2024 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW) | 979-8-3503-6460-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/IPDPSW63119.2024.00118

PCM Enabled Low-Power Photonic Accelerator for Inference and Training on Edge Devices

Juliana Curry
Dept. of Electrical and Computer Engineering
The George Washington University
Washington, D.C.
jrcurry97@gwu.edu

Ahmed Louri
Dept. of Electrical and Computer Engineering
The George Washington University
Washington, D.C.
louri@gwu.edu

Avinash Karanth
School of Electrical Engineering and Computer Science
Ohio University
Athens, Ohio
karanth@ohio.edu

Razvan Bunescu
Dept. of Computer Science
University of North Carolina at Charlotte
Charlotte, North Carolina
rbunescu@charlotte.edu

***Abstract*—The convergence of edge computing and artificial intelligence requires that inference is performed on-device to provide rapid response with low latency and high accuracy without transferring large amounts of data to the cloud. However, power and size limitations make it challenging for electrical accelerators to support both inference and training for large neural network models. To this end, we propose Trident, a low-power photonic accelerator that combines the benefits of phase change material (PCM) and photonics to implement both inference and training in one unified architecture. Emerging silicon photonics has the potential to exploit the parallelism of neural network models, reduce power consumption and provide high bandwidth density via wavelength division multiplexing, making photonics an ideal candidate for on-device training and inference. As PCM is reconfigurable and non-volatile, we utilize it for two distinct purposes: (i) to maintain resonant wavelength without expensive electrical or thermal heaters, and (ii) to implement non-linear activation function, which eliminates the need to move data between memory and compute units. This multi-purpose use of PCM is shown to lead to significant reduction in energy consumption and execution time. Compared to photonic accelerators DEAP-CNN, CrossLight, and PIXEL, Trident improves energy efficiency by up to 43% and latency by up to 150% on average. Compared to electronic edge AI accelerators Google Coral which utilizes the Google Edge TPU and Bearkey TB96-AI, Trident improves energy efficiency by 11% and 93% respectively. While NVIDIA AGX Xavier is more energy efficient, the reduced data movement and GST activation of Trident reduce latency by 107% on average compared to the NVIDIA accelerator. When compared to the Google Coral and the Bearkey TB96-AI, Trident reduces latency by 1413% and 595% on average.**

***Index Terms*—neural networks, photonic accelerators, phase change material, training, inference**

I. Introduction

The growth of the Internet of Things (IoT) and edge devices has led to significant progress in a wide range of applications such as image recognition, mobile augmented reality, and edge artificial intelligence (Edge AI) [18]. Most Edge AI relies on cloud computing and requires moving large amounts of data back and forth between the edge and the cloud, leading to issues of latency and privacy. Edge AI neural networks (NNs) depend on the implementation of matrix multiplications to compute layer outputs as well as nonlinear activation between layers [18]. Moving both inference and training to edge devices is primarily constrained by the size and power consumption limitations of edge devices [18]. Since edge devices are often battery-powered, low-power consumption is required. Therefore, as neural network model size increases to attain more accuracy, training and inference on edge devices while exploiting model parallelism and reducing energy consumption simultaneously, is a major challenge.

Emerging silicon photonics has been proposed for NN computation as it offers several advantages including higher performance-per-Watt, reduced energy consumption for data movement, higher bandwidth density, and overall improvement in execution time [6], [27], [33]. Photonic matrix multiplication and linear algebra operations can be performed in a single step, only inhibited by peripheral control operations such as modulation of filters or detection of optical signals [38]. As a result, photonic accelerators have demonstrated the potential to increase computing speed by 2-3 orders of magnitude [38]. Several photonic architectures have been proposed for NN inference operation, many of which are based on the broadcast-and-weight architecture [2], [9], [32]. The broadcast-and-weight architecture has been shown to perform multiply and accumulate (MAC) operations at frequencies up to five times faster than conventional electronics by using a tunable bank of microring resonators (MRRs) to encode NN weights [32]. In these architectures, the MRRs are tuned using the thermo-optic or the electro-optic effects to shift the resonant wavelength which consumes significant power in the range of 2 mW, to implement a single weight of the NN [9], [37].

Despite advancements in the use of photonics for inference, challenges with storing data in photonics have prevented in-situ training [7]. Many current photonic architectures train the NN with a digital model before mapping the trained network parameters to the optical hardware for inference where the weights will be static [2], [24], [30], [31]. This method is time-consuming and incurs a significant energy cost, and limits when weights can be updated. Additionally, digital models used at the time of training cannot capture all the manufacturing imperfections and variations of the physical hardware. The resulting mismatch between trained and implemented weights leads to sub-optimal accuracy at inference time [9]. These issues motivate a unified implementation of both training and inference on the same underlying optical hardware.

Phase change materials (PCMs) are an emerging technology that can be utilized as an alternative tuning method for MRRs. $Ge_2Sb_2Te_5$ (GST) is a PCM commonly used in optical storage

979-8-3503-6460-6/24/$31.00 ©2024 IEEE
DOI 10.1109/IPDPSW63119.2024.00118

and processing-in-memory because it exhibits distinct refractive indices and resistances as it switches between amorphous and crystalline states [25], [36]. The phase transition of GST is re-programmable and non-volatile, making it an energy-efficient tuning method when compared to electrical or thermal heaters. Due to the storage capabilities of GST, it can also implement a nonlinear activation function, another critical component of NN computations. The major contributions of this paper are:

