Optoelectronic Neural Systems

Device and Architecture Considerations for General Intelligence

Jeffrey M. Shainline[1]

[1] *National Institute of Standards and Technology, Boulder, CO, 80305*

March 1, 2019

Abstract

Contents

1 Introduction

Optoelectronic neural systems reside at the confluence of multiple disciplines. The subject is derived in large part from the foundations of communication theory and computation, as established by Turing [1], von Neumann [2], Shannon [3] and others. Yet the nature of information processing in neural systems is not wholly captured by the mathematical analysis formalized to describe serial communication and computation of digital signals. Concepts from neuroscience dating back to Ramon y Cajal [], Mountcastle [], Edeleman [], and many others during the flourishing of the last 40 years indicate that clean, serial data streams have much to gain from the network concepts of neural systems with complex spatial and temporal behavior, particularly if one seeks a machine that can think like an intelligent being. If one seeks complexity of performance, one must provide complexity in hardware. If one seeks comprehension in the face of ambiguity, the computation must be able to handle shades of gray. Probability and statistics are inherent to neural systems.

Conceptually, neural systems contain threads of communication theory, digital logic, probability and statistics. In practice, these systems combine the physics of many devices. The systems envisioned here utilize electrons to compute and photons to communicate. They leverage semiconductors to make light and superconductors to compute. At this confluence of physics and computing, it is possible to conceive of systems with complexity from the chip scale to the globe, employing the logic of neural systems for cognitive information processing across broad reaches of space and time.

Why do this now? Why this way? Since the inception of the EDVAC, its limitations were anticipated []. Yet the microelectronic march delayed the consequences by many decades. The scaling first charted by Moore [] left little impetus for revolutionary concepts. Now that scaling is reaching physical limitations [], there is an appetite for what new hardware may have to offer. And as computing machines have become deeply integrated in society, we are ready for a sea change in what we ask our machines to do. They can answer any question of fact, but that is no longer all we ask of them. We seek intelligent machines, computers that think, not superficially, but with as deep a wisdom as we can manage to construct. That is why we seek to combine the strengths of superconductors and light. In conjunction, we think, they will lead to the most intelligent machines. To put this discussion in context, we must revisit the origins of computing.

2 Historical context

Much of nature is best described by analog values. The radius of a planet's orbit can take any value across a broad range. The light output from a star is a smoothly varying function. Yet upon close inspection, that light is discrete. A photon is either there, or it is not. The ancient Chinese saw such dichotomies as central to the balance of nature, represented symbolically in the yin-yang. This concept led them to invent binary arithmetic as a representation of the interaction between mutually interdependent opposites [**?**]. Francis Bacon extended these ideas in 1623 to establish that two symbols were sufficient for encoding any communication [4], a concept that was matured by Claude Shannon during and after WWII [3]. Bacon was also well aware 400 years ago that optical communication was advantageous for such communication, although he probably did not anticipate the complex fiber-optic networks that now span the globe. Leibniz further developed these ideas [**?**] with the intention to utilize binary arithmetic for computing [**?**, 4]. In 1679, Leibniz proposed a mechanical apparatus for digital computing based on marbles passing through holes to perform logical operations.

From the earliest intellectual explorations into the digital domain, it has been clear that one of the great strengths of digital information processing is the simplicity of the binary format. Information provides the ability to answer questions, and a bit allows one to answer a single question. This or that? Yes or no? Yet despite the simplicity, a more rigorous mathematical foundation (provided by Alan Turing) as well as the motivations of war would be required before the Leibniz's vision of a digital computer would be realized.

In 1936, Turing published his seminal work "On Computable Numbers" [1], wherein he introduced the concept of a universal computing machine that, provided sufficient time and memory, could execute any computation that is possible, in principle, within the framework established by Gödel []. Turing's contribution was of fundamental mathematical importance in that it further strengthened the arguments made by Gödel regarding the existence of unsolvable problems (the "entscheidungsproblem", as laid out by Hilbert), but Turing's contribution was also of practical importance in that it provided a specific vision for the realization of a universal computer. Turing would go on to design a hardware manifestation of such an apparatus, but his system would never be built. Like the ancient Chinese, Bacon, Leibniz, Babbage, and many others to come, Turing was convinced of the necessity of implementing computation with binary symbols. As he stated in 1947, "Being digital should be of more interest than being electronic." [5]

Turing's work on decrypting messages encoded by the enigma machine during WWII led to advances in computing hardware, in part by leading to the construction of Colossus, a British machine used for analyzing wartime communications, but also by the influence Turing's ideas had on John von Neumann. Von Neumann appreciated the significance of Turing's universal computer, both for the insights regarding issues of completeness raised by Gödel, and also for the utility of enabling a single machine to be capable of performing any possible calculation. During the war, von Neumann found it imperative that the

United States develop an atomic weapon, and he appreciated the necessity of numerically modeling various aspects of the detonation of nuclear weapons. Beginning with the Manhattan Project in Los Alamos, and continuing at the Institute for Advanced Study at Princeton after the war, von Neumann was dedicated to creation of universal calculators in the mold of a Turing machine, the objective being to numerically analyze arbitrary differential equations, represented as difference equations with binary arithmetic. This objective led to the construction of apparatus with a centralized computing unit that could follow arbitrary (clearly articulated) instructions, and could read and store information in a separate memory unit. The memory unit could contain initial conditions, the results of calculations, and instructions that could be modified during the execution of the computation.

In 1953, a machine at the Institute for Advanced Studies employing this "von Neumann" architecture was primarily used for five types of calculations: nuclear explosions, shock waves, meterology, biological evolution, and stellar evolution. Already at that time, the universal machine was analyzing problems across 25 orders of magnitude in time and roughly the same number in space. During the subsequent 70 years, the von Neumann architecture has been used to answer questions regarding essentially every conceivable subject of any concern whatsoever to human beings. This trajectory represents tremendous success regarding von Neumann's initial goal. Nearly all modern computing is based on digital (binary) information processed in a von Neumann architecture, a scheme devised as a means to realize a Turing machine in electronic hardware.

Over the course of this time, hardware for computation has shifted from vacuum tubes, in which information is represented by the deflection of an electron beam under the influence of a voltage, to silicon microelectronics, in which information is represented by the flow of electrical current through a transistor under the influence of a voltage. This evolution of hardware has been perhaps as important as the original conception of the architecture in terms of enabling sustained technological evolution. While hardware has advanced tremendously, the underlying concepts regarding the form of computation performed have remained relatively static. In particular, two key traits of the von Neumann architecture have descended to this day from Turing ancestry: serial information processing, and separation of computation and memory.

John von Neumann would likely be surprised by the constancy of the architecture. As he stated in 1949, "There is reason to suspect that our predilection for linear codes, which have a simple, almost temporal sequence, is chiefly a literary habit, corresponding to our not particularly high level of combinatorial cleverness, and that a very efficient language would probably depart from linearity." [6] Von Neumann did not see the sequential processing performed by Turing's original conceptual apparatus as an ideal way to go about computing, but rather as a convenient way to get started, and certainly a useful tool for numerical investigation. Other limitations of the Turing machine and von Neumann architecture were also well known shortly after their conceptions. Julian Bigelow, the chief engineer of the Electronic Computer Project at the Institute for Advanced studies, articulated that, "If you actually tried to build a machine the way Turing described it, you would spend more time rushing back and forth to find places on a tape than you would doing actual numerical work or computation." Regarding the von Neumann architecture as implemented with vacuum tubes, Bigelow commented, "The design of an electronic calculating machine...turns out to be a frustrating wrestling-match with problems of interconnectivity and proximity in three dimensions of space and one dimension of time." Integrated silicon microelectronics enabled tremendous advances in computer, in large part because of the ability of lithographically defined wires to achieve extraordinary interconnectivity. Yet the challenges of routing information through space and time still form the core motivation for pursuing neural systems with optical interconnectivity.

Beyond hardware constraints, those at the dawn of electronic computation identified logical limitations as well. Describing generations of computers following the EDVAC, Bigelow stated, "The modern high speed computer, impressive as its performance is, from the point of view of getting the available logical equipment adequately engaged in the computation, is very inefficient indeed." Von Neumann went further, beginning to articulate his vision for new types of logic for computation. He claimed, "Whatever language the central nervous system is using, it is characterized by less logical and arithmetical depth than what we are normally used to." [7] Von Neumann anticipated the highly parallel computation performed by neural systems, and was likely influenced by the work of McCulloch and Pitts nearly a decade prior. In 1943, they showed that perceptrons (devices with computational properties loosely based on the nonlinear behavior of neurons in the brain) could be used to represent any logical expression, much like universal computation performed by a Turing machine [8].

While some were conceiving of how to utilize principles of neural information processing for computation, Alan Turing was contemplating whether any artificial machine could eventually embody the intelligence demonstrated in the brain. In 1950, Turing published "Computing Machinery and Intelligence", in which he argued that artificial intelligence should indeed be possible, and he offered a means to determine if it had been achieved, now referred to as the Turing Test [9]. After 70 years of development of hardware, with devices now defined at the nanometer scale, and architectures such as RISC-V that standardize instructions across Turing machines, silicon microelectronic hardware still appears far from achieving general intelligence. Turing was interested from the beginning in constructing a machine that could think, but the Turing Machine as originally conceived is not adept at the task. A silicon digital supercomputer can provide the answer to very hard math problems, and it can search databases, route information over switching networks, and control robots and automobiles. But such an architecture is not conducive to forming a broad concept of what is going on in the world. The digital system struggles with context, with the inter-relations between quantities. It is the intelligence Turing was attempting to emulate that is most difficult to achieve artificially with the serial operation of the Turing machine.

To achieve the intelligence of a human being, we must depart from the serial operation of a Turing machine, we must step away from the framework of the von Neumann architecture, and we must begin to speak in languages beyond binary. As expressed by von Neumann, "A new, essentially logical, theory is called for in order to understand high-complication automata and, in particular, the central nervous system. It may be, however, that in this process logic will have to undergo a pseudomorphosis to neurology." [10] This statement in 1951 anticipates the development of neuromorphic computing in calling for a form of complex information processing modeled after neural systems. After his untimely death in 1958, little effort was made to pursue this vision. As historian George Dyson explains, "The reliability of monolithic microprocessors and the fidelity of monolithic storage postponed the necessity for this pseudomorphosis far beyond what seemed possible in 1948." [4] The extraordinary success of silicon technology has made innovation in architecture and logic lack urgency.

The aspirations to achieve general intelligence and to incorporate the processes of the brain to enable new forms of computing have persisted. In 1990, Carver Mead introduced the concept of using silicon transistors to perform the operations of neurons [11]. Mead's idea was not to use digital logic to numerically step through the differential equations describing neurons, but rather to use the physics of transistors in the sub-threshold regime to embody an approximation of the dynamics inherent to neuronal operation. Such an approach operates MOSFETs as analog integrators. This idea has been important conceptually to establishing the field of neuromorphic computing, but in practice it has been difficult to achieve high performance, due in part to the challenge of achieving consistent performance from analog devices, the same challenge that led to the victory of digital over analog computing for numerical analysis.

Nevertheless, Mead's perspective marked the beginning of a trend that has been exponential since. The majority of the field of neuromorphic computing at present attempts to combine the silicon hardware that has been so successful with the highly parallel, spike-based computation that is observed in biological neural systems. The goal is to use hardware that can be manufactured economically at large scale to implement the new logic von Neumann anticipated when vacuum tubes were still in use. In this paper we argue that such hardware is not well equipped for neural information processing, but that with a different use of silicon microelectronic manufacturing, new devices and systems can be achieved economically that are tailored to this form of computation. Turing and von Neumann did not live to see the silicon revolution, nor were they privy to insights from modern neuroscience. From the perspective of the present day, we are in a much better position to design technologies that speak the language of the human brain to achieve the vision of machines that think.

Turing/von Neumann/digital: I'm going to give you these instructions and this input. I want you to generate an output that is the one and only (existence and uniqueness theorem) correct answer to a mathematically well-posed question.

brain/AGI: I'm going to give you an enormous amount of information regarding the world and all its parts. I want you to distill the essence, identify the salient relationships, and conceive of it all simultaneously across various spatial and temporal scales. And I want you to explain this world to me.

These are very different objectives. A Turing machine is a universal computer in the sense that if a procedure exists to calculate a given number, a Turing machine can do it. However, two aspects of the Turing model make it inefficient for the latter task. First, information processing is serial. Second, memory must be accessed as an independent step from change of internal state (processing).

3 Information processing in neural systems

Information processing in a Turing machine is a serial operation. The state of the machine is updated in a sequential manner based on instructions and the contents of memory. Neural computation departs markedly from this approach. Enormous numbers of operations in the brain occur simultaneously, with many neurons receiving communications from many neighbors and independently accessing local synaptic memory. The essence of neural information is to share the burden of computation across a large network of processors, while connecting them and signaling in such a way to enable the information of the disparate elements to be efficiently integrated in system-wide operations. Where a bit in a digital system enables one to answer a single binary question, the large number of synapses input to a neuron enable one to answer a large number of simple analog questions. Whereas the von Neumann architecture requires external memory to be read and written at each computational step, processing and memory are not distinct in neural systems. Whereas digital computing with a Turing apparatus can integrated information from separate calculations in a serial manner, neural information is, by its nature, integrated across a spatial and temporal hierarchy.

Perhaps the key concept behind neural cognition is the ability for systems to achieve differentiated local processing combined with information integration across space and time. Each neuron is receptive to a certain subset of information, and it will pulse in response to presentation of that information, while remaining quiescent under other stimuli. The response of the neurons in a network is differentiated, so they each express different signals, yet a simultaneous interpretation of a broad array of information can be gained through the network as a whole.

3.1 Spatial structure of neural systems

3.1.1 Adjacency matrix

In order for the neurons to differentiate their responses while efficiently exchanging information across the network, certain structural network considerations are pertinent [**?**]. The structures of neural systems are analyzed in the framework of network theory [12]. In network theory, a system is discussed as a set of nodes connected by a set of edges. To facilitate mathematical analysis, the system is represented by an adjacency matrix, $\mathbf{A}$. Each column of $\mathbf{A}$ represents a node in the network, and if an edge connects node i to node j, $A_{ij} = 1$, otherwise $A_{ij} = 0$. Connections can be directed, in which case $\mathbf{A}$ will be, in general, nonsymmetric. Connections can also we weighted, in which case the adjacency matrix becomes a weight matrix, $\mathbf{W}$, whose elements represent not only the presence of edges, but also their strength.

To analyze a network, we label each node with an index $1 \leq i \leq N$, where N is the total number of nodes in the network. If we know all the connections between the nodes in the network, we can proceed to construct the adjacency matrix. In neural systems, we may use network theory to analyze the system at various scales. At the smallest (device) scale, each neuron may be represented a node, and each synapse by a directed edge. Alternatively, at the system scale, brain regions may be represented by nodes and the connections between these regions by edges. Across these scales, certain network metrics are important for analyzing performance.

3.1.2 Node degree and degree distribution

A simple yet important quantity that can be calculated from the adjacency matrix is the number of connections made by each node. If the network has directed connections, the number of outgoing edges (out-degree) will, in general, be different from the number of incoming edges (in-degree). The out-degree is given by $k_i^{\mathrm{out}} = \sum_j A_{ij}$, and the in-degree is given by $k_j^{\mathrm{out}} = \sum_i A_{ij}$. At the scale of devices, the in-degree represents the number of synaptic connections terminating on a given neuron, and the out-degree represents the number of synaptic connections made by a given neuron. The function describing the degrees of all nodes in network is referred to as the degree distribution.

