SAND2015-8613C

3D Monolithic Resitive RAM

IEEE S3S Conference

Rohnert Park, CA

October 5, 2015

Matthew J. Marinella

Sandia National Laboratories

matthew.marinella@sandia.gov

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.

Outline

- Intro to ReRAM Device Technology
- 3D ReRAM Technology
- 3D ReRAM Challenges
- 3D ReRAM Applications
- Summary and Future Outlook

Outline

- Intro to ReRAM Device Technology
- 3D ReRAM Technology
- 3D ReRAM Challenges
- 3D ReRAM Applications
- Summary and Future Outlook

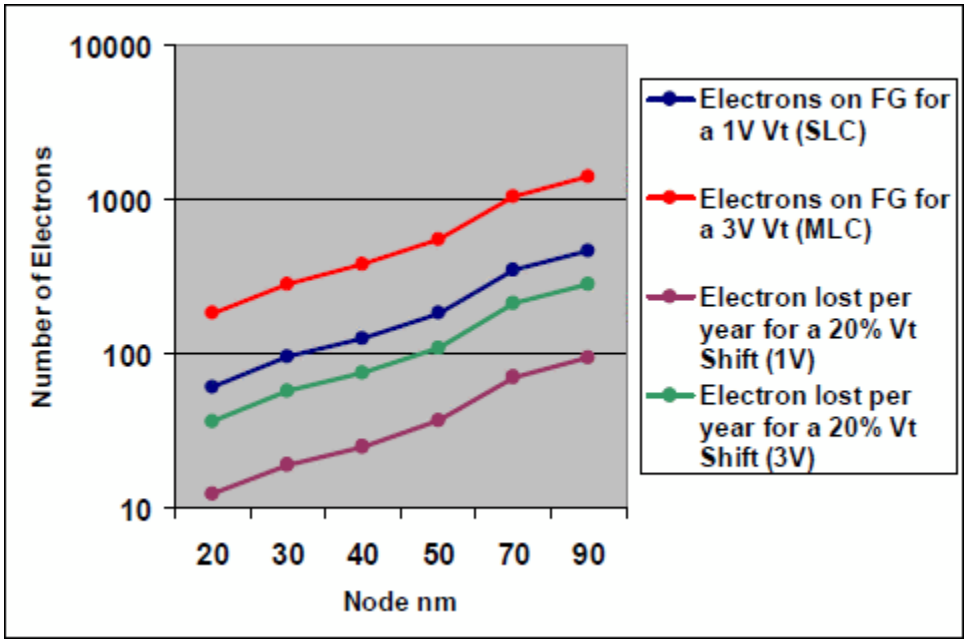
Emerging Memory

samsung.com

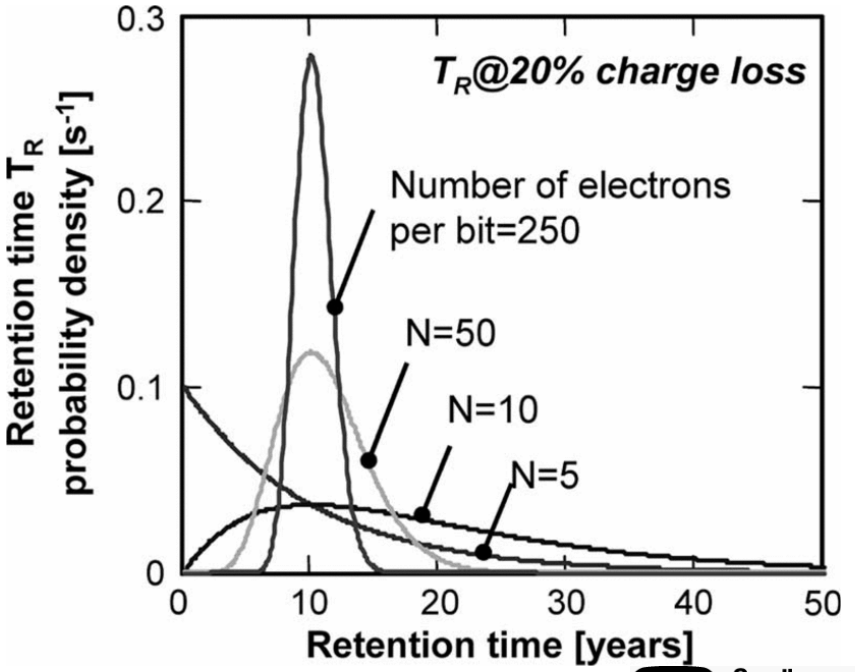
- This is a significant era for memory
- NAND Scaling and Status:
  - Planar announcements ended at 16nm
  - Reliability suffers with scaling; 12 nm is theoretical FG limit
  - Now: several companies have 256 Gbit NAND chips
    - Vertical scaling; 48 layer multi-level cell shipping
- DRAM Scaling and Status:
  - 20 nm GDDR5 shipping; 4 die TSV HBM in graphics card
  - Dielectric challenges for cells <20 nm
- Possible scaling limitations in sight for both
- Storage Class Memory
  - Major benefit to closing the storage→DRAM latency gap
- Near end of transistor scaling: no obvious replacement
- Near end of Flash/DRAM scaling: strong new technology candidates on the horizon!

NAND Flash Scaling

- Can have <100 electrons per gate in 2x and 1x NAND flash
- Significant fraction of electrons lost per year
  - Severe retention degradation
- 3D Vertical NAND has extended scaling for several years

