noise, i.e., stochastic magnitude and phase drift. Specifically, consider two optical operands x and y , we have $x' = \hat{x}_i e^{j\delta\phi_x} = (x + \delta x)e^{j\delta\phi_x}$ and $y' = \hat{y}_i e^{j\delta\phi_y} = (y + \delta y)e^{j\delta\phi_y}$, where δx and δy denote the magnitude drift, and $\delta\phi_x$ and $\delta\phi_y$ denote the phase drift. For simplicity, we extract the relative phase drift between the two operands and express it as an equivalent phase drift $\delta\phi_d = \delta\phi_y - \delta\phi_x$, following a Gaussian distribution $\delta\phi_d \sim \mathcal{N}(0, \sigma_\phi^2)$. The magnitude drift follows a Gaussian distribution $\delta x \sim \mathcal{N}(0, (\sigma_x|x|)^2)$, where the standard deviation depends on the absolute value $|x|$ we want to encode. With the encoding noise, the noisy transfer function of DDot is expressed as,

$$\begin{aligned} \begin{pmatrix} z_i^0 \\ z_i^1 \end{pmatrix} &= \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & j \\ j & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & e^{-j\pi/2} \end{pmatrix} \begin{pmatrix} \hat{x}_i \\ \hat{y}_i e^{j\delta\phi_d} \end{pmatrix} \\ &= \frac{1}{\sqrt{2}} \begin{pmatrix} \hat{x}_i - \sin\phi_i \hat{y}_i + j\hat{y}_i \cos\phi_i \\ \hat{y}_i \cos\phi_i + j(\hat{x}_i + \sin\phi_i \hat{y}_i) \end{pmatrix}, \end{aligned} \quad (7)$$

where $\phi_i = \delta\phi_d - \pi/2$ as a perturbed value around $-\pi/2$.

WDM dispersion. Our architecture leverages WDM to allow multiple wavelengths to share the same DDot unit. Nevertheless, even with the adoption of broadband devices (coupler, phase shifter), photonic circuits still exhibit slightly different responses to different wavelengths. Specifically, the wavelength-dependent transfer function of the directional coupler is described by $\begin{pmatrix} t & k \\ k & t \end{pmatrix}$, where $t = \sqrt{1 - \kappa(\lambda)}$ and $k = \sqrt{\kappa(\lambda)}$. $\kappa(\lambda)$ is the wavelength-dependent power coupling factor and computed as $\kappa(\lambda) = \sin^2((\pi L_c(\lambda_0))/4L_c(\lambda))$, where $L_c(\lambda)$ is the 100% coupling length. $\kappa(\lambda_0)$ is designed to be 1/2. Besides, the phase response of the phase shifter $\Delta\phi(\lambda) = 2\pi\Delta n_{\text{eff}}L/\lambda$, which is also wavelength-dependent. In this paper, we follow Dense WDM standard [24] with a 0.4 nm wavelength channel spacing and choose the center wavelength $\lambda_0 = 1.55$ nm. We sweep 25 wavelengths around λ_0 and show the corresponding κ and ϕ in Figure 3. The maximum relative difference of κ between λ_0 and the furthest wavelength is $\sim 1.8\%$. The maximum dispersion-induced phase difference is 0.28° , which is negligible compared to the 360° MZM tuning range.

By considering the impact of WDM dispersion, we have the

wavelength-dependent transfer function of DDot as

$$\begin{pmatrix} z_i^0 \\ z_i^1 \end{pmatrix} = \begin{pmatrix} (t_i \hat{x}_i - k_i \hat{y}_i \sin\phi_i) + jk_i \hat{y}_i \cos\phi_i \\ t_i \hat{y}_i \cos\phi_i + j(k_i \hat{x}_i + t_i \hat{y}_i \sin\phi_i) \end{pmatrix}, \quad (8)$$

where $\phi_i = \delta\phi_d - \pi/2 + \delta\phi_{\lambda_i}$ including dispersion induced phase drift $\delta\phi_{\lambda_i}$ in radian mode. Hence, the output photocurrent of DDot, considering both encoding noise and dispersion, is

$$\begin{aligned} I_o &\propto \sum_{i=0}^{N-1} \left(\frac{(k_i^2 - t_i^2)\hat{x}_i^2 + (t_i^2 - k_i^2)\hat{y}_i^2}{2} + 2t_i k_i \hat{x}_i \hat{y}_i (-\sin\phi_i) \right) \\ &\propto \sum_{i=0}^{N-1} \underbrace{(2k_i \sqrt{1 - k_i^2} (-\sin\phi_i) \hat{x}_i \hat{y}_i)}_{\text{multiplicative noise}} + \underbrace{(2k_i^2 - 1) \frac{\hat{x}_i^2 - \hat{y}_i^2}{2}}_{\text{additive noise}}, \end{aligned} \quad (9)$$

where each scalar product $\hat{x}_i \hat{y}_i$ has non-ideal scaling factors $\sin\phi_i$ and $2k_i \sqrt{1 - k_i^2}$, and an additive noise relative to $\hat{x}_i^2 - \hat{y}_i^2$. DDot is inherently robust to **1** multiplicative noise, as our design points $k = \frac{1}{\sqrt{2}}$ and $\phi = -\frac{\pi}{2}$ are at the local optima of $x\sqrt{1-x^2}$ and $\sin(\cdot)$ with minimized sensitivity to perturbations. DDot is robust to **2** additive noise. Our photonic design is also robust to the additive error as it can be naturally canceled out. In our DDot engine, MZM modulation range is $[-1, 1]$. Hence, before mapping \mathbf{x} and \mathbf{y} onto DDot, we need to scale them into $[-1, 1]$ with their maximum absolute value $\beta_x = \max(|x|)$ and $\beta_y = \max(|y|)$. This normalization effect naturally ensures \hat{x}_i and \hat{y}_i in Eq. (9) in a similar range, i.e., $[-1, 1]$, such that in a vector dot-product, the introduced multiple \hat{x}_i^2 and \hat{y}_i^2 can be self-compensated. Moreover, the square effect and scaling factor $\frac{1}{2}$ largely reduce the additive noise.

Overall, our design shows remarkable robustness to phase drift and WDM dispersion. The exceptional noise tolerance to dispersion enables DDot to scale with a large number of wavelengths, significantly enhancing the spectral parallelism.

Other Noises. To further simulate other noises in the system (DPTC), e.g., photo-detection noise, imperfect coupling ratios of direction couplers, we generally add a multiplicative noise to the computation outputs of DPTC, $\hat{I}_o = I_o(1 + \varepsilon)$, where $\varepsilon \sim \mathcal{N}(0, 0.05^2)$, when simulating the accuracy of running Transformer on our photonic designs.

IV. PROPOSED PHOTONIC ACCELERATOR ARCHITECTURE

In this section, we present the high-level architectural design of the proposed accelerator, Lightning-Transformer.

A. Overall System Design

The Lightning-Transformer architecture contains analog photonic computing units for General matrix multiplication (GEMM) acceleration, photonic interconnect for data broadcast, and underlying electronics for other operations, including signal conversion, data storage, nonlinearity functions, and SoftMax.

Architecture overview. Figure 4 illustrates the overall micro-architecture with a particular emphasis on the photonic part. We incorporate multiple photonic tensor cores (DPTC) into a single chip to increase the amount of parallel processing. We cluster N_c DPTC into one tile, and we have N_t tiles in total in

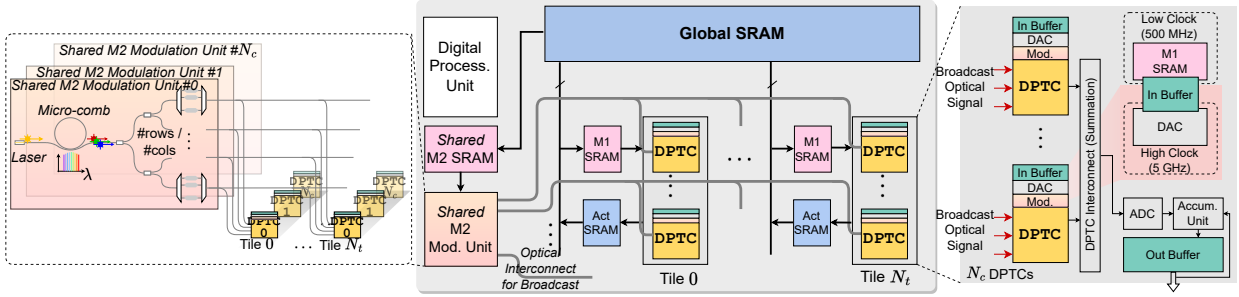


Fig. 4: High-level architecture of the proposed Lightning-Transformer. It has a three-level memory hierarchy, multiple photonic analog computing tiles/cores, on-chip multi-wavelength light sources, and optical interconnects for data broadcast.

the accelerator. All the photonic tensor cores DPTC are clocked at 5GHz for a conservative assumption. The operands of each DPTC undergo an electrical-to-optical conversion (E/O) with digital-to-analog converters (DAC), and the outputs of DPTC need analog-to-digital converts (ADC) to bring the analog result back to the digital domain. We have digital processing units to process non-GEMM operations needed in Transformer.

