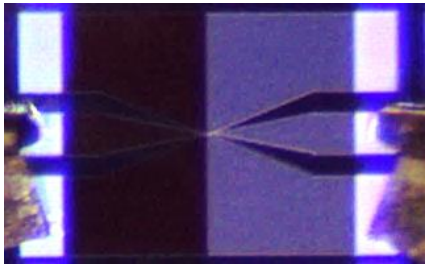
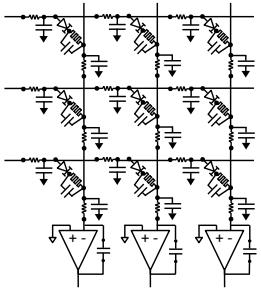
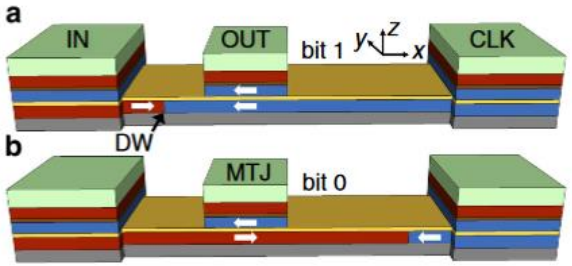
Exceptional service in the national interest

Sandia National Laboratories







Plasticity-Enhanced Domain-Wall MTJ Neural Networks for Energy-Efficient Online Learning

Christopher H. Bennett*,[1] T. Patrick Xiao [1], Can Cui,[2], Naimul Hassan [3], Otitoaleke Akinola [2], Jean Anne C. Incorvia [2], Alvaro Velasquez, Joseph S. Friedman [3], Matthew J. Marinella [1]

*cbennet@sandia.gov

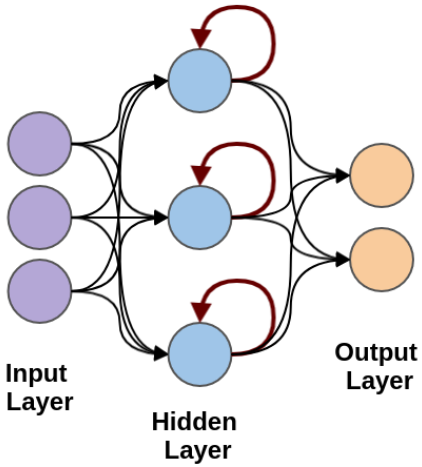
[1] Sandia National Laboratories [2] University of Texas, Austin [3] University of Texas, Dallas

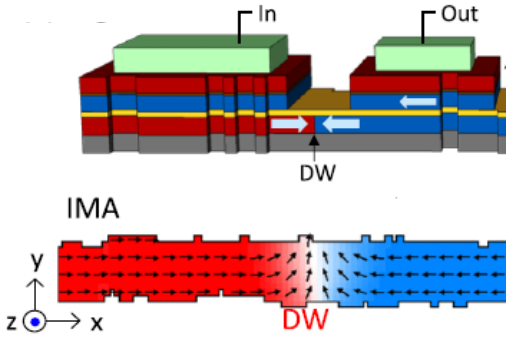
U.S. DEPARTMENT OF ENERGY | National Nuclear Security Administration | LDRD Laboratory Directed Research & Development

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

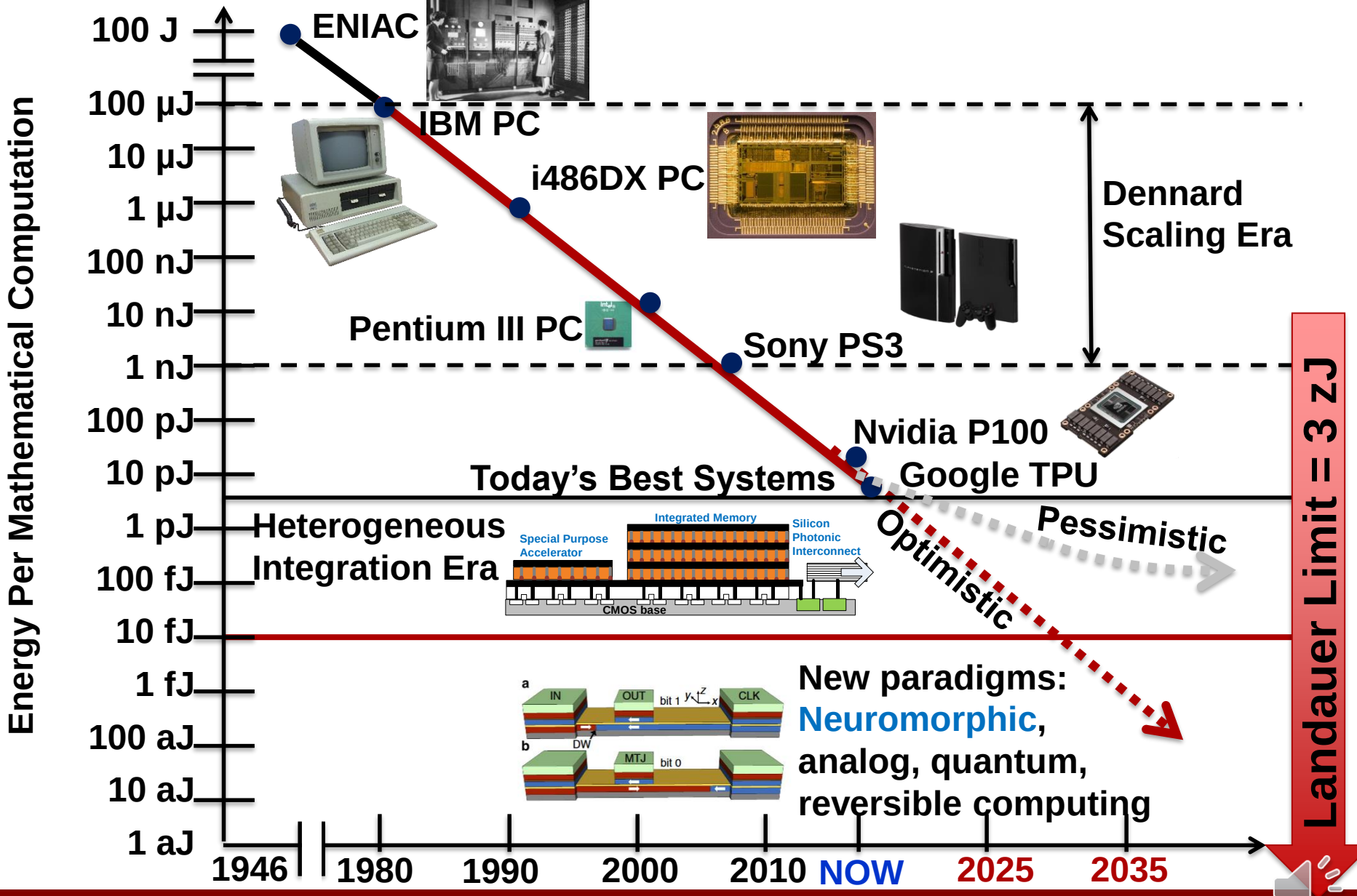
- **Intro and Motivation: Building Spintronic Accelerators**
- **Critical Bio-realistic Neuronal effects with DW-MTJs**
- **Implementation of Semi-supervised Learning**
- **Discussion & Future Work**





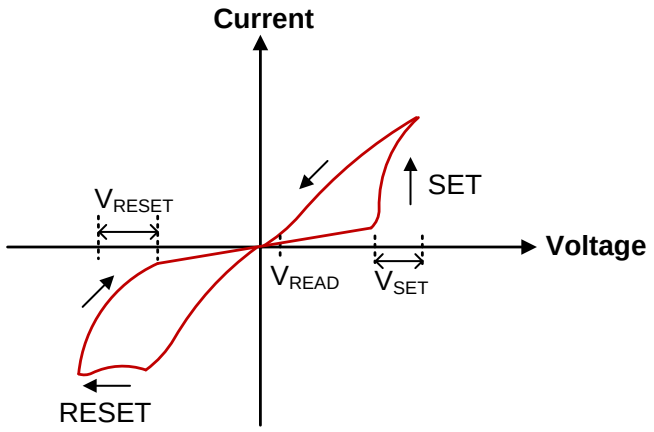
Evolution of Computing Machinery

Sandia National Laboratories



Realizing physical matrix kernels

- 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

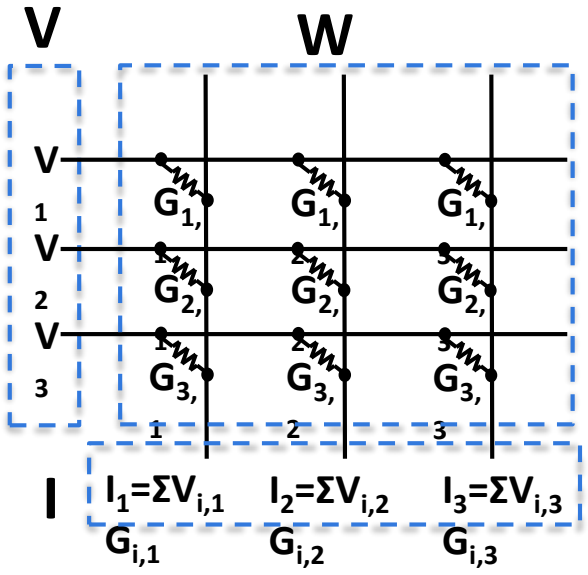


Mathematical

$$V^T W = I$$

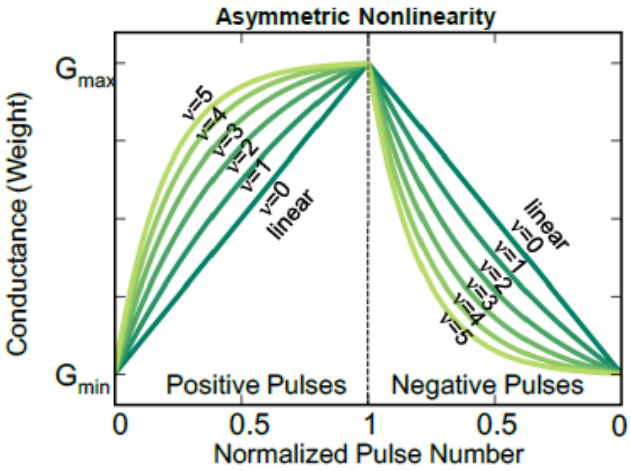
$$\begin{bmatrix} V_1 & V_2 & V_3 \end{bmatrix} \begin{bmatrix} W_{1,1} & W_{1,2} & W_{1,3} \\ W_{2,1} & W_{2,2} & W_{2,3} \\ W_{3,1} & W_{3,2} & W_{3,3} \end{bmatrix} = \begin{bmatrix} I_1=\Sigma V_{i,1}W_{i,1} & I_2=\Sigma V_{i,2}W_{i,2} & I_3=\Sigma V_{i,3}W_{i,3} \end{bmatrix}$$

