Neuromorphic Computing with Novel Devices

Lecturer: Kezhou Yang

Microelectronics Trust

2024.11

Outline

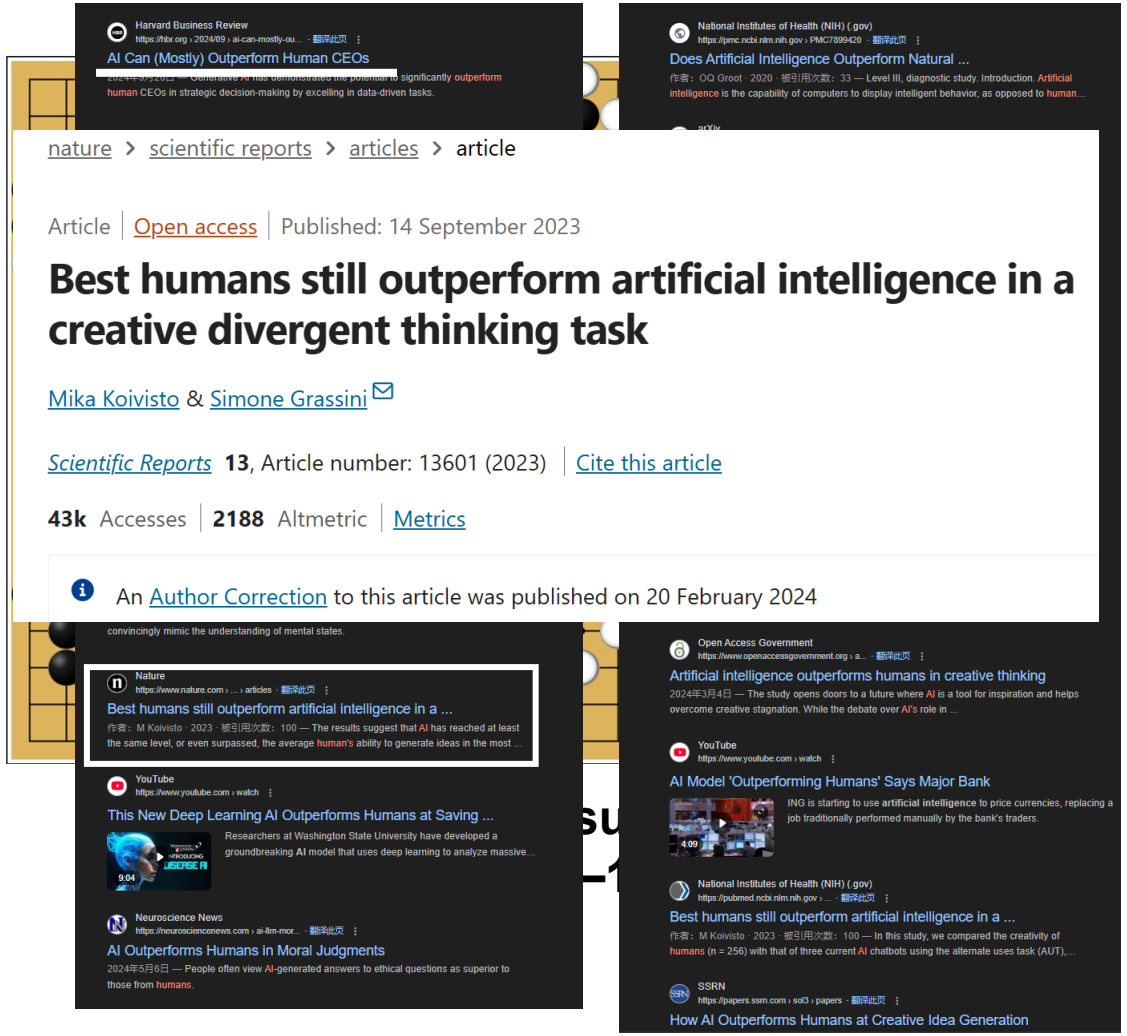
- Motivation: The problems we met
- What neuromorphic computing is
  - Concept and history
- Why we need neuromorphic computing
  - The properties of neuromorphic computing
- How we do neuromorphic computing
  - Algorithm: Spiking neural networks (SNNs)
  - Hardware: Novel devices

The problems we met with current computing paradigms

MOTIVATIONS OF THE FIELD

Deep Learning Outperform Human Beings…

- AlphaGo versus Lee Sedol (李世乭)
  - Go: Complex
  - AlphaGo: Neural network + Monte Carlo tree search
    - Calculate several future moves
- AI works better than human in many tasks

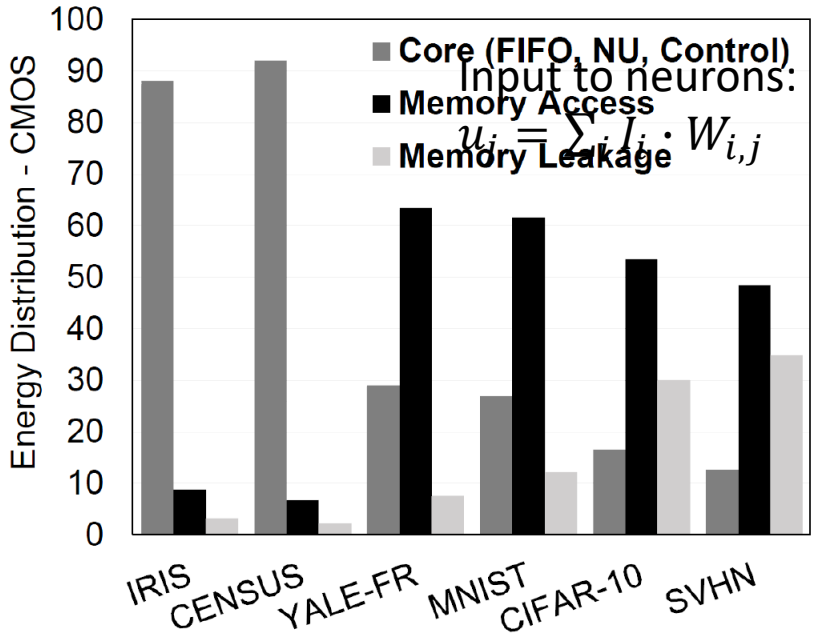


https://en.wikipedia.org/wiki/AlphaGo_versus_Lee_Sedol

At What Cost?

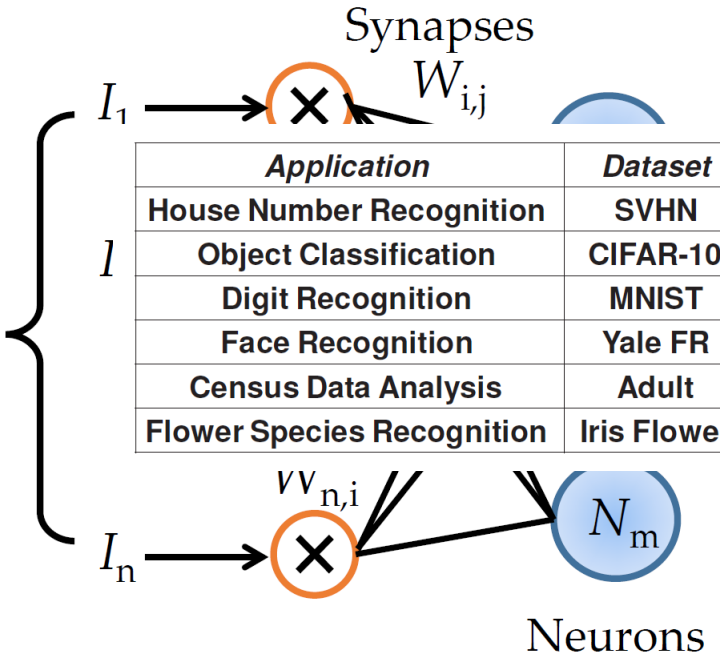
- AlphaGo
  - 1920 CPUs and 280 GPUs
  - Power consumption: $\sim 10^6 \mathrm{W}$ (5 seconds for each move)
- Human
  - Human brain power consumption: $\sim 20\mathrm{W}$
- Not feasible for power-constraint applications

