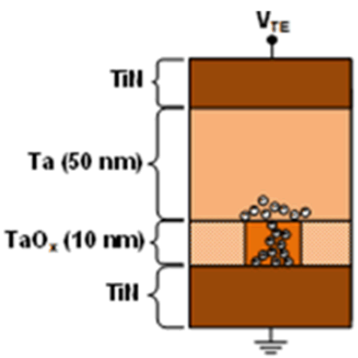
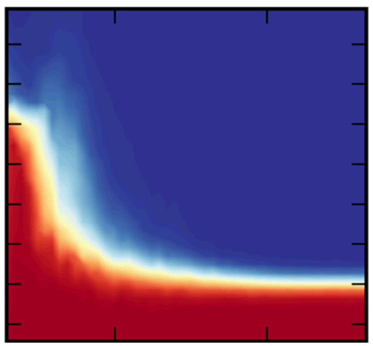
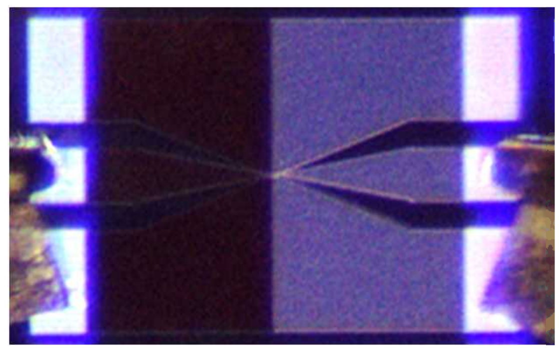
SAND2016-10494C

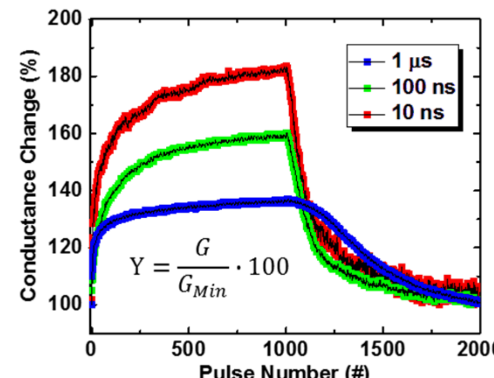
Exceptional service in the national interest

Sandia National Laboratories









Evaluating Resistive Memory Devices for Neuromorphic Computing Using Ultrashort Voltage Pulses

Robin Jacobs-Gedrim[1], David Hughart, Sapan Agarwal, Patrick Finnegan, Ronald Goeke, Michael Van Heukelom, Joshua Nowlin, Jamison Wagner, Conrad James, Matthew Marinella[2]

[1]rbjaco@sandia.gov

[2]mmarine@sandia.gov

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

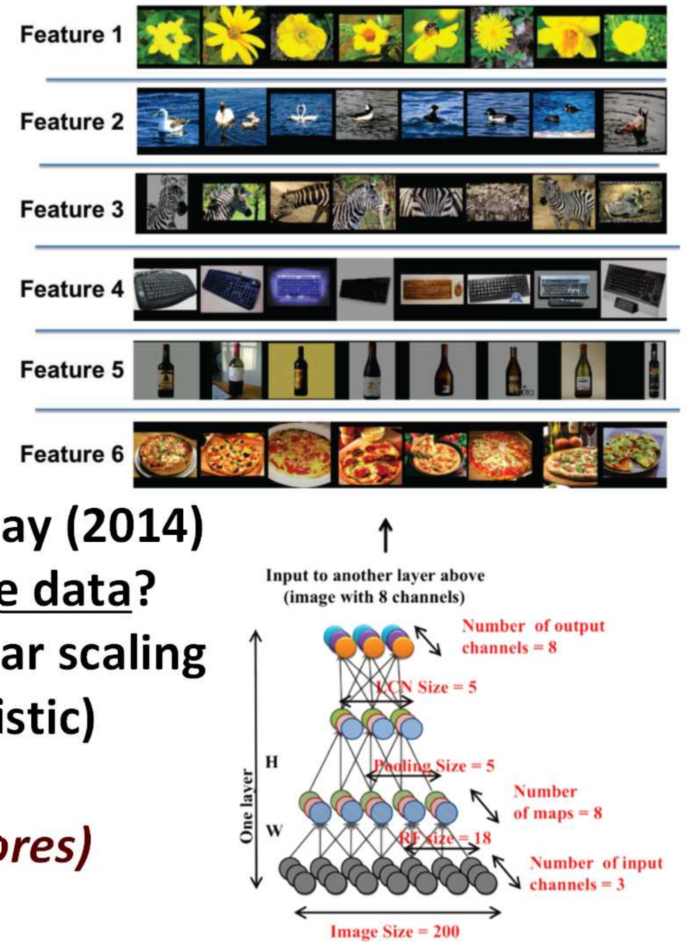
Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

Outline

- **Intro and Key Concepts**
- Experimental Results
  - Ultrafast Pulsing
  - CrossSim Training Results
- Conclusions and Future Work

