A Hybrid Optical-Electrical Analog Deep Learning Accelerator Using Incoherent Optical Signals

Mingdai Yang
University of Chicago

Mohammad Reza Jokar
University of Chicago

Junyi Qiu
University of Illinois Urbana-Champaign

Qiuwen Lou
University of Notre Dame

Yuming Liu
Tsinghua University

Aditi Udupa
University of Illinois Urbana-Champaign

Frederic T. Chong
University of Chicago

John M. Dallesasse
University of Illinois Urbana-Champaign

Milton Feng
University of Illinois Urbana-Champaign

Lynford L. Goddard
University of Illinois Urbana-Champaign

X. Sharon Hu
University of Notre Dame

Yanjing Li
University of Chicago

Abstract

We present a hybrid optical-electrical analog deep learning (DL) accelerator, the first work to use incoherent optical signals for DL workloads. Incoherent optical designs are more attractive than coherent ones as the former can be more easily realized in practice. However, a significant challenge in analog DL accelerators, where multiply-accumulate operations are dominant, is that there is no known solution to perform accumulation using incoherent optical signals. We overcome this challenge by devising a hybrid approach: accumulation is done in the electrical domain, while multiplication is performed in the optical domain. The key technology enabler of our design is the transistor laser, which performs electrical-to-optical and optical-to-electrical conversions efficiently to tightly integrate electrical and optical devices into compact circuits. As such, our design fully realizes the ultra high-speed and high-energy-efficiency advantages of analog and optical computing. Our evaluation results using the MNIST benchmark show that our design achieves 2214× and 65× improvements in latency and energy, respectively, compared to a state-of-the-art memristor-based analog design.

CCS Concepts

• **Hardware → Emerging optical and photonic technologies**.

Keywords

optical computing, deep learning accelerator

ACM Reference Format:
Mingdai Yang, Mohammad Reza Jokar, Junyi Qiu, Qiuwen Lou, Yuming Liu, Aditi Udupa, Frederic T. Chong, John M. Dallesasse, Milton Feng, Lynford L. Goddard, X. Sharon Hu, and Yanjing Li. 2021. A Hybrid Optical-Electrical Analog Deep Learning Accelerator Using Incoherent Optical Signals. In *Proceedings of the Great Lakes Symposium on VLSI 2021 (GLSVLSI '21), June 22–25, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3453688.3461531

1 Introduction

Deep learning (DL) accelerators are widely deployed in various application domains, and are expected to become more prominent. A number of emerging technologies (e.g., RRAM, FeFET) have been exploited to implement DL accelerators. Among them, optical DL accelerators implemented using coherent optical signals have attracted significant interest due to the following properties [14]: (1) matrix multiplication consumes no power; (2) the latency of optical compute/communication is very low; and (3) non-linear optical devices can be used to implement activation functions at low cost.

However, in practice it is challenging to compute with coherent optical signals due to their sensitivity to phase: even a slight shift in the optical phase (as a result of process and temperature variations) will affect the way splitters/combiners/modulators and other optical devices work, resulting in poor computation accuracy. For example, in a recent coherent optical DL accelerator [14], only 50% accuracy is achieved for a simple network with 2 layers, 4 inputs, and 4 outputs due to phase errors (vs. ∼ 76% without any phase error).

In this work, we focus on computing using incoherent optical signals, which does not require the phases to be aligned, thereby avoiding the major limitation of computing using coherent optical signals. Moreover, all the advantages of a coherent optical DL accelerator still apply, including low latency/energy and low-cost implementation of non-linear activation functions. However, currently there is no known solution to perform accumulation using incoherent optical signals.

To this end, we introduce a novel hybrid optical-electrical approach to implement multiply-accumulate (MAC) operations, the core compute primitive in all DL workloads, where accumulation is performed in the electrical domain, while multiplication is performed in the optical domain. Since addition/subtraction can be performed efficiently by simply joining two wires with currents flowing in the same/opposite directions in the electrical domain, this hybrid approach allows different operations to be performed in the domain that is best-suited for the corresponding operation.

GLSVLSI '21, June 22–25, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8393-6/21/06…$15.00
https://doi.org/10.1145/3453688.3461531

In our design, the conversions between electrical and optical domains – i.e., EtoO/OtoE conversions – must be performed in an ultra-efficient manner. Otherwise, the cost of EtoO/OtoE conversions can outweigh the benefits of the hybrid approach. Fortunately, an emerging technology called Transistor Laser (TL) [9, 16] exists for this exact purpose, which enables electrical and optical devices to be tightly integrated with ultra-high speed. As a result, our hybrid approach obtains the major advantages of optical and analog computing by paying only small costs for EtoO/OtoE conversions.

