Light in neural systems

Jeffrey M. Shainline

National Institute of Standards and Technology
Boulder, CO, 80305

December 13, 2018

Abstract

1 Introduction

Light is excellent for communication. If we are going to signal to extraterrestrial civilizations, it will almost certainly be with electromagnetic radiation. On our own planet, fiber optic links carry vast quantities of information across continents and between data centers. An important question in modern computing is: what is the shortest distance over which photonic communication is sufficiently advantageous and practical to merit displacement of electronic interconnects? Optical links between racks in data centers are becoming common. Major companies are investing seriously in photonics in the package. Monolithic optical links between processor and memory fabricated in a 45-nm CMOS node with no in-line changes have been demonstrated [**?**], with integration in 32-nm technology looking promising as well [**?**]. A primary challenge affecting further chip-scale electronic-photonic integration is the continued difficulty of achieving a light source implemented on silicon that is robust, efficient, and economical.

In parallel with the hardware considerations affecting optoelectronic integration are questions related to the future of computer architecture. A prominent theme emerging since clock speed leveled off in 2003 [] is parallelism. Computation is increasingly distributed among more processor cores. Many-core architectures continue to expand into complex on-chip networks (OCNs), in some cases resulting in highly distributed, brain-inspired systems []. As compute grows more distributed, communication demands more from interconnect networks. The demand for energy efficient communication bandwidth has been a major driver of on-chip photonics.

The major drivers for brain-inspired computers fall on a spectrum: energy and algorithmic efficiency for deployable applications (Internet of things, self-driving cars, mobile devices) reside on one side of the spectrum, and artificial general intelligence (AGI) resides on the other. Knowledge gained from neuroscience informs us that systems with general intelligence will benefit from very large numbers of computational elements as well as extreme communication between them. It is our perspective that hardware incorporating light for communication between electronic computational elements combined with an architecture of distributed optoelectronic spiking neurons will provide tremendous potential for AGI. Considerations pertinent to the realization of such a technology are the subject of this article.

2 Guided by neuroscience and VLSI

To guide the design of hardware for AGI, we must simultaneously consider the principles of neural information and principles of fabrication for very-large-scale integration. Regarding the principles of neural information, we know that computation in the brain makes efficient use of space and time by leveraging fractal scaling. [recap of general principles]

Based on these considerations, we expect a hardware platform capable of AGI to display at least six traits:

1. Plasticity mechanisms (excitation, inhibition, stdp, meta, homeostatic)
2. Dendritic nonlinearities (sequence detection and integration)
3. Spiking neuronal dynamics
4. Massive connectivity to enable short path lengths across big networks
5. The ability to efficiently use space from the scale of a single synapse up to systems limited by light-speed communication, and the ability to efficiently use time with network oscillations across a wide range of frequencies
6. Energy consumption and power density low enough to enable scaling to systems of this size

Regarding the fabrication of large systems, we assume success developing AGI is most likely if the infrastructure developed for digital computing with silicon electronics can be utilized. This leads us to at least five traits that hardware must display if extreme scalability is to be achieved:

1. Most likely based on performing lithography on silicon wafers
2. Patterning should not require features smaller than 193 nm lithography can achieve
3. Materials that are rare, volatile, or incompatible with conventional processing should be avoided
4. It is unlikely with any artificial technology that a human-level intelligence will fit on a single wafer. Straightforward assembly into multi-wafer modules should be achievable
5. For very large systems with photonic connectivity, the ability to leverage fiber optic technology will likely be highly advantageous

3 Optoelectronic synapses, dendrites, and neurons

Having made the choice to communicate with light, one must devise a means to perform the necessary device functions of neural systems. One such function is establishment of a synaptic weight. Synaptic weights are the primary form of long-term memory in neural systems, and they change the strength of connections between neurons [**?**, **?**]. In optoelectronic hardware leveraging light for communication, there are two general approaches to change the synaptic weight between two neurons: 1) in the photonic domain with variable attenuation of an optical signal; or 2) in the electronic domain with a variable electronic response upon detection of an optical pulse. This choice has several important ramifications for hardware and information processing. Regarding information processing, it is usually assumed that neural communication is digital: the presence or absence of an action potential is a binary one or zero, and the amplitude of the action potential is not encoding information. When adjusting the synaptic weight in the photonic domain, this is not the case. The number of photons reaching a neuron through a synaptic connection becomes an analog variable, and it is subject to shot noise, in addition to any noise mechanisms present in the detector. The signal-to-noise ratio of shot noise improves with $\sqrt{(N_{\mathrm{ph}})}$, where in this case N_{ph} is the average number of photons, so establishing weights in the photonic domain introduces an energy/noise tradeoff. Setting weights in the photonic domain also has the disadvantage that photons are discarded by attenuation at weak synaptic weights. Thus, by setting synaptic weights in the photonic domain, we place a burden on light sources to produce large numbers of photons to minimize shot noise, and we discard photons when they are attenuated at weak synapses. In this mode of operation, light is used for communication, but it is also used for the important computational operation of applying the synaptic weight.

