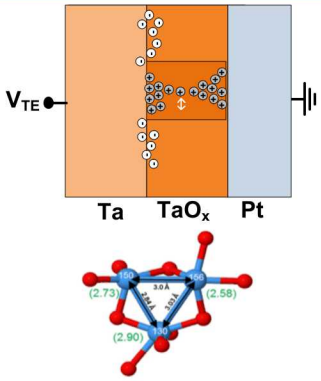
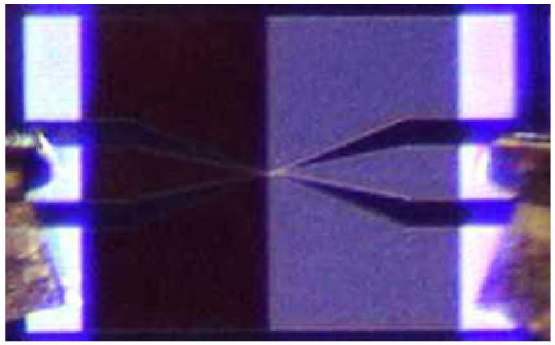
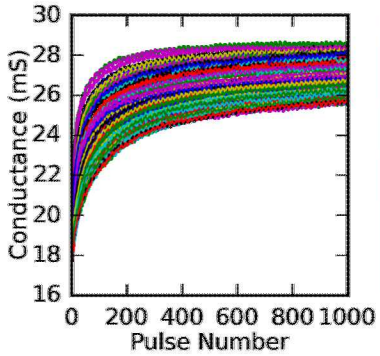
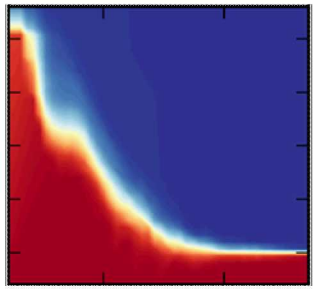
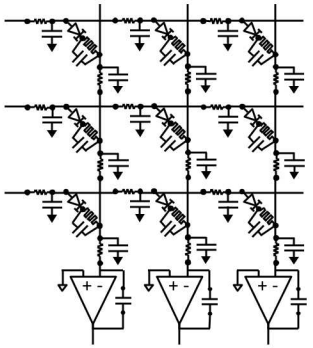
SAND2020-4465C











Efficient Neural Network Processing with Analog In-Memory Computing

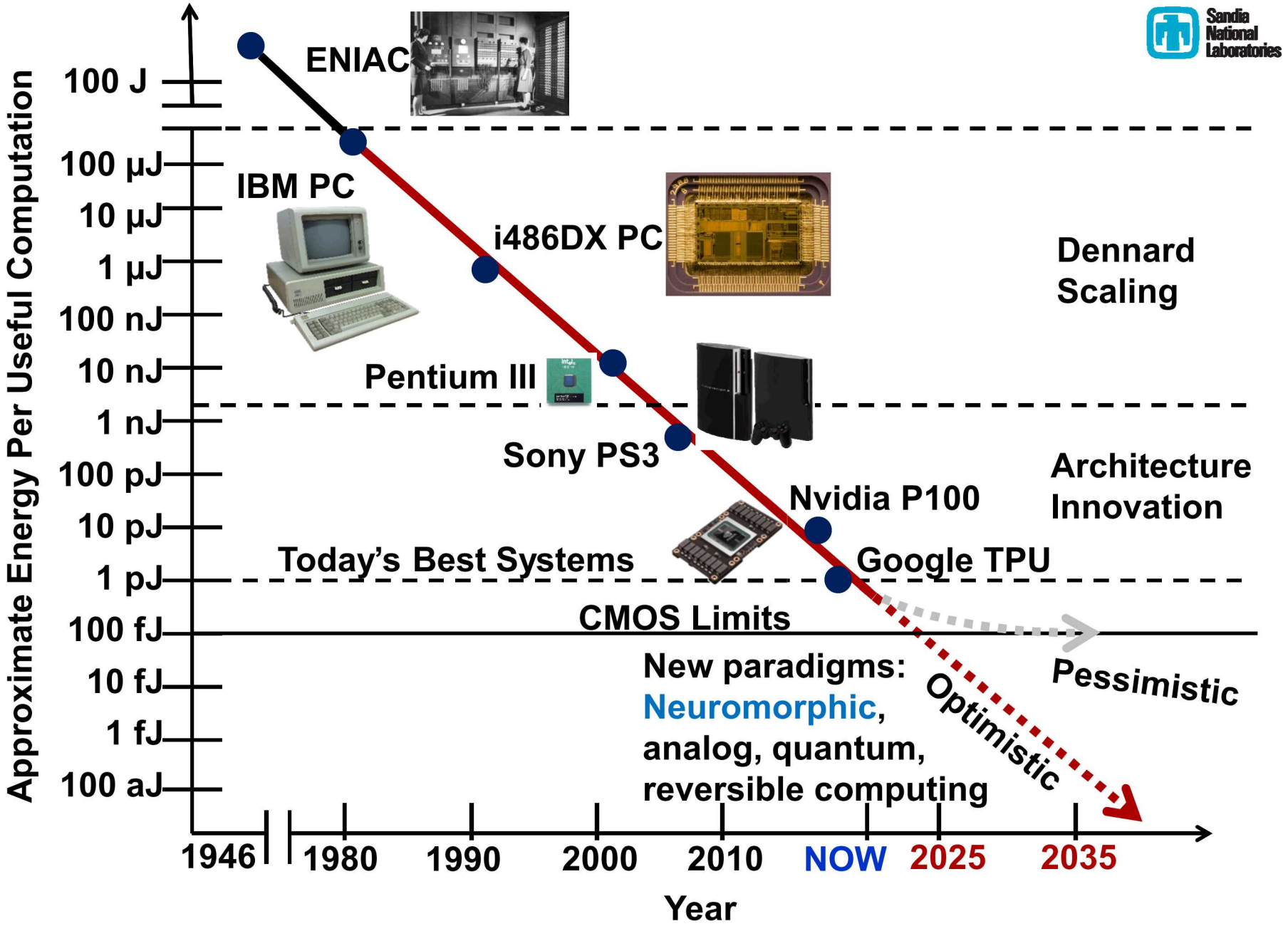
Matthew J. Marinella,
S. Agarwal, C. Bennett, P. Xiao, R. Jacobs-Gedrim, D.R. Hughart, E. Fuller, A.A. Talin,
J.A. Incorvia*, H Barnaby**
Sandia National Laboratories, Albuquerque, NM
*University of Texas, Austin, **Arizona State University

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

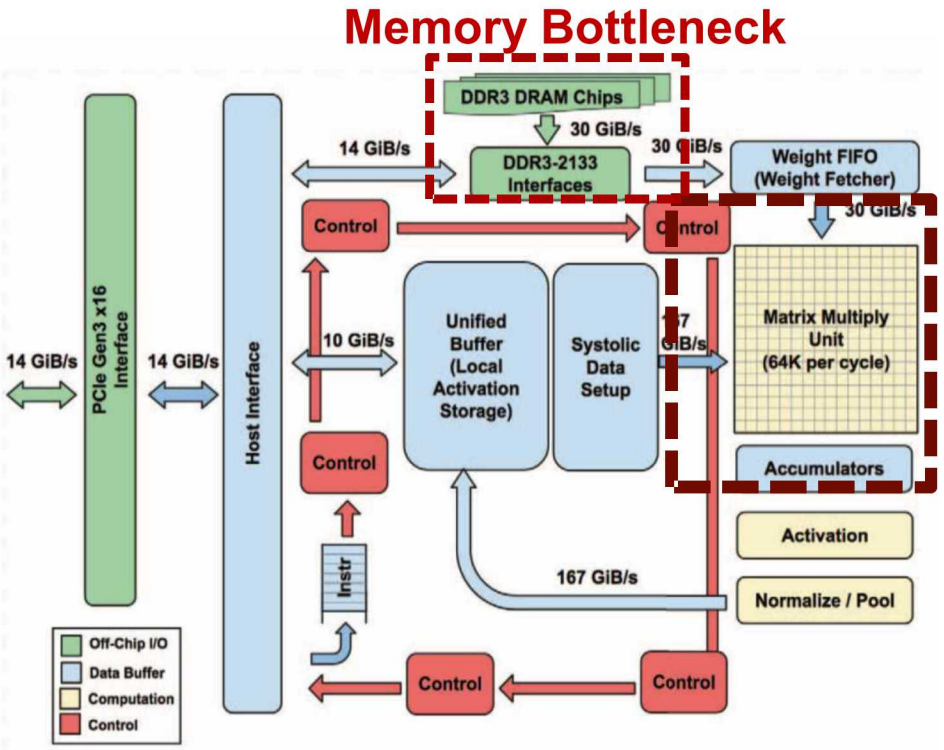
Sandia National Laboratories

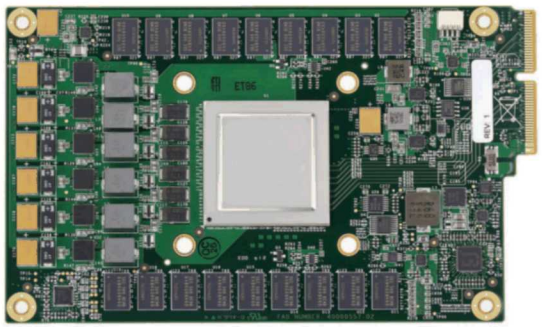
- **Motivation and Background**
- **Analog Accelerator Concepts & Performance Model**
- **Accelerator Accuracy Modeling & Characterization**
- **Future Directions**
- **Radiation Effects**
- **Summary**


Sandia National Laboratories
Approximate Energy Per Useful Computation
100 J
100 µJ
10 µJ
1 µJ
100 nJ
10 nJ
1 nJ
100 pJ
10 pJ
1 pJ
100 fJ
10 fJ
1 fJ
100 aJ
ENIAC
IBM PC
i486DX PC
Pentium III
Sony PS3
Nvidia P100
Today's Best Systems
Google TPU
CMOS Limits
Dennard Scaling
Architecture Innovation
New paradigms:
Neuromorphic,
analog, quantum,
reversible computing
Optimistic
Pessimistic
1946
1980
1990
2000
2010
NOW
2025
2035
Year

Digital Accelerators: Google TPU

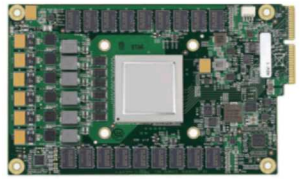
- Tensor Processing Unit: Machine learning accelerator in use by Google since 2015
- Accelerates neural network inference with TensorFlow
  - New versions may include training
- Core: matrix multiply unit





Digital Accelerators: Google TPU

- TPU v1 Performance results released at ISCA 2017
  - Die level performance of 2.3 TeraOps/W
  - → **~ 1 pJ per 8 bit operation!**
- Contemporary Intel Haswell die comparison:
  - 18 GigaOps/W
  - → 55 pJ per 8-bit operation

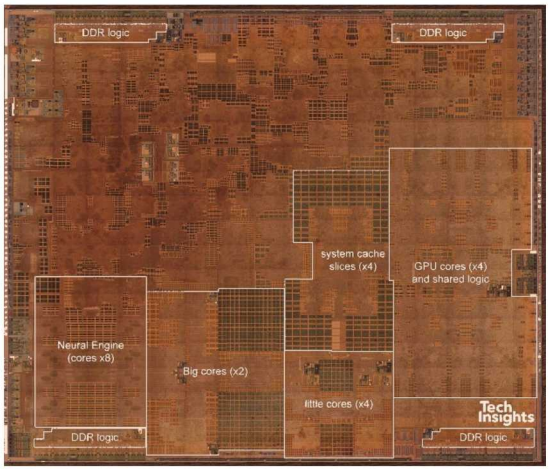


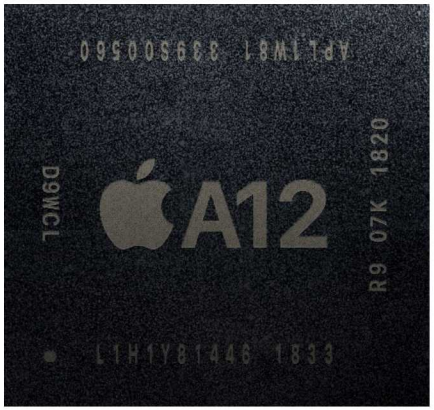
Model	*Die*										*Benchmarked Servers*				
	mm²	*nm*	*MHz*	*TDP*	*Measured*		*TOPS/s*		*GB/s*	*On-Chip Memory*	*Dies*	*DRAM Size*	*TDP*	*Measured*	
					Idle	*Busy*	8b	FP						*Idle*	*Busy*
Haswell E5-2699 v3	662	22	2300	145W	41W	145W	2.6	1.3	51	51 MiB	2	256 GiB	504W	159W	455W
NVIDIA K80 (2 dies/card)	561	28	560	150W	25W	98W	--	2.8	160	8 MiB	8	256 GiB (host) + 12 GiB x 8	1838W	357W	991W
TPU	<331*	28	700	75W	28W	40W	92	--	34	28 MiB	4	256 GiB (host) + 8 GiB x 4	861W	290W	384W

N. P. Jouppi *et al, ISCA*, Toronto, ON, 2017

State of the Art: Apple A12

- Apple's iPhone X main SoC processor
  - 7nm TSMC process
- Includes "neural engine" accelerator IP
  - Apple spec: 5 TeraOps/s (8 bit)
  - Power is ~2.5-5W
  - Hence, ~ 1-2 TeraOp/W,
- **State of the art neural accelerator is about 500fJ to 1pJ per 8 bit operation**
  - Nearing digital limits?
- Similar for published accelerator data, modern GPU, and Google TPU





apple.com, techinsights.com

Outline

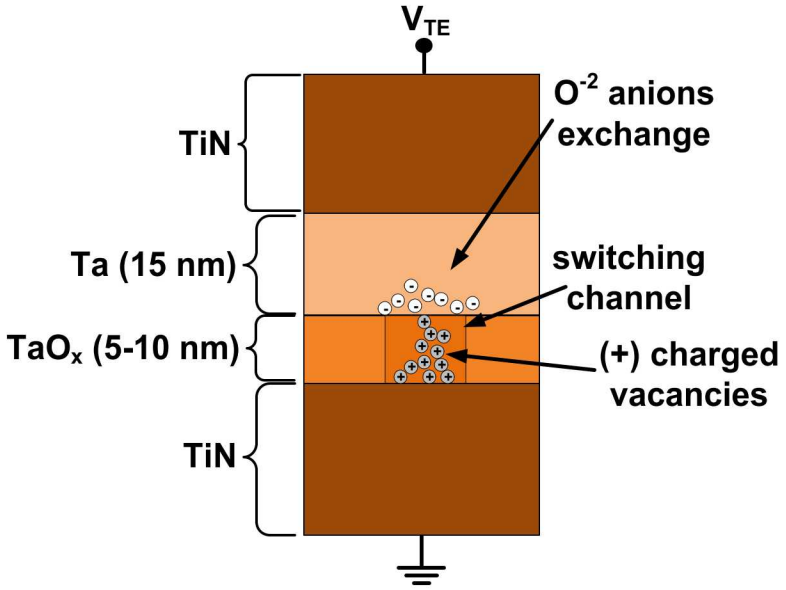
- Motivation and Background
- Analog Accelerator Concepts & Performance Model
- Accelerator Accuracy Modeling & Characterization
- Future Directions
- Radiation Effects
- Summary

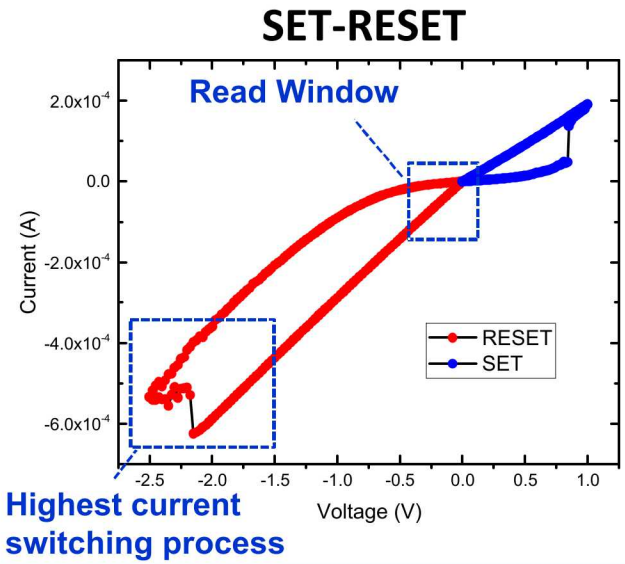
Exemplar Synapse: Oxide ReRAM

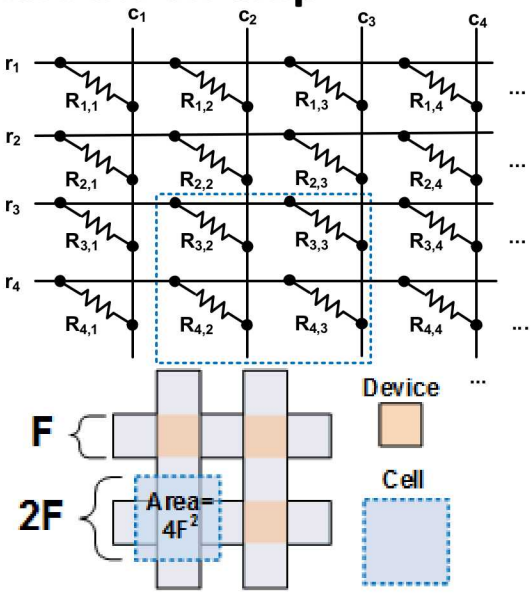
Sandia National Laboratories

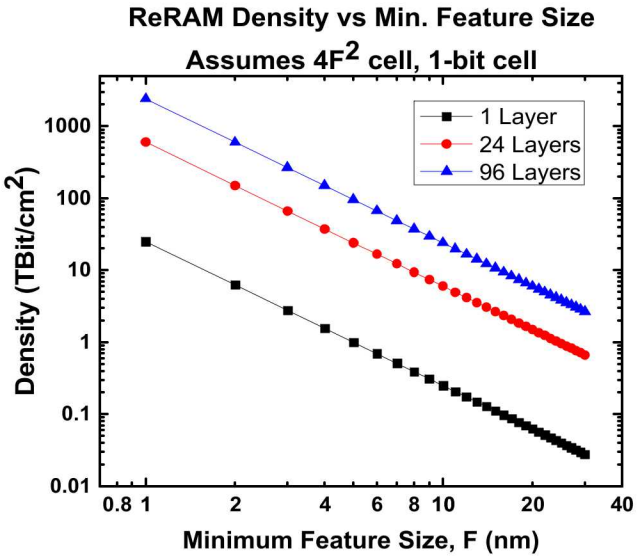
ReRAM (OxRAM)

- Starts as insulating MIM structure
- Forming: remove O^{2-} → soft breakdown
- Bipolar resistance modulation
- Excellent memory attributes: Switching in less than 1ns, less than 1 pJ demonstrated, scaling to 5nm, >10^{12} write cycles possible
- Potential for 100 Tbit of ReRAM on chip



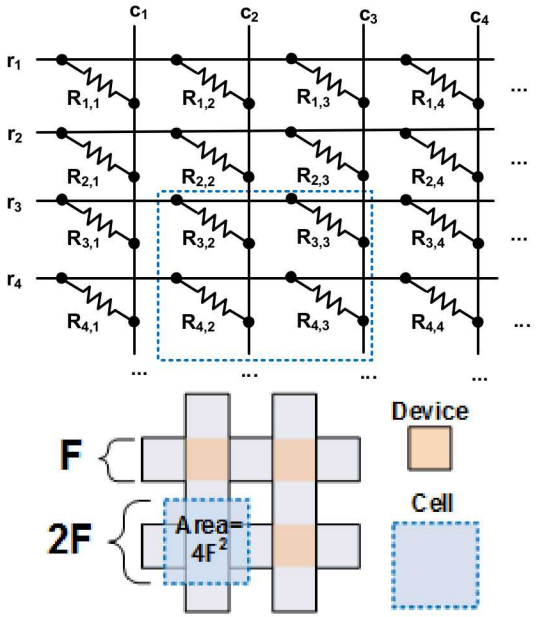


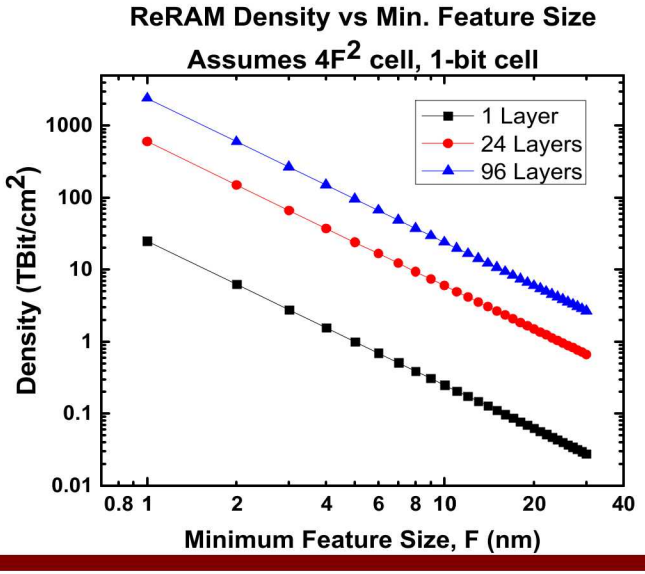




Crossbar Theoretical Limits

- Potential for 100 Tbit of ReRAM on chip
- What if we can compute with ReRAM?
- If each can perform 1M computations of interest per second (1 M-op):
  - 10^{12} active devices/chip x 10^{6} cycle per second →10^{18} comps per second per chip
  - Exascale-computations per sec on one chip!
- In order to not melt the chip, entire area must be limited to ~100W
- Allowed energy per operation = P x t/op = 100W / 10^{18} = 10^{-16} = 100 aJ/operation
- 10nm line capacitance = 10 aF
- Can charge line to 1V with 10 aJ





Analog Computation with a ReRAM Crossbar

- Electronic Vector Matrix Multiply

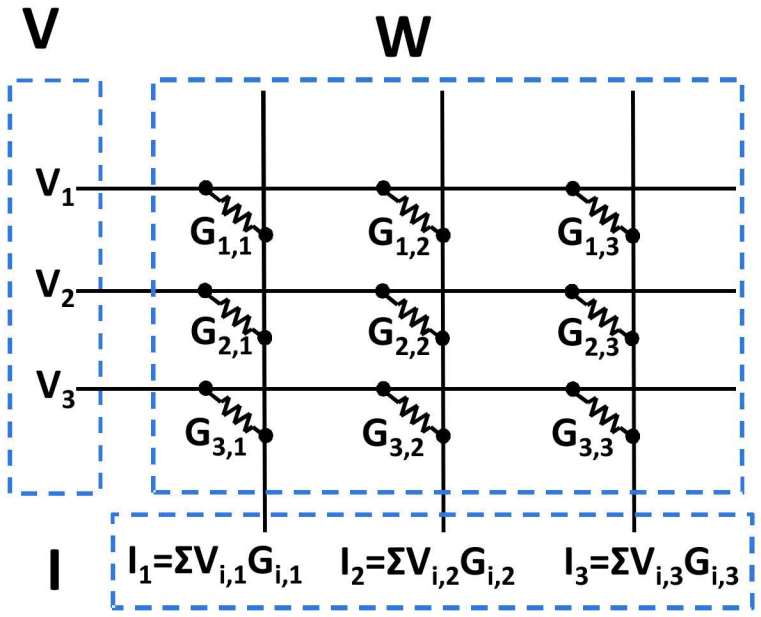
Mathematical

$$V^T W = I$$

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

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

Electrical

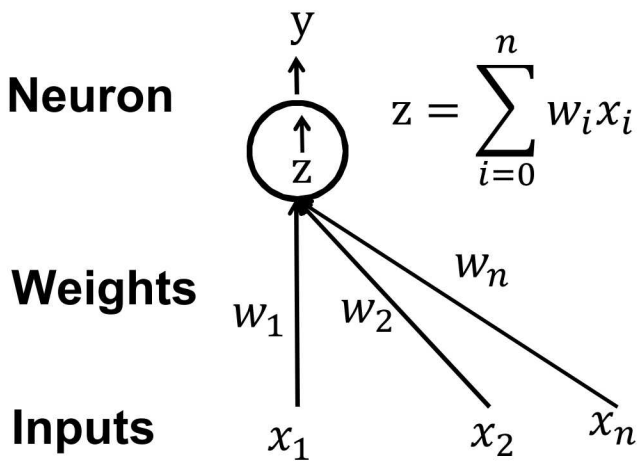


Basics of Neural Networks

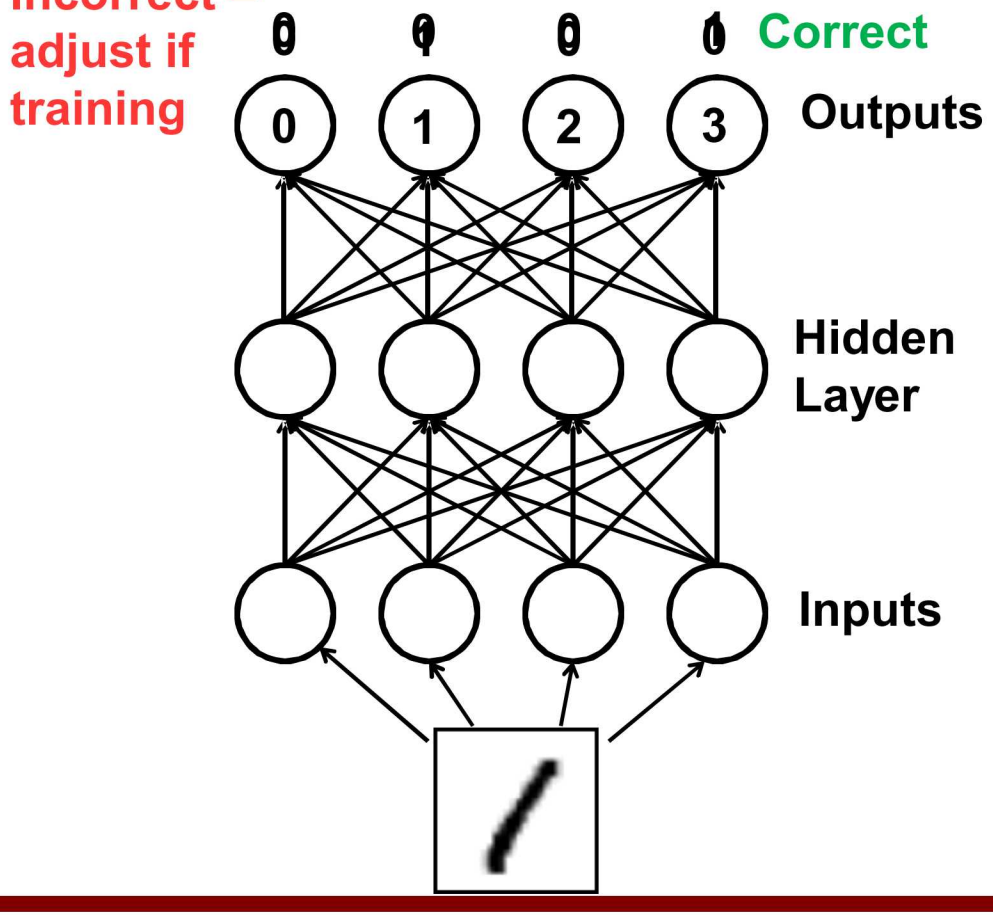
Sandia National Laboratories

Basic Building Block

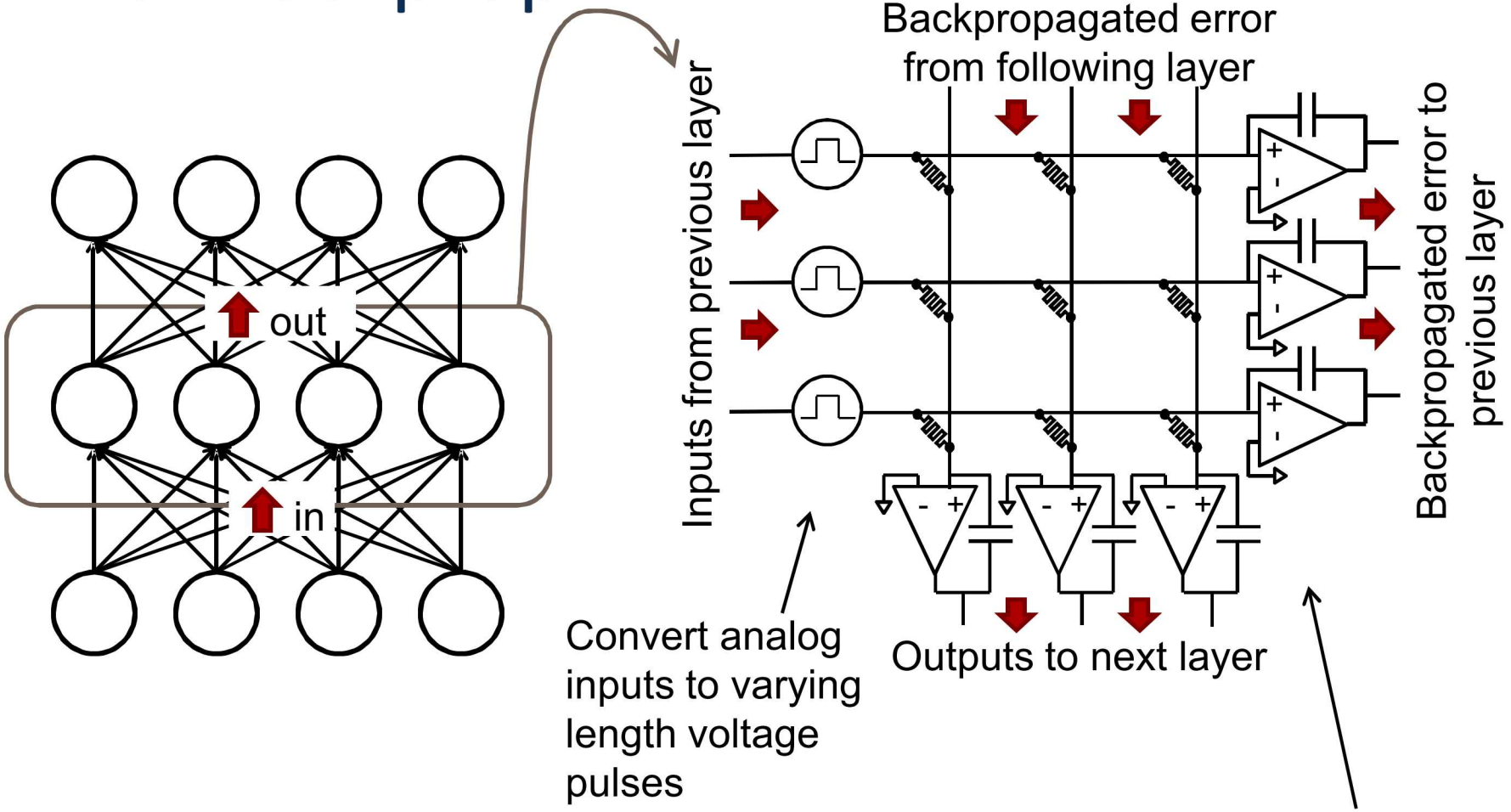
$$y = \frac{1}{1 + e^{-z}}$$



Simple Network: Backpropagation

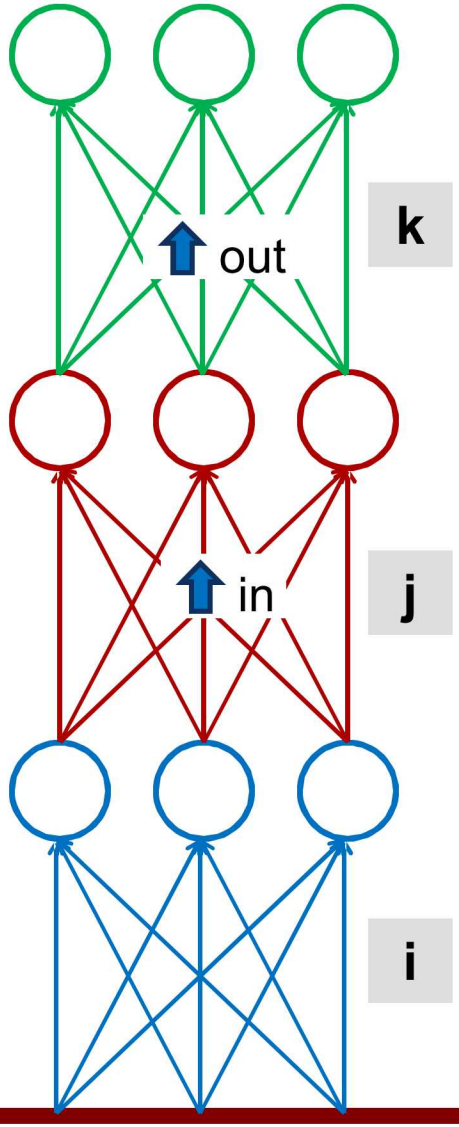


Mapping Neural Network to Crossbar for Backprop



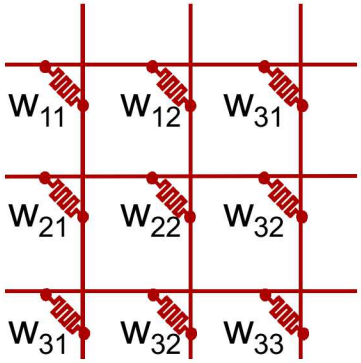
Vector Matrix Multiply, Matrix Vector Multiply, Rank 1 Update:
Key kernels used in many algorithms

