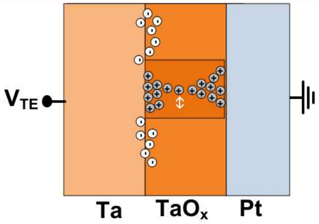
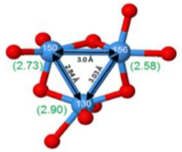
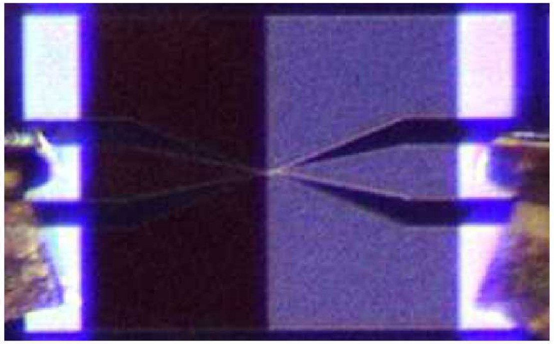
SAND2018-9678C

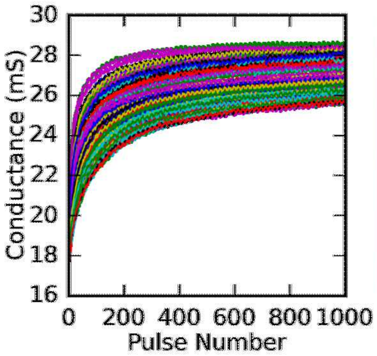
[redacted]

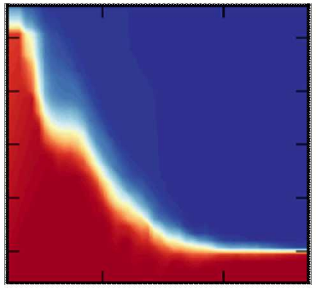
[redacted]

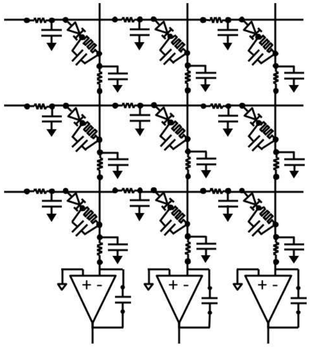












Energy Efficient Neuromorphic Algorithm Acceleration Enabled by Resistive Memory (ReRAM) Crossbars

Matthew J. Marinella*, S. Agarwal, R. Jacobs-Gedrim, D.R. Hughart, I. Richter, A. Hsia, E. Fuller, A.A. Talin, R. Goeke, S.J. Plimpton, and C.D. James

Sandia National Laboratories

**matthew.marinella@sandia.gov*

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.

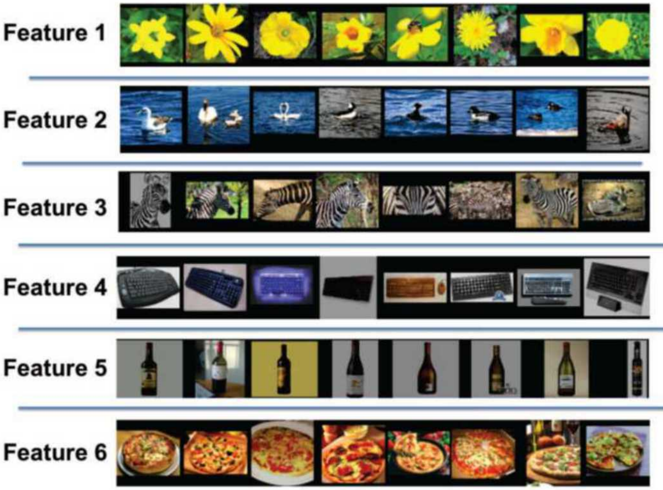
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

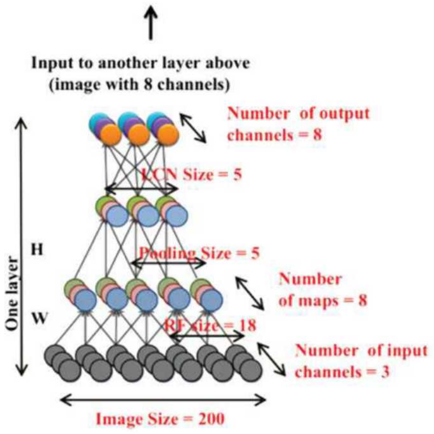
Outline

- Intro and Motivation
- ReRAM-Based Accelerator Key Concepts
- ReRAM-Based Accelerator Model
- Conclusion

Why do we need more efficient computers?

- **Google Deep Learning Study**
  - **16000 core, 1000 machine GPU cluster**
  - **Trained on 10 million 200x200 pixel images**
  - **Training required 3 days**
  - **Training set size set by what can be completed in less than one week**
- **What would they like to do?**
  - **~2 billion photos uploaded to internet per day (2014)**
  - **Can we train a deep net on <u>one day of image data</u>?**
  - **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)***
  - **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

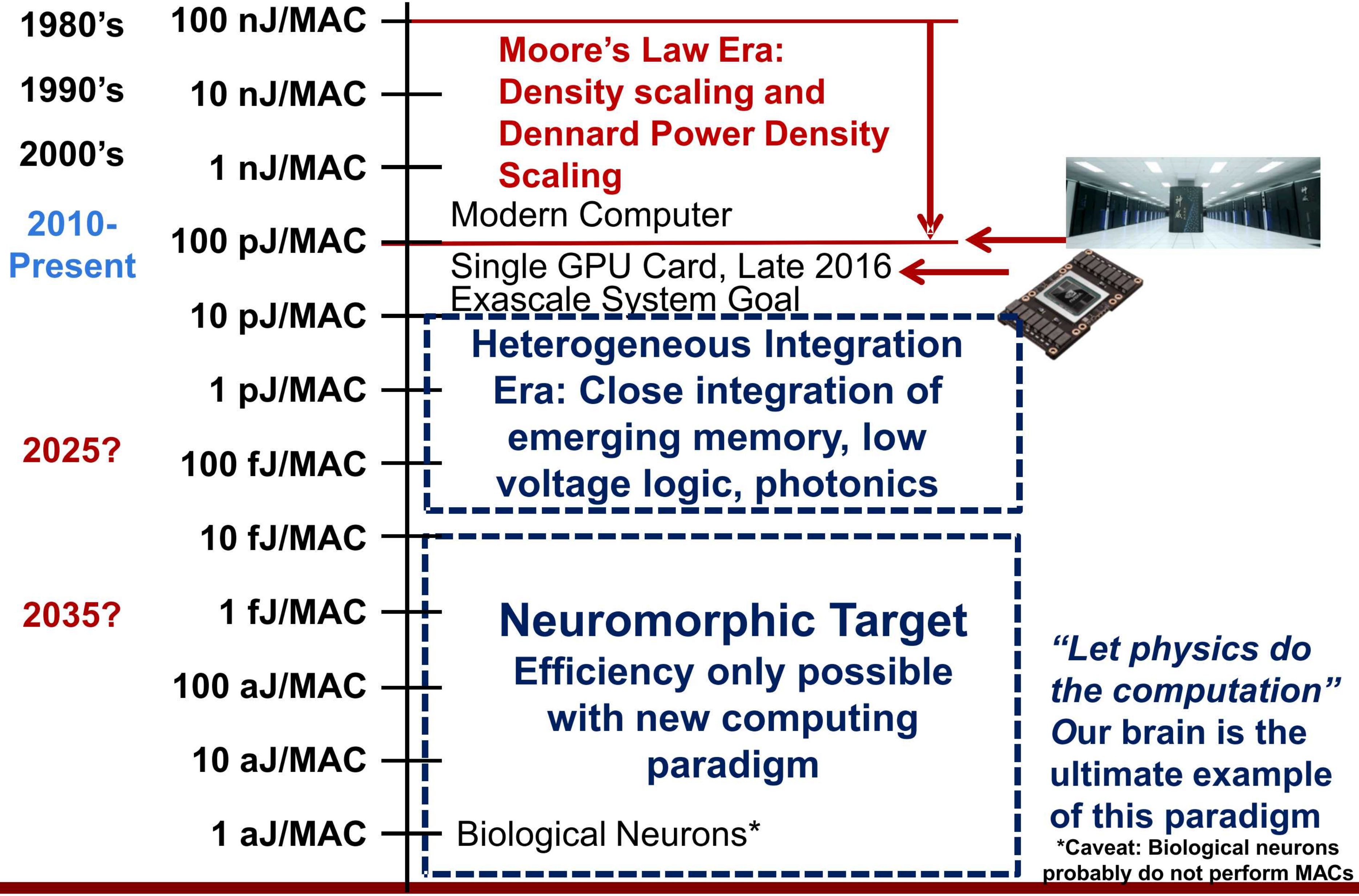
Where Are we Today?

- **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**

Evolution of Computing Machinery



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*

"Let physics do the computation"

Our brain is the ultimate example of this paradigm

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

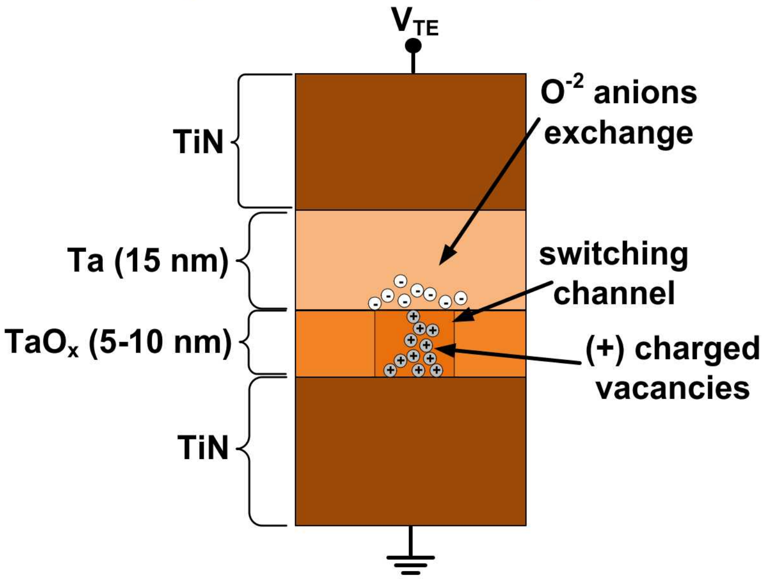
Inspired by Hasler and Marr, Frontiers in Neuroscience, 2013

Outline

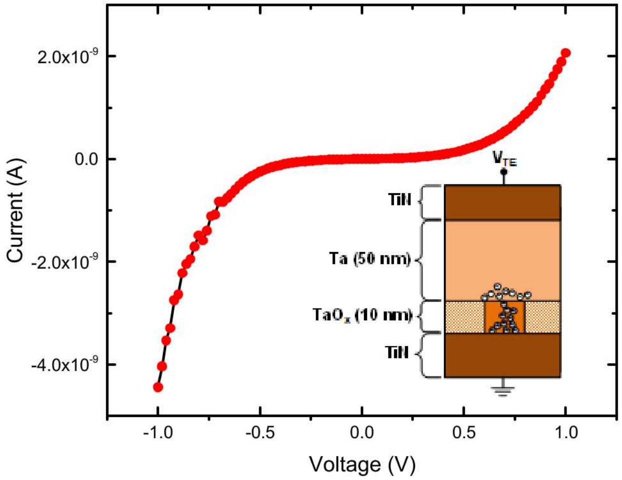
- Intro and Motivation
- ReRAM-Based Accelerator Key Concepts
- ReRAM-Based Accelerator Model
- Conclusion

Metal Oxide Resistive RAM (ReRAM)

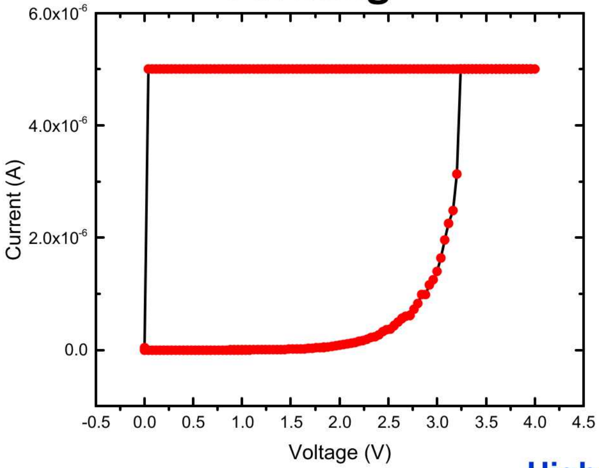
- Sandia TiN/Ta/TaOx/TiN example device
- Starts as insulating MIM structure
- Forming: remove O^{2-} → soft breakdown
- Bipolar resistance modulation
- Excellent memory attributes: Switching in less than 1ns, less than 1 pJ demonstrated, scaling to 5nm, $>10^{12}$ write cycles



