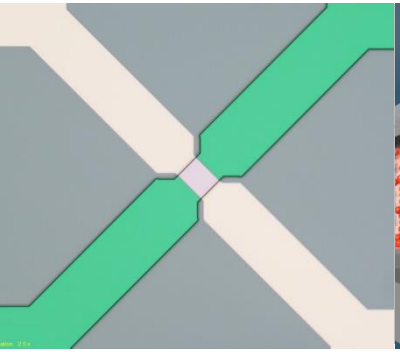
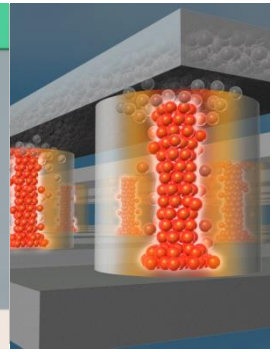
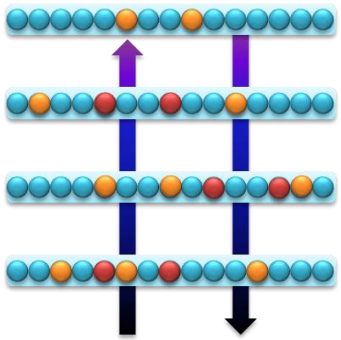
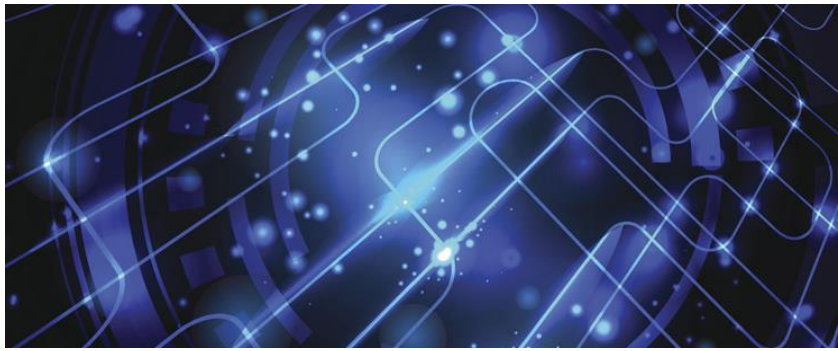
SAND2016-3965C

Exceptional service in the national interest

Sandia National Laboratories

Acceleration of Neural Algorithms using Emerging Nanoelectronic Devices

Matthew J. Marinella*, Sapan Agarwal, David Hughart, Steve Plimpton, Ojas Parekh, Tu-Thach Quach, Erik Debenedictis, Ron Goeke, Alex Hsia, Brad Aimone, Conrad James
Sandia National Laboratories
*matthew.marinella@sandia.gov

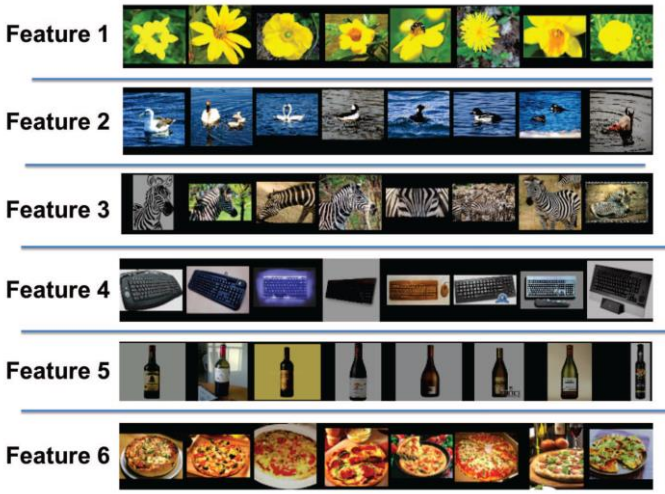
U.S. DEPARTMENT OF ENERGY | NNSA National Nuclear Security Administration | LDRD LABORATORY DIRECTED RESEARCH & DEVELOPMENT

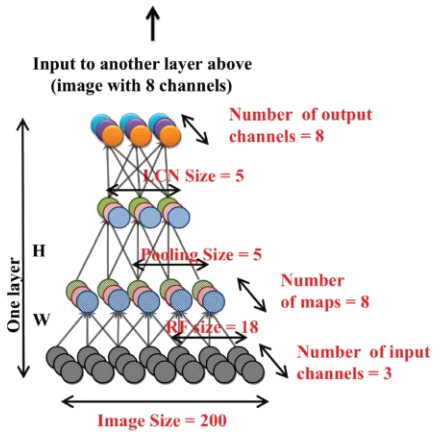
Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Why is Better Hardware Needed?

Sandia National Laboratories

- **Google Deep Learning Study**
  - 16000 core, 1000 machine GPU cluster
  - Trained on 10 million 200x200 pixel images
  - Required 3 days
  - Training set size set by what can be completed in less than one week, with available hardware
- **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 ZetaFLOPs to train in 3 days! (ZetaFLOPs=10^{15} FLOPs; ~5 billion modern GPU cores)*
  - Data is increasing exponentially with time
- **Solution: 10^{18} (Exa) Op-per-second on one chip**





