Efficient Neural Computing Enabled by Magneto-Metallic Neurons and Synapses

KAUSHIK ROY

ABHRONIL SENGUPTA, KARTHIK YOGENDRA, DELIANG FAN, SYED SARWAR, PRIYA PANDA, GOPAL SRINIVASAN, JASON ALLRED, ZUBAIR AZIM, A. RAGHUNATHAN

ECE, Purdue University

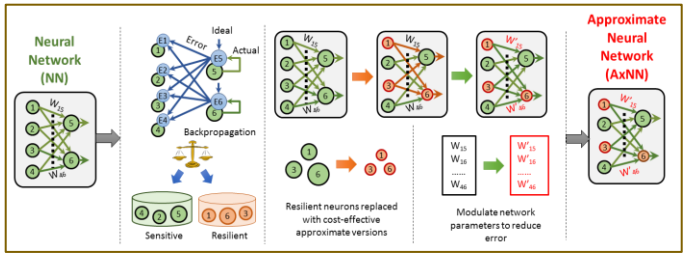
Presented By: Shreyas Sen, ECE, Purdue University

The Computational Efficiency Gap

IBM Watson playing Jeopardy, 2011

IBM Blue Gene supercomputer, equipped with 147456 CPUs and 144TB of memory, consumed 1.4MW of power to simulate 5 secs of brain activity of a cat at 83 times slower firing rates

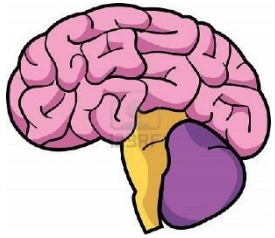
Neuromorphic Computing Technologies



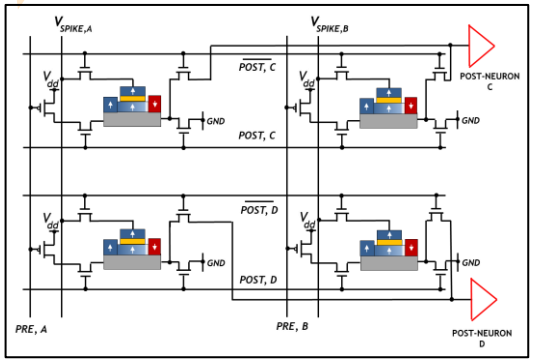
- Approximate Neural Nets, ISLPED ' 14
- Conditional Deep Learning, DATE 2016
- ….

Approximate Computing, Semantic Decomposition, Conditional DLN

Spintronics-Enabled

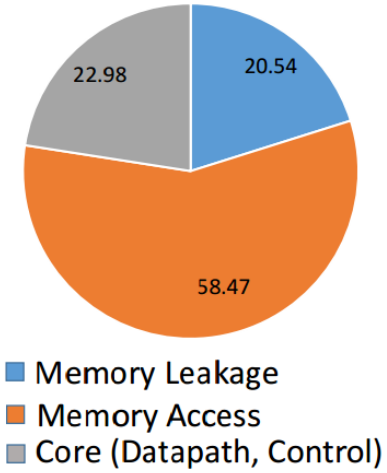


Hardware Accelerators



SW (Multicores/GPUs)

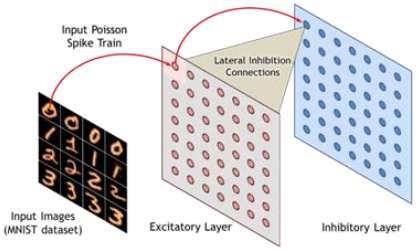
1 uJ/neuron



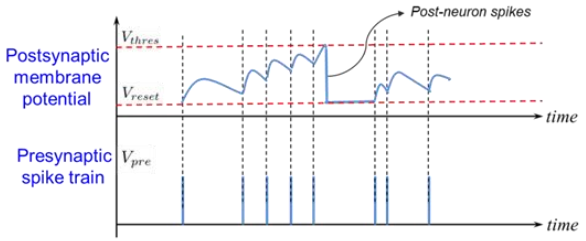
- Spin neuron, IJCNN ' 12, APL'15, TNANO, DAC, DRC, IEDM
- Spintronic Deep Learning Engine, ISLPED ' 14
- Spin synapse, APL ' 15
- ….

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

Top-Down

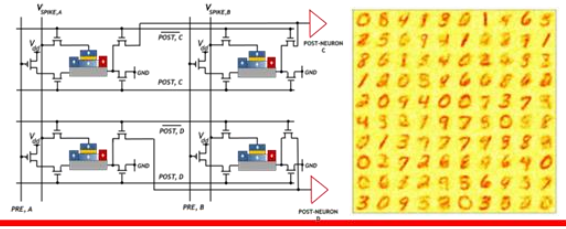


Investigate brain-inspired computing models to provide algorithm-level matching to underlying device physics

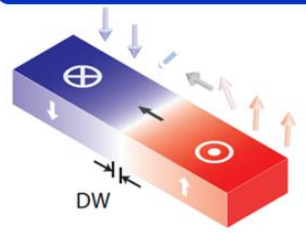


Device-Circuit-Algorithm co-simulation framework used to generate behavioral models for system-level simulations of neuromorphic systems

System Level Solution

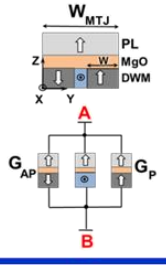


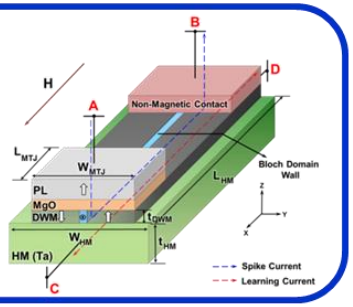
Bottom-Up



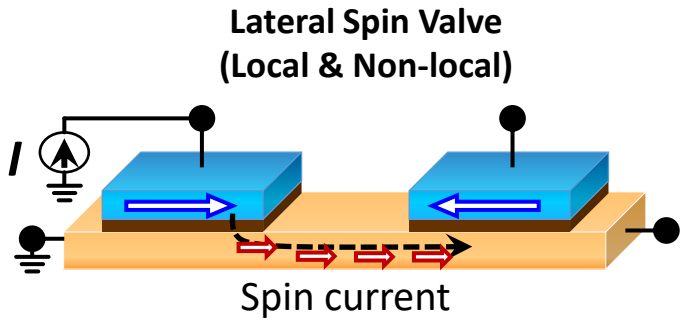
Investigate device physics to mimic "neuron/ synapse" functionalities

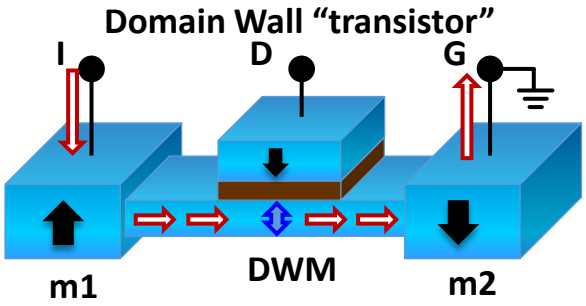
Calibration of device models with experiments

