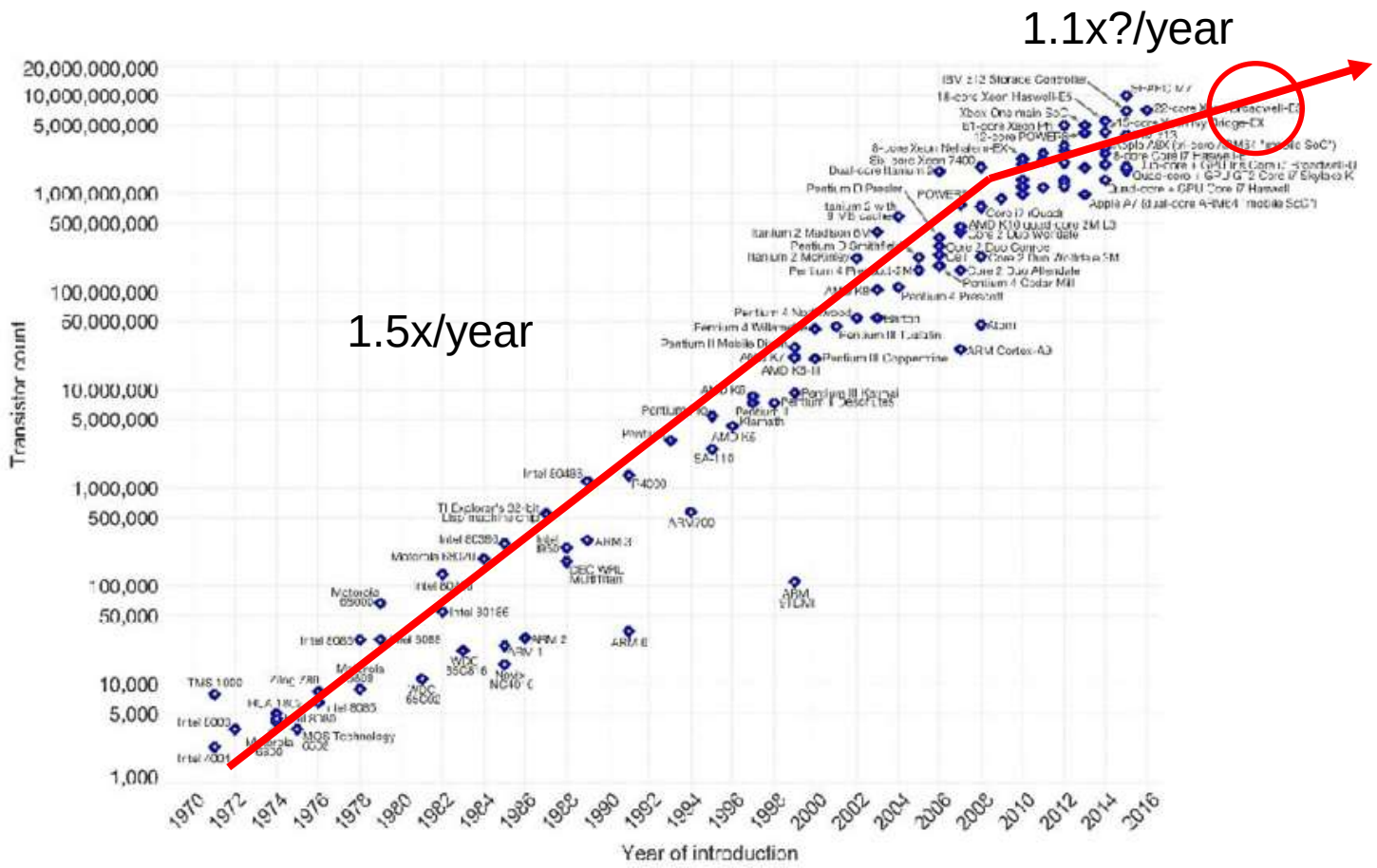
LUND UNIVERSITY

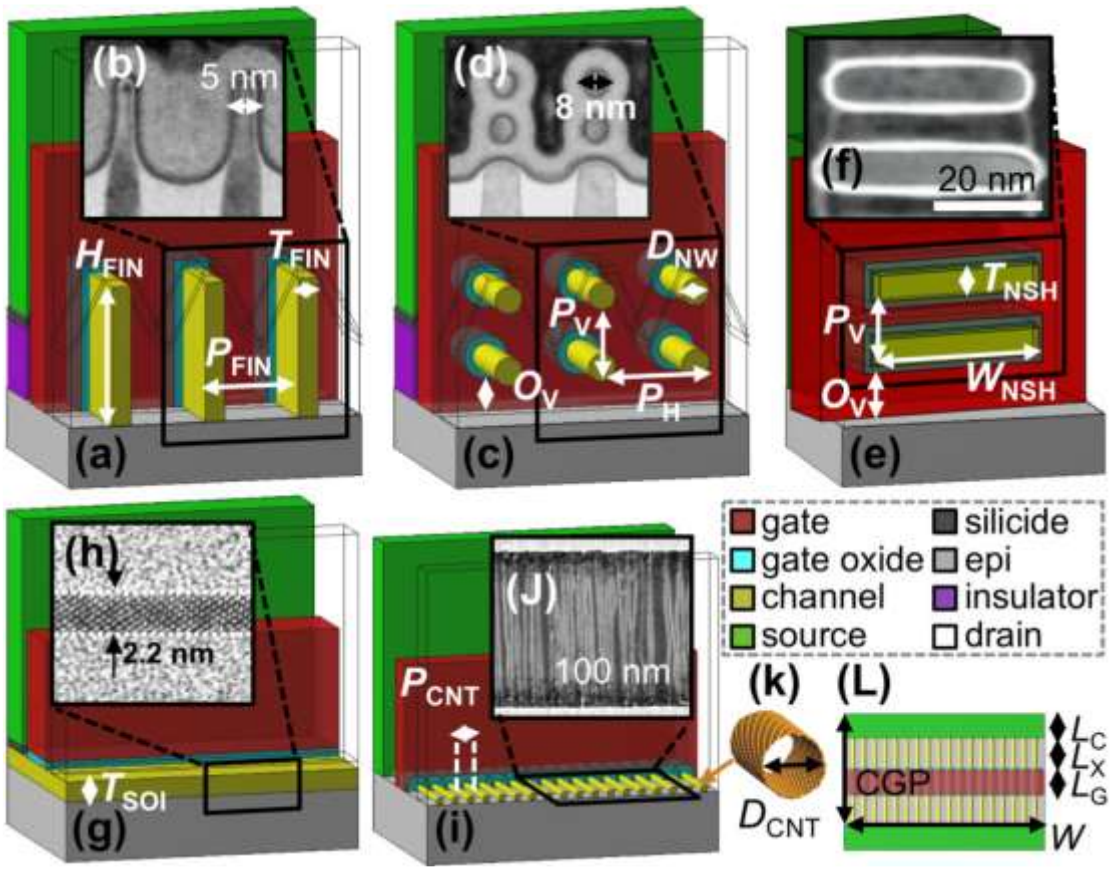
Emerging Memory Technologies

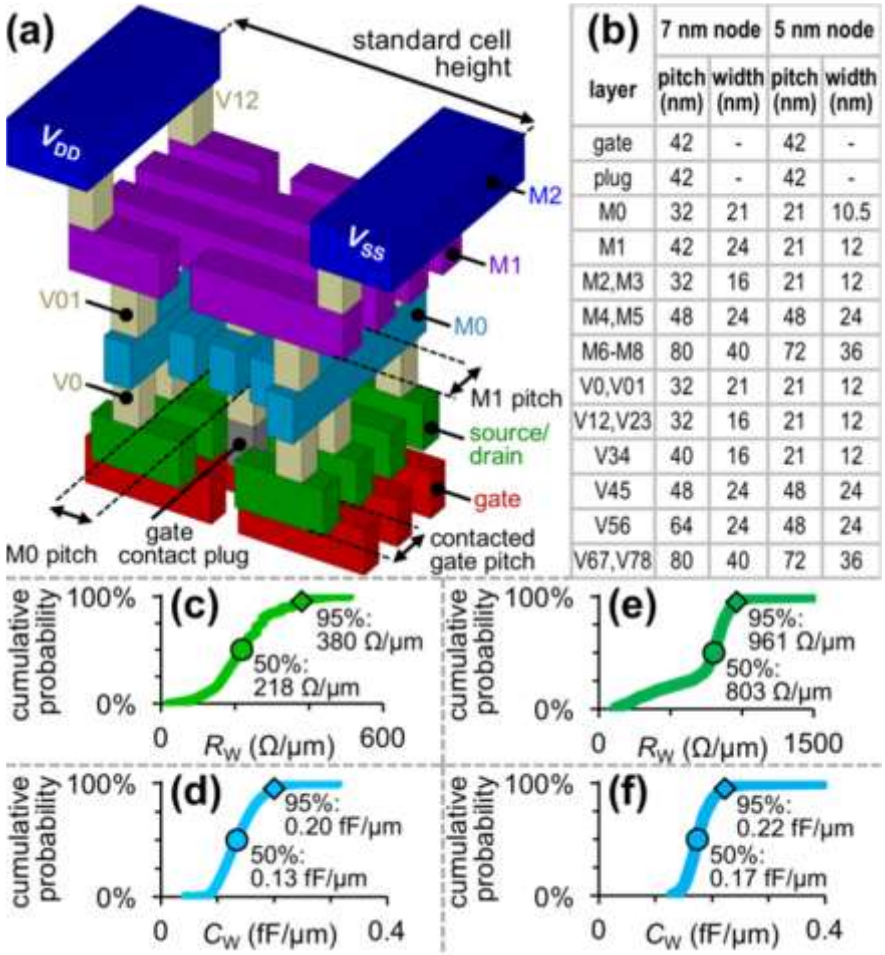
KARL-MAGNUS PERSSON

Future CMOS Development Irrelevant?



Technology Benchmarking



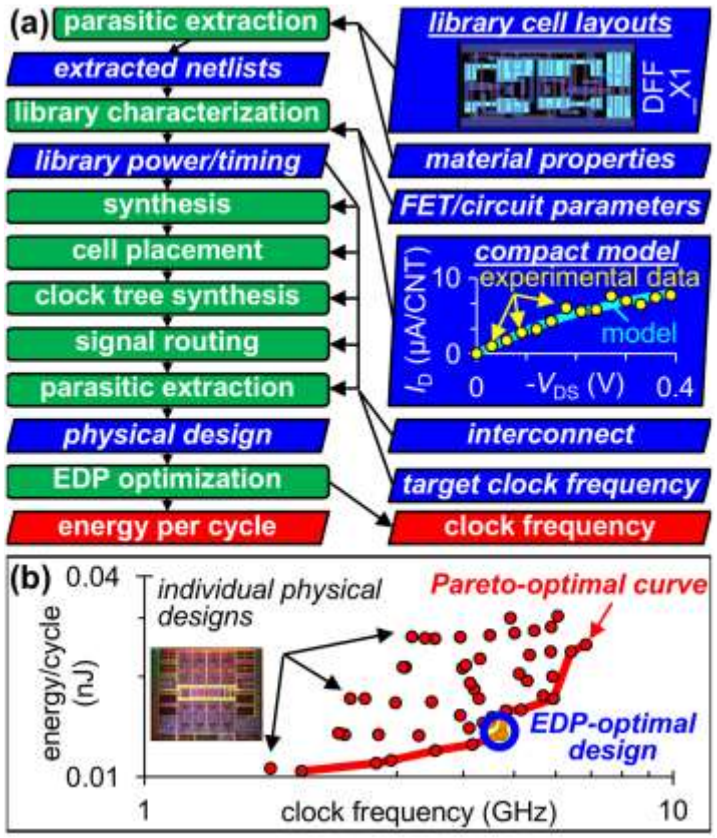


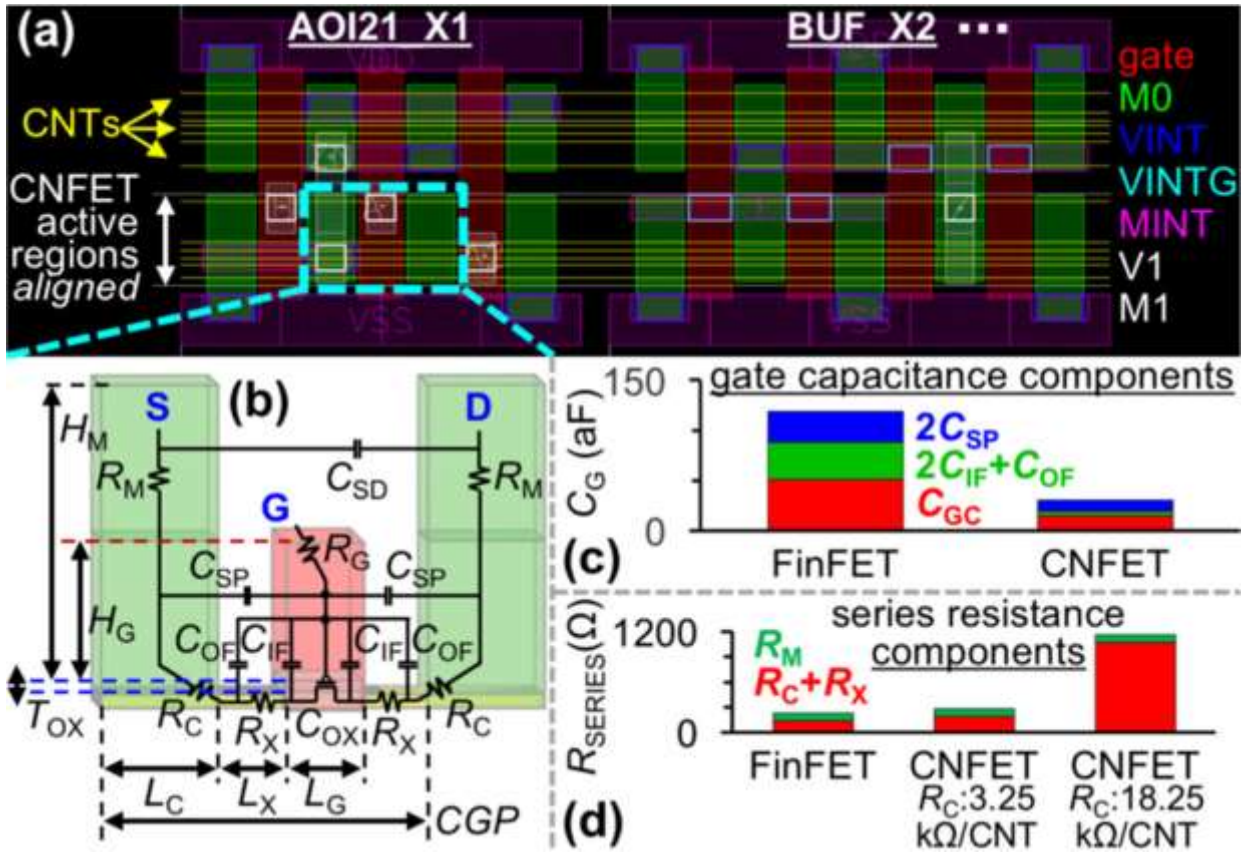
(b)

layer	7 nm node pitch (nm)	7 nm node width (nm)	5 nm node pitch (nm)	5 nm node width (nm)
gate	42	-	42	-
plug	42	-	42	-
M0	32	21	21	10.5
M1	42	24	21	12
M2,M3	32	16	21	12
M4,M5	48	24	48	24
M6-M8	80	40	72	36
V0,V01	32	21	21	12
V12,V23	32	16	21	12
V34	40	16	21	12
V45	48	24	48	24
V56	64	24	48	24
V67,V78	80	40	72	36