Energy Distribution during Computing



Input to neurons:

$u_j = \sum_i I_i \cdot W_{i,j}$



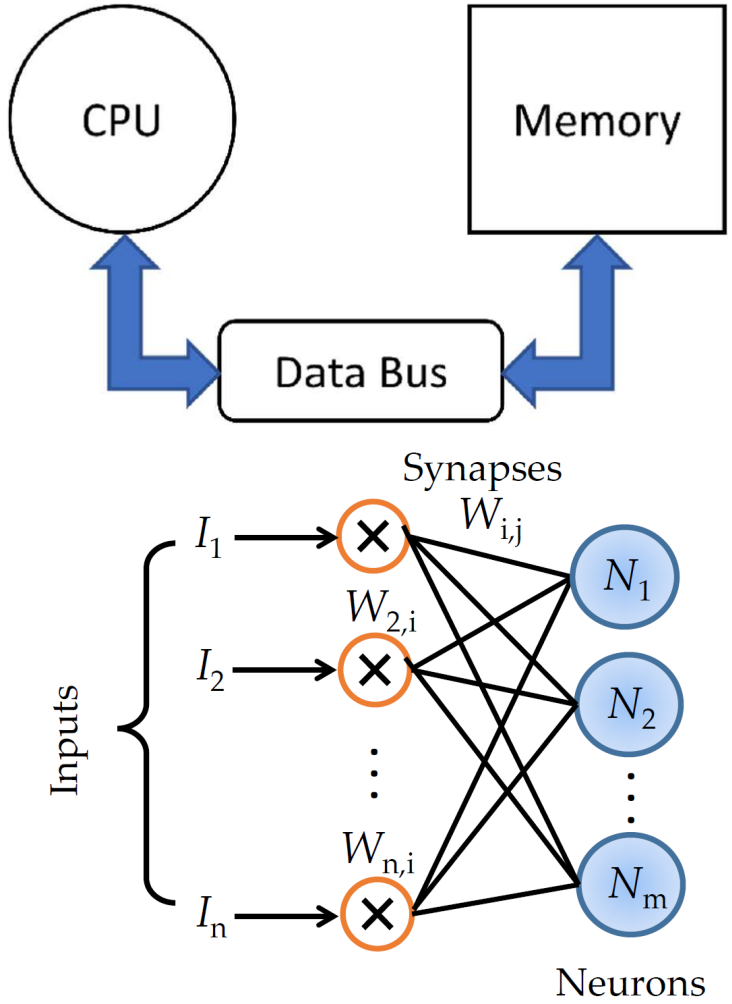
Application	*Dataset*	*Layers*	*Neurons*	*Synapses*
House Number Recognition	SVHN	6	9226	16787456
Object Classification	CIFAR-10	5	6666	12063744
Digit Recognition	MNIST	3	1546	1187328
Face Recognition	Yale FR	3	1039	794112
Census Data Analysis	Adult	2	1026	8192
Flower Species Recognition	Iris Flower	3	195	8384

- Significant energy is consumed in memory access and leakage

A. Sengupta, A. Ankit and K. Roy, "Performance analysis and benchmarking of all-spin spiking neural networks (Special session paper)," *2017 International Joint Conference on Neural Networks (IJCNN)*, Anchorage, AK, USA, 2017, pp. 4557-4563

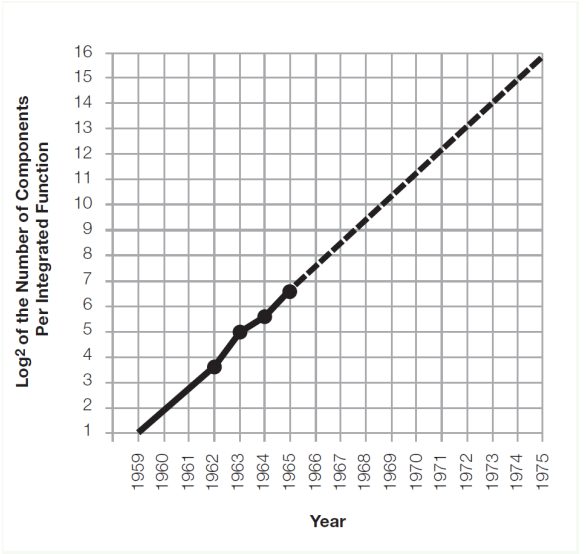
Von Neumann Bottleneck: Hardware-Software Mismatch

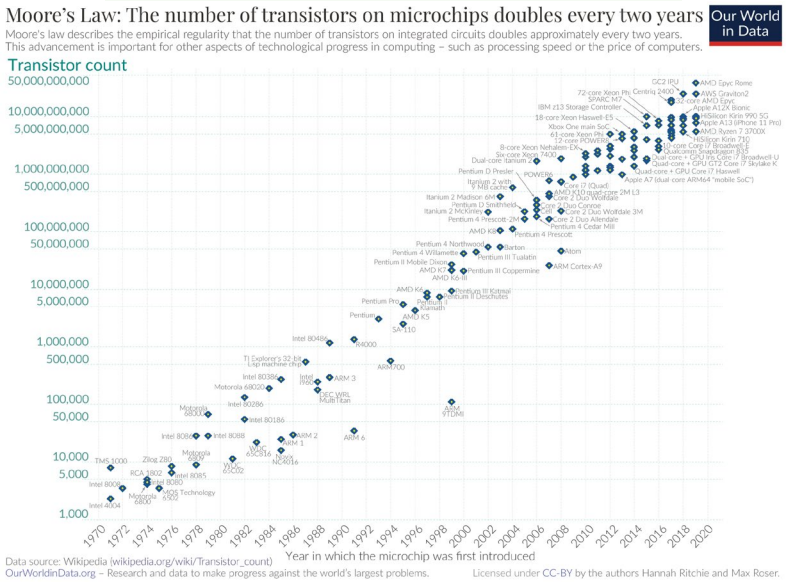
- Von Neumann computer
  - Separate CPU and memory
- Energy consumption
  - Data transportation: ~nJ
  - CPU computing: ~pJ
- Latency
  - Limited data bus throughput
  - CPU processing speed faster than memory
- Von Neumann bottleneck: Mismatch in architecture between algorithm and hardware



Moore’s Law is Dying

- Moore’s Law: Doubling of number of transistors on chip every 18 months.
- The problem
  - Physical limitation
    - Process size: Short channel effect
  - Cost
- We need to compute in a smart way: Neuromorphic computing





G. E. Moore, “Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff.,” IEEE Solid-State Circuits Society Newsletter, vol. 11, no. 3, pp. 33–35, Feb. 2009, doi: 10.1109/n-ssc.2006.4785860.
https://en.wikipedia.org/wiki/Transistor_count

