

Center for Spintronic Materials, Interfaces, and Novel Architectures

Spintronics Enabled Efficient Neuromorphic Computing: Prospects and Perspectives

KAUSHIK ROY

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

Purdue University



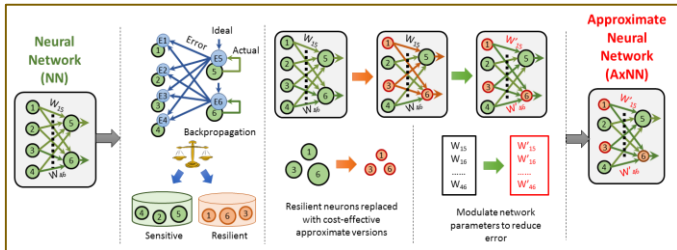
The Computational Efficiency Gap

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

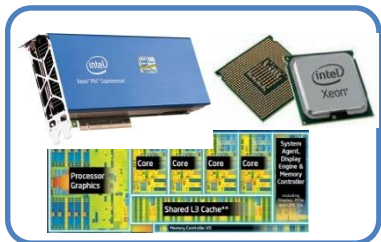
Neuromorphic Computing Technologies



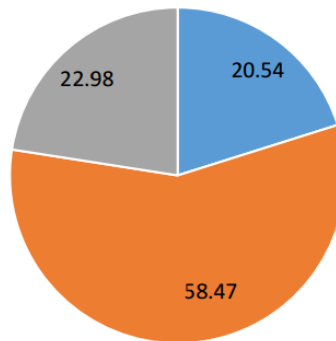
- Approximate Neural Nets, ISLPED '14
- Conditional Deep Learning, DATE 2016
-

Hardware Accelerators

SW (Multicores/GPUs)
1 uJ/neuron

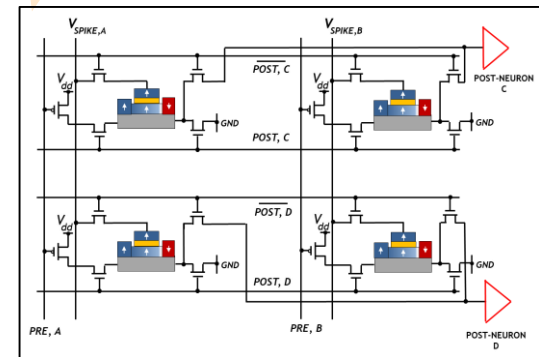


Approximate Computing, Semantic Decomposition, Conditional DLN

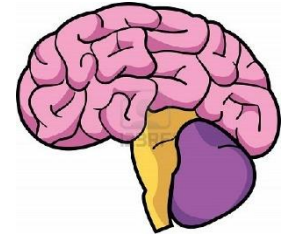


- Memory Leakage
- Memory Access
- Core (Datapath, Control)

Spintronics-Enabled

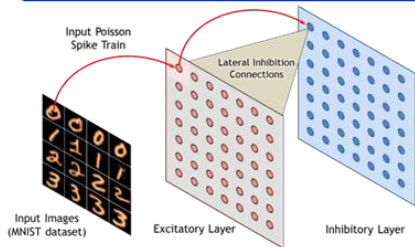


- Spin neuron, IJCNN '12, APL'15, TNANO, DAC, DRC, IEDM
- Spintronic Deep Learning Engine, ISLPED '14
- Spin synapse, APL '15
-

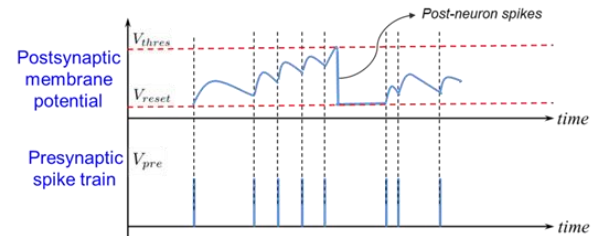


Device/Circuit/Algorithm Co-Design: Spin/ANN/SNN

Top-Down

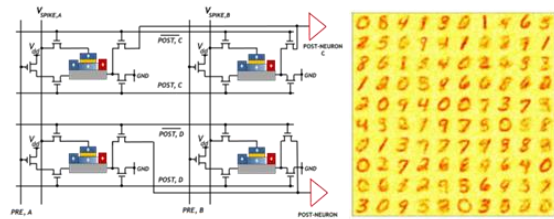


Investigate brain-inspired computing models to provide algorithm-level matching to underlying device physics

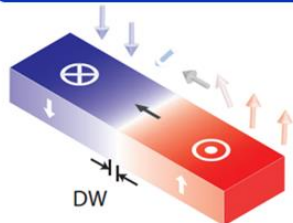


Device-Circuit-Algorithm co-simulation framework used to generate behavioral models for system-level simulations of neuromorphic systems

System Level Solution

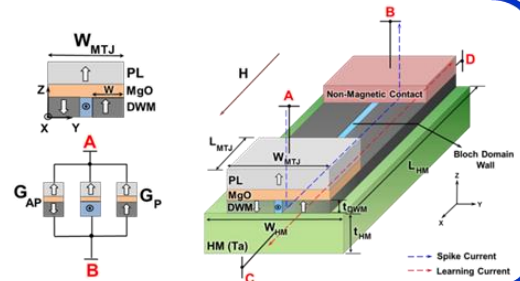


Bottom-Up



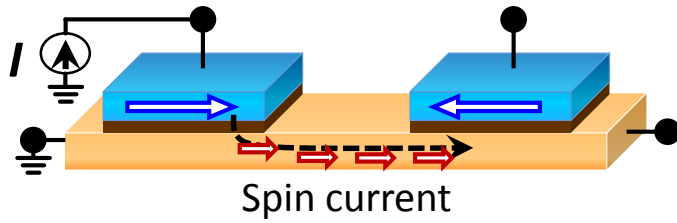
Investigate device physics to mimic "neuron/ synapse" functionalities

Calibration of device models with experiments

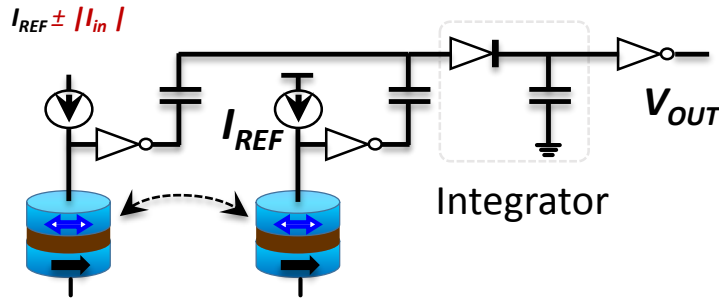
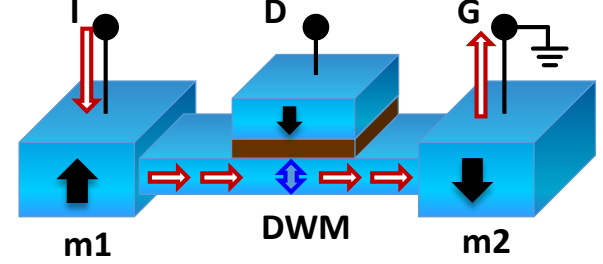