- **Trident**: A low-power photonic accelerator that uses PCM to effectively combine both training and inference for edge AI devices. By using the same hardware accelerator, we eliminate the mismatch between trained and implemented weights in prior designs.
- **Photonic Non-Linear Activation**: GST activation cell that allows the activation function to be stored within the processing element (PE), reducing electro-optic (E/O) and optoelectric (O/E) conversions as well as excessive memory accesses. The GST activation cell also eliminates the need for analog-to-digital converters (ADCs). This is a critical improvement because ADCs are a serious bottleneck that severely limits inference throughput per Watt in previous photonic accelerators [23]. Moreover, we also use PCM to maintain the resonant wavelength without expensive electrical or thermal heaters, thereby improving energy efficiency.
- **Performance Evaluations**: We evaluate the energy and latency performance of Trident on convolutional neural network (CNN) models GoogleNet, VGG-16, MobileNetV2, and ResNet-50 and compare the results with previous photonic accelerators. Trident improves energy efficiency by up to 43.5% on average and improves latency by up to 150.2% on average over previous photonic accelerators DEAP-CNN [2], CrossLight [31], and PIXEL [30]. Compared to electronic edge AI accelerators Google Coral [12] which utilizes the Google Edge TPU and Bearkey TB96-AI [22], Trident improves TOPS per Watt by 11.5% and 93.3% respectively. Despite the higher energy efficiency of the NVIDIA AGX Xavier, the reduced data movement and GST activation of Trident reduce latency by 107.7% on average compared to the NVIDIA AGX Xavier. Compared to the Google Coral and the Bearkey TB96-AI, Trident reduces latency by 1413.1% and 594.7% on average.

II. Background and Motivation

A. NN Basics for Inference and Training

Backpropagation (BP) is the most widely used algorithm to train neural networks. Training a neural network of N layers alternates between forward passes and backward passes where each layer receives some blame for how much it affected the resulting loss. Forward passes perform inference, which includes multiple layers of input vectors x multiplied with weight matrices W_k and non-linear activation. The result of a forward pass is compared with the input's label to compute an error signal. Backward passes send the error signal back to update the weights, essentially a continuous application of the chain rule to compute the gradient of the loss with respect to different weights and then use this value to do a gradient descent update.

TABLE I: Tuning Method Comparison

Tuning Method	Tuning Power	Speed
Thermal	1.02 nJ [9]	0.6 μs [9]
Electric	0.18 pm/V [15]	500 ns [15]
GST	660 pJ [37]	300 ns [13]

In a network where weights W_k connect layer $k-1$ to layer k and input x is considered layer 0, a logit is $h_k = W_k y_{k-1}$, and the activation is $y_k = f(h_k)$ the backpropagation update rule is as follows:

$$W_k = W_k - \beta * \delta W_k \tag{1}$$

where β is the learning rate,

$$\delta W_k = \delta h_k * y_{k-1}^T \tag{2}$$

and

$$\delta h_k = (W_{k+1}^T * \delta h_{k+1}) \odot f'(h_k) \tag{3}$$

B. Conventional Photonic Accelerator Devices

The majority of energy consumption and execution time incurred in existing photonic accelerators is rooted in the tuning of MRRs. Therefore, reducing the energy consumption for MRR tuning has the potential to improve the performance of photonic accelerators. There are several ways to actively tune the resonant wavelength of a ring resonator, compared in Table I. Electronic tuning is not widely used because the electro-optic effect has a limited range. Electronic tuning at 0.2 pm/V or 24.0 Hz/V requires applying DC voltage in the range of -100.0V to 100.0V to a 60.0 μm radius ring [15]. These wide voltage ranges and large rings complicate designs and quickly contribute to power and area consumption as the number of MRRs increases and becomes impractical for edge devices. Therefore, electronic tuning is not considered in this work. Thermal tuning requires individual heaters for each MRR which can shift an MRR's resonant wavelength within $\phi \pm 0.2$ to modulate the amplitude of the input signal [2], [4]. This approach of shifting the resonant wavelength is limited to avoid crosstalk from the adjacent channels in a multi-channel WDM system [2]. Crosstalk in thermally tuned MRRs results in a bit resolution of only 6 bits [10], meaning that training is not possible [34]. While effective, thermal and electrical methods of tuning are volatile and require the constant application of power. In addition to reducing energy consumption due to non-volatile tuning, PCM tuned MRRs have the added benefit of 8-bit resolution computation, enough for NN training.

III. Trident Architecture Design

A. Architectural Overview

The Trident architecture is shown in Figure 1. The architecture is used for both inference and training by encoding different values on the same hardware with an external control unit handling encoding. The proposed Trident architecture utilizes existing photonic devices to perform the MAC operations

