RE-ENGINEERING COMPUTING WITH NEURO-MIMETIC DEVICES, CIRCUITS, AND ALGORITHMS

Kaushik Roy

Abhronil Sengupta, Gopal Srinivasan, Aayush Ankit, Priya Panda, Xuanyao Fong, Deliang Fan, Jason Allred

School of Electrical & Computer Engineering
Purdue University

Center for Spintronics, National Science Foundation, DARPA, Vannevar Bush Fellowship, Office of Naval Research, Semiconductor Research Corporation, Intel

Human vs. Machine Chronicles

1997

IBM Deep Blue vs. Kasparov

IBM RS/6000 32-node server (Power2 + 8 dedicated chips)

~15000W

2011

IBM Watson vs. Brad Ritter & Ken Jennings

90 Power 750 Express servers (4 8-core CPUs)

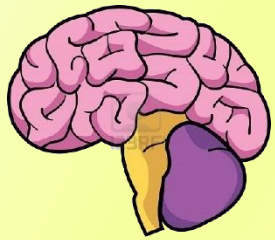
~200000W

2016

Google AlphaGo vs. Lee Sedol

(1920 CPUs, 280 GPUs)

~300000W



20W

WHAT ARE THE ISSUES WITH DLN?

- Requires massive amount of training data
  - Learn with much less data
- Supervised learning
  - Need for higher bio-fidelity
- Incremental/adaptive learning is difficult – catastrophic forgetting
  - Life-long learning
- Huge power consumption
  - Event driven evaluation can help; Approximate hardware
- Well-suited for image, speech, text recognition..
  - Need for cognitive systems to perform larger range of functions – not just sensory processing, but also reasoning and decision making
- Can neuro-mimetic devices help?
  - von-Neumann architecture not suitable
    - Need for in-memory-computing, efficient neurons and synapses

Neuromorphic Computing: An Overview

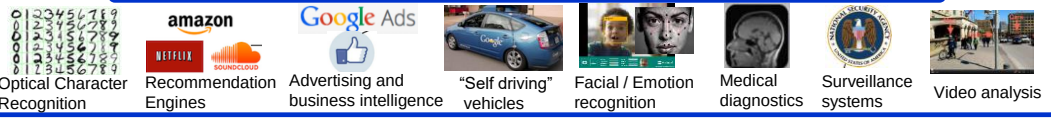
Explore neuromorphic computing models inspired from hierarchical layer arrangement and spiking nature of brain networks

Leverage from latest development in neuroscience on visual attention to design energy efficient hardware for deep learning neural networks

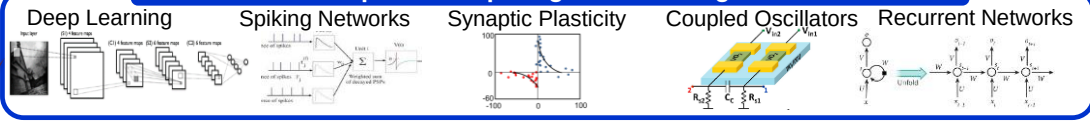
Design programmable and scalable hardware fabrics and explore circuit optimizations for achieving high connectivity

Investigate device physics to mimic "neuron/ synapse" functionalities

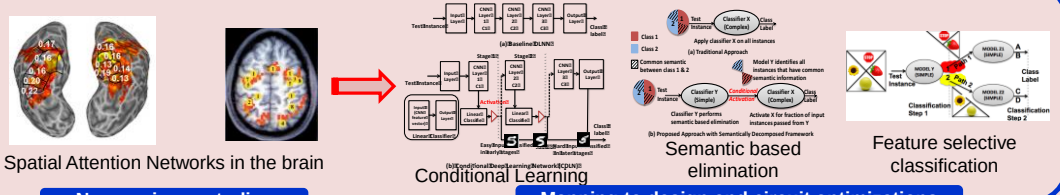
Big Data Analytics



Bio-Inspired Computing and Learning Models



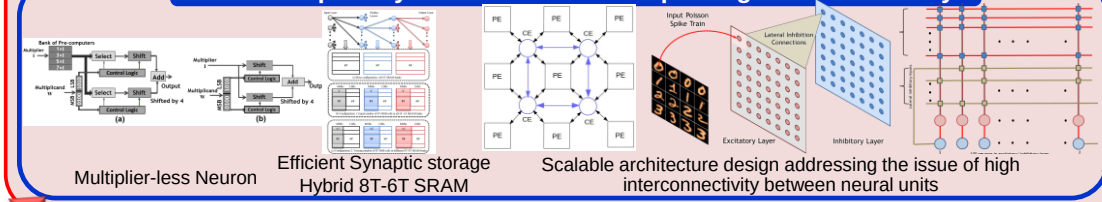
Visual-Attention Inspired Low-Power Algorithms/Hardware



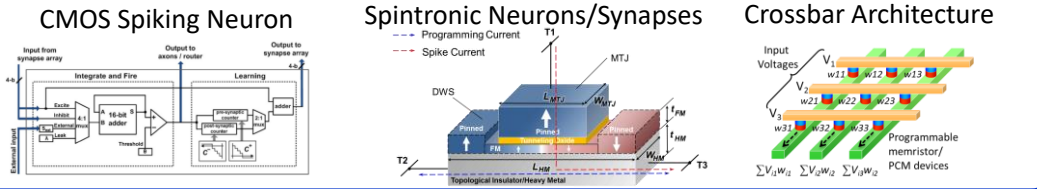
Neuroscience studies

Mapping to design and circuit optimizations

Neuromorphic System Architecture Exploiting Error-Resiliency



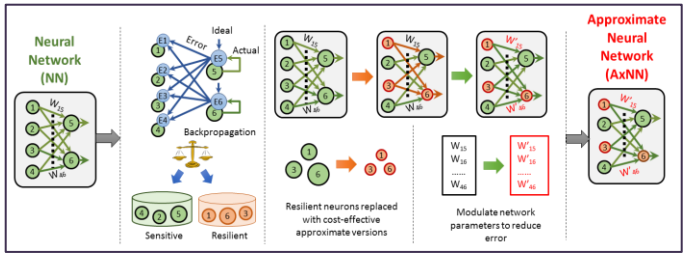
Neuro-mimetic CMOS & Post-CMOS Device Models

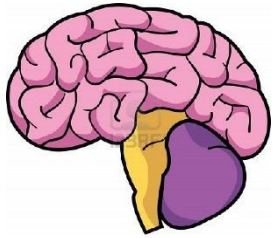


Experiments & Analysis from Neuroscience

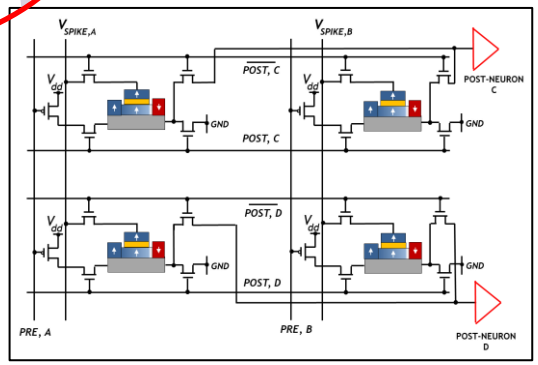
STUDENTS/Post-doc: Abhronil Sengupta, Priyadarshini Panda, Syed Shakib, Gopal Srinivasan, Chamika Liyanagendra, Bing Han, Jason Allred, Parami Wijesinghe, Ayush Ankit, Maryam Parsa, Shuhan Chen and others…

Cognitive Computing





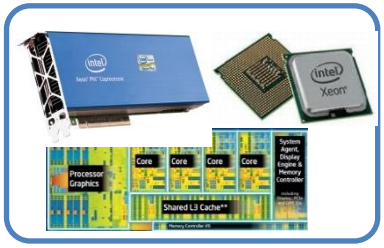
Emerging Devices: STT-devices, Resistive RAMs,..

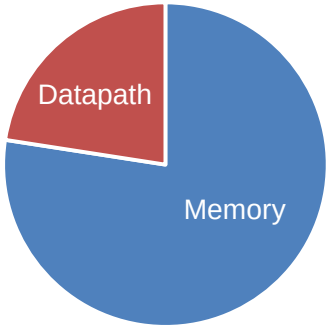


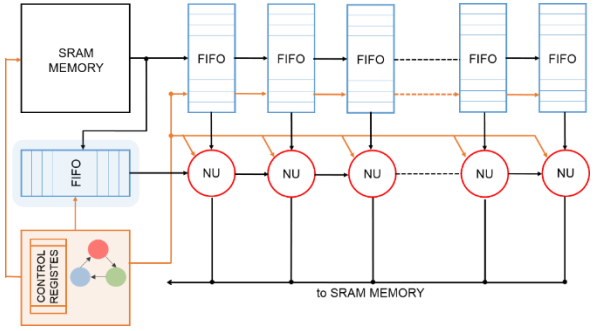
Recognition, Inference, Decision Making to Enable Intelligent Autonomous Systems, Life-long Learning, Neuro-inspired learning

Hardware Accelerators: Image, Speech and Text Recognition

SW (Multicores/GPUs)

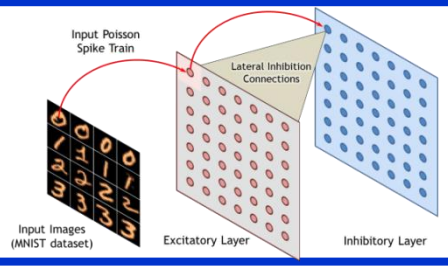




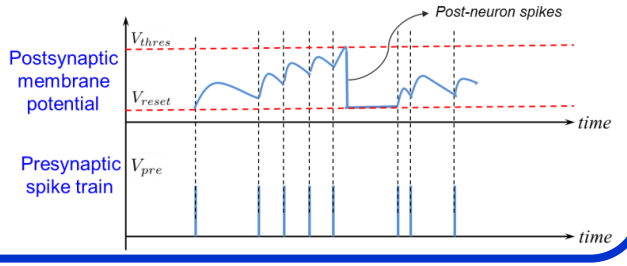


Device/Circuit/Algorithm Co-Design: Spin/ANN

Top-Down Perspective

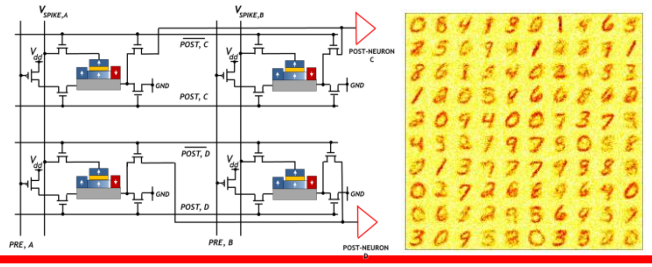


Investigate brain-inspired computing models to provide algorithm-level matching to underlying device physics

