P-DAC: Power-Efficient Photonic Accelerators for LLM Inference

Wen-Tse Chang*, Chun-Feng Wu*, Yun-Chen Lo†
*Department of Computer Science, National Yang Ming Chiao Tung University, Taiwan
†Department of Computer Science, Harvard University, USA
Corresponding Author: Chun-Feng Wu
E-mail: stupidrara123.cs12@nycu.edu.tw, cfwu417@cs.nycu.edu.tw, yunchenlo@seas.harvard.edu

***Abstract*—As traditional electronic hardware encounters the limitations of Moore's Law, optical computing is emerging as a promising alternative, delivering high data transmission rates, especially beneficial for big data and AI applications. Photonic accelerators, such as the Lightening-Transformer, utilize optical analog signals to accelerate Transformer-based models, achieving exceptional speed and low energy consumption. However, controlling modern optical intensity modulators (e.g., Mach-Zehnder Modulators) requires using electrical analog signals (e.g., voltage values) to adjust the optical signal intensity for realizing optical-based vector inner product calculations. Managing this modulation consumes significant power, as it involves selecting optimal electrical values through an electrical controller and converting digital signals to analog using digital-to-analog converters (DACs). In this work, we introduce P-DAC, a solution designed to reduce DAC power consumption, significantly enhancing the energy efficiency of optical accelerators for Transformer models.**

***Index Terms*—Photonic Accelerators, LLM Inference, Mach-Zehnder Modulators, DAC, Photonic DAC, Energy Efficient**

I. Introduction

Transformer-based large language models (LLMs) have gained significant attention in recent years, requiring substantial computing power and high data transmission bandwidth due to their scale and computational complexity [39], [14], [11], [29]. While modern systems primarily rely on electrical-based computation and interconnection, these devices are approaching the physical and economic limits of Moore's Law [35] and Dennard scaling [8], leading to increased power dissipation and slower performance gains. To address these challenges, integrated photonic accelerators have emerged as a promising alternative, offering ultra-high speed, parallelism, and energy efficiency [31], [20]. Unlike transistor-based chips, photonic devices leverage optical signals to achieve superior theoretical performance [9], [32], with the potential to significantly surpass electronic accelerators as technology advances. Photonic computing and interconnection could revolutionize LLM inference, making them a critical candidate for replacing traditional electronic systems [30], [45].

Photonic interconnections outperform electrical counterparts by offering higher bandwidth, lower latency, and faster, long-distance data transmission without signal degradation [36]. This technology has been widely applied to interconnect processing elements (PEs) in computing systems [16]. SPRINT [22] integrates photonic interconnections with Convolutional neural network (CNNs), leveraging broadcasting to share data across waveguides. PEs perform computations, write results to other waveguides, and send processed data to the global buffer. Similarly, SPACX [23] supports DNN inference by broadcasting data from a global waveguide to local waveguides, facilitating computation in PEs. Processed data is then transferred back to the global waveguide and buffer. Additionally, CAMON [41], a silicon photonic chiplet for manycore processors, reduces communication bottlenecks and enhances energy efficiency by optimizing cache and memory management in optical communication systems, especially in large-scale architectures.

In addition to interconnection, several research teams [21], [44] focus on accelerating CNN tasks, utilizing various photonic tensor core (PTC) architectures, such as the Mach-Zehnder interferometer (MZI) array [33]. The MZI requires singular value decomposition (SVD) and phase decomposition for operand mapping and is capable of performing arbitrary 2-D unitary matrix operations. However, it requires CPU to conduct task mapping, which is time-consuming. For example, mapping a 12×12 matrix takes approximately 1.5 ms for conducting SVD and phase decomposition. To substantially improve computation parallelism, recent researchers have focused on realizing the photonic dot productions in the analog space. For example, Albireo [34] and Lightening-Transformer [45] work on integrating analog-based photonic dot productions with CNN and Transformer, respectively. The key idea is to convert electrical digital signals to photonic analog signals by modulating the intensity and frequency and run dot productions via interleaving photonic signals.

Based on our investigation, analog photonic accelerators, such as Albireo and Lightening-Transformer, rely heavily on Mach-Zehnder Modulators (MZMs) for optical signal modulation. This process requires controllers to compute input voltage values and Digital-to-Analog Converters (DACs) to generate the corresponding voltages for driving the MZMs. However, the power consumption of DACs is significant and increases with higher bit precision, creating a bottleneck for energy-efficient optical computing [10], [43], [12], [25], [24], [3], [40], [38]. To address this, we propose the Photonic-DAC (P-DAC), which eliminates the need for traditional DACs by using approximation to convert optical digital signals into analog signals. Since the P-DAC operates with optical digital signals, it can leverage Wavelength Division Multiplexing (WDM) [13] to enhance the data rate by combining multiple wavelengths into a single waveguide. The P-DAC is particularly well-suited for Large Language Models (LLMs), whose inherent tolerance for numerical errors aligns with the P-DAC's design, often improving performance in such applications.

The key contributions of this work include:

- Proposing the P-DAC, a power-efficient alternative to traditional DAC-based systems, designed for optical accelerators.
- Demonstrating the feasibility of the P-DAC through mathematical derivation and experimental validation.
- Achieving significant energy savings in practical workloads, such as BERT [6] and DeiT [37], with up to 35.4% reduction in power consumption for 8-bit data sizes.

