Photonic spiking neural networks with event-driven femtojoule optoelectronic neurons based on Izhikevich-inspired model

Research Article | Vol. 30, No. 11 / 23 May 2022 / Optics Express 19360
Optics EXPRESS

Photonic spiking neural networks with event-driven femtojoule optoelectronic neurons based on Izhikevich-inspired model

YUN-JHU LEE, MEHMET BERKAY ON, XIAN XIAO, ROBERTO PROIETTI, AND S. J. BEN YOO*

Department of Electrical and Computer Engineering, University of California, Davis, CA 95616, USA
sbyoo@ucdavis.edu

Abstract: Photonic spiking neural networks (PSNNs) potentially offer exceptionally high throughput and energy efficiency compared to their electronic neuromorphic counterparts while maintaining their benefits in terms of event-driven computing capability. While state-of-the-art PSNN designs require a continuous laser pump, this paper presents a monolithic optoelectronic PSNN hardware design consisting of an MZI mesh incoherent network and event-driven laser spiking neurons. We designed, prototyped, and experimentally demonstrated this event-driven neuron inspired by the Izhikevich model incorporating both excitatory and inhibitory optical spiking inputs and producing optical spiking outputs accordingly. The optoelectronic neurons consist of two photodetectors for excitatory and inhibitory optical spiking inputs, electrical transistors' circuits providing spiking nonlinearity, and a laser for optical spiking outputs. Additional inclusion of capacitors and resistors complete the Izhikevich-inspired optoelectronic neurons, which receive excitatory and inhibitory optical spikes as inputs from other optoelectronic neurons. We developed a detailed optoelectronic neuron model in Verilog-A and simulated the circuit-level operation of various cases with excitatory input and inhibitory input signals. The experimental results closely resemble the simulated results and demonstrate how the excitatory inputs trigger the optical spiking outputs while the inhibitory inputs suppress the outputs. The nanoscale neuron designed in our monolithic PSNN utilizes quantum impedance conversion. It shows that estimated 21.09 fJ/spike input can trigger the output from on-chip nanolasers running at a maximum of 10 Gspike/second in the neural network. Utilizing the simulated neuron model, we conducted simulations on MNIST handwritten digits recognition using fully connected (FC) and convolutional neural networks (CNN). The simulation results show 90% accuracy on unsupervised learning and 97% accuracy on a supervised modified FC neural network. The benchmark shows our PSNN can achieve 50 TOP/J energy efficiency, which corresponds to 100 × throughputs and 1000 × energy-efficiency improvements compared to state-of-art electrical neuromorphic hardware such as Loihi and NeuroGrid.

1. Introduction

Machine Learning (ML) and Artificial Intelligence (AI) have already transformed our everyday lives–everything from scientific computing to shopping and entertainment involves some form of machine learning or intelligent algorithms. Such AI and ML systems typically use large von Neumann computing systems such as data centers, and they have shown remarkable capabilities to beat the human brain in some tasks, including the highly complex game of Go [1,2]. However, today's data centers consume megawatts of power (Google's AlphaGo utilized 1920 CPUs, and 280 GPUs, and Facebook Data Center electricity usage reached 5.1 Terawatt-hours in 2019 [3]), and the current deep neural network algorithms require labor-intensive hand labeling of large datasets. While such high power consumption of cloud computing is currently tolerated due to the consolidated and amortized economic model of massive users, the newly emerging

#449528 https://doi.org/10.1364/OE.449528
Journal © 2022 Received 20 Dec 2021; revised 28 Mar 2022; accepted 16 Apr 2022; published 18 May 2022

Optics EXPRESS

edge-computing [4] in the autonomous vehicles, drones, robots, smart-health, and the gateways of the Internet-of-Things (IoT) drives the need of intelligent, power-efficient, and high-throughput neuromorphic computing. Instead of using artificial neural networks (ANNs) in von Neumann architectures (e.g., utilizing non-spiking neural networks in GPU-based cluster), recent efforts towards spiking neuromorphic computing such as IBM's TrueNorth [5] and Intel's Loihi [6] processors have demonstrated significant energy-efficiency improvements compared to the non-spiking ANN counterparts. The bio-inspired neuromorphic hardware systems such as IBM's TrueNorth claim to achieve 176,000 times higher energy efficiency than the general-purpose Intel i7 based von-Neumann computing system for specific applications [5]. However, the scalability and performance of electronic neuromorphic hardware are typically constrained by the electrical interconnects with bandwidth limitations, thermal noise, electromagnetic interference, and relatively high loss and dispersion. On the other hand, photonic neuromorphic hardware can exploit optical interconnects with optical parallelism, high bandwidth, low-loss, low-noise, and energy-efficient communication independently of the distance.

Table 1 and Table 2 list recently published photonic neuromorphic hardware research. [7–29]. We select the research that includes at least two of the criteria in the following. (1) It's experimentally demonstrated. (2) The type of neuron model used in the research has applications in large-scale image classification. (3) The design of neurons shares similar features with our work such as both inhibitory excitatory light inputs, integrated laser output, nonlinear function by optoelectronic circuit, etc. (4) The research envisions how to scale to large photonic neural network implementation. The neuromorphic hardware consists of neurons and synaptic interconnect networks. The neuron is where the nonlinearity of the transfer function is implemented, and the synaptic interconnect networks apply weight values between the neurons. In large, there are four categories of photonic neural networks: photonic spiking neural networks (PSNNs) with optoelectronic neurons, PSNNs with spiking all-optical neurons, photonic neural networks (PNNs) with non-spiking optoelectronic neurons, and PNNs with non-spiking all-optical neurons. All-optical neurons utilize the nonlinear transfer function of optical materials or devices, for example, optical Kerr effects [28] or laser nonlinearity [12,15,16]. In contrast, optoelectronic neurons often exploit the nonlinearity in electronics or electronic circuits. Regarding energy efficiency, photonic spiking neural networks can be far more energy-efficient than non-spiking counterparts because they can incorporate event-driven communications with encoding methods such as spike-time-based coding and spiking-rate-based coding. This means that massively parallel and sparse neurons can be 'asleep' unless they have incoming signals which will 'wake' them to communicate with other neurons. Hence, if properly designed, PSNNs can incorporate event-driven neurons contrary to typical neurons that require a continuous supply of power to maintain neuron-like behaviors [8,10,12–19,23,25]. For instance, non-spiking neurons utilizing optical modulators require a constant supply of optical waves, typically from continuous-wave lasers [19,25], and even some spiking neurons utilize constantly powered lasers to extract spiking behaviors [8,12–16,20]. In this paper, we will design energy-efficient event-driven optoelectronic spiking neurons and achieve proof-of-principle demonstrations.

In a 2017 article [30], Miller reviews the possibilities of attojoule photonics and presents practical ~ 10 fJ/bit interconnect solutions with ~19 dB (80×) link loss budget exploiting quantum impedance conversion [31], where close integration of photonic-electronic integration leads to less than 1fF capacitance. Hence, it is possible to realize nanophotonic devices closely integrated with nanoelectronics to form a neuron at 10 fJ/bit energy efficiency with a fanout of 10-100. These nanoscale optoelectronic neurons can be far more energy-efficient than all-optical nonlinear neurons, which may require high optical energy per spike beyond 10 pJ/bit [17], and may suffer from challenges of isolating the input spikes [32]. As discussed in the following sections, an event-driven optoelectronic neurons can be designed to exploit quantum-impedance conversion utilizing nanolasers, nanophotodetectors, closely integrated with nanotransistors. Furthermore,

Optics EXPRESS

Table 1. state-of-the-art photonic non-spiking neuromorphic hardware research [9,18,19,23–29]

Neural network type	♦ Neuron mechanism (Nonlinear implementation) ☐ Neural network approach
Optoelectronic PNN	
Mitchell A. Nahmias et al. 2016 [19]	♦ Single PD accepts optical input and electrical wire to a laser. ☐ Wavelength-division multiplexed (WDM) network
Alexander N. Tait et al.2019 [23]	♦ Bi-PD accepts optical input and electrical wire to the MRR modulator with embedded p-n modulator and n-doped heater. ☐ Wavelength-division multiplexed (WDM) network
R. Amin et al. 2019 [25]	♦ Bi-PD accepts optical input and electrical wire to the ITO-based electro-absorption modulator. ☐ Wavelength-division multiplexed (WDM) network
I. A. D.Williamson et al. 2020 [24]	♦ The directional coupler directs a small portion of optical input to a single PD. The PD electrical wires to nonlinear signal conditioner and directly modulate the remaining portion of the original optical signal by a Mach-Zehnder Modulator (MZM). ☐ Mach-Zehnder interferometer (MZI) mesh network
Bin Shi et al. 2020 [26]	♦ Single PD receives optical input and sends the electrical signal to the external computer to perform the nonlinear function. ☐ Wavelength-division multiplexed (WDM) network
Xingyuan Xu et al. 2021 [27]	♦ N/A ☐ Time–wavelength multiplexing
All-optical PNN	☐
K. Alexander et al. 2013 [18]	♦ Realized sigmoidal artificial neurons by microdisk lasers ☐ N/A
S. Bhattacharya et al. 2015 [28]	♦ Realized neuron nonlinearity by Kerr type of centro-symmetric nonlinear medium. ☐ N/A
Xing Lin et al. 2018 [29]	♦ The nonlinear neuronlike behavior is implemented in a software-trained 3D-printed transmissive layer. ☐ Diffractive network
Yingjie Li et al. 2021 [9]	♦ The nonlinear neuronlike behavior is implemented in a software-trained 3D-printed transmissive layer. ☐ Multi-task diffractive network

the CMOS-compatible monolithic platform enables our design to be compatible with other Internet-of-Things (IoT) devices rather than operate in extreme working conditions [21].

As for the neuron model selection, Leaky-Integrate and Fire (LIF) or non-spiking artificial neuron models are appealing to photonic researchers due to the model's simplicity. In contrast, Intel announced that the electronic counterpart, Loihi2 [33] neuromorphic chip would support fully programable neuron models to enhance SNN's flexible learning capability. Thus, we were inspired by the Izhikevich model, which can implement various biological neuron behaviors. We pursue a neural network that could flexibly support various learning tasks in the future while demonstrating the performance using simple engineering machine learning tasks in order to compare and benchmark against published neuromorphic computing results. Our Izhikevich-inspired model preserves the biological neuron behavior by using three equations that govern the optoelectronic neurons. In addition, the optoelectronic neurons are designed to take excitatory and inhibitory inputs simultaneously. The inhibitory input serves as a suppress force to excitatory signals, a crucial neuroscience mechanism in the natural biological neurons in the brain. In

Optics EXPRESS

Table 2. state-of-the-art photonic spiking neuromorphic hardware research [7,8,10–17,20–22]

Neural network type	♦ Neuron mechanism (Nonlinear implementation) ☐ Neural network approach
Optoelectronic PSNN	
This work	♦ Bi-PD neuron with event-driven laser pump ☐ Mach-Zehnder interferometer (MZI) mesh network
Bruno Romeira et al. 2013 [7]	♦ Resonant tunneling diode (RTD) photodetector and laser diode (LD) semiconductor chips forming the RTD-LD excitable optoelectronic neuron ☐ N/A
J. M. Shainline et al. 2017 [21]	♦ Parallel nanowire detector (PND) and LED to form Integrate-and-fire neuron circuit ☐ Superconducting optoelectronic networks (SOENs)
Kengo Nozaki et al. 2019 [10]	♦ Applying photonic crystals (PhCs)-PD and PhC-electro-optic modulator (EOM) to achieve nonlinear functions ☐ N/A
Mitchell A. Nahmias et al. 2020 [8]	♦ Bi-PD neuron with two-section distributed feedback (DFB) laser requires continuous laser pumping. ☐ Wavelength-division multiplexed (WDM) network
M. Hejda et al. 2022 [22]	♦ RTD-PD (receiver) and RTD-LD (master) excitable optoelectronic neuron ☐ Shows all-to-one network layout
All-optical PSNN	
David Rosenbluth et al. 2009 [11]	♦ Nonlinear-fiber-based all-optical thresholder based on a modified nonlinear optical loop mirror as thresholding mechanism. ☐ N/A
Mitchell A. Nahmias et al. 2013 [12]	♦ VCSEL with an intracavity SA. ☐ Wavelength-division multiplexed (WDM) network
F. Selmi et al. 2014 [13]	♦ Micropillar laser with saturable absorber neuronlike excitable behavior on single and double pulse excitations. ☐ N/A
Bhavin J. Shastri et al. 2016 [14]	♦ Use graphene excitable fiber laser as a neuron. ☐ N/A
Joshua Robertson et al. 2017 [15]	♦ Presents VCSEL's inhibition behavior when the perturbation's arrival. ☐ N/A
Joshua Robertson et al. 2019 [16]	♦ Short temporal perturbations (stimuli) encoded in the applied bias current of a VCSEL triggers controllable spiking behavior. ☐ N/A
J. Feldmannet al. 2019 [17]	♦ Use the weighted input pulses to switch phase-change materials (PCMs) to perform integrate-and-fire functionality. ☐ Wavelength-division multiplexed (WDM) network
A. Jha et al. 2022 [20]	♦ Spiking neuron based on a hybrid graphene-on-silicon microring cavity ☐ Provide training method, but no photonic network presented

Optics EXPRESS

the following, Section 2 provides the details of the optoelectronic neuron model design and experimentally verifies by the proof-of-concept testbed neuron. Section 3 presents our nanoscale-PSNN design using the Section 2 neuron model, and Section 4 provides benchmarking for both the performance and energy consumption of our PSNN with other neuromorphic hardware.