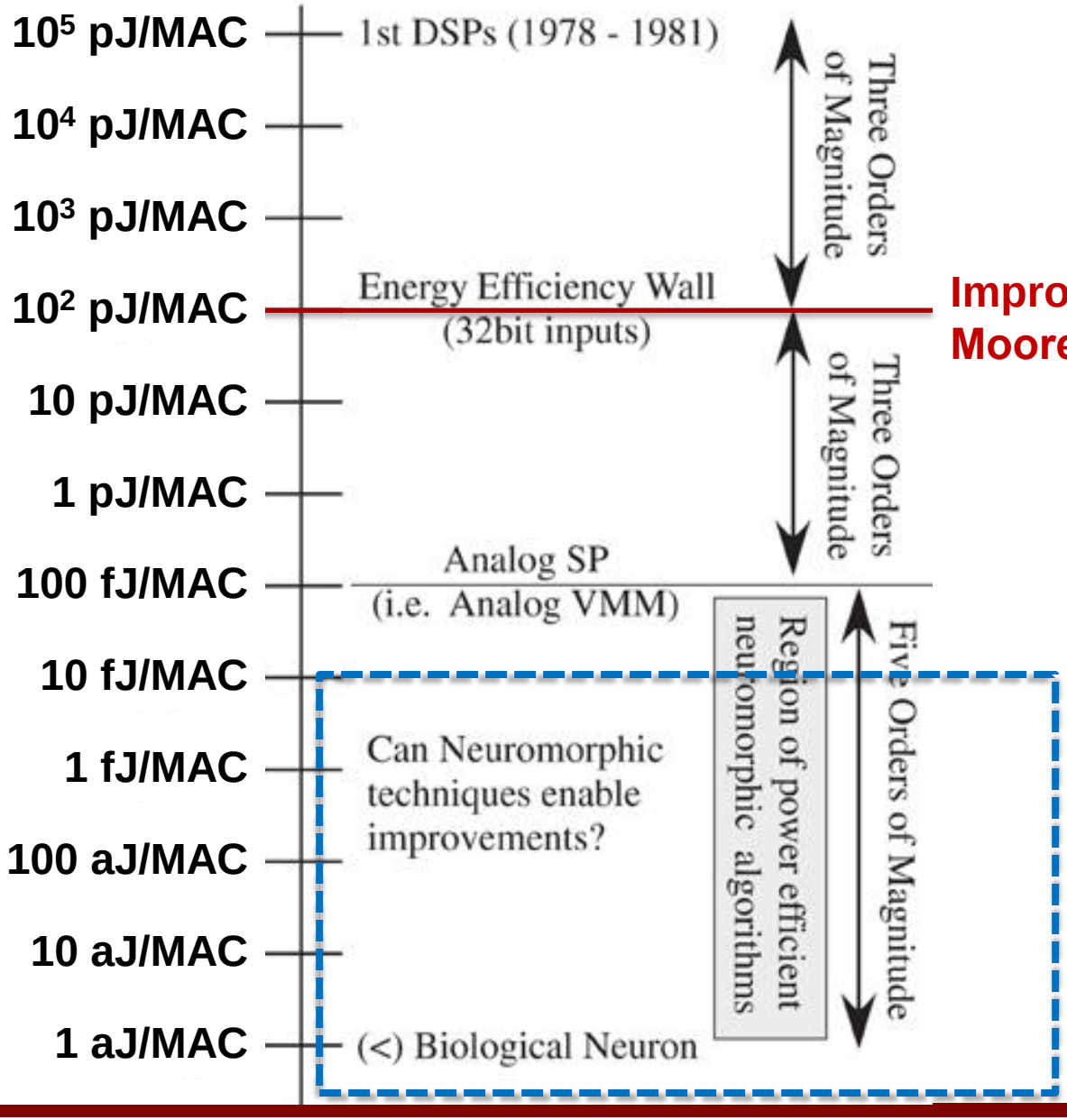
Q. Le, IEEE ICASSP 2013

Neural Processor Unit Estimated Final Design Specs; Comparison to Extreme Cases and TrueNorth

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

Power Efficiency Scaling

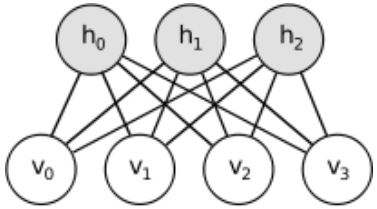


Modified from Hasler and Marr, Frontiers in Neuroscience, 2013

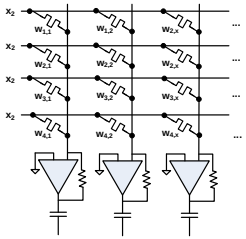
Improvements Due to Moore's Law

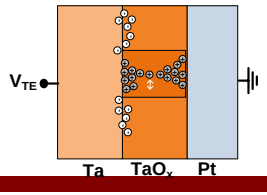
"Let physics do the computation" **Our brain is the ultimate example of this paradigm**

Neuromorphic Hardware Philosophy



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





Application

Algorithm

Computation Kernels

Architecture and Circuits

Device

Ex. Image/speech/pattern recognition, anomaly detection

Traditional: Backpropagation
Emerging: HTM, RBM, recurrent networks, deep learning

Traditional: Vector-matrix multiply, weight update
Emerging: STDP

Traditional: GPU, CPU
Emerging: Crossbar array, coupled oscillators

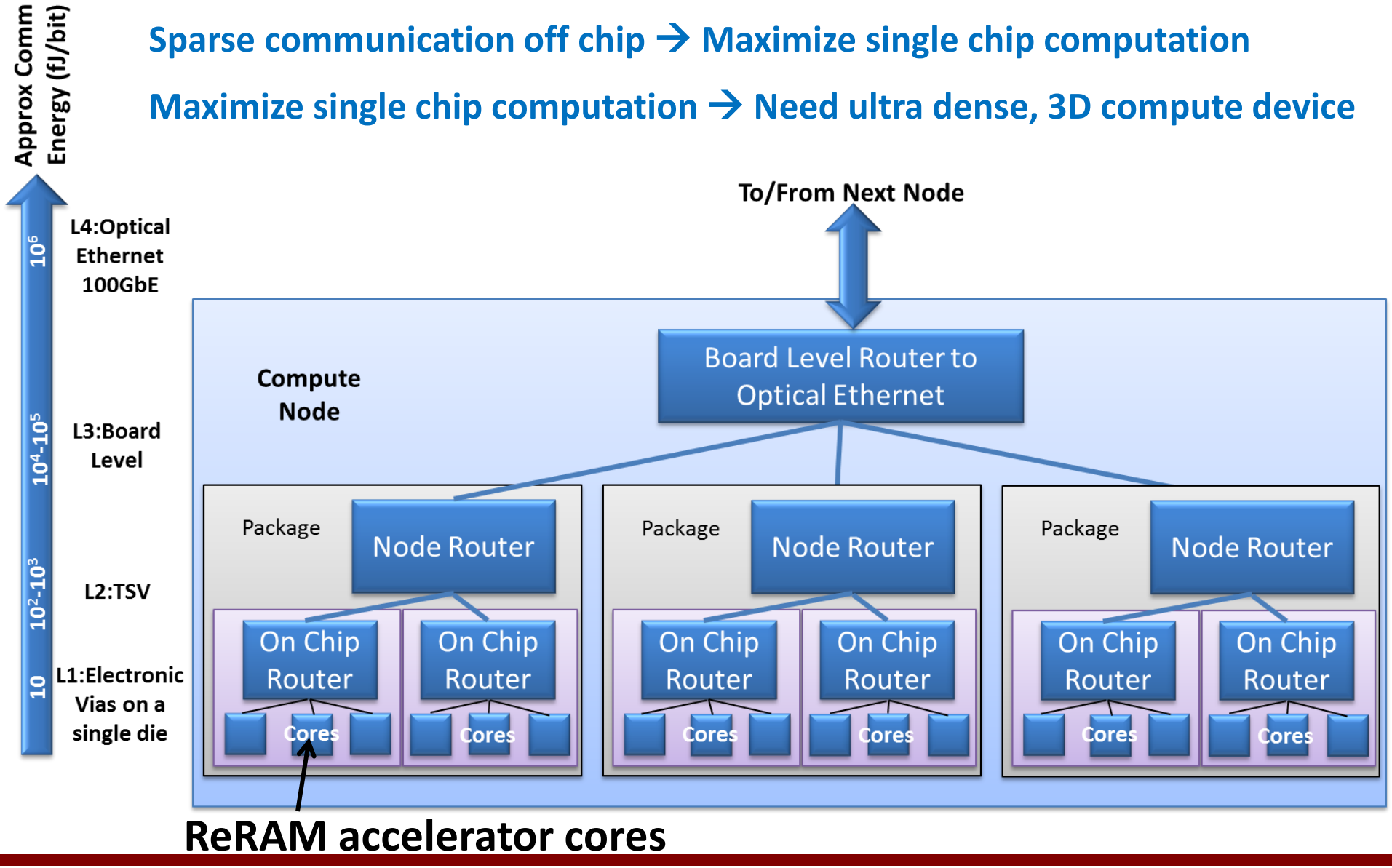
Traditional: FET
Emerging: ReRAM, STT/MeRAM, FeFET

