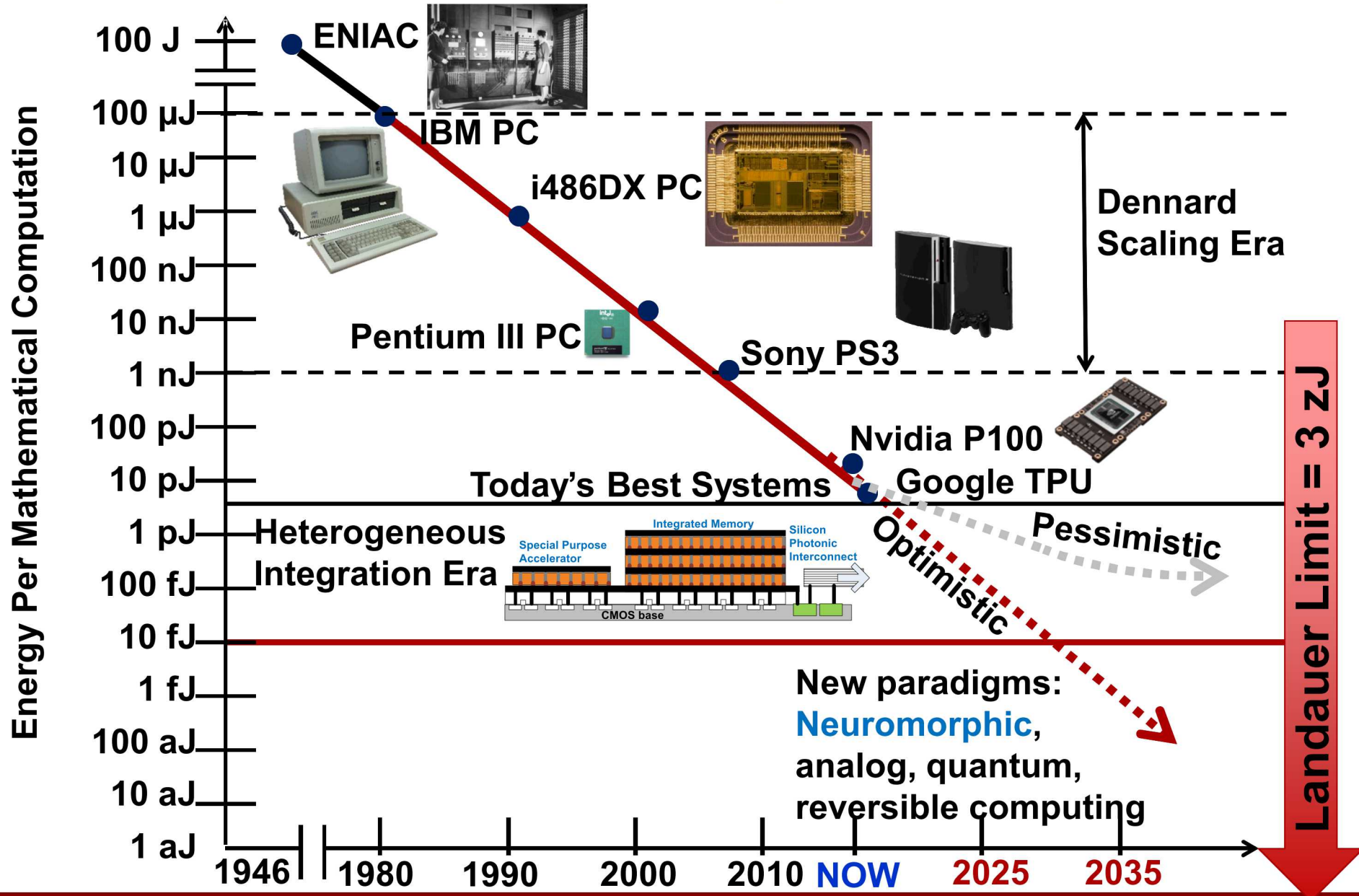


Designing and Modelling Analog Neural Network Training Accelerators

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

Evolution of Computing Machinery



AI Across Power Envelopes

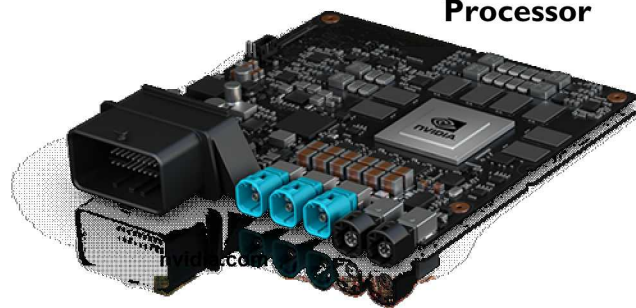
IoT, Edge, and
Mobile
Computing

Self Driving Cars,
Unmanned Arial
Vehicles, and Satellite
Computing

Datacenters, HPC



nest.com



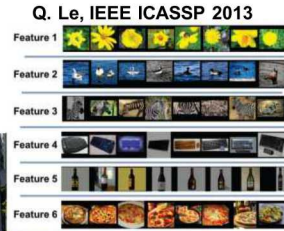
Nvidia Self-Driving
Processor



wikimedia.org



ASCI Red Supercomputer



1W

10W

10^2 W

10^3 W

10^4 W

10^5 W

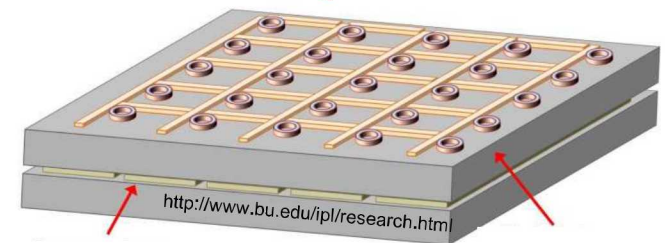
10^6 W

The Von Neumann Bottleneck



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

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



Processor Layer

Photonic Layer

Optical interconnects 100 fJ to 1 pJ

**Communications require
orders of magnitude more
energy!**

Use Resistive Memories for Local Computation

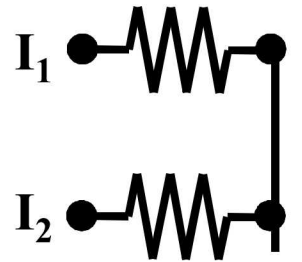


$$V = I \times R$$

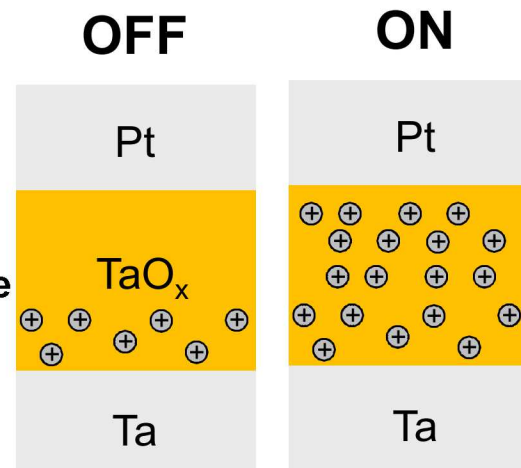
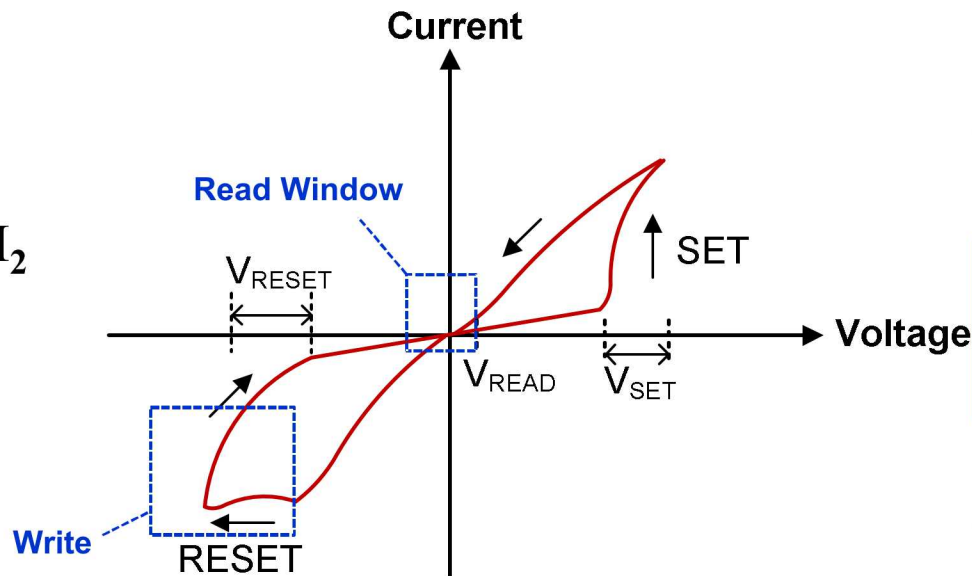
$$I = G \times V$$

multiplication

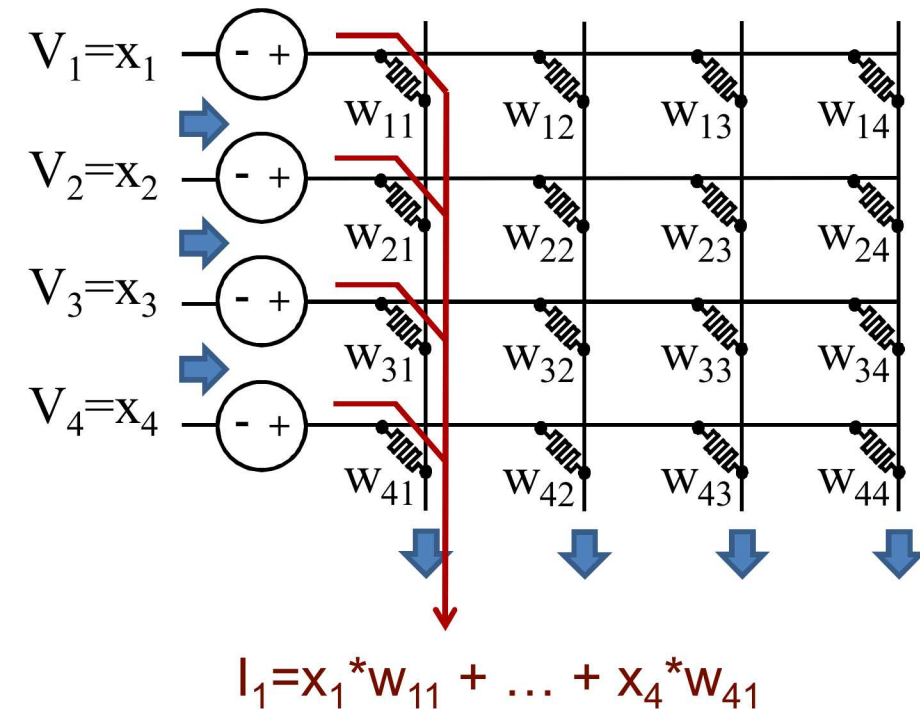
- A resistive memory or ReRAM is a programmable resistor
 - Apply small voltages allows the conductance to be read: $I = G \times V$
 - Apply large voltages to change the resistance



$$\text{Addition: } I = I_1 + I_2$$



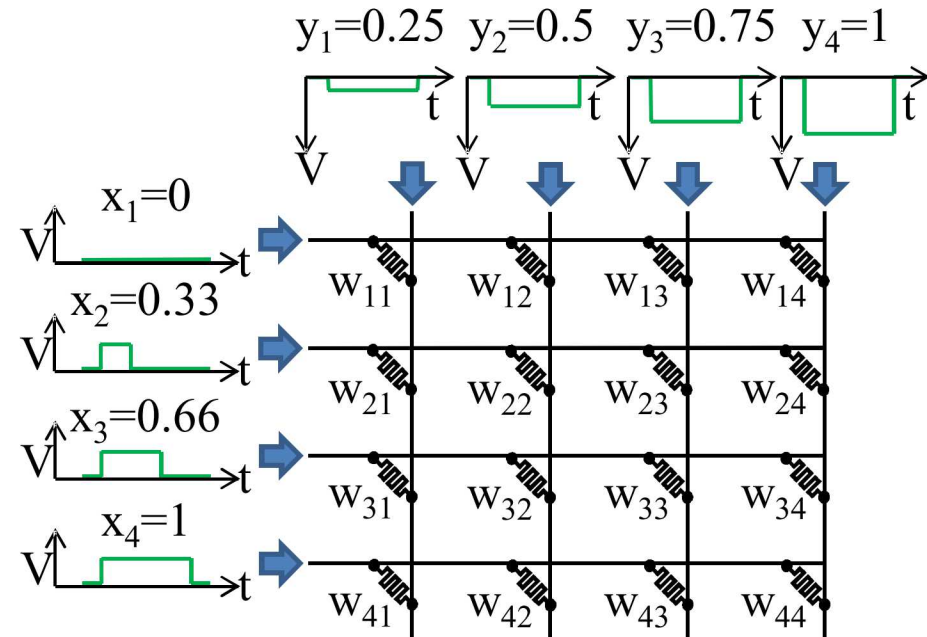
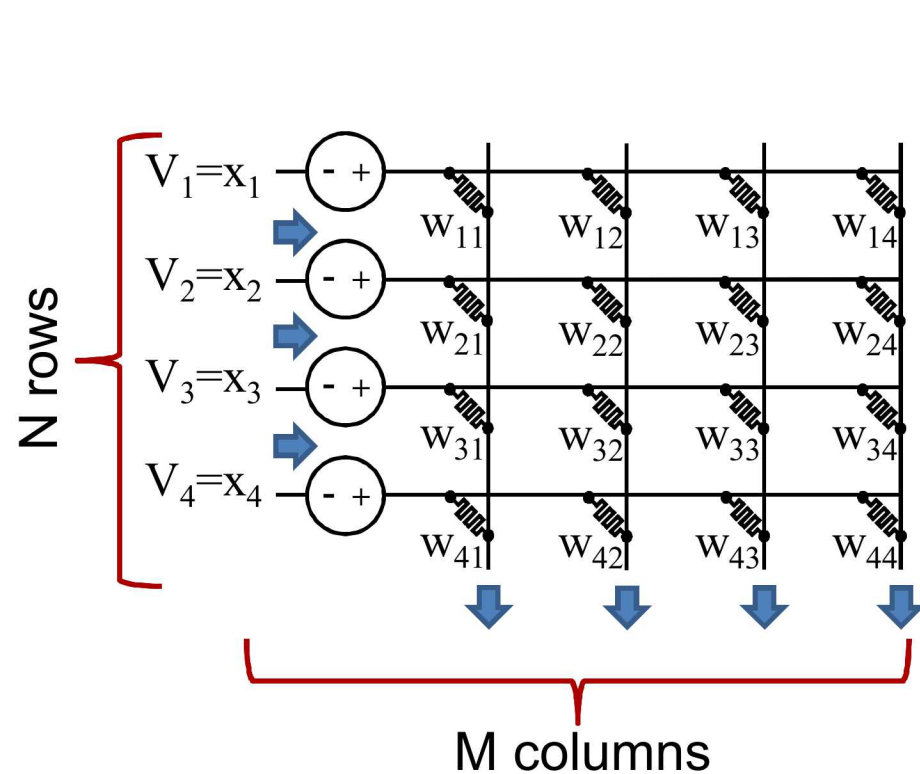
Directly Process in the Memory Itself



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

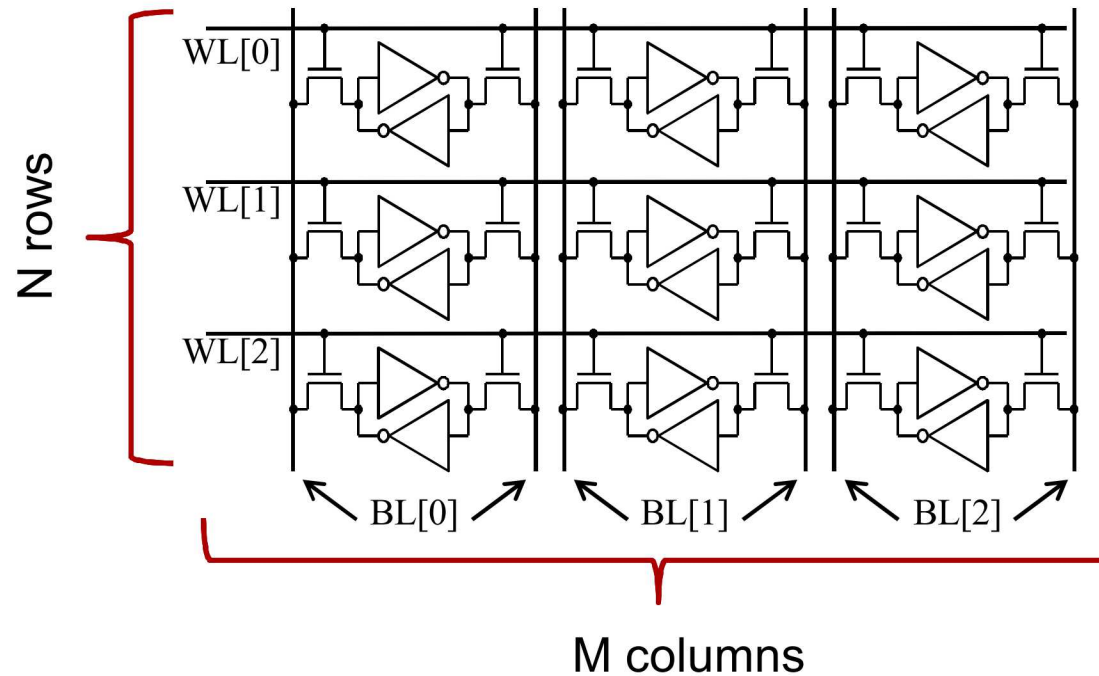
Crossbars Can Perform Parallel Reads and Writes



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

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

SRAM Arrays Require Charging Columns Multiple Times



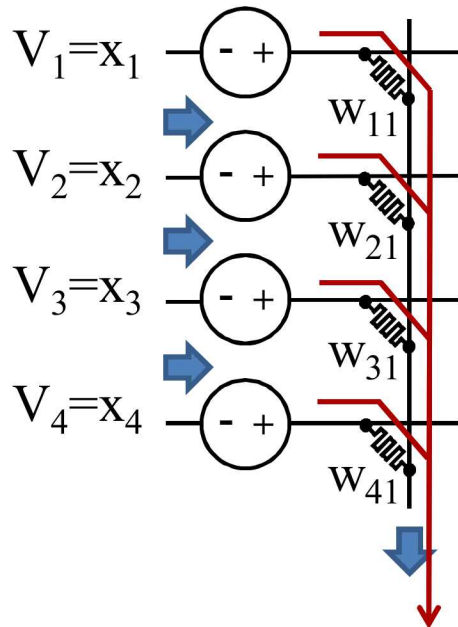
SRAMs must be read one row at a time, charging M columns
Each column wire length is $O(N)$.

Energy = N Rows $\times$ M Columns $\times$ $O(N)$ wire length

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

$O(N)$ times worse than a crossbar!

Need to Use Analog to Efficiently Discard Precision



Sum 1024 8 bit weights X 8 bit inputs:

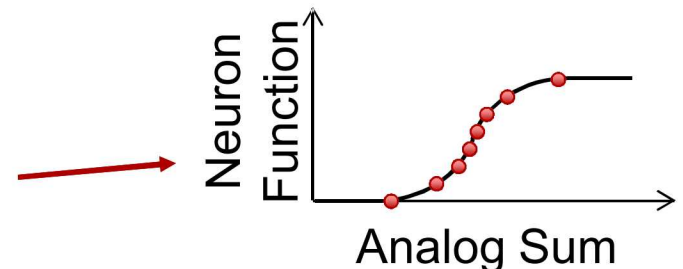
- Result has 26 bits of information!
- A 26 bit ADC would eliminate any analog advantage!

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

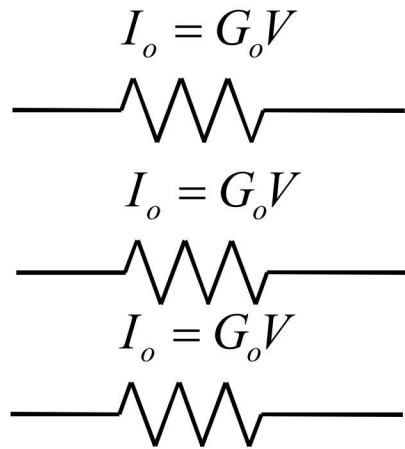
- i.e. digitize only 8 bits or fewer

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?

$$Energy = \underbrace{V^2 G_o \times N}_{\text{Power in each resistor} \times \text{number of resistors}} \times \frac{1}{\Delta f} \leftarrow \text{Determined by noise and SNR}$$

Power in each resistor ×
number of resistors

Determined by
noise and SNR

$$\text{Thermal Noise} = \langle \Delta I^2 \rangle$$

$$= N \times (4k_b T \times G_o \times \Delta f)$$

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

$$SNR^2 = \frac{(NI_o)^2}{\langle \Delta I^2 \rangle}$$

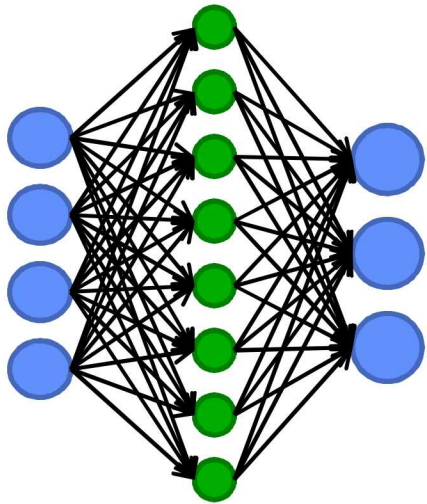
$$Energy = 4k_b T \times SNR^2$$

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

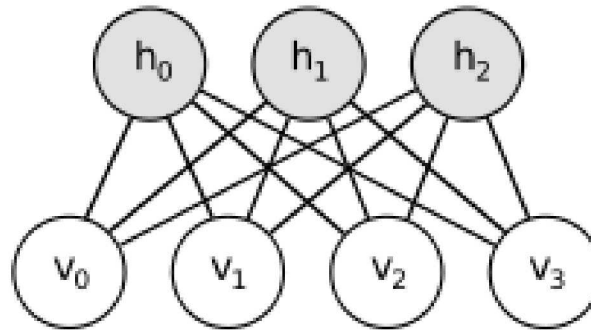
*we are assuming we need some fixed precision on the output, and don't need full floating point accuracy

Want To Accelerate Many Different Neural Algorithms

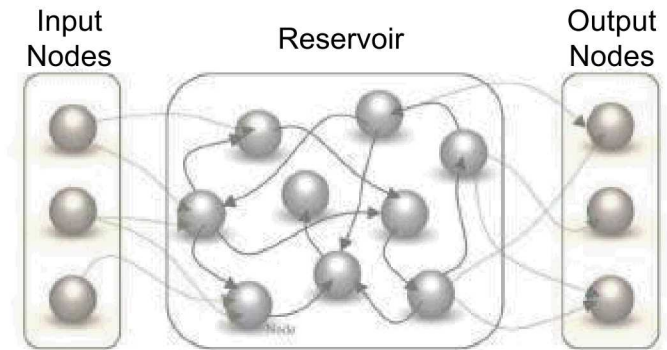
Backpropagation



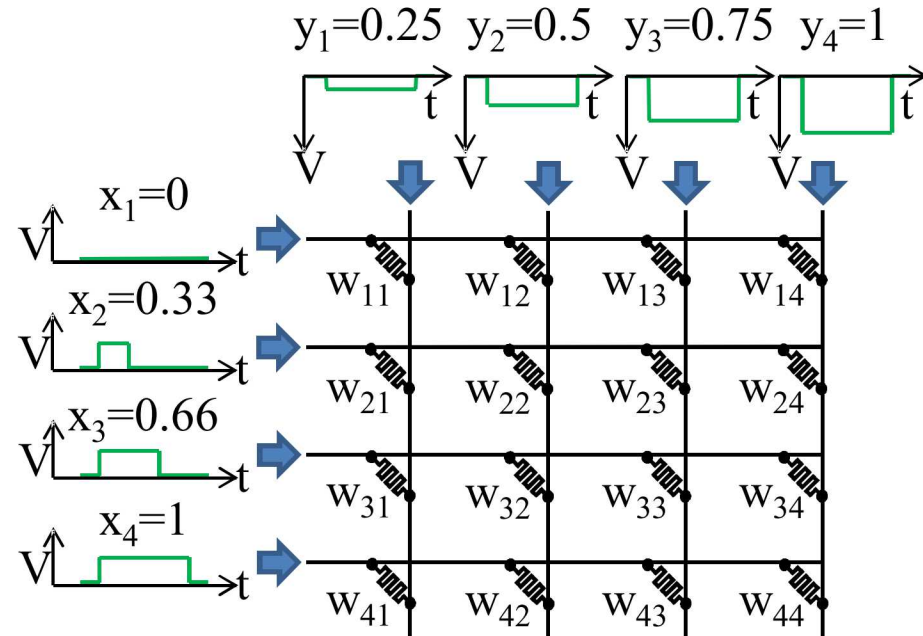
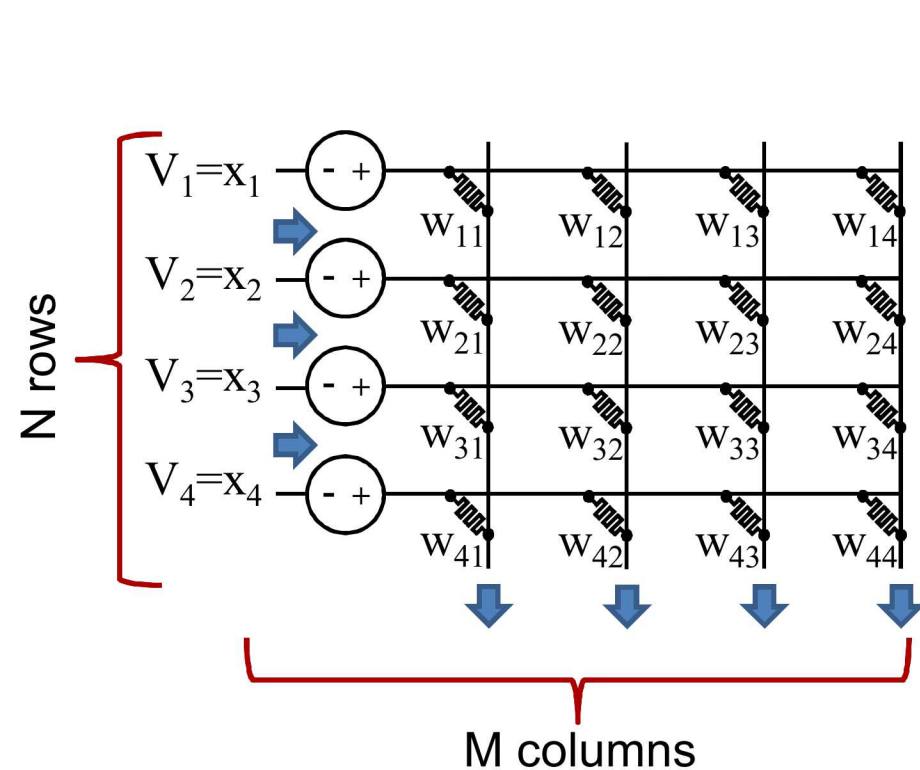
Sparse Coding