2. Demonstration of bio-inspired & event-driven optoelectronic neuron

2.1. Optoelectronic neuron model equation and behavior mechanism

There are many established models representing biological neuron behaviors. The Leaky-Integrate and fire (LIF) model [34], the Hodgkin-Huxley model [34], and the Izhikevich model [35,36] are among the most studied for neural networks simulations. In particular, previous studies of photonic spiking neurons [12,37] have primarily relied on the LIF model due to its simplicity. The LIF model is easier to realize on electronic circuits, but the LIF model's refractory part is not easily realizable. The Hodgkin-Huxley model utilizes four ordinary differential equations for four state variables to closely resemble biological neurons, including ion channels, which is extremely computationally complex. The Izhikevich model introduces two partial-differential equations [35] to model most biological neurons in mammals' nervous systems effectively. However, these simple equations can also cause instabilities in analog signals from the equivalent optoelectronic circuits emulating the simulated neurons. Thus, in this paper, we are inspired by the Izhikevich model and introduce the following three Eqs. (1)-(3) that govern the optoelectronic neurons design of Fig. 1(a).

The membrane potential of the neuron is expressed as:

$$R_1 C_1 \frac{dv}{dt} = R_1(I_{exc} - I_{inh}) \; - R_1 K_1 \max\{0, u - V_{th1}\}^2 - v \tag{1}$$

Whereas the refractory potential can be realized as:

$$R_2 C_2 \frac{du}{dt} = R_2 K_3 max\{0, v - V_{th3} - u\}^2 - u \tag{2}$$

A directly modulated laser implements the output of the optoelectronic neuron. The amplitude of the laser output signal is determined by $I_{laser,}$ which can be approximated as:

$$I_{laser} = K_2 \max\{0, v - V_{th2}\}^2 \tag{3}$$

R_1, R_2, C_1, and C_2 are the values of the resistors and the capacitors. K_1, K_2 and K_3 are transconductance gain of the field-effect transistor (FETs). V_{th1}, V_{th2}, and V_{th3} are the threshold of the field-effect transistor (FETs) shown in Fig. 1(a). To efficiently utilize the model in the neural network simulations, it neglects the subthreshold behaviors of the transistors and the lasers.

Figure 1(a) conceptualizes the working principle of the optoelectronic neuron. All transistors' operating points are set to the saturation condition. The optical spiking inputs detected by the excitatory photodetector *PD1_exc* will generate photocurrents to be integrated by the capacitor *C1* in the membrane potential circuit (MPC) and discharged through the resistor *R1*. As the voltage (the membrane potential) of the MPC build-up to the threshold of FET1 and FET2 to drive current through laser, it will fire output spikes, and the capacitance C2 in the refractory potential circuit (RPC) will start to charge up. As the refractory potential builds up to the threshold of FET3, the membrane potential is reset and kept at the reset voltage level until the refractory potential discharges below the threshold of FET3. Figure 1(b) shows the corresponding plot for the neuron mechanism. The process of determining the FET parameters can be found in the Appendix section 1.

In addition to the excitatory connection, a critical feature included in the designed neuron model is the inhibitory connection. The inhibitory connection can be interpreted as a negative input to

Optics EXPRESS

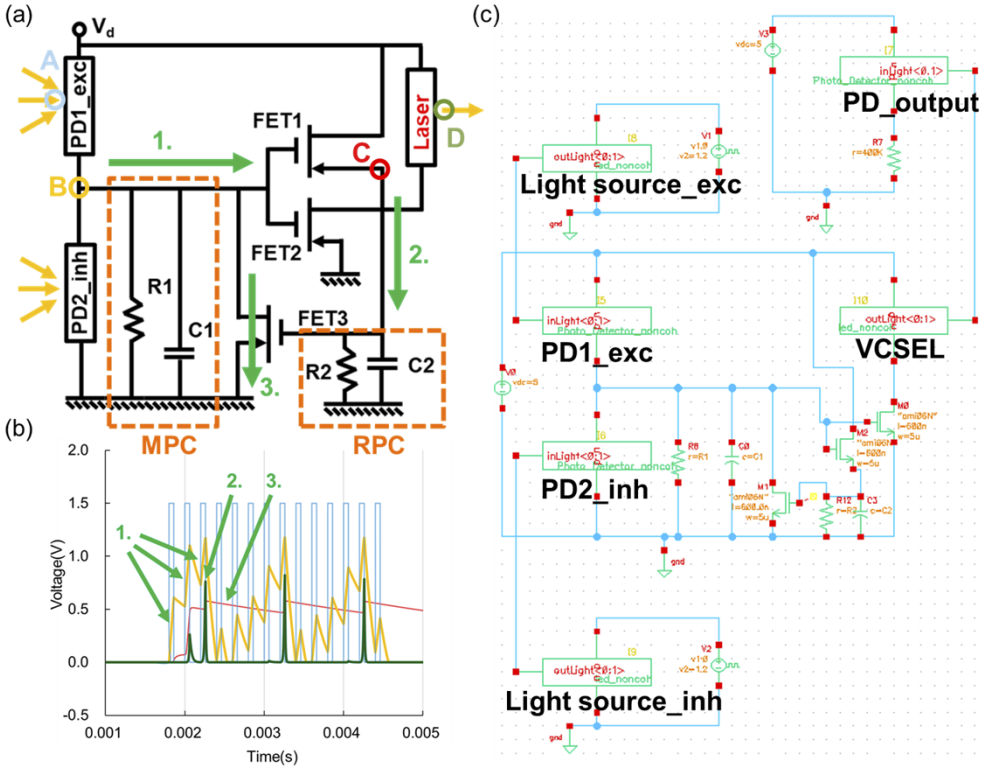

Fig. 1. (a) Optoelectronic neuron design consisting of three transistors (FET1, FET2, and FET3), one optical output (laser), and two optical inputs detected by photodetectors (PD1_exc for excitatory input and PD2_inh for inhibitory input). (step 1: Current from PD to MPC circuit, voltage built up on node B, step 2: Current from Vd to RPC circuit, voltage built up on node C, and step 3: Current from MPC circuit to GND) (b) corresponding neuron behavior from SPICE circuit simulations for steps 1-3. The color of each line represents the measurement point in the circuit. (A (Blue): optical excitatory input, B (Yellow): membrane potential (voltage at node B), C (Red): refractory potential (voltage at node C), D (Green): optical output) Step 1 leads to the accumulation of membrane potential. Step 2 leads to laser spike output. Step 3 leads to refractory potential increase and membrane potential decrease. (c) Cadence circuit-level simulation scheme using the Verilog-A model.

Optics EXPRESS

the neuron [38]. In SNN, inhibitory input signals decrease the accumulated membrane potential and make the neuron less responsive to the excitatory inputs. Contrary to the perceptron's negative inputs in ANNs, inhibitory inputs do not affect the neuron behaviors when the neurons are at the resting state. They are only effective when the neurons are excited beforehand. In the optoelectronic neuron hardware, excitatory and inhibitory inputs are received by the photodetectors pair (PD1_exc and PD2_inh) biased at the voltage V_d in reference to the ground. Therefore, inhibitory inputs are ineffective when the optoelectronic neuron is at its resting state ($v(t) = 0$, where $v(t)$ is measured at the Membrane Potential).

There are two main differences between the original Izhikevich model and the presented optoelectronic neuron model. First, the quadratic positive feedback part of the membrane potential in the Izhikevich model is truncated because it leads to the instabilities in the (optoelectronic neuron) circuit when the input signal is large. Here, the Izhikevich model considers the spike output solely as $v(t)$. In contrast, for optoelectronic neurons, another stage of the circuit involving FET2 is required to map the membrane potential voltage to drive electrical current into the laser. The second difference lies in its threshold behavior, as shown in Eq. (3). In the original Izhikevich model, both variables, $v(t)$ and $u(t)$ are immediately set to their reset values when the membrane potential reaches the threshold voltage. On the other hand, our optoelectronic neuron model will gradually reset the two variables in time by involving FET3, R2, and C2 in the circuit.

2.2. Optoelectronic neuron model simulation results

To pursue bio-inspired optoelectronic neuron designs inspired by Izhikevich model and simulate neuron behaviors, we built a two-level model consisting of the neuron circuit-level and the neural network-level simulation modules.

The circuit-level simulations focus on the neuron circuit behavior. As Fig. 1(c) illustrates, we built a compact model in Verilog-A for the optoelectronic neuron. Here, we included details of the physical device model parameters for transistors and the laser. We simulated the model using LTSpice and Cadence to emulate the optoelectronic spiking neuron behaviors under continuous and discrete arbitrary spiking patterns. As an initial simulation example, we set the neuron behavior parameters to accumulate three contiguous input spikes with spike width 60µs to cause the membrane potential to rise above the threshold for firing an output spike. We can easily change this neuron behavior by adjusting the capacitor and resistor values in the circuit to meet other neural network requirements.

Subsequently, we used the Nengo simulator [39] for the neural network-level simulations, which provides a platform for performing neural network simulations, including neuron models, spiking neural networks' learning algorithms, and synaptic interconnections. To test our neuron performance on a neural network for an actual application, we imported the Eqs. (1)-(3) and built our neuron model in Nengo. In addition to these equations, variables u and v are clipped between '0' (the ground level), and 'V_d' (the maximum voltage level), and the minimum spike width of the optoelectronic neuron is scaled up to be compatible with the Nengo simulation platform.

Figure 2 shows the simulated neuron behavior of our neuron model using different parameters. These behaviors match the various firing patterns of the Izhikevich neuron model [35]. For the following experiments, we only use the behavior of the regular spiking neuron (Fig. 2(a)) for engineering tasks. The simulated result in Fig. 2(a) matches nearly perfectly with experimental neuron results presented in Section 2.3. The maximum spiking rate is limited to 10 kHz for the neuron testbed demonstration due to the testbed setup wire connections and the available off-the-shelf transistors. As expected from Izhikevich's model [35], various spiking patterns—fast, slow, and burst spiking neurons can be implemented by tuning the parameters in Eqs. (1) and (2). The various parameters for emulating different neuron behaviors are actual physical parameters to be implemented in the optoelectronic neuron circuit. To verify the neuron responses in the simulations and experiments, we used an input spike train with four groups of spikes with a

Optics EXPRESS

maximum spiking rate of 10 kHz. The number of spikes in each group is 14, 5, 3, and 1 for the 1st, 2nd, 3rd, and 4th spike groups, respectively. There is a guard time of 3 ms between the groups. This guard time is to demonstrate that the discharge of the accumulated photocurrent shuts off the spiking output.

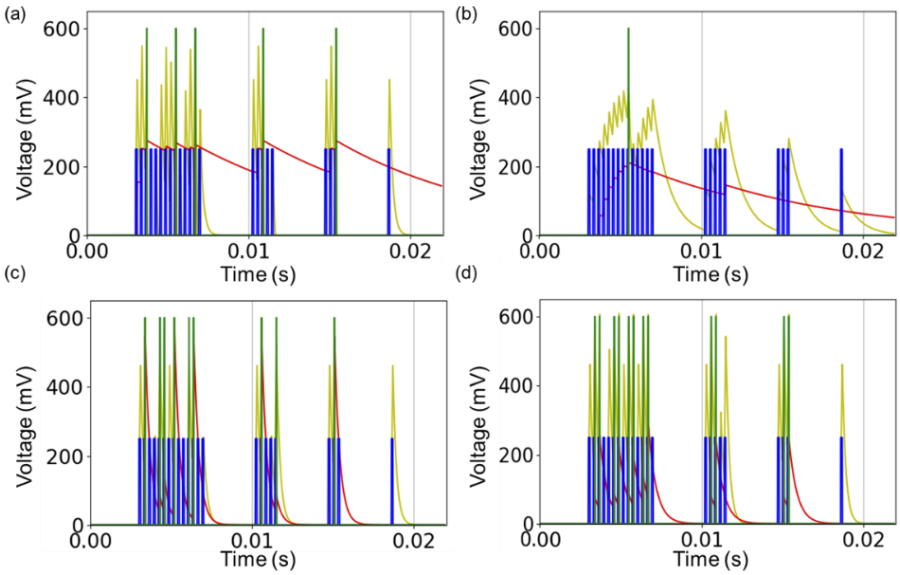

Fig. 2. Spiking neuron simulation performance. (a) is the regular spiking, which we will experimentally demonstrate in this paper. (b)(c)(d) is the neuron behaviors of our neuron model with various input spike parameters. (b) is low output spiking, (c) is fast-spiking, (d) is burst-spiking output spiking simulation results. The time scale is scaled down to match the experiment. For the following experiments, we only use the behavior of regular spiking (Blue: excitatory input, Red: refractory potential, Yellow: membrane potential, Green: neuron output)

2.3. Testbed experimental results

An experimental setup for our proof-of-principle neuron demonstration utilizing a commercial laser, commercial photodetectors *PD1_exc* and *PD2_inh*, and an electronic neuron circuit formed by discrete electronic components on a printed circuit board is built for testbed neuron. The detail of the experimental setup can be found in Appendix section 2.

Based on our circuit-level simulation utilizing the LTSpice tool and neural network level simulation, we conducted comparative studies of simulation vs. actual hardware experimental results of our optoelectronic neuron behavior on Fig. 3. For this purpose, we again created four different spiking groups as mentioned in the previous section— the number of spikes in each group is 14, 5, 3, and 1 for the 1st, 2nd, 3rd, and 4th spike groups, respectively, with a guard time of 3 ms between the groups as similar to the Nengo simulations in Section 2.2. Figure 3 (a)(b) summarizes the results with the excitatory input signal only. Figure 3 (c)(d) summarizes the results with both excitatory and inhibitory signal inputs.

