SemiSuperDraft

bryce.primavera1

August 2020

Abstract

1 Introduction

Recent years have seen a tremendous surge of interest in neuromorphic computing. Major chip makers have begun development of low-power neuromorphic chips for near-term deployment in edge AI applications. While the first practical applications for neuromorphic technology will be in small edge devices, the possibility of ultra large scale neuromorphic cognitive systems is tantalizing. Systems of such scale may be key to unraveling the mysteries of cognition, and some day might even be a platform for an artificial general intelligence that matches or even supersedes human capability.

With such lofty goals, it should not be surprising that the hardware for such massive systems will necessarily have significant deviations from current neuromorphic chips that target smaller scales. One of the most significant challenges in scaling up neuromorphic systems will be to design systems that efficiently enable the communication amongst billions or trillions of neurons with low latency. Most of the current neuromorphic systems use addressing of neurons as a way to time-multiplex spiking events amongst many neurons on to a small number of buses. This has been an extremely successful strategy and circumvents many of the physical issues that limit CMOS circuits to only a handful of outputs. In the near-term, addressed systems will continue to dominate the neuromorphic scene and their adaptability will continue to make them incredibly useful tools in testing hypotheses from the neuroscience and computer science communities. However, multiplexing inherently introduces trade-offs between connectivity and latency. Our contention is that multiplexing will be an untenable solution for the largest possible cognitive systems.

Optical communication is an extremely intriguing solution to the neural connectivity problem. Many of the same properties that have made light the medium of choice for communication in the largest artificial network yet constructed - the internet - make it attractive for large-scale neural systems. Unlike electrons, photons do not interact strongly with each other. This allows for very

low cross-talk and high fan-out. Additionally, waveguides and optical fibers enable low-loss communication that is nearly independent of the length of the link. These are excellent properties for implementing communication in large scale neural systems. While the lack of interaction between photons is a benefit for communication, it is a negative for implementing general computation. This has led to the hypothesis that the largest cognitive systems will be optoelectronic - utilizing optics for communication and electronics for computation.

There are a great variety of potential schemes for optoelectronic neuromorphic hardware. In this paper, we will restrict ourselves to a comparison of only two classes of networks designed around a couple basic assumptions:

1. Direct connections are superior to shared connections.

2. Optical communication should be binary.

These two conditions follow from the principles already mentioned. Optical communication is attractive precisely because it eliminates the need for shared communication lines by allowing for neurons with high fan-out. This means there will be no performance bottlenecks associated with network traffic and no overhead necessary to manage a massive address space. Additionally, when separate synapses are used for each connection, those synapses can be constructed with a wide variety of different properties. Creating many physical synapses for each neuron indeed costs substantial chip area, but in the supercomputing space that we are targeting, the physical size of the network is much less important than in the mobile applications that are often discussed in the neuromorphic community. The second constraint concerning binary optical communication stems from the hypothesis that computation is best done in the electronic domain. This also minimizes the amount of optical energy necessary per spiking event. If electrical to optical conversion is costly, then binary optical communication will be the most efficient solution.

With these general postulates established, a picture of the ideal hardware begins to emerge. There is a single optical transmitter at each neuron. This light emitter produces a short pulse of light each time the neuron spikes. The optical pulse is coupled into a waveguide and optical power is tapped off the waveguide for each downstream synapse. Each synapse contains a photodetector which registers an all or nothing spiking event. From there, all dendritic, weighting, summing, thresholding, and plasticity mechanisms are implemented in the electronic domain.

Following this model, a hardware platform known as SOENS (Superconducting OptoElectronic Networks) was proposed in 2016. SOENS exploits several exciting physical devices that are only possible at low temperatures. Superconducting Nanowire Single Photon Detectors (SNSPDs) are efficient, low-power detectors that allow the optical pulses reaching each synapse to be as faint as a few photons. Superconducting Josephson Junctions (JJs) provide an electronic circuit element that can compactly implement a wide variety neuronal computations. Lastly, the low temperature operation also makes integrated silicon light sources a viable option. An integrated silicon light source would make the

challenge of massive, wafer-level optoelectronic systems a much more realistic endeavour.

For all of its benefits, it is worth considering whether all of these exotic superconducting devices are truly superior to a more conventional semiconducting approach to optoelectronic neuromorphic hardware. The spirit of the hardware could potentially be preserved with a one-to-one replacement of each superconducting device with its semiconductor analog. Photodiodes could substitute for the SNSPDs. MOSFETs could play the role of JJs in implementing neuronal computation. Intriguingly, photodiodes and MOSFETs have none of the cryogenic constraints of their superconducting counterparts. Semiconductor systems can be envisioned that operate at 4K, 77K, or even room-temperature. Candidates for integrated light sources are available at all three of these suggested temperatures. This approach also clearly benefits from decades of development that have made the semiconductor industry what it is today.