G. Hills et al, Nanotech 2018

Technology Benchmarking

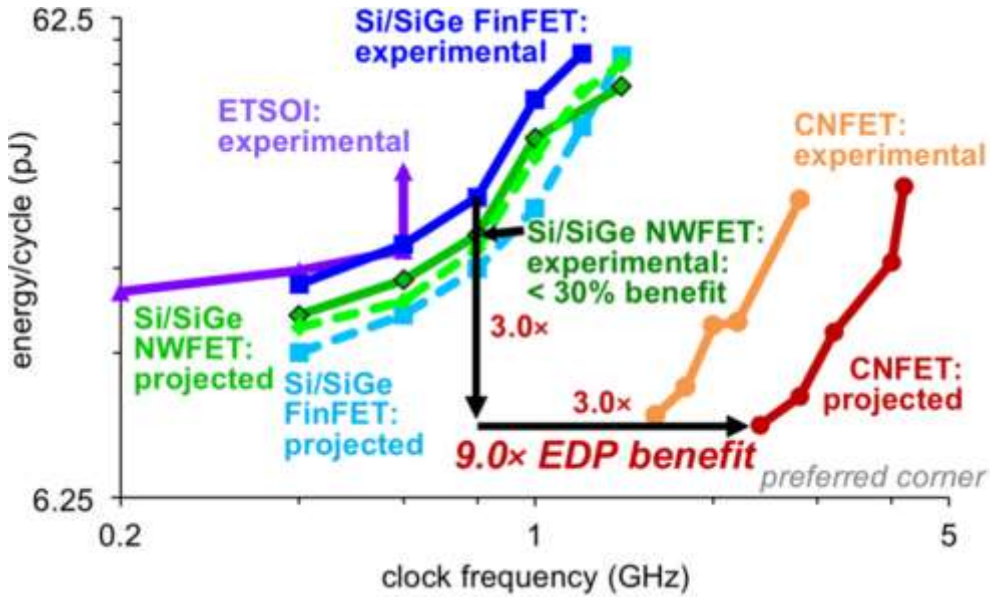


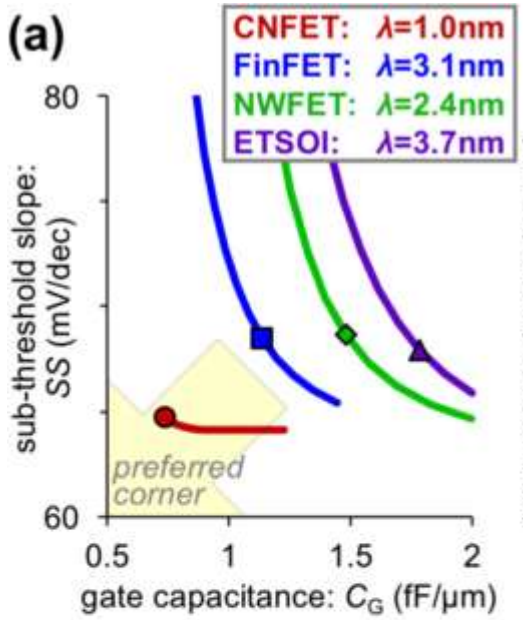


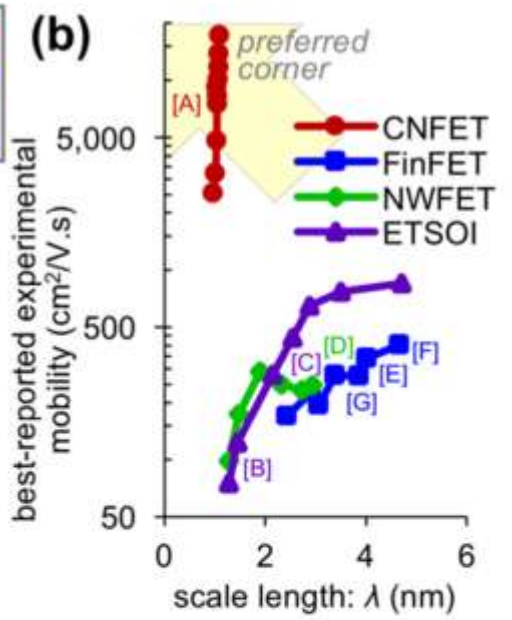
G. Hills et al, Nanotech 2018

Technology Benchmarking

NANO ELECTRONICS GROUP







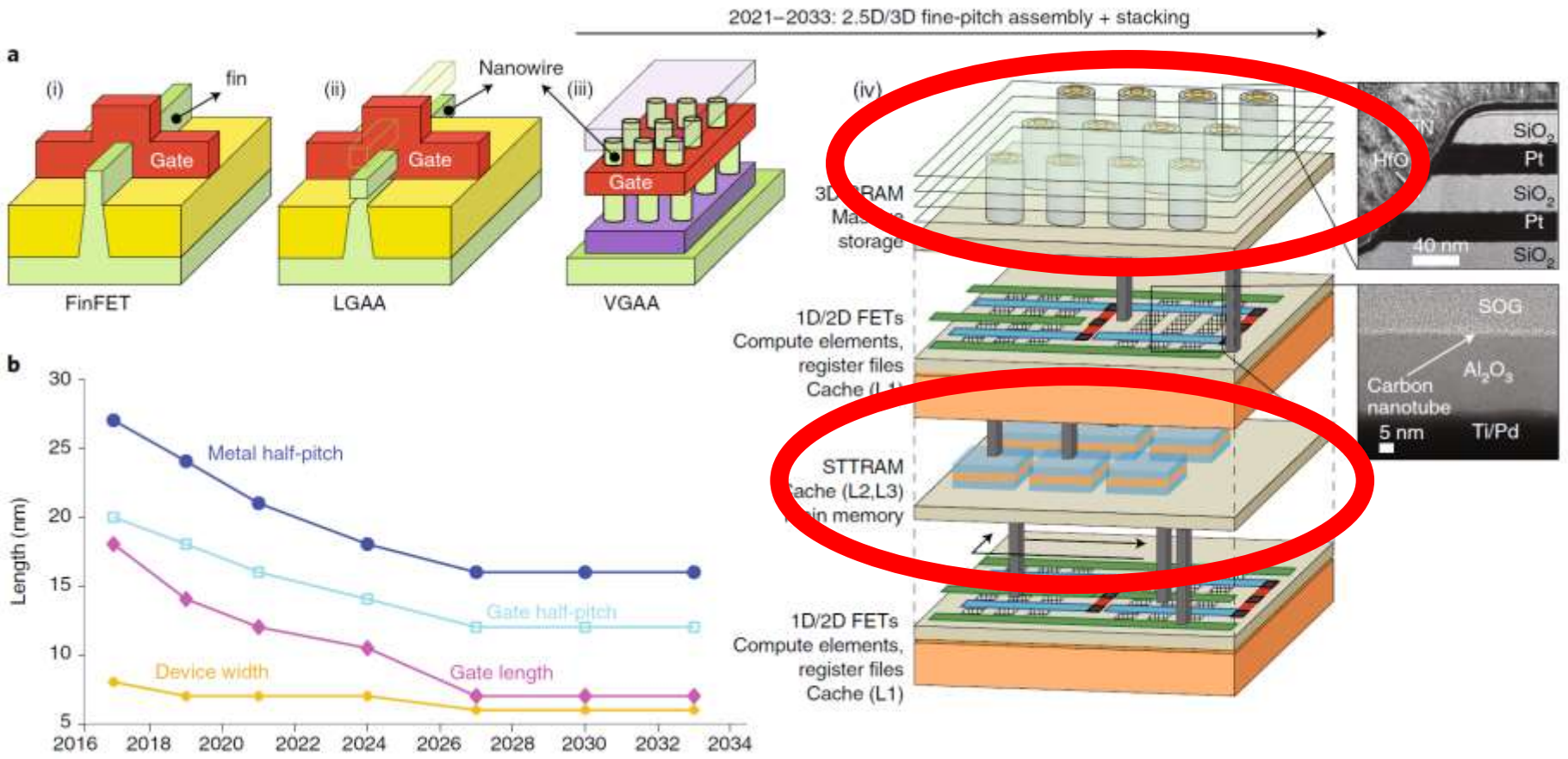
(c)

	FinFET	CNFET
λ (nm)	3.1	1.0
mobility (cm^2/V.s)	194	2,990
velocity (cm/s)	0.94	4.69
C_G (fF/μm)	1.13	0.74
SS (mV/dec)	68	65

(d)

VLSI circuit benefits	FinFET	CNFET
V_{DD}	1×	0.80×
I_{EFF}	1×	1.25×
C_{TOTAL}	1×	0.50×
I_{OFF}	1×	1×
power density	1×	1×

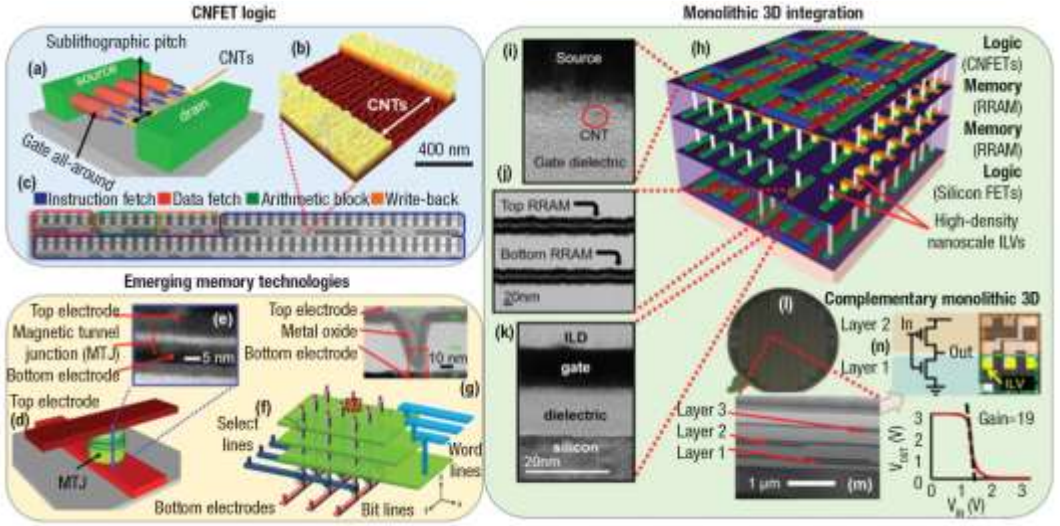
IRDS Review

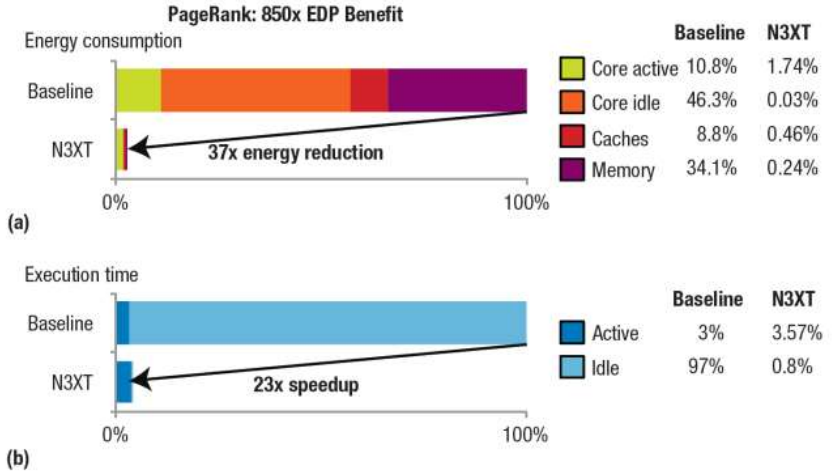


N. Orji et al - Nature Elec 2018

Research Trend I – Stacking Circuits and Mem

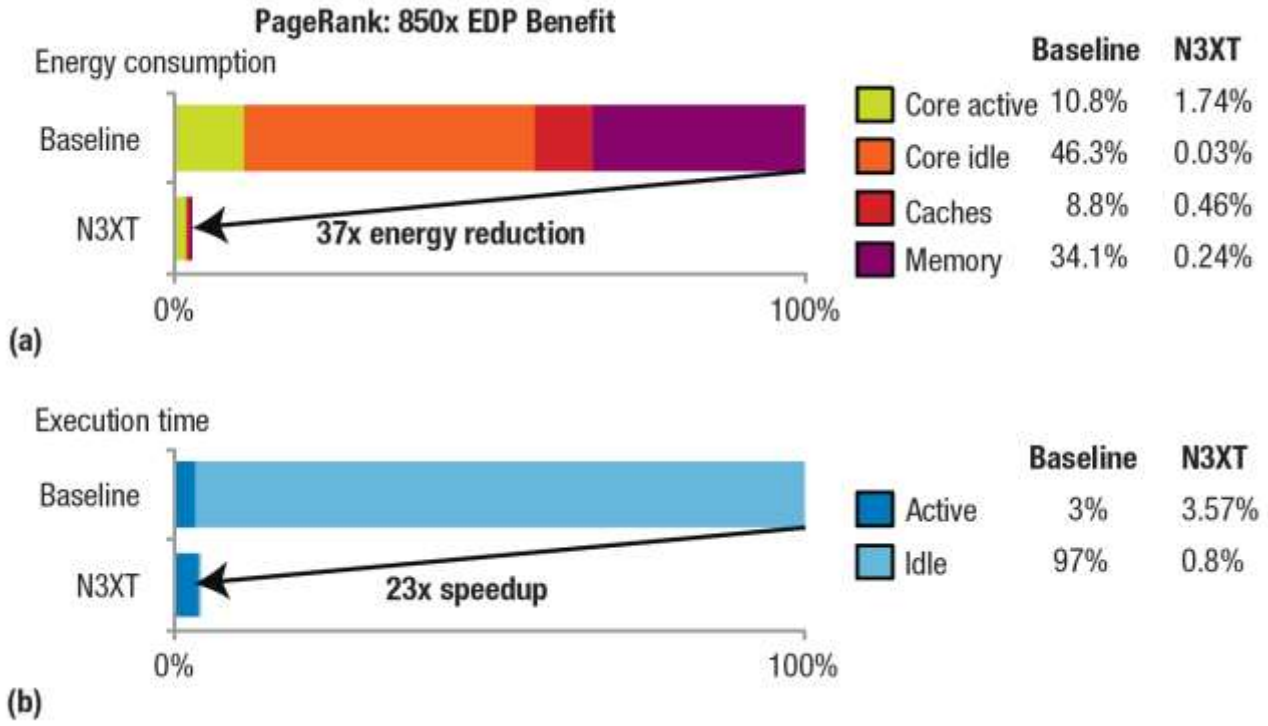
- Performance predictions of ReRAM with large scale circuit simulations using calibrated models
- Study shows benchmarks of a contemporary Intel Xeon Phi system VS a system with CNT-cores and STT-MRAM + 3D RRAM
- Proposed system shows up to 1000x gains in combined power and speed





M. Aly et al – IEEE computer 2015

Research Trend I – Stacking Circuits and Mem



Conventional CPU is idle 97% of the time!!

M. Aly et al – IEEE computer 2015

The Era of Big Data and Recognition

NANO ELECTRONICS GROUP

Machine Learning Hardware Implications

- Iterative re-programming of memory
- Large data sets → Limited by read/write of none-volatile-memory (NVM)

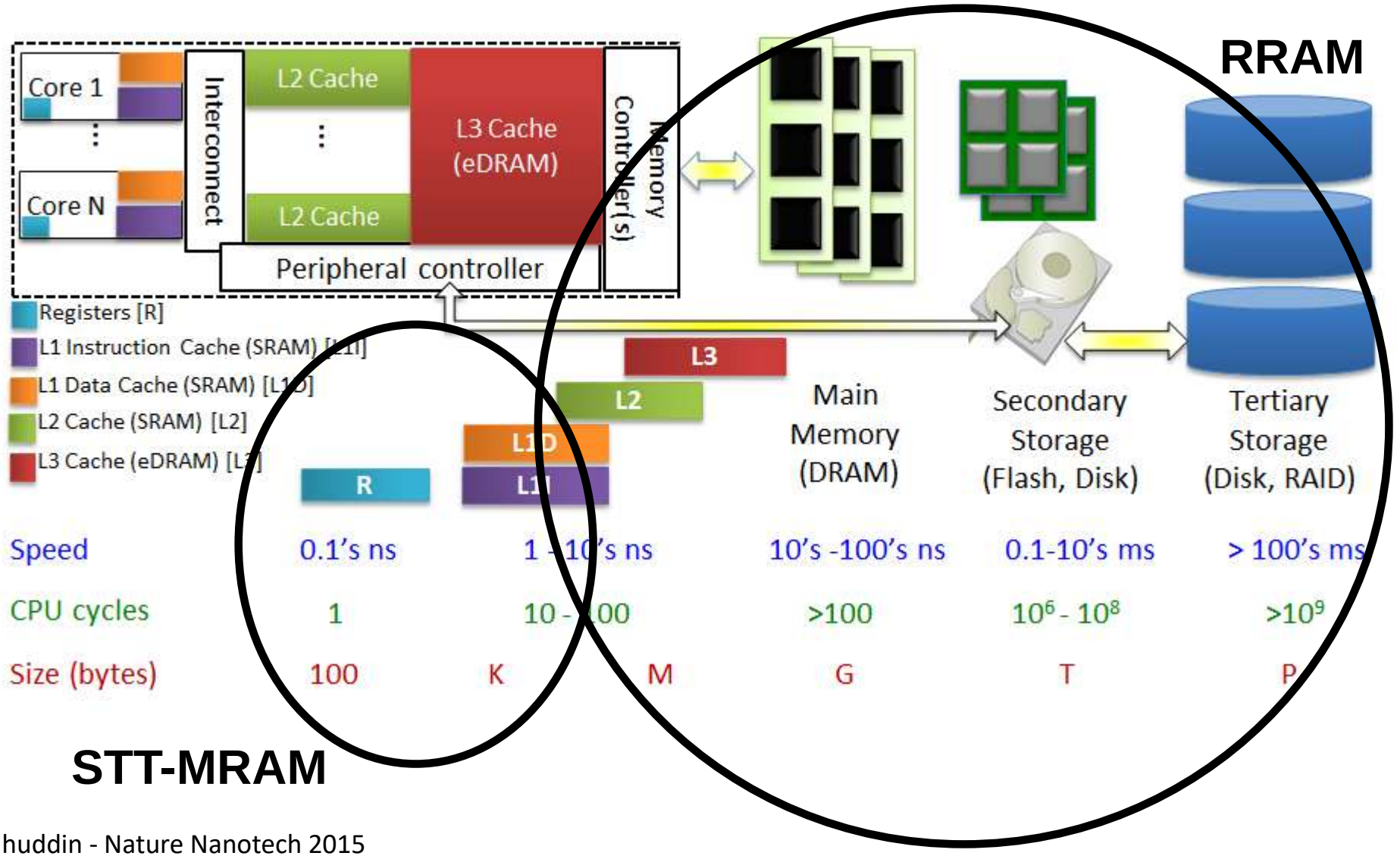
Hardware Challenges

- Component improvement stagnated - Moore's law has halted
- NVM technologies - 10,000x slower than computing
- Separate compute and memory circuitry infer large inefficiencies

Possible solutions

- New methods → co-integrated circuits and memory in 3D, introducing new materials
- True neuromorphic hardware → synaptic networks using computational units

The Memory Hiearchy

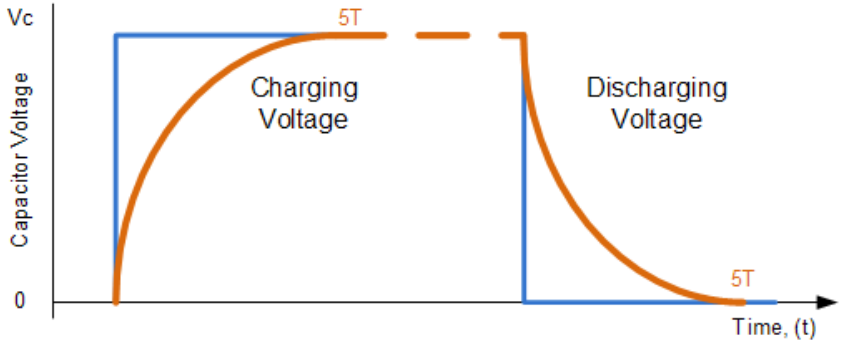


H.-S. P. Wong, S. Salahuddin - Nature Nanotech 2015

Limitations in NAND Flash

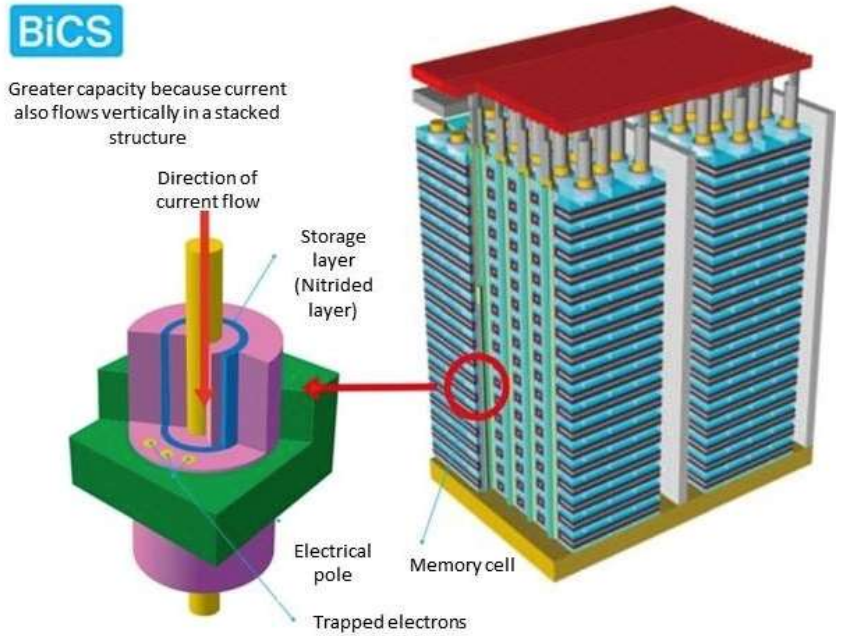
- **The upside**
  - None volatile
  - 3D integrated with 128 layers
  - Minimal feature size (F) down to 5 nm

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



RC delay

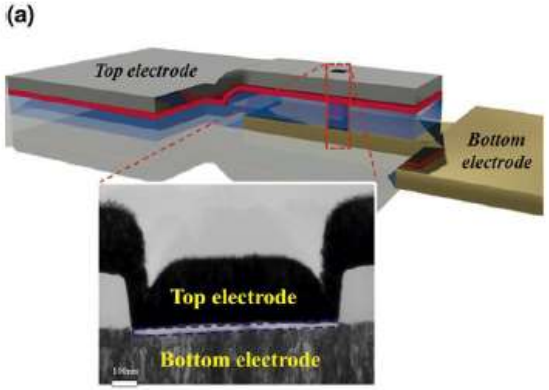
Bit Cost Scalable NAND Flash

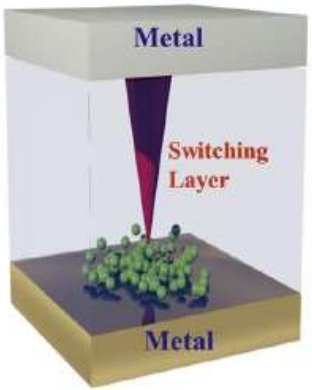


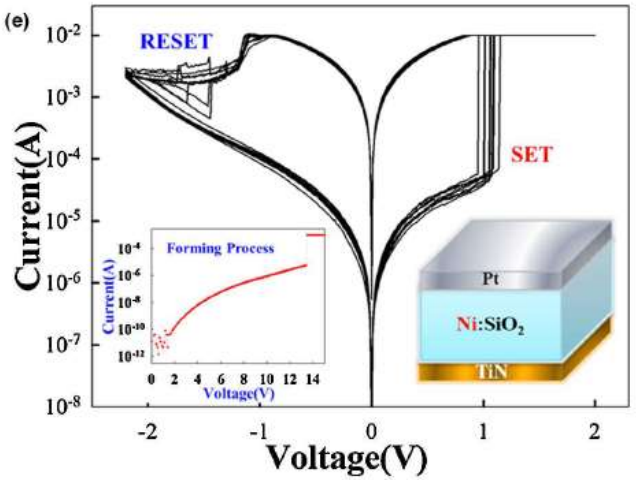
Schematic structure of planar NAND Flash