Figure 3(a)(b) illustrate simulated and experimental results, including the refractory potential and the optical output from the laser in addition to the optical excitatory input and the membrane potential. Here we observe that the optical output spikes fire when the membrane potential reaches the threshold, but more importantly, the refractory potential rises in response to the spike output. This indicates that the firing of the optical output spikes occurs only after the refractory

Optics EXPRESS

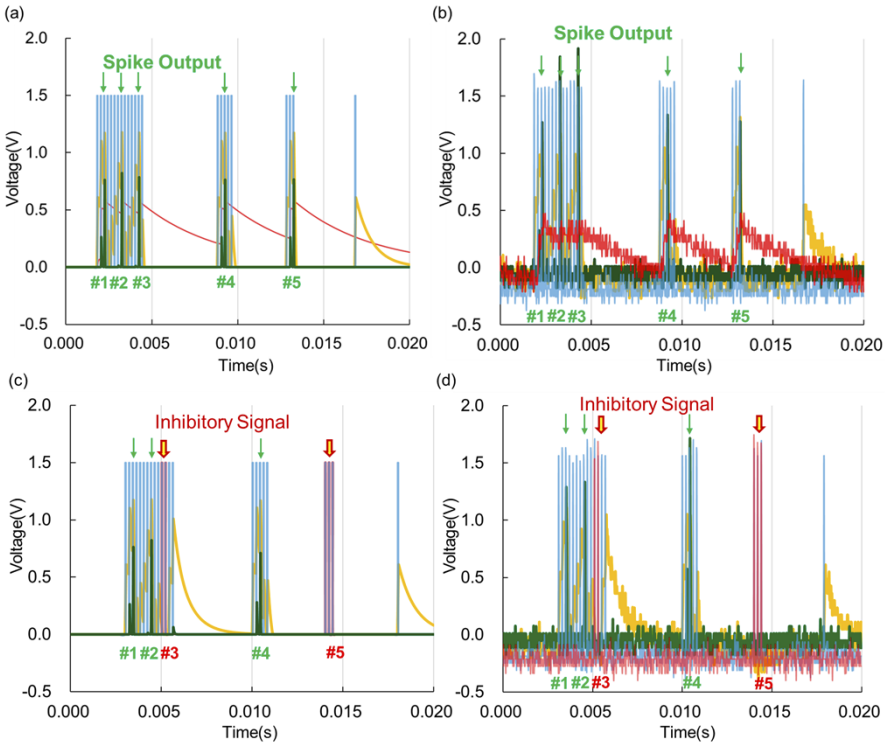

Fig. 3. Neuron spiking behavior simulation and experiment. The input spiking pattern consisting of the four groups in sequence (14, 5, 3, and 1 spike in each group). (a) simulated inputs, membrane potential values, refractory potential, and optical outputs with only excitatory signal input. (b) measured simulated inputs, membrane potential values, refractory potential, and optical outputs with only excitatory signal input. (Blue: optical excitatory input, Red: refractory potential, Yellow: membrane potential, Green: optical output) (c) simulated inputs, membrane potential values, refractory potential, and optical outputs with excitatory and inhibitory signal input. (d) measured simulated inputs, membrane potential values, refractory potential, and optical outputs with excitatory and inhibitory signal input (Blue: optical excitatory input, Red: optical inhibitory input, Yellow: membrane potential, Green: optical output)

Optics EXPRESS

period. This proves that our neuron model correctly represents the general dynamics of the Izhikevich model, including the refractory period. We can also see the experimental result in Fig. 3(b) shows the spike output behavior closely matching the LTSpice results in Fig. 3(a). We repeated the simulation and experiment using both excitatory and inhibitory input signals by adding the inhibitory signal to *PD2_inh*. Figure 3(c)(d) show the optical output spikes absent at #3 and #5 due to the presence of the inhibitory signal. The inhibitory signal (Red) cancels out the effect of the excitatory signal (Blue) to suppress the output spike (Green). This neuron behavior is consistent with the commonly seen functionality of biological inhibitory neurons. The detail of the testbed neuron experiment can be found in the Appendix section 2.

3. Optoelectronic photonic spiking neural networks (PSNNs)

Previous Sections show that the proof-of-concept testbed version of our optoelectronic neuron using off-the-shelf components can work well to emulate the bio-inspired neuron model with event-driven excitation of lasers. In this section, following the design procedure for testbed neurons discussed in Section 2, we considered two versions of the nanoscale-PSNN: foundry-PSNN and nano-PSNN. We also investigated the feasibility and detailed designs of PSNNs with nanoscale laser and transistors to quantify the potential benefit of energy efficiency and high throughput.

3.1. Foundry-PSNN

The foundry-PSNN is designed on a 90 nm CMOS platform with monolithically integrated silicon photonics. Figure 4 shows the detailed design of the 90nm Foundry-PSNN. The Foundry-PSNN architecture consists of a cascaded MZI mesh synaptic interconnect network and an event-driven spiking neuron layer which is designed to operate at the maximum 10 GHz spiking rate. Figure 4(a) is the 4×4 rectangular MZI mesh with embedded Bi-directional PD with a transimpedance amplifier (TIA) to support forward propagation and backpropagation training. Figure 4(b) is a neuron chip with multiple neuron designs. The neuron consists of 6 transistors (Fig. 4(e)) with disk or ring modulated laser or micro-transfer-printed III-V on silicon quantum dot (QD) lasers [40,41]. The QD laser will be fabricated by transfer printing after the silicon photonic fabrication. The details of the Verilog-A circuit and Nengo simulation of the foundry neuron can be found in the Appendix section 3.

3.2. Nano-PSNN

To envision future neuron and neural network design, we proposed an aggressive theoretical nano-PSNN. The nano-PSNN exploits attojoule photonics with quantum impedance conversion [42] by close integration with electronics with < 1 fF capacitance. Our neuron's nanoelectronics capacitance model includes the load capacitance on the photodetector, membrane capacitor, and transistor gate capacitance. The photodetector's load capacitance is around 0.1fF [4], and the simulated membrane capacitor is 0.5fF. The value of the transistor gate capacitance is derived from IRDS2020 [43]. Hence, it is possible to realize nanophotonic devices closely integrated with nanoelectronics to form a neuron at 10 fJ/bit energy efficiency with a fanout of 10-100 following the concept outlined by [30]. Furthermore, when using low-loss waveguides, the neuron is capable of communicating with other neurons nearly independently of the communication distance at high speeds (> 10 GHz).

Figure 5(a) illustrates the structural schematic of the proposed nano-optoelectronic neuron. We propose a low-Q nanophotonic crystal PD based on Ge/Si cavity which had similar work [44], as shown in Fig. 5(b). An ultra-low capacitance nano-cavity PD can generate sufficiently large voltage without an amplifier when combined with a high impedance load [31]. Based on this configuration, ~0.1fF capacitance is expected in the resonant nanophotonic PD. In addition to the ultra-compact size and extremely low capacitance, the extremely short electrical contact

Optics EXPRESS

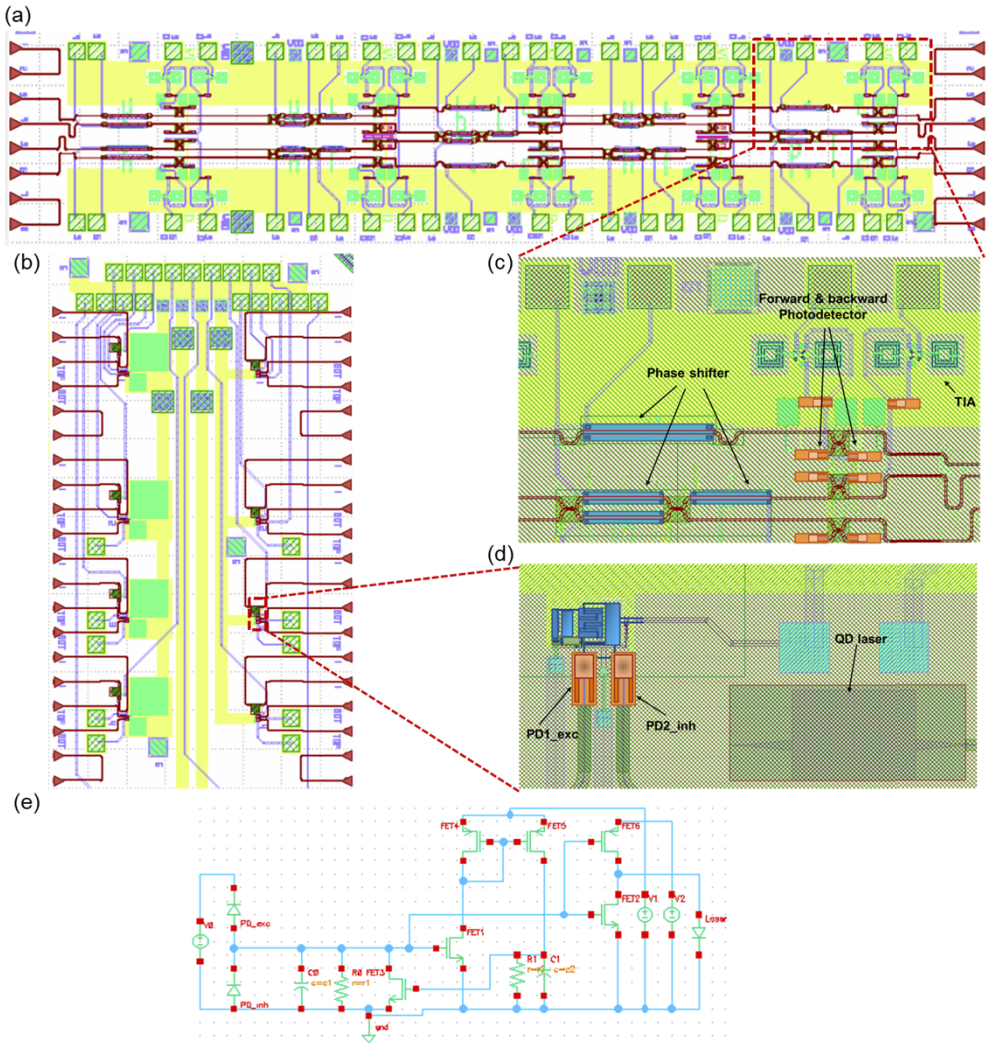

Fig. 4. The foundry-PSNN architecture consists of cascaded layers of a (a) MZI mesh synaptic interconnect network and (b) neuron layer. (c) is the detailed structure of 4×4 rectangular MZI mesh. The forward and backward PD is embedded for neural network training. (d) is one of our optoelectronic neuron designs used to connect micro-transfer-printed quantum dot lasers as neuron output. (e) is our foundry optoelectronic neuron circuit designs with Verilog-A model.

Optics EXPRESS

between PDs and next stage FET transistors can further guarantee exceptionally low circuit power consumption [30]. We anticipate such a system can operate beyond 10 GHz bandwidth with ultra-low energy consumption of < 1 fJ/bit.

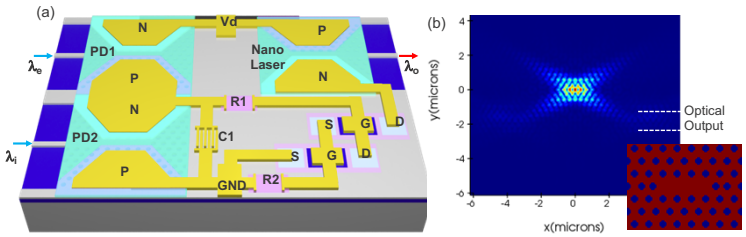

Fig. 5. (a) A schematic of the proposed optoelectronic neuron structure based on Fig. 1 including two Ge/Si photonic crystal enhanced photodiodes for excitatory and inhibitory inputs and two FETs on SOI for thresholding and spiking signal generation by triggering a quantum-dot laser with photonic crystal patterns etched for in-plane emission. (b) a photonic crystal cavity laser will be fabricated on silicon utilizing hybrid integration by transfer-printing to realize hybrid III-V/silicon nanophotonic devices.

An example of the proposed Nano-PSNN will consist of nanoscale-Optoelectronic Neurons, self-optimizing [45] nanophotonic synaptic interconnect network with 2×2 Nanoelectromechanical systems (NEMS)-MZI including tunable NEMS phase shifters [46]. An essential part of a neural computation scheme provides this necessary set of weighted connections from the nanoscale optoelectronic neuron outputs (lasers) to the inputs of the next layer of neurons (photodetectors). All devices, including the FET, Ge/Si nanodetectors, and waveguides, will be on a silicon-on-insulator (SOI) platform. At the same time, nanolasers will utilize hybrid InAs/AlGaAs quantum-dot on SOI structure with photonic crystal patterns etched in on silicon [47,48]. The hybrid InP Multi-Quantum-Well / silicon semiconductor optical amplifier demonstrated in [49] utilized a similar fabrication process. The absence of capacitive charge associated with the interconnect wires [30] can drastically reduce future nanophotonic neurons' power consumption.

3.3. *Nanoscale neuron energy consumption*

Neuron's power consumption varies differently among devices. Current state-of-the-art research neurons, such as [25], usually require 12-20 transistors [50] in a neuron circuit design to complete the spiking neuron behavior, while our optoelectronic neuron structure only requires fewer transistors to achieve neuron spiking. We can separate our neuron structure into three parts: photodetector, transistor circuit, and laser sections. The testbed version uses an off-the-shelf photodetector, transistor, and laser. At the same time, the Foundry-PSNN consists of 90 nm CMOS with monolithically integrated silicon photonics, including MEMS phase shifters for MZI synaptic interconnects realized by 90 nm silicon photonic CMOS process. Thus, Foundry-PSNN is a miniaturization of the current testbed-PSNN utilizing a commercial foundry with modified post-fabrication to realize MEMS MZI synaptic interconnects and micro-transfer-printed quantum dot lasers [51].

The power consumption of a neuron depends on the total energy required to generate certain spiking behaviors, such as spikes at a certain frequency and with a certain amplitude and duration. Therefore, we calculated both neurons' dynamic power and static power consumptions. The dynamic power is the power consumed when the neuron is generating spikes (transistors at ON state). The static power is the power consumed when the neuron is staying at the rest state (transistors at OFF state). Based on our calculation, foundry neuron and nano neuron requires a minimum of 21.09 fJ/spike and 200 aJ/spike, respectively, input energy to generate a spike output. The continuous spiking average power for foundry neuron and nano neuron is 714μW

Optics EXPRESS

and 8.14μW, respectively. The parameters, energy values, and power values of foundry and nano neurons based on the maximum possible efficiencies are listed in Table 3. The detailed energy calculation can be found in the Appendix section 5.

