SAND2016-7100C

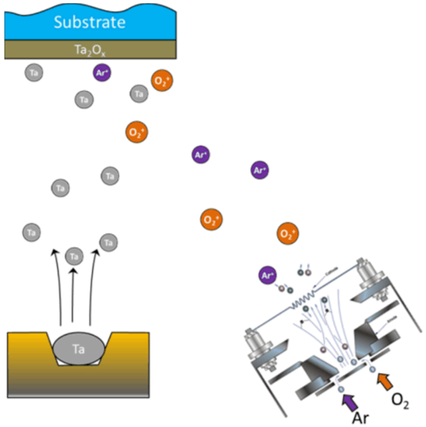
AVS 62nd International Symposium & Exhibition

Tantalum Oxide Resistive Memory Devices by IAD

EM-WeM-1, Wednesday 8:00AM

Ronald S. Goeke[1], M.J. Marinella[2], D.R. Hughart[2], and S.A. Decker[2]

Materials Science and Engineering Center[1]
Microsystems Science and Technology Center[2]
Sandia National Laboratories, Albuquerque, NM, USA

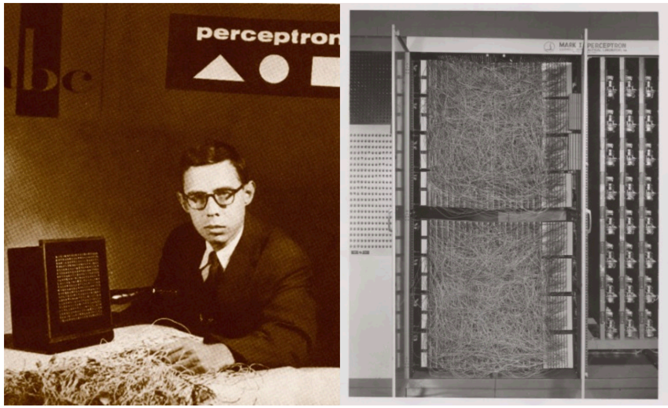


Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000

Neural-inspired Computing Hardware

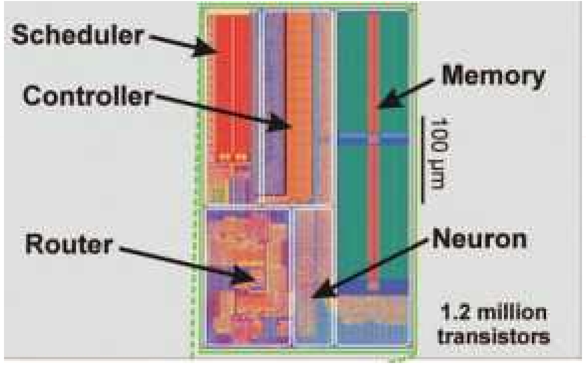
Revival of an Old Idea

Mark I Perceptron (Rosenblatt 1960):

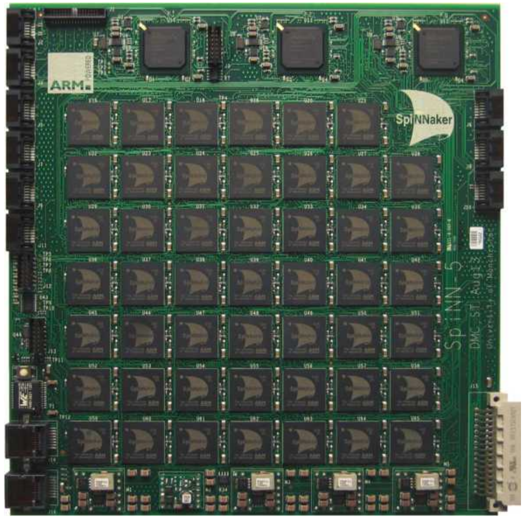


Arvin Calspan Advanced Technology Center; Hecht-Nielsen, R. *Neurocomputing* (Reading, Mass.: Addison-Wesley, 1990); Cornell Library;

DARPA , IBM TrueNorth (2014):



EU HBP, SpiNNaker (2014):

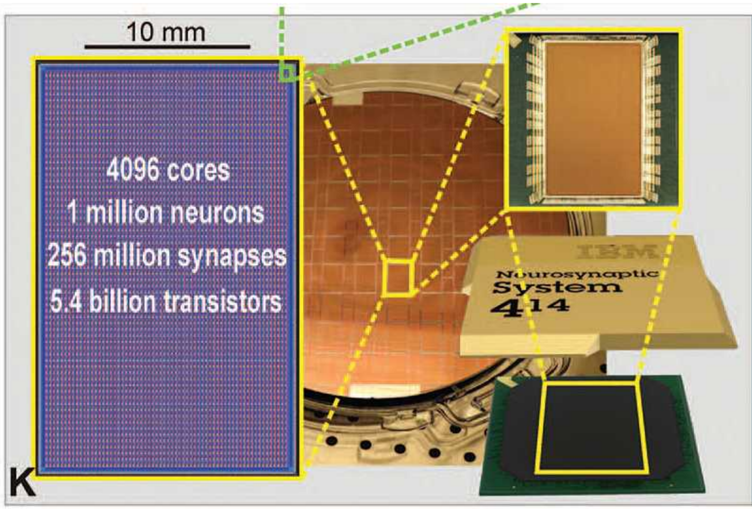


?

Neural Algorithm Computing

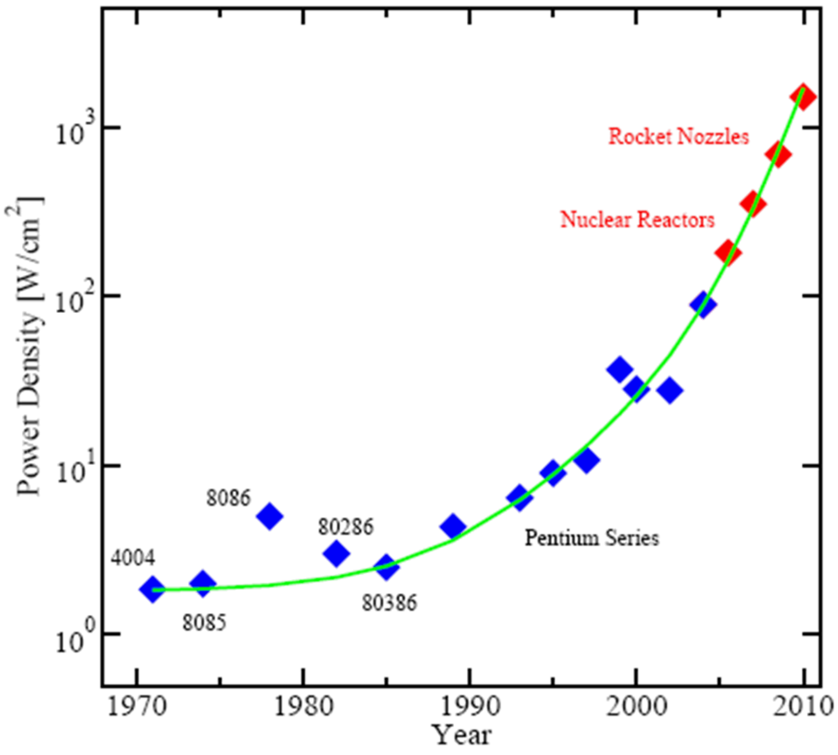
Current State of the Art

- **CPU/GPU**
  - Most general; common programming languages
  - Lowest power efficiency and performance
  - Memory separate from chip
  - Example: Google deep learning study (CPU→GPU)
- **FPGA**
  - General; requires hardware design language
  - Moderate performance and efficiency
- **Custom IC (Truenorth, Spinnaker)**
  - Specific: ex. executes STDP
  - Highest performance and efficiency
  - Expensive, 40MB local memory
  - Example: IBM Truenorth
- **Power ends up as a limiting factor**



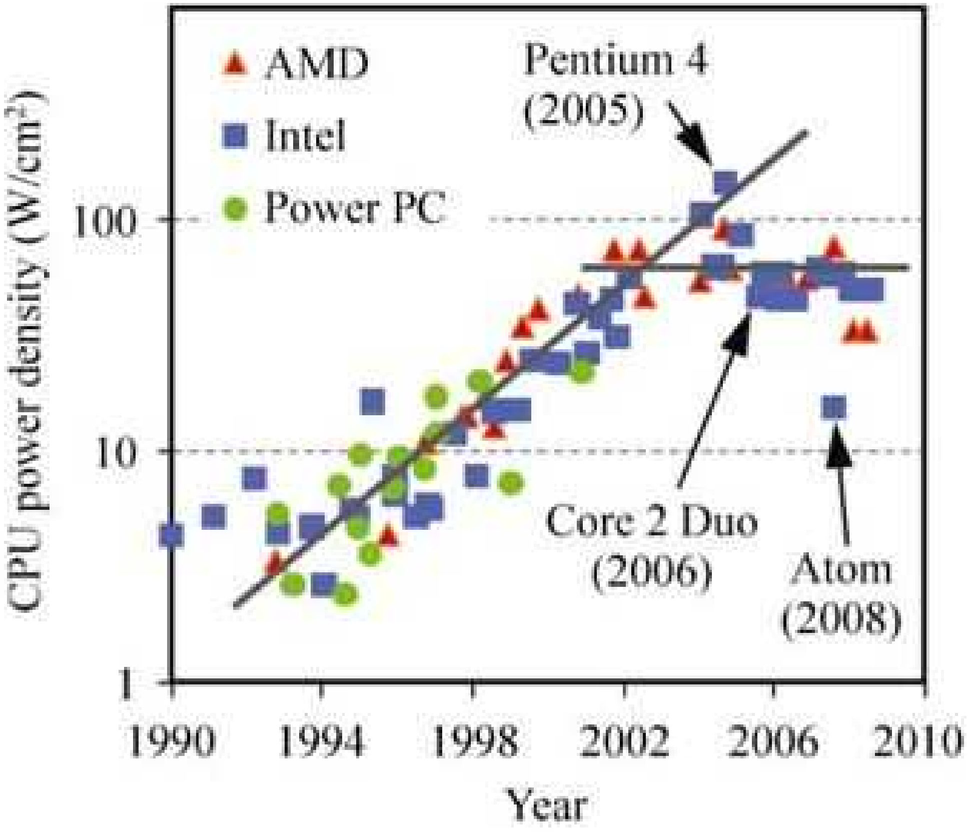
Heating During Computing is Enormous

Low power RAM needed



(a) Power loss density per die

http://www.iue.tuwien.ac.at/phd/holzer/node11.html



http://www.nanotechnologies.qc.ca/blog/industry/energy-dissipation-nanoscale-devices

Power Consumption is Also Enormous

Possible solution?

