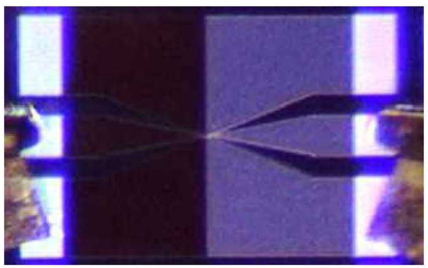
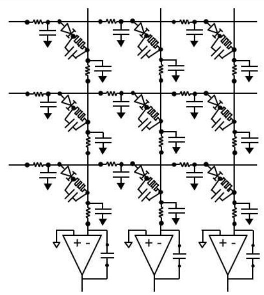
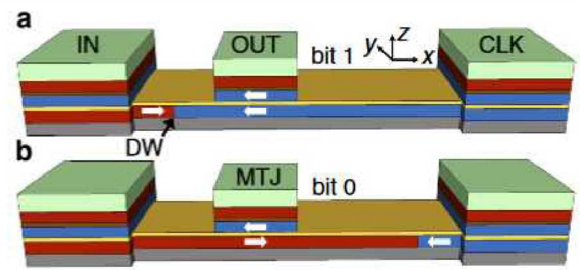
SAND2019-9563C







Semi-supervised learning and inference in domain-wall magnetic tunnel junction (DW-MTJ) neural networks

Christopher Bennett*,[1] Jean Anne C. Incorvia [2], Naimul Hassan [3], Xuan Hu [3], Joseph S. Friedman [3], Matthew J. Marinella [1]

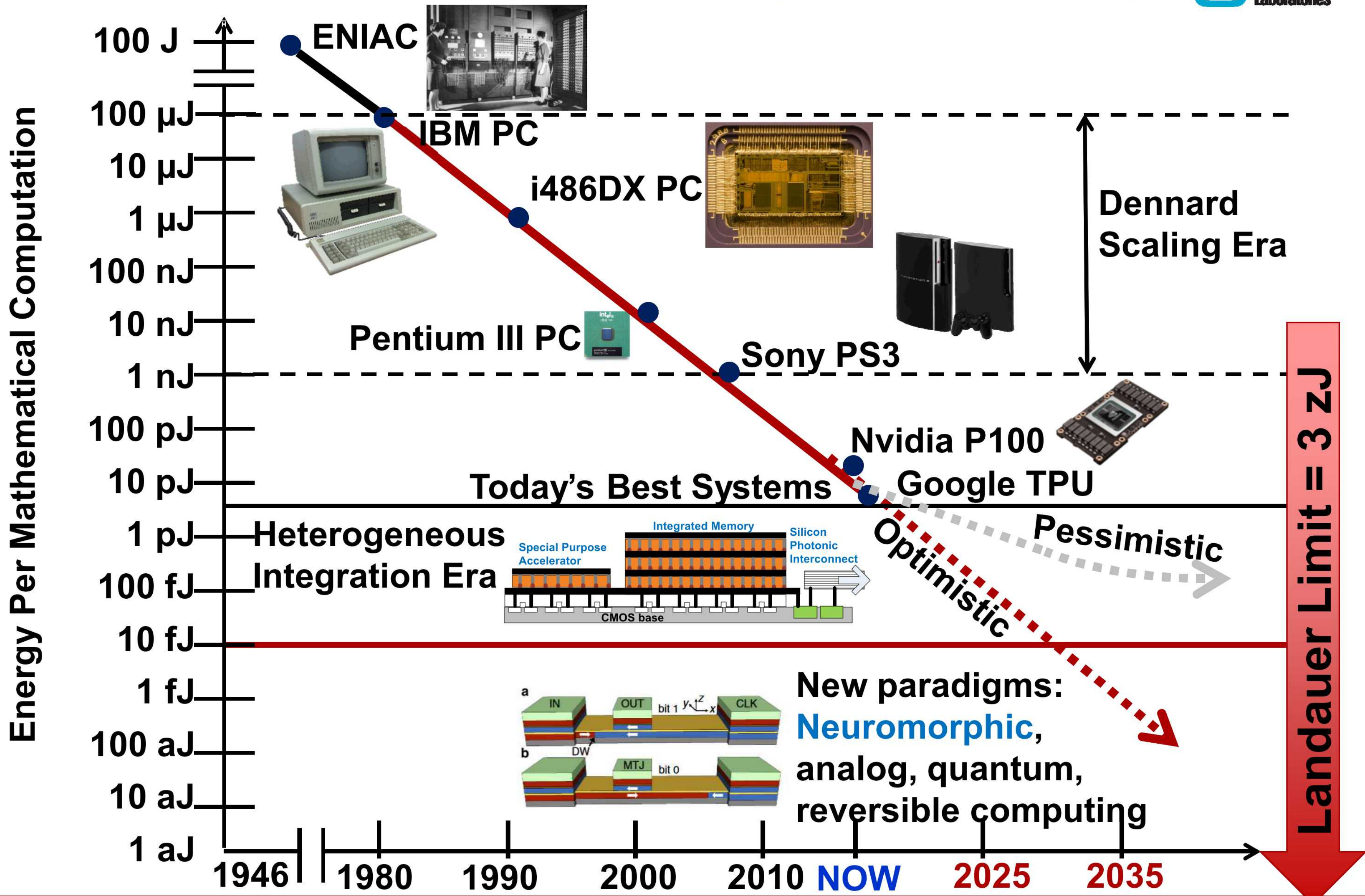
[1] Sandia National Laboratories [2] University of Texas, Austin [3] University of Texas, Dallas
*cbennet@sandia.gov

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Outline

- **Intro and Motivation: Building Accelerators**
- **Opportunities for Spintronic Neural Networks**
- **Free Bio-realistic Neuronal effects from DW-MTJs**
- **Applications: Supervised, Semi-supervised Learning**
- **Analysis: Coupling Strength + Comparison**
- **Summary & Future Work**

Evolution of Computing Machinery

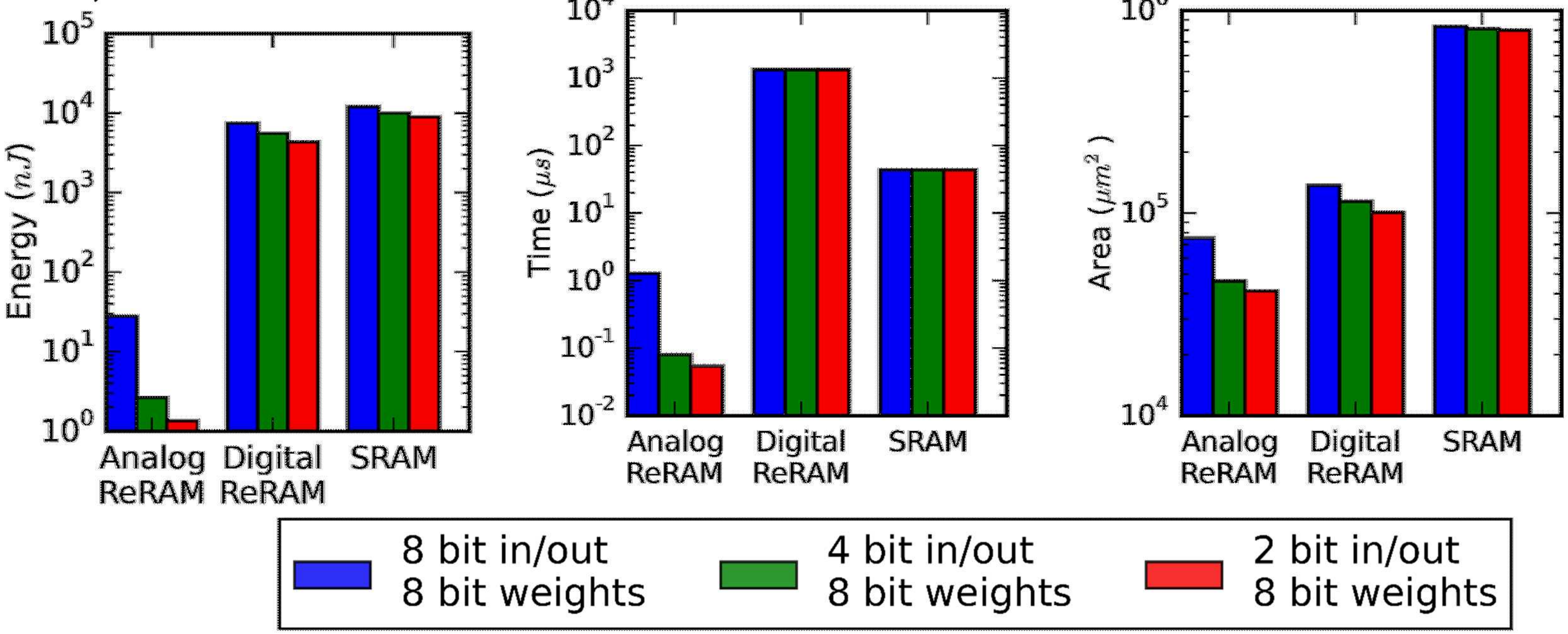


Why analog accelerators?