We apply our hybrid DL accelerator design approach to a 4-layer multilayer perceptron network for classifying the widely-used MNIST dataset as a proof-of-concept. We accurately model the latency, power, and area of our design using detailed device parameters and HSPICE simulations. Our results show that the latency and energy of our accelerator are 2.91 ns and 47.52 nJ, respectively. Compared to a state-of-the-art memristor-based accelerator running the same workload, our design provides a 2214× improvement in latency, and a 65× improvement in total energy. Such dramatic improvements are obtained since it is practical to build relatively large MAC compute units (e.g., compared to memristor crossbars) with high speed, which significantly reduces the need to store/re-fetch intermediate results to/from memories – an inefficient step that is typically required in other designs. We also show that the impact of noises/variations on the accuracy of our design is minimal ($< 0.25\%$).

The major contributions of this paper are the following. (1) We motivate and introduce the first hybrid optical-electrical analog DL accelerator using incoherent optical signals. (2) We present the detailed design of the hybrid accelerator. (3) We perform thorough evaluation to demonstrate that our idea is highly efficient.

2 Technology Background and Related Work

In this section, we discuss the background and related work on the key components of our hybrid design approach.

2.1 Using Incoherent Optical Signals to Perform MAC Operations

In existing optical DL accelerator designs [4, 6, 8, 14], a common approach to perform matrix multiplication is to send coherent optical signals through a network of Mach–Zehnder interferometers (MZIs), which consumes no power. However, as discussed in Sec. 1, the accuracy is extremely poor, which renders these designs infeasible in practice. Moreover, the MZIs are large and can adversely affect the latency, area, and scalability of these DL accelerator designs. Thus, we focus on incoherent optical signals instead.

To compute using incoherent optical signals, there exist low-cost optical devices – such as passive splitters which consume no power [1] and silicon nitride ring resonators with ultra-low (0.05 nW) leakage power [15] – that can be exploited to effectively perform multiplication operations. However, using incoherent optical signals for accumulation presents two significant challenges. First, subtraction operations are tricky because negative values cannot be represented using optical intensities (this is true for coherent optical signals as well). Second, for addition, simply combining two streams of incoherent optical signals does not work (unlike coherent optical signals), because combiners, which are the reverse of splitters, are passive reciprocal devices that work with the amplitude and phase of the electric field. Suppose a passive combiner could add two arbitrary incoherent optical signals, then the same device operating in reverse would be able to correctly split the power back. However, without a specific phase relationship of the signals, this is not possible given the lack of uniqueness of this inverse problem. We also explored combining optical signals of different wavelengths into a single waveguide. The fundamental problem is that we need many (e.g., hundreds of) wavelengths for a typical DL workload, which requires large optical devices (on the order of 300 um), and therefore, incurs prohibitive area and delay overhead. In short, currently there is no practical solution to perform add/subtract operations using incoherent optical signals.

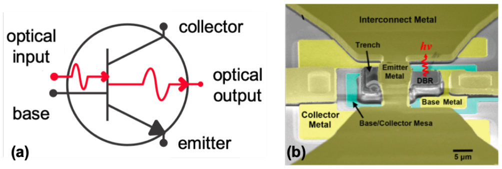


Figure 1: The transistor laser (TL). (a) Schematic. (b) SEM top view of a fabricated device prototype.

In contrast, addition and subtraction can be performed efficiently in the electrical domain by simply joining two wires with currents flowing in the same or opposite directions. This motivates us to devise a hybrid MAC unit, where multiplication/accumulation operations are performed in the respective optical/electrical domain that is best suited for each type of operations.

2.2 Using Transistor Lasers to Perform Efficient OtoE/EtoO Conversions

To support hybrid MAC operations, OtoE and EtoO converters are essential. Recent advances in optical technology have led to ultra-fast and low-cost converters – the transistor lasers (TLs).

A TL (depicted in Fig. 1) is an InGaP/GaAs heterojunction bipolar transistor (HBT) with the addition of quantum-wells for photon generation and optical cavity for optical output. When an electrical current is applied to the base, and if the collector-emitter voltage is higher than a threshold, a proportional optical output is generated and the TL performs EtoO conversion. When an optical input hits the base-collector junction, a photocurrent is generated and the TL functions as a photodetector (PD) to perform OtoE conversion.