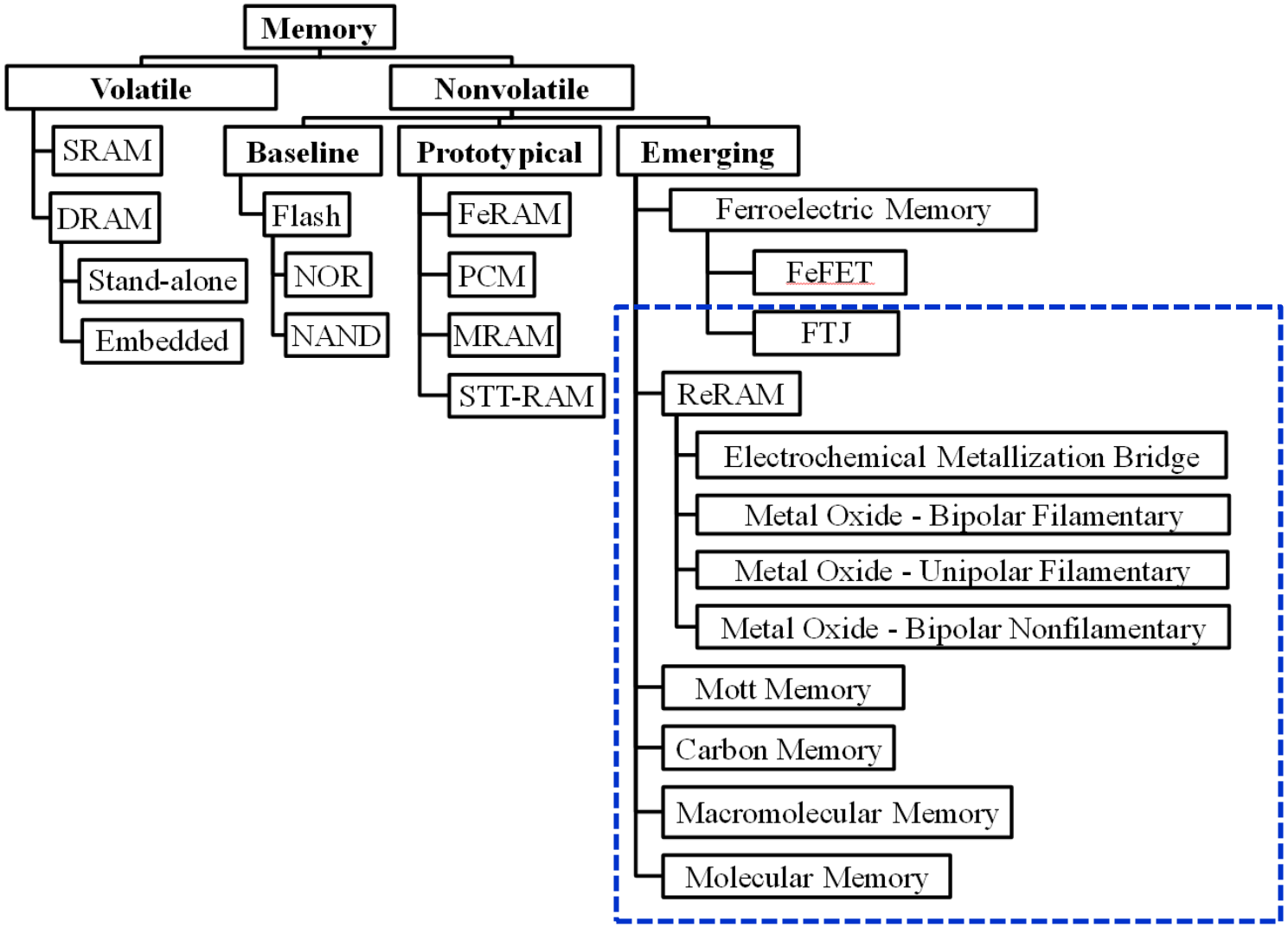


K Prall IMW 2008



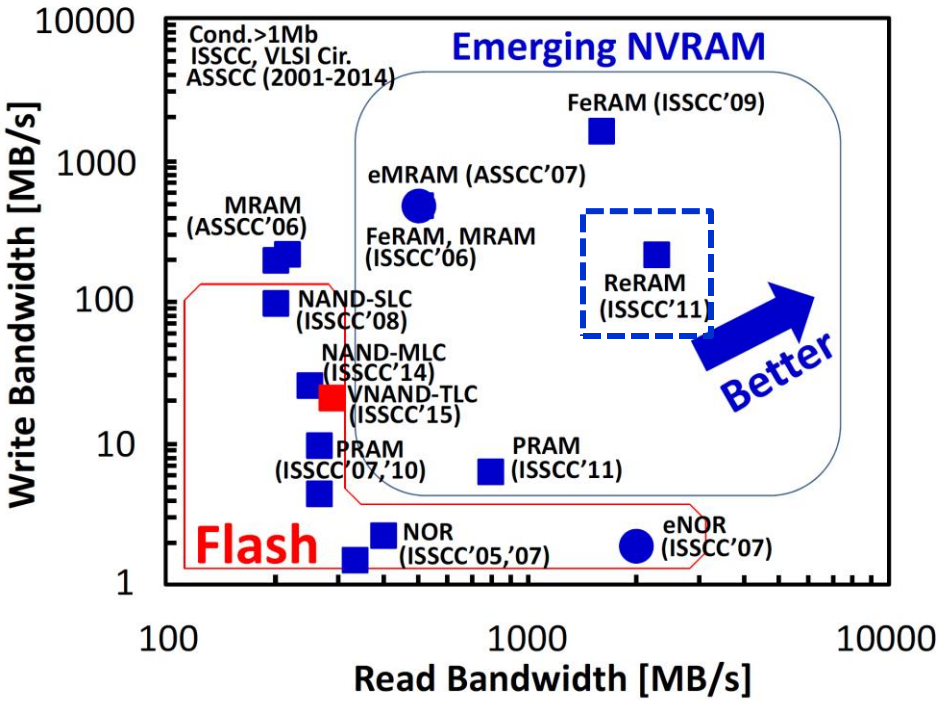
G Molas TED 2006

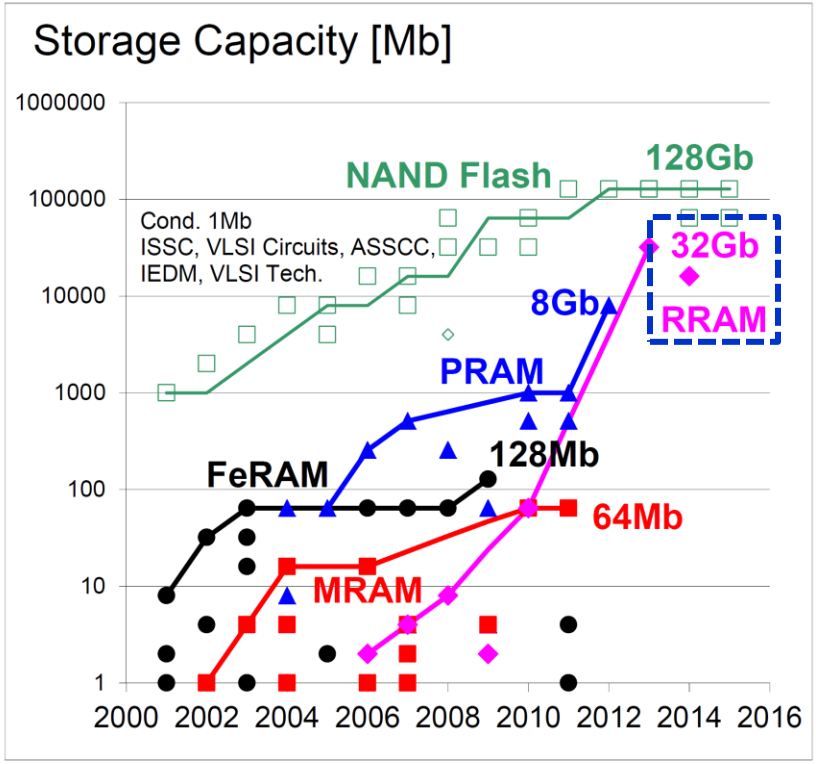
Emerging Memory Taxonomy



Resistive Emerging Memories

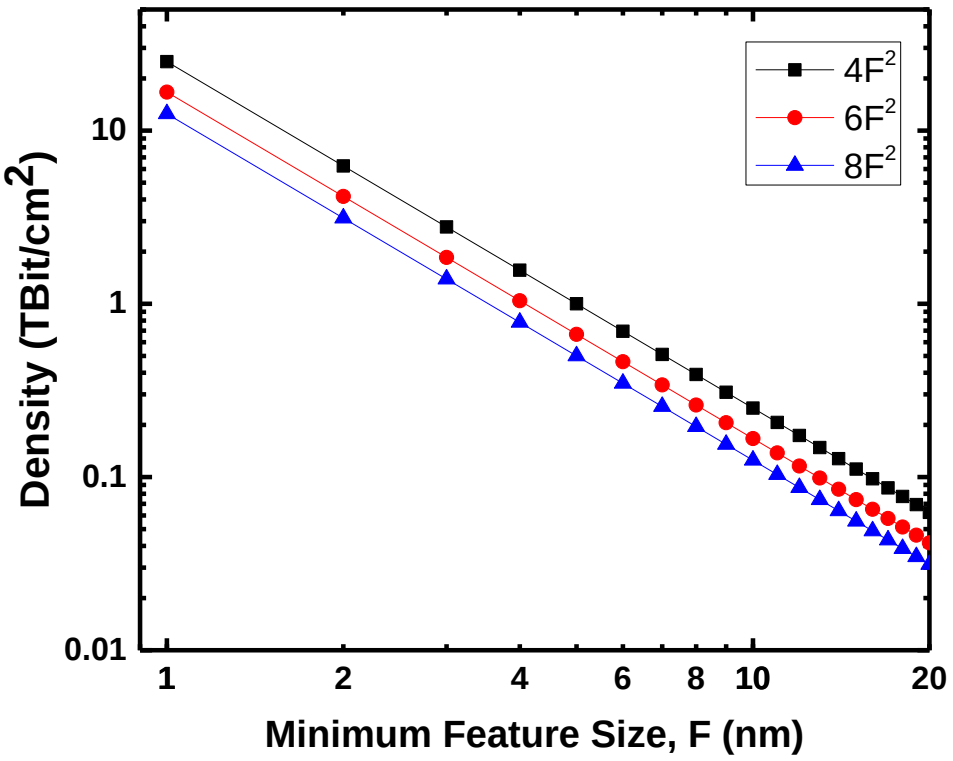
Emerging Memory Trends



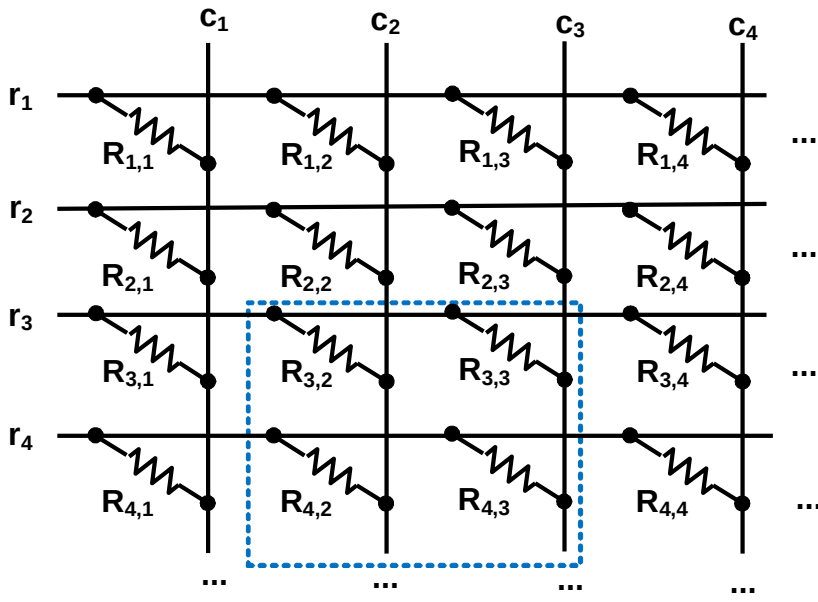


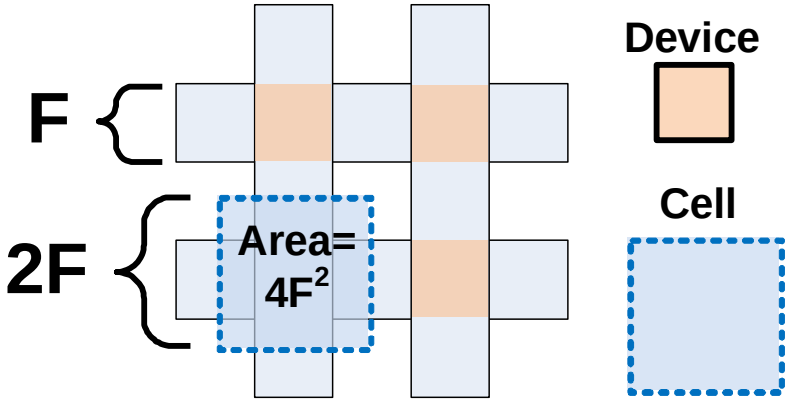
2D Resistive Crossbar Memories

- F=Feature size
- Max areal density possible → $4F^2$



Marinella and Zhrinov, in Emerging Nanoelectronic Technologies, Wiley, 2014.