Liquid State Machine



Crossbars Can Perform Parallel Reads and Writes

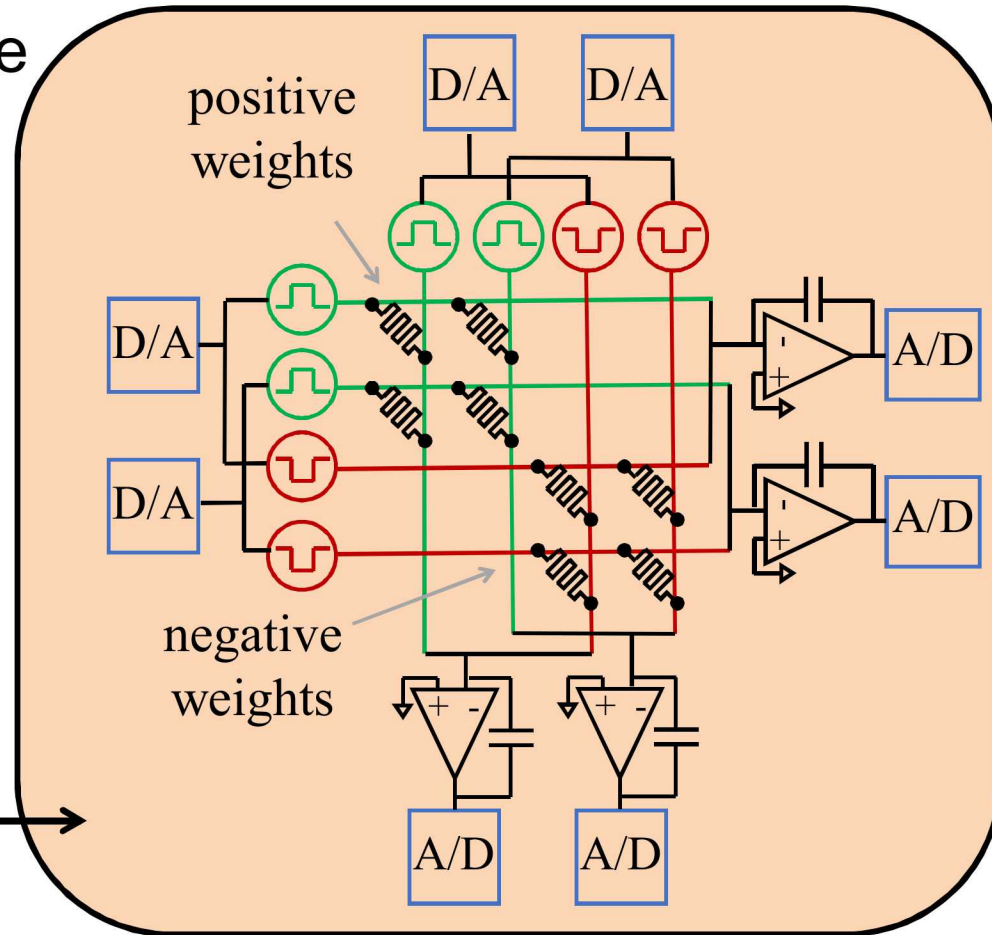
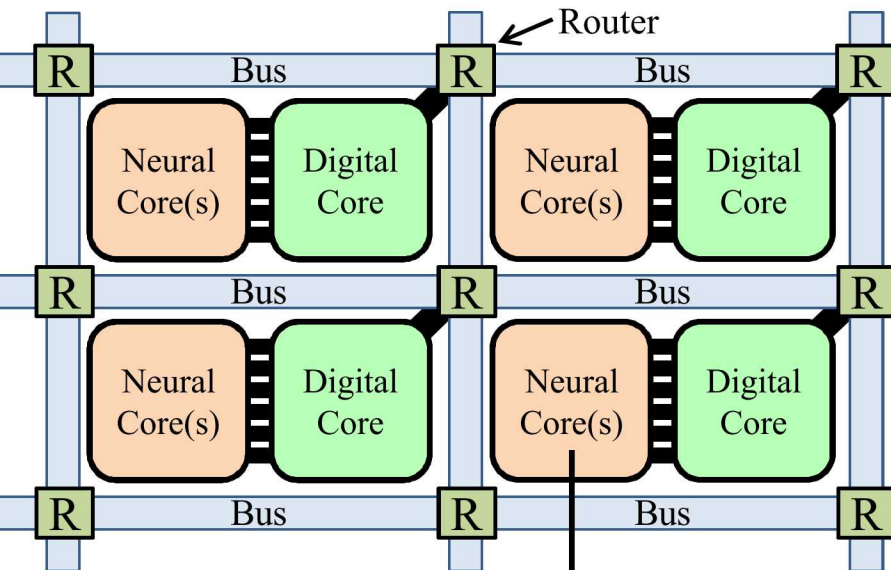


Energy to charge the crossbar is CV^2
 $E \propto C \propto \text{number of RRAMs} \propto N \times M$

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

General Purpose Neural Architecture

Run *any* neural algorithm on the same hardware



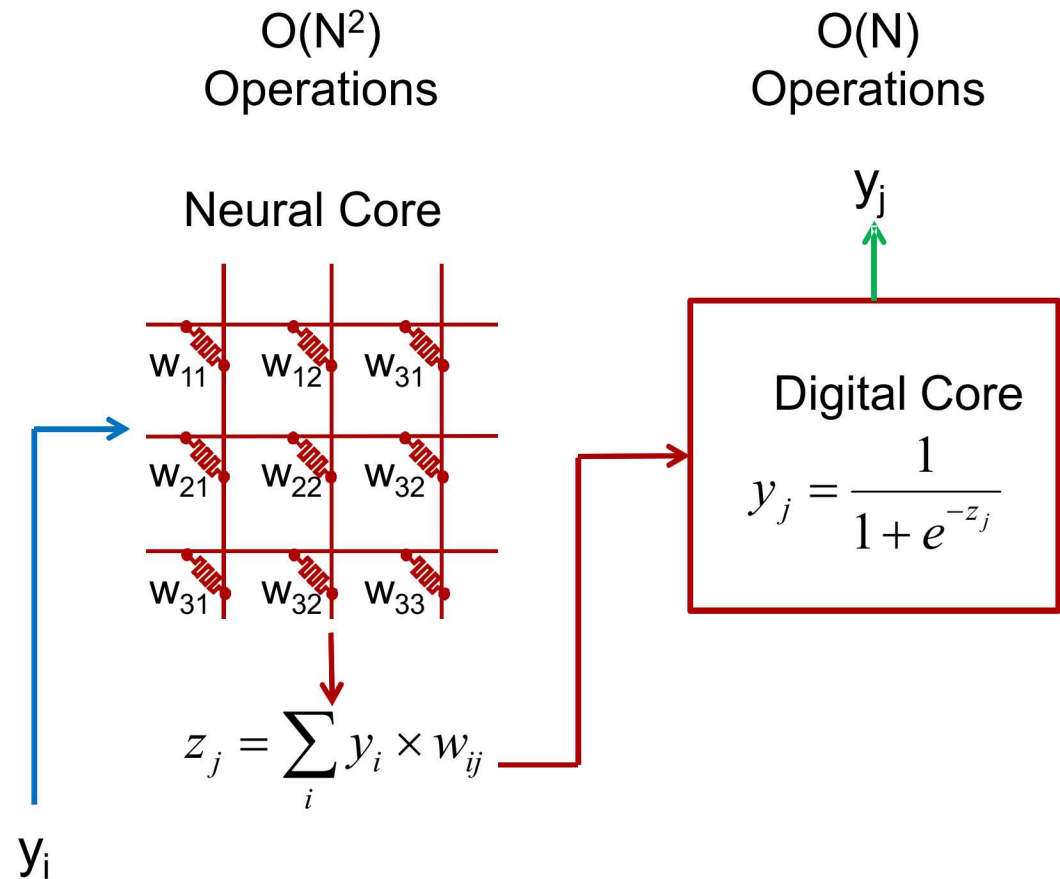
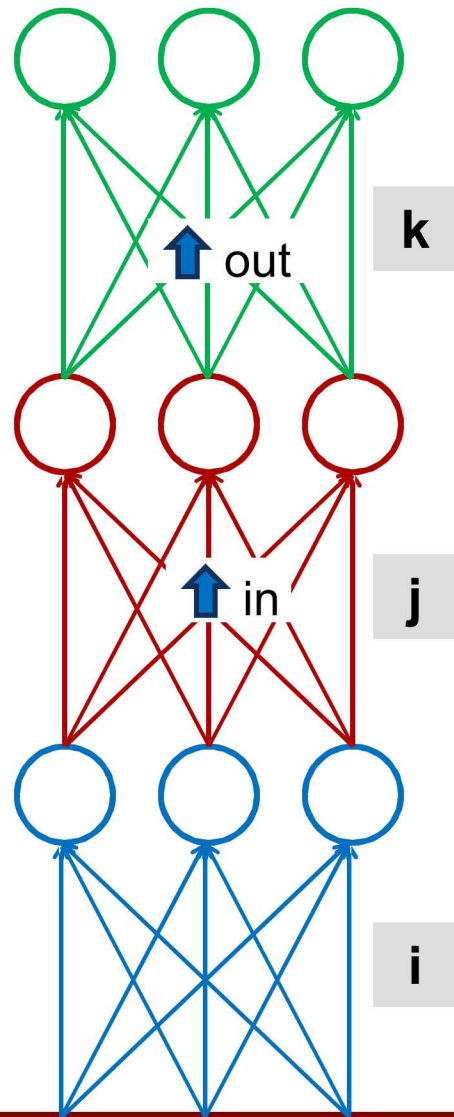
Neuromorphic core:

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

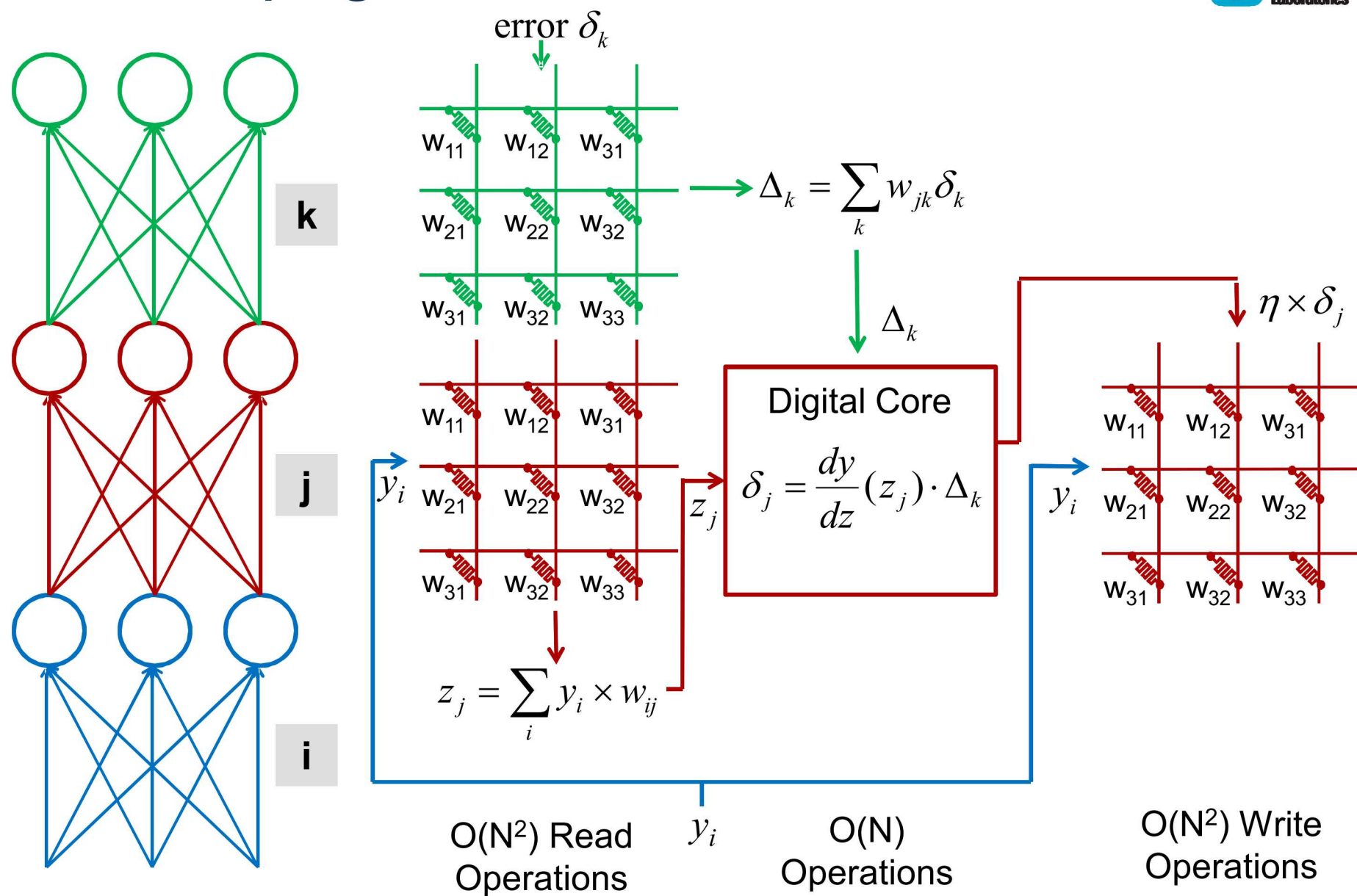
Digital Core:

- 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

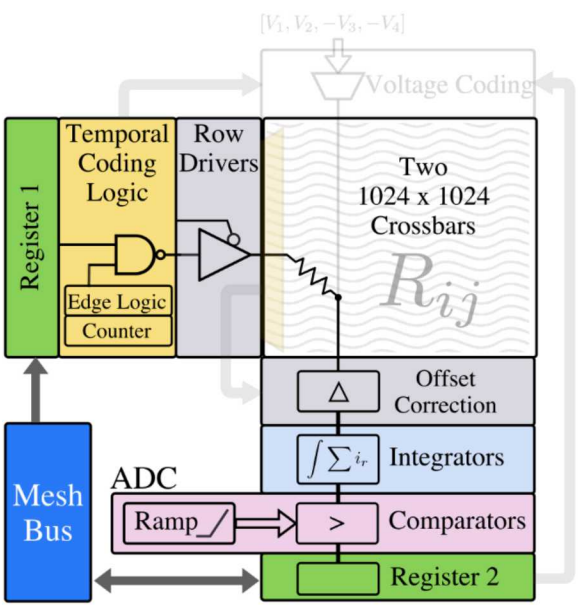


Back Propagation

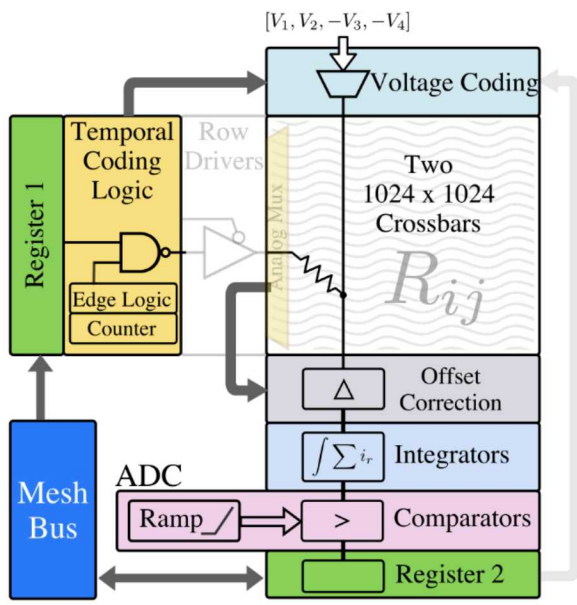


Design & Model Detailed Architecture

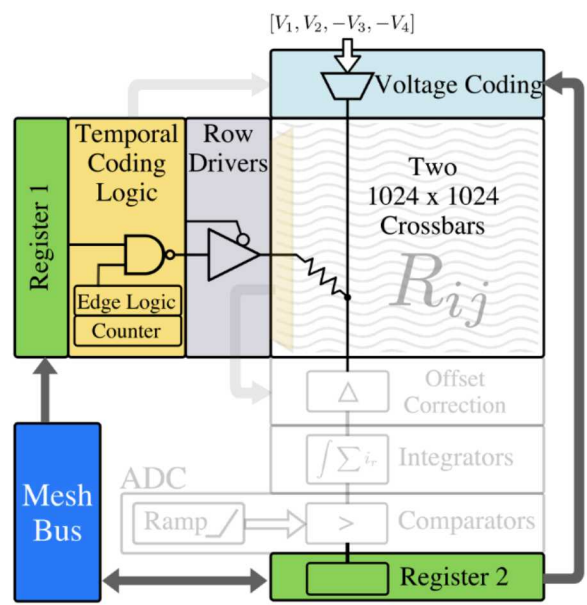
Vector Matrix Multiply



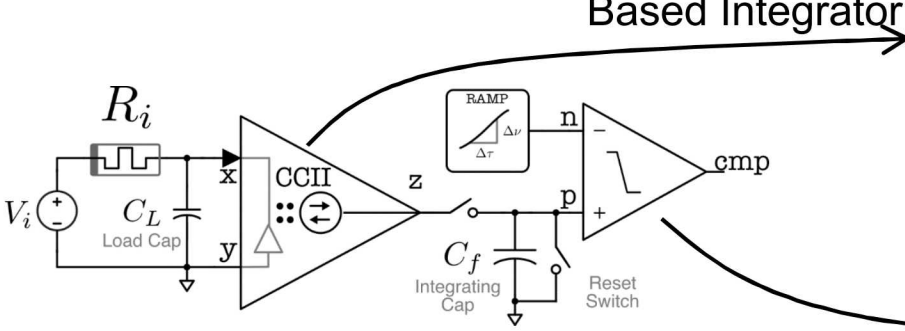
Matrix Vector Multiply



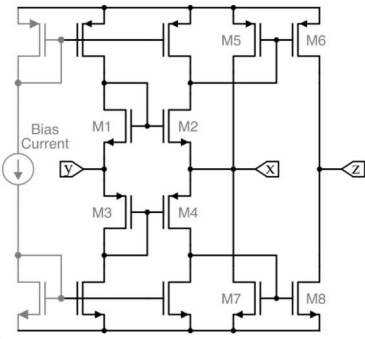
Outer product Update



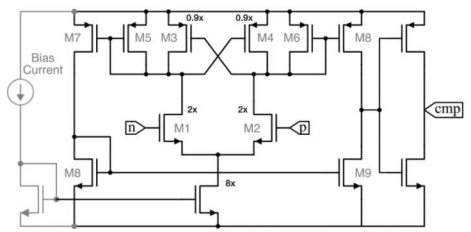
Neuron Circuitry



Current Conveyor Based Integrator

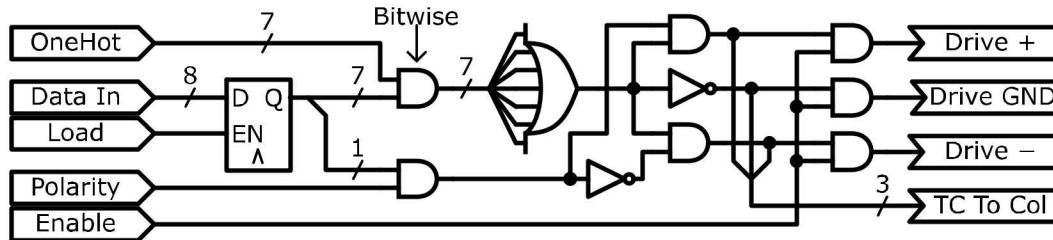


Comparator

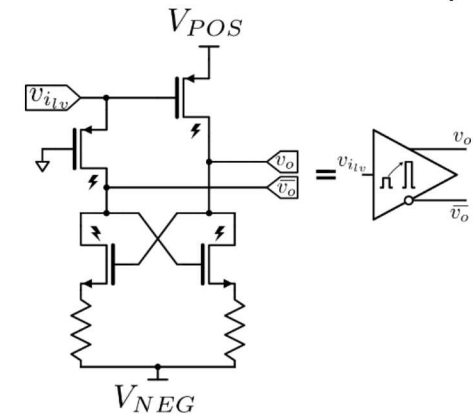


Row & Column Driver Circuitry

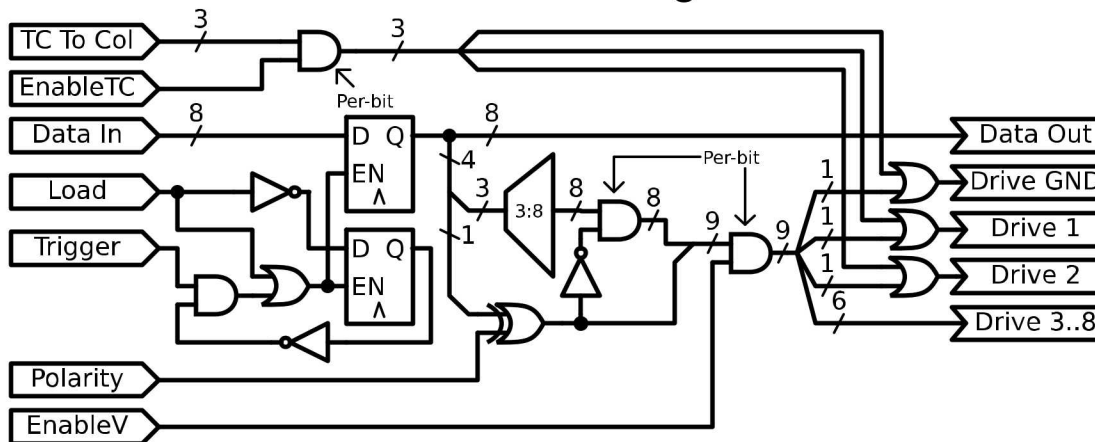
Row Driver Logic



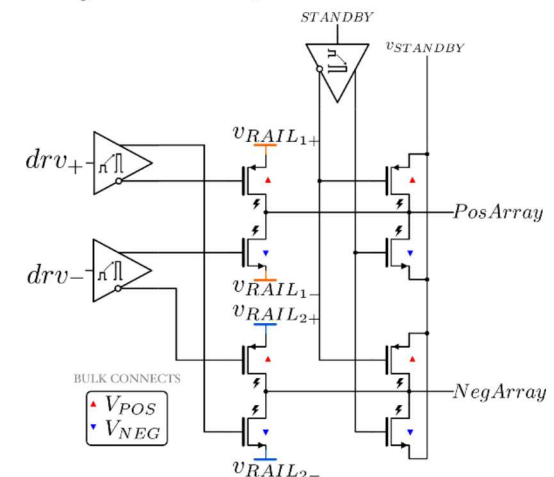
Voltage level shifter (drive high V transistor with low V)



Column Driver Logic



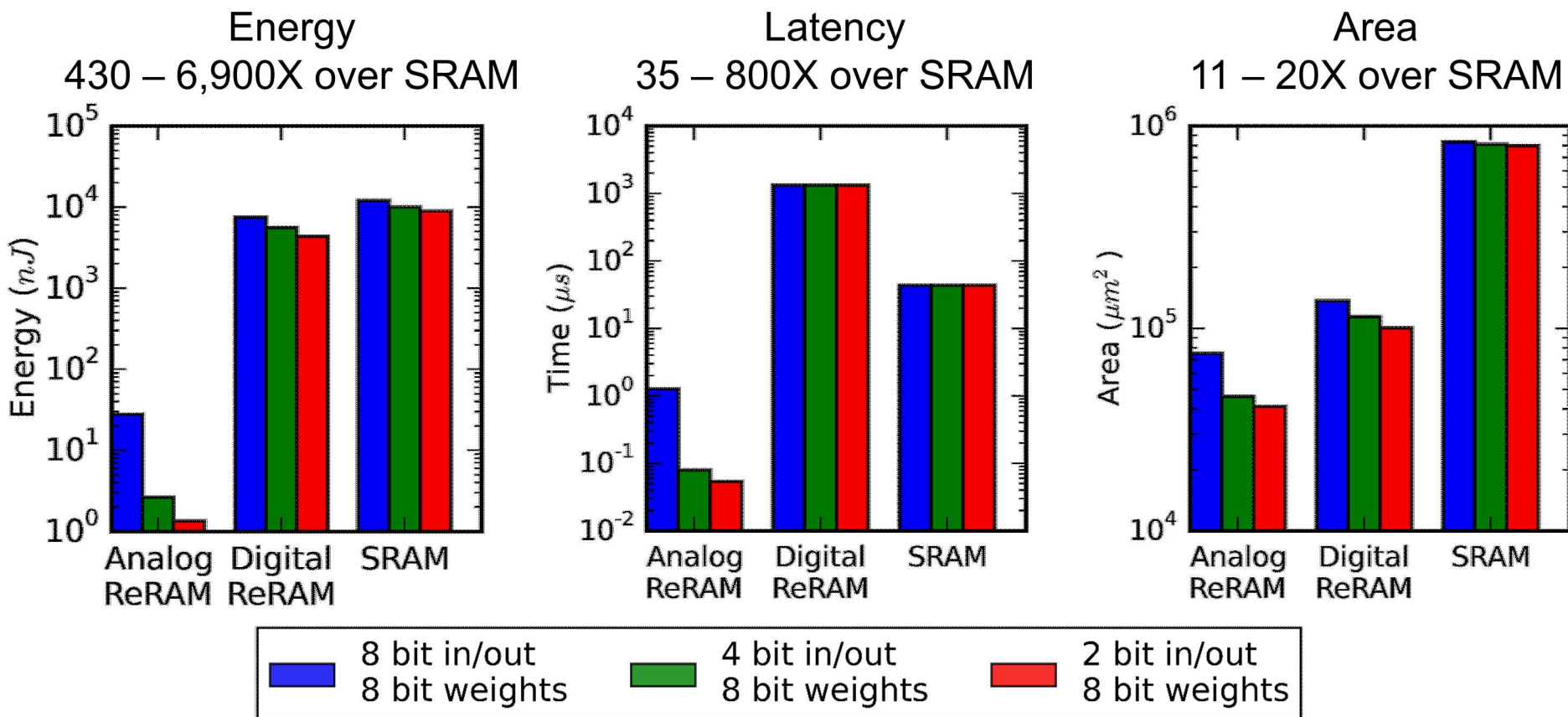
Array driver pass transistors



Compare Architectures

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

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



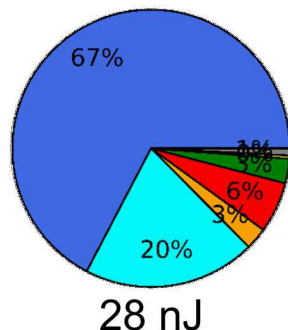
Used a commercial 14/16 nm PDK

***Requires 100 MΩ on state devices

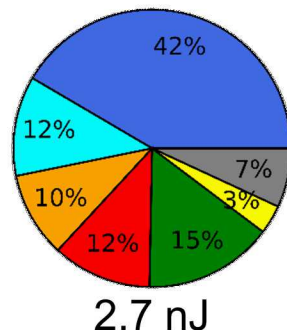
Neural Core Energy Analysis

Analog ReRAM

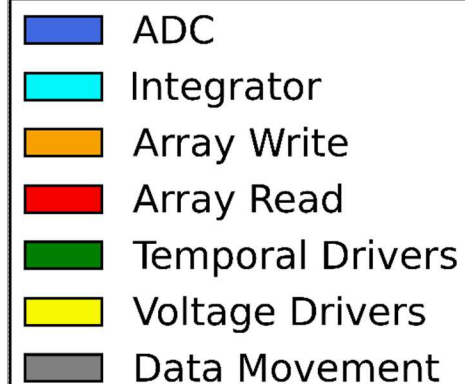
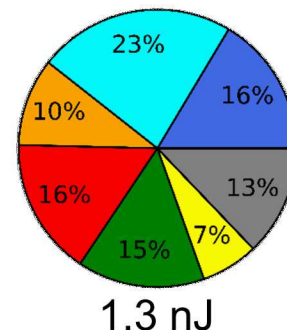
8 bits In/out
8 bit weights