Previous generations of TLs have been fabricated on GaAs substrates with minimal changes to the existing HBT flows [7, 16]. Detailed simulation and analysis results for the next generation of TLs are provided in Table 1. For EtoO conversions, TL's modulation rate is significantly higher than other lasers (e.g., VCSELs) because its spontaneous recombination lifetime is orders of magnitude lower [16], and the latency is only 1.93 ps. However, the power consumption is 2.5 mW. Therefore, we take careful consideration of the EtoO conversion power in our design so that the overall energy is minimized. For OtoE conversions, TLs are similar to PIN PDs [16], with a small latency of 1.93 ps and low power consumption of 3.7 uW. However, if TLs are already used for EtoO conversions, using them for OtoE conversions as well will reduce manufacturing complexity since the device structures are uniform.

Given TL's ultra-efficient EtoO/OtoE conversion capabilities, our hybrid optical-electrical approach is viable and promising.

Table 1: TL parameters [10]

Device parameters	latency=1.93 ps; data rate/frequency=60Gbps; wavelength=980 nm; voltage=2 V; junction capacitance=1 fF (with inline waveguides).
EtoO conversion parameters	power=2.5 mW; EtoO efficiency=0.4; max optical output power=1 mW.

OtoE conversion parameters	power=3.665 uW; responsivity=0.4 A/W;
	sensitivity=0.2uW.

3 Design of a Hybrid DL Accelerator

In this section, we present the details of our hybrid optical-electrical analog DL accelerator design.

3.1 Hybrid MAC Operations

A MAC operation consists of both multiplication and accumulation operations. To perform multiplication using incoherent optical signals, different intensity values of an optical signal (as measured in optical power) can be used to represent the value of a multiplicand. If the multiplier's value is ≤ 1, then a passive splitter [1] or a ultra-low-power ring resonator [15] can be used to perform multiplication by splitting an optical signal into two separate streams with a pre-specified split ratio. For example, if the multiplier is 0.3, then a split ratio of 30%:70% yields an output that is 0.3× the multiplicand. On the other hand, if the multiplier's value is > 1, a tunable optical amplifier is required, which consumes high power. Fortunately, for a DL workload, a multiplier corresponds to a weight value, and the weight values are already ≤ 1 in many representative DL workloads. If this is not the case, we can always retrain a neural network to make all weight values ≤ 1. Thus, multiplication can be performed in the optical domain with no or negligible power cost. Moreover, the split ratio of a ring resonator or a splitter (built using phase-change materials [11]) can be tuned, which means that we can program the weight values in our design to run different DL workloads. Note that, applying splitters/ring resonators for optical computing is a new contribution of our work.

Regarding accumulation, as elaborated in Sec. 2.1, addition and subtraction present significant challenges when incoherent optical signals are used, but can be achieved in the electrical domain by simply joining two wires. Therefore, MAC operations can be most efficiently executed when multiplication is performed in the optical domain and accumulation is performed in the electrical domain. In our hybrid MAC implementation, electrical inputs are first converted to optical signals using TLs, and then transmitted via waveguides to splitters/ring resonators, which perform the multiplication operations. Each {input, weight} product is converted back to the electrical domain using a PD (or a TL configured for OtoE conversion) to participate in accumulation.

In our design, the accumulation of all {input, weight} products are performed entirely in the electrical domain to minimize cross-domain conversions, which in turn minimizes power consumption and complexity. We have explored other partitioning schemes – e.g., first converting several partial sums back to optical signals, and then performing another OtoE conversion to calculate the final sum, which achieves higher communication speed because optical propagation delay is smaller than electrical wire delay. However, these alternatives impose much higher power/energy costs. Therefore, the partitioning scheme in our design achieves optimized tradeoffs between power, latency, and design complexity.

3.2 Hybrid Optical-Electrical Crossbar Design

In this section, we present the detailed hybrid circuit structure for MAC operations by first illustrating our approach using a small fully-connected (FC) layer example, and then elaborating on the generalized design. These circuits are formed as 2D crossbars similar to many RRAM-based designs (e.g., [13]); however, the underlying compute mechanisms are entirely different as our crossbar enables hybrid optical-electrical MAC operations.

FC Layer Example. Figure 2a shows a FC layer example, and the corresponding circuit structure is shown in Fig. 2b. Each column of the circuit derives one output of this FC layer. Each input is connected to a different row of the crossbar, and is distributed to all outputs using splitters, so that the three outputs can be calculated at the same time. We specify the evolution of the analog values along the first column of the circuit to clearly illustrate how the outputs are derived.

