Photonic In-Memory Computing Primitive for Spiking Neural Networks Using Phase-Change Materials

Indranil Chakraborty,[*] Gobinda Saha, and Kaushik Roy
School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana 47907, USA

(Received 3 August 2018; revised manuscript received 23 October 2018; published 30 January 2019)

Spiking neural networks (SNNs) offer an event-driven and more-biologically-realistic alternative to standard artificial neural networks based on analog information processing. This can potentially enable energy-efficient hardware implementations of neuromorphic systems that emulate the functional units of the brain; namely, neurons and synapses. Recent demonstrations of ultrafast photonic computing devices based on phase-change materials (PCMs) show promise for addressing limitations of electrically driven neuromorphic systems. However, scaling these stand-alone computing devices to a parallel in-memory computing primitive is a challenge. In this work, we use the optical properties of the PCM $Ge_2Sb_2Te_5$ to propose a photonic SNN computing primitive, comprising a nonvolatile synaptic array integrated seamlessly with previously explored "integrate-and-fire" neurons. The proposed design realizes an "in-memory" computing platform that leverages the inherent parallelism of wavelength-division multiplexing. We show that the proposed computing platform can be used to emulate a SNN inferencing engine for image-classification tasks. The proposed design not only bridges the gap between isolated computing devices and parallel large-scale implementation but also paves the way for ultrafast computing and localized on-chip learning.

DOI: 10.1103/PhysRevApplied.11.014063

I. INTRODUCTION

The phenomenal success in the field of deep learning using artificial neural networks (ANNs) based on analog information processing has had far-reaching consequences recently [1]. Machines driven by such networks have surpassed humans in various tasks ranging from pattern recognition to playing complex games such as Go [2] and chess [3]. However, the growing complexities of computational models involved in such multilayered neural networks have rendered the training and inferencing tasks extremely expensive in terms of memory and energy. The gulf between the energy efficiency of the brain and that of standard neural-network architectures has led researchers to explore a bioplausible alternative; namely, spiking neural networks (SNNs). The event-driven nature and sparse information encoding of SNNs make them more feasible for energy-efficient neuromorphic computing, thus paving the way toward unraveling the elusiveness of the brain. The fundamental operations performed by SNNs involve parallelized dot products through the synaptic network followed by subsequent integration and thresholding by the neurons. Neuromorphic systems attempting to leverage the sparse and event-driven nature of SNNs thus aim at efficient emulation of these functionalities.

The initial efforts [4–6] in hardware implementations of SNNs were based on standard von Neumann architecture [7] based on complementary-metal-oxide-seminconductor (CMOS) technology where the synaptic units of the neural networks are stored in the digital memory and are repeatedly fetched by the processor for computing operations. However, the overhead of frequent data transport between the memory and processor has led to a shift in the computing paradigm as "in-memory" computing platforms [8,9] attempt to emulate the "massively parallel" operations of the brain. Although the term "neuromorphic" was primarily coined [10] with CMOS technology in mind, this computing domain has branched out to nonvolatile-memory technologies such as oxide-based memristors [11], spintronics [12], and phase-change materials (PCMs) [13,14] in recent years. The natural ability of these resistive technologies to compute parallelized dot products using crossbar structures makes them promising candidates for neuromorphic systems. Despite the extensive efforts in nonvolatile-memory-based in-memory computing in the electrical domain, these technologies suffer from different drawbacks that manifest themselves in form of reduced energy efficiency, lower speeds and presence of sneak paths. Moreover, write latencies in memristors [15,16] are a major reason why memristive devices are not suitable for temporally scalable architectures. Thus, there is a need to explore a different memory technology that can enable computing as well as the possibility of shorter write times.

[*] ichakra@purdue.edu

Integrated photonics offers an alternative approach to standard microelectronic in-memory computing platforms and promises ultrafast neural computing and information processing. Recent advances in photonics-based neuromorphic computing have seen implementations of various kinds [17,18] of neural processing units on the photonic platform leveraging the inherent capability of matrix operations of integrated optical circuits. Spike-based processing systems have also been extensively explored with excitable lasers [19,20]. However, most of the photonic systems investigated in the context of neuromorphic computing are based on volatile information processing, which requires thermal tuners to maintain the modulation states, which might turn out to be energy expensive for large-scale systems. Nonvolatility offers the ability to write and erase information dynamically, which is desirable for large-scale implementations of neuromorphic systems. To that effect, recent demonstrations of subnanosecond writing speeds in PCM technology based on $Ge_2Sb_2Te_5$ (GST) through optical pulses has opened up a host of opportunities for in-memory computing in the photonic domain [21]. The ultrafast switching using light overcomes the long-standing obstacle of high "write" latencies [15] for PCMs in the electrical domain. The highly contrasting optical properties of GST in its crystalline and amorphous phases have led to implementations of all-photonic memories [22], switches [23], and reconfigurable nonvolatile computing platforms [24]. More recently, photonics-based GST devices have also been explored to emulate biologically plausible synapses [25], capable of undergoing spike-timing-dependent plasticity, and "integrate-and-fire" spiking neurons [26]. Despite these promising investigations into fast neural computing based on a nonvolatile platform, the challenge of scaling stand-alone devices to large-scale neuromorphic systems is enormous. Thus, there is a need to explore nonvolatile-memory primitives in the photonic domain, which can perform parallel computing. In this work, we propose an all-photonic SNN computing primitive, based on GST-based photonic neural elements, that attempts to bridge the gap between devices and system-level implementation of photonic neural networks. We leverage the inherent wavelength-division-multiplexing (WDM) [27] property of optical networks to propose a nonvolatile synaptic array, while exploring and mitigating the challenges arising from designs based on ring resonators of radii comparable to the wavelength of operation. Such a synaptic array can achieve higher densities than current state-of-the-art photonic computing systems. We show how the proposed synaptic computing platform can be seamlessly integrated with previously explored integrate-and-fire spiking neurons to realize an ultrafast and truly integrable spiking neural network. Finally, we evaluate the performance of the proposed photonic SNN in the task of classifying handwritten digits.

II. PHOTONIC SYNAPSES

The core computational units of any neural network are neurons and synapses. In SNNs, information is encoded in the form of spikes, and the neurons and synapses are capable of processing information through these spike trains. As shown in Fig. 1(a), the input trains of spikes are multiplied by the synaptic weights $w_1, w_2, \ldots, w_n$ and the weighted sum is received by an integrate-and-fire neuron. The internal state of the neuron, known as the "membrane potential" (V_{mem}) performs integration of the basis of the incoming weighted spikes and is compared with a threshold (V_{th}) at every time step. The neuron outputs a spike once V_{mem} reaches V_{th}. The synaptic functionality essentially corresponds to a multiplication operation of the inputs and the corresponding weights of the synapses. The basic operation performed by a single synapse can be represented as $I_i w_i$. We show how a single-bus microring resonator with a GST element embedded on top of it can operate as such a synapse. The device under consideration is a Si-on-insulator structure consisting of a rectangular waveguide and a ring waveguide as shown in Fig. 1(b). A GST element is deposited on one arm of the ring waveguide, which takes the shape of an arc, where the length of the arc is denoted as the length of the GST element (L_{GST}). The fabrication technique of building such

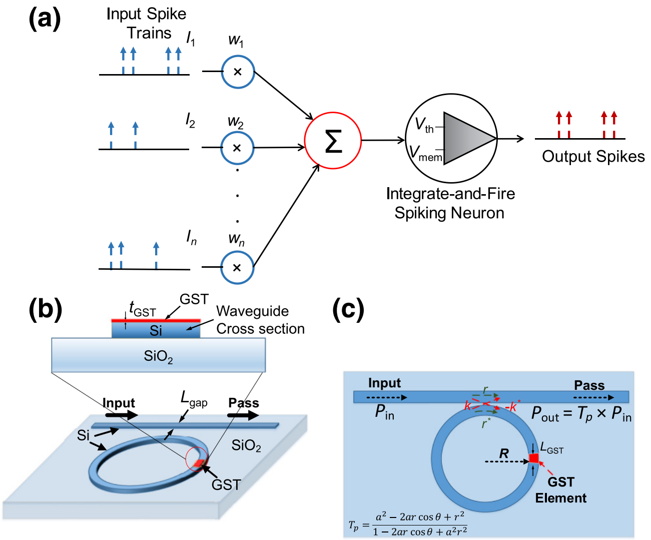

FIG. 1. (a) The basic functional elements of a SNN are spiking neurons and weighted synaptic connections. At each time instant, the inputs are weighted by the synaptic weights to produce a resultant output represented as $\sum_i P_i w_i$. The integrate-and-fire neuron's membrane potential (V_{mem}) is updated according to the weighted sum and compared with a threshold value (V_{th}). (b) GST-embedded single-bus-microring-resonator structure with Si waveguides on SiO_2 substrate. (c) Top view of the device illustrating the different parameters pertaining to the ring-resonator structure. The synaptic device performs an analog multiplication of input P_{in} and transmission T.

a structure has been well explored [23,24]. Waves in the rectangular waveguide are partially coupled to the ring and constructively interfere when the round-trip phase shift equals an integer multiple of 2π, leading to the resonant condition:

$$2\pi R_{\text{ring}} n_{\text{eff,wg}} = m\lambda_m, \tag{1}$$

where R_{ring} is the radius of the ring waveguide, $n_{\text{eff,wg}}$ is the effective refractive index of the ring waveguide, and λ_m is the resonant wavelength. The transmission through the "Pass" port is dependent on the device dimensions and material such that

$$T_p = \frac{a^2 - 2ar\cos\theta + r^2}{1 - 2ar\cos\theta + a^2r^2}, \tag{2}$$

where a is the attenuation factor, r is the self-coupling coefficient as shown in Fig. 1(c), and θ is the single-pass phase shift. Under resonance, θ equals 2π and the transmission is given by $T_{\text{min}} = [(a-r)/(1-ar)]^2$.

