Scalable Nanophotonic-Electronic Spiking Neural Networks

Luis El Srouji, Yun-Jhu Lee, Mehmet Berkay On, Li Zhang, *Member, IEEE*, and S. J. Ben Yoo, *Fellow, IEEE*

(Invited Paper)

***Abstract*—Spiking neural networks (SNN) provide a new computational paradigm capable of highly parallelized, real-time processing. Photonic devices are ideal for the design of high-bandwidth, parallel architectures matching the SNN computational paradigm. Furthermore, the co-integration of CMOS and photonic elements combineslow-loss photonic devices with analog electronics for greater flexibility of nonlinear computational elements. We designed and simulated an optoelectronic spiking neuron circuit on a monolithic silicon photonics (SiPh) process that replicates useful spiking behaviors beyond the leaky integrate-and-fire (LIF). Additionally, we explored two learning algorithms with the potential for on-chip learning using Mach-Zehnder Interferometric (MZI) meshes as synaptic interconnects. A variation of Random Backpropagation (RPB) was experimentally demonstrated on-chip and matched the performance of a standard linear regression on a simple classification task. In addition, we applied the Contrastive Hebbian Learning (CHL) rule to a simulated neural network composed of MZI meshes for a random input-output mapping task. The CHL-trained MZI network performed better than random guessing but did not match the performance of the ideal neural network (without the constraints imposed by the MZI meshes). Through these efforts, we demonstrate that co-integrated CMOS and SiPh technologies are well-suited to the design of scalable SNN computing architectures.**

***Index Terms*—Neuromorphic computing, spiking neural networks, nanophotonics, photonic integrated circuits, silicon photonics.**

I. Introduction

COMPUTATION using spiking neural networks (SNN) yields three major architectural advantages: (1) the sparsity of communication between elements which reduces energy cost, (2) the binarization of communication without discretization of messages (i.e., all-or-nothing spike responses), and (3) completely asynchronous operation of computational units. At the architectural level, the spiking paradigm requires several computational elements in common with the traditional artificial neural network (ANN)—weighted addition, nonlinearity, and learning algorithms—though with the additional complexity of computation spread through time. Traditional computational approaches based on the von Neumann computing architecture—including modern system architectures equipped with graphical processing units (GPUs)—are not well-suited for this computational paradigm due to the fundamental separation between computing and memory units and the resulting serialization of many processing tasks. In turn, the traditional computing paradigm cannot efficiently support the requisite computational elements without significant simplification or long latencies, thus warranting the development of new computer architectures. Neuromorphic design operates under the general principle that evolution has already produced a successful SNN architecture for operating under real-time, low-power conditions. Approaches to replicating this design employ a variety of digital, analog, or mixed-signal circuits based on electronic, photonic, or optoelectronic devices. Nonetheless, substantially more work is necessary to determine the optimal approach to abstract, apply, and improve upon this evolutionary design.

Digital neuromorphic processors (such as TrueNorth [1], Loihi [2], SpiNNaker [3], etc.) increase the parallelization of processing by including a large number of cores that allow asynchronous computation—in contrast to GPU architectures—though this approach is not unlike a specialized and monolithic form of cluster computing. Though each core completes its operation in parallel, a desire for determinism in digital electronics necessitates synchronization between simulated time steps. This determinism, in turn, limits fully asynchronous operation, which may prove to be prohibitive at biological network scales. On the other hand, analog electronic meshes can provide fully parallel computation, though the capacitance of electrical wire networks causes increases in both latency and power consumption.

Photonic and optical computing efforts have sought to exploit the nearly lossless and parallel communication capabilities of optical fibers into the domain of photonic integrated circuits (PICs). Several demonstrations have already shown matrix multiplication and convolutional processing using non-spiking photonic circuits [4], [5], [6]. These devices use a combination of wavelength-division multiplexing (WDM) and space-division multiplexing (SDM) to manage multiply-and-accumulate (MAC) operations in parallel; thus, these schemes are also compatible with spike processing in synaptic networks.

Manuscript received 1 July 2022; revised 11 October 2022; accepted 20 October 2022. Date of publication 25 October 2022; date of current version 2 November 2022. This work was supported in part by the Air Force Office of Scientific Research under Grant FA9550-181-1-0186 and in part by the Office of the Director of National Intelligence, Intelligence Advanced Research Projects Activity under Grant 2021-21090200004.

The authors are with the Electrical and Computer Engineering, University of California Davis, Davis, CA 95616-5270 USA (e-mail: lzelsrouji@ucdavis.edu; fexlee@ucdavis.edu; mbon@ucdavis.edu; zjuzhang@ucdavis.edu; sbyoo@ucdavis.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSTQE.2022.3217011.

Digital Object Identifier 10.1109/JSTQE.2022.3217011

Choices of nonlinearity in spiking elements vary widely from one approach to another, though a major division can be made between all-optical and optoelectronic approaches. Optical nonlinearities typically have shorter lifetimes and can potentially service higher-speed computation than electronic nonlinearities based on electronic charges or currents. However, the manipulation of these nonlinearities is governed mainly by material properties which are fixed after fabrication. Since biological neural networks operate over a range of time scales, it is preferable to have programmable elements in the neuron design. Optoelectronic approaches can take advantage of recent progress in the co-integration of CMOS circuitry with photonic devices to form flexible and programmable spiking neuromorphic computers.

A monolithic platform for CMOS and photonic elements allows for the close integration of programmable CMOS circuits alongside nearly lossless and highly parallelized photonic network technologies. Time-dynamics of the system can be programmed into mixed-signal electrical circuits while communication and matrix multiplication are handled by photonic interconnects. Whereas other approaches have relied on digital-to-analog (DAC) and analog-to-digital converters (ADC) to exchange input data and results between digital processors and photonic tensor cores [6], the approach reported here relies on analog photodetection for energy-efficient "impedance conversion" [7] between electrical nonlinearities (neuron dynamics) and photonic interconnects (synaptic networks). As such, the reduced capacitance, wire delay, and electrical losses of the co-integrated platform are critical for the energy efficiency of processing. Additionally, neurons in the human brain average upwards of 1,000 synapses per neuron [8]. Fanout in electrical circuits is limited by the wire and gate capacitance of downstream elements, and the capacitive coupling between wires places strict limits on the synaptic density of electrical neural networks. In contrast, photonic waveguides can carry hundreds of signals simultaneously using WDM. Furthermore, interferometry meshes in photonic circuits can perform unitary matrix multiplication without consuming any energy (as discussed in Section II-B). As such, electronics alone are not well-suited for brain-like circuits, and the mixed-signal, optoelectronic approach becomes warranted.

In addition to the architectural benefits, SNNs offer provable advantages in solving graph algorithms, constraint satisfaction, and other optimization problems [9], [10], [11], [12]. Incorporating learning and training using Hebbian [13] and spike-timing-dependent plasticity (STDP) [14] algorithms also allows for the application of SNNs in many of the same contexts as deep neural networks (DNN). These learning rules have the additional architectural advantage of using only locally available information to update each synapse. In principle, all weight updates within the network can be calculated entirely in parallel. With the appropriate network topology and training signals, Hebbian learning has also been shown capable of error-driven learning equivalent to backpropagation in deep and convolutional neural networks of moderate size [15], [16].

In this paper we will discuss the design of a nanophotonic-electronic neuromorphic architecture for native SNN computation with on-chip learning. First, Section II will provide a brief taxonomy of existing photonic and optoelectronic approaches to spiking neuron and optical matrix multiplication. Next, Section III will discuss the technologies and algorithms used while addressing scalability and remaining design challenges. Finally, Section IV will detail future directions and perspectives for the design of photonic neuromorphic processors.

II. Background and Survey

Spiking neural networks require two primary computational elements: (i) a nonlinear spiking unit that can integrate its inputs over time (the neuron) and (ii) a reconfigurable network to service weighted connections between these elements (the synaptic network). As previously alluded, the nonlinearities exploited for the design of spiking units can vary between all-optical and optoelectronic approaches, which limits the choice of network elements to service communications between units.

A. Spiking Nonlinearity

Excitability describes the ability of a system to quickly and temporarily deviate from its quiescent state following small perturbations and can be rigorously described through bifurcation analysis as done by Izhikevich [17]. Biological neurons are dynamical systems and have been classified into saddle-node and Andronov-Hopf bifurcations which correspond to *integrator* and *resonator* neurons, respectively. Simply put, integrator neurons integrate their inputs and will generate a spike upon reaching some dynamic threshold. In contrast, a resonator neuron undergoes some internal subthreshold oscillation with increased response and likelihood to generate a spike for inputs that fall at specific phases of a resonant frequency.

Computationally useful spiking neurons, however, need not be entirely biologically plausible. Instead, behavior is commonly summarized by the *leaky-integrate-and-fire* (LIF) neuron model. In the LIF model, the membrane potential constantly undergoes exponential decay towards its *resting potential* with discrete jumps at each input spike. When the membrane potential reaches a fixed threshold, the spike is generated, and the potential is instantaneously returned to a *reset potential*. LIF neurons are only able to represent integrator neurons and lose much of the complexity of behaviors seen in biological neurons. Alternatively, Izhikevich devised a neuron model which faithfully reproduces a wide range of biologically observed behaviors using only four parameters and two coupled differential equations [18]. For a brief summary of computationally relevant neuron behaviors and a comparison of neuron models, see [19]. Other taxonomies exist to classify neuron types according to these behaviors, though some evidence has shown that biological neurons may flexibly switch between these types based on the history of the cell [20]. As such, an ideal hardware implementation of spiking neurons would be capable of representing a range of neuron types for maximal computational ability.

Semiconductor lasers have been investigated for the isomorphisms between the time dynamics of material parameters of active photonic elements and the cellular mechanisms of biological neurons. Researchers have exploited the time dynamics of photocarriers, thermal diffusion, optical modes, and polarization competition to create excitable laser devices with varying degrees of faithfulness to their biological counterpart. Photonic spiking neurons can be most meaningfully divided into two categories based on whether the device can accept optical or electrical inputs—some devices can be modulated by either option, but electrical inputs provide advantages in system design as discussed in Section II-B.