Table 3. Foundry and nano neuron maximum possible energy and power consumption

	Foundry neuron	Nano neuron
Maximum spiking rate	10GHz	10GHz
Spike width	10ps	10ps
C1	60fF	500aF
PD load capacitance	2.1fF	100aF
PD responsivity	0.7A/W	1A/W
FET parasitic capacitance	6fF	1.1aF
Leakage current (voltage at power supply)	3.18μA (2V) 580pA (0.5V)	10nA (1.4V)
Neuron on current (voltage at power supply)	423.4μA (2V) 22.4μA (0.5V)	31.27μA (1.4V)
Minimum dynamic input energy to generate a spike output	21.09fJ/spike	200aJ/spike
Peak dynamic optical input power required at 10GHz	3.24dBm	-17dBm
Dynamic power in circuit when transistors are on	858μW	43.78μW
Static power when neuron is off	6.36μW	14nW
Continuous spiking average power	714μW	8.14μW
Waveguide loss in design platform	-1.6 dB/cm	-4 dB/m
Loss per MNIST experiment neural network layer	10dB	0.2dB
Peak dynamic optical input power required at 10GHz with network loss compensation	13.24dBm	-16.8dBm
MNIST handwriting recognition energy per image	5.2nJ	4.4pJ

3.4. Scalability and cascadeability of proposed PSNN

To address the possibility of scalability and cascadeability, we look into the photonic waveguide loss of our proposed PSNN. The photonic waveguide loss will serve as an indicator for the potential requirement on the neuron's output optical power. Our proposed PSNN will apply the compact tensorized optical neural network (TONN) exploiting the tensor-train decomposition architecture. According to the research [52], the proposed TONN uses 79× fewer Mach–Zehnder interferometers (MZIs) and 5.2× fewer cascaded stages of MZIs compared with the conventional ONN. This architecture has also proven robust to practical hardware imprecisions [53].

Suppose we apply our neuron to Diehl and Cook's MNIST handwriting recognition experiment [54] with 784-400-10 neuron network architecture, based on the parameter provided by the foundry, which the waveguide loss is around 1.6 dB/cm and MZI length 400μm per stage. In that case, it leads to a 10 dB loss per neural network layer. That means the foundry-PSNN neurons require 10× more dynamic optical output power to compensate for the network loss. As for nano PSNN, we assume waveguide loss is 4 dB/m [55] and MZI length 350μm per stage. That leads to a 0.2 dB loss per neural network layer, which means nano-PSNN neurons only require 1.05× more dynamic optical output power to excite the next layer's neurons successfully. As a result, the energy required per sample will be 5.2nJ and 4.4pJ, respectively. The parameters and power values for MNIST experiment are also listed in Table 3. The detail of the experiment energy calculation can be found in the Appendix section 4.

Optics EXPRESS

4. Benchmark of optoelectronic photonic spiking neural networks (PSNNs)

This benchmarking section includes two parts. In the first part, we simulate the training and the inference capability of optoelectronic neural networks, including the designed optoelectronic neuron model. Since the optoelectronic neurons serve as the neural network's activation function, the neuron's behavior will affect the inference and training accuracy. The second part of the benchmarking will address the optoelectronic neural networks' energy efficiency and throughput.

We will consider the case of optoelectronic neurons used in our actual testbed experiments employing bulky commercial lasers and transistors. More importantly, the new case of nanoscale optoelectronic neurons integrates nanotransistors and nanophotonics. Thus, we consider three types of Izhikevich-inspired optoelectronic neurons. The first is the optoelectronic neuron described for our experimental testbed demonstration (Testbed neurons). The second type is the foundry-implementation of an optoelectronic neuron that utilizes the same design process using the Izhikevich-inspired model but incorporates a commercial 90 nm silicon-CMOS-photonic foundry process (Foundry-neurons). The third is the envision nanoscale optoelectronic neuron (denoted as Nano-neurons) utilizing the quantum impedance conversion between nanoscale detectors and nanoscale (5nm) FET circuits driving nanoscale lasers (e.g., photonic crystal lasers). Further, for optical synaptic interconnections, we consider optical Mach-Zehnder interferometric meshes in neural networks [56] [57] capable of achieving near-zero static energy consumption by incorporating optical phase shifters with optical MEMS devices [58] or with optical phase change materials [42,59–62].

4.1. Performance of proposed photonic neural network

Figure 6 shows the neural network structure with the optoelectronic neurons and optical synaptic interconnects. Figure 6(a) is the conventional feedforward neural network structure with inhibitory and excitatory signal connections. Figure 6(b) has one-to-one inhibitory signal connections in the hidden layer instead. In Fig. 6(a), the optical synapses transmit both inhibitory and excitatory signals. In Fig. 6(b), only the excitatory signals are transmitted on the optical synapses. The inhibitory neuron only responds to a certain excitatory neuron and transmits the inhibitory signal to the rest. Thus, there are the same numbers of excitatory and inhibitory neurons in this topology in the hidden layer. A separate optical interconnect can realize the inhibitory signals without the synaptic weighting (can be realized by 1:N power splitters). The hidden layer will perform the winner-takes-all setting by sending inhibitory signals to other excitatory neurons.

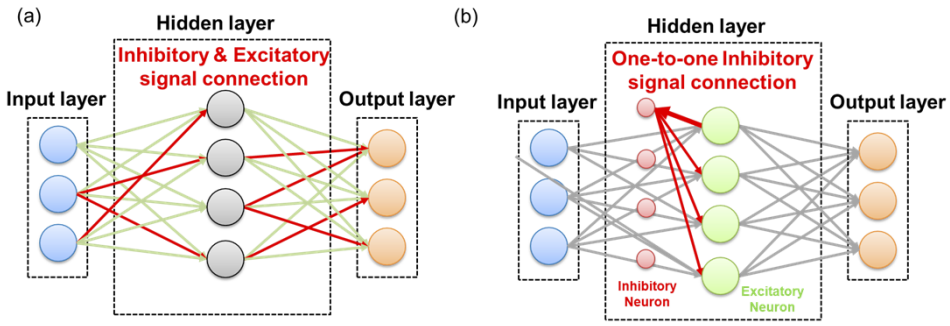

Fig. 6. Two neural network schemes. (a) is a neural network with inhibitory and excitatory signal coexist in the same interconnection (b) is a neural network with inhibitory and excitatory signals in separated interconnection. The inhibitory neuron only connects to one excitatory neuron (one-to-one inhibitory signal connection).

The synaptic interconnects between our optoelectronic neurons exploit the results from [63], which assign weight values by changing each mesh's phase shifter. Conventional multiport

Optics EXPRESS

interferometers use electro-optical or thermo-optical tuning that will require a supply of constant power to maintain the phase states. However, integrating non-volatile optical MEMS [31] into our optical synaptic design will only require power when weight value changes on the neural network. This approach can significantly reduce the power consumption during training and perform zero power consumption on PSNNs during real task work.

The feedforward neural network structure with inhibitory and excitatory signal connection is benchmarked with FC Nets and ConvNets [64–66] neural network architectures. Figure 6(a) shows the network architecture we used in Nengo and trained with supervised learning. Table 4 summarizes the inference benchmarking results with supervised learning. Here, we apply the ANN-to-SNN conversion training method. The FC Nets and ConvNets ANN are trained using the ReLU activation function with backpropagation methods. Next, we transferred the trained weight values to the proposed PSNN for inference testing. Table 4 compares the accuracy of Nengo's abstract LIF neuron, proof-of-concept testbed neuron, foundry neuron, and future energy-efficient optoelectronic neuron. The PSNN with future energy-efficient optoelectronic neurons uses an abstract model to be simulated in the Nengo simulator. Our PSNN with an optoelectronic neuron model can reach 97% accuracy and achieve comparable results to those with LIF neuron on feedforward network. But we fall short of LIF neuron on convolution network because ANN-to-SNN conversion training method in the Nengo simulator optimized for LIF neuron. In future works, this problem can be addressed by introducing proposed neurons' specific behaviors during the training stage of the SNN. On the other hand, as mentioned in the Introduction, our main goal is to design and demonstrate bio-plausible optioelectronic neurons for brain-derived neuromorphic computing caplable of learning flexibly on unexpected and various tasks. Future studies will include benchmarking of accuracies on such various learning tasks. The confusion matrix with the best accuracy (FC Nets 1000-500-10) is shown in Fig. 7

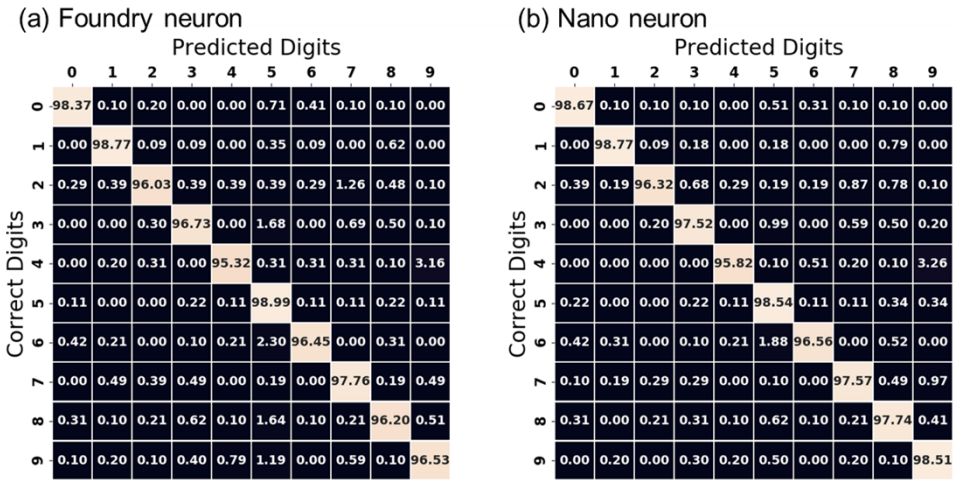

Fig. 7. confusion matrix for (a) foundry neuron and (b) nano neuron

For benchmarking of the one-to-one neural network of Fig. 6(b), we employed Diehl and Cook's MNIST handwriting recognition experiment [54] with 28 by 28 input neurons, 400 neurons in the hidden layer, and 10 neurons for output classification on our BRIAN simulator. We used unsupervised learning Spike-timing-dependent plasticity (STDP) [54] for training. We apply the winner-takes-all setting in the hidden layer. The excitatory neuron will send spike messages to its inhibitory neuron and provide feedback to suppress other excitatory neurons. By training with 60,000 MNIST datasets, we can achieve 90% accuracy [67] on this test (Table 5).

Optics EXPRESS

Table 4. Neural network performance results on Nengo simulator based on supervised learning [64–66] with ANN to SNN conversion

Type	Neuron Network Type	Artificial Neuron	LIF Spiking Neuron	Testbed Spiking Neuron	Foundry Spiking Neuron	Nano Spiking Neuron
FC Nets	300-100-10	95.3	97.3	96.24	90.28	96.44
	1000-500-10	99.51	97.55	97.04	97.12	97.55
	1500-1000-500-10	99.54	98.03	90.72	91.36	92.31
CONV Nets	LeNet-1	98.3	97.5	93.01	92	93.24
	LeNet-5	99.05	98.03	83.59	84.49	90.95

We expect that fine-tuning the synaptic weight values and increasing the number of layers or neurons could improve the accuracy. However, this is out of the scope of this paper.

Table 5. Neural network performance results on Brian simulator based on unsupervised spike-timing-dependent plasticity (STDP) learning [53]

Network \ Neuron Type	Diehl and Cook Test Accuracy	Optoelectronic Neuron Test Accuracy
784-400-10	95%	90%

The benchmarking results show that our optoelectronic neuron can obtain accuracy similar to state-of-the-art LIF neuron-based neural networks on supervised and unsupervised learning neural network architectures. These results support that our neuron design can preserve more biological neuron behaviors while keeping the competitive performance as the conventional simplified neuron model.

4.2. Benchmarking of energy efficiency and throughputs of proposed PSNN

We calculated the energy efficiency and the computing throughput of our neuromorphic computing platform compared to other electronic counterparts. As mentioned earlier, we considered both the neurons implemented in our current testbed, the foundry, and the future nanoscale optoelectronic neurons.

Our neural network's energy consumption has two main contributions: the photonic MZI mesh for all-to-all synaptic interconnect and the neuron for nonlinear function. Photonic MZI mesh networks have both dynamic and static power consumption. Conventional silicon photonic MZI mesh networks use thermo-optical-tuning of the phase-shifters, which typically consume a continuous 10 mW power supply to keep the desired state of the phase-shift (~10 mW static power per MZI). Our testbed version power benchmark is based on the thermo-optically tuned MZI mesh. The static energy degrades the overall benchmark power-efficiency performance. However, for optical phase shifters consisting of silicon photonic MEMS [58], the weight values can be remembered by latching the MEMS components with negligible static energy consumption. The impact of energy consumption on latching MEMS's reconfiguration is also negligible since such reconfigurations are expected to be infrequent (below 0.1% duty cycle). Similar to optical MEMS-based MZI mesh, optical phase change materials (OPCM) such as GeSbTe (GST) [68] or GeSbSeTe (GSST) [42] can also achieve zero static energy synaptic interconnects. In contrast, dynamic energy consumption is required during the training phase of the neural network. Once the training process is completed, there is no additional power needed to maintain the phase-shift states. In the following, we discuss benchmarking of energy efficiency, throughput, and accuracy of the neural networks available in the literature, our proof-of-concept testbed neural networks, our foundry implementation of the PSNN with MEMS MZIs (labeled as Foundry-PSNN), and

Optics EXPRESS

our futuristic nanoscale optoelectronic neuron based PSNN with MEMS MZIs (labeled as Nano-PSNN).

Table 6 lists the energy consumption collected from Table 2 spiking neuromorphic hardware research and adds electrical spiking neuromorphic hardware [69,70] for comparison. We include the energy consumption data on the neuron, whether the neuron required continuous power supply, neuron's maximum spiking rate, and energy per inference requirement for a neural network for comparison. Note that for neurons requiring a continuous power supply, there may be extra waste energy, which is not included in the calculation of neuron energy consumption. Our design has excellent energy efficiency per spike while maintaining a high spiking rate. However, there is no standard benchmarking method for spiking neural networks because, while artificial neural networks utilize synchronous multiply-and-accumulate (MAC) operations for benchmarking, SNN computations are based on spike events. Besides, as mentioned in the introduction, state-of-the-art photonic neuromorphic hardware usually requires a continuous laser source for chip operation, which most of the energy calculations in these papers did not consider. Furthermore, most of the current research shown in Table 1 and Table 2 are not implemented the entire end-to-end programmable photonic processor. That means the calculation can only be limited to chip-level benchmarking. As a result, to make the comparison fair, we considered three kinds of preliminary benchmarks that compare the neural network approaches against our PSNN based on MNIST dataset.