Communication Energy Analysis

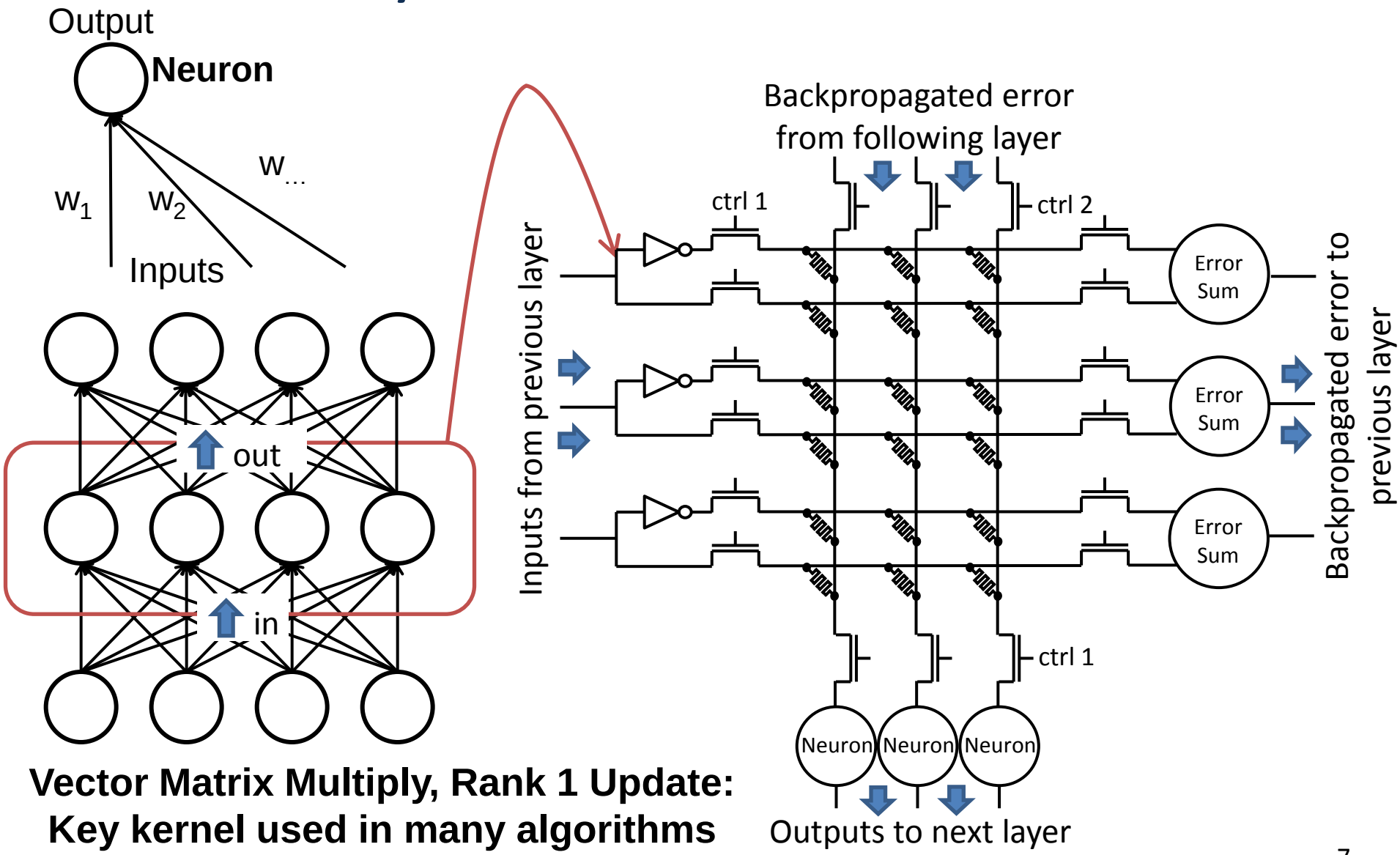
Sandia National Laboratories

Sparse communication off chip → Maximize single chip computation

Maximize single chip computation → Need ultra dense, 3D compute device

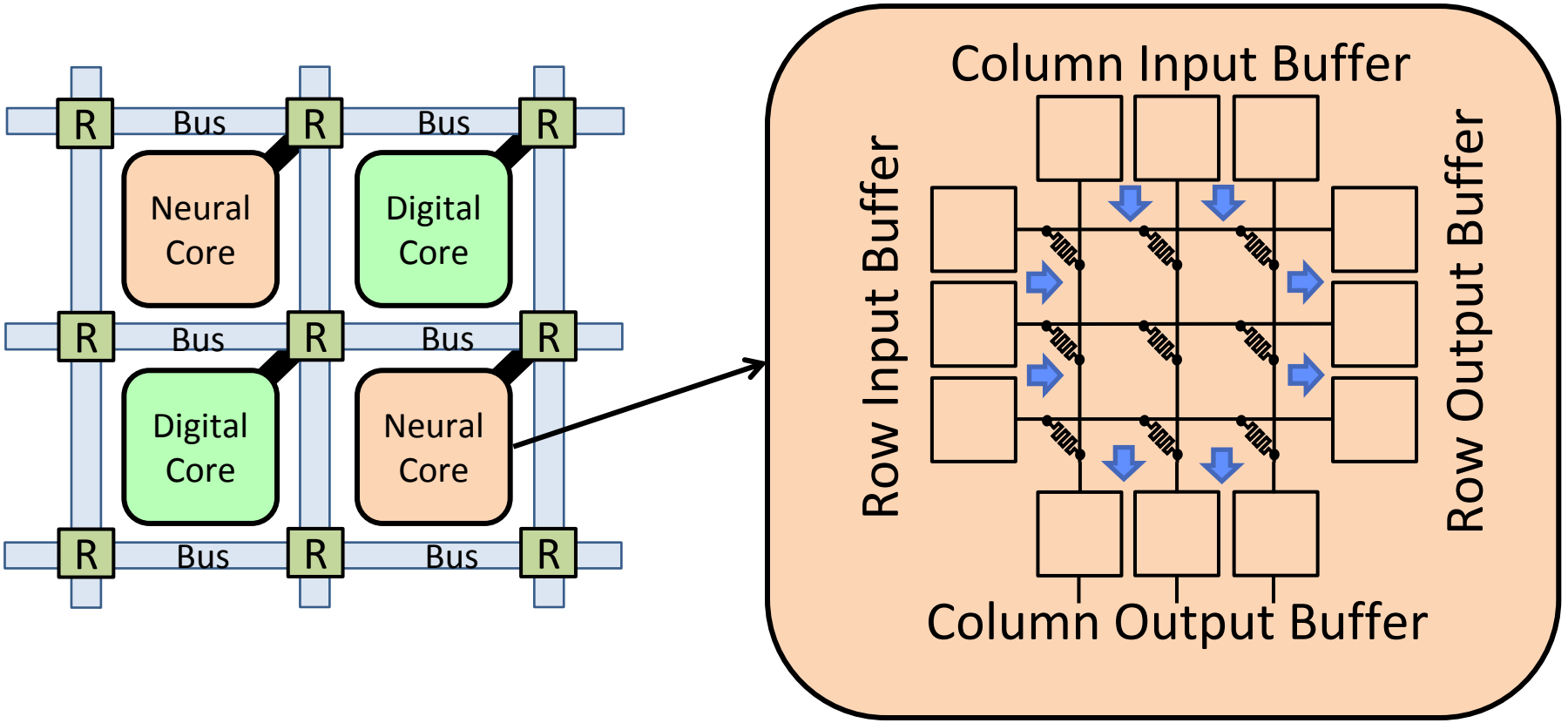


Case Study: Resistive Crossbar



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

Neural Accelerator Architecture

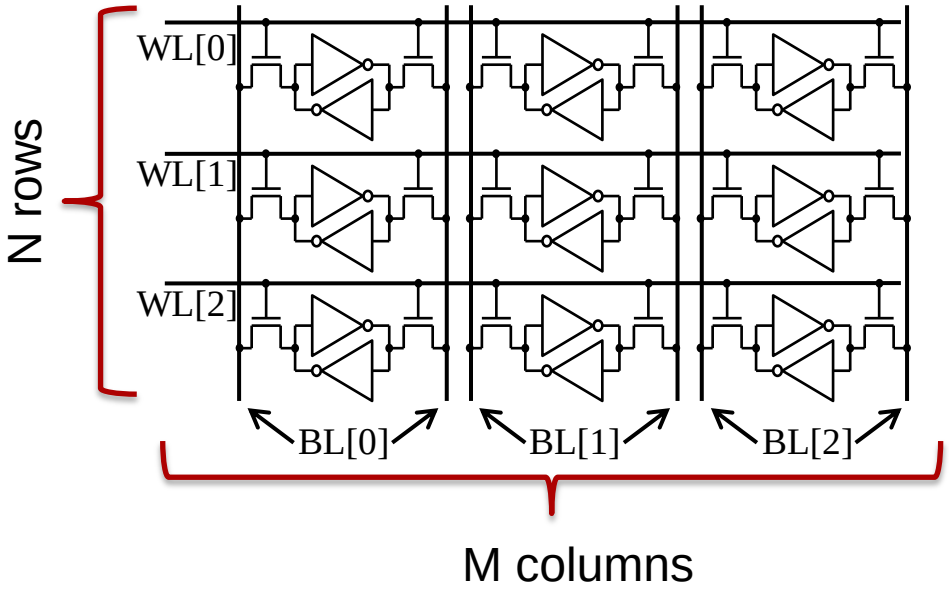


Run *any* neural algorithm on the same hardware

Computation Energy Analysis

Sandia National Laboratories

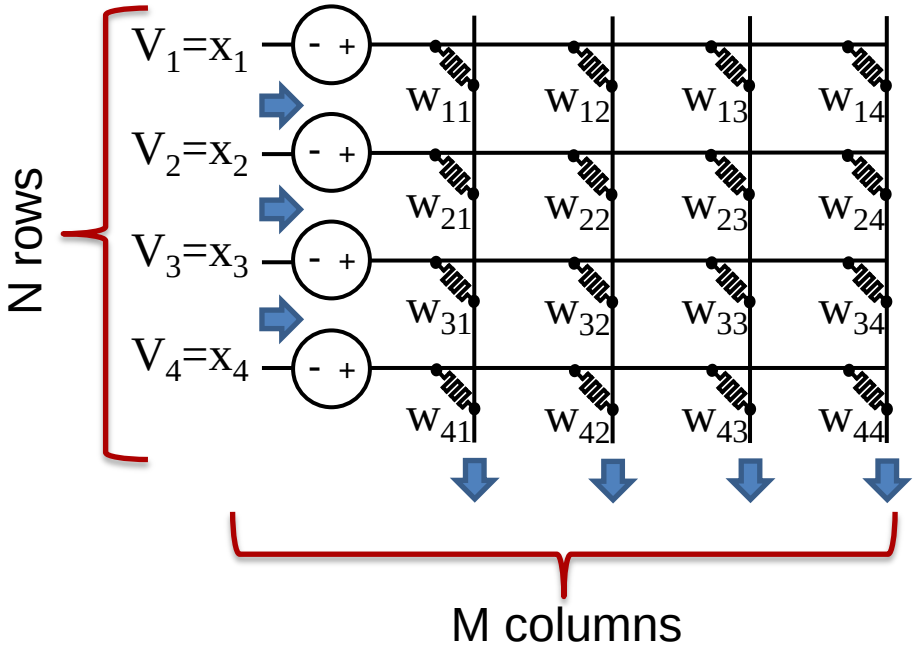
SRAM crossbar:



SRAMs must be read one row at a time, charging M columns;