We leverage the contrasting optical properties of GST in its amorphous (a-GST) and crystalline (c-GST) states to manipulate the attenuation in the ring waveguide and thus vary the transmission T_{min} at the resonance wavelength. The differing imaginary refractive indices of a-GST and c-GST lead to differential absorption of evanescently coupled light. The difference in optical absorption can be visibly observed through the cross-section view of the fundamental-mode profiles in the GST-embedded Si waveguide when excited by a TE-mode electromagnetic wave as shown in Fig. 2. c-GST introduces a significant change in waveguide mode in contrast to a-GST due to higher absorption in the GST element. The attenuation factor (a) in Eq. (2) can be related to the imaginary refractive index as

$$a = \exp\left(-\frac{2\pi\kappa_{\text{eff,GST}} L_{\text{GST}}}{\lambda} + \alpha\right), \tag{3}$$

where $\kappa_{\text{eff,GST}}$ is the effective imaginary refractive index of GST on the Si-SiO$_2$ stack, L_{GST} is the length of the GST element, and the term "loss" refers to other propagation losses, such as bending losses. The GST element can be programmed to partially crystallized levels such that multilevel states can be achieved [22,24]. From the perspective of neural networks, significant progress have been made toward proposing training algorithms [28,29] that preserve performance even with binarized synapses. Thus, although multilevel states would be desirable from a device point of view, modified training techniques can enable reasonable performance with low-precision synapses.

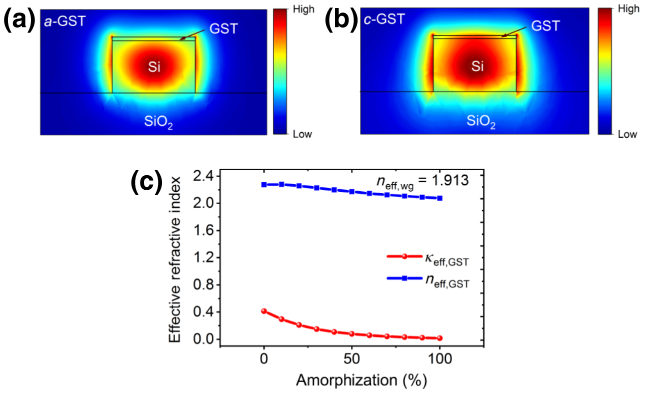

FIG. 2. Cross-section view of fundamental-mode profiles for a GST-embedded Si-SiO$_2$ waveguide section for (a) a-GST and (b) c-GST showing visible contrast in optical absorption for the two boundary states of GST. (c) The variation of the real ($n_{\text{eff,GST}}$) and imaginary ($\kappa_{\text{eff,GST}}$) refractive indices of GST with the degree of crystallization.

The refractive indices of partially crystallized GST can be calculated from effective permittivities approximated by an effective-medium theory [30,31]:

$$\frac{\epsilon_{\text{eff}}(p) - 1}{\epsilon_{\text{eff}}(p) + 2} = p \times \frac{\epsilon_c - 1}{\epsilon_c + 2} + (1-p) \times \frac{\epsilon_a - 1}{\epsilon_a + 2}, \tag{4}$$

where ϵ_c and ϵ_a are the complex permittivites of c-GST and a-GST, respectively, calculated from the refractive indices of GST [32] by $\sqrt{\epsilon(\lambda)} = n + i\kappa$, and p is the degree of crystallization. Thus, the different levels of crystallization of GST lead to various values of $\kappa_{\text{eff,GST}}$, thus leading to different levels of transmission. We leverage the multilevel transmission to implement an all-photonic synapse. For an incident optical pulse of power P_{in}, the synaptic functionality is realized such that the output power P_{out} is given by

$$P_{\text{out}} = T_{\lambda_m} P_{\text{in}}, \tag{5}$$

where T_{λ_m} is the transmission at resonant wavelength λ_m. T_{λ_m} represents the weight of the synapse, and the various levels of transmission with differing-degree-of-crystallization states of GST can be leveraged to represent an entire range of synaptic weights with appropriate discretization. We critically couple the resonator to the amorphous state such that the transmission is minimum in the amorphous state and increases with the degree of crystallization. While individual synapses represent a simple multiplication, the weighted inputs from multiple synapses are received by a neuron as shown in Fig. 1(a). To emulate such behavior, it is important to connect these synapses in an integrated fashion. Such a synaptic network would perform the most-ubiquitous functionality of any neural network, a dot product.

III. PHOTONIC DOT-PRODUCT ENGINE

We leverage the characteristics of the proposed nonvolatile photonic synaptic device to map the synaptic weights of a neural network in a photonic synaptic network capable of performing the dot product of the inputs and the weights.

A. Network design

We leverage the WDM technique to compute dot-product operations between incoming spikes and synaptic weights. We represent the synaptic weights in terms of the transmission T_λ of the microring resonator as discussed in the previous section. To represent multiple wavelengths, we use multiple ring resonators of increasing ring radius to represent different synapses in a row as shown in Fig. 3. The number of synapses (N) in each row is dependent on the free spectral range (FSR) of the ring resonator and this governs the dimension of the input vector of the dot-product engine. A WDM spike enters the straight waveguide through the input port, and the GST element on each ring resonator modulates the amplitude of the corresponding wavelength by the representative synaptic weight according to Eq. (5). Thus, at the output port we obtain a multiwavelength spike comprising different $T_{\lambda_i}P_i$ products corresponding to different wavelengths. This spike is then fed to a photodiode (PD) array, which produces a current given by the sum of all the amplitudes given by

$$I_{\text{out}} = R\sum_i T_{\lambda_i}P_i, \tag{6}$$

where R is the responsivity of the PD expressed as amperes per watt. This current is equal to the dot product of the input vector P and weight vector T_λ. The operation is illustrated in Fig. 3.

B. Synapse design constraints

Use of the WDM technique for the proposed photonic synaptic array imposes certain constraints on the design of the synaptic devices. For accurate dot-product operation, it is necessary to achieve significant isolation between the channels to minimize channel-to-channel interaction. The important parameters that constrain the design space of the synaptic device are finesse (F) and channel spacing (λ_{diff}). Finesse is the ratio of the FSR and the full width at half maximum (FWHM). For a single-bus ring resonator, the FWHM and FSR are expressed as [33]:

$$\Delta\lambda_{\text{FWHM}} = \frac{(1-ra)\lambda_m^2}{\pi n_g L\sqrt{ra}}, \tag{7}$$

$$\Delta\lambda_{\text{FSR}} = \frac{\lambda_m^2}{n_g L}, \tag{8}$$

$$F = \frac{\Delta\lambda_{\text{FSR}}}{\Delta\lambda_{\text{FWHM}}}, \tag{9}$$

where $L = 2\pi R_{\text{ring}}$ is the circumference of the ring, n_g is the group index, and the rest of the parameters have the same meaning as defined earlier. The interference due to adjacent channels can be modeled as

$$\begin{aligned} T'_{\lambda_i}|_{\lambda=\lambda_i} &= T_{\lambda_i}|_{\lambda=\lambda_i} \times T_{\lambda_i}|_{\lambda=\lambda_{i+1}} \times T_{\lambda_i}|_{\lambda=\lambda_{i-1}}, \\ T'_{\lambda_i}|_{\lambda=\lambda_i} &= \alpha_{\lambda_i}T_{\lambda_i}|_{\lambda=\lambda_i}, \end{aligned} \tag{10}$$

where $T'_{\lambda_i}|_{\lambda=\lambda_i}$ is the modified transmission due to interference from the adjacent resonant wavelengths, $T_{\lambda_i}|_{\lambda=\lambda_i}$, $T_{\lambda_i}|_{\lambda=\lambda_{i+1}}$, and $T_{\lambda_i}|_{\lambda=\lambda_{i-1}}$ are the transmissions of the ith ring at the ith, $(i+1)$th, and $(i-1)$th resonant wavelengths respectively, and α_{λ_i} represents the nonideal factor, which should ideally be close to 1; α_{λ_i} decreases with decreasing channel spacing (λ_{diff}) and increasing FWHM. For our design, we set the minimum radius of the ring to be 1.5 μm to achieve a high-density synaptic array for better scalability. Rings of similar size have been demonstrated previously [34] with certain modifications that we discuss next. The rest of the parameters concerning the synapses are chosen to maximize the number of rings in a single row (N) while maintaining α_{λ_i} close to 1 under the condition that $N\lambda_{\text{diff}} < \Delta\lambda_{\text{FSR}}$.

A number of challenges arise for rings of radius comparable to the wavelength of operation. Firstly, to achieve a critical coupling in the low-loss amorphous state, the power coupling gap between the bus and the ring waveguide needs to be small (less than 100 nm). This is because the interaction length between the ring and the straight waveguide is quite short and hence to achieve reasonable coupling, even to match the small intrinsic loss in

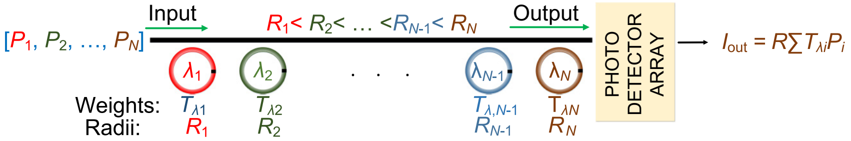

FIG. 3. Synaptic dot-product engine showing the arrangement of ring resonators with increasing radii representing the transmission vector $T_\lambda = \{T_{\lambda_1}, \dots, T_{\lambda_N}\}$. WDM signals are modulated by weights corresponding to the respective wavelength and the photodetector array collects the signals to generate a current I_{out} representing the dot product of the transmission vector T_λ and inputs $P = \{P_1, \dots, P_N\}$.

the ring in low-loss *a*-GST, we require a small power coupling gap. Such gaps are extremely difficult to fabricate. An alternative to using smaller gaps has been demonstrated [34] for rings of small radii. Reducing the width of the bus waveguide increases the spatial period of the propagating mode due to the lower effective refractive index. This results in a better phase match with the mode in the tightly curved ring waveguide. For the rest of our analysis, we use a bus waveguide of width 0.35 μm and a coupling gap of 135 nm.