What is neuromorphic computing

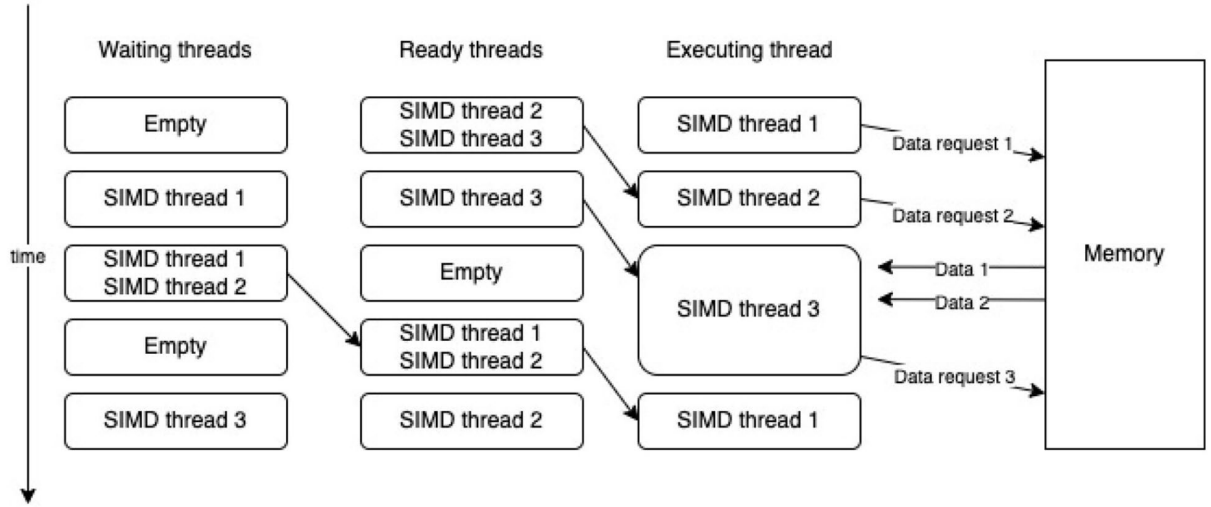
NEUROMORPHIC COMPUTING: A NEW COMPUTING PARADIGM

Neuromorphic Computing: Algorithm model

- Algorithm: Spiking neural networks (SNNs)
- Three generation of neural networks
  - 1st Gen: Perceptron network (Binary output)
    - Hopfield network
  - 2nd Gen: Analog neural network (Continuous activation)
    - Sigmoid, ReLU, …
    - Backpropagation
  - 3rd Gen: Spiking Neural work (Spikes)
    - Event-driven activation
    - Bio-plausible

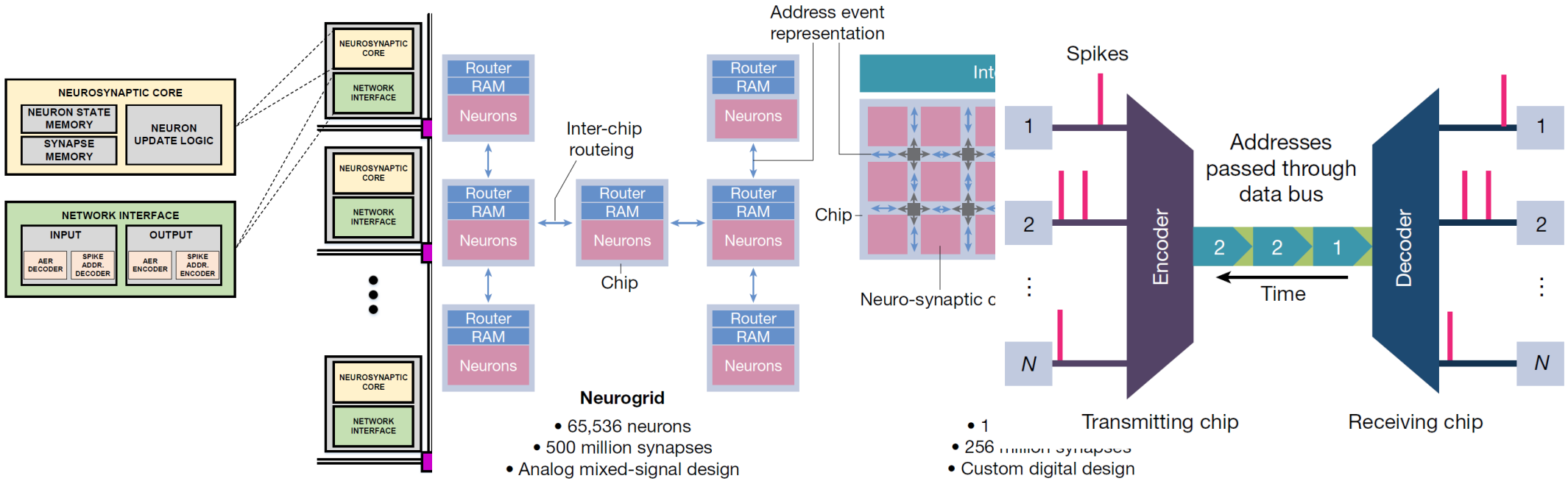
Neuromorphic Computing: Hardware

- Graphical Processing Unit (GPU)
  - Many parallel simple computing cores: Parallel computing
  - Memory bandwidth
- Tensor Processing Unit (TPU)
  - High volume low precision calculation
  - More input/output operations per unit Joule



Ivanov D, Chezhegov A, Kiselev M, Grunin A and Larionov D (2022) Neuromorphic artificial intelligence systems. Front. Neurosci. 16:959626. doi: 10.3389/fnins.2022.959626

Neuromorphic Chip



- Near-memory computing
- Asynchronous address event representation (AER)
- Network-on-chip (NOC)

Nguyen, Duy-Anh, Xuan-Tu Tran, and Francesca Iacopi. 2021. "A Review of Algorithms and Hardware Implementations for Spiking Neural Networks" Journal of Low Power Electronics and Applications 11, no. 2: 23. https://doi.org/10.3390/jlpea11020023

Roy, K., Jaiswal, A. & Panda, P. Towards spike-based machine intelligence with neuromorphic computing. *Nature* **575**, 607–617 (2019). https://doi.org/10.1038/s41586-019-1677-2

Neuromorphic Chip Project

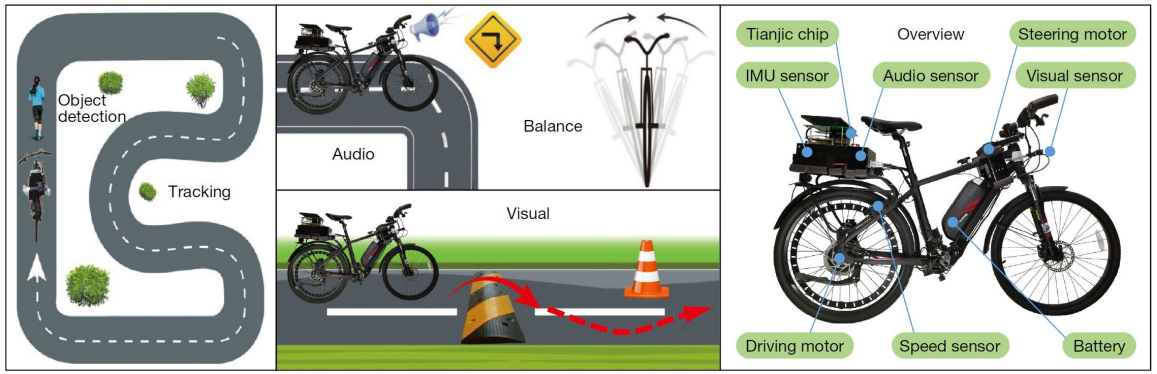
- TrueNorth (2014, IBM)
  - Digital
  - 4096 Neurosynaptic cores
    - 256 firing neurons, 256×265 synapses in each core
  - Inference for CNNs and RNNs
  - Demonstration: Gesture recognition
    - 10 gestures, 96.5% accuracy, 0.18W power consumption
- Neurogrid
  - Mixed digital-analog
  - 65536 neurons in 180 nm CMOS technology
  - Real-time biological brain simulation