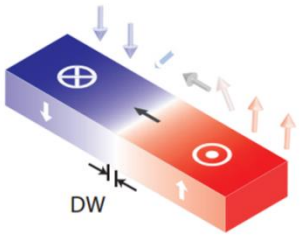


Device-Circuit-Algorithm co-simulation framework used to generate behavioral models for system-level simulations of neuromorphic systems

System Level Solution

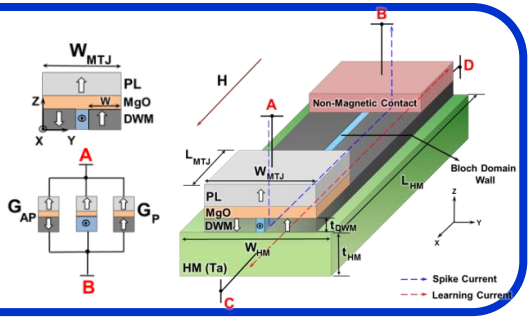


Bottom-Up Perspective



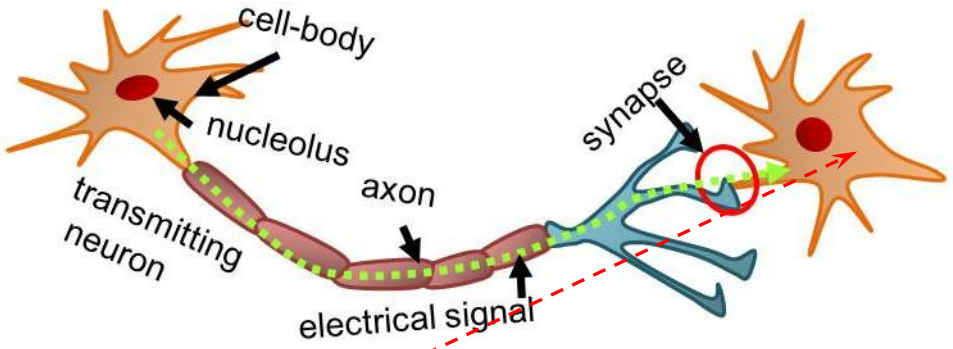
Investigate device physics to mimic "neuron/ synapse" functionalities

Calibration of device models with experiments



SPIKING NEURAL NETWORKS

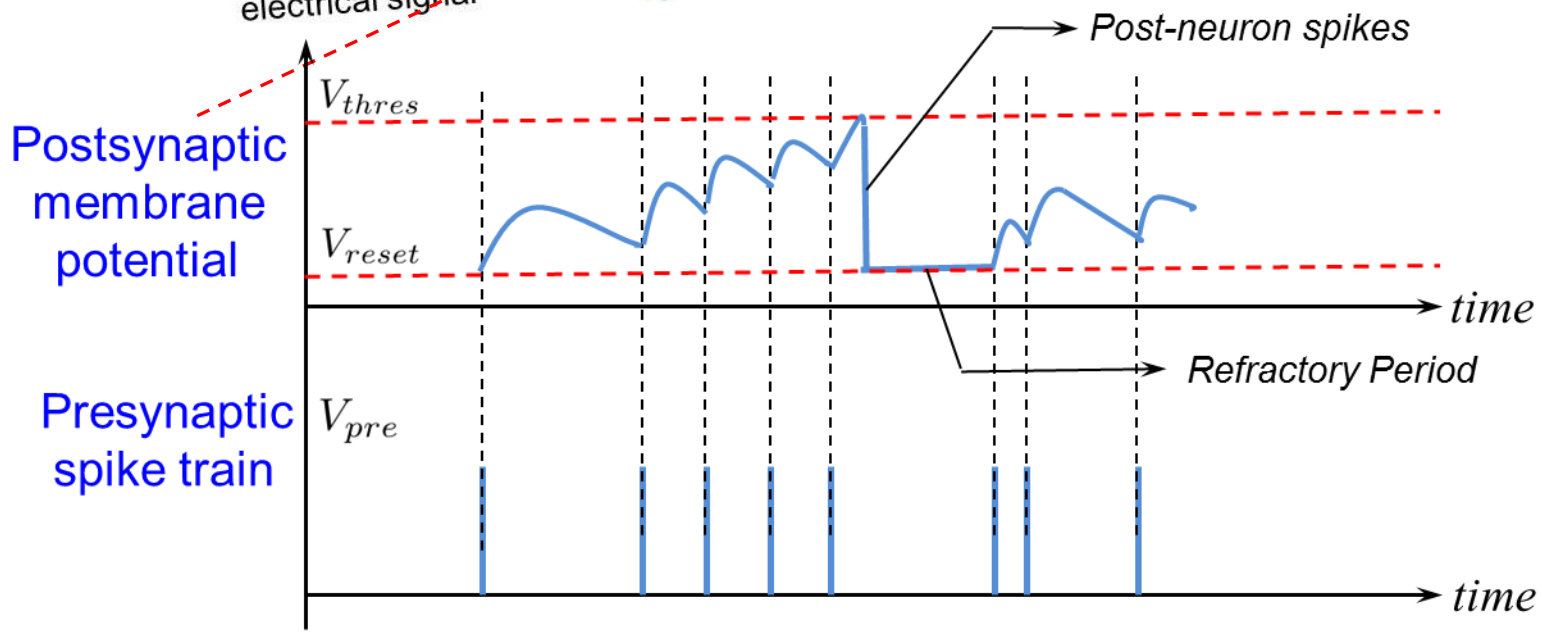
Spiking Neuron Dynamics



Postsynaptic neuron membrane potential *V* given by Leaky-Integrate-Fire model as follows:

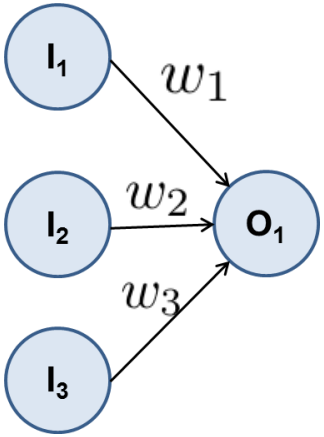
$$\tau \frac{V_{mem}}{dt} = -V_{mem} + R_{mem} \sum_i I_{post,i}$$

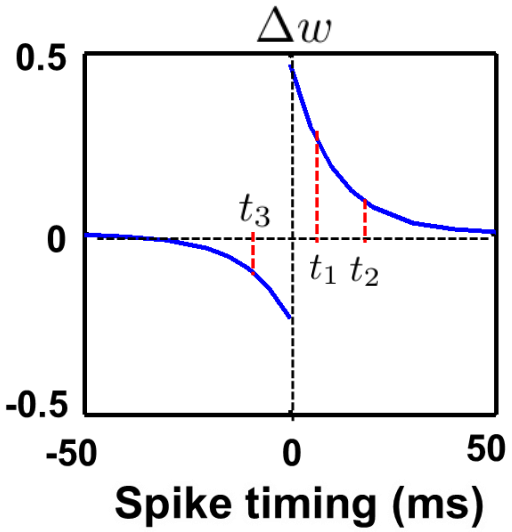
T denotes membrane time constant

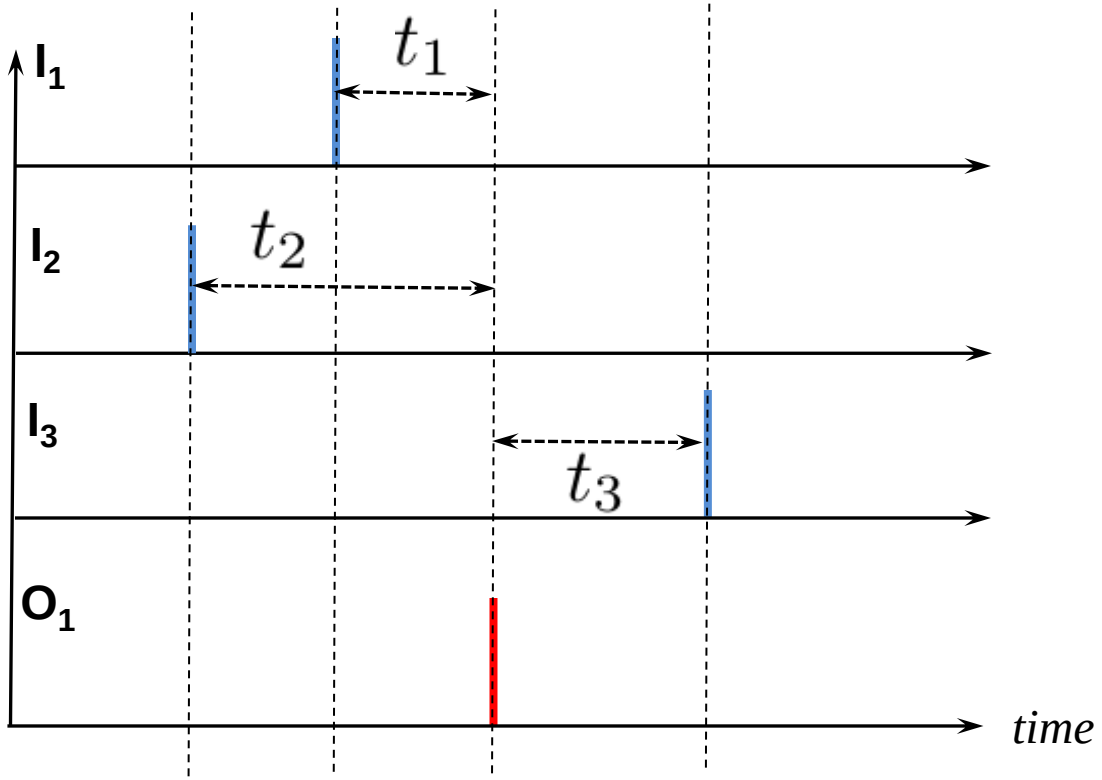


Postsynaptic neuron spikes when membrane potential crosses a certain threshold and gets reset

Spike Timing Dependent Plasticity: Example







Weight Update Equations

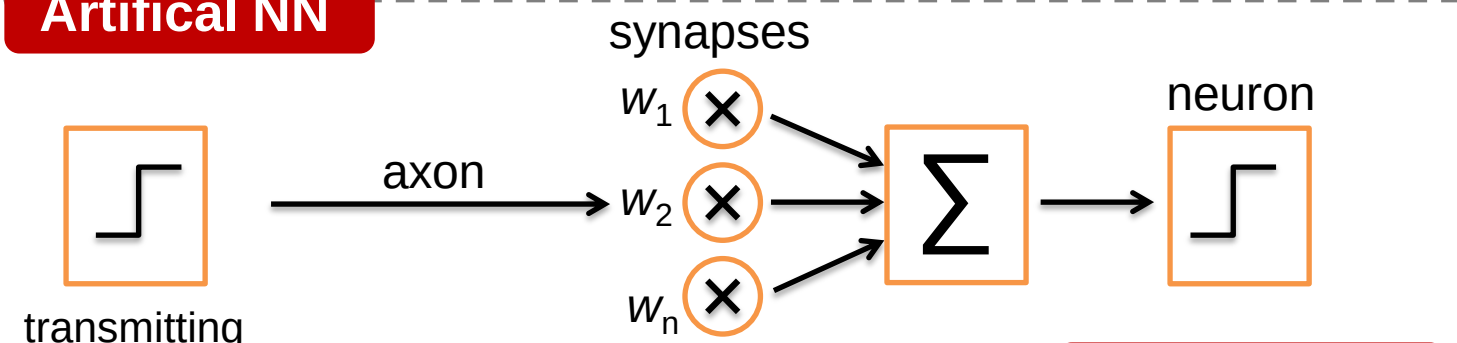
$$w_i^{new} = w_i^{old} + \Delta w(t_i) \times w_{max} \qquad i = 1, 2, 3$$

