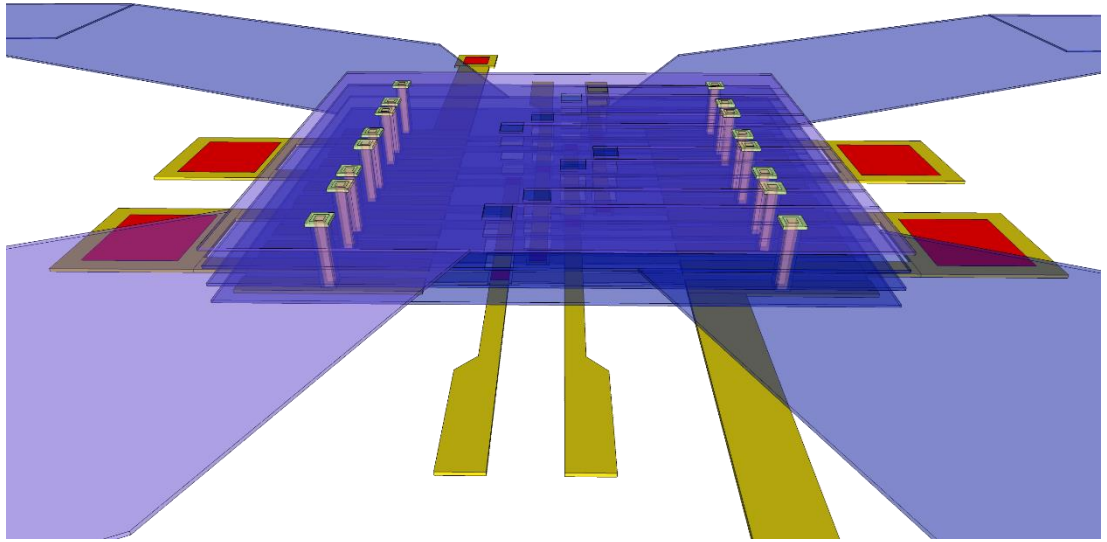
Neuromorphic Computing and Emerging Memory Technologies

KARL-MAGNUS PERSSON

Lecture Outline

Part I

- Introduction
- Neuromorphic computing
- Neural networks
- TrueNorth
- Contemporary memory systems
- ReRAM



Part II

- Overview of ReRAM technologies
- RRAM mechanics
- 3D RRAM integration
- RRAM research
- Outlook

The Era of Big Data and Recognition

NANO ELECTRONICS GROUP

Where we are headed

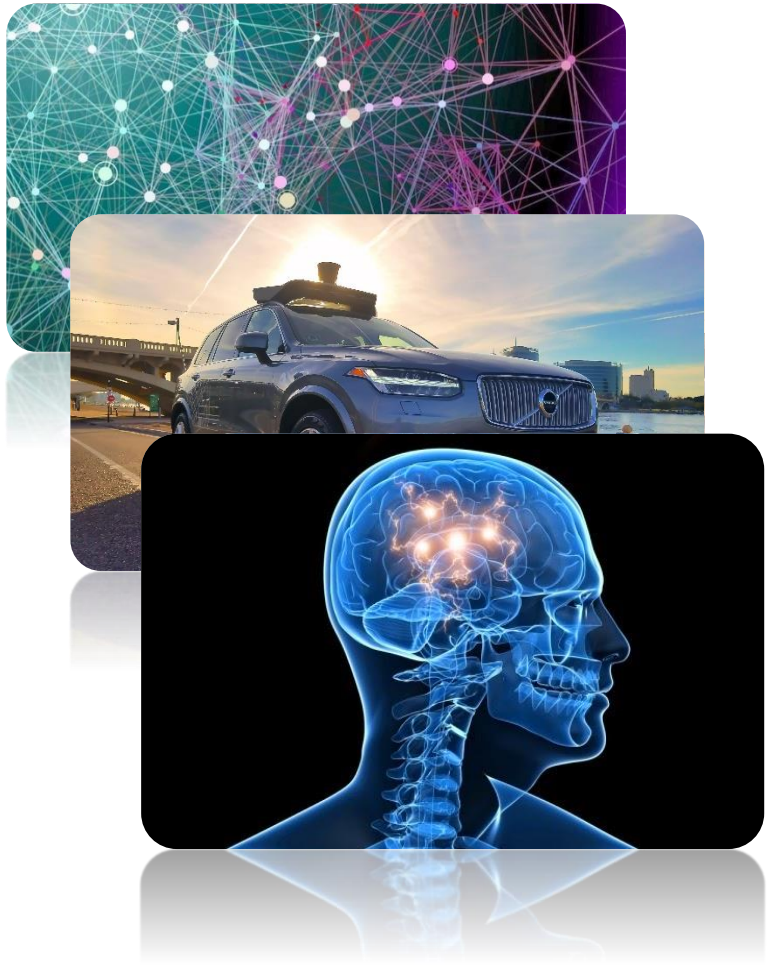
- Machine learning/AI - Improve and replace
- Early major disruptor - Autonomous driving

Hardware Challenges

- Moore's law has halted
- NVM technologies - 10,000x slower than computing
- Memory and computational circuits on different chips

Possible solutions

- Component improvement stagnated → creative systems
- Model the human brain → neuromorphic computing
- Replace silicon with more advantageous semiconductors

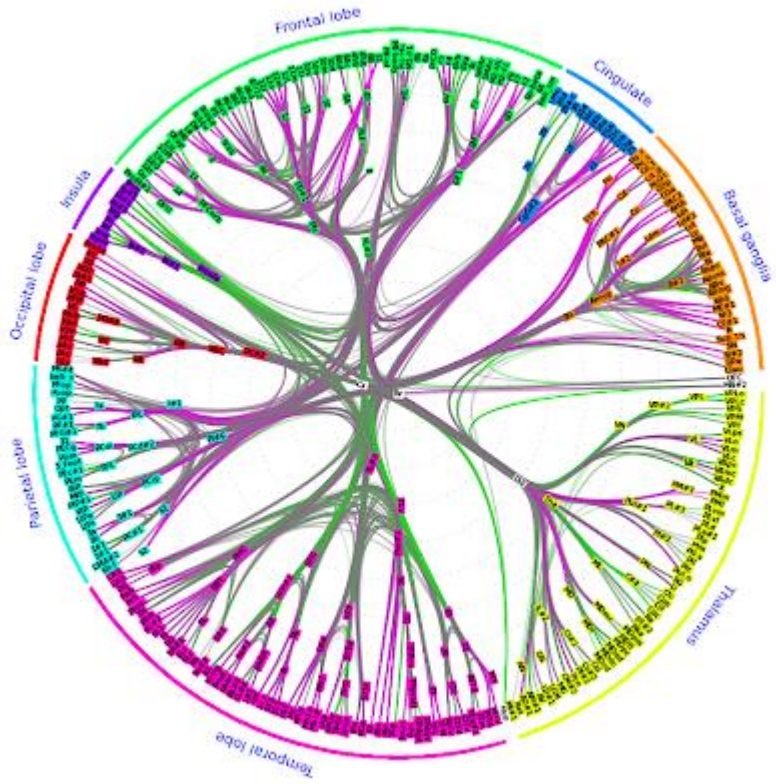


Neuromorphic Computing

- **von Neumann Computing**
  - Computations according to set instructions
  - A controller steers data between the CPU and the memory
  - Sequential, clock-based, high precision

- **Artificial biological systems**
  - Neuro-biological architectures
  - Mimic brain function with large scale circuits
  - Solving problems with synaptic networks

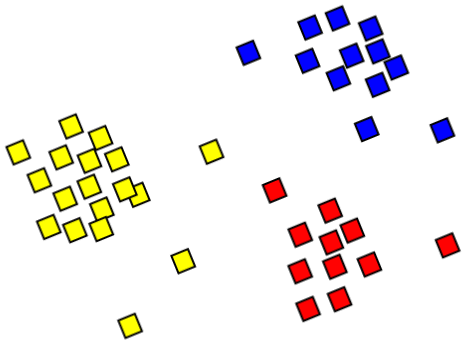
- **New ways of computing possible**
  - Parallelism over speed
  - Event driven computation instead of clock-cycles
  - Self-learning from prior experience
  - Lower precision, certain error rates acceptable



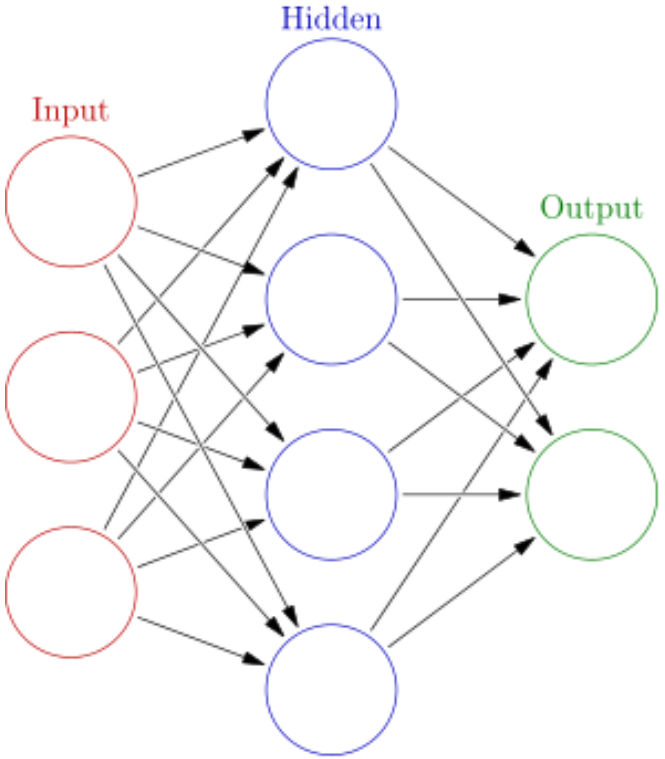
Network diagram of a human brain

Neural Networks

- **Synaptic networks of artificial neurons**
  - Modeled as inputs, outputs, and intermittent hidden layers
  - Neurons have both memory and switching capabilities
  - Switches have different thresholds
  - Summation of inputs are weighed differently
  - Errors can be propagated backwards
  - Iterations of adjusting the network accomplishes learning



Data categorization

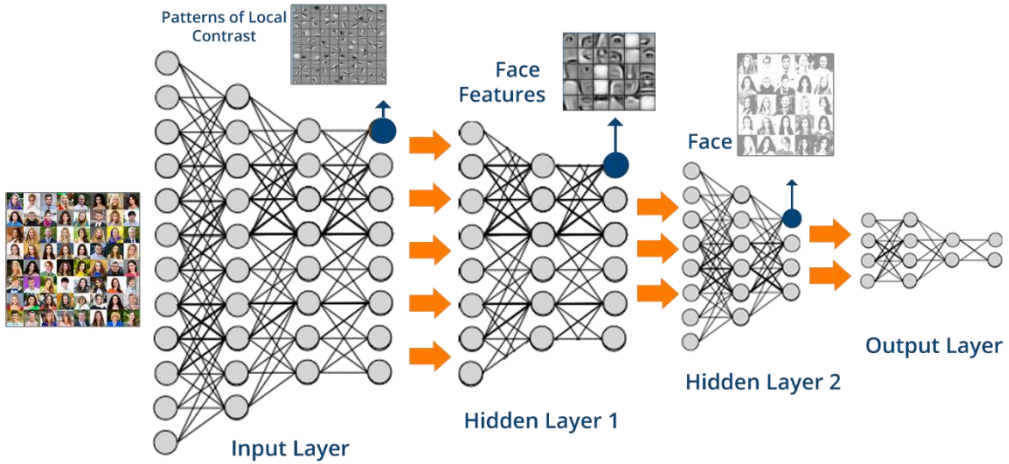


Neural network

Deep Neural Networks

- **Deep learning**
  - Subset of machine learning
  - Departure from pre-defined, task-specific algorithms
  - Supervised, semi-supervised, and unsupervised learning
  - Working with large sets of data
  - Efficient way to solve multivariable problems

