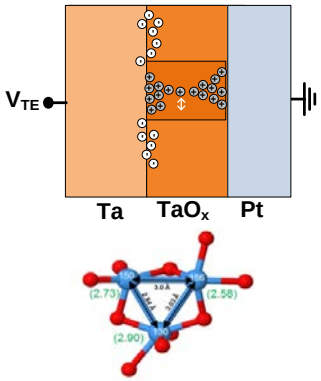
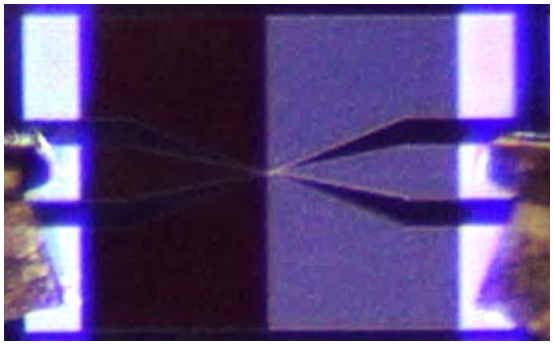
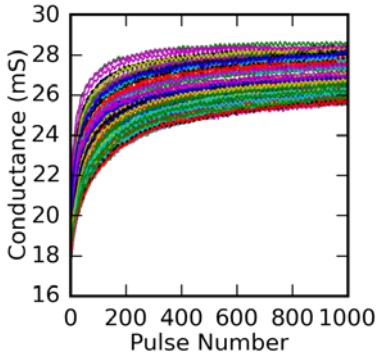
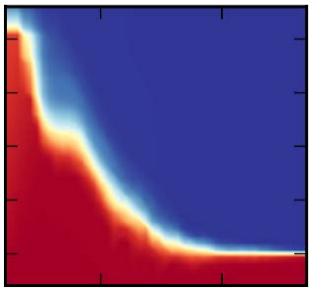
Exceptional service in the national interest

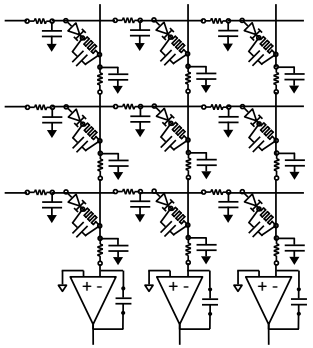
Sandia National Laboratories











Ultra-Efficient Neural Algorithm Accelerator Using Processing With Memory

Matthew J. Marinella, Sapan Agarwal, Robin Jacobs-Gedrim, Alex Hsia, David Hughart, Elliot Fuller, Steve J. Plimpton, A. Alec Talin, and Conrad D. James
Sandia National Laboratories
**matthew.marinella@sandia.gov*

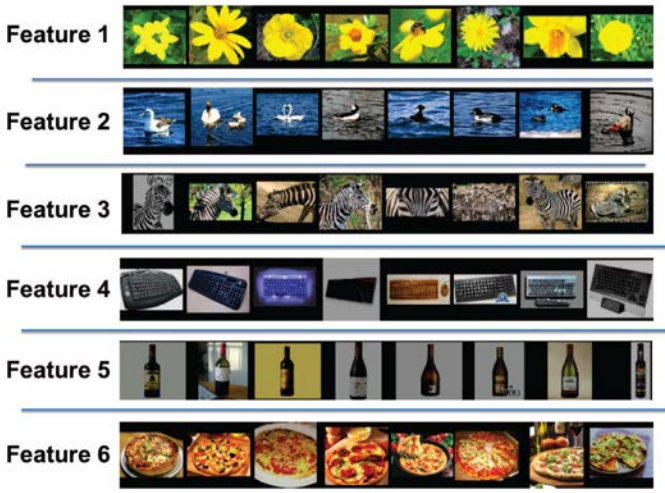
U.S. DEPARTMENT OF ENERGY | National Nuclear Security Administration | LDRD Laboratory Directed Research & Development

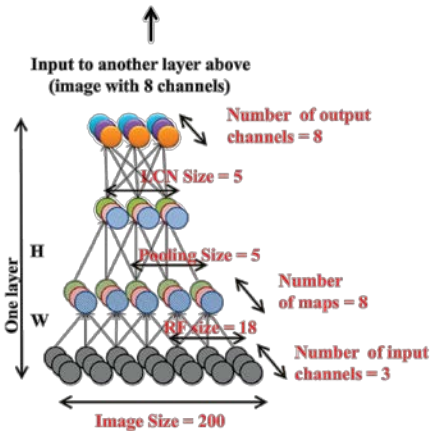
Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Why do we need more efficient computers?

Sandia National Laboratories

- **Google Deep Learning Study**
  - 16000 core, 1000 machine GPU cluster
  - Trained on 10 million 200x200 pixel images
  - Training required 3 days
  - Training dataset size: no larger than what can be trained in 1 week
- **What would they like to do?**
  - ~2 billion photos uploaded to internet per day (2014)
  - Can we train a deep net on one day of image data?
  - Assume 1000x1000 nominal image size, linear scaling (both assumptions are unrealistically optimistic)
  - *Requires 5 ZettaIPS to train in 3 days (ZettaIPS=10^{21} IPS; ~5 billion modern GPU cores)*
  - World doesn't produce enough power for this!
  - Data is increasing exponentially with time
- **Need >10^{16}-10^{18} instruction-per-second on 1 IC**
  - Less than 10 fJ per instruction energy budget





Q. Le, IEEE ICASSP 2013

Evolution of Computing Machinery

Sandia National Laboratories

1980's	100 nJ/MAC
1990's	10 nJ/MAC
2000's	1 nJ/MAC
2010-Present	100 pJ/MAC
	10 pJ/MAC
	1 pJ/MAC
2025?	100 fJ/MAC
	10 fJ/MAC
2035?	1 fJ/MAC
	100 aJ/MAC
	10 aJ/MAC
	1 aJ/MAC

Moore's Law Era: Density scaling and Dennard Power Density Scaling

Modern Computer

Single GPU Card, Late 2016

Exascale System Goal

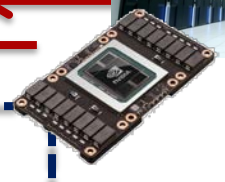
Heterogeneous Integration Era: Close integration of emerging memory, low voltage logic, photonics

Neuromorphic Target

Efficiency only possible with new computing paradigm

Biological Neurons*

1946 - ENIAC >100 J/MAC



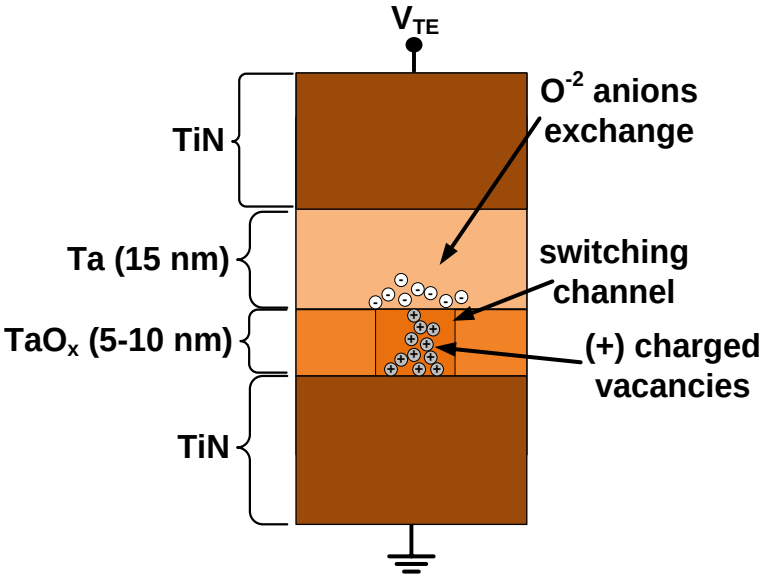
"Let physics do the computation"
Our brain is the ultimate example of this paradigm

***Caveat: Biological neurons probably do not perform MACs**

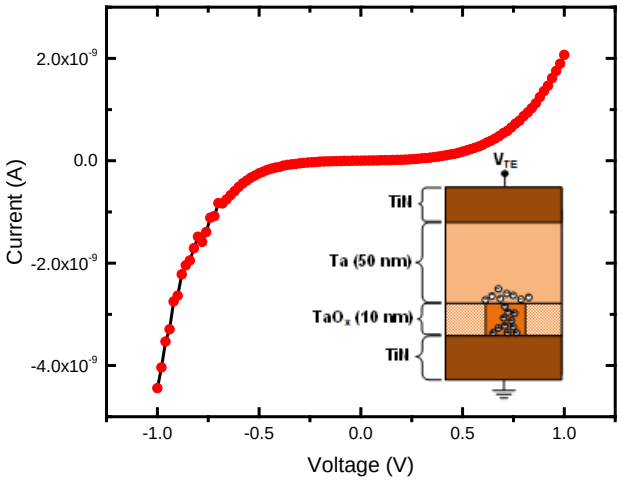
Metal Oxide Resistive RAM (ReRAM)

Sandia National Laboratories

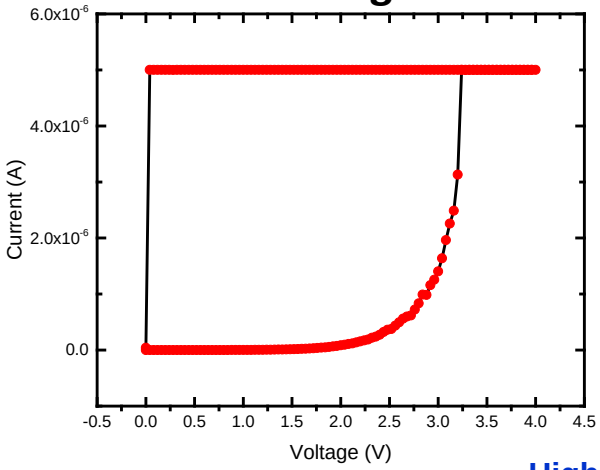
- **Example: Sandia TiN/Ta/TaOx/TiN device**
- **Starts as insulating MIM structure**
- **Forming: remove O^{2-} → soft breakdown**
- **Bipolar resistance modulation**
- **Excellent memory attributes: Switching in less than 1ns, less than 1 pJ demonstrated, scaling to 5nm, >10^{12} write cycles**



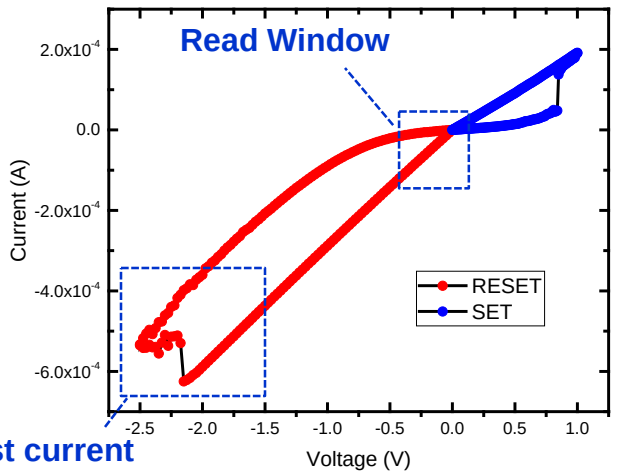
Pre-Form I/V



Forming



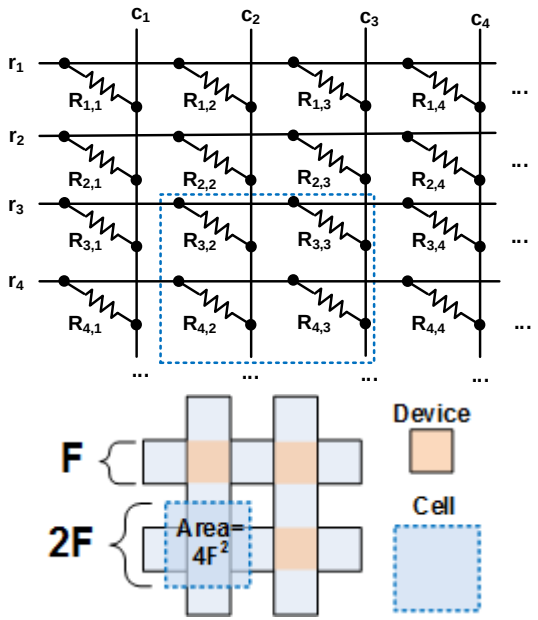
SET-RESET

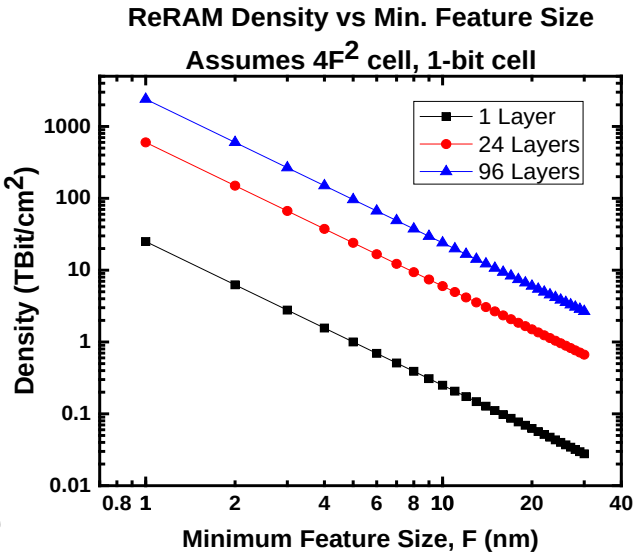


Highest current switching process

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**





How does a crossbar perform a useful computation per device?