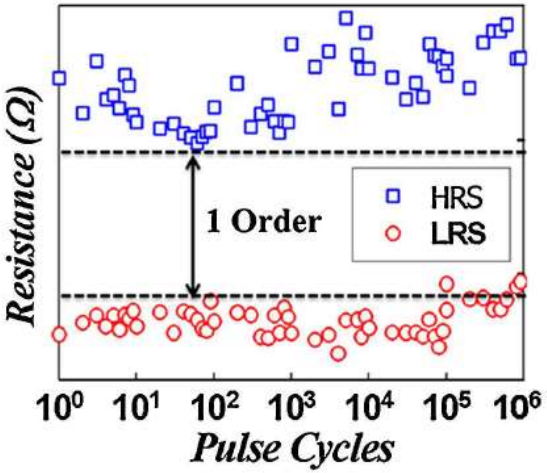
https://www.storagenewsletter.com

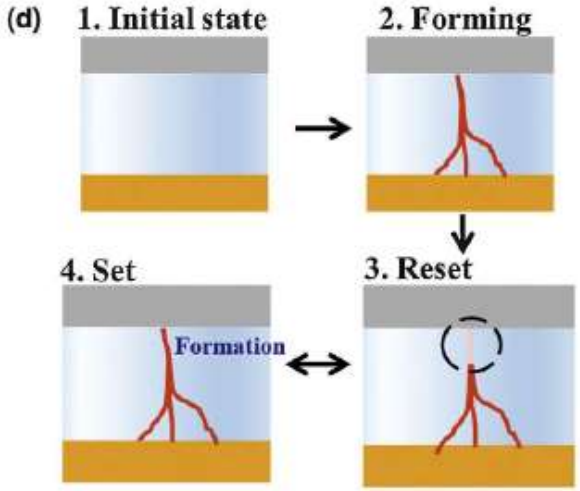
Resistive Memory



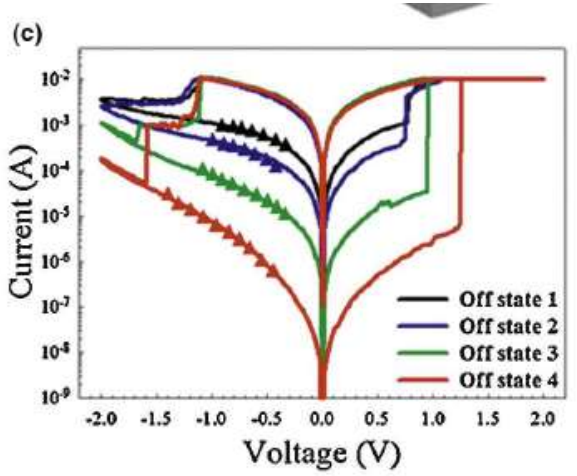








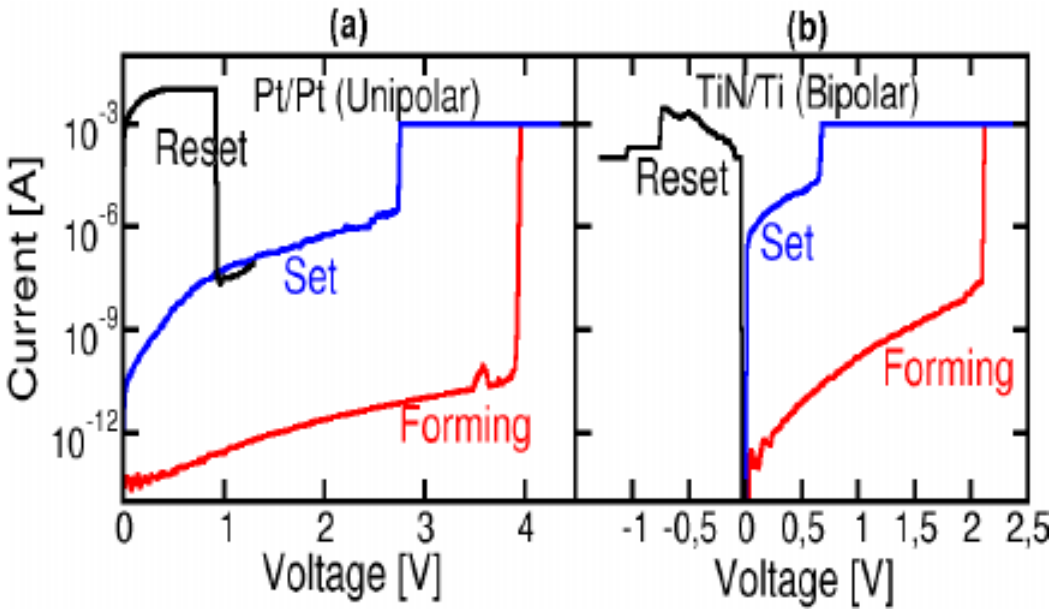
1. Filament Forming
2. Local Rupture (Reset)
3. Local Forming (Set)

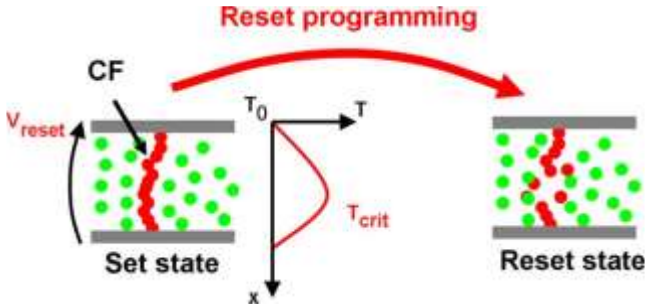


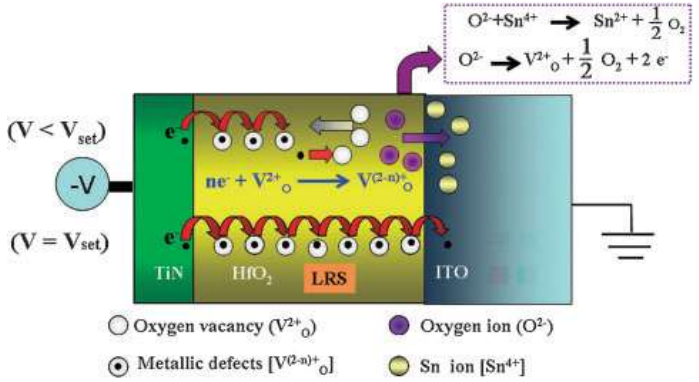
RRAM Mechanics

- Unipolar Switching
  - Joule heating
  - Ion movement due to diffusion
  - Vacancies dissolved into oxide
  - Occur at the center of the dielectric (hottest)
  - Large reset current required (most cases)

- Bipolar Switching
  - Bi-polar-bias requirement
  - Ion movement drift due applied field
  - Vacancies absorbed into metal
  - Occur at the oxide/metal interface







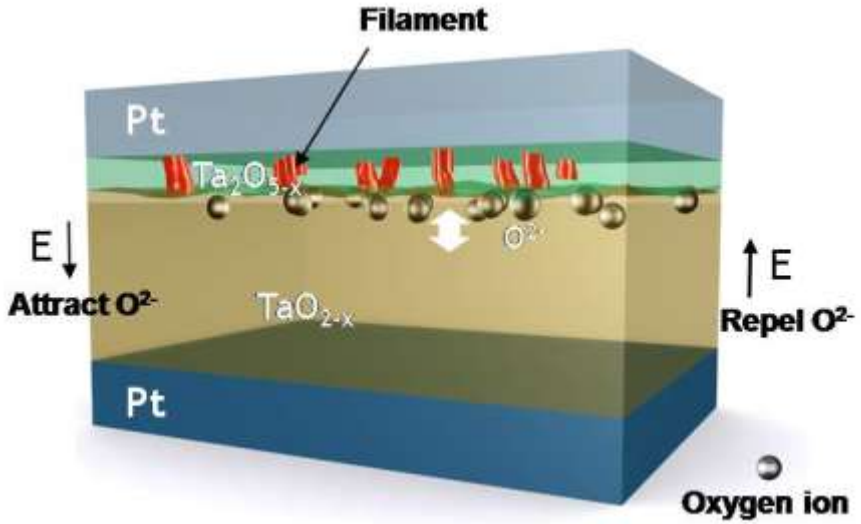
Overview of different ReRAM Technologies

- **RRAM (oxRRAM)**
  - Anode filament, oxygen vacancies form conductive path
  - High endurance, 1^{12} cycles at device level
  - 3D compatible

- **CBRAM**
  - Cathode filament, bridging with metal ion movement
  - Similar structure to RRAM
  - Endurance questionable (finite number of switches)

- **PCRAM**
  - Phase change memory, a flash heating switches dielectric film between amorphous and crystalline state
  - Endurance questionable (finite number of switches)

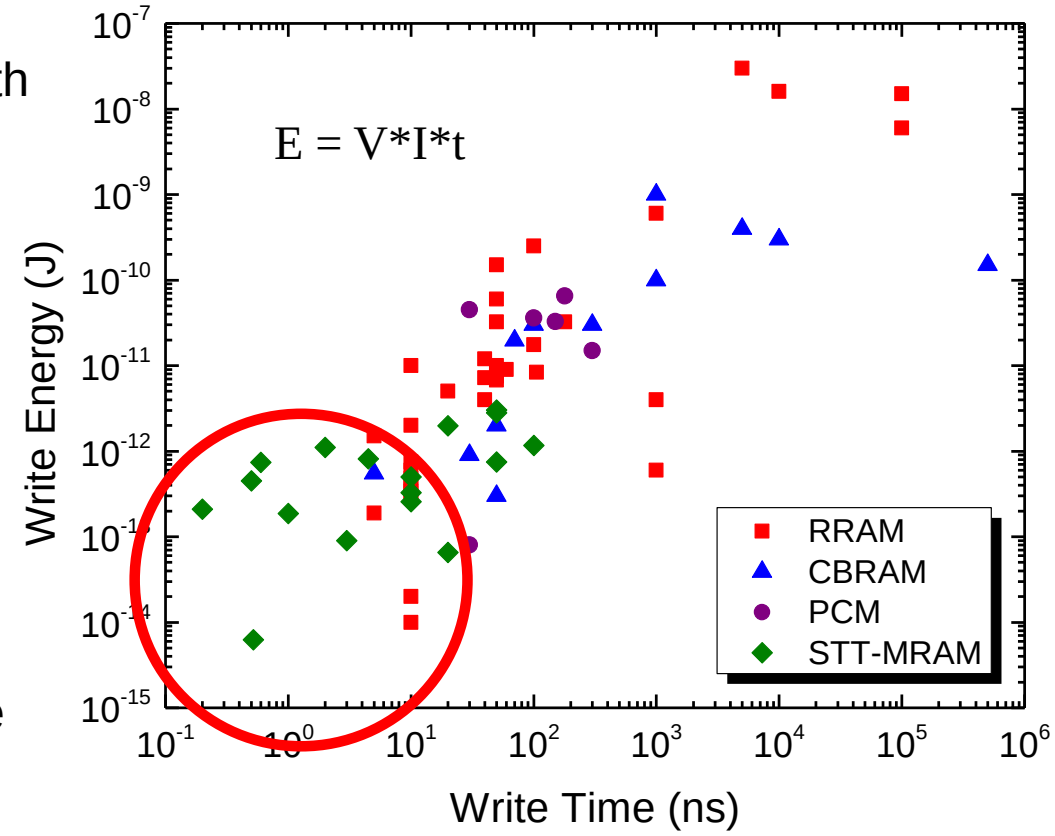
- **STT-MRAM**
  - Spin-transfer-torque magnetic RAM
  - Changing orientation of spin changes the conductivity
  - Advanced material stack, 3D compatibility unlikely



Samsung AIT News (2011)

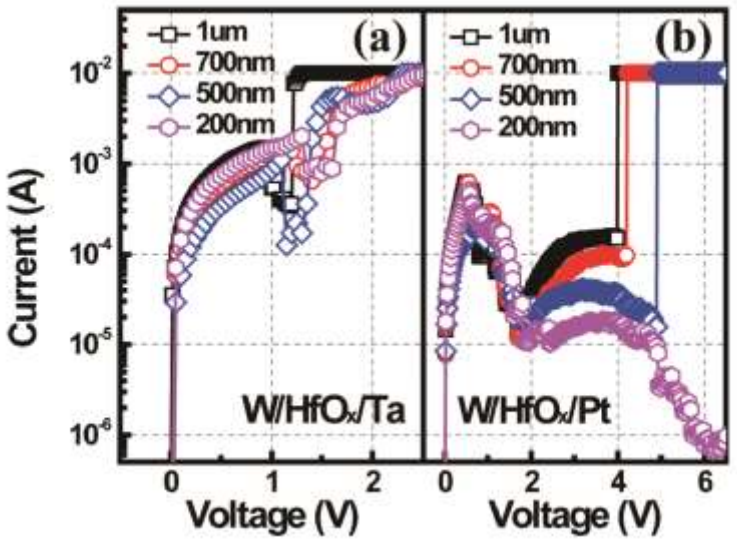
Overview of different ReRAM Technologies

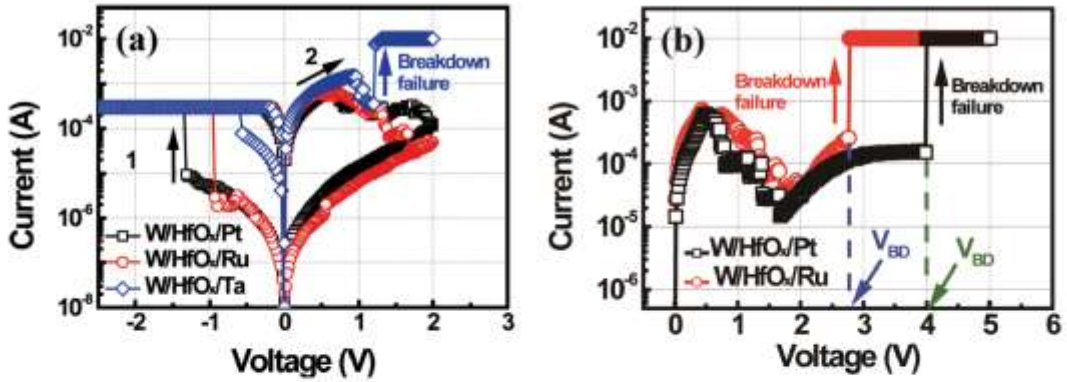
- **RRAM (oxRRAM)**
  - Anode filament, oxygen vacancies form conductive path
  - High endurance, 1^{12} cycles at device level
  - 3D compatible
- **CBRAM**
  - Cathode filament, bridging with metal ion movement
  - Similar structure to RRAM
  - Endurance questionable (finite number of switches)
- **PCRAM**
  - Phase change memory, a flash heating switches dielectric film between amorphous and crystalline state
  - Endurance questionable (finite number of switches)
- **STT-MRAM**
  - Spin-transfer-torque magnetic RAM
  - Changing orientation of spin changes the conductivity
  - Advanced material stack, 3D compatibility unlikely

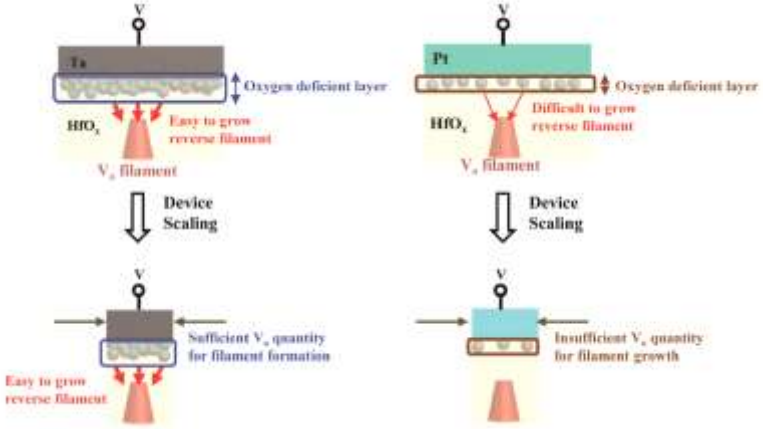


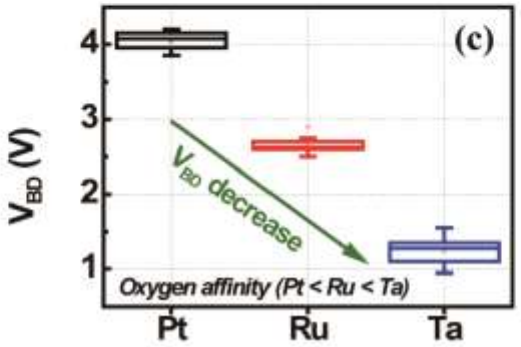
Stanford Memory Trends, H.-S. P. Wong et al
https://nano.stanford.edu/stanford-memory-trends

