Superconducting Optoelectronic Circuits for Neuromorphic Computing

Jeffrey M. Shainline,[*] Sonia M. Buckley, Richard P. Mirin, and Sae Woo Nam

National Institute of Standards and Technology, 325 Broadway, Boulder 80305, Colorado, USA

(Received 30 September 2016; revised manuscript received 9 February 2017; published 23 March 2017)

Neural networks have proven effective for solving many difficult computational problems, yet implementing complex neural networks in software is computationally expensive. To explore the limits of information processing, it is necessary to implement new hardware platforms with large numbers of neurons, each with a large number of connections to other neurons. Here we propose a hybrid semiconductor-superconductor hardware platform for the implementation of neural networks and large-scale neuromorphic computing. The platform combines semiconducting few-photon light-emitting diodes with superconducting-nanowire single-photon detectors to behave as spiking neurons. These processing units are connected via a network of optical waveguides, and variable weights of connection can be implemented using several approaches. The use of light as a signaling mechanism overcomes fanout and parasitic constraints on electrical signals while simultaneously introducing physical degrees of freedom which can be employed for computation. The use of supercurrents achieves the low power density ($1\ \mathrm{mW/cm^2}$ at 20-MHz firing rate) necessary to scale to systems with enormous entropy. Estimates comparing the proposed hardware platform to a human brain show that with the same number of neurons (10^{11}) and 700 independent connections per neuron, the hardware presented here may achieve an order of magnitude improvement in synaptic events per second per watt.

DOI: 10.1103/PhysRevApplied.7.034013

I. INTRODUCTION

Many foundational concepts in information theory and computing were developed beginning in the 1930s and 1940s through the work of Turing [1], von Neumann [2], Shannon [3], and others. Given the variety of proposed approaches to computing, it is somewhat surprising that the current landscape of computing technologies exclusively uses the von Neumann architecture. There has long been an interest in the relationship between information, computation, and cognition [4,5]. Computing architectures drawing inspiration from biological neural systems have been considered for decades [6], but investigation of novel architectures is only now becoming urgent as we reach the end of Moore's law scaling. The recent surge in deep learning and neural networks marked by advances in hardware [7–9], applications [10], and theory [11–13] has increased our understanding of the importance of such systems for solving complex problems.

Lin and Tegmark have recently argued [13] that the physics of our Universe is conducive to representation by neural networks. While there is an infinite number of possible functions a network may try to approximate, only a very limited subset will be of interest in our physical world. Additionally, it has been shown mathematically [11,12] that the ability of a neural network to accurately represent different kinds of functions (the expressivity of the network) scales as k^{mn}, where m is the dimension of the input, n is the number of hidden layers, and k is the number of nodes in each layer. This insight informs us that we can improve a network's ability to represent a broad range of functions both by increasing its width (k) and depth (n). Further, since the total information capacity of a computing system is proportional to the entropy, which scales with the number of distinct states which can be addressed by the system [14], computing systems based on complex interconnected networks, such as biological neural systems, offer extraordinary computational power.

To further maximize the information-processing capacity of such a system, it is desirable to fully utilize the time domain. For resilience to noise as well as temporally encoded information [15,16], signal communication via pulses, or spikes, is most advantageous, and such spike-encoded information is most powerful when many connections are established between processing units [16]. All of these findings taken together inform us that to implement neural networks most effectively in hardware, we should develop systems with a large total number of processing units, a large number of connections between units, and pulse-based communication.

Much like the von Neumann architecture has dominated modern computing, the hardware of silicon microelectronics has been similarly preeminent. It is possible that the ideal hardware platform for the next generation of computer architectures will also look very different. We make two conjectures which lead us to the hardware platform presented here. The first is that photons, based on their

[*] jeffrey.shainline@nist.gov

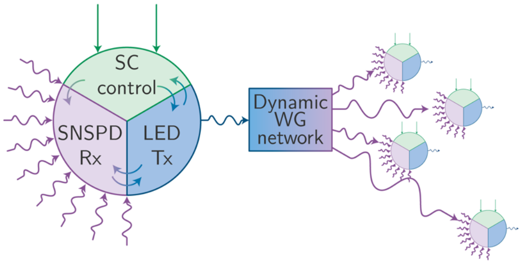

FIG. 1. Schematic representation of the proposed device concept. SC, superconducting; SNSPD, superconducting-nanowire single-photon detector; Rx, receive; Tx, transmit; WG waveguide.

noninteracting bosonic nature, will prove advantageous over electrons for achieving spike-based communication over networks with a large number of connections between nodes. That is to say, photonic fanout will overcome limitations of electronic fanout. The second conjecture is that superconducting circuits will enable lower power densities than semiconducting circuits, thereby leading to systems with a larger number of processing units and greater total complexity. In conceiving of a hardware platform integrating photonic with superconducting devices, we find a feasible route to highly scaled, multiphysical systems with extraordinary potential for computing complexity and experiments in information physics. A schematic representation of the concept is shown in Fig. 1.

The optoelectronic hardware platform is based on waveguide-integrated semiconductor light emitters working with superconducting detectors and electronics to implement weighted, directed networks [17]. Optical signals between neurons are communicated through reconfigurable nanophotonic waveguides. Utilization of light-emitting semiconductors allows efficient access to photonic degrees of freedom (frequency, polarization, mode index, intensity, statistics, and coherence), which achieve complex functionality analogous to chemical signaling in biological organisms and possibly with information-processing capabilities far beyond. Light enables massive interconnectivity with no need for time-multiplexing schemes that can limit the event rates of complementary metal-oxide-semiconductor (CMOS) systems [9,18]. Photonic signals are received and integrated by superconducting single-photon detectors. Firing thresholds and gain are controlled by a dynamic superconducting network, and neuron-generated photonic signals can reconfigure this current-distribution network. By employing superconducting electronics, we can approach zero static power dissipation [19], extraordinary device efficiencies, and utilize Josephson-junction circuits including single-flux-quantum devices [20–22].

Within this hardware platform, memory can be implemented via several means. These include temporally fixed synapses achieved with branching waveguides, synaptic weight variation via the actuation of locally suspended waveguides or through the use of magnetic Josephson junctions [23], or other magnetic and flux-storage components. The suspended waveguides that we explore in more detail in this work are reconfigurable on a time scale of 1 μs. None of these approaches draw power in the steady state.

The combination of efficient faint-light sources and superconducting-nanowire single-photon detectors interacting in an integrated-photonics environment enables neuronal operation with excellent energy efficiency, enormous intra- and interchip communication bandwidth, light-speed-limited latency, compact footprint, and relatively simple fabrication. The optoelectronic hardware platform is predicted to achieve 20 aJ/synapse event. By comparison, many CMOS systems are on the order of 20 pJ/synapse event [9,24,25], or in more recent work, hundreds of femtojoules per synapse event [26]. For these reasons, the proposed platform appears promising for advanced neuromorphic computing at the highest level of performance, while the compact nature and room-temperature operation of CMOS circuits will inevitably remain better suited for a wide range of neuromorphic applications.

The article is organized as follows. In Sec. II, we present the foundational neuronal optoelectronic circuits and consider each of the requisite constituent components. In Sec. III, we discuss the coupling of these circuits as well as mechanisms for reconfigurable memory enabling plasticity and learning. In Sec. IV, we discuss concrete applications of this hardware platform and consider the spatial and power scaling. We conclude with Sec. V. Details of the device design are presented in the appendixes.

II. OPTOELECTRONIC NEURONAL CIRCUITS

Information in neural systems is often referred to as "spike encoded," as interconnected neurons transmit information to one another in pulses [27]. An individual neuron (also referred to as a "processing unit," or simply, "unit") receives pulses from a number of upstream neurons. The neuron's input-output relation will be nonlinear, and if the integrated upstream signals exceed a certain threshold, the neuron may itself fire a pulse to its downstream connections. In this section, we describe superconducting optoelectronic circuits to emulate several biological neural responses. These circuits use integrated light-emitting diodes (LEDs) as transmitters with optical detectors as receivers. We next discuss the requirements for detectors and LEDs for this platform, and we motivate our choice from current technologies. Based on these choices, the energy per firing event is calculated.

A. Detector choice

A neuron that uses photonic signals requires both a source of photons and a photon detector. The choice of detector is critical to the design and analysis of the hardware platform. The central aim of this hardware platform is to achieve massive scaling to large numbers of interacting neurons. Therefore, simple waveguide integration, extreme energy efficiency, high yield, and small size are principal concerns. A review and comparison of single-photon detectors can be found in Ref. [28]. Of all existing detector options, only those based on superconductors allow single-photon detection in the infrared with zero static power dissipation and single-photon sensitivity to enable operation at the shot-noise limit. Because a system based on superconducting detectors will enable operation in this limit, it offers a useful platform to test the role of noise in learning and evolution of complex, dynamical systems.

There is an additional energy cost associated with cooling superconducting detectors to cryogenic temperatures necessary for operation. Therefore, an alternative is to move away from low-light levels and use integrated detectors such as Si [29–31], Si defect [32,33], Ge-on-Si [34–36], or III–V detectors, either bonded to Si [37] or on a fully III–V platform [38]. Such detectors have low signal-to-noise ratio requiring operation with significantly higher optical powers than if superconducting detectors are employed. While it may be possible to develop neuromorphic technology based on many of these detectors, we choose for this article to focus on superconducting-nanowire single-photon detectors (SNSPDs) due to the high efficiencies (>90%) [39] at wavelengths below the Si band gap, simple on-chip waveguide integration [40–46], compact size, and speed. While operation at cryogenic temperatures imparts a fixed energy cost, the energy cost per operation is significantly decreased by allowing integration with superconducting electronics. Therefore, cryogenic systems are of use in a subset of neuromorphic applications where the required system size is sufficiently large that the savings in chip power outweigh the cryocooling cost. Additionally, low-temperature operation allows the use of certain LED designs that are not possible at room temperature, as we discuss in Sec. II G.

B. Integrate-and-fire circuit

To encode information, the nodes of a neural network must have a nonlinear input-output relationship. In the proposed system, that nonlinearity is achieved via the transition of wires from the superconducting phase to the normal-metal phase. These phase transitions can be induced by absorption of a photon or by exceeding the critical current. A single SNSPD can be designed to fire with close to unity efficiency upon absorbing a single photon. We can think of this as an integrate-and-fire neuron in the limit of a single-photon threshold. In order to obtain an integrate-and-fire response with a threshold photon number larger than one, SNSPDs can be configured in parallel (step response) or series (continuous response). In Fig. 2(a), we show a circuit diagram of the parallel SNSPD array referred to as a parallel nanowire detector (PND) [47,48]. One example of an integrate-and-fire circuit is accomplished by placing the PND in parallel with a LED. The thresholding mechanism is explained pictorially in Figs. 2(b)–2(e). In the steady state, the PND is superconducting and has zero resistance. The semiconducting LED has finite resistance, and, therefore, all current from the source I_b flows through the PND. When a sufficient number of nanowires in the PND has been driven to the normal state by the absorption of photons, the critical current of the array is exceeded, the array becomes resistive, and current is diverted to the LED. This diversion of current and the subsequent production of light via carrier recombination constitutes the firing event. The LED fires with a step response, meaning that the LED output is independent of the exact number of photons absorbed and depends only on whether or not the threshold is exceeded. The diversion of current to the LED allows the PND to return to the superconducting state. Once this occurs, current ceases to flow through the LED, the production of light stops, and the device is reset.

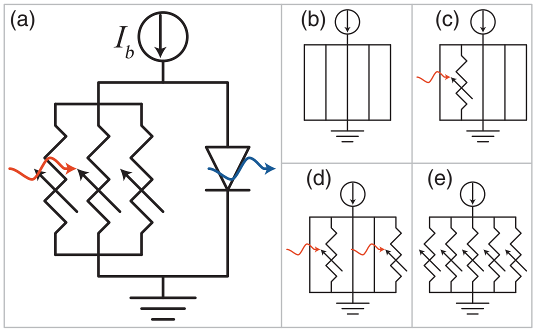

FIG. 2. (a) PND neuron circuit. (b) A PND with all wires superconducting. (c) A PND where one of the wires is driven normal by absorption of a single photon, redirecting the current through the other four. (d) A PND with two normal wires due to absorption of two photons. (e) A PND with all wires driven normal by exceeding the critical current. A LED in parallel with this PND now receives current, causing a firing event.

The minimum duration of a spike event is determined by the emitter lifetime. The integration time of the neuron can be engineered to be within the range of a few hundred picoseconds up to seconds. See Appendix A for a more detailed discussion of the temporal response of the circuits.

To model the spike probability of this circuit, we conduct Monte Carlo simulations of the device. The critical number of absorbed photons n_c is given by

$$n_c = N_{\mathrm{NW}} - \frac{I_b}{i_c}, \tag{1}$$

where N_{NW} is the number of nanowires in the array, I_b is the bias current for the entire array, and i_c is the critical current of a single wire. Equation (1) is derived in Appendix B. Although each individual firing event

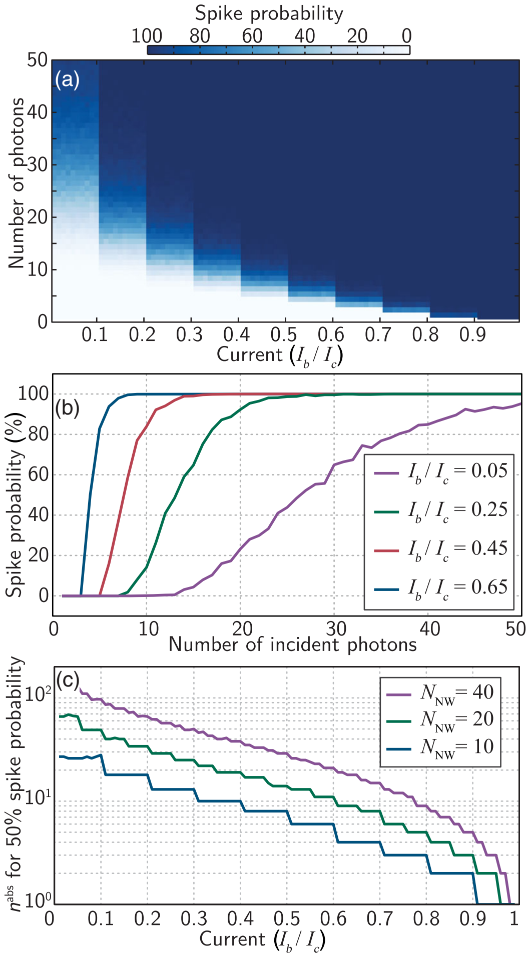

FIG. 3. Monte Carlo simulation of spike probability. (a) PND with ten SNSPDs. (b) The same simulation as (a) but with four traces isolated for clarity. (c) The number of absorbed photons which gives a 50% absorption probability plotted as a function of bias current. Traces for PNDs with 10, 20, and 40 nanowires are shown.

generates the same current pulse across the LED (i.e., a step response), a given number of input photons causes only the neuron to fire with some probability. This is due to the stochastic nature of the photon-absorption events, which we discuss in more detail in Appendix C. The results of these simulations are shown in Fig. 3. The probability of a spike occurring is plotted as a function of the number of photons incident on the device for various bias currents ranging from 0.01 of the array critical current (I_c) to 0.99 I_c in steps of 0.01 I_c. In Fig. 3(a), we show the behavior of an array with ten SNSPDs in parallel. Figure 3(b) shows the spike probability versus the number of incident photons for four values of bias current; these data are a subset of that shown in Fig. 3(a), plotted separately to illustrate the shape of the traces. The Monte Carlo simulations which produce these plots are conceptually based on the neuron design of Fig. 4 and proceed as follows. A given number of photons is assumed to be incident on a PND array. The pulse is assumed to pass each nanowire of the array in sequence. At each pass, a random number between zero and one is generated. If this random number is less than or equal to the assumed absorption probability (1% in these calculations), the number of photons in the pulse is reduced by one, and the state of that nanowire is set to nonsuperconducting. The photon pulse is allowed to pass each nanowire of the array 100 times. The number of photons in the pulse which cause Eq. (1) to be satisfied is recorded for each bias current. The result of 1000 such simulations is averaged to calculate the probability for spiking to occur.

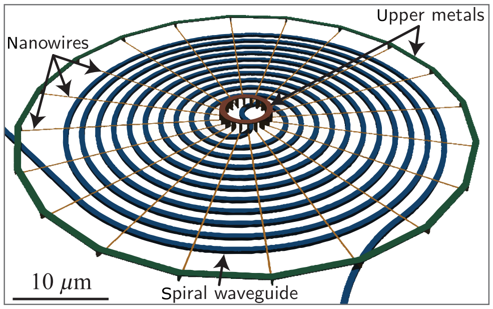

FIG. 4. The spiderweb neuron. The scale bar is shown for reference, but significantly more compact implementations of this device can be achieved.