By contrast, if we establish synaptic weights in the electronic domain, light is used exclusively for communication, and communication remains entirely digital. The presence of an optical signal can be used to represent an all-or-none communication event. In this case, the detector and associated electronics must be able to achieve a variable synaptic response to identical photonic pulses based on the configuration of the electronic aspects of the circuits. In this case, we expect that a neuron will send, on average, N_{ph} photons to each of its downstream synaptic connections. Due to shot noise, each downstream connection will receive $N_{\mathrm{ph}} \pm \sqrt{N_{\mathrm{ph}}}$ photons, and the detector circuit must be configured to implement a synaptic response if a threshold of N_{th} photons is detected. After detection, the electronic response must vary depending on the synaptic weight, independently of the precise number of photons that was detected. It is in this electronic response that the signal becomes analog again. Whereas setting the synaptic weights in the photonic domain places a larger burden on light sources, setting the synaptic weights in the electronic domain places a larger burden on detector circuits. One must achieve a detector circuit that converts light pulses to electrical current or voltage, and the amount of electrical signal must be largely independent of the number of photons in the pulse, depending instead on reconfigurable electrical properties of the circuit, such as bias currents or voltages. These reconfigurable bias currents or voltages then represent the synaptic weights, and the task of a neuron's light source is simply to provide a roughly constant number of photons to each of its downstream synaptic connections. For energy efficiency, the number of photons necessary to evoke a synaptic response from the detector (N_{th}) should be made as low as possible to make the job of the light source as easy as possible. N_{th} cannot be made lower than one, as the electromagnetic field is quantized into integer numbers of photons.

To reach this energy efficiency limit while setting the synaptic weight in the electronic domain, we have proposed utilizing optoelectronic circuits combining superconducting-nanowire single-photon detectors (SPDs) [**?**] in conjunction with Josephson junctions (JJs) [**?**] as synaptic receivers. These circuits have several desirable properties. First, they have very near zero dark counts, and their response is nearly identical whether they receive one or more than one photon within a short time window, so the quantity N_{th} can be made as low as physically possible. Second, because the circuit is not attempting to resolve the number of photons present in a pulse, shot noise on the photonic communication signals does not propagate beyond the synapse. Instead, the average number of photons arriving at a synapse simply must be high enough that the probability of a synapse receiving zero photons becomes tolerably small ($N_{\mathrm{ph}} = 5$ gives a 1% chance of receiving zero photons). Third, the synaptic weight in this circuit is set by a current bias across the Josephson junction, which has nearly zero effect on the behavior of the SPD, and can be straightforwardly adjusted to achieve a wide range of synaptic weights. The effect of this current bias is to change how much current gets added to the neurons integrated signal when a photon is detected. Fourth, in addition to the energy benefits derived from signaling with order one photon, superconducting detectors dissipate zero power in the steady state. Amidst these benefits lies a challenge: to realize these synapses, one must integrate SPDs with JJs in a scalable hardware platform. It is our perspective that such hardware will bring significant capabilities for AGI, in part due to the efficient, low-noise operation of these single-photon optoelectronic synapses.

Following the question of whether synaptic weights are set in the photonic or electronic domain lies the question of how the synaptic weights are modified. Synaptic weight modification is a primary means of signal processing and learning in neural systems. In machine learning, synaptic weights are often trained through backpropagation wherein the output of a network is compared to a desired

output, and synaptic weights are updated to minimize a cost function. For AGI systems interacting with complex, dynamic environments, cost functions associated with all inputs to the network generally cannot be defined, so supervised learning algorithms such as backpropagation are conceptually poorly matched to the learning requirements. Further, brain-scale systems can be expected to employ 10^{14} synapses connected in highly recurrent graphs, so scalable learning must be accomplished by mechanisms local to the synapses, and cannot afford to rely on an external supervisor making contact to each synapse independently. From neuroscience we know that a number of local synaptic plasticity mechanisms can enable the network to learn over time. These plasticity mechanisms include spike-timing-dependent plasticity (STDP), short-term plasticity, homeostatic plasticity, and metaplasticity. We now briefly discuss each of these mechanisms in the context of optoelectronic hardware.

In STDP, the timing between pulses from the pre-synaptic neuron and post-synaptic neuron change the strength of the synaptic connection. If a pulse from the pre-synaptic neuron arrives at the synapse within a short time window before a pulse from the post-synaptic neuron, it is inferred that the pulse from the pre-synaptic neuron contributed to the firing of the post-synaptic neuron, and the synaptic weight strengthens. This is referred to as Hebbian update. On the other hand, if a pulse from the pre-synaptic neuron arrives at the synapse just after a pulse from the post-synaptic neuron, the pulse from the pre-synaptic neuron may be arriving during the refractory period of the post-synaptic neuron, and is therefore not contributing to the activity of the post-synaptic neuron. The synaptic weight thus becomes weaker in a process referred to as anti-Hebbian update. Synapses capable of Hebbian and anti-Hebbian functions can perform STDP. When designing optoelectronic synapses for AGI systems, the ability to perform these operations based only on activity at each synapse is vital. For applications in deep learning, interferometric networks have been proposed to leverage light for matrix-vector multiplication []. In such an implementation, the phases applied throughout the network determine the synaptic weights, but changing a single phase changes multiple synaptic weights, in general. Thus, there appears to be no way to utilize this approach to achieving synaptic weights to implement STDP. Another approach to setting synaptic weights in the photonic domain utilizes variable attenuation of a phase-change material []. In this case, the absorption of photons can modify the atomic configuration of the material, providing a means by which the presence of optical signals can reduce material absorption and thereby modify the synaptic weight. It has been shown that this technique can be used to achieve Hebbian learning wherein the simultaneous arrival of photons from pre-synaptic and post-synaptic neurons reduces attenuation and increases synaptic weight. The materials demonstrated to date require billions of photons for this operation, so energy efficiency is a concern for scaling. Additionally, while Hebbian strengthening is straightforward with this approach, it may not be possible to decipher the order of arrival of the pulses, so anti-Hebbian weakening and full STDP may not be possible.

