Optoelectronic Intelligence

Jeffrey M. Shainline

National Institute of Standards and Technology, Boulder, CO, USA, 80305

jeffrey.shainline@nist.gov

March 4th, 2021

Abstract

General intelligence involves the integration of many sources of information into a coherent, adaptive model of the world. To design and construct hardware for general intelligence, we must consider principles of both neuroscience and very-large-scale integration. For large neural systems capable of general intelligence, the attributes of photonics for communication and electronics for computation are complementary and interdependent. Using light for communication enables high fan-out as well as low-latency signaling across large systems with no traffic-dependent bottlenecks. For computation, the inherent nonlinearities, high speed, and low power consumption of Josephson circuits are conducive to complex neural functions. Operation at 4 K enables the use of single-photon detectors and silicon light sources, two features that lead to efficiency and economical scalability. Here I sketch a concept for optoelectronic hardware, beginning with synaptic circuits, continuing through wafer-scale integration, and extending to systems interconnected with fiber-optic tracts, potentially at the scale of the human brain and beyond.

1 Introduction

General intelligence is the ability to assimilate knowledge across content categories and to use that information to form a coherent representation of the world. The brain accomplishes general intelligence through many specialized processors performing unique, complex computations [1, 2]. The information generated by these processors is communicated throughout the network via dedicated connections spanning local, regional, and global scales [3]. On the micro-scale, synapses, dendrites, and neurons are specialized processors comprising the gray matter computational infrastructure of the brain [4]. On the meso-scale, cortical minicolumns of 100 neurons act as specialized processors [5], and on the macro-scale, brain regions play that role [6]. Information is communicated between these modules via axonal fibers that comprise the white matter communication infrastructure of the brain. On short time scales, information processing occurs in synapses [7], dendrites [8], and within single neurons [9]. On longer time scales, the information generated by minicolumns is communicated across wider regions of the network so that the knowledge of specialized processors can combine in a comprehensive interpretation of a subject [10]. The utilization of many specialized processors combining their shared computational resources across many scales of space and time enables the brain to achieve general intelligence [1,2].

Computation and communication are the complimentary pillars of neural systems. Hardware for artificial general intelligence (AGI) will achieve the highest performance if complex, local processors can pool the information from their specialized computations through global communication. Electrons excel at computation, while light is excellent for communication. In silicon hardware, monolithic optical links between a processor and memory have been demonstrated [11]. These devices were fabricated in a 45-nm CMOS node with no in-line process changes, and off-chip light sources were utilized. Such work is driven by the desire for increased communication bandwidth in multi-core architectures. These architectures continue to expand into on-chip networks, in some cases resulting in highly distributed, brain-inspired systems implemented with CMOS electronics [12–17]. As computing grows more distributed, communication becomes a bottleneck. A primary challenge affecting further chip-scale electronic-photonic integration is the difficulty of achieving a light source on silicon that is robust, efficient, and economical [18, 19]. Lessons learned from very-large-scale integration (VLSI) inform us that economical fabrication of integrated circuits comprising simple components is necessary for scaling. In this regard, difficulties associated with integrated light sources are the most significant impediment to optoelectronic VLSI.

It is the perspective of our group at NIST that hardware incorporating light for communication between electronic computational elements combined in an architecture of networked optoelectronic spiking neurons may provide potential for AGI at the scale of the human brain. Spiking neurons are circuits that integrate signals over time and produce pulses when a threshold is reached. The spiking neurons discussed here are optoelectronic in that the pulses communicated from neurons to synapses consist of photons, while the computations performed within the neurons utilize electronics. Each neuron contains a light source, which is driven electrically upon reaching threshold. Each synapse contains a detector, which converts the optical signal to an electrical current or voltage upon receiving a photonic synapse event. Each neuron is a separate entity, and no hardware components are multiplexed to represent the operations of separate neurons at different times. While much of present-day computing infrastructure has evolved to implement a von Neumann architecture perfoming sequential operations in the model of a Turing machine, the functioning of neural systems departs

considerably from this model. Light has even more to offer in a neural computing context, because communication across scales is indispensable. Further, the spiking behavior of Josephson junctions combined with the efficiency of single-photon detectors make a compelling case for optical integration with superconducting electronics [20,21]. Such a choice necessitates low-temperature operation near 4 K. At this temperature, silicon light sources become available [22], indicating that a major impediment to optoelectronic VLSI many not be present in the superconducting domain. This article summarizes the reasoning behind the assertion that superconducting optoelectronic systems have unique potential to achieve general intelligence when considered from the perspectives of cognitive science and VLSI.

The unique cognitive capabilities of humans derive in part from the scale of the brain, including the number of neurons and the complexity of the communication network. While there is much to be gained from AI hardware at smaller scales, this article considers technological pathways to large cognitive systems, with tens to hundreds of billions of neurons, and communication infrastructure of commensurate complexity. Such technology will likely require many interconnected wafers, each packed densely with integrated circuits. We may refer to this field of research as "neuromorphic supercomputing". The effort is in some ways more akin to the construction of a fusion reactor or particle accelerator than a microchip, and potentially offering a similar scale of societal benefit in the form of an experimental test bed enabling the elucidation of the mechanisms of cognition and the exploration of the physical limits of intelligence.

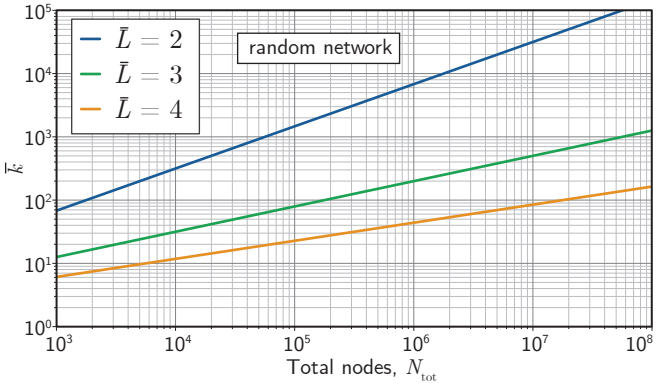

Figure 1: The average number of connections per node ($\bar{k}$) required to maintain a given average path length ($\bar{L}$) across a random network as a function of the total number of nodes in the system (N_{tot}).

2 Neuroscience as a guide

To guide the design of hardware for AGI, we must consider insights from neuroscience as to how neural systems integrate information across space and time to accomplish cognition [3,10,23,24]. A brief summary is provided here, highlighting aspects most pertinent to the design of hardware for cognition.

In the temporal domain oscillations and synchronization structure the activity of populations of neurons [10]. The spiking activity of neurons is observed to comprise nested oscillations across a range of frequencies [25]. On the fastest time scales of the brain, local clusters of neurons engage in transient dynamical activity induced by the present stimulus. These patterns of activity are referred to as gamma oscillations (80 Hz), and activity in this band is modulated by lower frequency oscillations [26,27] resulting from the combined activity of neurons across larger regions of the network [28]. These slower, broader patterns are referred to as theta oscillations (6 Hz), and neuronal communication across a network depends upon information present in gamma activity being structured into more complex syntax by dynamics on theta timescales [29,30]. This rich structuring of information in time is enabled by the spiking behavior of neurons. Computation and communication based on spikes facilitate a diversity of information coding schemes with resilience to noise while maintaining high energy efficiency due to sparse activity.

In the spatial domain a feature of neural systems that will recur in the present discussion is their modular, hierarchical construction [3, 23, 24, 31]. Neural systems are modular in that they are comprised of local regions of densely interconnected structures with sparser connectivity between such regions. Neural systems are hierarchical in that this pattern repeats across spatial scales in a fractal manner: minicolumns aggregate into columns, columns into complexes, etc. This fractal property is necessary to enable networks to scale arbitrarily, with dynamics constrained only by the physical hardware and spatial extent of the system rather than by the ability to communicate across the network [32]. Communication between distant modules is enabled by power-law scaling: the number of connections being sent to distant modules does not decay exponentially, but rather follows a power law [33,34]. The non-vanishing tail of long-range connections enables distant modules to quickly become correlated. In constructing hardware for artificial intelligence, it is imperative to enable rapid communication without traffic-dependent bottlenecks. Modules must be able to quickly engage in gamma activity, while signals from many interconnected modules at multiple levels of hierarchy must be able to simultaneously transmit across the complex network. The specific time scales defining behavior analogous to gamma and theta oscillations will be determined by the underlying computational devices.

In the form of gamma activity, clusters of neurons represent specific content, and the information from these clusters must be shared with other regions of the network to form a multifaceted representation of a stimulus. This computation and communication is facilitated by networks with a high clustering coefficient yet also an average path