Strength of the synapse should increase (decrease) as post and pre neurons appear to be temporally correlated (uncorrelated)

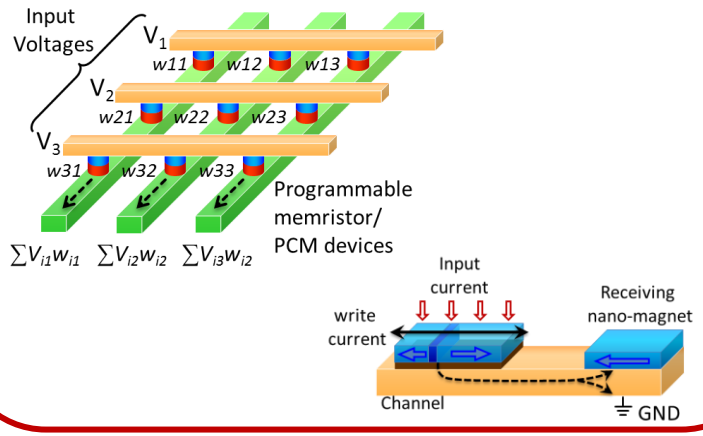
Bi and Poo, The Journal of Neuroscience, December 15, 1998.

Artificial Neural Networks: Simple Model

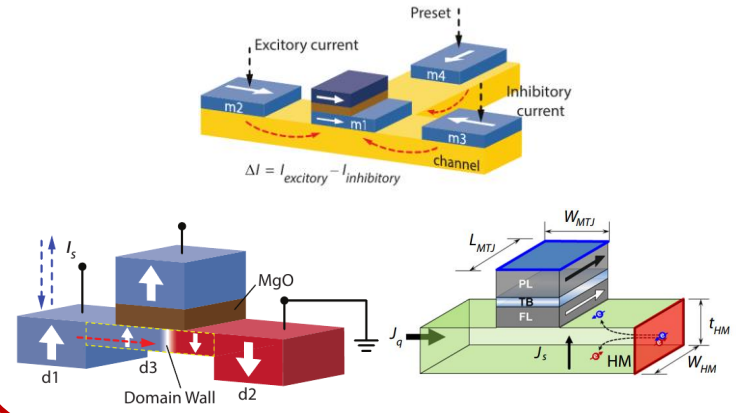
Artifical NN



Cross-bar array of programmable synapses



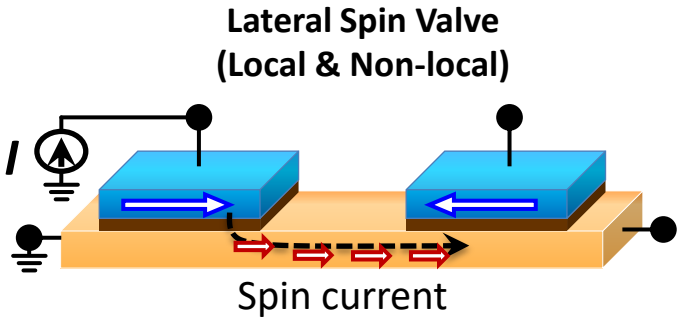
"Spin neurons"

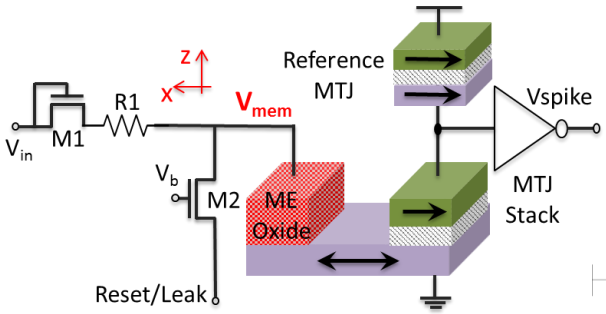


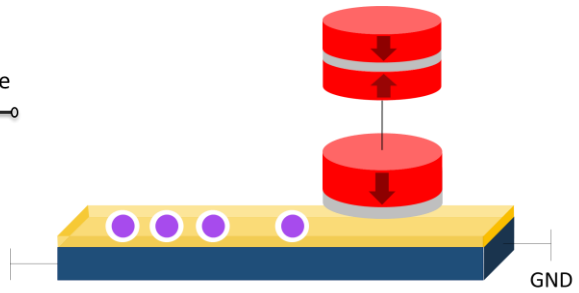
Spintronic neurons operating at ultra-low terminal voltages interfaced with resistive synapses lead to energy-efficient ANNs

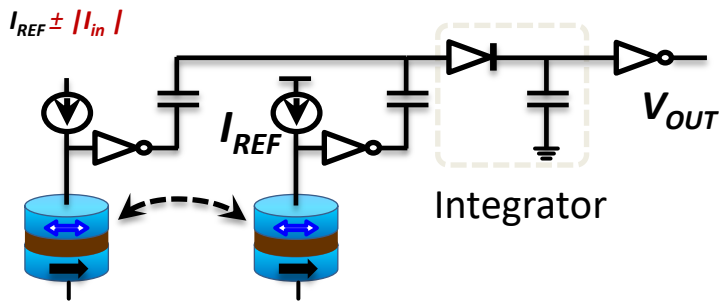
DEVICES: NEURONS, SYNAPSES, IN-MEMORY COMPUTING

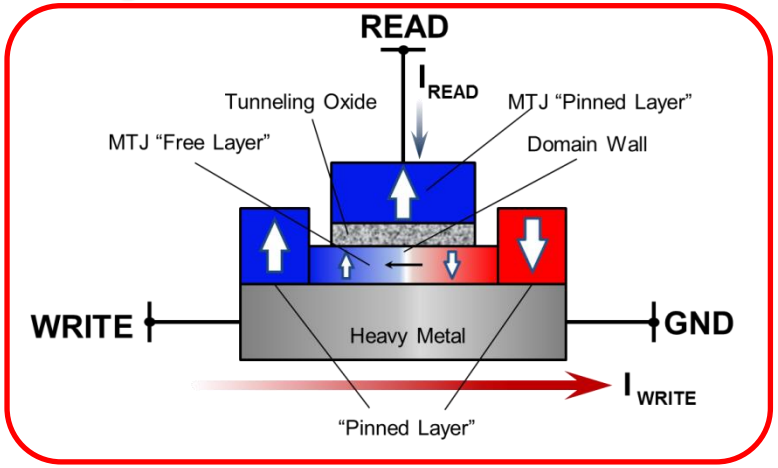
Building Primitives: Memory, Neurons, Synapses