II. Background, Observation, and Motivation

A. Background

1) Large Language Models (LLM) Inference: Large language models (LLMs) are advanced neural networks built on the transformer architecture, which integrates self-attention mechanisms with fully-connected layers to process and understand complex relationships in text. Self-attention enables the model to grasp global context by dynamically incorporating prior information when generating outputs. To produce the next output token, the model computes Query (Q), Key (K), and Value (V) vectors by multiplying the concatenated input and current output tokens with pre-trained weight matrices. Attention scores, derived by comparing Q with K, are used to weight

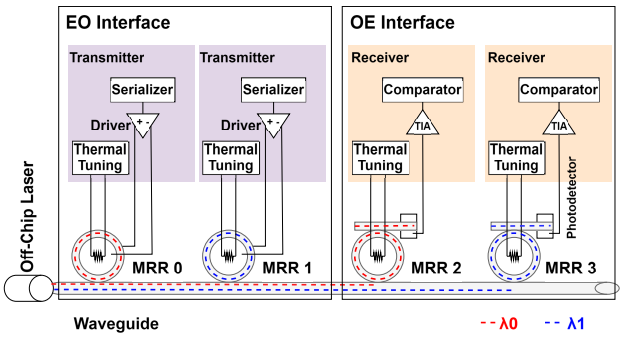


Fig. 1. WDM Technique

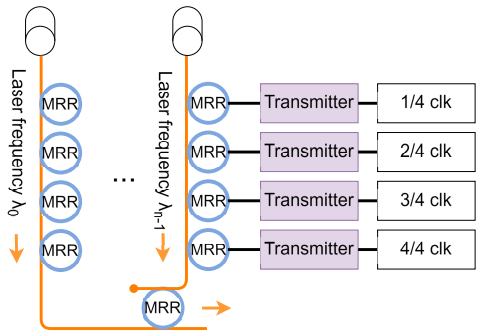


Fig. 2. A 4-bit EO Interface

the V vectors, capturing contextual relationships between tokens. During inference, the KV cache stores precomputed K and V vectors, allowing the model to reuse them efficiently for subsequent tokens without redundant calculations [18], [17], [11].

Running LLM inference demands substantial computing power and high data transmission bandwidth due to the vast scale of the model and the complexity of its computational processes. Large language models consist of billions (e.g., GPT-3), sometimes trillions (e.g., GPT-4 [28]), of parameters, necessitating extensive matrix multiplications for each token processed, particularly during the self-attention and fully-connected layers. The self-attention mechanism involves calculating Query, Key, and Value vectors for every token, performing dot products, and weighting operations, which are computationally expensive, especially for long sequences. Additionally, accessing the KV cache requires high data transmission bandwidth to efficiently retrieve and update cached values as new tokens are processed [17], [11]. Combined, these factors make LLM inference resource-intensive in terms of both computing and data transmission, posing challenges for optimization on modern hardware architectures.

2) Photonic Transmissions: Photonic (or optical) transmissions refer to data transmission using optical signals, with optical waveguides serving as the medium. They effectively overcome the limitations of electrical transmissions, offering high transmission rates and low delays, especially over long distances [42]. However, photonic systems require more complex components and incur additional costs due to the need for Electrical-to-Optical (E/O) and Optical-to-Electrical (O/E) conversions [19], [26], [5]. To achieve high transmission bandwidth, the Wavelength Division Multiplexing (WDM) technique is essential. WDM allows multiple data streams to be transmitted simultaneously over the same optical waveguide by using different wavelengths of light, thereby enabling photonic transmissions to offer significantly higher bandwidth — often one or even two orders of magnitude more than electrical transmissions [7], [1].

To better explain the functionality behind WDM technique, we illustrate a toy example in Figure 1. An MRR [4], [27] functions as a multiplexer (mux) or demux for WDM technique. An MRR filters and selects specific wavelengths by resonating at frequencies influenced by its structure, with precise tuning achieved through temperature adjustments. When the harmonic wavelength of the MRR matches an integer multiple of the wavelength in the waveguide, the light can be captured by the MRR. For example, as shown in figure 1, given the same laser source, MRR 0 and MRR 1 program optical signals to λ_1 and λ_2, while MRR 2 and MRR 3 receive optical signals as receivers to receive λ_1 and λ_2, respectively.

An EO interface typically includes transmitters that control the resonance frequency of each MRR through thermal tuning. CAMON [41] introduces a multi-bit EO interface that encodes n bits of data per laser frequency within a single clock cycle. Figure 2 illustrates a 4-bit EO interface, converting 4-bit electrical data to optical data. The clock cycle is divided into four intervals, with the transmitter modulating the MRR to write data at specific intervals (from $1/4$ to $4/4$ clk). This approach allows each laser frequency to store 4 bits of data within one cycle. On the OE interface side, each receiver includes a photodetector (PD) to convert incoming optical signals into electrical signals by generating current when photons interact with its sensitive material, typically via the photoelectric effect. A transimpedance amplifier (TIA) then amplifies the weak current from the PD into a usable voltage signal. Its output voltage can be expressed as follows:

$$V_{out} = R_f \times I_{in} \tag{1}$$

where R_f is the feedback resistor of the TIA.

3) Photonic Accelerators – Lightening Transformer: Among photonic accelerators, Lightening-Transformer [45] demonstrates the highest known performance in accelerating transformer-based applications, where we illustrate its architecture in Figure 3. Lightening-Transformer leverages the properties of optical analog signals and introduces two main optimizations specifically for the Transformer model: *1) it supports full-range inputs and outputs.* Since Transformer activations are not restricted to non-negative integers, Lightening-Transformer leverages the wave properties of light by treating optical field intensity as a numerical value and using the phase of light to determine positive and negative values, thus achieving full-range support. *2) It enables dynamic matrix multiplication.* "Dynamic" means that operands are generated in real time. Since the Transformer's Q, K, and V vectors are dynamically generated, mapping and device programming would cause significant system stalls on the photonic tensor core if they are much slower than the computation speed. Lightening-Transformer can rapidly compute dynamic input matrix multiplications, addressing this issue.

Lightening-Transformer works on the optical field which indicates the distribution and propagation of light waves in space, encompassing both amplitude and phase information. The amplitude determines the intensity of the light, while the phase determines the positions of the wave's peaks and troughs. Phase differences can cause interference, resulting in the reinforcement or cancellation of light waves. The architecture relies on three critical components: the Mach-Zehnder Modulator (MZM), Phase Shifter (PS), and Directional Coupler (DC). We illustrate all three structures in Figure 3.