Optical devices can be further classified into coherent and incoherent devices based on how incoming wavelengths are used to excite the active medium. In coherent excitable semiconductor lasers [21], [22], [23], [24], [25], the incoming signal interacts with a lasing cavity mode on the same wavelength to modulate the output signal directly. Excitability is induced by disturbing the balance between competing modes or polarizations which, with sufficient input energy, temporarily drive the extinction of one mode and amplification of the other. Bandwidth for such devices is bound by the cavity Q factor, with a time constant for energy dissipation given by $\tau = Q/\omega_0$. For incoherent devices [26], [27], [28], [29] the incoming signal interacts with some element within the cavity that indirectly modulates the output signal. This may take the form of optical pumping of the laser medium, or otherwise modulating the carrier populations which affect gain and saturation properties. Bandwidth for such approaches are limited by the dynamics of these carrier populations which are material dependent. Alternatively, optoelectronic approaches [30], [31] can allow for the design of analog circuitry with time-dynamics that can be fit to a variety of available neuron models, with lasers modulated by current injection in response to processed photodetector input. Optoelectronic designs are mainly limited by the total bandwidth of integrated photodetectors and electronics, though some estimates suggest that bandwidths upwards of 10 GHz can be expected; see [32] for a more in-depth review of various excitable semiconductor lasers with discussions of bifurcation paralleling Izhikevich's analysis.

B. Reconfigurable Networks

Given the ability of silicon waveguides to simultaneously support a wide range of wavelengths with negligible loss, on-chip optical networks are most efficiently parallelized using wavelength division multiplexing (WDM). Time-division multiplexing (TDM) offers another scheme for sharing computing resources over time, but the asynchronous and stochastic nature of SNNs is not likely to benefit from this technique. Using WDM, signals from each neuron can be routed according to wavelength, and resources for matrix multiplication may potentially be used for multiple independent operations to support weight sharing and convolution. To support such architectures, different neurons must be distinguishable by output wavelength. However, the system does not need a unique wavelength for each neuron since most SNN architectures group neurons into layers that provide an additional level of hierarchy for routing structures.

Using a WDM approach, arrayed waveguide grating routers (AWGR) can be used to support all-to-all routing schemes between neural layers [33], [34], [35]. Inputs to each layer would be passed through reconfigurable optical matrix multipliers such as cross-bar networks, micro-ring resonator (MRR) banks, and Mach-Zehnder interferometry (MZI) meshes. MZI meshes can perform unitary matrix transformations that correspond to lossless multiplication and are thus particularly suitable for low-power neuromorphic computing. See [36] for a longer discussion on the design trade-offs between each of these devices. Section III-B describes our MZI mesh architecture, while Section III-C details algorithms for training SNNs using MZI meshes.

III. Scalable Photonic SNN Technologies

A. Towards Attojoule Nanophotonic-Electronic Spiking Neurons

Neurons provide nonlinearity and signal regeneration between each neural network layer. Our previous work [37] presents an optoelectronic neuron design with projected energy efficiency on the order of 200 $aJ/$spike. Because the time scales of electrical circuits are more tunable than nonlinear photonic materials, the neuron is more easily programmable while still taking advantage of low-loss communication provided by photonic interconnects. The previous design closely matches the behavioral characteristics of the Izhikevich neuron model to achieve a variety of neural behaviors. We have updated this design for a more advanced foundry platform to move a step forward in realizing attojoule energy efficiencies.

Our previous foundry neuron design [37] also employs optoelectronics and a scalable MZI interconnect mesh; however, this design is not capable of the full range of neural behaviors described by the Izhikevich model. Using the GlobalFoundries (GF) 45SPCLO PDK, a new neuron was designed that can support a wider range of neural behaviors depending on applied voltage biasing. GF 45SPCLO is the successor of the GF 90WG PDK and preserves the same CMOS-silicon photonic co-integration with a more advanced process node and additional metal routing layers. Fig. 1 shows the GF 45SPCLO neuron circuit design. The pins labeled red mark voltage biasing nodes that can be adjusted to achieve the desired neuron behavior. These nodes correspond to the control of an adjustable positive bias (V_{bias}), spiking threshold (V_{th}), refractory feedback rate (V_{leak}), and adaptation rate (V_{leak2}). The function of these node voltages is divided between membrane potential control and feedback potential control. Membrane potential controls V_{bias} & V_{th} adjust the spiking threshold and determine the current flow into membrane potential for each spike input. Feedback potential controls, V_{leak} & V_{leak2}, determine the strength of negative feedback on the membrane potential and the length of the refractory period. Balanced photodetectors receive excitatory and inhibitory light input. The diode at the circuit output incorporates the I-V characteristics of the laser diode chosen for the design.

To demonstrate this design, we first simulate the basic spiking behavior in response to excitatory and inhibitory inputs simulated in Cadence Spectre (shown in Fig. 2). Next, the nodes of each measurement are matched to the color of each line in

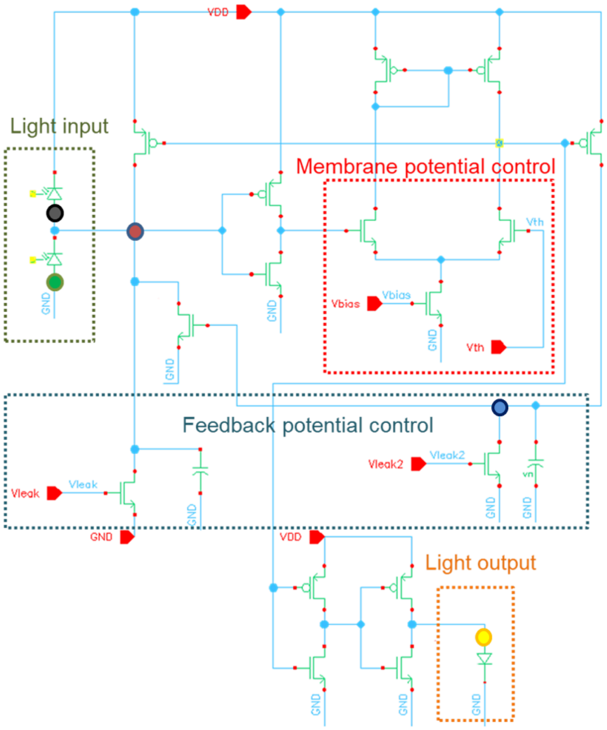

Fig. 1. The circuit diagram of 45SPCLO neuron design. The circuit mechanism of the optoelectronic neuron starts with converting light input to current. The membrane potential control section will decide the neuron threshold and feedback strength to the refractory feedback potential control section before sending the light out from the laser diode. The feedback potential control decides the refractory strength and the frequency of spiking.

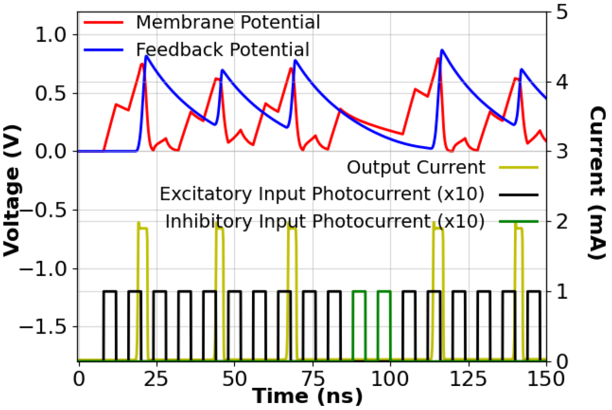

Fig. 2. Basic spiking behavior with excitatory and inhibitory input. Inhibitory inputs are assigned on the 11th and 12th spikes which oppose the excitatory input currents.

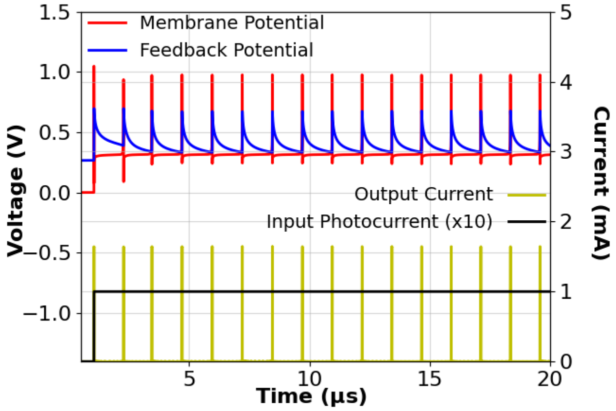

Fig. 3. Regular spiking neuron behavior. The step input shows that the circuit feedback mechanism properly functions and that the neuron is an excitable system. The spiking rate for regular spiking is set to the lower end of each voltage supply.

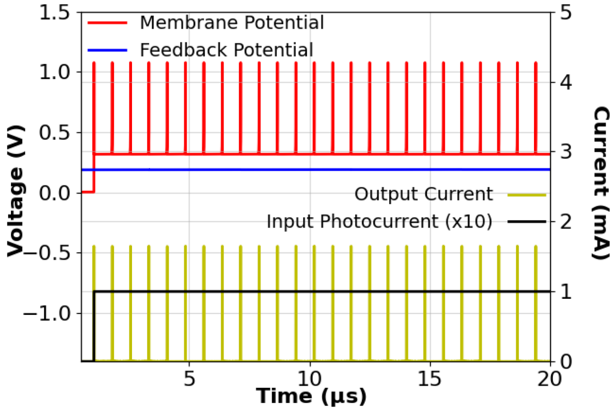

Fig. 4. Fast spiking neuron behavior. Each spike is 50 ns faster than a regular spike. The spiking speed can be adjusted by changing the voltage supplies. V_{th} has the most influence on spiking rate adjustment.

Fig. 1. Finally, we include inhibitory inputs on spike #11 and #12 and can confirm from Fig. 2 that inhibitory input suppressed the membrane potential and output, which matches our expectation.