In this example, since there are three outputs, each input is split into 3 equal streams using the *distribution splitters*. Each split input signal (e.g., v1 in Fig. 2b) then goes through another splitter called the *weight splitter*, which outputs the product of the split input and the weight value (e.g., v2). Next, each {split input, weight} product is converted to an electrical signal (v3) to participate in the analog accumulation operation (v4). Since weights can be positive or negative, the output of a weight splitter is connected to two PDs with opposite polarities (Fig. 2c). The sign bit of the weight is used to control the switches connected to the PDs: when PD+ is connected, current will flow downward along a copper wire to add to the final result; otherwise, current will flow upward and be subtracted from the accumulation result. The final accumulation result passes through an electrical current amplifier (*E_amplifier*) (v5), which is used to compensate for: (1) the input split fraction; (2) EtoO/OtoE conversion efficiency losses (*EtoO_loss* and *OtoE_loss*); and (3) other optical losses (*coupling_loss* and *splitter_loss*). The E_amplifier also performs the ReLU function, one of the most common activation functions in deep neural networks (DNNs), because a positive input current will result in a proportional amplified output, while a negative input will generate no current at the output. Finally, the output of each amplifier is passed through a TL to generate an optical output of this layer (e.g., out$_1$). Note that, the output values are bounded by the maximum power value that a TL can generate (1 mW in our current design). If the actual value exceeds this bound, it simply represents an overflow situation where the value is truncated, similar to digital implementations.

Generalized Structure. A general crossbar design with M inputs and N outputs is shown in Fig. 3. The underlying idea is similar to the FC layer example. One difference is that, if the number of outputs is large (e.g., >= 4), then one or more trees of splitters are used to distribute one input to participate in the computations of multiple outputs. This is necessary to achieve balanced splitter variations across all columns. Otherwise, if serial distribution splitters are used (as shown in Figure 2b), values distributed to farther columns will encounter higher variations that are cumulative of all previous splitters, and such a cumulative effect is severe because the number of outputs in a DL workload is typically quite large.

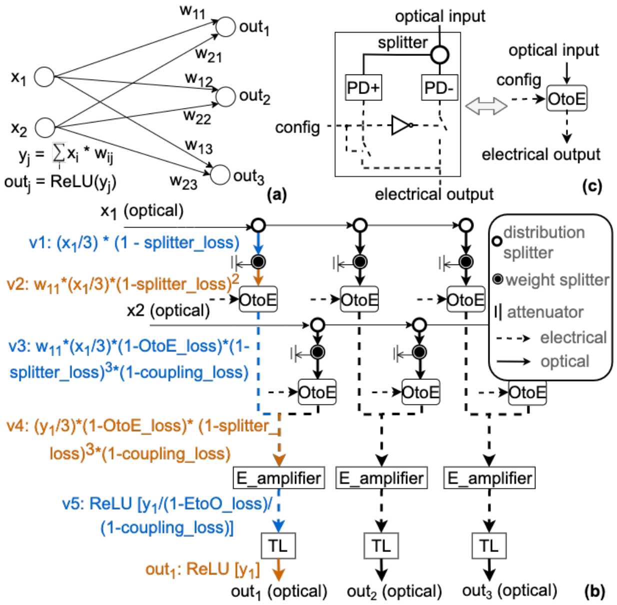


Figure 2: (a) A FC layer example. (b) The hybrid MAC circuit structure corresponding to (a). (c) The *OtoE* block in (b).

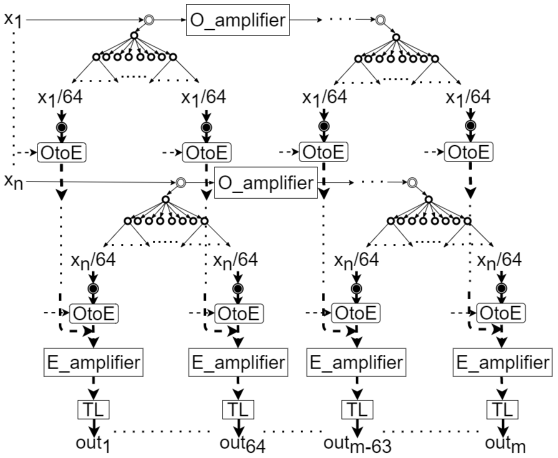


Figure 3: The generalized hybrid MAC crossbar design (same legend as Fig. 2b). The TLs are not needed in the last layer.

The largest number of splits that can be applied to a single input is limited by the PD's sensitivity – i.e., the minimal optical input power required for the PD to correctly sense the data. In our design, the maximum number of input splits is 64, which ensures that the smallest absolute values of all possible {split input, weight} products are greater than the PD's sensitivity requirement.

If the number of outputs is greater than 64, additional amplifiers (the *O_amplifiers* in Fig. 3) are inserted so that the inputs can be distributed to the rest of the crossbar columns. The O_amplifier consists of a PD (or TL) for OtoE conversion, an E_amplifier, and a TL for EtoO conversion.