Three orders magnitude energy savings w/ in-memory computing

- Vector Matrix Multiply (VMM)
- Matrix Vector Multiply (MVM)
- Outer Product Update (OPU)

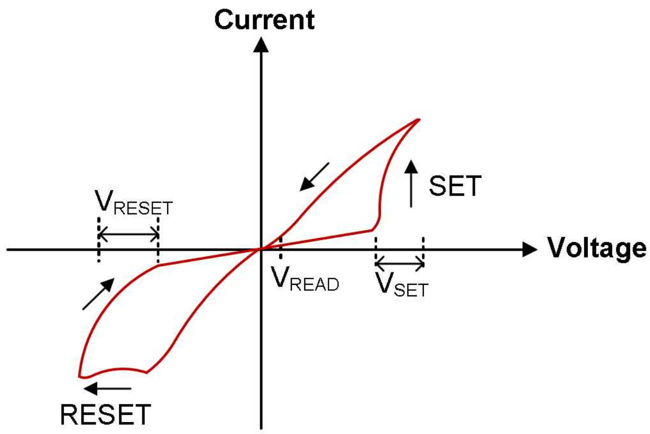
Energy 430 – 6,900X over SRAM | **Latency** 35 – 800X over SRAM | **Area** 11 – 20X over SRAM



Realizing physical matrix kernels

Sandia National Laboratories

- Ideal Vector-Matrix Mulitply :
  - Electrically realisable using Kirchoff's + Ohm's laws
- Programmable resistors - e.g. ReRAM/MRAM devices- key component
  - Small voltages to read (inference)
  - Large voltages to program



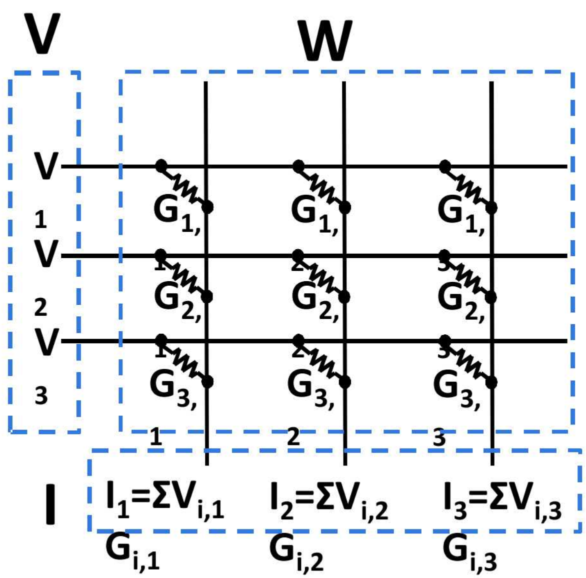
Mathematical

$$V^T W = I$$

$$\begin{bmatrix} V_1 & V_2 & V_3 \end{bmatrix} \begin{bmatrix} W_{1,1} & W_{1,2} & W_{1,3} \\ W_{2,1} & W_{2,2} & W_{2,3} \\ W_{3,1} & W_{3,2} & W_{3,3} \end{bmatrix} =$$

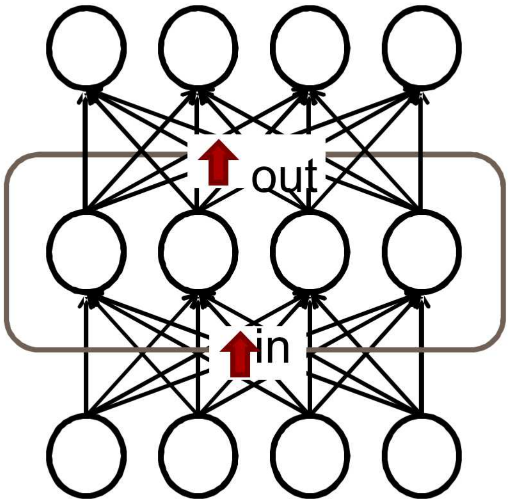
$$\begin{bmatrix} I_1=\Sigma V_{i,1} W_{i,1} & I_2=\Sigma V_{i,2} W_{i,2} & I_3=\Sigma V_{i,3} W_{i,3} \end{bmatrix}$$

