NEUROMORPHIC COMPUTING WITH MAGNETO-METALLIC NEURONS & SYNAPSES: PROSPECTS AND PERSPECTIVES

KAUSHIK ROY

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

ELECTRICAL & COMPUTER ENGINEERING

PURDUE UNIVERSITY

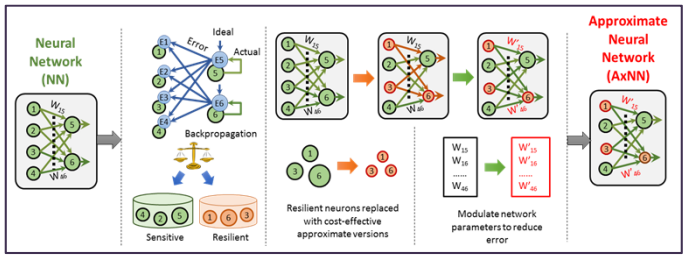
WEST LAFAYETTE, IN 47906, USA

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

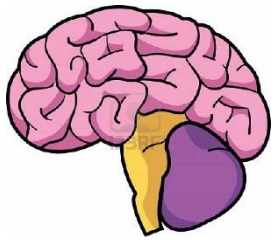
Neuromorphic Computing Technologies

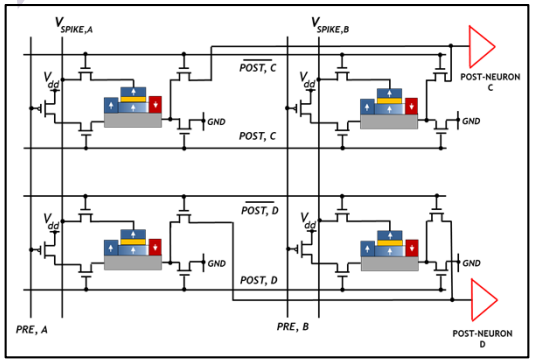


- Approximate Neural Nets, ISLPED ' 14
- Conditional Deep Learning, DATE 2016
- ….

Approximate Computing, Semantic Decomposition, Conditional DLN

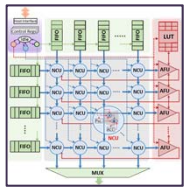
Post-CMOS Technologies

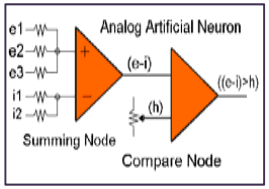




- Spin neuron, IJCNN ' 12, APL'15, TNANO, DAC, DRC, IEDM
- Spintronic Deep Learning Engine, ISLPED ' 14
- Spin synapse, APL ' 15
- ….

Hardware Accelerators

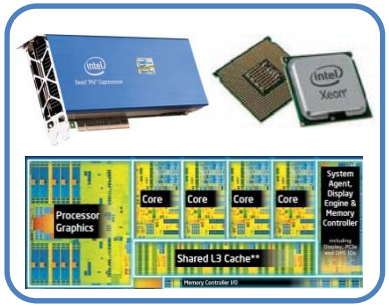




- QUORA, MICRO ' 13
- ….

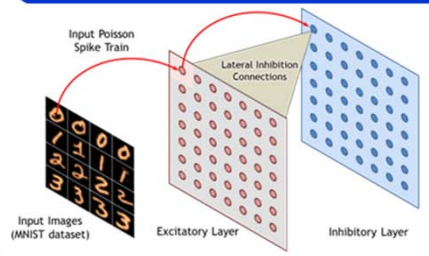
SW (Multicores/GPUs)

1 uJ/neuron

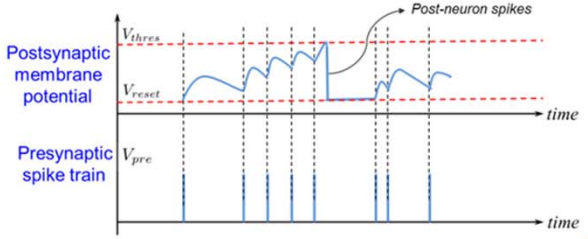


Device/Circuit/Algorithm Co-Design: Spin/ANN

Top-Down

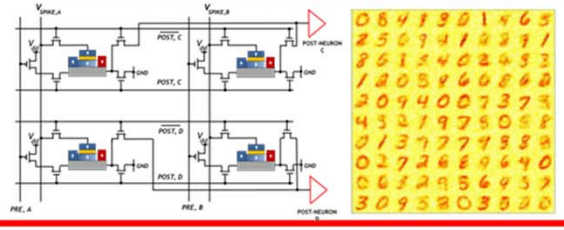


Investigate brain-inspired computing models to provide algorithm-level matching to underlying device physics

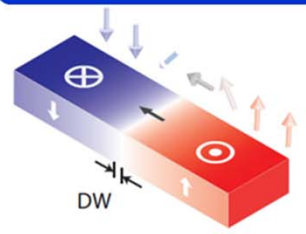


Device-Circuit-Algorithm co-simulation framework used to generate behavioral models for system-level simulations of neuromorphic systems

System Level Solution

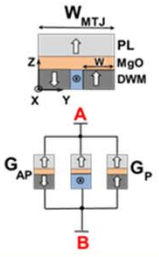


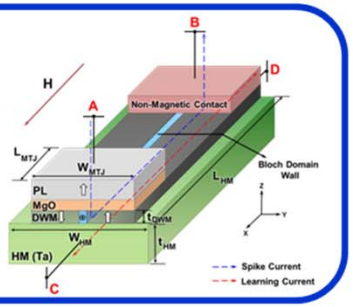
Bottom-Up



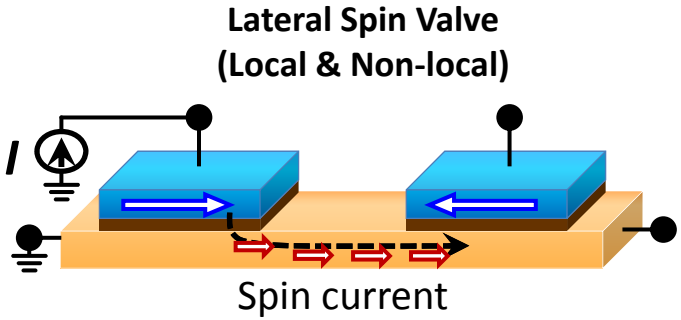
Investigate device physics to mimic "neuron/ synapse" functionalities