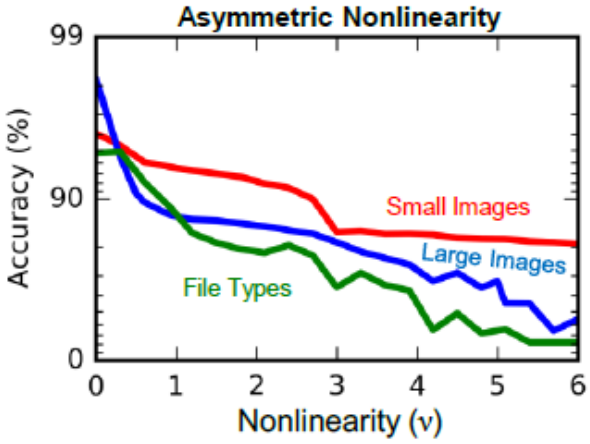
Electrical



Challenges for adaptive analog accelerators

- Emerging ReRAM : far from ideal , floating-point ‘weights’
- 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).*



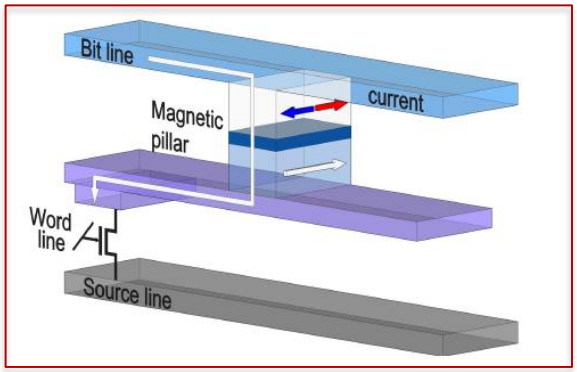


Agarwal et al, IJCNN 2016

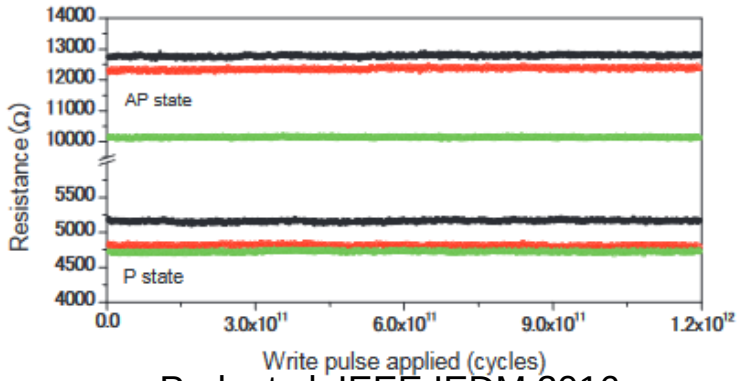
Major opportunity:

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 often not intrinsic (but can be designed)
- Additional Advantages:
  - Extreme endurance (important for online learning + inference)
  - Extremely fast (<5 ns read and write)
  - Low energy footprint : typically <1V programming, <50ns programming .
  - Extreme compactness and CMOS-compatible 1T1R array scaling (BEOL integration)



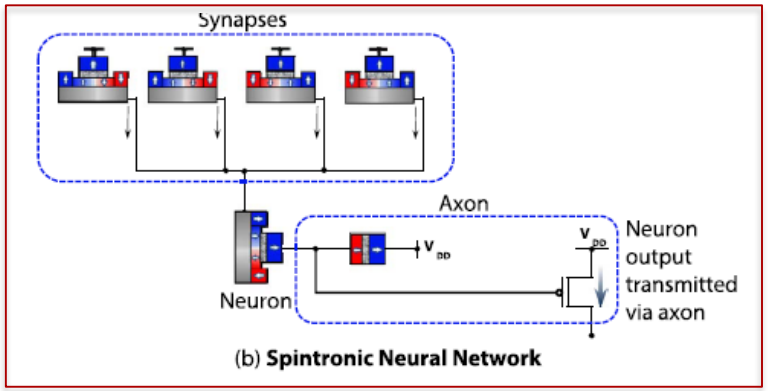
Makarov et al , IOP 2016



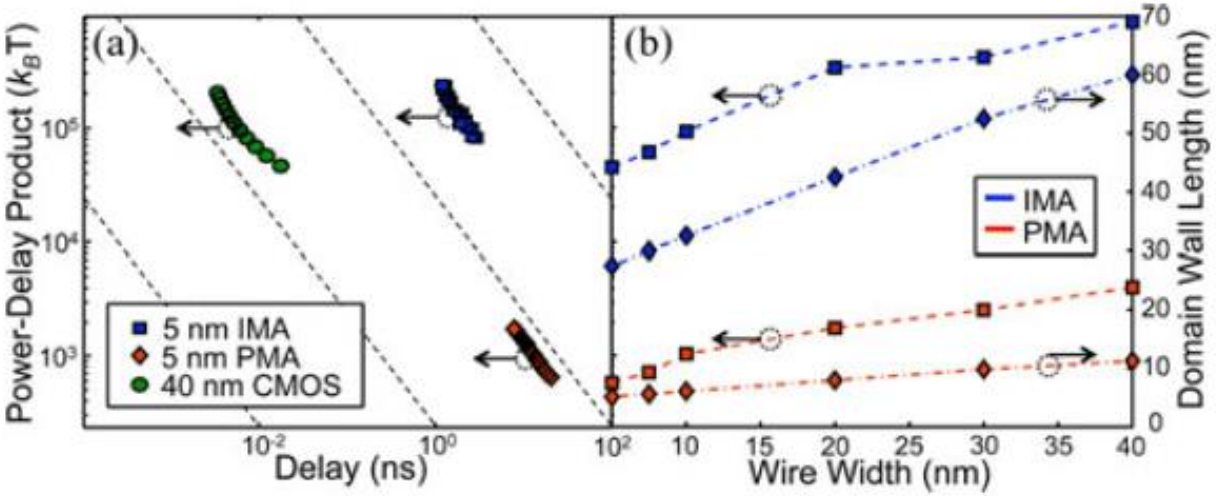
Park et al, IEEE IEDM 2016

Issues with existing spintronic NNs

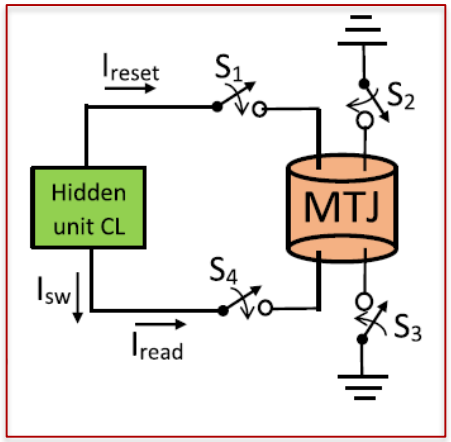
- Several existing spintronic NNs proposals over-use CMOS
  - Since CMOS will also be important at system-level (control blocks, routing…), may lose energy advantages.
- STT/SOT switches can be current heavy devices.
  - DW synapses/neurons -> path to *aJ* rather than *fJ* elementary switching costs!!



Sengupta et al, IEEE Biomedical Circuits & Systems 2016



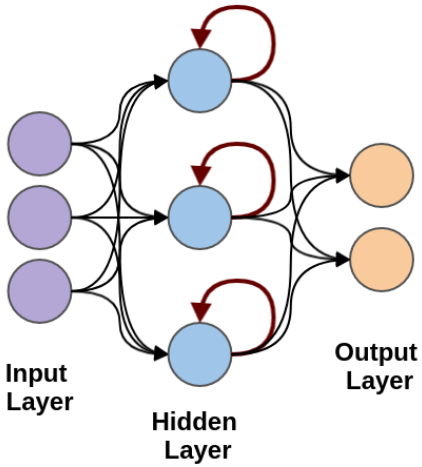
Currivan (Incorvia) et al, IEEE Magnetics Letters 2012