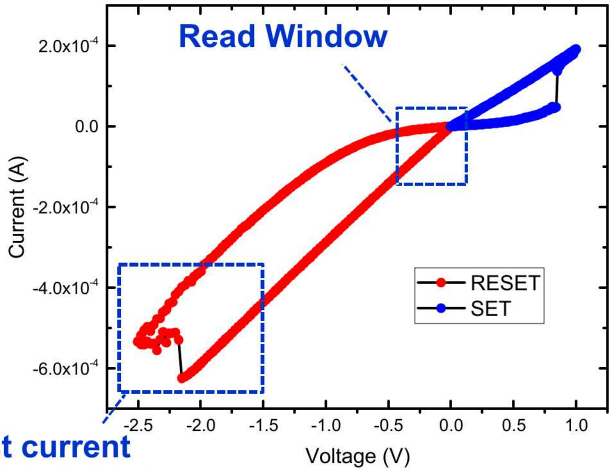
Pre-Form I/V



Forming



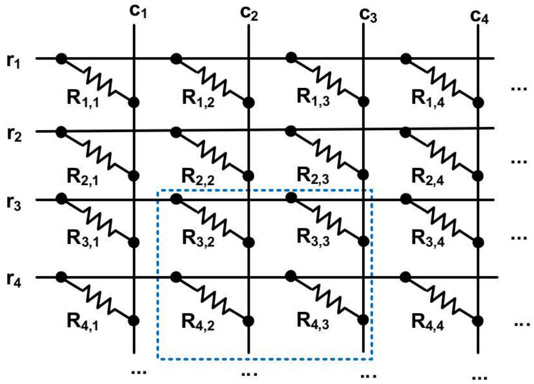
SET-RESET

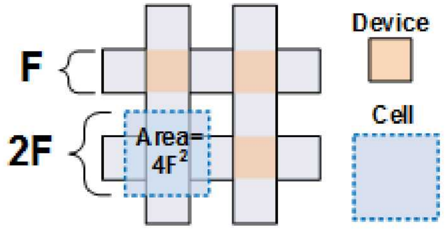


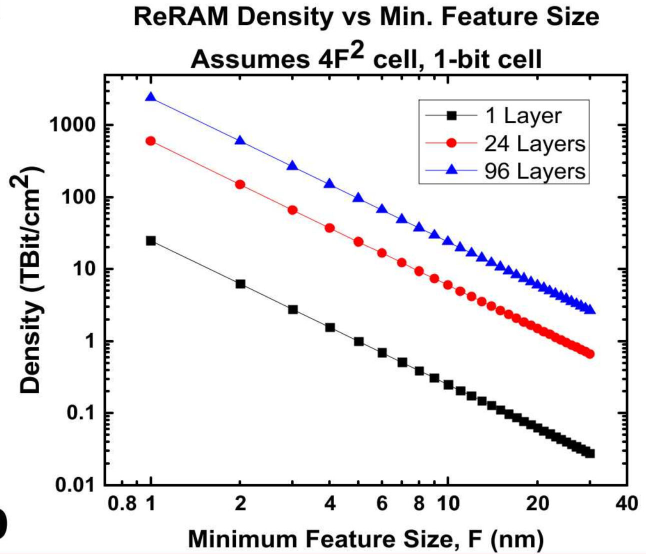
Highest current switching process

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





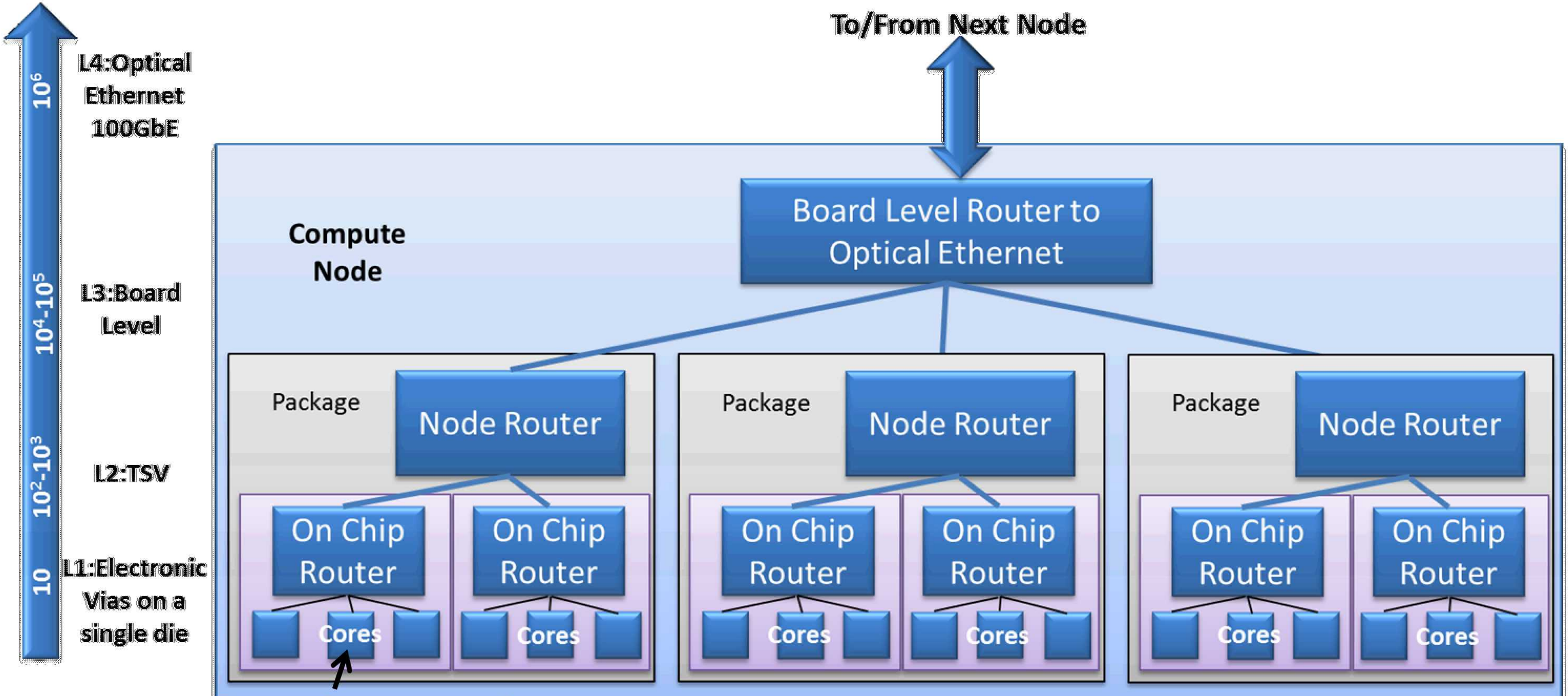


Why is it essential to cram so many computations on a single chip?

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



Crossbar accelerator cores

How does a crossbar perform a useful computation per device?

- Electronic Vector Matrix Multiply

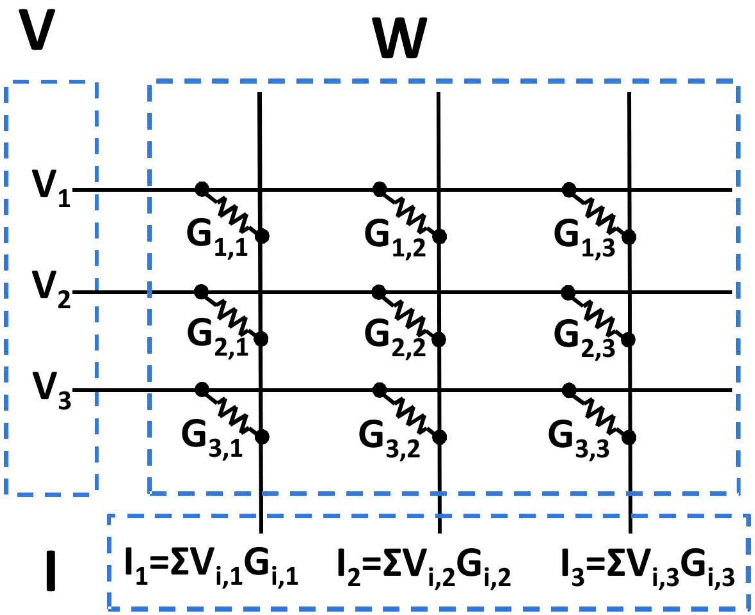
Mathematical

$$V^TW=I$$

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

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

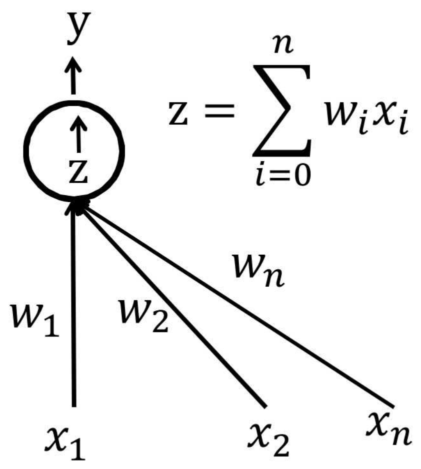
Electrical



Basics of Neural Networks

Basic Building Block

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



Neuron

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

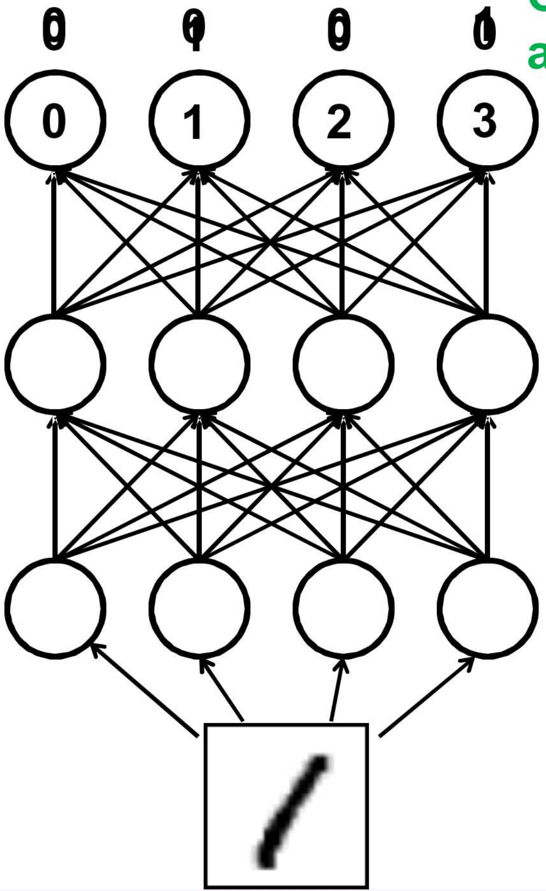
Weights

Inputs

Simple Network: Backpropagation

Incorrect – adjust

Correct – no adjustment

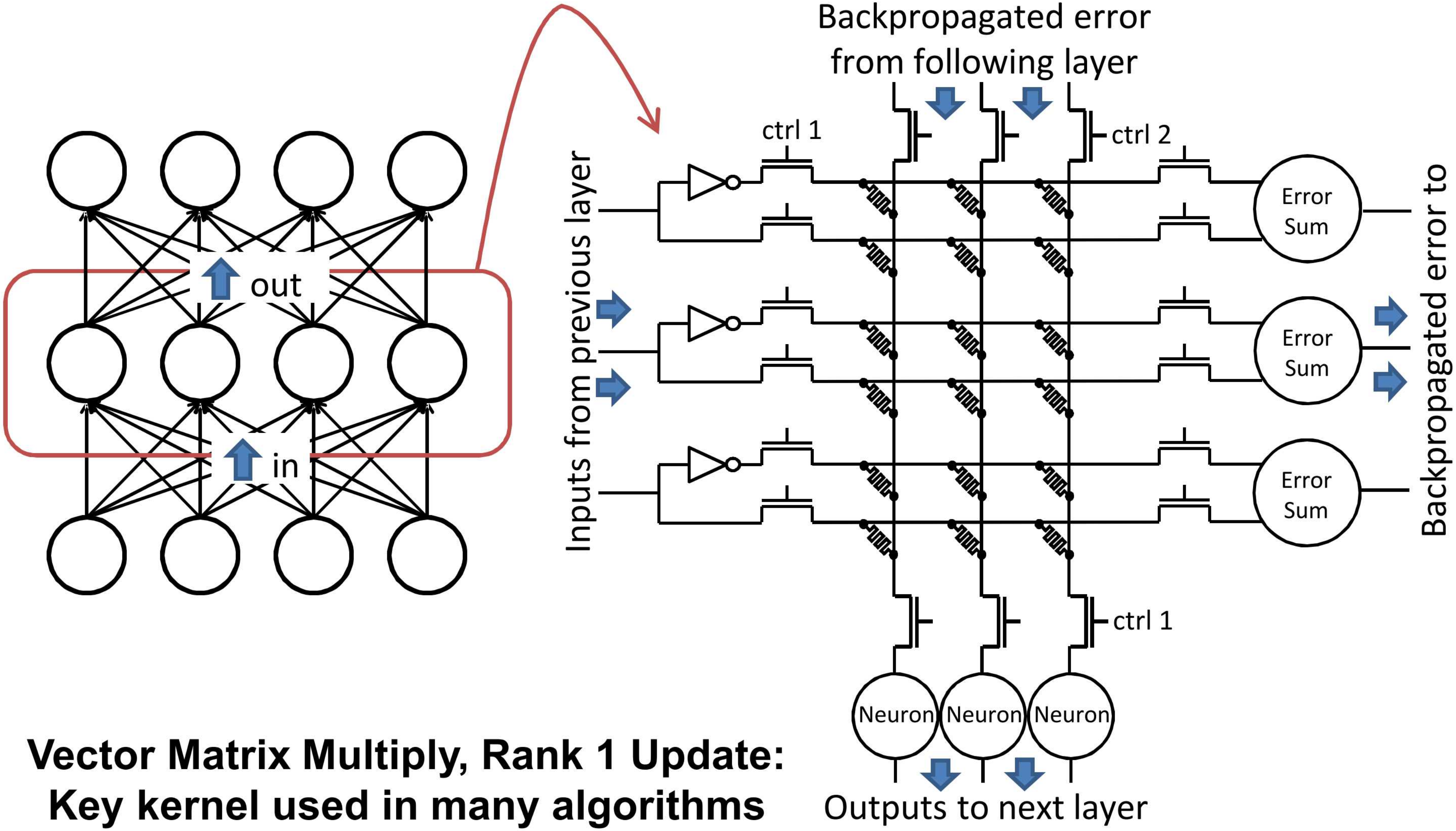


Outputs

Hidden Layer

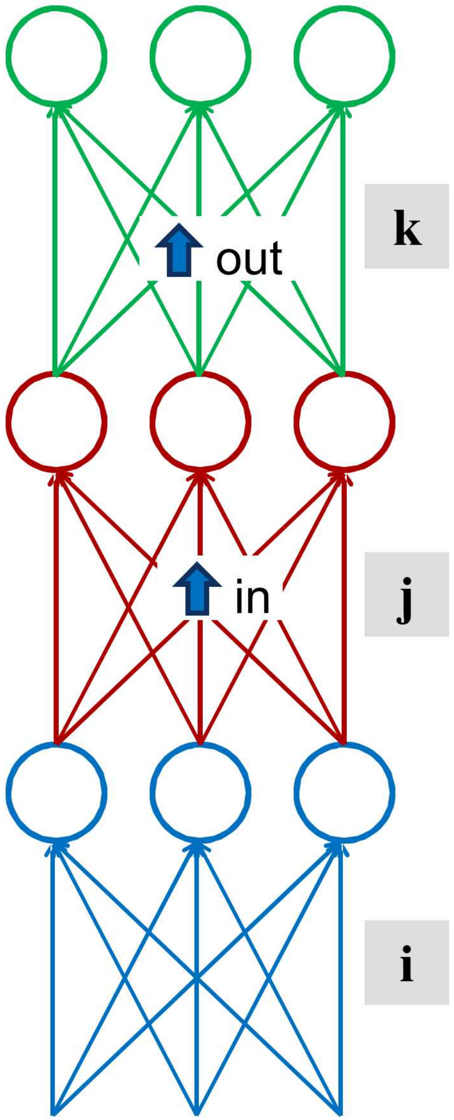
Inputs

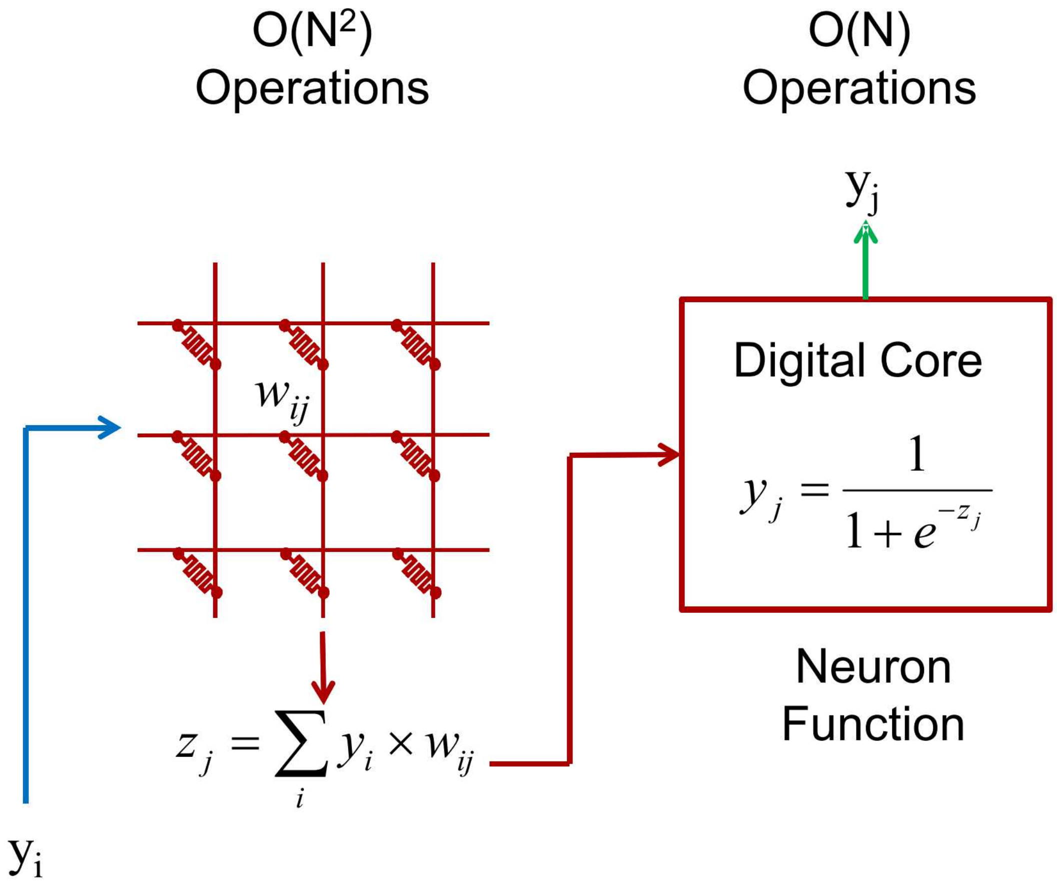
Mapping Backprop to a Crossbar


out
in
Backpropagated error from following layer
ctrl 1
ctrl 2
Inputs from previous layer
Error Sum
Error Sum
Error Sum
Backpropagated error to previous layer
ctrl 1
Neuron
Neuron
Neuron
Outputs to next layer