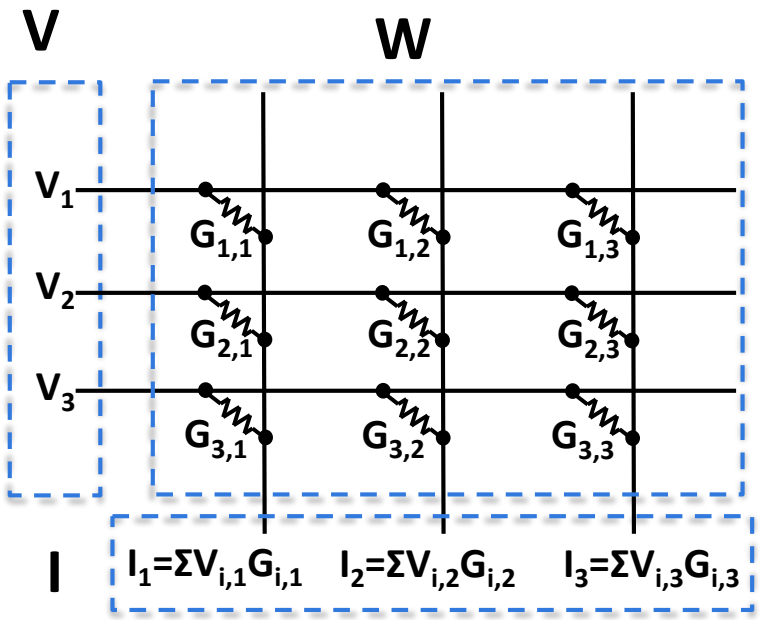
- Electronic Vector Matrix Multiply

Mathematical

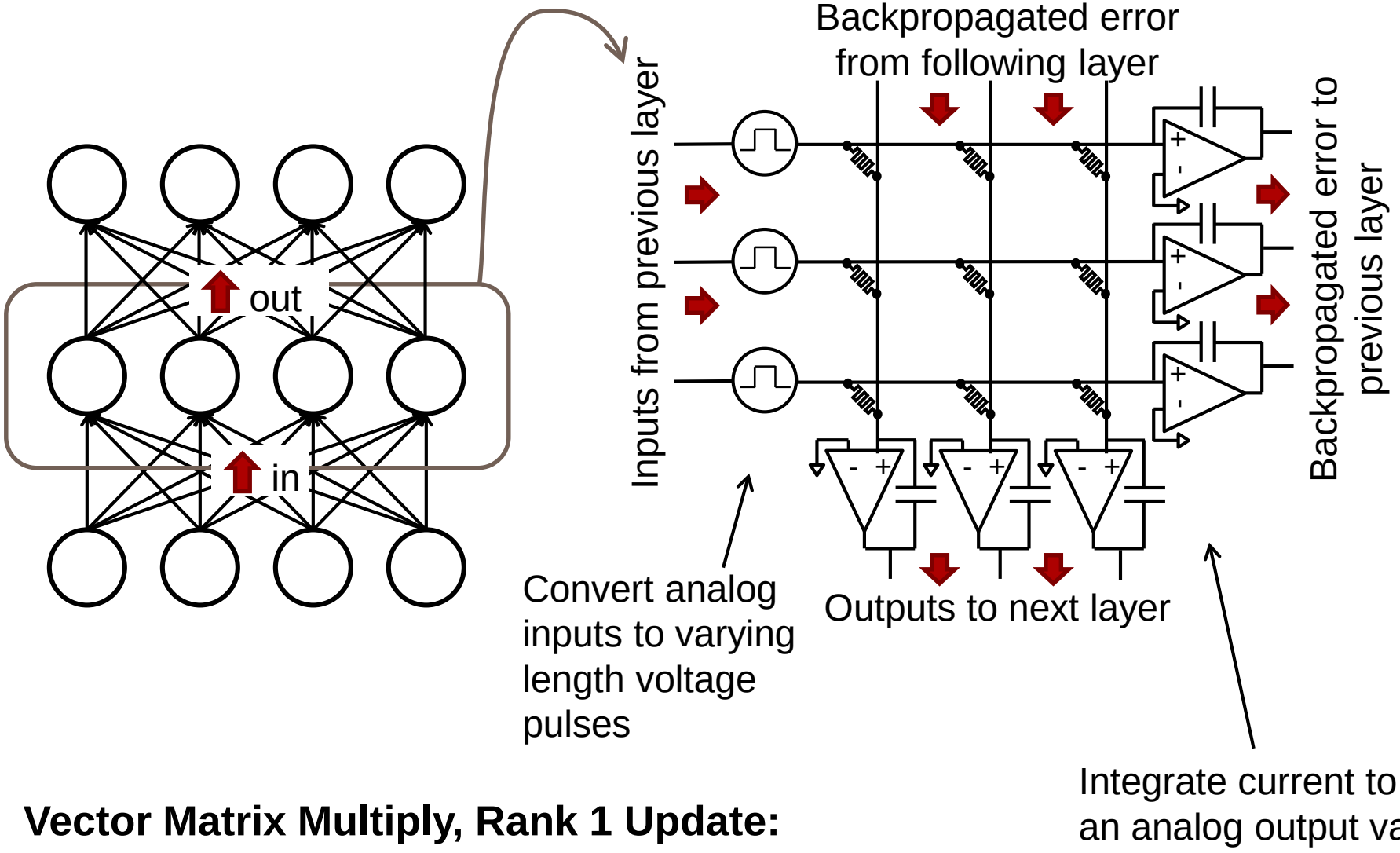
$$V^T W = I$$

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

Electrical

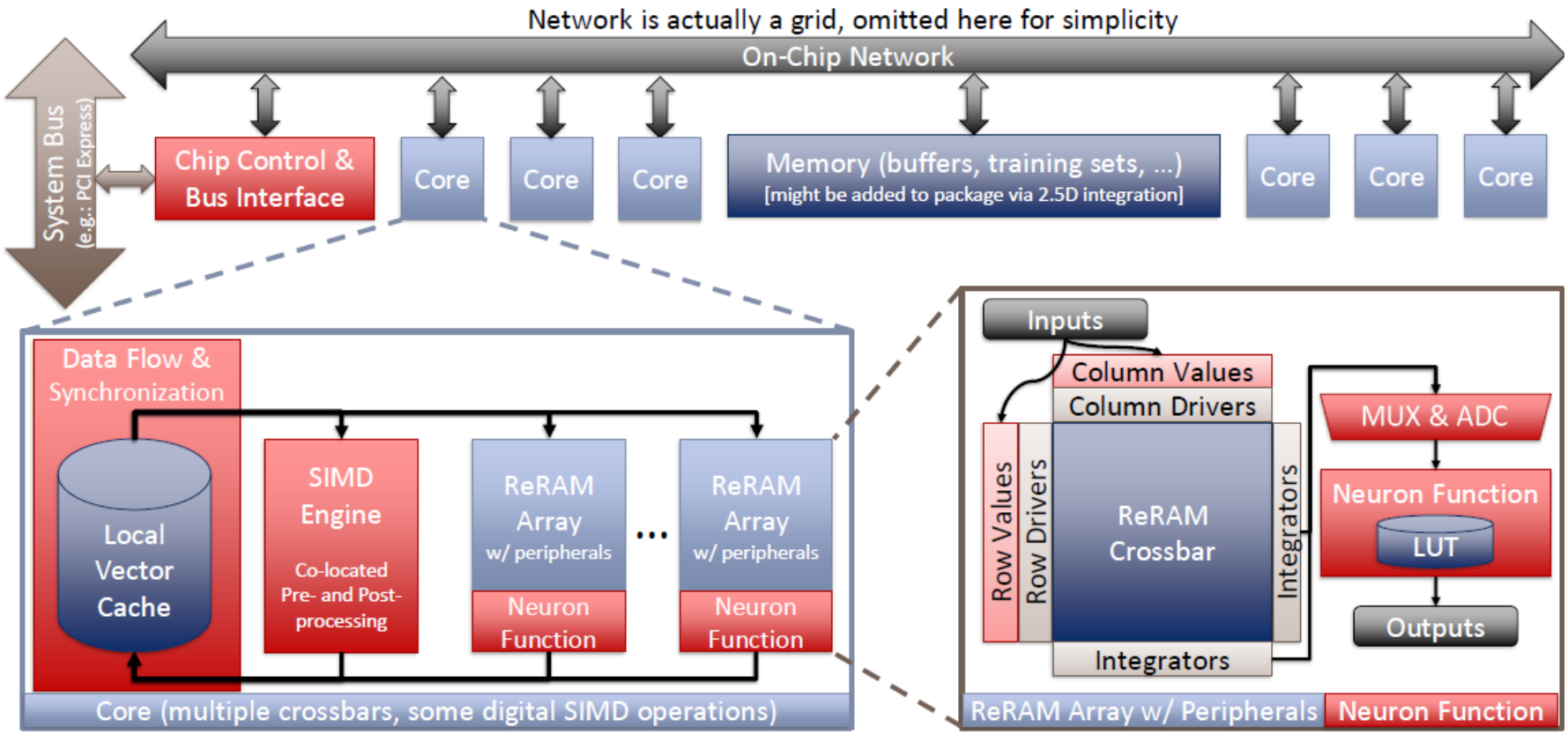


Mapping Backprop to a Crossbar

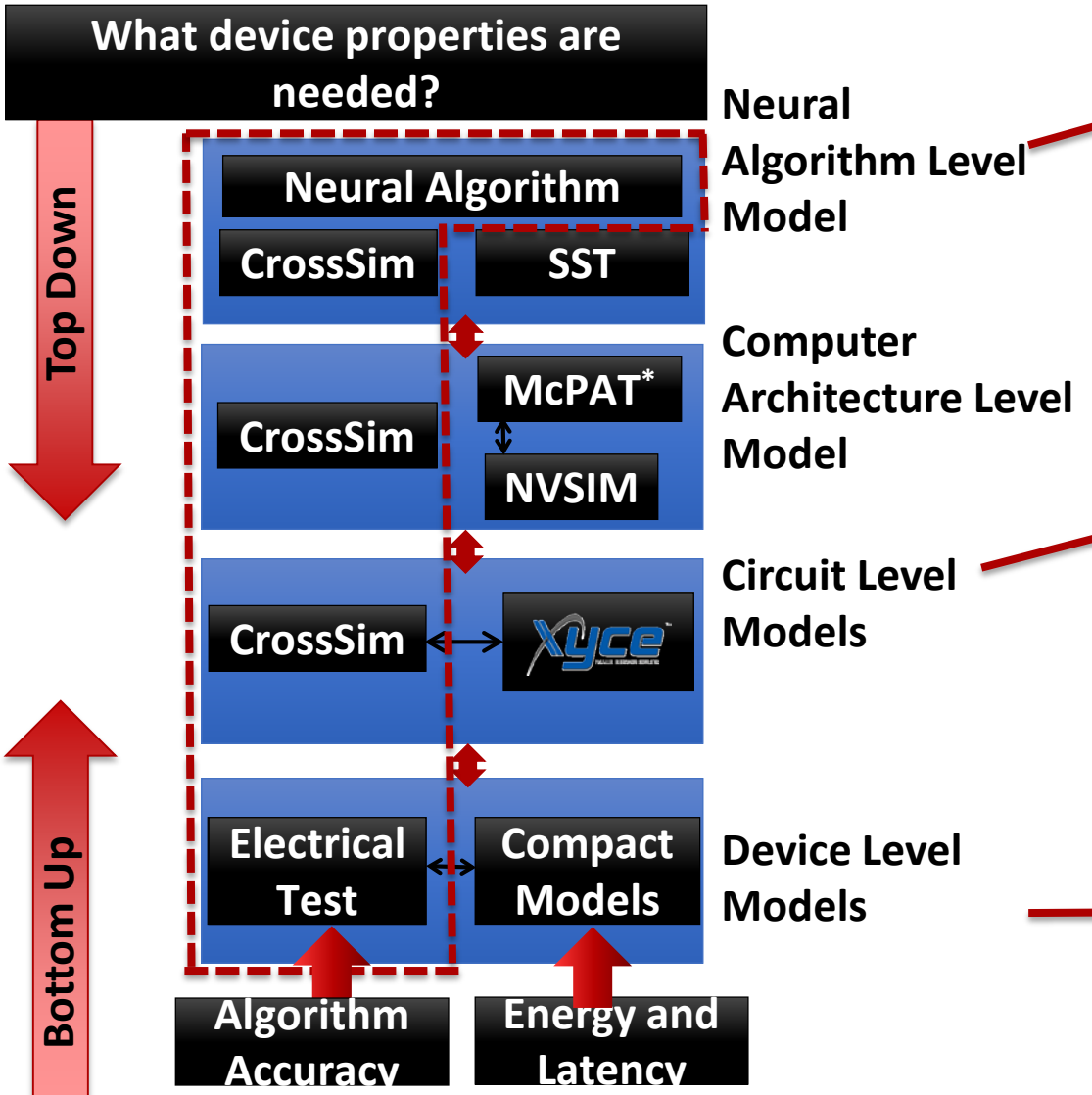


Vector Matrix Multiply, Rank 1 Update:
Key kernel used in many algorithms

Accelerator Architecture



Device to Algorithm Model

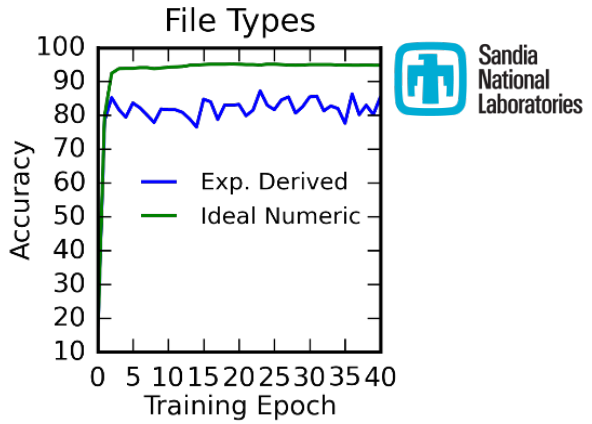


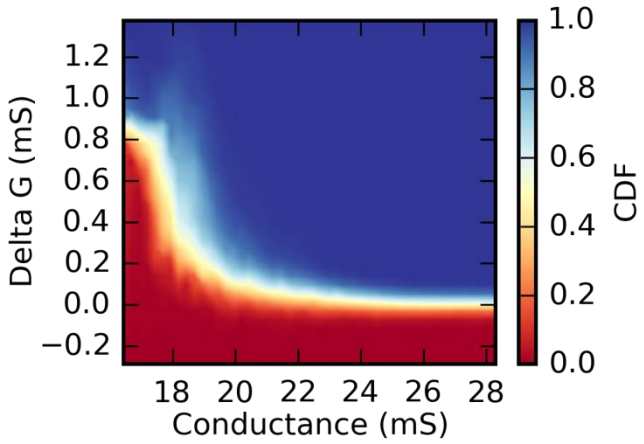
