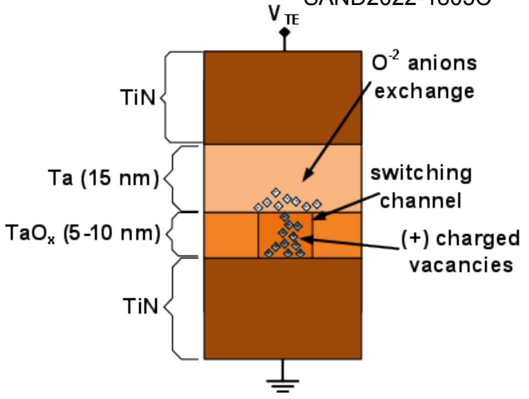
This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2022-1803C

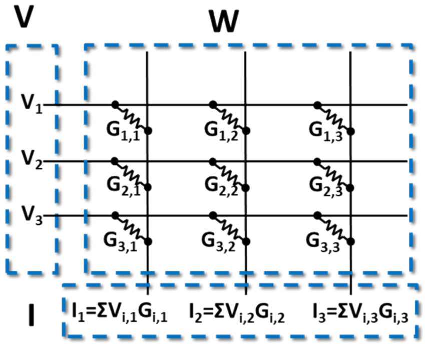
Sandia National Laboratories

Exceptional service in the national interest

Arizona State University



Energy Efficient Deep Neural Network Processing: Digital CMOS Limits and Prospects for Analog In-Memory Computing

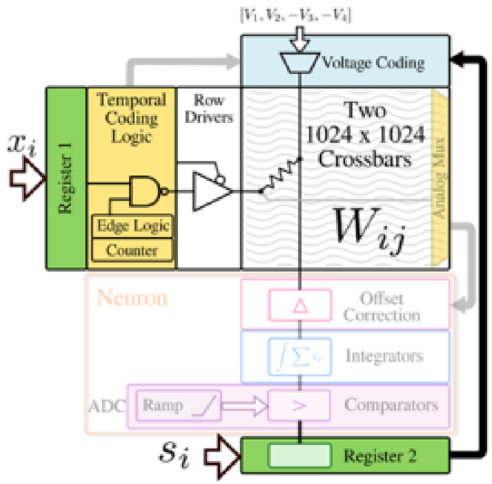


Matthew J Marinella[1], Patrick Xiao[2], Chris Bennett[2], Ben Feinberg[2], William Wahby,[2] Robin Jacobs-Gedrim[2], Vineet Agrawal[3], Helmut Puchner[3], Hugh Barnaby[1], and Sapan Agarwal[2]

1 – ARIZONA STATE UNIVERSITY, 2 – SANDIA NATIONAL LABORATORIES,

3 – INFINEON MEMORY SOLUTIONS

February 17, 2022



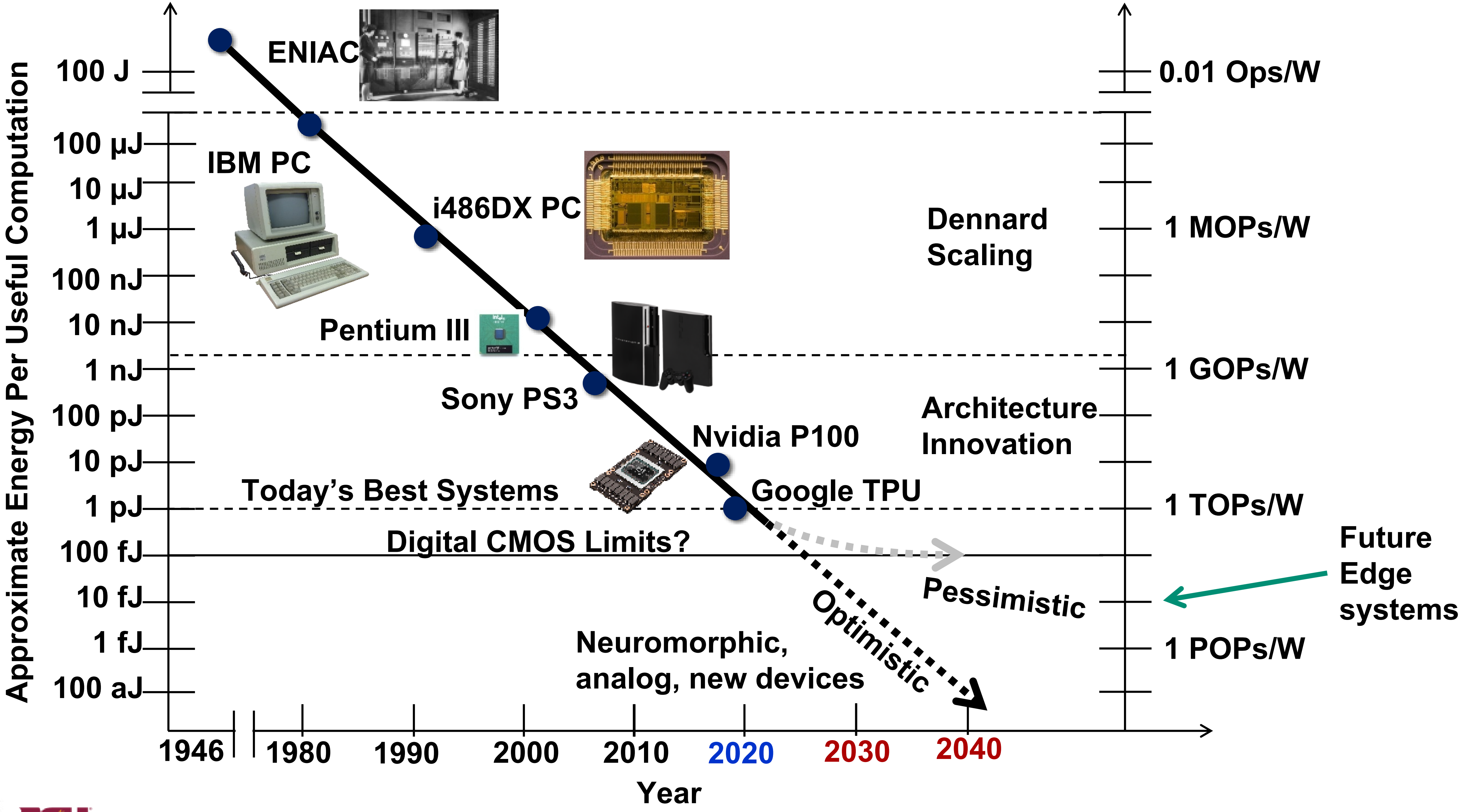
U.S. DEPARTMENT OF ENERGY NNSA

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.

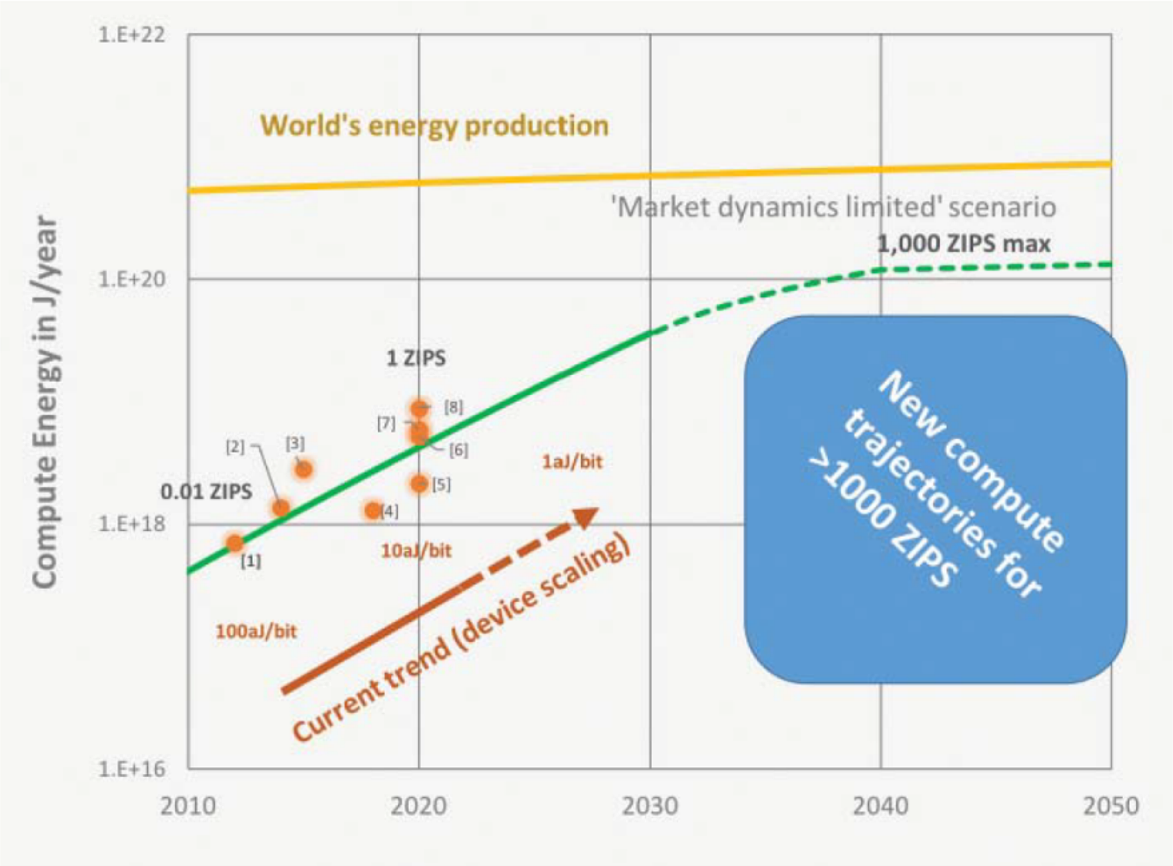
Outline

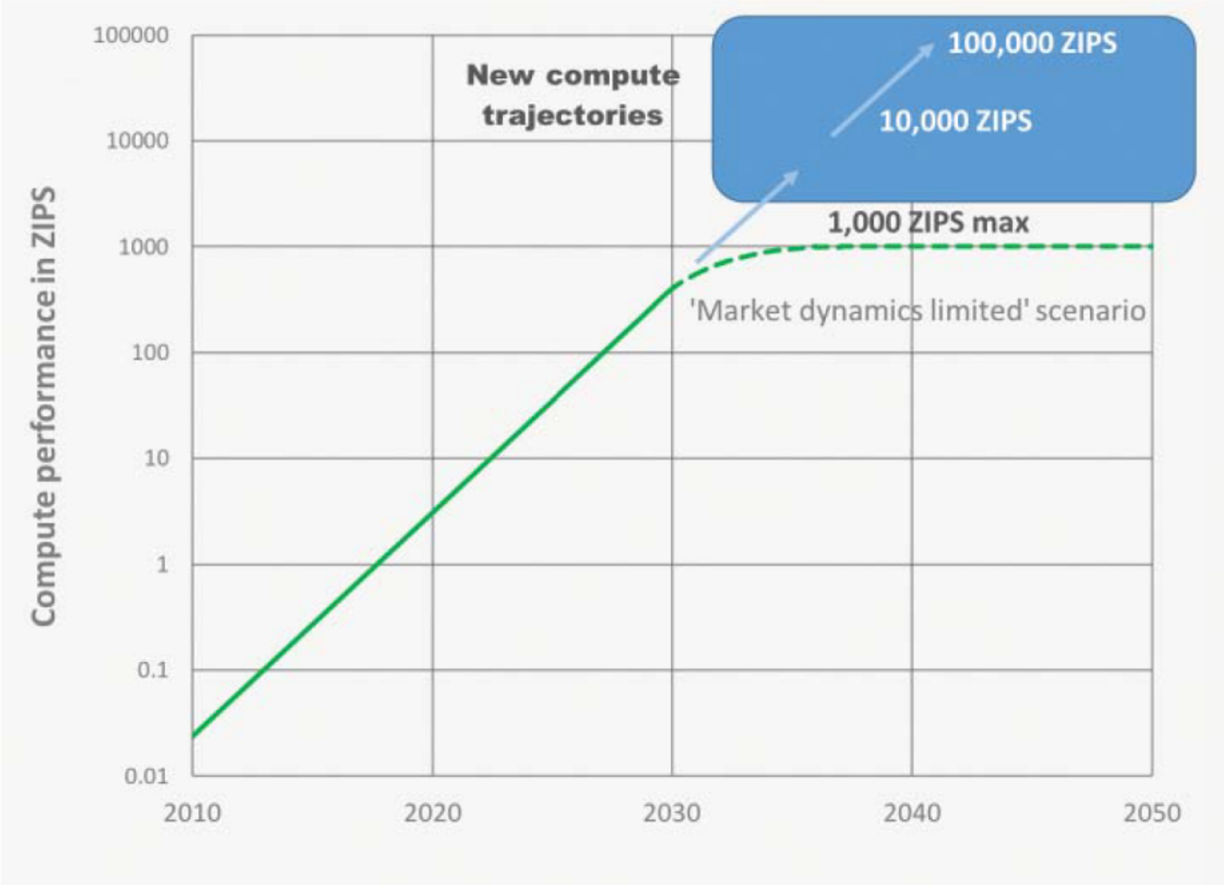
- **Motivation and Background**
- **Analog In-Memory Compute Energy & Latency**
- **Devices for Accurate Inference**
- **Conclusions**



Adapted from: **Marinella and Agarwal, Nature Electronics 2, 437, 2019**

View from the Semiconductor Industry



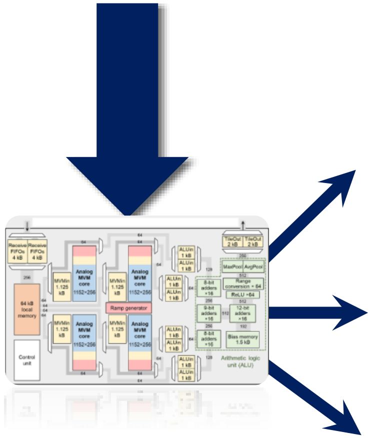


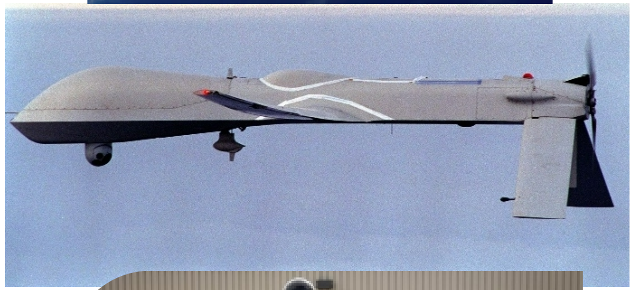
SRC Decadal Plan for Semiconductors, 2020

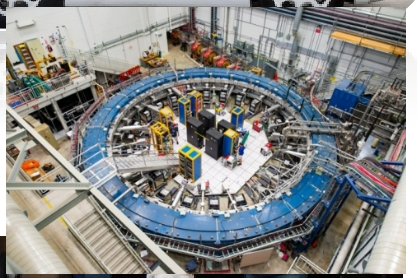
Semiconductor Research Corporation

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.

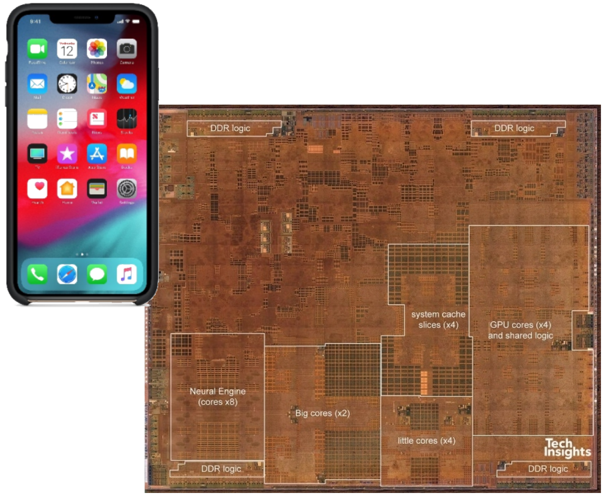




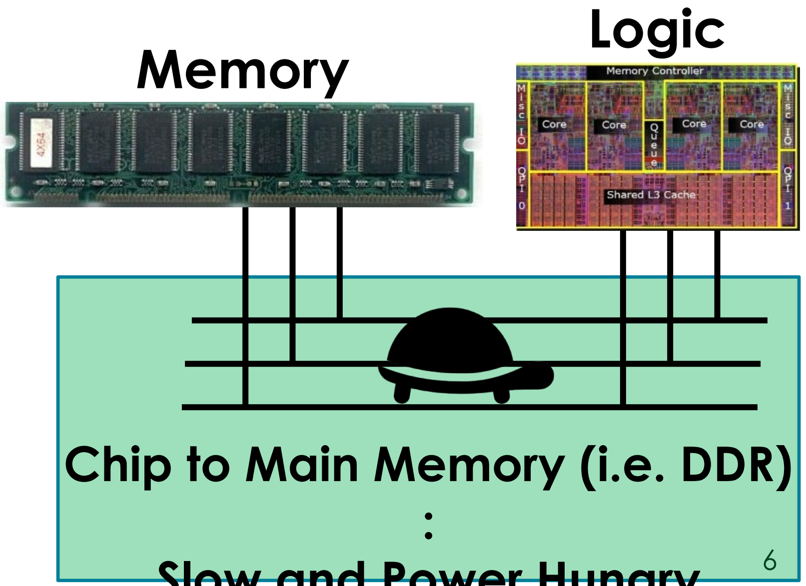


State of the Art CMOS Efficiency: **Apple A13**

- 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, techinsights.com



Outline

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

Keep Data in Memory & Exploit Physics for Computing

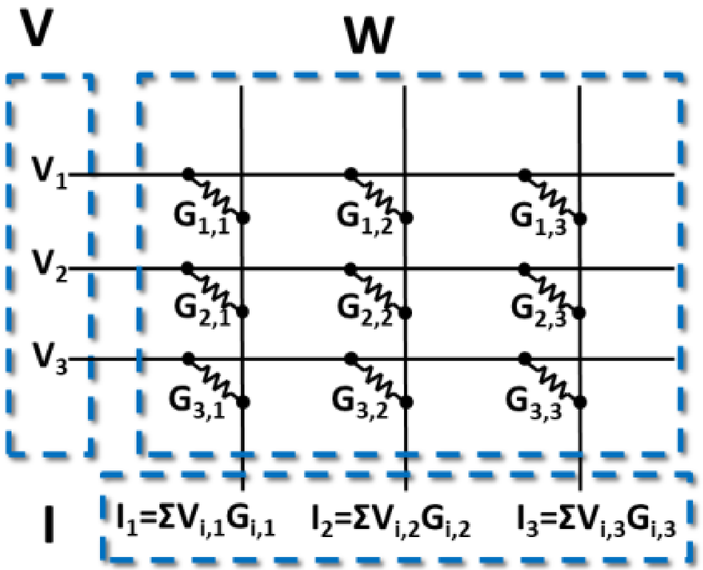
Mathematical

$$V^T W = I$$

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

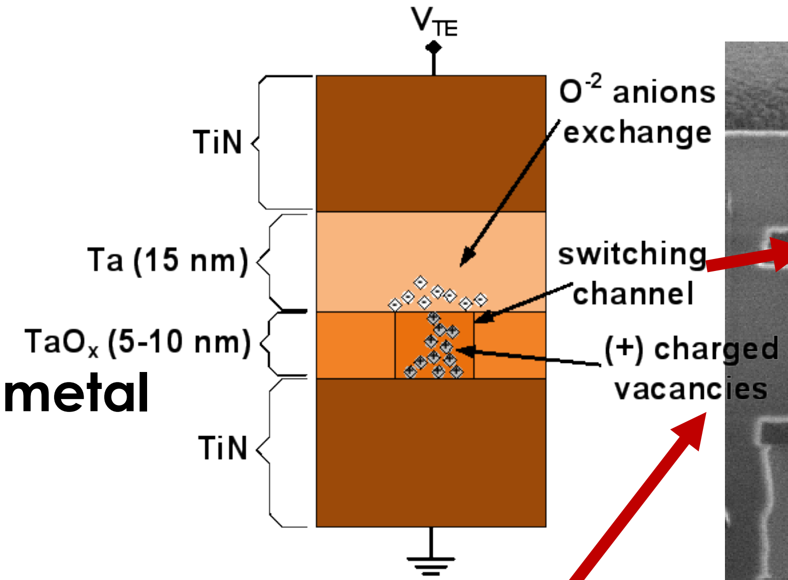
$$\begin{bmatrix} I_1 = \Sigma V_{i,1} W_{i,1} & I_2 = \Sigma V_{i,2} W_{i,2} & I_3 = \Sigma V_{i,3} W_{i,3} \end{bmatrix}$$