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 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)***
  - **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	10^5 pJ/MAC	Early DSP Demonstrations
1990's	10^4 pJ/MAC	
2000's	10^3 pJ/MAC	
2010-Present	10^2 pJ/MAC	Modern GPU, 32 bit precision
		Single GPU Card, Late 2016
2025?	10 pJ/MAC	Exascale System Goal
	1 pJ/MAC	**1 TMAC/W**
	100 fJ/MAC	Analog VMM Prototypes
	10 fJ/MAC	
2035?	1 fJ/MAC	
	100 aJ/MAC	
	10 aJ/MAC	
	1 aJ/MAC	Biological Neurons*

Improvements Due to Transistor scaling (Moore's Law and Denard Voltage Scaling)

Neuromorphic Target (Efficiency only possible with neural hardware & algorithms)

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

Modified from Hasler and Marr, Frontiers in Neuroscience, 2013

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

How does a crossbar perform a useful computation per device?

- Electronic Vector Matrix Multiply

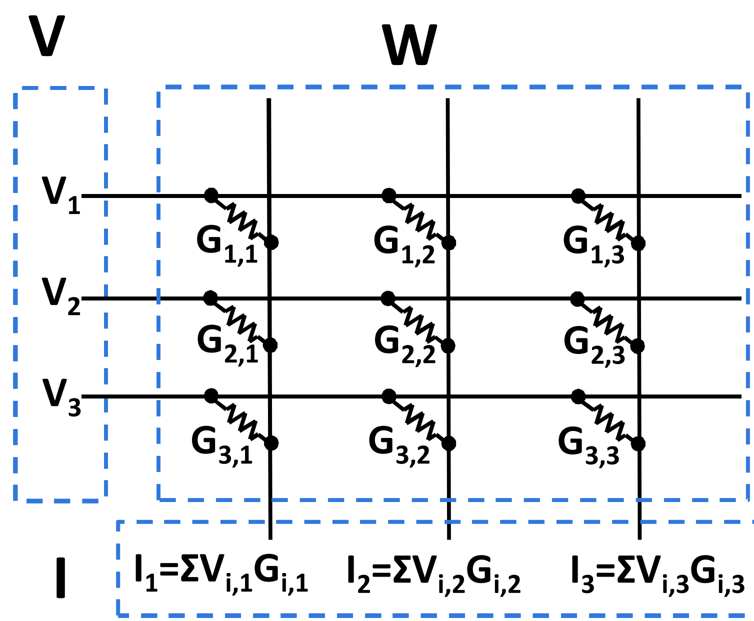
Mathematical

$$V^T W = I$$

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

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

Electrical



Neuromorphic Accelerator Device Requirements

Sandia National Laboratories

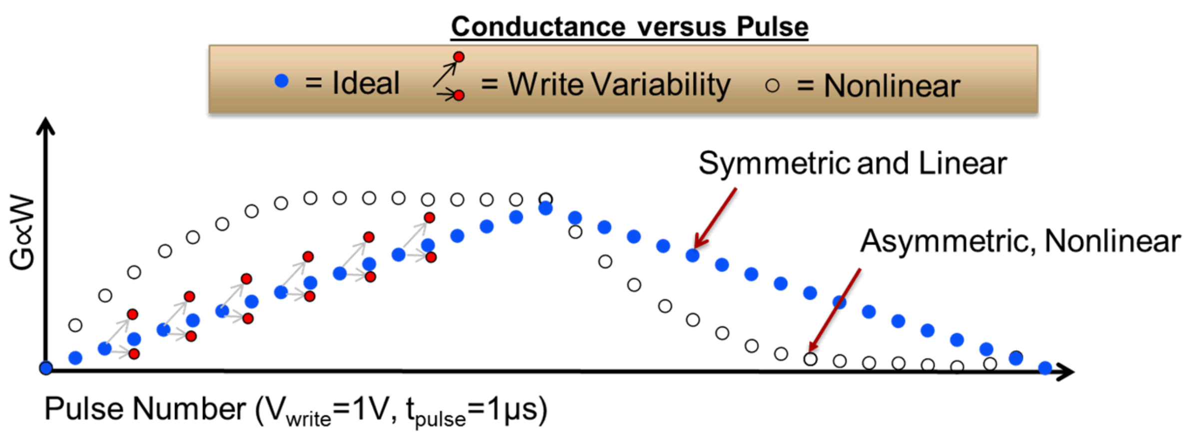
For scalable crossbars want:

- **Low operating current**
- **Low operating voltage**

	Small Images	Large Images	File Types
Read Noise σ (% Range)	3%	5%	9%
Write Noise σ (% Range)	0.3%	0.4%	0.4%
Asymmetric Nonlinearity (ν)	0.1	0.1	0.1
Symmetric Nonlinearity (ν)	>20	5	5
Maximum Current	160 nA	13 nA	40 nA
Minimum Retention (@ 85°C)	7 days	7 days	7 days
Minimum Nudge Endurance	10^7	10^7	10^7

For training accuracy want:

- **Low Write Variability**
- **Low Write Nonlinearity**
- **Low Asymmetry**
- **Low Read Noise**



S. Agarwal, et. al. "Resistive Memory Device Requirements for a Neural Algorithm Accelerator," *International Joint Conference on Neural Networks,* (2016)

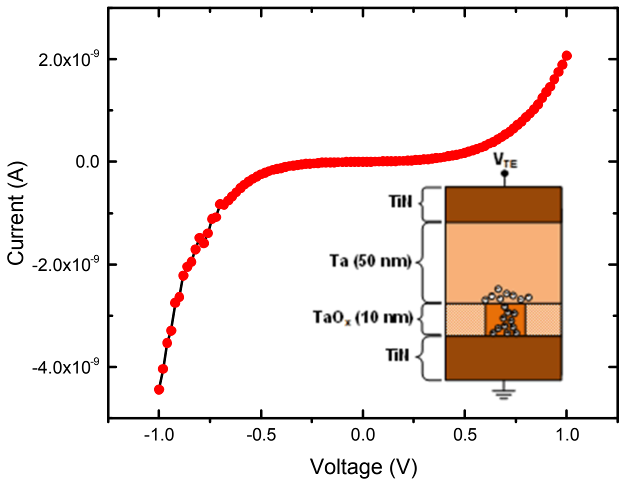
Outline

- Intro and Key Concepts
- **Experimental Results**
  - **Ultrafast Pulsing**
  - **CrossSim Training Results**
- Conclusions and Future Work

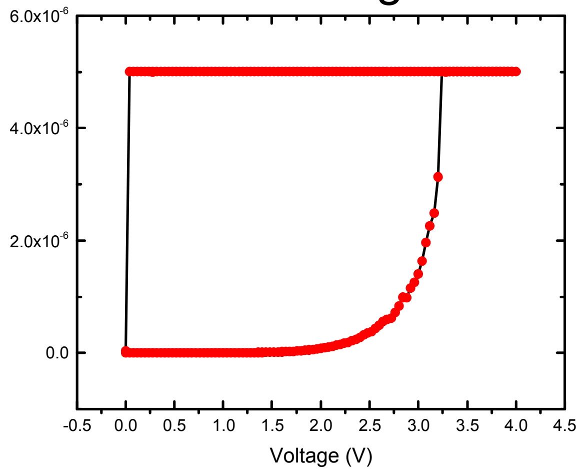
Resistive RAM (ReRAM)

- **Bipolar Metal Oxide Redox RAM**
- **Resistance is modulated in a metal oxide**
- **Scalable to 5nm, sub pJ switching energies**
- **TiN/Ta/TaOx/TiN Device Stack**