RRAM – BE Metal Considerations



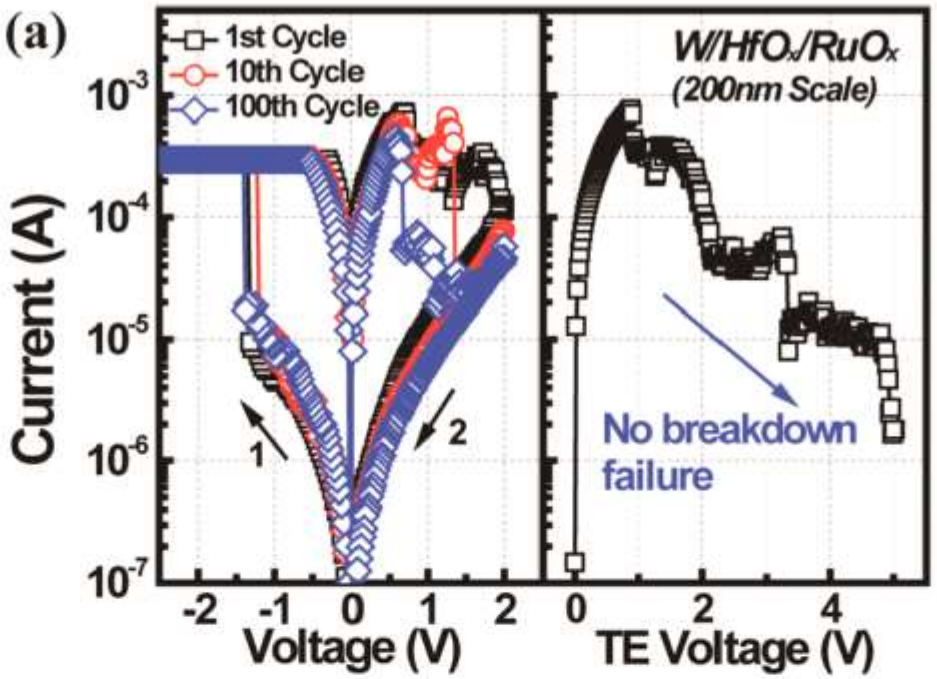


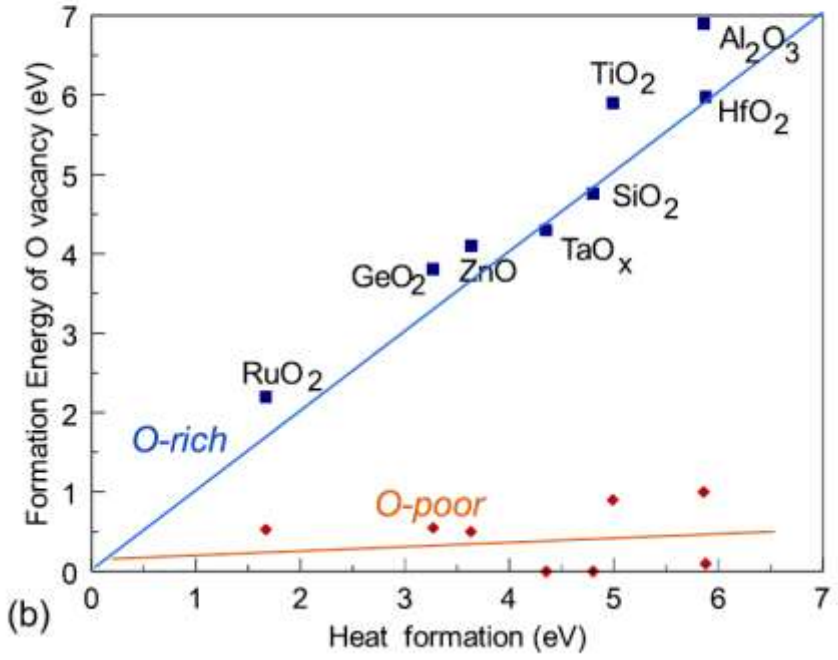




J Park et al – AIP 2016

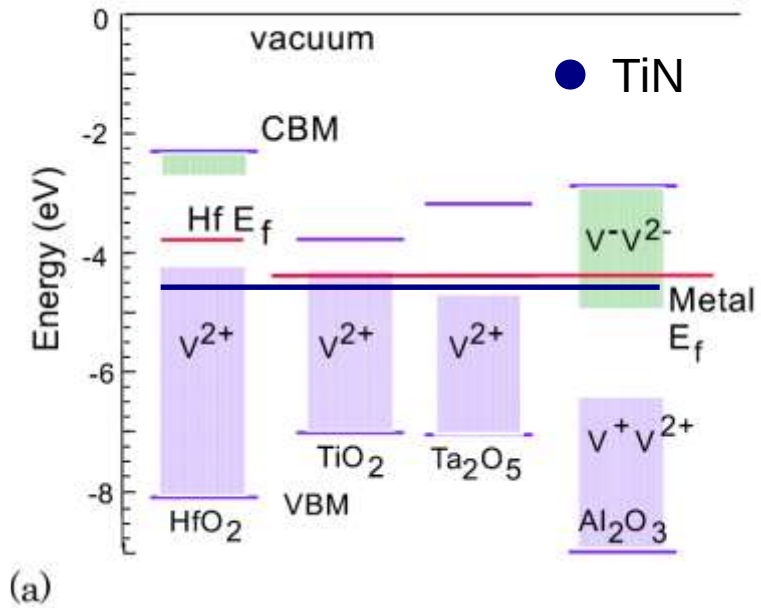
RRAM – BE Metal Considerations

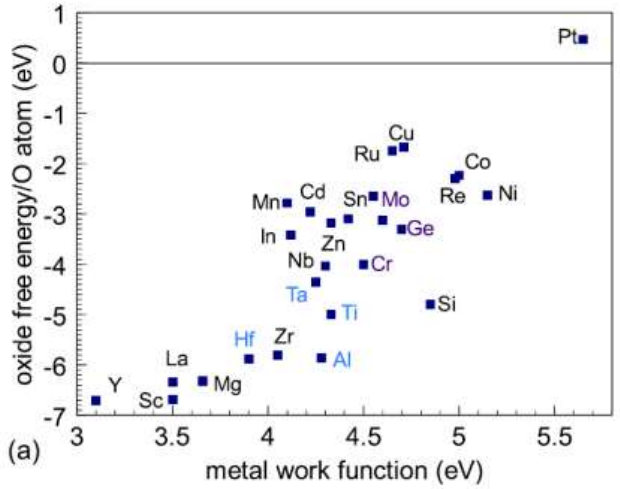


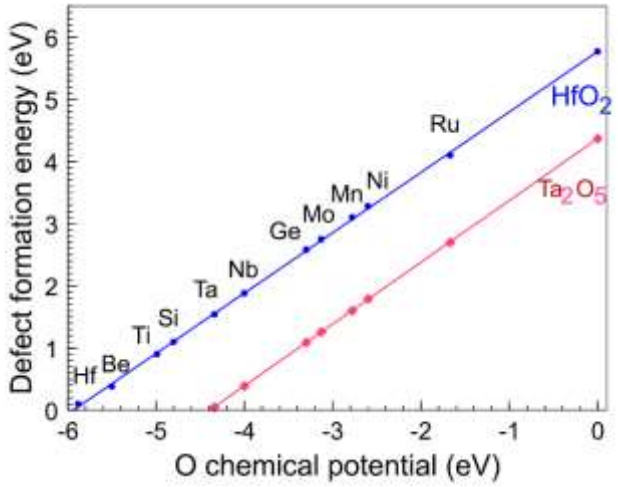


J Park et al – AIP 2016

RRAM Oxide and TE Properties



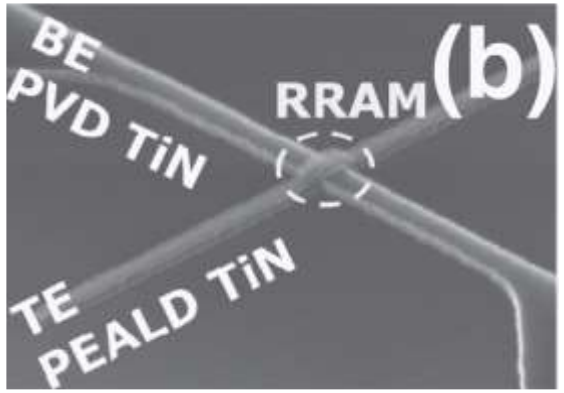


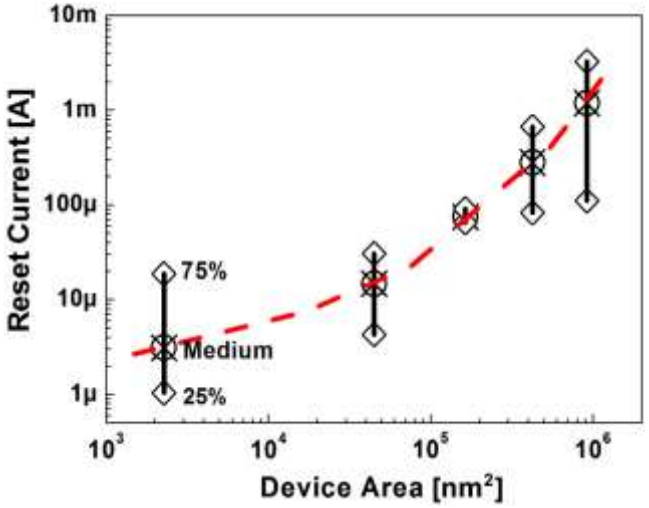


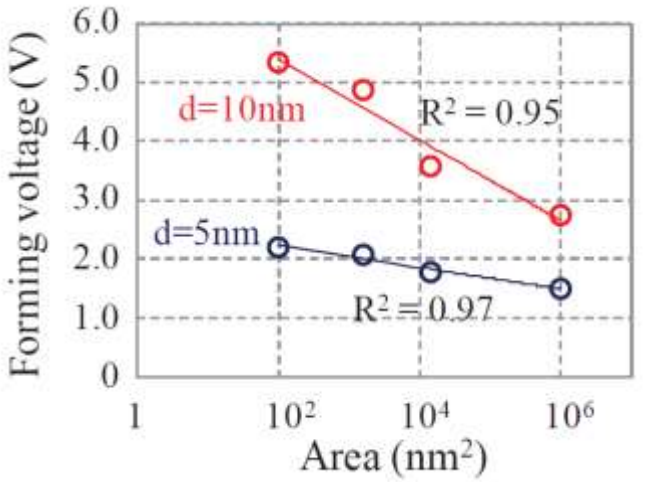
Yuzheng Guo et al – APL 2014

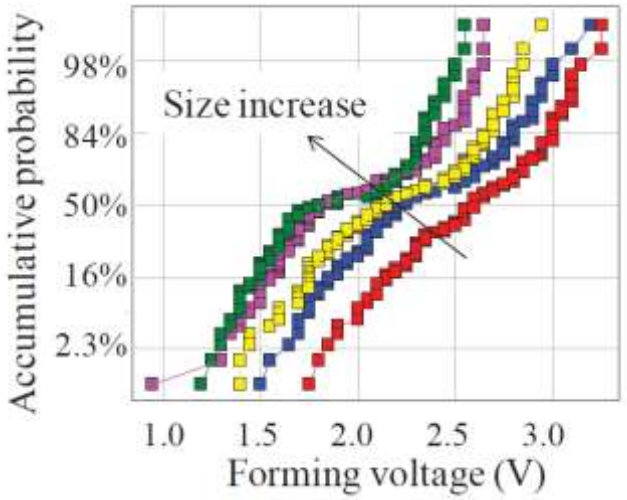
RRAM – Area Scaling

- RRAM switching is ideally independent of device area as only one filament forms
- Area dependence is instead partly coupled to self-capacitance, and a reduction in parasitic current discharge
- However, the probability to form a filament increase with area
- To reduce increased forming voltage and spread in the distribution, surface roughness and material quality at the interfaces are of crucial importance





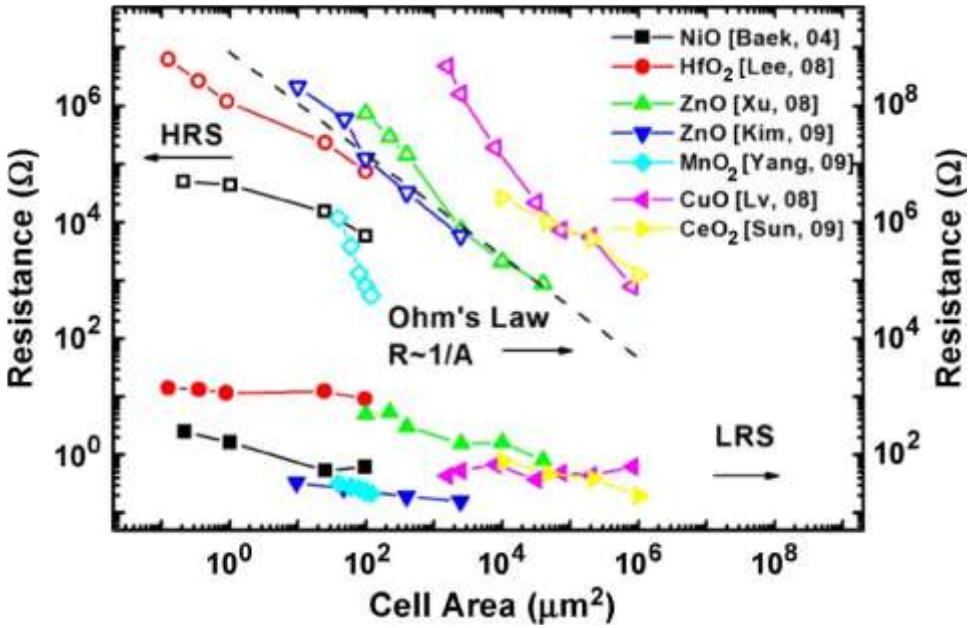


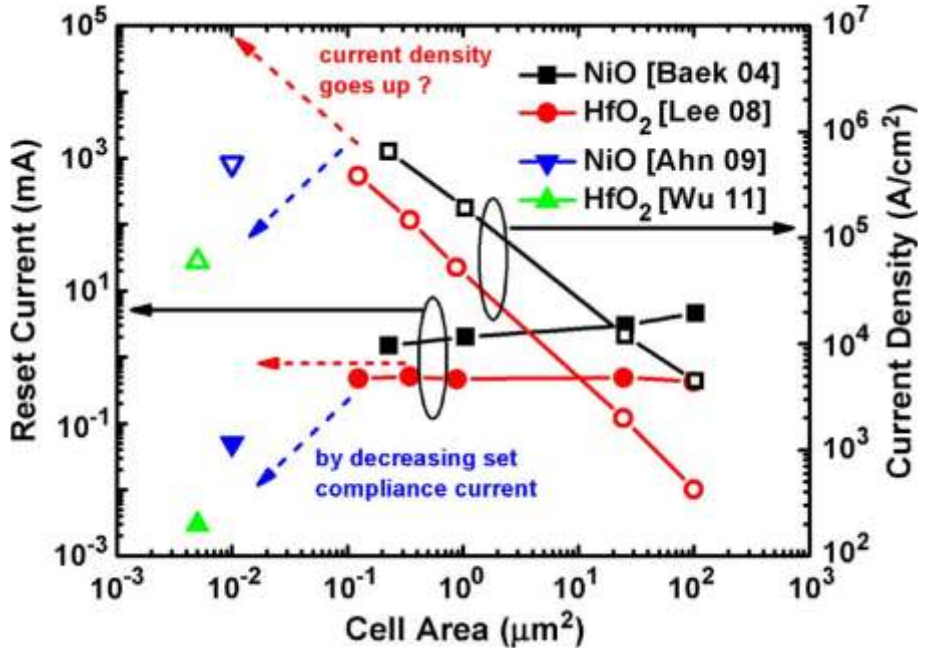


Y. Y. Chen – ME 2013

Ann Chen – Globalfoundires 2013

RRAM – Area Scaling

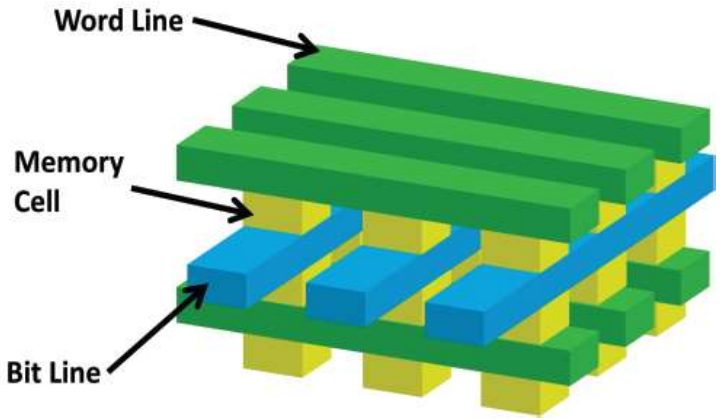


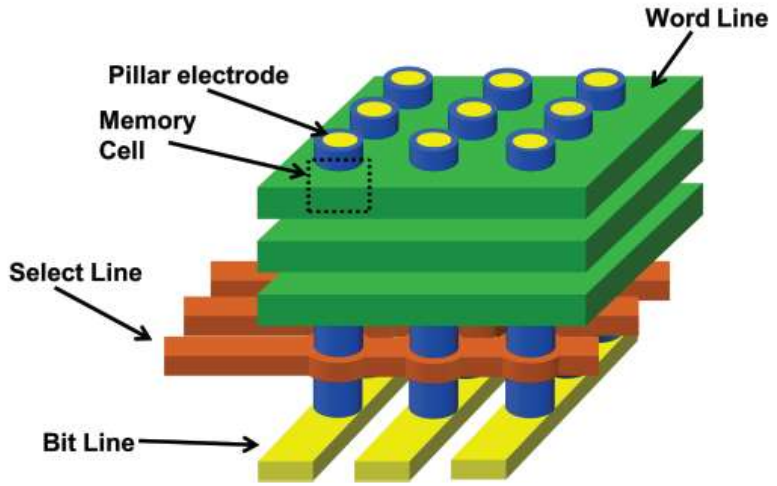


H.S. Philip Wong – IEEE Proceedings 2012

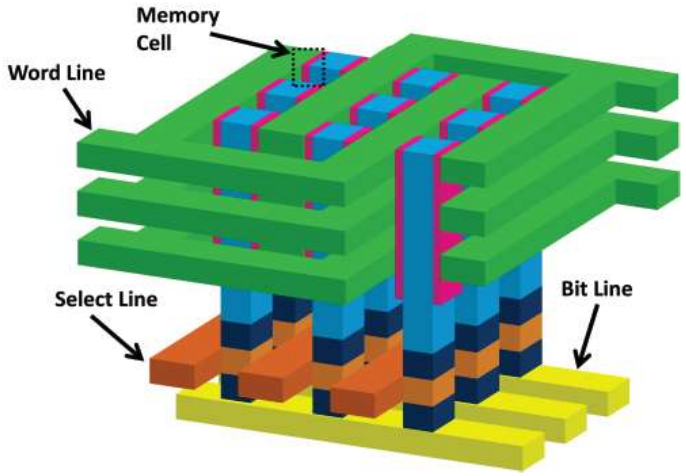
3D RRAM - Architectural Concepts

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



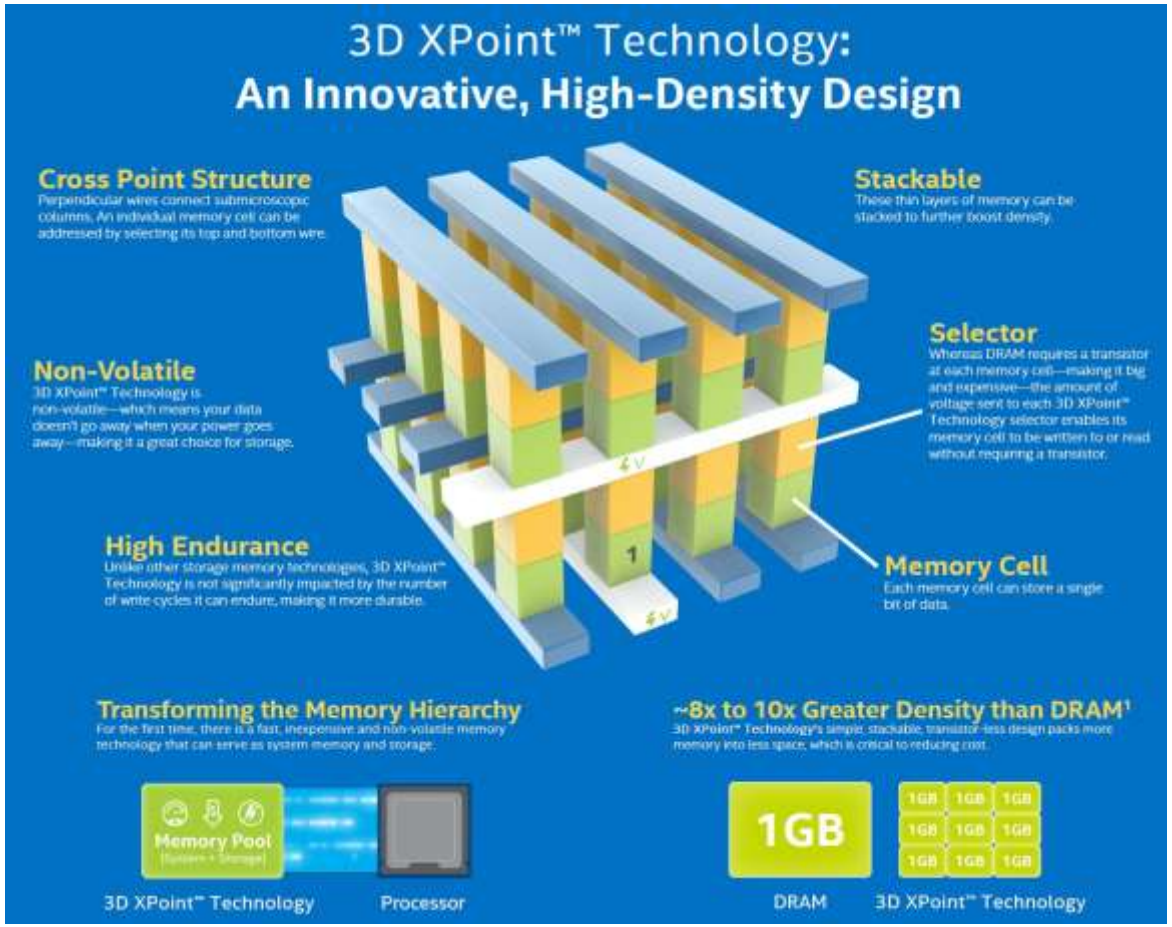
VRRAM type II
- Interconnect capacitance limited performance
- Litho-free stacking
- Bit performance improves with # layers
- Most scalable/cost efficient