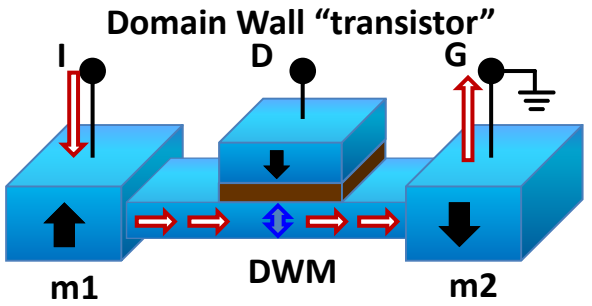
Calibration of device models with experiments

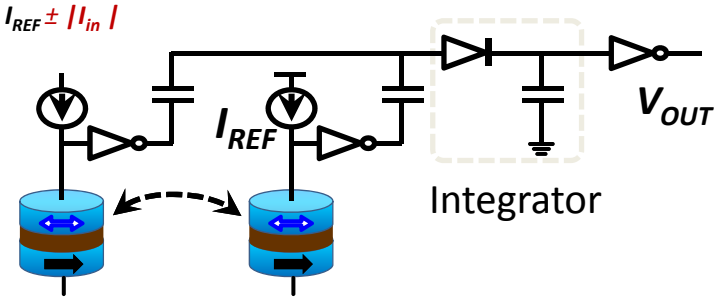


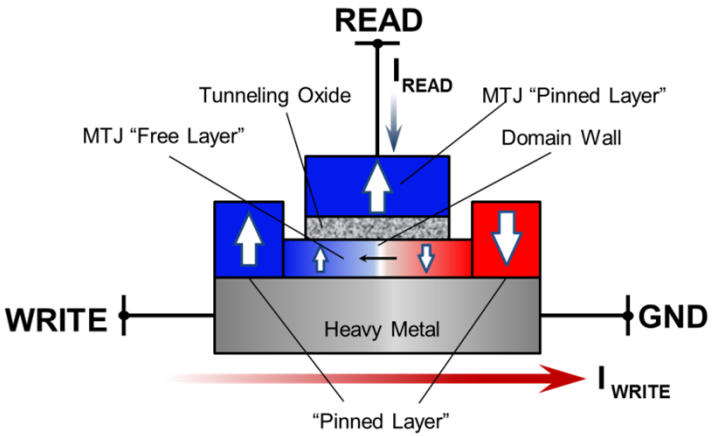


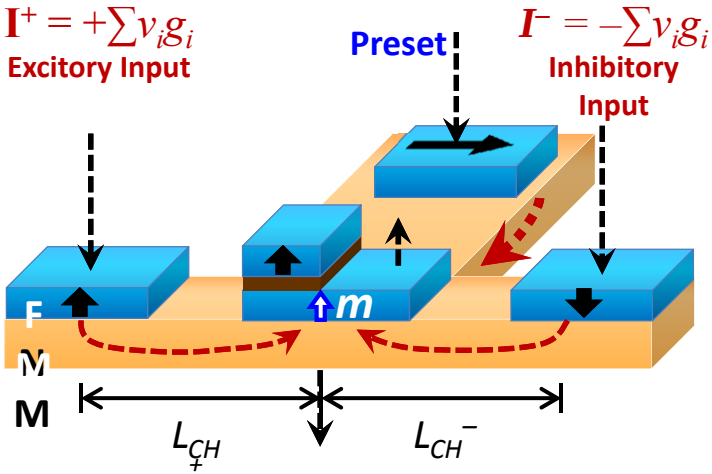
BUILDING BLOCKS: MEMORY, NEURONS, SYNAPSES