This paper is essentially a work of prognostication, seeking to determine which platform - superconductors or semiconductors - will be superior for implementing ultra large scale optoelectronic neuromorphic hardware. Attempting to predict the future is frequently an act of folly, but it is an inherent element to every serious proposal of novel technology. To make as accurate an assessment as is currently feasible, the two platforms are compared from a myriad of different viewpoints - computational ability, power consumption, feasibility of fabrication, and economic viability.

2 Communication

2.1 Minimum Optical Signal

The energy necessary for each optical spike is clearly an important characteristic for the overall energy consumption of these systems. If the light sources are inefficient, then the electrical to optical conversion for each spike may very well be the dominant source of power consumption. This would make it imperative to use as low energy of an optical signal as possible.

In the superconducting case, the minimum optical energy is limited primarily by random optical shot noise due to the quantized nature of light. This is the classic quantum limit calculation that is often treated in optical communication texts. However, unlike in semiconductor receivers, the high sensitivity of SNSPDs and low noise of a cryogenic environment make this quantum limit much more than just a theoretical curiosity. It is the main factor that keeps the energy per spike in SOENs from reaching the ultimate physical limit of one photon per spike. Additionally, we will use a conservative estimate of $\eta_D = 70\%$ for the detection efficiency of the SNSPDs.

The probability of measuring zero photons for a spiking event during a certain time window is given by:

$$P(0) = \sum_{k=0}^{\infty} \frac{N_{ph}^k e^{-N_{ph}}}{k!}(1-\eta_D)^k = e^{-N_{ph}\eta_D} \tag{1}$$

N_{ph} is the average number of photons per spiking event. Neural systems are known for remarkable robustness to noise. Detecting only 95% of spikes may be tolerable. From equation 1, this would correspond to roughly 5 photons (0.5 aJ for $\lambda = 1.5$ um) needed to reach the receiver. Of course, the total number of photons needed to be produced by the source will need to be higher to account for energy losses in the link. The total energy per spike, E_{spike}, will be:

$$E_{spike} = \frac{N_{ph} hc}{\eta \lambda} \tag{2}$$

η is the total energy efficiency of the optical link. It includes all optical losses and the inefficiency of the light source. This efficiency factor will be highly dependent on the specifics of the platform, but for now we will leave it as a free variable.

The semiconductor case, on the other hand, would operate in an entirely different regime. Poisson noise will be entirely negligible. In fact, Miller argues that for extremely low energy optical links, detectors will operate above even the thermal noise limit. The question of a minimum optical signal in this case, is not one about overcoming noise, but about being able to simply provide enough charge to switch a transistor.[1] Following the derivation in ref [1], and assuming the ideal case of unity quantum efficiency for the photodiode (each photon absorbed adds one electron to a photocurrent charging a capacitive load), the number of photons needed to produce a voltage swing of ΔV across a capacitance of C_{tot} is:

$$\Delta V = \frac{q N_{ph}}{C_{tot}} \tag{3}$$

Clearly, small capacitances are essential to minimize N_{ph}. C_{tot} must include the capacitance of the photodetector, the transistor gate, and any wiring parasitics. An aggressive estimate for C_{tot} is 300 aF. This would correspond to $N_{ph} = 1500$ for a .8V voltage swing. This corresponds to about 200 aJ for $\lambda = 1.55$ um. Achieving such a level of sensitivity is speculative, but it represents a likely best-case scenario for the semiconductor approach.

This analysis demonstrates that the number of photons required for a detector to register a spike is nearly 400 times lower for the superconducting platform than an analogous semiconductor platform. Coincidentally, this factor is very similar to the COP (coefficient of performance) of large-scale liquid Helium refrigerators. Modern 4.2K liquid helium cryostats typically require around 360W

[1]This is sometimes referred to as the "receiverlesss" approach. In ref [1], it is argued that this approach will likely be superior to systems using noise-limited amplifier for on-chip communication. The power cost of transimpedance amplifiers that are commonly used in long-distance optical communication is too high for this communication regime.

of cooling power to remove 1W of device power. [2] This means that if two optical communications links were identical in all measures (source efficiency, optical losses, etc.) except one was cooled to 4K with SNSPDs and the other operated at room-temperature with photodiodes, then communicating a spike would cost nearly the same energy in each link.

2.2 Light Sources

The previous section suggests that the increased sensitivity of SNSPDs provides virtually no advantage over state-of-the-art photodiodes operated at room-temperature once cooling power is considered.

References

[1] Miller Attojoule https://ee.stanford.edu/ dabm/448.pdf

[2] https://arxiv.org/pdf/1501.07154.pdf just below table 7