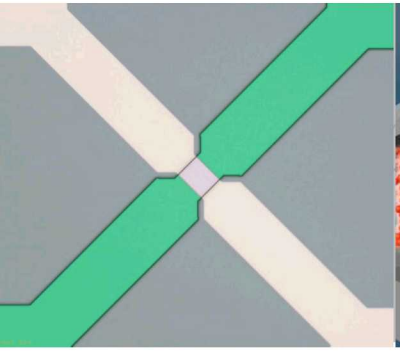
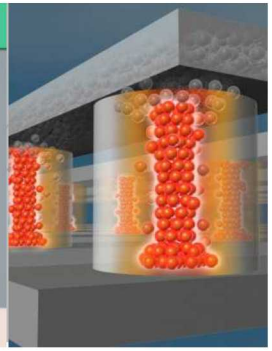
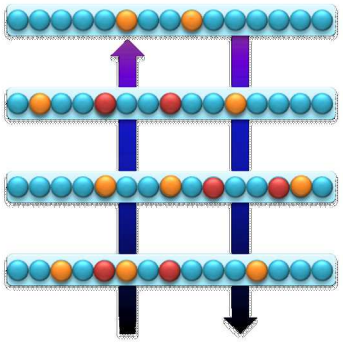
SAND2019-6932C

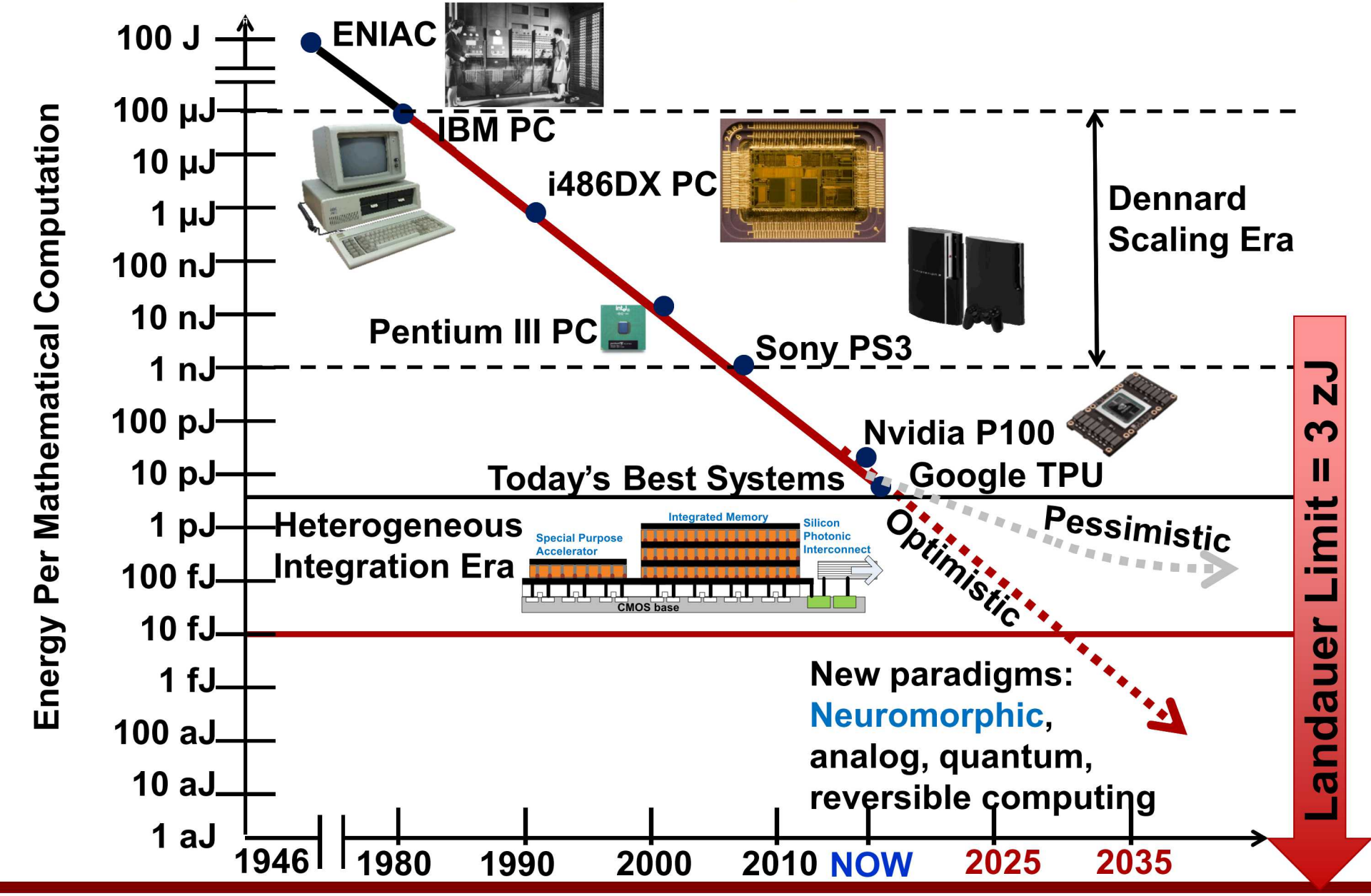





Designing and Modelling Analog Neural Network Training Accelerators

Sapan Agarwal, Robin B. Jacobs-Gedrim, Christopher Bennett, Alex Hsia, Michael S. Van Heukelom, David Hughart, Elliot Fuller, Yiyang Li, A. Alec Talin, Matthew J. Marinella
Sandia National Laboratories

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.

Evolution of Computing Machinery



AI Across Power Envelopes

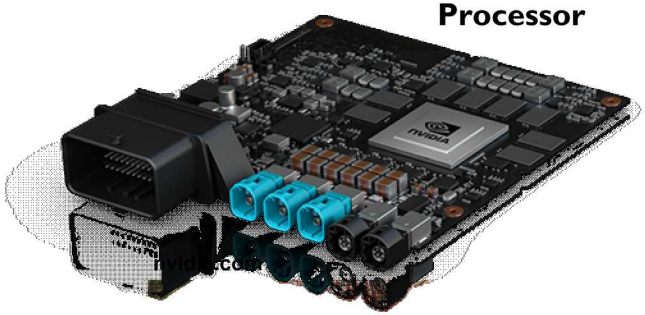
IoT, Edge, and Mobile Computing

Self Driving Cars, Unmanned Arial Vehicles, and Satellite Computing

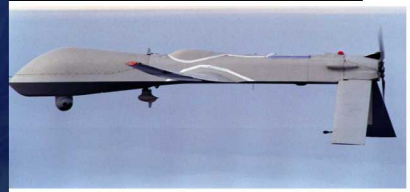
Datacenters, HPC

nest.com

Nvidia Self-Driving Processor

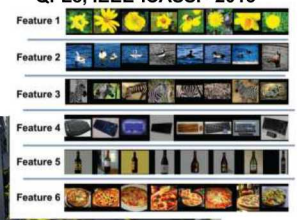


nvidia.com



wikimedia.org

Q. Le, IEEE ICASSP 2013



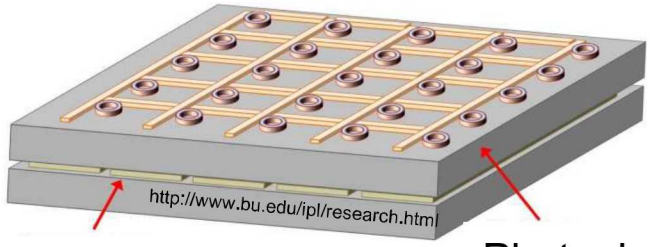
ASCI Red Supercomputer

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

The Von Neumann Bottleneck

Current Transistors ~ 10 aJ
40kT Noise Limit ~ 0.2 aJ

Cross chip communications ~ 1 pJ
DRAM Access >10 pJ
Ethernet ~ 1nJ



Processor Layer

Photonic Layer

Communications require orders of magnitude more energy!

Optical interconnects 100 fJ to 1 pJ

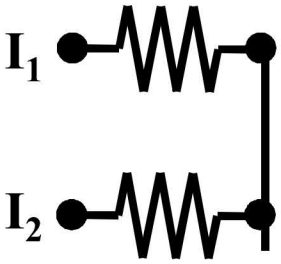
Use Resistive Memories for Local Computation

$$V = I \times R$$

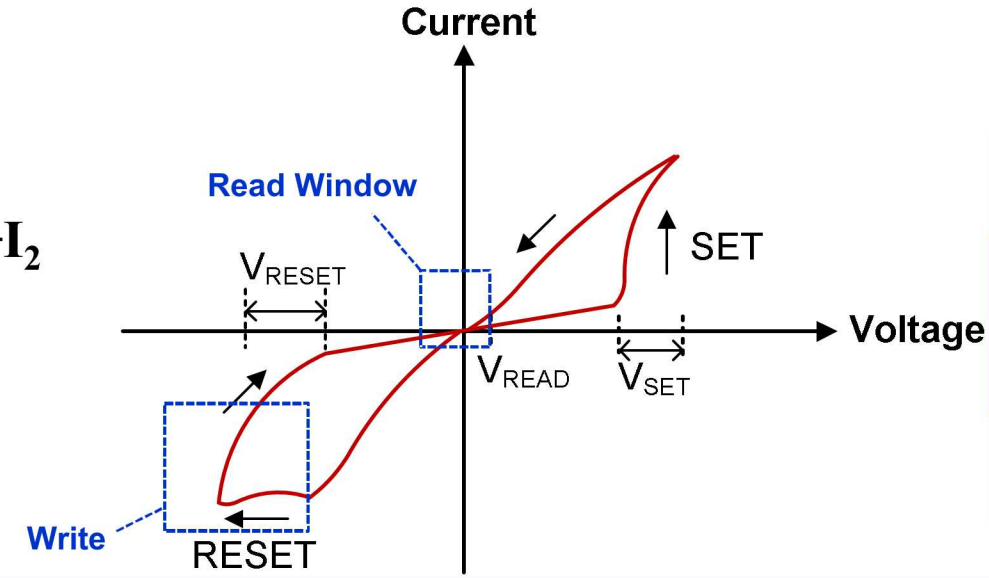
$$I = G \times V$$

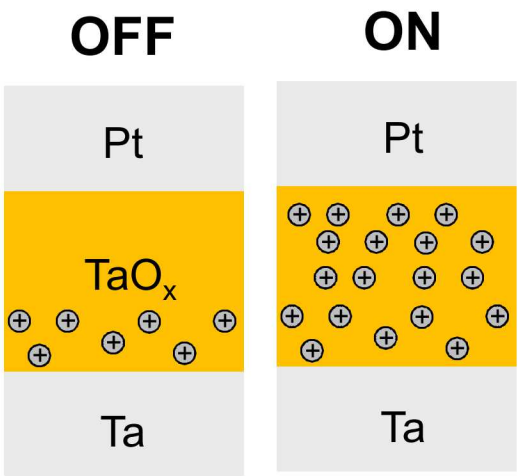
multiplication

- A resistive memory or ReRAM is a programmable resistor
  - Apply small voltages allows the conductance to be read: I = G × V
  - Apply large voltages to change the resistance

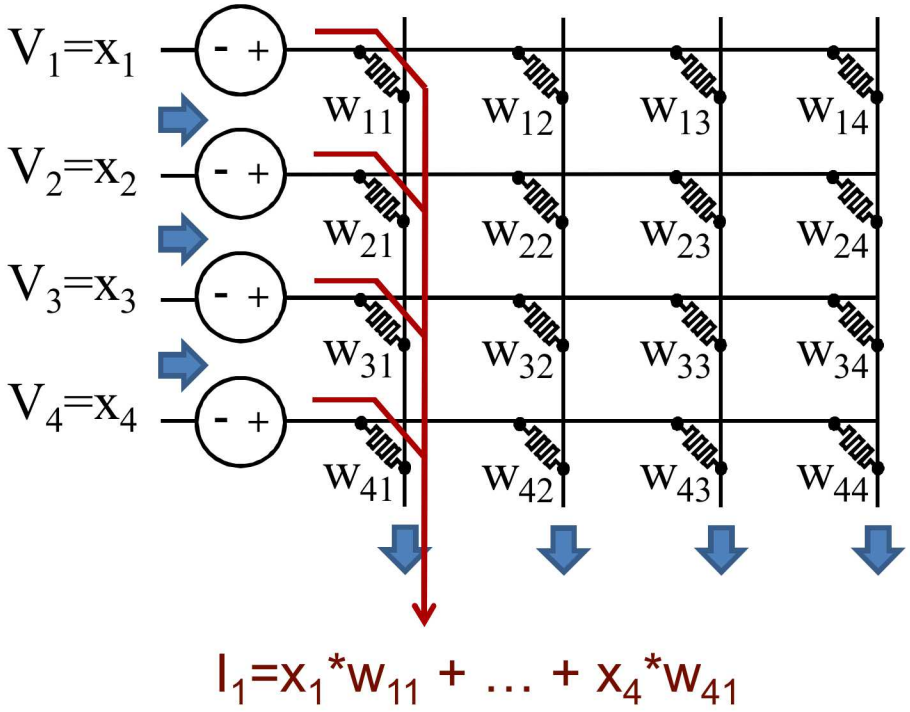


Addition: $I=I_1+I_2$





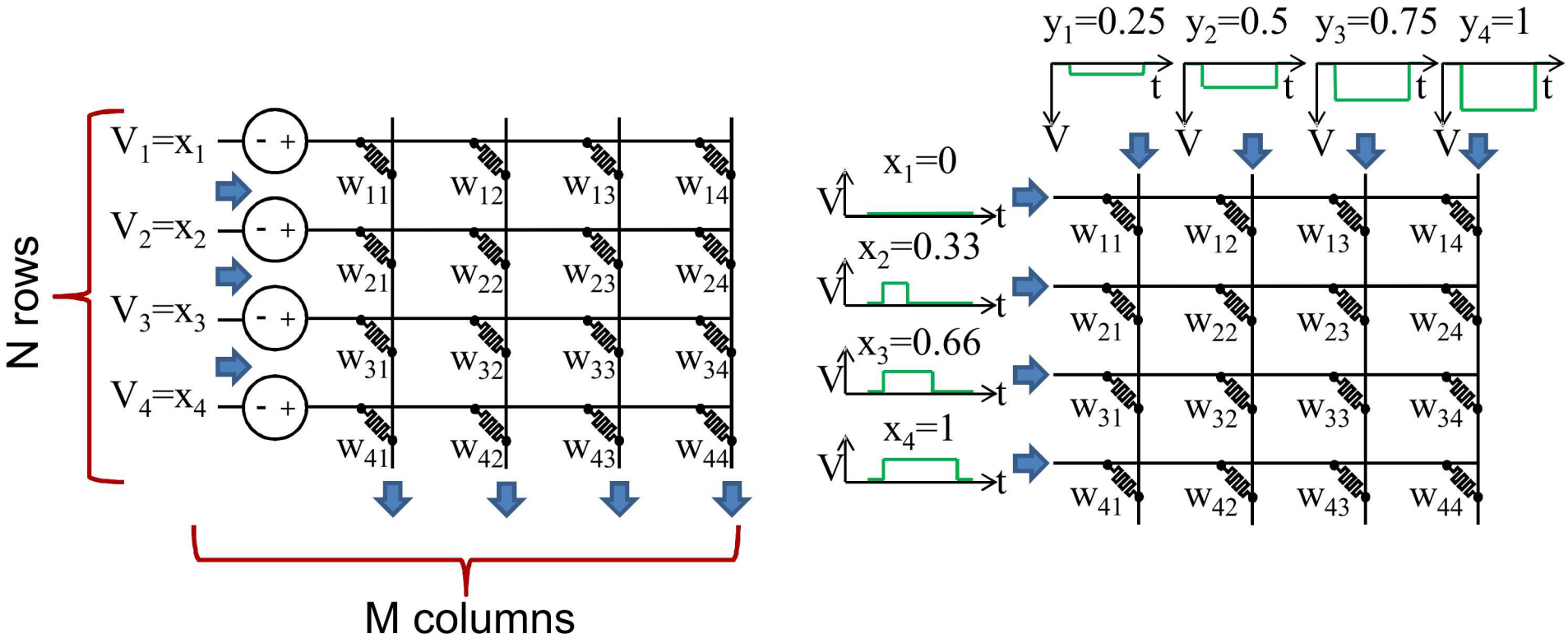
Directly Process in the Memory Itself



Analog is efficiently and naturally able to combine computation and data access

Effectively, large-scale processing in memory with a multiplier and adder at each real-valued memory location

Crossbars Can Perform Parallel Reads and Writes

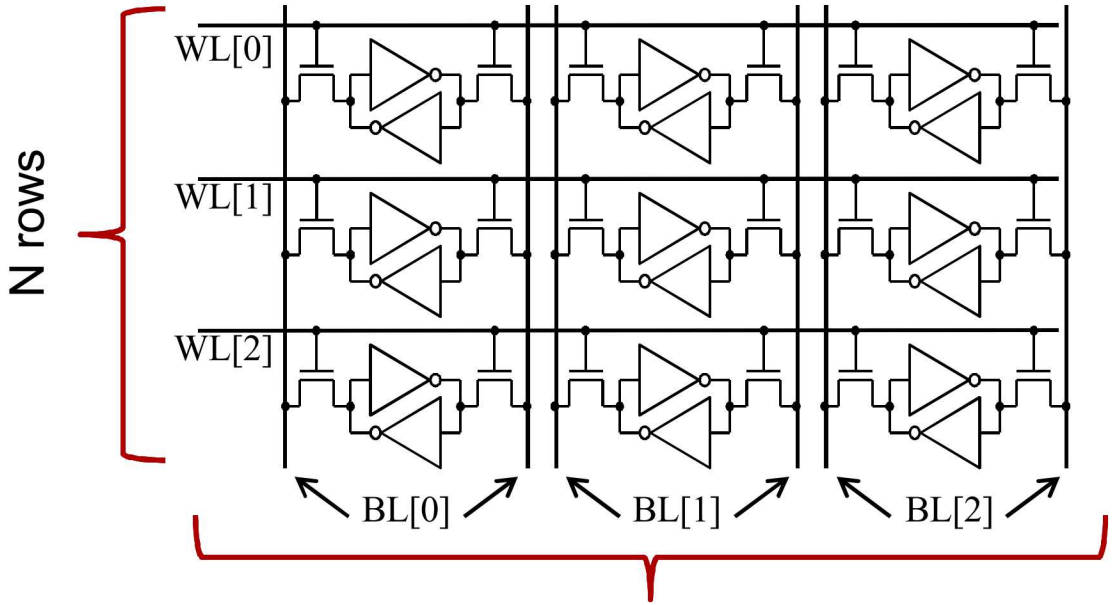


Energy to charge the crossbar is CV^2

$E \propto C \propto$ number of RRAMs $\propto N \times M$

$$E \sim O(N \times M)$$

SRAM Arrays Require Charging Columns Multiple Times



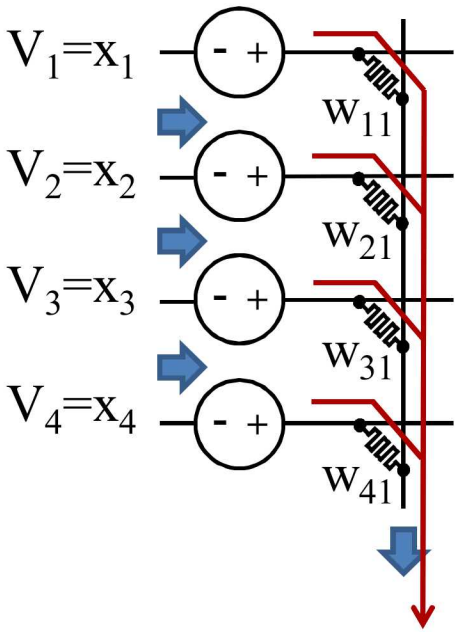
SRAMs must be read one row at a time, charging M columns
Each column wire length is O(N).

Energy = N Rows × M Columns × O(N) wire length

Energy ~ $O(N^2 \times M)$

O(N) times worse than a crossbar!

Need to Use Analog to Efficiently Discard Precision



Sum 1024 8 bit weights X 8 bit inputs:

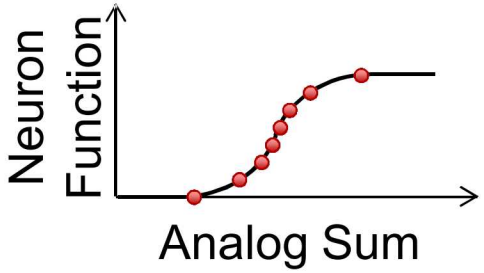
- Result has 26 bits of information!
- A 26 bit ADC would eliminate any analog advantage!

The sum can be done at full precision in analog, but a lower precision approximation is needed when digitizing

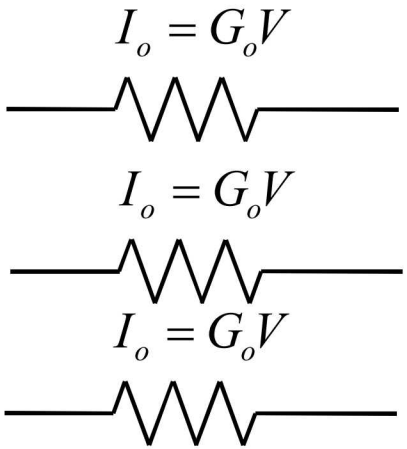
- i.e. digitize only 8 bits or fewer

To get the highest 8 bits of information, digital would need to keep a 26 bit intermediate result

Can design an ADC to choose non uniform values to digitize



The Noise Limited Energy to Read a Crossbar Column is Independent of Crossbar Size



Measure N resistors and determine the total output current with some signal to noise ratio (SNR)*

What is the minimum energy?

$$Energy = V^2 G_O \times N \times \frac{1}{\Delta f}$$

Power in each resistor × number of resistors

Determined by noise and SNR

$$\text{Thermal Noise} = \left\langle \Delta I^2 \right\rangle = N \times \left(4 k_b T \times G_o \times \Delta f\right)$$

$$SNR^2 = \frac{\left(N I_o\right)^2}{\left\langle \Delta I^2 \right\rangle}$$

$$\frac{1}{\Delta f} = 4 k_b T \times SNR^2 \times \frac{1}{V^2 G_o \times N}$$

If we double the number of resistors, we can double the speed to get the same energy and SNR.

