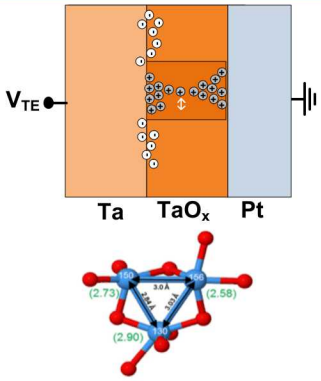
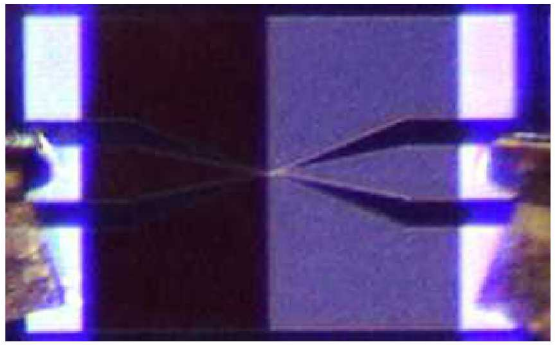
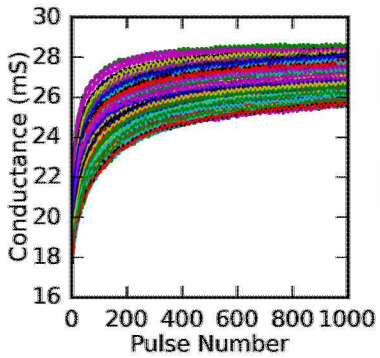
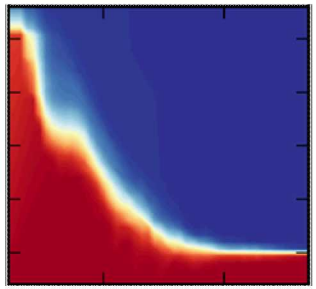
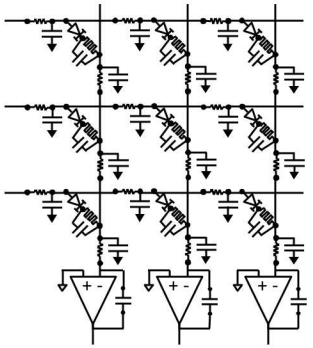
SAND2020-4460PE









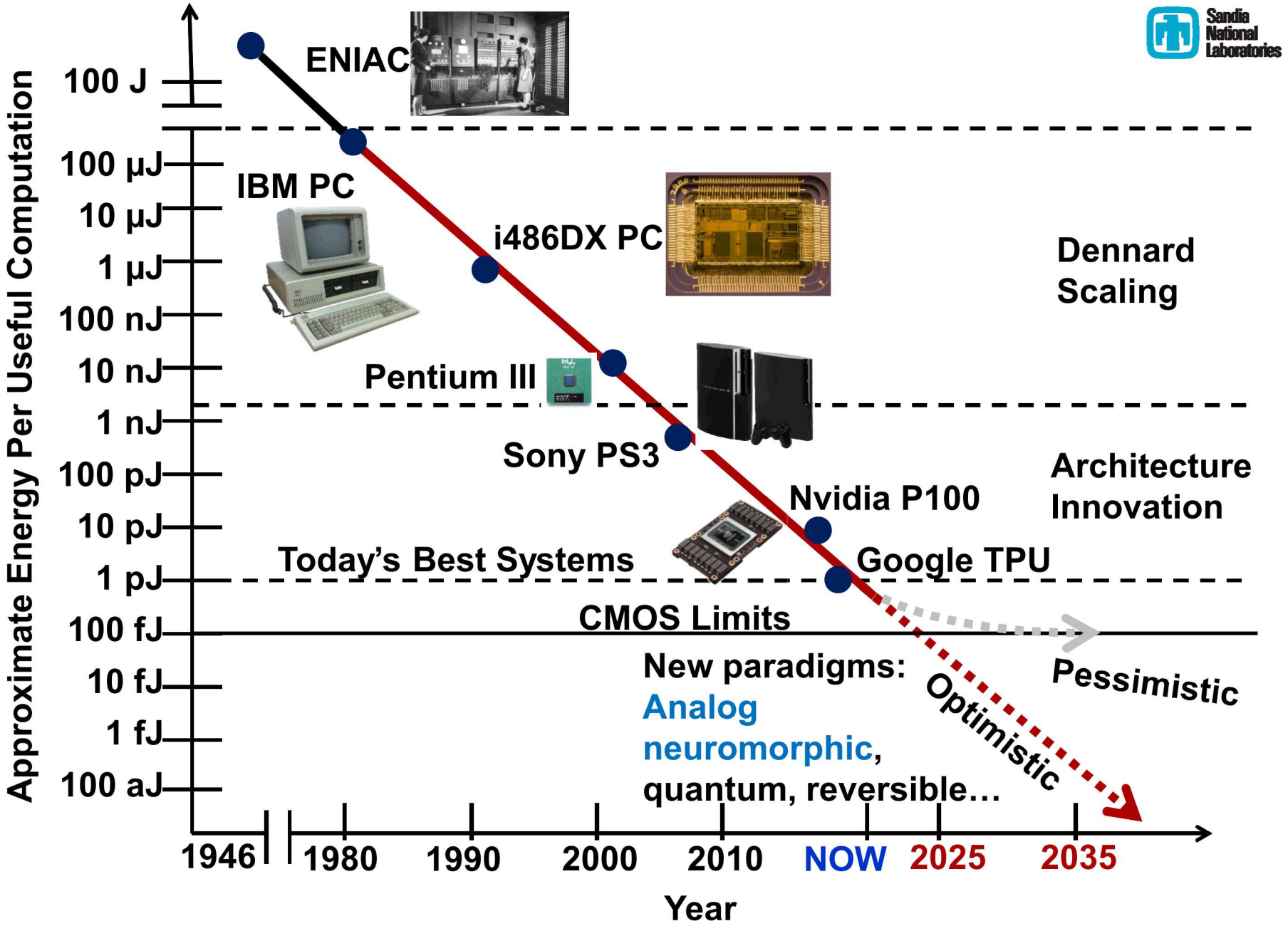


Neuromorphic Computing with Analog Nonvolatile Memory

Matthew J. Marinella
Sandia National Laboratories, Albuquerque, NM

Contributors: S. Agarwal, C. Bennett, R. Jacobs-Gedrim, E. Fuller, A.A. Talin, J.A. Incorvia (UT)

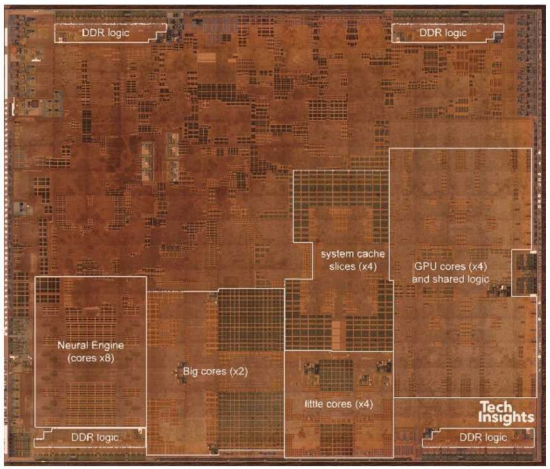
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

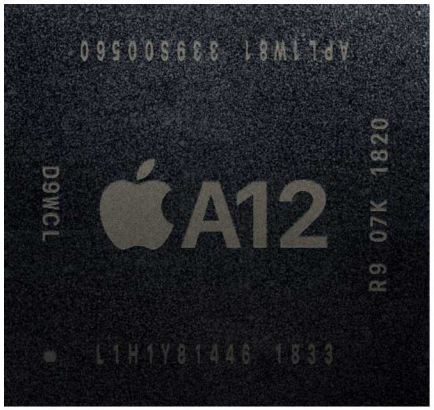

Sandia National Laboratories
Approximate Energy Per Useful Computation
100 J
100 µJ
10 µJ
1 µJ
100 nJ
10 nJ
1 nJ
100 pJ
10 pJ
1 pJ
100 fJ
10 fJ
1 fJ
100 aJ
ENIAC
IBM PC
i486DX PC
Pentium III
Sony PS3
Nvidia P100
Today's Best Systems
Google TPU
CMOS Limits
Dennard Scaling
Architecture Innovation
New paradigms: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

State of the Art: Apple A12

Sandia National Laboratories

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





apple.com, techinsights.com

Analog Computation with a Resistive Switch Crossbar

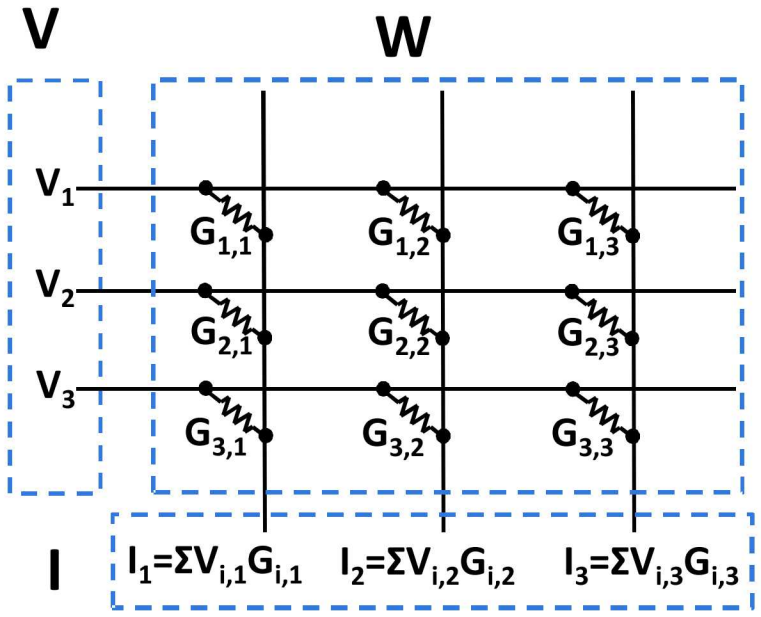
- Electronic Vector Matrix Multiply

Mathematical

$$V^T W = I$$

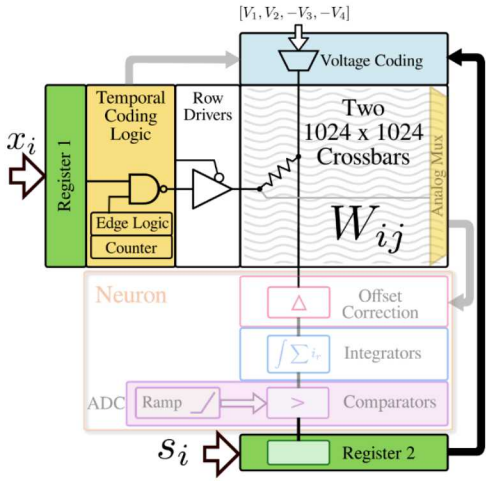
$$\begin{bmatrix} V_1 & V_2 & V_3 \end{bmatrix} \begin{bmatrix} W_{1,1} & W_{1,2} & W_{1,3} \\ W_{2,1} & W_{2,2} & W_{2,3} \\ W_{3,1} & W_{3,2} & W_{3,3} \end{bmatrix} = \begin{bmatrix} I_1=\Sigma V_{i,1}W_{i,1} & I_2=\Sigma V_{i,2}W_{i,2} & I_3=\Sigma V_{i,3}W_{i,3} \end{bmatrix}$$