Neuromorphic Chip Project

- Loihi (2018, Intel)
  - On-chip learning
  - 128 neurosynaptic cores, 3 Pentium processors
    - 1024 spiking neurons each core
  - Synaptic weight
    - 1 to 9 bits
    - Modifiable: Various learning rules (Supervised, non-supervised, reinforcing)
  - Applications
    - Recognition and segmentation of images; Processing data sequences; Proportional integral differential controller; Finding the shortest paths in a graph, …
- Loihi 2 (2021, Intel)
  - 3D multi-chip scaling
  - Analog spikes: 32-bit precision

Neuromorphic Chip Project

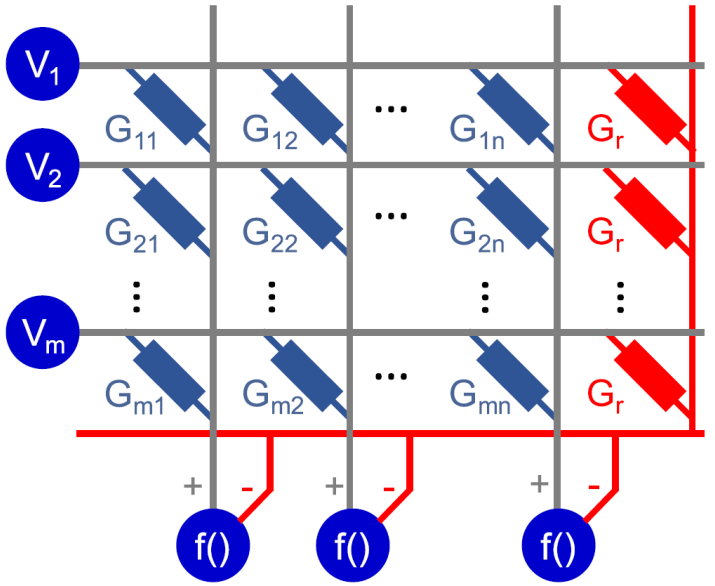
- Tianjic 天机(2019, Tsinghua 清华)
  - Hybrid chip for both ANNs and SNNs
    - 3% additional area consumption (Circuit reuse)
  - 156 cores for 40,000 neurons and 10,000,000 synapses
  - Faster and energy efficient than GPU for SNN and ANN application
  - Demonstration application: Bicycle motion control
    - Many different networks in different type with one chip



Pei, J., Deng, L., Song, S. et al. Towards artificial general intelligence with hybrid Tianjic chip architecture. Nature 572, 106–111 (2019). https://doi.org/10.1038/s41586-019-1424-8

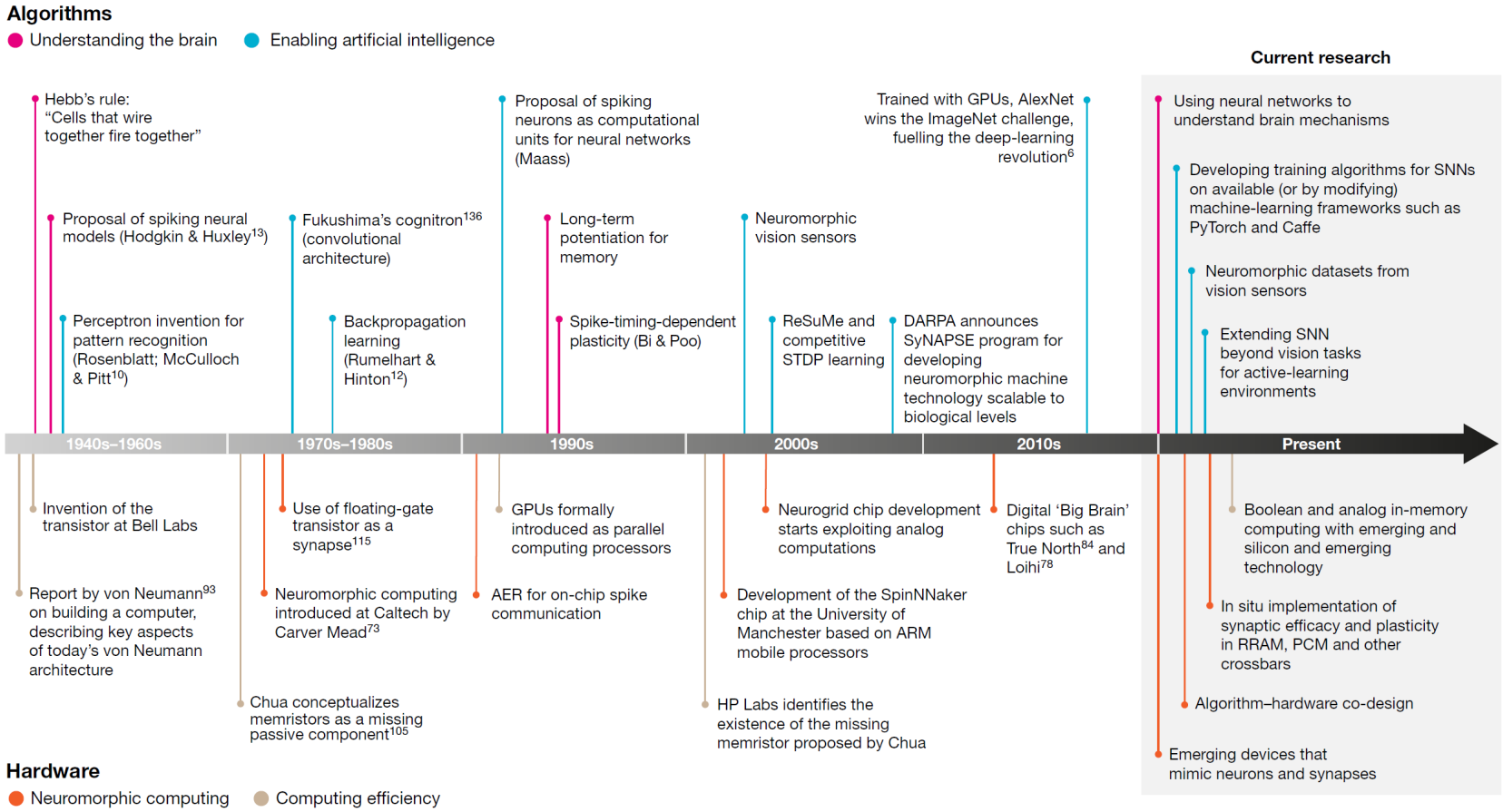
Non-Volatile Devices: In-Memory Computing

- Non-volatile devices for synapses: Similarity in mechanisms
  - Synaptic features
    - Plasticity: Weight modulation based on learning rule
    - Efficacy: Output generation based on input spikes
  - Synaptic weight: Device resistance
- Device technologies
  - Resistive RAM (RRAM); Phase-change memory (PCM); Magnetic RAM (MRAM),…
- Crossbar array: Dot-product calculation
  - Kirchhoff's law: $I_j = \sum_i V_i \cdot G_{i,j}$
- Neuron applications