In Figs. 3(a) and 3(b), we observe that by adjusting the bias current, we can adjust the shape of the firing function versus photon number. Yet, adjusting the bias current cannot tune the threshold with arbitrary accuracy. In Fig. 3(a), it is evident that the spike probability for a PND array with ten nanowires separates into ten bands. Therefore, to achieve higher-photon-number differentiation, more wires must be integrated. This point is illustrated in Fig. 3(c). Simulations similar to that of Fig. 3(a) are conducted for PND arrays with 20 and 40 nanowires, and the number of absorbed photons (n^{abs}) for which the spike probability reaches 50% is plotted versus the bias current. This figure further illustrates that the resolution of the PND array is limited by the number of nanowires in the array, resulting in discrete steps in the number of photons required for a spike event as a function of bias current. Because n_c and N_{NW} in Eq. (1) are both integers, the floor of the ratio I_b/i_c is effectively taken, and the utility of the current for setting the threshold is discretized. For the case of $N_{\mathrm{NW}} = 40$, the steps become quite small, and the curve is approximately continuous.

The simple model of Fig. 3 reveals that the PND array can achieve a high dynamic range in that the threshold can be tuned broadly in hardware by changing the number of wires in the array (from a single nanowire up to potentially thousands) as well as actively during operation by changing the bias current. The state space of the receiver, which scales as $2^{N_{\mathrm{NW}}}$, can be made quite large in the regime where thousands of nanowires comprise the PND.

Figure 4 presents a neuron design well suited to a system with a few tens and possibly hundreds of connections. We refer to this device as the spiderweb neuron. In this design, all upstream signals are combined on a single waveguide. This waveguide enters a spiral region in which it passes a number of SNSPDs which can be wired in series or parallel. Photon wave packets can pass several tens of SNSPDs several tens of times. The system can, thus, be engineered to spread the absorption probability evenly over the SNSPDs. In Fig. 3, the photons are assumed to pass each nanowire 100 times with a probability of absorption of 1% at each pass. The size of the detector portion of this neuron can be made as small as $10 \times 10\ \mu\mathrm{m}^2$ and depends on the thresholding number of photons. For a threshold of 1000 photons, the device is approximately $35 \times 35\ \mu\mathrm{m}^2$. We discuss the model in more detail in Appendixes A and C, and other neuron designs are discussed in Sec. III A. In the calculations of Fig. 3, we assume all photons arrive in a short pulse, so nanowire rebiasing dynamics can be neglected. The complex dynamics of the PND receiver array in the case of arbitrary photon-arrival times is the subject of future investigation.

C. Differentiable response circuit

In biological systems, the neuron response is not that of a step function but rather a nonlinear response taking the form of a sigmoid. For certain neural-network backpropagation algorithms, it is important that the response be continuous and differentiable [49]. Figure 5(a) shows the series-nanowire-detector (SND) [50] circuit which achieves a continuous and differentiable nonlinear response. In Fig. 5(b), we define a general optoelectronic circuit element symbolizing either the PND (Fig. 2) or the SND (Fig. 5). We envision the SND as a single length of superconducting wire with incident photons spread along the length of the wire. As in Fig. 2(a), the detector array is in parallel with the LED. When a single photon is absorbed by the SND, a length of normal wire called a hot spot emerges in series with the superconductor, leading to current redistribution between the two branches of the circuit. For common SNSPD materials, this resistance is approximately 1 kΩ for the typical wire width, while the length of the single hot spot is on the order of 100 nm [51,52]. As more photons are absorbed, more hot spots are created, and the resistance of the SNSPD increases. This resistance causes the voltage across the LED to increase, and sufficient current can be driven through the diode to produce an optical signal.

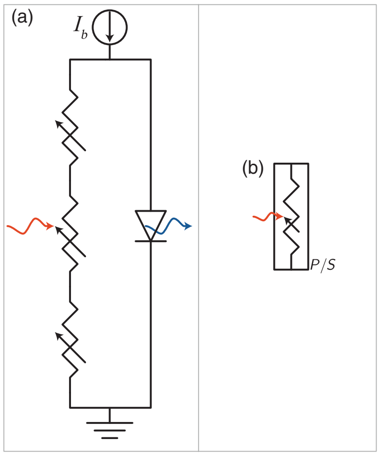

FIG. 5. (a) SND circuit. (b) Component diagram indicating either SND or PND array. This circuit symbol is used throughout this article.

While attempts have been made to utilize this effect for number-resolving single-photon detection [50], we emphasize that we propose to utilize this circuit in a very different operating regime. To detect a single photon with near-unity efficiency, a SNSPD is driven close to its critical current, and the ensuing voltage pulse is measured across a 50-Ω resistor in parallel with the SNSPD. When a photon is absorbed, a 1-kΩ hot spot is produced, and nearly all current is diverted to the 50-Ω load. For the application at hand, the device is not intended to observe events of one or a few photons but rather hundreds to thousands. Thus, diverting the current through a high-impedance diode with I-V relationship approximated by Eq. (D1) enables thresholding with some dynamic range for higher numbers of absorbed photons. The model of this SND-based neuron considers simple joule heating behavior in that each photon-absorption event results in the same hot-spot resistance, when in reality, the hotspot resistance depends on the current through that branch of the circuit, which depends on the temporal dynamics of the preceding absorption events. A thorough study of these dynamics is the subject of future work.

The electro-optic performance of the SND is analyzed in Fig. 6. The nanowire resistance as a function of the number of absorbed photons is shown in Fig. 6(a). In this model, we assume the photons are incident upon a length of out-and-back nanowire [40–46] with 100-μm attenuation length, and it is assumed that two photons absorbed at the same location along the nanowire give rise to the same resistance as a single photon absorbed at that location. For this reason, the nanowire resistance levels off as a function of the number of absorbed photons. The current-voltage relationship of the LED is highly nonlinear, as shown in the inset, but above a certain number of absorbed photons, the entire length of the absorbing region of the superconductor is driven normal, and the absorption of additional photons results in no additional resistance, as shown in Fig. 6(b) and 6(c). Hence, the device has an input-output relationship with an exponential turn-on when a threshold number of photons is absorbed followed by a flattening of the output when the entire SND is driven normal. Figures 6(b) and 6(c) show the photon input-output relationship for two different nanowire designs with critical currents of 4 and 8 μA, respectively, demonstrating the ability to tune the response in hardware. Note that the photon input-output

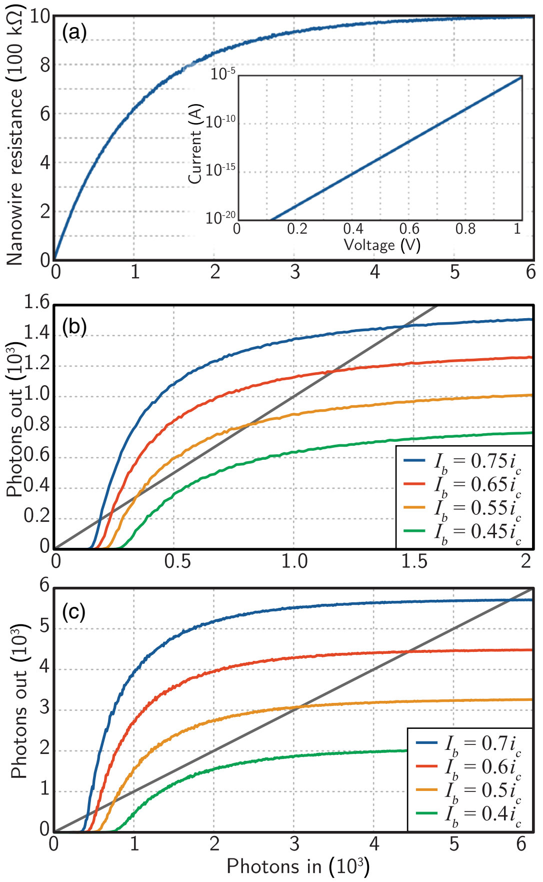

FIG. 6. Electrical characteristics for SND with $l_{\mathrm{wire}} = 100\ \mu$m. (a) Resistance versus number of photons for the SND. Inset shows the exponential current-voltage curve for the LED. Photons out versus photons in for SNDs with (b) $i_c = 4\ \mu$A, $\eta = 1\%$ and (c) $i_c = 8\ \mu$A, $\eta = 0.1\%$. Here, η is the efficiency of the LED.

relationship depends on the refractory period, as we discuss in Appendix A.

Based on the analysis of Fig. 6, in the SND-based neuron, the normal-state resistance of the SND and the applied bias determine the maximum voltage that can be achieved across the LED. This resistance and bias, in conjunction with the optoelectronic design of the LED, determines the number of photons generated, in contrast to the case of the PND where the number of photons generated is a step response determined by the bias current.

Both the PND-based integrate-and-fire circuit of Fig. 2(a) and the SND-based continuous-response circuit of Fig. 5(a) may offer utility for neuromorphic computing. For the case of the PND, the number of nanowires in the array is on the order of the number of photons required for threshold. This is also the order of the number of connections each processing unit makes to other units. Biological systems reveal that scaling to systems with thousands of connections per neuron is desirable [16]. To achieve this number of parallel receiver elements, several geometrical configurations can be utilized to arrange approximately 1000 micron-scale SNSPD elements, and the exploration of this design space is the subject of future work.

The SND device straightforwardly lends itself to hundreds or thousands of connections. In this case, we can expect the thresholding number of photons to be approximately 1000, and, therefore, we want a nanowire with the length of 1000 hot spots. Given the hot-spot length of 100 nm, the entire length of the nanowire is on the order of 100 μm, as simulated in Fig. 6. Such a length becomes quite compact when coiled in a spiral [see Fig. 11(b)], and as we discuss in Sec. III A, this configuration is well suited to receive inputs from hundreds to thousands of waveguides. We discuss the energy requirements of the SND and PND circuits in Sec. II F.

D. The nTron current amplifier

Introducing an amplifier into the circuits described in Secs. II B and II C allows decoupling of the firing threshold and LED gain. In a superconducting circuit, amplification can be achieved using the nTron, a three-terminal supercurrent amplifier [53]. When the current in the gate terminal exceeds the critical current, the path from the source to drain is driven normal, diverting the bias current to the parallel load. This recently developed device has been used to drive loads of tens of kilohms, making it suitable for this application.

In Fig. 7(a), we show a variation of the circuit of Fig. 2(a), but instead of driving the same current I_1 through the LED after firing, this circuit utilizes a nTron current amplifier to provide gain to the light emitter. The nTron allows us to decouple the current used to bias the receiver from the number of photons produced in the firing event. Note that in this configuration, I_2 can be less than I_1, making it possible to cover a broad range of input-output responses. The circuit of Fig. 7(a) also expands the state space in which information can be encoded.

E. Other neuromorphic circuits

We introduce the basic neuromorphic circuits in Secs. II B and II C. We now introduce several variants on those cells which enable diverse functionality desirable for neuromorphic computing.

Figure 7(b) shows an alternate configuration in which the LED is driven by current I_2 until a firing event occurs and cuts off the current supply. This circuit is shown with the LED below the nTron, but it can also be implemented without a nTron. This circuit is an example of an integrate-and-stop-firing neuron which can be useful in neuromorphic architectures to provide a means of stimulating various regions of the cortex until a certain level of activity is reached, at which point the firing neuron is quenched.

Another essential functionality of neuromorphic circuits is that of inhibitory connections [54,55]. Most neuronal connections provide feedforward excitation wherein an action potential produced by upstream neurons increases

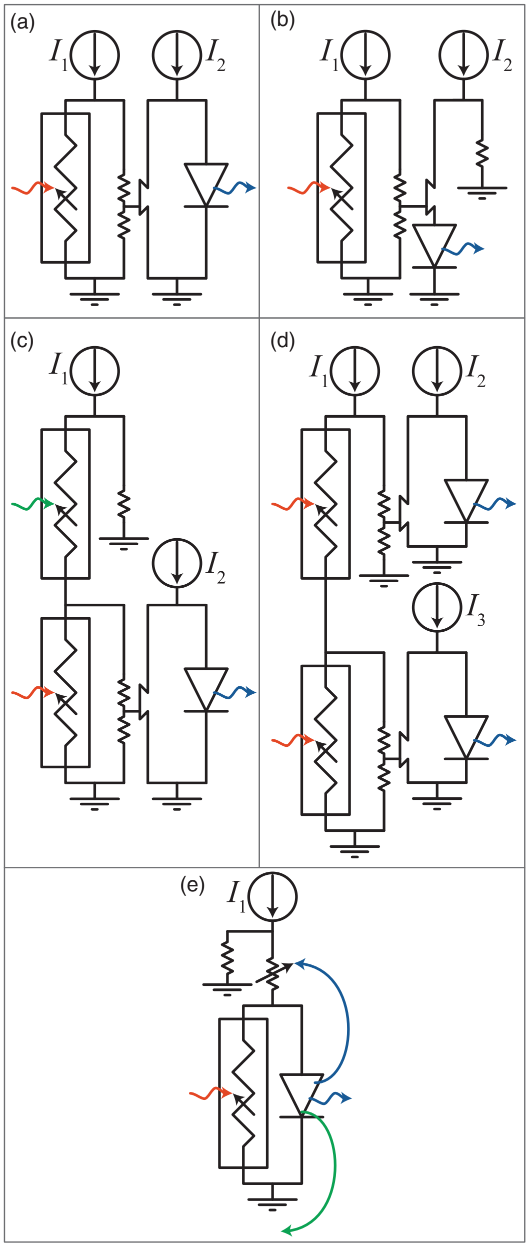

FIG. 7. Various neuromorphic circuit configurations. (a) PND with nTron amplifier. (b) Integrate-and-stop firing. (c) Neuron with the possibility for both excitatory and inhibitory excitation. In this figure, green corresponds to photons inhibiting firing and red to photons exciting firing. These photons can have different colors. (d) Firing of the upper neuron inhibits firing of the lower neuron. (e) Circuit for achieving self- and upstream feedback.

the probability of action potentials being produced by downstream neurons. But biological systems also exhibit connections wherein the firing of upstream neurons suppresses the probability of firing events by downstream neurons. Figure 7(c) shows a configuration which achieves this. The lower portion of the circuit is identical to that of Fig. 7(a), but the current I_1 feeding the receiver first passes through a preliminary nanowire array. Absorption of photons in this region of the circuit reduces the current through the primary receiver, increasing the threshold photon number. Waveguides from different upstream neurons can be routed to these two different ports to establish inhibitory or excitatory connections. In Fig. 7(c), the inputs to the two receivers are drawn with different colors, emphasizing the possibility that integrated-photonic filters placed before the neuron can be employed to route different frequencies to the two receivers. With this approach, we can employ the use of color to perform inhibitory or excitatory functionality in much the same way that different neurotransmitters perform inhibitory or excitatory functions in biological systems [54]. We note that low-loss spectral filters performing this function are commonplace in many integrated-photonic applications.

From an architectural standpoint, it may also be useful to establish purely electrical inhibitory connections. In Fig. 7(d), we show a circuit in which two neurons, each with only a single excitatory port, are connected in series. In this configuration, firing events in the upper neuron inhibit firing events in the lower neuron. Such a configuration is useful for moderating the net firing activity of groups of neurons.

It is also advantageous to have a means by which a single neuron can moderate its own firing activity. Such behavior is straightforward to implement, as is shown in Fig. 7(e). A power tap is added to the output of the LED, and some fraction of the produced light is incident upon a receiver in series with the current supply to the receiver array. The superconducting wire in this location may be wider than the integrating receiver, and it, therefore, may be designed to quench the current only when a large number of photons drives the superconducting wire normal.

In addition to self-feedback, biological neurons send both downstream signals as well as upstream signals when an action potential fires. The upstream signals are believed to be critical for spike-timing-dependent plasticity and synchronization of circuit behavior via threshold modification. To briefly hint at how self-feedback may be implemented in the proposed platform, the green arrow leaving the LED in Fig. 7(e) indicates that a power tap can also be used for upstream feedback. The color of this arrow is meant to remind us that it may be advantageous to use different frequencies of light for downstream and upstream signaling. A LED can be fabricated to emit at two distinct wavelengths or across some region of bandwidth, and integrated spectral filters can be employed to route the two signals. Alternatively, two different LEDs coupled to two different waveguides can be utilized.

In this section, we present several superconducting optoelectronic neuromorphic circuits covering a wide range of functions. We refer to members of this class of circuits as single-photon optoelectronic neurons (SPONs). We now proceed to discuss additional aspects of their performance.

F. Energy consumption

We introduce the basic SPON circuits of the proposed neuromorphic computing platform, and we are in a position to estimate the energy required for a firing event. A complete neuron firing event involves supplying current to the inductors associated with all superconducting wires (including the detectors), charging the capacitor associated with the LED *p-i-n* junction, and driving current through the LED to produce light. For the case of the PND circuit of Fig. 7(a), we analyze the energy consumption of each of these three contributions.

In this model, we assume one inductor L_{SNSPD} in the PND array for each photon, as well as a series inductance to achieve the desired temporal response (see Appendix A). We assume each element of the PND is 500 squares, while the entire receiver array is in series with 5000 squares of inductance. At low photon numbers, the energy consumption from inductance is dominated by the series inductance, but for higher numbers, it is dominated by the PND array and grows linearly. The energy required for photon production is calculated simply as $E_g n_\nu/\eta$, where E_g is the band gap of Si, n_ν is the number of photons created, and η is the efficiency. Thus, within this model, the contribution to energy consumption due to photon creation is linear throughout. We use E_g in this model because it is an upper bound on the photon energy. Any photon transmitted through a Si waveguide will have energy below the band gap. We assume a superconducting material with a sheet inductance of 400 pH/□ (such as WSi), and a parallel-plate capacitive model for the LED as described in Appendix D.