Next, we demonstrate three spiking patterns: regular spiking (RS), fast spiking (FS), and chattering (CH) in analogy to [18]. These behaviors can be achieved flexibly by modifying the voltages at each biasing pin, which allows a greater tolerance for mismatch between design and tapeout. These spiking patterns are shown in Fig. 3, Fig. 4, and Fig. 5 respectively. Input photocurrents are simulated as step functions from $0.0\,mA$ to $0.1\,mA$, and node voltages corresponding to each behavior are set as follows:

1) Regular spiking: bias (V_{bias}) low, threshold (V_{th}) low, refractory feedback (V_{leak}) low, and frequency adaptation (V_{leak2}) low.

2) Fast spiking: bias (V_{bias}) low, threshold (V_{th}) high, refractory feedback (V_{leak}) low, and frequency adaptation (V_{leak2}) high.

3) Chattering: bias (V_{bias}) medium, threshold (V_{th}) medium, refractory feedback (V_{leak}) high, and frequency adaptation (V_{leak2}) high.

These simulations verify the ability of the neuron circuit to achieve various spiking patterns on the more advanced 45SPCLO process. While the increased complexity of the spiking neuron creates difficulty for gradient approximation, learning algorithms can be designed that are agnostic to system parameters and nonlinearities. Section III-C will explore two such methods. Implementing these algorithms alongside spiking neurons

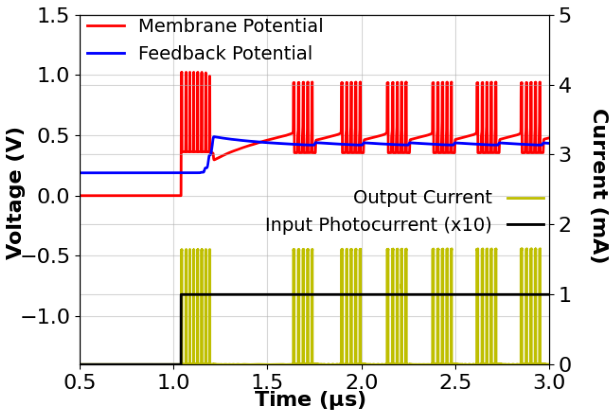

Fig. 5. Chattering neuron behavior. The neuron continuously fires for 0.3 μs and rests for 0.6 μs. This cycle is repeated with shorter firing and resting periods.

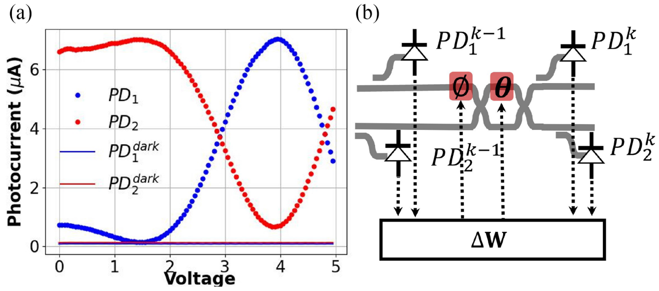

Fig. 6. (a) DC voltage sweep for phase shifter θ, (b) 2 × 2 MZI unit with power monitoring and local training features.

would bring the benefits of deep learning into the domain of event-driven processing.

B. Photonic MZI Mesh as Synaptic Network

The building block of an MZI Mesh is a 4-port device that consists of two 50:50 beam splitters and two-phase shifters, θ and ϕ as shown in Fig. 6(b). Inside the interferometer, the phase shifter θ controls the power splitting ratio. Meanwhile, outside of the interferometer, the phase shifter, ϕ, controls the relative phase difference between the two coherent input ports. As demonstrated in Fig 6(a), the tunable power splitting functionality is tested by sweeping applied DC voltage on the phase shifter θ. MZI Meshes can be arranged in several ways, with the most popular arrangements being the triangular [38] or rectangular [39] formations. Both of the formations can realize an arbitrary N × N unitary matrix. There are a variety of applications where MZI Meshes are employed, such as mode-division multiplexing [40], free-space beamforming [41], quantum computing [42], and photonic neural networks [43]. Our work utilizes MZI Meshes as synaptic interconnections for bio-inspired neural networks and aims to integrate learning algorithms on the same chip.

Calibration procedures of MZI Meshes are well-studied [44], but training MZI Meshes as neural network (NN) interconnects remains challenging. Hughes et al. [45] proposed an in-situ training to realize the traditional backpropagation algorithm for MZI meshes, and recently Pai et al. [46] experimentally demonstrated the method. This in-situ training requires additional forward and backward light propagation with power monitoring for each phase shifter element at each step. Pai et al. [46] utilized power tapping and grating couplers with an infrared camera to record the emitted power from MZI Meshes. Alternatively, Morichetti et al. [47] used a non-invasive power sensing device introduced for silicon waveguides. Ideally, neural network training algorithms are agnostic to the performance or transfer function of devices in the network. Done correctly, this avoids the need to calibrate elements within the network as it autonomously achieves a global minimum of its cost function. To this end, we exploited in-mesh 1:99 power taps and Ge photodetectors (PDs) within the Process Design Kit (PDK) elements of the active silicon photonic multi-project-wafer (MPW) runs from the AIM Photonic foundry. Information from these internal monitors were used with an implementation of random backpropagation (discussed in Section III-C1) to train a small neural network without MZI mesh calibration. Fig. 6(a) shows photocurrent changes on the monitoring PDs with respect to applied voltage on phase-shifter θ. Although we used thermo-optics as a simple and practical phase-shifting mechanism, it is also possible to utilize micro-electro-mechanical systems (MEMS) for even lower power consumption [48] in future designs.

Fig. 7(b) shows the fabricated and tested 6 × 6 rectangular MZI Mesh with power taps after each 2 × 2 MZI unit as shown in Fig. 6(b). At each output waveguide of the 6 × 6 mesh, a micro ring resonator (MRR) add-drop filter is placed with a PD on the drop ports, allowing for output monitoring by either optical or electrical means. When the MRR is at resonance, the output can be monitored and accessed through the electrical interface during the training. Alternatively, the MRR resonance wavelength can be tuned to let the optical signal propagate after the MZI Mesh. This way, multiple MZI Mesh layers can cascade for DNN-like implementation. All the components are available in AIM Photonic's PDK v4.0. The device is wire-bonded on a fanout printed-circuit board. A USB-interfaced multi-channel high current output digital-to-analog converter (DAC) unit drives the thermo-optic heaters and MRR add-drop filters. Similarly, the photocurrents are digitized by a USB-interfaced multi-channel 250kSps analog-to-digital converter (ADC), as shown in Fig. 7(a).

C. Training and Inference

We targeted a linear classification problem with 4-dimensional input vectors and two output classes for the on-chip training demonstration. We used the Iris flower dataset [49], consisting of 3 classes and 150 input samples. For simplicity in the proof-of-principle demonstration, we excluded one of the classes that is linearly separable from the other two classes. Therefore, a linear regression classifier can achieve a maximum of 94 true classifications over 100 samples. We use a single-mode cleaved fiber to couple a CW tunable laser source operating at 1553.7 nm to the chip. After the edge coupler, the first three 2 × 2 MZI stages act as tunable beam splitters and are used to generate coherent input vectors. First, the input generator phase shifters

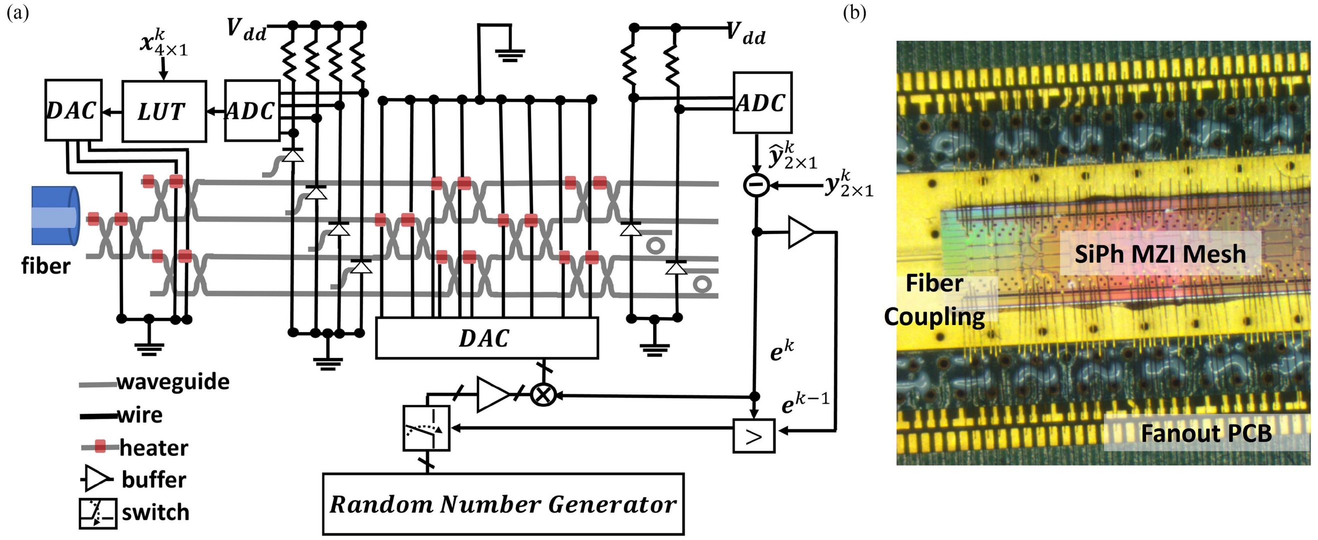

Fig. 7. (a) Random backpropagation experimental setup with optical and electrical components, (b) SiPh MZI Mesh, wirebonded on fanout PCB.

are optimized adaptively to create desired 100 input samples. Next, optimum voltage values are recorded in a look-up table (LUT) to recall in the training and interference cycles.

