



Exceptional service in the national interest

# Devices for Efficient Analog In-Memory Neural Computing at the Edge

Matthew J. Marinella, Patrick Xiao, Will Wahby, Robin Jacobs-Gedrim, Chris Bennett, Donny Robinson, Elliot Fuller, Alec Talin, Sapan Agarwal

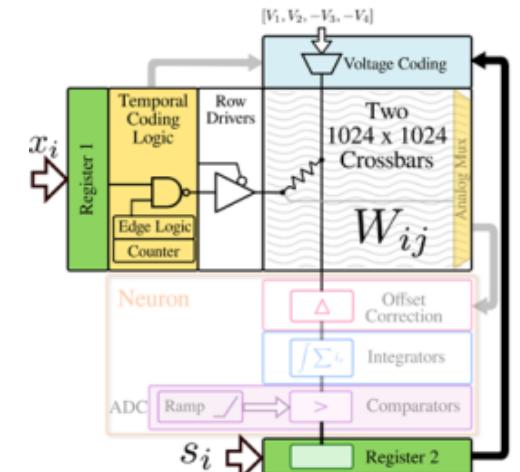
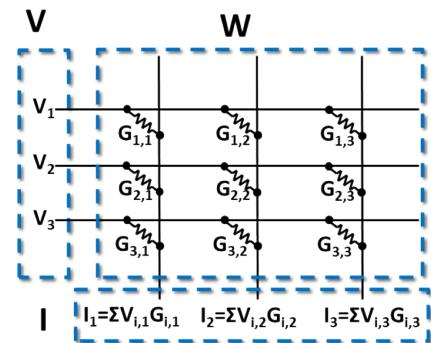
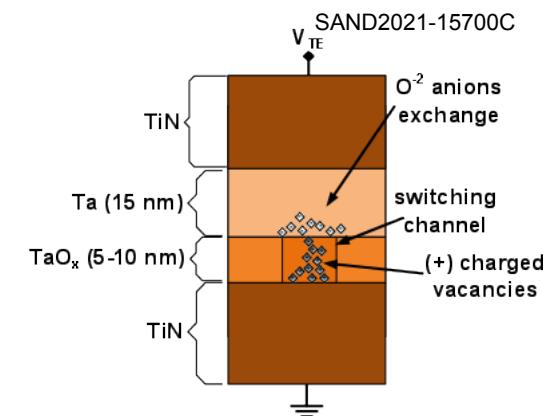
SANDIA NATIONAL LABORATORIES

December 7, 2021

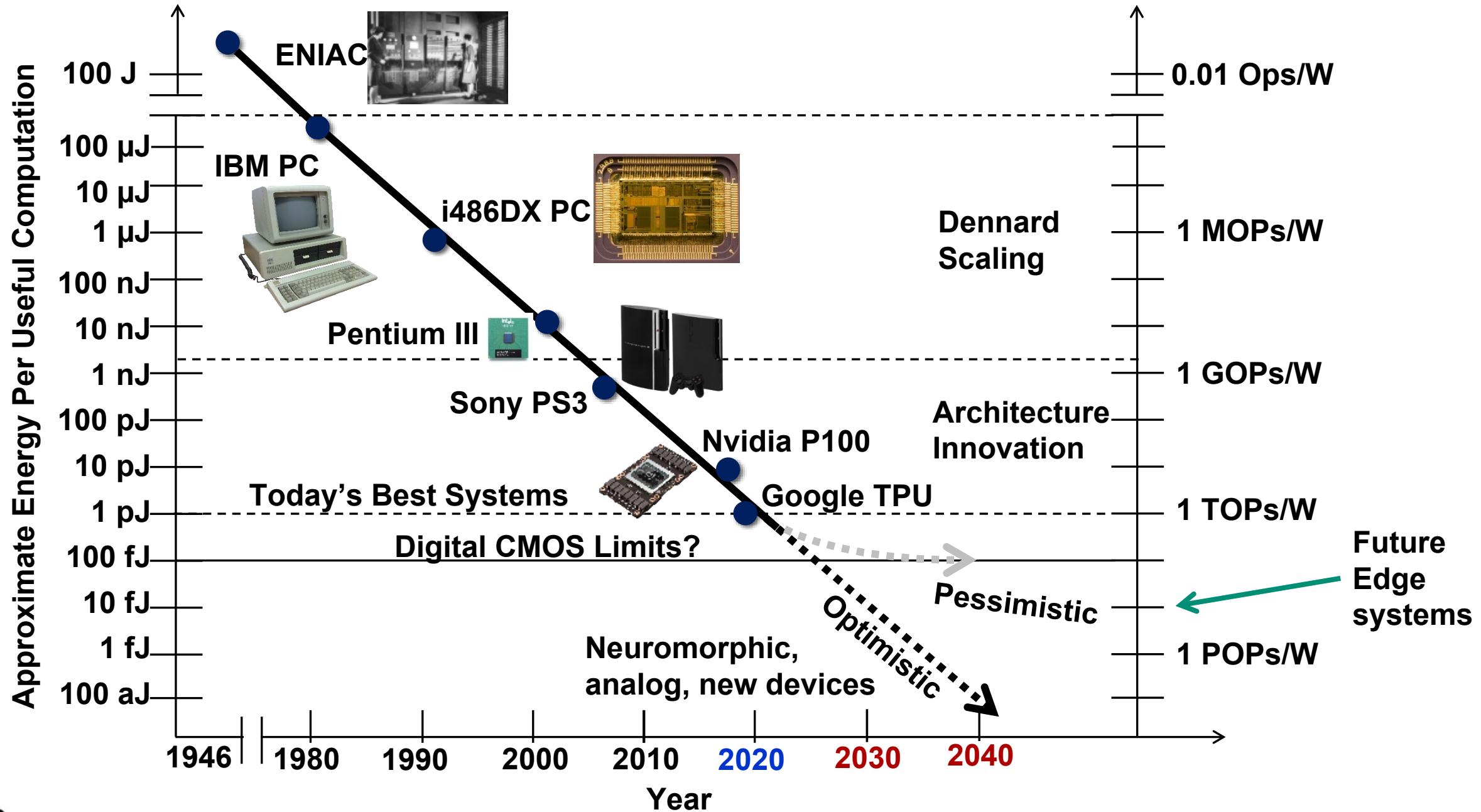


Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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.

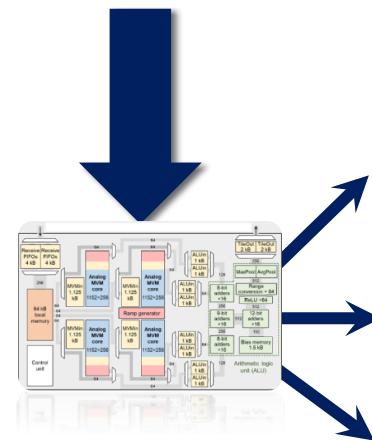


- **Motivation and Background**
- **Analog In-Memory Compute Energy & Latency**
- **Devices for Accurate Inference**
- **Devices for Accurate Training**
- **Conclusions**



# Revolutionary Systems

- What do we want in the future?
- **>10 TOPS/W:**
  - **→Supercomputing at the edge**
- Deep networks (100M+ parameters execute and train in the field)
- Lots of applications interested in this:  
Particle detectors, safe, full autonomous navigation in ground, air and space vehicles
- Getting to this goal may require imperfect hardware...and this might be ok.

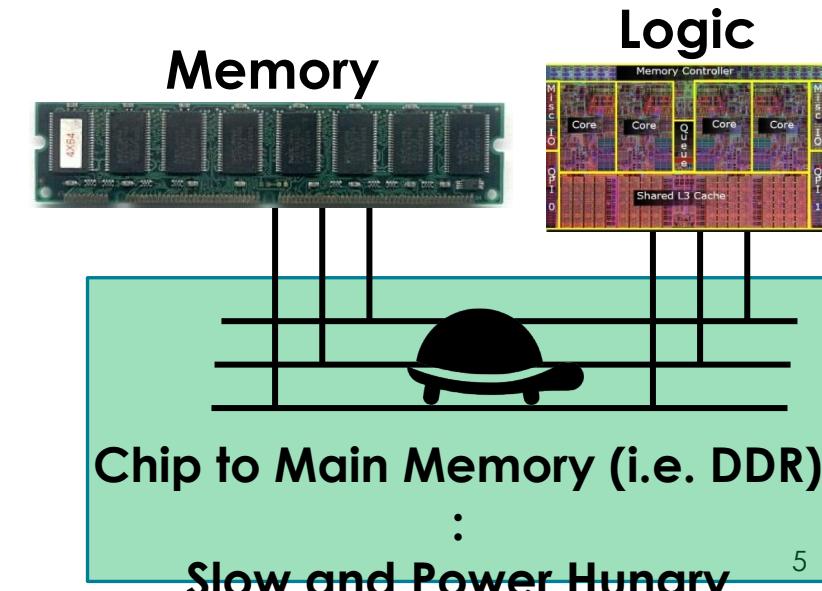


# State of the Art CMOS Efficiency: Apple A13

- Apple's iPhone 11 main SoC processor
  - 7nm+ TSMC process
- Lightening AMX 8-core Neural Engine accelerator IP
  - Apple spec: 5 TeraOps/s (TOPS) @ 8 bit precision
  - Power is ~2.5-5W
  - State of the art smartphone chip is ~ 1-2 TOPS/W
  - **~1pJ per 8 bit operation**
- von Neumann architecture has limitations, especially when off chip data movement is needed
- CMOS research is continuing to push efficiency with low voltage, weight on chip designs – how much more possible?
- **Where will the next orders of magnitude improvements come from?**



[apple.com](http://apple.com), [techinsights.com](http://techinsights.com)



# Outline

---

- Motivation and Background
- Analog In-Memory Compute Energy & Latency
- Devices for Accurate Inference
- Devices for Accurate Training
- Conclusions

# Keep Data in Memory & Exploit Physics for Computing

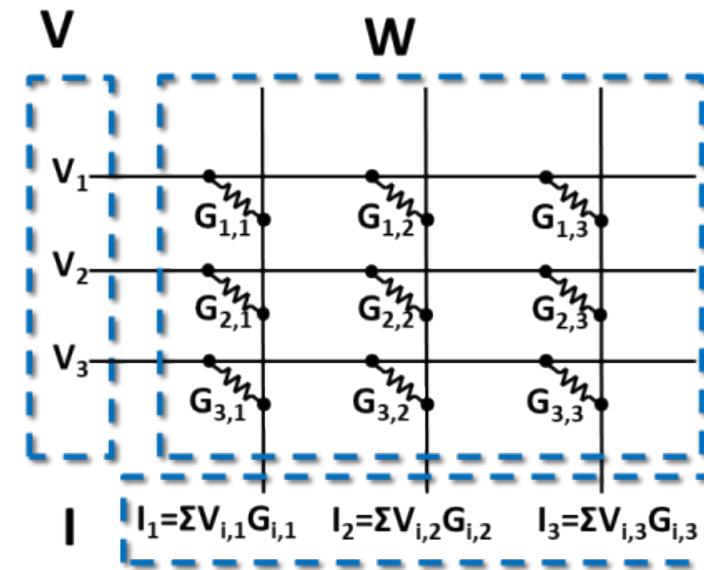
## Mathematical

$$V^T W = I$$

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

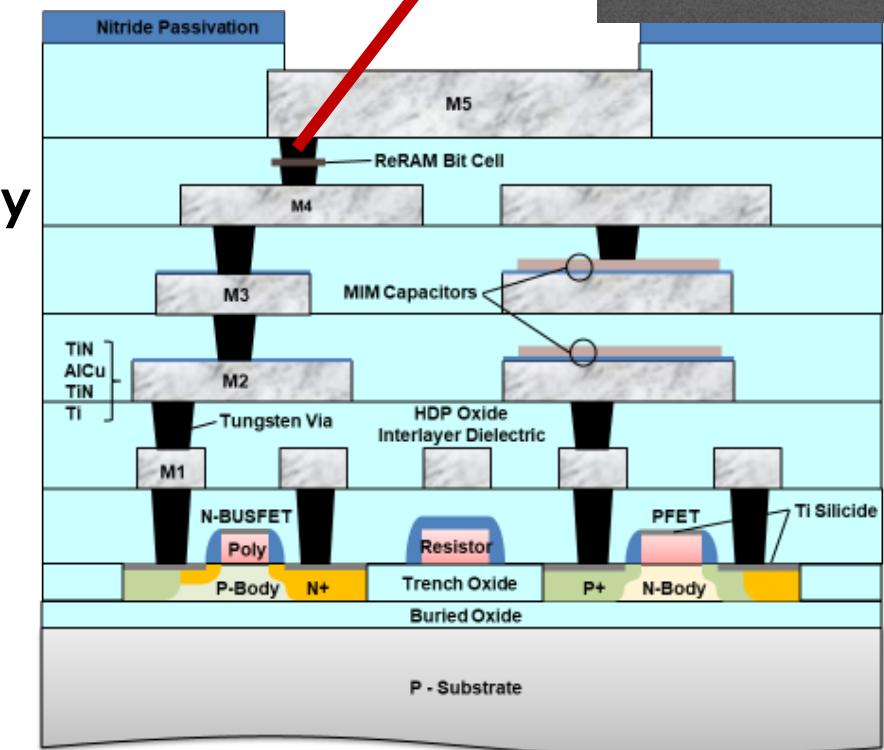
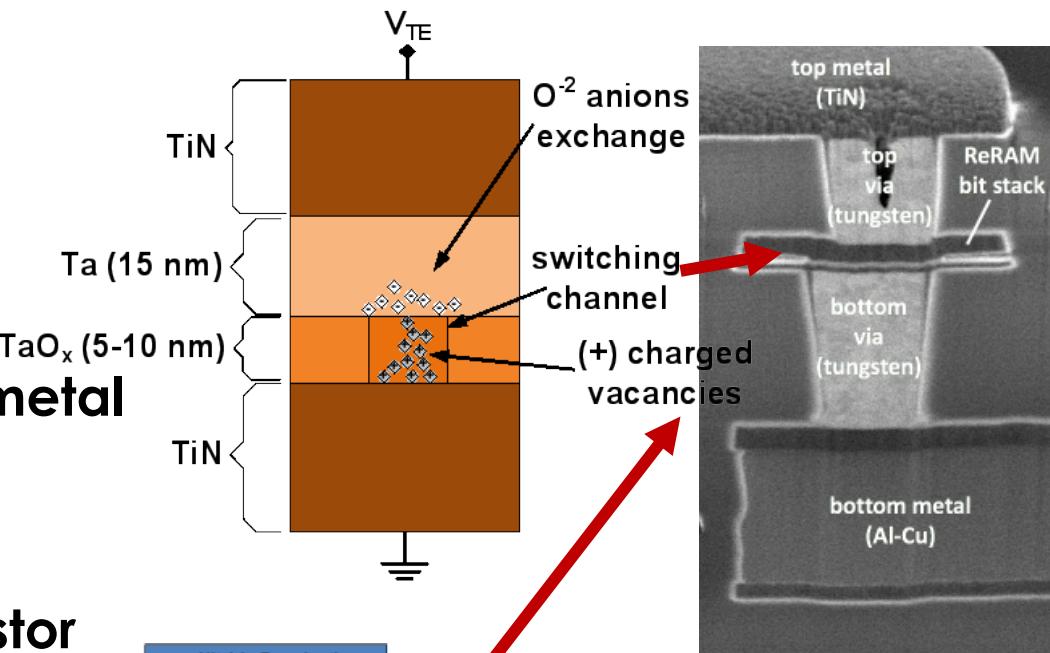
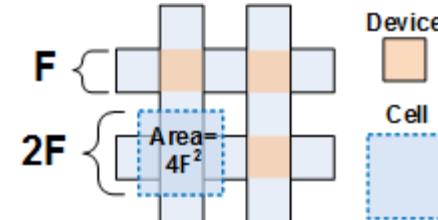
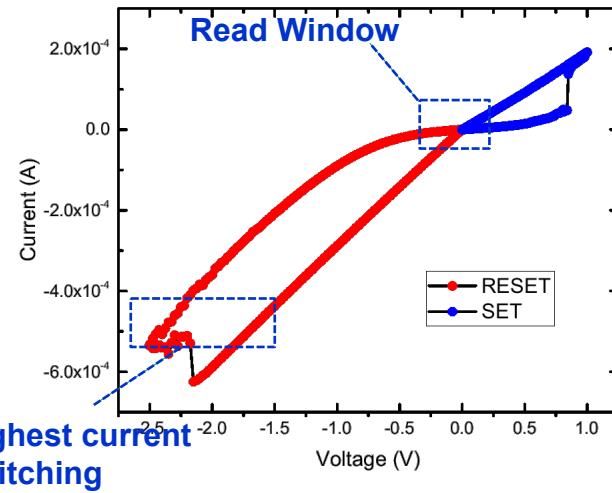
$$\begin{bmatrix} I_1 = \sum V_{i,1} W_{i,1} & I_2 = \sum V_{i,2} W_{i,2} & I_3 = \sum V_{i,3} W_{i,3} \end{bmatrix}$$

## Electrical



# Tunable Resistor: Oxide ReRAM

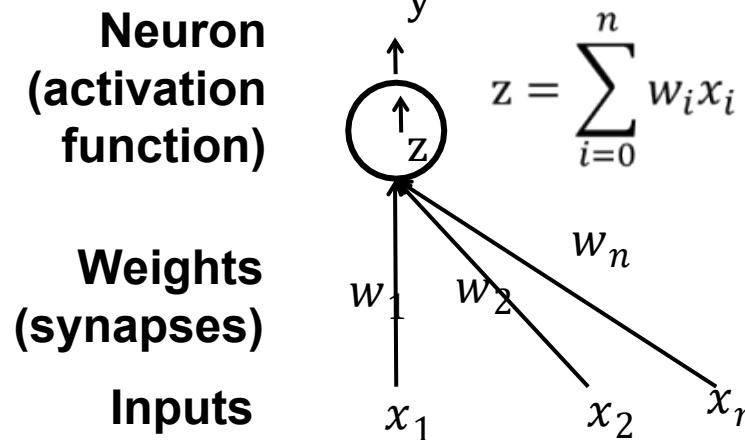
- Known as ReRAM, OxRAM, “memristor”
- Bipolar resistance modulation in metal-insulator-metal structure
  - +V pulse, R decreases. -V pulse, R increases
- Fast, scalable, low switching energy, tunable resistor
- Potential for 100 Tbit of ReRAM on chip
- Perfect Analog In-Memory Compute Energy & Latency candidate! SET-RESET



# Neural Network Basics

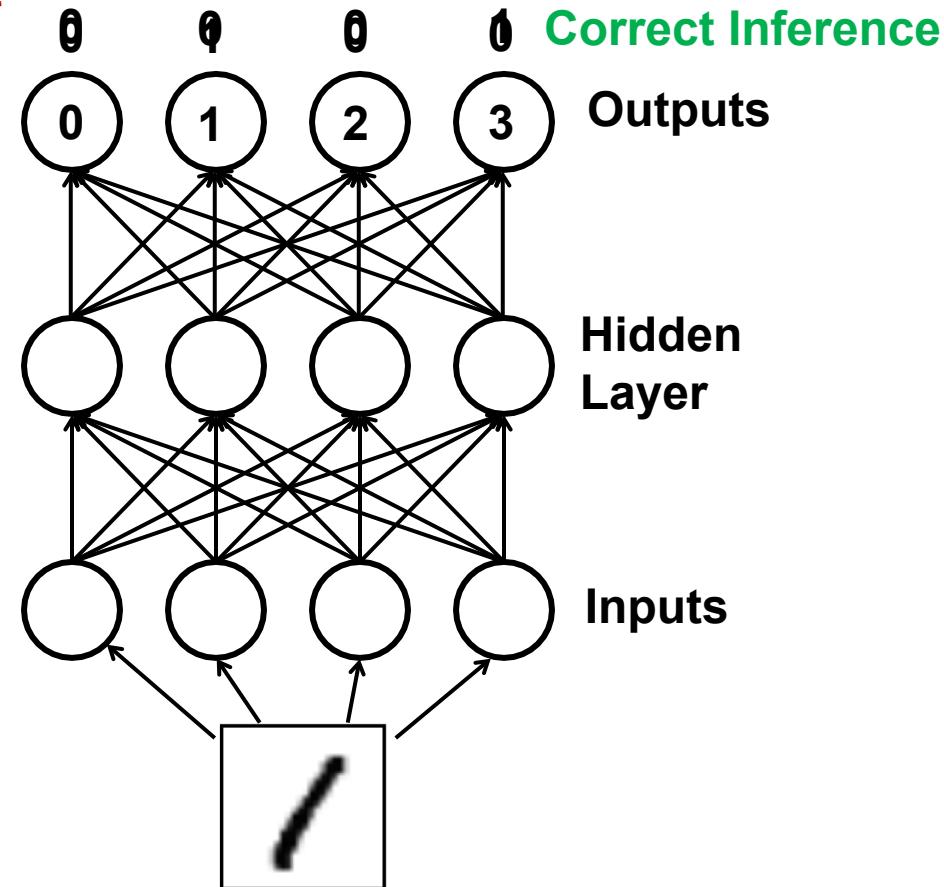
## Basic Building Block

$$y = \frac{1}{1 + e^{-z}}, \text{ReLU, etc.}$$

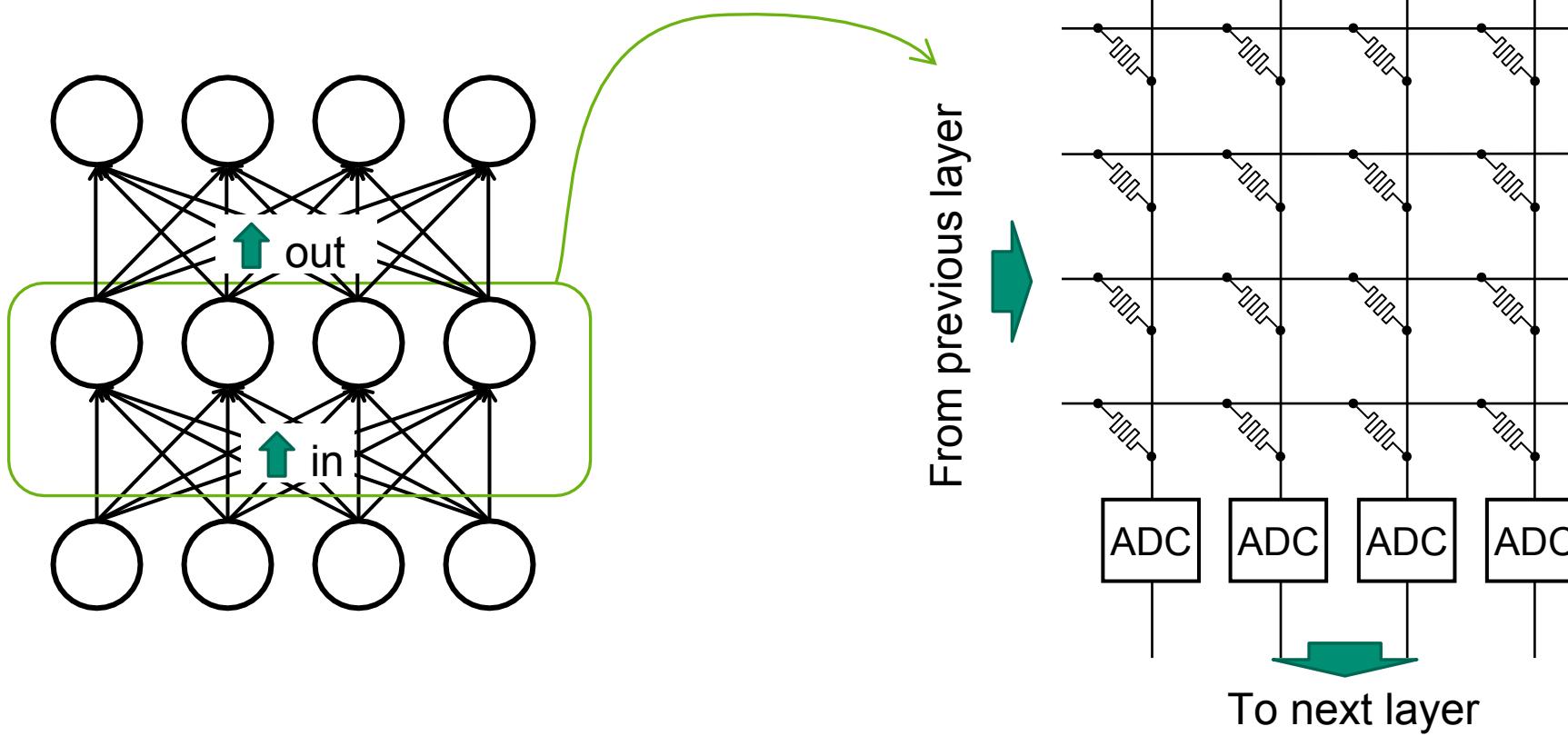


Incorrect –  
adjust if  
training

## Simple Network: Inference & Training (Backpropagation)



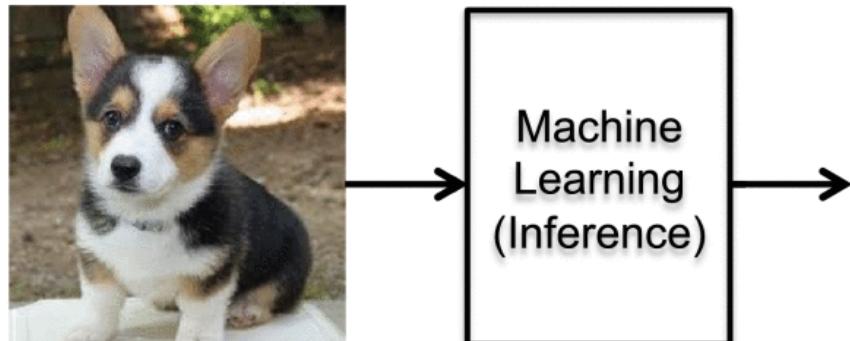
# Physically Mapping a Neural Network



# Neural Network Basics

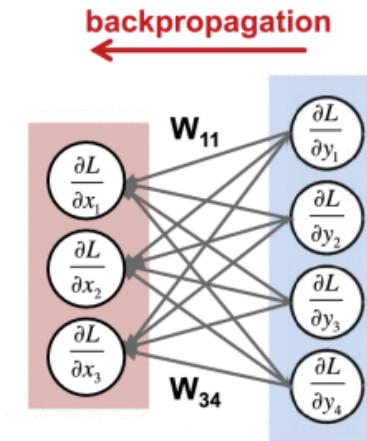
## Inference

- Feed forward operation of the network to perform task, i.e. classification
- Ex: Image recognition
- Computationally requires single feed forward pass through network
- **Typical device update through write-verify**