New devices and new computing paradigms will be needed in the near future.

- Tianhe-2 - World's Fastest Supercomputer
- ~3.4x10^{16} flops/s, **~20 MW, room-sized**
- Nuclear reactors will soon be needed for every supercomputer.

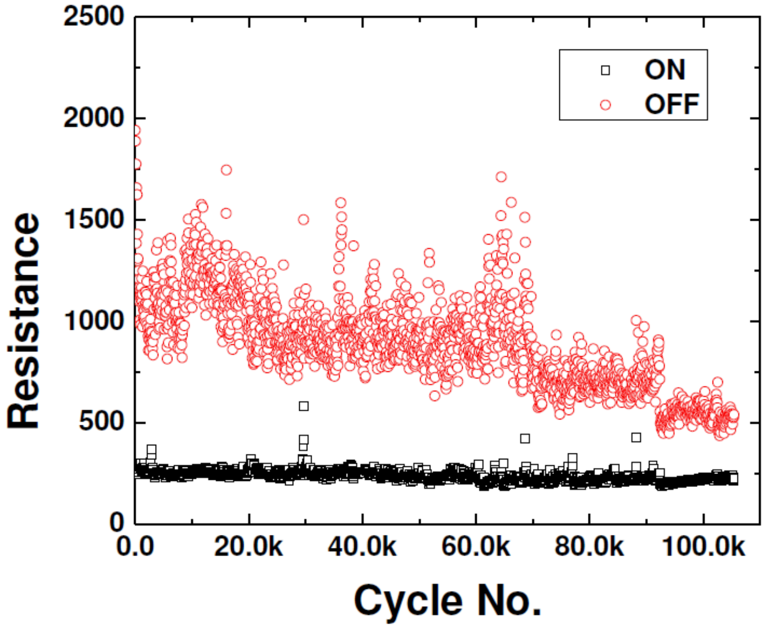
- The Human Brain
- **20W, 1200 cm^3**
- ~10^{18} flops/s are believed to be needed for the human brain project.

Can Memristors Help?

ReRAM

Resistive Memory

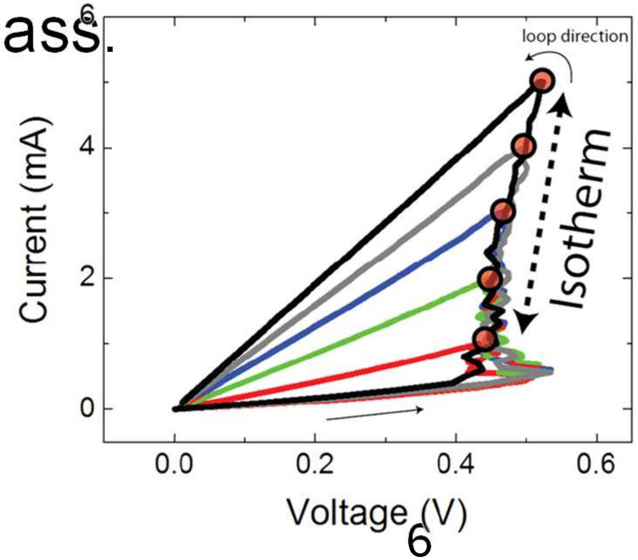
- A voltage or current can change the resistance to a 1 or a 0.



A.J. Lohn, et al, ECS Trans. **58**, 59-65 (2013).

Novel Circuit Element

- Using partial voltages causes only partial changes.
- Similarly to synapses - as current passes, it becomes easier for more current to pass.



A.J. Lohn, et al, Adv. Mater. **26**, 4486 (2014).

Why Do We Need an HW Accelerator?

Significant power savings

Problem: perceptron network training slow and extremely computationally intensive

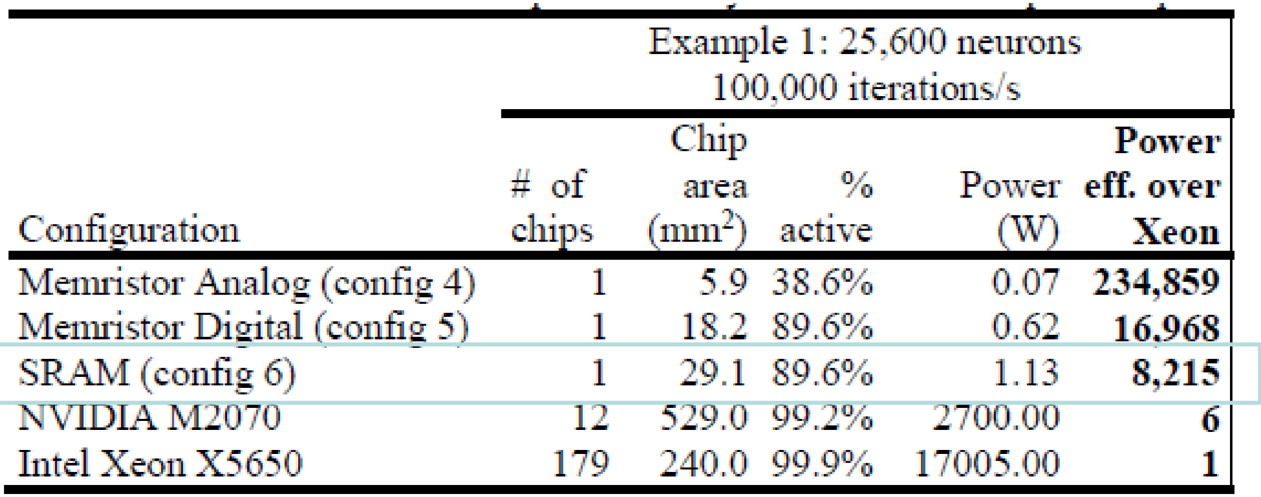
Use simulation results for similar algorithm as example

Significant power savings using a memristor-based HW accelerator :

16x reduction in power over SRAM ASIC

6x reduction in chip area over SRAM ASIC

Equivalent to 6x improvement in performance/area



	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 4)	1	5.9	38.6%	0.07	**234,859**
Memristor Digital (config 5)	1	18.2	89.6%	0.62	**16,968**
SRAM (config 6)	1	29.1	89.6%	1.13	**8,215**
NVIDIA M2070	12	529.0	99.2%	2700.00	**6**
Intel Xeon X5650	179	240.0	99.9%	17005.00	**1**

T. Taha, et al, IEEE IJCNN 2013.

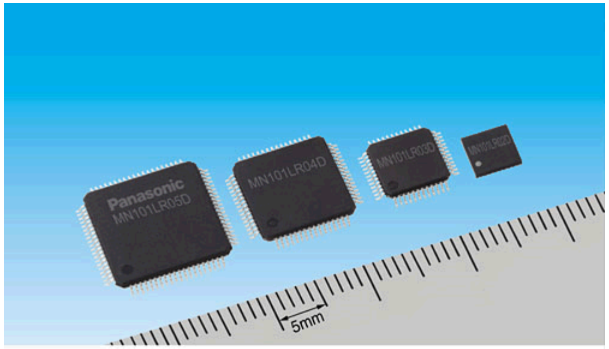
Memristors

Who’s interested

Resistive Memory

- Panasonic recently released the first product.

Panasonic
ideas for life



ReRAM Mounted Low Power Consumption Microcomputer, MN101LR Series

- They claim it is 5-10 times faster using half the power.
  - Probably saved in the wires.