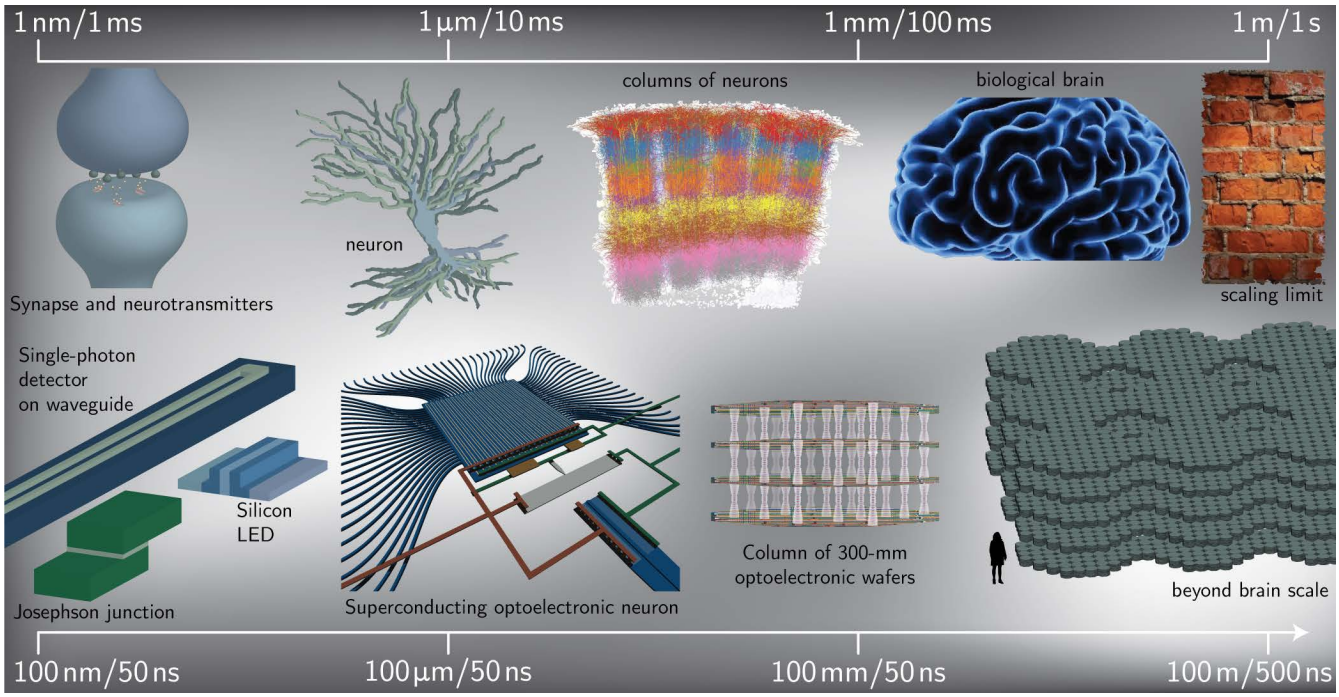

Figure 2: Structure across scales. Biological systems have functional components from the nanometer scale (neurotransmitters, axonal pores) up to the full brain (0.3 m linear dimension for full human cerebral cortex [35]). Speeds are limited by chemical diffusion and signal propagation along axons, which may ultimately limit the size of biological neural systems [10, 36]. The time constants associated with chemical diffusion and membrane charging/discharging span the range from 1 ms to 100 ms [4, 37] and dictate the speeds at which information processing occur. The wall at the right indicates the communication-limited spatial-scaling barrier. Optoelectronic devices rarely have components with critical dimension smaller than 100 nm, and optoelectronic neurons are likely to be on the 100-µm scale, with dendritic arbor extending for millimeters and axonal arbor in some cases spanning the system. The time constants of these components can be engineered in hardware across a very broad range with high accuracy through circuit parameters, enabling rapid processing as well as long-term signal storage. Optical communication enables optoelectronic systems to extend far beyond the limits imposed by the slow conduction velocity of axons.

length nearly as short as a random graph [38]. In the language of network theory, if node a is connected to node b, and b is connected to c, then clustering quantifies the probability that a will be connected to c. Path length quantifies the number of intermediate nodes that must be traversed to get from one node to another along the network connections. The average path length is determined by calculating this quantity over all pairs of nodes in the network, and taking the mean. A network with high clustering and low average path length is referred to as a "small-world network" [39]. Small-world networks are ubiquitous throughout the brain [3] and require long-range connections. In a random network, near and distant connections are equally probable, so the average path length across the network achieves a lower limit on path length for a given number of edges connecting a given number of nodes. Figure 1 shows the number of edges required per node to achieve a given average path length as a function of the number of nodes in the random network. For a modest network with one million nodes, each node must make several thousand connections to maintain a path length of two. For the case of a network with 100 million nodes, each node must make over one hundred thousand connections. This is similar to the hippocampus in the human brain, with nearly 100 million neurons, some with 50,000 or more nearly random synaptic connections [10]. Maintaining a short path length across the network is critical for information integration, and is an important motivator to use light for communication.

At the device level, dynamical behaviors thought to be necessary for attention, cognition, and learning, such as cross-frequency coupling [10] and synaptic plasticity [40–42], require complex capabilities. Dynamical synapses, dendrites, and neurons allow one structural network to realize myriad functional networks adapting on multiple time scales. While light is excellent for communication, electrical circuits are better equipped to perform these nonlinear, dynamical functions. For communication and computation, neural information processing will benefit immensely from optoelectronic integration.

Figure 2 charts the structures present on various scales for biological and optoelectronic hardware. The human brain has features spanning roughly eight orders of magnitude in size, from a nanometer to a tenth of a meter.

Across time, activity ranges from the 1 ms time scale of neurotransmitter diffusion across a synapse, through the 200 ms time scale of brain-wide theta oscillations, up to the memory retention time of the organism. The speeds of devices and communication in the brain are limited by the chemical and ionic nature of various operations. The maximum size of the brain may be limited by the slow conduction velocity of ionic signals along axons. If the brain were larger, signals would not have time to propagate between different regions during the period of theta oscillations, and system-wide information integration could not be efficiently achieved [10, 36].

Light and electronics together can enable communication and computation across spatial and temporal scales. We have proposed a specific approach we see as most conducive to large-scale implementation for AGI [20, 21, 36, 43]. The approach combines waveguide-integrated light sources and single-photon detectors for communication [20, 22] with Josephson circuits for synaptic, dendritic, and neuronal computation [21, 43]. As illustrated in Fig. 2, these optoelectronic networks will have features as small as 100 nm and potentially extend up to kilometers. Neuronal inter-spike intervals can be as short as 50 ns, while synaptic and dendritic processing occurs on the 50 ps time scale of Josephson junctions. Time constants can be chosen across many orders of magnitude, enabling information processing and memory across time scales. Figure 2 is intended to emphasize that if communication barriers can be removed, neural systems of extraordinary scale can be achieved.

Schematic illustration of the neurons and modular networks under consideration are illustrated in Fig. 3. A neuron with a complex dendritic tree is shown in Fig. 3(a). Neurons with excitatory ($\mathsf{S_e}$) and inhibitory ($\mathsf{S_i}$) synapses feeding into dendrites (D) and the neuron cell body (N). Upon reaching threshold, the transmitter (T) produces a pulse of light that fans out across a network of waveguides (not shown). Modular hierarchical construction is depicted in Fig. 3(b). The smallest blocks represent neurons, and their connections predominantly reside within their local module (blue). Yet important connections are made at all levels of hierarchy (red and dark green). Hardware for AGI must employ modular, hierarchical networks of complex neurons with rich dynamics that adapt on multiple timescales.

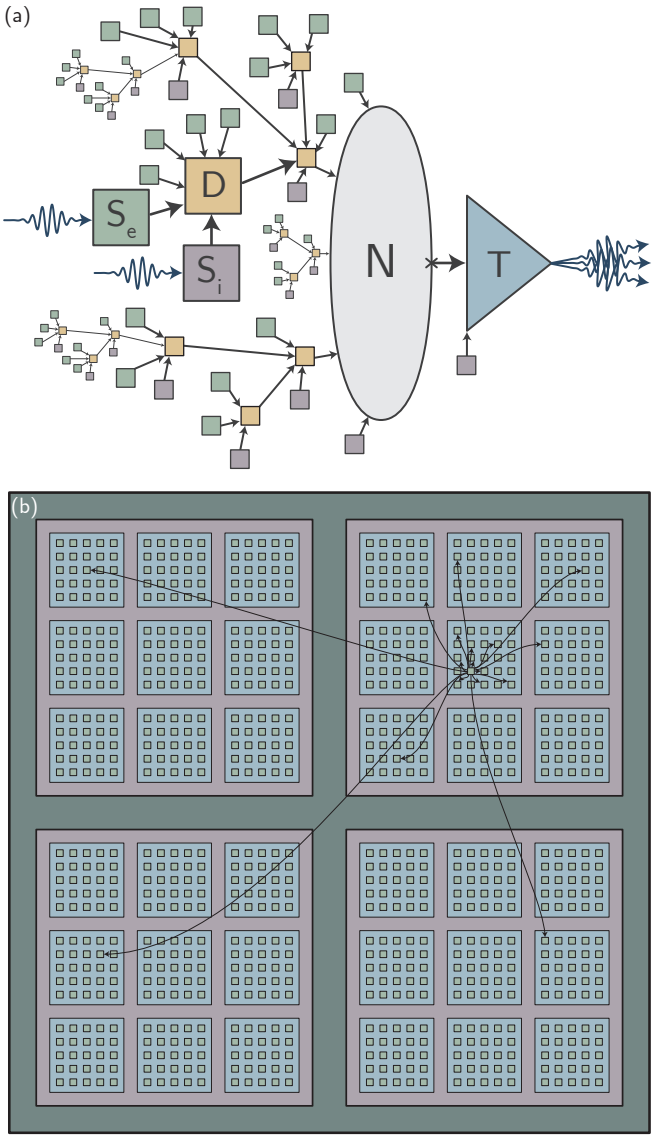

Figure 3: Block diagrams. (a) Optoelectronic neuron. Electrical connections are shown as straight, black arrows, and photons are shown as wavy, blue arrows. (b) Modular, hierarchical network construction. Here black arrows are photonic connections.

3 Superconducting optoelectronic synapses, dendrites, and neurons

Having chosen to communicate synaptic events with light, the quantum limit is a single photon per synaptic connection. We have designed a synapse (Fig. 4(a), [21, 36, 43]) that detects a single near-infrared photon and requires no power to retain the synaptic state, a feature enabled by the dissipationless nature of superconductors. The synapse utilizes a superconducting-nanowire single-photon detector (SPD), which is simply a current-biased strip of superconducting wire [44]. To achieve the desired synaptic operation, an SPD is combined in circuits with Josephson junctions (JJs) and superconducting loops to achieve the functions needed for neural information processing. In optoelectronic synapses of this design, the current bias across a single JJ establishes the synaptic weight (I_{sy} in Fig. 4(a)). This current bias can be dynamically modified through various photonic and electronic means based on control signals or network activity.

