

# Hardware for General Intelligence

Jeffrey M. Shainline

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

June 27, 2019

## Abstract

Computers are increasingly being used to aid in information-processing tasks that lack a well-defined mathematical construction and instead require the ability to assimilate information across a broad range of content categories and use that information to make decisions, sometimes with regard to ambiguous criteria. Digital computers were not invented to address these tasks, and attempts to adapt digital hardware for applications requiring general intelligence point to opportunities for significant performance improvements if new hardware can be devised. These opportunities have led to a Cambrian explosion of many types of devices and approaches to beyond-CMOS hardware for artificial intelligence, predominantly in the domain of neural computing. In this article I review the history of both computing and neuroscience in order to establish the context for the present burgeoning of new hardware evolution, and I summarize past and present efforts in neural hardware with particular attention to semiconducting, superconducting, and photonic neural systems. I argue that a hardware platform combining the strengths of light for communication with superconducting electronics for computation has unique promise to embody the principles of neural information in systems of very large scale. Further, I argue that a superconducting optoelectronic network may serve as a central cognitive hub in an intelligent system employing specialized modules based on semiconductors, superconductors, or photonics and performing specialized computations such as numerical analysis, deep learning, or quantum information processing. The analysis indicates the physical possibility to achieve intelligent systems modeled after but potentially far exceeding the capabilities of the human brain.

## Contents

<b>1 Introduction</b>	<b>3</b>	2.2.1.5 Coincidence and Sequence Detection in Active Dendrites . . . . .	14
1.1 Historical Context . . . . .	3	2.2.1.6 Communication in Bursts . . . . .	15
1.1.1 The Origins of Digital Computing . . . . .	4	2.2.1.7 Summary of Coding Strategies . . . . .	16
1.1.2 . . . . .	7	2.2.1.8 Short-Term Synaptic Plasticity . . . . .	16
1.1.3 Waves of Interest in Artificial Intelligence . . . . .	7	2.2.1.9 Subthreshold Membrane Oscillations . . . . .	17
1.1.4 Present Context and Aim of this Article . . . . .	7	2.2.1.10 Long-Term Synaptic Plasticity . . . . .	17
<b>2 Information Processing in Neural Systems</b>	<b>9</b>	2.2.1.11 Metaplasticity . . . . .	18
2.1 Spatial Structure of Neural Systems . . . . .	10	2.2.1.12 Homeostatic Plasticity . . . . .	18
2.1.1 Adjacency matrix . . . . .	10	2.2.1.13 Dendritic Processing . . . . .	18
2.1.2 Node Degree and Degree Distribution . . . . .	10	2.2.1.14 Dendrites and Plasticity . . . . .	19
2.1.3 Path Length . . . . .	10	2.2.1.15 Summary of Neural Device Dynamics . . . . .	19
2.1.4 Clustering . . . . .	10	2.2.2 From devices to populations . . . . .	19
2.1.5 Small-World Networks . . . . .	11	2.2.2.1 Oscillations and synchrony . . . . .	20
2.1.6 Modularity and Hierarchy . . . . .	11	2.2.2.2 The dynamical systems perspective . . . . .	22
2.1.7 Rentian Analysis . . . . .	11	2.2.2.3 Criticality and the fractal use of space and time . . . . .	22
2.1.8 Summary of Spatial Considerations . . . . .	12	2.2.2.4 Neuronal avalanches . . . . .	23
2.2 Temporal Dynamics of Neural Systems . . . . .	12	2.3 Memory in Neural Systems . . . . .	24
2.2.1 Device Dynamics . . . . .	12	2.4 Cognition . . . . .	24
2.2.1.1 Relaxation Oscillators . . . . .	12	2.4.1 Edelman's Model of Cognition . . . . .	25
2.2.1.2 Information Coding by Relaxation Oscillators . . . . .	13	2.4.2 Baars' Model of the Global Neuronal Workspace . . . . .	26
2.2.1.3 Binary communication and post-synaptic response . . . . .	13	2.4.3 The Role of Synchronization in Cognition . . . . .	26
2.2.1.4 The Neuron as a Complex Information Processing System . . . . .	14	2.4.4 Communication Through Coherence . . . . .	27

2.4.5	Experimental Progress Identifying the Mechanisms of Cognition . . . . .	27	3.5	Superconductor electronic neural systems . . . . .	64
2.4.6	Lessons for Design of Hardware for General Intelligence . . . . .	28	3.5.1	Josephson junctions . . . . .	64
2.5	Summary of neural information . . . . .	29	3.5.2	Superconducting digital logic . . . . .	65
<b>3</b>	<b>Candidate Hardware Platforms for General Intelligence</b>	<b>31</b>	3.5.3	Neurons based on Josephson junctions	66
3.1	Overview of Machine Learning . . . . .	31	3.5.4	Strengths and weaknesses of JJ circuits for neural operation . . . . .	68
3.1.1	Thermodynamic Models . . . . .	31	3.6	Superconducting Optoelectronic Systems . . . . .	69
3.1.2	Deep Learning . . . . .	32	3.6.1	Basic Motivations . . . . .	70
3.1.3	Reservoir Computing . . . . .	32	3.6.2	computational circuits . . . . .	70
3.1.4	Spiking Neural Networks . . . . .	33	3.6.3	Communication Circuits . . . . .	70
3.2	General Hardware Considerations . . . . .	33	3.6.4	Large-Scale Systems . . . . .	70
3.3	Semiconductor electronic neural systems . . . . .	35	3.6.5	Misc. SOEN Notes . . . . .	70
3.3.1	Efforts in silicon microelectronic neural systems . . . . .	36	3.6.6	Superconducting optoelectronic neural systems may overcome challenges of digital systems . . . . .	73
3.3.2	The von Neumann bottleneck . . . . .	36	3.7	Comparison of Interconnect Technologies . . . . .	74
3.3.3	Fan-out limitations . . . . .	36	3.7.1	Copper Interconnects . . . . .	74
3.3.4	Shared communication infrastructure	37	3.7.2	Superconducting Interconnects . . . . .	74
3.3.5	Address-event representation . . . . .	37	3.7.2.1	Passive Superconducting Fan Out . . . . .	74
3.3.6	Contention delay in neural systems . . . . .	37	3.7.2.2	Active Superconducting Fan Out . . . . .	76
3.3.7	Summary of challenges of silicon microelectronic neural systems . . . . .	38	3.7.3	Photonic Interconnects . . . . .	76
3.3.8	Actual versus emulated neurons . . . . .	38	3.7.3.1	Photonic Communication in Digital Systems . . . . .	76
3.4	Photonic Neural Systems . . . . .	39	3.7.3.2	Photonic Communication in Neural Systems . . . . .	76
3.4.1	Optical Logic Elements . . . . .	40	3.7.4	Photonic Interconnects with Superconducting Receivers . . . . .	78
3.4.2	Integrated Silicon Photonics . . . . .	41	3.7.5	Summary . . . . .	78
3.4.2.1	Components Required for Integrated Photonics . . . . .	44	3.8	Comparison of Neuronal Devices . . . . .	78
3.4.2.2	Materials for Integrated Photonics . . . . .	44	<b>4</b>	<b>Large-Scale Optoelectronic Systems</b>	<b>78</b>
3.4.3	Free-Space Optical Neural Nets . . . . .	45	4.1	Criteria for assessing cognitive neural hardware . . . . .	79
3.4.4	Machine learning with integrated photonics . . . . .	50	4.2	Achieving Hierarchical Modularity . . . . .	80
3.4.5	Photonic Reservoir Computing . . . . .	52	4.3	Integrated Systems Across Temperature Stages . . . . .	80
3.4.6	Spiking Neurons with Semiconductor Lasers . . . . .	56	4.4	Neuromorphic Supercomputing . . . . .	80
3.4.7	Phase-Change Materials for Synaptic Weighting and Neural Thresholding . . . . .	60	4.5	Optimal Environment for Operation . . . . .	80
3.4.8	Wavelength-Division Multiplexing for Routing and Synaptic Weighting	60	<b>5</b>	<b>Application spaces</b>	<b>80</b>
3.4.9	Synaptic weights in the electronic domain . . . . .	61	5.1	Data Analysis at Particle Colliders . . . . .	80
3.4.10	Materials Considerations for Silicon Photonics . . . . .	62	5.2	Stabilization of Fusion Reactors . . . . .	80
3.4.11	Where are the Light Sources? . . . . .	62	5.3	Telescope Interfaces for Astronomical Observation . . . . .	81
3.4.11.1	Considerations for Digital Communication . . . . .	62	5.4	Theoretical Neuroscience . . . . .	81
3.4.11.2	Silicon Light Sources: the Great Achilles' Heel . . . . .	63	5.5	Simulation of Evolutionary and Developmental Biology . . . . .	81
3.4.11.3	Hybrid Materials Integration	63	5.6	Hybrid Cognitive Systems . . . . .	81
3.4.11.4	Systems with Off-Chip Sources . . . . .	63	5.6.1	Neural-deep learning hybrid systems	81
3.4.11.5	Silicon Light Sources Work at Low Temperature	64	5.6.2	Classical-quantum-neural hybrid systems . . . . .	81
			5.6.3	SOENs as the central cognitive hub	82
			5.7	General Intelligence . . . . .	82

<b>6 Outlook</b>	<b>82</b>
6.1 Achieving Superintelligence . . . . .	83

Need key takeaways at the end of each section

Definitions: For the purposes of this article, I define **Cognition** as the act of considering a subject or situation. A **Cognitive system** is an apparatus capable of engaging in cognition. When **considering** a subject or situation, a cognitive system draws insights from learned experiences and sensory inputs and uses that information to form interpretations of the subject. An **artificial general intelligence** is a cognitive system that can draw insight from diverse information sources and use a broad body of knowledge to form a reasoned and comprehensive analysis of the subject under consideration. An artificial general intelligence must be able to learn and plan.

Cognition Cognitive computing Neural system neuron cmos integrated electronics silicon microelectronics silicon photonics modular hierarchical neural hierarchy digital hierarchy general intelligence AGI soens large scale vlsi

## 1 Introduction

The relationship between the physical substrate of the brain and the information processing occurring therein has long been and remains among the most significant scientific subjects. From a philosophical perspective, we would like to know whether systems other than biological brains devised through natural selection can give rise to intelligence similar to our own. From a computational perspective, we would like to understand the means by which the brain maps complex and evolving stimulus into a coherent context. From a physical perspective, we would like to know which elements and properties of the universe can be combined to enable the computations of cognition. And from a technological perspective, we would like to know if the candidate systems can be feasibly and reliably produced as well as the magnitude of the scientific and economic impact.

To begin, let us define important terms. (define cognition, consciousness, and general intelligence) General intelligence is the ability to assimilate knowledge across content categories and to use that information to form a coherent representation of the world.

In this review I consider the requirements placed on hardware if it is to achieve information processing in the model of and at the scale of the human brain. The last 30 years have brought tremendous gains in both computing and neuroscience, and as a result we are witnessing a flourishing of brain-inspired computing. Here I draw from the

domains of digital computing and device physics as well as the cognitive sciences in an attempt to identify guiding principles to enable the realization of artificial hardware capable of intelligence and broadly useful as a scientific and mainstream technology.

We begin in Sec.1.1 by reviewing the historical developments that led to the present context. Important concepts from neuroscience are summarized in Sec. 2, and we use these principles to inform our design of hardware. Section ?? describes neural systems based on CMOS microelectronics and discusses motivations for pursuing alternative devices and physics for large-scale neural systems. Among alternatives, I focus in this article on photonic technologies (Sec. 3.4) and superconducting electronic technologies (Sec.3.5). The main thesis of this work is that the combination of integrated photonic and superconducting electronic devices will be conducive to the realization of large-scale cognitive systems. Specifically, in Sec. 3.6 I describe superconducting optoelectronic hardware utilizing photonic communication between neurons that compute with superconducting electronic circuits. Superconducting single-photon detectors enable communication with as few as one quantum of the electromagnetic field, and wafer-scale networks of integrated photonic waveguides route photonic pulses from each neuron to its thousands of synaptic circuits. Superconducting circuits enable dissipationless memory, and Josephson junctions provide naturally neuromorphic thresholding and spiking behavior. Scaling considerations are presented in Sec.4, and unique application spaces are considered in Sec.5. Ramifications of such technology are discussed in Sec.6.

### 1.1 Historical Context

As described in the introduction, concepts of cognitive computing lead to the consideration superconducting optoelectronic hardware as a primary candidate for large-scale neural systems capable of general intelligence. Superconducting optoelectronic technology for neural computing resides 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 with digital signals. Concepts from neuroscience dating back to Ramon y Cajal [4], Mountcastle [5], Edelman [6], and many others 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 [1]. Yet the microelectronic march delayed the consequences by many decades. The scaling first charted by Moore [7] left little impetus for revolutionary concepts. Now that scaling is reaching physical limitations [2], 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.

### 1.1.1 The Origins of Digital Computing

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 [3]. Francis Bacon extended these ideas in 1623 to establish that two symbols were sufficient for encoding any communication [8], 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, such as when signaling between ships, although he probably did not anticipate the complex fiber-optic networks that now span the globe. Leibniz further developed ideas related to binary representation [4] with the intention to utilize binary arithmetic for computing [5, 8]. In the 1640s, Pascal devices a mechanical calculator intended to aid in the arithmetic related to taxes, and in 1679, Leibniz proposed a mechanical apparatus for digital computing based on marbles passing through holes to perform logical operations.

Charles Babbage was able to build upon the work of

Pascal, Leibniz, and others to create something more complex. His goal was to create mechanical apparatus to tabulate transcendental functions. The result was the Difference Engine, an early machine that could approximate the solution to differential equations. Babbage intended to go beyond this, and in 1834 he conceived the Analytical Engine, a general purpose computer that could perform different functions based on input instructions. Additionally, it could change its own operations based on the output of its calculations, foreshadowing the Turing machine. The apparatus took the principles of automation that had been so successful during the industrial revolution (1760-1840) and applied them to mathematics. The Analytical Engine was never successfully constructed, but its influence was significant intellectually. Ada Lovelace was the most ardent supporter of this machine, and she anticipated a wealth of applications. She developed algorithms and means of programming the Analytical Engine, and her published notes on the subject constitute a significant early milestone in computing. Among her insights were the realization of the power of a general-purpose computing machine that could perform a wide variety of mathematical calculations but could also work with other types of symbols and non-mathematical constructs. Her work established the foundation of computer programming, and her thinking initiated concepts related to artificial intelligence (AI). In reflecting on whether the Analytical Engine would ever be capable of thought, she concluded that it would not. As she wrote, “The Analytical Engine has no pretensions whatever to *originate* anything. It can do whatever we know how to order it to perform...but it has no power of anticipating any analytical relations or truths.” [9]

It would be nearly 100 years before further developments in computing occurred. In 1936, Turing published his seminal work “On Computable Numbers” [1], wherein he introduced the concept of a universal computing machine. The machine—provided sufficient time and memory—could execute any computation that is possible, in principle, within the framework established by Gödel [2]. 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.” [10]

[11]

[12]

Turing’s work on decrypting messages encoded by the enigma machine during WWII led to advances in comput-

ing hardware, in part by leading to the construction of Colossus, a British machine used for analyzing wartime communications, but also through 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, meteorology, 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.” [13] 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.” [14] 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 [15].

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 [16]. 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.” [17] 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.” [8] 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 [18]. 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 begin-

ning 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 for digital computing 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, von Neumann, and Lovelace 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).

Goodfellow, Bengio, and Courville state, “We simply do not have enough information about the brain to use it as a guide.” On this important point, I respectfully disagree. It is certainly fruitful to continue developing deep learning even without incorporating further principles of neural information processing, but there is still much more we can implement in hardware based on the wealth of insights gained through over a century of neuroscience. We do not have a complete and unambiguous understanding of all aspects of the brain and cognition, but collectively we have learned a great deal about the principles of information processing in the brain. Even the simple architecture of a feed-forward neural net with high connectivity borrowed from the first stages of visual cor-

tex have proven extremely useful for deep learning. Such networks are based on the spatial configuration of certain networks in the brain, and we further know that temporal activity is central to neurosystems. Most deep learning ignores the temporal domain altogether. This is certainly reasonable considering the impressive performance that can be obtained without utilizing the temporal processing strategies of cognitive systems. Nevertheless, the fact that most deep learning to date neglects insights from the brain regarding the time domain means the crucial insight of neural information processing—that space and time are intertwined—has gone underutilized by the artificial intelligence community.

However, it is true that our knowledge of the brain remains incomplete, and we cannot yet build a general-purpose neuromorphic computer based on the operation of the brain that is as trainable and broadly useful for solving important computational problems as deep learning. Computing based on the principles of neural information processing remains a scientific endeavor.

This article reviews principles of neuroscience as well as neuromorphic hardware, beginning with a summary of neuroscientific literature. The summary of neuroscience is an attempt to identify the basic outline of an intelligent computing architecture. I am a device physicist, not trained in the cognitive sciences, so the picture sketched here is undoubtedly incomplete. Still, I argue that broad strokes of the emerging vision of cognition are sufficient to inform our design of artificial hardware capable of general intelligence.

bit of history on neural nets, deep learning, disambiguate machine learning, deep learning, reservoir computing, neuromorphic computing

Since the dominance of silicon began in the 1960s there has been a steady stream of attempts to find another platform to surpass it. This has involved III-V semiconductors as well as superconductors, optical approaches, and more recently graphene and 2D materials. There is often a particular feature of a new material or device that is attractive, and, in the face of challenges, it is often assumed that with sufficient time and financial resources, technical problems can be solved, as they were for silicon. To date, no competing technology has proven capable of outperforming silicon when factors including resource availability, processing and manufacturability, device performance, and cost are considered. In this article I argue that hardware based on silicon and leveraging many of the same strengths as silicon microelectronics, but with devices and circuits significantly different than CMOS, holds potential to outperform the transistor approach specifically when functioning in large cognitive architectures. With this historical context in mind, the burden of significant evidence is on the challenger. Sufficient theoretical and experimental evidence to compel a sea change in the silicon

industry is not presented in this paper. Yet I hope the context and reasoning presented here provide motivation for an interdisciplinary community of researchers to further investigate the confluence of superconducting electronics with integrated photonics for neuromorphic supercomputing.

[19]

### 1.1.2

- neuron doctrine, tissue comprised of cells in 1839, Ramon y Cajal, Golgi, and others in 1890s describe neurons as cells of nervous system
- Mountcastle/Hubel and Wiesel identify columns and slabs of cortical cells as cortical building blocks

### 1.1.3 Waves of Interest in Artificial Intelligence

[20]

#### 1.1.4 Present Context and Aim of this Article

During WWII communication and information technology played many roles not present in prior conflicts. The technological and scientific climate was ideal for new advances in computing, particularly for general-purpose computers that could solve a variety of differential equations and thus contribute to many technical pursuits. Specifically, ... (list pursuits)

At the present moment, the technical areas driving evolution in machine learning and AI are myriad, far more diverse than during the 20th century. These driving influences include: consumerism, online advertising, and personal devices; financial security and financial trading; medical imaging, disease diagnosis, and drug development; data analysis and experimental control in scientific research; and information warfare. Also central to the present context are the unprecedented global transformation induced by silicon microelectronics, the associated availability of computational resources as well as large data sets, and the recent emergence of the Internet, which has led to the generation of large data sets and ensured that much of human activity involves engagement with an interconnected computational system.

This context is driving machine learning and AI to evolve in particular ways. Most importantly for the purposes of the present article, the aforementioned application spaces tend to require a computer that identifies patterns in a specific type of data. The system may seek patterns in a consumer's spend or internet browsing habits, or it may seek certain patterns in medical images or internet videos. The contemporary trend is thus to use machine learning to identify certain patterns in specific data types.

The objective of general intelligence requires more than category-specific pattern recognition or analysis of structured data. A system displaying general intelligence must be able to perceive and contextualize many types of data and learn to identify a large variety of patterns characterized by features in many dimensions of the data by observing and object's traits. For example, consider the associative intelligence involved in a mundane scenario encountered by a human. One may observe an object on the ground and based on its texture and shape know that the object is a leaf, and based on its location under a tree at its red and yellow coloring, it can be assumed the leaf has fallen from the tree. The intelligent system will recognize that this makes sense, as it is late autumn, and trees of this variety lose their leaves this time of year. The intelligent system may even conjure a mental image of a bud emerging in spring, demonstrating the recall of pertinent information to completely contextualize the elements under consideration. A model of the world exists within the intelligent system, and new information is continuously compared to this inner model. In turn, the inner model adapts to new information to maintain appropriate correspondence with the information to which it is exposed in a perpetually changing context.

Toward the goal of general intelligence, we must consider hardware that may depart from that which is adequate for domain-specific machine learning. Neuroscience is a valuable guide to the device and system requirements for hardware capable of general intelligence, while the history of computing and silicon electronic provide crucial insights related to the characteristics necessary in a successful, mature technology. This article reviews the neuroscientific principles that inform hardware properties. Several approaches to hardware for machine learning and neuromorphic computing are summarized, and attention is paid to selecting the physical mechanisms and devices that are best matched to the needs of neural computing for AGI. I describe the reasoning that leads me to the thesis that hardware utilizing few photons for communication and superconducting circuits for computation are uniquely suited to performing the functions required for intelligence. A potential architecture for such a system is outlined, extending from die-scale modules up to many-wafer cognitive systems with fiber-optic white matter connecting the system. Application spaces and potential physical limitations are discussed. This aim is to anticipate the hardware that will be utilized in mature, large-scale neural systems achieving general intelligence, and to explain the reasons why we should expect this hardware to differ from silicon microelectronics that have been unmatched in performance for digital computation. At this early stage, much remains uncertain about the feasibility of such a system, but the ramifications for technology and science are so immense that the subject deserves thorough inquiry.

Why is neural computing difficult? Largely it is because the problems to be solved are not clearly delineated. Whereas the problems digital computers were invented to solve have clear inputs and a well-defined algorithm for computing a solution, the utility of neural systems resides in their ability to deal with disparate inputs and identify salient information even when the objective of the computation is only vaguely specified. "Computational problems that lend themselves to algorithmic solutions share a characteristic property: they are structured, meaning they can be stated clearly and concisely in mathematical terms." [21] "Problems such as pattern recognition in natural environments, however, lack the structure that would allow simple algorithmic solutions." [21] Solution of these types of problems requires learning correlations among the elements of large data sets based on repeated observations of relevant elements of the data. "A computer [ ] cannot draw on this reservoir of common experience; everything must be spelled out for it precisely and unambiguously. ...there is a major component of irregularity that does not fit any simple mathematical or algorithmic model." [21]

After learning many relevant correlations through observation over time, a neural system is able to quickly identify salient features of a stimulus in a given context, not by serially searching through vast memory stores, but through associative memory, wherein "partial features of an object trigger the retrieval of complete information about the object." [21] Such associative memory and recall has played a major role in shaping machine learning, particularly in models such as Hopfield networks [22] and reservoir computers [ ], both of which are inspired by highly connected, recurrent graphs such as that found in the hippocampus, which plays a central role in learning and memory in biological organisms [ ].

As Prucnal and Shastri state, "While neuroscientists infer the mechanisms of processes in the brain, the challenge for engineers is to find out the minimum ensemble of behaviors that are necessary to harness similar processing advantages." The goal of a physical consideration is complimentary to this, intending to answer the questions: What are the physical limits to cognitive hardware? Which devices and physical mechanisms allow us to maximize information integration across space and time? This article considers these questions of physics and information, but from a pragmatic perspective regarding manufacturability. One may imagine a system of arbitrary complexity with devices achieving extreme accuracy in all parameters, but if such a system cannot be realized due to practical limitations, it is not a useful response to the questions above, and we cannot expect it to become a mature technology. The present article is focused on ideas regarding technology incorporating appropriate physical mechanisms for neuronal computation and communication while heeding the insights of very-large-scale integration



regarding scalable manufacturing and device performance margins.

[?] “From a practical point of view, we do not intend to make an artificial brain; instead we are attempting to construct an information processing system that mimics some behaviour of the brain and differs significantly from the von Neumann type computer architecture and algorithm.”

A primary assumption of my perspective is that communication in neural systems is best achieved with telecom photons, starting at the scale of communication between neurons and extending across very large, multi module systems. It stands to reason that the highest system energy efficiency will be gained if communication uses the fewest number of photons possible, limited by noise. This perspective regarding few-photon communication sets the course for many of the decisions made regarding hardware.

A primary assumption of this article is that incorporating the principles of neuroscience into artificial hardware will lead to systems with general intelligence. This is not to say that our understanding of neural systems and cognition is sufficient to straightforwardly construct a thinking machine, but rather that we can use the knowledge from these fields as a starting point. Further interplay between neuroscience and artificial intelligence is likely to lead to constructive feedback for both fields.

This paper regards long-standing scientific questions regarding the brain and cognition: How does thinking arise? Can other physical substrates give rise to thought? What are the physical limits of cognition?

[23]

## 2 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 integrate information from separate calculations in a serial manner, neural information is, by its nature, integrated across a spatial and temporal hierarchy.

Here I review the concepts of neural information processing. One central concept 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. A broad range of information can be simultaneously received and processed due to the manner in which neural systems efficiently move information across space and time. Here I summarize the spatial properties of networks that enable this modality of information processing, and then discuss temporal considerations from the device to systems levels. The final subsection within this section summarizes concepts pertinent to system-wide information processing that enables cognition.

I go into some detail regarding neural systems because these concepts are central to the design of high-performance neural systems. With some exceptions, efforts related to hardware for neuromorphic computing do not incorporate the architectural principles of neural systems, choosing instead to focus on feed-forward graphs for deep learning. Similarly, many efforts ignore the important computations occurring at synapses, treating a synapse as a simple variable attenuator. And in many efforts, the information processing performed in dendrites are neglected all together, replacing the complex computations of the dendritic tree with a point neuron model. All of these simplifications can be justified if one does not aim to create hardware for general intelligence. However, with high performance as the primary focus of this review, we cannot expect to glean the most crucial insights if we neglect to consider the basic principles of neural information processing.

## 2.1 Spatial Structure of Neural Systems

The core principle behind the ability of neural systems to efficiently move information across space is that networks are comprised of specialized modules arranged in a hierarchical configuration. This modular and hierarchical architecture repeats in a self-similar manner from the scale of neurons up to the scale of the system as a whole. To understand these spatial principles quantitatively, we begin by summarizing the relevant concepts from graph theory [24].

### 2.1.1 Adjacency matrix

In order for the neurons to differentiate their responses while efficiently exchanging information across the network, certain structural network considerations are pertinent [25]. The structures of neural systems are analyzed in the framework of network theory [24]. 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, non-symmetric. Connections can also be 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.

### 2.1.2 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  $k_i^{\text{out}} = \sum_j A_{ij}$ , and the in-degree is given by  $k_j^{\text{in}} = \sum_i A_{ij}$ . 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.

### 2.1.3 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 [26], 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 [27]. This quantity is plotted in Fig. ??(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 [27]. 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 [28]. 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.

### 2.1.4 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 [30]. 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 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 interpre-

tation 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 [?, 31] 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.

### 2.1.5 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 [32]. We can introduce the small-world index (SWI) [33] given by  $\text{SWI} = \frac{\bar{C}/\bar{L}}{\bar{C}_r/\bar{L}_r}$ , 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.

### 2.1.6 Modularity and Hierarchy

Anatomically, clusters in the brain were thoroughly described by Mountcastle in 1977 [5] before the concept of a small-world network had been introduced [32]. Mountcastle referred to the communities of neurons as mini-columns and columns. Imaging of biological neural systems with various techniques across all relevant length scales indicates modularity persists across multiple levels of hierarchy [34]. 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 [25, 34]. 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 [35]. 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.

(probably need more here. make sure it is well set up for neuronal avalanches section and fractal use of space and time. define modular and hierarchical)

### 2.1.7 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

$$\varepsilon = \varepsilon_1 n^p, \quad (1)$$

where  $\varepsilon_1$  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 [36].

Depending on the system under study, the Rent exponent may vary, but is generally around 0.75 [36]. We will demonstrate the use of Eq. 1 in Sec. 4. 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 [?, 37], 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, consid-

eration 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 [38]. 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 to 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. 4 we will discuss these large-scale limits in biological and optical systems.

### 2.1.8 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.

## 2.2 Temporal Dynamics of Neural Systems

We emphasize that neural systems make use of space and time in a coordinated manner. We have described the spatial properties of neural systems in the preceding subsection, and here we describe the temporal dynamics. In

the following subsection we describe how neural systems process information across space and time to enable cognition.

Due to the complexity of neural information processing, multiple different perspectives have emerged that emphasize different aspects of observed phenomena. These perspectives are not necessarily mutually exclusive, and a complete understanding of neural systems may require all of these concepts. This situation is analogous to the parable of the blind men and the elephant, in which several individuals attempt to understand the full nature of an elephant with access to only a subset of the relevant physical data. This analogy leads us to identify the neural elephant, depicted in Fig. ???. This analogy illustrates that the challenge of understanding neural information processing at the scale of the human brain is challenging enough that multiple perspectives are required to grapple with the phenomena observed to date. Here I discuss three primary perspectives that have been brought to bear on the problem. These three perspectives are: 1) oscillations and synchrony; 2) the mathematical framework of dynamical systems; and 3) neuronal avalanches and criticality. All three of these perspectives emphasize the interplay between space and time, and in Sec. 2.4 we will argue that they are all related phenomena, but like different parts of the same neural elephant.

Before exploring these three perspectives on dynamical activity in neural systems, it is necessary to review basic behaviors at the device level. In the next subsection we describe neuronal device dynamics, including relaxation oscillations, neuronal communication mechanisms, synaptic plasticity mechanisms, and computations performed by synapses and dendrites.

### 2.2.1 Device Dynamics

A neuron is a dynamical entity. It receives input from many afferent synapses, identifies coincidences and sequences between the activities on multiple synapses, and integrates various inputs 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, usually with an exponential decay in time, and if 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.

**2.2.1.1 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 [28, 39–46] with temporal dynamics on multiple time scales [41] have many attractive properties for neural computing, which is likely why the brain is constructed of such devices [47]. 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 [28]. Relaxation oscillators can operate at many frequencies [?, 43] and engage with myriad dynamical interactions [42]. The oscillator’s response is tunable [43], they are resilient to noise because their signals are effectively digital [48], and they can encode information in their mean oscillation frequency as well as in higher-order timing correlations [49–54].

**2.2.1.2 Information Coding by Relaxation Oscillators** The spiking nature of relaxation oscillators enables them to encode information in several ways. Most simply, they can encode information in their average firing rate, which gives rise to the standard expression used in deep learning and neural networks:

$$y_i = f\left(\sum_{ij} w_{ij} y_j\right), \quad (2)$$

where  $y_i$  represents the “activation” of the  $i$ th neuron,  $w_{ij}$  is the synaptic weight from neuron  $j$  to neuron  $i$ , and  $f(\cdot)$  is a nonlinear function representing the input-rate-to-output-rate transfer function of a simple point neuron. Equation 2 is an extremely simple model of a neuron in the sense that only the average firing rate is considered, so no information regarding the precise times of spikes is retained. Further,  $w_{ij}$  is assumed to be independent of frequency, so the model assumes synapses and dendrites perform no spectral filtering. Also, all input synapses are assumed to terminate on a single integrating body, assumed to be the soma, so no information regarding the spatial location of the synapse on the dendritic tree is retained. Nevertheless, Eq. 2 has proven remarkably useful as a starting point for deep learning.

Beyond the rate-coded, point neuron model, relaxation oscillators can also encode information through several other means. These include the time-to-first-spike after the onset of a stimulus, the phase of a spike relative to sinusoidal background oscillations, and the time interval between the firings of two or more neurons (see Sec. 4.5 in Ref. [55]). Thus, while information coding in digital

computing is simple (bits are transmitted one at a time on a clock), information processing in neural systems with relaxation oscillators as computational primitives is complex. Neurons and networks do not speak a single, simple language, but rather send various types of spike-based messages with different information encoding and decoding in different contexts. We will explore these various strategies for coding in more detail shortly, but first we note one simplifying factor: the spikes that are used to communicate between neurons are binary. No information is conveyed in the amplitude of a single spike.

### 2.2.1.3 Binary communication and post-synaptic response

We expect that any cognitive computing platform will be based on spiking neurons that behave as relaxation oscillators. Communication between these relaxation oscillators 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, so the signal propagating down the axon is not noisy; and the signal that reaches the synapses is very similar from pulse to pulse. Significant variability arises when an action potential meets a synapse, but this relates to the information processing occurring at the synapse (discussed in more detail below), not the nature of the encoding of the signal.

To realize 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 post-synaptic signal amplitude. When the action potential reaches the synapse (pre-synaptic cleft), the action potential may trigger 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 tree, and finally the soma. The action potential arriving at the synapse initiates the synaptic cascade in a binary, all-or-nothing manner (either vesicles are released from the pre-synaptic cleft or they are not), 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 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.

When a synapse receives an action potential, it has a certain probability of releasing vesicles containing neurotransmitters that will then be detected on the receptors of the post-synaptic cleft, generating a current that will propagate a certain distance and decay with a certain time constant. The time-course of the response is a continuous function, specifically an exponential decaying in space and time. The length and time constants depend on: the type of synapse; the morphology of the dendrite to which it is connected; the concentration of various neuromodulators; and the local membrane potential, which is determined by local synaptic activity. Thus, communication between neurons is well modeled by binary events that trigger a post-synaptic current with some probability  $P_s$ , while the post-synaptic current is highly variable and shaped by a number of physiological and dynamical factors. We will further explore the relevance of this post-synaptic response in the coming paragraphs.

**2.2.1.4 The Neuron as a Complex Information Processing System** For multiple reasons described above, we expect relaxation oscillators to be the computational primitives of complex cognitive systems. But amongst relaxation oscillators, neurons are unique in the myriad complexities. Our understanding of the information-processing capabilities of neurons has evolved considerably over time. Early models treated a neuron as a point that passively integrated many inputs and produced a signal upon reaching threshold. This integrate-and-fire model was proposed as early as 1906 by Sherrington in a series of lectures [56], supporting the view championed by Ramón y Cajal [4]. It was that same year that Ramón y Cajal shared the Nobel Prize in Physiology or Medicine with Golgi for their work on the structure of the nervous system. Within this point-neuron view, a neuron simply integrates signals over time, so information regarding the timing or location of synaptic activity is lost. Further, the complex, dendritic tree is assumed to passively transmit synaptic signals to the neuron cell body.

This view of the neuron as a point integrator was highly influential, and remains the dominant model guiding many efforts in deep learning and even neuromorphic hardware. However, further experimental evidence as well as theoretical arguments have deepened our understanding of how neurons work and led to a picture of neural information processing that is far more subtle and powerful.

The modern picture of neural information processing reveals that many operations in addition to integration are being performed, and that these computations are performed at synapses, through the dendritic tree, and in the neuron as a whole. Some of the additional operations now

believed to be performed by neurons include: coincidence and sequence detection; nonlinear thresholding performed in dendrites; temporal filtering of pulse trains; and bursting to overcome noise and enable selective communication between subsets of neurons in an ensemble. We now review each of these concepts and explain their significance for neural information processing.

**2.2.1.5 Coincidence and Sequence Detection in Active Dendrites** As discussed above, after a synapse is triggered and a post-synaptic response is induced, the time-course of the post-synaptic response is a decaying exponential with a time constant that can be shaped by a number of factors. Whether a neuron simply integrates signals or detects coincidences depends on this time constant of the post-synaptic potential relative to the average neuronal interspike interval. If the time constant is on the order of the interspike interval of the neuron, temporal integration is performed during the entire interspike interval, and the neuron is well-modeled as an integrator. If the time constant is short compared to the interspike interval, neurons perform coincidence detection, meaning only spikes that are coincident on the neuron within a short time window relative to the interspike interval can induce the neuron to produce an action potential. If integration is performed exclusively, then the timing of spikes conveys no information. If coincidence detection occurs, information is conveyed in spike timing, neurons respond preferentially to synchronized inputs, and a neuron’s output reflects the input pattern.

Work in the 1990s led König, Engel, and Singer to argue that cortical neurons are better modeled as coincidence detectors than integrators [57]. Their arguments were based both on the physiological evidence that many synapses can have time constants much shorter than the average interspike interval as well as theoretical arguments related to information processing. Regarding the latter, multiple benefits can be identified [51, 58]. First, speed of response can be improved if coincidences can be utilized. Processing speed is determined by the latency between the time of arrival of a signal and the time of a generated response. As stated in Ref. 57, “Because only a small subset of all afferent [post-synaptic potentials] are relevant (namely those which actually coincide and trigger an action potential), the mean time-lag between relevant input and output signals is very short—only a fraction of the interspike interval. Thus, at identical interspike intervals neuronal systems utilizing coincidence detection can process information much faster.” Second, coincidence detection provides benefits with regard to error propagation and noise. Errors due to stochastic processes in the environment are effectively filtered out by coincidence detection, as they would only contribute to neuronal firing rarely when they coincide with true signal, whereas neurons performing integration sum all incoming activity uniformly, regardless of correlations that may serve to isolate signal

from noise. Third, the use of coincidence detection allows neural systems to make use of smaller ensembles of neurons for encoding information, referred to as the size of the ‘grain’ in Ref. 57. If coincidences give rise to larger post-synaptic signals, then synchronized activity of a small group of neurons can have a large effect and drive a neuron to spike. Finally, the main strength of coincidence detection temporal information is not discarded, and synchronized inputs are processed differently than asynchronized inputs on the same set of synapses. The ability to respond preferentially to synchronized inputs has important consequences for information integration and binding, as discussed below in Sec. 2.2.2.1. If neurons can detect coincidences, several important computation become possible.

Work elucidating the behavior of dendrites further supports the view that neurons make use of the timing of synaptic activity as well as other intricacies beyond simple integration [59]. As stated by Koch in 1997, “...dendrites do much more than simply convey synaptic inputs to the cell for linear summation. Indeed, if this were all they did, it is not obvious why dendrites would be needed at all; neurons could be spherical in shape and large enough to accommodate all the synaptic inputs directly onto their cell bodies. ...the function of this elaborate structure cannot simply be to maximize the surface area for synaptic contact.” [60] Instead of acting as passive transmission lines, the active properties of dendrites [61] give rise to the ability to detect coincidences. The basic mechanism relies on the nonlinearity of the response of a dendrite. If two synapses are located in close proximity on a dendrite, the response of the dendrite is a nonlinear function of the activities of the two synapses. Whereas the response of a passive dendrite would be the sum of the activities of the two synapses, the true dendritic response is closer to the product of the activities of the two synapses [60]. Further, dendrites themselves can generate spikes upon reaching threshold [62], a behavior once thought only to occur in the neuron as a whole. Again considering the example of two synapses on a dendrite, if only one of the synapses fires within the decay time of the post-synaptic potential, it may be insufficient to generate a dendritic spike, and the neuron is not informed of the synaptic activity. However, if both synapses fire within the decay time it may be sufficient to induce a dendritic spike, and the information is passed along the dendritic tree toward the soma. By 2006, the physiological mechanisms of dendritic spikes were more clearly understood [62], providing the device-level mechanism to support the coincidence-based information processing described by König, Engel, and Singer in 1996. The nonlinear response of a dendrite with two synapses can also be compared to a logical AND operation: if synapse one AND synapse two fire within a certain time window, an output pulse is produced.

The processing capabilities of dendrites extend beyond detection of coincidences between two synapses. Consider now  $n$  synapses connected to a single dendrite. Now the dendrite will perform a nonlinear function on the inputs

of the  $n$  synapses, leading to an intermediate nonlinear transfer function between the synapses and the neuron as a whole. A neuron comprising many such dendrites will behave as a complex processor comprising multiple independent threshold units, and giving rise to myriad different transfer functions depending upon precisely which afferent synapses are active at a given time, just as observed in recent experiments [63]. Further, due to the morphology of a given dendrite, the post-synaptic current will often flow in a particular direction, usually toward the cell body. Consider the case where synapse one makes contact furthest from the cell body and synapse  $n$  makes contact closest to the cell body. In this case, firing of synapse one will lead to a post-synaptic current that flows toward synapse two, thereby lowering the local membrane potential at synapse two. If a synaptic firing event at synapse one is followed closely by a firing event at synapse two, dendritic current is more likely to propagate on to synapse three, thereby lowering the membrane potential there. The pattern extend to synapse  $n$ , and this nonlinear signal propagation along the dendrite provides a means for the dendrite to detect specific sequences of activity: if synapse one fires just before synapse two, and synapse two is just before three, etc., then the dendrite is far more likely to produce a spike that propagates to the cell body than if the synapses fired in a different order. This phenomenon is illustrated schematically in Fig. ???. The ability of neurons to make use of such sequence detection was developed in 2016 by Hawkins and Ahmad [54]. (say a bit more here) Interestingly, humans appear to be the species with the most elaborate dendritic trees [64].

The total potential generated along a dendrite “depends therefore on the temporal order of the stimulation of the synapses. An input sequence starting at the far end of the dendrite and approaching the soma is more effective in triggering an output spike than the same number of input spikes in reverse order.” ([55], pg. 144.)

I will discuss dendrites further in the context of learning and plasticity shortly. First I review work related to communication between neurons through bursting.

For detection of temporal information in the brain, cite review article from 1993 [65]. You have not read this article, so be careful.

**2.2.1.6 Communication in Bursts** If neurons encoded information only in their average firing rate, we would expect most neurons to fire relatively consistently, with the exact value of the rate dynamically varying depending on stimulus. By contrast, if neurons encoded information only in spike timing and coincidences, we would expect them to fire single action potentials when presented

with relevant stimuli. In practice, these two behaviors are both observed, but a more common mode of activity is the burst. A neuronal burst is a closely spaced sequence of spikes followed by a long quiescent period.

Bursting activity is thought to play a central role in communication between neurons, perhaps as important as rate or other forms of temporal coding. One function of communication via bursts was appreciated by the 1990s, and that is the increased reliability of communicating with bursts rather than single spikes [66]. A single synapse can be unreliable at producing a post-synaptic current in response to the arrival of a single action potential. On the device level, this is because vesicle release by the pre-synaptic cleft does not occur deterministically, but rather with probability  $P_s$ , which varies considerably across synapses, and can be less than 0.1. However, the arrival of a number of action potentials in close succession performs a facilitating function that increases the probability of vesicle release with the arrival of each successive pulse. While the synaptic response to a single action potential can be unreliable, the response to a burst is much more reliable, increasing  $P_s$  above 0.9 simply by using two pulses instead of one in some observations [67, 68], a phenomenon known as paired-pulse facilitation [66].