Forward Propagation – Inference



O(N²) Operations

Neural Core



$$z_j = \sum_i y_i \times w_{ij}$$

y_i

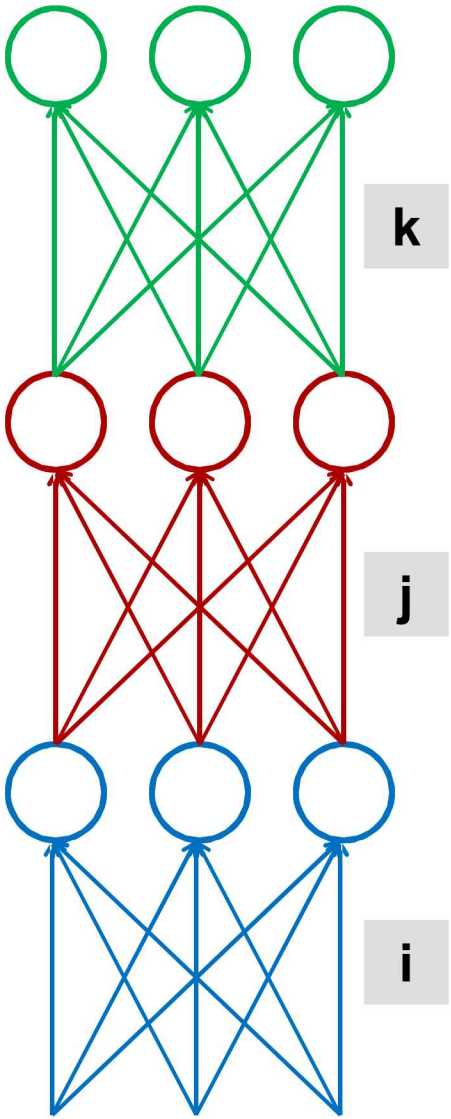
O(N) Operations

y_j

Digital Core

$$y_j = \frac{1}{1 + e^{-z_j}}$$

Back Propagation – Training



error δ_k

w_{11} w_{12} w_{31}
w_{21} w_{22} w_{32}
w_{31} w_{32} w_{33}

$$\Delta_k = \sum_k w_{jk}\delta_k$$

Δ_k

Digital Core

$$\delta_j = \frac{dy}{dz}(z_j) \cdot \Delta_k$$

$\eta \times \delta_j$

y_i

z_j

$$z_j = \sum_i y_i \times w_{ij}$$

y_i

y_i

w_{11} w_{12} w_{31}
w_{21} w_{22} w_{32}
w_{31} w_{32} w_{33}

w_{11} w_{12} w_{31}
w_{21} w_{22} w_{32}
w_{31} w_{32} w_{33}

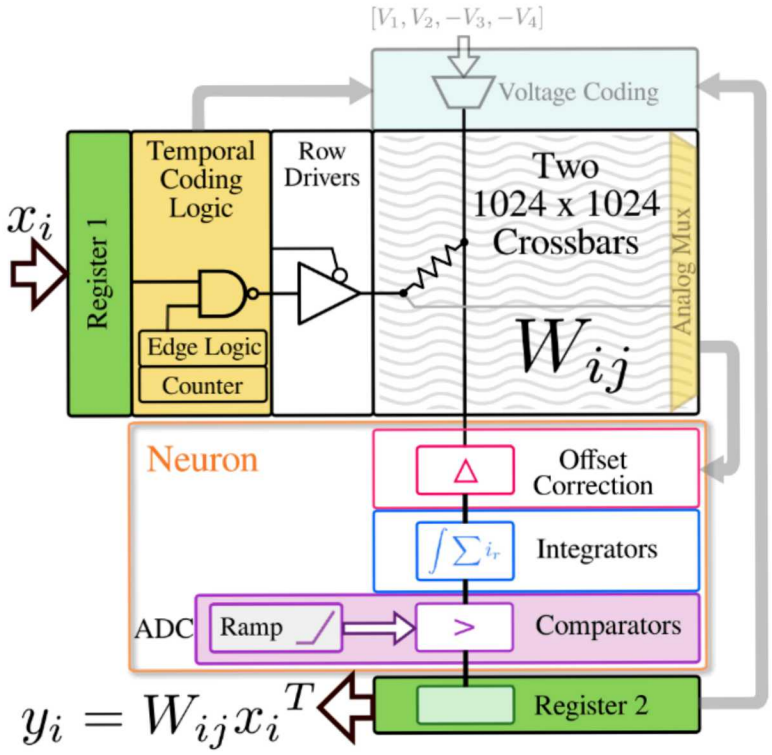
O(N²) Read Operations

O(N) Operations

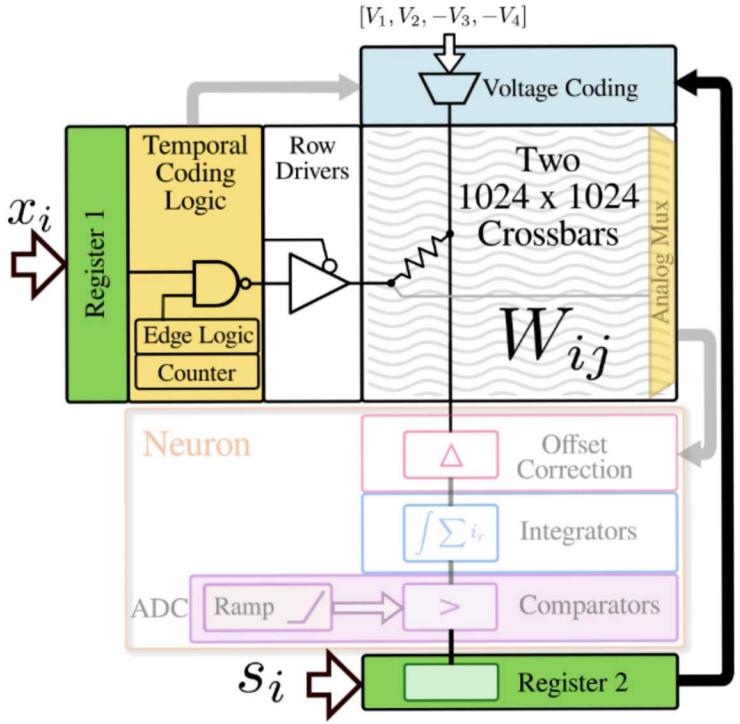
O(N²) Write Operations

Key Circuit Block/Kernel Analysis

Vector Matrix Multiply (Inference)



Rank-1 Update (Training)



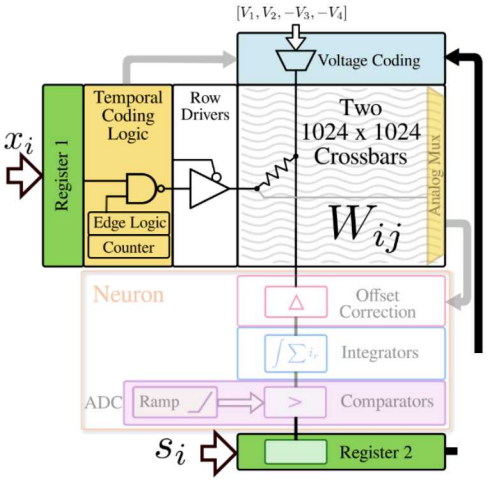
Marinella, Agarwal, et al, *IEEE JETCAS*, 2018

Device Assumptions for Analysis

Quantity		Value
Interconnect	Full Pitch (W_{M1_Pitch})	64 nm
	Capacitance	~200 aF/μm
	Resistance	~30 Ω/μm
Logic Transistor	Area	~0.04 μm^2
	Voltage	0.8 v
High-Voltage Transistor	Area	~0.35 μm^2
	Voltage	1.8 v
Crossbar	Dimensions $(n_{rows} \times n_{cols})$	1024 × 1024
	Minimum Pulse Width	1 ns
ReRAM & Select Device	ReRAM ON/OFF Ratio	10
	Capacitance (C_{ReRAM})	35 aF
Analog ReRAM & Select Device	On State Read Current	1 nA (R_{on} = 100 MΩ)
	On State Write Current	10.3 nA (R_{on} = 100 MΩ)
	Read Voltage	0.785 V
	Write Voltage	1.8 V

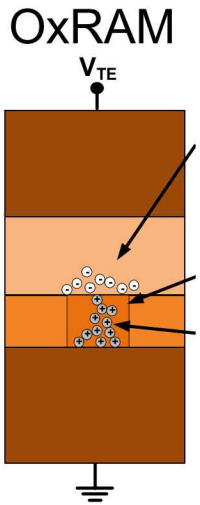
Marinella, Agarwal, et al, *IEEE JETCAS*, 2018.

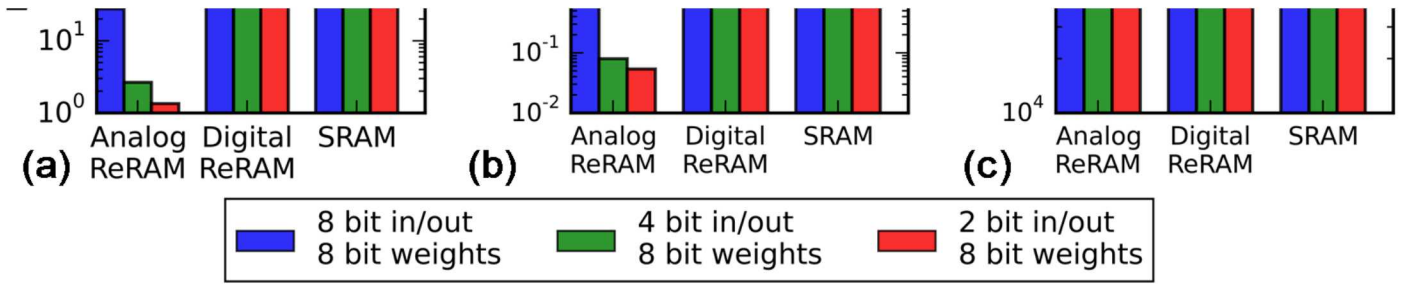
Matrix Energy/Latency Analysis



Component	Vector Matrix Multiply	Matrix Vector Multiply	Outer Product Update
Energy/Op ReRAM (fJ)	12.2	12.2	2.1
Energy/Op SRAM (fJ)	2718	4630	4102
Array Latency ReRAM (µs)	0.38	0.38	0.51
Array Latency SRAM (µs)	4	32	8

Based on a commercial 14/16 nm PDK





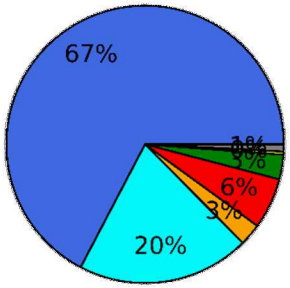
Two orders of magnitude beyond digital!

Neural Core Energy Analysis

Sandia National Laboratories

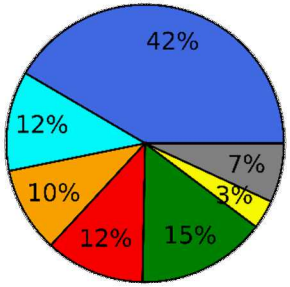
Analog ReRAM

8 bits In/out 8 bit weights



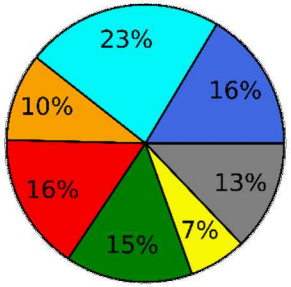
28 nJ

4 bits In/out 8 bit weights

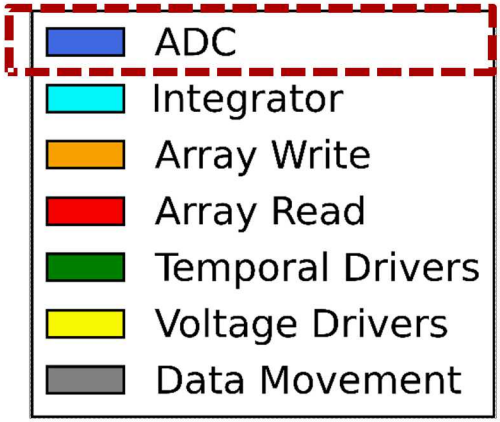


2.7 nJ

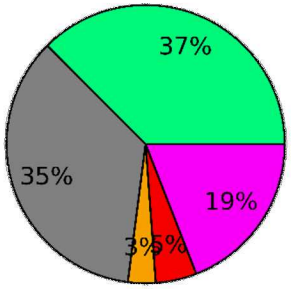
2 bits In/out 8 bit weights



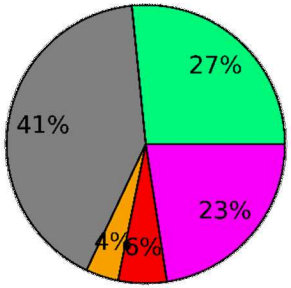
1.3 nJ



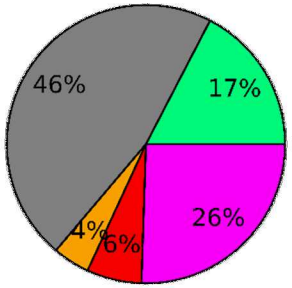
SRAM



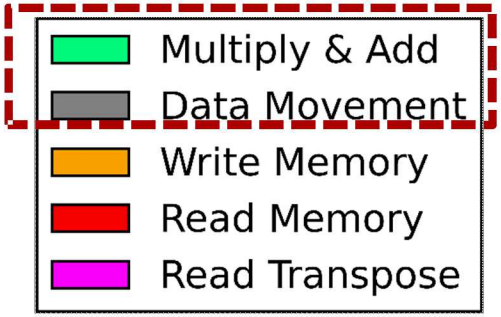
12,010 nJ



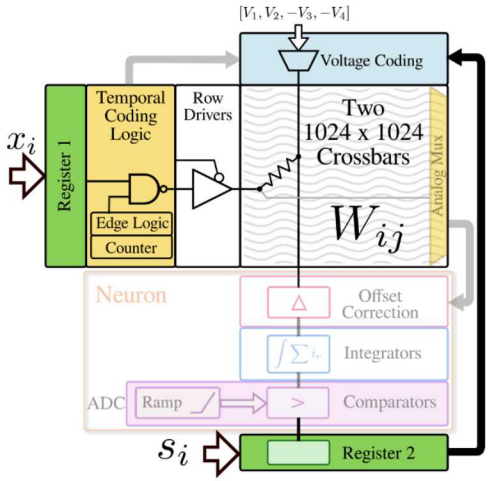
10,150 nJ



8,970 nJ



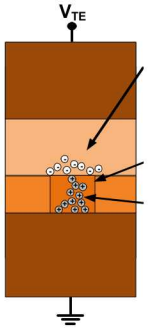
Analog Device Comparison



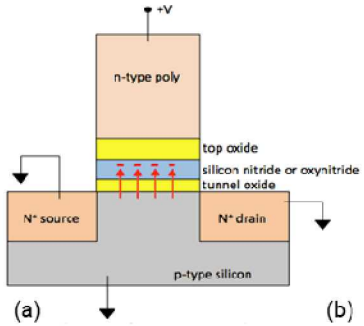
Marinella, et al, *IEEE JETCAS*, 2018

Component	Vector Matrix Multiply	Matrix Vector Multiply	Outer Product Update
Energy/Op IFG (fJ)	11.9	11.9	0.2
Energy/Op ReRAM (fJ)	12.2	12.2	2.1
Energy/Op SONOS (fJ)	13.7	13.7	68.2
Energy/Op SRAM (fJ)	2718	4630	4102
Array Latency IFG (µs)	0.39	0.39	1.9
Array Latency ReRAM (µs)	0.38	0.38	0.51
Array Latency SONOS (µs)	0.40	0.40	20
Array Latency SRAM (µs)	4	32	8

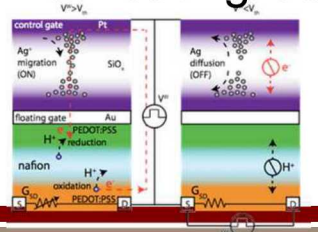
OxRAM

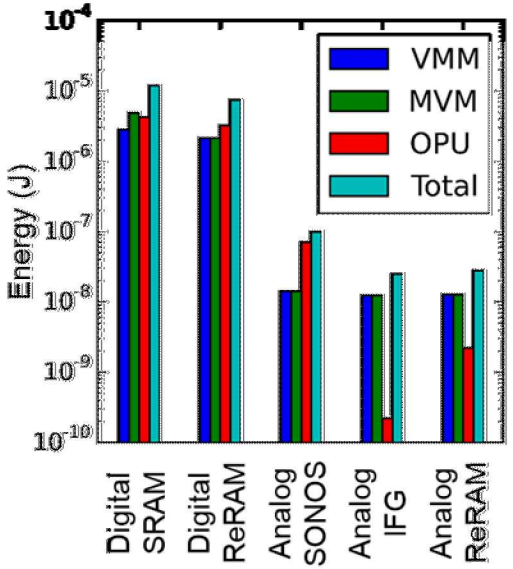


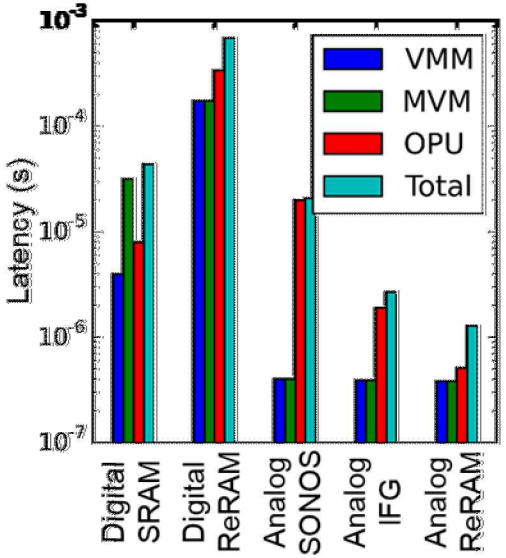
SONOS



Ionic Floating Gate





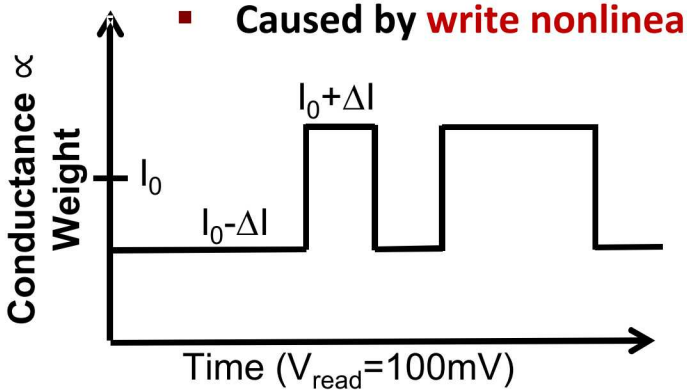


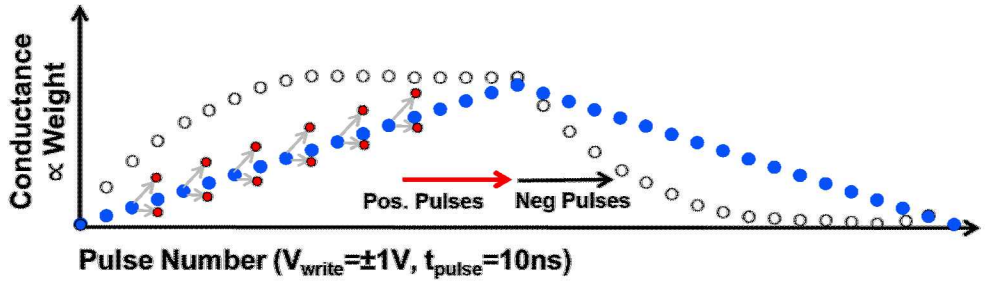
Outline

- Motivation and Background
- Analog Accelerator Concepts & Performance Model
- Accelerator Accuracy Modeling & Characterization
- Future Directions
- Radiation Effects
- Summary

Experimental Device Analog Nonidealities

- **Now we need to consider the challenges of analog processing**
- **Digital: During normal operation has perfect accuracy**
  - **Degraded device quality might mean reduced clock speed or reliability issues**
- **Analog: Device characteristics affect *algorithm accuracy*!**
  - **Research challenge: analog behavior must give acceptable algorithm-level results**
- **Inference challenges:**
  - **Measured device conductance not proportional to weight**
  - **Caused by current drift, noise**
- **Training challenges:**
  - **Actual analog device state change does not match intended weight update**
  - **Caused by write nonlinearity, asymmetry, stochasticity**

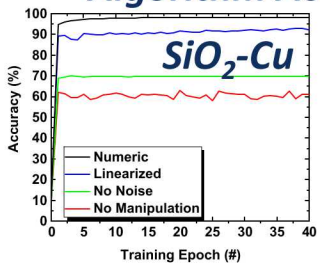


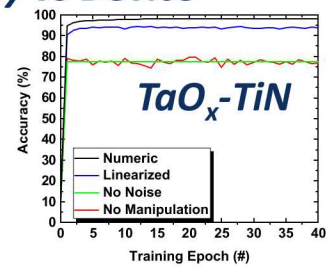


System Level Effects of Device Analog Properties: Multiscale Modeling

Sandia National Laboratories

Algorithm Accuracy vs Device



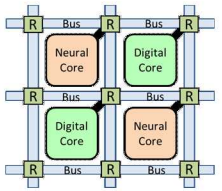


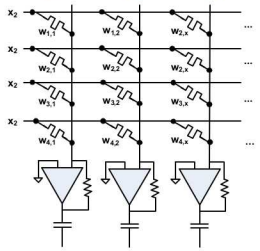
Target Algorithms

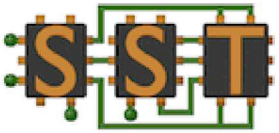
- Deep Learning (MLP)
- Deep Convolutional Nets

Algorithm Impact

Hardware System Models



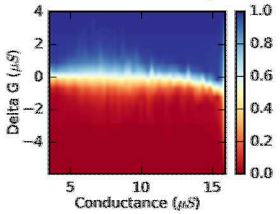




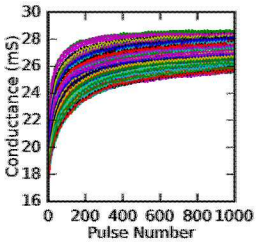
Sandia SST:
Model performance and energy requirements using HPC

Sandia Cross-Sim:
Translates device measurements and crossbar circuits to algorithm-level performance

CROSS SIM

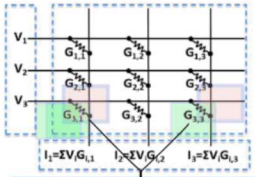


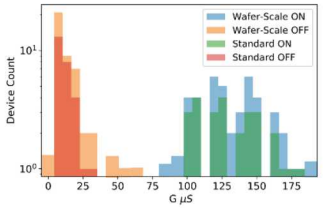
Circuit

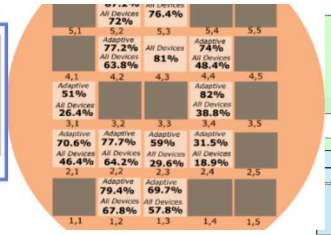


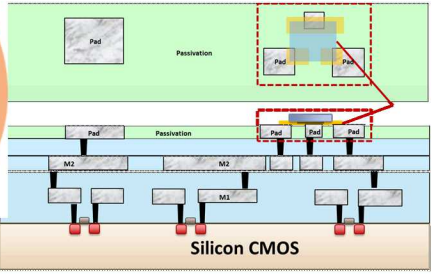
Memristor fabrication and measurements in MESAFab

Device

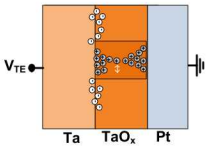


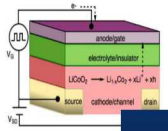


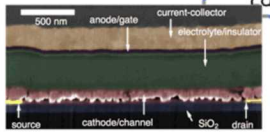




Advanced Materials

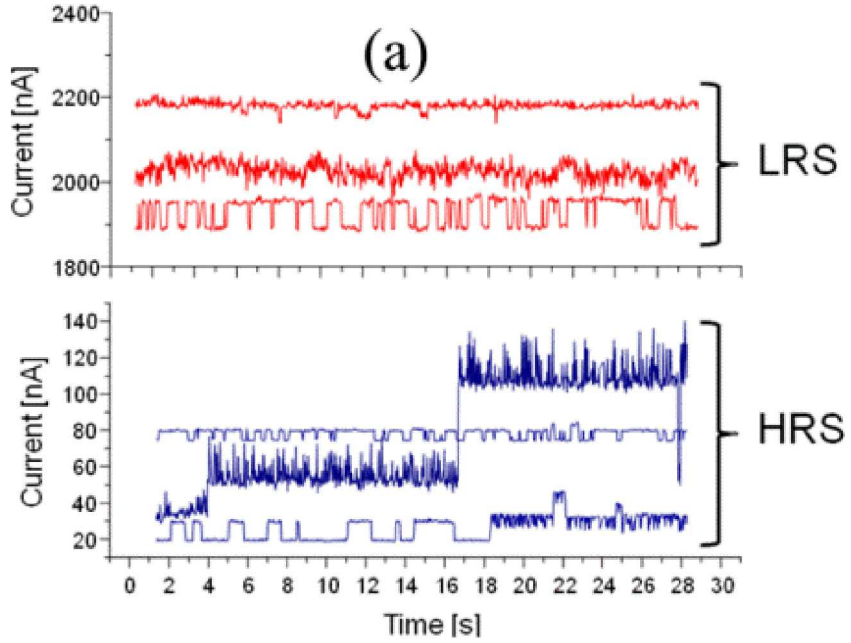




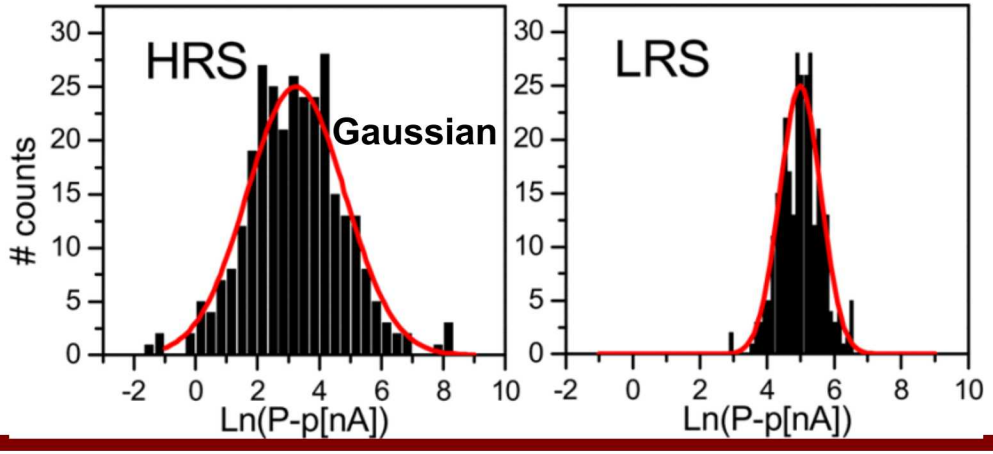


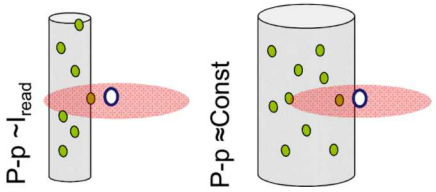
Modeling Effects of Noise on ReRAM Inference Accelerator