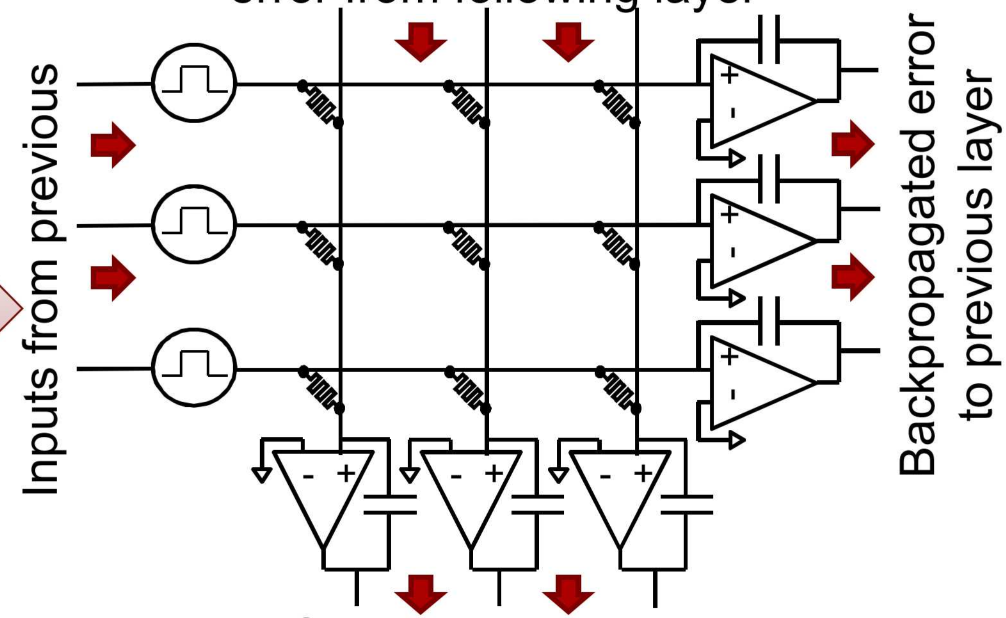
Electrical



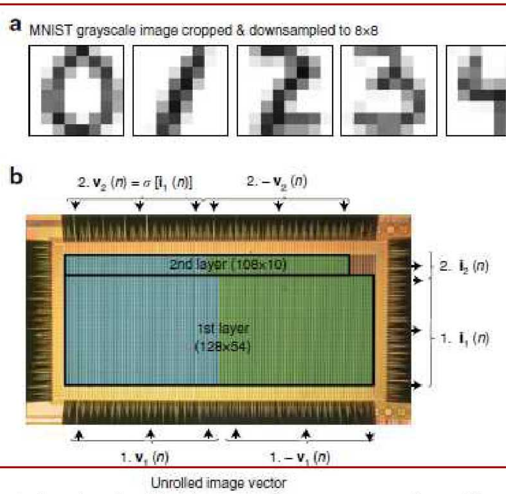
Mapping Neural Networks to Crossbars

Concept

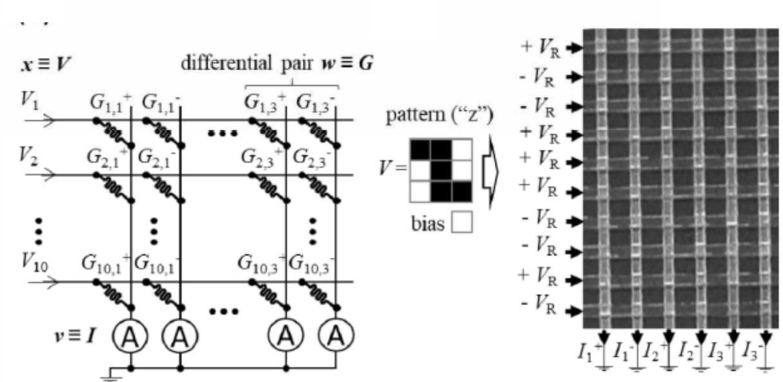




Prototype Experiments



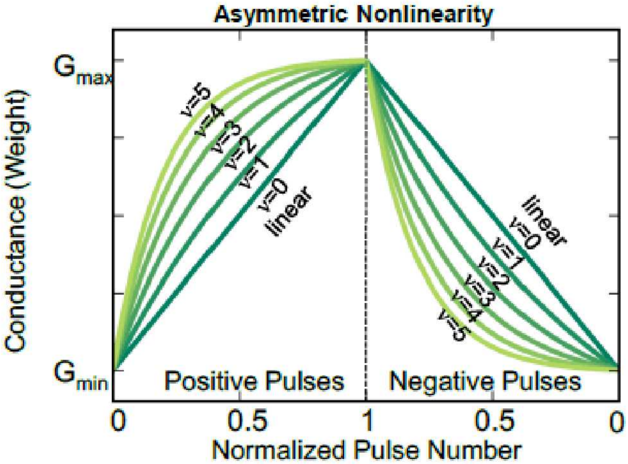
Source: Li et al, Nature Communications 2018

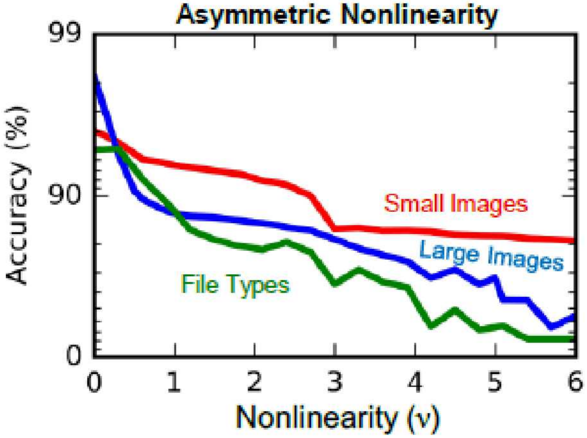


Source: Prezioso et al, Nature 2015

Challenges for adaptive analog accelerators

- Emerging ReRAM : far from ideal , floating-point <u>'weights'</u>
- Several key problems:
  - Limited resolution
  - Read and write noise
  - Device stochasticity
  - Device non-linearity
  - Device asymmetry
- Preliminary analysis: most severe impact from *asymmetric non-linearity*
- *How can we get around this??*
  - Increase bio-realism of learning accelerators -> lower synapse, neuron requirements
  - The brain does not use backprop *(at least as we currently apply it in ML).*





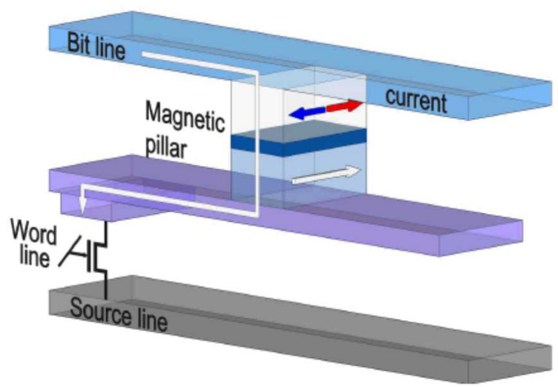
Agarwal et al, IJCNN 2016

Outline

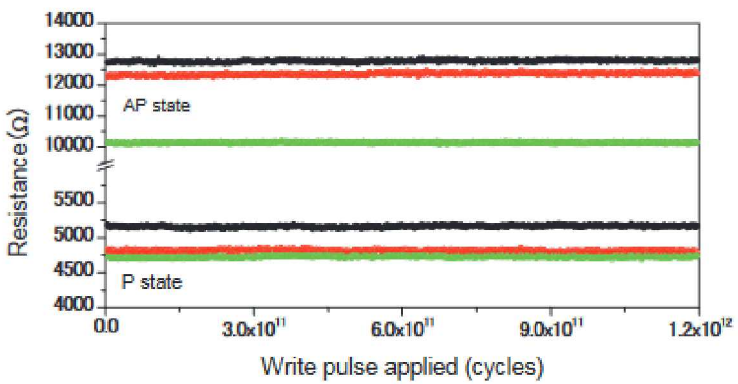
- **Intro and Motivation: Building Accelerators**
- **Opportunities for Spintronic Neural Networks**
- **Free Bio-realistic Neuronal effects from DW-MTJs**
- **Applications: Supervised, Semi-supervised Learning**
- **Analysis: Coupling Strength + Comparison**
- **Summary & Future Work**

Major opportunity:
Building neural networks with spintronics components

- Spintronic components alleviate signature device issues of ReRAM accelerators.
  - STT-MTJ/SOT-MTJ: intrinsically binary + stochastic -> non-linearity irrelevant.
  - Magnetic devices with analog behavior (Domain wall, skyrmionics) : different physics, non-linearity immune
- Additional Advantages:
  - Extreme endurance (important for online learning + inference)
  - Low energy footprint : typically <1V programming, <50ns programming .
  - Extreme compactness and CMOS-compatible 1T1R array scaling (BEOL integration)



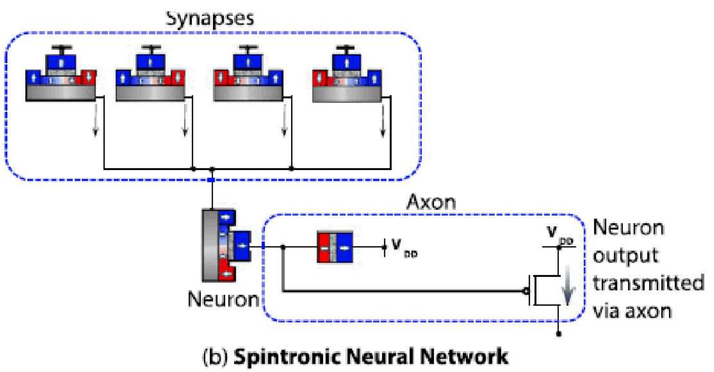
Makarov et al , IOP 2016



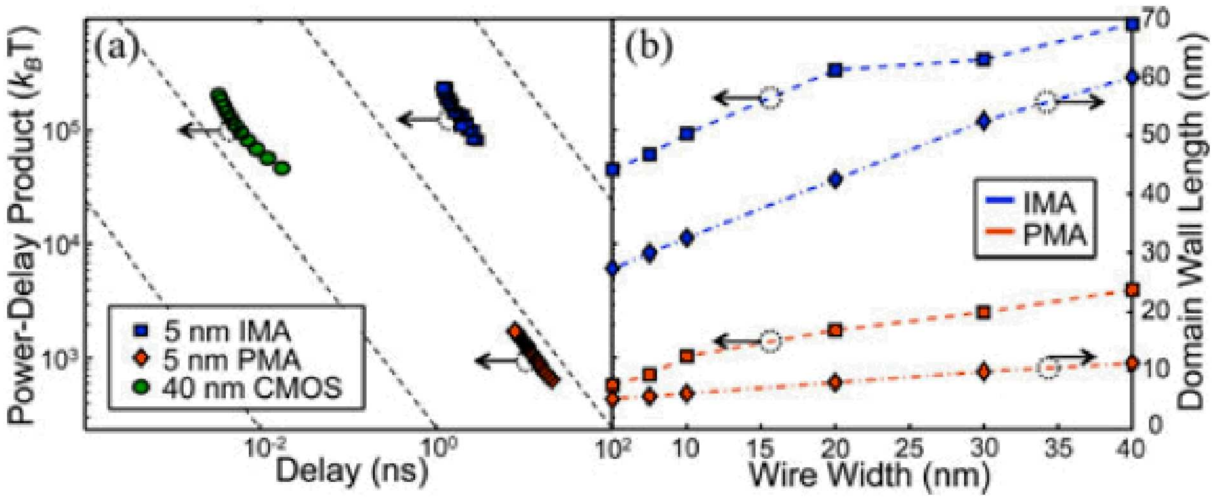
Park et al, IEEE IEDM 2016

Issues with existing spintronic NNs

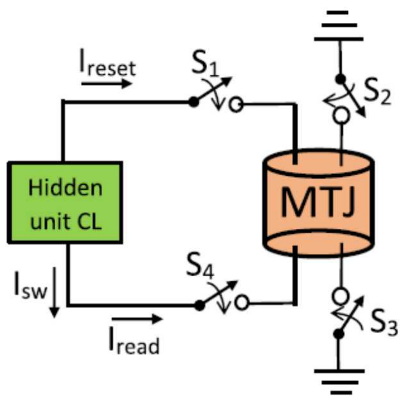
- Several existing spintronic NNs proposals over-use CMOS
  - Since CMOS will also be important at system-level (control blocks, routing…), may lose energy advantages.
- STT/SOT can be current heavy devices. DW synapses/neurons -> path to aJ rather fJ elementary switching costs!!