3.1.3 Path length

One reason the node degree is important is that it enables us to calculate the average minimum path length across the network. To calculate this quantity, we find the minimum path length (number of edges that must be traversed) for each pair of nodes in the network, and we take the average over all pairs. This average minimum path length is a coarse, global metric that provides information regarding the ability of information to be efficiently exchanged across the network. In constructing a neural system, one would like to keep path lengths as short as possible to facilitate efficient communication from any neuron to any other neuron. In general, network path lengths are minimized in random networks [13], wherein edges between nodes are assigned at random. Therefore, by considering the average minimum path length of a random network, we can estimate the connectivity required between neurons in order to maintain efficient communication across a neural system of a given size. This average path length can be calculated in closed form [14]. This quantity is plotted in Fig. 1(a), where we show the number of edges required per node ($\bar{k}$) to achieve a given average path length ($\bar{L}$) as a function of the number of nodes in the network [14]. Consider the case of a network with one million nodes. We see from this plot that if we wish to maintain a path length of two, each node must make, on average, one thousand connections. For the case of a network with 100 million nodes, each node must make 10,000 connections. This is similar to the case of the hippocampus in the human brain, with nearly 100 million neurons, each with 10,000 or more synaptic connections [15]. Maintaining a short average path length across the network is critical to enable efficient information integration, and this appears to be a major factor driving the extensive connectivity of biological neural systems, and our design of neural hardware for cognition must account for the requirement of many connections to enable efficient communication across the network.

3.1.4 Clustering

This analysis of average path length is relevant for quantifying a network's potential for information integration, but as we stated above, a neural system also depends on the ability for differentiated processing. Let us consider differentiation across various scales of the network. Ideally, no two neurons would behave identically, because this would not provide new information. Each neuron will fire preferentially in response to stimulus according to a tuning curve [18]. In practice, some redundancy is advantageous if a network is to rapidly gain confidence regarding the nature of a stimulus. Neural systems employ populations of

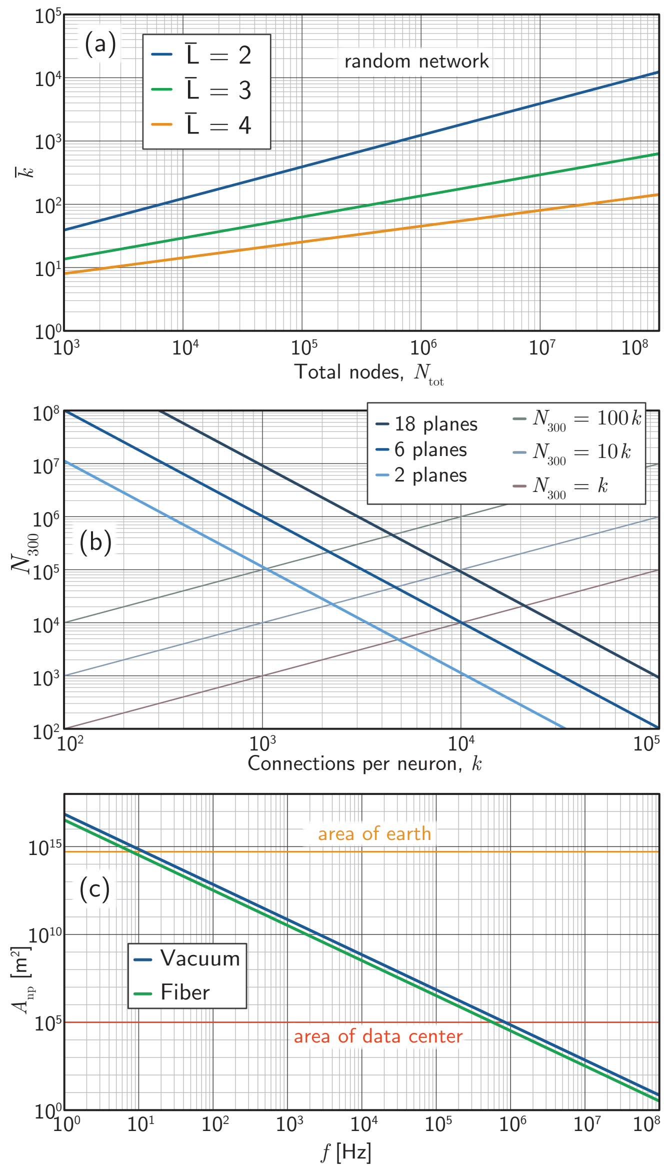

Figure 1: Scaling considerations for optoelectronic neural systems. (a) The average number of connections per node required to maintain a give average path length across a random network as a function of the total number of nodes in the system. (b) The total number of nodes that can fit on a 300 mm wafer as a function of the number of connections per node in the wire-limited regime [16]. (c) The area of the neuronal pool as a function of the frequency of neuronal oscillations assuming light-speed communication [17].

neurons to represent certain pieces of information. The net activity across the population represents the presence or absence of a stimulus, and the variation in activity across the population represents uncertainty about the interpretation of the stimulus. In order for a certain population of neurons to predominantly exchange information locally and rapidly establish a consensus interpretation of a stimulus, the neurons in that population should comprise a preponderance of connections within the population. Groups of neurons with an abundance of connections within the group relative to connections external to the group form a community, and a clustering coefficient [**?**, 19] is the simplest network metric. If node a is connected to node b and to node c, the clustering coefficient quantifies the fraction of cases in which node b is also connected to node c. This metric provides insight into the tendency of neurons to form specialized communities with information processing differentiated from other parts of the network, and is first step toward analyzing the modularity of the system.

3.1.5 Small-world networks

A random network has a low clustering coefficient, as the presence or absence of a connection between nodes b and c is, by definition, completely independent of the presence or absence of any connections to or from node a. Yet the networks of the brain—from the scale of neurons to the scale of connections between regions—show a high degree of clustering relative to a random network. As we have emphasized, neural information processing relies on both differentiated processing by local clusters and efficient integration of information across the network. We therefore expect neural systems to simultaneously achieve a high clustering coefficient and short average path length. A small-world networks is one with both of these traits [20]. We can introduce the small-world index (SWI) [21] given by $\mathrm{SWI} = \frac{\bar{C}/\bar{L}}{\bar{C}_{\mathrm{r}}/\bar{L}_{\mathrm{r}}}$, where $\bar{C}$ is the average clustering coefficient, $\bar{L}$ is the average shortest path, and the subscript, r, refers to a random graph. Whether analyzed at the scale of populations of a few thousand neurons or at the scale of large regions of the brain, the networks demonstrate large SWI, indicating that different populations of neurons represent different information, the activities of these populations can be efficiently communicated across the network, and giving hints that brain architecture is hierarchical and modular.

3.1.6 Modularity and hierarchy

Anatomically, clusters in the brain were thoroughly described by Mountcastle in 1978 [**?**] before the concept of a small-world network had been introduced [20]. Mountcastle referred to the communities of neurons as minicolumns and columns. Imaging of biological neural systems with various techniques across all relevant length scales indicates modularity persists across multiple levels of hierarchy [**?**]. The clustering coefficient discussed above can be generalized to the modularity, Q, quantifying the degree to which the connections of a network depart from what would be expected of a random network to form partitioned communities [?, ?]. For example, clusters of neurons are partitioned into mini-columns. At the next level of hierarchy, clusters of mini-columns are partitioned into columns. Clusters of columns are partitioned into brain regions. At the highest level of hierarchy in the brain, regions exchange information throughout the cerebral cortex and the thalamocortical complex that controls and coordinates operation of the largest modules comprising the brain. Neurons predominantly contribute to activity within their module, but they also must be able exchange information across partitions and up the information-processing hierarchy. Again we find differentiated, local processing combined with information integration across the hierarchical structure of the network. This neural information processing is enabled by a small-world architecture.

3.1.7 Rentian analysis

We can quantify the ability of information to be communicated up the hierarchy by analyzing the number of connections penetrating various partitions of the network, referred to as Rentian analysis. Rent's rule states that the number of edges crossing the boundary of a partition (k) is related to the number of nodes within the partition (n) by a power law of the form

$$\varepsilon = \varepsilon_1 n^p, \tag{1}$$

where ε_1 is the number of edges emanating from each node (first level of hierarchy), and p is referred to as the Rent exponent. Rent's rule (Eq. 1) was first observed in the context of VLSI circuits, and has also been shown to hold for biological neural circuits ranging from *C. elegans* to the human brain [22].

Depending on the system under study, the Rent exponent may vary, but is generally around 0.75 [22]. We will demonstrate the use of Eq. 1 in Sec. 8. The significance of the Rent exponent can be seen in its relation to the topological dimension, D. In Euclidean geometry, the surface area of a structure embedded in d-dimensional space is given by a power law of the form $s \sim v^p$. In general, the surface area s scales as a length to the power of $d-1$, while the volume scales as a length to the power of d. Thus, in Euclidean geometry, $p = 1 - 1/d$, or $d = (1-p)^{-1}$. The same expression holds in fractal geometry [?, ?], and in the case of Rentian scaling the topological dimension is related to the Rent exponent by $D = (1-p)^{-1}$, with $0 \leq p \leq 1$, so that $1 \leq D \leq \infty$. For $p > 2/3$, the topological dimension is larger than the embedding dimension, indicating that through a judicious implementation of input and output ports, information can flow through a network as if its components were connected in a higher-dimensional space. In the case of the brain, we find $D \approx 4$.

In any real network, it will only be possible to adhere to Eq. 1 across a certain domain of n. In the brain, consideration can apply from a single neuron up to the brain as a whole. Yet across this domain, Eq. 1 applies only piecewise, with discontinuities at certain partitions [23]. For example, we may expect Eq. 1 to hold from the scale of a single neuron up to the scale of a cortical column, but the connectivity patterns between columns are quite different than within, so we may expect a discontinuity at the scale of cortical columns, followed by a new expression of Eq. 1 with a different value of p, and perhaps a new interpretation of n as the number of columns within a partition. Similarly in the domain of microelectronics, a multicore processor on a chip may obey Eq. 1 within each processor, with a discontinuity at the scale of the connections between processors. The wiring organization of VLSI circuits and the brain is statistically fractal, but not infinitely so. The limitations are purely physical.

How does Rentian analysis inform our design of neural hardware? As we have emphasized, neural information is processed across a modular hierarchy. Rentian analysis quantifies the ability of information to be transmitted across partitions in the hierarchy. We conjecture that the ability for a neural system transmit more information across more levels of hierarchy will improve general intelligence. Therefore, between the neurons within a module, we expect that neural hardware should achieve high values of D over a large range of n before a discontinuity is necessary, so that multiple levels of hierarchy can be traversed efficiently within a module. Further, upon encountering a discontinuity, hardware must have a means of establishing another domain of Rentian scaling to efficiently collect and distribute information across more orders of hierarchy. Finally, we conjecture that the most intelligent neural hardware will provide this fractal scaling across as many levels of hierarchy as possible until the system finally reaches the limits set by the speed of communication. In Sec. 8 we will discuss these large-scale limits in biological and optical systems.

3.1.8 Summary of spatial considerations

The spatial structure of a neural system must comprise nodes with many edges to enable short path length across large networks. These nodes must also have high clustering to enable differentiated processing within modules. The information from these modules must be integrated across the hierarchy of the network, and the ability to do so is quantified by the topological dimension. Neural systems efficiently process information through the fractal use of space. Let us now consider their operation in time.

3.2 Temporal dynamics of neural systems

A neuron is a dynamical entity. It receives input from many afferent synapses, and it integrates this input over time. In a biological neuron, activity on a synapse results in a post-synaptic current into the receiving neuron. This current reduces the magnitude of the voltage across the neuron's cell membrane, and when that voltage is reduced below a threshold, the neuron produces an action potential, often referred to as a spike or pulse. This dynamical process of signal accumulation followed by bursting activity qualifies a neuron to be considered a relaxation oscillator. Before describing the temporal dynamics of neural systems in more detail, let us consider for a moment why relaxation oscillators are particularly well suited for cognition.

3.2.1 Relaxation oscillators

As we have mentioned, a defining aspect of cognitive systems is the ability to differentiate locally to create many sub-processors, but also to integrate the information from many small regions into a cohesive system, and to repeat this architecture across scales. A network of many dynamical nodes, each with the capability of operating at many frequencies, gives rise to a vast state space. As computational primitives that can enable such a dynamical system, oscillators are ideal candidates. In particular, relaxation oscillators [15,24–31] with temporal dynamics on multiple time scales [26] have many attractive properties for neural computing, which is likely why the brain is constructed of such devices [32]. We define a relaxation oscillator as an element, circuit, or system that produces rapid surges of a physical quantity or signal as the result of a cycle of accumulation and discharge. Relaxation oscillators are energy efficient in that they generally experience a long quiescent period followed by a short burst of activity. Timing between these short pulses can be precisely defined and detected [15]. Relaxation oscillators can operate at many frequencies [28] and engage with myriad dynamical interactions [27]. The oscillator's response is tunable [28], they are resilient to noise because their signals are effectively digital [33], and they can encode information in their mean oscillation frequency as well as in higher-order timing correlations [34–39].

3.2.2 Binary communication

For these reasons, we expect that any cognitive computing platform will be based on spiking neurons that behave as relaxation oscillators. Let us now consider some of the complex device functions that make neurons more capable of information processing than simpler relaxation oscillators. As mentioned above, communication between neurons is effectively binary—all or nothing. When a neuron produces an action potential, it propagates down the axon and branches throughout the axonal arbor. The signal propagates as a section of depolarization between the interior of the axon and the surrounding extracellular fluid. This depolarization opens pores in the membrane of the axon, allowing the flow of ions from the extracellular fluid into the axon, thus providing the electrical signal that will reach the synapses. Each time the action potential is generated, the behavior is nearly identical: the speed of propagation of the signal is set by the physical properties of the axon, the number of pores that open is very large and not noisy, and the signal that reaches the synapses is very similar from pulse to pulse. To further contribute to the digital nature of neuronal communication, the role of the action potential propagating down the axon is not to provide current to the post-synaptic neuron, but rather to begin a chemical cascade within the synapse that controls the signal amplitude. When the action potential reaches the synapse (pre-synaptic cleft), the action potential triggers the release of neurotransmitters into the synaptic cleft. These neurotransmitters diffuse through the fluid of the cleft, and bind to receptors on the post-synaptic cleft. These receptors then trigger the flow of current through the dendrite on which the synapse resides. This post-synaptic current carries the information that will be processed, first by the dendrite, then the dendritic arbor, and finally the soma. The action potential arriving at the synapse initiates the synaptic cascade in a binary, all-or-nothing manner, but the amount of current flowing into the post-synaptic dendrite depends on the state of the synapse, and it can take a continuum of values. Thus, communication in neural systems is digital (binary), yet information processing is analog. The synapse performs a digital-to-analog conversion, and the state of the synapse (which depends on many factors) determines the analog value entering into the computation performed by the dendrites and soma.

3.2.3 Short-term synaptic plasticity

The state of a synapse is affected by its activity over short and long time scales as well as external network factors. Neurons often signal in bursts (closely spaced sequences of spikes) [40], and within a burst, the time between spikes is referred to as the inter-spike interval. Changes of synaptic response over time scales on the order of the inter-spike interval are referred to as short-term synaptic plasticity [41]. One key effect of short-term plasticity is to perform a temporal filter on an afferent spike train. This can be a high-pass, low-pass, or band-pass filter, as shown in Fig. **??**. High-pass filtering results in only the first few pulses of a train being transmitted from the pre-synaptic axon to the post-synaptic dendrite. A synapse performing high-pass filtering reports to the post-synaptic neuron that the pre-synaptic neuron has begun to fire. Conversely, low-pass filtering results in synaptic response after the first several pulses of a train have occurred. A synapse performing low-pass filtering will not be active unless the pre-synaptic neuron produces a pulse train of a certain minimum duration, and therefore this synapse reports to the post-synaptic neuron when the pre-synaptic neuron has sustained bursting activity beyond a certain duration. Band-pass filtering combines these responses. A synapse performing band-pass filtering will only produce a response after an afferent pulse train exceeds a certain duration, and it will fall silent again after if the afferent pulse train continues beyond a certain duration.

These short-term filtering mechanisms enable synapses to report much more information to the neuron than simply the time-averaged rate of afferent activity. A neuron combining the signals from many synapses with various short-term responses has access to information regarding not just the average spike rates of the neurons from which it receives synaptic connections, but also regarding the initialization of bursting and the duration of spike trains.