- **Read noise referred to as random telegraph noise (RTN) often observed in ReRAM and CBRAM**
- **Example RTN noise measured in HfO RRAM**
- **Mechanism and magnitude depends on conductance state**



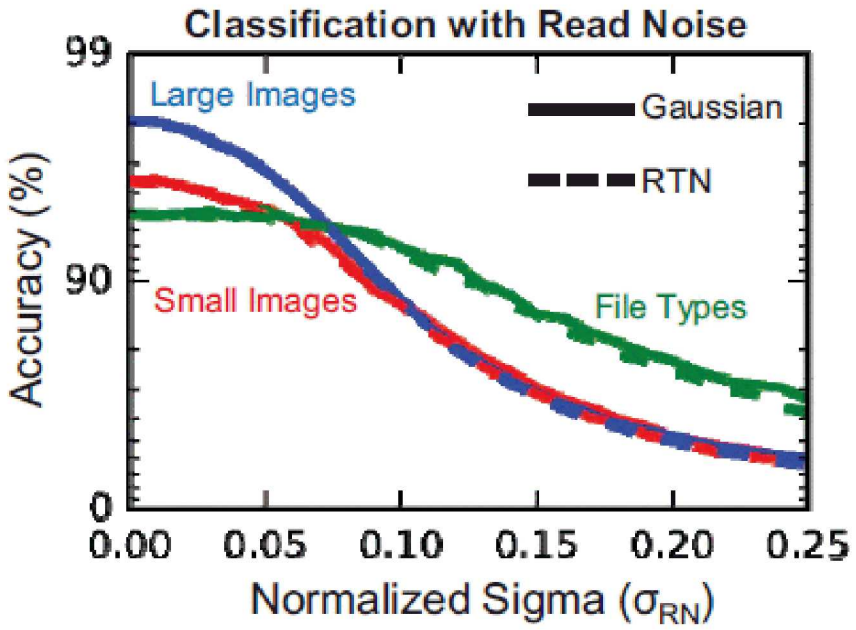
Example High Resistance and Low Resistance State Noise

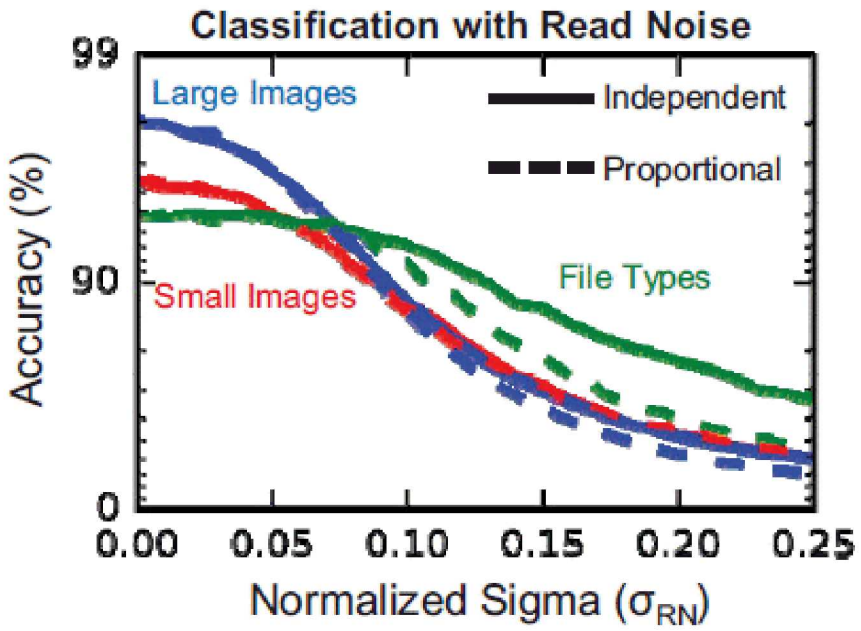




Modeling Effects of Noise on ReRAM Inference Accelerator

- **Trained with no noise, noise added during inference**
- **RTN and Gaussian similar**
- **Noise affects inference at about σ≈5% of dynamic range**

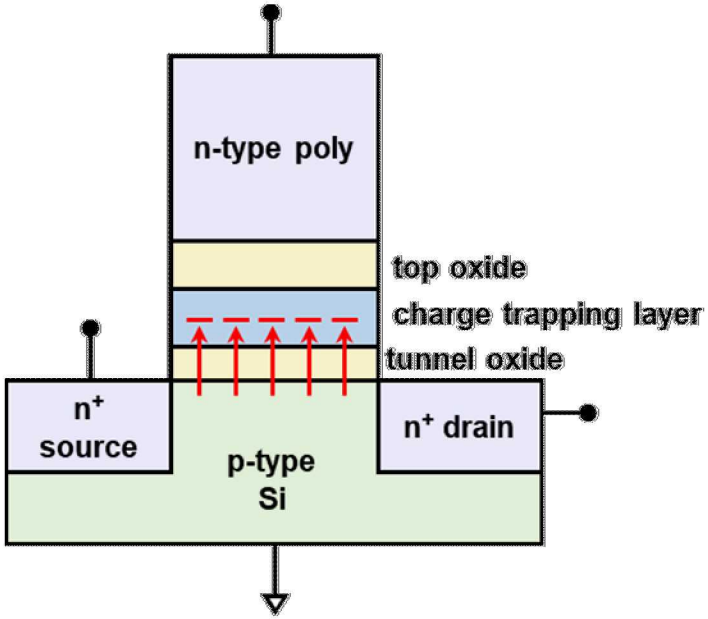


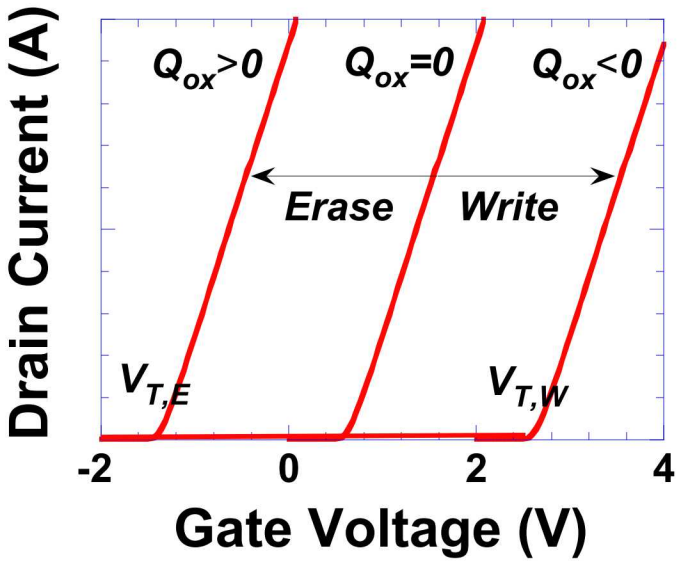


S Agarwal et al, IJCNN 2016

Modeling Effects of State Drift on SONOS Inference Accelerator

- Semiconductor-oxide-nitride-oxide-semiconductor (SONOS) and related charge trapping devices
  - One of the most common devices for flash memory
- MOSFET with a tunable threshold voltage based on the amount of charge in the charge trapping layer

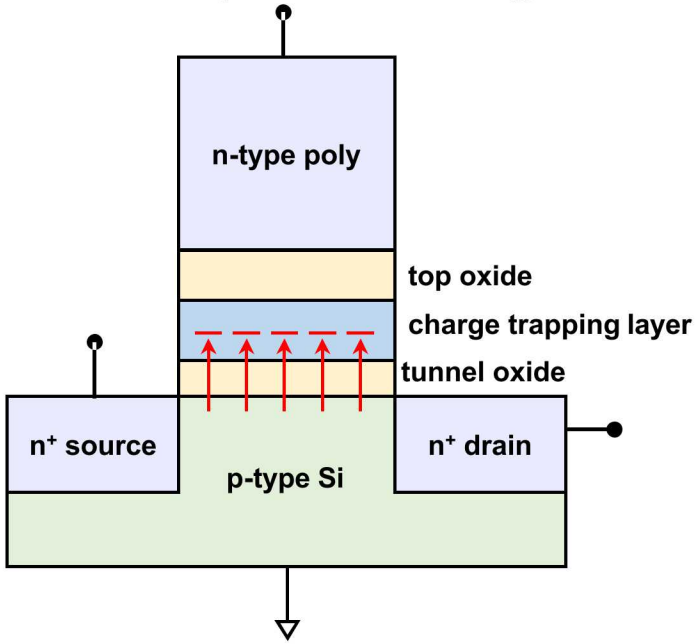


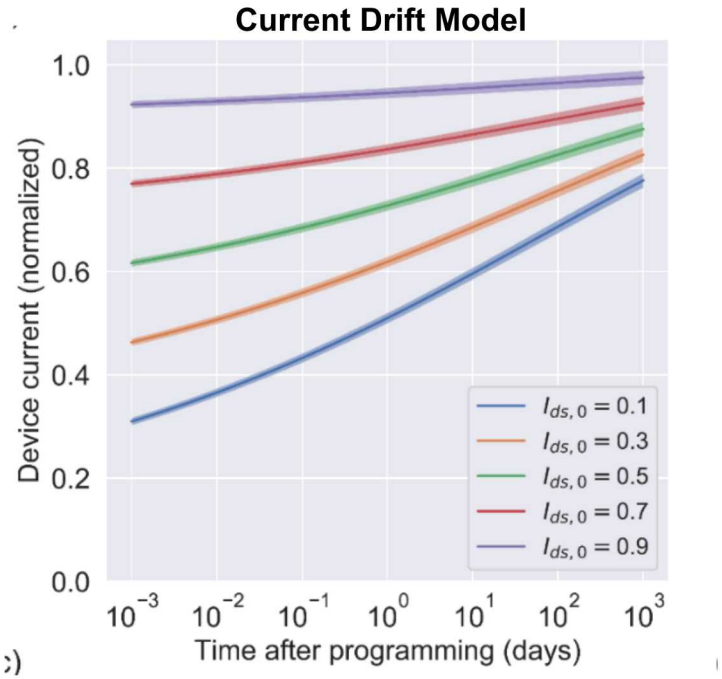


Courtesy DK Schroder

Modeling Effects of State Drift on SONOS Inference Accelerator

- SONOS will slowly lose charge and change from the programed to erased state (retention loss mechanism)
  - Clearly, this will affect the weights and inference accuracy
- Used simplified charge loss model to study affect on weights



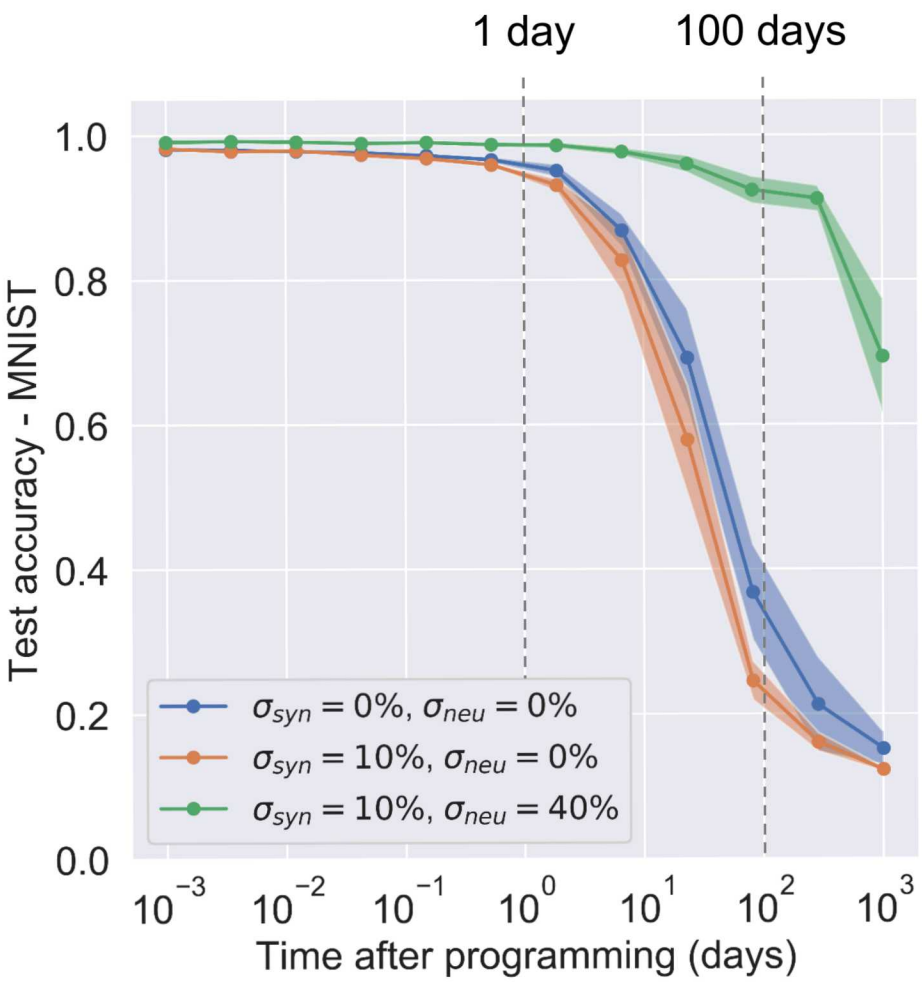


;)

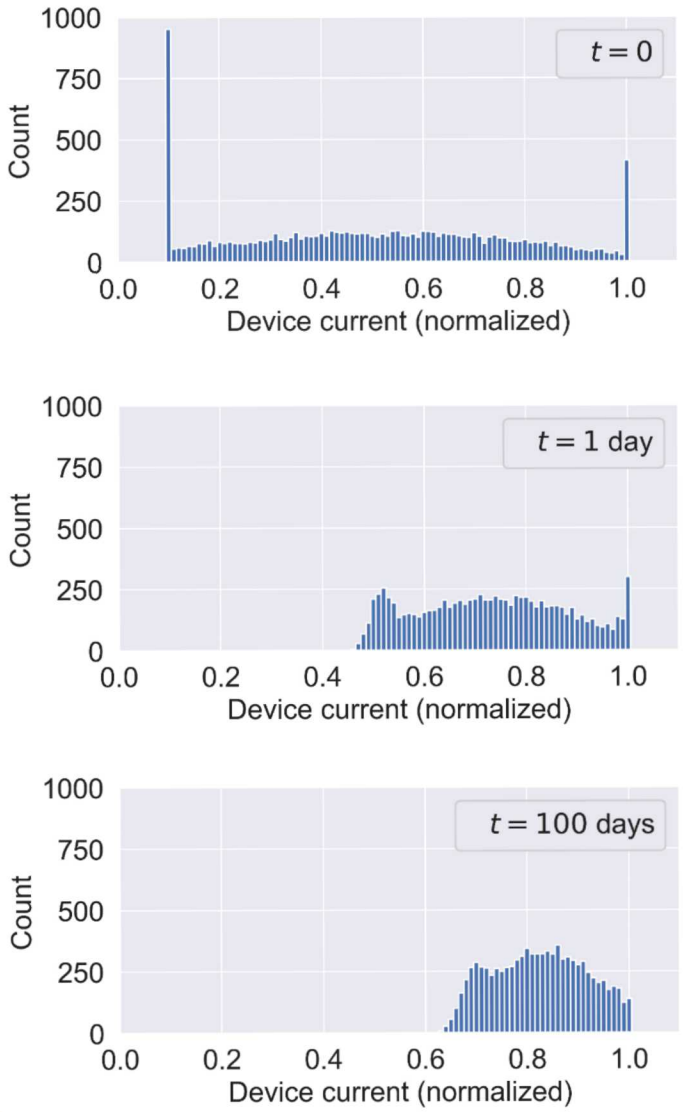
Modeling Effects of State Drift on SONOS Inference Accuracy

Sandia National Laboratories

Neural Network Performance Drift



Weight distribution drift:

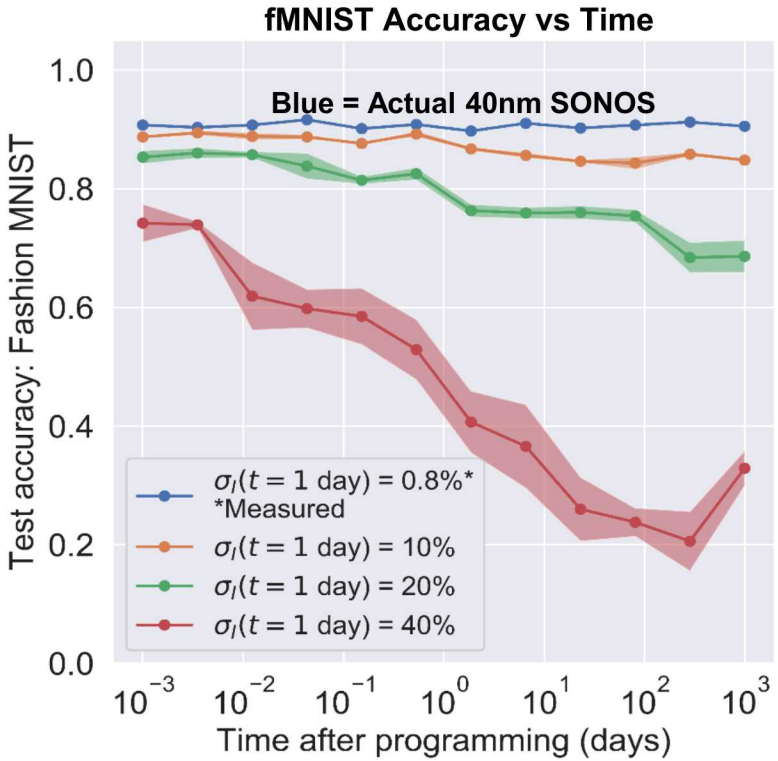


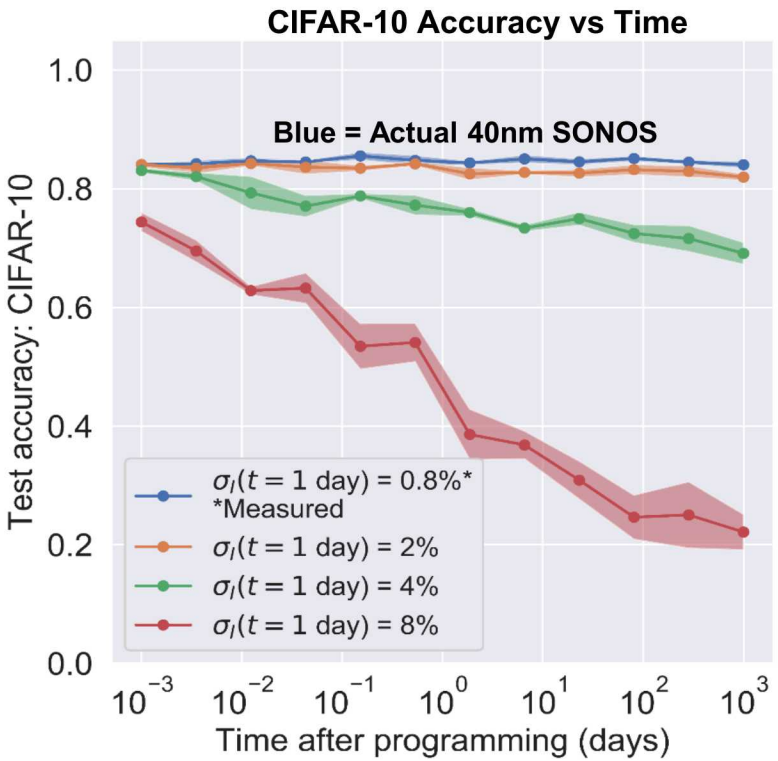
Modeling Effects of State Drift on SONOS Inference Accelerator

Sandia National Laboratories

Collaboration with CYPRESS EMBEDDED IN TOMORROW™

- Inference stability looks good for actual 40nm SONOS
  - Decay of 0.8% of dynamic range after 24hrs

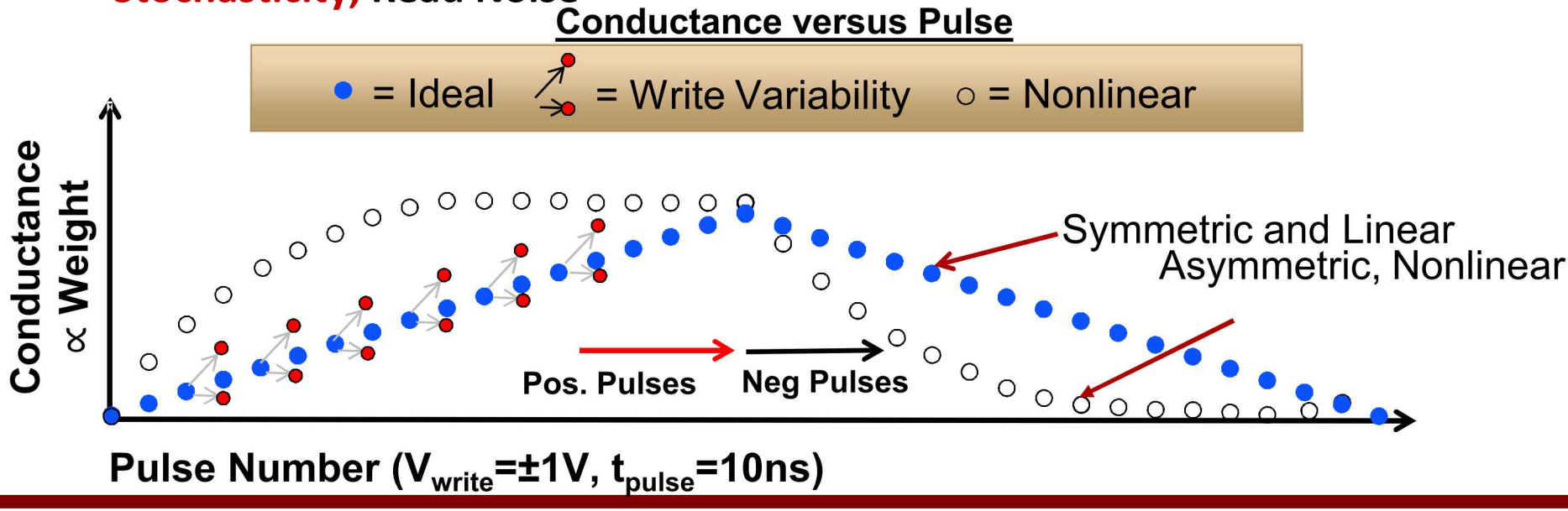




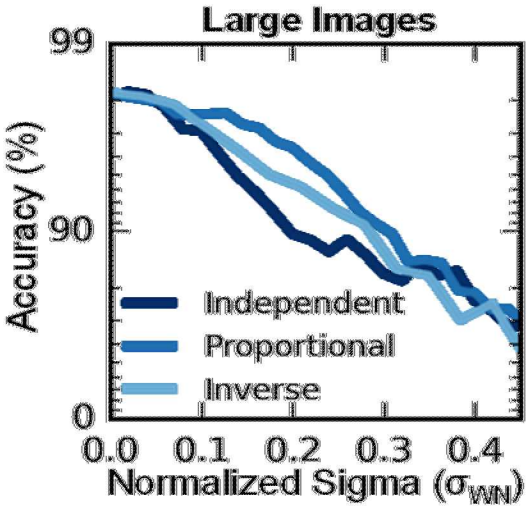
Bennett, Xiao, et al, IRPS 2020 (accepted, to be presented)

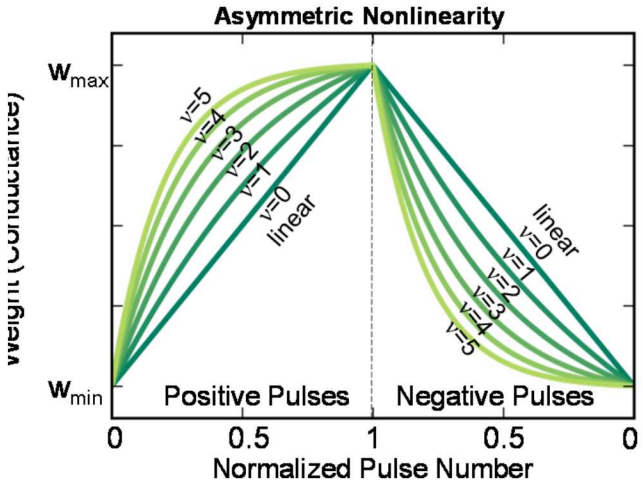
Experimental Device Analog Nonidealities: Training

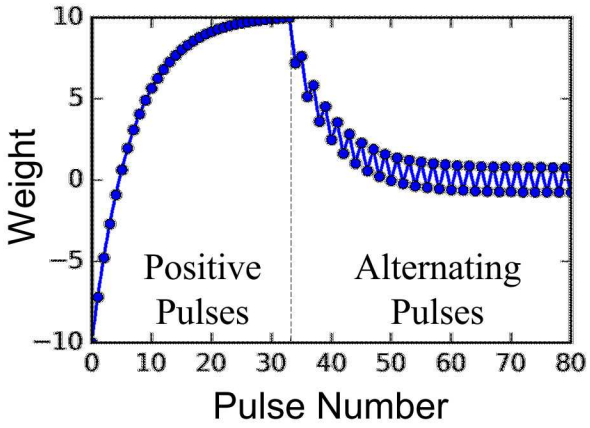
- **Analog architecture trade off digital precision for significant performance/watt and performance/area gains**
  - **Best suited for applications requiring ~8 bits precision or less**
- **Backprop training: Ideally weight would increase and decrease linearly proportional to learning rule result**
- **Key issue in experimental devices: altered the relationship between intended and actual update: Write Nonlinearity, Asymmetry, Stochasticity, Read Noise**



Device Requirements for Training

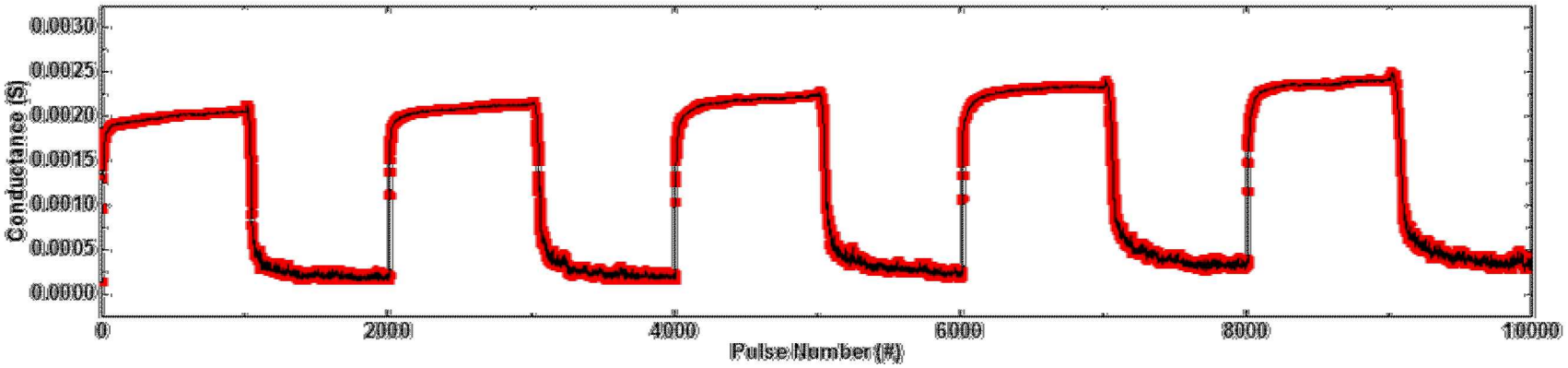


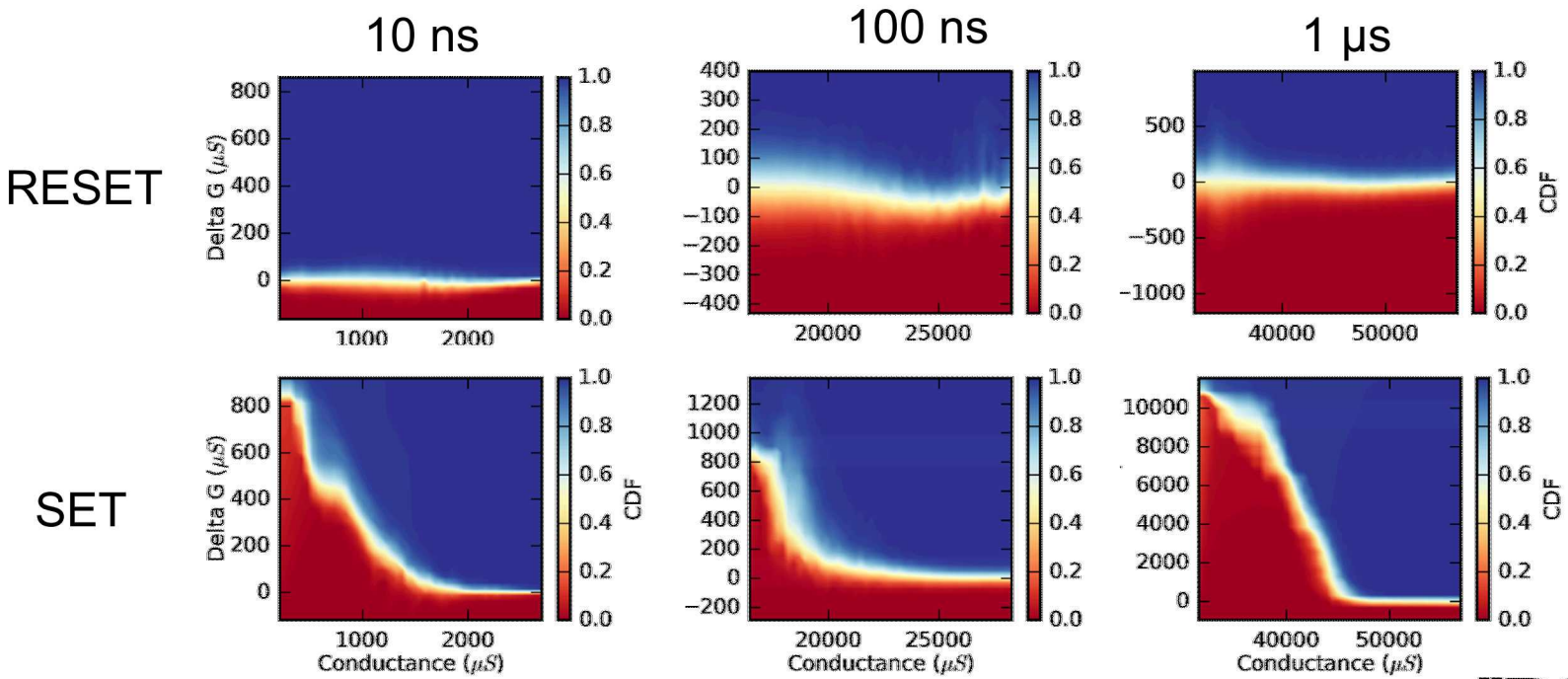




	Small Images	Large Images	File Types
Read Noise σ (% Range)	3%	5%	9%
Write Noise σ (% Range)	0.3%	0.4%	0.4%
Asymmetric Nonlinearity (ν)	0.1	0.1	0.1
Symmetric Nonlinearity (ν)	>20	5	5
Maximum Current	160 nA	13 nA	40 nA
Minimum Retention (@ 85°C)	7 days	7 days	7 days
Minimum Nudge Endurance	10^7	10^7	10^7