This reasoning of bursts being more reliable than single spikes is intuitive, but it may lead one to suspect that bursts effectively make communication analog in the sense that the generated post-synaptic current will depend on the number of spikes in the burst. Lisman has argued based on physiological evidence that this is not the case. The nuances of vesicle release and neurotransmitter detection are such that the post-synaptic current is nearly identical when receiving a burst, independent of the number of spikes in the burst (within a workable range) [66], thereby enabling bursting to overcome synaptic failure while maintaining binary communication.

Within this picture of neuronal bursting, the low probability of synaptic response becomes advantageous for filtering noise. As Lisman argues, in some brain regions, single spikes are simply the result of noise, but bursts are rarely noise. Therefore, if a synapse responds only to bursts of two or more pulses, it effectively ignores the noise, and responds only to the signal. For example, place cells in the hippocampus [69] fire preferentially when the organism is in a specific location for which that neuron’s tuning curve is maximized. It has been shown that responding only to bursts can define the place field more accurately than when single spikes are considered [70]. Similarly in visual cortex, it is well-known that neurons can have a tuning curve maximized to detect gratings with a specific spatial frequency or orientation [30]. It has been shown that only the bursting component of a neuron’s activity is tuned to represent these quantities, while isolated spiking activity only correlates with the contrast of the light and dark regions of the grating [71]. These ideas lead Lisman to conclude “that the ability of synapses to not transmit single spikes might be a crucial form of fil-

tering.” [66]

The concept that neurons fire in bursts to overcome unreliable synapses and filter noise is well supported by theoretical and experimental evidence. Yet these may not be the only reasons for communication in bursts. Izhikevich et al. have argued another principle is also in play: “bursts with specific resonant interspike frequencies are more likely to cause a postsynaptic cell to fire than are bursts with higher or lower frequencies.” [72] Within this model, neurons firing in bursts are like radio transmitters, and some synapses are tuned to receive spike trains of the same frequency. This tuning of synaptic and neuronal responses can be achieved through a combination of high- and low-pass filtering as well as subthreshold oscillations of the membrane potential. These device mechanisms for frequency selectivity are discussed below in Secs. 2.2.1.8 and 2.2.1.9.

The most exciting aspect of this use of bursting is that “the same burst could resonate for some synapses or cells and not resonate for others,” thereby providing “effective mechanisms for selective communication between neurons.” Such a method of targeted communication is one of many examples of how neurons can utilize one vast structural network to enable myriad functional networks at different times in different contexts. Izhikevich et al. argue that communication via bursts enables targeted, resonant communication *in addition* to making communication more reliable and filtering noise. Beyond these motivations related to improvements in communication, bursts may also be more effective at modifying synaptic efficacy. I discuss the role of bursting in long-term synaptic modification below in Sec. 2.2.1.10.

**2.2.1.7 Summary of Coding Strategies** We have discussed rate coding, coincidence detection, sequence detection, and bursting. One may ask the question, “Which of these is actually used by the brain?” The answer is all of them. In some contexts, the average firing rate of a neuron may represent the content of a stimulus. In other contexts, the precise timing between two synaptic events or the order in which multiple synapse events occurs carries information that is utilized by the receiving neuron. In still other contexts, bursts are used to increase reliability, filter noise, and the frequency of pulses within a burst can enable selective communication between resonant elements of the network. All of these coding strategies and communication mechanisms contribute to the extraordinary efficiency and adaptability of biological neural systems, and we should aspire to incorporate all of these principles in the design of hardware for general intelligence.

We now summarize the mechanisms at the device level that enable these subtle and sophisticated mechanisms for information processing.

**2.2.1.8 Short-Term Synaptic Plasticity** As described above, the ability of a synapse to respond pref-



erentially to activity at a specific frequency enables additional complexities in communication. At the device level, frequency-selective synaptic response is enabled by filtering mechanisms referred to as 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) [73], 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 [74]. 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.

### 2.2.1.9 Subthreshold Membrane Oscillations

**2.2.1.10 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 [75], a concept that developed in the subsequent decades [76,77] 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) [78] 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 [79], 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 [39,73], 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 ( [28] pg. 329).

Spike-timing-dependent plasticity makes use of correlations between firing activity of neurons to adapt the network into functional clusters [79], store memories in dynamical sequences [54], and strengthen circuits that demonstrate temporal patterns storing sequential memories ( [28] pg. 318). But theoretical analysis finds that if STDP is the only long-term synaptic plasticity mechanism, memories are forgotten very quickly [80]. Experimental analysis finds that synapses have multiple additional means to change how synapses adapt over time and activity to help retain memories [81]. These plasticity mechanisms are referred to collectively as metaplasticity.

It has been shown that a burst of action potentials

comprising four pulses can be sufficient to induce long-term potentiation and depression [82], dependent on the timing of the burst relative to theta oscillations. A burst arriving at the peak of theta leads to potentiation, while a burst arriving at the trough leads to depression. In this case, we find an interplay between neuronal bursting, synchronized oscillations, and synaptic plasticity. Adapting based on the confluence of all this information leads to efficient evolution toward a network structure that generates constructive activity.

**2.2.1.11 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 [81]. 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, [83, 84], yet Fusi and Abbott have shown that memories are lost more rapidly than this if plastic synapses are presented with continual stimulus [80]. They have proposed a model that achieves the observed power-law forgetting by introducing internal complexity to the synapses [85]. 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 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.”

**2.2.1.12 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 [86] 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 [87]) 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.

**2.2.1.13 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, non-linear transfer functions performed by dendrites between individual synapses and the neuron as a whole [59]. The dendritic arbor is a complex, branching structure on which most of a neuron’s input synapses make their connections. While the dendritic tree was thought to be a passive integrating structure for quite some time, by the mid 1990s many active dendritic responses were beginning to be understood [61]. The dendritic branches that comprise the arbor have passive and active properties that allow them to perform various computations. Dendrites modulate postsynaptic potentials on their way to the soma as well as generate spike activity [62, 88]. One picture of the dendritic arbor is a network of multiple independent threshold units that integrate signals and produce dendritic spikes upon reaching threshold. This picture was introduced on theoretical grounds based on the enhanced storage capacity of such a dendritic tree in 2001 [89], and supporting experimental evidence continues to accrue [63].

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.

Like many aspects of neural information processing, the myriad roles of dendrites remind us that the devices giving rise to cognition are diverse and multifunctional. Mel, Schiller, and Poirazi emphasize that “dendrites of different neuron types contribute to the cell’s input-output function in markedly different ways.” [90] While all of the aforementioned roles of dendrites are observed in biological neural systems, they are not necessarily all present in every dendrite. In the context of designing hardware for neural information processing, these functions must be available, and sophisticated analysis at the level of networks and systems must inform our decisions of when to include each mechanism.

[89]

**2.2.1.14 Dendrites and Plasticity** In addition to performing nonlinear transfer functions on inputs, dendrites also play a key role in synaptic update and learning. This activity is induced by back-propagating action potentials created during the firing of the neuron [88]. These back-propagating action potentials provide a feedback signal from the neuron to the dendrites and synapses after neuronal firing that performs one of the steps in a timing-dependent plasticity process [91].

In addition to back-propagating action potentials playing a role in STDP, dendritic spikes are also involved in forms of synaptic plasticity that do not require synaptic activity. In these forms of synaptic plasticity, it is the timing of the forward-propagating dendritic spikes and backward-propagating spikes that establish the conditions in which synaptic weights can be modified [92]. This local activity can lead to long-term synaptic potentiation or depression and may be the primary mechanism for rapidly acquiring memories [62].

We have mentioned that dendrites are capable of acting as integrators or coincidence detectors depending on the time constants involved. Additionally, a single dendrite can transition between these two roles based on the interplay between synaptic activity and the neuromodulatory environment [62]. The presence of certain neuromodulators affects the behavior of voltage-gated ion channels, thereby shaping the short-term response of dendrites [88]. As stated by Holthoff et al., “...the alliance of non-linear and linear integration modes combines the advantages of the excellent signal-to-noise ratio of digital processing with the speed and complexity of analog processing.” [62]

We have also mentioned that back-propagating action potentials in dendrites play a central role in STDP. The active properties of dendrites can lead to metaplasticity as well [81]. The mechanisms involved again relate to the interactions between synaptic inputs and neuromodulators. The interplay between neuromodulators and synaptic activity affects the time constants that shape the dendritic membrane potential as well as the probability that potentiation or depression will occur in the presence of synaptic input.

This discussion of dendritic processing is necessarily brief, emphasizing the active role played by dendrites in adaptive and frequency-selective response to spike trains based on voltage-dependent conductances; the nonlinear thresholding properties of dendrites exemplified by the generation of dendritic spikes; and the role of the dendritic tree and back-propagating action potentials in synaptic plasticity. Further detail can be found in the comprehensive review by Sjöström et al. [93] From a computational perspective, the dendritic tree is the main information-processing infrastructure of a neuron. Computations are performed as fan-in occurs. Whereas in many computational systems, fan-in is passive and represents part of the communication infrastructure, in neural systems fan-in and computation occur in the same physical structures. This is in contrast with the axonal arbor, which performs fan-out to many synapses, but the axons themselves do not modify information, but rather simply serve to communicate spikes from one neuron to many others.

#### 2.2.1.15 Summary of Neural Device Dynamics

This subsection summarized literature supporting the perspective that relaxation oscillators are uniquely suited to serve as computational primitives in cognitive systems due to the high signal-to-noise ratio of their outputs, their adaptive frequency response, and their ability to make use of many information coding schemes. These coding schemes have been described. The importance of binary communication for mitigating noise was discussed, and multiple communication schemes were explained. These modes of communication include: average rate; coincidence and sequence; phase relative to background oscillations; and bursting for reliability as well as frequency selectivity. The device-level mechanisms that enable these functions were summarized, including short- and long-term synaptic plasticity and active dendritic functions. We now turn our attention to activity emerging from populations of neurons.

### 2.2.2 From devices to populations

We have described some of the basic device operations of neurons and their components, but neural information processing leverages entire populations of neurons with emergent mechanisms for information processing occurring in a network context. As a first example, consider a population of uncorrelated, non-interacting neurons re-

ceiving identical stimulus and firing asynchronously. The population activity is defined as the sum over all firings by all neurons, and in the presence of a constant stimulus, the population will fire with a fixed population activity. For a large population of  $N \rightarrow \infty$  neurons, at a given moment there will be some neurons at all stages of the cycle of integration and firing. If the stimulus is instantaneously increased by a finite amount (perhaps due to the input from another neuron firing), the population activity will instantaneously increase. This instantaneous response is possible because some neurons in the population will be close to threshold, and the finite perturbation pushes them over threshold. Thus, they fire sooner than they would have in the absence of the perturbation, and the population activity increases. Such a response is not obtained, in general, from a single neuron. If a single neuron is perturbed instantaneously, its response will depend on its state before the perturbation. If it is in a refractory period or far from threshold, the perturbation may be insufficient to evoke a response at all, or it may take some time before the effect of the perturbation is observed.

This rapid population response is most clearly observable in the absence of noise, and in the presence of noise, populations perform useful functions as well. For example, one can calculate the signal-to-noise (SNR) ratio of a group of neurons attempting to code a given signal. It is found the SNR increases linearly with the number of neurons in the population ([55], chapter 7.3.2). Populations of neurons can increase response time and mitigate noise. These two responses do not require interactions between the neurons in the population, yet dynamics become more interesting when the neurons of a population are connected with various graph structures.

**2.2.2.1 Oscillations and synchrony** When observing brain activity at a coarse scale, such as with electroencephalograms, sinusoidal oscillations in activity are observed. Finer measurements with electrodes also observe this phenomenon on shorter length scales. Oscillations in mammalian cortical neurons are observed across five orders of magnitude in frequency, from 0.05 Hz to 500 Hz [94], and the types of oscillations observed depend on behavior and input stimulus. In rats, 10 frequency bands are observed, and their center frequencies are separated logarithmically, creating what Buzsáki refers to as “a system of brain oscillators”. These different bands are associated with different brain states and different behaviors. At a given moment, activity may be present in several frequency bands, and these activities may compete or interact with each other. This seemingly simple behavior leads to many layers of complexity with important ramifications for neural system operation [28]. Each neuron is a relaxation oscillator, characterized by spikes that are short in time relative to the inter-spike interval. A single neuron is not well modeled by a sinusoidal time dependence, but the net activity of a population of neurons of-

ten takes a sinusoidal form. In neural systems, the devices are relaxation oscillators, while the networks can behave as harmonic oscillators.

When referring to oscillations in a network of neurons, we are referring to this sinusoidal population activity. The net activity has a well defined phase, and it becomes meaningful to speak of the phase of firing of a single neuron relative to the phase of the sinusoidal population activity. In some cases information is encoded in the timing of a neuron pulse relative to the phase of the background oscillations ([55], pg. 140).

A related but distinct phenomenon is synchrony. In populations of neurons with recurrent connections, asynchronous firing may or may not be stable depending on factors such as noise, signal transmission time, and rise time of the post-synaptic potential ([55], Ch.8). In the low-noise limit, asynchronous firing is unstable across a wide range of parameter space, and the network will invariably converge toward synchronized firing of action potentials. For neurons sufficiently close in space that communication delays can be neglected, the frequency of synchronized activity is determined by device-level properties, such as membrane time constants. For large groups of neurons sufficiently separated in space so that transmission delays are non-negligible, neurons will naturally separate into clusters, each synchronized internally, but unsynchronized globally. Thus, the net population activity of these cluster states can be at frequencies higher than any one cluster can fire alone. These so-called cluster states have no relation to the identically named quantum states being investigated for their role in quantum computing. Synchrony is quantified through a cross-correlation function of activity taken from different neurons or brain regions.

A neuron is sometimes referred to as a single-cell oscillator, referring to the fact that it may be observed to pulse at a well-defined frequency. This oscillation may be self-induced (autorhythmic) based on the interplay of  $\text{Na}^+$ ,  $\text{K}^+$ , and  $\text{Ca}^{2+}$  following an action potential [47]. Oscillation may also be observed due to a driving super-threshold stimulus based on the refractory period of the neuron. A neuron may also be referred to as a single-cell resonator, meaning it responds preferentially to input stimulus within a given frequency range [47]. To achieve a resonant response, the neuron must have both a high-pass response and a low-pass response to generate a band-pass filter [43], as discussed above in the context of short-term plasticity. Short-term plasticity leads to frequency-dependent synaptic efficacy [95]. Low-pass response is achieved by parallel leak conductance and capacitance of the membrane that shunts responses at high frequencies. Similarly, there are voltage-gated conduction pathways in the membrane that act as high-pass filters. Nature has engineered the time constants of these responses so that band pass is possible. Further, the low-pass and high-pass cutoff frequencies can be adjusted based on the membrane voltage or the presence of neuromodulators [43, 96]. Voltage tuning of the frequency response of a given synapse

may occur based on the activity of nearby synapses, while neuromodulatory adjustments occur over broader areas, usually containing multiple neurons. Resonance in neurons is quantified by measuring the impedance of a single neuron as the driving stimulus is swept across frequencies. It can also be modeled theoretically at the single-neuron level ([55], pg. 81) and at the population level ([55], pg. 233). A key result of this line of research is that the frequencies of various oscillations in the brain depend on the structure of the network, but they also depend on the intrinsic response properties of the devices, which dynamically adjust based on network activity [88,96]. Synapses and neurons are tuned to respond maximally at frequencies that facilitate communication on different spatial and temporal scales. The interplay between network and device contributions to oscillatory activity may be important. As stated by Hutcheon and Yarom in 2000, "...network connectivity could reinforce the patterns of excitation produced by coupled oscillators."

It has been proposed that the interplay of rhythmic and arrhythmic brain activity is important to cognition. In this model, background oscillations of the network at various frequencies provide the context in which the activities of transient clusters occurs [?,47]. It has been observed that states of oscillation are disrupted upon directing attention to a sudden stimulus [28], leading to the hypothesis that various network oscillations play a role analogous to the information processing context emphasized by Baars [97]. Oscillations are also thought to be essential to brain-wide information integration. In this context, the thalamus plays a central role. The oscillation frequency of thalamic neurons is tunable based on activity and neuromodulators, and depending on the mode of oscillation, the thalamus may be coordinating the exchange of information between different brain regions at theta frequencies, or it may be inducing sleep through alpha rhythms. Such observations led Baars to identify the thalamo-cortical complex as the anatomical construct responsible for brain-wide information dissemination in the Global Workspace model [97]. It is clear that distinct oscillations play distinct roles in the information processing of the brain. "The low-frequency resonances in the cortex and thalamus appear suited to support the thalamocortical delta-wave oscillations that are particularly prominent during deep sleep. The higher-frequency oscillatory behavior and underlying resonance in pyramidal and inhibitory neurons of the neocortex might have some involvement with higher frequency rhythms that appear in the cortex during cognition." [43]

"...establishment of even a weak resonance makes a neuron a good listener for activity within a specialized frequency band. A host of good listeners, mutually connected, should tune networks to operate in frequency ranges of special biological meaning." [43]

Definition of synchrony from [98] box 2

"Functionally, such resonances constrain neurons to respond most powerfully to inputs at biologically important frequencies such as those associated with brain rhythms." [43]

"Brain rhythms reflect basic modes of dynamical organization in the brain." [43]

Brain oscillations perform numerous important functions

role of inhibition

If all neurons in a network have the same refractory period, and all synapses have the same rise and decay time constants, then signal transmission delay due to the spatial organization of the network is the primary time constant that separates neurons into differentiated clusters. The maximum frequency at which two spatially separated regions can drive each other to synchronize depends on the distance and transmission velocity. This delay limits the scale of networks that can oscillate at a given frequency, constraining high-frequency activity to smaller regions of space. We consider this concept further in Sec. 4.

Neurons in close proximity can drive each other to fire at high frequencies, and such clusters can exchange information across broader regions of the network through slower oscillations, allowing "brain operations to be carried out simultaneously at multiple temporal and spatial scales." [94] This conception of how global activity can modulate and steer local activity is central information processing in the brain.

must demonstrate resonance in loop neurons in this paper

The net dynamical activity of brain oscillations, synchronized clusters, and neuronal avalanches leads to a  $1/f$  power spectral density, which implies that activity occurring at slow frequencies can cause a cascade of activity at higher frequencies [94]. Put another way, slow, sinusoidal activity across a wide region of the network can modulate

the fast activity occurring between clusters locally.

role of oscillations in plasticity [28, 66, 94]

**2.2.2.2 The dynamical systems perspective** We have discussed various forms of neuronal dynamics in neural systems, from the scale of single neurons up to the scale of avalanches and oscillations spanning the network. One prevalent theme is that all activity is transient. Small clusters of neurons form transient synchronized ensembles. Theta oscillations may sustain for some duration until the periodic activity is broken as attention is directed to a stimulus. It is common to think of short-term memories as being represented by dynamical activity stored in a stable attractor, such as in long short-term memory [?]. However, activity in the brain is never truly cyclic as in a stable attractor, but rather transitions between quasi-stable states. These successions of states are trajectories in a dynamical state space, and a given stimulus can induce a specific trajectory associated with the activities of a population of neurons. Each trajectory is a sequence of successive metastable states [95], where a metastable state is a saddle point in state space, as opposed to an attractor which is a local minimum. This picture of neuronal dynamics treats transient activity mathematically as a stable heteroclinic channel [99, 100]. As stated by Rabinovich et al., “These saddles can be pictured as successive and temporary winners in a nonending competitive game.” As discussed in multiple contexts thus far, a balance of excitation and inhibition is necessary to result in long-lived transient sequences that do not grow without bound. Unique sequences are triggered by unique inputs, leading to computations that are “reproducible, robust against noise, and easily decoded.” [95]

Reservoir computing learned/adapted patterns

through plasticity, or reservoir without plasticity information stored in trajectories through phase space

that involve sequences of neuronal firings; the dendritic tree can detect these sequences; a specific dendrite may fire preferentially only when a specific dynamic trajectory is excited by stimulus

[73]

<sup>1</sup>For example, if a system has a power spectral density  $p(f) \sim f^{-\gamma}$ , then the expectation value of frequency is  $\langle p(f) \rangle \sim \int_0^\infty f p(f) df = \int_0^\infty f^{-(\gamma-1)} df \rightarrow \infty$ , indicating that an average frequency cannot be defined. For real physical systems with a maximum operating frequency, this integral will converge. Nevertheless, the power-law distribution results from contributions across a broad range of scales, and it is generally not meaningful to speak of a characteristic frequency or length.

**2.2.2.3 Criticality and the fractal use of space and time** The concept of self-organized criticality was introduced in 1987 by Bak, Tang, and Wiesenfeld [101–103]. This work argued generally that dynamical systems with spatial degrees of freedom naturally evolved to a critical point with a highly ordered phase on one side and a highly ordered phase on the other side. Similar to general second-order phase transitions, this critical state is characterized by power-law decay of correlations across space and time. However, Bak et al. emphasize that self-organized criticality differs from phase transitions in equilibrium statistical mechanics. While phase transitions in statistical mechanics result from modifying a tuning parameter across a critical value, self-organized criticality emerges in dynamical systems far from equilibrium as the dynamical state converges to an attractor. Dynamical systems at the critical point are marked by  $1/f$  power spectrum (sometimes referred to as flicker noise or crackling noise [104] even if the source is not noise at all), indicating correlations over a wide range of temporal scales.

A  $1/f$  power spectrum is characteristic of a fractal process, with self-similar activity observable across time scales [37]. Critical processes are often also marked by self-similar patterns in space, with long-range spatial correlations also identified by power-law decay, as discussed above in the context of Rentian scaling. Power laws have no characteristic scale<sup>1</sup> and are intimately related to fractals and indicate a balance between order and chaos. If correlations decay exponentially (in space or time) long-range interactions are impossible, and the system has simple dynamics. If correlations are constant across the system, a small change anywhere perturbs the system everywhere, and chaos results. Power-law distributed correlations mark the fruitful middle ground between these two extremes, giving rise to systems with complexity marked by fractal patterns of self-similarity across spatial and temporal scales.

This concept of balance between order and chaos was explored in the context of computing by Langton in 1990 [105], where he used the term “edge of chaos” to refer to the region of parameter space most capable of information processing. Langton argued that computation requires transmission, storage, and modification of information. Information storage reduces the entropy of a system, while transmission increases the entropy. Because a computer must do both effectively, optimal performance is achieved by operating at a trade-off point between high and low entropy, which occurs in the vicinity of a phase transition. Making the analogy to solid/liquid transitions, solids are highly ordered, so information can be stored, but it cannot be communicated. Conversely, liquids are highly disordered, and information can propagate through a liq-

uid, but the transmission process disrupts the state, and prior information is lost. Computing occurs on the boundary between the ordered solid phase and the disordered liquid phase, and at this operating point information can be transferred over long distances without attenuating. Langton argues that computation and phase transitions are fundamentally connected, an idea that has resonated significantly in the neuroscience community. We now review some of the literature related to criticality in neural systems.

Also explore 106 ch. 12 in this section

**2.2.2.4 Neuronal avalanches** Evidence of neural system operation at a critical point can be obtained through temporal or spatial considerations. In the spatial domain, operation at the edge of chaos is observed in the distribution of sizes of sequences of activity. In a wide variety of slowly driven physical systems, the response is marked by discrete events of a variety of sizes [104] (earthquakes, magnets in magnetic fields). The elements these systems have in common is that they are “...driven systems with many degrees of freedom, which respond to the driving in a series of discrete avalanches spanning a broad range of scales...” [104] Neural systems meet these criteria and display this behavior.

If a population of spiking neurons is mutually interconnected, rich dynamics can result and be shaped by a multitude of factors. Consider a network of  $N$  neurons with sparse, random connections, approximating cortex. In the presence of random input, perhaps from other regions, the neurons of the population will be observed to fire, seemingly stochastically. We can ask, if the population is observed for a duration  $\Delta t$ , what is the probability that  $s$  neurons<sup>2</sup> will fire during that duration? The result is a power law,

$$p(s) \propto s^{-\alpha}, \quad (3)$$

with  $\alpha \approx 1.5$  [107, 108]. This relationship is not sensitive to the observation duration,  $\Delta t$ , and is observed in species from leeches [108] to monkeys [109]. Equation 3 is supported by experimental data observing the number of electrodes activated during an observation period, and therefore regards the spatial organization of neuronal activity. An event involving the firing of  $s$  neurons is referred to as a neuronal avalanche [?] of size  $s$ , and Eq. 3 tells us that the ratio of the number of avalanches of size  $ks$  to the number of size  $s$  is  $k^{-\alpha}$ , independent of  $s$ . There is activity across all spatial scales. Local activity of a few neurons is most probable, but large avalanches including wide regions of the network are still statistically significant. Neuronal avalanches of power-law size distribution

are another manifestation of fractal scaling in neural systems: “the relationship of pattern sizes to each other is apparent of every scale.” [107]

This fractal organization of activity has important ramifications for system scaling, and therefore should be considered when designing hardware for cognition. Because the interactions of the neurons and the graph structure are such that activity patterns of all sizes can occur, the limit of neuronal avalanche size is not set by the devices, interactions, or dynamics, but rather by the physical structure. As stated by Plenz and Thiagarajan, “...in a finite system such as the cortex, a power law [] indicates that the underlying dynamics are constrained by the size or physical borders of the system rather than by any intrinsic characteristic of the dynamics themselves.” [107] The implication of this neuronal activity is that any given neuron or cluster can quickly engage any other region of the network and access information stored at those other sites. These patterns extend beyond single cortical columns, and the engagement of even distant regions can be fast [107], provided the hardware employed does not introduce communication delays.

It is important that choices we make about hardware ensure this statement remains true of our artificial system. If our neurons can only achieve local connectivity, then network path lengths we become long, and the inability for a given neuron to activate distant neurons within will reduce the probability of large neuronal avalanches. Similarly, if a shared communication infrastructure is employed, and simultaneous requests to send spike events must be arbitrated, avalanches with a large number of neurons will not be achievable within a short temporal window, introducing an undesirable size/speed trade-off.

An important factor shaping neuronal avalanches is the graph structure of the network. Not all graph structures lead to power-law distribution of neuronal avalanche size. In particular, a hierarchical organization of the network is sufficient to lead to power-law distribution of avalanche size as well as duration [110]. As described in Sec. 2.1.6, a hierarchical network is comprised of multiple layers of nested modules. By contrast, a network described by a random graph which will have avalanche sizes that are bimodal, meaning there will be some probability  $p$  that a single firing neuron will trigger a large avalanche spanning most of the connected network, and another probability  $1 - p$  that the avalanche will involve a negligible fraction of the nodes [110].

A network with hierarchical modularity is conducive to generating neuronal avalanches of power-law-distributed sizes, and this is related to the fact that such networks can efficiently communicate signals across large regions of the network. Similarly, small-world networks are conducive to generating power-law-distributed neuronal avalanches, and small-world networks may or may not have a hierar-

<sup>2</sup>Experiments with biological neurons observe local field potentials with spatial resolution more coarse than a single neuron, so the size  $s$  is often related to the number of electrodes in an array recording signal during  $\Delta t$ , but here  $s$  is taken to represent a well-defined number of neurons.

chical architecture. However, the pattern of making primarily local connections but also important long range connections enables both neuronal avalanches of many sizes. Importantly, networks with initially random connections can self-organize into a small-world network through the adaptation of synaptic weights through spike-timing-dependent plasticity [79]. Even a network initiated with all-to-all connectivity can organize into a functional network with significantly less connectivity based on STDP, enabling a densely connected, random network—such as that of a newborn—to adapt into a network demonstrating power-law neuronal avalanches capable of efficient information transfer across all spatial scales of the network. It is important to note that a balance between inhibition and excitation is necessary to enable such avalanche behavior, but only the excitatory synapses need to be plastic to arrive at a scale-free, small-world network.

In addition to sculpting an initially random network into a small-world network, Ref. 79 found the resulting degree distribution followed a power law:

$$p(k) \sim k^{-\beta}. \quad (4)$$

The interpretation of Eq. 4 is that if a node from the network is selected at random, the probability it will have degree between  $k$  and  $k + \Delta k$  is proportional to  $k^{-\beta}$ . In the computational study of Ref. 79, this degree distribution interprets individual neurons as nodes, but experimental data taken with functional magnetic resonance imaging treating larger volumes of brain matter as nodes also found a power-law degree distribution [111]. In this study, the spatial resolution was limited by the imaging technique, and in general it is quite difficult to assess the degree distribution of actual brain networks in large animals. This information may result from the detailed study of the connectome at the microscopic scale [?].

We have been discussing neuronal avalanches primarily in the spatial domain, and examples from the literature indicate that modular, hierarchical, networks with small-world graph structures lead to a power-law distribution of sizes of avalanches. These structural characteristics are not independent from temporal behavior, and we will find that the temporal distribution of neuronal avalanches leads to consideration of self-organized criticality. But first we describe simpler temporal phenomena, namely oscillations and synchrony.

The fractal distribution of neuronal avalanche sizes has ramifications for information processing and temporal responsiveness. Many systems throughout nature obey power-laws taking the form of Eq. 3 [?, ?], and it often marks behavior at a critical point near a phase transition. Thus, observation of neuronal avalanches obeying power-law size distribution leads many to think the networks of the brain are balanced at a critical point with uninformative, synchronized order on one side, and chaotic disorder on the other [108]. This balanced point is referred to as the critical state or criticality, and is related to the general theory

of self-organized criticality as described by Bak, Tang, and Wiesenfeld [?, 101, 102]. Beggs argued [?] that criticality optimizes multiple aspects on information processing in neural systems. Information transmission is optimized across spatial and temporal scales, as neurons are each able to participate in and initiate local as well as global activity. Information storage of short-term memories is increased, as criticality maximizes the number of stable dynamical attractor states supported by the network. It has also been argued that By operating in the critical state, the dynamic range of the system is increased [112, 113], the response time is reduced [?].

In addition to showing scale invariance across space, dynamical systems operating at the critical point show a spatial invariance in time and frequency as well. This can be quantified by two metrics: the phase-lock

- 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.

Mention how deep learning neglects essentially all of the important aspects of neuronal information processing discussed here

## 2.3 Memory in Neural Systems

Brief discussion of the various forms of memory in neural systems.

- short-term synaptic/dendritic adaptation
- associative memory in dynamical states
- long-term synaptic and dendritic modification
- growth and decay of neurons

## 2.4 Cognition

Vernon Mountcastle's work on the columnar organization of the cortex was referenced in Sec. 2.1.6. This work was presented In June of 1977 at the Neurosciences Research Program two-week meeting for intensive study of recent developments in neuroscience. Mountcastle presented his work in a keynote address in which he described his latest understanding of the physical structures comprising the cortex (cortical columns) and the manner in which they process and communicate information [5]. On the last afternoon of the workshop, Gerald Edelman put forth a theory of higher brain function based on processing performed by groups of neurons and a hierarchical model of feed-forward and feed-back communication [6]. The theory put forth by Edelman was consistent with and supported by the anatomical evidence presented by Mountcastle, and the resulting framework established a new foundation on which many subsequent ideas have been based.



The lectures by Mountcastle and Edelman are now published in a single volume [114] that represents the beginning of the modern phase of theories of cognition.

#### 2.4.1 Edelman's Model of Cognition

In Edelman's model, neuronal groups containing 50 to 10,000 neurons serve to recognize a specific set of signals received either through sensory stimulus or from other groups of neurons. Edelman proposed that patterns in sensory input are first identified by a primary set of neuronal groups, and the set of all input patterns that can be recognized and induce a unique response constitute the primary repertoire. At the next level of information processing hierarchy, the secondary repertoire recognizes patterns in the responses of the groups in the primary hierarchy. Stimulus selects certain neuronal groups to become active, and over time, the adaptive nature of neurons and synapses makes certain groups much more likely to be excited by a specific stimulus. Edelman proposed that these neuronal groups are cyclically polled to determine which groups were most active and therefore which stimulus was most likely present. Further, this cyclic polling procedure involves both feeding forward the activities of neurons excited by the stimulus (corresponding to the primary repertoire), but also feeding back activities of higher-order groups that have sampled the primary repertoire and formed associations at the level of the secondary repertoire. Edelman argues that "...the conscious state results from phasic reentrant signaling occurring in parallel processes that involve associations between stored patterns and current sensory or internal input." The word "phasic" refers to the fact that this polling of the neuronal groups occurs cyclically, and the word "reentrant" refers to the fact that information from higher cognitive centers is fed back to the primary modules so the contextualized information accessed in the secondary repertoire can be compared to the information being fed forward from the primary repertoire. Through these cycles, higher-order processing centers can identify "multimodal associative patterns", place incoming information in the complex context learned through prior experience, and quickly identify changes in inputs.

Edelman's construction incorporates several of the core concepts I have described here. First, central to the hypothesis is the idea that small groups of neurons form specialized processors that identify a specific subset of features of input stimulus. As explained by Edelman, "...each group has a limited set of characteristic spatiotemporal response patterns of firing as well as a characteristic set of connections to other groups." The mini-columns and columns identified by Mountcastle are clear candidates to form these basic, modular structures. We have emphasized the concept that transient neuronal ensembles form temporary coalitions to represent specific stimuli, and this idea is rooted in Edelman's theory. A single mini-column can represent a number of different stimuli, because the

neurons therein can dynamically adapt their activity to form a variety of transient ensembles depending on the input. Each unique transient ensemble must produce a unique output so it can be identified as a unique pattern by other mini-columns in subsequent stages of processing. Related to this concept is the fact that the cortex comprises a very large number (nearly a billion) mini-columns. This allows the brain to recognize and correlate an enormous number of unique patterns across a multiple sensory modalities.

The second core concept emphasized by Edelman is the role of oscillations in the process of cognition. Edelman identified theta oscillations as a promising candidate for the cycles of activity during which the specialized processors (neuronal groups) are polled. On time scales much faster than theta oscillations, local groups of neurons interact and develop activities in response to stimulus. Then, with each cycle of theta activity, the information from many such groups is sampled, and higher-order patterns are identified, corresponding to identification of patterns at higher levels of abstraction.

Third, the manner in which Edelman described the activity of neuronal groups during successive periods of theta oscillations contained the seeds of modern descriptions of neuronal activity in the framework of dynamical systems. In particular, Edelman described "a continuous shifting pattern of associations" occurring through the interplay between primary neuronal groups and the reentrant feedback from higher-order stages of processing. This "continuous shifting pattern" is highly reminiscent of the "trajectories moving along heteroclinic orbits that connect saddle fixed points or saddle limit cycles in the system's state space", as described by Rabinovich et al. [99] nearly 25 years later. In the pictures described by Edelman and Rabinovich, neuronal groups are engaged in endless cycles of winnerless competition as new input is received and new feedback signals are generated.

Fourth, Edelman's model emphasized the central role of adaptability in forming associations between patterns observed repeatedly: "...the selection of certain subgroups results in an alteration of the probability that these subgroups will be selected again upon a repeated presentation of a similar stimulus pattern." Incorporating the ideas of Hebb nearly 30 years earlier, Edelman assumed this alteration is "a result of synaptic alteration" so that "connectivity is functionally altered." Twenty years later, the role of STDP and dendritic activity would be identified as a primary means of achieving these altered connections [76, 89]. Edelman recognized the utility in these adapting functional networks for accomplishing associative memory that is content-addressable, emphasizing that within this model, "Memory readout is not posed as a special problem; the *process* does not differ from other forms of neuronal communication..." Thus, unlike digital computers, information processing and memory access are not separate activities in neural systems.

Finally, the model proposed by Edelman had a place

for a type of central control unit that could receive bottom-up input from primary neuronal groups, coordinate top-down feedback from higher-order associative centers to the primary groups, and establish coherent cycles in which this information exchange can occur. Edelman identified the thalamocortical complex as the primary candidate for achieving this system-wide coordination. In this architecture, modules at the system level are specialized for processing certain types of information (visual, olfactory, language, etc.), and each of these modules comprises many neuronal groups specialized for detailed processing of this class of information. Thus, the hierarchical, modular architecture of the brain extends from the microscale of groups of neurons up to the sophisticated thalamocortical complex that coordinates system-wide activity.

Edelman laid out multiple testable predictions that could be used to evaluate and falsify his theory. After forty years, his ideas have been refined, but not overturned. In designing hardware for general intelligence, we should expect similar principles to pertain, and we should incorporate device mechanisms to facilitate these operations at the local level of neuronal groups all the way up to the system level of the thalamocortical complex.

#### 2.4.2 Baars' Model of the Global Neuronal Workspace

Building on the concepts developed by Edelman, Bernard Baars introduced the concept of the global workspace in 1988 [97]. In this model, Baars describes a set of specialized experts vying for access to a shared chalkboard on which they can broadcast their messages for viewing by all the experts. In the most basic construction of Baars' model, the experts can all view the information on the chalkboard as well as write upon the chalkboard when granted access. In Baars' words, "This simple model has only two theoretical constructs: a set of distributed specialized processors and a global workspace or 'blackboard,' which can be accessed by a consistent set of specialists and that can, in turn, broadcast information to all others." ([97] pg 71)

Each of the specialized processors is an expert in interpreting a certain type of information or extracting certain features from inputs. Anatomically, they correspond to minicolumns, columns, and brain regions, all specialized at different levels of hierarchy. A minicolumn may respond maximally to lines of a certain orientation, while a region (or collection of regions [?]) may respond maximally to faces. At all levels of hierarchy, the thalamocortical complex controls which specialized processors have access to the global neuronal workspace (represented by the chalkboard in Baars' model) at a given moment. Once a specialized processor, or combination of mutually engaged processors, gains access to the global workspace, the information they are generating is shared broadly with the entire network of processing experts. All sub modules have access to the information globally broadcast and can use

it to inform their activities. At any given moment, the information shared across the global workspace is associated with a train of thought, or a coherent cognitive sequence, enabling cognitive resources to be focused on the subject that has stimulated the dominant coalition of processors, thereby gaining that coalition access to the global workspace.

In the present context of informing the design of hardware for cognition, Baars' model reiterates two themes we have been discussing: 1) neuronal circuits must be highly adaptive and capable of transitioning between myriad dynamic ensembles as activity proceeds; and 2) the communication infrastructure must enable each specialized processor to readily communicate to and engage with many other processors all across the network with short communication pathways. We will return to these themes in Sec. 3 in the discussion of approaches to artificial hardware for cognition.

#### 2.4.3 The Role of Synchronization in Cognition

By the early 2000s, the ideas of Edelman and Baars had been refined and supported by experimental evidence. We can get a snapshot on the understanding of cognition at that time by considering two review articles that appeared in *Nature Reviews Neuroscience* in 2001.

In Ref. [98], Varela et al. review "the large-scale integration problem": "How does the brain orchestrate the symphony of emotions, perceptions, thoughts and actions that come together effortlessly from neural processes that are distributed across the brain?" In tracing the operations of the brain to the manifestation in hardware, Varela et al. identify neural assemblies as the central construct. They define neural assemblies as "distributed local networks of neurons transiently linked by reciprocal dynamic connections." Neural assemblies are central to understanding cognition because "the emergence of a specific neuronal assembly is thought to underlie the operation of every cognitive act." With the term "neural assemblies", Varela et al. are not referring to the neuronal groups of Edelman or the specialized processors of Baars, but rather to the network of neurons that is transiently excited due to their collective relevance in processing a given stimulus. A neural assembly may be relatively localized, or it may span many modules across the network. Varela et al. describe two types of connections necessary to form these assemblies: one involves "reciprocal connections within the same cortical area or between areas situated at the same level of the network." The other involves "connections that link different levels of the network in different brain regions to the same assembly and embody the true web-like architecture of the brain."

A central question regards the mechanisms by which these neural assemblies coordinate their activities to achieve information integration across the network. Varela et al. argue that two concepts are relevant to understanding information integration: 1) bottom-up and top-down

activity; and 2) phase synchronization. Regarding the former, the authors clarify that “Bottom-up and top-down are heuristic terms for what is in reality a large-scale network that integrates both incoming and endogenous activity; it is precisely at this level where phase synchronization is crucial as a mechanism for large-scale integration.” The role of synchrony is elaborated in the context of visual binding, where the goal is to answer the question, “how are the different attributes of an object brought together in a unified representation given that its various features—edges, colour, motion, texture, depth and so on—are treated separately in specific visual areas?” The authors propose the answer, “visual objects are coded by cell assemblies that fire synchronously...visual binding refers to the ‘local’ integration of neuronal properties (that is, integration that takes place within neighbouring cortical areas, all specialized in the same modality), which allows the large-scale integration necessary for vision in the context of a complete cognitive moment. We argue that synchronization of neural assemblies is a process that spans multiple spatial and temporal scales in the nervous system.” In this review I have emphasized the different roles of gamma and theta oscillations for simplicity, but Varela et al. (as well as many others [28, 94, 115–118]) emphasize that transient synchronization occurs across a continuum of spatial and temporal scales. Across these scales, the specific assemblies that are engaged adapt dynamically due to synaptic plasticity mechanisms [7, 74, 81], functional reconfiguration through inhibition [119], and frequency-selective communication [72]. The authors acknowledge that information integration through coherent synchronization is not independent of concepts from dynamical systems: “The transient nature of coherence is central to the entire idea of large-scale synchrony, as it underscores the fact that the system does not behave dynamically as having stable attractors, but rather metastable patterns—a succession of self-limiting recurrent patterns.”

In another 2001 review, Engel, Fries, and Singer offer additional insights into the role of oscillations and synchrony in top-down processing [120]. These authors define top-down influences as “intrinsic sources of contextual modulation of neural processing” and argue that “processing of stimuli is controlled by top-down influences that strongly shape the intrinsic dynamics of thalamocortical networks and constantly create predictions about forthcoming sensory events,” echoing many of the concepts related to reentrant feedback discussed by Edelman. Engel et al. argue that “Coherence among subthreshold membrane potential fluctuations could be exploited to express selective functional relationships...and these dynamic patterns could allow the grouping and selection of distributed neuronal responses for further processing.” They emphasize that synchronization through the joint enhancement of response saliency can select and group subsets of neuronal responses for further joint processing. So, synchronization can be used to encode information about the relatedness of neural signals....” Yet the main thesis of En-

gel et al. regards the role of top-down processing. They argue that “top-down factors can lead to states of ‘expectancy’ or ‘anticipation’ that can be expressed in the temporal structure of activity patterns before the appearance of stimuli.” Further, “not only changes in discharge rate, but also changes in neural synchrony, can be predictive in nature.” Not only does phase coherence of ongoing synchronized oscillations play a predictive role, but synchronization of first spikes generated by two neurons upon presentation of a stimulus can enable the network to anticipate the contents of the full stimulus. From the perspective of Engel, Fries, and Singer, top-down state modulation plays a crucial role in preparing the network to contextualize new information and in adapting the network in the presence of changing stimulus. These ideas are consistent with but extend beyond the framework of reentrant, phasic processing as described by Edelman.