This is because the noise scales as sqrt(N) while the signal scales as N

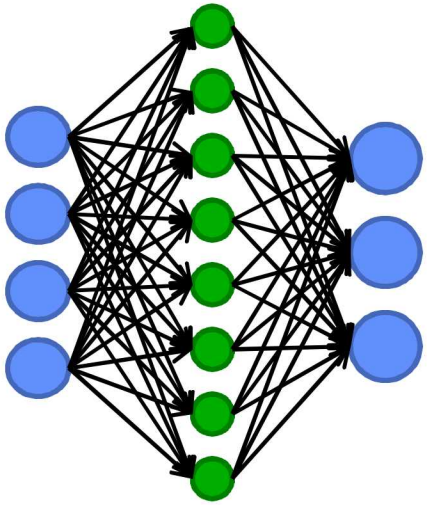
$$Energy = 4 k_b T \times SNR^2$$

*we are assuming we need some fixed precision on the output, and don't need full floating point accuracy

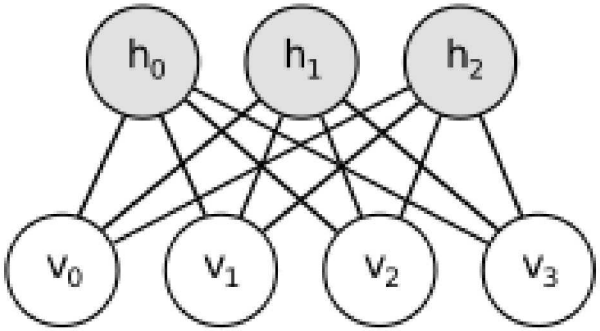
Want To Accelerate Many Different Neural Algorithms

Sandia National Laboratories

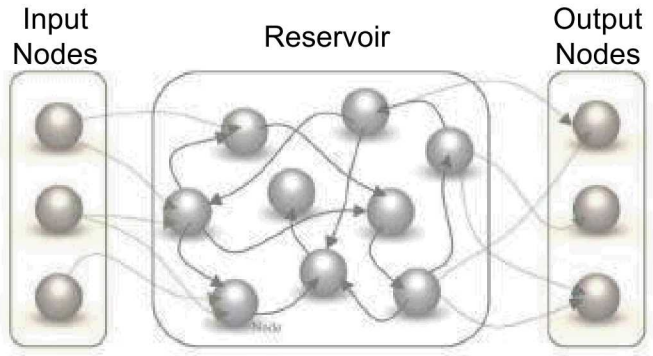
Backpropagation



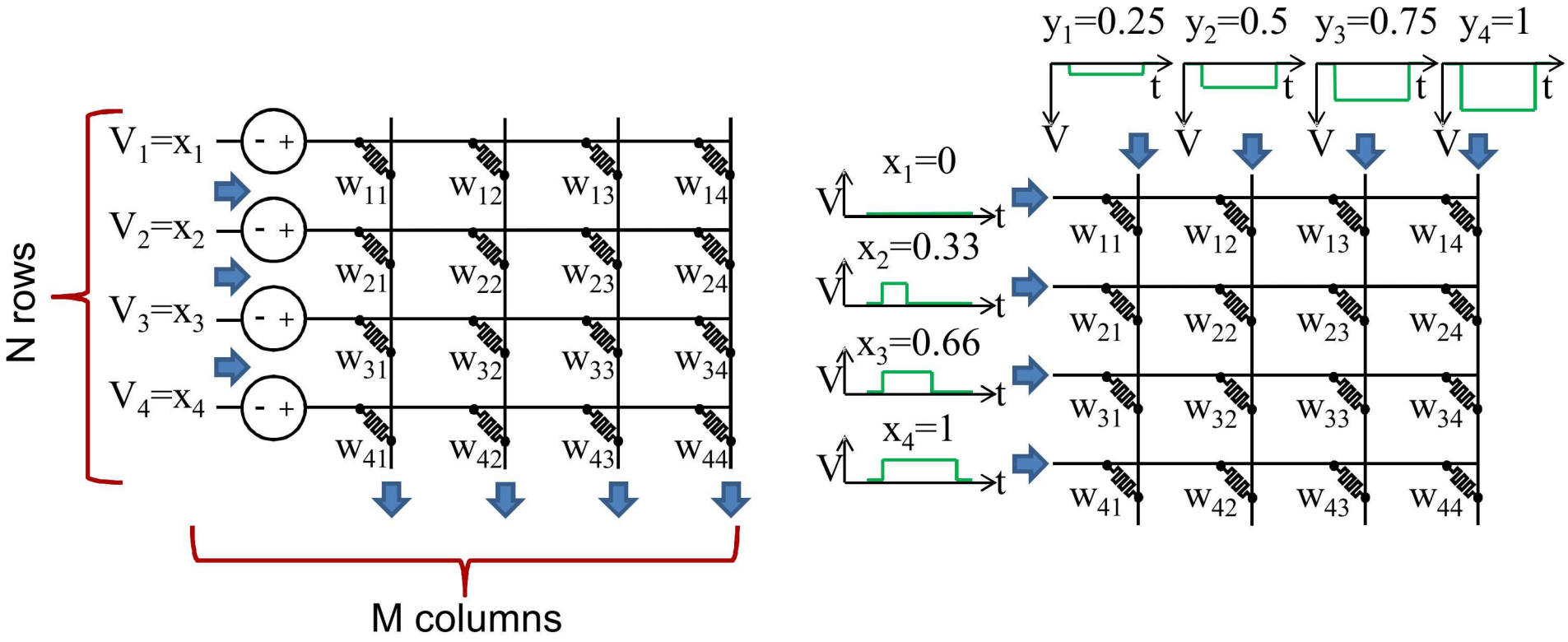
Sparse Coding



Liquid State Machine



Crossbars Can Perform Parallel Reads and Writes

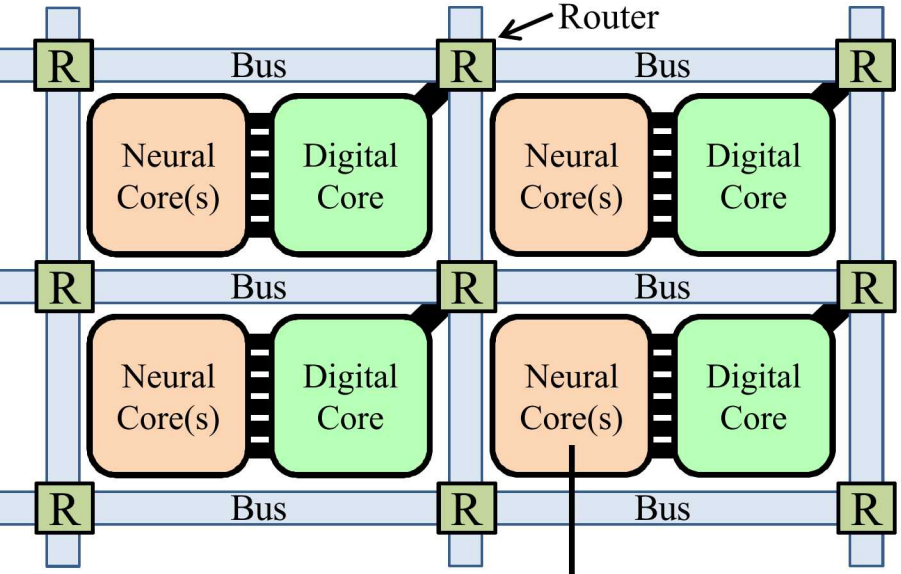


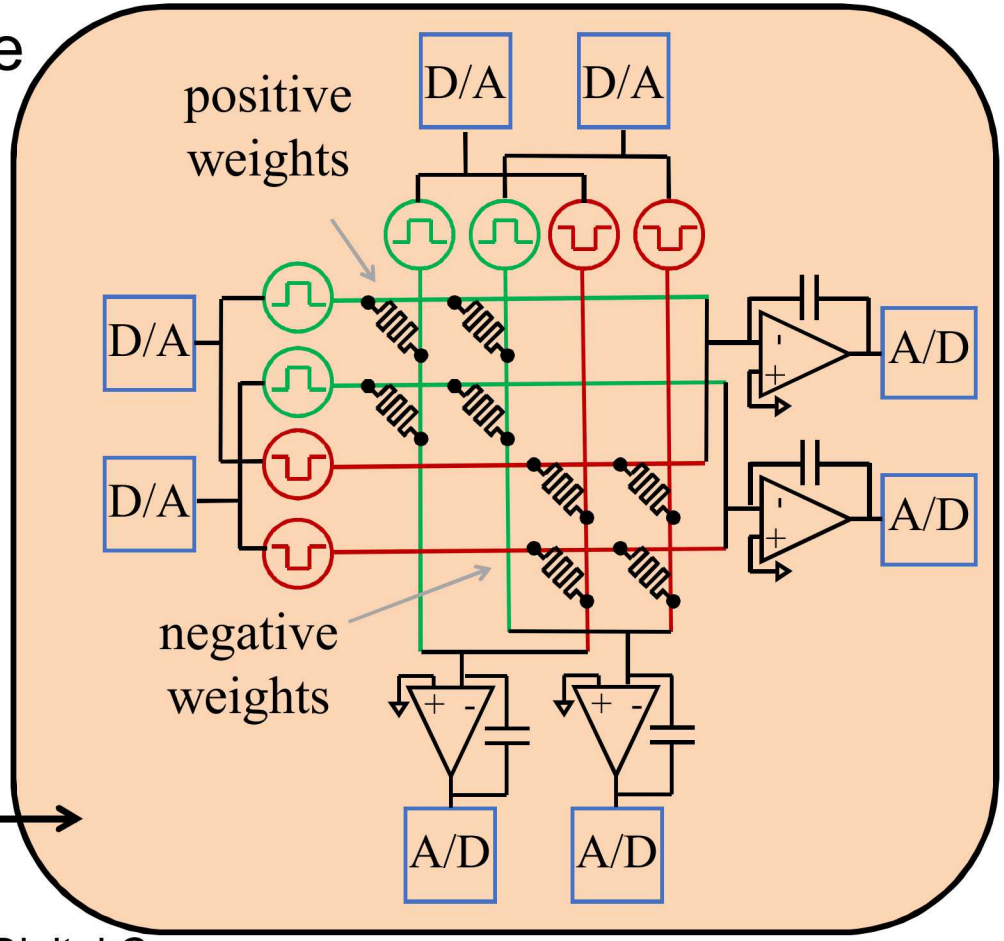
Energy to charge the crossbar is CV^2
$E \propto C \propto$ number of RRAMs $\propto N \times M$

$$E \sim O(N \times M)$$

General Purpose Neural Architecture

Run *any* neural algorithm on the same hardware





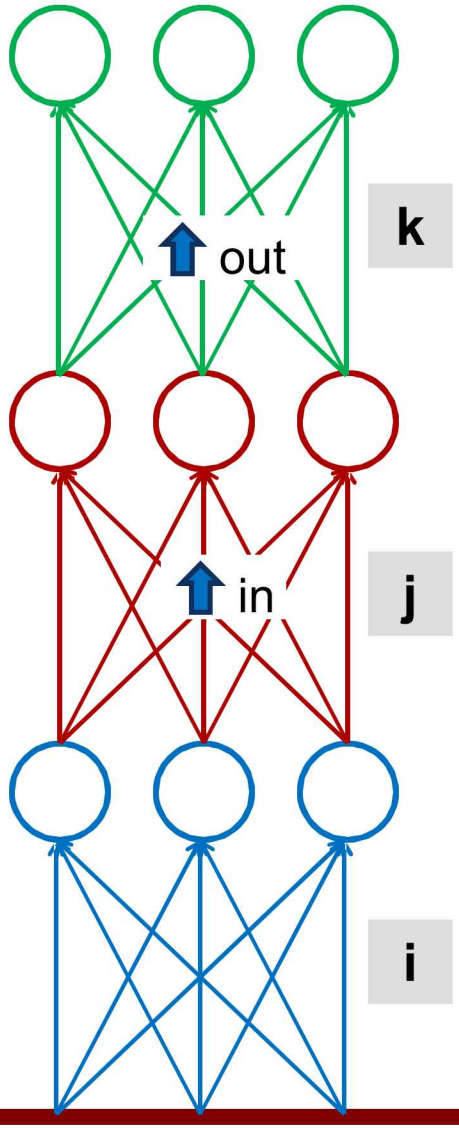
Neuromorphic core:

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

Digital Core:

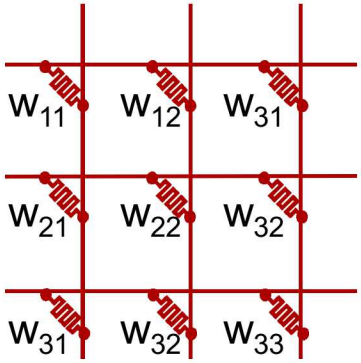
- Process neural core inputs/outputs
- For NxN crossbar, the crossbar accelerates $O(N^2)$ operations leaving only O(N) operations for the digital core

Can Run Neural Networks on this Architecture



O(N^2) Operations

Neural Core



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

y_i

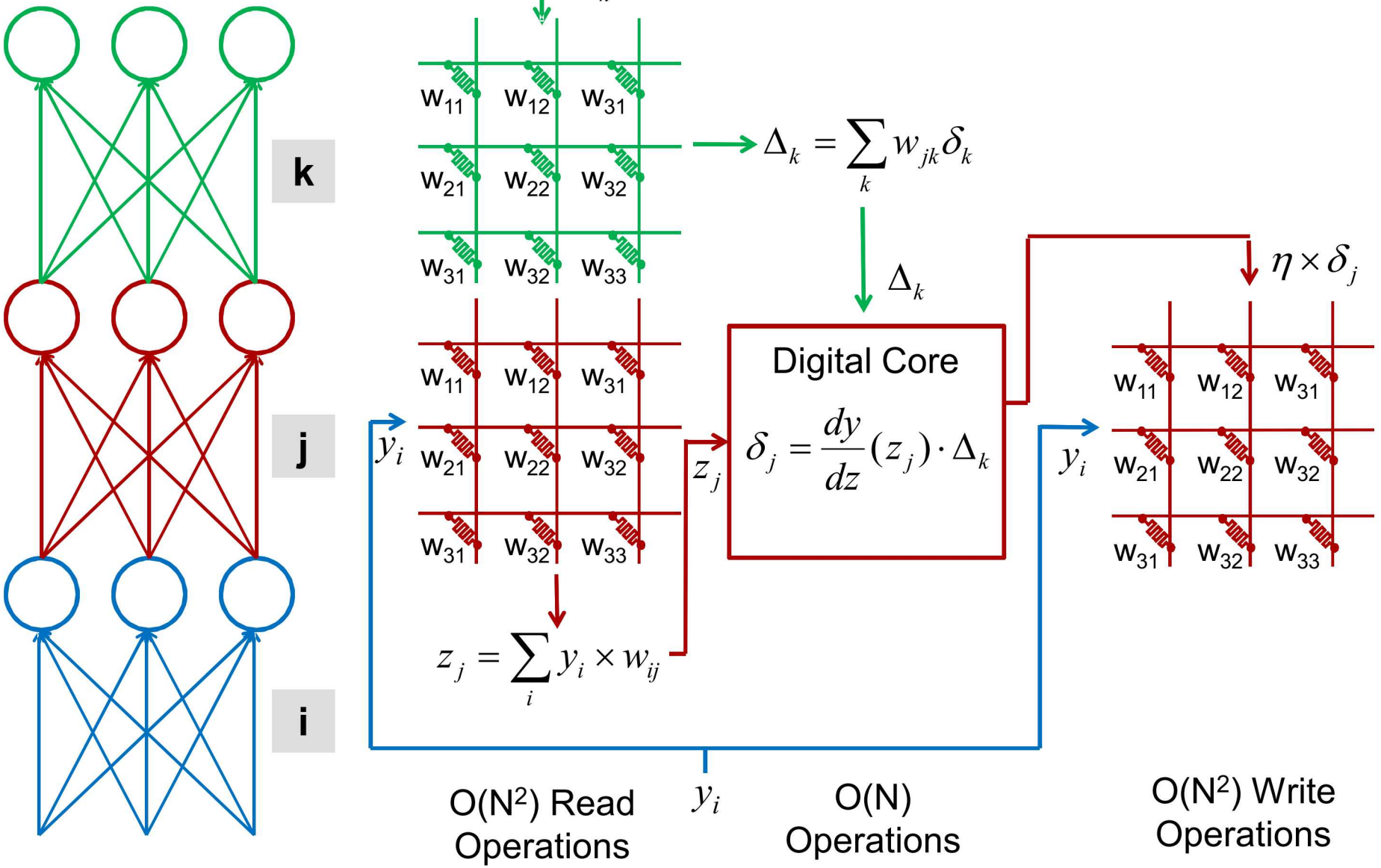
O(N) Operations

y_j

Digital Core

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

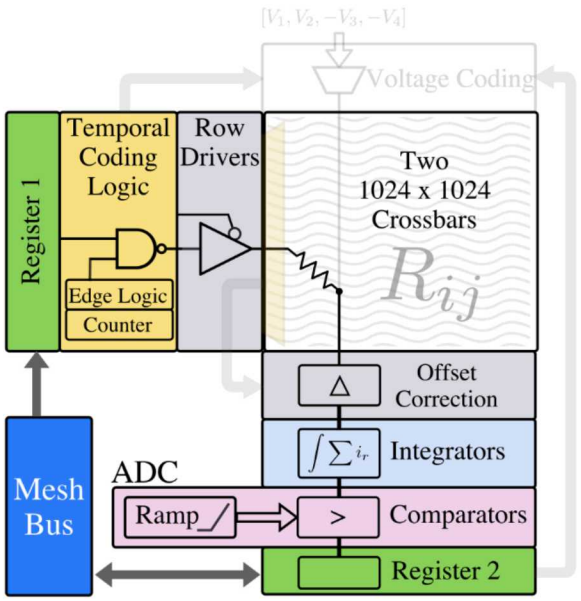
Back Propagation



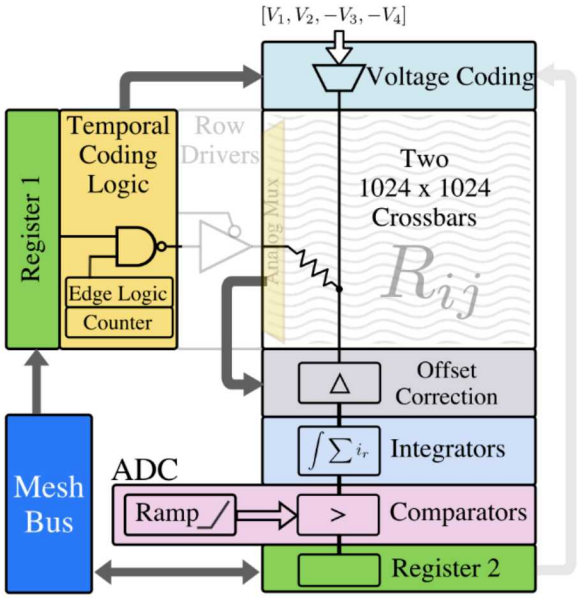
Design & Model Detailed Architecture

Sandia National Laboratories

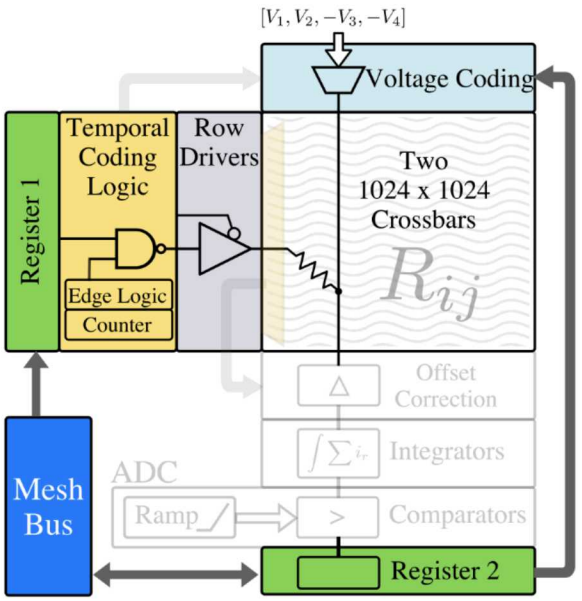
Vector Matrix Multiply



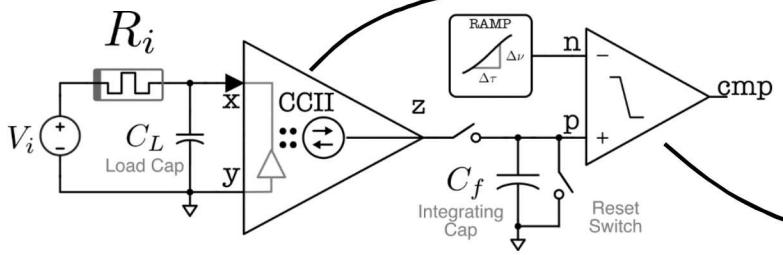
Matrix Vector Multiply



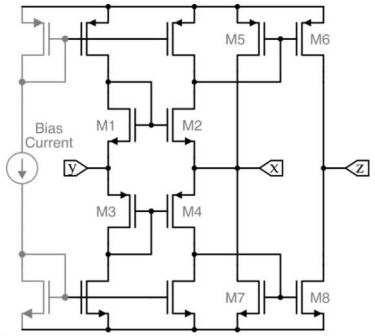
Outer product Update



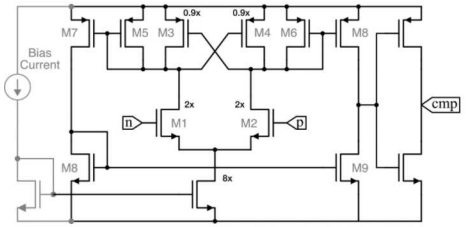
Neuron Circuitry



Current Conveyor Based Integrator



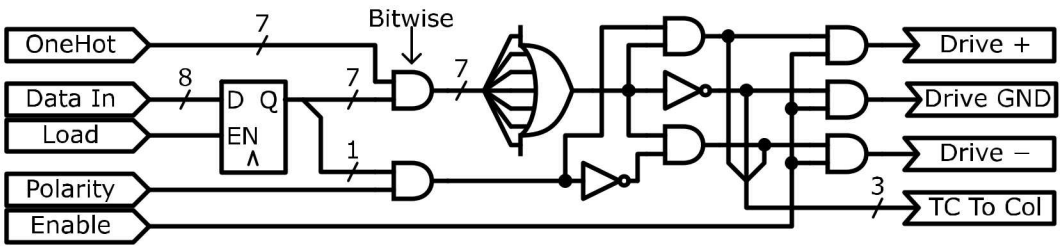
Comparator



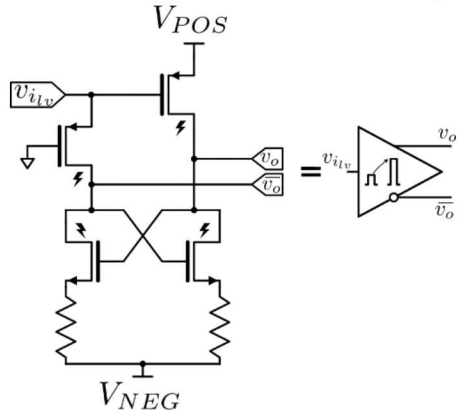
Row & Column Driver Circuitry

Sandia National Laboratories

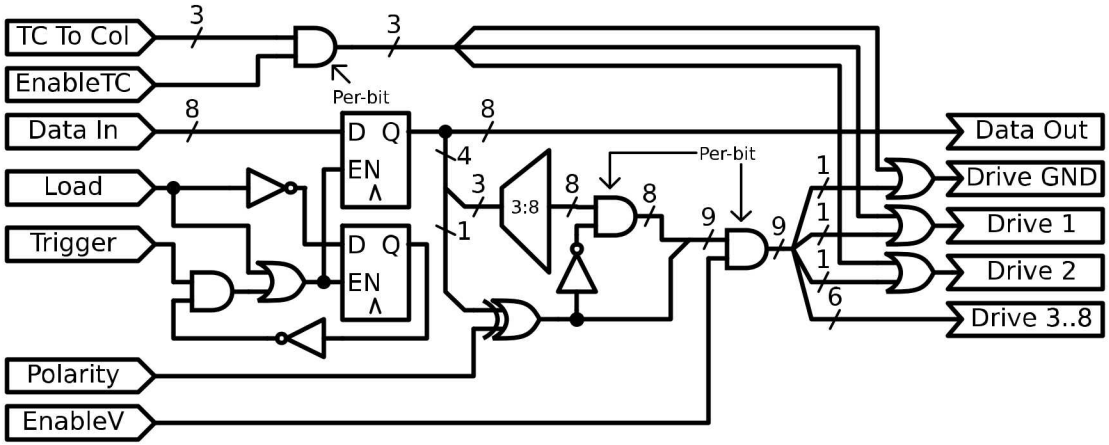
Row Driver Logic



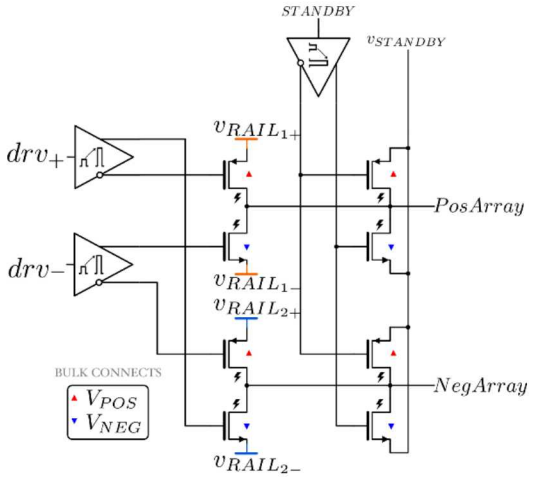
Voltage level shifter (drive high V transistor with low V)



Column Driver Logic



Array driver pass transistors

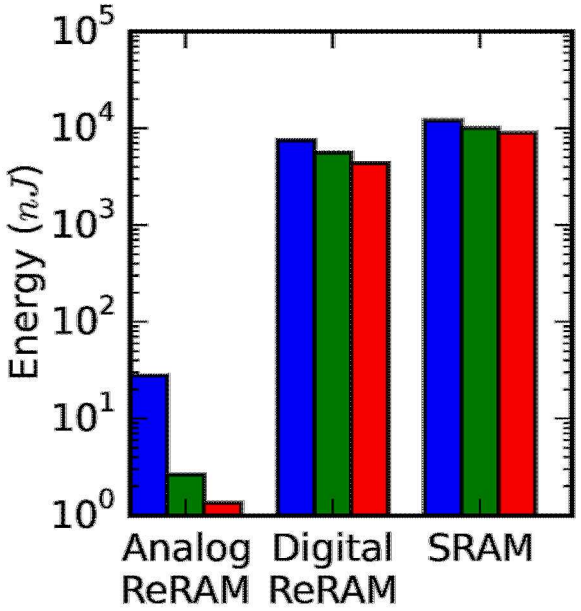


Compare Architectures

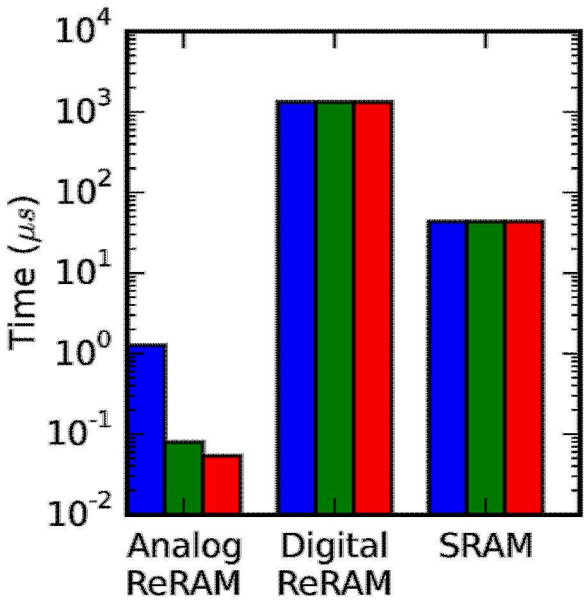
1024 x1024 = 1M array operations, sum over 1 training cycle, 3 operations:

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

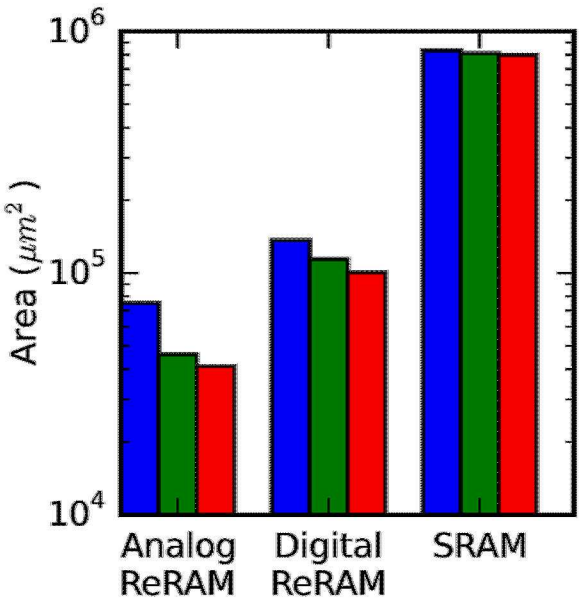
Energy
430 – 6,900X over SRAM

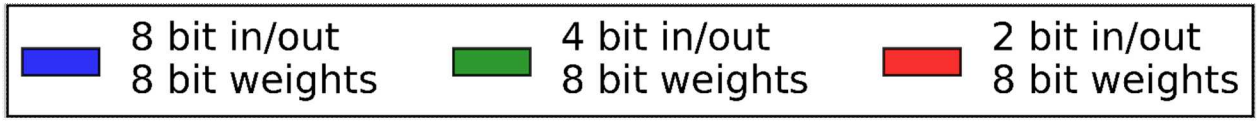


Latency
35 – 800X over SRAM



Area
11 – 20X over SRAM



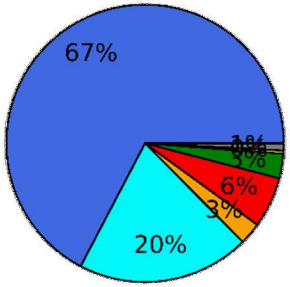
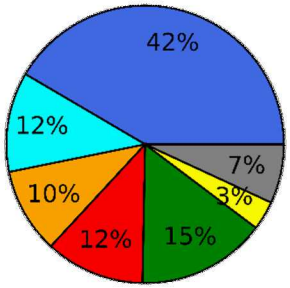
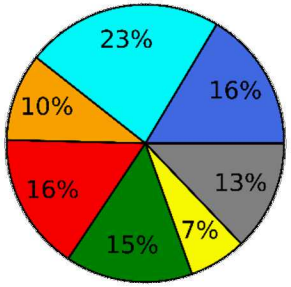
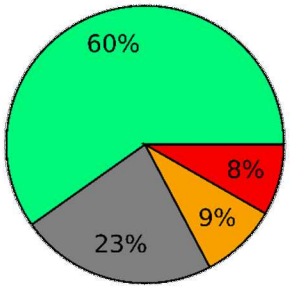
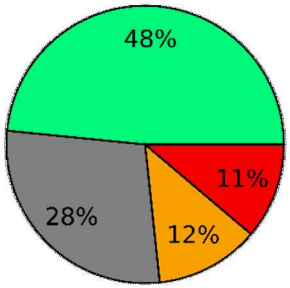
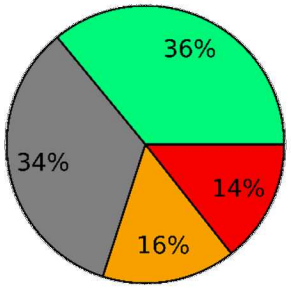
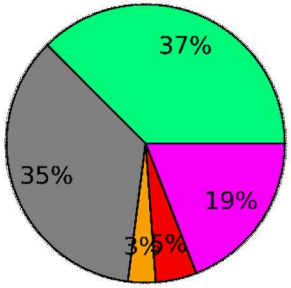
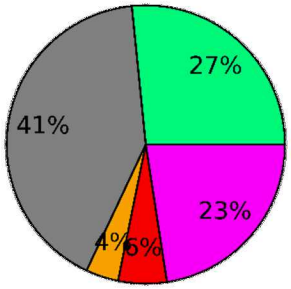
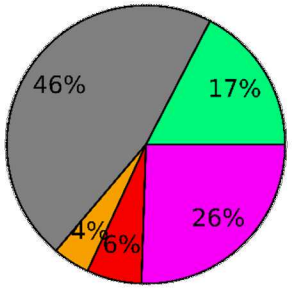


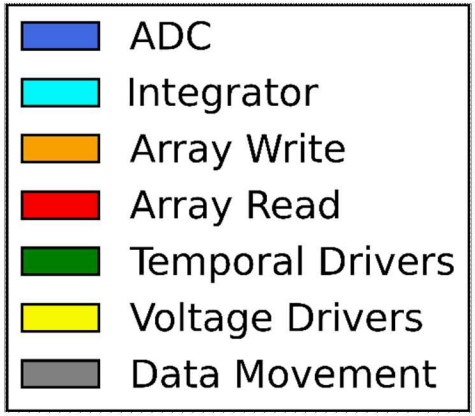
Used a commercial 14/16 nm PDK

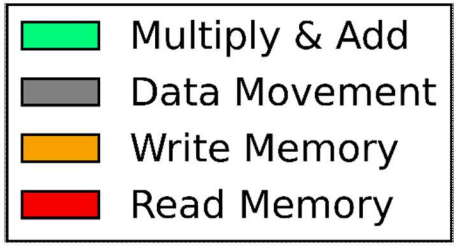
***Requires 100 MΩ on state devices

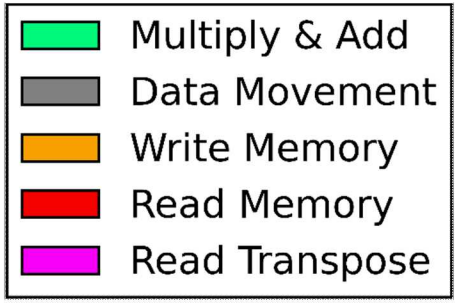
Neural Core Energy Analysis

Sandia National Laboratories

	8 bits In/out 8 bit weights	4 bits In/out 8 bit weights	2 bits In/out 8 bit weights
Analog ReRAM	 28 nJ	 2.7 nJ	 1.3 nJ
Digital ReRAM	 7,520 nJ	 5,580 nJ	 4,340 nJ
SRAM	 12,010 nJ	 10,150 nJ	 8,970 nJ

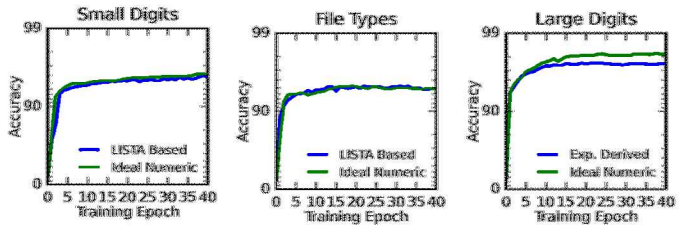






Multiscale Model of a Neural Training Accelerator

Sandia National Laboratories

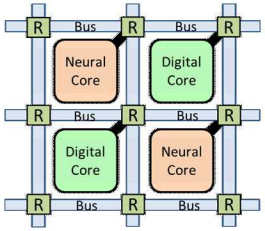


Target Algorithms

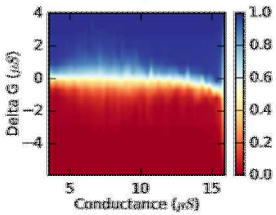
- Deep Learning
- Sparse Coding
- Liquid State Machines

Algorithms

Architecture

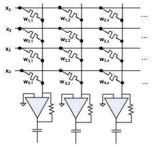


Modified McPAT/CACTI: Model performance and energy requirements



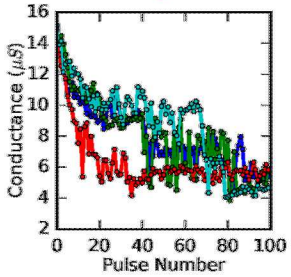
CROSS SIM

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



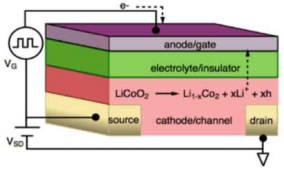
Circuits

Sandia's Xyce Circuit Sim: Simulate crossbar circuits based on our devices

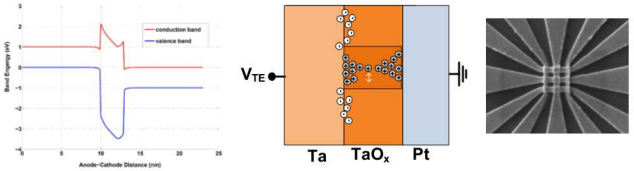


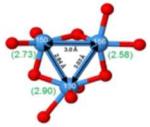
Memristor fabrication and measurements in MESAFab

Devices



Drift-diffusion model of ReRAM band diagram & transport (REOS, Charon)

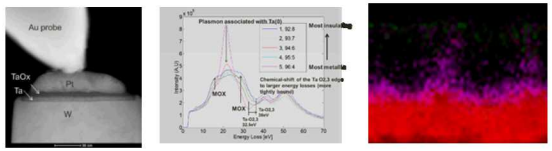




DFT of model of oxide physics, bands

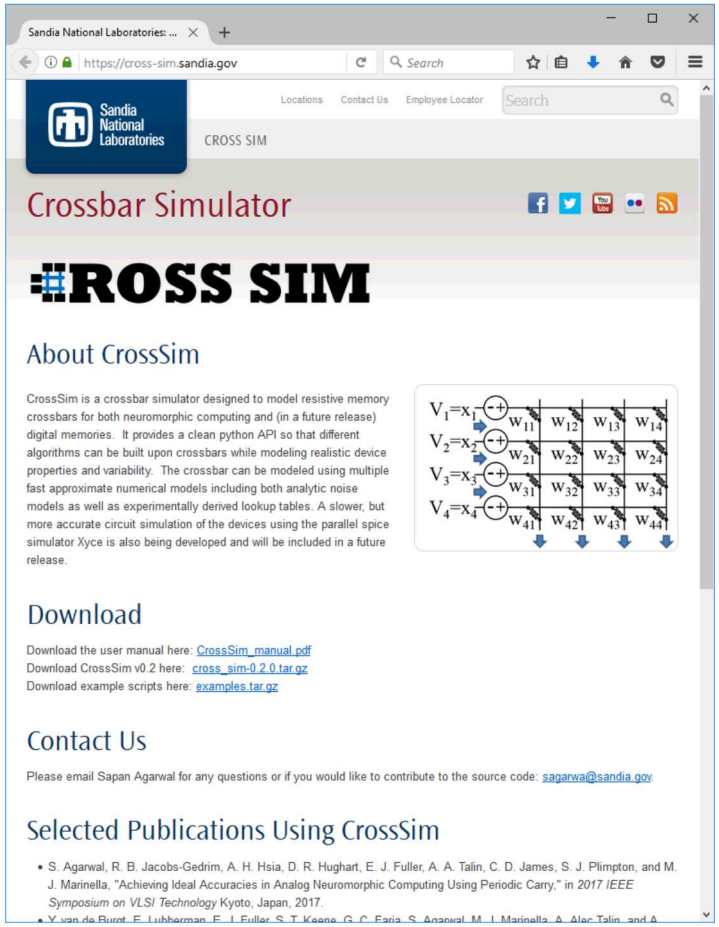
Materials

In situ TEM of filament switching: Use DFT model to interpret EELS signature



#ROSS SIM

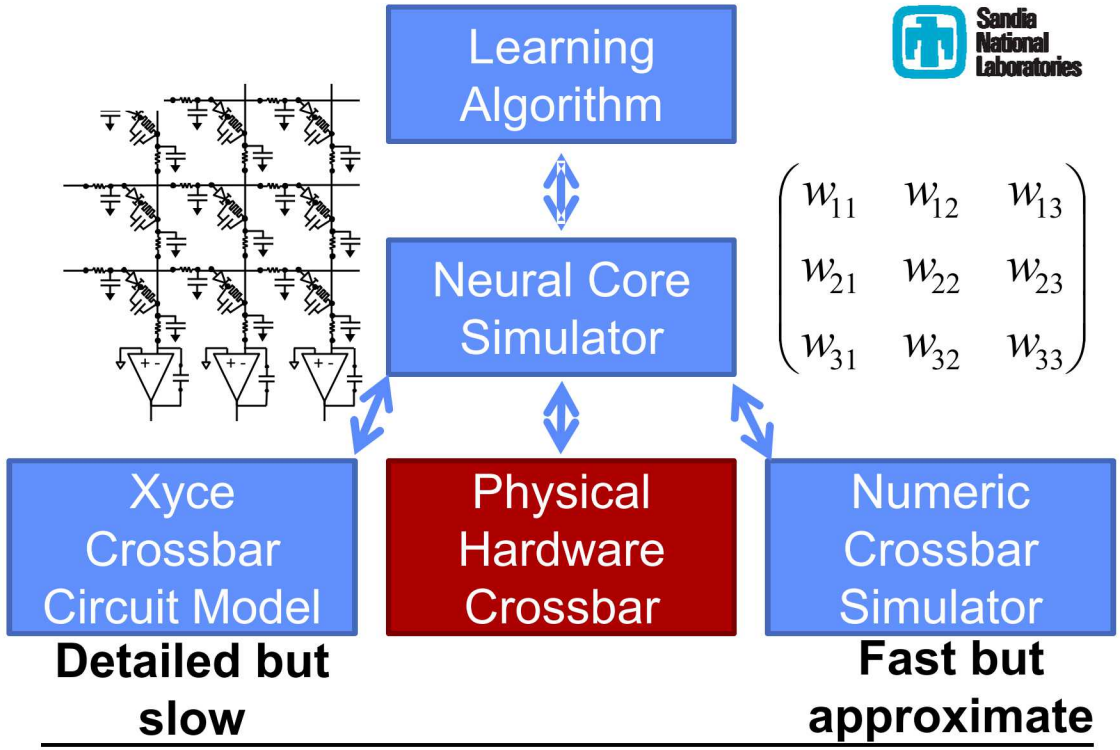
https://cross-sim.sandia.gov

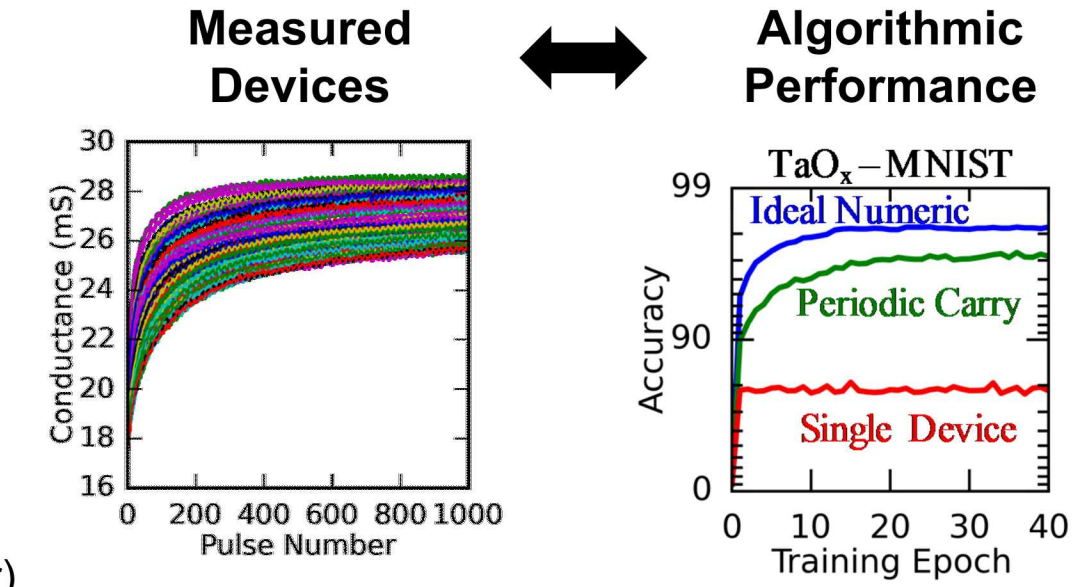


Simple Python API:

 # Do a matrix vector multiplication

result = neural_core.run_xbar_mvm(vector)





Simple API to model crossbars

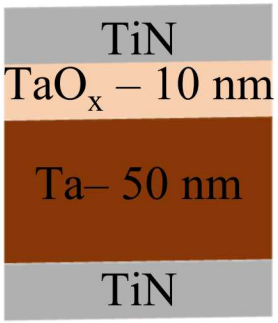
Sandia National Laboratories

```
# ************** set parameters defining the crossbar

params.algorithm_params.weights.sim_type = “XYCE” # Use a XYCE based sim
params.algorithm_params.weights.maximum = 10  # clipping limits
params.algorithm_params.weights.minimum = -10  # clipping limits
params.xyce_parameters.xbar.device.TAHA_A1 = 4e-4  # Xyce Parameters
…
…

# ************** API for running neural operations
#                        All crossbar details are transparent to the user

# Create a neural_core object that models a crossbar
neural_core = MakeCore(params=params)

neural_core.set_matrix(weights) # set the initial weights
result = neural_core.run_xbar_vmm(vector) # Do a vector matrix multiply
result = neural_core.run_xbar_mvm(vector) # Do the transpose, a matrix vector mult.
neural_core.update_matrix(vector1,vector2) # Do an outer product update
```

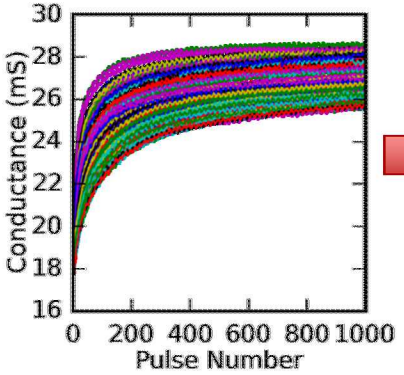
Go from Measurement to Accuracy

Sandia National Laboratories

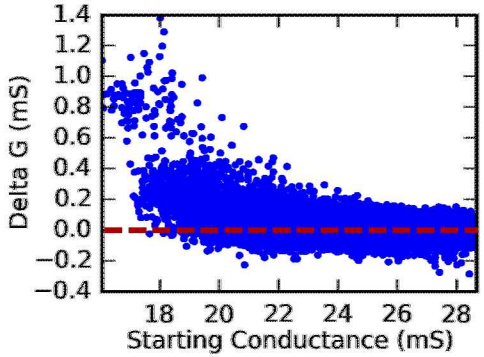
Fabricate Device



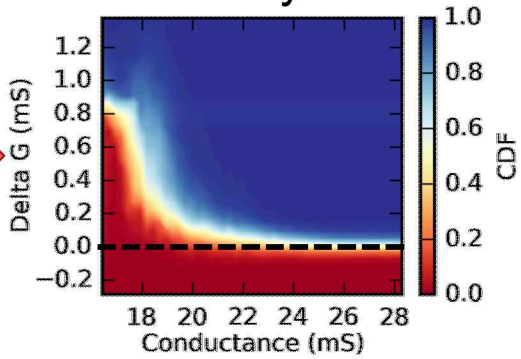
Measured Pulsing

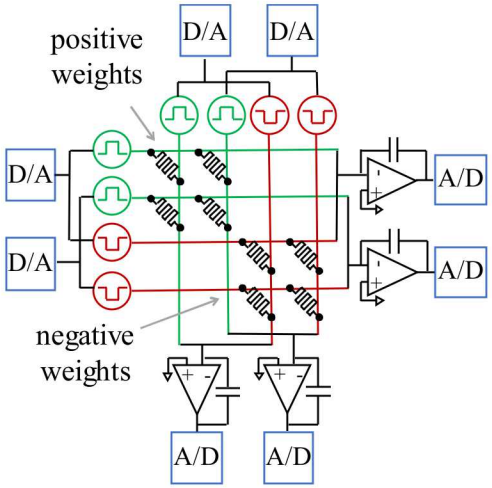


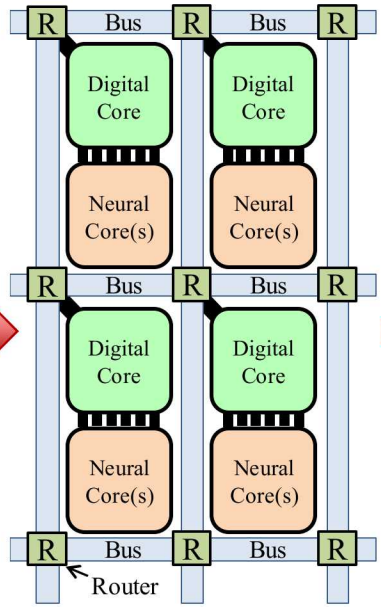
ΔG Scatterplot

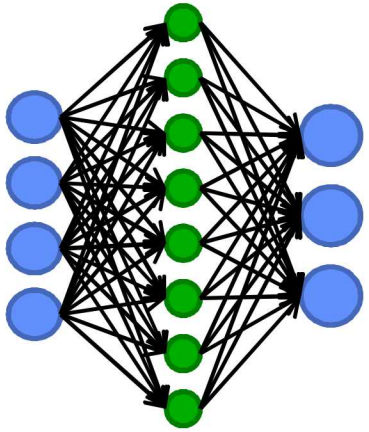


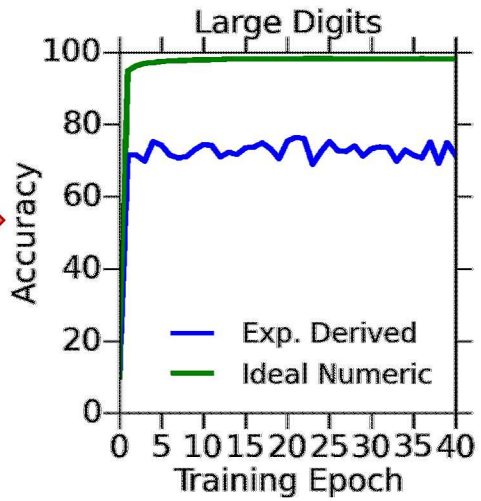
Cumulative Probability of ΔG





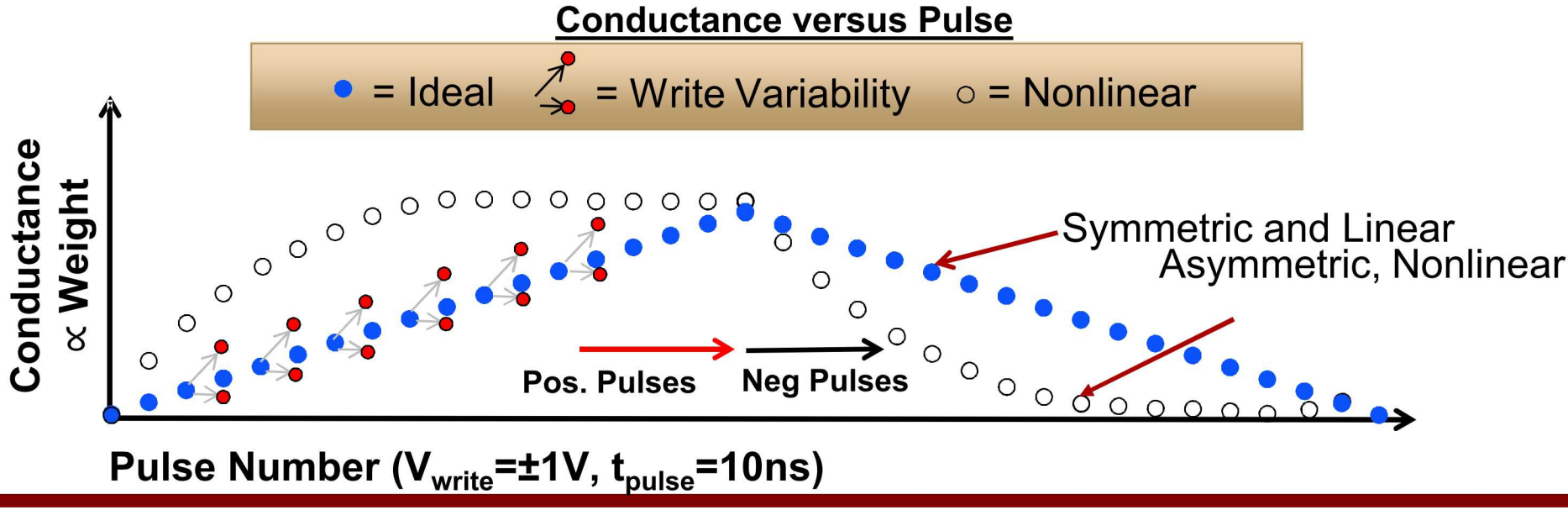




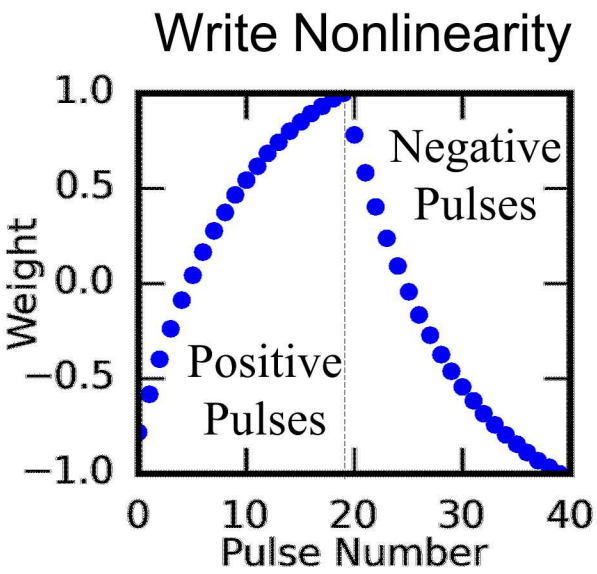


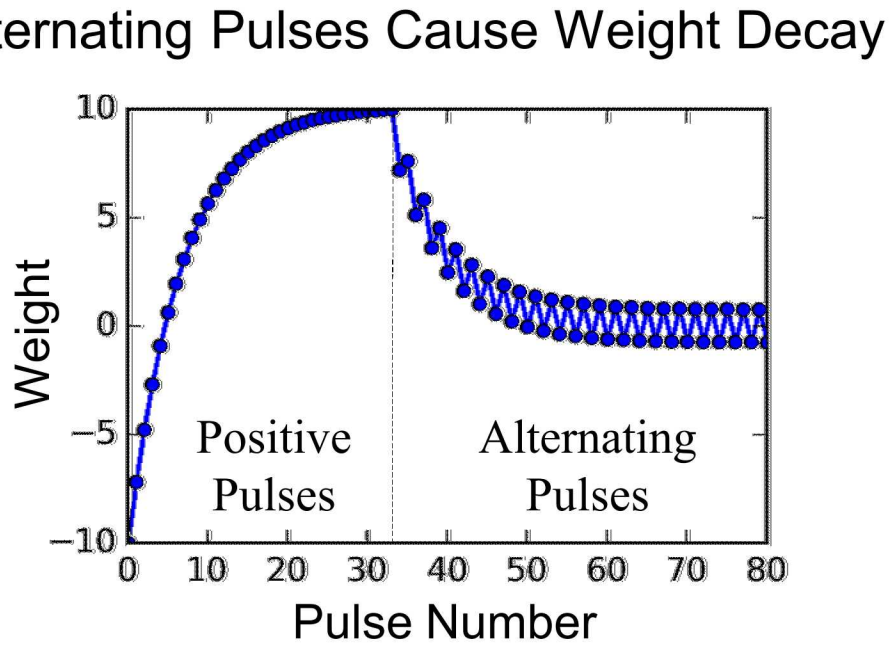
Experimental Device Nonidealities

- **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**

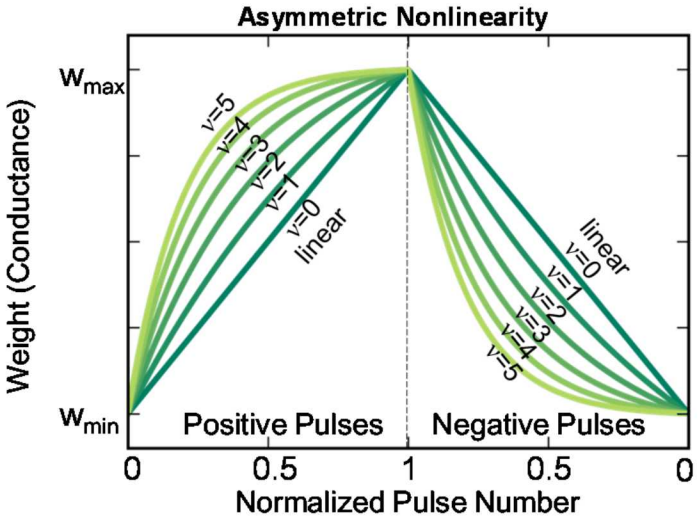


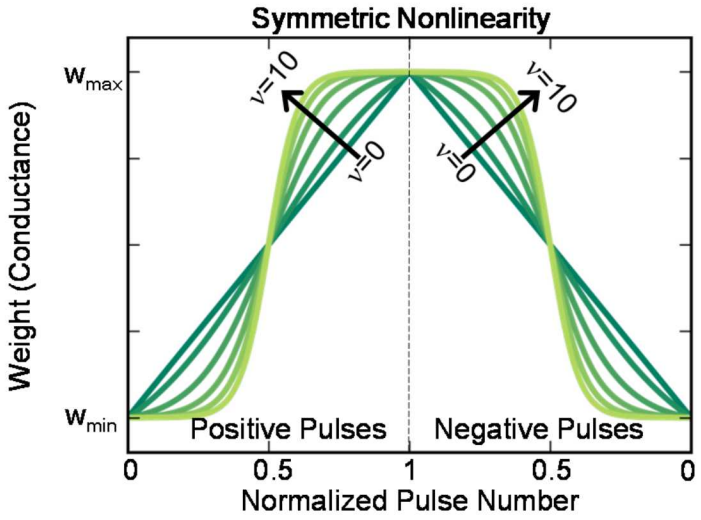
Asymmetric Write Nonlinearity Causes Weight Decay