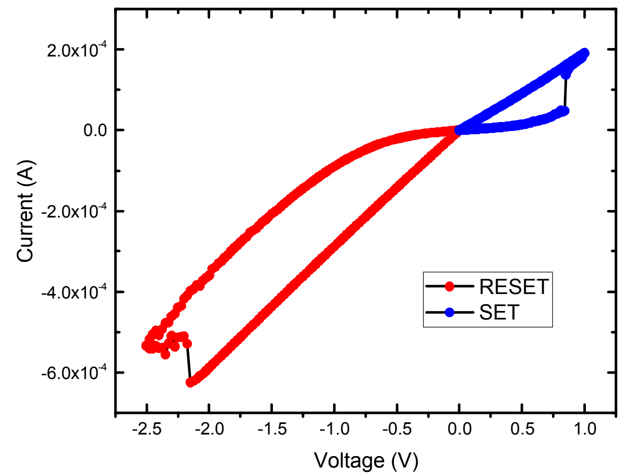
Pre-Form I/V



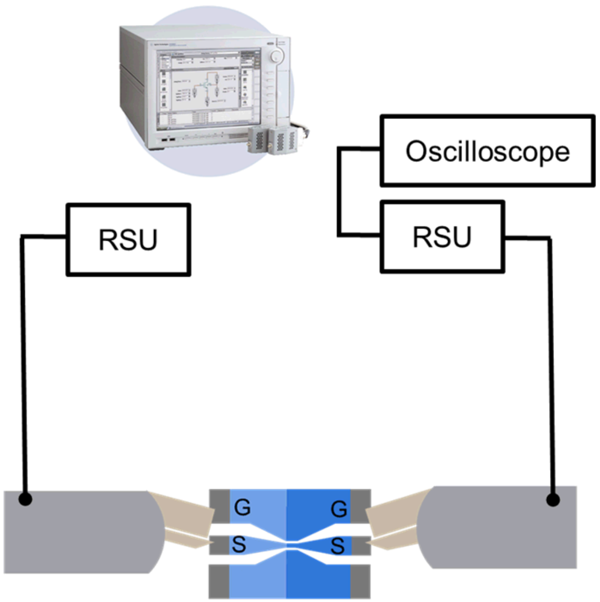
Forming

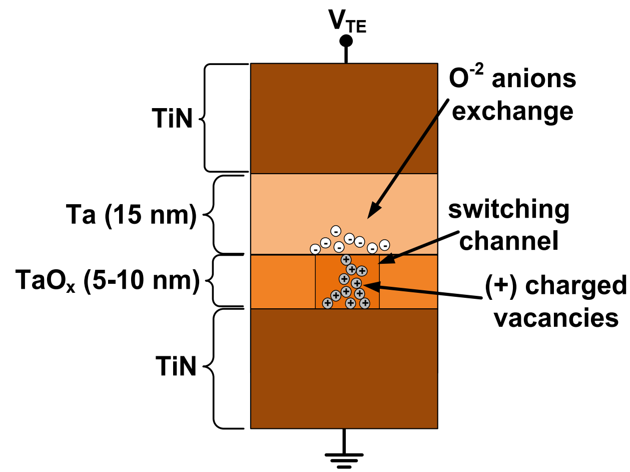


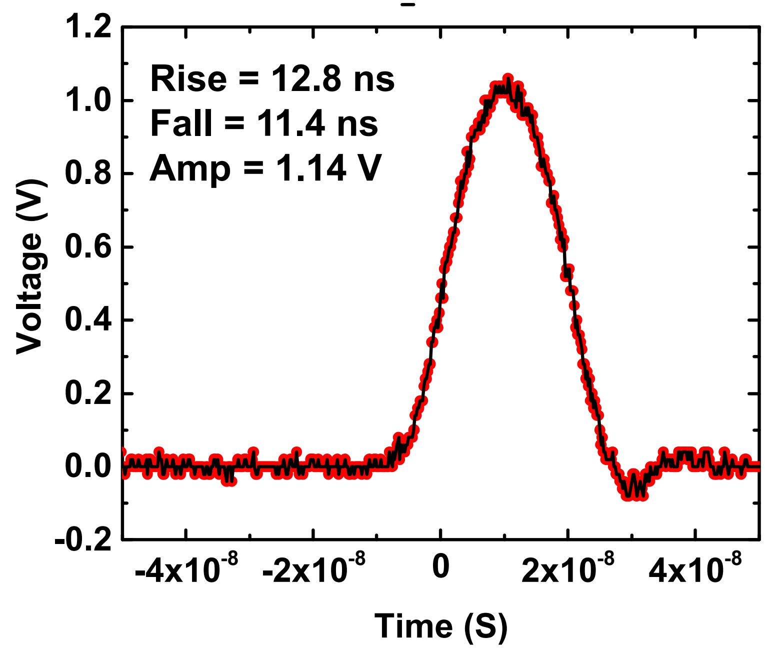
SET-RESET



Fast Pulsing / Coplanar Waveguides



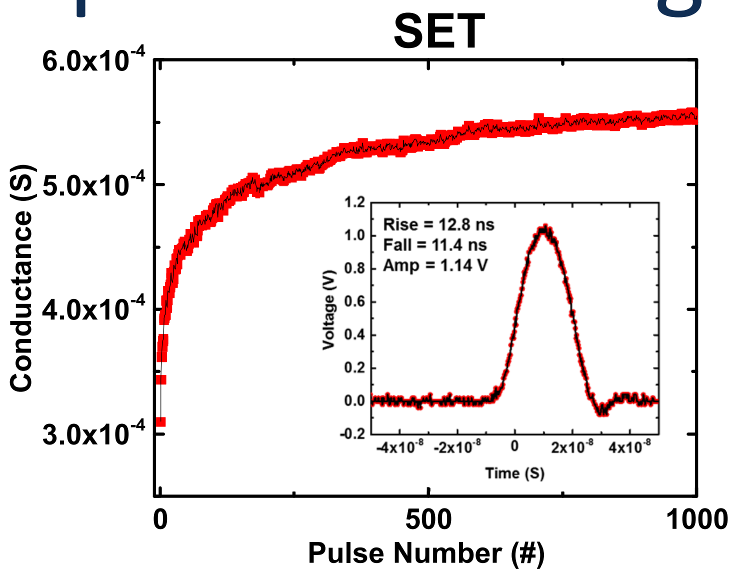


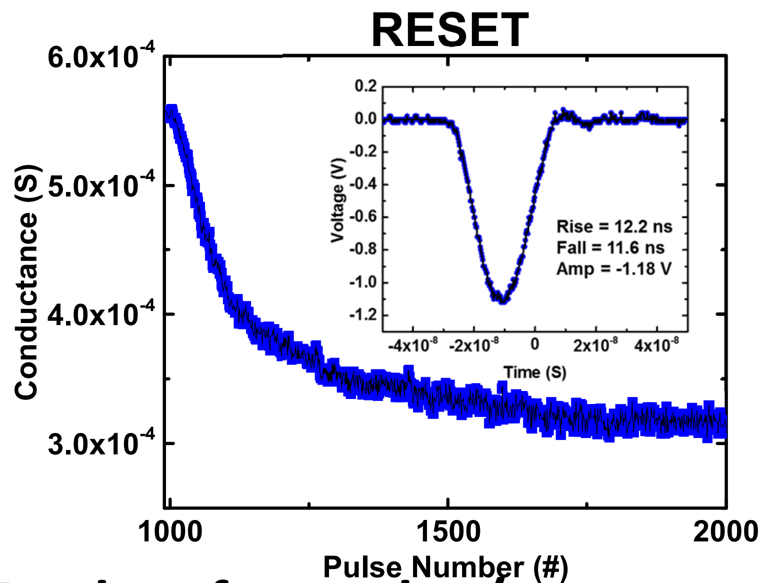


- **Ground Signal**
- **Conductor backed**
- **Agilent B1530 module**
- **10 ns RT/FT, 10 ns PW**
- **1 V nominal**
- **~140 mV overshoot**