BUILDING PRIMITIVES: MEMORY, NEURONS, SYNAPSES

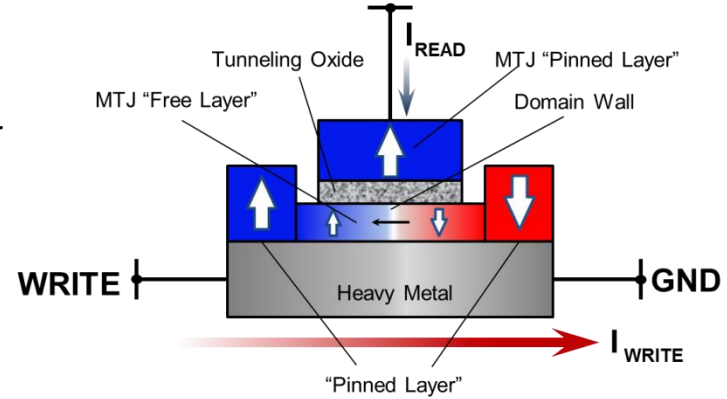
Lateral Spin Valve
(Local & Non-local)



Domain Wall "transistor"



READ



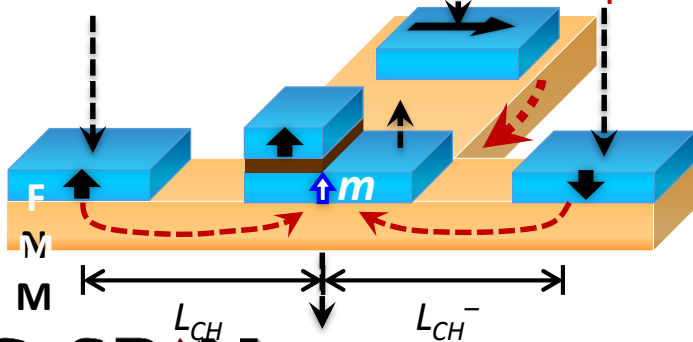
$$I^+ = +\sum v_i g_i$$

Excitatory Input

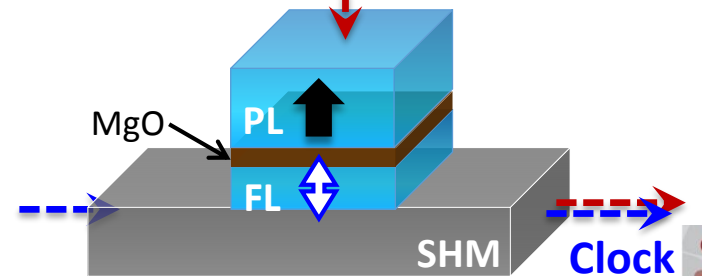
Preset

$$I^- = -\sum v_i g_i$$

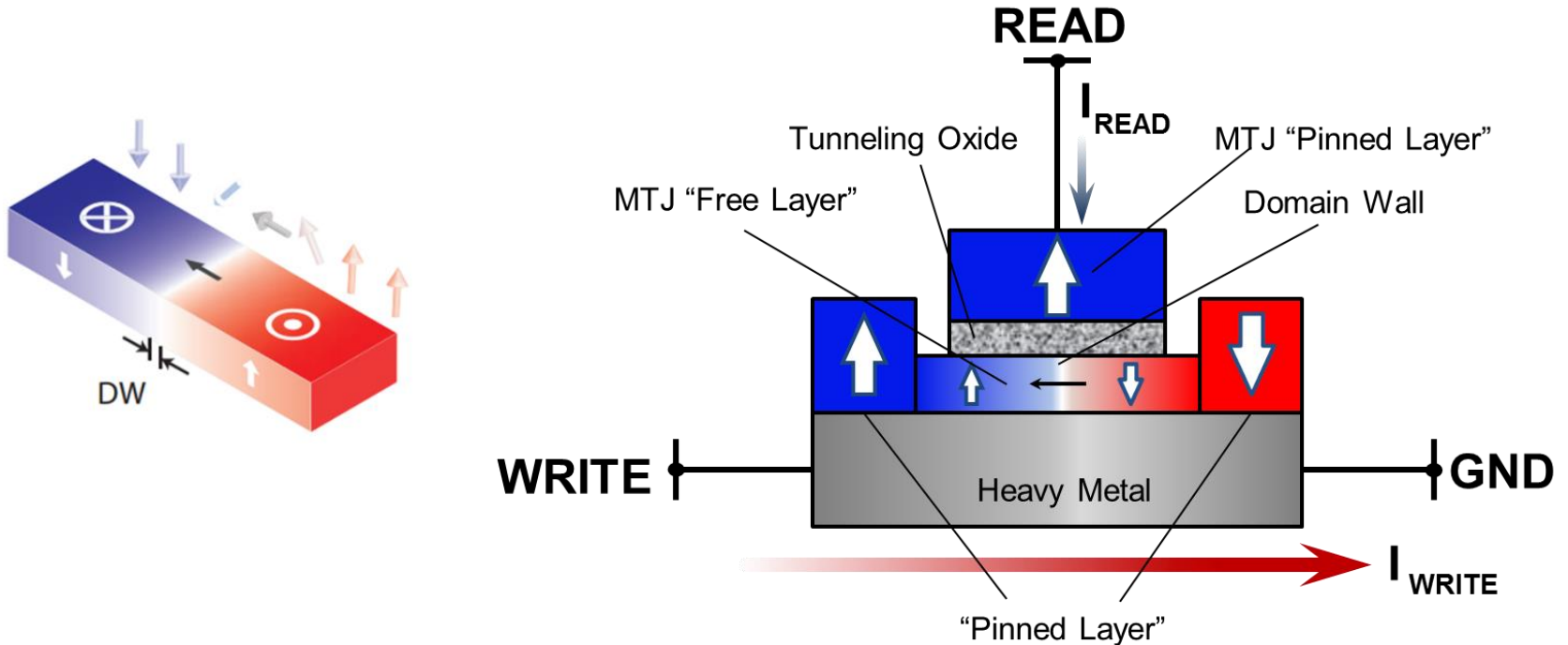
Inhibitory Input



Synaptic current
 $\Delta I_n = (I_n^+ - I_n^-)$



DW-MTJ: Domain Wall Motion/MTJ

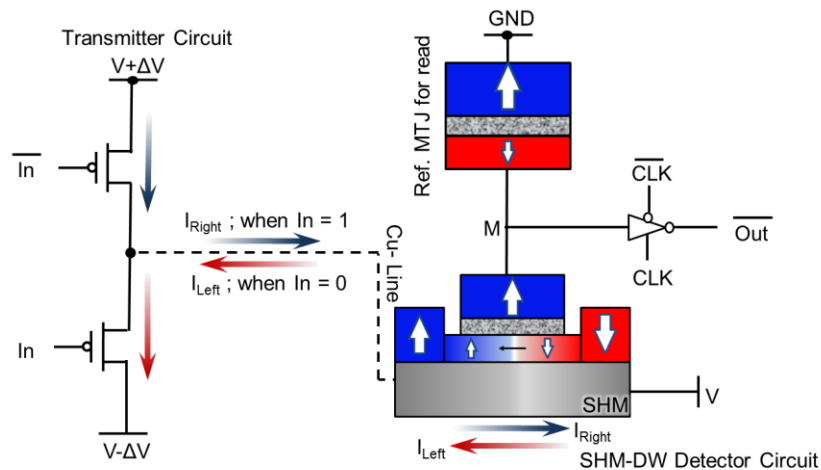


- Three terminal device structure provides decoupled “write” and “read” current paths
- Write current flowing through heavy metal programs domain wall position
- Read current is modulated by device conductance which varies linearly with domain wall position