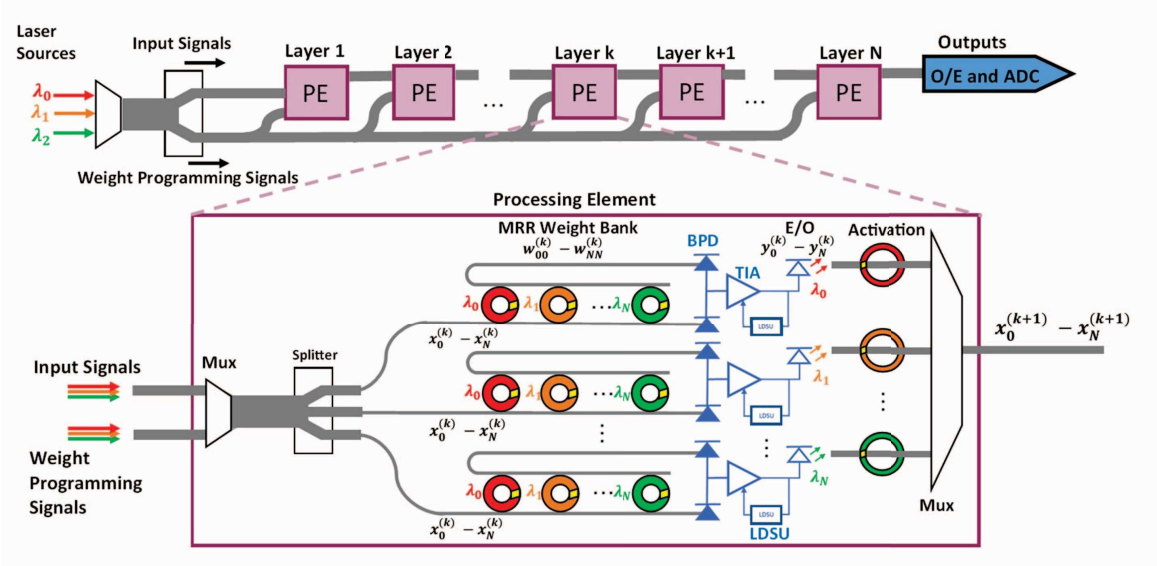


Fig. 1: The Trident Architecture Design includes a wavelength division multiplexed (WDM) waveguide (Figure 2a) that distributes the input laser sources to a chain of Processing Elements (PEs). The PE consists of a microring resonator (MRR) weight bank made up of MRRs with phase change material (PCM) cells (Figure 2b) that weight the inputs encoded onto the incoming laser signals before the data is accumulated by balanced photo-detectors (BPDs). The weights are pre-programmed using optical signals on the same WDM waveguide. The electronic portion (Figure 2c) consists of transimpedance amplifiers (TIAs), linear derivative storage units (LDSUs) (Figure 2d), and E/O lasers, all of which are included to enable training. After conversion back to the optical domain via the E/O lasers, the output signals of each layer are sent to $Ge_2Sb_2Te_5$ (GST) activation cells. The blue devices and connections indicate electronic devices and connections while all other colors represent photonic devices and connections.

necessary for NNs. After completing a MAC operation and non-linear activation on the first PE, the output of layer 1 forwarded directly to the input of the second PE where the weight bank has been pre-programmed and the second layer's MAC and non-linear activation can be performed. The output of each layer is forwarded to the next until the last layer is completed and the outputs can then be converted back to the digital domain and saved in memory. By assigning one PE to each layer of a NN, the weights can be pre-programmed for all the layers and stored inside the PCM of each MRR in the weight bank. Then, inference can be completed at the speed of light and forwarded between layers without any delay for fetching weights from memory or tuning the MRRs.

First, the input laser sources are combined and distributed to an array of PEs using a wavelength division multiplexed (WDM) waveguide such as the one shown in Figure 2a. This is accomplished by using MRRs (Figure 2b) to construct a weight bank with J rows and N columns. This structure allows for an input vector of size $Nx1$ to be multiplied with J rows of a weight matrix in parallel. Each input is assigned a wavelength and its value is encoded onto the amplitude of the laser. Each laser source has a wavelength corresponding to the different colors of MRRs with matching resonant wavelengths. The red λ_0, orange λ_1, and green λ_N rings in Figure 1 represent MRRs with different resonant wavelengths λ_0, λ_1, ..., λ_N spaced at least 1.6 nm apart, to correspond with the input laser sources being used to encode inputs x_0, x_1, ..., x_N [32]. With the resonant wavelength appropriately spaced, the intensity of λ_1, λ_2, ..., λ_N are ignored by the red λ_0 MRR on the left and passed through the WDM waveguide, until they are filtered by their corresponding MRR. The add-drop configuration is used for the MRRs so that both positive and negative weights $w \in [-1, 1]$ can be encoded [2]. Each row of the PE's MRR weight bank also includes a balanced photodetector (BPD), a transimpedence amplifier, an LDSU, an E/O laser, and finally a GST activation cell. The output of every row is encoded onto a different wavelength before being forwarded on to the next PE.

Weighted sum MACs have already been demonstrated on previous photonic accelerators [2], [9], [32]. The proposed design is novel because of the low-power, non-volatile GST tuning method for the MRR weight bank, and because of the photonic non-linear activation which allows an optical pulse to be fed into the next layer without the added delay of storing the output to an external memory unit using ADCs. Additionally, the GST tuning method allows for a higher bit resolution computations which, along with the photonic activation, makes training possible in a photonic accelerator.

1) Inference: To perform inference the weights are first pre-loaded into the MRR weight bank using optical weight programming signals that are sent in parallel to tune the weight bank MRRs. Similarly, the activation function is pre-loaded into the GST activation cell. Then, when the input signals are sent into the PE, the PCM of each MRR acts as a multiplier.

TABLE II: PE Hardware Devices Mapping

Device	Inference	Training Gradient Vector	Training Outer Product
Input Laser Sources	x_k	δh_{k+1}	δh_k
MRR Weight Bank	w_k	W_{k+1}^T	y_{k-1}^T
BPD Output	$y_k = w_k x_k$	$\delta h_k = (W_{k+1}^T * \delta h_{k+1})$	$\delta W_k = \delta h_k \cdot y_{k-1}^T$
TIA, E/O Laser Sources	y	$f'(h_k)$	δW_k

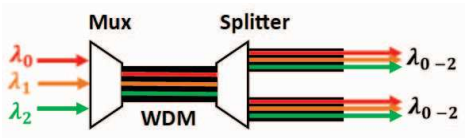
The balanced photodetector (BPD) collects the amplitude from the output of each PE on its row, effectively accumulating partial products. Therefore, the output of each BPD is a vector dot product. After passing through the BPD, the signal is amplified by the trans-impedance amplifier (TIA) and guided to the GST activation cell. The output of the activation cell is then forwarded to the next PE, where the weights have already been pre-loaded, to act as the input signals for a PE. Once a forward pass of inference has been completed, a backward pass for training can be performed.

2) Training: Training via backpropagation consists of the gradient vector computation and the weight update matrix computation. During the computation of the gradient vector $\delta h_k = (W_{k+1}^T * \delta h_{k+1}) \odot f'(h_k)$, the MRR weight bank is encoded with W_{k+1}^T, the input signals are equivalent to δh_{k+1}, and the TIAs after the BPD are tuned to $f'(h_k)$ to implement the necessary Hadamard product. Then, an outer product computation is performed to calculate the weight update matrix $\delta W_k = \delta h_k \cdot y_{k-1}^T$. To implement this, the MRR weight bank is encoded with y_{k-1}^T from N inputs, to utilize the entire weight bank and perform N outer products, and the input signals are equivalent to δh_k. For computation of the weight update matrix δW_k, no Hadamard product is needed, so the TIAs are used more generally to amplify the analog signal. The hardware device being used to represent each element of the equations used for inference and training is listed in Table II.