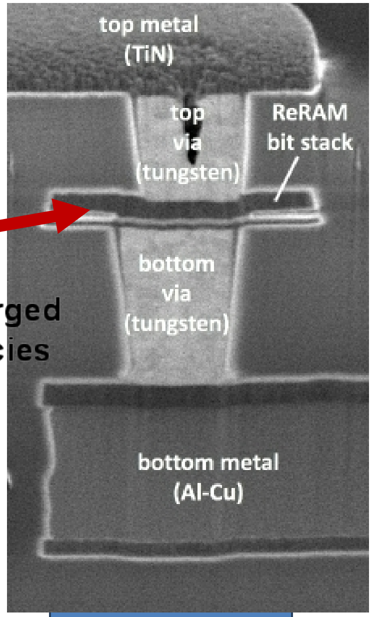
Electrical

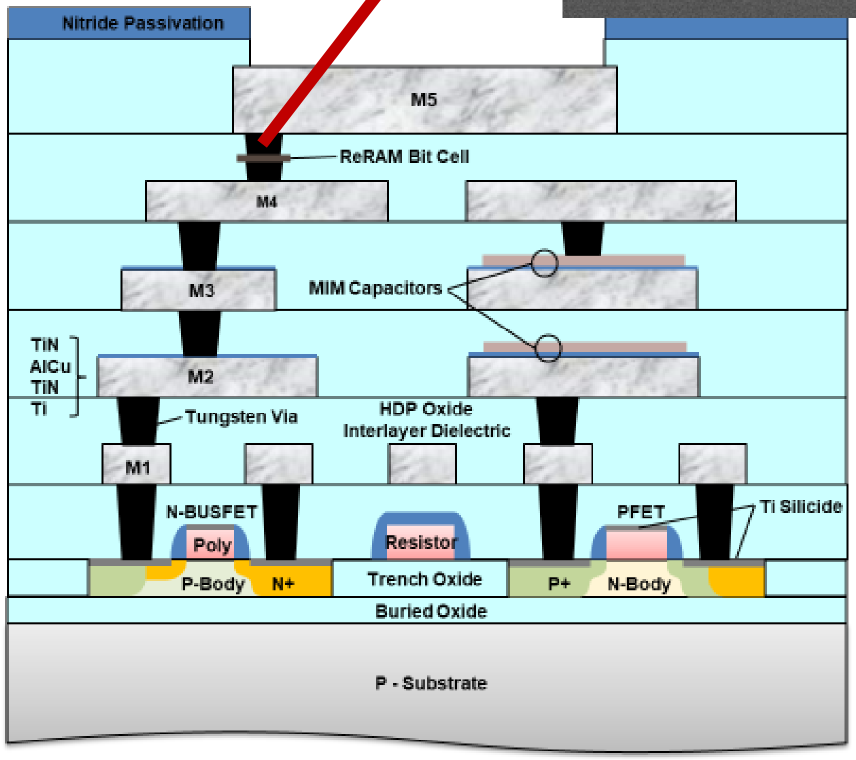


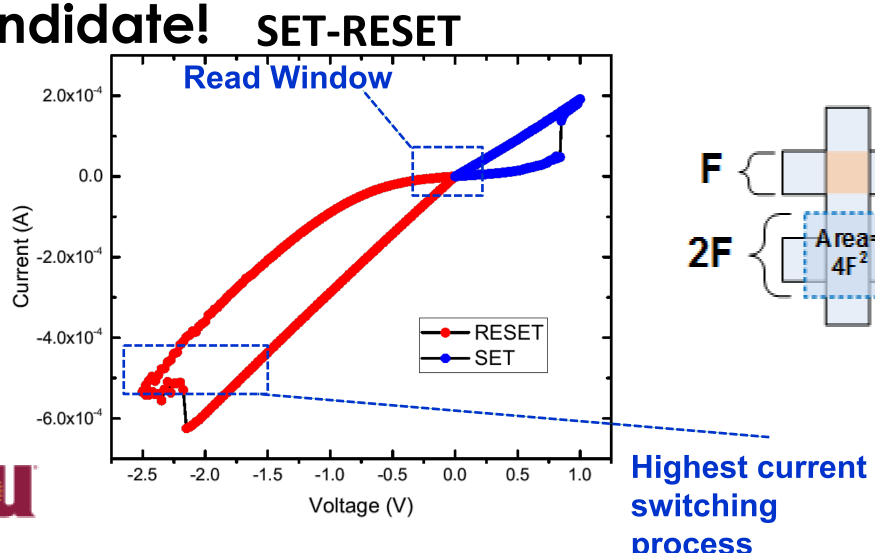
Tunable Resistor: Oxide ReRAM

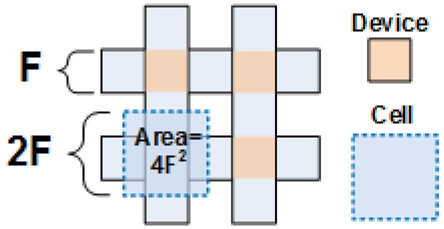
- **Known as ReRAM, OxRAM, "memristor"**
- **Bipolar resistance modulation in metal-insulator-metal structure**
  - **+V pulse, R decreases. -V pulse, R increases**
- **Fast, scalable, low switching energy, tunable resistor**
- **Potential for 100 Tbit of ReRAM on chip**
- **Perfect Analog In-Memory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







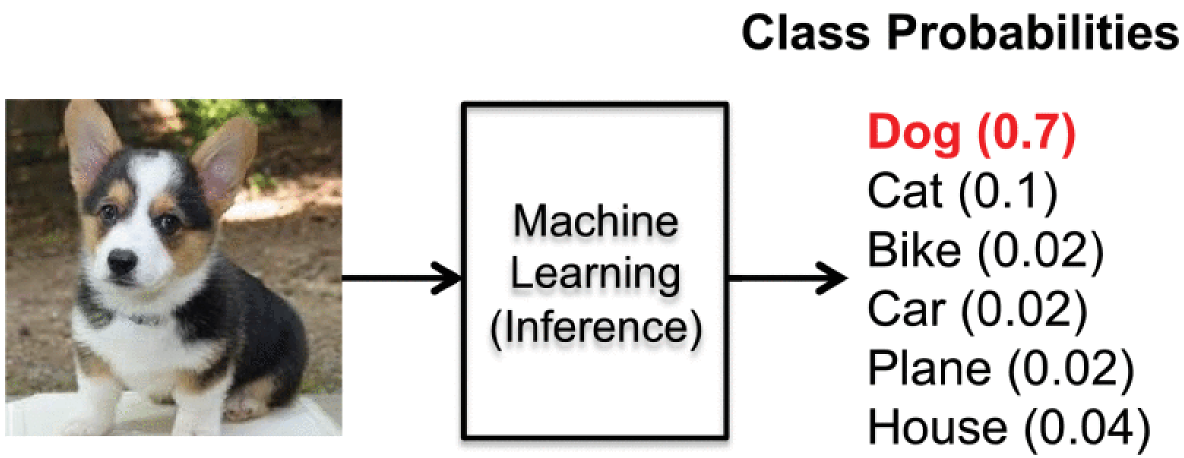




Neural Network Basics

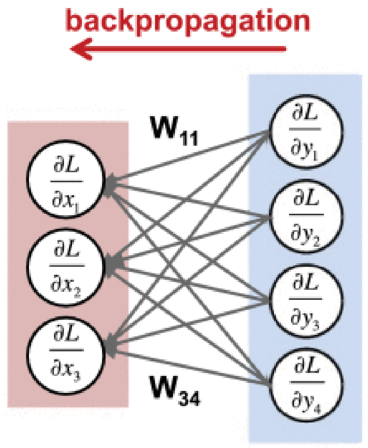
Inference

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



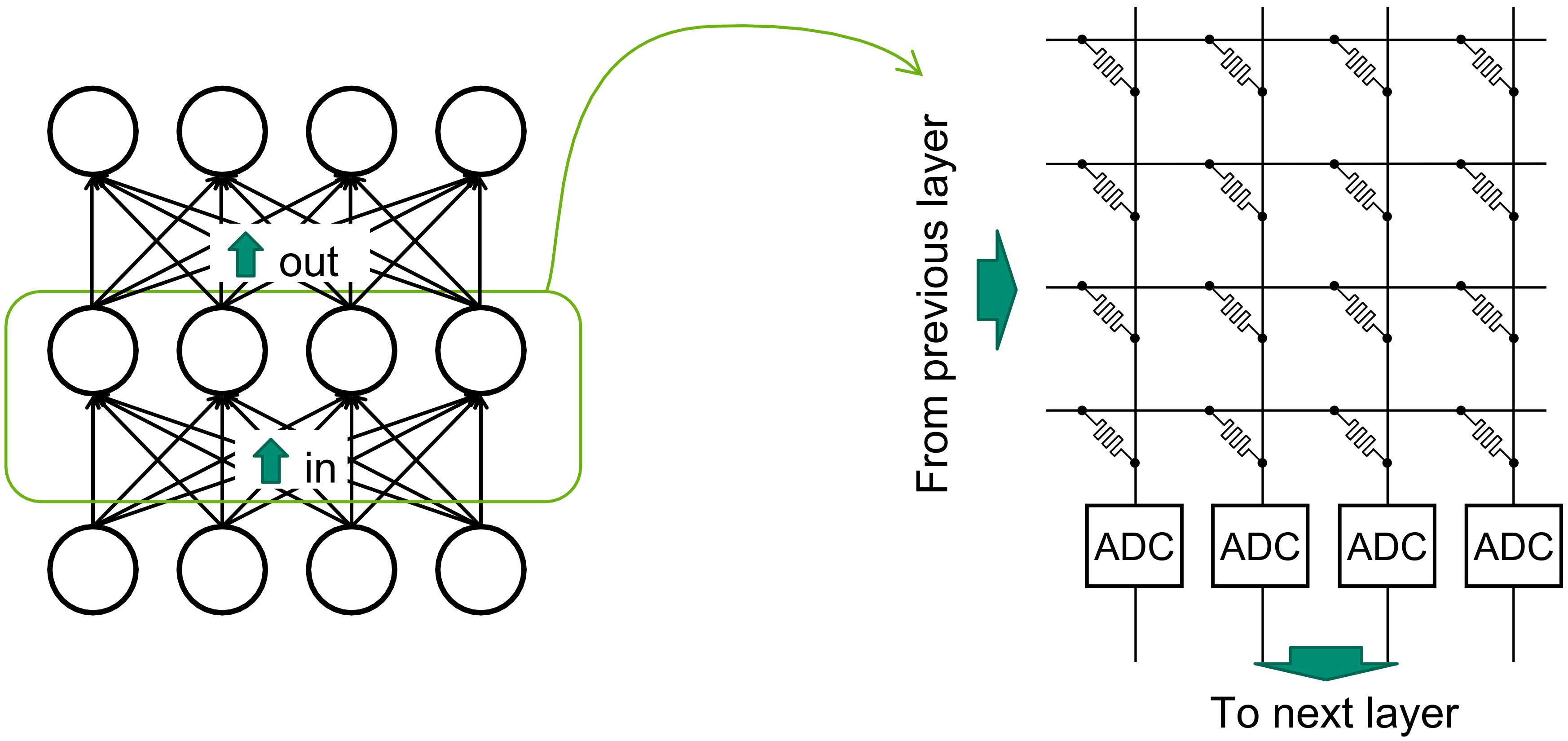
Training

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

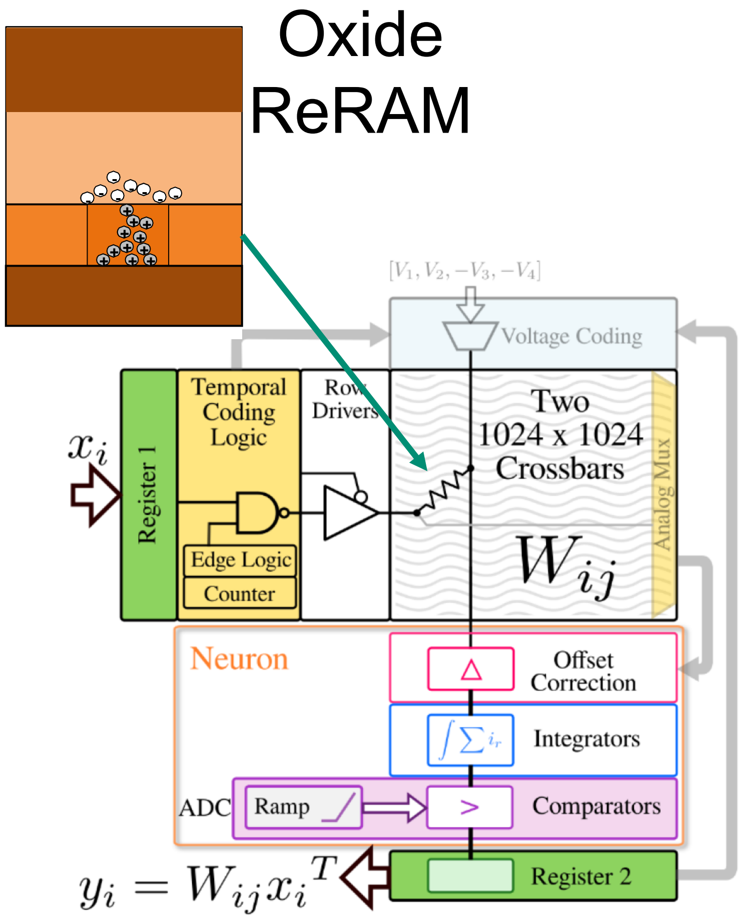


(b) Compute the gradient of the loss relative to the filter inputs

Physically Mapping a Neural Network to Resistive Array



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

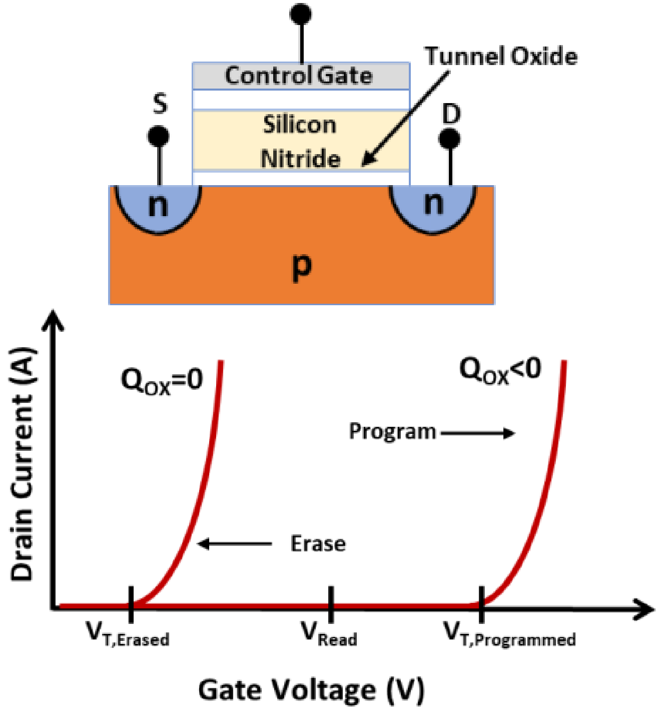
Initial results: two orders of magnitude beyond digital!

MJ Marinella, S Agarwal, et al, IEEE J. Emerging Topics in Circ. And Sys, 8, 2018.

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

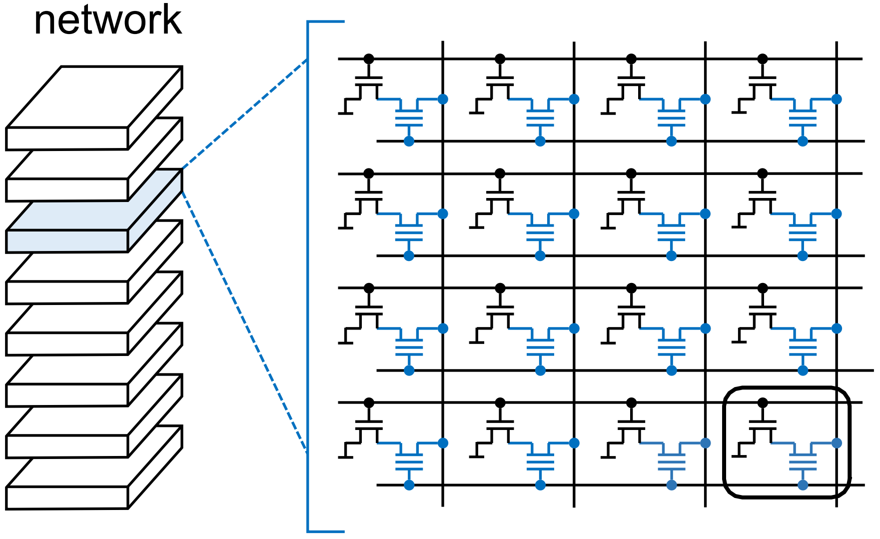
- Mature, commercial technology pioneered by Sandia in the 1980's
- Basis of modern SSD's (your iPhone uses SONOS)
- Can be used as resistive array similar to ReRAM
- Collaborating with Infineon to evaluate 40nm SONOS In Memory Computing

SONOS Device

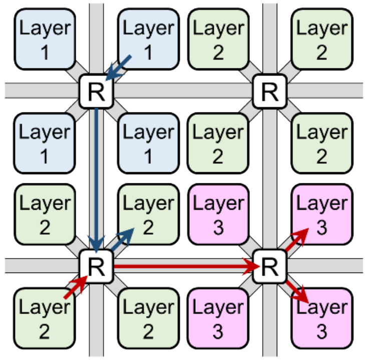


SONOS Analog VMM Array Implementation

Deep neural network

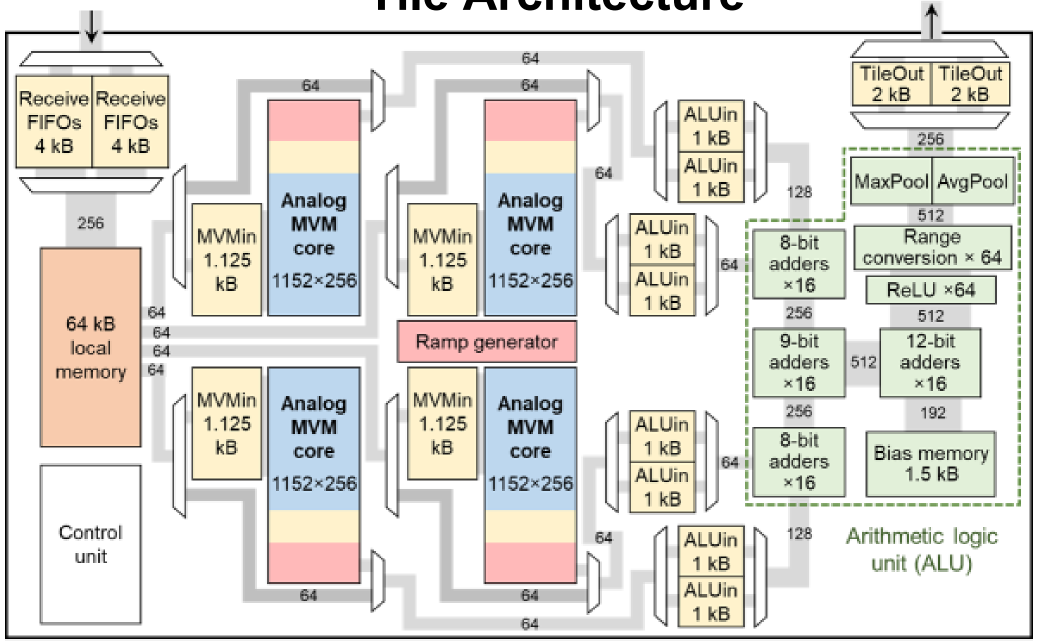


78 TOPS/Watt 8-bit Inference using 40nm SONOS



CNN Mesh

Tile Architecture



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!

T.P. Xiao et al, accepted, IEEE J. Circuits and Systems, 2021.