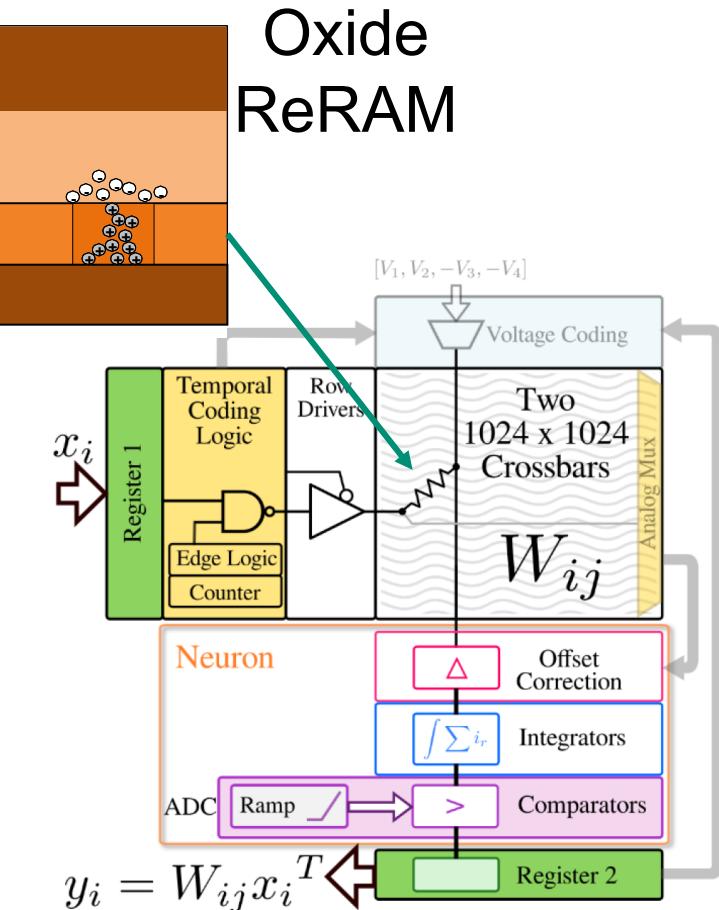
## Training

- Adjusting the weights to reduce error and improve
- Typically done with backprop
- **Parallel update possible on crossbar architecture**



(b) Compute the gradient of the loss relative to the filter inputs

# Tile Analysis



Component	Vector Matrix Multiply (8-bit, Inference)	Outer Product Update (8-bit, Training)
Energy/Op ReRAM (fJ)	12.2	2.1
Energy/Op Digital (fJ)	2718	4102
Array Latency ReRAM (μs)	0.38	0.51
Array Latency Digital (μs)	4	8

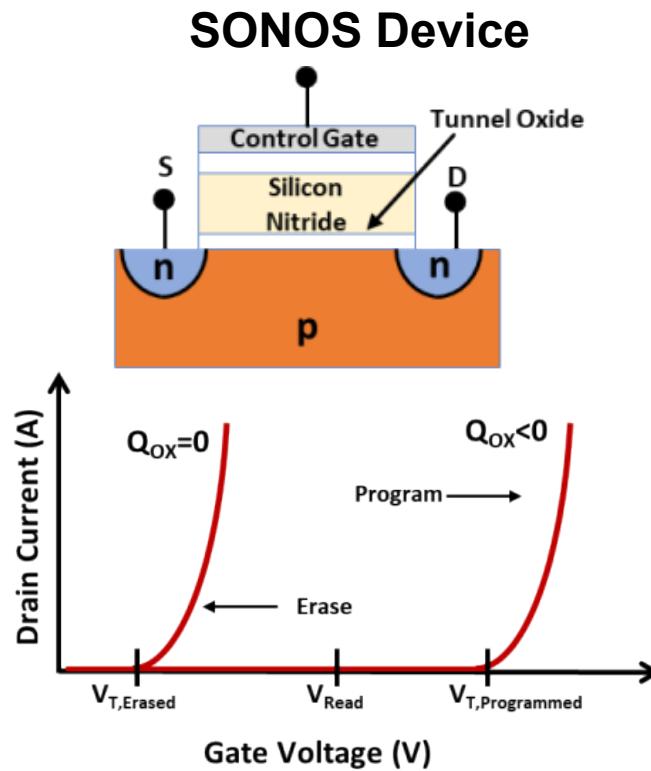
14nm  
PDK

Initial results: two orders of magnitude beyond digital!

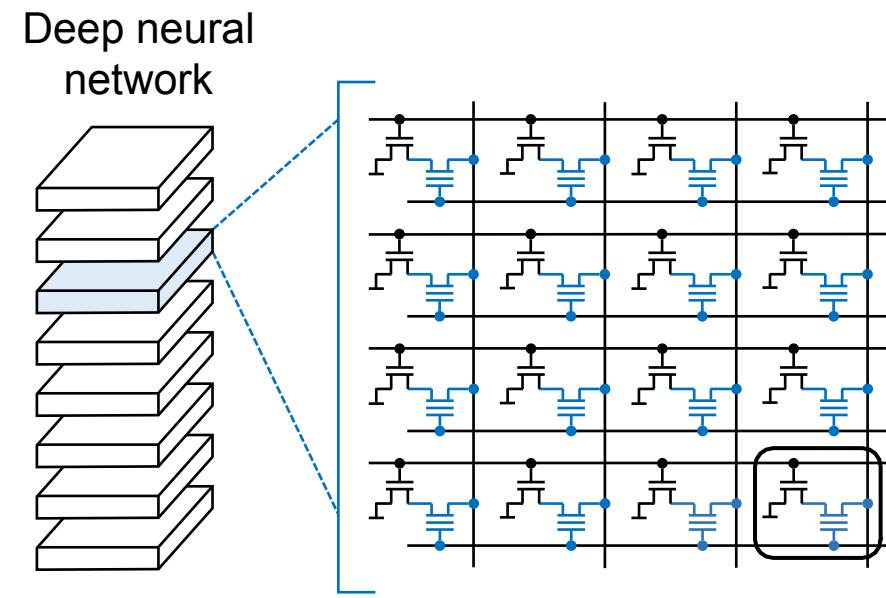


# Semiconductor-Oxide-Nitride-Oxide-Semiconductor (SONOS)

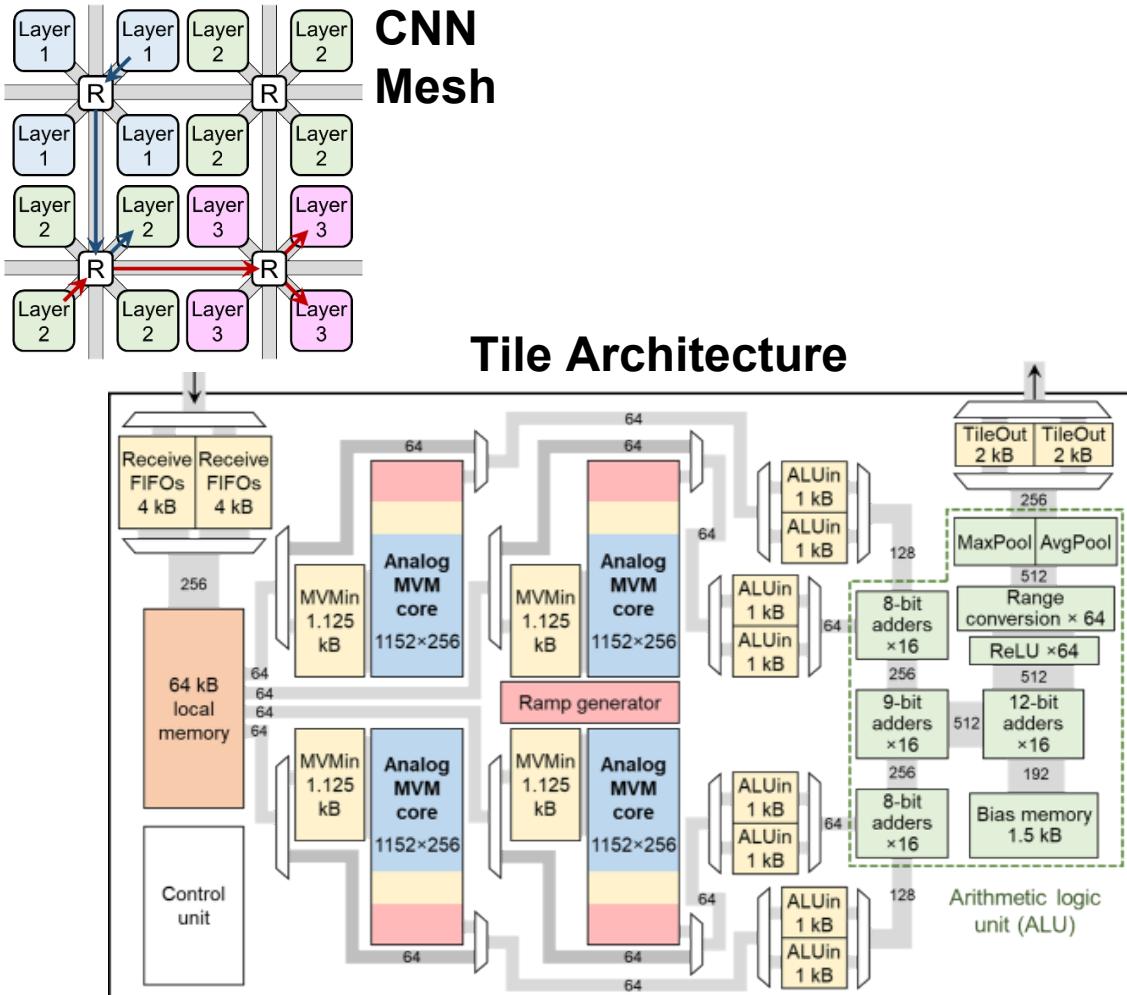
- Mature, commercial technology pioneered by Sandia in the 1980's
- Basis of modern SSD's (your iPhone uses SONOS)
- Can be used as resistive array similar to ReRAM
- Collaborating with Infineon to evaluate 40nm SONOS In Memory Computing



**SONOS Analog VMM Array Implementation**



# 78 TOPS/Watt 8-bit Inference using 40nm SONOS



ISAAC (2016)	Newton (2018)	This work
32 nm, ReRAM	32 nm, ReRAM	40 nm, SONOS
16 bits	16 bits	8 bits
0.63 TOPS/W (theoretical peak)	0.92 TOPS/W (theoretical peak)	21.8 TOPS/W (on ResNet-50) 55 TOPS/W (custom net, near peak)



**Commercial SONOS has excellent inference potential!**

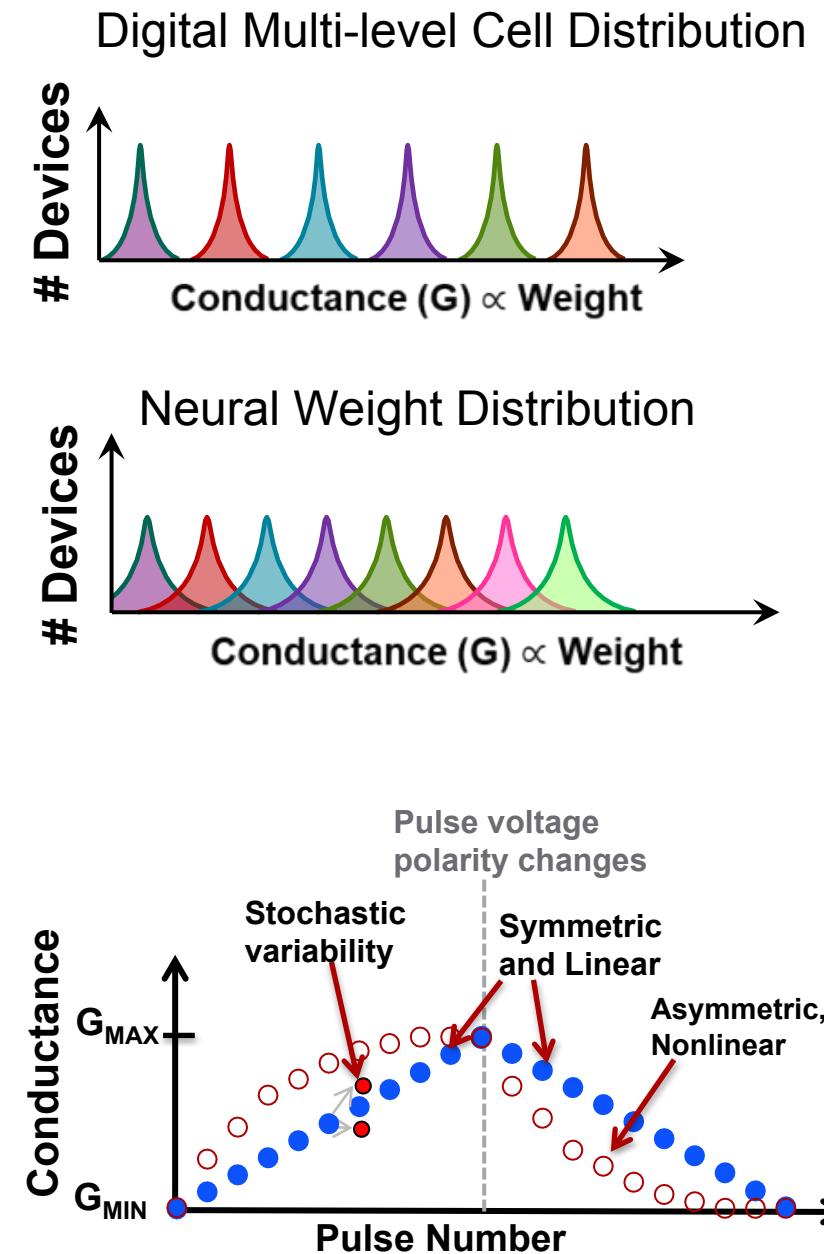
## Outline

---

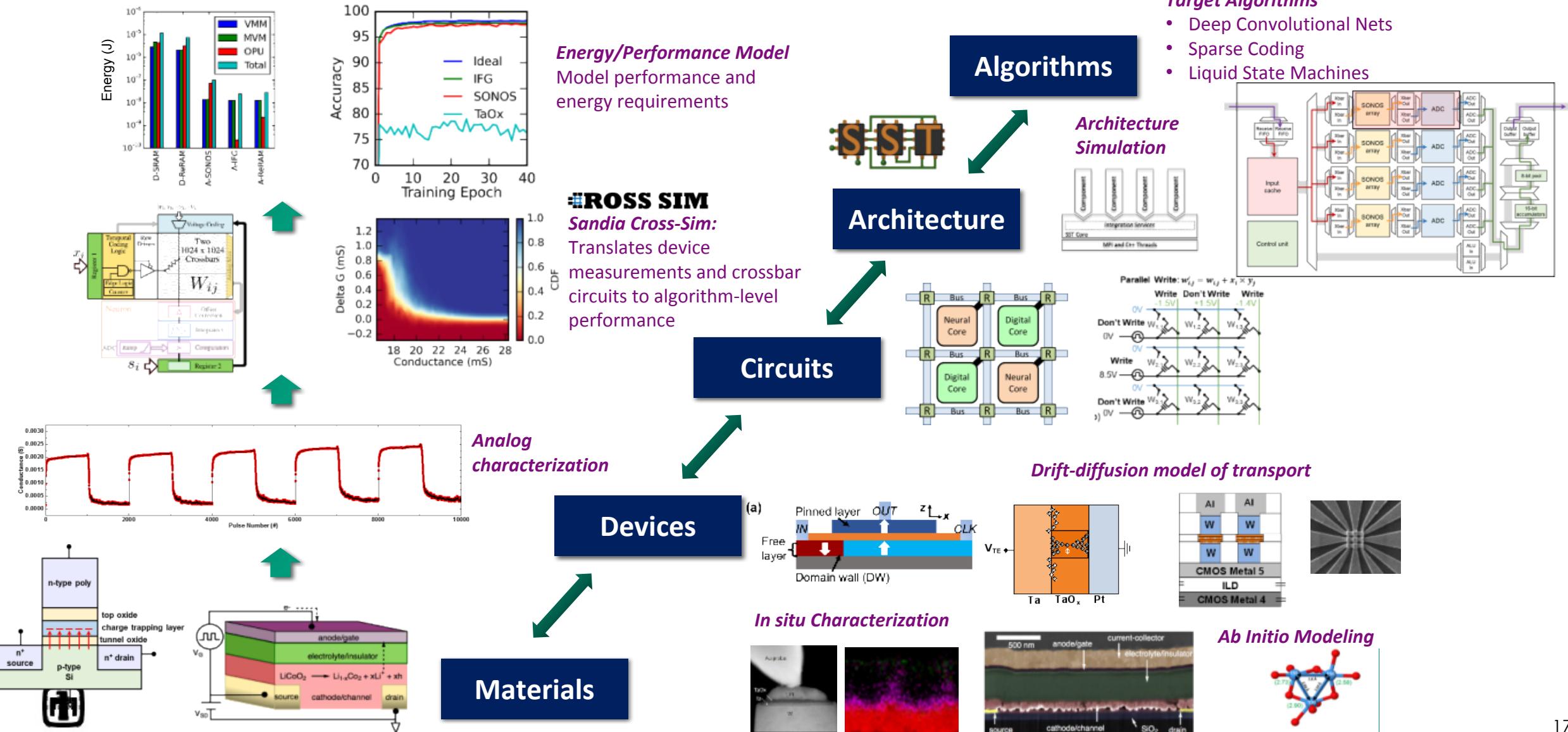
- Motivation and Background
- Analog In-Memory Compute Energy & Latency
- Devices for Accurate Inference
- Devices for Accurate Training
- Conclusions

# Analog Required a Paradigm Shift

- Analog processing offers great benefits...
- ...but comes with great challenges
- Digital: Deterministic, accurate results
- Analog: Device characteristics affect algorithm accuracy!
  - Research challenge: analog behavior must give acceptable algorithm -level results
- Inference Accuracy Challenges (this section)
  - Measured device conductance should be proportional to weight – but this is only approximately true
  - Caused by **analog programming accuracy versus state, current drift, read noise**
- Training Accuracy Challenges (next section)
  - Actual analog device state change does not match intended weight update
  - Caused by **write nonlinearity, asymmetry, stochasticity**
  - **Device to device variation**



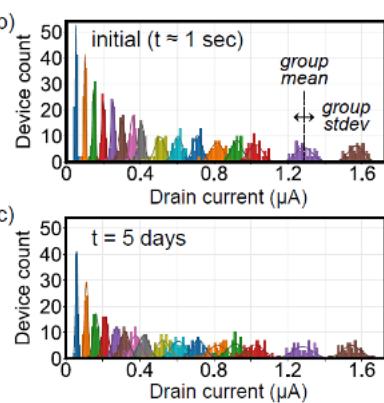
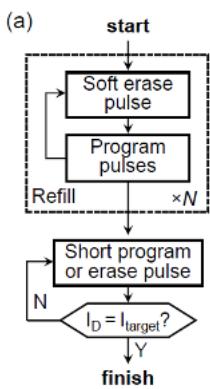
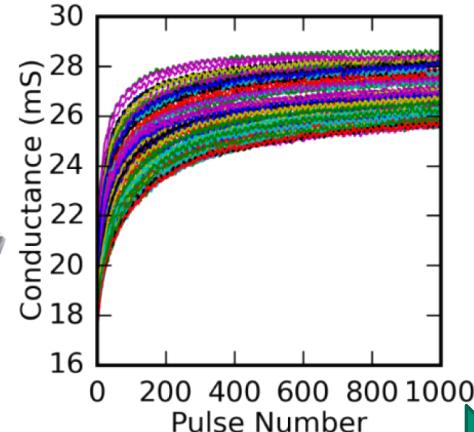
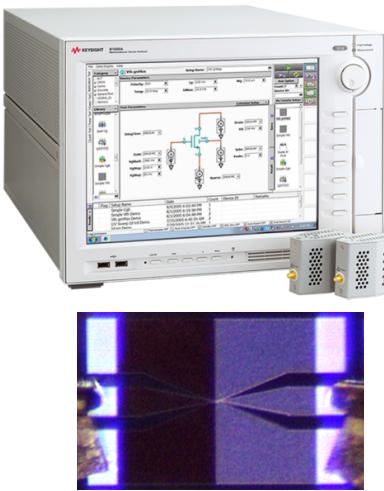
# Multiscale CoDesign Framework Enables Accuracy Prediction



# Compact Modeling Dataset for Neural Accuracy Model

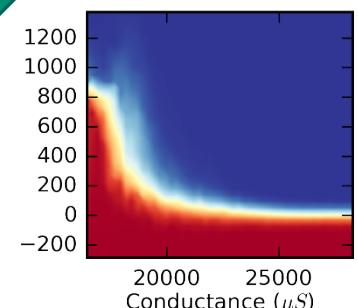
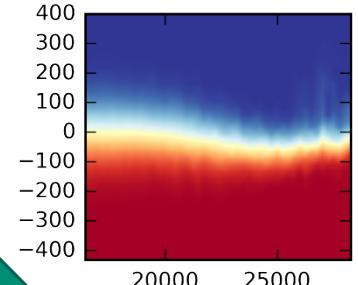
Assess Neural  
Algorithm  
Accuracy,  
Efficiency,  
Performance,  
Radiation

## Measure Devices



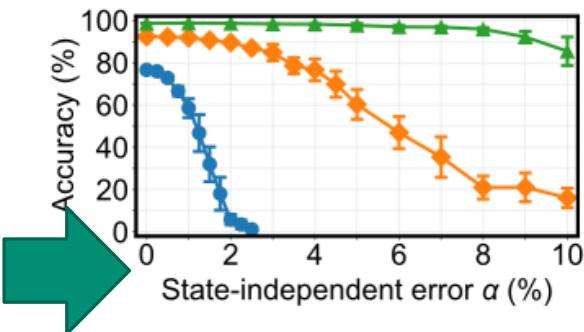
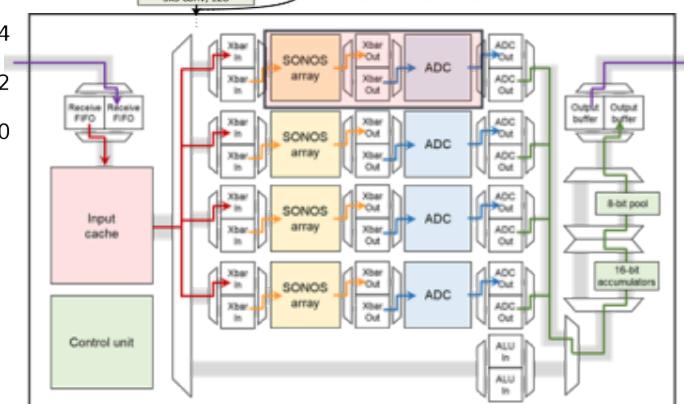
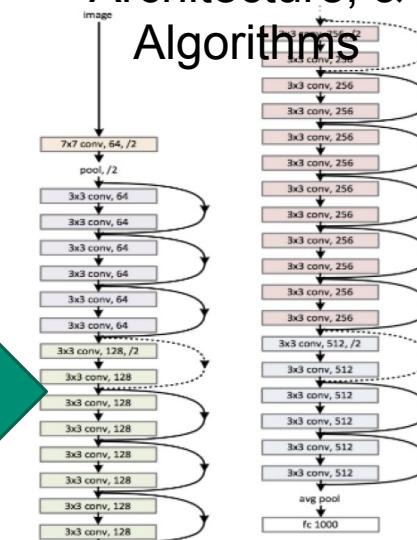
## Construct Lookup Tables