- **Hardware implications**
  - Iterative re-programming of memory
  - Performance limited by read/write of NVM
  - Separate compute and memory circuitry infer large inefficiencies

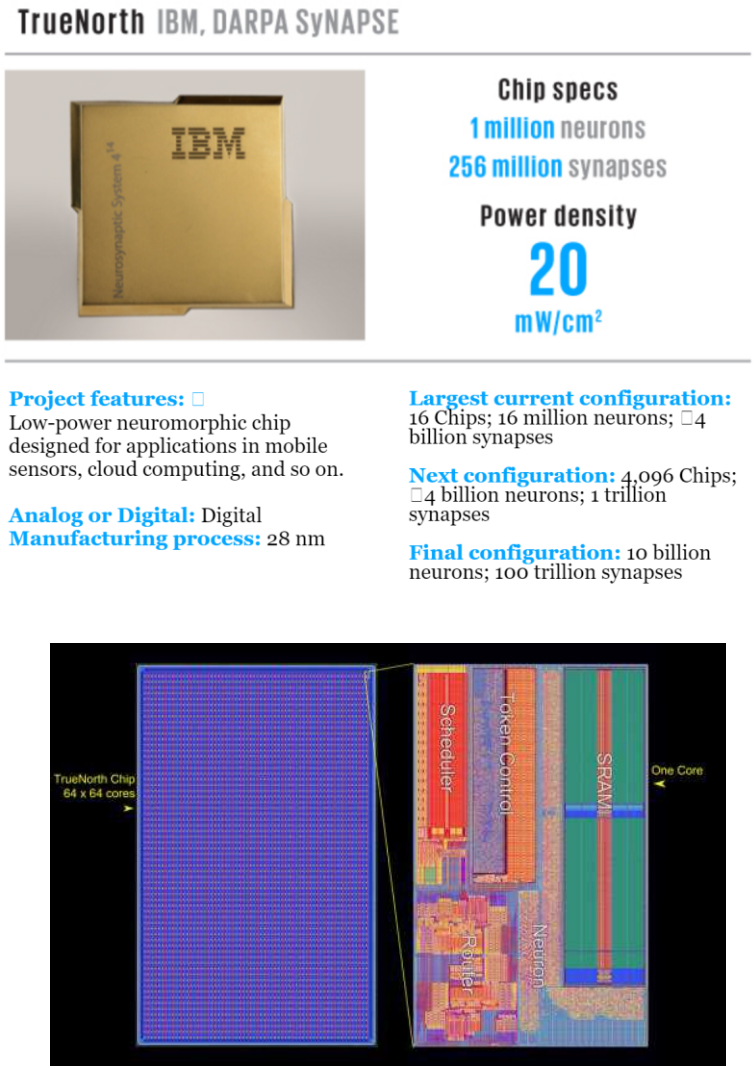


Deep neural network for face recognition

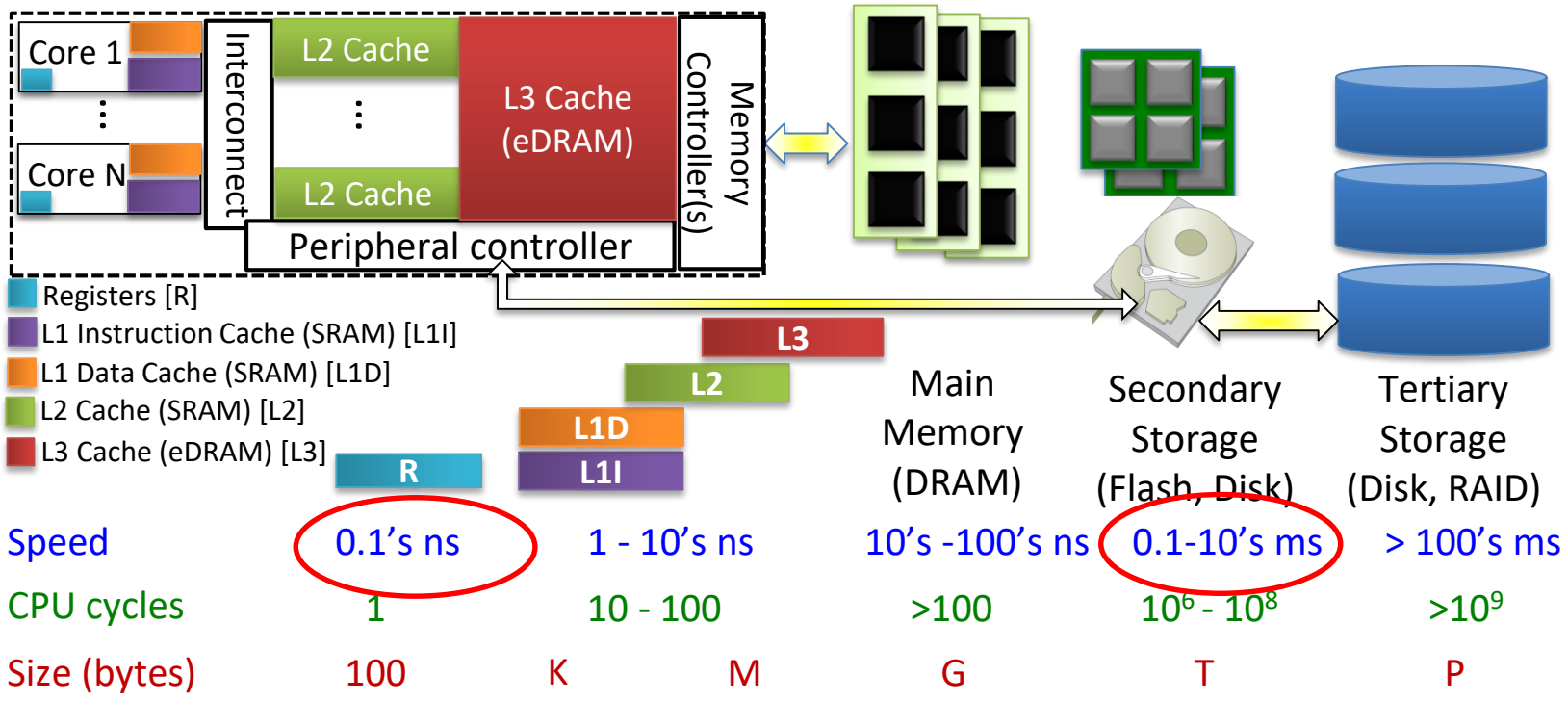
SyNAPSE and IBM TrueNorth

- **SyNAPSE**
  - DARPA funded program
  - Final aim is to replicate the neuron network in the human brain (100B neurons), using only 1 kW
  - Simulation of a brain like system currently takes 1.5M CPUs and 8 MW to run at 0.1% of the speed of the brain

- **TrueNorth**
  - IBM funded research chip
  - Fully digital, 1M neurons, 65 mW total (6,5 kW for 100B)
  - TrueNorth has 4096 computational cores
  - Separate compute circuitry and memory (SRAM)



Memory Hierarchy

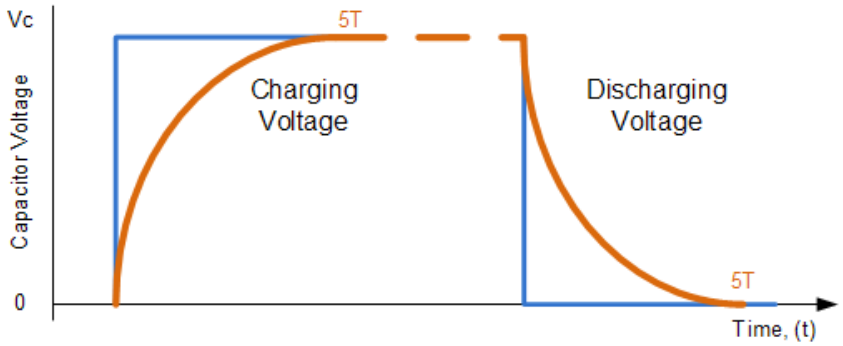


H.-S. P. Wong, S. Salahuddin, *Nature Nanotech* (2015)

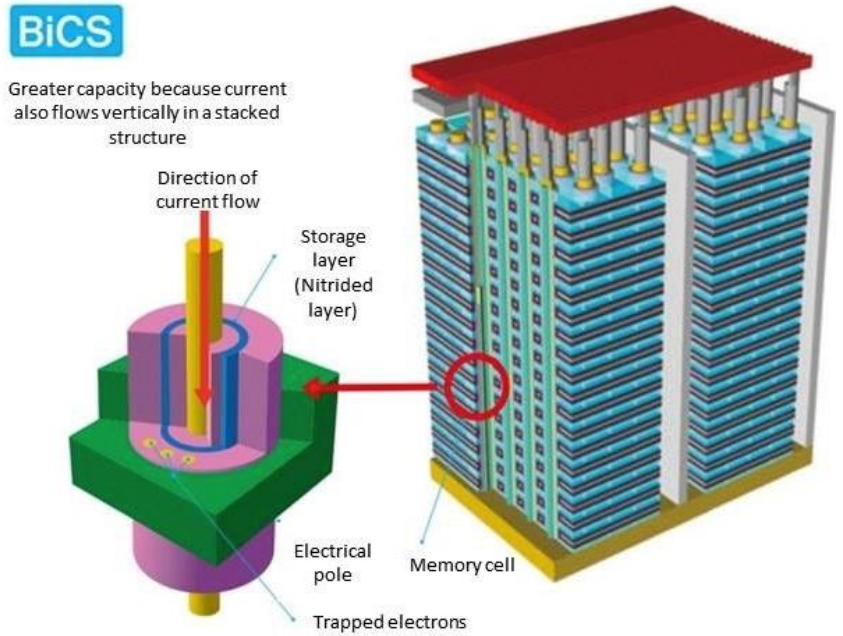
Limitations in NAND Flash

- **The upside**
  - 3D integrated with 128 layers in the near future
  - Minimal feature size down to 5 nm

- **The downside**
  - Read in ns but write in ms
  - Further scaling of the dielectric leads to electron leakage



RC delay



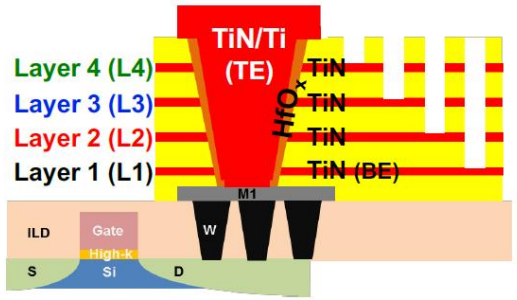
Schematic structure of planar NAND Flash

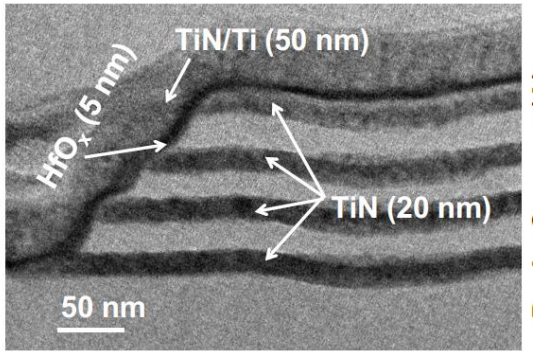
https://www.storagenewsletter.com

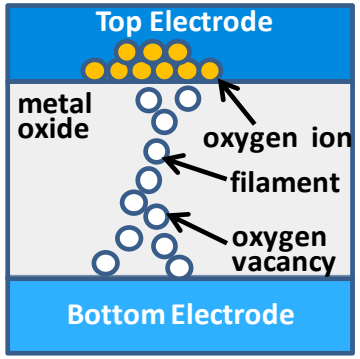
3D ReRAM – A Promising Candidate

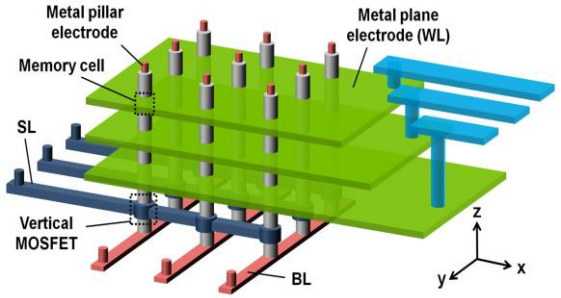
NANO ELECTRONICS GROUP

- **Power**
  - 1 – 2 V
  - 10's nA – 10's µA
- **Speed**
  - 10 ns read/write
- **Endurance**
  - > 1^9 cycles
  - 1^{12} cycles at device level
- **Scaling**
  - 1T1R ~ $6F^2$
  - F < 5 nm
- **3D ReRAM**
  - 128 layers
  - 64 Tb per chip