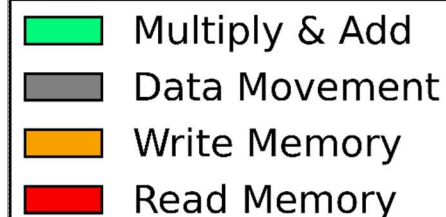
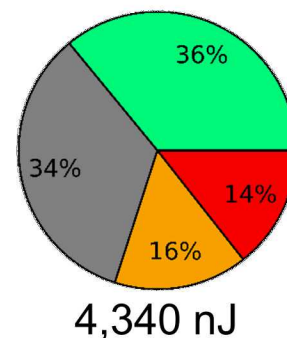
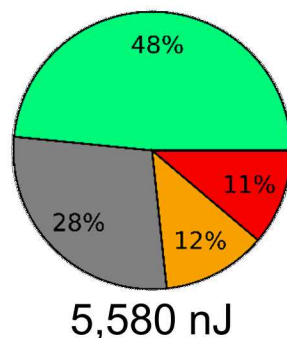
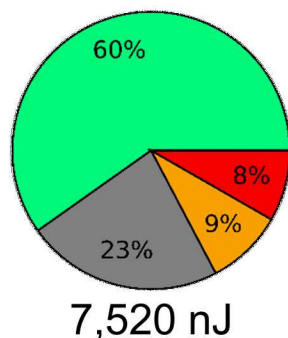
4 bits In/out
8 bit weights



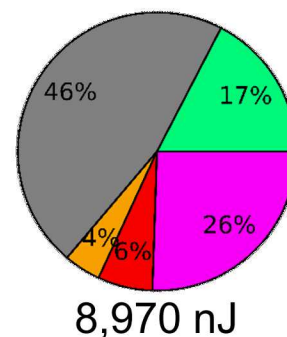
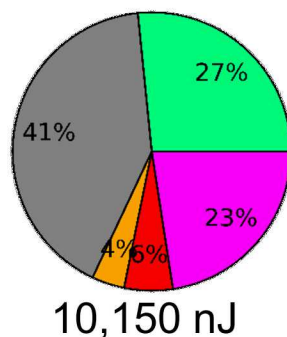
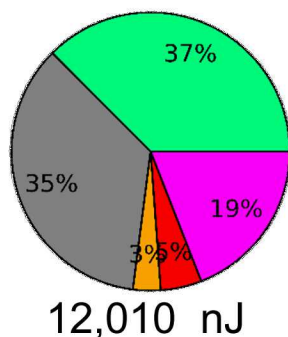
2 bits In/out
8 bit weights



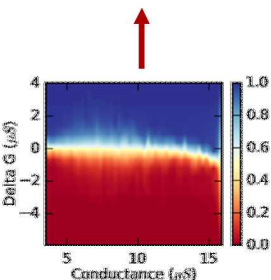
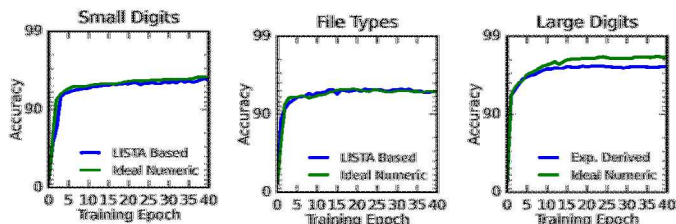
Digital ReRAM



SRAM



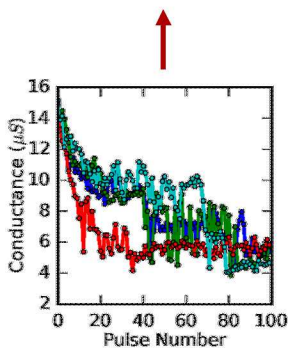
Multiscale Model of a Neural Training Accelerator



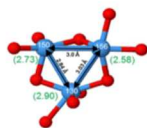
#ROSS SIM

Sandia Cross-Sim:

Translates device measurements and crossbar circuits to algorithm-level performance



Memristor fabrication and measurements in MESAFab



DFT of model of oxide physics, bands

Materials

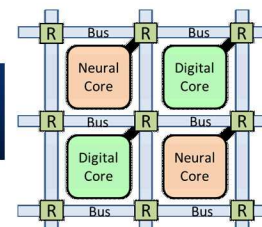
In situ TEM of filament switching: Use DFT model to interpret EELS signature

Target Algorithms

- Deep Learning
- Sparse Coding
- Liquid State Machines

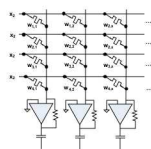
Algorithms

Architecture



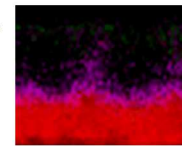
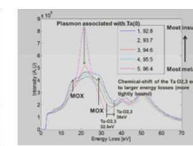
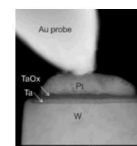
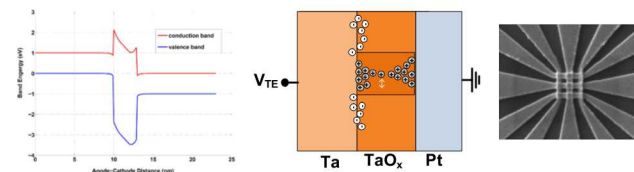
Modified McPAT/CACTI: Model performance and energy requirements

Circuits



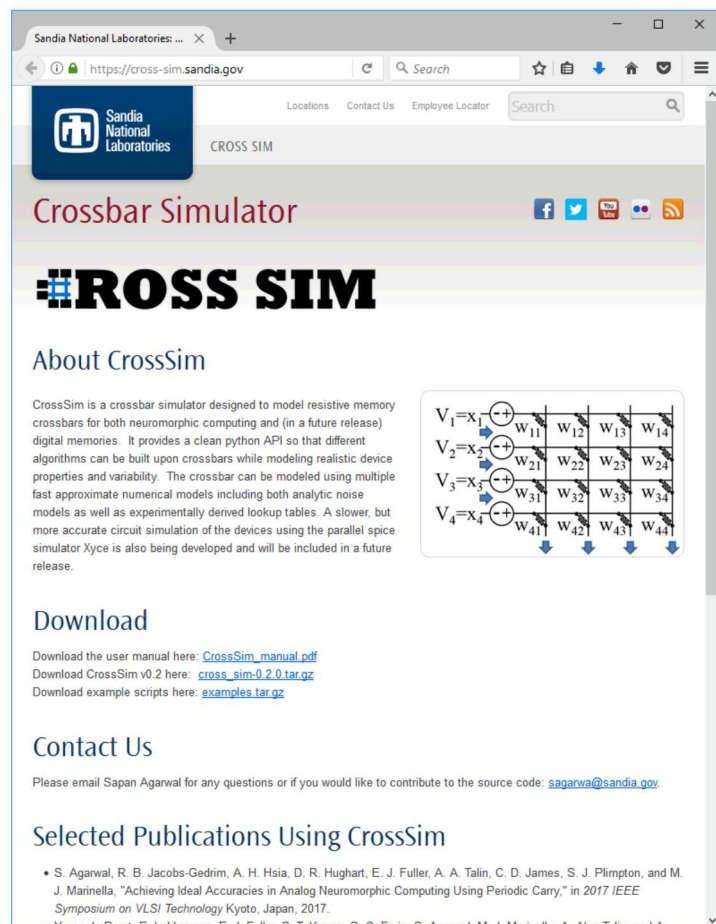
Sandia's Xyce Circuit Sim: Simulate crossbar circuits based on our devices

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



#ROSS SIM

<https://cross-sim.sandia.gov>



Crossbar Simulator

#ROSS SIM

About CrossSim

CrossSim is a crossbar simulator designed to model resistive memory crossbars for both neuromorphic computing and (in a future release) digital memories. It provides a clean python API so that different algorithms can be built upon crossbars while modeling realistic device properties and variability. The crossbar can be modeled using multiple fast approximate numerical models including both analytic noise models as well as experimentally derived lookup tables. A slower, but more accurate circuit simulation of the devices using the parallel spice simulator Xyce is also being developed and will be included in a future release.

Download

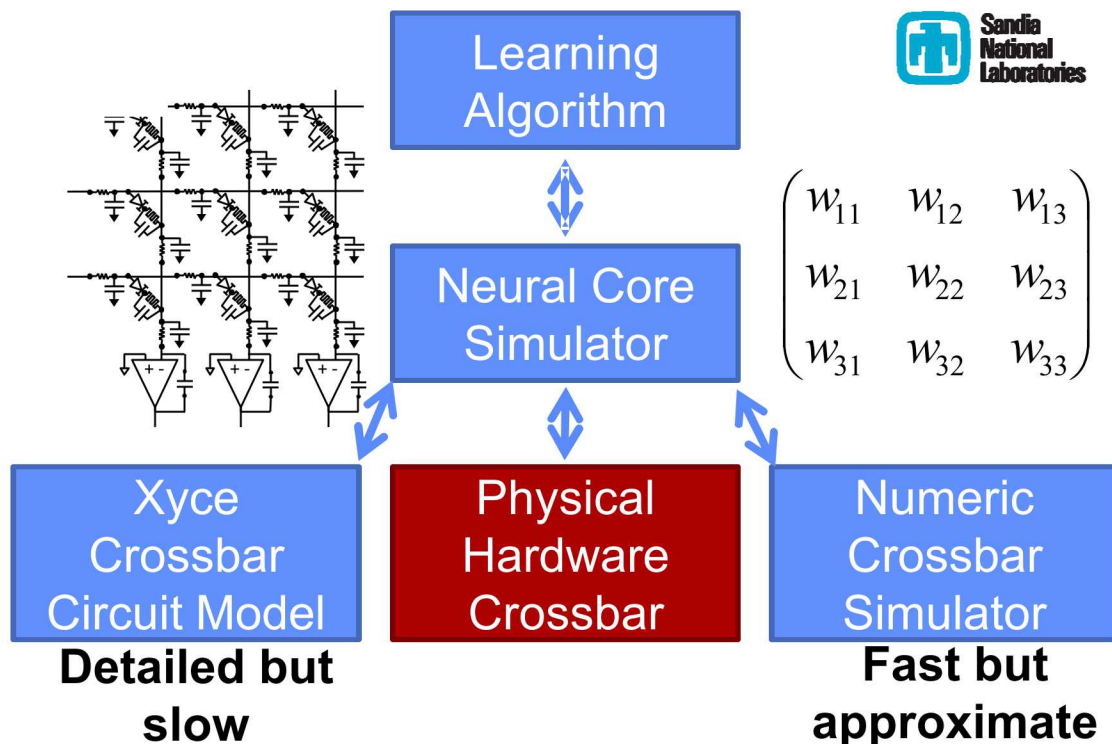
Download the user manual here: [CrossSim_manual.pdf](#)
 Download CrossSim v0.2 here: [cross_sim-0.2.0.tar.gz](#)
 Download example scripts here: [examples.tar.gz](#)

Contact Us

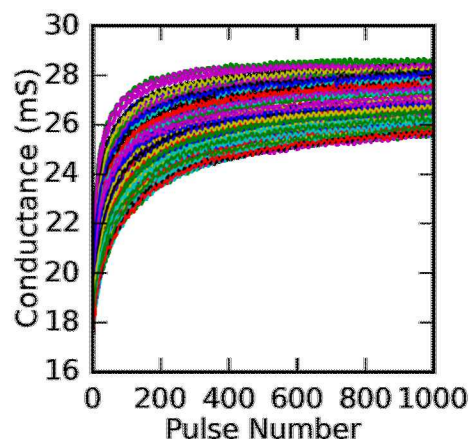
Please email Sapan Agarwal for any questions or if you would like to contribute to the source code: saganwal@sandia.gov.

Selected Publications Using CrossSim

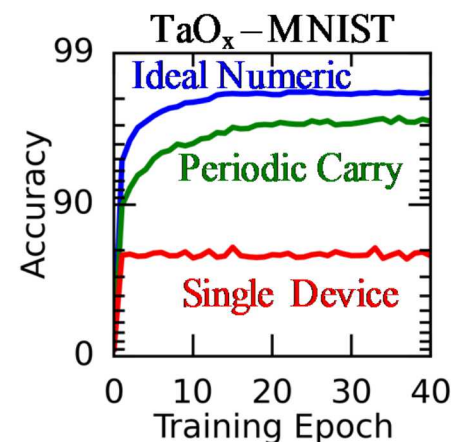
- S. Agarwal, R. B. Jacobs-Gedrim, A. H. Hsia, D. R. Hughart, E. J. Fuller, A. A. Talin, C. D. James, S. J. Plimpton, and M. J. Marinella, "Achieving Ideal Accuracies in Analog Neuromorphic Computing Using Periodic Carry," in 2017 IEEE Symposium on VLSI Technology Kyoto, Japan, 2017.
- Y. van de Burst, E. Lubberman, E. J. Fuller, S. T. Keane, G. C. Eria, S. Agarwal, M. J. Marinella, A. Alan-Talin, and A.



Measured Devices



Algorithmic Performance



Simple Python API:

Do a matrix vector multiplication

`result = neural_core.run_xbar_mvm(vector)`

Simple API to model crossbars

```
# ***** 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
...
...

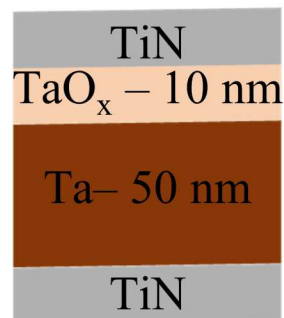
# ***** API for running neural operations
#           All crossbar details are transparent to the user

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

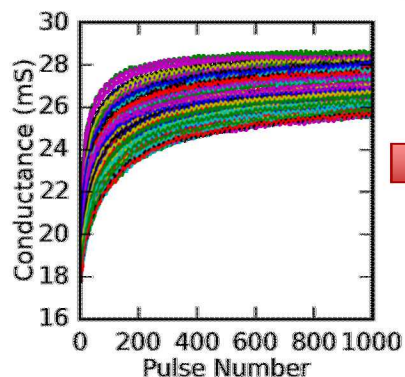
neural_core.set_matrix(weights) # set the initial weights
result = neural_core.run_xbar_vmm(vector) # Do a vector matrix multiply
result = neural_core.run_xbar_mvm(vector) # Do the transpose, a matrix vector mult.
neural_core.update_matrix(vector1,vector2) # Do an outer product update
```

Go from Measurement to Accuracy

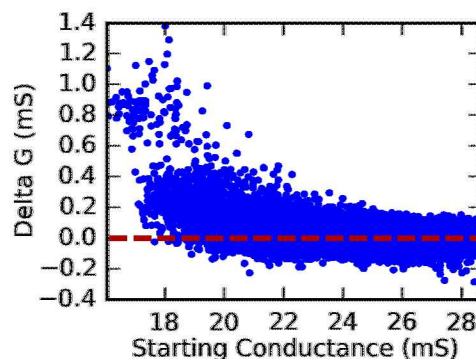
Fabricate
Device



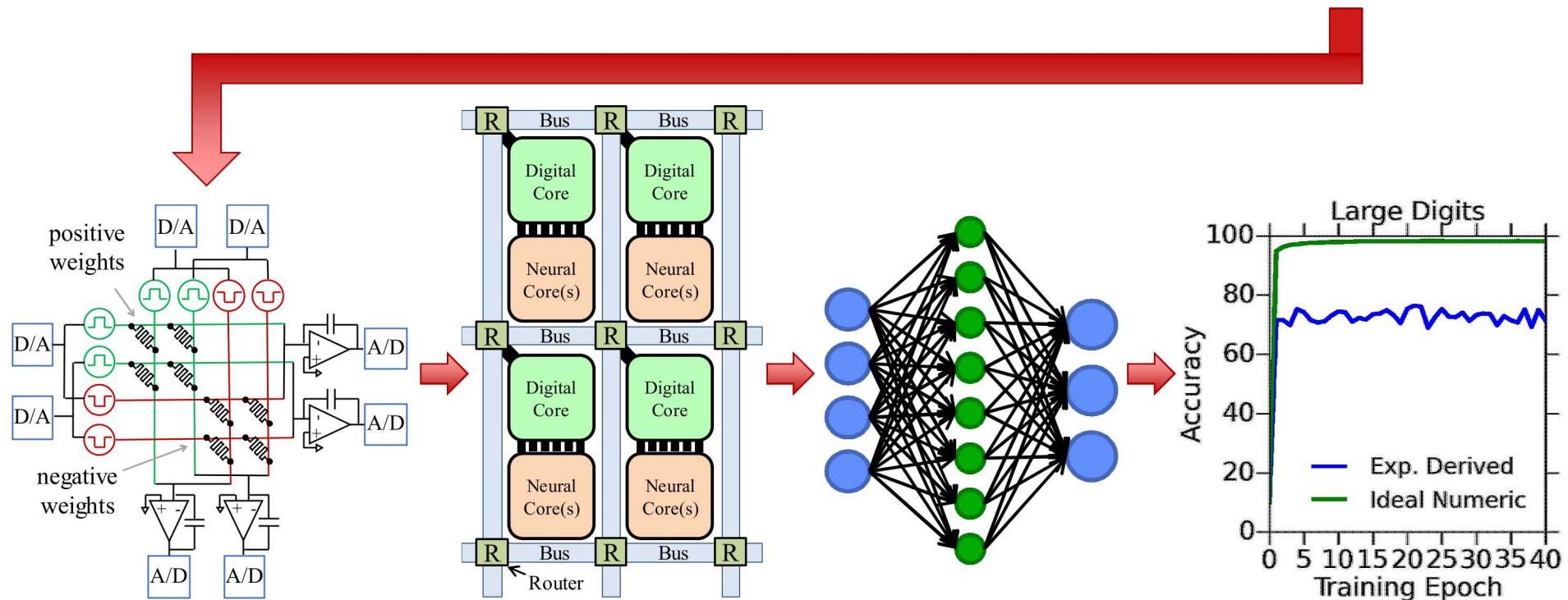
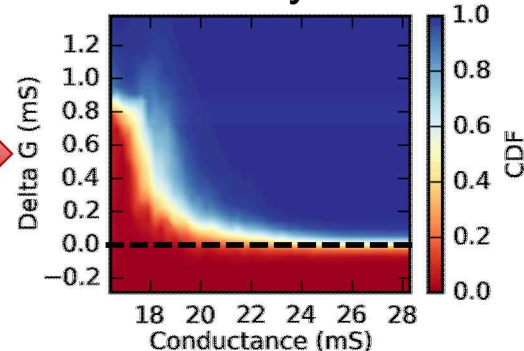
Measured Pulsing



ΔG Scatterplot

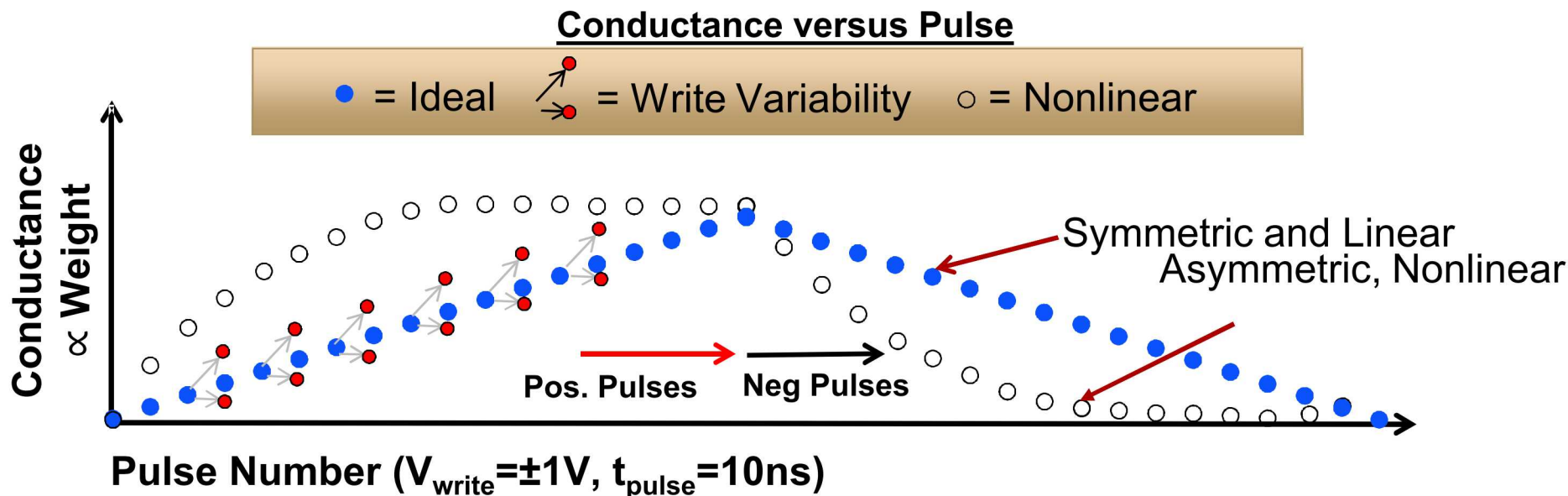


Cumulative
Probability of ΔG

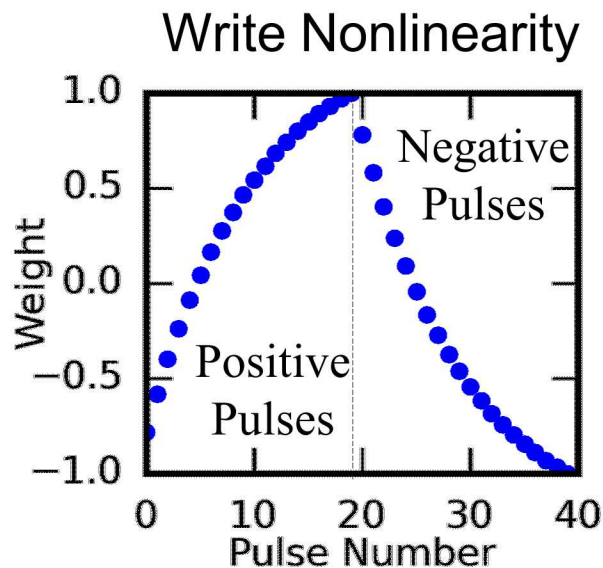


Experimental Device Nonidealities

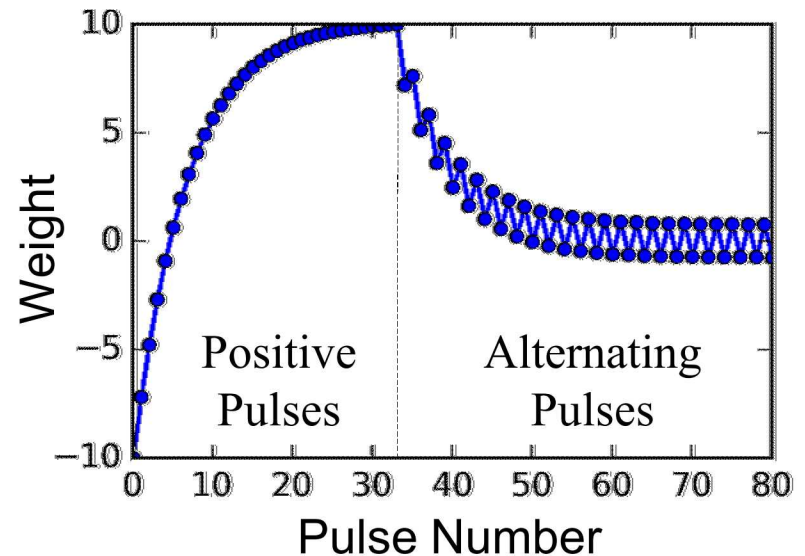
- Backprop training: Ideally weight would increase and decrease linearly proportional to learning rule result
- Key issue in experimental devices: altered the relationship between intended and actual update: **Write Nonlinearity, Asymmetry, Stochasticity**, Read Noise