100 ns



#ROSS SIM

## Model Array Circuitry, Architecture, & Algorithms

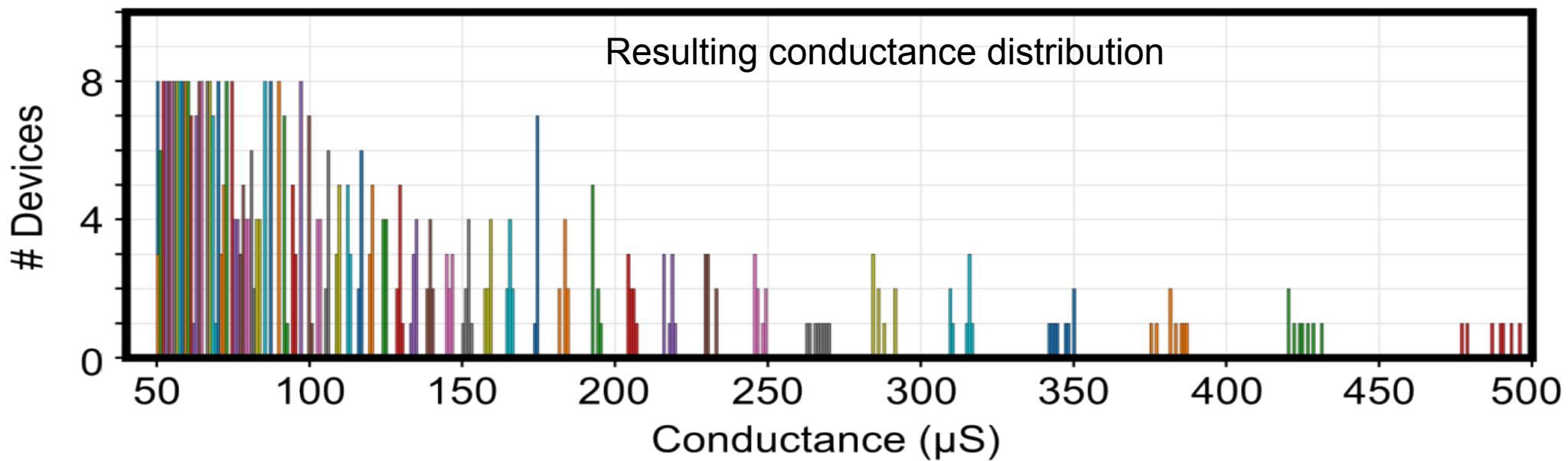
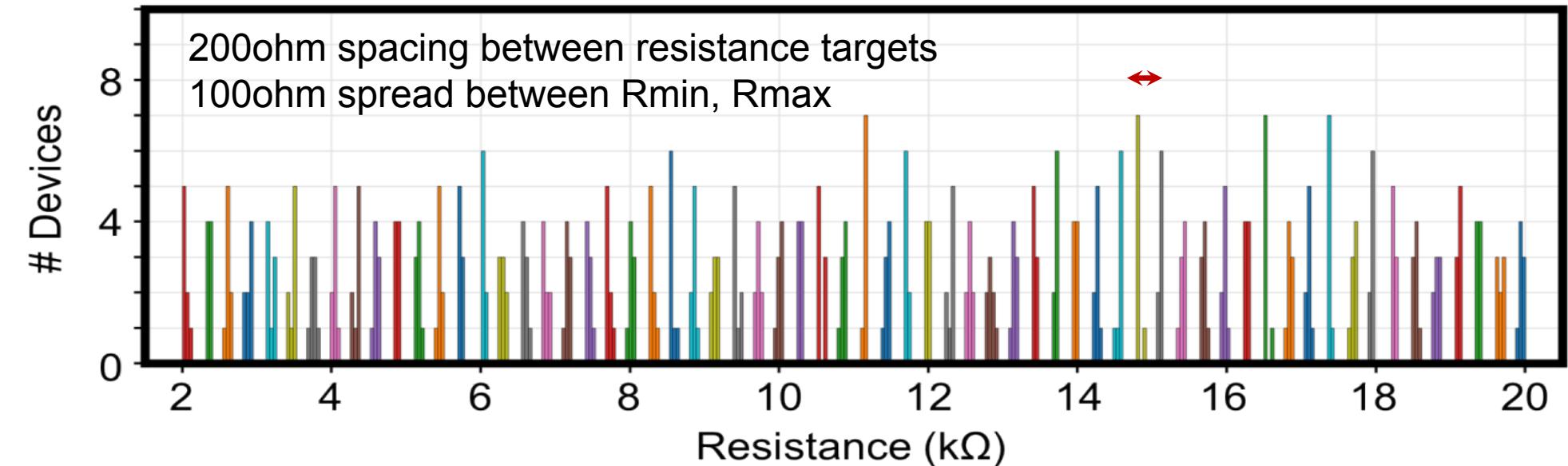
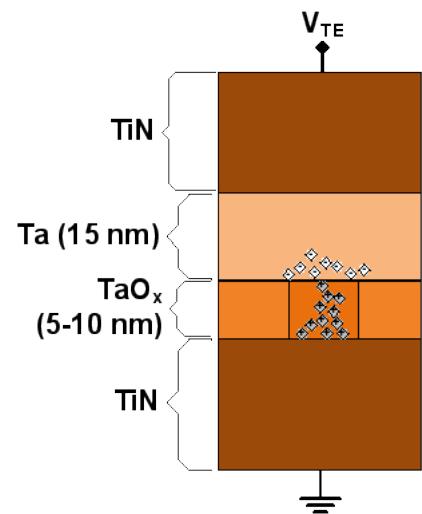


Component VMM OPU

Energy/Op ReRAM (fJ)	12.2	2.1
Array Latency ReRAM (μs)	0.38	0.51



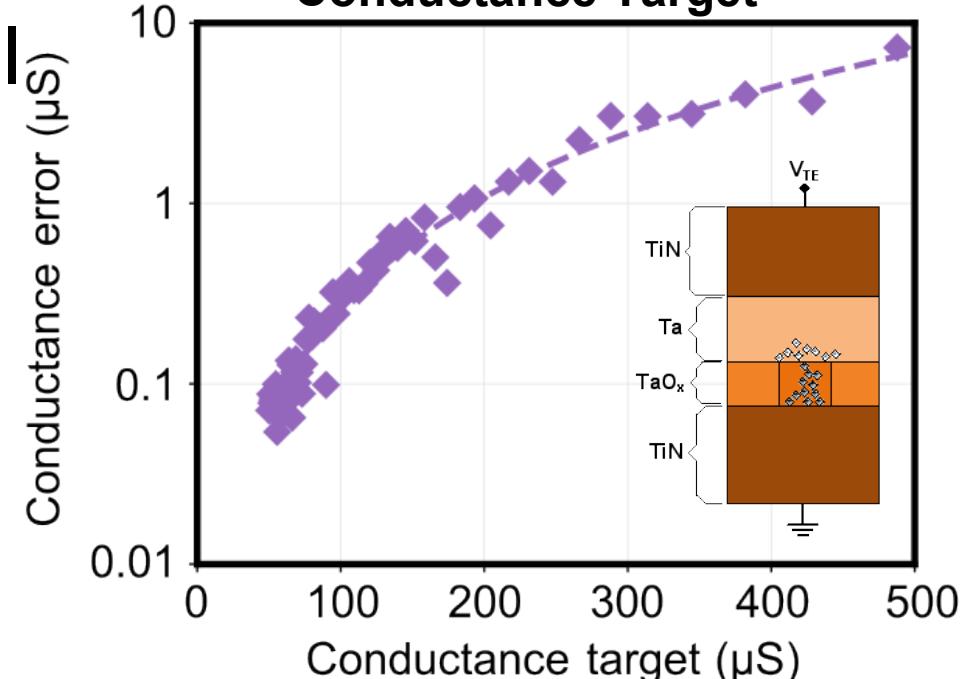
# Sandia TaOx ReRAM Inference Resistance Distributions



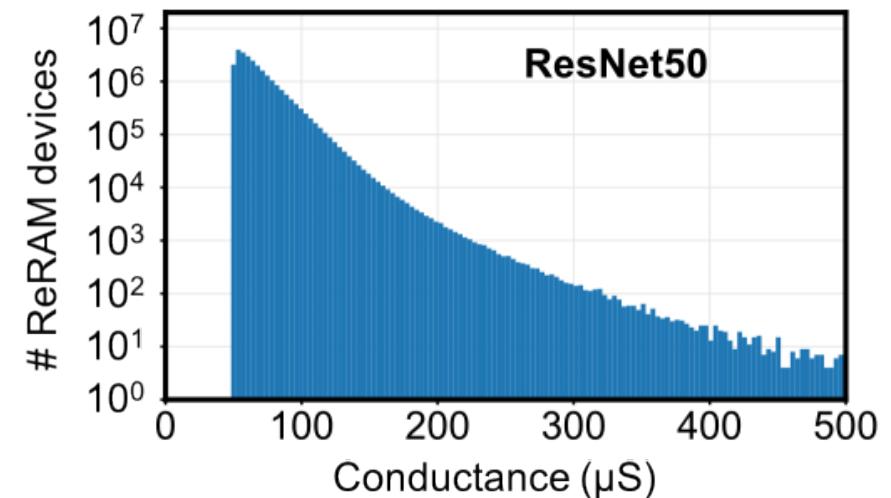
# Sandia TaOx ReRAM Error Model

- Conductance error approx parabolic with conductance target – this is ideal:
  - Lower conductances have lowest error and map to weights near zero.
  - Weights near zero hold most information, hence device error is minimized
- Modeled Accuracy in CrossSim Inference
  - ResNet50 CNN, ImageNet Dataset
  - 1000 image average
  - 8-bit ADC, 8-bit weight quant
  - Assume  $G_{ON}/G_{OFF} = 10$
- ReRAM accuracy on ImageNet:
  - Top-1 76.4%
  - Top-5 92.91%
- Compared to Digital (32 bit FP)
  - Top-1 77.18% (analog loss = 0.78%)
  - Top-5 93.06% (analog loss = 0.15%)
- Analog Inference predicted <1% loss!

Conductance Error as a Function of Conductance Target

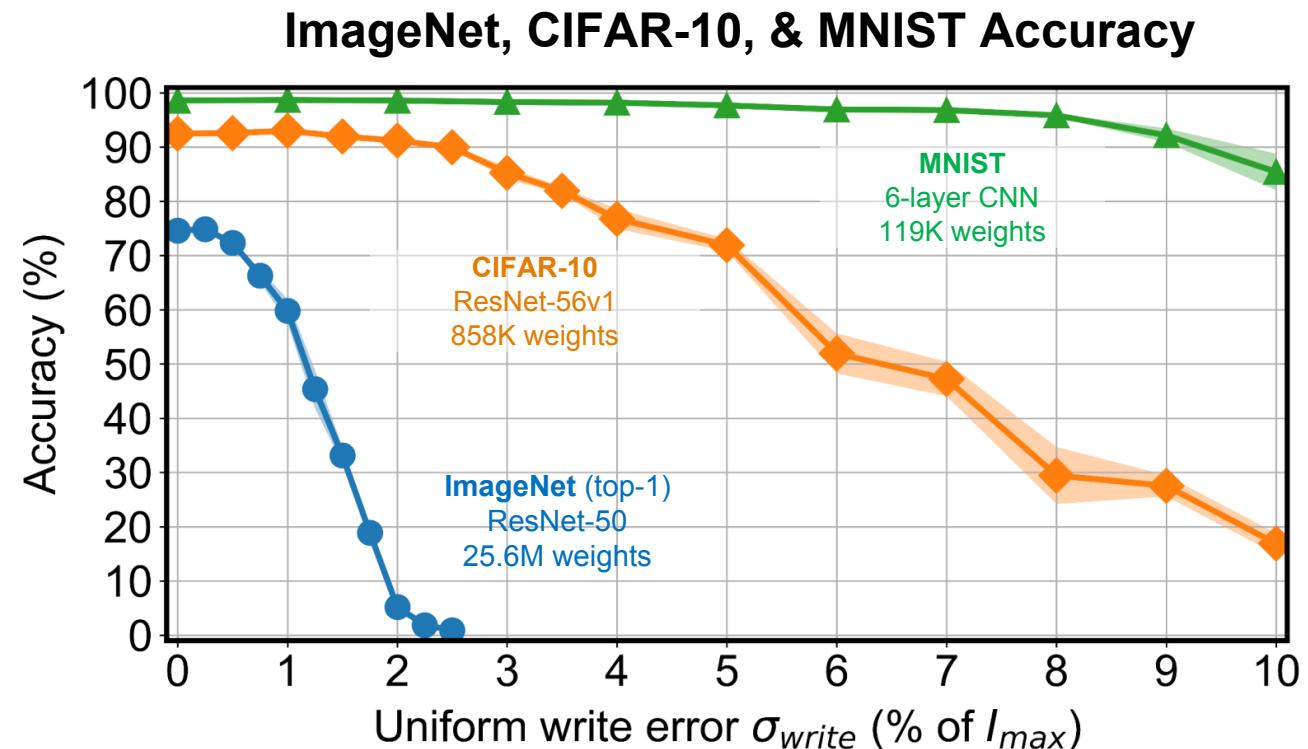


Conductance-Weight Distribution



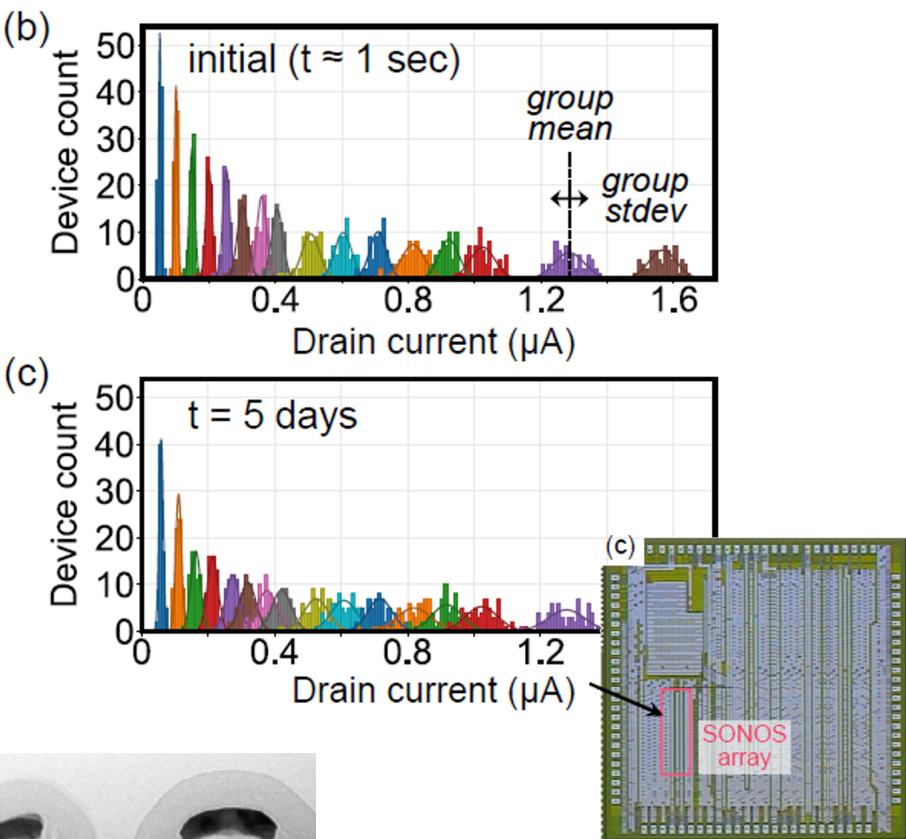
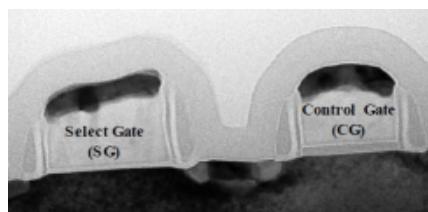
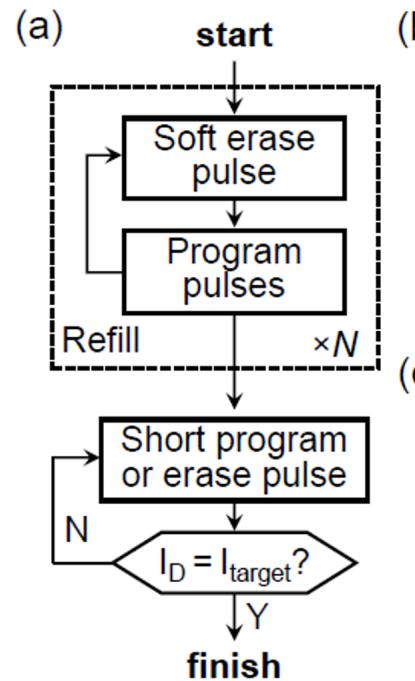
# Effect of Network and Dataset on Accuracy

- Different common datasets and CNN architectures often analyzed
- MNIST (uses simple CNN)
  - 28x28 pixel grayscale
  - 10 classes
  - 60k training images
  - 10k test images
- ImageNet (requires large CNN arch.)
  - 224x224 pixel color
  - 1000 classes
  - 1.3M training images
  - 100k test images
- ImageNet represents production-grade dataset
  - Sometimes smaller nets like MNIST are used due to computing constraints, esp for modeling training
  - Excellent accuracy on MNIST does not translate to excellent accuracy on ImageNet!

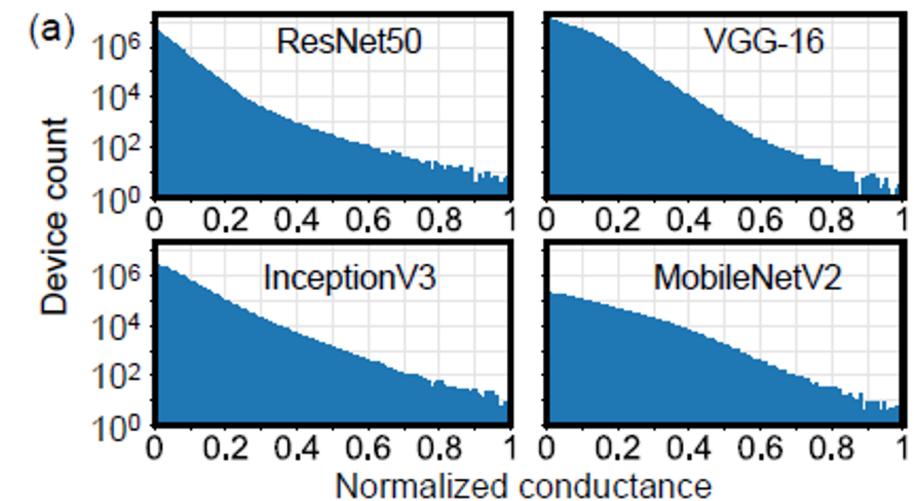
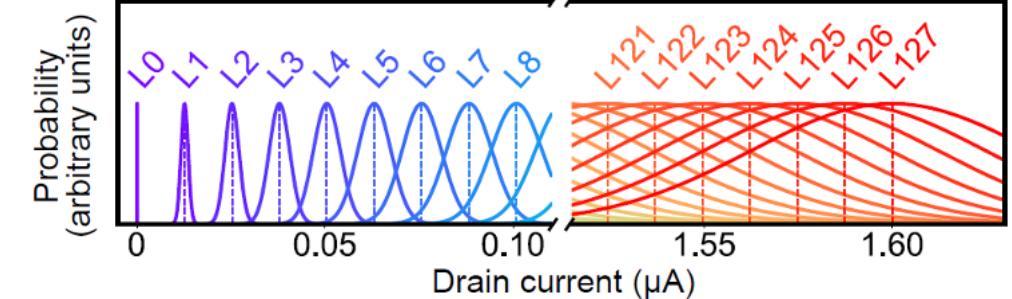


# 40nm SONOS Deep CNN Inference Modeling

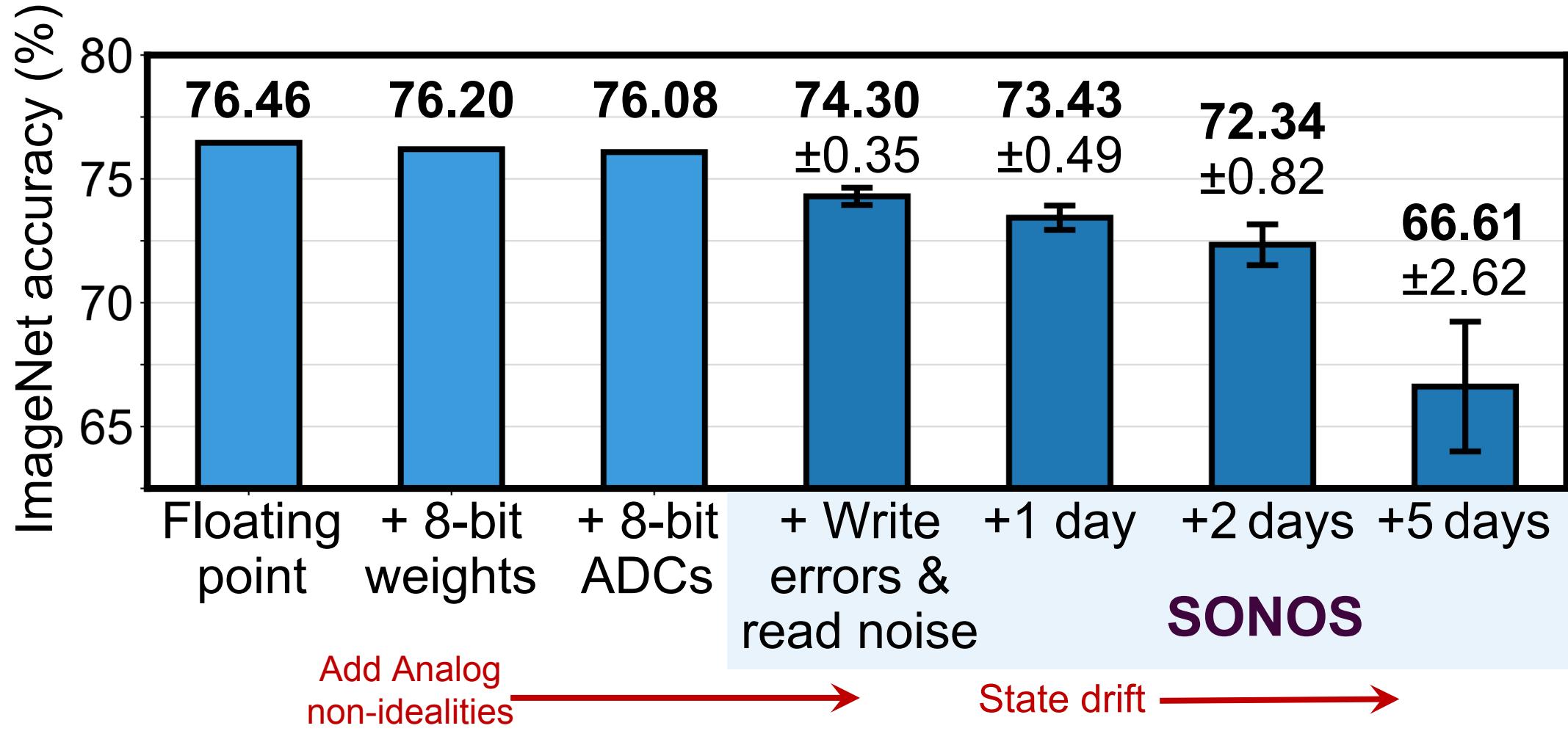
# Infineon 40nm SONOS Characterization Chip Data



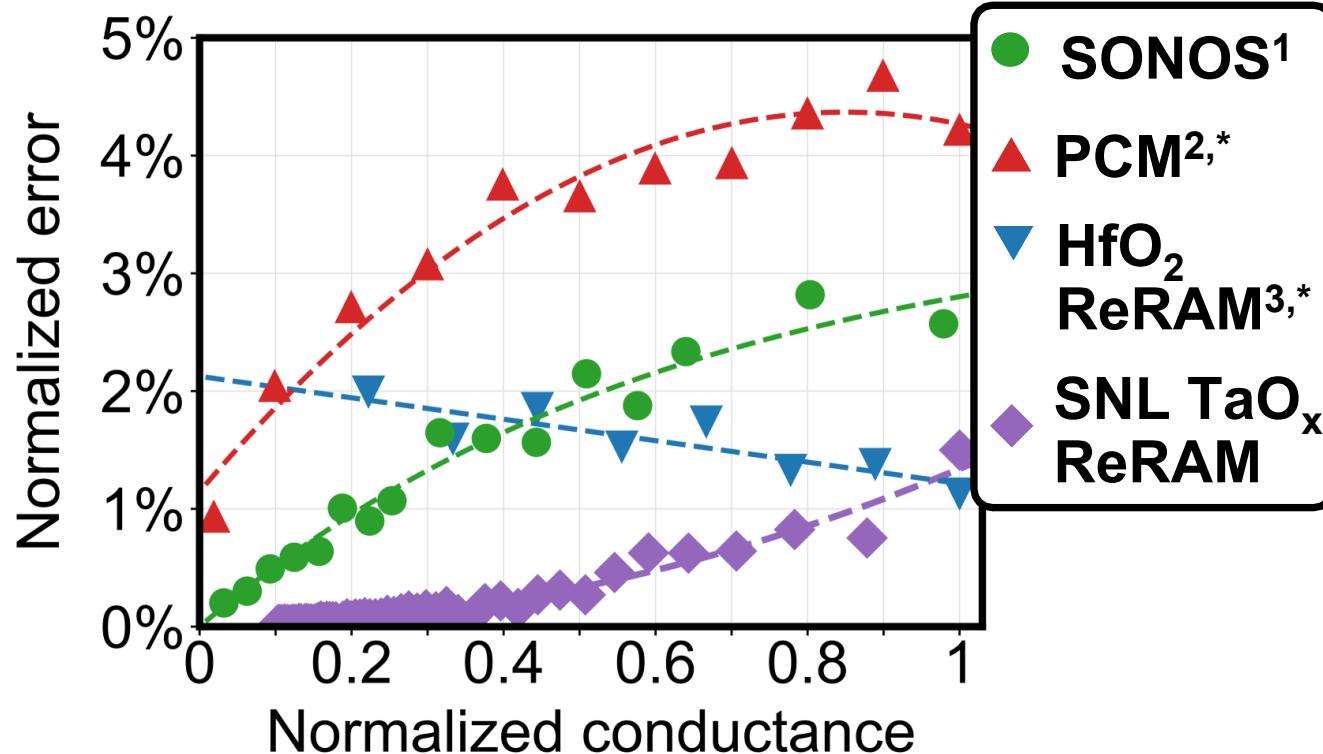
## Modeled 7-bit Weight Distribution and Mapping



# Infineon 40nm SONOS for CNN Inference - ImageNet



# Error and Inference Accuracy Summary: ReRAM, SONOS, PCM



## References and notes:

<sup>1</sup>T.P. Xiao et al, IEEE TCAS, in press, 2021.