Sengupta et al, IEEE Biomedical Circuits & Systems 2016



Currivan (Incorvia) et al, IEEE Magnetics Letters 2012



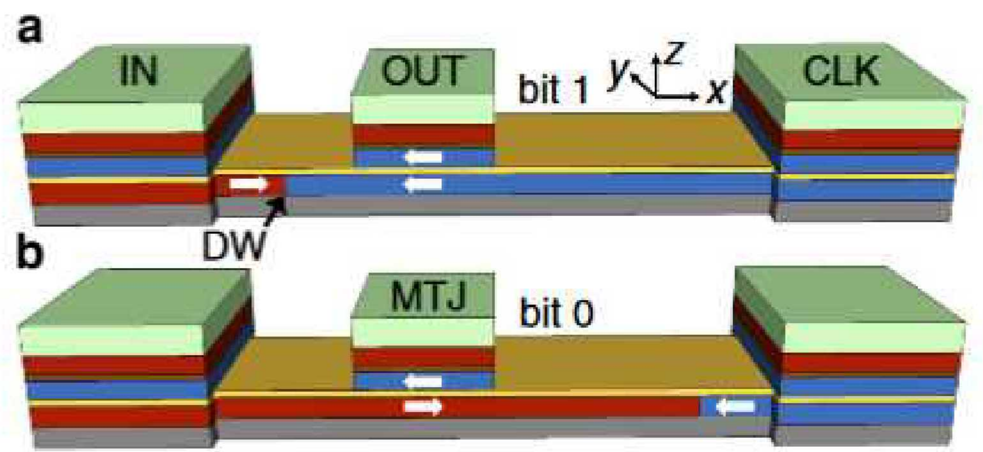
Mondal et al, ACM 2019

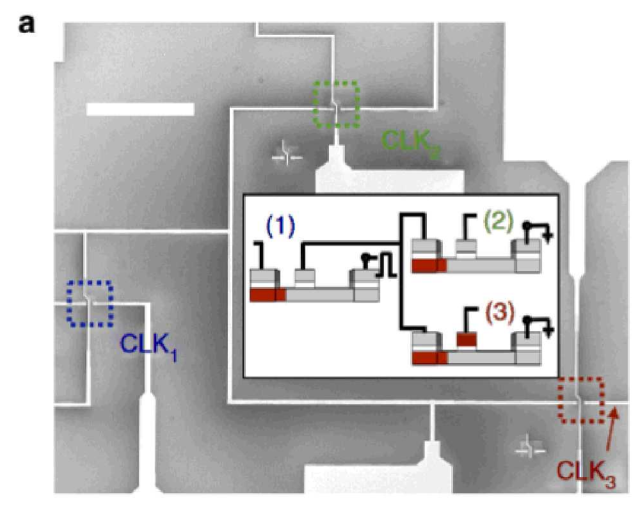
Outline

- Intro and Motivation: Building Accelerators
- Opportunities for Spintronic Neural Networks
- Free Bio-realistic Neuronal effects from DW-MTJs
- Applications: Supervised, Semi-supervised Learning
- Analysis: Coupling Strength + Comparison
- Summary & Future Work

DW-MTJ Basic Device Structure

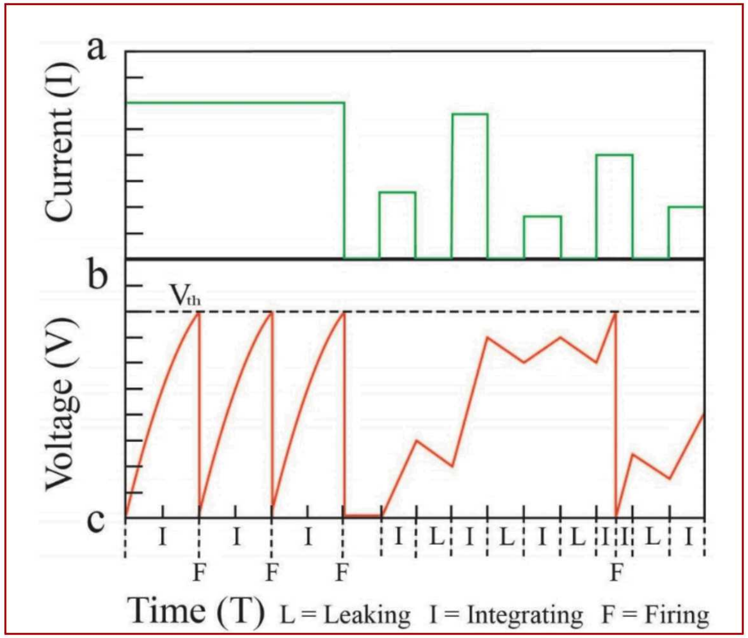
- Domain wall propagates through ferromagnet nanotrack/strip
- MTJ Output at center expresses:
  - Logic 1/high output if DW has moved past Output
  - Logic 0 / low output if DW has not moved past output.
- Pinned antiferromagnet terminal at end of track: for logic/clock
- Devices have been experimentally fabricated and co-integrated.



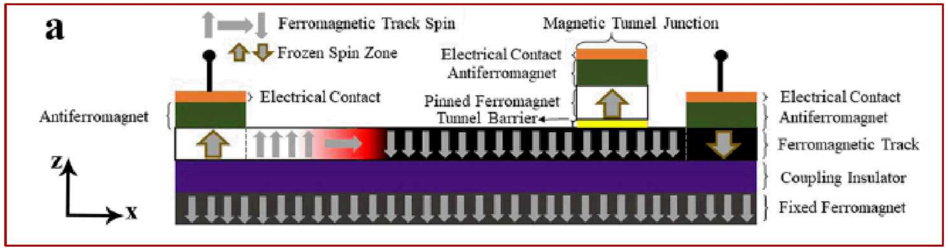


Integration and Leak Behavior

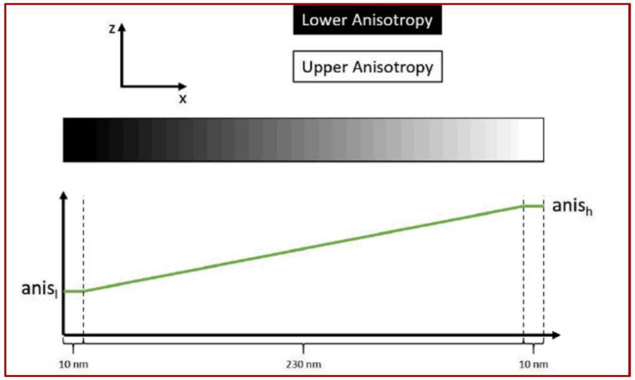
- Integration:
  - DW position integrates applied current and stores it (non-volatile)
- Leaking function:
  - Critical for neuron ‘ reset’/’spike’ function and dynamics (volatile)
  - Different methods for realizing leak: bottom fixed ferromagnet, trapezoidal shape, anisotropy gradient