Memory. For the memory part, we have a large 2MB global SRAM, which interacts with DRAM and is responsible for holding inputs, activations, and weights. The size of the global SRAM is designed to be sufficient for (1) storing single-layer largest activations for targeted low-bit BERT/DeiT Transformers’ single-batch inference. (2) double buffering for required off-chip data that is loaded chunk by chunk based on the tiling algorithm in Figure 5 to overlap data transfer time with computation. The corresponding off-chip HBM bandwidth is set to ensure the data access latency is hidden by computation latency. In addition, each tile has its own 4KB SRAM to hold the operands for its own matrix multiplication workload, as well as an activation SRAM to store the computed results. Since the photonic part of Lightning-Transformer runs at a clock speed of 5GHz, and our DPTC unit handles one small matrix multiplication in one clock cycle, the required data bandwidth is high. To address this, we follow [10] and decouple the large SRAM array into smaller 32KB sub-arrays. This improves data bandwidth by reading data from these sub-arrays with a small shifting interval. Data buffers are used for each photonic tensor core to communicate the DACs and SRAM in different clock domains.

Lightning-Transformer uses output-stationary (OS) dataflow and proposes multiple architecture-level optimization techniques to efficiently support GEMM operations, which will be elaborated on later.

Photonic tensor cores. Photonic tensor core DPTC works with necessary digital components, including DACs, ADCs, TIAs, and data buffers and operates at 4-bit precision by default. DPTC can implement $[N_h, N_\lambda] \times [N_\lambda, N_v]$ matrix multiplication in one cycle. Large GEMM operations in Transformer are mapped onto DPTC by using tiled matrix multiplication.

Photonic Interconnect. We leverage photonic interconnect to broadcast optical signals across photonic tensor cores when operand sharing opportunity exists.

Digital processing units. We assume all other non-GEMM

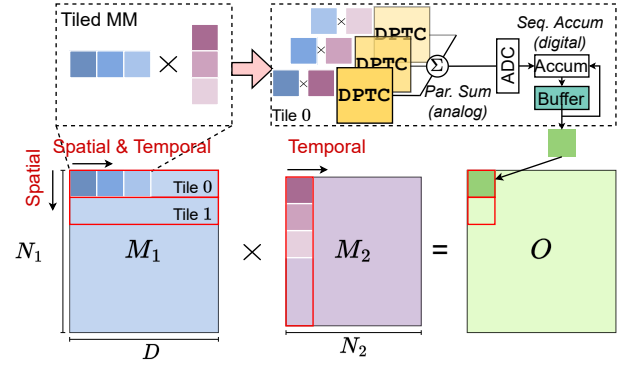


Fig. 5: Tiling and spatial/temporal mapping for processing GEMM. M_1 is the weight matrix when processing the linear layer, which is loaded chunk by chunk off-chip.

operations are implemented using digital electronics.

TABLE II: Table of notations used in the accelerator design

Notation	Definition
N_h	# input horizontal waveguides of each DPTC
N_v	# input vertical waveguides of each DPTC
N_λ	# wavelengths of each DPTC
N_c	# photonic tensor cores in each tile
N_t	# tiles in our accelerator

B. Dataflow

Our dynamically-operated DPTC frees the selection of dataflow by having both operands dynamically encoded. Most prior photonic accelerators typically map one operand into photonic circuit states that cannot be dynamically switched due to slow reconfiguration speed. This design concept limits its dataflow selection, making them only suitable for weight-stationary (WS) dataflow [10]. Note that the WS stationary dataflow in photonic accelerators differs from that of the digital electronics accelerator. They are constrained to only favor the WS dataflow, while the normal WS stationary dataflow is a design choice to explore data locality and data reuse.

To efficiently support GEMM on our multi-tile accelerator, we consider fine-grained tiling and carefully design the spatial/temporal mappings, as illustrated in Figure 5. We use the OS dataflow as the basic design principle for calculating MM between M_1 and M_2 . The OS dataflow enables us

to minimize the on-chip buffer size [33], [64] and adopt customized optimization techniques to reduce the cross-domain signal conversion overhead in our mixed-signal accelerator. Specifically, we tile matrix \mathbf{M}_1 along the N_1 dimension and map them to different tiles spatially. Each tile is responsible for calculating the multiplication result between one tiled row of \mathbf{M}_1 and matrix \mathbf{M}_2 temporally. At each cycle, each tile handles tiled matrix multiplication. Since we have multiple DPTC in each spatial unit, we distribute the tiled MM workload among these cores. The outputs of multiple cores are first accumulated by photocurrent summation in the analog domain, followed by A/D conversion. Then, the partial sums are sent to the output buffer for further sequential accumulation in the digital domain. This analog domain partial summation can not only reduce the A/D conversion cost but also avoid the precision loss during A/D conversion with full-precision photocurrent summation.

C. Architecture-Level Optimization

The expensive electrical-to-optical (E-O) conversion and optical-to-electrical (O-E) costs remain to be the key bottleneck for emerging photonic systems, which is also true in Lightning-Transformer. Therefore, we consider the two following optimization opportunities to reduce the E-O and O-E costs.

1) Inter-core Operand Broadcast via Optical Interconnect:

As shown in Figure 5, different tiles process the multiplication between different chunks of \mathbf{M}_1 and the same chunk of \mathbf{M}_2 . This creates an opportunity to share the common \mathbf{M}_2 part across multiple tiles. Leveraging the exceptional signal broadcast capabilities of optics, we encode the shared \mathbf{M}_2 in optical signals using global modulation units (as shown in the left part of Figure 4). These modulated signals are efficiently broadcasted to different DPTC units via optical interconnects, leading to an architecture-level $N_t \times$ reduction in data movement and modulation costs.

2) Analog-Domain Temporal Accumulation with Time Integral:

As our DPTC design supports efficient dynamic operand switching, we adopt OS dataflow that can enable analog-domain temporal accumulation [30], [68] via time integral to reduce the ADC overhead. This technique is not applicable to prior WS-static accelerators since their outputs are accumulated to different buffer locations across time steps, preventing utilizing temporal accumulation. Analog temporal accumulation accumulates results by using photodetectors and capacitors to store charges, which can be read at a later time. This allows ADC to operate at a lower frequency, thereby reducing ADC costs. Additionally, by performing accumulation before signal digitization, the partial summation is achieved in the analog domain in full precision, leading to accuracy [30]. In this paper, we set temporal accumulation depth to 3, i.e., A/D conversion happens once every three analog accumulation steps.

V. EVALUATION

A. Evaluation Setup

System setup. We build a Python-based simulator to simulate the latency, power, area, and energy efficiency of our proposed

TABLE III: Adopted component parameters in our paper. IL represents insertion loss, and FSR means free spectral range.

Device	Parameter	Value
DAC [7]	Precision	8-bit
	Power	50 mW (@14 GSPS)
	Area	11,000 μm^2
ADC [32]	Precision	8-bit
	Power	14.8 mW (@10 GSPS)
	Area	2,850 μm^2
TIA [43]	Power	3 mW
	Area	<50 μm^2
Microdisk [53]	Locking Power	0.275 mW [†]
	IL	0.93 dB
	Area	4.8 \times 4.8 μm^2
	FSR	5.6 THz (55.1 nm)
Microring Resonator	Tuning Power	0.21 mW
	Locking Power	1.2 mW/0.5FSR [49]
	IL	0.95 dB [39]
	Area	9.66 \times 9.66 μm^2 [39]
MZM	Tuning Power	2.25 mW [13]
	IL	1.2 dB [2]
	Area	260 \times 20 μm^2 [2]
Directional Coupler [63]	IL	0.33 dB
	Area	5.25 \times 2.4 μm^2
MEMS Phase Shifter [42]	IL	0.33 dB
	Area	100 \times 45 μm^2
	Response Time	2 μs
Photodetector [23]	Power	1.1 mW
	Sensitivity	-25 dBm
	Area	4 \times 10 μm^2
Y-branch [36]	IL	0.3 dB
	Area	1.8 \times 1.3 μm^2
Micro-comb [62]	Area	1,184 \times 1,184 μm^2
On-chip Laser	Wall-plug Efficiency	0.2 [58]
	Area	400 \times 300 μm^2

accelerator on actual Transformer inference. The area, leakage power, and access energy of the memory system are modeled using PRACTI [45] in 14 nm. We use high-bandwidth memory (HBM) to supply data to the photonic system with a bandwidth >1TB/s [37]. The energy cost of digital processing units is obtained from [21], [40], [59]. We use the ADC [32] and DAC [7] with similar technology nodes (14 nm and 16 nm), while their bit-widths and frequency do not precisely match our settings. We follow [26] to scale the ADC and DAC power according to the bit-width and frequency requirement of the photonic computing units. Table III lists the parameters of used photonic devices. The laser power is set to meet the minimum power requirement of the photodetector considering system loss [51] and is scaled based on the precision requirement and wall-plug efficiency. During simulation, the simulator implements the detail of the tiling algorithm and uses a batch size of 1. The deep pipeline of the photonic/digital processing unit is not adopted in this paper, which can be employed to further improve the system performance [10].

Functionality validation. We use Lumerical INTERCONNECT tools [34] with the AIM process design kit (PDK) to validate the functionality of our DDot engine. We inject optical encoding noise as discussed in Section III-C with input noise std. and phase noise std. being 0.03 and 2°, respectively. WDM dispersion is naturally considered in the simulation. 12 wavelengths are used with 0.4 nm wavelength channel spacing. The dot-product error of one random length-12 dot-product is 2.6% and 3.4% in 4-bit and 8-bit precision, respectively, as

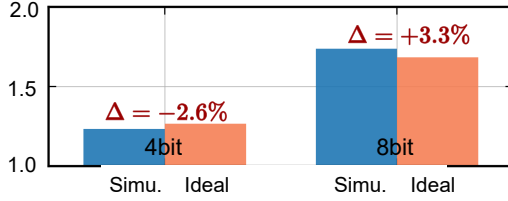


Fig. 6: Optical simulation results of 4-bit and 8-bit random length-12 dot-products on our DDot engine.