The signals from many synapses can be combined

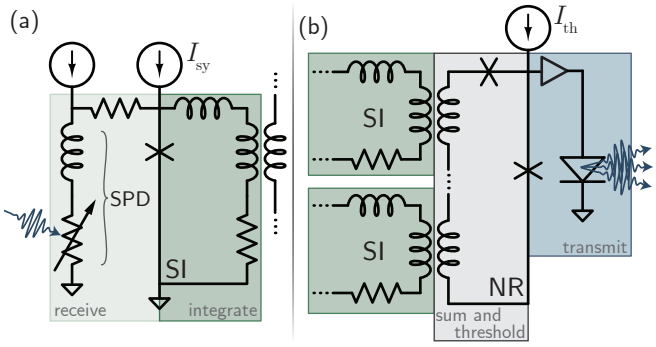

Figure 4: Circuit diagrams. (a) Superconducting optoelectronic synapse combining a single-photon detector (SPD) with a Josephson junction and a flux-storage loop, referred to as the synaptic integration (SI) loop. The synaptic bias current (I_{sy}) can dynamically adapt the synaptic weight. (b) Neuron cell body performing summation of the signals from many synapses as well as thresholding. Here the neuronal receiving (NR) loop is shown collecting inputs from two SI loops, but scaling to thousands of input connections appears possible. Upon reaching threshold, the transmitter circuit (amplifier [47] and LED [22]) produce a pulse of light that communicates photons to downstream synapses. The neuronal threshold current (I_{th}) can dynamically adapt the neuronal threshold.

through transformers coupled to dendrites or neurons (Fig. 4(b)). Neurons constructed in this manner are highly modular in that synapses, dendrites, and the neuron cell body itself are all based on the same core circuit, comprising a superconducting quantum interference device (SQUID) embedded in a flux-storage loop. SQUIDs are perhaps the most ubiquitous of all superconducting circuits [45, 46], often used as sensors due to their extraordinary sensitivity to magnetic flux and low-noise operation. These properties make SQUIDs ideal circuits for dendrites and neurons to perceive and respond to minute changes in analog signal levels. Dendritic and neuronal nonlinearities are a natural consequence of the JJ critical current, and can be shaped through the choice of circuit parameters, such as loop inductances and resistances, as well as dynamically with adaptive bias currents. Due to the prominent role of superconducting current storage loops, we refer to these as loop neurons. We refer to networks of loop neurons as superconducting optoelectronic networks (SOENs). In the operation of loop neurons, a single photon triggers a synaptic event, and spike-timing-dependent plasticity is induced by two photons—one from each neuron associated with the synapse.

In addition to the choice of SPDs as the detectors in the system, we must also select a light source, which must be fabricated across wafers by the millions. Because our choice of detectors dictates cryogenic operation, silicon light sources are an option. The light sources we have in mind are silicon LEDs [22], employing luminescence from defect-based dipole emitters. From the perspective of VLSI, achievement of a silicon light source as simple as a transistor would be the greatest contribution to the success of this technology. If cryogenic operation enables both single-photon detectors and silicon light sources, it will be worth the added infrastructure for cooling. We further justify this decision in Sec. 4.

To achieve complex neural circuits, we aim for monolithic integration of light sources, detectors, and superconducting circuit elements. Our group's experimental progress towards this end is summarized in Fig. 5. An important milestone was the demonstration of an all-silicon monolithic optical link. We measured waveguide coupling of light from micron-scale, all-silicon LEDs to integrated, silicon-based SPDs on a photonic chip (Fig. 5(a-c), [22]). Further progress on scalability and characterization of waveguide-integrated SPDs for use in the synapses under consideration was also presented in Ref. 50. The performance achieved in the first iteration of these optical links was not yet adequate. The observed efficiency was 5×10^{-7}, while 10^{-3} is desirable for large systems [36]. Yet the simplicity of both the source and detector made the fabrication and demonstration of a monolithic optical link far easier than if room-temperature operation were required. Subsequent work improved the brightness of the sources by two orders of magnitude through optimized fabrication procedures [51]. Additional gains may result from optimization of the diode structure used for electrical injection of carriers into the waveguide where electron-hole recombination at emissive centers produces waveguide-coupled luminescence. Elimination of etched surfaces and proper passivation in the active region may significantly reduce non-radiative recombination. Improvements to the optical structure may increase coupling efficiency from the emitters to the waveguide mode. For this application, the light sources are only required to produce incoherent pulses of 10,000 photons (1 fJ) at 20 MHz when operating at 4 K. Modest advances could enable silicon light sources to meet these specifications.

We have also demonstrated superconducting amplifiers capable of generating the voltage required to produce light from these sources (Fig. 5(d-f), [47]). Generating more than a millivolt with superconducting circuits is difficult, but the thin-film, micron-scale cryotron demonstrated in Ref. 47 leverages the extreme nonlinearity of the superconducting phase transition to rapidly generate high impedance and voltage with low energy, thus driving a semiconductor light source during each neuronal firing event. In Ref. 47 we demonstrated the use of these amplifiers to drive the LED-SPD link of Ref. 22. Fabrication of these devices appears compatible with silicon microelectronic manufacturing, provided the high-temperature steps required for dopant activation and contact annealing required for semiconductor devices are performed prior to the deposition of superconducting thin films.

Following a neuronal spike, the light produced by an amplifier driving an LED fans out across a network of

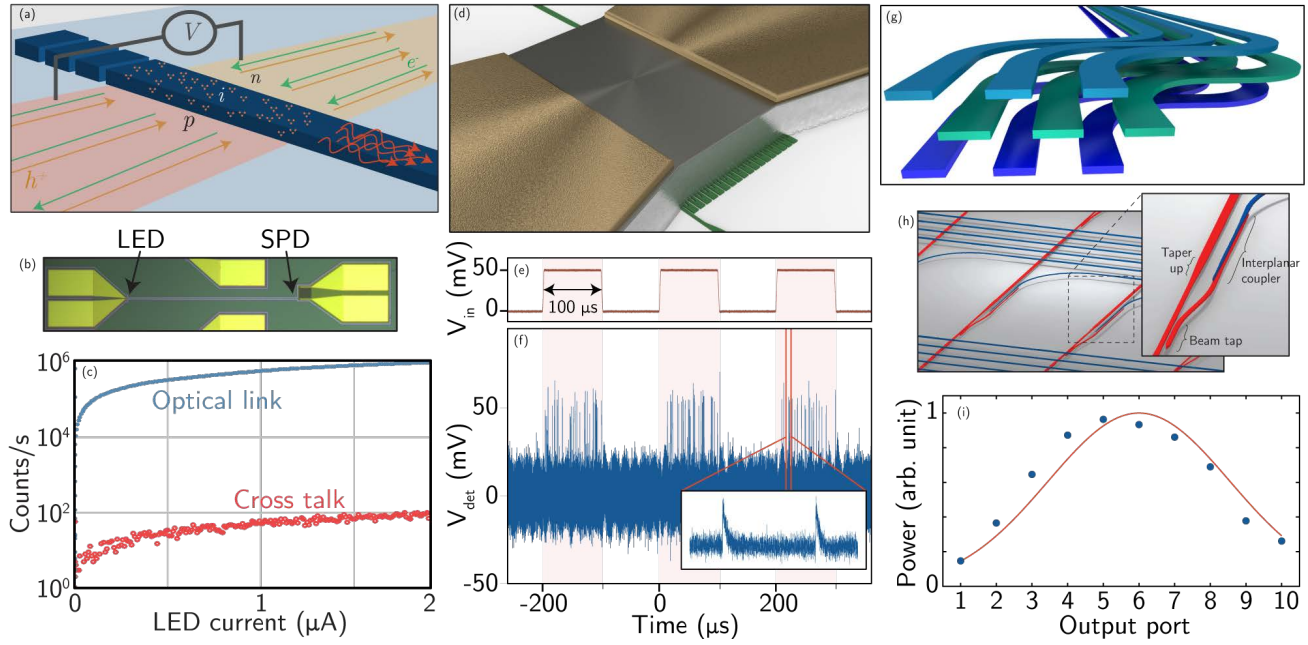

Figure 5: Experimental progress toward superconducting optoelectronic networks. (a) Schematic of waveguide-integrated silicon LED. (b) Microscope image of a silicon LED waveguide-coupled to a superconducting-nanowire detector. (c) Experimental data showing that light is coupled through the waveguide, while cross talk to an adjacent detector on the chip is suppressed by 40 dB. (a-c) Adapted from Ref. 22. (d) Schematic of the superconducting thin-film amplifier. (e,f) The resistive switch driving the LED. (e) Square pulses are driven into the switch gate. (f) When the switch is driven, light is produced from the LED and detected by the SPD. (d-f) Adapted from Ref. 47. (g) Schematic of multi-planar integrated waveguides for dense routing. (h) Schematic of feed-forward network implemented with two planes of waveguides. (i) Data from an experimental demonstration of routing between nodes of a two-layer feed-forward network with all-to-all connectivity. (g-i) Adapted from Refs. 48 and 49.

micron-scale dielectric waveguides terminating on the superconducting detectors at each synaptic connection. We have demonstrated multiple vertically integrated planes of these waveguides (Fig. 5(d-f), [48]), and used them to implement the architecture of a feed-forward neural network with two layers of 10 neurons per layer and all-to-all connectivity [49].

4 The landscape of research in photonic and superconducting neural systems

This approach to neural computing resides at the confluence of semiconductors, superconductors, and photonics, and should be contextualized with other work in these fields. It is clarifying to acknowledge that nearly all the efforts to use photonics or superconducting electronics for neural systems are focused on the entirely reasonable goal of doing useful computations with hardware that is available right now. These efforts are valuable and promising for their own ends without seeking brain-scale cognition. Comments here contrasting SOEN hardware with other current efforts are not criticisms of any work in the field, but rather explanation of the reasoning behind SOENs for large cognitive systems. A short summary of other photonic and electronic efforts is provided here, and comprehensive reviews of emerging neural hardware can be found in recent literature [52–54].

4.1 Semiconductor electronic neural systems

Given the extraordinary success of CMOS electronics, utilization of that hardware platform is the clear place to begin a search for artificial neural circuits. The history of exactly this pursuit is rich [55–57], accomplished [58–62], and exciting advances lie ahead [63]. So why advocate for an alternative? To further explain why we place optical communication at the center of hardware development, I briefly summarize the physical limitations of electrical interconnection networks [64]. It is impracticable in silicon electronics for a single device to source current to many other devices. A shared communication network must be employed. Switched media networks are used for this purpose. Each device must then only communicate to the nearest switch in the network. Because the communication infrastructure is shared, devices must request and wait for access to the switch network to transmit messages.

This approach to communication leverages the speed of electronic circuits to compensate for the challenge of direct communication. Limitations are reached when many devices must communicate with many other devices simultaneously. While neural activity is generally sparse, during high activity, such as coordinated gamma bursting at the peak of a theta oscillation (Sec. 2), many neurons must communicate simultaneously across the network. Due to the traffic-dependent bottleneck of shared interconnection infrastructure, as more neurons are added to the network, the average rate of neuronal firing events must decrease, and nested oscillations must shift to lower frequencies due to delays. Activity is limited to frequencies much slower than the brain in systems much smaller than the brain. Integration of information across the network is limited by the communication infrastructure.