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

Analog Core: Forward Propagation





O(N^2)
Operations

O(N)
Operations

w_{ij}

y_i

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

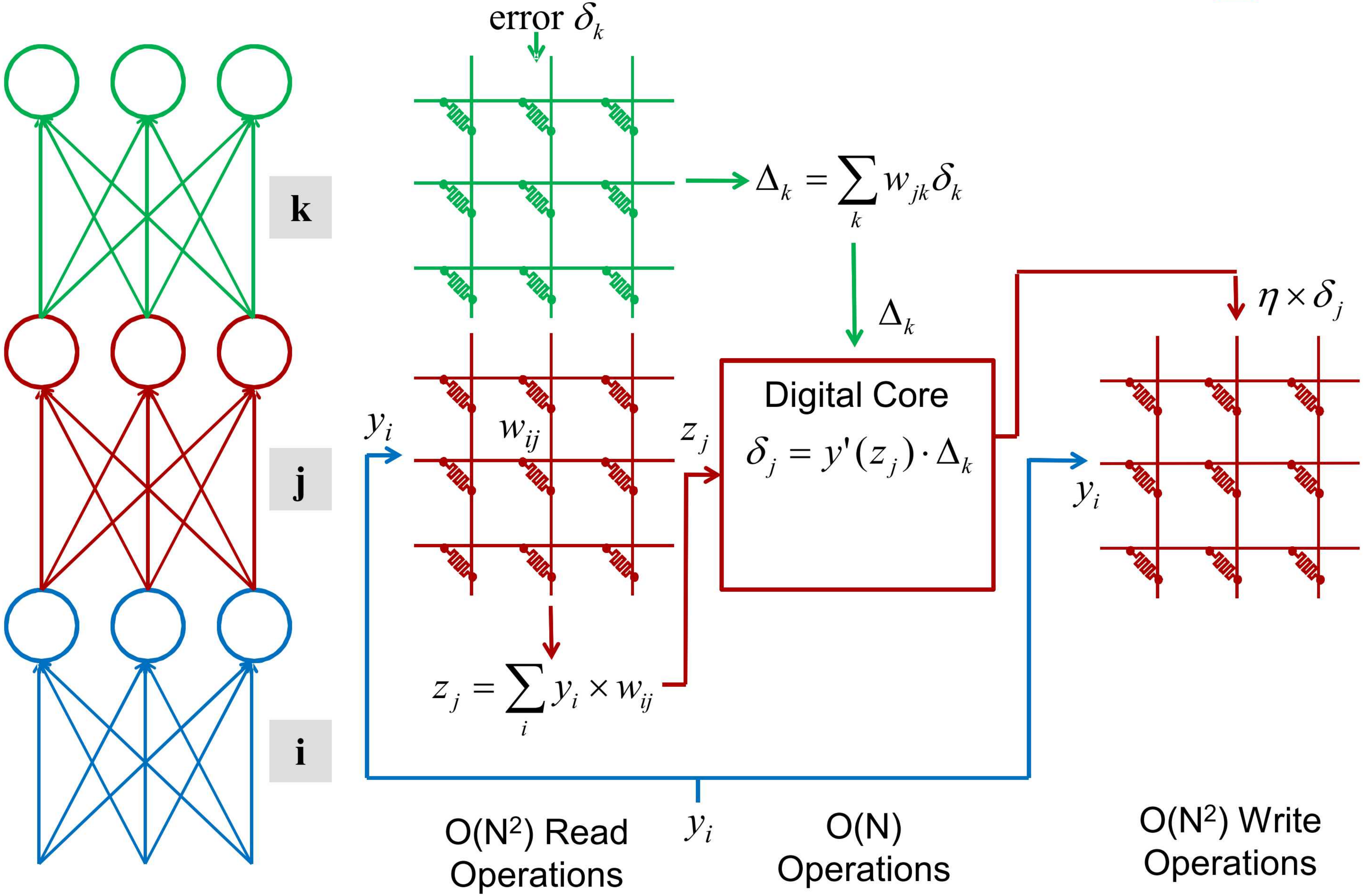
Digital Core

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

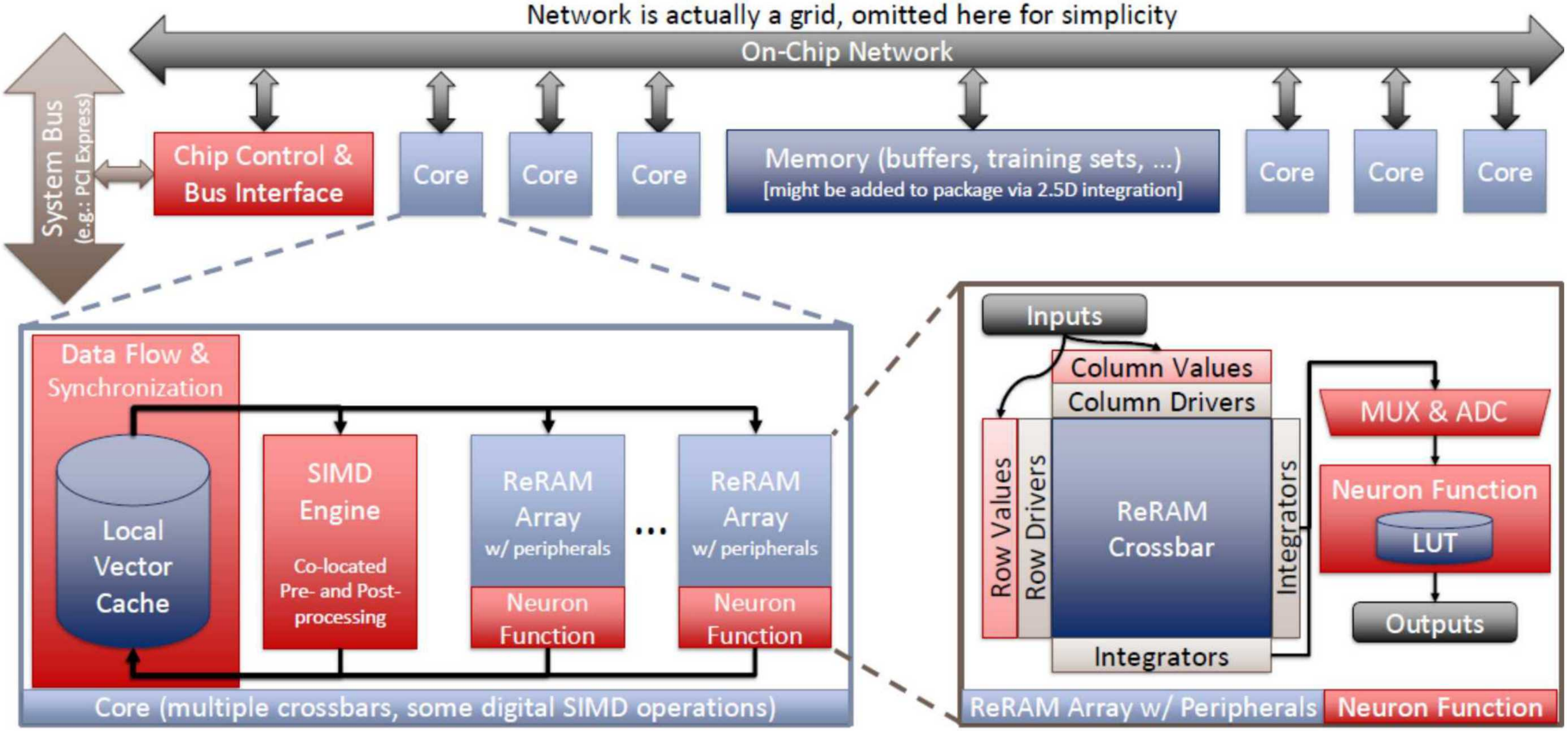
y_j

Neuron
Function

Analog Core: Back Propagation



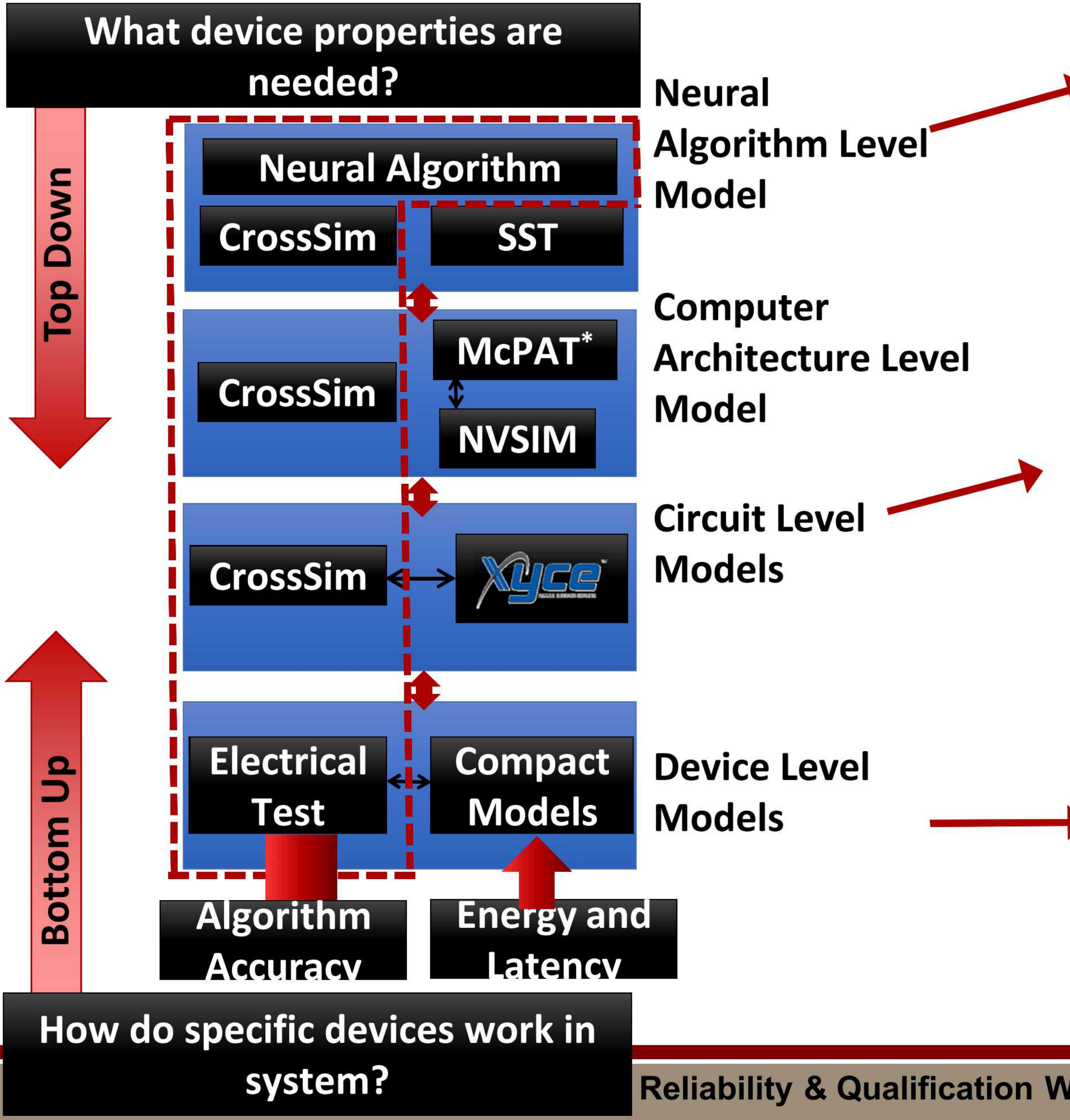
Accelerator Architecture


Network is actually a grid, omitted here for simplicity
On-Chip Network
System Bus
(e.g.: PCI Express)
Chip Control & Bus Interface
Core
Core
Core
Memory (buffers, training sets, …)
[might be added to package via 2.5D integration]
Core
Core
Core
Data Flow & Synchronization
Local Vector Cache
SIMD Engine
Co-located Pre- and Post-processing
ReRAM Array w/ peripherals
Neuron Function
…
ReRAM Array w/ peripherals
Neuron Function
Core (multiple crossbars, some digital SIMD operations)
Inputs
Column Values
Column Drivers
Row Values
Row Drivers
ReRAM Crossbar
Integrators
Integrators
MUX & ADC
Neuron Function
LUT
Outputs
ReRAM Array w/ Peripherals
Neuron Function

Outline

- **Intro and Motivation**
- **ReRAM-Based Accelerator Key Concepts**
- **ReRAM-Based Accelerator Model**
- **Conclusion**

Device to Algorithm Model



What device properties are needed?

Top Down

Neural Algorithm

CrossSim

SST

Neural Algorithm Level Model

CrossSim

McPAT*

NVSIM

Computer Architecture Level Model

CrossSim

Xyce

Circuit Level Models

Electrical Test

Compact Models

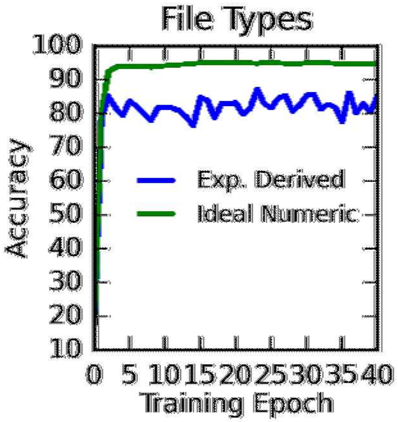
Device Level Models

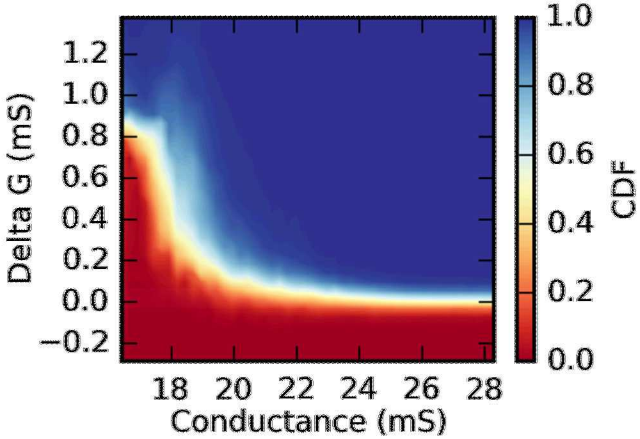
Bottom Up

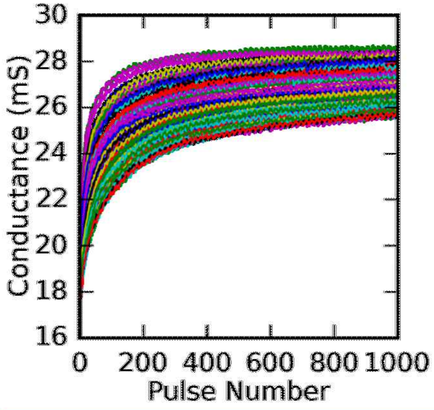
Algorithm Accuracy

Energy and Latency

How do specific devices work in system?