Neural Algorithm Level Model

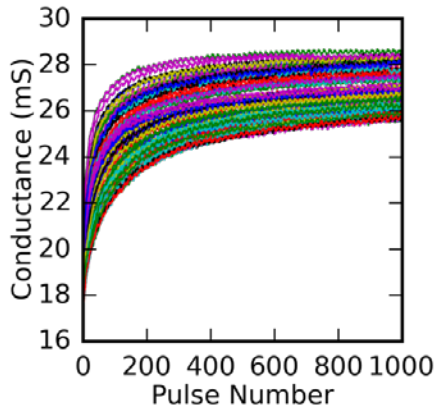
Computer Architecture Level Model

Circuit Level Models

Device Level Models



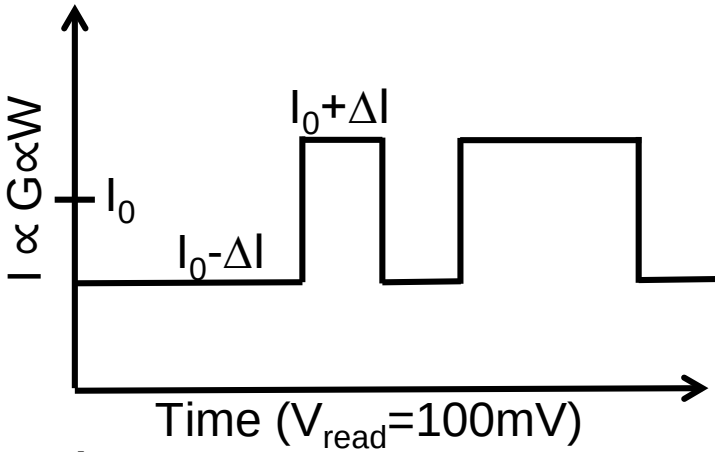




Experimental Device Nonidealities

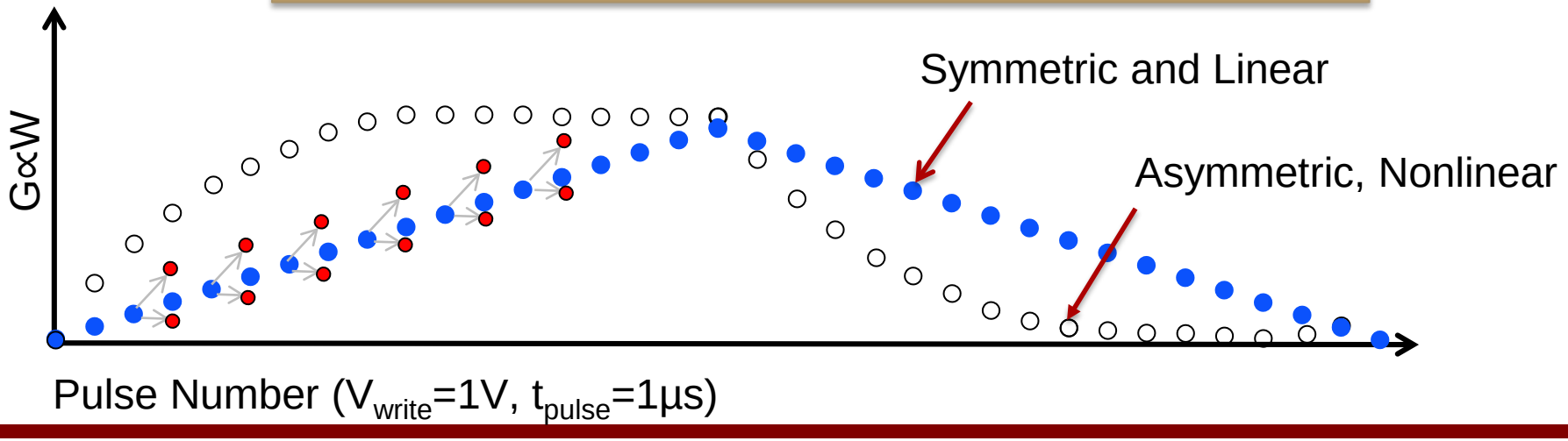
- **Ideally weight would increase and decrease linearly proportional to learning rule result**
- **Experimental devices have several nonidealities: Write Variability, Write Nonlinearity, Asymmetry, Read Noise**
- **Circuits also have A/D, D/A noise, parasitics**