4.2 Optical neural systems

One means to alleviate communication limitations is through the use of optics. The field of photonic neural systems began [65,66] with an implementation of the Hopfield model [67]. The objective was to combine the parallelism and interconnectability of optics, which are linear phenomena, with bistable optical devices to provide the thresholding nonlinearity of the Hopfield model. The hardware proposed combined compound-semiconductor LEDs with photodiodes and electronics for an initial implementation of nonlinearity to be replaced by optical bistable devices in subsequent generations. While LEDs and laser diodes have become mature technologies, bistable optical devices have not.

The field of photonic neural systems has since experienced an immense diversification, with myriad efforts using free-space optics [68], fiber components [69], and on-chip integrated photonics [70–72]. Along one branch of this tree, excitable lasers have been explored as spiking neurons [69]. These lasers integrate several optical inputs, and release a laser pulse upon reaching threshold. These devices can be extremely fast, but consume too much power for scaling to the level of the human brain. It is also difficult to tailor the neuronal responses, as they are primarily determined by carrier and cavity dynamics, which are dictated by basic physics and not easily adjusted with circuit parameters. Excitable lasers can be used as spiking neurons in the broadcast-and-weight architecture [73], wherein each neuron is assigned a wavelength, and synaptic weights are established with microring resonators that attenuate the optical signals, much like waveglength-division-multiplexed fiber-optic networks. In conventional silicon photonics [74], such multiplexing employs around 10 channels. It may be possible to extend this to 100 [73, 75], but even this limited number of channels would require cumbersome control circuits to hold synaptic weights stable. The requirement of precise control at every synapse as well as the non-monotonic, rapidly varying Lorentizian lineshape of microring resonances is not optimal for large-scale, unsupervised learning.

Phase change materials have also been explored for neuronal thresholding [76] and as a means of implementing variable attenuation of photonic signals to establish a synaptic weight [77]. This approach requires billions of photons to achieve synaptic weight modification. Relying on the properties of a material to achieve the complex computations occurring at a synapse limits functionality as compared to behaviors that can be tailored with integrated circuits.

Deep learning with continuous fields rather than spiking neurons is also receiving attention, and networks of on-chip, cascaded Mach-Zehnder interferometers are a prominent approach [70]. Such networks excel at feed-forward processing operations, but are not conducive to the recurrent networks employed by spiking neural systems nor the activity-dependent plasticity necessary for unsupervised learning. The challenge arises because in meshes of interferometers, adjustment of one phase modifies multiple synaptic weights. While such a technique may be suitable for specific training algorithms employed in supervised learning [78], it appears cumbersome for unsupervised learning in large neural systems, where local activity at each synapse updates that synaptic weight.

Another exciting and related application space of photonics is in reservoir computing. This field has been innovative and productive in recent years [79–83]. The objectives and hardware are only loosely related to the subject of large-scale cognition considered here, so further discussion is omitted.

As a broad point of contrast between the synapses discussed here and other systems using light for neural computing, most photonic neural systems encode information in the amplitude of optical signals received at a detector, and synaptic weights are established through modulation of the intensity of these optical signals. Whether phase modulation or direct amplitude modulation are leveraged, encoding synaptic weights in the intensity of light on a detector differs from the synaptic operations we are pursuing, where light is used for binary communication, and synaptic weights are established by electronic responses. This approach minimizes the optical power required and eliminates a source of noise. If synaptic weights are encoded in the intensity of an optical signal, noise from the light source is convoluted with the synaptic weight. With binary optical signaling the light level incident upon a synaptic detector does not influence the electronic response of the synapse, which is determined by the electronic circuits reading out the synaptic receiver. A binary response can be achieved with semiconductor receivers or superconducting circuits. In Ref. 50 we have shown that the response of a superconducting SPD is independent of the number of photons present in an incident pulse across four orders of magnitude of input intensity. Because no information is encoded in the light level, this form of optical communication does not suffer from typical shot noise. Provided one or more photons are received by the detector,

a synapse event is communicated. The Poisson distribution gives the probability that zero photons are received. With an average number of five or greater photons transmitted per synapse event, the probability of receiving zero photons is less than 1%, a considerably lower error rate than biological synaptic transmission [84]. We assume each neuronal light source will generate 10 photons per synaptic event to accommodate 3 dB of propagation loss while achieving 99% transmission success rate. All energy and power consumption estimates presented here use this value.

4.3 Superconducting electronic neural systems

Many approaches to neural computing using superconducting circuits leverage the nonlinear properties of Josephson junctions. The objective of early superconducting neural circuits was to perform the weighted summation and thresholding operations required in the computational primitives of ANNs [85, 86]. The circuits employed were similar to those utilized in superconducting digital logic, as were the basic concepts, such as using an up-down counter to implement synaptic weights [86]. From the beginning, and continuing to the present [87–89], attention is paid to sculpting a sigmoidal transfer function to implement back-propagation as well as alternative circuits for achieving Hebbian-type learning [85].

More recent efforts have broadened attention to consider also spiking neural systems, leveraging the inherent threshold and spike production of JJs [90, 91]. The most successful experimental effort to date demonstrated coupling of two neurons based on JJs, with inter-spike intervals on the order of tens of picoseconds [92]. Additional progress has been made in synaptic memory technology based on magnetic JJs, wherein magnetic nanoclusters embedded in the tunneling barrier of a JJ are re-oriented by current pulses, providing a means to modify the junction critical current and dynamically reconfigure the response of a synaptic circuit [91]. Such devices offer similar functionality to memristors being pursued for use in semiconductor-based neural systems [93].

The superconducting circuits discussed in the present context have much in common with other contemporary efforts in JJ-based neural systems, particularly in the use of SQUIDs as the primary active element [90, 94]. One point of contrast is that our emphasis is toward high-capacity, analog flux-storage loops with diverse time constants as well as utilization of complex neurons with a dendritic tree for hierarchical information processing within each neuron, whereas other efforts are primarily focused on the high-speed [92] and energy efficiency [88] enabled by the use of superconducting circuits. This distinction is minute in comparison to the difference introduced by the choice to employ photonic communication. The challenge with using superconducting electronics alone to enable large-scale cognitive system is communication. In superconducting circuits, direct fan-out is usually limited to two, so for neurons to make thousands of connections, many stages of pulse splitters and active transmission lines must be employed. This leads to a cumbersome communication network requiring many JJs and severe challenges for wiring and routing. Reference 95 analyzed fan-out and fan-in in these systems and argued there is no fundamental limit to fan-out. Fan-in was identified as a limiting factor. However, the use of analog synaptic integration loops with high inductance eliminates the fan-in bottleneck [36].

Fan-out challenges with superconducting circuits are not fundamental, and reasonable researchers in the field can disagree about the scale at which practical limits will be reached. For long-distance communication, pulses produced by JJs must be regenerated along active transmission lines. These transmission lines use JJs spaced periodically to re-transmit pulses, and the spacing of these JJs is set by inductance requirements. Using typical superconducting wires, a pitch of 100 µm between these junctions is expected, meaning a neuron trying to reach a synapse on the other side of a $1\,\mathrm{cm} \times 1\,\mathrm{cm}$ die will require 100 JJs for communication to that synapse. At the scale of a 300 mm wafer, 10,000 JJs would be required for long-range connections. Each of these JJs must be provided with a current bias. While many synaptic connections are local, long-distance connections are paramount, as described in Sec. 2. Systems containing billions of neurons spread across hundreds or thousands of wafers, extending over meters, connected with active, superconducting transmission lines do not appear promising to me. Such a communication network may not be fundamentally impossible, but if the hardware for passive photonic communication proves feasible, scaling to massively interconnected spiking neural systems will be greatly facilitated. Most researchers pursuing superconducting electronics for neural computing are not seeking this scale of system.

4.4 Optoelectronic neural systems

The superconducting optoelectronic approach to large cognitive systems described here utilizes similar superconducting circuits as Refs. 96, 92, and 91 for synaptic, dendritic, and neuronal computation, while leveraging light for communication, seeking scalability to massively interconnected systems. Optoelectronic integration may be most straightforward when combining superconducting circuits with silicon light sources operating at liquid helium temperature.

In addition to contrasting this approach to other existing work in the field, it is necessary to also consider what may seem a more straightforward route to optoelectronic intelligence. This route would involve spiking neurons based on waveguide-integrated light sources, as we have discussed, but instead of SPDs and JJs, semiconductor photodiodes and transistors would be employed. Pursuit of such hardware is impeded by the absence of light sources integrated with transistors. If there were a known

means to integrate light sources as simple as transistors with silicon microelectronics, the landscape of computing would differ radically. Nevertheless, the proposition that superconducting electronics are more promising than photodiodes and MOSFETs for this application requires justification.

Our choice to focus on the superconducting approach is based primarily on three factors. First, superconducting single-photon detectors dramatically reduce the brightness required of the light sources. While semiconducting detectors, such as avalanche photodiodes, can detect a single photon, the energy consumption negates the benefits of single-photon sensitivity in the system application under consideration. For scalable system integration, the semiconductor counterpart to a waveguide-integrated SPD working in conjunction with a JJ is a waveguide-integrated photodiode working in conjunction with a MOSFET. Such a semiconductor receiver is likely to require roughly 1000 photons to charge the capacitance of the MOSFET gate [97] to initiate a synapse event. This factor of 1000 in photon power is matched by the factor of 1000 incurred to cool the superconducting system (see Secs. 5 and 6), so the net power consumption for light generation in semiconductor and superconductor systems is roughly equivalent. Yet the important distinction is that the superconducting system dissipates this power off chip in a cryocooler, whereas the semiconducting system requires the light sources to produce this power in the form of photons. Optoelectronic neural systems leveraging superconductors can make due with light sources providing 10,000 photons within a few tens of nanoseconds (30 nW continuous-wave equivalent), while a semiconducting counterpart will require light sources 1000 times brighter to attain the same firing rate. Achieving the former appears possible with inexpensive silicon light sources, while the latter is likely to require further advances in III-V sources. While exciting progress continues to be made in III-V integration on silicon [98, 99], a central challenge remains to integrate these light sources intimately with electronics. The system under consideration requires fabrication of light sources by the millions across 300-mm wafers, which will surely be more cost effective if silicon devices as simple as transistors can be employed for light emission [22], a possibility that appears more likely with superconducting detectors and low-temperature operation.