IV. PHOTONIC INTEGRATE-AND-FIRE NEURONS

The proposed photonic dot-product engine (PDPE) needs to be interfaced with spiking neurons to realize a photonic SNN inferencing platform. In this work, we explore a photonic integrate-and-fire neuron that we proposed previously [26]. The neuron consists of an "integration unit" and a "firing unit". The "integration unit" of the neuron consists of two add-drop ring resonators with GST deposited on top of each as shown in Fig. 4(a). The purpose of the two ring resonators is to perform bipolar integration (i.e., the respective devices are fed by positive and negative weighted sums from the synapses to perform integration in the appropriate direction). The significance of positive and negative weighted sums will be clearer in the next section. The neuron operates in alternate "write" and "read" cycles. The GST elements on the ring resonators are initially in the crystalline state. With incident "write" pulses, the GST element begins to be partially amorphized. During the "read" phase, with partial amorphization, transmission at the "through" port of each ring resonator decreases and that at the "drop" port increases. Essentially, with incoming pulses, the transmission through the "drop" and "through" ports is positively and negatively integrated, respectively. These properties of the device can be combined to mimic the behavior of a bipolar integrate-and-fire neuron. The "drop" port and "through" port of the positive and negative integrating ring resonator, respectively, are connected to an inteferometer. The output of the interferometer represents the membrane potential of the spiking neuron. To perform the thresholding action, the membrane potential is fed to the "firing unit" of the neuron. This unit consists of an amplifier, a circulator, and a rectangular waveguide with GST deposited on top. During the "read" phase of the neuron, the resulting membrane potential, after being amplified and directed by the circulator toward the rectangular waveguide, attempts to amorphize the initially crystalline GST element on the rectangular waveguide. Initially, the output of the amplifier (P_{amp}) is insufficient to amorphize the GST on the rectangular waveguide and hence render it unable to transmit an output spike. However, when the membrane potential integrates enough to cross the threshold, on incidence of several "write" pulses, P_{amp} is ensured to be high enough to amorphize the GST on the rectangular waveguide, thus enabling it to transmit a spike. Once the neuron fires, a

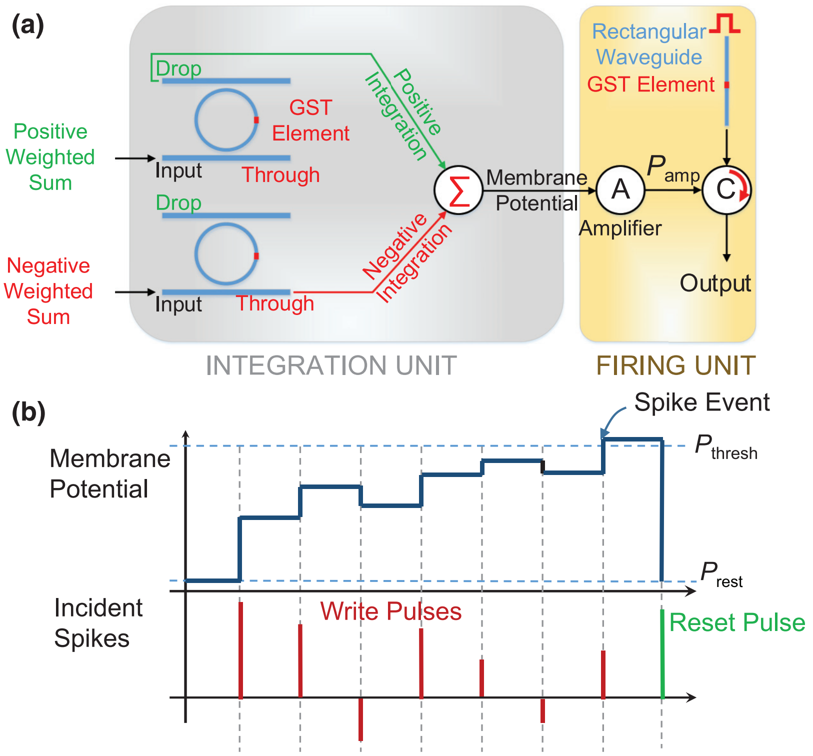

FIG. 4. (a) A bipolar integrate-and-fire neuron based on GST-embedded ring-resonator devices showing the integration and firing units. (b) Timing diagram showing the integration of membrane potential for various incident pulses demonstrating the operation of the proposed neuron.

"reset" pulse resets the states of the devices to their initial states and the membrane potential drops to the resting potential (P_{rest}) as shown in Fig. 4(b). Further details of the writing and reading schemes were presented in Ref. [26].

V. OPERATION OF AN ALL-PHOTONIC SPIKING NEURAL NETWORK

Implementation of a SNN based on the PDPE and integrate-and-fire neurons described above involves integration of the proposed structures. As elucidated above, the basic computational function of a neural network is a dot product. To realize parallel instances of such a functionality with use of the aforementioned PDPE, we use a splitter to feed the WDM input spikes to multiple PDPE rows with the input vector and obtain the dot products of each row from respective PD arrays as shown in Fig. 5. Essentially, the output vector thus obtained from the PD arrays gives us the multiplication of the vector of input spikes P_i with an $N \times M$ synaptic network T_{ij}. The M outputs I_j obtained from the PD arrays are fed to laser diodes, which convert the electrical current to optical spikes, thus completing the parallel dot-product operations, which can be represented as:

$$\begin{pmatrix} O_1 \\ O_2 \\ \vdots \\ O_M \end{pmatrix} \propto \begin{pmatrix} P_1 & P_2 & \dots & P_N \end{pmatrix} \begin{pmatrix} T_{11} & T_{12} & \dots & T_{1M} \\ T_{21} & T_{22} & \dots & T_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ T_{N1} & T_{N2} & \dots & T_{NM} \end{pmatrix}. \tag{11}$$

We now present how such a photonic synaptic network can be integrated with the proposed bipolar integrate-and-fire neurons to realize a photonic SNN. A schematic of such a photonic SNN is illustrated in Fig. 6. To account for negative weights in a neural network, we represent the element of the weight matrix T as comprising a positive and a negative component:

$$\begin{aligned} T_{ij} &= T_{ij}^{+} + T_{ij}^{-}, \\ T_{ij}^{+} &= T_{ij}, T_{ij}^{-} = T_{\mathrm{low}} \quad \text{when } T_{ij} > 0, \\ T_{ij}^{+} &= T_{\mathrm{low}}, T_{ij}^{-} = |T_{ij}| \quad \text{when } T_{ij} < 0, \end{aligned} \tag{12}$$

where T_{low} is the transmission corresponding to the lowest programmable state considered. Two PDPE arrays are used to map the positive and negative components as depicted in Fig. 6. The dot-product outputs from the laser-diode arrays of the two dot-product engine arrays can be represented as:

$$\begin{aligned} O_j^{+} &= \sum_i P_i T_{ij}^{+}, \\ O_j^{-} &= \sum_i P_i T_{ij}^{-}. \end{aligned} \tag{13}$$

These outputs from the jth rows are received by the jth integrate-and-fire neuron discussed earlier. The outputs from the positive and negative PDPE arrays are received by the positive and negative integrating ring resonators in the neuron, respectively. The two ring resonators integrate the membrane potential in the opposite direction on the basis of the two inputs, and the resulting integration mimics the desired integration that a biological integrate-and-fire neuron performs, given by

$$V_{\mathrm{mem},j}(t) = V_{\mathrm{mem},j}(t-1) + \sum_i P_i T_{ij}, \tag{14}$$

where $\sum_i P_i T_{ij} = \sum_i (P_i T_{ij}^{+} - P_i T_{ij}^{-})$ and $V_{\mathrm{mem},j}(t)$ is the internal state or the membrane potential of the jth neuron at time t. The resulting membrane potential is passed to a firing unit as described in Fig. 4 such that the neuron produces an output spike once $V_{\mathrm{mem},j}(t)$ reaches a threshold. The output spikes from all the neurons of the

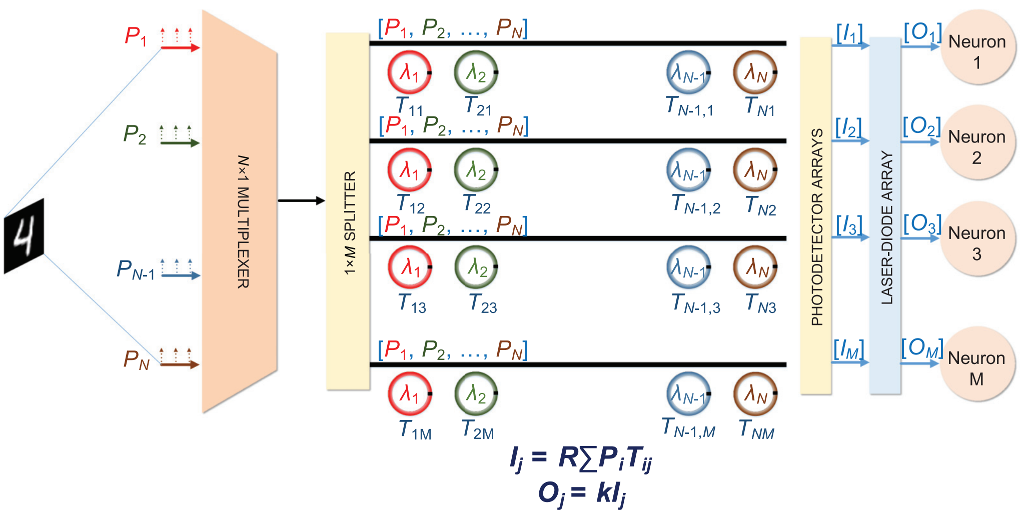

FIG. 5. Synaptic dot-product engine showing the arrangement of ring resonators with increasing radii representing the transmission vector $T_\lambda = \{T_{\lambda_1}, \dots, T_{\lambda_N}\}$. WDM signals are modulated by weights corresponding to the respective wavelength and the photodetector array collects the signals to generate a current I_{out} representing the dot product of the transmission vector T_λ and inputs $P = \{P_1, \dots, P_N\}$. k is an amplification factor.