Electrical



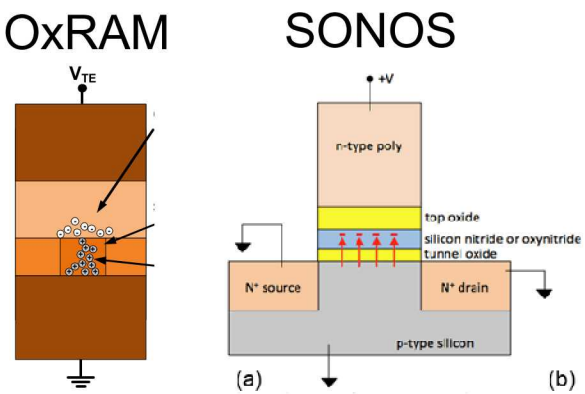
Let the physics do the computing

Matrix Energy/Latency Analysis

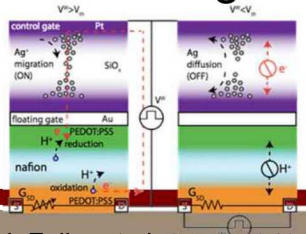


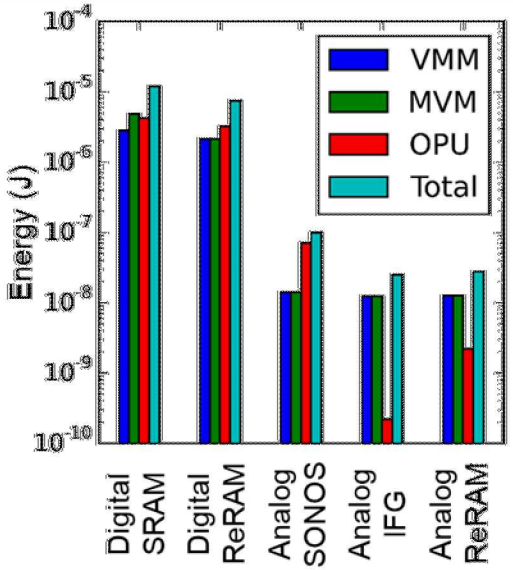
Marinella, et al, *IEEE JETCAS*, 2018

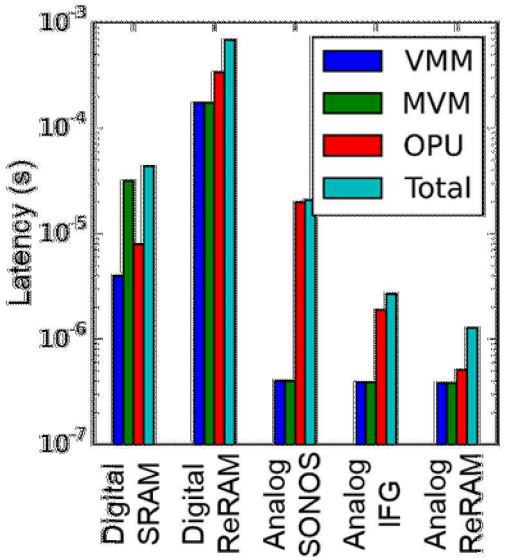
Component	Vector Matrix Multiply	Matrix Vector Multiply	Outer Product Update
Energy/Op IFG (fJ)	11.9	11.9	0.2
Energy/Op ReRAM (fJ)	12.2	12.2	2.1
Energy/Op SONOS (fJ)	13.7	13.7	68.2
Energy/Op SRAM (fJ)	2718	4630	4102
Array Latency IFG (µs)	0.39	0.39	1.9
Array Latency ReRAM (µs)	0.38	0.38	0.51
Array Latency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



Ionic Floating Gate



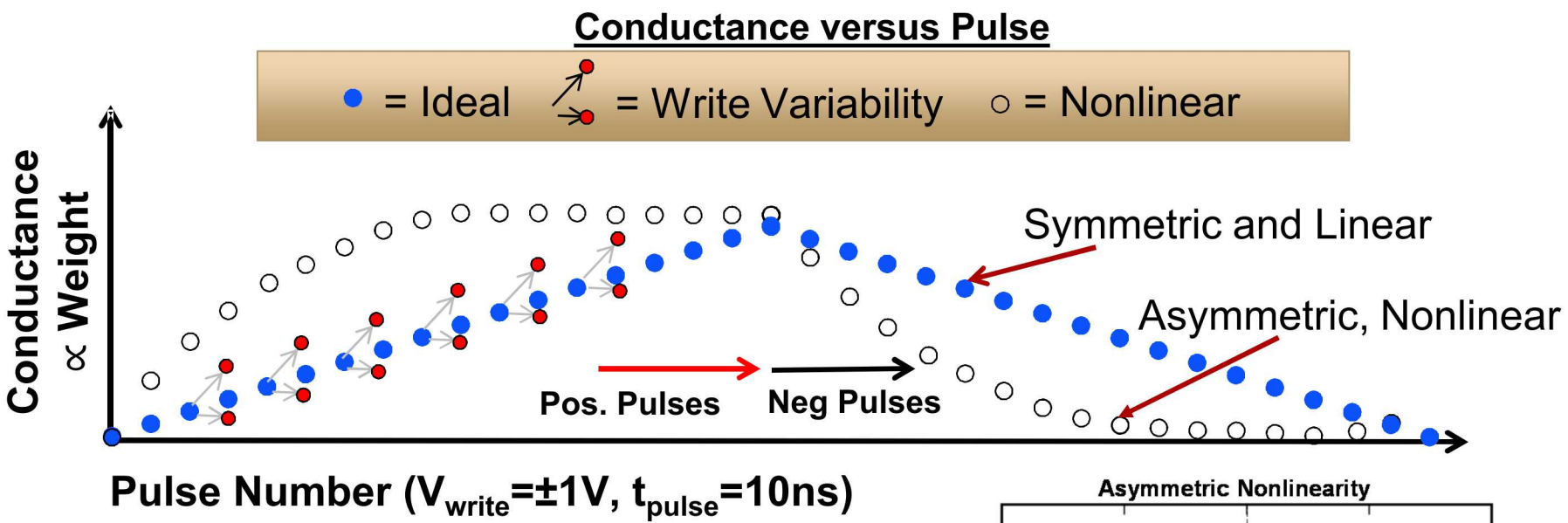




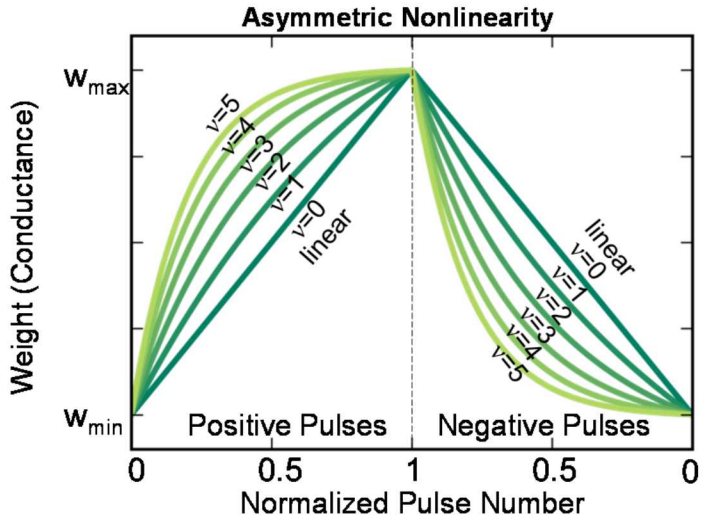
<2 fJ/op

Will enable training at the edge!

Analog Device Requirements

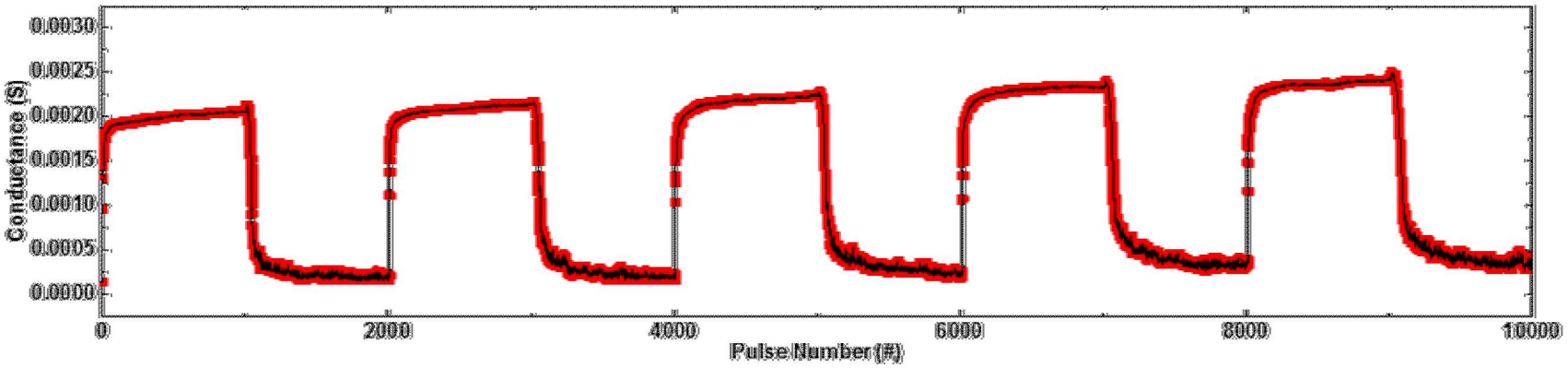


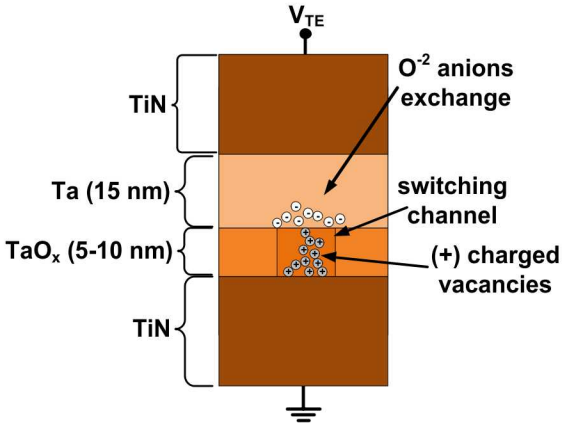
Read Noise σ (% Range)	5%
Write Noise σ (% Range)	0.4%
Asymmetric Nonlinearity (ν)	0.1
Symmetric Nonlinearity (ν)	5
Maximum Current	13 nA
Minimum Retention (@ 85°C)	7 days
Minimum Nudge Endurance	10^7