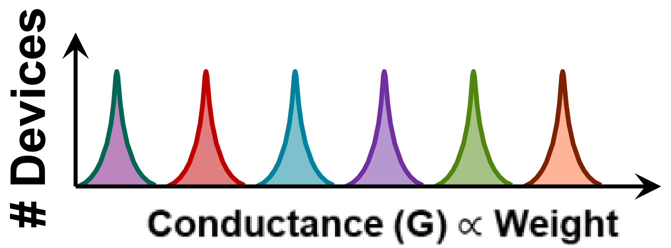
m
our commercial collaborator, Infineon

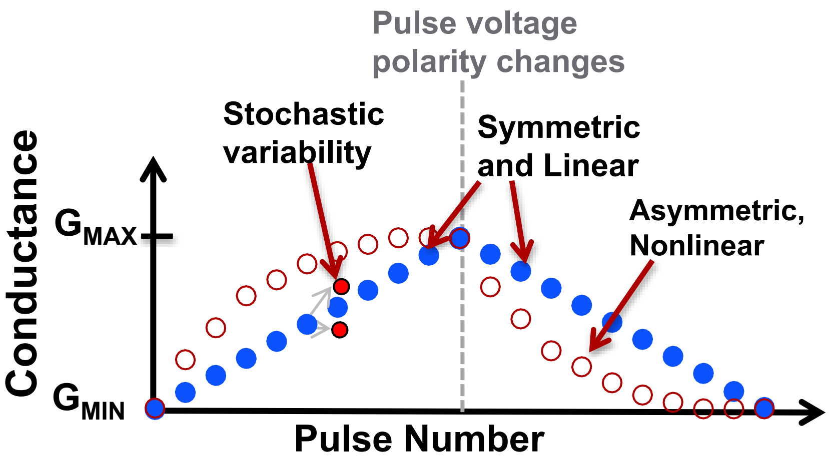
Outline

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

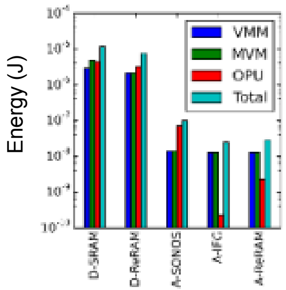
Analog Required a Paradigm Shift

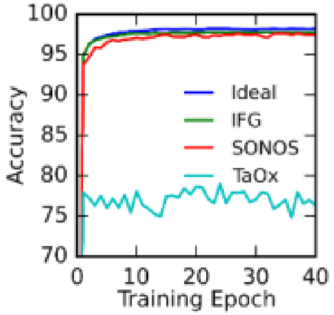
- **Analog processing offers great benefits…**
- **…but comes with great challenges**
- **Digital: Deterministic, accurate results**
- **Analog: Device characteristics affect *algorithm accuracy*!**
  - **Research challenge: analog behavior must give acceptable algorithm -level results**
- **<u>Inference Accuracy Challenges (this section)</u>**
  - **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**
- **<u>Training Accuracy Challenges (next section)</u>**
  - **Actual analog device state change does not match intended weight update**
  - **Caused by write nonlinearity, asymmetry, stochasticity**
  - **Device to device variation**



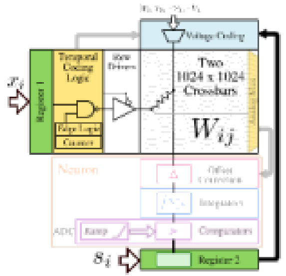


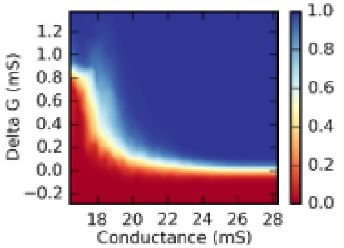
Multiscale CoDesign Framework Enables Accuracy Prediction



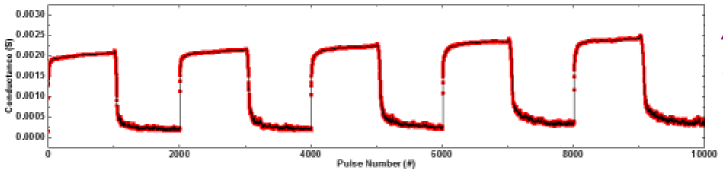


Energy/Performance Model
Model performance and energy requirements

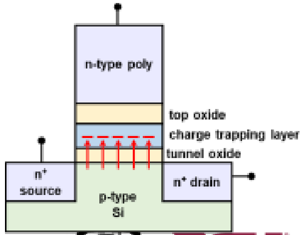


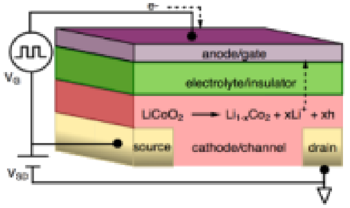


Sandia Cross-Sim:
Translates device measurements and crossbar circuits to algorithm-level performance



Analog characterization





Target Algorithms
- Deep Convolutional Nets
- Sparse Coding
- Liquid State Machines

Algorithms

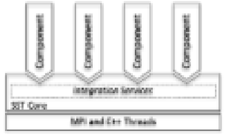
Architecture

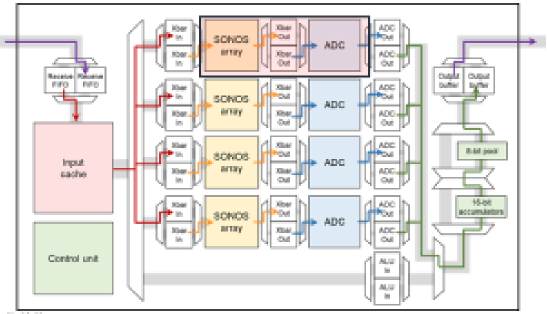
Circuits

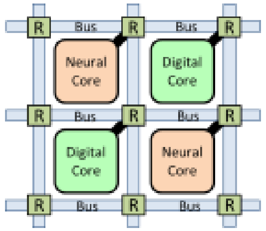
Devices

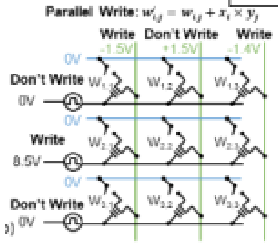
Materials

Architecture Simulation

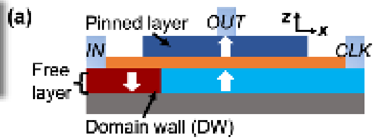


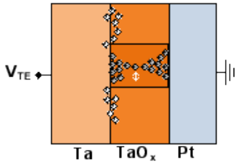


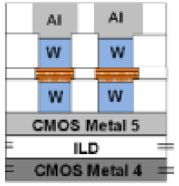


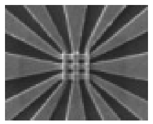


Drift-diffusion model of transport

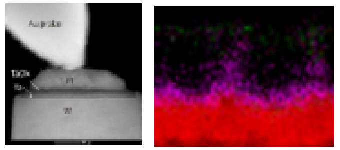


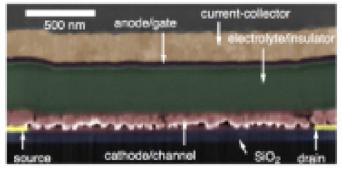




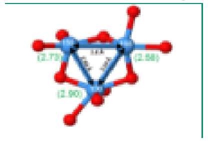


In situ Characterization



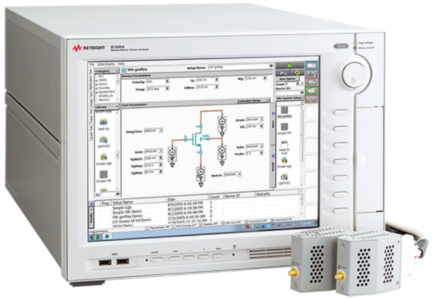


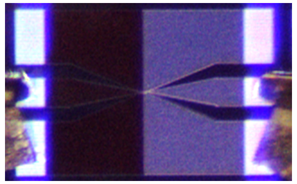
Ab Initio Modeling

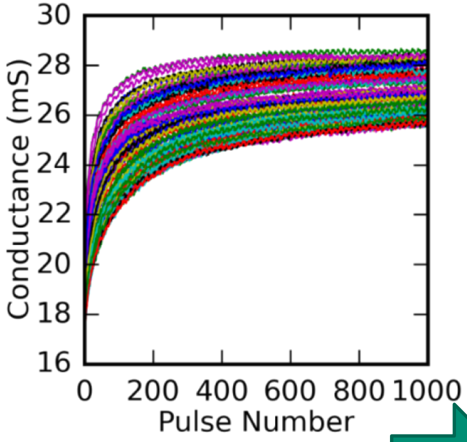


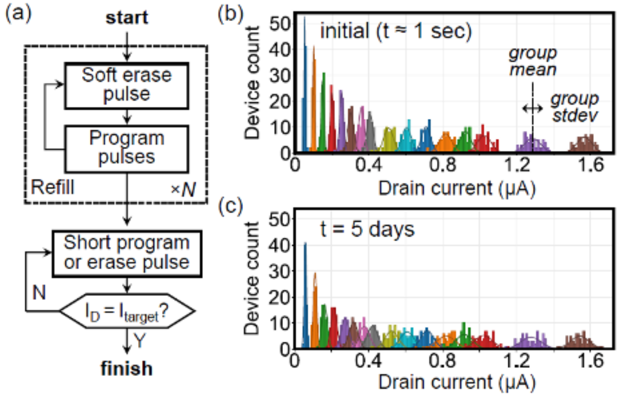
Compact Modeling Dataset for Neural Accuracy Model

Measure Devices

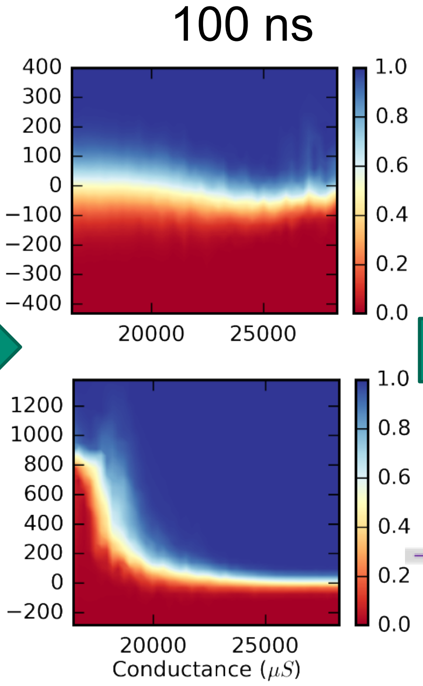






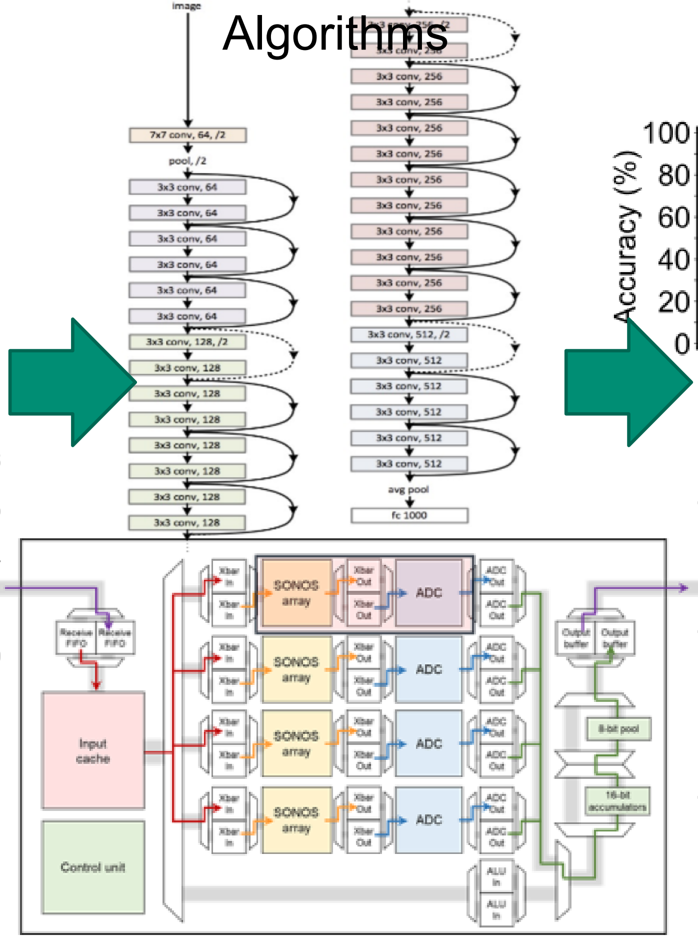


Construct Lookup Tables

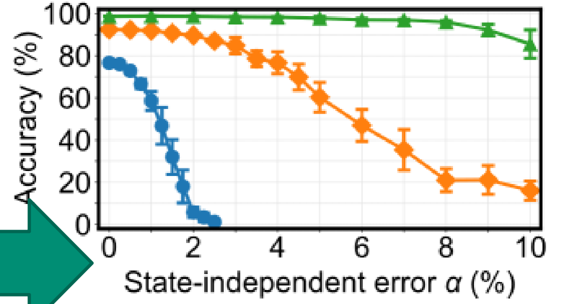


CROSS SIM

Model Array Circuitry, Architecture, & Algorithms



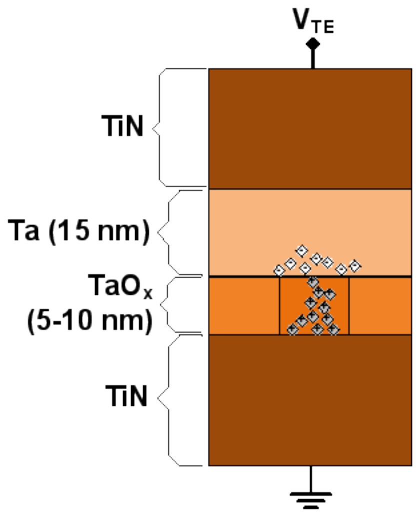
Assess Neural Algorithm Accuracy, Efficiency, Performance, Radiation

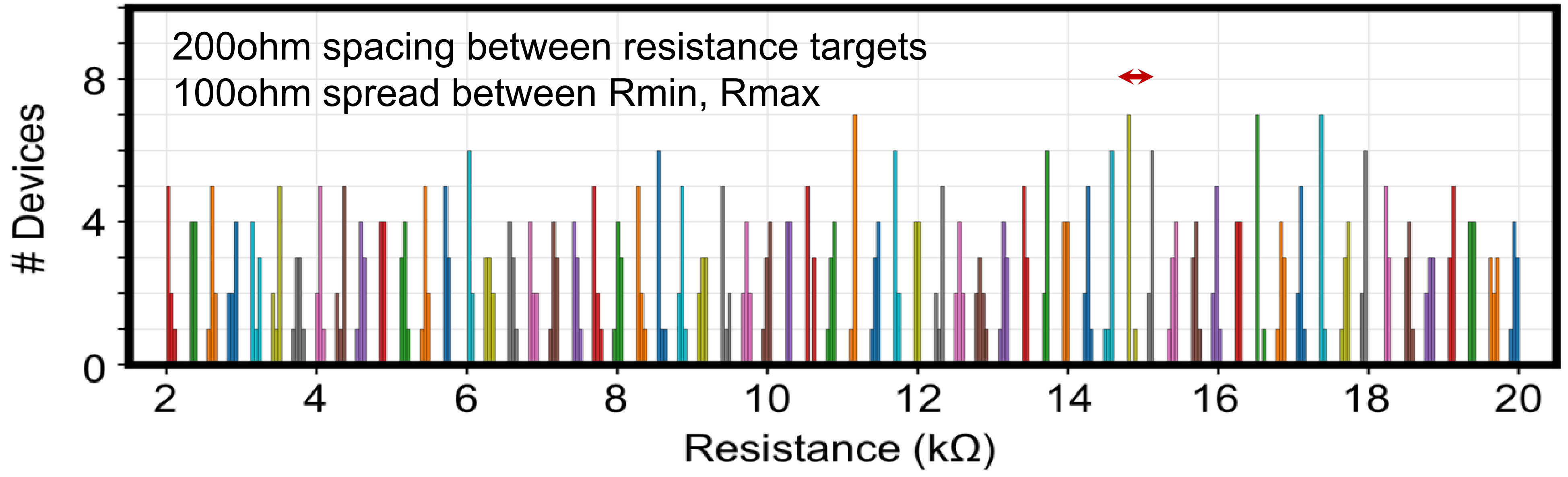


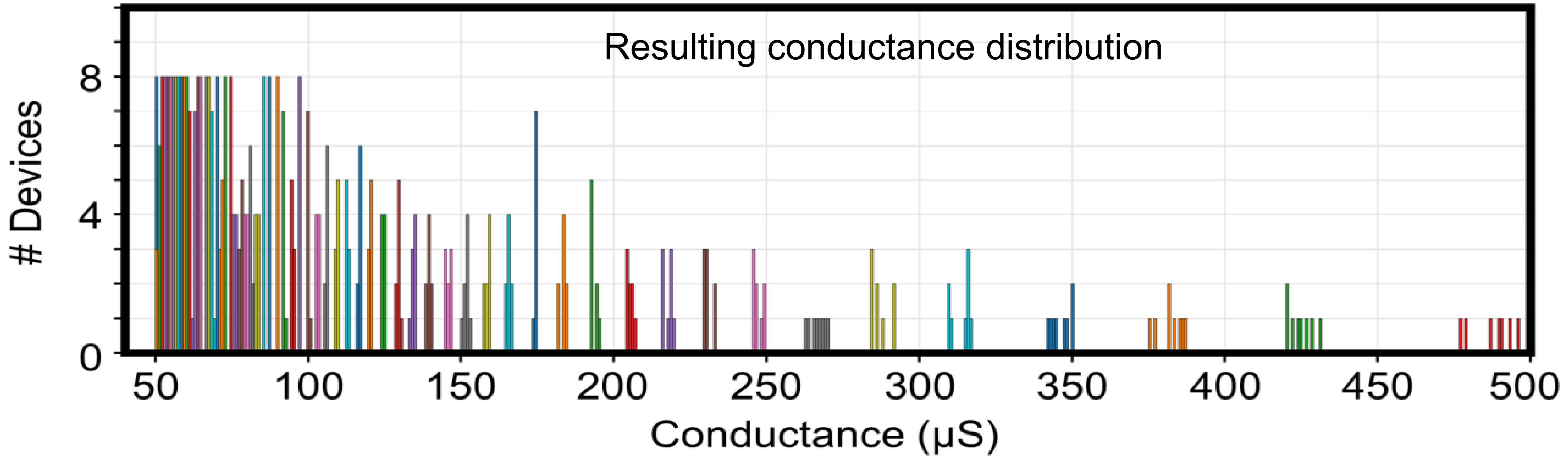
Component	VMM	OPU
Energy/Op ReRAM (fJ)	12.2	2.1
Array Latency ReRAM (μs)	0.38	0.51

Xiao et al TCAS, accepted, 2021.

Sandia TaOx ReRAM Inference Resistance Distributions



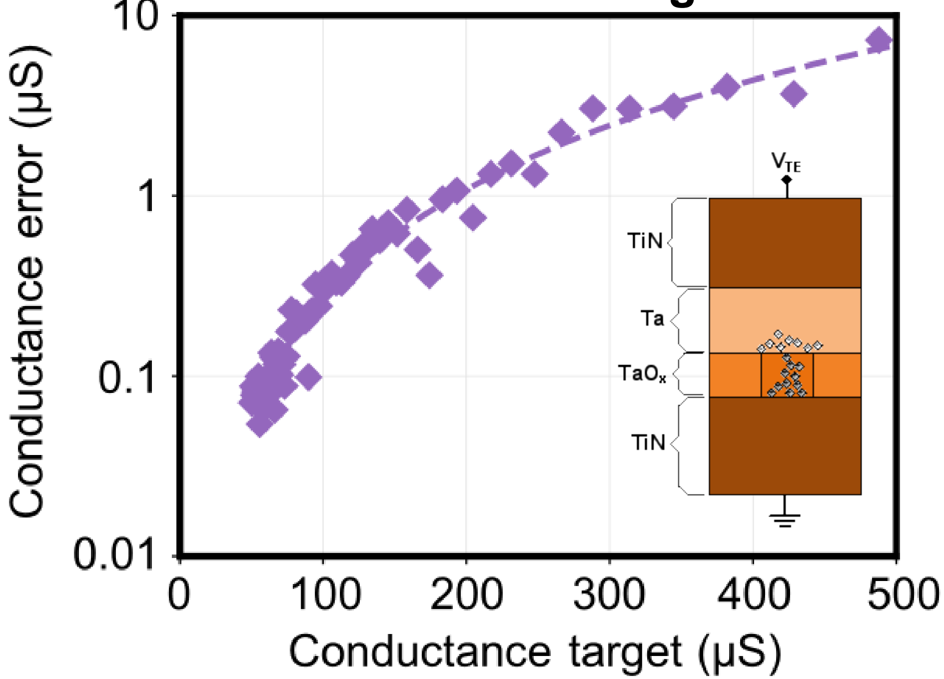




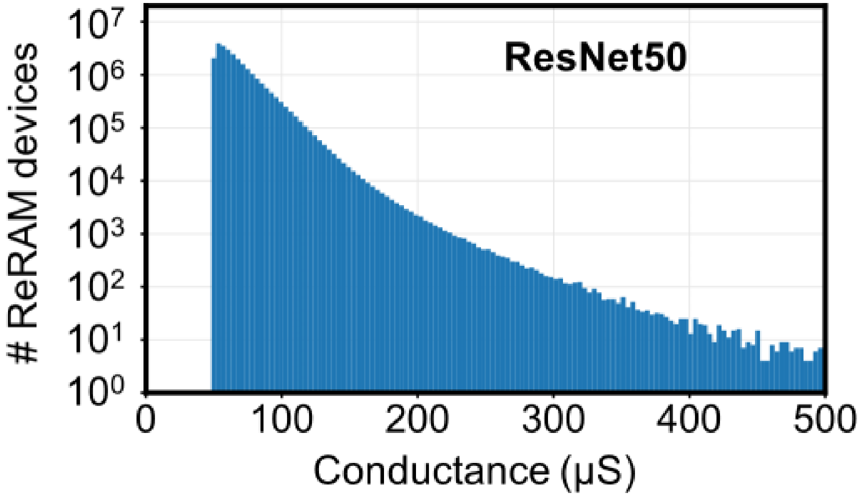
Sandia TaOx ReRAM Error Model

- **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!***
  - **Caveat: preliminary data – relaxation will degrade**