One of the challenges of using MZI Meshes as a synaptic weight matrix is that controllable variables (phase shifters) do not explicitly map to individual weight matrix entries. In other words, adjusting a single phase shifter will affect multiple weight matrix entries. Clements et al. [39] devised a decomposition method for rectangular meshes. In machine learning, however, the optimum weight matrix is unknown at the beginning of training and the additional resources for continual adjustment and decomposition become intractable. Hughes et al. [45] demonstrated a method of differentiating the weight matrix w.r.t. each phase shifter. However, this method requires two optical propagation steps in addition to the initial inference step: one forward, and one backward. Therefore, an external controller is required to schedule each propagation, and light sources must be bidirectional. Moreover, during the additional optical propagation steps, power must be monitored for every phase shifter element. The number of phase shifters in the MZI mesh scales as $N(N-1)$ for N × N weight matrices, meaning $\mathcal{O}(N^2)$ power monitoring is required. This presents remaining challenges for scalability in deep neural networks.

Here, we looked for more hardware-friendly solutions and, taking inspiration from biology, explored *random backpropagation* (RBP) and *contrastive Hebbian learning* (CHL) for MZI Meshes. In Section III-C1, we present an experimental demonstration of random backpropagation training for a linear classification task; Section III-C2 discusses the CHL algorithm and its relevance to human-like predictive error-driven learning.

1) Random Backpropagation: In RBP, global error is backpropagated electrically from the end of the network. As such, RBP does not require optical backpropagation or power monitoring for each 2 × 2 MZI unit. An important difference between conventional BP and RBP is the direction of the gradient. BP follows the steepest gradient direction, which requires error to multiply the conjugate transpose of the forward weight matrix. These forward weights are available in the memory unit in a digital computer, but for MZI Meshes, optical light would be physically backpropagated, as discussed earlier. The original researchers demonstrated that a random backward weight matrix could also guarantee learning unless random backward weights are exactly orthogonal to the steepest backward weights [50]. Further, neuroscience studies observed that backward synaptic connections of neural networks in mammals are not fully symmetric [51], [52] giving biological credibility to the RBP algorithm. *Direct feedback alignment*, a variant of RBP, has also been demonstrated for MRR-based photonic weight matrices [53]. Given that tunable elements in the MRR bank have a one-to-one mapping with the synaptic weight matrix, it is computationally easier to calculate the steepest gradient direction. Therefore, RBP can be more useful for MZI mesh training where this mapping is non-trivial. Nonetheless, MZI meshes are preferred for their ability to perform lossless matrix multiplication.

Appendix A summarizes our method of applying RBP on a SiPh MZI Mesh, while an illustration of our experimental setup is shown in Fig. 7(a). Python scripts realize multiplication, addition, comparator, and memory buffer operations in an external computer. Unlike conventional RBP, we draw a new random backward matrix for each iteration where the error is larger than the previous. With this modification, we empirically observed faster convergence to the classifier's highest accuracy and the ability to escape local minimums, as seen in the coarse search of Fig. 8(a). Note, however, this additional operation may not be necessary for a network with a larger number of parameters and multiple synaptic layers. For example, in the papers [50], [51], [52], [53], the authors use fixed random backward weights. Future efforts will involve the real-time implementation of these operations by integrated electronic circuits within the mesh.

Fig. 8(a) shows the interference accuracy of the SiPh MZI Mesh classifier for each epoch. In each epoch, 100 samples are forward propagated once. We use i.i.d. random backward

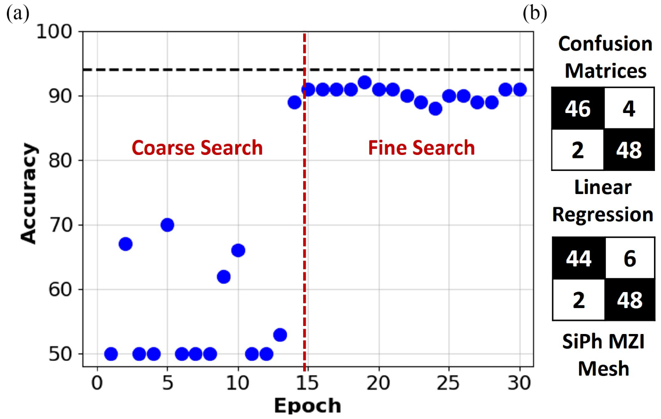

Fig. 8. (a) Interference accuracy during the random backpropagation training, (b) Confusion matrices for ideal linear regression and SiPh MZI Mesh classifer.

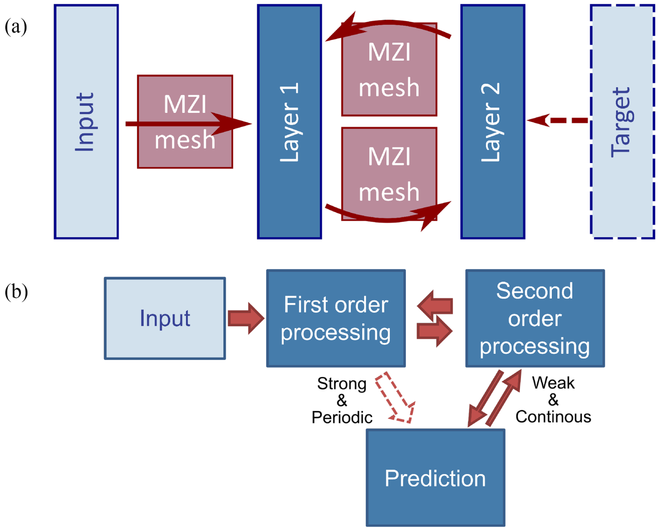

Fig. 9. (a) Schematic of two-layer CHL network structure. (b) Extension of this structure to predictive error-driven learning.

weights uniformly distributed in the interval $[-\mu, \mu]$. During the *coarse search* cycle ($\mu = 0.05$), the classifier searches different local minimums, and after some epochs, the interference accuracy decreases due to large variance on random weights. Therefore, we defined an accuracy limit (85 true labels among 100 samples) and switched to the *fine search* cycle ($\mu = 0.0025$) when the limit was reached. As seen in Fig. 8(a), the coarse search cycle ended when the classifier labeled 89 samples correctly, and in the fine search cycle, 92 true labels were achieved. The confusion matrix for the ideal linear regression classifier and SiPh MZI Mesh classifier are presented in Fig. 8(b). The SiPh MZI Mesh misclassified only two samples compared to the ordinary least squares linear regression model we built in the computer via *scikit-learn* Python package. We also implemented a numerical simulation for the MZI Meshes on the computer. From the simulation results, we observed that the SiPh MZI Meshes achieve the same accuracy as the linear regression model. Therefore, we concluded that the reason for the misclassification of two input samples was related to hardware imprecisions such as noise on the output PDs, electrical wires, thermal crosstalk between the phase shifters, etc.

Intuitively, traditional BP outperforms RBP in terms of convergence speed due to the steepest gradient direction. However, RBP is more hardware-friendly given that forward weights are unavailable and phase-shifter-to-weight mapping is not explicit in the MZI Meshes. Because the steepest direction for the gradient is not calculated, RBP does not require any power monitoring inside the MZI Meshes except for the input and output stages. Therefore, the PDs can scale with $\mathcal{O}(N)$ for N × N weight matrices. In the future, we plan to study RBP for larger SiPh MZI Meshes and more complex machine learning problems.

2) Contrastive Hebbian Learning: In contrast to backpropagation, where learning is based on credit towards global error, learning in biological systems is restricted to information local to a given synapse. Despite this, biological neural networks are able to autonomously develop expansive hierarchical abstractions of information useful for interpreting the environment. This represents a form of self-supervised learning that needs no explicit calculation of error but instead relies on chemical signals marking recent spiking activity local to a synapse.

O'Reilly [15] proved that differences in activity at two distinct phases of network computation could drive a class of temporal-difference learning rules equivalent to backpropagation and gradient descent of errors. This equivalence, however, only holds for a multi-layer perceptron (MLP) with recurrent feedback connections between each layer as in Fig. 9(a). The general learning rule has minor variations which have different properties, though an empirical test under common MLP tasks showed that the CHL variant often converges to a solution most quickly:

$$\Delta w_{ij} = \eta \left(a_i^+ a_j^+ - a_i^- a_j^- \right) \tag{1}$$

where a_i and a_j are variables representing the activity of the ith and jth neuron, and η dictates the rate of learning.

Superscripts denote the phase of activity that each variable represents. The minus phase of execution occurs first, and represents the network's natural response to the given input sample. Next, in the plus phase, the target activity is imposed on the output layer, and the network reaches a new equilibrium. For the fastest implementation, the duration of each phase should be the minimum time required for stable output activity. The product of sending and receiving neuronal activity roughly tracks their correlations during each phase. Taking the difference of this correlation in each phase forces the network to unlearn its natural response and learn the desired target activity. In a spiking network, activity in these phases can be represented by low pass filters of spike trains; however, non-spiking activity can be assumed to approximate a rate-coding of spiking activity that fits some non-linear activation function. Unlike backpropagation, however, the network architecture *requires* bidirectional synaptic connectivity (as shown between layers 1 and 2 of Fig. 9(a)) such that information propagates in both directions. Because each neuron is asynchronous, recurrence does