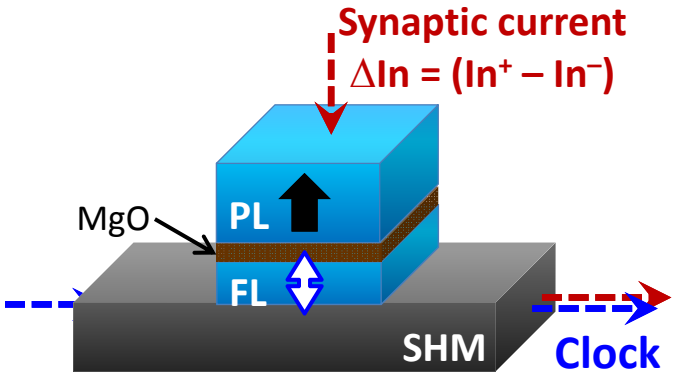




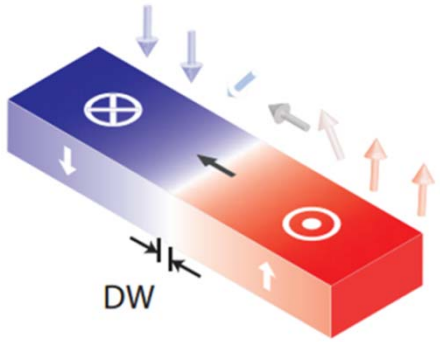


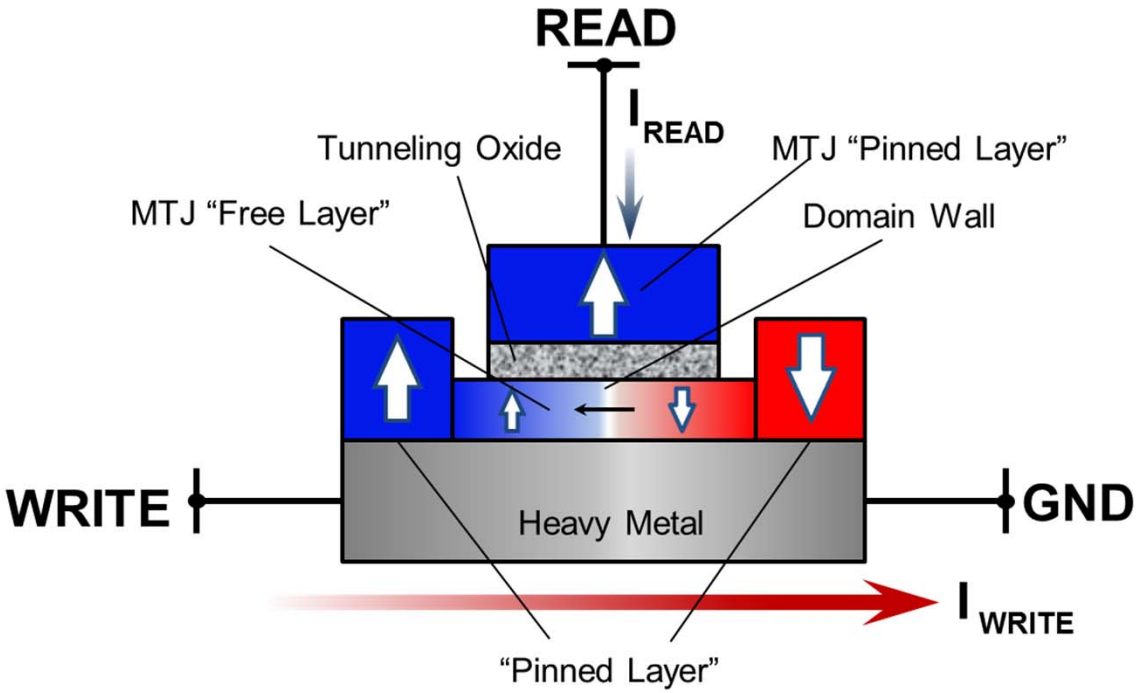






DW-MTJ: Domain Wall Motion/MTJ



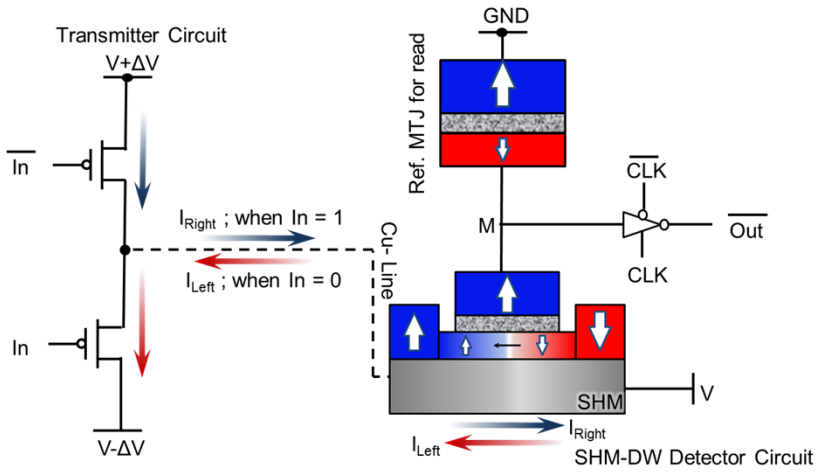


- Three terminal device structure provides decoupled "write" and "read" current paths
- Write current flowing through heavy metal programs domain wall position
- Read current is modulated by device conductance which varies linearly with domain wall position

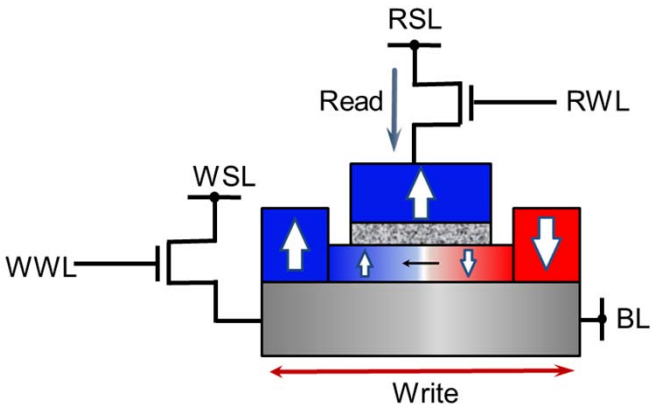
Universal device: Suitable for **memory, neuron, synapse, interconnects**

DW-MTJ for Interconnects/Memory

Interconnect Design



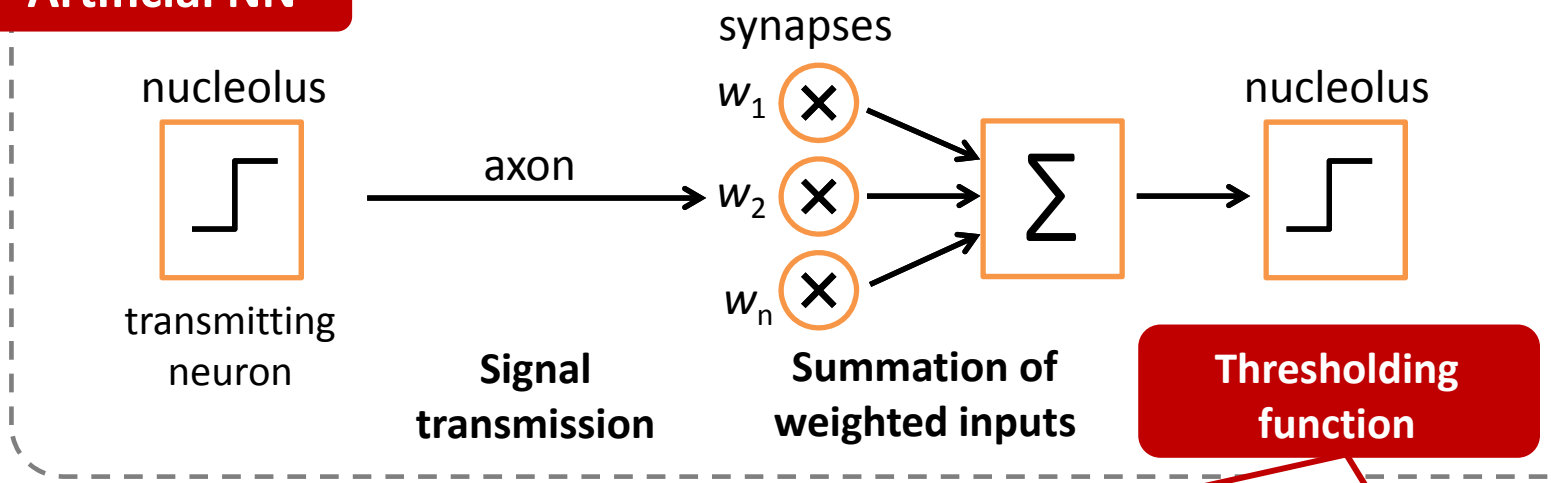
Memory Bit-Cell



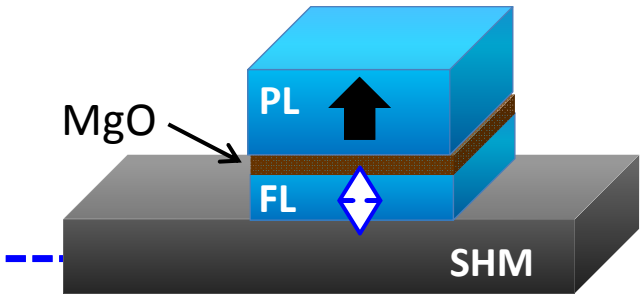
- Energy-efficient interconnect design can circumvent the energy and delay penalties in CMOS based global interconnects for scaled technology nodes
- DW-MTJ memory bit cell with decoupled “write” and “read” current paths