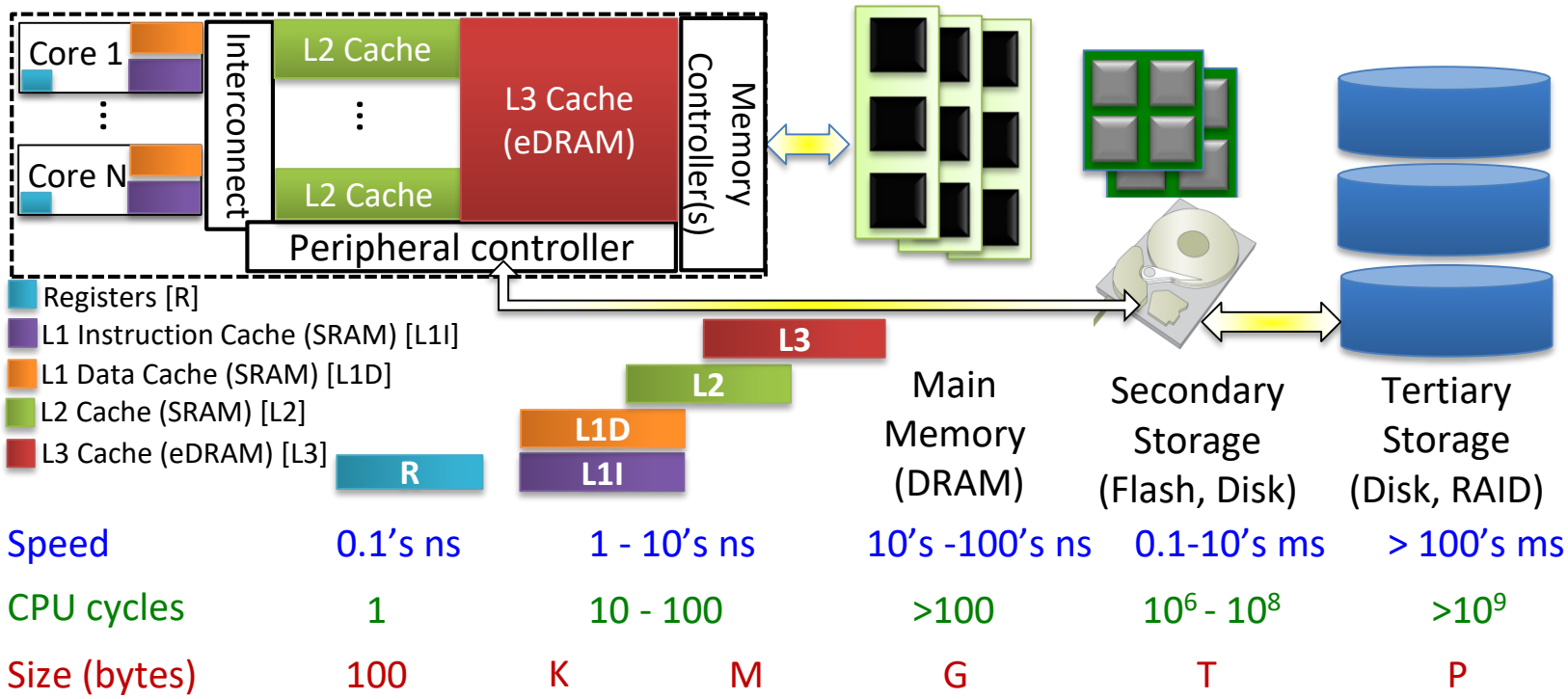


Stanford is a leading institution in the field of ReRAM under the supervision of Prof. H.-S. Philip Wong

3D ReRAM – A Promising Candidate

NANO ELECTRONICS GROUP

- **Power**
  - 1 – 2 V
  - 10's nA – 10's µA
- **Speed**
  - 10 ns read/write
- **Endurance**
  - > 1^9 cycles
  - 1^{12} cycles at device level
- **Scaling**
  - 1T1R ~ $6F^2$
  - F < 5 nm
- **3D ReRAM**
  - 128 layers
  - 64 Tb per chip

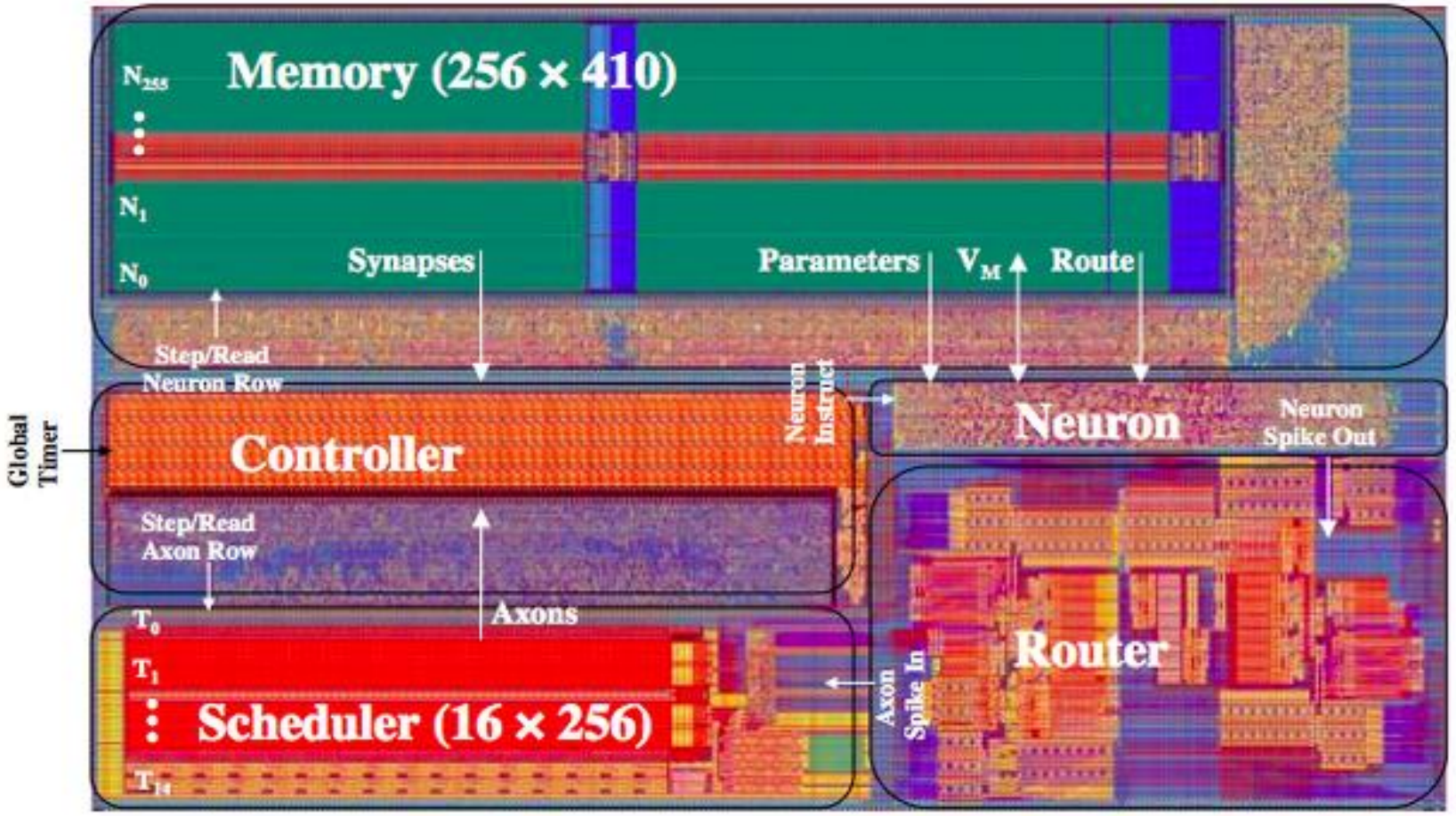


H.-S. P. Wong, S. Salahuddin, *Nature Nanotech* (2015)

3D ReRAM – A Promising Candidate

- **Power**
  - 1 – 2 V
  - 10's nA – 10's µA
- **Speed**
  - 10 ns read/write
- **Endurance**
  - > 1^9 cycles
  - 1^{12} cycles at device level
- **Scaling**
  - 1T1R ~ $6F^2$
  - F < 5 nm
- **3D ReRAM**
  - 128 layers
  - 64 Tb per chip

