Comparison of semiconducting and superconducting hardware for optoelectronic neuromorphic systems

Bryce Primavera and Jeff Shainline

National Institute of Standards and Technology
325 Broadway, Boulder, CO, USA, 80305

Tuesday 8th December, 2020

Abstract

Contents

1 Introduction

Lay out the problem:
- objective: neuromorphic supercomputing, scale of human brain
- machines that perform cognition
- not edge
- not focused on a single metric like speed or power, but rather on overall system scalability and performance
- seek the highest-performing artificial intelligence
- system considerations are paramount
- attempt to find physical limits of cognitive systems
- device features cannot be introduced if they are highly sensitive or require external tuning
- present-day neuromorphic cognitive systems using mosfets struggle in communication
- AER is a great way to make progress with existing hardware, but ultimately becomes a limiting factor
- optical communication may alleviate bottlenecks simply by avoiding wiring parasitics
- optical communication may enable the largest neuronal pool [1], which we assume correlates with high cognition
- we consider here an architecture with a single light source at each neuron
- this light source emits pulses playing the role of action potentials each time a neuron fires
- we do not consider any frequency or spatial multiplexing concepts from optical communications here,

as they tend to introduce requirements for precise device tolerances or active control of elements that are not scalable to the size of systems we seek
- action potentials are simply bursts of incoherent photons that a routed to all synaptic connections on a directional branching tree distribution network that taps off equal quantities of light to each synaptic connection
- photonic action potentials are received as binary communication signals
- synaptic weights and subsequent processing/computation is performed entirely in the electronic domain; light is used exclusively for communication
- the new challenge is that integrated light sources do not exist that can be placed at every neuron across a silicon wafer with modern VLSI technology
- the route to such an integrated device would be far easier if silicon light sources could be employed, an option if one accepts cryogenic operation
- here we assume such an option will be available at a future date
- the primary objective of the present study is to compare two approaches to the electronic circuits that would accompany such an optical communication network for large-scale cognitive systems
- the two approaches are semiconducting circuits and hybrid semiconducting/superconducting circuits
- in the semi case, photodetectors are waveguide-integrated semiconductor photodiodes, all computational circuits are based on mosfets, and light sources are waveguide-integrated semiconductor leds or lasers (focus on leds for processing/operation simplicity)
- in the super case, photodetectors are waveguide-integrated snspds, synaptic and dendritic circuits are based primarily on jjs coupled through mutual inductors, neuron circuits combine superconducting and semiconducting components, and the same light sources are employed
- we attempt to determine which of these approaches to neuromorphic hardware is likely to achieve superior cognitive performance by assessing their functionality in reference to established metrics from neuroscience and cognitive computing
- we describe these metrics in Sec. 2.
- we're not seeking biologically plausible time scales, but rather time scales from as long as possible to as fast as possible, working with the hypothesis that systems that have correlated dynamics and memory over as broad a range as possible will achieve optimal cognition

One way this paper could proceed is to compare mosfet-based circuits that implement the relevant operations as described in Ref. [2] with soens circuits as discussed in Refs. [3, 4]. Reference [2] describes mosfet circuits leveraging subthreshold operation principles to achieve short-term and long-term synaptic plasticity, adaptive threshold, spike production, and scaling to networks. Dendritic processing is not discussed. If we can do an excellent job understanding these circuits and determining how to couple the synaptic receivers to a photodetector and the spike production circuits to a light source, we should have most of what we need to do our comparison. The study could essentially be a comparison of the optoelectronic versions of the circuits in Ref. [2] with those in Refs. [3, 4]. A lot of the relevant work regarding soens has been done previously, but a few details may need to be considered, such as how the spiking activity of a neuron provides a feedback signal to all of its synapses for plasticity. Additional work may be required on the beyond-stdp learning rules.

Ref. [2] discusses hardware/software ecosystem. It appears likely that using photonic communication without AER will lead to significant reduction in complexity of network configuration, potentially significantly reducing the demands on software to control the system. This does come at the cost of reduced reconfigurability of networks.

The circuits in Ref. [2] are subthreshold current-mode circuits. Are there any basic physical statements we can make about using photodiodes (which are current-mode circuits) as inputs?