- **Mach-Zehnder Modulator (MZM):** By adjusting the path length difference in the split light paths, the recombined light can achieve either constructive or destructive interference [15]. Leveraging these principles, the MZM can modulate both the phase and intensity of the optical field. In analog signal processing, phase modulation enables the MZM to achieve full-range encoding (including positive and negative values), as shown in the following equation:

$$E_{out} = E_{in} cos\phi \tag{2}$$

where ϕ is the phase shift of light. Practically, the MZM is regulated by two input voltages, V_1 and V_2, to adjust the output optical field, the equation is as follows:

$$E_{out} = \frac{E_{in}}{2}\left((1+k)e^{j\frac{\pi V_1}{2V_\pi}} + (1-k)e^{j\frac{\pi V_2}{2V_\pi}}\right) \tag{3}$$

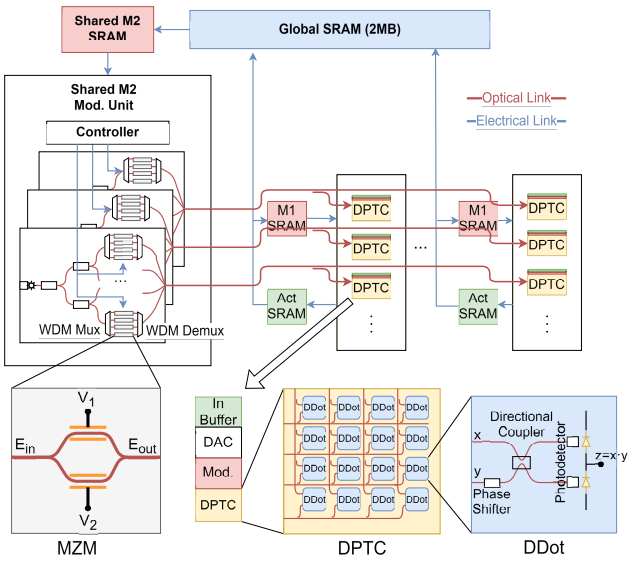


Fig. 3. Lightening-Transformer

where factor k represents the imbalance of splitting. Since the MZM can freely modulate light intensity, it serves as a key component for integrating computation into optical analog signals.

- **Phase Shifter (PS):** PS can change the phase of the light. The following equation represents its transformation formula:

$$x' = e^{j\phi}x \tag{4}$$

where ϕ is the phase shift of light, and x is the light signal.

- **Directional Coupler (DC):** DC can couples light of the same wavelength between two waveguides, transferring energy between them. The device comprises two closely positioned waveguides that enable energy transfer between them. Its transfer matrix of a 2-by-2 DC can be expressed as follows:

$$\begin{pmatrix} t & \sqrt{1-t^2}j \\ \sqrt{1-t^2}j & t \end{pmatrix} \tag{5}$$

where t is the transmission coefficient.

One of the key innovations of the Lightening-Transformer is its ability to compute dot products in the analog domain using a Dynamically-Operated Full-range Dot-Product Unit (DDot). DDot exploits the physical properties of light to efficiently compute the dot product $x \cdot y$. The operation follows the equation:

$$x \cdot y = \sum^{i}(x_i + y_i)^2 - \sum^{i}(x_i - y_i)^2 \tag{6}$$

In other words, by leveraging the physical properties of light to obtain

$$\sum^{i}(x_i + y_i)^2 \quad \text{and} \quad \sum^{i}(x_i - y_i)^2$$

and then taking their difference, we can derive $x \cdot y$. However, before attaining $\sum^{i}(x_i + y_i)^2$ and $\sum^{i}(x_i - y_i)^2$ we first need to find

$$x_i + y_i \quad \text{and} \quad x_i - y_i$$

DDot applies *WDM* technique and assigns each pairs x_i and y_i to the same wavelength. Thus, the light field intensities corresponding to wavelength i are x_i and y_i. The values $x_i + y_i$ and $x_i - y_i$ can be obtained by using a *phase shifter (PS)* and *directional coupler (DC)*:

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

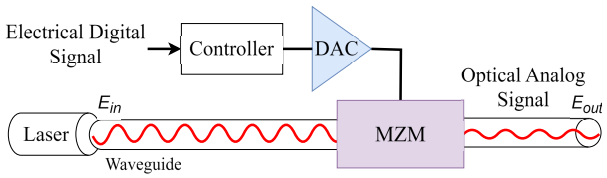


Fig. 4. MZM and DAC

where the leftmost matrix is the transformation matrix of the 50:50 DC, and the second matrix is the transformation matrix of the $-90°$ PS applied to the y-signal. After obtaining the two physical quantities $x_i + y_i$ and $x_i - y_i$, we apply the definition of light intensity:

$$I \propto \frac{1}{2}|E|^2$$

where I is the light intensity, E is the electric field amplitude. Since the photodetector can detect light intensity resulting from the superposition of multiple optical frequencies, we obtain the current by detecting both waveguides:

$$\sum^{i}(x_i + y_i)^2 \quad \text{and} \quad \sum^{i}(x_i - y_i)^2.$$

Finally, using Equation (6) enables us to attain the current intensity $x \cdot y$, and thus finish an inner product by DDot.

Combining DDot with WDM allows multiple wavelengths to utilize the same DDot unit, thereby enhancing computing density and parallelism. The PS and DC components of DDot are fully passive and fixed, which means there is no extra energy consumption because no need for external control and no issues with thermal crosstalk.

B. Observation on Lightening-Transformer's High Power Consumption & Motivation

To achieve high computational parallelism, Lightening-Transformer employs numerous MZMs to convert large volumes of data into optical analog signals simultaneously. Using WDM, these signals are transmitted to DDots for large-scale matrix multiplications. Activating many MZMs requires an equally large number of DACs to regulate their input voltages, as illustrated in Figure 4. However, driving these DACs can lead to significant power consumption. To validate this, we conducted an experimental analysis profiling the Lightening-Transformer's power usage. Results in Figure 5(a) and Figure 5(b) reveal that DACs contribute substantially to overall power consumption: 4-bit DACs in LT-B account for 21.8% (Figure 5(a)), while 8-bit DACs account for 50.5% (Figure 5(b)). These findings highlight that as bit precision increases to enhance model accuracy, DAC power consumption becomes a critical factor.