3.2.4 Long-term synaptic plasticity

Over time periods much longer than the inter-spike interval, the response of synapses can also change based on the activity of the two neurons involved in the synapse. A synapse that is more active will strengthen (long-term potentiation), and a synapse that is used less will weaken (long-term depression). This was the essential insight of Hebb in 1949 [42], a concept that developed in the subsequent decades [43, 44] to account for the fact that long-term potentiation only occurs if the pre-synaptic neuron fires just before the post-synaptic neuron, indicating the potential for causality, while long-term depression is induced when the pre-synaptic neuron fires just after the post-synaptic neuron. This spike-timing-dependent plasticity (STDP) [45] plays a central role in memory formation and network adaptation.

The physical mechanism responsible for STDP involves the growth and decay of neurotransmitter sources and receptors present at the synapse. These synaptic molecular machines (N-methyl-D-aspartate receptors, NMDARs) develop in response to the action potential arriving at the pre-synaptic terminal as well as the back-propagating signal from the post-synaptic neuron. The complex chemistry present at each synapse leads to a remarkable degree of diversity and adaptability in synaptic response.

By strengthening cooperative synapses, STDP adapts the structural network of neurons and their connections into functional networks embodying certain memories or computations learned over time based on the correlations of neuronal firing events. It has been shown that random networks with synaptic weights adjusted over time by STDP adapt into small-world networks [46], maintaining efficient communication, and adding functional clusters specialized for specific computations. This is one example of the structural network of a neural system can be used to manifest multiple functional networks.

The functional clusters established via STDP have spectral signatures. A given functional cluster of neurons will have a specific pattern of activity, that may repeat in time. The period of this repetition will depend on the specific parameters of the circuit, and a large structural network comprising many highly connected neurons will have the potential to establish a vast repertoire of functional clusters with oscillations at many frequencies. Through STDP, the network can increase the activity of certain oscillations corresponding to highly utilized functional modules. If a certain stimulus has a probability of activating a certain functional cluster, the overtime, the action of STDP will enable the network to correlate that stimulus with the dynamical response of the cluster, and the stimulus will evoke the activity of the cluster with higher probability. In the language of dynamical systems, a specific cyclical response of the functional cluster is referred to as a basin of attraction [24, 40], and STDP ensures that relevant stimuli lead the network to the appropriate basin of attraction. This is function of an autoassociator, and it is an important form of long-term memory ([15] pg. 329).

Spike-timing-dependent plasticity makes use of correlations between firing activity of neurons to adapt the network into functional clusters [46], store memories in dynamical sequences [39], and strengthen circuits that demonstrate temporal patterns storing sequential memories ([15] pg. 318). But theoretical analysis finds that if STDP is the only long-term synaptic plasticity mechanism, memories are forgotten very quickly [47]. Experimental analysis finds that synapses have multiple additional means to change how synapses adapt over time and activity to help retain memories [48]. These plasticity mechanisms are referred to collectively as metaplasticity.

3.2.5 Metaplasticity

While STDP adapts the strength of synapses (synaptic efficacy), metaplasticity adapts the rate at which synaptic efficacy changes over time and activity. For example, if a pre-synaptic neuron fires just before a post-synaptic neuron, the synapse connecting the two will be a candidate to experience long-term potentiation. But the *probability* that the synapse actually does potentiate can be controlled by chemical signals within the synapse. Additionally, the *amount* by which the synaptic efficacy changes is also subject to chemical modulation.

The function of metaplasticity is to control which neural circuits adapt at a given time [48]. The receptors mentioned above (NMDARs) can be controlled based on a variety of factors related to network activity so that adaptability may be turned on and off in certain regions at certain times. This functionality is required of plastic synapses to keep them from too quickly losing the trace of a memory that is still needed. Experiments with humans indicate that forgetting occurs as a power-law function of time, [49, 50], yet Fusi and Abbott have shown that memories are lost more rapidly than this if plastic synapses are presented with continual stimulus [47]. They have proposed a model that achieves the observed power-law forgetting by introducing internal complexity to the synapses [51]. In this model, each synapse has various states of efficacy (weak and strong synaptic weight), but it also has additional internal states with the same efficacy. The difference between these states is the probability with which the efficacy will change due to future plasticity events.

Metaplasticity provides a network with the means to to enable some regions to adapt at a given time, under a given stimulus, while other regions are unchanged at that time, under that stimulus. Further, metaplasticity provides a means by which some synapses within a region may change very rapidly to adapt to a new stimulus, while other synapses in the same region may change slowly or not at all when presented with the same stimulus. We expect that an intelligent neural system have the capability to immediately learn in response to new information, but also to maintain a lasting representation of all that has been learned through the network's existence even in the presence of continually varying input. Metaplasticity is an important means by which rapid learning in conjunction with long memory retention can be achieved. As stated by Abraham, "...these metaplasticity processes represent a major form of adaptation that helps to keep synaptic efficacy within a dynamic range and larger neural networks in the appropriate state for learning."

3.2.6 Homeostatic plasticity

To conclude this discussion of synaptic plasticity mechanisms, we note that short-term plasticity adapts based only on the activity of the pre-synaptic neuron, while STDP adapts based on correlations in the activity between pre- and post-synaptic neurons. The mechanism of homeostatic plasticity [52] adapts synaptic weights based only on the activity of the post-synaptic neuron. Homeostatic plasticity (also referred to as the Bienenstock-Cooper-Munro (BCM) model) adjusts the synaptic weights of synapses incident upon a given neuron based on a sliding temporal average of the recent firing activity of that neuron. Such a mechanism provides a means by which neuron and network activity can be maintained within useful limits and dynamic range can be maximized.

3.2.7 Dendritic processing

We have discussed how STDP leads to the formation of functional clusters within a network based on the history of correlated neuronal activity. But what if the network wishes to isolate specific functional clusters on time scales as short as an inter-spike interval? Or what if we wish to endow a neuron with the ability to respond not only to activity in single synapses, but rather to integrated activity from specific clusters of synapses, or to specific sequences of activity within a cluster of synapses? Dendritic processing enables these functions.

Dendritic processing refers to the intermediate, nonlinear transfer function performed by dendrites between individual synapses and the neuron as a whole [53]. The dendritic arbor is a complex, branching structure on which most of a neurons input synapses make their connections. The dendritic branches that comprise the arbor have passive and active properties that allow them to perform various computations. The dendritic arbor can thus be modeled as a network of multiple independent threshold units that integrate signals and produce dendritic spikes upon reaching threshold [54].

For example, consider a dendrite with two synapses. If the post-synaptic current into the dendrite is sufficient to produce a dendritic spike, and if this dendritic spike has the same form whether one or both synapses fire, the dendrite performs the OR operation. If both synapses are required to fire to produce a dendritic spike, it performs the AND operation. The current generated in a dendritic spike propagates only a short distance along the dendritic tree, into the next dendritic compartment closer to the cell body, and the current decays with an exponential time constant. Thus, dendrites can perform basic logical operations with a temporal component, and activity closer to the base of the dendritic tree at the cell body integrates information from a larger number of inputs.

In addition to the simple two-synapse Boolean operations described above, it has been proposed that ... sequences

Dendritic processing may even play a central role in learning as well. It appears to be the case that modification of NMDA receptors involved in STDP is governed by the production of dendritic spikes, and it is not required that the neuron fire in order for synaptic potentiation or depression to occur. These dendritic processing functions, including integration and thresholding, basic logical operations, sequence detection, and role in learning, indicate several means by which the devices comprising neurons and neural systems perform complex operations to maximize the ability of each neuron to utilize the information it is presented.

3.2.8 From devices to populations

3.2.9 Neuronal avalanches

3.2.10 Entrained oscillations and synchrony

3.2.11 Cross-frequency coupling

3.2.12 Fractal use of space and time

- get into a tad of history here, analogous to Turing/von Neumann

General intelligence: The ability to place a wide variety of information into a coherent context so that the behavior of the relevant parties can be understood and predicted.

3.3 Summary of neural information

Von Neumann suspected the existence of a more subtle and powerful language of information employed by the human brain. Neuroscience has elucidated many of the principles of this language. Let us attempt here to summarize the salient elements that should guide neural hardware design. Given the complexity of the subject and the rapidly evolving state of neuroscience, we expect time to bring corrections to these concepts, yet the foundations of these concepts do seem well established.

The model from neuroscience informing the hardware presented here is as follows. Each neuron attempts to gain access to as much information as is physically possible about the activities of the other neurons in the network. Each neuron gains access to pieces of this information based on the temporal filter it performs. For example, a given synapse (or pair) can pas on information about the rate, rising edges, falling edges, temporal correlations, or sequences output from any neuron (or pair) in the network. In the temporal domain, we assume the signals can each be given a distinct exponential decay constant. Each synapse then has the information to answer a question, such as, how much has neuron i been firing in the last τ_{ij} seconds? Or, how much has neuron i been bursting, and then quiescing, and then bursting again in the last τ_{ij} seconds?

The answers to these questions must pass through the dendritic arbor. Each dendrite contains information received from one or a number of synapses coupled to the dendrite. The net information contained in the dendrite may be able to answer a question such as, how much have neurons a, b, and c collectively been firing in the last τ seconds? Or, how many of a particular subset of 10 neurons in cluster q have stopped firing in the last τ seconds?

When under the influence of inhibitory neurons, a dendritic compartment will be quiet. Upon the release of inhibition, the dendritic compartment reports to the neuron the answer to the question it knows how to answer by transmitting an analog signal in the form of current that modifies the neuron's membrane potential. Each segment of the dendritic arbor performs a nonlinear transfer function on the signals from the synapses connected to that segment, and the neuron itself performs a nonlinear transfer function on the signals it receives from across its dendritic arbor. The neuron's nonlinear transfer function is to produce a pulse (an action potential) when the membrane potential of the soma reaches a certain threshold value. This pulse is communicated through the neuron's axonal arbor to all the neuron's downstream connections as a digital signal, wherein the presence of the pulse informs all downstream connections that under the present network conditions, the activities on all that neuron's dendrites were sufficient to induce firing, and the amplitude of the pulse is not used to encode information.

In this picture, the excitatory (pyramidal) neurons are a knowledge base that can be queried by the inhibitory interneuron network. The net objective of the network is to be able to identify as many correlations as is physically possible across space and time. In space, these correlations are limited by network path lengths. In time, correlations are detectable over time constants of synapses and dendrites. To identify correlations over longer times than this (such as the lifetime of the entity), the logic of synaptic plasticity and metaplasticity come into play [47,51]. Note that this model of inhibitory query of pyramidal neurons is readily scalable across arbitrary partitions of the network, so the basic informational principles are continuous from the scale of local networks up to the system as a whole. This is uniquely enabled by the fractal use of space and time.

We conjecture that the probability of observing a neuronal avalanche accessing the information in s dendritic compartments scales as $P(s) \sim s^{-\alpha}$. During oscillatory behavior, inhibitory neurons sample specific dendrites in an intentional, controlled manner at a frequency f_θ, so that the information contained in the collection of all synapses, dendrites, and neurons active in the functional network resonant at f_θ can be integrated across partitions of the network to be incorporated in computations a higher levels of network hierarchy. The mechanisms for this coherent information access and integration include cross-frequency coupling, wherein local activity occurring at higher frequencies f_γ is modulated by slower frequencies f_θ, with the phase of the higher frequency activity being well-defined relative to the phase of slower frequencies. Such network-wide information integration through multi-scale activity across space and time is thought to be necessary for cognition [15], perhaps by enabling access to the global neuronal workspace [55, 56].

To summarize the summary, a single neuron extracts as much information as it can from its neighbors, and it transmits as much as it can to its neighbors through its activity in various effective network contexts established by the state of the dendritic arbor as configured by the inhibitory interneurons. A cluster in the network attempts to answer as many questions as it can about its inputs, and it attempts to communicate this information across the network to as effectively as possible, and so on up the hierarchy of network partitions. A network of inhibitory neurons samples the information from synapses and dendrites in myriad combinations, in principle answering any question that could be reasonably posed regarding a stimulus that could be physically presented to the entity.

This model of neural information processing bears a resemblance to a Turing machine. Turing's goal was to make a machine that could answer any question that could be asked within the axioms of its system (universal while not violating Gödel), and the goal here is essentially the same. Yet in addition to the Turing machine behavior, wherein the network acts as an oracle, an intelligent neural system should also be able to ask its own questions by formulating an output that generates a response from an intelligent or inanimate agent so as to gain new information. In addition to generating an entity that is universal in the sense defined by Turing, we aspire to create machines that are intelligent in the sense that they can engage in self-directed learning. Such a machine will be able to answer our questions, but also have a mind of its own.

4 Semiconductor electronic neural systems

The origin of semiconducting devices is intimately intertwined with the history of computing. After WWII, vacuum tubes were a relatively mature technology, established for switching in phone networks and wartime radio communications. Thus, early computers developed shortly after the end of the war were based around the deflection of currents by voltages applied to the central conductor of the tube. The invention of the transistor in 1948 by Bardeen, Brattain, and Shockley replaced the bulky tubes, and the subsequent development of integrated microchips by Kilby and Noyce in 1959 initiated the technological revolution that has left the world utterly transformed. Innovations in lithography and processing led to the evolution captured by Moore's Law and Dennard scaling. After nearly 60 years, these scaling trends have neared the physical limits of transistors [], inspiring new creativity in devices and architectures. On the device side, use of photonic components for communication is gaining significant traction [57,58]. Architectural innovation has led to increased parallelism of computation, with brain-inspired concepts at the extreme end of this spectrum.

Need history of silicon electronics: why silicon? why Si instead of Ge? why Si instead of III-V? Oxide, cost, ease of manufacturing.

4.1 Efforts in silicon microelectronic neural systems

In arguing for utilization of light for communication in artificial neural systems, one may be perceived as adversarial toward purely electronic approaches. It is important to point out up front that artificial neural systems based on semiconductor electronics are the state of the art, and they will be for years to come. The immaturity of integrated photonic technology makes it a worthwhile enterprise to continue to push the limits of silicon microelectronic neural systems, and throughout the discussion below we hope the reader appreciates our respect for what these systems have been able to accomplish. Nevertheless, it is the goal of this work to consider asymptotic technological limits of artificial neural systems, so our task is to identify the bottlenecks present in conventional hardware and propose solutions to overcome those bottlenecks.

There are a number of large-scale efforts in silicon-microelectronic neural systems [], as well as a number of review articles summarizing those efforts [], so we relieve ourselves of the task of recapitulating that large body of important work, and instead focus on the common elements of all silicon microelectronic neural systems to date that limit the ability of those systems to achieve human-brain-scale cognition. Many of the efforts in neuromorphic computing do not intend to achieve cognitive systems, but rather intend to perform smaller-scale computational tasks with improved efficiency relative to other architectures [59,60] or to perform simulations of biological neural networks to advance our understanding of neuroscience [61,62]. This latter objective is entirely consistent with the original objective of using a universal Turing machine to process arbitrary differential equations to model an aspect of nature. Von Neumann did not intend for the EDVAC to actually produce shock waves, but rather to step through the differential equations modeling shock waves to enable the user of the system to predict the behavior of the physical system. By contrast, the objective of an artificial cognitive system is to physically achieve cognition in hardware, not just to model the behavior of a different physical system during cognition. It is not necessarily the case that a system modeling cognition would achieve cognition, particularly if the model represents only a subset of the true cognitive system.