The second factor driving our group to pursue the superconducting approach relates to multi-planar wafer-scale integration. Whether semiconductors or superconductors are used, artificial synaptic, dendritic, and neuronal circuits are not small. To accommodate millions of neurons and their synapses on a 300-mm wafer, on the order of 20 planes of photonic waveguides are required for communication, and a similar number of planes of electronic circuits are likely to be advantageous. For each plane of MOSFETs, high-temperature annealing steps are required for dopant activation, leading to processing challenges when integrating with metal wires, photonic waveguides, and light sources. This processing challenge is one reason extension of MOSFET processes to multiple stacked planes of transistors with copper interconnects has been difficult. Power dissipation and heat removal also come into play but may be less consequential in the context of spiking neurons with sparse activity. Superconducting electronic circuits are processed near room temperature, and the prospect of integrating many planes of JJs, SPDs, and waveguides appears to us to be less restrictive. Multiple planes of active SPDs [100] and JJs [101, 102] have been demonstrated.

The third factor steering us toward superconducting electronics relates to memory and learning. For a cognitive system of the scale under consideration, synaptic weight modification must be unsupervised and will be most readily realized if the signals that induce learning functions are the same signals, with the same current, voltage, or light levels, used for computing within neurons, and sent to synapses for communication. With superconducting circuits, single-flux quanta are used for computing, and single photons are used for communication. It appears possible for these same signals to update synaptic weights and enable learning, primarily by adjusting current biases to JJs. A close functional analogy would be to modify the voltage on the gate of a MOSFET in an analog manner, and indeed, this has long been the ambition of floating-gate MOSFETs for synaptic memory [103]. However, the voltages required to change the charge on the gate are much higher than typical voltages used for computation elsewhere within the circuit, making it difficult to implement unsupervised learning based only on the signals already present in the network. These persistent challenges with floating gates have led many to look elsewhere for suitable adaptive circuits [104]. While any one of these approaches may lead to the desired memory operations, it is our perspective that the path to systems with lifelong learning and a multitude of memory mechanisms appear less formidable with Josephson circuits.

Despite these arguments in favor of superconducting electronics, several valid counterpoints can be raised. The requisite silicon light sources remain to be proven. Massively multiplanar fabrication of superconducting optoelectronic wafers is an ambitious technological undertaking. For many readers, the requirement of cryogenic operation is the most disconcerting aspect of the project. Several comments are in order. Low-temperature operation eliminates such systems from consideration for applications that require low system power consumption, such as mobile devices. But for systems with a million neurons, existing cryogenic technologies drawing a kilowatt of wall power are suitable, comparable to a home air conditioner in power consumption and complexity, but with cooling based on the thermodynamic properties of liquid helium. For larger applications, cryogenic operation may prove an insurmountable obstacle, although the scale of cryogenics used in superconducting magnets for particle colliders offers hope. The field of quantum information also provides

an insightful lesson. Many types of qubits require operation at tens of millikelvin, necessitating the extra expense and complexity of dilution refrigerators. The environment at 4 K is comparatively balmy, and the required cryogenics are simpler and less expensive. Quantum information presently enjoys tremendous investment because these systems promise functions not otherwise possible. The same must be true of optoelectronic intelligence if it is to have a future. Anything that can be done with CMOS will be done with CMOS. If SOENs cannot achieve AGI that is otherwise unattainable, they will not be brought into existence. If they can attain unmatched cognition, someone is likely to be willing to pay for them, unless the expense is astronomical. The perspective presented here is that exactly this will come to pass: superconducting optoelectronic hardware will enable AI that simply cannot be achieved through other physical means. Low-temperature operation will be justified by the performance.

The vast majority of the universe is in thermal equilibrium with the cosmic microwave background at 2.7 K, below the proposed operating temperature of SOENs. In such a setting, all system power consumption estimates are reduced by a factor of 1000 from the numbers presented here, and the energy consumption per synaptic operation rivals that of the human brain, while enabling firing rates orders of magnitude faster. We should not expect technological intelligence to share our disposition to an environment where water is liquid; they may prefer to reside in an environment where helium condenses. If our goal is to answer scientific questions regarding the physical limits of cognition, low-temperature operation is not a fundamental impediment.

5 Scaling an optoelectronic system

The physics of light is complementary to that of electrons. Photons can co-propagate on a waveguide independently without capacitance. Waveguides can fan out without a charging penalty due to wiring. This is not to say photonic communication can address an arbitrarily large number of recipients without consequence. For each new recipient, the number of photons in a neuronal pulse must increase. As destinations get further away, more energy is dissipated to propagation loss. These realities notwithstanding, it is feasible for devices communicating with photons to make direct, independent connections to thousands of destinations, thereby eliminating the need for the shared communication infrastructure that is the primary impediment to achieving AGI with electrical interconnections.

Having made this claim, the burden is upon us to provide evidence of the feasibility of photonic communication in large-scale neural systems. The large wavelength of light relative to the size of electronic devices causes concern for the size of optoelectronic brain-scale networks. To build confidence for the feasibility of the endeavor, I sketch here a vision of how such an optoelectronic neural system may be constructed. At the foundation of this vision is the assumption that the technology will utilize the fabrication infrastructure of silicon electronics and photonics in conjunction with fiber optics for longer-range communication.

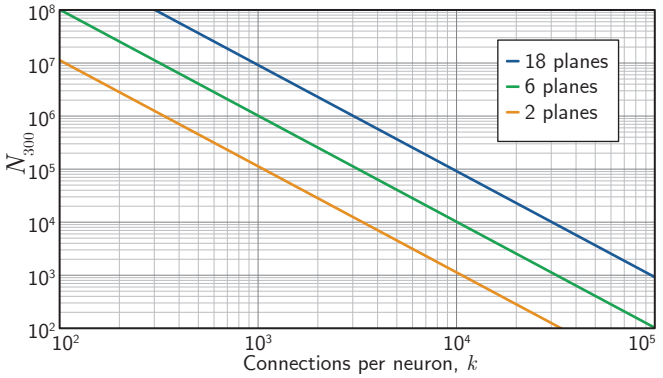

Figure 6: The total number of nodes that can fit on a 300 mm wafer (N_{300}) as a function of the number of connections per node (k) for various numbers of waveguide planes in the wire-limited regime [105].

At the wafer scale, light will be guided in multiple planes of dielectric waveguides [48, 49] (Fig. 7(a)), just as integrated electronics requires multiple wiring layers. To estimate the area of such photonic interconnection networks, we follow Keyes [105] and approximate the number of neurons that can be supported on a 300-mm wafer by $N = 2\sqrt{2}r^2\,(p/wk_{\mathrm{in}})^2$. Here, p is the number of planes of waveguides, w is the waveguide pitch (1.5 µm), k_{in} is the number of waveguides entering the neuron, and $r = 150$ mm. The prefactor results from assuming octagonal tiling. This expression is plotted in Fig. 6. The estimate informs us that a 300-mm wafer with six waveguide planes can support roughly one million neurons if they each have one thousand connections. More involved analysis finds more planes may be needed [36]. As a point of comparison to electrical neural systems, Ref. 106 finds that through multi-layer, wafer-scale integration of logic and memory, 250 million electrical neurons could fit on a 300 mm wafer. The trade-off is speed, as the shared communication network would limit the electrical neurons studied in Ref. 106 to 10 Hz operation when 1000 synaptic connections are made per neuron. Nevertheless, the message of Fig. 6 is that photonic routing results in large area consumption. An optoelectronic brain larger than that of a bumble bee will not fit on a single 300-mm wafer.

Optoelectronic intelligence will require communication between wafers. Wafers can be stacked vertically, and free-space optical links can send photons from a source on one wafer to a detector on a wafer above or below [107], as illustrated in Fig. 7(b). Assuming SPDs receiving vertical communication have a pitch of 25 µm, a 300-mm octagon could support 10^8 vertical communication links between two wafers. Considering wafers as laminar layers, as in