TABLE IV: Base (B) and Large (L) configurations for Lightning-Transformer.

Ours	N_t	N_c	N_h	N_v	N_λ	Global SRAM (MB)	area (mm^2)
LT-B	4	2	12	12	12	2	60.3
LT-L	8	2	12	12	12	4	112.82

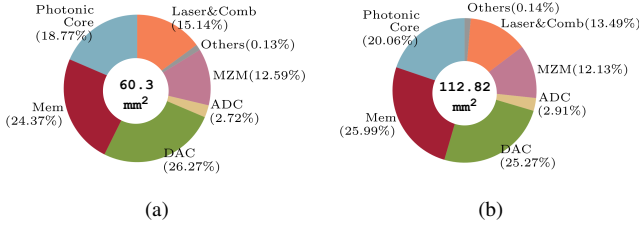


Fig. 7: Area breakdown of (a) LT-B and (b) LT-L.

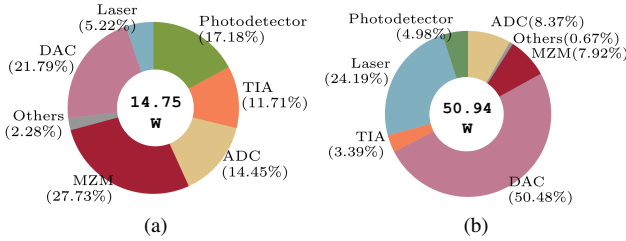


Fig. 8: Power breakdown of LT-B for (a) 4-bit (b) 8-bit precision. Others include memory system static power.

shown in Figure 6.

Models, datasets, training, and inference settings. We use DeiT [54] and BERT [11] to evaluate the efficiency and accuracy. Both are well-recognized Vision and NLP Transformers. We apply low-bit quantization on both weight and activation [15]. Noise-aware training is applied with encoding and systematical noise injected. We model the computation by using noisy analytic transformation (Eq. (9)) with all noises in Section III-C injected to test the inference accuracy.

Architecture configurations. We design two versions of Lightning-Transformer called LT-B (base version) and LT-L (large version) with detailed parameters in Table IV. LT-B has four tiles, where each tile has two DPTC. LT-L doubles the number of tiles to increase its throughput for large models.

B. System Efficiency Analysis

Area breakdown. Figure 7 shows the area breakdown of our LT-B and LT-L architectures. The LT-B has 60.3 mm^2 , which is around half of the area of the LT-L architecture (112.82 mm^2). For each architecture, the photonic core, memory, and

DAC contribute the largest portion of the area, with around 20%, 25%, and 25% share, respectively. The remaining components, such as laser, ADC, and MZM, account for less than 30% of the overall area.

Power breakdown. Figure 8 shows the power breakdown for LT-B for two data precision settings, 4-bit and 8-bit. The LT-L shares a similar power breakdown with 28.06 W (4-bit) and 95.92 W (8-bit) in total. In our Lightning-Transformer, all operands are dynamically encoded, resulting in a dominant operand encoding cost (DAC and MZM). The reported operand modulation cost has been optimized through operand sharing at intra-core and inter-core levels. The 8-bit LT-B consumes more than three times the power of the 4-bit one. This increase is primarily attributed to the largely increased power of high-bit DACs, which account for over 50% of the overall power in the 8-bit architecture. Also the laser power also significantly increases from 0.77 W to 12.3 W to satisfy higher output precision.

Area, power, latency, and performance scaling. In Figure 9, we show how the area, power consumption, and latency scale with the size of DPTC. Here, we don't consider input to be globally modulated, i.e., DACs are not shared across different tiles for better observing the scaling effect. The area increases from 5.9 mm^2 to 49.3 mm^2 when increasing the core size from 8 to 32. The ratio of each part roughly remains the same. Power consumption of one single core increases from 1.1 W to 17 W as the core size increases from 8 to 32, and the modulation, ADC, and DAC take the lion's share of the overall power consumption. The latency scaling shows a different pattern from the area and power consumption. The optics latency increases approximately linearly with the size as the optical path increases. The EO/OE latency remains almost the same. As shown in Figure 10, we further show the performance and efficiency scaling of our design. As the core size increases, the performance (TOPS), area efficiency (TOPS/mm^2), and energy efficiency (TOPS/W) increase while the energy efficiency per unit area ($\text{TOPS}/\text{W}/\text{mm}^2$) decreases due to the bottleneck of ADCs and DACs.

Wavelength scaling. Thanks to the superior robustness of our photonic design to WDM dispersion effects, we can increase the number of wavelengths to further boost processing parallelism. In our system, we use Microdisk [53] as the filter to implement the WDM MUX and WDM DEMUX. However, the microdisk imposes a free spectral range (FSR), which limits the number of wavelengths. Given its $\text{FSR}=5.6 \text{ THz}$ and design wavelength at 1550 nm , we can have its wavelength range as

$$\begin{aligned} \lambda_l &= c/(f_0 + \text{FSR}/2) = 1527.88 \text{ nm}; \\ \lambda_r &= c/(f_0 - \text{FSR}/2) = 1572.76 \text{ nm}. \end{aligned} \quad (10)$$

With a 0.4 nm channel spacing, we have up to 112 wavelengths.

C. Compare to State-of-the-art Photonic Accelerators

Baseline: We have built baseline systems based on incoherent MRR bank [52] and coherent MZI array [47]. For a fair comparison, we scale the number of PTC in baselines to match area of our base version of Lightning-Transformer

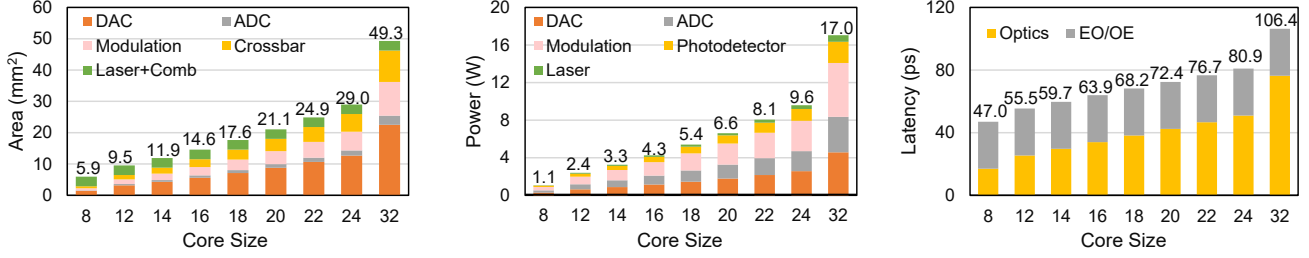


Fig. 9: Area, power, and latency scaling of single 4-bit DPTC core with increasing core size N , where $N_h = N_w = N_\lambda = N$.

(LT-B). Since the MZI array cannot efficiently support dynamic MM, we assume to use MRR bank to implement the MHA part for it. We adopt weight-static dataflow for baselines.

Energy cost model of PTC execution: Consider a GEMM between two matrices, $\mathbf{X} \in \mathbb{R}^{m \times d}$ and $\mathbf{Y} \in \mathbb{R}^{d \times n}$, and a PTC performing $[k, k] \times [k, k]$ MM at one time where $k' = 1$ when PTC supports MVM. k is typically small compared to the matrix dimension. Thus, tiled matrix multiplication is needed for large GEMM operations, executing PTC $T = \lceil \frac{m}{k} \rceil \lceil \frac{d}{k} \rceil \lceil \frac{n}{k'} \rceil$ times. If the energy to load (detect) one scalar input (output) is E_{load} (E_{det}) and the laser energy per cycle is E_{laser} , the PTC computation energy cost [28] (w/o data-movement cost) is:

$$\begin{aligned}
 E &= T \cdot E_{\text{laser}} + \underbrace{T \cdot k^2 E_{\text{load}}}_{\text{load X}} + \underbrace{T \cdot k k' E_{\text{load}}}_{\text{load Y}} + \underbrace{T \cdot k k' E_{\text{det}}}_{\text{output}} \\
 &\approx T \cdot E_{\text{laser}} + m d \lceil \frac{n}{k} \rceil (E_{\text{DAC}} + E_{\text{mod}}) \\
 &+ d n \lceil \frac{m}{k} \rceil (E_{\text{DAC}} + E_{\text{mod}}) + \lceil \frac{d}{k} \rceil m n (E_{\text{PD}} + E_{\text{amp}} + E_{\text{ADC}}).
 \end{aligned} \tag{11}$$

E_{load} contains the energy costs of DAC (E_{DAC}) and signal modulation (E_{mod}). E_{det} represents the cost of detecting optical signal (E_{PD}), amplifying it (E_{amp}), and performing analog-to-digital conversion (E_{ADC}). Note that in weight-static designs (MRR bank and MZI array), the DAC and dynamic modulation cost of the static operand can be amortized. However, MRR bank incurs an additional mW-level static locking power [10] to maintain the encoded value in the resonator, which cannot be amortized. The total locking cost scales with the total number of computations, i.e., mdn .