Read Noise

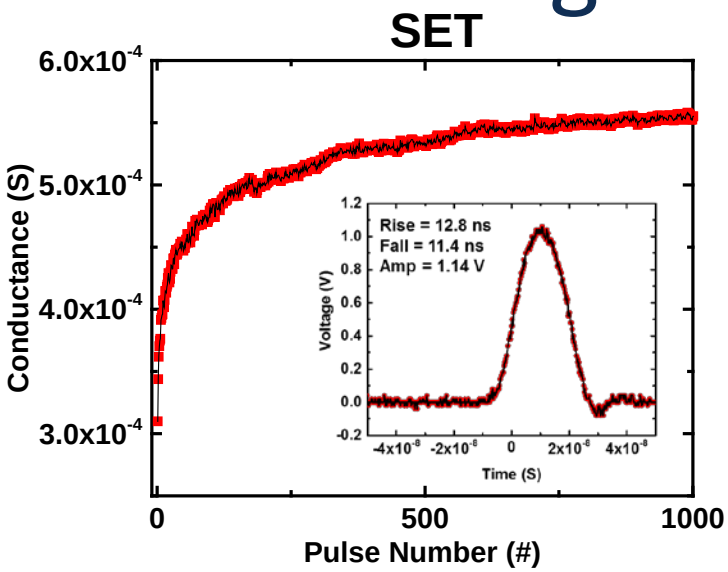


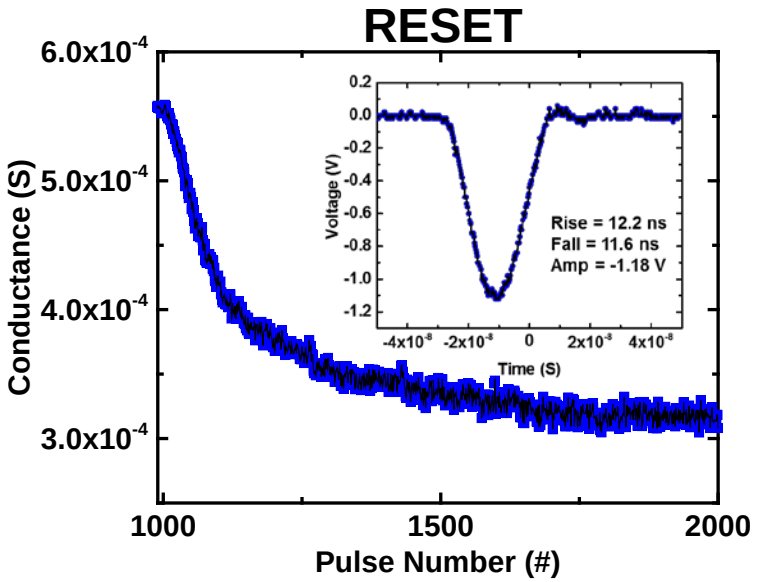
Conductance versus Pulse

● = Ideal ↗↘ = Write Variability ○ = Nonlinear

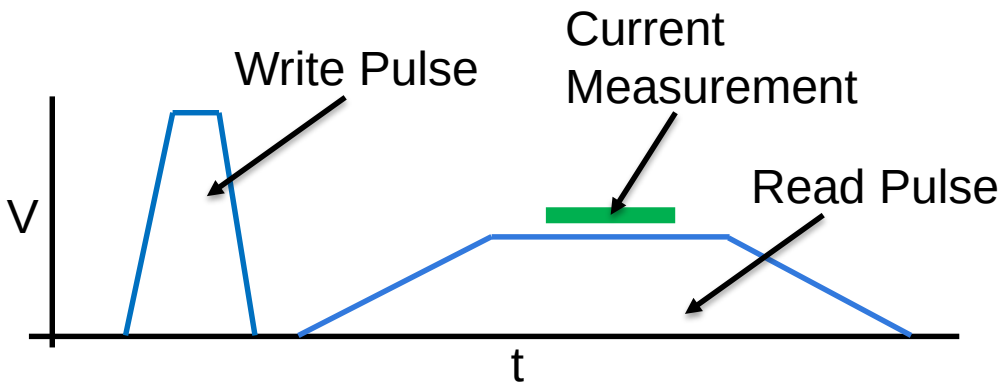


ReRAM Analog Characterization

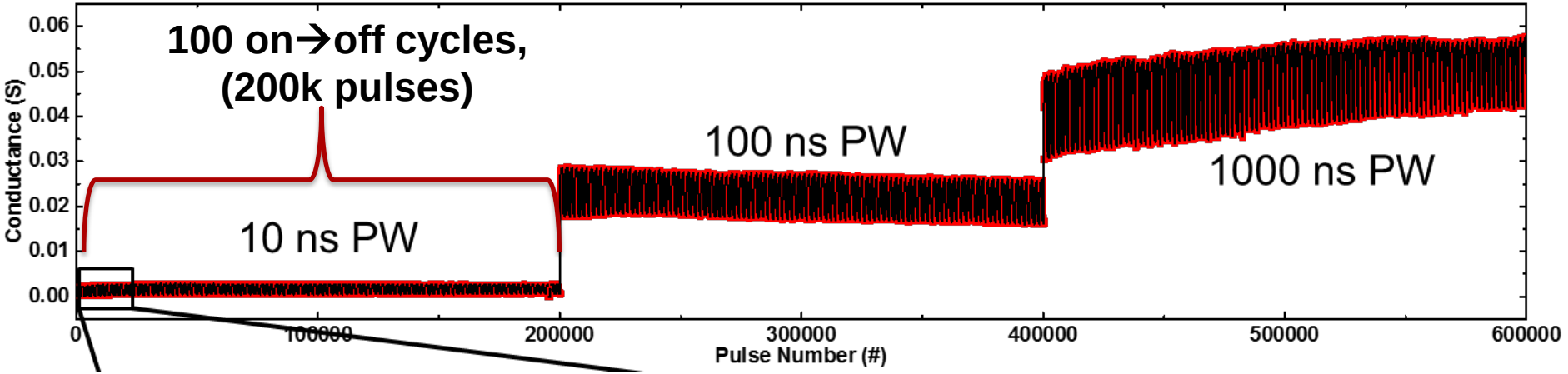


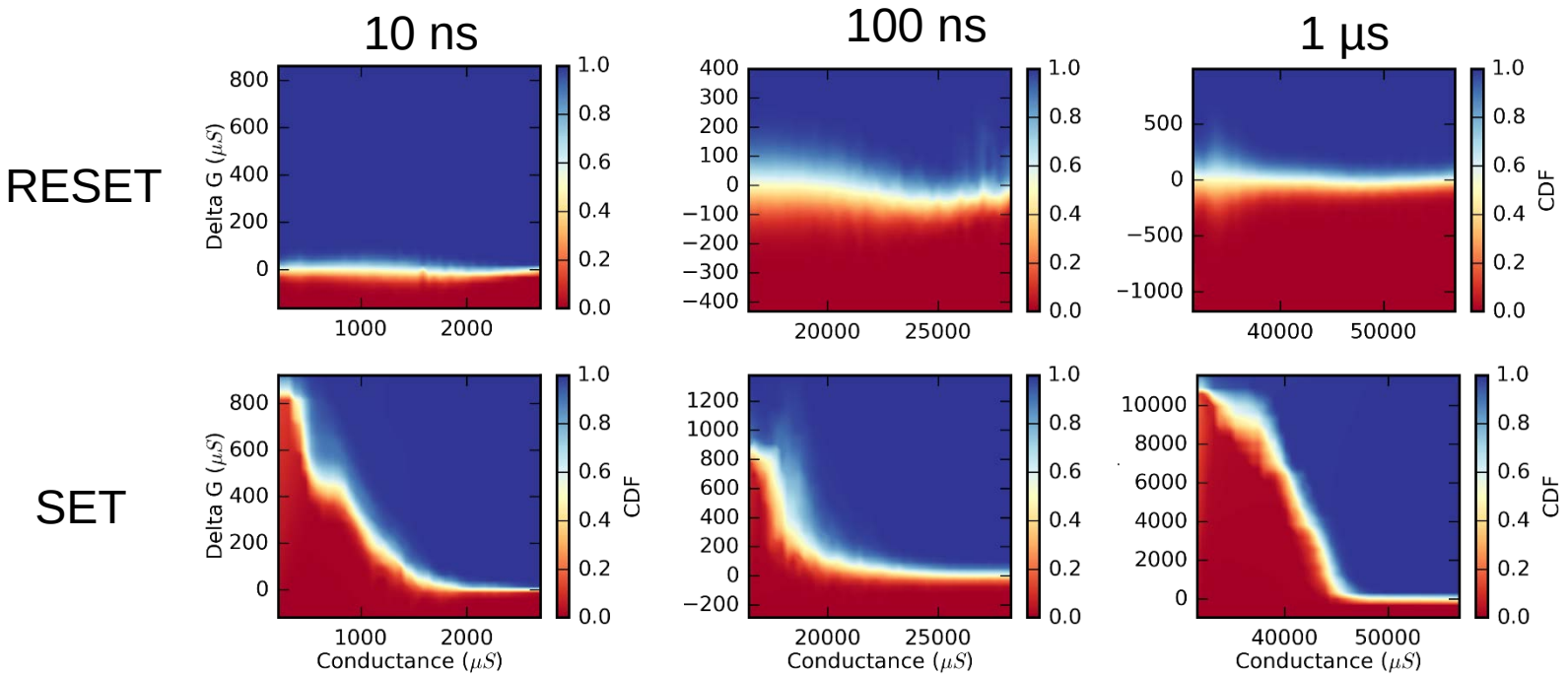


- **Use as a neuromorphic weight requires precise analog tuning**
- **Dataset requires 1000 repeated SET and RESET pulses**
- **Nominal pulse values**
  - **SET: +1V 10ns RT/PW/FT**
  - **RESET: -1V 10ns RT/PW/FT**
  - **READ: 100 mV 1 ms RT/PW/FT**



Repeated Pulsed Cycling

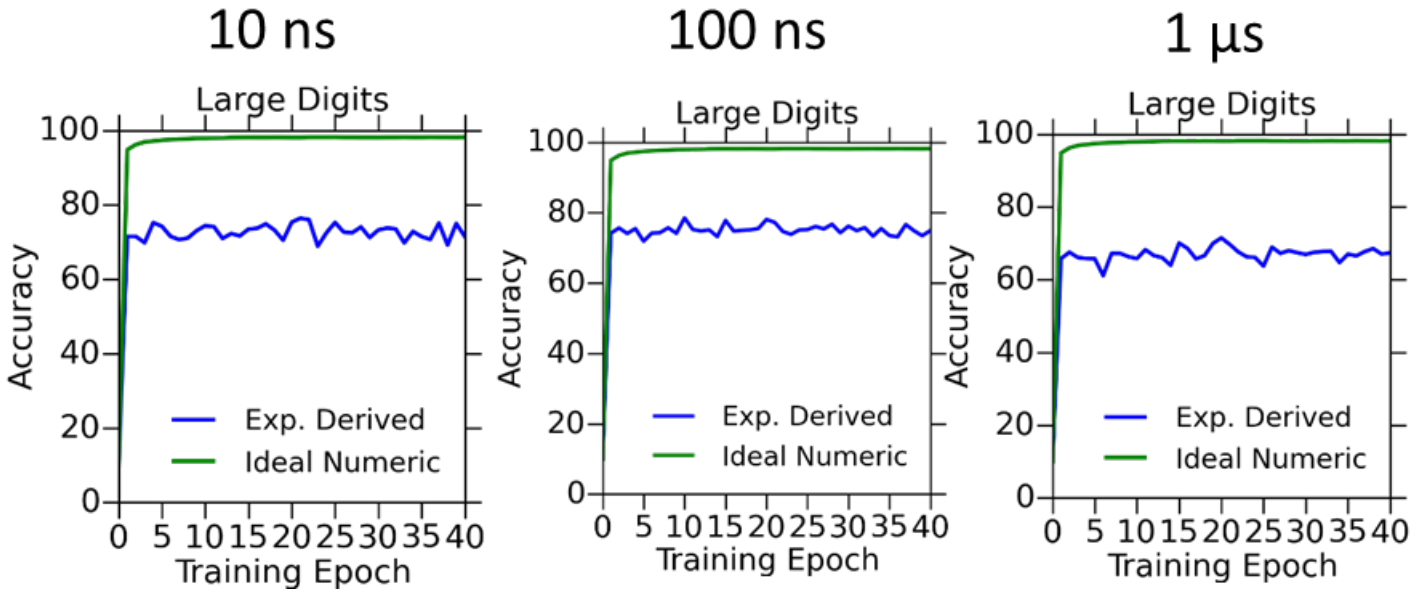




TaOx ReRAM in Backprop Training

Sandia National Laboratories

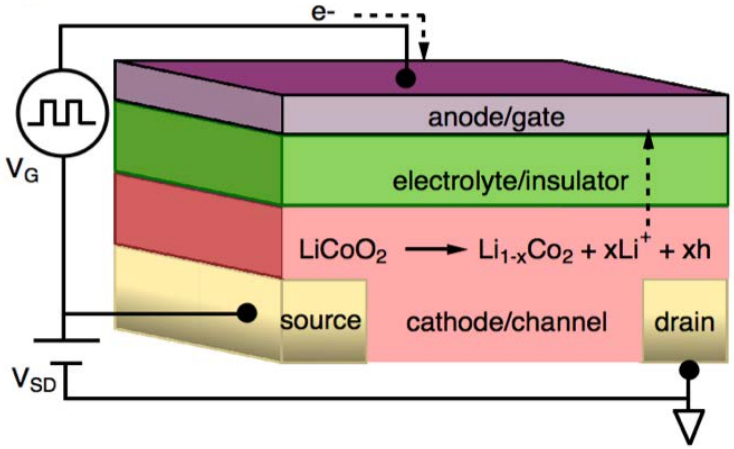
Increasing Network Size

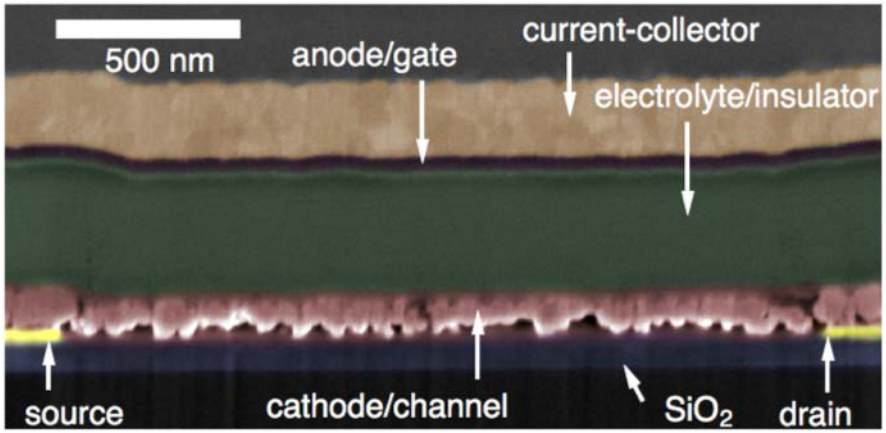


TaOx	Large Images	Small Images	File Types
10 ns	84.45%	71.40%	77.67%
100 ns	78.48%	89.48%	67.78%
1 us	71.48%	71.84%	56.33%

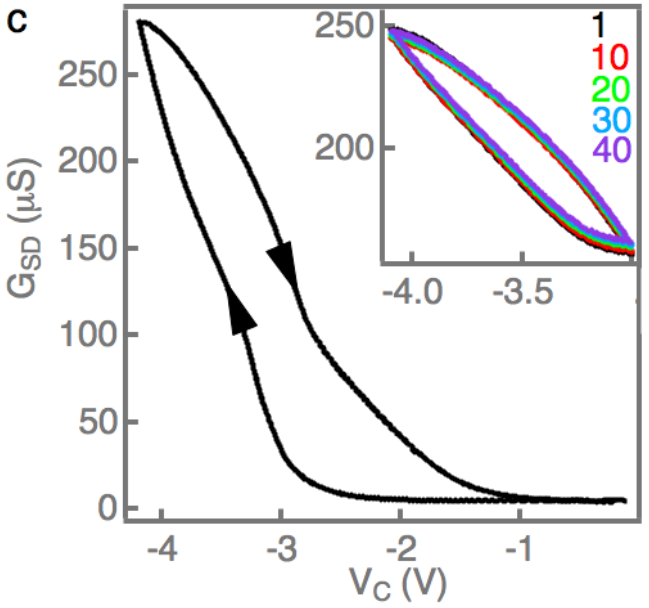
How can training accuracy be improved?