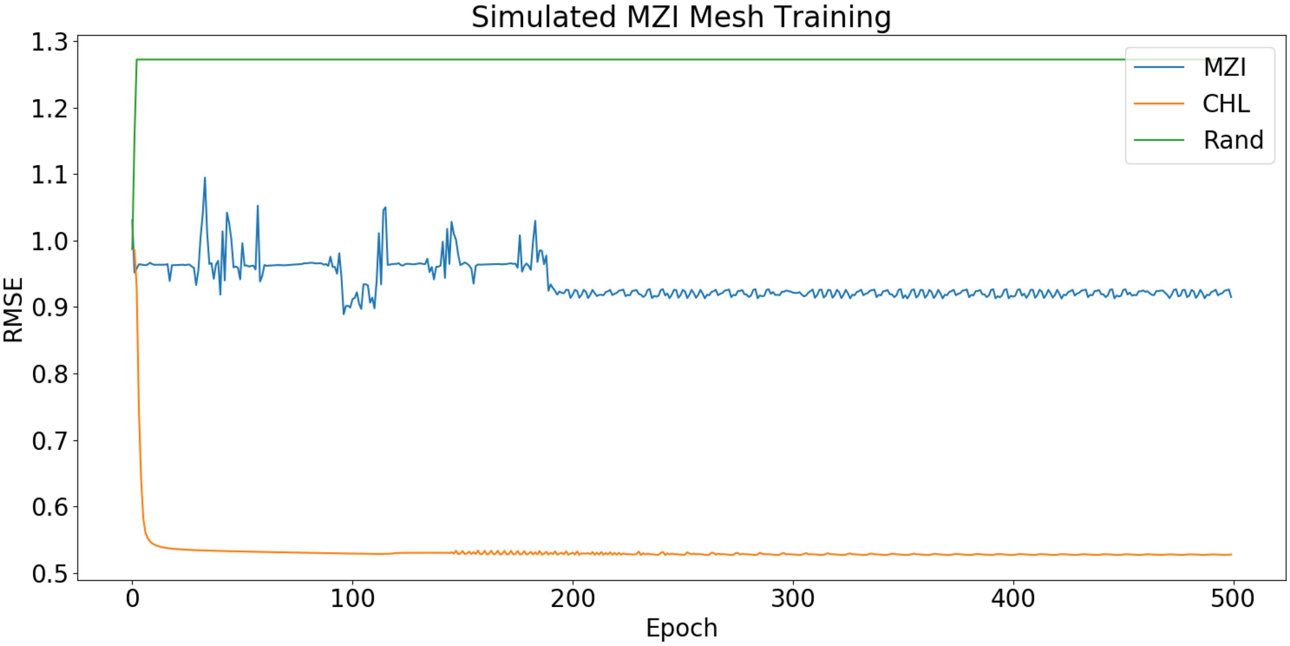

Fig. 10. Root-mean-squared-error (RMSE) of the MZI-mesh network (blue) compared to an ideal implementation of CHL (orange) and another implementation with completely random learning signals (green).

not increase computational complexity as it does on traditional computer architectures. Additionally, the locality of learning and agnosticism to the neuron nonlinearity is advantageous for spiking neuromorphic hardware.

Following the two-layer network structure depicted in Fig. 9(a), we simulated an implementation of CHL on an ideal MZI-mesh neural network. A set of 40 input-output pairs were generated from randomly-distributed, uniform-magnitude, four-dimensional vectors. Each layer was simulated with four rate-coded neurons with a sigmoidal activation function; as such, each MZI mesh was simulated as a 4×4 rectangular mesh. As in Fig. 6(b), it is assumed that each MZI unit of each mesh contains four PDs for input and output monitoring. For simplicity, it is assumed that each neuron injects light into the mesh on a separate wavelength and that the PD capacitance is large enough to reject the cross-term products between signals. Thus, the PD is assumed to linearly sum the power received from each wavelength. Because CHL assumes real-valued activation, phase shifter ϕ is neglected such that phase of each signal can be ignored. Following these assumptions, each MZI unit can be treated as a 2×2 sub-network that applies the following transformation to signal amplitude at each arm:

$$\boldsymbol{W} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} = \begin{bmatrix} \sin(\theta/2) & \cos(\theta/2) \\ \cos(\theta/2) & -\sin(\theta/2) \end{bmatrix} \quad (2)$$

Given that CHL is agnostic to the neural nonlinearity, Eq. 1 can be directly applied to the photodetector outputs as long as they are measured correctly at the plus and minus phases. However, as seen in, the MZI mesh cannot implement any arbitrary matrix. To resolve this, we can calculate derivatives that relate how a change in θ affects each individual weight. Next, we average the contribution from each Δw_{ij} to estimate the best overall change:

$$\Delta\theta = \frac{1}{4}\sum_{i,j}\left[\left(\frac{dw_{ij}}{d\theta}\right)^{-1}\Delta w_{ij}\right] \quad (3)$$

Note, we use $(dw_{ij}/d\theta)^{-1}$ because it is simpler to calculate than $d\theta/dw_{ij}$. Assuming that the plus and minus activity of each detector is recorded locally, this rule can be applied to every MZI in each mesh *all at once*. Fig. 10 shows the root-mean-squared-error (RMSE) over each epoch for the aforementioned two-layer 4×4 network, along with an ideal implementation (direct application of Eq. 1) and implementation with randomly selected Δw_{ij}. Learning is applied after each sample (not batched) with 500 epochs of training and a learning rate, $\eta = 0.1$. Each implementation is initialized to the same starting matrices.

Our MZI implementation of CHL showed an 11.21% decrease in RMSE over the course of training, which is indicative of learning. However, the ideal implementation showed a significantly larger decrease of RMSE at 46.53%. For comparison, the randomly varying network shows an *increase* of 28.87% in RMSE, giving more credibility to the idea that the MZI-CHL implementation is capable of learning—albeit at a much slower rate than the ideal implementation. It is clear from the stochastic nature of RMSE in Fig 10 that this implementation is prone to local minimums and instability. Nonetheless, this simple simulation illustrates the ability of the CHL rule to train local synaptic weights without regard for the other connections in the synaptic mesh and provides proof of concept for its use in MZI meshes. Additionally, the MZI mesh is restricted to unitary matrices, which preserve the magnitude of the input vector (before neural nonlinearity). In contrast, the ideal implementation allows independent gain and attenuation of each weight in the synaptic network. More work is needed to determine strategies for mitigating these restrictions and characterizing learning with more bio-realistic neural nonlinearities.

3) Predictive Error-Driven Learning: In the biology, however, target signals can only come from the network's own activity in response to its observations; even in the case of instructed learning, a biological brain must interpret perceptual stimuli (i.e., auditory and visual) and transform them into intelligible target signals for training. More recent work by O'Reilly et al. [54] has shown that the human brain may generate its own target training signals through cortico-thalamo loops which constantly undergo phases of prediction and observation to reduce future errors in prediction. O'Reilly et al. postulate that the alpha cycle (≈ 10 Hz) in the human brain demarcates iterations of such predictive error-driven learning, where plus and minus phases are separated by a bursting skip connection between primary processing regions and prediction-carrying regions (shown as the dotted connection in Fig. 9(b)) that fire with a 25% duty-cycle within the alpha rhythm; a simplified diagram of this neural network architecture can be seen in Fig. 9(b). Over many iterations of such prediction and observation, abstract representations can be learned that are capable of transformation-invariant object-recognition [54]. The learning rule in this model is more complicated than CHL to include additional biologically relevant terms, though the error-driven learning is captured sufficiently by the simpler rule.

Bursting is important to enforce the 25% duty-cycle and thus generate activity differences between the plus and minus phases of the CHL rule. The skip connection between first-order processing and the prediction layers allows the representation of the latter to more accurately match the ground-truth observation in the plus phase. Thus, without an explicit target signal, the network learns to better predict future inputs. Because subsequent inputs are governed by causality and are constantly occurring, the network is also constantly learning to better understand its environment. This structure can even be repeated for higher-order processing layers to hierarchically form even deeper, more abstract predictions of the input space. Future work is needed to identify an optimal implementation of the CHL rule within the MZI mesh structure and subsequently employ this style of self-supervised learning.

IV. Perspectives and Future Directions

A. Our Future System and Benchmarking

The nanophotonic-electronic spiking neuron is composed of three main components: a photodetector, a nonlinear electrical circuit, and a laser. The photodetector receives information from the synaptic network and converts the optical signal to electrical. The electrical circuit is the core of the neuron and processes the inputs to generate output spike responses. The laser output regenerates signal power after each layer to supply synaptic fanout to subsequent layers. Our team will exploit attojoule photonics with quantum impedance conversion [55] and closely integrate with low-capacitance ($< 1\,fF$) electronics for monolithic integration on a silicon-on-insulator (SOI) platform. Using photonic communication between each SNN layer reduces the capacitive charge associated with the interconnect wires [56] in comparable electronic circuits. Additionally, the photonic platform can allow neurons to communicate with other neurons at high speeds (∼10 GHz) independently of communication distance.

To calculate the projected energy consumption, we can examine the composition of each component in the attojoule nanophotonic-electronic spiking neuron design. The dynamic energy cost of the nonlinear electronic circuit and laser can be calculated by examining the transistor on-state currents and associated operation voltages and frequency. Meanwhile, the parasitic energy cost can be calculated from the total capacitance and the leakage current. According to our previous work [37], the electrical circuit current flow inside the maximum $10\,GHz$ spiking rate attojoule neuron is expected to be $31.27\,\mu A$ at $1.4\,V$ voltage supply when the neuron is in the ON state, while the leakage current is $10\,nA$ in the OFF state. The expected nanolaser energy consumption is 4.4 fJ per spike for a fanout of ~ 80 [56]. The parasitic capacitance includes the load capacitance on the photodetector, membrane capacitor, and transistor gate capacitance. According to the IRDS2020 [57] and [58], we expect the load capacitance of the photodetector to be around 0.1fF, and the simulated membrane capacitor to be 0.5fF. By considering closely integrated nanoelectronics at 10 fJ/bit energy efficiency and a fanout of 10-100 following the concept outlined by [56], the minimum dynamic input energy to generate a spike output is projected to be 200aJ/spike.

For input, the proposed attojoule neuron design will utilize a low-Q nanophotonic crystal photodetector with a Ge/Si cavity. The photonic crystal creates a resonant cavity that increases the confinement of light and reduces the size of the absorption medium [59], [60]. This allows for an ultra-low capacitance (∼0.1fF) nano-cavity PD that can generate sufficiently large voltage without amplification when combined with a high-impedance load [7]. In addition, minimizing the electrical wiring between PDs and the nonlinear electronic circuit also reduces power consumption [56]. Similarly, for spiking output, a hybrid InAs/AlGaAs quantum-dot nanolaser with a photonic crystal cavity can be employed.

The main sources of power consumption in the MZI-based synapses are waveguide loss, 2×2 MZI insertion loss, and phase shifters' power consumption. While standard silicon waveguide losses are around 2 dB/cm, SiN waveguides can further improve propagation loss to 0.45 dB/cm. Similarly SiN platforms can achieve 0.002 dB per waveguide crossing and 0.037 dB per 90° bending [61]. Unlike thermo-optical phase tuning, MOS capacitor (MOSCAP) phase shifters and phase change material (PCM) based optical phase tuning consume *'0'* mW static power [62]. With MOSCAP phase shifters 0.77 dB per 2×2 MZI can be achievable while PCM phase shifters provide <0.3 dB per 2×2 MZI unit [55].