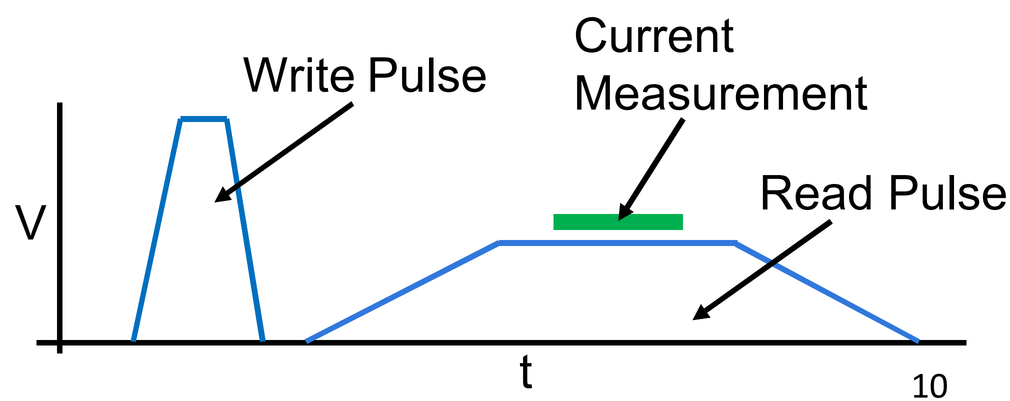
Repeated Pulsing

Sandia National Laboratories

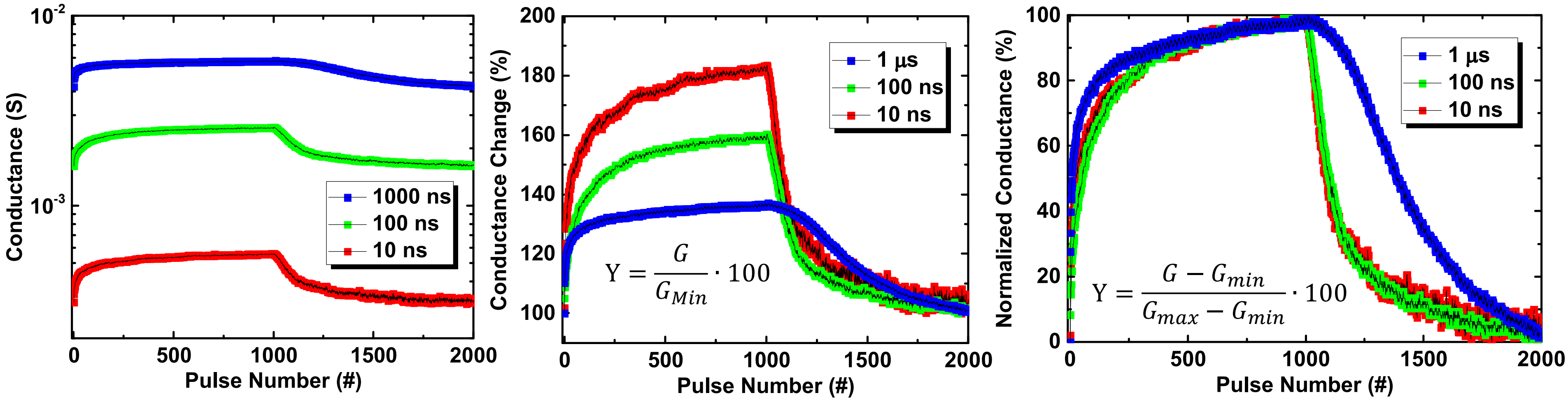




- **1000 repeated SET and RESET pulses for analog/neuromorphic operation**
- **Blind updates**
- **Nominal pulse values**
  - **SET: +1V 10ns RT/PW/FT**
  - **RESET: -1V 10ns RT/PW/FT**
  - **READ: 100 mV 1 ms RT/PW/FT**