Li-Ion Synaptic Transistor for Analog Computation (LISTA)



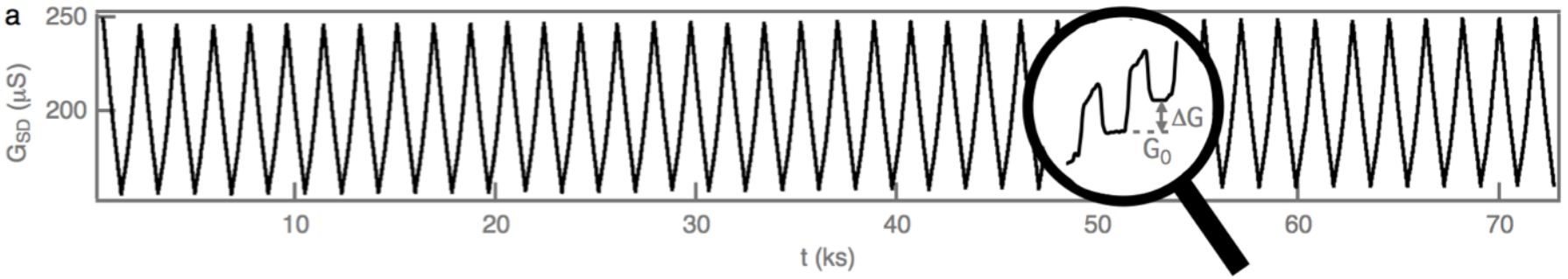


G-V for LISTA Transistor

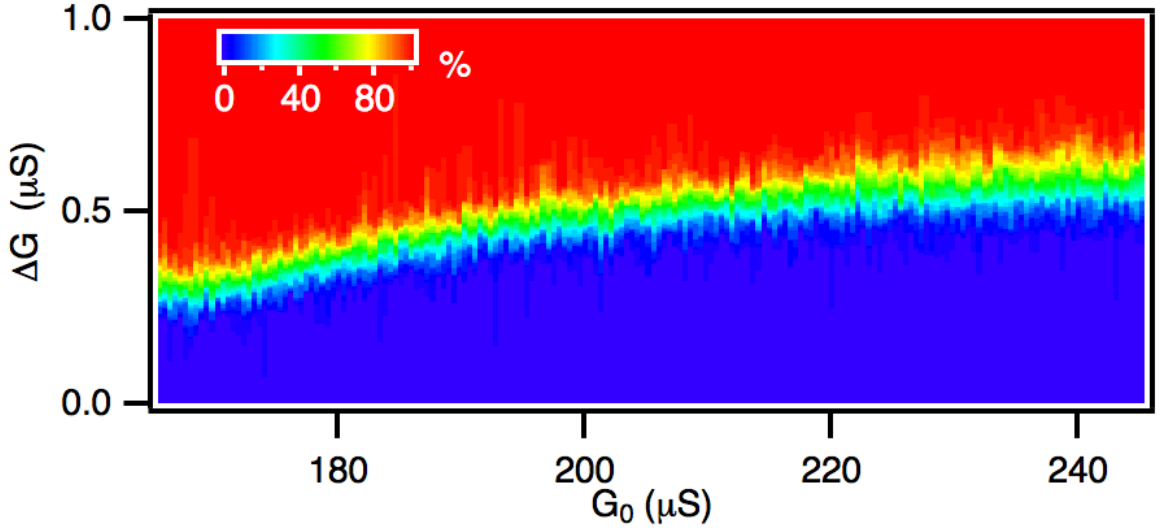


E. Fuller et al, *Adv Mater*, accepted 2017

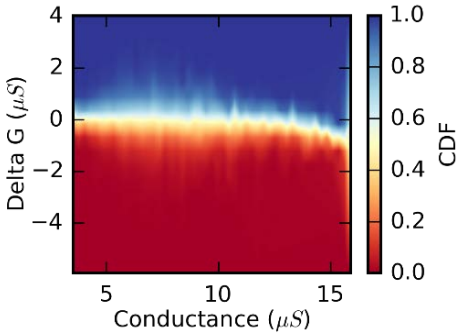
Analog State Characterization



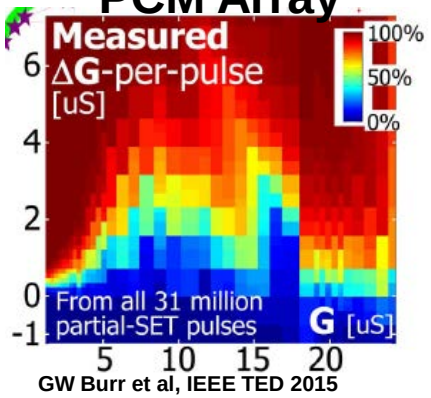
LISTA > 200 states



TaOx ReRAM



PCM Array



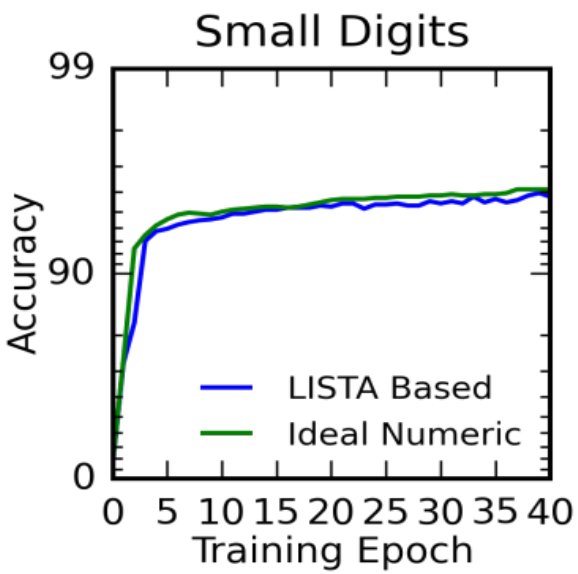
E. Fuller et al, *Adv Mater*, accepted 2017

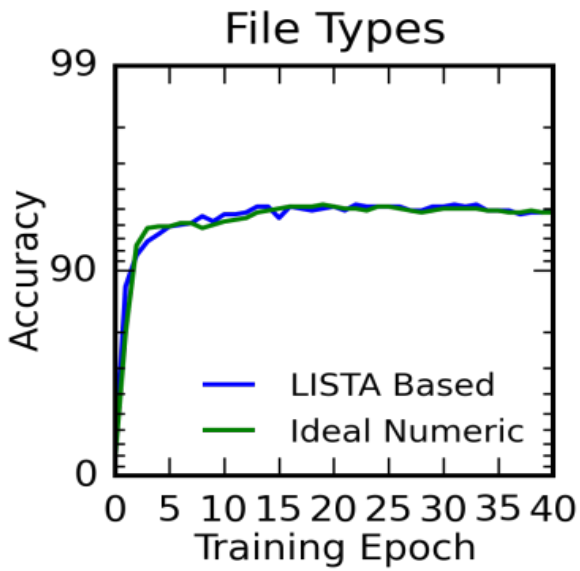
GW Burr et al, IEEE TED 2015

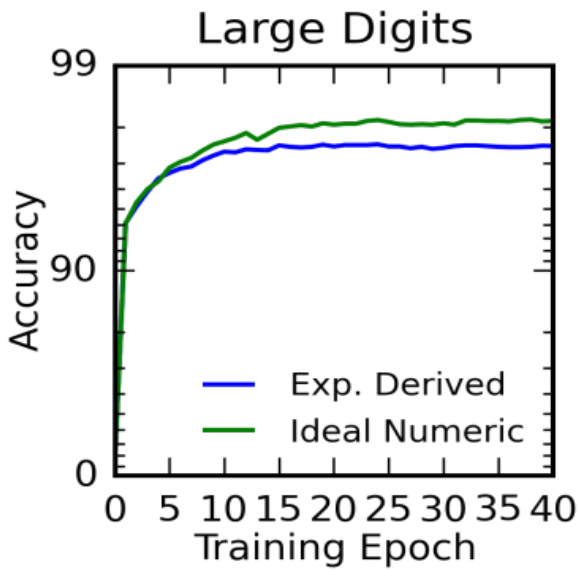
LISTA-device Performance for Backprop Algorithm

Sandia National Laboratories

Increasing Network Size





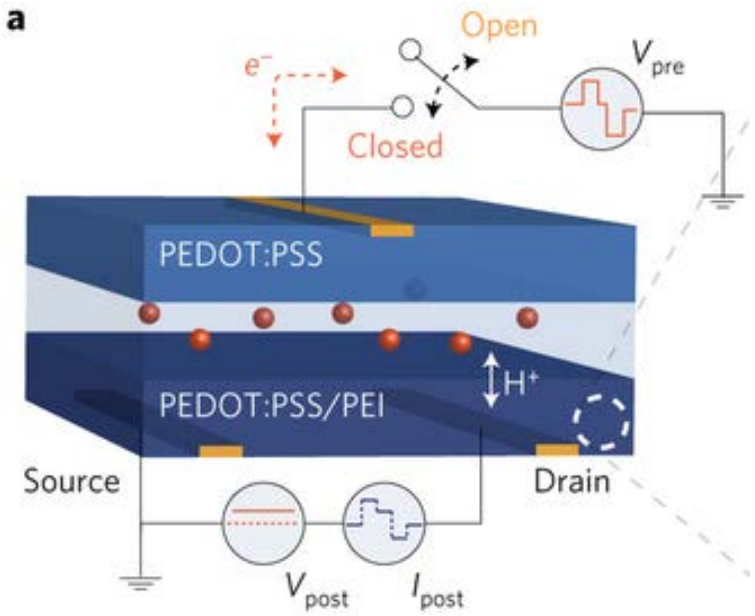


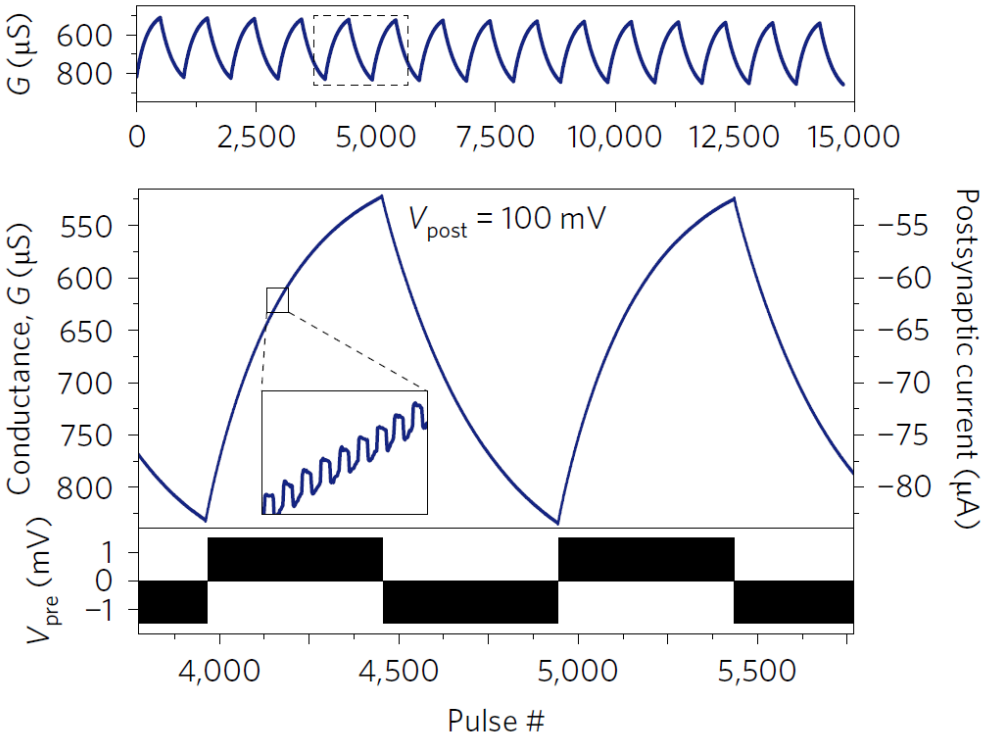
Data set	# Training Examples	# Test Examples	Network Size
UCI Small Digits[1]	3,823	1,797	64×36×10
File Types[2]	4,501	900	256×512×9
MNIST Large Digits[3]	60,000	10,000	784×300×10

See Poster for Detail!

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

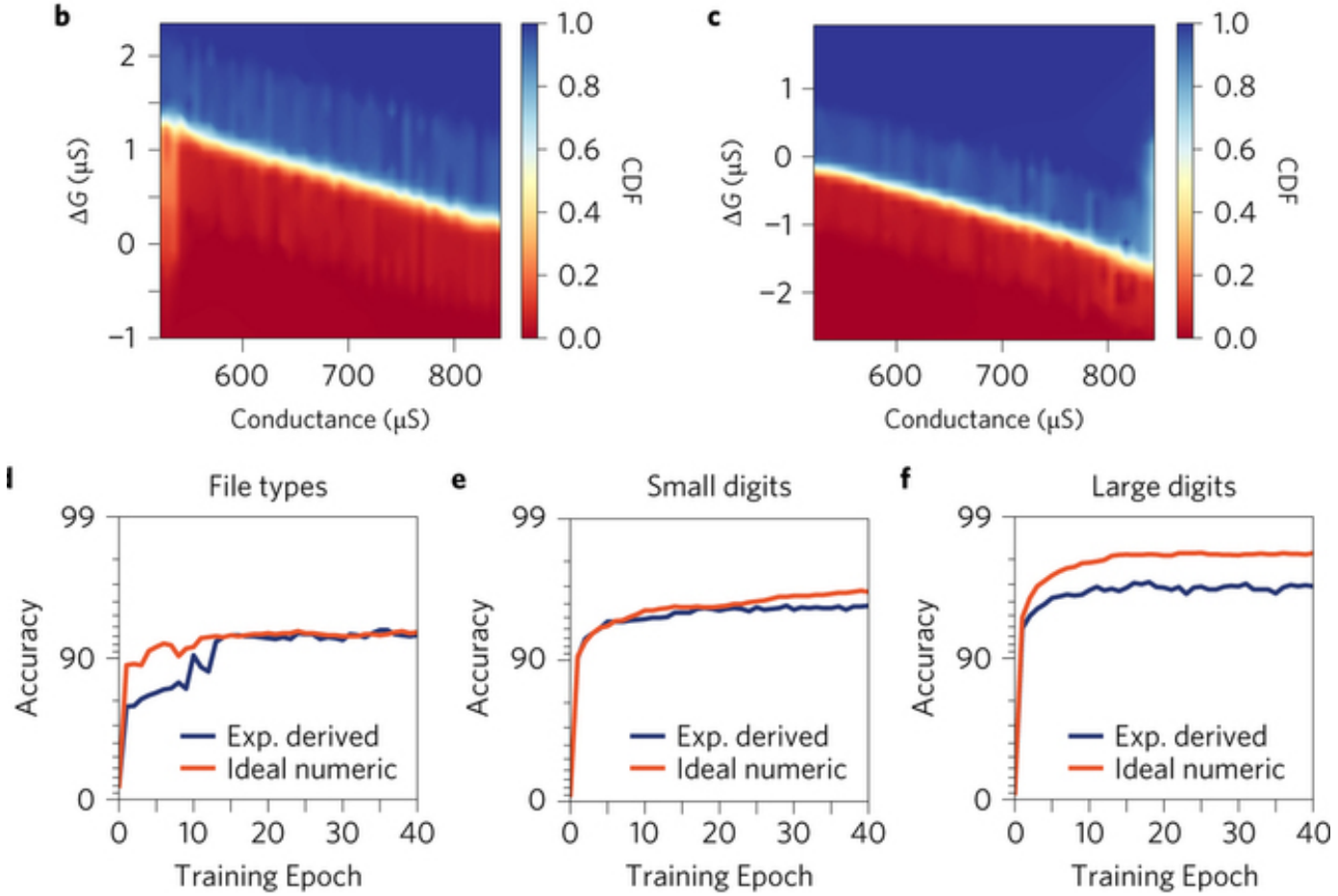
Electrochemical Neuromorphic Organic Device (eNode)





van de Burgt … Saleo, *Nature Mater.*, 2017 in press

Electrochemical Neuromorphic Organic Device (eNode)