Thresholding (Activation)

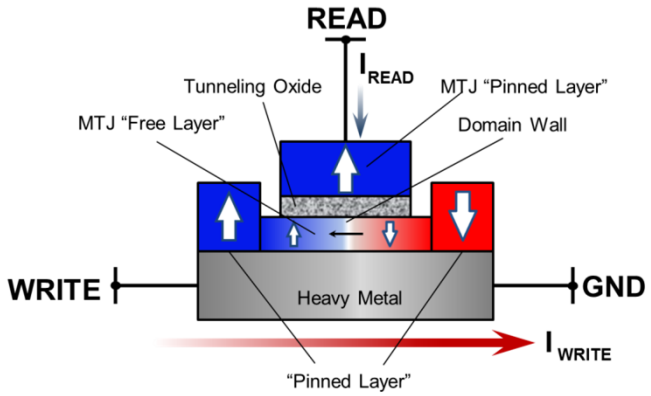
Artificial NN



Spin Hall based Switching

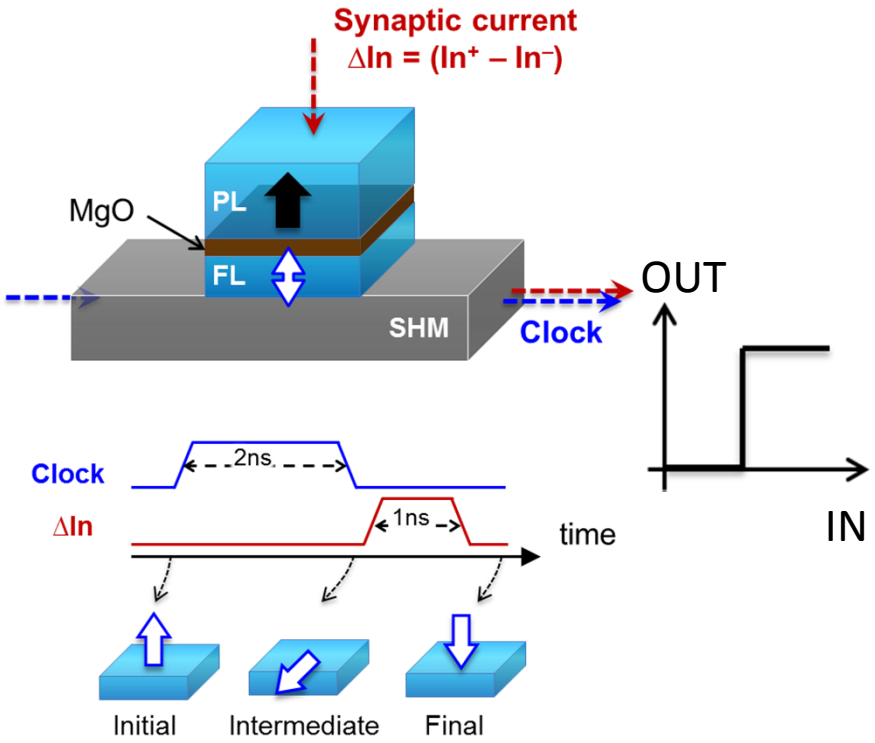


DW-MTJ

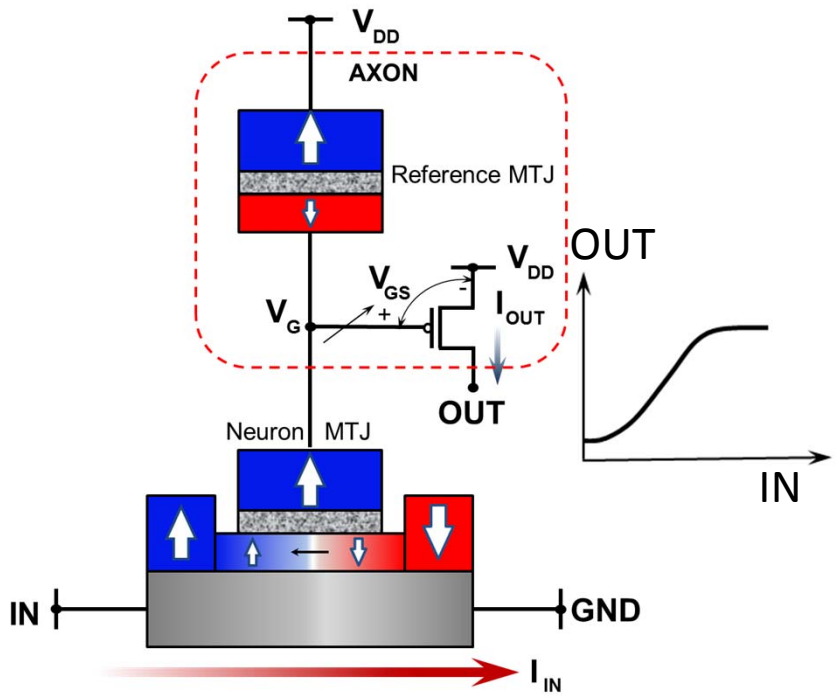


Switch a magnet using spin current, read using TMR effect

Step and Analog ANN Neurons