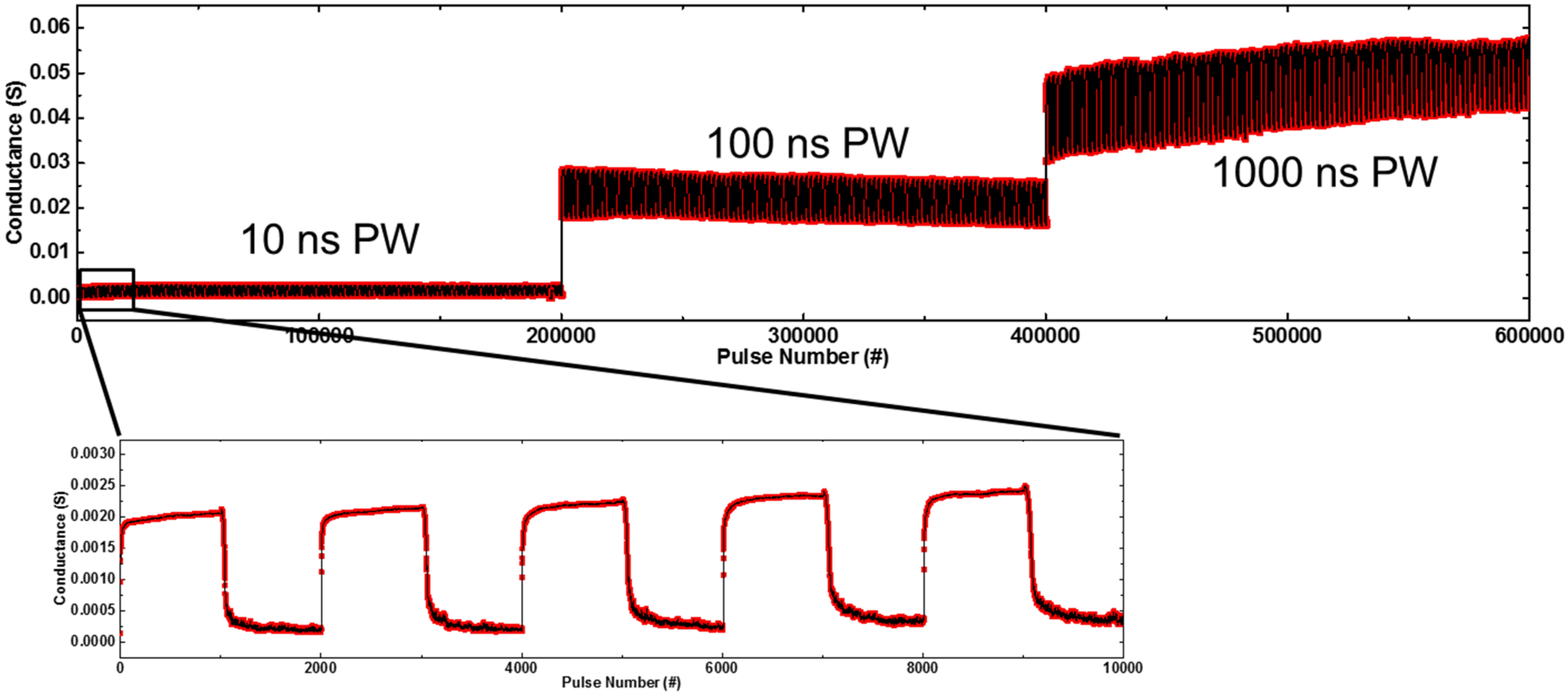
Effect of Pulse Width and Edge Time



- 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

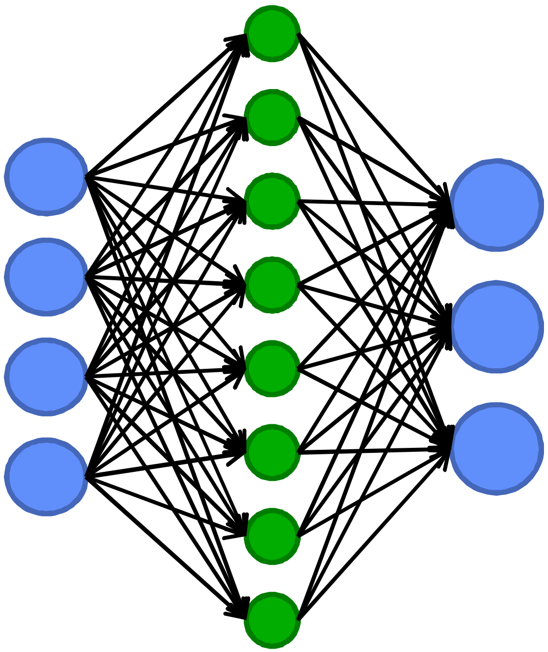
Nominal Pulse Voltage Values: SET: +1 V RESET: -1 V