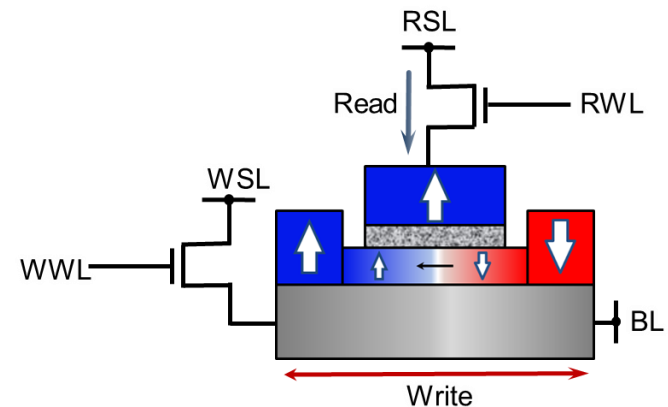
Universal device: Suitable for **memory, neuron, synapse, interconnects**

DW-MTJ for Interconnects/Memory

Interconnect Design



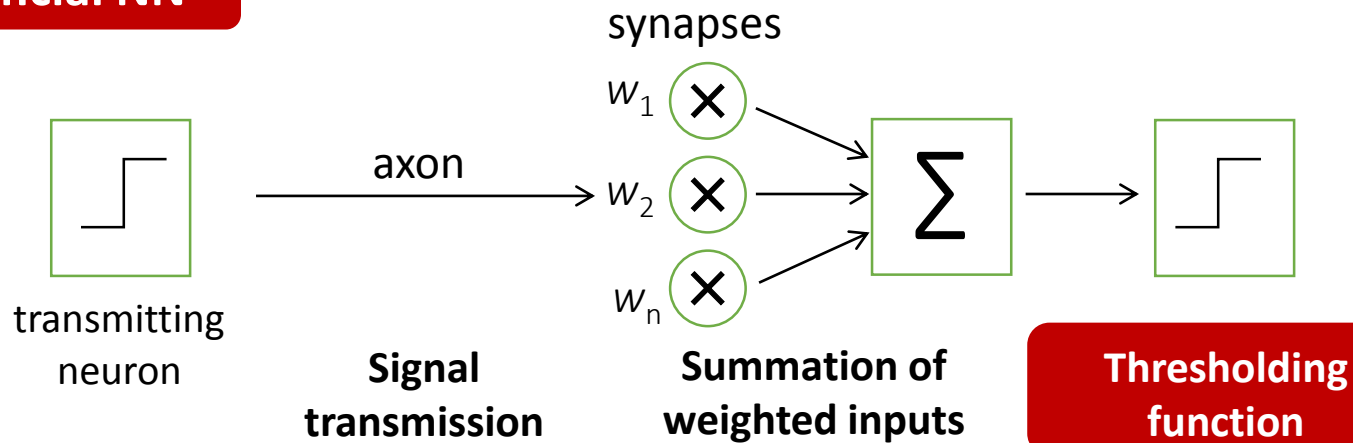
Memory Bit-Cell



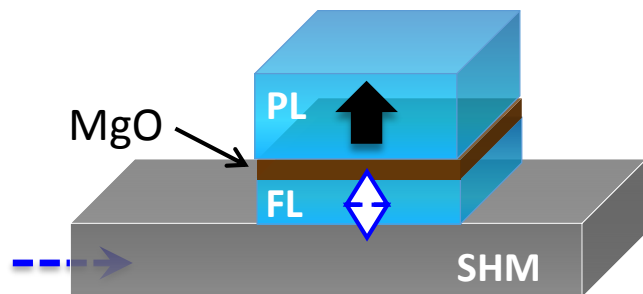
- Energy-efficient interconnect design can circumvent the energy and delay penalties in CMOS based global interconnects for scaled technology nodes
- DW-MTJ memory bit cell with decoupled “write” and “read” current paths