Conductance Error as a Function of Conductance Target

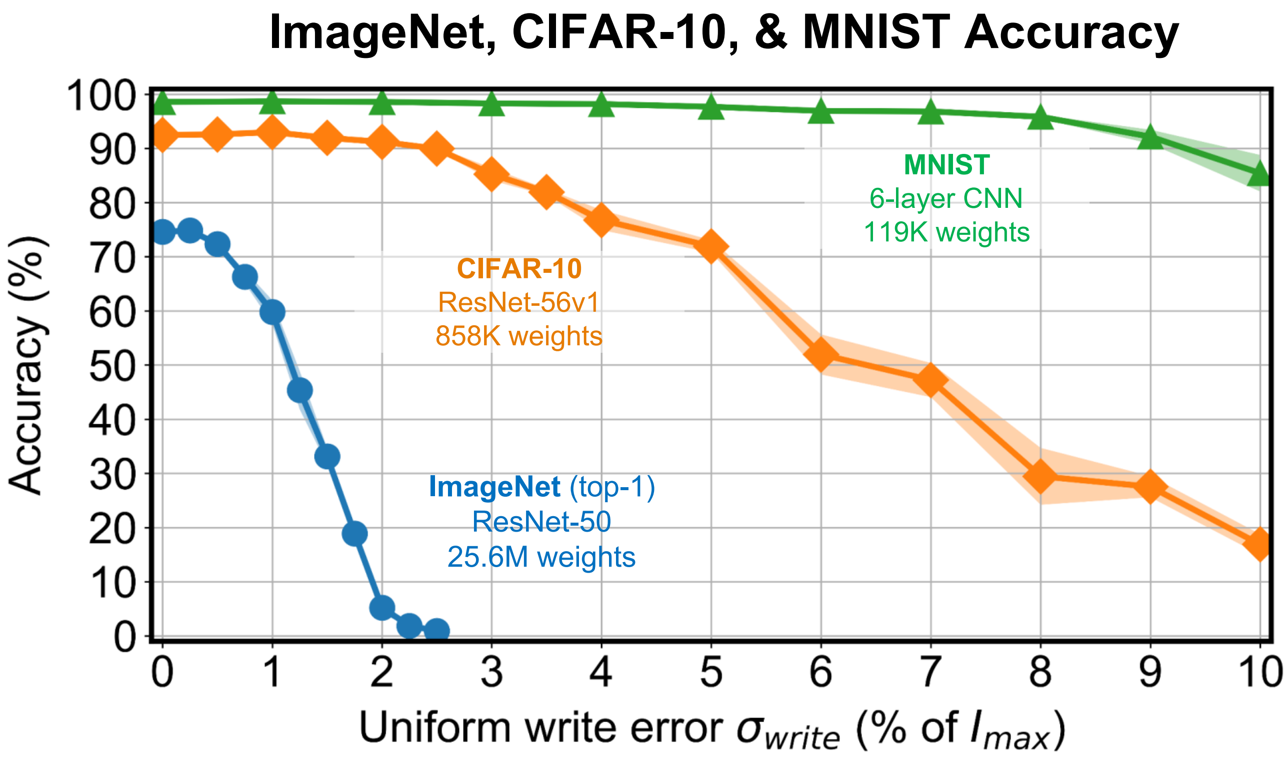


Conductance-Weight Distribution



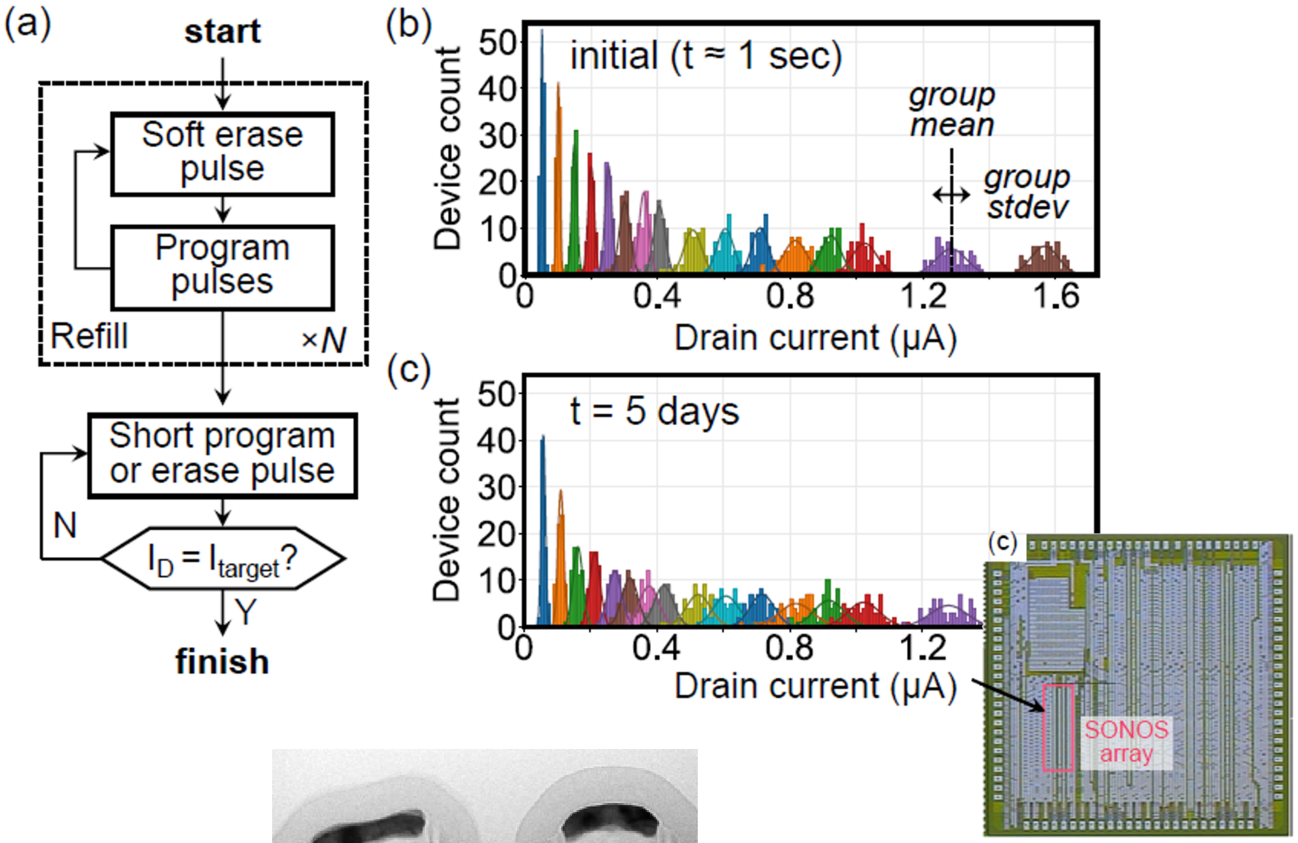
Effect of Network and Dataset on Accuracy

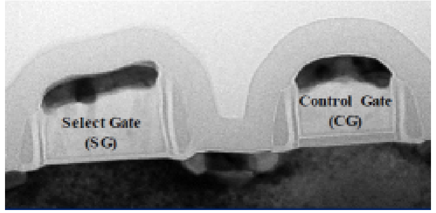
- Different common datasets and CNN architectures often analyzed
- MNIST (uses simple CNN)
  - 28x28 pixel grayscale
  - 10 classes
  - 60k training images
  - 10k test images
- ImageNet (requires large CNN arch.)
  - 224x224 pixel color
  - 1000 classes
  - 1.3M training images
  - 100k test images
- 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!



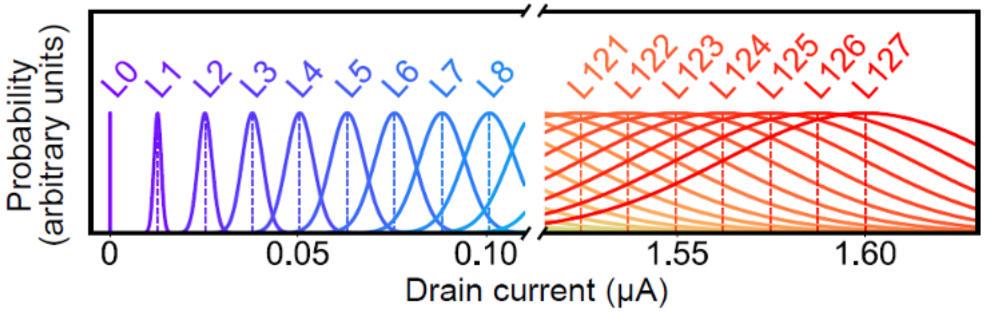
40nm SONOS Deep CNN Inference Modeling

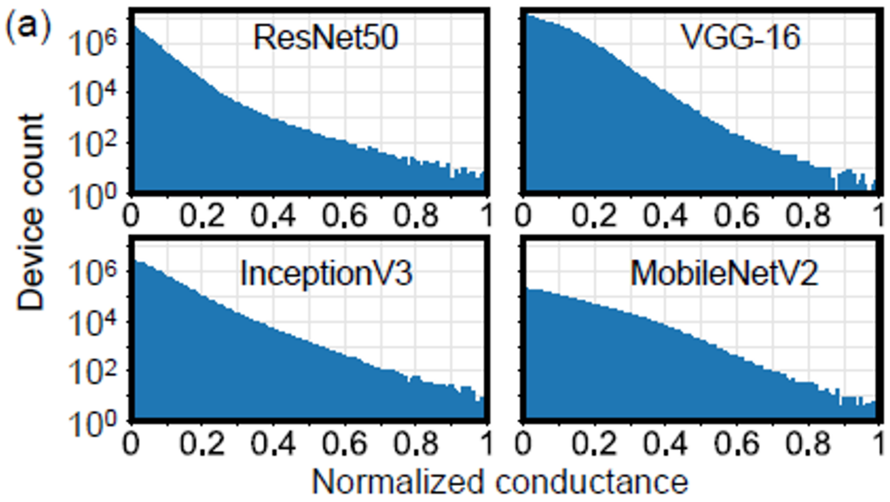
Infineon 40nm SONOS Characterization Chip Data





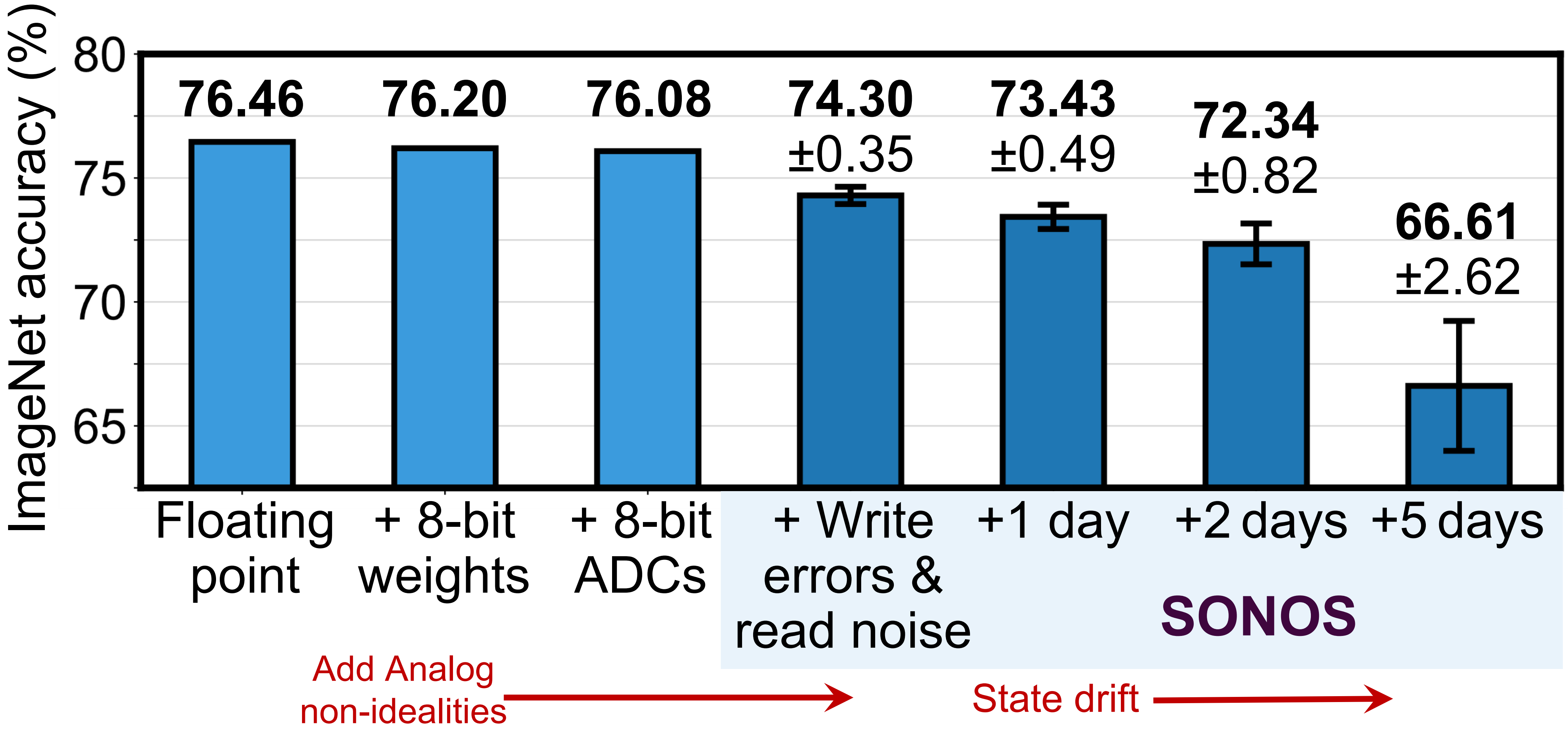
Modeled 7-bit Weight Distribution and Mapping





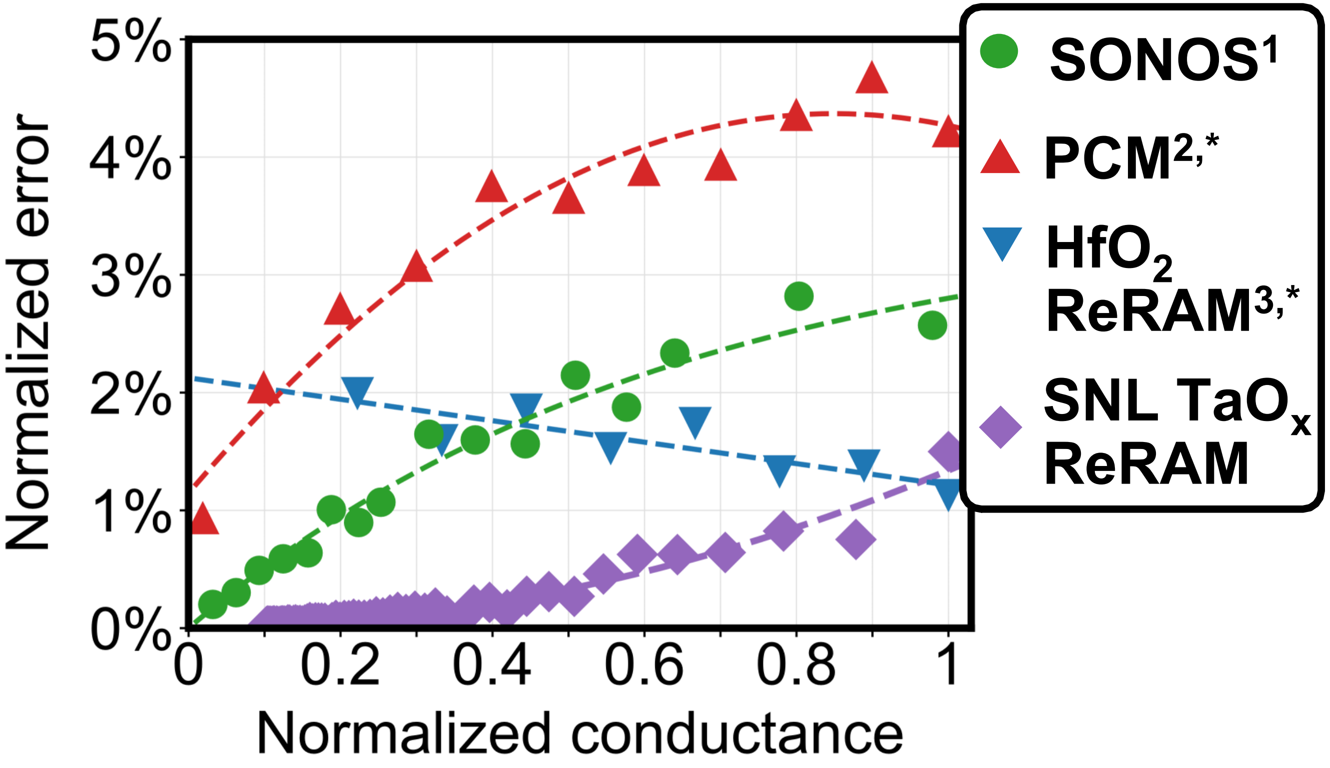
T.P. Xiao et al, IEEE TCAS, in press, 2021.

Infineon 40nm SONOS for CNN Inference - ImageNet



T.P. Xiao et al, IEEE TCAS, 2022.

Error and Inference Accuracy Summary: ReRAM, SONOS, PCM



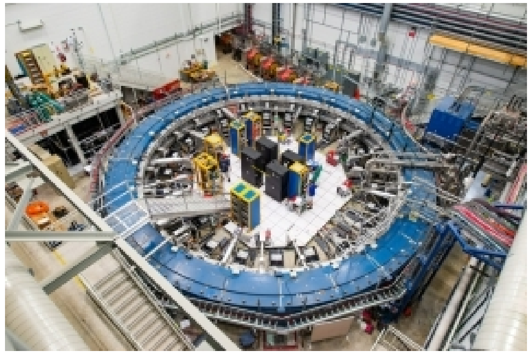
Technology[+]	Top-1 accuracy	Top-5 accuracy
Floating point digital	77.5%	93.3%
SNL TaOx ReRAM	76.4% ± 0.2%	93.3% ± 0.1%
SONOS[1]	74.0% ± 1.0%	92.5% ± 0.4%
PCM[2]	28.2% ± 6.4%	49.7% ± 7.8%

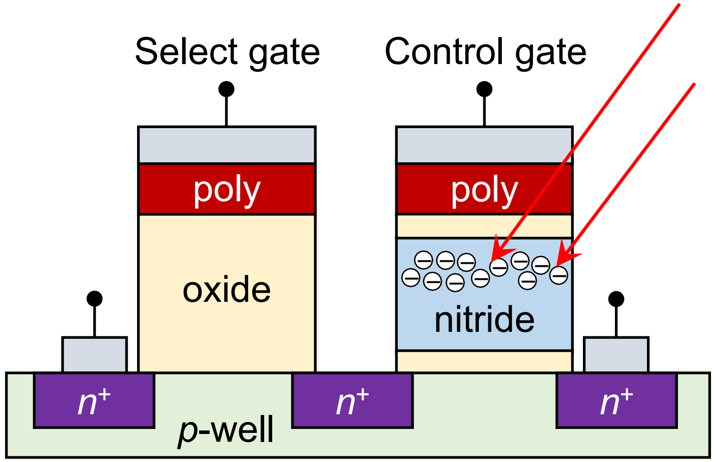
References and notes:
[1]T.P. Xiao et al, IEEE TCAS, 2022.
[2]V. Joshi et al, Nat Comm. 11, 2020.
[3]Milo et al, IEEE IRPS, 2021.

[+]All analog simulation also includes 8-bit weight quantization, 8-bit activations, and 8-bit ADCs
*PCM and HfO_2 error are modeled entirely from data and programming used in publication only.

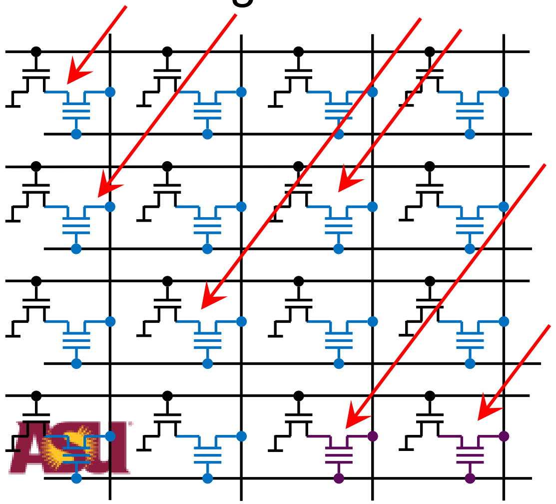
Device-Level Radiation Impacts Algo Accuracy

How will the accuracy degrade in radiation environments ?