4.2 The von Neumann bottleneck

While silicon microelectronics based on the field-effect transistor has made advances far beyond what was considered possible during the conception of the first electronic computers in the 1950s, modern CMOS neural systems still bear remarkable resemblances to early computing machines. In particular, the separation of processing and memory is present in many neural systems. Von Neumann understood that communication between processing and memory was likely to be a limitation, and this pinch point is still referred to as the "von Neumann bottleneck". This bottleneck is particularly problematic for implementing artificial neural systems, because processing and memory are not separate in neural systems. The synapses and dendrites that perform the first stages of computation are also the elements that store memory in their synaptic weights. Synaptic weights affect the dynamical operation of the neurons, and the dynamical operation of the neurons affects the synaptic weights. Therefore, when emulating the behavior of a neural system with a Turing machine employing the von Neumann architecture, significant communication between processors and memory is required. Some efforts side step this challenge by eliminating synaptic plasticity all together, leading to neuromorphic systems that perform inference, but do not learn [59]. Others include synaptic plasticity mechanisms to enable learning during operation, and bear the costs of reduced speed of network activity [60,62].

While the architecture of most silicon microelectronic neural systems show their von Neumann ancestry, such systems do not simply have one processor with one memory bank and one von Neumann bottleneck between. Instead, microelectronic neural systems employ massively multi-core architectures, wherein many processors with local memory are interconnected in a network. Such an approach improves upon the limitations of a single-processor/single-memory architecture, and spreads the communication burdens across many nodes. In this configuration, each processor simulates the activity of a number of neurons (usually a few hundred) by stepping through the differential equations that model the neurons' dynamics. With such an approach, each processor is a Turing machine employing the von Neumann architecture, and the information generated within each must be communicated to the rest of the network. While such an architecture mitigates the limits of a single von Neumann bottleneck, limitations still arise. As stated in Ref. [62], "...often the compute budget is dominated by input connections...which imposes an upper limit on the (number of neurons)×(number of inputs per neuron)×(mean input firing rate)." Furber et al. additionally state that plastic synapse models further burden the number of inputs a processor can manage. While the numbers are sufficiently high to be exciting for computational applications and neural simulations, these are some of the bottlenecks we hope to overcome with photonic communication.

4.3 Fan-out limitations

We have discussed some of the challenges of silicon microelectronic neural systems in terms of processor-memory communication bottlenecks, but it is illuminating to consider the physical origin of the problem. In all silicon microelectronic circuits, the transistor is the element that represents information. The presence or absence of a voltage applied to the gate of the transistor changes the state of the transistor, and in binary computing schemes, only two values of voltage are relevant. A transistor or circuit comprising transistors and wires has some capacitance, C, and the voltage applied to the circuit is given by $V = Q/C$. To switch the state of a silicon MOSFET, V must reach nearly 1 V. Capacitances can only be reduce so low, and in the context of neural circuits wherein significant connectivity is required, capacitance due to wiring dominates. As a rule of thumb, a wire in a CMOS process adds 200 aF/µm, so parasitic wire capacitance dominates when devices are separated by even a few microns [63]. Thus, if each neuron were to directly charge up the wires and transistors of a thousand target neurons, the amount of charge, Q, would be intractably large, requiring each neuron to source a prohibitive amount of current. In general, this physical limitation limits CMOS circuits to fan out of about four.

4.4 Shared communication infrastructure

This limited fan out is not specific to neural systems, and it has long been dealt with in various integrated circuits, initially through shared-media networks (the communication bus), and in contemporary circuits with switched-media interconnection networks [64]. In such a network, each node is connected locally to a switch fabric, and all nodes of the network share this communication infrastructure. Such switching networks enable nearly all integrated electronic systems, from networks on chip up to the internet, though the hardware implementing the switching varies with spatial scale.

The shared communication infrastructure of switched-media networks is an excellent solution to overcome the fanout limitations of silicon microelectronic devices. Each device must then only communicate to the nearest switch in the network. In a switched-media network, devices communicate with one another by sending packets of information. The packet contains routing information (the address of the recipient) as well as the data to be communicated. The interconnect network determines a valid route for the information traverse across the network (referred to as routing), and the switches are configured accordingly to achieve that physical route of information transfer.

Because the communication infrastructure is shared, devices must request access to the switch network to transmit messages. Multiple devices may request access simultaneously, in which case arbitration must be performed. Arbitration refers to the process of granting devices access to the switch network, and in general a packet will experience some delay while it waits in a queue to be granted access to the shared communication infrastructure. The process of serializing communication across a common interconnection network is referred to as time multiplexing. This approach to communication between electronic devices leverages the speed of electronic circuits to compensate for the difficulties in communication.

4.5 Address-event representation

For many applications, the latency incurred by the shared communication infrastructure is tolerable. The limitations are reached when many devices need to communicate with many other devices with a high frequency of communication events. Unfortunately, this is exactly the situation encountered in neural information processing. When implementing neural information processing with electronic communication infrastructure, neuron pulses are represented as packets of data called events. Some of the data in a packet representing an event must contain the addresses of the synapses to which the event should be communicated. This type of neural information processing is therefore referred to as address-event representation [65]. It is natural to adapt the von Neumann architecture to neural applications by assigning addresses to all elements of the network. This is a straightforward application of the way memory has been accessed since the early days of computing. As Julian Bigelow wrote in 1955, "...by means of explicit systems of tags characterizing the basically irrelevant geometric properties of the apparatus, known as 'addresses'. Accomplishment of the desired time-sequential process on a given computing apparatus turns out to be largely a matter of specifying sequences of addresses of items which are to interact." [66] We argue that for the most efficient communication and computation in neural systems, the geometrical properties of the apparatus are not irrelevant, and the burden of storing and communicating addresses in large neural systems would be advantageous to avoid.

One consequence of address-event representation is that as the size of the system grows, more information in each communication event must be allocated to specify addresses. This leads to increased burden on memory and processors. But the more severe challenge is introduced by the connectivity/speed trade-off. As more neurons, each with many synapses, are added to the network, the average frequency of neuronal firing events must decrease due to the limitations of the interconnection network to handle communication requests. For electronic systems with a few hundred thousand neurons, average event rates in the kilohertz range can be maintained []. Systems with a few hundred million neurons will likely be limited to operation at 10 Hz or below [].

4.6 Contention delay in neural systems

As stated in Ref. [64], "When the network is heavily loaded, several packets may request the same network resources concurrently, thus causing contention that degrades performance. Packets that lose arbitration have to be buffered, which increases packet latency by some *contention delay* amount of waiting time." In neural systems, many neurons must be able to communicate to many other neurons, and contention delay becomes severe. Contention delay is particularly limiting when large neuronal avalanches occur, or when many neurons from across the network form a transient synchronized ensemble (see Sec. 3). These are exactly patterns of network activity that are crucial for large-scale information integration and cognition. To employ hardware that suffers from contention delay places a limit on the network size, connectivity, and speed.

To put some numbers on in, in Ref. [67] demonstrated a network with 262 thousand neurons, each making one thousand virtual connections through the shared communication infrastructure. In that work, the neurons were able to maintain roughly 1 kHz average event rate per neuron. Reference [68] theoretically explored a network with a thousand times more neurons (250 million), again with one thousand connections per neuron, and found that communication would limit each neurons to less than 10 Hz average event rate. While the brain has 100 billion neurons, and the cerebral cortex has 10 billion, they each only fire, on average, at 1 Hz. Yet when necessary, they can burst at up to 100 Hz (for pyramidal neurons). For information processing in neural systems, it is important that many neurons be able to burst simultaneously and communicate across spatial scales. Even though neural network average activity is low, the ability for many neurons to simultaneously fire rapidly and communicate broadly is crucial to neural information processing. It is worthwhile to pursue hardware enabling this operation.

4.7 Summary of challenges of silicon microelectronic neural systems

The fan-out limitations due to charge-based parasitics necessitates the use of a shared communication infrastructure. For spiking neurons to send packets across this switching network, each neuron must have an address, and each processor and/or routing node must store in memory the addresses of all nodes in the network. As the size of the network grows, storing these addresses and communicating them in each transmitted packet places more stringent demands on processing, memory, and the von Neumann bottleneck between them at each node. With the address-event representation, spike events must be routed by the switch nodes of the interconnection network. Arbitration must be performed to handle collisions, and when network activity is high, contention delay occurs.

At the core of the challenges faced by microelectronic neural systems is the shared communication infrastructure. It is this shared infrastructure that forces the storage of addresses and contention delays at switches. The requirement of shared communication infrastructure physically results from the capacitance of wires and transistors that makes it impossible to directly connect each device to thousands of other devices. We consider the primary objective of using light for communication in neural systems to be alleviation of this physical limitation. Because light experiences no capacitance, inductance, or resistance, a pulse of light can fan out to as many destinations as there are photons in the pulse without requiring shared communication infrastructure. It is our perspective that if artificial neurons could communicate with light, and therefore establish direct, physical connections from each neuron to all of its synaptic connections without storing addresses or incurring contention delays, the benefits to cognitive performance would be immense. This would allow each neuron to fire at a maximum rate limited by its internal devices, completely independent of the size of the network or the number of incoming or outgoing connections. The new challenge that immediately becomes apparent is the size of the network of waveguides connecting the neurons—the white matter. To demonstrate feasibility of photonic communication between neurons, one must consider the spatial scaling of the network, a point we take up in Sec. 8.

There are important advantages to the shared switching network and address-event representation. Foremost is the adaptability. The same hardware can be reconfigured to emulate a variety of networks. This is useful if one wishes to design a single chip to perform a number of neural computation or to be used in the study of a number of neuroscientific investigations. Nevertheless, such adaptability carries a hardware premium. Any high-performance cognitive system must be adaptable in order to learn from experience. But we conjecture that it is advantageous for much of the communication infrastructure to be fixed to make more efficient use of limited space for hardware resources.

4.8 Actual versus emulated neurons

To close this section, we emphasize what we see as a more general lesson regarding hardware for artificial neural systems. In most contemporary silicon microelectronic neural systems, there are no actual neurons, and there are no actual axons connecting them. The approach follows closely the thinking laid out by Turing and established in hardware beginning with systems like the EDVAC: processors are not neurons, but they step through differential equations in time to arrive at outputs similar to what an actual neuron would produce. Each processor core follows the instructions in discrete time that cause it to behave as if it obeyed certain neural differential equations, but the underlying devices do not actually obey those equations. This approach of emulation is possible because a Turing machine is universal, but this does not mean it is efficient.

It is precisely in the neuromorphic context that the von Neumann architecture implementing a Turing machine is least suited to high performance. Neural systems utilize highly distributed memory, processing and memory access are not separate operations, processing occurs in parallel among many interconnected neurons and subnetworks, and communication on local and global scales is paramount. We should continue to explore silicon microelectronic neural systems in many different forms, but we should not be surprised if different hardware enabling different architectures is ultimately more efficient for large scale neural systems.

We conjecture that in efficient neural systems, the components will not perform a Turing-type emulation of neural behavior, but rather will physically manifest the differential equations of interest. This is an old idea [11] dating back to Mead in 1990, pre-dating even the early work on address-event representation by Boahen [65]. Carver Mead's original interest was in using the analog behavior of sub-threshold transistor to behave as neurons. He wished to utilize the isomorphism between the conductance of the transistor and membrane conductances in neurons. Mead attributes advantages "...to the use of elementary physical phenomena as computational primitives...". In the nearly 30 years since Mead coined the term "neuromorphic", many efforts have been made to utilize the analog properties of transistors to emulate neural operations [69]. Mixed analog and digital approaches have also been pursued []. While analog transistors may still rise to great performance, the basic limitation has been that the exponential dependence of transistor current on gate voltage leads to high device variability. Regardless, for the reasons listed above, device fan out in analog operation is still greatly limited, necessitating address-event representation. Most of the field has moved toward full embrace shared communication and digital emulation of neural dynamics. It is the most successful means of utilizing silicon microelectronics for brain-inspired computation.

Perhaps transistors are uniquely equipped for digital information processing. But are there other devices that may be more naturally suited to function as computational primitives in neural systems? The spikes produced by Josephson junctions in the form of single-flux quanta are a natural place to start looking.

5 Superconductor electronic neural systems

The first proposal to use superconducting components for computing occurred only a few years after the invention of the transistor. Beginning in 1950, Dudley Buck proposed to use the cryotron as a switch for digital logic [70, 71]. The principle of the cryotron is that a length of superconducting wire can be switched from zero impedance to very high impedance and back again by breaking and restoring superconductivity. Such functionality is matched to the needs of digital computing, and significant effort went into the development of computing systems based on cryotrons rather than vacuum tubes. This work extended into the 1960s, whence it became clear that many aspects of integrated silicon circuits would be superior for a number of reasons.

In 1962, Josephson explored tunneling between two superconducting wires separated by a thin barrier [72]. This led to the Josephson junction (JJ), the device that now dominates all information processing performed with superconducting electronics. Due to the significance of this device, it is worth a brief description of its functionality.

5.1 Josephson junctions

A JJ is created when two superconducting wires are separated by a thin tunneling barrier, as shown in Fig. 2(a). Excellent resources exist that describe the beautiful physics of JJs pedagogically [73–75], and it is our intention here to describe only very basic aspects of JJ operation relevant to the computing technology under discussion.

When realized in hardware, any Josephson junction has some shunting capacitance and resistance, leading to the effective circuit shown in Fig. 2(b), which is referred to as the resistively and capacitively shunted junction (RCSJ) model. While a JJ can, in general, be either voltage biased or current biased, for simplicity we restrict our attention to current-biased junctions as they are more relevant to the operations at hand. One can see that in this model of the junction, there are three conduction paths in parallel. Perhaps the most important aspect of a JJ when used as a classical information-processing element is the fact that the junction has a critical current. This means the central, superconducting path can carry a current $I \leq I_\mathrm{c}$ with exactly zero resistance and therefore zero voltage acorss the junction. However, if the current bias exceeds I_c, the fraction $\Delta I = I - I_\mathrm{c}$ cannot be carried through the supercon-

(a)

thin film diagram of JJ

(b)

RCSJ circuit diagram

(c)

JJ in operation, SFQ pulses being generated

Figure 2: Caption for JJ.

ducting channel, and instead must be carried through the parallel conduction pathways, with DC components being shunted through the resistor. Under a bias exceeding I_c, a voltage develops across the junction, and in general, this voltage will vary with time, even if the bias is constant, as we will describe in more detail shortly.

One of the reasons JJs are so fascinating is that depending on the choice of circuit parameters, JJs can demonstrate many behaviors. Because a JJ has an intrinsic inductance, the circuit effective JJ circuit of Fig. 2(b) can operate as an $L-R-C$ oscillator circuit, leading to ringing behavior. Alternatively, with other parameter choices, the junction can demonstrate latching behavior, wherein a junction biased above its critical current will enter the resistive state and stay there until the current bias is dropped well below the critical current, thus exhibiting hysteresis. For the information processing applications under consideration, it is more advantageous to operate near critical damping.

Let us now consider the operation of a simple circuit employing a single JJ...

Need to touch on speed, flux quantization (need to introduce SFQ terminology, fluxon), JTL, flux storage, energy, Φ_0, integral Vdt

5.2 Superconducting digital logic

Here we briefly review the history of using superconductors in digital computing. More detail can be found in Ref. [76]. The origin of superconducting computing is nearly concurrent with the origin of the rest of digital computing. In that post WWII context, the first components developed were switches for digital systems. The goal was to implement a von Neumann architecture, the same pursuit as vacuum tubes and transistors. In was in this setting that Buck developed a switch wherein a superconducting wire with high critical field is wrapped around a superconducting wire with lower critical field. A current passing through the coil wire could be used to break superconductivity in the core wire. Such an element could be used for switching and current amplification. Dudley Buck referred to this device as a cryotron.

The cryotron was a strong candidate as a switch for binary computing when compared to vacuum tubes, but with the development of integrated circuits this bulk component was not competitive for scaling. Yet in the late 1960s and early 1970s, IBM developed an integrated circuit element based on JJs that behaved like Buck's cryotron [77]. At that time, the materials for implementing superconducting integrated circuits were immature, and the device concepts for information processing were also under development. IBM chose to use Pb alloys as the material platform, and they chose to implement a latching logic [78]. This material platform was problematic, and nearly all contemporary efforts in superconducting logic utilize Nb as the predominant material for wiring and JJs (superconducting qubits are primarily based on Al). The latching logic IBM employed used the voltage across a JJ to represent information. As described above, a current-biased JJ can develop a voltage if the current bias exceeds $I \leq I_c$, and with a certain choice of the RCSJ circuit parameters, that junction can remain latched in the voltage state even after the current bias drops below $I \leq I_c$. Both the choice of Pb as a material platform and the choice to employ voltage-state logic ended up being problematic for IBM, and by the early 1980s IBM ceased its effort in superconducting digital logic.