STDP is an important mechanism that enables learning based on local activity of two neurons. In addition to this mechanism that modifies synaptic memory, it has been found that the ability to change not only synaptic weights, but also the rate at which synaptic weights are modified is an important capability to enable simultaneous acquisition of new knowledge and long-term memory retention []. Mechanisms that change the rate of synaptic weight update are referred to as metaplasticity []. Incorporating metaplasticity into artificial cognitive systems is necessary to enable the system to maintain a robust representation of past experiences while learning from a constantly changing environment. In optoelectronic hardware, the device-level operation that accomplishes metaplasticty will depend on how one has chosen to establish the synaptic weight in the first place. In the case of loop neurons the amount a synaptic weight is incremented during an STDP event depends on a bias current across a JJ, just as the synaptic weight itself depends on a bias current across a JJ. Metaplasticity is accomplished by modifying this synaptic update bias current, again based on detection of photons generated during neuronal firing events indicating the level of network activity. Like other operations in loop neurons, metaplastic update can be achieved with SPDs working in conjunction with JJs: photon detection events generate electrical current that changes the state of the synapse to make STDP more or less substantial.

In addition to STDP and metaplasticity, which affect synaptic weights and their rates of change over time scales long to relative to the time between pulses in a train (inter-spike interval), short-term plasticity is a crucial aspect of the behavior of a synapse that affects synaptic behavior on a time scale of the inter-spike interval. Short-term plasticity serves to filter input pulse trains. Short-pass, long-pass, and band-pass behavior have all been observed in biological neural systems. Short-pass filtering of spike trains causes the synapse to fire only at the beginning of an afferent spike train, while long-pass filtering leads to a synapse becoming active only after an multiple pulses of a train have been received. Band pass filtering ensures that a synapse does not respond to the first few pulses in a train (rising edge of the signal), but becomes responsive during a number of pulses in the middle of a sequence, before saturating and again going quiet during the final pulses of a train (falling edge of the signal). These rapid modifications of synaptic behavior are crucial to provide a neuron with a rich picture of the temporal activity input to its synapses, yet such complex behavior is difficult to achieve for many methods of establishing synaptic weights in photonic systems. It is our perspective that achieving STDP, metaplasticity, short-term plasticity, and other adaptive synaptic properties is significantly facilitated if electronic

circuits perform the adaptation, affecting the response of the electrical circuit to the detection of photonic communication events. In loop neurons, all these functions appear straightforward in simulations of superconducting circuits of modest complexity. Similar functions may be possible using CMOS devices in conjunction with semiconductor photodetectors. Whether at room-temperature or 4 K, optoelectronic synapses wherein electronic circuits receive photonic communication events and adapt synaptic behavior based on these optical signals are necessary for enabling complexity in intelligent optoelectronic neural systems. Light is excellent for communication, but electronics excel at computation.

While much computation in neurons occurs in synapses, the nonlinear response of the neuron itself is the primary computation performed on the inputs from all the neuron's synapses. In spiking neurons, this nonlinearity manifests as the production of a pulse when the integrated signal reaches threshold. Spiking neurons perform the a rate-in/rate-out transfer function. The leak rate of the integrated signal determines the turn on input rate, and the refractory period (reset time) determines the roll-over point of this transfer function. Thus, in addition to the nonlinearity associated with reaching threshold and producing a spike, neurons exhibit a nonlinear spike rate transfer function.

As emphasized above in the discussion of neuronal avalanches, neural systems display activity with structure across a wide range of temporal scales. Activity at lower frequencies will generally encompass information from a larger region of the network, while activities at higher frequencies will generally integrate information locally to build consensus amongst a smaller population of the network. Achieving a short refractory period enables fast oscillations, and is therefore important to achieving rapid interpretation of new stimulus. Photonic systems such as excitable lasers can achieve extremely fast reset times, and have thus received significant attention as relaxation oscillators for spiking neural systems [**?**, **?**]. However, it remains important for a neuron to maintain information regarding stimulus received in the past related to low-frequency, network-wide activity, so the leak rate of the neuron should be made as long as possible. If one attempts to utilize an optical cavity as an integrator of photons received from upstream connections, the leak rate will be limited by the photon lifetime in the cavity, $\tau = Q/\omega_0$, where Q is the cavity quality factor, and ω_0 is the angular frequency of electromagnetic wave. Q of 10^6 is difficult to achieve in integrated laser systems, and even with this high-Q we find $\tau = 1$ ns for photons at 1550 nm, indicating that integrating photonic signals in the optical domain makes it very difficult for a neuron to retain information relevant to lower-frequency, larger-area network activity. The response of such a neuron is independent of anything that happened in the network longer ago than τ, and the neuron can not identify correlations between local, rapid activity, and broader, slower activity.

Like the case of synaptic computations and plasticity, we reach the conclusion that the neuronal computation of signal integration is best performed in the electronic domain. This operation requires that a photodetector operate in conjunction with an electronic circuit to transduce photon detection events to stored electrical charge or current. If semiconducting circuit infrastructure is employed, this stored electrical signal will likely be charge on a capacitor resulting in voltage, and when the accumulated voltage reaches a given threshold, the electrical circuit must generate (or modulate) light, presumably via a semiconductor diode. If superconducting circuit infrastructure is employed, the stored electrical signal is supercurrent circulating in a loop, and the threshold is set by the critical current of a JJ. This superconducting implementation has the advantage that the signal can be maintained indefinitely (provided superconductivity is maintained), so the integration time τ can be chosen to be as long as theoretical analysis deems appropriate, and can be implemented in hardware with the L/r time constant of the integration loop. The integration time of the neuron is completely decoupled from the threshold of the neuron as well as all other aspects of the circuit. The primary challenge of this superconducting loop neuron design is to achieve the voltage necessary to produce light from a semiconductor diode when threshold is reached [**?**]. Fortunately, significant experimental progress has recently been made demonstrating the feasibility and efficiency of such an operation [**?**].