Comparison on an Attention: In Figure 11(left), we compare our LT-B and MRR baseline on an example MHA workload (\mathbf{QK}^T) in DeiT-T. We disable the architecture-level optimization of LT-B (i.e., photocurrent summation within the same tile, inter-core operand broadcast, and analog domain temporal accumulation) to purely compare the PTC design. We denote this version of LT-B as LT-crossbar-B. Although MRR bank can amortize the DAC cost of the static op1 the unamortized static operand locking power (op1-mod) contributes to >40% of total energy cost. Moreover, its inability to support full-range inputs requires decomposing them into non-negative ones and processing them separately(double T in Eq. (11)). The induced $2 \times$ input encoding (op2-DAC, op2-mod) and detection (det, ADC) cost eliminates the weight-static benefit. Our LT-crossbar-B has a $2.62 \times$ less energy cost with inherent

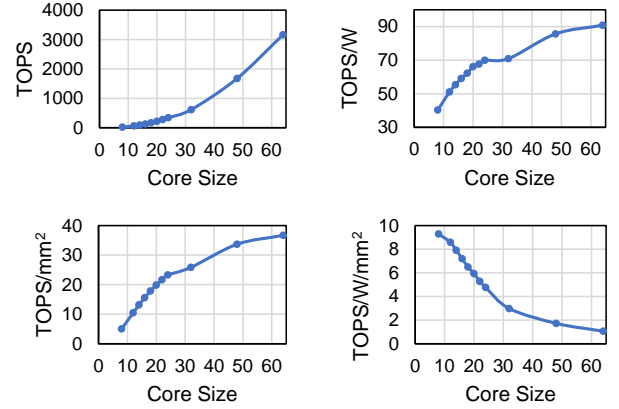


Fig. 10: Performance and efficiency of the optical computing part (ADC/DAC excluded) scale with larger core size.

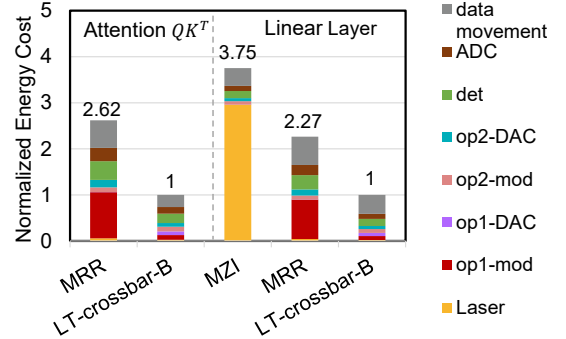


Fig. 11: Energy comparison and breakdown across main components (adder excluded) across LT-crossbar-B (LT-B without architecture-level optimization), MRR bank, and MZI array on example attention and linear layer workloads in DeiT-T. *mod* refers to the energy cost of dynamic modulation and operand locking. *det* refers to the energy costs of Photodetector and TIA. op1 is the static operand in MRR and MZI baselines when using weight-static dataflow.

full-range input support, no locking power, and maximized intra-core operand sharing to amortize encoding costs for both operands.

Comparison on a Linear layer: We highlight that our Lightning-Transformer is more efficient than prior designs, even on weight-static linear layer workload. In Figure 11(right), we compare the energy cost breakdown between MRR, MZI baselines, and LT-crossbar-B (LT-B

TABLE V: Comparison to prior photonic accelerators on in DeiT-T/B. The MHA (Qk^T and AV) and FFN modules are highlighted. We assume MRR bank implements MHA in the MZI array as it cannot support MHA. Our Lightning-Transformer is equipped with our architecture-level optimization discussed in Section IV-C, while even without, still better than baselines.

Precision	Model	Module	MZI-based Accelerator [47]			MRR-based Accelerator [52]			Lightning-Transformer-B			
			Energy↓ (mJ)	Latency↓ (ms)	EDP↓ (mJ*ms)	Energy↓ (mJ)	Latency↓ (ms)	EDP↓ (mJ*ms)	Energy w/o Arch Opt ↓ (mJ)	Energy ↓ (mJ)	Latency↓ (ms)	EDP↓ (mJ*ms)
4bit	DeiT-T	MHA	-	-	-	0.17	0.03	0.005	0.08	0.04	3.12e-3	1.33e-4
		FFN	1.47	6.27	9.19	0.89	0.14	0.12	0.39	0.22	1.04e-2	2.28e-3
		All	2.98	12.37	36.93	1.54	0.24	0.38	0.69	0.38	1.94e-2	7.44e-3
	DeiT-B	MHA	-	-	-	0.67	0.12	0.08	0.34	0.17	1.25e-2	2.13e-3
		FFN	23.46	100.24	2351.23	14.16	2.21	31.34	6.25	3.47	1.67e-1	5.81e-1
		All	44.91	190.46	8554.08	22.08	3.47	76.56	9.79	5.44	2.65e-1	1.44
Average Ratio			8.01	677.56	5426.27	4.03	12.85	51.79	1.80	1	1	1
8bit	DeiT-T	MHA	-	-	-	0.36	0.03	0.01	0.25	0.15	3.12e-3	4.76e-4
		FFN	19.21	6.27	120.32	1.83	0.14	0.25	1.09	0.68	1.04e-2	7.08e-3
		All	37.18	12.37	460.09	3.20	0.24	0.78	1.93	1.21	1.94e-2	2.34e-2
	DeiT-B	MHA	-	-	-	1.43	0.12	0.12	1.02	0.61	1.25e-2	7.61e-3
		FFN	307.27	100.24	30800.44	29.33	2.21	2.21	17.40	10.81	1.67e-1	1.81
		All	580.80	190.46	110620.44	45.77	3.47	3.47	27.33	16.98	2.66e-1	4.51
Average Ratio			32.46	675.67	21944.30	2.67	12.81	34.25	1.61	1	1	1

without architecture-level optimization) on the first linear layer of FFN block in DeiT-T. We emphasize that our is not merely designed for attention but for generic GEMM acceleration supporting ultra-parallel and energy-efficient full-range MM. Even though our design has two operands (W and x) being dynamically encoded, the resultant extra encoding cost (op1-mod and op1-DAC) for weight operand is marginal in our total energy since of our topology-enabled intra-core sharing. Hence, the efficiency superiority of Lightning-Transformer still holds and is well justified. The main reason that leads to the significantly worse energy efficiency of MZI-based PTC is its prohibitive laser power (laser) due to the huge insertion loss of deeply cascaded MZI array, which takes over 75% of its total energy cost. Note that its laser energy itself is already $2.9\times$ higher than our total energy. Moreover, given the limited bandwidth and bulky footprint of the MZI array, we can only fit a few cores on the chip, each supporting MVM using a single wavelength. Thus, it consumes much more cycles (longer latency) to finish an MM workload than our ultra-parallel, compact DPTC design. Based on the above two reasons, even if the weight-static MZI array sounds more efficient on the linear layer workloads than our dynamic tensor core, our Lightning-Transformer shows surprisingly better energy efficiency than it even on linear layer workloads. This conclusion seems counterintuitive but well-explained here and evidenced by the quantitative analysis in Figure 11(right).

Comparison of Lightning-Transformer variants: Figure 12 shows the energy cost breakdown on Attention/linear layers between MRR bank and different variants of the base Lightning-Transformer variants (LT-broadcast-B, LT-crossbar-B, LT-B) to highlight the contributions of introduced features to energy efficiency. LT-B activates all features in our accelerator design, i.e., crossbar-topology and architecture-level optimization. LT-crossbar-B turns off the architecture-level optimization while preserving our crossbar-style topology for our photonic tensor core DPTC. LT-broadcast-B adopts a similar DPTC topology to MRR, only broadcasting input operand to different dot-product units DDot. It doesn't employ the crossbar topology to share both

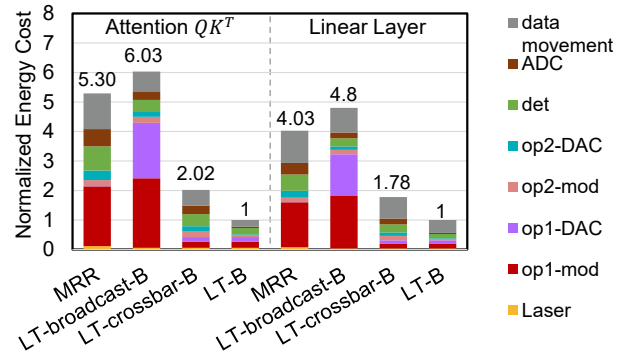


Fig. 12: Energy comparison and breakdown across main components between MRR and LT-B variants on example workloads (Qk^T and first layer of FFN) in DeiT-T. op1 is the static operand when MRR uses weight-static dataflow. LT-broadcast-B adopts a simple topology to only broadcast input operand (op1). LT-crossbar-B is our DPTC design that employs the crossbar topology to share both operands intra-core. LT is complete version that further adopts architecture-level optimization.

operands intra-core but still features dynamic, full-range operand support.

As in Figure 12, MRR bank exhibits significant static operand locking power (op1-mod) ($>40\%$ of the total energy), as weight-static dataflow can only amortize dynamic power. Although vanilla LT-broadcast-B features a slightly higher energy cost than MRR due to the unshared modulation cost of op1 (op1-mod and op1-DAC), the full-range feature results in $>2\times$ smaller energy cost for other parts (op2-DAC, det, ADC), as avoiding the need of running PTC twice for full-range inputs. By adopting the crossbar topology, LT-crossbar-B not only eliminates modulation overhead but also achieves over $2\times$ better energy efficiency than MRR bank.