S. Yu et al – ISCAS 2014

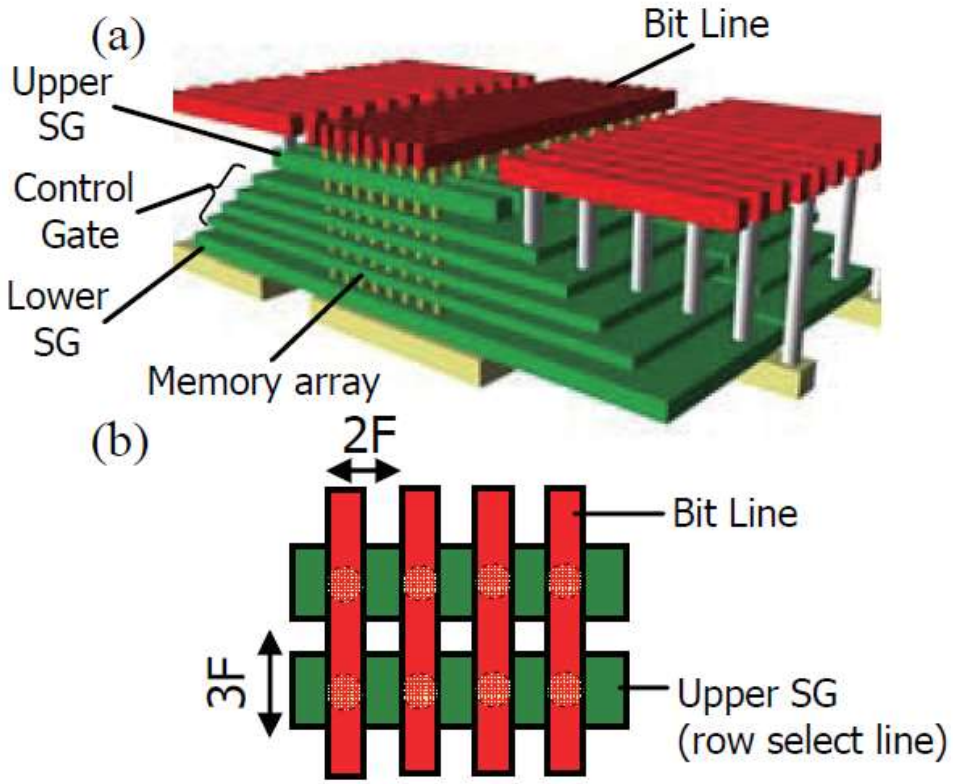
HRRAM – Commercialized

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

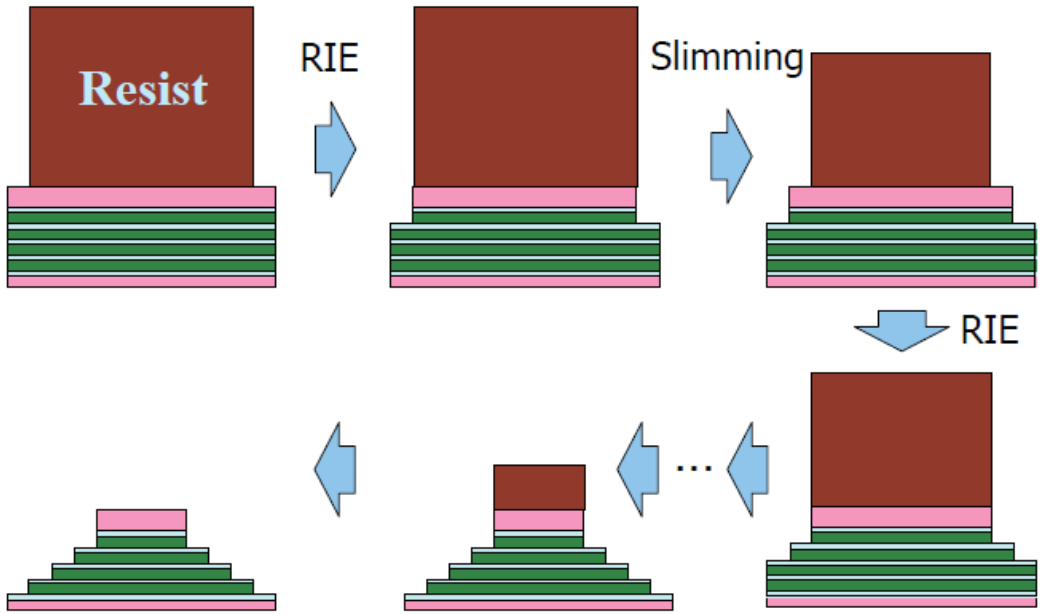
Intel 3D Xpoint
- 3x faster than NAND flash
- 5x more expensive
- Switching mechanism unknown
- Limited stacking potential due to diode selector



Vertical RRAM – Lithography Free



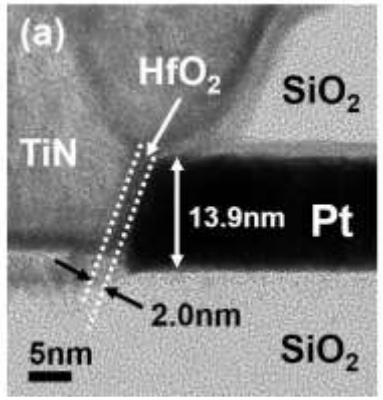
Litho-free formation of a stair-case structure

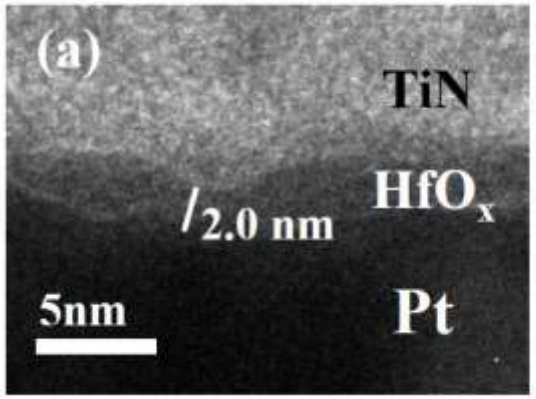


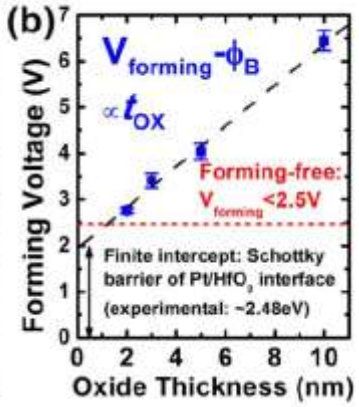
Tanaka et al – VLSI symposium 2007

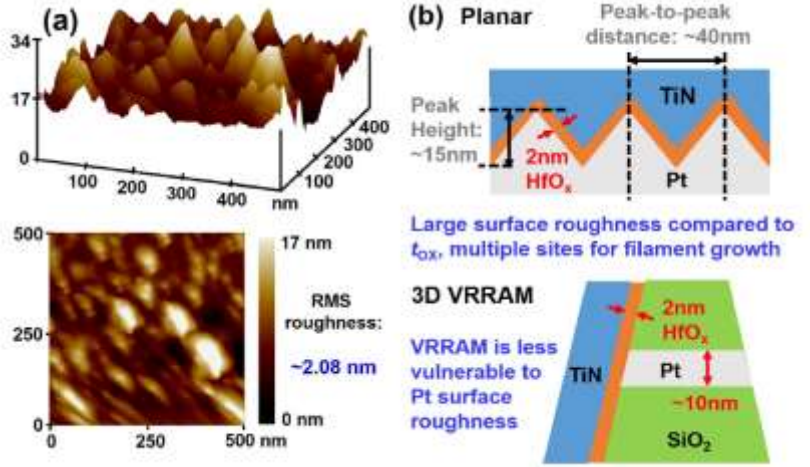
RRAM - Oxide Thickness Scaling

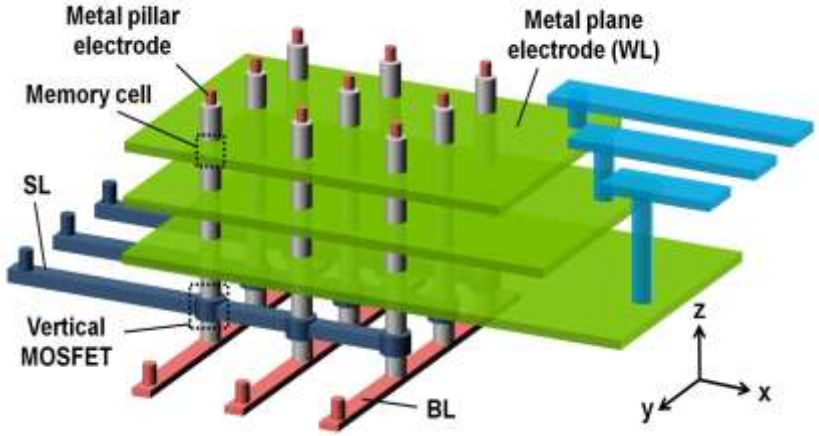
- Scaling the dielectric necessary to reduce minimum feature size
- Surface roughness affect the spread of the performance distribution
- Etched out vertical pillar have a smoother surface than deposited metal







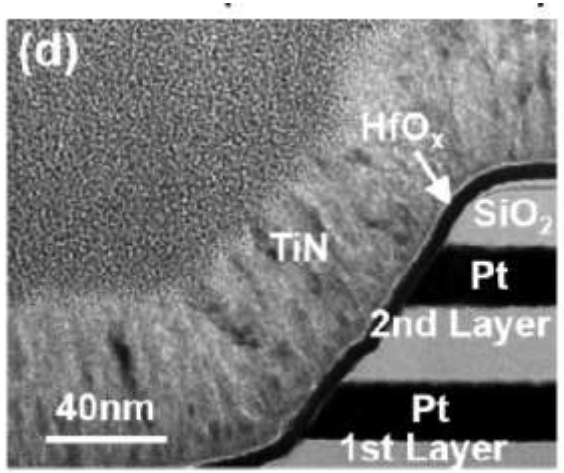


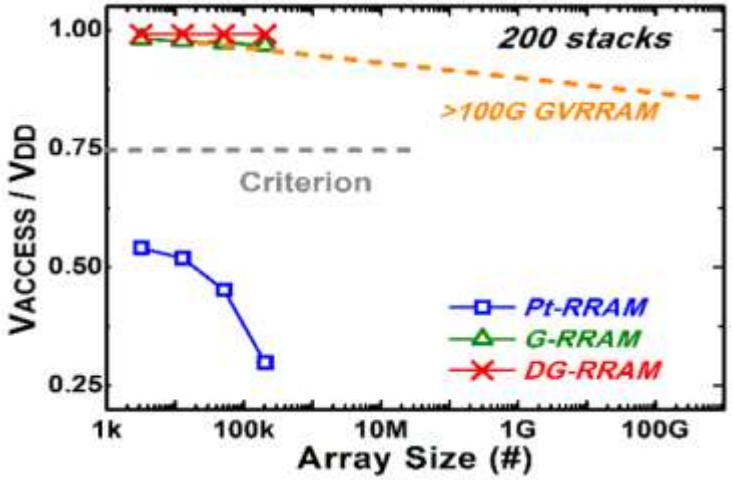


Zhao et al – IEDM 2014

Considerations on Scaled 3D Arrays

- Large arrays require MOSFET selectors to reduce leakage
- Vertical geometry allows for more aggressive thickness scaling as it reduces roughness
- Simulations show metal plane thickness will limit array size due to resistance of the vias, sub 6-nm metal is highly resistive
- Graphene and other 2D materials way become a viable way forward for large scaled arrays

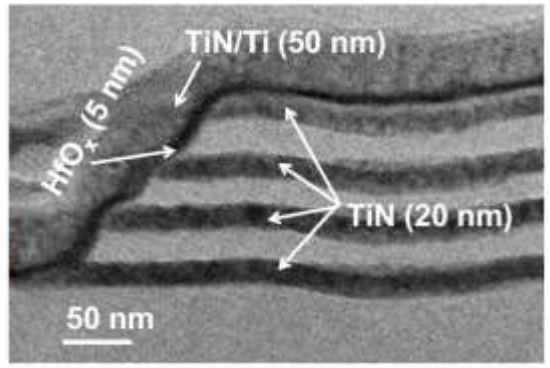


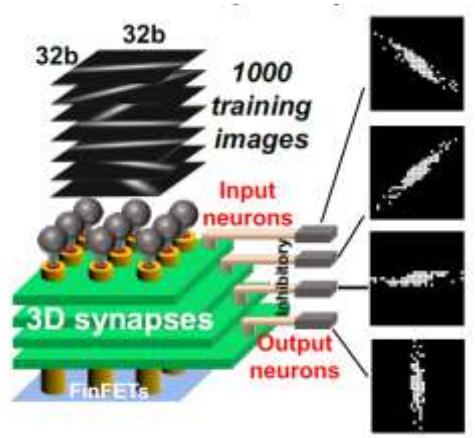


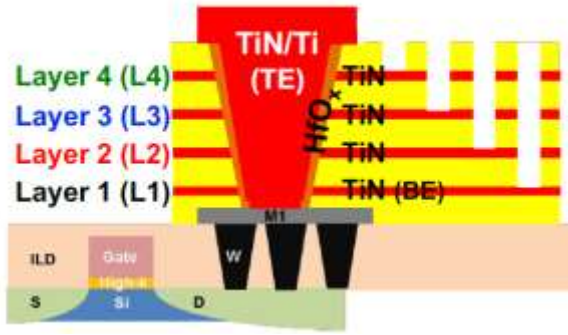
Joon Sohn – IEDM 2014

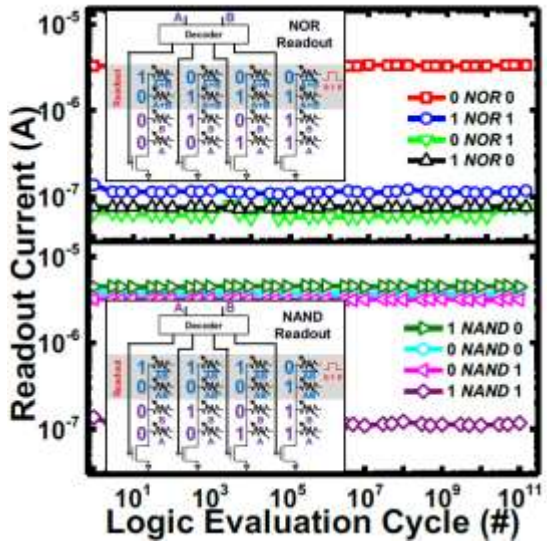
Research Trend II – In Memory Computations

- Dense, fast, and energy efficient memory opens for computations directly in the memory
- Permutations are made along the vertical pillars
- NOR and NAND operations can be accomplished with propagation of specific pulse trains
- In memory computations implementations could be task specific, drastically reduces integration area





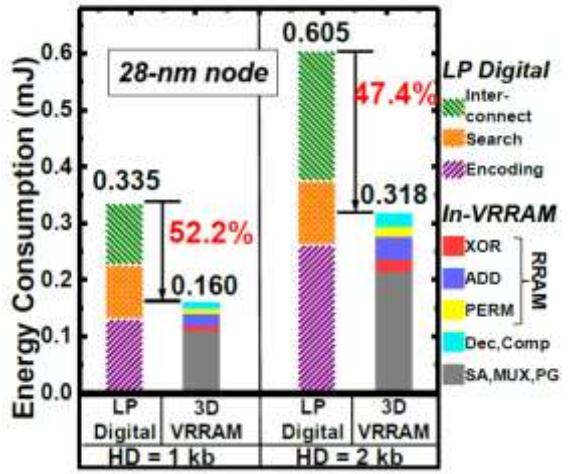


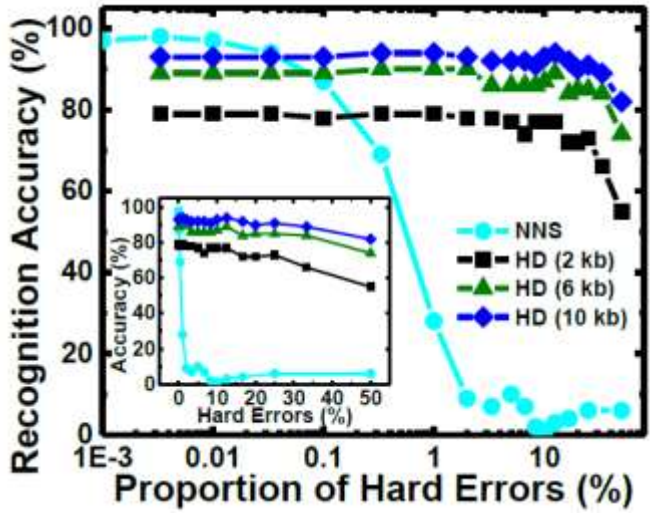


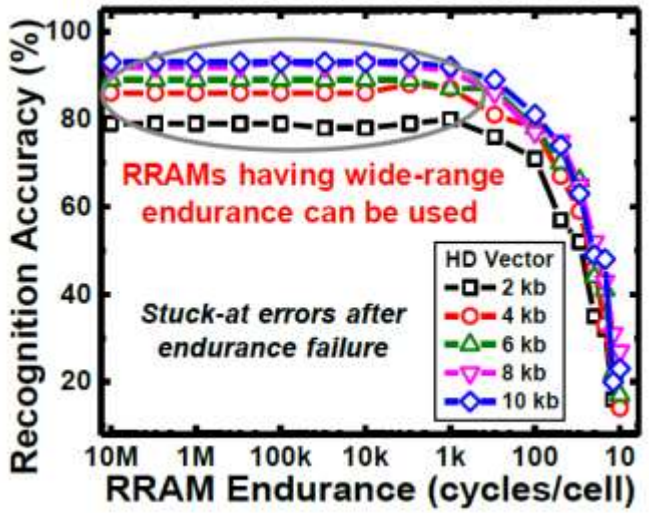
H. Li et al – VLSI 2016

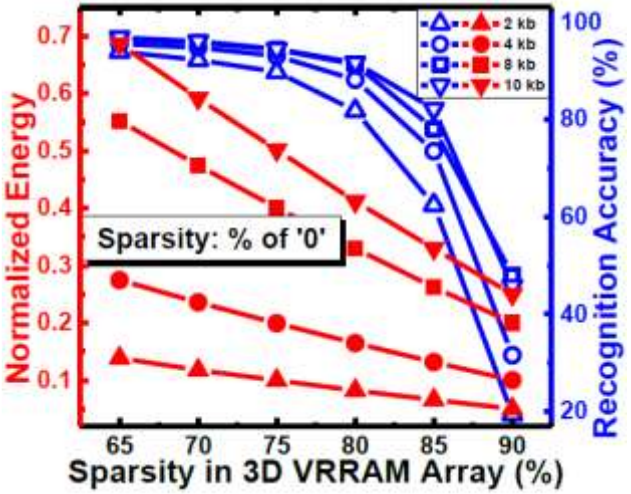
Research Trend III – Hyper-Dimensional Vectors