The effort at IBM began before silicon microelectronic technology had far surpassed other approaches to digital logic, Moore's law had only recently been formulated [79], and it certainly was not obvious which hardware would dominate for electronic computing. One of the motivations to use superconducting circuits was that they were simpler to fabricate than semiconducting circuits. Additionally, due to the intrinsic dynamics of JJs, they could be made to operate at extremely high speed with low energy per operation. Materials improvements led to the development of JJs based on Nb with AlO_x as a tunneling barrier, and several efforts continued to explore JJ-based circuits for computing. In particular, the work by Likharev and others in the late 1980s and early 1990s developed a new type of logic for digital computing (see Refs. [78] and [80] for technical details and Ref. [76] for a retrospective). Within this framework, the state of flux within a superconducting loop represents information: an empty loop represents a binary zero; a loop with one fluxon represents a binary one. This form of logic is referred to as flux-state logic, and it overcomes many of the weaknesses of voltage-state logic. Likharev and colleagues referred to this approach as rapid single-flux-quantum logic, and developed a family of gates to implement digital computing. Within this framework, a clock is distributed across all gates in the system, and logical operations proceed based on whether or not flux is present in certain loops during each clock cycle.

The foundational work on flux-state logic [76, 78, 80] occurred during the prime years of silicon microelectronic scaling. For this reason, it was difficult to foster a large effort in a competing technology, leading Likharev to later lament the oppressive impact incurred on other technologies as "CMOS continued its victorious march." []. Yet interest in superconducting electronics for computing has continued, particularly in Japan, and as the scaling of silicon transistor technology has reached new barriers, attention has returned. In the US, the resurgence led to the IARPA Cryogenic Computing Complexity program, started in 2014, to develop high-performance superconducting digital computers. This effort, however, has not quickly led to circuits outperforming CMOS for digital computing. There are four reasons for this. We list them in order of severity. The first reason is that it is difficult to implement large arrays of compact memory cells with superconducting electronics []. This is very problematic when implementing the von Neumann architecture, as these digital efforts have aspired to do. The secondary reason is that distribution of a high-speed clock across many logic gates encounters obstacles, and the timing jitter of the gates leads to errors if the clock is too fast, largely eliminating speed advantages relative to CMOS. The third reason is that a superconducting system resides inside a cryostat. A high-performance computer will need extensive I/O, and this cannot be achieved straightforwardly with co-axial cables due to the high heat load such cables transfer from room temperature the the 4.2 K stage where computation is performed. Optical approaches are being developed to overcome this challenge [], but optical sources or modulators often require on the order of a volt to encode a bit, while the SFQ pulses produced by JJs are on the order of a millivolt. This introduces the fourth problem limiting the success of superconducting electronics for digital computing: it is difficult for a superconducting circuit to change the state of a CMOS circuit. That is to say, it is difficult for a superconducting circuit to achieve sufficient voltage to switch the gate of a silicon transistor. Even if one intends to develop superconducting electronics to outperform silicon electronics specifically in the domain of digital logic, it is helpful (and perhaps vital) that the superconducting system be able to interface bi-directionally with semiconductors so the superconducting system can leverage the tremendous infrastructure of semiconductor integrated circuits.

5.3 Superconducting optoelectronic neural systems may overcome challenges of digital systems

Let us refer to these challenges with superconducting electronics as: 1) the memory problem; 2) the clock problem; 3) the I/O problem; and 4) the volt problem. We will argue below that a new type of cryotron, implemented in a compact, on-chip, thin-film device can overcome the volt problem, but only if one accepts slower speed than superconducting digital logic seeks, an acceptable trade-off in a neural system. Solving the volt problem this way enables us to generate light, thereby solving the "O" part of the I/O problem. The "I" part of the problem is solved by utilizing superconducting photon detectors, compact circuit elements developed since 2000 and integrated with photonic circuits within the last decade. With integrated light sources and detectors, superconducting systems can send and receive near-infrared photons over optical fibers, relieving the heat load of co-axial cables, and solving the volt problem at semiconductor photodetectors on the room-temperature side. The asynchronous nature of spiking neural systems eliminates the clock problem, provided synaptic, dendritic, and neuronal integration times can be made longer than the jitter of the circuits. Finally, while the memory problem is the most severe for superconducting digital systems, the prospects for memory in superconducting neural systems are one of the most exciting aspects of the technology. As discussed in Sec. 3, memory in neural systems involves short-term plasticity, long-term plasticity (STDP, for example), and metaplasticity (that adapts the rate with which synapses change). We find that JJ circuits very naturally implement these functions, and the distributed nature of synaptic memory avoids the difficulty of creating large arrays of addressable superconducting memory elements.

We will describe these superconducting optoelectronic circuits in more detail in Sec. **??**. Let us first turn our attention to JJ circuits that implement neural functions without optical communication.

5.4 Neurons based on Josephson junctions

- JJ neurons (Japan, Segal, Schneider, that recent theory paper)
- still, communication problems, fan-out

JJ neurons: [81] [82] [83] [84]
MJJs [85] [86] [87]

6 Integrated photonic neural systems

As was the case with semiconductors and superconductors, so it is with integrated photonics: the first application has been digital logic. In the case of photonics, the goal of the hardware augmentation is to aid in communication. Optical communication shows indisputable advantages over long distances, as exemplified in global fiber optic networks as well as local-area networks. On the chip scale, the advantages of optical communication must contend with the challenges of optoelectronic hardware integration.

6.1 Silicon photonics and superchips

In 1987, Soref and Bennett introduced the concept of using the shift in index of refraction that results from free carriers in silicon to achieve active optical components based on silicon waveguides [88]. This insight would have to wait until the development of silicon-on-insulator wafers in the early 2000s to be put into practice. Since then, an explosion of activity has occurred in the rapidly developing field of silicon photonics. Electro-optic effects have been used to make a variety of modulators [89] operating into the 10s of GHz based most commonly on Mach-Zehnder interferometers [90] or microring resonators [91]. In addition to the free-carrier electro-optic effects, in 1993 Soref also pointed to thermo-optic effects as a means to make dynamic photonic components on an optoelectronic chip [92]. The combination of electro-optic effects for fast index perturbation and thermo-optic effects for slow resonance tuning, in conjunction with etched silicon waveguide structures in silicon-on-insulator substrates, established a foundation of active components capable of signal switching, filtering, and modulation. In his 1993 paper, titled *Silicon-based optoelectronics*, Soref presented a more expansive view of the potential for what he termed "superchips" that combine the strengths of photonics and electronics monolithically on a single silicon chip. Silicon had long been the material of choice for integrated microelectronics, but Soref had identified a path to make silicon also a powerhouse in photonics as well.

To make use of silicon as a waveguiding medium so that the active components described above can be implemented, one must utilize light with photon energy less than the band gap of silicon ($E_\mathrm{g} = 1.17\,\mathrm{eV}/\lambda = 1.06\,\mu\mathrm{m}$ at $0\,\mathrm{K}$; $E_\mathrm{g} = 1.11\,\mathrm{eV}/\lambda = 1.12\,\mu\mathrm{m}$ at $300\,\mathrm{K}$). The buried oxide of silicon-on-insulator wafers becomes absorptive for $\lambda \gtrsim 2\,\mu\mathrm{m}$. Thus, the transparency window of silicon-on-insulator waveguides enables operation with wavelengths below $1.2\,\mu\mathrm{m}$, and includes the important telecom bands (O-band:1260 nm-1360 nm; C-band: 1530 nm-1565 nm), whose significance results from the very low attenuation of optical fibers at these wavelengths. Thus, silicon integrated photonic components can be interfaced with optical fibers for communication across long distances.

Yet if a material is transparent, it is not efficient for detecting light. To create photodetectors in silicon waveguides, two approaches are taken. One approach is to utilize SiGe regions patterned in Si waveguides, as the band gap of Si is narrowed by the incorporation of Ge. Germanium is present in many contemporary CMOS processes for strain engineering, and can be economically incorporated in the foundry because, like silicon, it is a group IV element, and therefore shares process compatibility and does not act as a dopant in Si. Waveguide-integrated [] and resonator-integrated [] SiGe detectors operating at the O-band and C-band. These detectors have been demonstrated with high efficiency approaching 1 A/W. The other approach is to introduce defects the silicon lattice, either through ion implantation or the use of poly-crystalline or amorphous silicon. These defects introduce absorptive states within the band gap. Detectors based on this principle have been demonstrated with x A/W responsivity [93].

6.2 Silicon light sources: the great Achilles' heel

So if photonic switches, modulators, filters, and detectors can all be implemented in silicon, why do all silicon microelectronic chips not have photonic components? There is one reason: a simple, inexpensive light source integrated with silicon waveguides operating at room temperature does not yet exist. Silicon has an indirect band gap, so optical emission requires a phonon for momentum conservation. This three-body process (electron, hole, phonon) is rare, so non-radiative recombination dominates. Regardless, if silicon is to be used as a passive and active waveguiding material for routing, switching, and modulation, a source emitting at a longer wavelength must achieved, just as detectors must absorb at longer wavelength, as described above. If detectors can be made to accomplish this, why is the same not true for sources? Despite efforts for decades [94], an economical, efficient, room-temperature, waveguide-integrated light source on silicon has not been discovered. To understand the source challenges, let us briefly consider three means by which researchers have attempted to create silicon light sources. More comprehensive surveys can be found in the literature [94–97].

Like the case of detectors, two approaches to creating light sources on silicon are band gap engineering with Ge alloys and introduction of states in the gap via lattice defects. While detectors based on SiGe have shown decent performance without extensive process development, the same cannot be said of SiGe sources. Poor material quality is not as problematic if the goal is to make an absorber, whereas non-radiative recombination pathways introduced by material defects greatly limit the efficiency of SiGe as a light source and lead to high threshold current for lasing [97]. Thus, despite the process compatibility of SiGe with CMOS, SiGe lasers to date have not high enough performance with low enough cost to find a market.

Similarly, light sources based on defects in silicon have been studied extensively for decades as the silicon microelectronics industry has matured [98]. While defect-based detectors have demonstrated useful performance and low cost at room temperature, defect-based light sources have not. To understand why, consider a three-level model of the processes of absorption and emission, as shown in

Fig. ??. The three levels involved are the ground state (E_0, electron in valence band, hole in conduction band), the first excited state (E_1, electron and hole bound to defect), and second excited state (E_2, electron in conduction band, hole in valence band). At room temperature, the two phonon mediated processes ($E_2 \to E_1$ and $E_1 \to E_2$) are both fast, with few-picosecond time constants (check Davies). The electric-dipole transition ($E_1 \to E_0$) is comparatively slower, with nanosecond to millisecond transitions depending on the specific defect []. In detection, the dipole transition ($E_0 \to E_1$) is pumped by the signal to be detected, and the excited electron-hole pair quickly transitions from E_1 to E_2, where a reverse-bias field sweeps the carriers out of the junction, resulting in detection. By contrast, in the emission process one pumps the $E_0 \to E_2$ transition (through electrical carrier injection in a $p-n$ junction), and the excited carriers quickly transition to E_1, but before they can make the slow transition from E_1 to E_0, they make the fast transition back from E_1 to E_2, and eventually recombine non-radiatively through a variety of pathways without making the slow, dipole transition required to generate light. Crucially for our story, this is not the case at low temperature. The $E_2 \to E_1$ transition involves emission of a phonon, so it remains fast, while $E_1 \to E_2$ involves absorption of a phonon. At liquid helium temperature (4.2 K), the relevant phonon states have low occupation, and the rate of the optical transition from E_1 to E_0 can be faster than the rate of transition back to the band edge, making silicon light sources possible based on this mechanism when operating at the same temperature required to enable superconducting circuits based on Josephson junctions.

In addition to these two approaches to light sources on silicon, a major effort has been undertaken in the last 15 years to achieve hybrid integration of III-V light sources on silicon. Process incompatibility and lattice mismatch make it difficult to grow III-V gain media directly on silicon. Independent processing of Si and III-V substrates followed by wafer bonding is being pursued, but contemporary CMOS is very comfortable at 300-mm-wafer scale, while III-V processing has stayed at 150 mm or below. Many such subtleties and complexities of process and materials integration have limited hybrid system performance and kept costs high. Many of the challenges are practical rather than fundamental, but nevertheless place real limits on the technologies that are achieved.

6.3 Systems with off-chip sources

Considerable work continues in the development of light sources on silicon. At present, many efforts are proceeding to demonstrate exciting systems on chip with optical communication based on external III-V lasers fiber-coupled to silicon optoelectronic processors. Such work began commercially with the founding of Luxtera in 2001 with the goal being to utilize integrated silicon photonics for network interconnects. More recently, the effort led by Stojanovic, Popovic, and Ram has led to the development of silicon photonic systems implemented in existing CMOS processes with zero changes to the process. This effort has demonstrated basic components, such as waveguides, filters and modulators in 45-nm [?, 93, 99] and 32-nm technology nodes []. This "zero-change" approach (initially funded by DARPA []) has matured to the point where all-optical communication with 11 wavelength-division multiplexed channels was used between a processor and DRAM [?, 57] in the same 45-nm silicon-on-insulator process that was used to create the IBM Power 7 processor (Watson, PlayStation 3). This feat represents a significant milestone in the technological trajectory connecting global photonic networks down to optoelectronic systems on a single chip, perhaps fulfilling Soref's vision of a superchip. In this work, the III-V light sources are external to the silicon chip with fiber coupling between. Some in the field contend this will remain the most tractable solution in the long term.

Significant commercial interest has persisted in this field since the founding of Luxtera, including major efforts by Intel [], and continuing with start-ups spinning out of the zero-change work []. All of these efforts attempt to use light as a means to communicate digital signals between electronic processors, whether it be at scale of a single chip [57], a server rack [] or a data center []. As has been the case in semiconductor electronics and superconductor electronics, the device and hardware infrastructure developed for digital information processing is now being explored for neuromorphic information processing.

6.4 Deep learning with silicon photonics

Like superconducting neural systems, the goal of nearly all efforts in optoelectronic neural systems and neuromorphic photonics is not to develop general intelligence, but rather to realize neural systems for specific tasks such as inference or control. For most efforts, the motivation for using light is the speed, either of laser cavity dynamics or optical communication. Device and hardware choices toward these ends may be different than for the focus of this article, which is general intelligence. We intend to explain why specific choices are not conducive to the present goal, even if they are suitable for other applications.

We consider deep learning to be based on feed-forward networks of non-spiking neurons trained through a supervised algorithm such as backpropagation. While markedly different from the recurrent networks of dynamical nodes that learn from experience through local plasticity mechanisms, the relative simplicity of deep learning makes it a natural place to begin utilizing principles of neural information processing. Feed-forward neural networks have been studied with free-space optics since the height of optical computing excitement in the late 1980s and early 1990s, and after the developments in silicon photonics following Soref, similar principles were developed in an integrated context.

The operation of synaptic weighting in deep learning reduces to matrix-vector multiplication. Such an operation can be achieved with an array of Mach-Zehnder interferometers. A recent demonstration accomplished this using thermo-optic phase shifters with silicon waveguides [100]. A network with four inputs and outputs was trained to classify four vowel sounds. The effort led to two start-up companies attempting to commercialize the technology to compete with specialized CMOS processors (such as tensor processing units) for deep learning. The photonic approach demonstrated so far made use of off chip light sources and detectors, and applied the nonlinearity in software. For such an approach to be competitive, significant system integration is required. The two senior authors of Ref. [100] have more recently moved back to a free-space approachin to deep learning [].