Simple ANN: Activation

Artificial NN

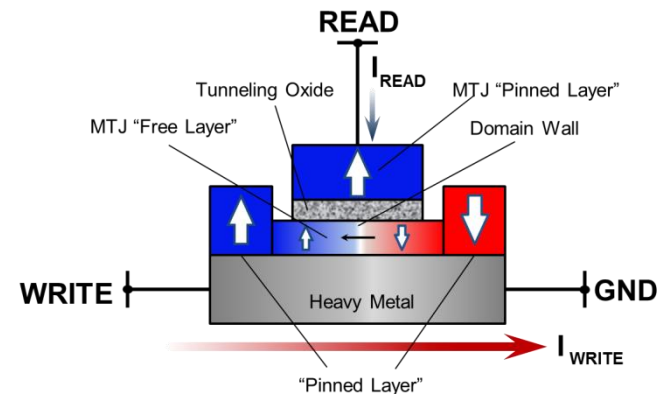


Spin Hall based Switching

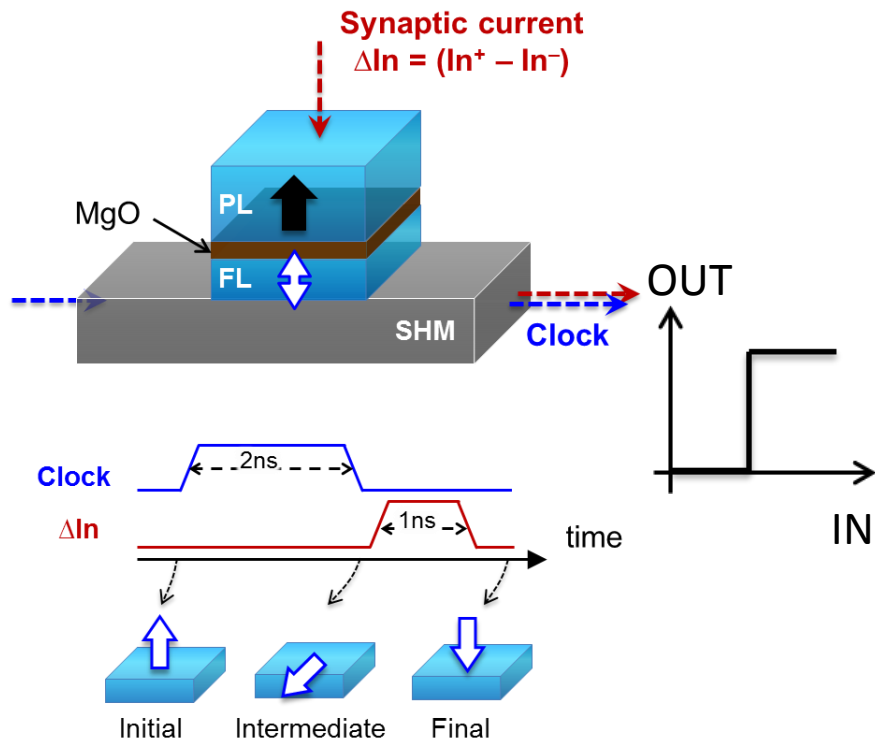


Switch a magnet using spin current, read using TMR effect

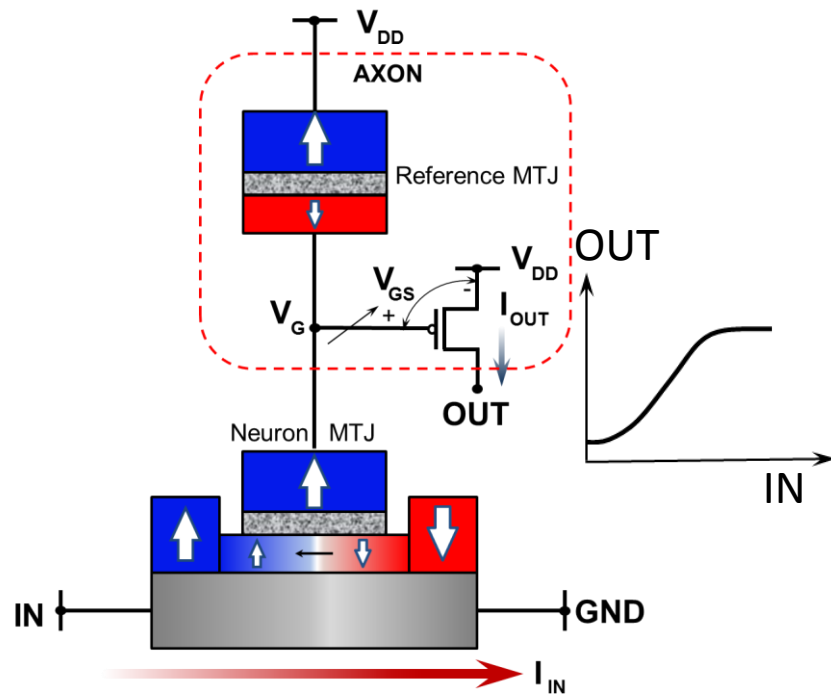
DW-MTJ



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 in Stochastic Neural

Benchmarking with CMOS Implementation

Neurons	Power	Speed	Energy	Function	technology
CMOS Analog neuron 1 [1]	$\sim 12\mu\text{W}$ (assume 1V supply)	65ns	780fJ	Sigmoid	/
CMOS Analog neuron 2 [2]	$15\mu\text{W}$	/	/	Sigmoid	180nm
CMOS Analog neuron 3 [5]	$70\mu\text{W}$	10ns	700fJ	Step	45nm
Digital Neuron [3]	$83.62\mu\text{W}$	10ns	832.6fJ	5-bit tanh	45nm
Hard-Limiting Spin-Neuron	$0.81\mu\text{W}$	1ns	0.81fJ	Step	/
Soft-Limiting Spin-Neuron	$1.25\mu\text{W}$	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

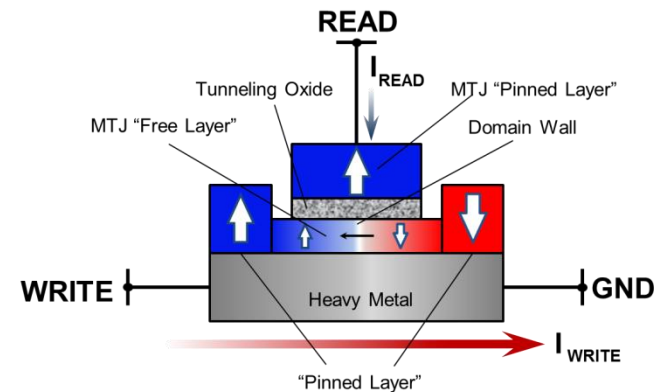
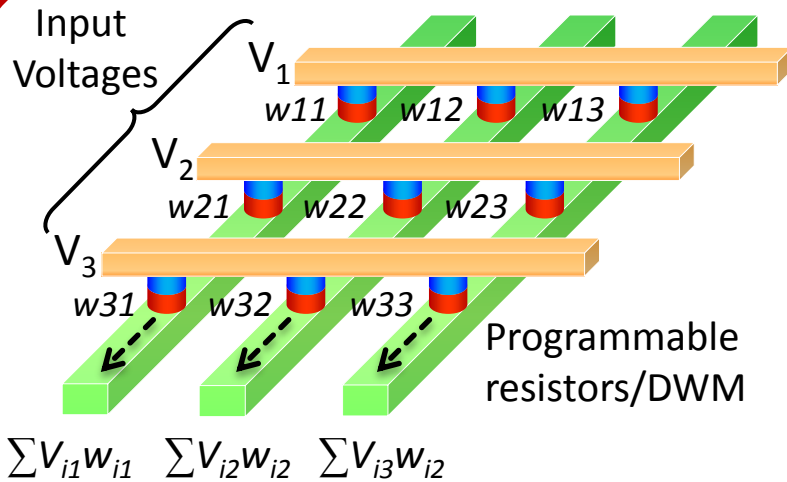
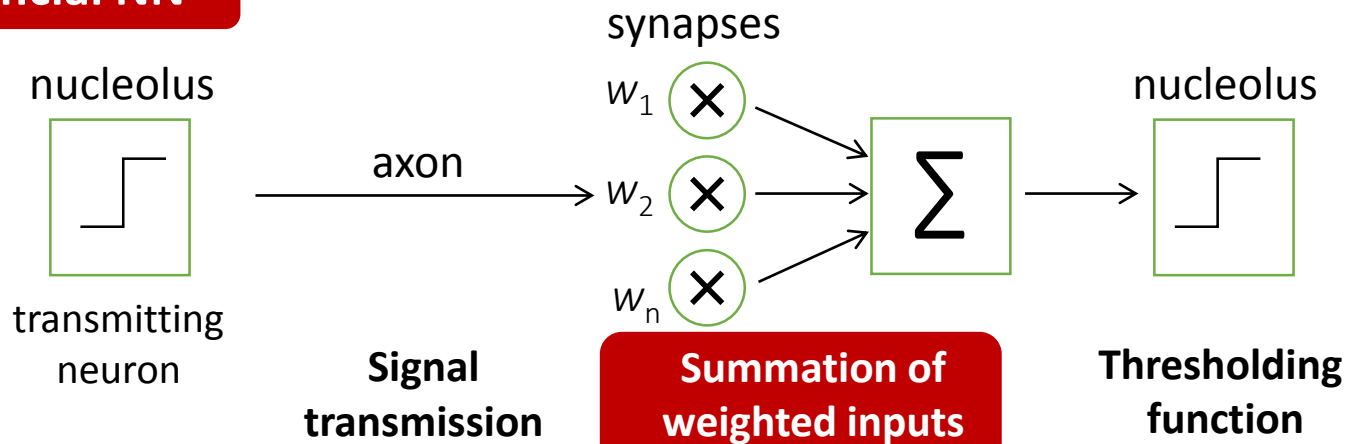
[3]: S. Ramasubramanian, et al., "SPINDLE: SPINtronic Deep Learning Engine for large-scale neuromorphic computing", ISLPED, 2014

[4]: D. Coue, etc "A four-quadrant subthreshold mode multiplier for analog neural network applications", TNN, 1996