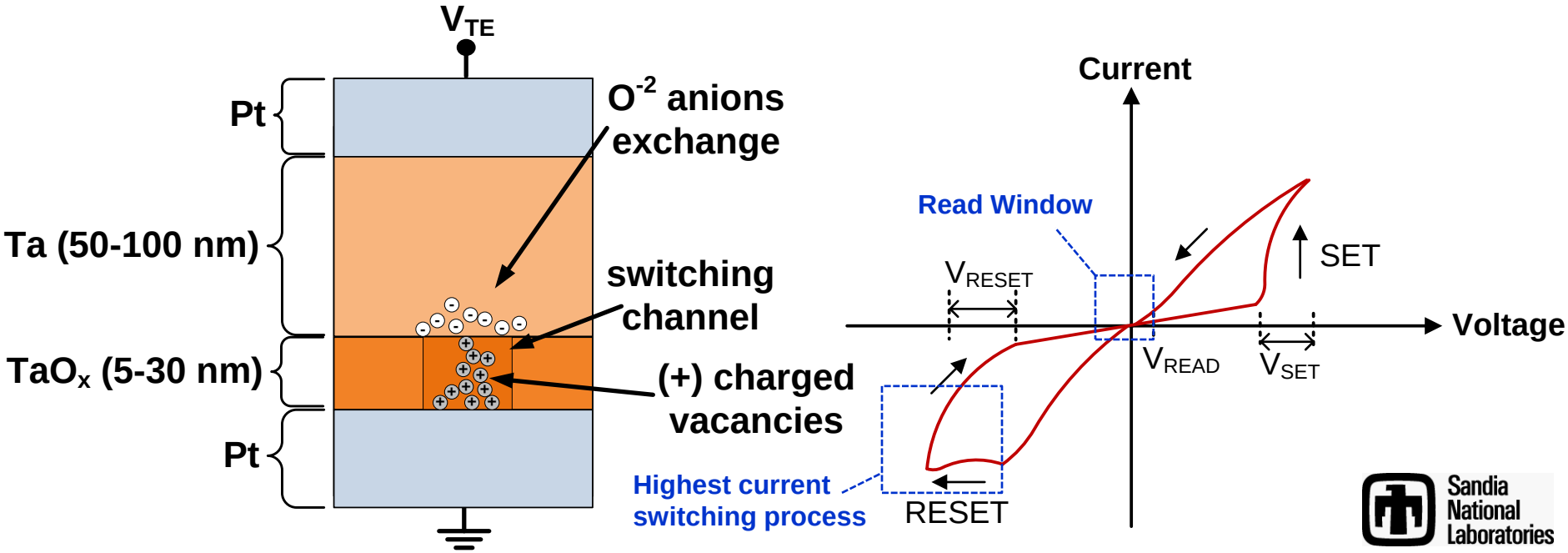




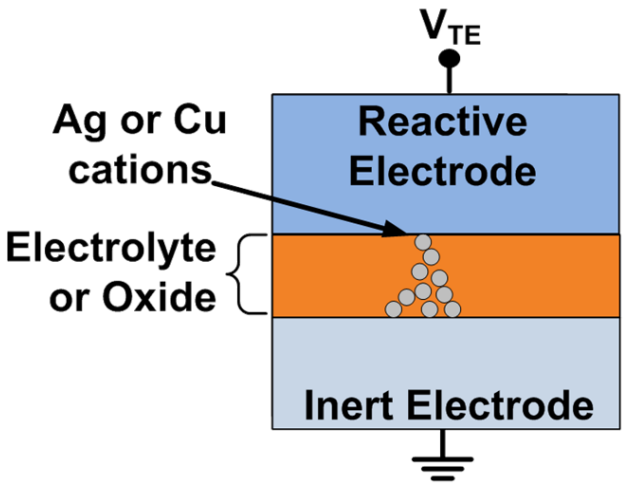
Sandia National Laboratories

Metal Oxide ReRAM Device

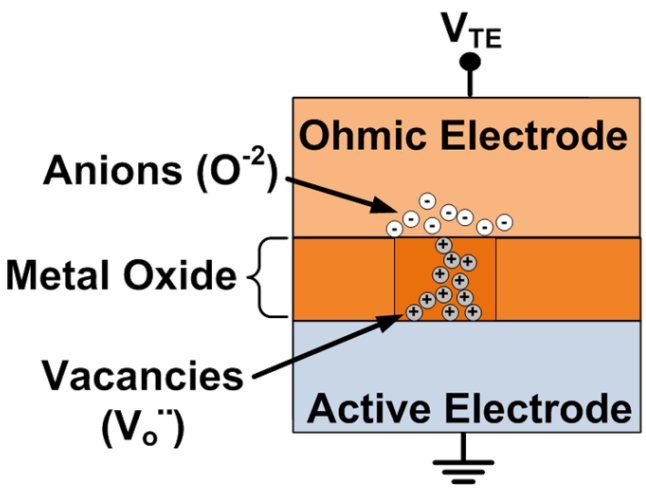
- "Hysteresis loop" is simple method to visualize operation
  - (memory operated through positive and negative pulses)
- Hypothesized oxide resistance switching mechanism
  - Positive V_{TE}: low R – O^{-2} anions leave oxide
  - Negative V_{TE}: high R – O^{-2} anions return
- Common switching materials: TaO_x, HfO_x, Al_2O_3, TiO_2…
- Despite progress, fine details of switching mechanism still debated



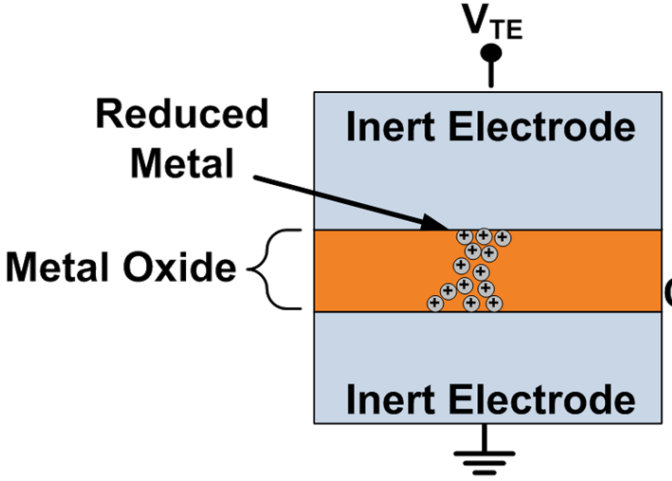
Electrochemical Metallization Bridge



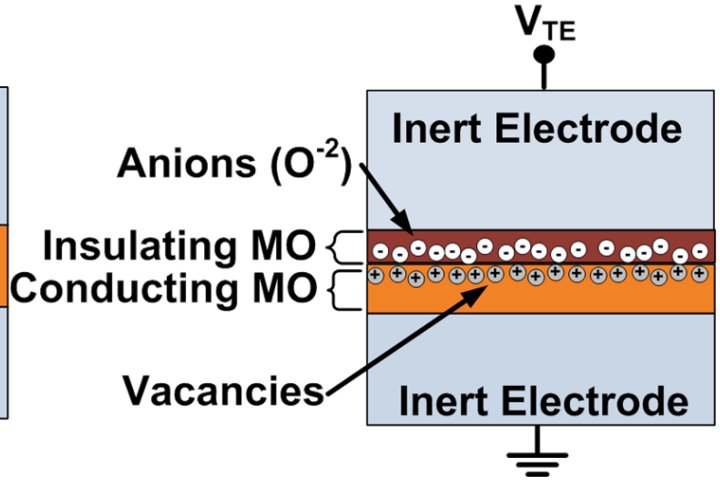
Metal Oxide: Bipolar Filamentary



Metal Oxide: Unipolar Filamentary



Metal Oxide: Bipolar Non-Filamentary

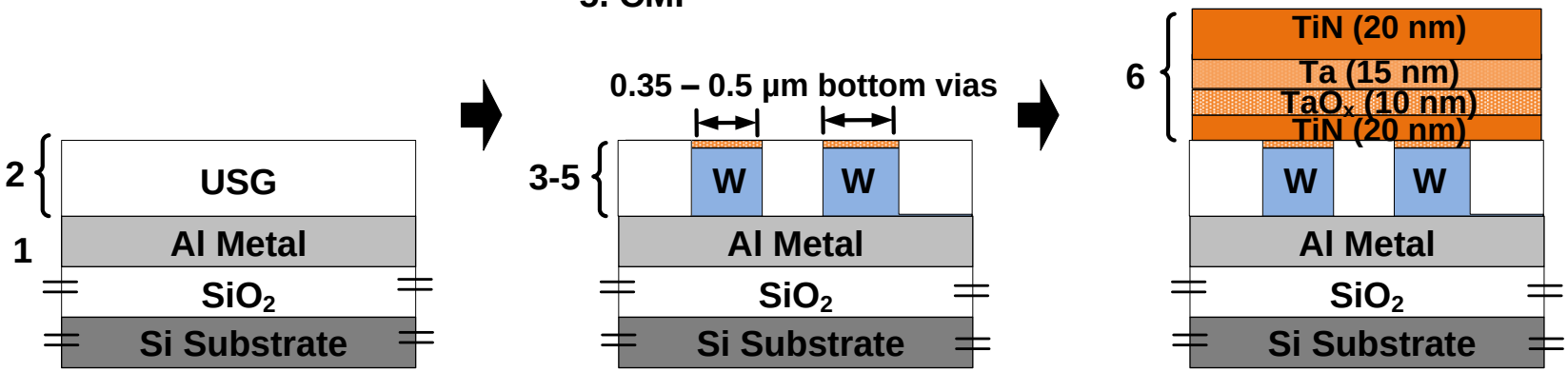


Sandia 2D ReRAM Process Flow

1. Deposit Bottom Metal (Al)
2. Deposit USG

3. Etch via holes in USG
4. Deposit W and TiN layers
5. CMP

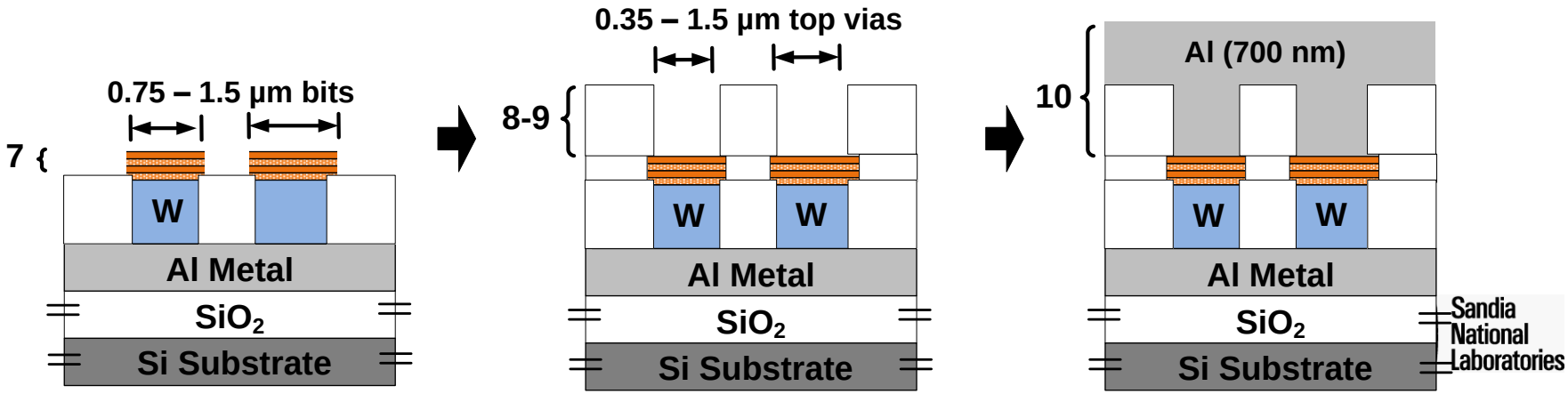
6. Deposit bit stack
(layers enlarged for clarity)