The approach of using 2-D arrays of interferometers for routing and synaptic weighting pursued in Ref. [100] is incompatible with large-scale cognitive systems for several reasons. One reason is that the index shifts induced by thermo-optic phase shifters are small, and power dependent, leading to either large structures, high power consumption, or both. Cross talk between thermal elements necessitates placing the waveguides far apart, and it is difficult to utilize the vertical dimension interferometer arrays, so attempting to scale results in networks that are sprawling in the plane. Further, as described in Sec. 3, an important mechanism of learning in spiking neural systems is through STDP, wherein the activity of the two neurons associated with a synapse leads to memory adaptation. With interferometer arrays, changing a single phase in the network will, in general, modify several synaptic weights. Therefore, while backpropagation can be implemented with such a network [], STDP cannot.

6.5 Spiking neurons with compound semiconductor lasers

While the interferometric approach to deep learning discussed above makes use of static neurons, several approaches to spiking neurons have been pursued as well. One class of spiking photonic neurons leverages the carrier dynamics in compound semiconductor laser cavities. It has long been known that the equation governing lasers with gain and saturable absorber regions are isomorphic to the leaky integrate-and-fire neuron [**?**], with the number of excited carriers in the laser playing the role of the membrane potential. This correspondence has led to several designs [**?**] and experimental efforts (see Ref. [102] and reference therein) to leverage this behavior to make spiking neurons that sum optical signals and produce optical pulses when a threshold has been reached. This work began in Er-doped fibers, and continues with on-chip implementations with III-V photonic systems, with much of the work being done in the Prucnal's group at Princeton. The refractory period of such neurons is set by the cavity photon decay time and is on the order of 10 ps, while the integration time is set by the carrier relaxation time, and is on the order of 100 ps. This short refractory period means such neurons can fire up to 10^9 times faster than biological neurons, yet the short integration time means temporal correlations amongst neuronal firing events is forgotten rapidly.

While the goal of these efforts in excitable lasers is to perform neuro-inspired computing very rapidly with small networks, and not to achieve brain-scale systems, we nevertheless point out two features of this approach to using light in neural systems that are not conducive to achieving large-scale systems. The first is power consumption. To properly set the threshold of these neurons, the gain region must be continuously pumped. This requires between 100 mW and 1 W per neuron, even when the neuron is not firing. For a system of 10^{10} neurons, a gigawatt would be consumed, even with the system at rest. The second limitation regards computation. As discussed in Sec. 3, neural information processing leverages many complex computations in synapses, dendrites, and neurons. In excitable lasers, all the computation occurs in the interaction between photons and carriers in the laser cavity. Multiply-accumulate operations can be performed with leak and threshold, but no path toward short-term synaptic plasticity or dendritic processing have been proposed. By relying on the exponential decay constants of photons and carriers, one is unable to tune the range of temporal information processing or supply the dendritic arbor with information across a wide range of temporal scales. These computations and time constants are more readily achieved in the electronic domain with circuits that can be engineered to perform complex functions rather than relying on material parameters, a point we revisit below.

6.6 Wavelength-division multiplexing for routing and synaptic weighting

In addition to the work on excitable lasers as spiking neurons, the Princeton group has also pioneered the use of concepts from wavelength-division multiplexing for both signal routing and synaptic weighting [**?**, **?**]. Within this framework, each neuron within a cluster produces or modulates light at a distinct wavelength upon firing. The signals from all neurons within the cluster are multiplexed onto a single broadcast waveguide, and all other neurons tap all colors from this waveguide and apply synaptic weights based on the frequencies of microring resonances relative to the neuron wavelengths. For a cluster of N neurons, N different colors of light must be generated, N microring filters must be used to multiplex these signals onto the broadcast waveguide, and each neuron must have $N-1$ microring filters to receive and weight the signals from all the other neurons. Thus, a cluster of N neurons requires N^2 microring resonators. This approach to communication between neurons is referred to as "broadcast-and-weight", and is closely related to the operation of wavelength-division multiplexing in fiber communication networks.

Again, the goal of the work from the Princeton group is not to achieve brain-scale systems, but rather to "...find out the minimum ensemble of behaviors that are necessary to harness similar processing advantages." [102] Nevertheless, adopting wavelength-division multiplexing concepts from larger-scale communication networks down to the chip scale is intuitive and aesthetically appealing, so it is worth pointing out why it ends up not being conducive to reaching large-scale cognitive systems. To begin, it is important to distinguish between using the wavelength of light for multiplexing multiple signals on a broadcast bus and the use of microring resonators to establish synaptic weights. The Princeton group uses both techniques, but it is possible to employ one or the other independently. When using wavelength for multiplexing, the advantage is that space can potentially be saved. Instead of each neuron having an independent axonal arbor to reach its downstream connections, many neurons share a single distribution waveguide. However, the area saved is significantly reduced by the fact that N^2 microring resonators must be employed. More important than area is power. Because microring resonances are so sensitive to minor variations in fabrication, each of the N^2 resonators must be actively aligned to the appropriate wavelength corresponding to the emission from the associated neuron. This typically requires on the order of 1 mW. For a brain-scale system of 10^{14} synapses, 100 GW would be required just to align the communication network. The power consumed for alignment limits scalability, but so does the procedure for carrying out the alignment. Each of the microrings must be aligned, and if thermal tuning is employed, significant cross-talk will occur. Implementing such alignment for systems of more than a few neurons becomes quite cumbersome. Additionally, the wavelengths of the neurons can only be spaced so closely if cross talk is to be avoided, and the gain bandwidth of the light sources is limited, so a limit of roughly 200 neurons within a cluster is encountered. One may think of such a cluster as analogous to a mini-column in the brain, but unfortunately communication between mini-columns is hindered by the use of wavelength for multiplexing. In order to communicate between mini-columns, a neuron must first communicate from its local cluster up to a higher level of hierarchy where the same colors are re-used, and then down again to the target cluster. Such a communication protocol severely limits the graph structures and path lengths that can be achieved (see Sec. 3. It is intuitive to leverage wavelength multiplexing in photonic neural systems to maximize use of bandwidth, but when used in this way wherein each neuron is uniquely identified by a color, scalability is severely hindered.

These considerations pertain to using wavelength for multiplexed routing, but there are independent reasons why using microring resonators to establish synaptic weights is not conducive to scaling. One challenge associated with microring weight banks is the fact that by changing a certain parameter (power delivered to heater, for example) the synaptic weight first increases, then saturates, the decreases as the resonance passes the target wavelength. This makes it very difficult for supervised or unsupervised learning to occur. Additionally, the shape of the resonance is nonlinear with very steep sections. Thus, to achieve uniform changes in synaptic weight, a nonuniform change in drive must be applied, and across much of the range of weights, the synaptic weight will be very noisy.

Microring weight banks and Mach-Zehnder interferometer networks have two things in common: they both require implementing phase shifts in photonic components (which usually draws power, even in the steady state), and neither is capable of implementing STDP or other unsupervised learning techniques. To achieve the largest-scale neural systems, it is highly advantageous if storage of a synaptic weight draws no power. For a system at the scale of the brain, if each synapse draws even 10 nW in the steady state, the system will consume 1 MW just to remember what it has learned.

6.7 Phase change materials for synaptic weighting

One technique for establishing synaptic weights between neurons signaling with light is to leverage phase-change materials [?]. Such materials have the property that the coefficient of optical absorption is different between the two phases. Therefore, a variable attenuator can be devised wherein the crystallization state of a small patch of phase-change material integrated on a waveguide determines how many photons are transmitted through the synapse. Reference [?] showed that such a synapse could be used to implement a form of Hebbian learning, wherein two pulses incident closely in time could strengthen the synaptic weight by adjusting the crystallinity of the material and reducing absorption.

Such Hebbian update in this system represents a novel route toward synaptic weighting in photonic neural systems. Unfortunately, the material studied in Ref. [?] requires billions of photons for Hebbian update, thereby exceeding the communication energy limit of a single photon by at least nine orders of magnitude. Additionally, the patch of phase-change material has no way of keeping track of the order in time or even the source of input pulses, so anti-Hebbian synaptic weakening cannot be achieved, and a route to full STDP has not been proposed.

6.8 Synaptic weights in the electronic domain

We have discussed here three approaches to establishing synaptic weights in photonic neural systems: interferometric networks; microring resonators; and phase change materials. These approaches all have one thing in common: they treat the synapse as a variable attenuator, and change the weight by varying the number of photons that pass through the synapse. Communication in biological neural systems is binary, and the synaptic weight is enacted based on how much post-synaptic current is generated, and is independent of the amplitude of the action potential reaching the pre-synaptic terminal. By contrast, if one establishes the synaptic weight in the photonic domain, communication is analog, and the number of photons in the pulse—analogous to the amplitude of the action potential—now carries information. This has two detrimental consequences. First, it requires that each neuron produce more photons that would be necessary for binary communication, and many photons are discarded at weak synapses. This is a power penalty. Second, setting the synaptic weights in the photonic domain means that any noise on the transmitting neuron light sources results in additional noise received by the neuron. This is an information-processing penalty.

The alternative is to set the synaptic weights in the electronic domain. The synaptic response is independent of the number of incident photons, and the synaptic weight is stored and implemented by an electronic circuit. Provided a synaptic terminal receives a photonic signal surpassing a certain threshold, a synaptic event is induced. The physical limit on the amplitude of this threshold signal is a single photon. Establishing the synaptic weight in this manner is most straightforward if each synapse is equipped with an independent photodetector. For integration with CMOS, the waveguide-integrated SiGe or defect detectors described above are good candidates. Logic circuits based on MOSFETs are the clear choice to implement synaptic, dendritic, and neuronal computations, and transistors operated in analog may play a role. Upon reaching threshold, the transistor circuits would drive a pulse through an on-chip laser, and the light thus produced would fan out to downstream connections. At those connections, as long as a number of photons greater than the threshold were received, the synaptic response would ensue, thus eliminating the effects of any noise on the photonic communication signal. The challenge here is the same at that mentioned above: it is hard to integrate light sources on silicon. If a million III-V or SiGe sources can be integrated on a 300-mm silicon optoelectronic wafer in a cost-effective manner, such an approach to optoelectronic networks will be viable.

To reach the physical limit of single-photon synaptic threshold, superconducting-nanowire single-photon detectors (SPDs) can be used. We will describe these detectors in more detail in the next section, but for the present discussion we point out that these detectors respond to single photons, and their response is nearly identical [] if one or more than one photon is detected. Thus, neuronal communication using these detectors enables the lowest possible communication signal level, and sources must produce only enough photons per synaptic connection so that even with noise, each synapse receives at least one photon, with a chosen tolerable error rate. Such communication appears to saturate a physical limitation for neuronal signaling with photons of a given wavelength. Whereas transistors were used for computation in the hardware example above, if SPDs are used for detection, circuits of JJs the clear choice for computation. Because SPDs and JJs both require operation near 4.2 K, optoelectronic hardware operating in this modality has the potential to utilize silicon light sources, potentially bringing a tremendous advantage in cost and scalability. In the next section we will describe the synaptic, dendritic, and neuronal functions of these circuits.

7 Superconducting optoelectronic systems

These detectors are wires of superconducting material [**?**], and they can be straightforwardly patterned atop on-chip dielectric waveguides [**?**, **?**, **?**, **?**]. The nanowire thickness is 4 nm-10 nm, its width is 80 nm-350 nm, and the interaction length along the waveguide can be as short as 10 µm. The wire is current biased in parallel with a resistive load (Fig. **??**), .

Because the detectors are superconducting, they draw very near zero power in the steady state.

- general concept: communication between neurons is photonic; when a neuron spikes it must either generate or modulate light; throughout, speed, size, power all co-optimized
- first key choice: generate or modulate
- modulate:
  - requires cw light running at all times ($x_{dB/cm} = 1; y_{dB/s} = 100 * x_{dB/cm} * c; q_{dB} = 3; t_s = q_{dB}/y_{dB/s}$, for 1 dB/cm propagation loss, 3 dB of the light is lost every 100 ps)
  - requires frequency tuning, most likely
  - cross talk of neurons on the same bus
- generate:
  - requires light source at every neuron
  - requires unprecedented optoelectronic integration, million sources and a billion detectors on a wafer
  - must be very low capacitance
  - seems like only a silicon light source will suffice, but this would require cryogenic operation
- second key choice: establish synaptic weight in the photonic or electronic domain?
- photonic domain:

- This choice has several important ramifications for hardware and information processing. Regarding information processing, it is usually assumed that neural communication is digital: the presence or absence of an action potential is a binary one or zero, and the amplitude of the action potential is not encoding information. When adjusting the synaptic weight in the photonic domain, this is not the case. The number of photons reaching a neuron through a synaptic connection becomes an analog variable, and it is subject to shot noise, in addition to any noise mechanisms present in the detector. The signal-to-noise ratio of shot noise improves with $\sqrt{(N_{\mathrm{ph}})}$, where in this case N_{ph} is the average number of photons, so establishing weights in the photonic domain introduces an energy/noise tradeoff. Setting weights in the photonic domain also has the disadvantage that photons are discarded by attenuation at weak synaptic weights. Thus, by setting synaptic weights in the photonic domain, we place a burden on light sources to produce large numbers of photons to minimize shot noise, and we discard photons when they are attenuated at weak synapses. In this mode of operation, light is used for communication, but it is also used for the important computational operation of applying the synaptic weight.
- these objections notwithstanding, to our knowledge, all except one optoelectronic neural approach proposed to date sets weight in photonic domain
- specific instances: mzi (no STDP, poor spatial scaling, cross-talk); wdm (limited number of channels, cross-talk with rings on master ring, demands on sources); mzi and wdm (thermal tuning hopeless for scaling, no plasticity mechanisms proposed); phase change synapses (at least don't dissipate steady state, still power lost due to variable attentuation, small footprint, Hebbian learning possible, but STDP not likely, meta, short term also doesn't look promising)

* electronic domain:
  - By contrast, if we establish synaptic weights in the electronic domain, light is used exclusively for communication, and communication remains entirely digital. The presence of an optical signal can be used to represent an all-or-none communication event. In this case, the detector and associated electronics must be able to achieve a variable synaptic response to identical photonic pulses based on the configuration of the electronic aspects of the circuits. In this case, we expect that a neuron will send, on average, N_{ph} photons to each of its downstream synaptic connections. Due to shot noise, each downstream connection will receive $N_{\mathrm{ph}} \pm \sqrt{N_{\mathrm{ph}}}$ photons, and the detector circuit must be configured to implement a synaptic response if a threshold of N_{th} photons is detected. After detection, the electronic response must vary depending on the synaptic weight, independently of the precise number of photons that was detected. It is in this electronic response that the signal becomes analog again. Whereas setting the synaptic weights in the photonic domain places a larger burden on light sources, setting the synaptic weights in the electronic domain places a larger burden on detector circuits. One must achieve a detector circuit that converts light pulses to electrical current or voltage, and the amount of electrical signal must be largely independent of the number of photons in the pulse, depending instead on reconfigurable electrical properties of the circuit, such as bias currents or voltages. These reconfigurable bias currents or voltages then represent the synaptic weights, and the task of a neuron's light source is simply to provide a roughly constant number of photons to each of its downstream synaptic connections. For energy efficiency, the number of photons necessary to evoke a synaptic response from the detector (N_{th}) should be made as low as possible to make the job of the light source as easy as possible. N_{th} cannot be made lower than one, as the electromagnetic field is quantized into integer numbers of photons.
  - only know of one system where electronic domain has been proposed: soens
  - basic functionality
  - stdp
  - meta
  - homeo
  - short-term
* neuronal computation: reaching threshold
  - differentiate between state-based and spiking
  - main considerations here are energy/power
  - how much energy is required to generate a pulse or drive a modulator?
  - how much light must be made/moved to drive all downstream synaptic connections?
  - how fast can pulses be generated (refractory period)?
  - how long can neurons remember (leak rate)?
  - what is range of spike rates? what is expected power?