In the first benchmark, we target spike-event-based computations, in which all the competitors use SNN for image classification. We defined an operation (OP) as one spike event in the neural network. One spike event starts from the neuron aggregating all the inputs spikes from the previous layer, and it ends with generating an output spike for the next layer. Using this approach, we eliminate the difference between classical von Neumann hardware and neuromorphic non-von Neumann hardware. Furthermore, to make the benchmarking fair between different hardware solutions, we only consider the power consumption for inference in the neural network rather than the training power consumption since different training algorithms and offline training systems for PSNN would lead to different results.

Figure 8(a) shows the benchmarking results based on our first benchmarking method. Compared to state-of-the-art neuromorphic hardware [5,6,71–74], our proof-of-concept testbed version of PSNN can achieve around 0.001 GOP/J energy efficiency at 0.001OP/s/mm^2 computing speed. Our current Foundry-PSNN version can achieve around 5×10^4 GOP/J energy efficiency at 10 GHz spike-event speed. If we further utilize sub-10 nm transistor and closer integration of nanophotonic and nanoelectronics, we can achieve over 10^6 GOP/J energy efficiency at 10 GHz spike-event speed. We also list several SNN hardware in the plot for comparison. Our Foundry-PSNN version outperforms all state-of-art spiking neural network hardware [5,6] by at least 1000x in terms of energy efficiency.

The second and third methods of benchmarking are previously presented at [75], which is based on calculating the conventional multiply-accumulate (MAC) operation and the total energy consumption for a specific task (inference benchmark). Especially inference benchmarking [70,76–81], which aims to analyze how much energy is required per image sample, is more objected in the spiking neural network benchmarking. We add our envision nano neuron into the plot and notice the correlation on Fig. 8(b) that higher accuracy results, such as TrueNorth (Case1), usually require more energy per image for the same target dataset. However, we can achieve lower energy per image in our foundry and future neuron version while keeping the same accuracy. Our currently developing foundry neuron version can reach 1 MNIST image per nJ and even higher in future neuron versions (10pJ). The details of the third benchmarking on calculating the MAC operation can be found in the Appendix section 5.

Optics EXPRESS

Table 6. benchmark of state-of-the-art photonic and electronic spiking neuromorphic hardware research

Neural network type	Neuron energy consumption[a]	Required continuous power supply?	Neuron max spiking rate	Neural network energy per inference
Optoelectronic PSNN				
This work (foundry neuron)	21.09 fJ/spike	N	10 GHz	5.2 nJ (MNIST)
Bruno Romeira et al. 2013 [7]	N/A	N	1 GHz	N/A
J. M. Shainline et al. 2017 [21]	20 fJ/spike[b]	N	20 MHz	N/A
Kengo Nozaki et al. 2019 [10]	42 aJ/bit	Y	40 G/bit	N/A
Mitchell A. Nahmias et al. 2020 [8]	260 fJ/MAC	Y	1 TMACs/s	N/A
M. Hejda et al. 2022 [22]	0.1 pJ/spike	N	10 GHz	N/A
Electrical PSNN				
Loihi [69]	226.3 pJ/MAC (81 pJ)	N	50 MMAC/s	2.47 mJ (MNIST)
TrueNorth [70]	1300 pJ/MAC	N	46 MMAC/s	268 nJ (case2 MNIST)
All-optical PSNN				
David Rosenbluth et al. 2009 [11]	N/A	Y	5 GHz	N/A
Mitchell A. Nahmias et al. 2013 [12]	N/A	Y	400 MHz	N/A
F. Selmi et al. 2014 [13]	N/A	Y	3.33 GHz	N/A
Bhavin J. Shastri et al. 2016 [14]	N/A	Y	500 MHz	N/A
Joshua Robertson et al. 2017 [15]	N/A	Y	2.83 GHz	N/A
Joshua Robertson et al. 2019 [16]	N/A	Y	2 GHz	N/A
J. Feldmannet al. 2019 [17]	300 pJ	Y	0.5 MHz	N/A
A. Jha et al. 2022 [20]	0.7 pJ	Y	40 GHz	N/A

[a]energy per spike is used by analog device and energy per bit or per MAC is used by digital device
[b]This data includes energy to maintain at 2 K temperature

Optics EXPRESS

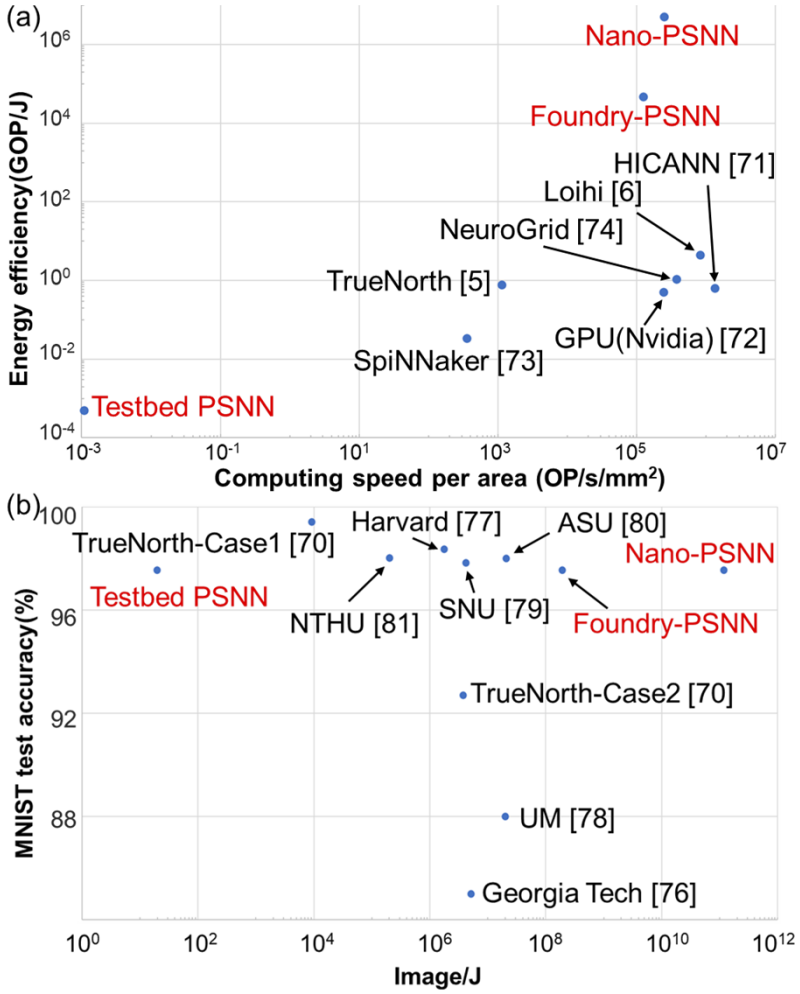

Fig. 8. (a) Energy efficiency benchmarking method based on spike-event, including the PSNN, described in this paper (denoted in red) in comparison with results described in the literature. (b) Inference benchmarking targeted at MNIST image dataset [70,76–81].

Optics EXPRESS

5. Conclusion

This paper presents brain-inspired neuron design, simulation, and experimental testbed demonstrations based on the Izhikevich's model implemented on optoelectronic hardware. We investigate neurons at three different implementations: (a) proof-of-principle testbed prototype neurons utilizing off-the-shelf components, (b) 90 nm monolithic silicon photonic-CMOS integrated circuits, and (c) future nanoscale (e.g., 3 nm) monolithic photonic-electronic integrated circuits operating in an event-driven-manner. We experimentally demonstrate that the testbed prototype neuron with excitatory and inhibitory photodetectors achieves spiking behaviors very closely resembling the theoretically simulated behaviors following the Izhikevich's model. We also design 90 nm photonic-CMOS IC neurons that exhibit spiking behaviors following the Izhikevich's model when carefully designed based on the 90 nm commercial foundry process-design-kit (PDK). Finally, we project the neuron behavior based on 3 nm photonic-CMOS IC neurons also following the Izhikevich's model. The foundry neuron showed 21.09fJ/spike energy efficiency, whereas the 3 nm photonic-CMOS IC neurons can achieve 0.2 fJ/spike energy efficiency when exploiting quantum impedance conversion. We then place these neurons in photonic synaptic interconnect networks to create PSNN equipped with non-volatile reconfigurable elements such as phase-change-materials in Mach-Zehnder interferometers. The preliminary estimations of energy efficiency and throughput comparisons based on the MNIST dataset showed that the proposed 90 nm silicon CMOS-photonics and 3 nm photonic-electronic PSNNs outperform any of the state-of-the-art neuromorphic hardware by orders of magnitude.

For illustration, authors Yun-Jhu Lee, Mehmet Berkay On, Xian Xiao, Roberto Proietti, AND S. J. Ben Yoo are represented below as YJL, MBO, XX, RP, and SJBY.

Appendix

A1. Optoelectronic neuron parameter selection process

The process of determining the parameters in the neuron circuit can be viewed in three steps. The first essential step is to decide what kind of transistors to use for FET1 and FET2. If we assume using the same type of transistor for FET1 and FET2, the transistor is required to support the current from the drain terminal to the source terminal higher than the threshold current of the laser to turn on the laser. Hence, our testbed neuron requires transistors that support at least 10 mA to excite the laser (due to the threshold current of the laser used for the testbed). The other crucial parameter for FET1 and FET2 transistors is the threshold voltage. The threshold voltages of these transistors need to be within the operating range of the circuit to perform charging and leaking the charges in the capacitors C1 and C2 through the R1C1 and R2C2 circuits. Note that the FET1 and FET2 transistors are need not be of identical types.

After finding suitable FET1 and FET2 transistors, the next step is to determine the membrane potential RC circuit's values (R1C1). The membrane potential RC circuit charges by a current input provided by the photodetector (PD1_exc) to make the circuit work. The amount of the current supplied by the photodetector is determined by the input light intensity and the photodetector's responsivity. By considering the amount of charge on each spike into the membrane potential RC circuit, the threshold voltage determined on the previous step and the RC value will determine the neuron's charging and leaking times. The charging and leaking times also specify the maximum operation speed on this neuron circuit.

The final step is to determine FET3 and the refractory potential RC circuit(R2C2). The capacitor value in the refractory potential RC circuit is the most crucial parameter in this step. It must keep the FET3 transistor in the ON state for a long enough time to drain the membrane potential. Thus, the capacitor(C2) and FET3 ON state voltage values need to be compatible with each other.

Optics EXPRESS

A2. Testbed neuron experiment

The experimental setup apparatus includes an FPGA generating arbitrary spiking patterns modulating the excitatory and the inhibitory lasers, *Laser1* and *Laser2* (commercially packaged pigtailed lasers), respectively. The outputs from *Laser1* and *Laser2* emulate the signals from upstream neurons. These outputs are directly coupled to *PD1_exc* (excitatory photodetector) and *PD2_inh* (inhibitory photodetector). The output of the neuron circuit utilizes a 1.3-micron wavelength VCSEL diode (prototype VCSEL from VERTILAS). The output spikes from the VCSEL are recorded by a lightwave converter with optical input ports. The experiment diagram is shown in Fig. 9.

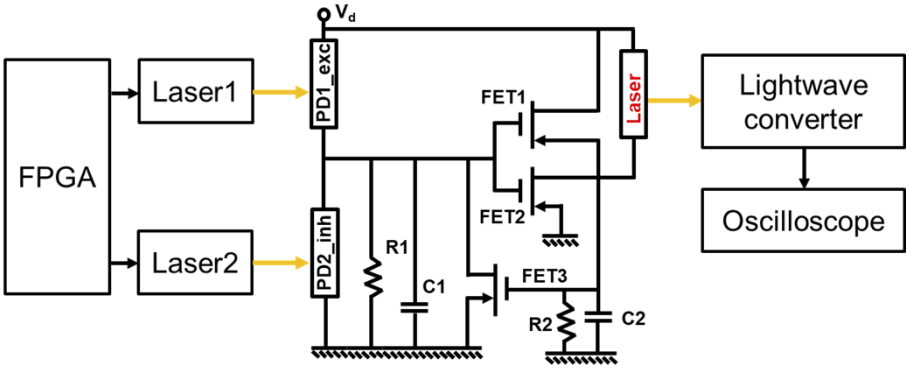

Fig. 9. An experimental setup for our proof-of-principle neuron utilizing a commercial laser, VCSEL, commercial photodetectors PD1_exc (excitatory photodetector) and PD2_inh (inhibitory photodetector), together with the electronic neuron circuit formed by discrete electronics on a printed circuit board. To test this neuron's optical input and output spiking performance, the apparatus includes an FPGA generating arbitrary spiking patterns into the excitatory and the inhibitory lasers, Laser1 and Laser2, respectively, directly coupling to PD1_exc and PD2_inh, respectively. An oscilloscope records the neuron's optical output of the laser.

Figure 10(a) is the LTSpice simulated input spiking pattern consisting of the four groups in sequence (14, 5, 3, and 1 spike in each group), and Fig. 10(b) is the measured optical spiking pattern output from *Laser1*. Figure 10(c) and (d) provide the simulated and measured membrane potential values measured at the membrane potential circuit of Fig. 1 by placing a monitor in the simulator and placing a probe in the actual experiment. As indicated by the arrows on Fig. 10(c) and (d), we observe that the simulated and the measured membrane potential values reach the threshold after three consecutive spike inputs. For the first spike group of 14 spikes, it reaches the threshold three times, and for both the second spike group of 5 spikes and the third spike group of 3 spikes, it reaches the threshold only once. For the fourth group of a single spike, it does not reach the threshold. Figure 10(e) and (f) illustrate simulated and experimental results, including the refractory potential and the optical output from the laser in addition to the optical excitatory input and the membrane potential. Here we observe that the optical output spikes fire when the membrane potential reaches the threshold, but more importantly, the refractory potential rises in response to the spike output. This indicates that the firing of the optical output spikes occurs only after the refractory period. This proves that our neuron model correctly represents the general dynamics of the Izhikevich model, including the refractory period. The experimental result in Fig. 10(f) shows the spike output behavior closely matching the LTSpice results in Fig. 10(e).