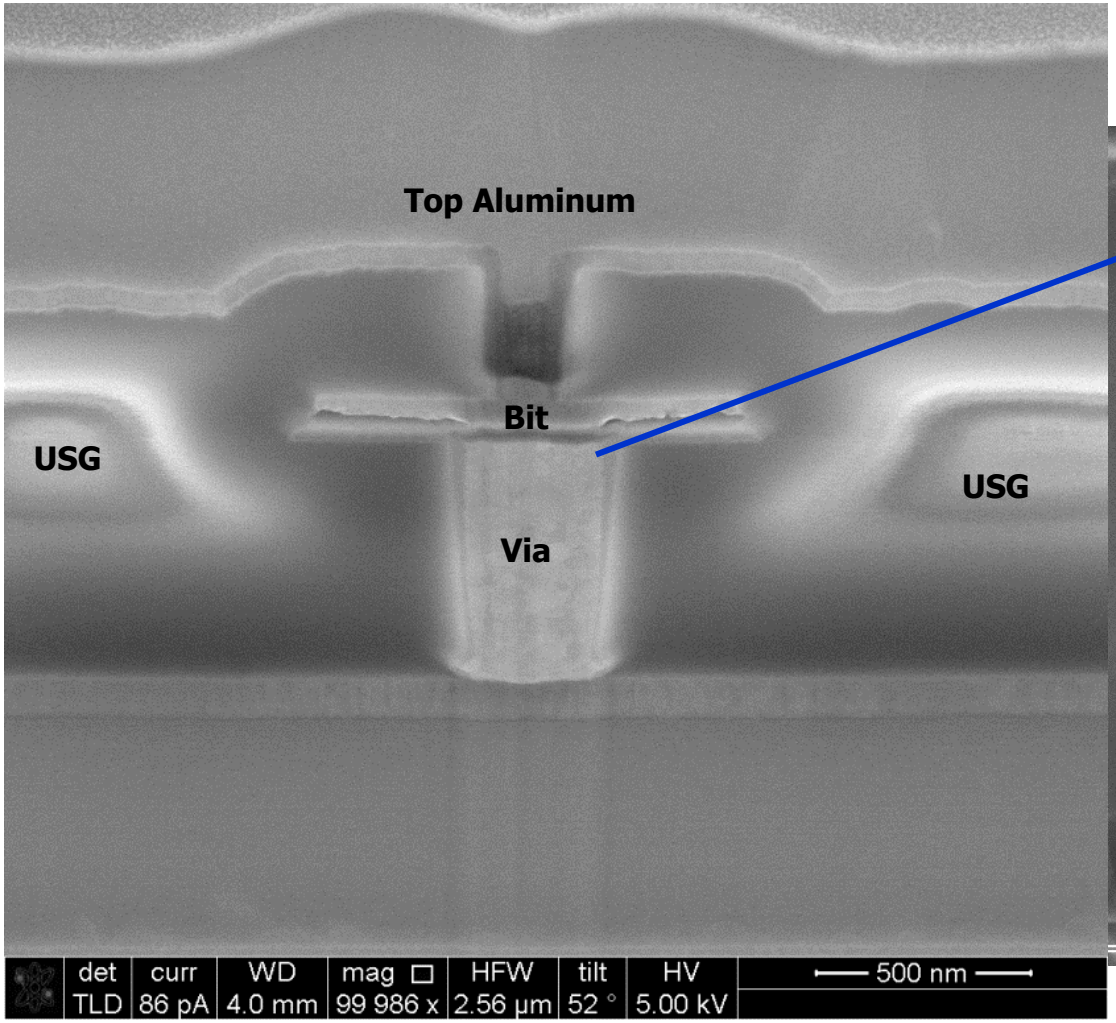
7. Etch bits

8. Deposit top USG
9. Etch top via holes in USG

10. Deposit top Al

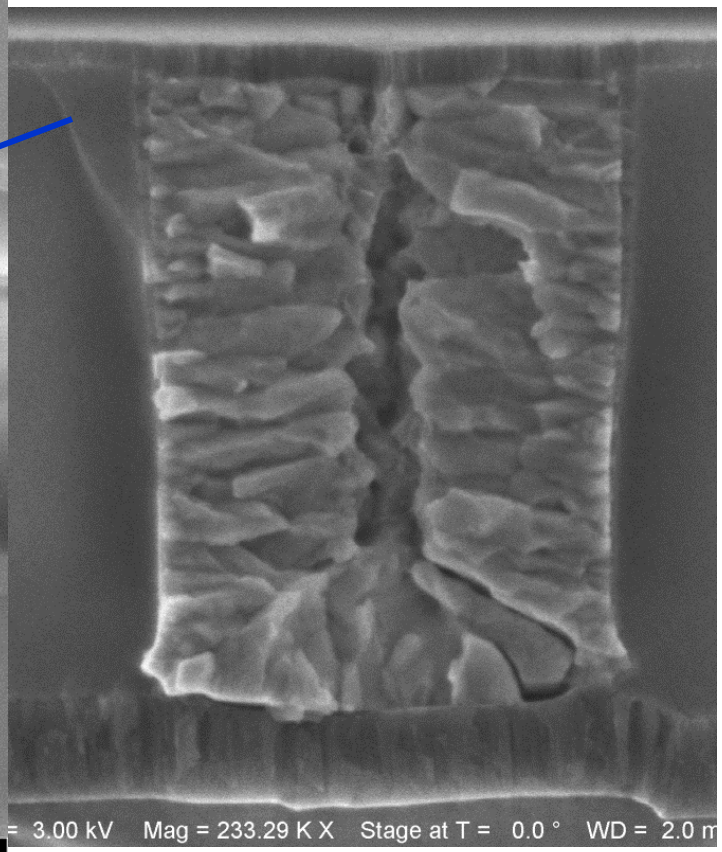


Final Structure



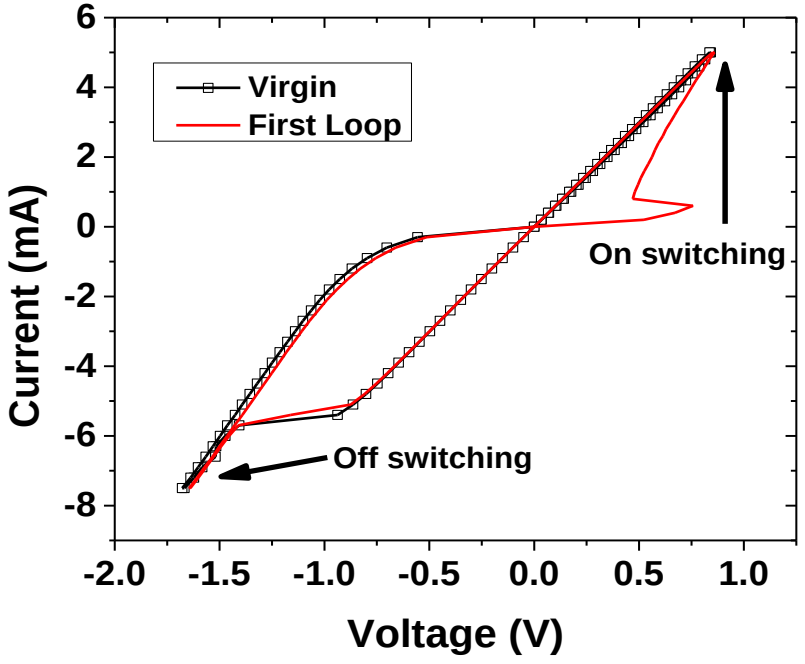
Important to have extremely flat surface under bit