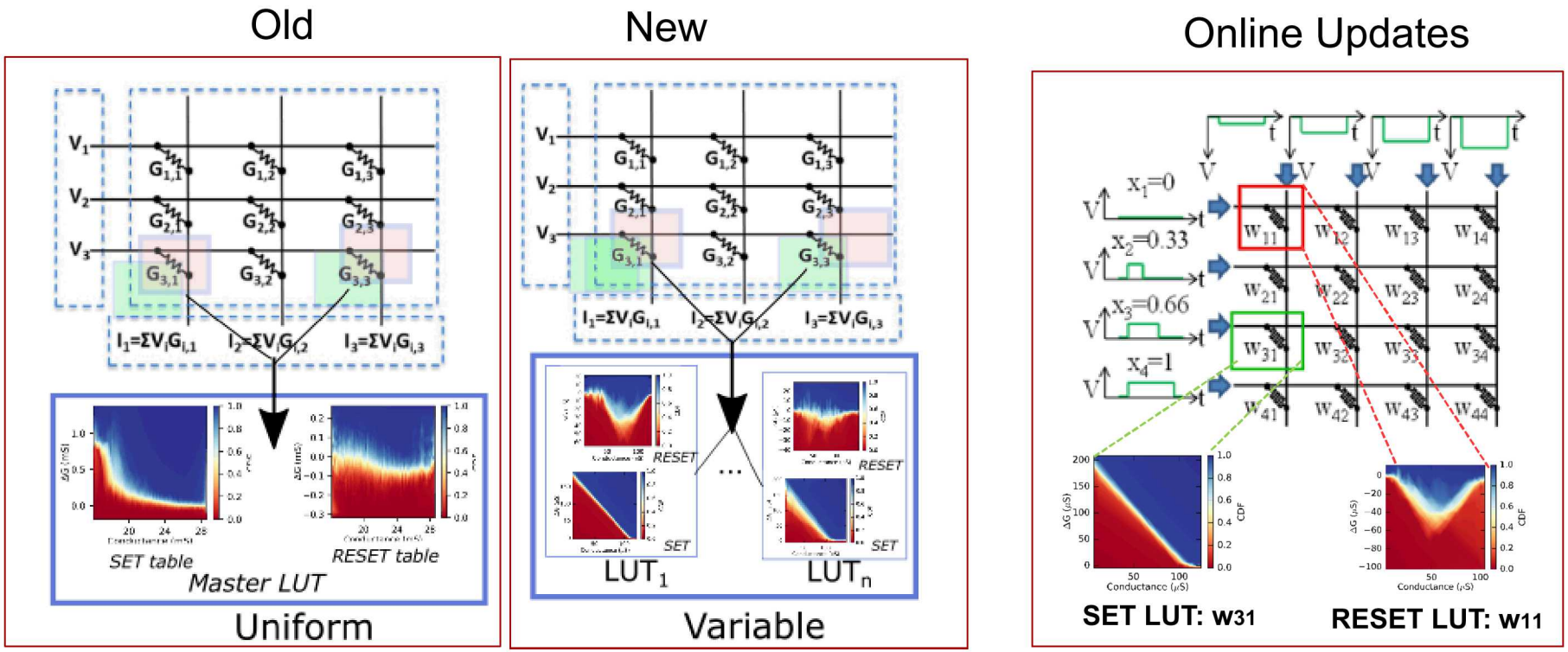
Training Lookup Table Generation





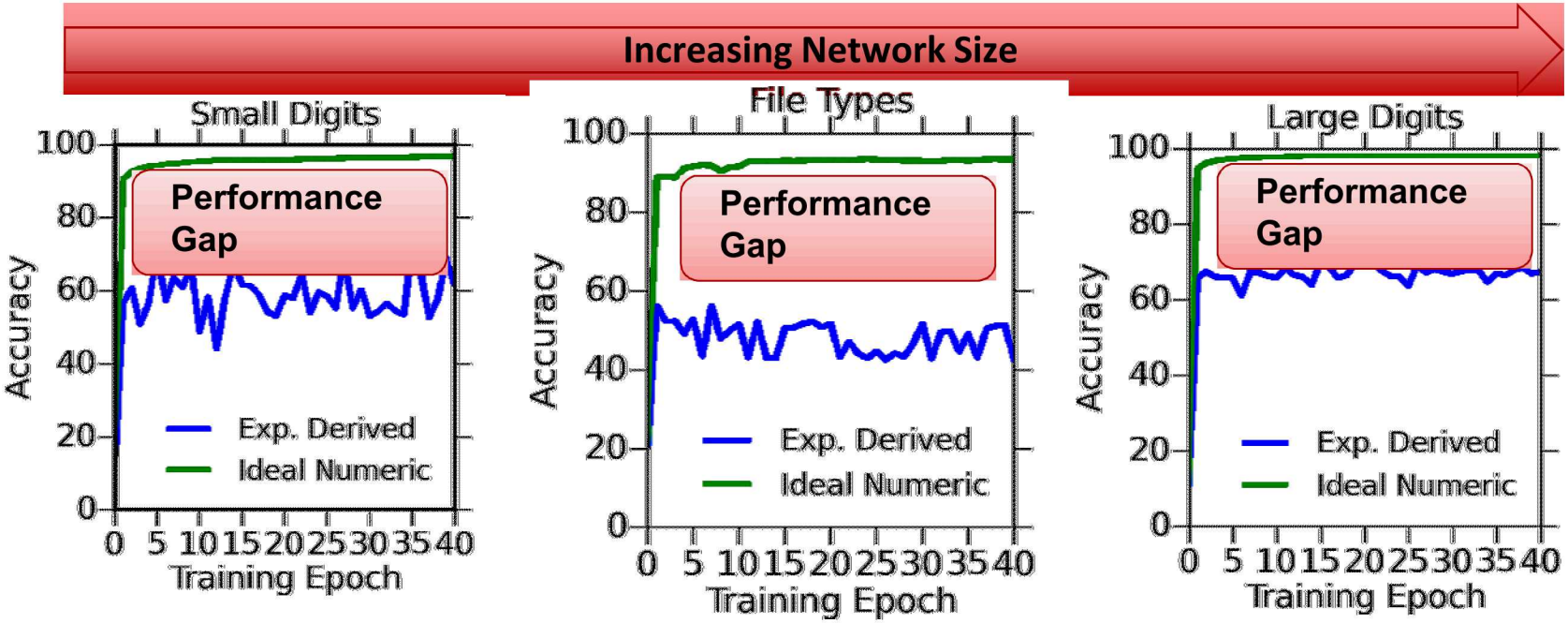
Device to Device Variation

- Extended the CrossSim platform to draw upon a library of look up tables (LUTs)
- LUTs assigned prior to online training and remain constant during it
- Individualized updates -> slowdown (somewhat mitigated)



CrossSim Model of TaOx ReRAM: MNIST, Backprop Training

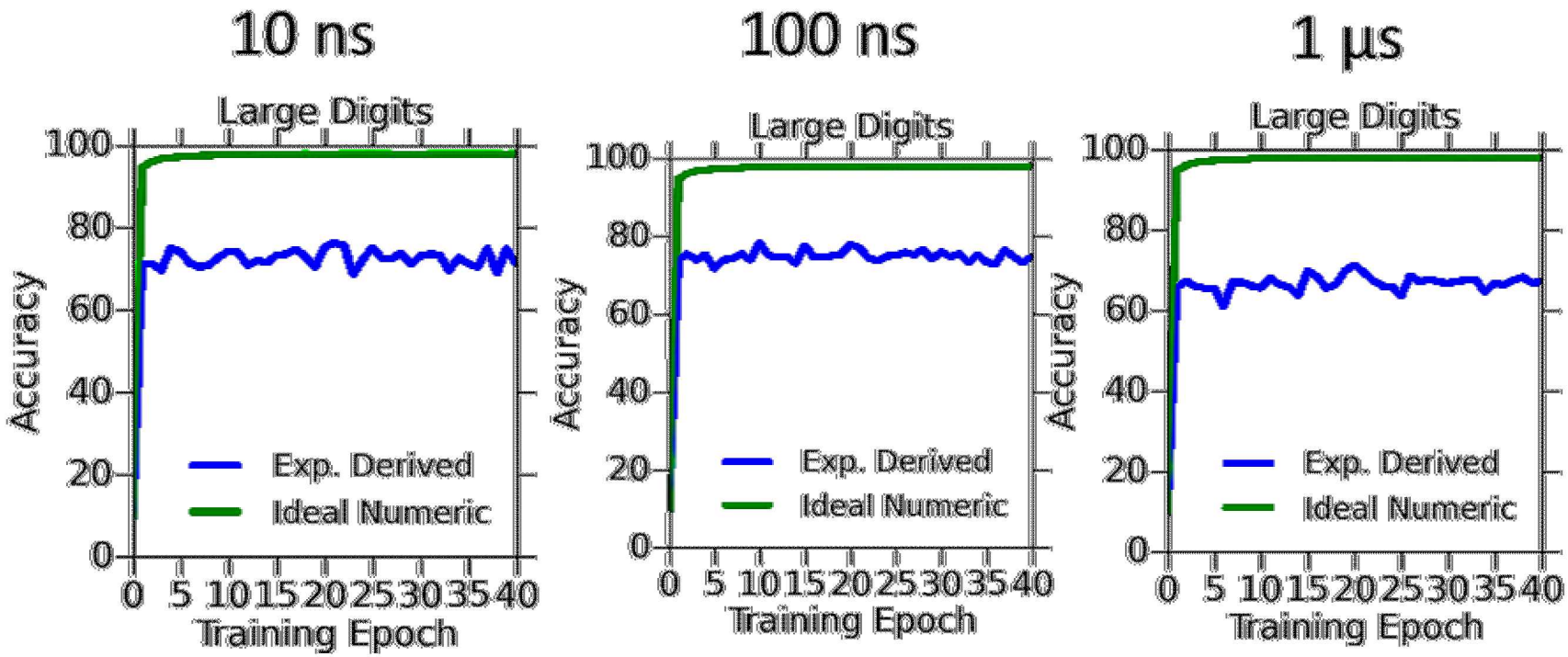
Sandia National Laboratories



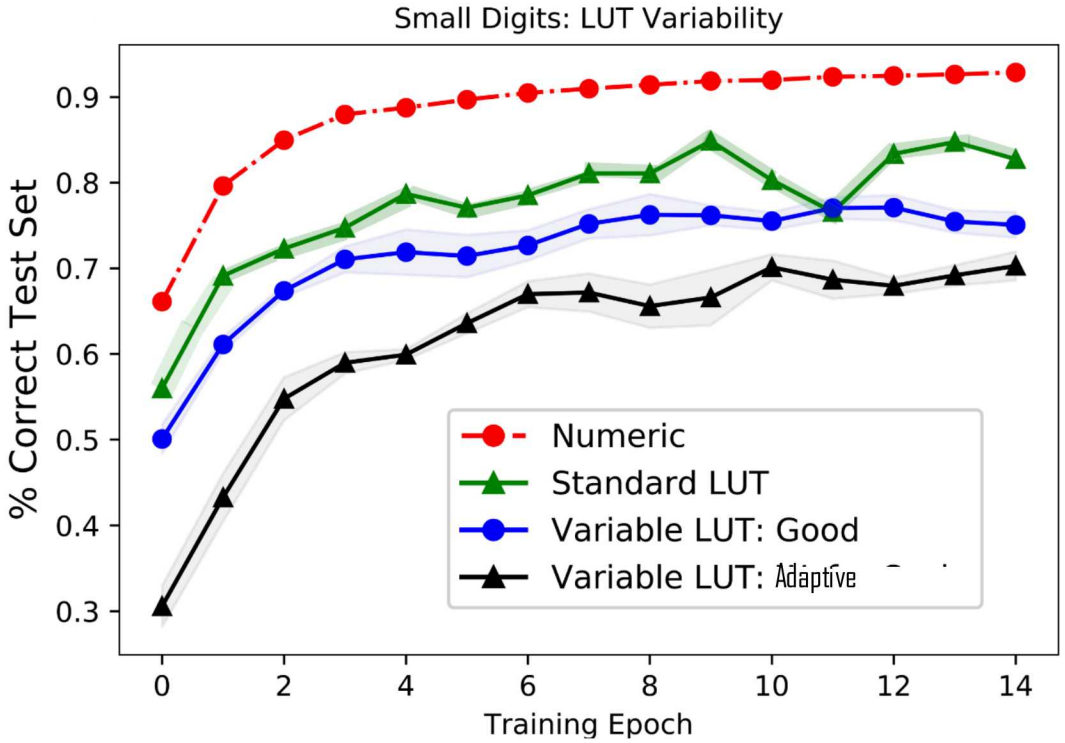
Data set	# Training Examples	# Test Examples	Network Size
UCI Small Digits[1]	3,823	1,797	64×36×10
File Types[2]	4,501	900	256×512×9
MNIST Large Digits[3]	60,000	10,000	784×300×10

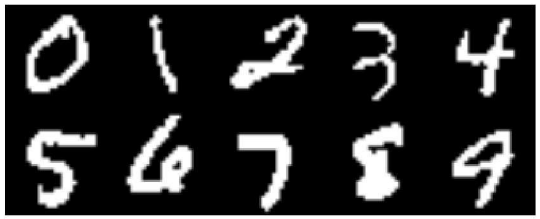
CROSS SIM

Pulse Width Optimization



ReRAM algorithm accuracies:
Small Digits (OCR Database)



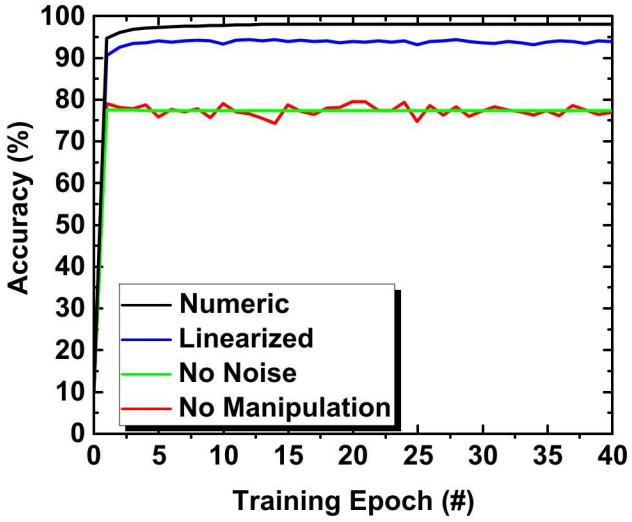


Data set	# Training Examples	# Test Examples	Network Size
UCI Iris Dataset [1]	100	50	4x8x3
UCI Small Digits[1]	3,823	1,797	64×36×10
File Types[2]	4,501	900	256×512×9
MNIST Large Digits[3]	60,000	10,000	784×300×10

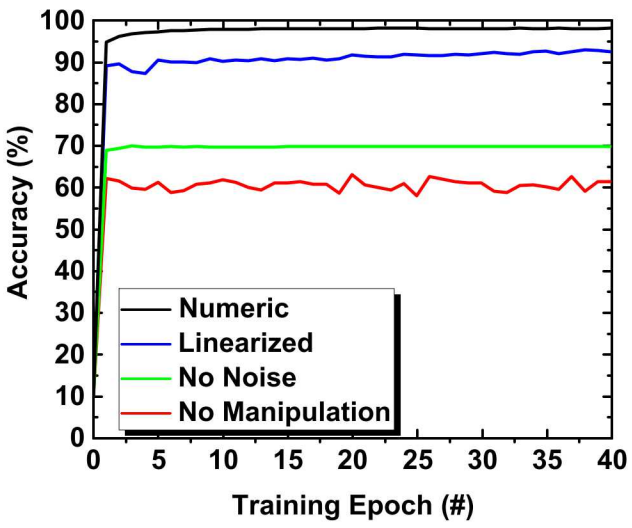
Comparison of Filamentary ReRAM

- TaO_x-Ta highest training classification accuracy initially
- Analyze effect of noise and nonlinearity on accuracy with CrossSim
- Nonlinearity is an inherent issue for each filamentary device

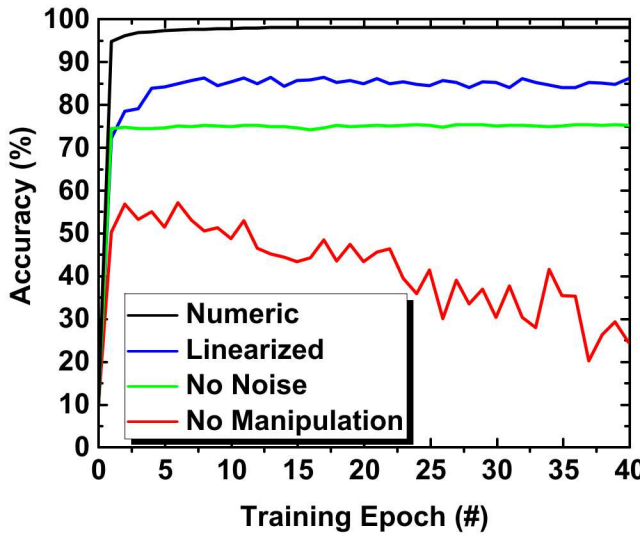
Sandia Baseline TaO_x-Ta



SiO_2-Cu



Ag-Chalcogenide



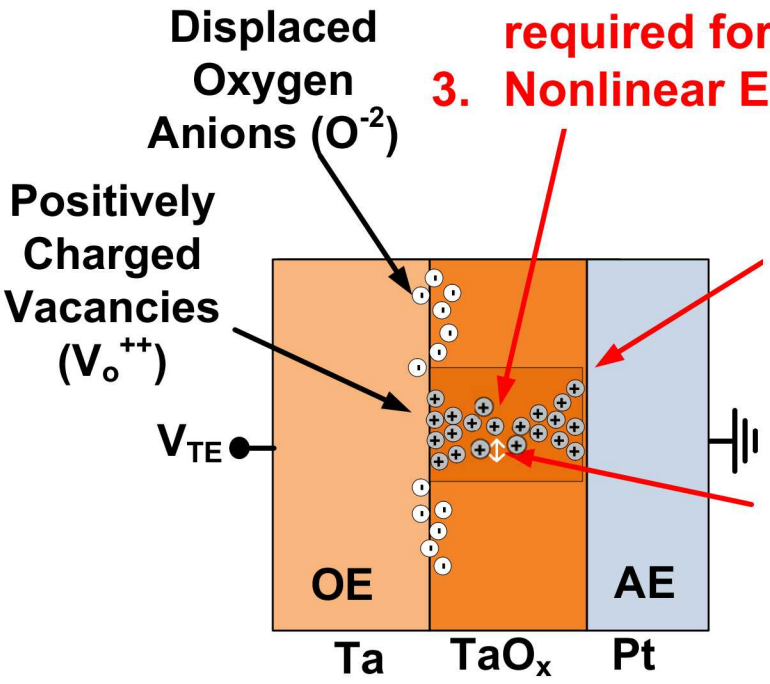
R. Jacobs-Gedrim et al, *IEEE-ICRC*, 2017

Problems with Filamentary Mechanism for Analog Training

Sandia National Laboratories

Nonlinearity

1. Tunneling current, esp in high resistances
2. Current crowding – high temperature required for change give runaway effect
3. Nonlinear E-field

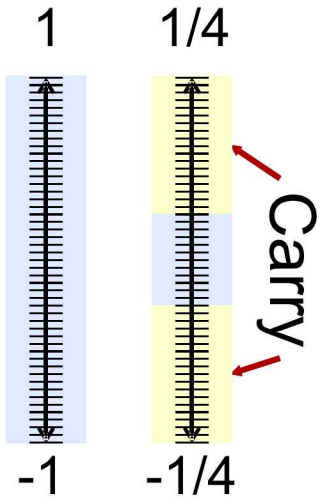


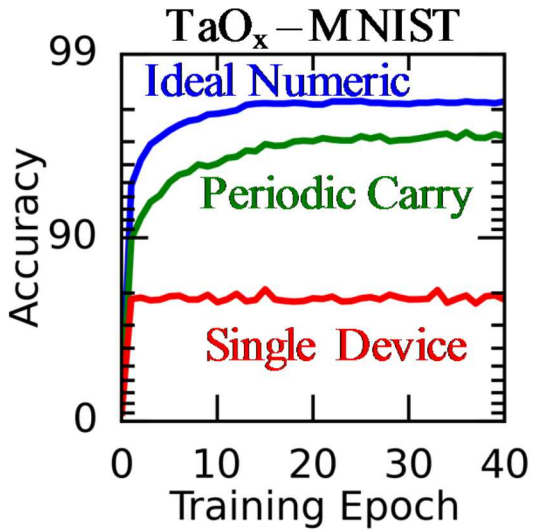
Asymmetry
Inherent property of bipolar device – Schottky-like and ohmic junctions

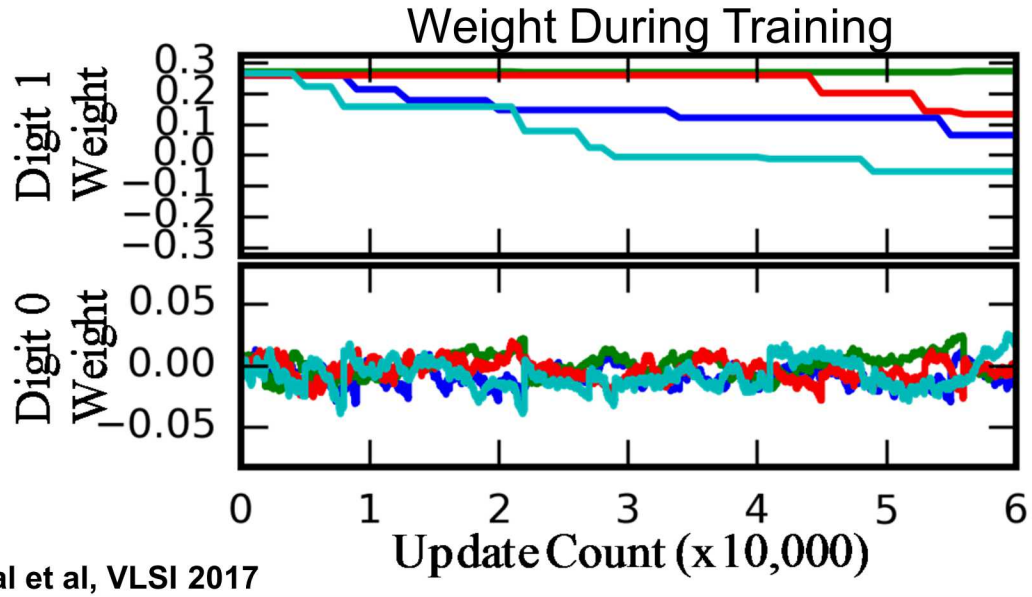
Stochasticity
G depends on position of a few atoms

Period Carry: Multi-Synapse Method High Accuracy in Nonideal Devices

- Example for TaOx: Each weight uses 2 synapses
- Carry once every 1000 updates for the LSB, and every 2 updates on others
- High variability TaOx device: accuracy improves from <89% to >97%, only ~1% under ideal
- Requires tradeoff of energy/latency for accuracy – exact tradeoff depends on algorithm reqs.

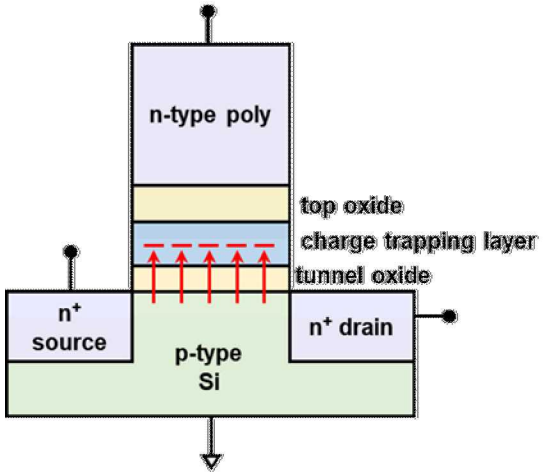


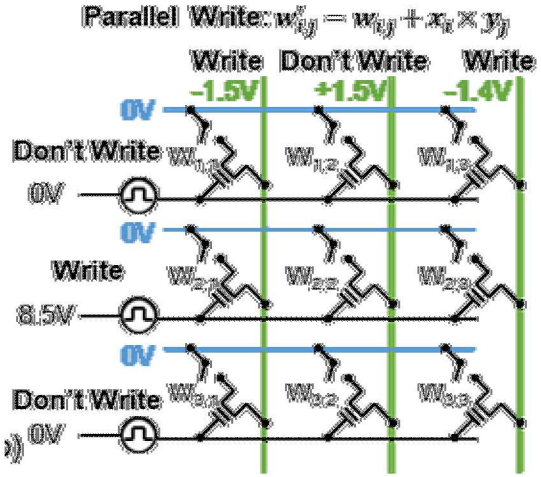


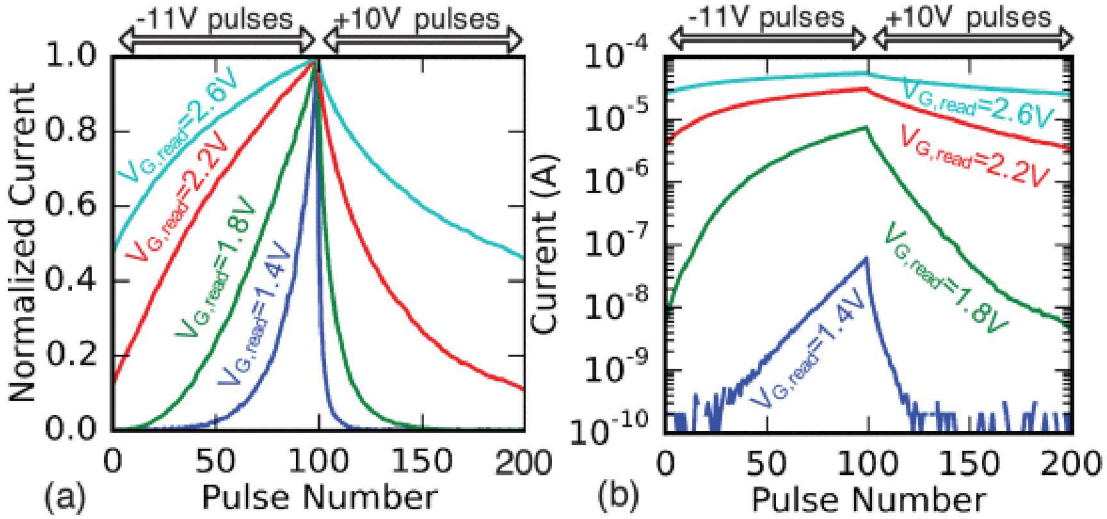


Agarwal et al, VLSI 2017

SONOS Training

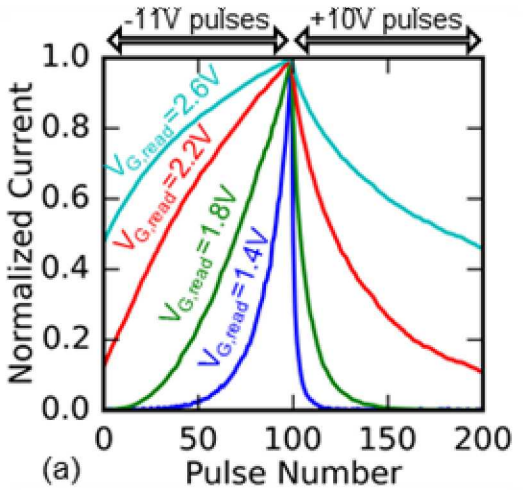


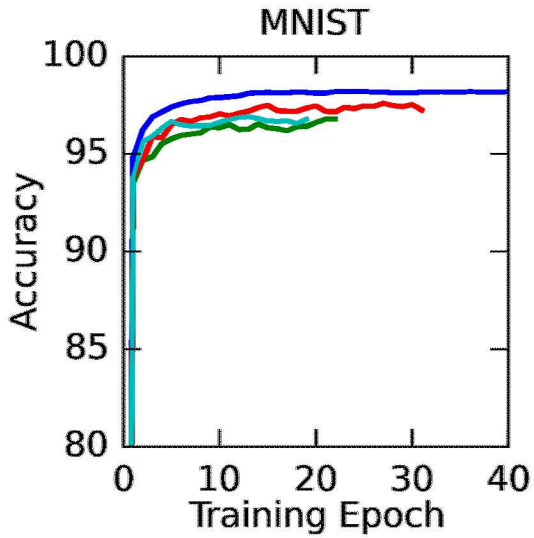




S Agarwal et al, IEEE J Exploratory Solid-State Computational Devices and Circuits, 5, 52-57, 2019.