- 10,000 bits/vector, any type of object can be represented
- Orthogonality factor between vectors determine object resemblance
- Doing in-memory reduces energy consumption slightly, but lowers chip area drastically (660x for a 2k-vector)
- Error resilience of HD computing mean that a wide range of RRAM technologies can be used
- Tricks such as generating random vectors with a high percentage ‘0’-s may be a way to reduce the power



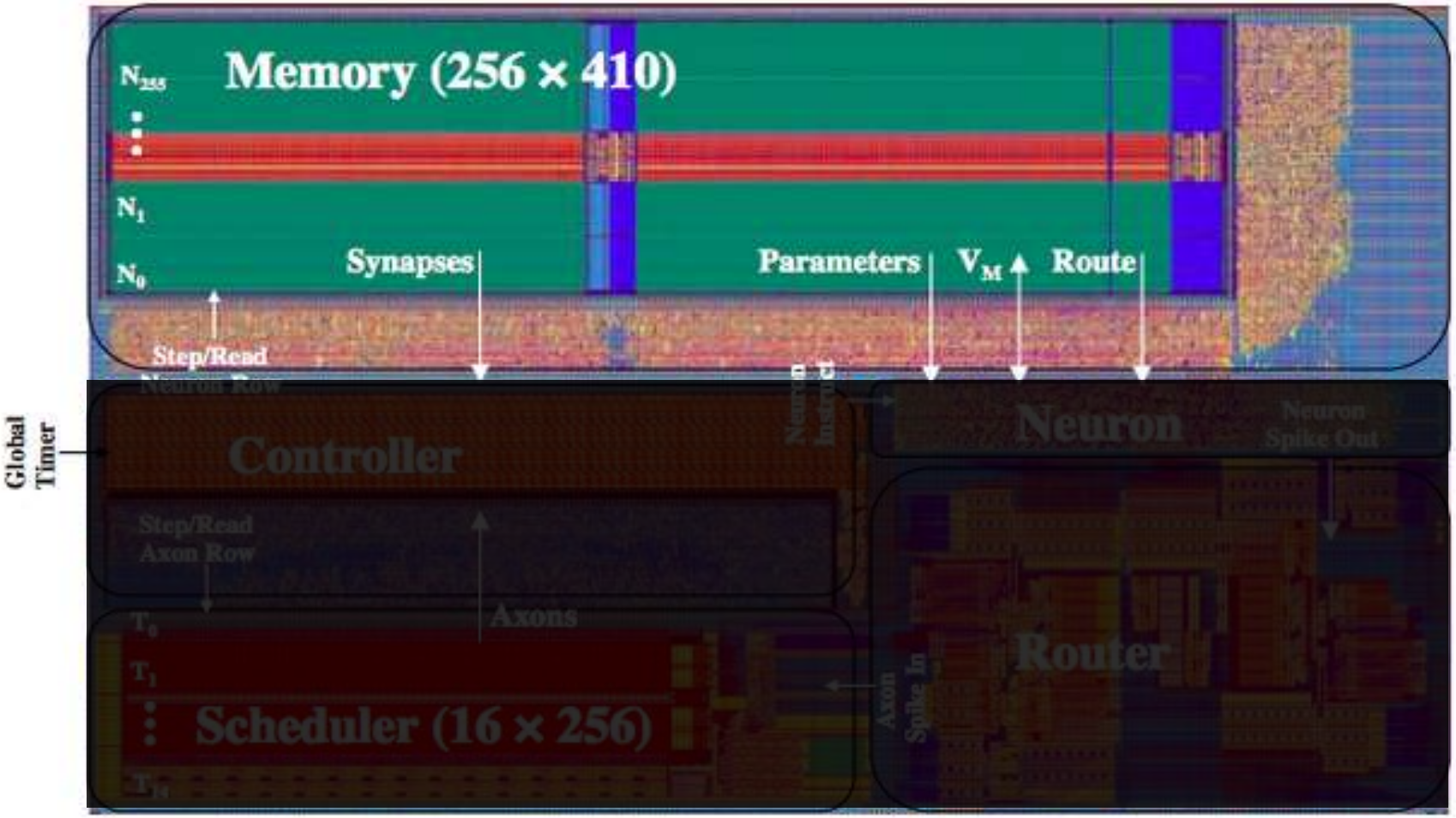






H. Li et al – IEDM 2016

Implications for Neuromorphic Computing



IBM TrueNorth - 1 unit chip layout (2014)

Implications for Neuromorphic Computing

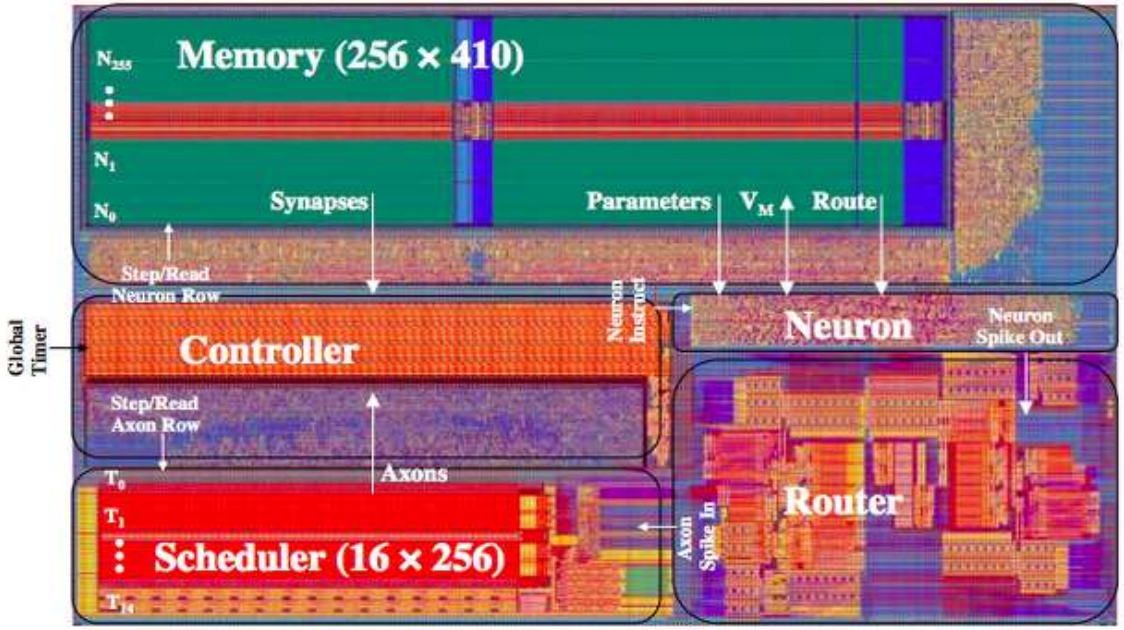
NANO ELECTRONICS GROUP

IBM TrueNorth

- SyNAPSE - DARPA funded initiative to simulate the brain
- Dedicated neuromorphic hardware
- 4096 computational units (1 unit pictured)
- Memory (SRAM) occupies about 40% of the chip area

RRAM

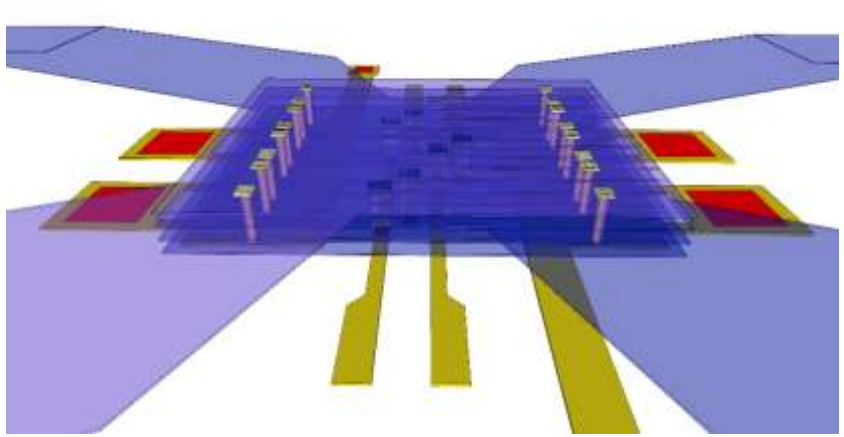
- 3D RRAM with 128 layers
- 64 Tb per chip

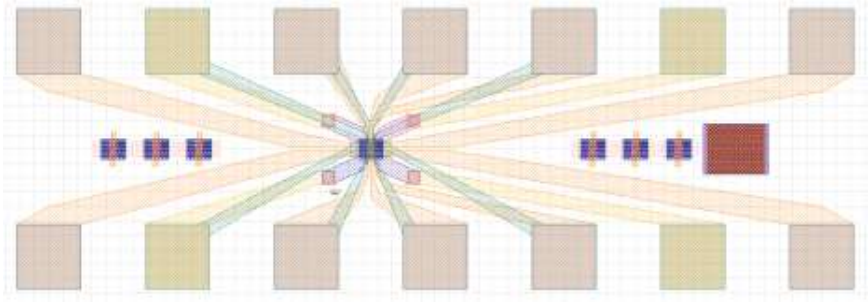


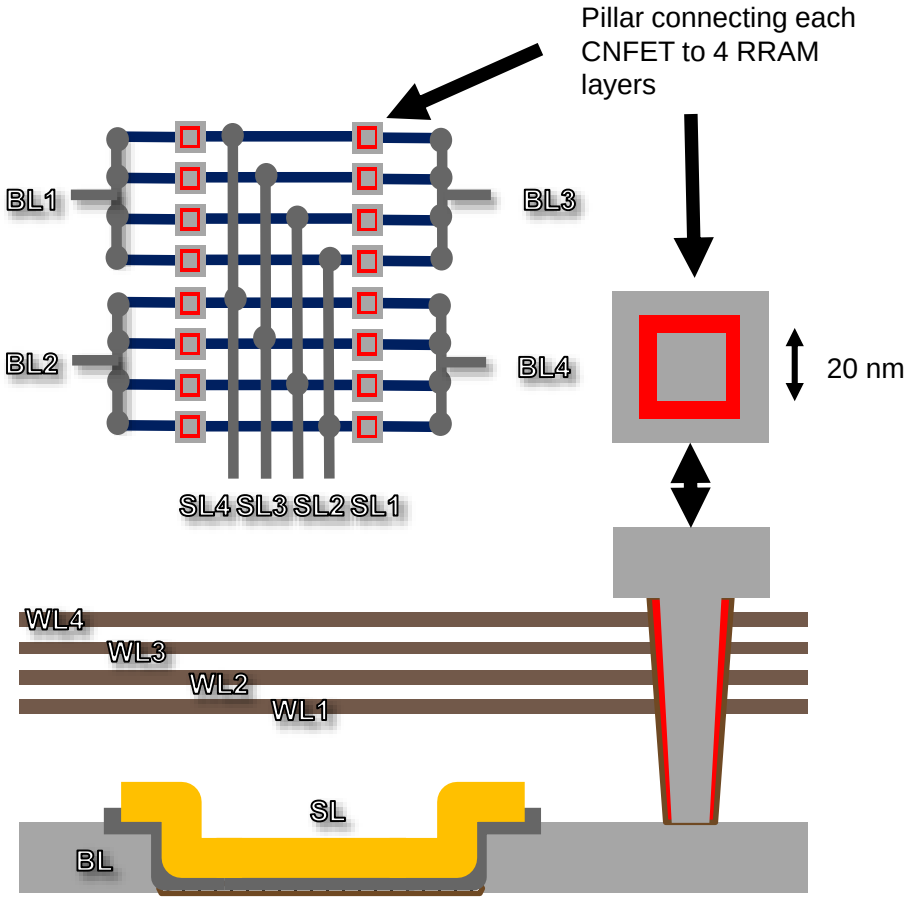
1 unit chip layout (2014)

SRAM ~ 120-140 F^2 → 1T1R RRAM 20x smaller!!

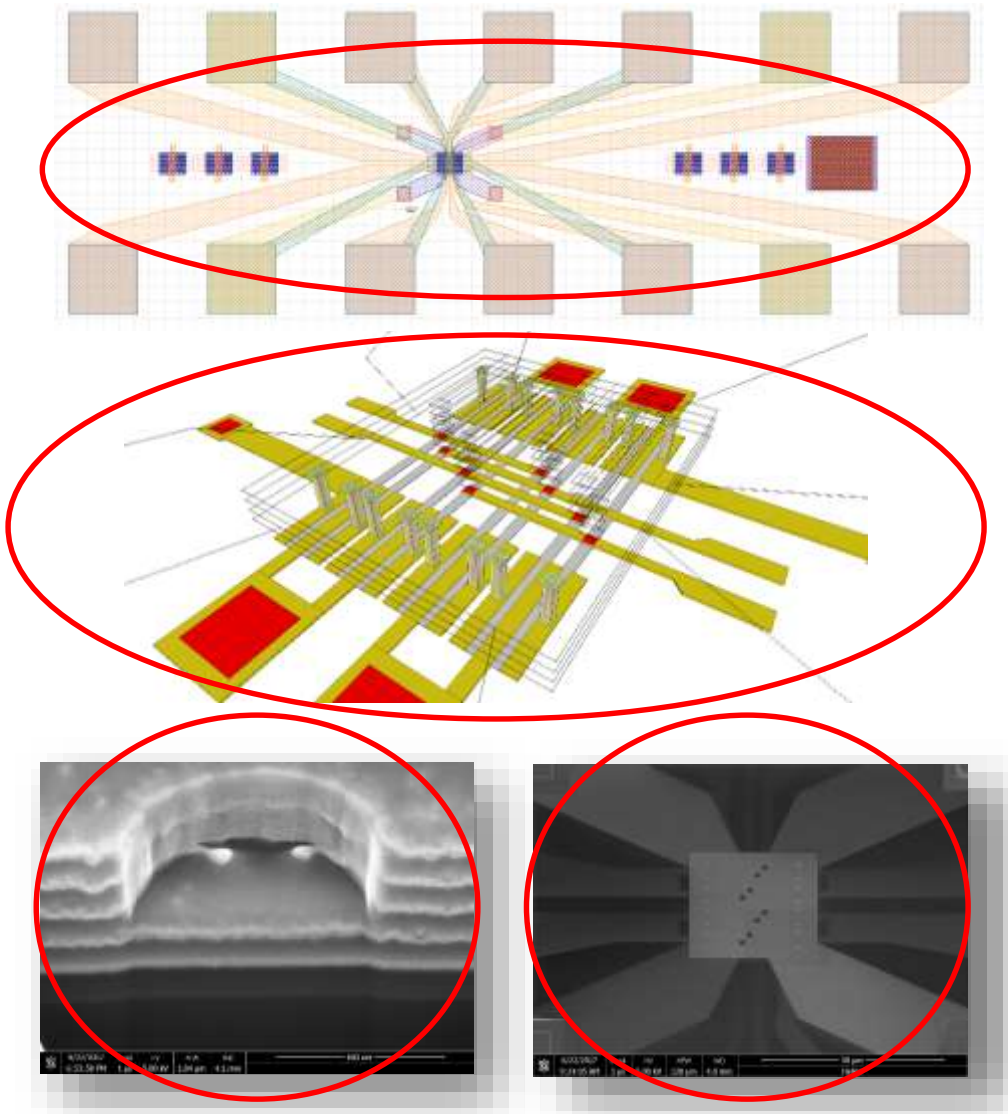
Stanford Research: CNTs with 3D RRAM

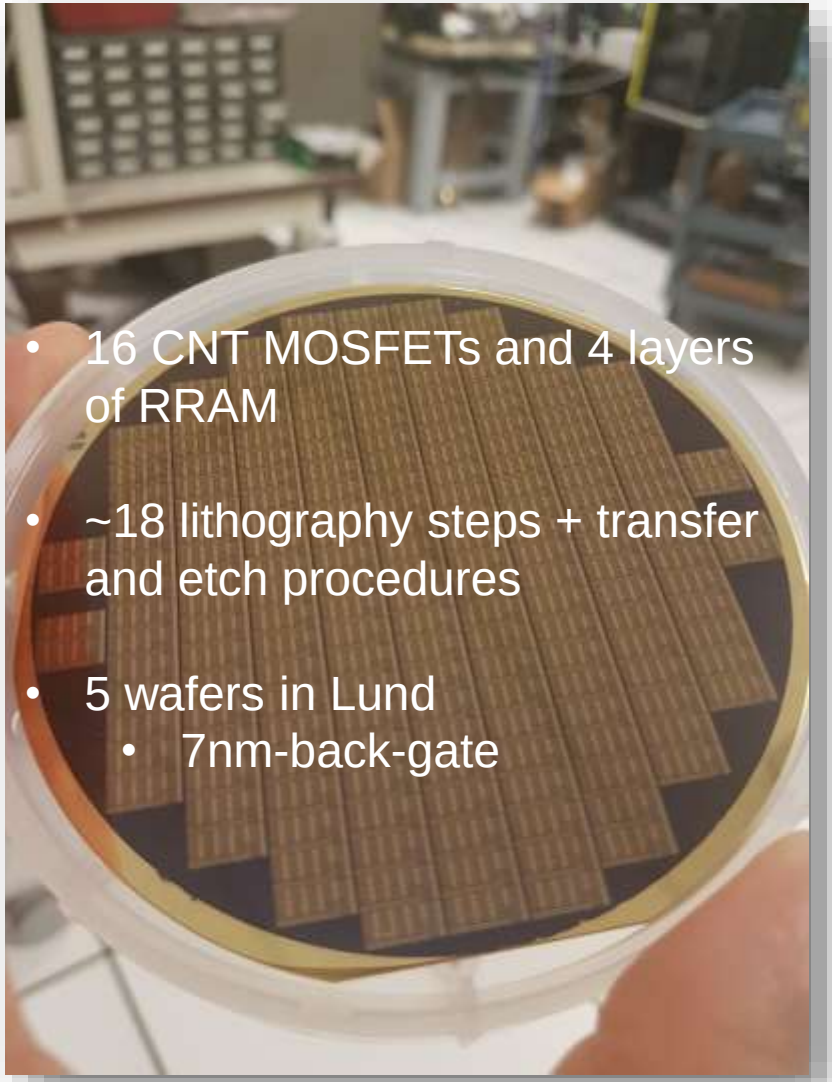






1T4R 64-bit Monolithic Memory Cell





- 16 CNT MOSFETs and 4 layers of RRAM
- ~18 lithography steps + transfer and etch procedures
- 5 wafers in Lund
  - 7nm-back-gate

2D Via-hole RRAM – Accepted at DRC

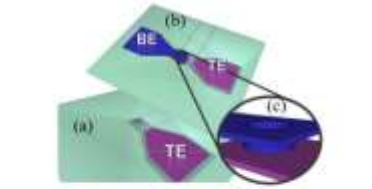


Fig. 1. (a) Schematic illustration of the RRAM structure post ALD of the electrical TE and the oxide, as well spin-coating, baking, and patterning of the lifting layer that has been opened down to the active RRAM interface area and to the TE. (b) Post the electrical BE definition, and (c), a magnification of the via-hole that constitutes the RRAM device.

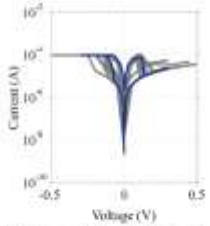


Fig. 2. 10 consecutive switching-cycles for a RRAM device corresponding to sample A. The switching is performed without any current limiting device in series and achieves a 10x high to low resistance ratio while switching at below ± 300 mV.

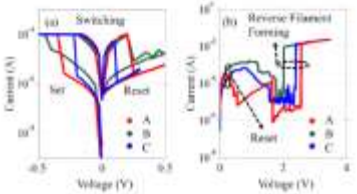


Fig. 3. (a) Low-voltage switching curves showing set (negative voltage) and reset (positive voltage) for three different devices of the different types, and where the current is limited by the self-compliant ITO electrode. (b) Devices biased at large positive voltages to the point where they form a reverse filament. Although they do not break due to the nature of the ITO, they have to go through extensive re-resetting to be able to switch as before.