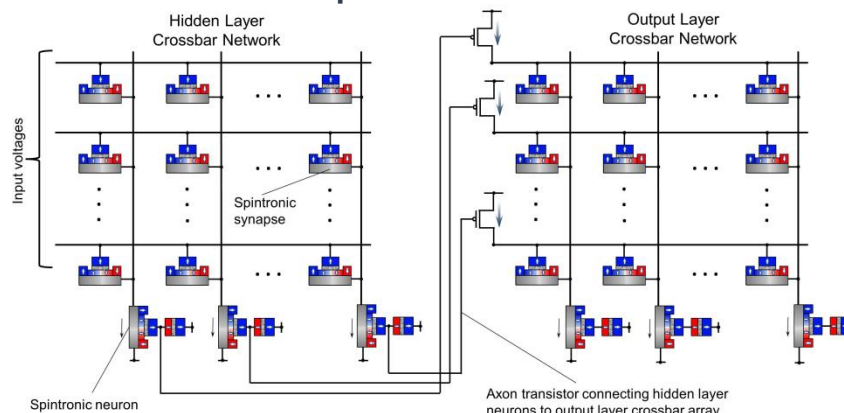
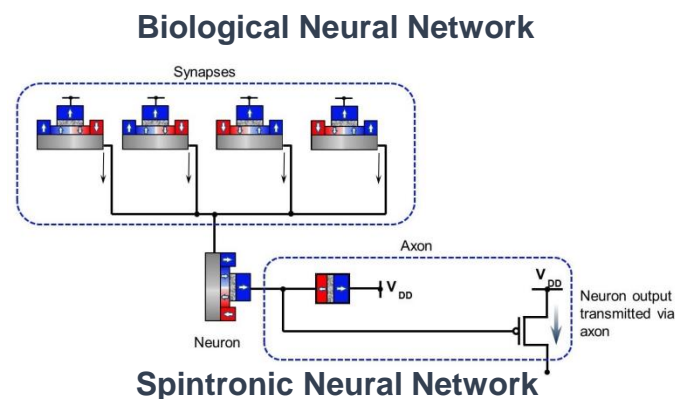
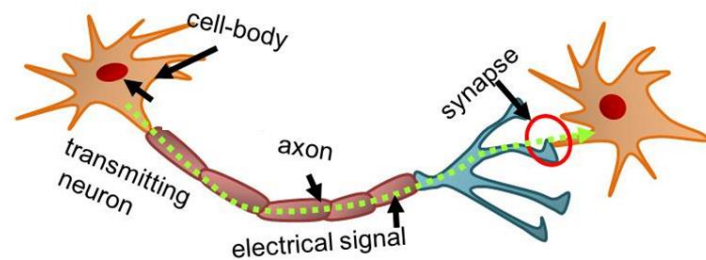
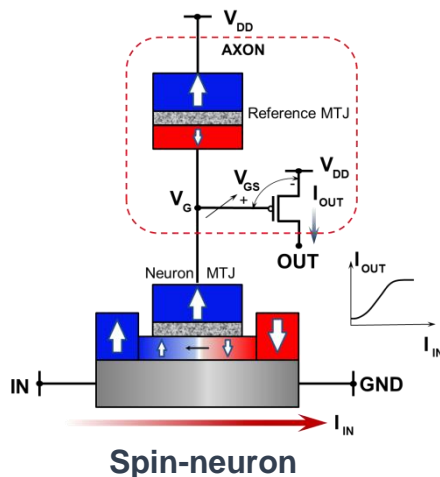
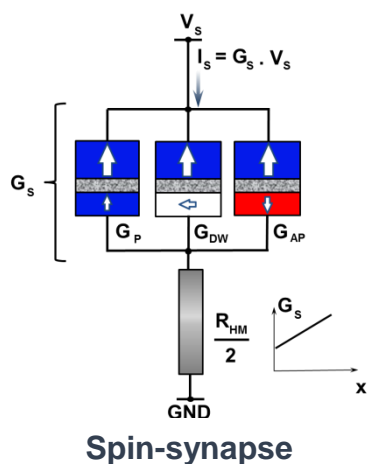
[5]: M. Sharad, etc, "Spin-neurons: A possible path to energy-efficient neuromorphic computers", JAP, 2013

In-Memory Computing (Dot Product)

Artificial NN



All-Spin Artificial Neural Network

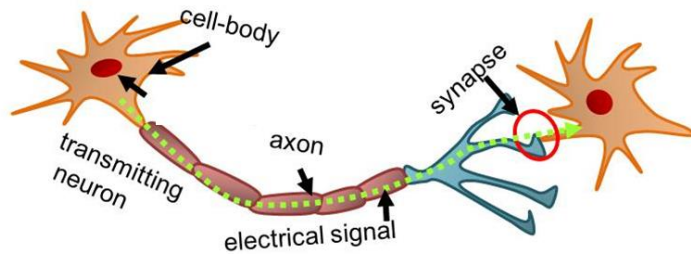


- All-spin ANN where spintronic devices directly mimic neuron and synapse functionalities and axon (CMOS transistor) transmits the neuron's output to the next stage
- Ultra-low voltage ($\sim 100\text{mV}$) operation of spintronic synaptic crossbar array made possible by magneto-metallic spin-neurons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

Spiking Neural Networks (Self-Learning)

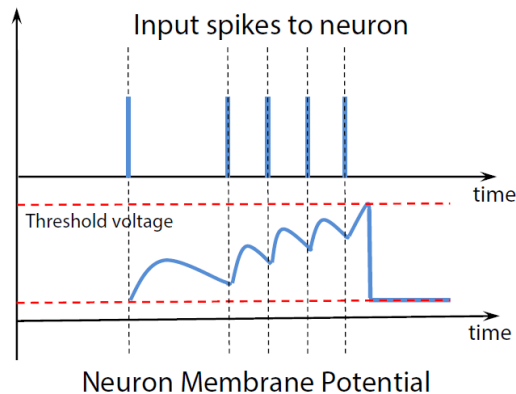
Spiking Neuron Membrane Potential

Biological Spiking Neuron

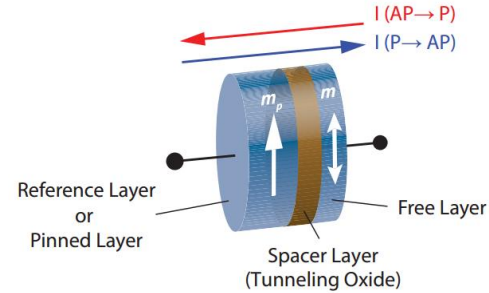


LIF Equation:

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

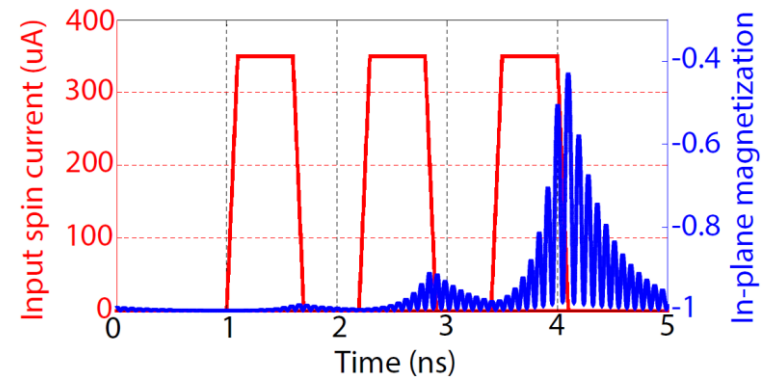


MTJ Spiking Neuron



LLGS Equation:

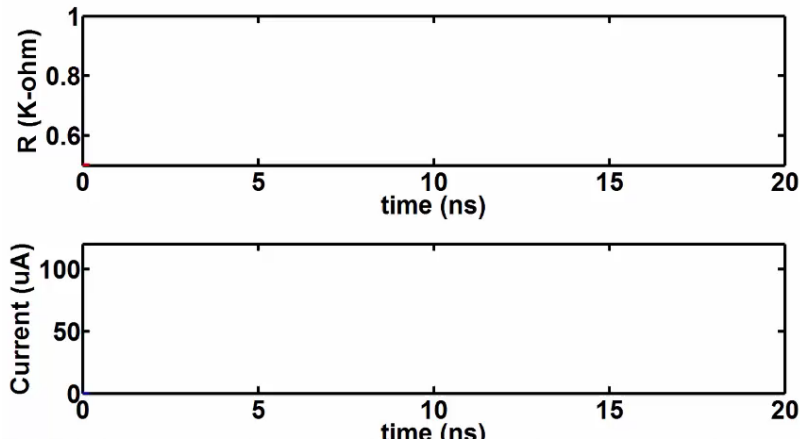
$$\frac{d\hat{\mathbf{m}}}{dt} = -\gamma(\hat{\mathbf{m}} \times \mathbf{H}_{eff}) + \alpha(\hat{\mathbf{m}} \times \frac{d\hat{\mathbf{m}}}{dt}) + \frac{1}{qN_s}(\hat{\mathbf{m}} \times \mathbf{I}_s \times \hat{\mathbf{m}})$$



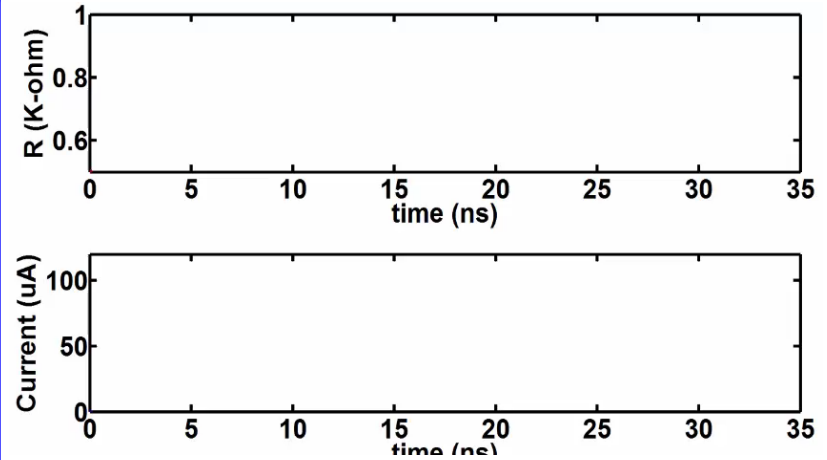
The leaky fire and integrate can be approximated by an MTJ – the magnetization dynamics mimics the leaky fire and integrate operation

MTJ as a Spiking Neuron

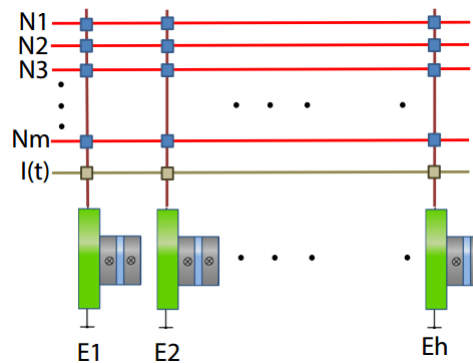
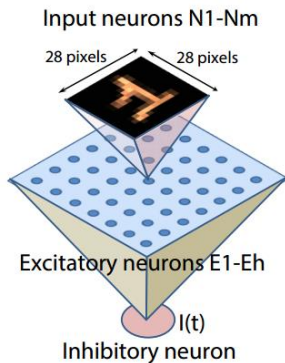
Spikes at 3ns interval



Spikes at 6ns interval

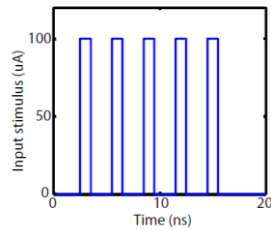


- **MTJ magnetization leaks and integrates input spikes (LLG equation) in presence of thermal noise**
- Associated “write” and “read” energy consumption is $\sim 1\text{fJ}$ and $\sim 1.6\text{fJ}$ per time-step which is much lower than state-of-the-art CMOS spiking neuron designs (267pJ [1] and 41.3pJ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