LT-B achieves the lowest energy cost. Compared to LT-crossbar-B, it further equips with inter-core operand broadcast with $\sim 4\times$ less input operand modulation cost (op2-DAC, op2-mod). It also leverages signal-domain-wise locality by

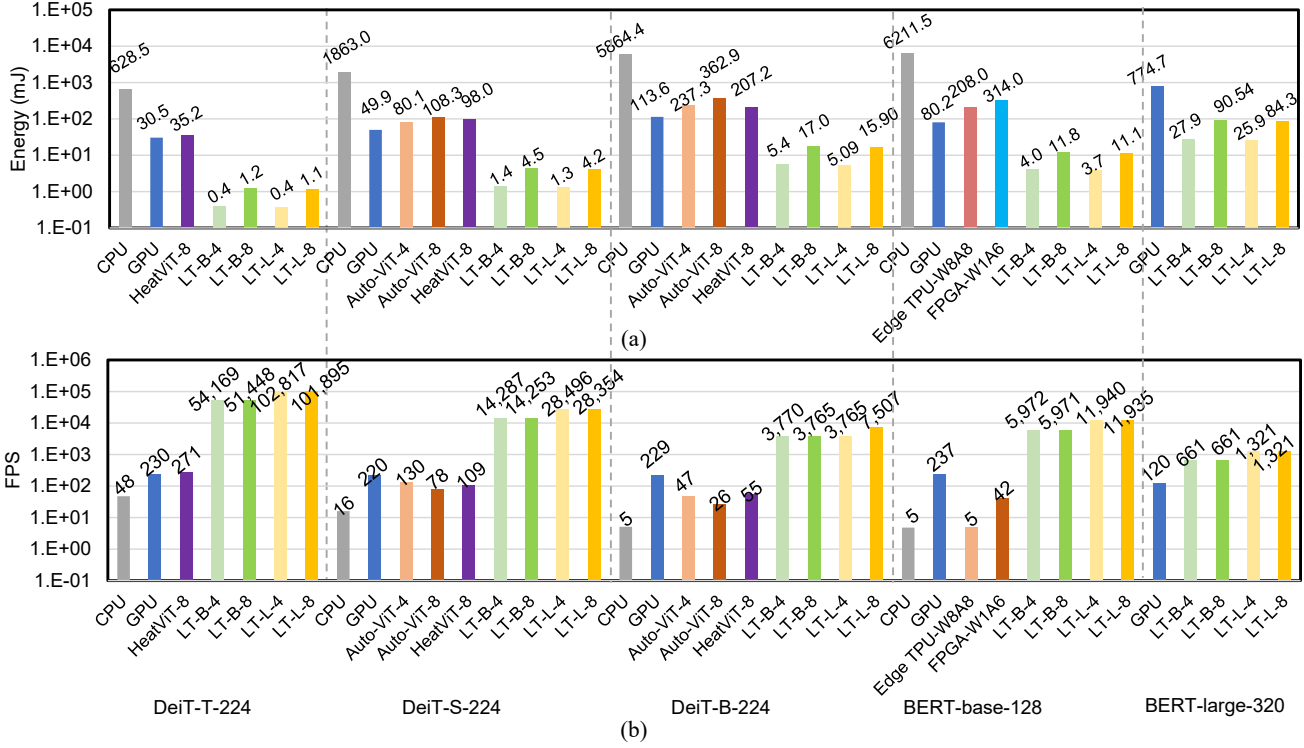


Fig. 13: Compare (a) energy consumption (mJ) and (b) frames-per-second (FPS) across different accelerator designs on various workloads (DeiT-series with ImageNet1K-224x224 and BERT-series with 128 and 320 sequence lengths) and two different bitwidth (4-bit and 8-bit).

applying photocurrent summation across PTCs in the same tile and achieving temporal partial summation before sending data to the ADC by time-integral, thus reducing ADC energy costs (ADC) cost by $\sim 6\times$. This time-integral technique is uniquely supported in our dynamically-operated DPTC design, as it can be only applied for accelerators supporting *output-stationary dataflow*.

Comparison on DeiT: Table V compares the base version of Lightning-Transformer (LT-B) with baselines on DeiT-T/B in 4-bit/8-bit precision. The results prove the superior performance of Lightning-Transformer in terms of energy, latency, and energy-delay product (EDP). Specifically, it outperforms the MZI array by over $8\times$, $675\times$, and $5000\times$ in terms of energy, latency, and EDP, respectively. Especially when scaling to high precision, the MZI baseline costs much more energy due to the exponentially increasing laser power. When compared to MRR bank, LT-B achieves energy, latency, and EDP savings of $2.6\times$, $12.8\times$, and $34.2\times$ on the 8-bit setting. On the 4-bit setting, LT-B achieves energy, latency, and EDP savings of $4\times$, $12.8\times$, and $51.7\times$ over the MRR bank. Even without architecture-level optimization in Sec. IV-C, LT-B still saves over $2\times$ energy compared to baselines, showing the efficiency of our DPTC design. Besides the architecture-level optimization, this superiority of Lightning-Transformer can be largely attributed to **1 High parallelism**. Our photonic tensor core performs ultra-parallel one-shot MM instead MVM, featuring significantly higher throughput and lower energy cost. **2 High efficiency**. DPTC features zero locking cost and

low laser power (low insertion loss) and exploits intra-core and inter-core operand sharing to reduce modulation overhead. **3 Natural full-range support** with no extra overhead. **4 Dynamic operand switch** at high speed (< 10 ps). The use of low-loss phase shifters in MZI array imposes a high latency term for *reconfiguring PTC* ($2 \mu\text{s}$) to the total latency when switching weight operand for tiled MM.

PTC-Level Takeaways

1 Exploring both spectral and spatial parallelism of optics brings significant performance. Besides spatial parallelism using multiple compact cores, a coherent broadband structure can further leverage multi-wavelength processing to fully unleash the spectral parallelism of optics, significantly increasing the computing parallelism and density.

2 Dynamic full-range operand encoding enables versatility. In contrast to prior weight-static designs with restricted applications, e.g., CNN with ReLU, a generic, highly-reprogrammable optical GEMM primitive, like our DPTC, can offer much wider applicability and higher efficiency to adapt to advanced and ever-evolving ML acceleration tasks.

3 Leveraging the natural broadcast capability of optics maximizes hardware sharing and energy efficiency. The crossbar topology largely amortizes the footprint cost of photonic devices via extensive hardware sharing while inducing minimum overhead due to the superior broadcasting and interconnecting capability of photonic waveguides.

Architecture-Level Takeaways

1 Combining photonic computing with photonic interconnect

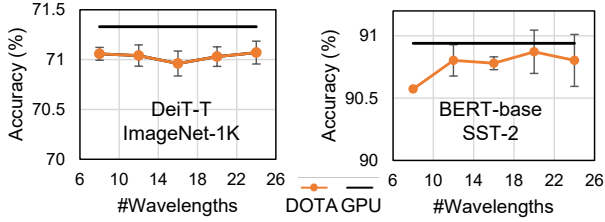


Fig. 14: Dispersion robustness evaluation with a range of wavelengths on 4-bit DeiT-T ImageNet1K and 8-bit BERT-base SST2 workload. The accuracy of quantized models running on GPU without any noise is shown. Input noise std. and phase noise std. are set to 0.03 and 2° .

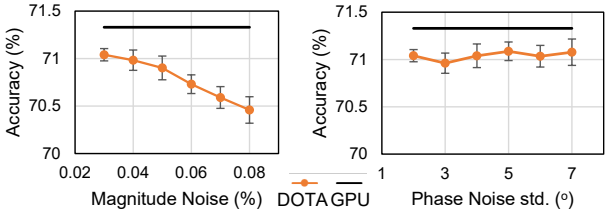


Fig. 15: Encoding magnitude and phase noise robustness evaluation with varying noise intensity on 4-bit DeiT-T ImageNet1K.

unleashes the power of optics. The integration of optical computing cores and cross-core optical interconnect allows for speed-of-light computing and communication with further operand sharing at the architecture level, simultaneously reducing the data movement and signal modulation cost.

② *Offloading more computing to the analog domain relaxes the A/D conversion bottleneck.* To reduce cross-domain signal conversion costs, it is critical to leverage the concept of *signal-domain-wise locality*. By offloading more computations within the analog domain, such as analog partial summation and temporal accumulation employed in Lightning-Transformer, the need for frequent A/D conversion is minimized, significantly alleviating power and latency bottlenecks from ADCs.

D. Compare to State-of-the-art Digital Accelerators

In Figure 13, we compare Lightning-Transformer to different hardware platforms to demonstrate our orders-of-magnitude performance improvements. Specifically, we make comparisons with (1) single Nvidia A100 GPU, (2) Intel Core i7-9750H CPU, (3) Cora Edge TPU [44], and (4) FPGA accelerators AutoViT [31], HEATViT [12]. For GPU and CPU, we use automatic mixed precision to run at least 100 inferences to calculate the averaged data with a warm-up period ahead. For the FPGA and Edge TPU, we use the results in the original paper. The results show that our Lightning-Transformer consistently achieves the lowest energy consumption, with over $300\times$, $6.6\times$, $18\times$, and $20\times$ reduction compared to CPU, GPU, Edge TPU, and other domain-specific Transformer accelerators. For throughput, Lightning-Transformer achieves the highest among all the platforms, even on the 4-tile LT-B system. We get 2 to 3 orders of magnitude lower energy-delay products for various benchmarks on our LT-L system.

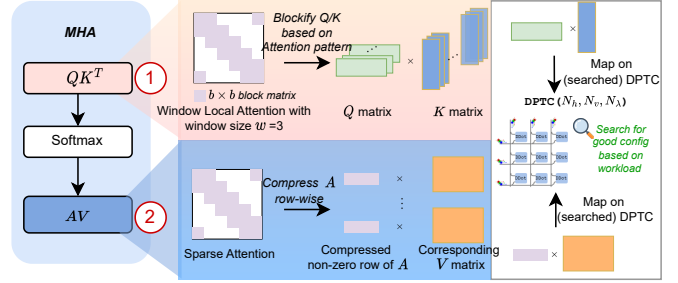


Fig. 16: Illustration of sparse Attention support on Lightning-Transformer using window-based local attention as an example. After blockification/compression, we transform the sparse computation to dense matrix/vector-matrix multiplication that DPTC can accelerate efficiently.