We repeated the simulation and the experiment when using both excitatory and inhibitory input signals by adding the inhibitory signal to *PD2_inh*. Figure 11(a) and (b) show the additional inhibitory input signal (red) as marked by arrows (red) for the simulation and the experiment

Optics EXPRESS

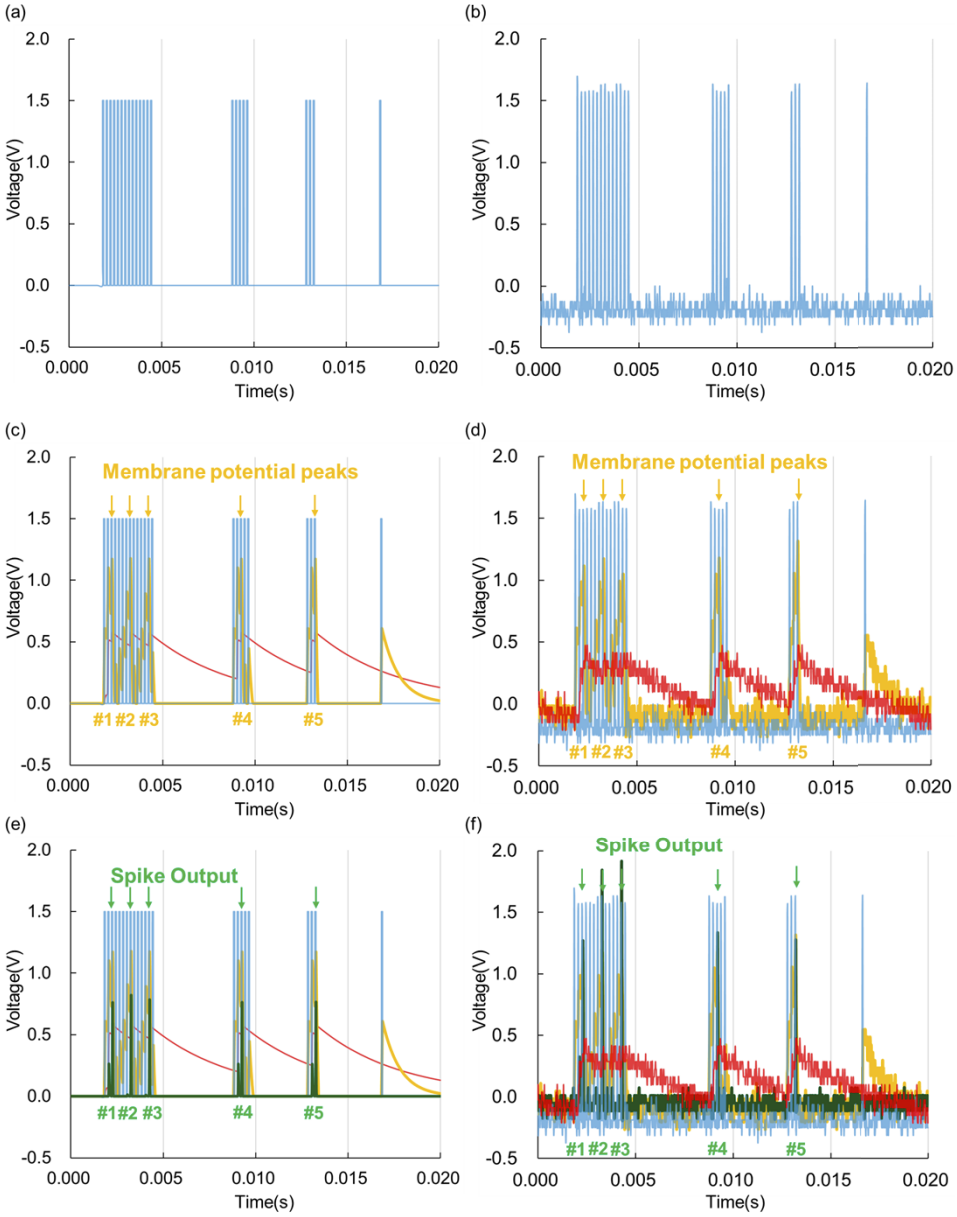

Fig. 10. Neuron spiking behavior with only excitatory signal input simulation and experiment. The input spiking pattern consisting of the four groups in sequence (14, 5, 3, and 1 spike in each group). (a) simulated inputs. (b) measured optical spiking pattern output from Laser1. (c) simulated inputs and membrane potential values. (d) measured inputs and membrane potential values. (e) simulated inputs, membrane potential values, refractory potential, and optical outputs. (f) measured simulated inputs, membrane potential values, refractory potential, and optical outputs. (Blue: optical excitatory input, Red: refractory potential, Yellow: membrane potential, Green: optical output)

Optics EXPRESS

results. As Fig. 11(c) and (d) demonstrate, the neuron behaves similarly to the behaviors shown in Fig. 10(c) and (d) when the inhibitory signal is absent (the membrane potential rises to the threshold for three consecutive spikes as labeled as #1, #2, and #4 in Fig. 11(c) and (d)). When the inhibitory signals are present, the membrane potential gets suppressed and cannot accumulate charge to generate spikes at #3 and #5. This behavior contrasts with the behavior seen at #3 and #5 of Fig. 10(c) and (d), where the membrane potential rose to the threshold in the absence of the inhibitory signals. Figure 11(e) and (f) show the optical output spikes absent at #3 and #5 due to the presence of the inhibitory signal.

From Fig. 10 and Fig. 11, we can find experimental results closely match the simulated results, and the optical output spikes are absent at #3 and #5 due to the presence of the inhibitory signal. The inhibitory signal (Red) cancels out the effect of the excitatory signal (Blue) to suppress the output spike (Green). This neuron behavior is consistent with the commonly seen functionality of biological inhibitory neurons.

A3. Foundry neuron simulation

We follow the same simulation process as testbed neuron with a two-level model consisting of the neuron circuit-level and the neural network-level simulation modules. The result is shown in Fig. 12. The timing of the input and output spike is the most important thing to focus on spiking neuron behavior. As a result, the Verilog-A model circuit behavior and the Nengo simulator can be viewed as a match.

A4. Foundry and nano neuron energy calculation

The foundry neuron is simulated with the design value used on Fig. 4(e) with a total capacitance of 68.1 fF (main capacitor (60 fF), photodetector load capacitance (2.1 fF), and transistors parasitic capacitance (6 fF)). The neuron behavior is set with three continuous spikes to charge to the threshold (0.65 V). Three spikes provide a total charge of $Q = C^*V = 44.27$ fC, which means one spike requires to contain 14.76 fC of charge. The foundry PD has a responsivity of 0.7A/W, leading to a neuron's dynamic input energy $E_{dynamic-in}$= 21.09 fJ/spike. If we assume the neuron is operating at its maximum spiking rate (10 GHz) with a spike width of 10 ps, the peak dynamic power $P_{dynamic-in} = \frac{E_{dynamic-in}}{T_{spike}}$ will be 2.11mW, which equals to 3.24 dBm. To support 10dB network loss due to silicon waveguides, we need an $E_{dynamic-out}$= 211fJ/spike. The output spike width is the same as the input spike width, which is 10 ps. That leads to a peak dynamic power of $P_{dynamic-in}$= 21.1mW, which equals to 13.24 dBm. As for the static power in our neurons, the only power consumed is caused by the leakage current when the neuron is OFF. The neuron's leakage current in the transistor circuit is 3.18μA at 2V power supply, and 580 pA at a 0.5V power supply, making the total P_{static}= 6.36 μW. When the neuron reaches its threshold (ON), it will turn the transistors into ON state. The current at 2 V power supply is 423 μA and 22.4 μA at a 0.5 V power supply, leading to $P_{dynamic-FET}$= 858μW.

Let us assume that the total neuron output spikes occupy t % of the time in a certain time slot. If we assume the optoelectronic neuron is spiking at the maximum rate of 10 GHz continuously, the neuron will turn on at most 3.33% of the time ($t = 3.33$). The average power is $P_{avg} = t * (P_{dynamic-in} + P_{dynamic-FET}) + (1-t) * P_{static}$= 714μW. According to Diehl and Cook's MNIST handwriting recognition experiment [54] and our simulation monitoring, the entire neural network only spikes around 8.6% of the time in the assigned time slot (sparse network), and average 27 spikes are required per sample. We exclude the energy consumption on the MZI training process and assume using non-volatile material to maintain the phase shifter state on MZI. As a result, the total energy consumption will only happen on neuron for this calculation. The neural network is based on 784-400-10 architecture, which states that the maximum MZI network size is 784×784. By applying tensor-train decomposition architecture, which is 5.2× fewer cascaded stages of MZIs, we require 151 MZI stages. Based on the parameter provided by the foundry,

Optics EXPRESS

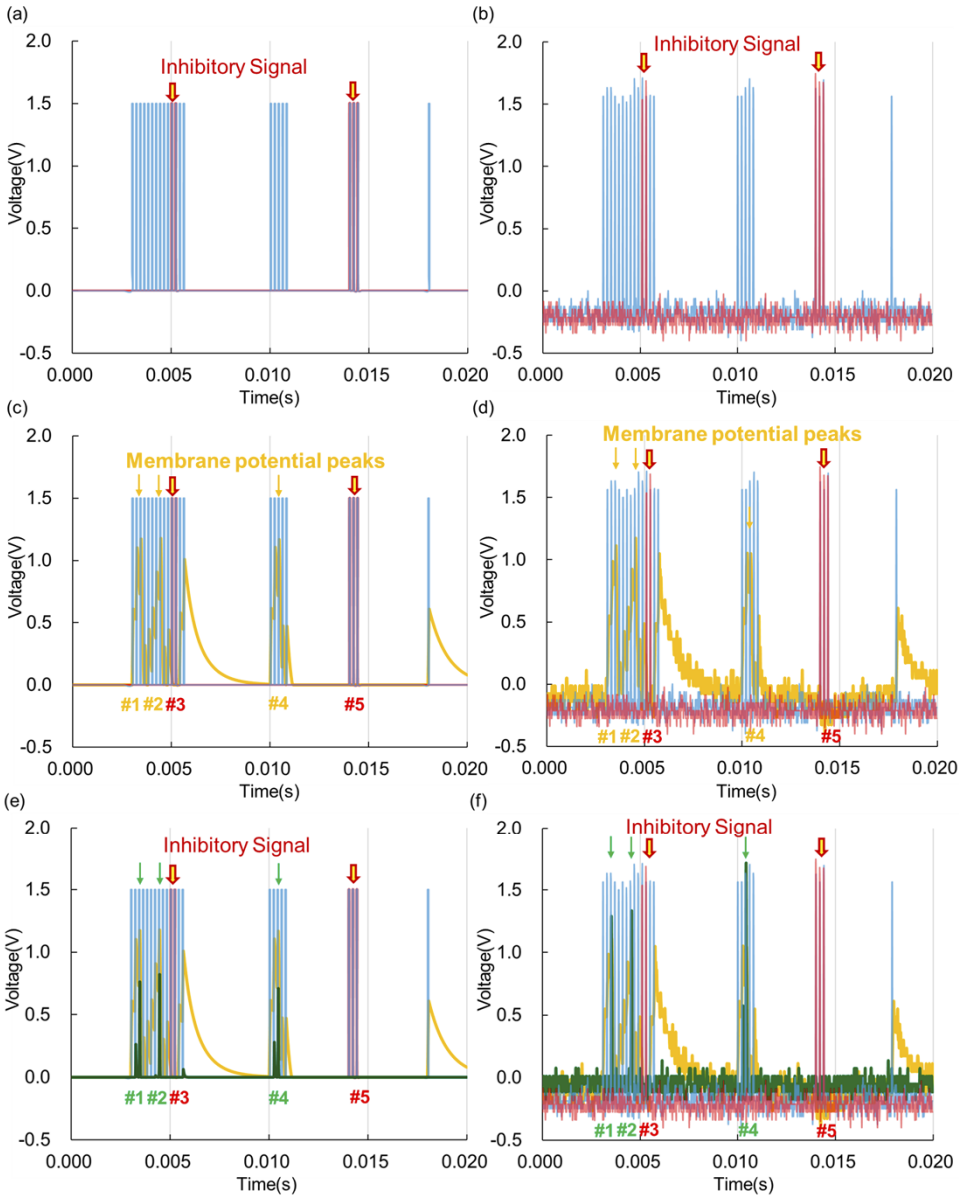

Fig. 11. Neuron spiking behavior with both excitatory and inhibitory signal inputs. The input spiking pattern for excitatory input is the same as Fig. 3, consisting of the four groups in sequence (14, 5, 3, and 1 spike in each group). An additional inhibitory signal is added on PD2_inh. (a) simulated excitatory and inhibitory inputs. Inhibitory inputs are label in red. (b) measured optical spiking pattern output from Laser1. (c) simulated excitatory and inhibitory inputs and membrane potential values. (d) measured excitatory and inhibitory inputs and membrane potential values. (e) simulated excitatory and inhibitory inputs, membrane potential values, and optical outputs (f) measured excitatory and inhibitory inputs, membrane potential values, and optical outputs. (Blue: optical excitatory input, Red: optical inhibitory input, Yellow: membrane potential, Green: optical output)