SRAM ~ 120-140 F^2
1 layer ReRAM 20x smaller

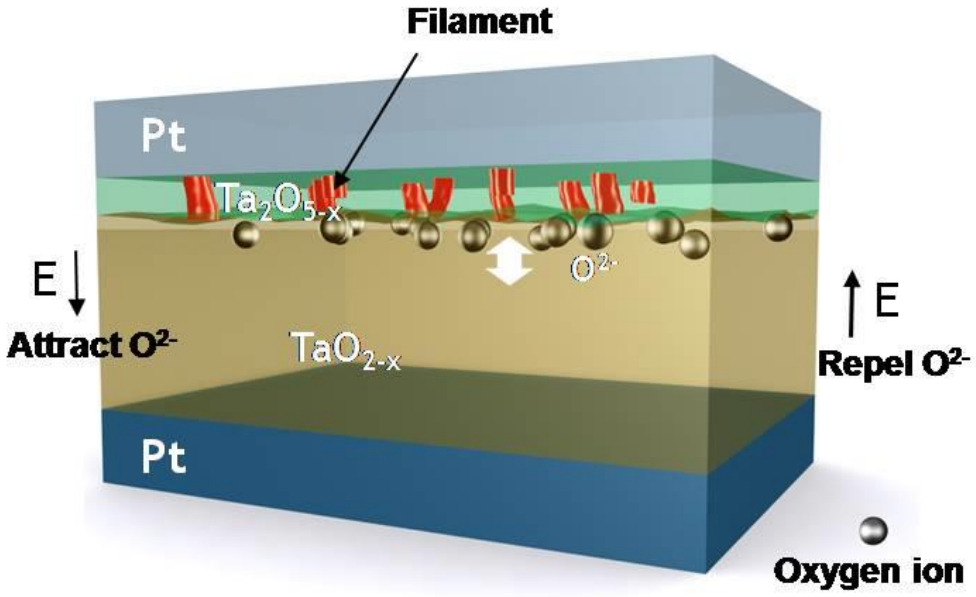


TrueNorth chip layout

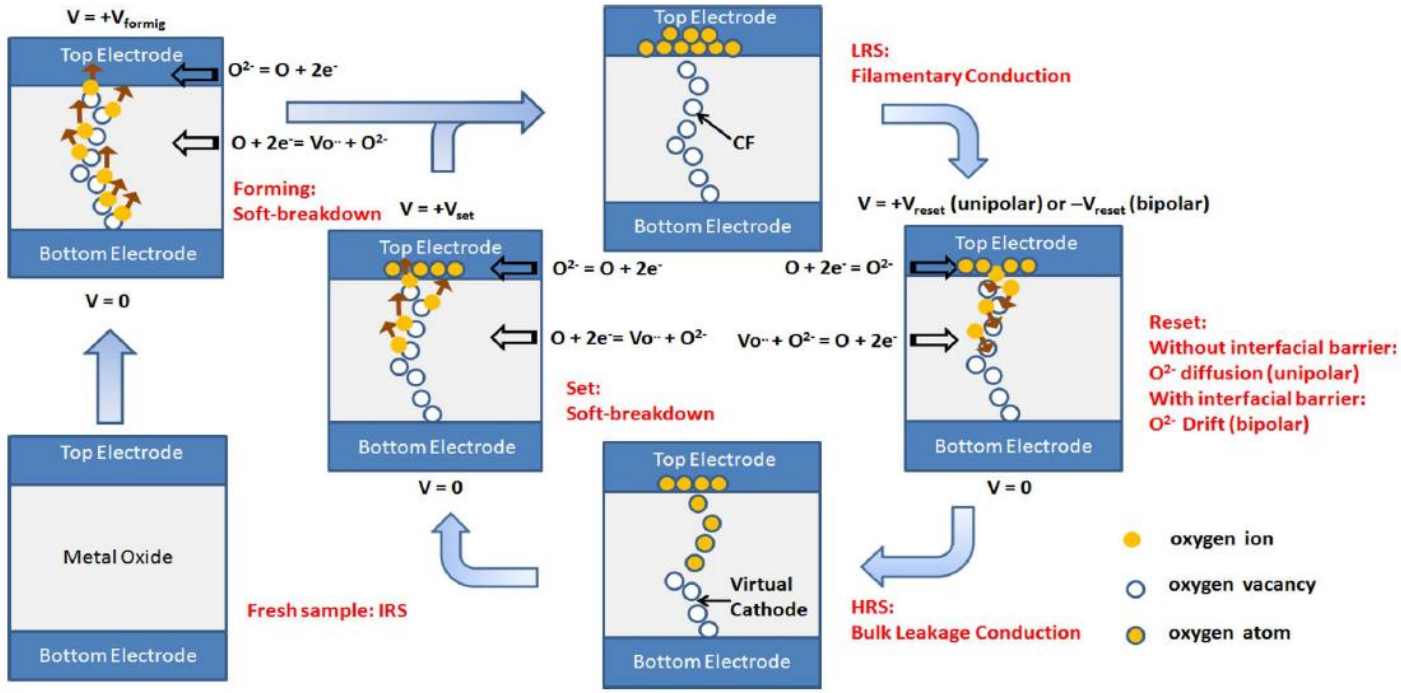
Overview of different ReRAM Technologies

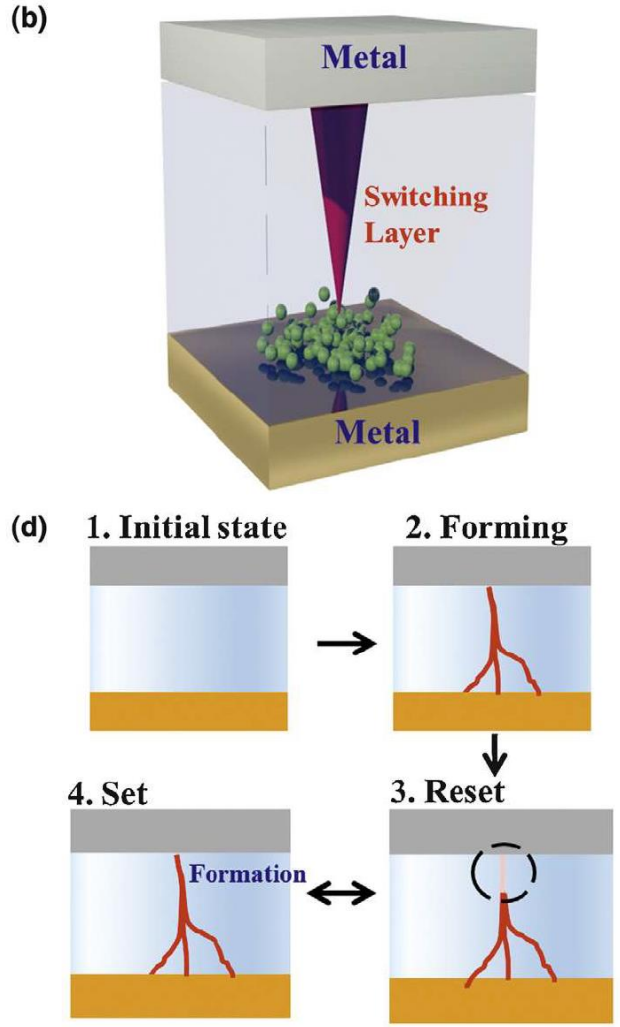
NANO ELECTRONICS GROUP

- **oxRRAM (RRAM)**
  - Anode filament
  - Oxygen vacancies form conductive path
  - Mobile oxygen ions responsible for switching
  - Simplest fabrication, currently most studied
  - 3D compatible
- **CBRAM**
  - Similar structure to RRAM
  - Cathode filament (conductive bridging)
  - Metal ions form a conductive path
  - 3D compatible
- **PCRAM**
  - Phase change memory where a flash heating switch dielectric film between amorphous and crystalline state
  - Retention questionable
- **STT-MRAM**
  - Spin-transfer-torque magnetic RAM
  - Changing orientation of spin changes the conductivity
  - Potentially very fast and energy efficient

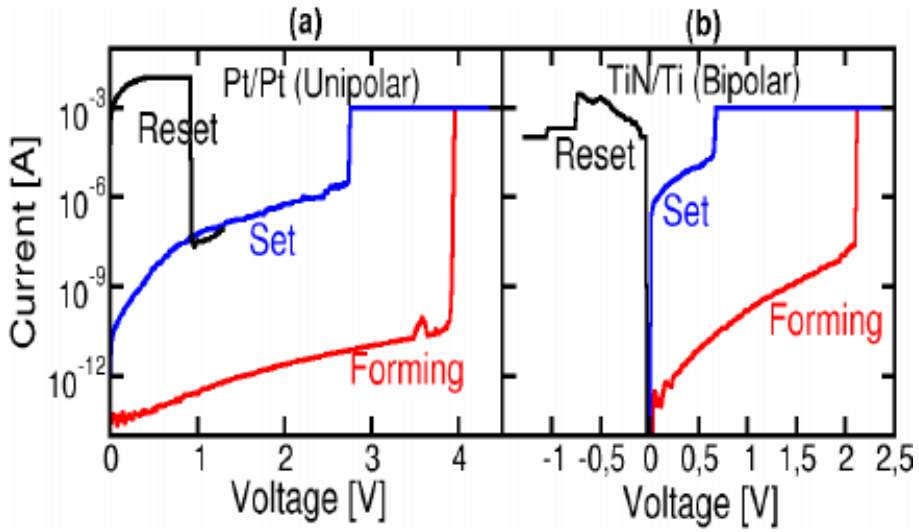


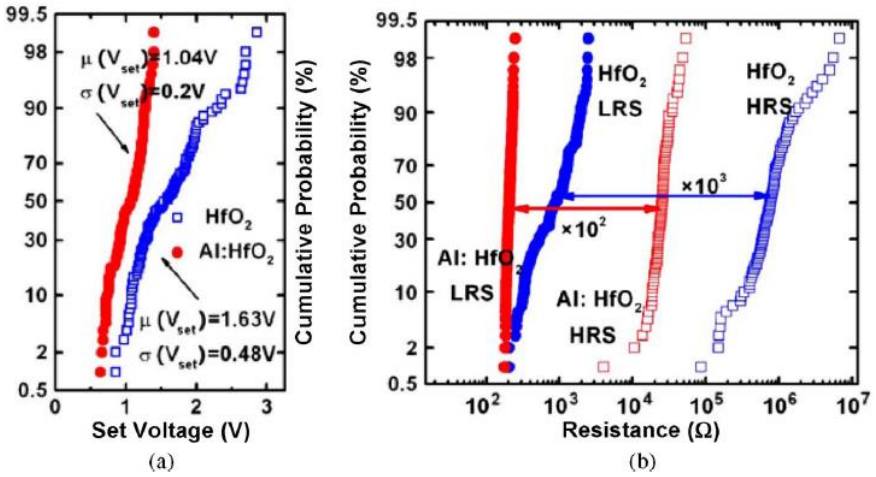
RRAM Mechanics

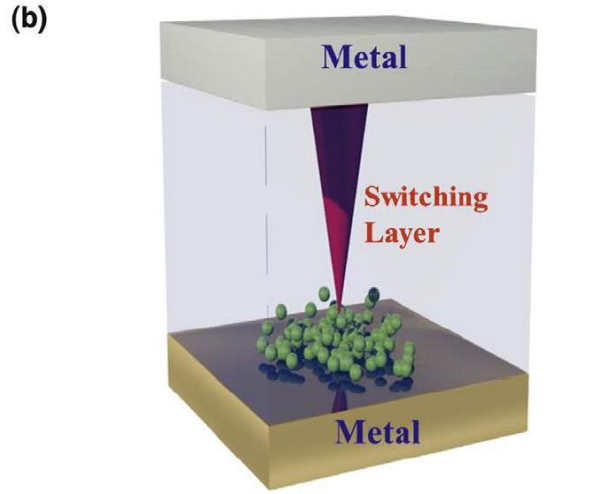


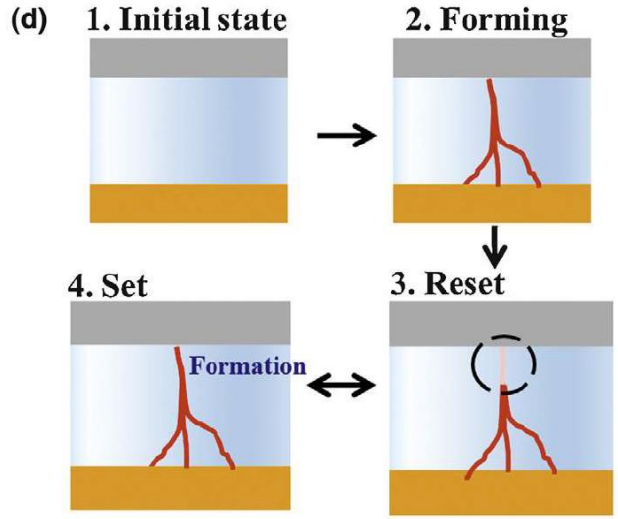


RRAM Mechanics





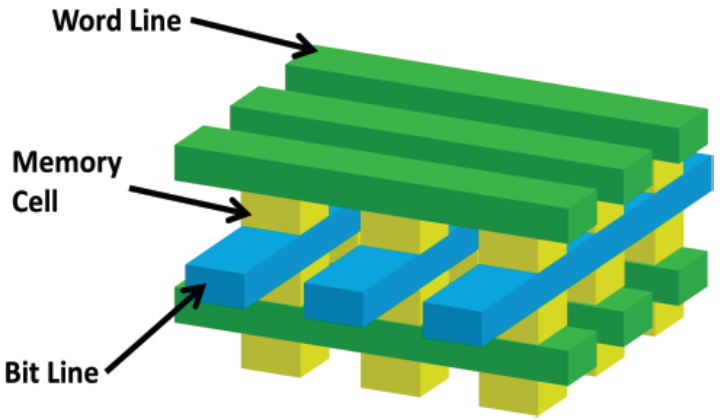


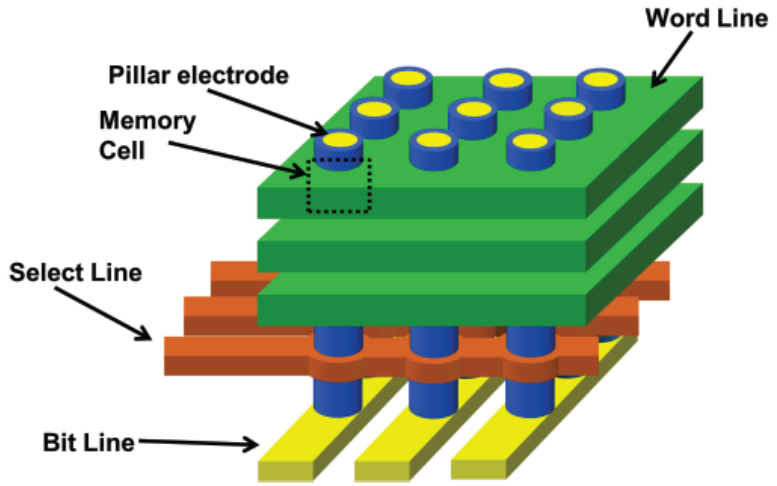


3D RRAM - Architectural Concepts

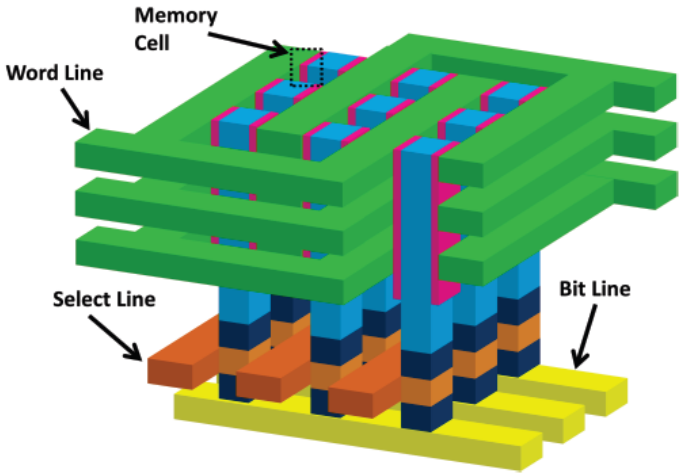
NANO ELECTRONICS GROUP

HRRAM
- Most simplistic
- Superior performance due to low RC interconnects
- Least cost efficient





