Challenges and Opportunities in Neuromorphic Computing

El. Towe and I. Kimukin

towe@cmu.edu

Carnegie Mellon University

Pittsburgh, PA 15213 USA

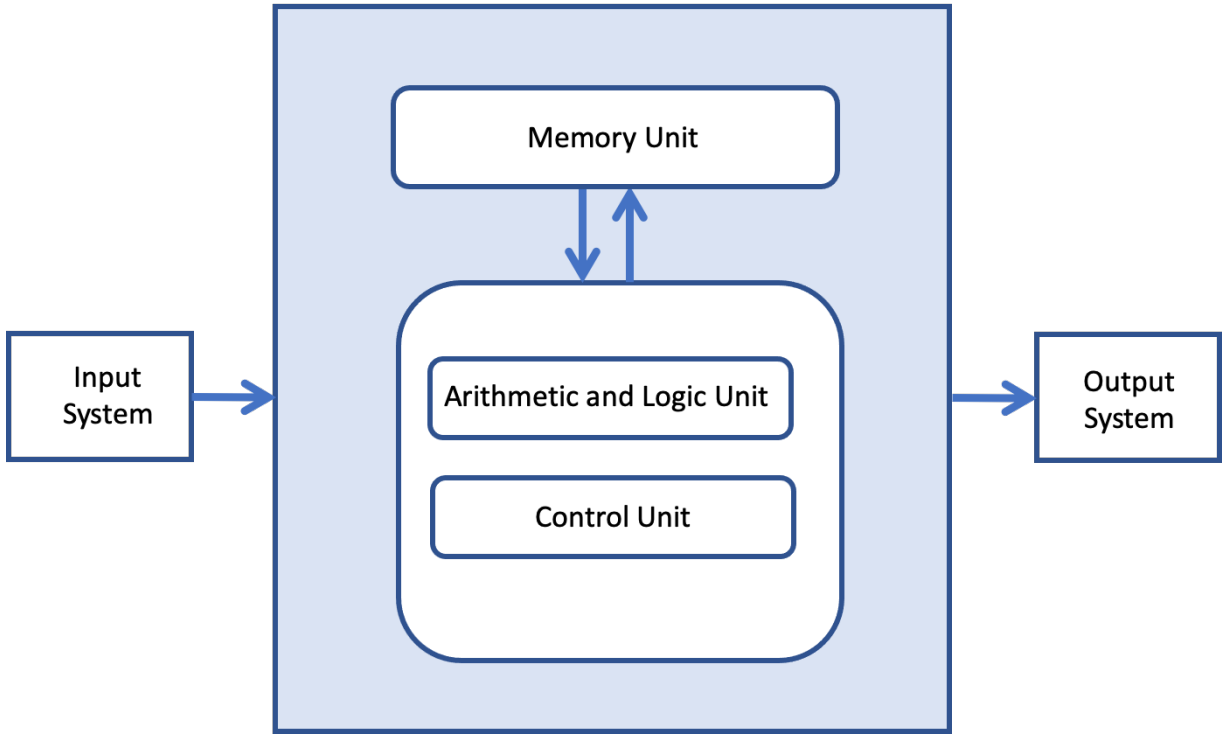
17th US-Korea Nanotechnology Forum, Seoul Korea

2023.04.03

Agenda

- Prevailing computing paradigm and its problems
- What is neuromorphic computing
  - Key differences between neuromorphic and classical computing
- Current implementations
  - Electronic neuromorphic computing
  - Photonic neuromorphic computing
- Opportunities and challenges

Prevailing von Neumann computing paradigm



- The von Neumann architecture of computing requires the shuttling of instructions (programs) and data back and forth between memory and processor;
- Most of the the time this is what a computing system is doing, thus creating the **von Neumann bottleneck.**
- Data shutting consumes a lot of energy, making cloud data centers some of the most energy intensive operations.

Neuromorphic Computing

- Neuromorphic computing takes inspiration from the brain, which we understand, is massively interconnected and processes information in parallel, and has:
  - About $\sim 10^{10}$ neurons, and
  - Each neuron is connected to about ~ 10^4 synapse;