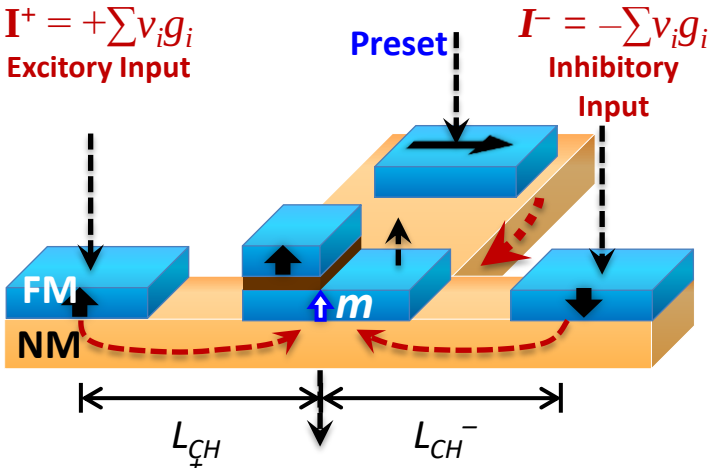


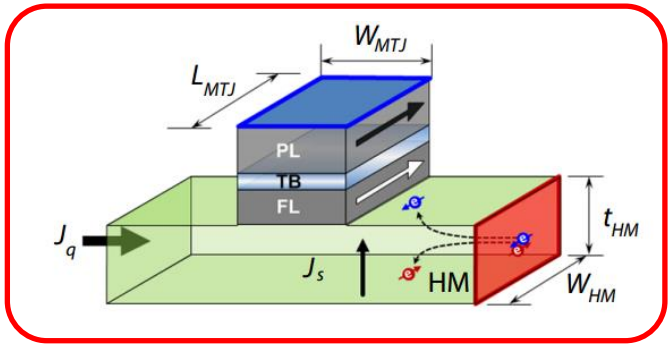






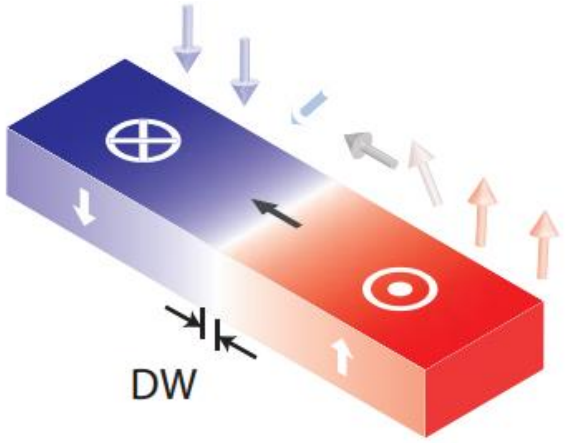


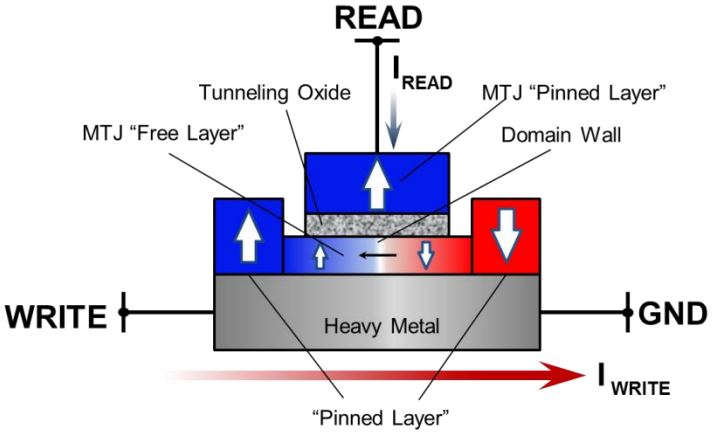




Spin Transfer Torque Induced Domain Wall Motion

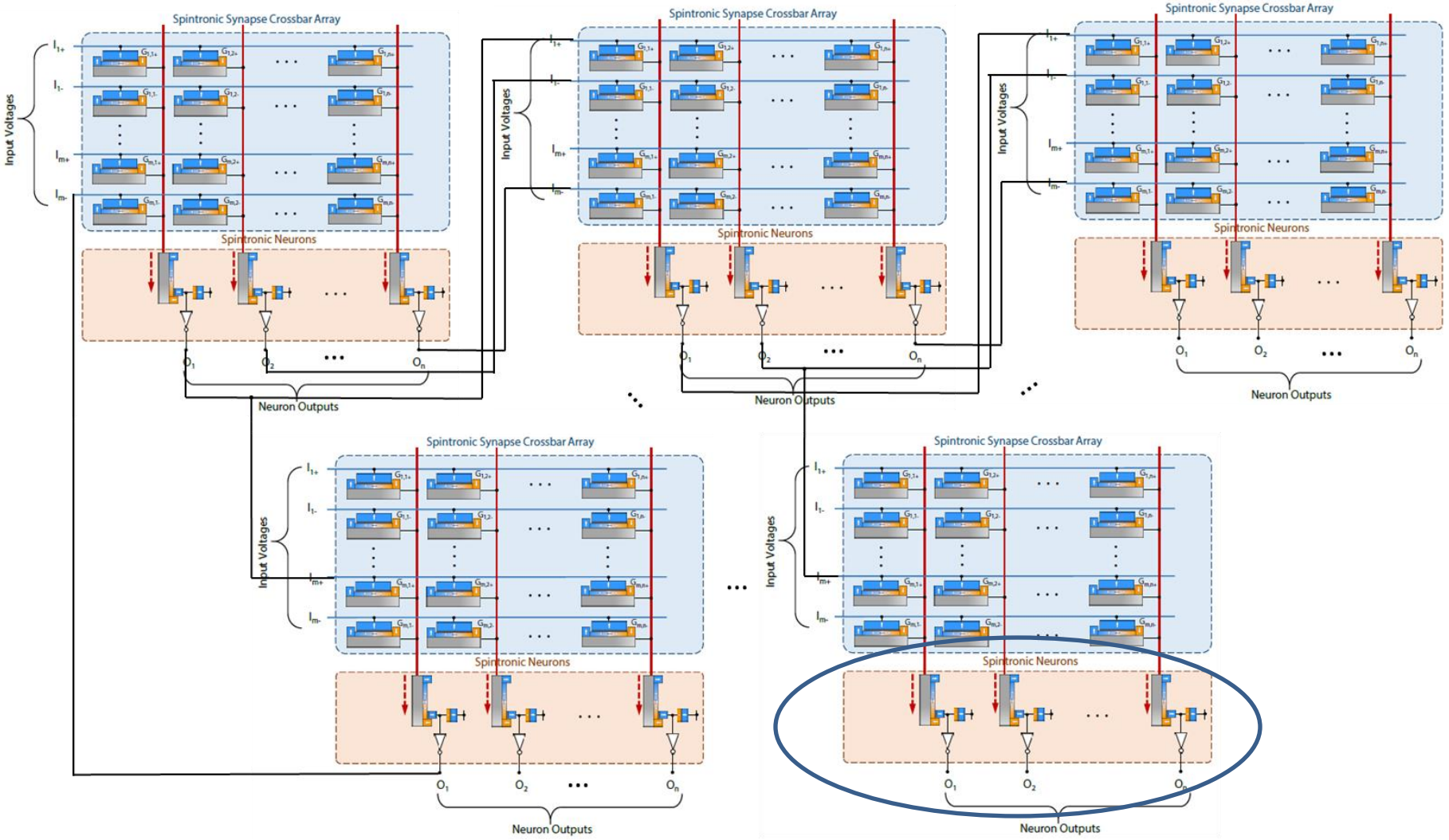
- **Multi-domain magnets consists of a domain wall (DW) separating regions with opposite magnetic polarizations**
- Domain wall can be moved in the direction of electron flow
- MTJ resistance varies with domain wall position
- Decoupled “write” and “read” current paths
- Low current induced Ne’el domain wall motion can be achieved by spin-orbit torque generated by spin-Hall effect from a heavy metal underlayer in presence of DMI



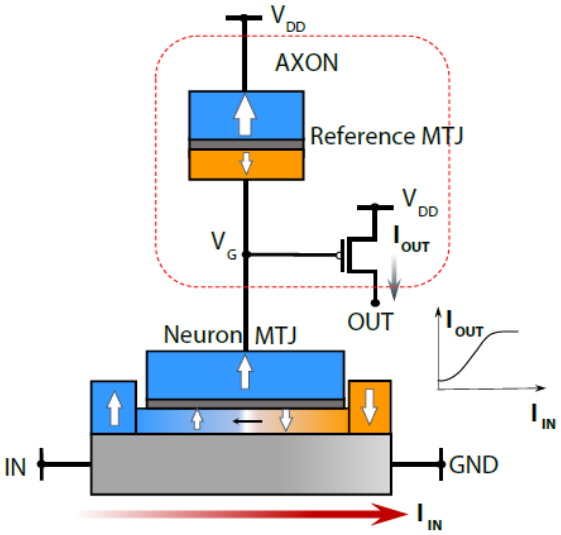


Universal device: Suitable for **memory, neuron, synapse, interconnects**

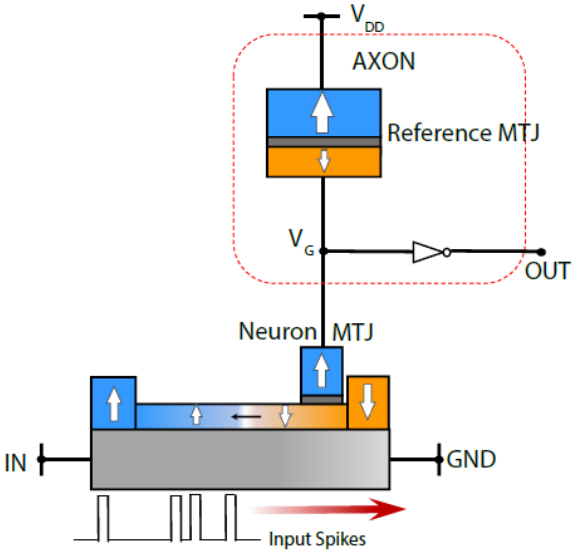
Core Computing Blocks



SHE Induced DW Motion: Neuron



Non-spiking Neuron

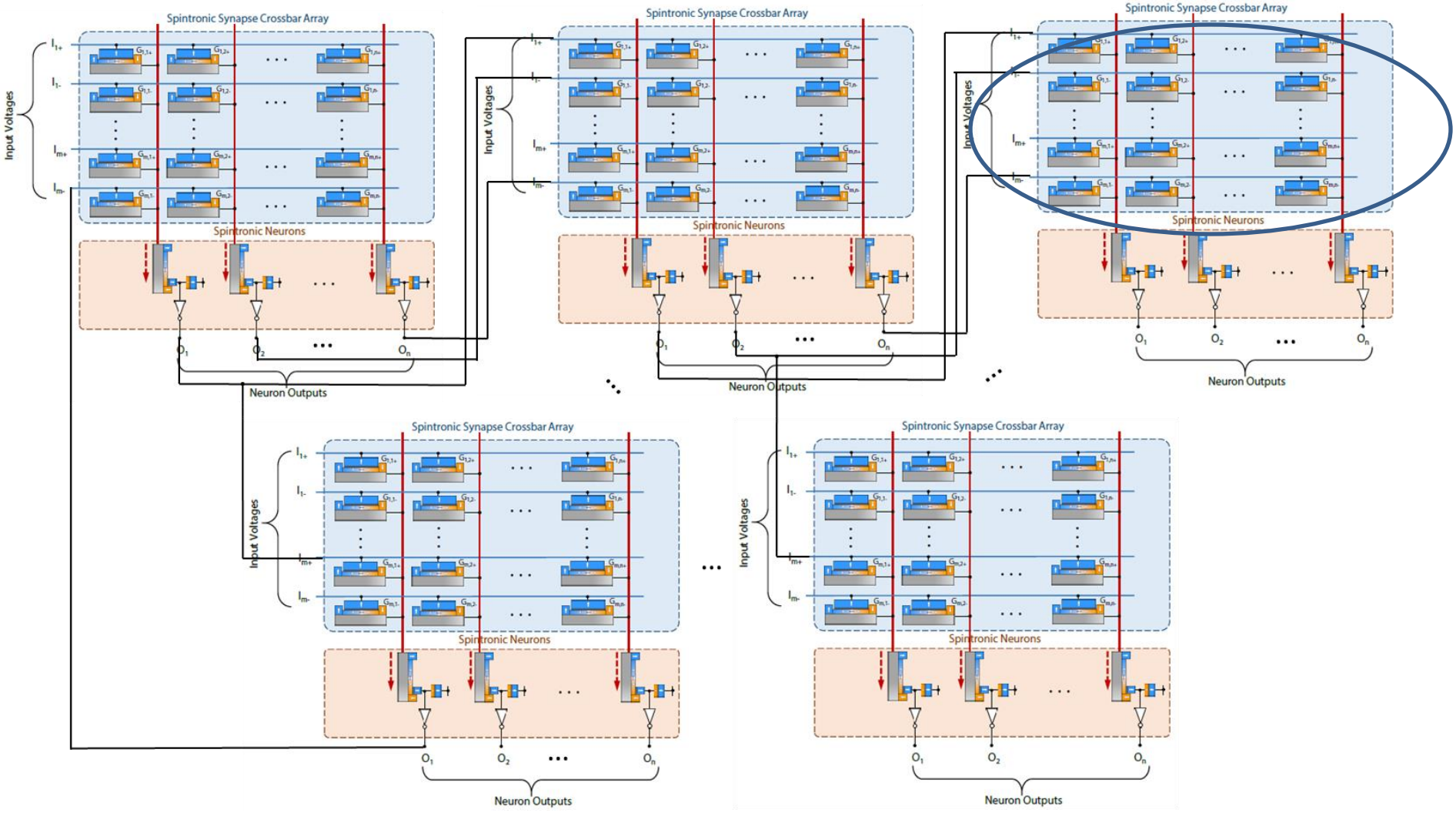


IF Spiking Neuron

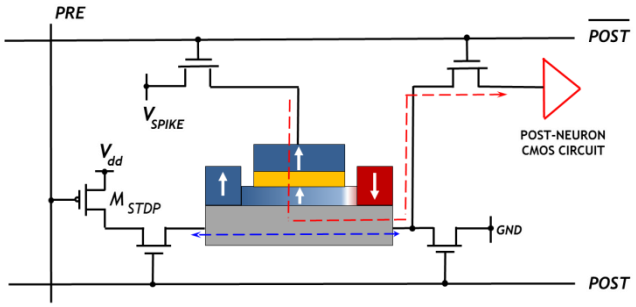
- **Three terminal spintronic device acting as a neuron (with different degrees of bio-fidelity) and synapse**
- The neuron is interfaced with the “axon” circuit to generate a corresponding analog output current with variation in the input current / Integrate-Fire “spiking” neuron can be implemented using a similar device structure where the MTJ is located at the extreme edge of the FM.
- Synapse, acting as the memory element, encodes the weight in its conductance value which is determined by the domain wall position