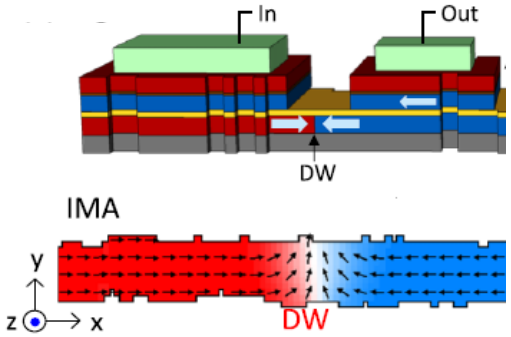


Mondal et al, ACM 2019

Outline

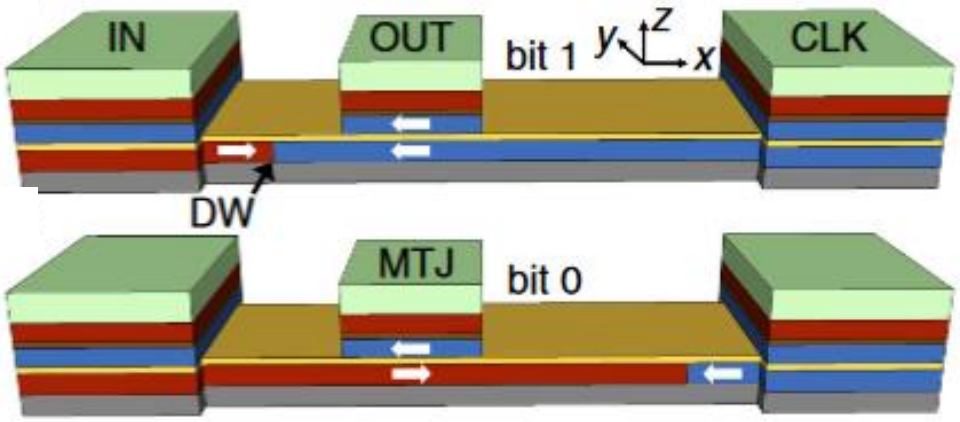
- **Intro and Motivation: Building Spintronic Accelerators**
- **Critical Bio-realistic Neuronal effects with DW-MTJs**
- **Implementation of Semi-supervised Learning**
- **Discussion & Future Work**

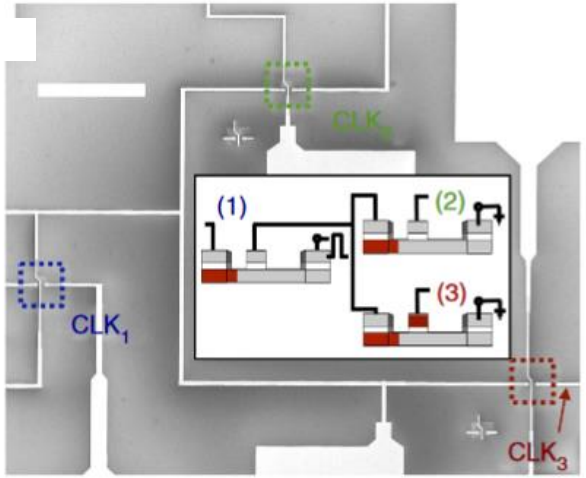




DW-MTJ Basic Device Structure

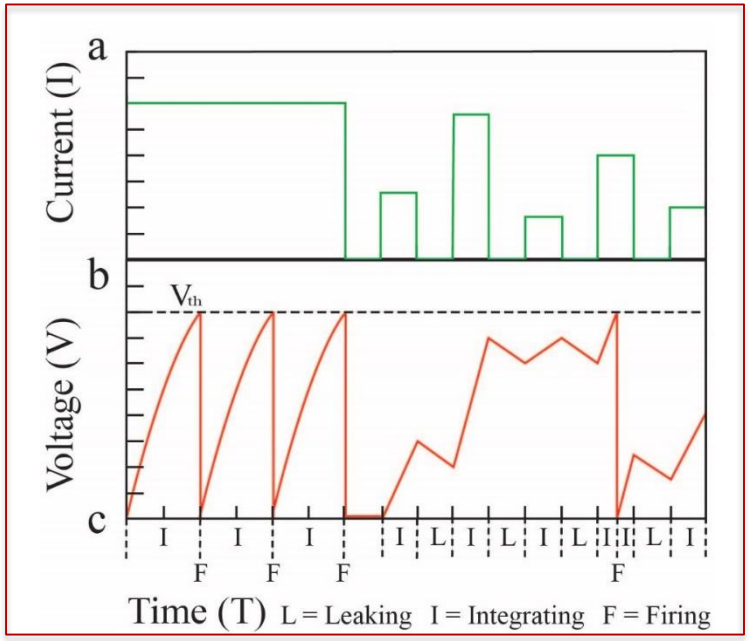
- Domain wall propagates through ferromagnet nanotrack/strip
- MTJ Output at center expresses:
  - Logic 0/low output if DW has moved past Output
  - Logic 1 / high output if DW has not moved past output.
- Pinned antiferromagnet terminal at end of track: for logic/clock
- Devices have been experimentally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



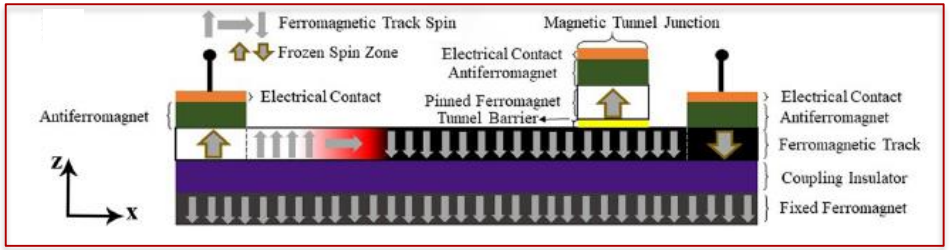


Integration and Leak Behavior

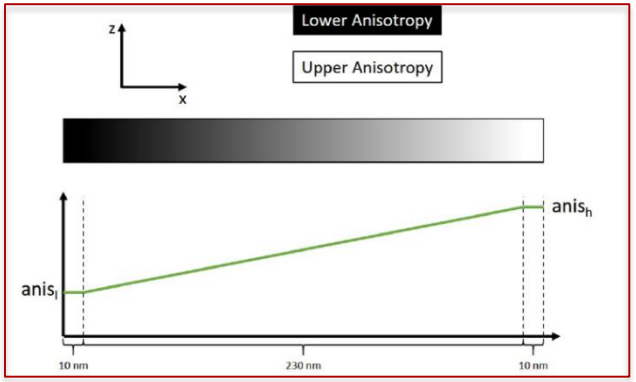
- Integration:
  - DW position integrates applied current and stores it (non-volatile)
- Leaking function:
  - Critical for neuron ‘ reset’/’spike’ function and dynamics (volatile)
  - Different methods for realizing leak: bottom fixed ferromagnet, trapezoidal shape, anisotropy gradient