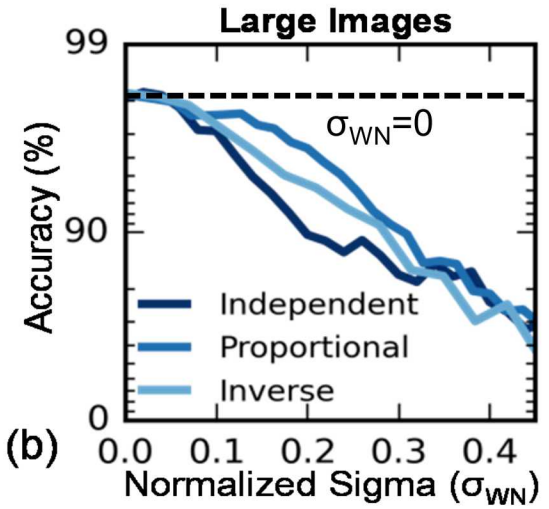


Modeling Device Requirements





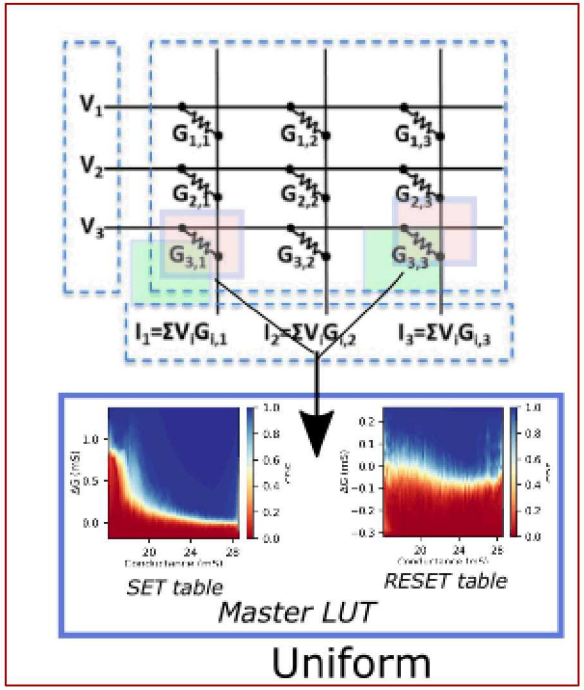
	MNIST
Read Noise σ (% Range)	5%
Write Noise σ (% Range)	0.4%
Asymmetric Nonlinearity (ν)	0.1
Symmetric Nonlinearity (ν)	5
Maximum Current	13 nA



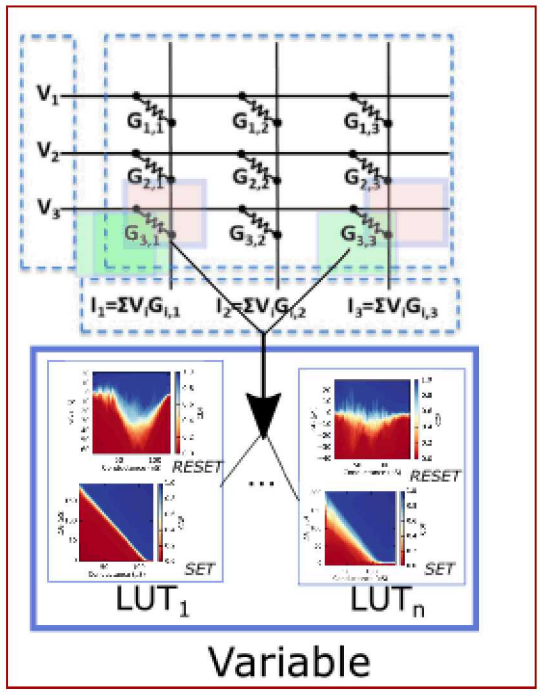
A realistic variability approach

- 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)

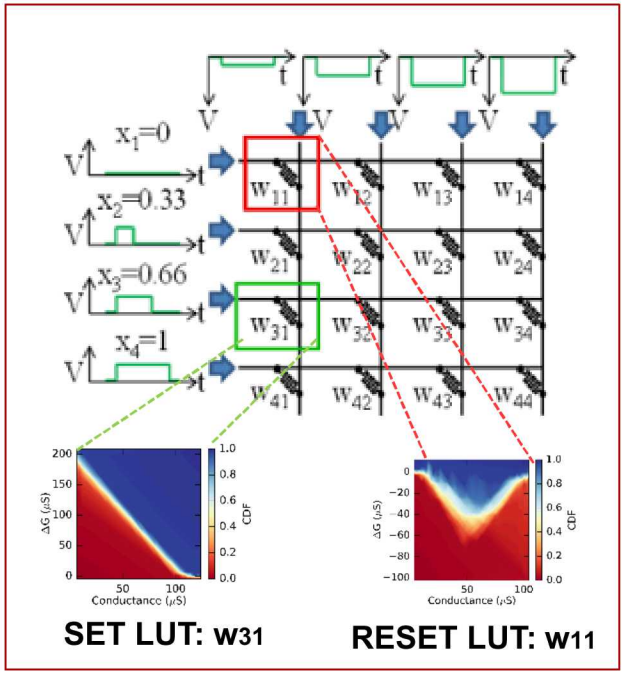
Old



New

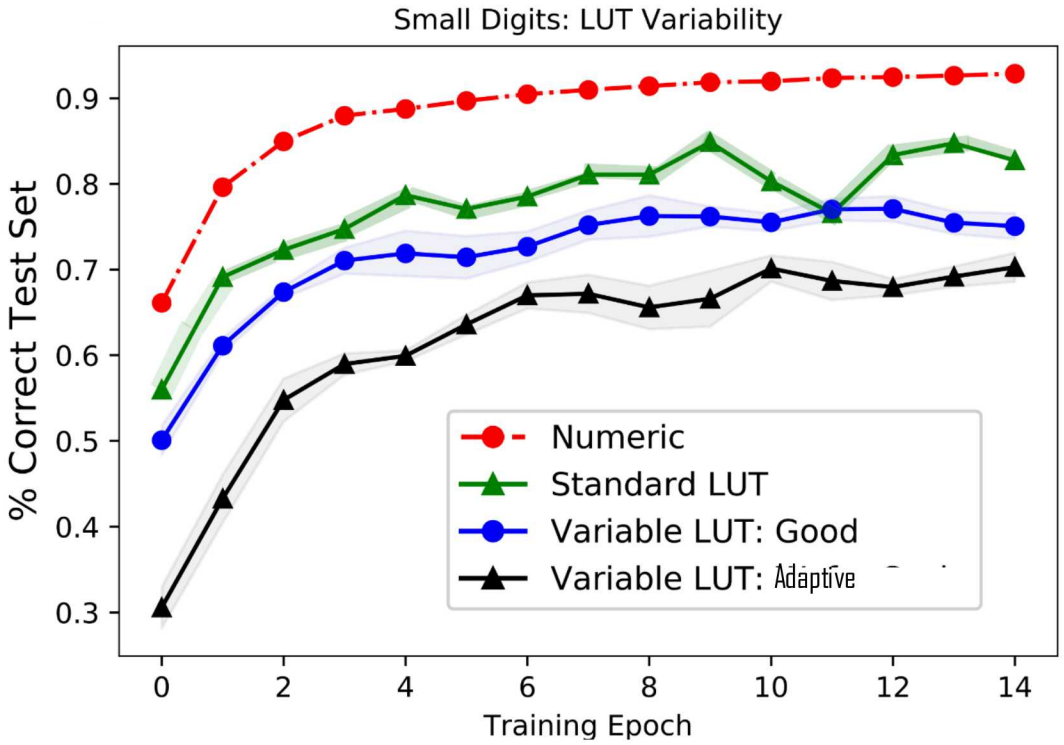


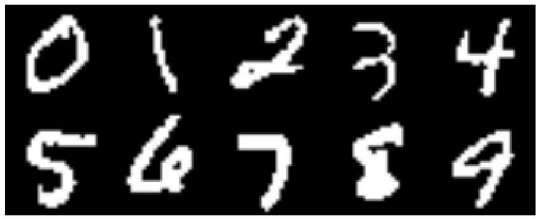
Online Updates



Implications on Neural algorithm accuracies:
Small Digits (OCR Database)

- Loss between uniform and Good Devices: **~4-7%**
- Loss between uniform and all functioning devices (adaptive): **~13-16%**

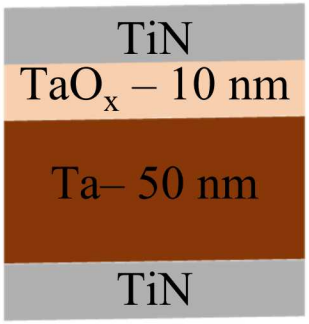




Data set	# Training Examples	# Test Examples	Network Size
UCI Iris Dataset [1]	100	50	4x8x3
UCI Small Digits[1]	3,823	1,797	64×36×10
File Types[2]	4,501	900	256×512×9
MNIST Large Digits[3]	60,000	10,000	784×300×10

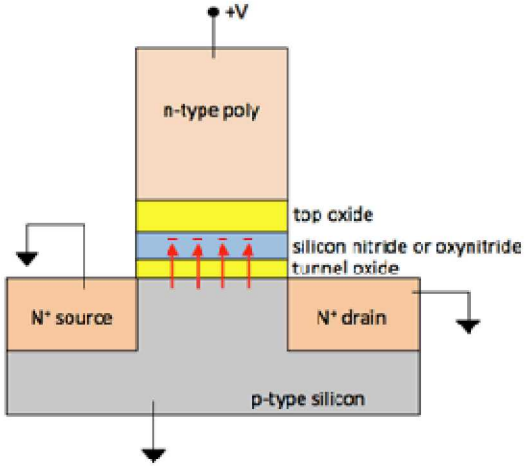
Compare Analog Devices

ReRAM



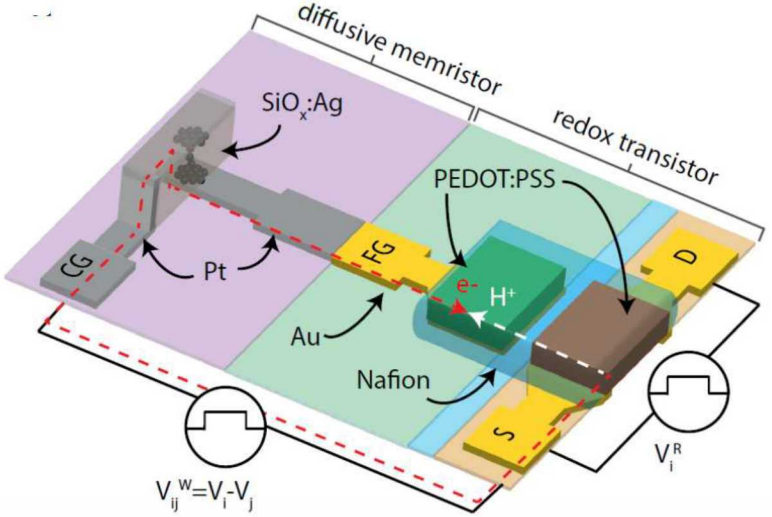
R. B. Jacobs-Gedrim *et al.*, "Impact of Linearity and Write Noise of Analog Resistive Memory Devices in a Neural Algorithm Accelerator," IEEE International Conference on Rebooting Computing (ICRC) Washington, DC, November 2017.

SONOS Silicon-Oxygen-Nitrogen-Oxygen-Silicon



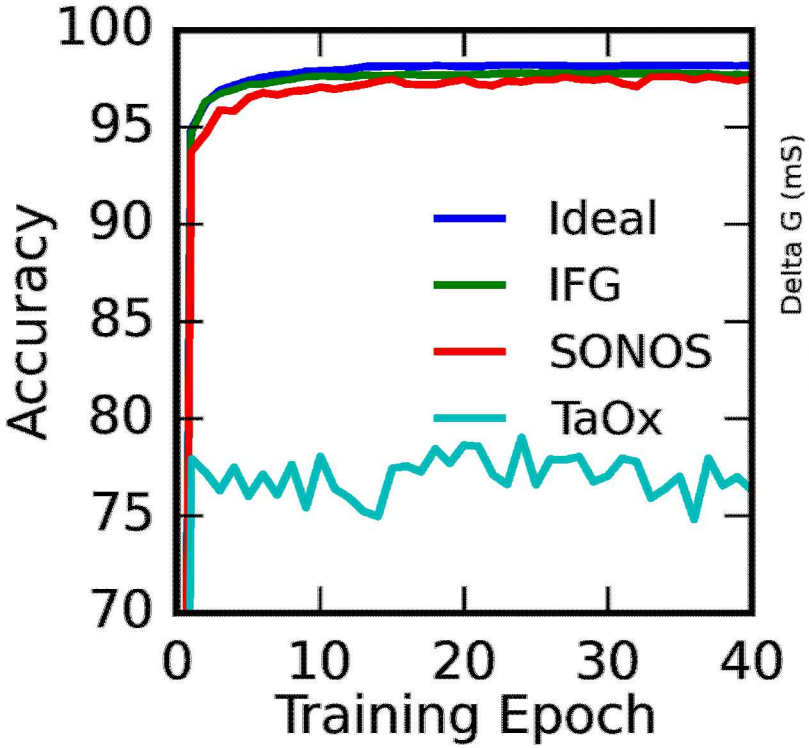
S. Agarwal *et al.*, "Using Floating Gate Memory to Train Ideal Accuracy Neural Networks," *IEEE Journal of Exploratory Solid-State Computational Devices and Circuits,* 2019

Ionic Floating-Gate Memory

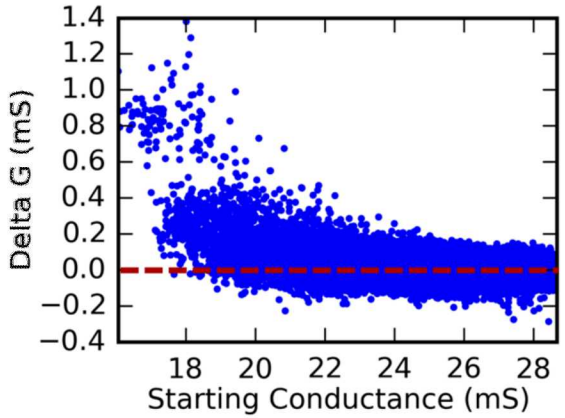


E. J. Fuller et al., "Li-Ion Synaptic Transistor for Low Power Analog Computing," *Advanced Materials*, vol. 29, no. 4, p. 1604310, 2017
E. J. Fuller et al., under review

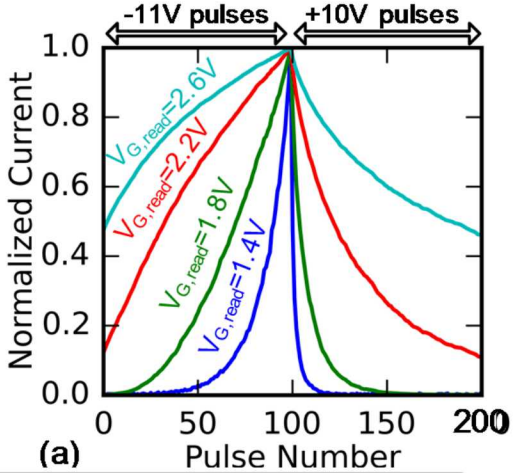
Three Terminal Devices Tend to Have Higher Accuracy

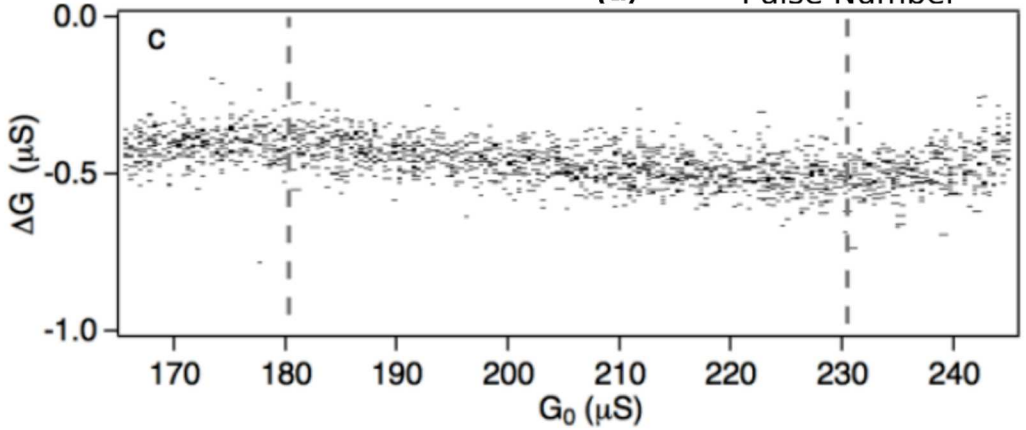


ReRAM



SONOS

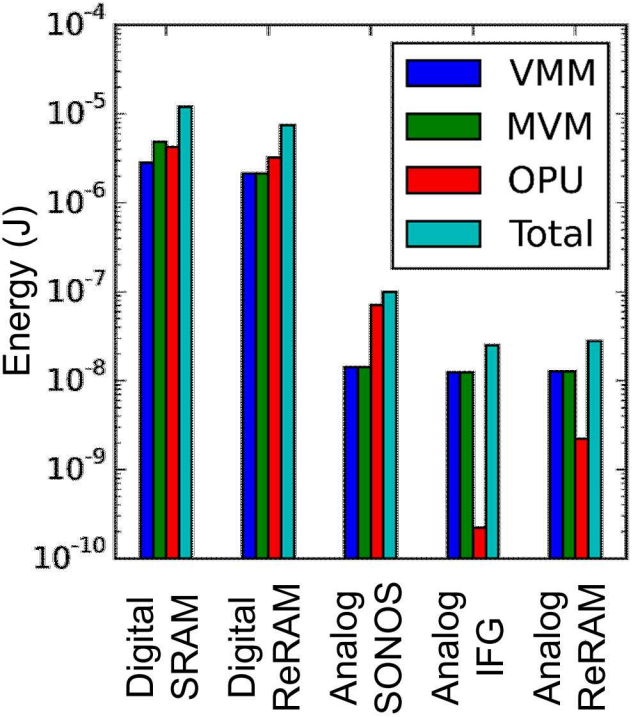




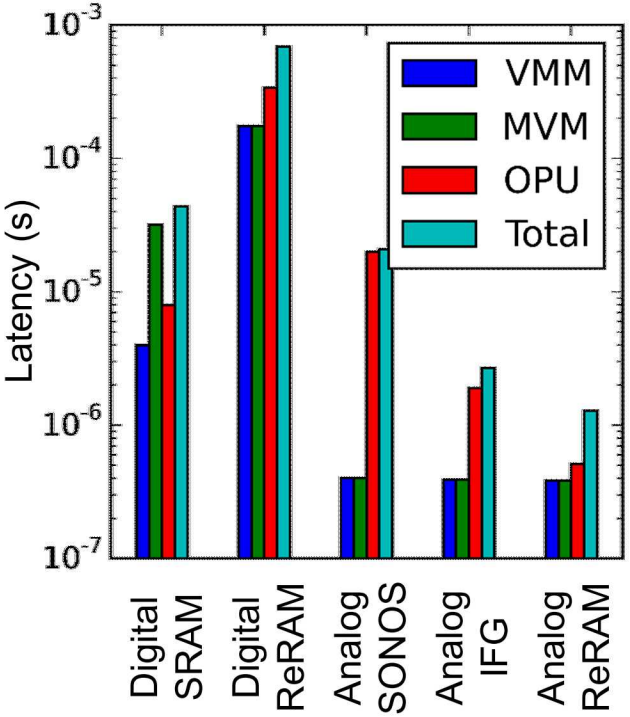
Ionic Floating-Gate

Compare Architectural Advantages

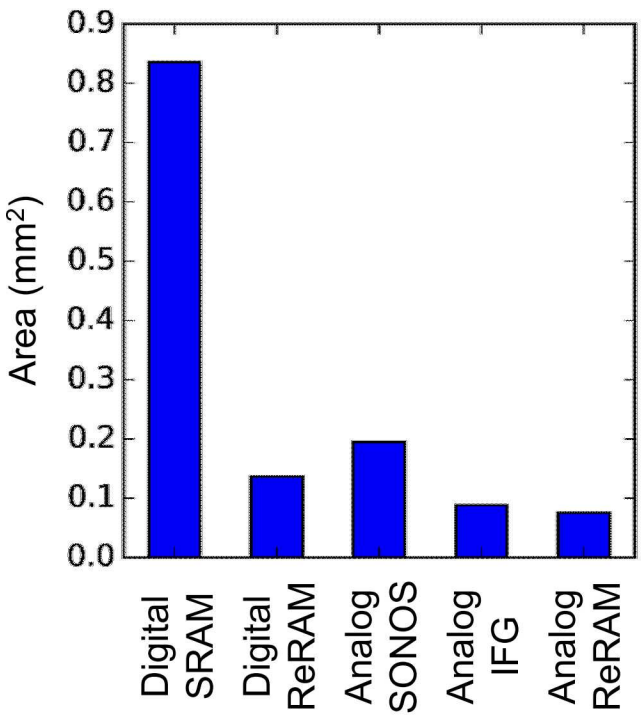
120-430X Energy Advantage



2-34X Latency Advantage



5-11X Area Advantage



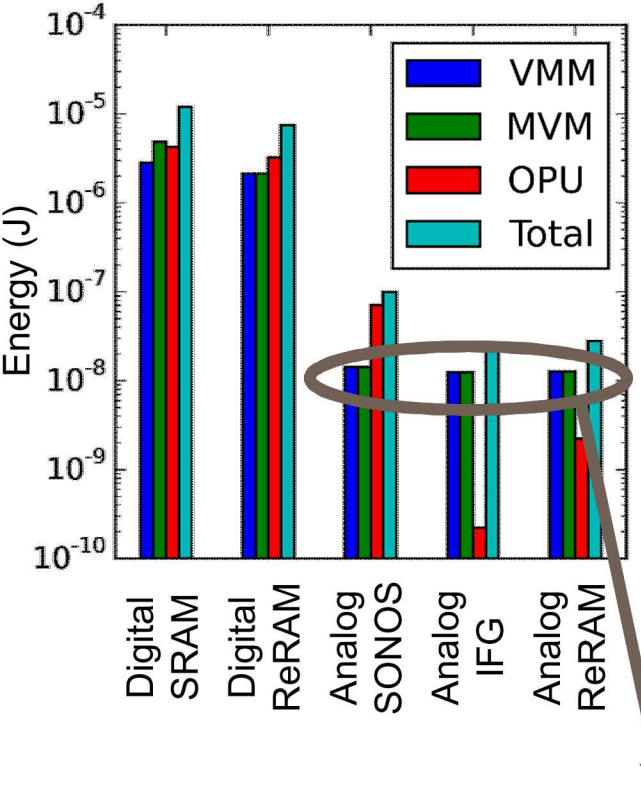
1024 x1024 = 1M array operations, sum over 1 training cycle, 3 operations:
- Vector Matrix Multiply - Matrix Vector Multiply - Outer Product Update