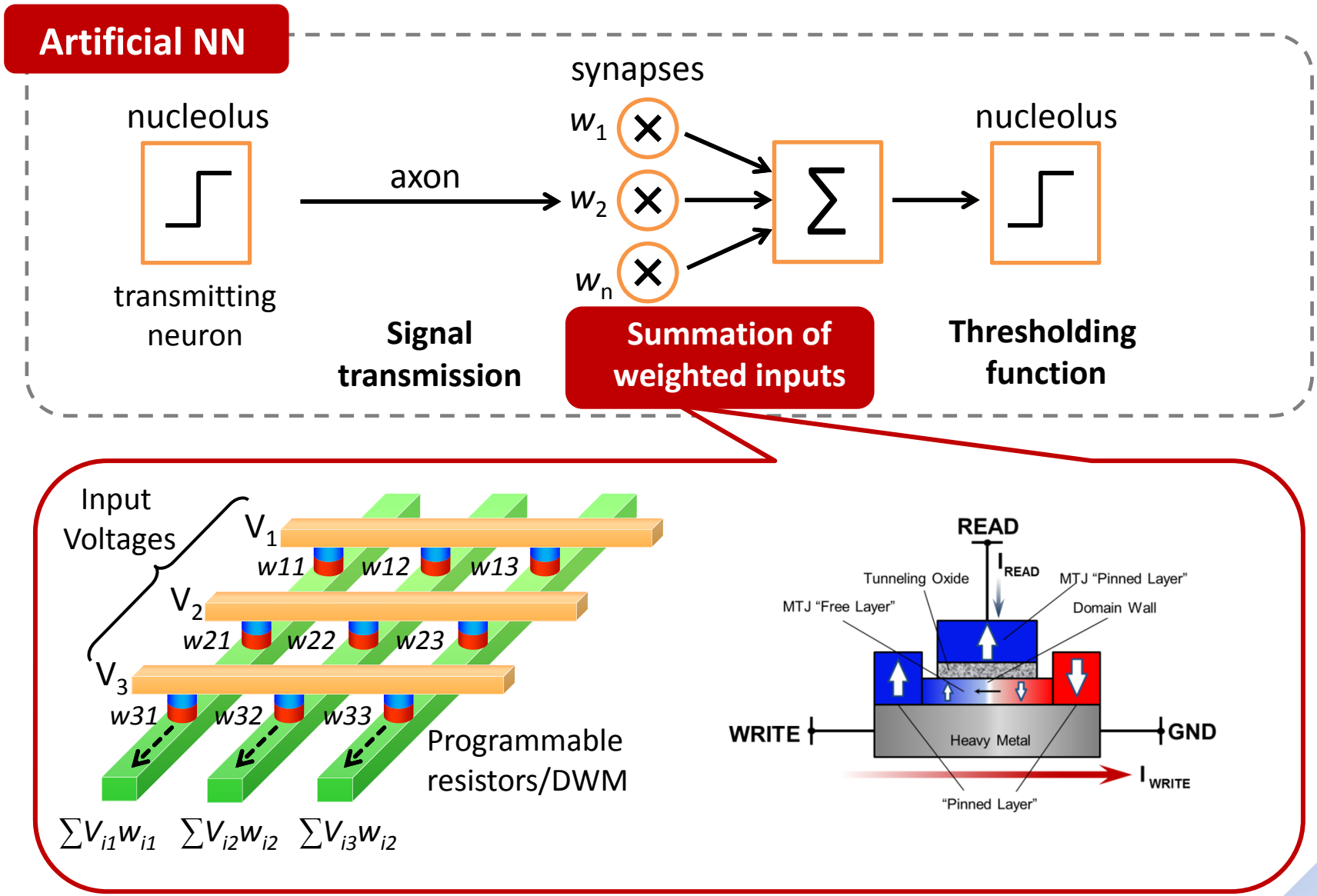
Step Neuron



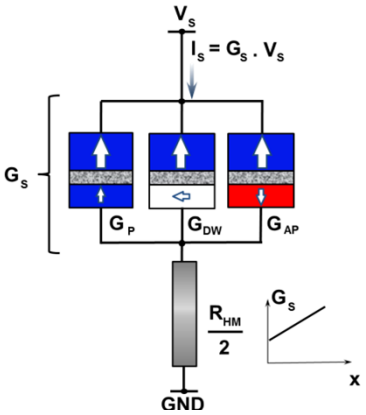
Analog Neuron

- 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 in Stochastic Neural nets

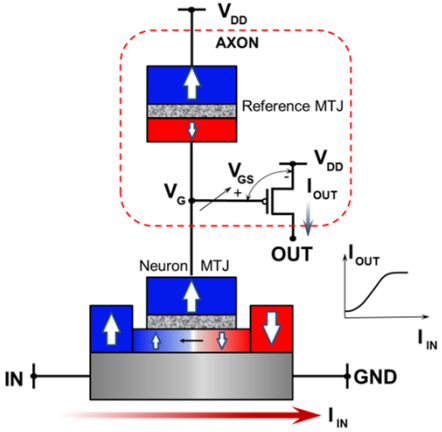
Sum of Weighted Inputs (Dot Product)