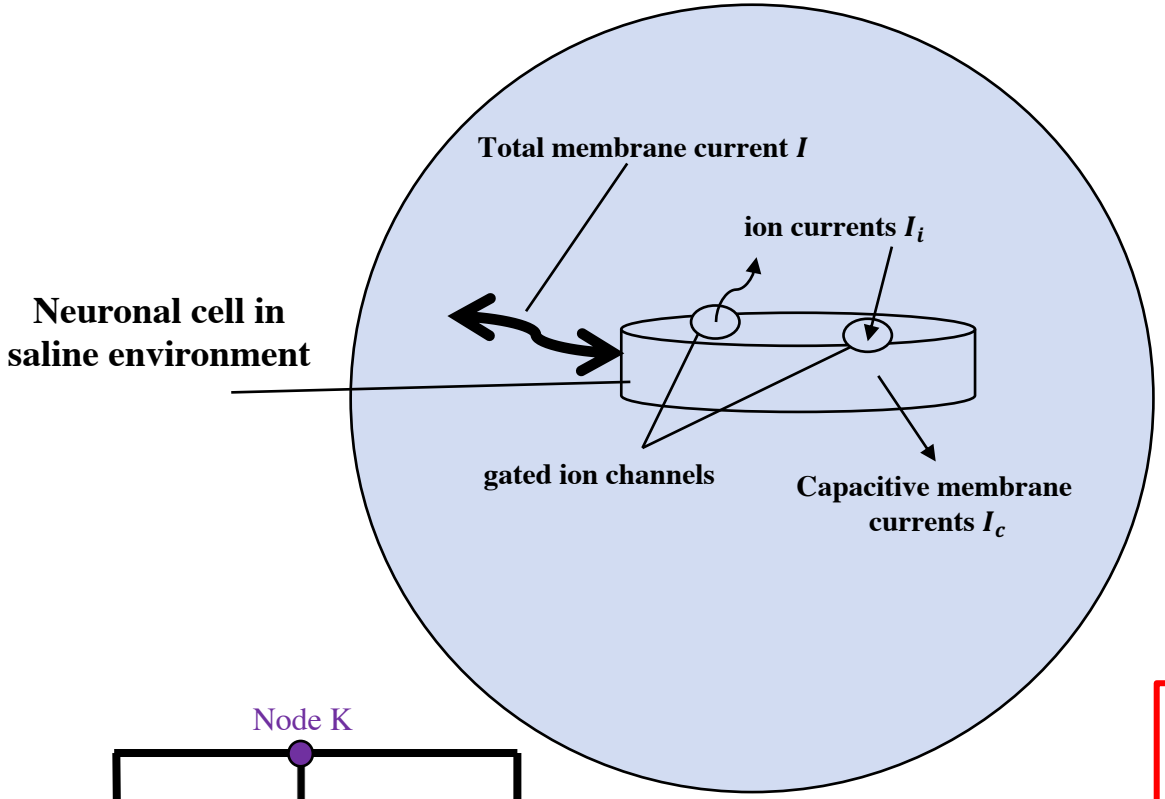
- Information processing is thought to take place in the synapses, which also store the information;

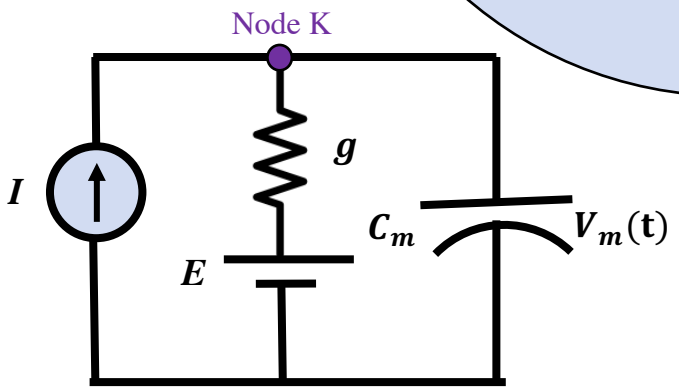
- Computing in the brain is therefore performed within the memory itself, unlike how it is done in classical computing;

- While performing all its remarkable feats, the brain consumes only about 20W.
  - No artificial processor is capable of this energy efficiency;

- The goal of neuromorphic computing is to develop systems inspired by the brain in the way they process information and use energy efficiently.

Dynamical electrical model of a neuronal cell





- **Cell membrane is an electrically insulating dielectric;**
  - Can be represented by capacitor , C_m;
- **Na^+, K^+, Cl^- pumps in membrane keep its inside below outside potential;**
  - Difference in potential is the resting potential, $E = E_{Na} + E_K + E_{Cl}$;
- **Membrane is not fully insulating – it leaks some (Na^+, K^+, Cl^-) ion currents;**
  - Can represent sum of leakages with conductance, $g = g_{Na} + g_K + g_{Cl}$;
- **Cell membrane receives charge q from external inputs or other external cells at some rate:** $\frac{dq}{dt} = I_{ext} + \sum I_{other\ cells} = I(t)$
- **Cell membrane electrical characteristics and parameters can be modeled by the circuit shown at bottom left corner.**