Used a commercial 14/16 nm PDK

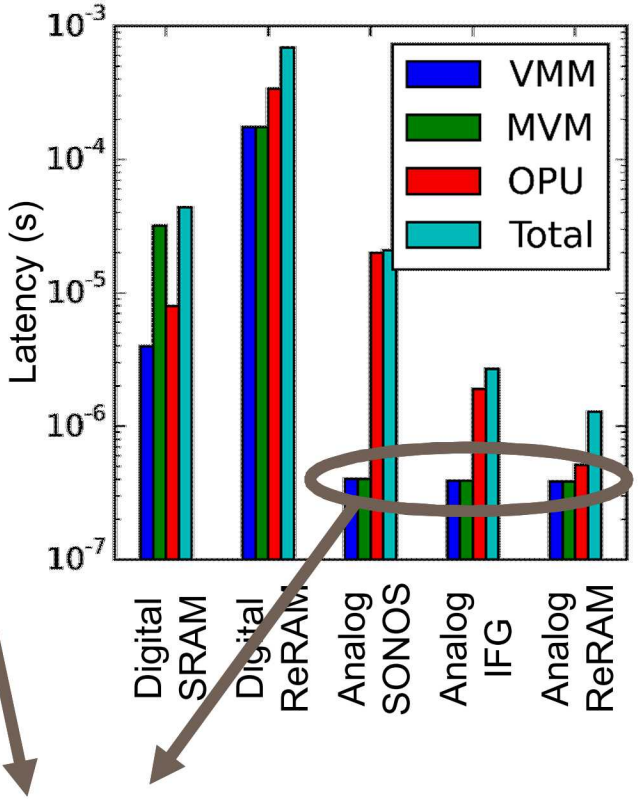
***Requires 100 MΩ on state devices

Compare Architectural Advantages: Vector Matrix Multiply

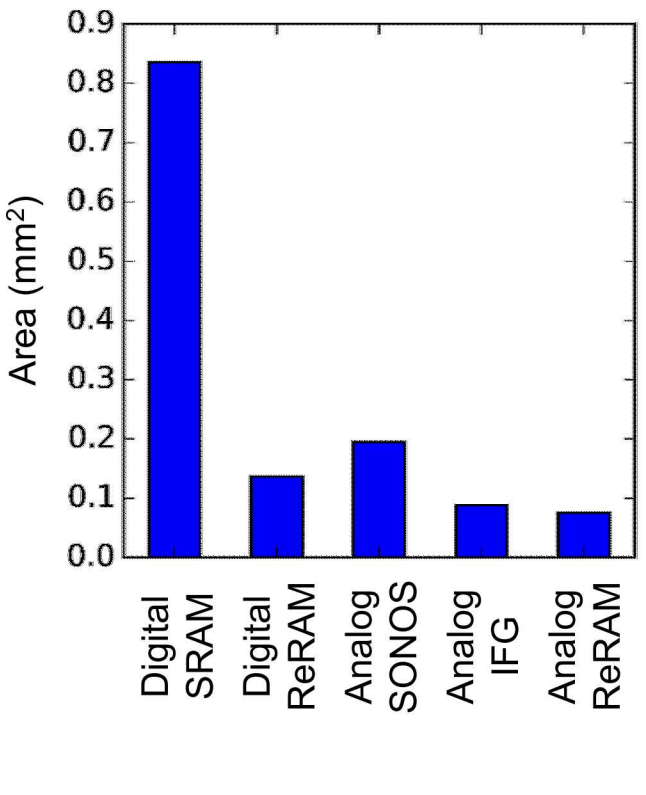
120-430X Energy Advantage



2-34X Latency Advantage



5-11X Area Advantage

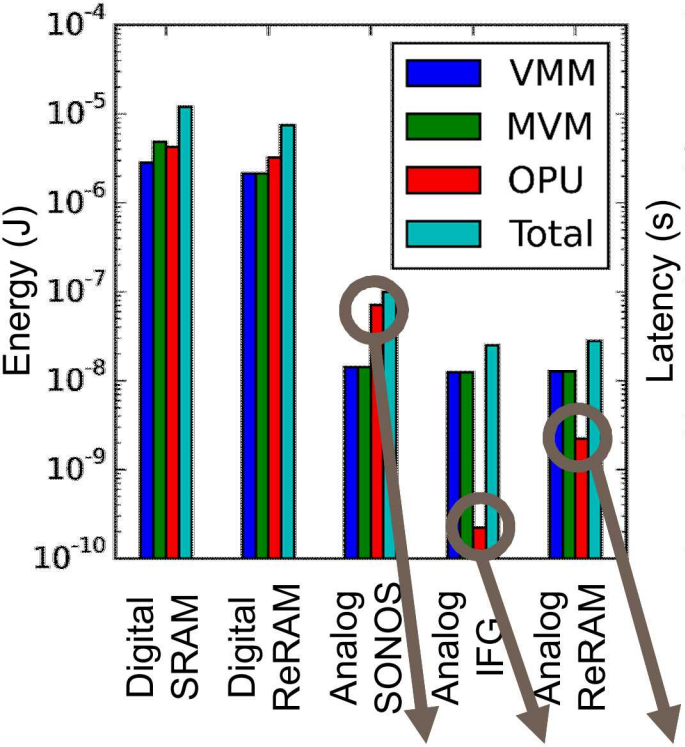


All Analog Vector Matrix Multiply and Matrix Vector Multiply have same energy and latency

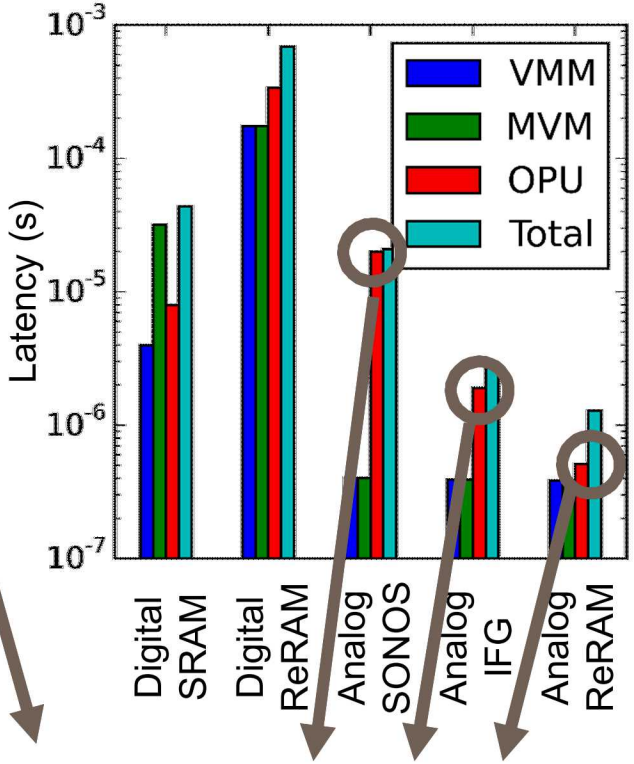
- Entirely dominated by ADC, device properties irrelevant

Compare Architectural Advantages: Outer Product Update

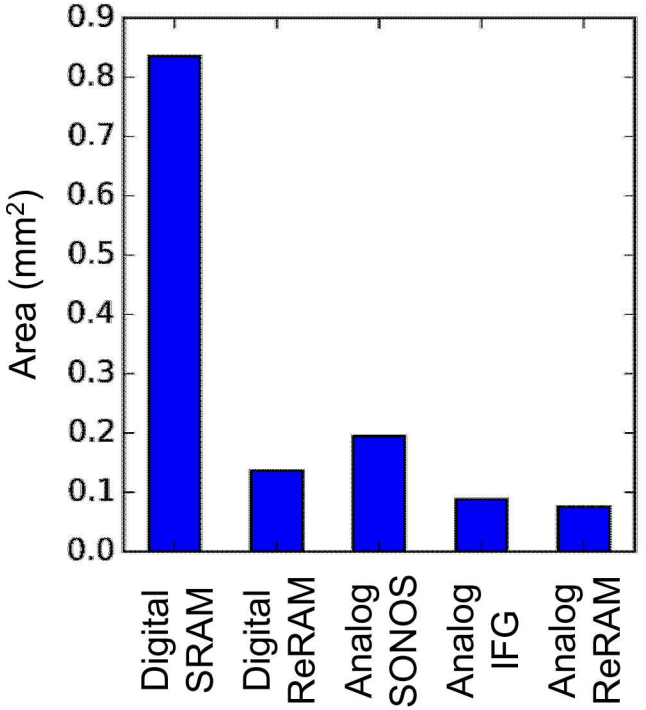
120-430X Energy Advantage



2-34X Latency Advantage



5-11X Area Advantage

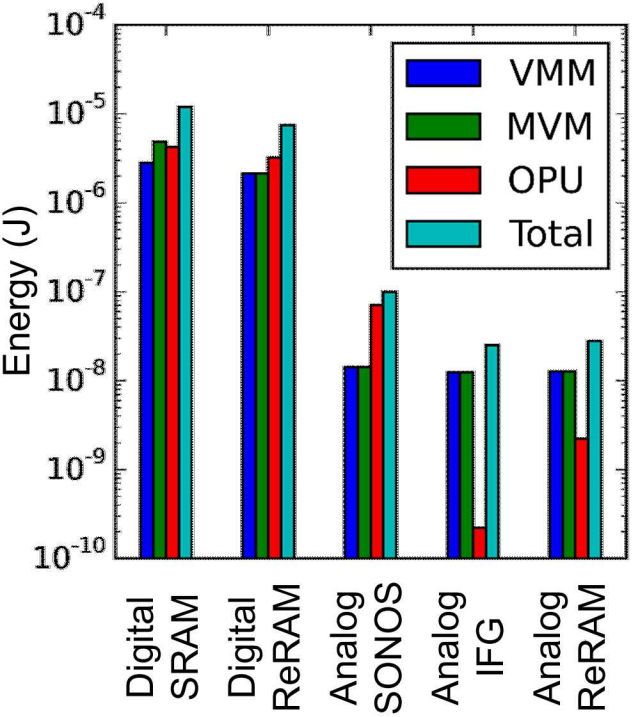


Outer Product Update is device dependent

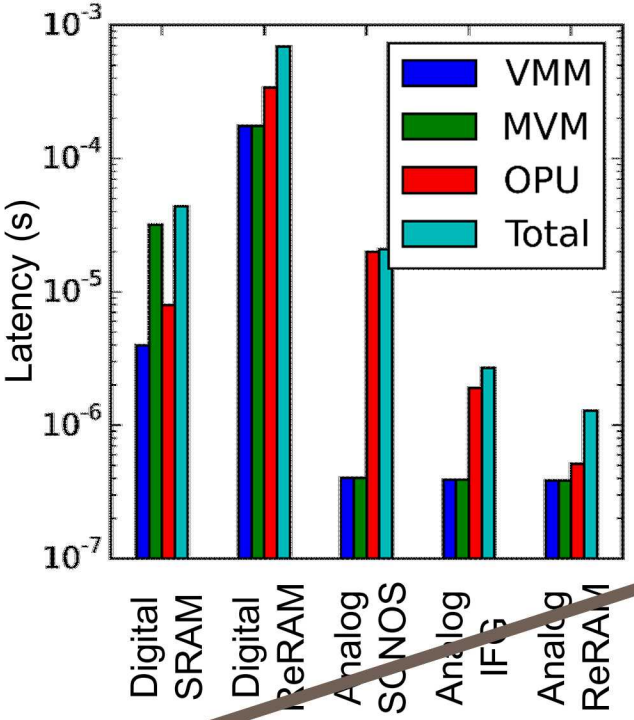
- SONOS has slow write (~1 ms) and high write voltage (11V)
- IFG and ReRAM write energy negligible compared to VMM
- IFG has extra delay over ReRAM for access device to turn off

Compare Architectural Advantages: Area

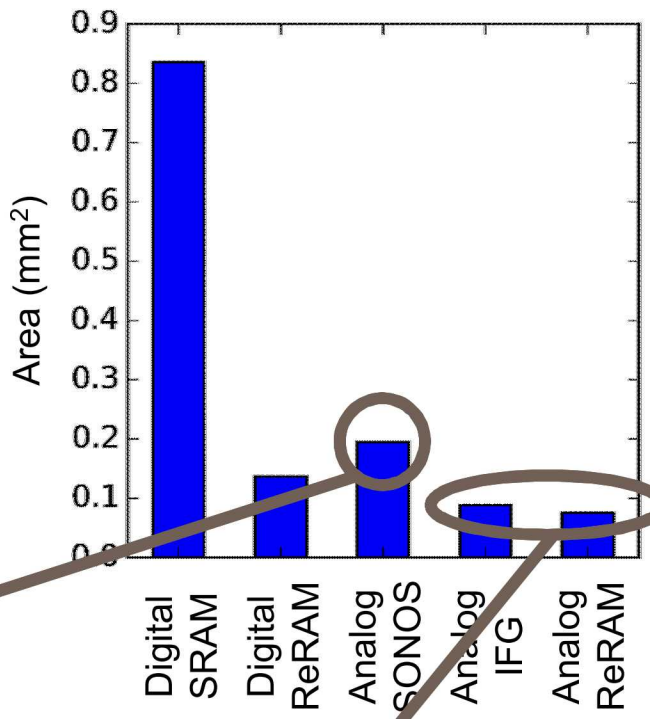
120-430X Energy Advantage



2-34X Latency Advantage



5-11X Area Advantage

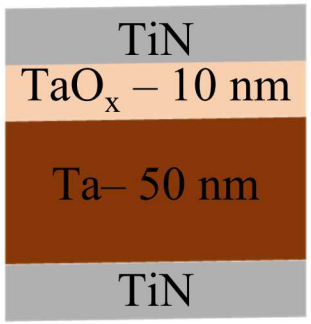


SONOS area cost reasonable, roughly doubles area

IFG and ReRAM go over transistors, area dominated by ADC and DAC

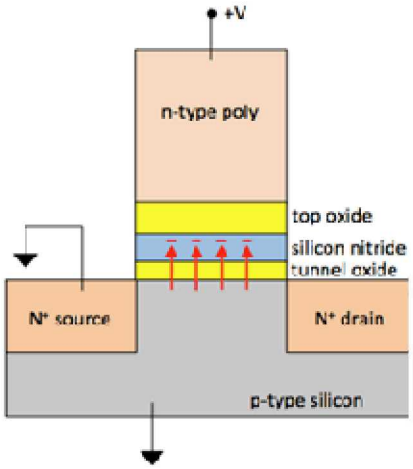
Analog Devices Summary for Training

ReRAM



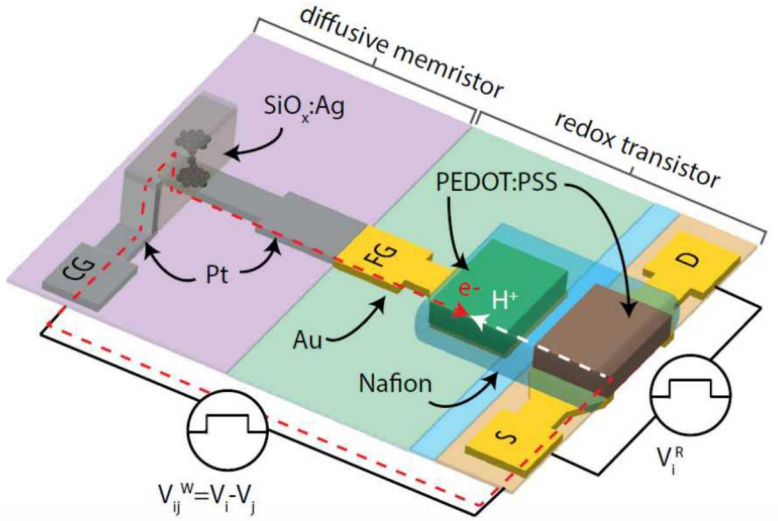
- Large Energy/Area/Latency advantage over digital
- Accuracy not good enough
- Back end of line compatible
- Under commercial development

SONOS
Silicon-Oxygen-Nitrogen-Oxygen-Silicon



- Moderate Energy/Area/Latency advantages over digital
- High Accuracy
- Commercially available
- Need to prove endurance and device to device variability

Ionic
Floating-Gate Memory

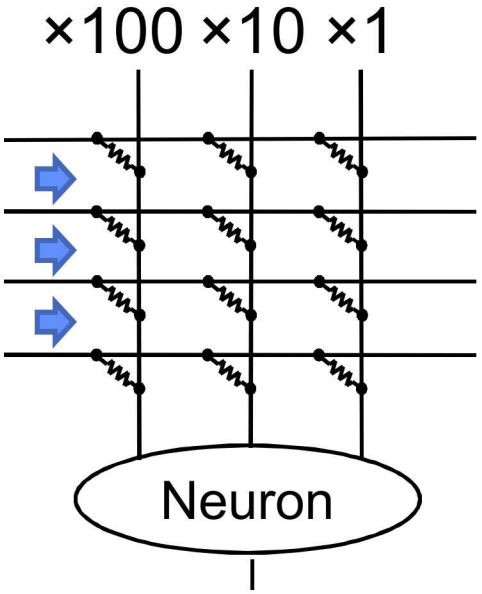


- Large Energy/Area/Latency advantages over digital
- High Accuracy
- Not clear how to integrate
- Has retention challenges

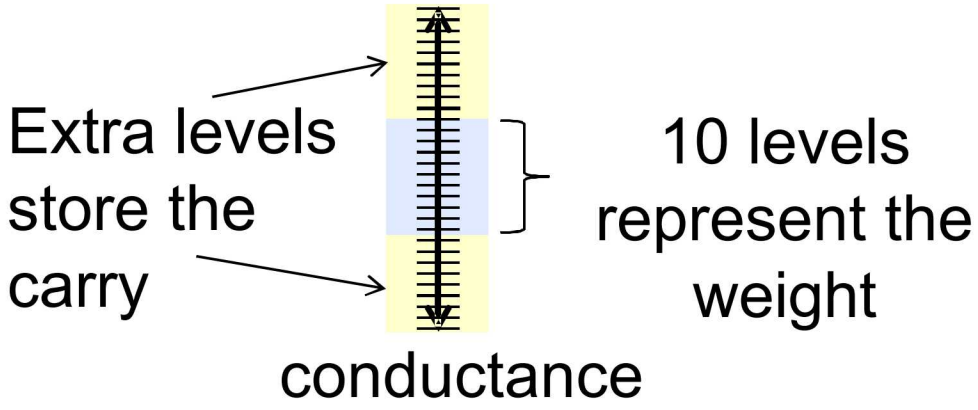
Multi-ReRAM Synapse: Periodic Carry

If we need more bits per synapse, use multiple memristors

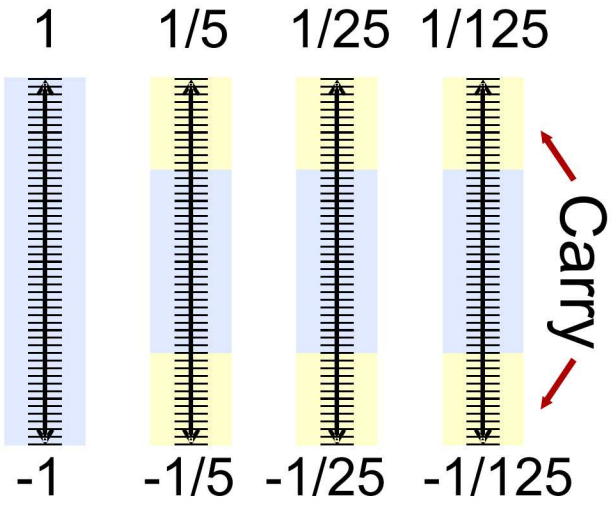
- Three 10 level ReRAMs could represent 1-1000!
- Adding to the weight requires reading every ReRAM to account for any carries and serially programming each ReRAM: VERY EXPENSIVE

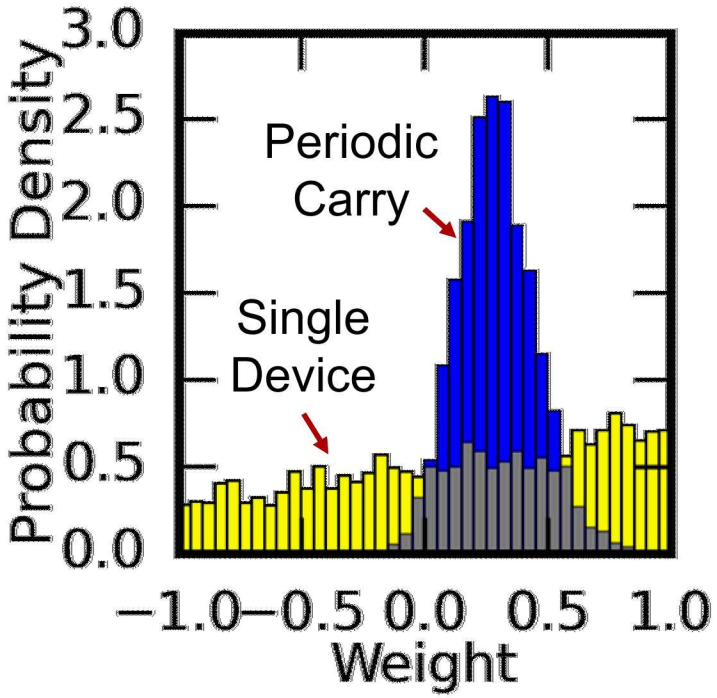


- Use >10 levels to represent a base 10 system
- Ignore carry and program the crossbar in parallel.
- Periodically (once every few hundred cycles) read the ReReRAM and perform the carry



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

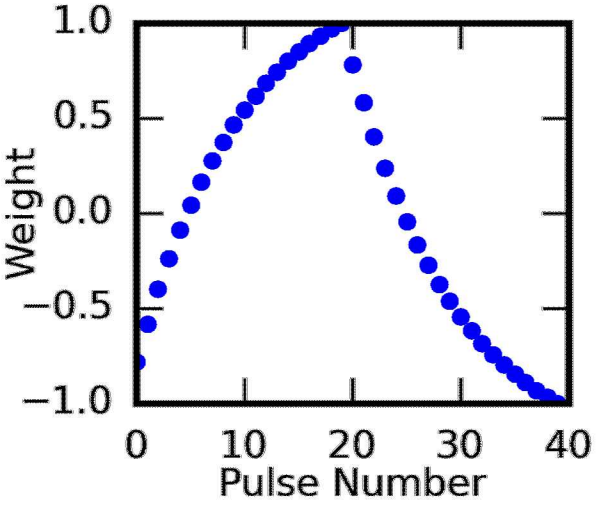
Learn from a 0.5% Signal

Noise Sigma = 1.4% for single device

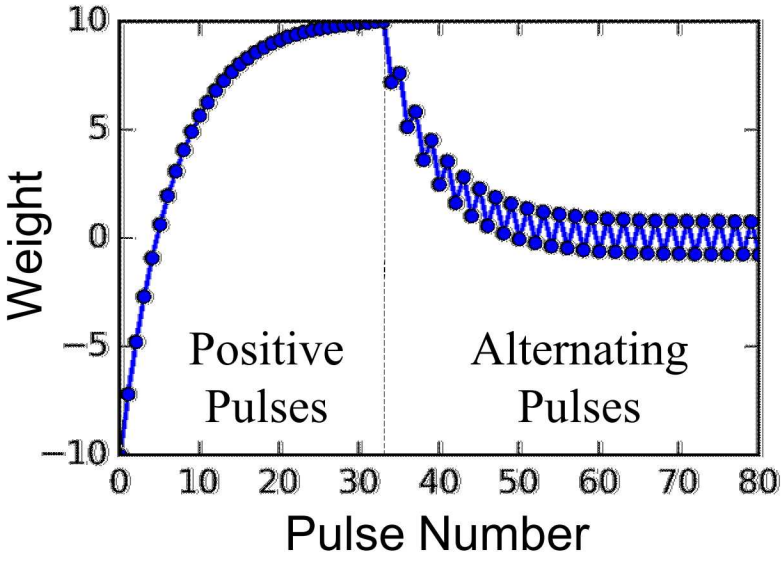
- (from $\sigma_{noise} / G_{range} = 0.1\sqrt{\Delta G / G_{range}}$)
- Write noise applied during updates and carries