To achieve power-efficient large-scale matrix multiplication with photonic accelerators, this paper introduces a new design to convert data directly into optical analog signals, bypassing electrical signal processing. We propose replacing traditional DACs with pure-photonic DAC components to regulate MZMs. The primary challenge is configuring MZMs using optical digital signals exclusively through photonic components, eliminating the need for electrical DACs.

III. Photonic Digital-to-Analog Converter (P-DAC)

A. Overview

We present our design in Figure 6, where we propose the P-DAC to control the Mach-Zehnder Modulator (MZM) without using electronic devices, integrating it with Lightening-Transformer. In this approach, we eliminate both the controller and DAC, which are typically required for regulating the MZM in Lightening-Transformer. Additionally, we leverage the high data rate of optical interconnections to efficiently propagate data from the shared M2 SRAM. The design concept behind P-DAC is detailed in Section III-B, with

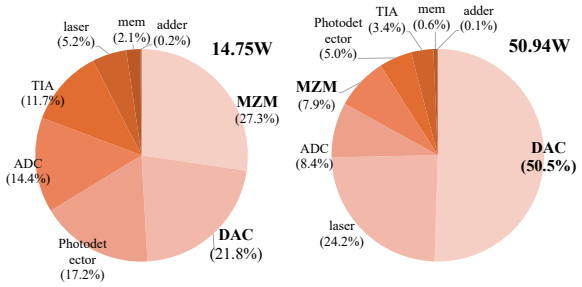


Fig. 5. Power breakdown of LT-B for (a) 4-bit (b) 8-bit precision.

mathematical derivations provided in Section III-B to demonstrate that P-DAC can fully replace the MZM regulation.

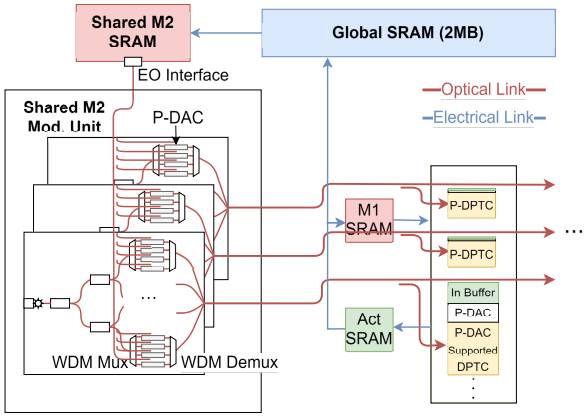


Fig. 6. Integration between Lightening-Transformer and P-DACs.

B. Design Concept behind P-DAC

In this section, we will explain how the Photonic-DAC (P-DAC) replaces traditional DACs, as well as its operational principles and advantages. Notably, in optical analog accelerators, Mach-Zehnder Modulators (MZMs) are typically used to convert electrical signals into optical analog signals. However, this process requires a controller to set the traditional DAC's output as the external voltage for the MZM, which, according to the experiments with the Lightening Transformer, results in significant power consumption from the DAC. In our design, we first convert the electrical digital signals into optical digital signals and then approximate them into optical analog signals, where this approximation process is key to the P-DAC's energy efficiency.

For the optical analog signal, Lightening-Transformer uses a controller to drive the DAC, which then applies the DAC's output voltage to the MZM. However, additional calculations are required to obtain the desired output voltage due to the non-linear nature of Equation (3). In our approach, we eliminate the need for both these calculations and the DAC. Instead, we directly use the P-DAC to convert the optical digital signal into an optical analog signal without prior computation.

It is worth noting that, since we initially convert the electrical digital signals into optical digital signals, we can also utilize the Wavelength Division Multiplexing (WDM) technique to pre-convert data from the memory side into optical digital signals and then transfer it to the P-DAC, thereby saving some energy.

We integrate the MZM into the P-DAC as the critical component for converting signals into optical analog format. Unlike traditional methods, we use only a few optical components to supply the external voltage to the MZM, significantly reducing the energy consumption

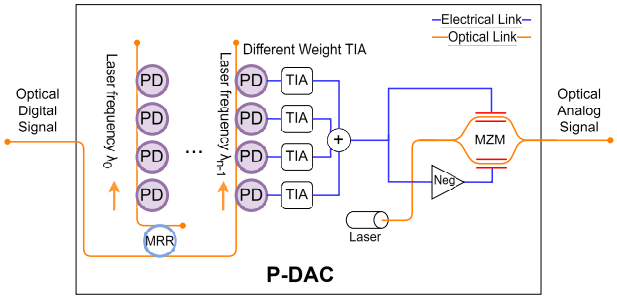


Fig. 7. P-DAC Structure: Integrated with OE interface and MZM

associated with traditional DACs. First, as shown in Figure 2, in the electrical-to-optical (EO) transformation, we adopt the multi-bit EO interface to convert signals into optical digital format. Then, as illustrated in Figure 7, once the optical digital data reaches the P-DAC, we refer to the multi-bit optical-to-electrical (OE) interface to apply different weights to each bit through a TIA and superimpose the voltages of each bit to serve as the MZM's input voltage.

Our P-DAC offers several advantages. It provides a simple implementation with high speed, functioning similarly to a Binary-Weighted DAC but ours is converting digital optical signals directly into analog optical signals. This design theoretically results in low power consumption, as the MZM operates based on relative voltage rather than absolute voltage, making its power usage dependent on the reference voltage. Note that the P-DAC exhibits slight numerical errors. However, since our target application is LLMs, which are inherently tolerant to minor inaccuracies, the P-DAC is perfectly suited for such use cases. By directly controlling the voltage weight of each bit, the input voltage can be calculated without the need for pre-calculations, thereby reducing controller (or CPU) overhead, streamlining the overall system design.