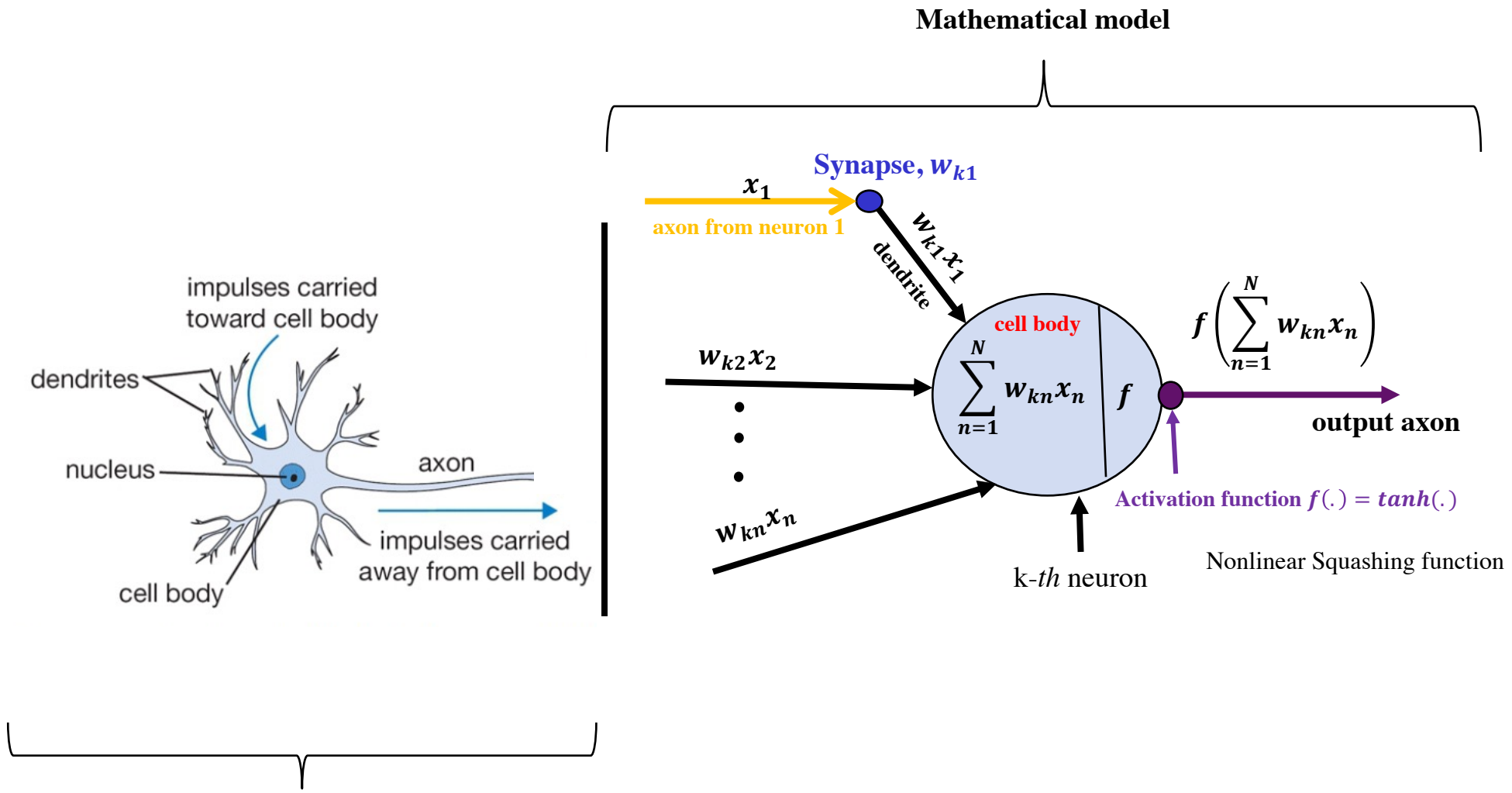
Kirchhoff's current conservation law at node K leads to:

$$C\frac{dV_m(t)}{dt} + g(V_m(t) - E) - I(t) = 0 \qquad (1)$$

This is the dynamical neuronal cell membrane *state* equation.

Electrical & Computer ENGINEERING

The biological neuron and the artificial model of it



Key differences between classical and brain computing

Human Brain	Digital Computer
Neurons and synapses are the basic building blocks.	Transistor logic blocks and memory are basic blocks.
Asynchronous communication (low energy without a central clock).	Synchronous communication (high energy with a single central clock)
Has high plasticity for dynamic reconfigurability.	Most digital circuits have limited reconfigurability.
Neurons have ready access to synapses (memory); no need to waste energy shuttling information back and forth between memory and processor since processing is in-situ.	Processor and memory are spatially separated, leading to lots of energy being consumed shutting data back and forth.
Neurons in brains are massively interconnected; a single neuron is connected to $\sim 10^4$ other neurons.	We couldn't do that using CMOS technology.
The human neocortex is estimated to have about $1.5 - 3.2 \times 10^{10}$ neurons and about 1.5×10^{14} synapses [Pakkenberg et al.].	1.2×10^{12} transistors in a Cerebras chip; 2.1×10^{10} transistors in Nvidia GV100 Volta card [Moore]

B. Pakkenberg, D. Pelviga, L. Marner, M. J. Bundgaard, H. J. G. Gundersen, J. R. Nyengaard, L. Regeur, "Aging and the human neocortex," Experimental Gerontology, 38(1-2), 95-99 (2003).

S. Moore, "Things to know about the biggest chips ever built," IEEE Spectrum, September 2, 2019.

Electronic implementation of synaptic devices

- Electronic synapses originate from Chua's research on a "missing" fundamental circuit element;
  - The well-known elements are the capacitor, inductor, and resistor; they are related by the following relations