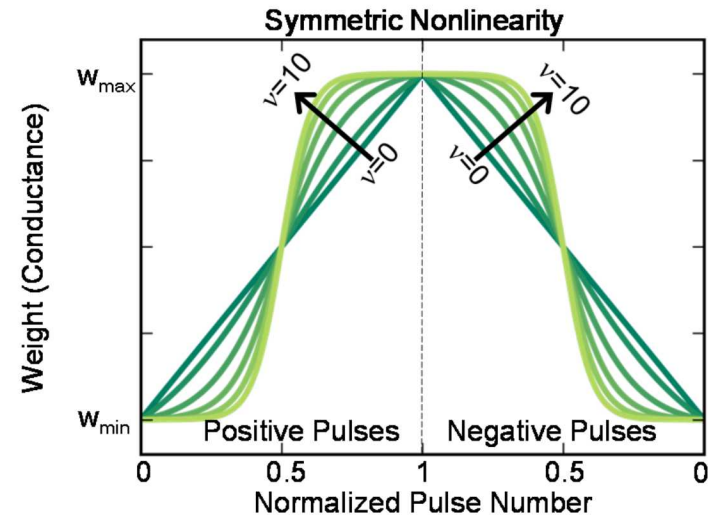
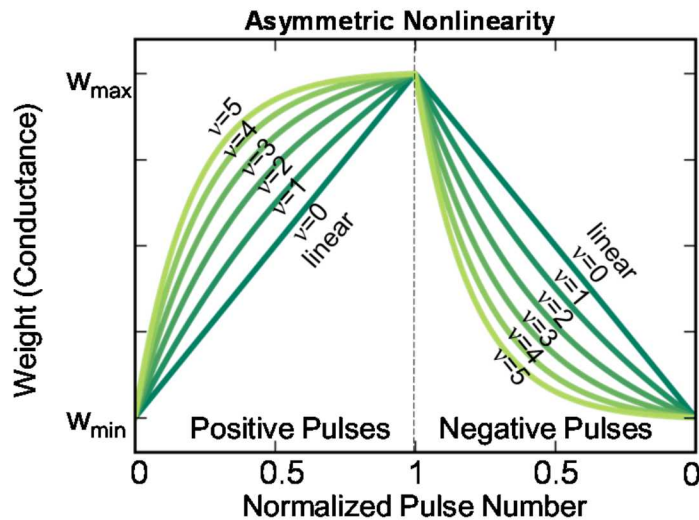
Asymmetric Write Nonlinearity Causes Weight Decay



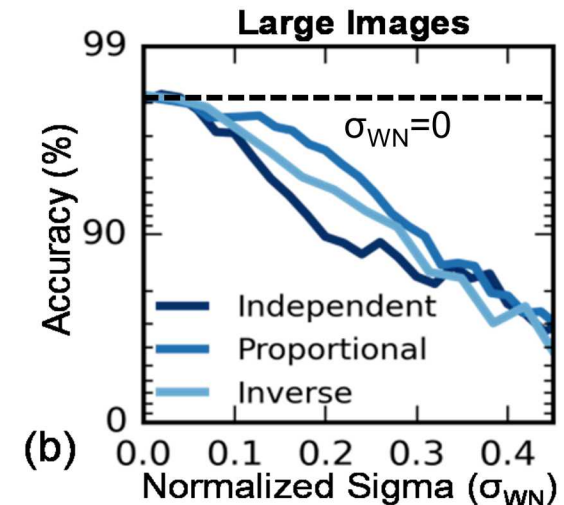
Alternating Pulses Cause Weight Decay



Modeling Device Requirements



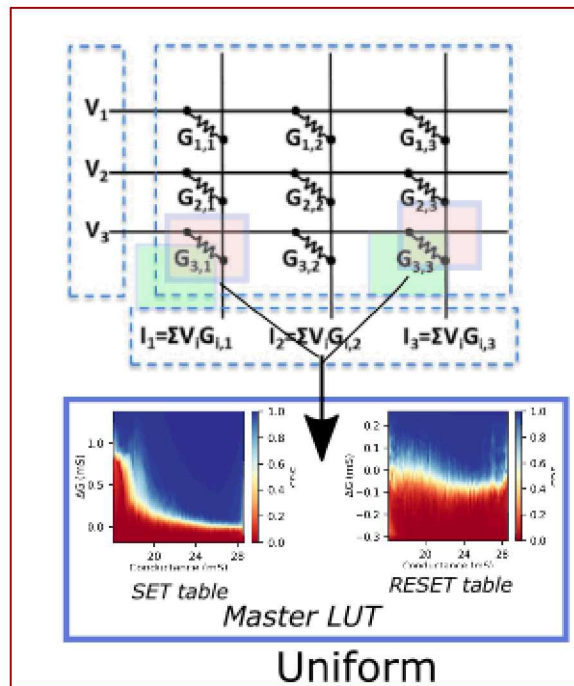
	MNIST
Read Noise σ (% Range)	5%
Write Noise σ (% Range)	0.4%
Asymmetric Nonlinearity (v)	0.1
Symmetric Nonlinearity (v)	5
Maximum Current	13 nA



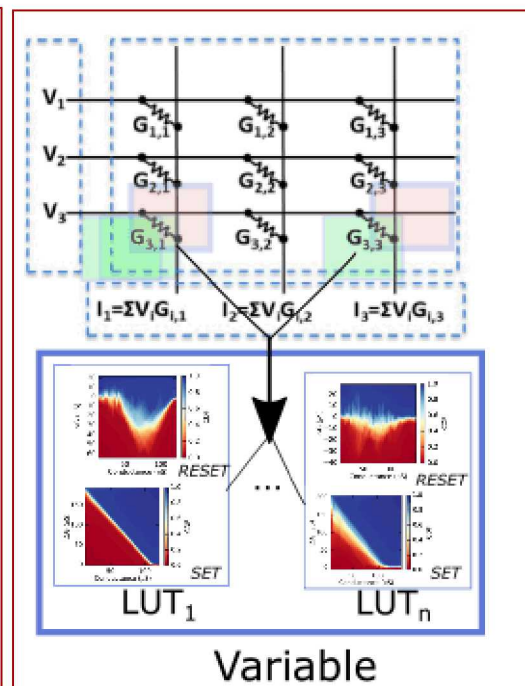
A realistic variability approach

- 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)

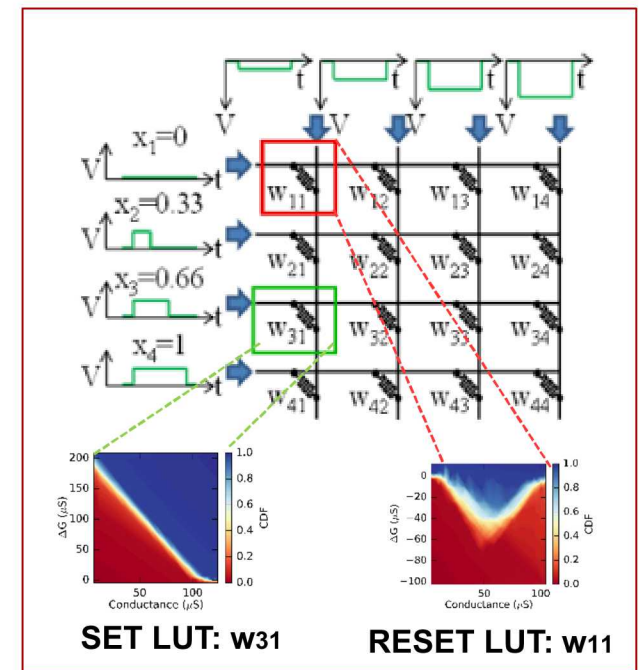
Old



New

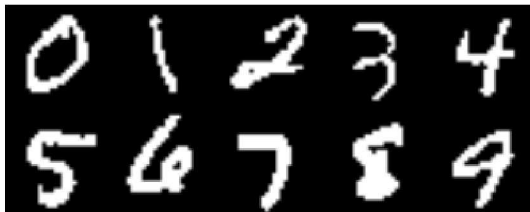
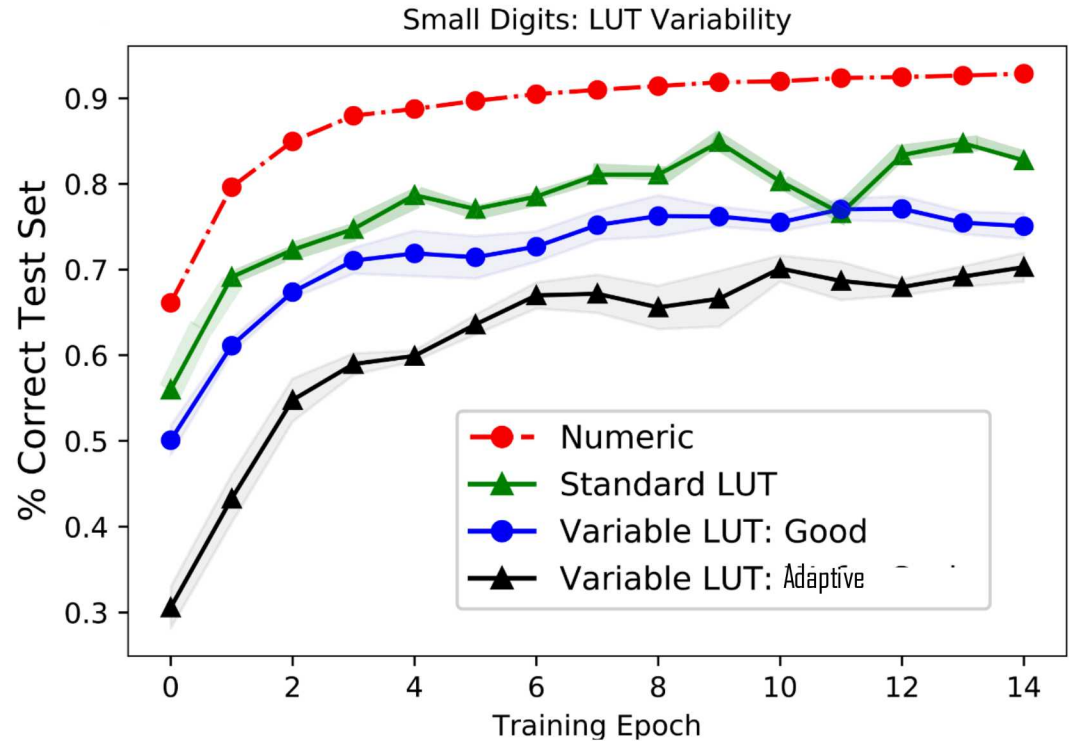


Online Updates



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

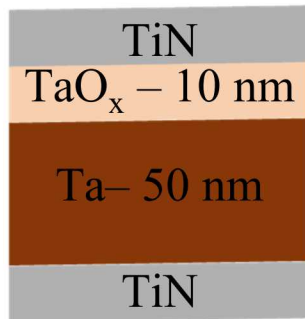
- Loss between uniform and Good Devices: **~4-7%**
- Loss between uniform and all functioning devices (adaptive): **~13-16%**



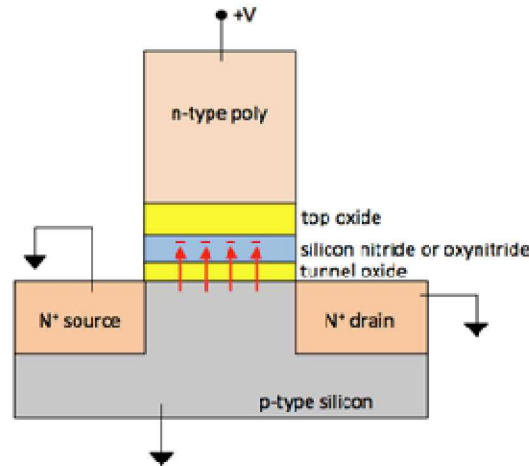
Data set	# Training Examples	# Test Examples	Network Size
UCI Iris Dataset [1]	100	50	4x8x3
UCI Small Digits[1]	3,823	1,797	64x36x10
File Types[2]	4,501	900	256x512x9
MNIST Large Digits[3]	60,000	10,000	784x300x10

Compare Analog Devices

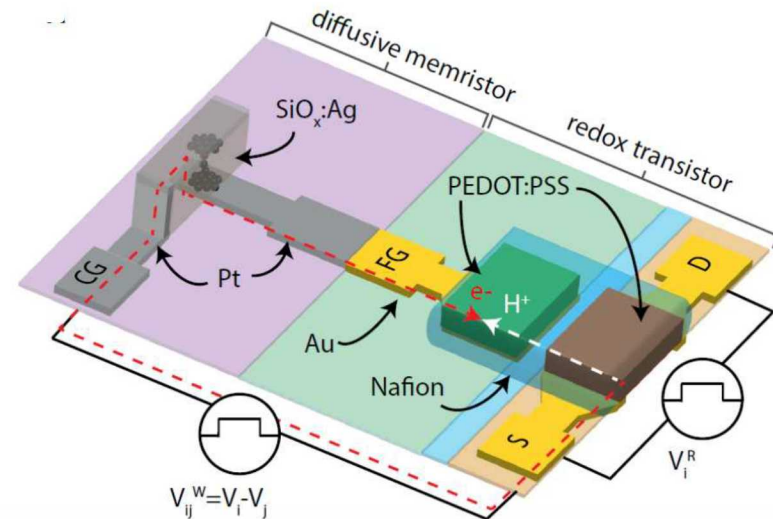
ReRAM



SONOS Silicon-Oxygen- Nitrogen-Oxygen-Silicon



Ionic Floating-Gate Memory

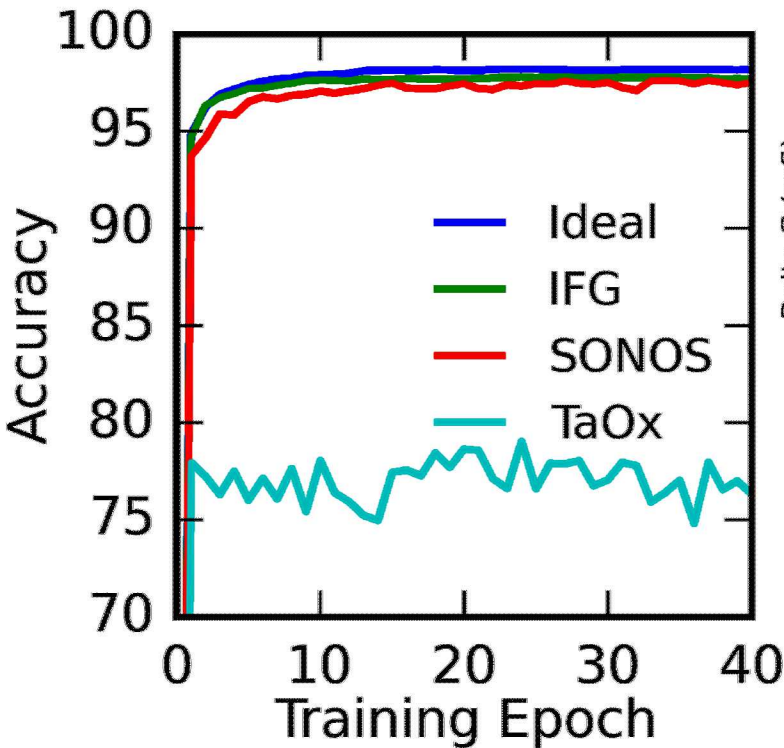


R. B. Jacobs-Gedrim *et al.*, "Impact of Linearity and Write Noise of Analog Resistive Memory Devices in a Neural Algorithm Accelerator," IEEE International Conference on Rebooting Computing (ICRC) Washington, DC, November 2017.

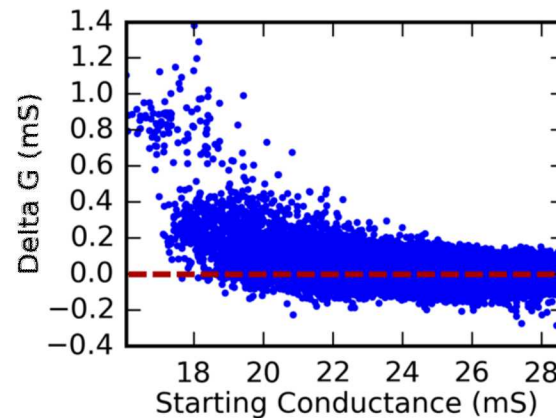
S. Agarwal *et al.*, "Using Floating Gate Memory to Train Ideal Accuracy Neural Networks," *IEEE Journal of Exploratory Solid-State Computational Devices and Circuits*, 2019

E. J. Fuller *et al.*, "Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing," *Science*, vol. 364, no. 6440, pp. 570-574, 2019.

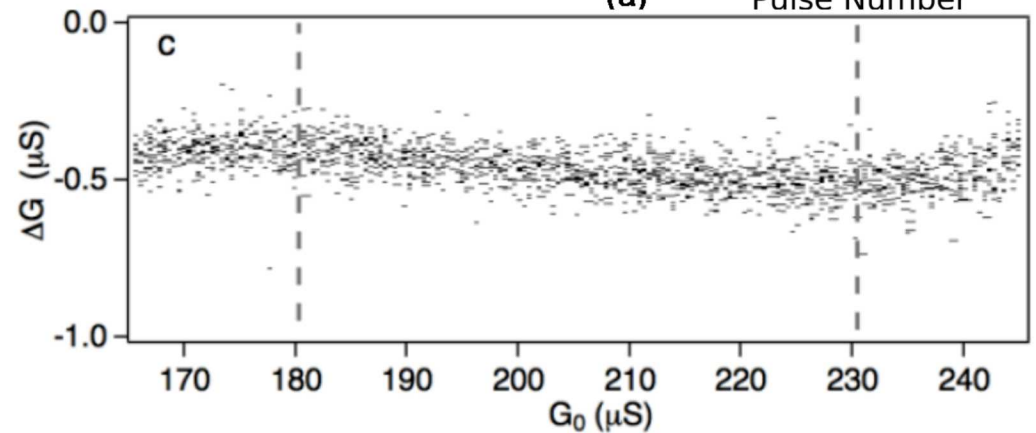
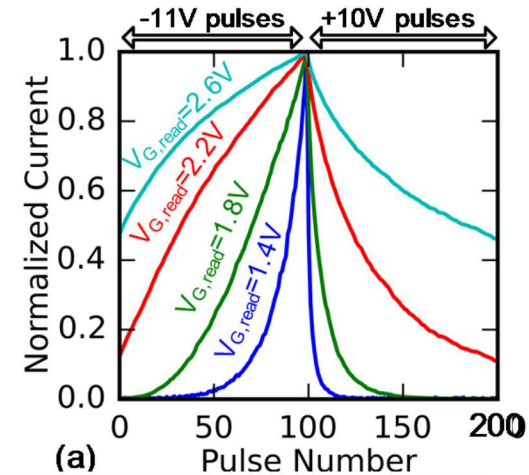
Three Terminal Devices Tend to Have Higher Accuracy



ReRAM



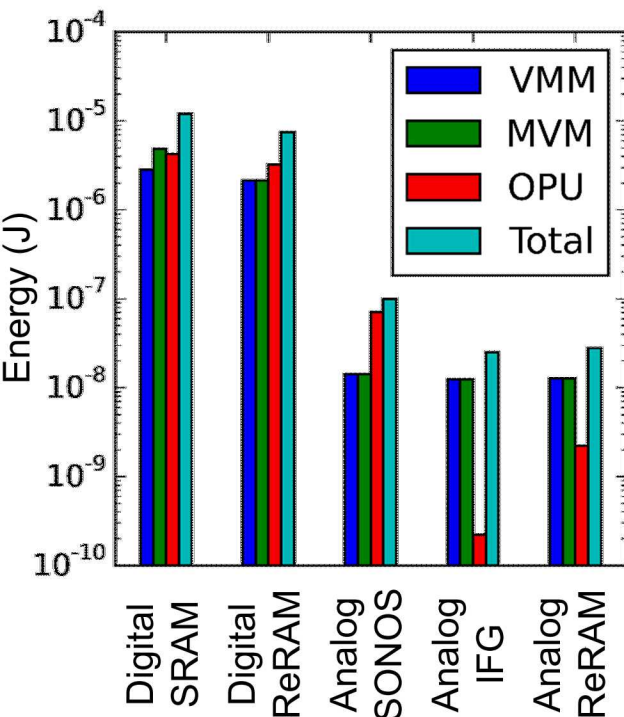
SONOS



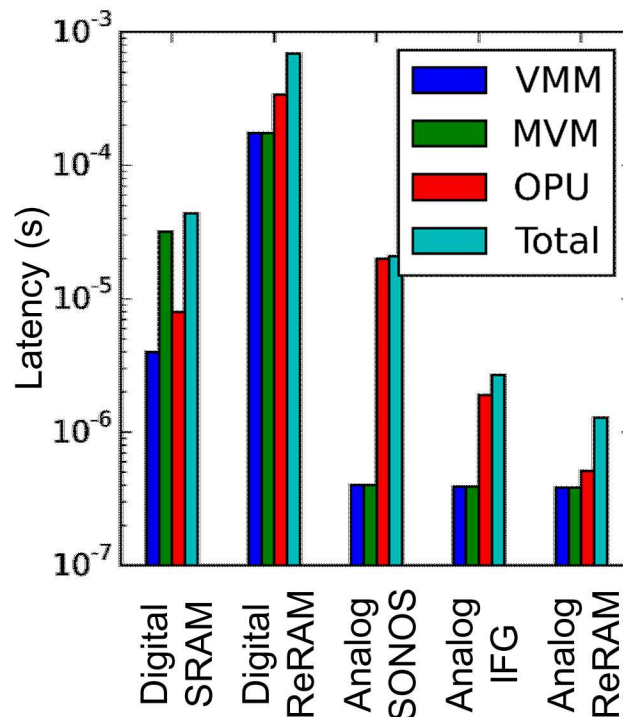
Ionic Floating-Gate

Compare Architectural Advantages

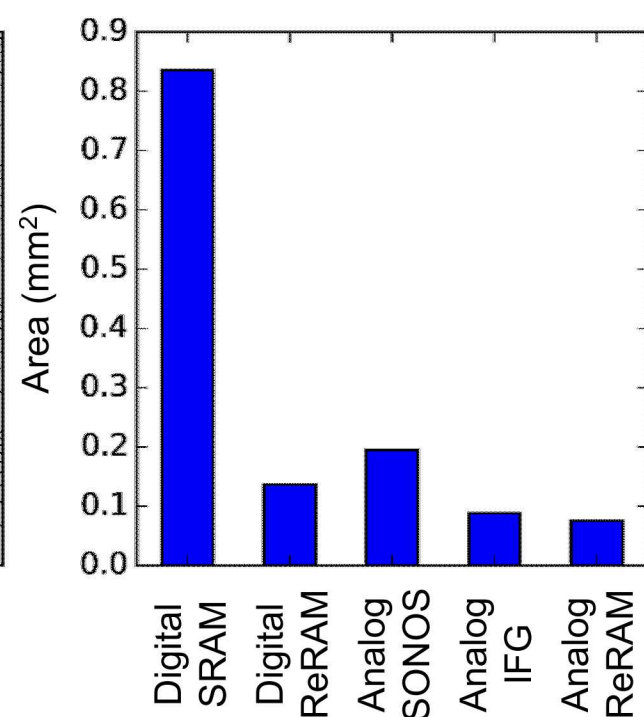
120-430X Energy Advantage



2-34X Latency Advantage



5-11X Area Advantage



1024 x 1024 = 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

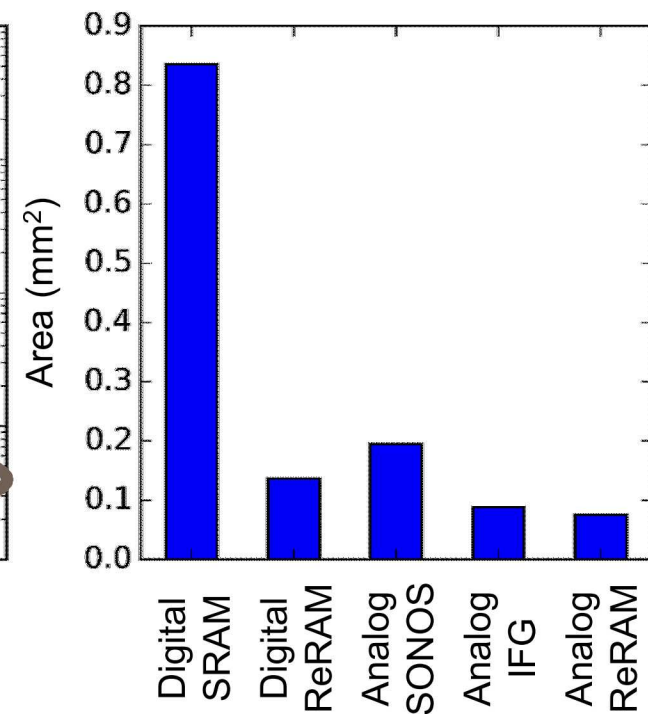
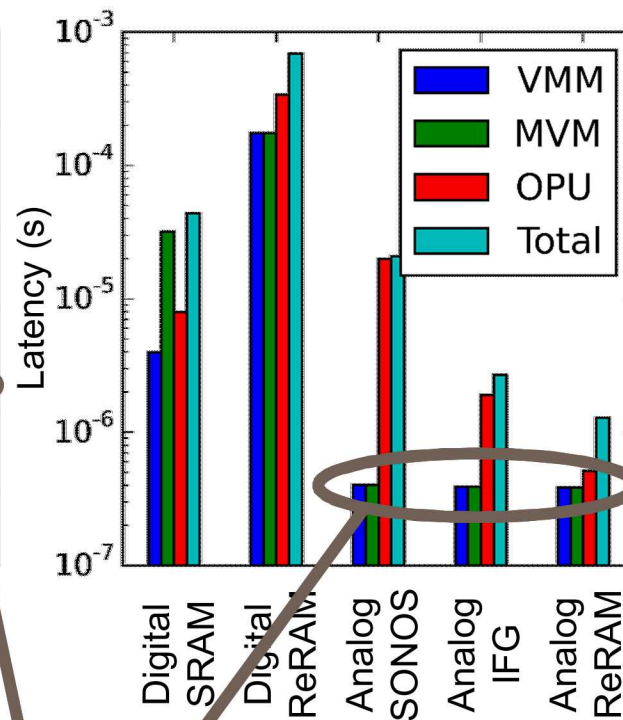
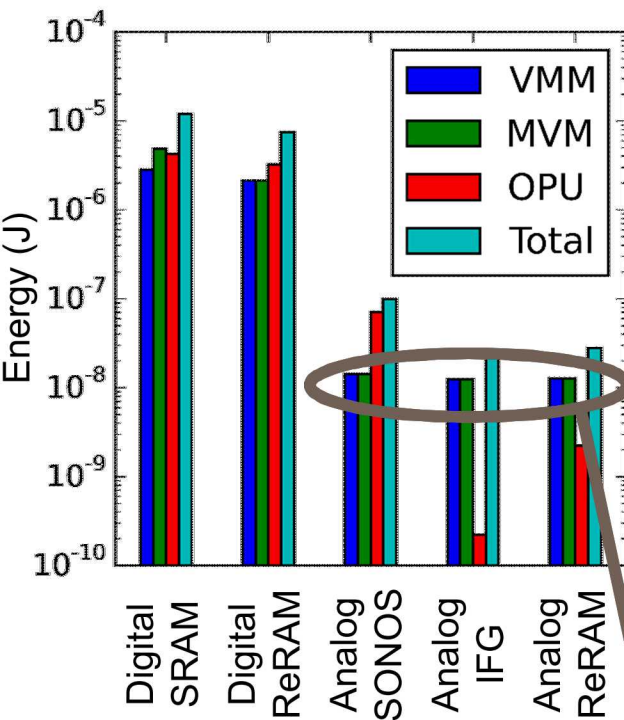
***Requires 100 M Ω on state devices