SONOS Accuracy and Energy





Component	Vector Matrix Multiply	Matrix Vector Multiply	Outer Product Update
Energy/Op ReRAM (fJ)	12.2	12.2	2.1
Energy/Op SONOS (fJ)	13.7	13.7	68.2
Energy/Op SRAM (fJ)	2718	4630	4102
Array Latency ReRAM (µs)	0.38	0.38	0.51
Array Latency SONOS (µs)	0.40	0.40	20
Array Latency SRAM (µs)	4	32	8

Low inference energy

Slow write

S Agarwal et al, IEEE J Exploratory Solid-State Computational Devices and Circuits, 2019.

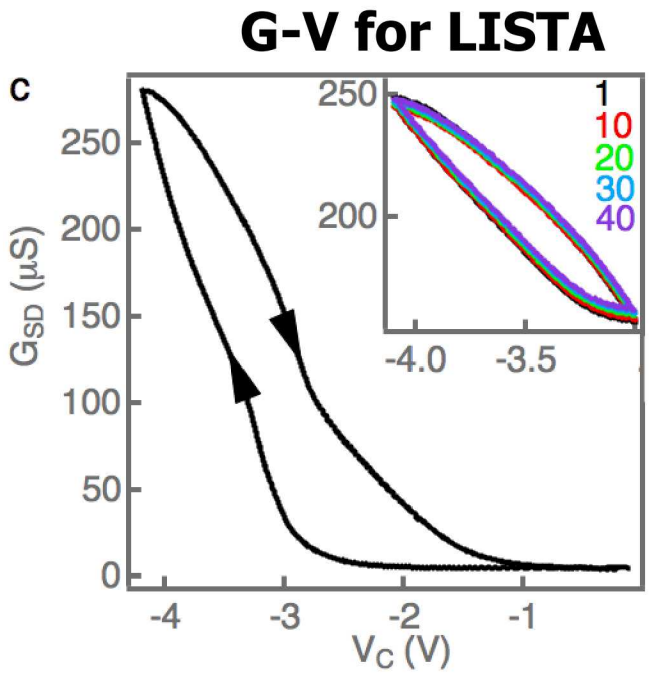
Outline

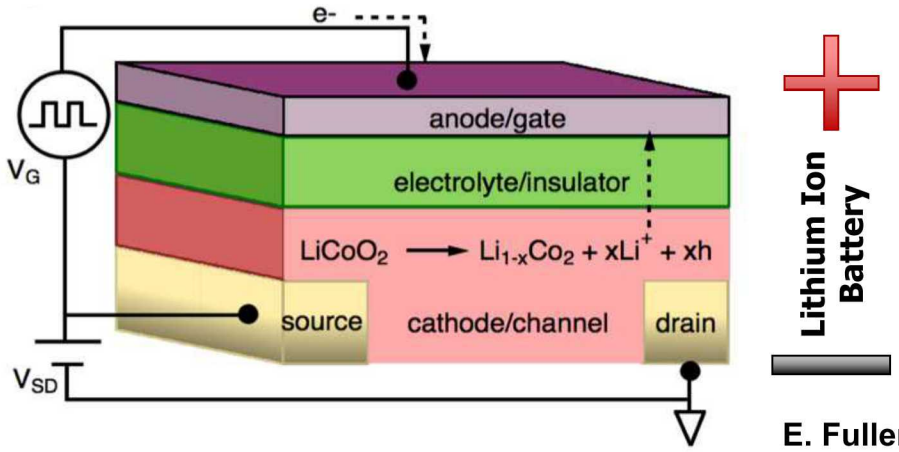
Sandia National Laboratories

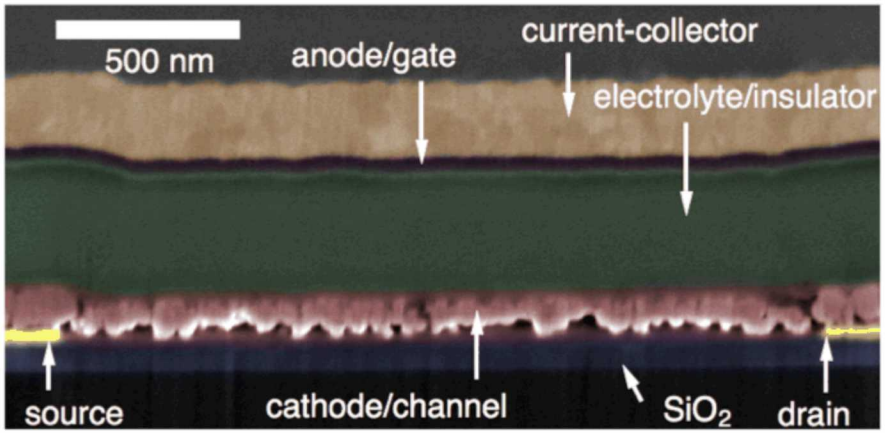
- Motivation and Background
- Analog Accelerator Concepts & Performance Model
- Accelerator Accuracy Modeling & Characterization
- Future Directions
- Radiation Effects
- Summary

Li-Ion Synaptic Transistor for Analog Computation (LISTA)

- Alternatively, novel devices may offer promise
- LISTA: modulate the doping of Lithium battery cathode
- Resistivity across cathode changes linearly with battery charge/discharge

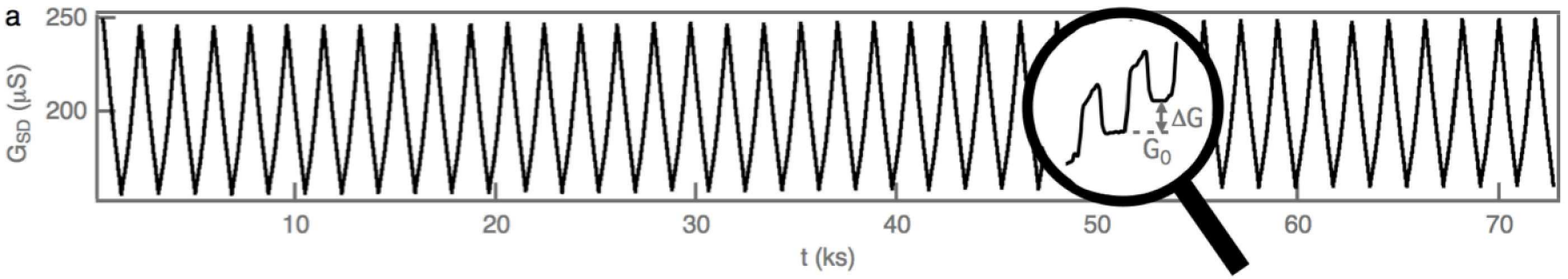




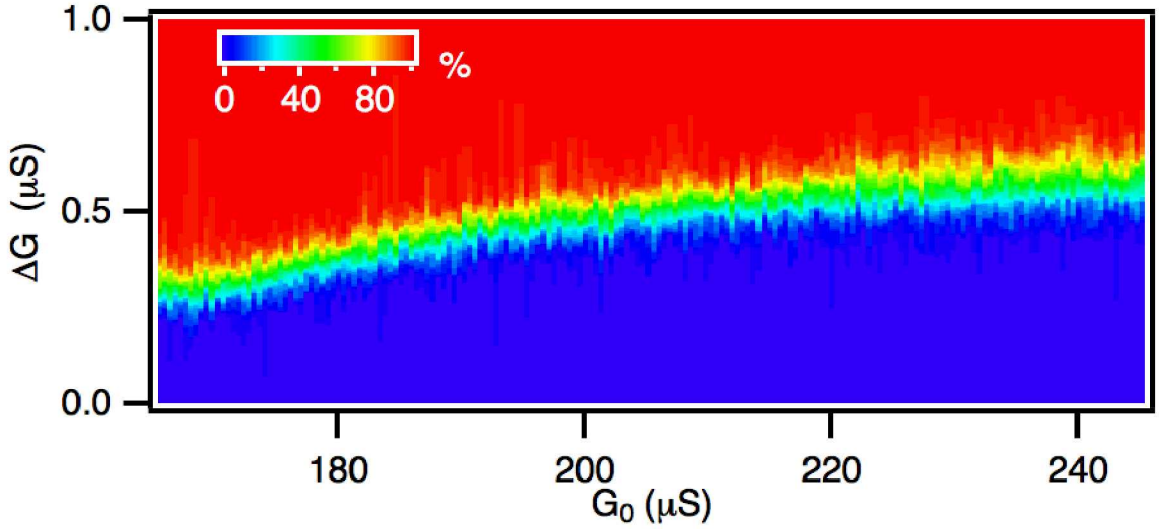


E. Fuller et al, *Adv Mater*, 2017

Analog State Characterization

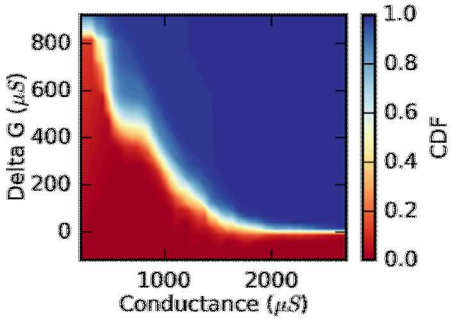


LISTA > 200 states

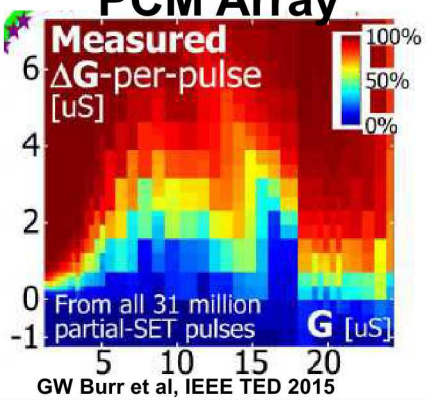


E. Fuller et al, *Adv Mater*, 2017

TaOx ReRAM



PCM Array

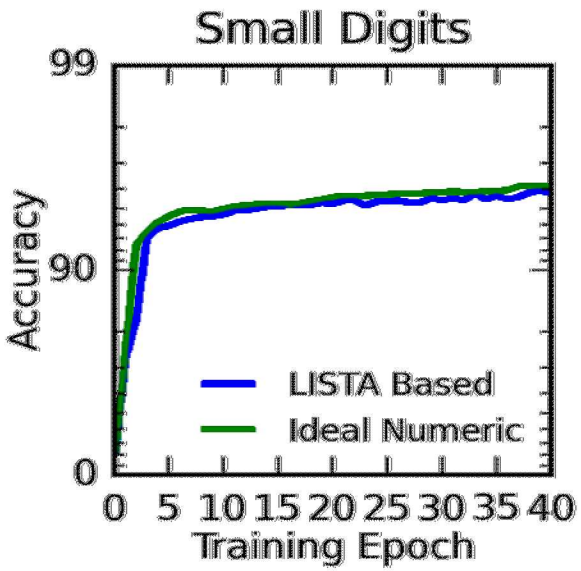


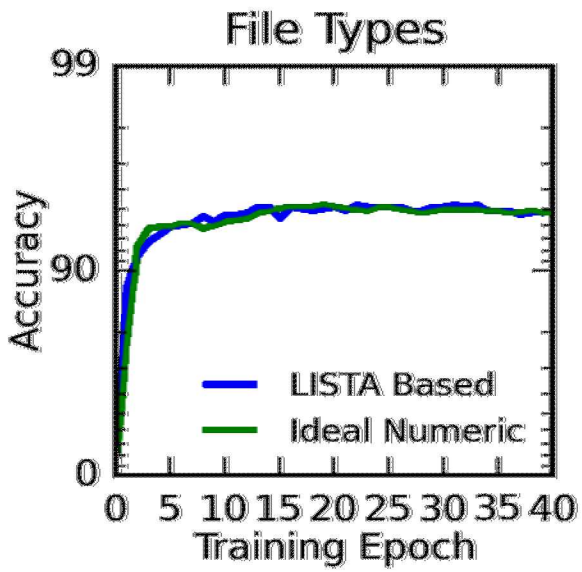
GW Burr et al, IEEE TED 2015

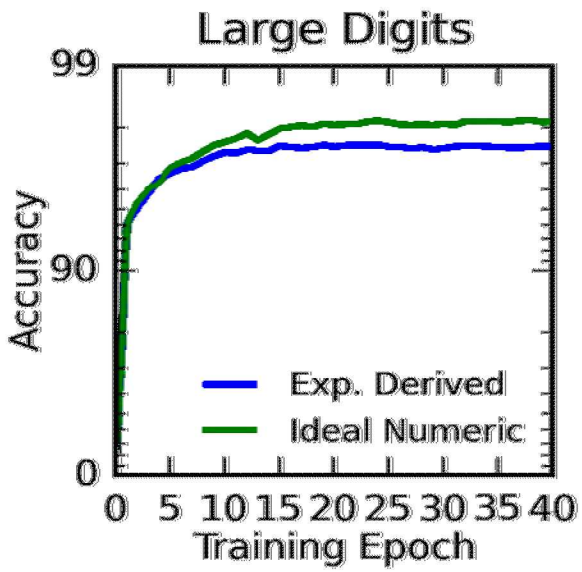
LISTA-device Performance for Backprop Algorithm

Sandia National Laboratories

Increasing Network Size



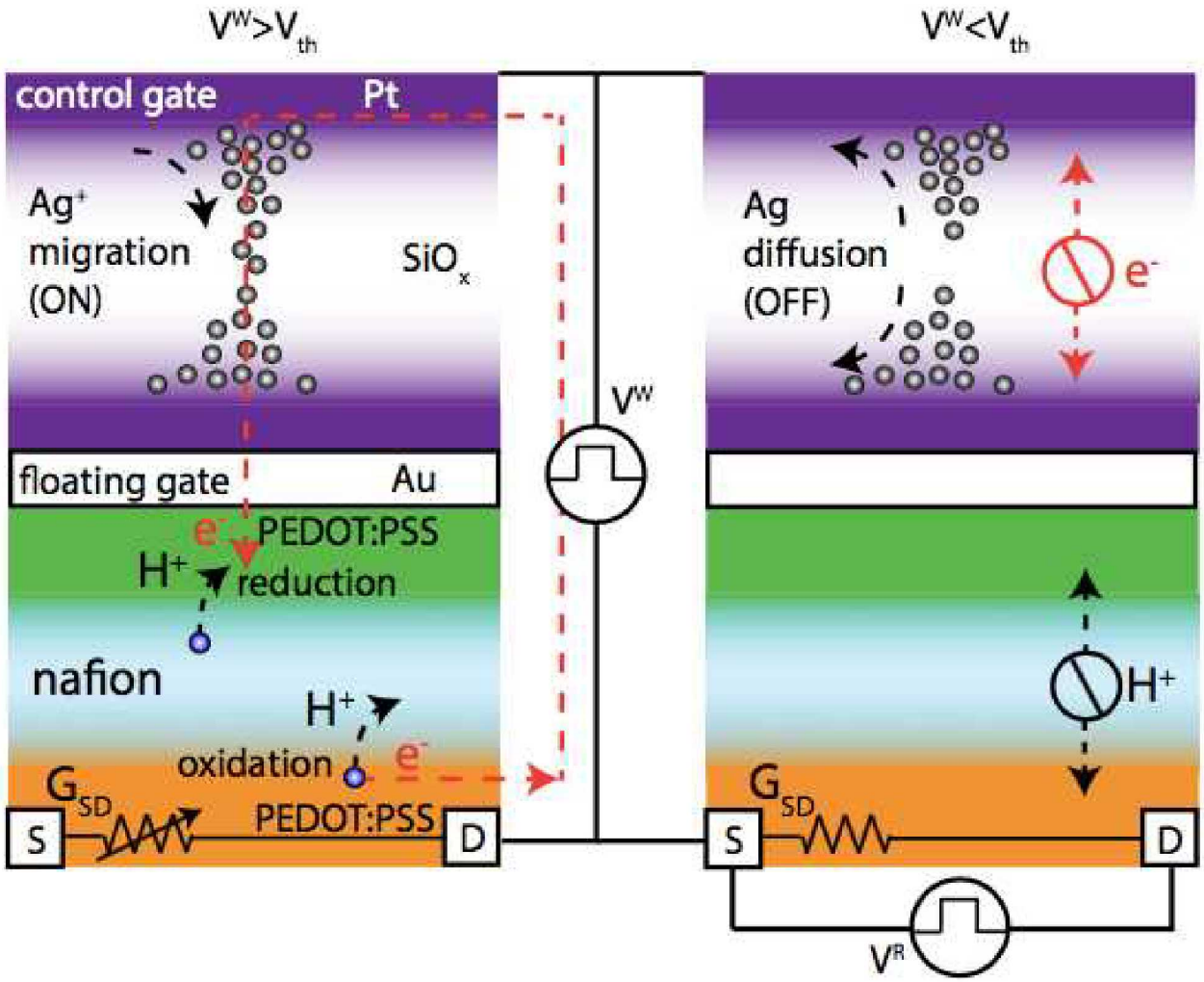




Data set	# Training Examples	# Test Examples	Network Size
UCI Small Digits[1]	3,823	1,797	64×36×10
File Types[2]	4,501	900	256×512×9
MNIST Large Digits[3]	60,000	10,000	784×300×10

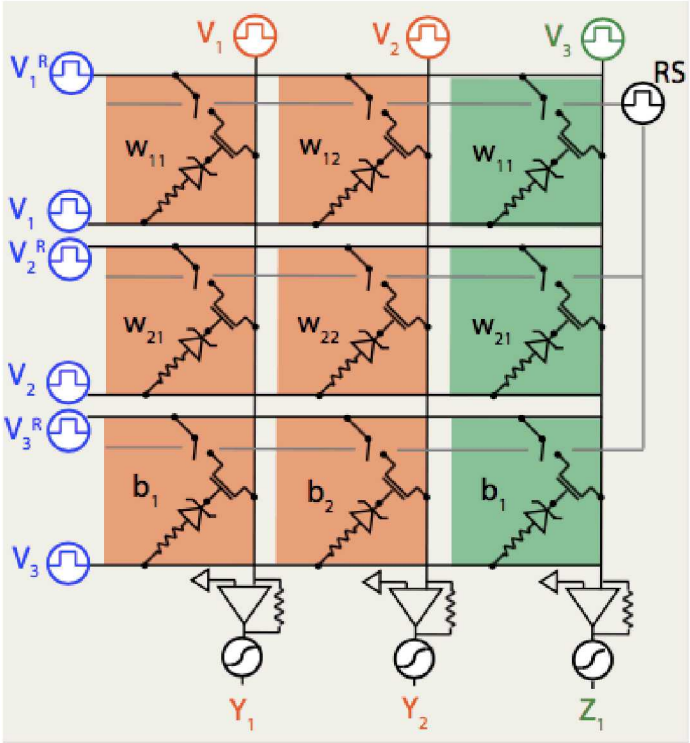
E. Fuller et al, *Adv Mater*, 2017

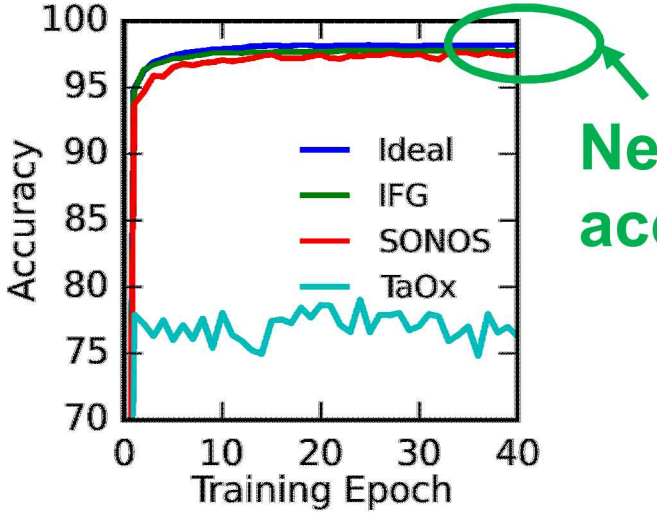
Ionic Floating Gate



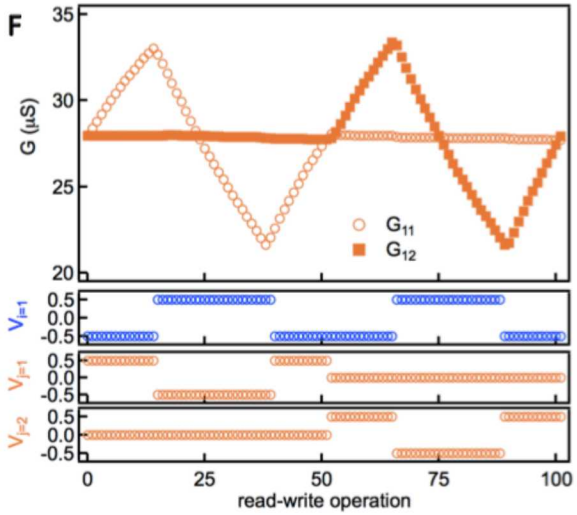
E. J. Fuller, S. T. Keene, A. Melianas, Z. Wang, S. Agarwal, Y. Li, Y. Tuchman, C. D. James, M. J. Marinella, J. J. Yang, A. Salleo, A. A. Talin, *Science* 364, 570, (2019).

IFG Array Demonstration



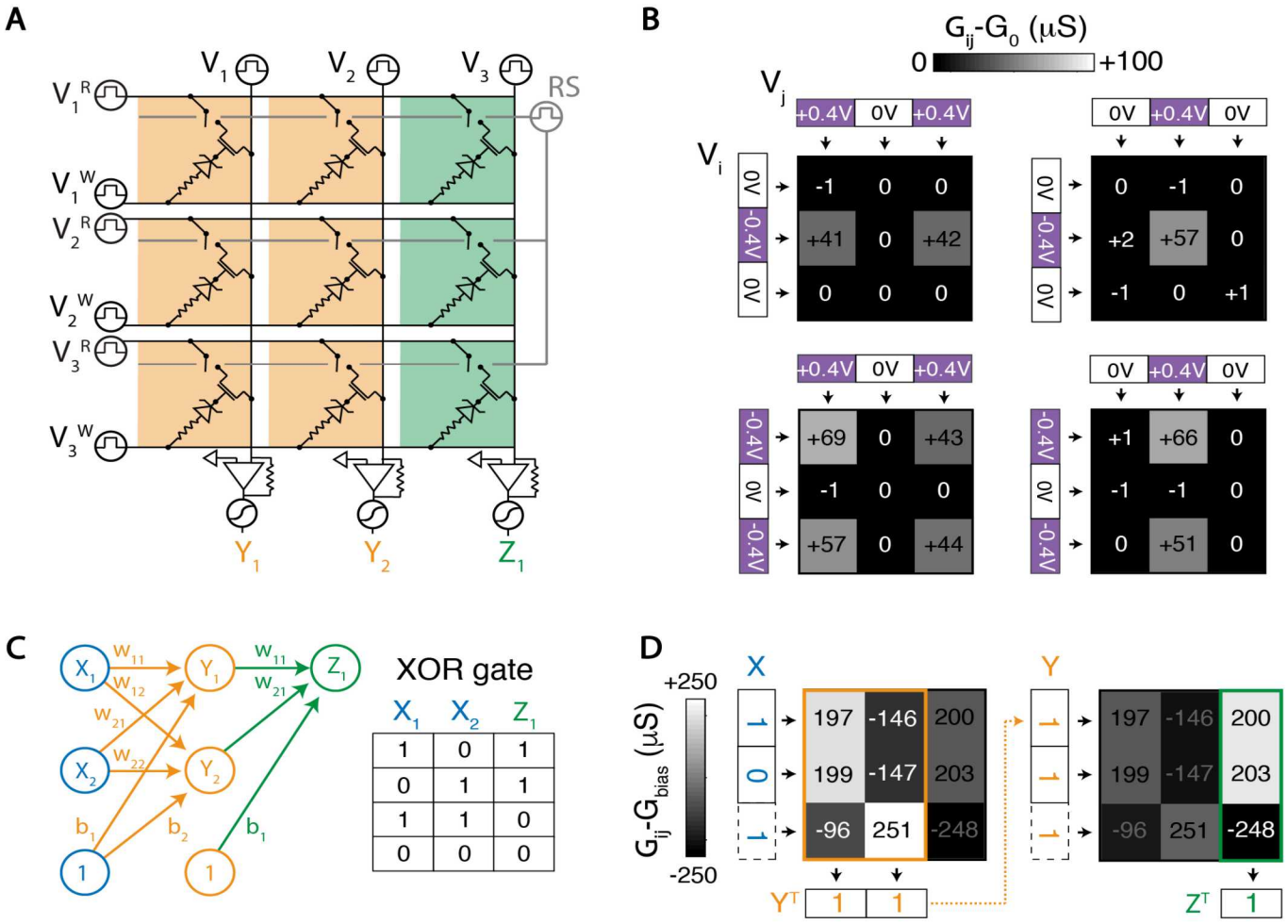


Near ideal accuracy



E. J. Fuller, S. T. Keene, A. Melianas, Z. Wang, S. Agarwal, Y. Li, Y. Tuchman, C. D. James, M. J. Marinella, J. J. Yang, A. Salleo, A. A. Talin, *Science* 364, 570, (2019).

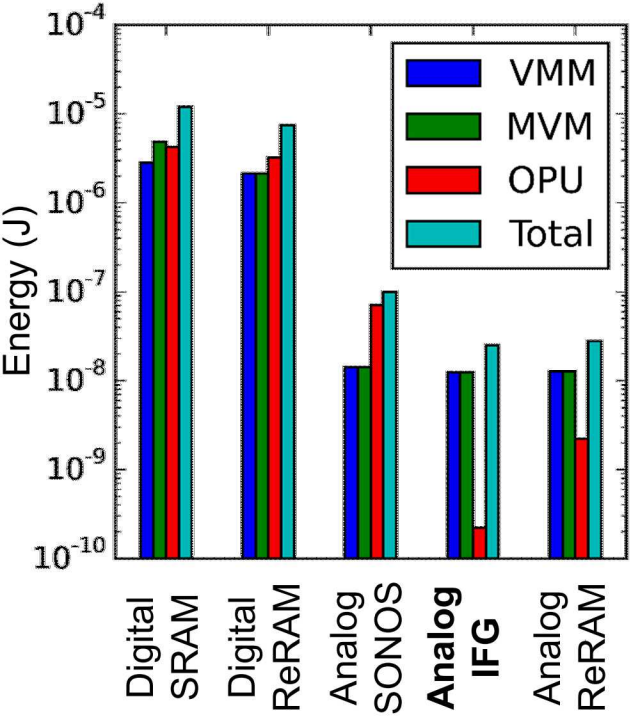
Programming Demonstration



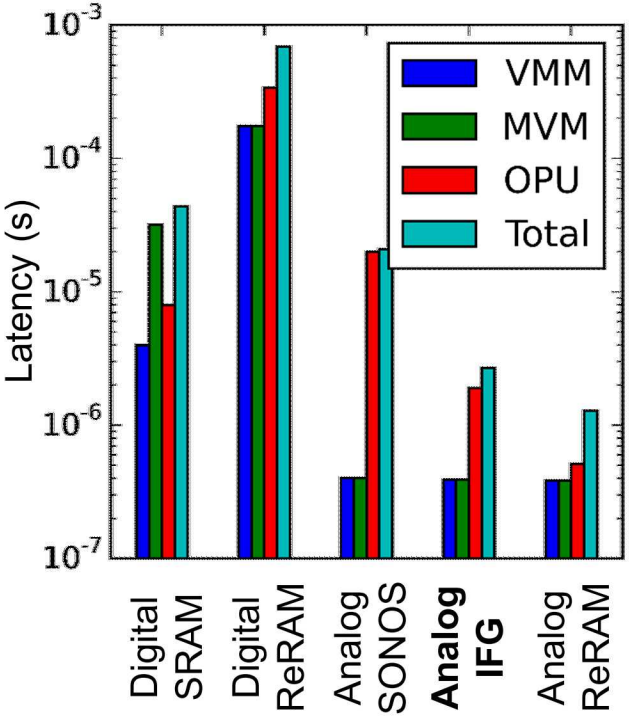
E. J. Fuller, S. T. Keene, A. Melianas, Z. Wang, S. Agarwal, Y. Li, Y. Tuchman, C. D. James, M. J. Marinella, J. J. Yang, A. Salleo, A. A. Talin, *Science* 364, 570, (2019).