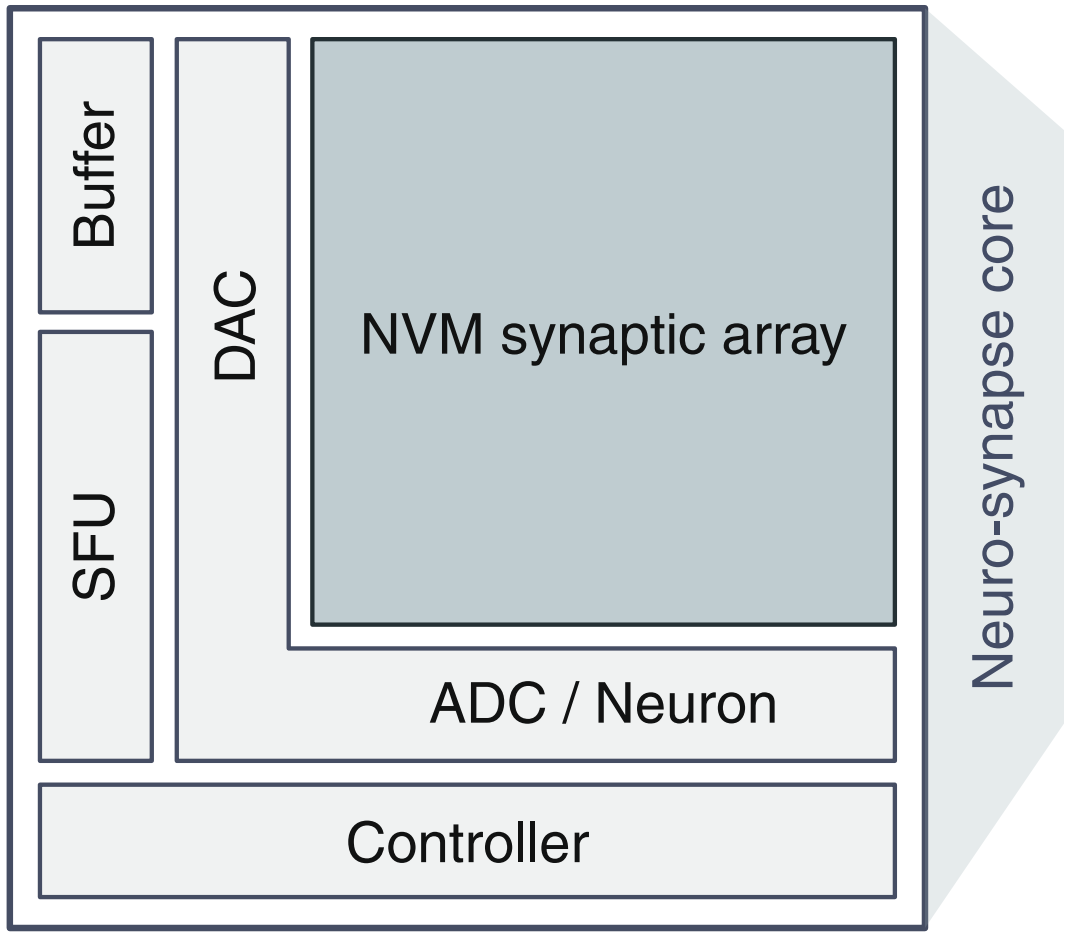
$$\begin{cases} dq = Cdv \\ d\phi = Ldi \end{cases} \quad \text{and} \quad \begin{cases} dv = Rdi \\ d\phi = R_M dq \end{cases}$$

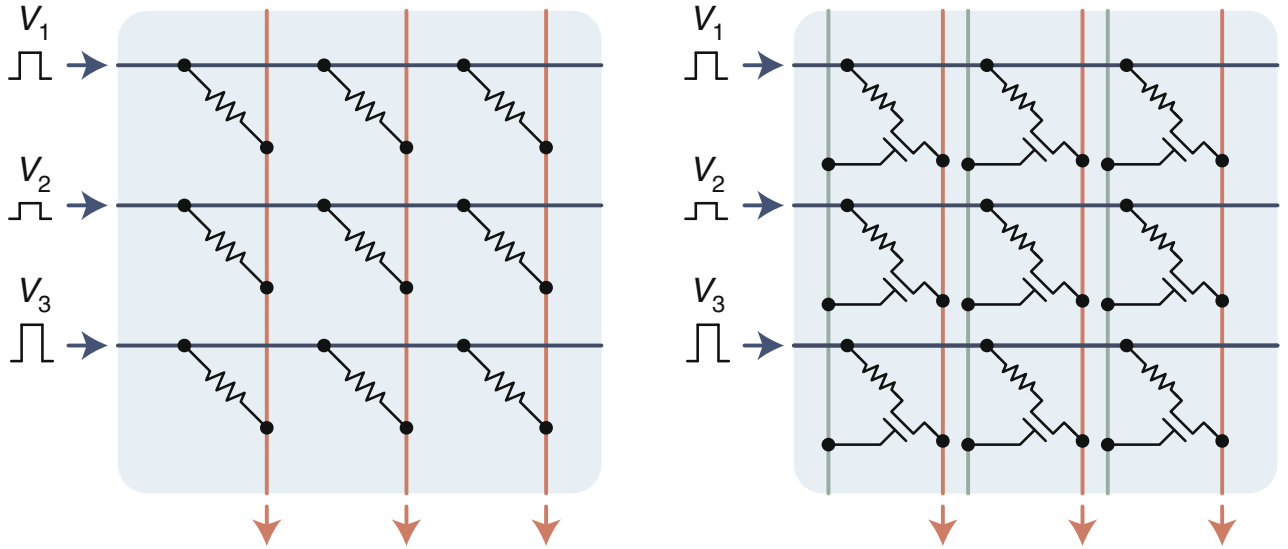
- Chua insisted that because of symmetry, there should be a fourth element, R_M, which he called the memory resistor, which today is called the memristor or the resistance switch;