E = N Rows × M Columns × *O(N)* wire length
~ ***O(N²×M)***

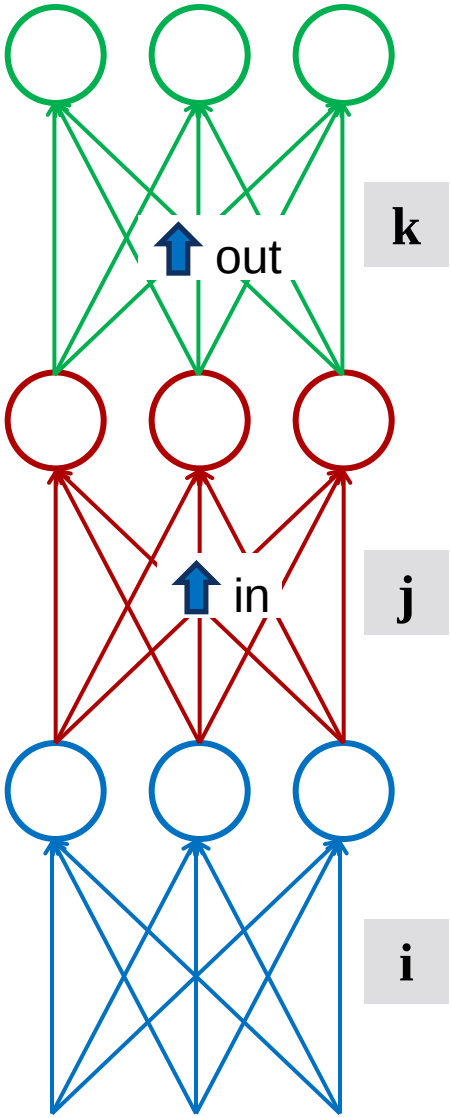
ReRAM crossbar:



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

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

Analog Core: Forward Propagation



O(N^2) Operations

O(N) Operations

y_j

Digital Core

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

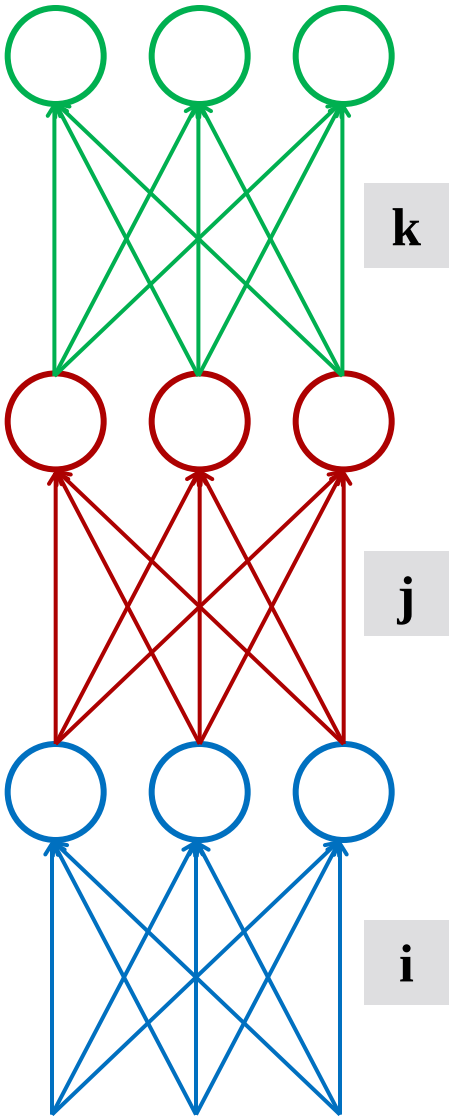
Neuron Function

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(N^2) Read Operations

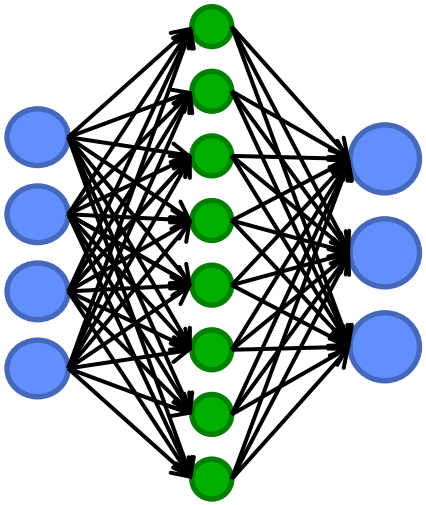
O(N) Operations

O(N^2) Write Operations

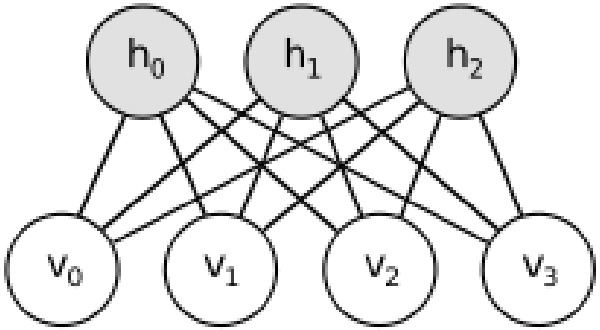
Architecture can Accelerate Many Different Neural Algorithms