In Fig. 8(a), we plot the total energy per photon as a function of the number of photons emitted for four values of LED efficiency. We find that with a unity-efficiency LED, the energy per photon can be as low as 2 aJ when larger photon numbers are created. This remarkably low number is still an order of magnitude greater than the 0.16 aJ stored in the $h\nu$ of the light quantum itself (assuming $\lambda = 1.22$ μm), with the extra energy going to supplying current to the inductors and charge to the capacitor. The figure reveals that producing LEDs with efficiency above 10% has only a modest benefit, as the contribution to energy consumption from inductance will become the limiting factor. However, for thresholding on larger photon numbers, as is desirable for neurons with more connections, the inductance per photon can likely be reduced. While a 100% efficient LED may not be realized, even a 1% efficient LED leads to 20 aJ/synapse event. This energy efficiency illustrates the promise of superconducting electronics and faint-light signals.

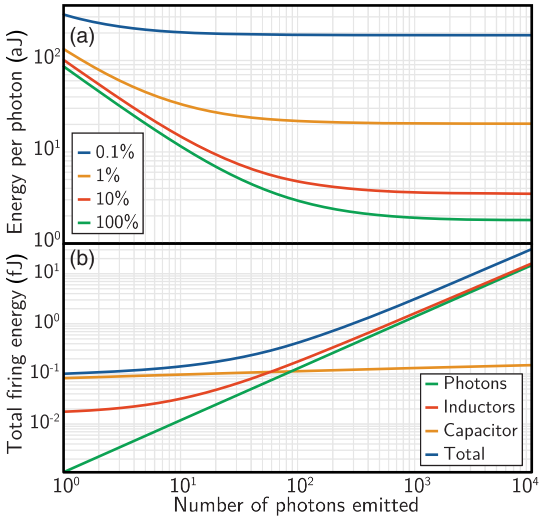

FIG. 8. (a) Energy required to generate a single photon versus number of photons emitted for four different LED efficiencies. (b) Contributions to total energy consumption for a 10% efficient LED.

In Fig. 8(b), we show the contribution to the total energy from the various circuit elements for the case of a 10% efficient LED. This efficiency is chosen for this plot because it is the value at which the contributions from inductance and photon production are nearly equal for photon numbers near or above 100. For low photon numbers, the dominant contribution is in charging the LED capacitor. Because of the highly nonlinear LED current-voltage relationship, a small increase in the voltage across the LED leads to a large gain in current. The capacitive energy is nearly constant across the range of photon numbers considered here, and for larger photon numbers, it makes a negligible contribution.

In the case of the SND circuit of Fig. 5 with parameters as shown in Fig. 6(b) driven at 0.6 I_c and receiving 10^3 photons, and assuming a hot-spot recovery time of 50 ns and a LED with 1% efficiency, the device achieves 100 aJ/synapse event. While not as efficient as the PND neuron, this device design still lends itself to massive scaling, as we discuss in Sec. IV B.

We believe a LED with 1% system efficiency is realistic in a nanophotonic environment at cryogenic temperature and with faint-light levels desired. Therefore, we use 20 aJ/photon as a representative number for what this platform can hope to achieve. We use the energy per photon as the energy per firing event per synapse (commonly referred to as the energy per synapse event), because the goal of the system is to produce neurons which threshold on a number of photons roughly equivalent to the number of connections made by the neuron. A neuron receiving 100 signals from upstream will threshold on 100 photons. It will produce 100 photons in a firing event and distribute them amongst 100 downstream synapses. Therefore, the energy per synapse event is calculated as the total energy of the firing event divided by the number of connections. In our case, for systems with 100 to 10 000 connections per unit, 20 aJ/synapse event is a realistic number.

The second law of thermodynamics informs us that to keep a system at 2 K, 150 W of cooling power is required per watt of power dissipated at 2 K. Assuming a 15% efficient cooling system, this gives an estimate of 1 kW of cooling power per watt of device power. Multiplying our conservative estimate of 20 aJ/synapse event by this factor of 10^3, the hardware achieves an energy consumption of 20 fJ/synapse event. Similarly, while the human brain uses 20 W to perform roughly 10^{14} synapse events per second, a power budget of 20 W corresponding to 20 mW of device power will enable our system to achieve 10^{15} synapse events per second. Success in developing LEDs with higher efficiency, reduction of the device inductance, and utilization of superconducting materials operating at higher temperatures will further increase the advantage. Additionally, while transistor technologies inevitably leak current, superconducting devices can be engineered to draw no power in the steady state and can be dc biased without loss using Josephson junctions [19].

G. Electrically injected light source

Having introduced the proposed optoelectronic neuronal circuits, we now proceed to analyze the operation and performance requirements of the LED. As we discuss in the previous section, we target operation efficiencies of around 10%. This efficiency is relatively easy to attain in III-V semiconductors such as GaAs and InP. However, for the application at hand, massive scaling is a priority, and massive scaling requires photonic electronic process integration. A single source with 100% efficiency is less desirable than the ability to scale to millions (and eventually billions) of sources each with 1% efficiency. We also require low-loss waveguides with the potential for reconfigurability (see Sec. III).

One option is to implement these devices on a GaAs or InP substrate. These have been the materials of choice for photonic integrated circuits where light sources are of the utmost importance. Quantum-dot-well LED lasers can be electrically injected with high efficiency on this platform [38] and combined with high-index (III-V) waveguides to form the synaptic connections described in Sec. III A. Another option is to implement the light sources in the III-V material and then couple to low-temperature deposited materials with low-loss waveguides [46,56] such as *a*-Si or SiN. A III-V platform has the advantage of high-efficiency light sources, but massive scaling on III-V substrates has historically been more difficult and expensive than on Si substrates. This drawback, while not fundamental, may prove significant in halting the development of this technology, especially since high emitter efficiencies are not a strict requirement for neuromorphic computing.

Another option is hybrid III-V silicon integration. Hybrid III-V silicon has followed one of three approaches 57]]: direct mounting, wafer bonding, or III-V material grown on Si. While direct mounting or wafer bonding are currently the preferred methods for optical interconnect applications, these applications typically require a single source that can be diverted to multiple components. For the proposed neuromorphic computing platform, we desire a separate electrically injected source for each neuron. Direct mounting, therefore, is not an option, but wafer bonding may be able to achieve the yield and reproducibility required for this application. Direct heteroepitaxial growth offers the most promise for hybrid integration with this system. In this case, the desired light source is templated III-V quantum dots grown in the intrinsic region of a lateral Si *p*-*i*-*n* junction. While great progress in this field has been made [58–64], additional effort is needed to achieve the waveguide-integrated sources required for this system. Promisingly, electrically injected single-photon emission has been demonstrated in these materials [58–61]. While single-photon emission is not a requirement for the present application, a desirable property of the emitters is that they have low-photon-number variance (defined as the standard deviation of the number of photons output for a given input current pulse over an ensemble of measurements). The fact that single-photon emission has been demonstrated in various systems indicates the possibility to bring this photon number variance down to the range of a few photons.

A major disadvantage of this heteroepitaxy approach is the significant cost and difficulty associated with growing these materials. As this approach matures, this material platform may become more desirable. Another similar approach using Ge [65] or Ge quantum dots [66] may also prove useful.

A commonly overlooked light source that may prove particularly promising for this application is emissive centers in Si [67]. These have proved unattractive for optical interconnects due to very low efficiencies at room temperature. Much work in this area was motivated by the prospect of room-temperature light sources [68] for CMOS and telecommunications [69] and, in particular, room-temperature lasers. This includes various point defects in Si including Er [70–73] and other emissive centers giving rise to electric-dipole-mediated transitions [67,74–82], as well as band-edge or Si nanocrystal-based emission processes [83–85]. While the efficiencies of many of these emitters fall off exponentially with increasing temperature, the SNSPDs required for this application operate at cryogenic temperatures where many point defects have suitable efficiencies. A large number of emissive centers are under consideration for this application [67].

The main challenge is the successful integration of large numbers of emitters with the ultimate goal being billions integrated in a system. Many emissive centers can be easily fabricated in a CMOS-compatible process via ion implantation and annealing [67,75,80–82,86,87]. A schematic of the desired device is depicted in Fig. 9. A *p*-*i*-*n* junction is created in a ridge waveguide. Emitters are located only in the ridge (intrinsic) region via lithographic patterning, and light is obtained from forward biasing the junction. While

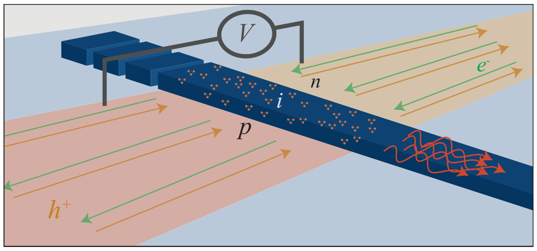

FIG. 9. Schematic of a monolithically integrated electrically injected emissive-center LED in Si for the proposed neuromorphic computing application.

this is a relatively standard configuration of a LED, for the application at hand it is important to keep the emitters localized only in the intrinsic region of the LED, as their presence elsewhere in the waveguides leads to intolerable loss. Thus, the ability to lithographically control the location of emitters is crucial.

With coimplantation of multiple impurities, it is possible to add additional (color) degrees of freedom to the platform. Similarly, on a III-V platform, we can take advantage of inhomogeneous broadening of the quantum-dot spectrum and tuning of dot size via templating or growth conditions.

We note that the neuromorphic computing platform proposed here is not tied to any one of these light sources, and indeed there are other possible light sources that we have not discussed. For the calculations throughout the present work, we assume LEDs with 1% efficiency at 1.22 μm in a waveguiding medium with index of 3.52 with a cladding of 1.46 above and below.

H. Summary

We have now presented several superconducting optoelectronic circuits capturing a broad range of neuromorphic behaviors. We have presented basic thresholding SPON circuits of Figs. 2 and 5, variants on these circuits as shown in Fig. 7 which enable gain, integrate-and-stop, and inhibitory connections, and circuits with self- and upstream feedback, as shown in Fig. 7(e). We now discuss the means by which we propose to connect these processing units.

III. CONNECTIVITY

Of central importance to the implementation of the proposed neuromorphic platform is the network of waveguides that connect the processing units. Optical waveguides offer the possibility for improved performance over electrical connections by allowing individual neurons to integrate signals from many sources without the need for time multiplexing. Because of the additional energy cost associated with the capacitance of additional wires [88], electrical neurons must utilize shared wires. Voltage pulses from different neurons on the same bus will interact. To prevent this, pulses must be delayed in time.

In the following section, we discuss how a network of optical waveguides can be implemented to form the connections between the SPON circuits presented in Sec. II. Each neuron has a waveguide exiting the LED and leading to many branching waveguides, which we liken to the axon and its arbor, and another set of integrating waveguides combining signals received from upstream neurons, which we liken to the dendritic arbor, as shown schematically in Fig. 1. The connections between these input and output waveguides act as synapses in this network. We outline a mechanism for varying the strength of the connections between various input and output waveguides, which is similar to varying synaptic weights in biological systems. We emphasize that other methods of connecting neurons in three dimensions using the same optoelectronic neurons are also possible. One can envision using gratings, flat lenses [89], metasurfaces [90,91], or optical phased arrays [92,93] to direct signals between neurons. Additionally, electrical means of changing synaptic weights at the receivers may prove useful.

A. The dendritic arbor

The dendritic arbor of a neuron collects signals from upstream neurons. For optoelectronic neurons, the equivalent of this is a waveguide network that combines optical signals from many other neurons to the neuron for detection. At each neuron, the device must be designed to combine the modes from a large number of waveguides on a PND or SND with low loss. There are likely many ways to achieve this functionality, and here we explore two.

A schematic of the first approach is presented in Fig. 10(a) showing the spiral waveguide receiver of the spiderweb SPON, the nTron, and the LED emitter. The major challenge of this device design is the merging of many single-mode waveguides into one multimode waveguide which enters the spiral. The proposed technique for accomplishing this is shown in Fig. 10(b). Two single-mode waveguides cannot be combined into one single-mode waveguide without significant loss [94]. However, two single-mode waveguides can be combined into one dual-mode waveguide nearly losslessly. In Fig. 10(b), several single-mode waveguides combine their power on a given main spine. That spine can receive at its input one single mode. As it continues to receive more modes, its width must grow. The lower-order modes of this adiabatically tapering multimode waveguide can pass each new single-mode input nearly losslessly as long as the width of the spine has grown to support an additional mode by the location of the next input waveguide. More detail regarding the optical design of this structure is given in Appendix E. Modal simulations reveal that a waveguide width of 2 μm in 200-nm-thick Si is sufficient to support several tens of modes at 1220-nm wavelength, each with tolerably small bending loss with a 10-μm radius of curvature. Therefore, this dendritic arbor and receiver design is suitable for the

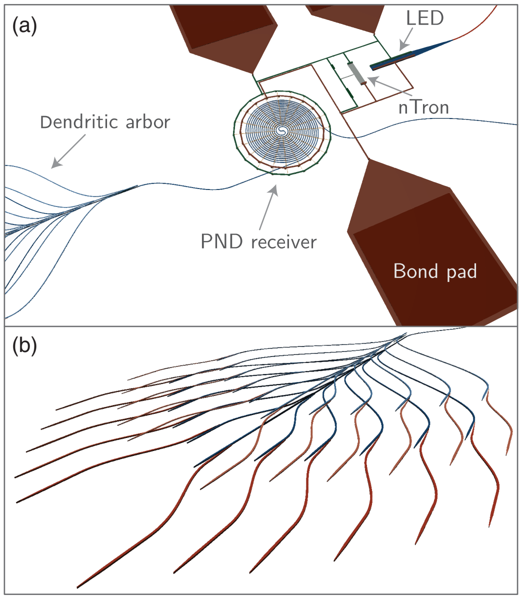

FIG. 10. The spiderweb neuron. (a) Overview of the device. (b) Dendritic arbor design which combines light from multiple neurons.

compact combining signals from approximately 40 upstream neurons.

The second proposed design is better suited to scaling to larger numbers of inputs. It is shown in Fig. 11. In this design referred to as the stingray SPON, the input waveguides are directly combined on a landing pad housing the PND or SND array. The implementation with a PND is shown in Fig. 11(a). As is shown in Appendix E, the minimum spacing required to avoid modal coupling is 600 nm at the input of the cell. From these input ports, the waveguides enter an array of sine bends where their spacing is reduced to enter the smaller landing pad containing the nanowires. In this sine region, intermodal coupling is tolerated (and perhaps even desirable to spread the photons across the nanowires), as all waveguides ultimately terminate on the detector array. Figures 11(b) and 11(c) show 2D finite-difference time-domain (FDTD) simulations of the structure. Figure 11(b) shows the propagation of light into the receiver body in the presence of absorbing nanowires, while Fig. 11(c) shows propagation without the absorbing nanowires. Here, 100 waveguides terminate on a receiver body with less than 0.2-dB insertion loss from any port, with the outermost ports giving the most loss, and the innermost ports achieving near-zero insertion loss. In this context, insertion loss refers to light entering and leaving the simulation without being absorbed in the nanowire array. Calculated quantitatively with pulsed excitation, we find the majority of loss is due to light scattering and not entering the detector array rather than being transmitted through the receiver due to inadequate absorption. The entire receiver of Fig. 11(b) occupies $30 \times 30\ \mu\mathrm{m}^2$.

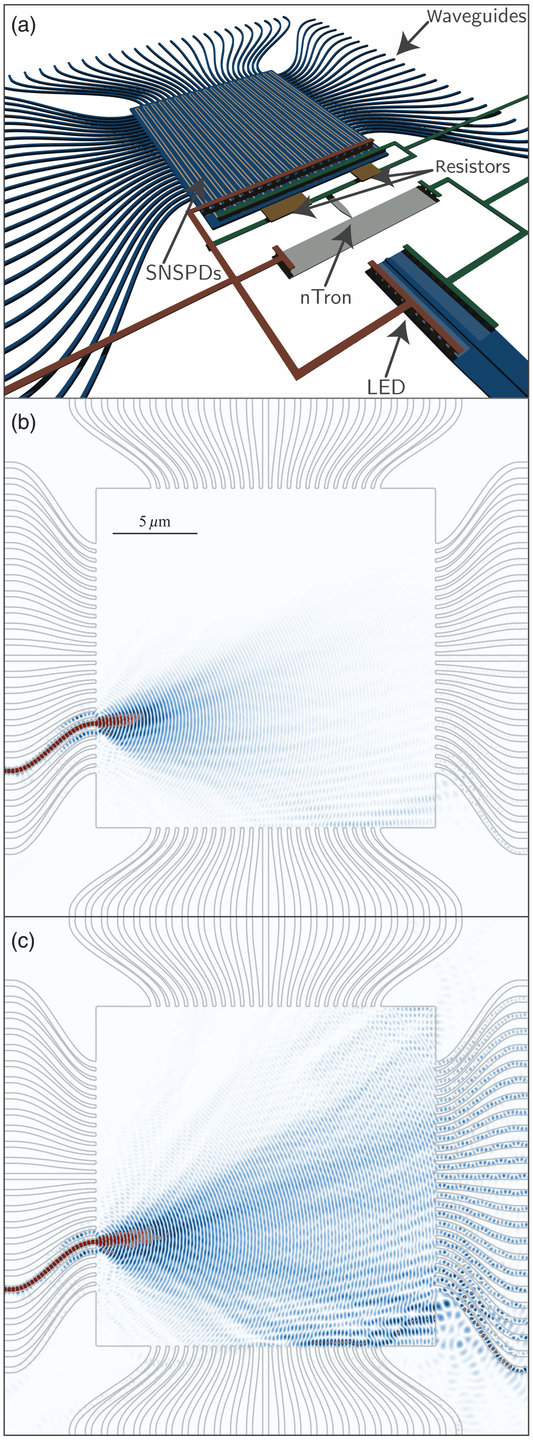

FIG. 11. (a) Schematic overview of the stingray neuron. (b) FDTD simulation of the dendritic arbor for the stingray neuron with SNSPDs present to absorb the light and (c) without SNSPDs present.