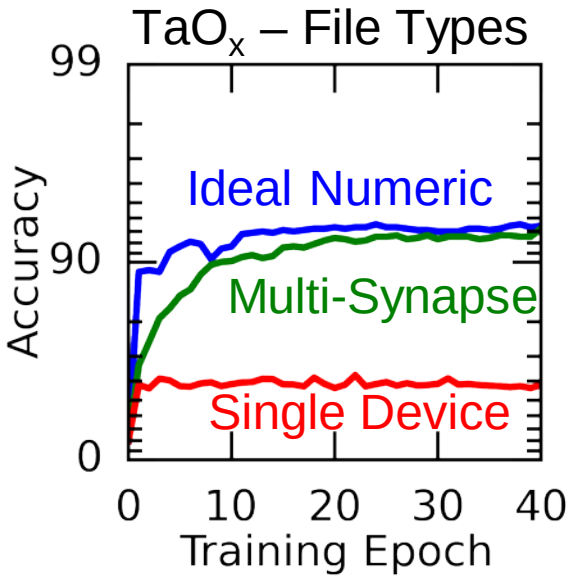


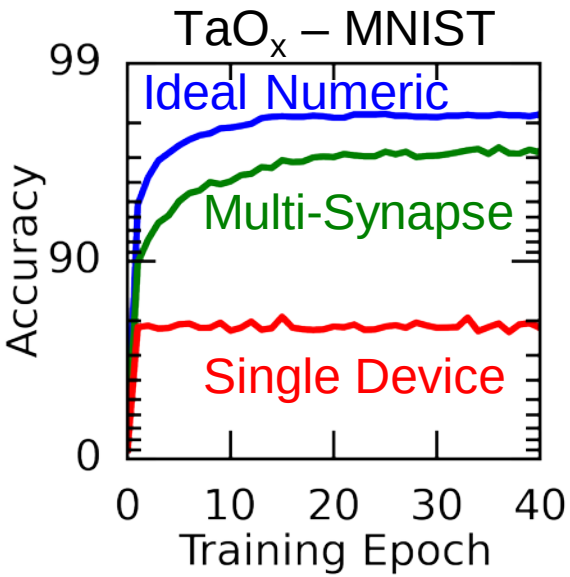
See Poster for Detail!

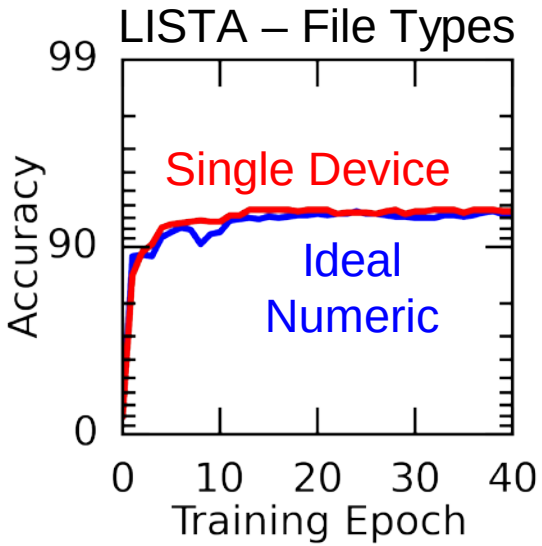
van de Burgt … Saleo, *Nature Mater.*, 2017 in press

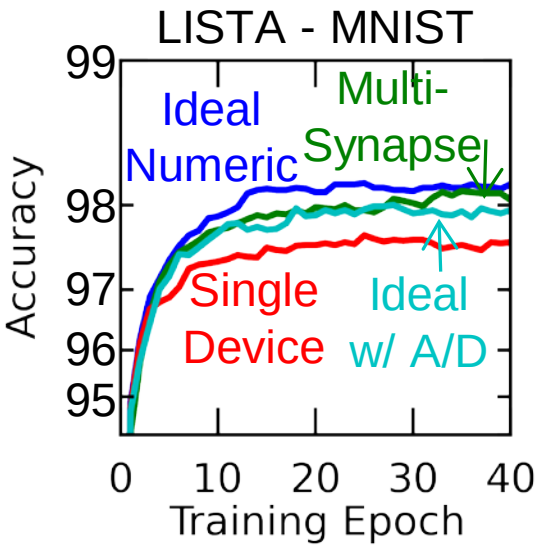
Circuit-Level Improvement

- Allows much closer to ideal with high variability TaOx device
- LISTA achieves essentially perfect accuracy
- Requires tradeoff of energy/latency for accuracy – exact tradeoff depends on algorithm reqs.









Agarwal et al, submitted 2017

Energy and Latency Analysis

Sandia National Laboratories

	SRAM	Digital ReRAM	Analog ReRAM
Equivalent Area ~450 1k × 1k matrices	400 mm^2	32 mm^2	11 mm^2
Total Time per 1-layer cycle 3 Ops: 2 reads, 1 write	~ 100 µs Transpose read dominated	~ 60 µs Update dominated: 10 ns write	~ 5 µs Temporal coding dominated: 256 levels
Total Energy per 1-layer cycle	~ 1000 nJ Multiply dominated	~ 700 nJ Multiply dominated	~ 15 nJ
Matrix Storage Area	95%	50%	17%
Periphery Area	5%	50%	100% crossbar on top of periphery
Matrices per 400 mm^2 Chip	~450	~5,500	~15,000
Energy / Operation	330 fJ	230 fJ	5 fJ
Operations/Second	14 TeraOps/S	270 TeraOps/S	9,000 TeraOps/S

Energy and Latency Analysis

Sandia National Laboratories

		SRAM	Digital ReRAM	Analog ReRAM Crossbar
Matrix Storage 1024×1024 [Values are per-array]	*Area*	800,000 µm²	35,000 µm²	Array: 4,300 µm² Periphery: 8,460 µm²
	Read	30 nJ / 15 µs	15 nJ / 4 µs	~ 3 nJ / ~ 1.5 µs
	Read Transpose	300 nJ / 65 µs	15 nJ / 4 µs	~ 3 nJ / ~ 1.5 µs
	Write	30 nJ / 15 µs	50 nJ / 45 µs	3 nJ / ~ 1.5 µs
Multiply Accumulators [256 in parallel]	*Area*	19,000 µm²		FREE
	Run: 1M ops	200 nJ / 4 µs		
Output LUT	*Area*	1,400 µm²		Uses Digital Methods
	Read	1 nJ / 1 µs		
Input/Output Buffers [8 bits]	*Area*	13,000 µm²		Uses Digital Methods
	Per Run	~ 0.1 nJ		
Vector Cache* 16 entries 1024x8-bit [Values are per 1024x8 vector]	*Area*	11,250 µm²	500 µm²	Uses Digital Methods
	Read	~ 0.1 nJ / ~ 0.2 µs	~ 1 nJ / 4 ns	
	Write	~ 0.1 nJ / ~ 0.2 µs	~ 1 nJ / 50 ns	

Conclusion

- Dennard (constant power density) scaling has ceased and Moore's law is slowing
- New paradigms like neuromorphic computing will be required for sub-fJ computing
- We now require a device through system design mentality
  - Motivation behind CrossSim
  - See poster for more detail on CrossSim
- Oxide-based resistive memory offers intriguing device options for both eras
- Novel LISTA and eNode devices, offer significant potential in the development of a low energy neural accelerator
  - See LISTA and eNode posters for more detail on these

Thank you!

Sandia National Laboratories

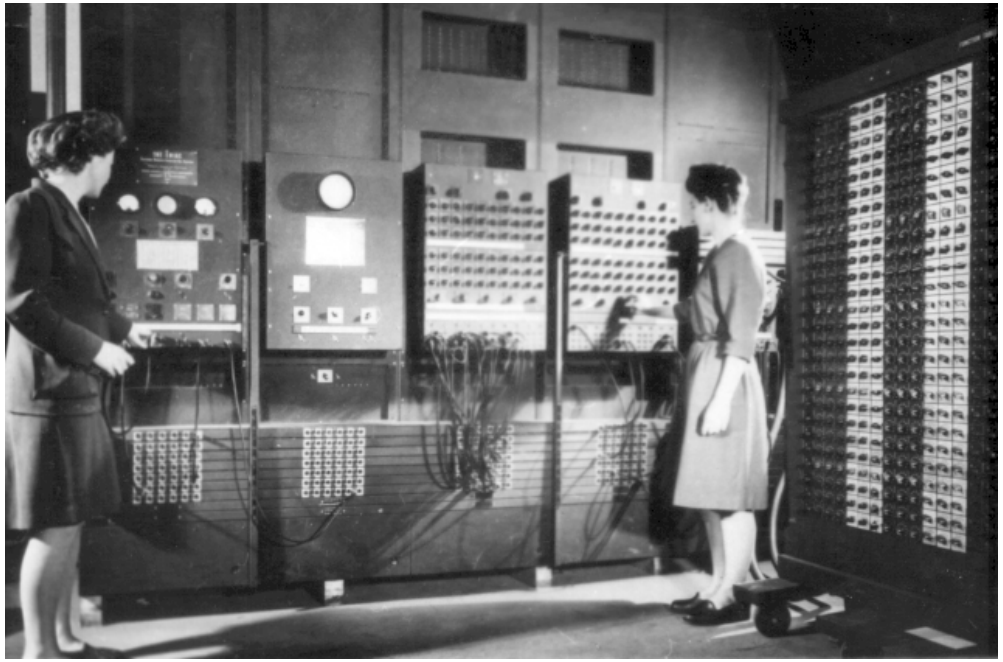
Acknowledgements

- This work is funded by Sandia's Laboratory Directed Research and Development as part of the Hardware Acceleration of Adaptive Neural Algorithms Grand Challenge Project
- Many shared ideas among collaborators:
  - Alberto Saleo, Yoeri van de Burgt, Stanford
  - Engin Ipek, U Rochester
  - Dave Mountain, Mark McLean, US Government
  - Stan Williams, John Paul Strachan, HPL
  - Dhireesha Kudithipudi, RIT
  - Jianhua Yang, U Mass
  - Hugh Barnaby, Mike Kozicki, Sheming Yu, ASU
  - Tarek Taha, U Dayton
  - Paul Franzon, NC State University
  - Sayeef Salahuddin, UC Berkeley
  - Dozens of others…
- **We are especially interested in collaborations on cross-sim!**

ENIAC

Sandia National Laboratories

Electronic Numerical Integrator And Computer
Developed by US Army/U Penn – 1946
150 kW, 357 FLOPs
400 J/FLOP (10 bit)



Where Are we Today?