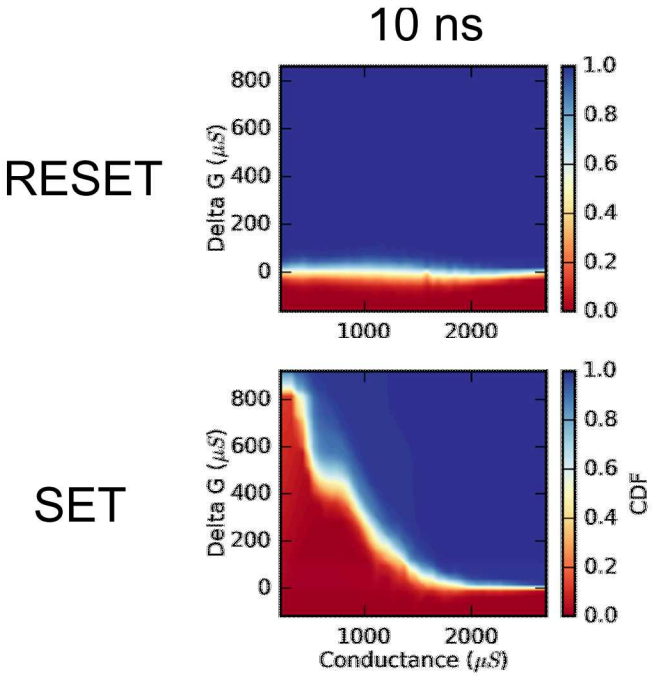


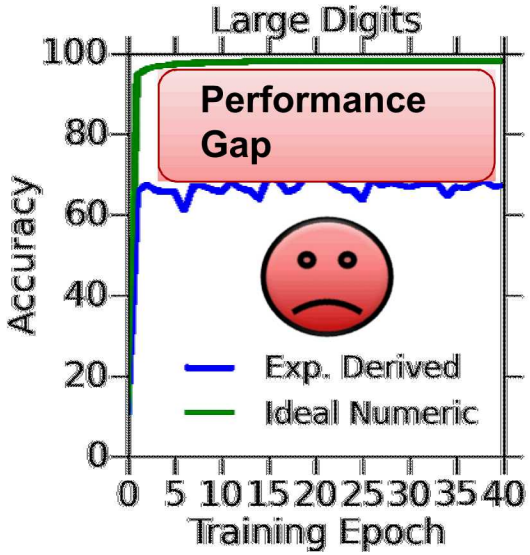
Device-Algorithm Interaction = Strict analog requirements

TaOx ReRAM Training Evaluation





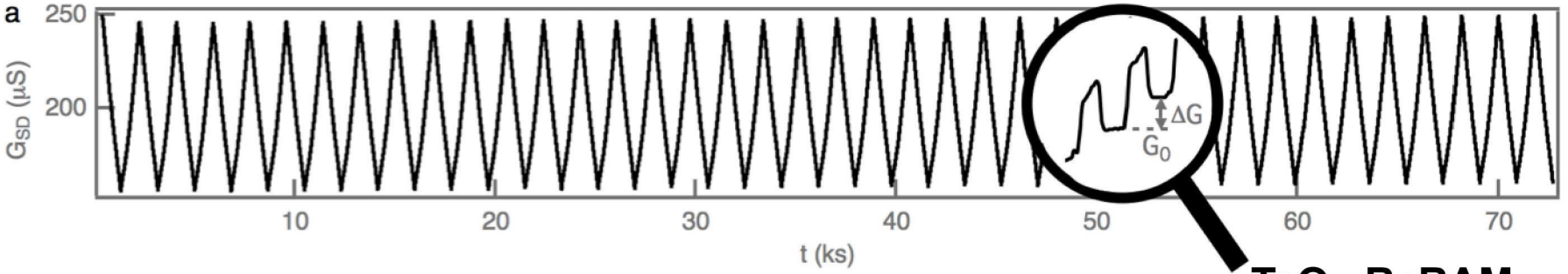




OxRAM does not meet analog training accelerator requirements

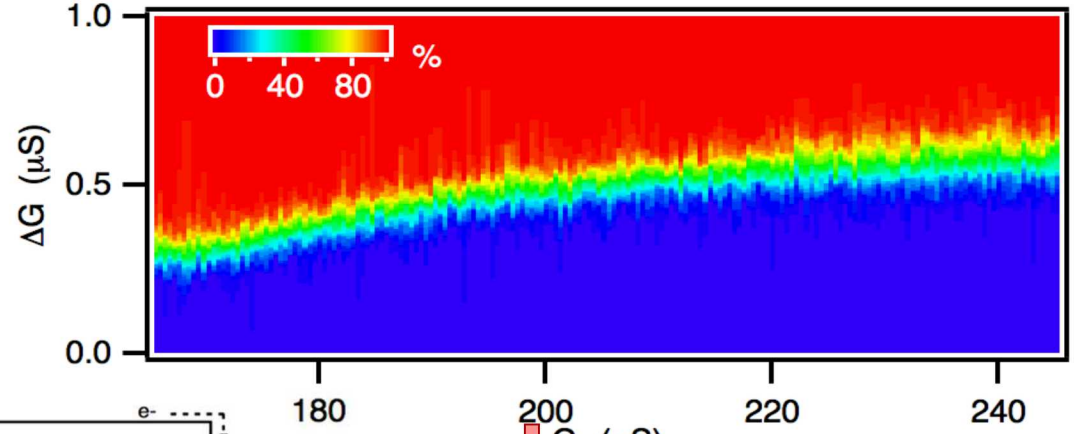
Ionic Transistor

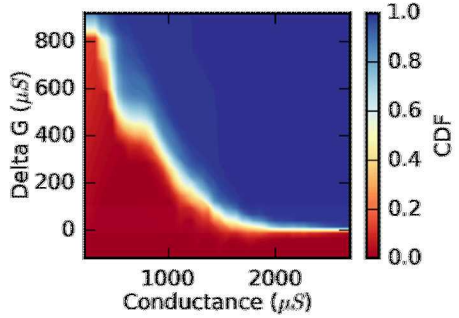
Sandia National Laboratories

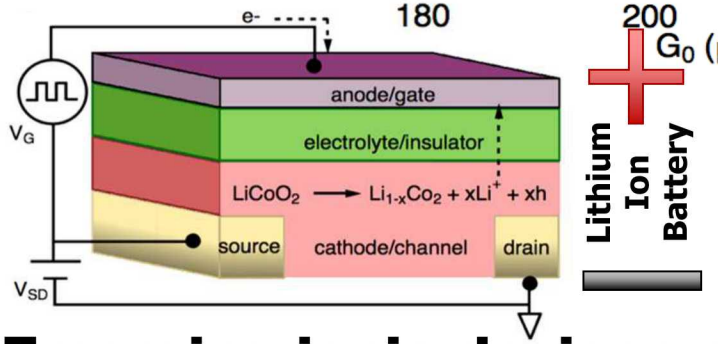


TaOx ReRAM

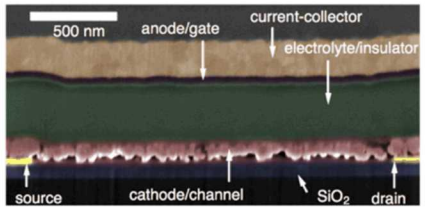
LISTA > 200 states



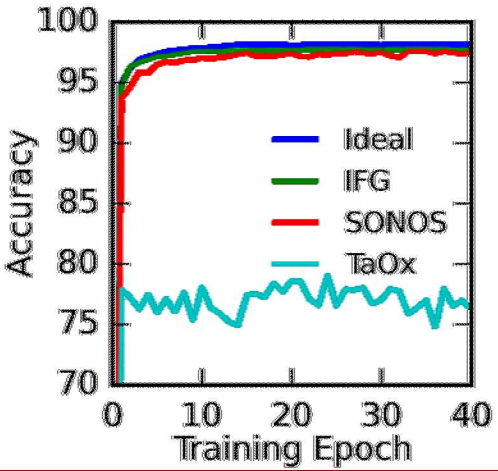




Lithium Ion Battery



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



Emerging ionic devices show promise for anal

Key Messages

- Energy efficiency critical at all scales, especially at the edge
- Digital CMOS nearing its limits
- Analog crossbar training operations 10-1000x more efficient than digital CMOS limits
  - This can enable training at the edge
- Training at the edge is desirable
  - Communication bandwidth is limited: cannot send all data to the cloud
  - Privacy: may not want to send all data to the cloud
- Device affects algorithm accuracy → challenge for analog
  - Ideal analog devices a topic of continued research
  - Compatible with algs which tolerate moderate precision