N. Hassan, J.Appl Phys 2018

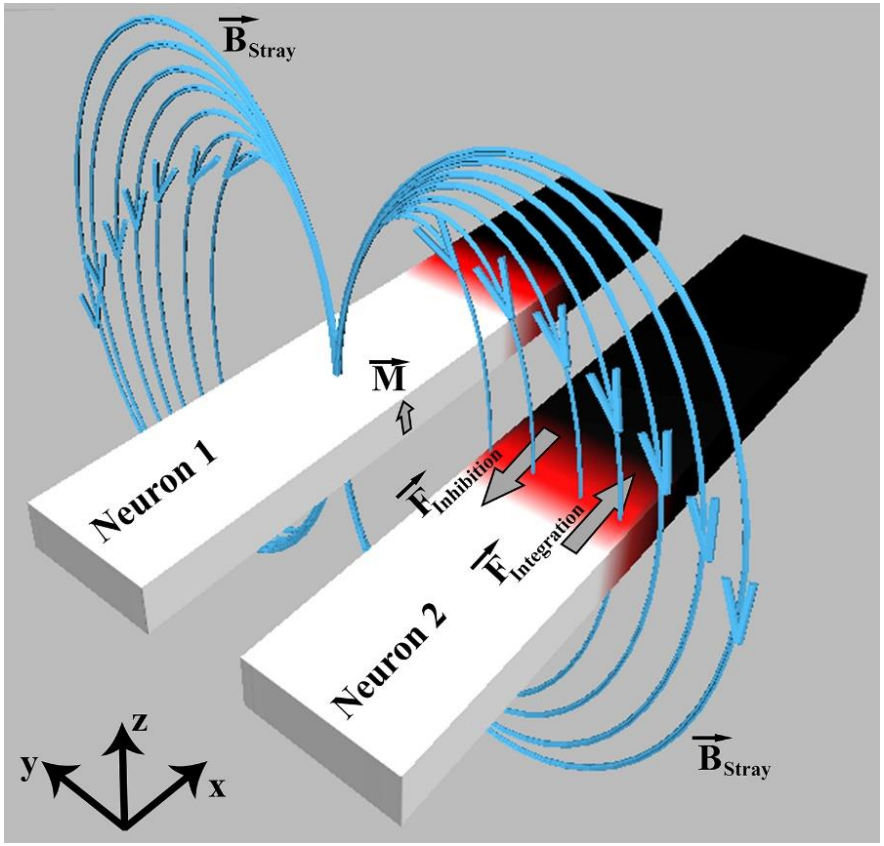


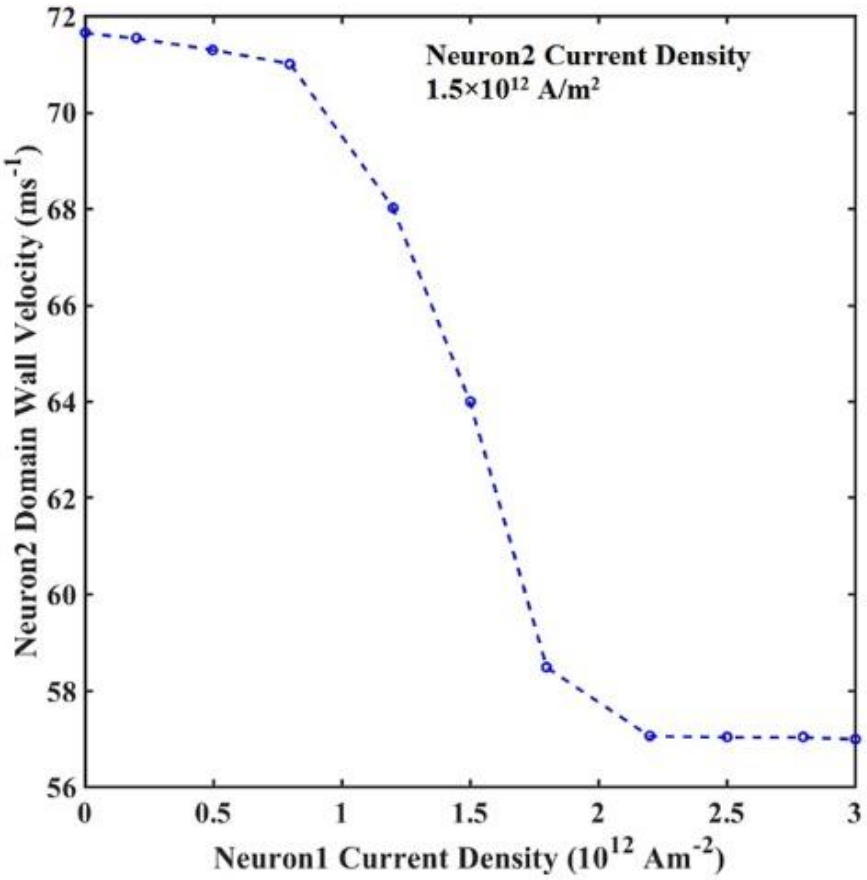
N. Hassan, J.Appl Phys 2018



W. Brigner, JxCDC 2019

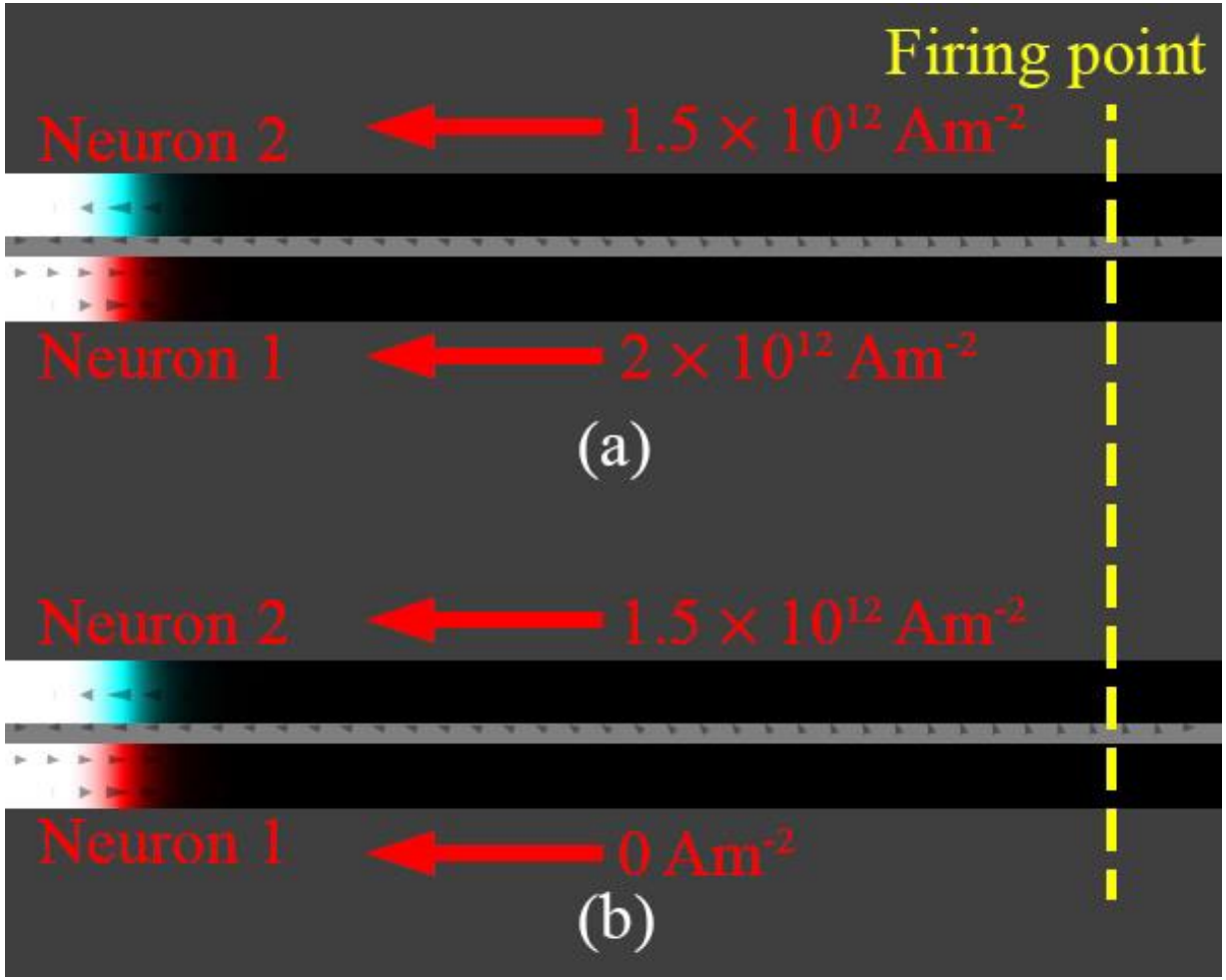
Lateral Inhibition between DW-MTJs





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

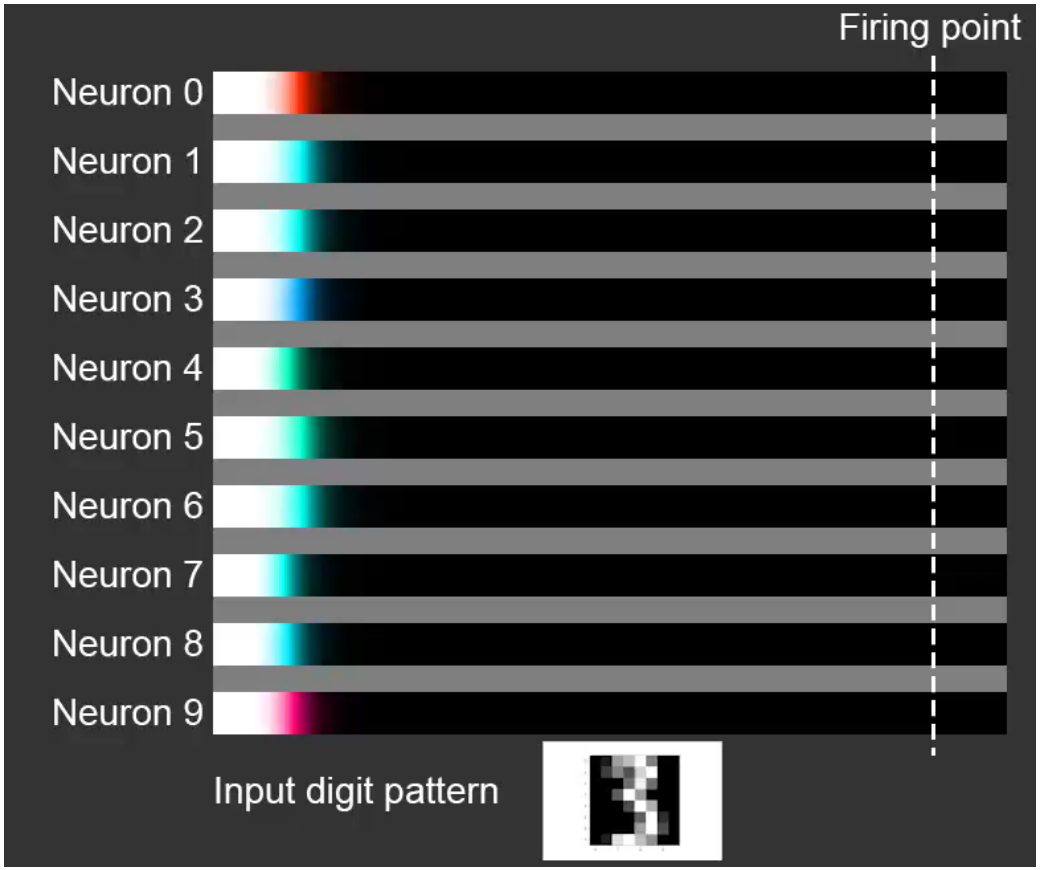
Lateral Inhibition: Demonstration



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

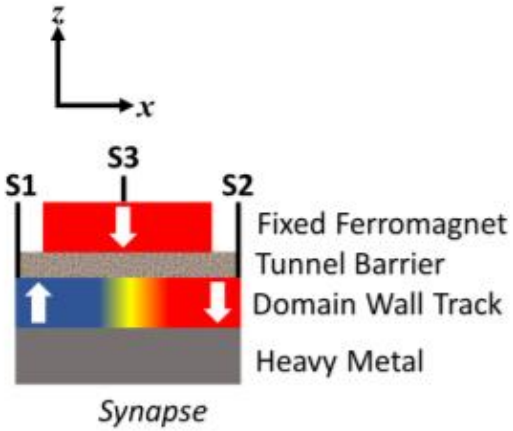
- Max-out operation was implemented in a perceptron (1 layer NN)
- Weights were pre-written before testing
- 94% success rate
- Inference works!
  - Very fast (<1us for entire test set)
  - Very low energy