One of the system-level design challenges is the thermal crosstalk between photonic and electronic circuits. Photonic circuits are sensitive the temperature due to optical phase changes. However, the proposed optoelectronic neuron circuitry is quiet for most of the response time thanks to the sparse nature of SNNs. The integrated laser and the MOS transistors are kept below the threshold when no spike event exists. Therefore, low power consumption and heat dissipation are achievable. Although optical phase tuning inside the interferometry provides weighting between the spiking neurons, proposed MZIs have symmetric arms. Therefore, unless the temperature changes are highly localized, optical power weighting by MZIs remains stable.

The scalability of interconnect is another critical design challenge. MZI meshes show nearly lossless multiplication that is particularly suitable for large-scale low-power neuromorphic computing. However, the number of tunable elements, $N \cdot (N-1)$ in an $N \times N$ MZI mesh, grows polynomially with the number of neurons in the layer. As such, a control circuit must be designed that scales with minimal additional computational complexity.

B. Footprint Efficiency

In the previous sections, we introduced and experimentally demonstrated bio-inspired on-chip training methods which improve the scalability of the SiPh MZI meshes for synaptic networks. We also simulated optoelectronic spiking neurons in GF 45SPCLO electronic-photonic hybrid platform and envisioned a scalable attojoule nanophotonic-electronic neuron design. However, one handicap of the proposed photonic neuromorphic system remains unaddressed: footprint efficiency. From our experience with commercial SiPh foundries, a 16 × 16 MZI mesh occupies a $12.5\mathrm{mm}^2$ chip area. Similarly, Lightmatter introduced their 64 × 64 SiPh AI accelerator occupying a $150\mathrm{mm}^2$ chip area [63], which incorporates billions of transistors. For context, many fabrication facilities have a reticle limit that falls between $400\mathrm{mm}^2$ and $800\mathrm{mm}^2$. As such, the increased size of photonic elements creates challenges for large on-chip neural networks. To improve footprint efficiency and enable deep and wide photonic neuromorphic systems, we propose two solutions: *Tensorized Photonic Neural Networks* (TPNN), which can reduce the number of MZIs by a factor of 582x [64], and *3D Electronic-Photonic Integrated Circuits* (3D EPICs) that support the chiplet design being pushed by industry leaders like Cadence [65].

1) Tpnn: There are three main methods to avoid over-parameterized neural networks and relieve hardware requirements such as *weight pruning*, *quantization*, and *model compression* [66]. Because photonic NNs are analog computers, available bit precision is already limited. Unlike electronics, a photonic system can easily offer all-to-all connectivity through wavelength and space-division multiplexing. Therefore, the benefits of weight pruning and quantization approaches are not significant. In contrast, model compression can result in fewer hardware resources and smaller footprints. We proposed and simulated an algorithm-hardware co-design approach: photonic tensorized neural networks [64]. Tensor-Train (TT) decomposition is a multi-dimensional array processing technique to represent large matrices in a low-rank approximation [67]. Although low-rank approximation may cause decreased performance in NNs, one could train NN models in TT-decomposed format so that performance degradation is minimized [68]. For some ML problems, low-rank approximation also serves as a regularization term and improves performance [69]. Moreover, in the simulations [70], we observed that TT-decomposed MZI meshes are more resilient to noise and hardware imprecision. Our simulations and benchmarks demonstrated that TPNN could improve the footprint-energy-efficiency product by 4 orders of magnitude by using 79× fewer 2 × 2 MZI units without decreasing accuracy below 95% in image classification tasks [62]. Future work will realize a SiPh end-to-end TPNN system and provide benchmarks for footprint-energy efficiency and performance.

2) 3D EPIC: 3D electronic ICs (EIC) promise low energy consumption, low noise, and high density because of shorter electrical wires [71]. The main enabling technology for 3D EICs is through-silicon vias (TSV). Although thermal relief and yield are the challenges, 3D integrated high bandwidth memories show clear advantages compared to 2D EICs. Similarly, 3D electronic-photonic ICs (EPICs) can achieve high density, low loss, and high bandwidth performance. Multi-layer silicon photonic devices are already available in commercial foundries. However, they rely on evanescent vertical couplers, which require relatively long taper lengths ($\sim 100\,\mu m$) and small layer distance ($\sim 1\,\mu m$) [61], [72]. As an alternative, our previous work demonstrates through silicon optical vias (TSOV) [73], [74] for 3D EPICS using 45° reflectors and silicon vias [74]. Ultrafast laser inscription also allows for freeform shaping of waveguides useful for routing in three dimensions. This technique has already been demonstrated for orbital-angular momentum multiplexing/demultiplexing and optical beam steering applications [75]. Furthermore, 3D EPICs provide devices to be stacked vertically, allowing for greater neuron density per area and thus the design of deeper and wider photonic neural networks.

C. Applications for SNNs

In relation to AI and machine learning, SNNs provide several advantages over modern computing paradigms for tasks that mimic the conditions in which they naturally evolved. Because SNNs process data over time in a continuous manner, they are well-suited to applications situated in real-time environments with single inference and learning instances presented at a time (such as event-based signal processing [76]). In addition, the spread of information over time allows multiple forms of memory at different time scales, similar to the human distinction between working [77], short-term [78], and long-term memories. Neuromorphic sensing and robotics are common applications of SNNs; for example, an adaptive robotic arm controller can provide reliable motor control as actuators wear down [79]. More speculatively, future devices might exploit these properties in the context of live audio and natural language processing for voice assistants, live-captioning services, or audio separation; similarly, SNNs can be used for live video and lidar processing in autonomous vehicles or surveillance systems. SNNs are not ideal for batched computation—in which multiple training samples are computed in parallel and averaged for parallelism in training—however, data centers may still make use of the increased computational parallelism in tasks like the nearest-neighbor search, which can be performed in constant time, $O(1)$, on neuromorphic chips like Loihi [80].

A major challenge of many modern DNN and reinforcement learning (RL) agents is the development of abstract, transformation-invariant representations of objects relevant to the task. For example, in classification tasks, a neural network must transform its input space into a representation that most clearly separates each labeled class. Similarly, RL agents must be able to process their input space into a representation that best accentuates the value of potential actions. Predictive error-driven learning, modeled after the work of O'Reilly [54], has the potential to autonomously build deep hierarchies of abstraction for a given input space. For example, a learning agent could implicitly learn physical properties of the world—such as gravity, buoyancy, and contact forces—simply by observing its environment. Combined with complementary learning systems for memory [81] and RL models based on the basal ganglia [82], a neuromorphic learning agent may be capable of replicating simple navigation and foraging behaviors that require the flexible application of knowledge and memory. Such a model could provide key insights for the development of self-motivated learning agents that exploit hierarchical representations to solve reinforced tasks. Developing dedicated spiking neuromorphic hardware and taking advantage of energy-efficient and scalable photonic devices will allow the development of larger models and new computational paradigms. These developments can be applied in dynamic, noisy environments that are not well-handled by today's machine learning efforts.

V. Conclusion

We have discussed the advantages of dedicated SNN hardware and highlighted the benefits of nanophotonic-electronic design within this computational paradigm. Additionally, we argued that the co-integration of photonic and electronic devices combines the high-bandwidth, low-power communication protocols of photonics with well-established and flexible CMOS circuitry. Towards constructing a photonic SNN computing architecture, we demonstrated an Izhikevich-inspired optoelectronic neuron design, implemented RPB on an MZI mesh, and simulated CHL on a rate-coded, MZI-mesh neural network. In addition, we proposed the construction of a powerful self-learning SNN computing architecture built from these technologies and based on predictive error-driven learning models of the human brain. Subsequently, we have discussed technologies for improving the scalability of neuron and network density through tensorization of large neural networks and 3D electronic-photonic integration. Finally, we discussed perspectives on the suitable applications of photonic SNNs and emphasized applications of interest for our efforts.

Future work is needed to establish the optimal design for brain-inspired spiking networks. Modern ANNs have oversimplified neural nonlinearities due to the limitations of the von Neumann computing architecture. Meanwhile, the heterogeneity of neural behaviors in different regions of the human brain provides various methods of encoding information. As such, a deeper exploration of these encodings is warranted to fully leverage the computing power of SNNs. Furthermore, modern learning algorithms are designed for sequential processing, which is not ideal for SNN hardware. As such, considerable work is necessary to determine the most efficient on-chip implementation of local learning rules like CHL. Nonetheless, the design challenges are well worth the effort to provide alternative routes for continued advances in computation and signal processing in the face of slowing progress of transistor scaling. Our continued work will focus on the characterization and design of nanophotonic-electronic spiking neurons and their incorporation within scalable, MZI-based neural networks capable of on-chip local learning.

Appendix
RBP Algorithm

Algorithm 1: Random Backprop on SiPh MZI Mesh.

```
Initialize resistor values $\mathbf{R}$, accuracy limit $L$, total number of samples $N$, MZI voltages $\mathbf{v}_{MZI}^{-1}$, error $\mathbf{e}^{-1} = \infty$, coarse and fine step sizes $\mu_c, \mu_f$, start with coarse search $\mu \leftarrow \mu_c$,, random backprop weights $\mathbf{B} \sim [-\mu, \mu]$
for Every epochdo
    for $k = 0$ through $N$do
        Find input generator voltages $\mathbf{v}_{in}^{k}$ for $\mathbf{x}^k$ in $LUT$
        Read input generator's PDs to verify $\mathbf{x}^k$
        Read output PDs $\mathbf{v}_{out}$
        Calculate photocurrent $\mathbf{i}_{out} = (v_{dd} - \mathbf{v}_{out})/\mathbf{R}$
        Normalize $\mathbf{i}_{out}$ to calculate $\hat{\mathbf{y}}^k$
        Calculate error $\mathbf{e}^k = |\hat{\mathbf{y}}^k - \mathbf{y}^k|^2$
        if $\mathbf{e}^k > \mathbf{e}^{k-1}$then
            Draw a new $\mathbf{B} \sim [-\mu, \mu]$
        end if
        Update $\mathbf{v}_{MZI}^{k} \leftarrow \mathbf{v}_{MZI}^{k-1} + \mathbf{B}\mathbf{e}^k$
    end for
    Calculate interference accuracy $a$
    for Every sample $\mathbf{x}^k$ do
        Find input generator voltages $\mathbf{v}_{in}^{k}$ for $\mathbf{x}^k$ in $LUT$
        Read input generator's PDs to verify $\mathbf{x}^k$
        Read output PDs $\mathbf{v}_{out}$
        Calculate photocurrent $\mathbf{i}_{out} = (v_{dd} - \mathbf{v}_{out})/\mathbf{R}$
        Decide class label $\hat{l}^k = \arg\max_{n} i_{out}[n]$
    end for
    $a = sum(\hat{\mathbf{l}} == \mathbf{l})$
    if $a \geq L$ then
        Switch to fine search $\mu \leftarrow \mu_f$
    end if
end for
```

Acknowledgment

The authors would like to thank GLOBALFOUNDRIES for providing silicon fabrication through the 90WG University Program and for their Technical Assistance in 45SPCLO MPW runs. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] P. A. Merolla et al., "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, 2014.