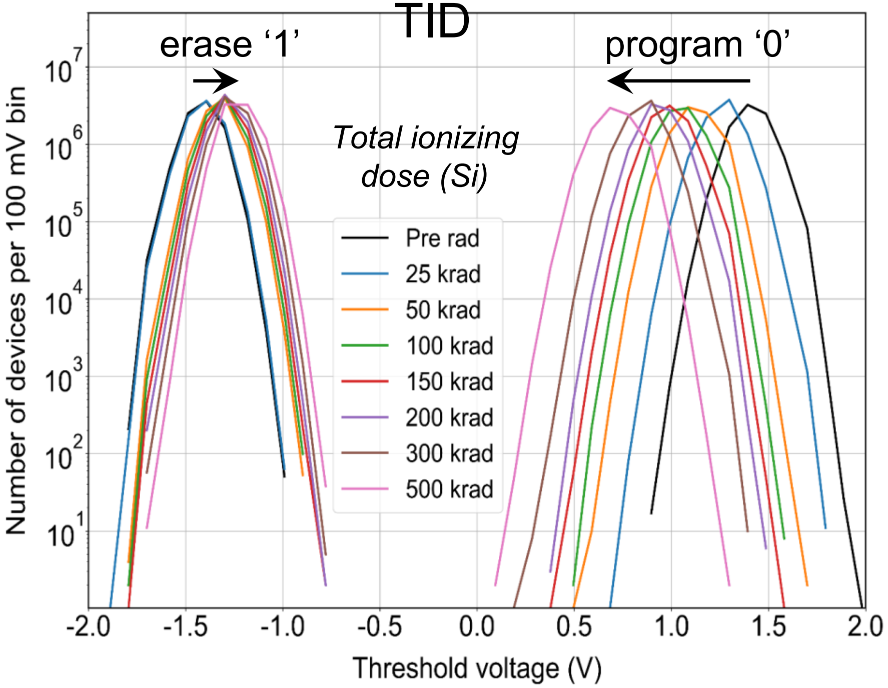




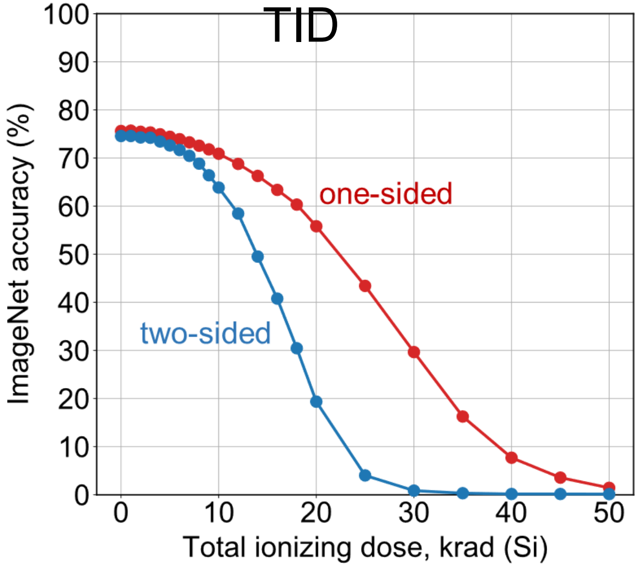
Ionizing Radiation

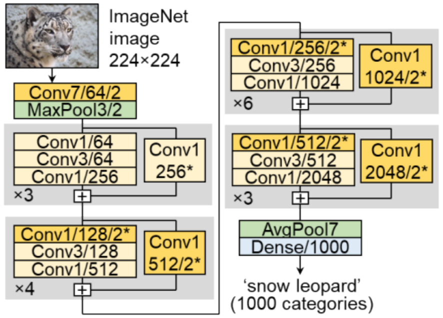


Threshold Distribution Shifts due to TID



Algorithm Accuracy Degradation due to TID





TP Xiao et al, IEEE Trans Nuclear Sci, 2021

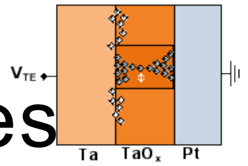
Outline

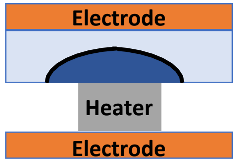
- Motivation and Background
- Analog In-Memory Compute Energy & Latency
- Devices for Accurate Inference
- Conclusions

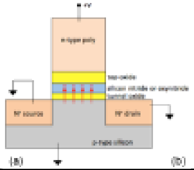
Analog Device Requirements

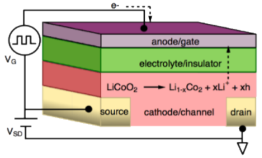
Property	Inference	Training
Analog programing error (w/ write verify)	Critical	Less Important
Long term retention	Important	Less Important
Read noise	Important	Less Important
Conductance Range	Important	Important
Short term state drift	Important	Important
Device to device variability	Important	Important
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
Inference	Analog programing error (w/ write verify)	🙂	😐	🙂	🙂
	Long term retention	🙂	🙂	🙂	😐
	Read noise	🙂	🙂	🙂	🙂
	Conductance range	😐	😐	🙂	🙂
Both	Short term state drift	😐	😐	😐	😐
	Device to device variability	☹️	😐	🙂	🙂
Training	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
  - Future work: 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

External Collaborators

Jean Anne Incorvia, UT

Stan Williams, TAMU

Hugh Barnaby, ASU

Alberto Salleo, Stanford

Yiyang Li, U Michigan

Helmut Puchner, Infineon

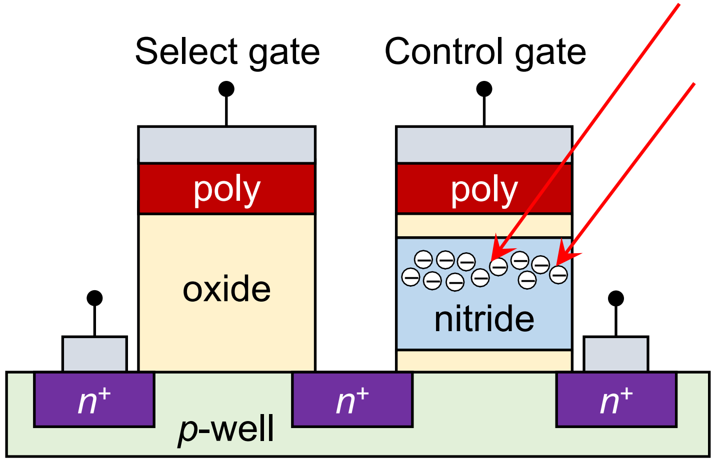
Vineet Agarwal, Infineon

John Paul Strachan, HPE

Victor Zhirnov, SRC

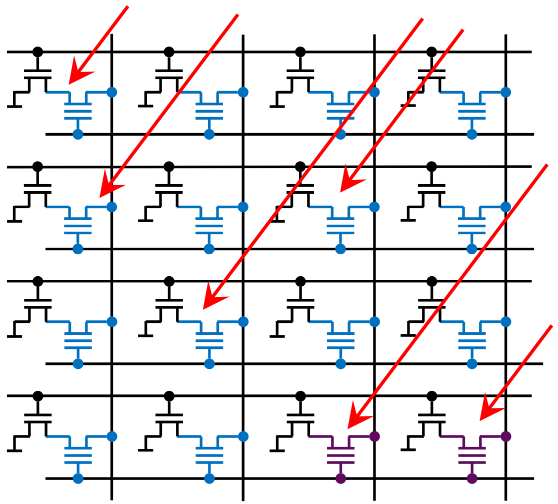
Jesse Mee, AFRL

Impact of Ionizing Radiation on Deep Net Accuracy

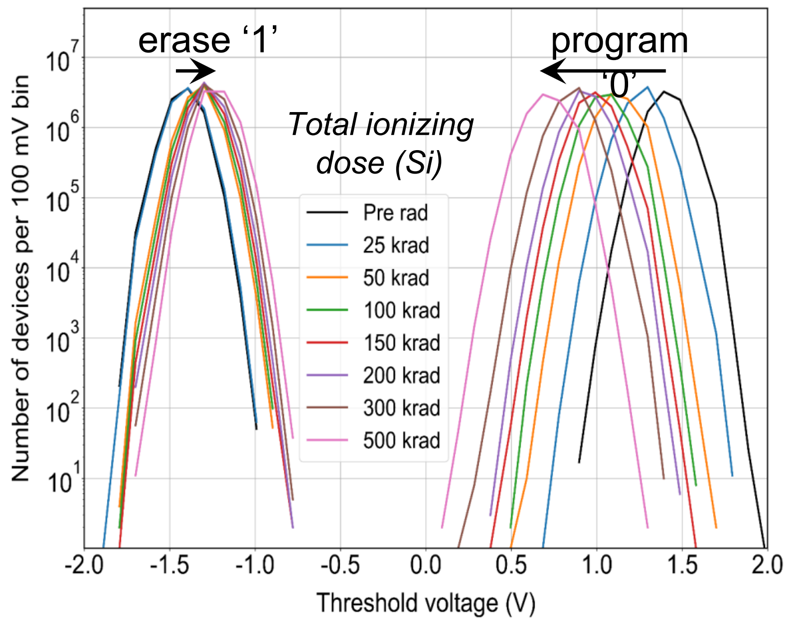


Ionizing Radiation

Uniform Gamma Irradiation

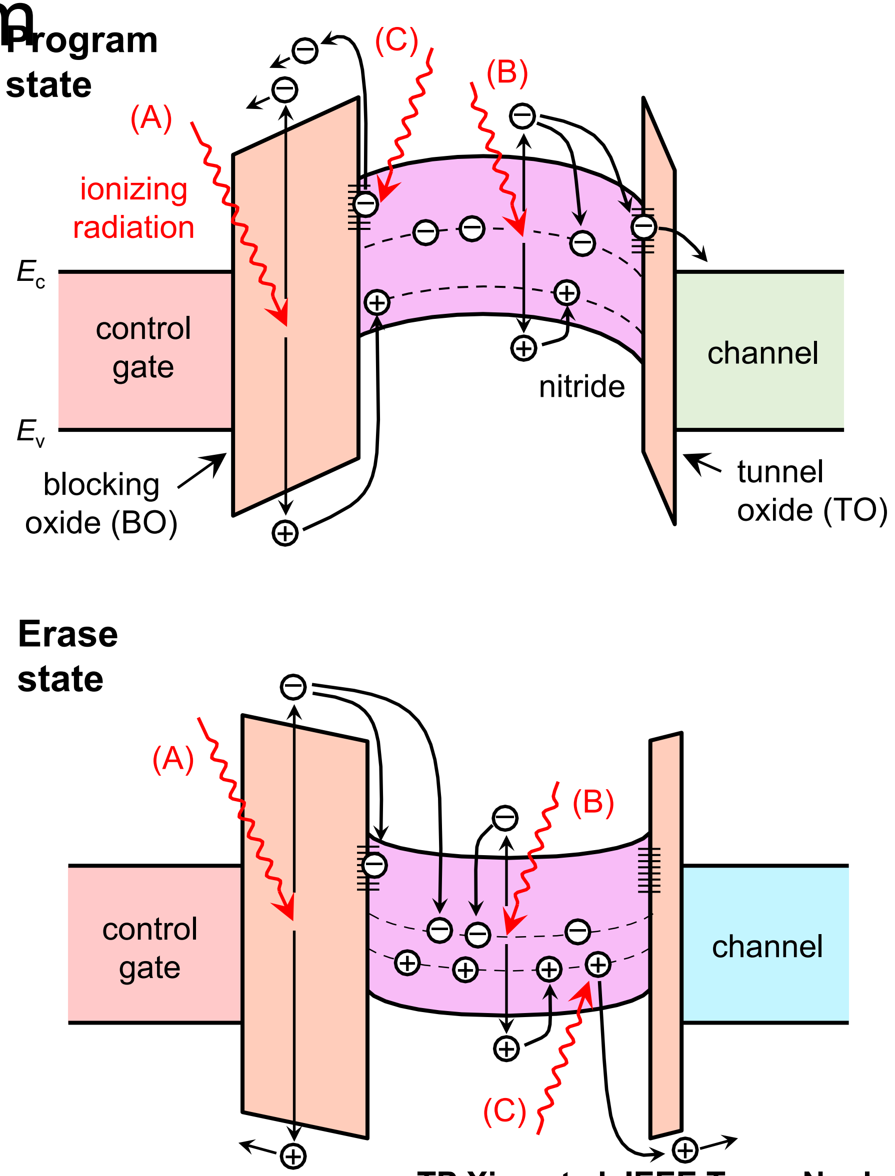


Threshold Distribution Shifts Across Array

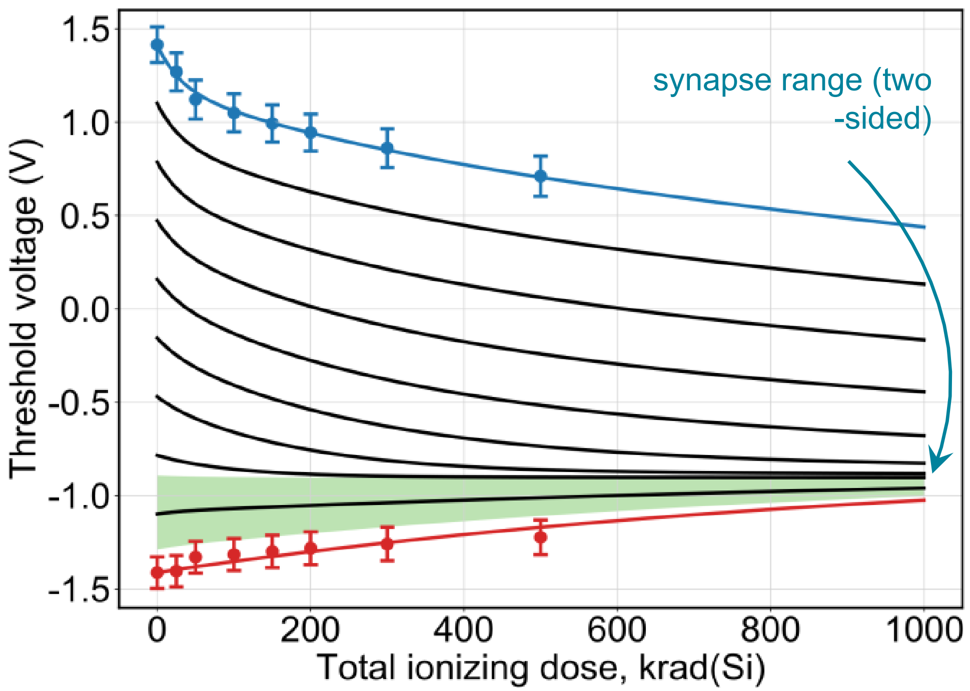


TP Xiao et al, IEEE Trans Nuclear Sci, 2021 (in press).

Analog Neuromorphic SONOS In Space: Physics to Algorithm



V_T versus Total Ionizing Dose: Model and Experiment

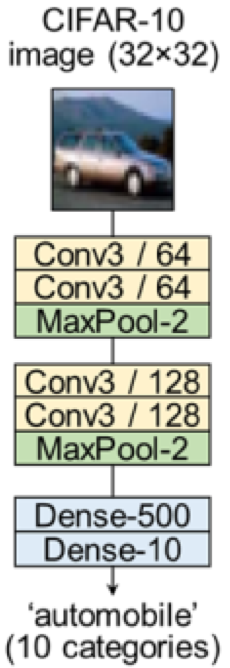


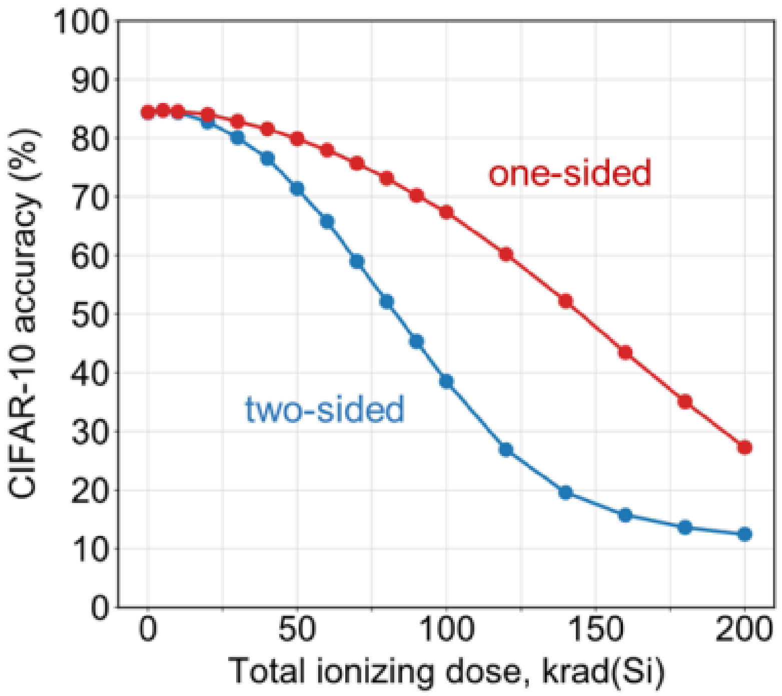
TP Xiao et al, IEEE Trans Nuclear Sci, 2021 (in press).

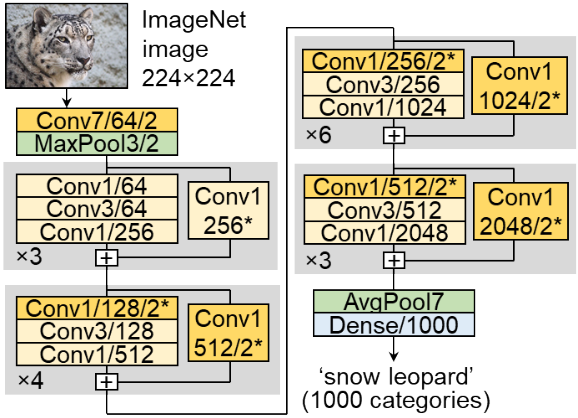
Analog Neuromorphic SONOS In Space: Physics to Algorithm

How will the accuracy degrade in space?

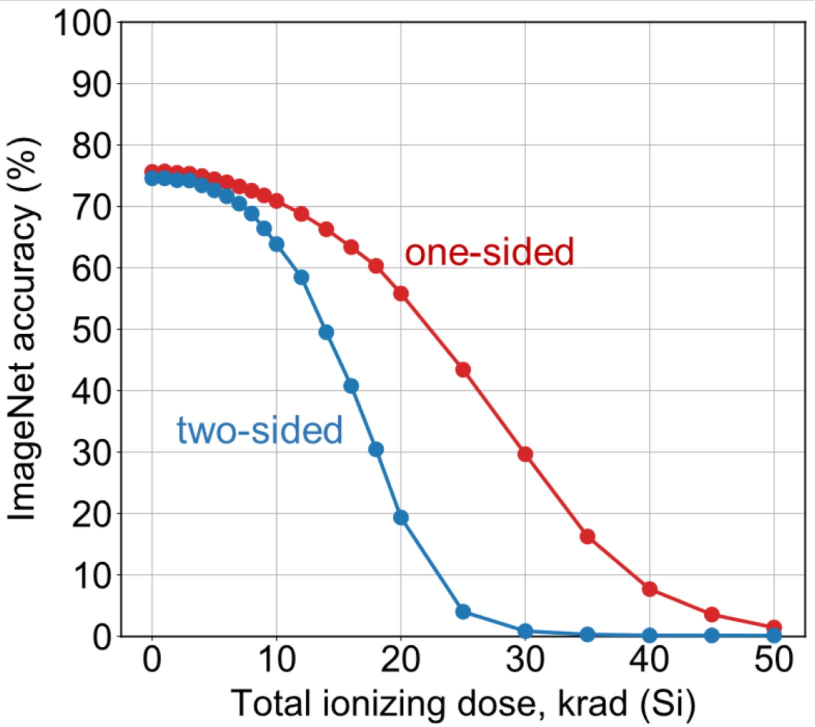
6-layer CNN for CIFAR-10
4.36M weights, 100.4M ops







ResNet-50 for ImageNet
25.6M weights, 4.1B ops



CoDesign provides insight for fielding neuromorphic devices

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

Neural Network Basics

Basic Building Block

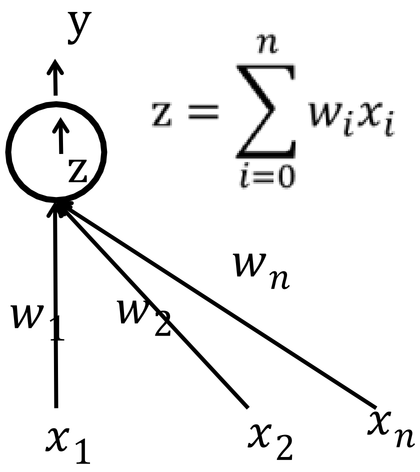
$$y = \frac{1}{1+e^{-z}}, ReLU, etc.$$

Neuron (activation function)