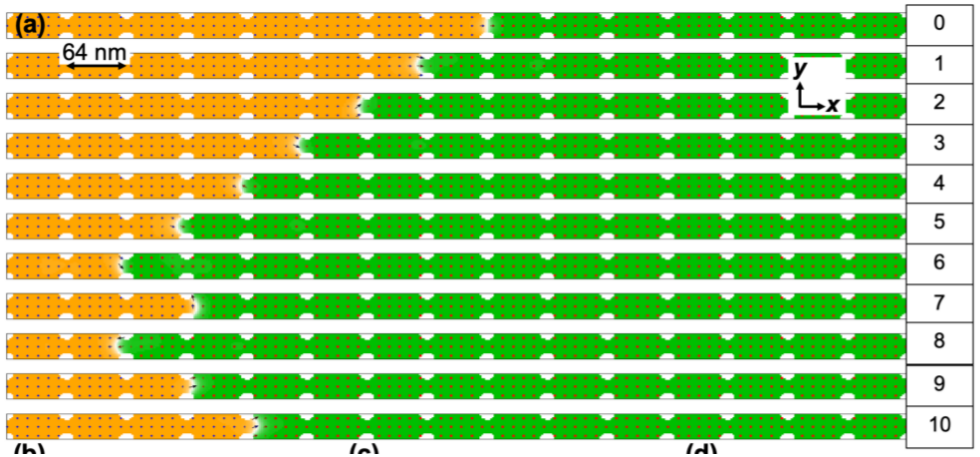
N. Hassan*, X. Hu*, L. Jiang-Wei, W. H. Brigner, O. G. Akinola, F. Garcia-Sanchez, M. Pasquale, C. H. Bennett, J. A. C. Incorvia, J. S. Friedman, *Journal of Applied Physics*, 2018

DW-MTJ Synapses

- DW-MTJ devices can be set to various levels of conductance as a function of the DW position under the fixed ferromagnet
  - In this configuration, output 3 (S3) is fabricated to be extremely long so as to encode maximal states
- For repeatable learning that adds resilience to bit-errors, notches can be added to DW-Synapse tracks



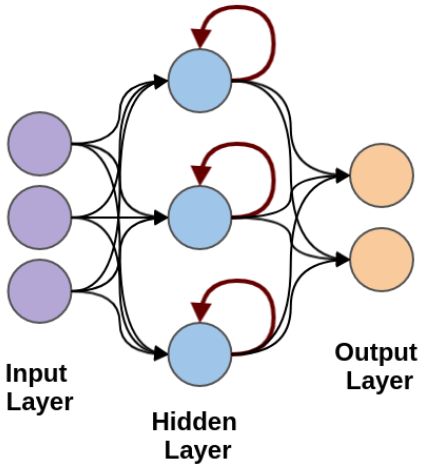
Source: https://arxiv.org/pdf/2003.11120.pdf

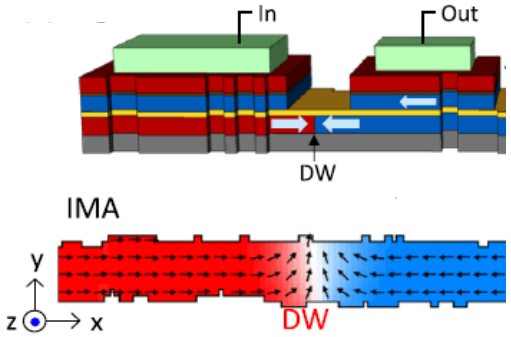


Source: https://iopscience.iop.org/article/10.1088/1361-6463/ab4157/pdf

Outline

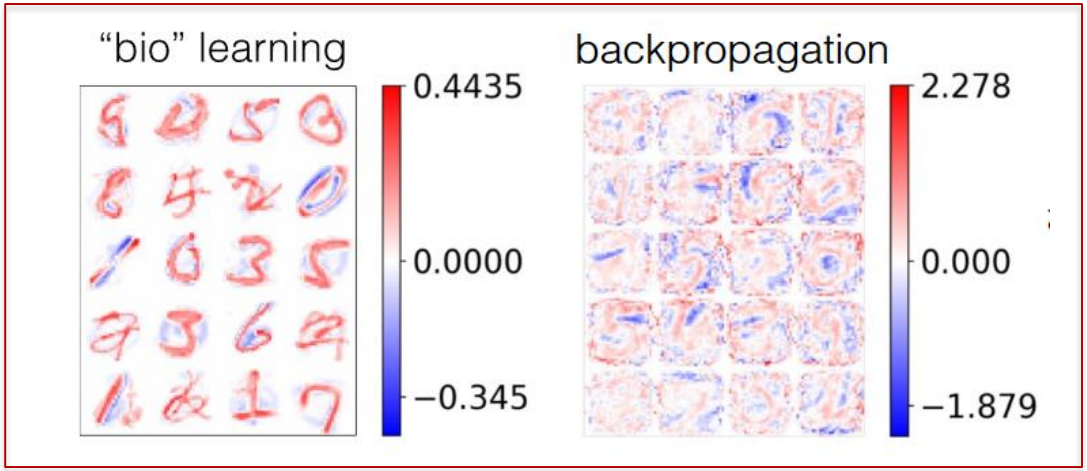
- **Intro and Motivation: Building Spintronic Accelerators**
- **Critical Bio-realistic Neuronal effects with DW-MTJs**
- **Implementation of Semi-supervised Learning**
- **Discussion & Future Work**



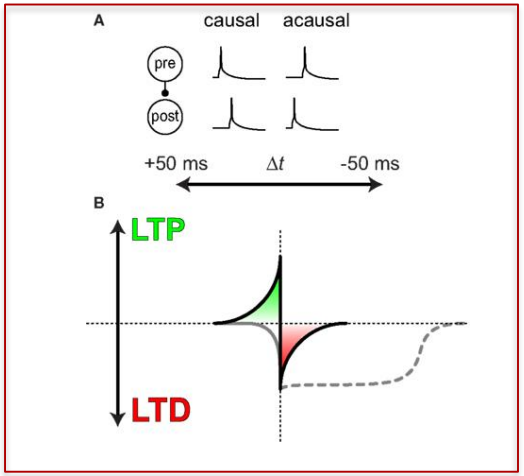


Why semi-supervised learning?

- Fully backpropagation-based learning require high CMOS-overhead and is memory expensive
  - Difficult to implement in hardware- even with emerging (spintronic) memory
- The brain uses an operation called spike-timing-plasticity (STDP) to encode local + simpler representations
- Significant interest in combining the approaches:
  - Less training examples needed -> less energy overhead



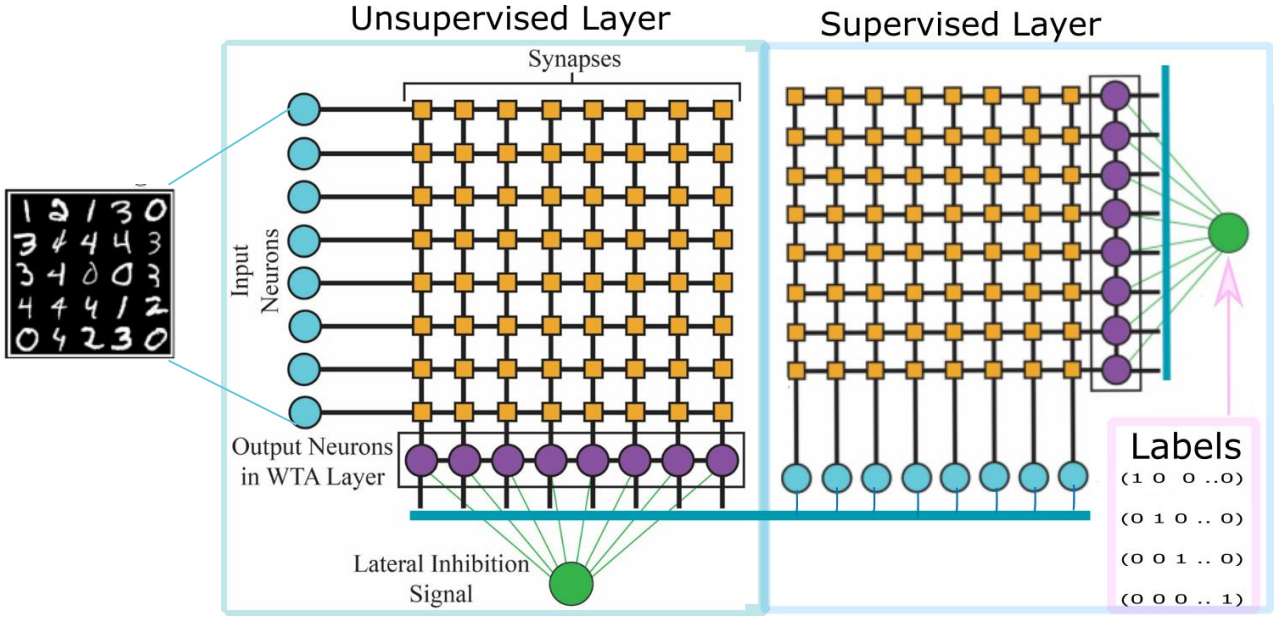
Krotov & Hopfield, PNAS, 2019



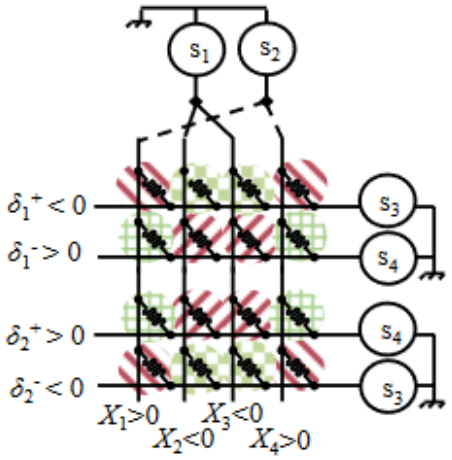
Makram et al, Frontiers, 2011

Semi-supervised learning: implementation I

- Dual online learning processes used to perform classification tasks on standard machine learning tasks
  - Phase 1 : *unsupervised clustering*: using k-WTA Algorithm
    - STDP weight updates
  - *Phase 2: Read out regression*
    - 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