Sandia National Laboratories

- **Single Unit: Nvidea Tesla P100 GPU**
  - **Most advanced GPU processor specs, released late 2016**
  - **Target's deep learning and neural applications**
  - **20 TFLOPs 16 bit peak performance w/ peak power dissipation of 300W**
  - **70 GFLOPs/watt or about 15 pJ/FLOP (16 bit)**

- **Supercomputer: Sunway TaihuLight (China)**
  - **Top supercomputer in the world**
  - **ShenWei processor**
  - **90 PFLOPs peak, 15 MW power**
  - **6 GFLOPs/W or about 170 pJ/FLOP**

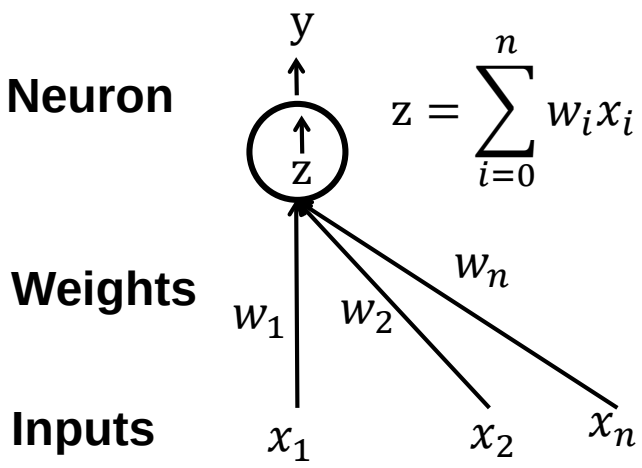
- **Need >1000x improvement to tackle *internet-scale* problems**

Basics of Neural Networks

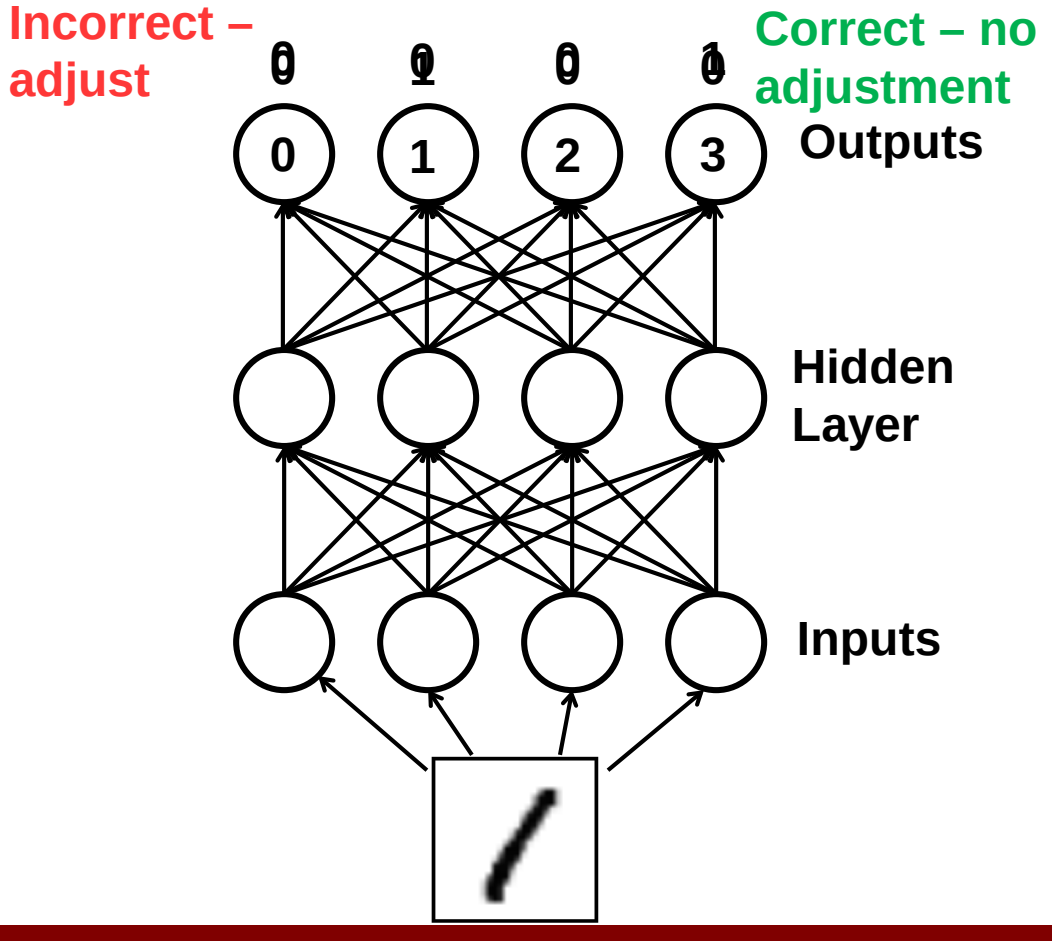
Sandia National Laboratories

Basic Building Block

$$y = \frac{1}{1 + e^{-z}}$$



Simple Network: Backpropagation

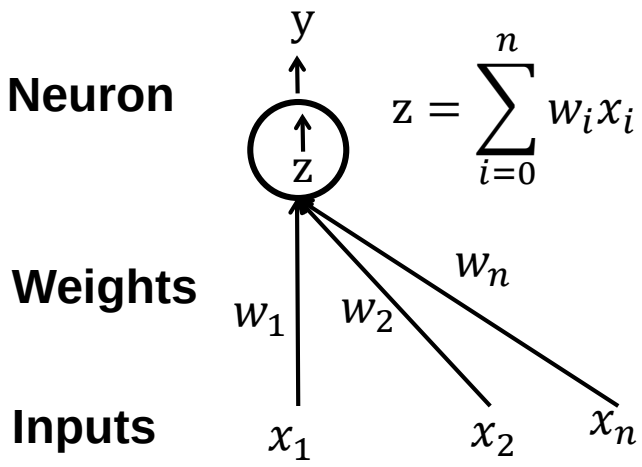


Another Analogy

Sandia National Laboratories

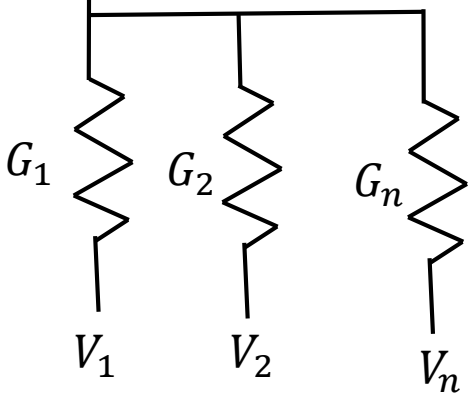
Mathematical

$$y = \frac{1}{1 + e^{-z}}$$



Electrical

$$I = \sum_{i=0}^{n} G_i V_i$$



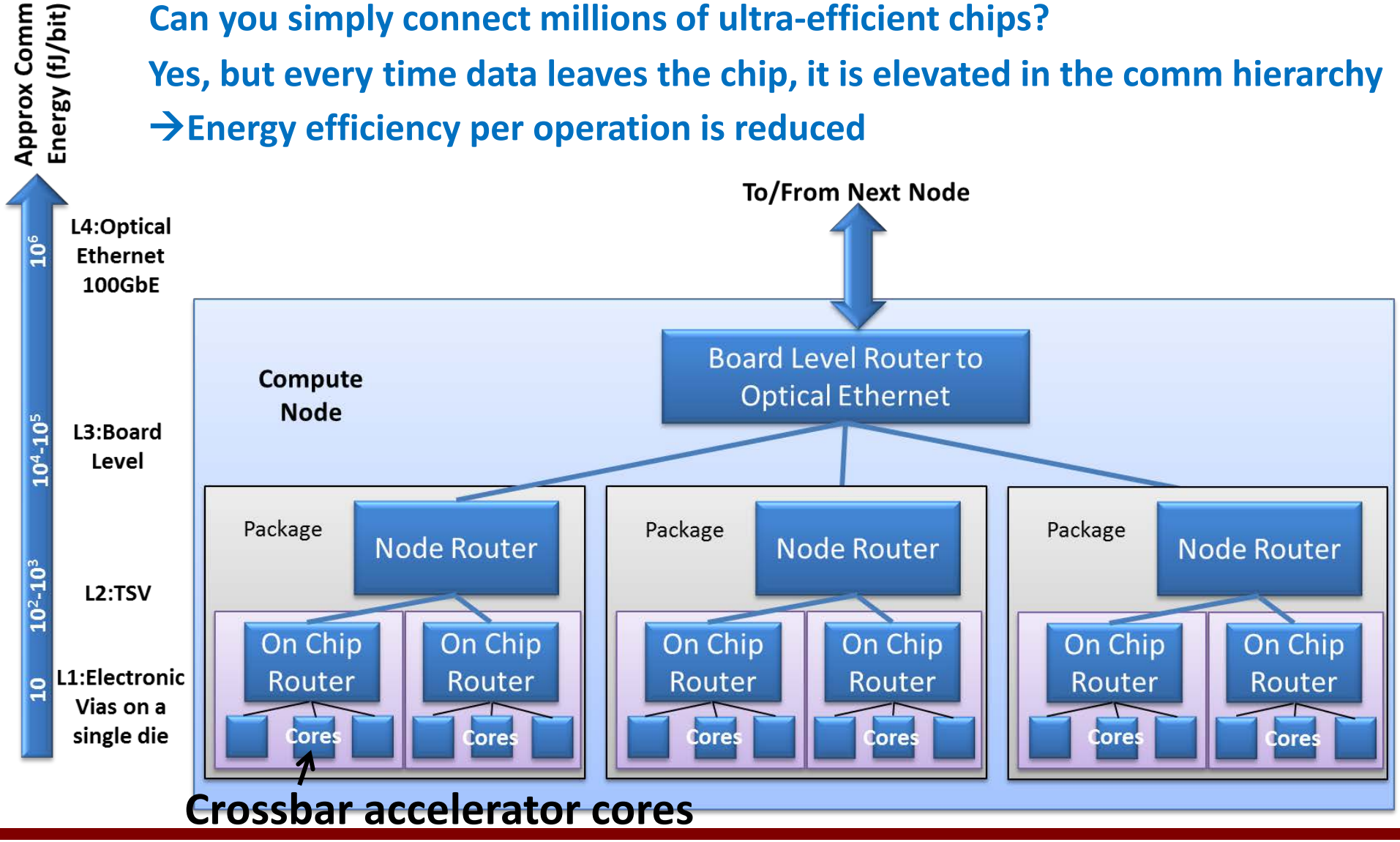
Why is it essential to cram so many computations on a single chip?

Sandia National Laboratories

Can you simply connect millions of ultra-efficient chips?

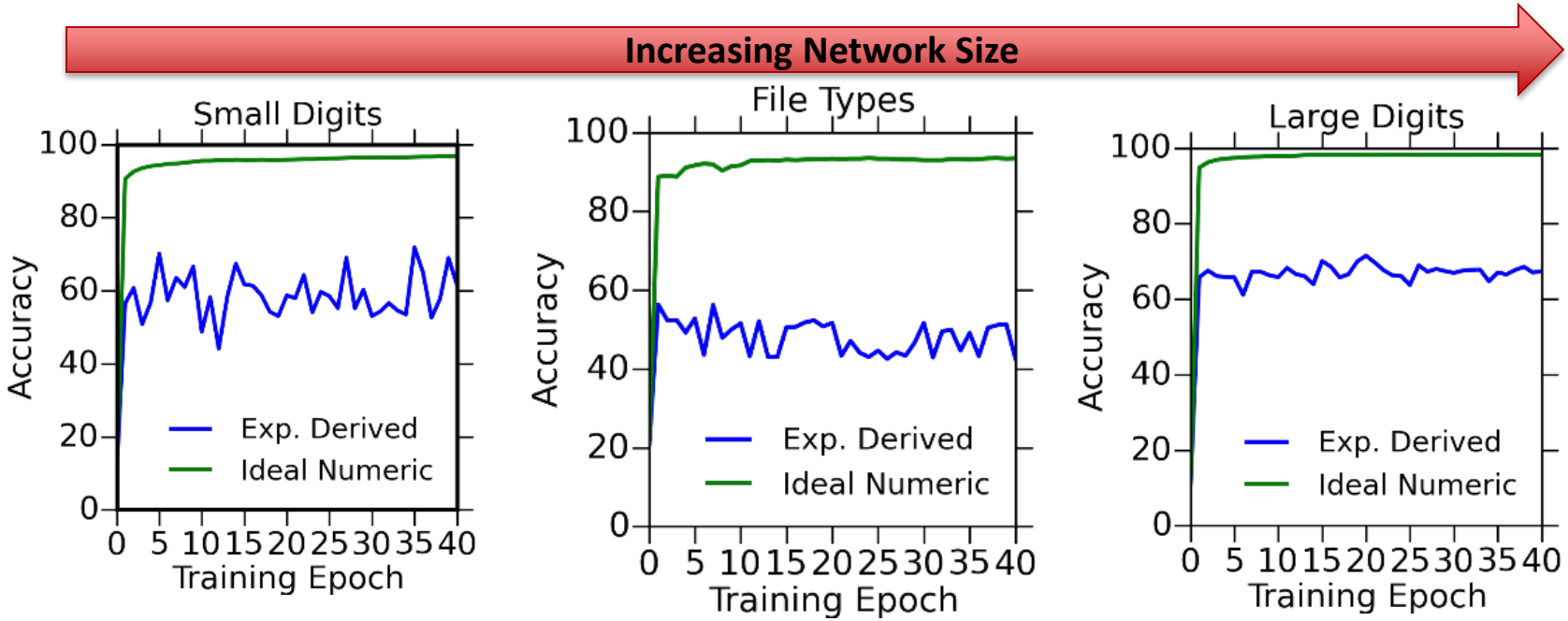
Yes, but every time data leaves the chip, it is elevated in the comm hierarchy