Novel Circuit Element

- Brain-inspired computing using memristors is receiving heavy investment from big companies.

hp

QUALCOMM

IBM

HRL LABORATORIES

intel

Sandia National Laboratories

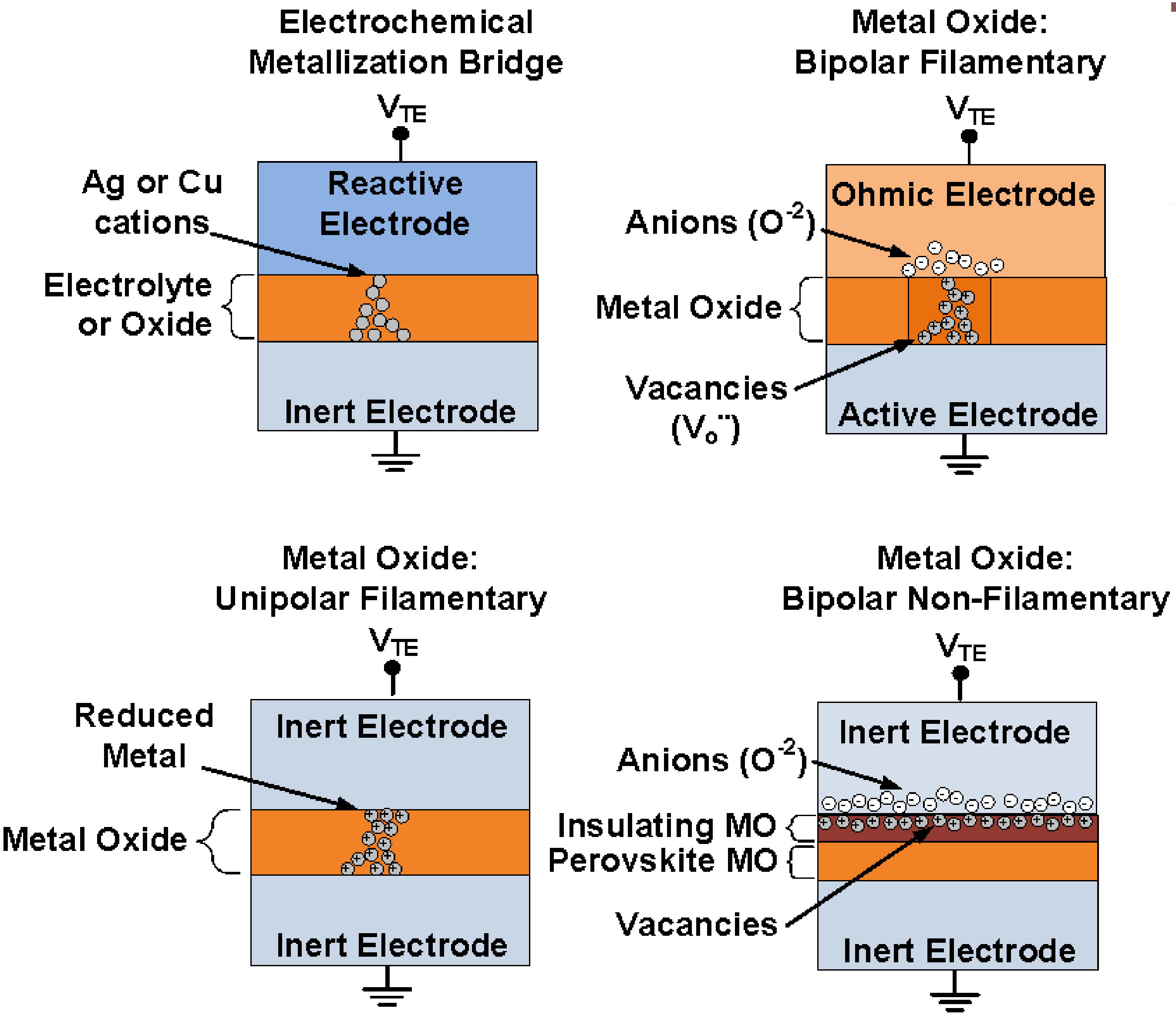
Brain Inspired Computing at Sandia

HAANA

- Sandia established a multidisciplinary LDRD project in neuromorphic computing project in October 2015
  - Hardware Acceleration of Adaptive Neural Algorithms (HAANA)
  - Algorithms research
  - Hardware and Device Architectures
  - Resistive Memory Devices
- Some of the initial work on development of resistive devices covered in this talk

Resistive Random Access Memory

ReRAM



Electrochemical Metallization Bridge

Metal Oxide: Bipolar Filamentary

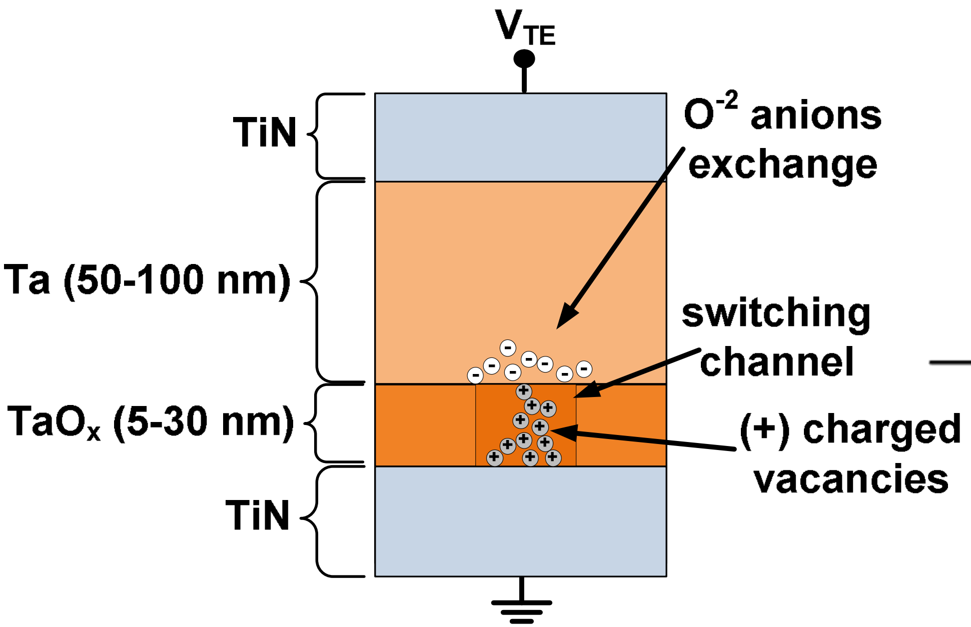
Metal Oxide: Unipolar Filamentary

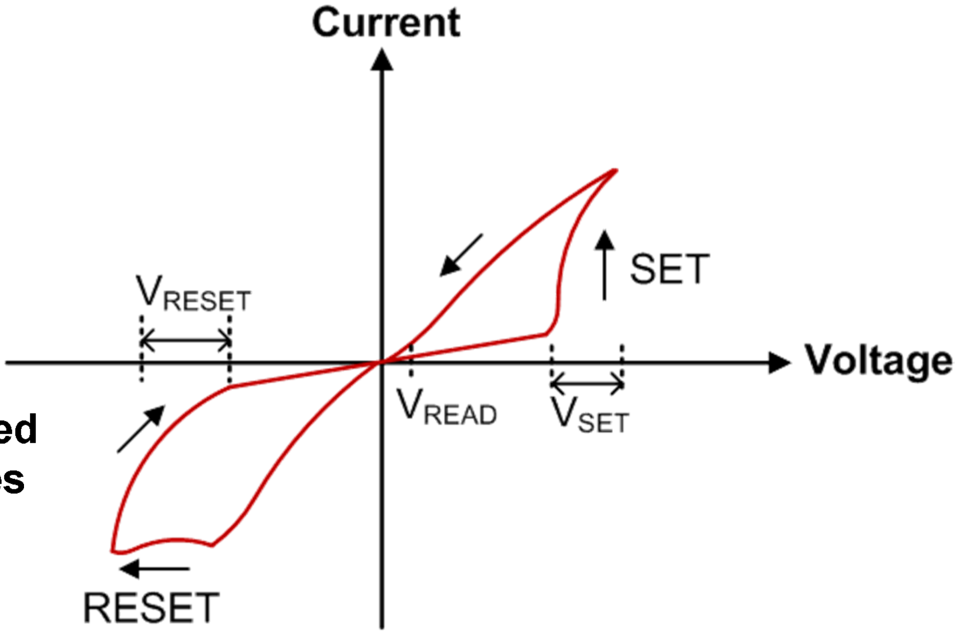
Metal Oxide: Bipolar Non-Filamentary

Valence Change ReRAM

TaO_x

- "Hysteresis loop" is simple method to visualize operation
  - (real operation through positive and negative pulses)
- Resistance Change Effect (polarities depend on device):
  - Positive voltage/electric field: low R – O^{-2} anions leave oxide
  - Negative voltage/electric field: high R – O^{-2} anions return
- Common switching materials: TaO_x, HfO_x, TiO_2, ZnO