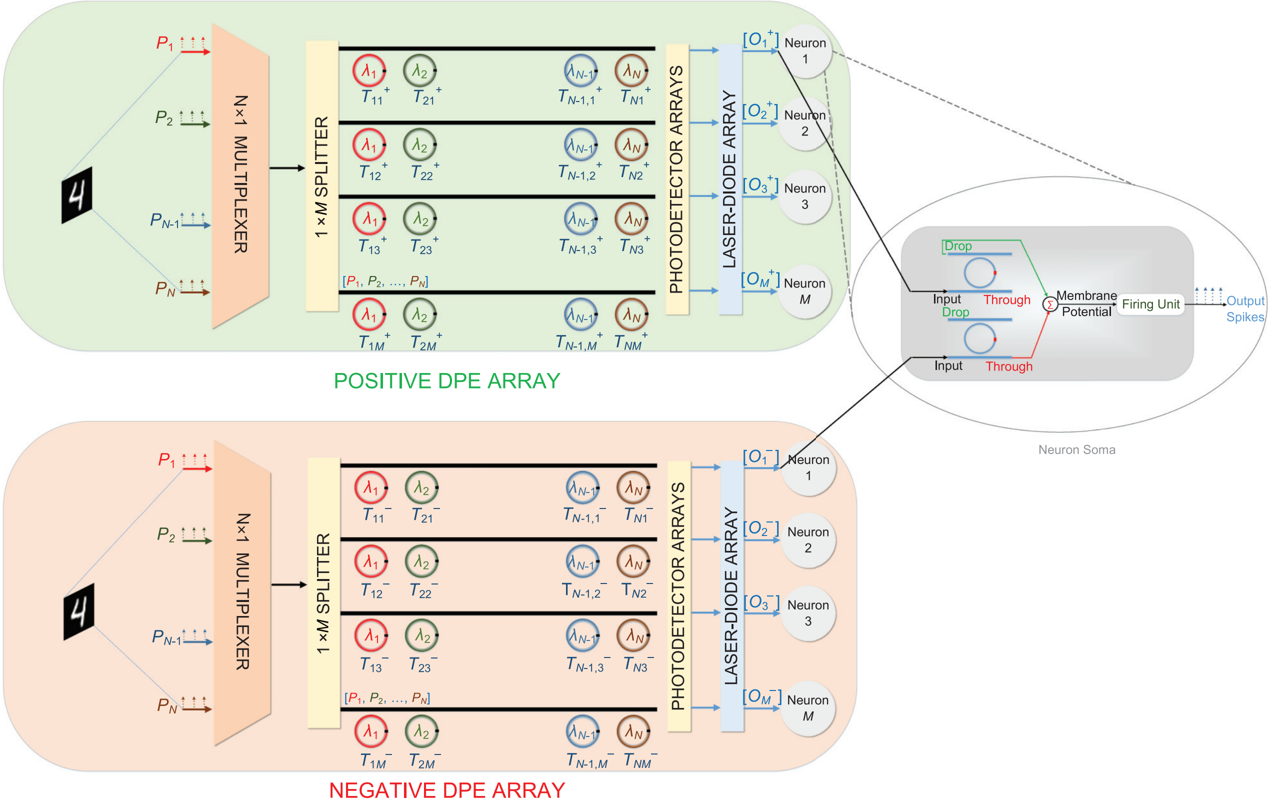
P_1
P_2
P_{N-1}
P_N
N×1 MULTIPLEXER
1 ×M SPLITTER
λ_1 λ_2 λ_{N-1} λ_N
T_{11}^+ T_{21}^+ $T_{N-1,1}^+$ T_{N1}^+
T_{12}^+ T_{22}^+ $T_{N-1,2}^+$ T_{N2}^+
T_{13}^+ T_{23}^+ $T_{N-1,3}^+$ T_{N3}^+
$[P_1, P_2, \ldots, P_N]$
T_{1M}^+ T_{2M}^+ $T_{N-1,M}^+$ T_{NM}^+
PHOTODETECTOR ARRAYS
LASER-DIODE ARRAY
$[O_1^+]$ Neuron 1
$[O_2^+]$ Neuron 2
$[O_3^+]$ Neuron 3
$[O_M^+]$ Neuron M
POSITIVE DPE ARRAY
Drop
Input
Through
Drop
Input
Through
Σ
Membrane Potential
Firing Unit
Output Spikes
Neuron Soma
T_{11}^- T_{21}^- $T_{N-1,1}^-$ T_{N1}^-
T_{12}^- T_{22}^- $T_{N-1,2}^-$ T_{N2}^-
T_{13}^- T_{23}^- $T_{N-1,3}^-$ T_{N3}^-
T_{1M}^- T_{2M}^- $T_{N-1,M}^-$ T_{NM}^-
$[O_1^-]$ Neuron 1
$[O_2^-]$ Neuron 2
$[O_3^-]$ Neuron 3
$[O_M^-]$ Neuron M
NEGATIVE DPE ARRAY

FIG. 6. An all-photonic spiking neural network. Two dot-product-engine (DPE) arrays are used to represent the positive and negative components of the weights. The outputs of the DPE arrays are converted to optical spikes and passed to integrate-and-fire neurons. The structure of an integrate-and-fire neuron is illustrated in the oval. Each neuron has two inputs, corresponding to outputs from the positive and negative DPE arrays. The neuron outputs a spike when the membrane potential crosses its threshold.

current layer are then fed to the next synaptic array layer. Figure 6 delineates the operation of basic building blocks of a neural network. We perform large-scale system-level simulations by emulating the behavorial model of the proposed spike processing system to assess the performance of neuromorphic systems based on this fabric.

It is important to consider the architecture-level facets of any computing primitive. The proposed design is analogous to memristive crossbars, where the high fan-in into the neurons is resolved by the inherent parallelism of the computing framework. In our design, each neuron receives two inputs, from the positive and negative synaptic arrays, and the output of that neuron is fed to one of the 16 inputs of the synaptic array of the next layer. In reality, neural networks are of far bigger sizes than the proposed design can accommodate. As a result, multiple instances of the proposed primitive can be used with time multiplexing to perform the entire vector-matrix multiplication operation. The partial sums from these instances are collected and added before being fed to the neuron. Output from a neuron again serves as inputs to the synaptic arrays storing the weights of the next layer of the neural network. Similar architectures have been explored with use of memristive technologies [16,35]. This work is concerned with device and circuit primitives of a spike-based photonic nonvolatile inferencing engine that will act as a computing core of a large-scale system similar to technologies in the electrical domain.

VI. RESULTS

A. Simulation framework

1. Device simulations

We evaluate the performance of the proposed all-photonic SNN fabric by designing a device-circuit-algorithm cosimulation framework. First, the device characteristics of each ring resonator in a dot-product-engine row are simulated for four different degrees of crystallization of the GST element with use of the simulator FDTD SOLUTIONS from Lumerical [36] based on the finite-difference time-domain (FDTD) method. The fixed parameters used for these simulations are listed in Table I. The mode profiles are obtained through electromagnetic simulations with the finite-element method in COMSOL MULTIPHYSICS [40].

2. Device-to-system framework

The device characteristics, obtained from the FDTD simulations, are analyzed and a Gaussian fit is applied

TABLE I. Simulation parameters.

Parameter	Value
Si-ring-waveguide cross section	$0.45 \times 0.25\ \mu\text{m}^2$
Si-bus-waveguide cross section	$0.35 \times 0.25\ \mu\text{m}^2$
Coupling gap (L_{gap})	$0.135\ \mu$m
GST length (L_{GST})	170–220 nm
GST thickness (t_{GST})	10 nm
GST width (W_{GST})	$0.44\ \mu$m
Si refractive index (n_{Si}) [37]	3.5
SiO_2 refractive index (n_{SiO_2}) [38]	1.4
c-GST refractive index ($n_{c\text{-GST}} + i\kappa_{c\text{-GST}}$) [39]	$7.2 + 1.9i$
a-GST refractive index ($n_{a\text{-GST}} + i\kappa_{a\text{-GST}}$) [39]	$4.6 + 0.18i$

on the data for interpolation. We develop a device-to-system codesign framework by building behavorial models of the proposed synapses and neurons based on the fitted device characteristics. The models are used to evaluate the inferencing performance of the standard neural-network topology on a standard digit-recognition task based on the MNIST dataset with use of the DEEP LEARNING TOOLBOX [41] in MATLAB. The MNIST dataset consists of 60 000 images in the training set and 10 000 images in the testing set.