Sandia National Laboratories

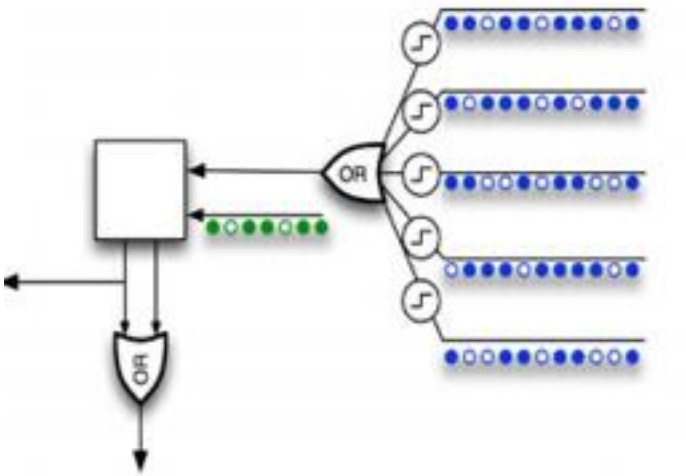
Backpropagation



Restricted Boltzmann Machines



Hierarchical Temporal Memory

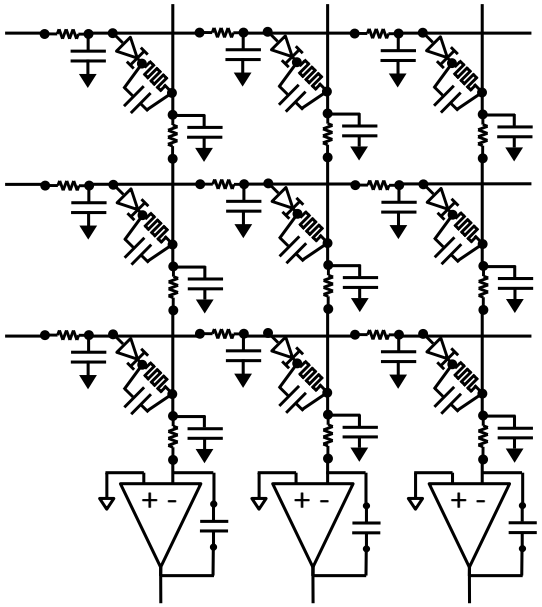


Next Step: System to Device Model

- **Right now, claims in the literature on possible energy efficiency with ReRAM Accelerator vary by 10^6**
- **Architecture-level simulations do not include device accurate models**
- **Device level simulations do not analyze system level attributes**
- **Solution:**
  - **Model and simulation based on all device-level variability data and compact models**
  - **Model all system level components**
  - **Circuit-level energy analysis**
- **Several groups have started this**

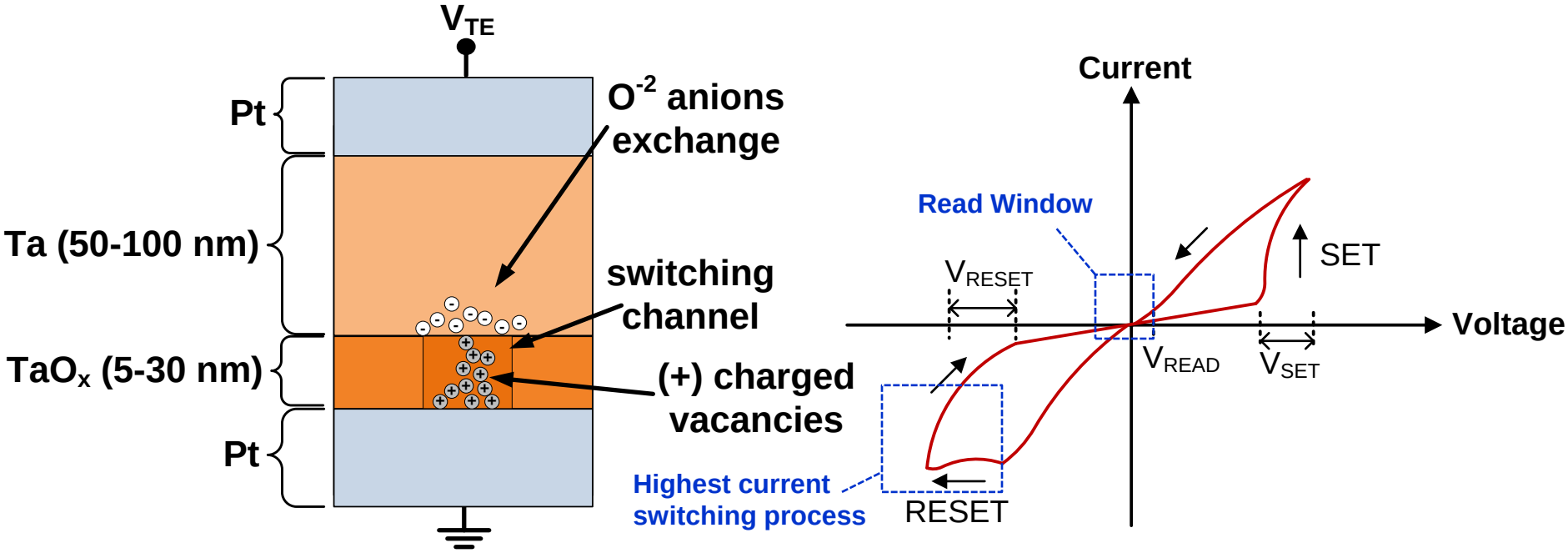
	Example 1: 25,600 neurons 100,000 iterations/s				
Configuration	# of chips	Chip area (mm^2)	% Active	Power (W)	**Power eff. over Xeon**
Memristor Analog (config 1)	1	3.7	9.7%	0.07	**253,489**
Memristor Digital (config 2)	1	9.7	60.8%	0.62	**27,546**
SRAM (config 3)	1	35.2	60.8%	1.13	**15,099**
NVIDIA M2070	12	529.0	99.2%	2700.00	**6**
Intel Xeon X5650	179	240.0	99.9%	17005.00	**1**

Taha et al IJCNN 2013



Metal Oxide ReRAM Device

- "Hysteresis loop" is simple method to visualize operation
  - (memory operated through positive and negative pulses)
- Hypothesized oxide resistance switching mechanism
  - Positive V_{TE}: low R – O^{-2} anions leave oxide
  - Negative V_{TE}: high R – O^{-2} anions return
- Despite progress, fine details of switching mechanism still debated
- Scalable to <5nm