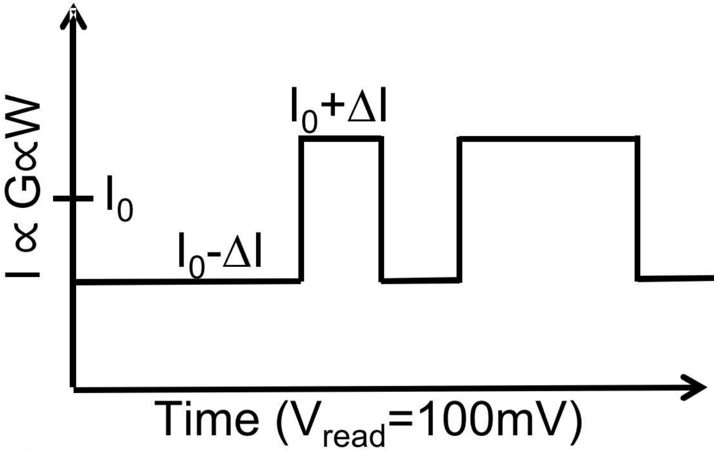




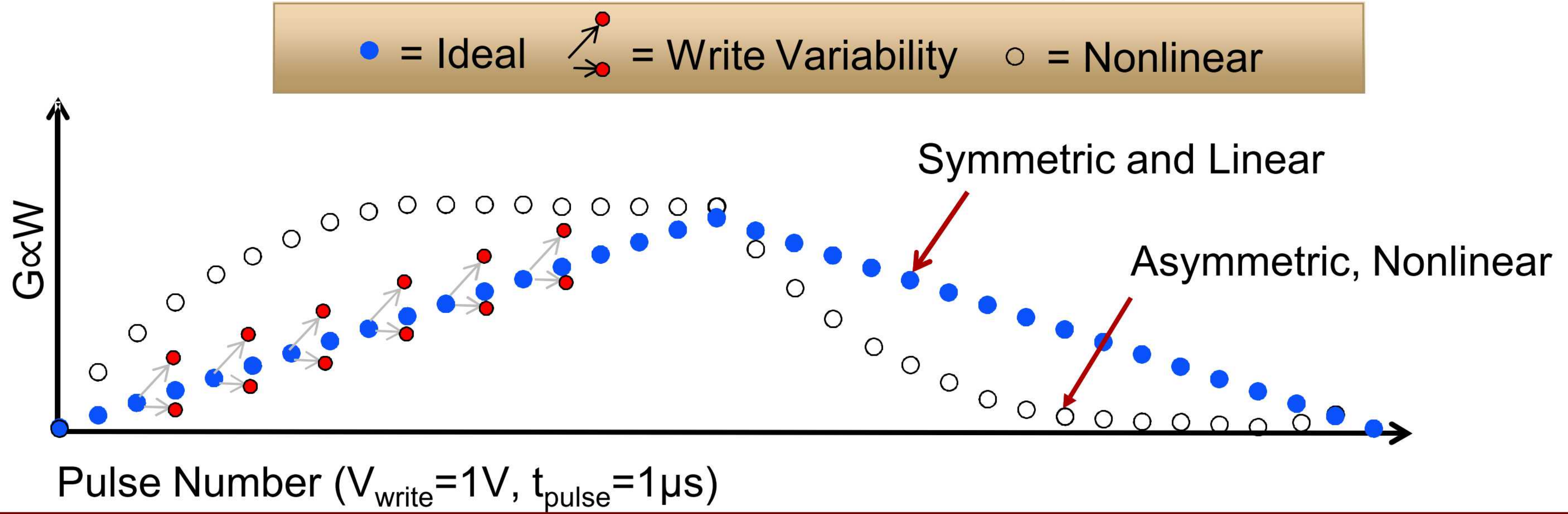
Experimental Device Nonidealities

- **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**

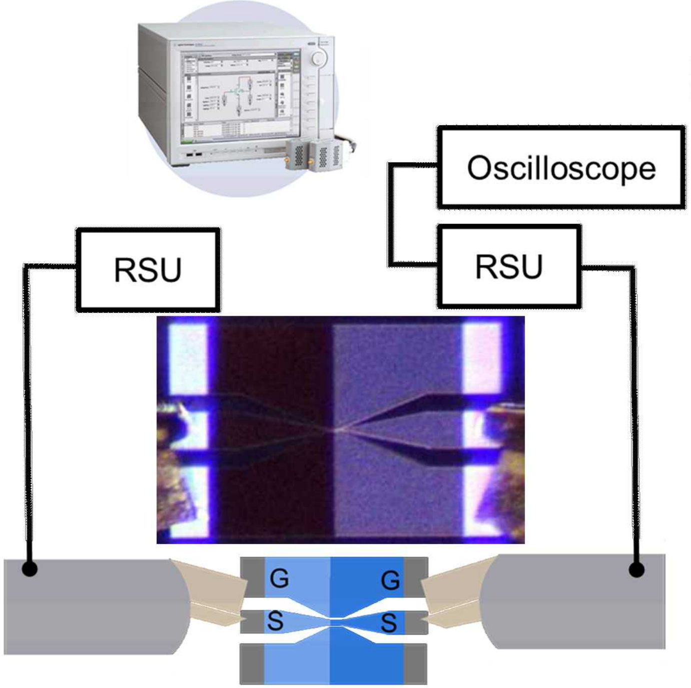
Read Noise



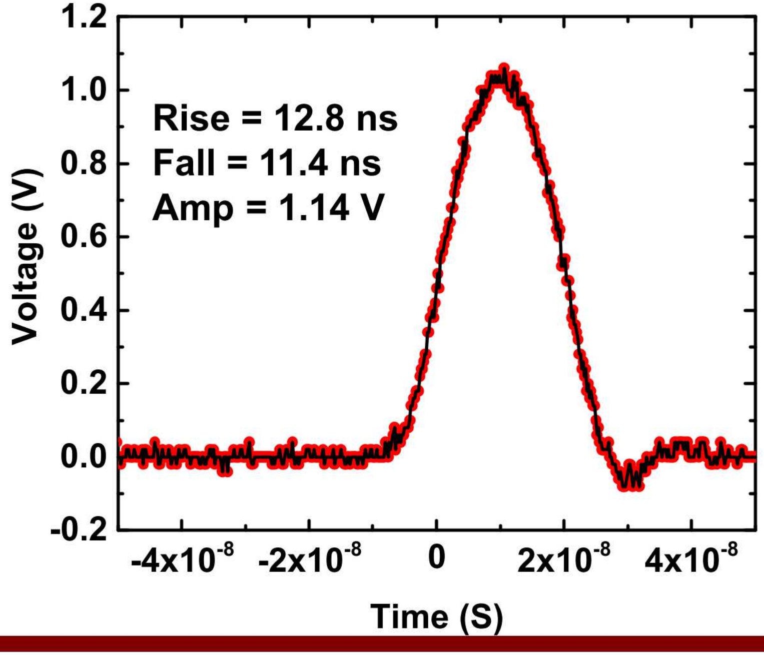
Conductance versus Pulse

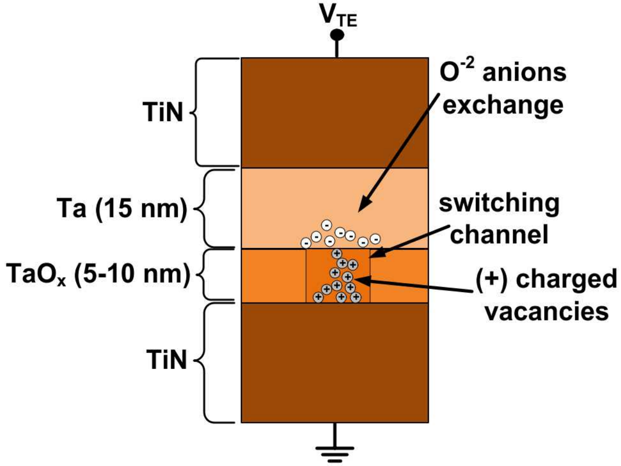


ReRAM Measurements

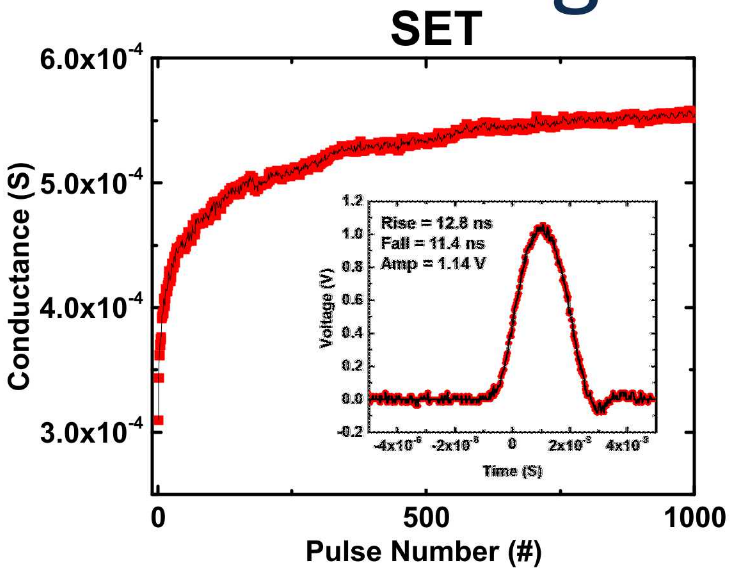


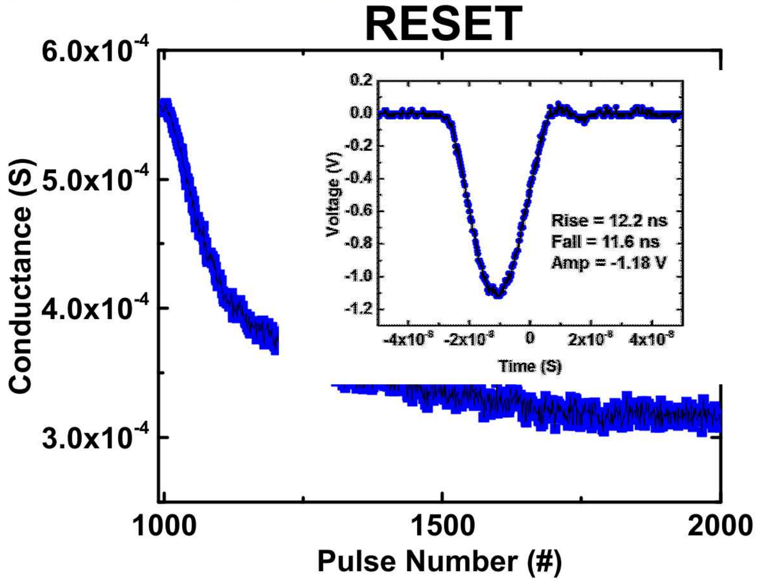
- DC Current-voltage “loops” sweeps are not time-controlled
  - Excessive heating and early wearout
  - Do not provide info on dynamics
- Physical switching < 10ns
- Need pseudo RF setup to measure
  - Ground/signal, conductor backed
  - Agilent B1530 module
  - 10 ns RT/FT, 10 ns PW
  - 1 V nominal,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



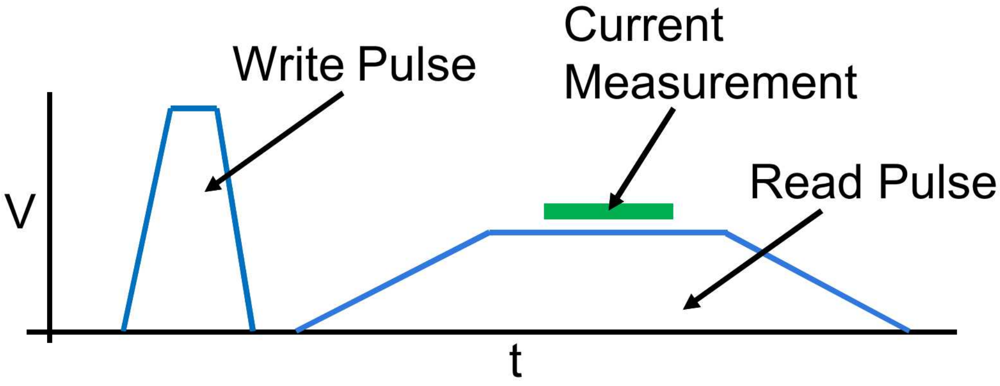


ReRAM Analog Characterization



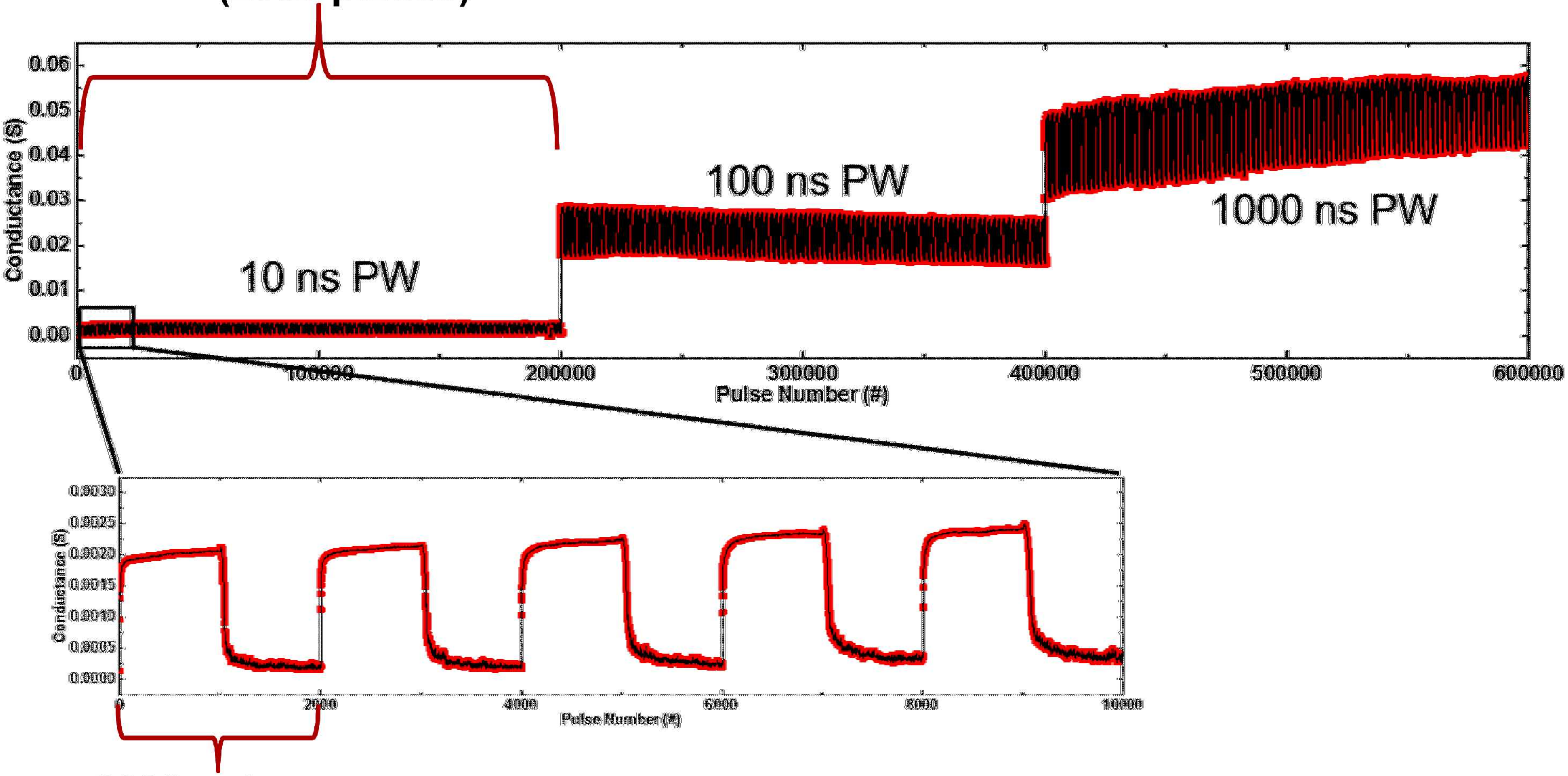


- **Use as a neuromorphic weight requires precise analog tuning**
- **Dataset requires 1000 repeated SET and RESET pulses**
- **Nominal pulse values**
  - **SET: +1V 10ns RT/PW/FT**
  - **RESET: -1V 10ns RT/PW/FT**
  - **READ: 100 mV 1 ms RT/PW/FT**



