Optoelectronic Neural Systems

Device and Architecture Considerations for General Intelligence

Jeffrey M. Shainline

NIST, Boulder, CO, 80305

December 21, 2018

Abstract

Contents

1 Introduction

Optoelectronic neural systems lie at the intersection of multiple fields of science and technology.

- computing
- cognitive sciences
- device physics
- neuroscience
- electronics
- photonics
- semiconductors
- superconductors

Now is the time because the limits of silicon electronics for implementing von Neumann architecture are upon us, and the same hardware used for digital computing to far surpass humans in that modality is not equipped to surpass us in neuro modality.

2 Historical context

- origins of modern computing intertwined with WWII
- Turing: interests, universal computation, computability, Turing machine, serial
- von Neumann: interests, universal computation, numerical investigation of numerous physical problems, digital computing, memory storing data and instructions, von Neumann bottleneck
- Shannon: communication, information in data streams, again focus is on serial information processing
- computing hardware: vacuum tubes, punched cards lead to silicon microelectronics, si uniquely suited to accomplishing digital computing, von Neumann architecture still going strong in si
- communication hardware: ethernet for pretty big networks, fiber-optic cables replacing telegraphs under the atlantic
- silicon photonics is where these two meet: light for commmunication, electronics for computation, maintaining the von Neumann architecture, WDM across the von Neumann bottleneck

[2] [3] [1]

3 Information processing in neural systems

- differentiated local processing with information integration across space and time
- average path length, small world
- hierarchical architecture
- neuronal avalanches
- rentian scaling
- cross-frequency coupling
- fractal use of space and time for scaling
- get into a tad of history here, analogous to Turing/von Neumann

4 Semiconductor electronic neural systems

- shared communication infrastructure
- address-event representation

- connectivity/speed tradeoff
- synaptic, dendritic, and neuronal functionalities: emulating neural behavior with digital systems.
- connect back to von Neumann: stepping through a differential equation in time with a Turing machine rather than leveraging devices that physically manifest the differential equations of interest
- hardware for universal computing with a Turing machine is not efficient for neural information
- memristors: really crappy synapses
- application spaces: deep learning, mobile devices, IoT; neuro-inspired, but not really neural computing
- mention Mead, sub-threshold transistor IandF isomorphism
- The distributed von Neumann approach still effectively steps through differential equations numerically. We advocate for hardware that actually lives through the response modeled by the diff eqs.

5 Superconductor electronic neural systems

- JJ basics
- IBM latching logic
- Likharev
- victorious march of CMOS
- third wave: IARPA/C3
- JJ neurons (Japan, Segal, Schneider, that recent theory paper)
- still, communication problems, fan-out

6 Optoelectronic synapses, dendrites, and neurons

- light easy to use for long haul, but chip scale?
- occurs in the context of silicon photonics evolution, soref and bennett, Luxtera, vladimir
- general concept: communication between neurons is photonic; when a neuron spikes it must either generate or modulate light; throughout, speed, size, power all co-optimized
- first key choice: generate or modulate
- modulate:
  - requires cw light running at all times ($x_{dB/cm} = 1; y_{dB/s} = 100 * x_{dB/cm} * c; q_{dB} = 3; t_s = q_{dB}/y_{dB/s}$, for 1 dB/cm propagation loss, 3 dB of the light is lost every 100 ps)
  - requires frequency tuning, most likely
  - cross talk of neurons on the same bus
- generate:
  - requires light source at every neuron
  - requires unprecedented optoelectronic integration, million sources and a billion detectors on a wafer
  - must be very low capacitance
  - seems like only a silicon light source will suffice, but this would require cryogenic operation
- second key choice: establish synaptic weight in the photonic or electronic domain?
- photonic domain:
  - This choice has several important ramifications for hardware and information processing. Regarding information processing, it is usually assumed that neural communication is digital: the presence or absence of an action potential is a binary one or zero, and the amplitude of the action potential is not encoding information. When adjusting the synaptic weight in the photonic domain, this is not the case. The number of photons reaching a neuron through a synaptic connection becomes an analog variable, and it is subject to shot noise, in addition to any noise mechanisms present in the detector. The signal-to-noise ratio of shot noise improves with $\sqrt{(N_{\mathrm{ph}})}$, 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.
  - these objections notwithstanding, to our knowledge, all except one optoelectronic neural approach proposed to date sets weight in photonic domain
  - specific instances: mzi (no STDP, poor spatial scaling, cross-talk); wdm (limited number of