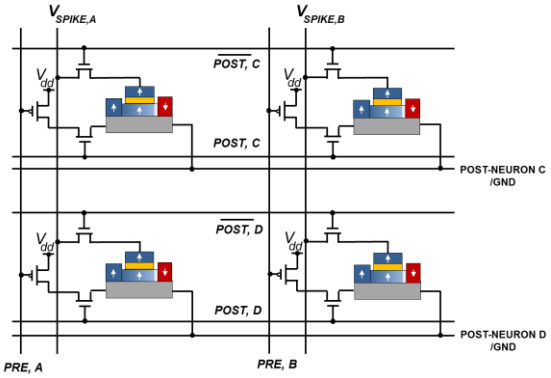
Sengupta, Roy, TBioCAS’16; Sengupta, Roy, TCAS’16

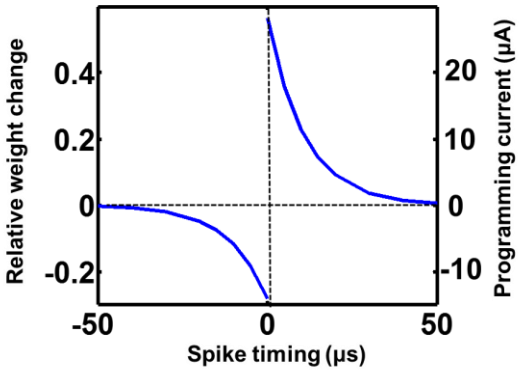
Core Computing Blocks



Self-learning in Spiking Neural Networks



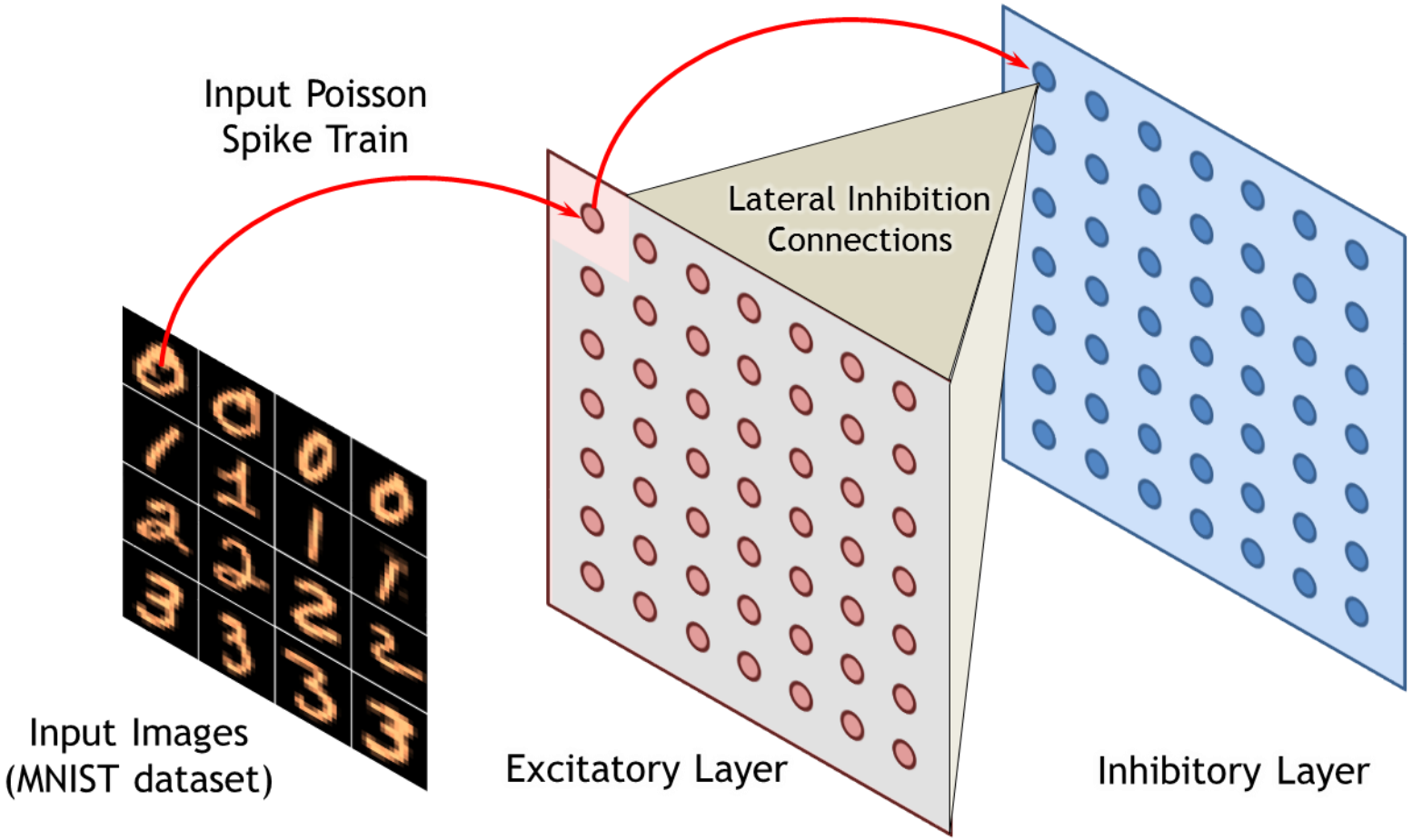




Spike-Timing Dependent Plasticity

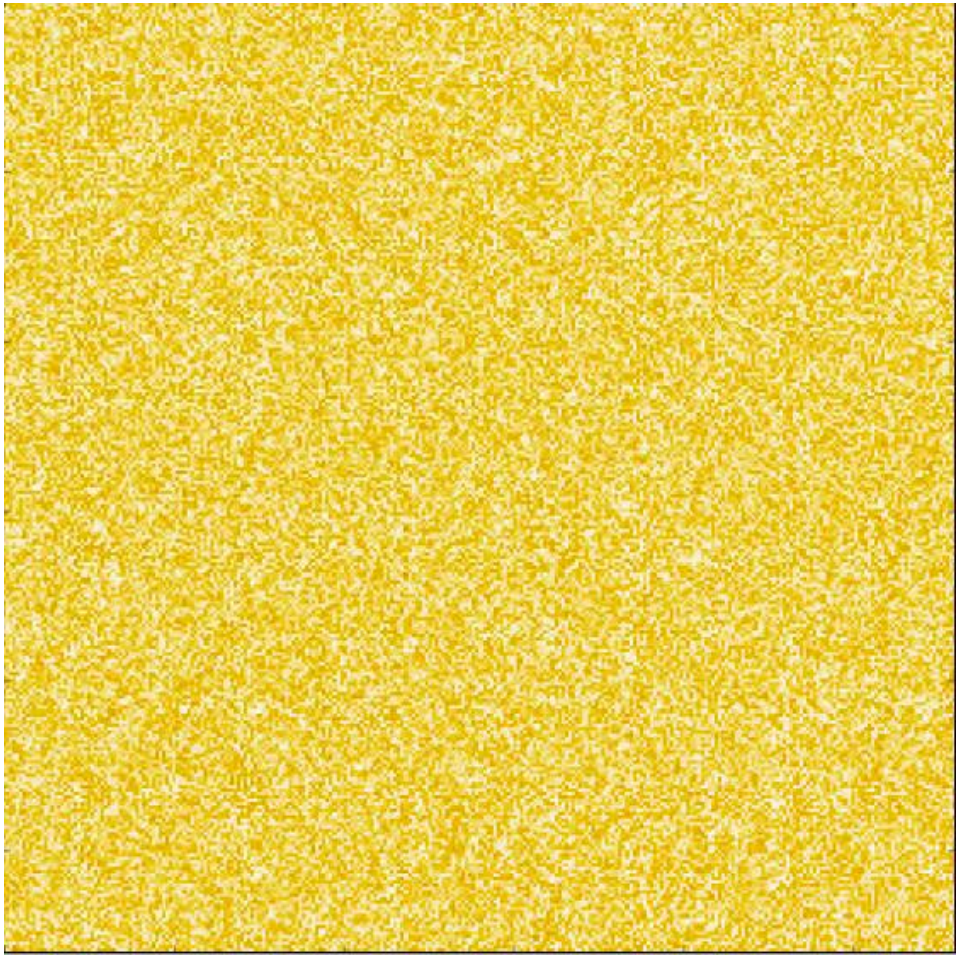
- Spintronic synapse in spiking neural networks exhibits spike timing dependent plasticity observed in biological synapses
- Programming current flowing through heavy metal varies in a similar nature as STDP curve
- Decoupled spike transmission and programming current paths assist online learning
- **48fJ energy consumption per synaptic event which is ~10-100x lower in comparison to emerging devices like PCM**

Network for Pattern Recognition

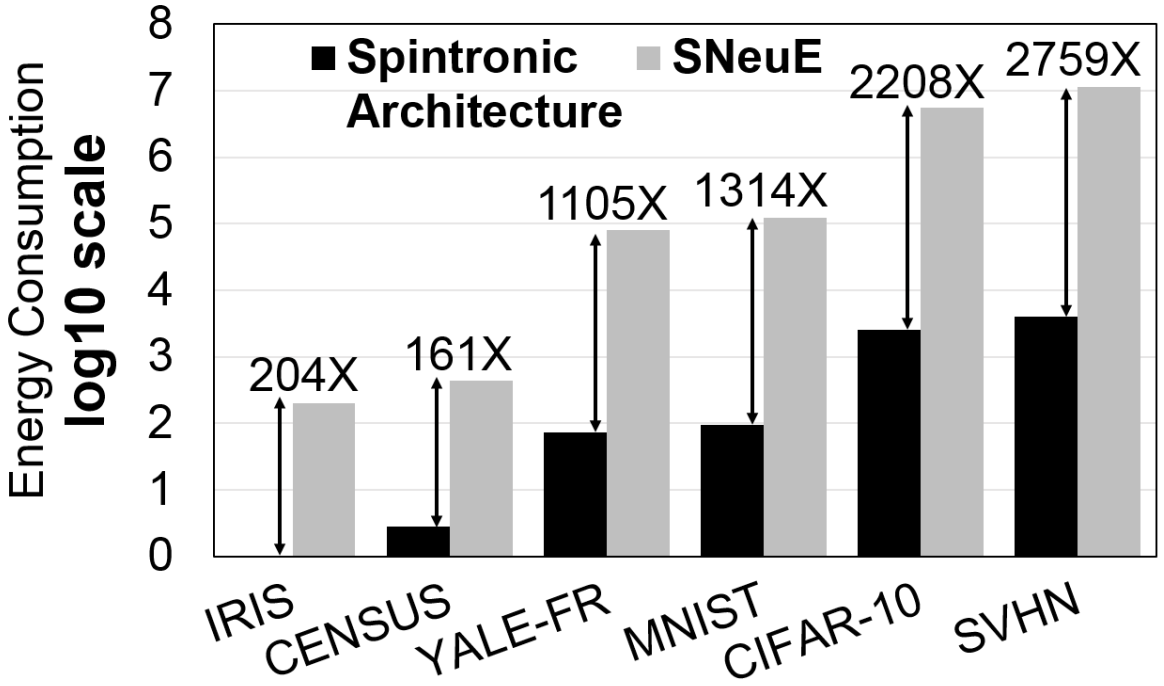


Pattern recognition performed in a network of excitatory spiking neurons in presence of lateral inhibition and homeostasis