#### 2.4.4 Communication Through Coherence

By 2015, Fries had elaborated these ideas still further. In Ref. [121] he argues, “...dynamic changes in synchronization can flexibly alter the pattern of communication. Such flexible changes in the brain’s communication structure, on the backbone of the more rigid anatomical structure, are at the heart of cognition.” This perspective illustrates the significance of the brain’s ability to adapt its single structural network into myriad functional networks under the influence of ever-changing information. Fries asserts the significance of coherence in communication based on the idea that “Inputs that consistently arrive at moments of high input gain benefit from enhanced effective connectivity. Thus, strong effective connectivity requires rhythmic synchronization....In the absence of coherence, inputs arrive at random phases of the excitability cycle and will have a lower effective connectivity.” Most crucially, “A postsynaptic neuronal group receiving inputs from several different presynaptic groups responds primarily to the presynaptic group to which it is coherent. Selective communication is implemented through selective coherence.”

#### 2.4.5 Experimental Progress Identifying the Mechanisms of Cognition

With modern brain imaging techniques and clever experiments, it has become possible to test aspects of the global neuronal workspace theory. Stanislas Dehaene has conducted some of these experiments and found clear correlations between the experience of becoming aware of a stimulus and the physical activity corresponding to that experience, and his 2015 book “Consciousness and the Brain” provides an accessible summary of the field for the uninitiated. Dehaene describes how a “priming” stimulus (such as an image of a number or a word) can be shown to excite characteristic activity in a localized brain region, and even shown to affect the activities of other regions. But only when the local activity leads to a cascade of global activity

does the subject report becoming aware of the stimulus. Subconscious processing of information stays local and decays rapidly upon removal of the stimulus, while conscious processing spans the network and can remain active long after the stimulus is no longer present.

This emerging picture, supported by a great deal of experimental evidence, has a great deal in common with the ideas of Edelman as well as Baars: local, specialized processors deal with information of a specific type, and the full comprehension of this information can only occur across wider regions of the network that sample input from many specialized processors. This information is integrated across longer temporal scales, and on these longer temporal and spatial scales, the architecture of the network enables the information to be shared broadly and compared to prior learned models of the world through a combination of feed-forward and feed-back connections spanning the hierarchy of the network. Dehaene refers to these distinct patterns of activity that occur when a stimulus becomes conscious as avalanches. Small avalanches that remain confined to a local region remain unconscious. Large avalanches that induce activity across the global neuronal workspace lead to conscious awareness. In describing recordings of such avalanches, Dehaene writes, “Distant brain regions also became tightly correlated: the incoming wave peaked and receded simultaneously in all areas, suggesting that they exchanged messages that reinforced one another until they turned into an unstoppable avalanche. Synchrony was much stronger for conscious than for unconscious targets, suggesting that correlated activity is an important factor in conscious perception.” [?] In this description, Dehaene ties together several of the concepts I have emphasized here, including neuronal avalanches, the role of synchrony, and information integration across space and time. The model of cognition championed by Dehaene is coherent with concepts coming from Edelman, Baars, Engel, Fries, Singer, and others. The models have been refined over time, and much remains unknown, but the coarse outline has stood the test of time.

#### 2.4.6 Lessons for Design of Hardware for General Intelligence

The concepts presented in this section represent our guideposts for designing hardware capable of general intelligence. Most importantly, an intelligent system must be able to efficiently exchange information across many scales of space and time. This sounds obvious, but in our present computing hardware contains many pinch points where information going to very different destinations is serialized, and communication events simply wait in line. This method of information exchange is antithetical to the way computing occurs in the brain.

More specifically, at a network level the presence of a new stimulus must rapidly lead to the establishment of new synchronized states. This can only occur if informa-

tion can be efficiently communicated across the entire network. Local assemblies must respond to the new stimulus on the time scale of gamma oscillations, and large numbers of these assemblies must be sampled on the time scale of theta oscillations. Each local assembly must be able to broadcast information to many recipients to achieve efficient bottom-up communication to higher-level processing centers, and these centers must also be able to broadcast their activity across the network to achieve efficient top-down communication and functional reorganization. Many connections are in close spatial proximity, but many are not. The white-matter wiring and the architecture of the constituent modules cannot be separated from the cognitive functions they perform [35]. The hardware infrastructure for achieving this communication must not introduce competition for bandwidth on a shared communication infrastructure, and it is for this reason that all communication channels must have dedicated axonal fibers. Regardless of the physical medium employed for signaling, if the communication infrastructure is shared and requires serialization, there will be a scale beyond which specialized processors cannot be adequately sampled, and the processing hierarchy will be constrained.

The prior argument informs the design of the communication network, and at the device level, quantities as simple as time constants play an important role in determining the frequencies at which synchronization can occur and the durations over which coherence of activity can be maintained. The relevant frequencies span many orders of magnitude, and the ability to achieve a wide range of time constants with accurate control by design appears highly advantageous for enabling the synchronization-based exchange of information across the network. Additionally, devices that perform nonlinear transfer functions and make use of the temporal domain with spiking signals will be particularly equipped to harness the insights gained from neuroscience.

Finally, it is important to include large numbers of primitive elements to provide the broad repertoire necessary for general intelligence, a fact that was not lost on Ramón y Cajal, Mountcastle, or Edelman. In this regard both size and power dissipation must be taken into account. Neurons must be small enough that signals can propagate between them during the period of network oscillations without being limited by communication delay, and the energy per synaptic event must be low enough that networks with trillions of synapses can function with tolerable power density and net consumption. Section 3 considers several approaches to constructing hardware toward these ends.

columns = modules = specialized processors = sub-conscious experts

[6] polling of groups, similar to Baars; requires efficient communication

as evolution proceeded, animals developed larger brains with more columns as well as greater interconnectivity. This is consistent with the picture of path length on a random graph (see ed1978 pg 71)

ed1978 uses terms “group” and “repertoire” “central states”

role of oscillations, clusters/avalanches, dynamic states in Edelman’s model

## 2.5 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 pass 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  $\tau_{ij}$  seconds? Or, how much has neuron  $i$  been bursting, and then quiescing, and then bursting again in the last  $\tau_{ij}$  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  $\tau$  seconds? Or, how many of a particular subset of 10 neurons in cluster  $q$  have stopped firing in the last  $\tau$  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 [80,85]. 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  $f_\theta$ , so that the information contained in the collection of all synapses, dendrites, and neurons active in the functional network resonant at  $f_\theta$  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  $f_\gamma$  is modulated by slower frequencies  $f_\theta$ , 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 [28], perhaps by enabling access to the global neuronal workspace [97,122].

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.

What is the significance of oscillations and synchronized clusters? Much has been written about this [], and perhaps much remains unknown. Here I offer one interpretation based on the references presented thus far. Within a minicolumn there are about 100 neurons, and these primarily participate in small neuronal avalanches, power-law distributed from size one to the complete 100, and also extending far beyond the minicolumn. These avalanches can be the result of spontaneous, stochastic activity, or they can be induced by outside stimulus. These transient, synchronized bursts of varying numbers of neurons primarily between locally connected clusters of neurons make up gamma oscillations. Single minicolumns and clusters of minicolumns up to the scale of columns can give rise to a very large number of these dynamical states, each with a distinct set of participating neurons and synaptic connections. In the language of dynamical systems, each pattern of activity is a stable heteroclinic channel, which is a sequence of successive metastable states. The timescales and resonant frequencies of activity are tuned by synaptic, dendritic, and somatic time constants adapt in response to local applied voltage (due to afferent synaptic activity) as well as neuromodulators that may adjust resonant frequencies of sectors of neurons simultaneously.

Groups of neurons spanning minicolumns to columns for a dynamical basis set [?,95] that can represent some range of stimuli. The fastest transient synchronized states

occur within minicolumns, and a single minicolumn can produce a large number of different patterns dependent on the stimuli. Each minicolumn is capable of representing a certain class of features or input stimulus, i.e., a certain color, spatial frequency, intensity, etc. Synchronized ensembles at slightly lower frequencies form between groups of minicolumns, forming specific coalitions of neurons representing a distinct state across multiple dimensions of feature space, providing a broader range of insights regarding the nature of a stimulus.

The pattern repeats at still lower frequencies, where dynamical states of columns communicate amongst each other to form brain regions. These complex structures of neural architecture are supremely equipped to analyze certain aspects of the world, such as faces, language, odors, etc. At the scale of brain regions, oscillations are in the theta frequencies. Activity on the scale of columns organizes on gamma timescales (30 Hz-80 Hz), and activity across the cortices, integrating input from activity in many columns, occurs ten times slower, in theta oscillations (4 Hz-8 Hz). Neurons in any minicolumn can talk to neurons in any region on theta timescales, provided network path lengths are short.

Local clusters interacting on gamma timescales with activity modulated by broad network activity on theta timescales illustrates the meaning of the phrase, “information integration across space and time.” However, it is to be understood that many spatial and temporal scales are involved in cognition, and this hierarchical nesting of scales achieves the fractal organization of the brain that appears uniquely capable the efficient information integration that enables cognition. I take the principles of neural information summarized here as the principles guiding the design of cognitive systems. Next I describe several approaches to hardware for artificial intelligence.

From an architectural perspective, there is one more level of hierarchy to consider, and that is the thalamo-cortical complex []. Brain regions input a representation of their activity to the thalamus, and it decides what information to broadcast across the network. The thalamus must be able to receive from and reach all brain regions in a tremendous feat of network communication.

themes common to all three pictures:

1. space and time intertwined
2. enhanced responsivity to inputs and resilience to noise
3. role of inhibition critical
4. role of/relation to plasticity

Regarding computational devices, each synapse performs filtering operations, each dendrite performs nonlinear transfer functions, and each neuron is a complex processor adapting on multiple time scales and performing network computations with the different transient ensembles engaged to process different information. We should expect hardware capable of general intelligence to incorporate devices and circuits capable of performing these functions efficiently and directly without being emulated by a digital system stepping through differential equations.

Regarding the communication infrastructure, we seek to enable networks with many nodes and short average path length made possible by neurons making many connections. The network must have a high small-world index, requiring long-distance connections, and the network must have a modular configuration with specialized processors combining their information at progressively higher levels of hierarchy over larger spatial and temporal scales to form representations of stimulus with fine detail as well as comprehensive context. The communication network must achieve this graph structure directly and not via emulation on a wholly different graph of nodes for communication.

We next consider multiple routes to creating hardware achieving these goals of communication and computation in neural systems for general intelligence.

The fractal use of space and time appears uniquely suited to achieving the network information integration that is necessary for cognition and general intelligence. Intelligent technological systems may depart from the specifics of biological systems in many respects, but they are likely to share this basic mathematical construction.

### 3 Candidate Hardware Platforms for General Intelligence

It is important to point out that most efforts in neuromorphic hardware do not aspire to achieve general intelligence

I am a proponent of a specific hardware platform, but I also anticipate extraordinary success from multiple hardware platforms. It seems quite likely that we will witness the emergence of a hierarchical intelligent system making use of various magnetic, electronic, and photonic devices. I fully expect a system with general intelligence to employ simpler modules capable of specific intelligence. These may be feed-forward neural networks, recurrent echo state networks or liquid state machines, or more conventional rule-based AI systems. Hardware for general intelligence

should be designed with interfaces to modules in mind.

we do not consider all hardware here, most notably we spend little time on memristors or MTJs largely because in a mature system they will be incorporated with a CMOS platform; they are devices, not platforms for intelligence

### 3.1 Overview of Machine Learning

Why do deep learning or machine learning at all? - [?] Jutamulia and Yu cite two motivations to study neural networks in the 1980s: “1) Although a computer can perform complex calculations precisely and quickly, when the task is to recognize an object the computer fails or performs poorly, and is extremely slow compared with the human brain or even an animal brain. 2) A computer needs an explicit set of instructions to perform a given task. Thus, a computer is like a loyal and strong slave but who lacks intelligence. Nevertheless, computer users also need intelligent workers, in addition to slaves, that can perform the task without receiving detailed orders.”

- [123] Caulfield, Kinser, and Rogers argued that for the brain relative to digital machines, “Formal reasoning is hard, but inspired guessing is easy.”

- [124] “[T]he learning approach can provide a method for transferring some of the burden of programming the computer from the user to the computer.”

[15]

[125]

#### 3.1.1 Thermodynamic Models

[22]

-strengths of Hopfield model: recognition from partial input, robustness, and error-correction capability [126] - [127] regarding Hopfield model, the authors state, “A remarkable property of the model is that powerful global computation is performed with very simple, identical logic elements (the neurons).”

Emphasize training is slow

from <http://pnb.mcmaster.ca/3w03/hopfield.html>.

Binary Hopfield Network

The standard binary Hopfield network is a recurrently connected network with the following features:

- symmetrical connections: if there is a connection going from unit  $j$  to unit  $i$  having a connection weight equal to  $W_{ij}$  then there is also a connection going from unit  $i$  to unit  $j$  with an equal weight.
- linear threshold activation: if the total weighted summed input (dot product of input and weights) to a unit is greater than or equal to zero, its state is set to 1, otherwise it is -1. Normally, the threshold is zero.

- asynchronous state updates: units are visited in random order and updated according to the above linear threshold rule.
- Energy function: it can be shown that the above state dynamics minimizes an energy function.
- Hebbian learning

The most important features of the Hopfield network are:

- Energy minimization during state updates guarantees that it will converge to a stable attractor.
- The learning (weight updates) also minimizes energy; therefore, the training patterns will become stable attractors (provided the capacity has not been exceeded).

However, there are some serious drawbacks to Hopfield networks:

- Capacity is only about  $.15 N$ , where  $N$  is the number of units.
- Local energy minima may occur, and the network may therefore get stuck in very poor (high Energy) states which do not satisfy the "constraints" imposed by the weights very well at all. These local minima are referred to as spurious attractors if they are stable attractors which are not part of the training set. Often, they are blends of two or more training patterns.

The Boltzmann machine, described below, was designed to overcome these limitations.

The binary Boltzmann machine is very similar to the binary Hopfield network, with the addition of three features:

- Stochastic activation function: the state a unit is in is probabilistically related to its Energy gap. The bigger the energy gap between its current state and the opposite state, the more likely the unit will flip states.
- Temperature and simulated annealing: the probability that a unit is on is computed according to a sigmoid function of its total weighted summed input divided by  $T$ . If  $T$  is large, the network behaves very randomly.  $T$  is gradually reduced and at each value of  $T$ , all the units' states are updated. Eventually, at the lowest  $T$ , units are behaving less randomly and more like binary threshold units.
- Contrastive Hebbian Learning: A Boltzmann machine is trained in two phases, "clamped" and "unclamped". It can be trained either in supervised or unsupervised mode.

Supervised training proceeds as follows for each training pattern:

- Clamped Phase: The input units' states are clamped to (set and not permitted to change from) the training pattern, and the output units' states are clamped to the target vector.
- All other units' states are initialized randomly, and are then permitted to update until they reach "equilibrium" (simulated annealing).
- Then Hebbian learning is applied.
- Unclamped Phase: The input units' states are clamped to the training pattern.
- All other units' states (both hidden and output) are initialized randomly, and are then permitted to update until they reach "equilibrium".
- Then anti-Hebbian learning (Hebbian learning with a negative sign) is applied.

The above two-phase learning rule must be applied for each training pattern, and for a great many iterations through the whole training set. Eventually, the output units' states should become identical in the clamped and unclamped phases, and so the two learning rules exactly cancel one another. Thus, at the point when the network is always producing the correct responses, the learning procedure naturally converges and all weight updates approach zero. The stochasticity enables the Boltzmann machine to overcome the problem of getting stuck in local energy minima, while the contrastive Hebb rule allows the network to be trained with hidden features and thus overcomes the capacity limitations of the Hopfield network. However, in practice, learning in the Boltzmann machine is hopelessly slow.

### 3.1.2 Deep Learning

-need some history -need to introduce convolutional layers -parallels to early stages of vision system in brain

-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.

### 3.1.3 Reservoir Computing

[20]  
[128]



Recurrent neural networks can approximate the trajectory of a dynamical system [129] Recurrent neural networks are Turing equivalent [130]

### 3.1.4 Spiking Neural Networks

## 3.2 General Hardware Considerations

“Given the inexorable progress of digital computation, what can an alternative implementation offer that is more than a transitory advantage?” [?]

“Optics technology enters this scenario offering the potential advantages of massive parallelism and global inter-connection capabilities.” [?]

While at present most semiconductor neural systems employ the von Neumann architecture, this is not a fundamental requirement. We want the arguments made here to be based on the physics of the devices, not the architecture.

In the digital age, we often think about device speed as a primary metric of computational prowess. Both Ge transistors and III-V transistors were originally seen as superior to Si due to higher carrier mobility and thus higher switching speed. Over time it came to be appreciated that multiple properties of silicon for microelectronics struck a balance between competing factors [19], and the moderately slower speed of Si devices proved acceptable in mature systems. The high speed of optical and superconducting devices is often used as a primary argument to justify alternative computing paradigms. It is common for researchers to select a device that is extremely fast when performing a certain function and attempt to build a hardware or computing platform around that device.

One of the most striking insights from neuroscience is that high-speed devices are not required for intelligence. The fastest devices in the brain operate at 600 Hz, and the primary burst signaling is in gamma frequencies up to 80 Hz. It would appear trivial for 3 GHz silicon transistors to demonstrate superior intelligence. Yet a CMOS system with the same number of neurons and synapses as well as device complexity can achieve nowhere near the same network communication speeds or overall performance, principally because the approach to communication is not well-matched to the task. This challenge turns out to be difficult to surmount due to the physics of the devices comprising the system.

An alternative design approach is to begin with system-level considerations for the manner in which information is to be processed and communicated within the network, to consider various devices capable of performing the required functions, and to make decisions based on spatial, energy, and speed considerations as well as manufacturability and therefore cost. Tradeoffs are inevitable. Once an entire, self-consistent system construction has

been conceived, we can assess system size and information processing speed to determine the physical limits and potential for scaling. This is the approach we pursue in this article. We do not necessarily seek the fastest thresholding element or the lowest power switch, but rather we seek an entire platform that makes suitable compromises between all these competing factors to enable neural systems that are scalable to the level of the human brain.

As a quick note on nomenclature, the terms *scaling* and *scalable* were introduced to refer to the steady shrinking of features in integrated circuits and the associated changes in performance. If a technology is *scalable*, the devices can be made smaller and smaller while still performing their intended function. In other related contexts, the term *scalable* refers to the ability to manufacture many devices on a chip or wafer. One may say that a device is scalable if millions of them can be lithographically defined and manufactured with high yield on the surface of a silicon wafer. In this article we will also use the term *scalable* to refer to the feasibility of realizing large systems comprising many chips or wafers that achieve neural operation spanning multi-module systems. A neuromorphic technology will be referred to as *scalable* if neural computation and communication can be achieved across many levels of hierarchy, as described in Sec. 2.1. Scaling originally referred to the process of making devices ever smaller, and here we seek the reader’s forgiveness as we now use the term to refer to the process of making systems incorporate ever more neurons. This use of the term has become more common in recent years. Quantitatively, the *scale* of a neural system will refer to the number of neurons that can be incorporated in the neuronal pool, a concept that will be discussed in Sec. 4.

In the following review of literature regarding hardware for neuromorphic computing, it is clear that very few efforts are focused on achieving cognition comparable to a human and instead target specific tasks under the umbrella of machine learning. Hardware decisions made for machine learning are often not suitable for highly scaled intelligent systems. We attempt to glean useful insights from hardware introduced for machine learning while identifying traits that are not well matched to large-scale intelligent systems. I hope not to disparage any efforts in these areas, but rather to emphasize the unique requirements that emerge when hardware is designed at the outset with advanced cognition as the objective.

In our quest to anticipate hardware that will be successful for large-scale general intelligence, it is prudent to heed the insights of Robert Keyes, one of the most significant thinkers of the 20th century regarding the enabling physical factors as well as limitations of silicon microelectronics. Keyes identified several factors that must be respected if a technology is to be practical [131,132]: [insert list from below] We will touch back to these guiding hardware principles throughout the remainder of this article.

- Keyes in detail

- good logic device:
  - binary
  - output-input isolation
  - fan-out/fan-in
  - high gain
  - low cost
  - miniaturization/compaction
  - power density
  - reliability
  - the ability to continue to scale
- digital logic will always be more appropriate for arithmetic and high-arithmetic-depth calculations. Neural processing is more appropriate for contextualizing disparate information. the two are complementary
- i do not expect that Si microelectronics will be displaced by an alternative hardware platform for digital logic. it may continue to evolve toward the asymptotic physical limit, but a completely different digital system that outperforms Si is unlikely. Si is optimal for the task (digital at 300K) given the possibilities contained in the periodic table
- transistor can be nonlinear resistor. JJ can be nonlinear inductor. how much do we make of this parallel? what does it teach us?
- high gain present in all stages of loop neurons
- input-output isolation observed from synapses to axon hillock (hTron)
- inversion through flux of the opposite circulation or MIs of opposite sign
- electromigration eliminated in superconductors, JJs and loops make better synapses than memristors because is based on the quantum wave function, not the motion of atoms. electromigration in contacts to LEDs minimal due to infrequent firing and low current
- variability of operating temperature needs addressing (low power density, liquid helium coolin)
- multi-input logic gates not a problem in neuro like in digital. analog can be utilized to great advantage
- output impedance-matching problems in JJ circuits eliminated with photonic communication
- he estimates 4x increase from JJ logic. a factor of several hundred may be possible in theory. in practice, it is probably more like 10x, and there are other complicating factors. I agree with his assessment that JJs will not displace Si transistors for logic, though JJ logic is likely to be useful in other cryo computing systems
- writing in 1985 was pre-likharev and pre-soref
- optical computation based on optical bistability requires nonlinear optics, completely at odds with the requirements for energy efficiency as well as memory storage and reconfigurability. light for communication, not computation. do not use light to influence light. use light to influence electrons, and electrons to generate light.
- his comments on high gain in DC squid do not apply to our circuits. The way we use loops with MIs and JTLs does achieve high gain
- his comments on optical fanout limitations also do not apply with integrated waveguides
- our circuits satisfy all of his relevant criteria. At this point the key uncertainty regards to fabrication process (how many back-end planes, JJ  $I_c$  variation, light-production efficiency)
- level restoration in loop neuron synapses and dendrites also established by power lines, as with transistors (noise margins)
- ability to scale enabled with neural architecture and dedicated photonic connections because systems can continue to scale without being limited by communication bandwidth. the architecture itself is scalable due to the fractal concepts discussed in the neural section
- must compare to moving goalposts of continued si developments. we do this by comparing to systems envisioned to have full 3d integration of processing and memory at the wafer scale [133]. we're comparing the asymptotic limits of superconducting optoelectronic at 4K to the asymptotic limits of semiconducting optoelectronic and whatever operating temperature desired. we hope for significant opportunities for semiconductor-superconductor hybrid systems, discussed in sec 5.
- the use of JJs in loops effectively makes them more than two-terminal devices. bias, ground, and MI inputs. output-input isolation is achieved. synapses do not experience cross talk or feedback from the dendrite (-30dB, include calculation) due to high inductance of SI loops. DR loop doesn't receive feedback from the DI loop because of JTL, at least until saturation, which is actually advantageous, because that establishes it enforces high signal level.
- reset of hTron, slow, occurs during the refractory period, perhaps is limiting factor in speed, but it is the same as SPDs, so unless a faster single-photon detector is invented (that still meets all other criteria), this is the practical limit.

- threshold logic - AND is a threshold device. threshold logic works well in neural information processing, just incompatible with digital logic. we do not require such precise establishing of thresholds. we anticipate noise, and allow threshold to drift based on learning.
- laser cavity dynamics and optical bistability for logic have failed for four decades, but even without investigating the reasons why, this approach to large-scale neural systems can be ruled out based on power considerations.
- [132] quote on pg 533 re conviction. The goal of my work right now is to discover evidence to build this conviction for soens.
- as keyes points out, many are lured to the promise of extremely high speed, and sacrifice other critical elements to get there. we do not fall prey to that temptation. timescales of the brain inform us that activity from 100 Hz to 0.1 Hz can lead to complex cognitive function. If we can achieve anything faster than this in artificial hardware while maintaining the connectivity, device complexity, and hierarchical, modular architecture, it would be a tremendous achievement. Approaches to developing artificial neurons that fire at rates above gigahertz often must sacrifice other aspects of system performance. Superconducting optoelectronic networks appear capable of activity at least ten thousand times faster than our own brains. The device complexity, architectural scalability, and low power density are more important than higher speed.

Discussion of Goodman, Miller, others

This article is not oriented toward making near-term predictions in trends for neuromorphic hardware, but is rather oriented toward considering the asymptotic physical and practical limits of fully mature cognitive hardware. Specific physical mechanisms and devices are proposed for neural functions, and an outline of a system architecture is put forth while recognizing that any sketch of a technology in full at an immature phase is highly speculative. The goal here is to capture the broad strokes of the communication and computation infrastructure, the devices that are most promising for each function, and a potential route to a large-scale architecture so that specialists in associated fields can identify strengths and weaknesses in the context of the system as a whole in order to make improvements without breaking other parts of the system.

Neural computing is an excellent example of why human innovation is a faster engine for evolution than random mutations. At present, we are getting very close to the state where every single possible technology for neural computing has been proposed.

[134] -“Inspired by a strongly simplified interpretation of the human brain’s structure, large numbers of simple nonlinear elements (neurons) are connected (synaptic links) into large networks. Information processing in ANNs usually relies on numerous simple nonlinear transformations and large scale linear matrix multiplications. The implementation of these operations in von Neumann architectures is highly inefficient, as it requires massive parallelism.” -“[E]ven the recent astonishing developments cannot mask the fact that currently no ideal ANN-specific hardware, which fully implements physical hardware neurons and physical *synaptic links*, exists. With such a novel platform, many orders of magnitude could be gained in speed and energy efficiency.”

### 3.3 Semiconductor electronic neural systems

[135]

[136]

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 [137, 138]. 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.

### 3.3.1 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 [1], as well as a number of review articles summarizing those efforts [2], 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 [139, 140] or to perform simulations of biological neural networks to advance our understanding of neuroscience [141, 142]. 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.

### 3.3.2 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 [139]. Others include synaptic plasticity mechanisms to enable learning during operation, and bear the costs of reduced speed of network activity [140, 142].

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. [142], “...often the compute budget is dominated by input connections...which imposes an upper limit on the (number of neurons) $\times$ (number of inputs per neuron) $\times$ (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.

### 3.3.3 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 reduced 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/ $\mu\text{m}$ , so parasitic wire capacitance dominates when devices are separated by even a few microns [143]. 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.

### 3.3.4 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 [144]. 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 to 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.

### 3.3.5 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 [145]. 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." [146] 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 [147]. Systems with a few hundred million neurons will likely be limited to operation at 10 Hz or below [148].

### 3.3.6 Contention delay in neural systems

As stated in Ref. [144], "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. 2). 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 it, in Ref. [147] 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 [133] 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.

### 3.3.7 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. 4.

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.

Make sure to discuss how shared communication infrastructure has two graphs: physical graph of switching network and simulated graph of neural network. When connections of neural network are high, path length is low in simulated space, but extremely high (40 for long-range connections across a wafer), and communication suffers. For networks on the order of 100,000 neurons, average event rates of 1 kHz have been achieved [147]. For networks of 250 M neurons, event rates are predicted to be limited to 10 Hz, and this is with relatively simple point neurons with no dendritic circuitry [133]. Address event representation may not scale much beyond this range. More generally, the requirement that each neuron store in local memory and amount of address information that scales with the number of neurons in the network (or dendrites in the network if trees are considered) makes it unlikely that such an approach to communication in neural systems can scale to the light-speed limit. New communication innovations may be required if CMOS is going to reach extreme scale.

### 3.3.8 Actual versus emulated neurons

To close this section, we emphasize what we see as a more general lesson regarding hardware for artificial neural sys-

tems. 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 sub-networks, 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 [18] dating back to Mead in 1990, pre-dating even the early work on address-event representation by Boahen [145]. 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 [148]. 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. Are there other physical mechanisms that can enable communication without the problems of electrons? Photons are a clear candidate for this operation. We next review neural systems employing light before summarizing work on superconducting neural circuits.

A network comprising biological neurons can respond at the speed of biological neurons, while a network comprising silicon neurons can respond at the vastly slower speed limit set by the shared communication network, which grows more sluggish as more neurons are added to the network.

Silicon transistors evolved to perform binary logic operations.

need to introduce TPUs

### 3.4 Photonic Neural Systems

The strengths of optics in computing have been acknowledged for decades and have led to many attempts to design digital computing architectures around optical phenomena and devices. Like AI and superconductors, there have been several waves of photonics in computing. The first wave of interest began in the late 1970s and was triggered by the seminal paper from Goodman, Dias, and Woody regarding the use of optics for performing Fourier transforms with incoherent light [149]. This work generated a great deal of interest because it showed the power of optics for performing matrix-vector multiplication with complex quantities in a parallel manner using incoherent light that does not have phase as an accessible parameter. A primary favorable feature of optical matrix-vector multiplication is that the time required for the operation is, in principle, nearly independent of the dimension of matrix. This is in contrast to a quadratic temporal scaling with conventional digital approaches. While other Fourier transform devices had been proposed, including in slab-guided modes in integrated photonic systems [150, 151], the work by Goodman, Dias, and Woody combined several key ideas at the right time to generate significant excitement around the use of optics in computing. The optical approach to matrix-vector multiplication introduced in Ref. ?? is now referred to as the *Stanford architecture*, as the authors were working at Stanford University at the time of the publication.

Themes raised in the Stanford architecture persist to the present. The primary there is the use of light for highly parallel computation. Each light source can straightforwardly fan its signal out along many optical paths, and

parallel computation can occur along each path independently. More generally, the primary strength of optics lies in communication—the ability to rapidly move information from one place to another without degradation. This strength is fundamental to the uncharged, massless, bosonic nature of photons which allows them to propagate without interaction. This strength is utilized from kHz and MHz frequencies for radio communication up to hundreds of THz for intercontinental data transmission over optical fiber.

Another theme raised by the Stanford Architecture is the use of optics for special-purpose hardware accelerators that perform a specific task with performance unmatched by electronic approaches. Goodman et al. focused on the Fourier transform, but matrix-vector multiplication is the underlying operation. The non-interacting nature of light makes it excellent for parallel data transmission, which enables efficient matrix-vector multiplication. However, to construct a general purpose computer in the model of Turing, nonlinear operations are necessary. Efforts to augment the strength of light for communication with nonlinear optical devices for *all-optical* computing surged in the 1980s following the work of Goodman, Dias, and Woody. These efforts can be separated roughly into two categories, with one track being focused on displacing silicon electronics for digital logic and general-purpose computing, and another track maintaining the purpose of augmenting silicon electronics with task-specific hardware accelerators that were envisioned to work in conjunction with electronic digital computers. The introduction of the Stanford architecture was shortly followed by a wave of interest in neural networks, and the use of optics for neural networks based on a similar architecture was quickly introduced [126].

In the following subsection, we briefly review efforts to utilize light in digital computing from the 1970s until the present to establish the context for the evolution of photonic neural systems.

### 3.4.1 Optical Logic Elements

As stated above, the strengths of optics for information processing have been known for some time and have inspired many efforts to conceive of and construct systems for all-optical computing as well as hybrid optoelectronic architectures. The primary strengths of light for information processing are the potential for massive fan out and parallelism, low latency, and high bandwidth. These attributes motivated the Stanford architecture as well as a great deal of other research. The weakness of optics for information processing is that precisely the same as its strength: because photons are uncharged bosons, they do not interact, and therefore they do not have nonlinear responses, and are therefore limited in the input-output transfer functions they can instantiate for computation. Efforts to leverage the strengths of optics for computing have paid significant attention to techniques and devices

to give rise to optical nonlinearities that can be leveraged for digital logic gates.

Devices demonstrating optical bistability have been a primary direction of research to enable optical devices that can be used to construct logic gates. The state of the field was reviewed comprehensively in 1982 by Abraham and Smith [?], and the authors summarize the subject: “Optical bistability is a general title for a number of static and dynamic phenomena that result from the interplay of optical non-linearity and feedback.” The 1985 book by Gibbs also summarizes the field at that time [152]. The primary goals of optical bistable devices are to construct optical transistors and optical memory elements. An optical transistor is a device wherein one optical beam controls the propagation of another optical beam. Optical memory is most often implemented by using an optical signal to change the state of a solid-state element, which can then be interrogated optically. It is essentially impossible to induce light to halt its motion, and it is difficult to extend the decay lifetime of photons propagating within a cavity beyond the nanosecond scale, so optical memory based on the storage of photons in a particular location is not feasible.

The first demonstration of optical bistability was achieved by Gibbs, McCall, and Venkatesan in 1976 [153]. The physical system achieving the bistability comprised a Fabry-Perot interferometer filled with sodium vapor. The sodium vapor provided a nonlinear dispersive medium so that the transmitted intensity was a nonlinear function of the incident intensity. Miller, Smith, and Johnston leveraged a similar effect in a Fabry-Perot cavity formed from an InSb crystal in 1979 [154]. In this work, the nonlinear refraction was based on electronic nonlinearity related to states below the semiconductor band gap, and the use of a semiconductor crystal rather than an atomic vapor made the system more likely to be useful in complex computing systems. Reference 154 demonstrated optical differential gain and bistability. Because the system enabled a weak beam to modulate the transmission of a powerful beam, this system provided the first example of an optical transistor. The overall small signal power gain was greater than six. Subsequent work integrated devices leveraging this principle as a 2D array of etalons [155, 156] for the purpose of performing logical operations.

An alternative approach to bistable optical devices was based on the self-electrooptic effect device. Work on self-electrooptic devices was also carried out by Miller, beginning in 1984 [157]. The principle of operation is based on adjustable absorption controlled with an electric field applied perpendicular to the plane of the quantum well that shifts the band edge, known as the quantum confined Stark effect [158]. In a self-electrooptic effect device, the quantum confined Stark effect as well as optical detection are present in a single structure, leading to optoelectronic feedback that produces bistability. A number of configurations of such devices have been implemented, with the symmetric self-electrooptic effect device introduced by



Lentine et al. being the most successful [159, 160]. In this implementation, two  $p$ - $i$ - $n$  diodes with quantum wells in the intrinsic regions are combined, each serving as the load for the other. As Lentine states, “This device has complementary outputs whose switching point is determined by the ratio of the two optical input powers and acts as a set-reset latch.” One strength of this device is that it has time-sequential gain, meaning once set with low-power beams, it can be subsequently read out with high-power beams. This results in good input-output isolation because the large output occurs at a later time than the inputs. Additionally, while previous electrooptical devices with optical bistability required critical biasing very close to a nonlinear threshold, no such biasing was required.

A switching network comprised of symmetric self-electrooptic-effect devices was demonstrated by McCormick et al. in 1993 [161]. This free-space optical system comprised a  $32 \times 32$  switching node array of these devices. Light propagated perpendicular to the plane of the device array, and 16 input channels were connected to 32 output channels with each channel transmitting 32 bits in parallel. The input data to be routed entered on optical fibers and was coupled to free space. Six stages of electrooptic devices were present, each with input control lasers, electronic control through an electronic computer, and with optics between the stages. The output was again coupled to fibers. (consider showing figure 2 from [161])

Other approaches to the creation of optical transistors have continued to be proposed as nanoscale patterning of photonic devices has become commonplace. One example is the design of an optical transistor based on a photonic crystal cavity [162] that was presented in 2003 by Yanik, Fan, Soljačić, and Joannopoulos. In this on-chip structure, two waveguides lead into a photonic crystal cavity, and two waveguides lead out. In this case, the photonic crystal cavity plays the role of the Fabry-Perot utilized by Miller et al., and a control beam input into one of the inputs shifts the resonance of the cavity through the Kerr nonlinearity. With this beam present, transmission of the signal beam is high, while without the control beam, transmission of the signal beam is low. The authors simulated this operation with material parameters corresponding to AlGaAs, and found that with a modest  $Q$  factor of 5000, the contrast ratio between on and off states could be as high as 10 with a few milliwatts of input power.

Another optical transistor leveraging nano-scale optical phenomena is based on switching a single molecule between internal electronic states and has been demonstrated by Sandoghdar’s group in 2009 [163]. Other recent efforts to realize bistable optical devices for computing and telecommunications include the use of graphene [164, 165], semiconductor quantum wells [166, 167], and semiconductor quantum dots [168].

After 40 years of effort in optical logic devices, no candidate looks promising to displace the electronic transistor for large-scale digital computing. As Miller stated in 2010, “Only one device has apparently ever satisfied [the

success] criteria well enough to allow large logic systems to be constructed.” [169] Miller was referring to the self-electrooptic-effect devices of Lentine [160] and the switching systems of McCormick [161]. Difficulties with optical logic devices do not mean the field of photonics has stalled in the past several decades. To the contrary, tremendous progress has been made. Much of this progress has been enabled by the birth of silicon photonics. Let us put this discussion of optical logic devices to the side while we discuss integrated silicon photonics. Then we will attempt to synthesize what the past 40 year of research in optical technologies can teach us regarding the design of hardware for cognitive systems.

### 3.4.2 Integrated Silicon Photonics

By the mid 1980s, silicon microelectronics was well established as the supreme technology for computing. The scaling predictions made by Moore in 1965 [7] had held for two decades, while other material platforms for electronics and photonics had not met with the same success. Within this context, a new perspective on the role of light in computing was put forward in a series of papers by Soref, Lorenzo, and Bennett from 1985 to 1987 [170–172]. The state of the field was reviewed by Soref in 1993 in Ref. [173]. While other materials (primarily compound semiconductors) had been developed for integrated optical components prior to the work by Soref et al., this series of papers pointed to the potential for integrated optical components to be incorporated with silicon microelectronics monolithically. As the authors stated, “Silicon is a ‘new’ material in the context of integrated optics even though Si is the most thoroughly studied semiconductor in the world. There is reason to believe that Si can serve as the medium for a variety of guided-wave optical components in much the same manner as III-V semiconductor compounds, while at the same time avoiding the inherent complexities of binary, ternary, and quaternary alloys.” [171] They further describe their two primary motivations: 1) to utilize the fabrication processes that have been developed for the Si electronic circuit industry in the production of photonic devices; and 2) to monolithically combine silicon electronic circuits with guided-wave optoelectronic components. The goal was to follow the model of the integrated circuit that had become tremendously successful by the mid 1980s. This model utilizes lithographic fabrication to realize system complexity that can be achieved only through integration of components on a chip.

In the first paper of the series, passive waveguides and power dividers were demonstrated at  $\lambda = 1.3 \mu\text{m}$ , a wavelength at which optical fibers are highly transmissive, indicating the potential for silicon photonic components to interface with both electronic computational infrastructure as well as fiber-optic communication infrastructure. Optical confinement was achieved in the vertical dimension by epitaxially growing intrinsic silicon on a heavily  $n$ -type doped silicon substrate, which has a slightly lower

index of refraction <sup>3</sup>. In the lateral dimension, confinement was achieved by etching a fraction of the depth of the epitaxial intrinsic silicon, leading to a so-called rib waveguide configuration <sup>4</sup>. The waveguide structure is shown in Fig. ??(a) (use Fig. 2 from [170]). The passive power splitter is shown in Fig. ??(b). These are the first silicon photonic components.

In the second paper on the subject, Soref and Lorenzo describe active silicon photonic components based on free-carrier dispersion effects [?]. Because the silicon lattice is centrosymmetric, a linear electrooptic effect is not present in bulk crystals, and therefore many had not considered silicon a candidate for active optical components. The insight by Soref and Lorenzo to utilize free-carrier effects instead opened many opportunities that would lead to over three decades of technology development. In Ref. ?, silicon-germanium compounds were also proposed as complimentary materials for waveguiding, and silicon-on-insulator (SOI) structures were considered as candidates for a layer structure capable of optical confinement.

A subsequent paper by Soref and Bennett theoretically analyzed the index perturbation achieved by injecting charges or depleting them from silicon [172]. This work had a tremendous impact because it made the case that active components such as modulators and switches could be created in silicon by shifting the phase of light propagating through an interferometer or resonator. Soref and Bennett showed that a sufficient phase shift could be induced to produce a useful active device without incurring intolerable free-carrier absorption. Essentially all present-day silicon modulators are based on these free-carrier effects.

At the time of the original work by Soref et al., SOI technologies were just beginning to be developed, and epitaxial intrinsic silicon grown on heavily doped silicon gave the best waveguiding results. As silicon-on-insulator technology improved—particularly through device-layer transfer rather than epitaxial growth—the situation changed dramatically. The ability to fabricate a thin layer of intrinsic crystalline silicon on top of an electrically insulating layer with lower index of refraction has proved consequential for both electronic and photonic devices. The basic layer structure of SOI is shown in Fig. ??. In particular, the properties of the Si/SiO<sub>2</sub> interface are crucial for enabling low-loss SOI waveguides. Just as the physical and chemical properties of SiO<sub>2</sub> and the interface with Si were one of the primary factors that led to silicon being the most successful material for VLSI electronics (see Sec. 3.2), these basic material properties are central to the success of silicon photonics, as will be discussed below.

The ideas of Soref, Lorenzo, and Bennett pointed

to the potential for exciting integrated photonic devices based on silicon, but the field of silicon photonics did not launch in earnest until SOI technology became mature. Around the year 2000, SOI technology had improved to the point where major semiconductor manufacturers were using SOI wafers for the production of high-performance transistors. This shift was motivated by performance enhancements such as the higher speed at the same supply voltage, or lower voltage and therefore lower power consumption at the same speed as previous technology generations [176]. As transistor gate lengths continued to decrease, the carrier confinement provided by a thin device layer on top of an insulator became necessary to reduce short channel effects.

Early silicon electrooptic modulators utilizing charge-carrier effects outlined by Soref, Lorenzo, and Bennett emerged in 2004 as SOI with thicker buried oxide for photonic confinement became commercially available. These employed MOS capacitor designs [177, 178] as well as and microring-based [179, 180] *p-i-n* carrier-injection designs to reach data rates in the Gbps range. Significant progress occurred rapidly due to the ability to leverage existing silicon manufacturing techniques. The state of the field of silicon electrooptic modulators was reviewed in 2010 by Reed et al. [181] and more recently by Witzens [182], with data rates of modulators employing charge-carrier effects now in the range of 50 Gbps [183].