- somewhere in here, comparison of detectors (going cold costs 500x for carnot, but gains 2000x for detector sensitivity)
- related, comparison of sources (going cold reduces how many photons must be made, but most importantly, if it means a silicon light source can work for this project, it is a game changer)
- inhibition, gotta have a plan
- dendritic processing
  - intermediate nonlinearities
  - direction attention with inhibition
  - sequence detection
  - how can any of this happen in the photonic domain?
- room temp vs cryo
  - sources (cryo enables Si sources. for large-scale integration, process simplicity brings tremendous advantage)
  - detectors (A SiGe photodetectors needs about 10^4 photons in 100 ps to respond; efficiency of SNSPDs, low-noise of SNSPDs, simplicity of fabrication, and excellent operation in conjunction with JJs)

8 Large-scale optoelectronic systems

8.1 Criteria for assessing cognitive neural hardware

- network metrics
  - total number of neurons in network
  - degree distribution
  - clustering coefficient on different scales
  - modularity analysis
  - rentian analysis
- synaptic metrics
  - range of achievable synaptic weights
  - number of achievable synaptic weights
  - max synaptic state retention time
  - analysis of short-term plasticity responses (filter properties, energy consumed, area)
  - analysis of stdp (range of values, update rates, energy required, time window control, area required)
  - analysis of metaplasticity (mechanisms, range of rates, energy required, area required)
  - total size of synapse
- dendritic metrics
  - analysis of dendritic intermediate nonlinear processing (range of time scales, I/O curves, energy, num synapses per dendrite)
  - analysis of dendritic sequence detection (time scales, number of synapses involved, energy)
  - available logic operations (Boolean, time of last firing, multisynaptic)
  - number of values achievable in readout
  - size of dendrites
- neuronal metrics
  - total number of dendrites
  - total number of synapses
  - analysis of the rentian fan-in of the dendrito-synaptic arbor
  - dynamic properties of threshold
  - refractory period
  - integration time (if different from dendritic integration)
  - energy of neuronal firing
  - timing jitter of neuronal firing
- communication metrics
  - total system bandwidth (at neuronal level, also including dendritic level where neuronal data rates are multiplied by information available to first layer of dendrites after any additional synaptic fanout)
  - I/O
- system operation metrics
  - operating temperature
  - power consumption
  - max spike events per second
  - power consumed during max spike events per second
  - power density
  - temperature variation during operation
  - total system size
- system production metrics
  - equipment required (i.e., technology node)
  - device yield/tolerance to variation
  - time required

  - packaging strategy
  - total material consumed
  - total cost to produce

- unprecedented integration of photonics and electronics in a scalable process that can be implemented with existing infrastructure–change a few implant conditions, swap out a few sputtering targets, improve BEOL dielectrics for photonics

- communication on various length scales, multi-planar on wafers, wafer-to-wafer vertical and lateral, fiber white matter

- feasibility of brain scale

- why si if no transistors?

  - III-V substrates should be pursued as well. Our group is working on this, initial anecdata indicates similar efficiency
  - big problem is fab. wafers are harder to scale, material harder to purify, oxide not as good for waveguide cladding. Similar consideration to mosfet gate. Overall manufacturability
  - may eventually use transistors for perhaps faster refractory period

- ultimate limits

For general networks, the algorithm by which partitions are identified can be made mathematically rigorous from a network theory perspective [23, 101]. For the analysis at hand, we consider the partitions of the network we have assigned in the hierarchy. For example ... (minicolumns, mesocolumns, macrocolumns on a wafer(s); multi-columnar modules, ...)

9 Application spaces

- original applications of computing

  - cryptography
  - weather
  - bombs
  - numerical solution to arbitrary diff eqs
  - from Turing, AI
  - now, apply to nearly all aspects of modern life

- what others in the field are pursuing

  - LASSO [60]
  - fast control [102]

- here, following neuroscience applications

  - vision systems
  - language processing
  - motor control
  - may lead to Turing's vision of an AGI one can interact with

- others, unique to large-scale neural systems and/or superconducting optoelectonic

  - internet monitoring/simulation
  - sociological simulation
  - genetic analysis/evo devo
  - neuroscience and dynamical systems
  - quantum/neural hybrid systems (Bayesian discussion)

9.1 Classical-Quantum-Neural hybrid systems

Quantum systems and neural systems have complimentary information processing capabilities. Quantum systems are fundamentally probabilistic, while neural systems are excellent for sampling probability distributions. Schemes to utilize quantum information are usually statistical, while populations of neurons can perform optimal Bayesian inference on samples drawn from statistical distributions. This reasoning leads us to consider the potential to utilize a neural system to perform quantum state tomography on large-scale quantum systems. The goals of the project are to construct a neural system capable of: 1) measuring the state of a network of qubits at the Heisenberg limit; 2) inverting the physical measurement through Bayesian inference to arrive at a quantum state reconstruction; and 3) reporting the reconstructed state over a classical communication channel as the qubits evolve in time, all implemented in scalable hardware.

Quantum information processing requires the ability to determine an unknown quantum state from a series of measurements performed on an ensemble of identically prepared systems. Performing measurements on many interacting qubits places severe demands on measurement hardware. To characterize a large quantum system, the number of measurements that must be performed can become intractably large if care is not taken to optimize the measurement protocol [1,2]. Additionally, the computational challenge of reconstructing the full quantum state from the set of measurements is formidable for large quantum systems. Developing hardware with classical, quantum, and neural capabilities presents an alternate route to develop scalable measurement techniques to extract Heisenberg-limited information [3] from a complex quantum system, to devise a method for a full quantum state to be efficiently reconstructed from measured data, and to ensure that the hardware implementation of this measurement/analysis procedure communicates efficiently to room temperature.

At present, the various elements of scalable quantum state tomography are maturing and beginning to combine. In hardware, control and measurement circuits operating at cryogenic temperature are being developed. Josephson circuits capable of detecting single microwave photons present an exciting new avenue for scalable qubit characterization [4], yet racks of control and readout electronics are still employed for interfacing to relatively small quantum systems. Regarding reconstruction of quantum states from measured data, statistical methods involving Bayesian inference have been developed in the context of quantum tomography over the last 30 years [5-9]. It has been shown that by reoptimizing the measurements to be performed as information about the quantum state is acquired, the total number of measurements can be reduced [10]. Modern techniques in machine learning have been applied to the problem, showing that a neural network can perform quantum state tomography on highly entangled states of a hundred qubits [11]. The neural network employed in the tomographic analysis of Ref. 11 is a conventional, feed-forward neural network implemented in software. The related field of spiking neural networks has found that populations of spiking neurons naturally perform Bayesian inference [12-14]. Networks of spiking neurons can be trained so the average firing rate of the population represents the expectation value of an observable, and the variance of the firing rates of the neurons represents the uncertainty. Bayesian analysis has been applied to a series of weak measurements to track the trajectory of a single qubit, but cumbersome measurement hardware infrastructure was utilized [15]. Software neural nets have been used to perform tomography on 100-qubit systems, but the measurements were performed conventionally [11]. The adaptive Bayesian formalism has been applied to a two-qubit system to minimize the number of measurements required for full state reconstruction, but explicit numerical calculations were performed on conventional computers between each measurement [2]. Spiking neural networks have been used to perform Bayesian inference on statistical distributions [16], but the systems under observation have all been classical [14]. It has been shown that neural networks can emulate quantum computation [17-19], can accurately measure quantum systems [20], and can perform quantum state tomography [11]. The proposed hybrid hardware would combine these advances in a measurement apparatus performing Heisenberg-limited measurements, conducting Bayesian inference in real time as information is received, and using all knowledge about the quantum system for optimization of measurement protocol and full state reconstruction. The metrological hardware we propose to develop would prepare highly entangled states of many qubits, perform the measurements and information processing necessary for tomography, and communicate the results of state reconstruction to room temperature with near-infrared photons over optical fiber [21].

This concept remains in the domain of thought experiments, but we will describe a route to make it real. We propose to model and construct the classical-quantum-neural (CQN) hybrid system depicted schematically in Fig. 1. The system comprises a classical control module, a quantum module containing a network of coupled qubits, and a neural module interfacing with both the classical and quantum systems. The envisioned operation of the CQN system is as follows. The classical system prepares the quantum system in a particular state. The quantum state is set by a static many-body Hamiltonian and a series of qubit drive pulses. The classical system also provides the neural system with information representing the static and drive Hamiltonians. The classical system generates microwave signals to probe the quantum system. We envision the measurement signals perform a series of weak measurements on a time scale short relative to the qubit decoherence and relaxation times [15], but projective measurements could be employed as well. The ability of weak measurements to give information as a function of time while a quantum state evolves fits nicely with the dynamical operation of spiking neurons. The output from these measurements is a faint microwave signal, and information about the state of the qubits is encoded in this signal. The neural system comprises an input layer, a computational reservoir, and an output layer. The input layer receives the faint photonic signals, and the dynamical state of the reservoir evolves in response to the varying signals received from the quantum system. To function as proposed, the input layer of the neural network should represent expectation values of the qubits in the quantum system, and the operation of the network should be to invert that information into a hypothesis regarding the density matrix [6,9], encoded in spike trains by the output neurons. Mathematically, we usually assume we know the density matrix and can therefore calculate the expectation value of any operator. In practice, one measures expectation values and infers the density matrix from the data. This is the inversion operation that will be performed by the neural system.

Superconducting circuits appear capable of realizing this CQN system. We propose to develop the quantum system based on transmon qubits operated in the dispersive regime, probed via microwave signals along transmission lines. Josephson arbitrary waveform synthesis will be utilized to generate the microwave qubit control and measurement signals, and a superconducting FPGA based on flux-quantum logic and magnetic Josephson junction memory elements will control the operation of the entire apparatus. The neural system will receive the microwave signals transmitted from the classical system through the quantum system, and therefore the input synapses to the neural system must respond to faint microwave signals to perform Heisenberg-limited observation of the quantum system. As the size of the quantum system grows, so must the neural system. To achieve the required communication across the large neural system, photonic connectivity is required, making superconducting optoelectronic loop neurons [21] promising as device primitives for the neural system. The optical signals from these neurons bring the added advantage that information is transduced to optical, and can be readily coupled to fiber for transmission out of the cryostat for further processing with CMOS circuits.

References: [1] C. Granade, J. Combes, and D.G. Cory, âĂIJPractical Bayesian tomographyâĂİ, New J. Phys. 18 (2016). [2] G.I. Struchalin, I.A. Pogorelov, S.S. Straupe, K.S. Kravtsov, I.V. Radchenko, and S.P. Kulik, âĂIJExperimental adaptive quantum tomography of two-qubit statesâĂİ, Phys. Rev. A 93 (2016). [3] A.A. Clerk, M.H. Devoret, S.M. Girvin, F. Marquardt, and R.J. Schoelkopf, âĂIJIntroduction to quantum noise, measurement, and amplificationâĂİ, arXiv:0810:4729 (2010). [4] A. Opremcak, I.V. Pechenezhskiy, C. Howington, B.G. Christensen, M.A. Beck, and R. McDermott, âĂIJMeasurement of a superconducting qubit with a microwave photon counterâĂİ, Science 361 1239 (2018). [5] K.R.W. Jones, âĂIJPrinciples of quantum inferenceâĂİ, Annals of Physics 207 (1991). [6] K.R.W. Jones, âĂIJFundamental limits upon the measurement of state vectorsâĂİ, Phys. Rev. A 50 (1994). [7] R. Derka, V. Buzek, G. Adam, and P.L. Knight, âĂIJFrom quantum Bayesian inference to quantum tomographyâĂİ, arXiv:quant-ph/9701029 (1997). [8] R. Schack, T.A. Brun, and C.M. Caves, âĂIJQuantum Bayes RuleâĂİ, Phys. Rev. A 64 (2001). [9] R. Blume-Kohout, âĂIJOptimal, reliable estimation of quantum statesâĂİ, New. J. Phys. 12 (2010). [10] F. Huszar and N.M.T. Houlsby, âĂIJAdaptive Bayesian quantum tomographyâĂİ, Phys. Rev. A 85 (2012). [11] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, âĂIJNeural-network quantum state tomographyâĂİ, Nat. Phys. 14 (2018). [12] W.J. Ma, J.M. Beck, P.E. Latham, and A. Pouget, âĂIJBayesian inference with probabilistic population codesâĂİ, Nature Neuroscience 11 (2006). [13] T. Yang and M.N. Shadlen, âĂIJProbabilistic reasoning by neuronsâĂİ, Nature 447 (2007). [14] J.M. Beck, W.J. Ma, R. Kiani, T. Hanks, A.K. Churchland, J. Roitman, M.N. Shadlen, P.E. Latham, and A. Pouget, âĂIJProbabilistic population codes for Bayesian decision makingâĂİ, Neuron 60 (2008). [15] K.W. Murch, S.J. Weber, C. Macklin, and I. Siddiqi, âĂIJObserving single quantum trajectories of a superconducting quantum bitâĂİ, Nature 502 (2013). [16] M.J. Barber, J.W. Clark, and C.H. Anderson, âĂIJNeural representation of probabilistic informationâĂİ, Neural Computation 15 (2006). [17] C. Wetterich, âĂIJQuantum computing with classical bitsâĂİ, arXiv:1816.05960 (2018). [18] C. Pehle, K. Meier, M. Oberthaler, and C. Wetterich, âĂIJEmulating quantum computation with artificial neural networksâĂİ, arXiv:1810.10335 (2018). [19] G. Carleo and M. Troyer, âĂIJSolving the quantum many-body problem with artificial neural networksâĂİ, Science 355 (2017). [20] D.T. Lennon, H. Moon, L.C. Camenzind, L. Yu, D.M. Zumbuhl, G.A.D. Briggs, M.A. Osborne, E.A. Laird, and N. Ares, âĂIJEfficiently measuring a quantum device using machine learningâĂİ, arXiv:1810.10042 (2018). [21] J.M. Shainline, S.M. Buckley, A.N. McCaughan, J. Chiles, A. Jafari-Salim, R.P. Mirin, and S.W. Nam âĂIJCircuit designs for superconducting optoelectronic loop neuronsâĂİ, J. Appl. Phys. 124 (2018).

$$\langle \mathcal{O} \rangle = \mathrm{Tr}(\mathcal{O}\rho) \tag{2}$$

10 Outlook

- circle back to Turing and von Neumann, their interests in machine intelligence and modeling computation after the brain
- circle back to digital vs neural, superconducting optoelectronics brings communication and spiking nonlinearities
- why go to all the trouble?
  - this technology will only be pursued if it can do something that nothing else can do
  - but it can, and what it can do is very important
    * exceptional complexity for experiments in network information, neuroscience models
    * quantum/neural hybrid systems
    * scaling beyond what is possible with other methods, perhaps the smartest machines on the planet
    * computing has shaped economy and society since its inception
    * powerful scientific tool
    * foundational questions about thought and consciousness amongst the most intriguing and important in modern science

We conjecture that Lovelace and Turing were both right. She was right that computing machines as they were known to her, and with the serial processing Turing proposed, really are not up to the challenge of thinking. And he was right that a machine can be capable of thought and learning like a child, but to do so, a modality of operation significantly different from the sequential instruction execution of the Turing machine must be employed.

Working in the field of beyond-CMOS computing hardware, one quickly absorbs the mantra: never underestimate CMOS. Working the the field of hardware for AI, one quickly absorbs the wisdom: never underestimate the brain. We recognize the audacity in proposing hardware to outperform CMOS for any task. Yet we think the arguments presented here make the case that it is worth pursuing silicon-based technology with superconductors, light sources, and waveguides instead of transistors and electrical interconnects for cognitive neural systems. Does this mean we are confident such hardware will lead to beyond-human intelligence? Not at all. We understand CMOS, and we know what its limitations are likely to be. But the brain maintains important secrets, even after hundreds of years of inquiry. We have laid out an architecture that achieves fractal scaling over many orders of magnitude, and appears promising for enabling communication across the hierarchy at speed far greater than biological systems. And we have tried to respect the complexity of synaptic, dendritic, and neuronal functionalities in our circuit concepts. But it is possible that the subtleties of neuronal devices and architectures are more clever than we presently comprehend, and the structures we have discussed—from circuits to systems—will not achieve the nuanced information processing that leads to advanced cognition. For example, at the device level, synapses communicate with many neurotransmitters that can be modulated independently and affect information processing differently. At the architecture level, the thalamus coordinates information processing and enables access to the global neuronal workspace in a masterful manner that unifies the signals from many brain regions into a coherent cognitive moment. It is not clear that the circuits presented here will achieve comparable complexity, and it is not clear that we will soon understand how, with optoelectronic systems, to implement something like the thalamocortical complex that integrates information across the entire network architecture. It is our perspective that progress beyond the present state requires a significant experimental effort. Hardware must be devised, and networks must be observed. Only then will we find the limits of what can be made and how well it can process information.