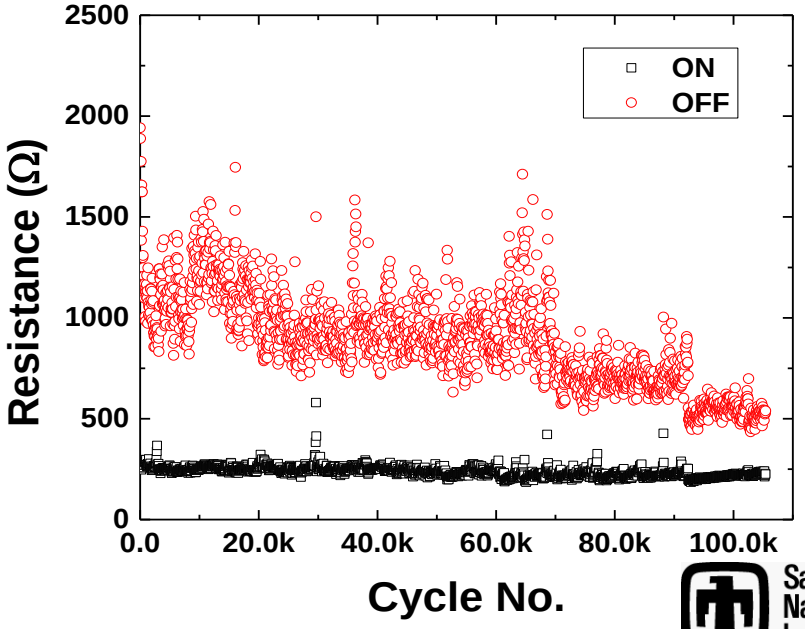
Polished TiN Surface



Basic Device Characterization

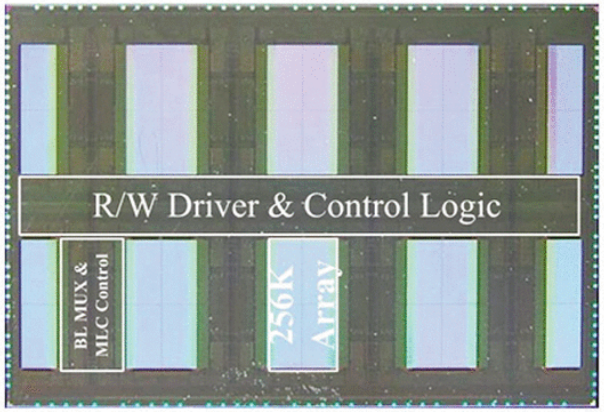
- Typical devices form at very low currents
- Appear “forming free” in current sweep mode
- Do not need > 1.8V; do not need a high voltage transistor
  - Drawback of floating gate NVM
- Resistance can be tailored by stoichiometry





Oxide ReRAM Macro Examples

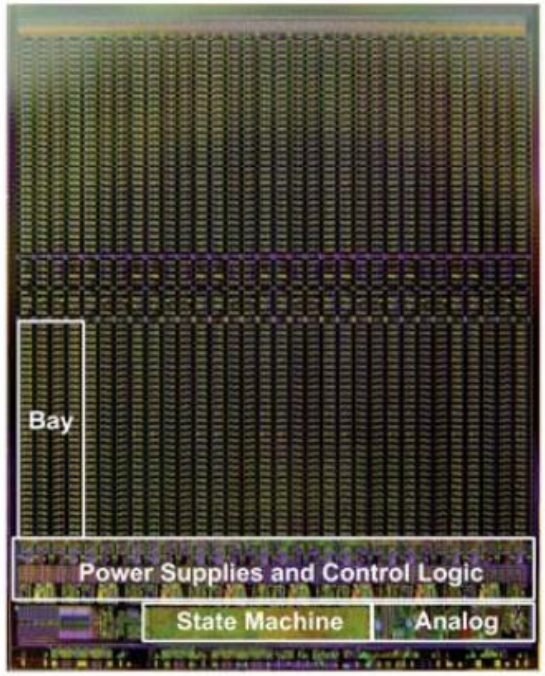
ITRI High Speed Macro



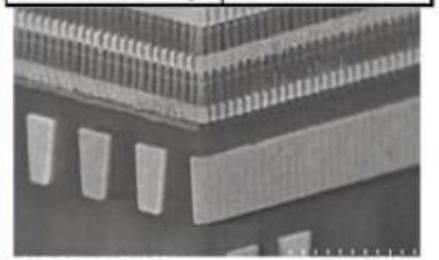
Process	CMOS: 0.18um 1P4M RRAM: 0.64um x0.48um
Memory Capacity	4Mb (32 x 128Kb sub-blocks)
Chip size	11310um x 16595um (with test-mode circuits)
Device	HV path: 3.3V device Cell array: 3.3V device Peripheral: 1.8V device
VDD	HV path: 3.3V Core: 1.8V
Read-Write Access Time (SLC-mode)	Random access: 7.2ns Burst-mode: 3.6ns

H.D. Lee et al (ITRI), VLSI 2012

Sandisk/Toshiba 32Gb, 230us write ReRAM



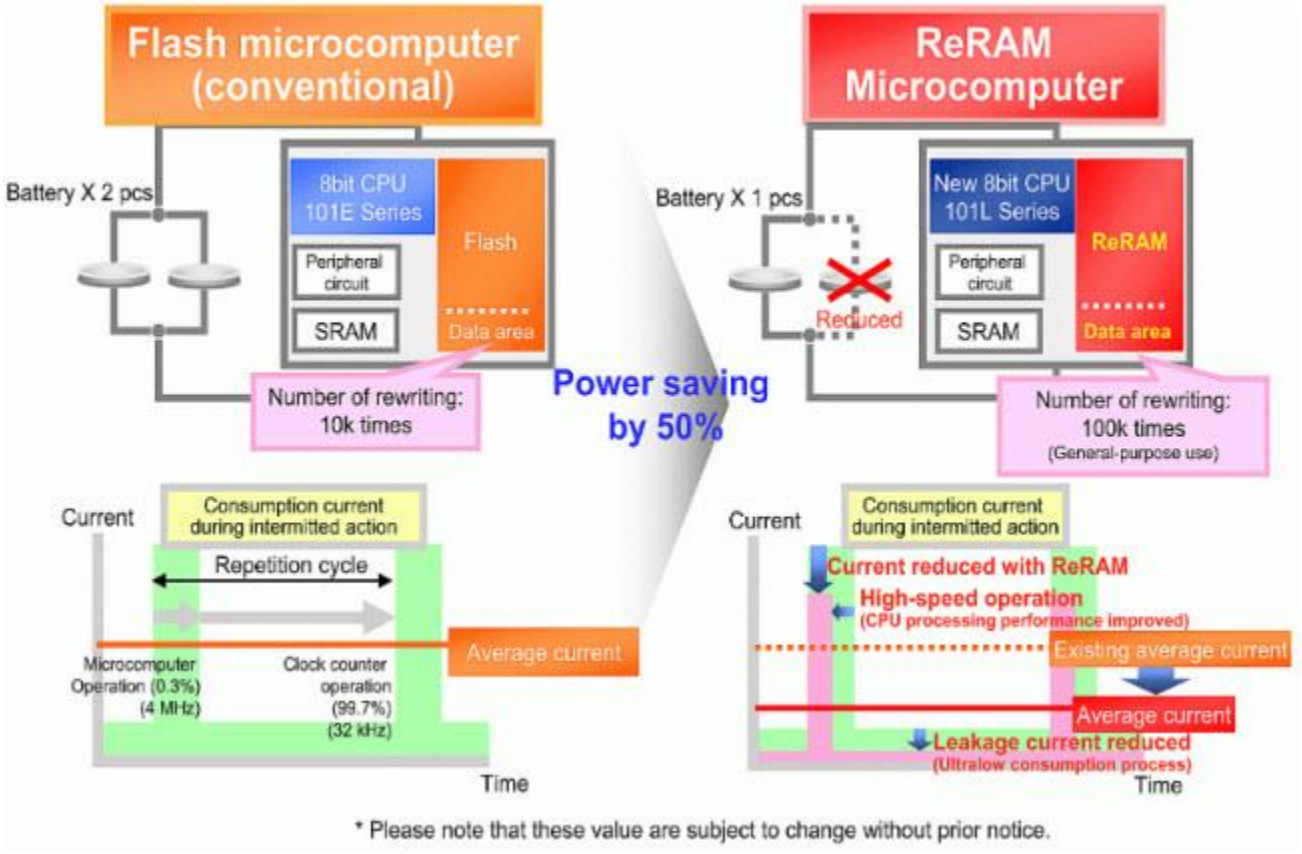
Density	32Gb
Cell Size	24nm x 24nm
Die Size	130.7mm^2
Interface	NAND-Compatible
Page Size	2KB
Read Latency	40us
Write Latency	230us

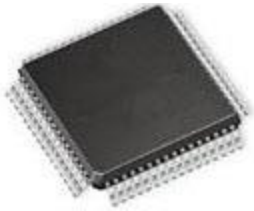


T-Y Liu et al (Sandisk, Toshiba), ISSCC 2013

First Commercial Oxide ReRAM Product

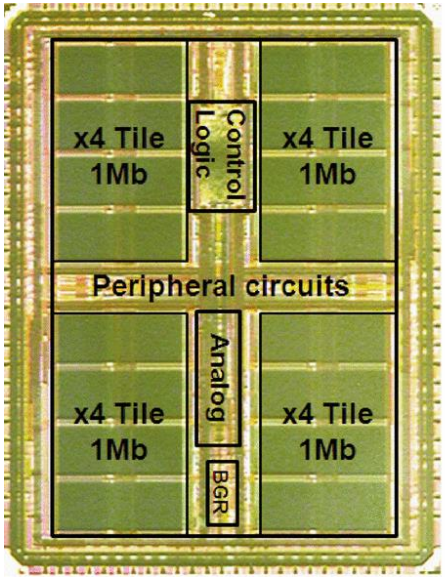
- **Panasonic MN101L ReRAM MCU**
- **Power and time saving over flash MCU**





Significant CBRAM Macro Demos

Sony 4MB High R/W CBRAM Macro

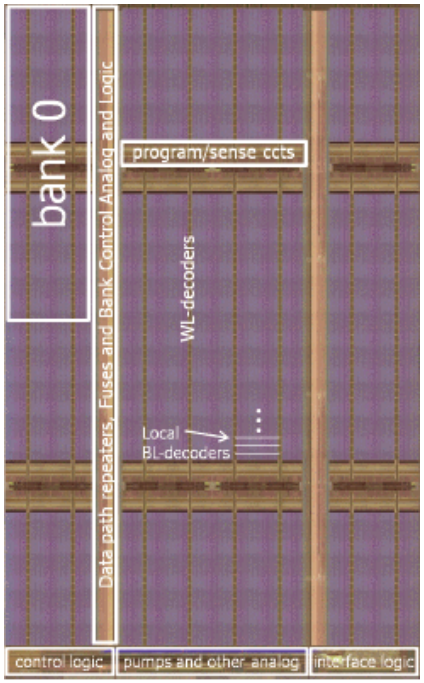


Capacity	4 Mb
Tile	256Kb
Process	180nm CMOS
Chip size	6.8x5.26mm 35.8 mm²
Cell architecture	1T-1R
Cell size	2.24 μm²
Power Supply	3.3 V,1.8 V
Memory / IF clock	125MHz
Read Size, Throughput	128Byte 2.3GB/s
Program Size, Throughput	16Byte 216MB/s