E. Accuracy & Robustness Analysis

We further show that Lightning-Transformer can realize digital-comparable accuracy with superior robustness against various non-ideality effects. We evaluate the accuracy of quantized models running on Lightning-Transformer by using the noisy analytic transformation in Eq. (9) during inference. As shown in Figure 14, compared to the accuracy of Transformers with equivalent bit-widths running on GPU, we maintain $<1\%$ accuracy loss ensured by our superior robustness. We also evaluate the robustness of our design against encoding the magnitude noise, phase noise, and dispersion effect. Figure 14 shows that our design is very robust to WDM dispersion even with more than 20 wavelengths, showing $<0.5\%$ accuracy drop. The reason is our design points of the directional coupler and phase shifter are at the local optima with minimized sensitivity to dispersion-induced perturbations, as discussed in Sec. III-C. Figure 15 shows that our design demonstrates high tolerance to random encoding magnitude and phase variations. On the DeiT-T ImagenetNet workload, the noise-induced accuracy degradation is within 0.5% with a small variance Noise-aware training is adopted in this paper, while more advanced noise-mitigation techniques [20], [56] can be applied to further boost the accuracy and robustness.

VI. DISCUSSION

A. Sparsity Support

The sparsity opportunity in Transformers mainly lies in three folds [12]. (1) **Redundancy of the attention head/token.** Attention head/token pruning directly removes unimportant heads or tokens entirely [57]. (2) **Redundancy of the token channels.** Token channel pruning removes some channels of the token embedding. (3) **Redundancy of the attention map.** A number of Transformers have proposed to explore sparsity inside the attention map so as to make the attention module more efficient. For example, BigBird [65] and BlockBERT [41] propose structured/block-wise sparse attention patterns, e.g., window- and global- attentions. The sparse attention features sparse computation in both QK^T and AV .

Our Lightning-Transformer can be easily extended to support the (1) and (2) opportunities, as they remove

head/token/channel entirely, resulting in regular dense GEMM. We can also support hardware-friendly structured/block-wise (3) sparse attention patterns by reformulating sparse computation into small chunked dense matrix-matrix/vector-matrix multiplication, as illustrated in Figure 16. To generate sparse attention $A = QK^T$ efficiently, we can blockify Q/K matrices based on structured sparse patterns and form groups of matrix-matrix multiplication. Take the block-wise window local attention [65] as an example. Assume the number of tokens is n , the window size is w and the block size is b . The token i will only attends to key matrix K with index $i - (w - 1)/2$ to $i + (w - 1)/2$. We can blockify the Q and K matrix based on the block size b , resulting in $\lceil n/b \rceil$ chunked Q and K matrices. Based on the window pattern, each chunked Q matrices will compute with only w chunked K matrices, still featuring dense matrix-matrix multiplication. To support AV , we can first compress the sparse attention A row by row. This approach generates dense A chunked matrices/vectors based on whether the sparsity is block-wise or not. Then, we can form vector-matrix/matrix-matrix multiplication between the compressed A with the corresponding rows of V . The above reformulated dense matrix-matrix/matrix-vector multiplication can be efficiently accelerated by DPTC. Moreover, we have the flexibility to explore heterogeneous DPTCs by having different/searched core sizes in Table II to better suit workloads with specific sparse patterns, avoiding low-utilization scenarios. For example, we can have a specific DPTC engine for vector-matrix multiplication by setting N_h to 1 to support vector-matrix multiplication featured by non-block-wise sparsity.

B. Large Language Model Support

Supporting large language models (LLMs) on our accelerator presents both challenges and opportunities, primarily because LLMs rely on parallel-unfriendly autoregressive Transformer models. These models generate tokens *one at a time* based on input and previous tokens, resulting in small-dimensional matrix multiplications with low operation intensity. This characteristic makes LLMs memory-bounded and underutilized the ultra-fast computing power offered by the photonic chips. Besides, they demand substantial on-chip memory for storing key and value tensors, also known as KV cache.

To adapt our accelerator for LLMs, several strategies can be considered. First, scaling on-chip memory or even building multi-chip systems are necessary to meet the memory/throughput demands of LLMs adequately. Secondly, we can batch multiple requests together to increase operation intensity, making better use of the accelerator’s computing capabilities. Additionally, we can trade excessive memory demands for cost-effective and rapid optical computation by recalculating Query (Q) and Key (K) values instead of fully caching them, as demonstrated in recent work [61]. Model compression techniques [57], [67] can also be applied to reduce memory usage by pruning unimportant tokens. Furthermore, optimizing our tiling algorithm, similar to the approach in [9], can help avoid materializing large full attention matrices, reducing

reliance on slow off-chip DRAM and improving overall efficiency.

By implementing these strategies and further tailoring our accelerator, we believe our work is poised to extend its capabilities to support Transformer-backed large language models (LLMs) in the future.

VII. CONCLUSION

The proliferation and increasing complexity of attention-based Transformers have spurred the need for specialized hardware accelerators. Photonic accelerators have shown promising efficiency and speed for CNN workloads, which require only weight-static linear operations. However, the current SoTA photonic accelerators face challenges in supporting Transformer with self-attention operations, primarily due to their inability to handle dynamic tensor multiplication and encode full-range operands. In this work, we introduce the first customized, high-performance, and energy-efficient photonic Transformer accelerator, *Lightening-Transformer*. We propose a novel photonic tensor core, a crossbar array of dynamic optical vector dot-product engines. This design overcomes the fundamental constraints of existing designs, enabling ultra-parallel matrix multiplication of two dynamic, full-range matrices. Our comprehensive evaluation shows that *Lightening-Transformer* achieves over $2.6\times$ energy and over $12\times$ latency reductions compared to prior photonic accelerators. Furthermore, it outperforms electronic Transformer accelerators with over $6\times$ energy reduction and 2 to 3 orders of magnitude lower energy-delay products with digital-comparable accuracy. Looking ahead, our work is poised to extend its capabilities to support Transformer-backed large language models (LLMs) in the future, further enhancing its applicability to cutting-edge machine learning tasks.

Our work highlights the potential of domain-specific photonic AI hardware for the efficient acceleration of advanced ML tasks. In the future, we anticipate significant advancements in optical computing technologies, enhancing their flexibility, applicability, and performance across a broader range of machine learning tasks. Ongoing research in innovative cross-layer co-design will push the boundaries of what photonic accelerators can achieve and result in the creation of highly efficient, low-latency next-generation electronic-photonic computing systems for increasingly complex ML workloads.

ACKNOWLEDGMENT

This work is supported in part by the Multidisciplinary University Research Initiative (MURI) program through the Air Force Office of Scientific Research (AFOSR) under contract #FA 9550-17-1-0071 and AFOSR project #FA9550-23-1-0452. We thank all anonymous HPCA reviewers for their insightful comments. Thanks to Zhixing Jiang and Shupeng Ning from the University of Texas at Austin for helpful suggestions and help during the artifact evaluation process.

APPENDIX

A. Abstract

Lightning-Transformer introduces the first customized photonic Transformer accelerator built upon a novel dynamically-operated photonic tensor core design.

The artifact contains three parts. The first is the noise-aware training/inference framework, enabling users to train/infer with the optical Transformer models built onto our photonic tensor core. We embed the analytic transformation of PTC in the forward path when computing matrix multiplication. We support quantization and non-ideality injection, including input phase/magnitude variation, WDM-induced dispersion, and a general systematic error term. The second part is the hardware simulator for our Lightning-Transformer accelerator. Our simulator simulates performance using behavior-level models. Our simulator can estimate the area and power of our Lightning-Transformer accelerator. Besides, it can predict the energy and latency when running Transformer workloads (DeiT/BERT) on Lightning-Transformer. The third part is profiling scripts for measuring the latency and power when running workloads on GPUs, such that you can compare the GPU performance with our light-empowered accelerators. We provide scripts for each part to validate our performance.

The minimal hardware requirement will be one CPU and one Nvidia GPU. The minimal software requirements will be CUDA and Python libraries such as PyTorch.

B. Artifact check-list (meta-information)

- **Program:** Python
- **Model:** Low-bit optical Transformer models.
- **Data set:** Publicly available Image classification dataset for DeiT evaluation.
 - ImageNet: <http://image-net.org/>: 160 GB.The dataset needs extra download and preparation, which is not included in our artifact, given its large size. Download and extract ImageNet using the [script](#).
- **Run-time environment:**
 - Ubuntu18 or above
 - Main software dependencies: Python, PyTorch, pytorch-image-models 0.3.2.
 - conda is required for package management.
- **Hardware:**
 - At least one Nvidia GPU for running inference of optical Transformer models with noise injection.
 - Training with optical Transformer models on ImageNet with quantization and modeled PTC behavior can be time-consuming and GPU-hungry. Hence, we provide a trained checkpoint for users to quickly validate accuracy under various non-idealities.
 - We tested on Nvidia A100 and A6000 GPUs.
- **Output:**
 - Classification accuracy of optical Transformer models on ImageNet.
 - Area and power report of our accelerator.
 - Latency and energy estimation of our accelerator for given workloads.
- **Experiments:** We prepare shell scripts for the following experiments.