Memristor: the electronic synapse

- HP reported the first experimental *two-terminal* memristor device in 2008 fabricated from titanium dioxide.
- Memristors can be made from oxide materials such Al_2O_3/TiO_2, $HfAl_yO_x/TaO_x$ and Ta/HfO_2, which are all compatible with silicon microfabrication technology.
- Common mechanism of operation of the memristor is the metal-insulator phase transition (there are other mechanisms: Mott-insulator, Ag-doped insulator for diffusive transport in memristor,
- In electronic neuromorphic computing, the memristor is typically used in a crossbar configuration designed to perform vector-matrix operation - one of the basic operations in deep learning algorithms.
-

Crossbar configuration of electronic memristors

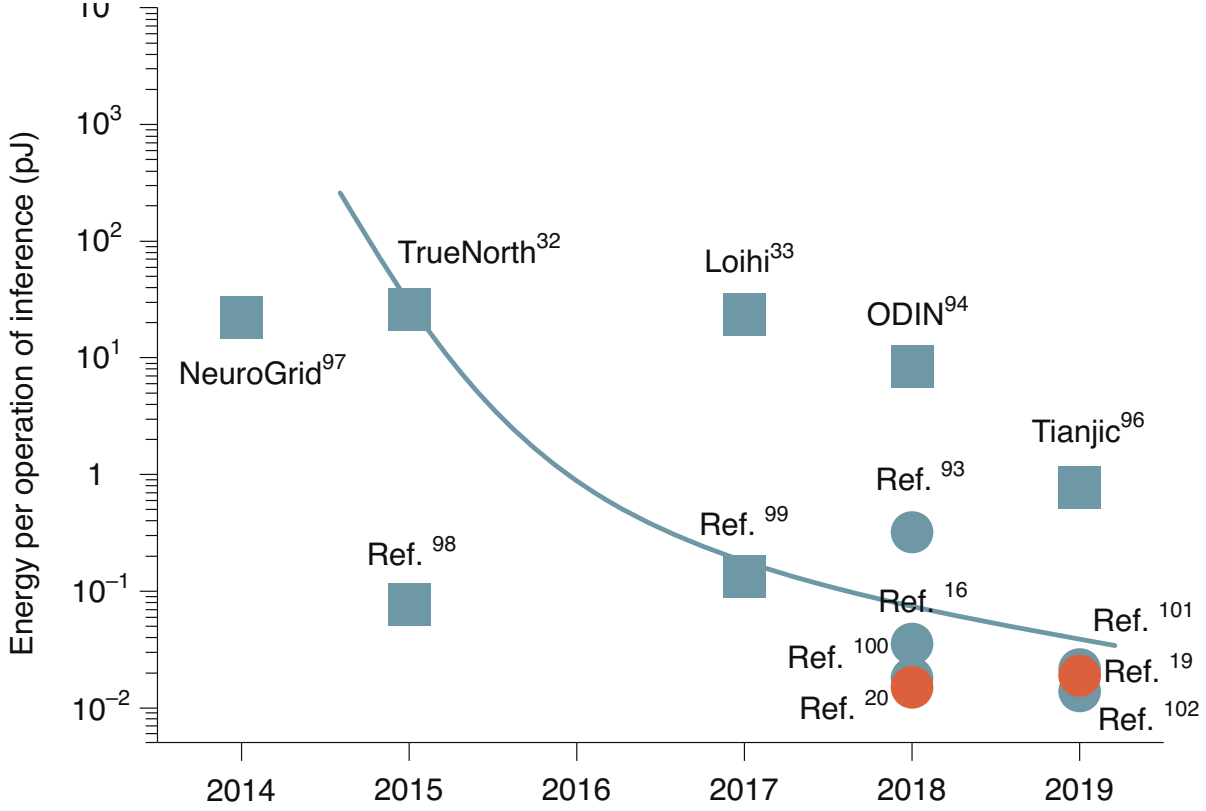




- Crossbar configuration requires input and output circuits for managing digital to analog input and analog to digital out;
- Some local buffer memory is also required;

Q. Xia and J. J. Yang, "Memristive crossbar arrays for brain-inspired computing," Nature Mat. **18** 309-312 (2019).

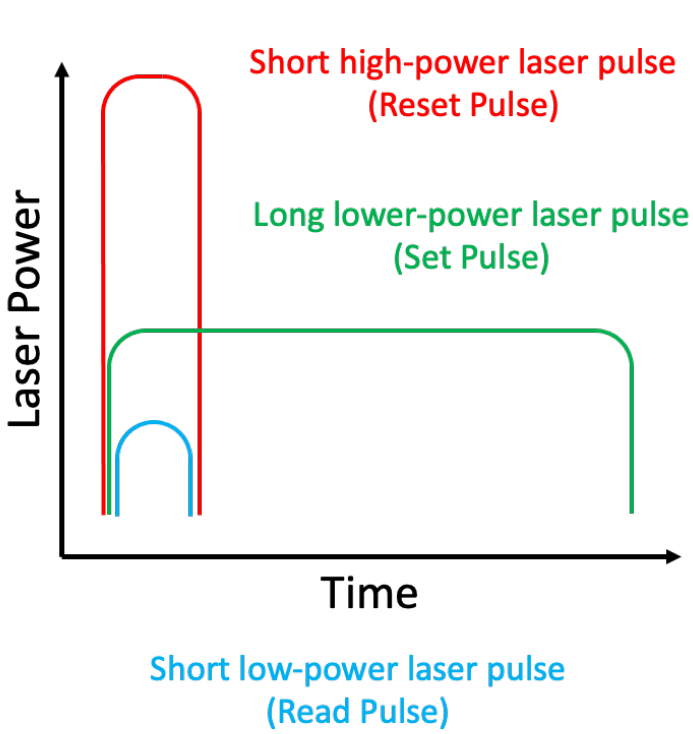
Milestones in electronic neuromorphic computing