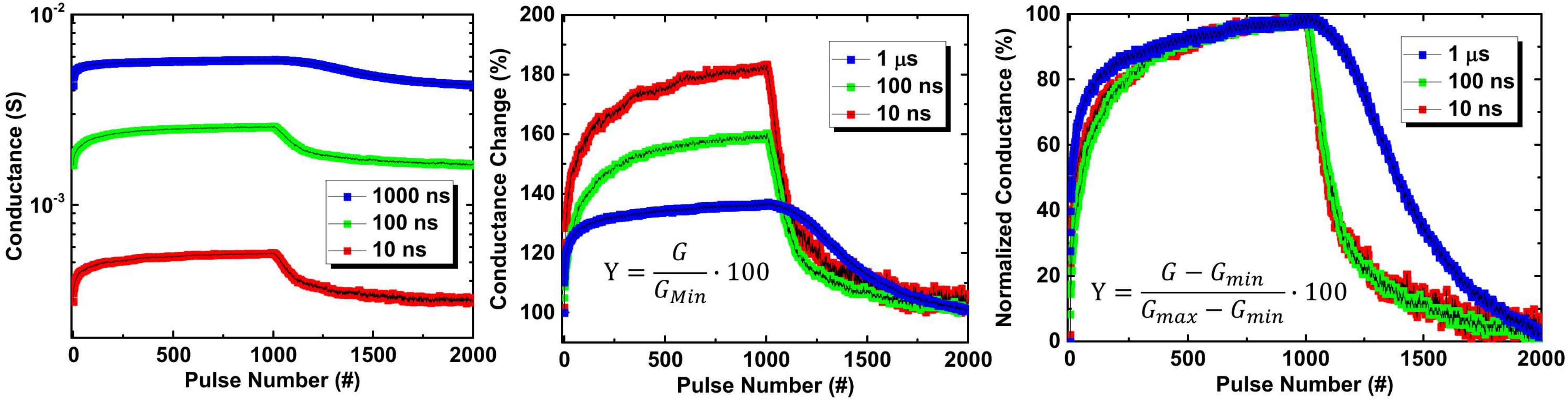
Pulse Width Analog Measurements

100 on→off cycles,
(200k pulses)



2000 pulses per
on→off cycle

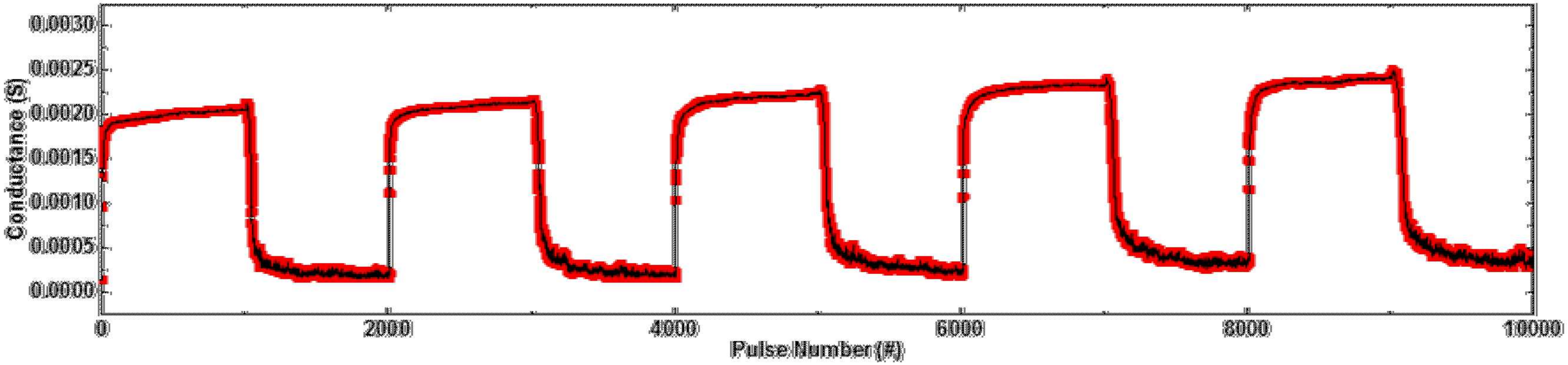
Effect of Pulse Width and Edge Time

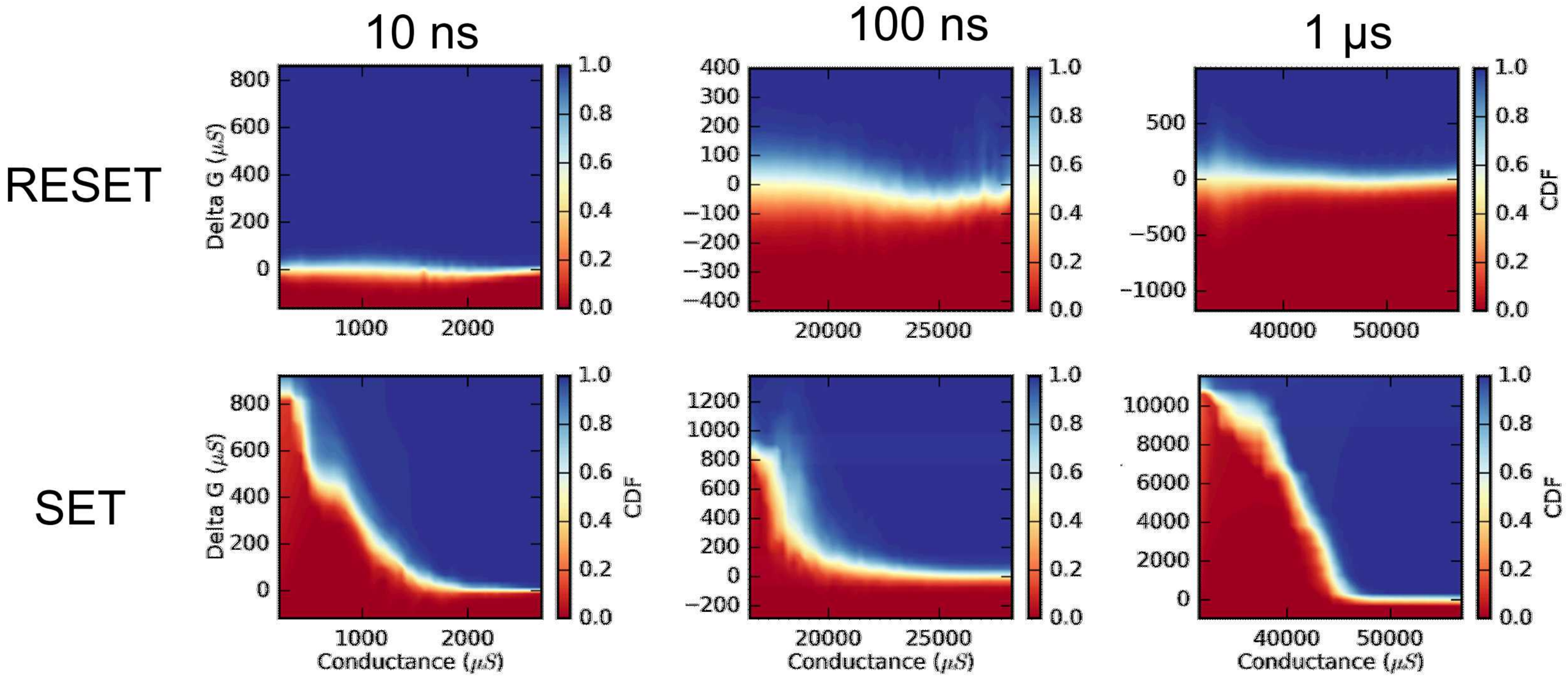


- **Shorter pulses may be employed to lower conductance switching range**
- **Linearity qualitatively similar across Pulse Width (PW) and Edge Time (ET)**
  - **Best for SET at 100 ns**
  - **Best for RESET at 1 us**
- **Relative conductance change increased with shorter Pulse Width / Edge Time**

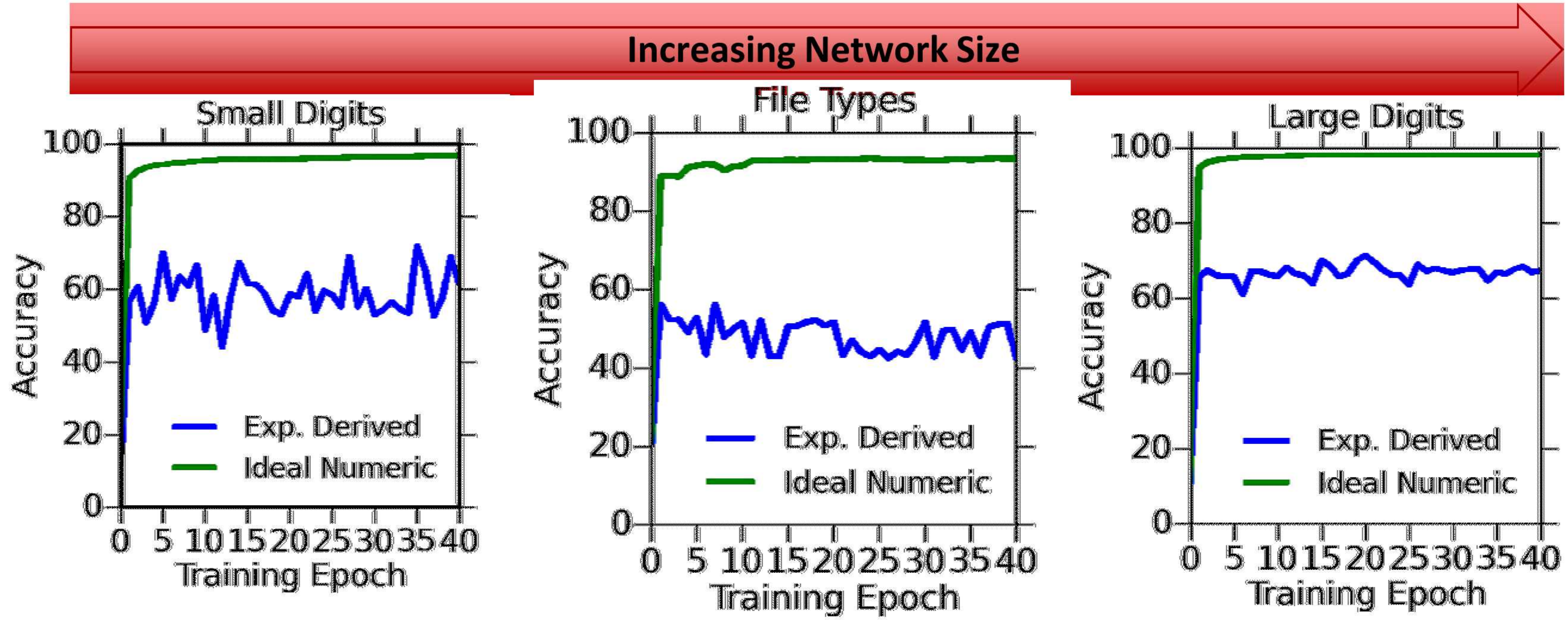
Nominal Pulse Voltage Values: SET: +1 V RESET: -1 V

Repeated Pulsed Cycling





TaOx ReRAM in Backprop Training

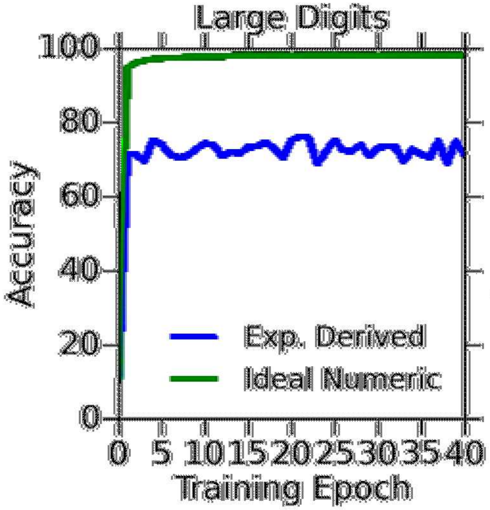


Data set	# Training Examples	# Test Examples	Network Size
UCI Small Digits[1]	3,823	1,797	64×36×10
File Types[2]	4,501	900	256×512×9
MNIST Large Digits[3]	60,000	10,000	784×300×10

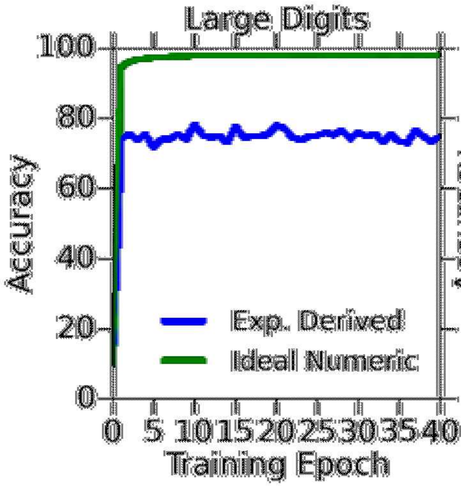
Modeling Effect of Pulse Time

Increasing Network Size

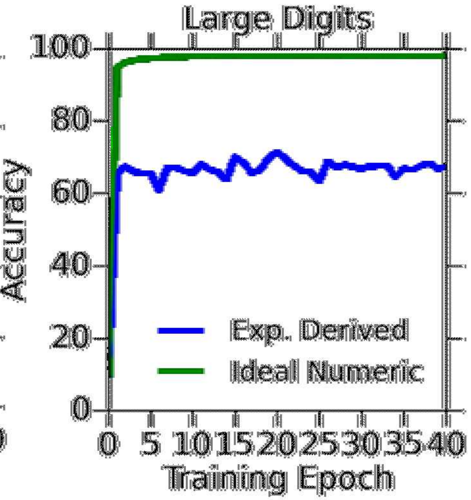
10 ns



100 ns



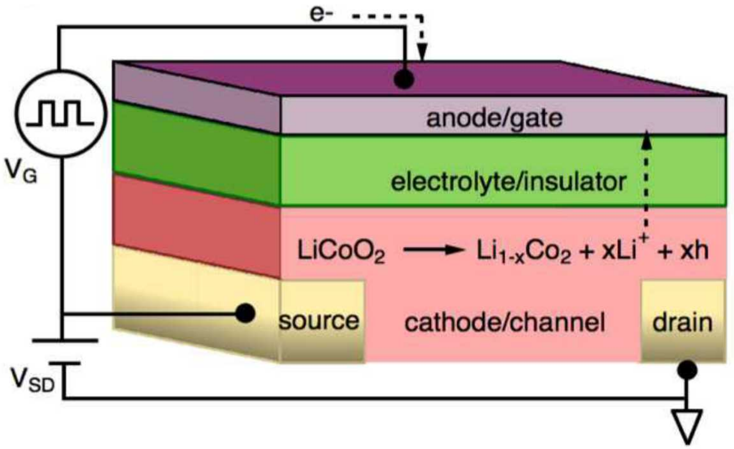
1 µs

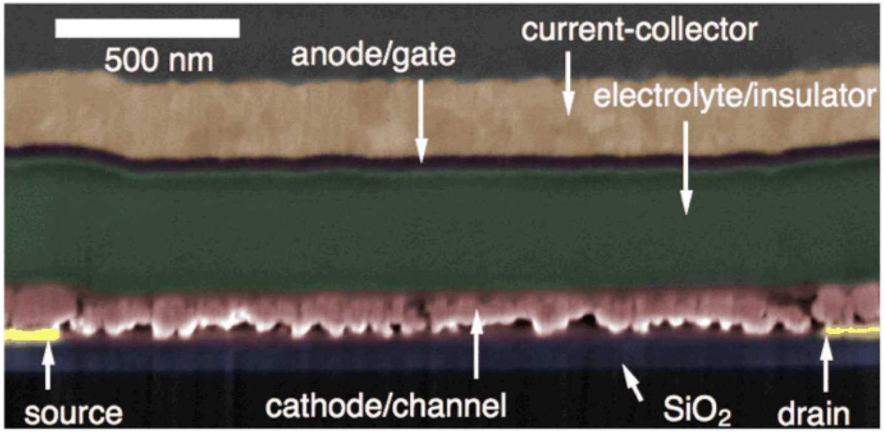


TaOx	Large Images	Small Images	File Types
10 ns	84.45%	71.40%	77.67%
100 ns	78.48%	89.48%	67.78%
1 us	71.48%	71.84%	56.33%