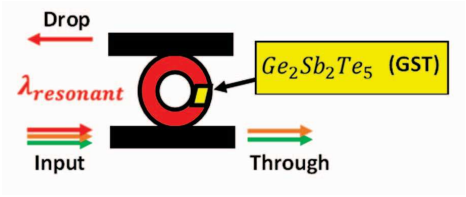
B. PCM-MRR Weight Bank

Instead of using conventional MRRs with thermo-optic or electro-optic tuning, Trident employs the emerging PCM tuning method, using GST as the PCM. GST has been applied widely in different fields including optical switching and routing [13], [21]. Crosstalk is not an issue for the GST tuning method since the resonant wavelength of each MRR is not being shifted. Instead, the GST acts as an attenuator on the optical signal, and bit resolution is dependent on the number of GST states. Current technology is capable of 255 levels for 8-bit resolution [5]. To program an integrated GST cell, optical pulses are used to switch between the crystalline and amorphous states or an intermediate state [8]. Optically tuning MRRs eliminates the area requirement for thermal heaters, as well as thermal crosstalk issues.

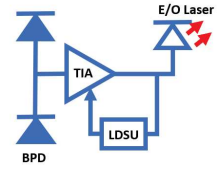
With the GST in the amorphous state, the photonic waveguide is highly transmissive, representing a large weight. In the crystalline state, most of the light is absorbed, leading to a small weight [37]. During a write cycle a high-power write



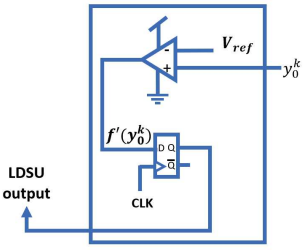
(a) WDM Waveguide



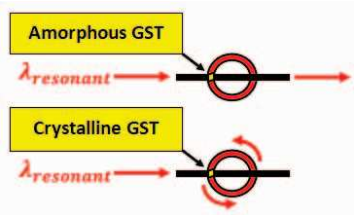
(b) MRR with PCM Weighting



(c) Electronic Elements



(d) LDSU



(e) GST Activation Cell

Fig. 2: Components of Trident architecture: (a) a wavelength division multiplexed (WDM) waveguide and (b) a microring resonator (MRR) with phase change material (PCM) cell made of $Ge_2Sb_2Te_5$ (GST). (c) Exploded view of a Linear Derivative Storage Unit (LDSU) which is comprised of an analog voltage comparator and a D-flip-flop. The LDSU stores the derivative of the GST Activation Function to enable training. (d) GST Activation Cell when the GST is in a fully amorphous state or a fully crystalline state.

pulse, greater than or equal to 660 pJ [37] and at the resonant wavelength of the MRR that is being written to, is injected from the MRR's input port. This high-power write pulse reduces the crystalline property of GST, and affects the transmittance of the drop and through ports of the MRR. During a read cycle, a short low-power optical pulse, about 20 pJ [8] and at the resonant wavelength of the MRR that is being read, is injected from the MRR's input port. The power consumption for tuning GST is 2.0 mW, slightly higher than the 1.7 mW of power needed to thermally tune an MRR. However, once the GST cell has been tuned, its state is non-volatile and is maintained

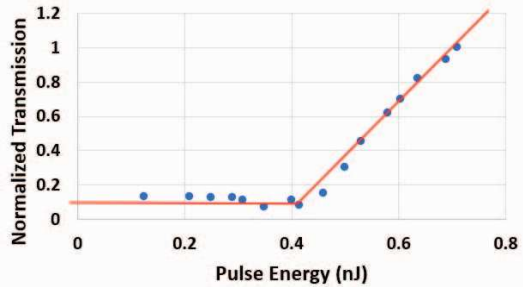


Fig. 3: The Output Function of the GST Activation Cell

until reset. Tuning a GST cell optically also only takes 0.3 μs, two times faster than thermally tuning an MRR. GST is reconfigurable, non-volatile for up to 10 years, can be written to and read from using optical pulses, and is capable of 8 bits of resolution, making it ideal for tuning MRRs to implement weights for both inference and training.

C. PCM Photonic Activation

During NN inference, a neuron applies a linear transformation to the input vector through a weight matrix, then a non-linear transformation is applied to the product through a non-linear activation function. The GST activation cell's response is similar to the commonly used rectified linear unit (ReLU) non-linear activation function. The GST activation cell, pictured in Figure 2e consists of a larger ring resonator (radius 60 μm) with an embedded GST cell at the intersection of the ring and the crossing waveguide. The GST at the intersection of the waveguide and the MRR has a non-linear response, shown in Figure 3. When the GST activation cell is in the crystalline state, the weighted sum pulse sent out of the output lasers couples strongly into the ring resonator, resulting in no observed output pulse. However, if the combined power of the weighted sum pulses is high enough to switch the GST activation cell to its amorphous state, the weighted sum pulse is no longer in resonance with the ring and will be transmitted beyond the ring, thus generating an output activation. As the switching of the GST activation cell only occurs above a certain threshold power, the neuron only generates an output pulse if the weighted sum exceeds this threshold. Thus, the system naturally executes non-linear activation over the optical power [8]. We note that all GST activation cells need to be recrystallized after each non-linear activation event. Therefore, the number of operation cycles is eventually limited by the endurance of the PCM cells. However, endurance is not a concern because individual PCM devices in endurance experiments have already shown the ability to perform a trillion switching cycles when fabricated to meet industry standards [17].