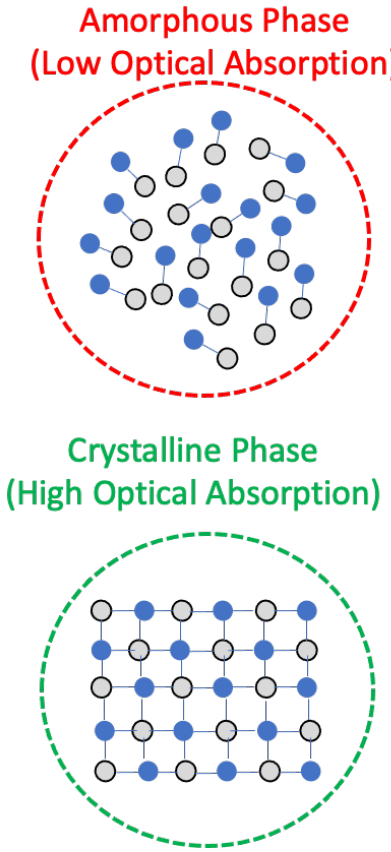


- Number of digital and analog electronic neurochips have been reported;
- The TrueNorth (IBM) and Loihi (Intel) chips are commercially available;

W. Zhang et al, "Neuro-inspired chips computing chips," Nature Eletronics, 3 371-382 (2020)

Phase change material: building block of the optical synapse



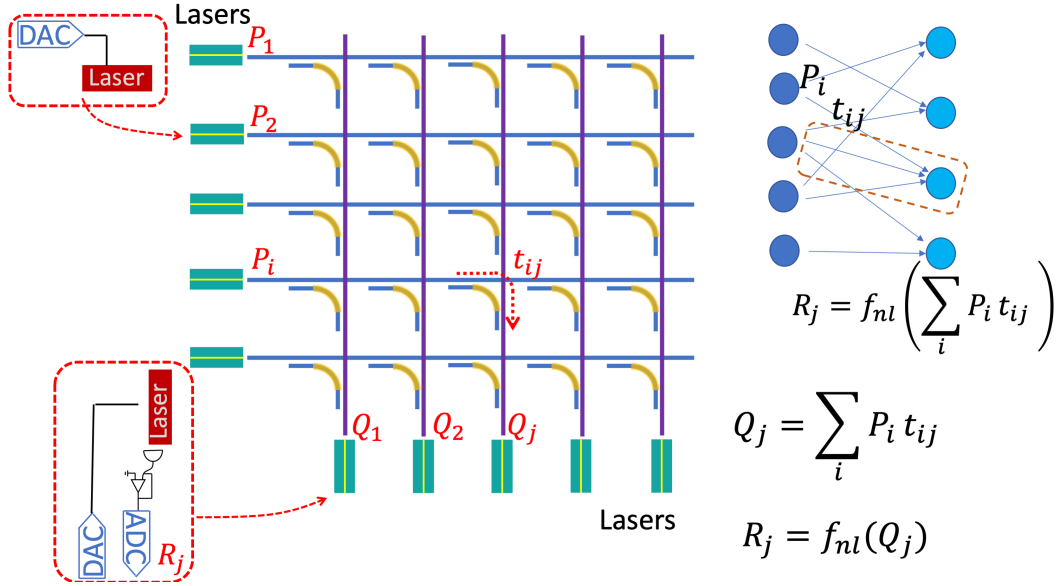


Chalcogenides materials:

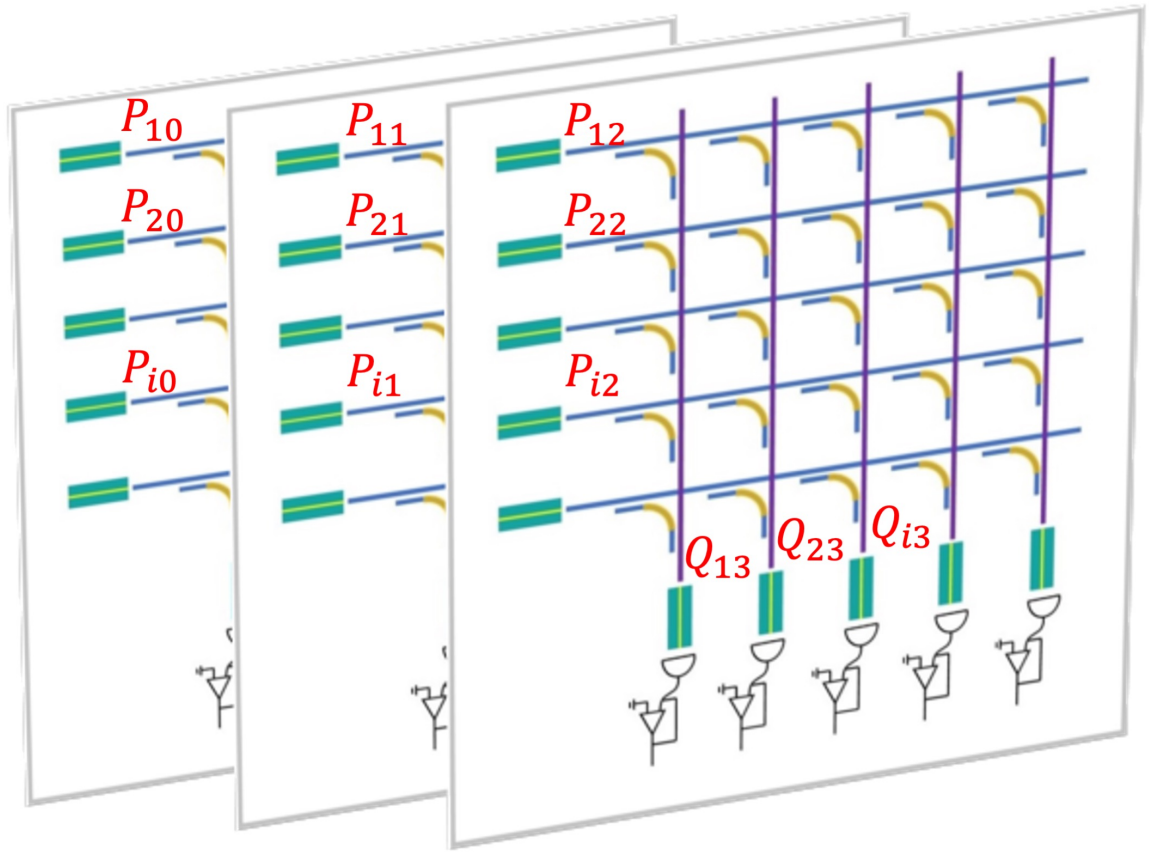
1. GeSbTe (GST)
2. GeSbSeTe (GSST)

Basic Architecture of an Optical Crossbar Multiplier

- Optical phase change arcs are fabricated at intersections of each crossbar waveguide. Each arc represents a synapse
- Horizontal waveguides fed with modulated light inputs from sources driven by digital-to-analog signal converters.
- Each vertical waveguide is an output, which is a sum of the horizontal inputs.
- Crossbar array performs the basic matrix-vector multiplication function of an artificial neuron.



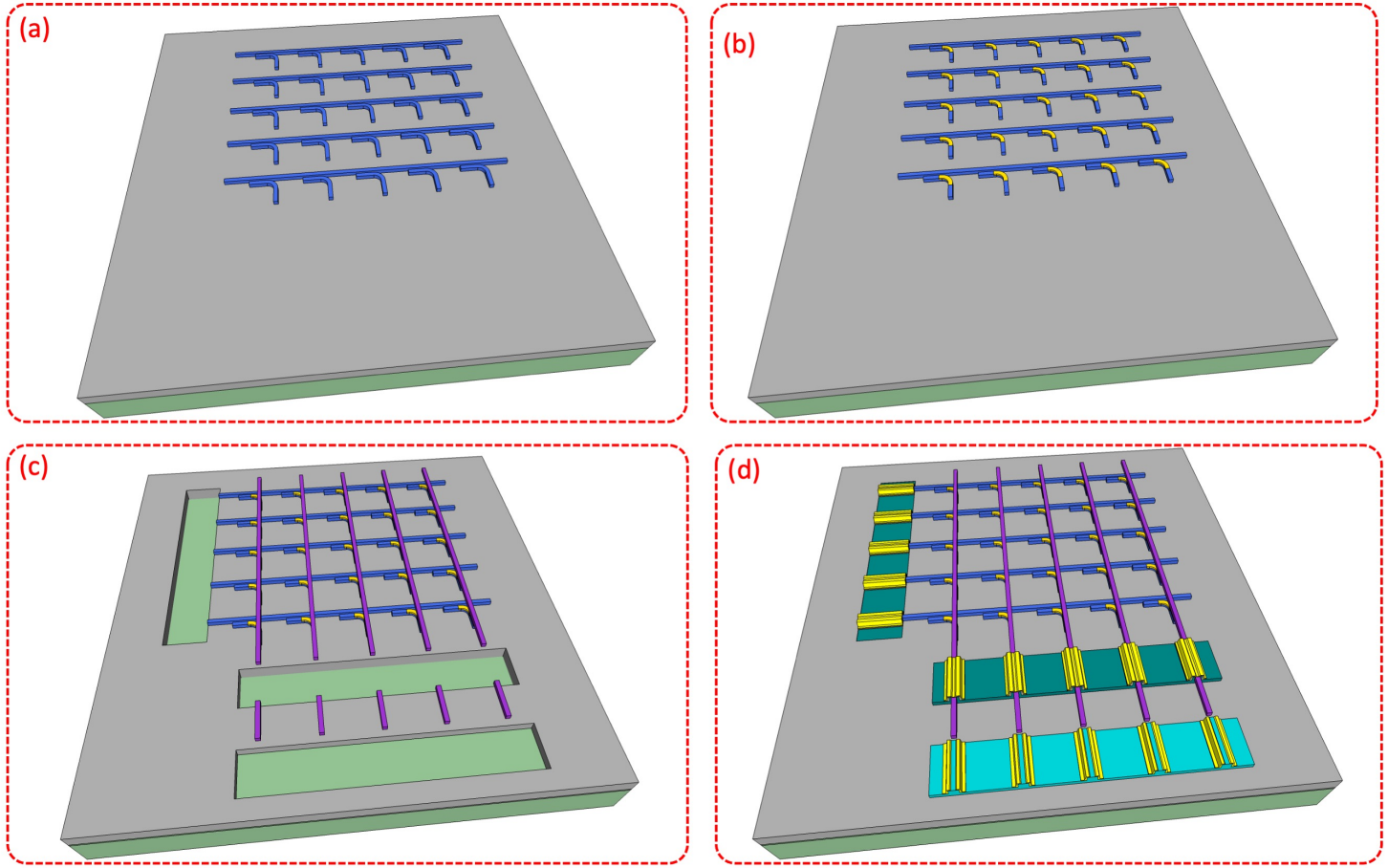
Potential Scaling of Optical Crossbar Multiplier