VRRAM type I
- Interconnect resistance limited performance
- More energy efficient than VRAM type II



VRRAM type II
- Interconnect capacitance limited performance
- Litho-free stacking
- Bit performance improves with # layers
- Most scalable/cost efficient

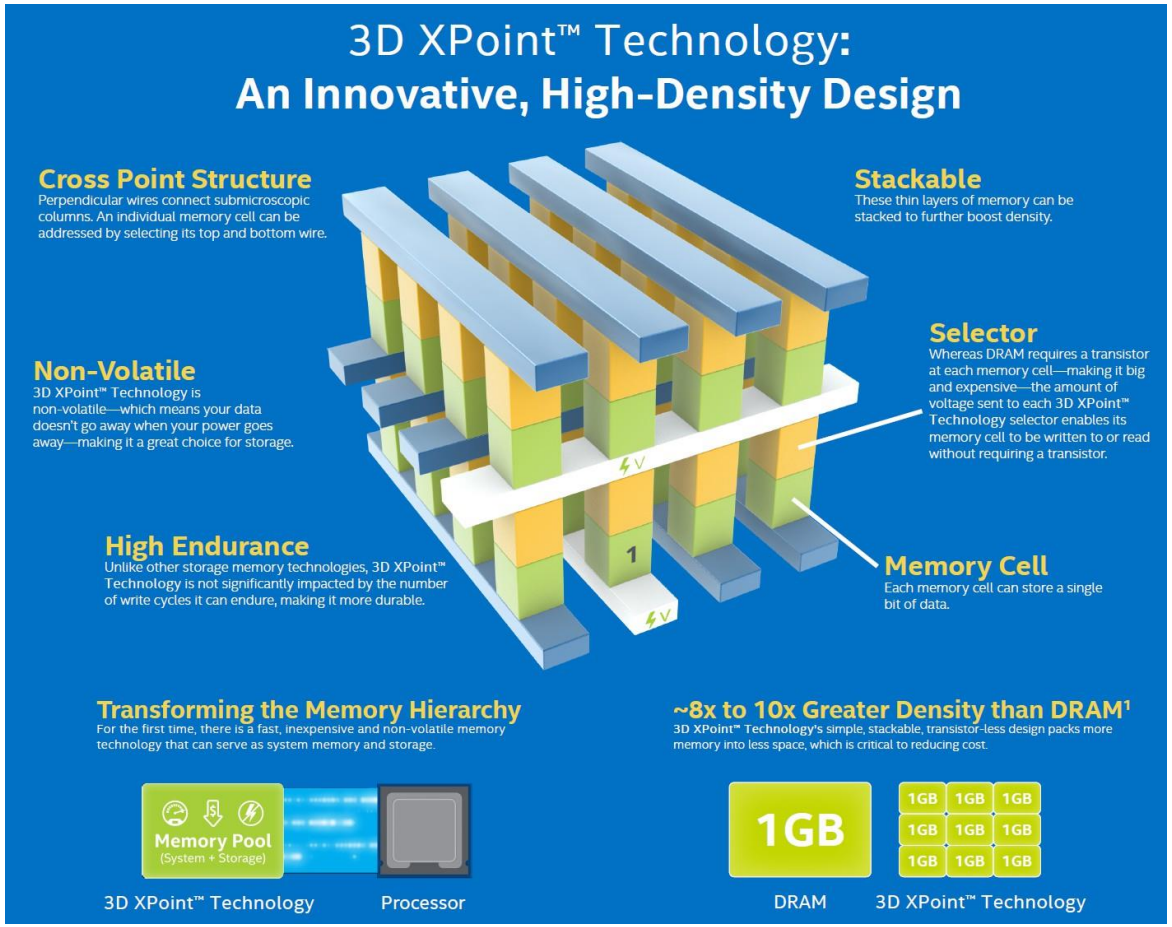
S. Yu et al – ISCAS 2014

HRRAM – Already Commercialized

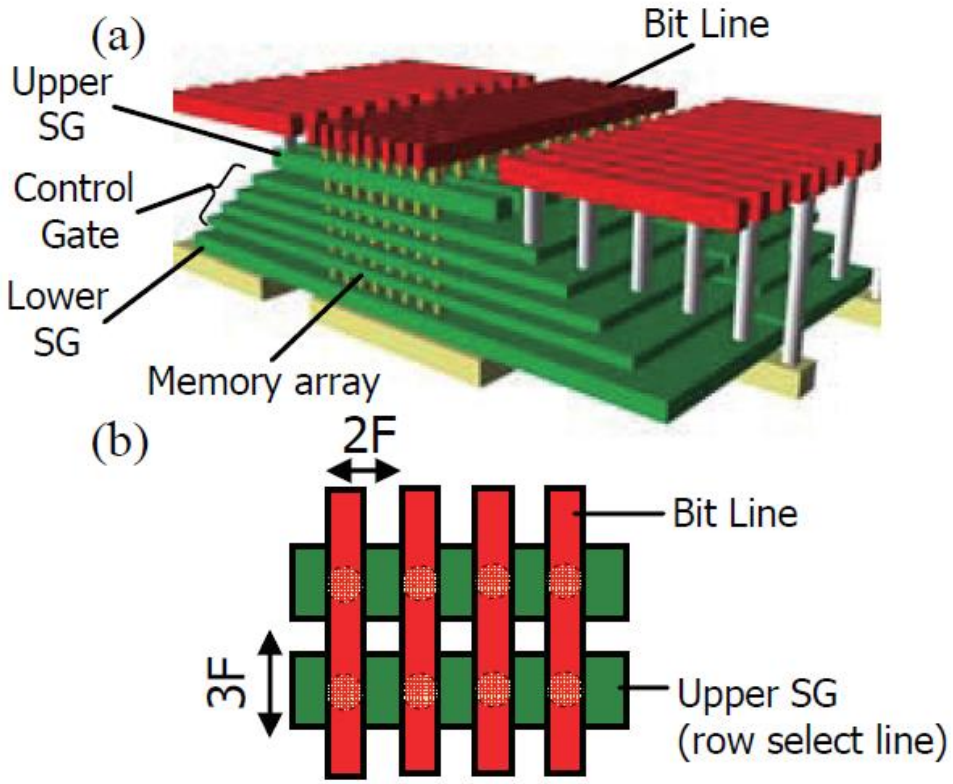
NANO ELECTRONICS GROUP

HRRAM
- Most simplistic
- Superior performance due to low RC interconnects
- Least cost efficient

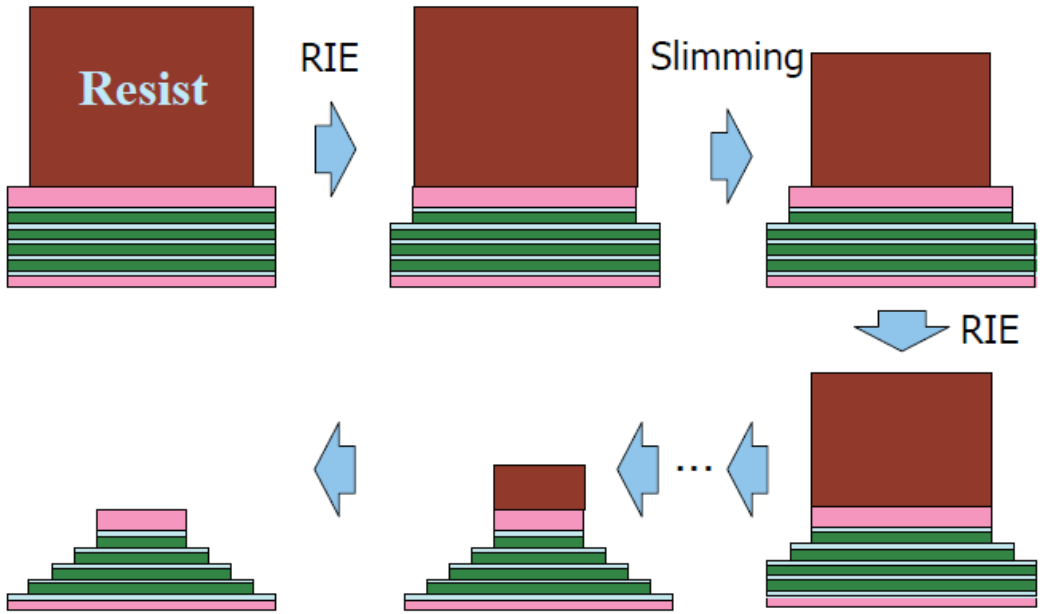
Intel 3D Xpoint
- 3x faster than NAND flash
- 5x more expensive
- Switching mechanism unknown
- Limited stacking potential due to diode selector



Vertical RRAM – Lithography Free



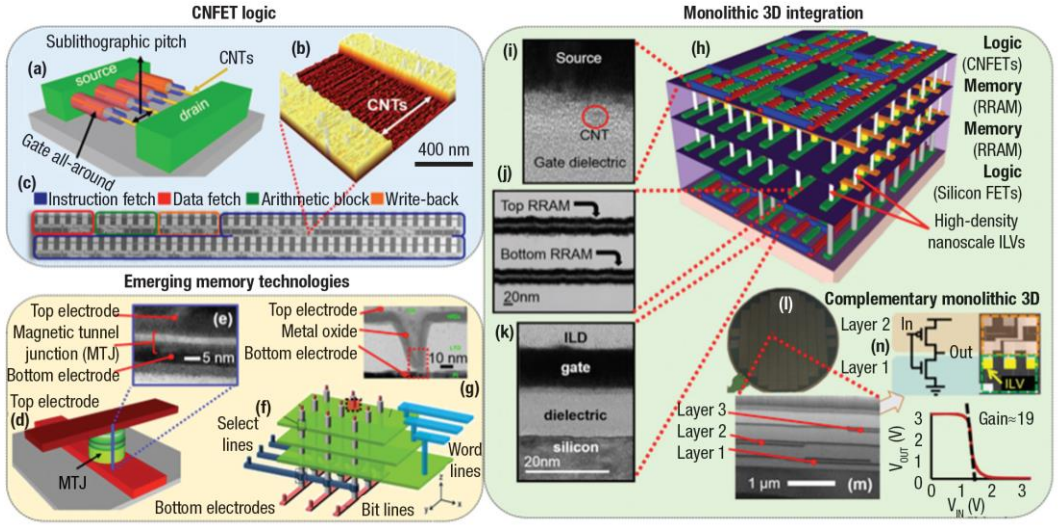
Litho-free formation of a stair-case structure