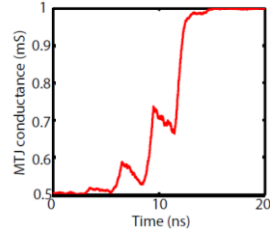


Spiking Neurons

LLGS Based Spiking Neuron

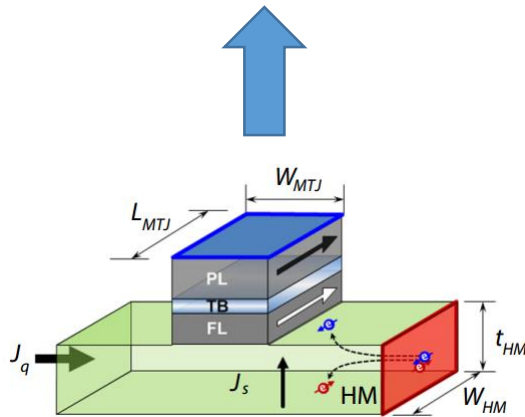


Input Spikes

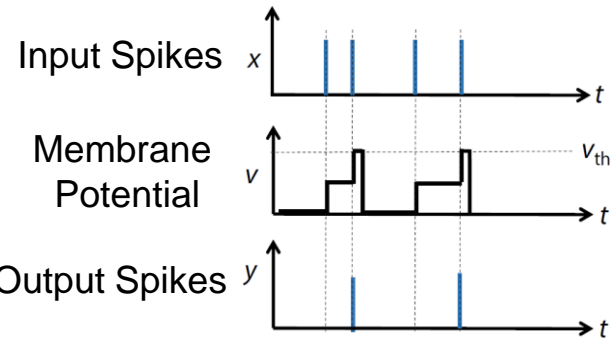


MTJ conductance

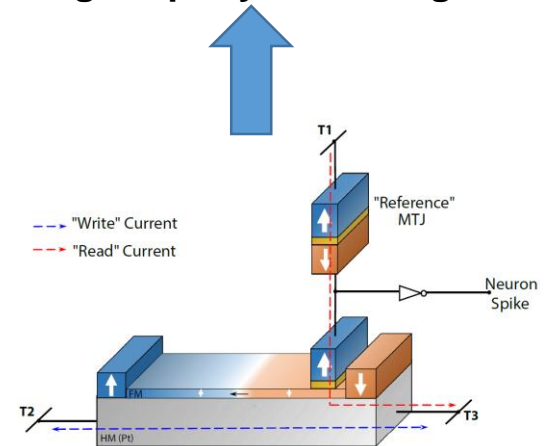
LLG Equation Mimicking Spiking Neurons



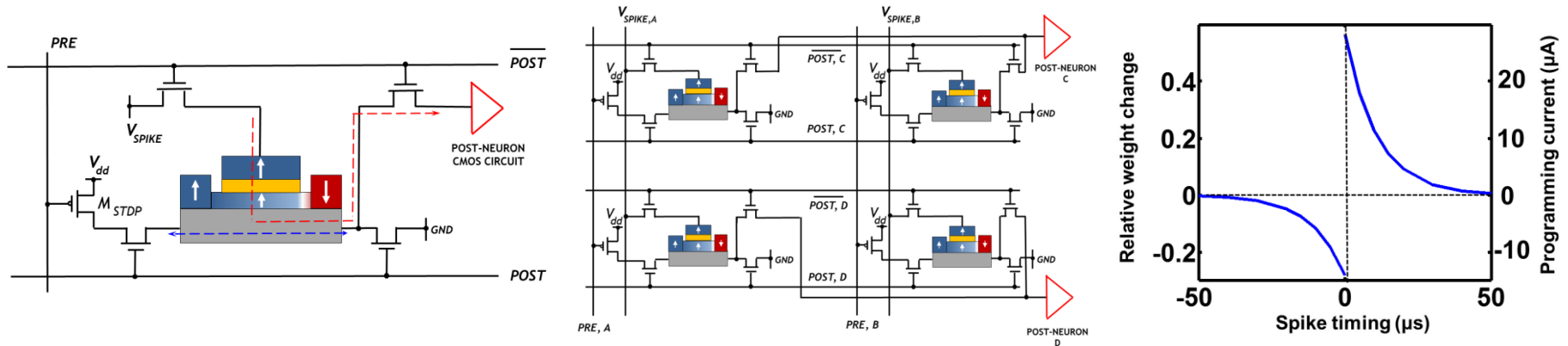
DW-MTJ base IF Neurons



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
- **15fJ energy consumption per synaptic event which is ~10-100x lower in comparison to SRAM based synapses /emerging devices like PCM**

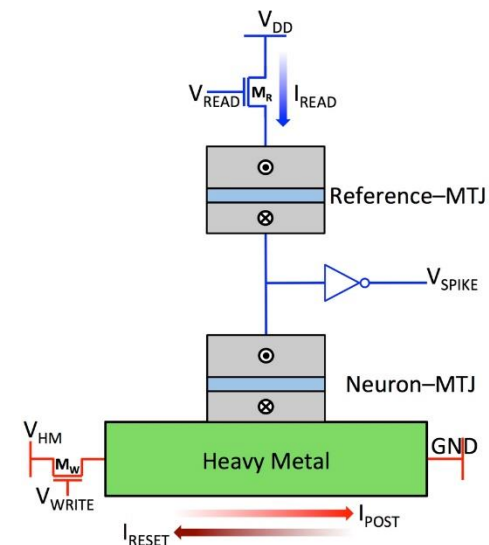
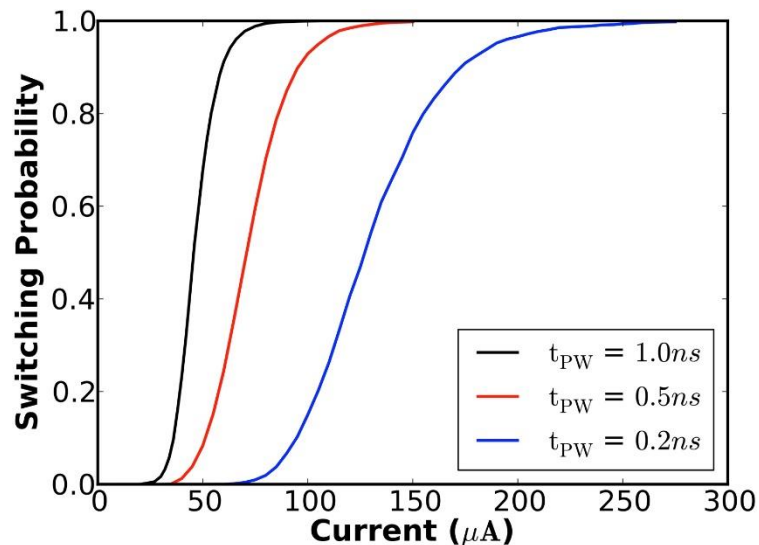
Comparison with Other Synapses

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)	100nm x 100nm	Threshold voltage ~2.2V	~300 μ s	2	Movement of Ag ions
FeFET	Y. Nishitani JJAP, 2013 (Panasonic, Japan)	Channel Length-3 μ m	Maximum gate voltage - 4V	10 μ s	3	Gate voltage modulation of ferroelectric polarization
Floating gate transistor	P. Hasler IEEE TBIOCAS, 2011 (GaTech)	1.8 μ m/0.6 μ m (0.35 μ m CMOS technology)	Vdd - 4.2V Tunneling Voltage - 15V	100 μ s (injection) 2ms (tunneling)	3	Injection and tunneling currents
SRAM synapse	B. Rajendran IEEE TED, 2013 (IIT Bombay)	0.3 μ m ² (10nm CMOS technology)	Average 328fJ for 4-bit synapse	-	-	Digital counter based circuits
Spintronic synapse	NRL Purdue	340nm x 20nm	Maximum 48fJ /event	1ns	3	Spin-orbit torque

MTJ Enabled All-Spin Spiking Neural Network

Probabilistic Spiking Neuron

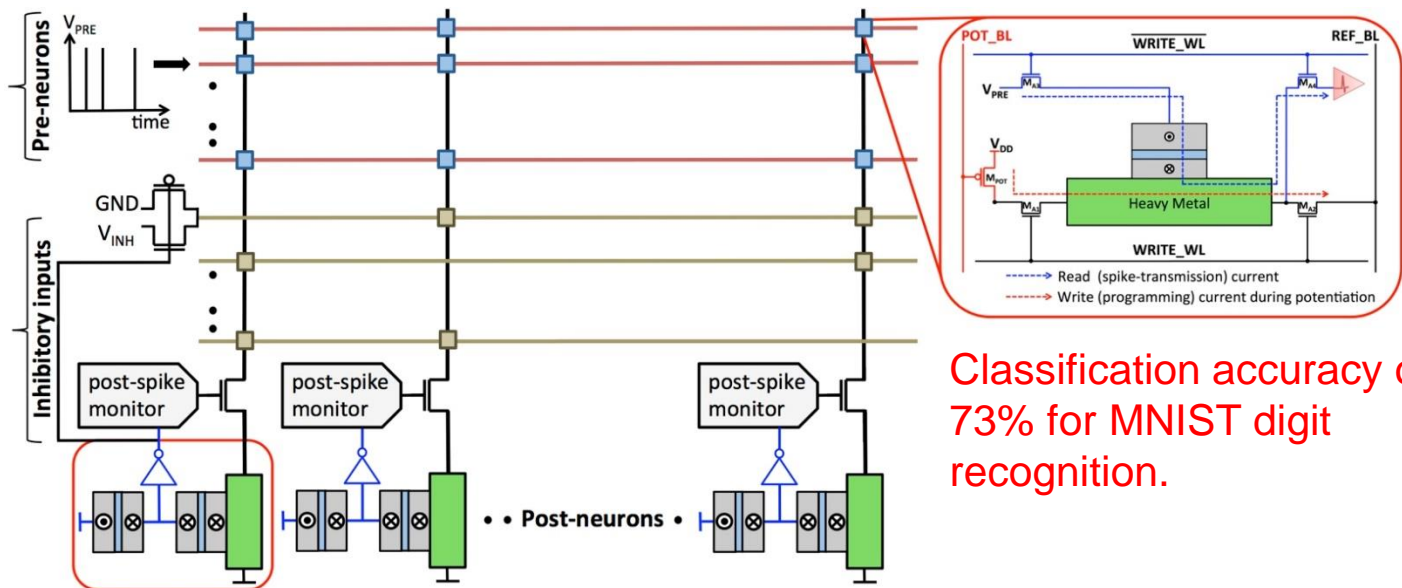
- A pre-neuronal spike modulated by synapse to generate current that controls the post-neuronal spiking probability.
- Exploit stochastic switching behavior of MTJ in presence of thermal noise.