N. Hassan, J.Appl Phys 2018

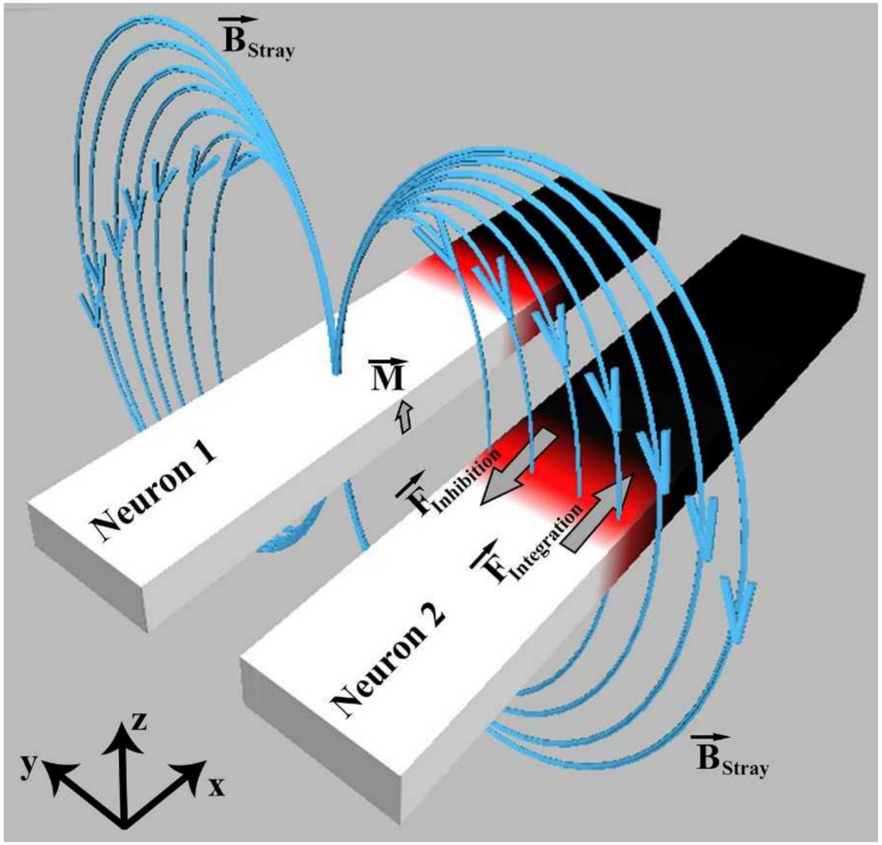


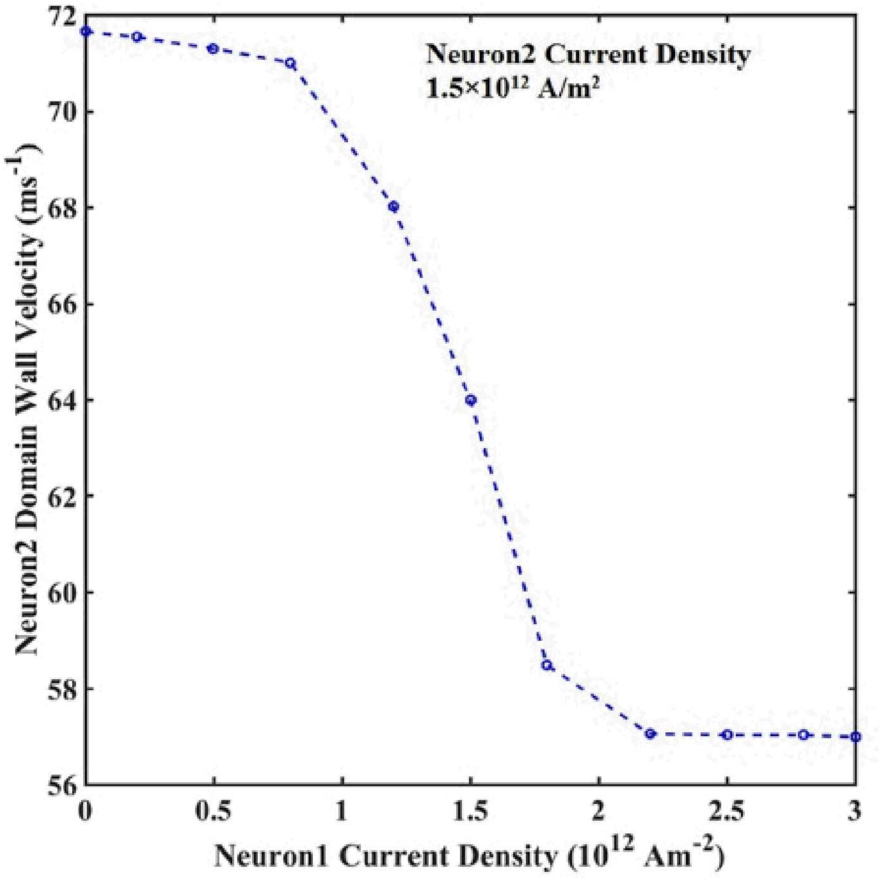
N. Hassan, J.Appl Phys 2018



W. Brigner, JxCDC 2019

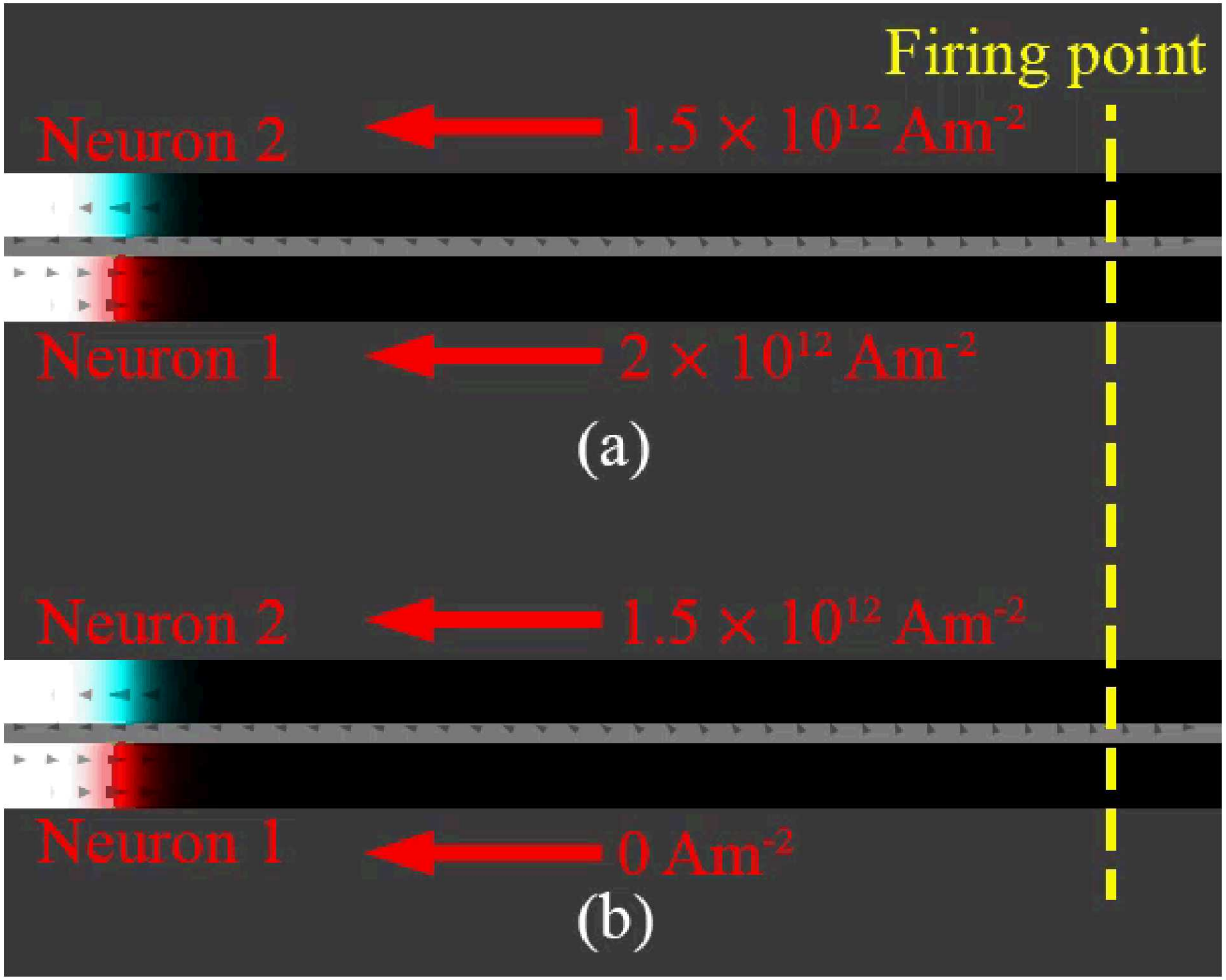
Lateral Inhibition between DW-MTJs





N. Hassan*, X. Hu*, L. Jiang-Wei, W. H. Brigner, O. G. Akinola, F. Garcia-Sanchez, M. Pasquale, C. H. Bennett, J. A. C. Incorvia, J. S. Friedman, *Journal of Applied Physics*, 2018