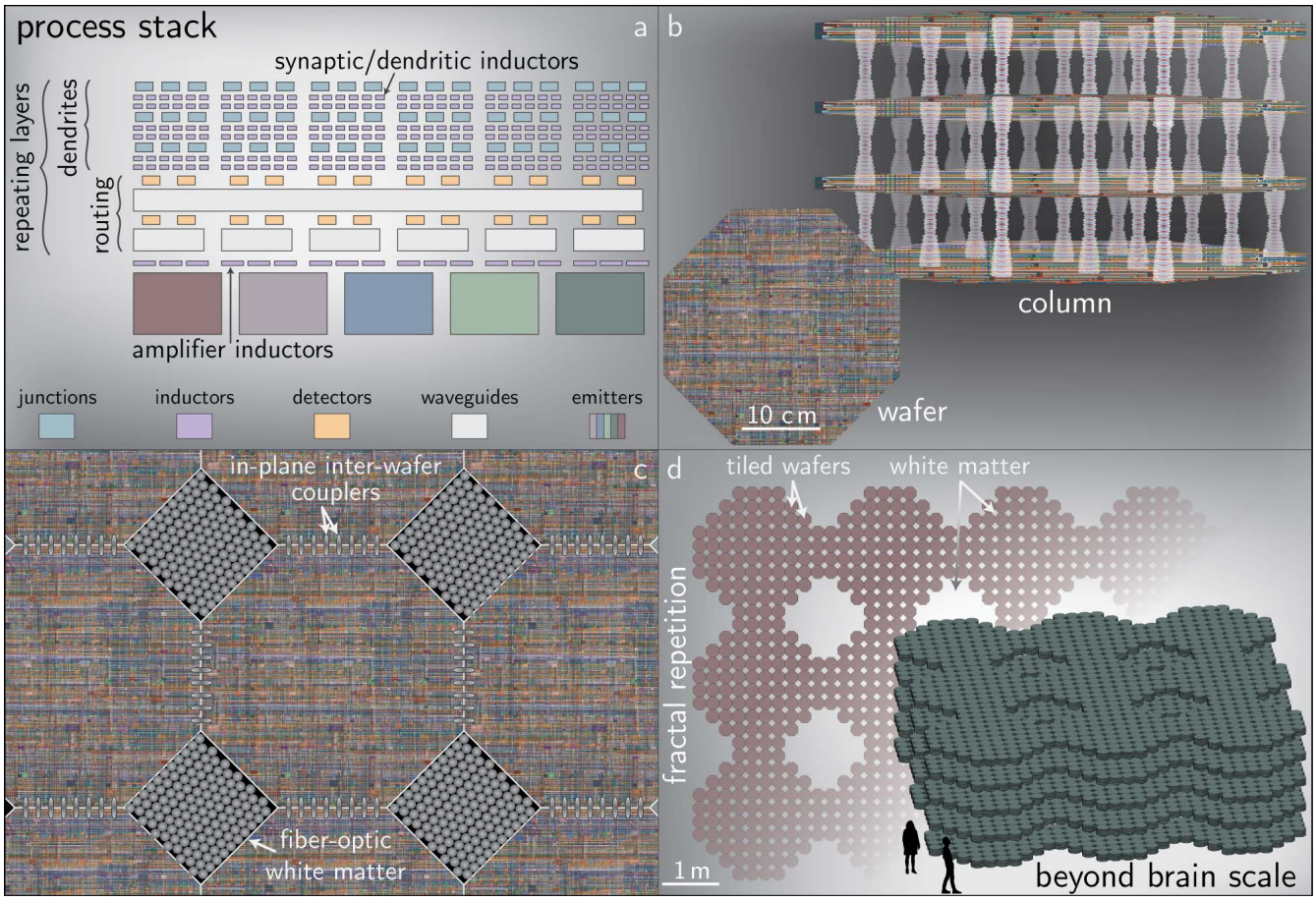

Figure 7: Hierarchical construction of optoelectronic neural systems. (a) Schematic of the process stack, with silicon light sources on a silicon-on-insulator wafer, waveguides and detectors above, followed by the Josephson infrastructure and mutual inductors for dendritic processing. (b) Vertical photonic communication between two stacked wafers. Liquid helium flows between the wafers of a column for cooling, and free-space links propagate without loss through the helium. The inset shows a schematic of a single 300 mm wafer, with neurons and routing, cut into an octagon for tiling. (c) Illustration of in-plane tiling. Lateral wafer-edge links connect wafers in a plane, and fiber optic bundles fill the voids between wafers for long-range communication. (d) A large neural system with multiple large modules, each containing hundred to thousands of wafers, enabled by photonic communication and the efficiency of superconducting detectors and electronics. Not shown is the fiber-optic white matter that would be woven through the voids between the octagons in this example hierarchical tiling.

cortex, such a configuration would result in roughly 5% inter-layer connectivity, similar to the fraction observed in mammals (Ref. 10, pg. 286).

In addition to feed-forward and feed-back free-space vertical coupling, lateral inter-wafer communication can be achieved at wafer edges, as shown in Fig. 7(c). In the tiling considered here, each wafer makes such connections to neighbors in the cardinal directions. With a 10 µm pitch, 11,500 wafer-edge couplers could be supported in each direction. Such a system would demonstrate strong connectivity within the vertical stack of the wafers, and weaker lateral connectivity. The reader may recognize the columnar organization of the cerebral cortex [5].

To achieve communication from within these columns to other regions of the network, optical fibers are ideal. Within the tiling under consideration, the square areas at diagonals between wafers can support fiber-optic bundles (Fig. 7(c)). These optical fiber tracts are analogous to white matter in the brain. One such region could house a million single-mode fibers of 125 µm diameter. These fibers will emanate from all wafers within the column, and if six wafers are stacked in a column, each wafer would have 167,000 output fibers to carry information to other regions. With one million neurons on a wafer, not every neuron would have access to a fiber for long-distance communication (unless wavelength multiplexing is employed). This again is consistent with brain organization, wherein the number of long-distance axons emanating from a region is smaller than the number of neurons within the region. Each of these fibers can branch as it extends through the white matter, so a neuron with access to a single wafer-edge fiber can establish multiple long-range connec-

tions. Recent progress in low-loss fiber-to-waveguide coupling [108] indicates a potential future direction for such integration of fibers with on-chip waveguides, but significant advances in manufacturing are required to realize the coupling of dense fiber bundles to 300-mm wafers.

With this columnar configuration in mind, one can assess the feasibility of constructing a system on the scale of the human cerebral cortex (10 billion neurons, each with thousands of synaptic connections). If a wafer holds a million neurons, a cortex-scale assembly requires 10,000 wafers. Assuming the volume of white matter scales as the volume of grey matter to the 4/3 power [109], the cortex-scale system would fit in a volume two meters on a side. While optoelectronic neurons are significantly bigger than their biological counterparts, it is not the absolute size that limits system performance. The relevant quantity for assessing scaling limitations is the size of a neuron divided by the velocity of communication [36]. Communication at the highest velocity in the universe more than compensates the large device size.

Regarding power, a single 300-mm wafer with a million neurons would dissipate one watt if the light production efficiency were $\eta = 10^{-4}$, a conservative estimate. For the cortex-scale system of 10,000 wafers, the device power consumption with $\eta = 10^{-4}$ would be 10 kW. A further cooling-power penalty of one thousand would be incurred if the system were operated in a background of 300 K. Thus, even in a conservative case of poor light production efficiency, an AGI on the scale of the human brain would consume 10 MW, the same order as a modern supercomputer. We are considering a system with roughly the same number of neurons and synapses as the human cerebral cortex, but with activity at 30,000 times the speed. While there is high uncertainty associated with scaling estimates of such an immature technology, these calculations indicate that artificial brain-scale systems with photonic communication and electronic computation may be feasible, a possibility with profound implications for the future of science and technology.

6 Summary and Discussion

I have argued that artificial neural hardware should be designed and constructed to leverage photonic communication while performing synaptic, dendritic, and neuronal functions with electronic devices. Superconducting optoelectronic circuits elegantly implement these functions, in part because of the utility of Josephson nonlinearities for neural computation, and also because superconducting detectors enable few-photon signals, approaching the lowest possible energy for optical communication. We have demonstrated all of the core components and are working toward complete integration.

The approach to optoelectronic hardware described here is not without limits, and different factors limit performance at different scales. Regarding speed, the synaptic response is limited by the reset time of the SPD, which is between 10 ns and 50 ns depending on the material used. A response time of 50 ns limits the maximum neuronal firing frequency of the neuron to 20 MHz. For the silicon light sources we have primarily been pursuing, the emitter lifetime is on the order of 40 ns [51], giving a maximum firing frequency comparable to the 20 MHz figure determined by the speed of SPDs. In biological neural systems, conduction delays are an important factor limiting speed. Using light for communication greatly alleviates this concern, yet there does exist a scale at which the speed of light becomes the limiting factor. Within the 50 ns reset time of the SPD or the comparable 40 ns lifetime of the silicon emitters, light can travel 10 m in fiber. A system of this linear extent would contain at least an order of magnitude more neurons than an entire human brain. The scale set by this speed limit does not represent the maximum possible scale of an optoelectronic neural system, but rather the maximum possible volume of neurons that can communicate within the highest frequency oscillations of the system.

Regarding power consumption, cryogenic cooling plays a key role. The power required for cooling contains two contributions: the base-level power required to keep the environment below the superconducting transition temperature, even when the devices are inactive, and the additional cooling power required to remove excess heat generated by the activity of the circuits. The first factor is a few hundred watts for small systems, while the second factor is typically about one kilowatt of extra cooling power per watt of power dissipated by the devices. For small systems comprising a few thousand neurons each with a few hundred synapses on a $1\,\text{cm} \times 1\,\text{cm}$ die, the devices will dissipate around a milliwatt [36], so the first factor dwarfs the second. The second factor does match the first until intermediate-scale systems with tens to hundreds of interconnected wafers, each dissipating 1 W when active. It is somewhere between the scale of a few thousand neurons on a die and a few million neurons interconnected across several wafers that we expect the performance of the system to exceed what can be accomplished without photonic communication and superconducting electronic computation. For large systems in excess of hundreds of interconnected wafers, the power dissipated by the active devices on the wafer and the associated cooling costs dominate. The power consumed by each wafer contains contributions from light sources, detectors at synapses, and JJs performing computations within dendrites and neurons. If light sources can be realized with 1% efficiency, each of these circuit components will contribute nearly equally to the total system power consumption [43].

Despite these limits, this approach to AGI appears possible for physical and practical reasons. Physically, due to photonic signaling, it is possible to achieve efficient communication across the network for systems with orders of magnitude more than the 10,000 wafers comprising a brain-scale system. Reference 36 explores the

communication-limited size of the system as a function of the frequency of network oscillations. Specialized processors with activity at 20 MHz (the gamma firing rate of loop neurons) can span an area 10 meters on a side before delays limit communication. Modules with activity at 1 MHz (the frequency of corresponding theta oscillations in this system) could integrate information across an area the size of a data center within a single theta cycle.

On the practical side, fabrication of SOENs at industrial scale appears feasible. All the proposed circuits can be created on 300-mm wafers with existing infrastructure, such as a 45-nm CMOS node. Ten thousand wafers move through such a foundry every day. If dedicated to fabrication of optoelectronic intelligence, a foundry could produce multiple brain-scale systems per year. While the devices employed here depart from conventional silicon microelectronics, the same fabrication infrastructure can be employed.

What are the next steps to realize loop neurons and SOENs? Low-cost source-detector integration at the wafer scale is required. Demonstration of requisite plasticity functions would be an important milestone. Multiplanar integration of superconducting electronics would further build momentum. Active devices must be augmented with improvements in deposited dielectrics to enable many planes of routing waveguides with low loss. Hardware improvements will not lead to AGI without further theoretical insights. Conceptual advances are required to achieve high-performance neural systems, train them, and make them intelligent.

7 Acknowledgements

I acknowledge and appreciate significant contributions to this project from team members Sonia Buckley, Jeff Chiles, Saeed Khan, Adam McCaughan, Bryce Primavera, and Alexander Tait. This work would not be possible without the group leadership of Sae Woo Nam, Richard Mirin, and the institutional support of NIST. I also thank three reviewers for valuable feedback.

Data available from the authors upon request.

References

[1] B.J. Baars. *A cognitive theory of consciousness*. Cambridge University Press, 1988.

[2] S. Dehane. *Consciousness and the brain*. Penguin, 2014.

[3] O. Sporns. *Networks of the Brain*. The MIT Press, Cambridge, Massachusetts, first edition, 2010.

[4] W. Gerstner and W. Kistler. *Spiking neuron models*. Cambridge University Press, Cambridge, first edition, 2002.

[5] V.B. Mountcastle. The columnar organization of the neocortex. *Brain*, 120:701, 1997.

[6] S.L. Bressler and V. Menon. Large-scale brain networks in cognition: emerging methods and principles. *Trends in cognitive sciences*, 14:277, 2010.