Optics EXPRESS

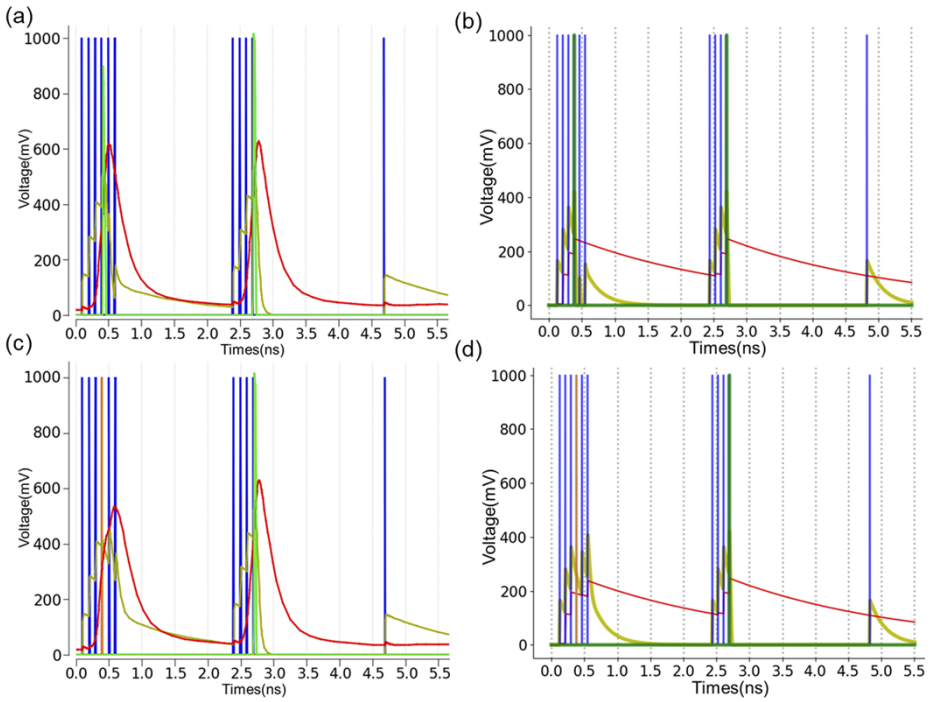

Fig. 12. The foundry-PSNN neuron simulation (a) is the foundry version neuron cadence simulation at a 10 GHz spiking rate. (Blue: optical excitatory input; Yellow: membrane potential; Red: refractory potential; and Green: laser modulation signals) (b) is the foundry version neuron spiking behavior used in Nengo simulation. Note that the spiking rate in the simulation is scaling down to meet Nengo simulator's time step and the difference of refractory potential won't affect the neuron behavior. (Blue: optical excitatory input; Yellow: membrane potential; Red: refractory potential; and Green: laser modulation signals) (c) is adding inhibitory input (on #4 spike start from left, labeled with color orange) on (a) cadence simulation. (d) is the corresponding Nengo simulation of (c)

which the waveguide loss is around 1.6 dB/cm and MZI length 400μm per stage. In that case, it leads to 10 dB loss per neural network layer. That means the foundry-PSNN neurons require 10× more dynamic optical output power to compensate for the network loss. As a result, the energy required per MNIST sample will be 5.2nJ for foundry-PSNN.

For the future nano-neuron-based PSNN, the IRDS2020 [43] report allows us to scale down the transistors($\leqq$3 nm) to the smaller capacitance (1.1 aF) and the lower threshold voltage (0.1 V). Based on the working mechanism explained in Section 1, we can assume the future neuron will have 0.601fF of total capacitance (main capacitor (0.5 fF), photodetector load capacitance (0.1 fF) [4], and transistors parasitic capacitance (1.1 aF)). Let us assume the neuron behavior is the same as our previous neuron version with three continuous input spikes charging the neuron to the threshold (0.1 V). Let us also assume that the photonic crystal PD has a responsivity 1 A/W. In that case, we estimate the future nano neuron's dynamic input energy per spike to be $E_{dynamic-in}$= 200 aJ/spike. If we follow the same operating condition as the foundry neuron, we estimate the following values: the peak dynamic power $P_{dynamic-in} = \frac{E_{dynamic-in}}{T_{spike}}$ = 20 μW, which equals to -17 dBm. The leakage current in the neuron will be 10 nA at 1.4 V power supply, making the total P_{static}= 14 nW. When the neuron is ON, the current at 1.4 V power supply is 31.27 μA, leading to $P_{dynamic-FET}$= 43.78 μW. The average power is P_{avg} = 8.14μW. For the power consumption of

Optics EXPRESS

Diehl and Cook's MNIST handwriting recognition experiment, we assume waveguide loss is 4 dB/m [55] and MZI length 350μm per stage. That leads to 0.2 dB loss per neural network layer, which means nano-PSNN neurons only require 1.05× more dynamic optical output power to excite the next layer's neurons successfully. As a result, the energy required per sample will be 4.4pJ.

A5. Additional benchmarking with nano-PSNN

This method of benchmarking is based on calculating the conventional multiply-accumulate (MAC) operation. In this comparison, we include both ANN and SNN to have a paramount view of energy efficiency between different approaches. One thing to notice here is that for SNN, the number of spikes per MAC would alter if we targeted a different task. Thus, the result shown here is based on applying PSNN on the MNIST dataset. We directly compare our PSNN with the energy efficiency results in [82]. The analog hardware is based on [5,71,74,83]. The digital hardware is based on [84–89]. The photonic hardware estimation is based on [82]. As shown in Fig. 13, we calculated the energy-efficiency-per-footprint as a product of the energy-efficiency for the workload (MAC/J) and the throughput per footprint (MAC/s/ mm^2). Our foundry-enabled version with O-E-O neural network can outperform most of all photonic approaches and achieve over 10^{26} $MAC^2/J/s/mm^2$ energy-efficiency-per-footprint. The future version can reach over 10^{29} $MAC^2/J/s/mm^2$ energy-efficiency-per-footprint with the above-mentioned implementation.

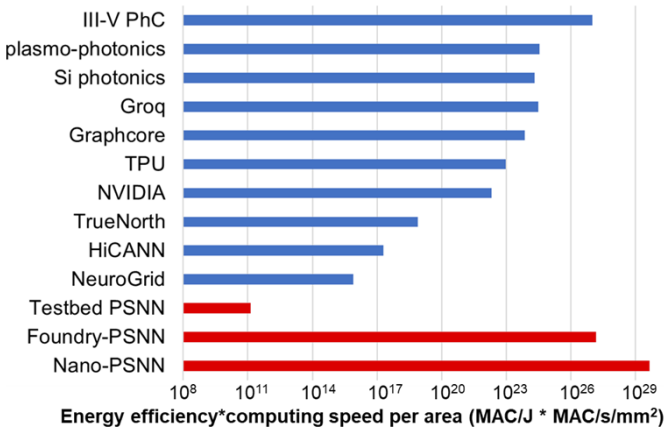

Fig. 13. Energy efficiency benchmarking based on conventional MAC operation.

Funding. Air Force Office of Scientific Research (FA9550-181-1-0186).

Acknowledgments. The authors would like to thank GLOBALFOUNDRIES for providing silicon fabrication through the 90WG university program.

Disclosures. All authors (YJL, MBO, XX, RP, and SJBY) are at the University of California, Davis, California, USA.

The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van denDriessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," Nature **529**(7587), 484–489 (2016).

Optics EXPRESS

2. D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of Go without human knowledge," Nature **550**(7676), 354–359 (2017).
3. B. Alves, "Facebook electricity use worldwide 2011-2019," https://www.statista.com/statistics/580087/energy-use-of-facebook/.
4. B. J. Shastri, A. N. Tait, T. Ferreira de Lima, M. A. Nahmias, H.-T. Peng, and P. R. Prucnal, "Neuromorphic Photonics, Principles of BT - Encyclopedia of Complexity and Systems Science," in R. A. Meyers, ed. (Springer, Berlin Heidelberg, 2018), pp. 1–37.
5. P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, "A million spiking-neuron integrated circuit with a scalable communication network and interface," Science **345**(6197), 668–673 (2014).
6. M. Davies, N. Srinivasa, T. Lin, G. Chinya, Y. Cao, S. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, Y. Liao, C. Lin, A. Lines, R. Liu, D. Mathaikutty, S. McCoy, A. Paul, J. Tse, G. Venkataramanan, Y. Weng, A. Wild, Y. Yang, and H. Wang, "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," IEEE Micro **38**(1), 82–99 (2018).
7. B. Romeira, J. Javaloyes, C. N. Ironside, J. M. L. Figueiredo, S. Balle, and O. Piro, "Excitability and optical pulse generation in semiconductor lasers driven by resonant tunneling diode photo-detectors," Opt. Express **21**(18), 20931–20940 (2013).
8. M. Nahmias, H. Peng, T. F. D. Lima, C. Huang, A. Tait, B. Shastri, and P. Prucnal, "A Laser Spiking Neuron in a Photonic Integrated Circuit.," arXiv Appl. Phys. (2020).
9. Y. Li, R. Chen, B. Sensale-Rodriguez, W. Gao, and C. Yu, "Real-time multi-task diffractive deep neural networks via hardware-software co-design," Sci. Rep. **11**(1), 11013 (2021).
10. K. Nozaki, S. Matsuo, T. Fujii, K. Takeda, A. Shinya, E. Kuramochi, and M. Notomi, "Femtofarad optoelectronic integration demonstrating energy-saving signal conversion and nonlinear functions," Nat. Photonics **13**(7), 454–459 (2019).
11. D. Rosenbluth, K. Kravtsov, M. P. Fok, and P. R. Prucnal, "A high performance photonic pulse processing device," Opt. Express **17**(25), 22767–22772 (2009).
12. M. A. Nahmias, B. J. Shastri, A. N. Tait, and P. R. Prucnal, "A Leaky Integrate-and-Fire Laser Neuron for Ultrafast Cognitive Computing," IEEE J. Sel. Top. Quantum Electron. **19**(5), 1–12 (2013).
13. F. Selmi, R. Braive, G. Beaudoin, I. Sagnes, R. Kuszelewicz, and S. Barbay, "Relative Refractory Period in an Excitable Semiconductor Laser," Phys. Rev. Lett. **112**(18), 183902 (2014).
14. B. J. Shastri, M. A. Nahmias, A. N. Tait, A. W. Rodriguez, B. Wu, and P. R. Prucnal, "Spike processing with a graphene excitable laser," Sci. Rep. **6**(1), 19126 (2016).
15. J. Robertson, T. Deng, J. Javaloyes, and A. Hurtado, "Controlled inhibition of spiking dynamics in VCSELs for neuromorphic photonics: theory and experiments," Opt. Lett. **42**(8), 1560–1563 (2017).
16. J. Robertson, E. Wade, and A. Hurtado, "Electrically Controlled Neuron-Like Spiking Regimes in Vertical-Cavity Surface-Emitting Lasers at Ultrafast Rates," IEEE J. Sel. Top. Quantum Electron. **25**(6), 1–7 (2019).
17. J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," Nature **569**(7755), 208–214 (2019).
18. K. Alexander, T. VanVaerenbergh, M. Fiers, P. Mechet, J. Dambre, and P. Bienstman, "Excitability in optically injected microdisk lasers with phase controlled excitatory and inhibitory response," Opt. Express **21**(22), 26182–26191 (2013).
19. M. A. Nahmias, A. N. Tait, L. Tolias, M. P. Chang, T. Ferreira de Lima, B. J. Shastri, and P. R. Prucnal, "An integrated analog O/E/O link for multi-channel laser neurons," Appl. Phys. Lett. **108**(15), 151106 (2016).
20. A. Jha, C. Huang, H. -T. Peng, B. J. Shastri, and P. Prucnal, "Photonic spiking neural networks and graphene-on-silicon spiking neurons," J. Lightwave Technol. **40**(9), 2901–2914 (2022).
21. J. M. Shainline, S. M. Buckley, R. P. Mirin, and S. W. Nam, "Superconducting Optoelectronic Circuits for Neuromorphic Computing," Phys. Rev. Appl. **7**(3), 034013 (2017).
22. M. Hejda, J. A. Alanis, I. Ortega-Piwonka, J. Lourenço, J. Figueiredo, J. Javaloyes, B. Romeira, and A. Hurtado, "Resonant Tunneling Diode Nano-Optoelectronic Excitable Nodes for Neuromorphic Spike-Based Information Processing," Phys. Rev. Appl. **17**(2), 024072 (2022).
23. A. N. Tait, T. deLima, M. A. Nahmias, H. B. Miller, H. -T. Peng, B. J. Shastri, and P. R. Prucnal, "Silicon Photonic Modulator Neuron," Phys. Rev. Appl. **11**(6), 064043 (2019).
24. I. A. D. Williamson, T. W. Hughes, M. Minkov, B. Bartlett, S. Pai, and S. Fan, "Reprogrammable Electro-Optic Nonlinear Activation Functions for Optical Neural Networks," IEEE J. Sel. Top. Quantum Electron. **26**(1), 1–12 (2020).
25. R. Amin, J. K. George, S. Sun, T. Ferreira de Lima, A. N. Tait, J. B. Khurgin, M. Miscuglio, B. J. Shastri, P. R. Prucnal, T. El-Ghazawi, and V. J. Sorger, "ITO-based electro-absorption modulator for photonic neural activation function," APL Mater. **7**(8), 081112 (2019).
26. B. Shi, N. Calabretta, and R. Stabile, "Deep Neural Network Through an InP SOA-Based Photonic Integrated Cross-Connect," IEEE J. Sel. Top. Quantum Electron. **26**(1), 1–11 (2020).