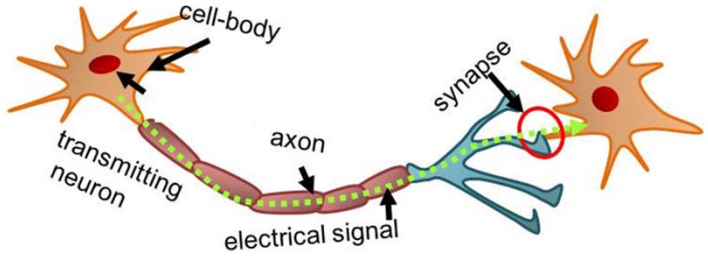
All-Spin Artificial Neural Network



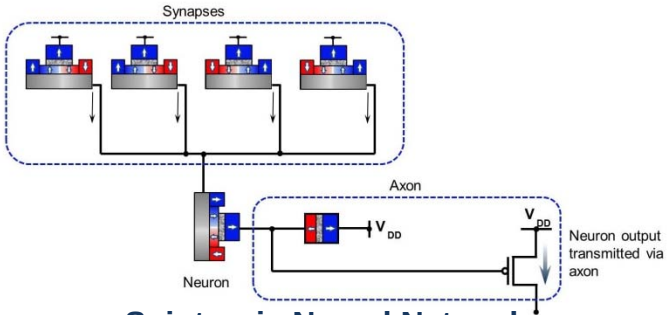
Spin-synapse



Spin-neuron

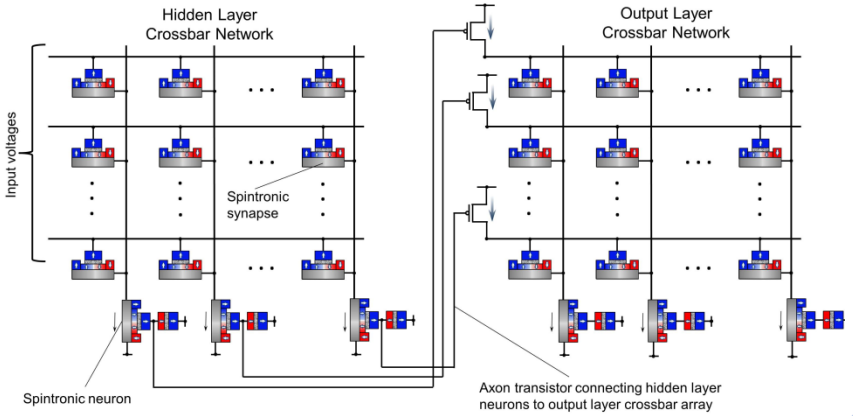


Biological Neural Network



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

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	/

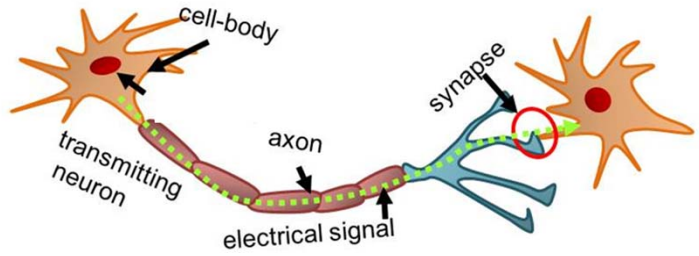
Compared with analog/ digital CMOS based neuron design, spin based neuron designs have the potential to achieve more than two orders lower energy consumption

[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

SPIKING NEURAL NETWORKS (SELF LEARNING)

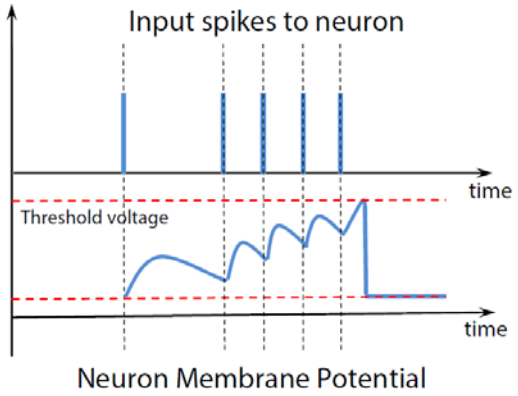
Spiking Neuron Membrane Potential

Biological Spiking Neuron

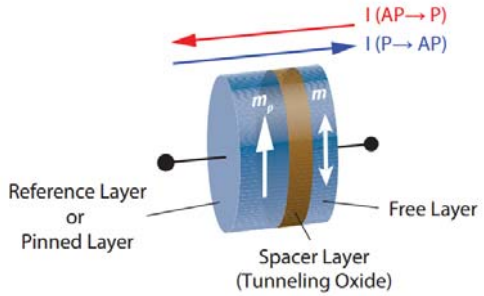


LIF Equation:

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

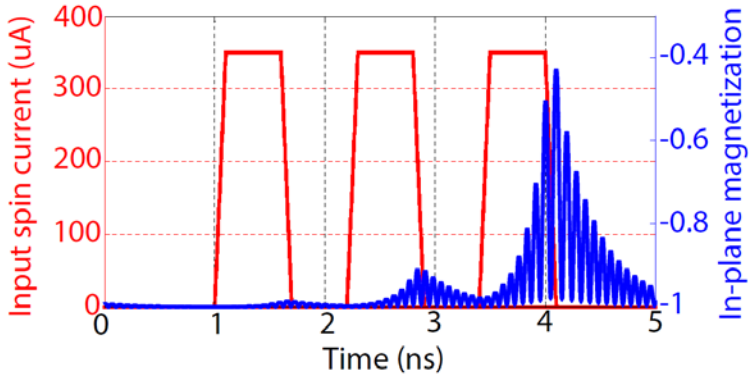


MTJ Spiking Neuron



LLGS Equation:

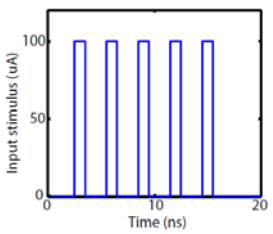
$$\frac{d\hat{\mathbf{m}}}{dt} = -\gamma(\hat{\mathbf{m}} \times \mathbf{H}_{eff}) + \alpha(\hat{\mathbf{m}} \times \frac{d\hat{\mathbf{m}}}{dt}) + \frac{1}{qN_s}(\hat{\mathbf{m}} \times \mathbf{I}_s \times \hat{\mathbf{m}})$$



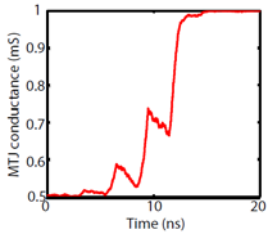
The leaky fire and integrate can be approximated by an MTJ – the magnetization dynamics mimics the leaky fire and integrate operation