C. Mathematical Derivation for Validating P-DAC

Determining the weight values for the TIA is complex because Equation (3) is nonlinear, so weights cannot be assigned proportionally. Fortunately, given the stochastic nature of LLMs, exact numerical precision is not as critical, as long as the output falls within an acceptable range for human perception. Thus, we determine TIA weights through approximation. Our goal is to control the MZM equation's V_1 and V_2 so that the value of E_{out} becomes the desired analog signal, so we focus on the Equation (3) first.

We first simplify Equation (3) to make equation more readable,

$$E_{out} = \frac{E_{in}}{2}\left(e^{jV_1'} + e^{jV_2'}\right) \tag{7}$$

where we define $V_1' = \frac{\pi V_1}{2V_\pi}$ and $V_2' = \frac{\pi V_2}{2V_\pi}$, assuming no imbalance in splitting so that the k factor can be ignored. Then, applying Euler's formula ($e^{jx} = cosx + jsinx$) to Equation (7):

$$E_{out} = \frac{E_{in}}{2}\left(\left(cosV_1' + jsinV_1'\right) + \left(cosV_2' + jsinV_2'\right)\right) \tag{8}$$

Then we let $V_2' = -V_1'$, allowing the expression to be further simplified:

$$\begin{aligned} E_{out} &= \frac{E_{in}}{2}\left(cosV_1' + jsinV_1' + cos\left(-V_1'\right) + jsin\left(-V_1'\right)\right) \\ &= \frac{E_{in}}{2}\left(cosV_1' + jsinV_1' + cosV_1' - jsinV_1'\right) \\ &= E_{in}cosV_1' \end{aligned} \tag{9}$$

As we can see, Equation (9) is equivalent to Equation (2) when $V_1' = \phi$. Although we successfully simplified Equation (3) to Equation (9), we still need a way to map the data to the desired input voltage after knowing the relationship between input voltage and phase shift. As

the goal of P-DAC is to convert optical digital data to optical analog data, our goal now is to map the analog data to V_1' such that

$$r \propto E_{out} = E_{in} cos V_1' \tag{10}$$

where r can seen as the analog data within the standard interval (-1, 1), which in turn corresponds to the desired analog value E_{out}. After normalizing E_{in} in Equation (10), we obtain:

$$r = cos V_1' \tag{11}$$

To proceed with the approximation, we first need to write down the trivial equation:

$$x = cos\left(cos^{-1}(x)\right) \quad \text{when} \quad -1 \leq x \leq 1 \tag{12}$$

After comparing Equation (11) and Equation (12), we obtain the equation as follows:

$$V_1' = cos^{-1}(r) \quad \text{when} \quad -1 \leq r \leq 1 \tag{13}$$

Thus, when MZM's input voltage V_1' is $cos^{-1}(r)$, from Equation (10), we can ensure that $E_{out} = rE_{in}$. In summary, by adjusting TIA weights of each bit to transform the optical digital data into $cos^{-1}(r)$ and using it as the input voltage to the MZM, we can ensure that $E_{out} = rE_{in}$. For example, if digital value is $0x40$ in 8-bit system, which analog value can be calculated as $\frac{0x40}{2^7-1} = 0.5$, you can set the variable $V_1' = cos^{-1}(0.5)$ and apply Equation (10) to achieve $E_{out} = 0.5E_{in}$, where 0.5 is the analog data of $0x40$ within the standard interval (-1, 1). The only problem now is how to assign TIA weights for $0x40$ to convert it to $V_1' = cos^{-1}(0.5)$. Note that assigning TIA weights is straightforward when the conversion function is linear (e.g., $f(r) = ar + b$), as the mapping is direct. To obtain an approximation, we expand $cos^{-1}(r)$ using a Taylor series:

$$cos^{-1}(r) = \frac{\pi}{2} - \left(r + \frac{1}{6}r^3 + ...\right) \tag{14}$$

by taking first-order approximation from Equation (14), we can get the approximation as follows:

$$cos^{-1}(r) \approx \frac{\pi}{2} - r = f(r) \quad \text{when} \quad 0 \leq r \leq 1 \tag{15}$$

the equation simplifies to a linear form; however, the greatest error occurs at $r = 1$ and $r = -1$, as demonstrated in the following approximation:

$$\left|\frac{1 - cos(f(1))}{1}\right| \approx 15.9\%$$
$$\left|\frac{(-1) - cos(f(-1))}{-1}\right| \approx 15.9\%$$

To further improve the accuracy of the approximation, we introduce an additional linear equation to combine with the result of Equation (15). Given that the highest error occurs at $r = 1$, we identify a linear function that passes through $(1, cos^{-1}(1))$ or $(1, 0)$ to minimize the total error. Note that we focus on the positive domain, as the function is symmetric. Combining this with the result from Equation (15), we derive the following function:

$$f(r) = \begin{cases} \frac{\pi}{2} - r & \text{when} \quad 0 \leq r \leq k, \\ \frac{k-\pi/2}{k-1}(1-r) & \text{when} \quad k < r \leq 1. \end{cases} \tag{16}$$

where $\frac{k-\pi/2}{k-1}(1-r)$ is the expression of linear equation that passes $(0, \pi/2)$ and $(1, 0)$, and k is the intersection of two expressions.