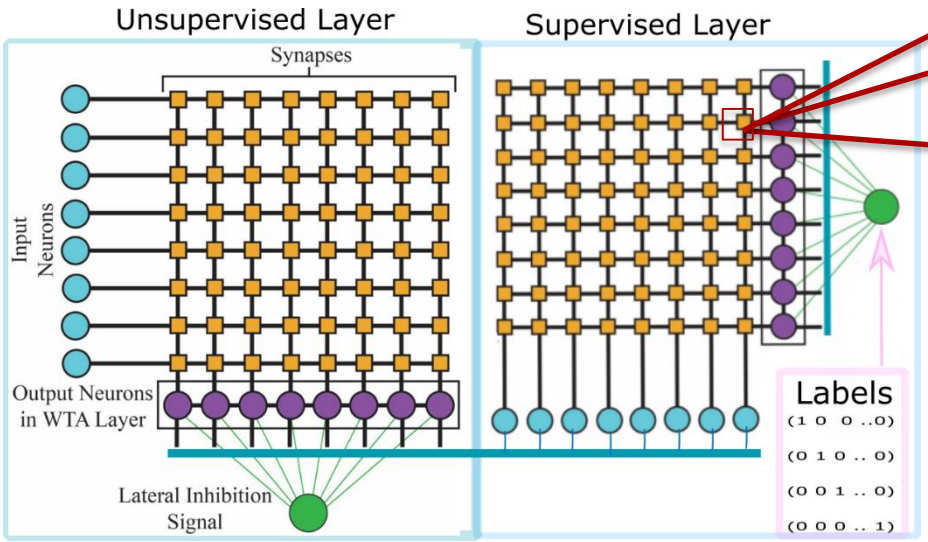
Zamandioost et al, IEEE WISP 2015

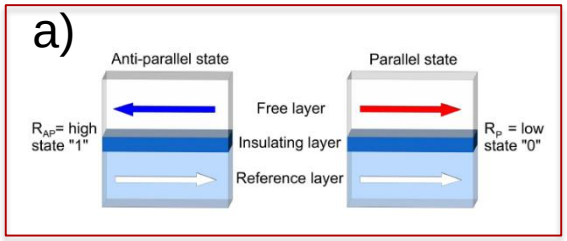
Source: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11090/110903I/Semi-supervised-learning-and-inference-in-domain-wall-magnetic-tunnel/10.1117/12.2530308.pdf

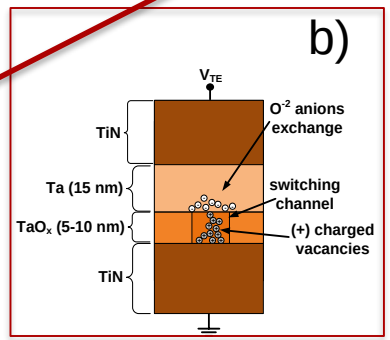
Semi-supervised learning: implementation II

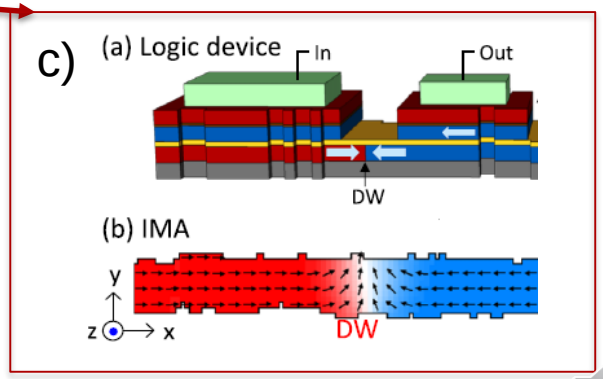
Sandia National Laboratories

- Neurons are always DW-MTJ devices.
  - First layer: leaky-integrate fire capability
  - Second layer: max-out
- Synapses can be :
  - A) 2 terminal magnetic synapse (STT-MTJ): Binary
  - B) 2 terminal resistive RAM (ReRAM): Binary or Analog
  - C) 3 terminal DW-MTJ: Binary or Analog
- Current work assumes 3T Notched synapse for energy #s



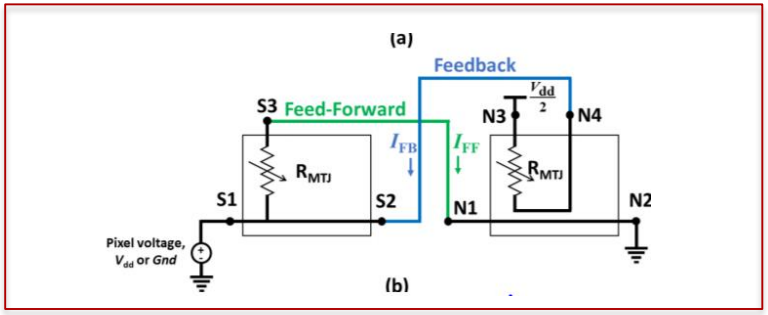




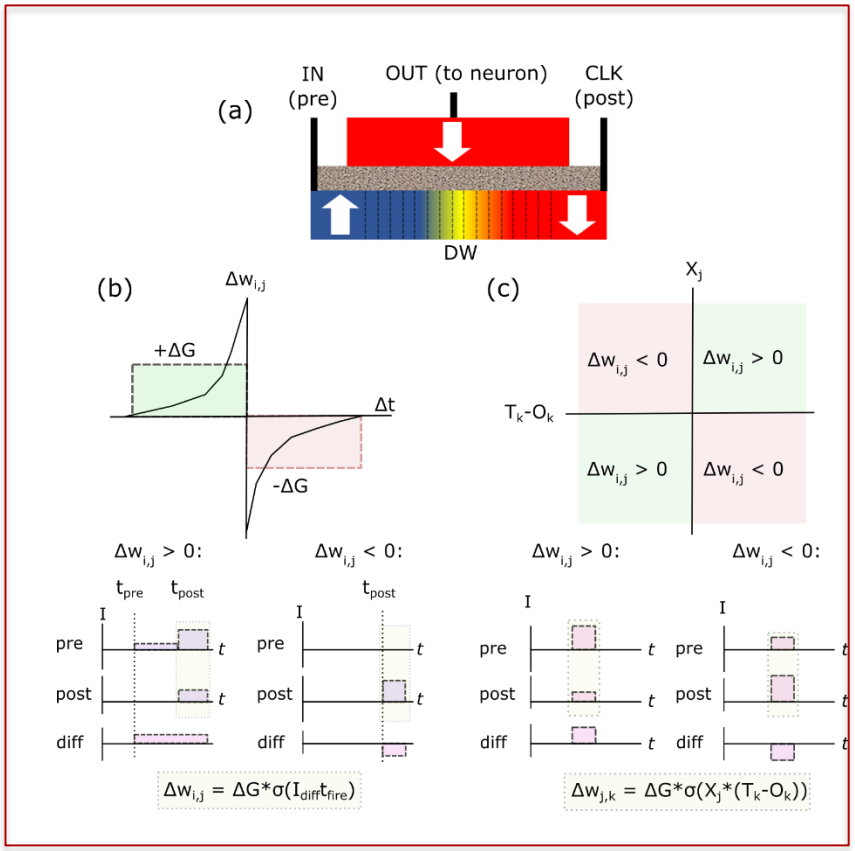


Detail on learning rule and synaptic implementation

- Current difference between pre and post synaptic ports determine plasticity event
- **I pre > I post :** *Weights increase*
  - *Correlation/Hebbian*
- **I pre < I post:** *Weights decrease*
  - *Anti-correlation/ Anti-Hebbian*
- The operation is realized via a special wiring which allows for feedback to occur.
- In the second layer , weights additionally move relative to the teacher signal
  - "Four quadrant learning rule " -- > *analogous to stochastic grad descent*



Source: https://arxiv.org/abs/2003.02357



Source: https://arxiv.org/abs/2003.02357

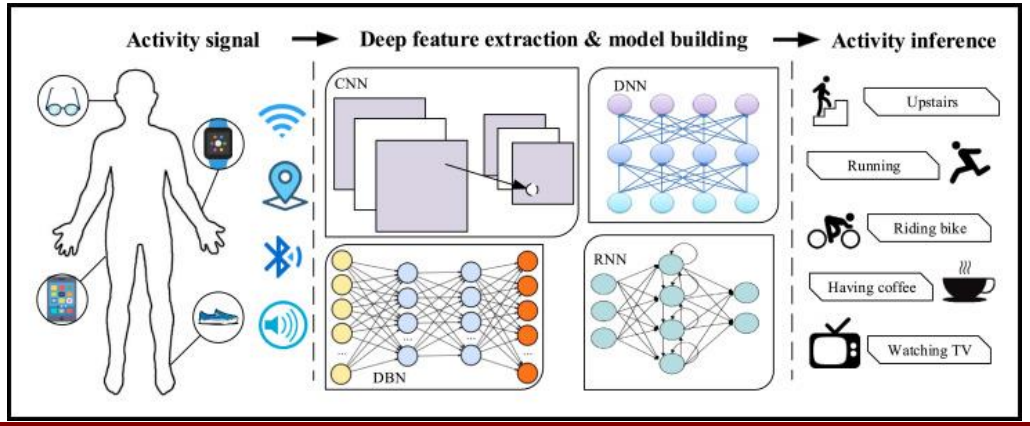
Battery of machine learning tasks

- In increasing order of difficulty:
  - Human Activity Recognition: sensor input
    - 21k training, 2.5k test
    - MLP typical result: 98%+
  - MNIST: small images
    - 60k training, 10k test
    - MLP typical result: 96%+
  - F-MNIST: small images
    - 60k training, 10k test
    - MLP typical result: 83%+