Periodic Carry Mitigates Write Nonlinearity

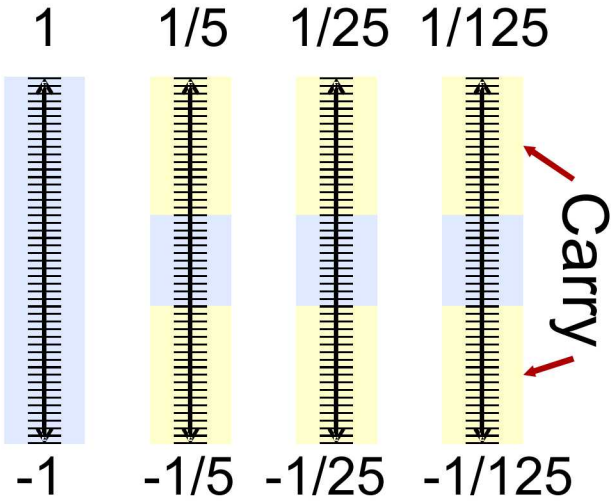
Write Nonlinearity

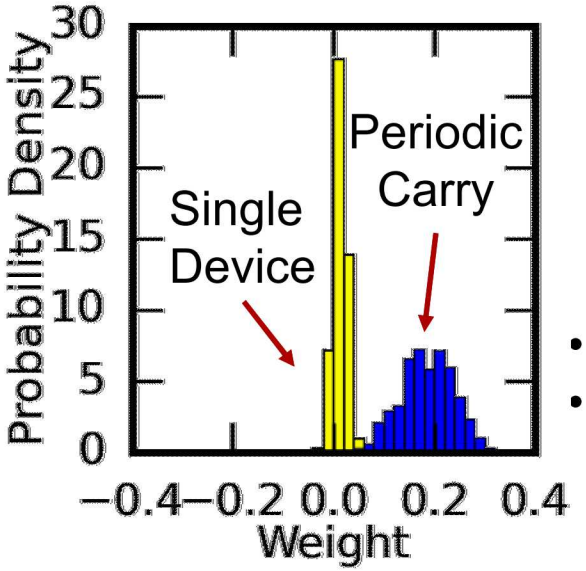


Alternating Pulses Cause Weight Decay



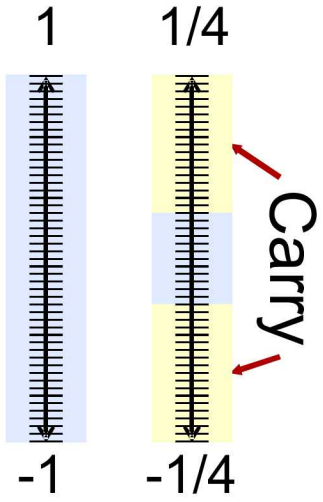
Use center linear range of weights



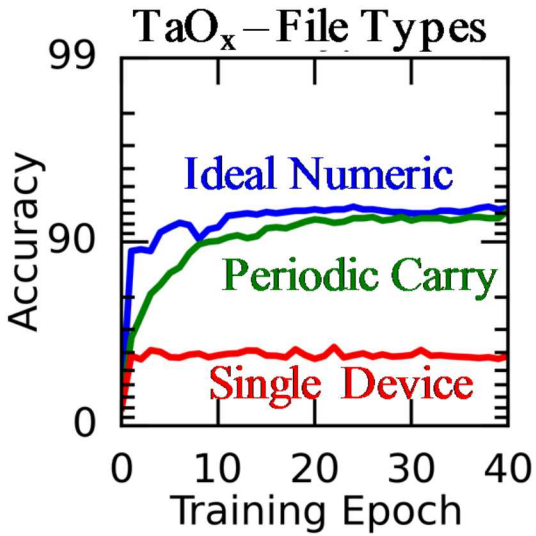


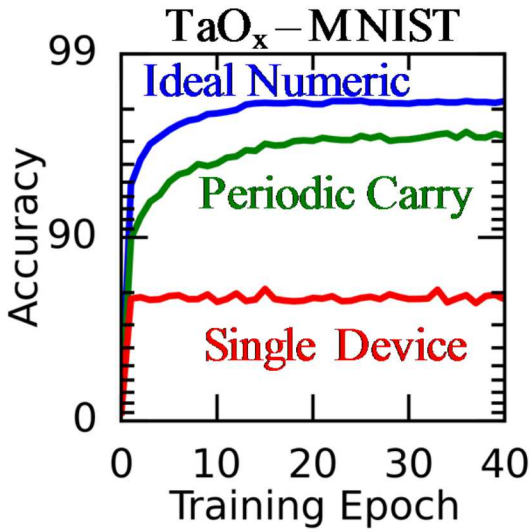
- Train with 1% signal
- Ideal result is 0.6

TaO_x Results



Carry once every 1000 updates

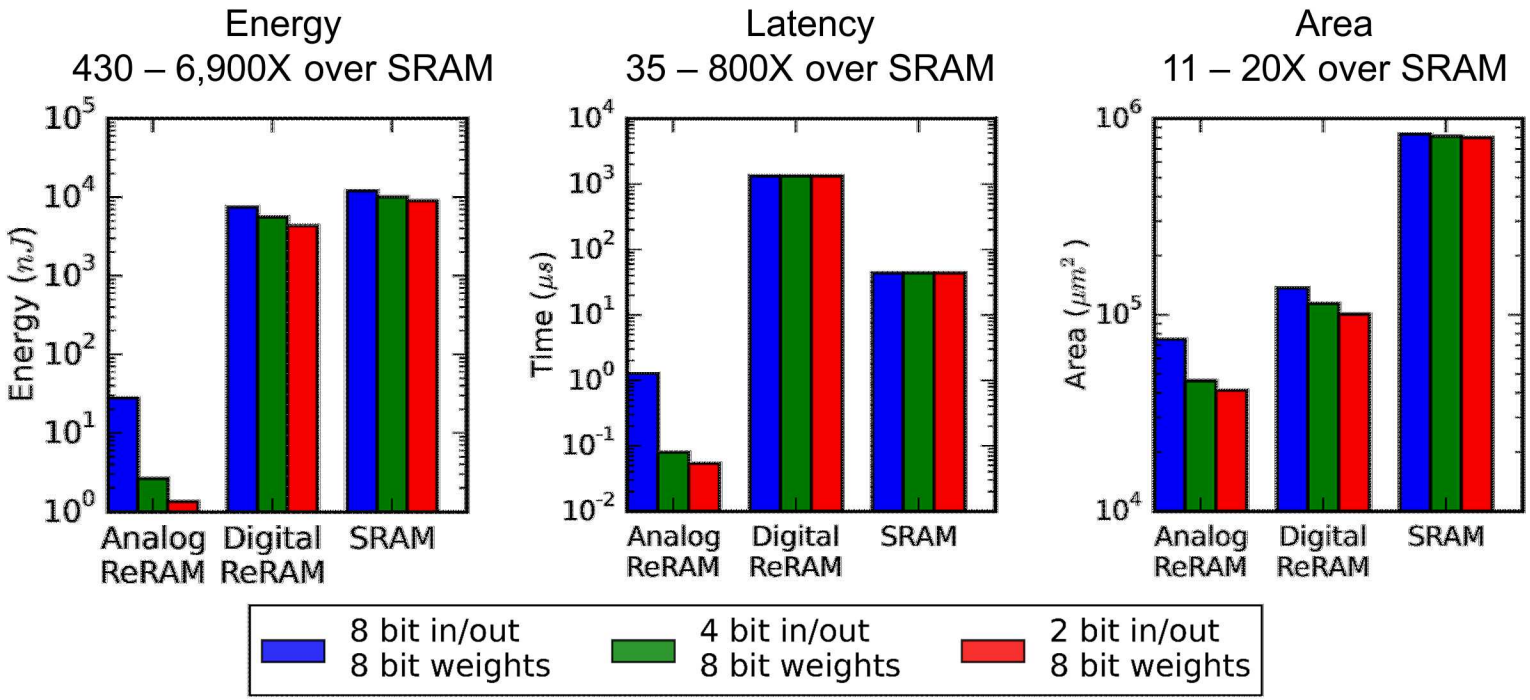




A/D and D/A is modeled, Serial operations modeled

- 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)

Summary



- Fundamental O(N) energy scaling advantage
- Use CrossSim to co-design materials to algorithms
  - Use periodic carry to overcome noise devices
- Need high resistance 10-100 MΩ Devices
- Need low write nonlinearities

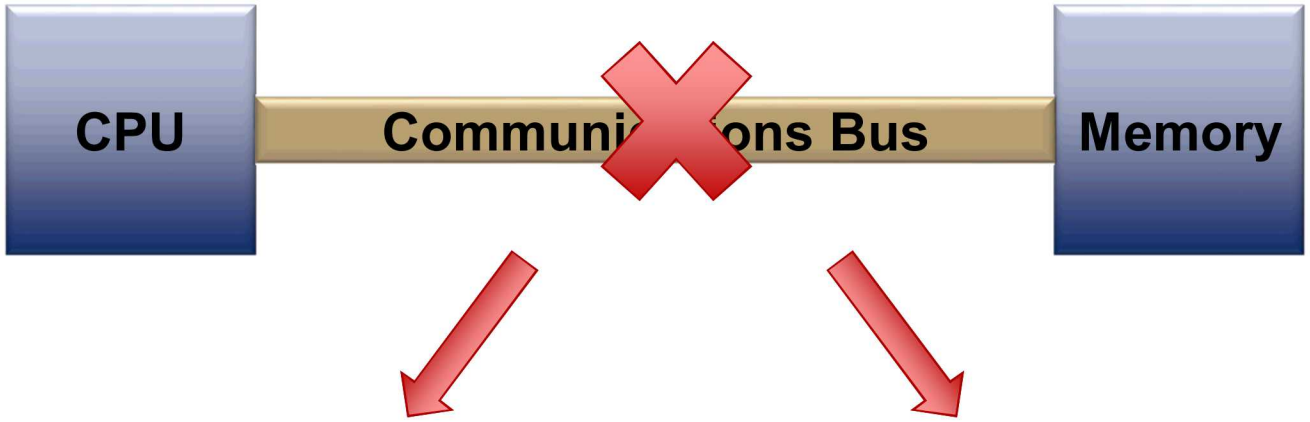
#ROSS SIM

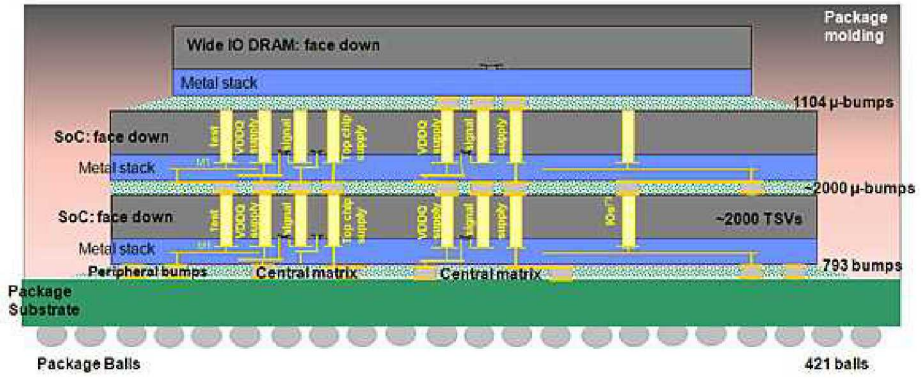
https://cross-sim.sandia.gov

Extra Slides

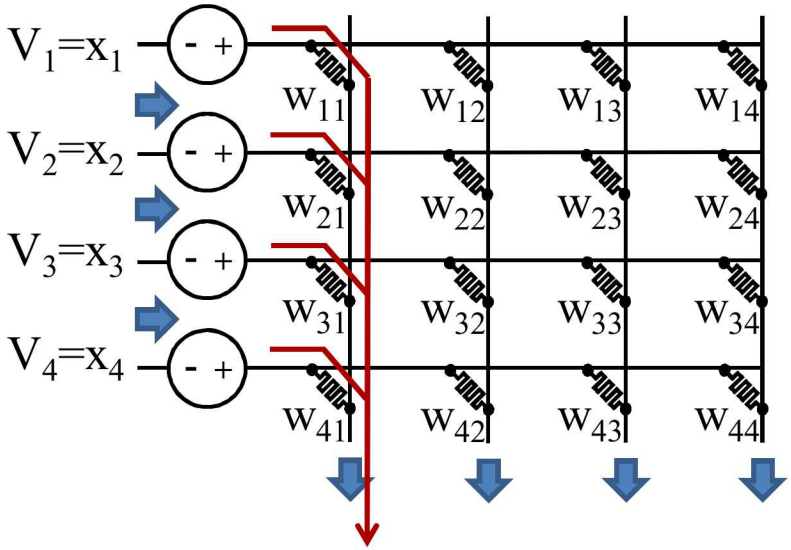
Sandia National Laboratories

Overcoming the Power Limit





Richard Goering, "Three Die Stack -- A Big Step "Up" for 3D-ICs with TSVs" Cadence blog



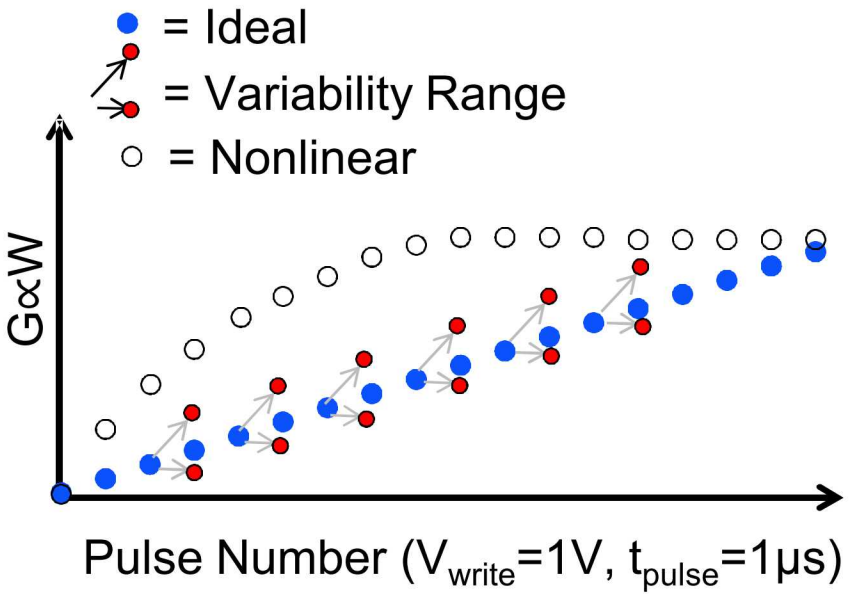
Integrate Processing and Memory

Experimental Device Non-idealities

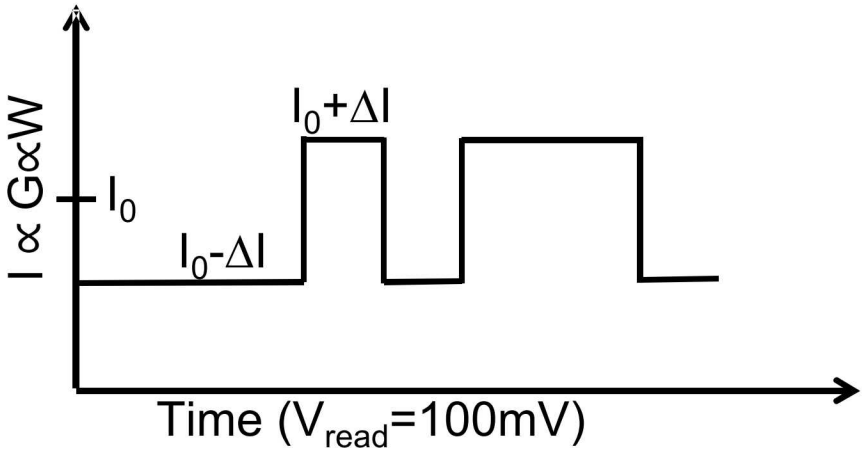
Device: Write Variability, Write Nonlinearity, Asymmetry, Read Noise

Circuit: A/D, D/A noise, parasitics

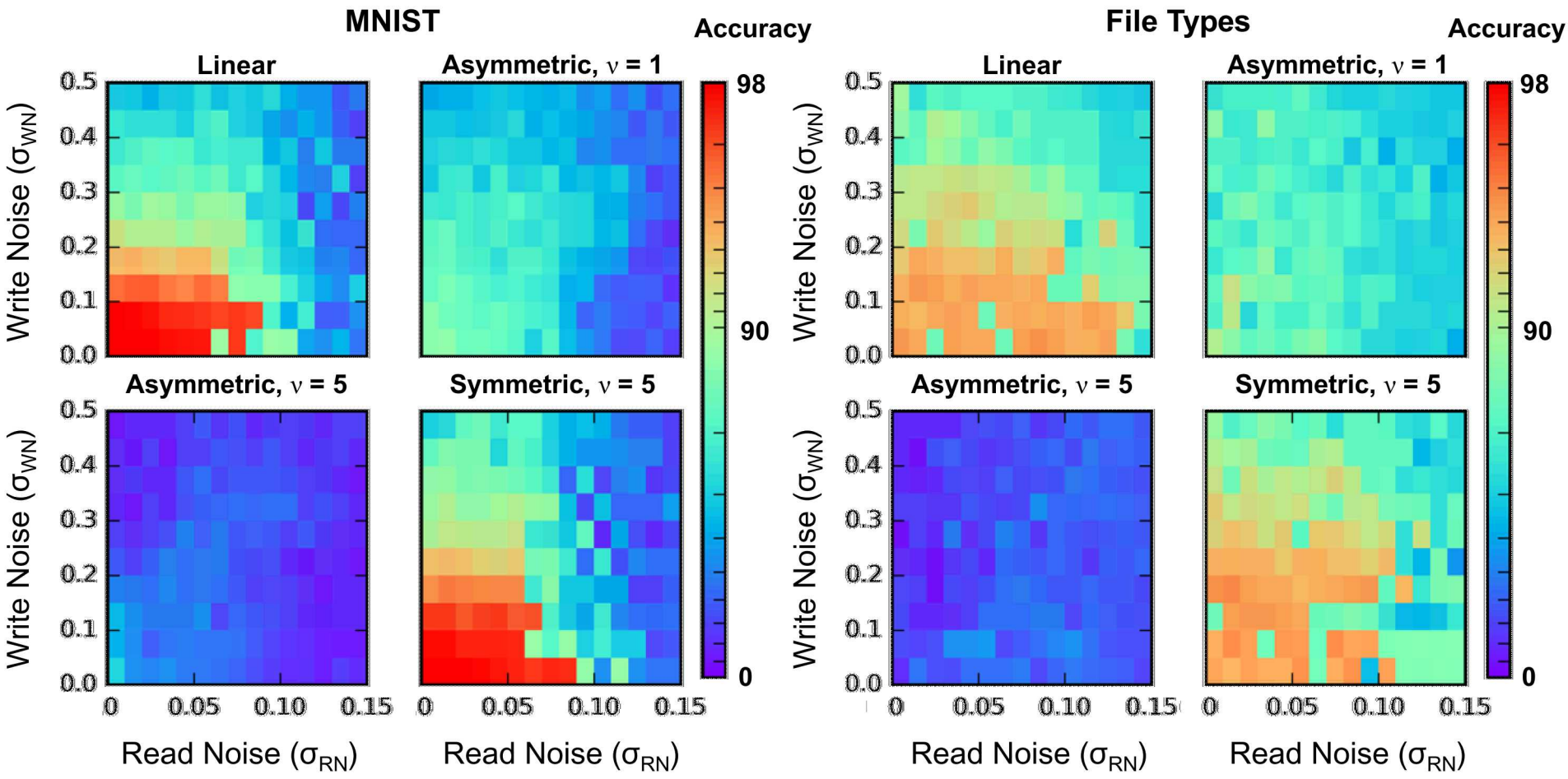
Variability and Nonlinearity



Read Noise

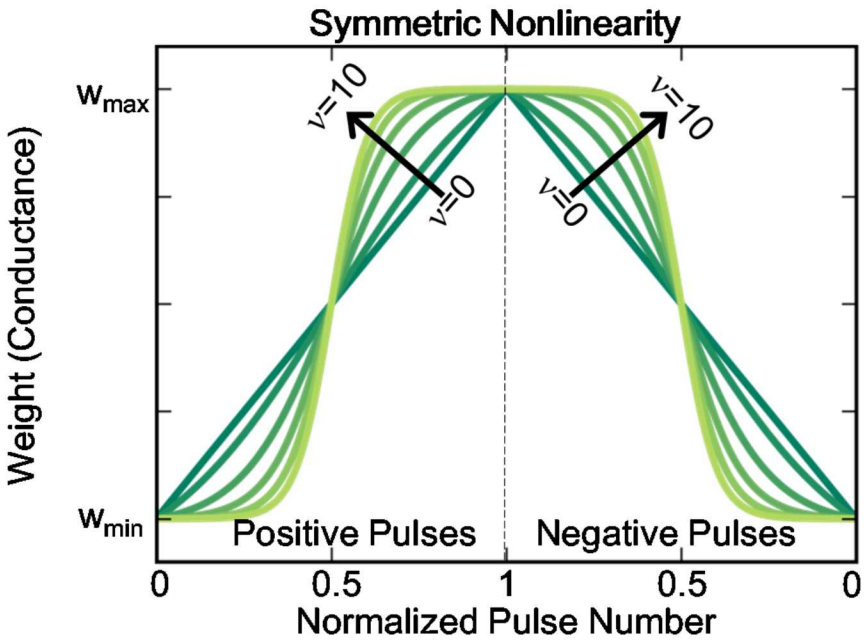


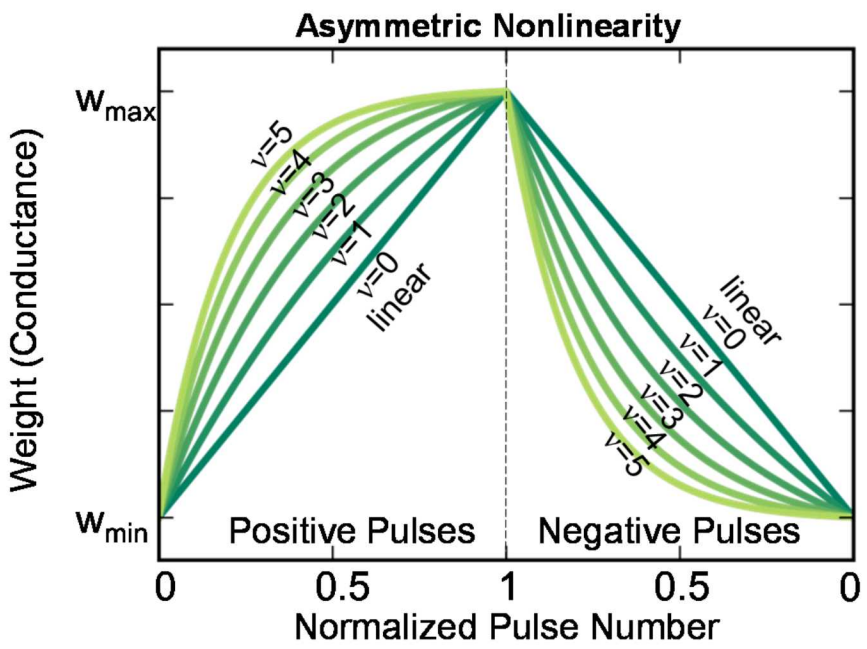
Combined Effects of Nonidealities



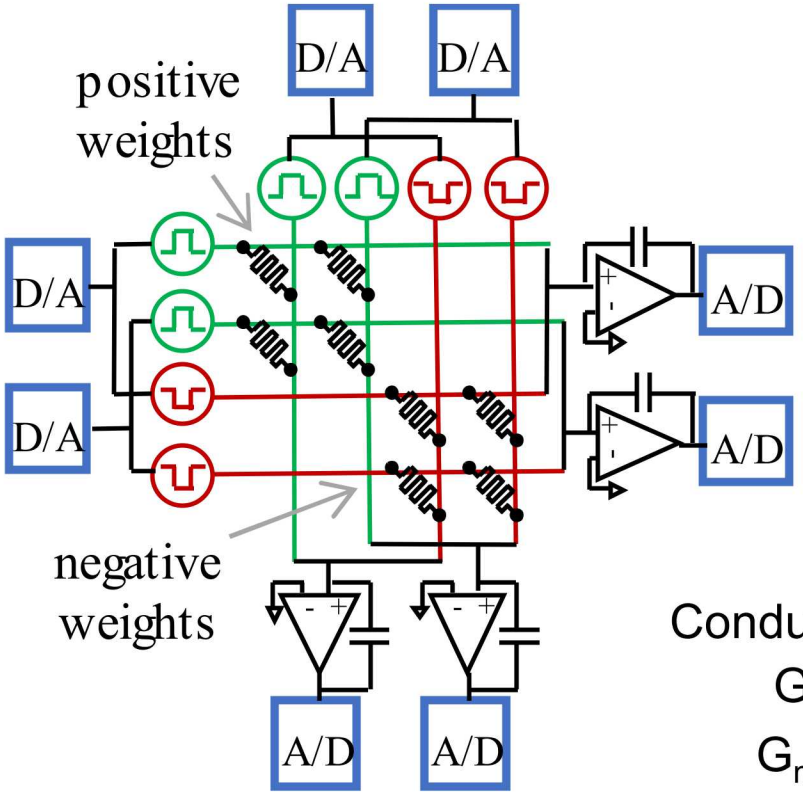
What are the Neural ReRAM Device Requirements?