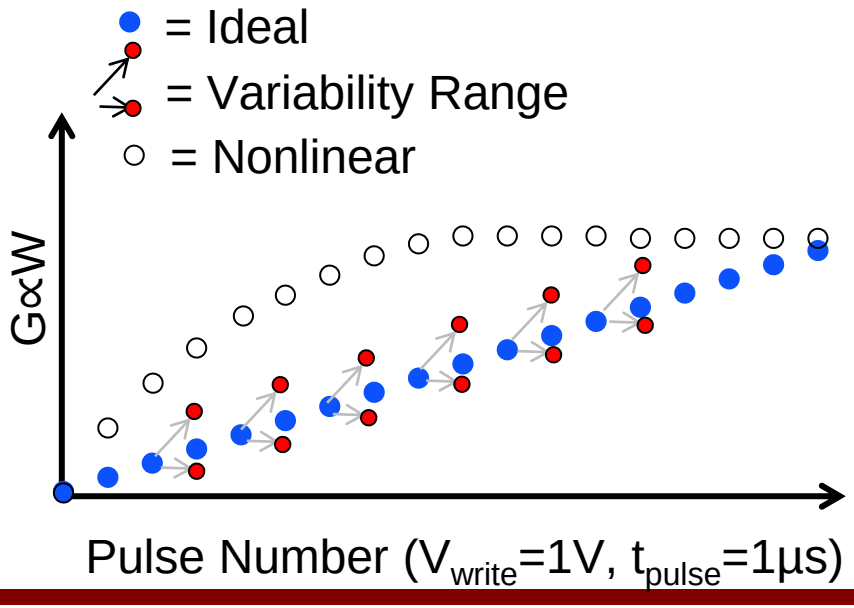
Experimental Device Characteristics

Sandia National Laboratories

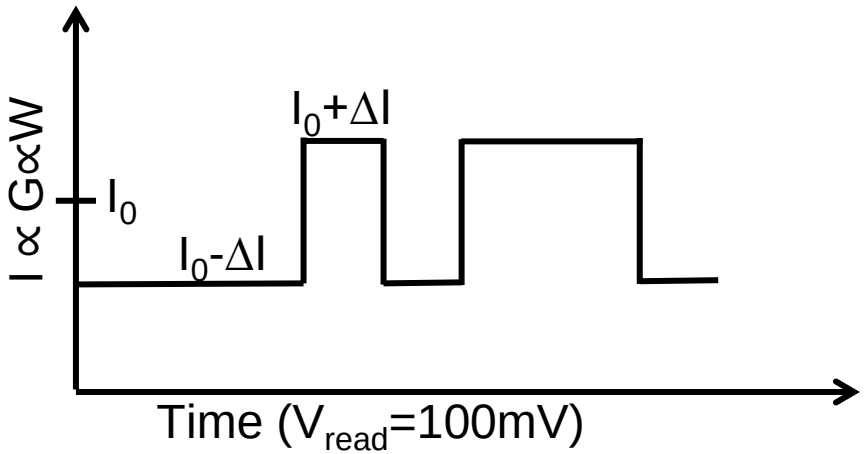
Incorporated experimental device and circuit level electrical characteristics to determine the design space criteria for algorithm convergence:

- **Device: Write Variability, Write Nonlinearity, Asymmetry, Read Noise**
- **Circuit: A/D, D/A noise, parasitics**