Ionic Floating Gate Energy

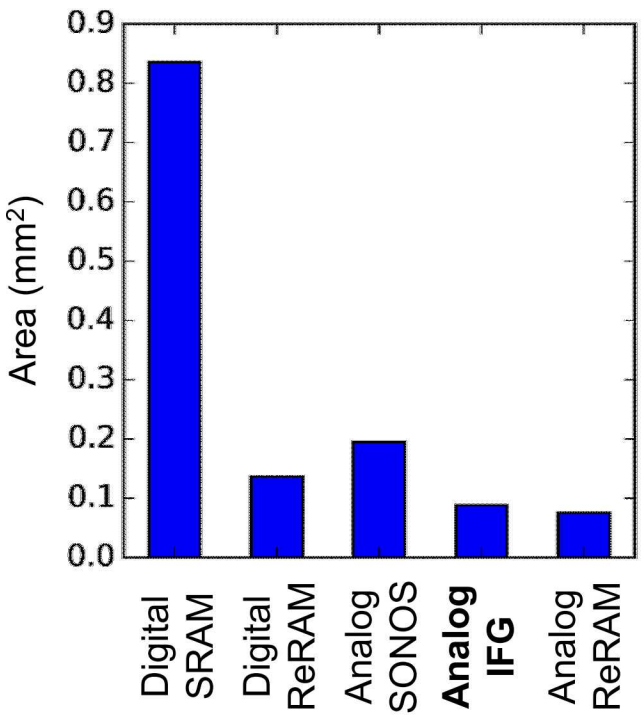
120-430X Energy Advantage



2-34X Latency Advantage



5-11X Area Advantage



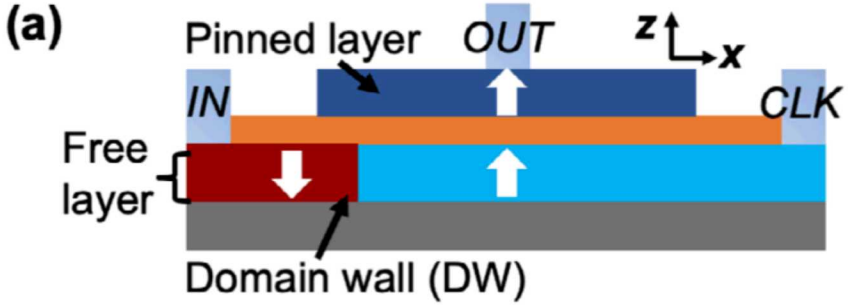
1024 x1024 = 1M array operations, sum over 1 training cycle, 3 operations:

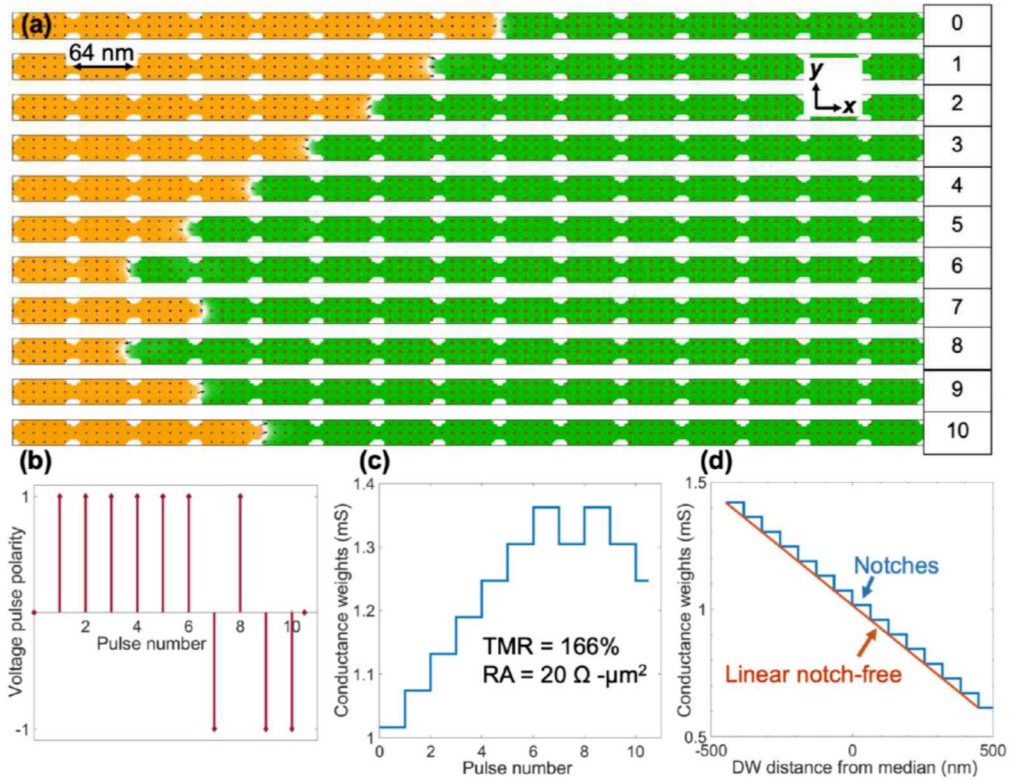
- Vector Matrix Multiply - Matrix Vector Multiply - Outer Product Update

Used a commercial 14/16 nm PDK ***Requires 100 MΩ on state devices

Agarwal et al, VLSI-TSA 2019

Magnetic Synapse Concept





Otitoaleke Akinola, Xuan Hu, Christopher H Bennett, Matthew Marinella, Joseph S Friedman, Jean Anne C Incorvia J. Phys. D, 2019.

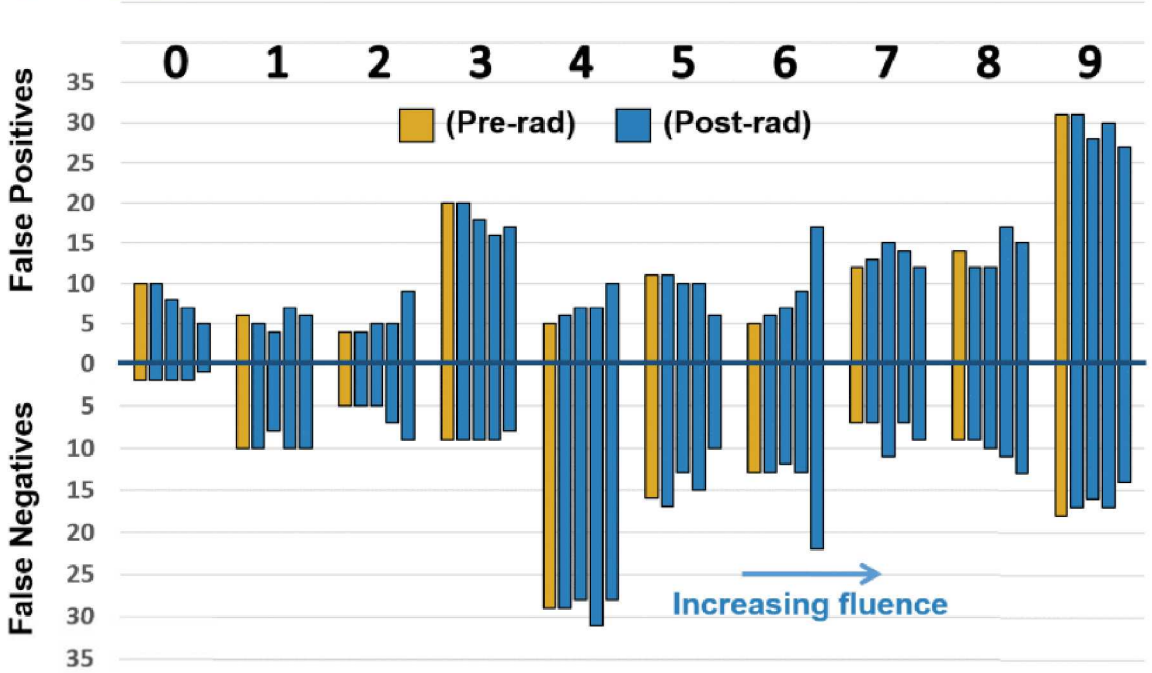
Outline

Sandia National Laboratories

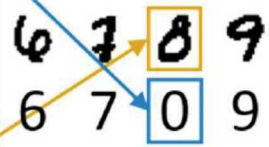
- Motivation and Background
- Analog Accelerator Concepts & Performance Model
- Accelerator Accuracy Modeling & Characterization
- Future Directions
- Radiation Effects
- Summary

Effects of Protons on TrueNorth Digital ASIC Inference

- MNIST inference on IBM TrueNorth ASIC
- Some SRAM flips occur
- Affects different digits differently
- On average accuracy not significantly affected



Digit	Number of False Negatives	Number of False Positives
0	2	10
1	10	6
2	5	4
3	9	20
4	29	5
5	16	11
6	13	5
7	7	12
8	9	14
9	18	31



R.M. Brewer et al, *TNS 67, 108-115, 2019*

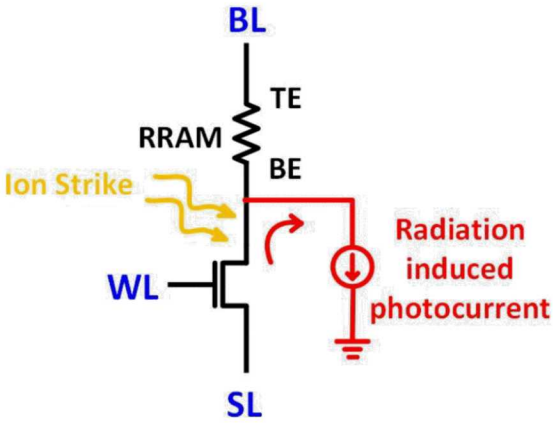
Effect of Single Events Upsets on Inference – ReRAM

- Based on SPICE model of SEU in 1T1R cell
- SEU has little effect on inference
- RRAM less susceptible to SEU than SRAM

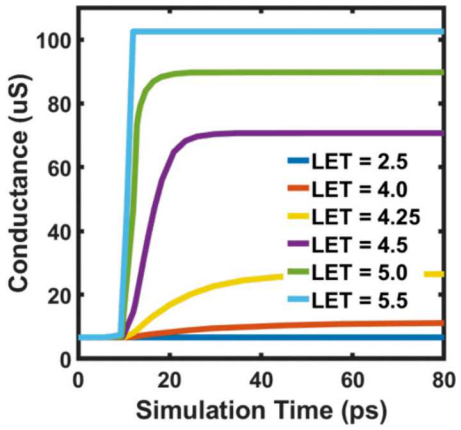
Effects of SEU: Smaller Network Topology

Simulation Variables	Value of Variables	Total Ion Number	Accuracy (%)
Operation years	0	0	97.26
	10	3	97.26
	20	6	97.26
	30	9	97.25
Technology node	45 nm	1	97.26
	65 nm	3	97.26
	130 nm	12	97.24
BL voltage	0.5 V	3	97.26
	0.7 V	3	97.26
	0.9 V	3	97.26
RRAM Switching voltage	0.55 V	3	97.25
	0.90 V	3	97.26
	1.25 V	3	97.26

1T1R Circuit Model



RRAM Conductance Transient



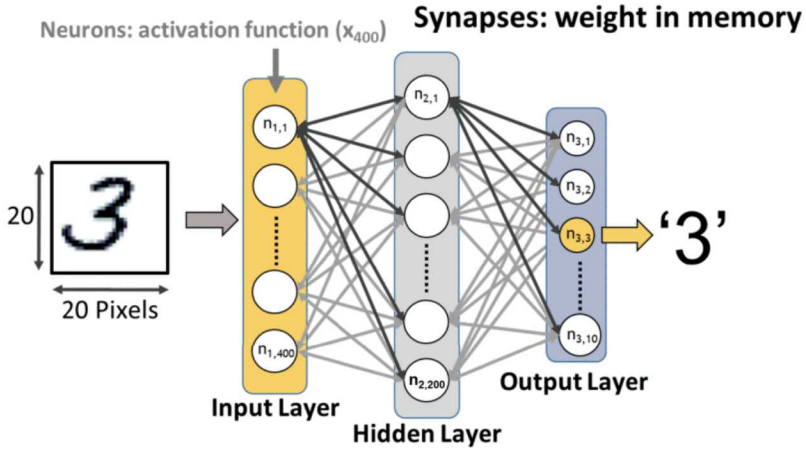
Effects of SEU: Larger Network Topology

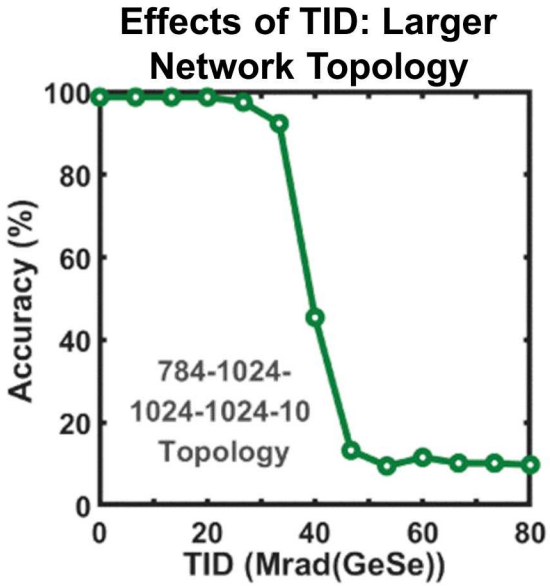
Simulation Variables	Value of Variables	Total Ion Number	Accuracy (%)
Operation years	0	0	98.74
	10	90	98.73
	20	180	98.72
	30	270	98.70
Technology node	45 nm	43	98.74
	65 nm	90	98.73
	130 nm	360	98.67
BL voltage	0.5 V	90	98.73
	0.7 V	90	98.74
	0.9 V	90	98.74
RRAM Switching voltage	0.55 V	90	98.66
	0.90 V	90	98.70
	1.25 V	90	98.73

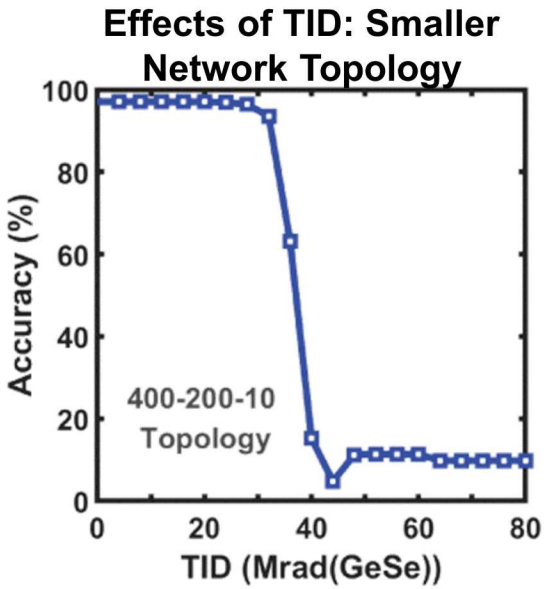
Ye et al, *TNS 66, 97-103, Jan 2019*

Effect of Total Ionizing Dose on Inference – CBRAM

- $Ag/Ge_{30}Se_{70}$ CBRAM exposed to 100 keV Li
- Primary mechanism is TID
- Only small accuracy degradation until fluences of ~30Mrad(GeSe)

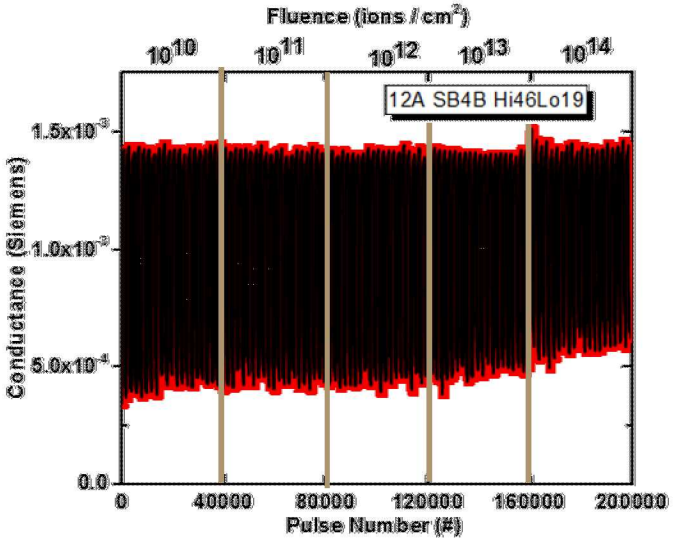


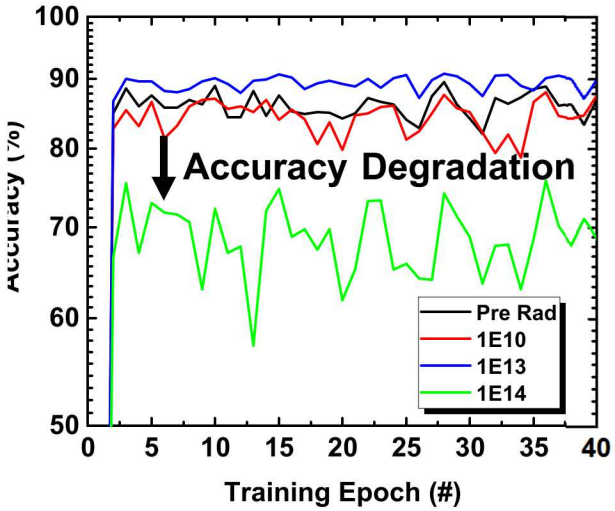


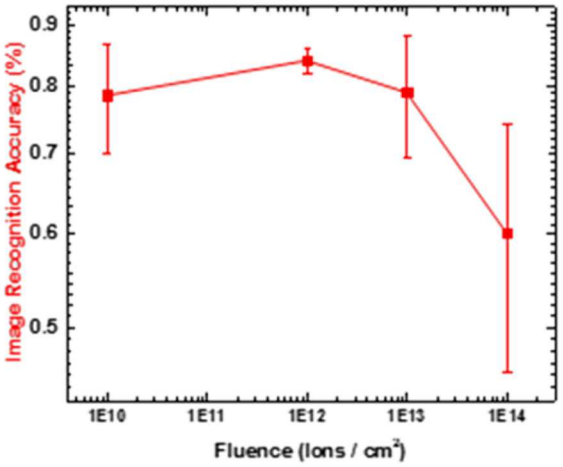


Effect of Displacement Damage on Training – ReRAM

- 2.5 MeV Si Ion irradiation, neural training between steps
- Training accuracy for higher resistance devices may be degraded following ion irradiation at high fluence
- In line with TaOx ReRAM heavy ion degradation results







R Jacobs-Gedrim, DR Hughart et al, *TNS, 2018*

Outline

Sandia National Laboratories

- Motivation and Background
- Analog Accelerator Concepts & Performance Model
- Accelerator Accuracy Modeling & Characterization
- Future Directions
- Radiation Effects
- Summary

Summary

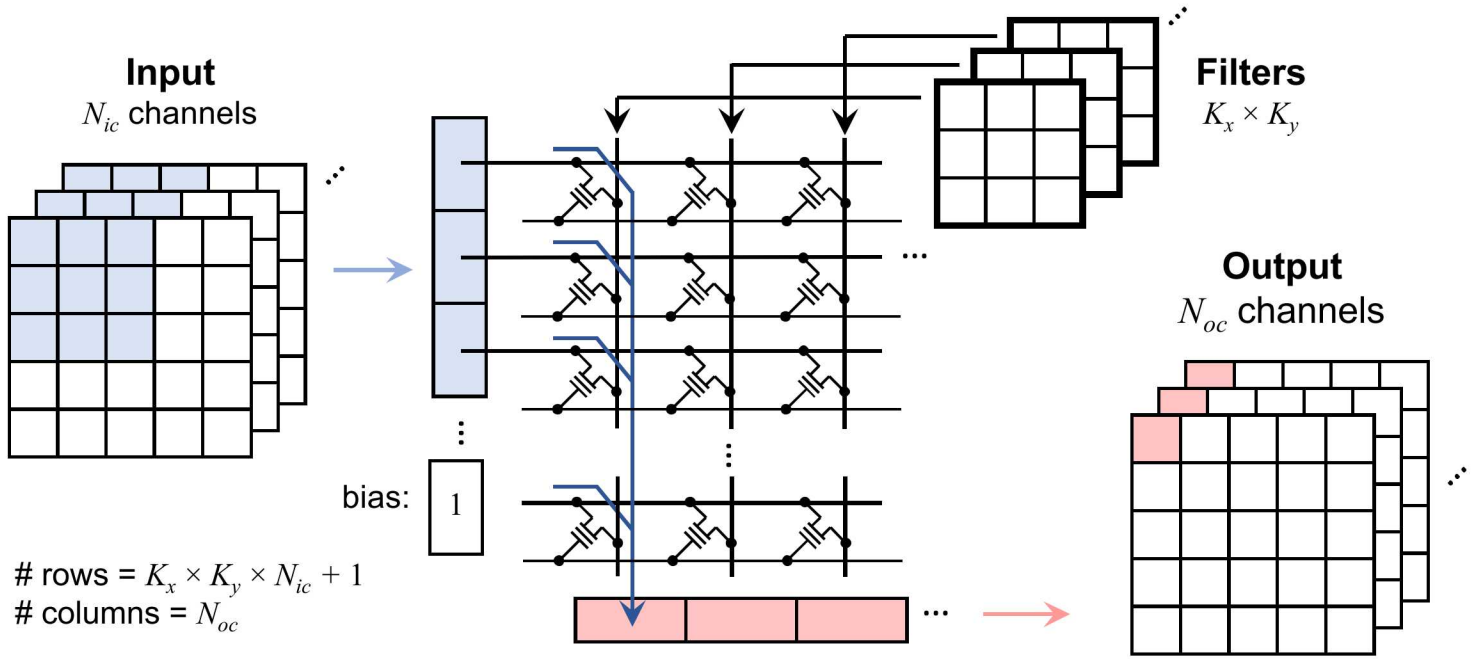
- As CMOS scaling slows (or ends), a new direction will be needed to achieve the continue the exponential improvements in performance per watt (or energy efficiency)
- Deep neural networks and related algorithms being adopted for many big data and embedded applications
- Analog crossbar VMM and training operations can extend performance per watt 10-1000x beyond digital CMOS limits
- Ideal analog "synapses" are a topic of continued research: Ionic floating gate and magnetic domain wall devices may mitigate issues:
  - IFG has demonstrated offers significant potential as a low energy, high accuracy neural network accelerator synapse
  - Future challenges: continue to improve and scale IFG – and demonstrate CMOS integration
- In-memory computing shows promise for radiation-induced fault tolerant computing
  - Additional research in this area is needed

Acknowledgements

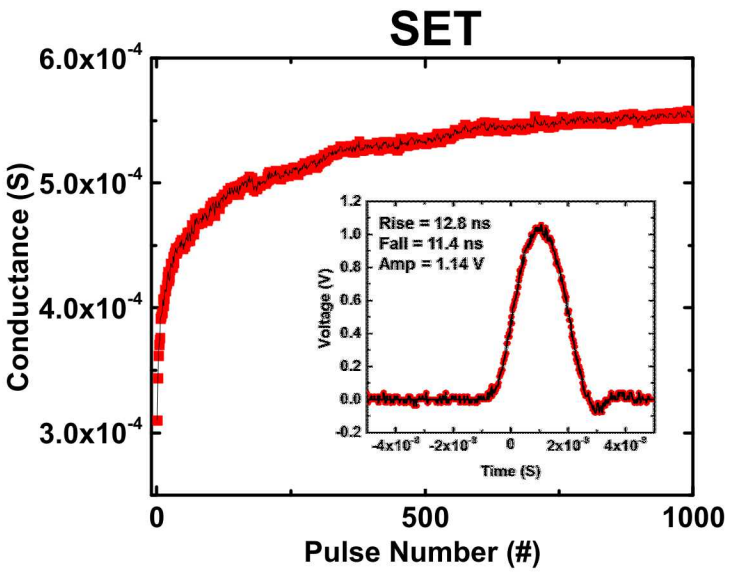
- Funding
  - Sandia National Laboratories Laboratory Directed Research and Development
  - Department of Energy Advanced Manufacturing Office (AMO)
  - Department of the Defense, Defense Threat Reduction Agency under grant no. HDTRA1-11-21-BRCWMD-BAA
- Many collaborators:
  - Alberto Saleo group, Stanford
  - Michael Kozicki, Jennifer Taggart, Sheming Yu, ASU
  - Engin Ipek, Isaac Richter, U Rochester
  - John Paul Strachan, Stan Williams (retired), HPE Labs
  - Jianhua Yang, Qiangfei Xia, U Mass
  - S. Salahuddin, UC Berkeley
  - Tarek Taha, C. U Dayton
  - Paul Franzon, NC State University
  - Dhireesha Kudithipudi, RIT
  - Dozens of others…

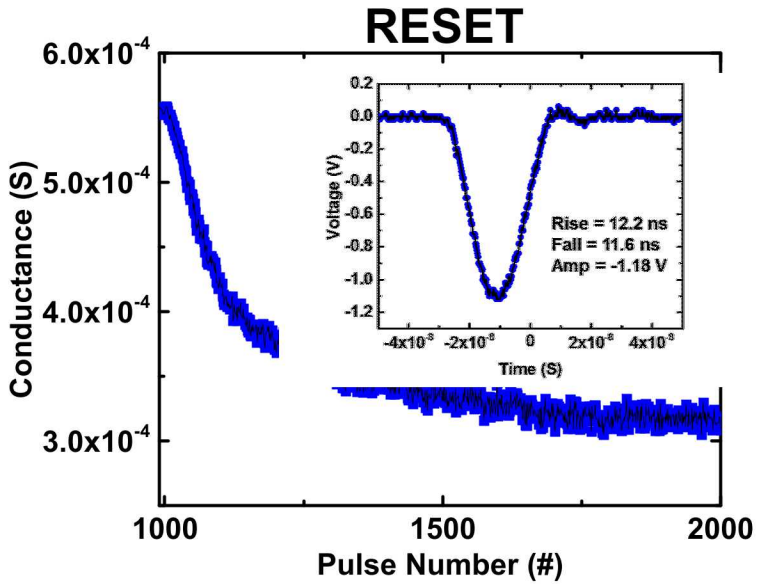
Thank you!

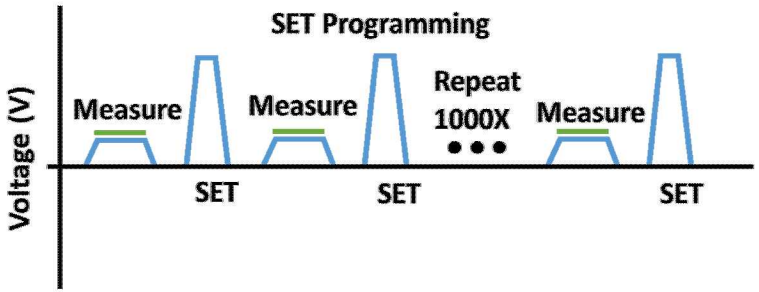
Conclusions and Future Work

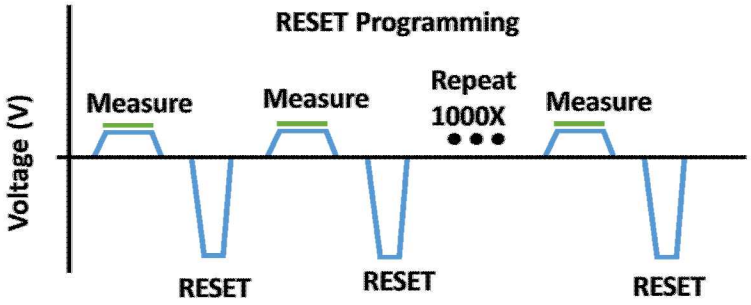


ReRAM Analog Characterization

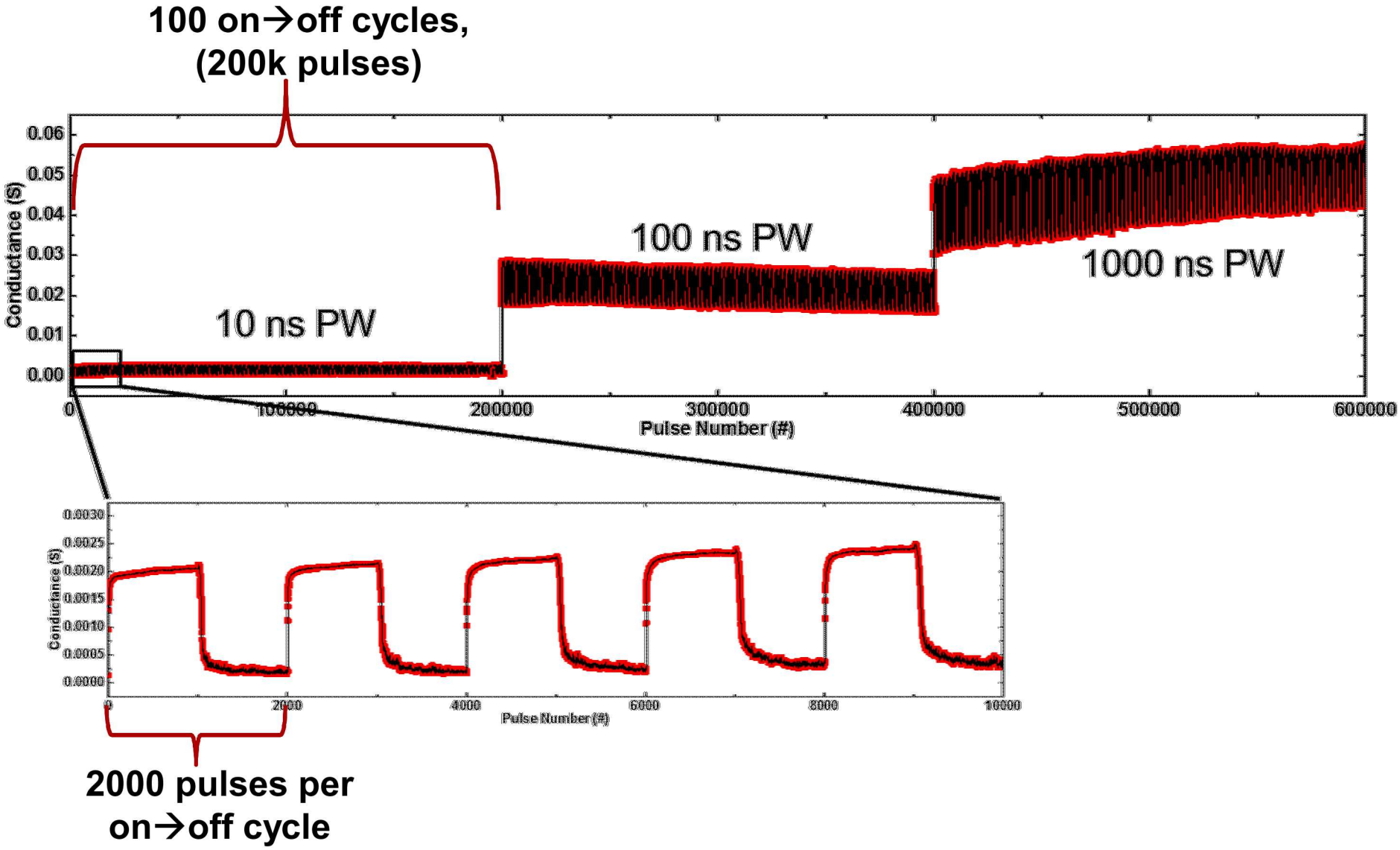








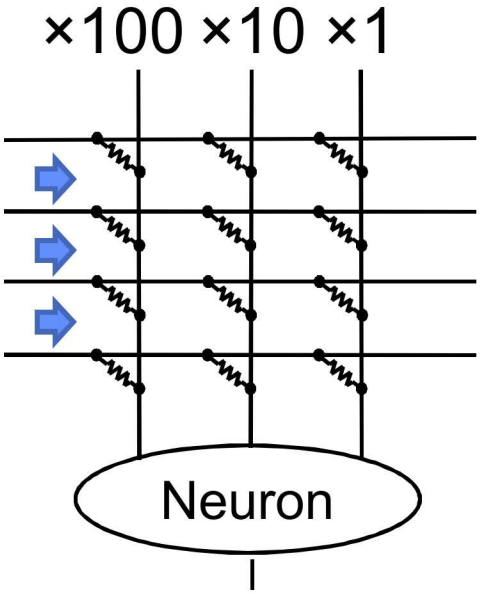
Pulse Width Analog Measurements



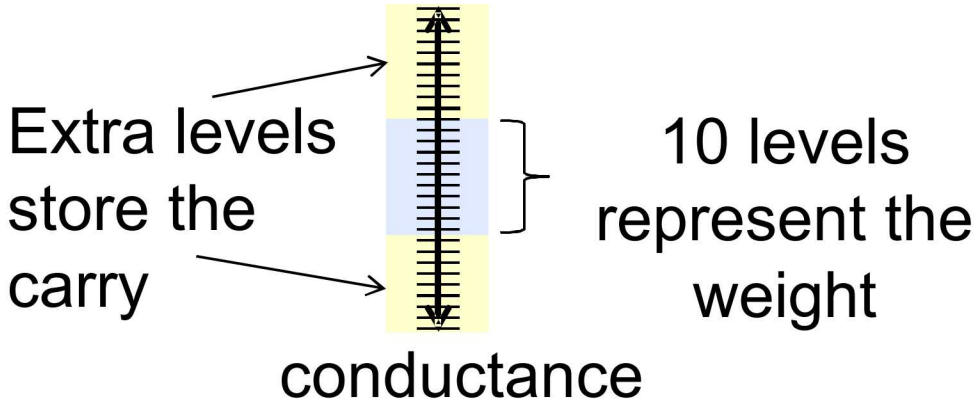
Multi-ReRAM Synapse: Periodic Carry

If we need more bits per synapse, use multiple memristors

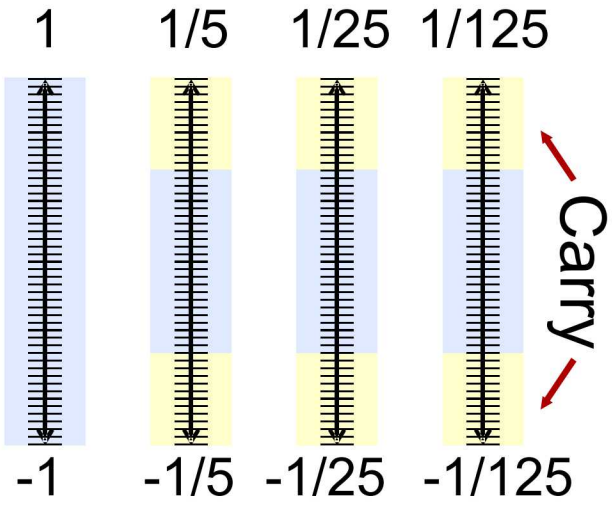
- Three 10 level ReRAMs could represent 1-1000!
- Adding to the weight requires reading every ReRAM to account for any carries and serially programming each ReRAM: VERY EXPENSIVE

