Light in neural systems

Jeffrey M. Shainline

National Institute of Standards and Technology
Boulder, CO, 80305

December 12, 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 by far the most 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 [2], with integration in 32-nm technology looking promising as well [1]. 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.

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

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 processing 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.

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

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.

- 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. 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.

$$P(s) = cs^{-\alpha} \tag{1}$$

$$e = cn^{-p} \tag{2}$$

$$d = \frac{1}{1-p} \tag{3}$$

References

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

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