To perform training, the derivative of the activation function, $f'(h_k)$, is stored during inference. This is accomplished using the linear derivative storage unit (LDSU) pictured in Figure 2d. For each fixed wavelength, the contrast and maximum transmission level of the output function can be adjusted. In [8] the GST activation cell was measured at a fixed wavelength, resulting in a nonlinear activation function. Figure 3 shows the activation at 1,553.4 nm. Because of the shape of the GST activation function, we can consider it as having two possible derivatives. If h_k is greater than the threshold, then $f'(h_k)$ is 0.34. If h_k is less than the threshold, then $f'(h_k)$ is 0. Since the result of a weighted sum MAC is electronic, a standard voltage comparator can be used to determine if each element of h_k is greater than or less than the activation threshold, 430.0 pJ. The result of the comparator is stored as a single bit in a D-flip-flop so that when it is time for the backward pass, the TIA can be programmed to $f'(h_k)$. Thus, a voltage comparator and D-flip-flop make up Trident's LDSU and enable training to be performed on the same architecture as inference. If the activation function is not needed for any layer, the GST activation cell can be set to a fully amorphous state, effectively eliminating the activation cell.

TABLE III: Trident Device Power Breakdown

Component	Power	Percentage
LDSU	$0.09mW$ [3], [16]	0.01%
E/O Laser	$0.032mW$ [28]	0.00%
GST MRR Tuning	$563.2mW$ [37]	83.34%
GST MRR Read	$17.1mW$ [8]	2.52%
GST Activation Function Reset	$53.3mW$ [8]	7.89%
BPD and TIA	$12.1mW$ [19]	1.78%
Cache	$30mW$ [30]	4.44%
Total	$0.67W$	100 %

Previous photonic accelerators [2], [20] only compute MAC operations using photonic hardware and use ADCs to store the result of a layer in digital memory. Then, the activation function is implemented digitally, and its result is encoded onto the photonic hardware which will carry out the MAC of the next layer. Trident's LDSU eliminates the need for ADCs between layers. The LDSU along with the GST activation cell reduce the latency of activation while avoiding the need to fetch $f'(h_k)$ from memory during training since it is already stored at the PE where the gradient vector is computed.

IV. Evaluation

We evaluate the performance of the proposed Trident architecture using current technology. The device parameters used for these estimates are shown in Table III. In the lifetime of an accelerator, inference will be performed much more often once a model has been successfully trained. Therefore, this analysis focuses on energy efficiency and latency during inference when comparing with existing photonic accelerators and electronic edge AI accelerators that can only perform inference.

We evaluate the performance of Trident on CNN models GoogleNet, MobileNet, VGG-16, AlexNet, and ResNet-50, all of which use the ReLU activation function. We perform a per-layer analysis using Maestro to yield latency and energy metrics for inference on these CNN models. The image input to each of these CNN models is assumed to have dimensions of $224 \times 224 \times 3$ and a weight stationary dataflow is used. We compare Trident with three recent photonic NN accelerators: DEAP-CNN [2], CrossLight [31], and PIXEL [30]. DEAP-CNN is a broadcast-and-weight-based accelerator that utilizes thermally tuned MRRs and digital activation function.

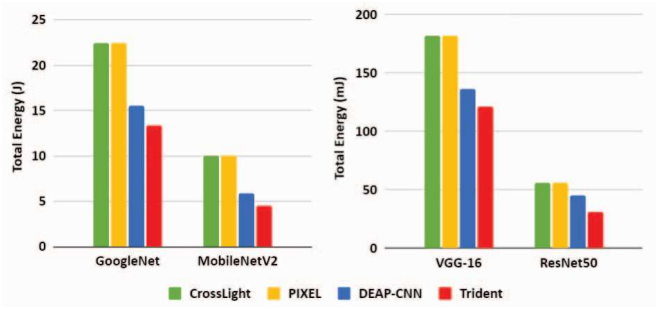


Fig. 4: Photonic Accelerators Total Energy Comparison

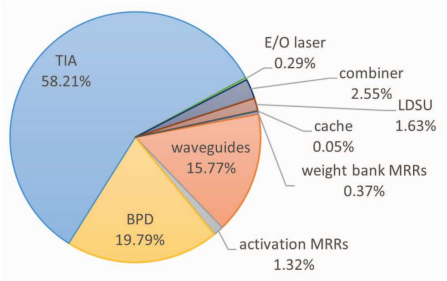


Fig. 5: Trident Chip Area Breakdown by Component

CrossLight also performs vector dot products using MRRs and summation using photodetectors but performs MRR tuning using a hybrid of thermo-optic and electro-optic tuning to reduce crosstalk. PIXEL is a mixed-signal photonic accelerator built using MRRs for bitwise logical operations and MZMs for analog accumulation. We compare against the 8-bit OO optical MAC unit from PIXEL. We apply the same device parameters in Table III to DEAP-CNN, CrossLight, PIXEL, and Trident and scale all four architectures to meet a 30 W power consumption threshold.

The threshold of 30 W was chosen to compare directly with state-of-the-art electronic edge AI accelerators NVIDIA AGX Xavier [1], Bearkey TB96-AI [22], and Google Coral Dev Board [12] which have a power draw of 30 W, 20 W, and 2 W respectively. These three electronic designs were chosen to represent the range of edge AI accelerators currently on the market. NVIDIA Jetson family was initially introduced by NVIDIA with TK1 in 2014 [26]. The family consists of 4 single board computers (SBCs) designed to deliver different price, performance, and power draw ratios at the edge of the network. All SBCs feature ARM CPU and NVIDIA GPU with CUDA support. Jetson AGX Xavier [1] is capable of training and is top of the Jetson line. TB96-AI is a SBC that features 2 possible RAM configurations - 3 GB (CPU 2GB + 1GB NPU) or 8GB (CPU 4GB + 4GB NPU) and allows running Android or Fedora OSs. Google Coral Dev Board is based on the Edge Tensor Processing Unit (Edge TPU) co-processor. This specialized ASIC, designed by Google, allows for inference of selected TF Lite models and to transfer-learn some pre-trained computer vision models [29]. However, it is not possible to train models from scratch. Edge TPU allows for up to 2 TOPS (int8) of AI performance per watt at a power draw of 2W (4 TOPS peak performance). Edge TPU is also sold as an accelerator module. However, the Google Coral Dev Board consumes up to 15 W, resulting in a maximum of 0.26 TOPS per Watt.