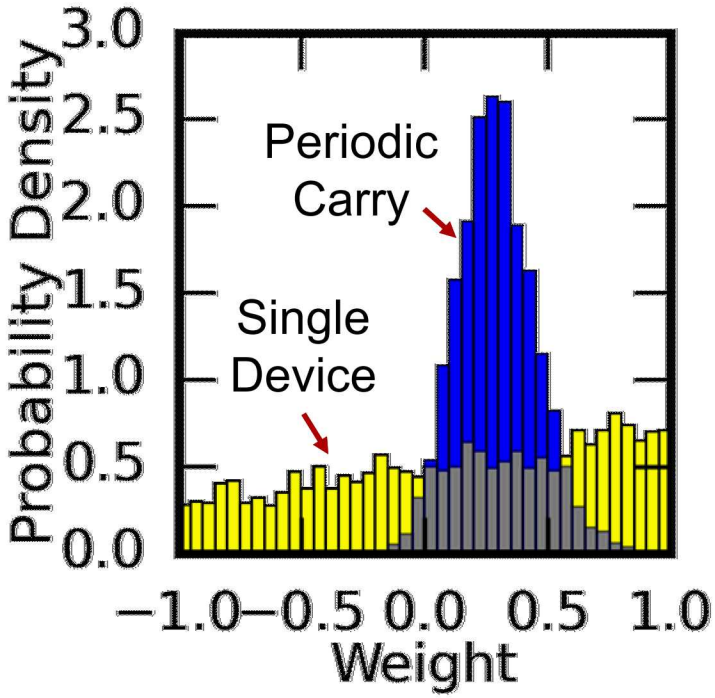


- Use >10 levels to represent a base 10 system
- Ignore carry and program the crossbar in parallel.
- Periodically (once every few hundred cycles) read the ReRAM and perform the carry



Periodic Carry Compensates for Write Noise





Read and reset every 100 pulses

Do 300,000 small (0.02% of weight range) updates

- net of 1500 positive training pulses

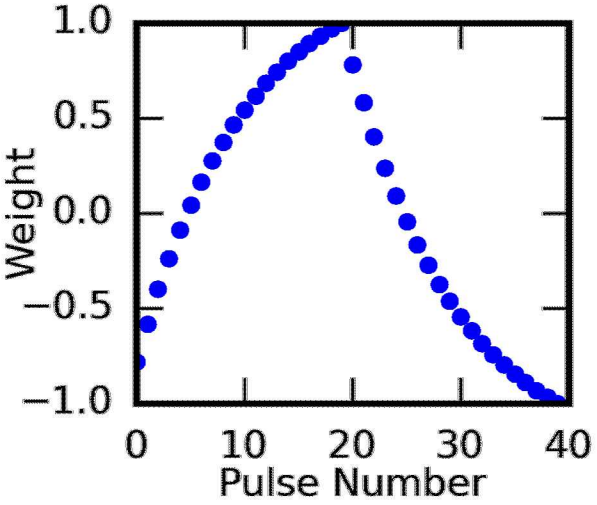
Learn from a 0.5% Signal

Noise Sigma = 1.4% for single device

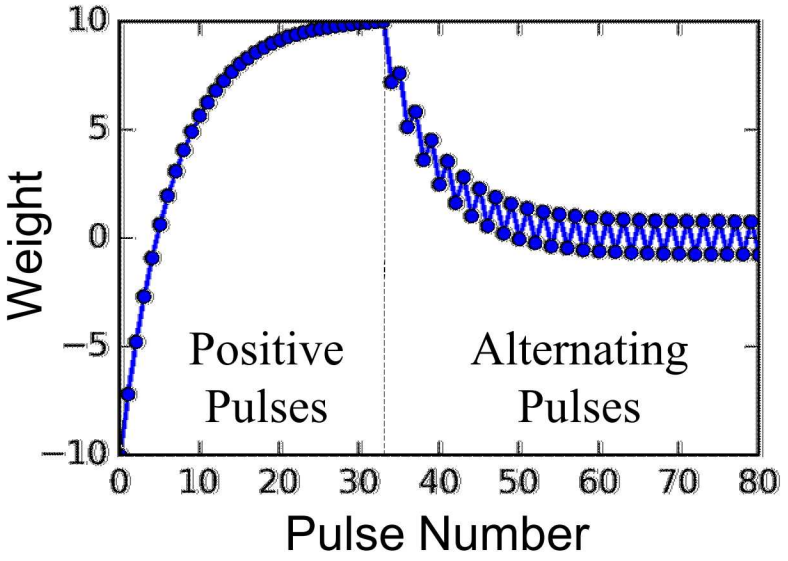
- (from $\sigma_{noise} / G_{range} = 0.1\sqrt{\Delta G / G_{range}}$)
- Write noise applied during updates and carries

Periodic Carry Mitigates Write Nonlinearity

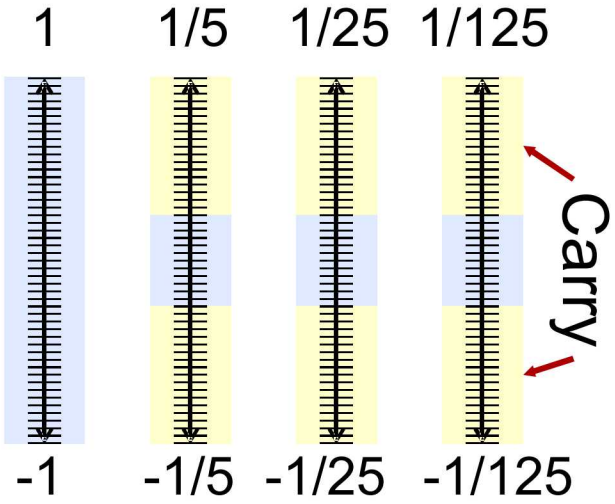
Write Nonlinearity

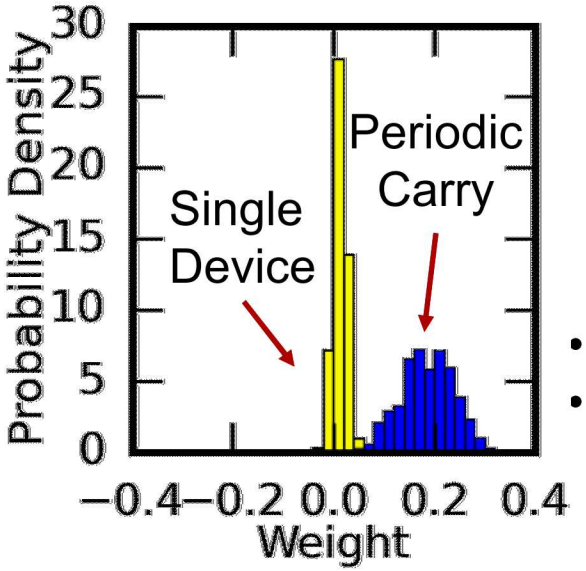


Alternating Pulses Cause Weight Decay



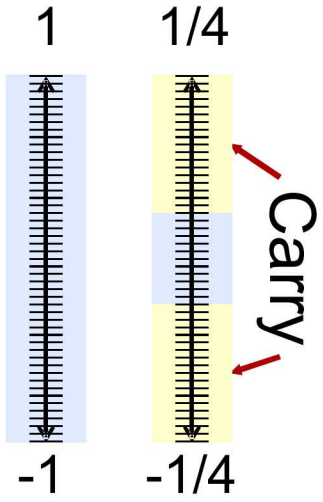
Use center linear range of weights



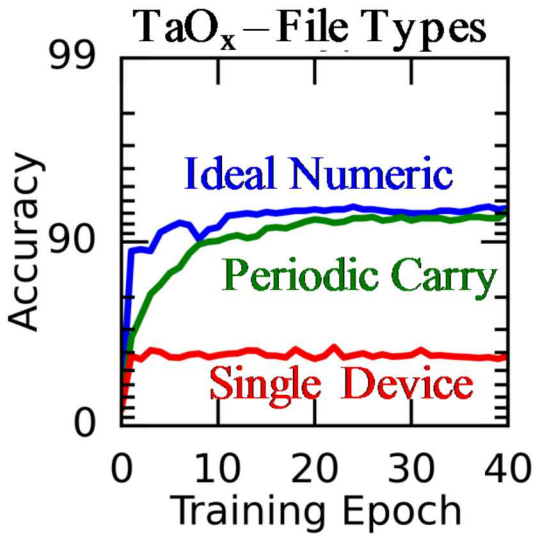


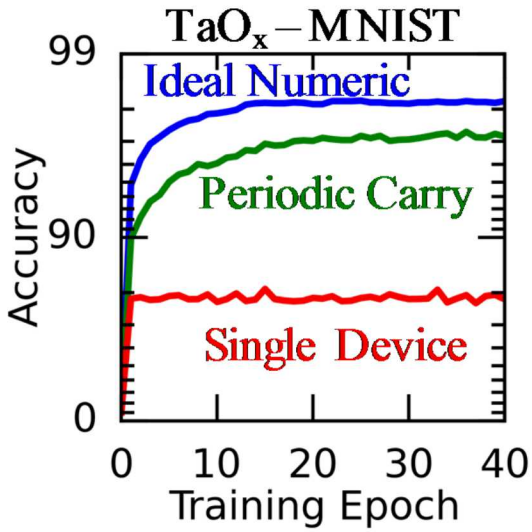
- Train with 1% signal
- Ideal result is 0.6

TaO_x Results



Carry once every 1000 updates

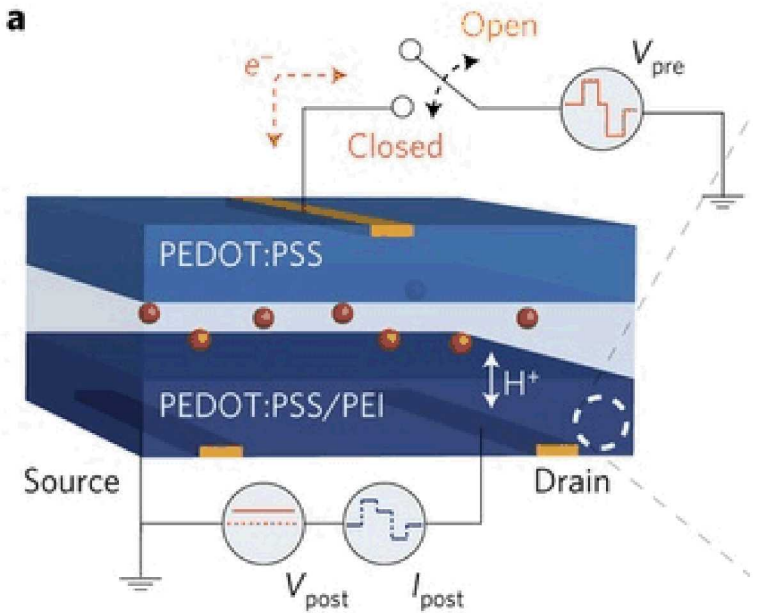


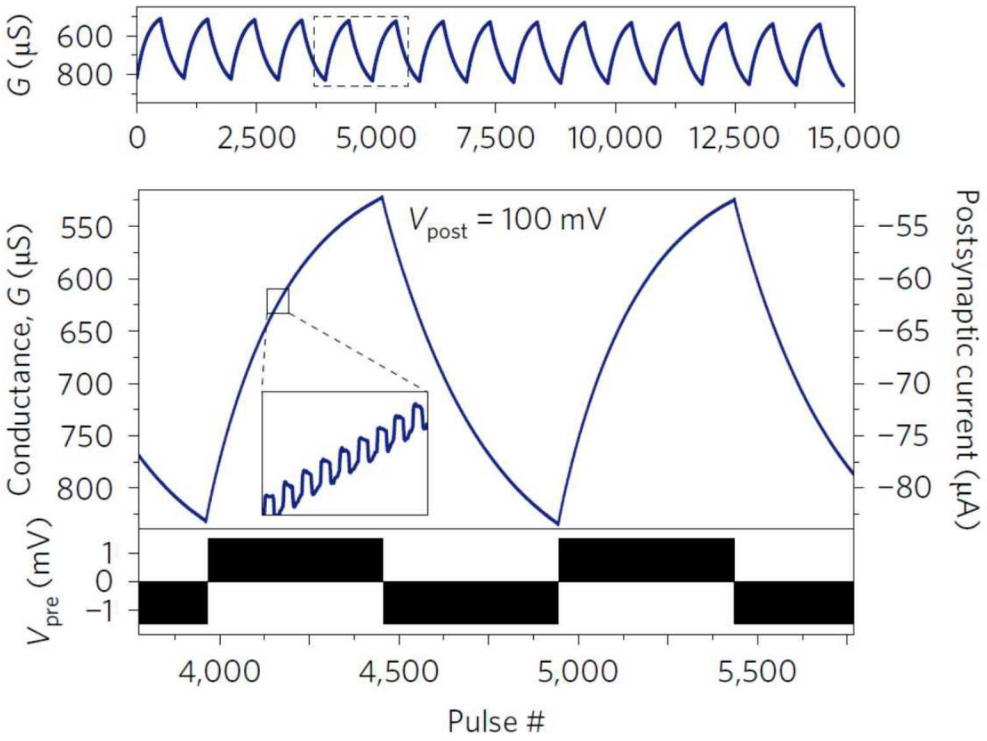


A/D and D/A is modeled, Serial operations modeled

- When resetting weight, need to adjust pulse size based on current state to compensate for nonlinearity
- When reading a single weight, need to adjust readout range to be smaller (change capacitor on the integrator)

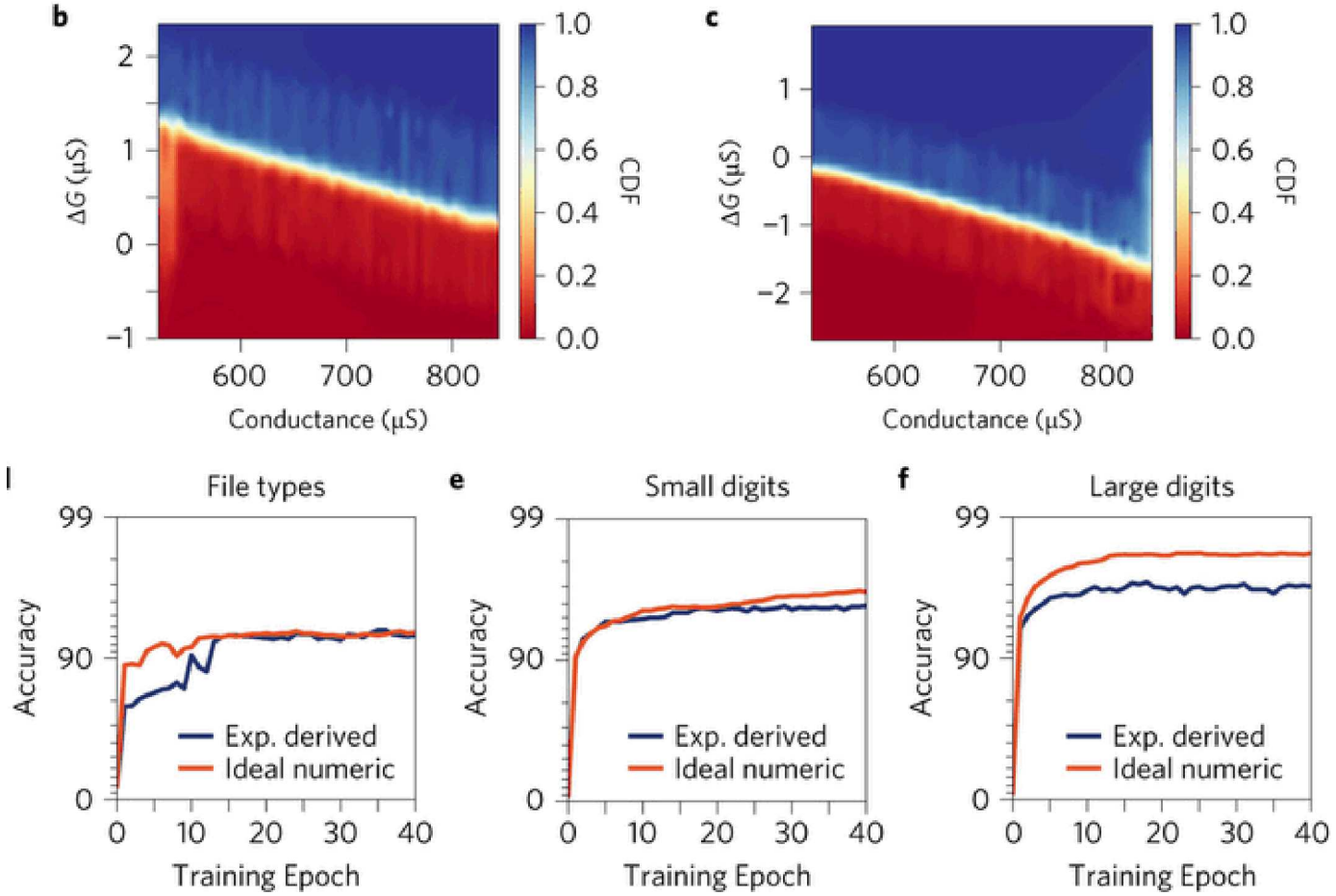
Electrochemical Neuromorphic Organic Device (eNode)





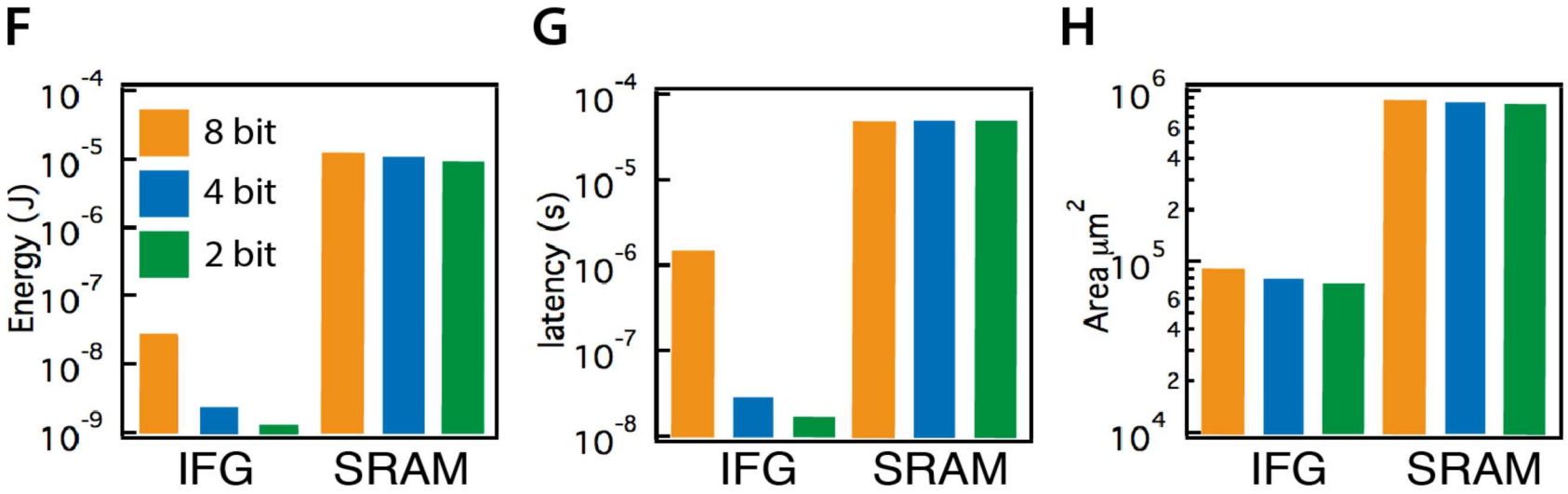
van de Burgt et al, *Nature Mater.*, 2017

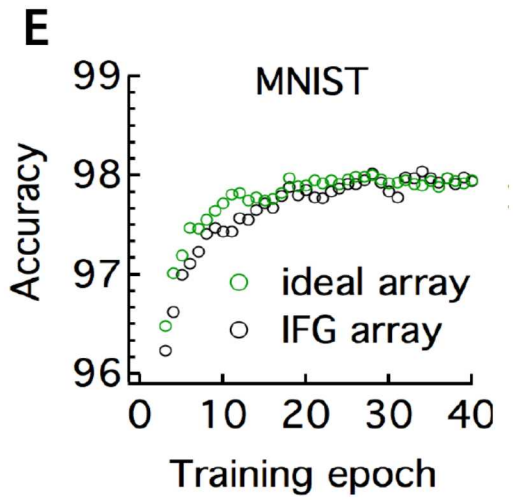
Electrochemical Neuromorphic Organic Device (eNode)



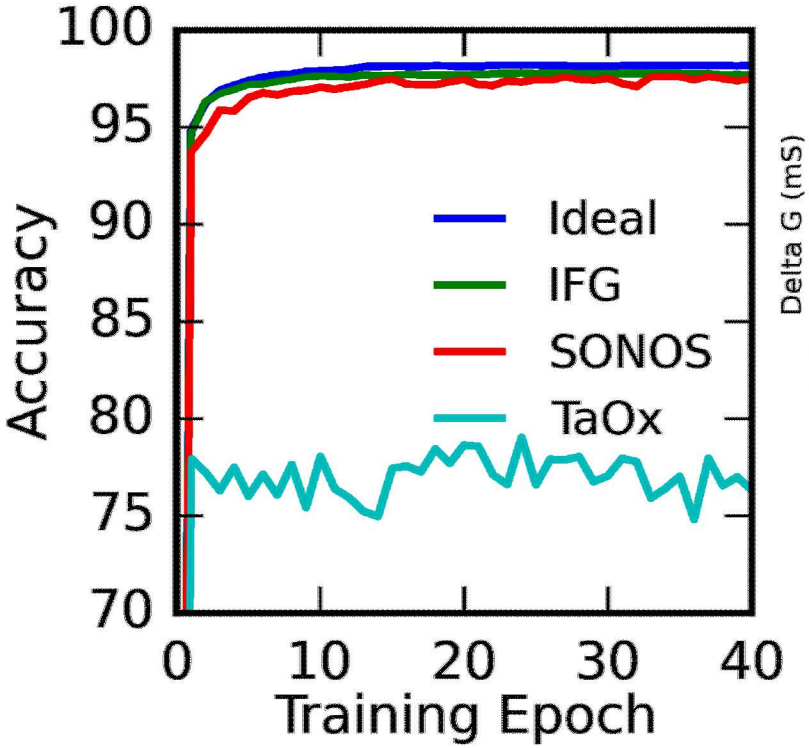
van de Burgt et al, *Nature Mater.*, 2017

Programming Demonstration

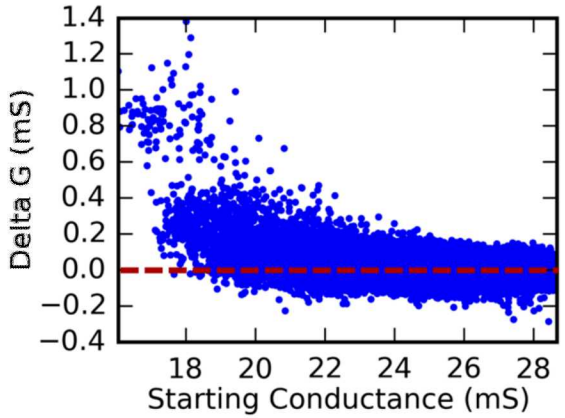




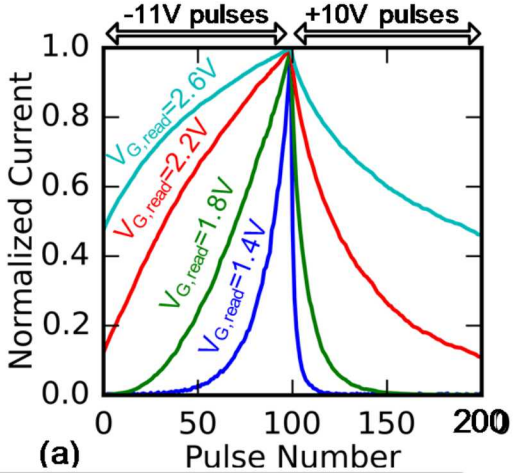
Three Terminal Devices Tend to Have Higher Accuracy

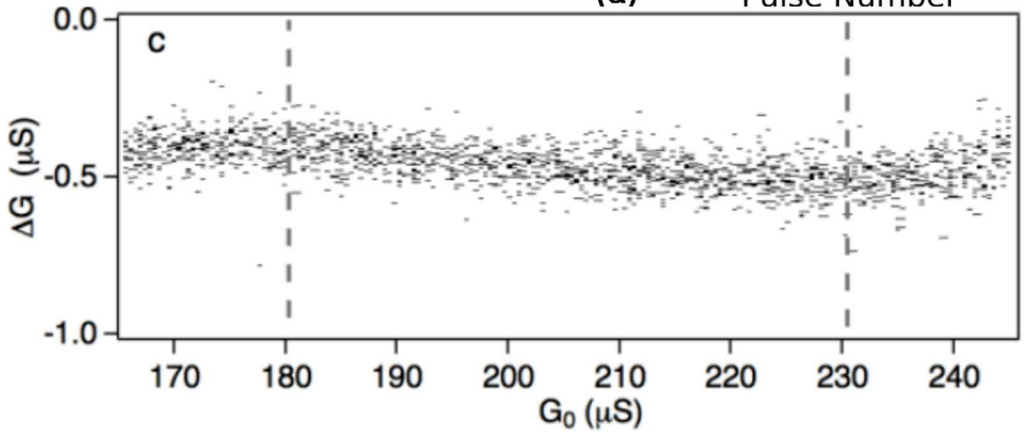


ReRAM



SONOS

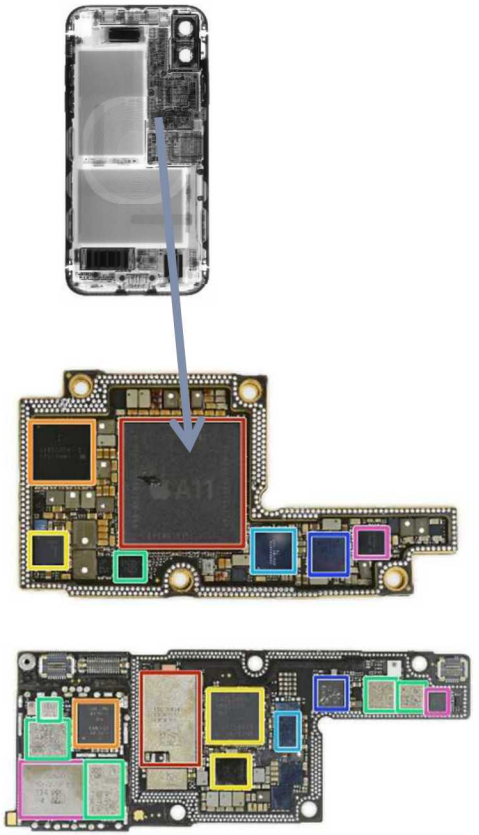




Ionic Floating-Gate

Digital Accelerators: Apple A11

- Apple's iPhone 8 and X main SoC processor
- TSMC 10nm process
- No formal technical papers or presentations given yet
- 600 GigaOps/s claimed, no other info
- Rough order of magnitude analysis still possible:
  - Assume "Op" is 8-bit fix point
  - Entire smartphone CPU <600mW
  - **Almost certainly > 1 TeraOp/W, or <1pJ/op**
- Next challenge: moving beyond this!