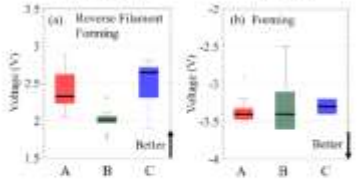


Fig. 4. (a) A boxplot of the RFF voltage distribution, where each box contains the data from 10 different devices. The RFF is lower for the device B, indicating that the film has more vacancies accumulated at the ITO-HfO_x interface. (b) A boxplot of the forming voltage showing little difference between the different samples.

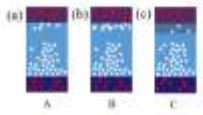
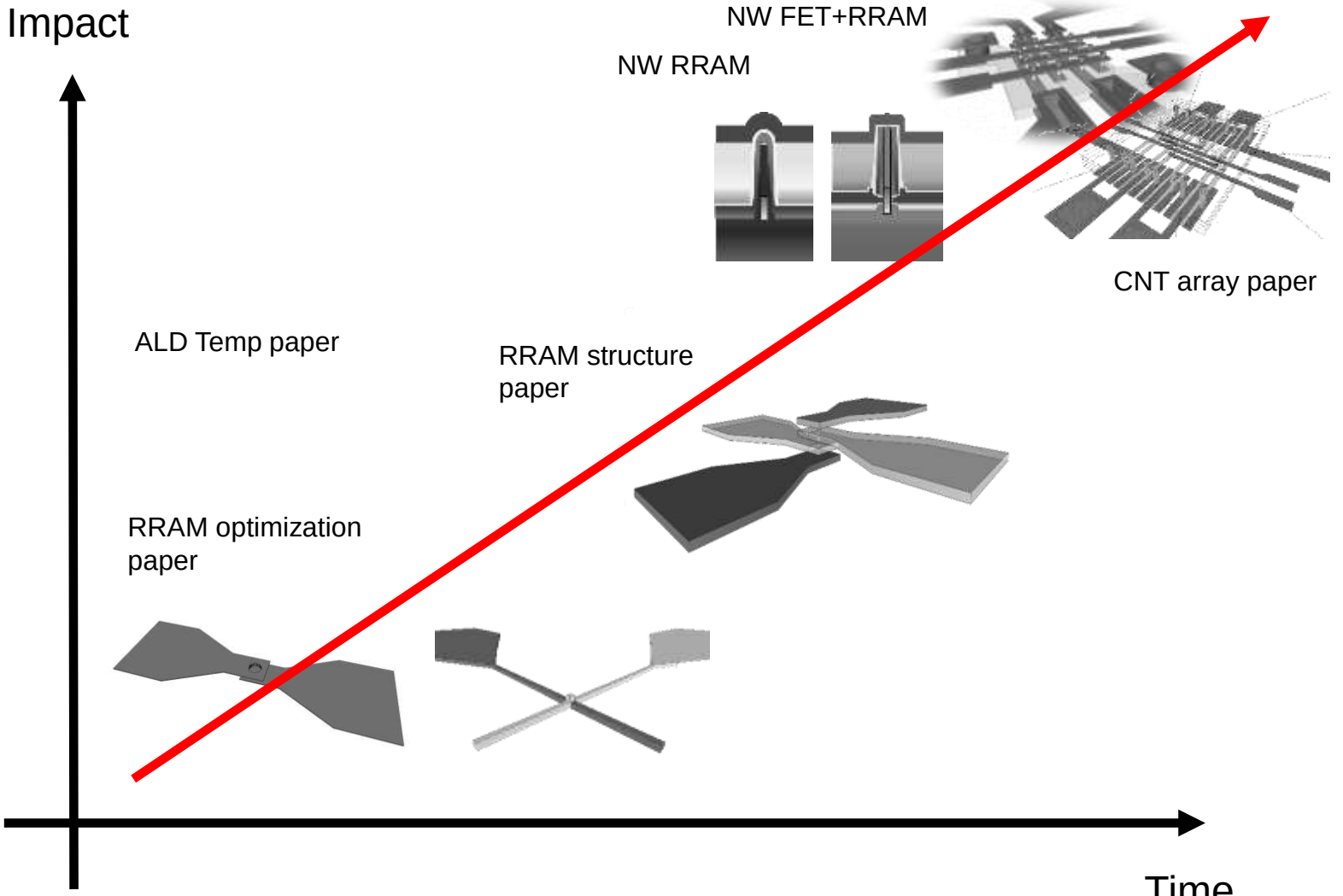


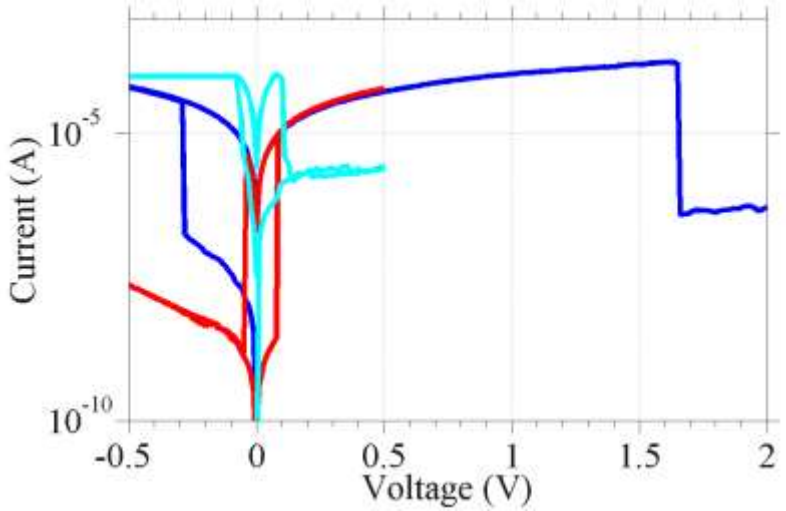
Fig. 5. (a) Schematic vacancy distributions for the different studied devices, where more vacancies at the electrical BE are likely to reduce the RFF voltage.

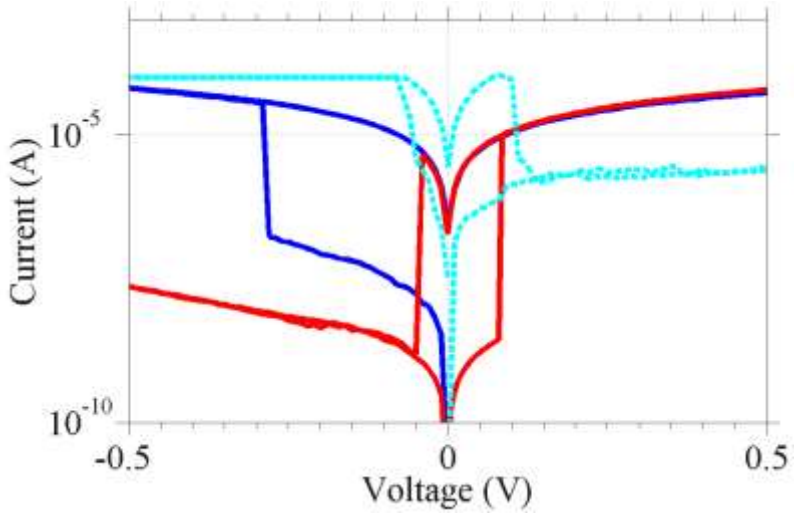
TABLE I: OXIDE DEPOSITION PARAMETERS

Sample	T (°C)	ALD cycles	t_{ox} (nm)
A ●	300	27 HF	3.2
B ●	300	27 HF	3.2
C ●	300	22 HF / 5 Al	3.2