Attention should be paid to device mismatch and tolerances, aspects that affect subthreshold mosfets as well as jjs. Can we determine whether one system will suffer less than the other? Does it matter? Can networks of adaptive, spiking neurons compensate?

MOSFET model: using predictive technology model (PTM) from ASU.

It is not our intention to argue that certain technologies should not be investigated. At this early stage in the development of this complex field, many ideas should be pursued. It is also not clear that in the long-term technological limit, a single, primary technology will dominate. It is possible that many different approaches to adaptive, intelligent, neuromorphic systems will reach maturity and excel in different contexts. Most starkly, one does not expect devices and systems optimized for small-scale, power-starved, edge applications to take the same form as devices and systems optimized for extreme cognitive performance, considering the objectives and constraints are significantly divergent.

It is our intention instead to assess whether superconducting optoelectronic neural systems deserve to be in the pool of contenders as a hardware platform for achieving extreme cognitive performance. The concept begins from the zeroth order, intuitive (to us) conjectures that 1) optical communication is physically optimal at large scale, 2) few-photon communication is advantageous from a power perspective. First-order elucidation that JJ circuits with diverse circuit parameters, spiking and thresholding behaviors, and convenient activity-based adaptive responses led us to the state of tempered optimism that such hardware might be capable of realizing large-scale spiking neural networks, potentially capable of general intelligence. Yet advocating for hardware combining semiconductors, superconductors, and photonics positions one at the tri-critical point, balanced between compatriot and pariah in all three communities. It is our objective to assess whether the superconducting optoelectronic approach to large-scale neural sytems for artificial intelligence withstands this round of second-order analyses where we subject the concept to scrutiny from numerous angles and quantitatively compare its modeled performance relative to a semiconductor optoelectronic approach that may be more palatable to the community, perhaps by 1/3. Our position would surely be less lonely if we could motivate others to join the soens pursuit, and we hope the study presented here represents prudent analysis that can give us confidence to second order that we are on the right track and can be justified when encouraging others to try playing in this sandbox.

2 Device requirements for neuromorphic systems

- synaptic
- dendritic
- neuronal
- communication network

3 Light Sources

Short section, pointing out we're trying to consider systems with similar light sources, hopefully silicon, operated at low temp. Temp could be 77K or 40K or 4K, but we'll try to assume it doesn't matter.

4 Interconnection network and fan-out

Short section. Similarly to 3, we want to assume semi and super systems are using the same concepts and hardware for the interconnection network. Fan-out (out degree) is constrained to be the same for the two systems, so light sources and transmitter driver circuits must be scaled to produce the appropriate number of photons.

With such an optical communication scheme, large networks of neurons can be implemented in an asynchronous manner without address requirements that lead to memory challenges in large networks. Each communication channel from a neuron to a synapse has a dedicated waveguide connection, so no time-multiplexing is required. One waveguide leaves each neuron, and it branches as needed to achieve the desired graph structure.

5 Detectors

Important section.

- Analyze responsivity and noise to determine how many photons must be incident per pulse on snspd and photodiode to achieve a certain error ($< 1\%$, for example)
- Consider energy consumption of detector per pulse
- also any quiescent power consumption
- area
- fabrication
- give reasons not to use APDs (Razavi [5], pg. 57)
- therefore, not using semiconductor single-photon detectors

Responsivity: $I_p = R_{ph} P_{op}$, where I_p is current generated by photodiode, R_{ph} is the responsivity, and P_{op} is the optical power.

quantum efficiency: $\eta = \frac{I_p/q}{P_{op}/(hc/\lambda)} = \frac{1.24 R_{ph}}{\lambda} = \frac{1}{1+\tau_r/\tau_{nr}}$

Receiver error rate: is Razavi's noise/error rate model correct for us? We need to pay special attention to two limiting cases: 1) a single spike is received after a long period of rest (string or zeros); and 2) a fast spike train is received (burst input) (one zero one zero ...) Compare Razavi's analysis to Bryce's calculation.

Here's Bryce's calculation:

5.1 Motivation

The ultimate goal of this project will be to compare the performance and feasibility of various incoherent optoelectronic platforms for neuromorphic computing. Semiconductor photodetectors have already become a ubiquitous technology and are consequently a natural starting point for a whole class of potential optoelectronic neuromorphic hardware. This short summary seeks to accomplish two goals: (1) To determine the minimum optical power required for the detector to register an optical spiking event and (2) to analyze both the static and dynamic energy dissipation in a detector.