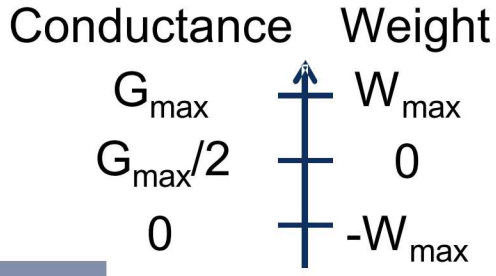
	Small Images	Large Images	File Types
Read Noise σ (% Range)	3%	5%	9%
Write Noise σ (% Range)	0.3%	0.4%	0.4%
Asymmetric Nonlinearity (v)	0.1	0.1	0.1
Symmetric Nonlinearity (v)	>20	5	5
Maximum Current	160 nA	13 nA	40 nA



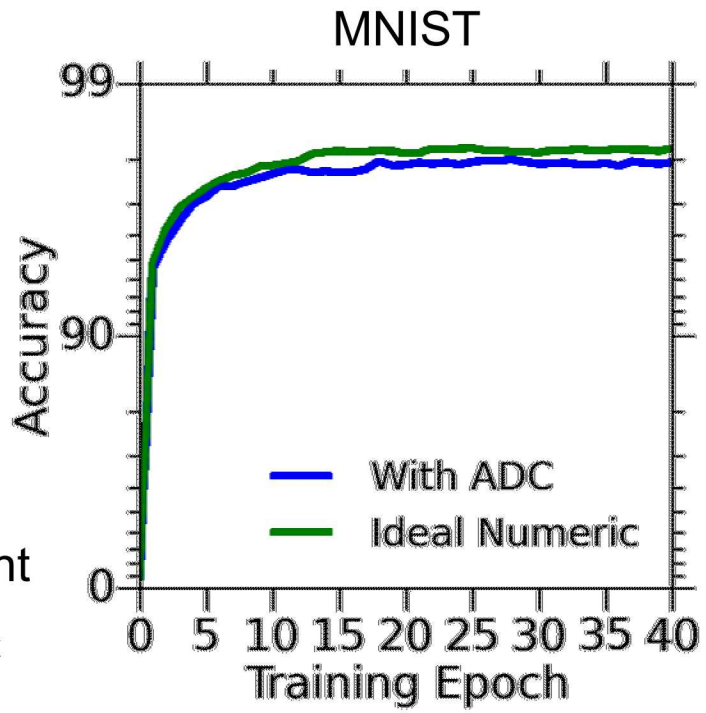


Full System Simulation





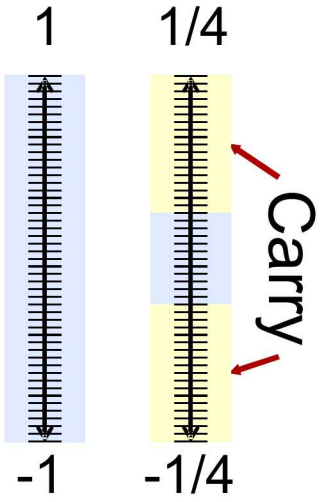
A/D & D/A Have Minimal Impact

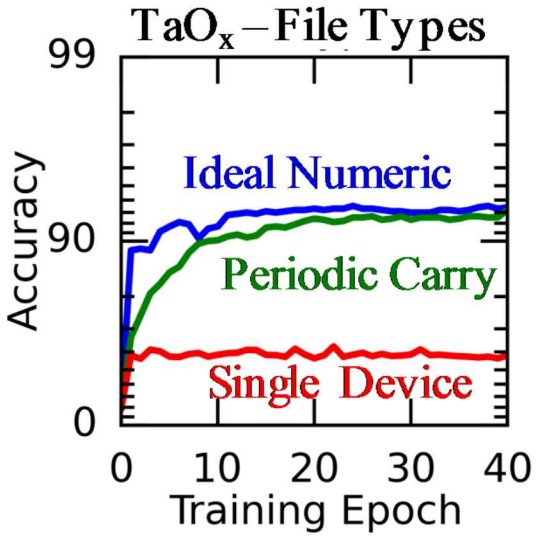


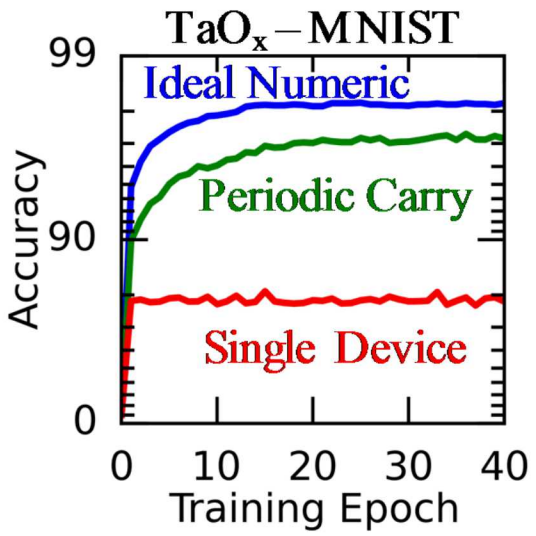
	Range	Bits
Row Input	-1 to 1	8
Col Output	-6 to 6	8
Col Input	-1 to 1	8
Row Output	-4 to 4	8
Row Update	-0.01 to 0.01	7
Col Update	-1 to 1	5

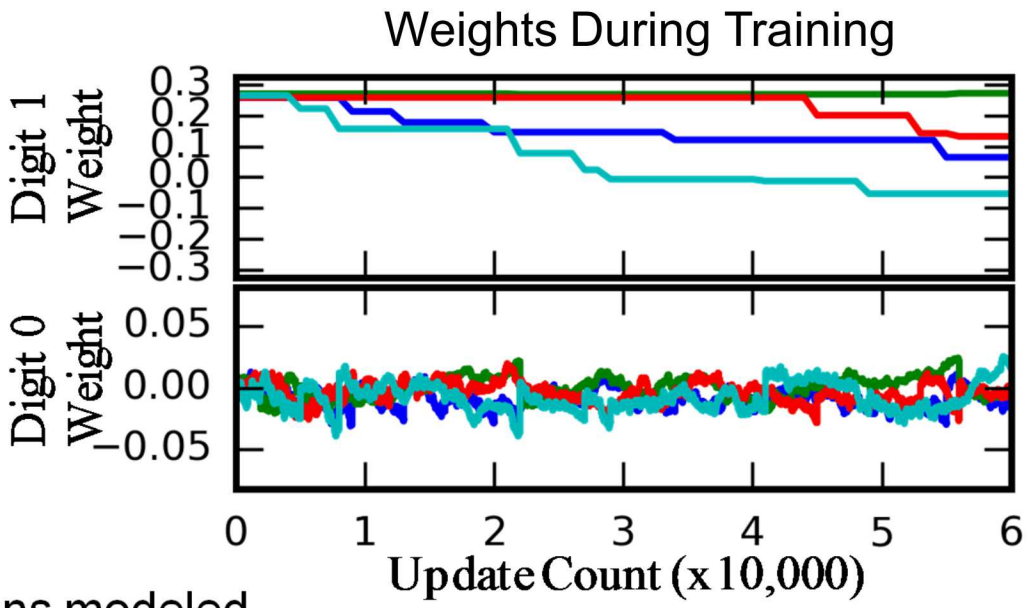
Data set	#Training/Test Examples	Network Size
File Types	4,501 / 900	256×512×9
MNIST	60,000 /10,000	784×300×10

TaO_x Results







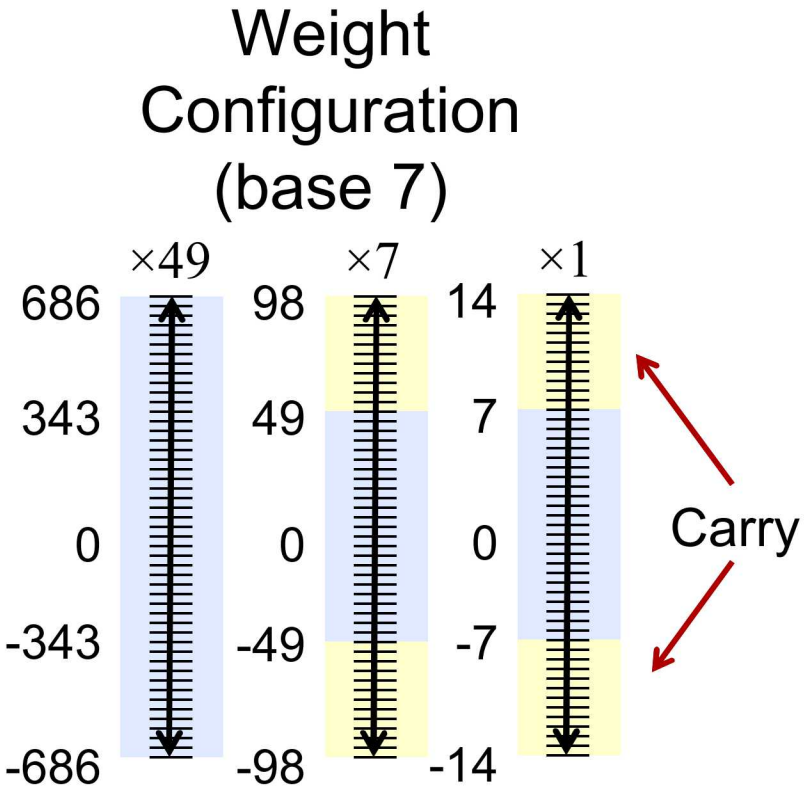


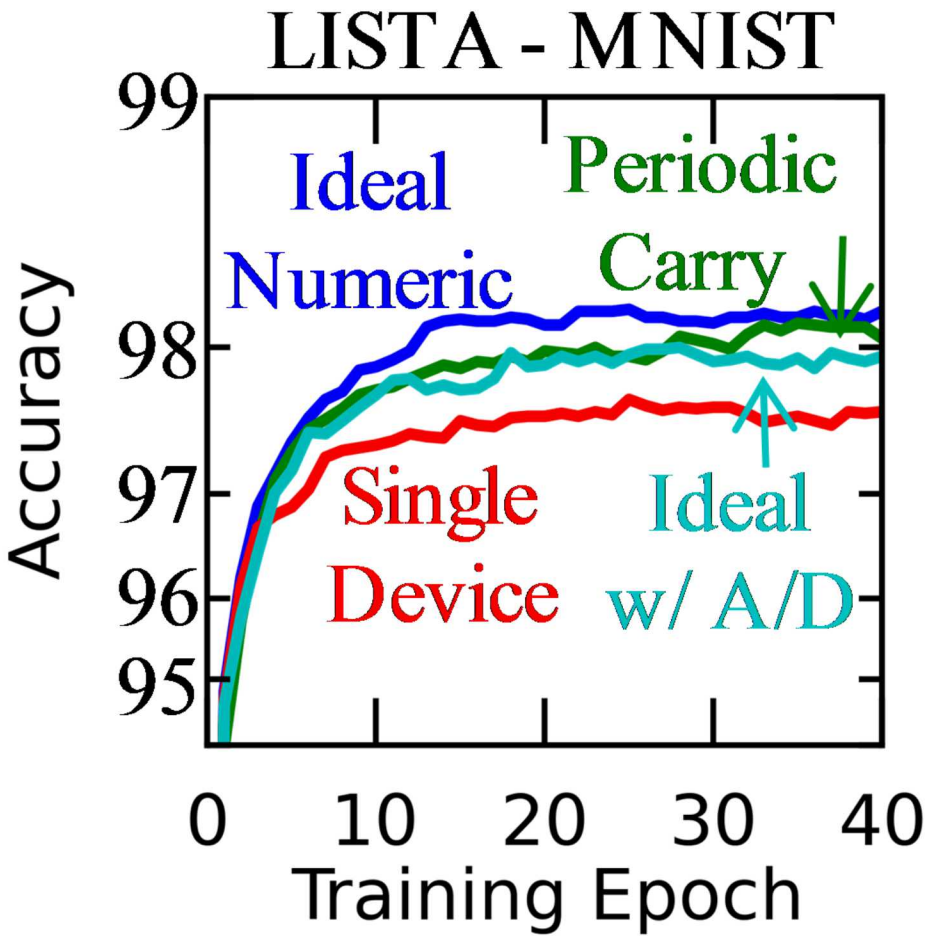
Carry once every 1000 updates for the LSB, and every 2 updates on others

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)

LISTA Results





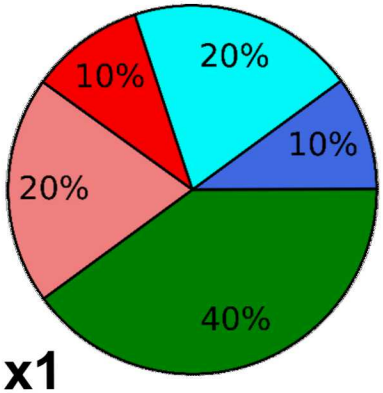
- Carry once every 1000 updates
- Use a single device per weight and subtract a reference current

Neural Core Latency Analysis

Sandia National Laboratories

Analog ReRAM

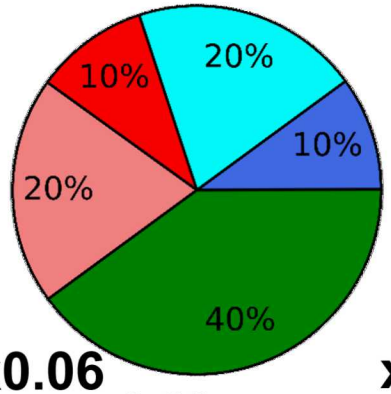
8 bit in/out



x1

1.28 µs

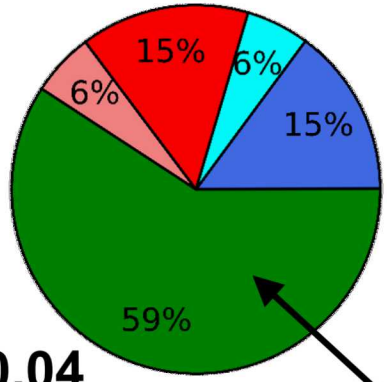
4 bit in/out



x0.06

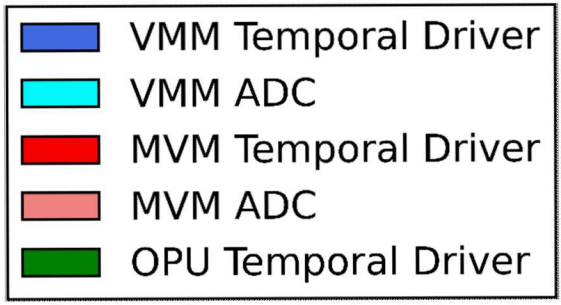
0.08 µs

2 bit in/out



x0.04

0.054 µs

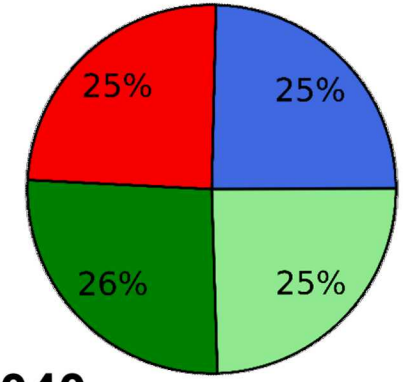


OPU = Outer Product Update

Min write time of 8 ns vs 1 ns incremental write

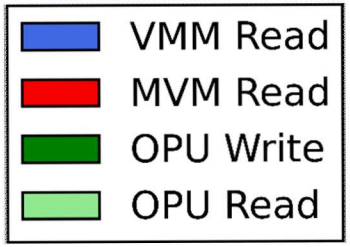
Digital ReRAM

All bit precisions



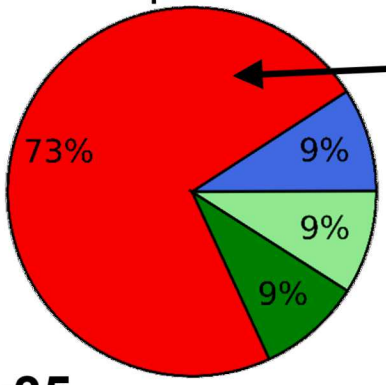
x1040

1335 µs



SRAM

All bit precisions

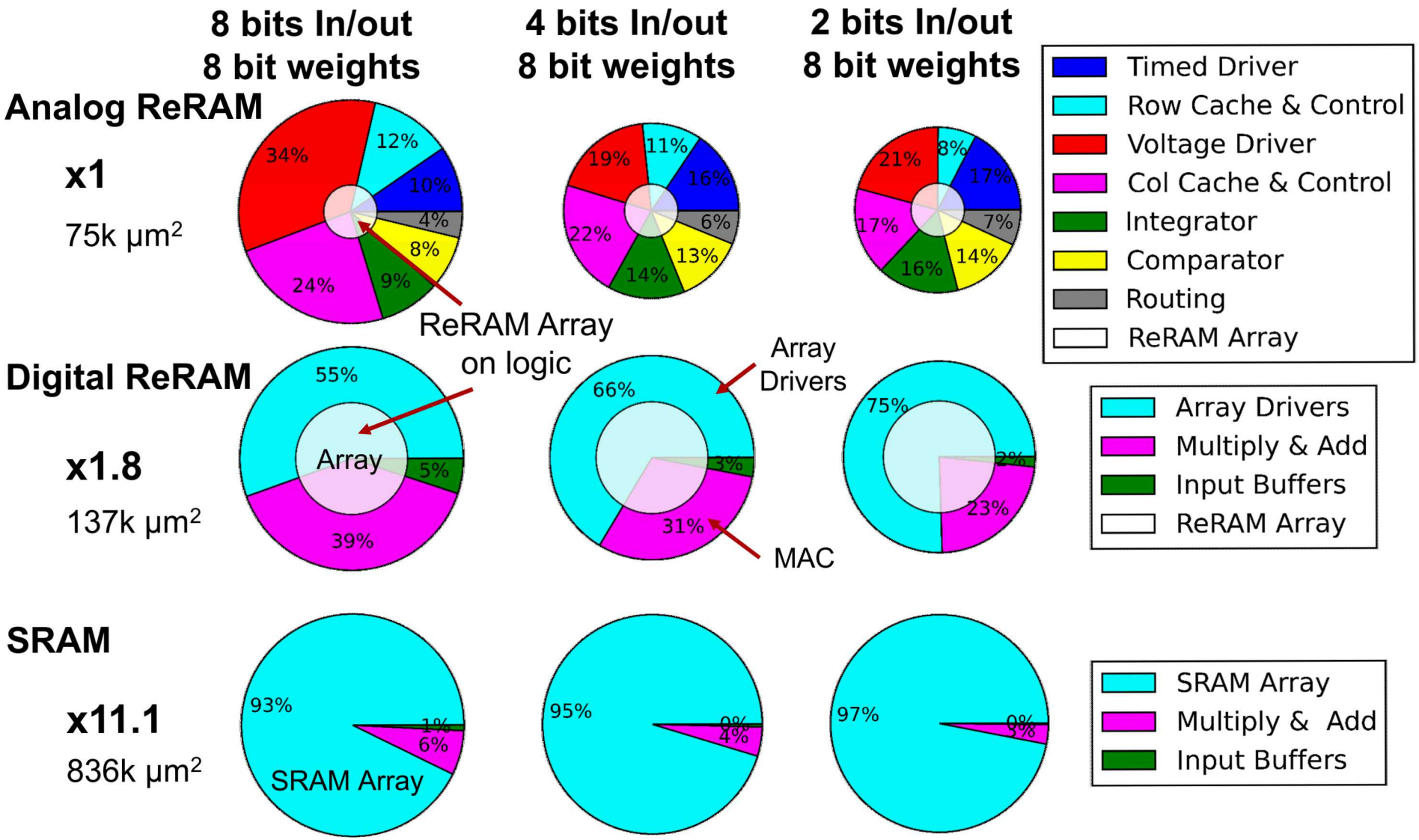


x35

44 µs

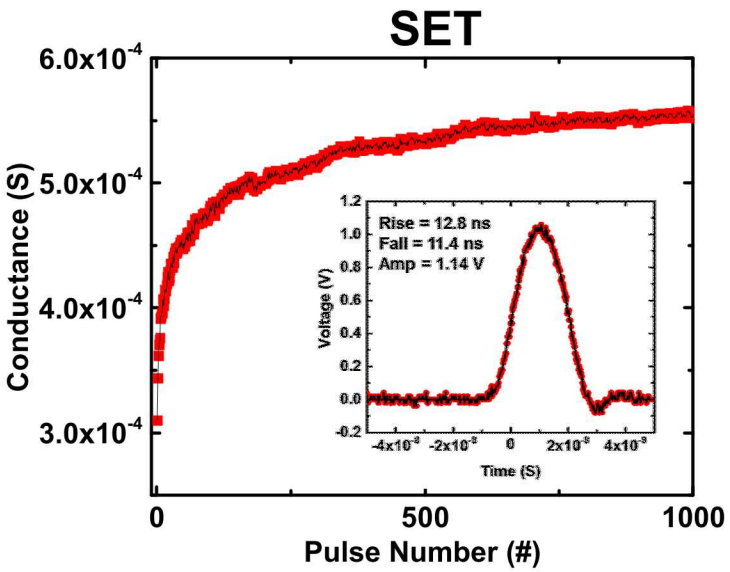
SRAM transpose read expensive

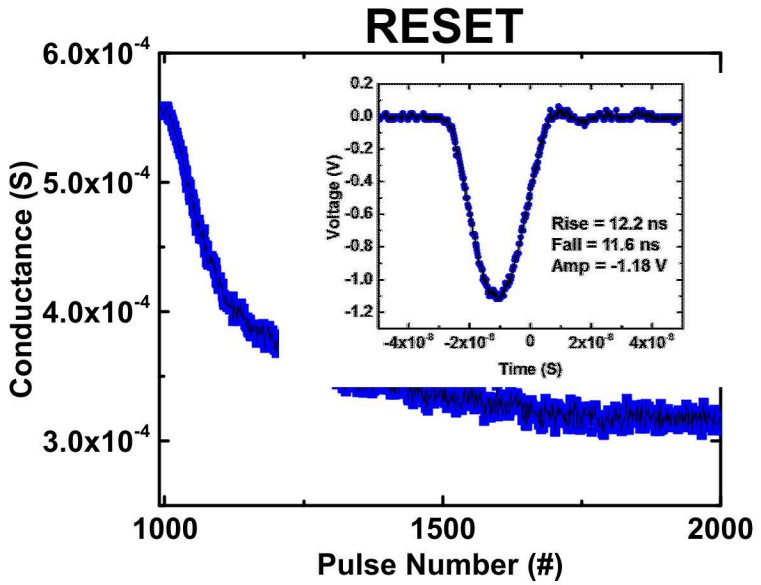
Neural Core Area Analysis

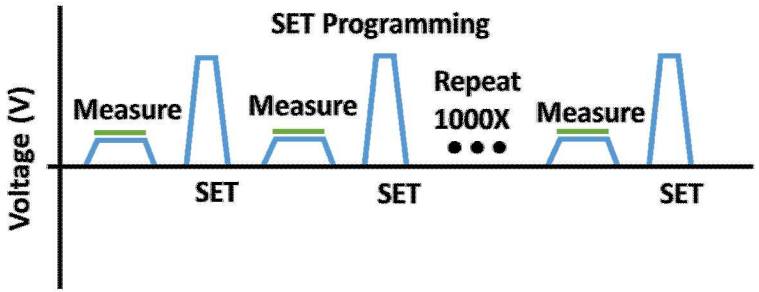


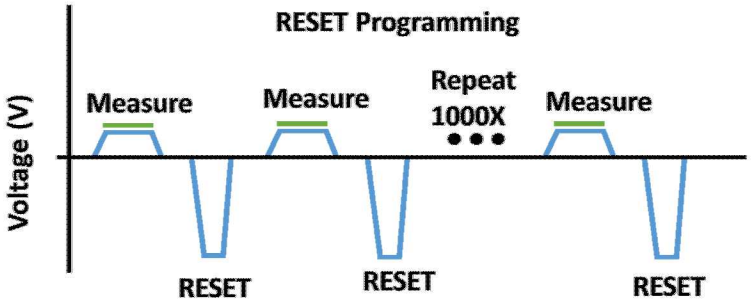
For the ReRAM, high voltage transistors require 8X area, improving this could give ~2X area savings