Backup Slides

Acknowledgements

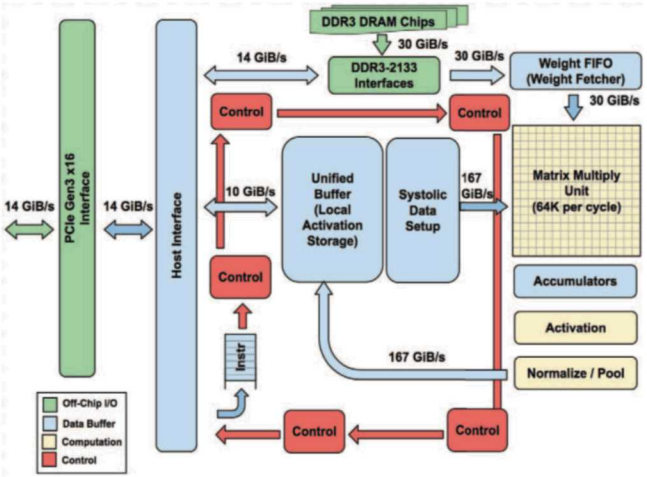
- Funding
  - Sandia National Laboratories Laboratory Directed Research and Development
  - Department of Energy Advanced Manufacturing Office (AMO)
  - Department of the Defense, Defense Threat Reduction Agency under grant no. HDTRA1-11-21-BRCWMD-BAA
- Many collaborators:
  - B. Draper, K. Knisely,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

Session Charge

- AI/ML has provided an explosion in new applications and features, but at a significant energy cost. Processing at the “edge” which is energy constrained is a unique challenge. What role will “analog” play in the AI/ML processing at the edge?
- What new devices are required for analog in machine learning?
  - Requirements in: precision, retention, linearity, speed etc.?
  - For example, could it be that, in future, you could have larger AI machines artificially produce data for training smaller AI machines?
- How scalable is Analog in Machine Learning?
  - Only for smaller “edge” applications?
  - Only for inference?
  - Learning on the fly?
- What should be the goals for Analog in Machine Learning for the next 3, 6 and 10 years?

State of the Art Digital

- TPU v1 Performance/Watt
  - Die level performance of 2.3 TeraOps/W
  - → **~ 1 pJ per 8 bit operation!**
- Intel Haswell die comparison:
  - 18 GigaOps/W
  - → 55 pJ per 8-bit operation
- Similar for mobile processors



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

N. P. Jouppi *et al, 2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*, Toronto, ON, 2017

AI Across Power Envelopes

Sandia National Laboratories

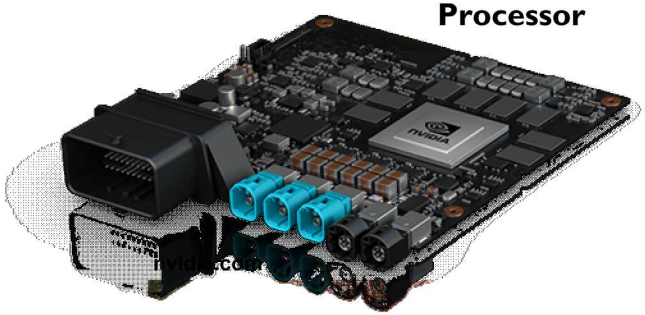
IoT, Edge, and Mobile Computing

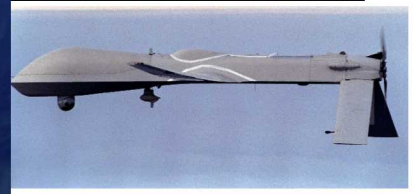
Self Driving Cars, Unmanned Arial Vehicles, and Satellite Computing

Datacenters, HPC

nest.com

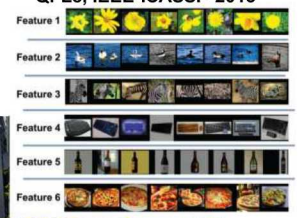
Nvidia Self-Driving Processor





wikimedia.org

Q. Le, IEEE ICASSP 2013



ASCI Red Supercomputer

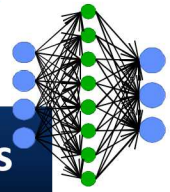
1W 10W 10^2W 10^3W 10^4W 10^5W 10^6W

	SONOS	RRAM	PCM	MRAM	FeRAM	Ionic
On/Off ratio	10^6	10-10^4	10^2-10^4	1.5-2	10^2-10^3	10^3
Retention	High	High	High	High	High	Medium
Drift	Weak	Weak	Strong	None	None	None
Endurance	10^6	10^8	10^6	>10^{12}	10^{10}	10^9
Switching Energy (Nudge)	100 fJ	100 fJ	10pJ	100fJ	100fJ	<100fJ
Switching Speed	1 µs	<10ns	10-100ns	<10ns	30ns	<10ns
Analog Tunability	High	Moderate	Low	Low	Low	High
DNN Inference Suitability	High	High	Low	Low	Low	High
DNN Training Suitability	Moderate	Moderate	Low	Low	Low	Moderate to High

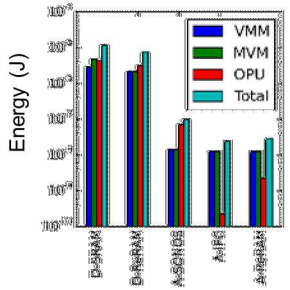
Device to Algorithm Model

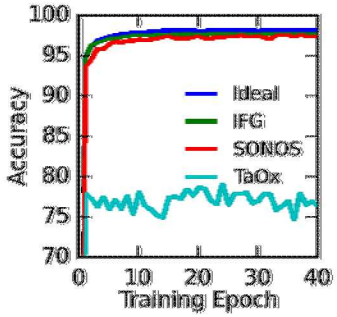
Sandia National Laboratories

Target Algorithms
- Deep Neural Nets
- Convolutional NN

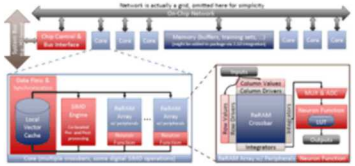


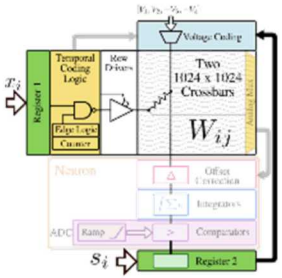
Algorithms

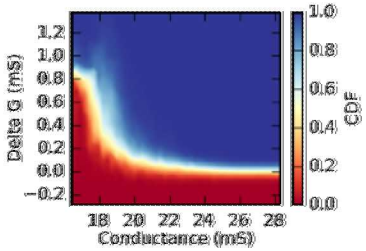




Energy/Performance Model
Model performance and energy requirements





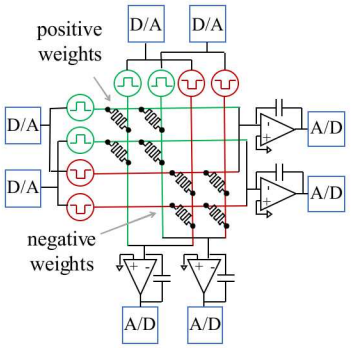


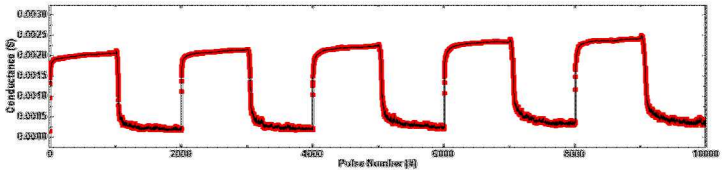
Architecture

MIHROSS SIM

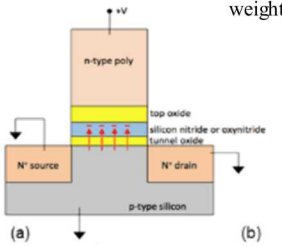
Translates device measurements and crossbar circuits to algorithm-level performance