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

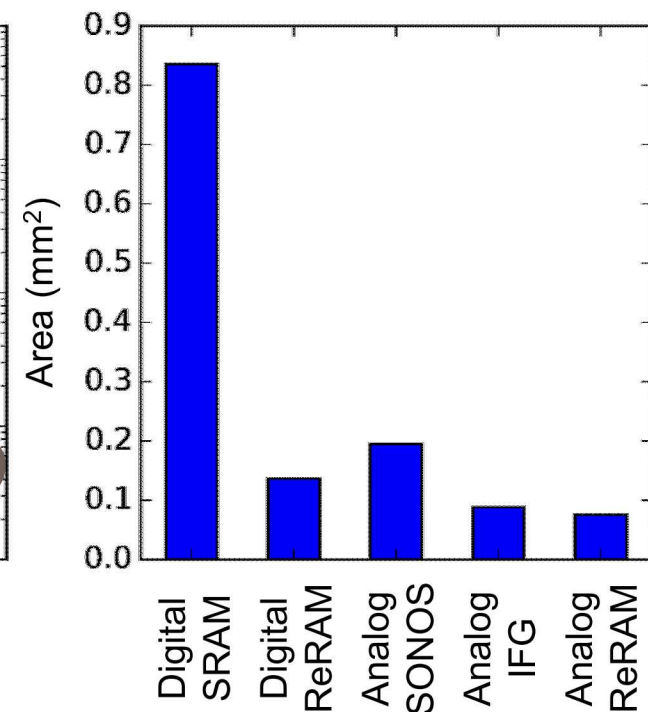
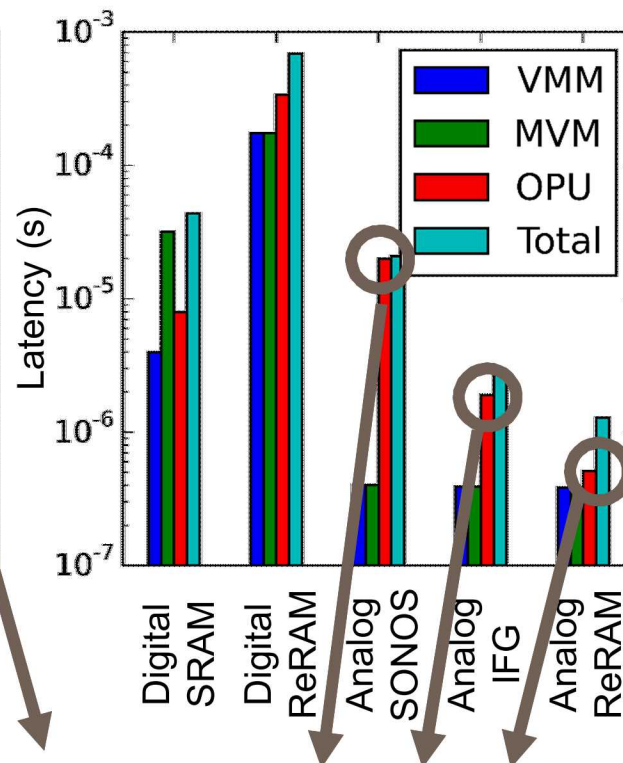
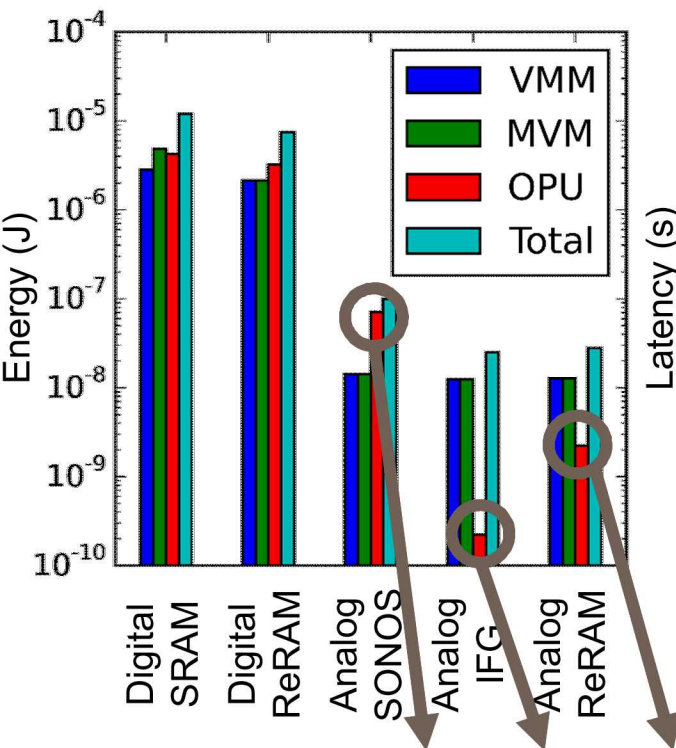
- Entirely dominated by ADC, device properties irrelevant

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

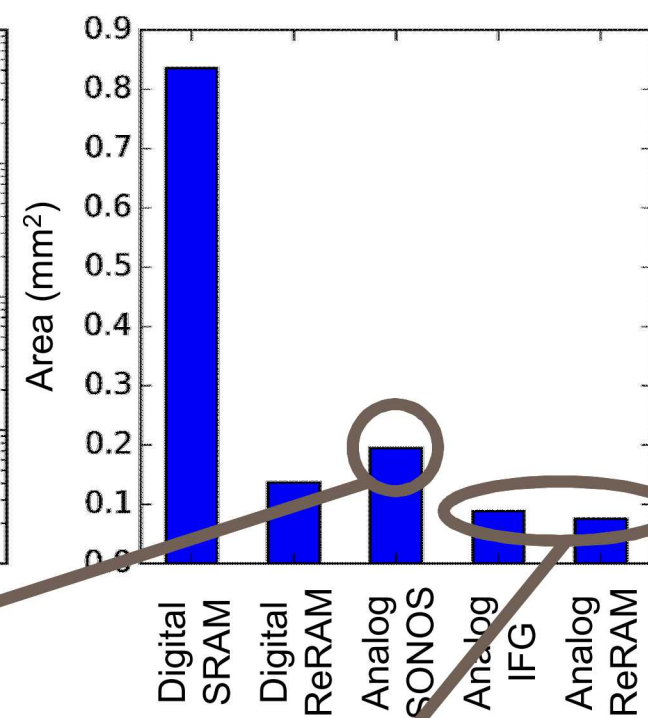
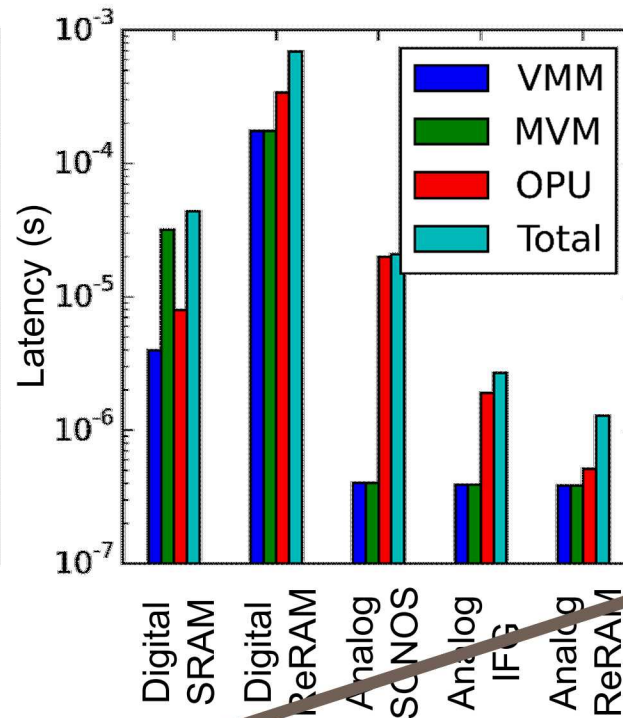
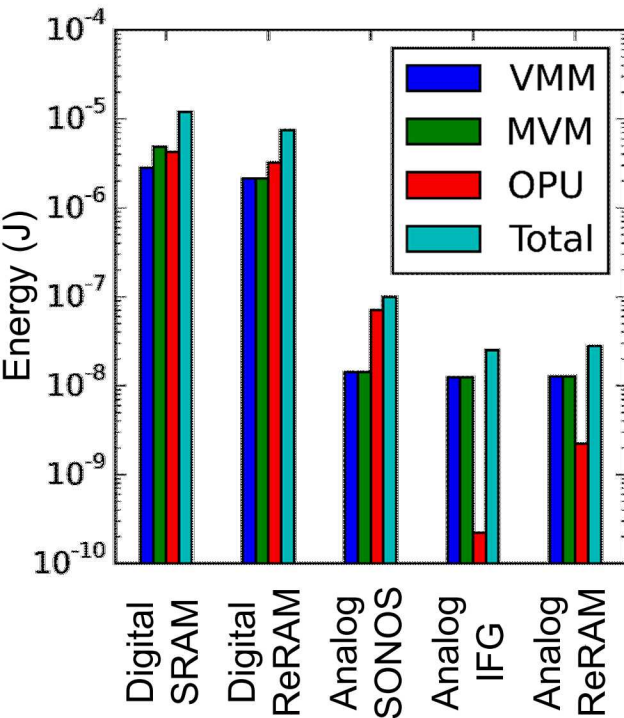
Compare Architectural Advantages:

Area

120-430X Energy Advantage

2-34X Latency Advantage

5-11X Area Advantage

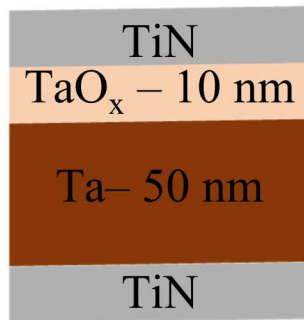


SONOS area cost reasonable, roughly doubles area

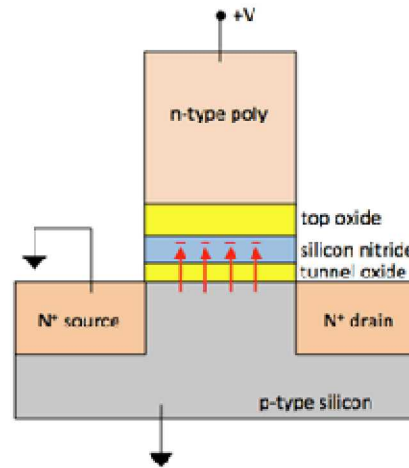
IFG and ReRAM go over transistors, area dominated by ADC and DAC

Analog Devices Summary for Training

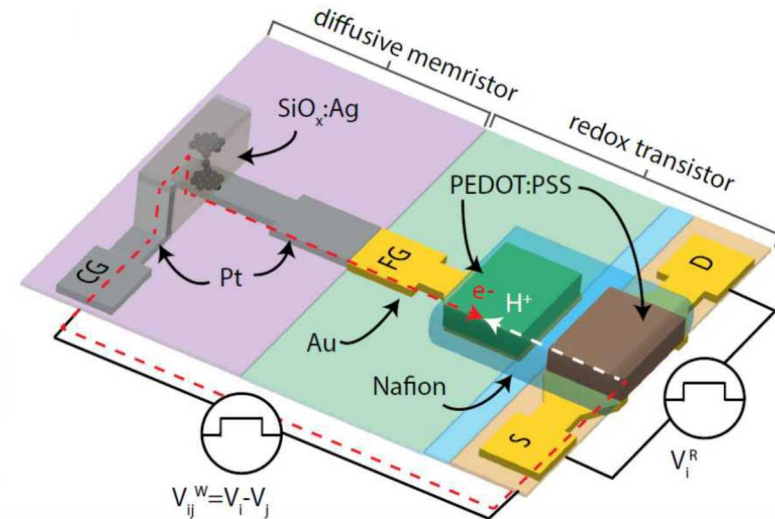
ReRAM



SONOS Silicon-Oxygen- Nitrogen-Oxygen-Silicon

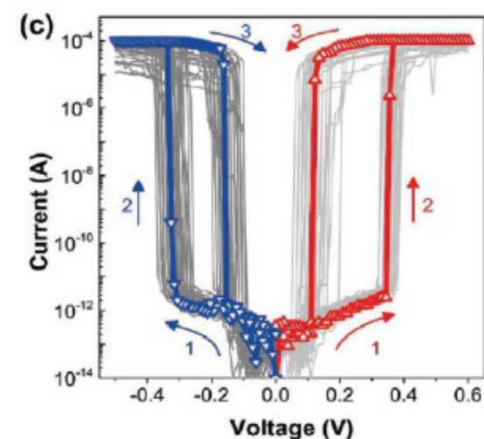
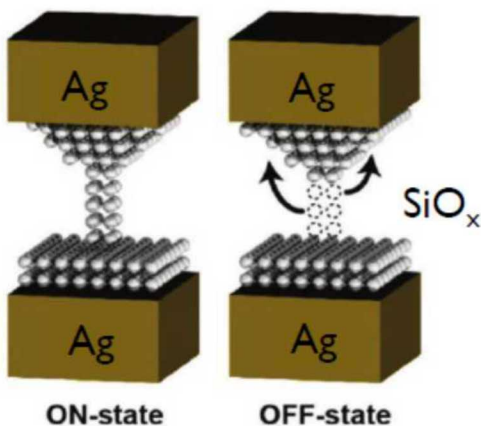
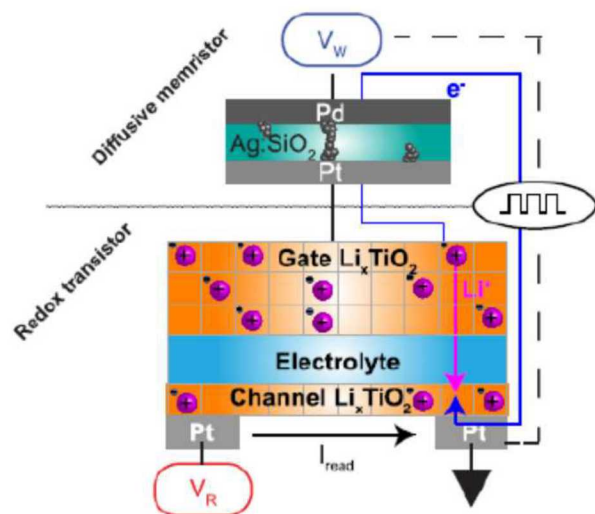


Ionic Floating-Gate Memory

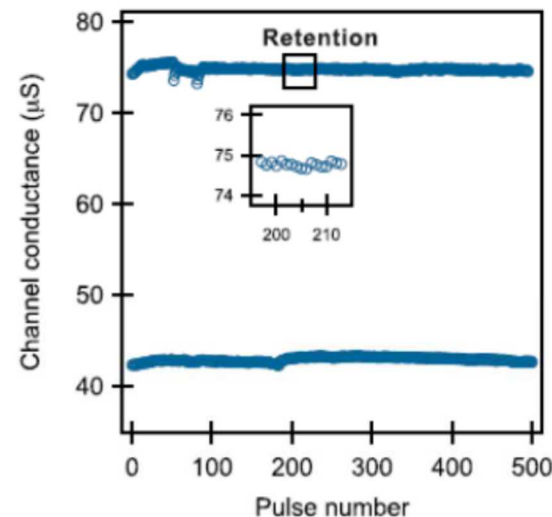
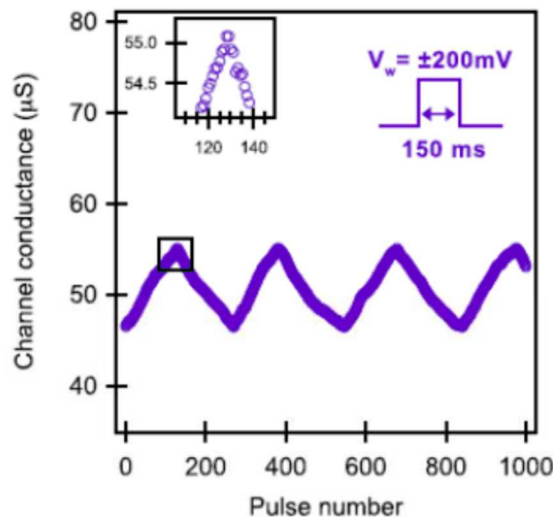
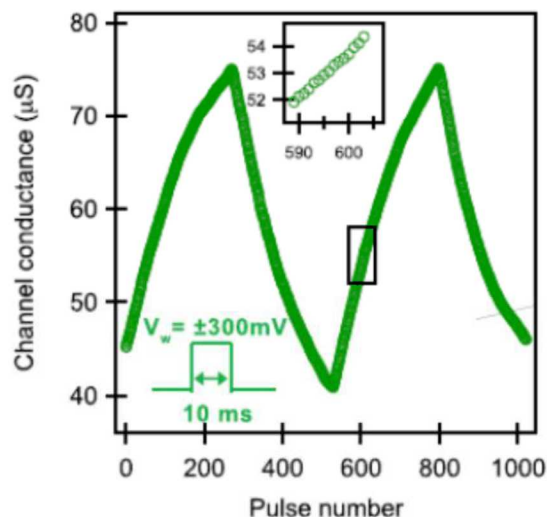


- | | | |
|---|---|--|
| <ul style="list-style-type: none"> • Large Energy/Area/Latency advantage over digital • Accuracy not good enough • Back end of line compatible • Under commercial development | <ul style="list-style-type: none"> • Moderate Energy/Area/Latency advantages over digital • High Accuracy • Commercially available • Need to prove endurance and device to device variability | <ul style="list-style-type: none"> • Large Energy/Area/Latency advantages over digital • High Accuracy • Not clear how to integrate • Has retention challenges |
|---|---|--|

Diffusive memristor for write-select, retention



Wang et. al. *Nature Materials* 16, 101-108 (2016)



Y. Li, E. J. Fuller, S. Asapu, S. Agarwal, T. Kurita, J. J. Yang, A. A. Talin, *ACS App Mat Int.*, DOI: 10.1021/acsami.9b14338

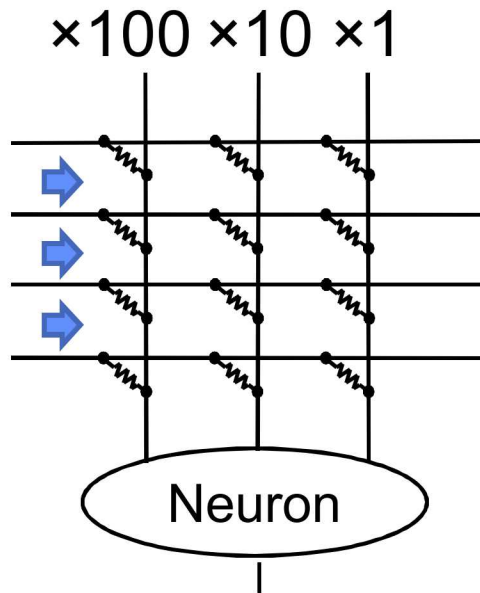
(Slide from A. Alec Talin)

Center 8300: Chemistry, Combustion and Materials Science

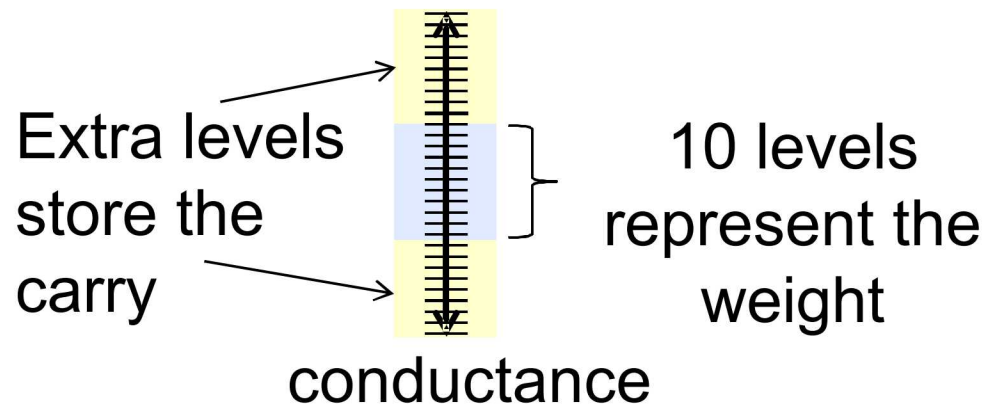
Multi-ReRAM Synapse: Periodic Carry

If we need more bits per synapse, use multiple memristors

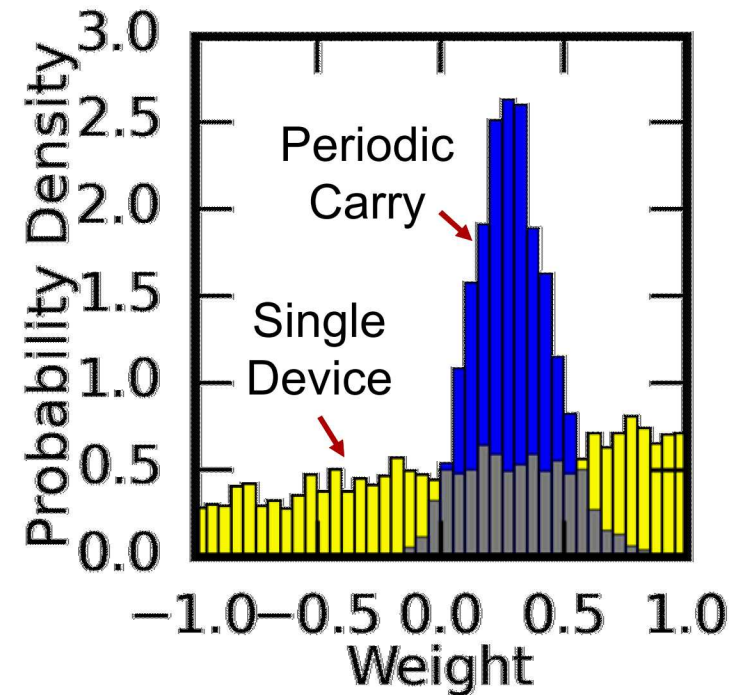
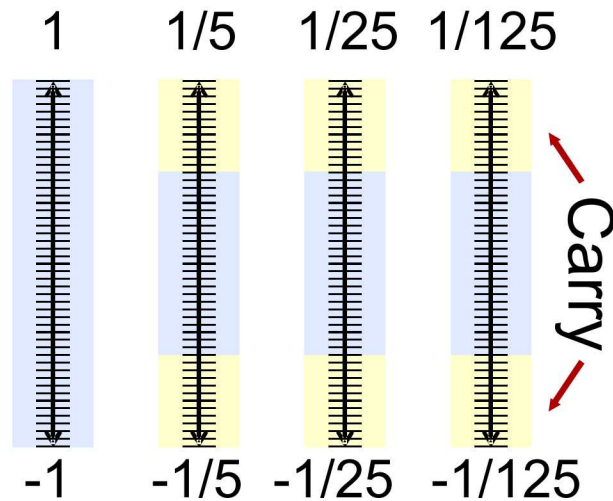
- Three 10 level ReRAMs could represent 1-1000!
- Adding to the weight requires reading every ReRAM to account for any carries and serially programming each ReRAM: VERY EXPENSIVE