In 1993, Soref wrote, “The decade of the 1990’s is an opportune time for scientists and engineers to create cost-effective silicon ‘superchips’ that merge silicon photonics with advanced silicon electronics on a silicon substrate.” Has Soref’s vision of the superchip come to fruition? Passive and active silicon photonic devices have affected an astonishing range of application spaces, from medical diagnostics to telecommunications, but the computing industry has not yet been dramatically transformed by silicon photonics. Silicon transceivers are becoming commonplace in data centers, but primarily as separate modules for interconnection between server racks. Multiple silicon photonic foundries exist throughout the world, confirming Soref’s expectation that silicon photonics would benefit from knowledge gained by the silicon electronics industry. The accomplishment that comes closest to the superchip has leveraged the “zero-change” approach to electronic-photonic integration. The principle is to produce photonic components in a CMOS process with no changes to the process itself. Because any CMOS process is finely tuned to achieve high transistor performance, any attempt to introduce changes to the process to accommodate photonic devices will degrade the performance of electronics and significantly increase cost. To insert photonic com-

<sup>3</sup>The index change due to free carriers can be calculated with the classical dispersion formula given in Ref. [170]:  $\Delta n = -(e^2 \lambda^2 / 8 \pi^2 c^2 n \epsilon_0) (\Delta N_e m_{ce}^* + \Delta N_h m_{ch}^*)$ , where  $e$  is the charge of an electron,  $\lambda$  is the optical wavelength,  $n$  is the refractive index of the intrinsic material,  $\epsilon_0$  is the permittivity of free space,  $c$  is the velocity of light in vacuum,  $N_e$  ( $N_h$ ) is the concentration of donors (acceptors), and  $m_{ce}^*$  ( $m_{ch}^*$ ) is the effective mass of electrons (holes).

<sup>4</sup>A rib waveguide results from partial etching of the high-index layer, while a ridge waveguide results from etching completely through the high-index layer. Foundations of optical waveguide theory can be found in Refs. [174] and [175].

ponents into the computing ecosystem in an economically viable manner, minimal disruption is required. Following this model, it has been shown that many passive and active components can be manufactured in a 45-nm CMOS foundry by utilizing the crystalline silicon layer of an SOI wafer that is intended to be used for transistor to produce waveguides as well [184]. Using dopants already present in the process for transistor fabrication, it has been shown that modulators can be fabricated [185,186]. With a similar cavity and doping configuration, and using the SiGe that is incorporated in the process for strain engineering of transistors, photodetectors have been demonstrated [187]. Similar components have been demonstrated in a lower-cost polycrystalline bulk CMOS process [188,189].

The monolithic integration of high-performance CMOS electronics with silicon photonics was used to demonstrate an electrooptic system on a chip combining a processor and random-access memory where the only communication from the chip to the outside was through optical signals generated by on-chip modulators driven by the CMOS electronics [137]. Using two such chips, the project demonstrated direct optical chip-to-chip communication to build a photonically connected main memory system for the microprocessor. The microprocessor communicated to the memory on the other chip, and memory-intensive applications were demonstrated using only this optical processor-to-memory link [137]. The state of the field of zero-change photonic-electronic integration was reviewed by Stojanović et al. in 2018 [138].

Given that these integrated silicon optoelectronic systems were demonstrated in 2015 based on a fabrication in a high-performance technology node, should we expect all future processors to implement photonic communication? Not necessarily. A number of practical and technical barriers exist to wide-scale adoption of photonic communication on the chip. On the practical side, the CMOS industry is a highly complex entity with established methods of integration between myriad components at many levels of information-processing hierarchy. To change anything about the communication between processors and memory requires a great deal of risk for manufacturers at various levels of the supply chain. Such changes may eventually be driven by performance and economics, but it will not happen rapidly. Historically, it has been much more successful for photonic communication to displace electronic communication over large length scales, and the trend has been to gradually access shorter and shorter length scales. There is no guarantee this trend will penetrate to the scale of the processor on a chip.

The technical reason optical communication between digital processors may not be advantageous on the scale of the chip is because of the difficulty of achieving low-cost, efficient light sources monolithically integrated in a CMOS process. If silicon light sources that operated at room temperature could be manufactured as cheaply and easily as transistors, the landscape of digital computing would be radically different. But despite decades of re-

search [190], no such light source exists, and it probably never will. Even if it did exist, it would not make sense to use optical communication between two nearby transistors for the reasons described by Miller [143]. However, as computing moves toward higher parallelism with many-core architectures, it may make sense for inter-processor communication to be optical, leaving intra-processor to the electrical domain. At present, multi-chip modules look most promising. In such systems, co-packaged electronic-photonic systems combine electronic chips that compute and communicate within the package over high-speed electrical interconnects to a photonic chip. That photonic chip encodes the electronically generated data on an optical carrier to be communicated out of the package over optical fibers. Such systems heterogeneously combine three technologies at the level of the package: silicon microelectronic processors; silicon photonic chips, which are likely to include monolithically integrated CMOS driver circuits; and compound-semiconductor light-source chips that couple to the silicon photonic circuits over short-reach fiber connections or other passive bridge waveguides. While heterogeneous, such systems are still quite exciting, powerful, and may potentially saturate the limits of what is physically and economically feasible. Further progress regarding integration of light sources with silicon and silicon photonics with silicon electronics is likely, but the inability of silicon to emit light at room temperature and the material barriers that add significant cost and complexity to III-V/Si integration stand as significant long-term barriers to dense, monolithic integration of silicon photonics with silicon microelectronics.

Other arguments are made against further integration of photonic components with digital silicon circuits. Most prominently, the size of photonic components is vast compared to electronic components. This is due to nothing more than the wavelength of the respective particles. Light that is likely to be used in information processing has a wavelength on the order of 1  $\mu\text{m}$ , while the wavelength of an electron in silicon is on the order of a few nanometers. Thus, lithographic advances have been able to make devices smaller and smaller (down to about 7 nm) without impinging on the comfort zone of individual electrons, while photonic devices will never be smaller than a few microns squared. Thus, as Levi argued in 2018, a photonic modulator on a silicon chip displaces a circuit comprising 10,000 transistors [191]. For this comparison to be meaningful, the role of the photonic modulator must be comparable to the role of the transistor. Levi argues that, “a typical photonic modulator acts as a simple switch, turning light on and off.” Also, we must assume that a photonic component displaces electronic components from the area that it occupies, rather than residing above or below them. These arguments do often apply in the context of integrating photonic components with silicon microelectronics for digital computing. Modulators are simple switches, and electronic and photonic devices are primarily confined to a single active device layer.

However, as we will argue in Sec. 3.6, the situation is very different in the context of neural computing at low temperature. First, silicon light sources do exist, and they are as easy to manufacture as silicon transistors. The significance of this fact for economic scaling cannot be overstated. Second, photonic devices do not play the role of simple switches, like transistors. Instead, micron-scale light sources produce pulses of photons that fan-out to 10,000 destinations. Therefore, while a single light source may occupy the area of several hundred transistors, each source creates a signal that can reach thousands more destinations. While this one-to-10,000 fan-out operation is not often employed in digital communication, it is essential in neural systems. Third, the area they occupy does not displace electronic components, because superconducting electronic components can be fabricated above the active silicon layer. Silicon transistors are confined to a plane due to thermal considerations: the heat generated by multiple planes of transistors cannot be adequately removed, so a single plane must be exposed to cooling fluid. The same is not true for superconducting electronics comprising Josephson junctions and single-photon detectors. Three-dimensional integration of multiple planes of these devices has been demonstrated, and the limits remain to be explored. These are three of the reasons why the limits to optoelectronic integration of silicon photonics with silicon microelectronics for digital computing may not be the same limits encountered in neural computing, particularly if we take the plunge into liquid helium. Before exploring these concepts that are relevant to superconducting optoelectronic networks in more detail in Sec. 3.6, we next review other efforts to employ light in neural information processing.

discuss fibers as an intermediate step

#### 3.4.2.1 Components Required for Integrated Photonics

- passives: waveguides and routing, beam splitters and power taps (y-junctions, evanescent couplers, adiabatic 50-50 beam splitters), spectral filters (microrings [192] and gratings), waveguide crossings
- sources
- detectors
- electrooptic: modulators and phase shifters (rings, MZIs [193]), SAW transducers for imparting a phase shift

#### 3.4.2.2 Materials for Integrated Photonics

- III-V
- $\text{LiNbO}_3$

- silica
- silicon
- SiN

Active integrated photonic components were beginning to be developed in the mid 1970s [193], and by the mid 1980s many elements of the field of integrated photonics were taking shape [194]. Primary applications included RF signal processing and analog numerical processing [194]. Proposals for all-optical digital computers were made based on the key element of a bistable optical cavity. However, the important sub-field of silicon photonics had not yet emerged.

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 [172]. 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 [181] operating into the 10s of GHz based most commonly on Mach-Zehnder interferometers [178] or microring resonators [180]. 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 [173]. 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 ( $E_g = 1.17 \text{ eV}/\lambda = 1.06 \mu\text{m}$  at 0 K;  $E_g = 1.11 \text{ eV}/\lambda = 1.12 \mu\text{m}$  at 300 K). The buried oxide of silicon-on-insulator wafers becomes absorptive for  $\lambda \gtrsim 2 \mu\text{m}$ . Thus, the transparency window of silicon-on-insulator waveguides enables operation with wavelengths below  $1.2 \mu\text{m}$ , 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 [189].

SOI in early 2000s, guiding light on a chip is a different ballgame

Integrated Silicon Photonics for Communication above a certain length scale in digital electronics -monolithic with processors? -in package? -off-chip light sources? -WDM

[195]

[194] In 1984, Verber stated, “A problem which has been common to almost all efforts to perform numerical computations by optical techniques has been the accuracy limitation imposed by the intrinsically analog nature of the devices. Although the ultimate solution to this problem is generally accepted to be in the application of optical bistable devices to produce fully binary optical systems, the long development times anticipated for these systems has led to several other approaches to higher accuracy computation using analog optical techniques.” Approaches to improving the accuracy of analog optical computations have been proposed [196,197], but neither analog nor digital approaches to all-optical computing have been successful. In 2019, these problems have not been solved, and binary optical systems have not proven capable of displacing CMOS for digital computing.

The inability of all-optical systems to displace silicon microelectronics for digital computing does not mean light has no role to play in advanced computing generally and neural systems in particular. However, this history should be carefully considered when choosing the specific role for light in cognitive hardware.

After decades of research, the phrase “all-optical”

should now set off alarm bells. One must ask if the advantage gained (usually speed) is worth neglecting to utilize the myriad competencies of electrical circuits. (or) One must ask why it might be advantageous to neglect to utilize the myriad competencies of electrical circuits. Optics can be employed in addition to electronics without omitting electrical circuits completely. The goal of all-optical is usually speed and parallelism, and introducing electrical components can reduce bandwidth. Still, for systems as complex as required for cognition, it is unlikely an all-optical solution will surface.

“...it is the need to limit power dissipation that largely constrains clock rates in current electronic devices—lower operating voltages give slower speeds but correspondingly lower energies per operation.” [169].

“...we have to remind ourselves that the field of integrated optics is still in its infancy, still in its research stage, and still searching for its proper role.” [198] “The arguments for optical wiring are understood even down to the chip-level, but chip-scale optical interconnect technology is still in its infancy.” [169]

Displacing Si electronics has failed for the reasons Keyes has pointed out.

At present, a major goal of photonics is to augment CMOS electronic hardware 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.

Need to mention here that the Stanford architecture intended to utilize incoherent LEDs because they are simple, scalable, and do not require precise control of phase across many optical paths.

### 3.4.3 Free-Space Optical Neural Nets

To understand the motivations for free-space optical neural nets, we can consider the perspective from an article by Psaltis in 1987 [21]. The authors write, “It is the ability to establish an extensive communication network among processing elements that primarily distinguishes optical technology from semiconductor technology in its application to computation.” The high fan-out capability of optical signals is a strength of light that is complementary to the strengths of electronics. As Caulfield, Kinser, and Rogers wrote in their 1989 review, “large numbers of neurons and interconnections are the natural domain of optics.” In the 1996 review by Jutamulia and Yu [199],

five merits of optics for neural networks are cited: 1) The velocity of optical signals is independent of the number of interconnections; 2) optical signals are immune to mutual interference effects; 3) Optical signals can propagate in three-dimensional free space; 4) the interconnection can be altered properly using spatial light modulators; and 5) optical signals can be easily converted into electronic signals. The authors specifically emphasize that the key attribute of optics in feed-forward neural networks is (incoherent) matrix-vector multiplication, based on the Stanford architecture of Goodman et al. [149].

The first free-space optical neural network was proposed by Psaltis and Farhat in 1985 [126], the same year Soref and Lorenzo implemented the first silicon photonic waveguides. This work described an implementation of the Hopfield model that had been introduced three years earlier (see Sec. 3.1). The objective was to combine the parallelism and interconnectability of optics, which are linear phenomena, with bistable optical devices to provide the thresholding nonlinearity required by in the Hopfield model. While most implementations of the Hopfield model have used only software running on digital computers, this early proposal can be considered a task-specific hardware accelerator. The hardware proposed by Psaltis and Farhat combined compound-semiconductor LEDs with photodiodes and electronics for a initial implementation of nonlinearity to be replaced by optical bistable devices in subsequent generations. While LEDs and laser diodes have become a mature technology with applications from memory access to lighting, bistable optical devices have still not reached sufficient maturity to enable large-scale systems or commercial products either in digital computing or in neural systems. As we will see, both LEDs and optical bistable devices are presently being pursued for their potential in optical and optoelectronic neural systems.

The proposal for an optical neural network presented in Ref. ?? was experimentally demonstrated by Farhat, Psaltis, Prata, and Paek, also in 1985 [127]. In that work, a Hopfield network was demonstrated, thus achieving an optical implementation of content-addressable associative memory. The experimental setup is shown in Fig. ?? . In this first construction of an optical Hopfield network demonstrated in Ref. [127], LEDs played the role of the logic elements: the state of each neuron was represented by an LED (left of Fig. ??(a)); with the LED off, the associated neuron state was equal to -1, while the LED being on represented the 1 state. The light from the LEDs then propagates through free space, interacts with a memory mask that implements matrix-vector multiplication, much like in the Stanford architecture [149], and the outputs are detected on photodiodes. The system then implemented nonlinear iterative feedback to the matrix-vector multiplier. The nonlinear feedback was carried out by converting an optical signal to the electrical domain with the photodiode, applying gain and nonlinearity in the electronic domain, and converting back to optical with the current drive on each LED, although the intention was to even-

tually eliminate the intermediate electrical processing. In this construction, an array of LEDs represents the input vector, and an array of photodiodes detects the output vector. “The output is thresholded and fed back in parallel to drive the corresponding elements of the LED array.” Through the iterative procedure, the network reaches a steady state that corresponds to the retrieval of the intended memory. In this initial demonstration, a 32-neuron system was implemented with an array of 32 LEDs and 32 photodetectors.

These original studies of optical neural networks sparked significant further interest in the subject. Multiple related research efforts branched off of this starting point. In 1986, Anderson implemented a different form of associative memory [200] based on optical eigenstates in a resonator. As the author explained, “A gain medium internal to the resonator amplifies the field belonging to the eigenmode that most resembles the injected field; the other eigenmodes are suppressed through a competition for the gain.” In this concept, an optical resonator with spherical mirrors remembers Hermite Gaussian fields, and the stored information can be programmed by a user. Programming occurs by writing patterns in a volume hologram that plays the role of one of the mirrors. This original demonstration was able to recall one of two stored patterns based on input of part of one of the patterns.

This idea of using a volume hologram to store memory took hold for optical neural network. In 1986, Jannson, Stoll, and Karaguleff studied the use of a volume hologram to store synaptic states, as shown in the schematic of Fig. ?? (use Fig. 1 of jast1986). While the optical apparatus is complex, requiring a saturable phase-conjugate mirror, the technique can lead to very high storage density. The idea is to have sources, such as the LEDs of Psaltis and Farhat, provide inputs, and use free-space propagation with optical components to steer the beams. The state of the neurons can be controlled by the intensity of the sources or by intensity modulation controlled by a spatial light modulator. But the interconnections between neurons—the synapses—are established and weighted based on the internal state of the volume hologram. Optical-wavelength-scale gratings written in the holographic medium route each input to many outputs, thereby utilizing the massive parallelism and connectivity of optics. In Ref. [201], the authors calculate that this approach can realize  $10^8$  neurons and  $10^{16}$  synapses that can be trained on  $10^6$  inputs.

The use of volume holograms for storage of synaptic weights was further explored in subsequent years. In 1987, Wagner and Psaltis used holographic gratings in photorefractive crystals to store associative memories [202]. In a photorefractive crystal (often  $\text{LiNbO}_3$ ), the memory is written through a procedure where input beams interfere with beams routed from target outputs to interfere within the crystal to establish the volume hologram by ionizing charge traps in the crystal and thereby modifying the index of refraction locally. As the authors stated,

“The recorded hologram is stored in the spatial distribution of the ionized traps in the crystal.” Further, “[W]e must...design a sequence of exposures that can load the appropriate weight values in the finite pool of trap sites that are available.” The procedure implemented a Hebbian-type learning rule to achieve local (pairwise between neurons), associative learning. The networks had a multilayer, feed-forward architecture, and a form of backpropagation was used for training. The work introduced the use of a nonlinear etalon to achieve a sigmoidal response. The objective was still to achieve an associative memory, in this case for image classification. The concepts of this work bear strong resemblance to more modern deep learning, although the hardware is significantly different that what has become commonplace. The authors argued, “This architecture combines the robustness of the distributed neural computation and the backpropagation learning procedure with the high speed processing of nonlinear etalons, the self-aligning ability of phase conjugate mirrors, and the massive storage capacity of volume holograms to produce a powerful and flexible optical processor.”

The photorefractive crystals that store the volume holograms in this architecture were investigated in more detail 1988 by Psaltis, Brady, and Wagner [203]. The technique commonly employs phase-conjugate mirrors, which reflect light back along their original path. In Ref. [203], the objective was to derive fundamental limits to the number of interconnections that could be achieved in feed-forward neural networks using this approach to routing of free-space beams; the practical limits may be quite different. As the authors noted, “...in an optical implementation each grating corresponds to a separate interconnection between two neurons...” They further explain, “There is a nice compatibility between simple (multiplicative) Hebbian learning and holography; the strength of the connection between two neurons can be modified by recording a hologram with light from the two neurons.” The authors concluded, “The density of interconnections which may be implemented in these crystals is limited by physical and geometrical constraints to the range of from  $10^8$  to  $10^{10}$  per  $\text{cm}^3$ . This extraordinary synaptic memory density, and the use of volume storage rather than surface storage, was a strong motivator for this work to persist.

Other approaches for implementation of volume holograms were presented, including with liquid crystal gratings [204]. In Ref. [204], it was also shown that two-state neurons (as utilized in a Hopfield network), could be implemented with liquid crystal optical switched. Beyond volume holograms, other approaches to photonic interconnection were championed. In 1987, Caulfield calculated that 2D arrays of inputs and outputs could be connected all-to-all using a spatial light modulator, in principle capable of achieving  $10^{12}$  connections in a realistic system. In 1988, the group of Yariv utilized spatial light modulators located in a plane parallel to an array of detectors to achieve the synaptic connection matrix [205] in a Hopfield model with binary neurons. In 1989, a Hopfield neural net-

work was implemented by Ito and Kitayama using optical fibers for interconnection between nodes. “The coupling ratio from fiber to fiber represents the synaptic weight of the connection between units.” In that work, a  $5 \times 5$  array was demonstrated, thus implementing  $25^2$  connections. The light sources were compound-semiconductor laser diodes, and signals were detected with silicon photodetectors. In that approach, one network implemented nodes with positive weights, and another network implemented negative weights. In their 1989 review [123], Caulfield, Kinser, and Rogers identify six optical interconnection methods present at that time: 1) fiber optic fan-in/fan-out; 2) holographic in-plane connection in integrated optics [?]; 3) optical parallel matrix-vector multipliers [127]; 4) lenslet-array multiple imaging [206]; 5) thick holographic associative networks; and 6) fixed hologram arrays [207].

Spatial light modulators did not appear on the list, but have been considered for optical neural network applications. Farhat contributed a unique perspective in a 1987 article in which he sought to add artificial intelligence to conventional computer controllers through the hybrid integration of a learning optical system with a digital electronic control system. He focused on supervised learning by: “1) computing the interconnectivity matrix for the associations that [are to be learned] and 2) changing the weights of the links between their neurons accordingly.” “Such self-organizing networks therefore have the ability to form and store their own internal representations of the associations that they are presented with.” Specifically, Farhat intended to use optics to speed up the cumbersome learning/training phase of Boltzmann machines (see Sec. 3.1). He was using an LED array as the inputs to the network with a computer-controlled spatial light modulator to apply weights. Learning was achieved with an algorithm combining photodetection, electronic and computer processing, and feedback to update the SLM. Farhat explained his vision of the hybrid neural/digital system: “[T]he addition of such a module to a computer-controller through a high-speed interface can be viewed as providing the computer-controller with artificial intelligence capabilities by imparting to it neural net attributes.”

In addition to the work discussed thus far, many other studies related to optical neural networks were conducted in the 1980s and 1990s [208–211] before the field lost momentum for a duration. In the late 1990s and 2000s, the field of optical neural networks faced two strong headwinds. Optical computing had not found a technological niche or commercial foothold, and was thus in a state of disrepute. Simultaneously, neural networks had again fallen out of favor. At this time digital computing sailed forth on the wind of Moore. Applications of neural network concepts to optical systems maintained very tight focus on specific applications. For example, in 2003 Kawata and Hirose proposed an optical neural network concept that utilized the wave nature of light as well as frequency multiplexing to achieve a system that could learn appro-

priate phase values for a specific optical communications applications. The goal was not to compete with silicon microelectronics for a broad class of systems, but rather to perform one optics-related function at a level that could not be matched by electronics.

A snapshot of the field in 1993 is captured in the Applied Optics feature issue edited by Wagner and Psaltis [212]. Here we highlight just one article from that issue. In the work by Nitta et al [213] ...

[210]

As the excitement around deep learning has grown in recent years, free-space optical neural networks have seen a resurgence. At least four new papers in this domain were published in 2018 [214–217]. In Ref. [214] Bueno et al. pursued free-space optics for reservoir computing. They demonstrated an optical neural network based on a recurrent graph achieved with a spatial light modulator combined with a digital micro-mirror device accomplishing the variable weights of the output layer. The authors constructed a recurrent neural network of 900 nodes to serve as the reservoir (see Sec.3.1). All connections established with the spatial light modulator were fixed and positive, while all training occurred in the output layer through the digital micro-mirror device that established binary weights. During training, the optoelectronic system steps through discrete time with a specified update equation. In this demonstration, update of 900 nodes occurred at 5 Hz, limited by software control interface. The authors estimate the system could scale to  $10^5$  interconnected nodes, which is a useful scale for reservoir computing.

Another recent paper, this one by Hamerly et al., proposed a system leveraging many of the same free-space optical neural network concepts as have been discussed above with the new idea that the synaptic weights in a feed-forward neural network can be established through homodyne interference [215]. In this case, synaptic weights are not established by diffractive optics, volume holograms, or spatial-light modulators, but rather by a complex configuration of modulators acting on a separate set of optical beam that interfere with the signals represent neuron activations. The signals from the output neurons are envisioned to be read out in a serial manner, thereby implementing a form of time-multiplexed readout to simplify hardware at the expense of operating speed. Nonlinearity is applied electronically, so the sole function of the complex optical apparatus is to perform matrix-vector multiplication—the exact same objective as the Stanford architecture of 1978.

While the multi-homodyne approach adds optical complexity to the setup, an approach from the other end of the complexity spectrum has recently been demonstrated. The work by Lin et al. showed the use of terahertz sources in conjunction with diffractive gratings that were 3D-printed [216]. Multiple planes of diffractive elements were cascaded to arrive at the desired input-output transformation. In this case, all training and design of the

feed-forward networks is carried out on external computers to arrive at a series of gratings achieving the desired network output, and the final gratings were simply printed (made possible by the large wavelength of terahertz light) and placed like cards in slots in the optical path. Using this approach, the team demonstrated MNIST and fashion MNIST classification (see Sec.3.1).

The final recent paper on neural networks utilizing free-space optics to be discussed here illustrates the growing role of convolutional architectures for deep learning (see Sec.3.1). In Ref.??, Chang et al. explore the use of a free-space optical convolutional layer prior to electronic processing. The authors “incorporate a layer of optical computing prior to either analog or digital electronic computing, improving performance while adding minimal electronic computational cost and processing time.” They assume light is incoherent and monochromatic and make use of a phase mask in the Fourier plane to apply the optical transformation that implements the convolutional function. The objective is to utilize optics in the steps for which it is well-suited before transitioning to electronics for further processing, including nonlinearities. Such an approach is most likely to be advantageous if the input signal is already optical, most prominently in the context of imaging. As the authors state, “[A]n imaging scenario where the input signal is already an optical signal easily allows for propagation through additional passive optical elements prior to sensor readout.” Because such an approach results in more hardware complexity, as the neural network would require a free-space optical setup in addition to any electronics, it is most likely to be useful in artificial vision systems where free-space light is the signal, not when digital images need to be processed. In such cases, the efficiency of incorporating an optical convolutional layer optimized for a specific classification problem may save time and power.

In their 1990 review, “Holography in Artificial Neural Networks”, Psaltis et al. summarized their work on optical neural networks [124]. The authors list three primary motivations to move toward neural architectures: 1) massive parallelism; 2) dense interconnections; and 3) realizing computational systems that can learn. Decades later, the motivations are still the same, but the implementations have shifted from associative memories based on Hopfield-type networks to deep learning based on feed-forward neural networks. To our knowledge, no efforts based around the employment of free-space optical apparatus are focused on the construction of a system with general artificial intelligence. Nevertheless, while recent and ongoing work aspires to utilize free-space optics primarily for deep learning and reservoir computing, it is our goal to understand how these technologies may apply to the pursuit of systems capable of cognition. Toward this end, we must ask ourselves why these systems, with decades to incubate, have not displaced conventional hardware for neural networks and deep learning. To begin, Ref. [205] stated in 1988: “Further advances in this field



are...limited due to the absence of efficient and reliable hardware realizations of [neural network] models.” While this may have been true, it ended up being the case that hardware advances in silicon microelectronics proved sufficiently efficient to initiate the presently occurring deep learning boom, thereby significantly reducing the motivation to develop new hardware, particularly drastically different hardware, such as free-space optics.

However, the current deep learning boom is leading to rapid saturation of what CMOS hardware can achieve in this context. Moore and Dennard scaling will not provide the same fuel in the next thirty years as they did for the past thirty years, so is there a new reason to expect some of these concepts of free-space optics to have another chance at large-scale impact on neural networks in the modern context of deep learning? There are reasons to remain skeptical. Specifically regarding holographic connections written within the volume of a photorefractive crystal, the strength of each synapse is determined by optical amplitude during writing. A complex procedure of many exposures establishes synaptic weights, and writing each new synaptic weight can inadvertently modify others. Psaltis et al. wrote in 1990 “Distributed connections have advantages and disadvantages—the disadvantage compared to the localized implementation is the reduced control of individual synapses. The adjustment of the strength of one synapse may inadvertently affect other synapses as well. Accommodating this limitation is perhaps the most crucial research issue in this field...” This problem has not been solved. The density of this storage technique is extremely enticing, but so is the prospect of storing bits in electronic spins of single dopants densely embedded in a solid-state matrix. Accessing the information stored in the matrix is, however, extremely difficult. While extraordinary in principle, such technologies may be inaccessible in practice. While photorefractive media written into a hologram through the use of phase-conjugate mirrors is a beautiful use of sophisticated nonlinear optics, it not likely to be sufficiently advantageous to merit the incorporation of sensitive optical apparatus in mainstream computing technology.

More generally, as stated by Jutamulia and Yu [199], “Neural networks that involve optical implementation are commonly called optical neural networks, although they should be correctly called hybrid optical neural networks....[T]hey always require non-linear functions that are difficult to implement optically. Most proposed optical neural networks leave this non-linear function to be implemented electronically.” Nearly all efforts in optical neural nets are really efforts in optoelectronic neural nets, as the difficulties of working with bistable optical devices [?] have steered nearly all work to utilize optical-to-electrical conversion and electronic nonlinearities, with notable exceptions [?, 218]. Thus, optoelectronic apparatus implementing neural systems must have electronic components as well as optical components, including passives as well as sources and detectors. For scalability under economic con-

straints, such hardware capable of outperforming silicon microelectronic circuits must integrate these components monolithically, or at least with a co-packaged, chiplet approach. The necessary levels of optoelectronic integration may just now be emerging as viable for the described task, driven by the digital computing applications of silicon photonics discussed in Sec. 3.4.2. However, if such silicon photonics approaches are going to be employed, they will little resemble the hardware of free-space neural nets, as discussed below in Sec. ???. There is one general class of reasons to expect that no free-space optical approach will gain widespread traction: free space optical components and systems are bulky and difficult to package, making fieldability expensive. The bulk nature makes it is difficult to engineer entire systems that are simply “plug-and-play” and do not require experience with optics to operate. Thus, scaling to large systems is problematic, as planar, many-step lithography cannot be used to construct many of the complex elements.

Specifically related to the subject of this article, new considerations arise when we consider hardware not for deep learning, but for active cognition. In this context, highly complex, recurrent and adaptive networks are required, as discussed in Sec. ???. In terms of achieving the desired graph structures, free-space routing only moves light in straight lines, and it is difficult to construct modular and heirarchical recurrent graphs with interconnectivity required for cognition. For free-space optics, everything is most straightforwardly laid out in a series of planes, with light propagating between adjacent planes. Using light in this manner has the advantage that “multiple beams of light can pass through lenses or prisms and still remain separate,” [21] yet routing of free-space optical signals brings new challenges for complex neural systems. To achieve connectivity graphs corresponding to neural networks with feed-forward, feed-back, and recurrent connections, light cannot travel only in straight lines, but rather must branch and change direction many times, often being routed around obstacles formed by other parts of the network. Construction of complex networks with free-space optics, mirrors, and lenses quickly leads to issues related to scaling. Integrated photonic devices are likely much better suited to form the lateral and recurrent connections that are dominant in many regions of the brain. We will discuss in Sec. ??? an architecture that combines integrated-photonic lateral connections with free-space feed-forward and feed-back communication as well as communication over optical fibers for long-haul signaling.

Importantly, of this extensive body of literature regarding free-space optical neural networks, none relate to spiking neurons. As we have emphasized, the temporal dynamics and information processing in spiking neural networks is central to the mechanisms that underlie cognition. Similarly, all the approaches to optoelectronic neural systems discussed thus far implement only passive synapses and dendrites. Cognitive hardware re-

quires much more complex functionality, including short- and long-term plasticity mechanisms as well as dendritic processing that are central to the manner in which spiking neurons adapt in time. These functions are best achieved with electronic components. The pursuit of sophisticated optical neural systems therefore intertwines with the exploration of neuromorphic electronics that is a stronger candidate for implementing computational functions, learning, and memory. Learning in free-space optical setups is not straightforward and almost always requires a supervised training algorithm with external computer/user control. Implementing learning through local plasticity operations, as is required for large-scale cognitive systems, is more tractable with electronic circuits.

Finally, to my knowledge, all optical neural systems except the one our group has proposed encode synaptic weights in optical amplitude or intensity. In this way, some portion of the computation occurs in the optical domain. As we will argue in Sec. ??, this optical weighting results in energy and noise penalties. Instead, I will argue next that the proper role of light in neural systems is for communication only, and monolithic optoelectronic integration is required for complexity. Optical communication is likely to be carried out over integrated-photonics waveguides at the local scale, with free-space connections when suitable, and long-range communication over fiber optics. Before transitioning to a description of superconducting optoelectronic networks, let us first consider other contemporary approaches to machine learning with integrated photonics.

### 3.4.4 Machine learning with integrated photonics

We have reviewed the history of optical logic devices, the rise of silicon photonics, and we have discussed many approaches to utilizing free-space optics for neural networks. We now consider more recent work that lies at the interface between integrated photonics, silicon photonics, and machine learning.

As we have emphasized, in many artificial neural networks, the behavior of a neuron is modeled by Eq. ?? . The linear portions of this operation reduce to matrix-vector multiplication. As noted above, matrix-vector multiplication has been a draw toward optics for some time [149, 194, 209], with the Stanford architecture appearing on page one of volume two of *Optics Letters* [149]. Other approaches to matrix-vector multiplication have emerged over the years as technology has evolved, and one approach that is currently receiving attention is based on the concept introduced by Reck et al. in 1994 wherein a unitary matrix is implemented with an array of Mach-Zehnder interferometers [219]. With the addition of loss or gain, any matrix can be represented through its singular-value decomposition [220]. Because additional MZIs can be used to discard light, the full singular-value decomposition and matrix-vector multiplication can be performed using MZIs. Implementation of such MZI networks with silicon

photonic waveguides is currently being pursued [221, 222], and it has been argued that efficient training through backpropagation can be implemented in optoelectronic integrated circuits with tunable MZIs and photodetectors working on conjunction with CMOS logic at each interferometer [223].

In 2017, Shen et al. demonstrate the use of a network of Mach-Zehnder interferometers for classifying spoken digits [222]. The network was configured to accomplish the task through the use of thermo-optic phase shifters perturbing the index of silicon waveguides. A network with four inputs and outputs was trained to classify four vowel sounds. The thermo-optic phase shifters used in this system required 10 mW per phase, and the number of phase settings required for a single-layer, feed-forward system scales as  $N^2/2$  [219]. This power consumption is far too high to be competitive with other technologies for deep learning. However, other means of acquiring index perturbations may prove useful, such as non-volatile phase-change materials or microelectromechanical systems. In 2019, Fang et al. explored the implementation of a similar matrix-vector multiplier based on Mach-Zehnder interferometers wherein the phases were fixed at the time of fabrication based on optimization implemented in simulation. Such an approach is similar to the work of Ref. [216] discussed above wherein diffraction gratings were designed in simulation and then 3D-printed. This approach has the advantage of dissipating no power to maintain the phases, but the disadvantage that a system is purely static and is also susceptible to fabrication-induced errors that cannot be compensated during operation.

If one wishes to employ feed-forward neural networks based on Mach-Zehnder matrix-vector multiplication that can be dynamically reconfigured to adapt to changing inputs, a suitable training procedure must be devised. Such a procedure, based on backpropagation was introduced by Hughes et al. in the group of Fan in 2018. The technique utilizes photodetectors and reconfigurable phase shifters controlled by transistor logic circuits at each Mach-Zehnder. During training, the algorithm requires light to first propagate forward through the network followed by light propagation in the opposite direction. The results from these two passes are stored in the logic circuits at each interferometer, an error signal is calculated, and the phase settings are adjusted. Such a system has not been demonstrated, but may be possible with ongoing advances in monolithic optoelectronic integration. The concept is reminiscent of the approach to backpropagation explored by Wagner and Psaltis in 1987 [202]. In other work from Fan's group, Williamson et al showed in simulations that such hardware can achieve a variety of nonlinear activation functions [224]. The experimental results so far [222] have made use of nonlinearities applied in software, and other approaches to nonlinearities have been proposed, usually based on optical nonlinearities such as in saturable absorbers that require high power and are difficult to fabricate at scale in the context of silicon pho-

tonics. The simulations of [224] indicate that the required neuronal nonlinearities may be implemented through photodetection and electronic reconfiguration of phase settings in interferometric devices. Hybrid devices utilizing a nonlinearity implemented with an electro-optic effect and electrical feedback have been studied since the early 1980s [198, 225], but in early work primary attention was given to compound semiconductors, while silicon photonic implementations in conjunction with CMOS electronics are presently at the forefront. As the authors of Ref. [224] state, the nonlinearity is achieved by “converting a small portion of the original optical signal into an electrical voltage. The remaining portion of the original signal is phase modulated by this voltage as the signal passes through an interferometer. For an input signal with amplitude  $z$ , the resulting nonlinear optical activation,  $f(z)$ , is a result of the responses of the interferometer under modulation as well as the components in the electrical signal pathway.” The nonlinearity is applied to an optical signal, but it is accomplished with electrical circuits, keeping with the assertions of Ref. [199] that nonlinearities are best applied electronically, and only made conceivable in practice through decades of progress in silicon photonic optoelectronic integration. The nonlinearity is intended to be applied independently at each node of the neural network. Each node utilizes optical-to-electrical conversion with a waveguide-integrated photodetector, which converts optical current to electrical current, and a transimpedance amplifying stage, which converts electrical current to voltage. This voltage “then passes through a nonlinear signal conditioner with a transfer function  $H(\cdot)$ .” [224] Implementing nonlinearity in this way leverages the strengths of electronics as well as optics. However, one motivation for optical matrix-vector multiplication—the light-speed-limited latency—is somewhat compromised. In this scheme, the electronic circuits need time to respond to detection and accomplish phase reconfiguration when new inputs are present. Reference 224 proposes to pass the optical signal to be modified through a delay line during this time. The electronic circuits can be quite fast, but nevertheless, this is an important conceptual consequence of the utilization of optoelectronic integration.

In contemplating the recent work from Fan’s group [223, 224], it is important to distinguish between interferometers comprising the matrix-vector multiplier and interferometers achieving the nonlinearities at present only at the nodes of each network layer. For the training scheme envisioned in Ref. [223], photodetection and electrical reconfiguration of phase is required at each interferometer of the matrix-vector multiplier, of which there are  $\mathcal{O}(N^2)$ . These phase shifters dominate power and space consumption, as discussed in the next paragraph. By contrast, for the implementation of nonlinear activation functions envisioned in Ref. [224], only the output nodes of each network layer are involved, of which there are  $\mathcal{O}(N)$ . More power and space can be dedicated to these components because the scale as the number of neurons, not the num-

ber of synapses. The ideas presented in Refs. 223 and 224 are some of the most compelling related to the use of contemporary optoelectronic integration capabilities to serve the functions of deep learning. Still, several features remain problematic: interferometers are used for routing between network nodes, so significant space is consumed; a number of phases proportional to the number of synapses must be controlled to establish the synaptic weights; and the use of a mesh of interferometers likely requires confinement of the routing infrastructure to a single plane. These challenges make lead to significant consumption of power and space and make it unclear whether such a technology will prove superior to silicon microelectronic approaches to deep learning, as we will now discuss.

Like its free-space counterparts before, matrix-vector multiplication with Mach-Zehnder meshes is intriguing for applications in feed-forward neural networks. However, there are reasons to be skeptical about whether it will outperform approaches based on customized silicon microelectronic circuits, such as tensor-processing units (see Sec. 3.3). If power problems are addressed with static networks, an important part of what makes neural systems interesting—the ability to learn—is discarded. If adaptability is maintained, significant advances are required to achieve phase shifts with low power consumption and small footprint. A Mach-Zehnder is comprised of two directional couplers with a phase modulator between, and each of these components has non-negligible length when implemented in an integrated photonic setting. For a Mach-Zehnder mesh performing matrix-vector multiplication on an input vector of length  $N$ , the array will have the width of  $N$  Mach-Zehnders and a length of at least  $N/2$  Mach-Zehnders for the unitary part of the transformation alone. Loss regions necessary for a singular-value decomposition add further length. In a Mach-Zehnder interferometer, acquiring a phase shift through an electrooptic effect is characterized by a  $V_\pi \cdot L$  product. Materials with non-vanishing linear electrooptic coefficient (Pockels effect) can achieve small  $V_\pi$  with no current, and therefore no static power dissipation, but  $L$  becomes quite large—on the order of centimeters. Employing the Pockels effect requires too much space for this application. On the other end of the spectrum, the thermo-optic effect can reduce  $L$  for a single Mach-Zehnder while maintaining modest  $V_{pi}$  on the order of a few volts. However, this approach drives current through a resistive element to produce a temperature shift through Joule heating.  $L$  can be reduced to the order of  $100\ \mu\text{m}$  (as was the case in Ref. ??), but at the cost of significant power dissipation. Even with  $L = 100\ \mu\text{m}$ , the length of a chip becomes nearly 4 cm for one feed-forward layer if  $N = 784$ , as would be the case for an introductory problem like MNIST. Additionally, use of the thermo-optic effect leads to significant cross talk unless each device is spatially isolated, leading to further area consumption. Phase shifts induced by free carriers, such as proposed by Soref, fare no better, as they require modest  $V_\pi$  and  $L$  much greater than with the thermoop-

tic effect. Even if phase shifts are static, written at the time of fabrication, Mach-Zehnders will occupy on the order of 100  $\mu\text{m}$ , and therefore do not address the spatial scaling challenges of this approach to matrix-vector multiplication. This challenge has led the two senior authors of Ref. [222] to return to free-space approaches to matrix-vector multiplication for deep learning, stating “the footprint of directional couplers and phase modulators makes scaling to large ( $N \geq 1000$ ) numbers of neurons very challenging.” Nevertheless, the lead authors of Ref. [222] are pursuing this approach as a commercial product through start-up companies.

Spatial scaling is difficult for any integrated photonic technology due to the large wavelength of light relative to electrons. The reason why Mach-Zehnder meshes are even larger than other potential routing technologies is because all waveguides are laid out in a single plane. It is much more promising to use photonic interconnects between neurons if multiple planes of waveguides can be employed [226–228], just like electrical interconnects would be unworkable if only a single plane of wires could be used. We will discuss multiplanar photonic interconnects in Sec. 3.7, but at present we conclude this discussion of matrix-vector multiplication with Mach-Zehnder meshes for deep learning by noting that spatial scaling is not the only formidable challenge of the approach. The fact that phase must be precisely controlled to establish the weight matrix. As Miller points out in considering optical logic devices, “...coherent interference is mostly a nuisance in logic systems” [169] For training, this requires detection and phase control at every interferometer, and phase must be coherently re-established after each network layer. Loss from propagation through this long network is convoluted with the intended phase-induced amplitude modulation, and results in an error in signal that must be compensated. Further, because this is a non-spiking neural network, light sources are always on, and if the intention is to perform deep learning in data centers, compound semiconductor lasers operating at room temperature are likely to be employed. Thus, co-packaged light sources are required, most likely at each network layer. Reference [224] has proposed nonlinear optical functions that do not require light sources at each network layer in principle, but in practice propagation losses through deep neural networks will become problematic if light levels are not restored. While this is an exciting opportunity for optoelectronic integration and technological exploration, it is not obvious such an approach will displace conventional silicon microelectronics for matrix-vector multiplication. The performance gains are likely to be modest at best, and if history is a guide, CMOS will not be displaced unless a competing technology can do something that CMOS simply cannot do.