Circuits

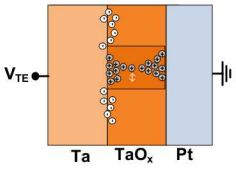


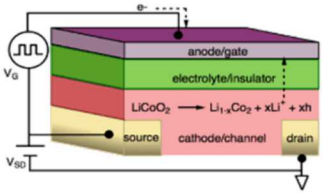


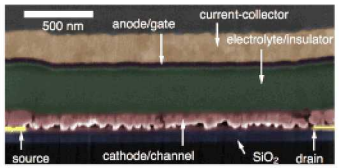
Analog characterization



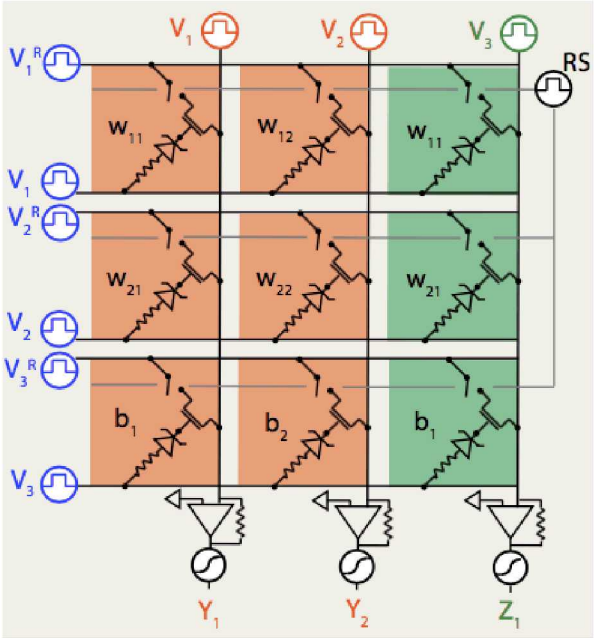
Devices

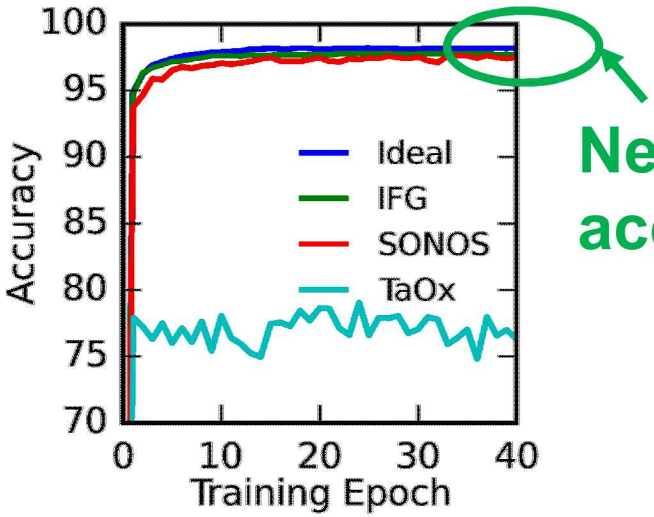




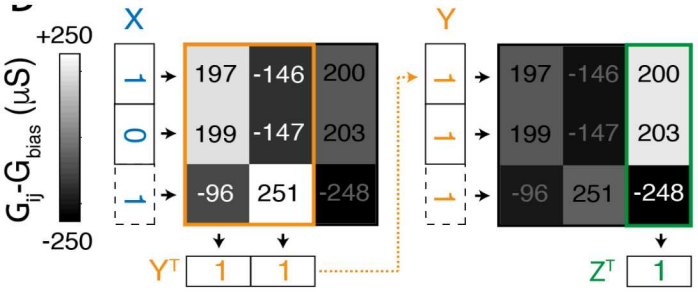


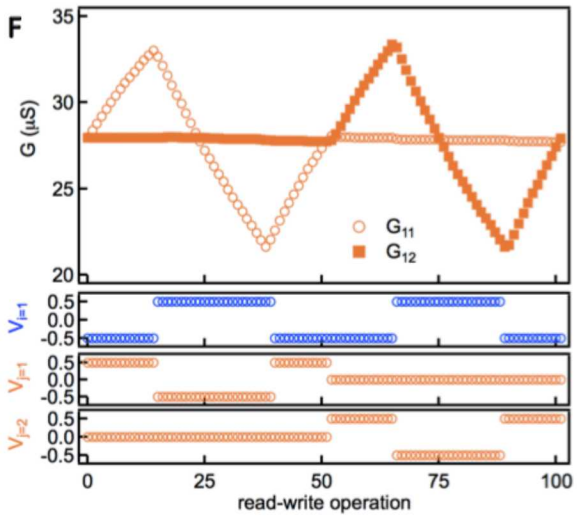
IFG Array Demonstration





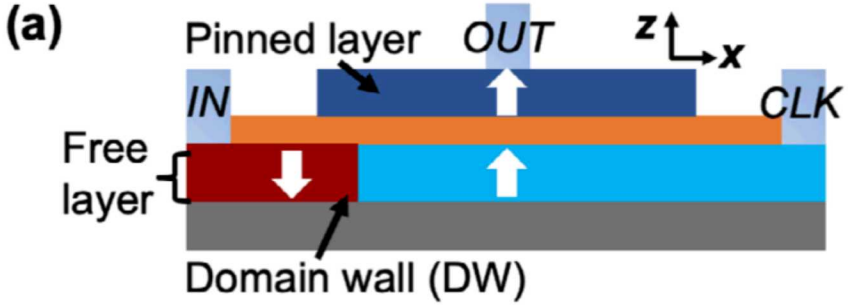
Near ideal accuracy

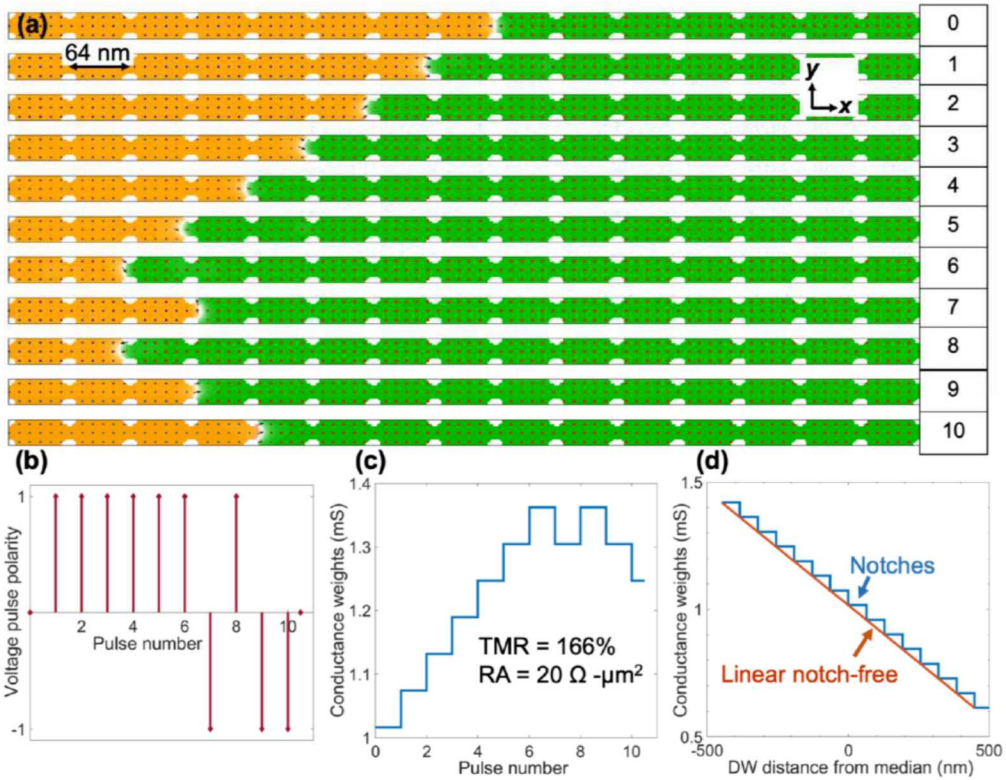




E. J. Fuller, S. T. Keene, A. Melianas, Z. Wang, S. Agarwal, Y. Li, Y. Tuchman, C. D. James, M. J. Marinella, J. J. Yang, A. Salleo, A. A. Talin, *Science* 364, 570, (2019).

Magnetic Synapse Concept



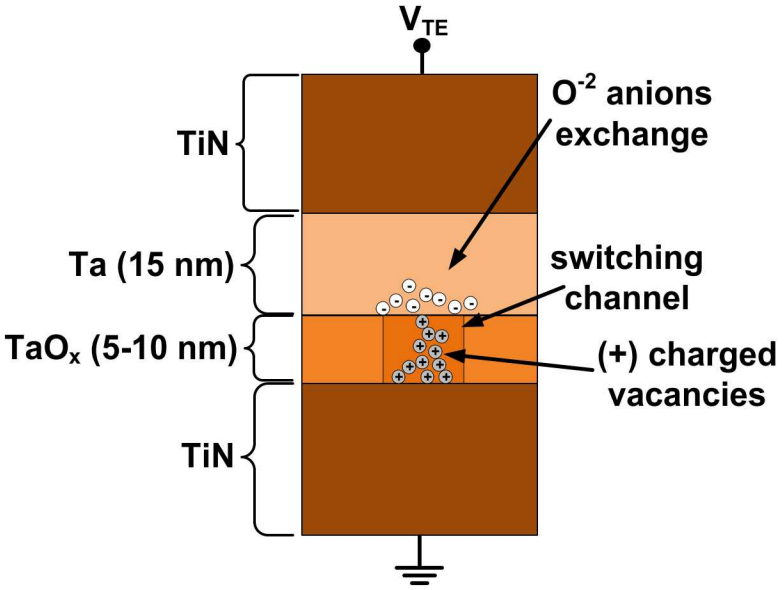


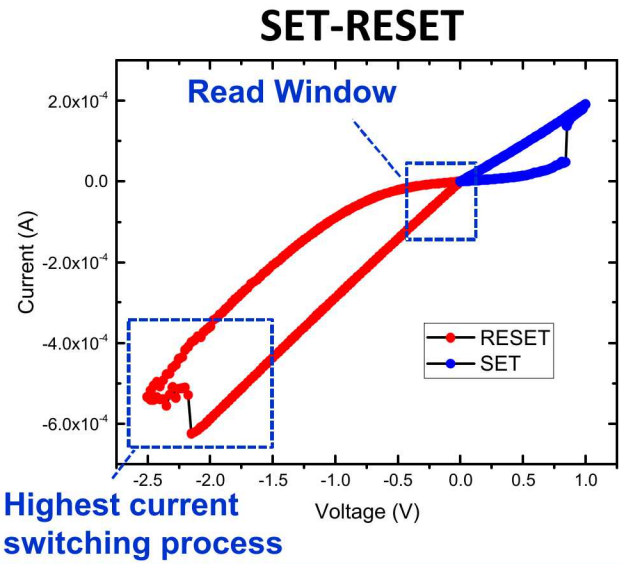
Otitoaleke Akinola, Xuan Hu, Christopher H Bennett, Matthew Marinella, Joseph S Friedman, Jean Anne C Incorvia J. Phys. D, 2019.

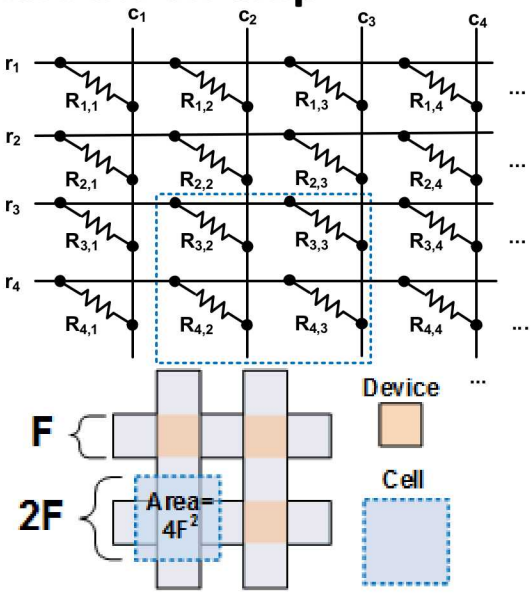
Exemplar Synapse: Oxide ReRAM

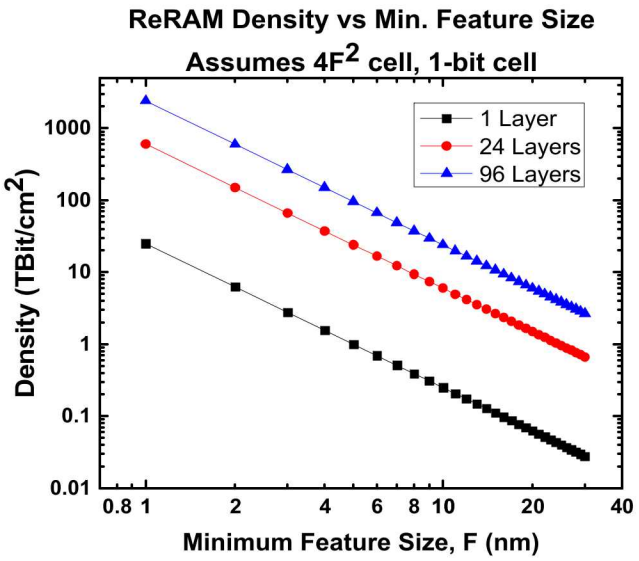
ReRAM (OxRAM)