A design with 204 input waveguides and less than 1-dB insertion loss with a footprint of $60 \times 60\ \mu\mathrm{m}^2$ is also found. For larger numbers of inputs, the simulations become cumbersome. Yet, scaling to larger systems is clearly possible.

For threshold-based computation, processing units with large numbers of connections are advantageous [16,95]. Biological systems achieve massive interconnectivity with 3D branching networks and dedicated wires for each connection. To achieve this level of massive interconnectivity, we propose the use of multilayer photonics. Recent work has demonstrated the utility of low-temperature-deposited dielectrics [46,96] and superconductors [46] for scalable integrated photonics. For future massive scaling, we propose the use of waveguide routing networks and dendritic arbors spanning several—and possibly up to tens—of photonic and superconducting layers. A hybrid of the aforementioned spiderweb and stingray neuron designs can be implemented in which higher vertical-mode orders are utilized as well as higher lateral-mode orders, and massively multimode waveguides deliver their photon pulses to SNSPD receivers. These receivers can be implemented between waveguiding layers. At present, the technical challenge of building networks with processing units supporting tens to hundreds of connections is a serious one, so we mention the fully 3D multilayer photonic approach [97] to emphasize that this neuromorphic platform holds promise for scaling far into the technological future, but such sophisticated processing is not required to implement even very advanced systems with 2D interconnectivity supporting hundreds of high-bandwidth connections per unit.

B. The axon and its arborization

The output waveguide (axon) from a unit's LED must split into as many branches as there are connections to be made. While such a power splitter may seem to be the time-reversed case of the dendritic arbor, the initial conditions make this device significantly easier to implement. In the case of the dendritic arbor, one cannot assume the optical field will populate the arbor modes in a particular manner. Thus, while a power splitter can readily couple from a single-mode waveguide into many other single-mode waveguides, multiple single-mode waveguides cannot simply merge their power into a single-mode waveguide unless a particular distribution of power is present in the input waveguides. Such power splitters [98] can be made with a small footprint and low loss. It is straightforward to generalize such power splitters into the third dimension with multilayer photonics, and such an implementation will enable thousands of synapses with a volume of 10 $\mu\mathrm{m}^3$/synapse.

C. Learning, reconfiguration, and plasticity

An important aspect of any neuromorphic computing system is the ability to establish the strength of interaction between the connected units. These connection strengths, often referred to as the weight matrix, are important for memory and learning. This weight matrix determines how much of the light from the firing of a particular neuron is coupled into any other neuron, analogous to the synaptic strength between two neurons in a biological system.

As a first implementation, fixed connection weights are quite useful for many computing applications [9]. This can be readily accomplished by branching the output waveguide from one neuron and routing those waveguide branches to various downstream target-neuron input waveguides.

However, while fixed interaction weights are useful as a preliminary tool, one wants to develop a system in which the interaction strengths are variable. This is challenging at cryogenic temperatures, where modulators that rely on either the thermo-optic effect or free-carrier injection are ineffective, while electro-optic switches require too much space for this application. We propose the employment of electromechanically actuated waveguide couplers schematically depicted in Figs. 12(a) and 12(b). The amount of light coupled from one waveguide to the other is determined by the distance between them. These waveguides can be coupled either vertically [Fig. 12(a)] or laterally [Fig. 12(b)]. This distance can be controlled electromechanically, and anywhere from 0% to 100% of the light can be coupled from one waveguide to the other. The minimum coupling will be set in hardware, as the gap at 0 V is the maximum. Any applied voltage (positive or negative) produces an attractive force between the two waveguides. We then want activity within the circuits to build up voltage between the waveguides and increase the strength of the synapse. Such couplers have recently been demonstrated [99] in a highly scaled configuration. In Ref. [99], 4096 such switches were operated with >60-dB

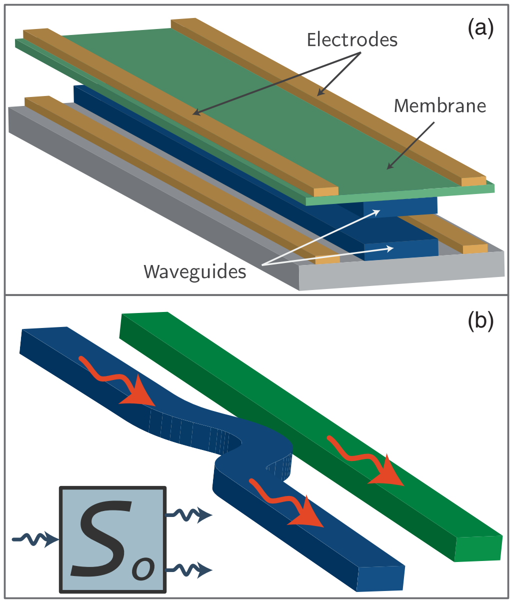

FIG. 12. Photonic synapses with electromechanically tunable coupling. (a) Interplane waveguide coupler. (b) Lateral waveguide coupler. The inset shows an abstract representation of the synaptic circuit element used in subsequent network diagrams.

extinction ratio and actuation voltage of 40 V. Because of the relaxed visibility requirements for this application, we expect much lower voltages will suffice.

To assess the utility of such synapses for neuromorphic computing, one must further specify the target application. To this end, we separate potential applications into two classes, which we refer to as supervised and unsupervised systems. For supervised systems, an input stimulus is injected into the system, the output is recorded, and the weight matrix is updated through a training algorithm to improve the output relative to a target. For such an application, one anticipates using control electronics to interface with the neuromorphic system, and arbitrary voltages can be applied to the various synaptic elements.

For more highly scaled implementations emulating the behavior of biological organisms, we turn our attention to unsupervised systems. Here it is important that each synapse be as small as possible to enable massive scaling, but it is also important that voltages be modest, as we want the activity in the circuits to be capable of reconfiguring the synapses. In particular, we want firing events from upstream neurons followed closely by firing events by downstream neurons to place charge on this MEMS capacitor (waveguide coupler) and thereby decrease the gap between the two waveguides and increase the optical coupling and, therefore, the synaptic strength. This coordinated charging of the membrane will accomplish spike-timing-dependent plasticity, an important learning and memory reinforcement mechanism in biological neural systems. In this mode of operation, we envision eliminating external control circuits and achieving the capacitor charging using integrated superconducting circuits to distribute current based on photon-absorption events. The storage of charge on a capacitor required for this device operation is very similar to dynamic random-access memory (DRAM), which is a mature technology. While implementing what is essentially spike-timing-dependent DRAM with suspended waveguide membranes presents a technical challenge, it offers a promising means to implement truly neuromorphic learning within this optoelectronic platform.

While the size of mechanical waveguide couplers and the voltages required for their operation are commensurate with the requirements for scaling this technology, an implementation of variable synaptic weights which does not rely on mechanically mobile components will be advantageous. It may be possible to implement synapses in the electronic domain by making use of superconducting circuit elements or magnetic elements such as magnetic tunnel junctions or magnetic Josephson junctions [23]. Such an approach to memory will be investigated in future work. Additionally, we note that a variable weight can be achieved with a tunable Mach-Zehnder interferometer. However, the size of such devices makes them poorly suited to highly scaled systems.

IV. NETWORKS AND SCALING

We have now discussed neural circuits based on optical signaling. We have discussed various means to connect these optical and electrical signals in a time-varying manner with event-based plasticity. In Fig. 13, we again show the inhibitory SPON of Sec. II C and introduce an abstract symbol to represent the circuit labeled N_o, which is used in the following sections as an element in networks. We refer to networks comprising interconnected SPONs as superconducting optoelectronic networks (SOENs). In this and the following schematics, we represent electrical inputs and outputs as black arrows running vertically and optical inputs and outputs as colored wavy arrows running horizontally. In Fig. 13, we emphasize that the optical processing unit can receive and transmit electrical and optical signals each in two ports. The electrical signals affect SPON threshold and gain, while the optical ports are either excitatory or inhibitory. This full functionality need not be employed, and as few as one optical input and output and one electrical input can be utilized.

We now illustrate how the circuits presented in Sec. II may be put to use in systems by considering the canonical example of the multilayer perceptron (MLP) in Sec. IV A. This leads us into a more general discussion of SOEN scaling in Sec. IV B.

A. Multilayer perceptron

Perhaps the most studied implementation of neural networks is the MLP [49] and its contemporary counterpart, the convolutional neural network (CNN) [100]. Our consideration of the MLP provides insight into other applications of this platform in terms of important quantities such as speed, size, and dynamic range.

Generally speaking, the MLP consists of a number of inputs incident on a weight matrix (array of synapses) which feeds into a layer of neurons. The output of this layer of neurons projects to at least one more layer of weights and

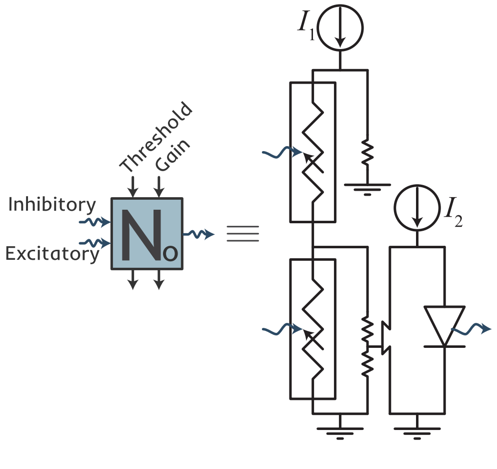

FIG. 13. Abstract symbol definition for general neuron with inhibition and gain.

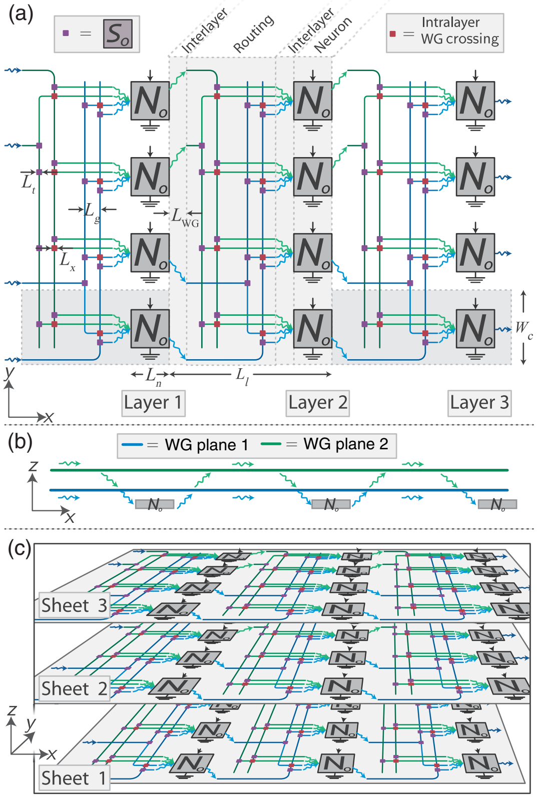

FIG. 14. (a) Schematic of the MLP implemented with the SOEN platform. (b) Cross section in the x-z plane. (c) Three-dimensional schematic of stacked die. (a) illustrates layers of neurons in the network, (b) illustrates planes of routing waveguides, and (c) illustrates sheets of stacked die.

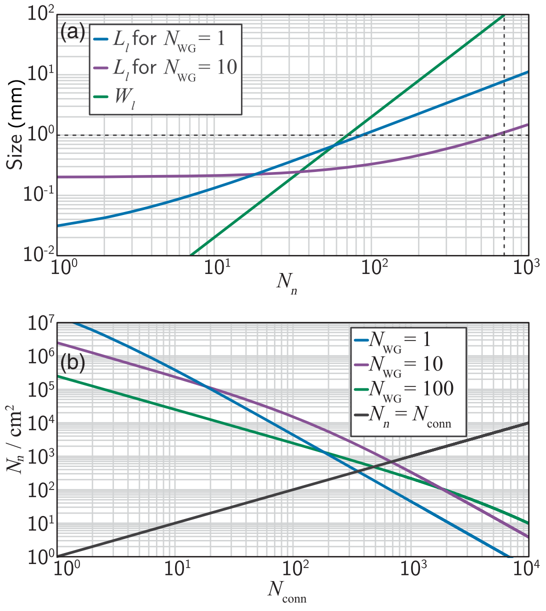

FIG. 15. (a) Length and width of layer versus number of neurons in a layer assuming each neuron in a given layer is connected to each neuron in the next layer. (b) Number of neurons per centimeter squared versus the number of connections per neuron.

neurons, and often several, before being output from the system. In Fig. 14(a), we show a schematic diagram of how such a MLP is likely to be implemented. Such a MLP can be achieved with a single plane of routing waveguides or many such planes. Here, we use “plane” to refer to vertically stacked dielectric layers to avoid confusion with the processing layers of the MLP progressing horizontally in Fig. 14(a). The processing layers of the MLP are labeled in Fig. 14(a), and the cross-sectional view of planes of routing waveguides is shown in Fig. 14(b). Stacked sheets of die are illustrated in Fig. 14(c).

Several factors determine the functionality of a MLP. These include the dynamic range of the inputs, the speed with which the inputs can be received, the bit depth of the synaptic weights, and the speed with which the weights can be reconfigured. From Fig. 6(c), we see that for 0.7 I_c, the response turns on at around 500 photons, and it roughly levels out by 3000 photons. For this case, the dynamic range of the inputs is, therefore, $\log_2(2500) \approx 11$ bits. The speed with which inputs can be received is limited by the device reset time of 50 ns, so a 20-MHz input rate is achievable. The bit depth of the weights depends on the number of discrete values of coupling achievable between the two waveguides involved in a synapse, and further investigation is required to report a valid estimate for this number. The speed with which the weights can be changed is at least 1 MHz [99].

The number of inputs, the number of connections per neuron, and the number of MLP layers all affect the size and complexity of MLP that can be fabricated on a given die. In Figs. 15(a) and 15(b), we consider a model of these factors to estimate what may be achieved with reasonable size. Figure 15(a) assesses the length L_l and width W_l of a single MLP layer, as given by Eq. (F1) as a function of the number of neurons in a MLP layer N_n for two different values of the number of vertically stacked waveguide planes N_{WG}. The model assumes a feedforward configuration wherein every neuron in a given MLP layer is connected to every neuron in the next MLP layer with a variable-weight connection. The total width of a MLP layer is also plotted. See Appendix F for more information. If we assume that a 10×10 cm^2 die is the largest we want to fabricate, we find the width limits the number of connections per neurons to 700, and we are, thus, considering MLP layers with 700 inputs and 700 neurons per layer. For the case with $N_{\mathrm{WG}} = 10$, the length of a MLP layer with 700 connections per neuron is 1 mm. We can, therefore, fit 100 such MLP layers on the 10×10 cm^2 die. The total number of neurons is 70 000. A MLP or CNN with 700 inputs, 700 connections per neuron, and 100 layers receiving inputs at 20 MHz with weight reconfiguration speed of 1 MHz is a very powerful tool. While it is not necessarily optimal to work with a neural network of 100 layers, as shallower networks are advantageous for several reasons [95], we present this model to quantify SOEN spatial scaling keeping in mind that network depth can be traded for a larger number of inputs or larger connectivity. As a point of comparison, the recent demonstration of a computer defeating the world champion Go player input the state of the board as a 19 × 19 matrix (361 inputs) to the 13-layer-deep neural network [10]. The bit depth of the synapses proposed here is unlikely to reach the 32 bits utilized in software implementations running on modern graphics processing units, but there are likely many applications in which such a constraint is minor compared to the system advantages of speed, complexity, and connectivity.

B. Scaling

To further pursue this discussion of the scaling of the MLP (or other similar neuromorphic computing systems), we consider the number of neurons in an area of 1 cm^2 versus the number of connections per neuron, N_{conn}. Figure 15(b) shows the results of the model of Eq. (F1) for $N_{\mathrm{WG}} = 1$, 10, and 100. If $N_{\mathrm{conn}} = 10$ is sufficient for a given application, we can achieve a neuron density of 400 000 neurons per centimeter squared. Because of the size of the interlayer couplers, this is achieved more compactly with $N_{\mathrm{WG}} = 1$ than with $N_{\mathrm{WG}} = 10$. For N_{conn} in the range of 100 to 1000, it becomes advantageous to utilize $N_{\mathrm{WG}} = 10$. For $N_{\mathrm{conn}} = 100$, over 10 000 neurons will fit within a centimeter squared, and for $N_{\mathrm{conn}} = 1000$, 300 neurons fit within a centimeter squared. It does not become advantageous to use $N_{\mathrm{WG}} = 100$ until $N_{\mathrm{conn}} = 2000$, and even then the gain is modest. To achieve 10 000 connections per neuron (comparable to a mammalian brain), only a few devices fit within a centimeter squared (given the present model), and we are left in awe of the massive interconnectivity and scaling achieved by the bottom-up nanofabrication of biological organisms.