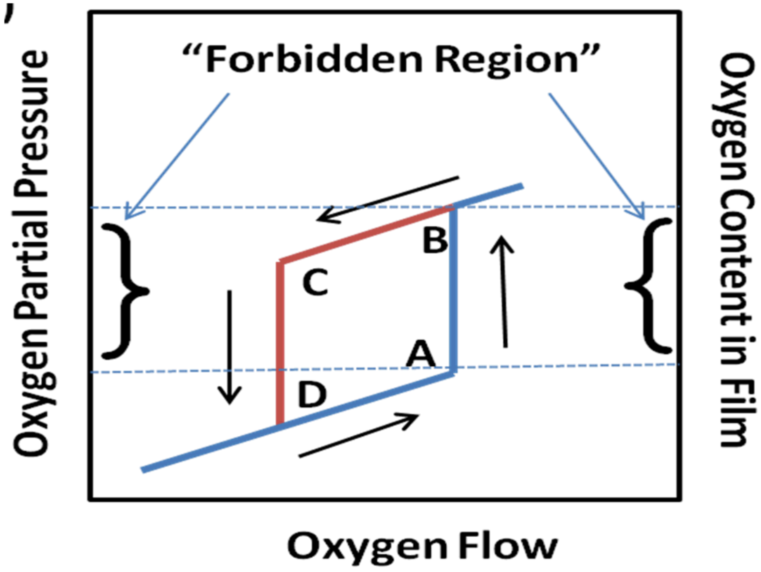




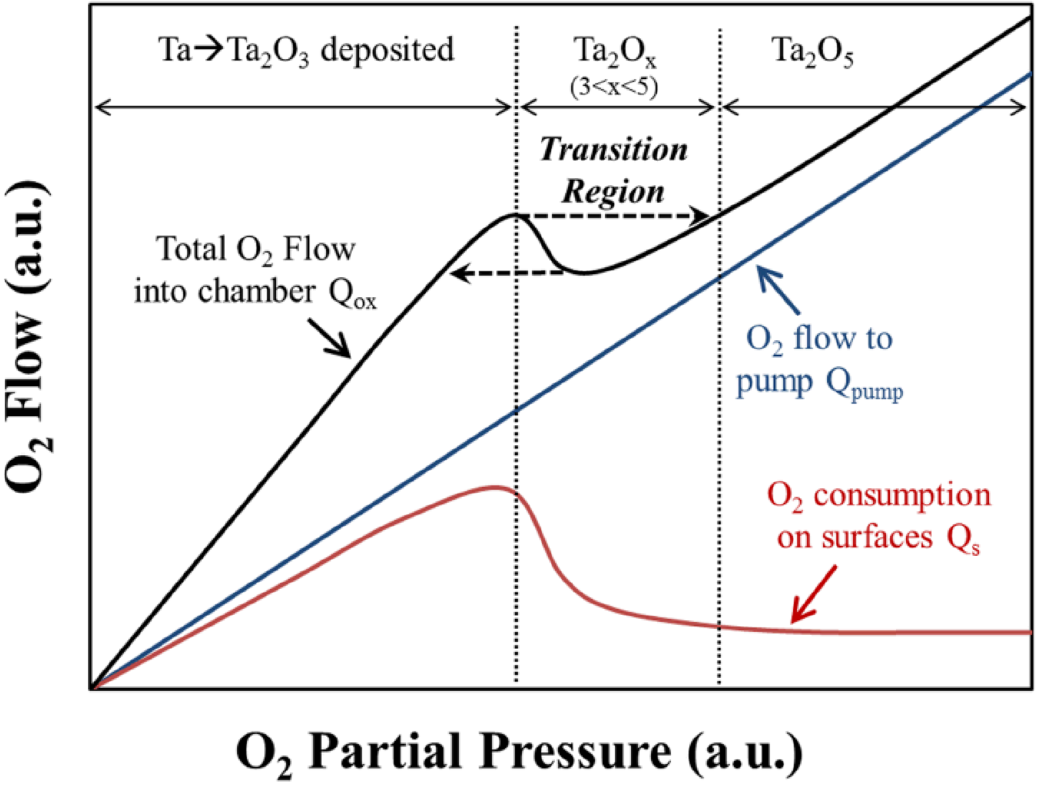
Switching Film Development:

Reactive Sputtering "Forbidden Region"

- One of the parameters that we vary is oxygen content
- Forbidden oxygen flow-pressure region occurs due to target poisoning
    - This is the region we need to be in to get ideal ReRAM stoichiometry



A.J. Lohn et al APL 103, 063502 (2013)

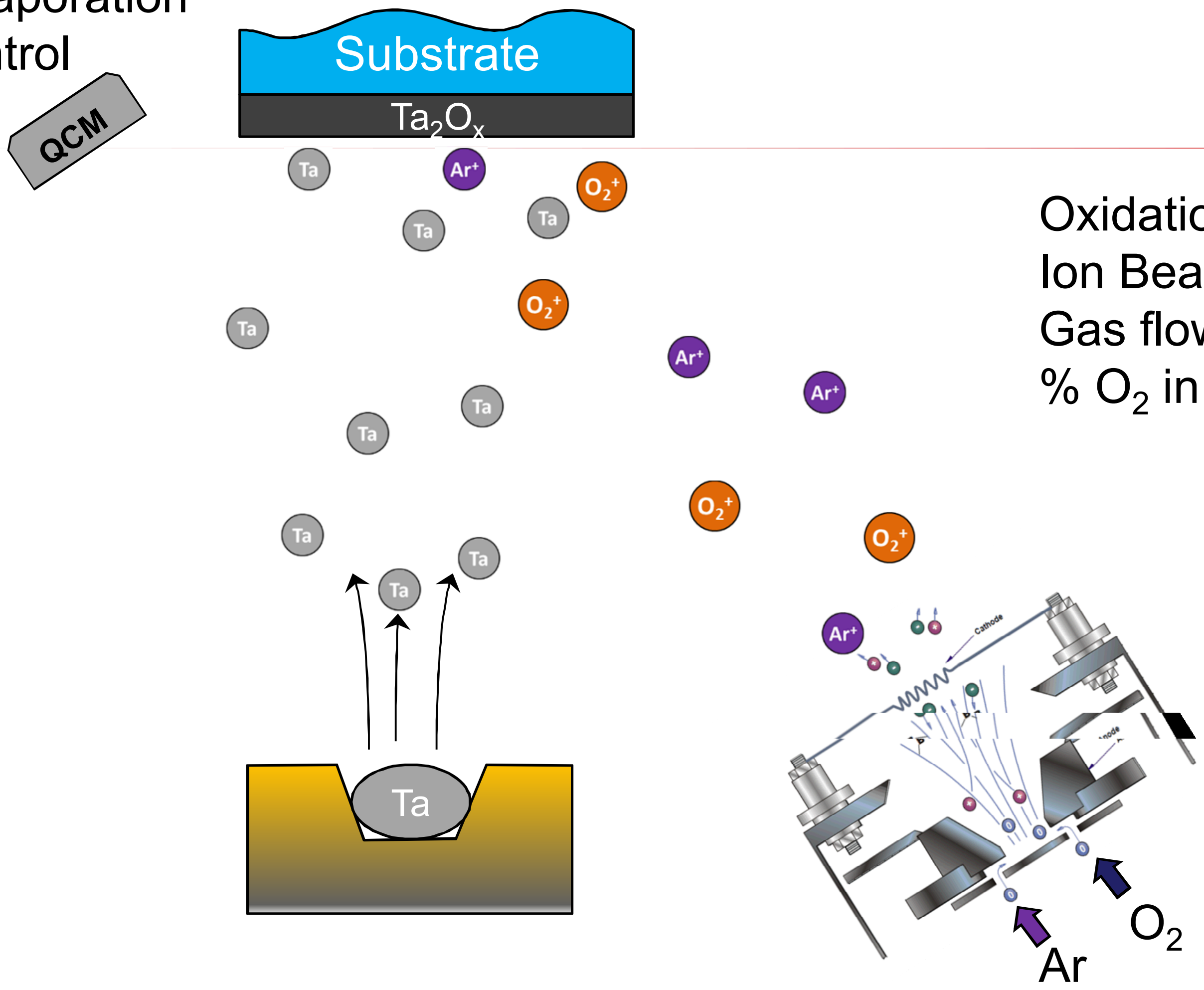


J.E. Stevens et al, J. Vac. Sci. Technol. A 32, 021501 (2014)

Ion Assisted Deposition (IAD)

Stoichiometry control

Metal Evaporation
Rate Control



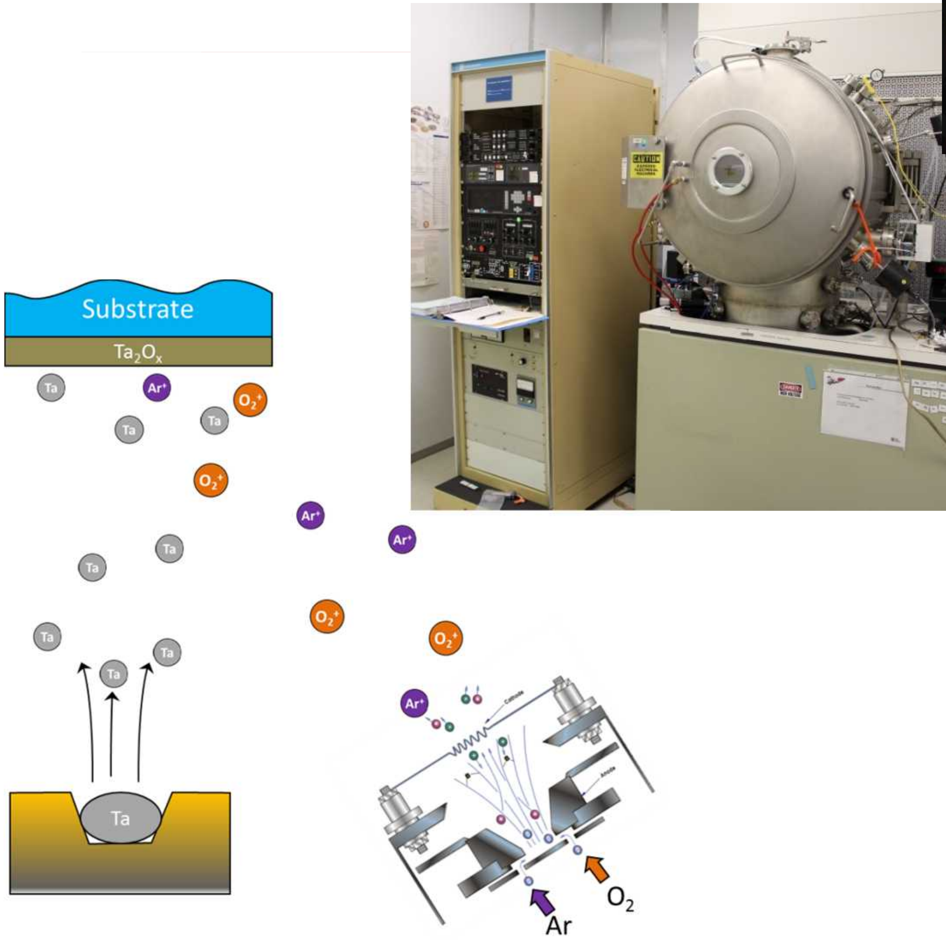
Oxidation control
Ion Beam Current
Gas flow rate
% O_2 in Ar/O_2 gas mixture