Repeated Pulsed Cycling

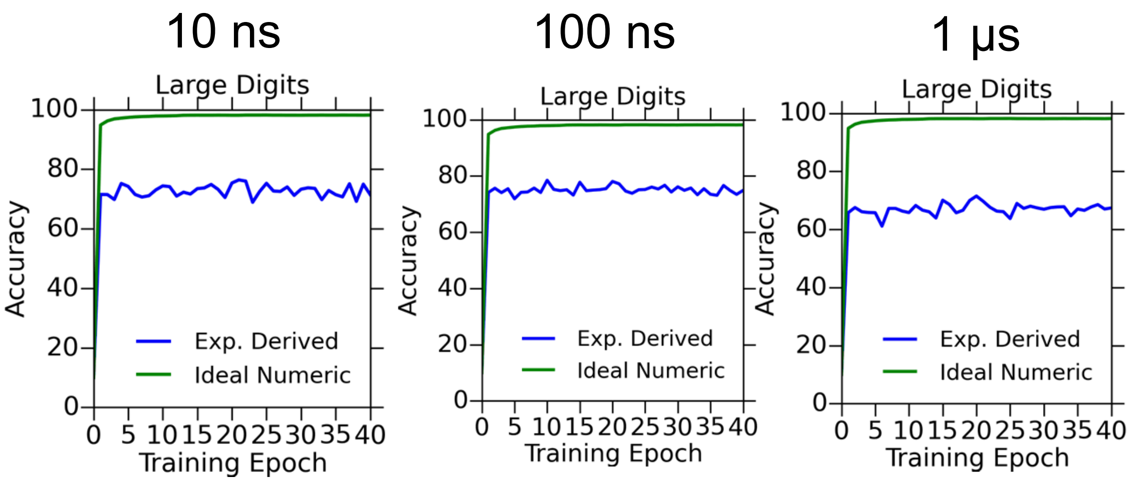


- Tested 100 cycles of 1000 SET+RESET pulses with 10, 100, 1000 ns PW/ET
- Devices stable over large number of pulsing cycles
- Large discontinuity in conductance when changing PW and ET is not from device instability over time/repeated cycling

CrossSim Training Results



Data set	# Training Examples	# Test Examples	Network Size
MNIST Large Digits	60,000	10,000	784×300×10



- **Trained CrossSim on MNIST Large Digit set of handwritten images**
- **Higher image recognition after training for devices operated at 10 ns and 100 ns than 1 µs**

Outline

- Intro and Key Concepts
- Experimental Results
  - Ultrafast Pulsing
  - CrossSim Training Results
- **Conclusions and Future Work**

Conclusions

- **Relative conductance change increased with shorter Pulse Width / Edge Time**

- **Ultrafast pulses may be employed to lower conductance switching range without reducing overall conductance change**

  Ultrafast pulses down to 100s of ps are possible to implement in-silico!

- **Achieved repeated analog operation of TaO_x/Ta devices for neuromorphic computation**

- **Progress has been made toward "fab friendly" devices suitable for image recognition, but still a long way to go before achieving numerical equivalent**

Acknowledgements

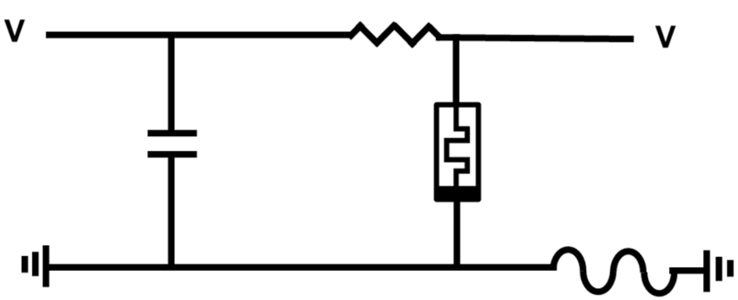
- Sapan Agarwal
- David Hughart
- Patrick Finnegan
- Ronald Goeke
- Michael Van Heukelom
- Joshua Nowlin
- Jamison Wagner
- Conrad James
- Matthew Marinella

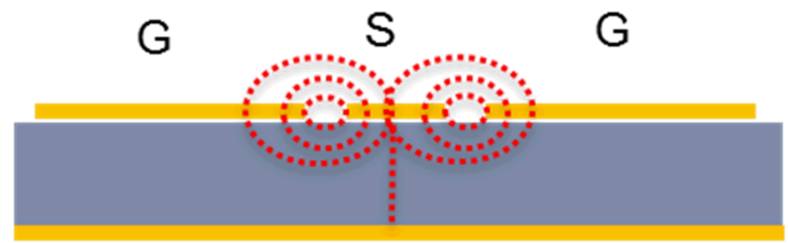
This work is funded by Sandia's Laboratory Directed Research and Development as part of the Hardware Acceleration of Adaptive Neural Algorithms Grand Challenge Project

Thank you!

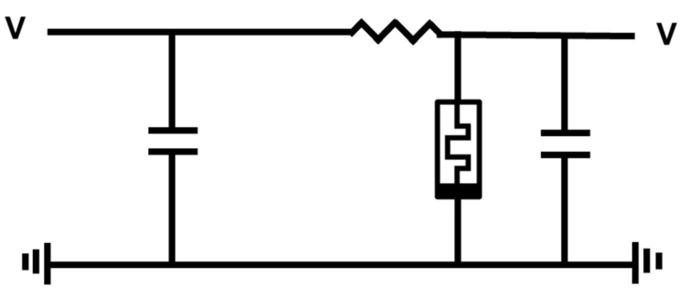
Questions?