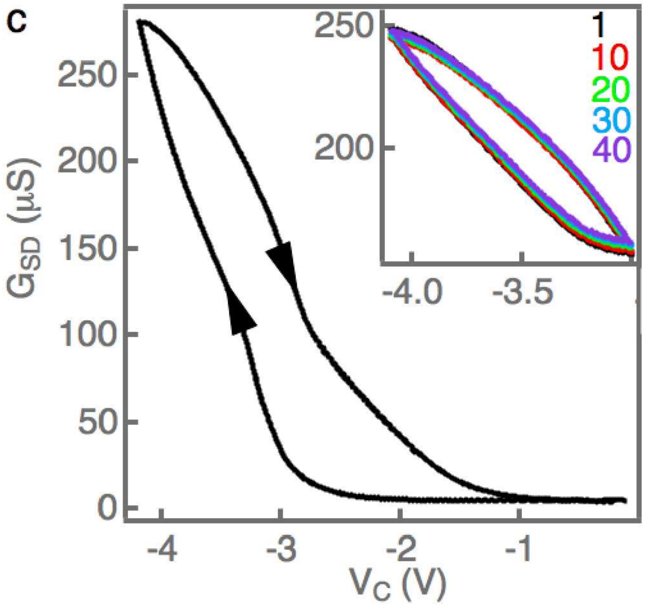
How can training accuracy be improved?

Li-Ion Synaptic Transistor for Analog Computation (LISTA)



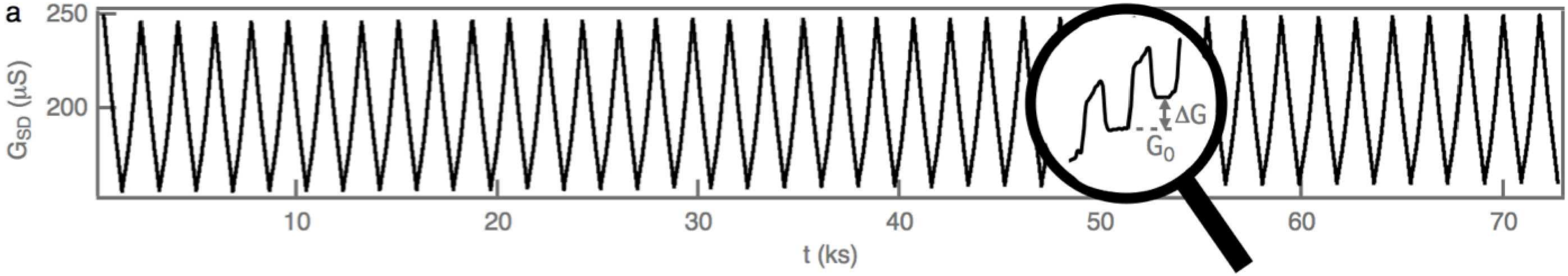


G-V for LISTA Transistor



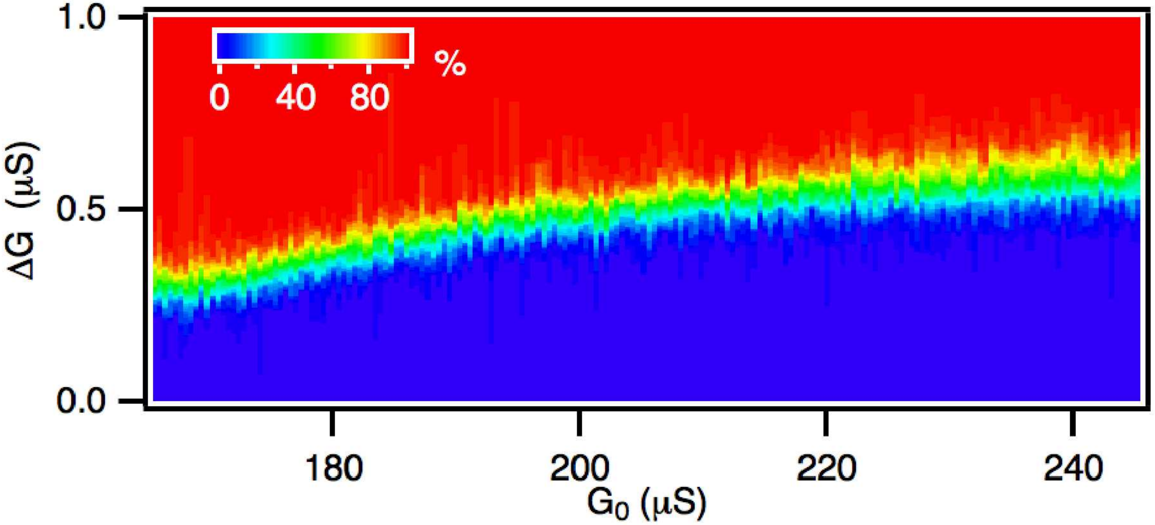
E. Fuller et al, *Adv Mater*, accepted 2017

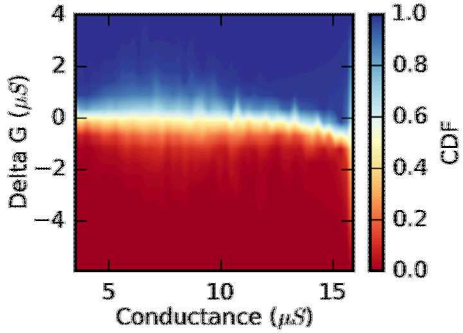
Analog State Characterization



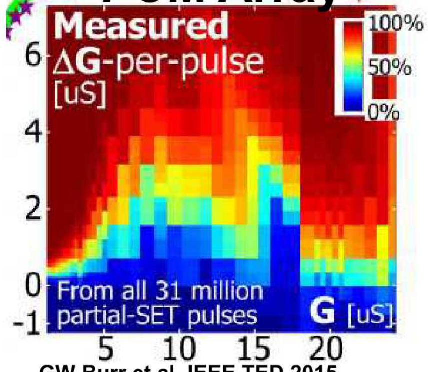
TaOx ReRAM

LISTA > 200 states





PCM Array

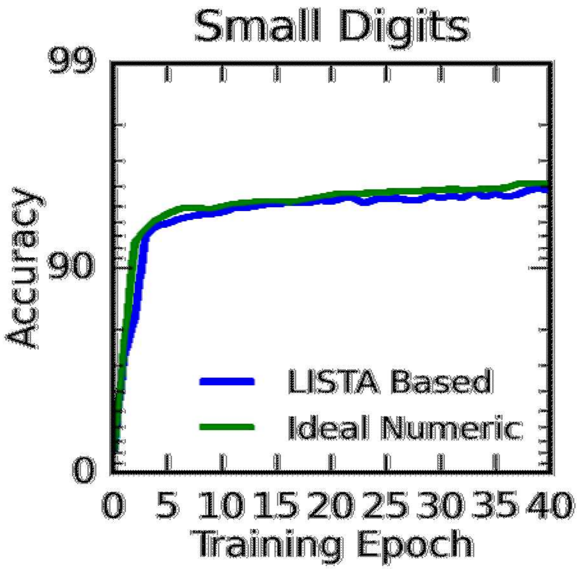


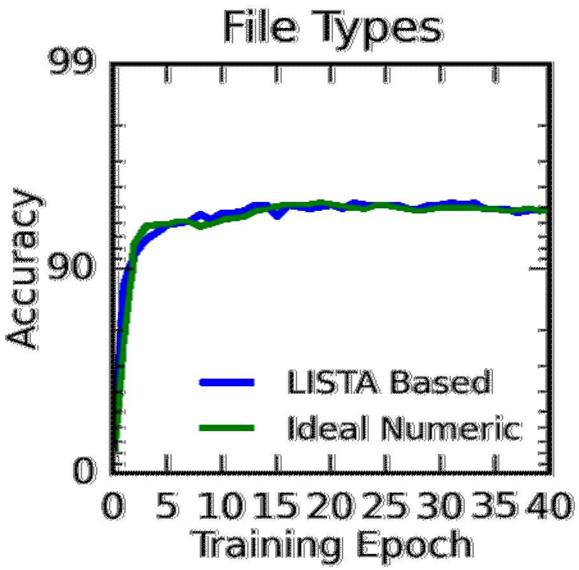
E. Fuller et al, *Adv Mater*, accepted 2017

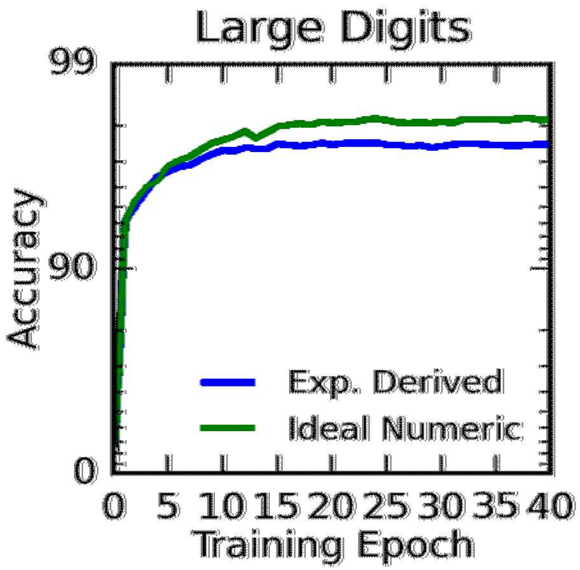
GW Burr et al, IEEE TED 2015

LISTA-device Performance for Backprop Algorithm

Increasing Network Size





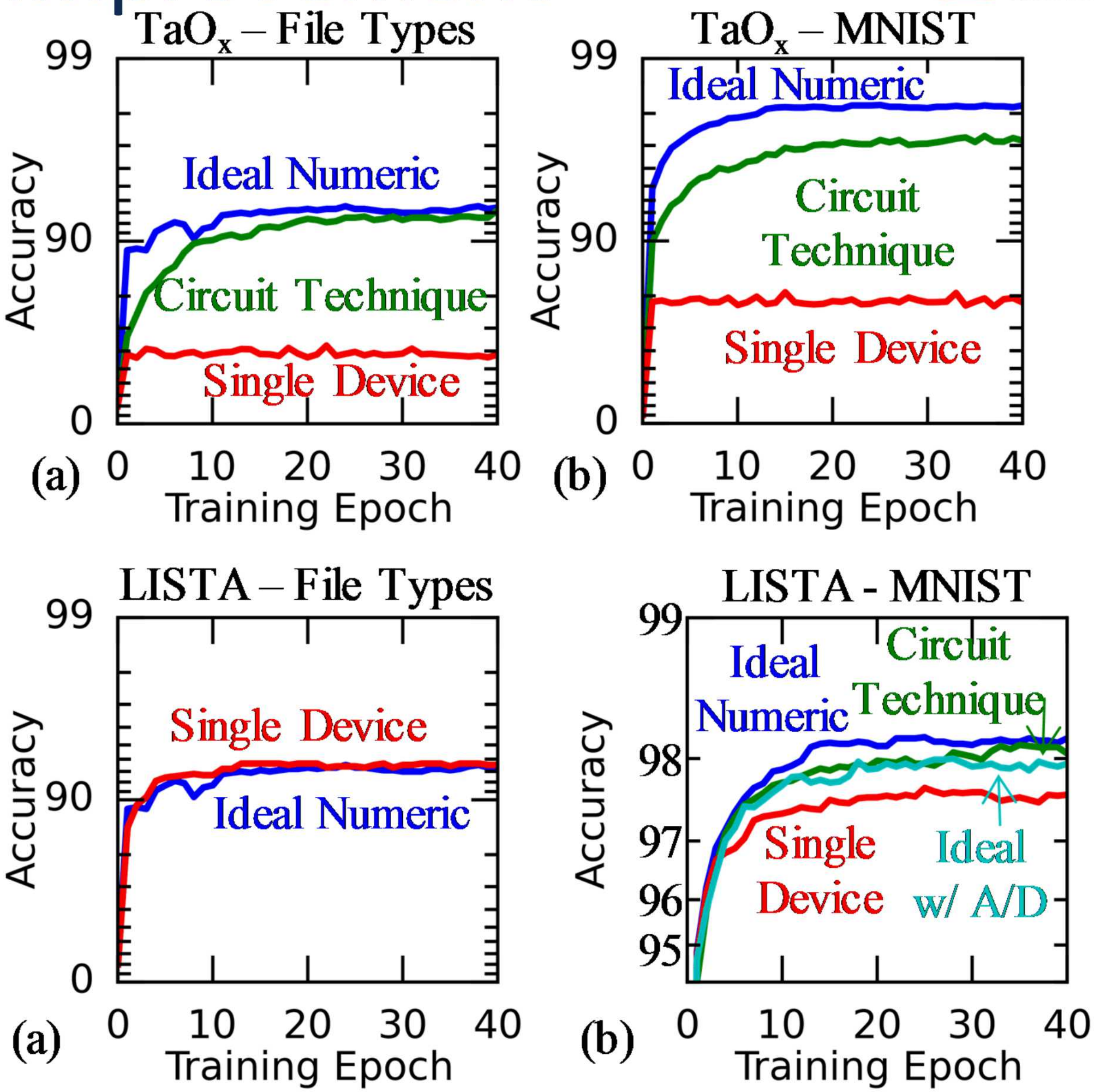


Data set	# Training Examples	# Test Examples	Network Size
UCI Small Digits[1]	3,823	1,797	64×36×10
File Types[2]	4,501	900	256×512×9
MNIST Large Digits[3]	60,000	10,000	784×300×10

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

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 Comparison

Overview		Digital SRAM	Digital ReRAM	Analog ReRAM Crossbar
Equivalent Area ~450 1k× 1k matrices		400 mm^2	32 mm^2	11 mm^2 [64nm pitch]
Total Time *[per cycle]*		~ 100µ s	~ 60µ s	~ 5µ s
Total Energy *[per cycle]*		~ 1000 nJ	~ 700 nJ	~ 15 nJ
Matrix Storage Area		95%	50%	17%
Periphery Area		5%	50%	100% (crossbar is above periphery)
Matrices per 400 mm^2 Chip		~450	~5,500	~15,000
The above figures do not include a SIMD engine or on-chip routing fabric, and are based on a 14nm FinFET process.				