In addition to these synaptic and neuronal computations, the significance of intermediate dendritic nonlinear processing is increasingly being emphasized by the neuroscience community [**?**].

Dendrites:

- Dendritic nonlinearities can in principle be in either photonic of electronic domain. Photonic domain, nonlinearities require quite a bit of power (saturable absorbers, frequency conversion)

Neurons:

- Spiking neural dynamics requires that, upon reaching threshold, pulses of light must be produced
- this has led some to propose excitable lasers as neurons. light detected by photodetector, fan-in, time constants, power required
- integration should be performed in the electronic domain, and upon reaching an electronic threshold, a voltage pulse can be produced to either generate light or modulate an externally provided cw light stream
- for communication across a branching network of passive photonic waveguides, coherence of light source is not required
- choice of light source is crucial (on or off chip, material, wdm or not, wavelength)

discussion of frequency (each neuron cannot have its own frequency due to practical limitations, selection and routing of neuronal communication signals cannot be achieved by frequency alone)

4 Communication with guided light

The central premise of our work is that photonic signals are superior to electronic signals for communication across large-scale neural systems. To explain why we place this conjecture at the center of hardware development, we briefly summarize the physical limitations of electrical interconnection networks [**?**].

4.1 Electrical interconnection networks

When using silicon electronics, information is represented by voltages. A voltage is achieved physically by accumulating charge on a capacitor. For one device to communicate information to another device, it must provide charge in the form of electrical current. This current must charge up the capacitors representing information at the target device as well as the capacitance of the wire over which the charge is transferred. There are fundamental and practical limits to how small the capacitance of the devices and wires can be []. As a rule of thumb, a wire in a CMOS process adds 200 aF/µm, so parasitic wire capacitance dominates when devices are separated by even a few microns. It becomes impracticable for a single device to source current to many other devices, so in practice a single transistor rarely drives more than four other transistors (fanout of four). For many applications, it is crucial for each device to be able to send information to many more than four destinations. Because all the devices requiring communication with one another cannot be directly wired to each other, a shared communication network is employed. In contemporary computing, switched media networks are used for this purpose. Each device must then only communicate to the nearest switch in the network. In a switched media network, devices communicate with one another by sending packets of information. The packet contains routing information (the address of the recipient) as well as the data to be communicated. The interconnect network determines a valid route for the information traverse across the network (referred to as routing), and the switches are configured accordingly to achieve that physical route of information transfer.

Because the communication infrastructure is shared devices must request access to the switch network to transmit messages. Multiple devices may request access simultaneously, in which case arbitration must be performed. Arbitration refers to the process of granting devices access to the switch network, and in general a packet will experience some delay while it waits in a queue to be granted access to the shared communication infrastructure. This 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. For many applications this is adequate. 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. One consequence of this approach is that as the size of the system grows, more information in each communication event must be allocated to specifying destinations. This leads to increased burden on memory and processors. But the more severe challenge is introduced by the connectivity/speed trade-off. As more neurons, each with many synapses, are added to the network, the average frequency of neuronal firing events must decrease due to the limitations of the interconnection network to handle communication requests. For electronic systems with a few hundred thousand neurons, average event rates in the kilohertz range can be maintained []. Systems with a few hundred million neurons will likely be limited to operation at 10 Hz or below [].

4.2 Photonic interconnection networks

The requirement of utilizing a shared communication infrastructure results from the physics of electrons. Because electrons interact with each other due to their charge, they can be used to establish a voltage on a capacitor, which is useful for computing. But the charge-based interaction is also what makes it difficult to source sufficient current to directly communicate between many devices. The physics of light is complimentary. Photons, being uncharged bosons, interact with one another only through quantum interference, such as Hong-Ou-Mandel interference. Light-matter interaction is generally weak, especially at low light levels. Thus, photons can co-propagate on waveguides independently of one another without wiring capacitance. This enables a pulse of photons to fan out to many destinations without a charging penalty due to wiring. This is not to say photonic communication can address an arbitrarily large number of recipients without consequence. For each new recipient, the number of photons in the initial pulse must increase, and as destinations get further away, more energy is dissipated to propagation loss. These realities notwithstanding, it appears feasible for devices communicating with photons to make direct, physical, point-to-point connections to many thousands of destinations, thereby eliminating the need for shared communication

infrastructure that we see as the primary impediment to achieving AGI with electrical interconnections.

Having made this claim, the burden is upon us to provide evidence of the feasibility of photonic communication in large-scale neural systems. Much like electrical interconnection networks utilize different technologies at the scales of chips versus data centers, photonic interconnection technologies must enable communication from the scale of an on-chip network to a multi-wafer system the size of a data center. Further, the specific architectures suitable for neural information processing require that neurons be capable of communicating seamlessly across all scales of this network hierarchy. At the outset, the large wavelength of light relative to the size of electronic devices (and relative to the size of devices in the brain) lead us to be concerned for the size of brain-scale networks. To begin building confidence for the feasibility of the endeavor, we sketch a vision of how a general optoelectronic neural system may be constructed.