channels, cross-talk with rings on master ring, demands on sources); mzi and wdm (thermal tuning hopeless for scaling, no plasticity mechanisms proposed); phase change synapses (at least don't dissipate steady state, still power lost due to variable attentuation, small footprint, Hebbian learning possible, but STDP not likely, meta, short term also doesn't look promising)

- electronic domain:
  - By contrast, if we establish synaptic weights in the electronic domain, light is used exclusively for communication, and communication remains entirely digital. The presence of an optical signal can be used to represent an all-or-none communication event. In this case, the detector and associated electronics must be able to achieve a variable synaptic response to identical photonic pulses based on the configuration of the electronic aspects of the circuits. In this case, we expect that a neuron will send, on average, N_{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.
  - only know of one system where electronic domain has been proposed: soens
    - basic functionality
    - stdp
    - meta
    - homeo
    - short-term
- neuronal computation: reaching threshold
  - differentiate between state-based and spiking
  - main considerations here are energy/power
  - how much energy is required to generate a pulse or drive a modulator?
  - how much light must be made/moved to drive all downstream synaptic connections?
  - how fast can pulses be generated (refractory period)?
  - how long can neurons remember (leak rate)?
  - what is range of spike rates? what is expected power?
- somewhere in here, comparison of detectors (going cold costs 500x for carnot, but gains 2000x for detector sensitivity)
- related, comparison of sources (going cold reduces how many photons must be made, but most importantly, if it means a silicon light source can work for this project, it is a game changer)
- inhibition, gotta have a plan
- dendritic processing
  - intermediate nonlinearities
  - direction attention with inhibition
  - sequence detection
  - how can any of this happen in the photonic domain?

7 Achieving large-scale optoelectronic systems

- unprecedented integration of photonics and electronics in a scalable process that can be implemented with existing infrastructure–change a few implant conditions, swap out a few sputtering targets, improve BEOL dielectrics for photonics
- communication on various length scales, multi-planar on wafers, wafer-to-wafer vertical and lateral, fiber white matter
- feasibility of brain scale
- why si if no transistors?

  - III-V substrates should be pursued as well. Our group is working on this, initial anecdata indicates similar efficiency
  - big problem is fab. wafers are harder to scale, material harder to purify, oxide not as good for waveguide cladding. Similar consideration to mosfet gate. Overall manufacturability
  - may eventually use transistors for perhaps faster refractory period
- ultimate limits

8 Outlook

- circle back to Turing and von Neumann, their interests in machine intelligence and modeling computation after the brain
- circle back to digital vs neural, superconducting optoelectronics brings communication and spiking nonlinearities
- why go to all the trouble?
  - this technology will only be pursued if it can do something that nothing else can do
  - but it can, and what it can do is very important
    * exceptional complexity for experiments in network information, neuroscience models
    * quantum/neural hybrid systems
    * scaling beyond what is possible with other methods, perhaps the smartest machines on the planet
    * computing has shaped economy and society since its inception
    * powerful scientific tool
    * foundational questions about thought and consciousness amongst the most intriguing and important in modern science

References

[1] C. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379, 1948.

[2] A. Turing. On computable numbers, with an application to the entscheidungsproblem. *J. of Math*, 58:230, 1936.

[3] J. von Neumann. A first draft of a report on the edvac. 1945.