Record Throughput

W Otuka et al (Sony), ISSCC 2011

Sony/Micron 16 Gb CBRAM Macro

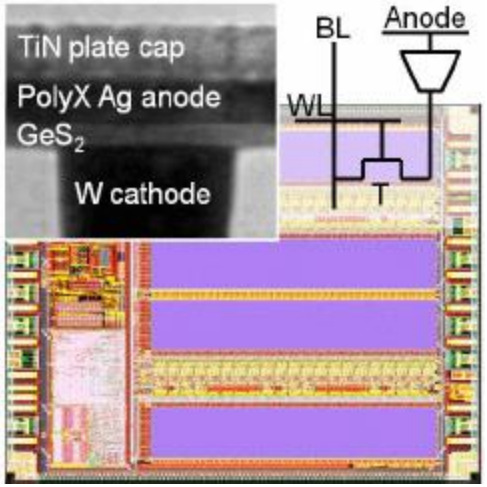


Summary Table		
Density		16 Gb
Tech node (nm)		27
Cell Size (nm²)		4374 (6F²)
Die Size (mm²)		168
Selector		Buried WL MOS selector
Read Performance	BW (MB/s)	1000
	Latency (uS)	2
Write Performance	BW (MB/s)	200
	Latency (uS)	10

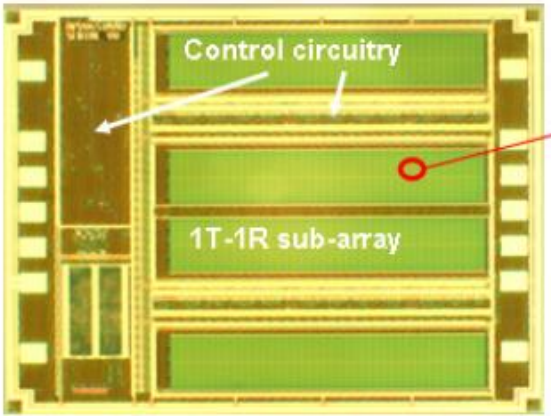
R Fackenthalet al (Micron, Sony), ISSCC 2014

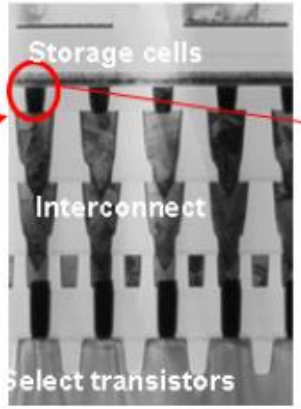
First Commercial CBRAM Product

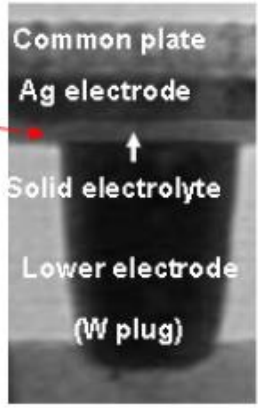
- Available 2012 by Adesto Tech
- Ag/GeS2/W cell (Gen 1)
- Integrated with BEOL of Altis CMOS
- I^2C and SPI-based serial memories, up to 512kbit
- 25µs byte write
  - >100x improvement over similar EEPROM



adesto™ TECHNOLOGIES







Sandia National Laboratories

Valov et al, Nanotechnology 2011, Jameson et al IEDM 2013

Outline

- Intro to ReRAM Device Technology
- 3D ReRAM Technology
- 3D ReRAM Challenges
- 3D ReRAM Applications
- Summary and Future Outlook

3D Schemes

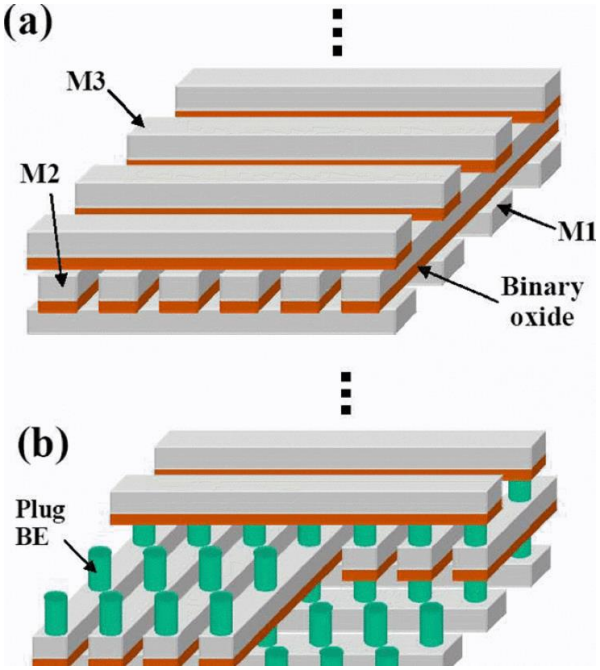
Two Main Schemes:

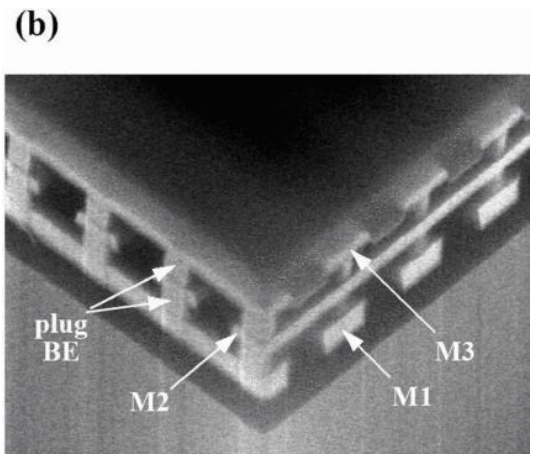
1. **3D Crosspoint**
   - Straightforward: layer the 2D crosspoint
   - Repeats process sequence N times for N layers
   - Easier to fabricate reliable cell and select device
   - Best performance
   - High cost for large stacks

2. **Vertical RRAM (VRRAM)**
   - Memory layers deposited first without lithographic definition
   - High aspect ratio vertical holes etched in a layered structure
   - *Number of critical masks independent of layers*

Layered Crosspoint Arrays

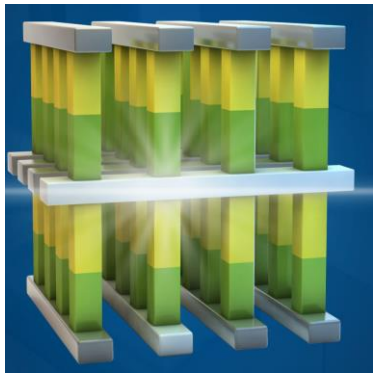
- Simplest and earliest method of 3D monolithic ReRAM
- 2 layer stack ReRAM demonstrated IEDM 2005 by Samsung
  - Many prototypes demonstrated since
- Similar architecture of recently publicized: Intel/Micron "3D Xpoint" memory





IG Baek (Samsung) IEDM 2004

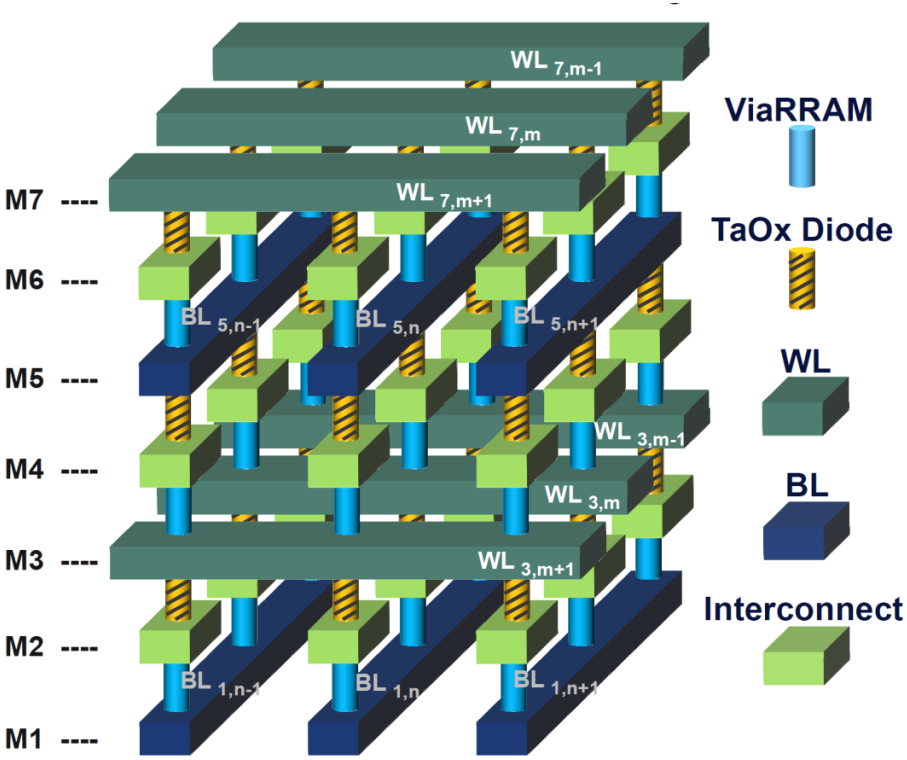
Intel/Micron 3D Xpoint
(Press Release Picture)