B. Device simulations

We consider 16 ring resonators of radius linearly increasing from 1.5 to 1.59 μm in any particular dot-product-engine row. The choice of the number of devices, N, in a single row was discussed earlier. The length of the GST element is increased accordingly and chosen iteratively to ensure uniform transmission characteristics across the wavelength range of operation. We perform FDTD simulations for each device with four different degrees of crystallization of GST (30%, 50%, 80%, 100%); the observed transmission characteristics for the rings are shown in Fig. 7(a). As expected, the transmission for each device decreases with decreasing degree of crystallization. The observed FSR is 53.1 nm, and the difference between the highest and lowest resonant wavelengths is 47 nm, which is well within the FSR, thus ensuring no interference from resonant wavelengths beyond the region of operation. Figures 7(b) and 7(c) show the contrast in electric field absorption by the GST element in the ring resonator for 30% and 100% crystallized GST. We observe certain variations across different wavelengths, which can be minimized by further adjustment of the length of the GST element. However, from the perspective of neuromorphic applications, these variations prove to be insignificant. We explore the impact of such variations in our evaluation of the proposed neuromorphic processing engine. We use the dependence of transmission on the degree of crystallization to realize the synaptic behavior of the rings. Figure 8(a) shows the Gaussian fit of the simulated data across degrees of crystallization ranging from 0% to 100%. The Gaussian fit provides a fairly accurate representation of the observed data and is a powerful tool to speed up our analysis in light of the computationally expensive FDTD simulations. It can be observed that transmission has a nonlinear relationship with p, and hence operation of the rings as synapses

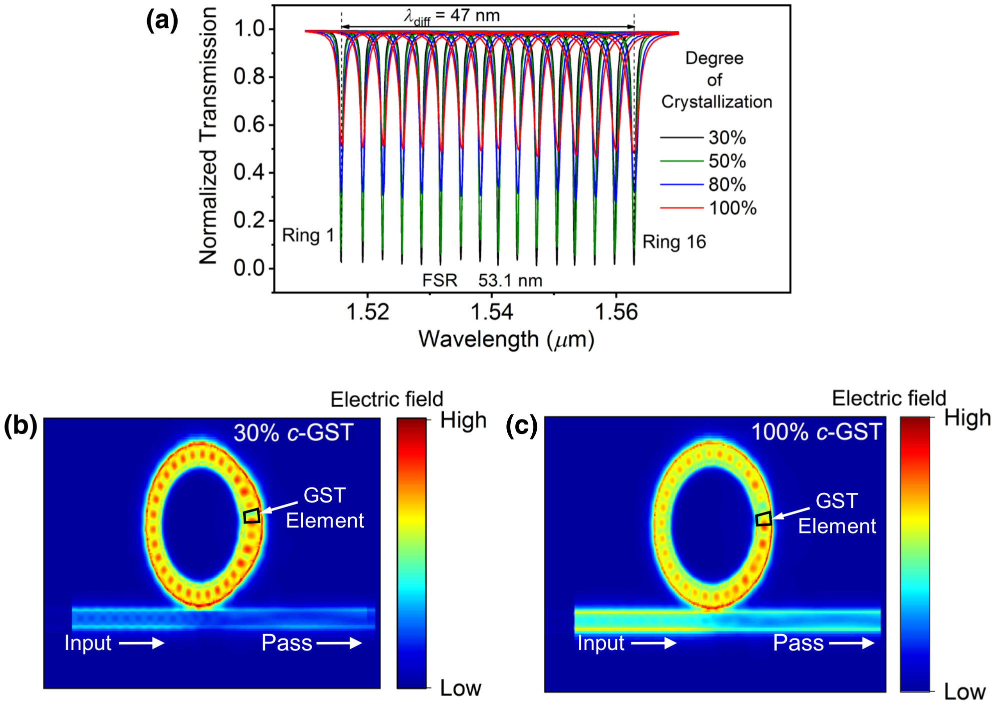

FIG. 7. (a) Normalized transmission for 16 different rings for four degrees of crystallization (30%, 50%, 80%, 100%) showing a decreasing trend with decreasing degree of crystallization. The wavelength range for the 16 rings is less than the FSR for the design. (b),(c) Electric field profile in the ring-resonator system showing visible contrast in optical absorption and field transmission at the “pass” port in the GST element for *c*-GST and 30% *c*-GST, respectively.

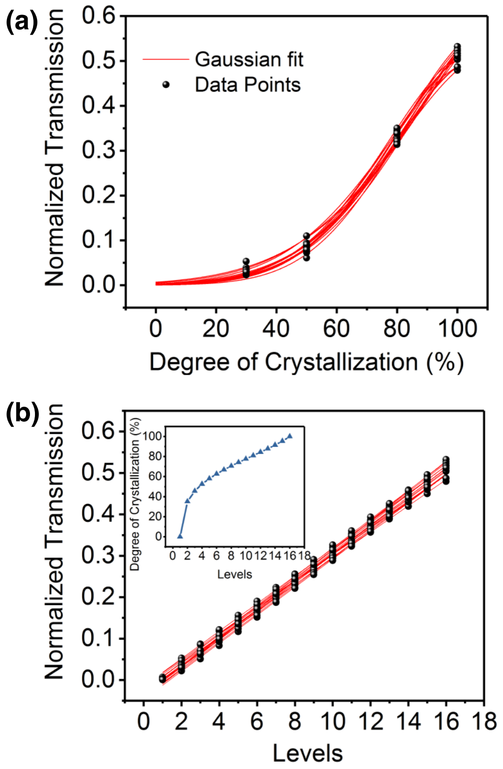

FIG. 8. (a) Gaussian fit of simulated data points for degree of crystallization ranging from 0% and 100%. (b) Linearly varying transmission across 16 different programmable states (levels) of the GST. The inset shows the degrees of crystallization corresponding to the levels.

would require the GST element to be programmed to states with nonlinearly increasing p. This can be achieved with an appropriate amplitude of the programming stimulus. Figure 8(b) shows the transmission levels for each ring corresponding to 16 discretized programmable states or levels. The degree of crystallization, p, for each state is shown in the inset in Fig. 8(b). The linear relationship between transmission and levels is a necessity for the target application (i.e., a dot-product operation for neuromorphic computing), which leads us to the choice of programmable states with the nonlinear distribution of p.

C. Interference errors

The transmission characteristics of the different rings for different states of the GST element are used to evaluate the accuracy of the dot-product operation performed with the proposed synaptic network. The error in the computation stems from the premise of overlapping frequency response between adjacent channels. The advantage of the proposed implementation over electrical counterparts is that in the electrical domain the losses due to line resistance are a function of the input and the weights, thus rendering them difficult to model. The impact of the error in this setup is dependent only on the weight level and hence can be easily

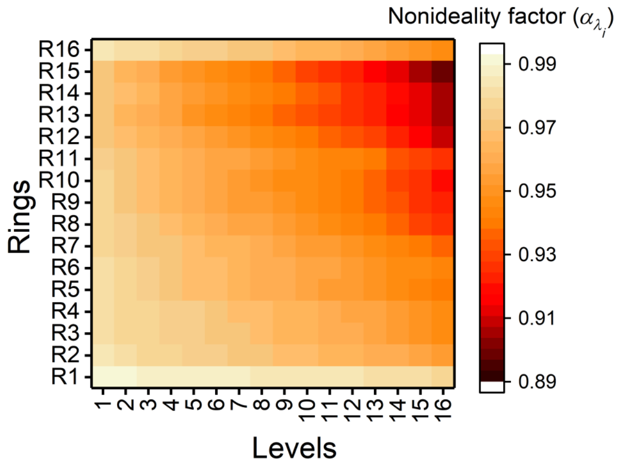

FIG. 9. Map of nonideality factor (α_{λ_i}) arising due to interference from adjacent rings for each ring in the dot-product-engine row.

modeled, analyzed, and even corrected in light of the proposed application. In Eq. (9), we formulate a behavioral model of the error arising from interference due to adjacent channels. Figure 9 shows the map of the nonideality factor α_{λ_i} for all 16 rings for 16 different levels. This is calculated by our fitting of the extracted α_{λ_i} from Fig. 7(a) on the basis of Eq. (9). We observe that the errors are highest for rings of greater radius and for the highest levels. This can be attributed to greater FWHM for rings of greater radius due to the longer lengths of the GST element used to achieve uniform transmission levels across the operating wavelength range. We include these error characteristics corresponding to each ring for our system-level evaluation of the proposed photonic SNN inferencing framework.

D. System-level SNN performance

We develop a device-to-algorithm-level framework to perform system-level analysis of the photonic SNN implementation. A SNN, like any other neural network, consists of multiple layers of neurons connected through synapses. The unique property of SNNs is that the inputs to the network are discretized spike events instead of analog values. The synapses act as weights that are multiplied by the amplitude of the incoming stimulus and the resulting weighted sum (i.e., dot product of all impulses coming from different synapses) is received by the neuron. We map the device characteristics of each individual synapse and integrate-and-fire spiking neurons discussed previously to explore the validity of operation of the proposed devices as synapses and neurons in such a SNN. We now explain how we perform the evaluation of a SNN on the proposed PCM-based photonic inferencing framework. We consider a fully connected neural network consisting of three layers—the input layer, the hidden layer, and the output layer—as shown in Fig.