3.3 Whole Accelerator Design

Multiple hybrid MAC circuits are integrated to implement a DL accelerator (Fig. 4). The most efficient implementation is to use one crossbar-like structure to calculate the outputs of one DNN layer.

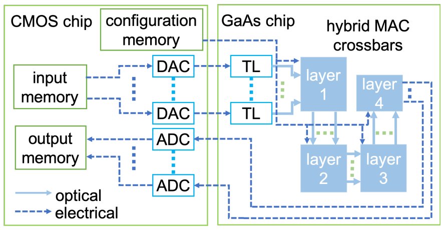


Figure 4: Overview of our hybrid DL accelerator design.

Then the outputs of a crossbar are routed to the inputs of the next crossbar to execute multiple layers of a DNN. Moreover, to minimize routing complexity, the crossbars can be oriented as shown in Fig. 4 so that the waveguide distances between the crossbars implementing any adjacent DNN layers are minimized. If it is not possible to use one hybrid MAC crossbar to implement each DNN layer (e.g., due to cost constraints), intermediate results of the crossbars need to be stored/re-fetched to/from memory. Fortunately, given the fast computation speed, and since the main elements in the hybrid MAC circuit are passive splitters/low-power ring resonators and low-cost PDs, it is practical to build large hybrid MAC crossbars to minimize memory access overheads.

The input interface of the accelerator consists of the input memory which stores the inputs of a DL workload, DACs (digital-to-analog converters), and TLs to convert electrical inputs to optical signals. The outputs of the accelerator drive the ADCs (analog-to-digital converters) to generate digital signals, which are subsequently written back to the output memory. The weight values of a DNN (stored in the configuration memory) control the split ratios of weight splitters/ring resonators and the switches of the PDs.

Similar to DL accelerators implemented using other emerging technologies (e.g., RRAM), the weight values of a trained DNN need to be configured into the cross-point elements only once. In our design, the configuration overhead is small (e.g., it takes <4 ns to configure multiple elements and consumes 0.5 pJ per element if the ring resonators in [15] are used). Moreover, if a hybrid MAC crossbar can be dedicated to each DNN layer, such configuration costs can be amortized across a large number of inference tasks.

Regarding the physical design, the hybrid accelerator consists of a CMOS chip containing SRAMs, DACs, and ADCs, and a GaAs chip containing TLs, splitters, and E_amplifiers (see Fig. 4). CMOS and GaAs chips can be bonded through hybrid integration [5].

4 Evaluation Methodology and Results

To evaluate our hybrid accelerator approach, we compose a proof-of-concept design based on a 4-layer multilayer perceptron (MLP) network for classifying the widely-used MNIST dataset. This network consists of one 784×256 input layer, two 256×256 hidden layers, and one 256×10 output layer. Correspondingly, our accelerator consists of four MAC crossbars (one per layer). This design operates on reduced-precision fixed-point numbers, which is common for DL accelerators. The input and weight values are quantized to 16 and 8 levels, respectively. The activation function is ReLU.

We obtain accurate latency, power, and energy results for our design, and quantitatively compare our accelerator against a state-of-the-art memristor-based design. Moreover, since analog computing is subject to noises/variations (in both optical and electrical

Table 2: Device and circuit parameters.

Waveguide/splitter parameters [1, 3]

1:64 splitter tree_height/width=66 um per 2 inputs/220 um[1];
waveguide_propagation_speed=1.763e5 um/ns;
longest waveguide delay between two layers=21.6 ps;
waveguide propagation loss=0.3 dB/cm; splitter loss=0.2 dB;
coupling loss=0.5 dB; bending radius=30 um;

Electrical component parameters from HSPICE simulations

E_amplifier_latency[2]=7.72 ps; E_amplifier_power=1.4 mW;
E_amplifier_max_gain=800×;
wire_propagation_speed=2.22e4 um/ns[3];
DAC_power/latency = 0.267 mW/88.5 ps;
ADC_power/latency = 1.2 mW/294 ps.

[1] We design the splitter tree using pixelated topology optimization in which we minimize a loss function by varying which pixels are Si_3N_3 or SiO_2. The selected pattern affects how light couples from the single input waveguide into intermediate modes and finally to the outputs, thereby enabling us to achieve a more compact splitter than a conventional multi-mode interferometer. Moreover, for a design with two waveguide layers, half of the splitter trees are placed in each layer to reduce the vertical distance of the crossbar.

[2] The E_amplifier is implemented using 4 GaAs bipolar junction transistors, and outputs a maximum current of 1.25 mA (which, when serves as an input to a TL, corresponds to an optical output with the maximum optical power of 1 mW).