While the scaling to 10 000 connections per neuron is formidable, the range of $N_{\mathrm{conn}} = 100$–1000 is promising and technologically consequential. As is the case for scaling CMOS neuromorphic platforms, utilization of die tiling [9] plays a crucial role in this technology. For this purpose, the SOEN platform is in an excellent position. Die can be tiled in 2D with several types of connectivity to adjacent die including electrical, single-flux-quantum, and photonic communication over interdie bridge waveguides. Additionally, tiling in the third dimension is possible with the usual bump-bonding approach for electrical connectivity as well as with free-space optical signals sent from one chip using vertical grating couplers and received by a chip above or below using SNSPD arrays [101]. Information over such links can be encoded temporally, spatially, or in frequency with forgiving alignment tolerances. From Fig. 15(b), we find that 700 neurons with 700 connections per neuron can fit on a 1 × 1 cm^2 die if ten waveguiding planes are utilized.

To analyze long-term scaling, we consider a system on the scale of the human brain. To this end, we envision tiling a 215 × 215 array of these die in a sheet to build a system with 32×10^6 neurons. Such a sheet will be approximately 1 mm thick. To achieve the scale of the brain, 2150 such sheets need to be stacked with intersheet coupling to construct a cube 2.15 m on a side and with a total volume of 10 m^3. The system then comprises 7×10^{10} neurons or roughly the number contained in the human brain.

To achieve such a system, we envision sheets of die mounted in trays with in-plane fiber-optic connections leaving from the perimeter of the trays and out-of-plane free-space grating-to-SNSPD interconnects, thus, enabling the trays to slide laterally. Achieving intersheet connectivity without physical bonds enables access to die within the volume of the cube for diagnostics, repair, and local iteration and evolution. Massive interconnectivity between neurons on different die can be accomplished using such grating interconnects [89–93].

Of greater importance than the size of highly scaled systems is the power consumption. We again consider a system of SPONs with 700 connections each. Such a device consumes 2×10^{-17} J/synapse event, and with 700 connections, each firing event consists of 700 synapse events. Information processing in neuromorphic systems requires sparse event rates, so for the SOEN hardware wherein 20 MHz is achievable based on device limitations, 20 kHz represents a sparse rate. Note that this rate is a factor of (2×10^4)–(2×10^5) faster than biological event rates and a factor of 1000 faster than the CMOS demonstration which achieved 26 pJ/synapse event and was limited by time multiplexing [9]. For the system under consideration, we have 7×10^{10} processing units which we consider to be firing at this rate with this energy per firing event, giving a total device power consumption of 20 W. These numbers give 5×10^{16} synapse events per second per watt. The system must be kept around 2 K, so we also include an additional 1 kW of cooling power per watt of device power, as we discuss in Sec. II F. While this cooling power does not affect the power density (which ultimately limits scaling), and this 20 kW is minuscule compared to the tens of megawatts of a modern supercomputer, if we include this additional power in the calculation, we find that we achieve 5×10^{13} synapse events per second per watt.

To put this in perspective, the human brain also uses 20 W of device power, but by analogy to the inclusion of the cooling power in the above calculation, one must include the human's total power of 100 W which is necessary to sustain the brain's operational state. The brain has roughly 10^{11} neurons with roughly 7×10^3 synapses per neuron firing between 0.1 and 1 Hz [54,102–104]. For the purposes of this calculation, we generously assume the rate is 1 Hz. This equates to 7×10^{12} synapse events per second per watt. Even with the 1-kW/W cooling power of the cryostat, we find that the number of synapse events per second per watt of the SOEN system exceeds that of the brain by an order of magnitude. The size of the SOEN system (10 m^3, 2.15 m on a side) is, however, much larger than the biological brain.

Importantly, because signaling occurs predominantly in the optical domain, firing events can be directly imaged with a camera. For massively scaled systems, this direct optical imaging becomes a powerful metrological tool. Such a measurement technique can be used to monitor device and system performance across spatial and temporal scales in a manner analogous to functional magnetic resonance imaging of biological organisms.

To close this discussion of scaling, we address the cryogenic requirements of a 1-m^3 SOEN system. We seek a ^{4}He sorption refrigerator capable of cooling a 1 m^3 volume to 2 K with 20 W of cooling power. While this is a relatively large cryostat, it is certainly well within the realm of possibility. No new physical principles of operation need to be developed; it is simply a question of scaling up existing ^{4}He cryogenic systems. Additionally, if suitable SNSPD materials can be found which operate at 4 K with high yield, 20 W of cooling power is straightforward to achieve. We are of the opinion that with the advancement of single-flux-quantum processors, superconducting qubit devices, and SOENs, large-scale cryogenic technology will advance significantly in the coming years. Presently, many conversations in advanced computing debate whether the technology which proves victorious will operate within a cryostat or at room temperature. We speculate that a supercomputer of the future will leverage optoelectronic devices on various material platforms to employ quantum principles, neuromorphic principles, and digital logic principles across various temperature stages. The device designer is faced with the task of optimizing hardware performance at each temperature stage, and the architect is liberated to dream with von Neumann [5] far beyond the architecture that now bears his name.

V. DISCUSSION AND OUTLOOK

We have explained the proposed devices and their functions, analyzed their performance, and considered their scaling. Here we consider possibilities for utilization of this platform for neuromorphic applications.

A. Advantages of optoelectronic neural networks

The unparalleled performance of the brain emerges from the enormous number of connections between neurons and the numerous complex signaling mechanisms available to the neurons. Optical signaling has an advantage over electronics in terms of the ability to route noninteracting signals in three dimensions without wiring parasitics. These strengths have been recognized for many years, and early implementations utilized reconfigurable holographic gratings [105,106] for forming connections between optoelectronic neurons [107]. We envision utilizing multilayer waveguides [97] in a deposited-photonics process [46] to implement photonic fanout [98] and routing from each neuron to its many downstream connections.

The two key components to enable photonic fanout and routing at an intradie level are multilayer waveguide power dividers and in-plane waveguide crossings. Both of these devices occupy a small area and operate with low loss and no *RC* penalty. As we mention in Sec. IV B, implementing these devices with roughly ten waveguiding layers appears optimal, comparable to the number of back-end-of-line metal layers used in CMOS for interconnect. With ten waveguiding layers, the desired routing between optoelectronic neurons still requires in-plane waveguide crossings [108]. The ability to implement multilayer power dividers and in-plane waveguide crossings with low loss and low cross talk allows dedicated communication lines for each interneuron connection.

On the receiving end, signals from an arbitrary number of SPONs can be received simultaneously, and time multiplexing is unnecessary. The system is conducive to encoding of information in both spike rate and timing. On an electronic platform, the length of an electronic signal line increases as the number of connections grows, resulting in a larger *RC* time constant. This increase in *RC* time constant with number of connections forces a speed or connectivity trade-off, leading most electronic neuromorphic implementations to share communication lines. Such a shared interconnect can transmit only a single voltage pulse within a time window, and this limits both the number of connections between neurons and the firing rate of each neuron.

Other approaches that leverage phenomena unique to optics for neuromorphic computing [109–114] have employed optical devices such as lasers and integrated microresonators. Laser cavities with strong light-matter interaction can be leveraged to realize complex nonlinear dynamics which can emulate the behavior of neurons [109,111,114]. The frequency selectivity of integrated ring resonators can be used to achieve synaptic weights [112]. Optical neural networks [113,115] and spiking neurons [111,116–119] based on these effects have been proposed and demonstrated. Optical reservoir computing has also recently been demonstrated [120–122] as another way in which inherently optical phenomena can be leveraged for advanced computing. The distinction of the proposed SOEN platform is that it operates in the few-photon regime with compact, energy-efficient components, enabling a large degree of scalability. Thus, at present, many electronic and photonic technologies appear promising for neuromorphic computing, and the most suitable hardware platform is likely to depend on the application.

B. The visual cortex

While we describe in detail in Sec. IV A how a simple neural network (the MLP) can be built with SPONs, the

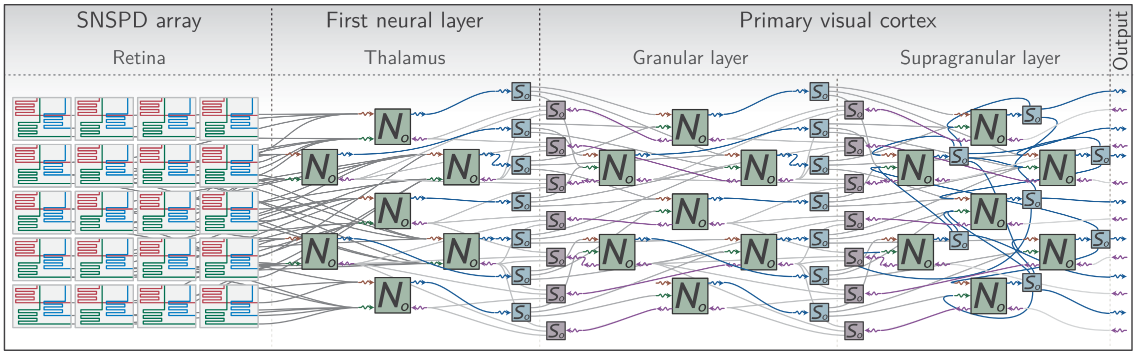

FIG. 16. Schematic of a SOEN model of the mammalian visual cortex.

potential of the SOEN platform for more complex systems should not be overlooked. The visual cortex is the most thoroughly studied region of the mammalian brain [123], yet there is still a great deal to be understood about information encoding from the retina through the thalamus and on to the visual cortex. A nonbiological experimental test bed is highly desirable to explore hypotheses [8,124]. Biologically realistic supercomputer simulations of the brain can simulate only a small fraction of the brain cells in a small mammal at significantly reduced speed [125,126]. The massive parallelism enabled by a scalable, biologically realistic hardware implementation of the many thousands of neurons involved in the visual system can provide more quick and efficient simulations [18,126,127], which may give further insight into the visual system, while also offering potential for image-processing applications.

We propose a hardware platform with the potential for a built-in retina, manifest as integrated SNSPDs, which can be used in pixel arrays [101] for monolithic image acquisition and analysis. In Fig. 16, we show a schematic of how such a SNSPD array can be integrated with a multilayer neural network to emulate the visual system. To illustrate the key points of such an experimental system, we break the visual system into three parts: the retina, the thalamus, and the primary visual cortex. In biological systems, the primary visual cortex is highly sophisticated, being organized into six layers each with their own sublayers [123]. For the purpose at hand, we treat the primary visual cortex as being composed of two layers referred to as the granular layer and the supragranular layer.

At the left of Fig. 16, the SNSPD array receives light from the environment and converts it to signals to be sent to the first layer of neurons in the thalamus, in direct analogy with a biological retina. Much like the cones in one’s eye, the pixels of the SNSPD array can be designed to be more sensitive to particular frequencies simply by varying the thickness of an antireflection coating locally above each pixel.

From the retina, a small number of pixels project to each neuron in the thalamus without a large amount of branching. Similarly, the neurons of the thalamus project to the first layer of the visual cortex with minimal branching. Importantly, some of these connections are inhibitory and some are excitatory. While inhibitory connections are known to play a central role in information encoding in the visual system, the full scope of that role remains the subject of investigation. The biologically realistic mechanism for implementing inhibitory connections, as illustrated in Fig. 7(c) is of great utility in using SOENs to study information encoding in the visual system. In the thalamus, there is little if any recurrence, meaning the neurons in that layer project forward but do not form synapses on each other. The thalamic neurons do, however, receive feedback from the granular layer of the visual cortex. The ability to straightforwardly implement feedback with SOENs, as illustrated in Fig. 7(e), is another feature of great utility in using SOENs to model the visual system.

The granular layer receives feedforward signals from the thalamus, projects feedforward signals to the supragranular layer, and receives feedback from the supragranular layer. While still only minimally recurrent, neurons in the granular layer branch more heavily to form a larger number of connections across more neurons in the supragranular layer. The supragranular layer projects its output to other regions of the cortex and is also heavily recurrent. At the right of Fig. 16, we show the neurons in the supragranular layer making connections with other neurons within the layer.

For an initial SOEN visual system, we envision implementing the retina and thalamus on a single die, with a separate chip of 700 neurons being employed for the granular layer and a third chip of 700 mutually interacting neurons representing the supragranular layer. This experimental test bed may offer insight into outstanding questions such as how and why concentric circular patterns of retinal response are mapped to bars for processing in the visual cortex. With a simple system like that illustrated in Fig. 16, it will be possible to conduct experiments related to object recognition, edge detection, the perception of motion and spatial frequency, as well as many other subjects in contemporary visual system research.

C. High-performance application spaces

One strength of neuromorphic systems is their ability to find trends and extract features from large and noisy data sets, thereby reducing the dimensionality of those data sets [128]. They can learn over time based on the temporal evolution of the data under consideration. Several societal challenges require this type of analysis of large numbers of complex, interacting units—exactly the type of system for which neuromorphic computing excels. These applications include monitoring of markets, Internet traffic metrology, detection of hacking attacks, modeling of climate systems, and phenotypic prediction from genomic data. For these applications, supercomputers at the limit of what is possible with CMOS implementations of the von Neumann architecture are presently in use. Yet, greater performance is still required. For the most demanding computational tasks of this class, massively scaled systems employing parallel computation in a neuromorphic architecture are likely to play a central role. It is for these applications that we envision the SOEN platform making the largest impact.

Another likely solution to the current bottlenecks facing supercomputers is superconducting electronics. In particular, Josephson-junction processors with single-flux-quantum logic are poised for use in the next generation of supercomputers. These processors can provide an improvement over CMOS in speed by roughly a factor of 100 with extremely high-energy efficiency. Our proposed platform will integrate well into such supercomputers, offering neuromorphic capability to von Neumann implementations [129] and additional degrees of freedom to neuromorphic Josephson-junction systems [20,21,130], which are purely electronic. In addition, the SOEN platform may offer a means to transduce single-flux quantum pulses to the optical domain, for interconnects between chips and with the outside world (cryostat I/O) via photonic signaling.

D. Summary

We describe a hardware platform combining superconducting single-photon detectors and electronics with semiconducting faint-photon sources to operate as a massively interconnected information-processing system. The SOEN platform consists of neurons that exhibit complex signaling and efficient access to photonic degrees of freedom such as frequency, polarization, mode index, intensity, and coherence, in analogy to the complex signaling mechanisms in the brain. The proposed networks of connections based on reconfigurable waveguides offer advantages over electronic connections in terms of speed, connectivity, and energy efficiency.

In the present paper, we argue that through the use of networks of neurons consisting of semiconductor LEDs, superconducting-nanowire single-photon detectors, and reconfigurable optical waveguides, we can build advanced computing systems. Such networks can achieve states of enormous entropy through massive interconnectivity and the interaction of multiple physical degrees of freedom. We further show that the integrate-and-fire operation of superconducting optoelectronic neurons can be used for spike-encoding information. Such spike-encoded information is highly advantageous for high-bandwidth information processing with temporal information encoding and resilience to noise. These concepts have recently been placed on a solid theoretical foundation [11–13,16], so we should not be surprised to find that the brain's computing mechanisms employ all of these concepts. The fundamental principles of information theory which enable reasoning, decision, innovation, and consciousness are currently incompletely understood. To date, we know of one computing platform which can accomplish these tasks: the biologically evolved neural system. We do not appear to be close to a complete understanding of the information theory describing such a complex system. Yet, by exploring alternative physical systems with comparable complexity, we stand to learn a great deal about the fundamentals of information science.

ACKNOWLEDGMENTS

We wish to thank Professor Ryan Tung, Dr. Amanda Casale, Dr. Aaron Hagerstrom, Dr. Ann Hermundstad, Dr. Matthew Phillips, Professor Gert Cauwenberghs, and Ryan Uhlenbrock for fruitful discussions.

APPENDIX A: INTEGRATION TIME AND REFRACTORY PERIOD

The integration time of a SPON is the time from the absorption of a photon until the receiver no longer has a memory of that absorption event. The behavior of integrate-and-fire devices with integration times less than infinity are referred to as leaky integrate-and-fire neurons. In the context of SPON devices, in the most basic case, this integration time is determined by the hot-spot relaxation time of the superconductor, which depends on the material quasiparticle dynamics which are governed by the electron-phonon coupling and the thermal conduction to the substrate. This thermal relaxation is a material-dependent quantity and can be as fast as 200 ps in NbN [52]. In WSi, it is closer to 1 ns [52], and there may be materials for which it is even slower. Additionally, the bias current is shown to have a significant effect on the quasiparticle recombination time [52]. Therefore, the choice of superconducting material and substrate may be leveraged to tune the integration time to a desired value in hardware, and the bias current may be used to modify it dynamically.