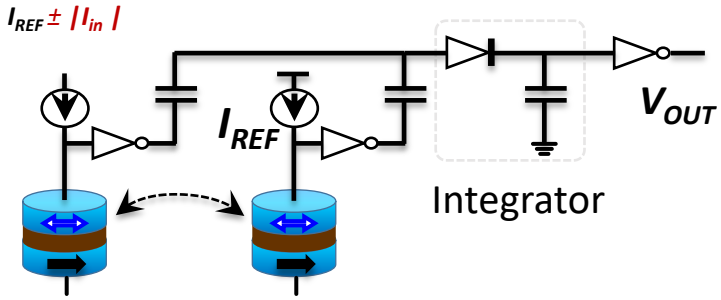


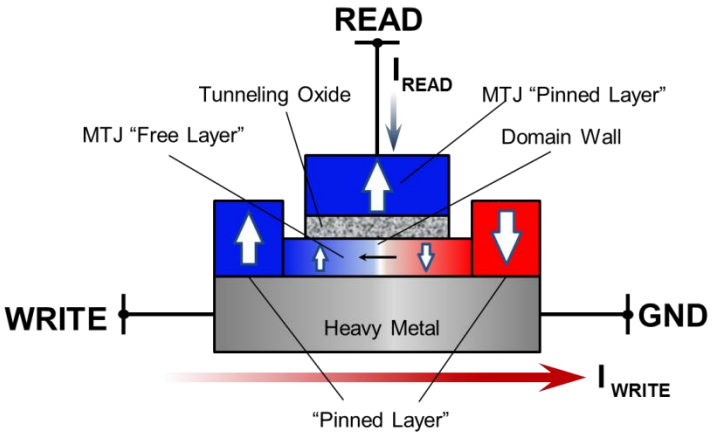


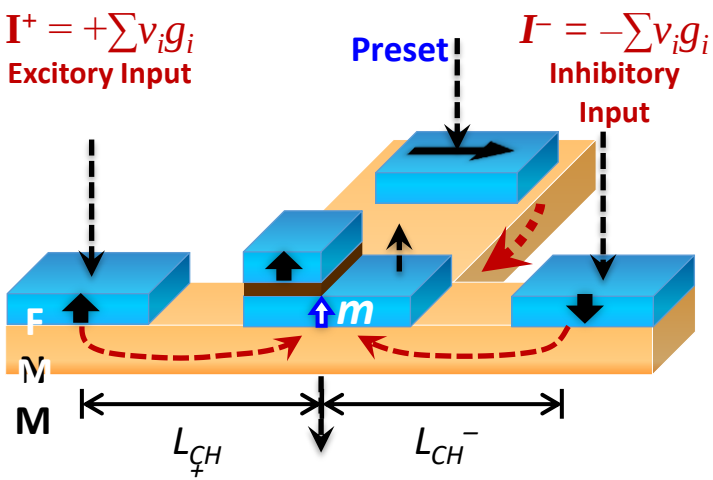
BUILDING PRIMITIVES: MEMORY, NEURONS, SYNAPSES