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





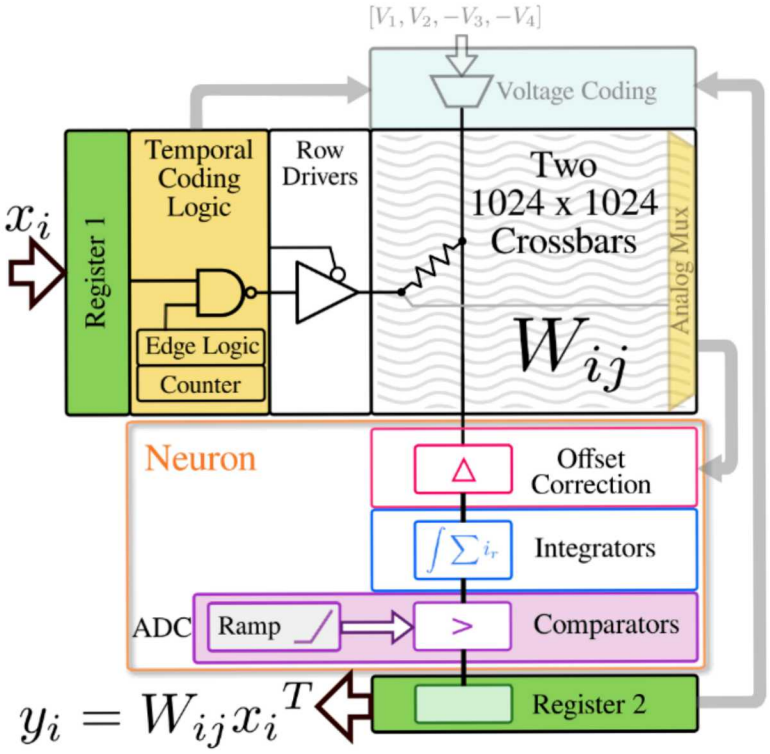




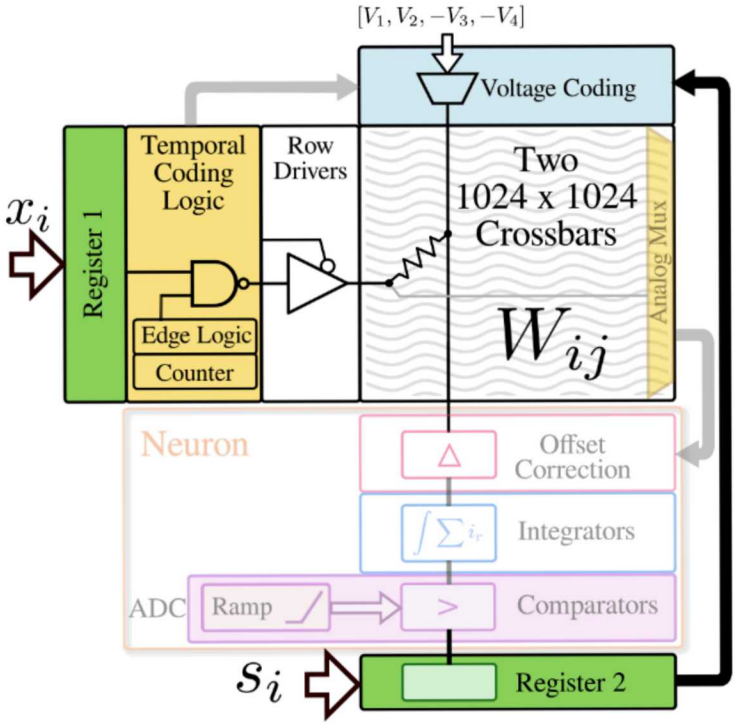
Key Circuit Block/Kernel Analysis

Sandia National Laboratories

Vector Matrix Multiply (Classification)

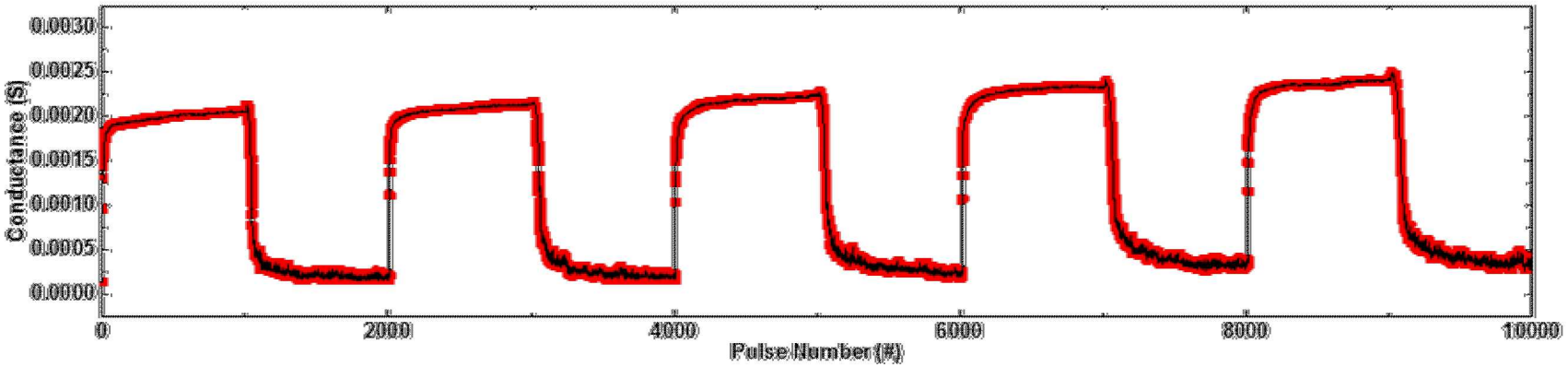


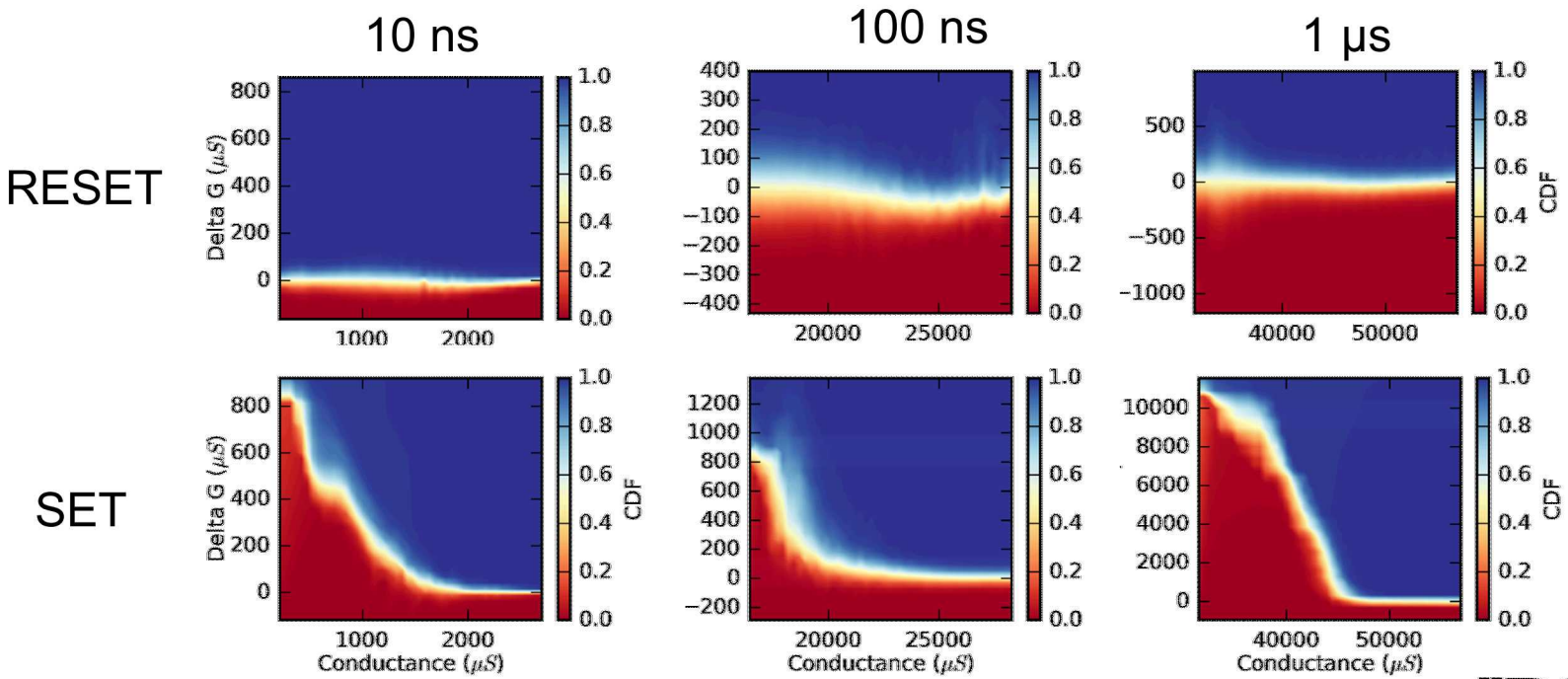
Rank-1 Update (Training)



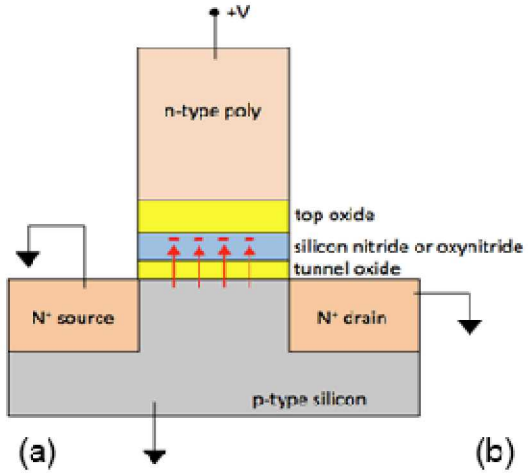
Marinella, Agarwal, et al, *IEEE JETCAS*, 2018