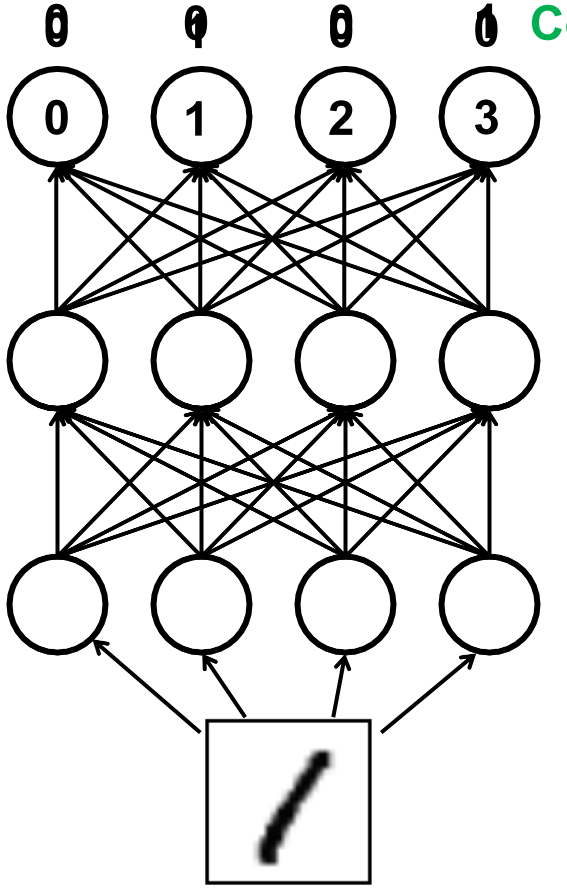
$$z = \sum_{i=0}^{n} w_i x_i$$

Weights (synapses)

Inputs

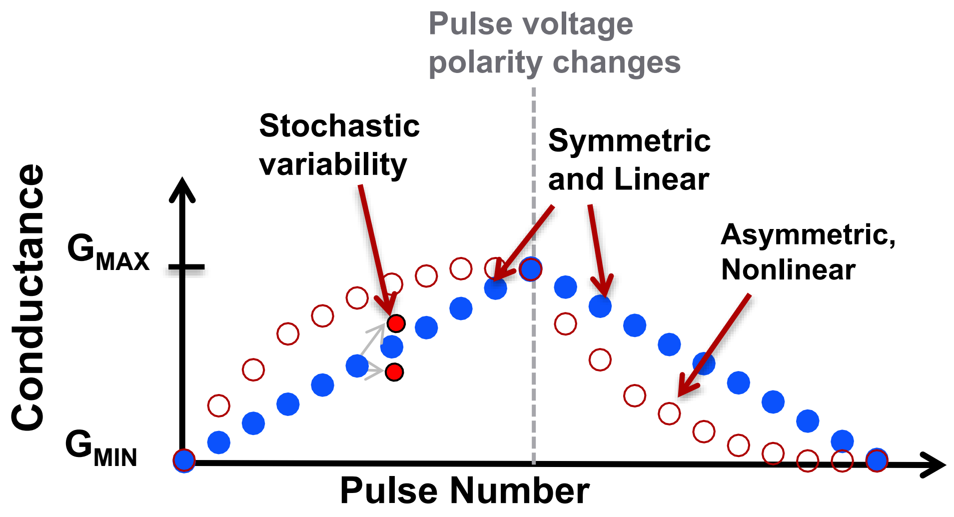


Simple Network: Inference & Training (Backpropagation)

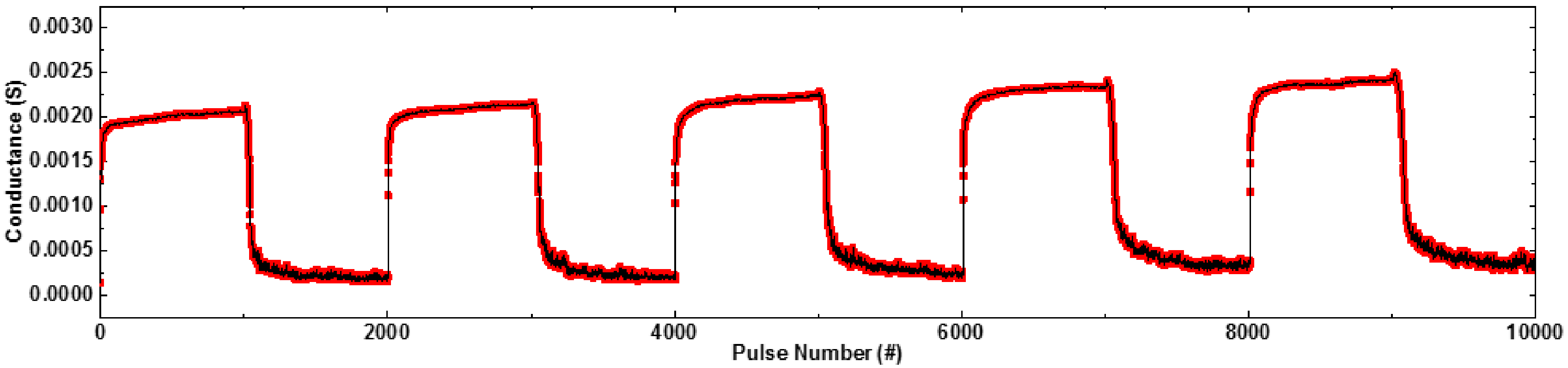


Device Challenges for Training

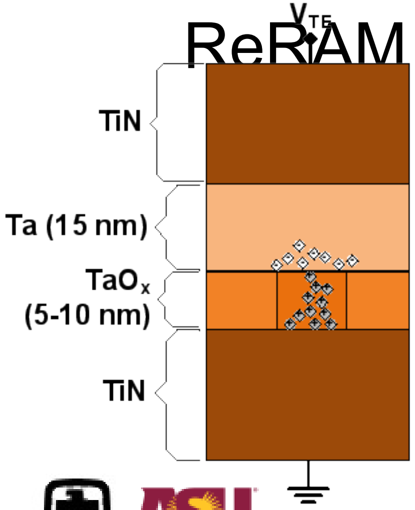
- **Training has an overlapping set of challenges**
- **Ideally weight increases and decreases linearly proportional to learning rule result**
- **Issue for open loop nonvolatile memory: altered the relationship between intended and actual update**
  - **Nonlinear and asymmetric state change**
  - **Cycle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



Characterization for Training

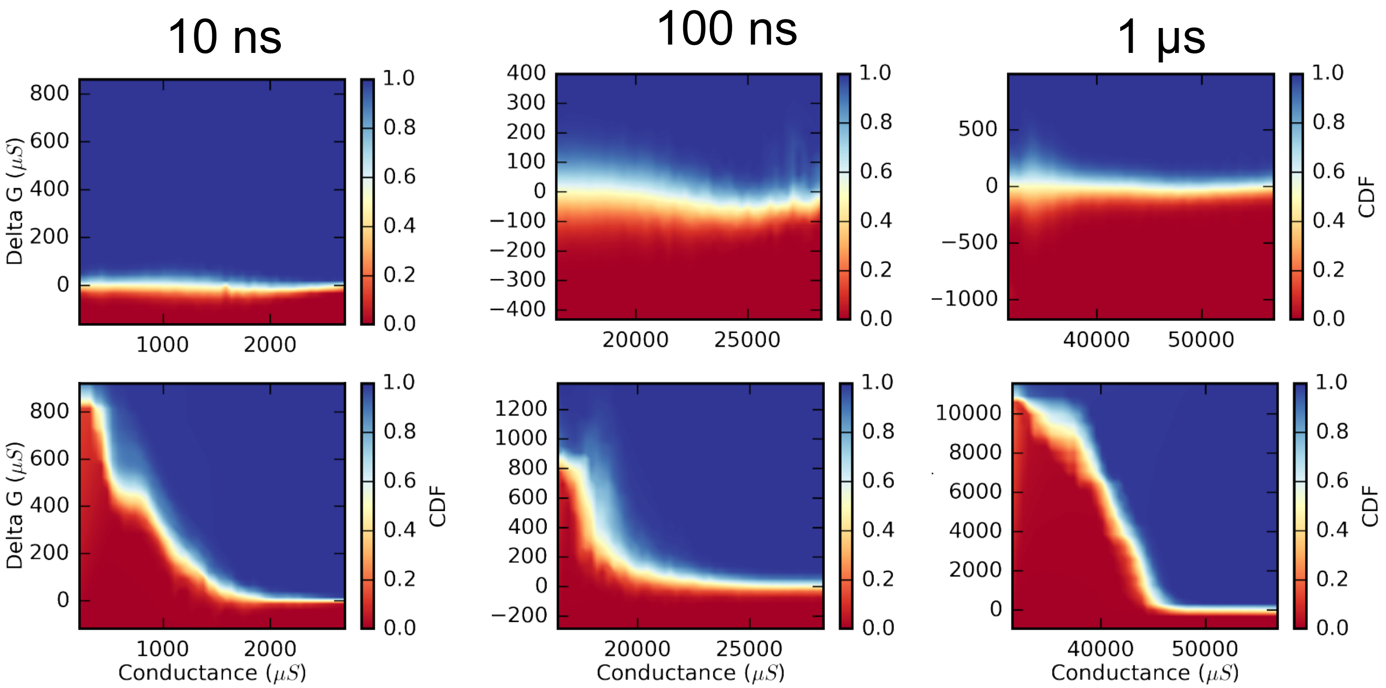


SNL TaOx ReRAM

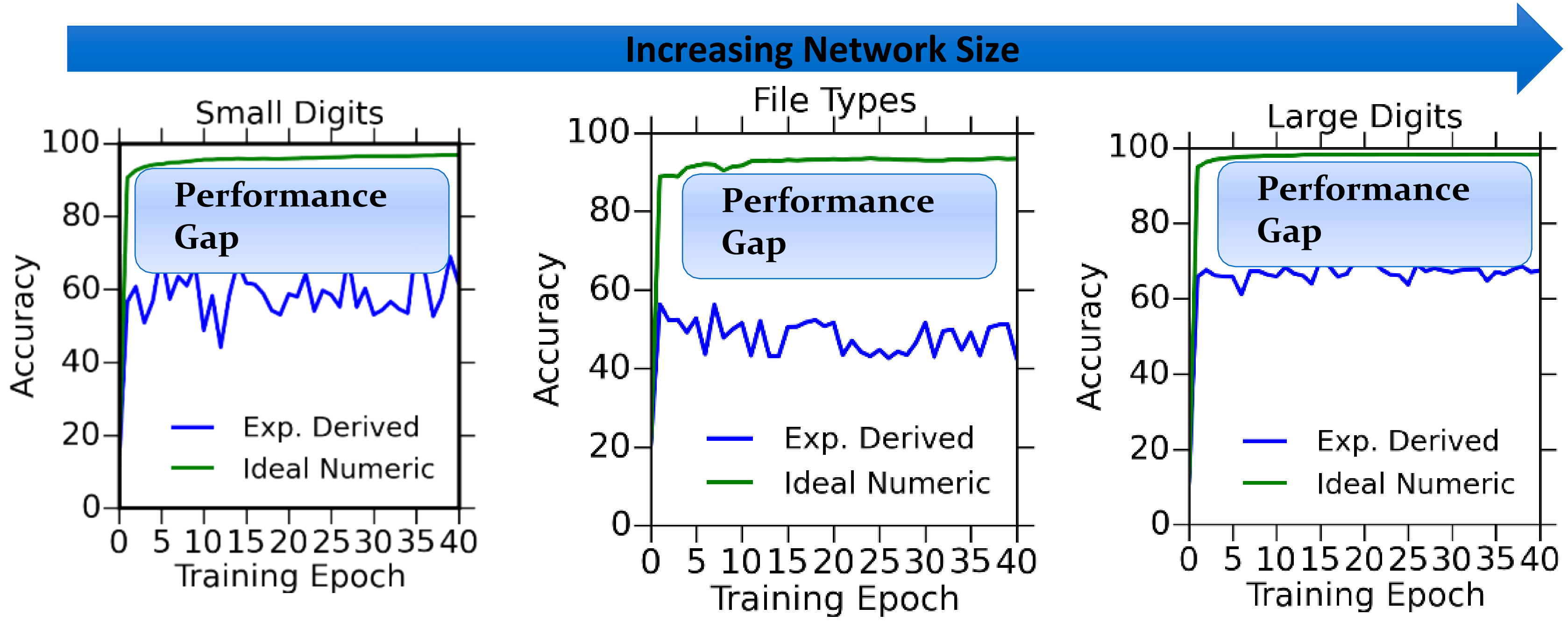


RESET

SET

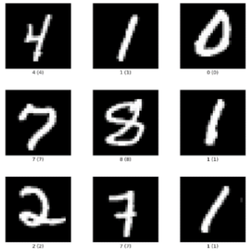


Initial TaOx ReRAM Training Accuracy Modeling (MNIST)



Unexpected results:

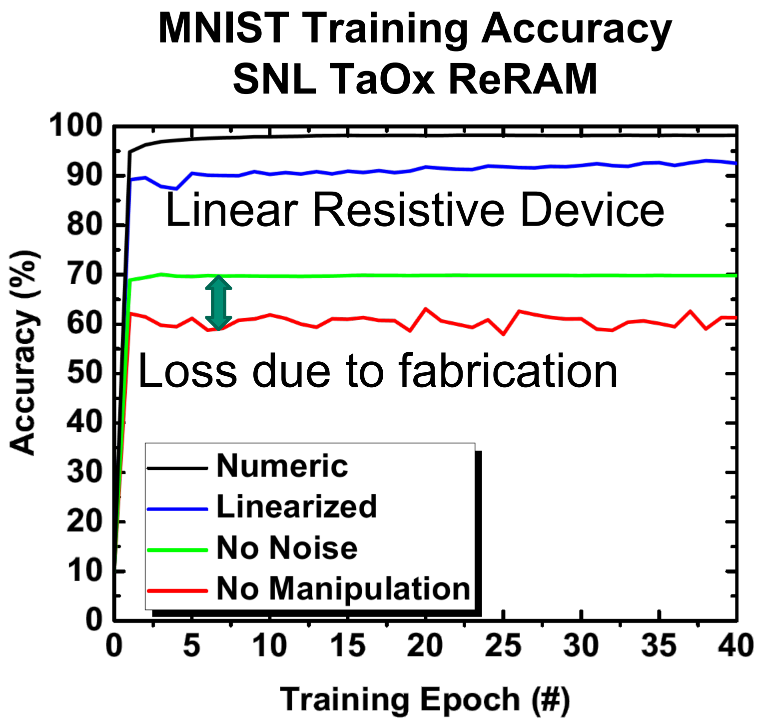
TaOx ReRAM was not ideal device for open loop training



ROSS SIM

Training is significantly more challenging than inference!

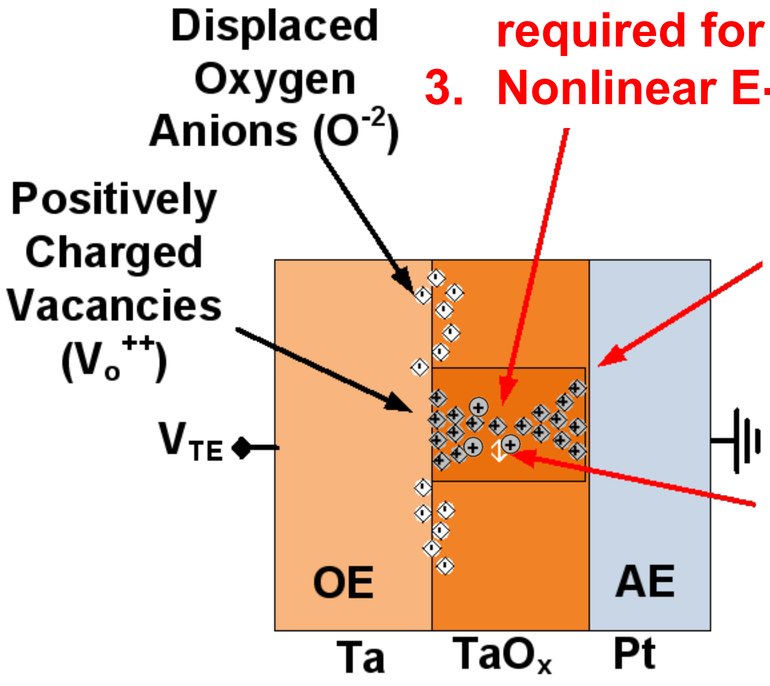
Physical Insight from Multiscale Model - CrossSim
Challenges using Filamentary ReRAM for Training



CROSS SIM

Nonlinearity

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



Asymmetry

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

Stochasticity

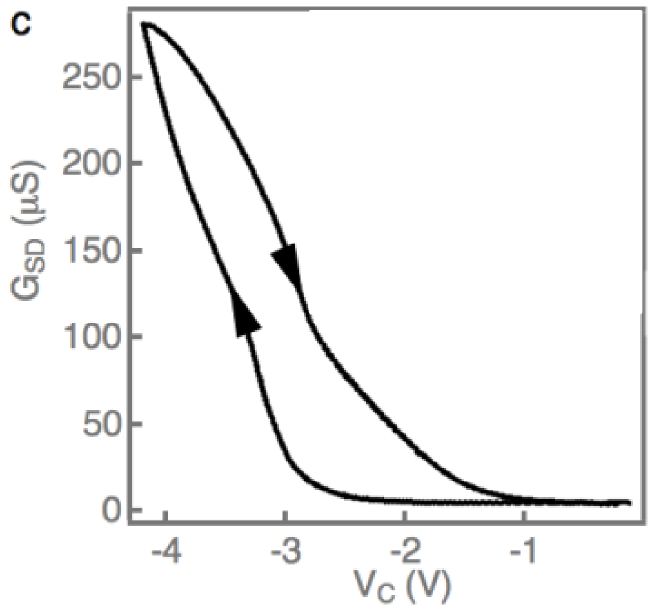
G depends on position of a few atoms

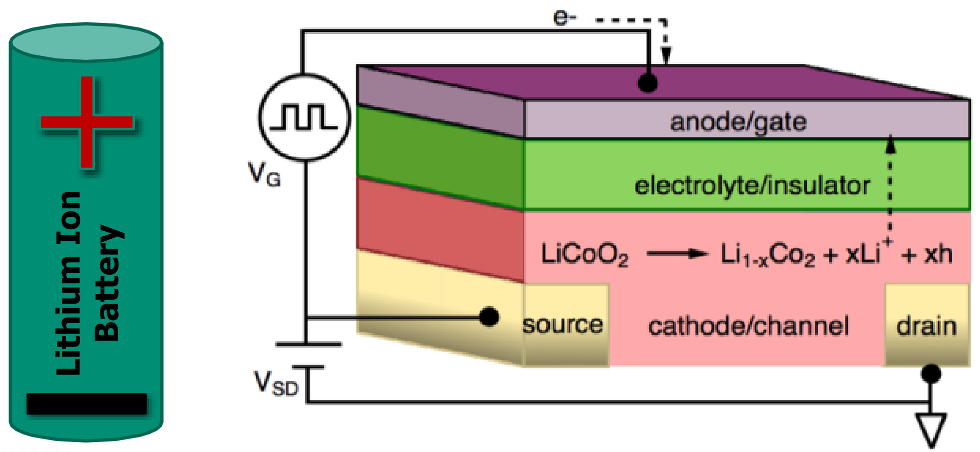
R. Jacobs-Gedrim et al, Proc. 2017 IEEE ICRC, 2017.