Optics EXPRESS

27. X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, R. Morandotti, A. Mitchell, and D. J. Moss, "11 TOPS photonic convolutional accelerator for optical neural networks," Nature **589**(7840), 44–51 (2021).
28. S. Bhattacharya, S. N. Patra, and S. Mukhopadhyay, "An all optical prototype neuron based on optical Kerr material," Optik (Munich, Ger.) **126**(1), 13–18 (2015).
29. X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, "All-optical machine learning using diffractive deep neural networks," Science **361**(6406), 1004–1008 (2018).
30. D. A. B. Miller, "Attojoule Optoelectronics for Low-Energy Information Processing and Communications," J. Lightwave Technol. **35**(3), 346–396 (2017).
31. D. A. B. Miller, "Optics for low-energy communication inside digital processors: quantum detectors, sources, and modulators as efficient impedance converters," Opt. Lett. **14**(2), 146–148 (1989).
32. S. J. BenYoo, "Wavelength conversion technologies for WDM network applications," J. Lightwave Technol. **14**(6), 955–966 (1996).
33. Intel Lab, "Taking Neuromorphic Computing to the Next Level with Loihi 2 Technology Brief," https://www.intel.com/content/www/us/en/research/neuromorphic-computing-loihi-2-technology-brief.html.
34. W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, Neuronal Dynamics (n. d.).
35. E. M. Izhikevich, "Simple Model of Spiking Neurons," IEEE Trans. Neural Networks **14**(6), 1569–1572 (2003).
36. E. M. Izhikevich, "Polychronization: computation with spikes," Neural Comput. **18**(2), 245–282 (2006).
37. P. R. Prucnal, B. J. Shastri, T. F. deLima, M. A. Nahmias, and A. N. Tait, "Recent progress in semiconductor excitable lasers for photonic spike processing," Adv. Opt. Photon. **8**(2), 228–299 (2016).
38. P. Jonas and G. Buzsaki, "Neural inhibition," Scholarpedia **2**(9), 3286 (2007).
39. T. Bekolay, J. Bergstra, E. Hunsberger, T. DeWolf, T. Stewart, D. Rasmussen, X. Choo, A. Voelker, and C. Eliasmith, "Nengo: a {Python} tool for building large-scale functional brain models," Front. Neuroinform. **7**, 1–13 (2014).
40. M. Liao, S. Chen, J. -S. Park, A. Seeds, and H. Liu, "III–V quantum-dot lasers monolithically grown on silicon," Semicond. Sci. Technol. **33**(12), 123002 (2018).
41. J. Zhang, G. Muliuk, J. Juvert, S. Kumari, J. Goyvaerts, B. Haq, C. Op deBeeck, B. Kuyken, G. Morthier, D. VanThourhout, R. Baets, G. Lepage, P. Verheyen, J. VanCampenhout, A. Gocalinska, J. O'Callaghan, E. Pelucchi, K. Thomas, B. Corbett, A. J. Trindade, and G. Roelkens, "III-V-on-Si photonic integrated circuits realized using micro-transfer-printing," APL Photonics **4**(11), 110803 (2019).
42. F. DeLeonardis, J. Hu, R. Soref, V. M. N. Passaro, and Y. Zhang, "Broadband Electro-Optical Crossbar Switches Using Low-Loss $Ge_2Sb_2Se_4Te_1$ Phase Change Material," J. Lightwave Technol. **37**(13), 3183–3191 (2019).
43. "THE INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS: 2020," IEEE (2020).
44. J. Song, S. Yuan, C. Cui, Y. Wang, Z. Li, A. X. Wang, C. Zeng, and J. Xia, "High-efficiency and high-speed germanium photodetector enabled by multiresonant photonic crystal," Nanophotonics **10**(3), 1081–1087 (2021).
45. D. A. B. Miller, "Self-configuring universal linear optical component [Invited]," Photonics Res. **1**(1), 1–15 (2013).
46. T. Grottke, W. Hartmann, C. Schuck, and W. H. P. Pernice, "Optoelectromechanical phase shifter with low insertion loss and a 13π tuning range," Opt. Express **29**(4), 5525–5537 (2021).
47. B. Ellis, M. A. Mayer, G. Shambat, T. Sarmiento, J. Harris, E. E. Haller, and J. Vučković, "Ultralow-threshold electrically pumped quantum-dot photonic-crystal nanocavity laser," Nat. Photonics **5**(5), 297–300 (2011).
48. G. Shambat, B. Ellis, J. Petykiewicz, M. A. Mayer, A. Majumdar, T. Sarmiento, J. S. Harris, E. E. Haller, and J. Vuckovic, "Electrically Driven Photonic Crystal Nanocavity Devices," IEEE J. Sel. Top. Quantum Electron. **18**(6), 1700–1710 (2012).
49. S. Cheung, Y. Kawakita, K. Shang, and S. J. BenYoo, "Highly efficient chip-scale III-V/silicon hybrid optical amplifiers," Opt. Express **23**(17), 22431–22443 (2015).
50. J. H. B. Wijekoon and P. Dudek, "Compact silicon neuron circuit with spiking and bursting behaviour," Neural Networks **21**(2-3), 524–534 (2008).
51. A. Osada, Y. Ota, R. Katsumi, K. Watanabe, S. Iwamoto, and Y. Arakawa, "Transfer-printed quantum-dot nanolasers on a silicon photonic circuit," Appl. Phys. Express **11**(7), 072002 (2018).
52. X. Xiao, M. B. On, T. VanVaerenbergh, D. Liang, R. G. Beausoleil, and S. J. BenYoo, "Large-scale and energy-efficient tensorized optical neural networks on III–V-on-silicon MOSCAP platform," APL Photonics **6**(12), 126107 (2021).
53. M. B. On, Y. -J. Lee, X. Xiao, R. Proietti, and S. J. BenYoo, "Analysis of the Hardware Imprecisions for Scalable and Compact Photonic Tensorized Neural Networks," *2021 Eur. Conf. Opt. Commun.* 1–4 (2021).
54. P. Diehl and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," Front. Comput. Neurosci. **9**, 99 (2015).
55. M. A. Tran, D. Huang, T. Komljenovic, J. Peters, A. Malik, and J. E. Bowers, "Ultra-Low-Loss Silicon Waveguides for Heterogeneously Integrated Silicon/III-V Photonics," Appl. Sci. **8**(7), 1139 (2018).
56. Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, "Deep learning with coherent nanophotonic circuits," Nat. Photonics **11**(7), 441–446 (2017).
57. S. Pai, B. Bartlett, O. Solgaard, and D. A. B. Miller, "Matrix Optimization on Universal Unitary Photonic Devices," Phys. Rev. Appl. **11**(6), 064044 (2019).
58. H. Sattari, T. Graziosi, M. Kiss, T. J. Seok, S. Han, M. C. Wu, and N. Quack, "Silicon Photonic MEMS Phase-Shifter," Opt. Express **27**(13), 18959–18969 (2019).

Optics EXPRESS

59. J. Faneca, L. Trimby, I. Zeimpekis, M. Delaney, D. W. Hewak, F. Y. Gardes, C. D. Wright, and A. Baldycheva, "On-chip sub-wavelength Bragg grating design based on novel low loss phase-change materials," Opt. Express **28**(11), 16394–16406 (2020).
60. M. Shalaginov, S. An, Y. Zhang, F. Yang, P. Su, V. Liberman, J. Chou, C. Roberts, M. Kang, C. Rios, Q. Du, C. Fowler, A. Agarwal, K. Richardson, C. Rivero-Baleine, H. Zhang, J. Hu, and T. Gu, "Reconfigurable Non-Volatile High-Performance Metalens," in *2020 Conference on Lasers and Electro-Optics (CLEO)* (2020), pp. 1–2.
61. Y. Zhang, J. B. Chou, M. Shalaginov, C. Rios, C. Roberts, P. Robinson, B. Bohlin, Q. Du, Q. Zhang, J. Li, M. Kang, C. Gonçalves, K. Richardson, T. Gu, V. Liberman, and J. Hu, "Reshaping light: reconfigurable photonics enabled by broadband low-loss optical phase change materials," in *Micro- and Nanotechnology Sensors, Systems, and Applications XI*, T. George and M. S. Islam, eds. (SPIE, 2019), Vol. 10982, pp. 98–105.
62. M. Miscuglio, J. Meng, O. Yesiliurt, Y. Zhang, L. J. Prokopeva, A. Mehrabian, J. Hu, A. VKildishev, and V. J. Sorger, "GSST-based photonic memory multilevel perceptron," in *Conference on Lasers and Electro-Optics* (Optical Society of America, 2020), p. JF3A. 2.
63. W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walsmley, "Optimal design for universal multiport interferometers," Optica **3**(12), 1460 (2016).
64. D. C. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Deep, Big, Simple Neural Nets for Handwritten Digit Recognition," Neural Comput. **22**(12), 3207–3220 (2010).
65. Y. Lecun, L. D. Jackel, L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. A. Muller, E. Sackinger, P. Simard, and V. Vapnik, "Learning algorithms for classification: A comparison on handwritten digit recognition," in *Neural Networks*, J. H. Oh, C. Kwon, and S. Cho, eds. (World Scientific, 1995), pp. 261–276.
66. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE **86**(11), 2278–2324 (1998).
67. Y. -J. Lee, M. B. On, X. Xiao, and S. J. BenYoo, "Demonstration of an Optoelectronic Excitatory & Inhibitory Neuron for Photonic Spiking Neural Networks," in *Conference on Lasers and Electro-Optics* (Optical Society of America, 2020), p. SM1E. 6.
68. C. Williams, N. Hong, M. Julian, S. Borg, and H. J. Kim, "Tunable mid-wave infrared Fabry-Perot bandpass filters using phase-change GeSbTe," Opt. Express **28**(7), 10583 (2020).
69. A. Shrestha, H. Fang, D. P. Rider, Z. Mei, and Q. Qiu, "In-Hardware Learning of Multilayer Spiking Neural Networks on a Neuromorphic Processor," in *2021 58th ACM/IEEE Design Automation Conference (DAC)* (2021), pp. 367–372.
70. S. K. Esser, R. Appuswamy, P. Merolla, J. V. Arthur, and D. S. Modha, "Backpropagation for Energy-Efficient Neuromorphic Computing," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds. (Curran Associates, Inc. , 2015), Vol. 28, pp. 1117–1125.
71. Kirchhoff-Institut für Physik, "HiCANN," https://www.kip.uni-heidelberg.de/%0Avision/previous-projects/facets/neuromorphic-hardware/waferscaleintegration-%0Asystem/hicann/.
72. J. C. Knight and T. Nowotny, "GPUs Outperform Current HPC and Neuromorphic Solutions in Terms of Speed and Energy When Simulating a Highly-Connected Cortical Model," Front. Neurosci. **12**, 941 (2018).
73. E. Stromatias, F. Galluppi, C. Patterson, and S. Furber, "Power analysis of large-scale, real-time neural networks on SpiNNaker," *2013 Int. Jt. Conf. Neural Networks* 1–8 (2013).
74. B. V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J. Bussat, R. Alvarez-Icaza, J. V. Arthur, P. A. Merolla, and K. Boahen, "Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations," Proc. IEEE **102**(5), 699–716 (2014).
75. Y. -J. Lee, M. B. On, X. Xiao, and S. J. Ben Yoo, "Energy-Efficient Photonic Spiking Neural Network on a monolithic silicon CMOS photonic platform," in *Optical Fiber Communication Conference (OFC) 2021*, P. Dong Kani, J. Xie, C. Casellas, R. Cole, and C. Li, eds., OSA Technical Digest (Optical Society of America, 2021), p. Tu5H. 5.
76. D. Kim, X. She, N. M. Rahman, V. C. K. Chekuri, and S. Mukhopadhyay, "Processing-In-Memory-Based On-Chip Learning With Spike-Time-Dependent Plasticity in 65-nm CMOS," IEEE J. Solid-State Circuits **3**, 278–281 (2020).
77. P. N. Whatmough, S. K. Lee, D. Brooks, and G. Wei, "DNN Engine: A 28-nm Timing-Error Tolerant Sparse Deep Neural Network Processor for IoT Applications," IEEE J. Solid-State Circuits **53**(9), 2722–2731 (2018).
78. F. N. Buhler, P. Brown, J. Li, T. Chen, Z. Zhang, and M. P. Flynn, "A 3. 43TOPS/W 48. 9pJ/pixel 50. 1nJ/classification 512 analog neuron sparse coding neural network with on-chip learning and classification in 40 nm CMOS," in *2017 Symposium on VLSI Circuits* (2017), pp. C30–C31.
79. J. Park, J. Lee, and D. Jeon, "7. 6 A 65 nm 236. 5nJ/Classification Neuromorphic Processor with 7. 5% Energy Overhead On-Chip Learning Using Direct Spike-Only Feedback," in *2019 IEEE International Solid- State Circuits Conference - (ISSCC)* (2019), pp. 140–142.
80. S. Yin, S. K. Venkataramanaiah, G. K. Chen, R. Krishnamurthy, Y. Cao, C. Chakrabarti, and J. Seo, "Algorithm and hardware design of discrete-time spiking neural networks based on back propagation with binary activations," in *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (2017), pp. 1–5.
81. P. -Y. Tan, P. -Y. Chuang, Y. -T. Lin, C. T. Wu, and J. -M. Lu, "A Power-Efficient Binary-Weight Spiking Neural Network Architecture for Real-Time Object Classification," (2020).
82. A. R. Totović, G. Dabos, N. Passalis, A. Tefas, and N. Pleros, "Femtojoule per MAC Neuromorphic Photonics: An Energy and Technology Roadmap," IEEE J. Sel. Top. Quantum Electron. **26**(5), 1–15 (2020).
83. J. Hsu, "IBM's new brain [News]," IEEE Spectr. **51**(10), 17–19 (2014).
84. E. Stromatias, "Scalability and robustness of artificial neural networks," The Univ. Manchester (2016).

Optics EXPRESS

85. "Groq," https://groq.com/.
86. P. Teich, "Tearing apart Google's TPU 3. 0 AI coprocessor," https://www.nextplatform.com/2018/05/10/%0Atearingapart-googles-tpu-3-0-ai-coprocessor/.
87. R. Smith, "Nvidia volta unveiled: Gv100 GPU and tesla v100 accelerator announced,".
88. S. Knowles, "Scalable silicon compute," in *Proc. Workshop Deep Learn. Supercomputer Scale* (2017).
89. P. Wijesinghe, A. Ankit, A. Sengupta, and K. Roy, "An All-Memristor Deep Spiking Neural Computing System: A Step Toward Realizing the Low-Power Stochastic Brain," IEEE Trans. Emerg. Top. Comput. Intell. **2**(5), 345–358 (2018).