5.2 Photodetector Basics and Noise

The most basic parameter for a photodetector is the *responsivity*, $\mathcal{R}$. The responsivity is simply the ratio of photocurrent to input optical power. It is related to the external efficiency, η as shown below:

$$\mathcal{R} = \frac{q\eta\lambda}{hc} \tag{1}$$

where λ is the wavelength of the incoming light. η, and therefore $\mathcal{R}$, can be expected to be strong functions of wavelength. With responsivity defined, it is straightforward to predict the amount of photocurrent (I_{ph}) produced for a certain input optical power (P_{opt}):

$$I_{ph} = \mathcal{R} P_{opt} \tag{2}$$

In order to produce a measurable response, the photocurrent must be larger than the RMS fluctuations of the noise current. There are two primary sources of this noise in photodiodes - dark current noise and Johnson Noise from the load resistor. The noise from the dark current will be considered first. The RMS fluctuations can be given as:

$$< I^2 >= 2qI\Delta f \tag{3}$$

I is simply the total current flowing through the diode and Δf is the bandwidth of the detector. This is already an interesting detail, as the noise can potentially be reduced by operating at a lower bandwidth (I guess this corresponds to something like integrating the signal?). For this paper, I will take the bandwidth to be 20 MHz, for sake of comparison with SOENS. In principle all sources of current - ideal diode current, SHR, band-to-band recombination, and photocurrent - should be included in I.

We can approximate $I = I_D + I_{ph}$, where I_D is the dark current. There is an additional Johnson Noise associated with the load resistor, R_L. This leads to a total noise, I_N of:

$$I_N = \sqrt{2qI\Delta f + \frac{4k_BT\Delta f}{R_L}} \tag{4}$$

Equating the photocurrent with the RMS noise fluctuations gives an estimate of the bare minimum detectable optical power.

$$P_{min} = \frac{I_N}{\mathcal{R}} \tag{5}$$

Or in terms of the minimum necessary photocurrent, I_{ph}:

$$I_{ph} = \sqrt{2q(I_D + I_{ph})\Delta f + \frac{4k_BT\Delta f}{R_L}} \tag{6}$$

Solving the quadratic equation for I_{ph} yields:

$$I_{ph} = q\Delta f(1 + \sqrt{1 + \frac{2I_D+}{q\Delta f} + \frac{4k_BT}{q^2R_L\Delta f}}) \tag{7}$$

where $I_R = \sqrt{\frac{4k_BT\Delta f}{R_L}}$, the contribution to the noise current from the Johnson noise of the resistor. The minimum optical power, P_{min} follows from the responsivity:

$$P_{min} = \frac{I_{ph}}{\mathcal{R}} = \frac{q\Delta f}{\mathcal{R}}(1 + \sqrt{1 + \frac{2I_D}{q\Delta f} + \frac{4k_BT}{q^2R_L\Delta f}}) \tag{8}$$

The next task is to determine the appropriate temporal length of an optical spike. For a photodetector operating at its maximum rate, it suffices to use $1/\Delta f$ as the duration of spike. Of course, this duration will likely be set by the speed of the LED, but for now I'm going to just assume that the bandwidth of our receiver circuit was designed to coincide with the bandwidth of the LED. In this case, with photon energy E_{ph}, the necessary number of photons per spike (N) is then:

$$N = \frac{P_{min}}{E_{ph}\Delta f} \tag{9}$$

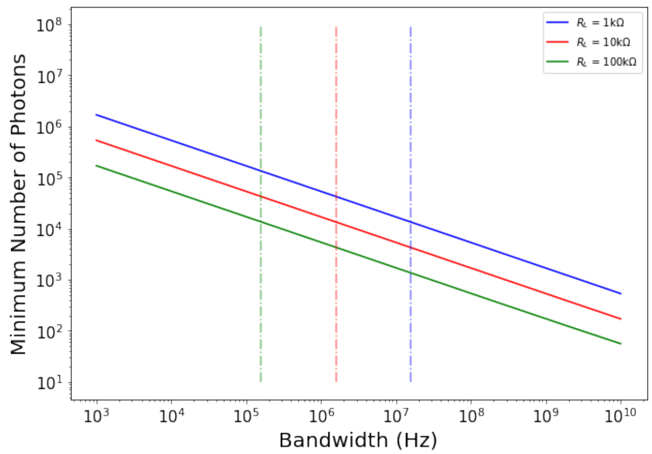

Figure 1: A plot of the minimum number of photons needed per spike as a function of Bandwidth. $I_D = 10$nA, $\mathcal{R} = .5$, and $\lambda = 1.3\mu$m. The dotted vertical lines correspond to the maximum frequency achievable without a transimpedance amplifier for a 10pF photodiode.

A plot of the minimum number of photons per spike vs bandwidth is presented in Figure 1. From experimenting with some plausible parameters, it seems likely that noise will be dominated by the Johnson Noise in the load resistor. This noise can be reduced by utilizing a larger resistor. However, a larger R_L will degrade the bandwidth of the circuit. A lower bandwidth requires the light source to be on for longer and increases the necessary optical power. In Figure 1, the color-coded vertical lines represent the maximum frequency that a 10pF photodiode could respond at. In fact, since the bandwidth is inversely proportional to R_L, it can be seen from equation 8 that in the resistor dominated noise regime, N will become independent of R_L. Fortunately, this bandwidth limit can be beaten by using a transimpedance amplifier as we will see in following sections.

Finally, we can also see a simple example of explicit temperature dependence. I_D is a strong function of temperature and will play a role when we get to static power dissipation, but the Johnson noise in the resistors is more important for determining the minimum optical signal. It has a simple square root temperature dependence, which can be seen in figure 2.

5.3 Transimpedance amplifier

6 Synaptic circuits and weighting

Important section.

- comparison of possible functionalities
  - short-term facilitating
  - short-term depressing
  - long-term event-based potentiation/depression (STDP)

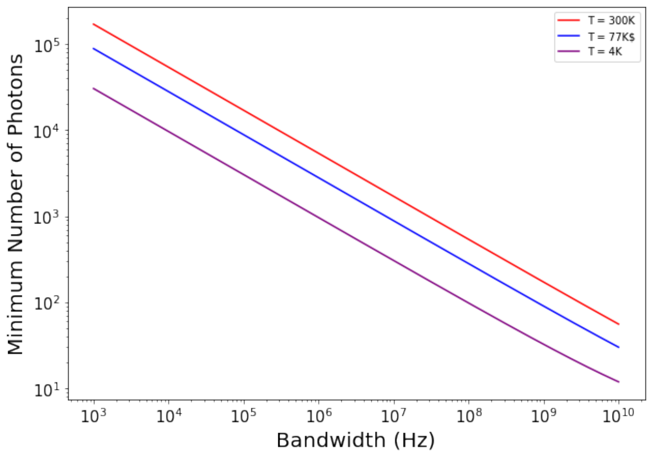

Figure 2: Temperature dependence of the minimum optical signal. $R_L = 100k\Omega$

- for each, consider information-processing metrics such as range of operation, bit depth
- for each, also consider area, power
- also sensitivity (exponential weight dependence on voltage in subthreshold mosfet vs polynomial (almost linear) weight dependence on current in soens)
- discuss in the context of Fusi's work ([6–8])
- it may just be that semiconductor-compatible approaches to memory appear weaker because they have had more time to be scrutinized

7 Dendrites and fan-in

8 Neural integration and threshold

9 Synaptic, dendritic, and neuronal adaptation

Discuss options for short- and long-term synaptic plasticity;, adaptive dendritic functions; and neuronal refractory period, spike-frequency adaptation, and homeostatic plasticity (threshold adaptation).