Energy Analysis

Per-Component Breakdown		Digital SRAM	Digital ReRAM	Analog ReRAM Crossbar
Matrix Storage 1024× 1024 Digital: 8 bits/value Analog: 1 cell/value [Values are per-array]	*Area*	**800,000 µm^2**	35,000µ m^2	10,000µ m^2
	Read	30 nJ / 15µ s	15 nJ / 4µ s	~ 3 nJ / ~ 1.5µ s
	Read Transpose	300 nJ / **65 µs**	15 nJ / 4µ s	~ 3 nJ / ~ 1.5µ s
	Write	30 nJ / 15µ s	50 nJ / **45 µs**	~ 3 nJ / ~ 1.5µ s
Multiply Accumulators [256 in parallel]	*Area*	19,000µ m^2		Performed by crossbar
	Run [1M ops]	**200 nJ** / 4µ s		
Output LUT [8 bit→ 16 bit]	*Area*	1,400µ m^2		Uses Digital Methods
	Read [1K values]	1 nJ / 1µ s		
Input/Output Buffers [8 bits]	*Area*	13,000µ m^2		
	Per Run	~ 0.1 nJ		
128 Entry 1024x8 Vector Cache (8 matrices per cache) [Values are per vector]	*Area*	90,000µ m^2	4,000µ m^2	
	Read	~ 0.1 nJ / ~ 0.2µ s	~ 1 nJ / ~ 4 ns	
	Write	~ 0.1 nJ / ~ 0.2µ s	~ 1 nJ / ~ 50 ns	

Digital ReRAM based on output from X. Dong, et. al., *NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory*, in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 31, no. 7, pp. 994-1007, July 2012.

Outline

- Intro and Motivation
- ReRAM-Based Accelerator Key Concepts
- ReRAM-Based Accelerator Model
- Conclusion

Conclusion

- Dennard (constant power density) scaling has ceased and Moore’s law is slowing
- As this slows, a new direction will be needed to achieve the continue the exponential improvements in performance/watt (aka energy efficiency)
- New paradigms like neuromorphic computing will be required for sub-fJ computing
- We now require a device through system design mentality
  - Motivation behind CrossSim
- Oxide-based resistive memory offers intriguing device options for both eras
- Novel lithiated device LISTA and circuit techniques offer significant potential in the development of a low energy neural accelerator

Thank you!

Acknowledgements

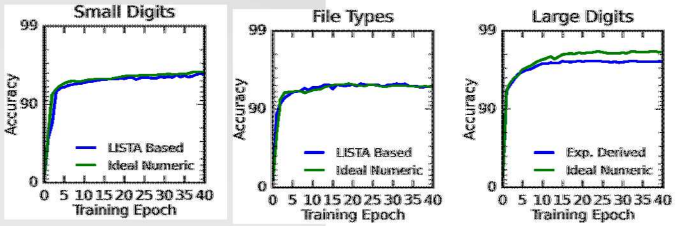
- This work is funded by Sandia's Laboratory Directed Research and Development as part of the Hardware Acceleration of Adaptive Neural Algorithms Grand Challenge Project
- Many shared ideas among collaborators:
  - DOE BIS: John Shalf, Ramamoorthy Ramesh, Patrick Nealeau
  - Dave Mountain, Mark McLean,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

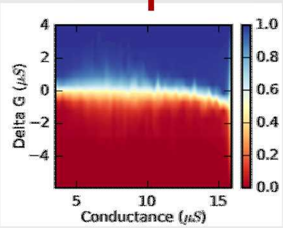
Backup Slides

Energy Analysis

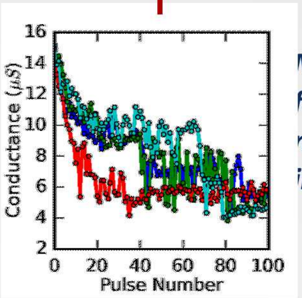
Analog Breakdown Values are per indicated operation	Area	Energy	Latency
Array [1024× 1024]	4,300μ m^2	~ 0.2 nJ read ~ **2 nJ** write	~ 1 ns (propagation)
Temporal Drivers [1024 rows]	460μ m^2	~ 2 pJ read ~ 0.3 nJ write	**1 ns× 2^{bits}**
Voltage Drivers [1024 cols; 16 voltages]	5,000μ m^2	~ 2 pJ read ~ 0.3 nJ write	≤ 1 ns
Integrators/ADCs (reads only)	3,000μ m^2	~ **2 nJ**	**1 ns× 2^{bits}**

Multiscale CoDesign Model: Neuromorphic Crossbar Accelerator

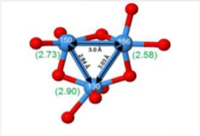




Sandia Cross-Sim: Translates device measurements and crossbar circuits to algorithm-level performance



Memristor fabrication and measurements in MESAFab



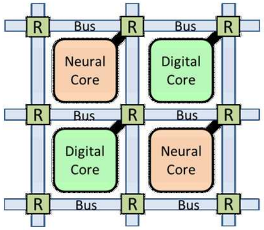
DFT of model of oxide physics, bands

Target Algorithms

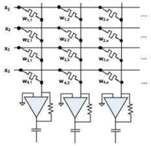
- Deep Learning
- Sparse Coding
- Liquid State Machines

Algorithms

Architecture



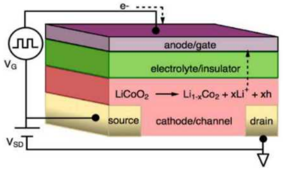
Modified McPAT/CACTI: Model performance and energy requirements



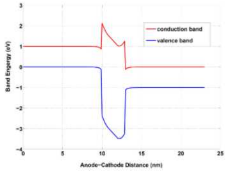
Circuits

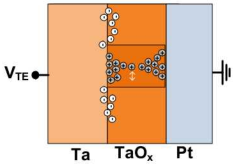
Sandia's Xyce Circuit Sim: Simulate crossbar circuits based on our devices

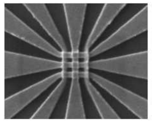
Devices



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

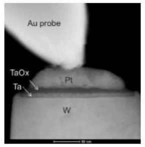


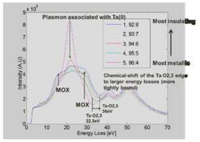


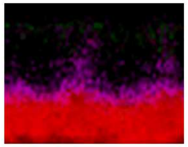


Materials

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







Beyond Moore Co-design Framework

Modeling

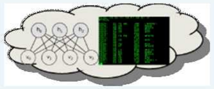
10,000x improvement: 20 fJ per instruction equivalent

Experimental

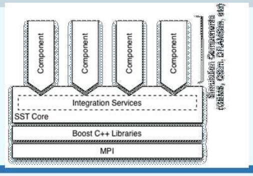
Algorithms → Architectures → Devices → Materials

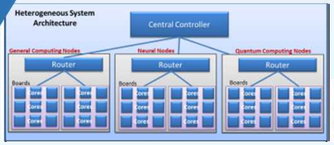
Algorithms & SW Environments

Algorithms and Software Environments
- Application Performance Modeling



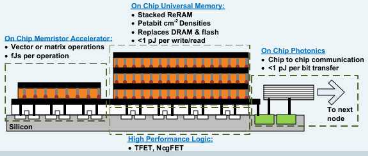
Computer System Architecture Modeling
- Next generation of Structural Simulation Toolkit
- Heterogeneous systems HPC models

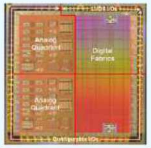




Hardware & Circuit Architectures

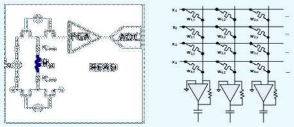
Microarchitecture Models
- McPAT, CACTI, NVSIM, gem5





Component Fabrication
- Processors, ASICs
- Photonics
- Memory

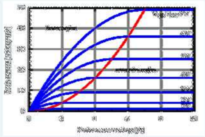
Circuit/IP Block Design and Modeling
- SPICE/Xyce model

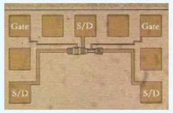


Test Circuit Fab and Measurement
- Subcircuit measurement

Comm., Memory & Computation Devices

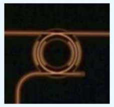
Compact Device Models
- Single device electrical models
- Variability and corner models

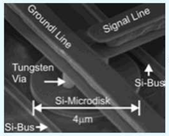


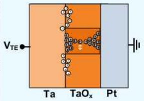


Device Measurements
- Single device electrical behavior
- Parametric variability

Device Physics Modeling
- Device physics modeling (TCAD)
- Electron transport, ion transport
- Magnetic properties



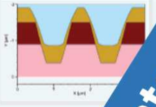


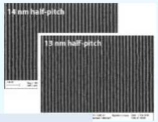


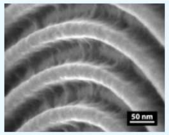
Device Structure Integration and Demonstration
- Novel device structure demonstration

Materials

Process Module Modeling
- Diffusion, etch, implant simulation
- EUV and novel lithography models

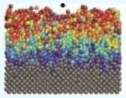




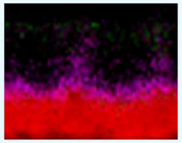


Process Module Demonstrations
- EUV and novel lithography
- Diffusion, etch, implant simulation

Atomistic and Ab-Initio Modeling
- DFT – VASP, Socorro
- MD – LAMMPS



Example activities within a MSCD framework



Fundamental Materials Science
- Understanding Properties/Defects via Electron, Photon, & Scanning Probes
- Novel Materials Synthesis

Integrated Storage Class Memory

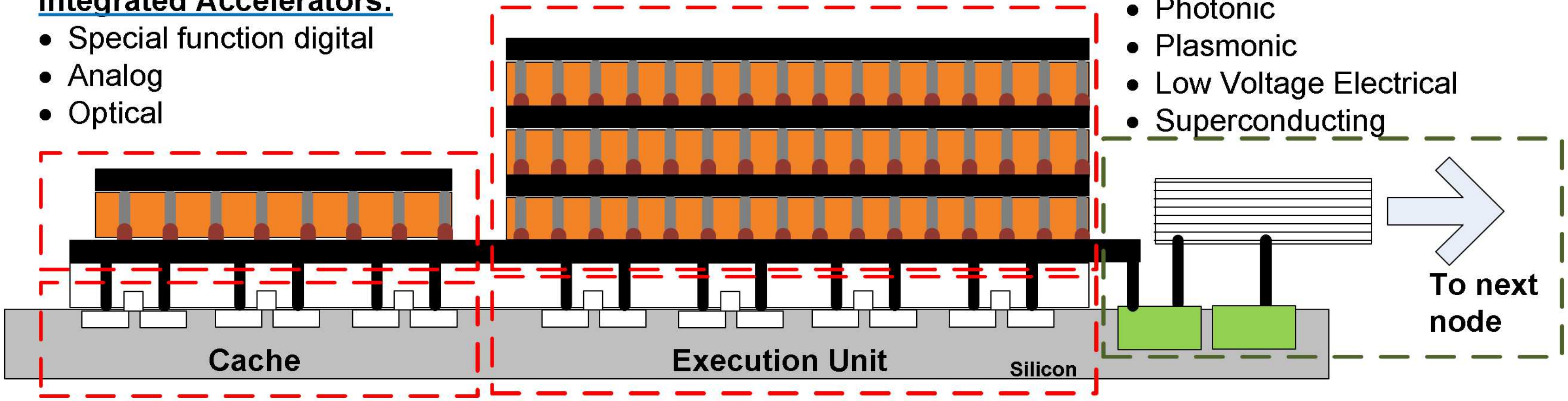
- ReRAM
- STT Magnetic RAM
- CBRAM
- Ferroelectric RAM

Integrated Accelerators:

- Special function digital
- Analog
- Optical

Integrated Communication Devices

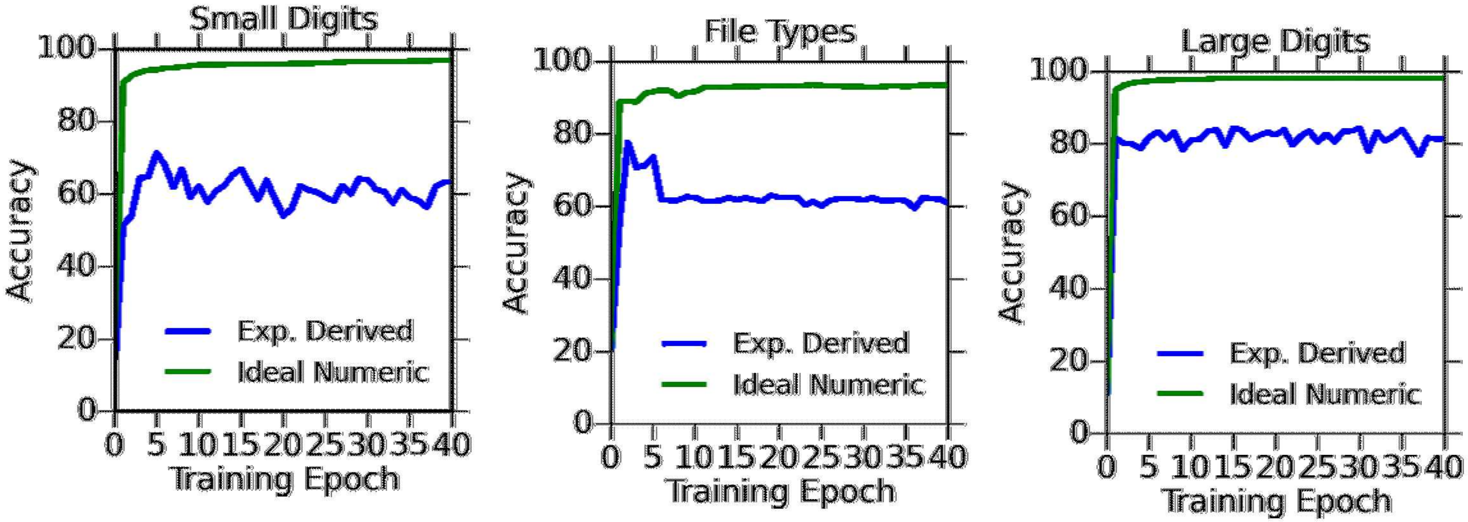
- Photonic
- Plasmonic
- Low Voltage Electrical
- Superconducting



Low voltage high performance logic:

- Tunnel-FET
- Negative Cg FET
- Single Electron Transistor

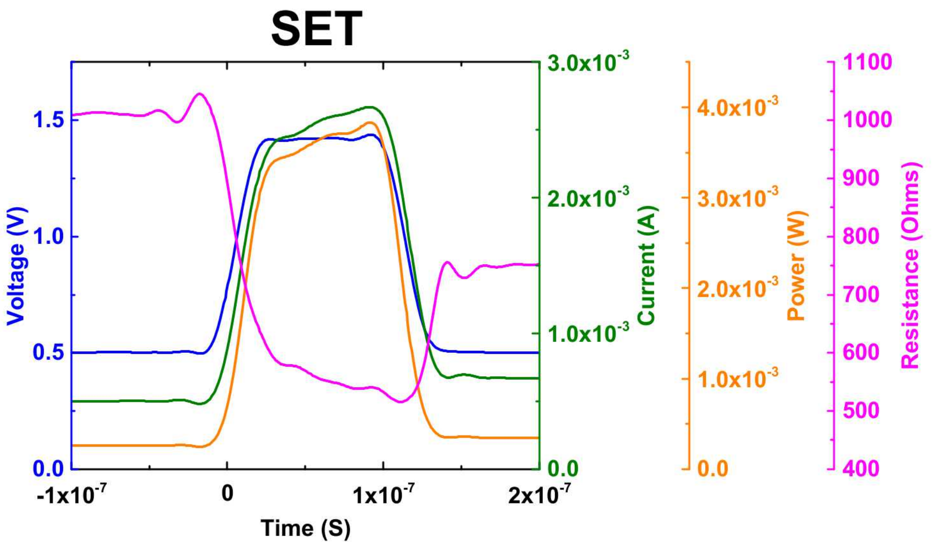
TaOx ReRAM in Backprop Training (10ns)

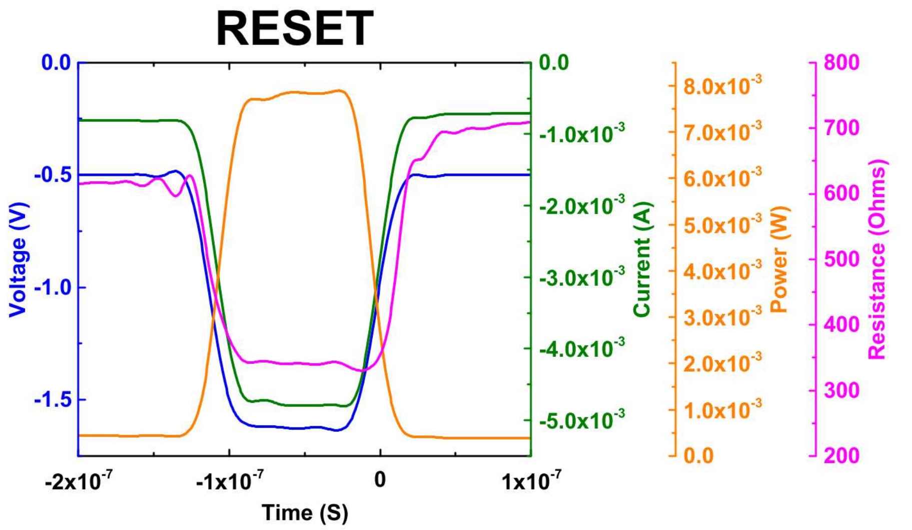


Data set	# Training Examples	# Test Examples	Network Size
UCI Small Digits[1]	3,823	1,797	64×36×10
File Types[2]	4,501	900	256×512×9
MNIST Large Digits[3]	60,000	10,000	784×300×10

How can training accuracy be improved?