We emphasized above that a successful neural technology must leverage the fabrication infrastructure of silicon electronics. We therefore conjecture that silicon photonics technology will be utilized to move light between neurons. At the wafer scale, light will be guided in dielectric waveguides. Silicon photonics provides three primary dielectric materials that can be used for these passive waveguides: Si, SiN, and SiO_2. To two significant figures, the indices of refraction of these materials close to 1550 nm are 3.5, 2.0, and 1.5, respectively. Perhaps not coincidentally, these are the three primary dielectrics used in CMOS technology as well. The highest index contrast of these materials is achieved if we construct waveguides with Si for the core and SiO_2 for the cladding. This will allow the smallest waveguide pitch and tightest bends, enabling dense connectivity. The trade-off is that with higher index contrast comes higher Rayleigh scattering from waveguide line-edge roughness. Si waveguides can be expected to have 1 dB/cm propagation loss, so 30 dB attenutation would be incurred propagating across a wafer. It may be possible to simply turn up the power of light sources to accommodate this loss, but it is more likely that longer-distance connections will be achieved with SiN waveguides clad with SiO_2. SiN waveguides are likely to have 0.2 dB/cm propagation loss []. With this loss, 6 dB of loss would be incurred propagating across a 300 mm wafer. If this is too high, lower index contrast can be achieved with $SiON_x$ waveguides. Because we can choose to either turn of the power of the light sources while using smaller, high-index waveguides or save power by using smaller, low-index waveguides, we are faced with a power/size trade-off. Such a trade-off must be address after system-level considerations.

Achieving the dense, complex routing required to connect large numbers of neurons on a wafer will require multiple planes of waveguides, just as integrated electronics requires multiple wiring layers. We anticipate that optoelectronic neural systems will utilize dielectric waveguide layers deposited in the back-end-of-line in the fabrication process, with lower layers having higher index and being utilized for local connections, and higher layers having gradually lower index used for more distant connections. We wish to approximate the area of such photonic interconnection networks. Following Keyes [**?**], we approximate the area required for the waveguides entering a neuron as $A_\mathrm{n} = (n_\mathrm{in} w/k)^2$, where n_in is the number of waveguides entering the neuron (in-directed synaptic connections), w is the waveguide pitch, and k is the number of planes of waveguides. In general, w will depend on index contrast, but for this analysis we treat it as constant and estimate area by considering low- and high-index limits. As will be discussed shortly, for tiling multi-wafer assemblies, wafers diced into Octagons may be advantageous, so we take the area of a wafer to be $A_8 = 2\sqrt{2}r^2$ with $r = 150$ mm. The number of neurons that can be supported on a 300-mm wafer is given by the ratio,

$$N_8 = \frac{A_8}{A_\mathrm{n}} = 2\sqrt{2}r^2 \left(\frac{k}{w n_\mathrm{in}}\right)^2. \qquad (1)$$

This expression is plotted in Fig. 1. This estimate informs us that a 300-mm wafer can support roughly one million neurons if they each have one thousand connections. As a point of comparison, Ref. [**?**] finds that through multi-layer, wafer-scale integration of logic and memory, 250 million electrical neurons could fit on a 300-mm wafer. The trade-off is, of course, speed, as the shared communication network would limit the electrical neurons studied in Ref. **??** to 10 Hz operation. Nevertheless, the message of Fig. 1 is that photonic routing results in large area consumption. If two planes of routing waveguides are used, 100,000 neurons with 1000 connections each can fit on a wafer. With six planes, one million such neurons can fit on a wafer, and with 18 routing planes, the number is close to 10 million. The human cortex contains over 10 billion neurons. An optoelectronic brain larger than a bumble bee will not fit on a wafer.

Optoelectronic intelligence will require communication between wafers. This can be accomplished through multiple means. Most simply, wafers can be stacked vertically, and free-space optical links can send photons from a source on one wafer, directed vertically by a beam-shaping grating coupler, to a detector on a wafer above or below, as illustrated in Fig. 2(a). Such 3D techniques are being developed for electronics, but the ability of light to propagate through free space and the multi-micron alignment tolerances enabled by wide-area photodetectors [**?**] make such 3D integration feasible for photonic communication as well. Assuming the photodetectors receiving vertical communication have a pitch of 25 µm, a 300-mm octagon could support 10^8 vertical communication links between two wafers. Assuming half of each such pixel is for feed-forward communication from the lower wafer to the upper wafer, and half is for feed-back from the upper wafer to the lower wafer. This would result in 5×10^7 synaptic connections originating from neurons on one wafer and

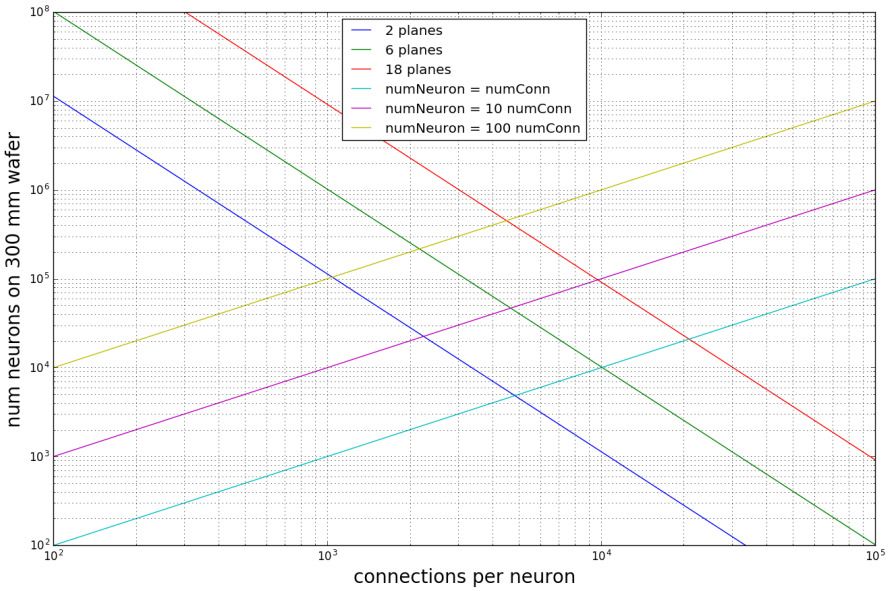

Figure 1: Number of neurons on a wafer.