Further, the PND circuit shown in Fig. 2 can be modified so that each wire in the PND array is in parallel with a small shunt resistor. In this configuration, the L/R time constant of each receiving wire can be chosen to set the integration time. In this case, the hot-spot relaxation time represents a lower limit on the integration time, but the integration time can be extended to very long times relative to other time scales of the system simply by adjusting the L/R value.

Recent studies [131,132] reveal that nonuniform current distribution in the PND as drawn in Fig. 2(a) is problematic for number-resolving photon detection. To avoid this, the cylindrically symmetric nanowire arrays of Figs. 4 and 19 are proposed. In this geometry, no nanowire occupies an edge, so supercurrent is evenly distributed after each firing event. Also shown in Ref. [132] is the fact that a PND can trap flux after a photon-absorption event. To utilize this to extend the integration time to infinity, the geometry of Fig. 4 is proposed. If one wishes to dissipate flux to reduce the integration time, the topological variant of Fig. 19 is proposed. The differing circuit designs of these two devices are shown in Fig. 20. In the flux-dissipating configuration shown in Figs. 19 and 20(b), flux-trapping superconducting loops are avoided, and all locations where hot spots can be created are on a boundary with the normal environment. Therefore, vortices created during absorption events are not trapped.

We note that in biological systems, the integration time is set by the RC time constant of the membrane and is typically approximately 1 ms or approximately 10^{-4}–10^{-5} the firing period. Taking the 1-ns quasiparticle lifetime as the integration time, this corresponds to operating the system with (10–100)-kHZ event rates, a range that is straightforward to achieve.

The refractory period of a neuron refers to the time following a firing event during which the neuron cannot fire again. For a standard SNSPD, this dead time is governed by the L/R time constant of the series inductance of the SNSPD and the resistance across which the voltage pulse is being measured. In the case of WSi, this L/R time constant is usually 50 ns [133]. This resistance is usually 50 Ω, but in the present case, it is the impedance of the LED, which will be several kilohms, giving a shorter refractory period. If an application requires a longer refractory period, an additional series inductance can be added to achieve the desired delay. We note that in some SNSPD material systems, the L/R time constant must be chosen sufficiently large to avoid latching, while in the present application, the feedback circuit of Fig. 7(e) can also be utilized to avoid latching and control the refractory period.

APPENDIX B: THRESHOLD CONDITION FOR THE PND ARRAY

Here we derive the expression of Eq. (1). We begin by defining all the quantities of interest. The number of nanowires in the PND array is denoted by N_{NW}. The number of nanowires driven normal by photons is denoted by n^{abs}. The critical number of nanowires driven normal is denoted by n_c^{abs}. The bias current through the entire array is denoted by I_b. The current through a single wire of the array is denoted by i. The critical current of a single wire is denoted by i_c.

In the steady state, before any photons are absorbed, $n^{\mathrm{abs}} = 0$, and $i = I_b/N_{\mathrm{NW}}$. Upon absorption of a single photon, $n^{\mathrm{abs}} = 1$ and $i = I_b/(N_{\mathrm{NW}} - 1)$. In the general case that n nanowires are driven normal by photons, $n^{\mathrm{abs}} = n$ and $i = I_b/(N_{\mathrm{NW}} - n)$.

The condition for n_c^{abs} is $i = i_c = I_b/(N_{\mathrm{NW}} - n_c^{\mathrm{abs}})$. Rearranging gives $n_c^{\mathrm{abs}} = N_{\mathrm{NW}} - (I_b/i_c)$.

APPENDIX C: INTEGRATION OF SUPERCONDUCTING-NANOWIRE DETECTORS

To properly understand the behavior of the SNSPD receivers, we must analyze the optical absorption and statistical behavior of waveguide-integrated SNSPDs [40–46]. We first calculate the attenuation of light as a function of propagation length for 200-nm-thick waveguides (t_{WG}) in the asymptotic slab regime. The waveguide refractive index is 3.52, the cladding index is 1.46, and our calculations are at a wavelength of 1220 nm. The nanowire is assumed to be 4 nm thick, 300 nm wide with a 50% fill factor, and $n = 3.25 + 2.19i$. In Fig. 17, we show the

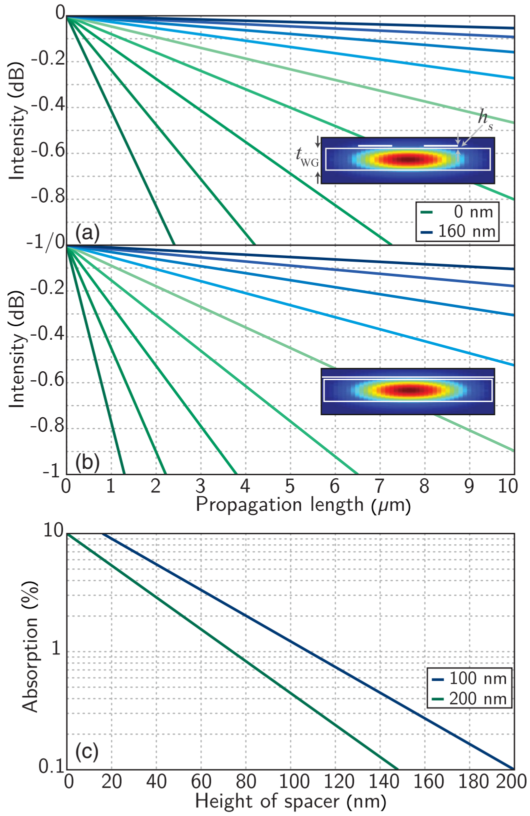

FIG. 17. Absorption of light propagating in a waveguide with SNSPD on top in (a) parallel and (b) perpendicular configurations, for different spacer heights between the SNSPD and waveguide. (c) Absorption in waveguides of different thicknesses for different spacer heights.

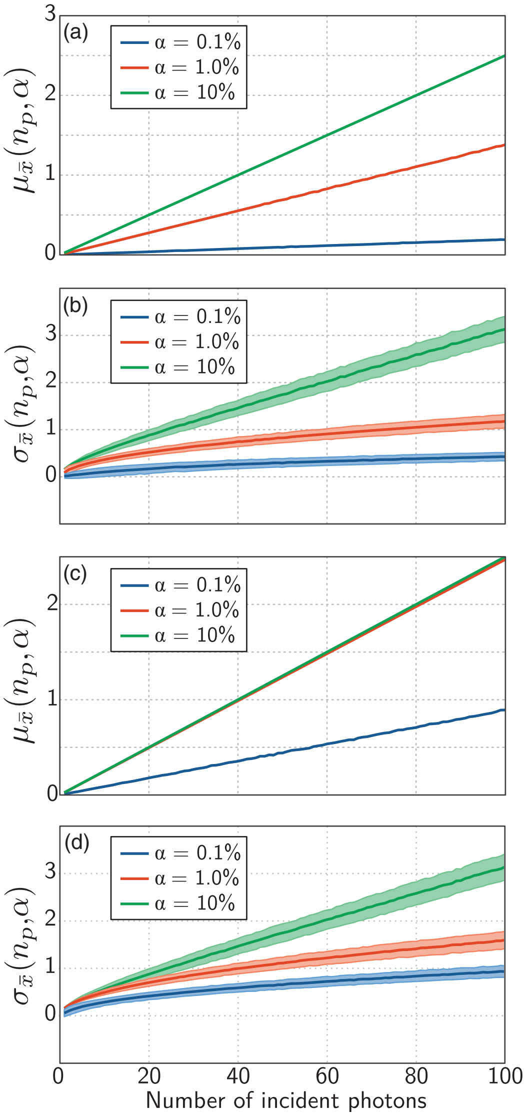

FIG. 18. Mean number (a) and standard deviation (b) of absorbed photons versus number of incident photons for neuron designs where light is directed past each nanowire once (single pass). Mean number (c) and standard deviation (d) of absorbed photons versus number of incident photons for neuron designs where light is directed past each nanowire ten times.

results for the common out-and-back configuration [light propagating parallel to the nanowire, Fig. 17(a)] and the slab configuration [light propagating perpendicular to the nanowire, Fig. 17(b)]. In each case, the various traces are for different spacer thicknesses, (h_s, refractive index 1.46) between the waveguide and nanowire, ranging from zero to 160 nm in steps of 20 nm. The modal distribution is shown in the inset. The data in Figs. 17(a) and 17(b) are fractal in nature, so an increase of the x axis by one decade is accompanied by an increase in the y axis by a decade (on the decibel scale). From these plots, one can see that for both the parallel and perpendicular configurations, a wide range of attenuation coefficients can be achieved.

In Fig. 17(c), we show the probability of absorption after a single pass by a nanowire as a function of spacer thickness for waveguides with 100 and 200 nm thickness, illustrating another degree of freedom for tuning the absorption. It is important to be able to engineer the statistical distribution of absorption across the SNSPD receiver. For the case of the PND, each SNSPD should absorb an average of one photon each, as an additional photon absorption in the same SNSPD will not contribute to the spike event. For the case of the SND, the requirement is less stringent, but one still wants to spatially distribute absorption events so that hot spots do not overlap until a certain (large) number of photons is absorbed.

To address the design requirements of the PND, we consider the absorption statistics as calculated via Monte Carlo simulations. We perform 1000 trials each for different photon numbers incident on a PND with 40 SNSPDs. Figure 18(a) shows the mean number of photons absorbed (out of 1000 trials) in the PND as a function of the number of incident photons for different absorption probabilities, in the case where only a single pass by each nanowire occurs. This behavior may be achieved with a design like that of Fig. 11. For each of the 1000 simulations, the arithmetic mean of the number of photons absorbed per nanowire is calculated for each value of incident photon number as

$$\mu_x(n_p,\alpha) = \frac{1}{N_{\mathrm{NW}}}\sum_{i=1}^{N_{\mathrm{NW}}} x_i, \tag{C1}$$

where x_i is the number of photons absorbed in the ith nanowire. From these values, the mean number of photons absorbed per nanowire $\mu_{\bar{x}}$ is then calculated as the mean of the means (grand mean) in Eq. (C1).

One then engineers the absorption probability in the PND such that the mean number of absorbed photons per nanowire per pulse and the standard deviation of this number are both less than or equal to 1. In Fig. 18(b), we show the standard deviation data for the single-pass case. For each of the 1000 trials, the standard deviation of the number of absorbed photons is calculated as

$$\sigma_x(n_p,\alpha) = \sqrt{\frac{1}{N_{\mathrm{NW}}}\sum_{i=1}^{N_{\mathrm{NW}}}(x_i-\mu_x)^2}, \tag{C2}$$

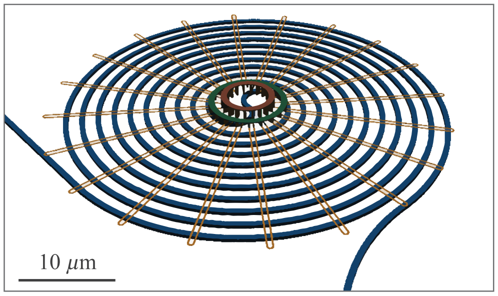

FIG. 19. Flux-dissipating version of the spiderweb neuron.

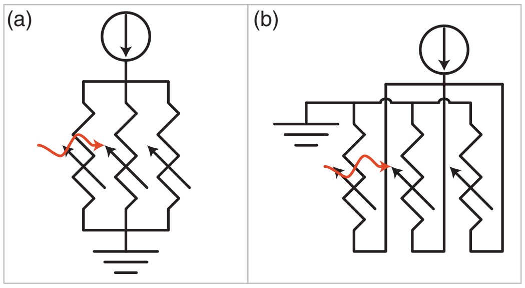

FIG. 20. (a) Flux-trapping PND circuit. (b) An alternative PND design which avoids flux trapping.

where μ_x is given by Eq. (C1). The mean of these standard deviations over the 1000 Monte Carlo trials ($\sigma_{\bar{x}}$) is calculated, as is the standard deviation of the standard deviations. The center trace of each curve in Fig. 18(b) is $\sigma_{\bar{x}}$ for a given value of α, and the width of the trace is calculated by adding and subtracting the standard deviation of the standard deviations. The standard deviation with $\alpha = 10\%$ is roughly three photons. Thus, such large absorption is undesirable for this purpose, as the initial wires tend to absorb more than a single photon, and the latter wires absorb zero photons. For the one-pass case, 1% absorption appears to be close to ideal. The mean number of absorbed photons is close to 1, as is the standard deviation. The standard deviation for $\alpha = 0.1\%$ is even lower, yet the mean number of absorbed photons is only approximately 0.2. Therefore, many photons are passing through the array without being absorbed.

In Figs. 18(c) and 18(d), we show results for the case where ten passes by each nanowire occur, as may be achieved with the spiderweb neuron design of Fig. 10. For the case of ten passes, $\alpha = 0.1\%$ performs much better, although all photons are still not absorbed.

Consider the case where 40 photons are incident. We want all 40 of these photons to be absorbed by the 40 nanowires of the array, and, therefore, we want $\mu_{\bar{x}}$ to be near unity. In Fig. 18(c), we see that we achieve this for both $\alpha = 1\%$ and 10%, yet in the case of $\alpha = 10\%$, all photons are absorbed on the first pass [as seen in Fig. 18(a)], so the mode of the distribution is greater than 1, and the standard deviation is larger than desired. By comparing the standard deviations for the $\alpha = 1\%$ and $\alpha = 0.1\%$ cases in Fig. 18(d), we find that $\alpha = 0.1\%$ gives a more desirable spread of absorption events (smaller standard deviation). From this analysis, we find that for the PND receiver array, it is desirable to operate with low α and a high number of passes.

APPENDIX D: *p-n* JUNCTION MODEL OF THE LIGHT-EMITTING DIODE

To model the performance of the emitters, we work with an analytical model of a *p-n* junction [134]. Within this model, the current-voltage relationship for the junction is given by

$$I_{p-n}(V) = eA\left(\sqrt{\frac{D_p}{\tau_p}}p_n + \sqrt{\frac{D_n}{\tau_n}}n_p\right)(e^{eV/k_BT} - 1). \quad \text{(D1)}$$

In Eq. (D1), the electron and hole diffusion coefficients are given by $D_n = \mu_n(kT/e)$ and $D_p = \mu_p(kT/e)$, where μ_n (μ_p) is the mobility of electrons (holes). The electron and hole lifetimes are given by τ_n and τ_p, respectively, which we take to be 40 ns. n_p is the concentration of electrons on the p-doped side of the junction, and p_n is the concentration of holes on the n-doped side of the junction. To achieve low-temperature operation, we assume degenerate doping, and, therefore, a low mobility is to be expected. We use a value of 100 $\mathrm{cm}^2/(\mathrm{V\,s})$ [135] for both electron and hole mobilities. Because this value will be limited by ionized impurity scattering, it is likely to change little as the temperature decreases to 1 K.

From the electronic current, we calculate the photonic current as

$$I_\nu(V) = \eta\frac{I_{p\text{-}n}(V)}{e}. \quad \text{(D2)}$$

This model for the current through the diode is derived for an abrupt *p-n* junction, yet for the waveguide-integrated LED, one employs a *p-i-n* junction. Also, the present model breaks down at low temperature. We use $T = 300$ K in Eq. (D1) because our measurements inform us that in the degenerate doping regime, the behavior is relatively constant to low temperature. Therefore, we use this model only as an approximation, and a more thorough numerical and experimental investigation of the devices to be used in the platform is the subject of future investigation. With this in mind, we approximate the capacitance of the junction using a simple parallel-plate model where the capacitance is given by $C = \epsilon A/d$, where ϵ is the material permittivity, A is the capacitor area, and d is the distance between the plates. We assume $\epsilon = 12\epsilon_0$, $A = 10\ \mu\mathrm{m} \times 100$ nm, and $d = 300$ nm. The energy associated with charging this capacitor is then calculated as $E_c = 1/2CV^2$. We note that for all values of photon number generated by the LEDs within this model, the applied voltage is below the built-in potential of the junction, so true forward-bias operation is not required. We anticipate that for the case of a *p-i-n* junction, the voltages required to achieve the same number of photons will increase slightly, but this can easily be accommodated by utilizing nanowires with larger critical currents.

APPENDIX E: WAVEGUIDE DESIGN FOR THE DENDRITIC ARBOR

In Fig. 21(a), we show effective indices at 1220 nm for slab thicknesses up to 600 nm to illustrate that many vertical modes can be present with high effective indices with only modest film thicknesses. We find that for < 200

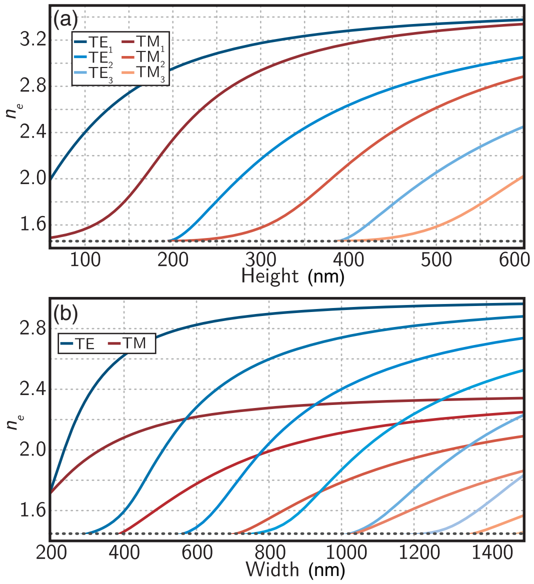

FIG. 21. Effective indices of refraction for various guided modes in a waveguiding layer with index of refraction $n = 3.52$ and cladding $n = 1.46$. (a) Slab mode calculations of both TE and TM modes for different film thicknesses showing different vertical mode orders. (b) TE and TM modes for different waveguide widths in a film of height 200 nm. The cladding index is shown as the dashed line in both (a) and (b).

thick waveguides, only the first vertical order TE and TM modes are present. Therefore, in Sec. III A, we assume a waveguide height of 200 nm. For massive scaling even beyond that presented in Sec. IV B, it may be necessary to use multimode waveguides with higher vertical as well as lateral modes and both polarizations.