- Inference of the 4-bit DeiT-T (checkpoint provided) on ImageNet dataset with various variations injected.
- Training of low-bit optical DeiT model (Optional).
- Area and power profiling for our accelerator Lightning-Transformer.
- Latency and energy estimation on our accelerator Lightning-Transformer for DeiT/BERT workloads.
- Workload profiling on GPU.

- **How much disk space required (approximately)?:** < 10GB without considering the ImageNet data.
- **How much time is needed to prepare workflow (approximately)?:** 1 hour.
- **How much time is needed to complete experiments (approximately)?:** 24 hours without training.
- **Publicly available?:** Yes.
 - <https://github.com/zhuhanqing/Lightning-Transformer>
 - <https://doi.org/10.6084/m9.figshare.24899037.v1>
- **Code licenses (if publicly available)?:** GNU GPL3.0
- **Data licenses (if publicly available)?:**
 - ImageNet. BSD 3-Clause License
- **Workflow framework used?:** PyTorch.
- **Archived (provide DOI)?**
<https://doi.org/10.6084/m9.figshare.24899037.v1>.

C. Description

1) *How to access:* The artifact is available at [Github repo](#). Check the `readme.md` for the detailed descriptions. Basically, it has three parts

- `software_model`: algorithm codes for training/running optical Transformers on our photonic accelerator, which models the matrix multiplication with the analytical transformation of our unique photonic tensor core. Quantization and noise injection are supported.
- `hardware_simulator`: our behavior-level simulator.
- `profile`: profiling scripts for latency and power measurement on GPU.

2) *Hardware dependencies:* Python scripts are deployed on the server with a dedicated Nvidia GPU (CUDA>11.x). These Python files are implemented for command run on the server. Our implementations have been evaluated with Nvidia A100 and A6000.

- One power GPU machine is desired.
- The training of DeiT-T on 4 A100 takes ~ 2 days on ImageNet due to the quantization operations and dedicated simulation of GEMM workloads on our photonic tensor core.
- Hence, we provide a checkpoint that can be evaluated for accuracy validation. It takes 30 mins to obtain the inference accuracy on a single A100, with variation injection and quantization.

3) *Software dependencies:* The artifact is implemented in Python and requires several packages, such as PyTorch and Timm. The detailed install step can follow the `readme.md` in the artifact. We use `conda` to manage packages.

4) *Data sets:* Image classification datasets ImageNet. Download and extract ImageNet following [facebookresearch/deit](#). This [script](#) provides a step-by-step instruction.

5) *Models*: In this artifact, we provide the codes for the optical DeiT model with quantization and various non-idealities injection. We support injecting input magnitude encoding variation, input phase encoding variation, output variation, and WDM-induced dispersion, as discussed in Sec. III-C.

D. Installation

You can first download the repo to the local machine. Then, enter the folder and install the required packages following the instructions in the README.md.

E. Experiment workflow

We provide multiple examples to run our artifact to reproduce the main results of our papers.

1) *Inference with the optical Transformer model*: Enter the `software_model` directory for this part. We provide a script (`evaluate_quant_transformer_scan_noise.sh`) to construct and infer an optical DeiT model with sweeping noises such that you can obtain the results in Fig. 14 and Fig. 15

Check the `./software_model/readme.md` for more details if you want to customize experiments, e.g., launching training jobs or inference with a different noise value.

2) *Area and power profiling of our accelerator*: Enter the `hardware_simulator` directory.

- Run `entry_area_power_profile.py` to obtain area and power report.
- Results is dumped out in CSV format. It should deliver the same results as in Figure 7 and Figure 8.

We provide a batch run script such that you can profile the area and power of all `Lightning-Transformer` variants. See `./scripts/area_power_all.sh` All results will be in `./results/area_power_all/`.

3) *Latency and energy estimation of running workloads in our accelerator*:

- Run `./scripts/energy_latency_onns_deit.sh` to compare with photonic baselines.
- Run `./scripts/energy_latency_all.sh` to obtain energy and latency of our accelerator on various DeiT/BERT models.

4) *Profile latency and energy cost of inference on GPU*: Enter the `profile` directory and follow the `readme.md` to measure the latency and energy cost for DeiT/BERT models.

F. Evaluation and expected results

The expected results should match what we have in the experimental session and demonstrate the orders-of-magnitude better energy efficiency of our `Lightning-Transformer` accelerator over baselines.

G. Experiment customization

This framework provides various customization.

- Inference of optical DeiT model with different noise levels.
- Estimate energy and latency of different DeiT/BERT models. We support DeiT-T/S/B and BERT-B/L, where you can adjust the sequence lengths for BERT.

H. Methodology

Submission, reviewing and badging methodology:

- <https://www.acm.org/publications/policies/artifact-review-badging>
- <http://cTuning.org/ae/submission-20201122.html>
- <http://cTuning.org/ae/reviewing-20201122.html>
- <https://sc21.supercomputing.org/submit/reproducibility-initiative/ad-ae-appendix-process-badges/index.html#section16>

REFERENCES

- [1] S. Affi, F. Sunny, M. Nikdast, and S. Pasricha, "Tron: Transformer neural network acceleration with non-coherent silicon photonics," *arXiv preprint arXiv:2303.12914*, 2023.
- [2] S. Akiyama, T. Baba, M. Imai, T. Akagawa, M. Takahashi, N. Hirayama, H. Takahashi, Y. Noguchi, H. Okayama, T. Horikawa *et al.*, "12.5-gb/s operation with 0.29- ν cm ν π I using silicon mach-zehnder modulator based on forward-biased pin diode," *Optics express*, vol. 20, no. 3, pp. 2911–2923, 2012.
- [3] M. G. Anderson, S.-Y. Ma, T. Wang, L. G. Wright, and P. L. McMahon, "Optical transformers," *arXiv preprint arXiv:2302.10360*, 2023.
- [4] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
- [5] T. 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*, vol. 33, pp. 1877–1901, 2020.
- [6] —, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [7] P. Caragiulo, O. E. Mattia, A. Arbabian, and B. Murmann, "A compact 14 gs/s 8-bit switched-capacitor dac in 16 nm finfet cmos," in *2020 IEEE Symposium on VLSI Circuits*. IEEE, 2020, pp. 1–2.
- [8] N. 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.
- [9] T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré, "Flashattention: Fast and memory-efficient exact attention with io-awareness," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 344–16 359, 2022.
- [10] C. Demirkiran, F. Eris, G. Wang, J. Elmhurst, N. Moore, N. C. Harris, A. Basumallik, V. J. Reddi, A. Joshi, and D. Bunandar, "An electro-photonic system for accelerating deep neural networks," *arXiv preprint arXiv:2109.01126*, 2021.
- [11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [12] P. Dong, M. Sun, A. Lu, Y. Xie, K. Liu, Z. Kong, X. Meng, Z. Li, X. Lin, Z. Fang *et al.*, "Heatvit: Hardware-efficient adaptive token pruning for vision transformers," in *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2023, pp. 442–455.
- [13] P. Dong, W. Qian, H. Liang, R. Shafiqi, D. Feng, G. Li, J. E. Cunningham, A. V. Krishnamoorthy, and M. Asghari, "Thermally tunable silicon racetrack resonators with ultralow tuning power," *Optics express*, vol. 18, no. 19, pp. 20 298–20 304, 2010.
- [14] A. 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. [Online]. Available: <https://openreview.net/forum?id=YicbFdNTTy>
- [15] S. K. Esser, J. L. McKinstry, D. Bablani, R. Appuswamy, and D. S. Modha, "Learned step size quantization," in *International Conference on Learning Representations*, 2020.
- [16] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja *et al.*, "Parallel convolution processing using an integrated photonic tensor core," *Nature*, 2021.
- [17] C. 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*, vol. 9, no. 12, pp. 3906–3916, 2022.
- [18] J. Gu, C. Feng, H. Zhu, R. T. Chen, and D. Z. Pan, "Light in ai: Toward efficient neurocomputing with optical neural networks—a tutorial," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 6, pp. 2581–2585, 2022.
- [19] J. Gu, Z. Zhao, C. Feng *et al.*, "Towards area-efficient optical neural networks: an FFT-based architecture," in *Proc. ASPDAC*, 2020.
- [20] J. Gu, Z. Zhao, C. Feng, H. Zhu, R. T. Chen, and D. Z. Pan, "ROQ: A noise-aware quantization scheme towards robust optical neural networks with low-bit controls," in *Proc. DATE*, 2020.
- [21] H. Guo, L. Peng, J. Zhang, Q. Chen, and T. D. LeCompte, "Att: A fault-tolerant rram accelerator for attention-based neural networks," in *2020 IEEE 38th International Conference on Computer Design (ICCD)*. IEEE, 2020, pp. 213–221.
- [22] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.
- [23] Z. Huang, C. Li, D. Liang, K. Yu, C. Santori, M. Fiorentino, W. Sorin, S. Palermo, and R. G. Beausoleil, "25 gbps low-voltage waveguide si–ge avalanche photodiode," *Optica*, vol. 3, no. 8, pp. 793–798, 2016.
- [24] S. V. Kartalopoulos, "Introduction to dwdm technology," (*No Title*), 1999.
- [25] J. 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.
- [26] Y. Kim, H. Kim, D. Ahn, and J.-J. Kim, "Input-splitting of large neural networks for power-efficient accelerator with resistive crossbar memory array," in *Proceedings of the International Symposium on Low Power Electronics and Design*, 2018, pp. 1–6.
- [27] F. Koyama and K. Iga, "Frequency chirping in external modulators," *Journal of Lightwave Technology*, vol. 6, no. 1, pp. 87–93, 1988.
- [28] M. Li, Z. Yu, Y. Zhang, Y. Fu, and Y. Lin, "O-has: Optical hardware accelerator search for boosting both acceleration performance and development speed," in *2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*. IEEE, 2021, pp. 1–9.
- [29] S. Li, H. Yang, C. W. Wong, V. J. Sorger, and P. Gupta, "Photofourier: A photonic joint transform correlator-based neural network accelerator," in *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2023, pp. 15–28.
- [30] —, "Photofourier: A photonic joint transform correlator-based neural network accelerator," in *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2023, pp. 15–28.
- [31] Z. Lit, M. Sun, A. Lu, H. Ma, G. Yuan, Y. Xie, H. Tang, Y. Li, M. Leiser, Z. Wang *et al.*, "Auto-vit-acc: An fpga-aware automatic acceleration framework for vision transformer with mixed-scheme quantization," in *2022 32nd International Conference on Field-Programmable Logic and Applications (FPL)*. IEEE, 2022, pp. 109–116.
- [32] J. Liu, M. Hassanpourghadi, and M. S.-W. Chen, "A 10gs/s 8b 25fj/cs 2850um 2 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. IEEE, 2022, pp. 160–162.
- [33] L. Lu, Y. Jin, H. Bi, Z. Luo, P. Li, T. Wang, and Y. Liang, "Sanger: A co-design framework for enabling sparse attention using reconfigurable architecture," in *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*, 2021, pp. 977–991.
- [34] I. Lumerical Solutions, "Lumerical interconnect," <https://www.lumerical.com/products/interconnect>, 2003.
- [35] G. Mourgas-Alexandris, M. Moralis-Pegios, A. Tsakyridis, S. Simos, G. Dabos, A. Totovic, N. Passalis, M. Kirtas, T. Rutirawut, F. Gardes *et al.*, "Noise-resilient and high-speed deep learning with coherent silicon photonics," *Nature Communications*, vol. 13, no. 1, p. 5572, 2022.
- [36] D. P. Nair and M. Ménard, "A compact low-loss broadband polarization independent silicon 50/50 splitter," *IEEE Photonics Journal*, vol. 13, no. 4, pp. 1–7, 2021.
- [37] M. O'Connor, N. Chatterjee, D. Lee, J. Wilson, A. Agrawal, S. W. Keckler, and W. J. Dally, "Fine-grained dram: Energy-efficient dram for extreme bandwidth systems," in *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture*, 2017, pp. 41–54.
- [38] OpenAI, "Gpt-4 technical report," 2023.
- [39] P. Pintus, M. Hofbauer, C. L. Manganello, M. Fournier, S. Gundavarapu, O. Lemonnier, F. Gambini, L. Adelmini, C. Meinhart, C. Kopp *et al.*, "Pwm-driven thermally tunable silicon microring resonators: design, fabrication, and characterization," *Laser & Photonics Reviews*, vol. 13, no. 9, p. 1800275, 2019.
- [40] H. Prashanth and M. Rao, "Somalib: Library of exact and approximate activation functions for hardware-efficient neural network accelerators," in *2022 IEEE 40th International Conference on Computer Design (ICCD)*. IEEE, 2022, pp. 746–753.
- [41] J. Qiu, H. Ma, O. Levy, W.-t. Yih, S. Wang, and J. Tang, "Blockwise self-attention for long document understanding," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2555–2565.
- [42] N. Quack, A. Y. Takabayashi, H. Sattari, P. Edinger, G. Jo, S. J. Bleiker, C. Errando-Herranz, K. B. Gylfason, F. Niklaus, U. Khan *et al.*,