- The optical crossbar multiplier can be scaled by 3D layering.
- Waveguide made with SiN, whose index of refraction is $n = 1.99$. interlayer cladding is SiO_2 with index $n = 1.44$.
- The 3D integration will require heterogenous integration processes for inclusion of DAC/ADC and optoelectronic sources and detectors.

Key Fabrication Process Steps

- Basic building block is a silicon substrate, which serves as an optical bench for assembling the crossbar multiplier.
- Other materials compatible with silicon include silicon nitride, and silicon dioxide.
- Heterogeneous components include III-V lasers, detectors, and Si-based DACs/ADCs.



Opportunities

- New collaborations that bring together expertise in (i) materials science, (ii) electrical and computer engineering, (iii) computer science, and (iv) neuroscience;

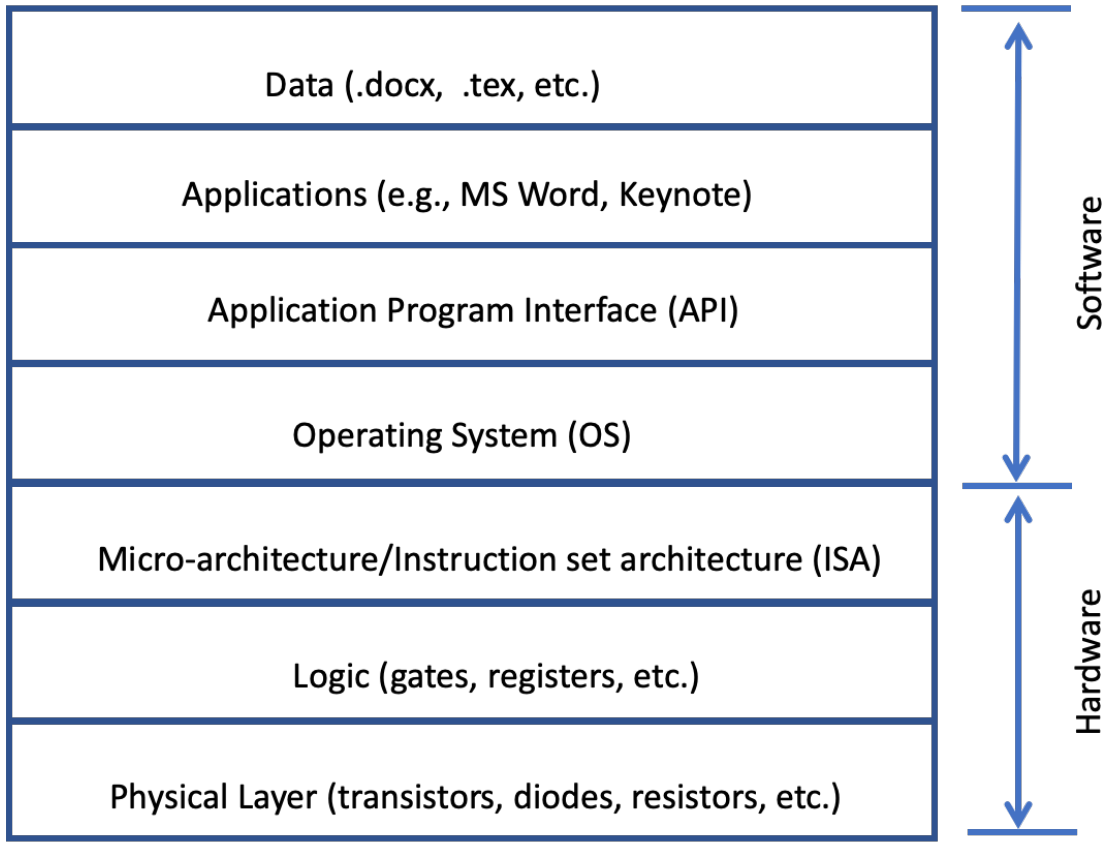
- Neurochips offer potential to develop fast, low power application-specific accelerators for machine learning algorithms;

- Neuromorphic computing presents a new computing paradigm that leverages classical computing architectures and existing silicon microfabrication infrastructure;

Challenge of the incumbent computing framework

- To overcome the drag of the incumbent computing framework (for a toolchain stack), neuromorphic computing requires its own clear definition of a layered abstraction that defines how the hardware and the software shoukd to interact.

- Framework will facilitate effective design and collaborations among the diverse teams needed to develop the field.



Summary

- Reviewed neuromorphic computing
  - Discussed how it is currently implemented
- Discussed some of the opportunities and challenges