Daniele Ielmini and Stefano Ambrogio 2020 Nanotechnology 31 092001

History and Development of Neuromorphic Hardware



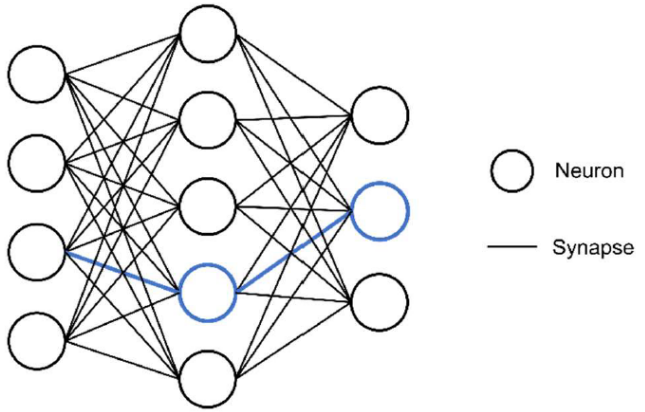
Roy, K., Jaiswal, A. & Panda, P. Towards spike-based machine intelligence with neuromorphic computing. Nature 575, 607–617 (2019). https://doi.org/10.1038/s41586-019-1677-2

Why we need neuromorphic computing

PROPERTIES OF NEUROMORPHIC COMPUTING

Properties of Neuromorphic Computing

- Connectionism: The capability of learning by linking a large number of simple units
  - Learning: Finding the appropriate weight of synapses
- Parallelism
  - Massive number of ‘weak’ units work parallelly
- Asynchrony
  - Synchrony increases overhead, limiting the efficiency
    - Clock tree consumes 20~45% of power



Properties of Neuromorphic Computing

- Spiking nature of information processing
  - Spiking neural network model
    - Spike times and delay convey information
  - Advantages
    - Data can be transferred asynchronously
    - Good for dynamic data processing
    - Non-linearity
    - Energy efficient
  - Challenge: Training and topology
- On-device learning

Properties of Neuromorphic Computing

- Local learning: Update weight without global information
  - Training in non-spiking neural networks: Backpropagation
    - Weight transport problem
    - Update locking problem
  - Local learning method: Synaptic weight updated by connected neuron activity
    - Spike timing dependent plasticity (STDP)
- Sparsity: Very few neurons activate simultaneously
  - Spikes generated only due to dramatic changes in signal : Temporal sparsity
    - Event-based cameras
  - Spike generated only when membrane potential exceeds threshold: Spatial sparsity
    - ReLU in non-spiking networks
  - Sparsity in connection: Structural sparsity

Properties of Neuromorphic Computing

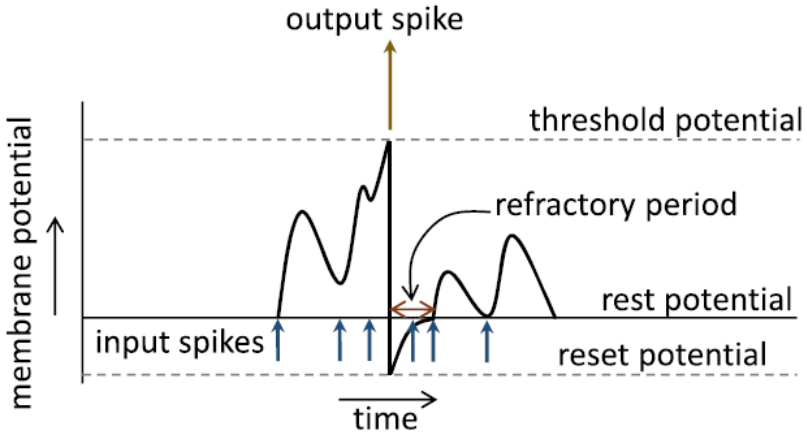
- Analog computing
  - Less computational elements
  - Purposes to be analog
    - Neuron dynamic
      - Faster and more energy efficient with analog system
      - One-one correspondence
      - Disadvantage: Difficult to configure and debug
    - Synaptic operations
      - Ohm’s and Kirchhoff’s law
- In-memory computing
  - Neuron: Memory + computation
  - Free of von Neumann bottleneck
  - Hybrid approach: Near-memory computing

How to do neuromorphic computing

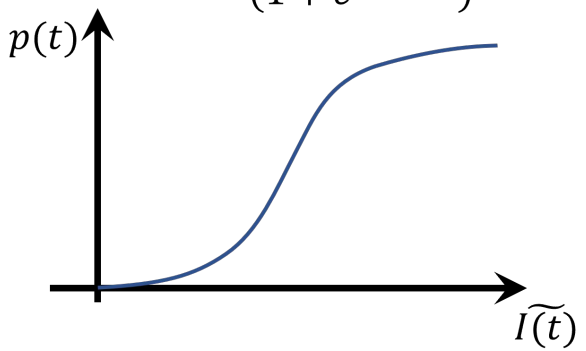
MORE DETAILS IN ALGORITHMS AND HARDWARE

Neuron Models

- Key problems
  - How spikes are generated
  - How information is encoded
- Neuron model: Spike generation
  - Leaky Integrate and Fire (LIF) model
    - Can be derived from Hodgin-Huxley model
    - Neuron state: Membrane potential
      - Lifted by input spikes
      - Leak when no input arrives
    - Spike generation: Membrane potential exceeds threshold
  - Stochastic neuron model
    - Probability of firing follows a non-linear function



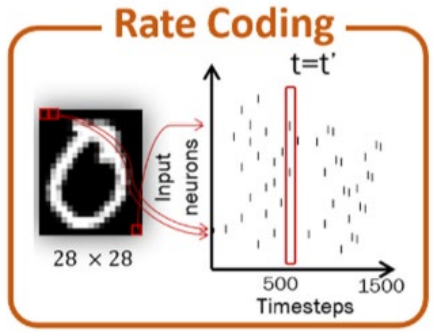
$$p(t) = \frac{1}{(1 + e^{-w \cdot \widetilde{I(t)}})}$$

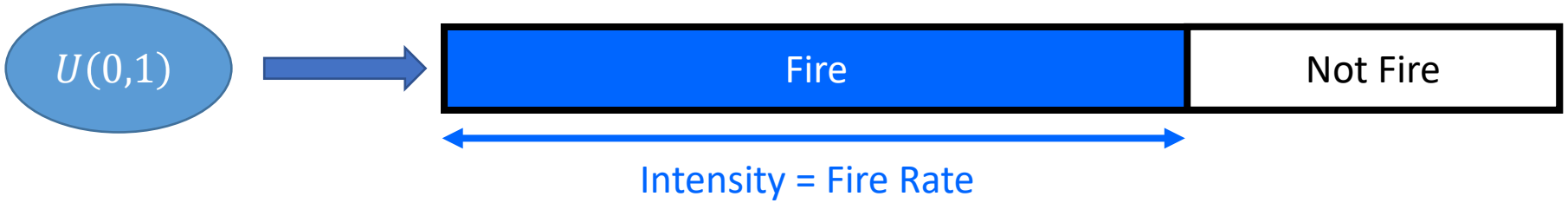


Nitin Rathi, Indranil Chakraborty, Adarsh Kosta, Abhronil Sengupta, Aayush Ankit, Priyadarshini Panda, and Kaushik Roy. 2023. Exploring Neuromorphic Computing Based on Spiking Neural Networks: Algorithms to Hardware. ACM Comput. Surv. 55, 12, Article 243 (December 2023), 49 pages. https://doi.org/10.1145/3571155