<sup>2</sup>V. Joshi et al, Nat Comm. 11, 2020.

<sup>3</sup>Milo et al, IEEE IRPS, 2021.



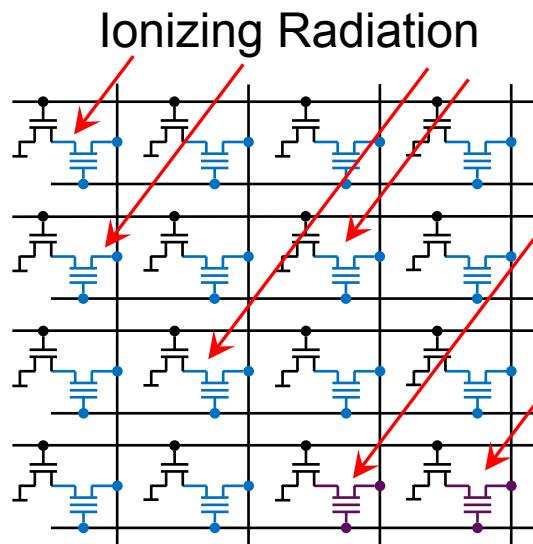
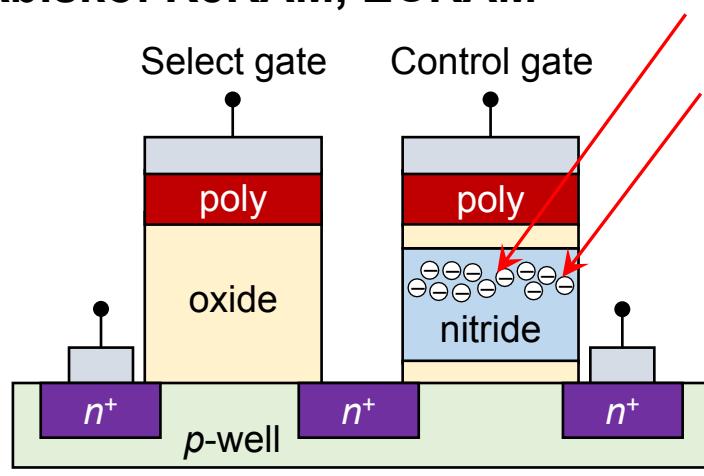
<sup>+</sup>All analog simulation also includes 8-bit weight quantization, 8-bit activations, and 8-bit ADCs

<sup>\*</sup>PCM and HfO<sub>2</sub> error are modeled entirely from data and programming used in publication only.

Technology <sup>+</sup>	Top-1 accuracy	Top-5 accuracy
Floating point digital	77.5%	93.3%
SNL TaO <sub>x</sub> ReRAM	76.4% $\pm$ 0.2%	93.3% $\pm$ 0.1%
SONOS <sup>1</sup>	74.0% $\pm$ 1.0%	92.5% $\pm$ 0.4%
PCM <sup>2</sup>	28.2% $\pm$ 6.4%	49.7% $\pm$ 7.8%

# Device-Level Radiation Impacts Algo Accuracy (T)

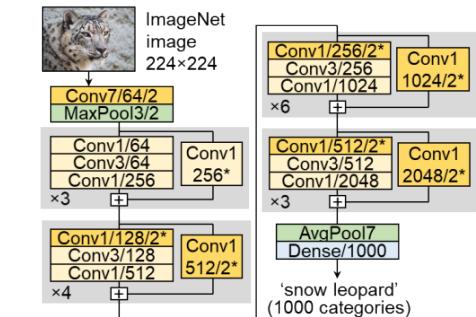
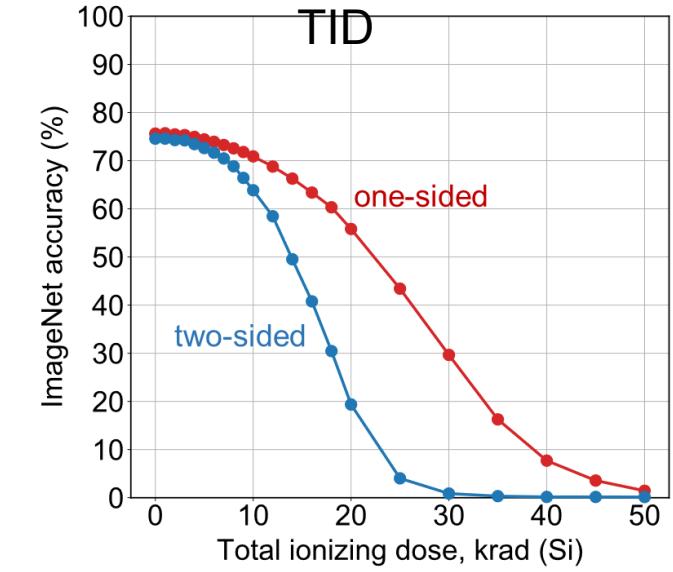
Prior Work: Commercial SONOS  
Abisko: ReRAM, ECRAM



How will the accuracy  
degrade in radiation  
environments ?



Algorithm Accuracy Degradation due to



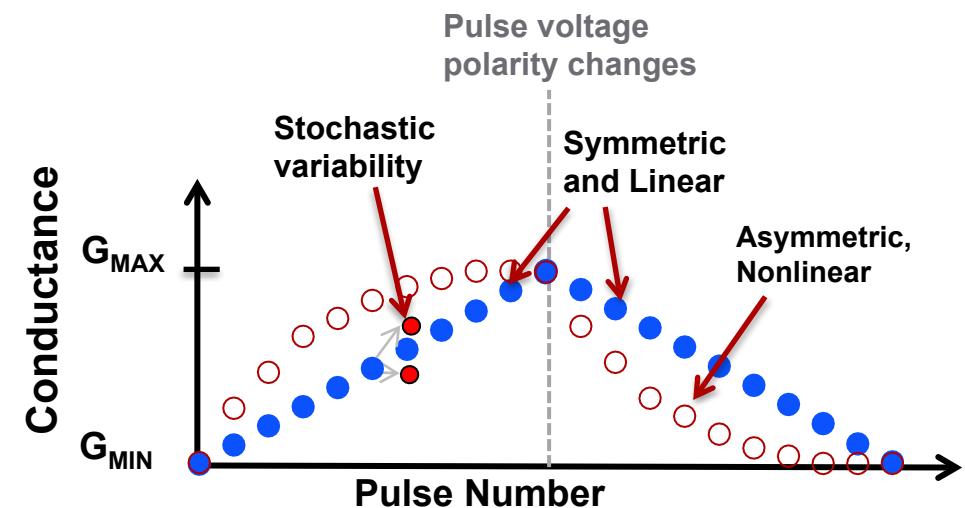
## Outline

---

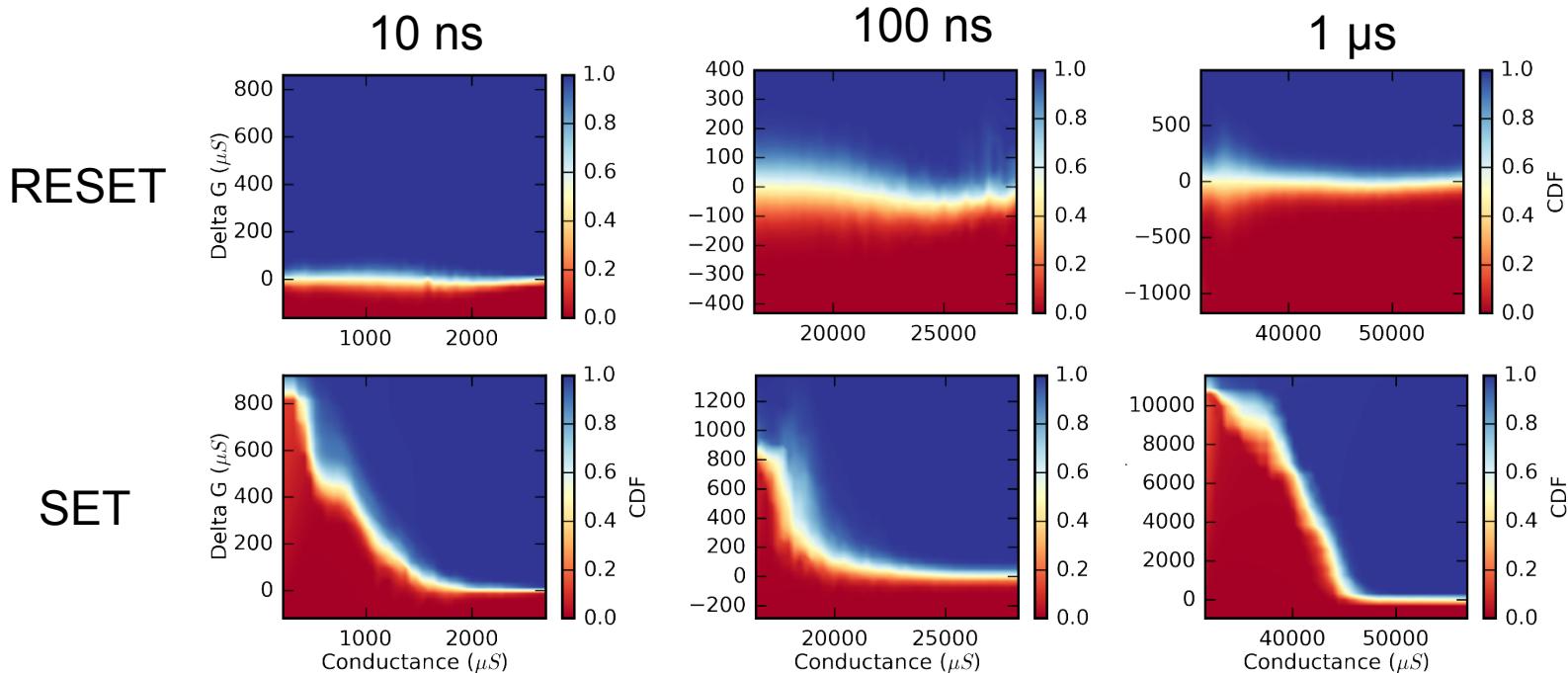
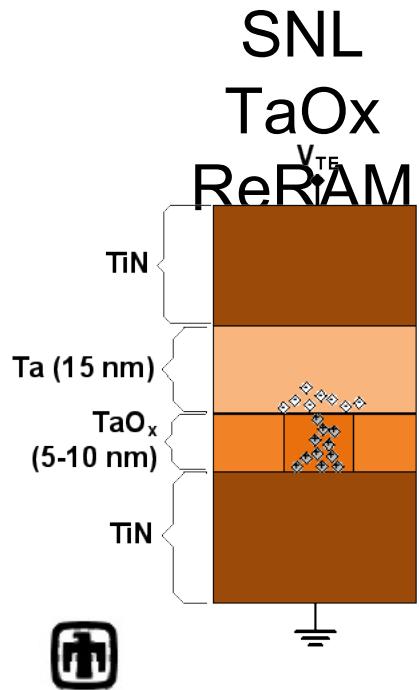
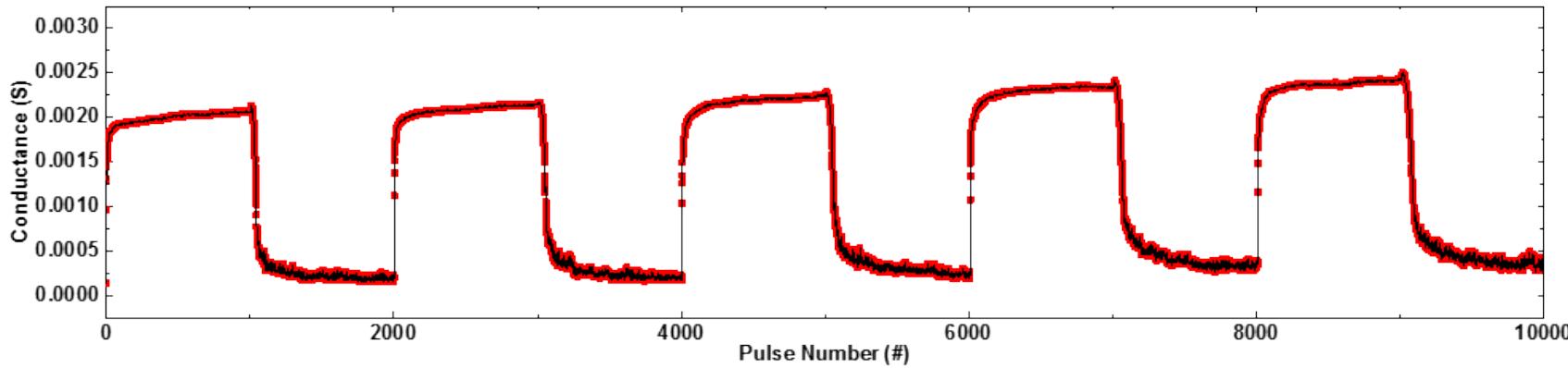
- Motivation and Background
- Analog In-Memory Compute Energy & Latency
- Devices for Accurate Inference
- Devices for Accurate Training
- Conclusions

# Device Challenges for Training

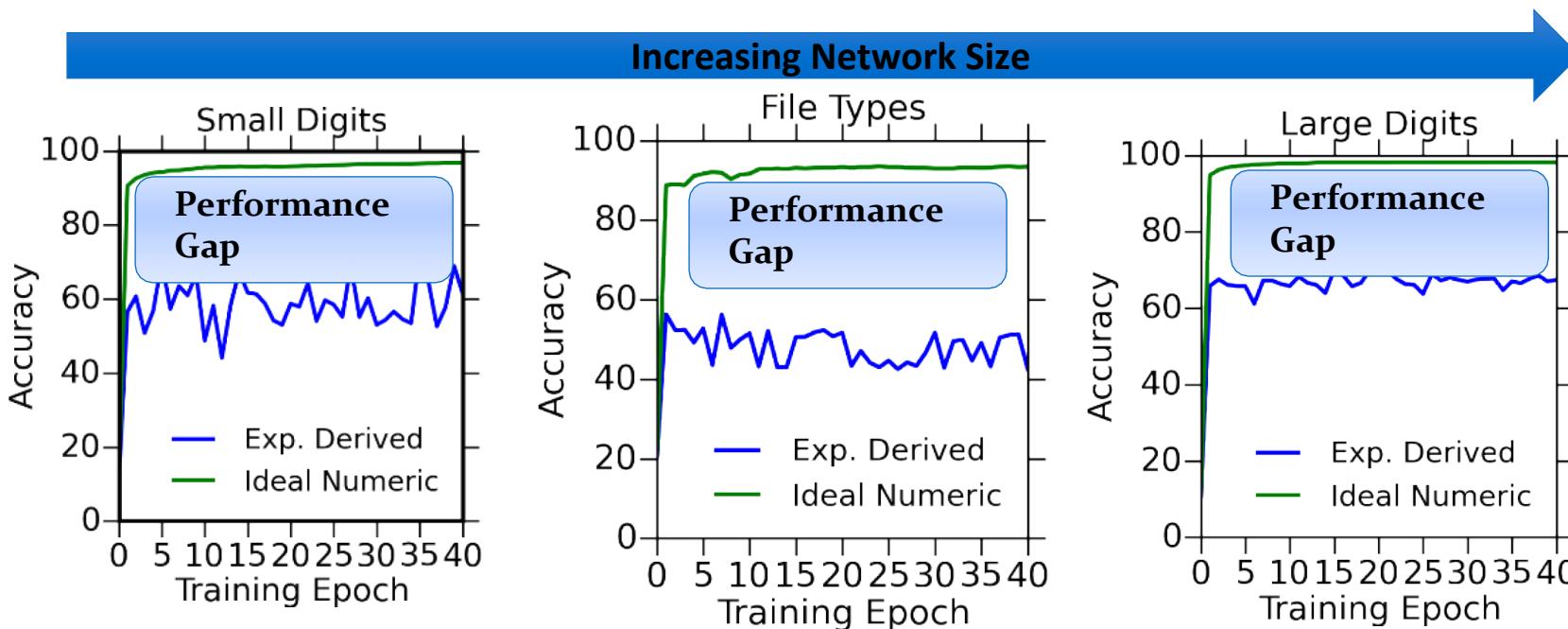
- Training has an overlapping set of challenges
- Ideally weight increases and decreases linearly proportional to learning rule result
- Issue for open loop nonvolatile memory: altered the relationship between intended and actual update
  - Nonlinear and asymmetric state change
  - Cycle to cycle random variability (write stochasticity)
  - Device to device random variability
- Also: very high endurance ( $>10^{12}$ )



# Characterization for Training



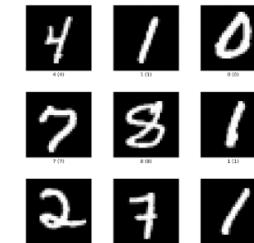
# Initial TaOx ReRAM Training Accuracy Modeling (MNIST)



#ROSS SIM

## Unexpected results:

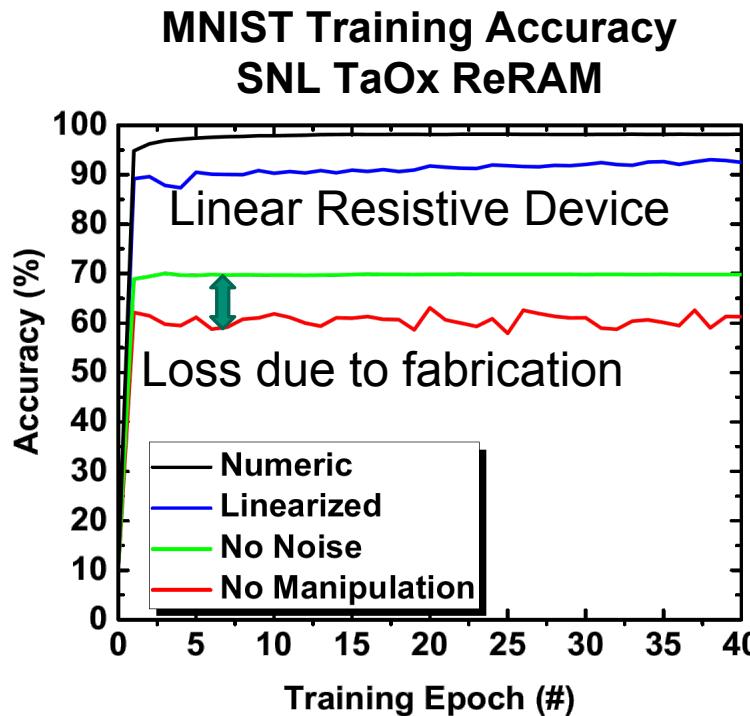
**TaOx ReRAM was not ideal device for open loop training**



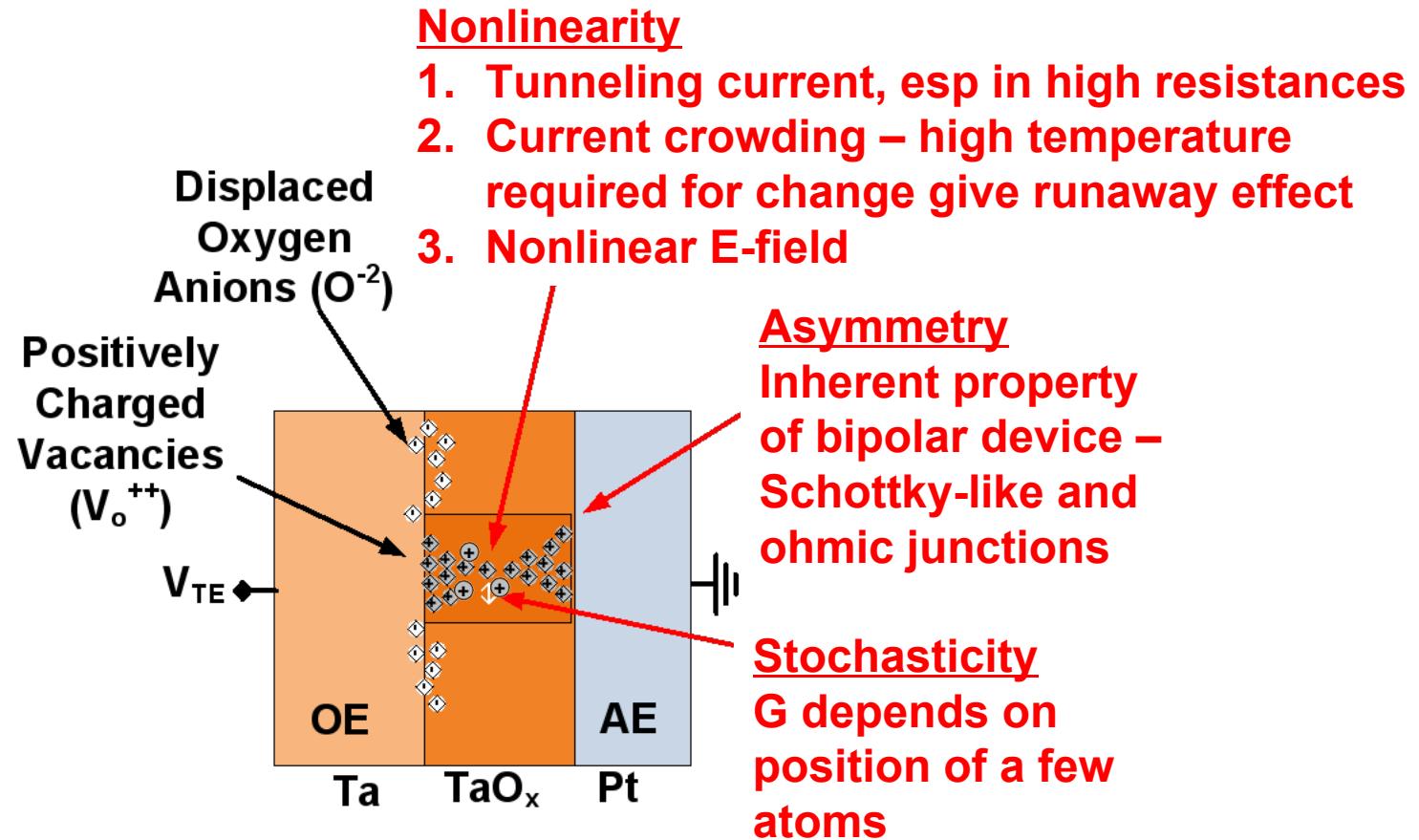
Training is significantly more challenging than inference!