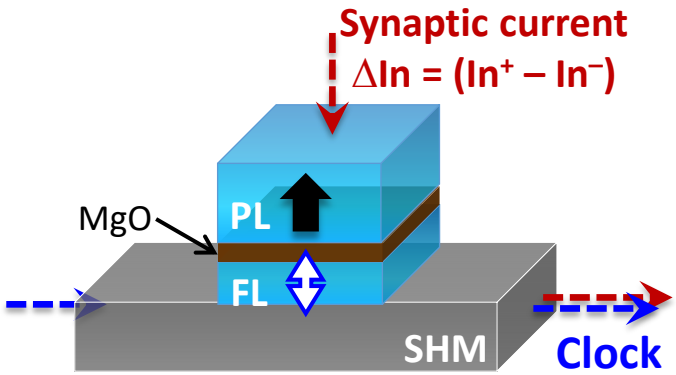




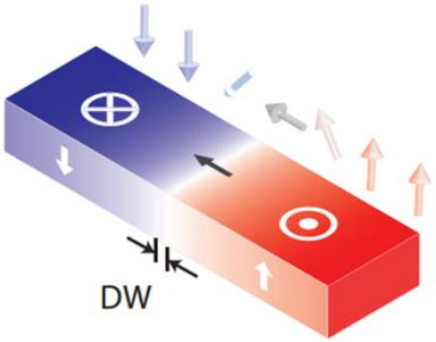


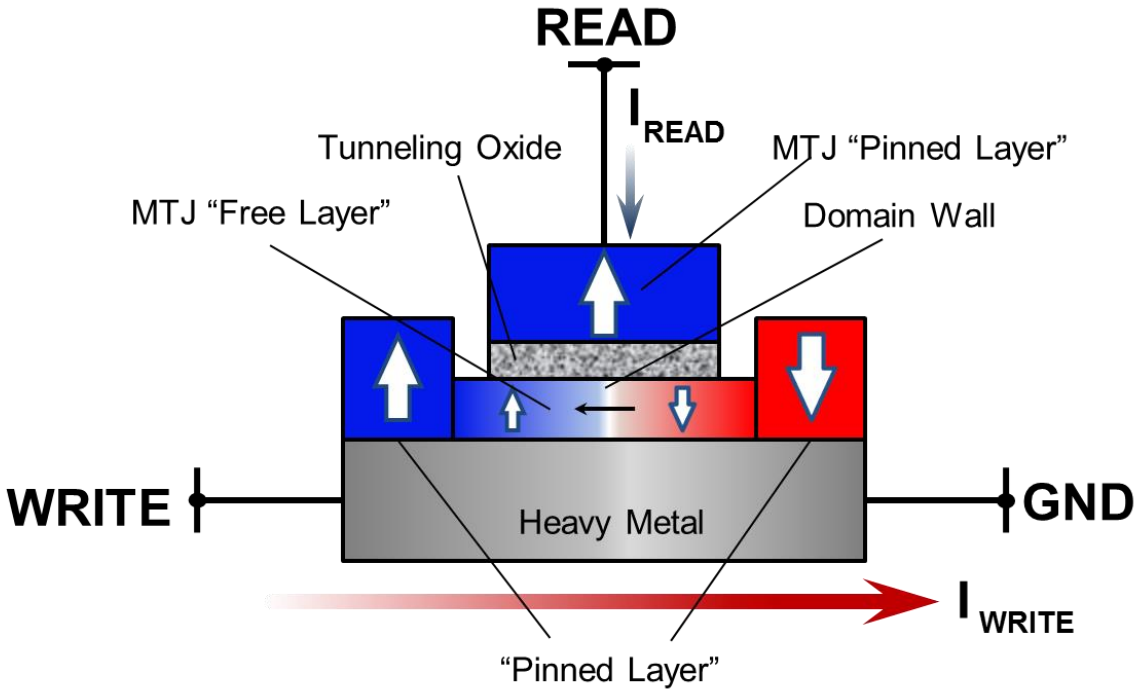






DW-MTJ: Domain Wall Motion/MTJ



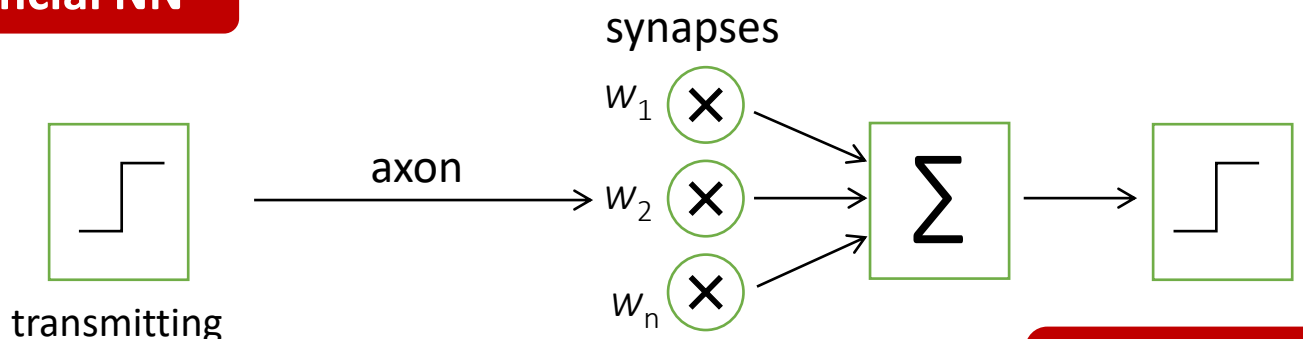


- Three terminal device structure provides decoupled "write" and "read" current paths
- Write current flowing through heavy metal programs domain wall position
- Read current is modulated by device conductance which varies linearly with domain wall position

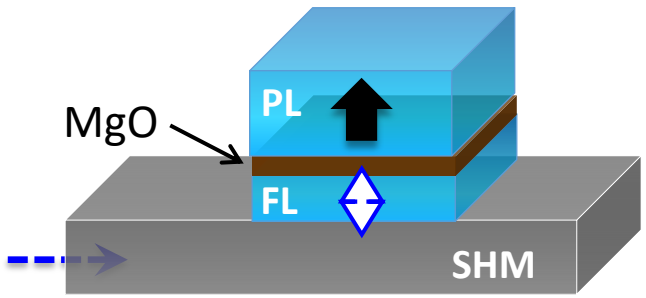
Universal device: Suitable for **memory, neuron, synapse, interconnects**

Simple ANN: Activation

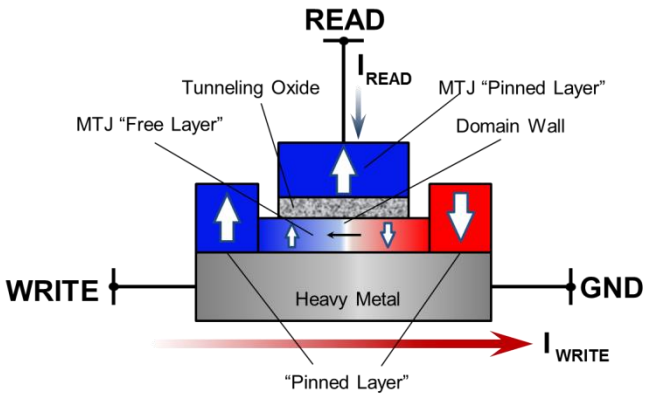
Artificial NN



Spin Hall based Switching

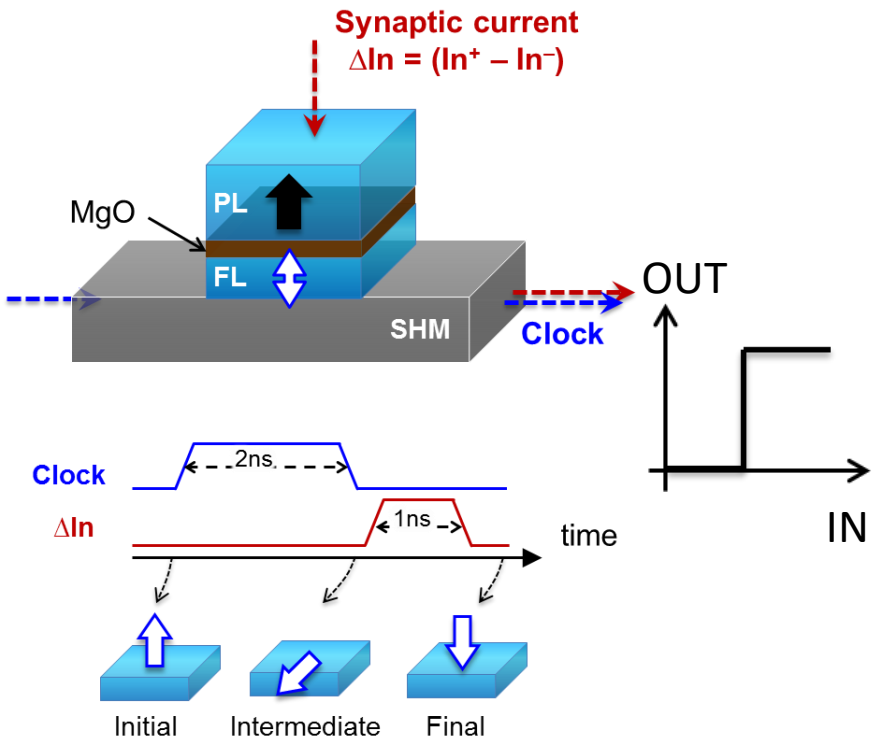


DW-MTJ

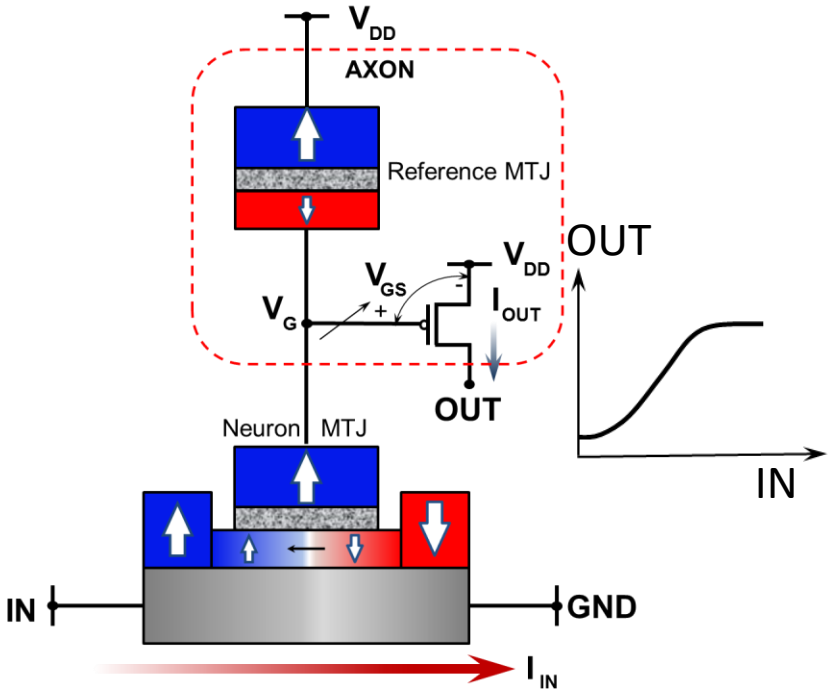


Switch a magnet using spin current, read using TMR effect

Step and Analog ANN Neurons



Step Neuron



Analog Neuron

- Neuron, acting as the computing element, provides an output current (I_{OUT}) which is a function of the input current (I_{IN})
- Axon functionality is implemented by the CMOS transistor
- Note: Stochastic nature of switching of MTJ can be used in Stochastic Neural nets