9.1 Memristors

- Synaptic adaptation occurs due to atomic motion within the material. This means all time constants and signal levels necessary for adaptation are strongly coupled to material parameters. Rather than designing a circuit with essentially arbitrary time constants and signal levels that are exactly the same as the signal levels used elsewhere in the system for information processing, one must explore many materials and find the ones that work closest to the signal levels used for the rest of dendritic and neuronal processing.
- Motion at the atomic level also risks degradation over time. If such a synapse performs short-term adaptation (with conductance changing on the time scale of the inter-spike interval), even materials that can survive billions of cycles will reach the end of life in a matter of days, assuming synapses receive inputs on average at 1 kHz, as expected for optoelectronic neurons
- One may argue synaptic plasticity occurs due to structural changes at the nanoscale (particularly long-term adaptation), so memristor mechanisms are similar to biology in this regard. However, the assembly of molecular machines at the nanoscale via protein synthesis and construction leads to the potential for growth and regeneration in a manner that will be difficult for memristors to match. Memristors will suffer from the second law of thermodynamics, eventually succumbing to a homogeneous, disordered state, while the organizing capabilities of biological synapses locally circumvent this inconvenient propensity.
- The plasticity mechanisms proposed in soens are based around achieving desired functionality through circuit design, with complexity and responses limited primarily by area, while the plasticity mechanisms leveraged in memristors are based on material properties, and are therefore limited primarily by the physics of the universe. There is a large parameter space to explore in both cases, but circuit design provides much more opportunity for flexibility and implementation of diverse responses across the network, while diverse responses from memristors may require a large number of different materials, which will be difficult to fabricate for practical reasons.
- The strength of memristors is exactly the weakness of soens: memristors can be made very small.
- It may be helpful to have metaplastic neuromodulatory control, so a region of synapses can adapt more or less rapidly at various times. With the circuit approach this can be accomplished by varying common current or voltage biases. Is there an analogous mechanism for memristors? Can you change their rate of adaptation collectively, not based on their individual history of activity, but by other mechanisms based on network activity across larger regions?

10 Transmitter driver circuits

11 Time constants and subthreshold oscillations

12 Biasing

13 Power consumption, cooling, and system considerations (including fabrication and production)

14 Notes

Differences between neural and digital optical communication:

- neural: mux/demux not required
- neural: high power optical signals not necessary (nor tolerable)
- neural: not point to point, one to many
- neural: asynchronous, no clock, no phase-locked loop, no clock recovery on receive
- neural: 1s and 0s not equally common; signals are sparse
- neural: TIA + limiting amplifier + decision circuit likely uses too much power
- neural: noise is more tolerable, decision circuit still potentially useful
- neural: speed can be much lower, as demonstrated by biology
- neural: with lower light levels, light-source driver circuits don't need to deliver as much current
- multi-chip partitioning required for digital due to high speed and sensitivity to timing jitter, multi-chip not tolerable for neural (cannot have multiple chips for each neuron) Tx and Rx amplifiers cannot remain in isolation ([5] pg. 5)
- neural: bits are not sampled on a clock

other notes:

- in conventional optical communication systems, package parasitics limit speed. optoelectronic integration crucial for overcoming this limitation ([5] pg. 5)
- for long time constants, semiconductors can augment RC by op amp gain: $RC \to (1+A)RC$, where A is the op amp gain, which can be enormous, like 300,000. thus, essentially arbitrarily long time constants can be achieved. the price is power.
- regarding subthreshold oscillations, RLC behavior in semiconductors can be achieved with op amps. in this case, there is no inductor, and that role is played by the active op amp. the price is power

15 Acknowledgements

This is a contribution of NIST, an agency of the US government, not subject to copyright.

A Appendix One

Appendix One

References

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

[2] E. Chicca, F. Stefanini, C. Bartolozzi, and G. Indiveri. Neuromorphic Electronic Circuits for Building Autonomous Cognitive Systems. *IEEE J. Sel. Top. Quant. Electron.*, 26:7700315, 2020.

[3] J.M. Shainline, S.M. Buckley, A.N. McCaughan, J. Chiles, A. Jafari-Salim, M. Castellanos-Beltran, C.A. Donnelly, M.L. Schneider, R.P. Mirin, and S.W. Nam. Superconducting Optoelectronic Loop Neurons. *J. Appl. Phys.*, 126:044902, 2019.

[4] J.M. Shainline. Fluxonic Processing of Photonic Synapse Events. *IEEE J. Sel. Top. Quant. Electron.*, 26:7700315, 2020.

[5] B. Razavi. *Design of Integrated Circuits for Optical Communications*. Wiley, second edition, 2012.

[6] D.J. Amit and S. Fusi. Learning in neural networks with material synapses. *Neural Computation*, 6:957, 1994.

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

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