Variability and Nonlinearity

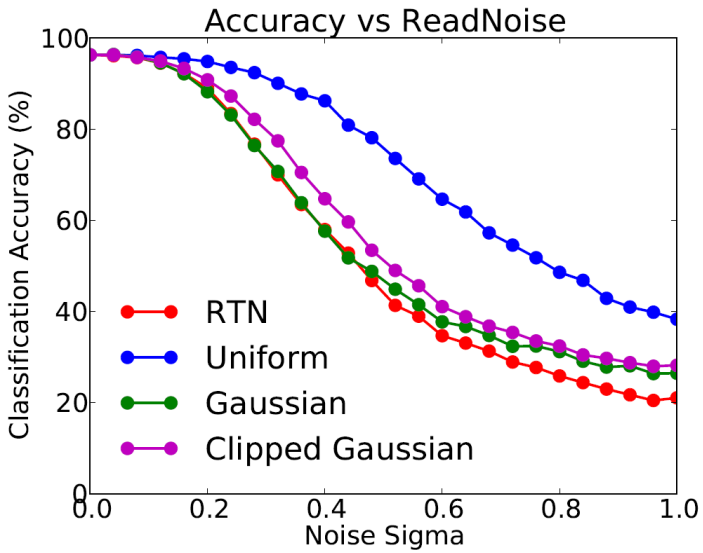


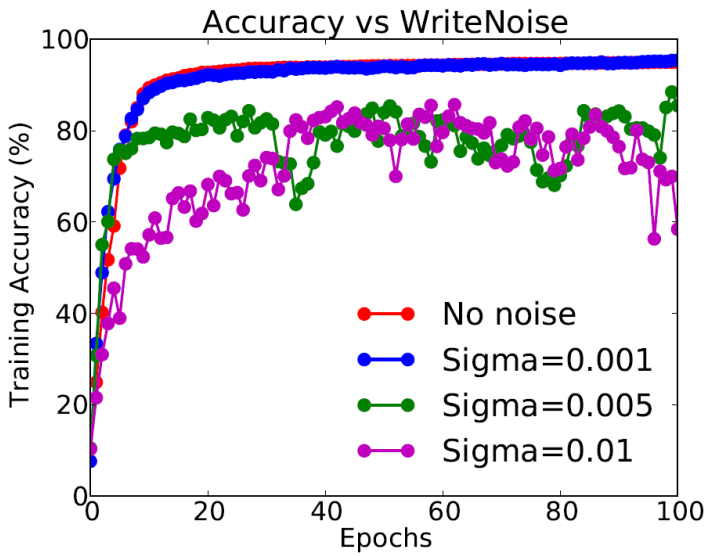
Read Noise

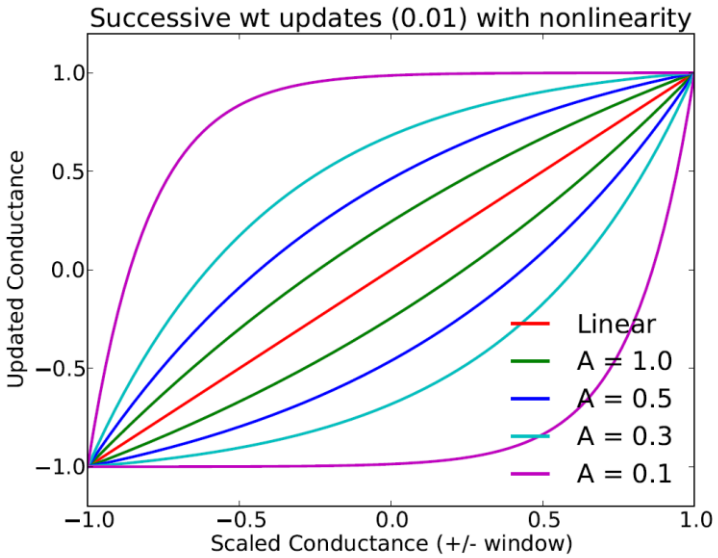


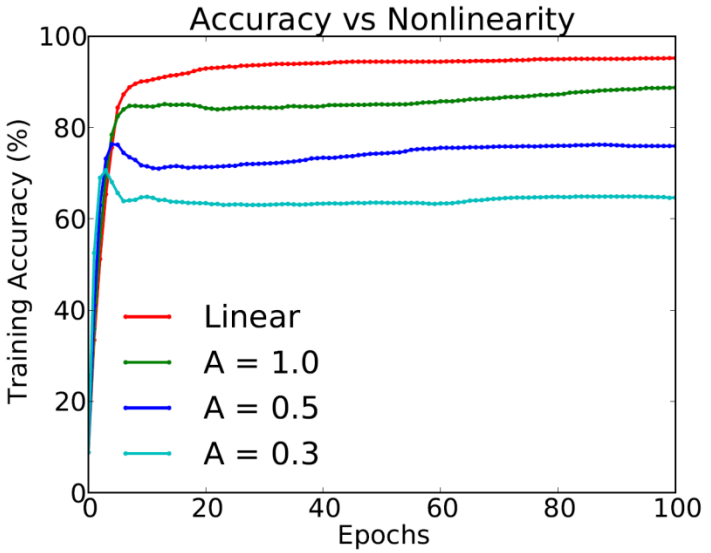
Effects of Read and Write Noise

Sandia National Laboratories

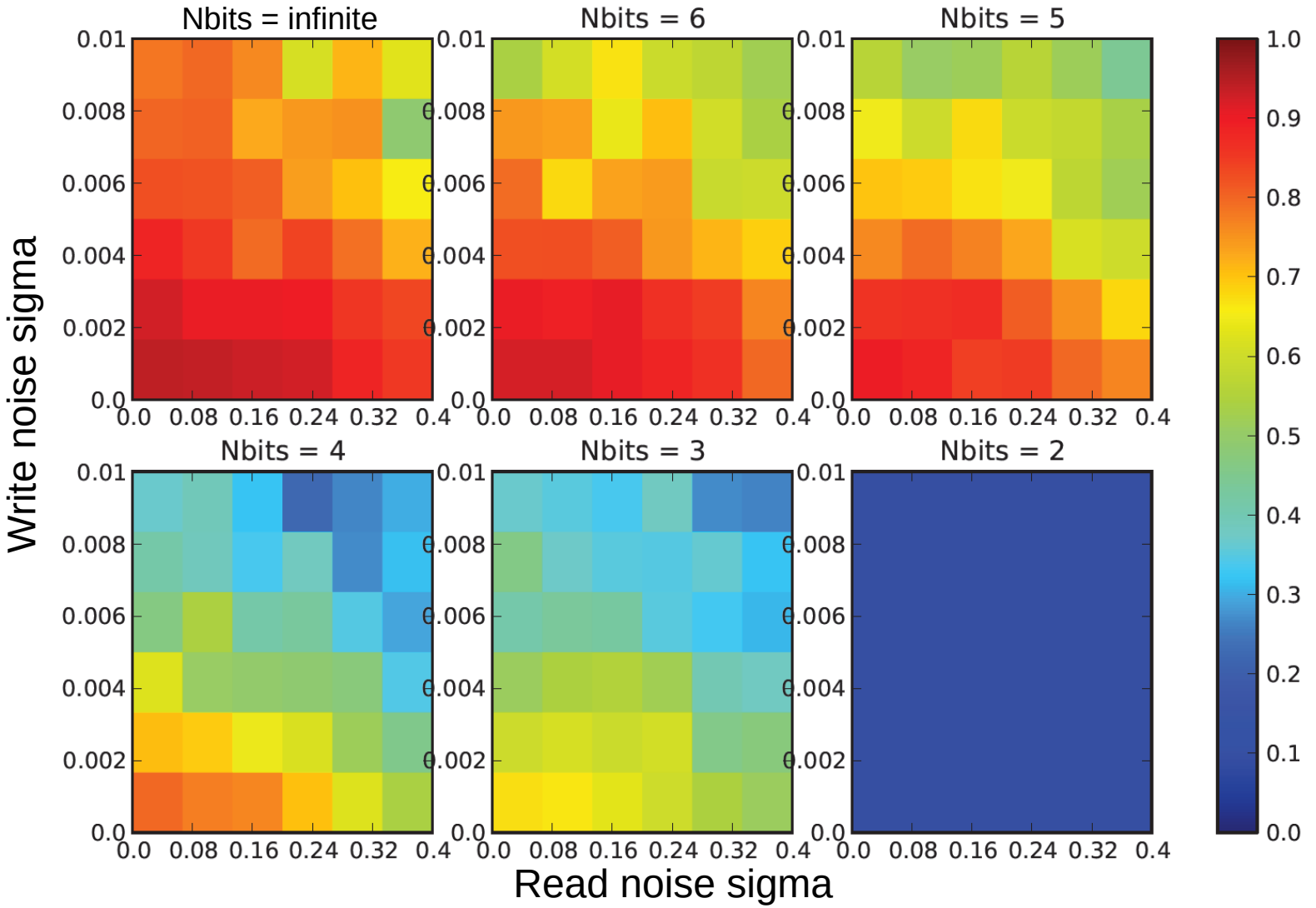








Training Accuracy vs Read and Write Noise for Different Bit Precisions

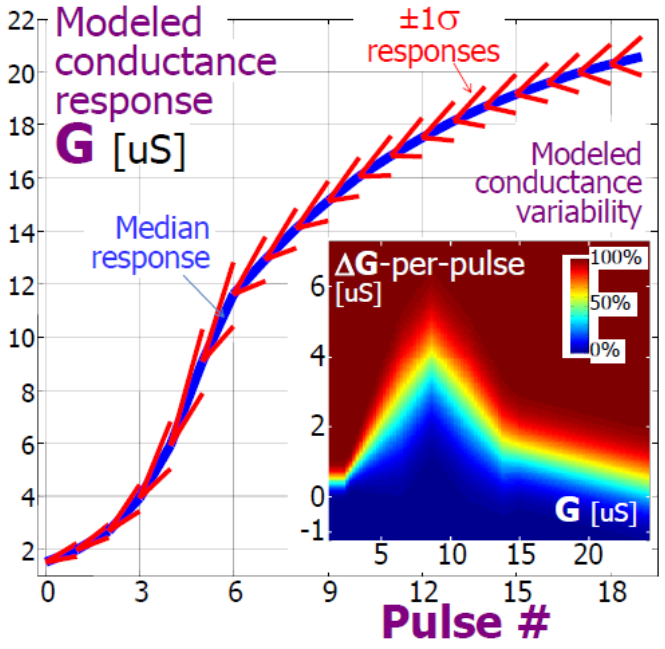


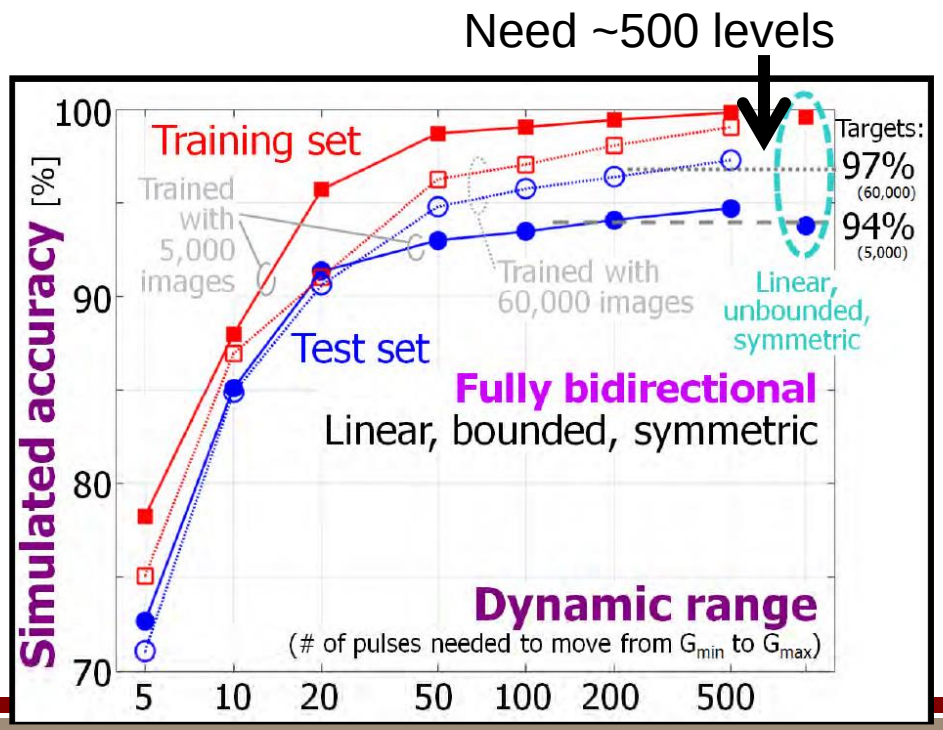
Fewer bits may be possible with better input/output ranges and stochastic rounding

IBM Backprop Training with PCRAM

Sandia National Laboratories

- Thorough study; simulation trained MNIST on actual PCRAM chip and achieved ~85% accuracy w/ 20 level
- Full numerical accuracy reached with ~200-500 levels (7-9 bit)





Device Requirements

Sandia National Laboratories

Requirements set by Simulation:

1. **Need write noise, σ_{write}, <0.05-0.001**
2. **Linearity factor > 1**
3. **Need symmetric conductance changes**

Never demonstrated experimentally

- **10MΩ on-state resistances required.**
- **Read noise (σ_{noise}<0.10) typically within experimental device capabilities**
- **Scalability <10nm**

Device Research Tasks:

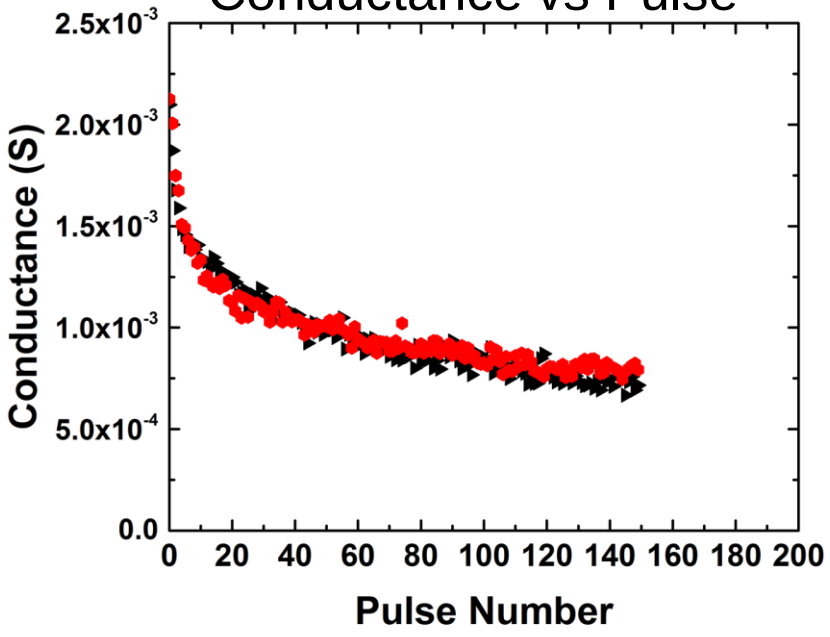
- **Measure and assess experimental devices against these requirements**
- **Provide data to the system level model**
- **Learn from system assessment and improve devices**

Device Characterization and Modeling

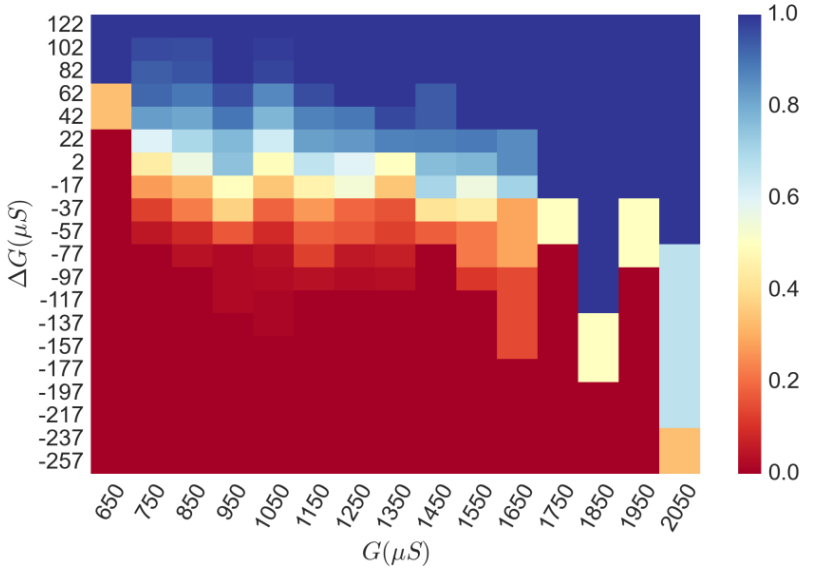
Sandia National Laboratories

1. **Characterize repeated pulsing behavior**
2. **Arrange data as required by circuit-level model and look-up table**
3. **Create analytical noise model for higher-level numerical simulation**

Conductance vs Pulse



Delta Conductance vs Conductance State

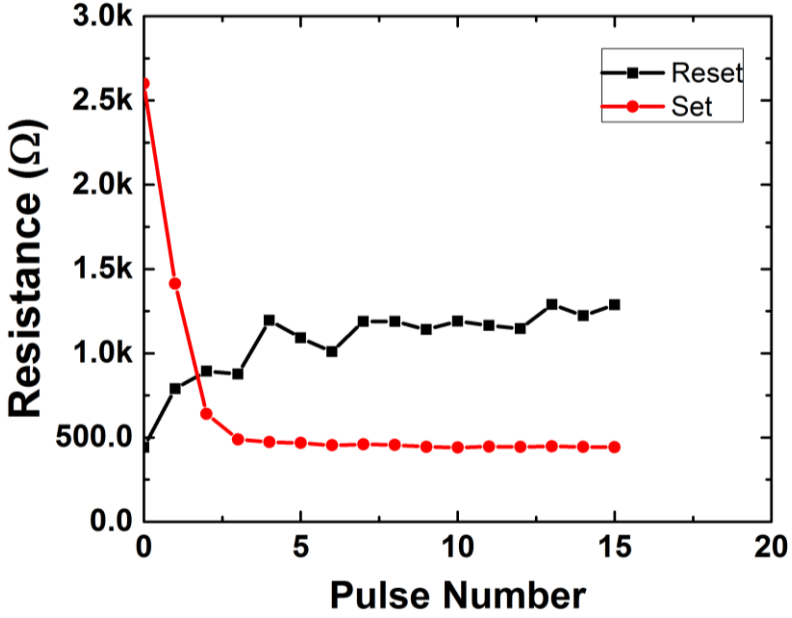


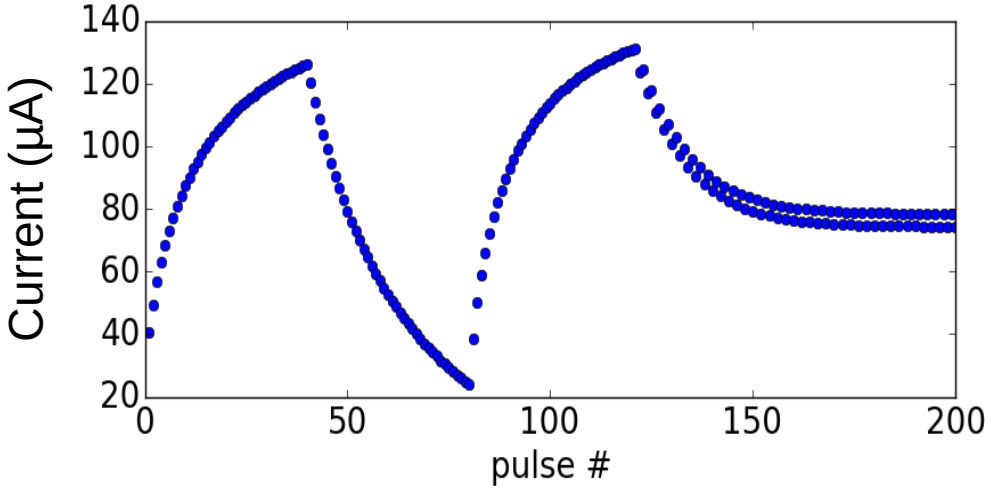
$$G = G_0 + \Delta G + N(\sigma_1)$$

Original state — Target update — Noise

Asymmetry & Nonlinearity

- **Assymetric analog resistance change in filamentary TiN/TaO_x cell**
  - **SET abrupt – thermal runaway?**
  - **RESET – gradual transition**
- **Nonlinear, asymmetric G vs pulse curve**
- **Need device improvements!**





Key Points

Sandia National Laboratories

- **Brain inspired computing has experienced a resurgence of interest…**
  - **Has it found a killer app?**
- **New devices and hardware *are* needed**
  - ***The brain is a piece of hardware*!**
  - **Currently cannot apply neural algorithms to true internet-scale datasets**
- **New algorithms should be developed in parallel with hardware**
  - **Coordination of multidisciplinary research needed: EE, CS, neuroscientists, material sci, etc**

Possible Discussion Questions

Sandia National Laboratories

- **What is the "killer app" for brain inspired computing? (asked in earlier RCS session)**
- **What are the key kernels?**
  - **Matrix operations? Spike timing dependent plasticity?**
- **What are the key new devices and circuits?**
  - **Resistive crossbars?**
- **What level of energy efficiency and/or performance improvement is needed to justify a new device technology?**
  - **Tech development is very expensive!**