Benchmarking with CMOS Implementation

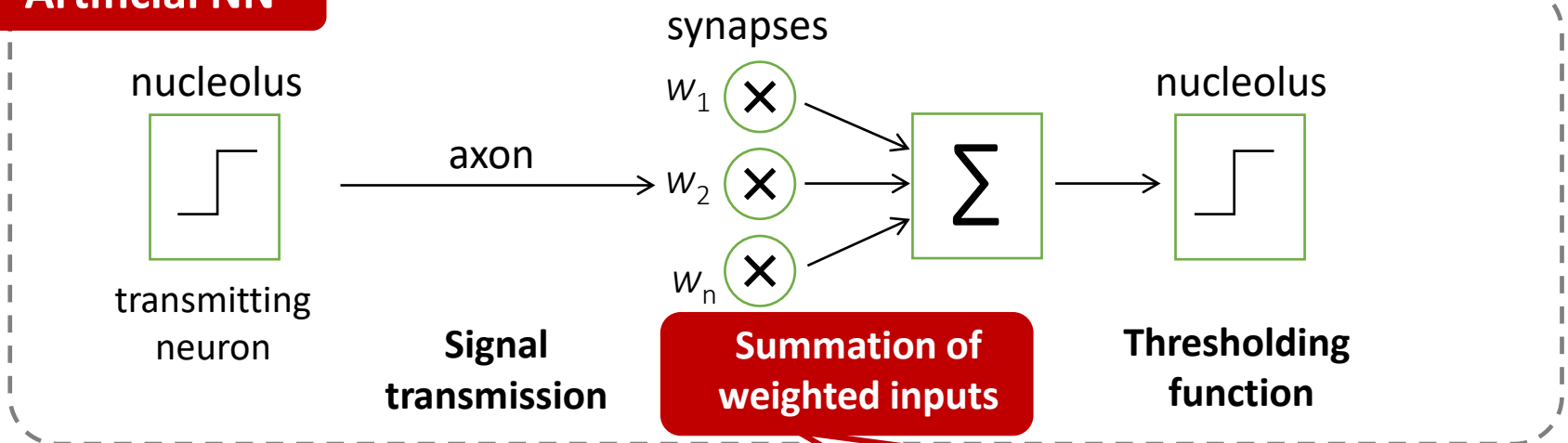
Neurons	Power	Speed	Energy	Function	technology
CMOS Analog neuron 1 [1]	~12μW (assume 1V supply)	65ns	780fJ	Sigmoid	/
CMOS Analog neuron 2 [2]	15μW	/	/	Sigmoid	180nm
CMOS Analog neuron 3 [5]	70μW	10ns	700fJ	Step	45nm
Digital Neuron [3]	83.62μW	10ns	832.6fJ	5-bit tanh	45nm
Hard-Limiting Spin-Neuron	0.81μW	1ns	0.81fJ	Step	/
Soft-Limiting Spin-Neuron	1.25μW	3ns	3.75fJ	Rational/ Hyperbolic	/

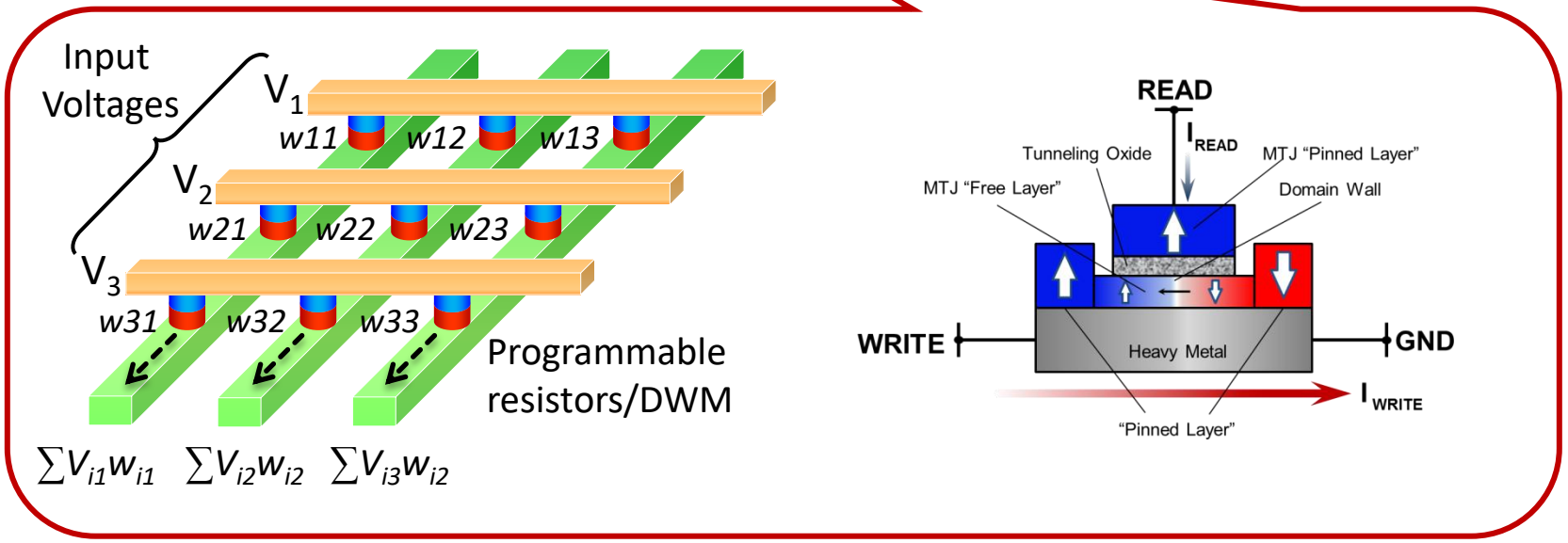
Compared with analog/ digital CMOS based neuron design, spin based neuron designs have the potential to achieve more than two orders lower energy consumption