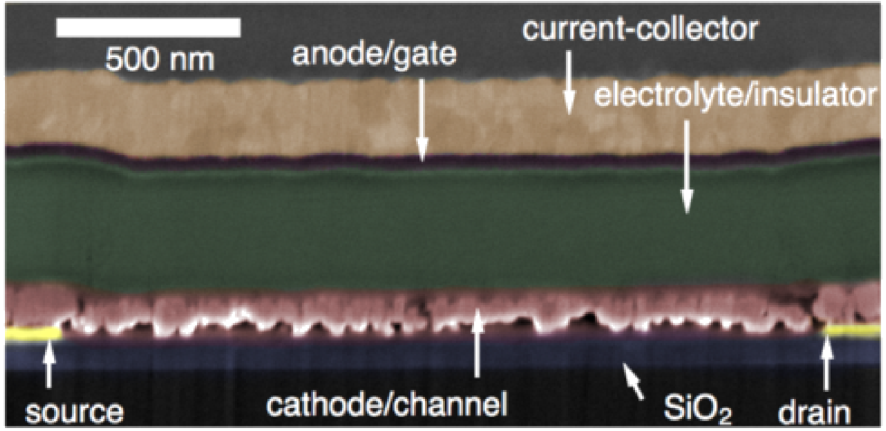
Electrochemical RAM (ECRAM) Synapse

- 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

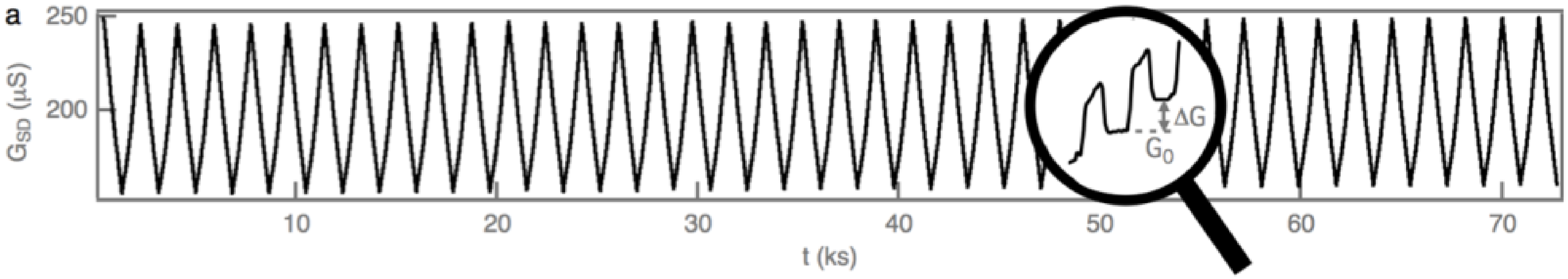




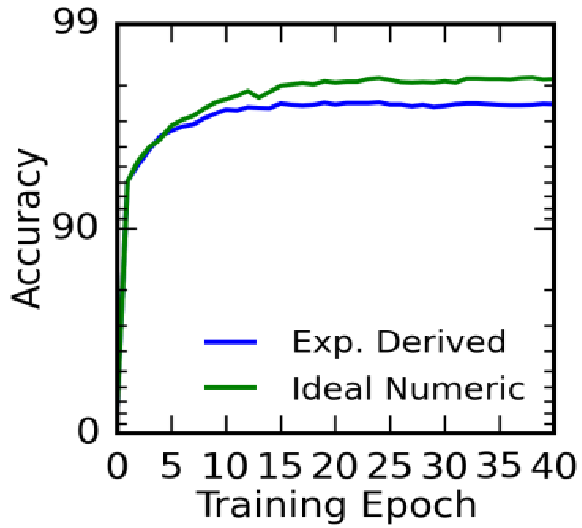


E. Fuller et al, *Adv Mater*, 2017

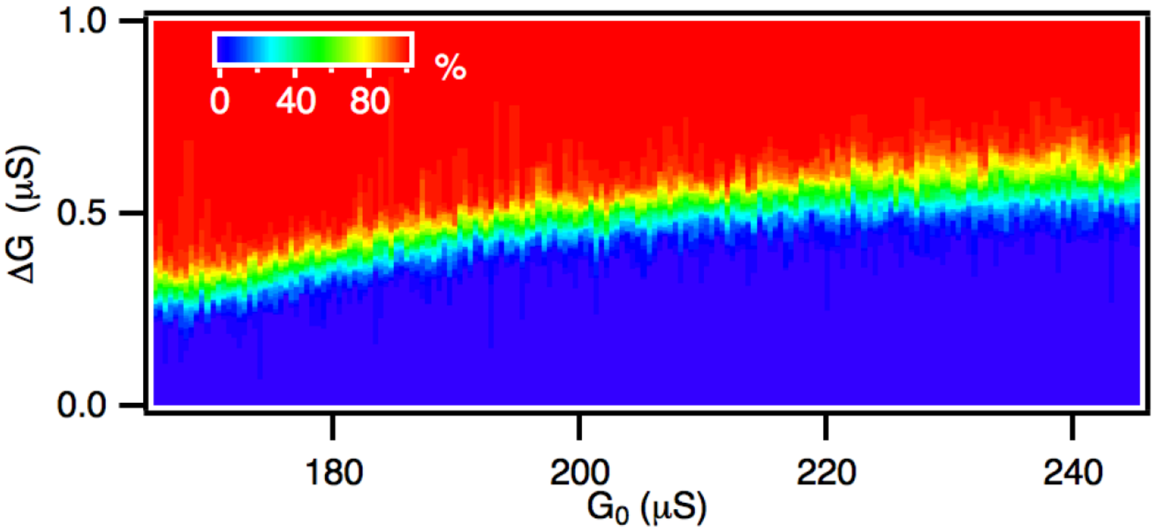
Analog State Comparison



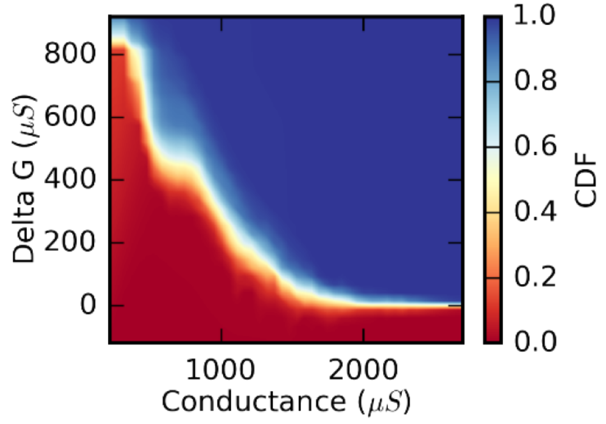
ECRAM-MNIST



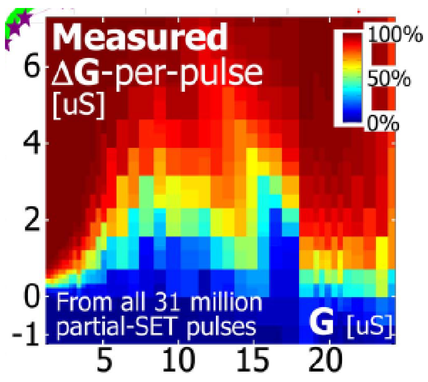
ECRAM



TaOx ReRAM



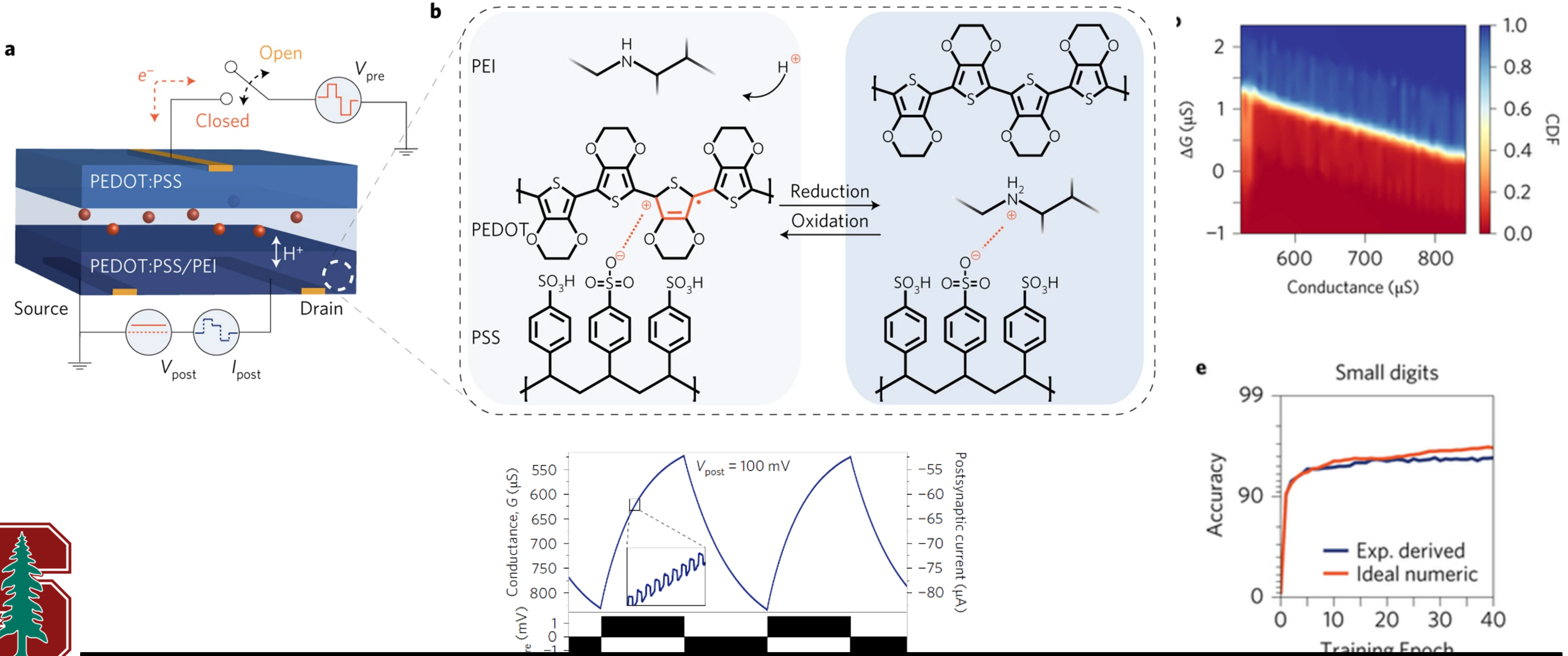
PCM Array



GW Burr et al, IEEE TED 2015

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

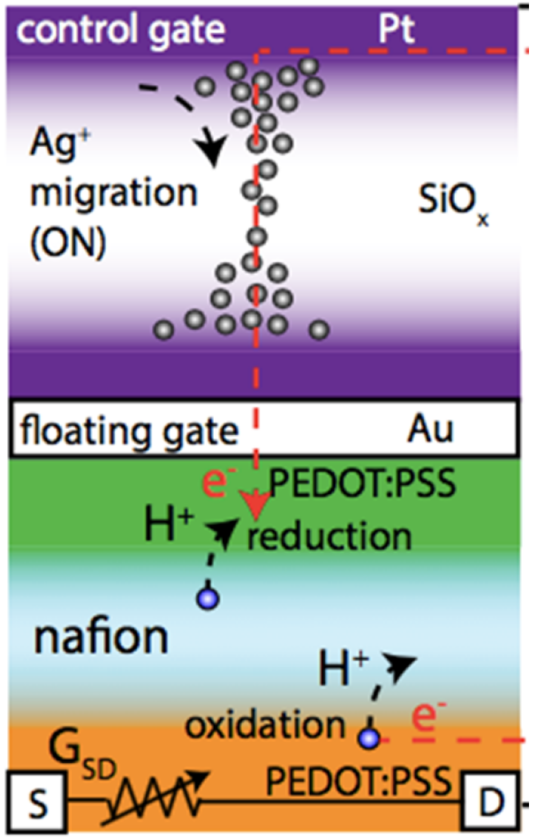
Electrochemical Neuromorphic Organic Device (eNode)

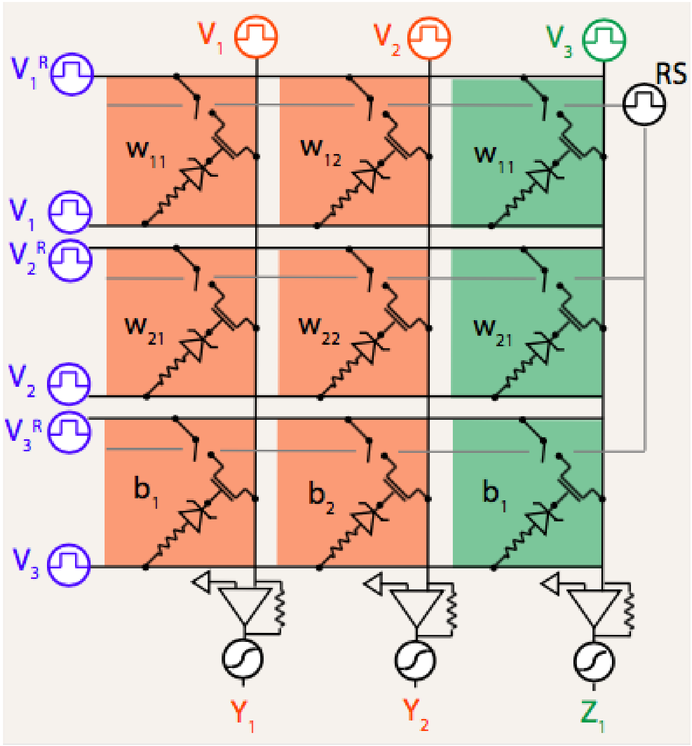


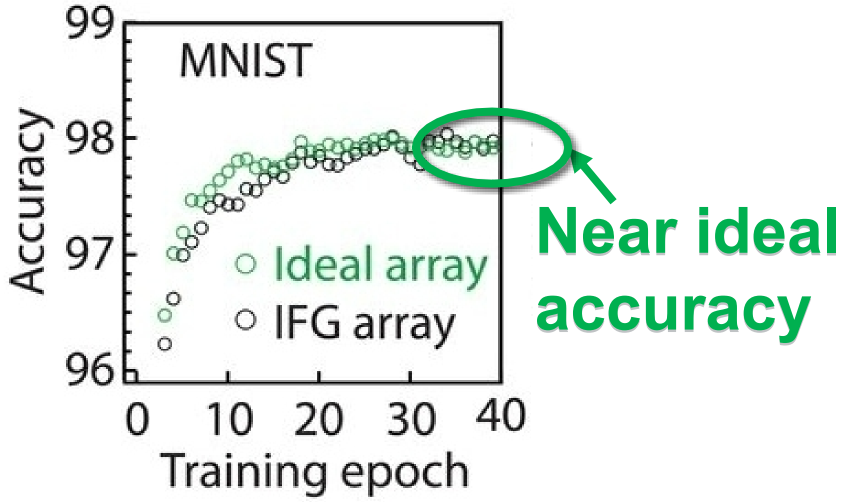
Proton-based polymer ECRAM synapse: fast, better endurance

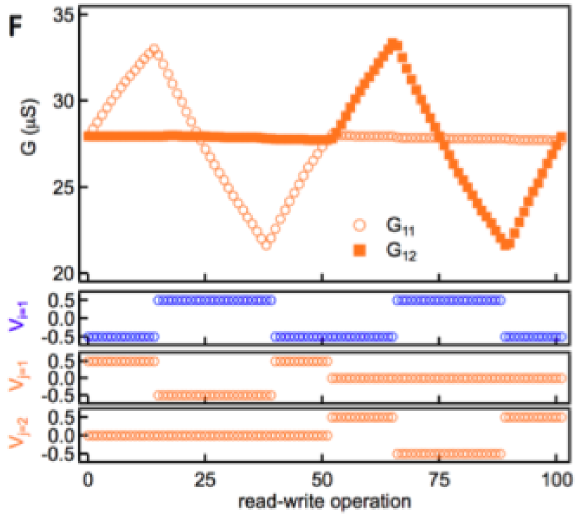
van de Burgt et al, *Nature Mater.*, 16, 414, 2017

ECRAMs Array Parallel Update Training Demonstration





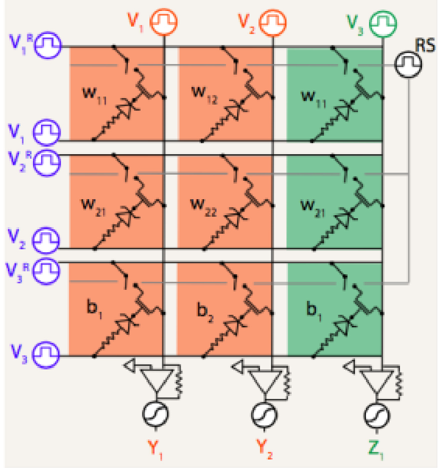


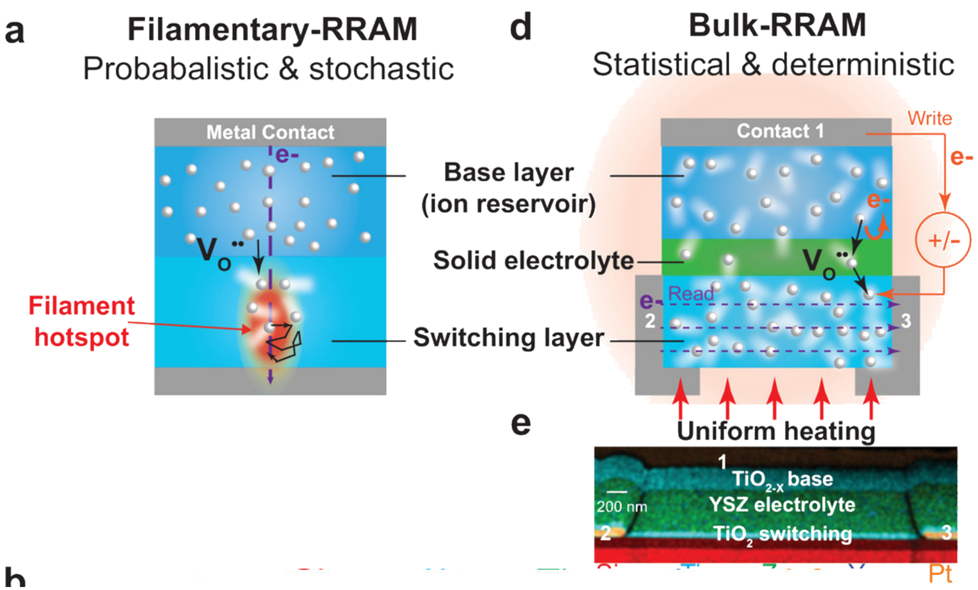


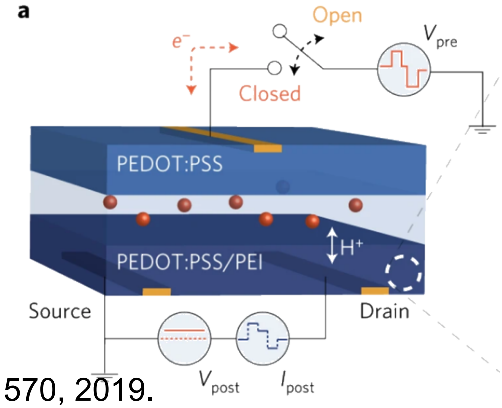
E. J. Fuller, S. T. Keene, A. Melianas, Z. Wang, S. Agarwal, Y. Li, Y. Tuchman, C. D. James, M. J. Marinella, J. J. Yang, A. Salleo, A. A. Talin, *Science* 364, 570, (2019).

Novel Devices for Accurate Inference/Training

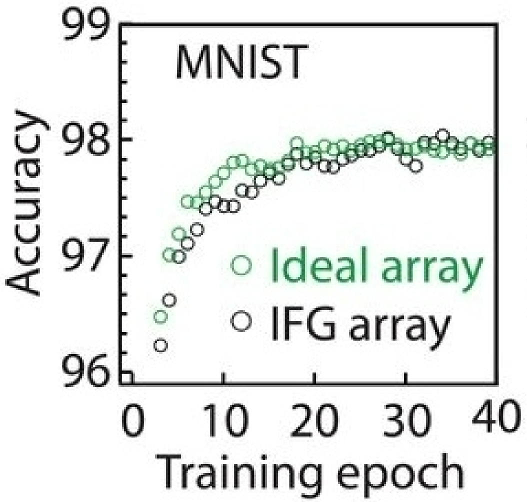
- Battery Inspired Devices
  - **Bulk ReRAM**
  - Bulk nanoionic devices
  - Two terminal charge tunnel junction
  - Organic and bio compatible switches
- Linear write: dG does not depend on G.
  - Ideal for training



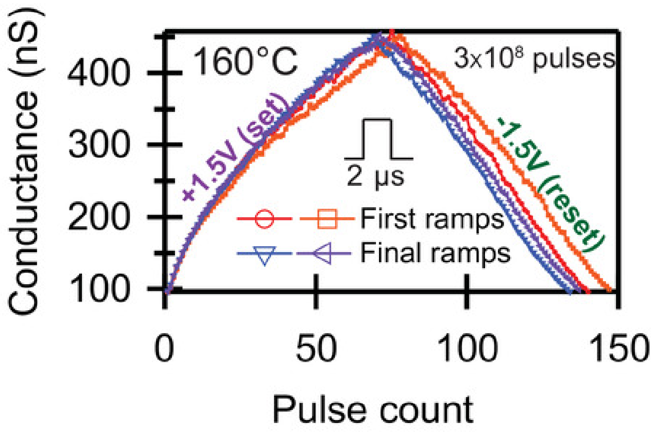




E. J. Fuller et al, *Science* 364, 570, 2019.



Y. Li et al, Adv Mater. 2020.



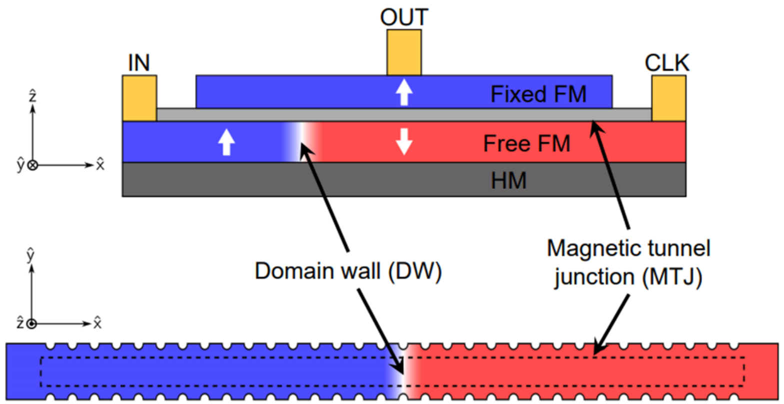
Y. Li et al, Adv Mater. 2020.