Input Encoding Framework

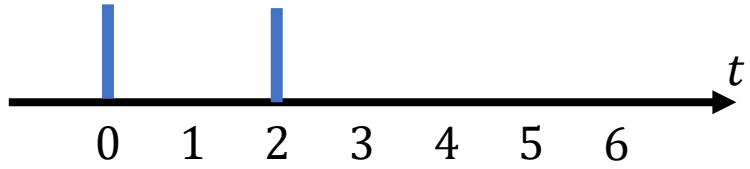
- Encoding framework: How analog values are encoded in spike trains
  - Quantization error
- Rate encoding: Mean firing rate
  - Poisson encoding
    - Discretization error determined by number of timesteps
      - High latency for high accuracy



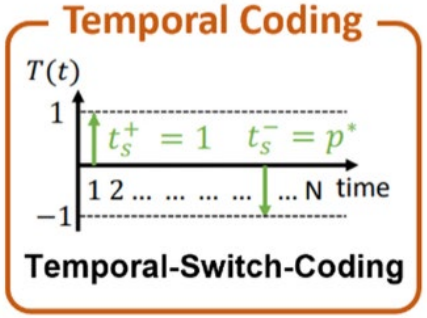


Nitin Rathi, Indranil Chakraborty, Adarsh Kosta, Abhronil Sengupta, Aayush Ankit, Priyadarshini Panda, and Kaushik Roy. 2023. Exploring Neuromorphic Computing Based on Spiking Neural Networks: Algorithms to Hardware. ACM Comput. Surv. 55, 12, Article 243 (December 2023), 49 pages. https://doi.org/10.1145/3571155

Input Encoding Framework



- Temporal encoding: Time instances for encoding
  - Logarithmic Temporal Coding
    - Example: $5 = 0000101_2$
  - Rank Order Coding: Firing order of neurons
    - Time-to-First-Spike: Each neuron fires once
  - Temporal Switch Coding: Time difference between spikes
    - Good energy-efficiency
  - Challenge: Lack of training algorithm
- Encoding layer: Train the encoding process
  - Encoding function: Neural network to convert analog values to spike trains



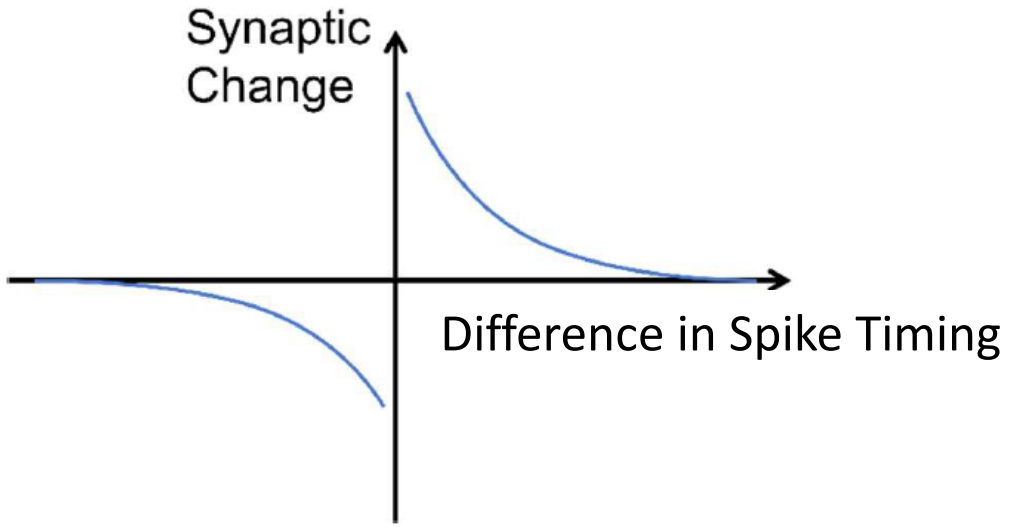
Nitin Rathi, Indranil Chakraborty, Adarsh Kosta, Abhronil Sengupta, Aayush Ankit, Priyadarshini Panda, and Kaushik Roy. 2023. Exploring Neuromorphic Computing Based on Spiking Neural Networks: Algorithms to Hardware. ACM Comput. Surv. 55, 12, Article 243 (December 2023), 49 pages. https://doi.org/10.1145/3571155

Input Encoding Framework

- Standard cameras
  - No internal time information
  - Low, fixed sampling rate: Motion blur
  - Bad for low light, high dynamic range environment
- Event-based sensors
  - Log-intensity changes at each pixel
    - Fires a spike when change exceeds threshold
    - Idle when no change detected
  - High temporal resolution, high dynamic change, low power consumption
    - Real-time human-machine interface systems, robotics, wearable electronics, vision-based edge-devices,…
    - Object detection and tracking, gesture recognition, ego motion estimation,…

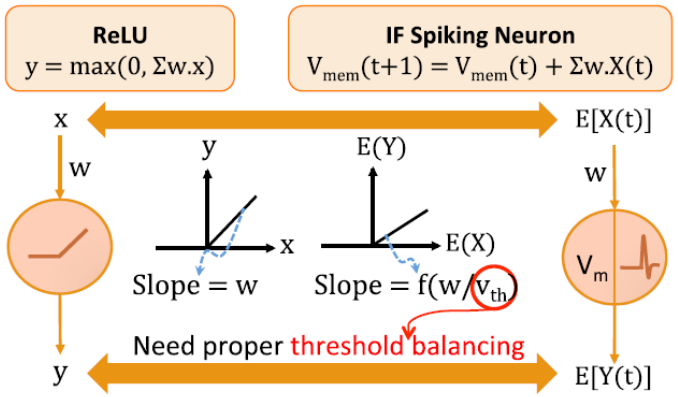
Unsupervised Learning of SNNs: STDP

- Spike Timing Dependent Plasticity (STDP)
  - Weight update according to difference in spike timing of the connected neurons
  - Each synapse updated independently
    - Stability problem
- Limitation
  - Not working for high-level features: Shallow networks



Supervised Learning: ANN-to-SNN Conversion

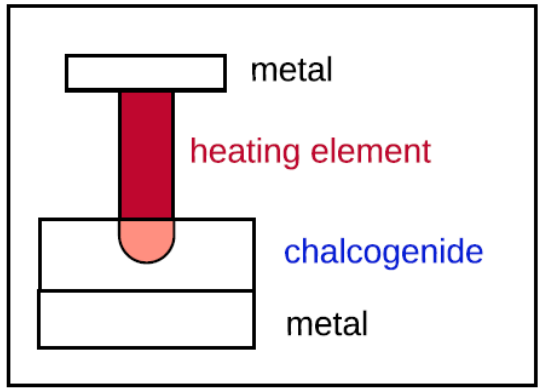
- ANN-to-SNN conversion procedure
  - Train ANN with ReLU neurons
    - Restrictions: No bias, average pooling, no batch normalization
  - Iso-architecture SNN initialized with the trained weight
    - ReLU can be mapped to IF neuron with small loss
      - Threshold balancing
- Major draw back: Time information is not utilized.