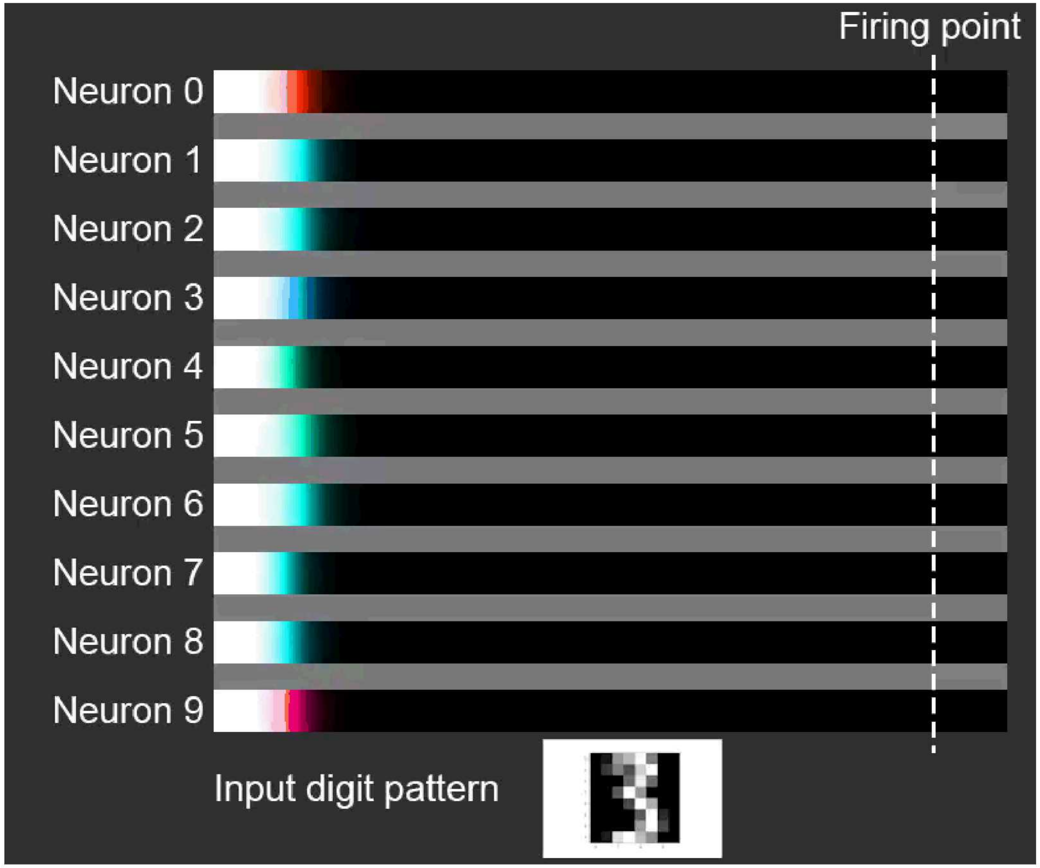
Lateral Inhibition: Demonstration



N. Hassan*, X. Hu*, L. Jiang-Wei, W. H. Brigner, O. G. Akinola, F. Garcia-Sanchez, M. Pasquale, C. H. Bennett, J. A. C. Incorvia, J. S. Friedman, *Journal of Applied Physics*, 2018

Application of LIF DW-MTJs

- Max-out operation was implemented in a perceptron (1 layer NN)
- Weights were pre-written before testing
- 94% success rate
- Inference works!
  - Very fast (<1us for entire test set)
  - Very low energy



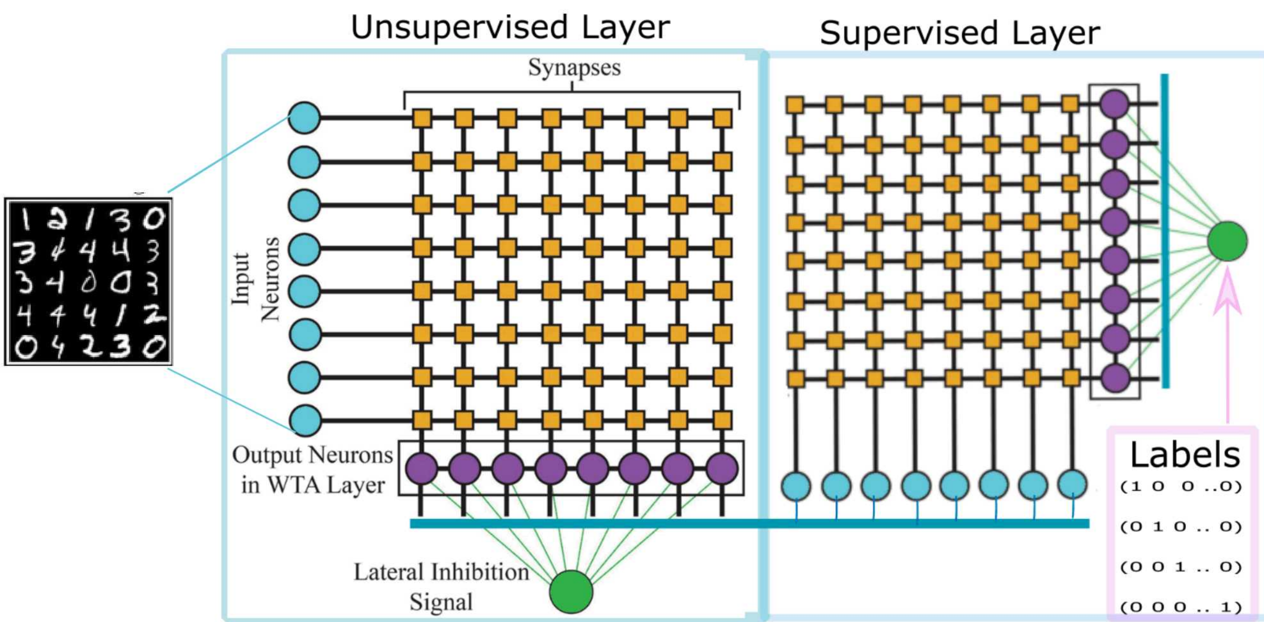
N. Hassan*, X. Hu*, L. Jiang-Wei, W. H. Brigner, O. G. Akinola, F. Garcia-Sanchez, M. Pasquale, C. H. Bennett, J. A. C. Incorvia, J. S. Friedman, *Journal of Applied Physics*, 2018

Outline

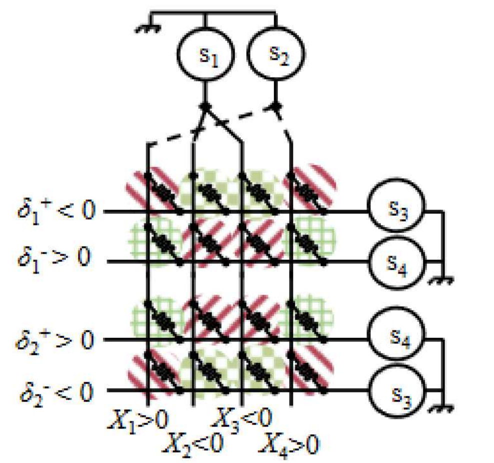
- Intro and Motivation: Building Accelerators
- Opportunities for Spintronic Neural Networks
- Free Bio-realistic Neuronal effects from DW-MTJs
- Applications: Supervised, Semi-supervised Learning
- Analysis: Coupling Strength + Comparison
- Summary & Future Work

Proposed DW-MTJ Spintronic NN I

- Previous work showed written weights (e.g. inference only) performance.
  - Not online learning!
- Current work- dual online learning used to recognize MNIST database:
  - A first NN layer uses DW-MTJ devices for unsupervised clustering: k-WTA Algorithm
  - A second NN layer uses DW-MTJ devices for max-out operation used in the supervised learning process (same as previous).
  - Weight updates: Widrow-Hoff (same as 'delta' rule used in backprop).