→Energy efficiency per operation is reduced



TaOx ReRAM in Backprop Training

Sandia National Laboratories



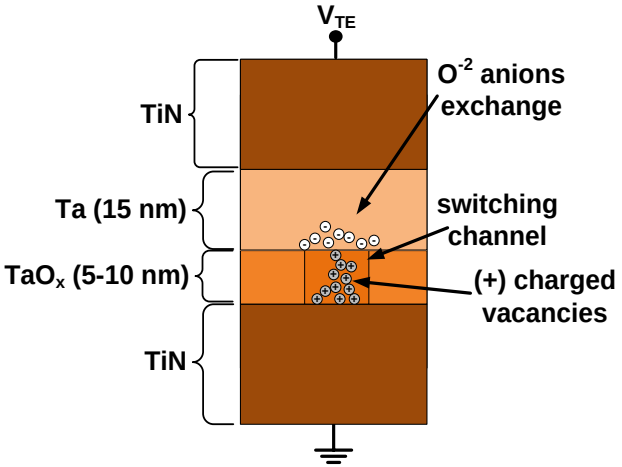
Data set	# Training Examples	# Test Examples	Network Size
UCI Small Digits[1]	3,823	1,797	64×36×10
File Types[2]	4,501	900	256×512×9
MNIST Large Digits[3]	60,000	10,000	784×300×10

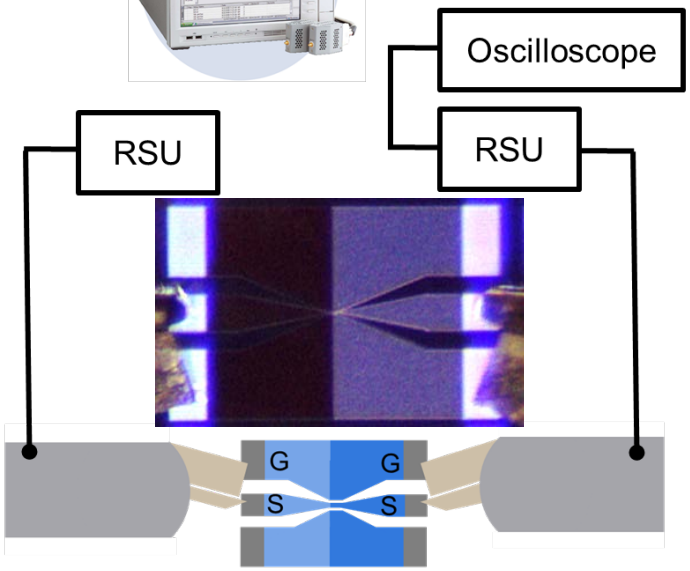
1 µs pulses

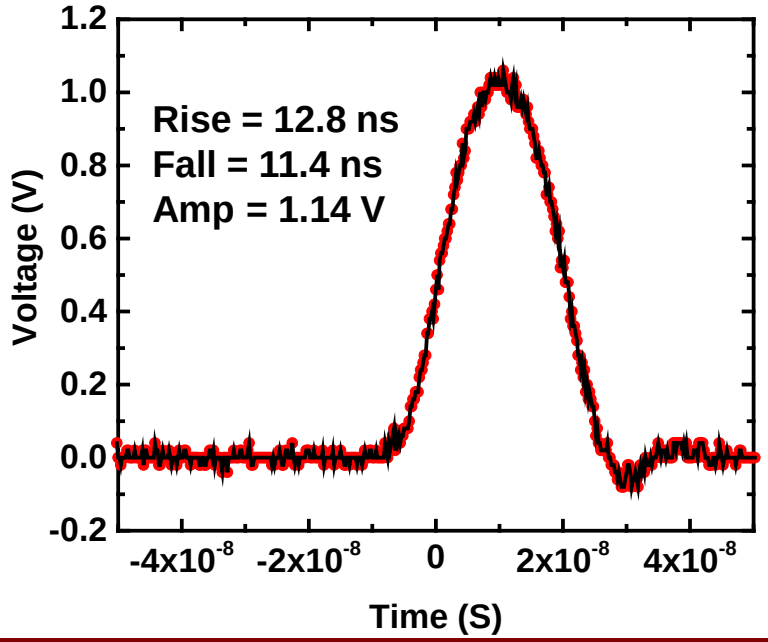
ReRAM Measurements

Sandia National Laboratories

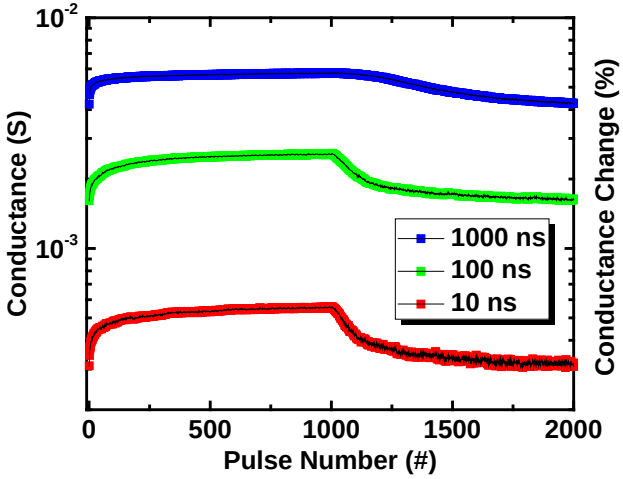
- **DC Current-voltage "loops" sweeps are not time-controlled**
  - **Excessive heating and early wearout**
  - **Do not provide info on dynamics**
- **Physical switching < 10ns**
- **Need pseudo RF setup to measure**
  - **Ground/signal, conductor backed**
  - **Agilent B1530 module**
  - **10 ns RT/FT, 10 ns PW**
  - **1 V nominal, ~140 mV overshoot**

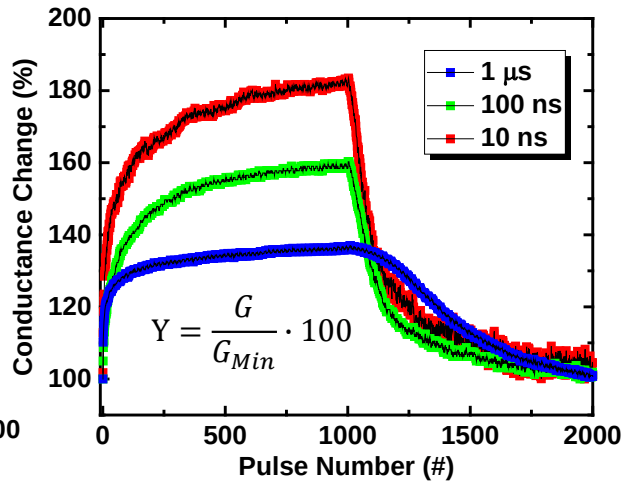


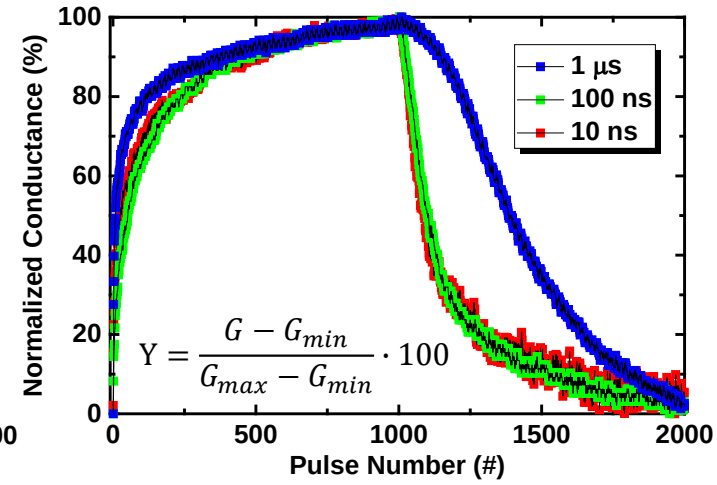




Effect of Pulse Width and Edge Time



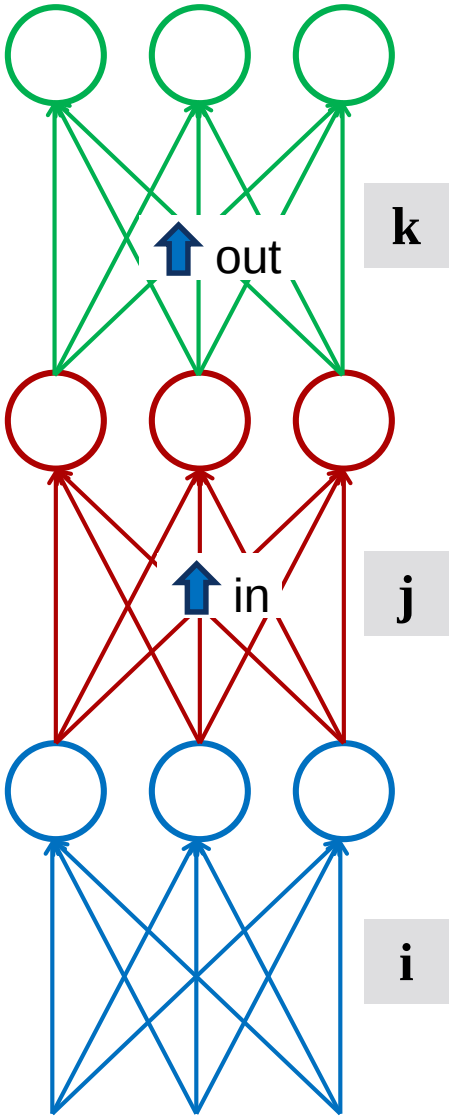




- **Shorter pulses may be employed to lower conductance switching range**
- **Linearity qualitatively similar across Pulse Width (PW) and Edge Time (ET)**
  - **Best for SET at 100 ns**
  - **Best for RESET at 1 us**
- **Relative conductance change increased with shorter Pulse Width / Edge Time**

Nominal Pulse Voltage Values: SET: +1 V RESET: -1 V

Analog Core: Forward Propagation



O(N²) Operations

O(N) Operations

y_j

Digital Core

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

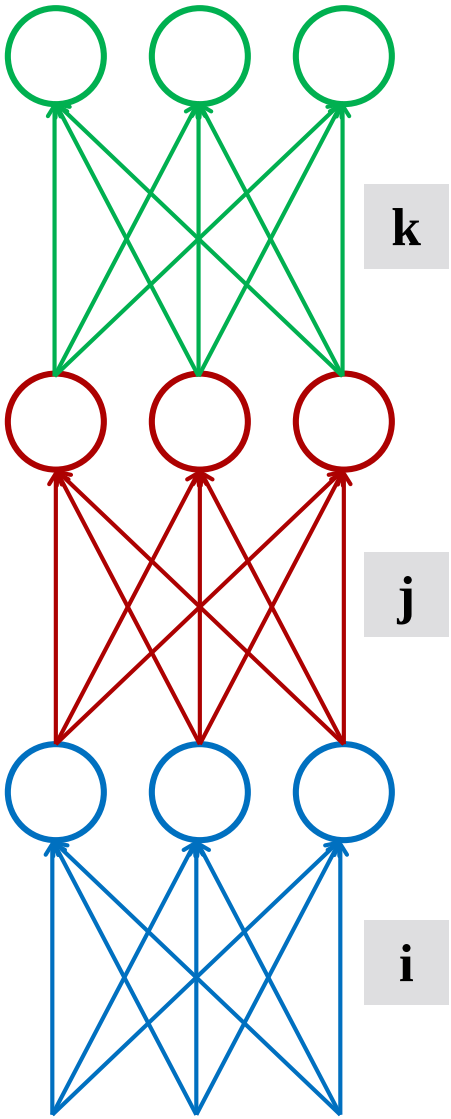
Neuron Function

w_{ij}

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

y_i

Analog Core: Back Propagation



error δ_k

$\Delta_k = \sum_k w_{jk}\delta_k$

Δ_k

$\eta \times \delta_j$

y_i

w_{ij}

z_j

Digital Core

$\delta_j = y'(z_j) \cdot \Delta_k$

y_i

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

y_i

O(N²) Read Operations

O(N) Operations

O(N²) Write Operations