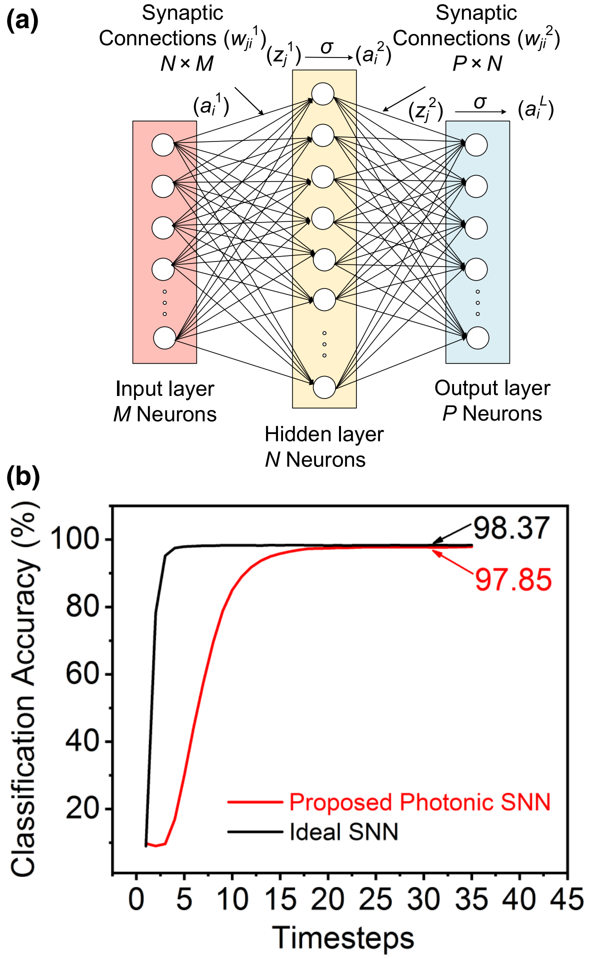

FIG. 10. (a) Fully-connected-neural-network topology consisting of an input layer (M), a hidden layer (N), and an output layer (P) of neurons. The resulting synaptic networks are of size $N \times M$ and $P \times N$. (b) Evolution of classification accuracy of handwritten-digit-recognition task based on the MNIST dataset comparing the performance of our proposed photonic SNN with ideal-SNN performance. Here "ideal SNN" corresponds to software-level functionalities without consideration of device characteristics.

10(a). This type of topology has been well explored [42]. For our analysis, we consider a network with $M = 784$, $N = 500$, and $P = 10$. We analyze the accuracy of such a network in a standard handwritten-digit-recognition task based on the MNIST dataset [43]. A popular way of implementing spike-based inferencing systems is to train a network as an ANN and then convert it to a SNN by well-explored conversion algorithms [42,44]. The weights of the network are trained with the backpropagation algorithm [45] as in the case of ANNs. The neurons in ANNs are usually nonlinear mathematical functions, such as rectified linear units (ReLU) [46], sigmoid functions or hyperbolic tangent functions, with ReLUs being the most-popularly-chosen neuron functionality. During conversion, an artificial neuron with ReLU functionality can be directly converted to an integrate-and-fire neuron mathematically [42]. The details of the operation of the integrate-and-fire neuron were elucidated in our earlier work [26]. The trained weights of the network after the ANN are converted to a SNN and mapped to the observed characteristics of each synaptic device in the proposed synaptic network. The synaptic network has the provision of operating 16 synapses simultaneously. To perform the dot product of larger dimensions, the synaptic network needs to be time multiplexed, as discussed earlier. To simulate large-dimension operations with the proposed synaptic network, we repeat the device characteristics every 16 synapses. The weights of the network can be negative. To account for negative weights, two dot-product engines are used, shown in Fig. 6, as described earlier.

The pixels of input images of size 28 × 28 are divided into streams of spikes whose frequency is proportional to the pixel intensity. At every time step, the input can either be "0" when there is no spike or "1" in the event of a spike. The behavorial model of the SNN inferencing framework described above is implemented with the MATLAB DEEP LEARNING TOOLBOX [41] with the network topology shown in Fig. 10(a). The network is evaluated at every time step by our passing the inputs through the forward path from the input layer to the output layer through the synaptic network and recording the activity of the network. Finally, the output neuron with the highest spiking activity is compared with the label of the input image to determine the accuracy of the recognition system. The classification performance of the proposed photonic SNN is compared with that of an ideal SNN in Fig. 10(b). Here "ideal SNN" essentially means software-level evaluation without device characteristics being taken into consideration. There is a degradation in accuracy of 0.52% after 35 time steps from the ideal case arising from the different variations in device characteristics discussed earlier. The concept of time steps here corresponds to how many times we evaluate the network over the Poisson-distributed input spikes generated from the image. The duration of a time step is not relevant in this context as we do not include any temporal dynamics in the system. We further attempt to isolate the contribution of synaptic device variations to the observed degradation in accuracy by considering a comparison test case: ideal synapses with proposed neurons. The accuracy degradation amounted to 0.1% after 35 time steps. This implies 0.42% degradation due to synaptic variations.

We evaluate the energy consumption of the basic building blocks for our system, the synaptic array and the neurons. The energy consumed by each synapse can be estimated by the transmission (or the weight) of the synaptic device. As the information being processed is based on spike events, the input can either be "1" or "0." Experimental demonstrations [22] have shown that read-out for GST-based Si photonic devices can be achieved by pulse energies of 0.48 pJ. For our case, because of smaller GST footprints, we consider input “1” to correspond to a pulse of amplitude 0.25 mW. The power consumed by the synapse is thus given by $(1 - T)$ mW, where T is the transmission of the synapse. As these read pulses will eventually write into the neurons, we choose a pulse width of 200 ps, which is the minimum pulse width required to write into the GST, as we observed previously [26]. Considering these metrics for the read pulses and power calculations for each synapse, we estimate the energy consumption of the entire classification operation described above. The resulting average energy consumption for first layer of the neural network in the synaptic array is calculated to be approximately 12.5 fJ per synapse per time step of evaluation. For the second layer, the energy consumption is approximately 1.6 fJ per synapse per time step. The difference is energy consumption in the two layers is due to sparser spiking activity in the second layer. The energy consumed by each neuron was calculated in our previous work to be 5 pJ per time step. The writing energies for PCM devices of similar feature sizes [47,48] in the electrical domain can amount to 14–19 pJ while operating at speeds of 40–100 ns. The total energy consumption for an image classification is calculated to be approximately 261 nJ (178 nJ consumed by the synaptic operations and 83 nJ consumed by the neurons). Although the energy consumption is comparable to that of CMOS technology [49], photonics potentially offers a faster operation at subnanosecond speeds. In this work, we have consider a significantly-high-amplitude read pulse (0.25 mW) through the synapses, which is reflected in the high energy per inference operation. The proposed synapses can be potentially read with a pulse of lower amplitude on the basis of the sensitivity of the photodetectors and that will significantly reduce the energy requirements of the system. Moreover, the speed of operation in the photonic domain is significantly greater since read latencies of neuromorphic systems based on memristors are usually on the order of nanoseconds. These benefits encourage us to further explore the possibility of neuromorphic hardware design based on this technology.

VII. DISCUSSION

The proposed photonic SNN inferencing framework fills a major void of scaling from device to systems in current state-of-the-art photonic neuromorphic studies based on PCMs. However, a few challenges that stand in the way of physical demonstration of the proposal need to be overcome. Firstly, reconfigurability of the proposed nonvolatile synaptic array is a necessity. Various reconfigurability schemes have been explored on phase-change-based photonic platforms [24,32]. We explore the possibility of adding an input bend waveguide (WG_{write}) as a writing port for each synapse at a distance such that the inferencing framework is unaffected. The width of WG_{write} (W_{write}) is intentionally considered to be much lower than that of the ring waveguide of the synaptic device. This is done to achieve asymmetric coupling such that during writing the wave leaks out of WG_{write} appropriately for efficient writing, while during the standard inferencing operation, the wave remains mostly confined within the ring. Figure 11(a) shows the structure and arrangement of WG_{write} adjacent to the proposed synaptic device. t_{gap} denotes the distance between the ring waveguide and WG_{write}. We observe that error in transmission during normal inferencing operation due to the presence of WG_{write} is around 0.5% for $t_{\text{gap}} \sim 300$ nm. For the same distance, we calculate the transient field coupling from WG_{write} to the ring to be 70%. Thus, this writing scheme is a viable option for achieving reconfigurability in the proposed network.

The dimensions chosen for our analysis are aimed at our achieving desirable functionality for ring resonators of small radius of approximately 1.5 μm. The main motivation behind use of small ring resonators is to achieve high area density for scalability. We explore a number of

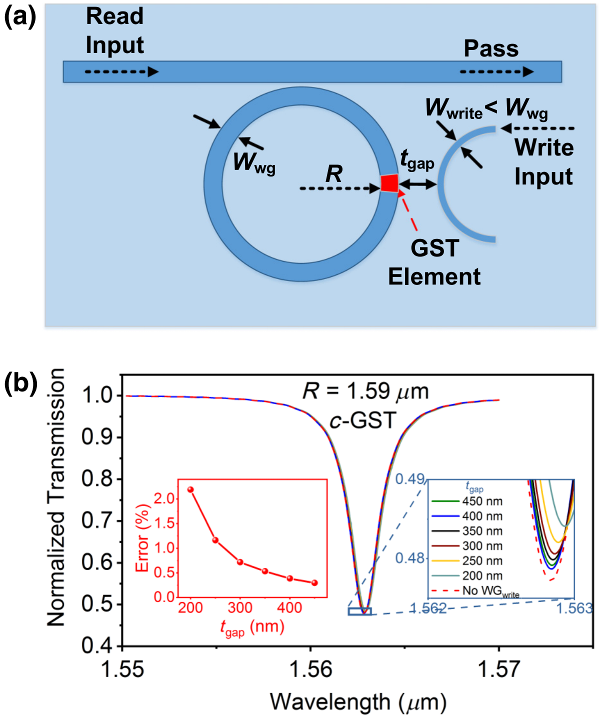

FIG. 11. (a) Structure and arrangement of input write waveguide at a distance t_{gap} to the synaptic device. The width of the write waveguide (W_{write}) is smaller than that of the ring waveguide (W_{wg}) for asymmetric coupling. (b) Transmission characteristics of a 1.59-μm ring for different values of t_{gap} compared with the case without a write waveguide. The inset on the right shows an enlarged view of the transmission characteristics to show the different cases clearly. The inset on the left shows the variation of the percentage error in transmission at a read wavelength of 1562.85 nm with t_{gap}.

challenges arising from such small rings such as nonuniform bending and coupling losses across the range of wavelength and fabrication difficulties to achieve critical coupling. We attempt to mitigate such challenges by appropriate design. Further, we delineate the design constraints for scaling individual synapses to a network of synapses, which is necessary for large-scale neuromorphic systems. GST-based photonic platforms also experience a small resonance shift between the different programmable states of the PCM. The resonance shift between the any two states can be quantified by [23]

$$\frac{\Delta\lambda_m}{\lambda_{m,\text{in}}} = \frac{\Delta n_{\text{eff,GST}}}{n_{g,\text{eff}}}\frac{L_{\text{GST}}}{2\pi R_{\text{ring}}}, \tag{15}$$

where $\lambda_{m,\text{in}}$ is the resonant wavelength in the initial state, $\Delta n_{\text{eff,GST}}$ is the difference in effective refractive index between the states, and $n_{g,\text{eff}}$ is the group index. For our case, it amounts to approximately 0.012 nm. In addition to the variations arising from device characteristics, we also explore errors arising due to interference from adjacent channels and their impact on the performance of the proposed photonic SNN. From our analysis, it can be observed that the network size, N, considered in our synaptic fabric is rather conservative. N can be further increased, which would result in higher errors. However, the effect of such variations is modeled in Eq. (9), and the resulting accuracy degradation can be recovered by modification of the training algorithm as explored for memristive technologies [50].