Switching Power & Energy Measurement

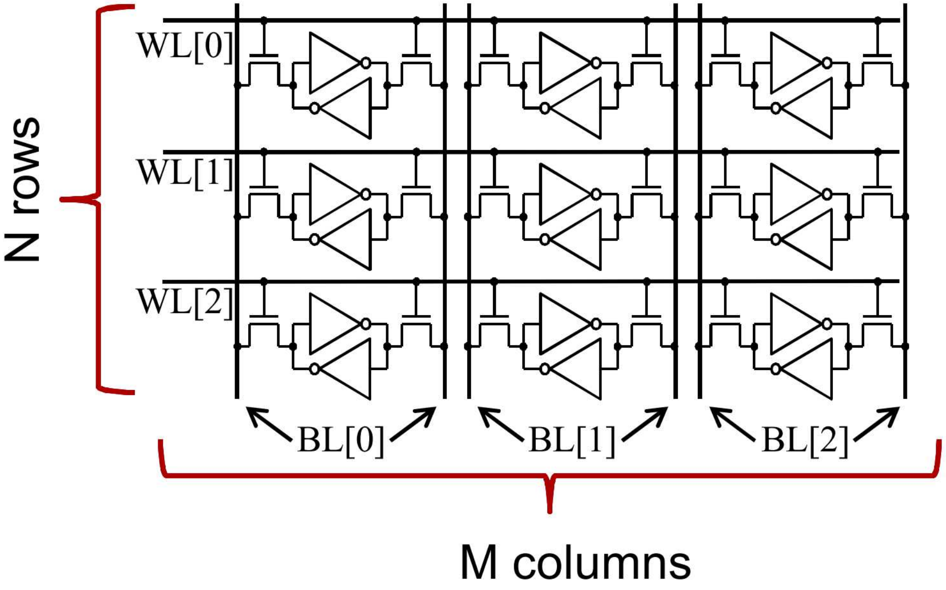




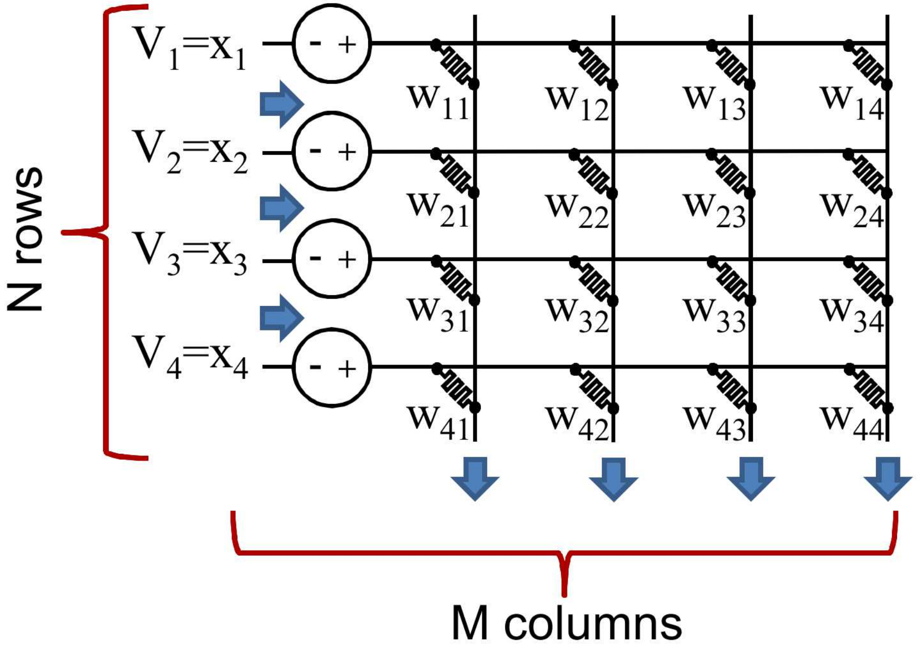
- **Energy determination requires fast pulsed measurements:**
- **Can measure resistance change during pulsed switching with pulsewidths > 100 ns and edgetimes > 10 ns**
- $E = \int_0^t P(t)$
  - **≈800 pJ (RESET)**
  - **≈400 pJ (SET)**
- **Wasted power/energy past first ~1ns of pulse**
- **Lower energy with high resistance devices, sub-ns pulse**
  - **> 1pJ demonstrated @ <1ns in similar TaOx device (by HP)**

Theoretical Efficiency Analysis

SRAM crossbar:



ReRAM crossbar:

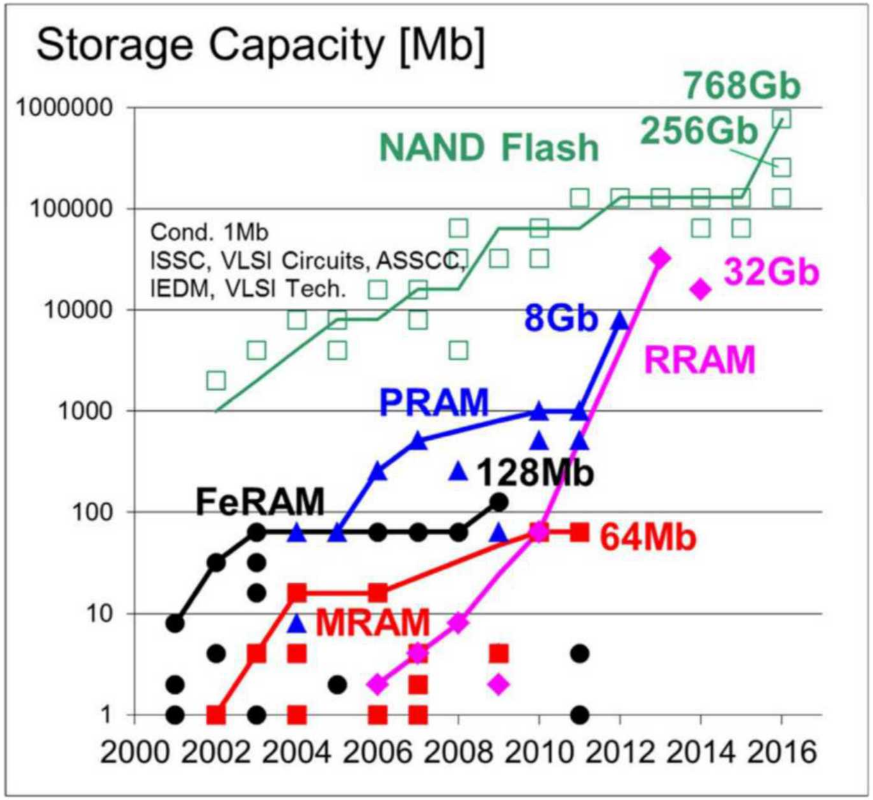


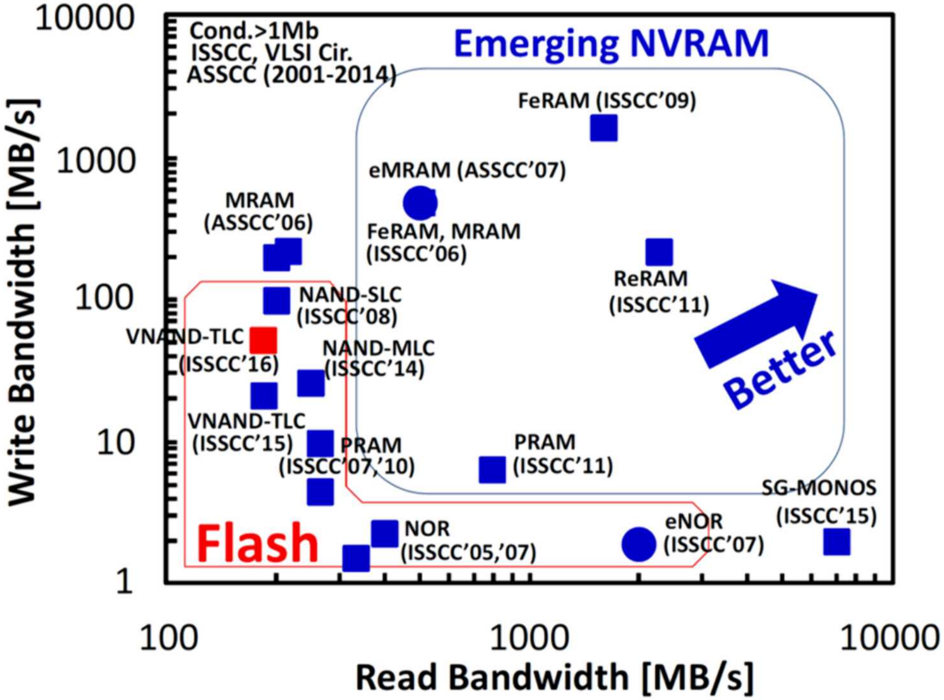
SRAMs must be read one row at a time
→charges M columns;
E = N Rows x *O(N)* wire length x M Columns
~ *$O(N^2 \times M)$*

Energy to charge the crossbar is CV^2;
E ∝ C ∝ number of RRAMs ∝ N×M
~ *O(N×M)*

Implication: Crossbar is *O(N)* better than SRAM in energy consumption for vector-matrix multiply computations

Technological Considerations: Trends

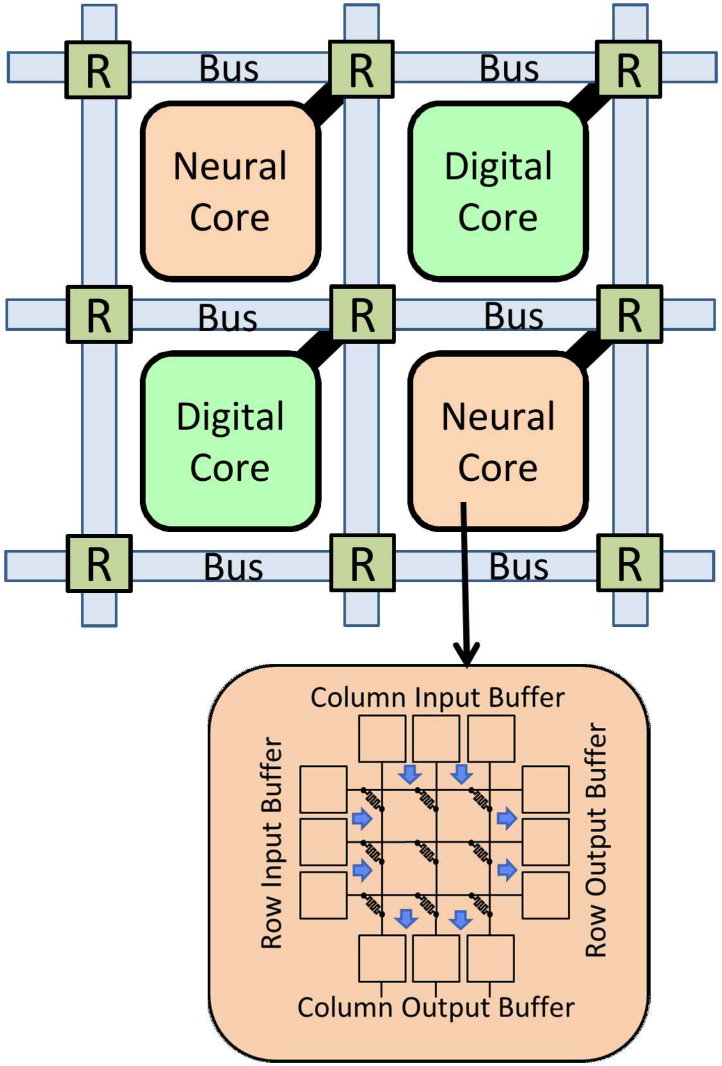




ISSCC 2016 Trends Report

HAANA Crossbar Accelerator Design

- **Initial work by several groups indicates order of magnitude energy efficiency gains are possible using a ReRAM accelerator**
- **The assumptions and outcomes of these models vary significantly**
- **HAANA goal: develop a Multiscale CoDesign Framework which can evaluate our neural crossbar accelerator algorithms, architectures, and devices on a "level playing field"**
- **Evaluate architectures and devices for accuracy, energy, perf.**
- **Once a clear energy advantage demonstrated, move forward with technology development**



How can we get to fJ computing?

Description	**NPU-1**	**NPU-2**	**NPU-3**	**TrueNorth**
System clock frequency	100kHz	1 MHz	10 MHz	1 kHz
Synapses per neuron	500	500	500	256
Average energy per device update	1 fJ	1 fJ	10 aJ	26 pJ
Energy per update op cycle (per core)	250pJ	250pJ	2.5pJ	
Operations per second (per core)	250 GOPs	250 GOPs	250 GOPs	
Single core max power	25 uW	250 uW	25 uW	
Chip Area	4 cm^2	4 cm^2	4 cm^2	4.3 cm^2
Cores per layer	800 k	800 k	800 k	4 k
Layers per chip	10	100	10	1
Neurons per chip	4 B	200 B	4 B	1 M
Chip Max Power	200 W	**10 kW**	200 W	70 mW
Chip Max operations per second	**0.2 ExaMACS**	**10 ExaMACS**	**20 ExaMACS**	28 GigaOps
Operations per second per watt	**10^{15} MACS/W**	**10^{15} MACS/W**	**10^{17} MACS/W**	4x10^{11}Ops/W

MACS = Multiply Accumulate per Second

How do we get to 10 fJ per inst?

- CMOS scaling not providing significant energy efficiency gains
- Many algorithmic, architectural, and device answers:
  - Neuromorphic algorithms
  - Analog accelerators
  - mV switch (e.g. TFET, NgcFET)
  - Superconducting electronics, quantum computing…
- Which horse should we bet on??
- Well…studies for each approach "prove" each respective option to be the best path forward
- Winner not yet clear, most will require major development efforts to realize full potential ($$)
- Need systematic, universal method to determine best approaches for further investment…