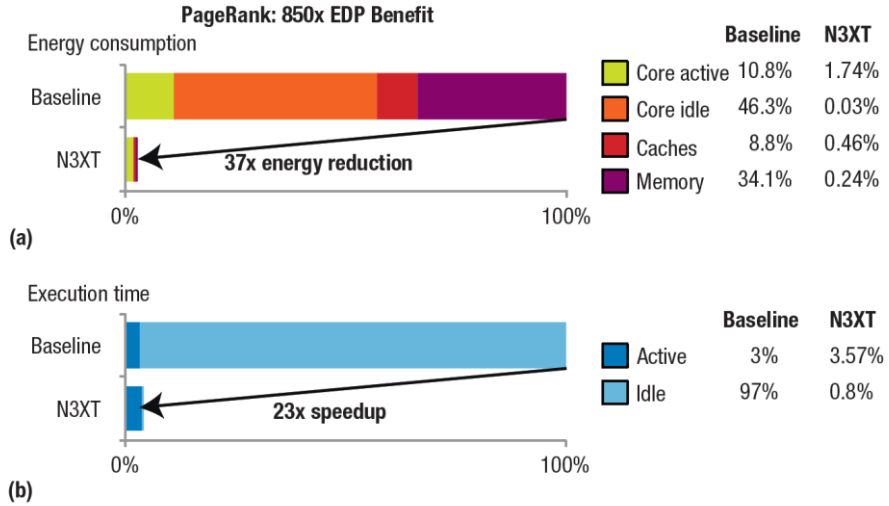


Tanaka et al – VLSI symposium 2007

Research Trend I – Stacking Circuits and Mem

- To predict the potential benefits of introducing new types of memory, projected and demonstrated performance metrics have been put into models
- Benchmarks of a contemporary Intel Xeon Phi system VS a system with CNT-cores with STT-MRAM + 3D RRAM show large advantages
- Major part of the improved performance comes from the new memory technology (conventional CPU is idle 97% of the time)
- Proposed system shows up to 1000x gains in combined power and speed

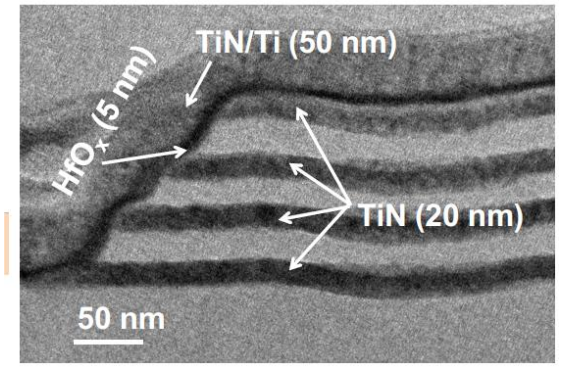


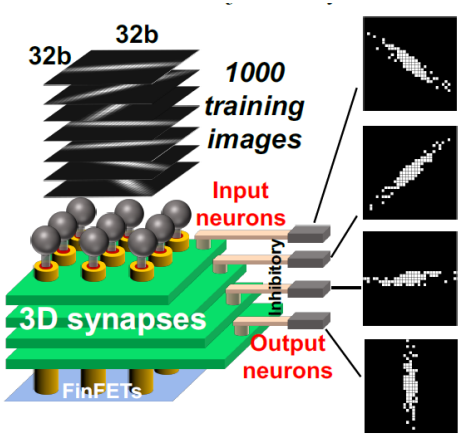


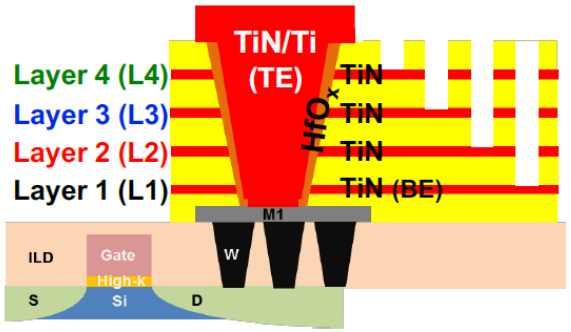
M. Aly et al – IEEE computer 2015

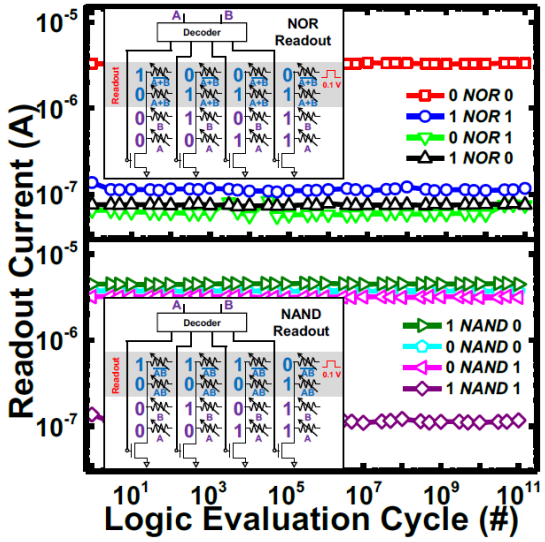
Research Trend II – In Memory Computations

- One interesting aspect of 3D RRAM is the possability to do computations directly in the memory, having hyper-dimensional vectors (3D vectors)
- NOR and NAND operations can be accomplished with specific pulse trains
- Not determined if in-memory computations will be a viable way forward