[1]: A. J. Annema, “Hardware realisation of a neuron transfer function and its derivative”, Electronics Letters, 1994
[2]: M. T. Abuelma’ati, etc, “A reconfigurable satlin/sigmoid/gaussian/triangular basis functions”, APCCAS, 2006
[3]: S. Ramasubramanian, et al., "SPINDLE: SPINtronic Deep Learning Engine for large-scale neuromorphic computing", ISLPED, 2014
[4]: D. Coue, etc “A four-quadrant subthreshold mode multiplier for analog neural network applications”, TNN, 1996
[5]: M. Sharad, etc, “Spin-neurons: A possible path to energy-efficient neuromorphic computers”, JAP, 2013

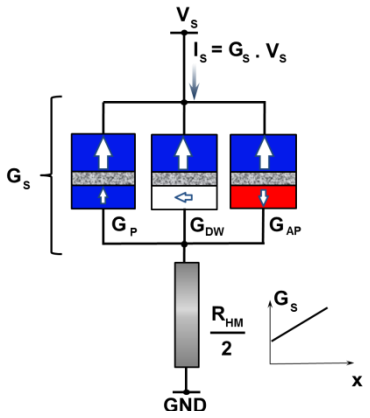
In-Memory Computing (Dot Product)

Artificial NN

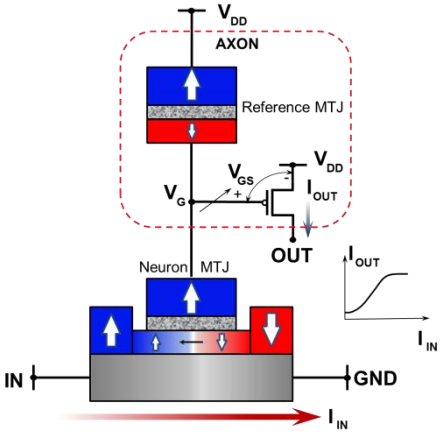




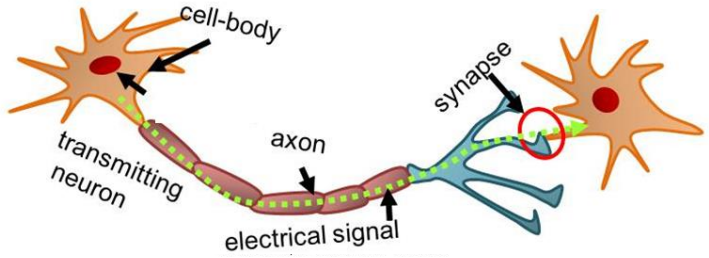
All-Spin Artificial Neural Network



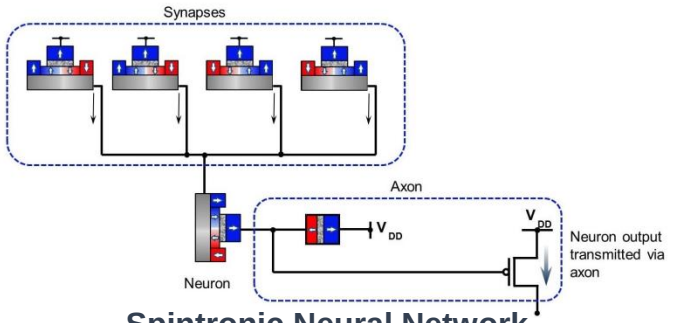
Spin-synapse



Spin-neuron

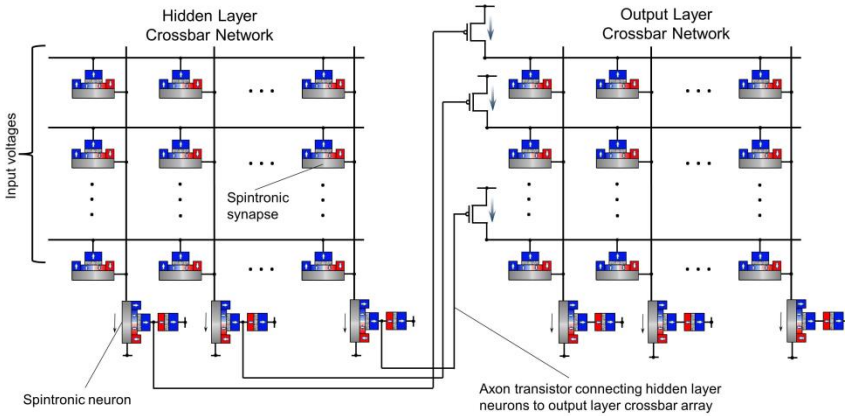


Biological Neural Network



Spintronic Neural Network

- All-spin ANN where spintronic devices directly mimic neuron and synapse functionalities and axon (CMOS transistor) transmits the neuron's output to the next stage
- Ultra-low voltage (~100mV) operation of spintronic synaptic crossbar array made possible by magneto-metallic spin-neurons
- **System level simulations for character recognition shows maximum energy consumption of 0.32fJ per neuron which is ~100x lower in comparison to analog and digital CMOS neurons (45nm technology)**

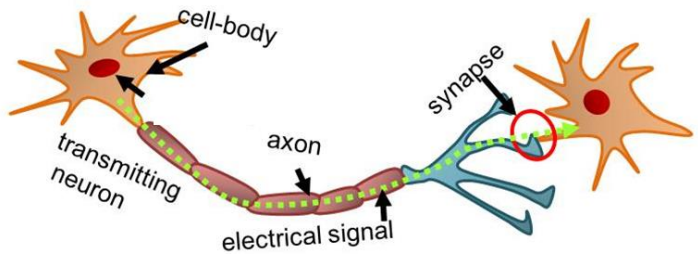


All-spin Neuromorphic Architecture

Spiking Neural Networks (Self-Learning)

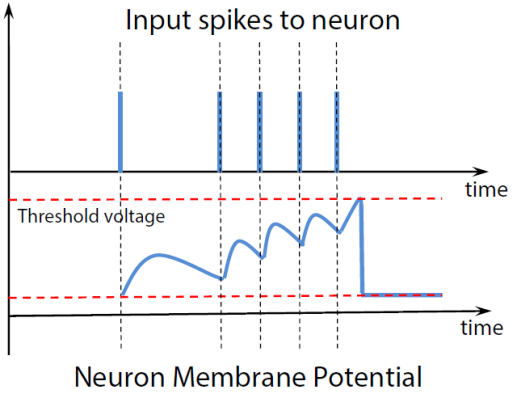
Spiking Neuron Membrane Potential

Biological Spiking Neuron

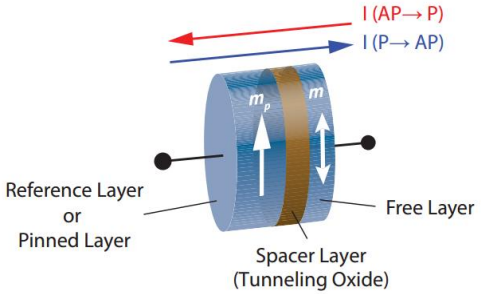


LIF Equation:

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

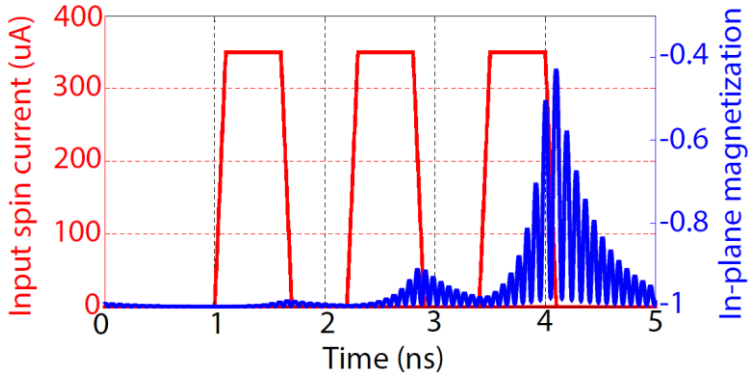


MTJ Spiking Neuron



LLGS Equation:

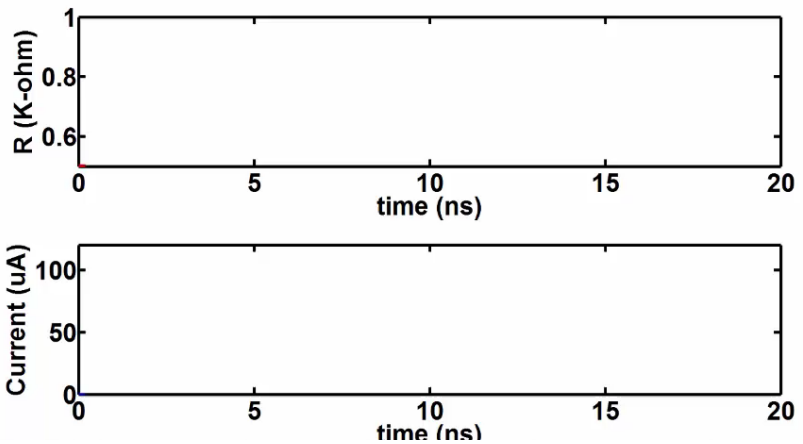
$$\frac{d\hat{\mathbf{m}}}{dt} = -\gamma(\hat{\mathbf{m}} \times \mathbf{H}_{eff}) + \alpha(\hat{\mathbf{m}} \times \frac{d\hat{\mathbf{m}}}{dt}) + \frac{1}{qN_s}(\hat{\mathbf{m}} \times \mathbf{I}_s \times \hat{\mathbf{m}})$$



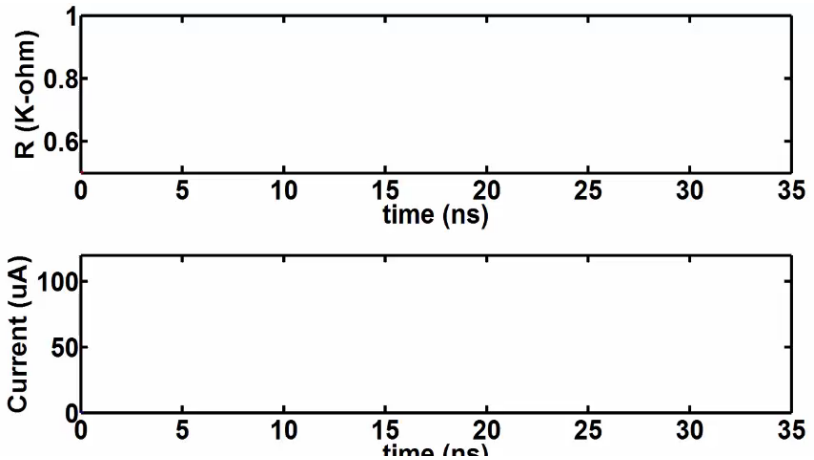
The leaky fire and integrate can be approximated by an MTJ – the magnetization dynamics mimics the leaky fire and integrate operation

MTJ as a Spiking Neuron

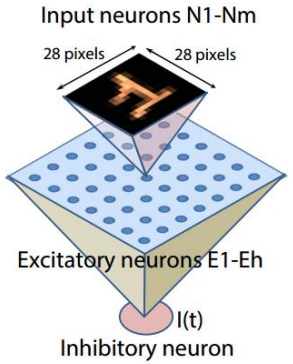
Spikes at 3ns interval

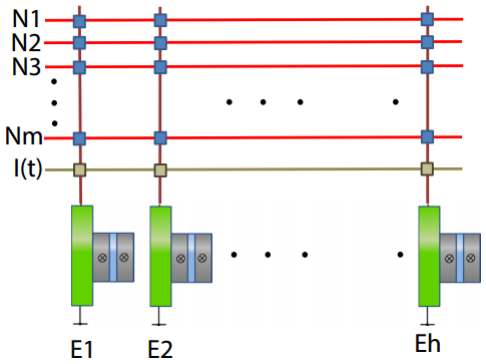


Spikes at 6ns interval



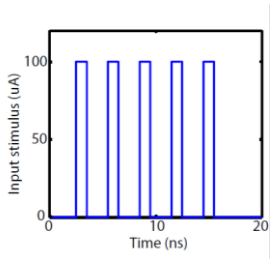
- **MTJ magnetization leaks and integrates input spikes (LLG equation) in presence of thermal noise**
- Associated "write" and "read" energy consumption is ~ 1fJ and ~1.6fJ per time-step which is much lower than state-of-the-art CMOS spiking neuron designs (267pJ [1] and 41.3pJ [2] per spike)