- “Integrated silicon photonic mems,” *Microsystems & Nanoengineering*, vol. 9, no. 1, p. 27, 2023.
- [43] M. 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.
- [44] B. C. Reidy, M. Mohammadi, M. E. Elbtity, and R. Zand, “Efficient deployment of transformer models on edge tpu accelerators: A real system evaluation,” in *Architecture and System Support for Transformer Models (ASSYST@ ISCA 2023)*, 2023.
- [45] A. Shafaei, Y. Wang, X. Lin, and M. Pedram, “Fincacti: Architectural analysis and modeling of caches with deeply-scaled finfet devices,” in *2014 IEEE Computer Society Annual Symposium on VLSI*. IEEE, 2014, pp. 290–295.
- [46] B. J. Shastri, A. N. Tait, T. F. de Lima, W. H. P. Pernice, H. Bhaskaran, C. D. Wright, and P. R. Prucnal, “Photonics for artificial intelligence and neuromorphic computing,” *Nature Photonics*, 2021.
- [47] Y. Shen, N. C. Harris, S. Skirlo *et al.*, “Deep learning with coherent nanophotonic circuits,” *Nature Photonics*, 2017.
- [48] K. Shiflett, A. Karanth, R. Bunescu, and A. Louri, “Albireo: Energy-efficient acceleration of convolutional neural networks via silicon photonics,” in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2021, pp. 860–873.
- [49] M. Streshinsky, R. Ding, Y. Liu, A. Novack, Y. Yang, Y. Ma, X. Tu, E. K. S. Chee, A. E.-J. Lim, P. G.-Q. Lo *et al.*, “Low power 50 gb/s silicon traveling wave mach-zehnder modulator near 1300 nm,” *Optics express*, vol. 21, no. 25, pp. 30 350–30 357, 2013.
- [50] M. Sun, H. Ma, G. Kang, Y. Jiang, T. Chen, X. Ma, Z. Wang, and Y. Wang, “Vaqf: fully automatic software-hardware co-design framework for low-bit vision transformer,” *arXiv preprint arXiv:2201.06618*, 2022.
- [51] F. Sunny, A. Mirza, M. Nikdast, and S. Pasricha, “Crosslight: A cross-layer optimized silicon photonic neural network accelerator,” in *2021 58th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 2021, pp. 1069–1074.
- [52] A. N. Tait, T. F. de Lima, E. Zhou *et al.*, “Neuromorphic photonic networks using silicon photonic weight banks,” *Sci. Rep.*, 2017.
- [53] E. Timurdogan, C. M. Sorace-Agaskar, J. Sun, E. S. Hosseini, A. Biberman, and M. R. Watts, “An ultralow power athermal silicon modulator,” *Nat Commun*, 2014.
- [54] H. 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. 10 347–10 357.
- [55] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [56] W. Wan, R. Kubendran, C. Schaefer, S. B. Eryilmaz, W. Zhang, D. Wu, S. Deiss, P. Raina, H. Qian, B. Gao *et al.*, “A compute-in-memory chip based on resistive random-access memory,” *Nature*, vol. 608, no. 7923, pp. 504–512, 2022.
- [57] H. Wang, Z. Zhang, and S. Han, “Spatten: Efficient sparse attention architecture with cascade token and head pruning,” in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2021, pp. 97–110.
- [58] H. Wang, R. Zhang, Q. Kan, D. Lu, W. Wang, and L. Zhao, “High-power wide-bandwidth 1.55- μ m directly modulated dfb lasers for free space optical communications,” in *CLEO: Science and Innovations*. Optica Publishing Group, 2019, pp. JTu2A–72.
- [59] M. Wang, S. Lu, D. Zhu, J. Lin, and Z. Wang, “A high-speed and low-complexity architecture for softmax function in deep learning,” in *2018 IEEE asia pacific conference on circuits and systems (APCCAS)*. IEEE, 2018, pp. 223–226.
- [60] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis, “Efficient streaming language models with attention sinks,” *arXiv*, 2023.
- [61] —, “Efficient streaming language models with attention sinks,” *arXiv preprint arXiv:2309.17453*, 2023.
- [62] X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, R. Morandotti *et al.*, “11 tops photonic convolutional accelerator for optical neural networks,” *Nature*, vol. 589, no. 7840, pp. 44–51, 2021.
- [63] C. Ye and D. Dai, “Ultra-compact broadband 2×2 3 db power splitter using a subwavelength-grating-assisted asymmetric directional coupler,” *Journal of Lightwave Technology*, vol. 38, no. 8, pp. 2370–2375, 2020.
- [64] H. You, Z. Sun, H. Shi, Z. Yu, Y. Zhao, Y. Zhang, C. Li, B. Li, and Y. Lin, “Vitcod: Vision transformer acceleration via dedicated algorithm and accelerator co-design,” in *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2023, pp. 273–286.
- [65] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang *et al.*, “Big bird: Transformers for longer sequences,” *Advances in neural information processing systems*, vol. 33, pp. 17 283–17 297, 2020.
- [66] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia *et al.*, “Glm-130b: An open bilingual pre-trained model,” *arXiv preprint arXiv:2210.02414*, 2022.
- [67] Z. Zhang, Y. Sheng, T. Zhou, T. Chen, L. Zheng, R. Cai, Z. Song, Y. Tian, C. Ré, C. Barrett *et al.*, “H₂o: Heavy-hitter oracle for efficient generative inference of large language models,” *arXiv preprint arXiv:2306.14048*, 2023.
- [68] S. Zheng, J. Zhang, and W. Zhang, “Scalable optical neural networks based on temporal computing,” *arXiv preprint arXiv:2303.01287*, 2023.
- [69] M. Zhou, W. Xu, J. Kang, and T. Rosing, “Transpim: A memory-based acceleration via software-hardware co-design for transformer,” in *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2022, pp. 1071–1085.
- [70] H. Zhu, J. Zou, H. Zhang, Y. Shi, S. Luo, N. Wang, H. Cai, L. Wan, B. Wang, X. Jiang *et al.*, “Space-efficient optical computing with an integrated chip diffractive neural network,” *Nature communications*, vol. 13, no. 1, p. 1044, 2022.