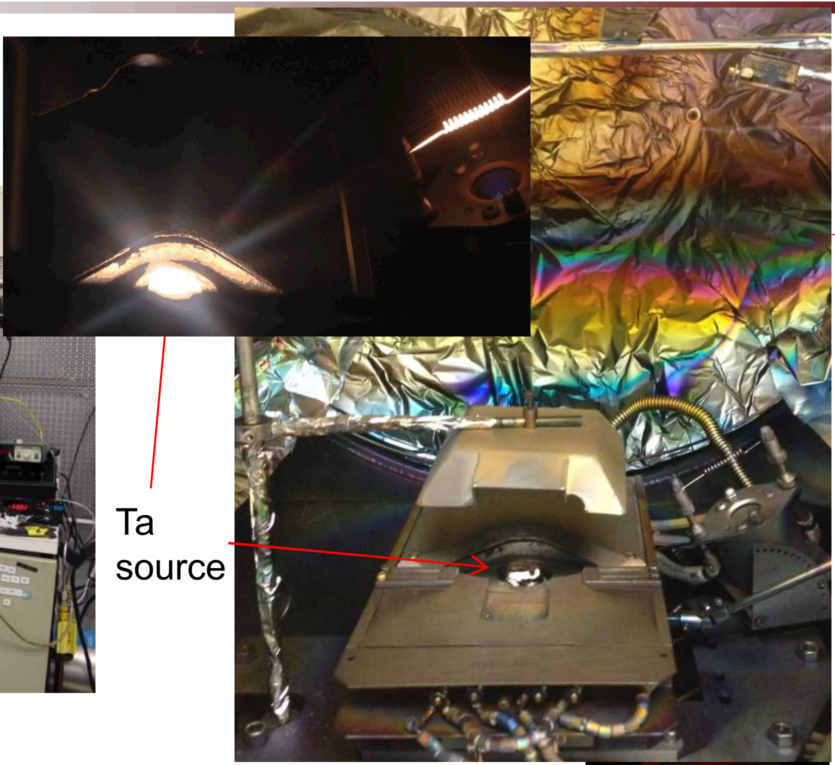
Ion Assisted Deposition (IAD)

System used for Ta_2O_x deposition

0.5Å/s growth rate

33cm source to substrate



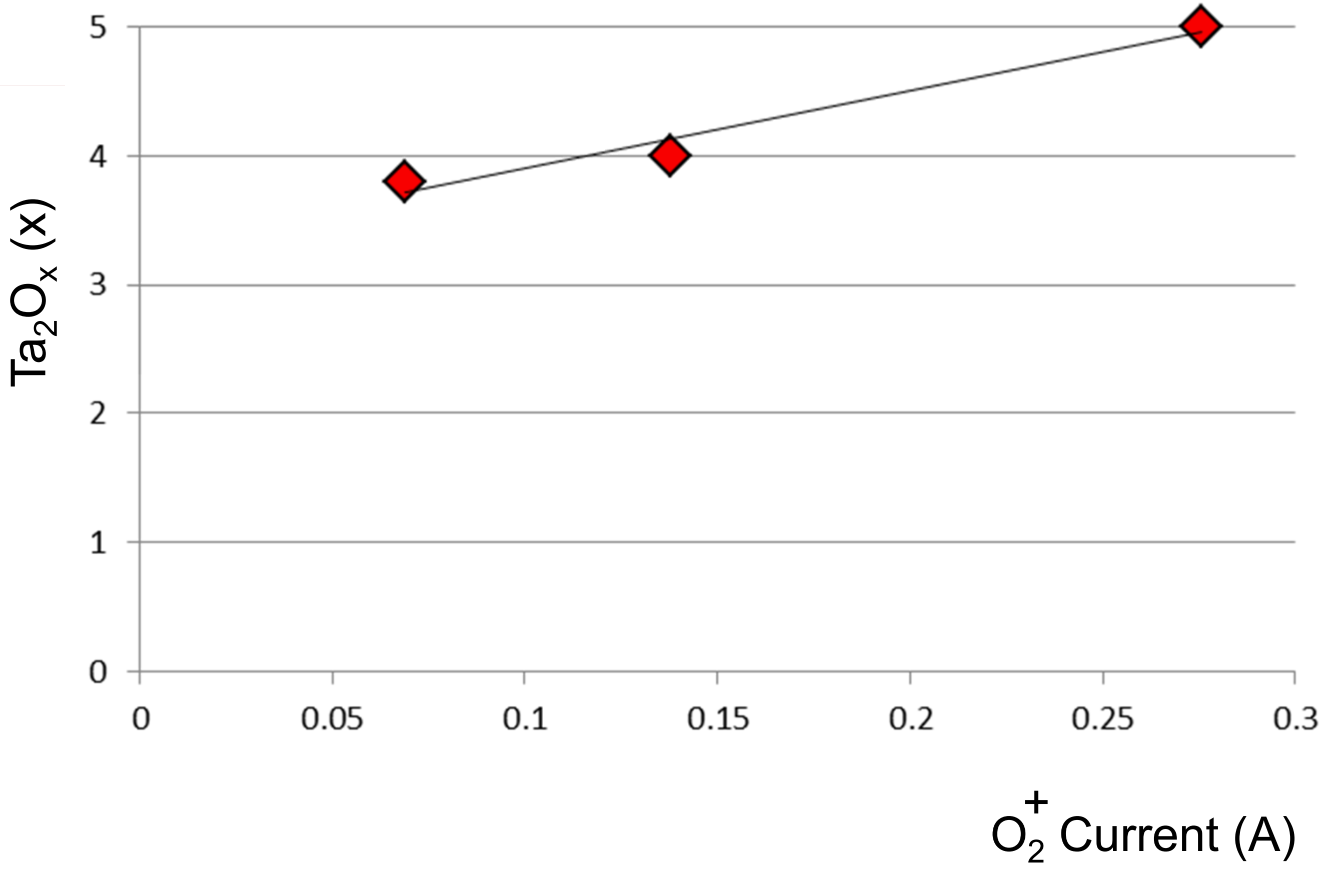


Four pocket rotary turret 10kV electron beam gun, and molten material during deposition