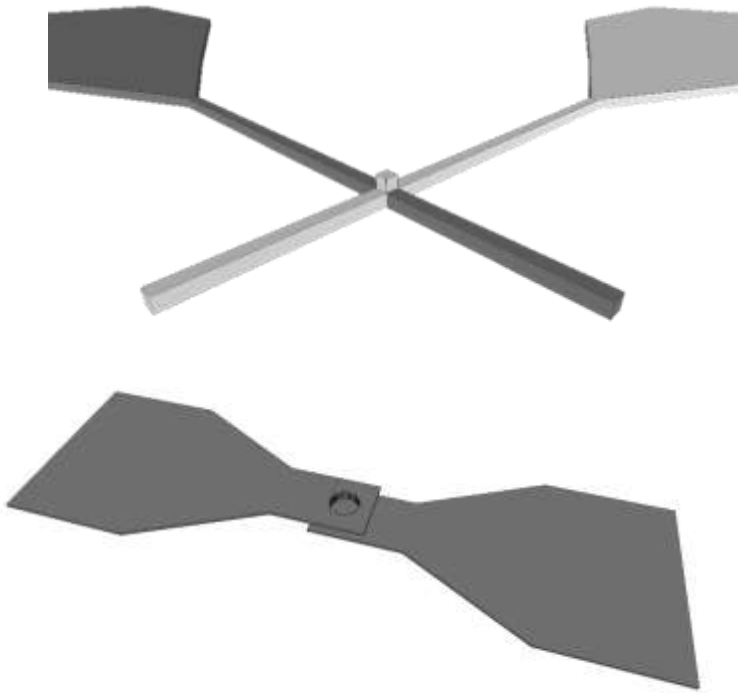
Lund RRAM Roadmap – 5-Å-Al/22-Å-Hf

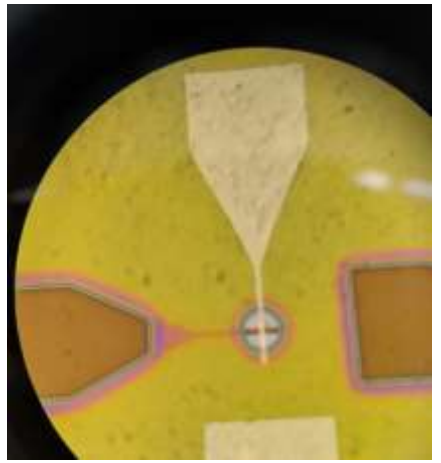




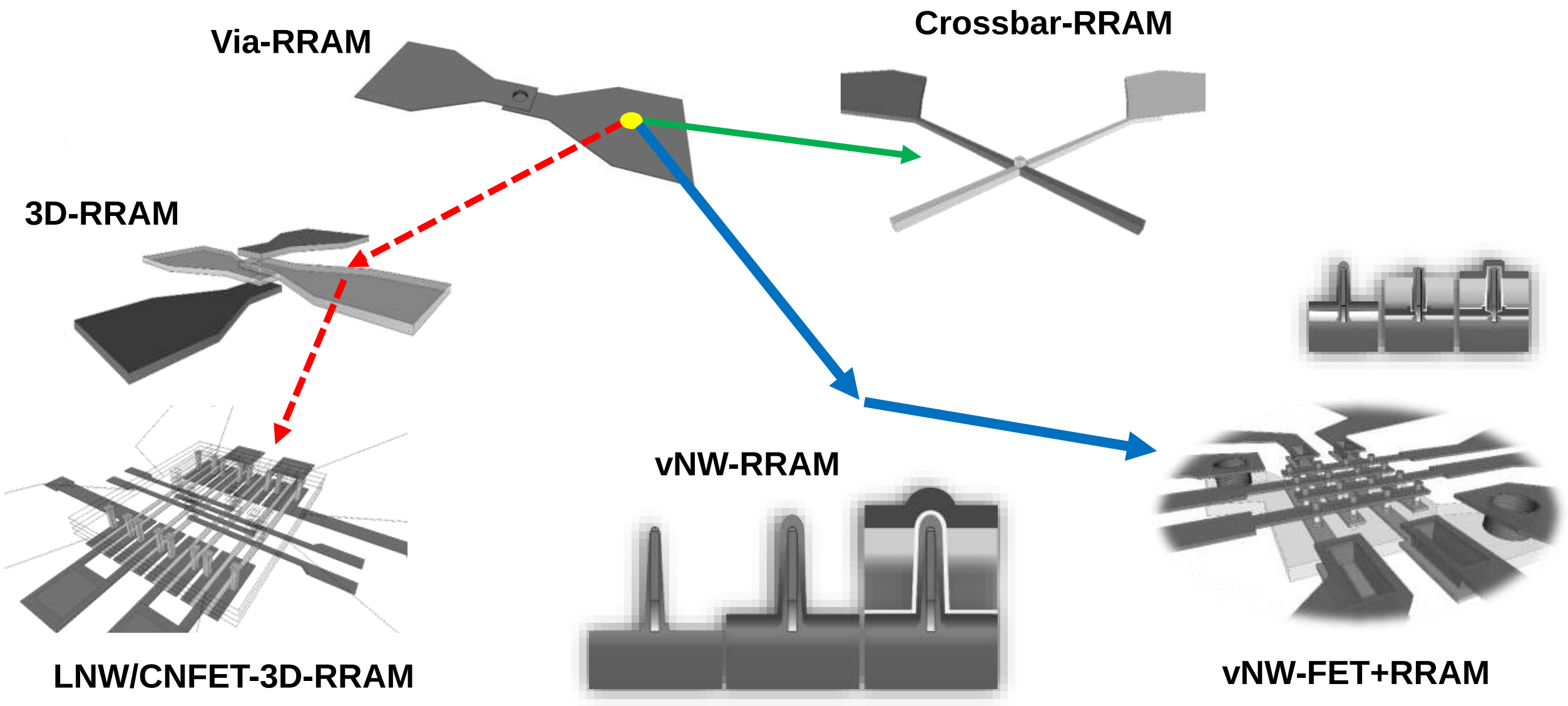
Forming
Ti - 1.55 V
TiN - 2.3 V

TiN - Cross
TiN - Via
Ti - Cross

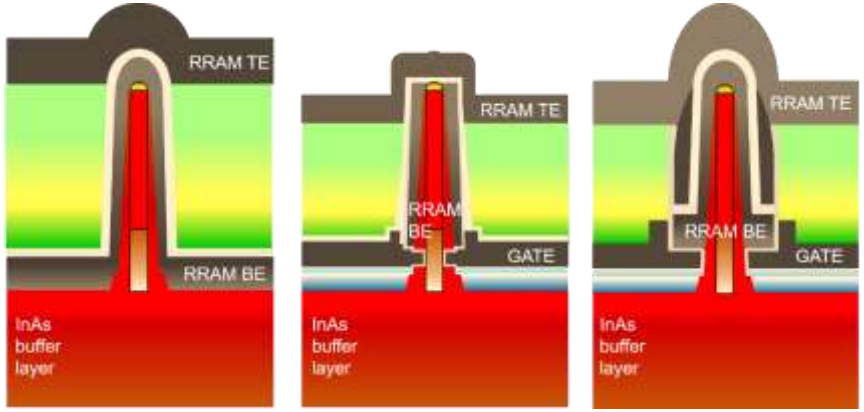


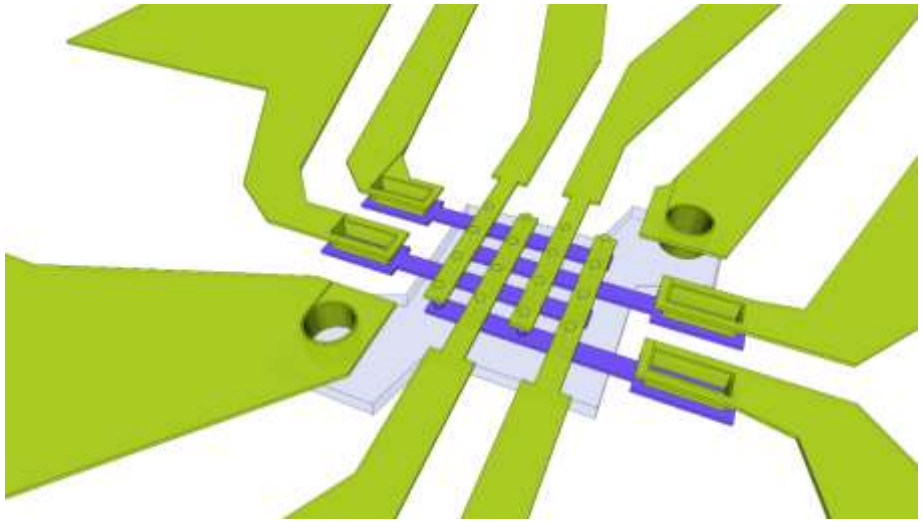


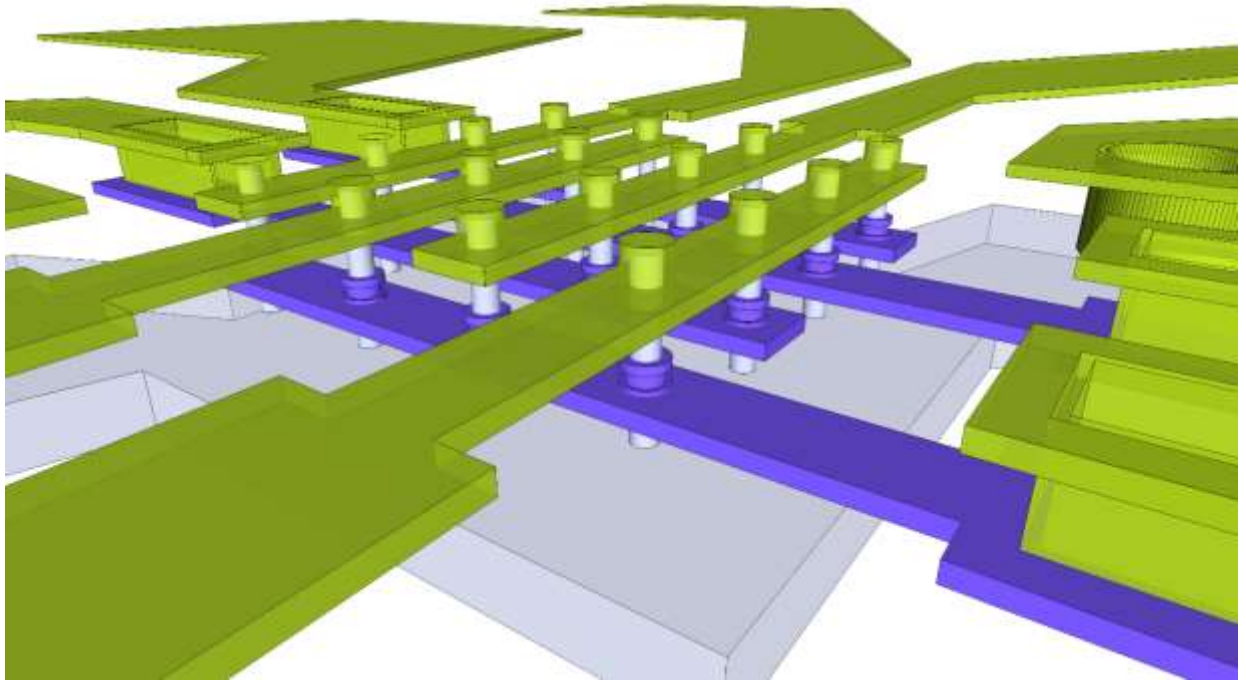
Lund RRAM Roadmap - Structures



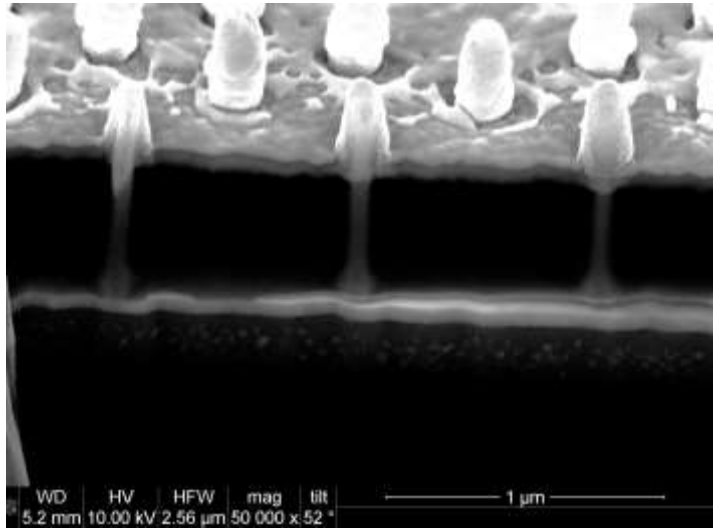
NWFET-RRAM 4x4 Cross-bar Array

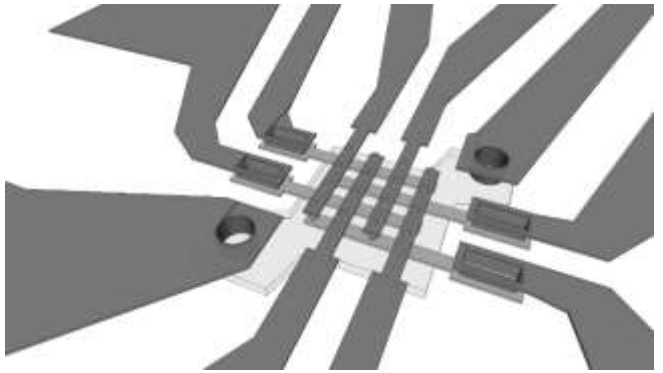






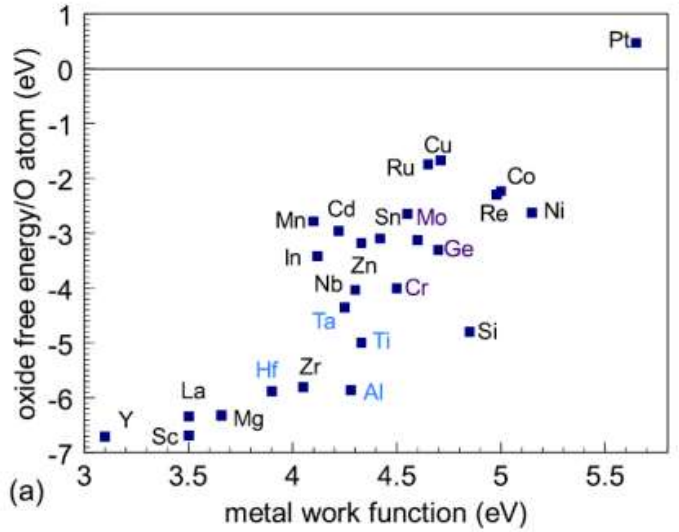
FET Controlled RRAM Measurement – NW1

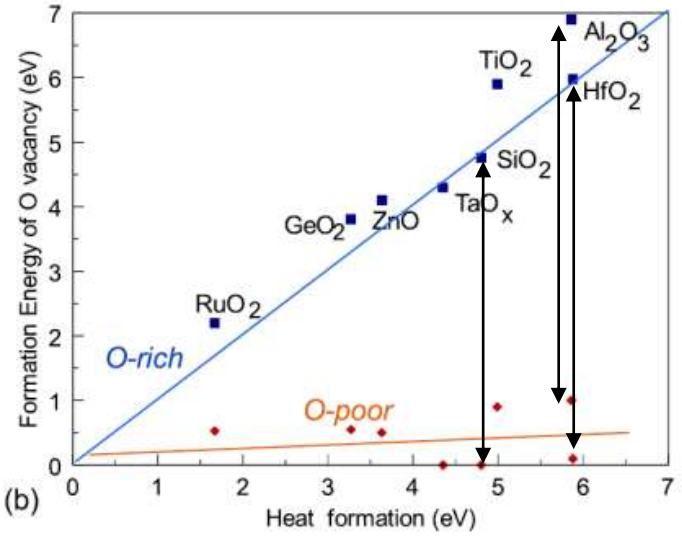




RRAM Oxide and TE Properties

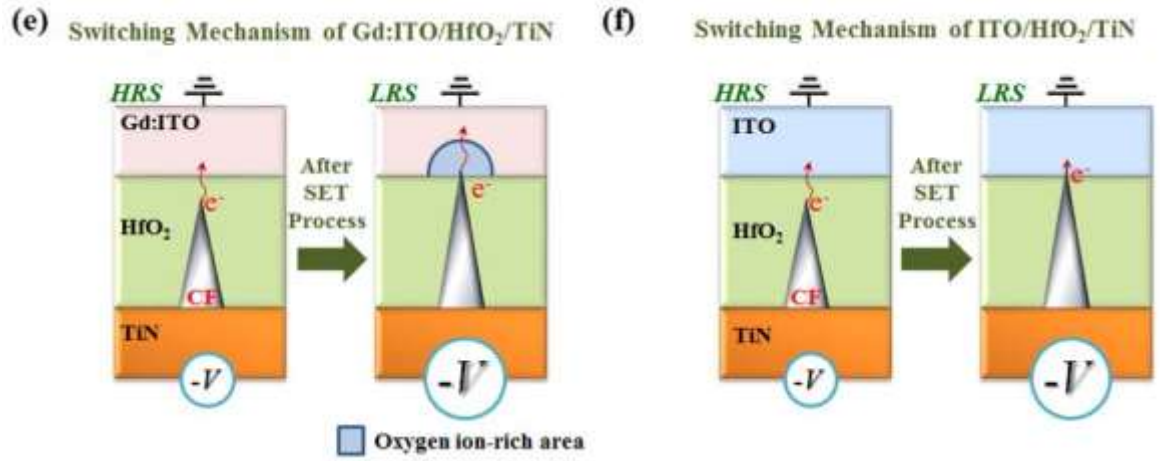
Compound	Stable condition*	Structure	Heat of formation of compond (eV)	Formation energy of oxygen vacancy** (eV)		Band gap in perfect cystral (eV)	Defect level (eV)		
				$q = 0$	$q = +2$		$\varepsilon(+2/0)$ (eV)	$\varepsilon(+2/+1)$ (eV)	$\varepsilon(+1/0)$ (eV)
Al_2O_3	hypo. NP	bixbyite	−16.15	7.14	0.18	6.09	3.48	*3.80*	*3.16*
	normal	corundum	−16.35	7.08	1.32	6.90	2.88	*3.07*	*2.69*
In_2O_3	normal	bixbyite	−7.81	1.53	0.29	1.05	*0.62*	0.56	0.67
	HP	corundum	−7.67	1.63	0.23	1.00	*0.70*	0.68	0.72
SnO_2	normal	rutile	−6.04	3.49	2.41	1.26	*0.54*	0.50	0.58
	HP	fluorite	−5.05	−0.45	−0.41	0.29	*−0.02*	−0.05	0.01

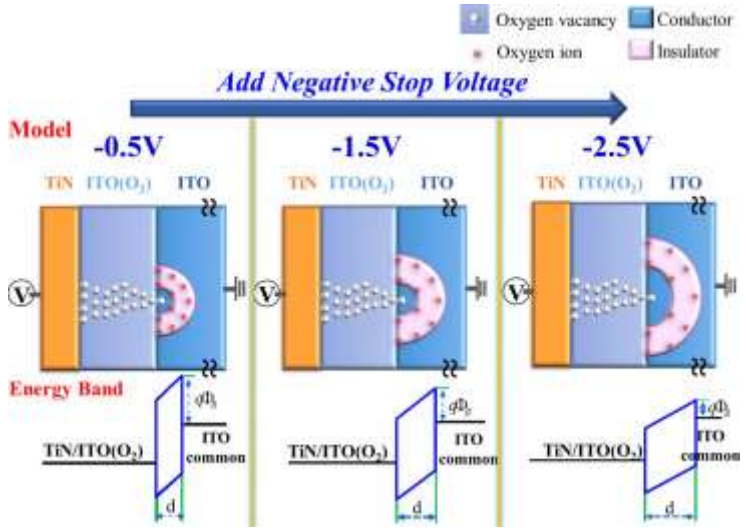


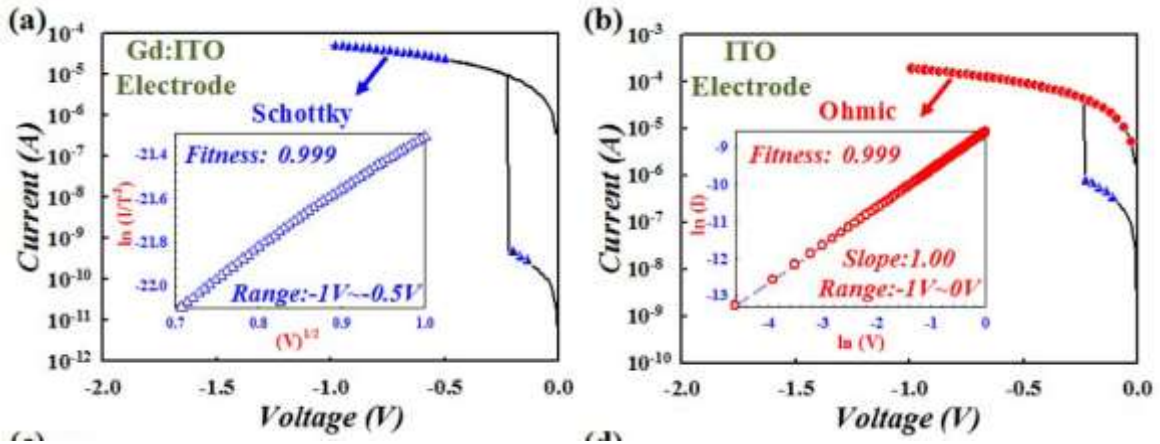


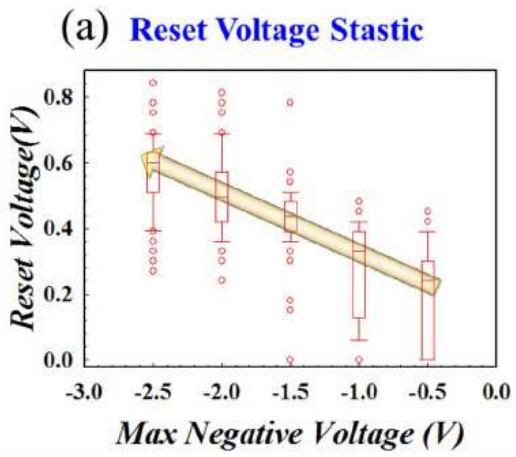
Yuzheng Guo et al – APL 2014

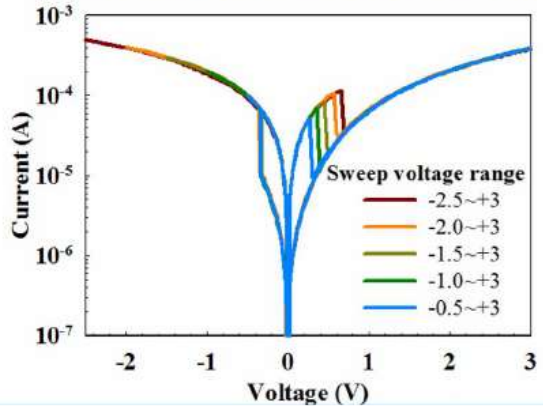
ITO RRAM Considerations



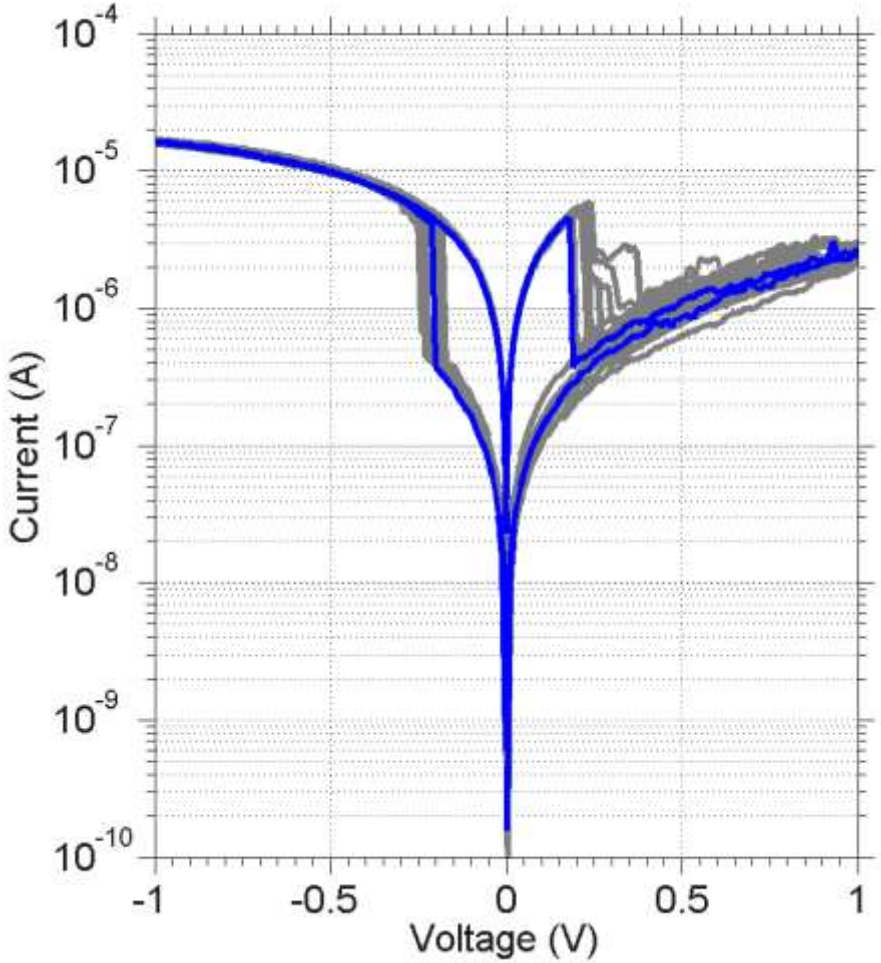


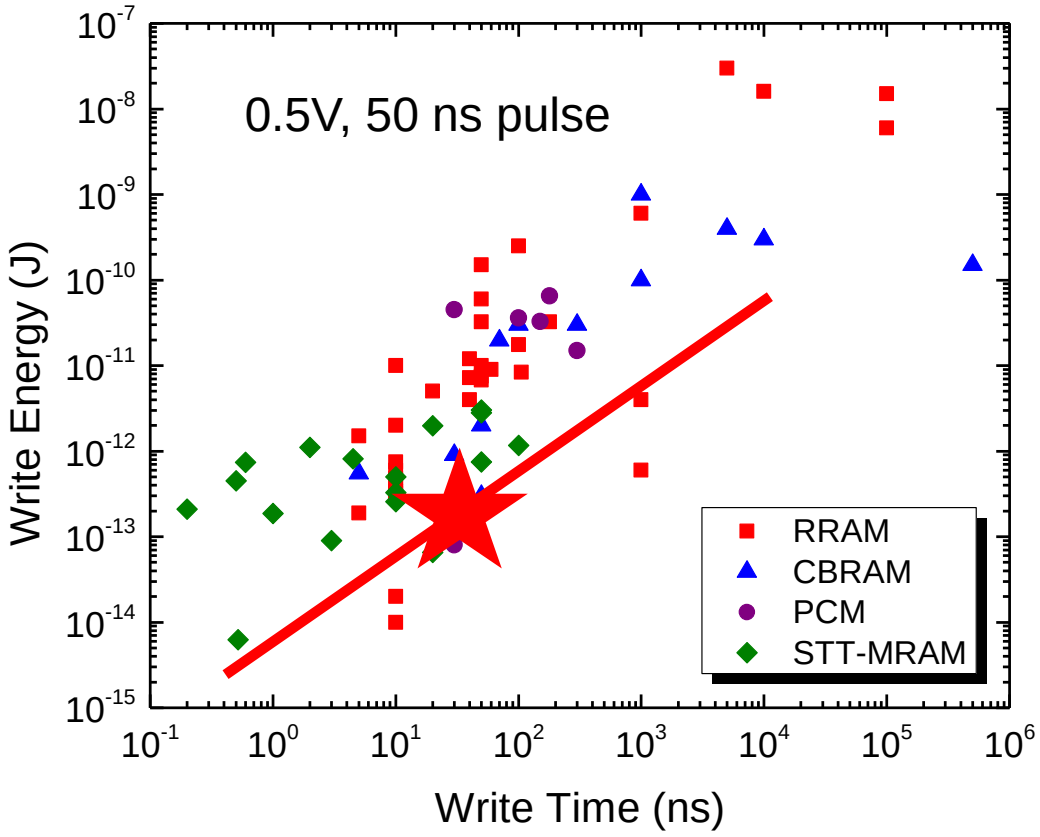






ITO-NWFET-RRAM Memory Technology

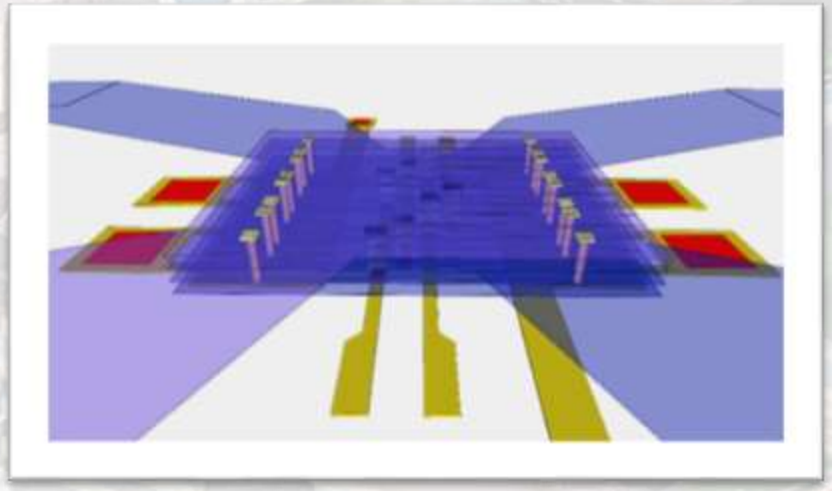


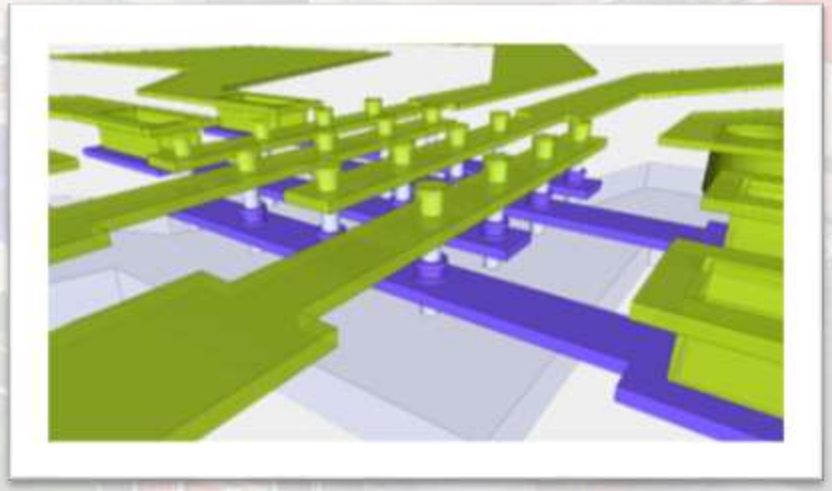


Stanford Memory Trends, H.-S. P. Wong et al
https://nano.stanford.edu/stanford-memory-trends

Future of Computing

- Buzzwords + buzzwords
- 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 neuromorphic computing
- Enormous space for innovation and new startups, specialized hardware to complement the software needed





LUND
UNIVERSITY

NANO
ELECTRONICS
GROUP