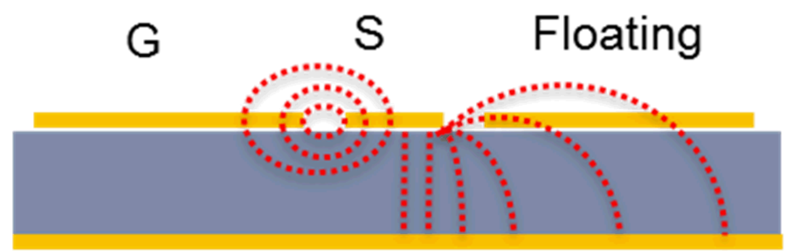
Advantage of Coplanar Waveguides





GSG shields as coplanar waveguide

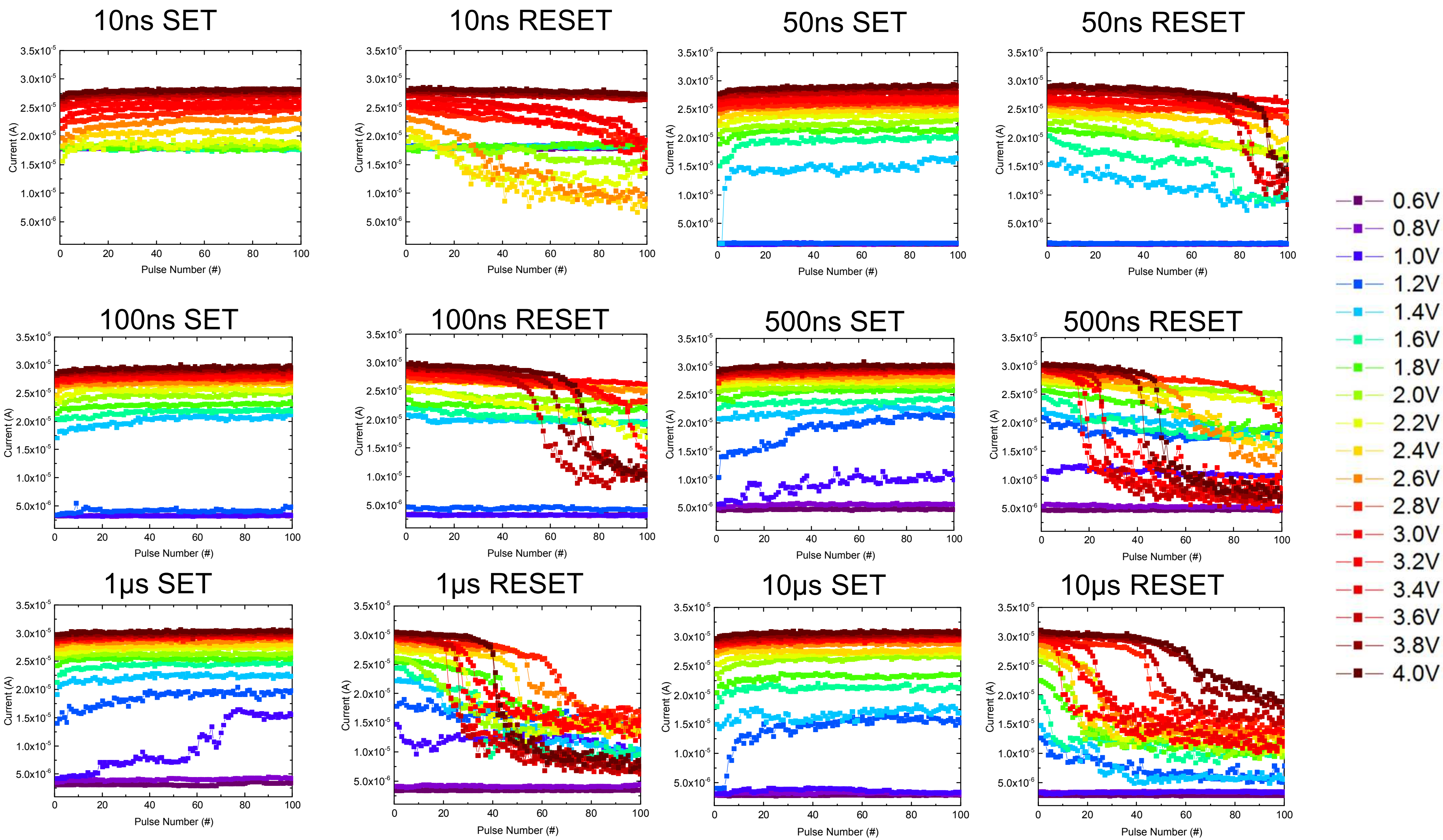




GS Field lines go to Si layer/chuck

- **Without coplanar waveguide - Path to ground is through large inductor**
- **With coplanar waveguide, inductor is replaced by small capacitors**
- **If full GSG configuration, should be able to carry RF signals up to 40 Ghz**
- **Since only using GS probes, frequency is limited, but only need 1 Ghz**

JS100614A_W-11 Parametric Test



100 ns Pulswidth delta G / G