To find the smallest total error, we have to find the minimum of following expression:

$$\left(\int_{r=0}^{k} \left|\frac{cos(\frac{\pi}{2} - r) - r}{r}\right|\right) + \left(\int_{r=k}^{1} \left|\frac{cos\left(\frac{k-\pi/2}{k-1}(1-r)\right) - r}{r}\right|\right) \tag{17}$$

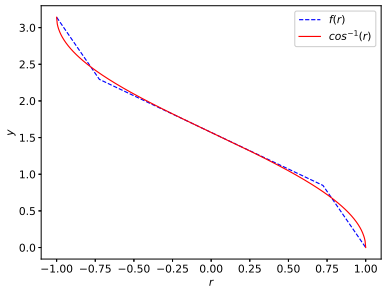


Fig. 8. The Plot of $f(r)$ and $cos^{-1}(r)$. When $r = \pm 0.7236$ has maximum error 8.5%

where $0 \leq k \leq 1$. After running the program to find the optimal k value, we determined that the smallest result occurs when $k \approx 0.7236$ has smallest value. To extend this analysis, we performed the same calculation for negative values of r, and the resulting function is as follows:

$$f(r) = \begin{cases} -3.0651r + 0.07648 & \text{when} \quad -1 \leq r \leq -0.7236, \\ \frac{\pi}{2} - r & \text{when} \quad -0.7236 < r \leq 0.7236, \\ -3.0651(r-1) & \text{when} \quad 0.7236 < r \leq 1. \end{cases} \tag{18}$$

the function plot is shown in Figure 8. The maximum error is at $r \pm 0.7236$, as demonstrated in the following calculation:

$$\left|\frac{-0.7236 - cos(f(-0.7236))}{-0.7236}\right| \approx 8.5\%$$
$$\left|\frac{0.7236 - cos(f(0.7236))}{0.7236}\right| \approx 8.5\%$$

Finally, the function in (18) is now linear, allowing us to easily assign the TIAs' weights using the above function. Note that the function in the P-DAC hardware can be easily decomposed into three parts by adding logic gates in the circuit (e.g., leq).

IV. Performance Evaluation

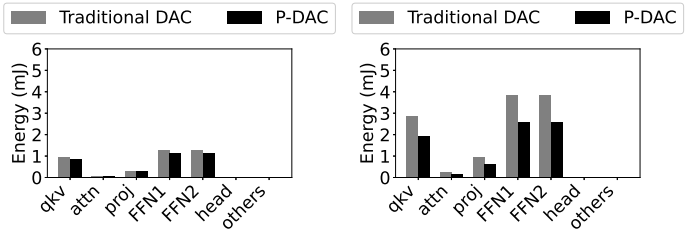
A. Performance Metrics and Evaluation Setup

In this section, we evaluate the performance of the proposed P-DAC design using power consumption and specific workload energy consumption as key metrics. The total power consumption is calculated by accounting for contributions from different modules, measured under typical operating conditions. The proposed design is compared with Lightening-Transformer with using traditional DAC [2], to highlight the advantages of the P-DAC. The workloads used in this evaluation are BERT [6] and DeiT [37]. Note that the power of the P-DAC is calculated based solely on its components and does not include circuit power. Also, since the MZM's power usage depends on the reference voltage, this suggests that circuit power can be further reduced, so we assume that circuit power can be neglected in this analysis.

This work utilizes the source code of Lightening-Transformer (written in python), which is distributed under the GNU General Public License (GPL). Modifications were made to adapt the DAC implementation for our proposed P-DAC.

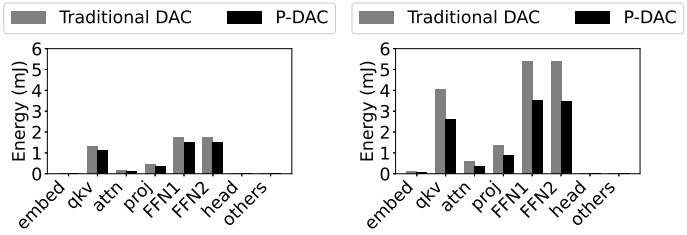
B. Evaluation Results

1) Model Workloads Energy Evaluation: Figure 9 presents the energy breakdown results for BERT-base with a sequence length of 128, while Figure 10 shows the energy breakdown for DeiT with ImageNet1K-224×224 and 197 tokens. Both figures compare the performance of the Lightening-Transformer using a traditional DAC versus our P-DAC. The x-axis represents different operations during inference, while the y-axis indicates the energy consumption



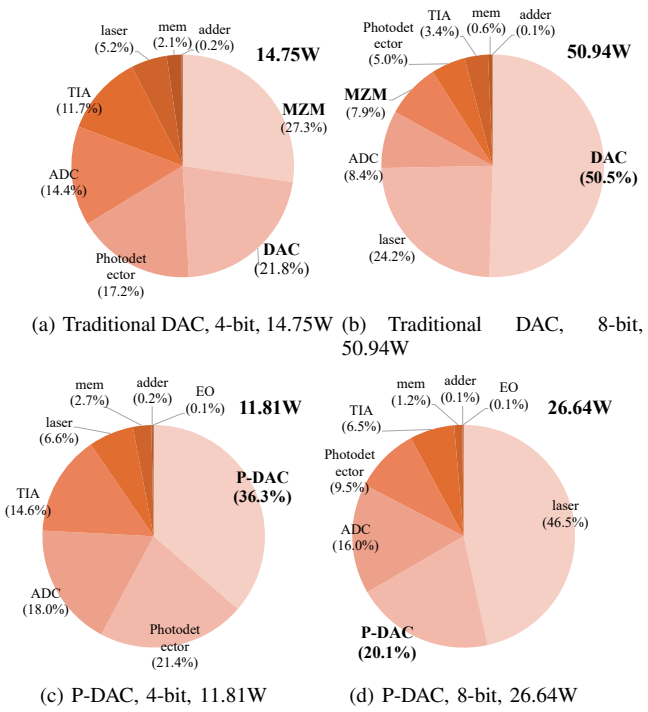
(a) 4-bit, Total: 3.951mJ, 3.510mJ (b) 8-bit, Total: 11.767mJ, 7.961mJ