- 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

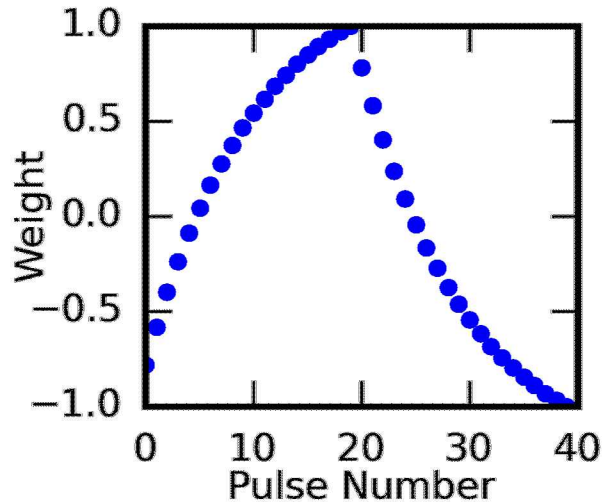
Noise Sigma = 1.4% for single device

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

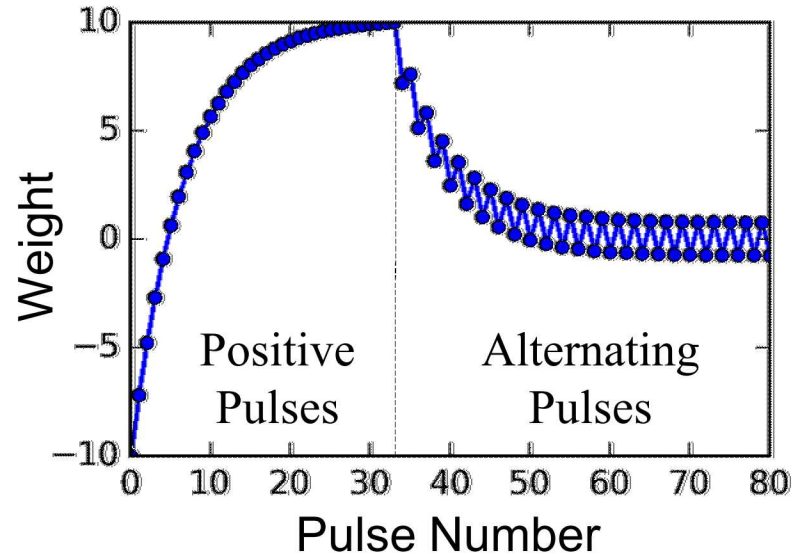
Learn from a 0.5% Signal

Periodic Carry Mitigates Write Nonlinearity

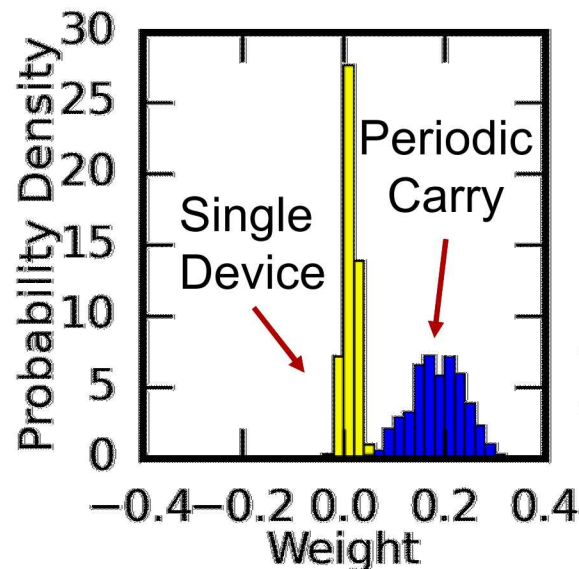
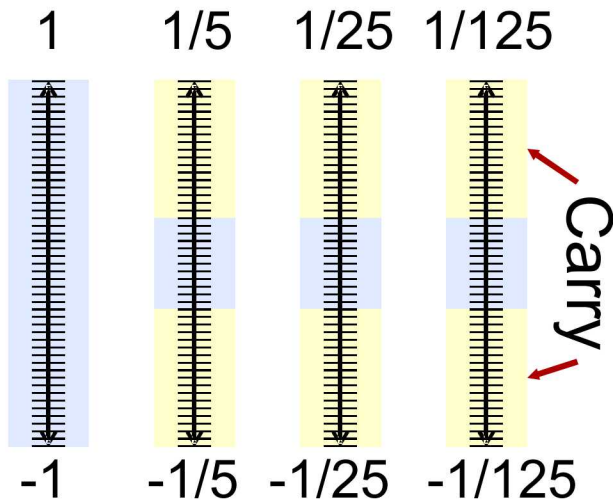
Write Nonlinearity



Alternating Pulses Cause Weight Decay

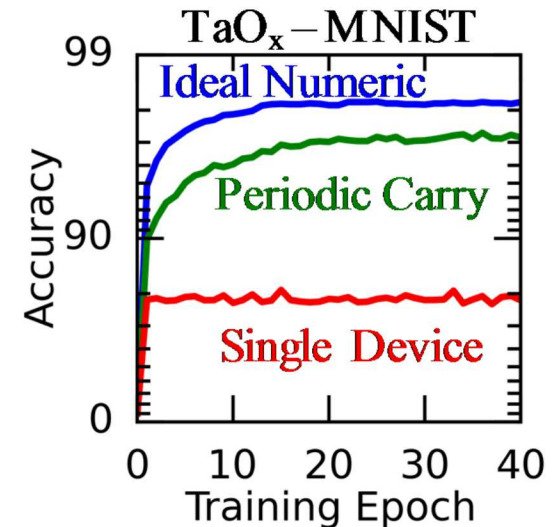
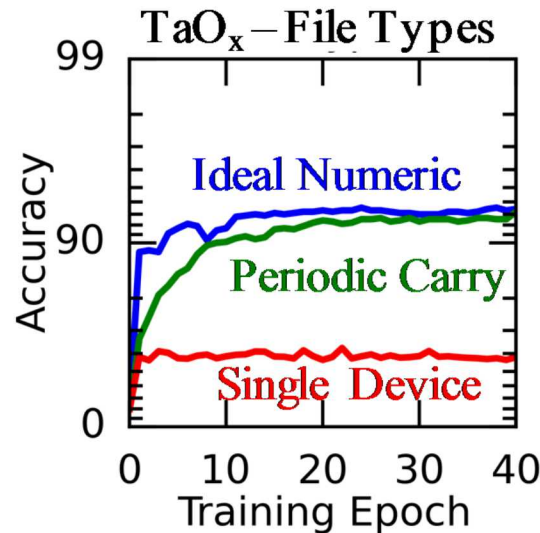
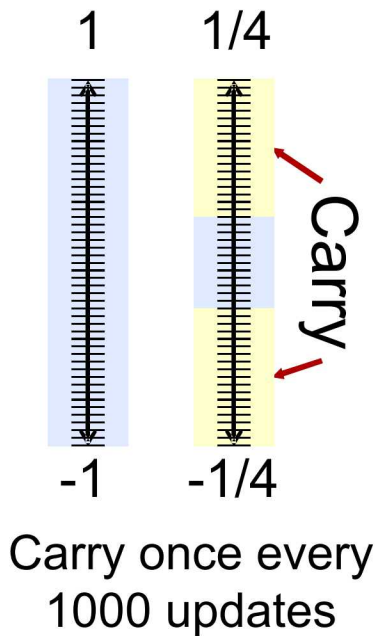


Use center linear range of weights



- Train with 1% signal
- Ideal result is 0.6

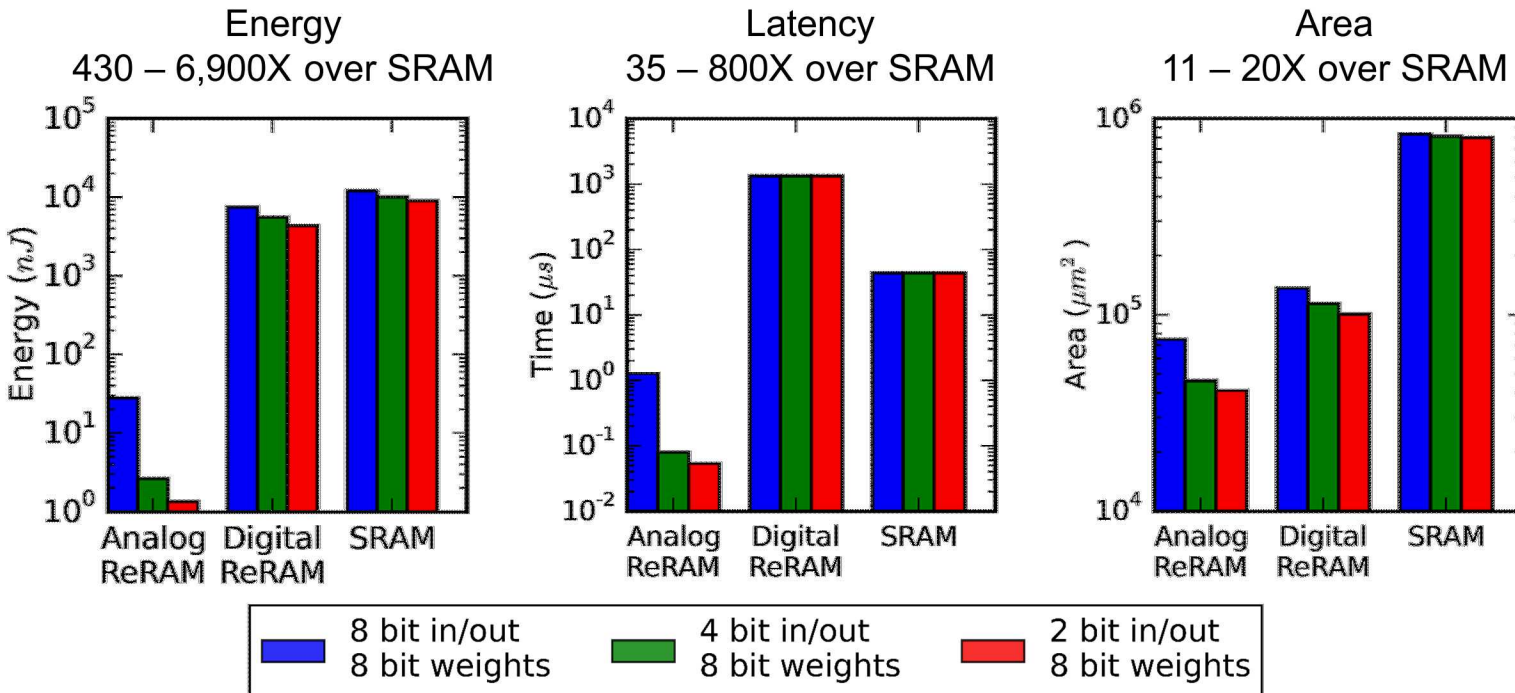
TaO_x Results



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 M Ω Devices
- Need low write nonlinearities

CROSS SIM

<https://cross-sim.sandia.gov>

Sandia is Hiring! sandia.gov/careers



~12,500 people in
Albuquerque, NM



~1,800 people in
Livermore, CA

Fulfilling Our National Security Mission



Nuclear Deterrence



*Defense Nuclear
Nonproliferation*



*National Security
Programs*



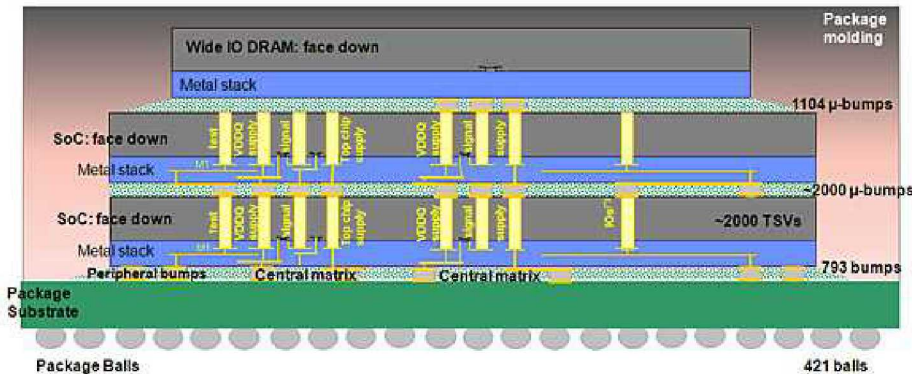
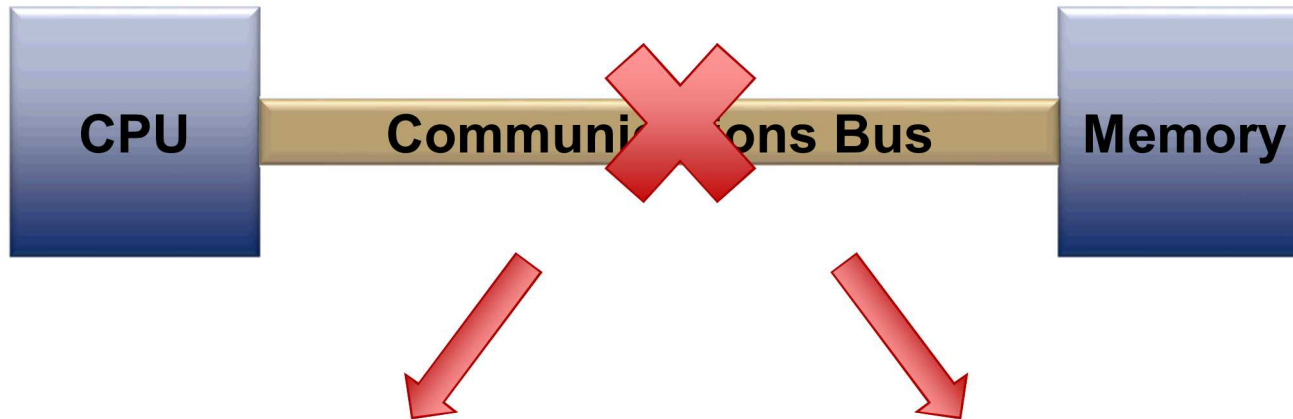
*Energy & Homeland
Security*



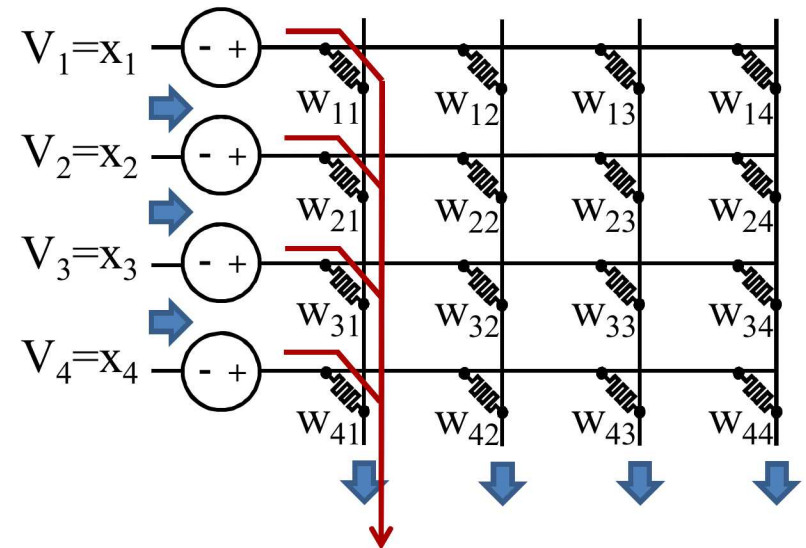
*Advanced Science &
Technology*

Extra Slides

Overcoming the Power Limit



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



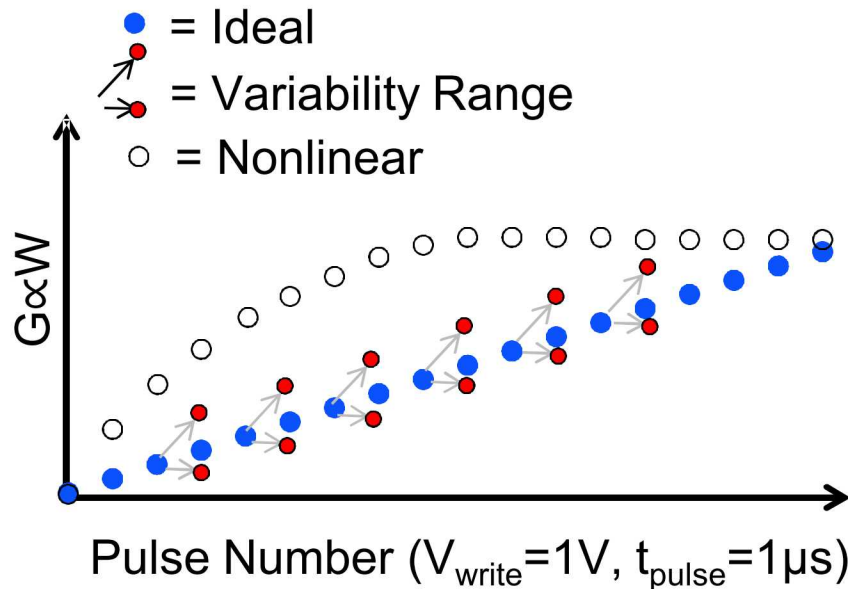
Integrate Processing and Memory

Experimental Device Non-idealities

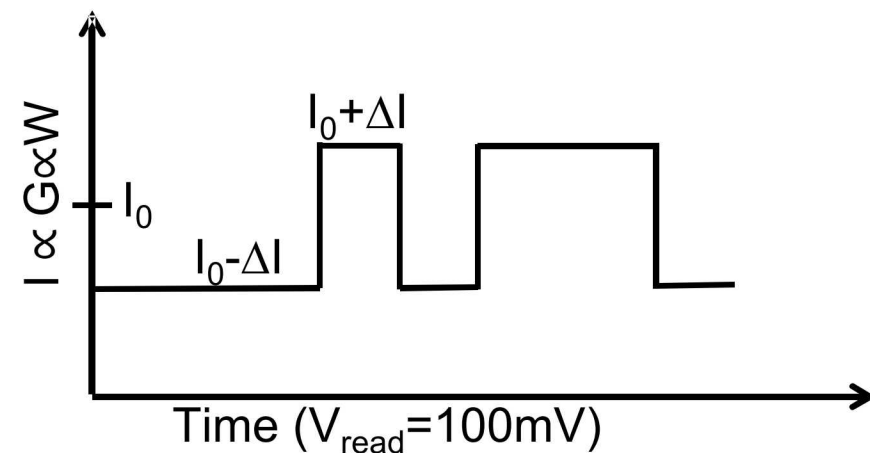
Device: **Write Variability, Write Nonlinearity, Asymmetry, Read Noise**

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

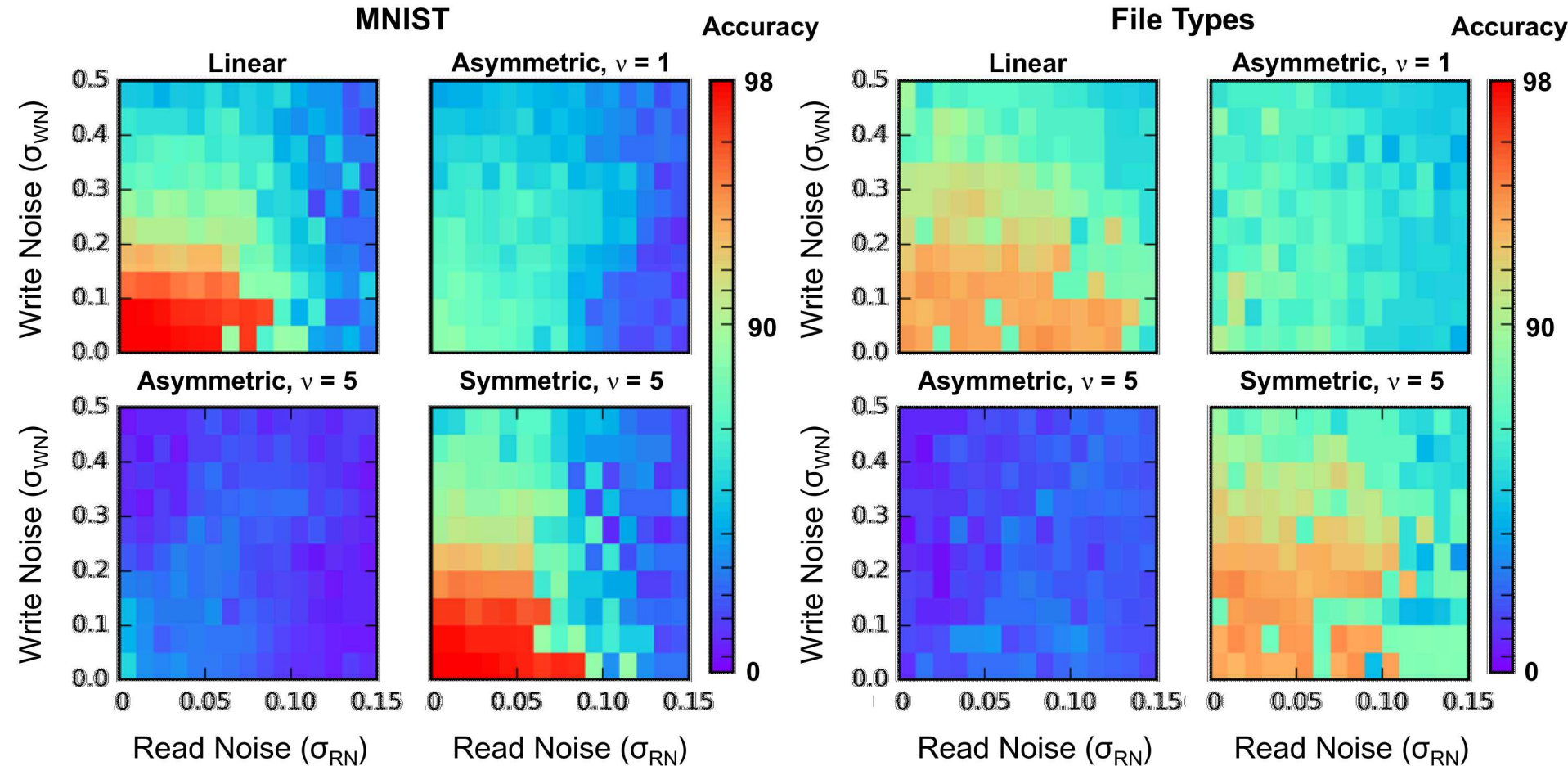
Variability and Nonlinearity



Read Noise

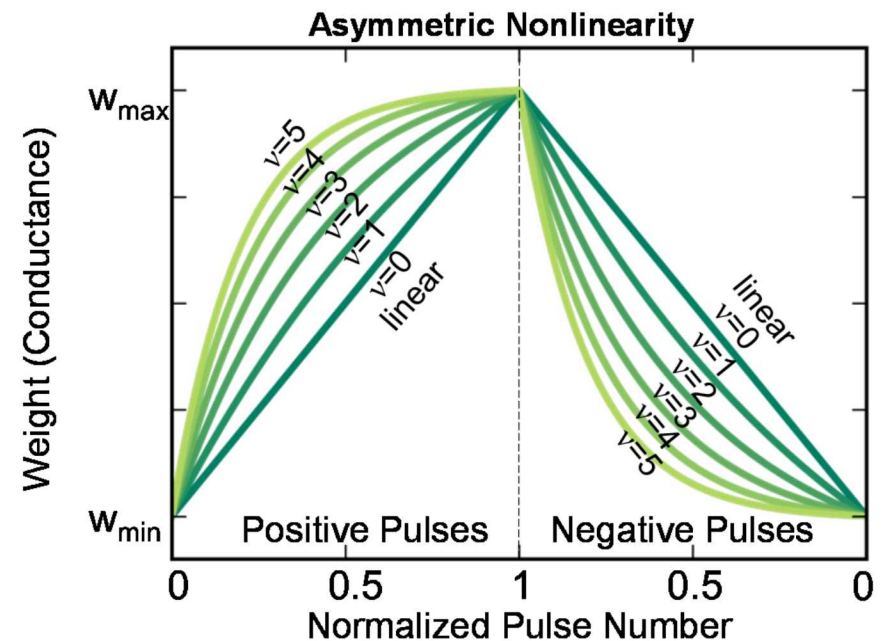
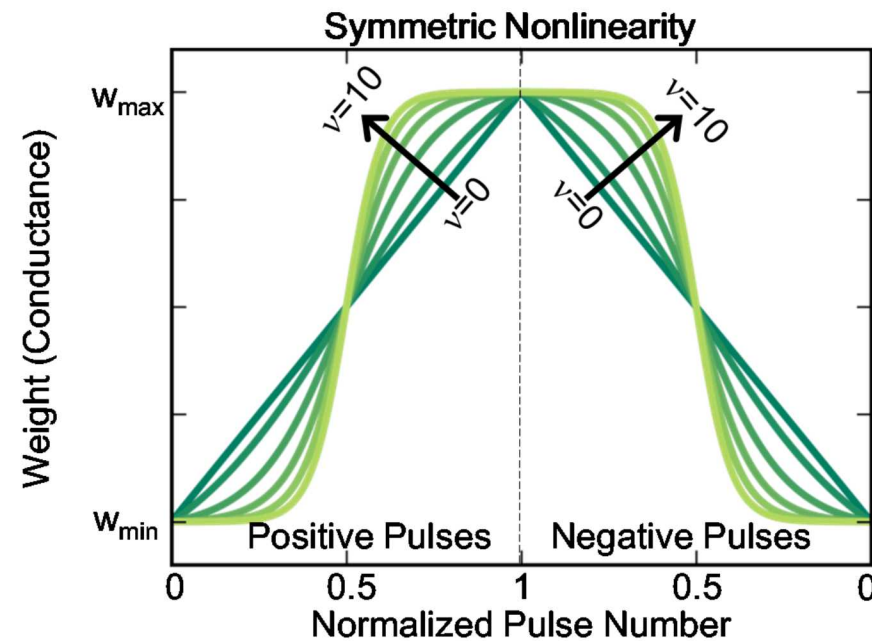


Combined Effects of Nonidealities

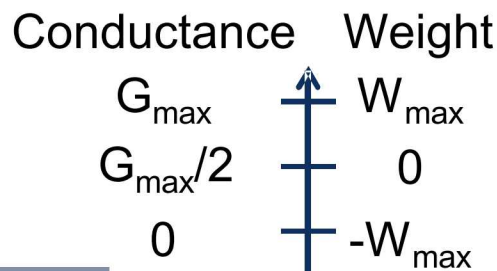
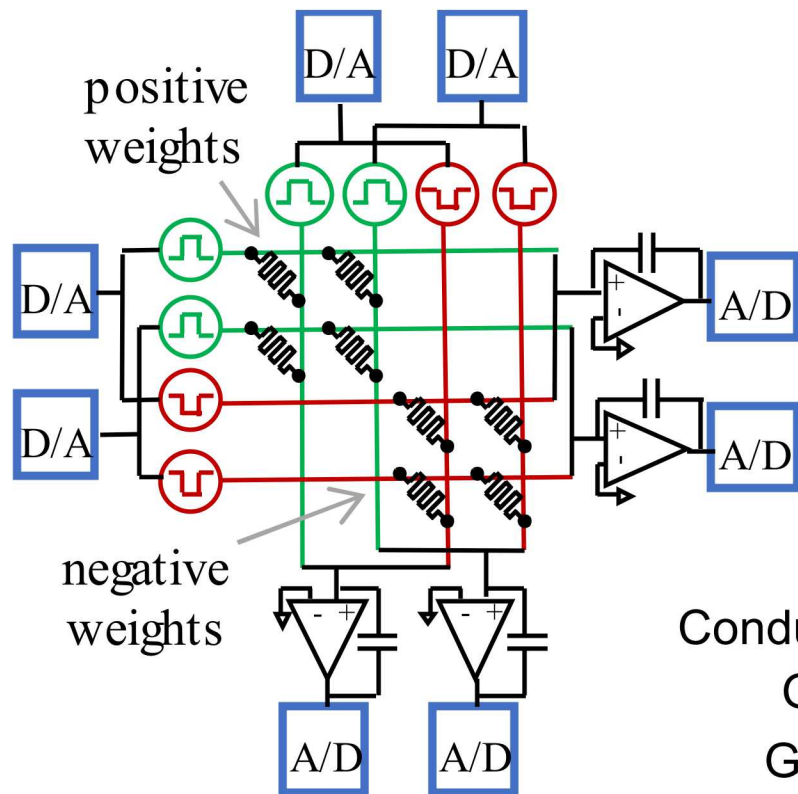


What are the Neural ReRAM Device Requirements?