MNIST Task

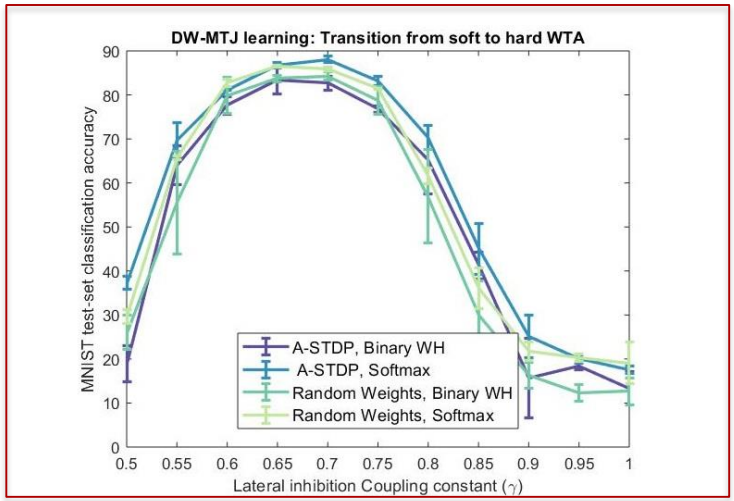
Fashion-MNIST Task

HAR Task

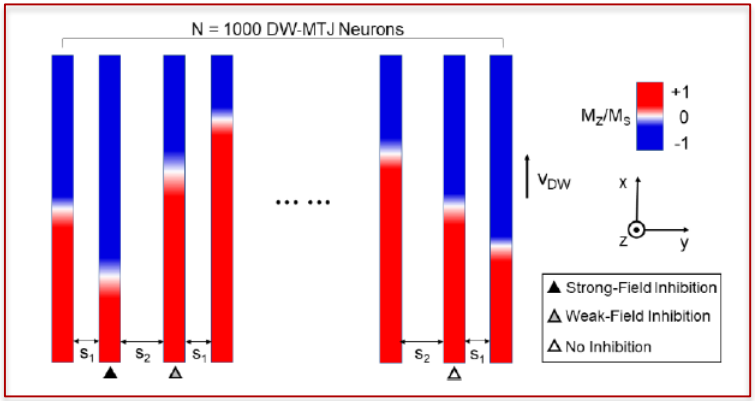


Role of lateral interaction parameter

- Interaction parameter gamma γ defines the interaction of neighboring DW-MTJ competing neurons/nanotracks
- **γ is too high:** *neurons over-compete [current provided by the vector matrix-multiply is noisy]*
  - *Relates to hard WTA case : too few neurons fire (<10%)*
- **γ is too low:** *neurons under-compete [current provided by the vector matrix-multiply is not modified by plasticity]*
  - *Relates to (very) soft WTA case : too many neurons fire (>50%)*
- *At optimal* **γ** values, around 15-30% of neurons seem to fire at each moment.
  - Simulations ongoing to see if this is experimentally feasible.



Source: https://arxiv.org/abs/2003.02357



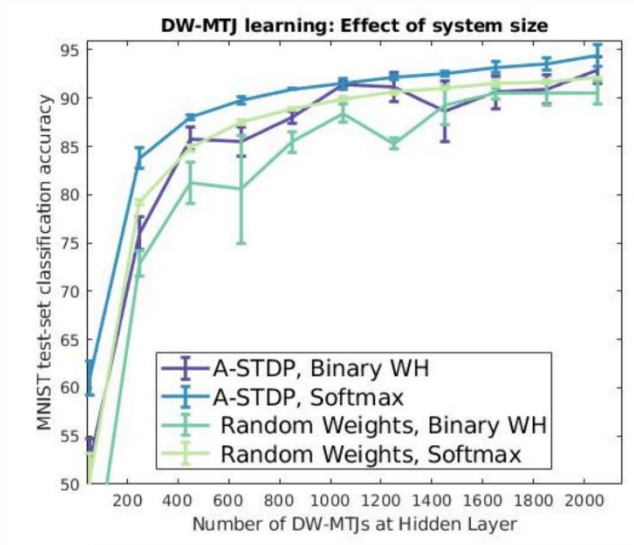
Source: https://arxiv.org/abs/1912.04505

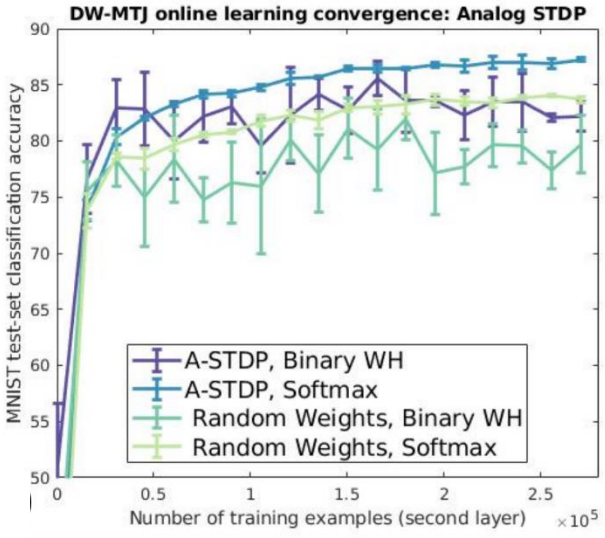
Plasticity-enhanced learning is rapid

- Standard MLP (multi-layer-perceptron) using BP typically requires 200-500k samples
- Our preliminary results demonstrate that plasticity-enhanced learning is extremely rapid (<100k samples required)
  - This is task dependent
- Our current results approach BP performance on all three considered tasks at appropriate **γ** levels

MNIST Task

Policy	Result
Random Weights	93.5%
Perceptron (1 layer NN)	88%
MLP	96.5%
Our Approach	95%



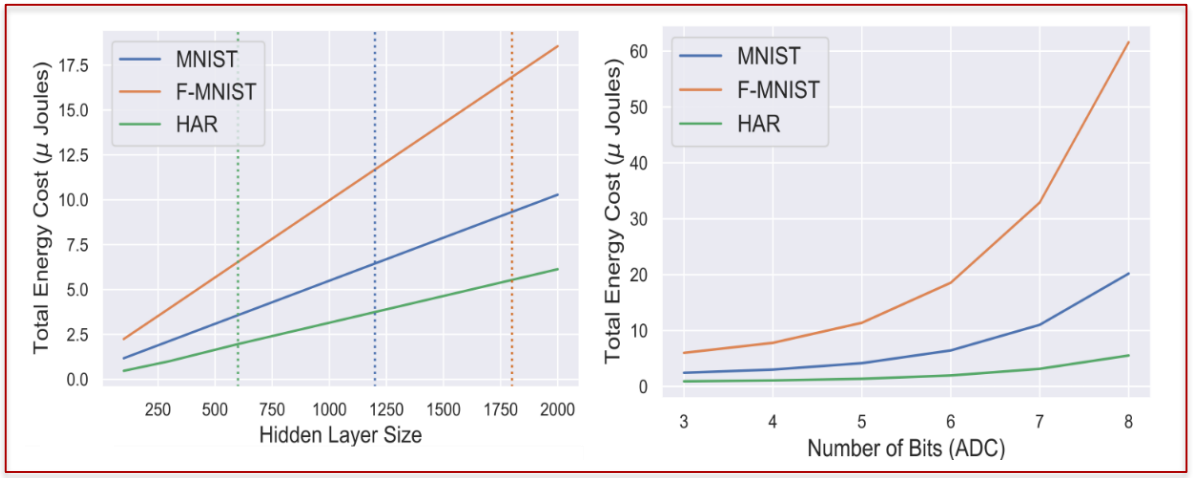


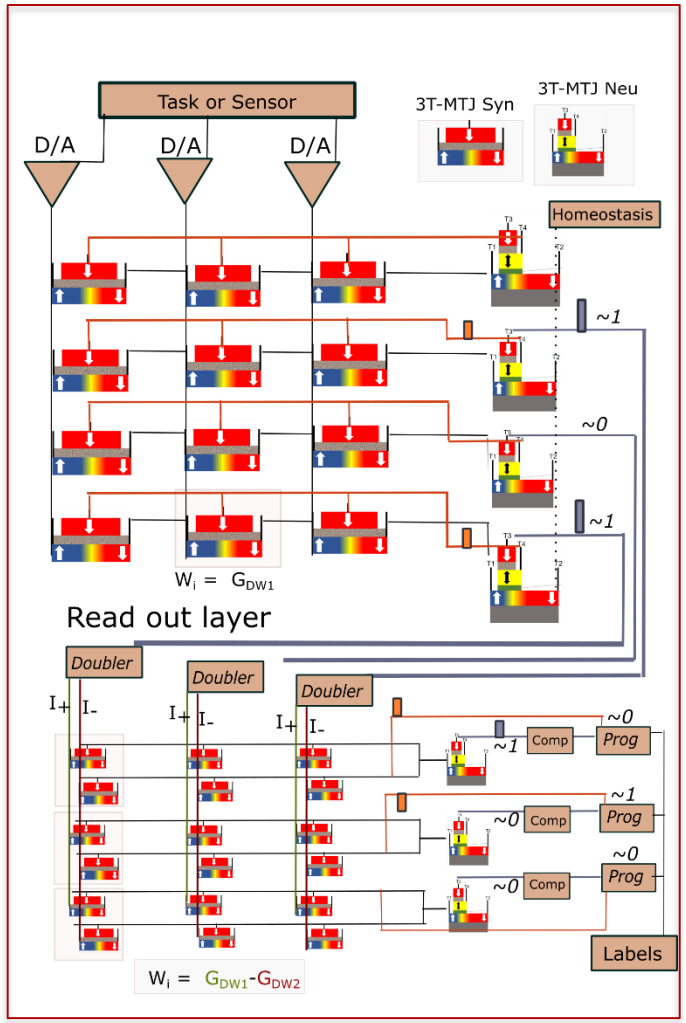
F-MNIST Task

Policy	Result
Random Weights	76%
Perceptron (1 layer NN)	71%
MLP	81.5%
Our Approach	80%

Neural network crossbar architecture

- Required components
  - D/A converters
  - Current mirrors at ‘hidden neurons’
  - Ramp ADC for output in second crossbar
  - Teacher signal generation
  - Logic block and clock for control of second crossbar
- Favorable overall footprint! ☺
  - Ramp ADC drives power costs; *can be pushed down further if we want less output bits*