# Physical Insight from Multiscale Model - CrossSim Challenges using Filamentary ReRAM for Training



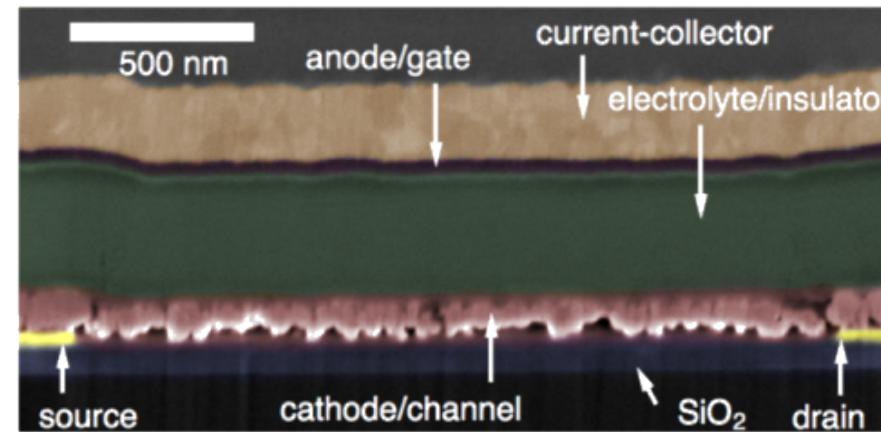
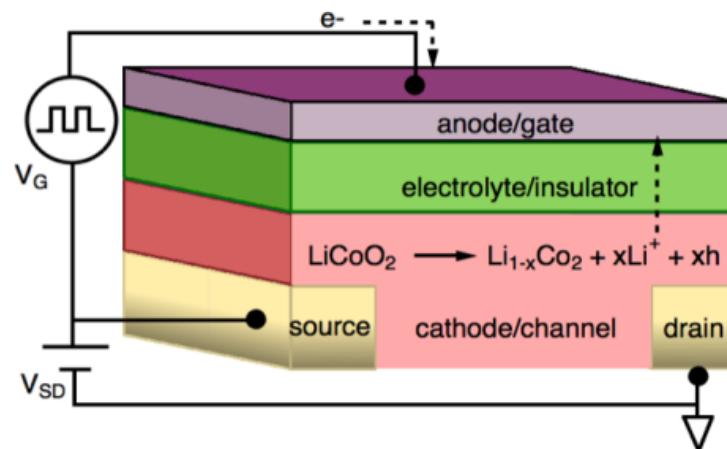
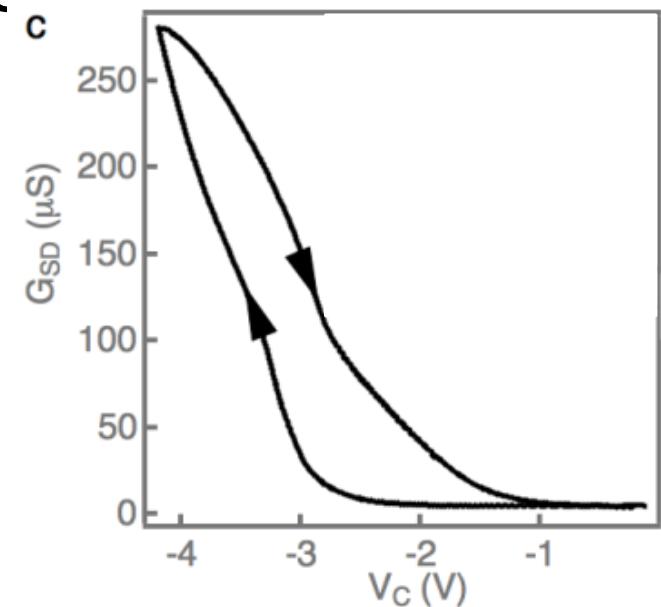
**#ROSS SIM**



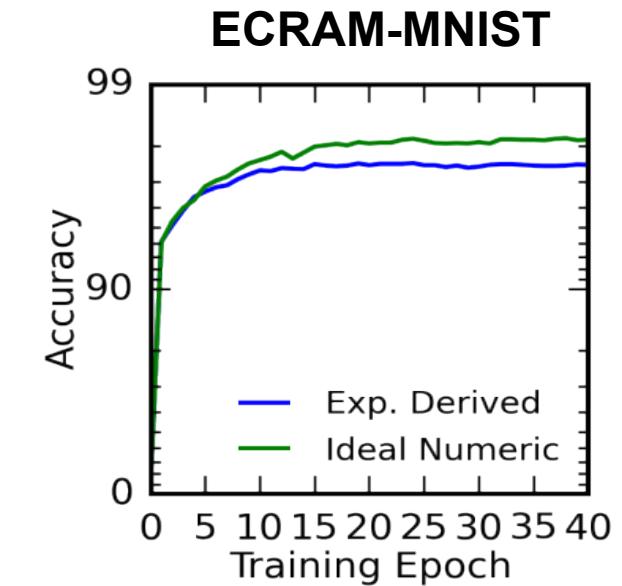
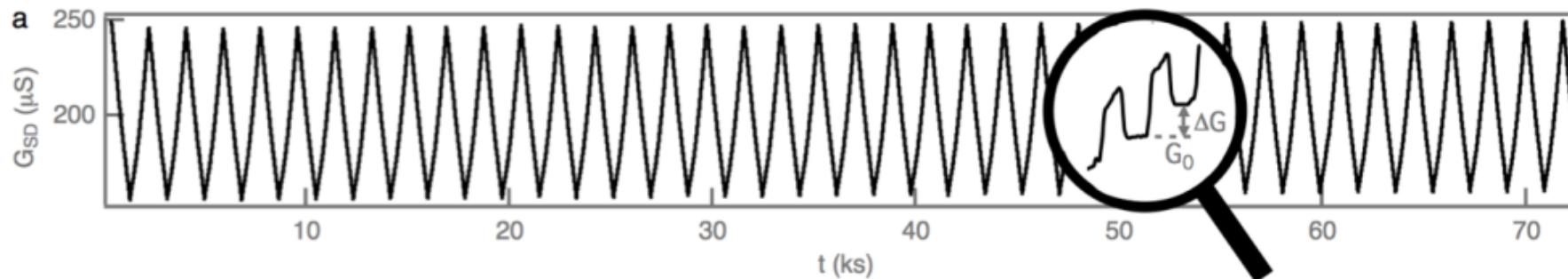
# Electrochemical RAM (ECRAM) Synapse

- As we used codesign to understand the challenges with ReRAM, Sandia was exploring doping modulation of Lithium battery cathode
- Novel device discovery: resistivity across cathode changes linearly with battery charge/discharge
- Battery can function as an analog nonvolatile transistor!

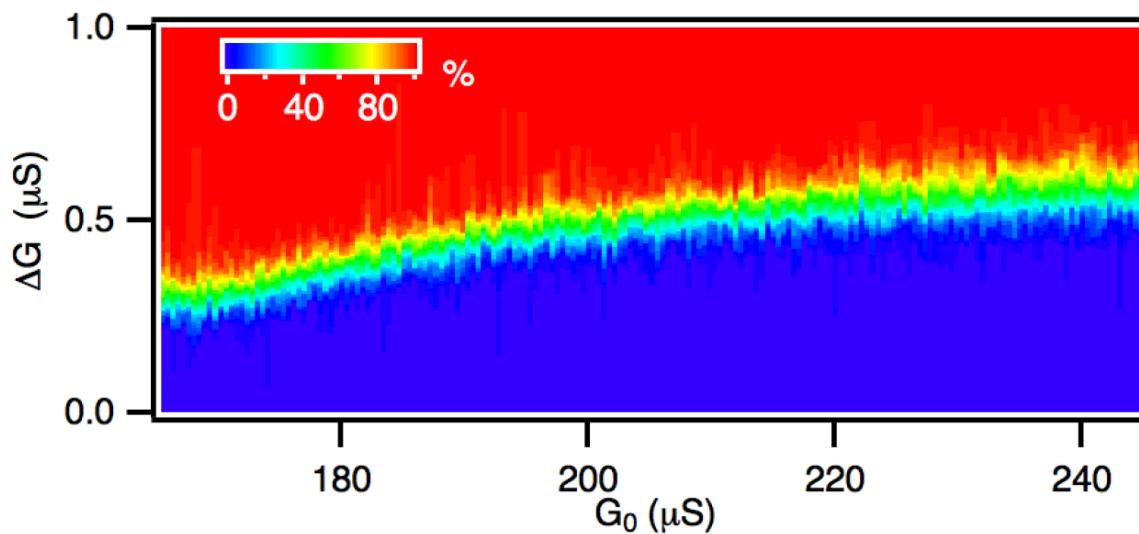
Conductance vs Voltage



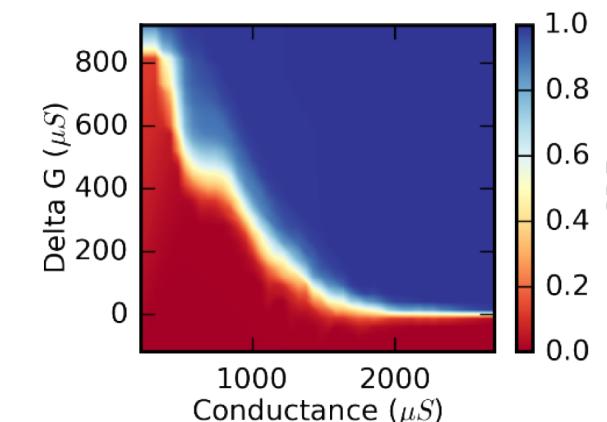
# Analog State Comparison



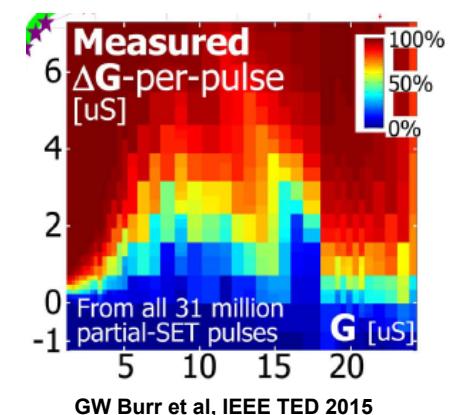
### ECRAM



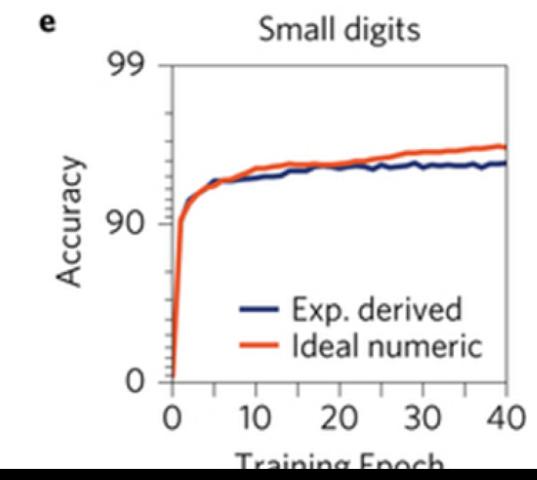
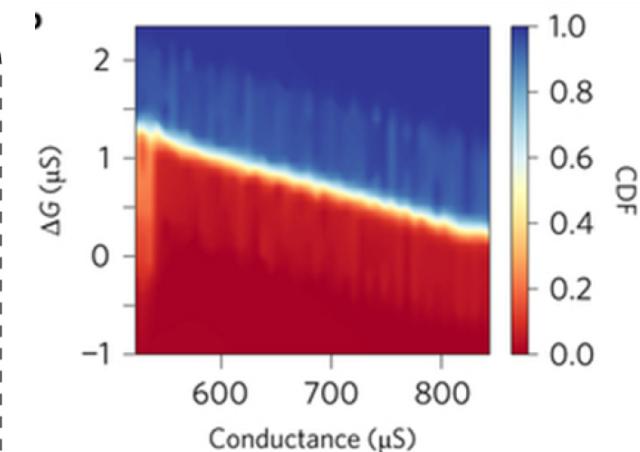
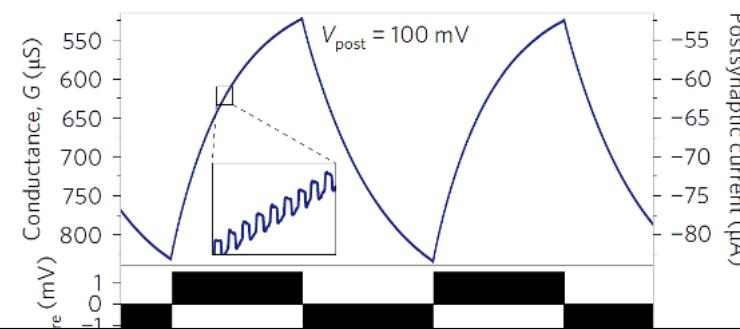
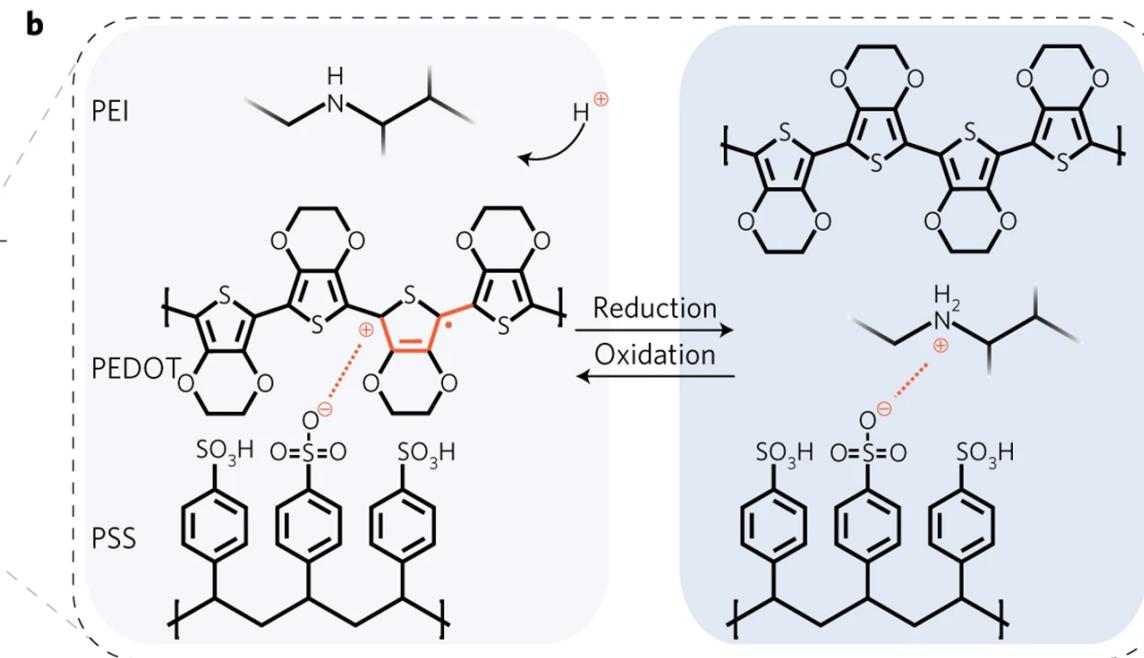
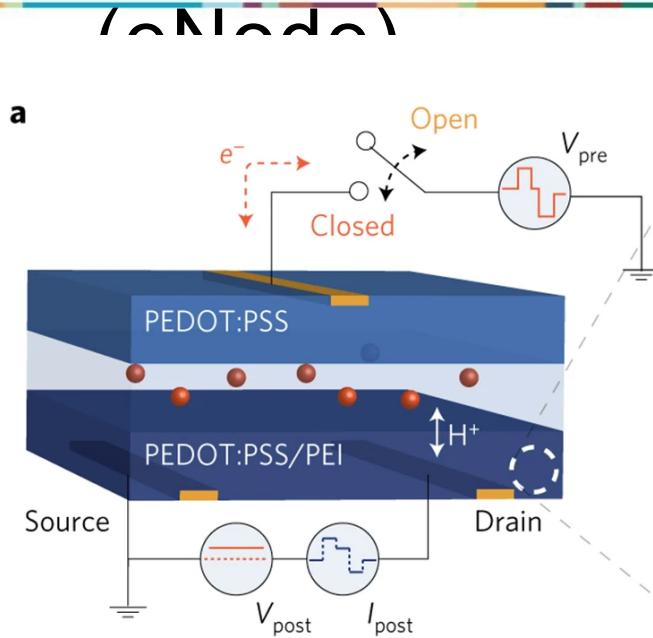
### TaOx ReRAM



### PCM Array



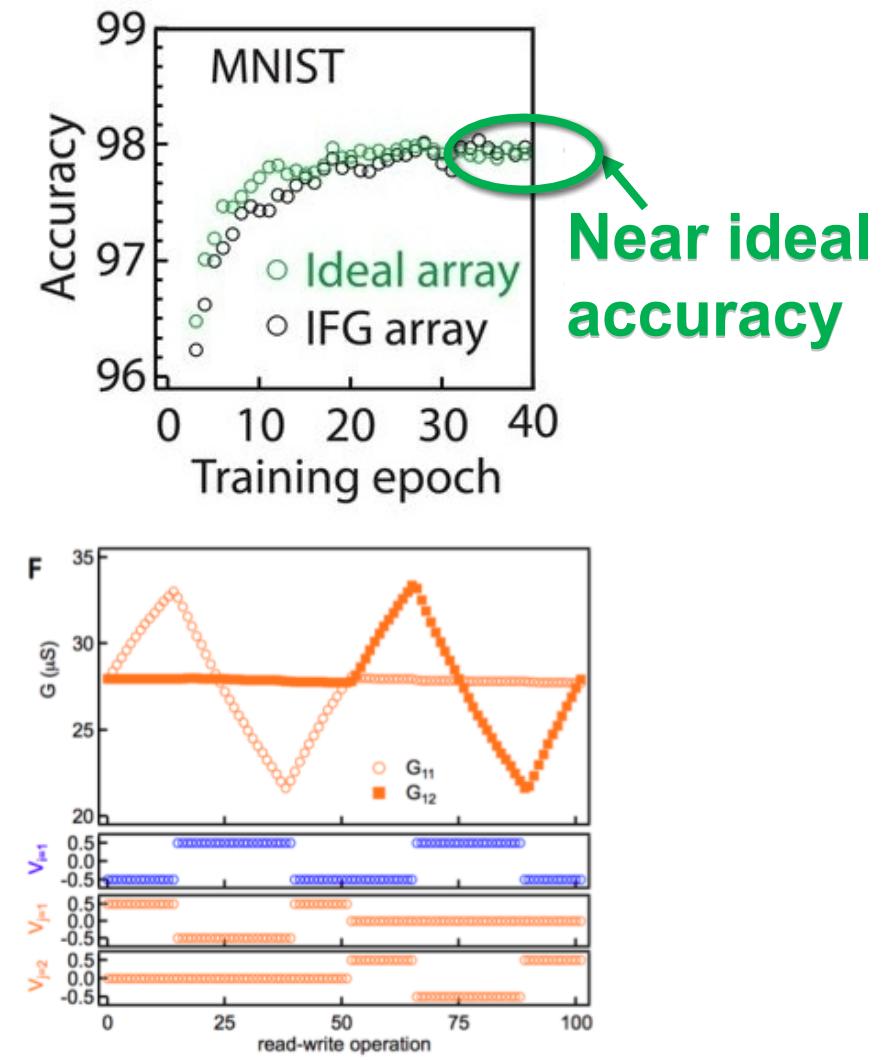
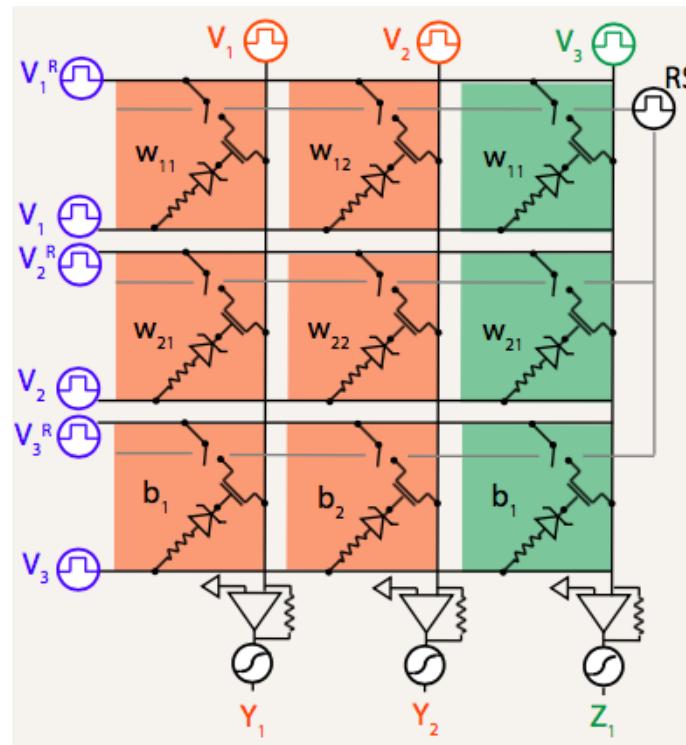
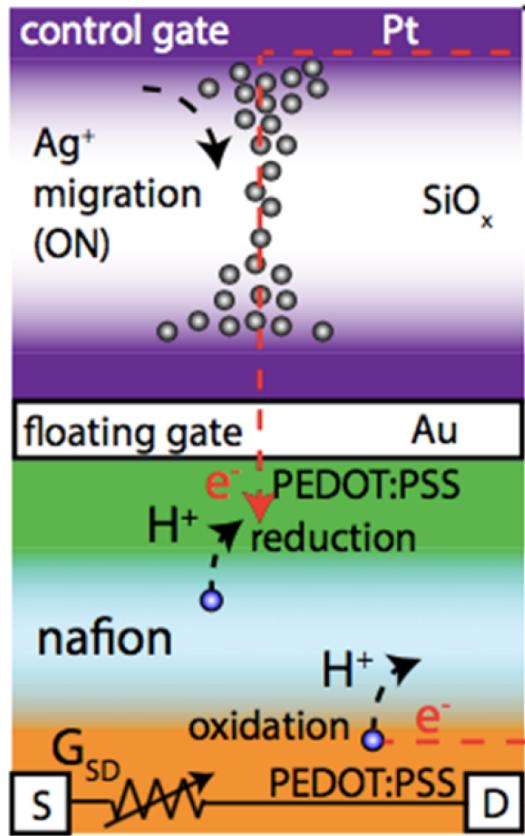
# Electrochemical Neuromorphic Organic Device



Proton-based polymer ECRAM synapse: fast, better endurance



# ECRAMs Array Parallel Update Training Demonstration



# Outline

---

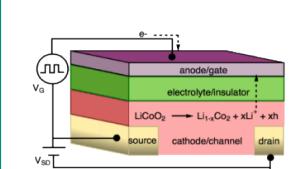
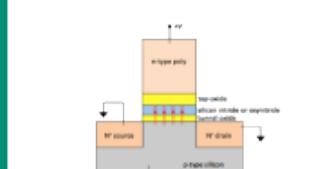
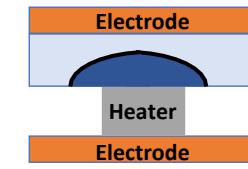
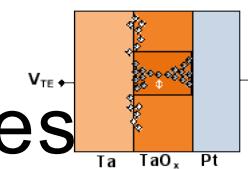
- Motivation and Background
- Analog In-Memory Compute Energy & Latency
- Devices for Accurate Inference
- Devices for Accurate Training
- Conclusions