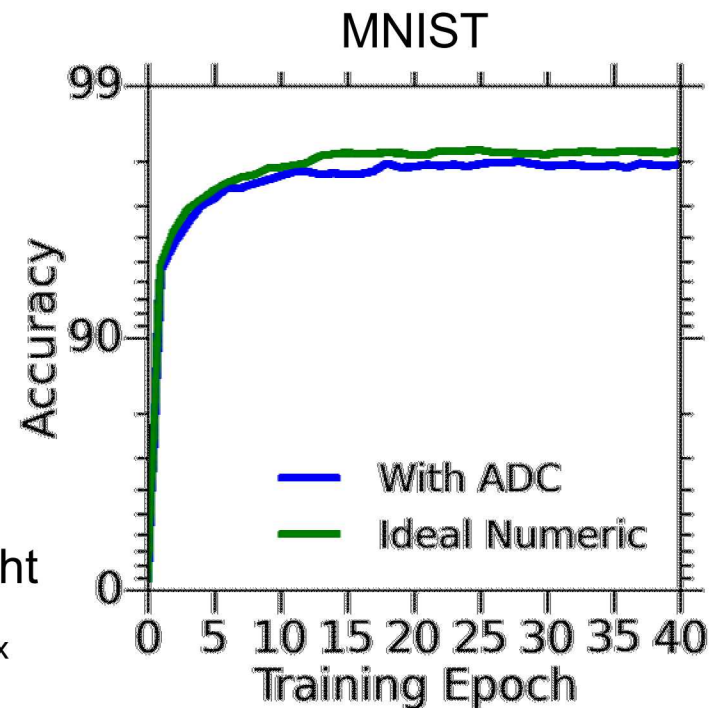
	Small Images	Large Images	File Types
Read Noise σ (% Range)	3%	5%	9%
Write Noise σ (% Range)	0.3%	0.4%	0.4%
Asymmetric Nonlinearity (v)	0.1	0.1	0.1
Symmetric Nonlinearity (v)	>20	5	5
Maximum Current	160 nA	13 nA	40 nA



Full System Simulation



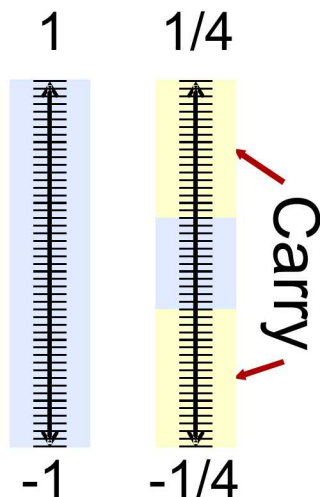
A/D & D/A Have Minimal Impact



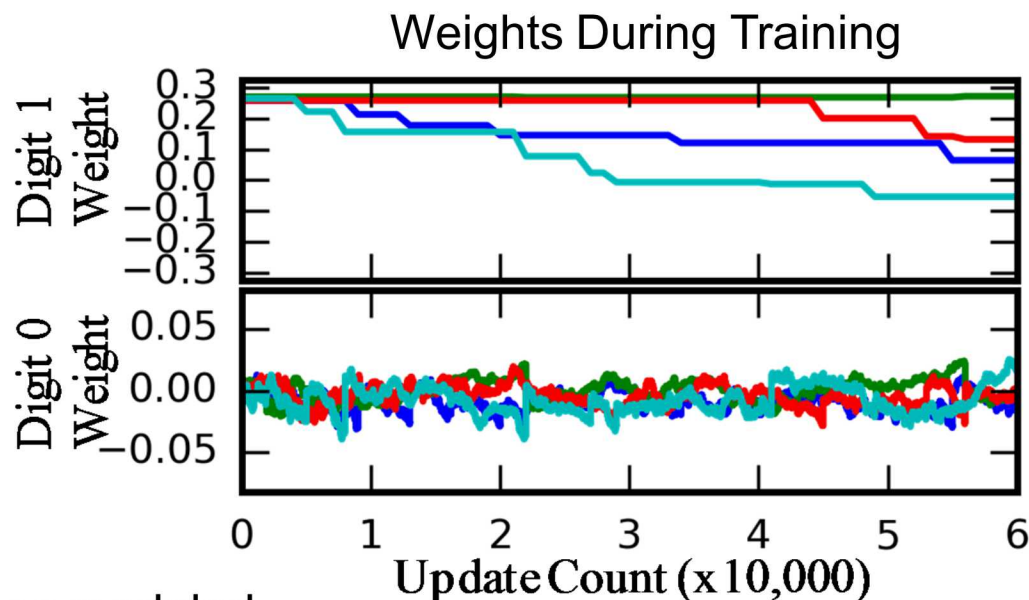
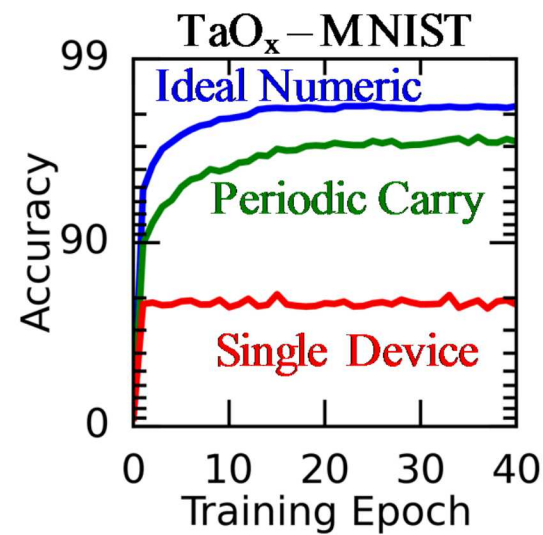
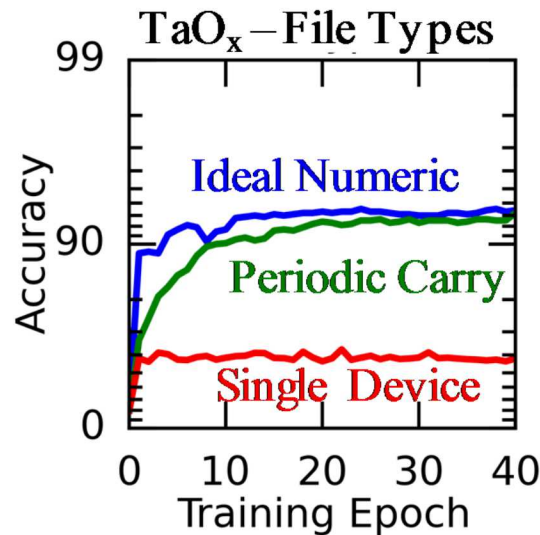
	Range	Bits
Row Input	-1 to 1	8
Col Output	-6 to 6	8
Col Input	-1 to 1	8
Row Output	-4 to 4	8
Row Update	-0.01 to 0.01	7
Col Update	-1 to 1	5

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



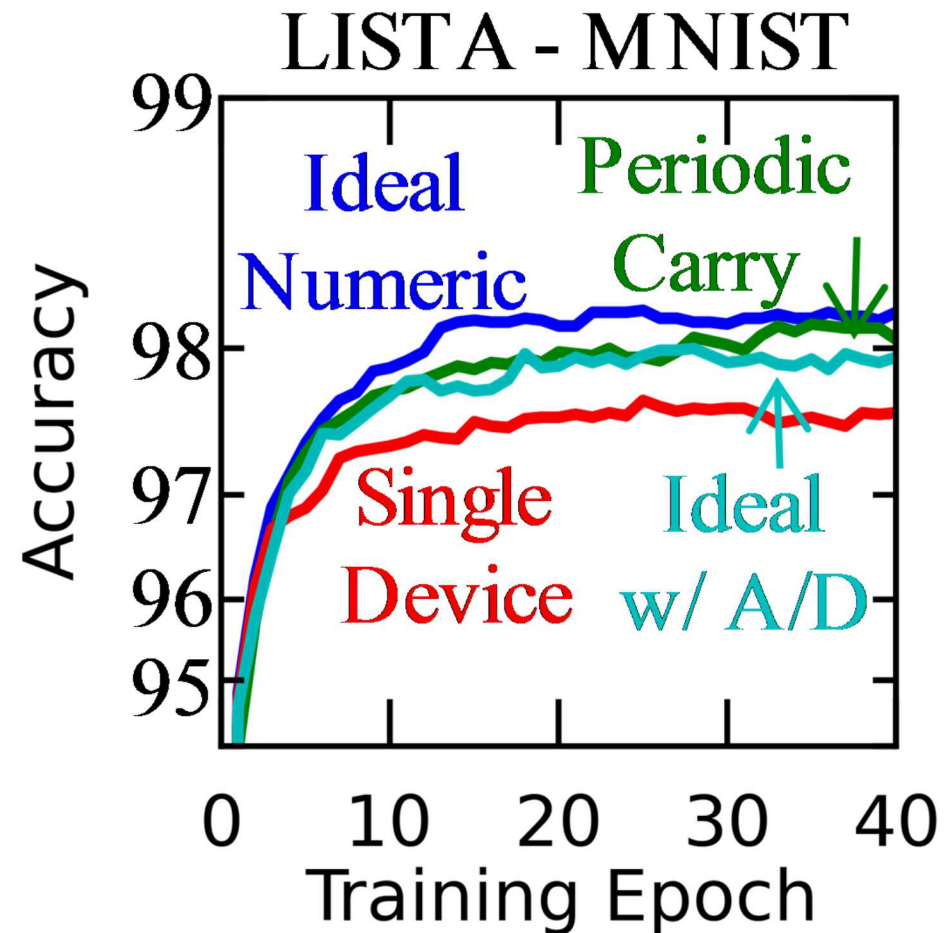
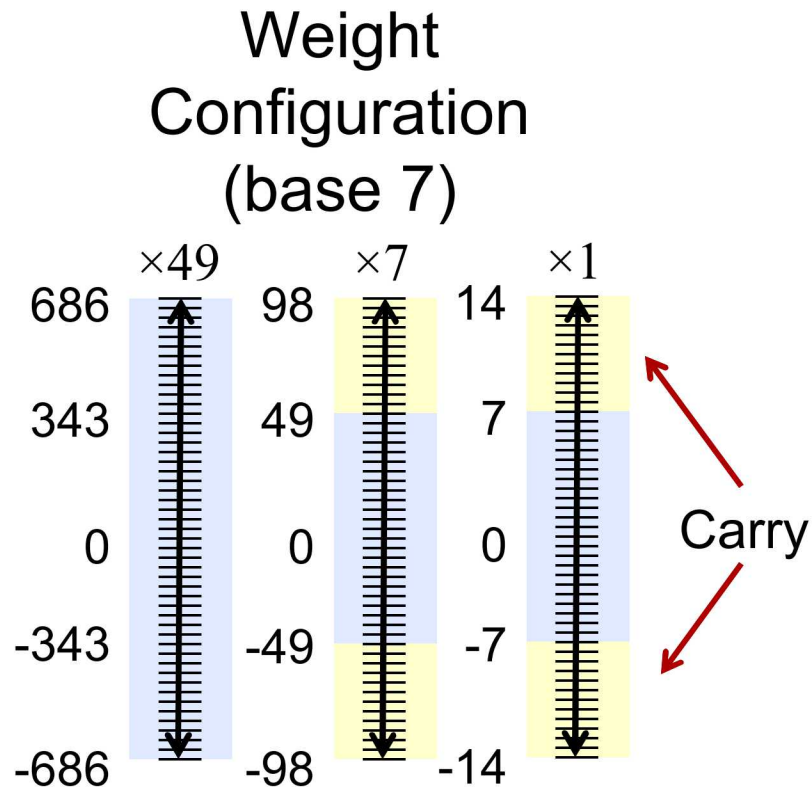
Carry once every 1000 updates for the LSB, and every 2 updates on others



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)

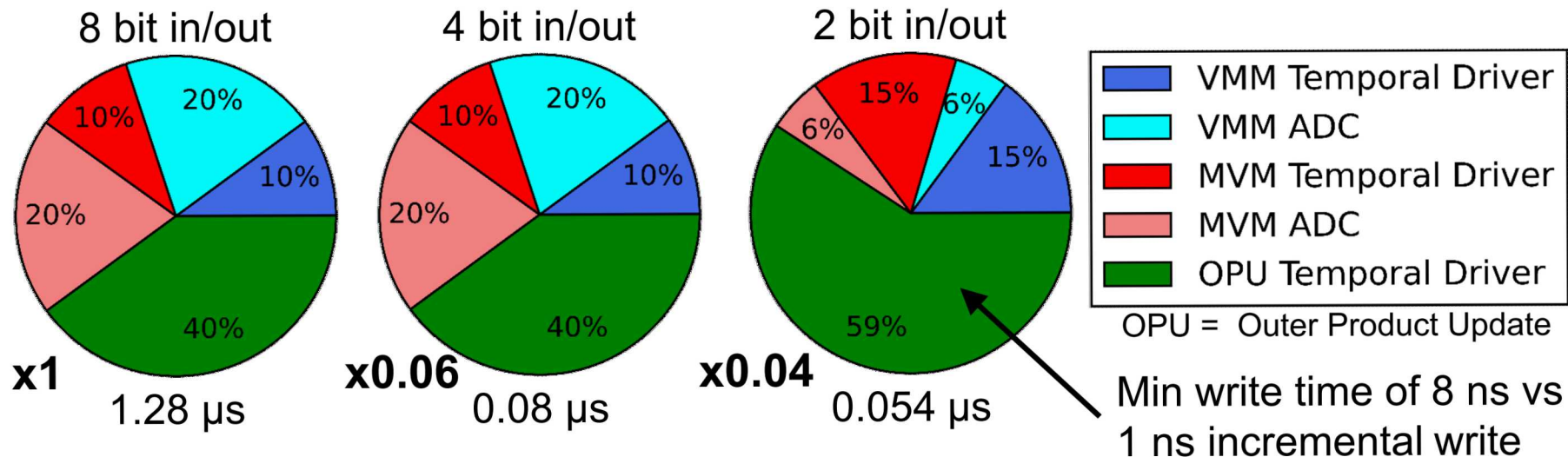
LISTA Results



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

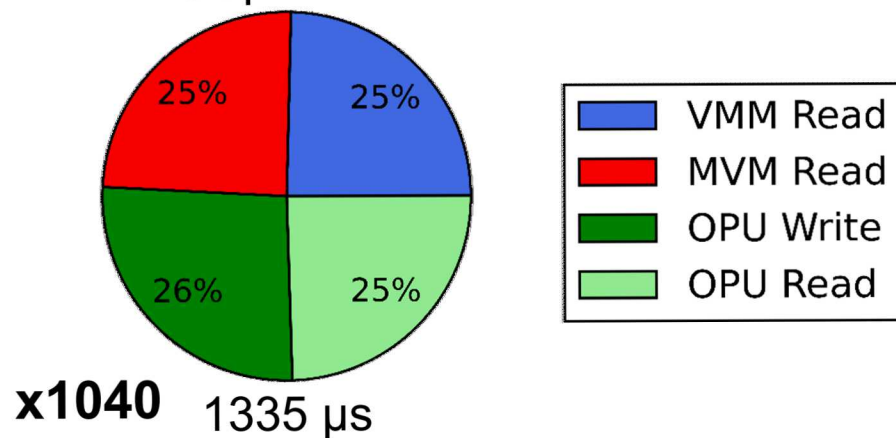
Neural Core Latency Analysis

Analog ReRAM



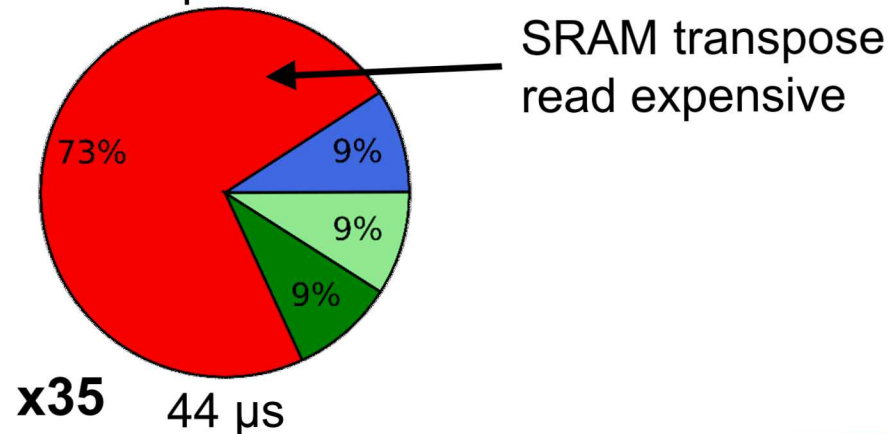
Digital ReRAM

All bit precisions

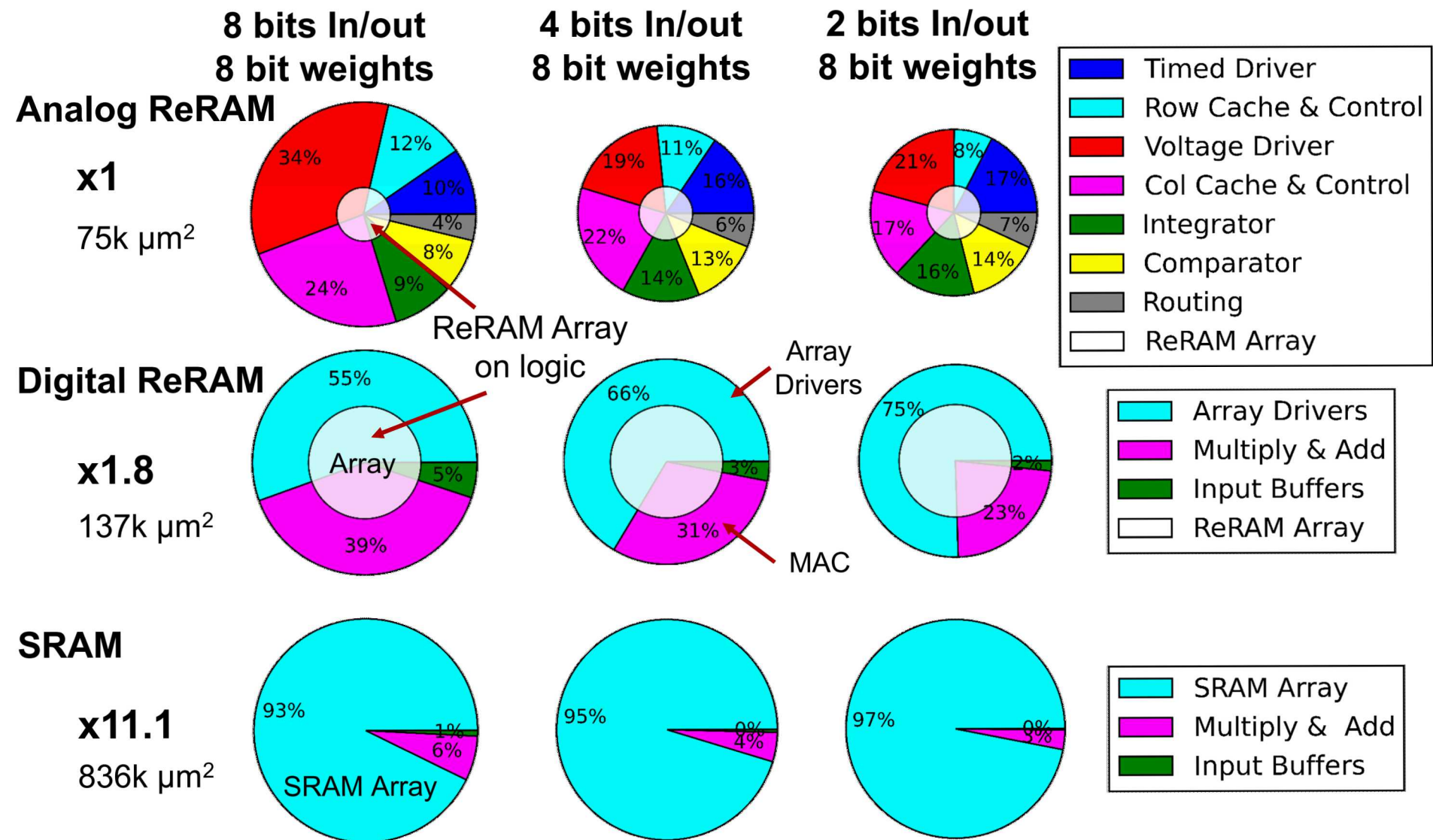


SRAM

All bit precisions

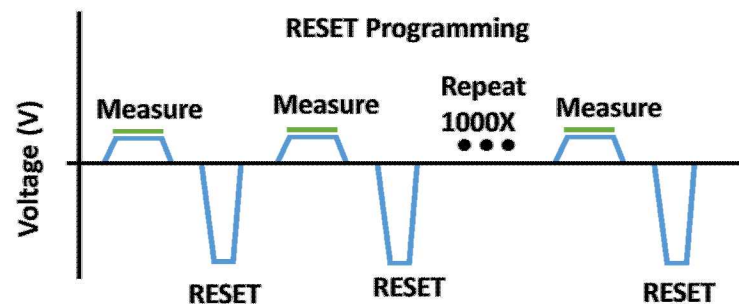
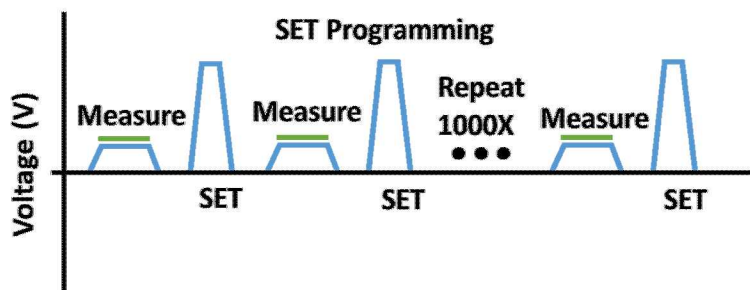
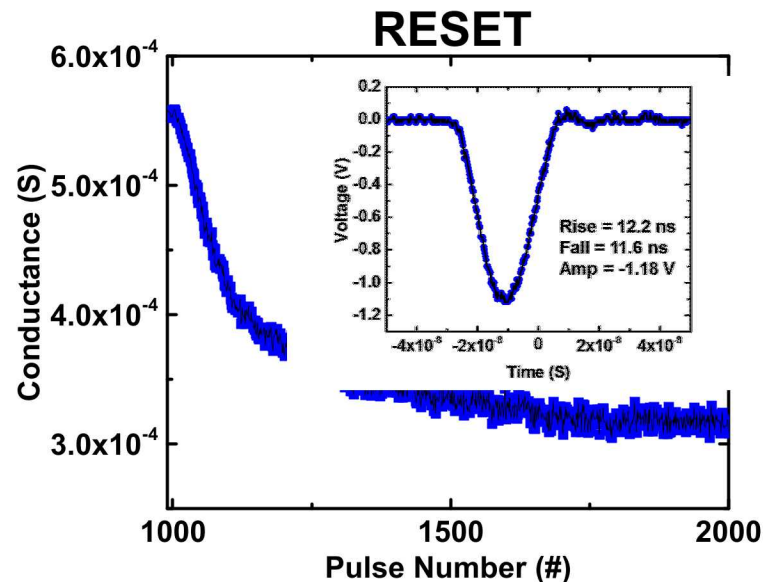
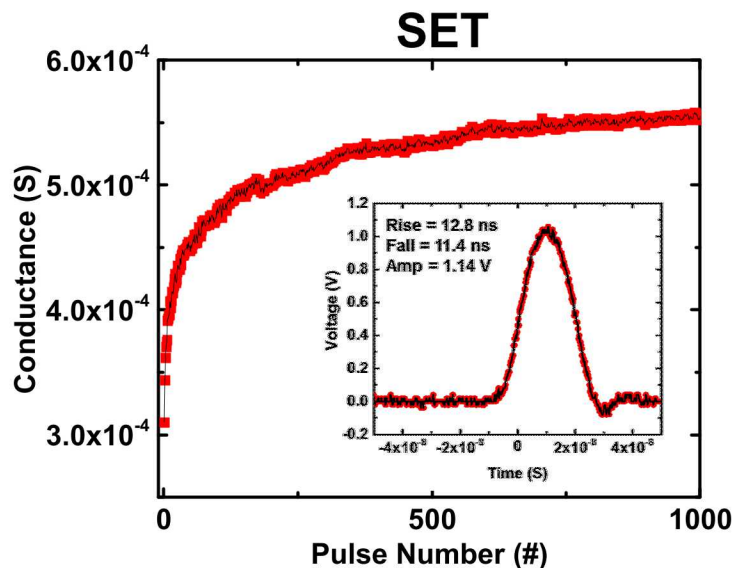