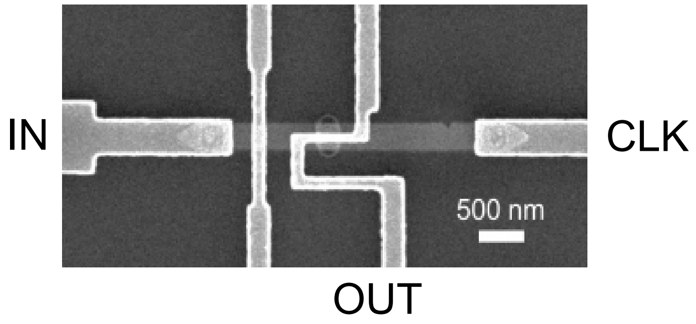
Collaboration with

Exploration of Novel Magnetic Synapses for Training

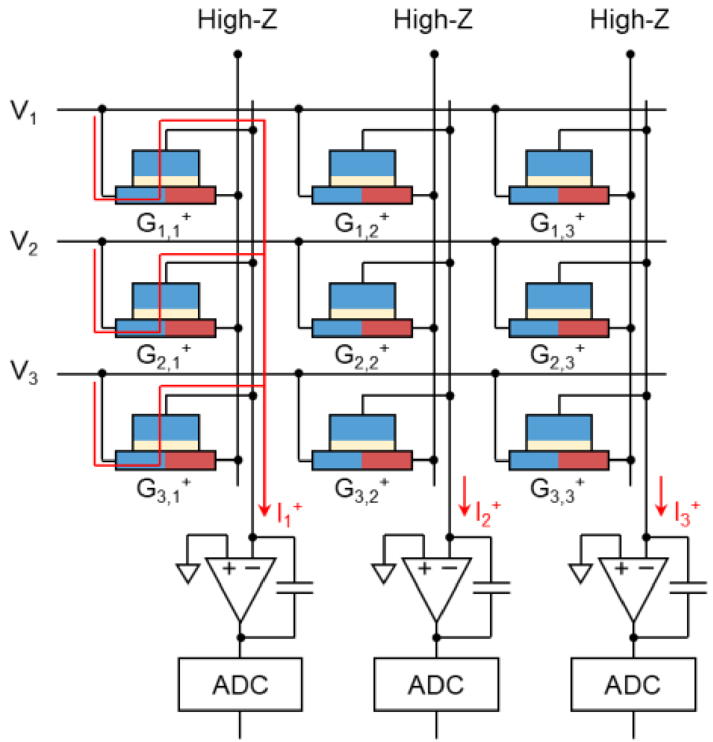
INC Lab
Professor
Jean Anne Incorvia

Low-energy domain wall synapses

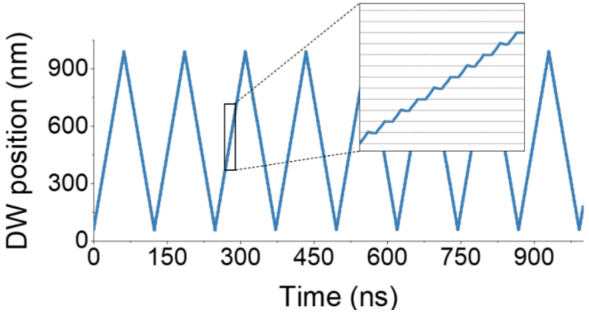


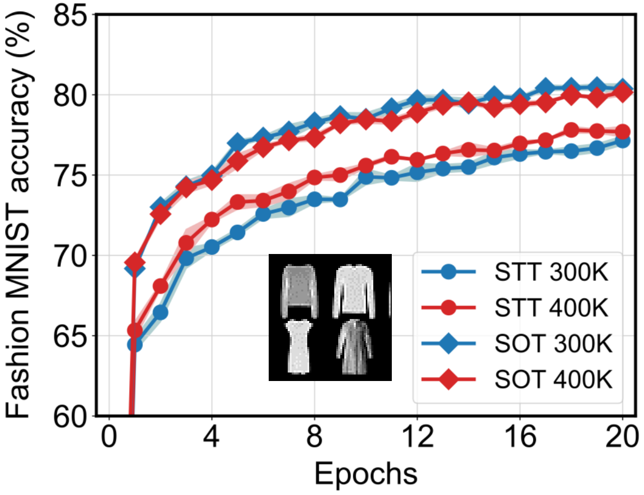


Integration in a crossbar array



Demonstrating linear updates and neural network accuracy





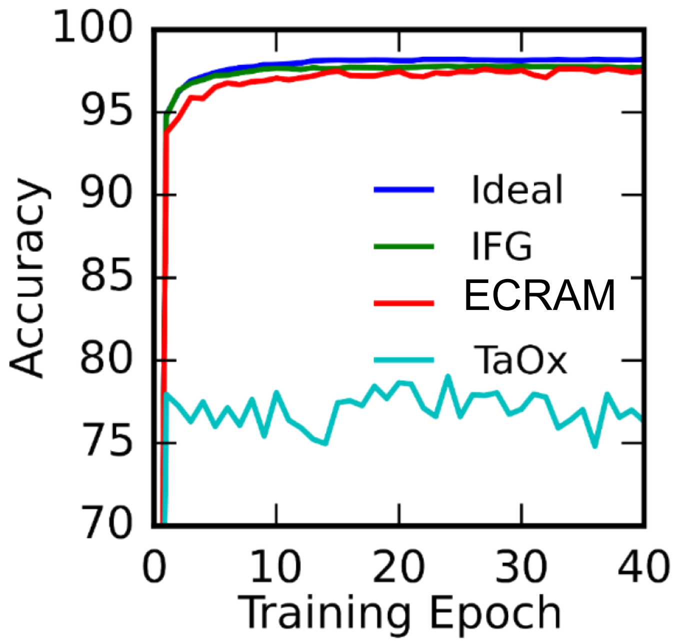
S. Liu et al, Apl Phys Lett, 2021 (under revision)
M. Almadar, Appl. Phys Lett, 2021 (in press)

Training Accuracy and Tile Energy/Summary

ECRAM: Use for training & inference

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

ReRAM: Training is not accurate: better for inference

Comparison of State of the Art Accelerators

TABLE II. Comparison of selected digital and mixed-signal neural network inference accelerators from industry and research.[a] 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^2 values are based on the die area, where provided.

	NVIDIA T4[175]	Google TPU v1[22,b]	Habana Goya HL-1000[176]	DaDianNao[44]	UNPU[51]	Reference 122 mixed-signal[c]
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[d]	1 bit
Clock speed	2.6 GHz	700 MHz	2.1 GHz (CPU)	606 MHz	200 MHz	10 MHz
Benchmarked workload	ResNet-50[177] (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^2)	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

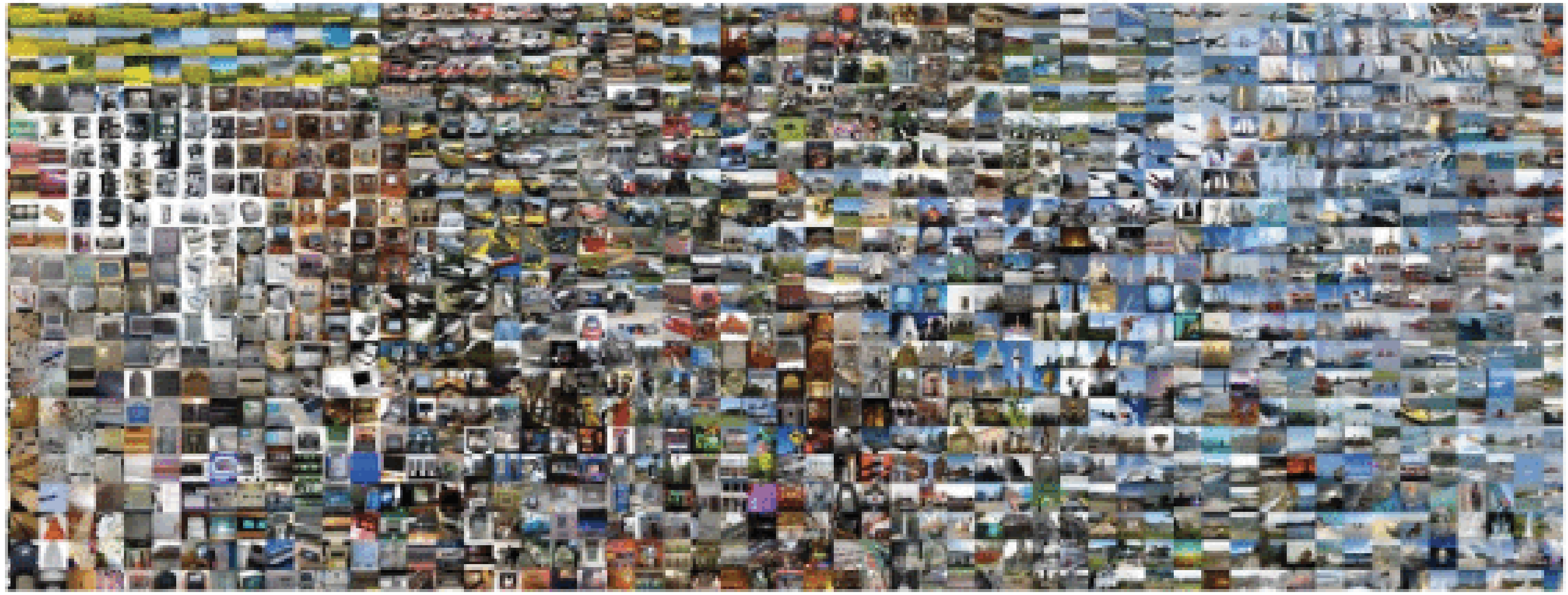
[a]To enable performance comparisons across a uniform application space, we did not consider accelerators for spiking neural networks.
[b]The TPU v2 and v3 chips, which use 16-bit floating point arithmetic, are commercially available for both inference and training on the cloud. MLPerf inference benchmarking results for the Cloud TPU v3 are available,[179] but power and area information is undisclosed. The TPU v1 die area is taken to be the stated upper bound of 331 mm^2; the listed TOPS/mm^2 values are therefore a lower bound.
[c]The mixed-signal accelerator in Ref. 122 performs multiplication using digital logic and summation using analog switched-capacitor circuits.
[d]The UNPU architecture flexibly supports any weight precision from 1 to 16 bits. The results are listed for 1-bit weights.

Example Standard Visual Recogntion Datasets

MNIST

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

ImageNet



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

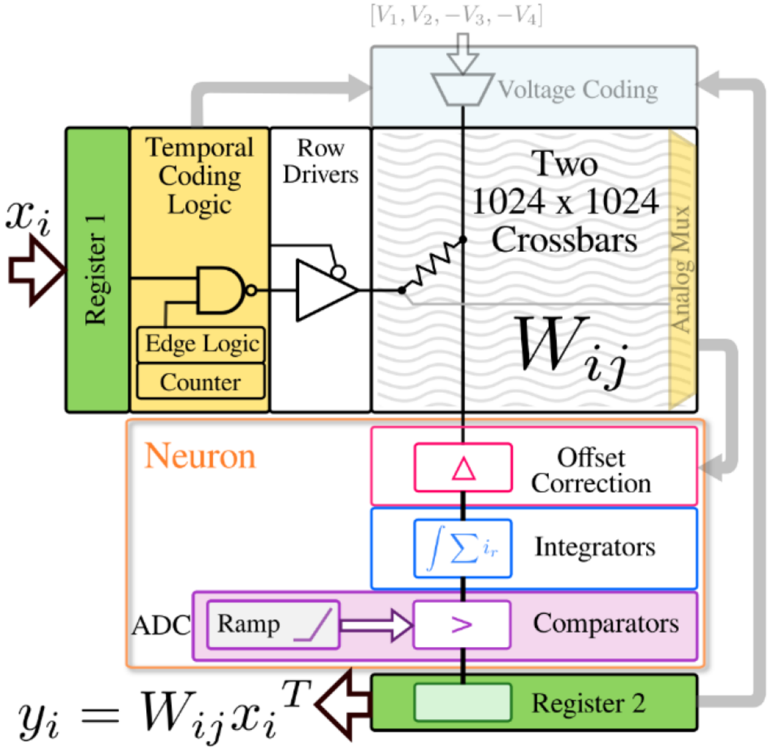
VV. Sze, Y. Chen, T. Yang and J. S. Emer, Proc IEEE, vol. 105, no. 12, pp. 2295-2329, Dec. 2017

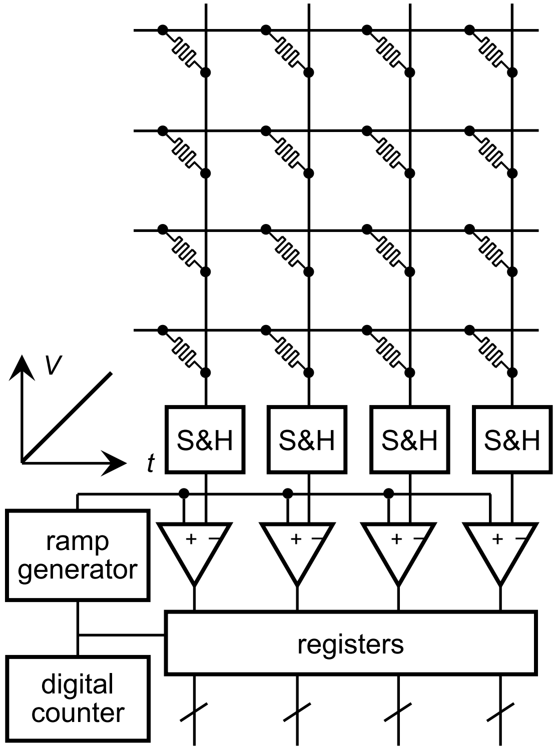
How much computing needs to be done?

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

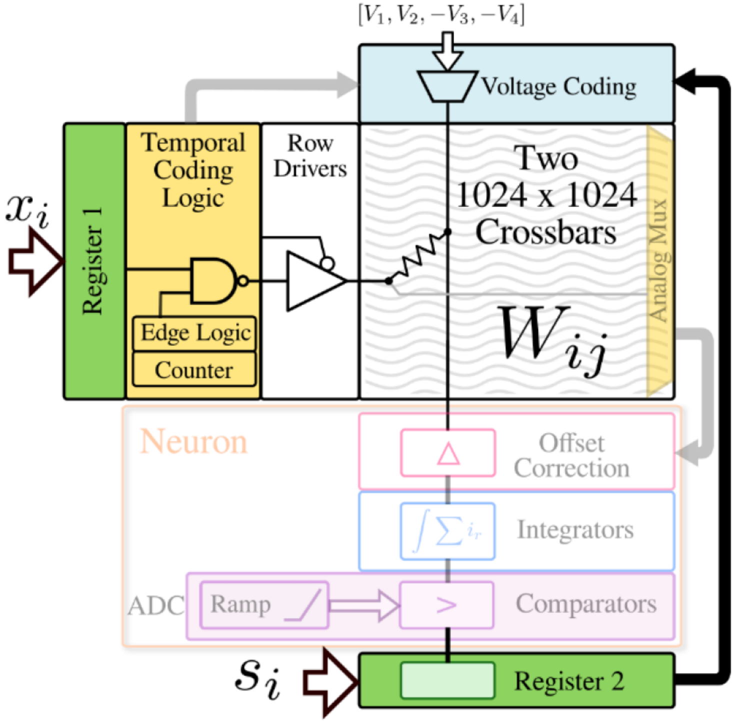
Key Circuit Block/Kernel Analysis

Vector Matrix Multiply (Inference)





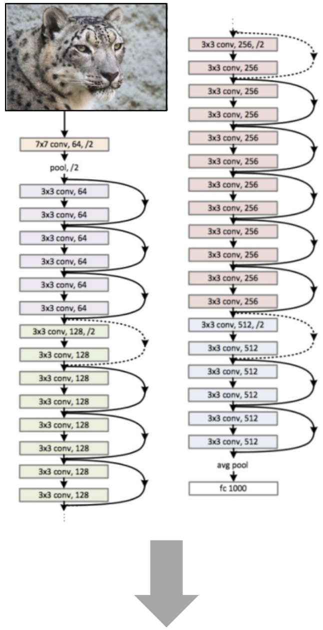
Rank-1 Update (Training)



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

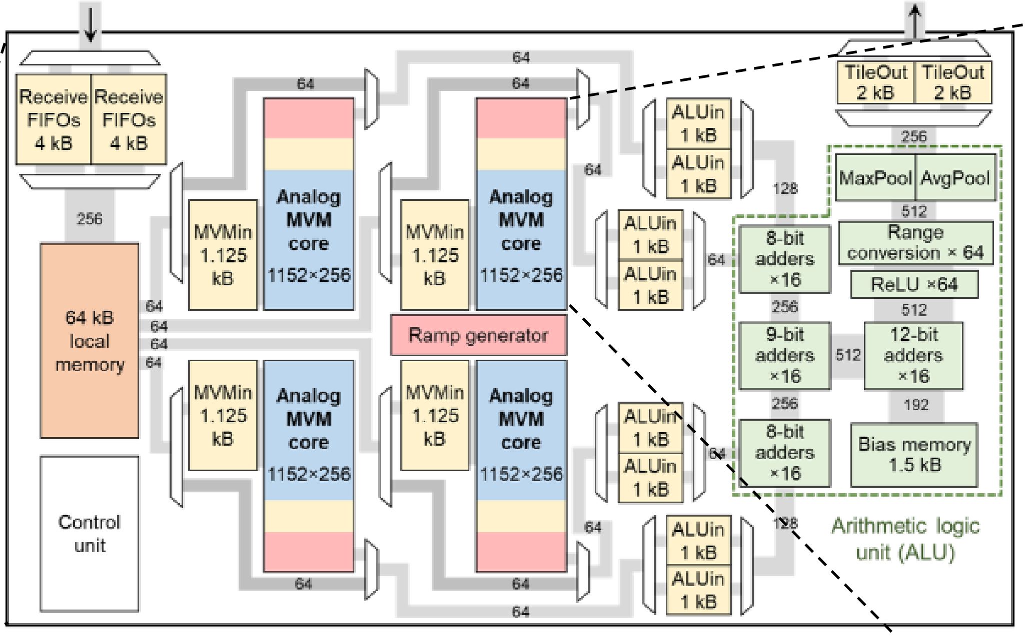
Neural Network Inference Architecture

Neural network

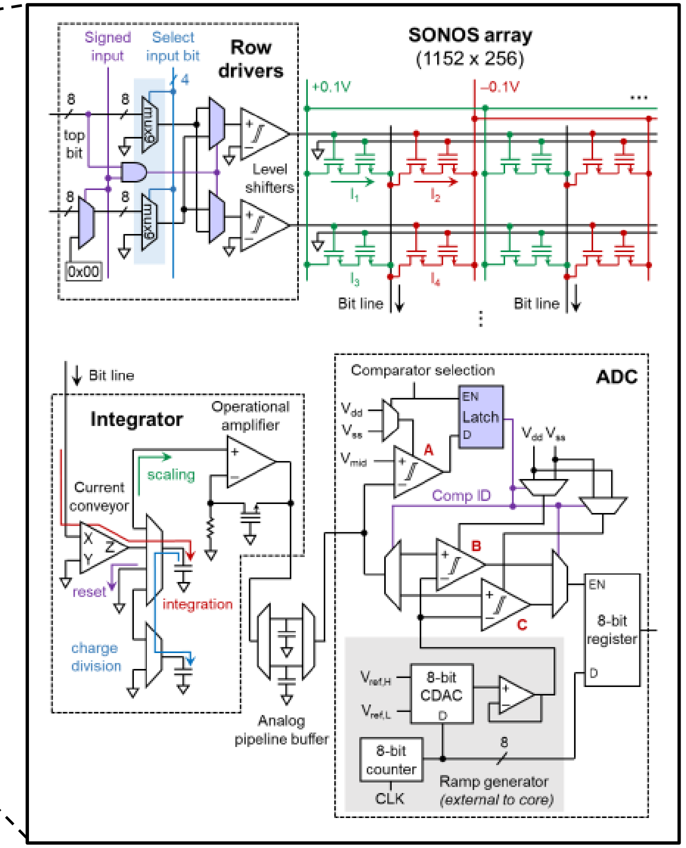


Mesh architecture

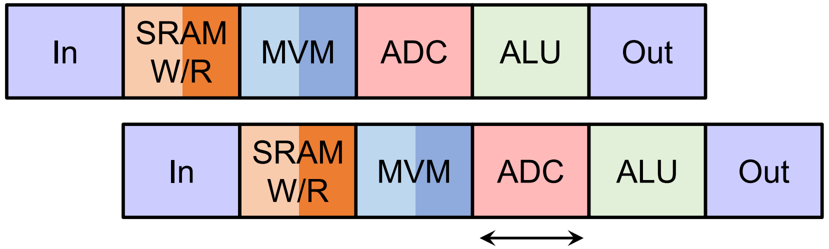
Pipelined MVM tile



Analog MVM core



Data batch 1



Data batch 2

295 clock cycles

Circuits designed and simulated using commercial 40nm PDK

T.P. Xiao et al, In preparation, IEEE J. Circuits and Systems, 2021.