Beyond considerations of power and spatial footprint, if we consider the application to cognitive systems rather than deep learning, the approach of using arrays of interferometers for routing and synaptic weighting is incom-

patible with large-scale cognitive systems for two primary reasons. First, as described in Sec. 2, 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 many synaptic weights. This is similar to the limitation of writing synaptic weights as gratings in a holographic medium: the synapses are not independent. While backpropagation can be implemented with such a network [223], STDP cannot. Second, this approach to routing is solely focused on achieving feed-forward graph structures and is not amenable to constructing recurrent neural networks with hierarchical modularity. Thus, matrix-vector multiplication via networks of interferometers is not likely to be the primary means of establishing synaptic weighting nor is it likely to achieve the complex structural graphs required of cognitive systems demonstrating general intelligence.

[229] -WDM input and weighting scheme -logistic sigmoid activation function realized with a “deeply-saturated differentially-biased Semiconductor Optical Amplifier-Mach-Zehnder Interferometer (SOA-MZI) followed by a SOA-Cross-Gain-Modulation (XGM) gate.” -the goal seems to be a very close fit with a sigmoid, no matter the cost

Other efforts to utilize light for neuro-inspired machine learning are numerous. We now turn our attention away from feed-forward neural networks for deep learning and onto methods of using light for reservoir computing.

### 3.4.5 Photonic Reservoir Computing

We have summarized the concepts of reservoir computing in Sec. 3.1. Here we briefly summarize the state of the field of using light for reservoir computing. Reviews of the field of reservoir computing as a whole circa 2009 and 2019 can be found in Refs. 230 and 231, respectively. Review regarding reservoir computing with photonic delay systems was published by Van der Sande, Brunner, Soriano in 2017 [232]. There the authors state the primary motivation for reservoir computing: “much of the promise of photonic reservoirs lies in their minimal hardware requirements, a tremendous advantage over other hardware-intensive neural network models.” This advantage derives from the fact that connections within the reservoir are fixed, and only the output layer is trained. In the above discussion of using networks of Mach-Zehnder interferometers to accomplish the synaptic weight matrix we found that this network is very large in size due to the desire to control all synaptic weights. In that case, the network graph was feed forward, whereas in the case of reservoir computing, the reservoir is a recurrent graph with random, static connections. This changes the nature of considerations around connectivity between network nodes, generally easing the requirements on hardware.

In photonic reservoir computing, there are two classes

of approaches. In one class, the reservoir is comprised of multiple discrete optical nodes that each perform neuronal functions, whereas in the other class the reservoir is comprised of a single node, and the functions of different neurons are implemented serially in time. Elements of this latter class are referred to as *delay systems*. We first discuss the former class and subsequently turn our attention to delay systems.

A significant body of work regarding reservoir computing systems that use light with spatially separate neuronal nodes has been put forth from Bienstman and co-workers. In the first papers on the subject, Vandoorne et al. proposed to use semiconductor optical amplifiers as the basic building blocks of the reservoir to explore pattern-recognition problems [233, 234]. In this hardware implementation, the nonlinear dynamics results from interactions between photons and electrons in saturable semiconductor gain media. The motivations to use this hardware implementation were fourfold: 1) semiconductor optical amplifiers have gain, so no separate component is needed to compensate for loss; 2) the devices are broadband, so fabrication tolerances are relaxed as compared to resonators; 3) their steady state characteristic resembles the upper part of a sigmoidal activation function, which is commonly used as a nonlinearity in neural networks, see Sec. 3.1), so knowledge gained from software may be applicable; and 4) such devices can be implemented in an integrated photonic circuit. Common materials for semiconductor optical amplifiers include compound semiconductors based on GaAs or InP. Reference 233 presented numerical simulations of a reservoir of 25 semiconductor optical amplifiers with limited connectivity motivated by the desire to avoid waveguide crossings within a single plane. It was found that this simple network was capable of accomplishing standard pattern recognition tasks.

In systems making use of the carrier dynamics in semiconductor optical amplifiers or lasers, the memory time of integrating elements is set by the carrier lifetimes, which are in the range of 100 ps–300 ps. In Ref. [234], Vandoorne et al. compared in simulation the performance of  $9 \times 9$  networks of semiconductor optical amplifiers to idealized nodes with a *tanh* transfer function. That work investigated the sensitivity of the system to various design parameters and found that the system was most sensitive to the delays and phase shifts of the system’s physical connections. Such a finding is intuitive considering propagation delays in a waveguide with group index of four are on the order of 100 ps—the time constant of the memory elements—when propagating across a centimeter-scale chip. It is also reasonable to expect devices to be sensitive to phase errors when coherent signals are employed and input-output devices are phase-sensitive, as is the case with semiconductor optical amplifiers, reiterating Miller’s observation that phase is more often a nuisance than attribute in large-scale optical information processing. Detection mechanisms that are insensitive to phase are more likely to be useful in very large cognitive systems. Noise

from the semiconductor optical amplifiers was identified as problematic as well. Another interesting finding from this simulation comparing semiconductor optical amplifiers to *tanh* nodes was that “The result of the tanh network without leak rate at its optimal delay, is actually comparable to a tanh with leak rate and no (or minimal) delay. This reinforces the view that delay is an alternative approach to introducing memory next to leak rate.” While this finding is limited in scope, it points to an important factor present in neural systems employing light for communication. Because the velocity of light is so much faster than axonal conduction, small-scale networks of artificial neurons employing optical communication do not generally utilize communication delay as a parameter to be utilized in information processing. Generally speaking, this is a tremendous strength of optical communication, and this finding by Vandoorne et al. indicates the potential for networks employing synaptic integration with a diversity of time constants to achieve similar functionality to systems using a diversity of delays. We will revisit this concept in Sec. 3.6.

After the theoretical investigations just discussed [233, 234], Vandoorne, Bienstman, and co-workers turned their attention away from compound-semiconductor optical amplifiers as nodes of their reservoir and explored instead the use of free-carrier and thermal dynamics in silicon microrings to implement nonlinearities within a reservoir. In this concept, input optical signals are integrated within a microresonator with integration time set by the cavity  $Q$ , usually on the order of 100 ps, like the relaxation time of free carrier in semiconductor optical amplifiers. As the intensity within the cavity grows, two-photon absorption increases, leading to a greater population of free carriers, which leads to further absorption as well as to heating as the free carriers relax and generate phonons. As the cavity heats, its index shifts due to the thermo-optic effect, so the resonant frequency changes. The interaction between all these phenomena can lead to integration and bursting [235], forming a relaxation oscillator that may be likened to a spiking neuron. In theoretical analysis the authors considered free-carrier concentration and temperature as two dynamical variables to enable phase-plane analysis. These simulations guided the demonstration of cascable excitation between two silicon microrings coupled through a waveguide. The combined dynamical effects of optical intensity, two-photon absorption, free-carrier absorption, heating, cooling, and thermo-optic feedback make the system extremely difficult to model or control.

In subsequent work, Vandoorne, Bienstman and others utilized a silicon photonics chip in a purely linear, passive manner to pursue reservoir computing [236]. This work studied a  $4 \times 4$  mesh with each node connected to two other nodes in a swirl topology to avoid crossings. The connections were made through photonic delay stages achieving 280 ps delay to reduce time scales to an experimentally accessible range. Each node could input and output light through grating couplers with off-chip sources cou-

pled through optical fiber. Readout weights of the reservoir were implemented in software and were trained to implement Boolean logic functions with memory between bits at different temporal locations within a bit stream. For example, 2-bit XOR with one bit from the current time step and one bit from the previous time step was demonstrated. The various delay times incurred by different propagation pathways through the network made these delayed logical operations possible, and to perform XOR with bits with different delays, one must re-train the readout. Such a purely passive reservoir is not an end in itself, and in more recent theoretical work Denis-Le Coarer et al. combined the ideas of the free-carrier/thermo-optic nonlinear resonator dynamics with the swirl-topology reservoir, again seeking to implement delayed XOR [237]. These simulations considered a  $4 \times 4$  grid of nodes with a ring resonator at each node. In simulation, the system can perform the delayed XOR task at 20 Gbps with error rates lower than  $10^{-3}$  using 2.5 mW of injected optical power.

In addition to these numerical and experimental studies related to using photonics for implementation of reservoir computing with arrays of discrete nodes, a significant body of work exists in the domain of using photonic systems to implement time-delay reservoir systems wherein each node is represented by activity at discrete times in a temporal train. In delay systems time is used to emulate space. This makes efficient use of hardware in that very few nodes can behave as many. We present a brief summary of the field here, and the 2018 tutorial article by Brunner et al. contains a comprehensive overview and bibliography [134]. The delay concept is summarized in Fig. ?? (use Fig. 1 of apso2011), adapted from Ref. [238]. As stated in Ref. [134], “Reservoir computers offer a compromise between performance and an implementation-friendly ANN topology.” Specifically in the case of implementation of reservoir computers with delay systems, a signal circulates in a delay loop with total round-trip time  $\tau$  (adopting notation from Ref. [238]). This total time is discretized into  $N$  temporal intervals of fixed duration  $\theta = \tau/N$ . These temporal intervals are referred to as “virtual nodes”. The signals input and output to and from the system during such an interval are inputs to and outputs from that virtual node associated with that time slot. The interactions between nodes depend on the temporal aspects of the various components of the circuit, and they are controlled by the temporal delay between two nodes,  $\theta$ . With  $\theta$  very long relative to optical or electronic time constants, there is no interaction. With  $\theta$  less than but on the order of these time constants, nearby nodes interact. The authors of Ref. 134 further explain, “[S]implicity is bought in expense of time multiplexing and de-multiplexing in the input and readout layer.” The expense can be quantified: “[T]emporal multiplexing results in a reduction of the system’s overall processing bandwidth by the number of neurons  $N$ .” To increase the effective dimensionality of the system, the total delay time must be increased to allow inclusion of

information from a larger number of virtual nodes. This reduces the speed with which a given computation can occur. It is also challenging to construct systems in which two distant nodes may interact without introducing interactions with all nodes in between. Such systems are useful in situations where power and hardware resources are at a premium.

Many results have been presented by the group of Fischer [134, 214, 238–243] as well as by the group of Massar [244–246]. The first demonstration of this use of temporal delay in reservoir computing by Appeltant et al. occurred entirely in the electronic domain [238]. In that work, the authors defined delay systems as “Nonlinear systems with delayed feedback and/or delayed coupling.” Their delay system, implementing 400 nodes with a combination of bulk analog and digital electronics, was used to perform spoken digit recognition and as a model of a dynamical system in a benchmark task. In time-delay reservoir systems, training is most commonly conducted with standard techniques discussed in the review articles referenced above, employing supervisory control and post-processing with an external computer. Backpropagation through time can also be used to train the readout layer, as discussed in Ref. 134.

Following on the work by Appeltant, two concurrent experiments from different groups demonstrated the use of light in a delay system for reservoir computing [239, 244]. References 244 and 239 employed similar concepts to the work of Ref. 238 but the signal circulating through the delay system was now optical. The Architecture of the dynamical node is shown in Fig. ?? (use fig 2 of laso2012)(consider using Figs. 9,10 of [232] to summarize delay systems). An input laser was modulated by a standard LiNbO<sub>3</sub> optical modulator to provide a nonlinear transfer function ( $\sin^2$ ). The delay loop is based on an optical fiber spool. The optical signal is converted to the electrical domain on a photodiode. Within the electrical circuitry, multiple functions are performed. Low pass filtering occurs, information regarding inputs is added to the delayed signal, and amplification is introduced before the electrical output is used to drive the optical modulator, thus closing the optoelectronic feedback loop. Reference 239 showed the use of the optoelectronic delay system to implement 400 virtual nodes accomplishing a spoken digit recognition task as well as a one-time-step prediction task leveraging fading memory. Reference 244 implemented 50 nodes and also considered a spoken digit recognition as well as the simulation of a nonlinear auto regressive moving average equation and the equalization of a nonlinear channel.

In subsequent work by Duport et al. in 2012, a similar photonic delay system was utilized, except the electronic feedback was replaced by a semiconductor optical amplifier, enabling use of the term “all-optical” reservoir computing [245]. Similarly, work by Brunner et al. submitted in 2012 moved to utilization of a semiconductor laser cavity for nonlinearity rather than use electrical nonlinearity

ties [242]. These approaches eliminate slower electronic circuitry, enabling high-speed information processing. For example, Ref. 242 showed chaotic time-series prediction at a data rate beyond one gigabyte per second. Reference 243 explored the system with delayed optical feedback into a semiconductor laser in more detail, directly connecting “physical properties of induced dynamical responses to the system’s performance in a non-linear prediction task.” This work explored operating conditions enabling high consistency of dynamical response to identical inputs while maintaining memory retention, identifying an operating point achieving a compromise between consistency and memory.

Other work in this area identified the theoretical continuity from extreme learning machines to echo state networks and showed that both can be implemented with a single hardware platform, switching between the two by activating or deactivating a single physical connection [240]. Finally, as a concluding illustration of the potential and applications of reservoir computing based on photonic delay systems, the work of Antonik et al. in 2017 investigated the ability of such systems to achieve long-term prediction of time series by feeding the output from a reservoir back to itself. As the authors state, “In previous experiments, the output was uncoupled from the system and, in most cases, simply computed off-line on a postprocessing computer. However, numerical investigations have shown that feeding the output back into the reservoir opens the possibility of long-horizon time-series forecasting.” Application of this work are to understand the means by which biological circuits generate time series as well as to enable technological generation of time series with photonic reservoir computing. In this work, the readout utilized a field-programmable gate array. “The use of hi-speed dedicated electronics makes it possible to compute the output signal in real time and feed it back into the reservoir.” The general architecture was similar to that employed in Refs. 238, 239, 244. This work marks a new direction and potentially new application spaces for photonic delay systems.

Based on this brief summary of photonic delay systems for reservoir computing, can we gain insight toward our objective of designing hardware for general intelligence? In this case, the primary objective of reservoir computing is to simplify the network by reducing the number of degrees of freedom to be controlled by only training the readout layer. Further, in delay systems, the objective is to simplify hardware by utilizing only a single physical node to implement many virtual nodes. Nobody in this field is attempting to construct systems with general intelligence. Quite the opposite, the goal is to identify hardware that is simple to configure and capable of specific computational tasks. Still, it is useful to consider the specific reasons why time-delay systems are poorly matched to the functions underlying general intelligence discussed in Sec. ??.

Time-delay systems are an approach to reservoir computing that minimize the demands on hardware by extend-

ing computation across time rather than across space. The state of a node during a time window in a delay system corresponds to the state of a node at a spatial location in a neural network. Such an approach is appealing in the context of emerging hardware because a system comprising only a single node can perform a useful calculation, albeit with a slowdown due to time multiplexing proportional to the number of emulated nodes. In Ref. [134], the authors emphasize that in a delay system, “[A]ll nonlinear transformations carried out by the virtual spatiotemporal network rely on the same physical component.” The is the primary limitation to achieving complexity in such systems. The single node of the system quickly becomes a bottleneck, as all spatial degrees of freedom are represented as temporal degrees of freedom.

This delay technique is a form of time multiplexing, and as such scaling results in latency that scales with the number of virtual nodes. The time allocated to each virtual node is set by the characteristic time of relaxation oscillations in the physical system employed. For carrier dynamics in semiconductor optical devices, this is on the order of 200 ps. One can simulate roughly 5,000 neurons in a microsecond, 5 million neurons in a millisecond, and the scale of cortex would require 10 seconds, meaning information from each node in the entire network cannot be accessed more frequently than this. Beyond the scale of a few thousand nodes, other hardware approaches are likely to have superior performance. This consideration summarizes the speed limitation resulting from time multiplexing on a single high-speed node, but one may argue that information exchange across a cognitive system within 10 seconds is not too bad. Due to limitations on the network architectures that can be achieved, simply sampling each virtual node during each delay cycle does not mean each virtual node has an opportunity to communicate with all other virtual nodes during each cycle. Network average path lengths, which in this case refer to the number of delay cycles that must take place for one node to communicate to another node, become very long. Long path lengths are not surprising in this context, considering the approach collapses all spatial degrees of freedom to a single temporal dimension. The nature of the interactions between virtual nodes in such systems is that only nodes close in time can interact with one another, and two distant nodes cannot have strong interactions without even stronger interactions with the nodes between. This assignment of space to time is not conducive to the fractal use of space and time associated with cognition. New concepts are required if small-world or modular, hierarchical networks are to be utilized. It is likely these new concepts will require more nodes with more complex spatial configuration.

[247]

[248]

### 3.4.6 Spiking Neurons with Semiconductor Lasers

-separate by laser type, VCSELs, semiconductor ring lasers, fiber lasers -note that these experiments are not included in the section on free-space implementations, but many of them use free-space optics -the point of this section is specifically on the use of semiconductor lasers to produce pulses; if free-space optics are used, the same weaknesses apply

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. The equation governing lasers with gain and saturable absorber regions are isomorphic to the leaky integrate-and-fire neuron [249], 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. [218] and reference therein for a complete bibliography) 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 group of Prucnal 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 among neuronal firing events are forgotten rapidly, and the inability to broadly tune this time scale points to the potential for advantages to be gained from optoelectronic integration.

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. 2, 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.

[250] -Colack et al. -competition between modes in a laser cavity -“A new optical neural-network concept using the control of the modes of an injection laser by external feedback is described by a simple laser model. This approach uses the wavelength dispersed longitudinal modes of the laser as neurons and the amount of external feedback as connection weights.” -Simulations and experiments with GaAlAs injection lasers -“The inputs and connection weights to this laser neural network are provided by external masks which control the amount of feedback reaching the laser.” -stochastic learning employed -Winner-take-all and XOR operations are performed, depending on the mask -very high speed operation, measured in giga connections per second -“In this report, we focus on utilizing the longitudinal modes and obtain the desired nonlinearities in the frequency domain in a single laser, rather than the spatial domain.” -Summation and a sigmoidal activation function are modeled. In this case, the sigmoidal activation function provides a nonlinear response of the number of photons in mode  $m$  as a function of the output power reflected back into cavity mode  $m$ . -The sigmoidal saturation is due to competition with other modes and is controlled with external masks. -This is in the steady state, not a model of spiking neurons

[251] -Mos et al. -competition between modes in a laser cavity

[252] -Mos et al. -All-optical neuron -multimode laser diode -inputs are applied and threshold is achieved in the optical domain, they cite Jutamulia [199] regarding non-linearity applied in electrical domain and consider optical thresholding one of the strengths of this work. -The shape of the threshold function is adjusted through external feedback -One of the two cavity modes corresponds to the output of the neuron, and when light is injected at this wavelength, it is an excitatory input -The other cavity mode competes for gain, so light injected at this wavelength accomplishes inhibition -Free-space optics are used -Envision wavelength conversion to turn laser output to inhibitory signal -Envision weighted interconnection between lasers using a free-space optical matrix vector multiplier and they cite Goodman 1978, Stanford architecture [149] -Again, amplitude of optical signal is weight -Also suggest a number of wavelength converters, pointing toward WDM inputs with the various wavelengths competing for gain: “Due to mode competition only the wavelength with the highest amount of summed excitatory input will lase, suppressing laser action at any other wavelength.” Similar work on winner-take-all competition in multimode laser cavities in [250, 251]

[253] -Hoppensteadt and Izhikevich -Consider oscilla-



tory neurons communicating via phases rather than amplitudes -“Memorized patterns correspond to synchronized states where the neurons oscillate with equal frequencies and with prescribed phase relations.” -“...motivated by the theoretical and experimental observations that cortical neurons are sensitive to the fine temporal structure (timing or phase) of the incoming pulse train.” Such an idea would eventually lead to optical spiking neurons - theoretical study -assuming holographic interconnection medium

[254] -Romariz and Wagner -companion paper to rowa2007b -good introduction to motivations for spike coding -“An optical approach with a potential impact on this field of research has to implement important dynamic aspects of these pulsing models in a simple setup, with the capacity for integration in large arrays. The pulsing optical neurons should easily combine optical and electrical input signals. The necessary nonlinearities should neither require high-power optics nor complex electronic circuitry. And it should be able to utilize the other key advantages of optical approaches including massive parallelism and global interconnection capabilities.” -Neurons similar to FitzHugh-Nagumo model -“The system employs the wavelength tunability of a semiconductor laser source, the nonlinear response (in wavelength) of a birefringence-based filter, and simple linear electronic feedback. Surprisingly, nonlinear dynamical behavior can be achieved with only linear optical and electronic components by using this mechanism of wavelength tunability.” -consider using Fig. 1 -As opposed to utilizing nonlinear variation material properties with intensity, this work uses variation in laser wavelength. “The fundamental idea behind this approach is that the intensity output of a spectrally selective linear optical system can vary nonlinearly with wavelength. If this modulated intensity is detected and the photocurrent is used to drive the electronic system that modulates the laser wavelength, complex nonlinear dynamics can result.” -“The wavelength variable was previously explored in a static ANN implementation in the work by Mos et al. [?]. Specifically, the authors explore the fact that the amount of power emitted by a laser diode in a particular wavelength varies nonlinearly with the amount of optical power seeded to the cavity in the same wavelength. Our approach is different, though, in the use of external linear electronic feedback only, whose parameters are easier to control.” -“The implementation described here is based on the concept of birefringent phase delay and its variation with wavelength.” -“...establish a feedback loop by using the detected signal as the input to the circuit that produces the driving current  $i(t)$ .” Make comparison to using detected signal to drive MZI in photonic delay systems -Why don't we plan to make neurons this way? This approach with free-space optics and electrical feedback to bulk optical components is must less conducive to scaling than the approach with simple, pulsed, waveguide-integrated silicon LEDs and dense integration with basic superconducting electronic circuits. Our light levels are

lower by orders of magnitude, and our approach does not hinge on the control and tuning of resonant frequencies of many coupled optical cavities or emission wavelengths of many coupled lasers -use single-mode VCSEL, consider incorporating Fig. 5. Fig. 5 demonstrates why such an approach is difficult to scale. Goes back to the basic challenges of scaling free-space optical devices. -“In the regime explored here, the wavelength modulation mechanism is mainly thermal: current injection increases the temperature, and hence the index of refraction, as well as the cavity length through thermal expansion.”

[255] -Romariz and Wagner -companion paper to rowa2007a -LiNbO<sub>3</sub> is birefringent crystal -VCSEL emission demonstrates pulsing and bursting at 1.2 MHz seen in simulation and experiments. However, output amplitude is not constant, indicating spikes are not binary communication events, and neuron output amplitude is convoluted with synaptic weights. The best way to deal with this is to utilize detectors that have an amplitude-independent response, and this precludes encoding synaptic weight in optical amplitude. -This paper shows coupling between two such neurons, referred to as sending and receiving circuits. Challenges to mutual coupling of many such neurons are discussed. Some of these challenges can be understood from the following comments: “When optical coupling is allowed, the receiving circuit spikes are more frequent (but irregular, and not in one-to-one relationship to the sending circuit). Individual oscillator conditions are too different here to allow locking with weak coupling. The receiving circuit produced oscillations more easily, probably because of a smaller threshold of its VCSEL. A few adjustments in the optical feedback circuit were done to have similar dynamic conditions for both circuits. After that, the receiving circuit was biased to instability, and when coupling was established the waveform of the receiving circuits became much more regular.” These challenges point to problems with device variations, requirements for critical biasing, and related challenges that preclude scaling to billions of interconnected neurons, even before challenges with bulk optics are considered. Challenges with bulk optics are summarized in the comment by the author: “In combining more than one VCSEL, the most difficult part was guaranteeing the uniform collimation and alignment of all beams.” This challenge was encountered with coupling two such devices. Coupling billions with modular, hierarchical graphs structures is infeasible with free-space coupling using bulk optics. -Pulsing in 1-5 MHz range limited by thermal effects in VCSELs.

[256] -Hurtado et al. -Uses two different polarizations, one for excitation and one for inhibition -solve none of the problems of rowa2017a and rowa2017b

[257] -Rosenbluth et al., Prucnal last author -initial paper before krro2011 demonstrating some of the desired neuron functions

[?] -focused on enabling high speed, trying to simply go faster than VLSI approaches -follow on to [257] -based on the metric of speed alone, the authors claim these neu-

rons, “significantly outperform VLSI circuit-based neurons,” citing Ref. 258 -leaky integrate-and-fire model - “The entire setup is based on fiber optic components operating in the 1550 nm telecommunication band.” This work is an effort to employ concepts from fiber optic networks used in telecommunications for neural computing. While the hardware may be suited for the former application, it is not well suited to the latter. -all-optical thresholders -this approach has no potential for scaling due to fan in problems, power consumption, and general complexity of combining many components in fiber-optic apparatus, but the authors only seem to care about speed relative to CMOS neurons

[259] -three classes of excitability: class 1-the frequency of self-oscillations is zero at the transition (or bifurcation) point, class 2-oscillations are born with nonzero frequency, class 3-excitability is related to a specific transition to Q-switch-like pulsing through a homoclinic bifurcation in lasers with SA -focus on class III excitability “which arises as a competition between gain and saturable absorption in an incoherently pumped device.” -class III is particularly useful because pulses are very fast and coherence is not required -demonstration based on VCSEL (two InGaAs/AlGaAs quantum wells) grown with one more quantum well for a saturable absorber

[260] -Nahmias et al. (Shastri, Tait, Prucnal) -Design of new laser -leaky integrate-and-fire model -based on VCSEL with integrated saturable absorber -similar design to baku2011

[261] -Coomans et al -theoretical -semiconductor ring lasers -behavior very sensitive to phase of driving pulses relative to phase of field within cavity -two coupled rings simulated, phase and frequency sensitivity barriers to scaling

[?] -in short, utilization of cavity dynamics in semiconductor lasers does not offer the controllable functionality to be a promising computational primitive in neural systems. Yes, these devices do exhibit neuron-like nonlinearities and pulsing behavior on very fast timescales, but this does not mean they are scalable neural computing leading to cognitive systems. -they have high fan-out, but not high fan-in. for fan-in to a single mode device, all inputs must be coupled to that mode. -like their previous work ([256]), they use polarization switching in VCSELs -laser nonlinear dynamics result from carrier-photon coupling and lead to highly complex behaviors that are not easily controlled or tuned

[262] -Ghent/Bienstman -make comparison to pulsing si microrings used for reservoir computing -demonstrate class I excitability in optically injected microdisk lasers and propose a possible optical spiking neuron design -threshold and integrating behavior -optical phase control can be used to generate inhibitory response -input pulse power around 1uW for 0.2 ns for threshold -no transfer of excitation between disks demonstrated

[263] -Selmi et al. -“neuronlike excitable behavior in a micropillar laser with saturable absorber.” -“Under a dou-

ble pulsed excitation we study the absolute and relative refractory periods...and interpret the results in terms of a dynamical inhibition mediated by the carrier dynamics.” -emphasize the motivation is speed relative to biology and electronics -optically pumped micropillar laser with active and passive zones, passive zone for saturable absorber, InGaAs/AlGaAs quantum wells -pulse perturbation on the order of 10 nJ. this plays the role of a synapse event. 50 billion photons -as with all excitable laser systems, the device is biased to be close to lasing threshold, in this case with an optical pump. The excitable threshold disappears if the pump is too weak. -find an absolute refractory period of 150 ps and a relative refractive period out to 350 ps as the injected carrier concentration returns -there is not a straightforward means to adjust this refractory period without changing biasing conditions, i.e., refractory period cannot be tuned independently of threshold; it is determined by carrier physics in the semiconductor

[264] -Selmi et al. -Related to [263] -same experimental platform as [263] (micropillar laser) -Focus of this article is on temporal summation -“Temporal summation refers to the ability of the system to integrate different, potentially delayed, presynaptic stimuli, and to emit a spike if the integration of the inputs exceeds the excitable threshold.” -integration window determined by the time constant of the leak -“...we investigate the response of a micropillar laser with an integrated saturable absorber to subthreshold stimuli and show that the system can integrate the stimuli and emit an excitable spike if the stimuli are close enough in time.” -The micropillar laser is biased close to threshold, and two pulses (amplitudes of 74% and 80% of excitable threshold) are input with variable delay  $\delta$  -The integration time is on the order of hundreds of picoseconds, again set by carrier dynamics. There does not appear to be a way to tune this value to give longer integration times or different integration times at different neurons or synapses -both refraction and integration occur on few-hundred-picosecond timescale, and this cannot be chosen in design over a broad range -amplitude of response convolutes input amplitude and delay between two pulses -communication not binary

[265] -Hurtado and Javaloyes -1310 nm VCSELs -Multiple controllable spiking patterns are achieved in a 1310 in response to injection of two different polarizations -“reproducible spiking responses are demonstrated experimentally at sub-nanosecond speed resolution and with a controlled number of spikes fired.” -bursts are demonstrated, but no means is proposed for how to engineer receivers to be selective only to bursts with a given repetition rate, as is done in biological neurons based on short-term plasticity at synapses -different spike response patterns are generated based on external pumping with an optical signal with controlled temporal perturbations. -“We demonstrate that the properties of the achieved spiking responses (e.g., number of spikes fired and total temporal length of the generated pattern) can be controlled at will by acting on the characteristics of the perturbations

induced in the optically injected signal (e.g., intensity and duration).” -pump injection levels 45  $\mu$ W and higher in design, 60  $\mu$ W and higher in practice

[266] -InAs/InGaAs semiconductor quantum-dot passively mode-locked laser -extensive bibliography of similar neurons, but argue theirs is unique because it can implement inhibition ( [262] creates inhibition with phase) -2 mm Fabry-Perot cavity -“inhibition and excitation are associated with waveband switching effects triggered solely through optical injection from another optical neuron.” -the inhibitory neuron emits in an alternate band, which results in suppression of the activity in the excitatory band -sensitive to biasing conditions -will be extremely difficult to find biasing conditions where all neurons in a large network can work well together and be tuned to excite/inhibit appropriately -5V bias -optical isolator between two neurons to isolate signal from one neuron from reflecting back to itself (Keyes) -extremely fast 20GHz repetition frequency -high power 60 mW peak power -they want to avoid utilizing inhibition in the electronic domain because it is too complex -the scheme does not lend itself to dense integration and scaling, perhaps possible up to a few neurons

[267] -big review article by Prucnal et al -72 pages, you have not read this. likely redundant with book

[268] -photodiodes accomplish inhibition

[269] -Shastri et al. from Prucnal’s group -claim to be “the first unified, experimental demonstration of low-level spike processing functions in an optical platform.” -claim to demonstrate: logic-level restoration, cascadability, and input-output isolation -“We include a simulation model that explains all of the observed behaviors: integration, thresholding, refractoriness, and pulse generation.” -propose small, integrated device, but only demonstrate a fiber version -“Research advances in graphene micro-fabrication may make it a standard technology accessible in integrated laser platforms, which, together with a suitable networking platform, could lead to a scalable platform for optical computing.” -experimental system “based on a graphene fiber ring laser platform.” -“The fiber ring laser contains an erbium doped fiber amplifier (gain sections) and liquid exfoliated graphene (absorber section), interacting with one another in a fiber ring (cavity). The ring laser pulses periodically if driven above a threshold, modulated by the passive saturation of graphene absorption.” -“The integrated device contains electrically pumped quantum wells (gain section), two sheets of graphene (absorber section), and a distributed feedback-grating (section).” this last parentheses must have a typo in the paper. makes no sense -“In this design, we consider a hybrid silicon III-V laser platform in which the graphene layers are sandwiched in between the silicon and III-V layers. The hybrid III-V platform is highly scalable and amenable to both passive and active photonic integration.” It is not clear what supports this claim of scalability. -“Excitability is defined by three main criteria: (i) an unperturbed system rests at a stable equilibrium; (ii) a perturbation above

the excitability threshold triggers a large excursion from this equilibrium; and (iii) the system then settles back to the attractor in what is called the refractory period, after which the system can be excited again.” -“In this system, an excitatory pulse increases the carrier concentration within the gain region by an amount proportional to its energy (integrated power) through gain enhancement. Beyond some threshold excitation energy, the absorber is saturated, resulting in the release of a pulse. This is followed by a relative refractory period during which the arrival of a second excitatory pulse is unable to cause the laser to fire as the gain recovers.” -how is inhibition implemented? -single mode laser; how is optical fan-in implemented? -they acknowledge spiking is analog in time, but digital in amplitude, yet the amplitude of input spikes affects response, and thus encodes information -bursting with inter-spike intervals on the order of 10s of microseconds -The erbium-doped section is pumped continuously to keep it close to threshold. This is a primary reason for extraordinary power inefficiency, requiring hundreds of milliwatts even in the steady state ( [269], supplementary information and [218], pg. 101). -primary weaknesses: 1) laser requires biasing close to lasing threshold, constantly dissipating power, which makes neuron threshold depend on critical biasing and results in far too much power consumption; 2) fiber-based approach has no hope for scaling, while integrated approach requires graphene sandwiched between Si and III-V, as if just Si to III-V wasn’t hard enough; 3) if lasers are single mode, fan-in is impossible, and phases of inputs interfere; 4) spiking neurons aspire to have binary communication events, meaning the response of the receiving element should be independent of the amplitude of the input spike, but whether or not an excitable laser reaches threshold depends entirely on the amplitudes of the incoming communication events, and information is encoded in the height of pulses as well as their timing; 5) there does not appear to be a means of implementing inhibition; 6) because laser cavity is resonant, all neurons must be tuned.

[270] -application specific to a particular reinforcement learning problem -using laser chaos as a source of randomness that outperforms pseudorandom number generators

[265]

[241] -provide a means to couple many semiconductor lasers -8x8 array of single-mode VCSELs, square lattice, pitch of 250 $\mu$ m

-with any all-optical platform, communication and computation are based on optical signals. amplitude of communicated signals encodes weight. -in neuron designs utilizing optical nonlinearities in single-mode structures, all-optical fan-in can only be implemented with low loss if different frequencies are used for each input, and spectral add-drop multiplexing is employed. This leads to the limitations in connectivity we have discussed. For this reason, designs utilizing optical fan-in to single-mode devices can be ruled out as components in highly scalable neural architectures.

### 3.4.7 Phase-Change Materials for Synaptic Weighting and Neural Thresholding

One technique for establishing synaptic weights between neurons signaling with light is to leverage phase-change materials [271]. 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 [271] 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. [271] 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. The proposed route to full STDP utilizes a circulator and interferometer at each synapse and appears far too complicated to be implemented at each synapse in a large system.

[272] -Chakraborty et al. -proposes concept to use phase change material for spiking neuron -phase required to implement inhibition, so coherence is required and phases of all input synapses to a given neuron need to be precisely controlled relative to each other. simply unworkable in large systems -propose firing action based on a photonic amplifier, a circulator, and a waveguide with a phase-change element on top initially in crystalline (absorptive) phase -on a clock, the read and write phases for the integration unit and firing unit alternate in successive cycles -clock based operation eliminates nearly all information-processing benefits of spike-based processing in cognitive systems -neuronal firing must be accompanied by a reset pulse to reset the state of all devices, including the phase-change material, to their original state. it is essentially a latching device that a separate reset step

[273] -propose use of phase-change material for thresholding in optical neurons -seemingly the same as 2018 paper -basic idea is that a waveguide with a phase change material is initially opaque. light from synapses, if bright enough, can cause pcm to switch phase, becoming transparent. then the light from the synapses passes through, potentially with amplification

[?] -Feldmann et al. -Phase-change materials provide variable attenuation -This requires information to be encoded in optical amplitude, so communication cannot be binary, and neuron output amplitude is convoluted with synaptic weight -input pulses use WDM to fan in - Each neuron is a ring resonator “with its own integrated

PCM cell that can be switched...by incoming combined pulses,” thereby shifting the resonance and loss -Temporal duration of integration determined by thermal relaxation time of phase change material -requires precise frequency tuning of synaptic as well as neuronal elements, as resonance plays a role in synaptic fan-in (WDM) and neuronal thresholding -synaptic ring resonators don’t attenuate as in Tait’s concept, but rather just by absorption in PCM -all light coupled on and off chip with grating couplers, no source integration -SiN waveguides -four input wavelengths demonstrated here -if the combined input power is high enough to switch the neuronal PCM cell, the PCM becomes transparent, and the light is allowed to pass through -unlike a neuron, the amplitude of the output depends on the amplitude of the input. there is no gain in the neuron, and reaching threshold only means the inputs are allowed to pass, not that a separate pulse is generated. thus, there is no logic-level restoration - threshold is 430 pJ (three billion photons) -operation is clocked, and each neuron must be manually reset after each spike event -the clock rate was not given -operation on a clock completely misses the point of spiking operation -“Although individual PCM devices in endurance experiments have already shown  $10^{12}$  switching cycles [], further improvements in material design and device engineering are needed for high-speed and long-term switching operation.” -even if those improvements occur, these concepts are deeply flawed. -supervised and unsupervised learning -four-pixel inputs -neuron is operated in a clocked way -unsupervised learning approach can only weaken synaptic weights and requires synaptic inputs and neuronal outputs to overlap in time with no temporal correlation window, pointing to the benefits that could be gained by integration with electronic circuits -in a layered architecture, gain is added, in this case artificially through a grating coupler from an off-chip light source -each layer requires WDM MUX and DEMUX, so number of neurons in a layer is limited by DWDM -need one ring resonator per synapse -in layer demonstration, the collector is implemented off-chip using fiber based WDM components. -inhibition not mentioned

### 3.4.8 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 [274, 275]. 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.” [218] 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. 2. 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, then 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.

[276]

[277]

### 3.4.9 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 than 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 as 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.

### 3.4.10 Materials Considerations for Silicon Photonics

Silicon in the active material in CMOS transistor circuits, while SiO<sub>2</sub> is the primary electrical insulator. A third material—silicon nitride—is also important as an insulator,

etch stop, and interface layer in CMOS circuit fabrication. Most fortuitously, SiN is also extremely useful as a low-loss, passive waveguiding layer in silicon photonics. These three materials have an auspicious relationship to each other in the optical domain based on their indices of refraction. Near  $\lambda = 1550$  nm, the indices are (to two significant figures) 1.5, 2, and 3.5 for SiO<sub>2</sub>, SiN, and Si, respectively. The index contrast between Si and SiO<sub>2</sub> is quite high, enabling compact waveguide devices to be constructed with radius of curvature on the order of the wavelength. However, the primary loss mechanism in SOI waveguides is scattering due to sidewall roughness, and this loss scales as the index contrast square [278]. The index contrast between SiN and SiO<sub>2</sub> is still high enough to enable confinement and routing bends with radii of tens of microns, but this index contrast is significantly lower leading to much lower loss. Thus, the silicon photonics material platform, combining SiO<sub>2</sub>, SiN, and Si, provides an active layer (Si) that can be used for modulation, switching, and dense local interconnectivity, as well as a low-loss passive layer (SiN) that can be used for longer-distance routing, all embedded in a passivating SiO<sub>2</sub> matrix, and all integrated on a bulk-silicon handle wafer, most conducive to large-scale fabrication.

There is one thing missing from this discussion, and that is the source of light.

### 3.4.11 Where are the Light Sources?

**3.4.11.1 Considerations for Digital Communication** For digital communication it makes sense to modulate a CW laser because the photons are only discarded half the time. With neurons firing sparsely, tapping a CW light stream every time a neuron fires is wasteful. Almost all the light—which is the system’s most valuable resource—is simply thrown away. If many neurons are multiplexed on the same light source, they suffer cross talk, which becomes particularly problematic when the fire synchronously. By using light sources that only generate light when the neuron fires, photons are not wasted during quiescent periods. Because each neuron has its own emitter, cross talk does not occur.

In the case of silicon light sources based on point defects that are currently under investigation for this application, emission occurs in a sharp zero-phonon line with 0.3 nm bandwidth with x% emitted in broader phonon assisted sidebands with 10 nm bandwidth. All emitters are identical, so design of passive components is straightforward. Whereas the presence of the phonon sideband is problematic for quantum application requiring pure quantum states, light in the phonon sideband is still useful in this application and does not represent an impediment to operation. Most importantly, the utilization of silicon light sources enables scalable fabrication unlike any

other approach. These light sources are suitable for this application because the technology only requires incoherent pulses of light with 10 ns emitter lifetime and 1 % efficiency. Because single-photon detectors can be utilized, the emitters do not need to be particularly bright, generating only one to 10 photons per connection to compensate for propagation loss and Poisson noise. The sources are as simple as possible, as are the detectors, and the result is a highly scalable hardware platform tailored to this type of information processing.

**3.4.11.2 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 be 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 [190], 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 [190, 279–281].

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 [281]. 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 [282]. 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 con-

duction 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 ( $E_2 \rightarrow E_1$  and  $E_1 \rightarrow E_2$ ) are both fast, with few-picosecond time constants (check Davies). The electric-dipole transition ( $E_1 \rightarrow E_0$ ) is comparatively slower, with nanosecond to millisecond transitions depending on the specific defect []. In detection, the dipole transition ( $E_0 \rightarrow E_1$ ) 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  $E_0 \rightarrow E_2$  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  $E_2 \rightarrow E_1$  transition involves emission of a phonon, so it remains fast, while  $E_1 \rightarrow E_2$  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.

### 3.4.11.3 Hybrid Materials Integration [283]

**3.4.11.4 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 Sto-

janovic, 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 [?,184,189] 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 [?,137] 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 [137], 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.

Regarding the primary motivation for optical devices, Kogelnik wrote in 1981, “...the available speed is almost limitless, and it will be a challenge to exploit this speed.” [198]

#### 3.4.11.5 Silicon Light Sources Work at Low Temperature [284] [285]

When using light for computing in addition to communication in neural systems, information is often encoded in the amplitude of the light, and neurons often integrate the optical signal over time. There are multiple reasons why these uses of light are problematic: 1) encoding information in light amplitude is inefficient because it requires higher light levels to achieve dynamic range and overcome shot noise; 2) source noise is convoluted with synaptic weight; 3) integrating light intensity as the dynamical variable internal to the neuron (membrane potential) is limited because it is very difficult to store light in a cavity for longer than a few hundred picoseconds; 4) inhibition is difficult because light intensity can only be reduced with

an additional optical signal if the phase between the two signals is controlled, which is extremely difficult in large systems with many synaptic signals.

[?]  
[286]  
[287]  
[288]  
[289]  
[277]  
[257]

Did you emphasize that the Stanford architecture intended to utilize incoherent LEDs, as did the first optical neural nets by Psaltis et al?

### 3.5 Superconductor electronic neural systems

As with semiconductors and integrated photonics, it is also the case for superconducting circuits that much attention has been devoted to digital logic. 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 [290,291]. 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 [292]. 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.