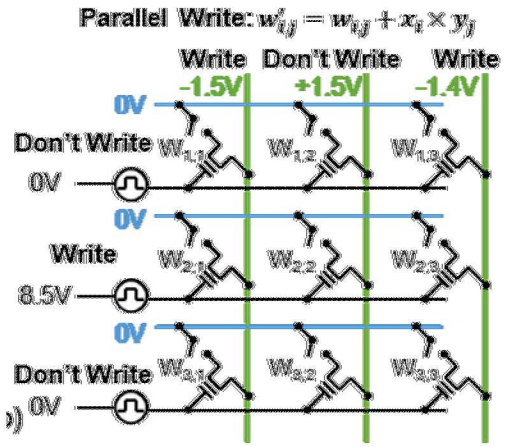
Lookup Table Generation

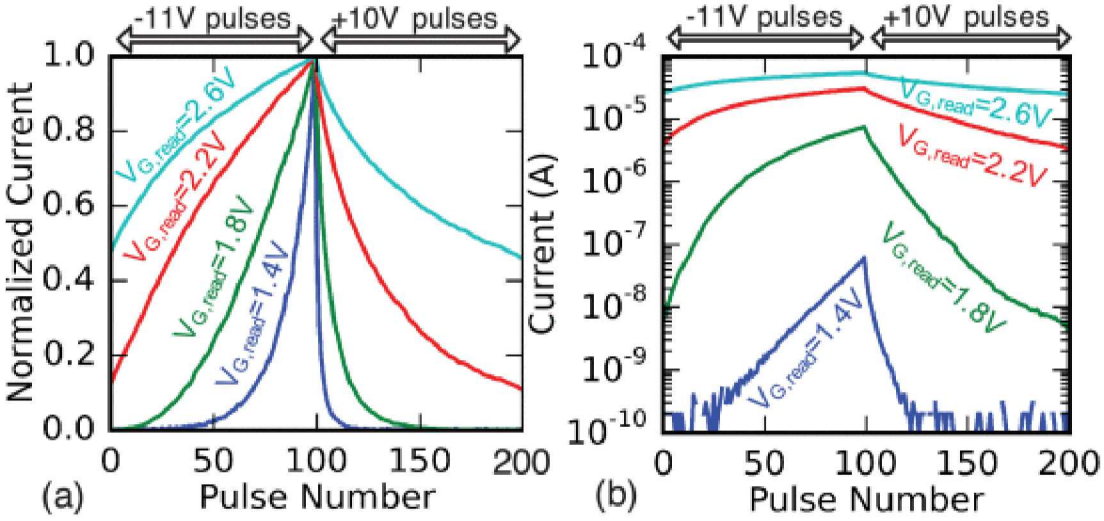




Semiconductor-Oxide-Nitride-Oxide-Semiconductor (SONOS)

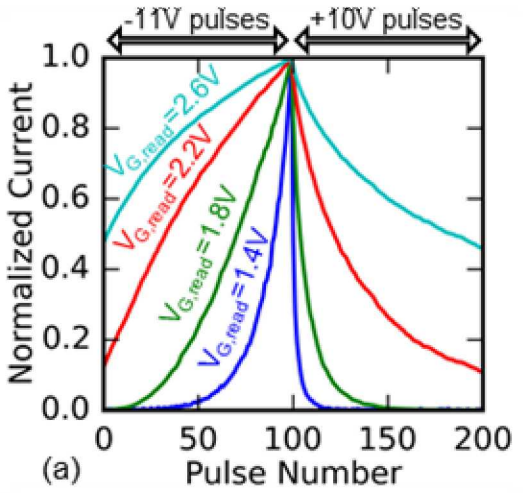


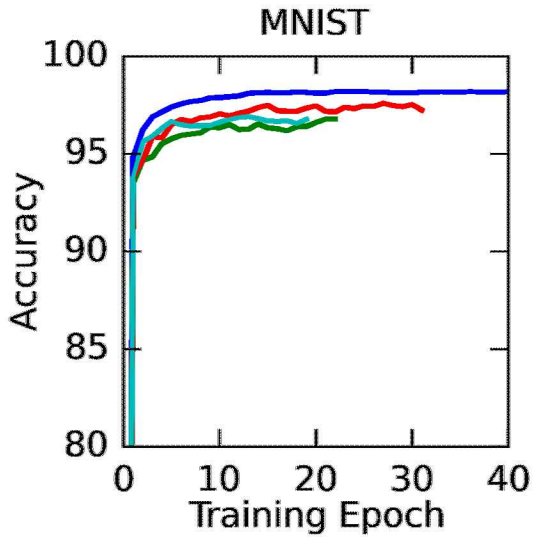




S Agarwal et al, IEEE J Exploratory Solid-State Computational Devices and Circuits, 5, 52-57, 2019.

SONOS Accuracy and Energy





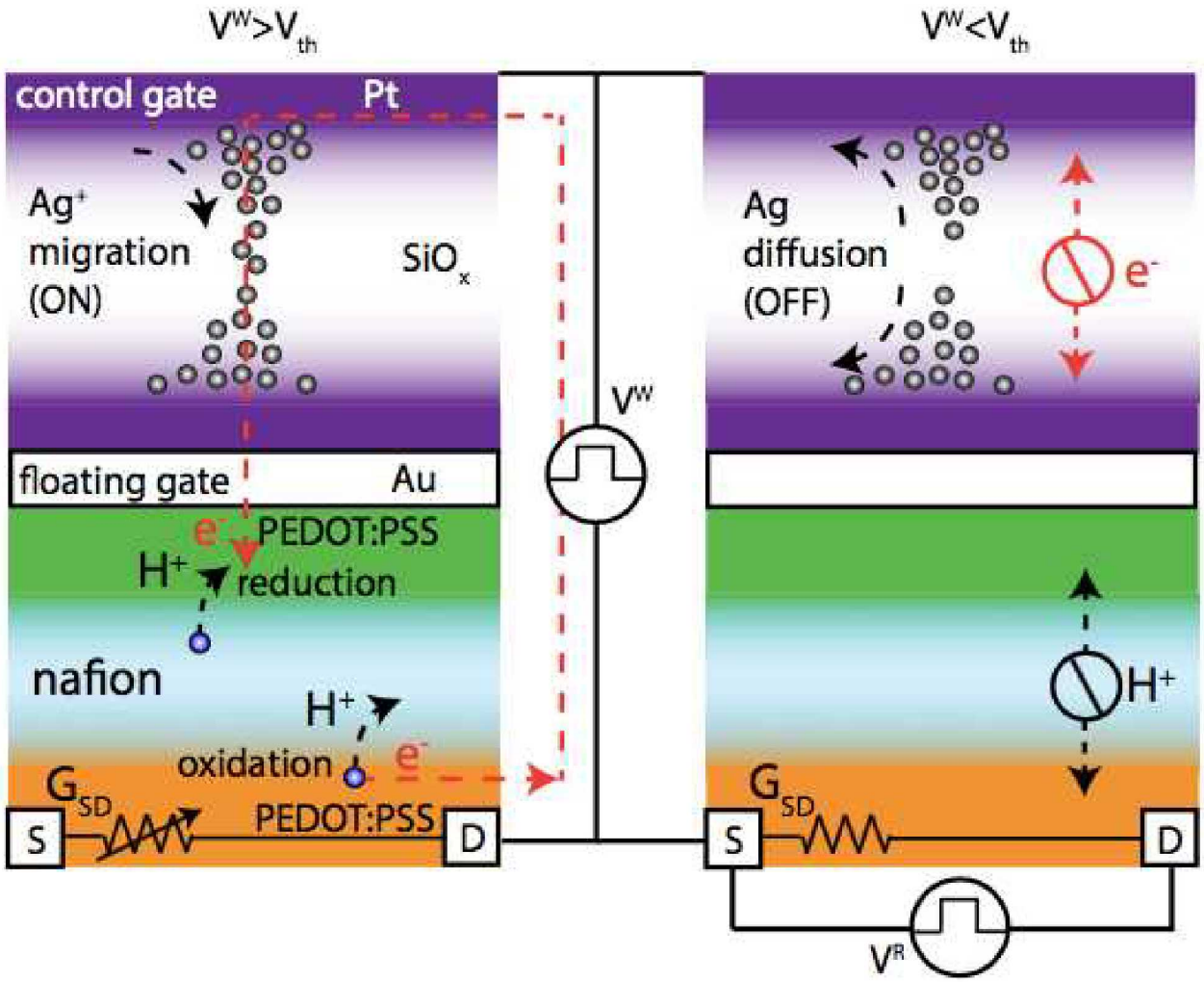
Component	Vector Matrix Multiply	Matrix Vector Multiply	Outer Product Update
Energy/Op ReRAM (fJ)	12.2	12.2	2.1
Energy/Op SONOS (fJ)	13.7	13.7	68.2
Energy/Op SRAM (fJ)	2718	4630	4102
Array Latency ReRAM (µs)	0.38	0.38	0.51
Array Latency SONOS (µs)	0.40	0.40	20
Array Latency SRAM (µs)	4	32	8