Maintaining the 30 W power limit for edge devices, a maximum of 44 PEs can be utilized, each with 256 MRRs for weight matrix-vector multiplication. All 44 PEs consume an area of $604.6mm^2$, less than 1 square inch, resulting in a compact and area efficient design suitable for many edge devices. Most of that area is consumed by the TIAs as shown in the area breakdown in Figure 5 and most of the power draw is consumed by tuning the weight bank MRRs as shown in Table II. That is why the non-volatile property of GST is so important. Once the weights are tuned in a PE, the power draw is reduced by 83.34% from 0.67 W to 0.11 W for the next MAC that uses the same weights. Each PE has a cache with 16 kB of storage and a footprint of $0.092 \times 0.085\ mm^2$ and shared L2 cache size 32 MB to handle storing data. We also assume a 1.37GHz maximum clock rate.

V. Results

A. Inference

Trident outperforms all three photonic accelerators in terms of energy efficiency during inference. As shown in Figure 4, Trident improves energy efficiency by and average of 16.4% compared to DEAP-CNN, 43.5% compared to CrossLight, and 43.4% compared to PIXEL. The energy advantages of Trident are greater when compared to CrossLight and PIXEL since CrossLight uses an additional VCSEL and MRR for summation and PIXEL uses power-hungry MZMs. Despite the added power of the LDSU and photonic activation reset, Trident is more energy efficient overall because tuning the weight bank MRRs monopolizes power consumption. In addition, the photonic non-linear activation and LDSU remove the need for ADCs between PEs, further reducing power consumption for Trident.

To compare Trident's energy efficiency against current edge AI accelerators we calculate TOPS per Watt. The most time consuming operation of performing a MAC operation is the tuning time of GST cells inside weight matrix MRRs, which takes 300 ns. However, all of the MRRs can be tuned in parallel so that weights are pre-loaded, after which inference can be performed on many inputs without re-tuning. Because photonics propagation time is equal to the speed of light and all the MRRs can be tuned in parallel MAC operations can be computed at a rate of 7.8 TOPS resulting in 0.29 TOPS per Watt. State-of-the-art edge AI accelerators such as the NVIDIA AGX Xavier [1], Bearkey TB96-AI [22], and Google Coral Dev Board [12] are capable of 1.1, 0.15, and 0.26 TOPS per Watt respectively. Trident outperforms the Bearkey TB96-AI [22] in terms of TOPS per Watt by 93.3%, and Google Coral by 11.5%. The NVIDIA AGX Xavier is more energy efficient than Trident at 1.1 TOPS per Watt [11]. However, Trident is capable of both inference and training, including activation, in the photonic domain.

The Bearkey TB96-AI and the Google Coral both are only capable of inference using pre-trained models and cannot be

TABLE IV: Performance of Trident vs. Electronic Accelerators

Accelerator	TOPS	Watts	TOPS per W	Training
NVIDIA AGX Xavier	32	30	1.1	Yes
Bearkey TB96-AI	3	20	0.15	No
Google Coral	4	15	0.26	No
Trident	7.8	30	0.29	Yes

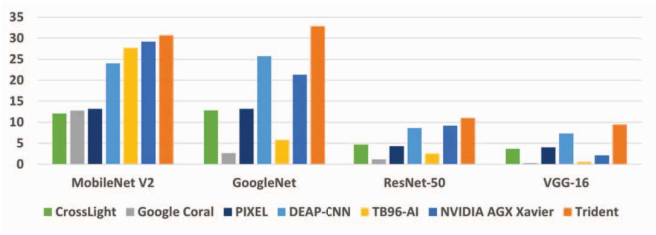


Fig. 6: Edge Accelerators Inferences per Second Comparison

used for training. Unlike NVIDIA AGX Xavier and the rest of the NVIDIA Jetson series, Trident stores and performs the activation function within the accelerator processing elements, reducing data movement between memory and compute units as well as performing the activation function at the speed of light. Because of this, Trident can perform more inferences per second across various NN models as shown in Figure 6. Trident reduces latency by an average of 107.7% compared to the NVIDIA AGX Xavier, 1413.1% compared to the Google Coral Dev Board, and 594.7% compared to the Bearkey TB96-AI. When compared to photonic accelerators DEAP-CNN, CrossLight, and PIXEL, Trident improves inferences per second by 27.9%, 150.2%, and 143.6% on average respectively. The reduced latency of Trident is mostly due to the GST tuning method. Compared to the thermal heaters used to tune DEAP-CNN and PIXEL, the GST tuning used in Trident is 2× faster. Additionally, the more energy efficient tuning method allows Trident to scale to more PEs than other photonic accelerators while remaining within the 30 W power requirement.

B. Training Latency

We evaluate training latency on the two accelerators capable of training, Trident and the NVIDIA AGX Xavier, by calculating the amount of time it would take to train various NN models on 50,000 training images. We use the throughput during inference of these models to estimate throughput during training instead of relying on pure TOPS to account for data movement and resource sharing latency. The training time for each NN model varies based on the number of parameters, from 4 million for GoogleNet to 138 million for VGG-16. However, Trident maintains a higher average throughput and therefore improves training latency by 8.5% for MobileNetV2, 15.9% for ResNet-50, and 38.5% for VGG-16. There is a 10.6% increase in training time when using Trident for GoogleNet; however, it only amounts to a 6 second difference.