Neural Core Area Analysis

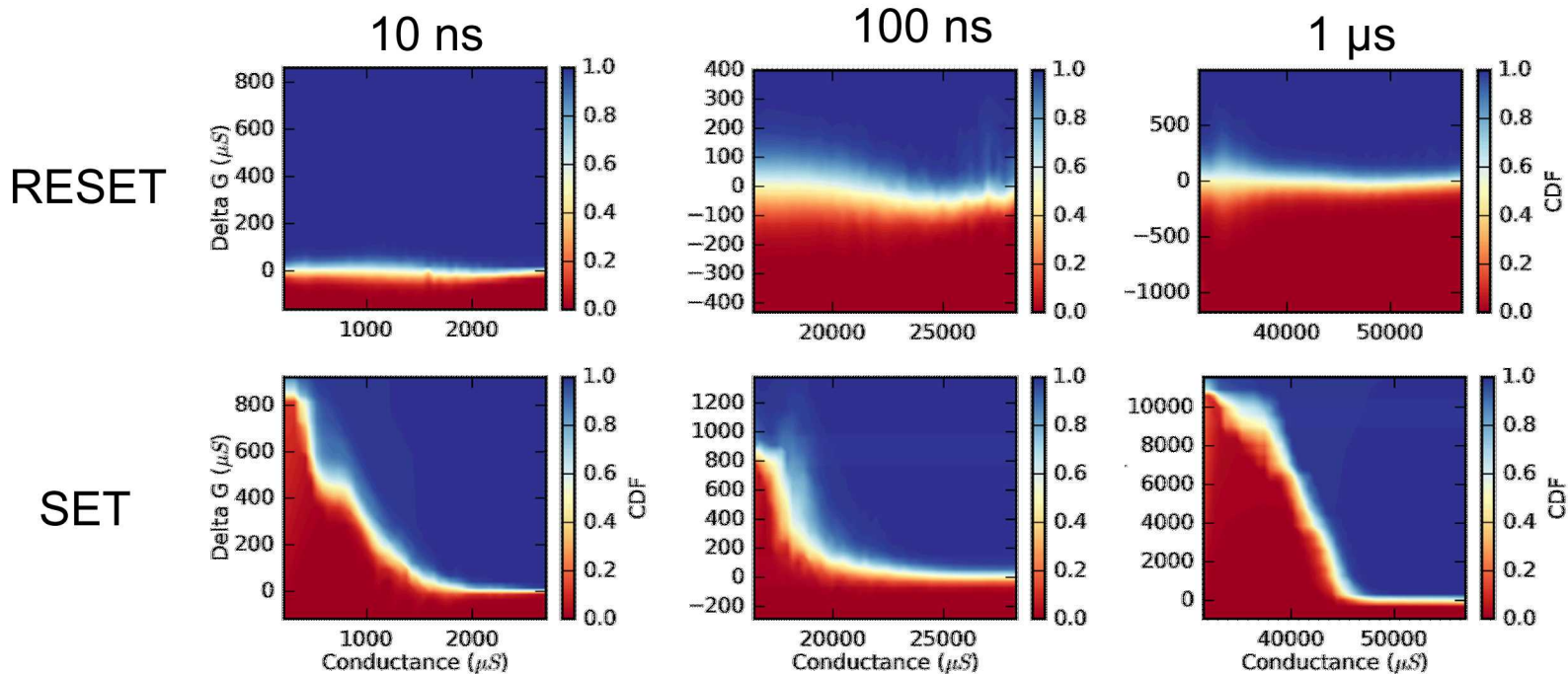
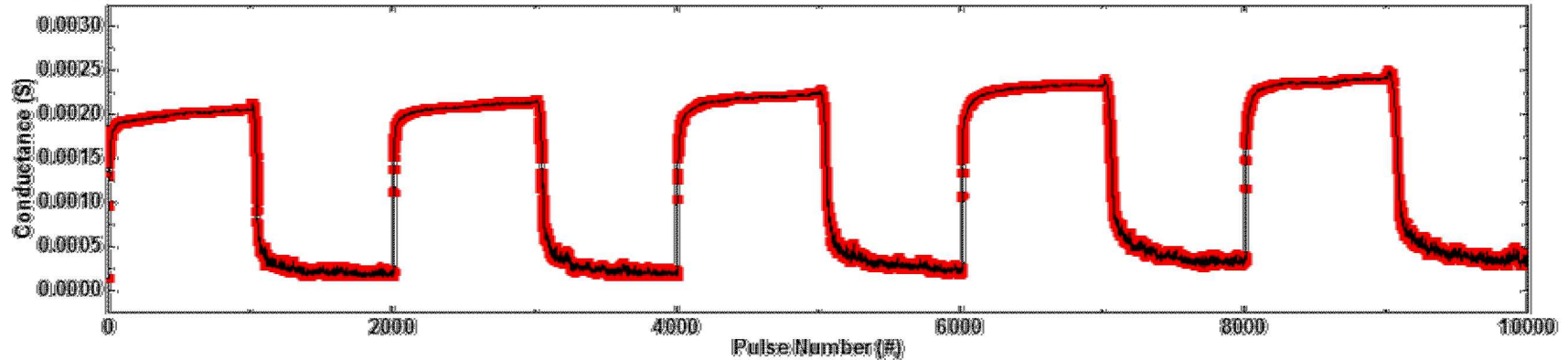


For the ReRAM, high voltage transistors require 8X area, improving this could give ~2X area savings

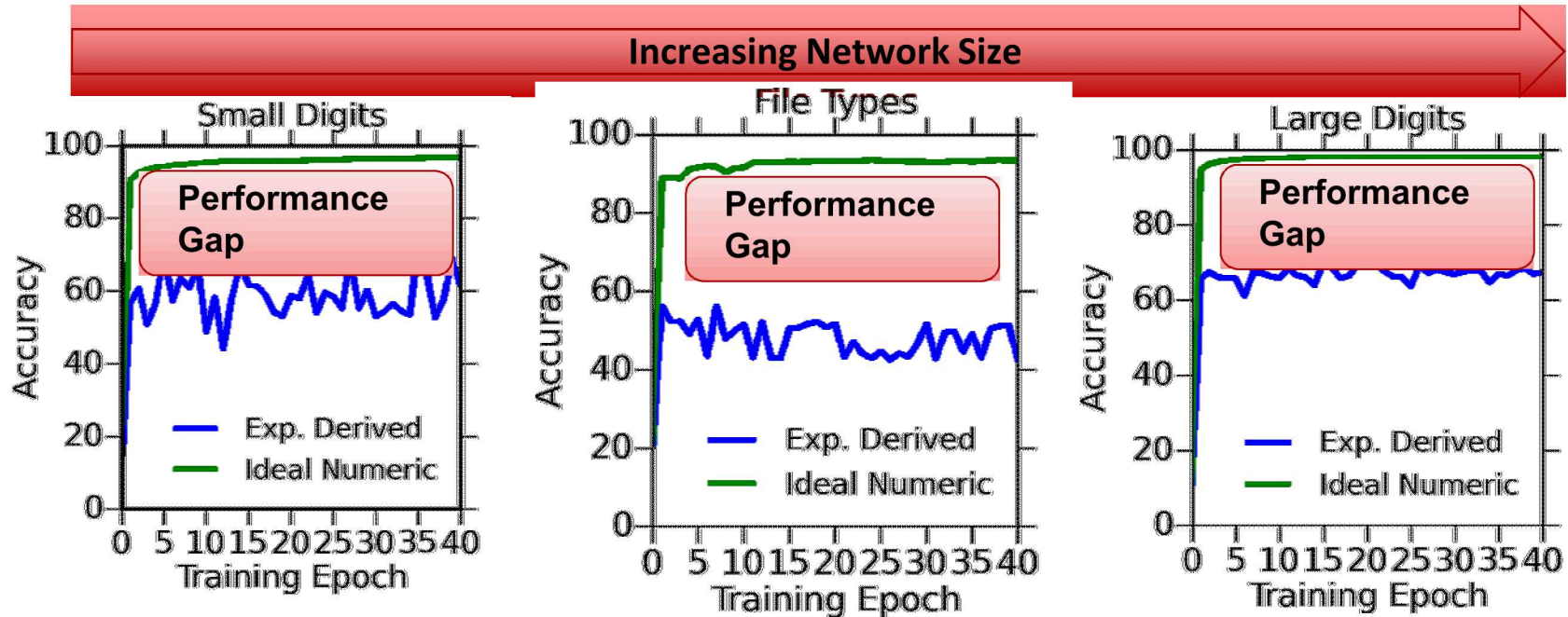
ReRAM Analog Characterization



Modeling Pulsed Cycling Characteristics

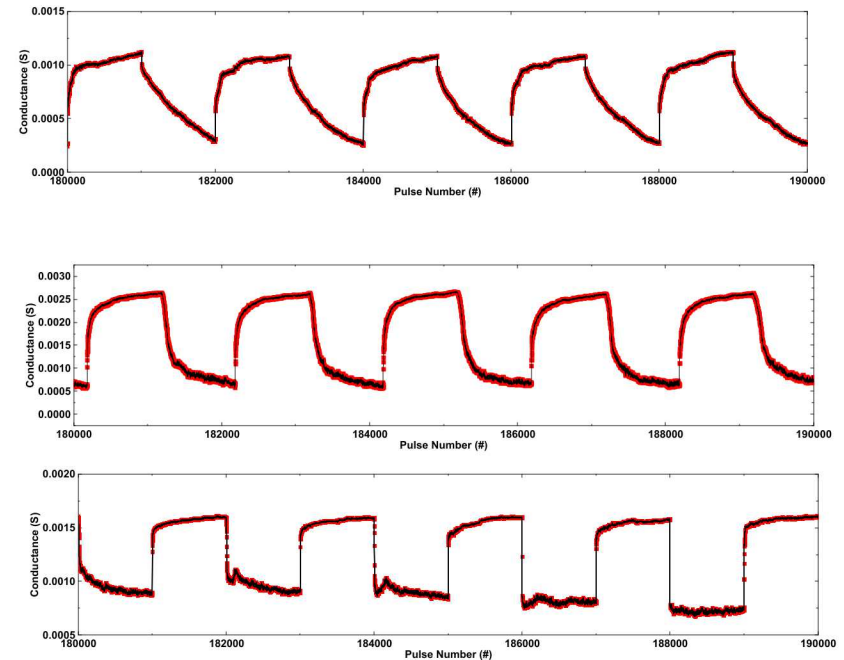
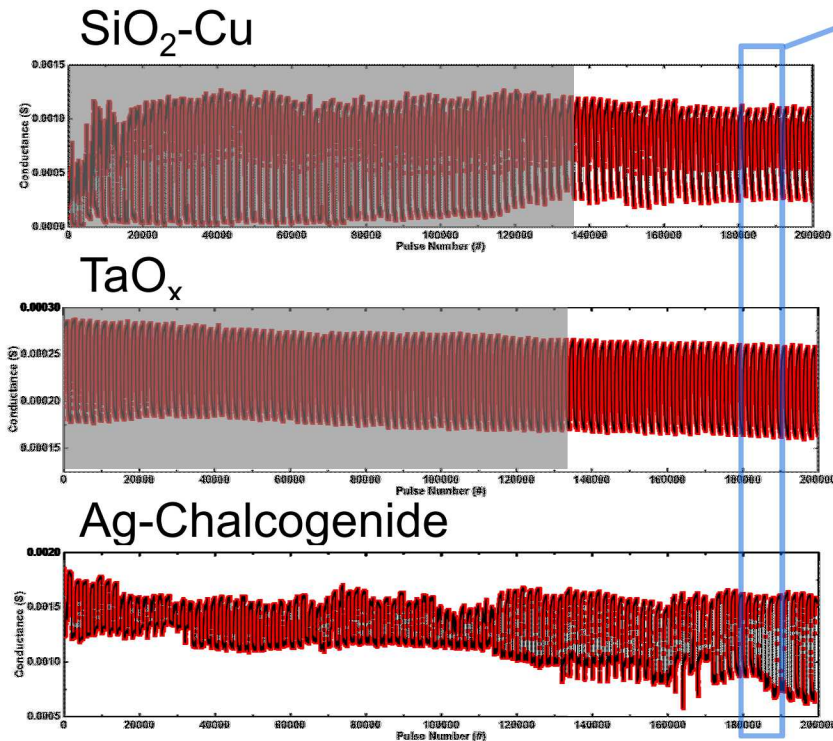


TaOx ReRAM in Backprop Training



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

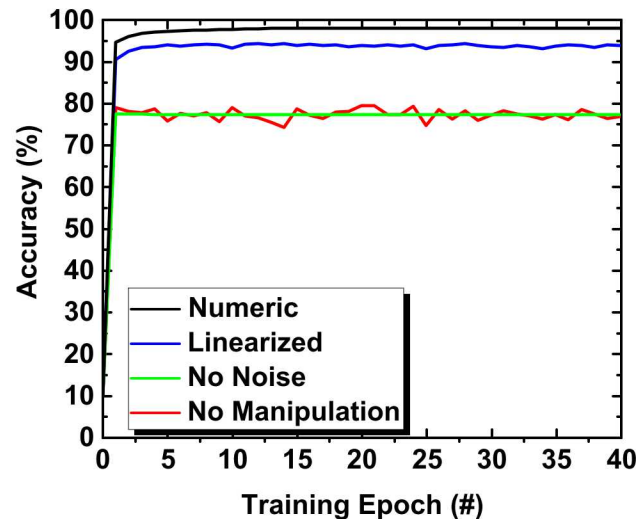


- For comparison all devices at 100 ns due to impedance limitation
- Device operation voltages found by increasing amplitude by 0.1 V until switching occurred – must survive 200,000 nudges so lowest possible voltage used
- Chalcogenide SET = +0.8 V RESET = -0.8V
- SiO₂-Cu SET = +1.4 V RESET = -1.6 V
- TaO_x SET +1.0 V RESET = -1.0 V

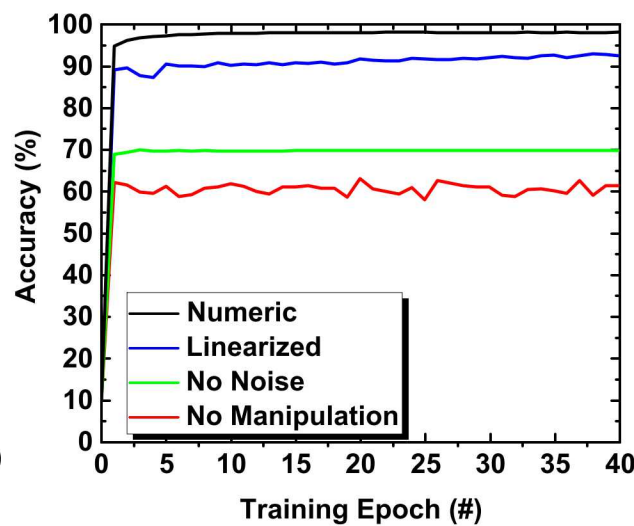
Comparison of Filamentary ReRAM

- 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

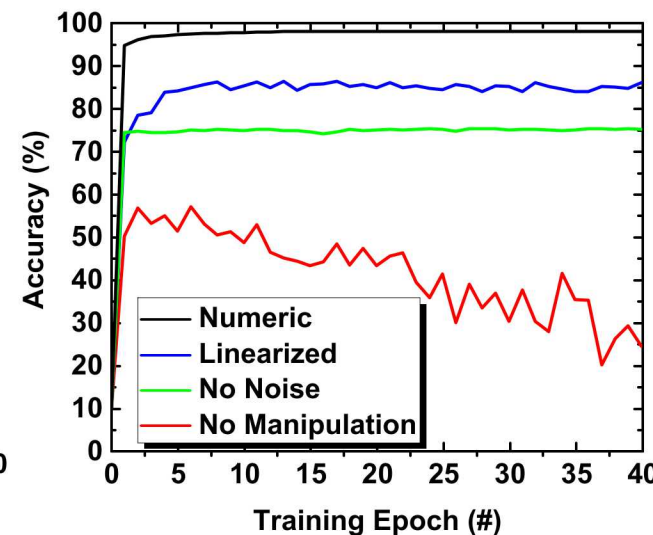
Sandia Baseline TaO_x -Ta



SiO_2 -Cu



Ag-Chalcogenide

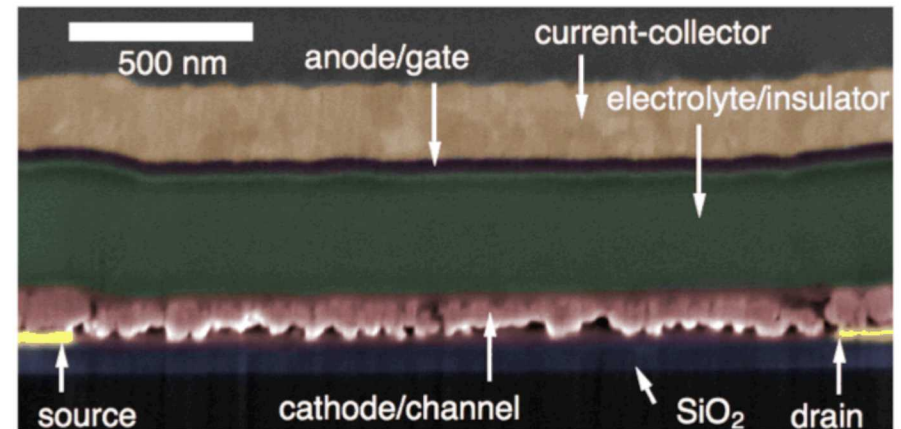
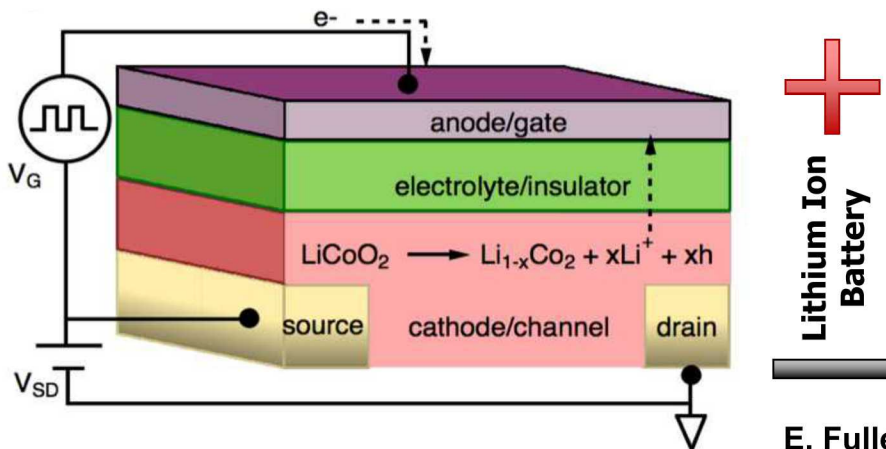
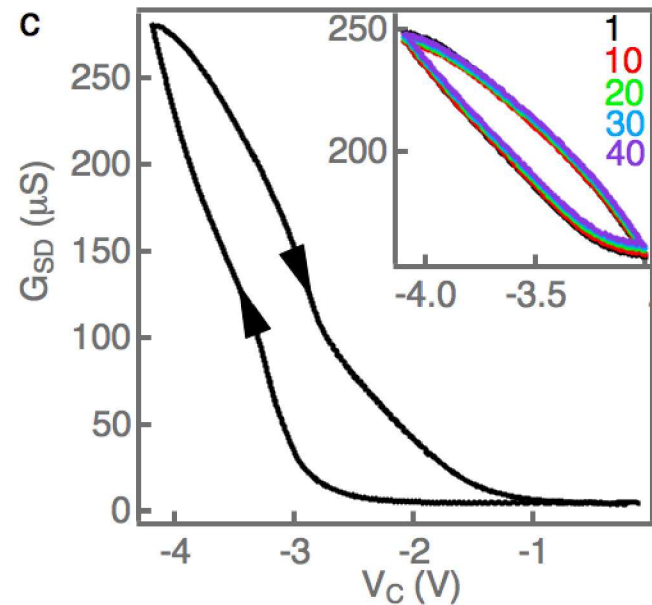


R. Jacobs-Gedrim et al, *IEEE-ICRC*, 2017

Li-Ion Synaptic Transistor for Analog Computation (LISTA)

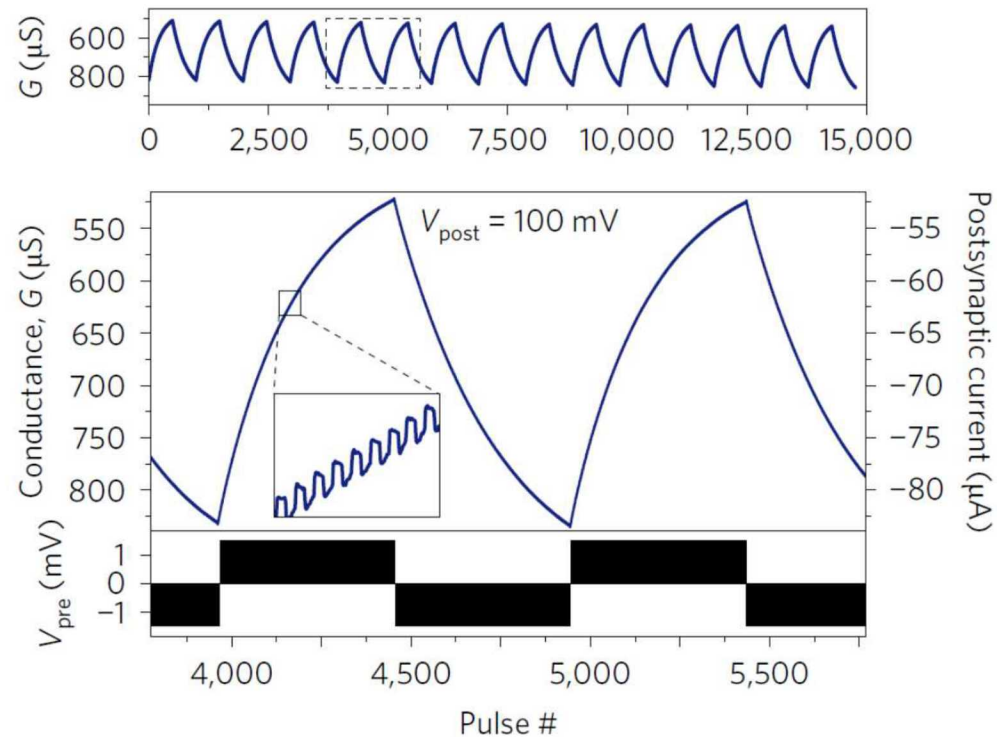
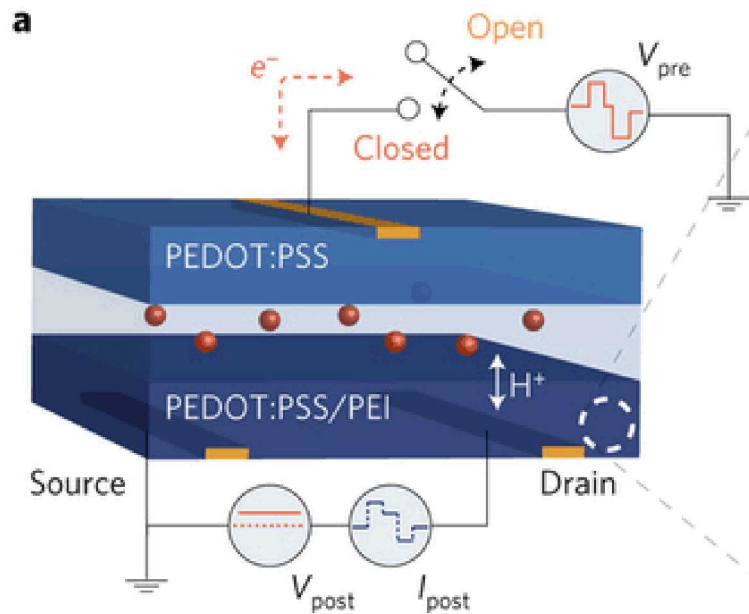
- Alternatively, novel devices may offer promise
- LISTA: modulate the doping of Lithium battery cathode
- Resistivity across cathode changes linearly with battery charge/discharge

G-V for LISTA



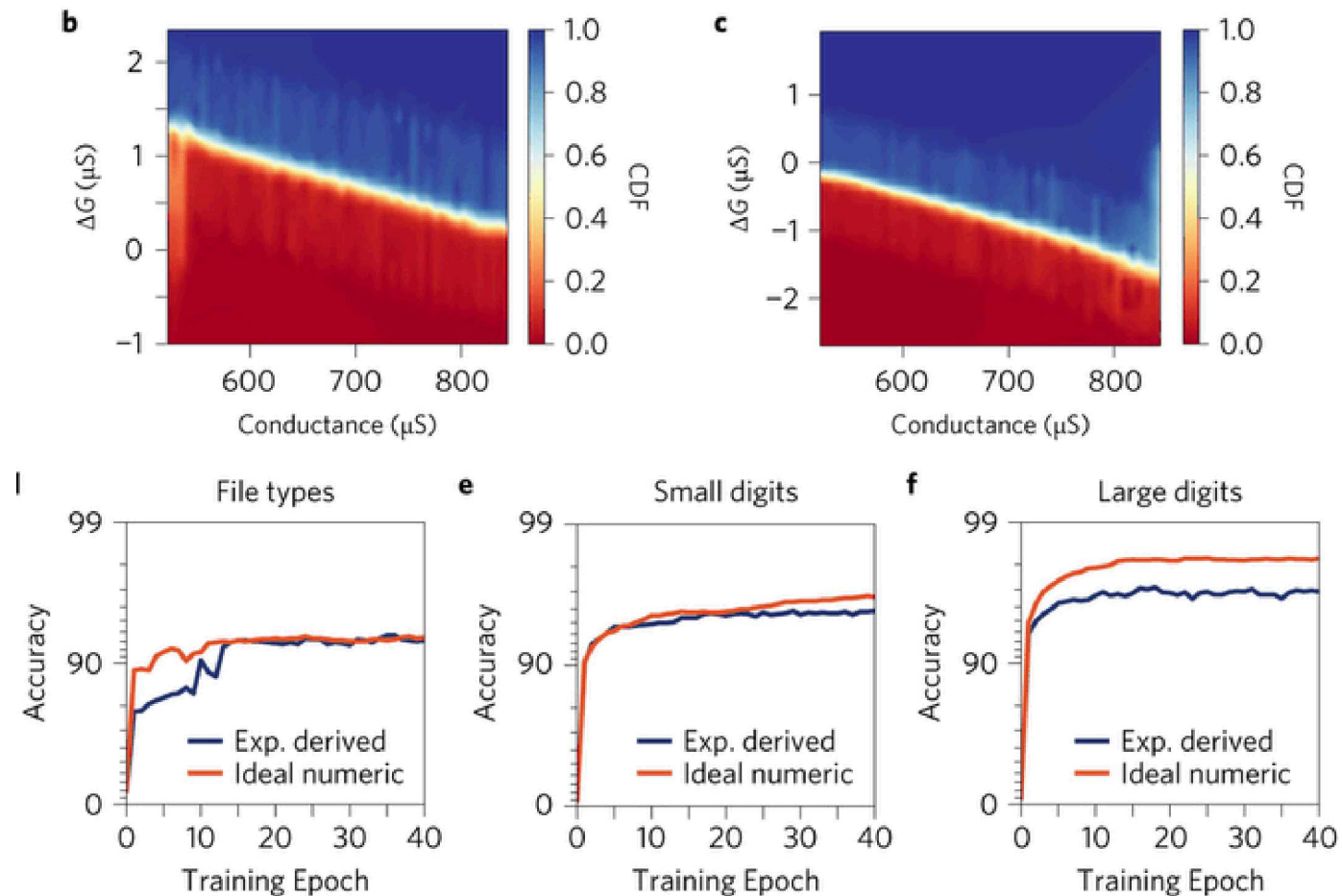
E. Fuller et al, *Adv Mater*, 2017

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