Having selected 200 nm as our waveguiding layer thickness, we consider the lateral mode spectrum, as shown in Fig. 21(b). Here we see that the second-lateral-order TE mode emerges above the cladding index around 350 nm; we choose this as the single-mode width for the dendritic arbor simulations. From Fig. 21(b), we also find that a large number of higher-lateral-order modes are present with high effective index and modest waveguide width. For the dendritic arbor design presented in Fig. 10(b), it is important that a compact multimode waveguide be achievable. From this analysis, we find that a waveguide with tens of modes can be achieved while still maintaining a compact bend radius.

In addition to choosing the single-mode width, we also need to choose the minimum interwaveguide gap that avoids undesired coupling of modes in space. To do this, we calculate the supermode propagation constants as a function of the waveguide gap, as shown in Fig. 22. We see the splitting between the symmetric and antisymmetric modes is quite large for a gap of 100 nm, but both modes converge to the uncoupled value for a gap of 600 nm. The fractional splitting $\Delta\beta/\beta_0$ is shown in the inset. Here, $\Delta\beta$ is the difference between the propagation constants of the symmetric and antisymmetric supermodes, and β_0 is the uncoupled propagation constant. Based on this analysis, we choose 600 nm to be the interwaveguide gap for the dendritic arbor design of Fig. 11 and the value used in the scaling analysis of Sec. IV B.

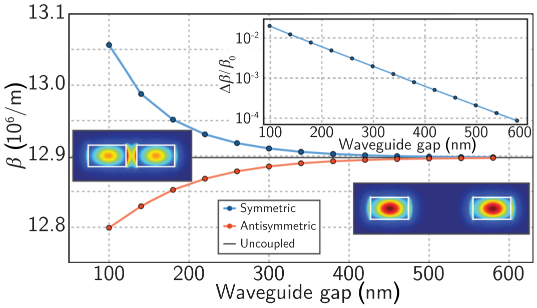

FIG. 22. Supermode propagation constants for 200-nm-thick, 350-nm-wide waveguides with 3.52 core index and 1.46 cladding index at $\lambda = 1220$ nm. The inset shows the fractional splitting, and the mode profiles show the symmetric mode for gaps of 100 and 600 nm.

APPENDIX F: SCALING

In reference to Sec. IV B, the length of a MLP layer is given by

$$L_l = (L_t + L_g + L_x)\frac{N_n}{N_{\mathrm{WG}}} + 2L_{\mathrm{WG}}N_{\mathrm{WG}} + L_n, \qquad \text{(F1)}$$

where L_t is the length of a single tap (or synapse) taken to be 10 μm; L_g is the length of a gap between two vertically running waveguides taken to be 5 μm, which is sufficiently wide to allow for undercut of the mechanically mobile synapses; L_x is the length of an intraplane waveguide crossing taken to be 3 μm; N_n is the number of neurons in a MLP layer [four in Fig. 14(a)]; N_{WG} is the number of vertically stacked waveguide planes used for routing; L_{WG} is the length of an interplane coupler between two waveguiding planes taken to be 10 μm. L_n is the length of a single neuron as shown in Fig. 11. L_n is determined predominantly by the number of inputs and, therefore, is taken to be the interwaveguide gap, 600 nm $\times$ N_n. The width of a single neuron is taken to be equal to its length, and within this model, we assume each neuron in a given layer has a synapse connecting to each neuron of the next layer.

APPENDIX G: INFORMATION

The application of Shannon's theory of communication [3] to neural systems enables the quantification of information-processing capacity. The mutual information (in bits) between a neural system and a stimulus can be represented as [27]

$$I_m = \int ds \int dr P[s] P[r|s] \log_2 \left(\frac{P[r|s]}{P[r]} \right). \tag{G1}$$

In Eq. (G1), $P[r]$ is the probability of spike rate r occurring given a stimulus s, $P[s]$ is the probability of stimulus s occurring from the set of all possible stimuli, and $P[r|s]$ is the conditional probability of response rate r being evoked when the system is presented with stimulus s. With a neuromorphic computing platform, one wants to maximize the mutual information. Because I_m within this model is calculated simply as a double integral over stimuli and response rates, we can maximize this quantity by increasing the limits of the integral. Because the proposed devices can operate at 20 MHz—and potentially up to 1 GHz by employing superconductors with faster thermal recovery—they can achieve response rates as well as receive stimulus across this entire bandwidth. The intrinsic speed of SPONs is greater than biological systems by a factor of 10^4, and this affects both the stimulus and response bandwidths in the double integral.

In addition to increasing the double integral by increasing the bandwidths, we can also maximize the bit depth. As we discuss in Sec. IV B, signals can be discretized into roughly 11 bits. However, it is possible to increase this number further at the expense of size and efficiency.

We discuss the s and r in Eq. (G1) with the photonic input to the receiver array and photonic output pulse rate of the transmitter in mind, but the neuron of Fig. 13 can receive more stimulus and generate more output. For example, if one considers not only the photons incident upon the receiver as stimulus but also the current through the SNSPD, the bit depth of the discernible stimuli increases further.

Equation (G1) is derived by considering the difference between the entropy of a neuron's responses to a given stimulus and the noise entropy. As such, it is a measure of the information content at the device level and not at the system level. A full analysis of the information content of population-encoded information is beyond the scope of this work. At a minimum, we point out that the information content of a population grows with the size of that population. Therefore, the high bandwidth of SPON devices, the ability to scale to units with large numbers of connections, and the ability to scale to systems with large numbers of units while maintaining a low power density points to the potential for complex systems with enormous information content. We note that these attributes are enabled by photonic signaling and superconducting electronics.

[1] A. M. Turing, On computable numbers, with an application to the entscheidungsproblem, Proc. London Math. Soc. **42**, 230 (1936).

[2] J. von Neumann, *First Draft of a Report on the EDVAC*, IEEE Annals of the History of Computing Vol. 14 (1945), p. 27.

[3] C. E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. **27**, 379 (1948).

[4] A. M. Turing, Computing machinery and intelligence, Mind **59**, 433 (1950).

[5] J. von Neumann, *The Computer and the Brain* (Yale University Press, New Haven, CT, 1958).

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

[7] *Silicon Neurons That Compute* (Springer, New York), Vol. 7552.

[8] Proceedings of 2014 IEEE Biomedical Circuits and Systems Conference (BioCAS), 2014.

[9] 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 scalable communication network and interface, Science **345**, 668 (2014).

[10] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, 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 (London) **529**, 484 (2016).

[11] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, On the expressive power of deep neural networks, arXiv:1606.05336.

[12] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, Exponential expressivity in deep neural networks through transient chaos, arXiv:1606.05340.

[13] H. W. Lin and M. Tegmark, Why does deep and cheap learning work so well?, arXiv:1608.0822.

[14] J. Machta, Entropy, information, and computation, Am. J. Phys. **67**, 1074 (1999).

[15] S. Panzeri, S. R. Schultz, A. Treves, and E. T. Rolls, Correlations and the encoding of information in the nervous system, Proc. R. Soc. B **266**, 1001 (1999).

[16] J. Hawkins and S. Ahmad, Why neurons have thousands of synapses, a theory of sequence memory in neocortex, Front. Neural Circuits **10**, 23 (2016).

[17] J. Bang-Jensen and G. Gutin, *Digraphs* (Springer-Verlag, Berlin, 2007).

[18] J. Hasler and B. Marr, Finding a roadmap to achieve large neuromorphic hardware systems, Front. Neurosci. **7**, 118 (2013).

[19] D. E. Kirichenko, S. Sarwana, and A. F. Kirichenko, Zero static power dissipation biasing of RSFQ circuits, IEEE Trans. Appl. Supercond. **21**, 776 (2011).

[20] T. Hirose, T. Asai, and Y. Amemiya, Pulsed neural networks consisting of single-flux-quantum spiking neurons, Physica (Amsterdam) **463C–465C**, 1072 (2007).

[21] P. Crotty, D. Schult, and K. Segall, Josephson junction simulation of neurons, Phys. Rev. E **82**, 011914 (2010).

[22] S. E. Russek, C. Donnelly, M. Schneider, B. Baek, M. Pufall, W. H. Rippard, P. F. Hopkins, P. D. Dresselhaus, and S. P. Benz, Stochastic single flux quantum neuromorphic computing using magnetically tunable Josephson junctions, in *Proceedings of IEEE International Conference on Rebooting Computing* (IEEE, New York, 2016).

[23] I. V. Vernik, V. V. Bol'ginov, S. V. Bakurskiy, A. A. Golubov, M. Y. Kupriyanov, V. V. Ryazanov, and O. A. Mukhanov, Magnetic Josephson junctions with superconducting interlayer for cryogenic memory, IEEE Trans. Appl. Supercond. **23**, 1701208 (2013).

[24] G. Indiveri *et al.*, Neuromorphic silicon neuron circuits, Front. Neurosci. **5**, 73 (2011).

[25] X. Wu, V. Saxena, K. Zhu, and S. Balagopal, A CMOS spiking neuron for brain-inspired neural networks with resistive synapses and *in-situ* learning, arXiv:1505.07814.

[26] F. Merrikh Bayat, X. Guo, M. Klachko, M. Prezioso, K. K. Likharev, and D. B. Strukov, Sub-1-μs, sub-20-nJ pattern classification in a mixed-signal circuit based on embedded 180-nm floating-gate memory cell arrays, arXiv:1610.02091.

[27] P. Dayan and L. F. Abbott, *Theoretical Neuroscience* (MIT Press, Cambridge, MA, 2001).

[28] M. D Eisaman, J. Fan, A. Migdall, and S. V Polyakov, Invited review article: Single-photon sources and detectors, Rev. Sci. Instrum. **82**, 071101 (2011).

[29] T. K. Woodward and A. V. Krishnamoorthy, 1-Gb/s integrated optical detectors and receivers in commercial CMOS technologies, IEEE J. Sel. Top. Quantum Electron. **5**, 146 (1999).

[30] Woo-Young Choi, Myung-Jae Lee, and Jin-Sung Youn, Si integrated photoreceivers, in *Proceedings of 2010 IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)* (IEEE, New York, 2010), pp. 77–81.

[31] Takasumi Tanabe, Hisashi Sumikura, Hideaki Taniyama, Akihiko Shinya, and Masaya Notomi, All-silicon sub-Gb/s telecom detector with low dark current and high quantum efficiency on chip, Appl. Phys. Lett. **96**, 101103 (2010).

[32] Jason J. Ackert, Silicon-on-insulator microring resonator defect-based photodetector with 3.5-GHz bandwidth, J. Nanophoton. **5**, 059507 (2011).

[33] K. K. Mehta, J. S. Orcutt, J. M. Shainline, O. Tehar-Zahav, Z. Sternberg, R. Meade, M. A. Popović, and Rajeev J. Ram, Polycrystalline silicon ring resonator photodiodes in a bulk complementary metal-oxide-semiconductor process, Opt. Lett. **39**, 1061 (2014).

[34] Solomon Assefa, Fengnian Xia, and Yurii A Vlasov, Reinventing germanium avalanche photodetector for nanophotonic on-chip optical interconnects, Nature (London) **464**, 80 (2010).

[35] Solomon Assefa, Fengnian Xia, William M. J. Green, Clint L. Schow, Alexander V. Rylyakov, and Yurii A. Vlasov, CMOS-integrated optical receivers for on-chip interconnects, IEEE J. Sel. Top. Quantum Electron. **16**, 1376 (2010).

[36] Jurgen Michel, Jifeng Liu, and Lionel C. Kimerling, High-performance Ge-on-Si photodetectors, Nat. Photonics **4**, 527 (2010).

[37] Joost Brouckaert, Gunther Roelkens, Dries Van Thourhout, and Roel Baets, Thin-film III-V photodetectors integrated on silicon-on-insulator photonic ICs, J. Lightwave Technol. **25**, 1053 (2007).

[38] R. Nagarajan, M. Kato, J. Pleumeekers, P. Evans, S. Corzine, S. Hurtt, A. Dentai, S. Murthy, M. Missey, R. Muthiah, R. A. Salvatore, C. Joyner, Richard Schneider, M. Ziari, F. Kish, and D. Welch, InP photonic integrated circuits, IEEE J. Sel. Top. Quantum Electron. **16**, 1113 (2010).

[39] F. Marsili, V. B. Verma, J. A. Stern, S. Harrington, A. E. Lita, T. Gerrits, I. Vayshnker, B. Baek, M. D. Shaw, R. P. Mirin, and S. W. Nam, Detecting single infrared photons with 93% system efficiency, Nat. Photonics **7**, 210 (2013).

[40] J. P. Sprengers, A. Gaggero, D. Sahin, S. Jahanmirinejad, G. Frucci, F. Mattioli, R. Leoni, J. Beetz, M. Lermer, M. Kamp, S. Höfling, R. Sanjines, and A. Fiore, Waveguide superconducting single-photon detectors for integrated quantum photonic circuits, Appl. Phys. Lett. **99**, 181110 (2011).

[41] W. Pernice, C. Schuck, O. Minaeva, M. Li, G. Goltsman, A. Sergienko, and H. Tang, High speed travelling wave single-photon detectors with near-unity quantum efficiency, Nat. Commun. **3**, 1325 (2012).

[42] S. Ferrari, O. Kahl, V. Kovalyuk, G. N. Goltsman, A. Korneev, and W. H. P. Pernice, Waveguide-integrated single- and multi-photon detection at telecom wavelengths using superconducting nanowires, Appl. Phys. Lett. **106**, 151101 (2015).

[43] F. Najafi, J. Mower, N. C. Harris, F. Bellei, A. Dane, C. Lee, X. Hu, P. Kharel, F. Marsili, S. Assefa, K. K. Berggren, and D. Englund, On-chip detection of nonclassical light by scalable integration of single-photon detectors, Nat. Commun. **6**, 5873 (2015).

[44] D. Sahin, A. Gaggero, J.-W. Weber, I. Agafonov, M. A. Verheijen, F. Mattioli, J. Beetz, M. Kamp, S. Höfling, M. C. M. van de Sanden, R. Leoni, and A. Fiore, Waveguide nanowire superconducting single-photon detectors fabricated on GaAs and the study of their optical properties, IEEE J. Sel. Top. Quantum Electron. **21**, 3800210 (2015).

[45] C. Schuck, X. Guo, L. Fan, X. Ma, M. Poot, and H. X. Tang, Quantum interference in heterogeneous superconducting-photonic circuits on a silicon chip, Nat. Commun. **7**, 10352 (2016).

[46] J. M. Shainline, S. M. Buckley, N. Nader, C. M. Gentry, K. C. Cossel, M. Popović, N. R. Newbury, S. W. Nam, and R. P. Mirin, A versatile, inexpensive integrated photonics platform, arXiv:1611.02346.

[47] A. Divochiy, F. Marsili, D. Bitauld, A. Gaggero, R. Leoni, F. Mattioli, A. Korneev, V. Seleznev, N. Kaurova, O. Minaeva, G. Gol'tsman, K. G. Lagoudakis, M. Benkhaoul, F. Lévy, and A. Fiore, Superconducting nanowire photon-number-resolving detector at telecommunication wavelengths, Nat. Photonics **2**, 302 (2008).

[48] F. Marsili, D. Bitauld, A. Gaggero, S. Jahanmirinejad, R. Leoni, F. Mattioli, and A. Fiore, Physics and application of photon number resolving detectors based on superconducting parallel nanowires, New J. Phys. **11**, 045022 (2009).

[49] C. M. Bishop, Neural networks and their applications, Rev. Sci. Instrum. **65**, 1803 (1994).

[50] S. Jahanmirinejad and A. Fiore, Proposal for a superconducting photon number resolving detector with large dynamic range, Opt. Express **20**, 5017 (2012).

[51] A. Engel, J. Lonsky, X. Zhang, and A. Schilling, Detection mechanism in SNSPD: Numerical results of a conceptually

simple, yet powerful detection model, IEEE Trans. Appl. Supercond. **25**, 2200407 (2015).

[52] F. Marsili, M. J. Stevens, A. Kozorezov, V. B. Verma, C. Lambert, J. A. Stern, R. D. Horansky, S. Dyer, S. Duff, D. P. Pappas, A. E. Lita, M. D. Shaw, R. P. Mirin, and S. W. Nam, Hotspot relaxation dynamics in a current-carrying superconductor, Phys. Rev. B **93**, 094518 (2016).