VI. Related Work

Silicon photonic accelerators capable of in-situ training are part of a new research effort to design the next generation of scalable and energy-efficient processors for NNs. In [14], an MZM mesh is used to perform inference and in-situ training. However, this is not as area-efficient as Trident because it uses large MZMs which take up a lot of area on the chip and are slow, while Trident uses MRRs. Additionally, [14] does not implement a non-linear activation function optically. In [9], a broadcast-and-weight-based architecture is used along with Direct Feedback alignment (DFA) to perform training in photonics. The architecture is evaluated on the MNIST dataset using a two-layer fully connected network implemented on a 1x4 MRR array. However, DFA is not effective for training convolutional layers [35]. The reliance on thermally tuned MRRs also limits the number of MRRs in a weight bank and constrains the bit resolution to a maximum of 6. Since Trident uses PCM to tune MRRs, it can achieve 8-bit resolution for more accurate training and can scale to larger numbers of MRRs to support larger NN models. In [8], an all-optical spiking network is introduced that uses PCM-tuned MRRs and a PCM activation function. However, the approach is limited to positive weights, which means it would not be able to implement asymmetric backpropagation which relies heavily on sign concordance. The ability to implement asymmetric backpropagation is an added advantage of Trident since a main benefit of optics is parallelism. The architectures described above were evaluated on small-scale NN models consisting of just a few layers, whereas Trident has been evaluated with much larger NN models that are used in practice.

TABLE V: Edge Accelerators Time to Train 50,000 Images

NN Model	NVIDIA AGX Xavier	Trident	Percent Change
MobileNetV2	32.5 s	29.7 s	-8.5%
GoogleNet	57.1 s	63.2 s	10.6%
ResNet-50	365.7 s	307.2 s	-15.9%
VGG-16	1293.8 s	796.1 s	-38.5%

VII. Conclusion

This paper proposes a unified photonic accelerator architecture for both training and inference in NNs. The Trident PE utilizes phase change material GST to tune MRRs in the broadcast-and-weight style MRR weight bank to achieve 2× speedup compared to thermally tuned MRR weight banks. We show how both training and inference can be executed on the same PE. We also proposed an LDSU and a GST activation cell for each row of the MRR weight bank. These components facilitate a photonic activation function and its derivative to be stored within the PE, reducing data movement, and eliminating the use of ADCs between PEs while enabling in-situ training. Trident provides up to 43% energy savings and up to 150% reduction in latency on average when compared to the photonic accelerators DEAP-CNN, CrossLight, and PIXEL. When compared to electronic edge AI accelerators Google Coral and Bearkey TB96-AI, Trident saves 11% and 93% energy on average respectively and reduces latency by 1413% and 595% on average respectively. Despite the higher energy efficiency of the NVIDIA AGX Xavier, the reduced data movement and GST activation of Trident improves inference latency by 107% and

training latency by 13% on average compared to the NVIDIA AGX Xavier.

ACKNOWLEDGMENT

This research was partially supported by NSF grants CCF-1901165, CCF-1901192, CCF-1936794, CFF-1953980, CCF-2311543, CCF-2311544, CCF-2324644, CCF-2324645, and CNS-2321224. We sincerely thank the anonymous reviewers for their excellent feedback.

REFERENCES

[1] H. A. Abdelhafez, H. Halawa, K. Pattabiraman, and M. Ripeanu, "Snowflakes at the edge: A study of variability among nvidia jetson agx xavier boards," in *Proceedings of the 4th International Workshop on Edge Systems, Analytics and Networking*, 2021, pp. 1–6.

[2] V. Bangari, B. A. Marquez, H. Miller, A. N. Tait, M. A. Nahmias, T. F. De Lima, H.-T. Peng, P. R. Prucnal, and B. J. Shastri, "Digital electronics and analog photonics for convolutional neural networks (deap-cnns)," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 26, no. 1, pp. 1–13, 2019.

[3] A. Bhardwaj, V. Chauhan, and M. Kumar, "Design of cmos based d flip-flop with different low power techniques," in *2019 6th International conference on signal processing and integrated networks (SPIN)*. IEEE, 2019, pp. 834–839.

[4] W. Bogaerts, P. De Heyn, T. Van Vaerenbergh, K. De Vos, S. Kumar Selvaraja, T. Claes, P. Dumon, P. Bienstman, D. Van Thourhout, and R. Baets, "Silicon microring resonators," *Laser & Photonics Reviews*, vol. 6, no. 1, pp. 47–73, 2012.

[5] X. Chen, Y. Xue, Y. Sun, J. Shen, S. Song, M. Zhu, Z. Song, Z. Cheng, and P. Zhou, "Neuromorphic photonic memory devices using ultrafast, non-volatile phase-change materials," *Advanced Materials*, p. 2203909, 2022.

[6] Q. Cheng, J. Kwon, M. Glick, M. Bahadori, L. P. Carloni, and K. Bergman, "Silicon photonics codesign for deep learning," *Proceedings of the IEEE*, vol. 108, no. 8, pp. 1261–1282, 2020.

[7] S. K. Esser, P. A. Merolla, J. V. Arthur, A. S. Cassidy, R. Appuswamy, A. Andreopoulos, D. J. Berg, J. L. McKinstry, T. Melano, D. R. Barch *et al.*, "From the cover: Convolutional networks for fast, energy-efficient neuromorphic computing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 41, p. 11441, 2016.

[8] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," *Nature*, vol. 569, no. 7755, pp. 208–214, 2019.

[9] M. J. Filipovich, Z. Guo, M. Al-Qadasi, B. A. Marquez, H. D. Morison, V. J. Sorger, P. R. Prucnal, S. Shekhar, and B. J. Shastri, "Silicon photonic architecture for training deep neural networks with direct feedback alignment," *Optica*, vol. 9, no. 12, pp. 1323–1332, 2022.

[10] M. J. Filipovich, Z. Guo, B. A. Marquez, H. D. Morison, and B. J. Shastri, "Training deep neural networks in situ with neuromorphic photonics," in *2020 IEEE Photonics Conference (IPC)*. IEEE, 2020, pp. 1–2.