Misc. notes:

- origins of modern computing intertwined with WWII
- Turing: interests, universal computation, computability, Turing machine, serial, cryptography
- von Neumann: interests, universal computation, numerical investigation of numerous physical problems, numerically solving differential equations, digital computing, memory storing data and instructions, von Neumann bottleneck
- cryptography leads to creation of Turing machines one side of the Atlantic, numerical analysis of nuclear weapons leads to creation of Turing machines on the other (Dyson, pg 257)
- Shannon: communication, information in data streams, again focus is on serial information processing
- computing hardware: vacuum tubes, punched cards lead to silicon microelectronics, si uniquely suited to accomplishing digital computing, von Neumann architecture still going strong in si
- communication hardware: ethernet for pretty big networks, fiber-optic cables replacing telegraphs under the atlantic
- silicon photonics is where these two meet: light for commmunication, electronics for computation, maintaining the von Neumann architecture, WDM across the von Neumann bottleneck
- Turing's discussion of ingenuity and intuition (Dyson pg. 252): digital all ingenuity, brute force search; neuro brings intuition back and honors its role; populations of neurons enable intuition to be based on Bayesian inference rather than random guesses.
- Turing says "ingenuity replaced by patience". This is very much what happens in digital neuro. Brute search takes to long to enable neural information processing.
- computing, communication theory, and cryptography all advanced significantly during and in response to WWII. The 80 years from 1938 to 2018 have seen the emergence of transformative technologies in these fields. Much contemporary work follows in these veins. For example, the goals of a universal quantum computer are very similar to Turing's universal calculator, with the addition of quantum physics to dramatically increase the speed of certain algorithms. Because quantum states are fragile and subject to decoherence, quantum systems strike us as very poorly suited to perform the serial operations of a Turing processor requiring writing and reading to reliable memory. Nevertheless, the requirement for cryptography in a world where trust is unfounded is sufficient motivation for many to pursue quantum computers, if for no other reason than to perform Schor's algorithm.
- Grover's search algorithm is another motivator, again following Turing's line of reasoning to replace intuition with ingenuity, and ingenuity with brute search. The problem is the physics and hardware we have at our disposal make it very difficult to realize machines capable of performing these operations efficiently.
- In addition to limitations resulting from the fact that it is hard to implement quantum computers in hardware, some problems simply do not map well onto Turing machines, no matter the machine's complexity. Embracing the duality of ingenuity and intuition,

as a neural system does, is increasingly useful for solving many of our present problems, including those of national security and defense, and extending into realms of medicine and science.

- “The paradox...to understand.” [4] pg. 263.

To put the present discussion of cognitive neural systems in context, we must revisit the origin of computing. Nearly all modern computing is based on digital (binary) information processed in a von Neumann architecture, which was devised as a means to realize a Turing machine in electronic hardware.

Turing’s 1936 paper, indisputably a record of historical brilliance, has been so impactful in part due to the simplicity of the concept. It is tractable to contemplate the potential operations of a single read/write head following one-dimensional instructions. Describing this model, Turing was able to produce formal proofs about the universality of the apparatus. Yet questions of efficiency remained.

It is far less tractable to mathematically model the capabilities of a system of a hundred billion interconnected dynamical nodes in a network of high topological dimension, as we find in the brain. It would not have been sensible to pursue neuromorphic computing systems until the limits of the Turing machine were reached, particularly considering Turing proved his machine could determine the result of any computable function.

By following the evolutionary history of the concept of a Turing machine followed by the implementation with the von Neumann architecture, it is natural to pursue neural systems with many processors following the instructions to compute the results of the equations used to model neurons. The downside of this approach to numerical emulation of neural systems is the inefficiency relative to the performance achieved by hardware embodying neural operations based on the physics of the constituent devices.

Silicon photonics provides three primary dielectric materials that can be used for these passive waveguides: Si, SiN, and SiO_2. The indices of refraction of these materials are 3.5, 2.0, and 1.5, respectively, for λ close to 1550 nm. These are the three primary dielectrics used in CMOS technology as well.

At different times, a neuron firing is known by other neurons to mean different things.

Reasons to publish (addressing concerns of Bostrom)

1. New/early/infancy; significant development required, both concept and hardware 2. If soens ever did prove feasible, crossover would be Bostrom’s slow category 3. Will require concerted effort, probably at least 100s of people, money, foundry 4. A specific hardware proposal has the potential to offer a useful case study, perhaps leading to preparedness 5. I am an employee of the Federal Government in service of the US taxpayers, and I have an obligation to publish my research. 6. If a superintelligence powerful enough to rapidly overcome us decides in fact to do so, it may have good reason to do so 7. I strongly doubt if a fast takeover transpires. That would 8. If such a technology is a threat, the sooner we are aware of it’s potentiality the better

Reasons to expect soens to be slow in achieving superintelligence (at least decades for superintelligence; more rapid, interesting progress on smaller scale on time scale of year or so)

1. Need new hardware just to determine if these circuit concepts will perform well and scale, at least 10 years for maturity 2. Need device and architecture improvements, theory and experimental capabilities, breakthroughs in understanding how to use such systems 3. Expensive, at least $1B for human-brain scale 4. Progress will come in distinct hardware generations. We can ensure we don’t produce the next iteration until we are ready 5. It will take a movement of historical proportions to realize beyond-human intelligence with soens, there is no risk of stumbling abruptly across the finish line

other concepts to address:

- neuromodulators
- gap junctions

Folks to contact:

- Likharev
- Furber
- Modha
- Srinivasa
- Davies
- Segall
- Tait
- Prucnal
- Harris
- Miller

- Aimone
- Kadin
- Van Duzer

In light of all this complexity, how should we mathematically model these devices?

.0.1 Which model of neuronal dynamics to use?

References

[1] A.M. Turing. On computable numbers, with an application to the entscheidungsproblem. *J. of Math*, 58:230, 1936.

[2] J. von Neumann. A first draft of a report on the edvac. 1945.

[3] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379, 1948.

[4] G. Dyson. *Turing's Cathedral*. Vintage Books.

[5] A. Turing. Lecture to the london mathematical society on 20 february 1947. 1947. https://www.vordenker.de/downloads/turing-vorlesung.pdf.

[6] J. von Neumann. *Problems of Hierarchy and Evolution*. Unknown.

[7] J. von Neumann. *The computer and the brain*. Yale University Press.

[8] W.S. McCulloch and W.H. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115, 1943. http://www.cse.chalmers.se/ coquand/AUTOMATA/mcp.pdf.

[9] A.M. Turing. Computing machinery and intelligence. *Mind*, 59:433, 1950.

[10] J. von Neumann. *The general and logical theory of automata*. Wiley.

[11] C. Mead. Neuromorphic electronic systems. *Proc. IEEE*, 78:1629, 1990.

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

[13] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.

[14] A. Fronczak, P. Fronczak, and J.A. Holyst. Average path length in random networks. *Phys. Rev. E*, 70:056110, 2004.

[15] G. Buzsaki. *Rhythms of the brain*. Oxford University Press, 2006.

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

[17] J.M. Shainline. The largest cognitive systems will be optoelectronic. In *IEEE International Conference on Rebooting Computing*. IEEE, Nov 2018.

[18] P. Dayan and L.F. Abbott. *Theoretical Neuroscience*. The MIT Press, 2001.

[19] G. Fagiolo. Clustering in complex directed networks. *Phys. Rev. E*, 76:026107, 2007.

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

[21] M.D. Humphries and K. Gurney. Network 'small-world-ness': a quantitative method for determining canonical network equivalence. *PLOS One*, 3:2051, 2008.

[22] 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.

[23] H.M. Ozaktas. Information flow and interconnections in computing: extensions and applications of rent's rule. *J. Parallel Distrib. Comput.*, 64:1360, 2004.

[24] S. Strogatz. *Nonlinear dynamics and chaos*. Westview Press, 2015.

[25] R.E. Mirollo and S.H. Strogatz. Synchronization of pulse-coupled biological oscillators. *SIAM J. Appl. Math*, 50:1645, 1990.

[26] D. Somers and N. Kopell. Rapid synchronization through fast threshold modulation. *Biological cybernetics*, 68:393, 1993.

[27] E.D. Lumer, G.M. Edelman, and G. Tononi. Neural dynamics in a model of thalamocortical system. i. layers, loops and the emergence of fast synchronous rhythms. *Cerebral Cortex*, 7:207, 1997.

[28] B. Hutcheon and Y. Yarom. Resonance, oscillation and the intrinsic frequency preferences of neurons. *Trends in Neuroscience*, 23:216, 2000.

[29] Jean-Marc Ginoux and Christophe Letellier. Van der pol and the history of relaxation oscillations: Toward the emergence of a concept. *Chaos*, 22:023120, 2011.

[30] F.L. Vernon Jr. and R.J. Pedersen. Relaxation oscillations in josephson junctions. *J. Appl. Phys.*, 39:2661, 1968.

[31] N. Calander, T. Claeson, and S. Rudner. A subharmonic josephson relaxation oscillator - amplification and locking. *Appl. Phys. Lett.*, 39:504, 1981.

[32] R.R. Llinas. The intrinsic electrophysiological properties of mammalian neurons: insights into central nervous system function. *Science*, 242:1654, 1988.

[33] R.B. Stein, E.R. Gossen, and K.E. Jones. Neuronal variability: noise or part of the signal? *Nature Neuroscience*, 6:389, 2005.

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

[35] S. Thorpe, A. Delorme, and R. Van Rullen. Spike-based strategies for rapid processing. *Neural Networks*, 14:715, 2001.

[36] E. Salinas and T.J. Sejnowski. Correlated neuronal activity and the flow of neural information. *Nature Reviews Neuroscience*, 2:539, 2001.

[37] K.M. Stiefel and T.J. Sejnowski. Mapping function on neuronal morphology. *J. Neurophysiol.*, 98:513, 2007.

[38] T. Branco, B.A. Clark, and M. Hausser. Dendritic discrimination of temporal input sequences in cortical neurons. *Science*, 329:1671, 2010.

[39] J. Hawkins and S. Ahmad. Why neurons have thousands of synapses, a theory of sequence memory in neocortex. *Frontiers in Neural Circuits*, 10:23, 2016.

[40] E.M. Izhikevich. *Dynamical systems in neuroscience*. MIT Press.

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

[42] D.O. Hebb. *Organization of behavior: a neuropsychological theory*. John Wiley and Sons.

[43] G.-Q. Bi and M.-M. Poo. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, 18:10464, 1998.

[44] S. Song, K.D. Miller, and L.F. Abbott. Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience*, 3:919, 2000.

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

[46] C.-W. Shin and S. Kim. Self-organized criticality and scale-free properties in emergent functional neural networks. *Phys. Rev. E*, 74:045101(R), 2006.

[47] S. Fusi and L.F. Abbott. Limits on the memory storage capacity of bounded synapses. *Nature Neuroscience*, 10:485, 2007.

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

[49] J.T. Wixted and E. Ebbesen. On the form of forgetting. *Psychol. Sci.*, 2:409, 1991.

[50] J.T. Wixted and E. Ebbesen. Genuine power curves in forgetting. *Mem. Cognit.*, 25:731, 19978.

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

[52] L.N. Cooper and M.F. Bear. The bcm theory of synapse modification at 30: interaction of theory with experiment. *Nature Reviews Neuroscience*, 13:798, 2012.

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

[54] S. Sardi, R. Vardi, A. Sheinin, A. Goldental, and I. Kanter. New types of experiments reveal that a neuron functions as multiple independent threshold units. *Scientific reports*, 7:18036, 2017.

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

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

[57] 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.

[58] V. Stojanovic, R.J. Ram, M. Popovic, S. Lin, S. Moazeni, M. Wade, C. Sun, L. Alloatti, A. Atabaki, F. Pavanello, N. Mehta, and P. Bhargava. Monolithic silicon-photonic platforms in state-of-the-art cmos soi processes. *Opt. Express*, 26:13106, 2018.

[59] 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.

[60] 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.

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

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

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

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

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

[66] J. Bigelow. *Physical and Physiological Information Processes and Systems*. unknown.

[67] 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.

[68] 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.

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

[70] D.A. Buck. The cryotron–a superconductive computer component. *Proc. IRE*, 44:482, 1956.

[71] https://spectrum.ieee.org/tech-history/heroic-failures/dudley-bucks-forgotten-cryotron-computer. *IEEE Spectrum*.

[72] B.D. Josephson. Possible new effects in superconductive tunneling. *Physics Letters*, 1:251, 1962.

[73] M. Tinkham. *Introduction to Superconductivity*. Dover, second edition, 1996.

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

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

[76] K.K. Likharev. Superconductor digital electronics. *Physica C*, 482:6, 2012.

[77] W. Anacker. Josephson computer technology: an ibm research project. *IBM Journal of research and development*, 24:107, 1980.

[78] 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.

[79] G.E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38, 1965.

[80] P. Bunyk, K. Likharev, and D. Zinoviev. Rsfq technology: physics and devices. *Int. J. High Speed Electron. Sys.*, 11:257, 2001.

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

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

[83] K. Segall, S. Guo, P. Crotty, D. Schult, and M. Miller. Phase-flip bifurcation in a coupled josephson junction neuron system. *Physica B*, 455:71, 2014.

[84] 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.

[85] 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.

[86] 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 *IEEE International Conference on Rebooting Computing*. IEEE, Oct 2016.

[87] 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.

[88] R. Soref and B. Bennett. Electrooptical effects in silicon. *IEEE J. Quant. Elect.*, 23:123, 1987.

[89] G. Reed, G. Mashanovich, F.Y. Gardes, and D.J. Thomson. Silicon optical modulators. *Nat. Photon.*, 4:518, 2010.

[90] L. Liao, D. Samara-Rubio, M. Morse, A. Liu, and D. Hodge. High speed silicon mach-zehnder modulator. *Opt. Express*, 13:3129, 2005.

[91] Q. Xu, S. Manipatruni, B. Schmidt, J. Shakya, and Michal Lipson. 12.5 gbit/s carrier-injection-based silicon micro-ring silicon modulators. *Opt. Express*, 15:430, 2007.

[92] R. Soref. Silicon-based optoelectronics. *Proceedings of the IEEE*, 81:1687, 1993.

[93] 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.

[94] J.M. Shainline and J. Xu. Silicon as an emissive optical medium. *Laser and Photonics Reviews*, 1:334, 2007.

[95] M. Lipson. Guiding, modulating, and emitting light on silicon-challenges and opportunities. *J. Lightwave Technology*, 23:4222, 2005.

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

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

[98] Gordon Davies. The optical properties of luminescence centres in silicon. *Physics Reports*, 176:83–188, 1989.

[99] J.S. Orcutt, B. Moss, C. Sun, J. Leu, M. Georgas, J. Sun M. Weaver J. Shainlineand E. Zgraggen, H. Li, S. Urosevic, M. Popovic, R.J. Ram, and Vladimir StojanoviĂĞ. Open foundry platform for high-performance electronic-photonic integration. *Opt. Express*, 20:12222, 2012.

[100] 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.

[101] H.M. Ozaktas. Paradigms of connectivity for computer circuits and networks. *Optical Engineering*, 31:1563, 1992.

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