[7] L.F. Abbott and W.G. Regehr. Synaptic computation. *Nature Reviews*, 431:796, 2004.

[8] G.J. Stuart and N. Spruston. Dendritic integration: 60 years of progress. *Nature Neuroscience*, 18:1713, 2015.

[9] C. Koch. Computation and the single neuron. *Nature*, 385:207, 1997.

[10] G. Buzsaki. *Rhythms of the Brain*. Oxford University Press, 2006.

[11] C. Sun, M.T. Wade, Y. Lee, J.S. Orcutt, L. Alloatti, M.S. Georgas, A.S. Waterman, J.M. Shainline, R.R. Avizienis, S. Lin, B.R. Moss, R. Kumar, F. Pavanello, A.H. Atabaki, H.M. Cook, A.J. Ou, J.C. Leu, Y.-H. Chen, K. Asanović, R.J. Ram, M.A. Popović, and V.M. Stojanović. Single-chip microprocessor that communicates directly using light. *Nature*, 528:534, 2015.

[12] K.A. Boahen. Point-to-point connectivity between neuromorphic chips using address events. *IEEE Tran. Circ. Sys. II*, 47:416, 2000.

[13] T. Pfeil, A. Grubl, and K. Meier. Six networks on a universal neuromorphic computing substrate. *Frontiers in Neuroscience*, 7:11, 2013.

[14] 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.

[15] S.B. Furber, F. Galluppi, S. Temple, and L.A. Plana. The spinnaker project. *Proceedings of the IEEE*, 102:652, 2014.

[16] J. Park, T. Yu, S. Joshi, C. Maier, and G. Cauwenberghs. Hierarchical address event routing for reconfigurable large-scale neuromorphic systems. *IEEE Trans. Neural Networks and Learning Systems*, 28:2408, 2017.

[17] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S.H. Choday, and G. Dimou. Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*, 1:82, 2018.

[18] D. Liang and J.E. Bowers. Recent progress in lasers on silicon. *Nature Photonics*, 4:511, 2010.

[19] Z. Zhou, B. Yin, and Jurgen Michel. On-chip light sources for silicon photonics. *Light: science and applications*, 4:e358, 2015.

[20] J.M. Shainline, S.M. Buckley, R.P. Mirin, and S.W. Nam. Superconducting optoelectronic circuits for neuromorphic computing. *Phys. Rev. App.*, 7:034013, 2017.

[21] J.M. Shainline, S.M. Buckley, A.N. McCaughan, J. Chiles, A. Jafari-Salim, R.P. Mirin, and S.W. Nam. Circuit designs for superconducting optoelectronic loop neurons. *J. Appl. Phys.*, 124:152130, 2018.

[22] S. Buckley, J. Chiles, A.N. McCaughan, G. Moody, K.L. Silverman, M.J. Stevens, R.P. Mirin, S.W. Nam, and J.M. Shainline. All-silicon light-emitting diodes waveguide-integrated with superconducting single-photon detectors. *Appl. Phys. Lett.*, 111:141101, 2017.

[23] R.F. Betzel and D.S. Bassett. Multi-scale brain networks. *NeuroImage*, 160:73, 2017.

[24] A.N. Khambhati, M.G. Mattar, N.F. Wymbs, S.T. Grafton, and D.S. Bassett. Beyond modularity: Fine-scale mechanisms and rules for brain network configuration. *NeuroImage*, 166:385, 2018.

[25] G. Buzsaki and A. Draguhn. Neuronal oscillations in cortical networks. *Science*, 304:1926, 2004.

[26] R.T. Canolty, E. Edwards, S.S. Dalal, M. Soltani, S.S. Nagarajan, H.E. Kirsch, M.S. Berger, N.M. Barbaro, and R.T. Knight. High Gamma Power is Phase-Locked to Theta Oscillations in Human Neocortex. *Science*, 313:1626, 2006.

[27] O. Jensen and L.L. Colgin. Cross-frequency coupling between neuronal oscillations. *Trends in Cognitive Sciences*, 11:268, 2007.

[28] A. von Stein and J. Sarnthein. Different frequencies for different scales of cortical integration: from local gamma to long range alpha/theta synchronization. *Int. J. Psychophysiology*, 38:301, 2000.

[29] P. Fries. Rhythms for Cognition: Communication Through Coherence. *Neuron*, 88:220, 2015.

[30] G. Buzsaki. *The Brain from Inside Out.* Oxford University Press, 2019.

[31] D. Meunier, R. Lambiotte, and E.T. Bullmore. Modular and hierarchically modular organization of brain networks. *Frontiers in Neuroscience*, 4:200, 2010.

[32] D. Plenz and T.C. Thiagarajan. The organizing principles of neuronal avalanches: cell assemblies in the cortex? *Trends in Neuroscience*, 30:101, 2006.

[33] D.S. Bassett, D.L. Greenfield, A. Meyer-Lindenberg, D.R. Weinberger, S.W. Moore, and E.T. Bullmore. Efficient physical embedding of topologically complex information processing networks in brains and computer circuits. *PLoS Computational Biology*, 6:1, 2010.

[34] M.M. Sperry, Q.K. Telesford, F. Klimm, and D.S. Bassett. Rentian scaling for the measurement of optimal embedding of complex networks into physical space. *J. Complex Networks*, 5:199, 2017.

[35] H.G. Schnack, N.E.M. van Haren, R.M. Brouwer, A. Evans, S. Durston, D.I. Boomsma, R.S. Kahn, and H.E. Hulshoff Pol. Changes in thickness and surface area of the human cortex and their relationship with intelligence. *Cerebral Cortex*, 25:1608, 2014.

[36] J.M. Shainline, S.M. Buckley, A.N. McCaughan, J. Chiles, A. Jafari-Salim, M. Castellanos-Beltran, C.A. Donnelly, M.L. Schneider, R.P. Mirin, and S.W. Nam. Superconducting optoelectronic loop neurons. *J. Appl. Phys.*, 126:044902, 2019.

[37] C. Koch, M. Rapp, and I. Segev. A Brief History of Time (constants). *Cerebral Cortex*, 6:93, 1996.

[38] E. Estrada and P.A. Knight. *A First Course in Network Theory.* Oxford, Oxford, United Kingdom, first edition, 2015.

[39] D.J. Watts and S.H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440, 1998.

[40] H. Markram, W. Gerstner, and P.J. Sjostrom. Spike-timing-dependent plasticity: a comprehensive overview. *Frontiers in Synaptic Neuroscience*, 4:2, 2012.

[41] W.C. Abraham. Metaplasticity: tuning synapses and networks for plasticity. *Nature Neuroscience*, 9:387, 2008.

[42] S. Fusi, P.J. Drew, and L.F. Abbott. Casdcade models of synaptically stored memories. *Neuron*, 45:599, 2005.

[43] J.M. Shainline. Fluxonic processing of photonic synapse events. *IEEE J. Sel. Top. Quant. Electron.*, 26:7700315, 2019.

[44] 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. Photon.*, 7:210, 2013.

[45] T. Van Duzer and C.W. Turner. *Principles of superconductive devices and circuits*. Prentice Hall, USA, second edition, 1998.

[46] Alan M. Kadin. *Introduction to superconducting circuits*. John Wiley and Sons, USA, first edition, 1999.

[47] A.N. McCaughan, V.B. Verma, S.M. Buckley, A.N. Tait, S.W. Nam, and J.M. Shainline. A superconducting thermal switch with ultrahigh impedance for interfacing superconductors to semiconductors. *Nature Electronics*, 2:451, 2019.

[48] J. Chiles, S. Buckley, N. Nader, S.W. Nam, R.P. Mirin, and J.M. Shainline. Multi-planar amorphous silicon photonics with compact interplanar couplers, cross talk mitigation, and low crossing loss. *APL Photonics*, 2:116101, 2017.

[49] J. Chiles, S.M. Buckley, S.W. Nam, R.P. Mirin, and J.M. Shainline. Design, fabrication, and metrology of 10 x 100 multi-planar integrated photonic routing manifolds for neural networks. *APL Photonics*, page 106101, 2018.

[50] S. Buckley, A.N. Tait, J. Chiles, A.N. McCaughan, S. Khan, R.P. Mirin, S.W. Nam, and J.M. Shainline. Integrated-Photonic Characterization of Single-Photon Detectors for Use in Neuromorphic Synapses. *Phys. Rev. Applied*, 14:054008, 2020.

[51] S. Buckley, A.N. Tait, G. Moody, B. Primavera, S. Olsen, J. Herman, K.L. Silverman, S. Papa Rao, S.W. Nam, R.P. Mirin, and J.M. Shainline. Optimization of photoluminescence from W centers in silicon-on-insulator. *Opt. Express*, 29:16057, 2020.

[52] C.D. Schuman, T.E. Potok, R.M. Patton, J.D. Birdwell, M.E. Dean, G.S. Rose, and J.S. Plank. A survey of neuromorphic computing and neural networks in hardware. *arXiv*, page arXiv:1705.06963v1, 2017.

[53] K. Berggren, Q. Xia, K.K. Likharev, D.B. Strukov, H. Jiang, T. Mikolajick, D. Querlioz, M. Salinga, J.R. Erickson, and S. Pi et al. Roadmap on emerging hardware and technology for machine learning. *Nanotechnology*, 32:012002, 2020.

[54] B.J. Shastri, A.N. Tait, T. Ferreira de Lima, W.H.P. Pernice, H. Bhaskaran, C.D. Wright, and P.R. Prucnal. Photonics for artificial intelligence and neuromorphic computing. *Nat. Photonics*, 15:102, 2021.

[55] C. Mead. *Analog VLSI and Neural Systems*. Addison Wesley.

[56] C. Mead. Neuromorphic Electronic Systems. *Proc. IEEE*, 78:1629, 1990.

[57] S.-C. Liu, T. Delbruck, G. Indiveri, A. Whatley, and R. Douglas, editors. *Event-based neuromorphic systems*. John Wiley and Sons, 2015.

[58] R.J. Vogelstein, U. Mallik, J.T. Vogelstein, and G. Cauwenberghs. Dynamically reconfigurable silicon array of spiking neurons with conductance-based synapses. *IEEE Trans. Neural Networks*, 18:253, 2007.