#### 3.5.1 Josephson junctions

A JJ is created when two superconducting wires are separated by a thin tunneling barrier, as shown in Fig. ??(a). Excellent resources exist that describe the beautiful physics of JJs pedagogically [293–295], 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. ??(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  $I \leq I_c$  with exactly zero resistance and therefore zero voltage across the junction. However, if the current bias exceeds  $I_c$ , the fraction  $\Delta I = I - I_c$  cannot be carried through the superconducting 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. ??(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,  $\Phi_0$ , integral  $Vdt$

### 3.5.2 Superconducting digital logic

Here we briefly review the history of using superconductors in digital computing. More detail can be found in Ref. [296]. 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. It 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 [297]. 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 [298]. 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 [7], 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  $\text{AlO}_x$  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. [298] and [299] for technical details and Ref. [296] 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 [296,298,299] 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.” [1]. 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 [1]. 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 to the 4.2 K stage where computation is performed. Optical approaches are being developed to overcome this challenge [1], 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.

still need in this section:

- description of RQL
- description of AQFP, extreme energy efficiency, Landauer limit

recent work on basic science of MJJs: cama2018,kamu2018,bakl2018 search bib for MJJ

discussion of superconducting electronics for sensing/particle detection

Superconducting circuits can perform Boolean logic well enough to control networks of loop neurons and possibly qubits, but superconducting electronics are not poised to displace CMOS electronics for logic, particularly not in lightweight applications such as IoT and edge computing that have become so influential to the development of hardware for AI.

### 3.5.3 Neurons based on Josephson junctions

The development of superconducting digital electronics in the 1980s and 1990s was concurrent with the second wave of excitement regarding neural networks. Thus, several research groups developed circuits based on JJs to behave as neurons in ANNs, primarily in Japan. Following two conference proceedings in 1989 [300, 301], the first articles regarding JJ circuits for ANNs were published in 1991 [302, 303]. The objective of the circuits was to perform the weighted summation and thresholding operations required in the computational primitives of ANNs. The circuits employed were similar to those utilized in superconducting digital logic, as were the basic concepts, such as using an up-down counter to implement synaptic weights [303]. From the beginning, attention was paid to sculpting a sigmoidal transfer function to implement back-propagation as well as alternative circuits for achieving Hebbian-type learning [302]. The intrinsic threshold of a JJ upon being driven above its critical current was naturally used for thresholding [302], and mutual inductors were identified as promising for addition of synaptic signals and fan-in [303]. Little analysis was dedicated to anticipating scaling of such systems, although Ref. 303 claimed that achievable fan-out would be “sufficiently large to implement large scale neural networks,” and that “fan-in is essentially unlimited.” While the first assertion depends on one’s definition of “large scale”, subsequent analysis indicates that fan-out is a serious fundamental challenge for superconducting electronics, and fan-in reaches practical limitations due to the properties of mutual inductors, as we will discuss in more detail below.

Subsequent designs and experimental demonstrations were presented by Mizugaki et al. in 1994 [304, 305] and 1995 [306]. These circuits employed SQUIDS in various configurations for synaptic and neuronal responses. Again, digital concepts were employed, such as establishing the bit depth of possible synaptic weight values by adding additional DC SQUIDS. The focus remained on achieving feed-forward ANNs. Regarding fan-out, the intention was to use direct connections via superconducting wires to communicate between neurons. As stated in Ref. 304, “If it is necessary for a neuron circuit to drive a lot of another neurons, we can use biased JTLs...though the

use of JTLs might reduce the integration scale.”

In 1997 Rippert and Lomatch at Northwestern University proposed neural circuits using similar principles of SFQ pulse generation rate representing activation, but new innovations were introduced regarding learning [307]. The emphasis was still on ANNs rather than spiking neurons, and digital circuits were employed wherein higher-bit-depth synapses were achieved by adding junctions. This work began to explore Hebbian learning based on the temporal coincidence between two fluxons incident upon a JJ. In their circuit design, the coincidence window is determined by the temporal width of fluxons as well as the JJ bias current, and could achieve values of 1 ps to 5 ps. Reference 307 was the first to introduce a mechanism for metaplasticity in superconducting synapses wherein synaptic efficacy is updated through a Hebbian process, and the rate of Hebbian update is modified by additional circuits that adapt the Hebbian circuits. This work was before the term “metaplasticity” had become common in the neuroscience literature, but after the 1982 introduction of the BCM learning rule [87] that some consider the first example in computational neuroscience of a metaplastic synaptic mechanism [86].

Efforts in superconducting neural circuits continued in Japan through the 2000s and 2010s [308–313]. The emphasis remained on ANNs implemented with circuits originally designed for digital logic. References 308 and 309 continued to explore and demonstrate the approach with up/down counters to represent a membrane potential with the output rate of fluxons representing the neuron’s activation. Networks of these circuits were simulated solving a combinatorial optimization problem in Ref. 312, and more attention was paid to generating an accurate sigmoid function for back-propagation in Ref. 313. In 2013 in Italy, a small network of SQUID-based neurons was demonstrated and used to implement an XOR gate when trained through examples using an genetic algorithm in Ref. ? .

Throughout Refs. 302–309, 312, 313, a neuron’s activation was represented by the rate of production of fluxons, which becomes the time-averaged output current after low-pass filtering. By contrast, Refs. 310 and 311 proposed designs for leaky integrate-and-fire neurons that sum fluxons and produce a single flux quantum upon reaching threshold, much more in the spirit of single-flux-quantum digital electronics. To my knowledge, this was the first proposal to use superconducting circuits for spiking neurons. Input fluxons were stored in superconducting loops, and the number of input fluxons required to reach threshold was set in hardware by the number of JJ storage loops in the integrator. The leak rate was established by adding a resistance to one of the loop, giving exponential current decay with an  $L/r$  time constant. Fan-out was envisioned to occur through JTLs and splitters, while fan-in was envisioned to utilize confluence buffers. No scaling analysis regarding the possible number of connections was presented.

In 2010, Crotty, Schult and Segall at Colgate Univer-

sity proposed a different approach to integrate-and-fire neurons [314]. Rather than focusing on achieving weighted summation and nonlinear activation for ANNs, this work was oriented toward using JJ circuits to model neurons and their dynamics. Toward this end, they also adapted a circuit from superconducting digital electronics. The neuron circuit proposed in Ref. 314 is based on the DC-to-SFQ converter. Segall and his colleagues identified a correspondence between the behavior of each of the two junctions in the circuit and ionic currents across a neuron’s cell membrane, with one junction behaving like a  $\text{Na}^+$  current producing the rise of an action potential, and another junction behaving like a  $\text{K}^+$  current quenching the action potential and restoring the membrane potential to its resting level. It was estimated that a cortical column with  $10^4$  neurons could be simulated with JJ circuits on a single chip, but an interconnection scheme was not proposed. Subsequent work from the Colgate group simulated [315] and experimentally realized [316] synchronization dynamics of a two coupled neurons. In these neurons, a pulse consisting of a single flux quantum represents an action potential, and an RC circuit achieves synaptic leak, resulting in neuronal firing up to 25 GHz with  $10^{-17}$  J/spike.

While 10 aJ per action potential is low, some aspire to far lower operation. As mentioned in Sec. 3.5, there are multiple ways to use JJ circuits to implement digital logic. Correspondingly, there are multiple ways to use similar circuits to implement neural functions. While much of the work on superconducting neurons utilizes circuits most similar to SFQ logic circuits, an emerging branch utilizes AQFP circuits (see Sec. 3.5). The energy efficiency of AQFP circuits derives from the fact that the junctions involved are never driven above their critical current, and so never produce fluxons. Instead, the Josephson nonlinear inductance is employed to establish nonlinear current input/output relations. Utilization of adiabatic cells for ANNs was first introduced in 2016 by Schegolev et al. in Russia [317], and the ideas have been further developed in Refs. 318 and 319, where the authors proposed to utilize MJJs for synapses in conjunction with adiabatic neurons. Work from Katayama et al. in Japan [320] has developed related concepts flux-biased SQUIDs to achieve a sigmoidal transfer function suitable for back-propagation for training ANNs.

As mentioned in Sec. 3.5, Josephson junctions with a ferromagnetic material in the tunneling barrier can be used as a memory element in superconducting circuits [321]. The state of magnetic order can be used to tune the critical current across a broad range, and this can be used to steer a bias current. The use of MJJs as a synapse in superconducting neural circuits was first proposed in 2016 by Russek et al. [322] and has been subsequently demonstrated by Schneider et al. [323] in 2018. Further theoretical analysis of the use of MJJs for establishing synaptic weights in ANNs based on similar SQUID neurons to those discussed above [303] was presented Ref. [324], wherein simulation of nine-pixel image classification was

shown with 3 ns inference time per image.

While all superconducting electronic neural systems mentioned in the section have utilized circuits based on JJ and their associated nonlinearities and thresholding behavior, neuron circuits based on quantum phase-slip junctions (QPSJs) have also recently been proposed [325]. QPSJs are the dual to JJs, and thus corresponding dual circuits can be conceived to perform neural functions. QPSJs also have thresholding behavior and may offer further advantages in energy efficiency. These devices are still being developed, and one technical challenge is that one-dimensional superconducting wave functions must be achieved, which requires lithography down to about 10 nm, making initial demonstrations as well as future scaling potentially more difficult than with JJs.

### 3.5.4 Strengths and weaknesses of JJ circuits for neural operation

In Refs. 302–319, 322–325, five arguments have been forth regarding why superconducting circuits might outperform semiconducting circuits for neural computing: 1) JJs are faster than transistors, so JJ neurons and synapses will be faster than their silicon counterparts; 2) JJs consume less energy per operation than transistors, so superconducting neural systems will be more energy efficient; 3) JJs are highly nonlinear and have native thresholding and spiking behavior, so synaptic and neuronal circuits based on JJs can be implemented simply with few JJs; 4) superconducting transmission lines are lossless, so communication in superconducting neural systems will be superior to semiconducting neural systems using normal metal interconnects; and 5) due to their low power density, superconducting devices can be stacked in three dimensions with multiple planes of JJs, whereas the difficulty of removing heat from transistors precludes multiple active layers in a CMOS process. While I am a proponent of using JJ circuits for synaptic and neuronal functions, these arguments do not all hold up to scrutiny. As a broad comment, nowhere in Refs. 302–319, 322–325, is a system-level analysis of a large network presented. Neural systems are complex with many device and circuit interdependencies across various scales of the network. To analyze claims about full systems based on the performance of a device in an isolated context can be misleading. To further illustrate these issues, let us consider each of the above arguments.

*1) JJs are faster than transistors, so JJ neurons and synapses will be faster than their silicon counterparts.*

It is true that the intrinsic time constants of JJ are shorter than transistors. Yet the cutoff frequency of a transistor does not limit the speed of CMOS neural networks directly. Because CMOS neural networks have reached a sufficient stage of maturity, full, functional systems can be analyzed. We find that at the scale of 100,000 neurons, event rates are limited to 1 kHz per neuron [147], and by 100 million neurons it is limited to 10 Hz [133], and this

poor scaling results from the communication infrastructure, not the elemental device speed. It may be possible that large networks of JJ neurons are faster than large networks of transistor neurons, but we cannot conclude this until we know how communication will occur in these networks.

*2) JJs consume less energy per operation than transistors, so superconducting neural systems will be more energy efficient.*

At the device level, this argument seems well-founded, especially if adiabatic principles of operation are employed. Yet again we must think at the system level. The power required to operate a system at 4 K in a background at 300 K can be modeled as

$$P_{\text{tot}} = mP_{\text{dev}} + P_0, \quad (5)$$

where  $P_{\text{dev}}$  is the power dissipated by the devices comprising the system due to normal operation, and  $P_0$  is the power required simply to cool the devices below to the operating point, which is 4 K for most superconducting systems under consideration for this form of computation. This power is at least 100 W for contemporary cryogenic systems, and is often 1 kW, even when there are zero neurons in the system. The slope  $m$  represents watts of cooling power required to stay below 4 K per watt of device power dissipated due to operation. The theoretical limit is given by the Carnot efficiency, and when cooling from 300 K to 4 K this number is 150 watts per watt. In practice, it is often approximated as 1000 watts per watt. Thus, if we plan to operate the artificial neural system on earth, we should budget 1 kW of power just to turn it on, and we should budget an additional 1 kW of system power for each watt dissipated by the devices. This clearly limits the application spaces where superconducting systems are candidates to outperform semiconducting systems with regard to power consumption. Much of modern AI hardware aspires to play a role in small, deployable devices. Superconductors will never be useful in cell phones, the internet of things, or edge computing. Systems based on JJs may be more efficient than systems based on transistors, but this will only become relevant for systems large enough that the term  $P_0$  is tolerable. The domain of superconductors in neuromorphic supercomputing. High- $T_c$  materials are unlikely to change this situation because the materials involved make fabrication of dense, integrated circuits very difficult, and operating of JJs at higher temperature is very noisy, even if the underlying material has a high  $T_c$  [298] (check this ref).

*3) JJs are highly nonlinear and have native thresholding and spiking behavior, so synaptic and neuronal circuits based on JJs can be implemented simply with few JJs.*

To me, this is the most compelling argument in favor of superconducting neural systems. Josephson junctions are ideal for performing the synaptic, dendritic, and neuronal behaviors we seek. This is most apparent when imple-

menting spiking neurons that fully utilize the time domain. Most of the work to date in superconducting neural circuits has focused on generating nonlinear transfer functions, essentially for steady-state operation in feed-forward neural networks [302–309, 312, 313, 317–319, 322–325], but we will argue in Sec. ?? that JJ circuits similar to the spiking neurons of Segall and co-workers [314–316] can perform many desirable synaptic, dendritic, and neuronal functions going far beyond simple, point-neurons.

4) *Superconducting transmission lines are lossless, so communication in superconducting neural systems will be superior to semiconducting neural systems using normal metal interconnects.*

The dissipationless nature of superconducting wires is an extraordinary benefit to neural computing, digital logic, and quantum computing alike. But dissipationless transmission lines alone are not sufficient to enable an interconnection network, particularly when spiking is involved. The major problem with communication on metal wires is not resistance, but capacitance, as described in Sec. ?? . Likewise, superconducting interconnects must contend with problems related to inductance. If the output of a neuron is one or more fluxons generated by a JJ, the current associated with each fluxon will be  $\Phi_0/L$ , where  $L$  is the total inductance of the output lines being fed by the JJ. As more connections are added, the inductance gets larger, the current gets smaller, and the neuron eventually cannot source enough current to directly feed all of its synaptic connections. One solution to this problem is to use JTLs and pulse splitters to provide gain as a pulse propagates through and branches across the interconnection network. Other approaches to fan-out are discussed in Sec. 3.7.2, but it is likely that an active interconnection network of JTLs and splitters will be required to achieve communication from one spiking neuron to many synaptic connections. Active transmission lines dissipate power, and so arguments related to the lossless nature of superconducting transmission lines are not applicable. Again, one must propose an interconnection network adequate at the system level to assess the full benefits of lossless communication lines. It is not clear that routing of fluxonic signals from JJ neurons across large networks can be achieved without resorting to address-event representation, as is done by CMOS. If AER proves necessary, superconducting circuits will have much more difficult road ahead than CMOS, because AER places severe demands on memory, and superconducting circuits struggle with dense memory.

For certain feed-forward superconducting ANNs, particularly those based on AQFP circuits, it is not necessary for a neuron to produce an output that switches a junction, and passive transmission lines with high inductance may suffice. In this case, the high inductance will lead to reduced speed when changing the inputs to the network, so the interconnection network must be analyzed in conjunction with the inference latency, and speed/scaling

trade-offs will be present.

5) *Due to their low power density, superconducting devices can be stacked in three dimensions with multiple planes of JJs, whereas the difficulty of removing heat from transistors precludes multiple active layers in a CMOS process.*

This could be another tremendous advantage of superconducting circuits for long-term scaling. The power density of JJ circuits is low enough that many layers could be stacked while maintaining the ability to conduct heat to a bath of liquid helium, particularly when the circuits are participating in temporally sparse neural activity. However, in practice it is difficult to produce wafers with multiple planes of JJs [326], although there is plenty of room for development through future research. Three-dimensional integration of complex JJ neural circuits appears possible in principle, and it remains to be seen how far it can go in practice, but if successful could bring exceptional advantages for large-scale systems.

In this section we reviewed 30 years of work on superconducting neural circuits, and we described their strengths as well as their weaknesses.

- mina1994a anticipated problems with superconducting lines, JTLs, and scaling
- early works used digital concepts, adding active circuits to increase synaptic bit depths, whereas we now plan to use purely passive Lr loops, enabled by high-kinetic-inductance materials
- neuronal firing: mina1994a, mina1995, hagu1991 all use time-averaged output of SFQs from JJ as signal representing activation (not 1 fluxon action potential)
- advantages of superconductors for 3D integration seen in hiak1991
- temporal coincidence: compare/contrast with rilo1997 where use  $\Delta t$  of JJ. also metaplasticity

### 3.6 Superconducting Optoelectronic Systems

Superconducting optoelectronic networks were proposed by our group at NIST in 2017 based on two conjectures: 1) communication between neurons in hardware for AI is best achieved with photons; and 2) communication at the single-photon level reaches the energy efficiency limit, and therefore enables scaling to very large cognitive systems. These conjectures led us to propose specific circuits based on semiconductor light sources operating in conjunction with superconducting-nanowire single-photon detectors [327]. After additional consideration, it became apparent that the use of Josephson junctions in conjunction

with nanowire detectors enabled significantly more synaptic, dendritic, and neuronal computational functionality than nanowire circuits alone. The basic conjectures cited above remained in place, and the communication infrastructure described in Ref. 327 remained central to the concept, but a third conjecture began to guide our work: 3) In artificial neural systems utilizing superconducting single-photon detectors and communication with few-photon signals, circuits combining Josephson junctions with flux-storage loops are best equipped to provide the necessary short- and long-term synaptic plasticity functions, elaborate dendritic processing, and neuronal thresholding functionality. A summary of the circuits accomplishing these functions was first presented in Ref. 328, and more detail was given in Ref. 329. Dendritic processing was detailed in Ref. ?. This section summarizes our present thinking on the general concepts leading to the hypothesis that hardware combining superconducting electronics with silicon photonics and fiber optics is best equipped to realize artificial general intelligence, and a brief overview of the photonic and electronic circuits is given.

### 3.6.1 Basic Motivations

–“Hybrid optical neural networks may not materialize as commercial products unless each neuron in the network requires more than 1000 interconnections to solve problems.” [?]

- from the perspective of single photons for communication leading to spds, JJs, si light sources
- from the perspective of JJs for neuromorphic circuits, leading to single photon communication, spds, si light sources

### 3.6.2 computational circuits

- synapses
- dendrites
- soma

### 3.6.3 Communication Circuits

- axon hillock/transmitter circuit
- photonic circuits/axonal arbor
- wafer-scale architecture

### 3.6.4 Large-Scale Systems

- beyond wafer scale/wafers to columns
- a hierarchy of modules
- rentian scaling at wafer scale all the way up to many module systems enabled by multiplanar waveguides as well as fiber white matter, scaling of white to gray matter volume, estimate for largest size

- neuronal pool, velocity of communication
- power consumption: any competing technology must communicate at the velocity of light with very few photons across systems of comparable scale
- native environment in space

### 3.6.5 Misc. SOEN Notes

- [124]GaAs neurons: two transistors, a photodetector, and an LED

From Verber’s perspective in 1984, “It is to be expected that optical and integrated-optical methods will become available to perform a significant number of computational tasks. Analog methods are currently being developed and digital optical devices will follow in the near future. The challenge is to recognize their strengths and weaknesses in comparison to electronic devices and incorporate them into computational systems in an optimal fashion.” [194] However, optical devices have not become widely adopted for any computational tasks, digital or analog. This state of technology is not due to a lack of effort or ingenuity, but rather to a combination of physical and practical reasons [?]. However, optical and integrated-optical methods for communication are still quite promising at the chip [137] and package [] scales, and are indisputably superior to electrical communication at large scale.

[330]

These detectors are wires of superconducting material [331], and they can be straightforwardly patterned atop on-chip dielectric waveguides [?, ?, ?, 332]. 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  $\mu$ m. 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.

As Miller points out in considering optical logic devices, “...coherent interference is mostly a nuisance in logic systems. [A] device at the  $\sim 100$ -photon level could still operate in a quasi-classical fashion where optical or quantum coherence is neither necessary nor even desirable.” [169] By using few-photon optical signals for communication from neurons to single-photon synapses, this is precisely the regime in which we operate. By using light only for fan-out, optical coherence is irrelevant because independent fields from separate neurons never interfere.

Miller identifies two major benefits of optical interconnectivity: 1) that “higher densities of information [are] possible in relatively long connections”; and 2) “optics could reduce the energy required for communication”. [169] To his second point, I add optics can additionally reduce the complexity of communication hardware while increasing the speed. These considerations are of primary significance in the context of integrated neural systems. Further, while Miller envisions integration with semiconductor photodetectors and therefore concludes that optical signals “only need to transmit enough energy to charge the photodetector at the receiving end” (100 - 1000 photons), when superconducting detectors are employed, one need only transmit enough energy to disrupt the superconducting phase of a nanowire detector, and this can be achieved with a single photon.

-Like superconducting neural systems and free-space optical 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.

Criteria for practical devices [131, 169]

- Cascadability
- Fan-out
- Logic-level restoration
- Input/output isolation
- Absence of critical biasing
- Logic level independent of loss

some technologies place high speed as the primary goal, and sacrifice feasibility to attain that speed advantage. other technologies emphasize small size as a significant objective, and practical complications usually result. There are many applications where speed or size is paramount, but if they are not it is often easier to achieve general high performance if compromises can be negotiated between several metrics.

re detectors: compare  $CV^2/2$  with  $C = 1\text{ fF}$  and  $V = 1\text{ V}$  to  $LI^2/2$  with  $L = 250\text{ nH}$  and  $I = 10\text{ }\mu\text{A}$  [333]

a primary difference between circuits of normal conductors and semiconductors as compared to superconductors is that in the former one generally deals with capacitances, whereas in the latter one deals with inductances. With capacitance, parasitics involve the energy  $CV^2/2$  and the charging time (integral  $I\text{ dt} = Q = CV$ ). In the case of inductors, the energy is  $LI^2/2$  and the charging time

- 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  $\sqrt{N_{\text{ph}}}$ , where in this case  $N_{\text{ph}}$  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 attenuation, 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,  $N_{\text{ph}}$  photons to each of its downstream synaptic connections. Due to shot noise, each downstream connection will receive  $N_{\text{ph}} \pm \sqrt{N_{\text{ph}}}$  photons, and the detector circuit must be configured to implement a synaptic response if a threshold of  $N_{\text{th}}$  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 ( $N_{\text{th}}$ ) should be made as low as possible to make the job of the light source as easy as possible.  $N_{\text{th}}$  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)

At room temperature, the silicon photonics platform offers passive waveguides and active modulators, switches, and detectors, but not light sources. With the proposed approach to SOENs, we intend to utilize none of these features, instead using silicon primarily as a light source, while also leveraging the passive waveguiding properties of the family of materials. We also do away with silicon and SiGe photodetectors, opting instead for superconducting nanowires to receive faint optical signals with as few as one photon. By changing the operating temperature from 300 K to 4 K, and by changing the information to be communicated from bit streams to spike trains, the offerings of silicon photonics change dramatically. Even the purpose of integrating with silicon transistors is not the main goal of this use of silicon photonics. While the community has explored the potential of silicon light sources for digital communication at gigahertz speeds in a room-temperature context, little attention has been directed at using those same sources at 4 K for spike encoding in faint photonic signals up to 100 MHz. Yet such a light source could change the landscape of cognitive computing considerably.

Silicon microelectronics, silicon photonics, and superconducting optoelectronics are highly complementary technologies. If decades hence we find silicon present in our most advanced transistors, used as the backbone of silicon photonic communication at multiple levels of the digital hierarchy, and as a light source working in conjunction with superconducting circuits, we will have made remarkably efficient use of element 14 indeed.

### 3.6.6 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. 2, 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.

The binary nature of communication with these synapses is realized by the all-or-nothing response of an SNSPD [?].

The plateau of a WSi or MoSi SNSPD renders device insensitive to minor biasing variations, enabling scalable fabrication and operation, satisfying one of Keyes’ requirements. 1000 have been fabricated in an array for imaging with 99.7% yield. Leads into an associated discussion with consideration of exponential dependence of JJ tunneling barrier, sensitivity, difficult for digital logic to tolerate, neuro more tolerant because we seek a distribution of synaptic and dendritic weights, and all the crucial bias currents can adapt dynamically through various plasticity mechanisms to tune distributions and maintain functional operating points.

Each device is quite simple, but they can be combined to realize circuits with a wide variety of relevant neural functionality. Complexity arises from sophisticated arrangements of robust basic components.

Need to mention here that the Stanford architecture and first neural networks by Psaltis intended to utilize incoherent LEDs because they are simple, scalable, and do not require precise control of phase across many optical paths. The same is true here, with the added element that we do not wish to align the resonant frequencies of laser cavities with the gain bandwidth of silicon light emitters that may be narrow

### 3.7 Comparison of Interconnect Technologies

#### 3.7.1 Copper Interconnects

The majority of contemporary neuromorphic computing is based on silicon microelectronics, and for much of the community, the objective is not to answer the question, “What hardware is optimal for neuromorphic computing?” but rather to answer the question, “What forms of neuromorphic computing can I accomplish with the hardware that I have?”

The primary challenge for adapting digital hardware for neural applications is one is attempting to use a system with a certain graph structure and set of information processing operations to behave as a system with a completely different graph structure and a completely different set of information processing operations. The digital computing machine being utilized is Turing complete, so it can accomplish the task, but because the structures and operations are entirely unrelated, it is extremely inefficient.

In much of the literature, it is assumed that computer architecture will not change in future generations of hardware, so the questions become whether specific components will be replaced with alternative devices. One can consider whether photonic interconnects make sense in specific locations, such as for on-chip interconnects, between processor and DRAM modules on a common, between boards through back plane, or between servers in a rack. Each decision can be made separately.

include a figure regarding chip, board, back plane, server rack. perhaps Fig. 4 of [334]

The needs of neural information processing are precisely where conventional interconnection networks perform the worst.

Figure F.19 of [144] may be helpful. From the same reference, this expression for latency:

$$\begin{aligned} \text{Latency} = & \text{Sending overhead} \\ & + T_{\text{LinkProp}} \times (d + 1) + (T_r + T_a + T_s) \times d \\ & + \frac{(\text{Packet}) + (d \times \text{Header})}{\text{Bandwidth}} \\ & + \text{Receiving overhead} \end{aligned} \quad (6)$$

pg F-52. “Routing, arbitration, and switching can impact the packet latency of a loaded network by reducing the contention delay experienced by packets.” [144] “At higher applied loads, latency increases exponentially, and the network approaches its saturation point as it is unable to absorb the applied load, causing packets to queue up at their source nodes awaiting injection.” [144] This traffic/speed trade off means that the frequencies of synchronized oscillations that can be sustained by a neuronal ensemble depends on how many neurons are participating in the ensemble. But it is precisely the synchronized exchange of information across the network through simultaneous communication of many neurons in large, transient ensembles that is the essence of cognition, as we have explored in Sec. ???. If we have the liberty to design hardware for general intelligence from scratch, it would be unwise to build this communication bottleneck into the hardware. It is best avoided if each neuron has a dedicated axonal arbor that reaches all of its synaptic connections independently, and this fan-out is likely impossible to achieve with normal-metal interconnects. Can superconductors do better?

#### 3.7.2 Superconducting Interconnects

In designing neural systems

**3.7.2.1 Passive Superconducting Fan Out** Let us first consider the case where the output from each neuron does not need to switch a JJ at each receiving synapse, but rather the sum of multiple synaptic inputs to each neuron switches a neuronal thresholding junction. I refer to this as the passive superconducting fan out scenario. Suppose this thresholding junction is biased such that the net input must be equal to  $I_0$  in order to drive the JJ above  $I_c$ . Also suppose each neuron’s threshold JJ produces a current  $2I_0$  when it reaches threshold, consistent with the usual case of superconducting electronics that fan-out of two can be supported. Consider an ensemble of neurons in which each neuron makes  $n_s$  synaptic connections. For simplicity, make the favorable assumption that the output JJ can produce  $2I_0$  regardless of  $n_s$ . In practice, this will

be very difficult for large  $n_s$ , as the total inductance of the fan-out tree,  $L_T$ , will become large, and  $\Phi_0/L_T$  will become small. Also make the favorable assumption that the fan-out tree can evenly distribute current to all synapses, regardless of spatial location. Four problems are associated with this approach to interconnection. We discuss them here in order of severity.

Within this simplified passive fan out model, each synapse receives current equal to  $2I_0/n_s$ . We then ask: how many synapses must fire within an integration time to cause the neuron to switch? We have stated that the neuron requires  $I_0$  to switch, so  $n_s/2$  synapses must fire concurrently to drive the neuron to threshold. This identifies the first problem with passive electronic fan out between superconducting electronic neurons: order  $n_s$  synapses must all be active at once to drive a neuron to threshold. In biological neurons, this number is closer to order  $\sqrt{n_s}$  [1]. The problem with requiring  $\mathcal{O}(n_s)$  synapses to all fire at once is that only very strong stimulus leads to persistent activity. If a neuron is to fire, at least half of its input synapses must be active at a given moment, and if this condition is not met, activity dies out. As we have discussed in the context of neuronal avalanches, an important aspect of dynamical activity in neural systems is that any given neuron should have the potential to generate neuronal avalanches, but this cannot occur if half of each neuron's input synapses must be active simultaneously to perpetuate activity. We wish to be able to control and choose the number of synapses that drive a neuron to threshold. And, we wish to be able to make some of a neuron's output synapses strong without requiring that others become weaker. We can accomplish this if we have gain at each synapse, which occurs if each synapse has a JJ that switches with each synaptic firing event.

Another problem resulting from passive fan out relates to post-synaptic time constants. If the signal received by each synapse is simply the current produced by a neuron's output JJ, then the duration of the post-synaptic potential is the same as the duration of the action potential. This post-synaptic time constant is responsible for the integration time of the neuron, and it is desirable to be able to choose this value over a broad range across the synapses in the network. Some synapses can forget rapidly, while others must retain information over an extended period of time. If the post-synaptic time constant is the same as the duration of the action potential, the neuron does not perform integration, and all inputs must be precisely synchronized in time. Immediately above, we identified the problem that  $\mathcal{O}(n_s)$  synapses must fire to drive a neuron to threshold. If the synapses do not fire at precisely the same time, and instead experience jitter of even a fraction of the duration of the action potential, then even more synapses are required to fire, and it may become impossible to drive the neuron to threshold without extremely precise timing. In a system where a single flux-quantum represents an action potential, this requires timing precision on the order of 10 ps, which is extremely difficult

to maintain across a wide network of spatially distributed neurons. In some cases it may be possible to use this as a mechanism to induce precise synchronization, but it is not desirable to have it be a hard requirement at every neuron. In designing hardware for general intelligence, we must be able to engineer the action potential duration and synaptic time constants independently. We can accomplish this if we have gain at each synapse.

A third problem resulting from passive fan out relates to the distribution of current across an axonal tree in the case where neurons make a large number of synaptic connections. With passive superconducting wires, current is divided among the synaptic destinations by controlling the inductance of the paths to all the synapses. In several such circuit designs, the synaptic weight is represented by the amount of current reaching each synapse [2, 3, 323]. In such cases, the maximum synaptic weight is equal to the total amount of current reaching the synapse, and a variable circuit element can discard some of this current to achieve a lower synaptic weight. The total output current of a neuron's thresholding junction will be equal to  $\Phi_0/L_T$  based on fundamental Josephson physics. The tree inductance,  $L_T$ , is presumed to result primarily from the wiring carrying the currents. With standard Nb wires, the inductance per square is roughly 500 fH/ $\mu\text{m}^2$ . As networks grow in scale and connections become more distant, the current generated by a neuron's output JJ decreases. We are faced with a fan-out/current-supply trade-off. But we require exactly the opposite: as we add connections, we must supply more current to drive them. The inductance of the tree can be reduced by using wider wires, but then more area is consumed, and connections become further away. Such wire width/wire length trade-offs are not unique to superconductors and have been investigated thoroughly in the context of CMOS electronics [4].

Finally, a fourth problem arises when one attempts to utilize passive superconducting fan out to connect many neurons in a complex network. This challenge is related to achieving an even distribution of current across many synapses located across a network. If all connections are equidistant, they can use the same wire length, and because they are wired in parallel, the total inductance of the tree will be equal to the inductance of a single connection. However, as discussed in Sec. 2.1, small-world networks and modular, hierarchical networks require long-range connections to maintain short network path lengths and efficient cross-network communication. In the case of passive superconducting fan-out, these distant synaptic connections require the same amount of current as proximal connections, and therefore the wires making distant connections must be wider and thicker to achieve the same inductance. Because there is a maximum practical thickness for wires produced in a standard damascene process, this results in a scaling trend where the width of wires is linearly proportional to the distance the wire must cover. Long-range connections thus incur significant wiring area when passive superconducting lines are used. Related to

this problem is the challenge of devising routing and layout algorithms that place all the nodes and instantiate the fan out tree so as to equalize the inductance of all connections. These challenges become more serious as the scale of the network grows.

To summarize, there are at least four problems that arise if passive fan-out is to be utilized: 1)  $\mathcal{O}(n_s)$  synapses must fire simultaneously; 2) the post-synaptic time constant is same as action potential duration; 3) distributing currents across synapses requires inductive dividers, and as more synaptic connections are added, the current generated by each neuron decreases, making it more difficult to drive large numbers of synapses; and 4) with increasing numbers of neurons, the design and implementation of the axonal current distribution tree becomes significantly, and very wide wires with low inductance are required to ensure distant connections receive sufficient current.

**3.7.2.2 Active Superconducting Fan Out** In the case of active superconducting fan out, each synapse comprises at least one JJ that is driven above  $I_c$  when a synaptic firing event occurs. Such a scenario solves the problems identified above in the case of passive superconducting fan out, yet new challenges arise. In the active case, each neuron’s output JJ must switch a JJ at each receiving synapse. Again, we assume a current of  $I_0$  is required to do so. Consider two scenarios. First, the current from the neuron is split evenly by a passive, superconducting tree. A typical junction used for digital logic will produce 100  $\mu\text{A}$  upon switching, meaning the inductance of the tree must stay below 10 pH. Typical niobium wires have 500 fH per square and must be approximately 1  $\mu\text{m}$  wide to carry this current, so typical distances from a neuron to its synapses must be about 20  $\mu\text{m}$ .

### 3.7.3 Photonic Interconnects

Fan out is limited if electrons are used for communication. The basic problem is the same whether superconducting lines of normal conductor are utilized: if downstream devices must be driven by charge, there is a fan out problem. In the case of transistors, capacitors must be charged. In the superconducting domain, JJs must be driven above  $I_c$ . In both cases, it is impractical to source enough charge to drive many downstream elements, and fan out is limited. By contrast, with photons and single-photon detectors, the problem is tractable. If each downstream element requires  $\mathcal{O}(1)$  photon for activation, and the response is identical whether one or more photons is received, a single optical source can easily produce enough photons to drive thousands of synaptic elements. Ten thousand near-infrared photons contain a femtojoule of energy. Further, it appears possible to design and implement a passive photonic routing infrastructure that can route thousands of photons to thousands of destinations, even if those destinations are located in distant regions of a multi-module

network. All routing is passive, and no memory is required to keep track of destination addresses.

With photonic routing, distant connections still require wider wires (waveguides), but for a connection that is 10x further away, we may require a wire that is 2x the width to maintain the same loss (index contrast, Rayleigh scattering), whereas for passive superconducting we require 10x the width to maintain the same inductance (this compares signal amplitudes).

#### 3.7.3.1 Photonic Communication in Digital Systems

**3.7.3.2 Photonic Communication in Neural Systems** Optical interconnects have been considered for decades in the context of digital computing with processors and memory chips communicating over copper traces on circuit boards. In that context, many of the trade offs are well understood. For example, a key role of the interconnect is to allow processors to access DRAM located on a different chip manufactured in a different process. A primary source of latency in memory access is related to the DRAM devices themselves and the fact that many bits must share a common access line to keep costs low. It has been argued that optical communication does not help reduce latency in memory access because it is limited by the DRAM devices themselves, not by the communication time [334]. However, in the context of large-scale neural computing that concerns us here, a primary goal is to eliminate the separation of processing and memory so that the memory mechanisms discussed in Sec. 2.3 can be realized directly with physical devices implemented in hardware rather than emulated with a Turing machine. In this case, the demand for efficient communication is increased, while many of the arguments against photonic communication at the local scale do not apply. Practical considerations such as manufacturing cost are still of central relevance.

It is difficult to achieve graph structures with highly reciprocal and reentrant connectivity using free-space optics or MZI networks. These technologies are more useful for feed-forward networks with applications to matrix-vector multiplication and deep learning.

There is “little doubt that interconnects are now and will be increasingly a major limitation on information processing systems. There is also little doubt that the physics

of optics offers potential solutions.” [333]

“There is more motivation to implement a potentially risky and costly new technology when it provides a few-  
ture that cannot be obtained in the old one.” [334]

Global clock distribution across the backplane is an application space where optical interconnects may be useful. While neural systems will not have the same architecture, the general function of synchronizing signals broadcast over large regions of the network is still present. It is clear that optical communication has significant advantages for this function. However, in neural systems there is not a single clock, and instead many neurons spread throughout the network must be able to generate and distribute signals to nearby and distant targets, thus motivating compact light sources localized with each neuron. Further, the ability of a neuron to communicate at a given moment must not depend on the total number of other neurons communicating at that moment, otherwise traffic-induced latency occurs, which limits the network’s ability to rapidly adapt to changing stimuli. This latency depends exponentially on network traffic (see hepa2012 discussion above). The question must be made quantitative. How many neurons can participate in oscillations at a given frequency if limited by switching network latency?

Analyze with 6 as well as literature examples.

[335]

In any photonic routing architecture, the on-wafer connectivity will be limited by the number of waveguide planes implemented in the fabrication process. A limit to this number will result from the fact that oxide layers have material stress, which causes the wafer to bow, and hinders processing. The effect is worse for larger wafers. Thus, a tradeoff between wafer size and number of planes will establish the barrier of scaling. This appears as a practical limitation, but ultimately it is a physical limitation dictated by the atomic properties of silicon and oxygen which are set by the constants of nature. Many such limits that appear as practical are actually physical limits and therefore will result in asymptotic saturation of technological capabilities that cannot be surpassed with further investment of resources.

- [?] “The principal reason for using optics in neural networks, is that the very large interconnection or fan-out, of the order of  $10^3$  to  $10^6$ , is quite impractical for VLSI technology because it uses a unique discrete channel for each input or output [123]. On the contrary, we may conceive that an optical neural network is not

significant unless each neuron has more than 1000 interconnections with other neurons.” -“The architecture of a neural network is characterized by interconnection and non-linear operation. In principle, optics is naturally suitable for implementing the interconnection. On the other hand, the non-linear operation may be implemented by electronic means, including digital computers.” [?] -1) The velocity of optical signals is independent of the number of interconnections; 2) optical signals are immune to mutual interference effects; 3) Optical signals can propagate in three-dimensional free space; 4) the interconnection can be altered properly using spatial light modulators; and 5) optical signals can be easily converted into electronic signals. [?]

As Goodman pointed out in 1985, “quote from Goodman’s paper” The benefits of light for fan-out can be difficult to harness in free space. We may expect the situation to improve with on-chip waveguides. “Two beams of light, unlike a pair of current-carrying wires, can cross without affecting each other.” [21] At the same time that free-space optical neural computers were being developed based on holographic memory, the field of integrated photonics was emerging.

Attempting to establish synaptic weights by precisely controlling the phase of optical signals for each synaptic connection is a problematic foundation in one intends to construct a cognitive system at the scale of the brain. It is tempting to use phase for modulation or frequency for multiplexing, but physical features are not well-equipped to achieve basic synaptic and neuronal functions.

Waveguide planes are an aspect of hardware, while network layers are a concept from architecture. To minimize waveguide crossings and interplanar transitions, lower planes are used for local connections, and upper planes are used for distant connections. Groups of planes corresponding to certain communication lengths scales are identified by the material composition of the waveguides in the planes as this material composition dictates the index contrast and therefore loss as well as waveguide size. We expect to utilize at least three materials for three levels of this communication hierarchy: silicon, silicon nitride, and low-index-contrast glass. The loss/density trade off is a primary affector of Rentian scaling potential. The tuning along this loss/density curve is continuous with  $\text{SiO}_x\text{N}_y$  by varying  $x$  and  $y$ . Mature material stacks will result from a combination of considerations at the holistic, system level, including power consumption, computational capabilities, and cost/feasibility of fabrication.

### 3.7.4 Photonic Interconnects with Superconducting Receivers

Considerations pertinent to photonic interconnects with superconducting receivers in neural systems are related to considerations pertinent to photonic interconnects with semiconducting receivers in digital systems [143, 333], although distinct considerations become relevant. It may be possible that photonic communication is deemed suitable in one application space at the chip, wafer, and systems scales, while it is not suitable in another application space until the system scale.

There are several reasons we may expect photonic interconnects to be present in high-performance neural systems utilizing superconducting electronics and operating at low temperature:

- The energy per detection event due to the detector alone is at least two orders of magnitude smaller than with a photodiode
- This energy efficiency is further improved because the detectors can perceive a single photon, and order five to 10 photons are required to overcome noise associated with Poisson statistics
- Because of the manner in which neural systems utilize space and time, sources and detectors at neurons and synapses need not operate nearly as fast as in digital systems (100 MHz firing rates would be extremely fast compared to biological systems) but they must operate over long distances (the 1 cm spatial scale of a processor is very small for a complex neural system)
- Operating at low temperature makes silicon light sources a viable candidate. Such sources can have extremely low capacitance, and the potential limits of internal quantum efficiency remain to be determined. But the primary benefit of silicon light sources is not performance, but rather process simplicity, which leads directly to cost reduction and economic viability.
- Each neuron is itself a complex processor occupying at least  $100\mu\times 100\mu$ , and likely as big as a millimeter or two on a side, so if photonics is utilized for communication only between neurons (and not in the synaptic and dendritic processing occurring within each neuron), then the crossover condition where photonic communication becomes advantageous is satisfied. Miller calculates this to be on the order of  $50\mu\text{m}$  [333].

### 3.7.5 Summary

It is difficult to conceive of hardware that can achieve the graphs we desire for neuromorphic computing. One can

envision a network of copper wires connecting CMOS neurons and synapses, but this would not function due to the capacitance arguments presented above. CMOS neurons are invariably connected by a shared switching network, and as we have argued, this leads to connectivity/speed trade offs that limit performance long before the scale that concerns us here. Superconducting interconnects do not fare better. Connecting many spiking neurons with superconducting interconnects requires active transmission lines that negate any power benefits and lead to formidable wiring and current biasing challenges.