Spiking Neurons

LLGS Based Spiking Neuron

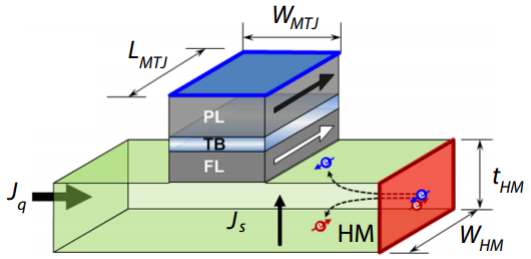


Input Spikes

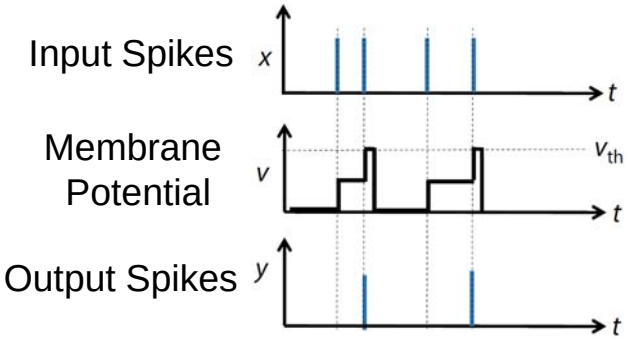


MTJ conductance

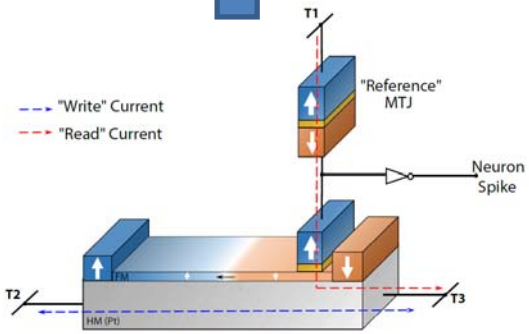
LLG Equation Mimicking Spiking Neurons



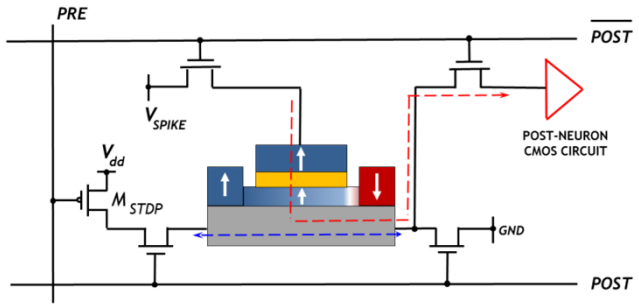
DW-MTJ base IF Neurons

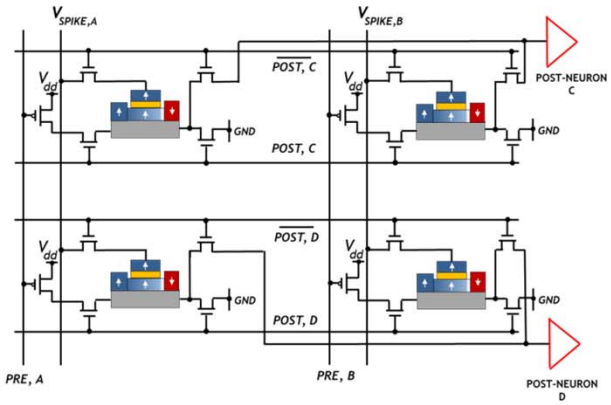


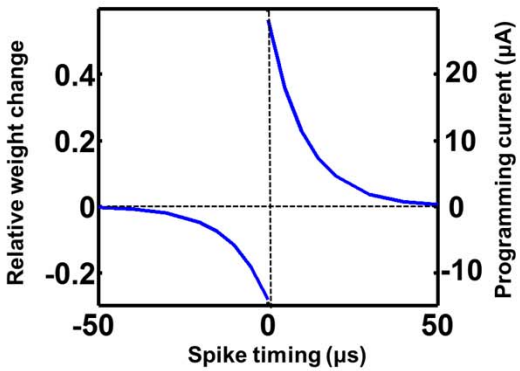
DW Integrating Property Mimicking IF Neuron



Arrangement of DW-MTJ Synapses in Array for STDP Learning



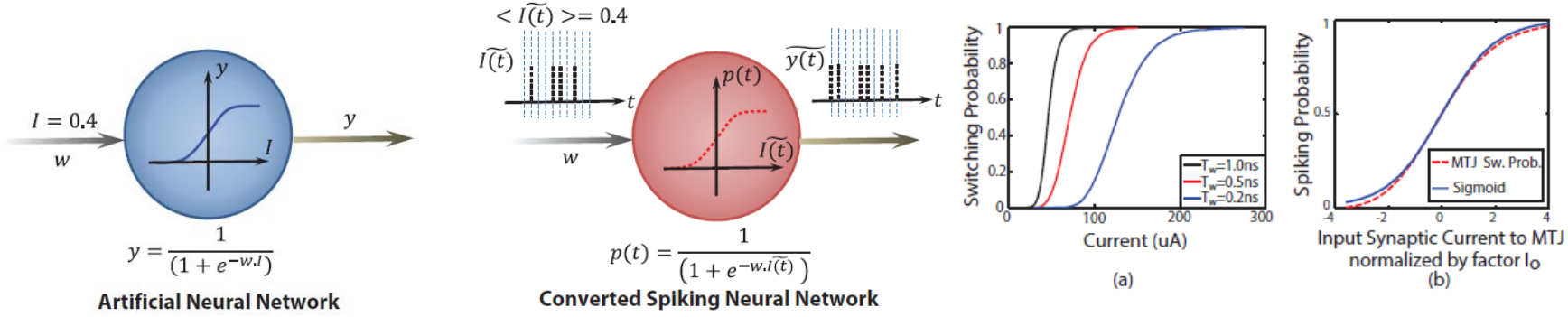




Spike-Timing Dependent Plasticity

- 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**

Stochastic SNN

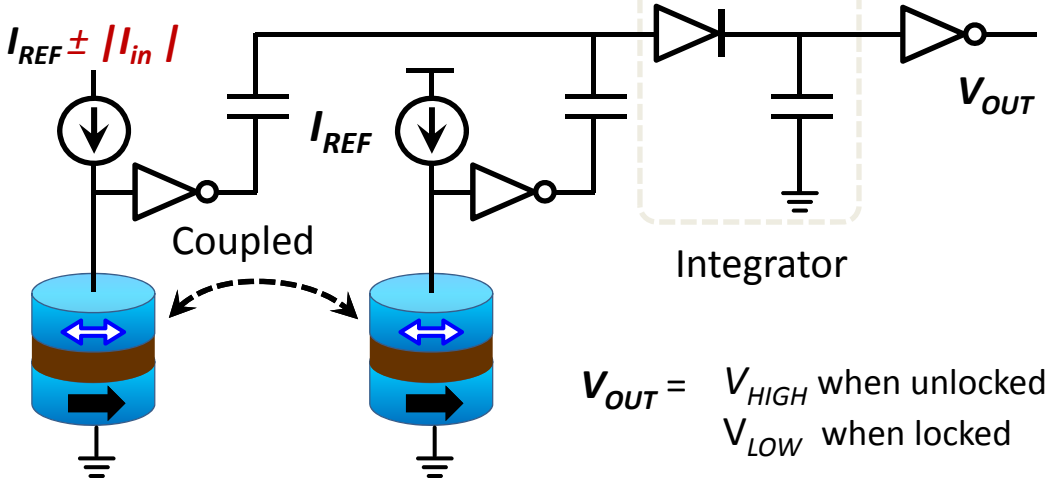


- We propose ANN-SNN conversion where the neural transfer function is interpreted as the spiking probability of the neuron in a particular time-step
- Such a functionality is enabled by the stochastic device physics of switching in a Magnetic Tunnel Junction in presence of thermal noise
- **System-level simulations indicate energy consumption of 19.5nJ per image classification at the end of 50 time-steps of SNN simulation (>97% accuracy on MNIST dataset)**