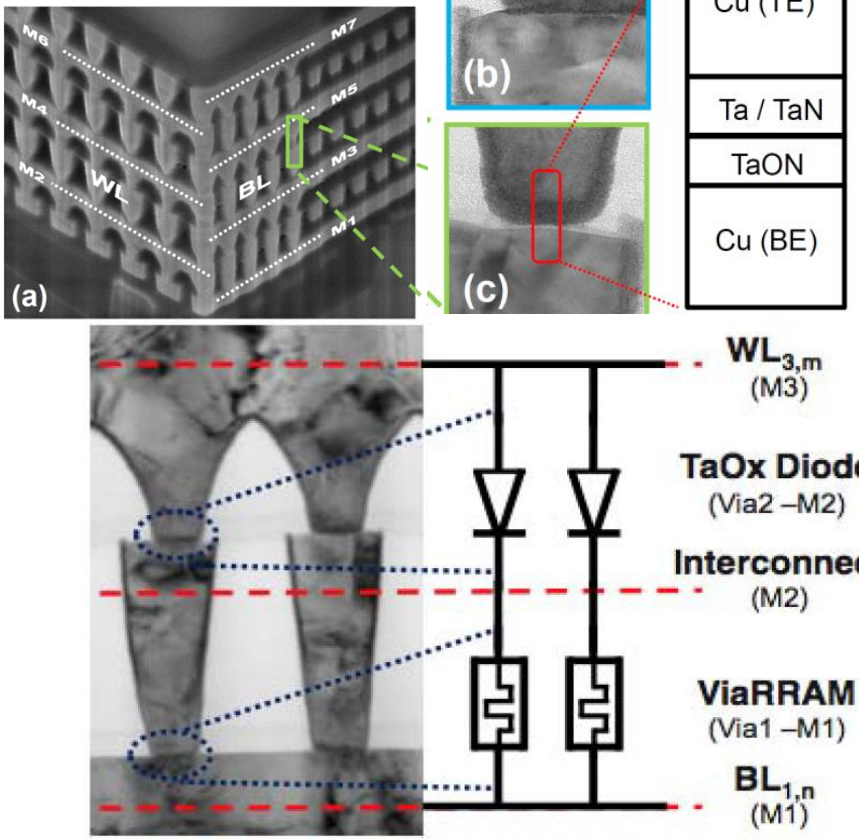


intel.com

Layered Crosspoint Array: ViaRRAM

- 30 nm x 30 nm "ViaRRAM" cell in TSMC 28nm process
  - True $4F^2$ unit cell, with F = design rule
- Unipolar ReRAM cell (RESET current >100μA)

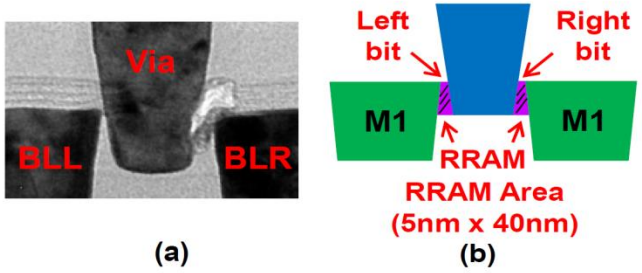


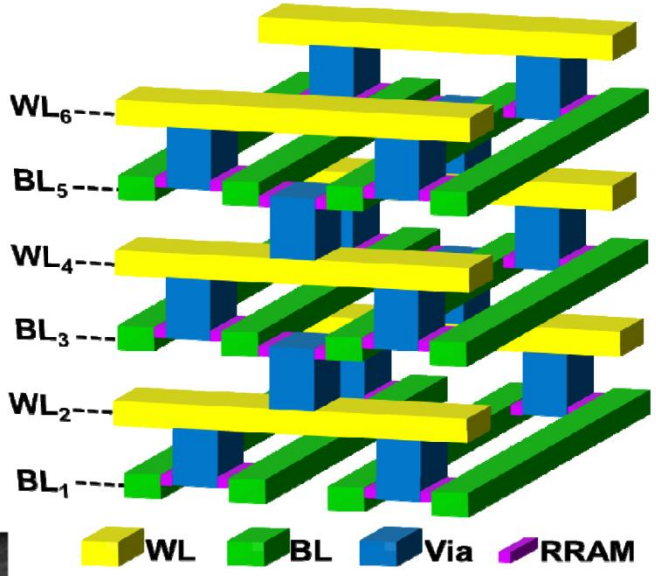


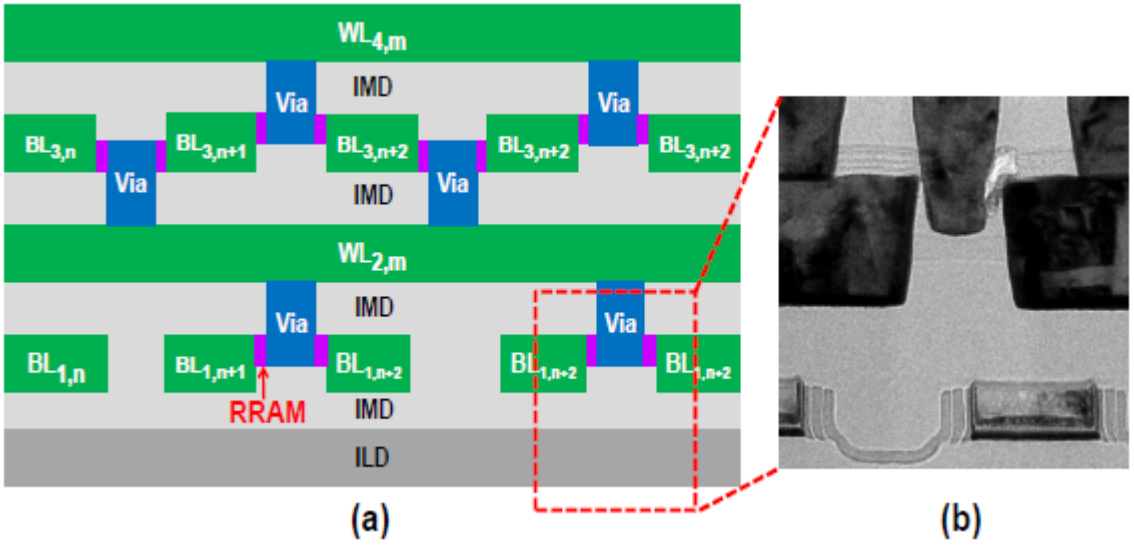
Hsieh et al (NTHU, TSMC, ITRI) IEDM 2013

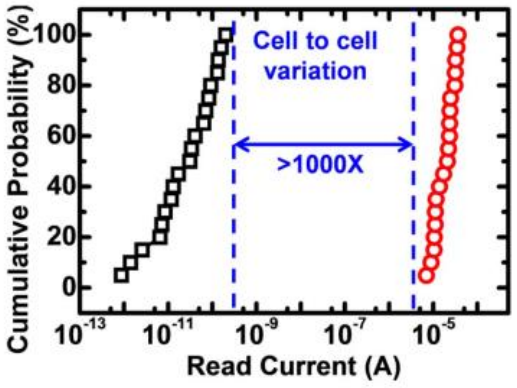
Layered Crosspoint Array: "Interweaved" Crosspoint Array

- Scheme enables 2 bits per cell with via between to wires





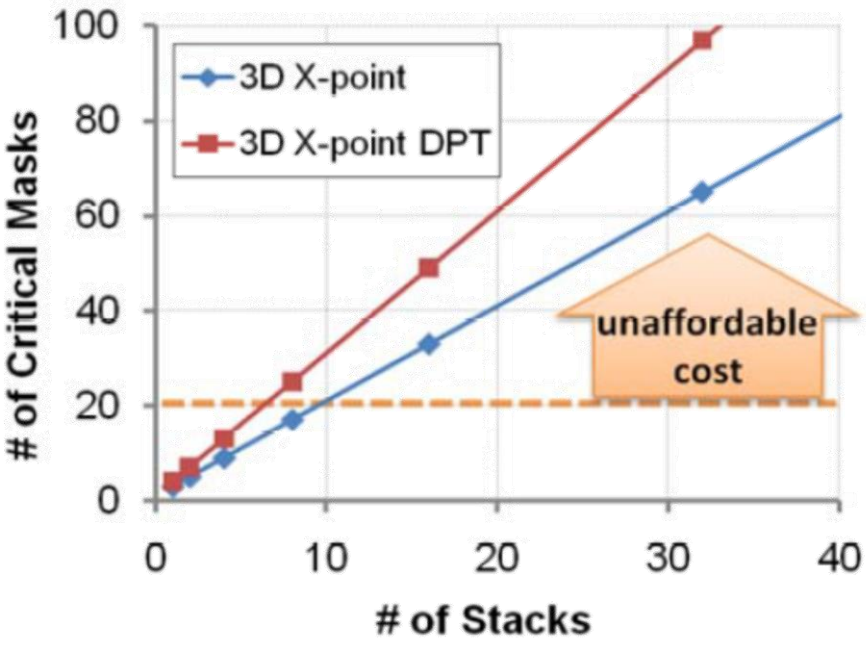




Chin et al (NTHU, TSMC) IEDM 2014

Problem with Stacking Horizontal Crosspoint Arrays

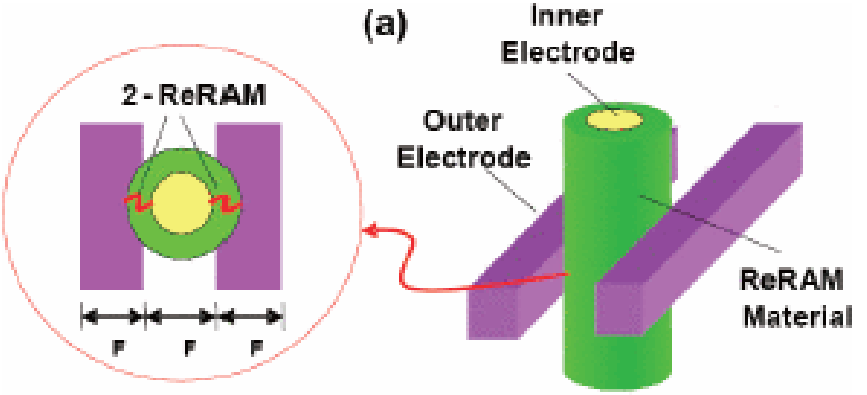
- Horizontal crosspoint stacking not cost effective past ~8 layers
  - Double patterning for HP less than ~40 nm
- No scaling path: doubling number of layers may double critical masks (and cost!)
- Need another method to continue layering…