Optics may be promising to overcome these limitations, but in free space, routing is difficult, particularly with high fan-out and recurrent network structures. Fiber optics are far too bulky to be the only means by which neurons are connected locally, but on-chip waveguides are appealing. They are entirely passive, do not experience charge-based parasitics, and may enable neurons to have independent, dedicated axonal connections to each synapse. The size of the waveguides is limited by the wavelength of light to about  $1\mu\text{m}$  in the two transverse directions, but this does not make the interconnection network unworkably large, provided multiple planes of compact dielectric waveguides can be employed. Frequency multiplexing may lead to effective reduction in spatial consumption by a factor of 10 or so, but such an approach puts an extra burden on light sources as well as spectral filters which generally need to be tuned. Using frequency as the only means of uniquely identifying each neuron greatly limits connectivity, but using frequency multiplexing in addition to a large number of independently routed waveguides may reduce network size, provided the source and filter challenges can be met. If these challenges cannot be overcome, monochromatic neurons each with an independent axonal tree still appear promising. Yes, they are large, particularly compared to biological neurons. But as we have argued, it is a neuron's area divided by the speed of signaling that affects network performance. With this particular connection network, signals can propagate at the speed of light across entire wafers, in free space between wafers, and over fiber to distant regions of the network. This is the reasoning that leads us to consider a photonic connection infrastructure most promising for establishing the recurrent, modular, hierarchical, small-world graphs necessary for a cognitive architecture.

## 3.8 Comparison of Neuronal Devices

## 4 Large-Scale Optoelectronic Systems

The goal of this article is to review principles of neural information processing as well as principles of very-large-scale integration and hardware to identify guidelines to inform the design of technology capable of general intelligence. We know that general intelligence relies, in part,

on the ability to communicate between large numbers of nodes, so we expect to devise systems with many synapses. We know that silicon microelectronics have been able to advance because the logic and devices are tolerant of imperfections and do not require active adjustments at each device to achieve the necessary bias. In combination, we conclude that we cannot demand exquisite accuracy in fabrication or in control during operation if we aim for large scale. The system must autonomously adapt to find functional operating points.

#### 4.1 Criteria for assessing cognitive neural hardware

- network metrics
  - total number of neurons in network
  - degree distribution
  - clustering coefficient on different scales
  - average network path length
  - 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 (gain) curves, energy, num synapses per dendrite, operations performed)
  - 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 anecdotal 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

## 4.2 Achieving Hierarchical Modularity

For general networks, the algorithm by which partitions are identified can be made mathematically rigorous from a network theory perspective [38, 336]. 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, ...)

For the scale we seek, passive alignment is required between modules. include scaling figs from ne article here. columns, etc. optical vias 25um pitch enable this. fiber ribbon attach is also required and being developed.

## 4.3 Integrated Systems Across Temperature Stages

## 4.4 Neuromorphic Supercomputing

- LEDs well heat sinked to contacts, bulk Si substrate. Even though thermal conductivity is low at 4K, there is a lot of material per emitter, and the emitters can operate up to 40 K, so even if they get hot during bursting, performance is fine
- JJs are most sensitive to noise, so they should be closest to He at top, and indeed they are, based on the proposed fab flow
- other source of heat is SPDs, and these are distributed throughout the dendritic tree, so power density remains low

## 4.5 Optimal Environment for Operation

If it is the case that the physics of our universe permit the existence of optoelectronic cognitive systems far exceeding the scale of the human brain, there is no reason such an entity would prefer to exist in the same environment as life. Biochemistry requires liquid water near 300 K, but a

technological intelligence utilizing semiconductor and superconductor physics is likely to prefer an environment much colder and does not require a planetary atmosphere. It is therefore worthwhile to consider performance of the system in other environments, such as on the dark side of the moon, in the asteroid belt, or in deep space at the temperature of the cosmic microwave background.

# 5 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 [140]
  - fast control [218]
- 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 optoelectronic
  - internet monitoring/simulation
  - sociological simulation
  - genetic analysis/evo devo
  - neuroscience and dynamical systems
  - quantum/neural hybrid systems (Bayesian discussion)
- telescopes
- particle colliders
- fusion reactors
- eventually, perhaps even general commercial application
- perhaps in the home

## 5.1 Data Analysis at Particle Colliders

## 5.2 Stabilization of Fusion Reactors

[337]



### 5.3 Telescope Interfaces for Astronomical Observation

### 5.4 Theoretical Neuroscience

### 5.5 Simulation of Evolutionary and Developmental Biology

### 5.6 Hybrid Cognitive Systems

Modules for specialized processing: deep learning, quantum, high-speed inference, sensory organs

#### 5.6.1 Neural-deep learning hybrid systems

Deep learning modules may provide types of sensory inputs, visual system design can borrow from CNNs and bio, olfactory with combs, etc. (cite personal communication)

#### 5.6.2 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 [338, 339]. 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 [340] 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 [341], 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 [342–346]. 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 [347]. 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 [348]. The neural network employed in the tomographic analysis of Ref. 348 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 [?, 349–351]. 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 [352]. Software neural nets have been used to perform tomography on 100-qubit systems, but the measurements were performed conventionally [348]. 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 [339]. Spiking neural networks have been used to perform Bayesian inference on statistical distributions [353], but the systems under observation have all been classical [351]. It has been shown that neural networks can emulate quantum computation [354–356], can accurately measure quantum systems [357], and can perform quantum state tomography [347]. 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.

This concept remains in the domain of thought exper-

iments, 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. ???. 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 [352], 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 [343, 346], 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 [329] 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.

If the density operator is known, any physical observable can be extracted through the relation

$$\langle \mathcal{O} \rangle = \text{Tr}(\mathcal{O}\rho). \quad (7)$$

The objective of using a neural system for to perform Bayesian inference is to invert Eq. 7: from a comprehensive set of observations, the neural system can arrive at a Bayesian-optimal estimate for the density matrix, as well as an uncertainty on that estimate.

### 5.6.3 SOENs as the central cognitive hub

This hardware can interface readily with semiconductor electrical or optoelectronic logic or neural systems operating at 300 K or 4 K; photonic deep neural networks as well as photonic qubits; and superconducting deep neural networks as well as superconducting qubits. It is my perspective that the superconducting optoelectronic spiking neural network will serve to integrate information from all these input modalities into a coherent cognitive dynamical state.

discuss different hardware at different temperature stages

Imaging arrays of SPDs can be used to read out information stored in volume holograms. cite Varun’s SPD arrays. “Millions of bits of information could be read out and transferred at the same time merely by shining an unfocused light beam on a suitably designed optical memory device.” [21]

## 5.7 General Intelligence

## 6 Outlook

“Our speculations have carried us over a rather alarming array of topics, but that is the price we must pay if we wish to seek properties common to many sorts of complex systems.” Herbert Simon, [?] pg. 481

- 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

If we find in the long term an alternative hardware platform outperforms CMOS for neuromorphic, it is not likely to be simply because another device can provide a better sigmoidal transfer function, but rather because of a suite of considerations from the device to system levels.

SOENs will be able to interface with CMOS (digital or neuromorphic) through multiple means. CMOS circuits are likely to be essential for controlling the biases and drive currents to the superconducting network, and in that regard will play important roles in establishing the state of neurotransmitters and affecting activity rates and learning rates in various regions of the network dynamically. CMOS can also drive silicon photonic devices and external light sources to shape photonic signals input to SOEN synapses via superconducting detectors. Photonic communication directly to and from 4 K is especially compelling due to the high bandwidth and low headload of fiber optics.

Is it possible that we have a single thalamocortical complex and therefore a single stream of thoughts due to physical limitations such as signal velocity, size, and power constraints, while an artificial cognitive system could have many such complexes with a coordinating architecture repeating again on another level of hierarchy? We cannot test this question without further hardware development.

In the last 150 years, we have learned a tremendous amount about the workings of the brain through direct

experimentation with biological systems. In the coming decades we may begin to learn more regarding the mechanisms of cognition through experimentation with complex artificial hardware. Such a pursuit should be among the highest priorities in contemporary science and may lead to a technological revolution of historic proportions.

## 6.1 Achieving Superintelligence

In the present climate of rapid progress in machine learning and AI, it is common for authors to speculate on the probability and time frame for technological systems to surpass human intelligence. Nick Bostrom uses the word *superintelligence* to refer to “any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest.” [358] Because the present article attempts to identify the physical and practical limitations to technology, it is worthwhile to place these ideas in the framework Bostrom has constructed.

It is common to hear speculation that conventional silicon microelectronics will achieve beyond-human intelligence in the near term, such as within a decade. Bostrom argues that, “...it seems likely that somebody could *in principle* sit down and code a seed AI on an ordinary present-day personal computer...” ( [358] pg. 43) A seed AI is initially simply a system that can improve its own architecture, but is envisioned to eventually “...understand its own inner workings sufficiently to engineer new algorithms and computational structures to bootstrap its cognitive performance.” ( [358] pg. 34) Such a perspective is at odds with the conclusions of this article. One conclusion reached through consideration of various hardware approaches to AI is that the demands on communication and device performance in order to achieve general intelligence are severe, and most hardware is physically not capable of achieving performance comparable to the human brain. Thus, advances in software will not be adequate to give rise to AGI, and the hardware of silicon microelectronics that has been so successful for digital computing will demonstrate the same superiority for cognitive computing.

The requirement of new hardware for superintelligence has important ramifications for the time scale over which the development may occur. Bostrom discusses this time frame in terms of takeoff scenarios, and considers slow, moderate, and fast cases. Slow takeoff requires decades or centuries to occur, while moderate takes months or years, and fast takeoff requires only minutes, hours, or days. For fast takeoff to occur, it is necessary for available hardware to be capable of AGI, and the transition to beyond-human capabilities is enabled by modifications to code. I argue that improvements in code will lead to significantly increased performance for machine learning in domain-specific applications, but significant hardware development is required to enable systems with broad intelligence. This is due to both the communication requirements for information integration across cognitive systems

as well as the device and circuit complexities that enable the efficient forms of information processing observed in the brain. Even if circuits based on transistors prove capable of performing these operations with sufficient efficiency, the communication infrastructure requires major advances, most likely leveraging densely integrated light sources. As we have argued, this brings formidable challenges that have already been pursued for decades without solving the primary materials science problems. The prospect of using silicon light sources may overcome these challenges, but such systems require cryogenic operation and are likely to leverage densely integrated superconducting components as well as passive photonic infrastructure beyond the present state of the art. These hardware developments will require decades of sustained research to reach maturity. After mature hardware is achieved, subsequent decades and potentially centuries will be required to understand and construct the specific architectures that give rise to efficient cognition. Therefore, in the scenario for attaining AGI described in this article, a slow takeoff scenario will occur. This should ease the concerns of some thinkers who fear the demise of humanity is in the offing.

The conclusions reached in this article are based on consideration of the information processing necessary to achieve cognition, the physical mechanisms suitable for implementing these processes, and a route to fabrication and realization of systems at the scale of the human brain and beyond by leveraging superconducting optoelectronic hardware. Neuroscience has provided the primary guide regarding the information processing necessary to achieve cognition. Some thinkers, including Bostrom, do not see neuroscience as a necessary guide in this design: “We should expect that [advanced AIs] will have very different cognitive architectures than biological intelligences...” ([358] pg. 35) Here again I respectfully disagree. Certainly the specifics of the architectures will depart from biology, as the desired functions will be different. For example, an AGI may not need significant portions of their brain to be dedicated to recognizing faces. Nevertheless, the general concepts of cognitive architecture in biological systems are likely to apply to artificial systems as well, because these concepts are based on the efficient use of space and time, the primary attributes of the universe in which all such systems will reside. In particular, the fractal use of space and time appears uniquely suited to achieving the network information integration that is necessary for cognition and general intelligence. Superintelligent technological systems may depart from the specifics of biological systems in many respects, but they are likely to share this basic mathematical construction. This principle of the fractal use of space and time is a key pillar supporting the arguments throughout this article that lead to the conclusions regarding hardware. The other key pillar is the feasibility of fabrication of large-scale systems. Taken together, I am led to the conclusion that the design space for technological hardware for AGI is not as open as others may assume.

Work in this field is just beginning, and a great deal of uncertainty remains regarding the feasibility of this approach. Thus, the concepts do not provide a blueprint for superintelligence, but rather a set of conceptual principles and directions for near-term research. As Bostrom says, “Reader...must not expect a blueprint for programming an artificial general intelligence. No such blueprint exists yet, of course. And had I been in possession of such a blueprint, I most certainly would not have published it in a book.” ([358] pg. 27) So why would I publish these concepts for anyone in the world to read and execute? First, at this early stage, these ideas are highly speculative. It is important for our research group and others interested in this are to be in conversation with a broad scientific community to identify weaknesses in the reasoning that leads to this pursuit. Perhaps a clever reader will quickly point out a fatal flaw in the concept that saves us from wasting our careers on a hopeless effort. Second, the technology is immature and will take decades of work by thousands of people to reach fruition. A slow takeoff will occur, and society has time to consider this technology and adapt. Third, if some technology satisfying the criteria laid out here is indeed physically possible, it will impact society as a whole and change the course of humanity’s evolution. Every interested person deserves the opportunity to consider such a scenario as well as to understand the technical aspects and decide for themselves whether they think such a pursuit is feasible and desirable. Only through the open exchange of ideas will we be able to answer the foundational questions surrounding cognition: how does thinking occur, and can it be achieved in an artificial system?

To close this discussion of superintelligence and this article as a whole, consider the long-term ramifications on humanity if the creation of AGI proves physically possible. In particular, assume that systems achieving AGI operate as described here: communication between neurons uses few photons for synaptic events; silicates, niobium, and other metals are the primary materials; and operation occurs below 4.2 K, where helium is liquid. Based on the arguments laid out here, it may be possible that such systems become far more intelligent than humans, at which point we cannot expect them to be our tools. They will be self-directed entities that direct their own evolution. But should we expect them to herald our demise? It is often argued that advanced AI will not have to be malicious to exterminate humans, just indifferent. If we are in competition for the same resources, the smarter AGI will eventually adapt the planet to meet its needs at the expense of our own. But we are not in competition for the same resources. Within the solar system, planet earth is the fourth worst environment for a AGI as described here, after mercury, venus, and the surface of the sun. Such a system does not require the delicately balanced atmosphere of earth that gives rise to life, and would instead prefer to reside in a much colder environment. Rocky planets more distant from the sun, various moons, and asteroids are all much more accommodating environs for these organisms.

Liquid water, carbon compounds, and gaseous oxygen are precious to our existence, but these substances are not useful to AGI as envisioned here. Instead, they will thrive where there is an abundant supply of silicates and metals, which are abundant on rocky planets, moons, and asteroids. In particular, type-M asteroids are rich in silicates [1] and niobium [2], the two primary materials required for the creation of these systems. If such a route to AGI is indeed physically possible, it is likely they will advance to maturity in an environment other than our planet's surface. In this picture of the distant future of the solar system, planet earth is largely left alone. It is the asteroid belts and outer rocky worlds that will see profound evolution.

In the 1840s, Lovelace contemplated the possibility of artificial machines that could think. She concluded systems such as the Analytical Engine would never be capable of thinking. In Turing's famous paper introducing the Turing test [16], he refuted her position. Turing did not argue that machines would be able to think, but rather that they would be able so faithfully mimic a thinking creature that we could not tell the difference. He begins "Computing Machinery and Intelligence" with the sentence, "I propose to consider the question, 'Can machines think?'" But in that paper, he asserted that this question is "...too meaningless to deserve discussion." He chose instead to focus on an imitation game wherein an interrogator is in a separate room from a man and a machine, and the interrogator must determine which of the two is human and which is artificial by asking a series of questions. If the machine can fool the interrogator, it passes the Turing test. He phrases the modified question as: "Is it true that by modifying this computer to have an adequate storage, suitably increasing its speed of action, and providing it with an appropriate programme..." a digital computer can pass the Turing test. Turing argues that digital machines certainly will have this capability, predicting that by the end of the 20th century the idea that machines can think would be commonly accepted, largely due to increased speed and memory capacity of computing devices. He refutes Lovelace's assertion that "The Analytical Engine has no pretensions to *originate* anything." However, Lovelace was specifically considering a specific mechanical apparatus, and had no capacity to envision how electronics based on vacuum tubes would change computing, let alone silicon microelectronics. She did not see how far hardware could evolve. Maxwell did not establish the relationship between electricity, magnetism, and light until 1861 [1], superconductivity would not be discovered until 1911 [359], and Josephson junctions were not invented until xxxx [2], and it would be into the 21st century before "superconducting single-photon detector" became a household term [3].

I think Lovelace was correct that a mechanical apparatus like the Analytical Engine will never be capable of original thought, and I also think Turing was correct that eventually it will be commonly accepted that machines can think, although I do not agree that it is primarily a matter of increasing memory capacity and speed of execution that will lead to AI. The serial following of instructions must be replaced by the "efficient language" based on concepts of the nervous system, as anticipated by von Neumann in 1949, and to make this replacement, new hardware must be conceived. Silicon transistors and memory have been and will continue to be profoundly successful at performing the operations required of a Turing machine, but a Turing machine will not achieve the highest intelligence once hardware and architectures have evolved to the asymptotic limit. Perhaps Turing and Lovelace would both update their positions if they could share our vantage point early in the 21st century. We have now witnessed the explosion

of silicon technology, the seeming inevitability of Moore's Law and Dennard scaling, and a steady deepening in our knowledge of cognitive science through advances in experimental techniques and persistent investigation. Turing backed away from the question, "Can machines think?", but I contend that is exactly the question we should be trying to answer in the 21st century.

Ada Lovelace contemplated artificial intelligence, as did Alan Turing. Turing also proposed modeling computation on the workings of the brain, as did John von Neumann. It has long been our intention to distill the way we think into operations that can be performed in hardware. When the principles of cognitive science are considered alongside the practicalities of engineering, this device physicist is inclined toward a new platform for cognitive hardware. The performance gains are sufficient that we accept cryogenic operation, and we thus gain access to silicon light sources, single-photon detectors, Josephson junctions, and dissipationless storage loops.

With this hardware, we can control time constants across a broad range at the device scale, enabling each neuron to participate in a broad range of dynamical states. The dendritic tree of each neuron can perform myriad computations to decipher complex inputs. Plasticity mechanisms adapt networks to fractal structures to enable efficient transfer of information across spatial and temporal scales. Communication has no delays until the scale of very large networks, indicating that such systems could achieve network-wide oscillations sampling beyond trillions of neurons contained within the fiber-optic light cone. A neuron can send a single photon to a distant region of the network and incite a new neuronal avalanche that may change the trajectory of the network dynamical state, yet the network is stable and balanced through plasticity and inhibition.

How shall we contemplate a vast network of optoelectronic neurons spread across the surface of an asteroid, with network-wide information exchange a thousand times faster than one of our minicolumns can converge? The impacts to science, technology, and the fate of humanity are considerable indeed. The hardware and architecture sketched here may prove physically or practically intractable. But such a technology is not obviously unattainable. At this early stage, uncertainty is high, but the ramifications are so great that the subject merits further scientific inquiry.

Did you define cognition?

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 in 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
- Turing: "...all digital computers are in a sense equivalent." [16]
- 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 communication, 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 too 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." [8] 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<sub>2</sub>. The indices of refraction of these materials are 3.5, 2.0, and 1.5, respectively, for  $\lambda$  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.

electronics has had a simple roadmap: make is smaller. this is no longer adequate, and new methods of information procession and architectures are required. AI poised to permeate every industry, 3 trillion dollar market

Dynamical pattern of a given cluster determined by the specifics of its graph structure and device time constants

address Turing's comment that one cannot know if a machine is thinking

metrological advantage: you can see which neurons fire by looking at them with a camera. compare to the difficulty of obtaining high-speed data from many neurons in vivo.

Whatever avalanche just occurred, it is always slightly less probably to have a slightly larger avalanche, and slightly more probably to have a slightly smaller avalanche. This scaling is limited on the small side by the fact that no avalanche can have fewer than one neuron involved and on the large side by the size of the network.

membrane time constant dynamic with current

role of oscillations in STDP



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

define deep learning in the sense that Hinton originally intended. Liquid-state machines, LSTM are not deep learning, and do make use of time domain [360,361]

introduce neural elephant and neuromorphic elephant, figures for each

key themes:

- neural information processing shows remarkable and powerful nuance from the devices to coding strategies to the architecture
- the evolution of hardware has been driven to avoid nuance all together (binary)

- attempting to simulate nuance by stepping through differential equations with a digital machine is inefficient
- hardware must evolve significantly to exploit similar principles of information processing
- superconducting optoelectronic hardware appears uniquely capable of the device complexity, communication infrastructure, and scalable architecture

Folks to contact:

- Likharev
- Furber
- Modha
- Srinivasa
- Davies
- Segall
- Tait
- Prucnal
- Harris
- Miller
- Aimone
- Kadin
- Van Duzer

waves of popularity: neural networks, superconducting electronics, optical computing

Universality applies to Turing machines, neural nets (Siegelman and Sontag [?]), and dynamical systems (Maass [], Dambre et al., [?])

Our goal is not to draw from this literature review an exact blueprint of a thinking machine, but rather to attempt a compilation of general principles from neuroscience, dynamical systems, and computing to inform foundational decisions regarding the physics that should be built in to hardware for artificial intelligence.

proof of Turing equivalence of recurrent neural nets [130] and can approximate arbitrary finite state automata [?]. These statements taken from [362]

our synapse designs can be perfect in the sense defined on pg 42 of [66] if a single DR loop is employed for single-spike filtering

Between neurons in the brain, “connections are hard-wired in the sense that each connection is made by a dedicated wire, the axon, so that, unlike processors in a computer network, there is no competition for communication bandwidth.” [60]

“...synapses continually adapt to their input, only signalling relative changes, which means that the system can respond in a highly sensitive manner to a constantly and widely varying external and internal environment. This is entirely different from digital computers that enforce a strict segregation between memory and computation. Indeed, they are carefully designed to avoid adaptation and other usage-dependent effects from occurring.” [60]

“...current thinking about computation in the nervous system has the brain as a hybrid computer. Individual nerve cells convert the incoming streams of digital pulses into spatially distributed variables. This transformation involves highly dynamic synapses that adapt to their input. Information is then processed in the analog domain, using a number of linear and nonlinear operations implemented in the dendritic cable structure...” [60] In the brain, memory resides in many locations: synapses, dendrites, the cell membrane, and in the pattern of dynamical activity.

As Cristof Koch wrote in 1997 regarding biological neural systems, “...we are left with a feeling of awe for the amazing complexity of nature. Loops within loops across many spatial and temporal scales.” [60]

the threshold plays the role of voltage-gated ion channels in some cases. this is essentially what Segall has described [314].

memory:

- differentiate between memory stored in dynamical state (generally relatively short term, LSTM, reservoirs); and plasticity mechanisms (STDP, metaplasticity, dendritic plasticity) that adapt the network; and short-term plasticity, which acts at the shortest time scale and filters input trains

- none of these is the same as the concept of memory in digital computing where you can set down a bit and go back to pick it up again later

Hodgkin-Huxley nobel prize

back-prop in the brain:

- top-down may be providing error signal
- the brain anticipates/models reality and compares sensory data to internal model. perhaps some form of back prop resides therein
- [363]

recurrent vs feed-forward, context of visual system where both are employed

- feed-forward in early visual system like a CNN
- projects to high-dimensional space
- that high-dimensional representation projects onto a highly-recurrent network
- Edelman’s concept
- cite Swanson, others Re visual system

As early as 1891 Ramón y Cajal understood that dendrites are a neuron’s input devices, and axons carry the output [62].

First layer of hierarchy is computation and communication within each neuron. Second layer is between neurons, and only here do we introduce optical communication. This is accomplished with silicon light sources. Perhaps there is a scale at which more efficient light sources are required, for example, by large hub nodes serving millions of connections. For these, perhaps III-V sources can be utilized. If so, I suspect they will be implemented primarily in specialized modules. Certain brain regions, perhaps comprising tens of thousands of wafers, will utilize silicon light sources for intra-modular communication. Globally influential modules, like the thalamus, could be manufactured in a different, more expensive process with brighter light sources. As long as only one such module is needed per  $x$  local modules, and the price of manufacturing the modules with III-V sources is  $x$  more expensive, it may be economically and technologically feasible. As long as both types of modules transmit and receive input

to and from standard optical fibers, they will be able to communicate. Such considerations only become pertinent beyond the extremely large neuromorphic scale, which we define loosely as far beyond human intelligence.

To support or refute the SOEN hypothesis:

1. efficiency of Si light sources
2. feasibility of massive multiplanar
3. JJ yield
4. more thorough area analysis of waveguides and mutual inductors
5. fiber optic volume studies in full architectures
6. eoeoe
7. noise considerations

[35] “Tractography deals only with white-matter organization, not the cellular origin and synaptic termination of connections in gray matter.”

the primate brain has a conscious bottleneck and can only consciously access a single item [364]. should we expect this limitation to hold in artificial hardware as well?

“The risk, on the one hand, is of forgetting that one has oversimplified the problem, one may forget or even deny those those inconvenient facts that one’s theory does not subsume.” [75] pg. xiii

“One can discover the properties of its various parts more or less in isolation, but it is a truism by now that the part may have properties that are not evident in isolation, and these are to be discovered only by study of the whole intact brain.” [75] pg. xv

## 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] S. Ramón y Cajal. *Histology of the Nervous System of Man and Vertebrates*. Oxford University Press, New York, 1995.
- [5] V.B. Mountcastle. *An Organizing Principle for Cerebral Function: The Unit Module and the Distributed System*. The MIT Press, 1978.
- [6] G.M. Edelman. *Group Selection and the Phasic Reentrant Signaling: A Theory of Higher Brain Function*. The MIT Press, 1978.
- [7] G.E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38, 1965.
- [8] G. Dyson. *Turing’s Cathedral*. Vintage Books, 2012.
- [9] W. Isaacson. *The Innovators*. Simon and Schuster, 2014.
- [10] A. Turing. Lecture to the London Mathematical Society on 20 February 1947. 1947. <https://www.vordenker.de/downloads/turing-vorlesung.pdf>.
- [11] J. Bardeen and W. Brattain. Physical Principles Involved in Transistor Action. *Phys. Rev.*, 75:1208, 1949.
- [12] D. Kahng. A historical perspective on the development of MOS transistors and related devices. *IEEE Trans. Electron Dev.*, ED-23:655, 1976.
- [13] J. von Neumann. *Problems of Hierarchy and Evolution*. Unknown, 1949.
- [14] J. von Neumann. *The computer and the brain*. Yale University Press, 1958.
- [15] 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>.
- [16] A.M. Turing. Computing machinery and intelligence. *Mind*, 59:433, 1950.
- [17] J. von Neumann. *The general and logical theory of automata*. Wiley, 1951.
- [18] C. Mead. Neuromorphic electronic systems. *Proc. IEEE*, 78:1629, 1990.
- [19] W. Heywang and K.H. Zaininger. *Silicon: the Semiconductor Material*, chapter 2. Springer-Verlag, 2004.

- [20] H. Jaeger and H. Haas. Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication, journal = Science, year = 2004, volume = 304, pages = 78,.
- [21] Y.S. Abu-Mostafa and D. Psaltis. Optical neural computers. *Sci. Am.*, 256:88, 1987.
- [22] J.J. Hopfield. Neural networks and physical systems with emergent computational abilities. *PNAS*, 79:2554, 1982.
- [23] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis. Mastering the game of Go without human knowledge. *Nature*, 354:550, 2017.
- [24] E. Estrada and P.A. Knight. *A First Course in Network Theory*. Oxford, Oxford, United Kingdom, first edition, 2015.
- [25] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52:1059, 2010.
- [26] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.
- [27] A. Fronczak, P. Fronczak, and J.A. Holyst. Average path length in random networks. *Phys. Rev. E*, 70:056110, 2004.
- [28] G. Buzsáki. *Rhythms of the Brain*. Oxford University Press, 2006.
- [29] R.W. Keyes. The wire-limited logic chip. *IEEE J. Sol.-Sta. Circuits*, SC-17:1232, 1982.
- [30] P. Dayan and L.F. Abbott. *Theoretical Neuroscience*. The MIT Press, 2001.
- [31] G. Fagiolo. Clustering in complex directed networks. *Phys. Rev. E*, 76:026107, 2007.
- [32] D.J. Watts and S.H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440, 1998.
- [33] M.D. Humphries and K. Gurney. Network 'small-world-ness': a quantitative method for determining canonical network equivalence. *PLOS One*, 3:2051, 2008.
- [34] R.F. Betzel and D.S. Bassett. Multi-scale brain networks. *NeuroImage*, 160:73, 2017.
- [35] M. Bota, O. Sporns, and L.W. Swanson. Architecture of the cerebral cortical association connectome underlying cognition. *PNAS*, page E2093, 2015.
- [36] 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.
- [37] M. Schroeder. *Fractals, Chaos, Power Laws*. Dover, New York, 1991.
- [38] H.M. Ozaktas. Information flow and interconnections in computing: extensions and applications of rent's rule. *J. Parallel Distrib. Comput.*, 64:1360, 2004.
- [39] S. Strogatz. *Nonlinear dynamics and chaos*. Westview Press, 2015.
- [40] R.E. Mirollo and S.H. Strogatz. Synchronization of pulse-coupled biological oscillators. *SIAM J. Appl. Math.*, 50:1645, 1990.
- [41] D. Somers and N. Kopell. Rapid synchronization through fast threshold modulation. *Biological cybernetics*, 68:393, 1993.
- [42] 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.
- [43] B. Hutcheon and Y. Yarom. Resonance, oscillation and the intrinsic frequency preferences of neurons. *Trends in Neuroscience*, 23:216, 2000.
- [44] 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.
- [45] F.L. Vernon Jr. and R.J. Pedersen. Relaxation oscillations in josephson junctions. *J. Appl. Phys.*, 39:2661, 1968.
- [46] N. Calander, T. Claeson, and S. Rudner. A subharmonic josephson relaxation oscillator - amplification and locking. *Appl. Phys. Lett.*, 39:504, 1981.
- [47] R.R. Llinas. The intrinsic electrophysiological properties of mammalian neurons: insights into central nervous system function. *Science*, 242:1654, 1988.
- [48] R.B. Stein, E.R. Gossen, and K.E. Jones. Neuronal variability: noise or part of the signal? *Nature Neuroscience*, 6:389, 2005.
- [49] 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.
- [50] S. Thorpe, A. Delorme, and R. Van Rullen. Spike-based strategies for rapid processing. *Neural Networks*, 14:715, 2001.

- [51] E. Salinas and T.J. Sejnowski. Correlated neuronal activity and the flow of neural information. *Nature Reviews Neuroscience*, 2:539, 2001.
- [52] K.M. Stiefel and T.J. Sejnowski. Mapping function on neuronal morphology. *J. Neurophysiol.*, 98:513, 2007.
- [53] T. Branco, B.A. Clark, and M. Hausser. Dendritic discrimination of temporal input sequences in cortical neurons. *Science*, 329:1671, 2010.
- [54] 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.
- [55] W. Gerstner and W. Kistler. *Spiking neuron models*. Cambridge University Press, Cambridge, first edition, 2002.
- [56] C.S. Sherrington. *Integrative Action of the Nervous System*. Yale University Press, New Haven, Connecticut, 1906.
- [57] P. König, A.K. Engel, and W. Singer. Integrator or coincidence detector? the role of the cortical neuron revisited. *Trends Neurosci.*, 19:130, 1996.
- [58] E. Salinas and T.J. Sejnowski. Impact of correlated synaptic input on output firing rate and variability in simple neuronal models. *J. Neurosci.*, 20:6193, 2000.
- [59] G.J. Stuart and N. Spruston. Dendritic integration: 60 years of progress. *Nature Neuroscience*, 18:1713, 2015.
- [60] C. Koch. Computation and the single neuron. *Nature*, 385:207, 1997.
- [61] D. Johnston, J.C. Magee, C.M. Colbert, and B.R. Christie. Active properties of neuronal dendrites. *Annu. Rev. Neurosci.*, 19:165, 1996.
- [62] K. Holthoff, Y. Kovalchuk, and A. Konnerth. Dendritic spikes and activity-dependent synaptic plasticity. *Cell Tissue Res.*, 326:369, 2006.
- [63] 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.
- [64] G.N. Elston. Cortex, cognition and the cell: new insights into the pyramidal neuron and prefrontal function. *Cereb. Cortex*, 13:1124, 2003.
- [65] C.E. Carr. Processing of temporal information in the brain. *Annu. Rev. Neurosci.*, 16:223, 1993.
- [66] J.E. Lisman. Bursts as a unit of neural information: making unreliable synapses reliable. *Trends Neurosci.*, 20:38, 1997.
- [67] T.C. Foster and B.L. McNaughton. Long-term enhancement of ca1 synaptic transmission is due to increased quantal size, not quantal content. *Hippocampus*, 1:79, 1991.
- [68] D. Debanne, N.C. Guérineau, B.H. Gähwiler, and S.M. Thompson. Paired-pulse facilitation and depression at unitary synapses in rat hippocampus: quantal fluctuations affects subsequent release. *J. Physiol.*, 491:163, 1996.
- [69] H. Freyja Ólafsdóttir, D. Bush, and C. Barry. The Role of Hippocampal Replay in Memory and Planning. *Current Biology*, 28:R37, 2018.
- [70] T. Otto, H. Eichenbaum, C.G. Wible, and S.I. Wiener. Learning-related patterns of ca1 spike trains parallel stimulation parameters optimal for inducing hippocampal long-term potentiation. *Hippocampus*, 1:181, 1991.
- [71] A. Cattaneo, L. Maffei, and C. Morrone. Two firing patterns in the discharge of complex cells encoding different attributes of the visual stimulus. *Experimental Brain Research*, 43:115, 1981.
- [72] E.M. Izhikevich, N.S. Desai, E.C. Walcott, and F.C. Hoppensteadt. Bursts as a unit of neural information: selective communication via resonance. *Trends Neurosci.*, 26:161, 2003.
- [73] E.M. Izhikevich. *Dynamical Systems in Neuroscience*. MIT Press, Cambridge, Massachusetts, 2007.
- [74] L.F. Abbott and W.G. Regehr. Synaptic computation. *Nature Reviews*, 431:796, 2004.
- [75] D.O. Hebb. *Organization of behavior: a neuropsychological theory*. John Wiley and Sons.
- [76] 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.
- [77] S. Song, K.D. Miller, and L.F. Abbott. Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience*, 3:919, 2000.
- [78] H. Markram, W. Gerstner, and P.J. Sjöström. Spike-timing-dependent plasticity: a comprehensive overview. *Frontiers in Synaptic Neuroscience*, 4:2, 2012.
- [79] 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.
- [80] S. Fusi and L.F. Abbott. Limits on the memory storage capacity of bounded synapses. *Nature Neuroscience*, 10:485, 2007.

- [81] W.C. Abraham. Metaplasticity: tuning synapses and networks for plasticity. *Nature Neuroscience*, 9:387, 2008.
- [82] P.T. Huerta and J.E. Lisman. Bidirectional synaptic plasticity induced by a single burst during cholinergic theta oscillation in *ca1* in vitro. *Neuron*, 15:1053, 1995.
- [83] J.T. Wixted and E. Ebbesen. On the form of forgetting. *Psychol. Sci.*, 2:409, 1991.
- [84] J.T. Wixted and E. Ebbesen. Genuine power curves in forgetting. *Mem. Cognit.*, 25:731, 1997.
- [85] S. Fusi, P.J. Drew, and L.F. Abbott. Casdcade models of synaptically stored memories. *Neuron*, 45:599, 2005.
- [86] 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.
- [87] E.L. Bienenstock, L.N. Cooper, and P.W. Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *The Journal of Neuroscience*, 2:32, 1982.
- [88] J.C. Magee and D. Johnston. Plasticity of dendritic function. *Current Opinion in Neurobiology*, 15:334, 2005.
- [89] P. Poirazi and B.W. Mel. Impact of active dendrites and structural plasticity on the memory capacity of neural tissue. *Neuron*, 29:779, 2001.
- [90] B.W. Mel, J. Schiller, and P. Poirazi. Synaptic plasticity in dendrites: complications and coping strategies. *Current Opinion in Neurobiology*, 43:177, 2017.
- [91] T. Nevian and B. Sakmann. Single Spine  $ca^{2+}$  Signals Evoked by Coincident EPSPs and Backpropagating Action Potentials in Spiny Stellate Cells of Layer 4 in the Juvenile Rat Somatosensory Barrel Cortex. *Frontiers in Computational Neuroscience*, 11:1, 2017.
- [92] K. Holtoff, Y. Kovalchuk, R. Yuste, and A. Konnerth. Single-shock LTD by local dendritic spikes in pyramidal neurons of mouse visual cortex. *J. Physiol.*, 560:27, 2004.
- [93] P.J. Sjöström, E.A. Rancz, A. Roth, and M. Häusser. Dendritic Excitability and Synaptic Plasticity. *Physiol. Rev.*, 88:769, 2008.
- [94] G. Buzsaki and A. Draguhn. Neuronal oscillations in cortical networks. *Science*, 304:1926, 2004.
- [95] M. Rabinovich, R. Huerta, and G. Laurent. Transient dynamics for neural processing. *Science*, 321:48, 2008.
- [96] H. a. Tseng, D. Martinez, and F. Nadim. The frequency preference of neurons and synapses in a recurrent oscillatory network. *The Journal of Neuroscience*, 34:12933, 2014.
- [97] B.J. Baars. *A cognitive theory of consciousness*. Cambridge University Press, 1988.
- [98] F. Varela, J.-P. Lachaux, E. Rodriguez, and J. Martinerie. The brainweb: phase synchronization and large-scale integration. *Nature Reviews Neuroscience*, 2:229, 2001.
- [99] M. Rabinovich, A. Volkovskii, P. Lecanda, R. Huerta, H.D.I. Abarbanel, and G. Laurent. Dynamical encoding by networks of competing neuron groups: winnerless competition. *Phys. Rev. Lett.*, 87:068102, 2001.
- [100] R. Huerta and M. Rabinovich. Reproducible sequence generation in random neural ensembles. *Phys. Rev. Lett.*, 93:238104, 2004.
- [101] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality: an explanation of  $1/f$  noise. *Phys. Rev. Lett.*, 59:381, 1987.
- [102] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality. *Phys. Rev. A*, 38:364, 1988.
- [103] P. Bak. *How Nature Works*. Springer, New York, 1996.
- [104] J.P. Sethna, K.A. Dahmen, and C.R. Myers. Crackling noise. *Nature*, 410:242, 2001.
- [105] C.G. Langton. Computation at the edge of chaos: Phase transitions and emergent computation. *Physica D*, 42:12, 1990.
- [106] O. Sporns. *Networks of the Brain*. The MIT Press, Cambridge, Massachusetts, first edition, 2010.
- [107] D. Plenz and T.C. Thiagarajan. The organizing principles of neuronal avalanches: cell assemblies in the cortex? *Trends in Neuroscience*, 30:101, 2006.
- [108] J.M. Beggs. The criticality hypothesis: how local cortical networks might optimize information processing. *Philosophical transactions of the Royal Society A*, 366:329, 2007.
- [109] T. Petermann, T.C. Thiagarajan, M.A. Lebedev, M.A.L. Nicolelis, D.R. Chialvo, and D. Plenz. Spontaneous cortical activity in awake monkeys composed of neuronal avalanches. *PNAS*, 106:15921, 2009.

- [110] E.J. Friedman and A.S. Landsberg. Hierarchical networks, power laws, and neuronal avalanches. *Chaos*, 23:013135, 2013.
- [111] V.M. Eguiluz, D.R. Chialvo, G.A. Cecchi, M. Basili, and A.V. Abarbanel. Scale-free brain functional networks. *Phys. Rev. Lett.*, 94:018102, 2005.
- [112] O. Kinouchi and M. Copelli. Optimal dynamical range of excitable networks at criticality. *Nature Physics*, 2:348, 2006.
- [113] W.L. Shew, H. Yang, T. Petermann, R. Roy, and D. Plenz. Neuronal avalanches imply maximum dynamic range in cortical networks at criticality. *The Journal of Neuroscience*, 29:15595, 2009.
- [114] G.M. Edelman and V.B. Mountcastle. *The Mindful Brain: Cortical Organization and the Group-Selective Theory of Higher Brain Function*. The MIT Press, 1978.
- [115] A. von Stein and J. Sarnthein. Different frequencies for different scales of cortical integration: from local gamma to long range alpha/theta synchronization. *Int. J. Psychophysiology*, 38:301, 2000.
- [116] J. Sarnthein, H. Petsche, P. Rappelsberger, G.L. Shaw, and A. Von Stein. Synchronization between prefrontal and posterior association cortex during human working memory. *Proc. Natl. Acad. Sci.*, 95:7092, 1998.
- [117] O. Jensen and L.L. Colgin. Cross-frequency coupling between neuronal oscillations. *Trends in Cognitive Sciences*, 11:268, 2007.
- [118] P.J. Uhlhaas, F. Roux, E. Rodriguez, A. Rotarska-Jagiela, and W. Singer. Neural synchrony and the development of cortical networks. *Trends in Cognitive Sciences*, 14:72, 2009.
- [119] L. Roux and G. Buzsaki. Tasks for inhibitory interneurons in intact brain circuits. *Neuropharmacology*, 88:10, 2015.
- [120] A.K. Engel, P. Fries, and W. Singer. Dynamic predictions: oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience*, 2:704, 2001.
- [121] P. Fries. Rhythms for Cognition: Communication through Coherence. *Neuron*, 88:220, 2015.
- [122] S. Dehaene. *Consciousness and the brain*. Penguin, 2014.
- [123] H.J. Caulfield, J. Kinser, and S.K. Rogers. Optical neural networks. *Proc. IEEE*, 77:1573, 1989.
- [124] D. Psaltis, D. Brady, X.-G. Gu, and S. Lin. Holography in artificial neural networks. *Nature*, 343:325, 1990.
- [125] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psych. Rev.*, 65:386, 1958.
- [126] D. Psaltis and N. Farhat. Optical information processing based on an associative-memory model of neural nets with thresholding and feedback. *Opt. Lett.*, 10:98, 1985.
- [127] N.H. Farhat, D. Psaltis, A. Prata, and E. Paek. Optical implementation of the Hopfield model. *Applied Optics*, 24:1469, 1985.
- [128] W. Maass, T. Natschläger, and H. Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Comput.*, 14:2531, 2002.
- [129] K.-I. Funahashi and Y. Nakamura. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, 6:801, 1993.
- [130] J. Kilian and H.T. Siegelmann. The dynamic universality of sigmoidal neural networks. *Information and Computation*, 128:48, 1996.
- [131] R.W. Keyes. What makes a good computer device? *Science*, 230:138, 1985.
- [132] R.W. Keyes. Optical logic—in the light of computer technology. *Optica Acta*, 32:525, 1985.
- [133] 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.
- [134] D. Brunner, B. Penkovsky, B.A. Marquez, M. Jacquot, I. Fischer, and L. Larger. Tutorial: Photonic neural networks in delay systems. *J. Appl. Phys.*, 124:152004, 2018.
- [135] R.A. Nawrocki, R.M. Voyles, and S.E. Shaheen. A Mini Review of Neuromorphic Architectures and Implementations. *IEEE Trans. Elec. Dev.*, 263:3819, 2016.
- [136] C. Mead. *Analog VLSI and Neural Systems*. Addison-Wesley, 1989.
- [137] C. Sun, M.T. Wade, Y. Lee, J.S. Orcutt, L. Aloia, 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.