Fig. 9. Energy breakdown of BERT-base with 128 sequence length



(a) 4-bit, Total: 5.437mJ, 4.740mJ (b) 8-bit, Total: 16.982mJ, 10.974mJ

Fig. 10. Energy breakdown of DeiT with ImageNet1K-224 × 224, 197 tokens



(a) Traditional DAC, 4-bit, 14.75W (b) Traditional DAC, 8-bit, 50.94W

(c) P-DAC, 4-bit, 11.81W (d) P-DAC, 8-bit, 26.64W

Fig. 11. Power breakdown of LT-B

for a single inference run. For BERT in 4-bit data size (Figure 9(a)), replacing the original architecture with the P-DAC design achieves an energy reduction of 11.2%, while for an 8-bit data size(Figure 9(b)), the reduction increases to 32.3%. For DeiT in 4-bit data size (Figure 10(a)), replacing the original architecture with the P-DAC design achieves an energy reduction of 11.2%, while for an 8-bit data size(Figure 10(b)), the reduction increases to 32.3%. The experimental results demonstrate that the proposed P-DAC design offers substantial energy efficiency improvements over traditional DAC-based architectures. It is worth noting that the data movement in the attention accounts for a smaller proportion compared to the FFN, resulting in a larger energy reduction in attention. This is because P-DAC does not affect the energy consumption associated with data movement. For instance, in BERT, the attention achieves an energy savings of 18.3% for 4-bit data and 42.1% for 8-bit data, while the FFN achieves savings of 11.0% for 4-bit and 32.1% for 8-bit. Similarly, in DeiT, the attention saves 19.0% for 4-bit and 42.3% for 8-bit, while the FFN saves 12.6% for 4-bit and 35.1% for 8-bit.

2) Full Compute-bound Scenario Power Evaluation: Figure 11 illustrates the power breakdown of each hardware component, with the P-DAC's power including the integrated MZM. This metric reflects a fully compute-bound scenario where hardware performance is not limited by memory bandwidth. As shown in Figure 11(a) and Figure 11(c), the P-DAC reduces power consumption by 19.9% compared to traditional DAC-based systems for a 4-bit data size. For an 8-bit data size, as depicted in Figure 11(b) and Figure 11(d), the power savings increase substantially to 47.7%.

This setup effectively highlights the capabilities of the P-DAC and provides a projection of its energy consumption under scenarios with sufficient memory bandwidth in the future. While we expect that higher bit sizes will enable more accurate responses from LLMs, the P-DAC significantly reduces energy consumption, yet the majority of the energy consumption remains constrained by the laser. This is because the energy savings provided by the P-DAC are so significant that the relative power consumption of the laser becomes more prominent. Therefore, we anticipate that with future advancements in laser technology, overall power consumption can be further reduced.

V. Conclusion

In this paper, we proposed the P-DAC (Photonic Digital-to-Analog Converter) design as an energy-efficient alternative to traditional DAC-based systems for optical signal processing. Our evaluation demonstrated that the P-DAC significantly reduces power consumption compared to conventional DAC systems, especially as the data size increases. Under a fully compute-bound scenario, the P-DAC achieved an impressive 47.7% reduction for 8-bit data. We also demonstrated the feasibility of the P-DAC design through mathematical inference. For both BERT and DeiT workloads, we observed considerable energy savings in the attention mechanism, with reductions as high as 35.4% for 8-bit data. These improvements highlight the potential of the P-DAC to enable more energy-efficient optical computing systems.

VI. Acknowledgement

We would like to thank Prof. David Brooks and Prof. Gu-Yeon Wei from Harvard University for their thoughtful comments and insightful suggestions. This work was supported in part by the National Science and Technology Council under grant nos. 113-2628-E-A49-021, 113-2640-E-A49-012 and Ministry of Education under Yushan Young Fellow Program.

References

[1] A. M. Alatwi, A. N. Zaki Rashed, and E. M. El-Gammal. Wavelength division multiplexing techniques based on multi transceiver in low earth orbit intersatellite systems. *Journal of Optical Communications*, 45(1):125–135, 2024.

[2] 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*, pages 1–2. IEEE, 2020.

[3] H.-Y. Cheng, C. Hakert, K.-H. Chen, Y.-H. Chang, J.-J. Chen, C.-L. Yang, T.-W. Kuo, et al. Future computing platform design: A cross-layer design approach. In *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 312–317. IEEE, 2021.

[4] I. Chremmos, O. Schwelb, and N. Uzunoglu. *Photonic microresonator research and applications*, volume 156. Springer, 2010.

[5] 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. *ACM Journal on Emerging Technologies in Computing Systems*, 19(4):1–31, 2023.

[6] J. Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] M. M. Eid, A. N. Z. Rashed, S. El-Meadawy, and M. A. Habib. Best selected optical fibers with wavelength multiplexing techniques for minimum bit error rates. *Journal of Optical Communications*, (0):000010151520200239, 2020.

[8] D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H.-S. P. Wong. Device scaling limits of si mosfets and their application dependencies. *Proceedings of the IEEE*, 89(3):259–288, 2001.

[9] 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*, 69(6):2581–2585, 2022.

[10] S. Jain, A. Sengupta, K. Roy, and A. Raghunathan. Rxnn: A framework for evaluating deep neural networks on resistive crossbars. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 40(2):326–338, 2020.

[11] Y. Jin, C.-F. Wu, D. Brooks, and G.-Y. Wei. s^3: Increasing gpu utilization during generative inference for higher throughput. *Advances in Neural Information Processing Systems*, 36:18015–18027, 2023.

[12] Y.-W. Kang, C.-F. Wu, Y.-H. Chang, T.-W. Kuo, and S.-Y. Ho. On minimizing analog variation errors to resolve the scalability issue of reram-based crossbar accelerators. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 39(11):3856–3867, 2020.

[13] S. V. Kartalopoulos. Introduction to dwdm technology. *(No Title)*, 1999.

[14] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[15] F. Koyama and K. Iga. Frequency chirping in external modulators. *Journal of Lightwave Technology*, 6(1):87–93, 1988.