$$\Delta W_{i,k} = \Delta G sign(X_i(T_k - O_k)),$$

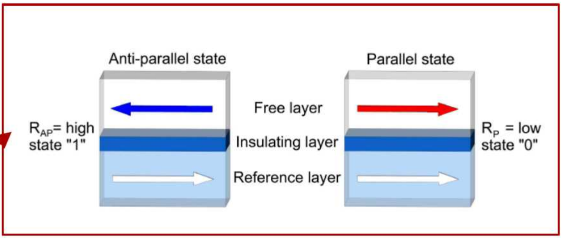


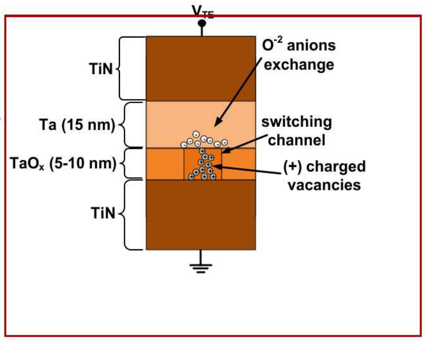
Lin et al, Scientific Reports 2016

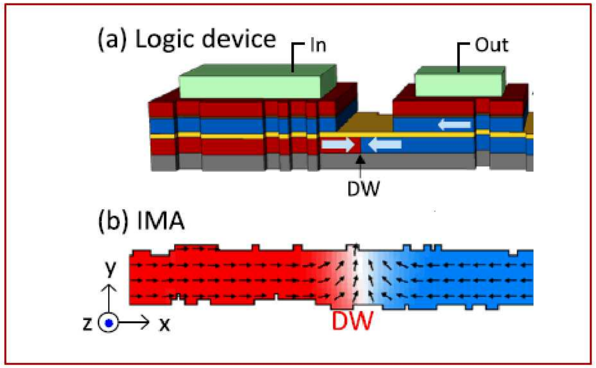
Zamandioost et al, IEEE WISP 2015

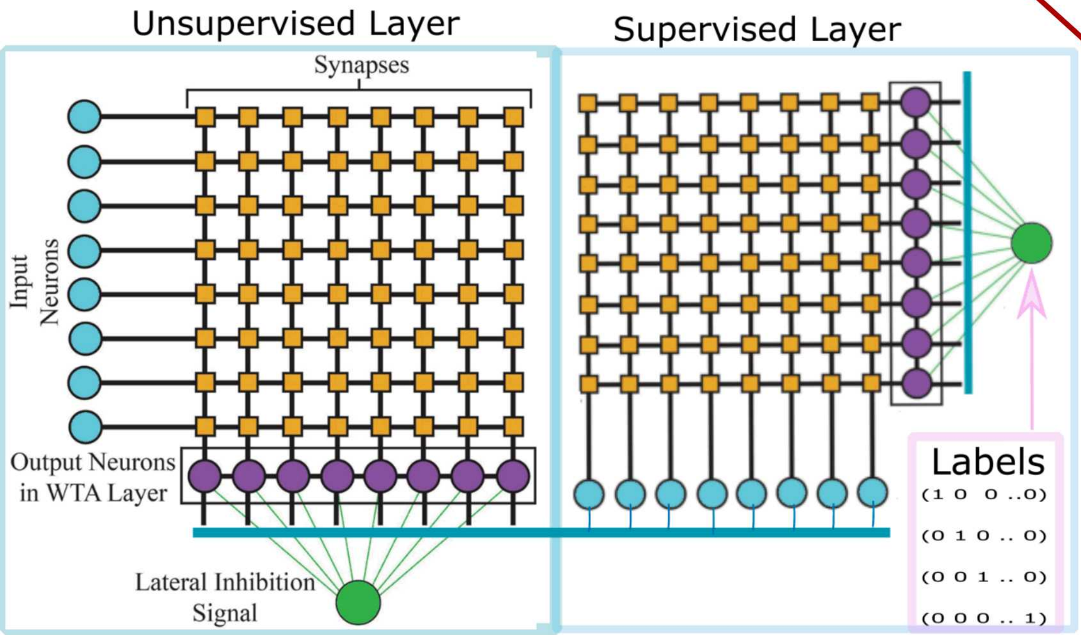
Proposed DW-MTJ Spintronic NN II

- Neurons are always DW-MTJ devices.
- Synapses can be :
  - 2 terminal magnetic synapse (STT-MTJ): Binary
  - 2 terminal resistive RAM (ReRAM): Binary or Analog
  - 3 terminal DW-MTJ: Binary or Analog





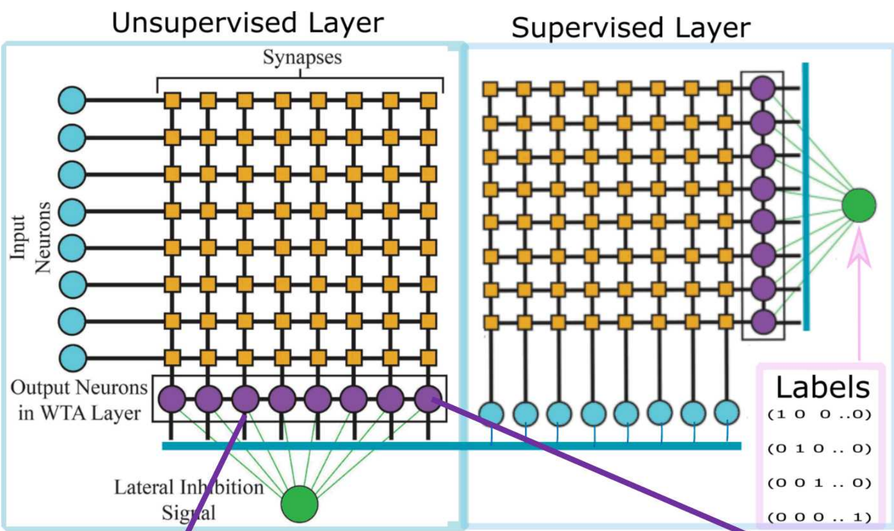


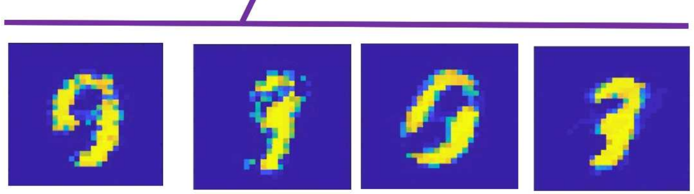


Proposed DW-MTJ Spintronic NN III

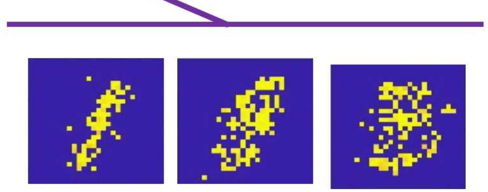
- Unsupervised first-layer variations considered:
  - Random weights (no plasticity operation – control case)
  - Binary STDP (plasticity updates are constant/sign based)
  - Analog STDP (plasticity updates are scaled/numeric)

Policy	Result (Size: 200 DW-MTJs)
Random	78% test set correct
B-STDP	88% “”
A-STDP	92% “”