Self-Learning in Spiking Neural Network

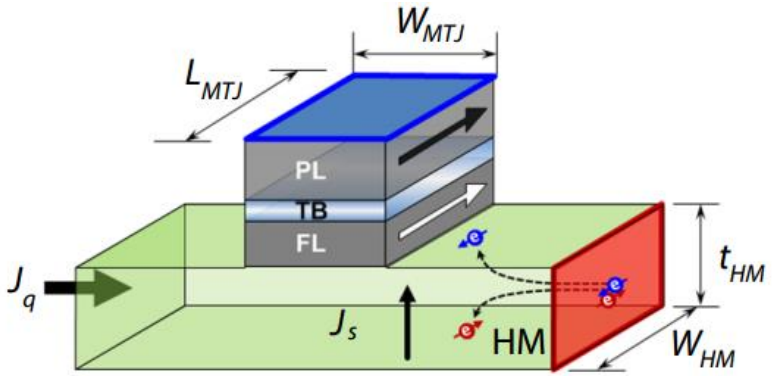


Energy Comparison – Spintronic Engine & CMOS Baseline



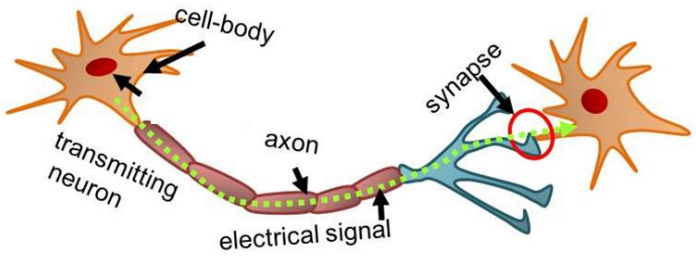
- **Spintronic architecture achieves energy improvements of 204X – 2759X with respect to SNeuE (CMOS baseline) across all benchmarks.**
- Spintronic crossbar enabled in-memory processing enables to overcome the memory domination/bottlenecks in the CMOS engine.
- Spintronic neurons interfaced with spintronic crossbar (SCA) allows energy-efficient inner-product computations.

STOCHASTIC NEURAL NETWORKS: SHE BASED MTJ



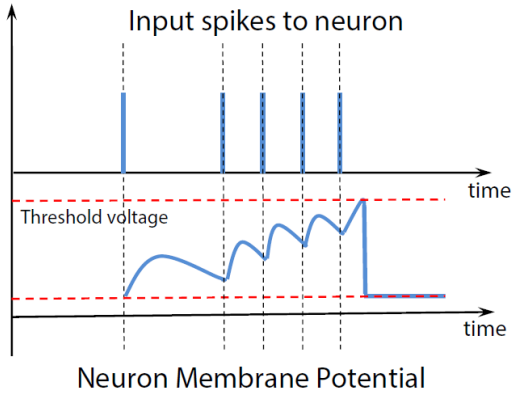
Spiking Neuron Membrane Potential

Biological Spiking Neuron

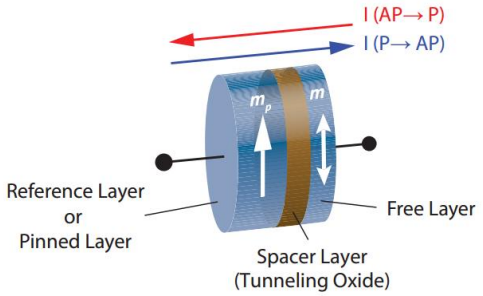


LIF Equation:

$$C\frac{dV}{dt} = -\frac{V}{R} + \sum_j w_j I_{post,j}$$

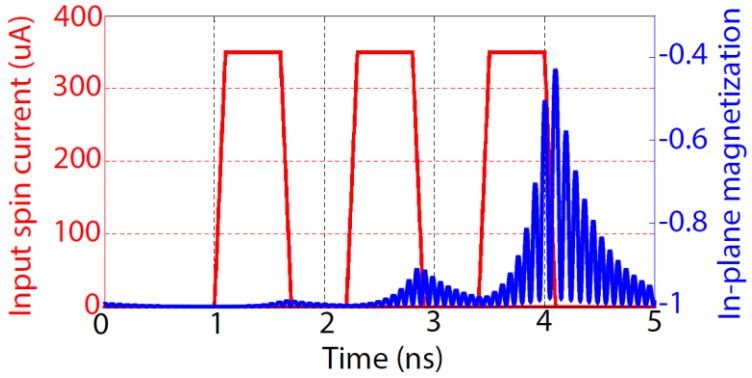


MTJ Spiking Neuron



LLGS Equation:

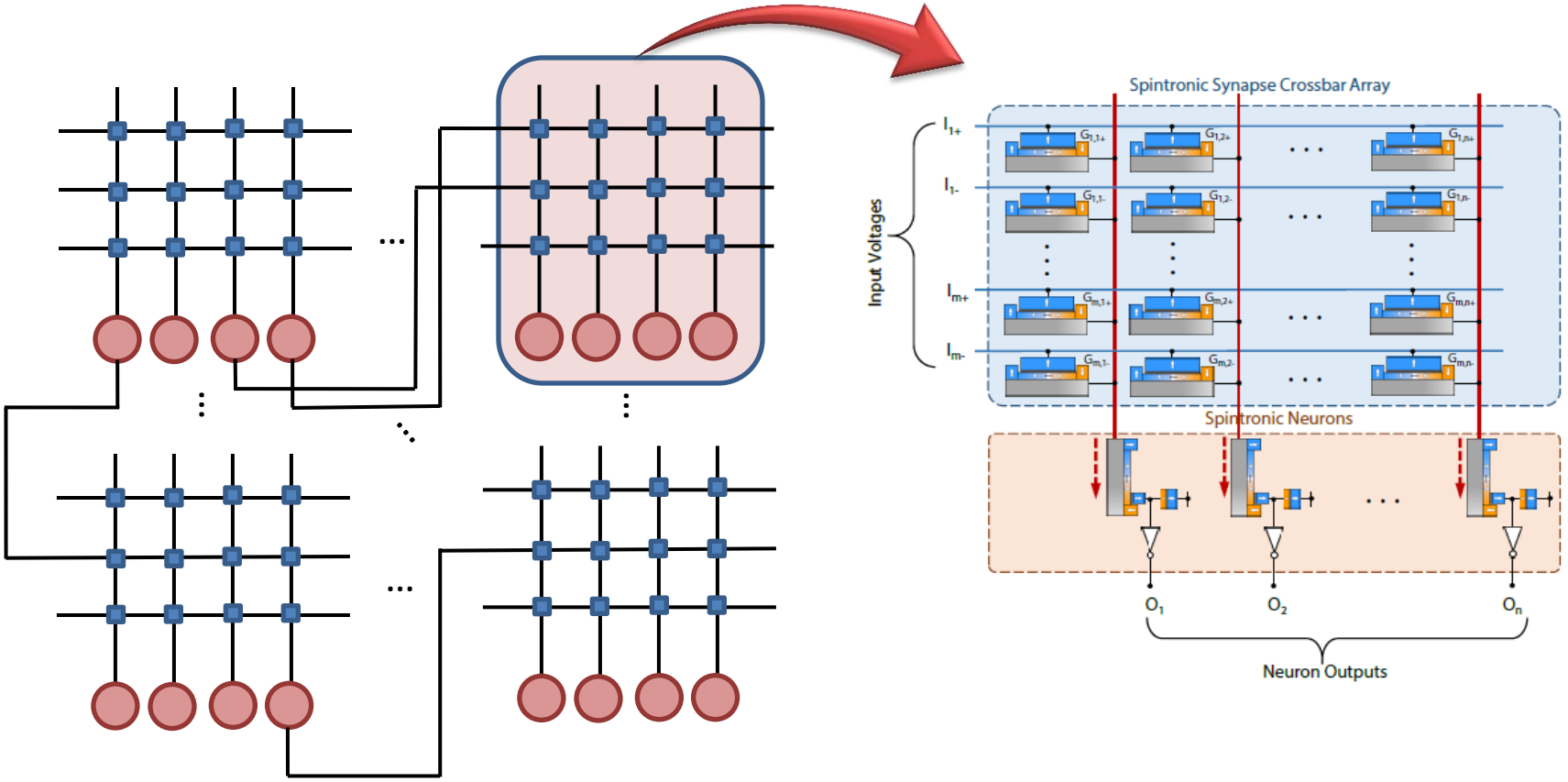
$$\frac{d\hat{\mathbf{m}}}{dt} = -\gamma(\hat{\mathbf{m}} \times \mathbf{H}_{eff}) + \alpha(\hat{\mathbf{m}} \times \frac{d\hat{\mathbf{m}}}{dt}) + \frac{1}{qN_s}(\hat{\mathbf{m}} \times \mathbf{I}_s \times \hat{\mathbf{m}})$$



The leaky fire and integrate can be approximated by an MTJ – the magnetization dynamics mimics the leaky fire and integrate operation