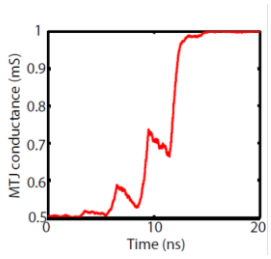


Spiking Neurons

LLGS Based Spiking Neuron

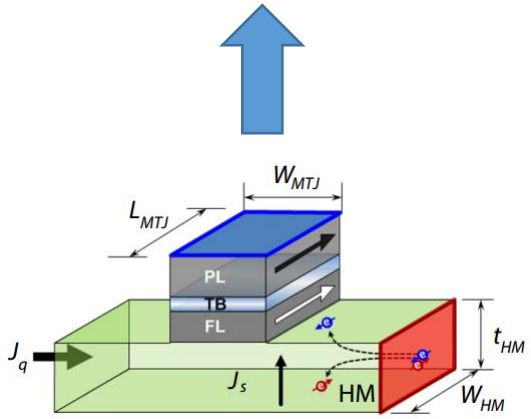


Input Spikes

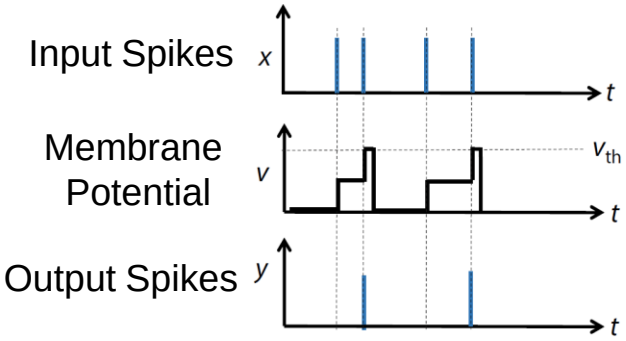


MTJ conductance

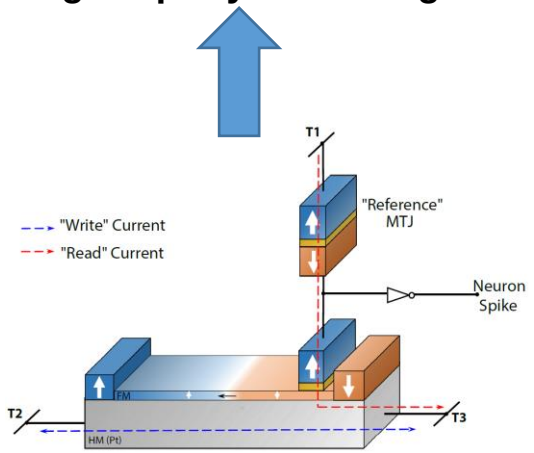
LLG Equation Mimicking Spiking Neurons



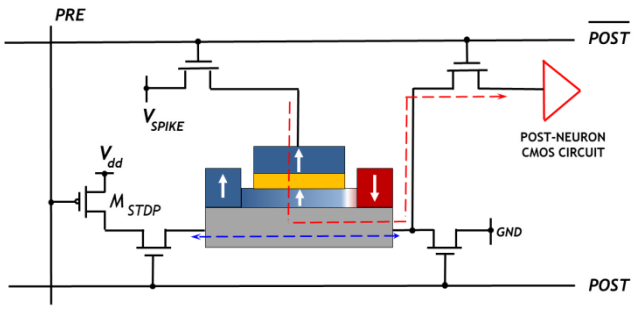
DW-MTJ base IF Neurons

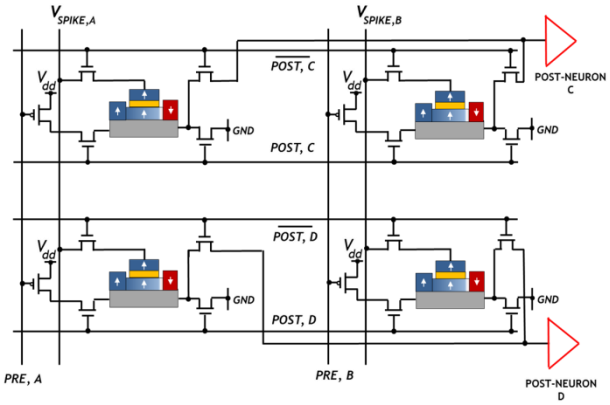


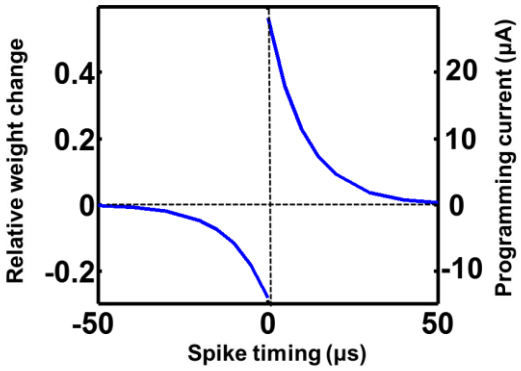
DW Integrating Property Mimicking IF Neuron



Arrangement of DW-MTJ Synapses in Array for STDP Learning







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 SRAM based synapses /emerging devices like PCM**

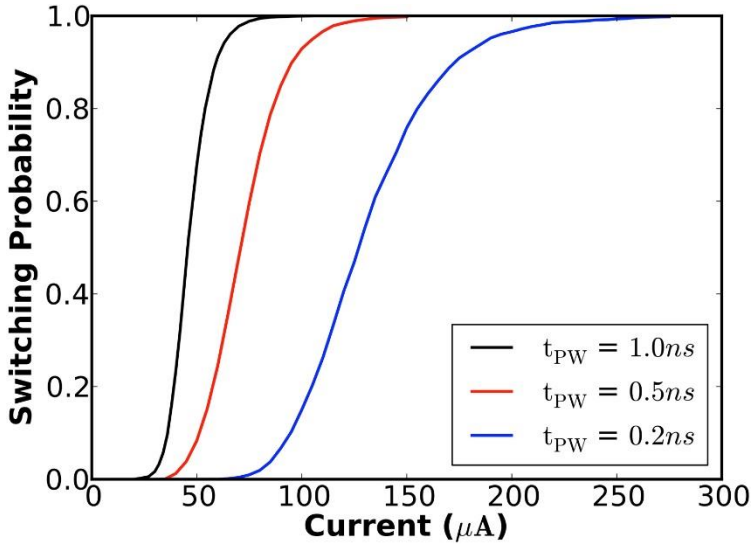
Comparison with Other Synapses

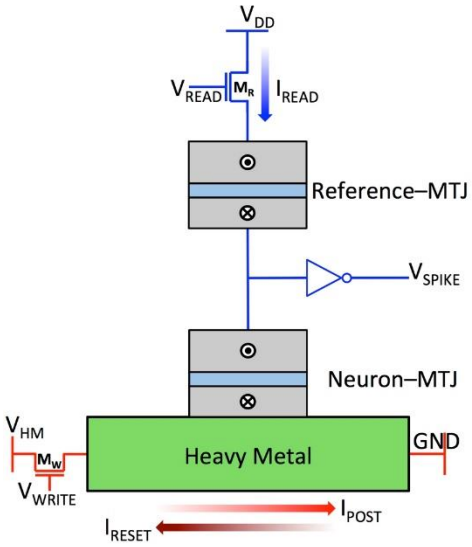
Device	Reference	Dimension	Prog. Energy	Prog. Time	Terminals	Prog. Mechanism
GeSbTe memristor	D. Modha ACM JETCAS, 2013 (IBM)	40nm mushroom and 10nm pore	Average 2.74 pJ/ event	~60ns	2	Programmed by Joule heating (Phase change)
GeSbTe memristor	H.-S. P. Wong Nano Letters, 2012 (Stanford)	75nm electrode diameter	50pJ (reset) 0.675pJ (set)	10ns	2	Programmed by Joule heating (Phase change)
Ag-Si memristor	Wei Lu Nano Letters, 2010 (U Michigan)	100nmx100nm	Threshold voltage~2.2V	~300µs	2	Movement of Ag ions
FeFET	Y. Nishitani JJAP, 2013 (Panasonic, Japan)	Channel Length-3µm	Maximum gate voltage – 4V	10µs	3	Gate voltage modulation of ferroelectric polarization
Floating gate transistor	P. Hasler IEEE TBIOCAS, 2011 (GaTech)	1.8µm/0.6µm (0.35µm CMOS technology)	Vdd - 4.2V Tunneling Voltage – 15V	100µs (injection) 2ms (tunneling)	3	Injection and tunneling currents
SRAM synapse	B. Rajendran IEEE TED, 2013 (IIT Bombay)	0.3µm^2 (10nm CMOS technology)	Average 328fJ for 4-bit synapse	-	-	Digital counter based circuits
Spintronic synapse	**NRL Purdue**	**340nmx20nm**	**Maximum 48fJ /event**	**1ns**	**3**	**Spin-orbit torque**

MTJ Enabled All-Spin Spiking Neural Network

Probabilistic Spiking Neuron

- A pre-neuronal spike modulated by synapse to generate current that controls the post-neuronal spiking probability.
- Exploit stochastic switching behavior of MTJ in presence of thermal noise.