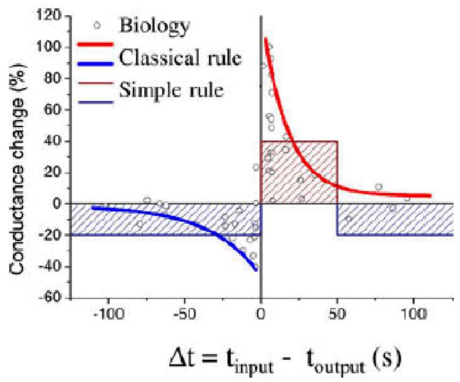




Analog STDP: reconstructed filters



Binary STDP: filters



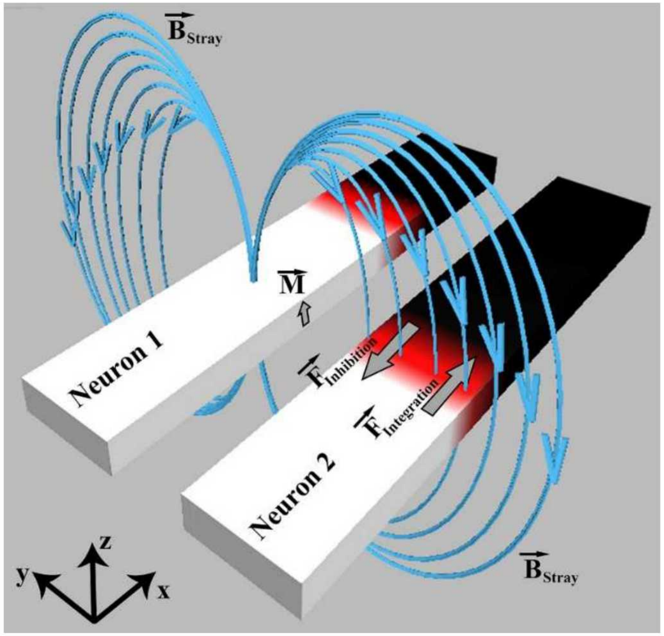
Querlioz et al, IEEE Transactions Nanotechnology, 2013

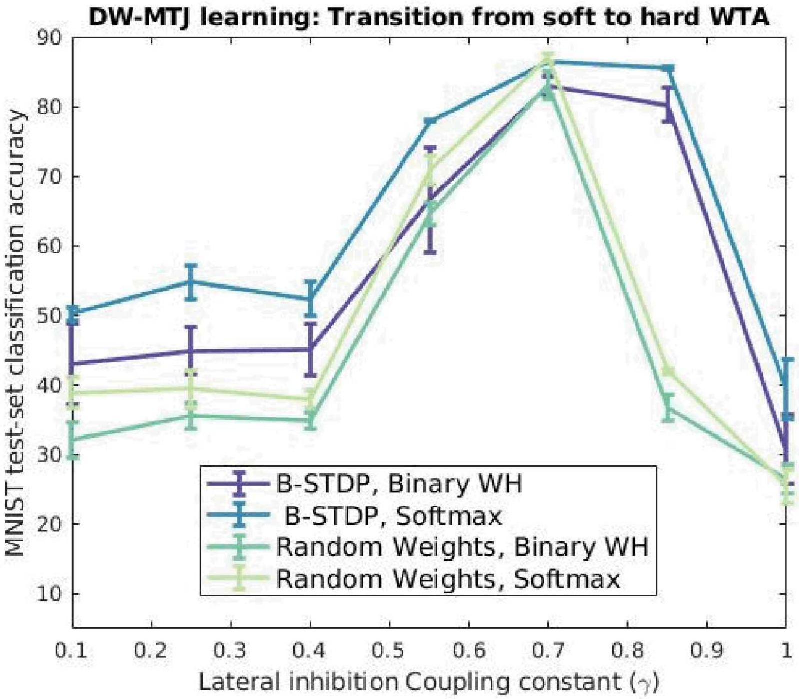
Outline

- **Intro and Motivation: Building Accelerators**
- **Opportunities for Spintronic Neural Networks**
- **Free Bio-realistic Neuronal effects from DW-MTJs**
- **Applications: Supervised, Semi-supervised Learning**
- **Analysis: Coupling Strength + Comparison**
- **Summary & Future Work**

Importance of inhibition schemes

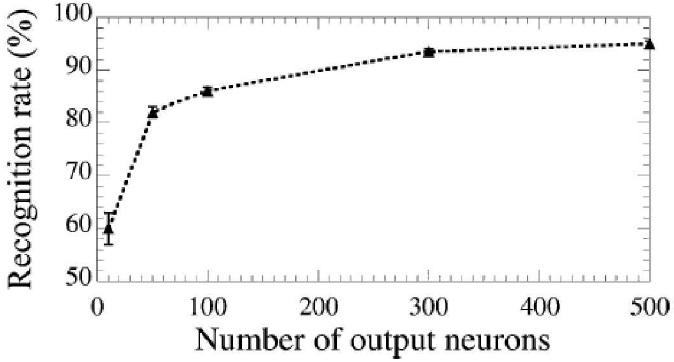
- The lateral coupling between DW-MTJs is important
  - Too few neurons fire (too many inhibited : representations too sparse
  - Too many neurons fire (too few inhibited): representations too noisy
- At early stage of evaluating if we can fabricate wires at correct dimensions + spacings
- More elaborate physics model being built to inform NN simulations



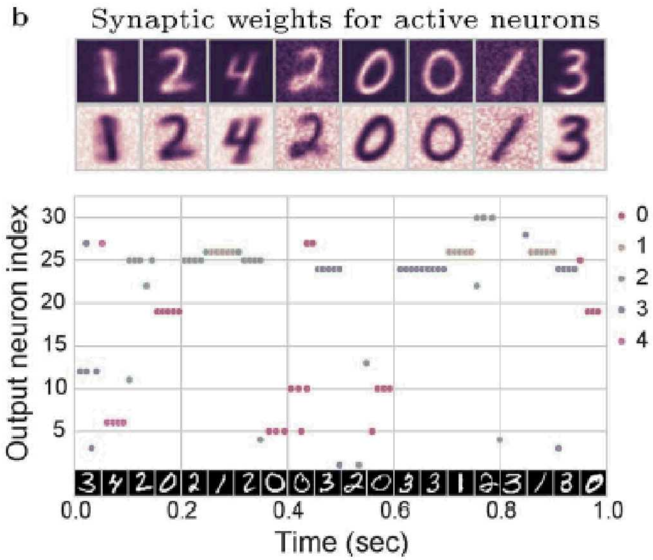


Comparisons to other LIF learning systems

- STDP results comparable to best reported results (93%) [1],[2] for NN combining supervised and unsupervised approaches
  - However, we use a more realistic + energy-efficient read-out method
- STDP results are superior to those obtained using memristor +ReRAM LIF emulator neurons (78%) [3]
  - LIF circuit also had a high level of complexity



Source: [1]



Source: [3]

[1] Querlioz et al, IEEE Transactions Nanotechnology 2013
[2] Bennett et al, IEEE IJCNN, 2016
[3] Al-Shedivat et al, IEEE Jetcas, 2015

Outline

- **Intro and Motivation: Building Accelerators**
- **Opportunities for Spintronic Neural Networks**
- **Free Bio-realistic Neuronal effects from DW-MTJs**
- **Applications: Supervised, Semi-supervised Learning**
- **Analysis: Coupling Strength + Comparison**
- **Summary & Future Work**

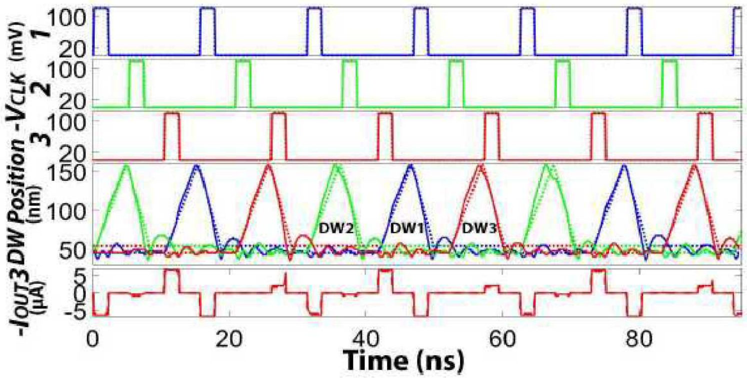
Summary

Sandia National Laboratories

- DW-MTJ devices are a promising nanodevice to implement analog accelerator/NN systems
  - Not susceptible to classical issues with ReRAM synapses + CMOS Neurons
  - Ultra-low energy budgets and rich physics allows LIF behavior: CMOS-free
- Early simulations suggest LIF behavior and plasticity behaviors (STDP) lead to promising generalization (test-set accuracy) on real tasks ☺

Next Steps

- Better analyze upper boundaries of NN performance and compare to MLP, CNN
  - Can benchmark results against Cross-Sim software package
  - In principle should be able to stack/combine unsupervised layers
- Obtain accurate energy estimates using DW-MTJ SPICE model [1]
- Integrate physics-rich estimates of lateral inhibition effects



https://cross-sim.sandia.gov

[1] Hu et al, IEEE Transactions Electron Devices, 2019

Thank you!