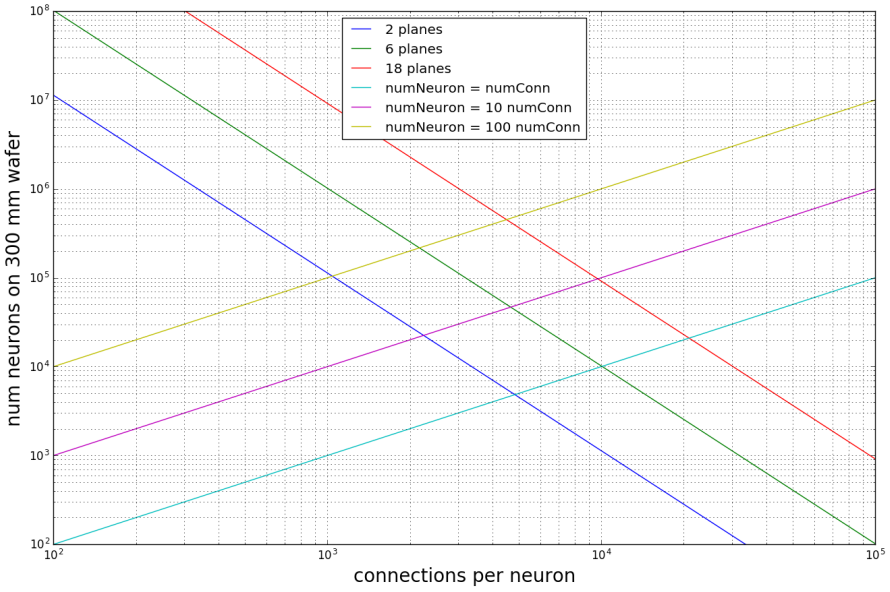

Figure 2: Photonic interconnection on various length scales.

terminating in neurons on a vertically adjacent wafer. If each wafer had one million neurons with one thousand connections per neuron within the wafer, the total number of intra-wafer synaptic connections would be 10^9. Therefore, the number of synapses present in a layer of this network that originated on a previous layer would be 5%, similar to the fraction observed in the laminar structure of biological cortex [**?**, **?**].

In addition to free-space vertical coupling, inter-wafer communication can be achieved at wafer edges with in-plane waveguide couplers, as shown in Fig. 2(c). In the octagonal (truncated square) tiling used here for illustration, each wafer could make such connections to neighbors in the cardinal directions. With a 10 µm pitch of these wafer-edge couplers, 11,500 could be supported in each of the cardinal directions with 46,000 total in-plane, lateral connections. Such a system would demonstrate strong connectivity within the vertical stack of the wafers, and weaker lateral connectivity. We begin to see that such an architecture very much resembles the columnar organization of cortex, perhaps even quantitatively in the degree of local, lateral, and vertical connections. There is much room for fine-tuning through device and system engineering. The columns of the brain typically have six distinct laminar layers for processing. It is somewhat unclear how the number of vertically integrated layers affect computation or what determines how many layers are suitable. It may be the case that timing errors on action potentials accumulate through layers, limiting the utility of additional feed-forward computation []. Presumably there will be some maximum number of vertically integrated wafers that is useful for neuronal information processing, and this modular extent of vertical wafer stacking must be considered in conjunction mechanical considerations regarding the structure of the assembly.

The wafer tiling we have just described leads to a picture of optoelectronic networks with vertically stacked columns of wafers with horizontal connectivity around the wafer perimeter. To achieve communication from within these columns to other (perhaps distant) regions of the network, optical fibers are ideal. Within the truncated square tiling under consideration, the square areas at diagonals between wafers can support fiber-optic bundles. These optical fiber tracts would be analogous to white matter in the brain. One such region could house roughly a million standard single-mode fibers of 125 µm diameter. These fibers will emanate from all wafers within the column, so the number of outputs available to each wafer will depend on how many vertically integrated wafers are utilized in a column. To be concrete, assume six wafers are stacked in a column, each wafer would have roughly 167,000 output fibers to carry information to distant regions of the network. With one million neurons on a wafer, this would mean not every neuron on the wafer would be able to couple to a fiber for long-distance communication. This again is consistent with brain organization wherein the number of long-distance axons emanating from a region is smaller than the number of neurons within the region.