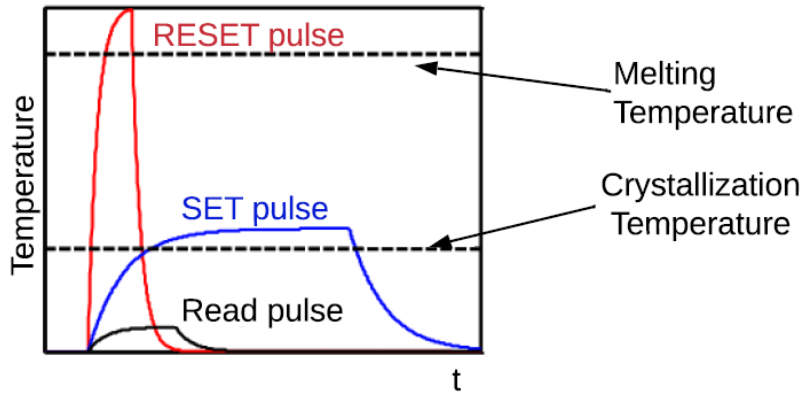


Nitin Rathi, Indranil Chakraborty, Adarsh Kosta, Abhronil Sengupta, Aayush Ankit, Priyadarshini Panda, and Kaushik Roy. 2023. Exploring Neuromorphic Computing Based on Spiking Neural Networks: Algorithms to Hardware. ACM Comput. Surv. 55, 12, Article 243 (December 2023), 49 pages. https://doi.org/10.1145/3571155

Hardware: Phase Change Memory

- Phase-Change Memory (PCM)
  - Structure: Metal electrodes/chalcogenide/resistive electrode/metal electrode
  - State
    - Amorphous: High resistance
    - Crystalline: Low resistance
- Integrate-and-Fire neuron function
  - Set: Apply medium voltage pulse
  - Reset: Apply high voltage short pulse
- Advantage: No need of erasing

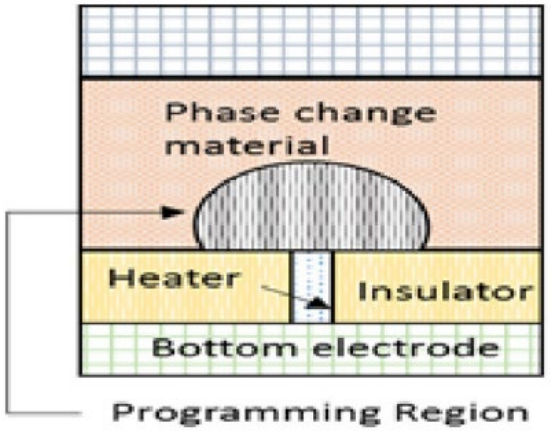


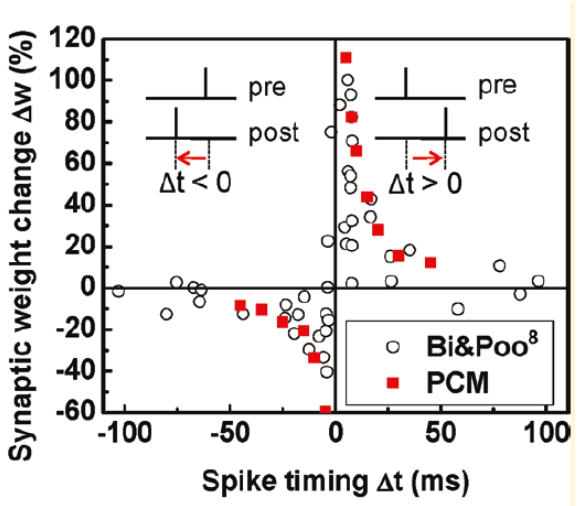


Varshika, M.L.; Corradi, F.; Das, A. Nonvolatile Memories in Spiking Neural Network Architectures: Current and Emerging Trends. Electronics 2022, 11, 1610. https://doi.org/10.3390/electronics11101610

Hardware: Phase Change Memory

- Synaptic function
  - Intermediate states: Proportion of amorphous and crystalline region
  - STDP: Pre- and post-synaptic pulses applied at top and bottom electrodes
- Challenge
  - Reliability
  - Resistance drift over time
  - Intermediate state in small PCM
  - Endurance



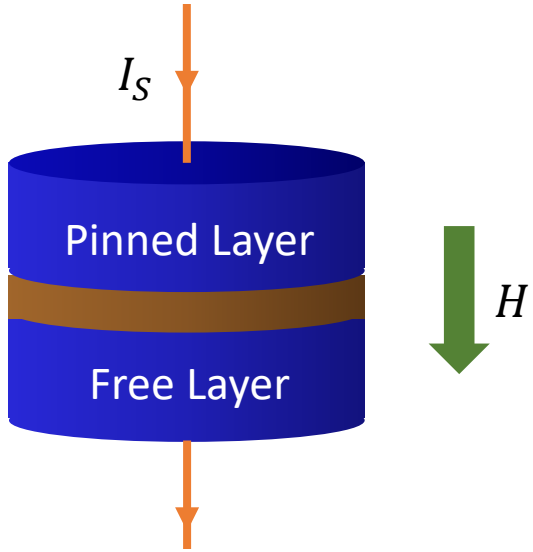


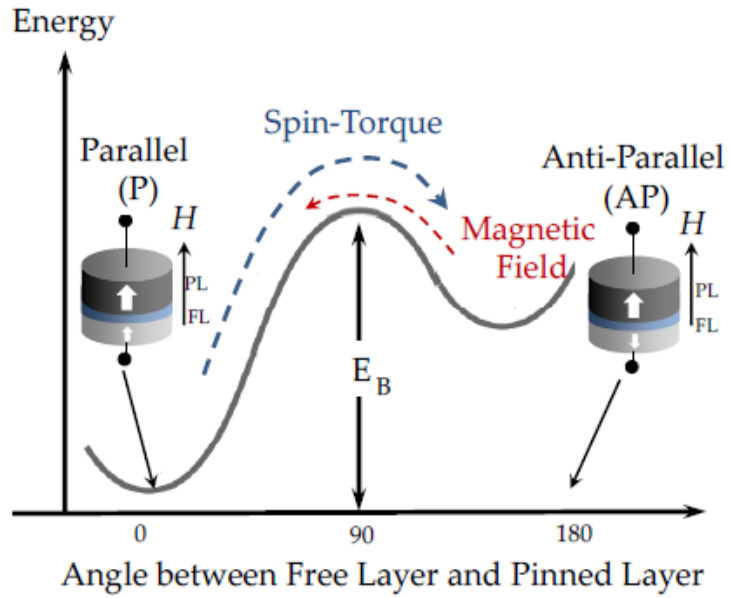
Jangra, P., Duhan, M. Performance-based comparative study of existing and emerging non-volatile memories: a review. J Opt 52, 2395–2409 (2023). https://doi.org/10.1007/s12596-022-01058-w

Kuzum, Duygu, et al. "Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing." *Nano letters* 12.5 (2012): 2179-2186.