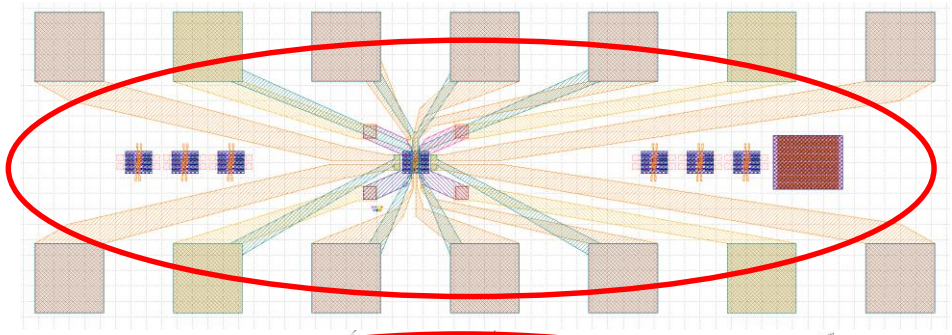


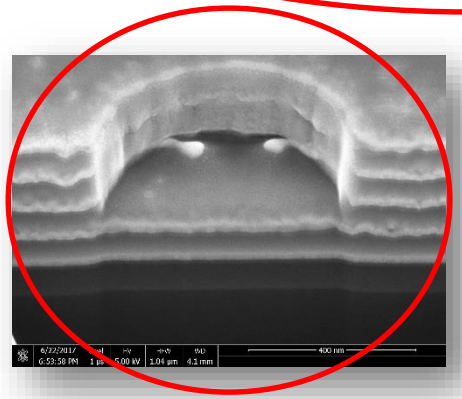


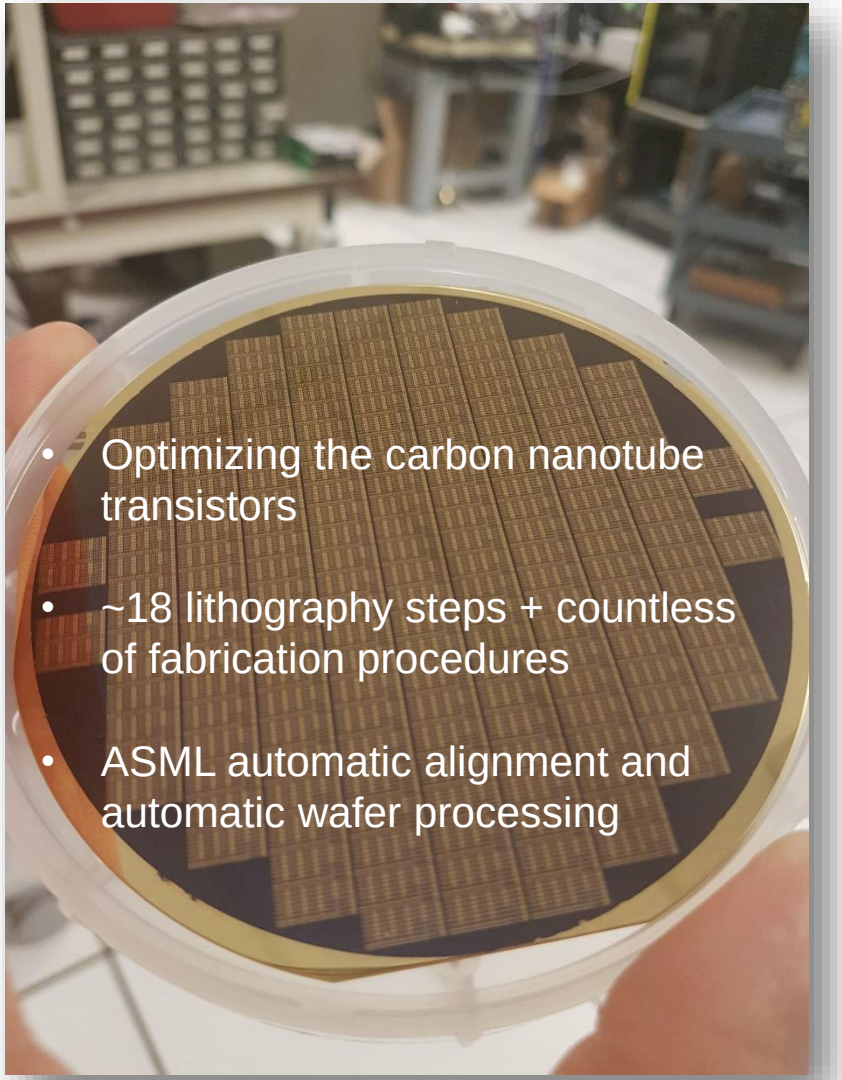
H. Li et al – VLSI 2016

Research - From Idea to Realization

NANO ELECTRONICS GROUP

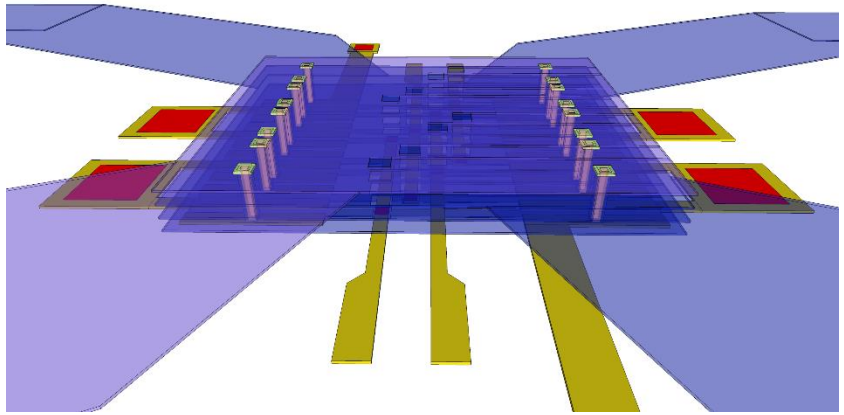


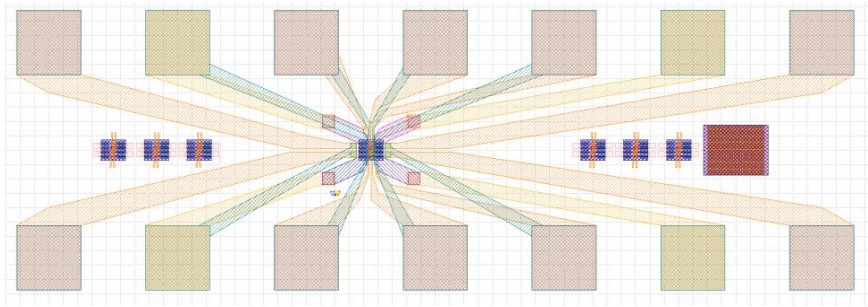


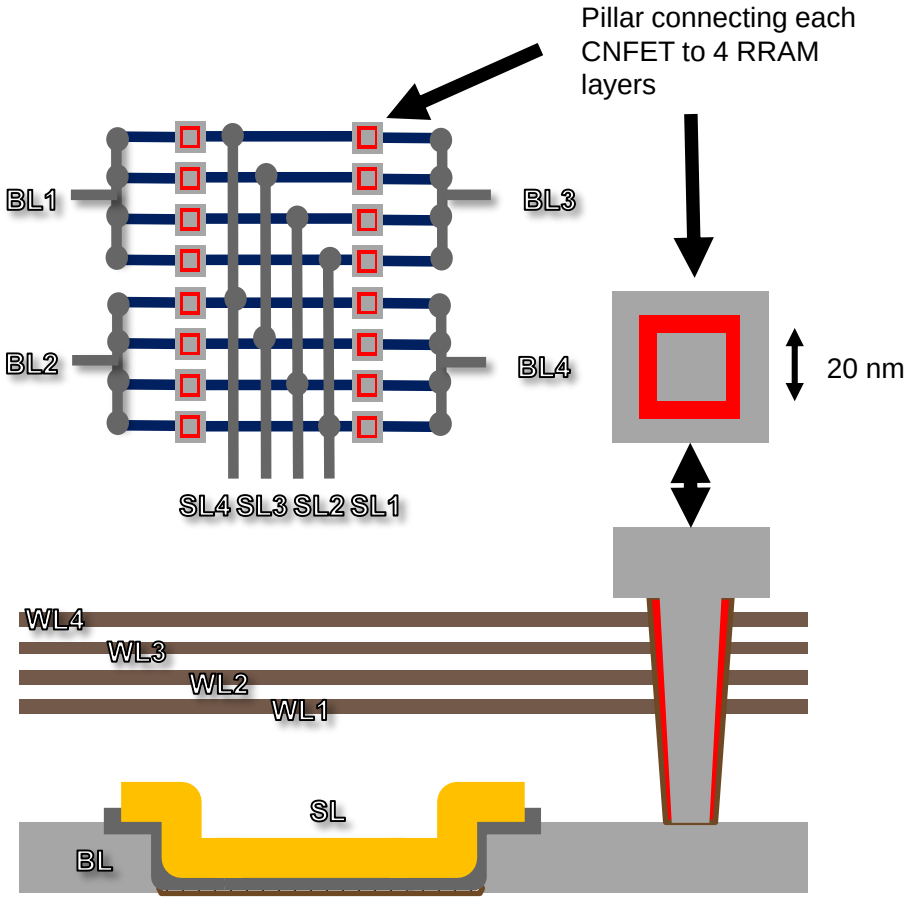


Research - From Idea to Realization

NANO ELECTRONICS GROUP

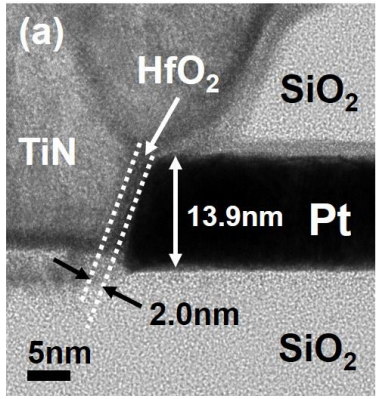


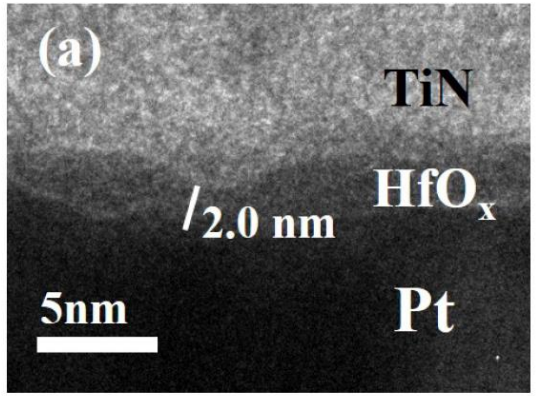


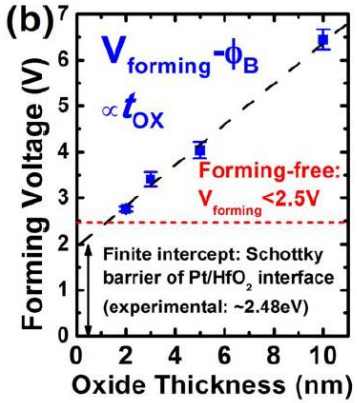


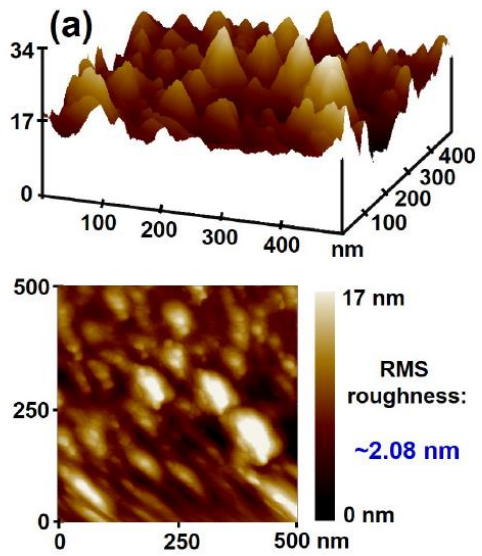
M-Ox Thickness Considerations

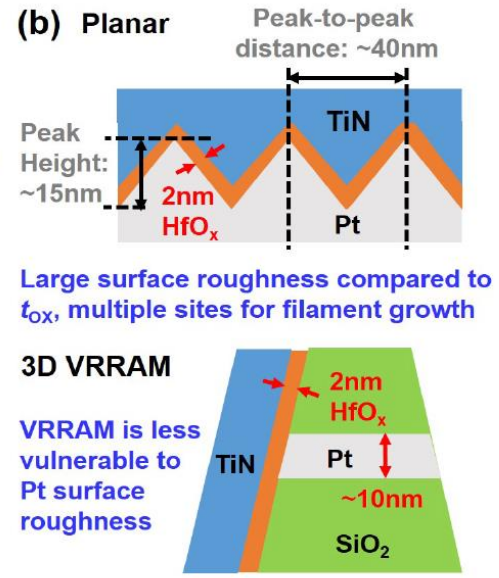
- Scaling the dielectric desirable
- Surface roughness is a limitation
- Vertical pillar have smoother surface