KRI EH200 ion gun
Gas mixture of Ar & O2 @ 18 sccm

Stoichiometry Control

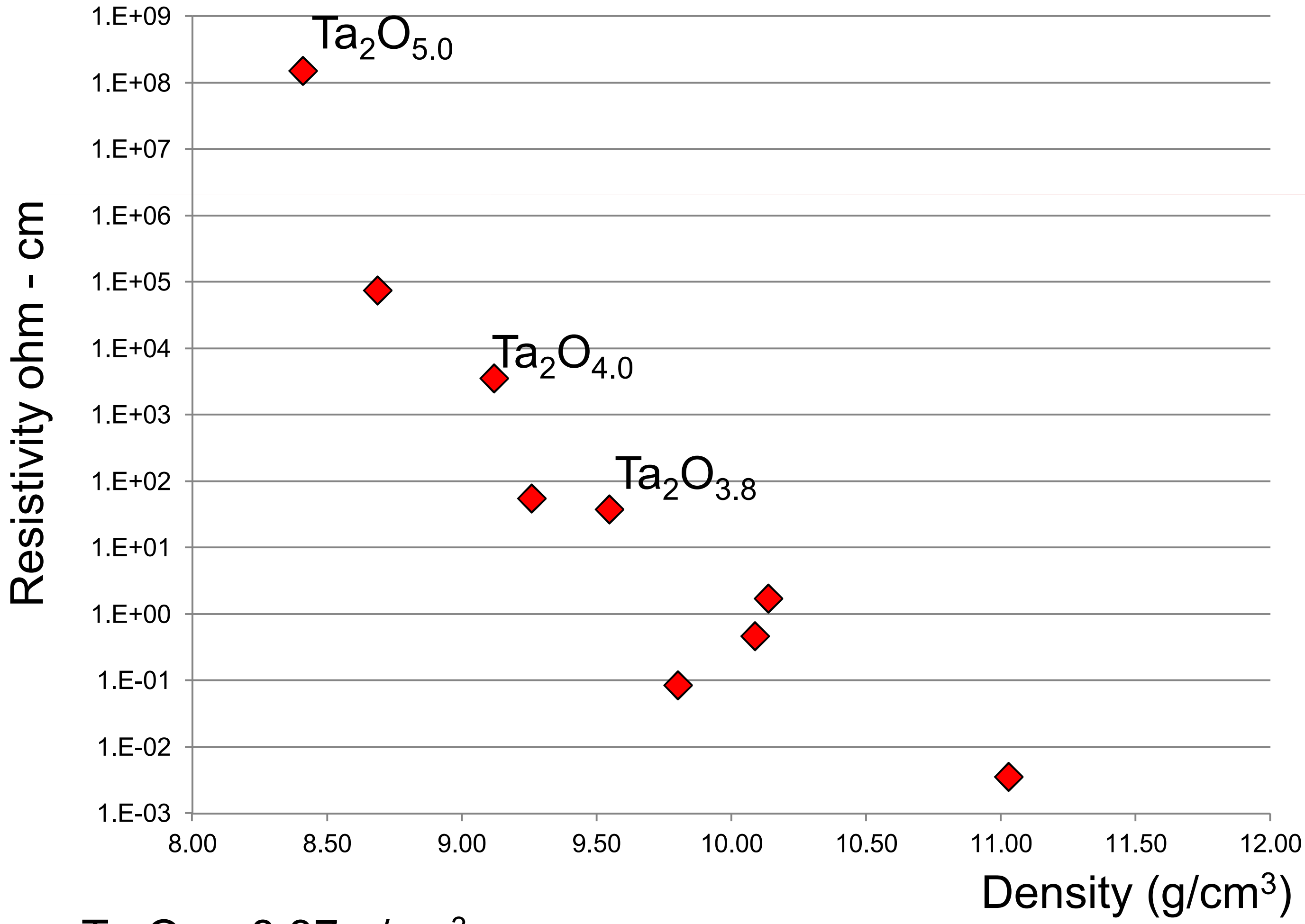
Ions and O_2 partial pressure



Ions contribute to system pressure and also react with growing film

IAD Film Resistivity

Density, resistivity and stoichiometry correlated

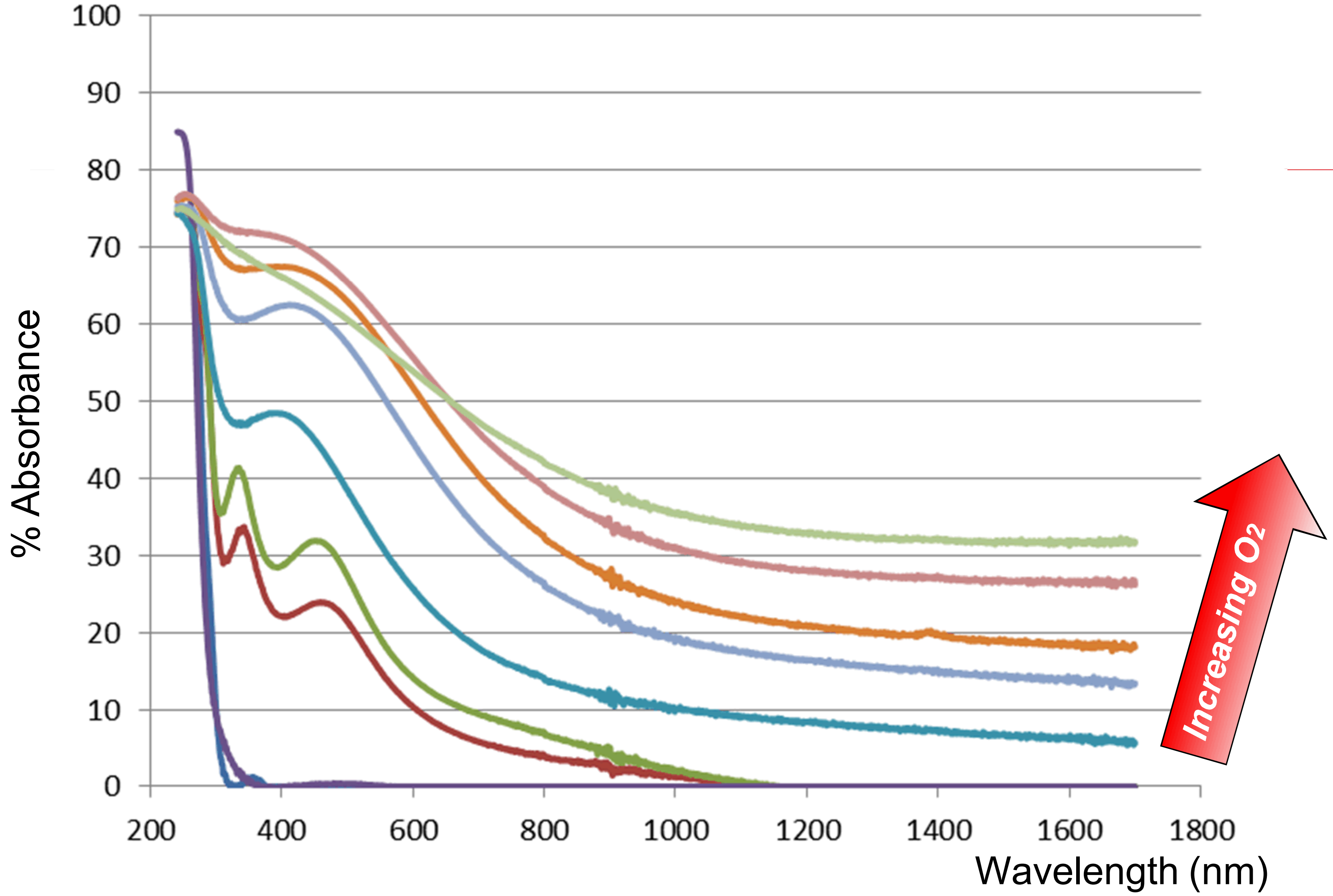


Ta_2O_5 = 8.37 g/cm^3

Ta = 16.69 g/cm^3

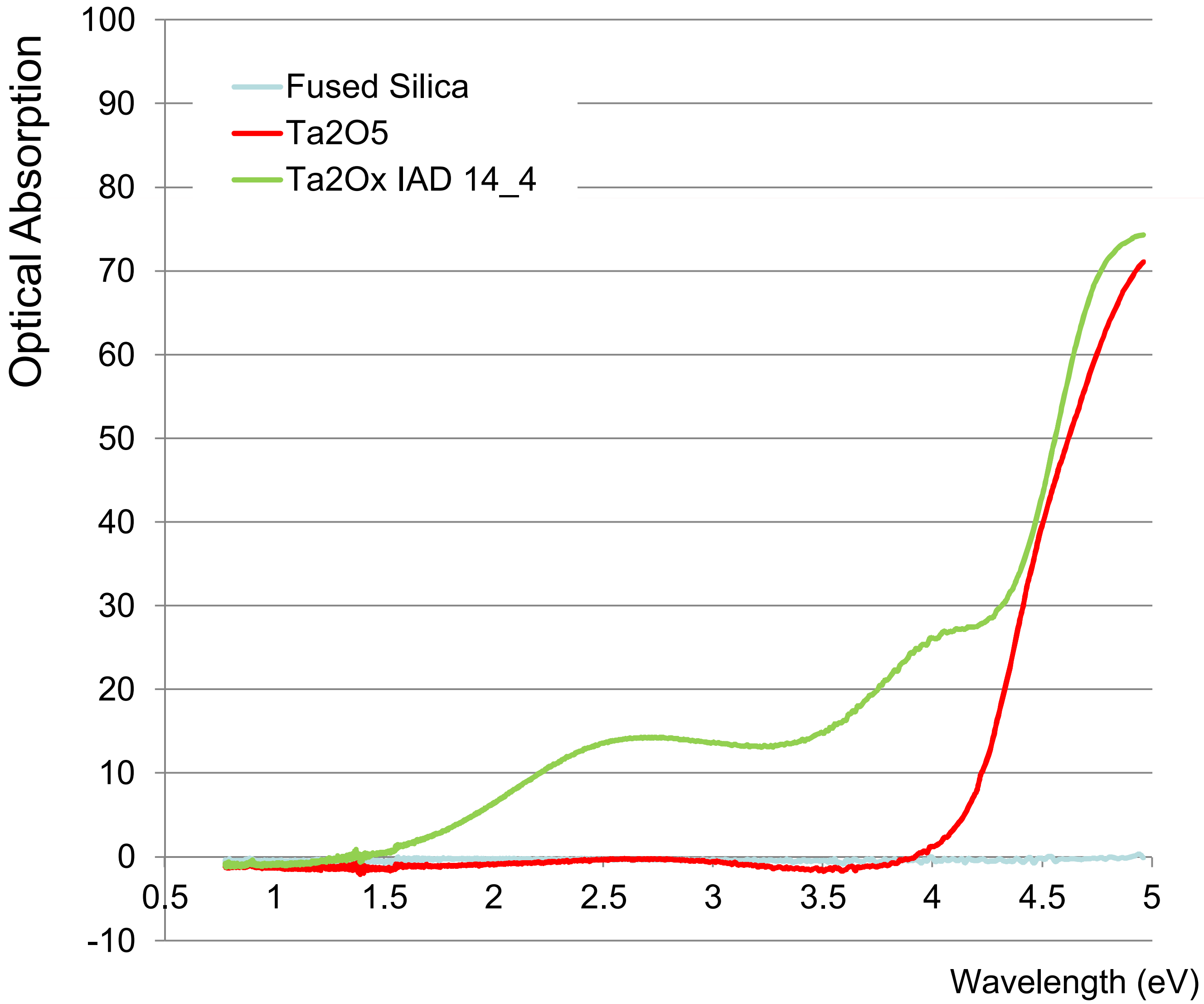
Optical Absorption

Spectrophotometry and Ellipsometry



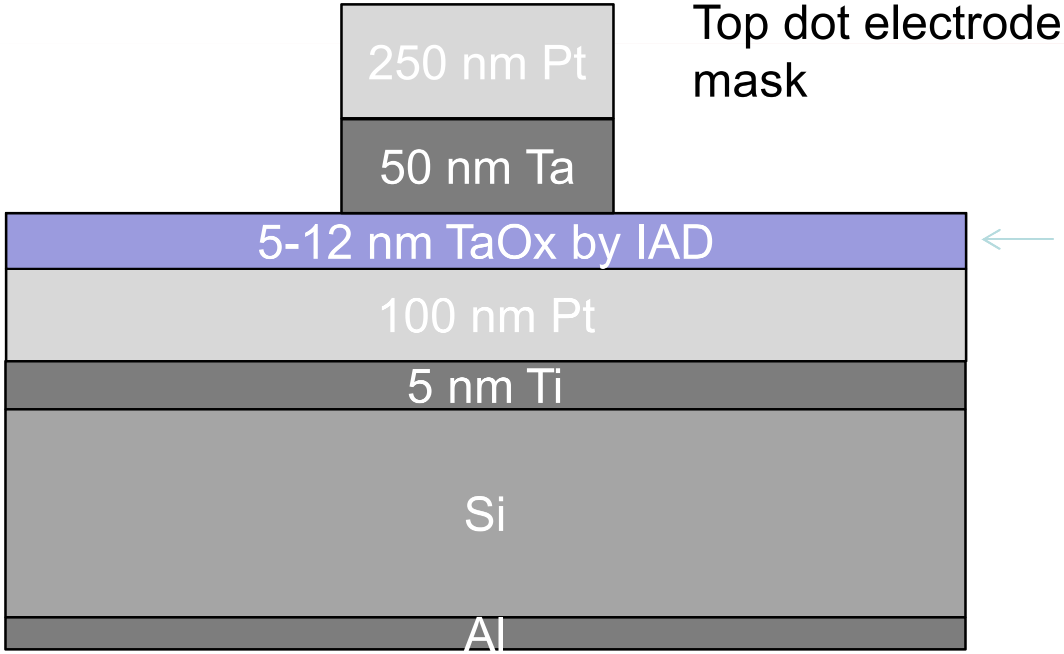
Band Gap Shift

Optical Absorbance



Test Device

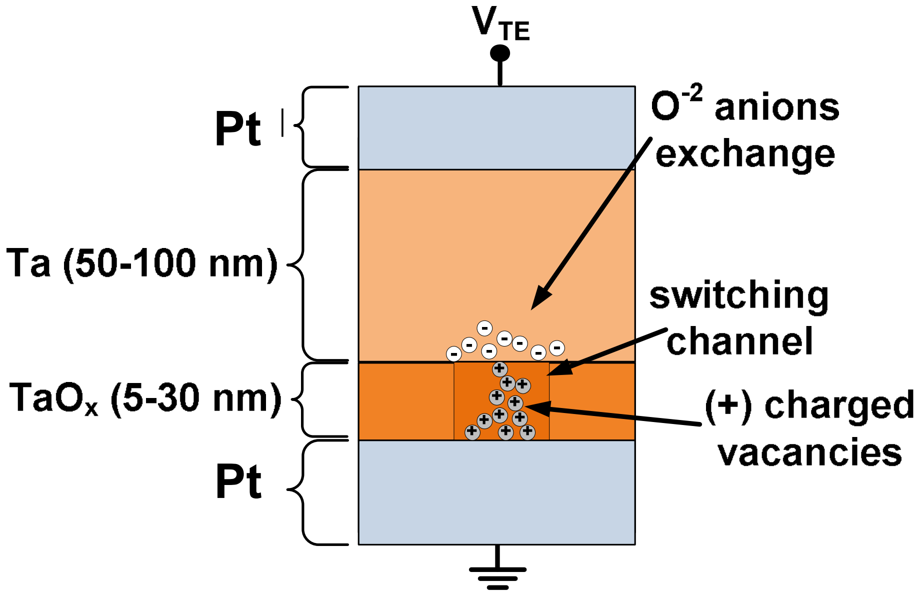
Layer structure