[2] M. Davies et al., "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, Jan./Feb. 2018.

[3] E. Painkras et al., "SpiNNaker: A multi-core system-on-chip for massively-parallel neural net simulation," in *Proc. IEEE Custom Integr. Circuits Conf.*, 2012, pp. 1–4.

[4] A. Mehrabian, Y. Al-Kabani, V. J. Sorger, and T. El-Ghazawi, "PCNNA: A photonic convolutional neural network accelerator," in *Proc. IEEE 31st Int. Syste.-Chip Conf.*, 2018, pp. 169–173.

[5] X. Xu et al., "11 TOPS photonic convolutional accelerator for optical neural networks," *Nature*, vol. 589, no. 7840, pp. 44–51, 2021.

[6] J. Feldmann et al., "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, no. 7840, pp. 52–58, 2021.

[7] D. A. Miller, "Optics for low-energy communication inside digital processors: Quantum detectors, sources, and modulators as efficient impedance converters," *Opt. Lett.*, vol. 14, no. 2, pp. 146–148, 1989.

[8] J. Hawkins and S. Ahmad, "Why neurons have thousands of synapses, a theory of sequence memory in neocortex," *Front. Neural Circuits*, vol. 10, 2016, Art. no. 23.

[9] C.-N. Chou, K.-M. Chung, and C.-J. Lu, "On the algorithmic power of spiking neural networks," *Leibniz Int. Proc. Inf.*, vol. 124, 2018.

[10] S. J. Verzi et al., "Computing with spikes: The advantage of fine-grained timing," *Neural Comput.*, vol. 30, no. 10, pp. 2660–2690, 2018.

[11] J. Kwisthout and N. Donselaar, "On the computational power and complexity of spiking neural networks," in *ACM Int. Conf. Proc. Ser.*, 2020, vol. 17, pp. 1–7.

[12] J. B. Aimone et al., "Provable advantages for graph algorithms in spiking neural networks," in *Proc. Annu. ACM Symp. Parallelism Algorithms Architectures*, 2021, pp. 35–47.

[13] W. Gerstner and W. M. Kistler, "Mathematical formulations of Hebbian learning," *Biol. Cybern.*, vol. 87, no. 5, pp. 404–415, 2002.

[14] N. Caporale and Y. Dan, "Spike timing–dependent plasticity: A Hebbian learning rule," *Annu. Rev. Neurosci.*, vol. 31, no. 1, pp. 25–46, 2008, doi: 10.1146/annurev.neuro.31.060407.125639.

[15] R. C. O'Reilly, "Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm," *Neural Comput.*, vol. 8, no. 5, pp. 895–938, 1996.

[16] G. Amato, F. Carrara, F. Falchi, C. Gennaro, and G. Lagani, "Hebbian learning meets deep convolutional neural networks," *Lecture Notes Comput. Sci.*, vol. 11751, pp. 324–334, 2019.

[17] E. M. Izhikevich, "Neural excitability, spiking and bursting," *Int. J. Bifurcation Chaos*, vol. 10, no. 6, pp. 1171–1266, 2000.

[18] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Trans. Neural Netw.*, vol. 14, no. 6, pp. 1569–1572, Nov. 2003.

[19] E. M. Izhikevich, "Which model to use for cortical spiking neurons?," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1063–1070, Sep. 2004.

[20] M. Steriade, "Neocortical cell classes are flexible entities," *Nature Rev. Neurosci.*, vol. 5, no. 2, pp. 121–134, 2004.

[21] M. Giudici, C. Green, G. Giacomelli, U. Nespolo, and J. R. Tredicce, "Andronov bifurcation and excitability in semiconductor lasers with optical feedback," *Phys. Rev. E*, vol. 55, no. 6, 1997, Art. no. 6414.

[22] W. Coomans, L. Gelens, S. Beri, J. Danckaert, and G. Van Der Sande, "Solitary and coupled semiconductor ring lasers as optical spiking neurons," *Phys. Rev. E.*, vol. 84, no. 3, 9 2011, Art. no. 036209.

[23] M. Brunstein et al., "Excitability and self-pulsing in a photonic crystal nanocavity," *Phys. Rev. A.*, vol. 85, no. 3, 2012, Art. no. 031803.

[24] J. Dambre et al., "Excitability in optically injected microdisk lasers with phase controlled excitatory and inhibitory response," *Opt. Exp.*, vol. 21, no. 22, pp. 26182–26191, 2013.

[25] B. Garbin et al., "Incoherent optical triggering of excitable pulses in an injection-locked semiconductor laser," *Opt. Lett.*, vol. 39, no. 5, pp. 1254–1257, 2014.

[26] M. A. Nahmias et al., "A leaky integrate-and-fire laser neuron for ultrafast cognitive computing," *IEEE J. Sel. Topics Quantum Electron.*, vol. 19, no. 5, Sep./Oct. 2013, Art. no. 1800212.

[27] F. Selmi et al., "Relative refractory period in an excitable semiconductor laser," *Phys. Rev. Lett.*, vol. 112, no. 18, 2014, Art. no. 183902.

[28] A. Hurtado and J. Javaloyes, "Controllable spiking patterns in long-wavelength vertical cavity surface emitting lasers for neuromorphic photonics systems," *Appl. Phys. Lett.*, vol. 107, no. 24, 2015, Art. no. 241103.

[29] F. Selmi et al., "Temporal summation in a neuromimetic micropillar laser," *Opt. Lett.*, vol. 40, no. 23, pp. 5690–5693, 2015.

[30] B. Romeira et al., "Excitability and optical pulse generation in semiconductor lasers driven by resonant tunneling diode photo-detectors," *Opt. Exp.*, vol. 21, no. 18, pp. 20931–20940, 2013.

[31] A. N. Tait, B. J. Shastri, M. A. Nahmias, P. R. Prucnal, and T. F. d. Lima, "Excitable laser processing network node in hybrid silicon: Analysis and simulation," *Opt. Exp.*, vol. 23, no. 20, pp. 26800–26813, 2015.

[32] A. N. Tait, B. J. Shastri, M. A. Nahmias, P. R. Prucnal, and T. F. d. Lima, "Recent progress in semiconductor excitable lasers for photonic spike processing," *Adv. Opt. Photon.*, vol. 8, no. 2, pp. 228–299, 2016.

[33] C. Mitsolidou et al., "Silicon photonic 8×8 cyclic arrayed waveguide grating router for O-band on-chip communication," *Opt. Exp.*, vol. 26, no. 5, pp. 6276–6284, 2018.

[34] Y. Zhang et al., "Foundry-enabled scalable all-to-all optical interconnects using silicon nitride arrayed waveguide router interposers and silicon photonic transceivers," *IEEE J. Sel. Topics Quantum Electron.*, vol. 25, no. 5, Sep./Oct. 2019, Art no. 8300409.

[35] X. Xiao, R. Proietti, S. Werner, P. Fotouhi, and S. J. Yoo, "Flex-LIONS: A scalable silicon photonic bandwidth-reconfigurable optical switch fabric," in *Proc. IEEE 24th OptoElectron. Commun. Conf./Int. Conf. Photon. Switching Comput.*, 2019, pp. 1–3.

[36] L.E. Srouji et al., "Photonic and optoelectronic neuromorphic computing," *APL Photon.*, vol. 7, no. 5, 2022, Art. no. 051101.

[37] Y.-J. Lee, M. B. On, X. Xiao, R. Proietti, and S. J. B. Yoo, "Photonic spiking neural networks with event-driven femtojoule optoelectronic neurons based on izhikevich-inspired model," *Opt. Exp.*, vol. 30, no. 11, pp. 19360–19389, 2022.

[38] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, "Experimental realization of any discrete unitary operator," *Phys. Rev. Lett.*, vol. 73, no. 1, pp. 58–61, 1994.

[39] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley, "Optimal design for universal multiport interferometers," *Optica*, vol. 3, no. 12, pp. 1460–1465, 2016.

[40] K. Choutagunta, I. Roberts, D. A. Miller, and J. M. Kahn, "Adapting Mach–Zehnder mesh equalizers in direct-detection mode-division-multiplexed links," *J. Lightw. Technol.*, vol. 38, no. 4, pp. 723–735, Feb. 2020.

[41] M. Milanizadeh et al., "Multibeam free space optics receiver enabled by a programmable photonic mesh," vol. 12, 2021.

[42] X. Qiang et al., "Large-scale silicon quantum photonics implementing arbitrary two-qubit processing," *Nature Photon.*, vol. 12, no. 9, pp. 534–539, 2018.

[43] Y. Shen et al., "Deep learning with coherent nanophotonic circuits," *Nature Photon.*, vol. 11, no. 7, pp. 441–446, 2017.

[44] S. Pai et al., "Parallel fault-tolerant programming of an arbitrary feedforward photonic network," vol. 9, 2019.

[45] 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, 2018, Art. no. 864.