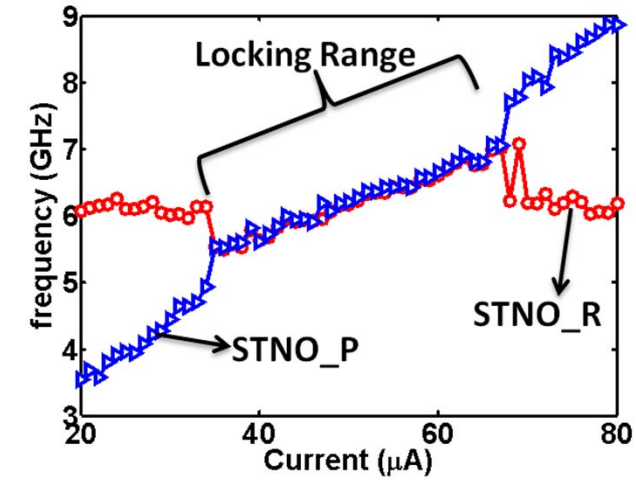
Sengupta, Roy *et al.*, IEEE Trans. On Electron Devices, 2016

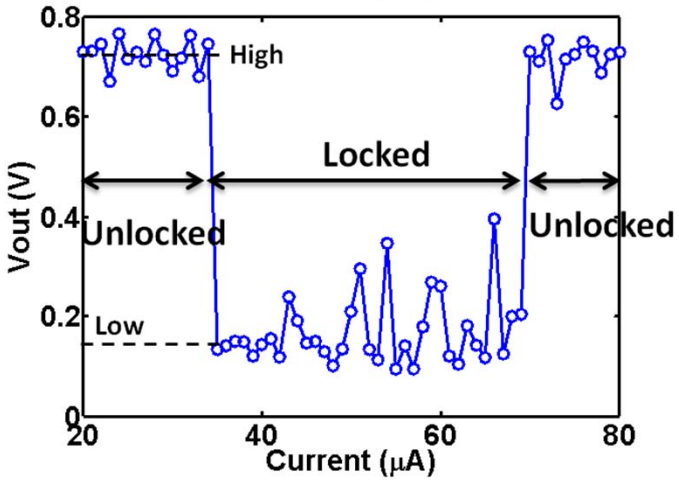
Computing with Coupled STNOs

Spin-Neurons & Synapses: Coupled Spin-Torque Oscillators



V_{OUT} = V_{HIGH} when unlocked
V_{LOW} when locked

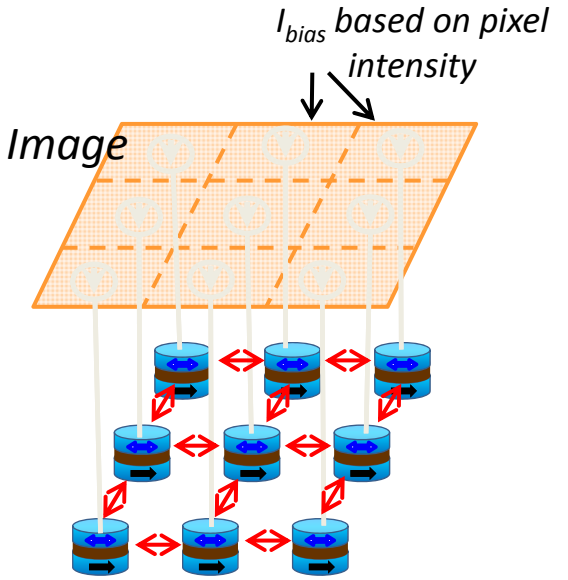




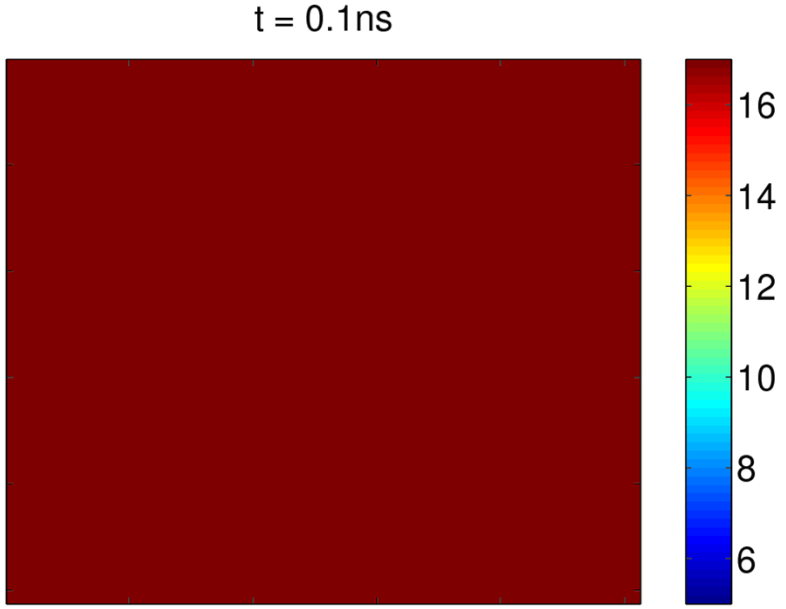
When $|I_{in}|$ is within the locking range (neuron threshold), two oscillators are locked. Otherwise, they are unlocked.

STNOs can be used to provide thresholding functionality with tunable threshold

Edge Detection using STNOs



Gilbert damping constant (alpha) = 0.01
Saturation magnetization = 800 emu/cc
Magnet volume (IMA) = 20x20x2 nm3
Eb = 30kT
lambda = 2
epsilon prime = 0
P = 0.9
Hext = 11k Oe at 0.45 degrees from normal to the plane; Ibias 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 computin
  - Need for new leaning techniques suitable for emerging devices
  - Materials research, new physics, new devices, simulation models
- A long (but interesting) path ahead…