[11] A. Frumusanu. (2019) Investigating nvidia's jetson agx: A look at xavier and its carmel cores. [Online]. Available: https://www.anandtech.com/show/13584/nvidia-xavier-agx-hands-on-carmel-and-more/2

[12] Google. (2020) Dev board datasheet — coral. [Online]. Available: https://coral.ai/docs/dev-board/datasheetdocument-revisions

[13] P. Guo, A. M. Sarangan, and I. Agha, "A review of germanium-antimony-telluride phase change materials for non-volatile memories and optical modulators," *Applied sciences*, vol. 9, no. 3, p. 530, 2019.

[14] T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, "Training of photonic neural networks through in situ backpropagation and gradient measurement," *Optica*, vol. 5, no. 7, pp. 864–871, Jul 2018.

[15] H. Jung, K. Y. Fong, C. Xiong, and H. X. Tang, "Electrical tuning and switching of an optical frequency comb generated in aluminum nitride microring resonators," *Optics letters*, vol. 39, no. 1, pp. 84–87, 2014.

[16] S. Kumar and A. Kumar, "A brief review on antlion optimization algorithm," in *2018 international conference on advances in computing, communication control and networking (ICACCCN)*. IEEE, 2018, pp. 236–240.

[17] D. Kuzum, R. G. Jeyasingh, B. Lee, and H.-S. P. Wong, "Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing," *Nano letters*, vol. 12, no. 5, pp. 2179–2186, 2012.

[18] Y.-L. Lee, P.-K. Tsung, and M. Wu, "Techology trend of edge ai," in *2018 International Symposium on VLSI Design, Automation and Test (VLSI-DAT)*. IEEE, 2018, pp. 1–2.

[19] K. Li, S. Liu, X. Ruan, D. J. Thomson, Y. Hong, F. Yang, L. Zhang, C. Lacava, F. Meng, W. Zhang, P. Petropoulos, F. Zhang, and G. T. Reed, "Co-design of a differential transimpedance amplifier and balanced photodetector for a sub-pj/bit silicon photonics receiver," *Opt. Express*, vol. 28, no. 9, pp. 14 038–14 054, Apr 2020.

[20] 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.

[21] H. Liang, R. Soref, J. Mu, A. Majumdar, X. Li, and W.-P. Huang, "Simulations of silicon-on-insulator channel-waveguide electrooptical 2×2 switches and 1×1 modulators using a $ge_2sb_2te_5$ self-holding layer," *Journal of Lightwave Technology*, vol. 33, no. 9, pp. 1805–1813, 2015.

[22] L. Limited. (2020) Tb-96ai. [Online]. Available: https://www.96boards.org/product/tb-96ai/

[23] W. Liu, W. Liu, Y. Ye, Q. Lou, Y. Xie, and L. Jiang, "Holylight: A nanophotonic accelerator for deep learning in data centers," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2019, pp. 1483–1488.

[24] A. Mehrabian, Y. Al-Kabani, V. J. Sorger, and T. El-Ghazawi, "Pcnna: A photonic convolutional neural network accelerator," in *2018 31st IEEE International System-on-Chip Conference (SOCC)*. IEEE, 2018, pp. 169–173.

[25] A. Narayan, Y. Thonnart, P. Vivet, A. Coskun, and A. Joshi, "Architecting optically controlled phase change memory," *ACM Transactions on Architecture and Code Optimization*, vol. 19, no. 4, pp. 1–26, 2022.

[26] Nvidia. (2021) Jetson tk1: Modile embedded supercomputer takes cuda everywhere – nvidia developer blog. [Online]. Available: https://developer.nvidia.com/blog/

[27] H.-T. Peng, M. A. Nahmias, T. F. De Lima, A. N. Tait, and B. J. Shastri, "Neuromorphic photonic integrated circuits," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 6, pp. 1–15, 2018.

[28] G. Römer and P. Bechtold, "Electro-optic and acousto-optic laser beam scanners," *Physics procedia*, vol. 56, pp. 29–39, 2014.

[29] K. Seshadri, B. Akin, J. Laudon, R. Narayanaswami, and A. Yazdanbakhsh, "An evaluation of edge tpu accelerators for convolutional neural networks," in *2022 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, 2022, pp. 79–91.

[30] K. Shiflett, D. Wright, A. Karanth, and A. Louri, "Pixel: Photonic neural network accelerator," in *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2020, pp. 474–487.

[31] 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.

[32] A. N. Tait, T. F. De Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," *Scientific reports*, vol. 7, no. 1, pp. 1–10, 2017.

[33] A. R. Totović, G. Dabos, N. Passalis, A. Tefas, and N. Pleros, "Femtojoule per mac neuromorphic photonics: An energy and technology roadmap," *IEEE Journal of selected topics in Quantum Electronics*, vol. 26, no. 5, pp. 1–15, 2020.

[34] N. Wang, J. Choi, D. Brand, C.-Y. Chen, and K. Gopalakrishnan, "Training deep neural networks with 8-bit floating point numbers," *Advances in neural information processing systems*, vol. 31, 2018.

[35] M. B. Webster, J. Choi *et al.*, "Learning the connections in direct feedback alignment," 2020.

[36] G. Yang, C. Demirkiran, Z. E. Kizilates, C. A. R. Ocampo, A. K. Coskun, and A. Joshi, "Processing-in-memory using optically-addressed phase change memory."

[37] H. Zhang, L. Zhou, J. Xu, L. Lu, J. Chen, and B. Rahman, "Silicon microring resonators tuned with gst phase change material," in *2018 Asia Communications and Photonics Conference (ACP)*. IEEE, 2018, pp. 1–3.

[38] H. Zhou, J. Dong, J. Cheng, W. Dong, C. Huang, Y. Shen, Q. Zhang, M. Gu, C. Qian, H. Chen *et al.*, "Photonic matrix multiplication lights up photonic accelerator and beyond," *Light: Science & Applications*, vol. 11, no. 1, p. 30, 2022.