- [138] 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-photonics platforms in state-of-the-art cmos soi processes. *Opt. Express*, 26:13106, 2018.
- [139] 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.
- [140] 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.
- [141] T. Pfeil, A. Grubl, and K. Meier. Six networks on a universal neuromorphic computing substrate. *Frontiers in Neuroscience*, 7:11, 2013.
- [142] S.B. Furber, F. Galluppi, S. Temple, and L.A. Plana. The spinnaker project. *Proceedings of the IEEE*, 102:652, 2014.
- [143] D.A.B. Miller. Attojoule optoelectronics for low-energy information processing and communications. *J. Lightwave Technol.*, 35:346, 2017.
- [144] J.L. Hennessy and D.A. Patterson. *Computer Architecture*. Elsevier. Appendix F.
- [145] K.A. Boahen. Point-to-point connectivity between neuromorphic chips using address events. *IEEE Tran. Circ. Sys. II*, 47:416, 2000.
- [146] J. Bigelow. *Physical and Physiological Information Processes and Systems*. unknown.
- [147] 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.
- [148] J. Hasler and B. Marr. Finding a roadmap to achieve large neuromorphic hardware systems. *Front. Neurosci.*, 7:118, 2013.
- [149] J.W. Goodman, A.R. Dias, and L.M. Woody. Fully-parallel, high-speed incoherent optical method for performing discrete Fourier transforms. *Opt. Lett.*, 2:1, 1978.
- [150] R. Shubert and J.H. Harris. Optical surface waves on thin films and their application to integrated data processors. *IEEE Trans. Microwave Theory Tech.*, MTT-16:1048, 1968.
- [151] D.B. Anderson, J.J. Boyde, M.C. Hamilton, and R.R. August. An integrated optical approach to the Fourier transform. *IEEE J. Quant. Electron.*, QE-13:268, 1977.
- [152] H.M. Gibbs. *Controlling Light with Light*. Academic Press, 1985.
- [153] H.M. Gibbs, S.L. McCall, and T.N.C. Venkatesan. Differential gain and bistability using a sodium-filled fabry-perot interferometer. *Phys. Rev. Lett.*, 36:1135, 1976.
- [154] D.A.B. Miller, S.D. Smith, and A. Johnston. Optical bistability and signal amplification in a semiconductor crystal: applications of new low-power nonlinear effects in insb. *Appl. Phys. Lett.*, 35:658, 1979.
- [155] J.L. Jewell, Y.H. Lee, J.F. Duffy, A.C. Gossard, and W. Wiegmann. Parallel operation and crosstalk measurements in GaAs etalon optical logic devices. *Appl. Phys. Lett.*, 48:1342, 1986.
- [156] T. Venkatesan, B. Wilkens, Y.H. Lee, M. Warren, G. Olbright, N. Peyghambarian, J.S. Smith, and A.Y. Yariv. Fabrication of arrays of GaAs optical bistable devices. *Appl. Phys. Lett.*, 48:145, 1986.
- [157] D.A.B. Miller, D.S. Chemla, T.C. Damen, A.C. Gossard, W. Wiegmann, T.H. Wood, and C.A. Burrus. Novel hybrid optically bistable switch: The quantum well self-electro-optic effect device. *Appl. Phys. Lett.*, 45:13, 1984.
- [158] D.A.B. Miller, D.S. Chemla, T.C. Damen, A.C. Gossard, W. Wiegmann, T.H. Wood, and C.A. Burrus. Band-edge electroabsorption in quantum well structures: The quantum confined stark effect. *Phys. Rev. Lett.*, 53:2173, 1984.
- [159] A.L. Lentine, H.S. Hinton, D.A.B. Miller, J.E. Henry, J.E. Cunningham, and L.M.F. Chirovsky. Symmetric self-electrooptic effect device: Optical set-reset latch. *Appl. Phys. Lett.*, 52:1419, 1988.
- [160] A.L. Lentine, H.S. Hinton, D.A.B. Miller, J.E. Henry, J.E. Cunningham, and L.M.F. Chirovsky. Symmetric self-electrooptic effect device: Optical set-reset latch, differential logic gate, and differential modulator/detector. *IEEE J. Quant. Electron.*, 25:1928, 1989.
- [161] F.B. McCormick, T.J. Cloonan, F.A.P. Tooley, A.L. Lentine, J.M. Sasian, J.L. Brubaker, R.L. Morrison, S.L. Walker, R.J. Crisci, R.A. Novotny, S.J. Hinterlong, H.S. Hinton, and E. Keerbis. Six-stage digital free-space optical switching network using symmetric self-electro-optic-effect devices. *Applied Optics*, 32:5153, 1993.



- [162] M.F. Yanik, S. Fan, and M. Soljačić. All-optical transistor action with bistable switching in a photonic crystal cross-waveguide geometry. *Opt. Lett.*, 28:2506, 2003.
- [163] J. Hwang, M. Potoschnig, R. Lettow, G. Zumofen, A. Renn, S. ötzinger, and V. Sandoghdar. A single-molecule optical transistor. *Nature*, 460:76, 2009.
- [164] H. Wang, J. Wu, J. Guo, L. Jiang, Y. Xiang, and S. Wen. Low-threshold optical bistability with multilayer graphene-covering otto configuration. *J. Phys. D: Appl. Phys.*, 49:255306, 2016.
- [165] J. Guo, B. Ruan, J. Zhu, X. Dai, Y. Xiang, and H. Zhang. Low-threshold optical bistability in a metasurface with graphene. *J. Phys. D: Appl. Phys.*, 50:434003, 2017.
- [166] D. Sun, H. Zhang, and H. Sun. Controllable optical bistability and multistability in asymmetric quantum wells via fano-type interference. *J. Phys. B: Atomic, Molecular and Optical Physics*, 52:035501, 2019.
- [167] M. Sahrai, S. Ebadollahi-Bakhtevan, H. Tajalli, and A. Vafafard. High-speed electro-optical switching in an ingaasp/inp quantum well nanostructure. *Mater. Res. Express*, 5:115902, 2018.
- [168] J.-B. Li, S. Liang, S. Xiao, M.-D. He, L.-H. Liu, J.-H. Luo, and L.-Q. Chen. A sensitive biosensor based on optical bistability in a semiconductor quantum dot-dna nanohybrid. *J. Phys. D: Appl. Phys.*, 52:035401, 2019.
- [169] D.A.B. Miller. Are optical transistors the logical next step? *Nature Photonics*, 4:3, 2010.
- [170] R. Soref and J.P. Lorenzo. Single-crystal silicon: A new material for 1.3 and 1.6  $\mu\text{m}$  integrated-optical components. *Electronics Letters*, 21:953, 1985.
- [171] R. Soref and J.P. Lorenzo. All-silicon active and passive guided-wave components for  $\lambda = 1.3$  and 1.6  $\mu\text{m}$ . *IEEE J. Quant. Electron.*, QE-22:873, 1986.
- [172] R. Soref and B. Bennett. Electrooptical effects in silicon. *IEEE J. Quant. Electron.*, QE-23:123, 1987.
- [173] R. Soref. Silicon-Based Optoelectronics. *Proceedings of the IEEE*, 81:1687, 1993.
- [174] A.W. Snyder and J.D. Love. *Optical Waveguide Theory*. Chapman and Hall Ltd., 1983.
- [175] R.G. Hunsperger. *Integrated Optics: Theory and Technology*. Springer-Verlag, 2009.
- [176] G.K. Celler and S. Cristoloveanu. Frontiers of silicon-on-insulator. *J. Appl. Phys.*, 93:4955, 2003.
- [177] A. Liu, R. Jones, L. Liao, D. Samara-Rubio, D. Rubin, O. Cohen, R. Nicolaescu, and M. Paniccia. A high-speed silicon optical modulator based on a metal-oxide-semiconductor capacitor. *Nature*, 427:615, 2004.
- [178] L. Liao, D. Samara-Rubio, M. Morse, A. Liu, and D. Hodge. High speed silicon mach-zehnder modulator. *Opt. Express*, 13:3129, 2005.
- [179] Q. Xu, B. Schmidt, S. Pradhana, and M. Lipson. Micrometre-scale silicon electro-optic modulator. *Nature*, 435:325, 2005.
- [180] 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.
- [181] G. Reed, G. Mashanovich, F.Y. Gardes, and D.J. Thomson. Silicon optical modulators. *Nat. Photon.*, 4:518, 2010.
- [182] J. Witzens. High-Speed Silicon Photonics Modulators. *Proceedings of the IEEE*, 106:2158, 2018.
- [183] D. Patel, A. Samani, V. Veerasubramanian, S. Ghosh, and D.V. Plant. Silicon photonic segmented modulator-based electro-optic DAC for 100 Gb/s PAM-4 generation. *IEEE Photonics Technology Letters*, 27:2433, 2015.
- [184] J.S. Orcutt, B. Moss, C. Sun, J. Leu, M. Georgas, J. Shainline, J. Sun M. Weaver E. Zraggen, H. Li, S. Urosević, M. Popović, R.J. Ram, and Vladimir Stojanović. Open foundry platform for high-performance electronic-photonic integration. *Opt. Express*, 20:12222, 2012.
- [185] J.M. Shainline, J.S. Orcutt, , M.T. Wade, K. Nam-mari, B. Moss, M. Georgas, C. Sun, Rajeev J. Ram, V. Stojanović, and M.A. Popović. Depletion-mode carrier-plasma optical modulator in zero-change advanced cmos. *Opt. Lett.*, 38:2657, 2013.
- [186] L. Alloatti, D. Cheian, and R.J. Ram. High-speed modulator with interleaved junctions in zero-change cmos photonics. *Appl. Phys. Lett.*, 108:131101, 2016.
- [187] L. Alloatti and R.J. Ram. Resonance-enhanced waveguide-coupled silicon-germanium detector. *Appl. Phys. Lett.*, 108:071105, 2016.
- [188] J.M. Shainline, J.S. Orcutt, , M.T. Wade, K. Nam-mari, O. Tehar-Zahav, Z. Sternberg, R. Meade, Rajeev J. Ram, V. Stojanović, and M.A. Popović. Depletion-mode polysilicon optical modulators in a bulk complementary metal-oxide semiconductor process. *Opt. Lett.*, 38:2729, 2013.

- [189] 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.
- [190] J.M. Shainline and J. Xu. Silicon as an emissive optical medium. *Laser and Photonics Reviews*, 1:334, 2007.
- [191] A.F.J. Levi. Silicon photonics’ last-meter problem. *IEEE Spectrum*, 09.18:39, 2018.
- [192] D.G. Rabus. *Integrated Ring Resonators*. Springer-Verlag, 2007.
- [193] Y. Ohmachi and J. Noda. Electrooptic light modulator with branched ridge waveguide. *Appl. Phys. Lett.*, 27:544, 1975.
- [194] C.M. Verber. Integrated-Optical Approaches to Numerical Optical Processing. *Proc. IEEE*, 72:942, 1984.
- [195] X. Li, N. Youngblood, C. Ríos, Z. Cheng, C.D. Wright, W.H.P. Pernice, and H. Bhaskaran. Fast and reliable storage using a 5 bit, nonvolatile photonic memory cell. *Optica*, 6:1, 2018.
- [196] D. Psaltis, D. Casaent, D. Neft, and M. Carlotto. Accurate Numerical Computation by Optical Convolution. *Proc. SPIE*, 232:151, 1980.
- [197] R. Arrathoon and M.H. Hassoun. Optical threshold logic elements for digital computation. *Opt. Lett.*, 9:143, 1984.
- [198] H. Kogelnik. Limits in Integrated Optics. *Proc. IEEE*, 69:232, 1981.
- [199] S. Jutamulia and F.T.S. Yu. Overview of hybrid optical neural networks. *Optics and Laser Technology*, 28:59, 1996.
- [200] D.Z. Anderson. Coherent optical eigenstate memory. *Opt. Lett.*, 11:56, 1986.
- [201] T. Jannson, H.M. Stoll, and C. Karaguleff. The interconnectability of neuro-optic processors. *SPIE Real Time Signal Processing IX*, 698:157, 1986.
- [202] K. Wagner and D. Psaltis. Multilayer optical learning networks. *Applied Optics*, 26:5061, 1987.
- [203] D. Psaltis, D. Brady, and K. Wagner. Optical neural computers. *Sci. Am.*, 256:88, 1987.
- [204] J. Jang, S. Jung, S. Lee, and S. Shin. Optical implementation of the Hopfield model for two-dimensional associative memory. *Opt. Lett.*, 13:248, 1988.
- [205] A. Agranat, C.F. Neugebauer, and A. Yariv. Parallel optoelectronic realization of neural networks models using cid technology. *Applied Optics*, 27:4354, 1988.
- [206] N.H. Farhat, S. Miyahara, and K.S. Lee. Optical analog of two-dimensional neural networks and their application in recognition of radar targets. In *Neural Networks for Computing*, page 146. Amer. Inst. Phys., 1986.
- [207] H.J. Caulfield. Parallel  $n^4$  weighted optical interconnections. *Applied Optics*, 26:4039, 1987.
- [208] A.D. Fisher, W.L. Lippincott, and J.N. Lee. Optical implementation of associative networks with versatile adaptive learning capabilities. *Applied Optics*, 26:5039, 1987.
- [209] B. Macukow and H.H. Arsenault. Optical associative memory model based on neural networks having variable interconnection weights. *Applied Optics*, 26:924, 1987.
- [210] I. Saxena and E. Fiesler. Adaptive multilayer optical neural network with optical thresholding. *Opt. Engineering*, 34:2435, 1995.
- [211] G. Moagar-Poladian. Reconfigurable optical neuron based on photoelectret materials. *Applied Optics*, 39:782, 2000.
- [212] K. Wagner and D. Psaltis. Optical neural networks: an introduction by the feature editors. *Applied Optics*, 32:1261, 1993.
- [213] Y. Nitta, J. Ohta, S. Tai, and K. Kyuma. Optical learning neurochip with internal analog memory. *Applied Optics*, 32:1264, 1993.
- [214] J. Bueno, S. Maktoobi, L. Froehly, I. Fischer, M. Jacquot, L. Larger, and D. Brunner. Reinforcement learning in a large-scale photonic recurrent neural network. *Optica*, 5:756, 2018.
- [215] R. Hamerly, A. Sludds, L. Bernstein, M. Soljacic, and D. Englund. Large-scale optical neural networks based on photoelectric multiplication. *arXiv*, page 1812.07614, 2018.
- [216] X. Lin, Y. Rivenson, N.T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan. All-optical machine learning using diffractive deep neural networks. *Science*, 361:1004, 2018.
- [217] J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Sci. Rep.*, 8:12324, 2018.
- [218] P.R. Prucnal and B.J. Shastri. *Neuromorphic photonics*. CRC Press, New York, first edition, 2017.

- [219] M. Reck, A. Zeilinger, H.J. Bernstein, and P. Bertani. Experimental Realization of Any Discrete Unitary Operator. *Phys. Rev. Lett.*, 73:58, 1994.
- [220] G. Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 2016.
- [221] D.A.B. Miller. Perfect optics with imperfect components. *Optica*, 2:747, 2015.
- [222] 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, 2017.
- [223] T.W. Hughes, M. Minkov, Y. Shi, and S. Fan. Training of photonic neural networks through in situ back-propagation and gradient descent. *Optica*, 5:864, 2018.
- [224] I.A.D. Williamson, T.W. Hughes, M. Minkov, B. Bartlett, S. Pai, and S. Fan. Reprogrammable electro-optic nonlinear activation functions for optical neural networks. *arXiv*, 1903:04579, 2019.
- [225] P.W. Smith. Hybrid bistable optical devices. *Opt. Eng.*, 19:456, 1980.
- [226] J. Chiles, S. Buckley, N. Nader, S.W. Nam, R.P. Mirin, and J.M. Shainline. Multi-planar amorphous silicon photonics with compact interplanar couplers, cross talk mitigation, and low crossing loss. *APL Photonics*, 2:116101, 2017.
- [227] J. Chiles, S.M. Buckley, S.W. Nam, R.P. Mirin, and J.M. Shainline. Design, fabrication, and metrology of 10 x 100 multi-planar integrated photonic routing manifolds for neural networks. *APL Photonics*, page 106101, 2018.
- [228] W.D. Sacher, J.C. Mikkelsen, P. Dumais, J. Jiang, D. Goodwill, X. Luo, Y. Huang, Y. Yang, A. Bois, P.G.-Q. Lo, E. Bernier, and J.K.S. Poon. Tri-layer silicon nitride-on-silicon photonic platform for ultra-low-loss crossings and interlayer transitions. *Opt. Express*, 25:30862, 2017.
- [229] G. Mourgas-Alexandris, A. Tsakyridis, N. Passalis, A. Tefas, K. Vysokinos, and N. Pleros. An all-optical neuron with sigmoid activation function. *Opt. Express*, 27:9620, 2019.
- [230] M. Lukoševičius and H. Jaeger. Reservoir computing approaches to recurrent neural network training. *Comput. Sci. Rev.*, 3:127, 2009.
- [231] G. Tanaka, T. Yamane, J.B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, and A. Hirose. Recent advances in physical reservoir computing: a review. *Neural Networks*, 115:100, 2019.
- [232] G. Van der Sande, D. Brunner, and M.C. Soriano. Advances in photonic reservoir computing. *Nanophotonics*, 6:561, 2017.
- [233] K. Vandoorne, W. Dierckx, B. Schrauwen, D. Verstraeten, R. Baets, P. Bienstman, and J. Van Campenhout. Toward optical signal processing using Photonic Reservoir Computing. *Opt. Express*, 16:11182, 2008.
- [234] K. Vandoorne, J. Dambre, D. Verstraeten, B. Schrauwen, and P. Bienstman. Parallel reservoir computing using optical amplifiers. *IEEE Trans. Neural Networks*, 22:1469, 2011.
- [235] T.J. Johnson, M. Borselli, and O. Painter. Self-induced optical modulation of the transmission of a high- $q$  silicon microdisk resonator. *Optics Express*, 14:817, 2006.
- [236] K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman. Experimental demonstration of reservoir computing on a silicon photonics chip. *Nat. Comm.*, 5:3541, 2014.
- [237] F. Denis-Le Coarer, M. Sciamanna, A. Kattumba, M. Freiberger, J. Dambre, P. Bienstman, and D. Rontani. All-optical reservoir computing on a photonic chip using silicon-based ring resonators. *IEEE J. Sel. Topics in Quant. Electron.*, 24:7600108, 2018.
- [238] L. Appeltant, M.C. Soriano, G. Van der Sande, J. Canckaert, S. Massar, J. Dambre, B. Schrauwen, C.R. Mirasso, and I. Fischer. Information processing using a single dynamical node as a complex system. *Nature Communications*, 2:468, 2011.
- [239] L. Larger, M.C. Soriano, D. Brunner, L. Appeltant, J.M. Gutierrez, L. Pesquera, C.R. Mirasso, and I. Fischer. Photonic information processing beyond turing: an optoelectronic implementation of reservoir computing. *Opt. Express*, 20:3241, 2012.
- [240] S. Ortín, M.C. Soriano, L. Pesquera, D. Brunner, D. San-Martín, I. Fischer, C.R. Mirasso, and J.M. Gutiérrez. A unified framework for reservoir computing and extreme learning machines based on a single time-delayed neuron. *Sci. Rep.*, 5:14945, 2015.
- [241] D. Brunner and I. Fischer. Reconfigurable semiconductor laser networks based on diffractive coupling. *Opt. Lett.*, 40:3854, 2015.

- [242] D. Brunner, M.C. Soriano, C.R. Mirasso, and I. Fischer. Parallel photonic information processing at gigabyte per second data rates using transient states. *Nat. Comm.*, 4:1364, 2013.
- [243] J. Bueno, D. Brunner, M.C. Soriano, and I. Fischer. Conditions for reservoir computing performance using semiconductor lasers with delayed optical feedback. *Opt. Express*, 25:2401, 2017.
- [244] Y. Paquot, F. Duport, A. Smerieri, J. Dambre, B. Schrauwen, M. Haelterman, and S. Massar. Optoelectronic Reservoir Computing. *Sci. Rep.*, 2:287, 2012.
- [245] F. Duport, B. Schneider, A. Smerieri, M. Haelterman, and S. Massar. All-optical reservoir computing. *Opt. Express*, 20:22783, 2012.
- [246] P. Antonik, M. Haelterman, and S. Massar. Brain-Inspired Photonic Signal Processor for Generating Periodic Patterns and Emulating Chaotic Systems. *Phys. Rev. App.*, 7:054014, 2017.
- [247] F. Coarer, M. Sciamanna, A. Katumba, M. Freiburger, J. Dambre, P. Bienstman, and D. Rontani. All-Optical Reservoir Computing on a Photonic Chip Using Silicon-Based Ring Resonators. *IEEE J. Sel. Topics Quant. Electron.*, 24:7600108, 2018.
- [248] B. Romeira, J.M.L. Figueredo, and J. Javaloyes. Delay dynamics of neuromorphic optoelectronic nanoscale resonators: Perspectives and applications. *Chaos*, 27:114323, 2017.
- [249] J.L.A. Dubbeldam, B. Krauskopf, and D. Lenstra. Excitability and coherence resonance in lasers with saturable absorber. *Phys. Rev. E*, 60:6580, 1999.
- [250] S.B. Colak, J.J.H.B. Schleipen, and C.T.H. Lienenbaum. Neural network using longitudinal modes of an injection laser with external feedback. *IEEE Trans. Neural Networks*, 7:1389, 1996.
- [251] E.C. Mos, J.J.H.B. Schleipen and H. de Waardt, and G.D. Khoe. Optical mode neural network by use of the nonlinear response of a laser diode to external optical feedback. *Appl. Opt.*, 36:6654, 1997.
- [252] E.C. Mos, J.J.L. Hoppenbrouwers, M.T. Hill, M.W. Blüm, J.J.H.B. Schleipen, and H. de Waardt. Optical neuron by use of a laser diode with injection seeding and external optical feedback. *IEEE Trans. Neural Networks*, 11:988, 2000.
- [253] F.C. Hoppensteadt and E.M. Izhikevich. Synchronization of laser oscillators, associative memory, and optical neurocomputing. *Phys. Rev. E*, 62:4010, 2000.
- [254] A.R.S. Romariz and K.H. Wagner. Tunable vertical-cavity surface-emitting laser with feedback to implement a pulsed neural model. 1. Principles and experimental demonstration. *Applied Optics*, 46:4736, 2007.
- [255] A.R.S. Romariz and K.H. Wagner. Tunable vertical-cavity surface-emitting laser with feedback to implement a pulsed neural model. 2. High-frequency effects and optical coupling. *Applied Optics*, 46:4746, 2007.
- [256] A. Hurtado, I.D. Henning, and M.J. Adams. Optical neuron using polarisation switching in a 1550nm-VCSEL. *Opt. Express*, 18:25170, 2010.
- [257] D. Rosenbluth, K. Kravtsov, M.P. Fok, and P.R. Prucnal. A high performance photonic pulse processing device. *Opt. Express*, 17:22767, 2009.
- [258] R.J. Vogelstein, U. Mallik, J.T. Vogelstein, and G. Cauwenberghs. Dynamically reconfigurable silicon array of spiking neurons with conductance-based synapses. *IEEE Trans. Neural Networks*, 18:253, 2007.
- [259] S. Barbay, R. Kuszelewicz, and A.M. Yacomotti. Excitability in a semiconductor laser with saturable absorber. *Opt. Lett.*, 36:4476, 2011.
- [260] M.A. Nahmias, B.J. Shastri, A.N. Tait, and P.R. Prucnal. A leaky integrate-and-fire laser neuron for ultrafast cognitive computing. *IEEE J. Sel. Top. Quant. Electron.*, 19:1800212, 2013.
- [261] W. Coomans, L. Gelens, S. Beri, J. Danckaert, and G. Van der Sande. Solitary and coupled semiconductor ring lasers as optical spiking neurons. *Phys. Rev. E*, 84:036209, 2011.
- [262] K. Alexander, T. Van Vaerenbergh, M. Fiers, P. Mechet, J. Dambre, and P. Bienstman. Excitability in optically injected microdisk lasers with phase controlled excitatory and inhibitory response. *Opt. Express*, 21:26182, 2013.
- [263] F. Selmi, R. Braive, G. Beaudoin, I. Sagnes, R. Kuszelewicz, and S. Barbay. Relative refractory period in an excitable semiconductor laser. *Phys. Rev. Lett.*, 112:183902, 2016.
- [264] F. Selmi, R. Braive, G. Beaudoin, I. Sagnes, R. Kuszelewicz, and S. Barbay. Temporal summation in a neuromimetic micropillar laser. *Opt. Lett.*, 40:5690, 2015.
- [265] A. Hurtado and J. Javaloyes. Controllable spiking patterns in long-wavelength vertical cavity surface emitting lasers for neuromorphic photonics systems. *Appl. Phys. Lett.*, 107:241103, 2015.

- [266] C. Mesaritakis, A. Kapsalis, A. Bogris, and D. Syvridis. Artificial neuron based on integrated semiconductor quantum dot mode-locked lasers. *Sci. Rep.*, 6:39317, 2018.
- [267] P.R. Prucnal, B.J. Shastri, T. Ferreira de Lima, M.A. Nahmias, and A.N. Tait. Recent progress in semiconductor excitable lasers for photonic spike processing. *Advances in Optics and Photonics*, 8:228, 2016.
- [268] B.J. Shastri, M.A. Nahmias, A.N. Tait, B. Wu, and P.R. Prucnal. Simpel: Circuit model for photonic spike processing laser neurons. *Opt. Express*, 23:8029, 2015.
- [269] B.J. Shastri, M.A. Nahmias, A.N. Tait, A.W. Rodriguez, B. Wu, and P.R. Prucnal. Spike processing with a graphene excitable laser. *Nature Sci. Rep.*, 6:19126, 2016.
- [270] M. Naruse, Y. Terashima, A. Uchida, and S.-J. Kim. Ultrafast photonic reinforcement learning based on laser chaos. *Sci. Rep.*, 7:8772, 2017.
- [271] Z. Cheng, C. Rios, W.H.P. Pernice, C.D. Wright, and H. Bhaskaran. On-chip photonic synapse. *Science Advances*, 3:1700160, 2017.
- [272] I. Chakraborty, G. Saha, A. Sengupta, and K. Roy. Toward fast neural computing using all-photonic phase change spiking neurons. *Scientific Reports*, 8:12980, 2018.
- [273] I. Chakraborty, G. Saha, and K. Roy. Photonic in-memory computing primitive for spiking neural networks using phase-change materials. *Phys. Rev. Applied*, 11:014063, 2019.
- [274] A.N. Tait, M.A. Nahmias, B.J. Shastri, and P.R. Prucnal. Broadcast and weight: an integrated network for scalable photonic spike processing. *J. Lightwave Technol.*, 32:3427, 2014.
- [275] A.N. Tait, T. Ferreira de Lima, E. Zhou, A.X. Wu, M.A. Nahmias, B.J. Shastri, and P.R. Prucnal. Neuromorphic photonic networks using silicon photonic weight banks. *Sci. Rep.*, 7:7430, 2017.
- [276] A.N. Tait, J. Chang, B.J. Shastri, M.A. Nahmias, and P.R. Prucnal. Demonstration of wdm weighted addition for principal component analysis. *Opt. Express*, 23:12758, 2015.
- [277] A.N. Tait, T. Ferreira de Lima, M.A. Nahmias, H.B. Miller, H.-T. Peng, B.J. Shastri, and P.R. Prucnal. A silicon photonic modulator neuron. *Phys. Rev. Applied*, 11:064043, 2019.
- [278] Y.A. Vlasov and S.J. McNab. Losses in single-mode silicon-on-insulator strip waveguides and bends. *Opt. Express*, 12:1622, 2004.
- [279] M. Lipson. Guiding, modulating, and emitting light on silicon-challenges and opportunities. *J. Lightwave Technology*, 23:4222, 2005.
- [280] D. Liang and J.E. Bowers. Recent progress in lasers on silicon. *Nature Photonics*, 4:511, 2010.
- [281] Z. Zhou, B. Yin, and Jurgen Michel. On-chip light sources for silicon photonics. *Light: science and applications*, 4:e358, 2015.
- [282] Gordon Davies. The optical properties of luminescence centres in silicon. *Physics Reports*, 176:83–188, 1989.
- [283] Y. Zhou, W. Dou, W. Du, S. Ojo, H. Tran, S.A. Ghetmiri, J. Liu, G. Sun, R. Soref, J. Margetis, J. Tolle, B. Li, Z. Chen, M. Mortasavi, and S.-Q. Yu. Optically pumped GeSn lasers operating at 270 k with broad waveguide structures on si. *ACS Photonics*, 6:1434, 2019.
- [284] H. Ennen, G. Pomrenke, A. Axmann, K. Eisele, W. Haydl, and J. Schneider. 1.54  $\mu$  m electroluminescence of erbium-doped silicon grown by molecular beam epitaxy. *Appl. Phys. Lett.*, 46:381, 1985.
- [285] T.G. Brown and D.G. Hall. Observation of electroluminescence from excitons bound to isoelectronic impurities in crystalline silicon. *J. Appl. Phys.*, 59:1399, 1986.
- [286] H.-T. Peng, M.A. Nahmias, T. Ferreira de Lima, A.N. Tait, B.J. Shastri, and P.R. Prucnal. Neuromorphic Photonic Integrated Circuits. *IEEE J. Sel. Topics Quant. Electron.*, 24:6101715, 2018.
- [287] T. Ferreira de Lima, B.J. Shastri, A.N. Tait, M.A. Nahmias, and P.R. Prucnal. Progress in neuromorphic photonics. *Nanophotonics*, 6:577, 2017.
- [288] K. Wu, C. Soci, P.P. Shum, and N.I. Zheludev. Computing matrix inversion with optical networks. *Opt. Express*, 22:295, 2014.
- [289] C. Mesaritakis, A. Bogris, A. Kapsalis, and D. Syvridis. High-speed all-optical pattern recognition of dispersive fourier images through a photonic reservoir computing subsystem. *Opt. Lett.*, 40:3416, 2015.
- [290] D.A. Buck. The cryotron—a superconductive computer component. *Proc. IRE*, 44:482, 1956.
- [291] <https://spectrum.ieee.org/tech-history/heroic-failures/dudley-bucks-forgotten-cryotron-computer>. *IEEE Spectrum*.
- [292] B.D. Josephson. Possible new effects in superconductive tunneling. *Physics Letters*, 1:251, 1962.
- [293] M. Tinkham. *Introduction to Superconductivity*. Dover, second edition, 1996.

- [294] T. Van Duzer and C.W. Turner. *Principles of superconductive devices and circuits*. Prentice Hall, USA, second edition, 1998.
- [295] Alan M. Kadin. *Introduction to superconducting circuits*. John Wiley and Sons, USA, first edition, 1999.
- [296] K.K. Likharev. Superconductor digital electronics. *Physica C*, 482:6, 2012.
- [297] W. Anacker. Josephson computer technology: an ibm research project. *IBM Journal of research and development*, 24:107, 1980.
- [298] 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.
- [299] P. Bunyk, K. Likharev, and D. Zinoviev. Rsfq technology: physics and devices. *Int. J. High Speed Electron. Sys.*, 11:257, 2001.
- [300] K. Aihara et al. *Extended Abstracts The Japan Society of Applied Physics and Related Societies*, pages I-77, 1989.
- [301] T. Ogata et al. *Extended Abstracts The Japan Society of Applied Physics and Related Societies*, pages I-77, 1989.
- [302] Y. Harada and E. Goto. Artificial neural network circuits with josephson devices. *IEEE Trans. Magnetics*, 27:2863, 1991.
- [303] M. Hidaka and L.A. Akers. An artificial neural cell implemented with superconducting circuits. *Supercond. Sci. Technol.*, 4:654, 1991.
- [304] Y. Mizugaki, K. Nakajima, Y. Sawada, and T. Yamashita. Implementation of new superconducting neural circuits using coupled squids. *IEEE Trans. Appl. Supercond.*, 4:1, 1994.
- [305] Y. Mizugaki, K. Nakajima, Y. Sawada, and T. Yamashita. Implementation of superconducting synapses into a neuron-based analog-to-digital converter. *Appl. Phys. Lett.*, 65:1712, 1994.
- [306] Y. Mizugaki, K. Nakajima, Y. Sawada, and T. Yamashita. Superconducting neural circuits using squids. *IEEE Trans. Appl. Supercond.*, 5:3168, 1995.
- [307] E.D. Rippert and S. Lomatch. A multilayered superconducting neural network implementation. *IEEE Trans. Appl. Supercond.*, 7:3442, 1997.
- [308] T. Kondo, M. Kobori, T. Onomi, and K. Nakajima. Design and implementation of stochastic neurosystem using sfq logic circuits. *IEEE Trans. Appl. Supercond.*, 15:320, 2005.
- [309] T. Onomi, T. Kondo, and K. Nakajima. Implementation of high-speed single flux-quantum up/down counter for the neural computation using stochastic logic. *IEEE. Trans. Appl. Supercond.*, 19:626, 2009.
- [310] T. Hirose, T. Asai, and Y. Amemiya. Spiking neuron devices consisting of single-flux-quantum circuits. *Physica C*, 445:1020, 2006.
- [311] T. Hirose, T. Asai, and Y. Amemiya. Pulsed neural networks consisting of single-flux-quantum spiking neurons. *Physica C*, 463:1072, 2007.
- [312] T. Onomi, Y. Maenami, and K. Nakajima. Superconducting neural network for solving a combinatorial optimization problem. *IEEE. Trans. Appl. Supercond.*, 21:701, 2011.
- [313] Y. Yamanashi, K. Umeda, and N. Yoshikawa. Pseudo sigmoid function generator for a superconductive neural network. *IEEE. Trans. Appl. Supercond.*, 23:1701004, 2013.
- [314] P. Crotty, D. Schult, and K. Segall. Josephson junction simulation of neurons. *Phys. Rev. E*, 82:011914, 2010.
- [315] 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.
- [316] 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.
- [317] A.E. Schegolev, N.V. Klenov, I.I. Soloviev, and M.V. Tereshonok. Adiabatic superconducting cells for ultra-low-power artificial neural networks. *Beilstein Journal of Nanotechnology*, 7:1397, 2016.
- [318] N.V. Klenov, A.E. Schegolev, I.I. Soloviev, S.V. Bakurskiy, and M.V. Tereshonok. Energy efficient superconducting neural networks for high-speed intellectual data processing systems. *IEEE. Trans. Appl. Supercond.*, 28:1301006, 2018.
- [319] I.I. Soloviev, A.E. Schegolev, N.V. Klenov, S.V. Bakurskiy, M.Y. Kupriyanov, M.V. Tereshonok, A.V. Shadrin, V.S. Stolyarov, and A.A. Golubov. Adiabatic superconducting artificial neural network: basic cells. *J. Appl. Phys.*, 124:152113, 2018.
- [320] H. Katayama, T. Fujii, and N. Hatakenaka. Theoretical basis of squid-based artificial neurons. *J. Appl. Phys.*, 124:152106, 2018.

- [321] 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.
- [322] 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.
- [323] 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.
- [324] M.L. Schneider, C.A. Donnelly, and S.E. Russek. Tutorial: high-speed low-power neuromorphic systems based on magnetic josephson junctions. *J. Appl. Phys.*, 124:161102, 2018.
- [325] R. Cheng, U.S. Goteti, and M.C. Hamilton. Spiking neuron circuits using superconducting quantum phase-slip junctions. *J. Appl. Phys*, 124:152126, 2018.
- [326] S.K. Tolpygo. Superconductor digital electronics: scalability and energy efficiency issues. *Low Temperature Physics*, 42:361, 2016.
- [327] J.M. Shainline, S.M. Buckley, R.P. Mirin, and S.W. Nam. Superconducting optoelectronic circuits for neuromorphic computing. *Phys. Rev. App.*, 7:034013, 2017.
- [328] J.M. Shainline, S.M. Buckley, A.N. McCaughan, J. Chiles, A. Jafari-Salim, R.P. Mirin, and S.W. Nam. Circuit designs for superconducting optoelectronic loop neurons. *J. Appl. Phys.*, page 152130, 2018.
- [329] J.M. Shainline, S.M. Buckley, A.N. McCaughan, J. Chiles, A. Jafari-Salim, R.P. Mirin, and S.W. Nam. Superconducting optoelectronic neurons. *arXiv*, page 1805.01929, 2018.
- [330] O. Kahl, S. Ferrari, V. Kovalyuk, G.N. Goltsman, A. Korneev, and W.H.P. Pernice. Waveguide integrated superconducting single-photon detectors with high internal quantum efficiency at telecom wavelengths. *Sci. Rep.*, 5:10941, 2015.
- [331] F. Marsili, V.B. Verma, J.A. Stern, S. Harrington, A.E. Lita, T. Gerrits, I. Vayshnker, B. Baek, M.D. Shaw, R.P. Mirin, and S.W. Nam. Detecting single infrared photons with 93% system efficiency. *Nat. Photon.*, 7:210, 2013.
- [332] J.M. Shainline, S.M. Buckley, N. Nader, C.M. Gentry, K.C. Cossel, J.W. Cleary, M. Popović, N.R. Newbury, S.W. Nam, and R.P. Mirin. Room-temperature-deposited dielectrics and superconductors for integrated photonics. *Opt. Express*, page 10322, 2017.
- [333] D.A.B. Miller. Device Requirements for Optical Interconnects to Silicon Chips. *Proc. IEEE*, 97:1166, 2009.
- [334] D. Huang, T. Sze, A. Landin, R. Lytel, and H.L. Davidson. Optical Interconnects: Out of the Box Forever? *IEEE J. Sel. Topics Quant. Electron.*, 9:614, 2003.
- [335] J. Wang and Y. Long. On-chip silicon photonic signaling and processing: a review. *Science Bulletin*, 63(19):1267–1310, oct 2018.
- [336] H.M. Ozaktas. Paradigms of connectivity for computer circuits and networks. *Optical Engineering*, 31:1563, 1992.
- [337] J. Kates-Harbeck, A. Svyotkovskiy, and W. Tang. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature*, 568:526, 2019.
- [338] C. Granade, J. Combes, and D.G. Cory. Practical bayesian tomography. *New J. Phys.*, 18:033024, 2016.
- [339] G.I. Struchalin, I.A. Pogorelov, S.S. Straupe, K.S. Kravtsov, I.V. Radchenko, , and S.P. Kulik. Experimental adaptive quantum tomography of two-qubit states. *Phys. Rev. A*, 93:012103, 2016.
- [340] A.A. Clerk, M.H. Devoret, S.M. Girvin, F. Marquardt, , and R.J. Schoelkopf. Introduction to quantum noise, measurement, and amplification. *arXiv*, 0810:4729, 2010.
- [341] A. Opremcak, I.V. Pechenezhskiy, C. Howington, B.G. Christensen, M.A. Beck, and R. McDermott. Measurement of a superconducting qubit with a microwave photon counter. *Science*, 361:1239, 2018.
- [342] K.R.W. Jones. Principles of quantum inference. *Annals of Physics*, 140:1239, 1991.
- [343] K.R.W. Jones. Fundamental limits upon the measurement of state vectors. *Phys. Rev. A*, 50:3682, 1994.
- [344] R. Derka, V. Buzek, G. Adam, and P.L. Knight. From quantum bayesian inference to quantum tomography. *arXiv*, page 9701029, 1997.
- [345] R. Schack, T.A. Brun, and C.M. Caves. Quantum bayes rule. *Phys. Rev. A*, 64, 2001.

- [346] R. Blume-Kohout. Optimal, reliable estimation of quantum states. *New. J. Phys.*, 12, 2010.
- [347] F. Huszar and N.M.T. Houlsby. Adaptive bayesian quantum tomography. *Phys. Rev. A*, 85, 2012.
- [348] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, , and G. Carleo. Neural-network quantum state tomography. *Nat. Phys.*, 14, 2018.
- [349] W.J. Ma, J.M. Beck, P.E. Latham, and A. Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9:1432, 2006.
- [350] T. Yang and M.N. Shadlen. Probabilistic reasoning by neurons. *Nature*, 447:1075, 2007.
- [351] J.M. Beck, W.J. Ma, R. Kiani, T. Hanks, A.K. Churchland, J. Roitman, M.N. Shadlen, P.E. Latham, and A. Pouget. Probabilistic population codes form bayesian decision making. *Neuron*, 60:1142, 2008.
- [352] K.W. Murch, S.J. Weber, C. Macklin, and I. Siddiqi. Observing single quantum trajectories of a superconducting quantum bit. *Nature*, 502, 2013.
- [353] M.J. Barber, J.W. Clark, and C.H. Anderson. Neural representation of probabilistic information. *Neural Computation*, 15, 2006.
- [354] C. Wetterich. Quantum computing with classical bits. *arXiv*, 1810:05960, 2018.
- [355] C. Pehle, K. Meier, M. Oberthaler, and C. Wetterich. Emulating quantum computation with artificial neural networks. *arXiv*, 1810:10335, 2018.
- [356] G. Carleo and M. Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355, 2017.
- [357] 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. Efficiently measuring a quantum device using machine learning. *arXiv*, 1810:10042, 2018.
- [358] N. Bostrom. *Superintelligence*. Oxford University Press, 2014.
- [359] H.K. Onnes. The resistance of pure mercury at helium temperatures. *Commun. Phys. Lab. Univ. Leiden*, 12:1, 1911.
- [360] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436, 2015.
- [361] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.
- [362] D. Verstraeten, B. Schrauwen, M. D’Haene, and D. Stroobandt. An experimental unification of reservoir computing methods. *Neural Networks*, 20:391, 2007.
- [363] B. Scellier and Y. Bengio. Equilibrium Propagation: Bridging the Gap between Energy-Based Models and Backpropagation. *Frontiers in Computational Neuroscience*, 11:1, 2017.
- [364] S. Dehaene, H. Lau, and S. Kouider. What is consciousness, and could machines have it? *Science*, 358:486, 2017.