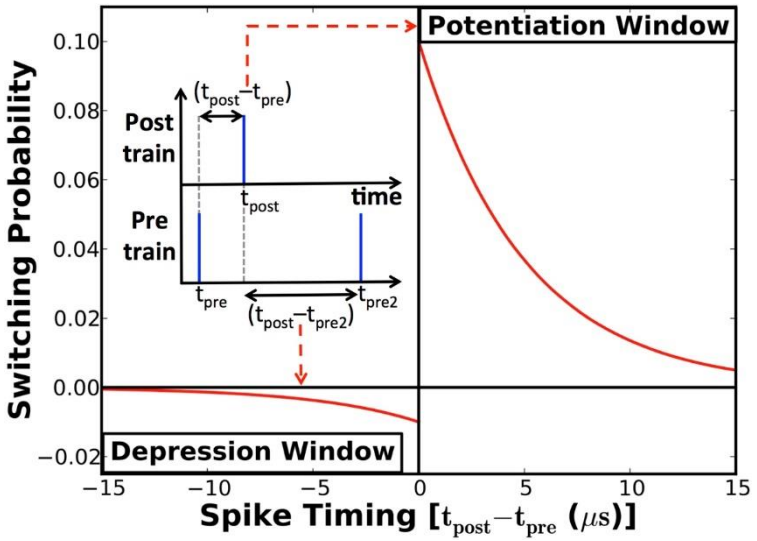


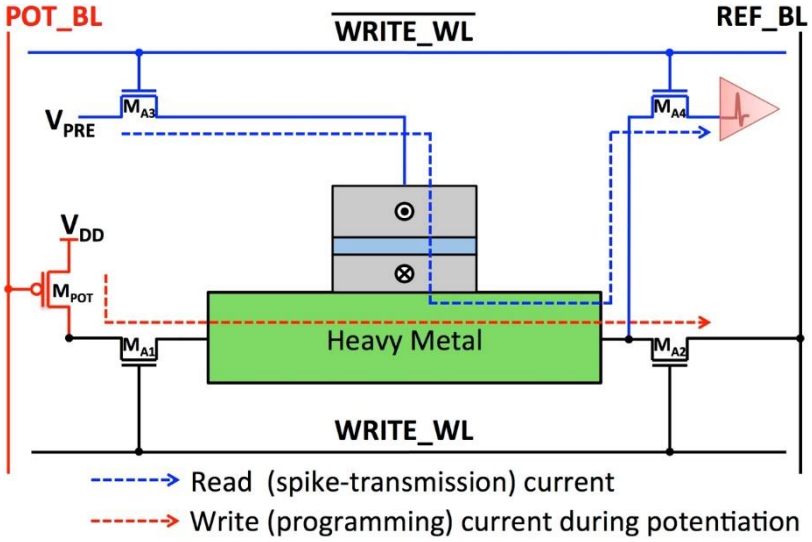


MTJ Enabled All-Spin Spiking Neural Network

Stochastic Binary Synapse

- Synaptic strength proportional to temporal correlation between pre- and post-spike trains.
- Stochastic STDP – Synaptic learning embedded in the switching probability of binary synapses.

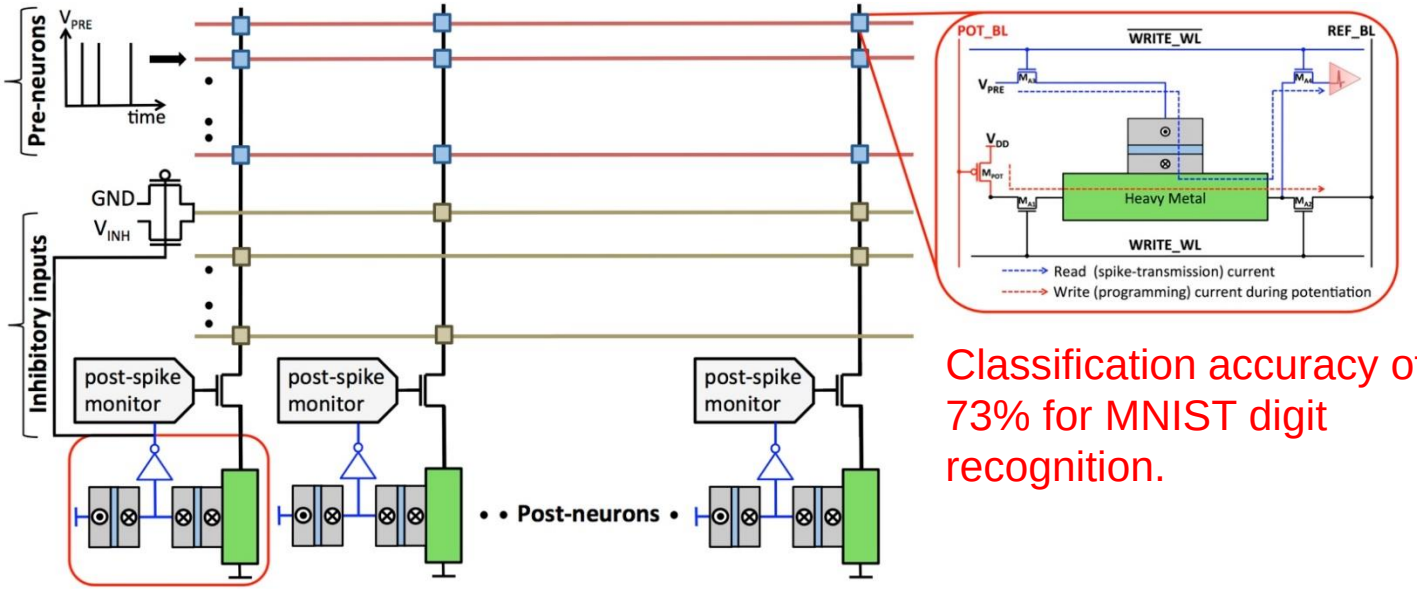




MTJ Enabled All-Spin Spiking Neural Network

Stochastic SNN Hardware Implementation

- Crossbar arrangement of the spin neurons and synapses for energy efficiency.
  - Average neuronal energy of 1fJ and 1.6fJ per timestep for write and read operations, and 4.5fJ for reset.
  - Average synaptic programming energy of 70fJ per training epoch.



Classification accuracy of 73% for MNIST digit recognition.

Summary

- Spintronics do show promise for low-power non-Boolean/brain-inspired computing
  - Need for new leaning techniques suitable for emerging devices
  - Materials research, new physics, new devices, simulation models
- An exciting path ahead…