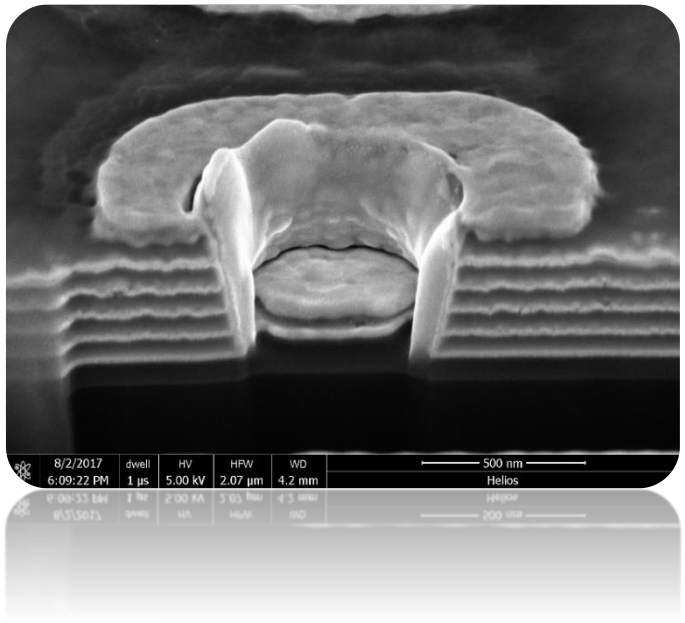


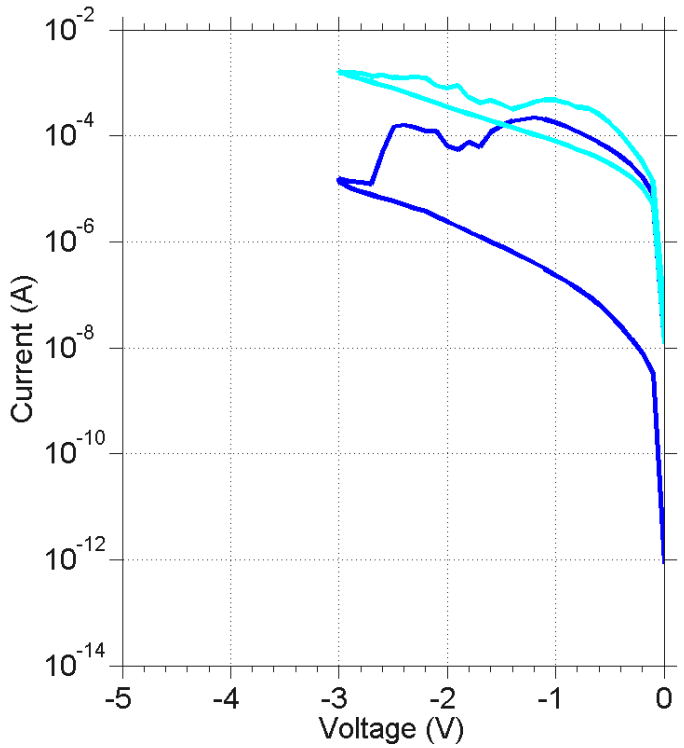


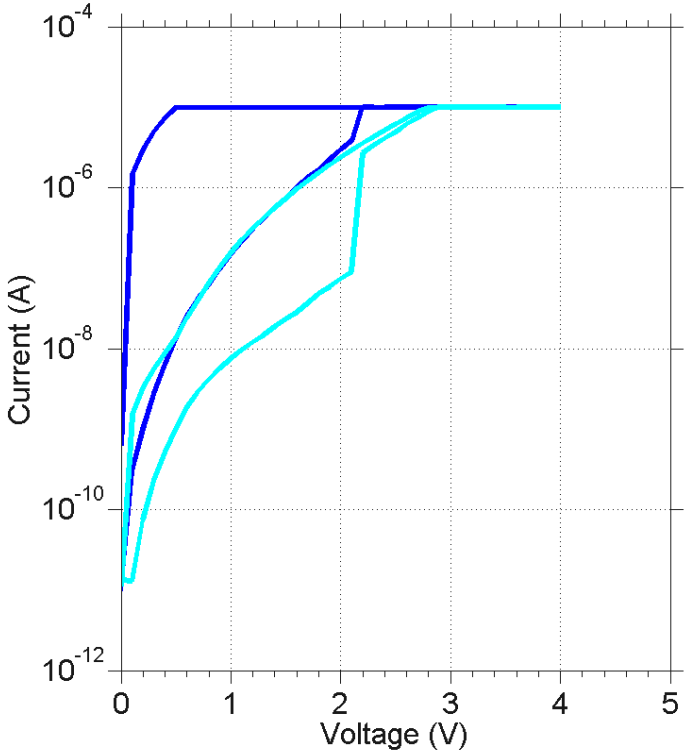


Zhao et al – IEDM 2014

TiN RRAM Layer 1 and 2 – Form and Reset

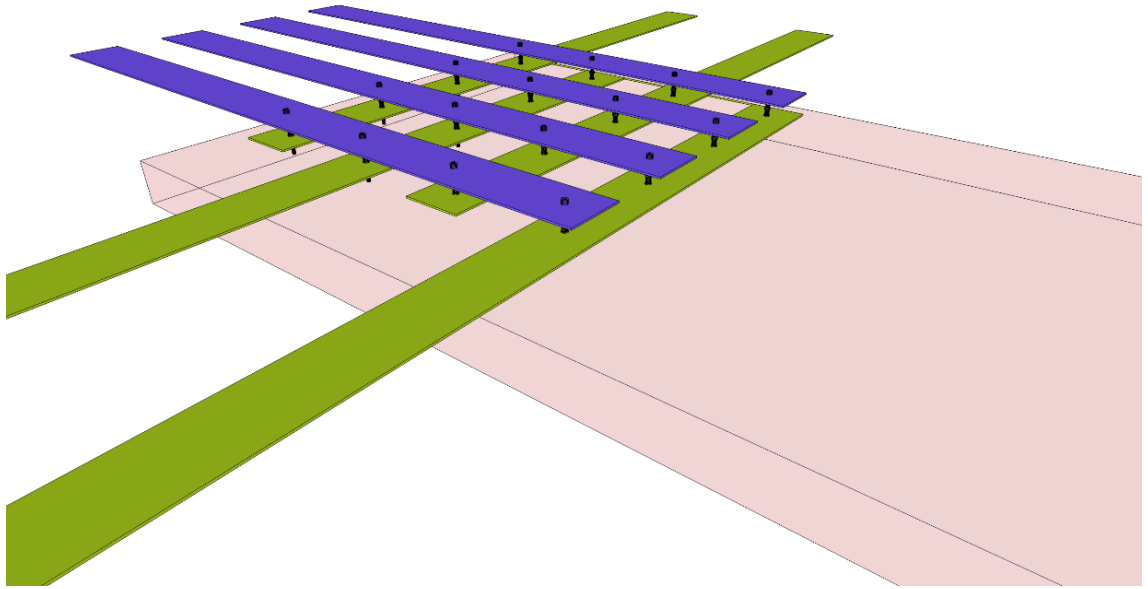


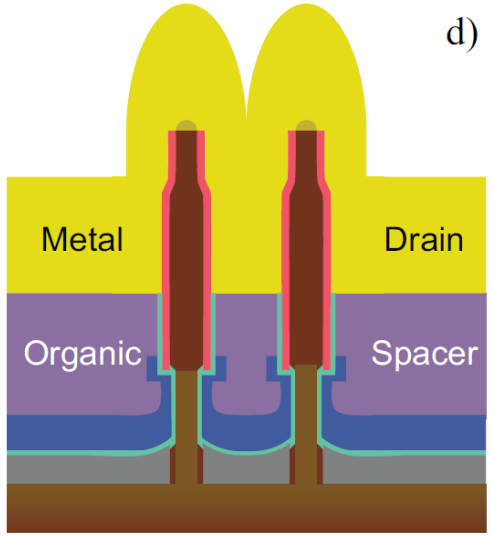


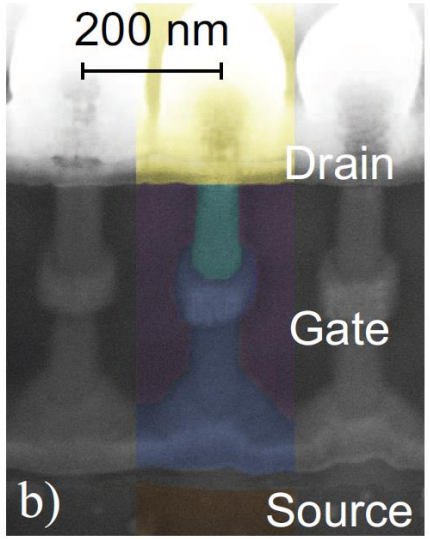


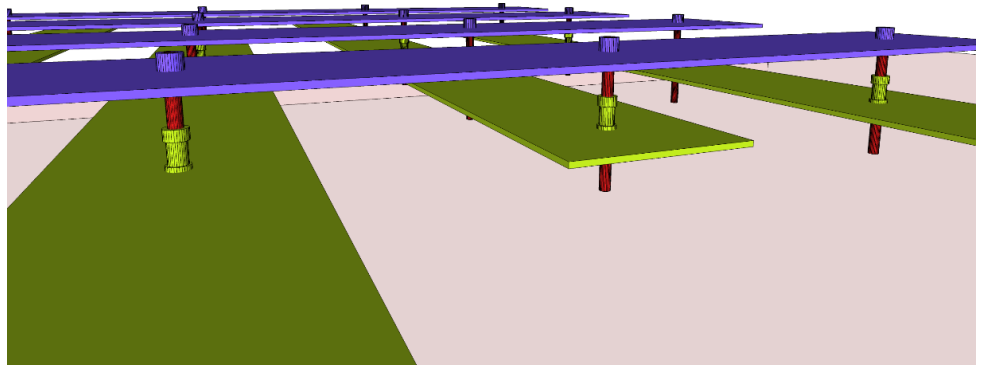
Lund Research – RRAM with NWFETs

NANO ELECTRONICS GROUP



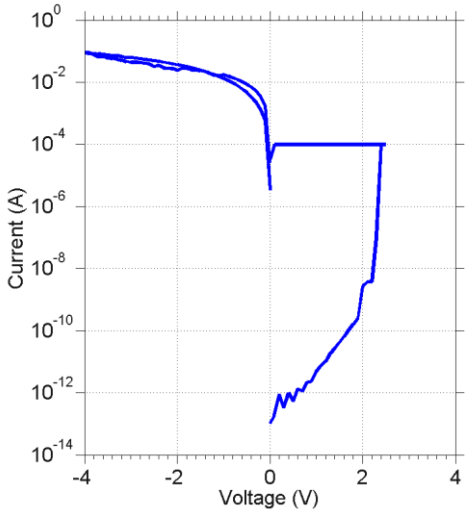




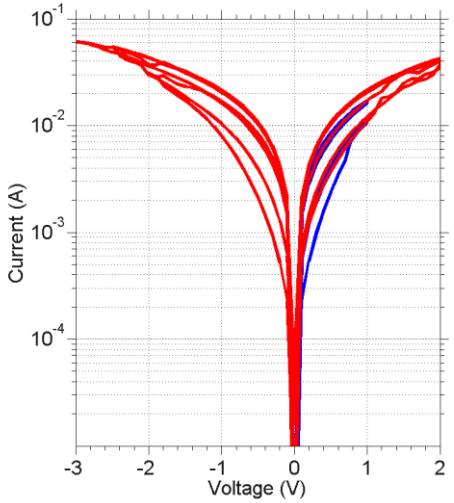


Initial 2D RRAM Tests

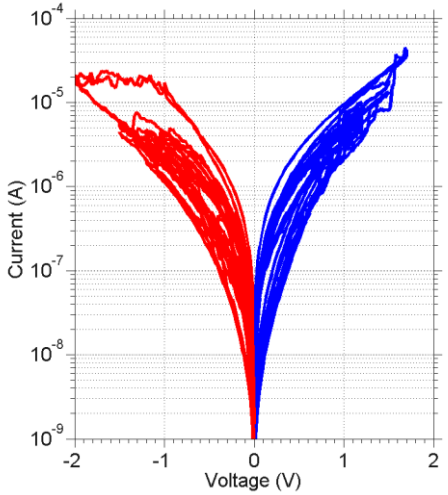
Ag/HfO2/W



Ag/HfO2/Ti/W

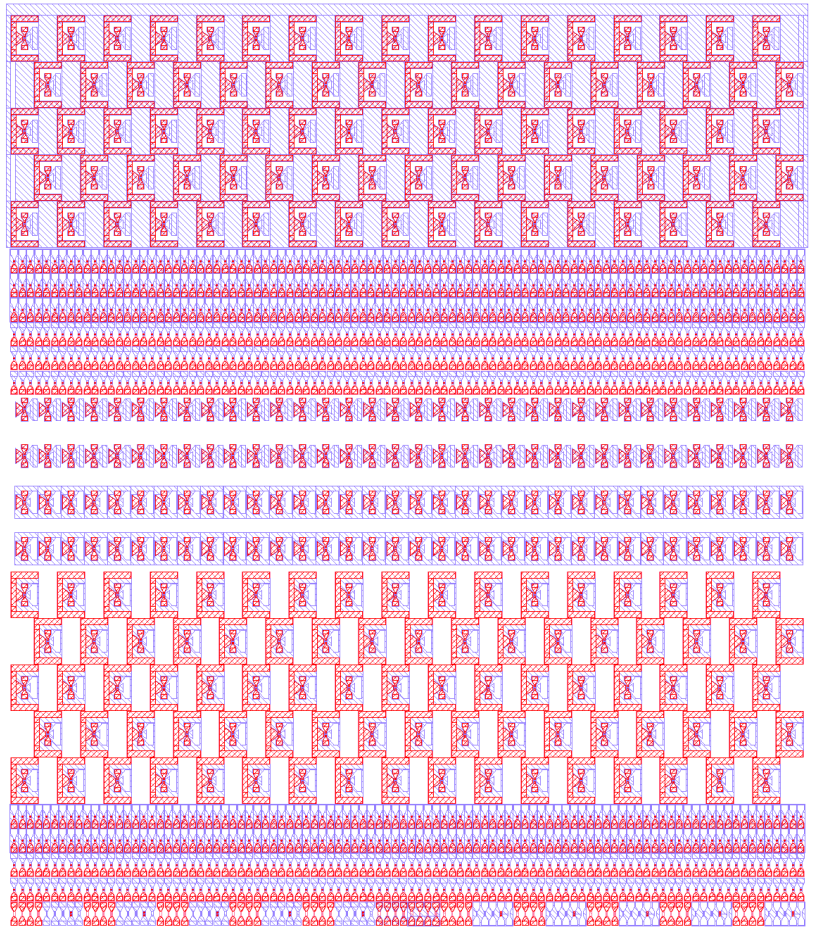


ITO/SiO2/ITO



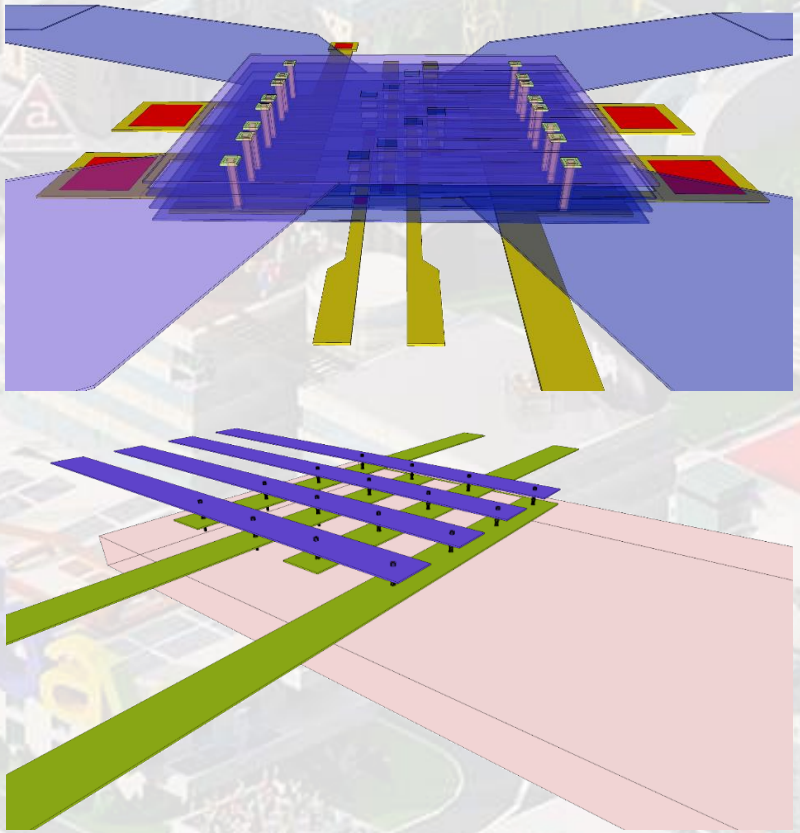
How to Proceed

1. 2D RRAM characterization
    1. Voltage envelope
    2. Performance metrics
2. 3D RRAM on vMOSFET
    1. Voltage envelope
    2. Performance metrics
3. 3D RRAM on vTFET
    1. If MOSFET evaluation goes well
4. 3D RRAM arrays
    1. Common plane or cross-bar
    2. Bottom-up or top-down
5. 3D RRAM with nFET only muxers circuitry → ~100 FETs for a 4x4x4 64 bit cell



Future of Neuromorphic Computing

- Buzzwords + buzzwords
- Conventional computers ill-fitted for deep-learning applications
- Contemporary memory technologies are not on par with the development in CPU speed
- 3D stacking circuits and ReRAM could potentially improve efficiency for data intense computing by 1000x
- Rapidly growing research area due to the promise of machine learning and AI
- Enormous space for innovation and new startups



LUND
UNIVERSITY

NANO
ELECTRONICS
GROUP