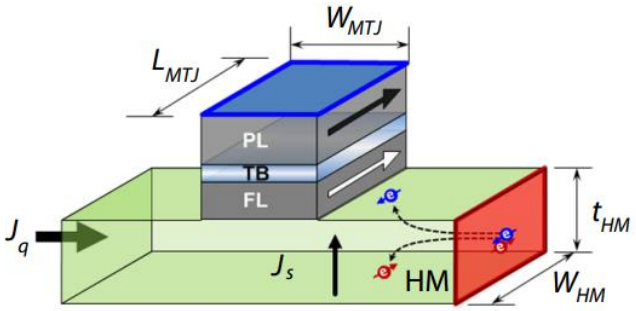
Sengupta, Roy, Scientific Reports'16

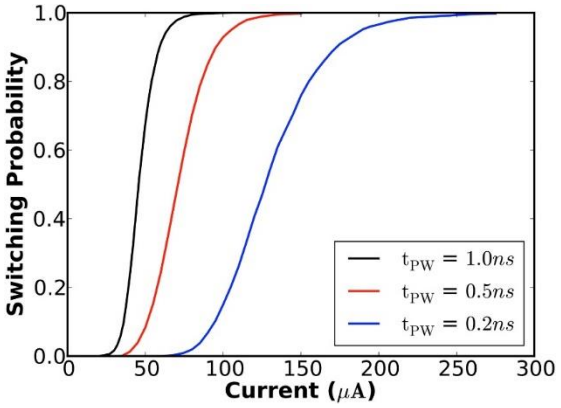
Interconnected Crossbars for NNs

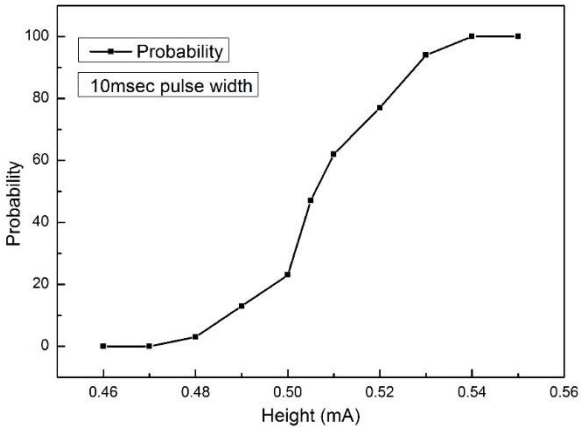


SHE Based MTJ: Stochastic Neuron & Synapse

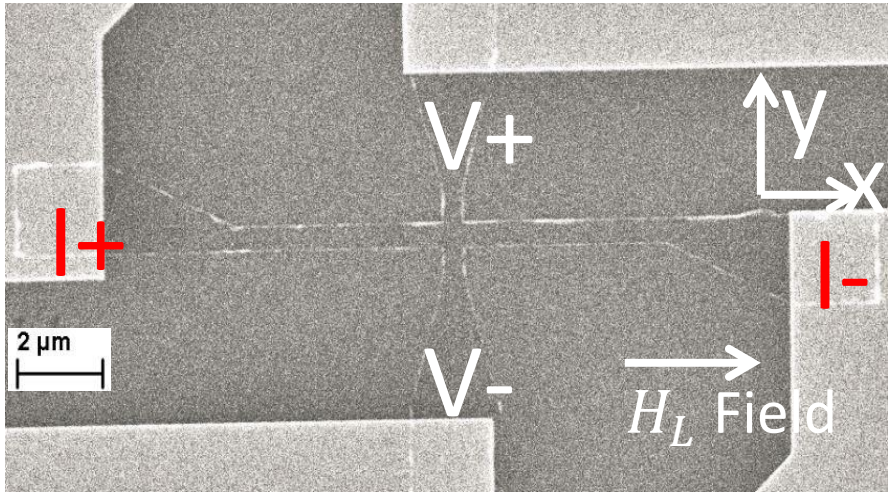
- Exploit the stochastic switching behavior of a mono-domain MTJ in the presence of thermal noise – sigmoidal function
- Stochastic sigmoidal neuron
- Replaces multi-bit synapses with a stochastic single bit (binary) synapse.

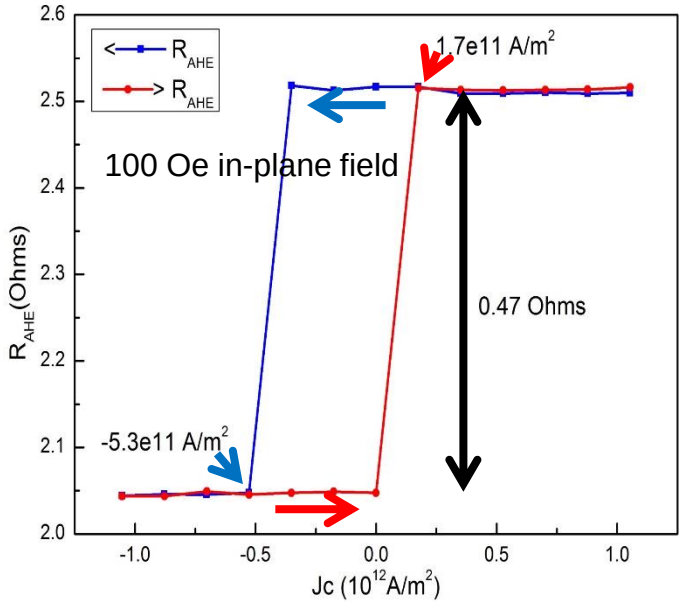




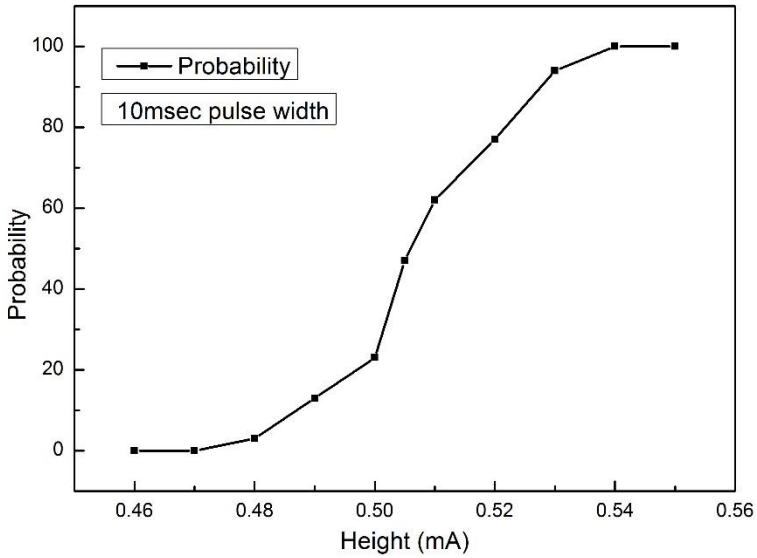


SHE-Based Switching: Experiments



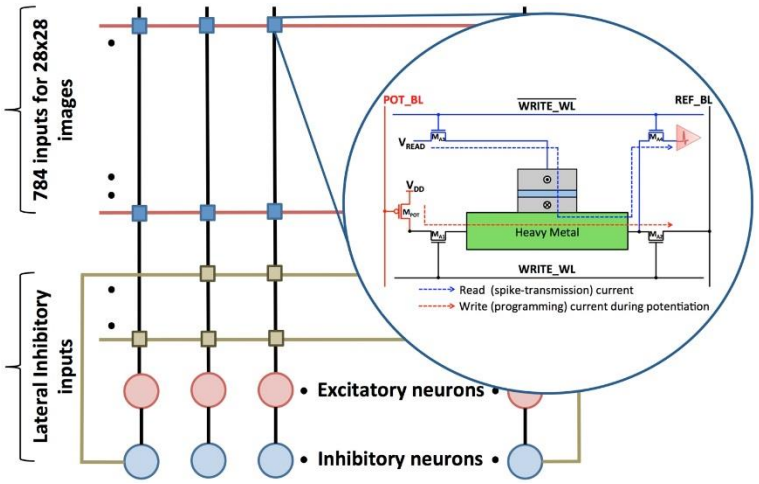


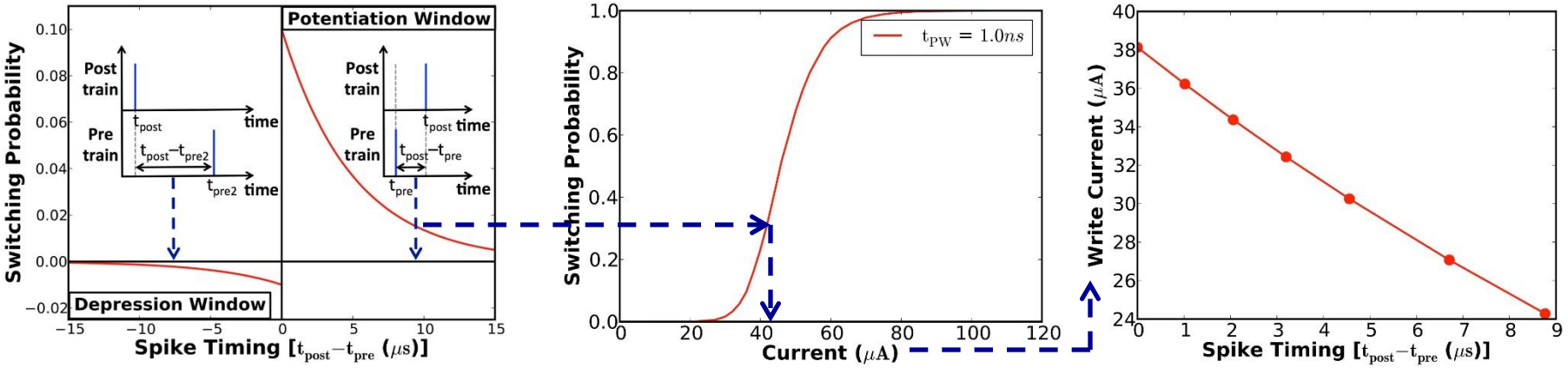
- Anomalous Hall effect (AHE) in FM layer: Hall resistances change abruptly when the magnetization switches in FM.
- Ta(5nm)/CoFeB(1.3nm)/MgO(1.5nm)/Ta(5 nm) (bottom to top) Hall bar structure; CoFeB shows PMA.
- In-plane field H_L applied
- A current pulse group with changeable magnitude is sent through I+/-. Then resistance change is measured across V+/-