# Requirements for Inference versus Training

Property	Inference	Training
Analog programing error (w/ write verify)	Critical	Less Important
Long term retention	Important	Less Important
Read noise	Important	Less Important
Conductance Range	Important	Important
Short term state drift	Important	Important
Device to device variability	Important	Important
Write stochasticity	Less Important	Important
Write speed	Less Important	Important
Write linearity	Less Important	Important
Write symmetry	Less Important	Critical
Endurance	Less Important	Critical



# Perspective: IMC Devices



Property	ReRAM	PCRAM	SONOS/FG	ECRAM
Analog programing error (w/ write verify)	😊	😐	😊	😊
Long term retention	😊	😊	😊	😐
Read noise	😊	😊	😊	😊
Conductance range	😐	😐	😊	😊
Short term state drift	😐	😐	😐	😐
Device to device variability	😢	😐	😊	😊
Write stochasticity	😢	😢	😊	😊
Write speed	😊	😊	😢	😐
Write linearity	😢	😢	😊	😊
Write symmetry	😢	😢	😊	😊
Endurance	😐	😐	😐	😐



Inference ☺

Inference ☺

Training ☺

# Final Thoughts

---

- Traditional digital CMOS computing is hitting disruptive roadblocks for continuing energy efficiency
- Analog In Memory Computing offers path to  $>10$  TOPS/W
  - Idea for deep neural nets/convolutional nets
- Analog In Memory Computing has significant new challenges
  - Algorithm accuracy depends on the device
  - Inference and training have distinct challenges, with some overlap.
  - Inference: excellent behavior predicted with commercial SONOS and ReRAM
  - Training: more challenging, future devices such as ECRAM and related nonfilamentary devices may provide a path forward

# Thank You – Questions?

---



# Acknowledgements

## Sandia Contributors

Patrick Xiao

Chris Bennett

Will Wahby

Sapan Agarwal

Alec Talin

Robin Jacobs-Gedrim

David Hughart

Elliot Fuller

Ben Feinberg



**Hewlett Packard  
Enterprise**



## External Collaborators

Jean Anne Incorvia, UT

Stan Williams, TAMU

Hugh Barnaby, ASU

Alberto Salleo, Stanford

Yiyang Li, U Michigan

Helmut Puchner, Infineon

Vineet Agarwal, Infineon

John Paul Strachan, HPE

Victor Zhirnov, SRC

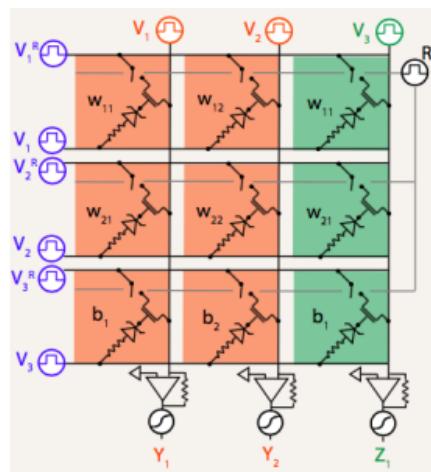
Jesse Mee, AFRL



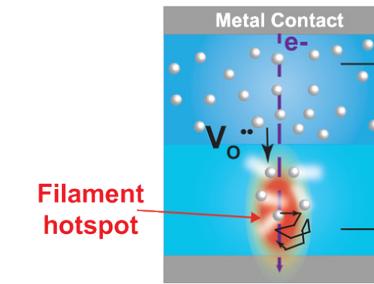
Semiconductor  
Research  
Corporation

# Novel Devices for Accurate Inference/Training

- Battery Inspired Devices
  - **Bulk ReRAM**
  - Bulk nanoionic devices
  - Two terminal charge tunnel junction
  - Organic and bio compatible switches
- Linear write:  $dG$  does not depend on  $G$ .
  - Ideal for training

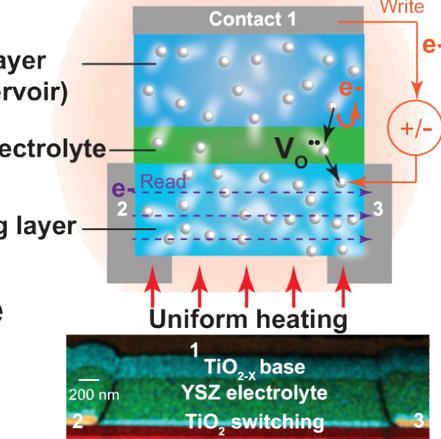


**a** Filamentary-RRAM  
Probabalistic & stochastic

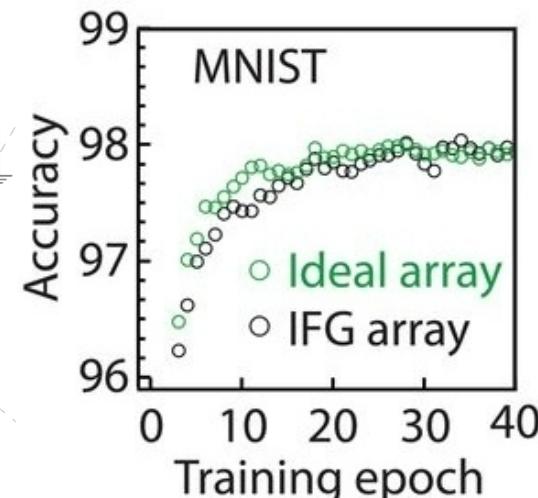
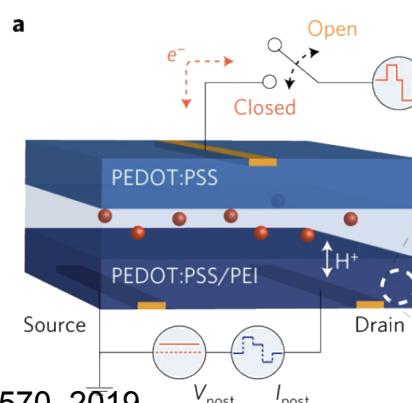


**d**

**Bulk-RRAM**  
Statistical & deterministic

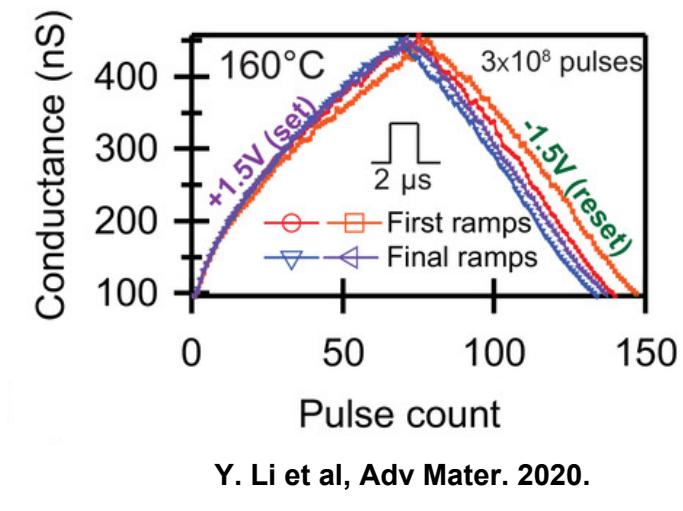


**e**



E. J. Fuller et al, *Science* 364, 570, 2019.

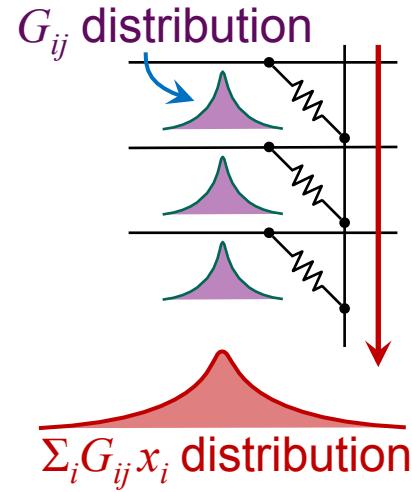
Y. Li et al, *Adv Mater.* 2020.



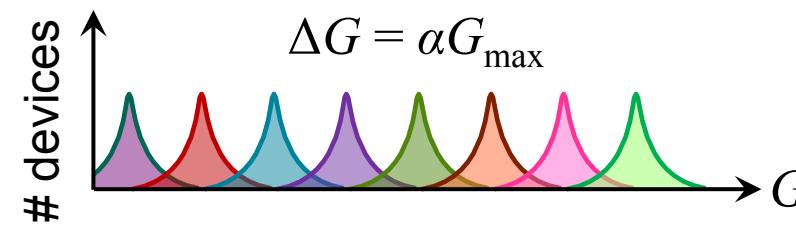
Y. Li et al, *Adv Mater.* 2020.

# Physical Weight Properties is Critical

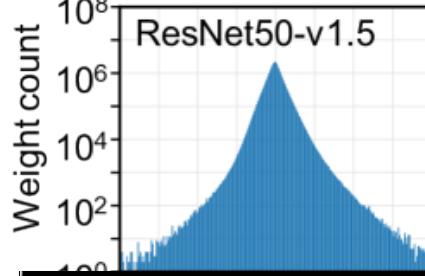
Error accumulation



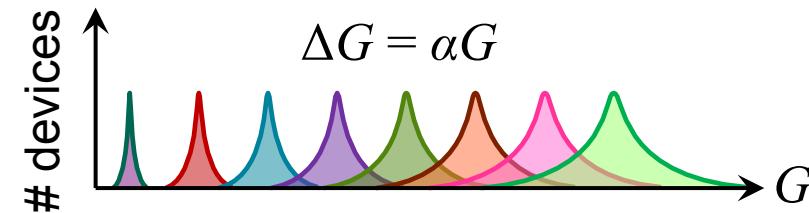
State-independent error



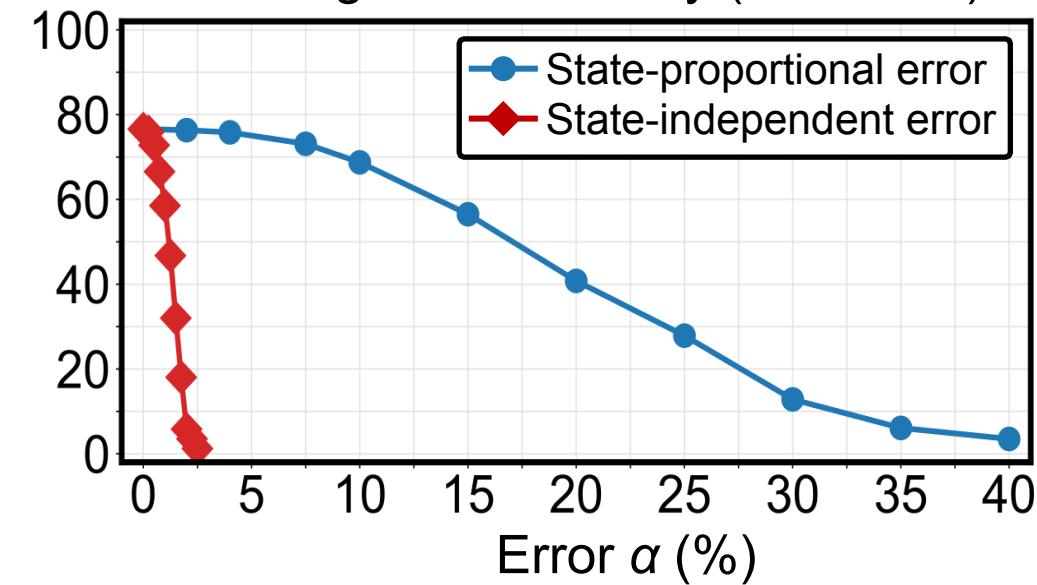
Typical Weight Distribution



State-proportional error



ImageNet Accuracy (7 bits/cell)

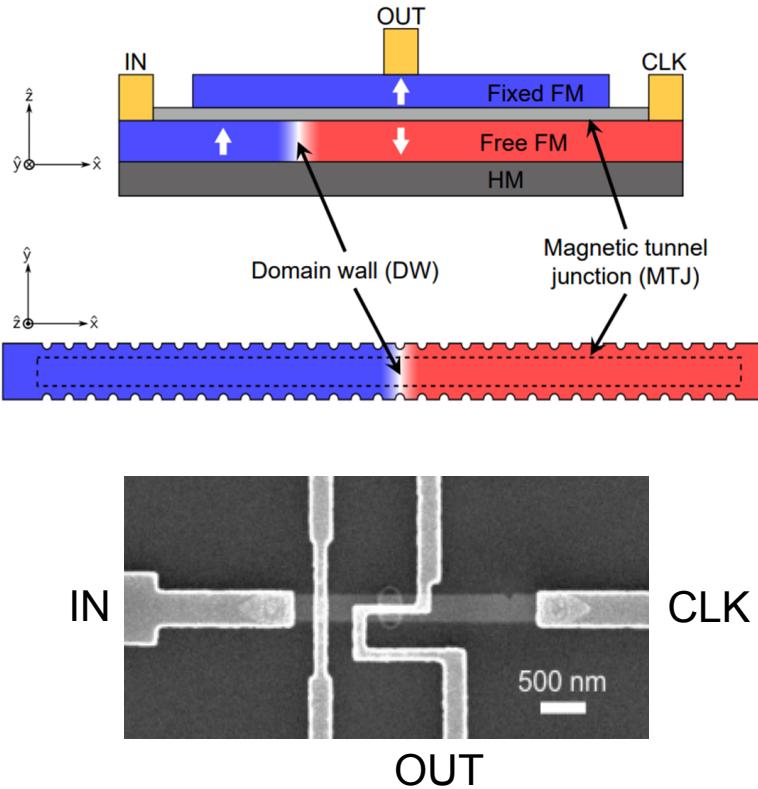


State-proportional error is ideal

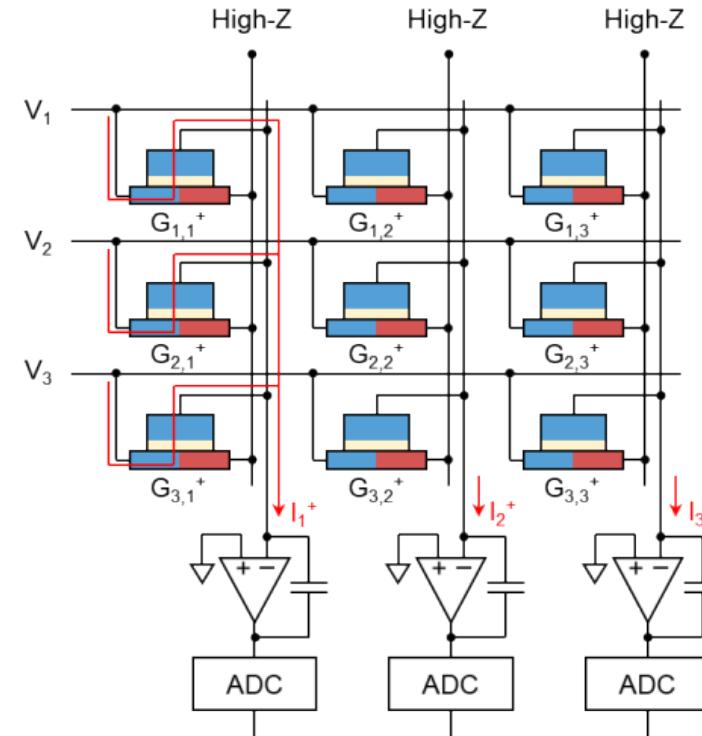


# Exploration of Novel Magnetic Synapses for Training

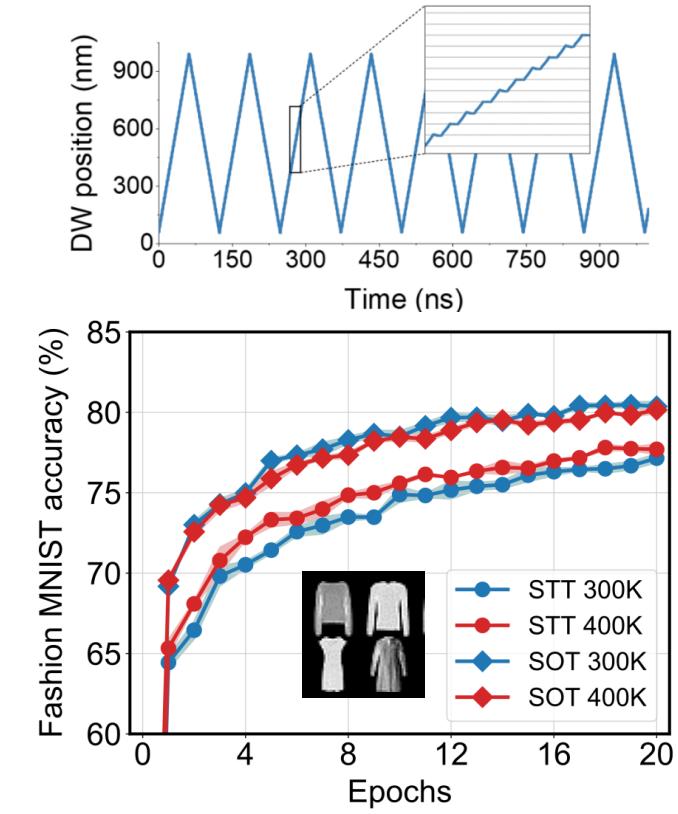
## Low-energy domain wall synapses



## Integration in a crossbar array



## Demonstrating linear updates and neural network accuracy



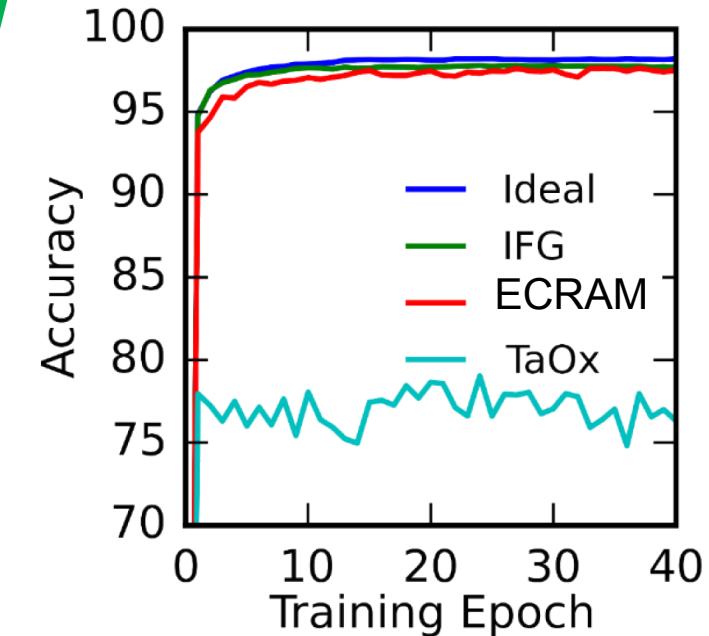
# Training Accuracy and Tile Energy/Summary

Codesign to Model Performance & Energy

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

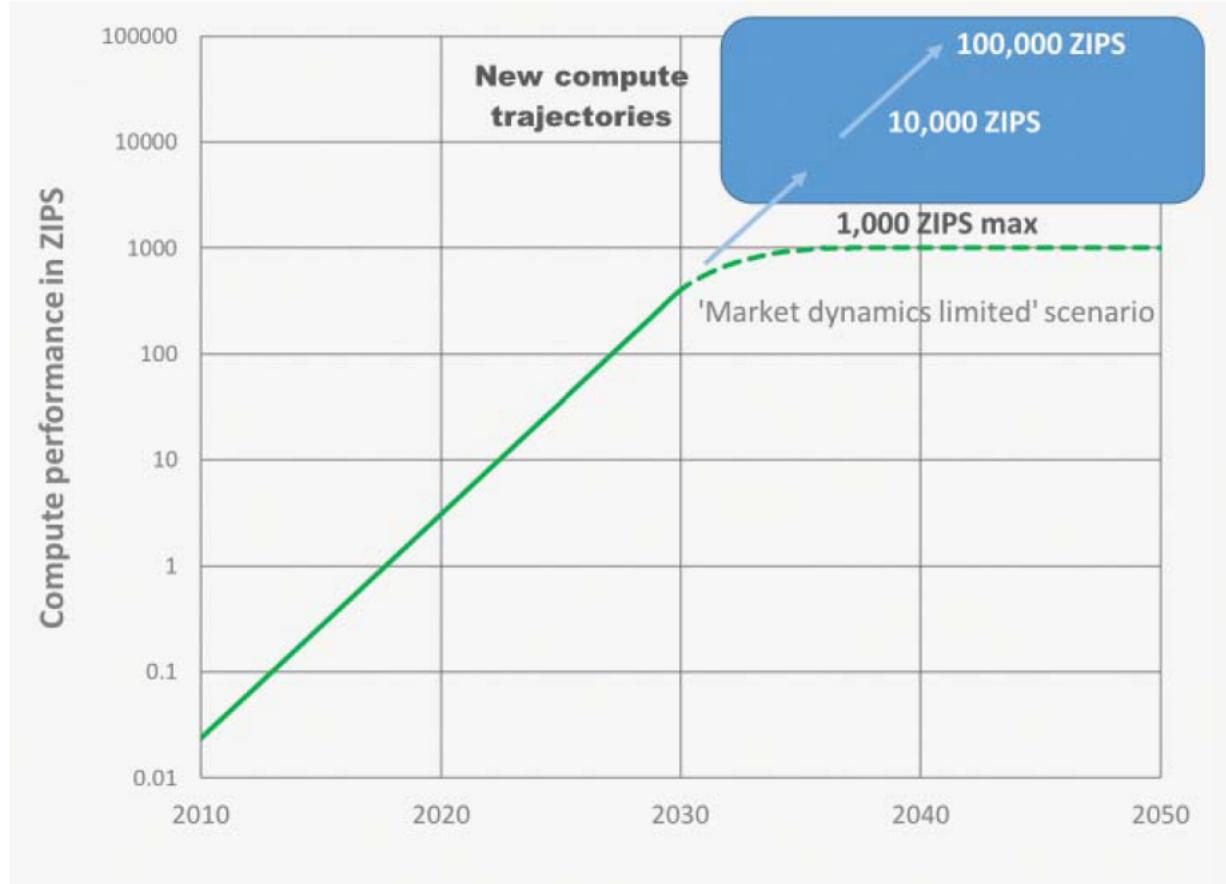
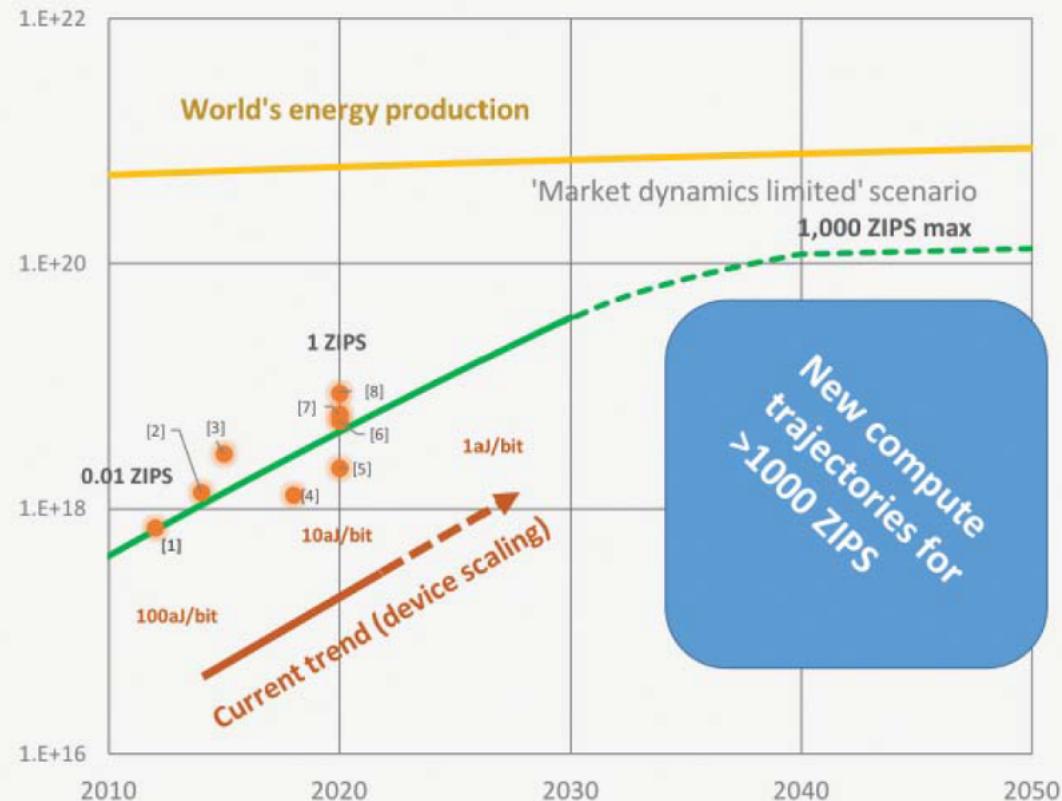
**SONOS: While accuracy, program  
is slow: use for inference**