The errors arising due to interference between adjacent rings essentially stem from the use of WDM-based computation; therefore, the limitations of array size due to WDM merits discussion. WDM, while introducing parallelism in the system, is constrained by the finesse of the rings. In this work, we show that we can use 16 rings in a single dot-product-engine row, which implies that the array can process 16 inputs in parallel. The size of the array is thus limited to $16 \times N$, where N is limited by the area and not design constraints. However, analogous computing units in the electrical domain using memristive crossbars are also limited in size due to electromigration limits, sneak paths, and line resistances. The photonic array on the other hand, although limited in one direction due to finesse, can be possibly extended to larger sizes in the direction of N. Moreover, time multiplexing is a popular practice when one is implementing large-scale neural networks on memristive networks, as alluded to earlier. The possibility of fast writing into PCMs can potentially make these photonic arrays more suitable for temporally scalable architectures.

An alternative way to implement photonic neural networks is through the use of inteferometers [18], where the weights of the network are controlled through phase shifters. Such phase shifters can consume a significant amount of power per synapse to maintain the weight. On the other hand, nonvolatile elements based on PCMs can potentially encode the weights without requiring any power to maintain their states. However, we do not use the concept of phase shift for our design. We encode the weights in terms of levels of partial crystallization. Nonvolatility is necessary for large-scale neuromorphic systems for primarily two reasons: (i) it eliminates the need for phase shifters as constant tuning is not required and (ii) it provides a platform for in-memory computing rather than storing the synaptic weights in a separate memory. In this work, the intention to use a nonvolatile-material-based memory primitive is to eliminate the need for thermal tuners. We propose of a photonic neuromorphic platform from a scalable-system point of view based on a nonvolatile-memory primitive. Recent proposals [51,52] have looked at scalable systems to realize complex neural dynamics for dynamic learning. However, the flux-based memory in such systems is dependent on temperature and also on the run time of operation. Such detailed neurobiological functionalities make them more suitable for brainlike simulations similar to NeuroGrid simulations [53] in the electrical domain. In this work, we do not incorporate complex biological dynamics of SNNs in our system and rather focus on leveraging the inherent sparsity of spike-based processing while performing image classification for energy efficiency. The primary motivation behind our exploring this primitive stems from the building of a potentially reconfigurable neuromorphic system that performs energy-efficient inferencing. To build such neuromorphic platforms to perform spike-based processing in standard architectures, in-memory computing offers significant promise. To that effect, nonvolatile-memory primitives are quintessential and more suitable as they potentially eliminate the need for off-chip DRAM accesses, thus alleviating memory bottlenecks.

A popular way of implementing such spike-based inferencing systems is to train a network as an ANN and then convert it to a SNN by well-explored conversion algorithms [42]. This method has seen considerable success [44] in image classification, far beyond the scope of spike-based training algorithms. The neurons in ANNs are usually nonlinear mathematical functions, such as ReLUs, sigmoid functions, or hyperbolic tangent functions, with ReLUs being the most-popularly-chosen neuron functionality. During conversion, an artificial neuron with ReLU functionality can be directly converted to an integrate-and fire neuron mathematically [44]. This explains why we choose integrate-and-fire neurons as the spiking neurons in our proposal. Integrate-and-fire neurons are not associated with time constants as they do not include leak factors and the operations are fairly simple, unlike with other spiking neurons. The proposal concerns our building a spike-based photonic neuromorphic inferencing platform for image classification task. The neuron does not bear exact resemblance to a biological neuron; however, the design leverages the event-driven behavior of biological neurons. The aim of this work is to build a fast neuromorphic inferencing platform in the spiking domain to perform machine-learning tasks such as image classification. Several studies [53] previously explored brainlike neuron and synaptic functionalities with more-significant resemblance for complex neural simulations, albeit in the electrical domain.

The major advantage of building neuromorphic systems based on photonics rests in the speed of operation. The primary bottleneck in "write" latencies arises from the programming time of the integrate-and-fire neuron, which can be as low as 200 ps for the technology explored. Although the current technology is power expensive during writing, the speed of writing still enables us to achieve a reasonable energy efficiency. With further optimization of switching techniques or by use of alternative PCMs with lower switching power, further energy benefits can also be aimed for to achieve energy consumption comparable with that of other technologies in the electrical domain. In turn, the proposed photonics computing platform eliminates various drawbacks usually faced in the electrical counterparts, such as metal-wire resistance, electromigration, and sneak paths. Despite the inherent challenges in the design and implementation, our proposed SNN framework based on GST-on-silicon photonic neuromorphic fabric enables parallelism through integration of a synaptic network with integrate-and-fire neurons. Such a design paves the way for scalable photonic architectures suitable for large-scale neuromorphic systems able to perform fast computations.

VIII. CONCLUSION

We propose a photonic SNN computing primitive through seamless integration of nonvolatile synapses and integrate-and-fire neurons based on phase-change materials. The microring-resonator devices explored for such synapses and neurons leverage the differential optical absorption of GST for nonvolatility. We use the WDM technique to scale individual synapses into a large-scale synaptic array capable of performing parallelized dot products. Our design is based on ring resonators of radius comparable to the wavelength of operation to achieve high area density while maintaining performance. We explore several challenges involved in such small ring resonators and propose certain design modifications to achieve uniform and desirable characteristics across the entire operating wavelength range. Finally, we develop a device-to-system-level framework to evaluate the performance of the proposed photonic in-memory computing primitive and integrate-and-fire neurons as a SNN inferencing engine by building behavioral models of the photonic neuromorphic fabric and achieve performance comparable to that of an ideal network. Neuromorphic systems based on integrated photonics offer an alternative dimension to the current wave of exploring beyond-von Neumann computing frameworks, and our proposed photonic SNN inferencing engine achieves a significant step toward proposing individual nonvolatile devices capable of performing in-memory computing and scaling to a network of such devices to realize a truly integrated spiking neural network.

ACKNOWLEDGMENTS

This work was supported in part by the Office of Naval Research Multidisciplinary University Research Initiative, the National Science Foundation, Intel Corporation, and the Department of Defense Vannevar Bush Faculty Fellowship.

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, Deep learning, Nature **521**, 436 (2015).

[2] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis, Mastering the game of go with deep neural networks and tree search, Nature **529**, 484 (2016).

[3] Murray Campbell, A. Joseph Hoane, and Feng hsiung Hsu, Deepblue, Artif. Intell. **134**, 57 (2002).

[4] R. Serrano-Gotarredona, M. Oster, P. Lichtsteiner, A. Linares-Barranco, R. Paz-Vicente, F. Gomez-Rodriguez, L. Camunas-Mesa, R. Berner, M. Rivas-Perez, T. Delbruck, Shih-Chii Liu, R. Douglas, P. Hafliger, G. Jimenez-Moreno, A. C. Ballcels, T. Serrano-Gotarredona, A. J. Acosta-Jimenez, and B. Linares-Barranco, CAVIAR: A 45k neuron, 5M synapse, 12G connects/s AER hardware sensory–processing–learning–actuating system for high-speed visual object recognition and tracking, IEEE Trans. Neural Netw. **20**, 1417 (2009).

[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**, 668 (2014).

[6] Steve B. Furber, Francesco Galluppi, Steve Temple, and Luis A. Plana, The SpiNNaker project, Proc. IEEE **102**, 652 (2014).

[7] John Von Neumann, *The Computer and the Brain* (Yale University Press, New Haven, CT, USA, 2012).

[8] Avishek Biswas and Anantha P. Chandrakasan, in *2018 IEEE International Solid-State Circuits Conference (ISSCC)* (IEEE, 2018).

[9] Akhilesh Jaiswal, Indranil Chakraborty, Amogh Agrawal, and Kaushik Roy, 8T SRAM cell as a multi-bit dot product engine for beyond von-Neumann computing, arXiv:1802.08601.

[10] C. Mead, Neuromorphic electronic systems, Proc. IEEE **78,** 1629 (1990).

[11] Can Li, Miao Hu, Yunning Li, Hao Jiang, Ning Ge, Eric Montgomery, Jiaming Zhang, Wenhao Song, Noraica Dávila, Catherine E. Graves, Zhiyong Li, John Paul Strachan, Peng Lin, Zhongrui Wang, Mark Barnell, Qing Wu, R. Stanley Williams, J. Joshua Yang, and Qiangfei Xia, Analogue signal and image processing with large memristor crossbars, Nat. Electron. **1**, 52 (2017).