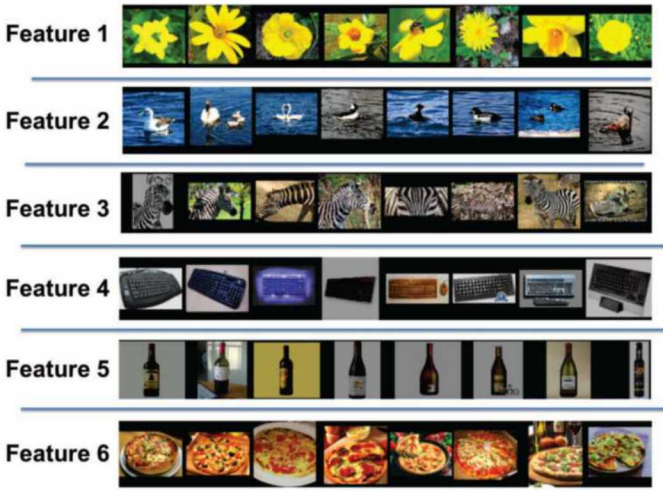


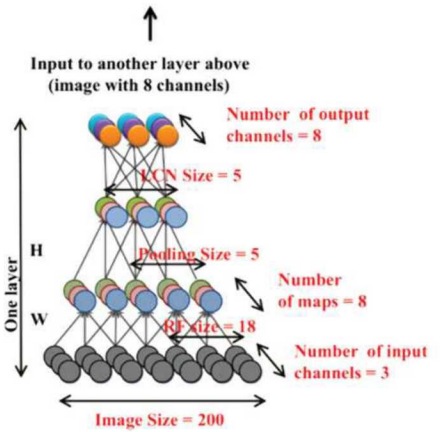
www.ifixit.com

Why should we continue these gains?

Sandia National Laboratories

- **Google Deep Learning Study**
  - 16000 core, 1000 machine GPU cluster
  - Trained on 10 million 200x200 pixel images
  - Training required 3 days
  - Training dataset size: no larger than what can be trained in 1 week
- **What would they like to do?**
  - ~2 billion photos uploaded to internet per day (2014)
  - Can we train a deep net on one day of image data?
  - Assume 1000x1000 nominal image size, linear scaling (both assumptions are unrealistically optimistic)
  - *Requires 5 ZettaIPS to train in 3 days (ZettaIPS=10^{21} IPS; ~5 billion modern GPU cores)*
  - World doesn't produce enough power for this!
  - Data is increasing exponentially with time
- **Need >10^{16}-10^{18} instruction-per-second on 1 IC**
  - Less than 10 fJ per instruction energy budget

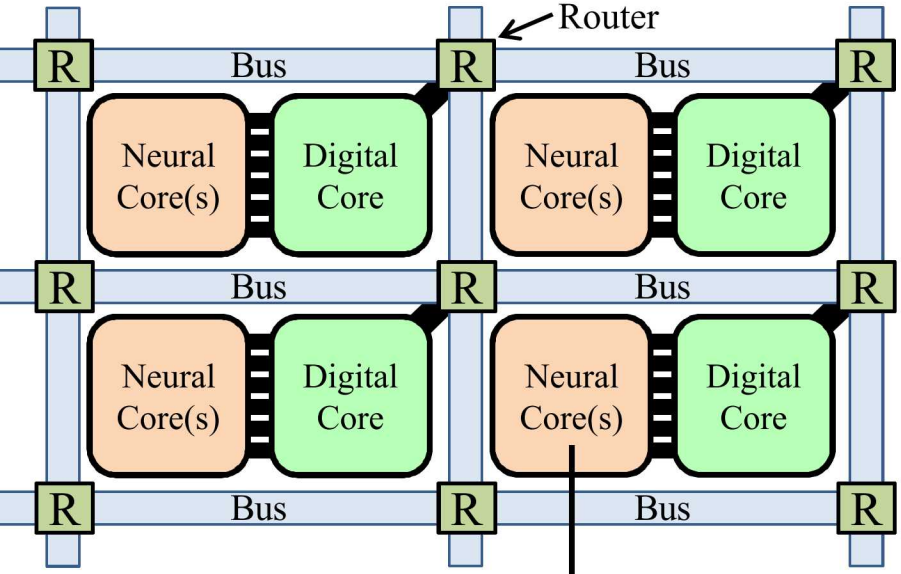


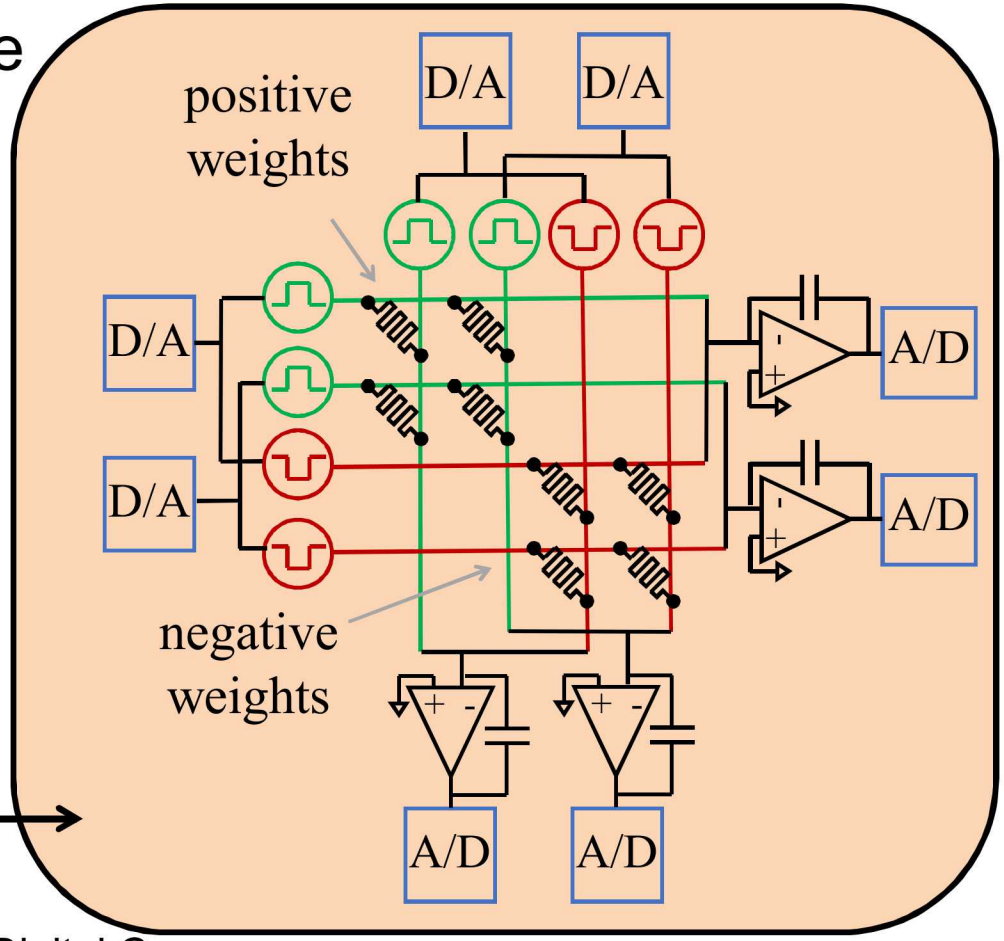


Q. Le, IEEE ICASSP 2013

General Purpose Neural Architecture

Run *any* neural algorithm on the same hardware





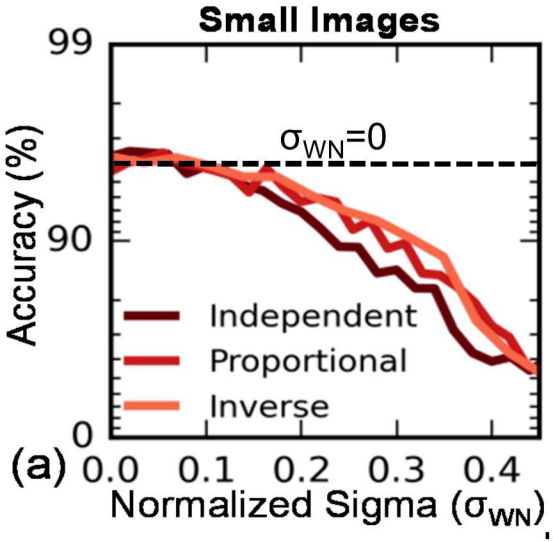
Neuromorphic core:

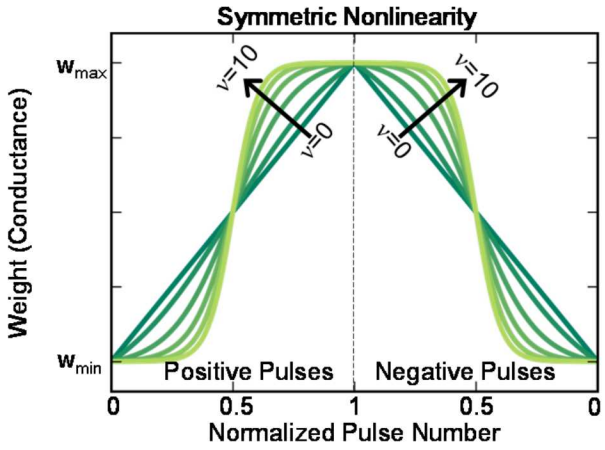
- Evaluate vector matrix multiplies along rows or columns
- Train based on input vectors

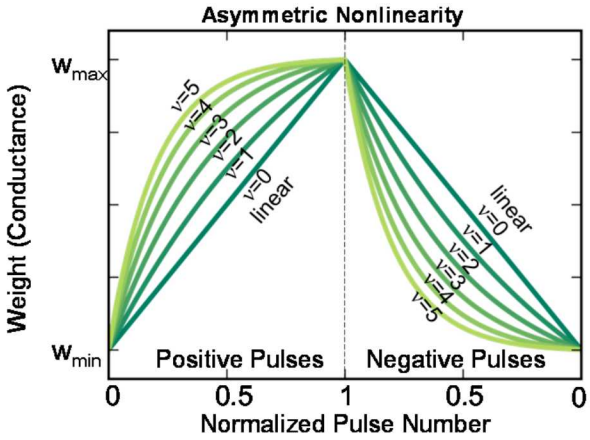
Digital Core:

- Process neural core inputs/outputs
- For NxN crossbar, the crossbar accelerates $O(N^2)$ operations leaving only O(N) operations for the digital core

Modeling Device Requirements



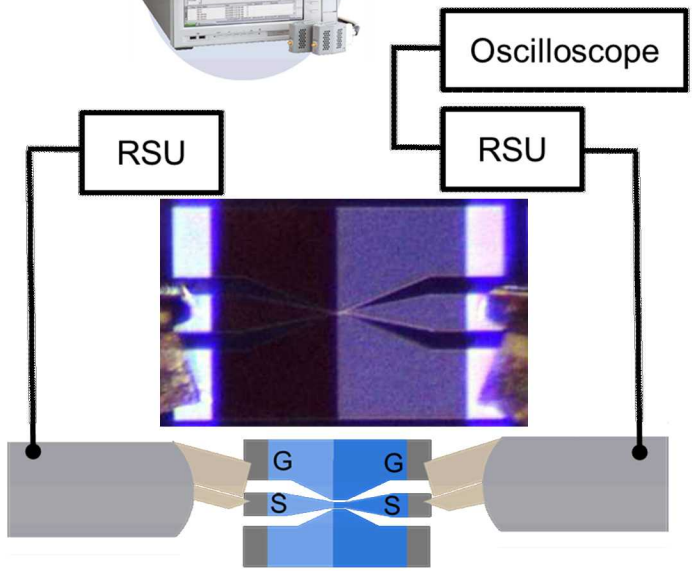


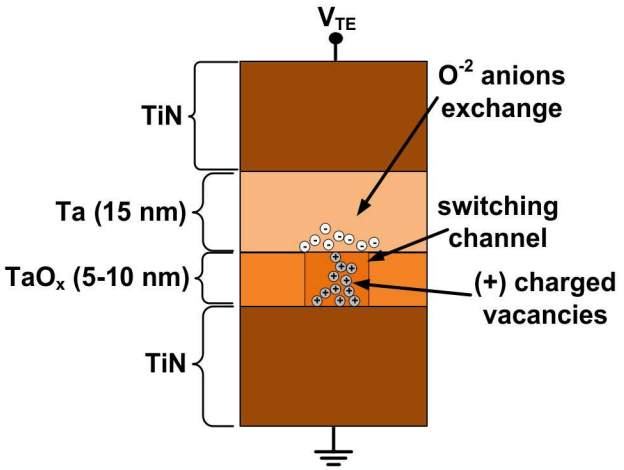


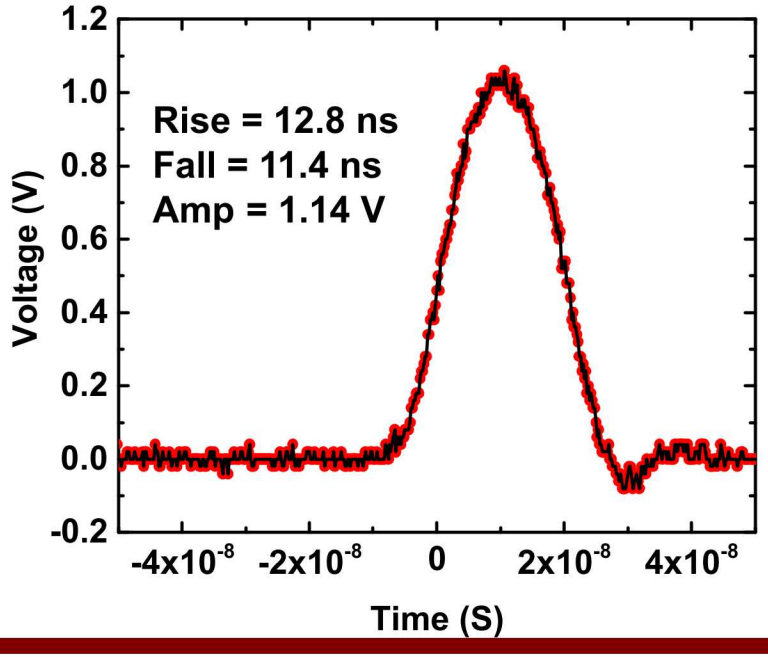
	Small Images	Large Images	File Types
Read Noise σ (% Range)	3%	5%	9%
Write Noise σ (% Range)	0.3%	0.4%	0.4%
Asymmetric Nonlinearity (ν)	0.1	0.1	0.1
Symmetric Nonlinearity (ν)	>20	5	5
Maximum Current	160 nA	13 nA	40 nA
Minimum Retention (@ 85°C)	7 days	7 days	7 days
Minimum Nudge Endurance	10^7	10^7	10^7

ReRAM Measurements

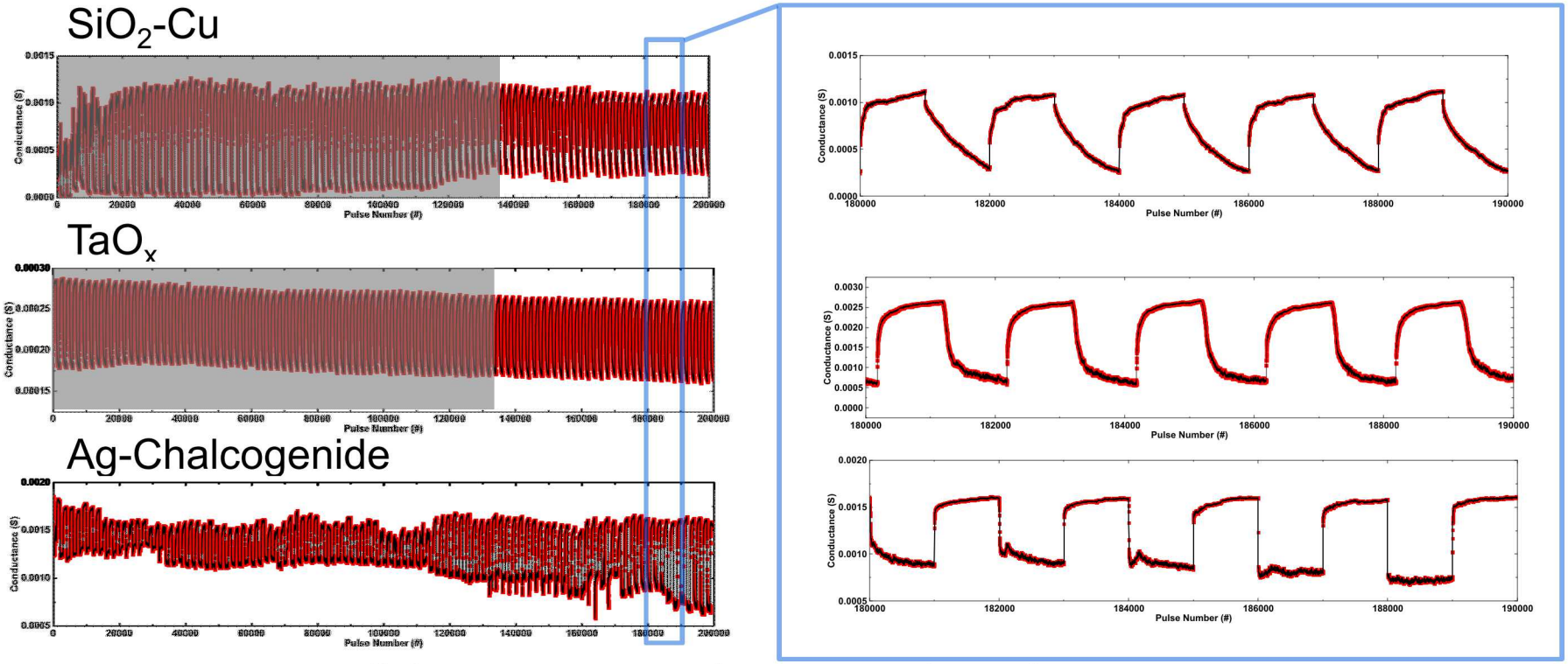
- **DC Current-voltage "loops" sweeps are not time-controlled**
  - **Excessive heating and early wearout**
  - **Do not provide info on dynamics**
- **Physical switching < 10ns**
- **Need pseudo RF setup to measure**
  - **Ground/signal, conductor backed**
  - **Agilent B1530 module**
  - **10 ns RT/FT, 10 ns PW**
  - **1 V nominal, ~140 mV overshoot**







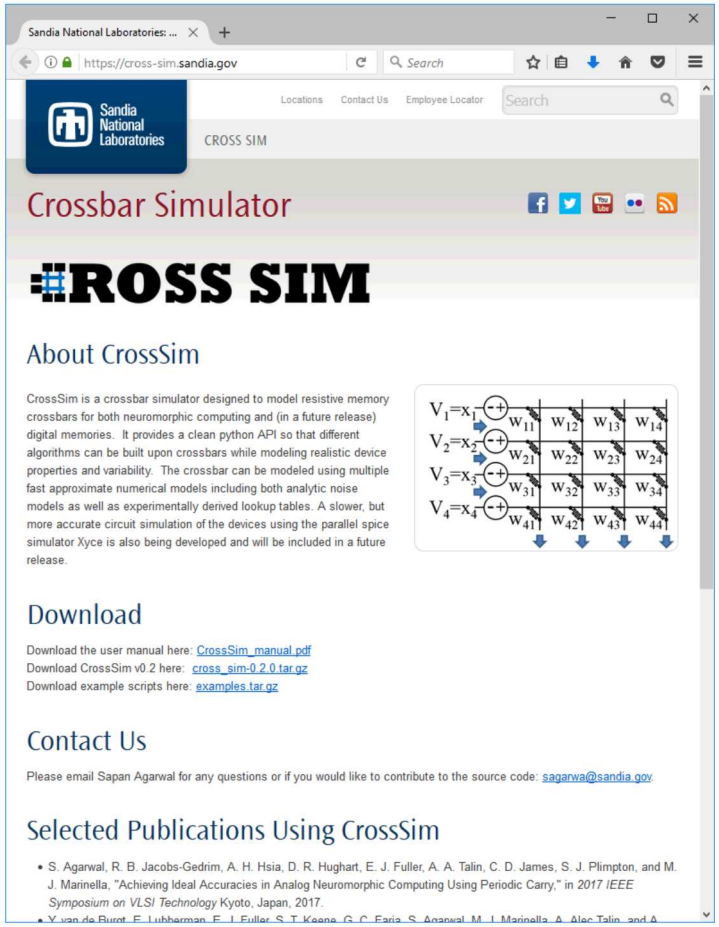
Analog Device Comparison



- **For comparison all devices at 100 ns due to impedance limitation**
- **Device operation voltages found by increasing amplitude by 0.1 V until switching occurred – must survive 200,000 nudges so lowest possible voltage used**
- **Chalcogenide SET = +0.8 V RESET = -0.8V**
- **SiO_2-Cu SET = +1.4 V RESET =-1.6 V**
- **TaO_x SET +1.0 V RESET = -1.0 V**

#ROSS SIM

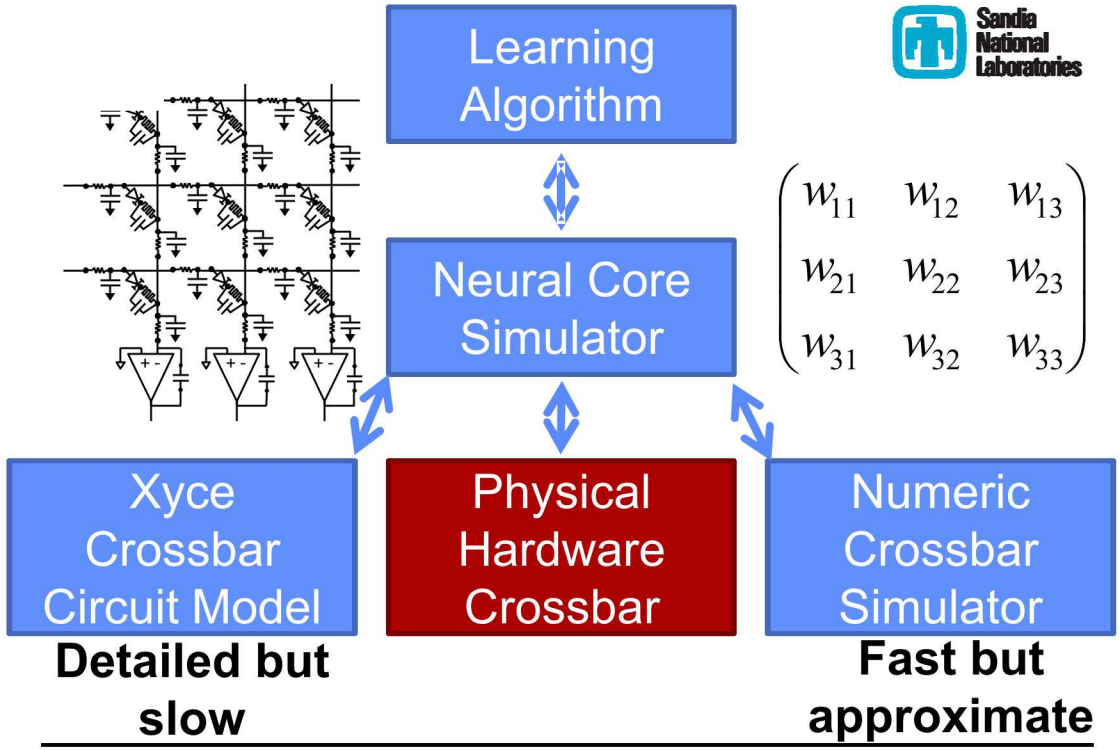
https://cross-sim.sandia.gov



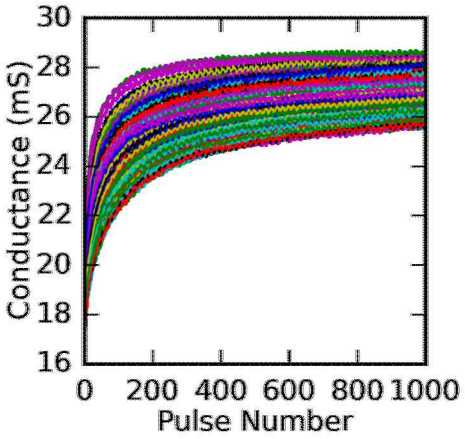
Simple Python API:

 # Do a matrix vector multiplication

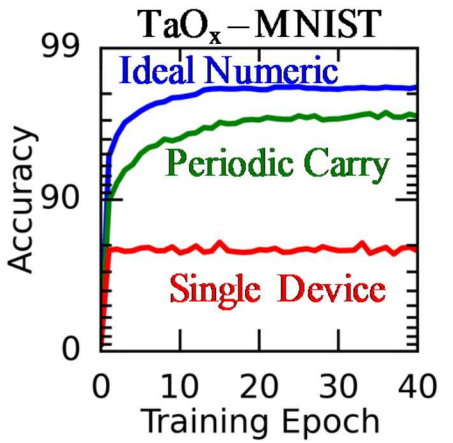
result = neural_core.run_xbar_mvm(vector)



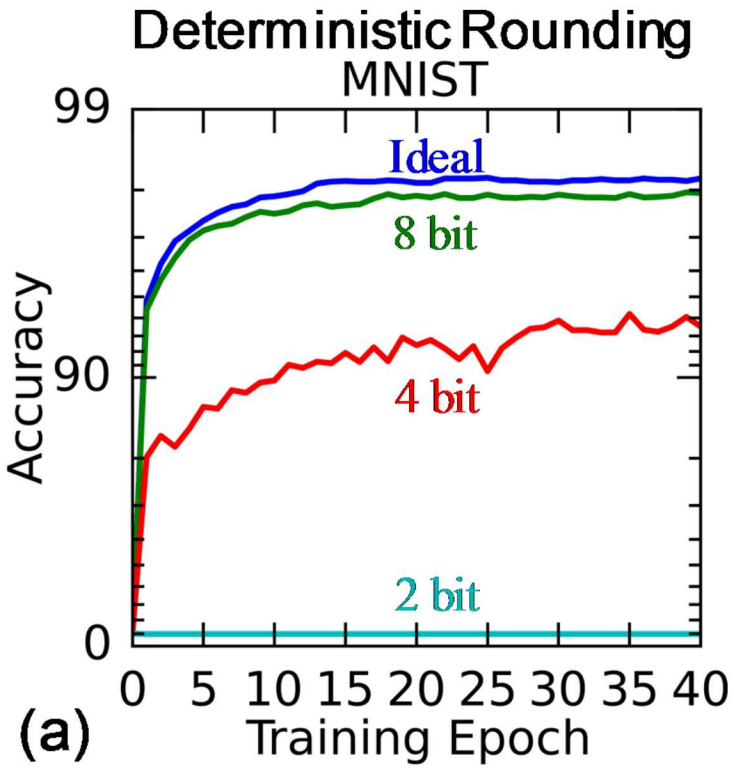
Measured Devices

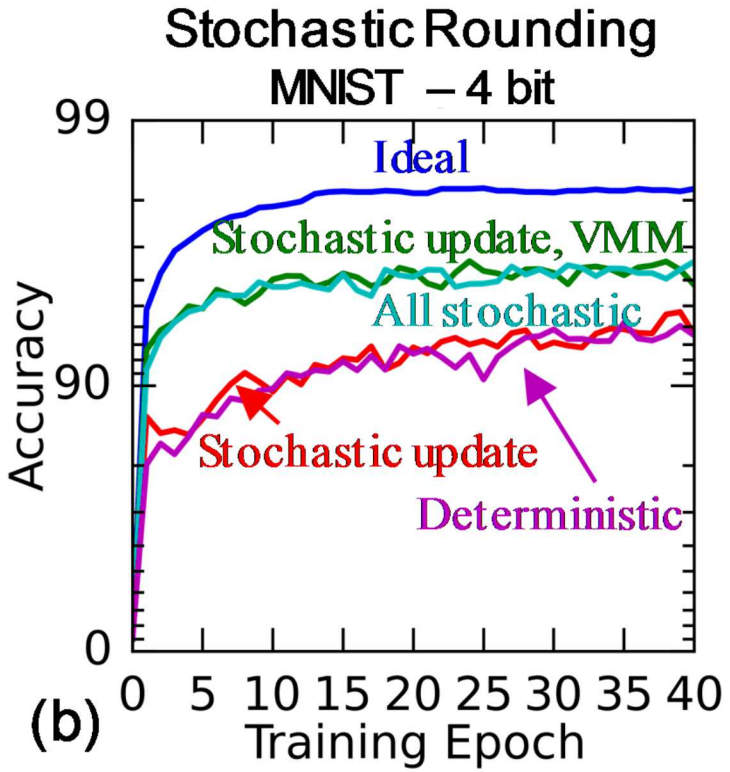


Algorithmic Performance



Effect of System Precision on Backpropagation Classification Accuracy





Agarwal et al, IEEE E3S, 2017

AI Across Power Envelopes

Sandia National Laboratories

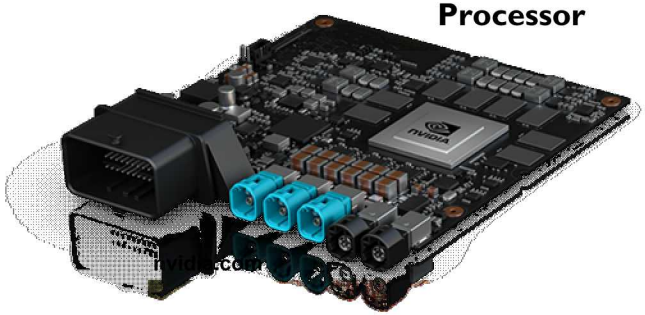
IoT, Edge, and Mobile Computing

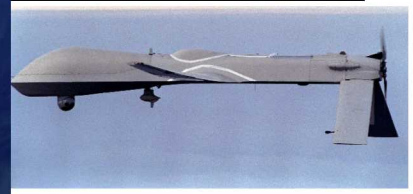
Self Driving Cars, Unmanned Arial Vehicles, and Satellite Computing

Datacenters, HPC

nest.com

Nvidia Self-Driving Processor





wikimedia.org

ASCI Red Supercomputer

1W 10W 10^2W 10^3W 10^4W 10^5W 10^6W