MTJ Enabled All-Spin Spiking Neural Network

Stochastic SNN Hardware Implementation

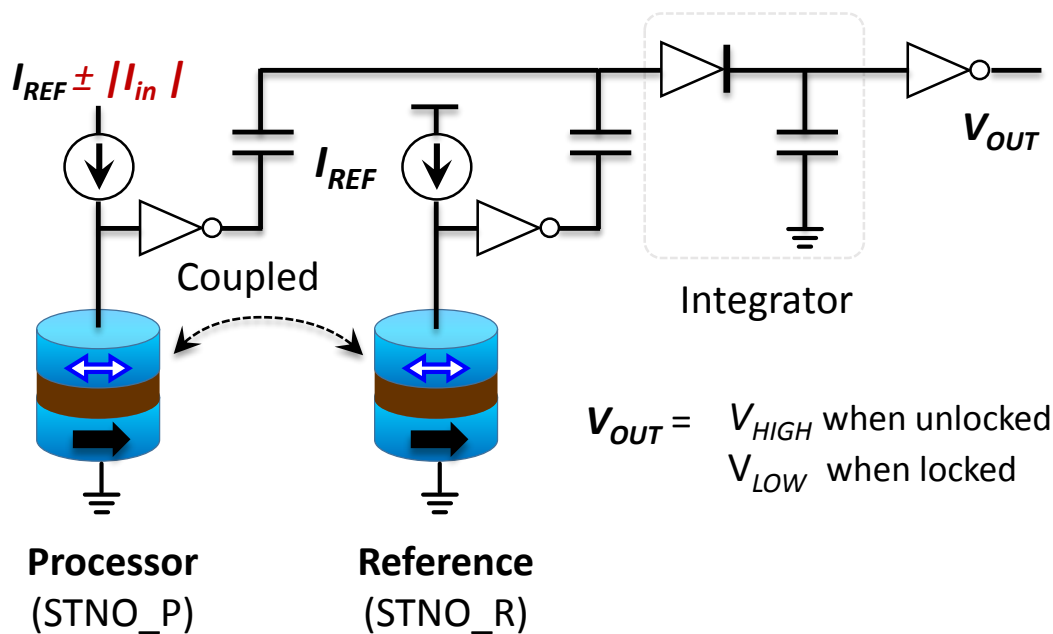
- Crossbar arrangement of the spin neurons and synapses for energy efficiency.
 - Average neuronal energy of 1fJ and 1.6fJ per timestep for write and read operations, and 4.5fJ for reset.
 - Average synaptic programming energy of 70fJ per training epoch.



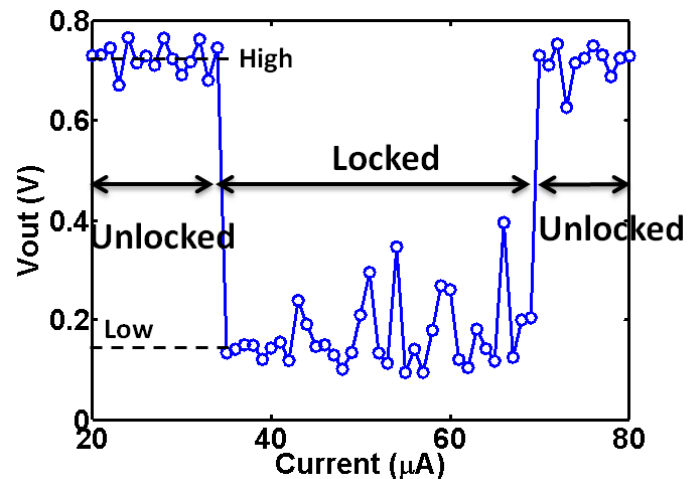
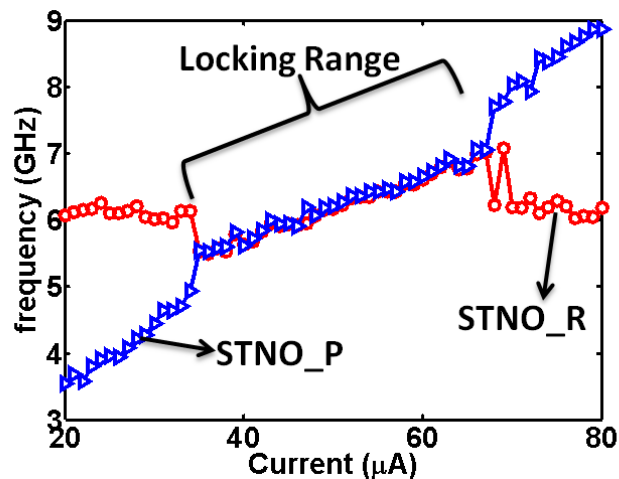
Classification accuracy of 73% for MNIST digit recognition.

Computing with Coupled STNOs

Coupled Spin-Torque Oscillators

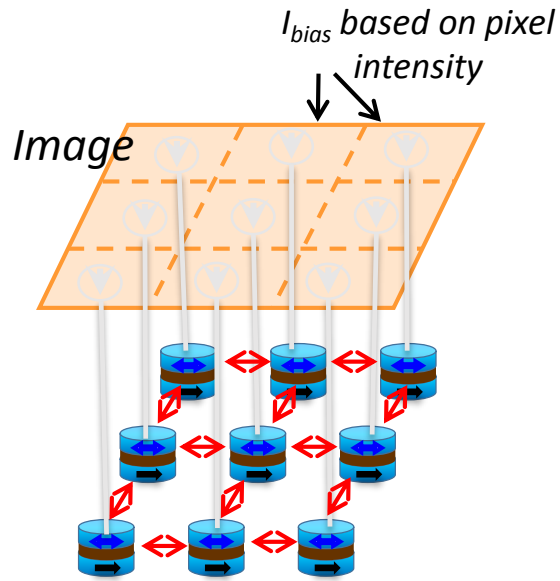


When I_{in} is within the locking range (neuron threshold), two oscillators are locked. Otherwise, they are unlocked.



STNOs can be used to provide thresholding functionality with tunable threshold

Edge Detection using STNOs



$t = 0.1\text{ns}$



Gilbert damping constant (α) = 0.01
 Saturation magnetization = 800 emu/cc
 Magnet volume (IMA) = 20x20x2 nm³
 $E_b = 30\text{kT}$
 $\lambda = 2$
 $\epsilon' = 0$
 $P = 0.9$
 $H_{\text{ext}} = 11\text{k Oe}$ at 0.45 degrees from normal to the plane;
 I_{bias} range = 10uA to 50uA (i.e 10uA for black and 50uA for white pixels);
 Distance between STOs = 70nm (for coupling)

Summary

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