[53] A. N. McCaughan and K. K. Berggren, A superconducting-nanowire three-terminal electrothermal device, Nano Lett. **14**, 5748 (2014).

[54] *Fundamental Neuroscience*, edited by L. Squire, D. Berg, F. Bloom, S. du Lac, A. Ghosh, and N. Spitzer (Elsevier, New York, 2008).

[55] T. Nowotny and M. I. Rabinovich, Dynamical Origin of Independent Spiking and Bursting Activity in Neural Microcircuits, Phys. Rev. Lett. **98**, 128106 (2007).

[56] Aleksandr Biberman, Kyle Preston, Gilbert Hendry, Nicolás Sherwood-Droz, Johnnie Chan, Jacob S. Levy, Michal Lipson, and Keren Bergman, Photonic network-on-chip architectures using multilayer deposited silicon materials for high-performance chip multiprocessors, ACM J. Emerging Technol. Comput. Syst. **7**, 1 (2011).

[57] Zhiping Zhou, Bing Yin, and Jurgen Michel, On-chip light sources for silicon photonics, Light Sci. Appl. **4**, e358 (2015).

[58] Z. Yuan, B. E. Kardynal, R. M. Stevenson, A. J. Shields, C. J. Lobo, K. Cooper, N. S. Beattie, D. A. Ritchie, and M. Pepper, Electrically driven single-photon source, Science **295**, 102 (2002).

[59] A. J. Bennett, P. Atkinson, P. See, M. B. Ward, R. M. Stevenson, Z. L. Yuan, D. C. Unitt, D. J. P. Ellis, K. Cooper, D. A. Ritchie, and A. J. Shields, Single-photon-emitting diodes: A review, Phys. Status Solidi B **243**, 3730 (2006).

[60] A. J. Shields, Semiconductor quantum light sources, Nat. Photonics **1**, 215 (2007).

[61] N. Mizuochi, T. Makino, H. Kato, D. Takeuchi, M. Ogura, H. Okushi, M. Nothaft, P. Neumann, A. Gali, F. Jelezko, J. Wrachtrup, and S. Yamasaki, Electrically driven single-photon source at room temperature in diamond, Nat. Photonics **6**, 299 (2012).

[62] A. Y. Liu, C. Zhang, J. Norman, A. Snyder, D. Lubyshev, J. M. Fastenau, A. W. K. Liu, A. C. Gossard, and J. E. Bowers, High performance continuous wave 1.3 μm quantum dot lasers on silicon, Appl. Phys. Lett. **104**, 041104 (2014).

[63] Y. Wan, Q. Li, Y. Geng, B. Shi, and K. M. Lau, InAs/GaAs quantum dots on GaAs-on-V-grooved-Ci substrate with high optical quality in the 1.3 μm band, Appl. Phys. Lett. **107**, 081106 (2015).

[64] H. Schmid, M. Borg, K. Moselund, L. Gignac, C. M. Breslin, J. Bruley, D. Cutaia, and H. Riel, Template-assisted selective epitaxy of III-V nanoscale devices for co-planar heterogeneous integration with Si, Appl. Phys. Lett. **106**, 233101 (2015).

[65] Jifeng Liu, Xiaochen Sun, Rodolfo Camacho-Aguilera, Lionel C Kimerling, and Jurgen Michel, Ge-on-Si laser operating at room temperature, Opt. Lett. **35**, 679 (2010).

[66] Cheng Zeng, Yingjie Ma, Yong Zhang, Danping Li, Zengzhi Huang, Yi Wang, Qingzhong Huang, Juntao Li, Zhenyang Zhong, Jinzhong Yu, Zuimin Jiang, and Jinsong Xia, Single germanium quantum dot embedded in photonic crystal nanocavity for light emitter on silicon chip, Opt. Express **23**, 22250 (2015).

[67] Gordon Davies, The optical properties of luminescence centres in silicon, Phys. Rep. **176**, 83 (1989).

[68] D. Recht, F. Capasso, and M. J. Aziz, On the temperature dependence of point-defect-mediated luminescence in silicon, Appl. Phys. Lett. **94**, 251113 (2009).

[69] C. Sun *et al.*, Single-chip microprocessor that communicates directly using light, Nature (London) **528**, 534 (2015).

[70] H. Ennen, G. Pomrenke, A. Axmann, K. Eisele, W. Haydl, and J. Schneider, 1.54 μm electroluminescence of erbium-doped silicon grown by molecular beam epitaxy, Appl. Phys. Lett. **46**, 381 (1985).

[71] J. Palm, F. Gan, B. Zheng, J. Michel, and L. C. Kimerling, Electroluminescence of erbium-doped silicon, Phys. Rev. B **54**, 17603 (1996).

[72] Hak-Seung Han, Se young Seo, and Jung H. Shin, Optical gain at 1.54 μm in erbium-doped silicon nanocluster sensitized waveguide, Appl. Phys. Lett. **79**, 4568 (2001).

[73] Purnawirman, J. Sun, T. N. Adam, G. Leake, D. Coolbaugh, J. D. B. Bradley, E. Shah Hosseini, and M. R. Watts, *C*- and *l*-band erbium-doped waveguide lasers with wafer-scale silicon nitride cavities, Opt. Lett. **38**, 1760 (2013).

[74] S. T. Pantelides, The electronic structure of impurities and other point defects in semiconductors, Rev. Mod. Phys. **50**, 797 (1978).

[75] G. Davies, E. C. Lightowlers, and Z. E. Ciechanowaska, The 1018 meV (W or I_1) vibronic band in silicon, J. Phys. C **20**, 191 (1987).

[76] P. L. Bradfield, T. G. Brown, and D. G. Hall, Electroluminescence from sulfur impurities in a *p*-*n* junction formed in epitaxial silicon, Appl. Phys. Lett. **55**, 100 (1989).

[77] S. G. Cloutier, P. A. Kossyrev, and J. Xu, Optical gain and stimulated emission in periodic nanopatterned crystalline silicon, Nat. Mater. **4**, 887 (2005).

[78] Efraim Rotem, Jeffrey M. Shainline, and Jimmy M. Xu, Enhanced photoluminescence from nanopatterned carbon-rich silicon grown by solid-phase epitaxy, Appl. Phys. Lett. **91**, 051127 (2007).

[79] E. Rotem, J. M. Shainline, and J. M. Xu, Enhanced photoluminescence from nanopatterned carbon-rich silicon grown by solid-phase epitaxy, Appl. Phys. Lett. **91**, 051127 (2007).

[80] Jiming Bao, Malek Tabbael, Taegon Kim, Supakit Charnvanichborikarn, James S. Williams, Michael J. Aziz, and Federico Capasso, Point defect engineered Si sub-band gap light-emitting diode, Opt. Express **15**, 6727 (2007).

[81] Y. Yang, J. Bao, C. Wang, and M. J. Aziz, Sub-bandgap luminescence centers in silicon created by self-ion implantation and thermal annealing, J. Appl. Phys. **107**, 123109 (2010).

[82] H. Sumikura, E. Kuramochi, H. Taniyama, and M. Notomi, Ultrafast spontaneous emission of copper-doped silicon enhanced by an optical nanocavity, Sci. Rep. **4**, 5040 (2014).

[83] Wai Lek Ng, M. A. Lourenco, R. M. Gwilliam, S. Ledain ang G. Shao, and K. P. Homewood, An efficient

room-temperature silicon-based light-emitting diode, Nature (London) **410**, 192 (2001).

[84] Martin A. Green, Jianhua Zaho, Aihua Wang, Peter J. Reese, and Michael Gal, Efficient silicon light-emitting diodes, Nature (London) **412**, 805 (2001).

[85] Robert J. Walters, George I. Bourianoff, and Harry A. Atwater, Field-effect electroluminescence in silicon nanocrystals, Nat. Mater. **4**, 143 (2005).

[86] B. J. Coomer, J. P. Goss, R. Jones, S. Oberg, and P. R. Briddon, Interstitial aggregates and a new model for the I_1/W optical centre in silicon, Physica (Amsterdam) **273B–274B**, 505 (1999).

[87] J. M. Shainline and J. Xu, Silicon as an emissive optical medium, Laser Photonics Rev. **1**, 334 (2007).

[88] Joseph W. Goodman, Fan-in and fan-out with optical interconnections, Opt. Acta **32**, 1489 (1985).

[89] David A. Fattal, Jingjing Li, Zhen Peng, Marco Fiorentino, and Raymond G Beausoleil, Flat dielectric grating reflectors with focusing abilities, Nat. Photonics **4**, 466 (2010).

[90] Alexander V Kildishev, Alexandra Boltasseva, and Vladimir M Shalaev, Planar photonics with metasurfaces, Science **339**, 1232009 (2013).

[91] Nanfang Yu and Federico Capasso, Flat optics with designer metasurfaces, Nat. Mater. **13**, 139 (2014).

[92] J. K. Doylend, M. J. R. Heck, J. T. Bovington, J. D. Peters, L. A. Coldren, and J. E. Bowers, Two-dimensional free-space beam steering with an optical phased array on silicon-on-insulator, Opt. Express **19**, 21595 (2011).

[93] J. Sun, T. N. Adam, G. Leake, D. Coolbaugh, J. D. B. Bradley, E. Shah Hosseini, and M. R. Watts, *C*- and *L*-band erbium-doped waveguide lasers with wafer-scale silicon nitride cavities, Opt. Lett. **38**, 1760 (2013).

[94] E. J. Stanton, M. J. R. Heck, J. Bovington, A. Spott, and John E. Bowers, Multi-octave spectral beam combiner on ultra-broadband photonic integrated circuit platform, Opt. Express **23**, 11272 (2015).

[95] Pierre Baldi and Santosh S. Venkatesh, *On Properties of Networks of Neuron-Like Elements*, edited by Dana Z. Anderson (American Institute of Physics, New York, 1988).

[96] Z. Shao, Y. Chen, Y. Zhang, F. Zhang, J. Jian, Z. Fan, L. Liu, C. Yang, L. Zhou, and S. Yu, Ultra-low temperature silicon nitride photonic integration platform, Opt. Express **24**, 1865 (2016).

[97] W. D. Sacher, Y. Huang, G.-Q. Lo, and J. K. S. Poon, Multilayer silicon nitride-on-silicon integrated photonic platforms and devices, J. Lightwave Technol. **33**, 901 (2015).

[98] X. Li, H. Xu, X. Xiao, Z. Li, J. Yu, and Y. Yu, Compact and low-loss silicon power splitter based on inverse tapers, Opt. Lett. **38**, 4220 (2013).

[99] T. J. Seok, N. Quack, S. Han, R. S. Muller, and M. C. Wu, Large-scale broadband digital silicon photonic switches with vertical adiabatic couplers, Optica **3**, 64 (2016).

[100] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Netw. **61**, 85 (2015).

[101] M. S. Allman, V. B. Verma, M. Stevens, T. Gerrits, R. D. Horansky, A. E. Lita, F. Marsili, A. Beyer, M. D. Shaw, D. Kumor, R. Mirin, and S. W. Nam, A near-infrared 64-pixel superconducting nanowire single photon detector array with integrated multiplexed readout, Appl. Phys. Lett. **106**, 192601 (2015).

[102] P. Lennie, The cost of cortical computation, Curr. Biol. **13**, 493 (2003).

[103] S. B. Laughlin and T. J. Sejnowski, Communication in neuronal networks, Science **301**, 1870 (2003).

[104] "Neuron firing rates in humans".

[105] H. J. Caulfield, J. Kinser, and S. K. Rogers, Optical neural networks, Proc. IEEE **77**, 1573 (1989).

[106] D. Psaltis, D. Brady, X. G. Gu, and S. Lin, Holography in artificial neural networks, Nature (London) **343**, 325 (1990).

[107] D. Psaltis, A. A. Yamamura, K. Hsu, S. Lin, X.-G. Gu, and D. Brady, Optoelectronic implementations of neural networks, IEEE Commun. Mag. **27**, 37 (1989).

[108] Y. Liu, J. M. Shainline, X. Zeng, and M. Popović, Ultra-low-loss CMOS-compatible waveguide crossing arrays based on multimode Bloch waves and imaginary coupling, Opt. Lett. **39**, 335 (2014).

[109] W. Coomans, L. Gelens, S. Beri, J. Danckaert, and G. Van der Sande, Solitary and coupled semiconductor ring lasers as optical spiking neurons, Phys. Rev. E **84**, 036209 (2011).

[110] Damien Woods and Thomas J. Naughton, Optical computing: Photonic neural networks, Nat. Phys. **8**, 257 (2012).

[111] 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**, 1800212 (2013).

[112] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, Broadcast and weight: An integrated network for scalable photonic spike processing, J. Lightwave Technol. **32**, 4029 (2014).

[113] A. N. Tait, M. A. Nahmias, Y. Tian, B. J. Shastri, and P. R. Prucnal, *Nanophotonic Information Physics*, edited by M. Naruse (Springer, New York, 2014), Chap. 8.

[114] 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**, 19126 (2016).

[115] E. C. Mos, J. J. H. B. Schleipen, and H. de Waardt, Optical-mode neural network by use of the nonlinear response of a laser diode to external optical feedback, Appl. Opt. **36**, 6654 (1997).

[116] Konstantin S. Kravtsov, Mable P. Fok, Paul R. Prucnal, and David Rosenbluth, Ultrafast all-optical implementation of a leaky integrate-and-fire neuron, Opt. Express **19**, 2133 (2011).

[117] W. Coomans, L. Gelens, L. Mashal, S. Beri, G. Van der Sande, J. Danckaert, and G. Verschaffelt, Semiconductor ring lasers as optical neurons, Proc. SPIE Int. Soc. Opt. Eng. **8432**, 84321I (2012).

[118] A. Hurtado, K. Schires, I. D. Henning, and M. J. Adams, Investigation of vertical cavity surface emitting laser dynamics for neuromorphic photonic systems, Appl. Phys. Lett. **100**, 103703 (2012).

[119] Thomas Van Vaerenbergh, Martin Fiers, Pauline Mechet, Thijs Spuesens, Rajesh Kumar, Geert Morthier, Benjamin Schrauwen, Joni Dambre, and Peter Bienstman, Cascadable excitability in microrings, Opt. Express **20**, 20292 (2012).

[120] L. Larger, M. C. Soriano, D. Brunner, L. Appeltant, J. M. Gutierrez, L. Pesquera, C. R. Mirasso, and I. Fischer,

Photonic information processing beyond Turing: An optoelectronic implementation of reservoir computing, Opt. Express **20**, 3241 (2012).

[121] Y. Paquot, F. Duport, A. Smerieri, J. Dambre, B. Schrauwen, M. Haelterman, and S. Massar, Optoelectronic reservoir computing, Sci. Rep. **2**, 287 (2012).

[122] 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).

[123] B. S. Lankow and W. M. Usrey, *Visual Processing in the Monkey*, edited by R. M. Williams (Nova Science Publishers, 2011), Chap. 9.

[124] M. Riesenhuber, *Object Categorization in Man, Monkey, and Machine*, edited by S. J. Dickinson, A. Leonardis, B. Schiele, and M. J. Tarr (Cambridge University Press, Cambridge, England, 2009).

[125] Henry Markram, The blue brain project, Nat. Rev. Neurosci. **7**, 153 (2006).

[126] Shih-Chii Liu and Tobi Delbruck, Neuromorphic sensory systems, Curr. Opin. Neurobiol. **20**, 288 (2010).

[127] Chi-Sang Poon and Kuan Zhou, Neuromorphic silicon neurons and large-scale neural networks: Challenges and opportunities, Front. Neurosci. **5**, 108 (2011).

[128] G. E. Hinton and R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science **313**, 504 (2006).

[129] K. K. Likharev and V. K. Semenov, RSFQ logic/memory family: A new Josephson-junction technology for sub-terahertz-clock-frequency digital systems, IEEE Trans. Appl. Supercond. **1**, 3 (1991).

[130] K. Segall, S. Guo, P. Crotty, D. Schult, and M. Miller, Phase-flip bifurcation in a coupled Josephson junction neuron system, Physica (Amsterdam) **455B**, 71 (2014).

[131] A. Casaburi, R. M. Heath, M. G. Tanner, R. Cristiano, M. Ejrnaes, C. Nappi, and R. H. Hadfield, Current distribution in a parallel configuration superconducting stripline detector, Appl. Phys. Lett. **103**, 013503 (2013).

[132] A. Casaburi, R. M. Heath, M. Ejrnaes, C. Nappi, R. Cristiano, and R. H. Hadfield, Experimental evidence for photoinduced vortex crossing in current carrying superconducting strips, Phys. Rev. B **92**, 214512 (2015).

[133] J. K. W. Yang, A. J. Kerman, E. A. Dauler, V. Anant, K. M. Rosfjord, and K. K. Berggren, Modeling the electrical and thermal response of superconducting nanowire single-photon detectors, IEEE Trans. Appl. Supercond. **17**, 581 (2007).

[134] B. G. Streetman and S. K. Banerjee, *Solid State Electronic Devices*, 6th ed. (Pearson Prentice Hall, Upper Saddle River, NJ, 2006).

[135] N. D. Arora, J. R. Hauser, and D. J. Roulston, Electron and hole mobilities in silicon as a function of concentration and temperature, IEEE Trans. Electron Devices **29**, 292 (1982).