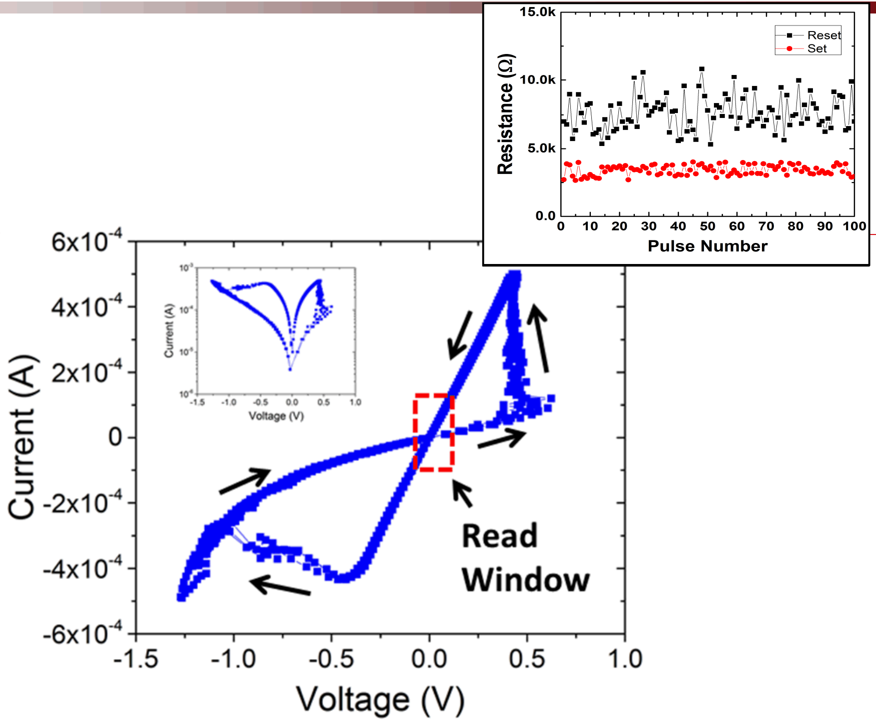


Top dot electrode created by shadow mask

Memristor I-V Characteristics

- Resistive RAM stores state in the form of resistance
- Applied current and voltage can change resistance state
  - Hysteresis loop
- Low voltages can read state
  - Read window



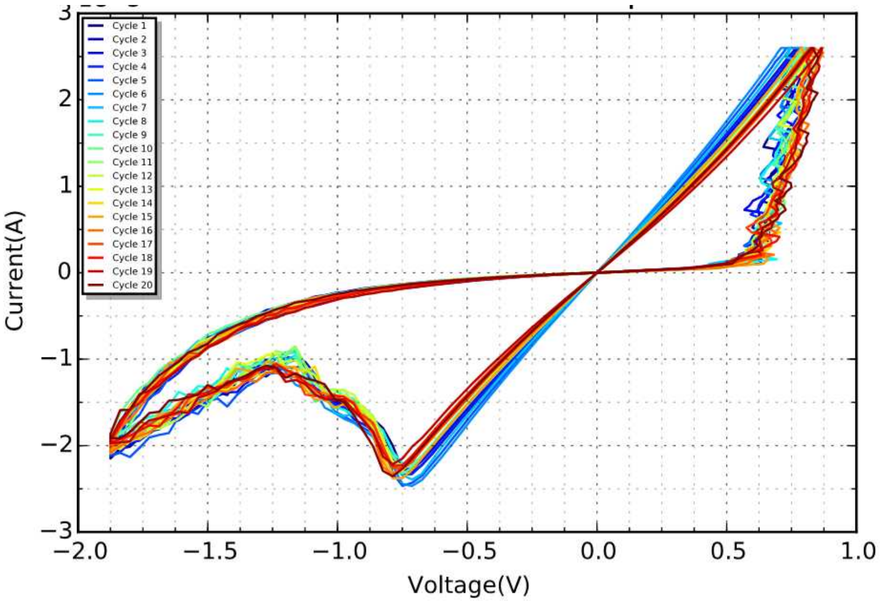


- Resistive switching
  - Oxygen vacancies
- TaO_x
  - Oxygen anions

IAD Improvements

High Repeatability

- More precise control over film thickness, stoichiometry and reduction in surface roughness
  - Improvements in yield and uniformity
  - Process adjustments easier due to higher repeatability



Ta_2O_x by IAD:

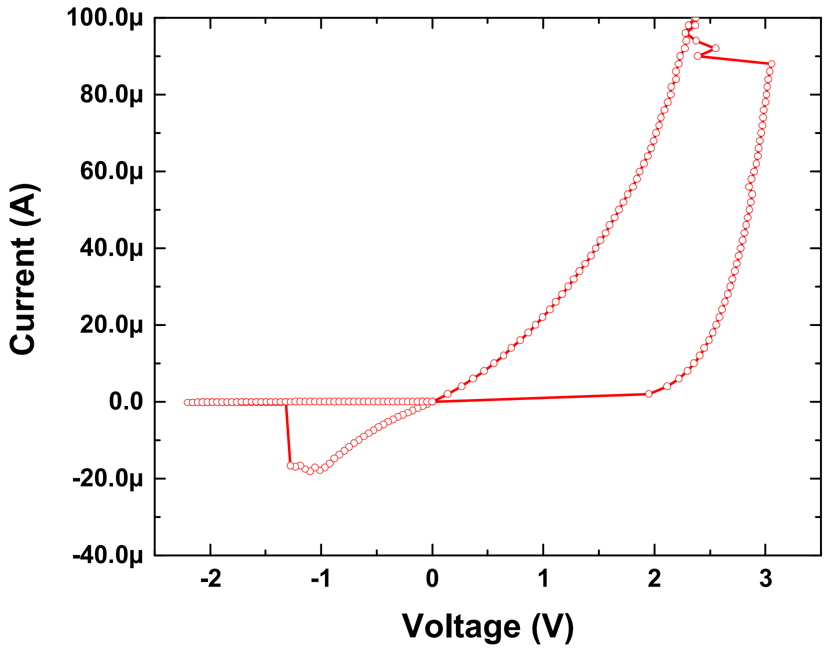
Device Results

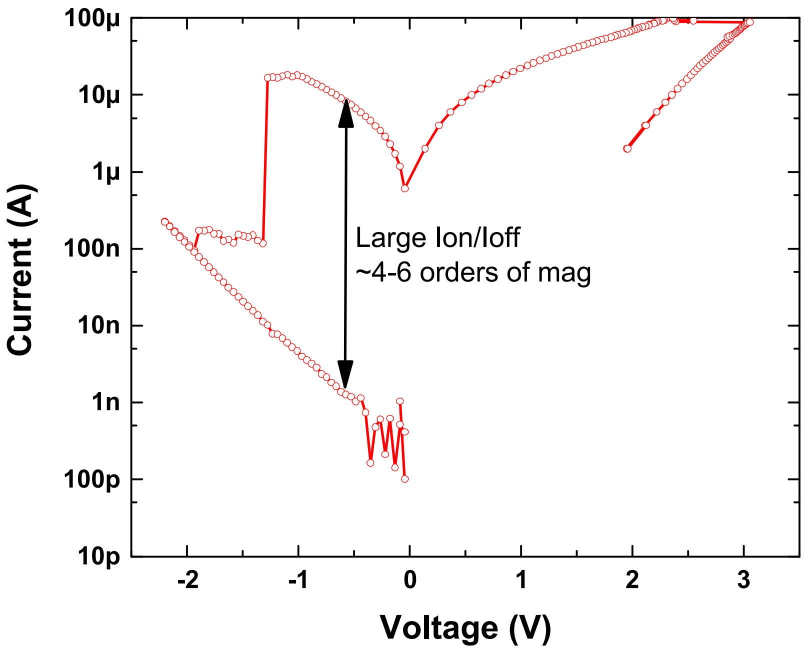
High yield; some batches 100% working devices

High resistance operation

Very high Roff/Ron ratio; as high as 6 orders of magnitude – critical for analog applications

Highly nonlinear IV curve – may provide self-selection





Large Ion/Ioff
~4-6 orders of mag

Conclusions

- Previous Metal Oxide ReRAM cells have relatively high cycle to cycle variability, which may significantly limit the resolution of an analog accelerator
- Ion Assisted Deposition of TaOx shows promise for significantly reducing that variability and improving device performance.
- IAD demonstrated excellent control over stoichiometry

Acknowledgements:
This project is funded by Sandia's Laboratory Directed Research and Development (LDRD) Program.
Carl Smith for film deposition.

Neural Algorithm Computing

Next Generation

Problem: neural algorithm training requires significant memory and logic interaction

What is the most efficient way to combine memory, logic and interconnects?

SRAM: on chip cache memory is limited to ~40MB digital (Intel E7)

- ns latency, max regardless of CPU, GPU or ASIC

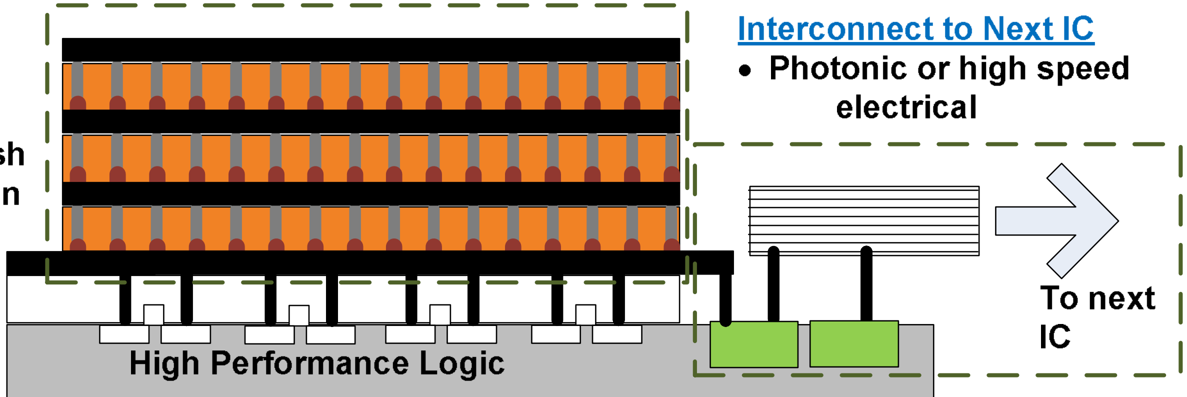
Off chip communication to DRAM costs >100 pJ/op, ~10ns latency

Resistive memory on chip: can be stacked to >TB/cm^2, >100 layers

- On chip access, <pJ per op and <1ns latency possible
- Terabit densities on single chip – on chip wiring is low energy!
- Sub 1V switching – minimal CV^{2}f loss (DRAM 2-5V)

Significant power savings using a ReRAM based HW accelerator

- **Example: Taha found 16x reduction in power, 6x improvement in perf per chip over SRAM**



ReRAM Layers:

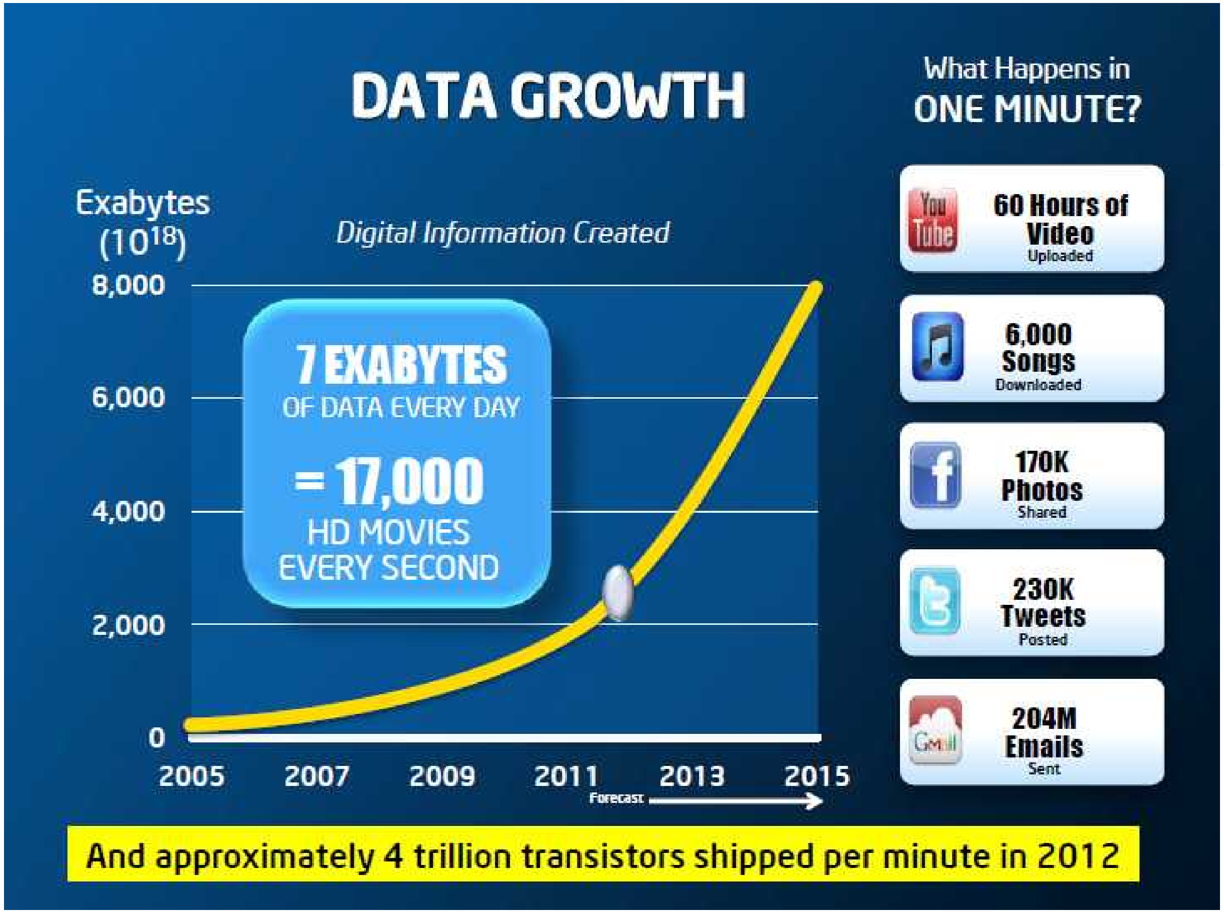
- Terabit cm^{-2} per layer
- Replaces DRAM & flash
- <1 pJ, <10 ns operation

Interconnect to Next IC

- Photonic or high speed electrical

Data Centers are Expanding Rapidly

Problems for analysis and energy consumption



And approximately 4 trillion transistors shipped per minute in 2012

M. Mayberry, Intel, IEEE VLSI Symposium 2012