ReRAM Analog Characterization

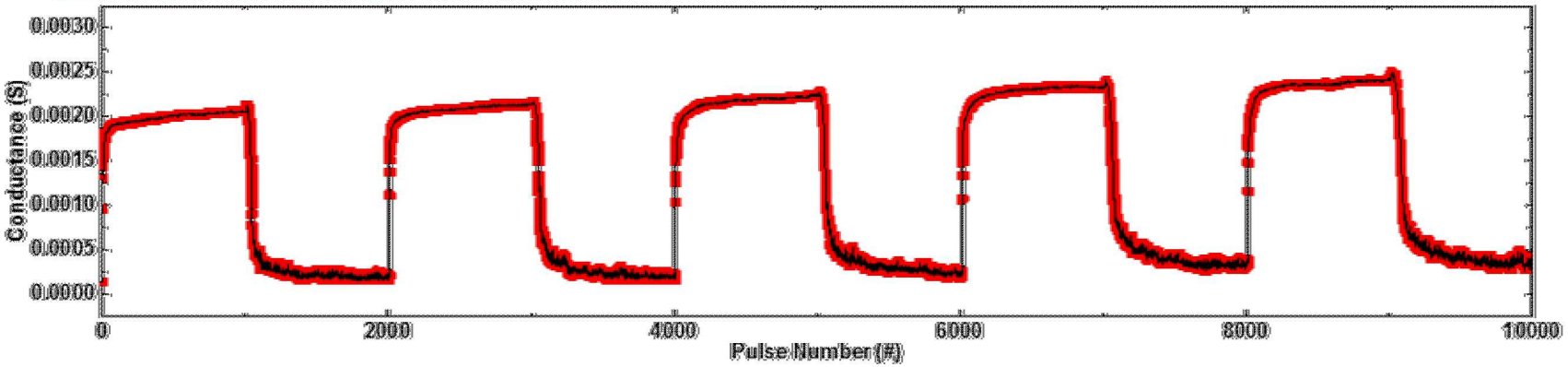


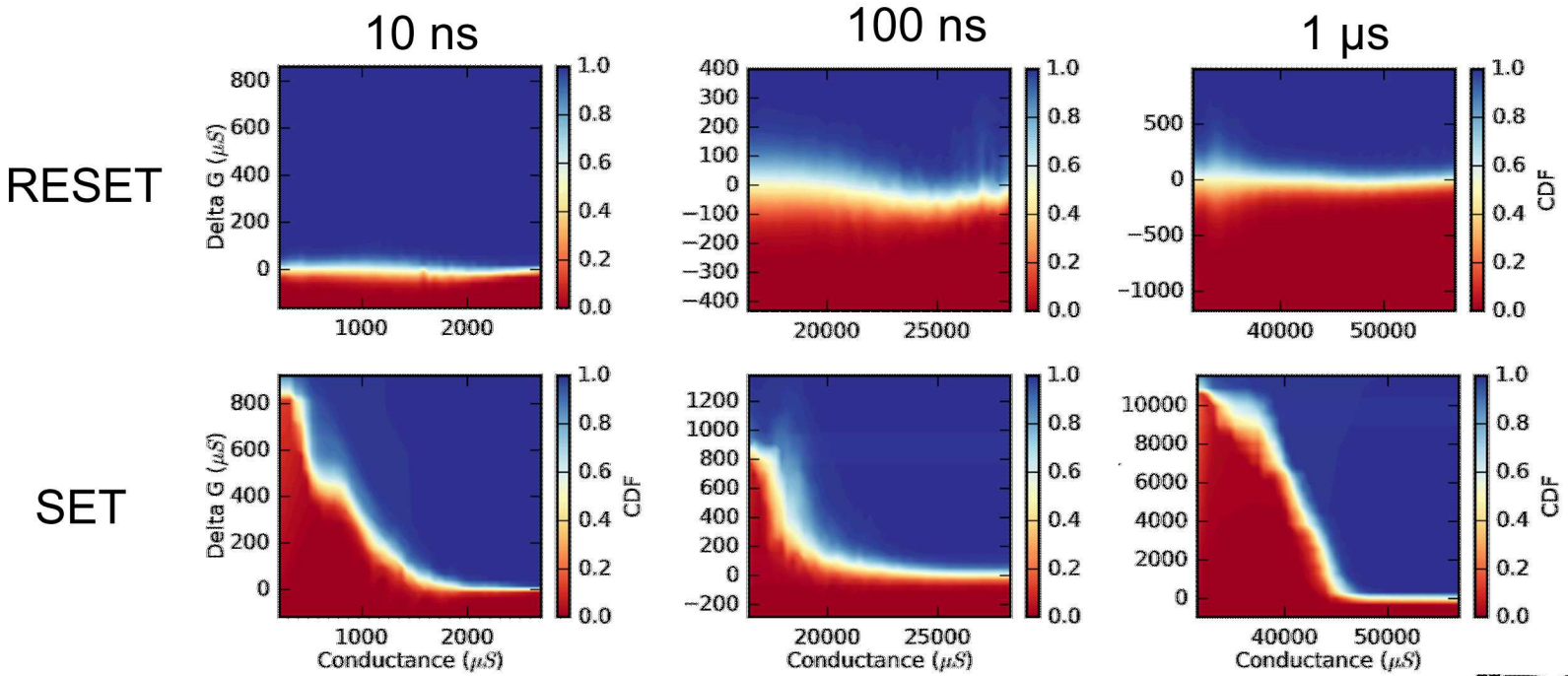




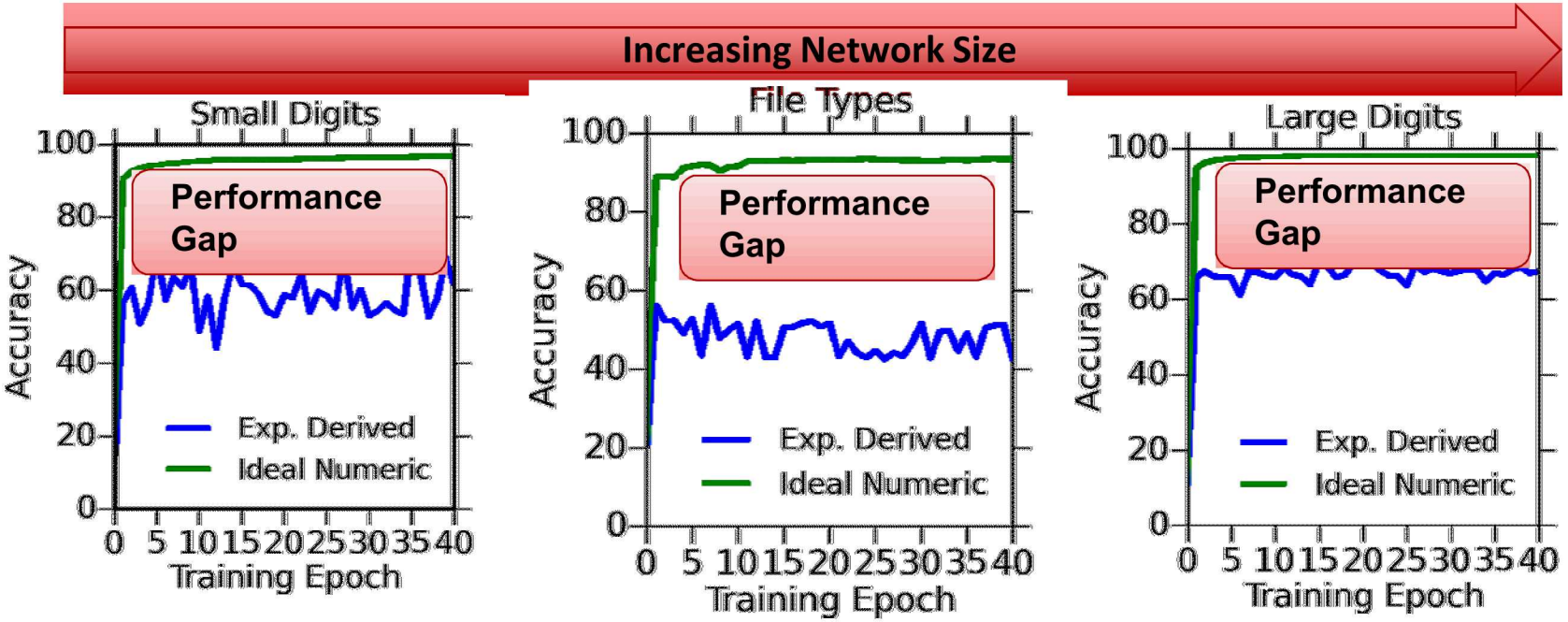


Modeling Pulsed Cycling Characteristics



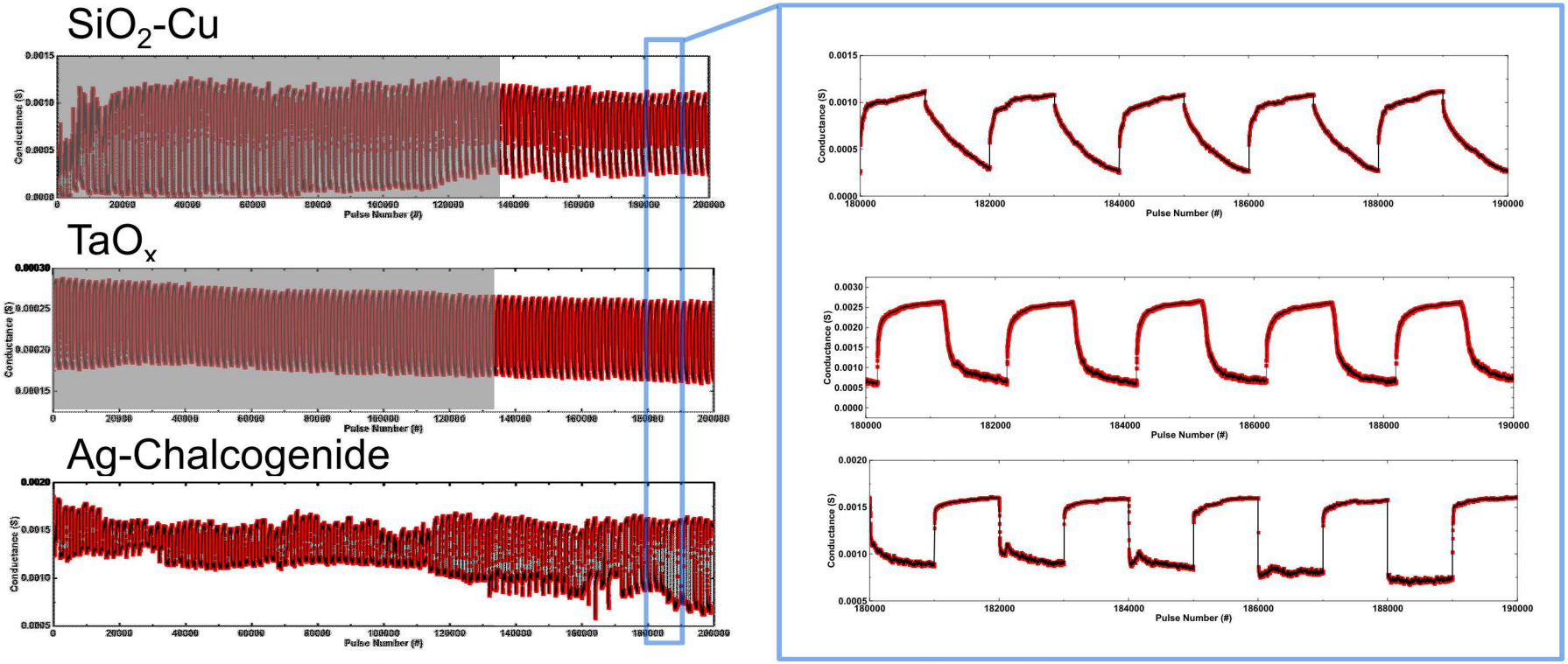


TaOx ReRAM in Backprop Training



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

Analog Device Comparison

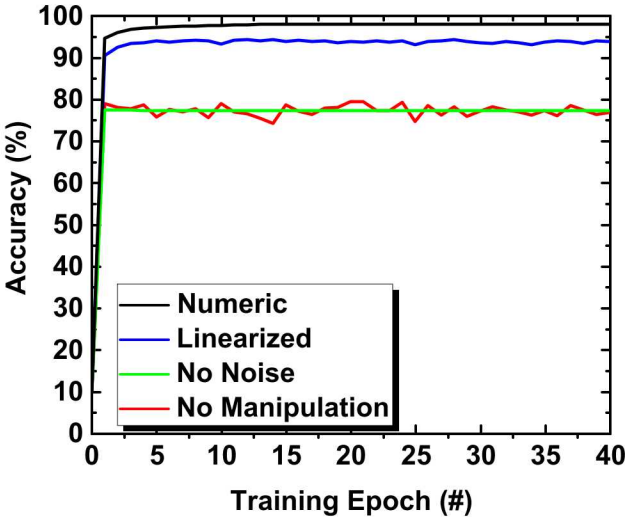


- **For comparison all devices at 100 ns due to impedance limitation**
- **Device operation voltages found by increasing amplitude by 0.1 V until switching occurred – must survive 200,000 nudges so lowest possible voltage used**
- **Chalcogenide SET = +0.8 V RESET = -0.8V**
- **SiO_2-Cu SET = +1.4 V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

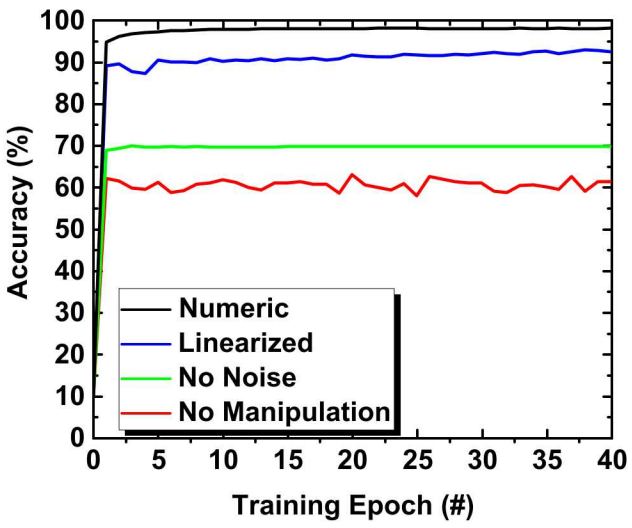
Comparison of Filamentary ReRAM

- 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

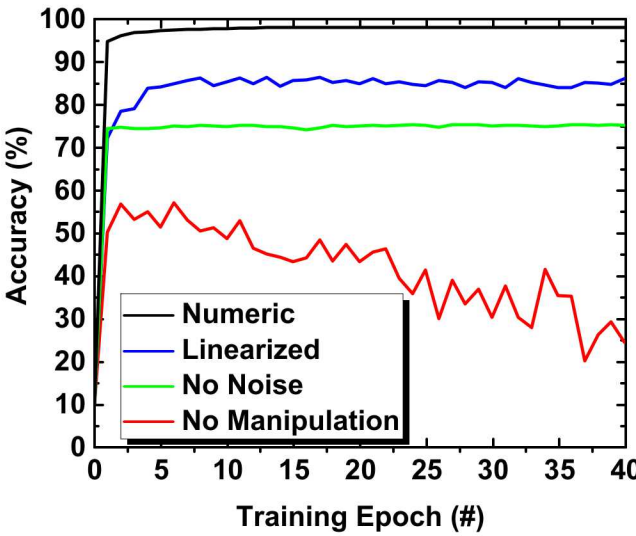
Sandia Baseline TaO_x-Ta



SiO_2-Cu



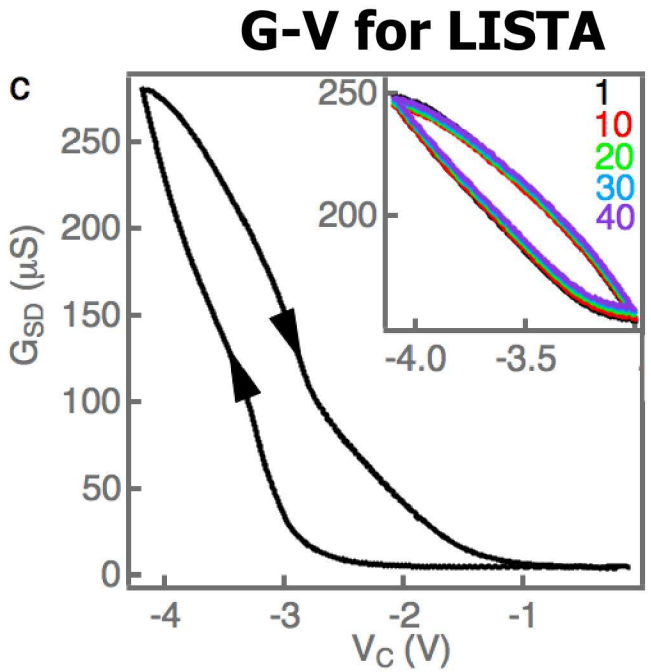
Ag-Chalcogenide

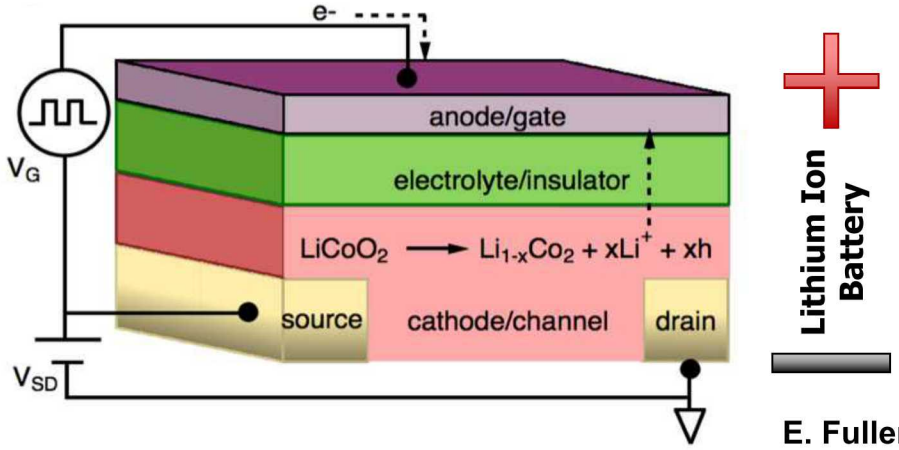


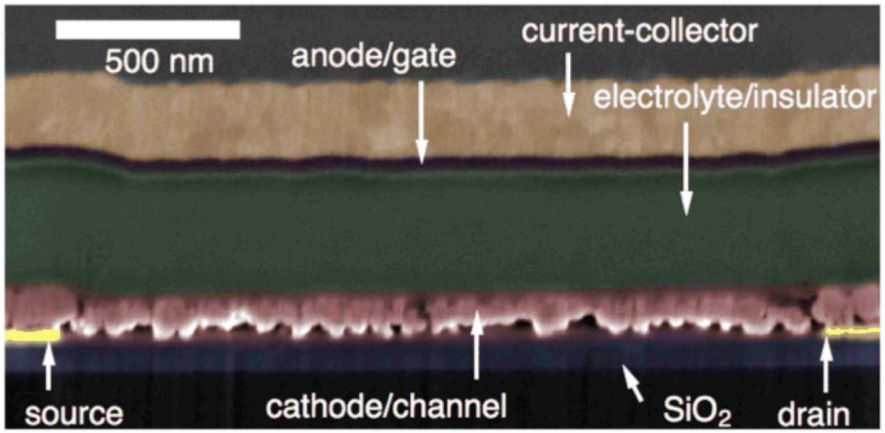
R. Jacobs-Gedrim et al, *IEEE-ICRC*, 2017

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

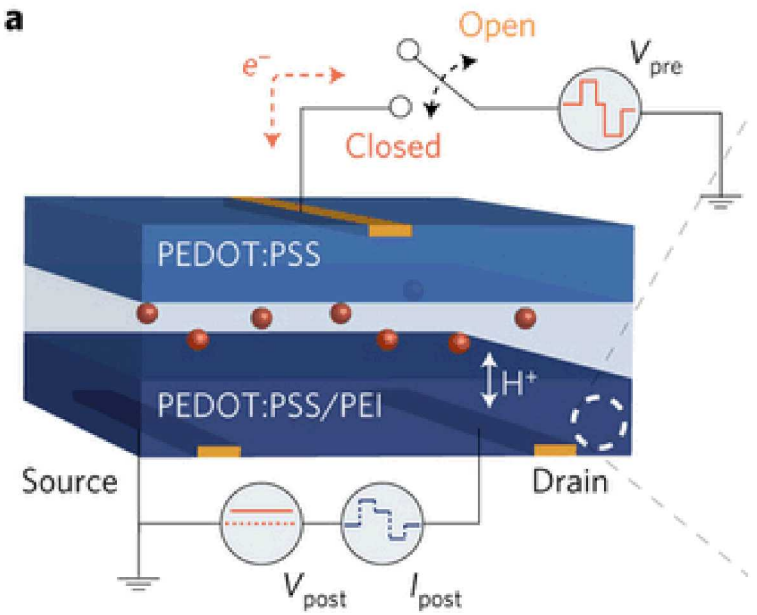


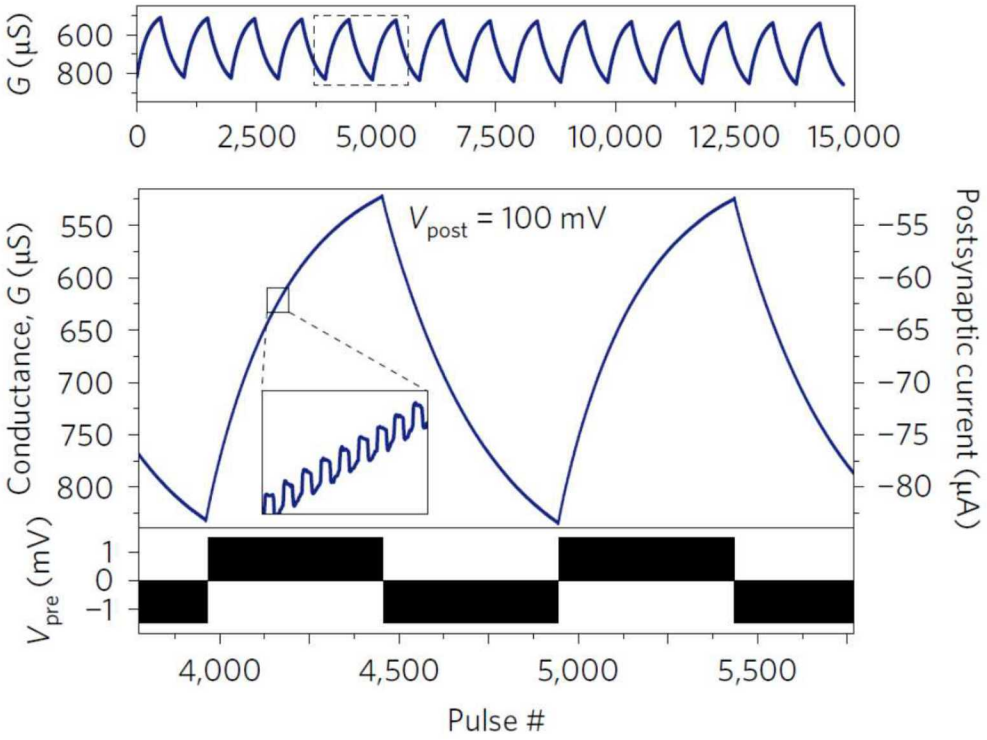




E. Fuller et al, *Adv Mater*, 2017

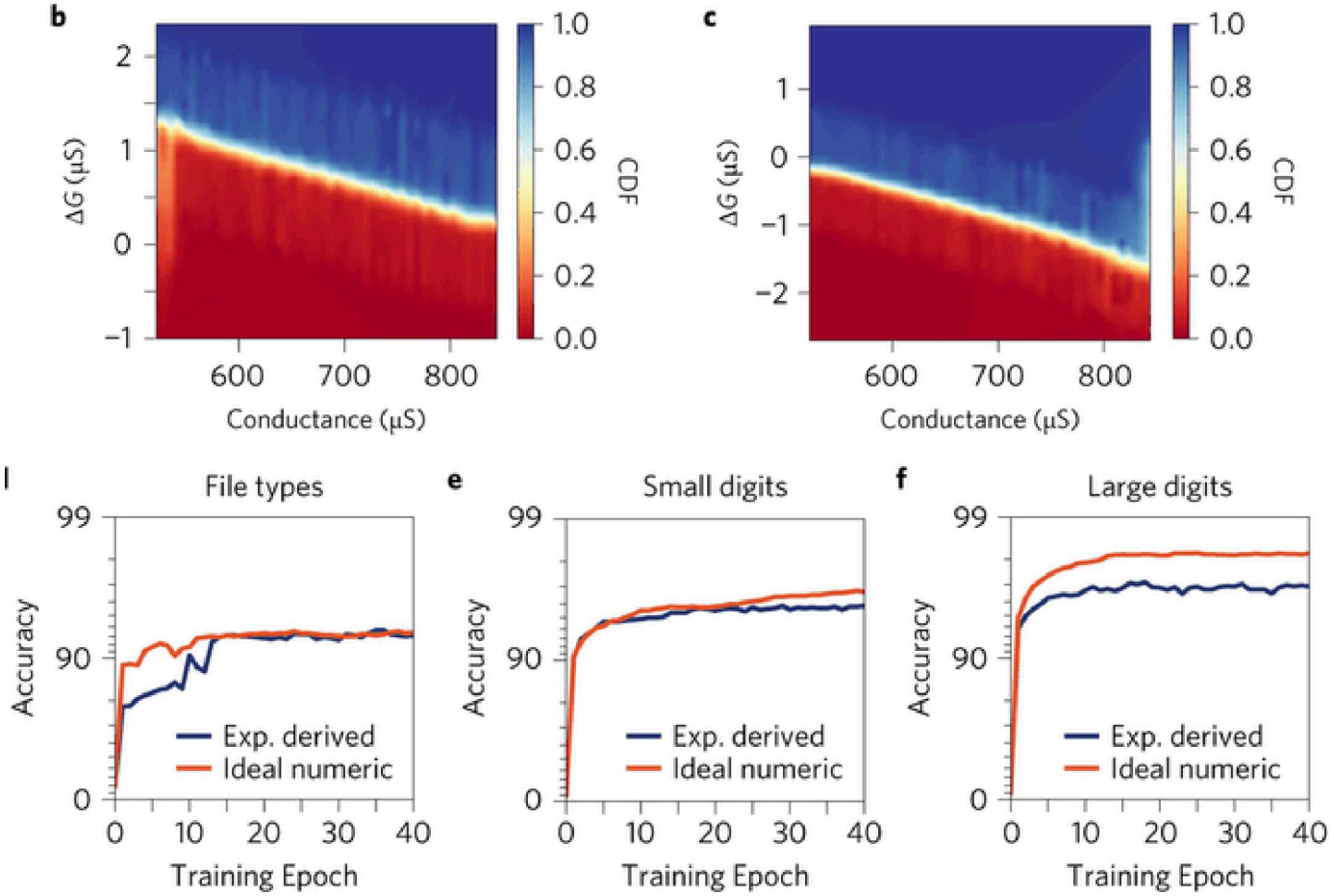
Electrochemical Neuromorphic Organic Device (eNode)





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

Electrochemical Neuromorphic Organic Device (eNode)



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