[12] Abhronil Sengupta and Kaushik Roy, Encoding neural and synaptic functionalities in electron spin: A pathway to efficient neuromorphic computing, Appl. Phys. Rev. **4**, 041105 (2017).

[13] Sukru B. Eryilmaz, Duygu Kuzum, Rakesh Jeyasingh, SangBum Kim, Matthew BrightSky, Chung Lam, and H.-S. Philip Wong, Brain-like associative learning using a nanoscale non-volatile phase change synaptic device array, Front. Neurosci. **8**, 205 (2014).

[14] Tomas Tuma, Angeliki Pantazi, Manuel Le Gallo, Abu Sebastian, and Evangelos Eleftheriou, Stochastic phase-change neurons, Nat. Nanotechnol. **11**, 693 (2016).

[15] Bipin Rajendran, Yong Liu, Jae-sun Seo, Kailash Gopalakrishnan, Leland Chang, Daniel J. Friedman, and Mark B. Ritter, Specifications of nanoscale devices and circuits for neuromorphic computational systems, IEEE Trans. Electron Devices **60**, 246 (2013).

[16] Ali Shafiee, Anirban Nag, Naveen Muralimanohar, Rajeev Balasubramonian, John Paul Strachan, Miao Hu, R. Stanley Williams, and Vivek Srikumar, Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars, ACM SIGARCH Comput. Architect. News **44**, 14 (2016).

[17] Kristof Vandoorne, Pauline Mechet, Thomas Van Vaerenbergh, Martin Fiers, Geert Morthier, David Verstraeten, Benjamin Schrauwen, Joni Dambre, and Peter Bienstman, Experimental demonstration of reservoir computing on a silicon photonics chip, Nat. Commun. **5**, 3541 (2014).

[18] Yichen Shen, Nicholas C. Harris, Scott Skirlo, Mihika Prabhu, Tom Baehr-Jones, Michael Hochberg, Xin Sun, Shijie Zhao, Hugo Larochelle, Dirk Englund, and Marin Soljacic, Deep learning with coherent nanophotonic circuits, Nat. Photonics. **11**, 441 (2017).

[19] Alexander N. Tait, Mitchell A. Nahmias, Bhavin J. Shastri, and Paul R. Prucnal, Broadcast and weight: An integrated network for scalable photonic spike processing, J. Lightwave Technol. **32**, 3427 (2014).

[20] Alexander N. Tait, Thomas Ferreira de Lima, Ellen Zhou, Allie X. Wu, Mitchell A. Nahmias, Bhavin J. Shastri, and Paul R. Prucnal, Neuromorphic photonic networks using silicon photonic weight banks, Sci. Rep. **7**, 7430 (2017).

[21] Carlos Rós, Nathan Youngblood, Zengguang Cheng, Manuel Le Gallo, Wolfram H. P. Pernice, C. David Wright, Abu Sebastian, and Harish Bhaskaran, In-memory computing on a photonic platform, arXiv:1801.06228.

[22] Carlos Rios, Matthias Stegmaier, Peiman Hosseini, Di Wang, Torsten Scherer, C. David Wright, Harish Bhaskaran, and Wolfram H. P. Pernice, Integrated all-photonic non-volatile multi-level memory, Nat. Photonics **9**, 725 (2015).

[23] Matthias Stegmaier, Carlos Rios, Harish Bhaskaran, C. David Wright, and Wolfram H. P. Pernice, Nonvolatile all-optical 1 × 2 switch for chipscale photonic networks, Adv. Opt. Mater. **5**, 1600346 (2016).

[24] Jiajiu Zheng, Amey Khanolkar, Peipeng Xu, Shane Colburn, Sanchit Deshmukh, Jason Myers, Jesse Frantz, Eric Pop, Joshua Hendrickson, Jonathan Doylend, Nicholas Boechler, and Arka Majumdar, GST-on-silicon hybrid nanophotonic integrated circuits: A non-volatile quasi-continuously reprogrammable platform, Opt. Mater. Express **8**, 1551 (2018).

[25] Zengguang Cheng, Carlos Rós, Wolfram H. P. Pernice, C. David Wright, and Harish Bhaskaran, On-chip photonic synapse, Sci. Adv. **3**, e1700160 (2017).

[26] Indranil Chakraborty, Gobinda Saha, Abhronil Sengupta, and Kaushik Roy, Toward fast neural computing using all-photonic phase change spiking neurons, Sci. Rep. **8**, 12980 (2018).

[27] Lin Yang, Ruiqiang Ji, Lei Zhang, Jianfeng Ding, and Qianfan Xu, On-chip CMOS-compatible optical signal processor, Opt. Express **20**, 13560 (2012).

[28] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio, in *Advances in Neural Information Processing Systems* (Curran Associates, 2016), pp. 4107–4115.

[29] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi, in *Computer Vision ECCV 2016* (Springer International Publishing, Amsterdam, The Netherlands, 2016), p. 525.

[30] Yiguo Chen, Xiong Li, Yannick Sonnefraud, Antonio I. Fernández-Domńguez, Xiangang Luo, Minghui Hong, and Stefan A. Maier, Engineering the phase front of light with phase-change material based planar lenses, Sci. Rep. **5**, 0 (2015).

[31] Nikolai V. Voshchinnikov, Gorden Videen, and Thomas Henning, Effective medium theories for irregular fluffy structures: Aggregation of small particles, Appl. Opt. **46**, 4065 (2007).

[32] Wolfram H. P. Pernice and Harish Bhaskaran, Photonic non-volatile memories using phase change materials, Appl. Phys. Lett. **101**, 171101 (2012).

[33] Wim Bogaerts, Peter De Heyn, Thomas Van Vaerenbergh, Katrien De Vos, Selvaraja Shankar Kumar, Tom Claes, Pieter Dumon, Peter Bienstman, Dries Van Thourhout, and Roel Baets, Silicon microring resonators, Laser Photon. Rev. **6**, 47 (2012).

[34] Qianfan Xu, David Fattal, and Raymond G. Beausoleil, Silicon microring resonators with 1.5-μm radius, Optics. Express **16**, 4309 (2008).

[35] Aayush Ankit, Abhronil Sengupta, Priyadarshini Panda, and Kaushik Roy, in *Proceedings of the 54th Annual Design Automation Conference 2017* (ACM, 2017), p. 27.

[36] Lumerical, Lumerical Inc. (2017).

[37] David E. Aspnes and A. A. Studna, Dielectric functions and optical parameters of Si, Ge, GaP, GaAs, GaSb, InP, InAs, and InSb from 1.5 to 6.0 eV, Phys. Rev. B **27**, 985 (1983).

[38] I. H. Malitson, Interspecimen comparison of the refractive index of fused silica, JOSA **55**, 1205 (1965).

[39] Sang-Youl Kim, Sang J. Kim, Hun Seo, and Myong R. Kim, in *Optical DataStorage'98* (International Society for Optics and Photonics, 1998), Vol. 3401, p. 112.

[40] Comsol, Multiphysics Reference Guide for COMSOL 4.2 (2011).

[41] Rasmus Berg Palm, Prediction as a candidate for learning deep hierarchical models of data, Technical University of Denmark **5** (2012).

[42] Peter U. Diehl, Daniel Neil, Jonathan Binas, Matthew Cook, Shih-Chii Liu, and Michael Pfeiffer, in *2015 International Joint Conference on Neural Networks (IJCNN)* (IEEE, 2015), pp. 1–8.

[43] MNIST handwritten digit database, http://yann.lecun.com/exdb/mnist/.

[44] Abhronil Sengupta, Yuting Ye, Robert Wang, Chiao Liu, and Kaushik Roy, Going deeper in spiking neural networks: VGG and residual architectures, arXiv:1802.02627.

[45] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, Learning representations by back-propagating errors, Nature **323**, 533 (1986).

[46] Vinod Nair and Geoffrey E. Hinton, in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (2010), p. 807.

[47] Benjamin C. Lee, Engin Ipek, Onur Mutlu, and Doug Burger, in *ACM SIGARCH Computer Architecture News* (ACM, 2009), Vol. 37, p. 2.

[48] H.-S. Philip Wong, Simone Raoux, SangBum Kim, Jiale Liang, John P. Reifenberg, Bipin Rajendran, Mehdi Asheghi, and Kenneth E. Goodson, Phase change memory, Proc. IEEE **98**, 2201 (2010).

[49] Abhronil Sengupta, Maryam Parsa, Bing Han, and Kaushik Roy, Probabilistic deep spiking neural systems enabled by magnetic tunnel junction, IEEE Trans. Electron Devices **63**, 2963 (2016).

[50] Indranil Chakraborty, Deboleena Roy, and Kaushik Roy, Technology aware training in memristive neuromorphic systems for nonideal synaptic crossbars, IEEE Trans. Emerging Topics Comput. Intell. **2**, 335 (2018).

[51] Jeffrey M. Shainline, Adam N. McCaughan, Sonia M. Buckley, Christine A. Donnelly, Manuel Castellanos-Beltran, Michael L. Schneider, Richard P. Mirin, and Sae Woo Nam, Superconducting optoelectronic neurons III: Synaptic plasticity, arXiv:1805.01937.

[52] Jeffrey M. Shainline, Jeff Chiles, Sonia M. Buckley, Adam N. McCaughan, Richard P. Mirin, and Sae Woo Nam, Superconducting optoelectronic neurons V: Networks and scaling, arXiv:1805.01942.

[53] Ben Varkey Benjamin, Peiran Gao, Emmett McQuinn, Swadesh Choudhary, Anand R. Chandrasekaran, Jean-Marie Bussat, Rodrigo Alvarez-Icaza, John V. Arthur, Paul A. Merolla, and Kwabena Boahen, Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations, Proc. IEEE **102**, 699 (2014).