[3] Because the splitter trees occupy a large area, we are able to increase the width of the wire to 20 um to optimize for propagation speed for wire length of up to 10 mm, while accounting for scattering and skin effects.

domains), we develop a software simulation framework to assess the accuracy of our hybrid accelerator.

4.1 Latency/Power/Energy Models and Results

We derive accurate latency, area, power, and energy results of our design using device/circuit parameters obtained from detailed device/HSPICE simulations, as summarized in Tables 1 and 2.

Latency/Area. The total execution latency is obtained by summing up the latencies of all crossbars, inter-crossbar latencies, and the input/output interface latencies. The latency of each crossbar is bounded by the delay for a signal to travel from the first input to the last output, and can be estimated using Equations (1)-(3).

$$\mathit{crossbar\_latency} = \mathit{horizontal\_latency} + \mathit{vertical\_latency} \tag{1}$$

$$\begin{aligned} \mathit{horizontal\_latency} &= (\mathit{tree\_width} \times \mathit{num\_columns}/64)/ \\ \mathit{waveguide\_propagation\_speed} &+ (\lceil \mathit{num\_columns}/64 \rceil - 1) \times \\ &(\mathit{TL\_latency} + \mathit{E\_amplifier\_latency} + \mathit{PD\_latency}) \end{aligned} \tag{2}$$

$$\begin{aligned} \mathit{vertical\_latency} &= \mathit{PD\_latency} + \mathit{E\_amplifier\_latency} + \\ \mathit{TL\_latency} &+ \mathit{tree\_height}/\mathit{waveguide\_propagation\_speed} + \\ &(\mathit{tree\_height} \times (\mathit{num\_rows}/2 - 1))/\mathit{wire\_propagation\_speed} \end{aligned} \tag{3}$$

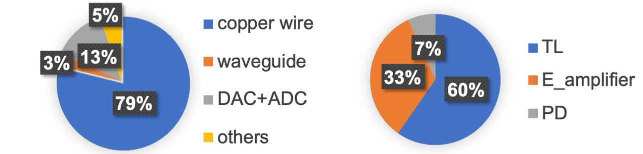
Thanks to the fast computation speed and efficient inter-crossbar data communication, an MNIST inference task takes only 2.91 ns in our hybrid design, as shown in Fig. 5a and Fig. 5b. We can see that the electrical wire propagation delay dominates the overall latency, even though the wire geometry has already been scaled to maximize propagation speed (see Table 2). However, as discussed in Sec. 3.1, using electrical wires throughout still yields the best energy efficiency. Moreover, all peripheral electrical circuitry (SRAMs, DACs, and ADCs) only needs to operate at a clock cycle time of 2.91 ns, so they do not impose latency bottlenecks.

The horizontal and vertical dimensions of each crossbar provided in the delay model are used to determine the total area of each crossbar. The four crossbars together occupy an area of 39.6mm^2.

Power/Energy. The active and inactive power of a hybrid crossbar circuit is estimated using Equations (4) and (5), respectively. Note that, when the TLs are configured as lasers to perform EtoO conversions, they cannot be power-gated off since it takes a long time for them to be turned on. The power results and breakdown are summarized in Fig. 5a and Fig. 5c.

	layer 1	layer 2 / 3	layer 4	input/output interfaces	inter-crossbar	overall
Latency (ns)	1.22	0.43	0.39	0.38	0.06	2.91
Active power (W)	12.88	4.24	0.02	2.25	0	16.33 (avg. power)
Inactive power (W)	8.48	2.56	0	2.04	0	

(a) Latency/power breakdown per operation



(b) Latency breakdown per component type

(c) Layer 1 crossbar active power breakdown per component type

Figure 5: Detailed latency and power results.

Combining the latency and power results, the energy of an inference task for the MNIST MLP network is 47.52 nJ, calculated using Equation (6). In addition to the energy of the hybrid MAC crossbars (which accounts for the majority of the total energy consumption), we also include the energy consumption of DACs/ADCs (derived using the HSPICE simulation results reported in Table 2) and the energy to fetch inputs and store outputs from/to SRAMs (which is only 0.22 nJ per inference task, derived using CACTI [2]).