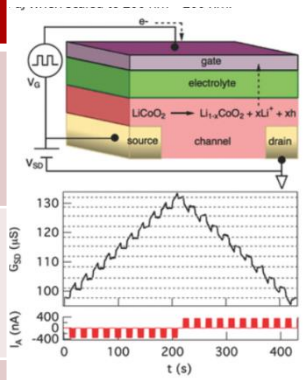
Source: https://arxiv.org/abs/2003.02357

Benchmarking learning overhead energy

- Power reqs are state-of-the-art for magnetic neural networks,
  - Superior to all except for an ionic device
  - But this device requires programming times 1000x+ slower than the DW-MTJ devices!
- Further levers exist for pushing down these numbers even lower:
  - Increasing DW speed
  - Enhancing SOT switching interface efficiency
  - Reducing bits of ADC components
    - Reduced bit activation can be used at training-time to reduce stringency

Result + Energy on MNIST Task *(This work)*

Policy	Result	Energy
Random Weights	93%	7.41 µJ
Our Approach	96% “”	7.41 µJ
MLP	95% “”	53.11 µJ



Power estimates for MNIST Task *(Benchmark)*

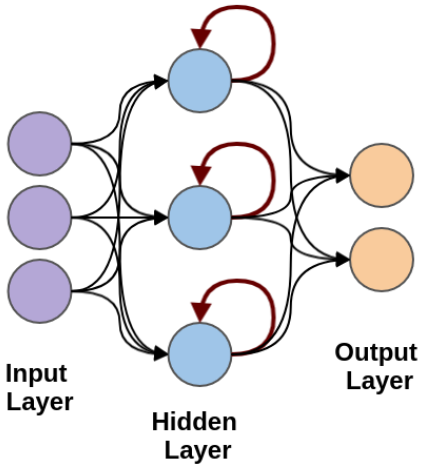
System	Ferroelectric	TaOx	LISTA
NoProp, SGD[a]	$6.28mJ$	$0.053J$	$1.8\mu J$
NoProp, Batch[a]	$127.68\mu J$	$1.09mJ$	$1.78\mu J$
MLP, SGD Analog[b]	$0.75J$	$6.31J$	$7.94mJ$
MLP, SGD Digital[c]	$3.41J$	$28.86J$	$184.39mJ$
MLP, Batch Analog[b]	$15.3mJ$	$70.9mJ$	$7.91mJ$
MLP, Batch Digital[c]	$34mJ$	$0.288J$	$91.52\mu J$

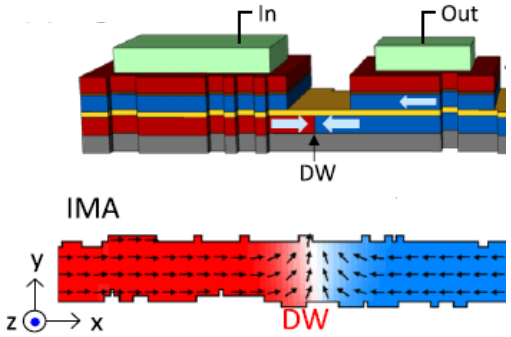
Every value shows the total energy cost of learning considering core

Bennett, Christopher H., et al. "Contrasting advantages of learning with random weights and backpropagation in non-volatile memory neural networks." *IEEE Access* 7 (2019): 73938-73953.

Outline

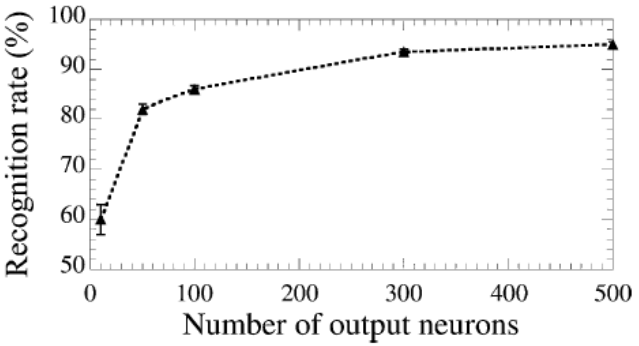
- **Intro and Motivation: Building Spintronic Accelerators**
- **Critical Bio-realistic Neuronal effects with DW-MTJs**
- **Implementation of Semi-supervised Learning**
- **Discussion & Future Work**



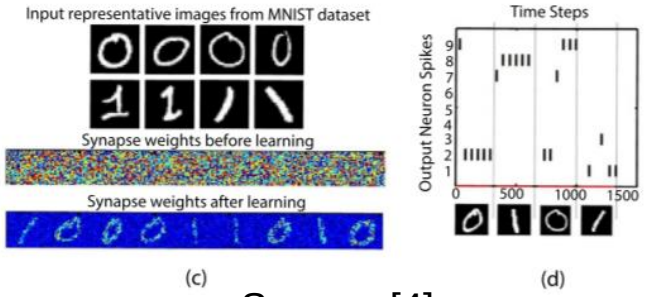


Comparisons to other STDP learning systems

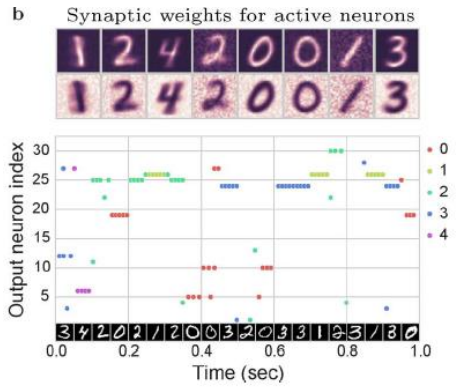
- STDP results comparable to best results (93%) [1],[2] for NN combining supervised + unsupervised approaches
  - A more realistic + energy-efficient read-out method
- STDP results superior to those obtained using memristor +ReRAM LIF emulator neurons (78%) [3]
  - LIF circuit also had a high level of complexity
- STDP results superior to those using rate-encoding techniques with stochastic MTJs [4]



Source: [1]



Source: [4]

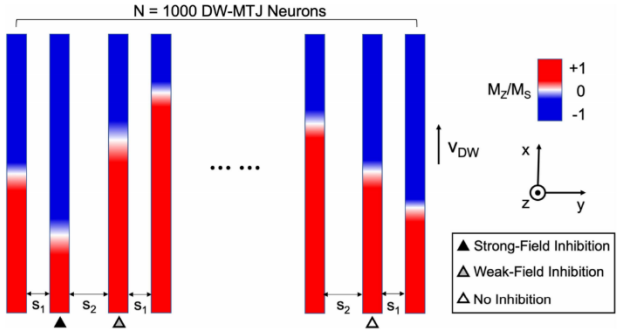


Source: [3]

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

Future work and physical realizability

- Since correct learning relies on high interaction, more work is done to realize tight coupling between sampling neurons
  - Early works suggests that alternating inter-neuron spacings can realize better WTA performance by maximizing stray-field interactions



Source: https://iopscience.iop.org/article/10.1088/1361-6528/ab86e8/pdf

- Max bit resolution of DW-MTJ synapses needs to be increased for state of the art tasks
  - 4-5 bits is OK on fMNIST/MNIST, but CIFAR and ImageNet tasks collapse at this level.
  - DW-MTJ tracks may only support 4-5 bits; an alternative to realize 7+ bit is bit-slicing.

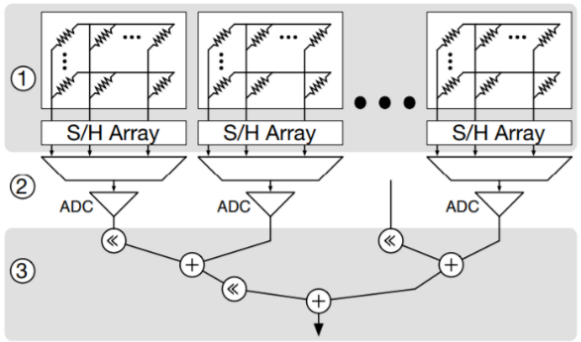


Figure 3. Cluster organization.

Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8416841

Summary

Sandia National Laboratories

- DW-MTJ devices have been co-integrated as neurons and synapses in a semi-supervised learning system
  - Fares better than ReRAM synapses + CMOS Neurons systems and competing spintronic proposal
  - Actual learning performance on tasks is acceptable (> 90% for MNIST) but can be improved
- Careful circuit and algorithm choices implement efficient learning with low-latency
  - Ground-breaking energy accuracy of : <10 µJ energy cost is reached by using a Ramp ADC, current mirrors at the hidden layer, and low-bit precision
  - This energy budget will need to increase on more complex machine learning tasks.

Next Steps

- Real DW-MTJ Synapse and neuron devices are being fabricated and benchmarked against simulation
  - Will eventually allow for a cross-bar realistic simulation for full backprop learning using CrossSIm
  - In principle should be able to stack/combine unsupervised layers (compare to Boltzmann Machines)
- Continue to evaluate lateral inhibition effects and possible clustering implementations
  - This is a hard constraint for effective sub-sampling in k-WTA
  - CMOS-supported clustering is a lot more energy expensive.

CROSS SIM

https://cross-sim.sandia.gov

Thank you! Questions?

Collaborators:

Prof. Joseph S. Friedman, U.T. Dallas
Naim Hassan & Xuan Hu, U.T. Dallas

Prof . Jean Anne C. Incorvia,
Can Cui, O. Akinola, U.T. Austin

Dr. Matthew J. Marinella, Sandia Labs
Dr. T. Patrick Xiao, Sandia Labs

Appendix: Energy and performance data on all tasks

Sandia National Laboratories

TABLE I
CLASSIFICATION AND REGRESSION TASK PERFORMANCE

Task	Learning Style		
	Random, Ana-BP	STDP, Bin-BP	STDP, Ana-BP
HAR	96.13%	95.83%	97.93%
MNIST	93.52%	93.12%	94.92%
f-MNIST	76.52%	77.52%	79.52%