Low inference energy

Slow write

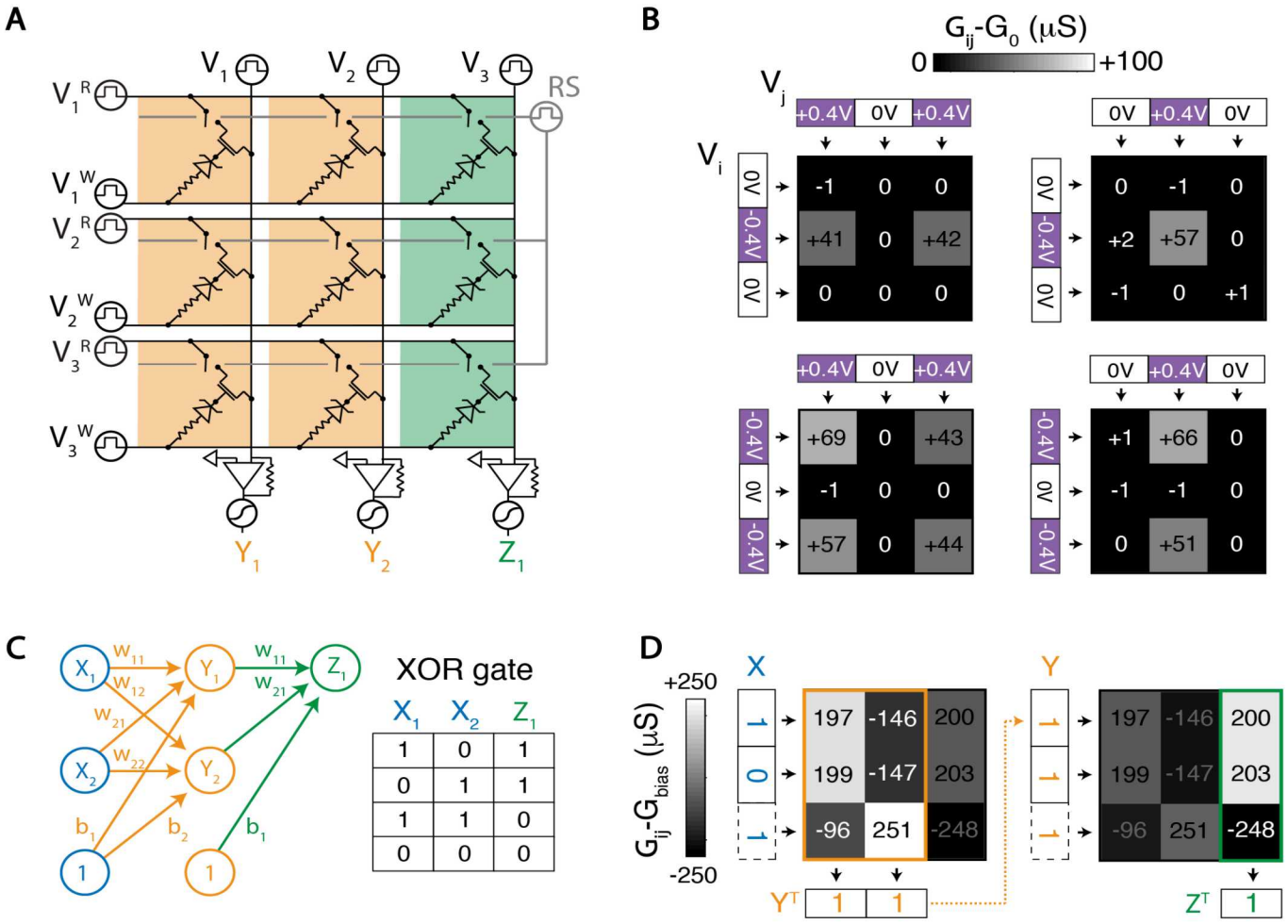
S Agarwal et al, IEEE J Exploratory Solid-State Computational Devices and Circuits, 2019.

Ionic Floating Gate



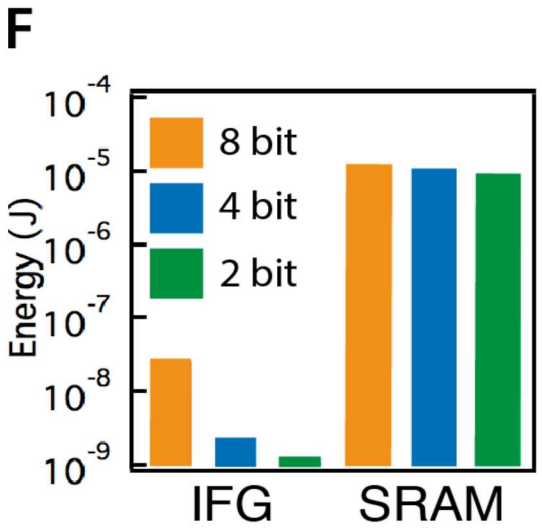
E. J. Fuller, S. T. Keene, A. Melianas, Z. Wang, S. Agarwal, Y. Li, Y. Tuchman, C. D. James, M. J. Marinella, J. J. Yang, A. Salleo, A. A. Talin, *Science* 364, 570, (2019).

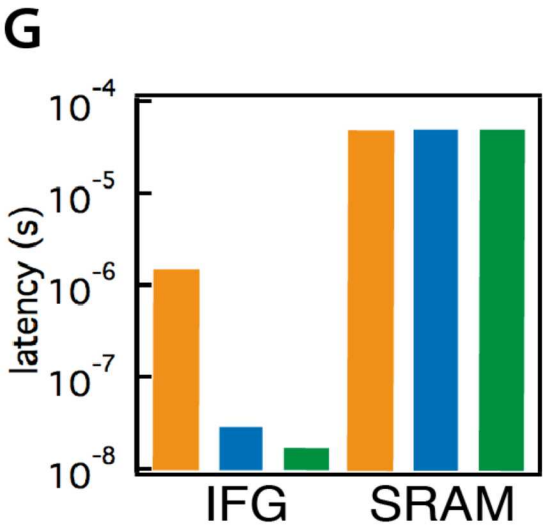
Programming Demonstration

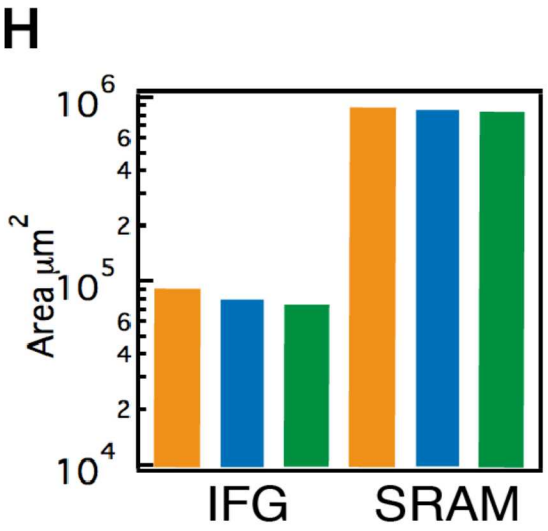


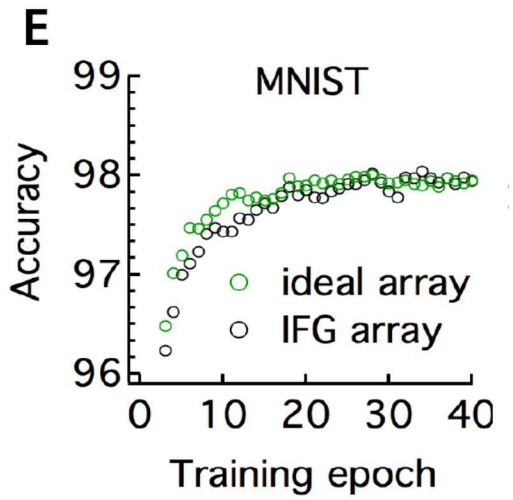
E. J. Fuller, S. T. Keene, A. Melianas, Z. Wang, S. Agarwal, Y. Li, Y. Tuchman, C. D. James, M. J. Marinella, J. J. Yang, A. Salleo, A. A. Talin, *Science* 364, 570, (2019).

Programming Demonstration



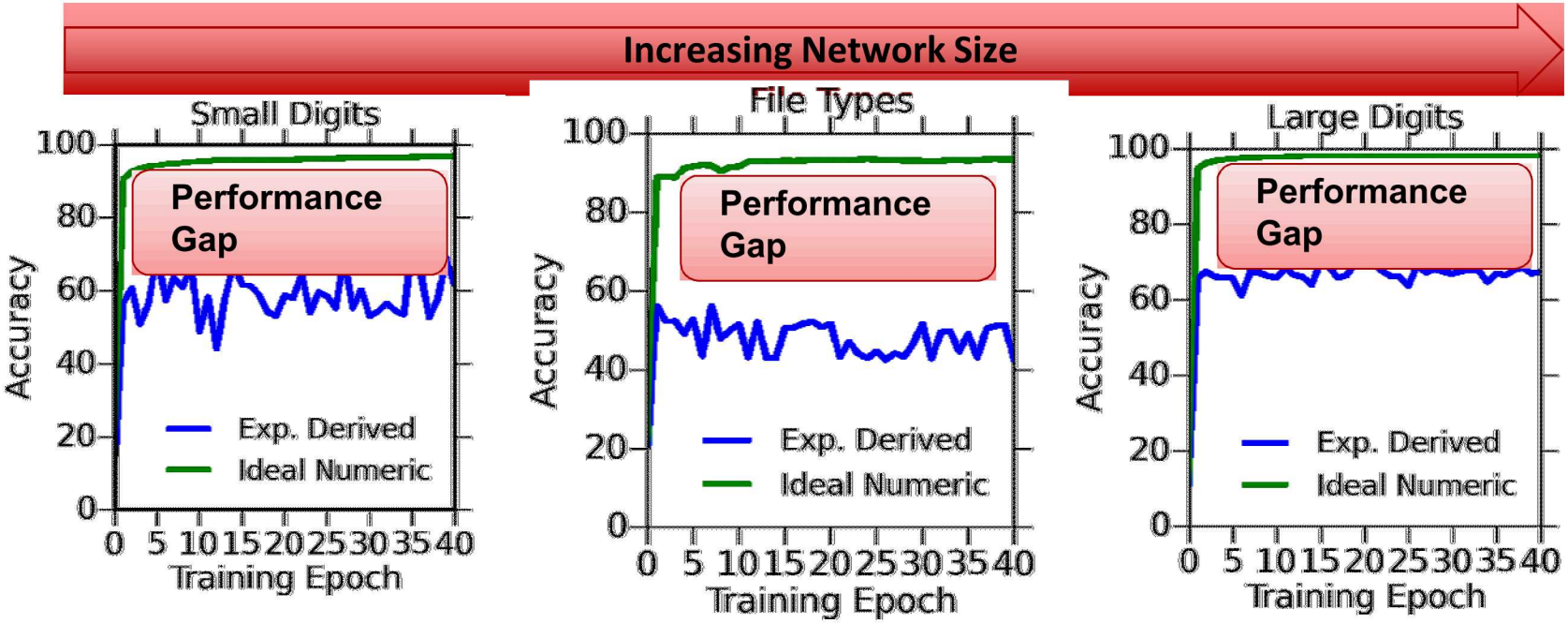






CrossSim Model of TaOx ReRAM: MNIST, Backprop Training

Sandia National Laboratories



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

CROSS SIM

Crossbar Theoretical Limits

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