[16] A. V. Krishnamoorthy, R. Ho, X. Zheng, H. Schwetman, J. Lexau, P. Koka, G. Li, I. Shubin, and J. E. Cunningham. Computer systems based on silicon photonic interconnects. *Proceedings of the IEEE*, 97(7):1337–1361, 2009.

[17] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023.

[18] W. Lee, J. Lee, J. Seo, and J. Sim. {InfiniGen}: Efficient generative inference of large language models with dynamic {KV} cache management. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pages 155–172, 2024.

[19] C. Li, M. Browning, P. V. Gratz, and S. Palermo. Luminoc: A power-efficient, high-performance, photonic network-on-chip. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 33(6):826–838, 2014.

[20] C. Li, X. Zhang, J. Li, T. Fang, and X. Dong. The challenges of modern computing and new opportunities for optics. *PhotoniX*, 2:1–31, 2021.

[21] 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)*, pages 15–28. IEEE, 2023.

[22] Y. Li, A. Louri, and A. Karanth. Sprint: A high-performance, energy-efficient, and scalable chiplet-based accelerator with photonic interconnects for cnn inference. *IEEE Transactions on Parallel and Distributed Systems*, 33(10):2332–2345, 2021.

[23] Y. Li, A. Louri, and A. Karanth. Spacx: Silicon photonics-based scalable chiplet accelerator for dnn inference. In *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 831–845. IEEE, 2022.

[24] T.-S. Lo, C.-F. Wu, Y.-H. Chang, T.-W. Kuo, and W.-C. Wang. Space-efficient graph data placement to save energy of reram crossbar. In *2021 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, pages 1–6. IEEE, 2021.

[25] S. Ma, D. Brooks, and G.-Y. Wei. A binary-activation, multi-level weight rnn and training algorithm for adc-/dac-free and noise-resilient processing-in-memory inference with envm. *IEEE Transactions on Emerging Topics in Computing*, 11(2):292–302, 2023.

[26] A. Narayan, Y. Thonnart, P. Vivet, A. Coskun, and A. Joshi. Architecting optically controlled phase change memory. *ACM Transactions on Architecture and Code Optimization*, 19(4):1–26, 2022.

[27] A. Narayan, Y. Thonnart, P. Vivet, and A. K. Coskun. Prowaves: Proactive runtime wavelength selection for energy-efficient photonic nocs. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 40(10):2156–2169, 2020.

[28] OpenAI. Gpt-4 technical report, 2023.

[29] R. Pope, S. Douglas, A. Chowdhery, J. Devlin, J. Bradbury, J. Heek, K. Xiao, S. Agrawal, and J. Dean. Efficiently scaling transformer inference. *Proceedings of Machine Learning and Systems*, 5:606–624, 2023.

[30] S. Rumley, M. Bahadori, R. Polster, S. D. Hammond, D. M. Calhoun, K. Wen, A. Rodrigues, and K. Bergman. Optical interconnects for extreme scale computing systems. *Parallel Computing*, 64:65–80, 2017.

[31] J. Shalf. The future of computing beyond moore's law. *Philosophical Transactions of the Royal Society A*, 378(2166):20190061, 2020.

[32] B. J. Shastri, A. N. Tait, T. Ferreira de Lima, W. H. Pernice, H. Bhaskaran, C. D. Wright, and P. R. Prucnal. Photonics for artificial intelligence and neuromorphic computing. *Nature Photonics*, 15(2):102–114, 2021.

[33] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, et al. Deep learning with coherent nanophotonic circuits. *Nature photonics*, 11(7):441–446, 2017.

[34] 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)*, pages 860–873. IEEE, 2021.

[35] T. N. Theis and H.-S. P. Wong. The end of moore's law: A new beginning for information technology. *Computing in science & engineering*, 19(2):41–50, 2017.

[36] C. A. Thraskias, E. N. Lallas, N. Neumann, L. Schares, B. J. Offrein, R. Henker, D. Plettemeier, F. Ellinger, J. Leuthold, and I. Tomkos. Survey of photonic and plasmonic interconnect technologies for intra-datacenter and high-performance computing communications. *IEEE Communications Surveys & Tutorials*, 20(4):2758–2783, 2018.

[37] 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*, pages 10347–10357. PMLR, 2021.

[38] C.-L. Tsai, C.-F. Wu, Y.-H. Chang, H.-W. Hu, Y.-C. Lee, H.-P. Li, and T.-W. Kuo. A digital 3d tcam accelerator for the inference phase of random forest. In *2023 60th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2023.

[39] A. Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[40] Z. Wang, T. Luo, C. Liu, W. Liu, R. S. M. Goh, and W.-F. Wong. Enabling energy-efficient deployment of large language models on memristor crossbar: A synergy of large and small. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[41] Z. Wang, Z. Wang, J. Xu, Y.-S. Chang, J. Feng, X. Chen, S. Chen, and J. Zhang. Camon: Low-cost silicon photonic chiplet for manycore processors. *IEEE transactions on computer-aided design of integrated circuits and systems*, 39(9):1820–1833, 2019.

[42] S. Werner, J. Navaridas, and M. Luján. Designing low-power, low-latency networks-on-chip by optimally combining electrical and optical links. In *2017 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 265–276. IEEE, 2017.

[43] Q. Zheng, Z. Wang, Z. Feng, B. Yan, Y. Cai, R. Huang, Y. Chen, C.-L. Yang, and H. H. Li. Lattice: An adc/dac-less reram-based processing-in-memory architecture for accelerating deep convolution neural networks. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2020.

[44] S. Zheng, J. Zhang, and W. Zhang. High-throughput optical neural networks based on temporal computing. *arXiv preprint arXiv:2303.01287*, 2023.

[45] H. Zhu, J. Gu, H. Wang, Z. Jiang, Z. Zhang, R. Tang, C. Feng, S. Han, R. T. Chen, and D. Z. Pan. Lightening-transformer: A dynamically-operated optically-interconnected photonic transformer accelerator. In *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 686–703. IEEE, 2024.