[59] G. Indiveri, B. Linares-Barranco, T.J. Hamilton, A. van Schaik, R. Etienne-Cummings, T. Delbruck, S.-C. Liu, P. Dudek, P. H afliger, S. Renaud, J. Schemmel, G. Cauwenberghs, J. Arthur, K. Hynna, F. Flolwosele, S. Saighi, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang, and K. Boahen. Neuromorphic silicon neuron circuits. *Front. Neurosci.*, 73:73, 2011.

[60] E. Chicca, F. Stefanini, C. Bartolozzi, and G. Indiveri. Neuromorphic Electronic Circuits for Building Autonomous Cognitive Systems. *Proc. IEEE*, 102:1367, 2014.

[61] S.A. Aamir, Y. Stradmann, P. Müller, C. Pehle, A. Hartel, A. Grübl, J. Schemmel, and K. Meier. An Accelerated LIF Neuronal Network Array for Large-Scale Mixed-Signal Neuromorphic Architecture. *IEEE Trans. Circuits Sys.*, 65:4299, 2018.

[62] P.A. Bogdan, A.G.D. Rowler, O. Rhodes, and S.B. Furber. Structural Plasticity on the SpiNNaker Many-Core Neuromorphic System. *Front. Neurosci.*, 12:434, 2018.

[63] D. Strukov, G. Indiveri, J. Grollier, and S. Fusi. Building Brain-Inspired Computing. *Nat. Comm.*, 10:4838, 2019.

[64] J.L. Hennessy and D.A. Patterson. *Computer Architecture*. Elsevier. Appendix F.

[65] D. Psaltis and N. Farhat. Optical information processing based on an associative-memory model of neural nets with thresholding and feedback. *Opt. Lett.*, 10:98, 1985.

[66] N.H. Farhat, D. Psaltis, A. Prata, and E. Paek. Optical implementation of the Hopfield model. *Applied Optics*, 24:1469, 1985.

[67] J.J. Hopfield. Neural networks and physical systems with emergent computational abilities. *PNAS*, 79:2554, 1982.

[68] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund. Large-Scale Optical Neural Networks Based on Photoelectric Multiplication. *Phys. Rev. X*, 9:021032, 2019.

[69] P.R. Prucnal and B.J. Shastri. *Neuromorphic photonics*. CRC Press, New York, first edition, 2017.

[70] Y. Shen, N.C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljacic. Deep learning with coherent nanophotonic circuits. *Nature Photonics*, 11:441, 2016.

[71] K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman. Experimental demonstration of reservoir computing on a silicon photonics chip. *Nat. Comm.*, 5:3541, 2014.

[72] A.N. Tait, T. Ferreira de Lima, M.A. Nahmias, H.B. Miller, H.-T. Peng, B.J. Shastri, and P.R. Prucnal. Silicon photonic modulator neuron. *Phys. Rev. Applied*, 11:064043, Jun 2019.

[73] 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.*, 32:3427, 2014.

[74] M. Lipson. Guiding, Modulating, and Emitting Light on Silicon-Challenges and Opportunities. *J. Lightwave Technology*, 23:4222, 2005.

[75] K. Preston, N. Sherwood-Droz, J.S. Levy, and M. Lipson. Performance Guidelines for WDM Interconnects Based on Silicon Microring Resonators. In *Conference on Lasers and Electrooptics*, page CThP4. Optical Society of America, 2011.

[76] I. Chakraborty, G. Saha, A. Sengupta, and K. Roy. Toward fast neural computing using all-photonic phase change spiking neurons. *Scientific Reports*, 8:12980, 2018.

[77] Z. Cheng, C. Rios, W.H.P. Pernice, C.D. Wright, and H. Bhaskaran. On-chip photonic synapse. *Science Advances*, 3:1700160, 2017.

[78] T.W. Hughes, M. Minkov, Y. Shi, and S. Fan. Training of photonic neural networks through in situ backpropagation and gradient descent. *Optica*, 5:864, 2018.

[79] K.-I. Funahashi and Y. Nakamura. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, 6:801, 1993.

[80] K. Vandoorne, J. Dambre, D. Verstraeten, B. Schrauwen, and P. Bienstman. Parallel reservoir computing using optical amplifiers. *IEEE Trans. Neural Networks*, 22:1469, 2011.

[81] S. Ortín, M.C. Soriano, L. Pesquera, D. Brunner, D. San-Martín, I. Fischer, C.R. Mirasso, and J.M. Gutiérrez. A unified framework for reservoir computing and extreme learning machines based on a single time-delayed neuron. *Sci. Rep.*, 5:14945, 2015.

[82] G. Van der Sande, D. Brunner, and M.C. Soriano. Advances in photonic reservoir computing. *Nanophotonics*, 6:561, 2017.

[83] D. Brunner, B. Penkovsky, B.A. Marquez, M. Jacquot, I. Fischer, and L. Larger. Tutorial: Photonic neural networks in delay systems. *J. Appl. Phys.*, 124:152004, 2018.

[84] J.E. Lisman. Bursts as a unit of neural information: making unreliable synapses reliable. *Trends Neurosci.*, 20:38, 1997.

[85] Y. Harada and E. Goto. Artificial neural network circuits with josephson devices. *IEEE Trans. Magnetics*, 27:2863, 1991.

[86] M. Hidaka and L.A. Akers. An artificial neural cell implemented with superconducting circuits. *Supercond. Sci. Technol.*, 4:654, 1991.

[87] A.E. Schegolev, N.V. Klenov, I.I. Soloviev, and M.V. Tereshonok. Adiabatic superconducting cells for ultra-low-power artificial neural networks. *Beilstein Journal of Nanotechnology*, 7:1397, 2016.

[88] N.V. Klenov, A.E. Schegolev, I.I. Soloviev, S.V. Bakurskiy, and M.V. Tereshonok. Energy efficient superconducting neural networks for high-speed intellectual data processing systems. *IEEE. Trans. Appl. Supercond.*, 28:1301006, 2018.

[89] I.I. Soloviev, A.E. Schegolev, N.V. Klenov, S.V. Bakurskiy, M.Y. Kupriyanov, M.V. Tereshonok, A.V. Shadrin, V.S. Stolyarov, and A.A. Golubov. Adiabatic superconducting artificial neural network: basic cells. *J. Appl. Phys.*, 124:152113, 2018.

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

[91] M.L. Schneider, C.A. Donnelly, S.E. Russek, B. Baek, M.R. Pufall, P.F. Hopkins, P. Dresselhaus, S.P. Benz, and W.H. Rippard. Ultralow power artificial synapses using nanotextured magnetic josephson junctions. *Science Advances*, 4:1701329, 2018.

[92] K. Segall, M. LeGro, S. Kaplan, O. Svitelskiy, S. Khadka, P. Crotty, and D. Schult. Synchronization dynamics on the picosecond time scale in coupled josephson junction networks. *Physical Review E*, 95:032220, 2017.

[93] S.G. Kim, J.S. Han, H. Kim, S.Y. Kim, and H.W. Jang. Recent Advances in Memristive Materials for Artificial Synapses. *Advanced Materials Technologies*, 3:1800457, 2018.

[94] M.L. Schneider, C.A. Donnelly, and S.E. Russek. Tutorial: high-speed low-power neuromorphic systems based on magnetic josephson junctions. *J. Appl. Phys.*, 124:161102, 2018.

[95] M.L. Schneider and K. Segall. Fan-out and Fan-in Properties of Superconducting Neuromorphic Circuits. *J. Appl. Phys.*, 128:214903, 2020.

[96] T. Hirose, T. Asai, and Y. Amemiya. Pulsed neural networks consisting of single-flux-quantum spiking neurons. *Physica C*, 463:1072, 2007.

[97] D.A.B. Miller. Attojoule optoelectronics for low-energy information processing and communications. *J. Lightwave Technol.*, 35:346, 2017.

[98] M. Tand, J.-S. Park, Z. Wang, S. Chen, P. Jurczak, A. Seeds, and H. Liu. Integration of III-V Lasers on Si for Si Photonics. *Prog. Quant. Electron.*, 66:1, 2019.

[99] Y. Han and K.M. Lau. III-V lasers selectively grown on (001) silicon. *J. Appl. Phys.*, 128:200901, 2020.

[100] V.B. Verma, F. Marsili, S. Harrington, A.E. Lita, R.P. Mirin, and S.W. Nam. A three-dimensional, polarization-insensitive superconducting nanowire avalanche photodetector. *Appl. Phys. Lett.*, 101:251114, 2012.

[101] S.K. Tolpygo, V. Bolkhovsky, R. Rastogi, S. Zarr, A.L. Day, E. Golden, T.J. Weir, A. Wynn, and L.M. Johnson. Planarized Fabrication Process With Two Layers of SIS Josephson Junctions and Integration of SIS and SFS π-Junctions. *IEEE Trans. Appl. Supercond.*, 29:1101208, 2019.

[102] T. Ando, S. Nagasawa, N. Takeuchi, N. Tsuji, F. China, M. Hidaka, Y. Yamanashi, and N. Yoshikawa. Three-dimensional adiabatic quantum-flux-parametron fabricated using a double-active-layered niobium process. *Supercond. Sci. Technol.*, 30:075003, 2017.

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

[104] N.K. Upadhyay, H. Jiang, Z. Wang, S. Asapu, Q. Xia, and J.J. Yang. Emerging Memory Devices for Neuromorphic Computing. *Advanced Materials Technologies*, 4:1800589, 2019.

[105] R.W. Keyes. The wire-limited logic chip. *IEEE J. Sol.-Sta. Circuits*, SC-17:1232, 1982.

[106] A. Kumar, Z. Wan, W.W. Wilcke, and S.S. Iyer. Toward human-scale brain computing using 3d wafer scale integration. *ACM Journal on Emerging Technologies in Computing Systems*, 13:45, 2017.

[107] M. Cabezon, I. Garces, A. Villafranca, J. Pozo, P. Kumar, and A. Kazmierczak. Silicon-on-insulator chip-to-chip coupling via out-of-plane vertical grating couplers. *Applied Optics*, 51:8090, 2012.

[108] S. Khan, S.M. Buckley, J. Chiles, R.P. Mirin, S.W. Nam, and J.M. Shainline. Low-loss, high-bandwidth fiber-to-chip coupling using capped adiabatic tapered fibers. *Appl. Phys. Lett. Photonics*, 5:056101, 2020.

[109] K. Zhang and T.J. Sejnowski. A universal scaling law between gray matter and white matter in cerebral cortex. *PNAS*, 97:5621, 2000.