$$\begin{aligned} \mathit{active\_crossbar\_power} &= \mathit{max\_PD\_power} \times \mathit{num\_columns} \times \\ \mathit{num\_rows} &+ (\mathit{E\_amplifier\_power} + \mathit{TL\_power}) \times \mathit{num\_columns} + \\ &(\mathit{max\_PD\_power} + \mathit{E\_amplifier\_power} + \mathit{TL\_power}) \times \\ &\mathit{num\_rows} \times (\lceil \mathit{num\_columns}/64 \rceil - 1) \end{aligned} \tag{4}$$

$$\begin{aligned} \mathit{inactive\_crossbar\_power} &= \mathit{TL\_power} \times (\mathit{num\_columns} + \\ &\mathit{num\_rows} \times (\lceil \mathit{num\_columns}/64 \rceil - 1)) \end{aligned} \tag{5}$$

$$\begin{aligned} \mathit{Total\_energy} = &\sum_{\mathit{layer}} [\mathit{inactive\_crossbar\_power}(\mathit{layer}) \times (2.91\ \mathit{ns} - \\ &\mathit{crossbar\_latency}(\mathit{layer})) + \mathit{active\_crossbar\_power}(\mathit{layer}) \times \\ &\mathit{crossbar\_latency}(\mathit{layer})] + \mathit{ADC\_power} \times \mathit{ADC\_latency} \times 10 + \\ &\mathit{DAC\_power} \times \mathit{DAC\_latency} \times 784 + \mathit{TL\_power} \times 2.91\ \mathit{ns} \times 784 + \\ &\mathit{energy\_to\_fetch\_inputs} + \mathit{energy\_to\_store\_outputs} \end{aligned} \tag{6}$$

Quantitative Comparison. We perform quantitative comparison between our design and an ISAAC-like memristor-based DL accelerator [13] designed for the same 4-layer MLP network. The memristor-based DL accelerator consists of: (1) 22 128×128 crossbars (occupying an area of 3.1mm^2), which allow all weights to be mapped so there is no need to reprogram; (2) SRAMs for storing inputs, partial results generated by the crossbars, and final outputs; (3) an H-tree structure for connecting the crossbars and SRAMs as well as analog repeaters for reducing H-tree wire latency; and (4) peripheral components. The latency, power, and area of the crossbars are derived based on the RRAM devices in [17]. Any inactive crossbars are power-gated off to minimize power consumption. To derive the total latency and power due to SRAM accesses, we obtain the per-access latency and power using CACTI [2], and combine them with the total access count obtained using EvaNN [12]. We estimate the power and latency of the H-tree interconnects, the repeaters, and other peripheral components with HSPICE simulation.

The comparison results are summarized in Table 3. Our hybrid accelerator achieves a 2214× improvement in latency. Thus, even

Table 3: Quantitative comparison results.

	latency (ns)		energy (nJ)	
	our design	memristor design	our design	memristor design
MAC crossbar	2.47	30.47	46.40	16.47
inter-crossbar data communication	0.06	6411.34	0.88	3063.20
peripheral	0.38	1.53	0.24	2.21
total	2.91	6443.34	47.52	3081.88
	2214x difference		65x difference	

though our design incurs higher power cost, it is still able to achieve a 65× improvement in energy.

Note that, in current memristor-based designs, the majority of the latency and energy are spent on inter-crossbar communication due to two limitations. First, RRAM crossbar sizes are rather limited (typically no larger than 128×128) since larger ones are more susceptible to device variations. Consequently, the computations of one layer of the MLP network must be carried out using multiple crossbars. Second, to minimize wire complexity and overhead, the results of one crossbar array are typically first written to SRAMs, and then read by the subsequent crossbars [13]. In contrast, the size of our hybrid crossbars can be tuned for a given workload (as discussed in Sec. 3.3). Larger crossbars simply incur longer delays but are not much more susceptible to variations, because the number of the E_amplifiers, which incur the largest amount of noises based on HSPICE results, is fixed at one per column. Moreover, the outputs of one hybrid crossbar can be *directly* routed to the inputs of another by leveraging fast inter-layer communication in the optical domain. Therefore, our design is able to achieve significantly lower latency and higher energy efficiency than memristor-based designs.

4.2 Noise/Variation Models and Results

In addition to optical losses which we have accounted for by factoring them into the amplification ratio of the E-amplifiers, our design is also subject to noises and variations (e.g., manufacturing variations in splitters, noises in amplifiers, temperature variations, and so on). To evaluate their impact on DL workload accuracy, we modify the forward propagation steps in TensorFlow and Keras to model various noises/variations, and perform Monte Carlo simulations which consist of 3 million experiments. In each simulation, we vary each weight value (to model splitter variations) by Δ_1 and each accumulation result (to model the cumulative variations of TLs, PDs, wires, and E_amplifiers) by Δ_2, where Δ_1 and Δ_2 are random variables uniformly distributed on $[-x, x]$. Our accelerator is designed with a x=5% target. In addition, we also report x=3% and 7% to show sensitivity. Our results (Fig. 6) show that the degradation of accuracy due to noises/variations is minimal (< 0.25%).