4.3 Scaling to large systems

The analogies with biological neural systems has been emphasized: densely connected neurons on wafers can be arranged in columns, comprising the grey matter responsible for computation. Waveguides on the wafer and optical fibers between wafers comprise the white matter supporting communication. Notice that we have arrived at this construction based on no specific assumptions about the operation of optoelectronic neurons. We expect the intra- and inter-wafer architecture of optoelectronic neural systems to evolve toward a system like that illustrated in Fig. 2 based solely on the practical considerations of routing light from many sources to many destinations under the constraints of waveguide materials and wafer-scale fabrication. This analysis, while qualitative, elucidates the inevitability of discontinuities of Rentian scaling at certain boundaries. For example, the dense connectivity enabled by high-index dielectric waveguides on a wafer will result in a certain Rent exponent when analyzing only the sub-network contained on a wafer. This exponent may be close to unity, indicating the ability of the network to efficiently integrate information at the wafer scale. However, at the wafer boundary, connectivity is limited, and it will not be possible to maintain the same Rent exponent across partitions of the network containing multiple wafers. Instead, a distinct Rentian analysis is appropriate at this scale wherein it is more suitable to discuss the number of wafers within a partition rather than the number a neurons. A new Rentian exponent will characterize a range of spatial scales from a wafer to perhaps thousands of wafers. Perhaps this Rent exponent will again be close to unity, indicating the ability for each wafer to efficiently communicate with every other wafer within this module. Yet again at some point we will reach a limit, there will be a discontinuity in Rentian scaling, and it will be more appropriate to analyze partitions containing various numbers of these modules, each of which comprises thousands of wafers with millions of neurons per wafer.

This modular architecture appears to be inevitable for any computational entity occurring in nature, and it has consequences for how information is processed. The neuroscience community from Mountcastle to the present as well as the VLSI community across similar decades have grappled with the ramifications. Information processing must be simultaneously locally specialized (populations of neurons code for specific stimuli) and globally integrated (network-wide activity affects individual neural behavior). It is the simultaneous local processing of neuronal clusters combined with the global communication amongst neuronal populations that manifests as neuronal avalanches with power-law size distribution. Neuronal avalanches and fractal scaling indicate that increases in network capacity are possible without fundamental reorganization of the system [**?**]. This form of self-similar information process-

ing has no size limits, in principle. Ultimate limitations will be practical, and related to the inability to maintain continuous Rentian scaling across space and time. The most intelligent system will be the one that can integrate information most effectively across many scales of Rentian hierarchy. It is in this regard that optical communication has the most to offer. On a wafer, photonic fan-out across dielectric waveguides enables neurons to make thousands of direct connections without the limits of a shared switching network. Free-space and wafer-edge couplers enable significant inter-wafer communication conducive to columnar information processing. Such columns can communicate to one another locally and globally over fiber optic links.

With this configuration in mind, we can assess the feasibility of constructing systems on the scale of the human cortex, with 10 billion neurons, each with thousands of synaptic connections. If a wafer holds a million neurons, such an assembly requires 10,000 wafers and would easily fit in a volume a few meters on a side. Several aspects of this technology make the challenge likely to succeed. First and foremost, due to photonic signaing, it remains possible to achieve efficient communication across the network for systems with orders of magnitude more than 10,000 wafers. Information integration with photonic communication is physically possible. It is also practically possible, because all the proposed circuits (for loop neurons or otherwise) can be fabricated at the wafer scale with existing infrastructure (300 mm silicon-on-insulator substrates, 193 nm immersion lithography, equivalent of 45-nm CMOS node). A 45-nm CMOS foundry processes 10,000 wafers per day. If such a foundry were dedicated to fabrication of optoelectronic intelligence, it may be able to produce multiple brain-scale systems per year. Assembly of the wafers into a functional system would be difficult, but probably not much more difficult than the construction of a contemporary supercomputer, and likely much easier than a particle collider. The requirement for liquid-helium cooling does not appear to us to be a major impediment to implementation. From our perspective, the greatest unknown is the light source. If the silicon light sources that have already been demonstrated in cryogenic optical links [**?**] can be produced with internal quantum efficiency (η_{qe}) of 10^{-3}, we anticipate this project will be economically viable. At present, $\eta_{\mathrm{qe}} = 5 \times 10^{-7}$ has been demonstrated in the first attempt with no optimization. These light sources need not achieve high performance. We only require they achieve incoherent pulses of 10,000 photons ($\approx$ 1 fJ) at 20 MHz. If no silicon light source operating at 4 K can meet these criteria, and integration of III-V light sources on silicon wafers is required, the cost and complexity of a project at this scale may become prohibitive.

To summarize, we are asking the community to believe in a project that depends critically on wafer-scale integration, operation at 4 K, and silicon light sources. This is a lot to ask, but society has much to gain. Achieving truly intelligent artificial systems was never going to be easy. From our perspective, integrated optoelectronic hardware makes it possible.

5 Discussion

At present, the challenge of attaining an artificial intelligence rivaling a human appears formidable with the use of silicon electronics alone. The primary challenge arises because direct signaling between large numbers of neurons is not possible due to the charging requirements of wires and devices. Silicon electronic networks must use shared communication infrastructure, resulting in a connectivity/speed tradeoff. It is our perspective that the use of photonic communication will successfully mitigate this tradeoff, despite the increased size of photonic interconnect networks. Photonic fanout enables direct connections between large numbers of neurons, and the velocity of light enables communication across ten-meter systems before communication limits network speed.

Light is excellent for communication, while electronics excel at computation. It is our perspective that artificial neural hardware should be designed and constructed to maximally leverage photonic communication while performing synaptic, dendritic, and neuronal functions with electronic circuits for complexity of computation. From our perspective, superconducting optoelectronic circuits appear to naturally implement these functions, in part because light sources and detectors work so much better at low temperature. For mobile applications, superconducting circuits are irrelevant. But we think low-temperature (4 K) operation will make the creation of large-scale optoelectronic cognitive hardware easier when considering device to system scales. The construction of AGI will require many optoelectronic wafers with superconducting, semiconducting, and photonic components immersed in liquid helium. The fabrication infrastructure is the same as contemporary CMOS. The construction, facilities, and cost are likely comparable to a contemporary supercomputer. Understanding the principles of network informaiton processing and designing the architecture on the scale of an intelligent brain are the true grand challenges.