[46] S. Pai et al., "Experimentally realized in situ backpropagation for deep learning in nanophotonic neural networks," 2022, *arXiv:2205.08501*.

[47] F. Morichetti et al., "Non-invasive on-chip light observation by contactless waveguide conductivity monitoring," *IEEE J. Sel. Topics Quantum Electron.*, vol. 20, no. 4, Jul./Aug. 2014, Art. no. 8201710.

[48] A. Yuji et al., "MORPHIC: Programmable photonic circuits enabled by silicon photonic MEMS," *Silicon Photon.*, vol. 11285, 2020, Art. no. 1128503, doi: 10.1117/12.2540934.

[49] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[50] T. P. Lillicrap, D. Cownden, D. B. Tweed, and C. J. Akerman, "Random feedback weights support learning in deep neural networks," 2014, *arXiv:1411.0247*.

[51] A. Van Schaik et al., "Event-driven random back-propagation: Enabling neuromorphic deep learning machines," *Front. Neurosci.*, vol. 11, 2017, Art. no. 324.

[52] G. Detorakis, T. Bartley, and E. Neftci, "Contrastive Hebbian learning with random feedback weights," *Neural Netw.*, vol. 114, pp. 1–14, 2019.

[53] A. N. Trondheim, "Direct feedback alignment provides learning in deep neural networks," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 1045–1053.

[54] R. C. O'reilly, J. L. Russin, M. Zolfaghar, and J. Rohrlich, "Deep predictive learning in neocortex and pulvinar," *J. Cogn. Neurosci.*, vol. 33, no. 6, pp. 1158–1196, 2021.

[55] F. De Leonardis, R. Soref, V. M. Passaro, Y. Zhang, and J. Hu, "Broadband electro-optical crossbar switches using low-loss Ge_2 Sb_2Se_4 Te_1 phase change material," *J. Lightw. Technol.*, vol. 37, no. 13, pp. 3183–3191, Jul. 2019.

[56] D. A. Miller, "Attojoule optoelectronics for low-energy information processing and communications," *J. Lightw. Technol.*, vol. 35, no. 3, pp. 346–396, Feb. 2017.

[57] "The International roadmap for devices and systems," *IEEE*, 2020.

[58] B. J. Shastri et al., *Neuromorphic Photonics, Principles of BT - Encyclopedia of Complexity and Systems Science*, R. A. Meyers, Ed., Berlin, Germany: Springer, 2018, pp. 1–37.

[59] Y. El-Batawy, F. M. Mohammedy, and M. J. Deen, *13 - Resonant Cavity Enhanced Photodetectors: Theory, Design and Modeling*. B. B.T.P. Nabet, Ed., Sawston, U.K.: Woodhead Publishing, 2016, pp. 415–470.

[60] K. Nozaki et al., "InGaAs nano-photodetectors based on photonic crystal waveguide including ultracompact buried heterostructure," *Opt. Exp.*, vol. 21, no. 16, pp. 19022–19028, 2013.

[61] W. D. Sacher et al., "Tri-layer silicon nitride-on-silicon photonic platform for ultra-low-loss crossings and interlayer transitions," *Opt. Exp.*, vol. 25, no. 25, pp. 30862–30875, 2017.

[62] X. Xiao et al., "Large-scale and energy-efficient tensorized optical neural networks on III–V-on-silicon MOSCAP platform," *APL Photon.*, vol. 6, no. 12, 2021, Art. no. 126107.

[63] C. Ramey, "Silicon photonics for artificial intelligence acceleration: HotChips 32," in *Proc. IEEE Hot Chips 32 Symp.*, 2020, pp. 1–26.

[64] X. Xiao and S. J. B. Yoo, "Scalable and compact 3D tensorized photonic neural networks," in *Proc. Opt. Fiber Commun. Conf. Exhib.*, 2021, pp. 1–3.

[65] "3D-IC Design Solution—Cadence,".

[66] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 1135–1143.

[67] I. V. Oseledets, "Tensor-train decomposition," *SIAM J. Sci. Comput.*, vol. 33, no. 5, pp. 2295–2317, 2011.

[68] A. Novikov, D. Podoprikhin, A. Osokin, and D. Vetrov, "Tensorizing neural networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 442–450.

[69] C. Hawkins and Z. Zhang, "Bayesian tensorized neural networks with automatic rank selection," *Neurocomputing*, vol. 453, pp. 172–180, 2021.

[70] M. B. On, Y.-J. Lee, X. Xiao, R. Proietti, and S. Ben Yoo, "Analysis of the hardware imprecisions for scalable and compact photonic tensorized neural networks," in *Proc. IEEE Eur. Conf. Opt. Commun.*, 2021, pp. 1–4.

[71] J. U. Knickerbocker et al., "Three-dimensional silicon integration," *IBM J. Res. Develop.*, vol. 52, no. 6, pp. 553–569, Nov. 2008.

[72] K. Shang et al., "Low-loss compact multilayer silicon nitride platform for 3D photonic integrated circuits," *Opt. Exp.*, vol. 23, no. 16, pp. 21334–21342, 2015.

[73] Y. Zhang, A. Samanta, K. Shang, and S. J. B. Yoo, "Scalable 3D silicon photonic electronic integrated circuits and their applications; scalable 3D silicon photonic electronic integrated circuits and their applications," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 2, Mar./Apr. 2020, Art. no. 8201510.

[74] Y. Zhang, Y.-C. Ling, Y. Zhang, K. Shang, and S. J. B. Yoo, "High-density wafer-scale 3-D silicon-photonic integrated circuits," *IEEE J. Sel. Topics Quantum Electron.*, vol. 24, no. 6, Nov./Dec. 2018, Art. no. 8200510.

[75] S. J. Ben Yoo, B. Guan, and R. P. Scott, "Heterogeneous 2D/3D photonic integrated microsystems," *Microsyst. Nanoeng.* vol. 2, no. 1, pp. 1–9, 2016.

[76] P. Blouw and C. Eliasmith, "Event-driven signal processing with neuromorphic computing systems," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 8534–8538.

[77] M. Giulioni et al., "Robust working memory in an asynchronously spiking neural network realized with neuromorphic VLSI," *Front. Neurosci.*, vol. 5, 2012, Art. no. 149.

[78] A. Rao, P. Plank, A. Wild, and W. Maass, "A long short-term memory for AI applications in spike-based neuromorphic hardware," *Nature Mach. Intell.*, vol. 4, no. 5, pp. 467–479, 2022.

[79] T. DeWolf, T. C. Stewart, J. J. Slotine, and C. Eliasmith, "A spiking neural model of adaptive arm control," *Proc. Roy. Soc. B: Biol. Sci.*, vol. 283, no. 1843, 2016, Art. no. 20162134.

[80] E. P. Frady et al., "Neuromorphic nearest neighbor search using intel's pohoiki springs," in *Proc. ACM Int. Conf. Proc. Ser.*, 2020, pp. 1–10.

[81] R. C. O'Reilly, R. Bhattacharyya, M. D. Howard, and N. Ketz, "Complementary learning systems," *Cogn. Sci.*, vol. 38, no. 6, pp. 1229–1248, 2014.

[82] D. Rasmussen, A. Voelker, and C. Eliasmith, "A neural model of hierarchical reinforcement learning," *PLoS One*, vol. 12, no. 7, 2017, Art. no. e0180234.

Luis El Srouji received the B.S. degree in applied physics with an emphasis in physical electronics from the University of California, Davis, CA, USA, in 2020. He is currently working toward the Ph.D degree in electrical engineering with the University of California, Davis. His research interests include the design of bio-physically accurate analog neuron circuits, development of learning algorithms for optoelectronic spiking neural networks, and fabrication of on-chip laser sources.

Yun-Jhu Lee received the B.S. degree in life science from the National Taiwan University, Taiwan. He is currently working toward the Ph.D degree in electrical and computer engineering with the University of California, Davis, CA, USA. His research interests include neuromorphic computing, integrated photonics, MEMS, and control system.

Mehmet Berkay On received the B.S. in electrical and electronics engineering from the Bilkent University, Ankara, Turkey, in 2018. He is currently working toward the Ph.D. degree in electrical and computer engineering with the University of California, Davis, CA, USA. His research interests include energy-efficient photonic neuromorphic systems, RF-photonic signal processing, fiber-optic communication, and compressive sensing.

Li Zhang (Member, IEEE) received the B.S. degree in electronics and information technology and instrumentation from Zhejiang University, Hangzhou, China, in 2016. He is currently working toward the Ph.D. degree in electrical engineering with the University of California at Davis, Davis, CA, USA. His research interests include ultra-wideband transceiver, trans-impedance amplifier and optical driver.

S. J. Ben Yoo (Fellow, IEEE) received the B.S. degree in electrical engineering with distinction, the M.S. degree in electrical engineering, and the Ph.D. degree in electrical engineering with a minor in physics, from Stanford University, Stanford, CA, USA, in 1984, 1986, and 1991, respectively. He is currently a Distinguished Professor of electrical engineering with University of California (UC) at Davis, Davis, CA, USA. Prior to joining UC Davis in 1999, he was a Senior Research Scientist with Bellcore, leading technical efforts in integrated photonics, optical networking, and systems integration. He led the MONET testbed experimentation efforts, and participated in ATD/MONET systems integration and a number of standardization activities. Prior to joining Bellcore in 1991, he conducted research on nonlinear optical processes in quantum wells, a four-wave-mixing study of relaxation mechanisms in dye molecules, and ultrafast diffusion-driven photodetectors with Stanford University. His research interests with UC Davis includes 2D/3D photonic integration for future computing, communication, imaging, and navigation systems, micro/nano systems integration, and the future Internet. His research interests with Bellcore included the next-generation internet, reconfigurable multiwavelength optical networks (MONET), wavelength interchanging cross connects, wavelength converters, vertical-cavity lasers, and high-speed modulators. He was the recipient of the DARPA Award for Sustained Excellence (1997), the Bellcore CEO Award (1998), the Mid-Career Research Faculty Award (2004 UC Davis), and the Senior Research Faculty Award (2011 UC Davis). He is the Fellow of OSA, NIAC.