**ECRAM: Use for training  
& inference**

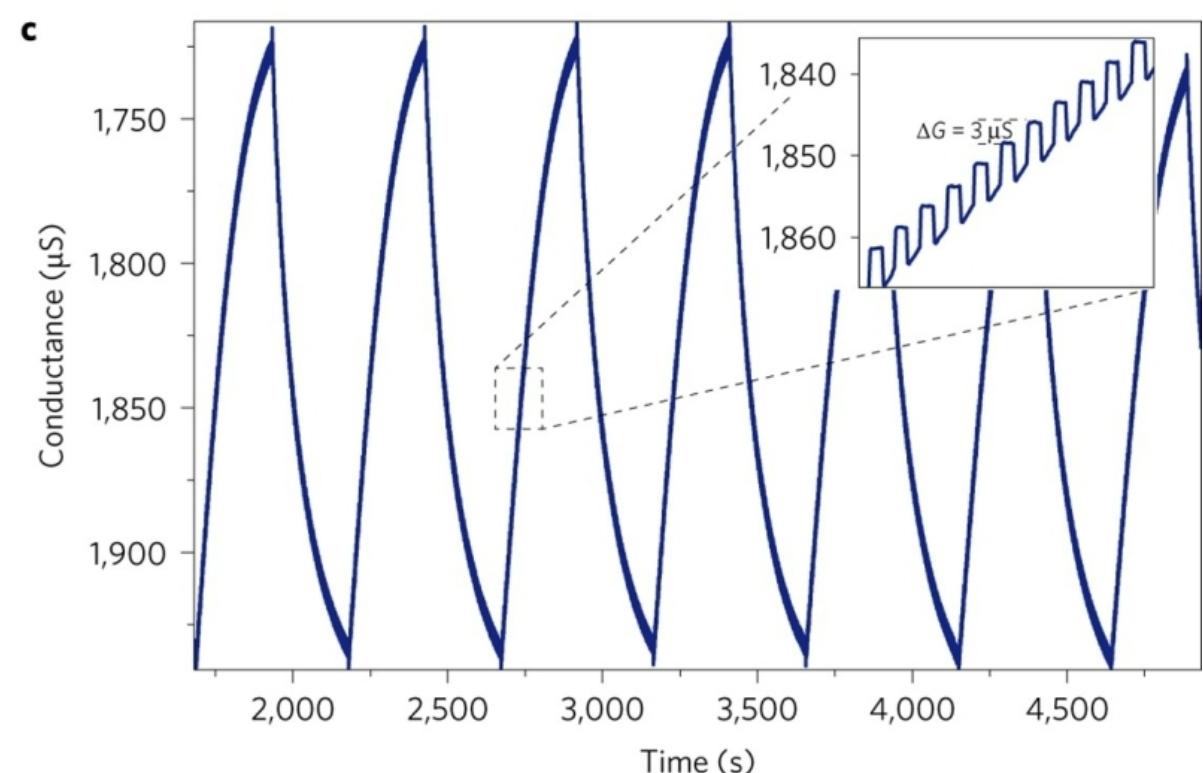
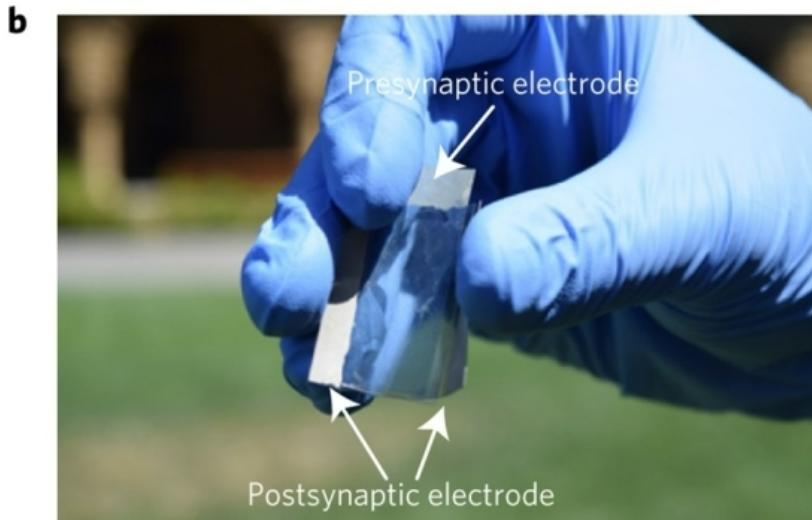
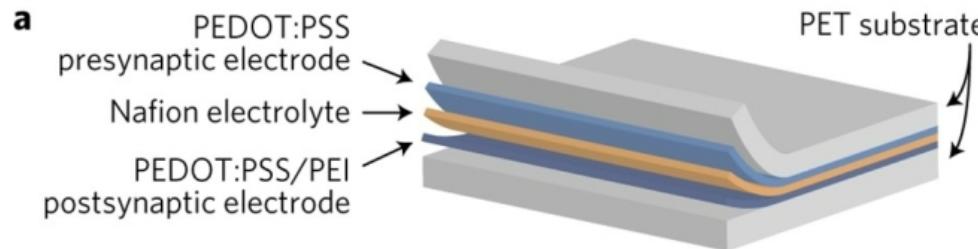


**ReRAM: Training is not  
accurate: better for  
inference**

# View from the Semiconductor Industry



# Flexible eNode



# Comparison of State of the Art Accelerators

TABLE II. Comparison of selected digital and mixed-signal neural network inference accelerators from industry and research.<sup>a</sup> TOPS: Tera-Operations per second. We have counted MACs as single operations where possible. Note that performance (TOPS) is measured at the specified level of weight and activation precision, which differs between accelerators. The results for NVIDIA T4, TPU, Goya, UNPU, and Ref. 122 are measured; others are simulated. TOPS/mm<sup>2</sup> values are based on the die area, where provided.

	NVIDIA T4 <sup>175</sup>	Google TPU v1 <sup>22,b</sup>	Habana Goya HL-1000 <sup>176</sup>	DaDianNao <sup>44</sup>	UNPU <sup>51</sup>	Reference 122 mixed-signal <sup>c</sup>
Process	12 nm	28 nm	16 nm	28 nm	65 nm	28 nm
Activation resolution	8-bit int	8-bit int	16-bit int	16-bit fixed-pt.	16 bits	1 bit
Weight resolution	8-bit int	8-bit int	16-bit int	16-bit fixed-pt.	1 bit <sup>d</sup>	1 bit
Clock speed	2.6 GHz	700 MHz	2.1 GHz (CPU)	606 MHz	200 MHz	10 MHz
Benchmarked workload	ResNet-50 <sup>177</sup> (batch = 128)	Mean of six MLPs, LSTMs, CNNs	ResNet-50 (batch = 10)	Peak performance	Peak performance	Co-designed binary CNN (CIFAR-10)
Throughput (TOPS)	22.2, 130 (peak)	21.4, 92 (peak)	63.1	5.58	7.37	0.478
Density (TOPS/mm <sup>2</sup> )	0.04, 0.24 (peak)	0.06, 0.28 (peak)	...	0.08	0.46	0.10
Efficiency (TOPS/W)	0.32	2.3 (peak)	0.61	0.35	50.6	532

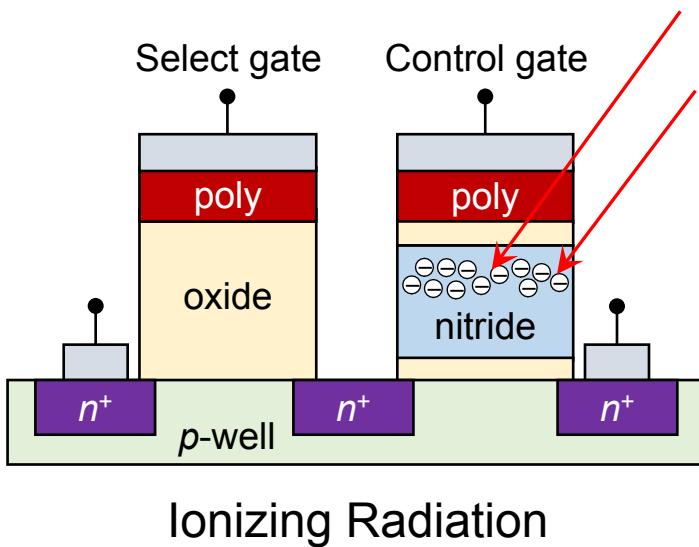
<sup>a</sup>To enable performance comparisons across a uniform application space, we did not consider accelerators for spiking neural networks.

<sup>b</sup>The TPU v2 and v3 chips, which use 16-bit floating point arithmetic, are commercially available for both inference and training on the cloud. MLPerf inference benchmarking results for the Cloud TPU v3 are available,<sup>179</sup> but power and area information is undisclosed. The TPU v1 die area is taken to be the stated upper bound of 331 mm<sup>2</sup>; the listed TOPS/mm<sup>2</sup> values are therefore a lower bound.

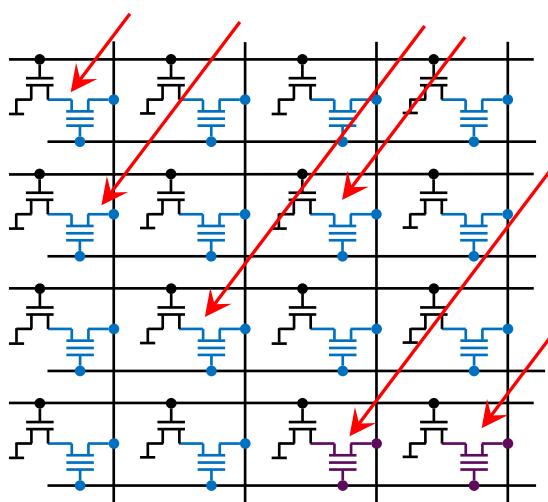
<sup>c</sup>The mixed-signal accelerator in Ref. 122 performs multiplication using digital logic and summation using analog switched-capacitor circuits.

<sup>d</sup>The UNPU architecture flexibly supports any weight precision from 1 to 16 bits. The results are listed for 1-bit weights.

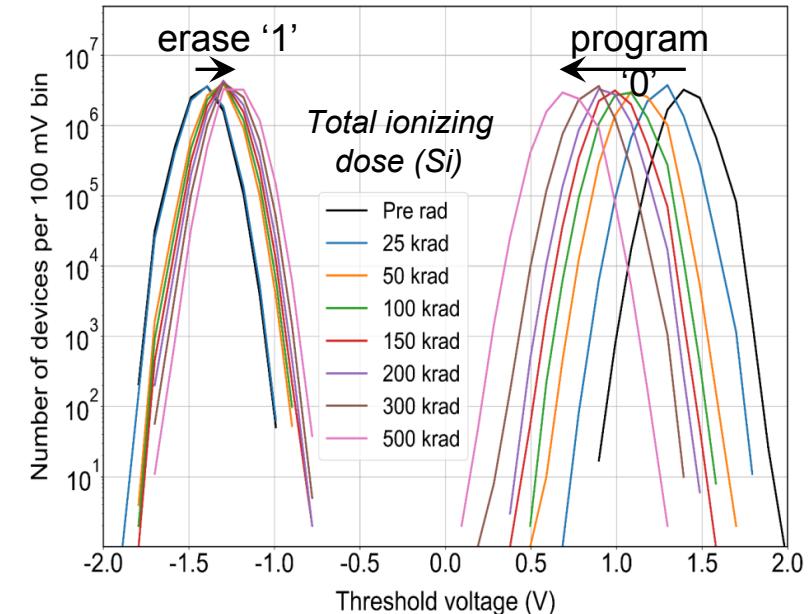
# Impact of Ionizing Radiation on Deep Net Accuracy



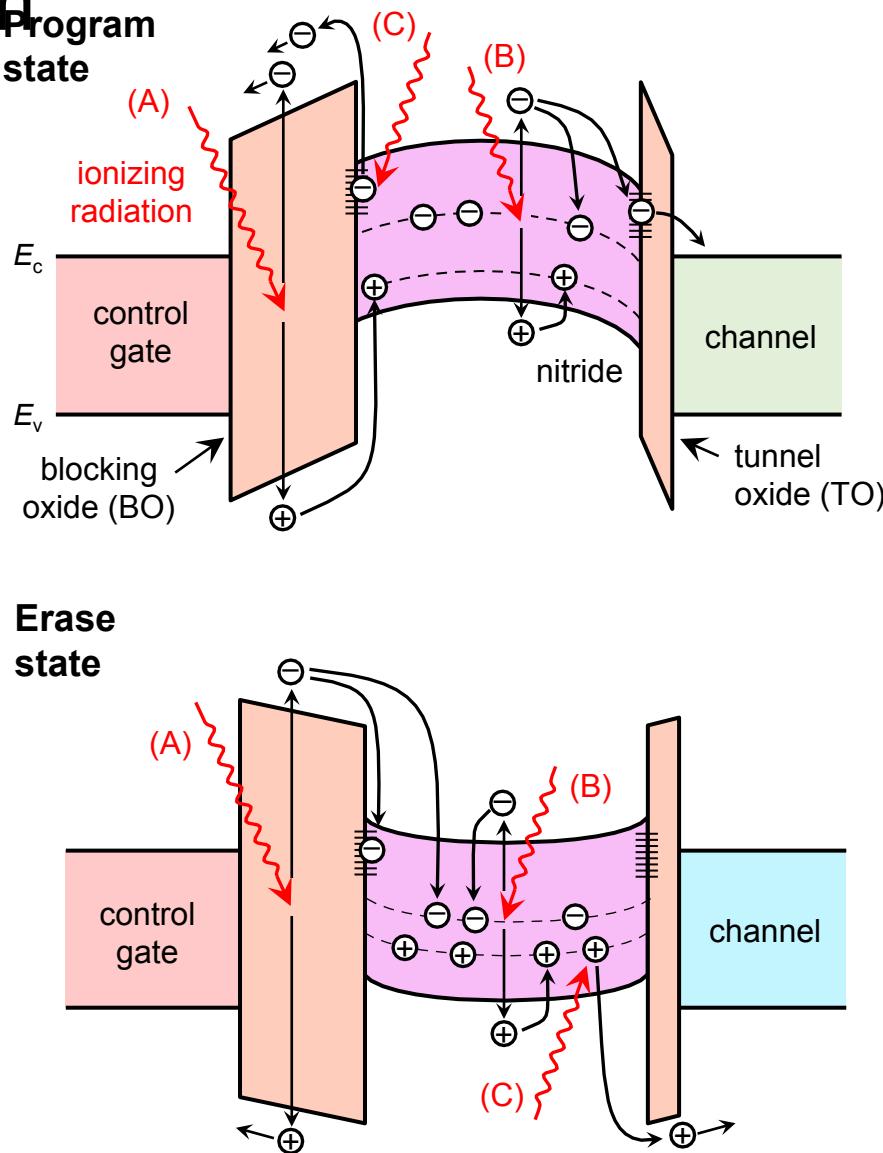
Uniform Gamma Irradiation



Threshold Distribution Shifts Across Array

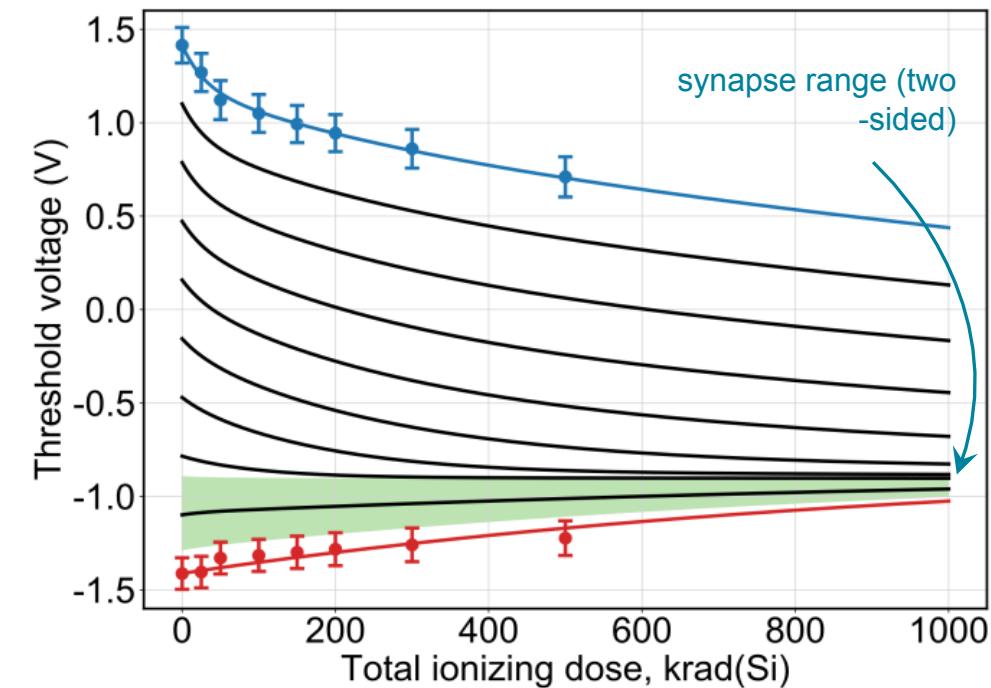


# Analog Neuromorphic SONOS In Space: Physics to Algorithm



TP Xiao et al, IEEE Trans Nuclear Sci, 2021 (in press).

**$V_T$  versus Total Ionizing Dose: Model and Experiment**

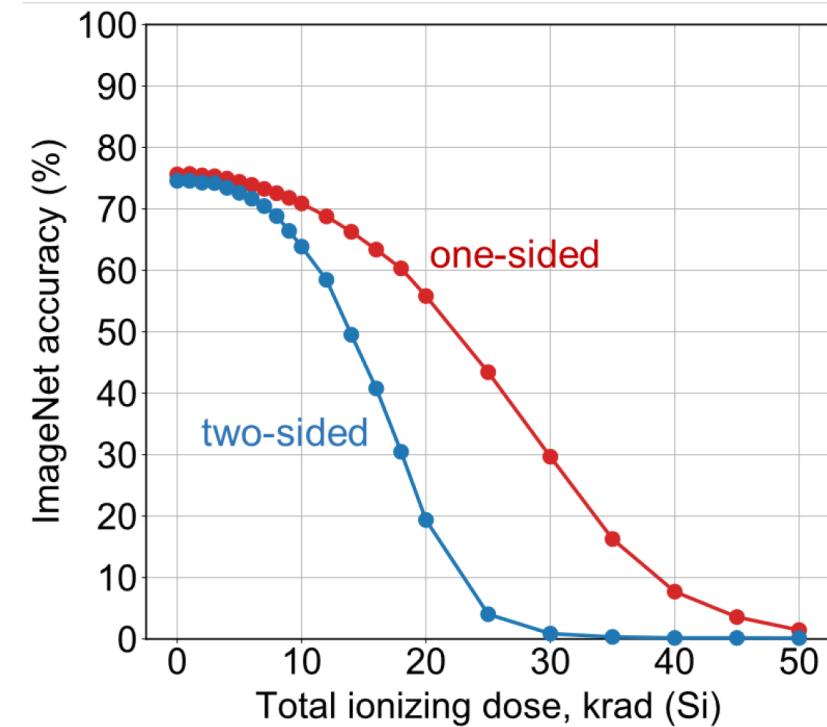
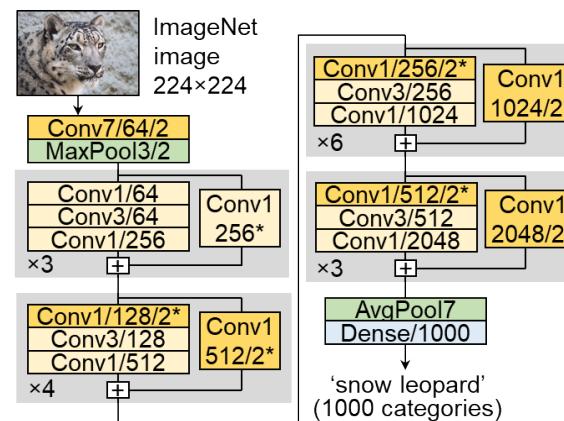
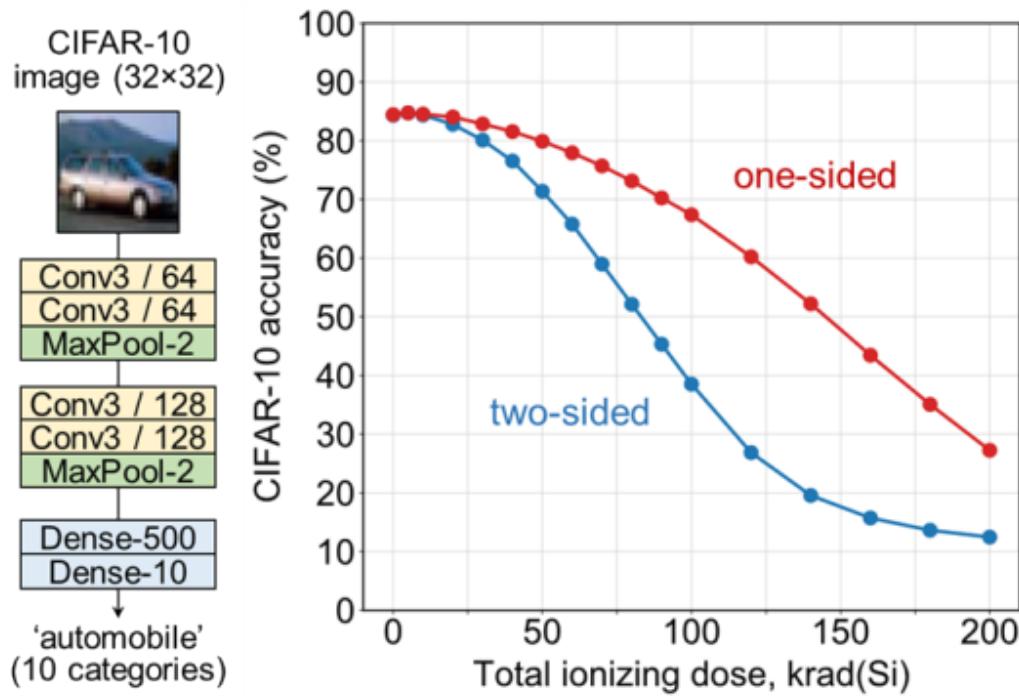


# Analog Neuromorphic SONOS In Space: Physics to Algorithm

How will the accuracy degrade in space?