Hardware: Spintronic Devices

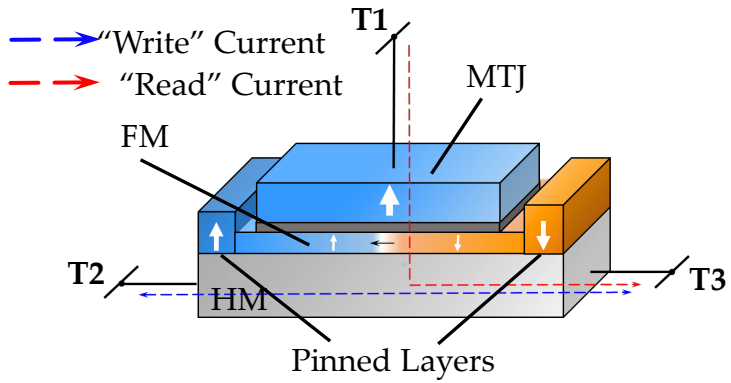
- Magnetic tunnel junction (MTJ) structure
  - Pinned Layer/ Oxide Layer/ Free Layer
- Device States
  - Parallel state – Low resistance
  - Anti-parallel state – High resistance
  - Difference in Resistance allows informatio encoding
- Manipulation
  - Magnetic field
  - Current: Spin injection

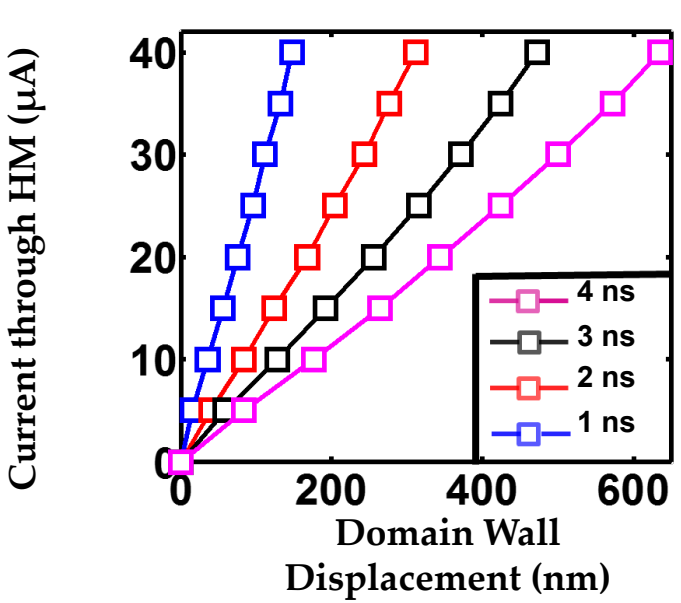


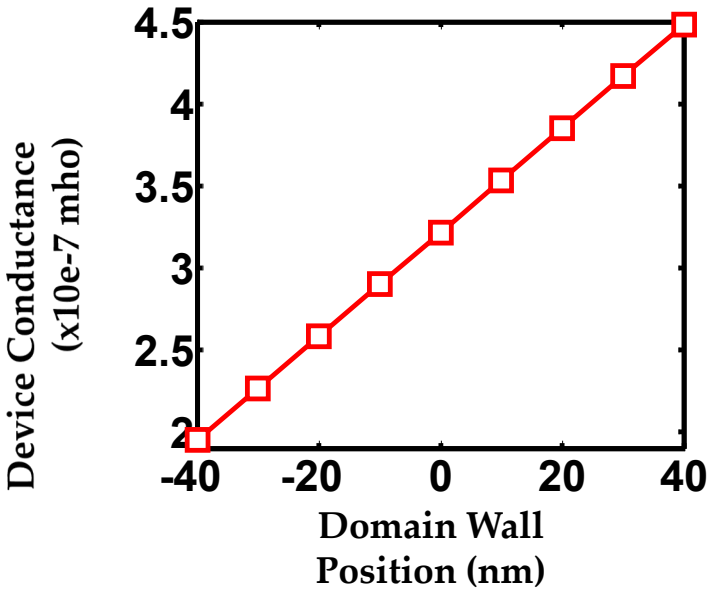


Hardware: Spintronic Devices

- Large MTJs: High barrier height
  - Multiple programmable states: Synapse device





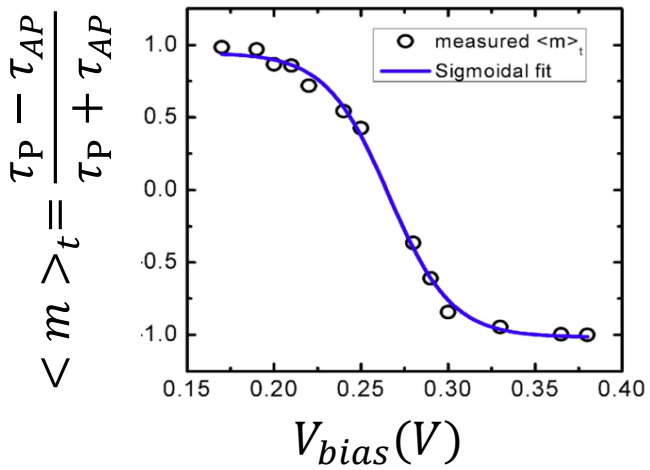


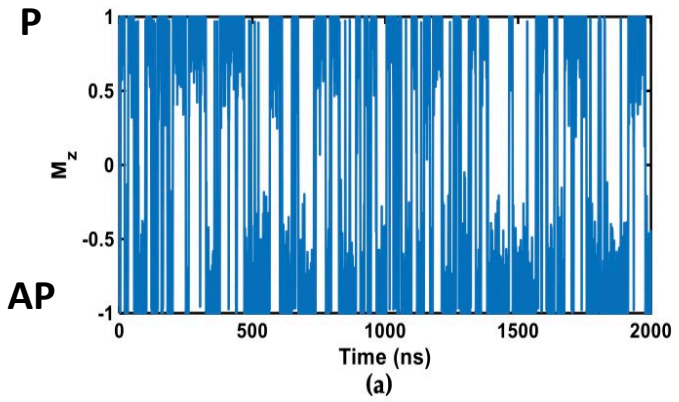
Kaushik Roy, Abhronil Sengupta, and Yong Shim , "Perspective: Stochastic magnetic devices for cognitive computing", Journal of Applied Physics 123, 210901 (2018) https://doi.org/10.1063/1.5020168

Hardware: Spintronic Devices

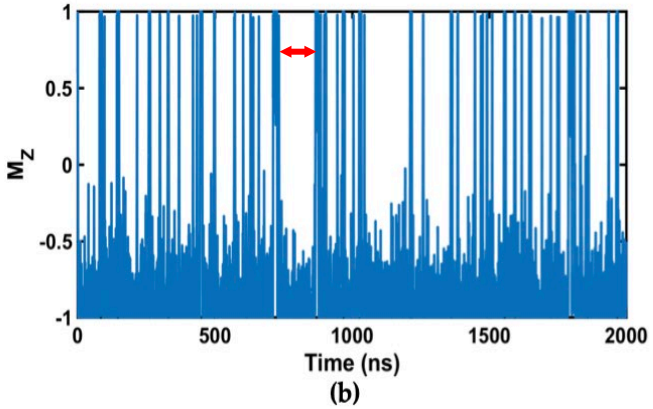
- Mono-domain – Single bit
- Low barrier height
  - Stochastic switching
  - Telegraphic switching

$$< m >_t = \frac{\tau_P - \tau_{AP}}{\tau_P + \tau_{AP}}$$





$V_I = 0$ mV
$\tau_{AP} = 3.14$ ns
$\tau_P = 3.10$ ns



$V_I = 10$ mV
$\tau_{AP} = 7.60$ ns
$\tau_P = 0.68$ ns

Mukund Bapna and Sara A. Majetich , "Current control of time-averaged magnetization in superparamagnetic tunnel junctions", Applied Physics Letters 111, 243107 (2017) https://doi.org/10.1063/1.5012091

THANK YOU