We also obtain the MNIST workload accuracy of the memristor-based design (introduced in Sec. 4.1) using a PyTorch framework that simulates memristor crossbar structures and assuming the same input/weight quantization levels as our design. The accuracy of the memristor-based design ranges from 97.54% (for 7% RRAM device variations) to 97.87% (with no device variation), which is comparable to our accuracy results.

5 Conclusions

In this paper, we present a hybrid optical-electrical DL accelerator, the first work that allows incoherent optical signals to be used for performing analog MAC operations. Grounded by the TL's ultra-efficient EtoO/OtoE conversion capabilities, our idea is to perform the accumulation operations, which cannot be performed using incoherent optical signals, in the electrical domain instead. As a result, our design is not only realizable in practice, but also achieves unprecedented latency and energy improvements over existing DL accelerators. The highly promising results of this work encourage us to investigate new design approaches that will allow incoherent optical signals to be used to compute a wide variety of DL workloads.

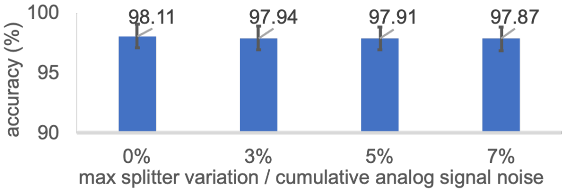


Figure 6: Average MNIST accuracy vs. % of noises/variations.

Acknowledgments

We thank all anonymous reviewers for their comments. This work is sponsored in part by NSF grants 1640192 and 1405959, and also E2CDA-NRI, a funded center of NRI, a Semiconductor Research Corporation (SRC) program sponsored by NERC and NIST.

References

[1] A. Arbabi et al. 2011. Realization of a narrowband single wavelength microring mirror. *Applied Physics Letters* 99, 9 (2011).

[2] Rajeev Balasubramonian et al. 2017. CACTI 7: New Tools for Interconnect Exploration in Innovative Off-Chip Memories. *ACM Trans. Archit. Code Optim.* 14, 2 (2017).

[3] D. J. Blumenthal et al. 2018. Silicon Nitride in Silicon Photonics. *Proc. IEEE* 106, 12 (Dec 2018), 2209–2231.

[4] Julian Bueno et al. 2018. Reinforcement learning in a large-scale photonic recurrent neural network. *Optica* 5, 6 (2018), 756–760.

[5] John Carlson et al. 2019. Epitaxial Bonding and Transfer Processes for Large-Scale Heterogeneously Integrated Electronic-Photonic Circuitry. *J. Electrochem. Soc.* 166, D3158 (2019).

[6] Julie Chang et al. 2018. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Scientific reports* 8, 1 (2018), 12324.

[7] Milton Feng et al. 2017. Resonance-free optical response of a vertical cavity transistor laser. *Applied Physics Letters* 111, 12 (2017).

[8] Tyler W Hughes et al. 2018. Training of photonic neural networks through in situ backpropagation and gradient measurement. *Optica* (2018).

[9] M. R. Jokar et al. 2019. Direct-modulated optical networks for interposer systems. In *Proceedings of the 13th IEEE/ACM International Symposium on Networks-on-Chip*.

[10] M. R. Jokar et al. 2020. Baldur: A Power-Efficient and Scalable Network Using All-Optical Switches. In *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. 153–166.

[11] Haibo Liang et al. 2015. Electro-optical phase-change 2x2 switching using three- and four-waveguide directional couplers. *Appl. Opt.* (Jul 2015).

[12] I. Palit et al. 2019. A Uniform Modeling Methodology for Benchmarking DNN Accelerators. In *2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. 1–7.

[13] A. Shafiee et al. 2016. ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars. In *ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*.

[14] Yichen Shen et al. 2017. Deep learning with coherent nanophotonic circuits. *Nature Photonics* 11, 7 (2017), 441.

[15] P. R. Stanfield et al. 2019. CMOS-compatible, piezo-optomechanically tunable photonics for visible wavelengths and cryogenic temperatures. *Opt. Express* 27, 20 (Sep 2019), 28588–28605.

[16] Han Wui Then, Milton Feng, and Nick Holonyak. 2013. The transistor laser: Theory and experiment. *Proc. IEEE* 101, 10 (2013), 2271–2298.

[17] M. Zhao et al. 2017. Investigation of statistical retention of filamentary analog RRAM for neuromophic computing. In *2017 IEEE International Electron Devices Meeting (IEDM)*. 39.4.1–39.4.4.