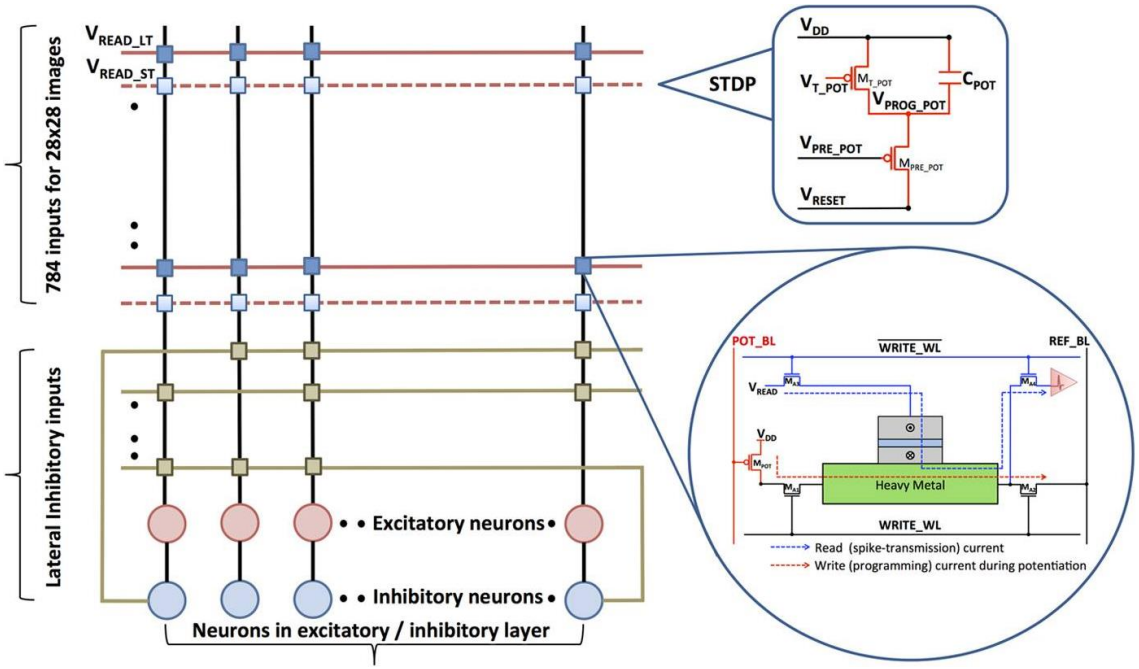
Stochastic Binary Synapse

- Synaptic strength is updated based on the temporal correlation between pre- and post-spike trains
- Synaptic learning is embedded in the switching probability of binary synapses
- Switch the MTJ based on spike timing by passing the required write current



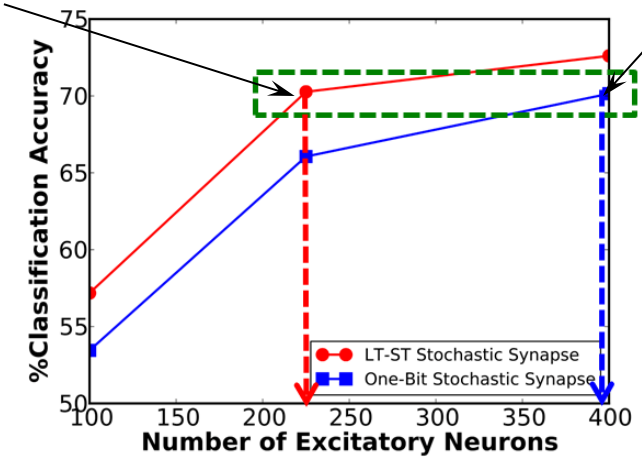


Spiking Neuromorphic Architecture with LT-ST Synapses



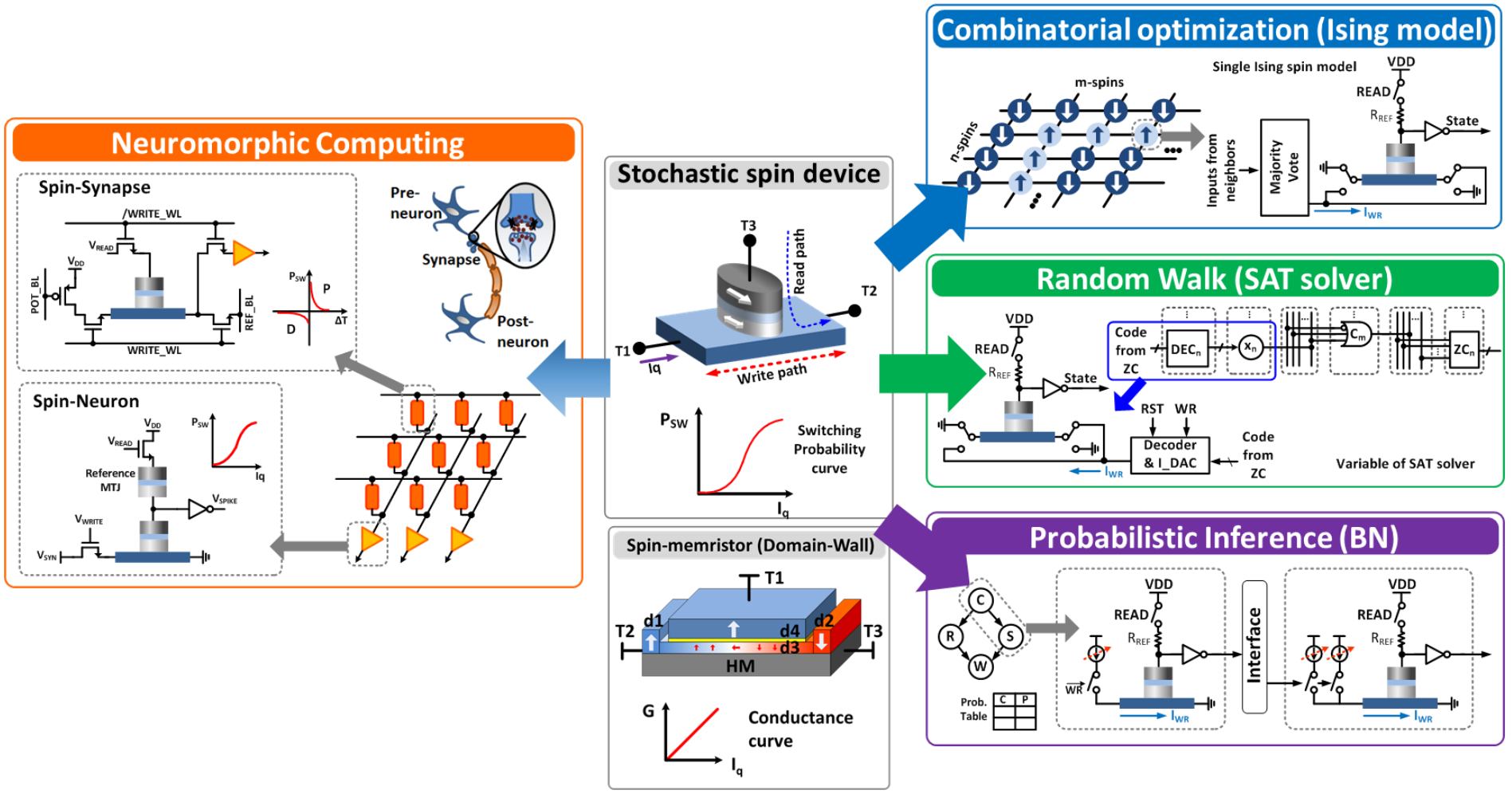
Prog. Energy = 10uJ

Prog. Energy = 23uJ

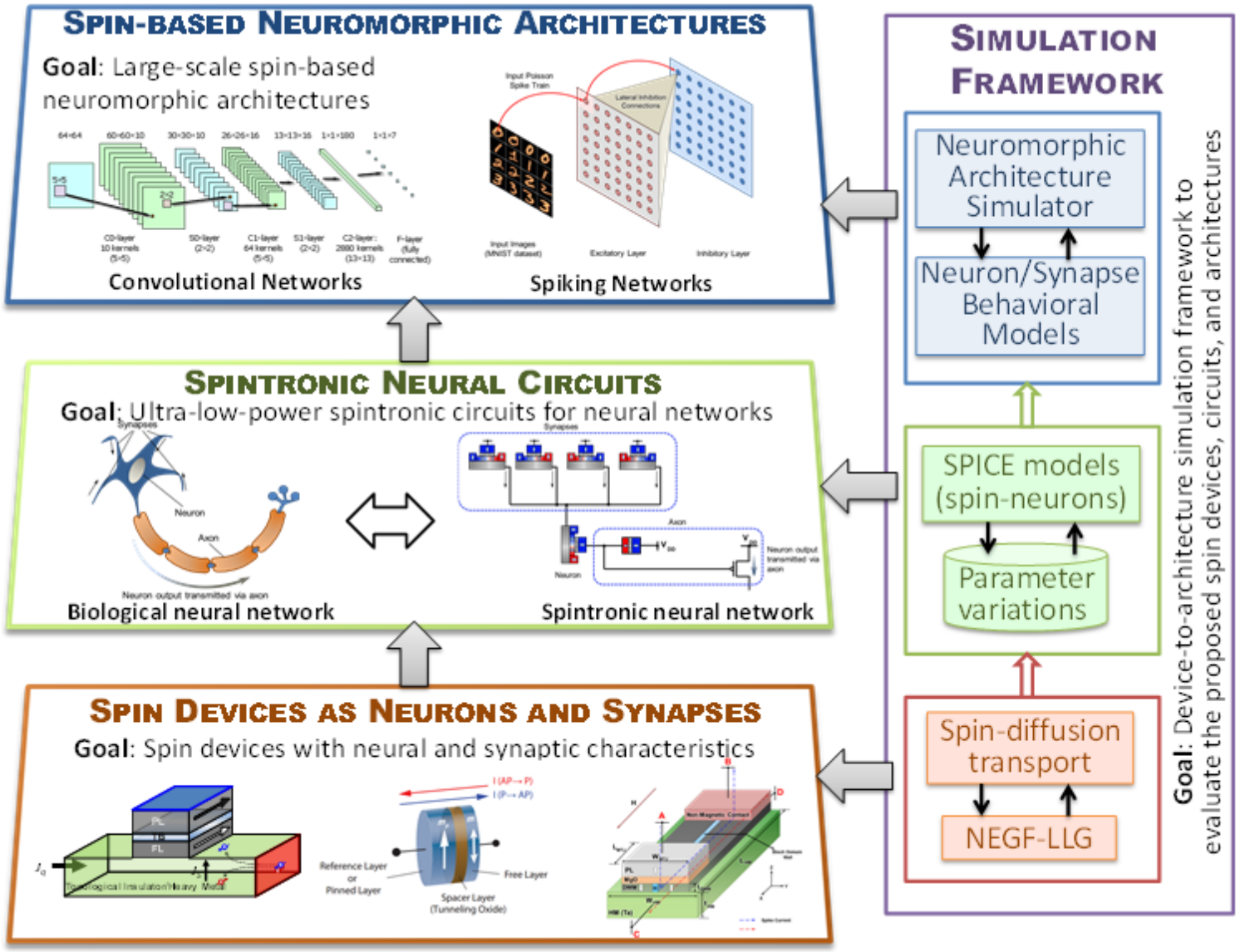


- Crossbar arrangement of the LT-ST synapses with CMOS neurons
- The significant LT synapses are driven by a higher read voltage
- Network of LT-ST synapses provides 5% improvement in the classification accuracy over one-bit synapses
- Under iso-accuracy conditions, the LT-ST synapses offer a 2X reduction in the synaptic write energy

From Devices to Circuits and Systems



Simulation Framework



Conclusions

- Other than memory, STT devices also show promise for a class of computing models such as “brain-inspired computing”, stochastic computing

- STT-devices as “neurons” and “synapses” for both ANN, SNN show the possibility of large improvement in energy compared to CMOS implementation