**6-layer CNN for CIFAR-10**  
4.36M weights, 100.4M ops



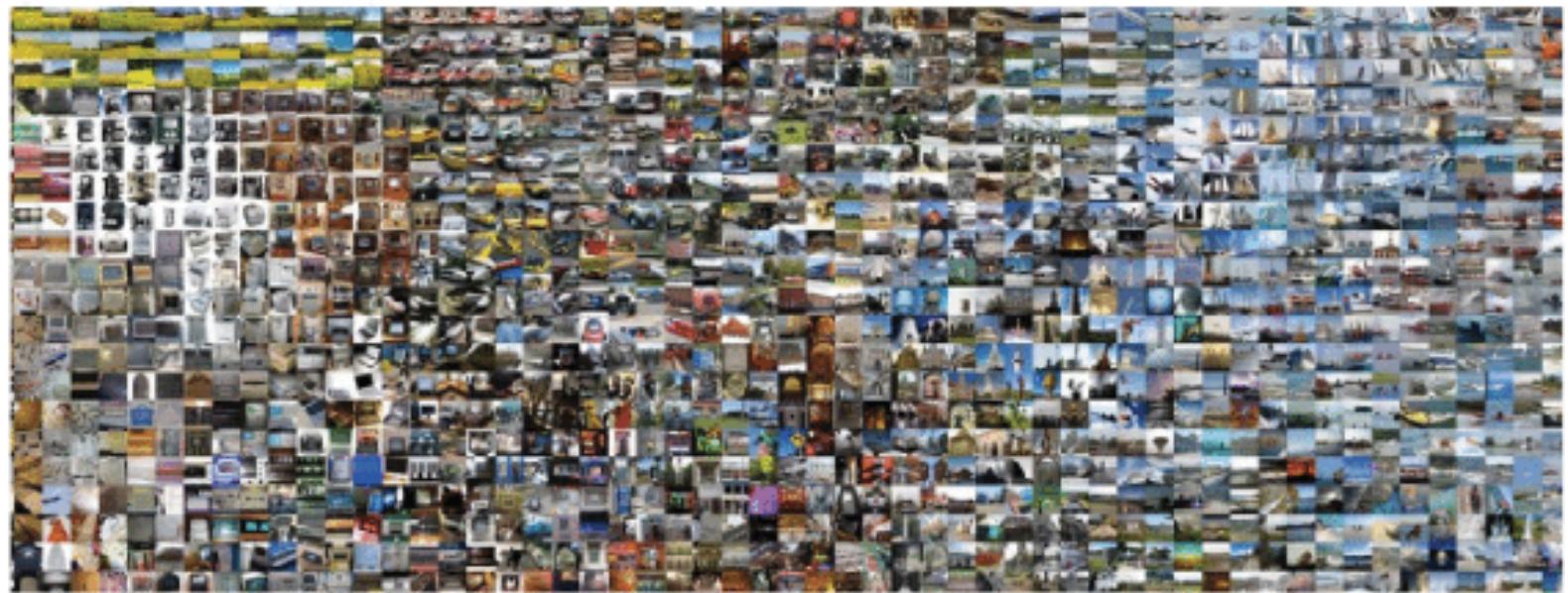
CoDesign provides insight for fielding neuromorphic devices

# Example Standard Visual Recognition Datasets

**MNIST**

3 6 8 1 7 9 6 6 9 1  
6 7 5 7 8 6 3 4 8 5  
2 1 7 9 7 1 2 8 4 6  
4 8 1 9 0 1 8 8 9 4  
7 6 1 8 6 4 1 5 6 0  
7 5 9 2 6 5 8 1 9 7  
1 2 2 2 2 3 4 4 8 0  
0 2 3 8 0 7 3 8 5 7  
0 1 4 6 4 6 0 2 4 3  
7 1 2 8 7 6 9 8 6 1

**ImageNet**



- 28x28 pixel grayscale
- 10 classes
- 60k training images
- 10k test images

- 256x256 pixel color
- 1000 classes
- 1.3M training images
- 100k test images



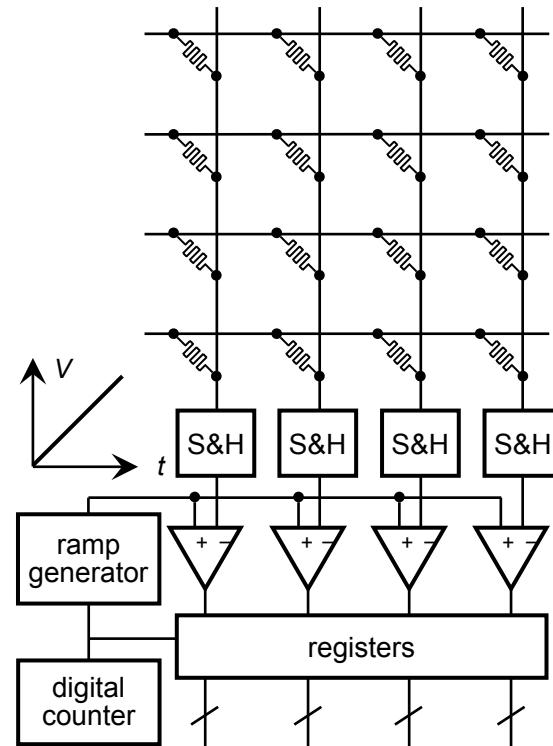
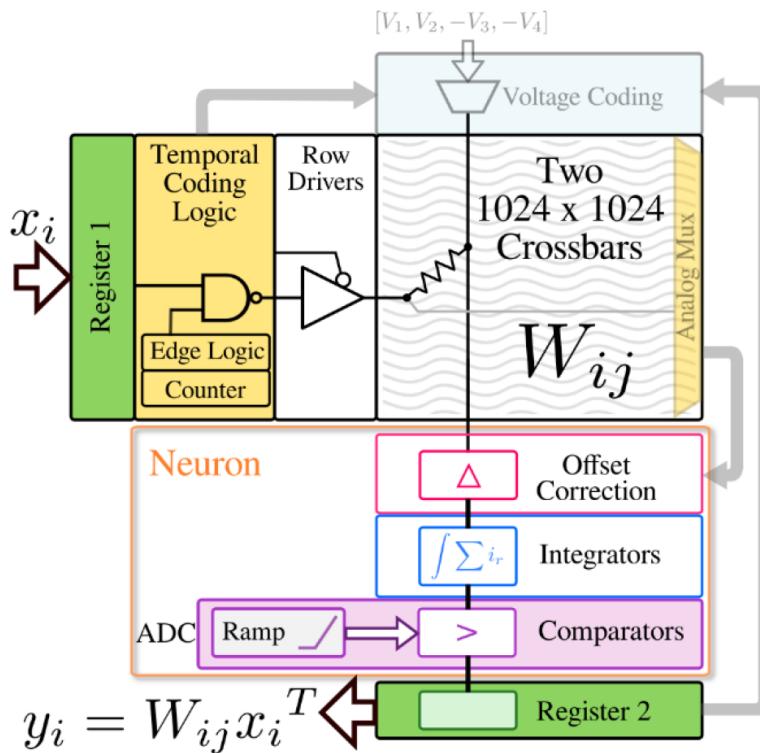
# How much computing needs to be done?

Metrics	LeNet 5	AlexNet	Overfeat fast	VGG 16	GoogLeNet v1	ResNet 50
<b>Top-5 error<sup>†</sup></b>	n/a	16.4	14.2	7.4	6.7	5.3
<b>Top-5 error (single crop)<sup>†</sup></b>	n/a	19.8	17.0	8.8	10.7	7.0
<b>Input Size</b>	$28 \times 28$	$227 \times 227$	$231 \times 231$	$224 \times 224$	$224 \times 224$	$224 \times 224$
<b># of CONV Layers</b>	2	5	5	13	57	53
<b>Depth in # of CONV Layers</b>	2	5	5	13	21	49
<b>Filter Sizes</b>	5	3,5,11	3,5,11	3	1,3,5,7	1,3,7
<b># of Channels</b>	1, 20	3-256	3-1024	3-512	3-832	3-2048
<b># of Filters</b>	20, 50	96-384	96-1024	64-512	16-384	64-2048
<b>Stride</b>	1	1,4	1,4	1	1,2	1,2
<b>Weights</b>	2.6k	2.3M	16M	14.7M	6.0M	23.5M
<b>MACs</b>	283k	666M	2.67G	15.3G	1.43G	3.86G
<b># of FC Layers</b>	2	3	3	3	1	1
<b>Filter Sizes</b>	1,4	1,6	1,6,12	1,7	1	1
<b># of Channels</b>	50, 500	256-4096	1024-4096	512-4096	1024	2048
<b># of Filters</b>	10, 500	1000-4096	1000-4096	1000-4096	1000	1000
<b>Weights</b>	58k	58.6M	130M	124M	1M	2M
<b>MACs</b>	58k	58.6M	130M	124M	1M	2M
<b>Total Weights</b>	60k	61M	146M	138M	7M	25.5M
<b>Total MACs</b>	341k	724M	2.8G	15.5G	1.43G	3.9G
<b>Pretrained Model Website</b>	[56] <sup>‡</sup>	[57, 58]	n/a	[57-59]	[57-59]	[57-59]

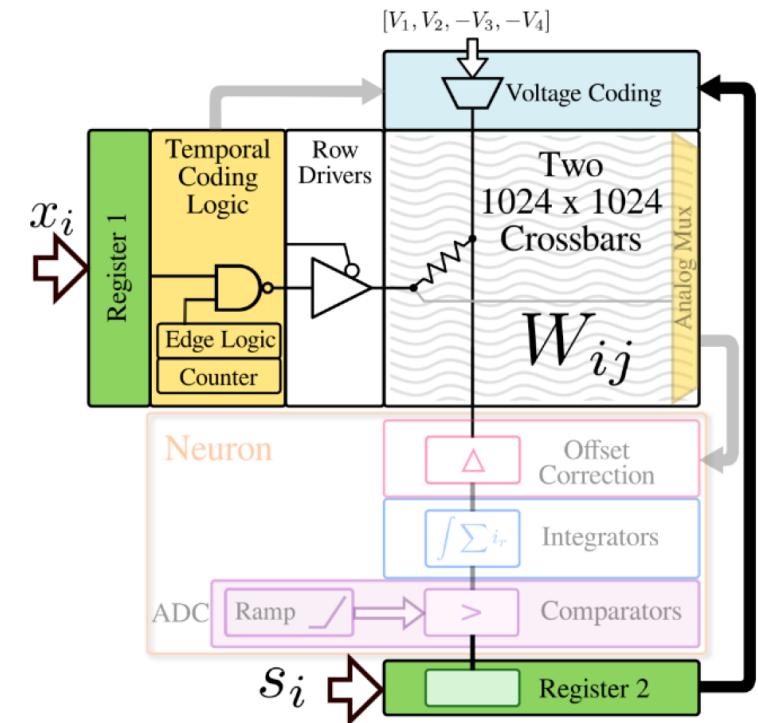


# Key Circuit Block/Kernel Analysis

## Vector Matrix Multiply (Inference)

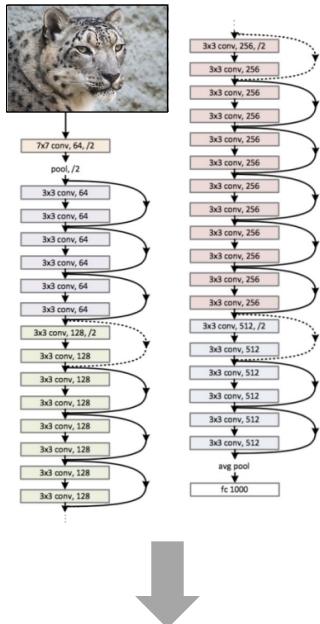


## Rank-1 Update (Training)

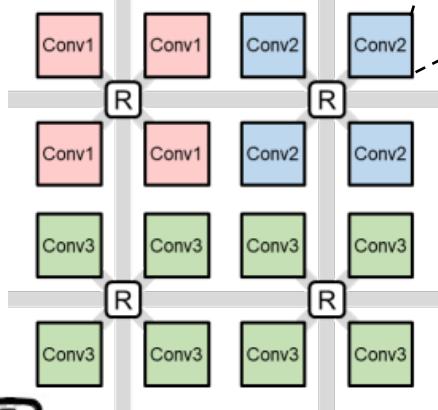


# Neural Network Inference Architecture

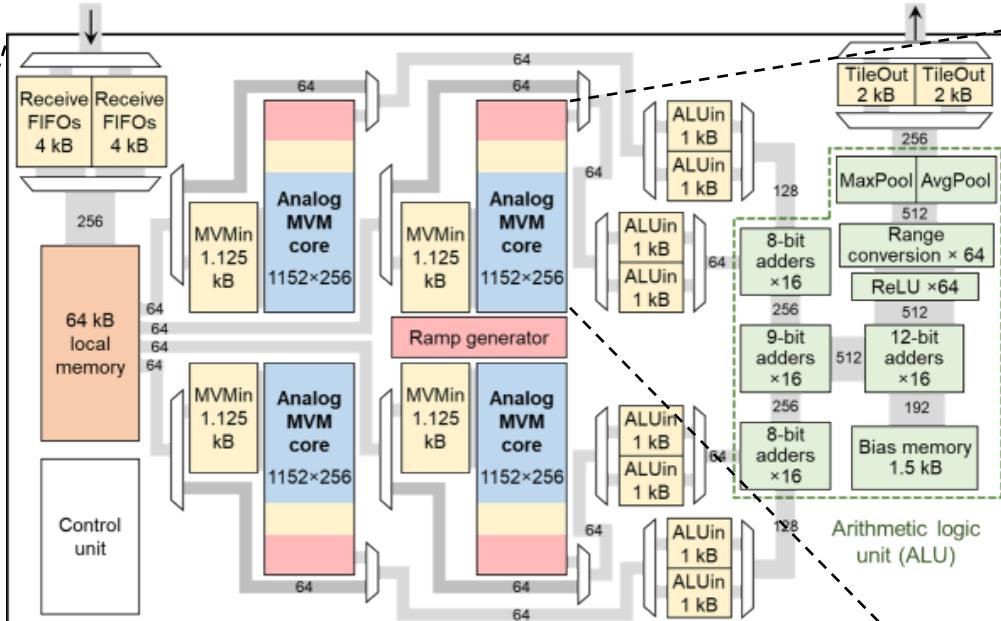
## Neural network



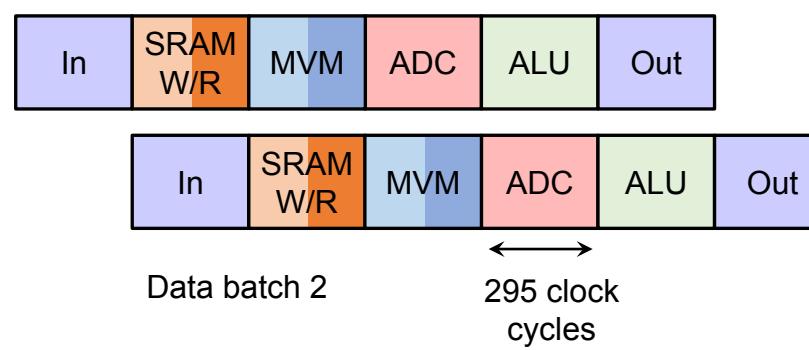
## Mesh architecture



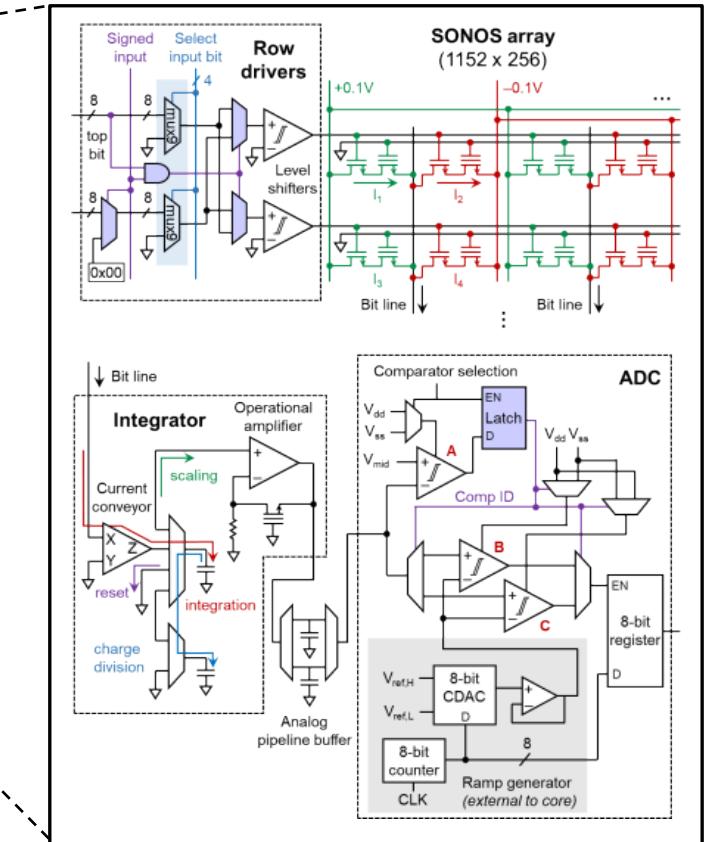
## Pipelined MVM tile



Data batch 1



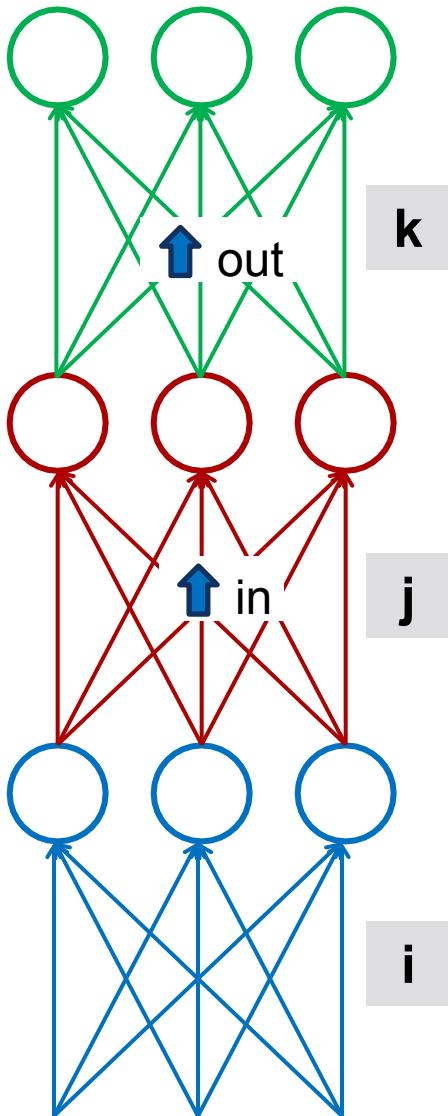
## Analog MVM core



Circuits designed and simulated using commercial 40nm PDK



# Forward Propagation – Inference



$O(N^2)$   
Operations

Neural Core



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

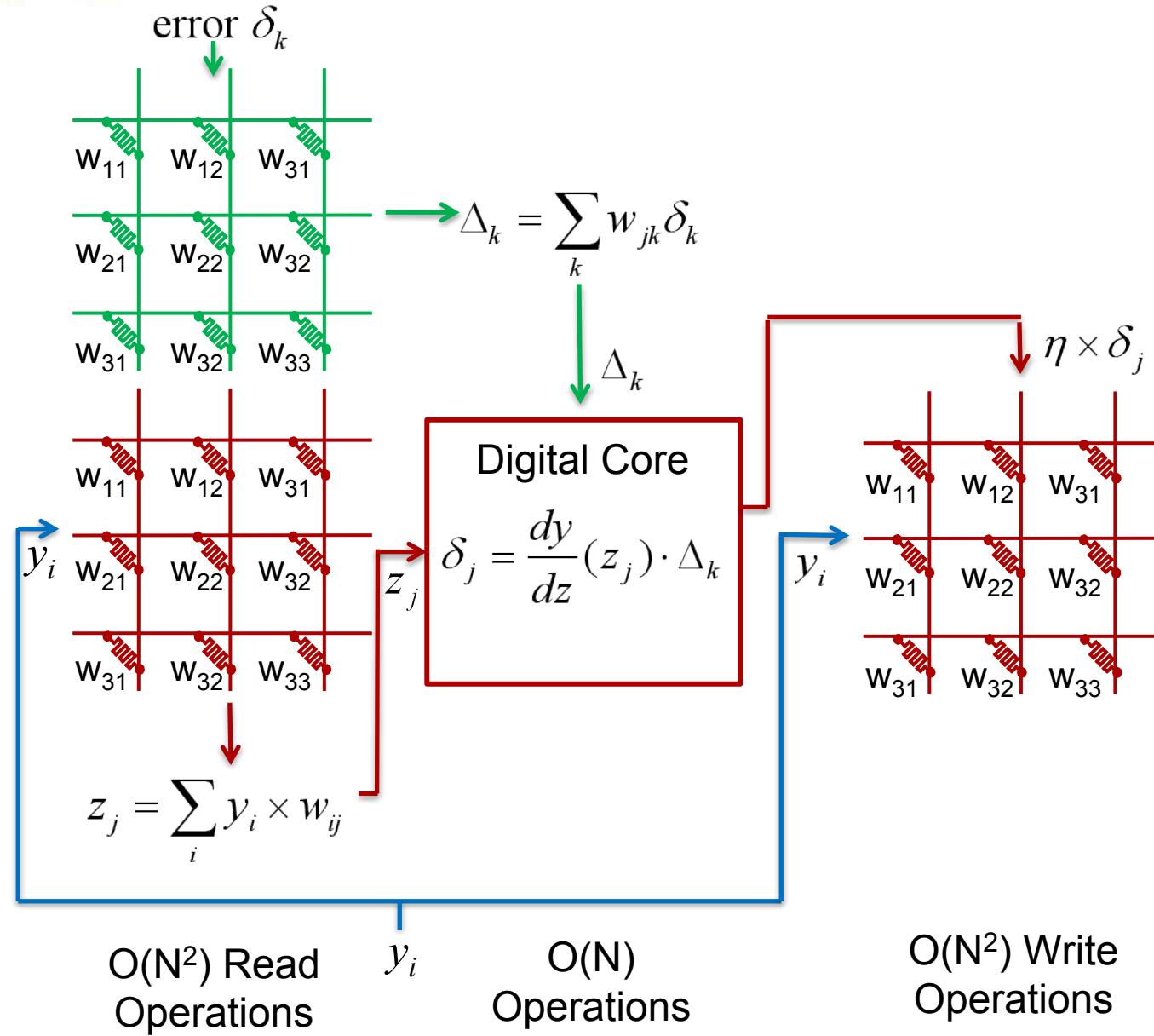
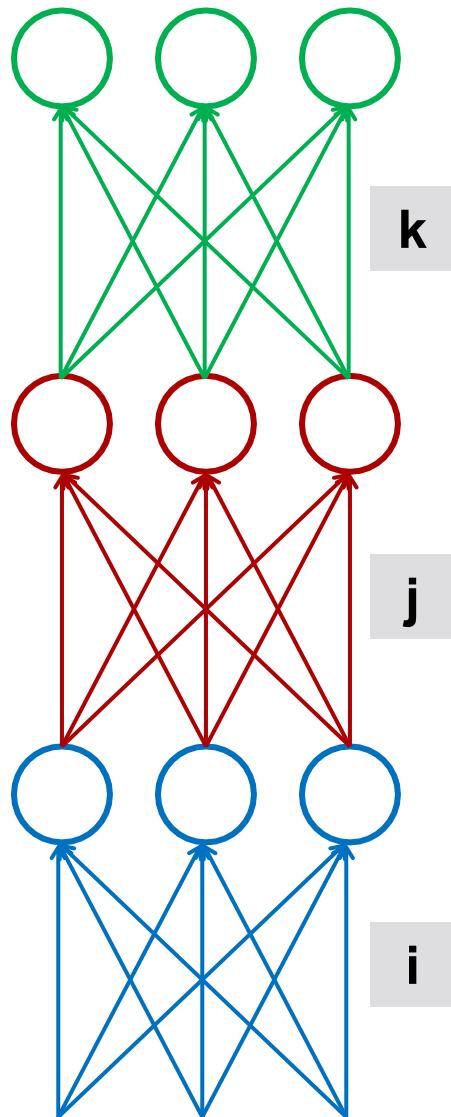
$O(N)$   
Operations

$y_j$

Digital Core

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

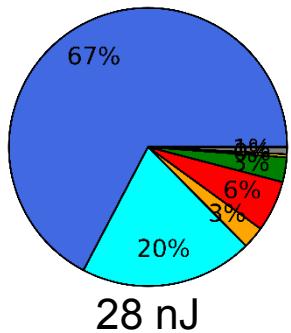
# Backward Propagation – Training



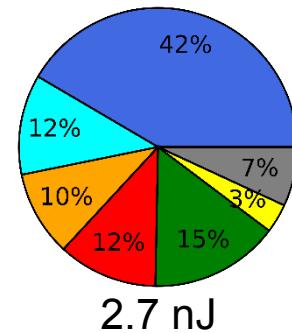
# Neural Core Energy Analysis

Analog ReRAM

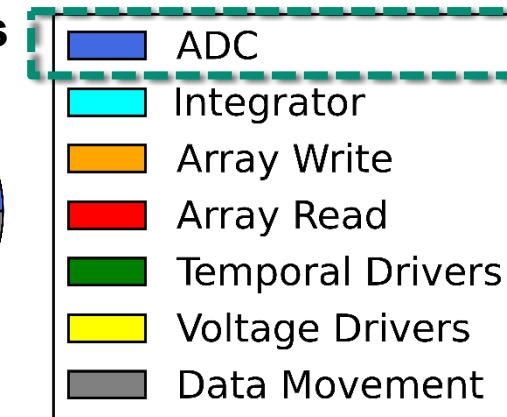
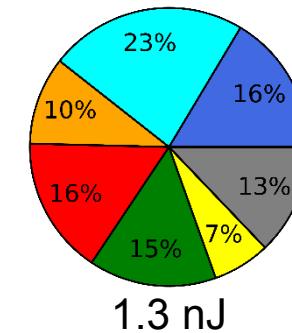
8 bits In/out  
8 bit weights



4 bits In/out  
8 bit weights



2 bits In/out  
8 bit weights



SRAM