Amdahl's law draws our attention to the general principle that we should not try too hard to optimize one aspect of a system if performance will only be limited by another aspect. One is hyper aware of this principle when contemplating new hardware for neural systems, precisely because neural systems have intricacies that are interdependent across many functions and scales. Synapses must be designed with a variety of plasticity mechanisms in mind, while simultaneously ensuring nonlinear processing in dendrites and neurons. Computation must be considered alongside communication, and information integration across space must be considered simultaneously with information integration over time. A specific device or mechanism may appear suitable to perform a given function when that function is considered in isolation, but to be

successful in cognitive computing, each component must perform well in isolation and integrate well with the rest of the computational hierarchy. Speed matters, but it must be considered in the context of information propagation across the network. Extremely fast oscillators do not bring their full advantage if they cannot communicate effectively to large numbers of other oscillators or if they cannot retain information regarding the history of their inputs. Power consumption is significant, but one may be willing to burn more if performance can be substantially increased, provided the network can be adequately cooled. Size is important, but it must be compared to the distance signals can travel in the period of a network-wide oscillation cycle. The two attributes we find most fundamental are that neural systems require excellent communication, and neural devices require complex computation. Together, these requirements inform our perspective that advanced neural hardware requires optoelectronic integration.

Notes: Must explain: level of devices $\longrightarrow$ synaptic, dendritic, neuronal functionalities for enabling diverse dynamical states, extracting maximal information from spike trains, and efficient retrieval and storage of memories.

network level $\longrightarrow$ neuronal avalanches integrate information across space and time and rely on fractal scaling. for large-scale systems, this means fractal scaling must be supported across spatial and temporal scales. This requires efficient communication without activity-related bottlenecks or connectivity/speed tradeoffs. In the spatial domain, this is enabled by networks that maintain small-world architecture locall as well as globally. In the temporal domain, this is enabled by direct, point-to-point communication unburdened by a shared switching infrastructure. Neuron A must be able to spike and communicate to all of its connections at any time and at any frequency up to its device-limited maximum relaxation oscillation frequency, irrespective of the activity of any other Neuron B in the network.

Scale-free network activity is only possible from the smallest to largest scales (in space and time) allowed by hardware. It is our perspective that this range can be maximized if electronic and photonic physics are both utilized.

- general principles apply to any integrated photonic neural technology
- intra-wafer
  - dense routing on a chip or wafer is best accomplished with high-index-contrast dielectric waveguides
  - massive connectivity requires multiple planes of dielectric waveguides
  - index contrast can be generally lowered for larger reach connections with SiNO
- inter-wafer
  - wafer-scale fabrication will place a limit on the number of neurons that can be monolithically integrated.
  - inter-wafer communication must be possible.
  - free-space: beam divergence?
  - fiber optic white matter
- ultimate limits
  - at largest scale, all cognitive systems must be limited by the distance light can travel during a network oscillation at a given frequency
  - speed of light limits what can be thought in our universe
  - rentian scaling from die-sized neuronal clusters to planet-sized cognitive systems
  - most significant weakness of photonic systems is size, but ultimately more than compensated by the speed of light
- further comparison to biology
  - constituents of brain matter vs artificial (glial cells/silicon matrix, dendritic arbor/superconducting circuitry, axonal arbor/waveguide interconnects, spiking neurons/pulsing light sources, neuromodulators/control currents, water/He)
  - contrast axon scaling, delay with photonic system
  - complexity across spatial and temporal scales
  - system-wide design co-optimization
- what's next for photonic neural circuits

- similar to what is needed for optoelectronic integration for digital logic
- low-cost source-detector integration at the wafer scale
- improvements in BEOL photonic routing
- demonstration of photonic neural systems beyond a single variable attenuator or relaxation oscillator (Princeton leads the way)
- improved fiber-to-chip coupling and multi-chip/wafer modules
- scaling up: wafer-scale integration and beyond
- further theoretical analysis at device, circuit, and system levels
- similar to other artificial neural systems, need theoretical analysis to understand how to use these systems, train them, make them intelligent
- AGI requires significant hardware improvements, but hardware alone will not be smart. that requires insight from device to architecture, neuron to network, across space and time.

There are 10^{14} synapses in the human cortex. If it takes even 10 nW to maintain the state of each synapse, the system will consume a megawatt just to remember what it has learned. This is one reason superconducting synapses are attractive: they can maintain a memory indefinitely with no static power dissipation. Additionally, the strength of a synapse can be increased or decreased based on single-photon detection events. STDP appears possible with a single photon for each step of the update process.

Suppose we could get around fabrication considerations and construct arbitrary neural systems based on biological neurons. Would we then choose to pursue that technology for beyond-human intelligence instead of developing optoelectronic hardware? We think not, because the slow conduction velocity of axons presents a limit to communication, and this limit is likely already saturated near the scale of the human brain. If this argument is correct, we should not expect genetic engineering to achieve much smarter people than have already walked the earth.

$$P(s) = cs^{-\alpha} \tag{2}$$

$$e = cn^{-p} \tag{3}$$

$$d = \frac{1}{1-p} \tag{4}$$