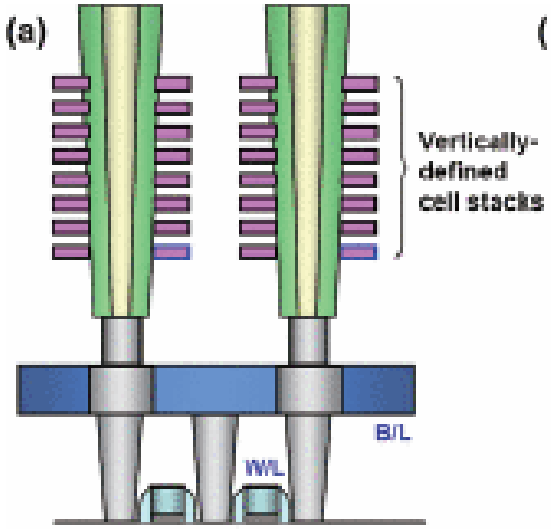


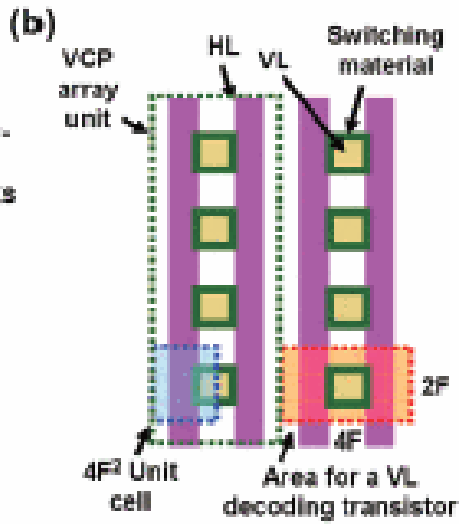
IG Baek et al (Samsung), IEDM 2011

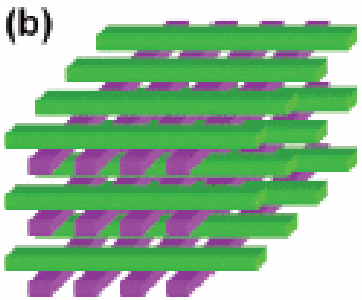
Samsung Initial Vertical Architecture

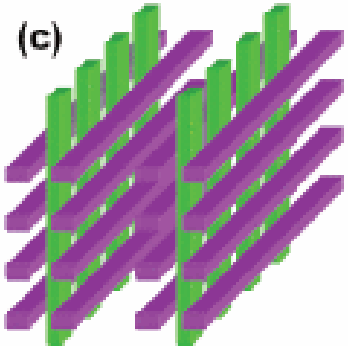
- 2009 VLSI: Samsung presents first vertical RRAM demonstration
- Ru/NiO/(WOx)/W cell
- Reset current > 1 mA
  - Due to unipolar cell







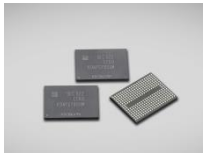




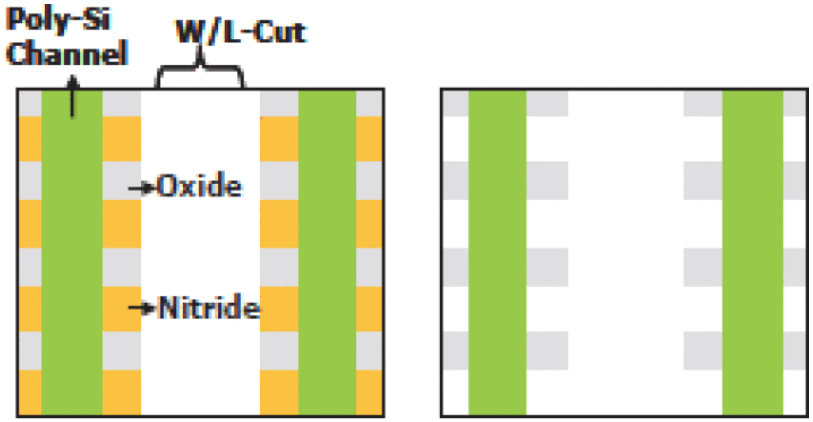
Yoon et al, VLSI Tech 2009

Samsung Vertical-NAND Flash

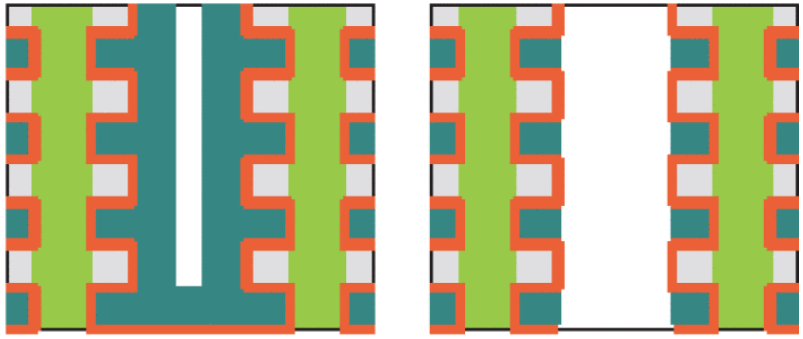
256 Gbit chip!
48 layer, MLC



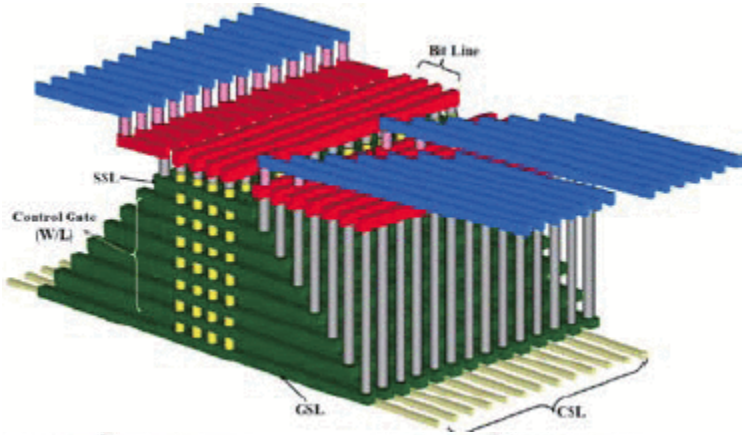
- **"Terabit Cell Array Transistor"**



(a) After 'W/L cut' dry etch (b) Wet removal of nitride

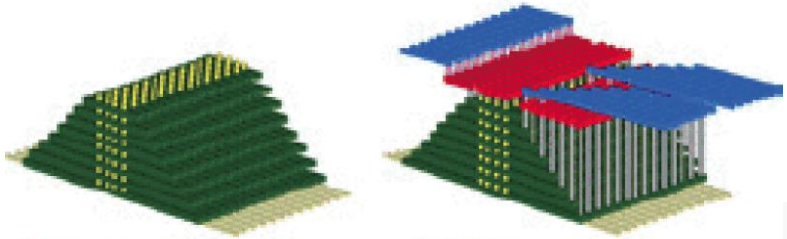


(c) Deposition of gate dielectric and tungsten (d) Gate node separation



(a) Oxide/Nitride Multi- Layer Deposition (b) Channel Hole

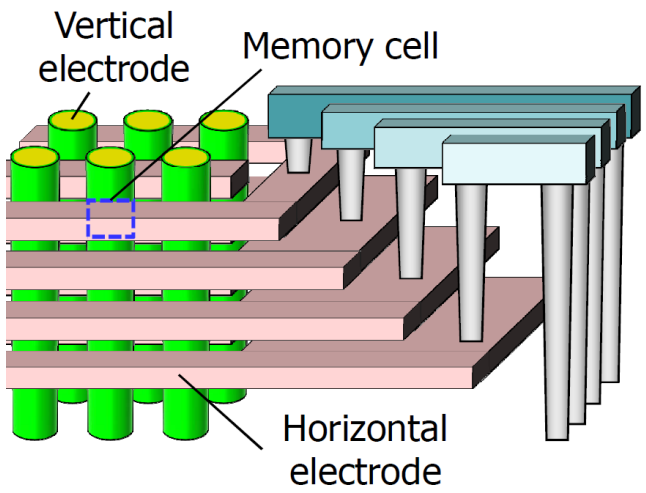
(c) Gate Pad (d) W/L Cut Etch

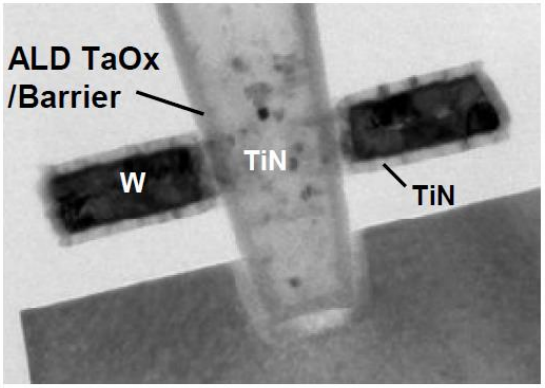


(e) After Gate Replacement Process & CSL Implant (f) BEOL

Jang et al, VLSI Tech 2009

Samsung Vertical ReRAM Process





(a) Mold (SiO_2/Si_3N_4) Depo.

(b) Hole D/E and TMO/VE Depo.

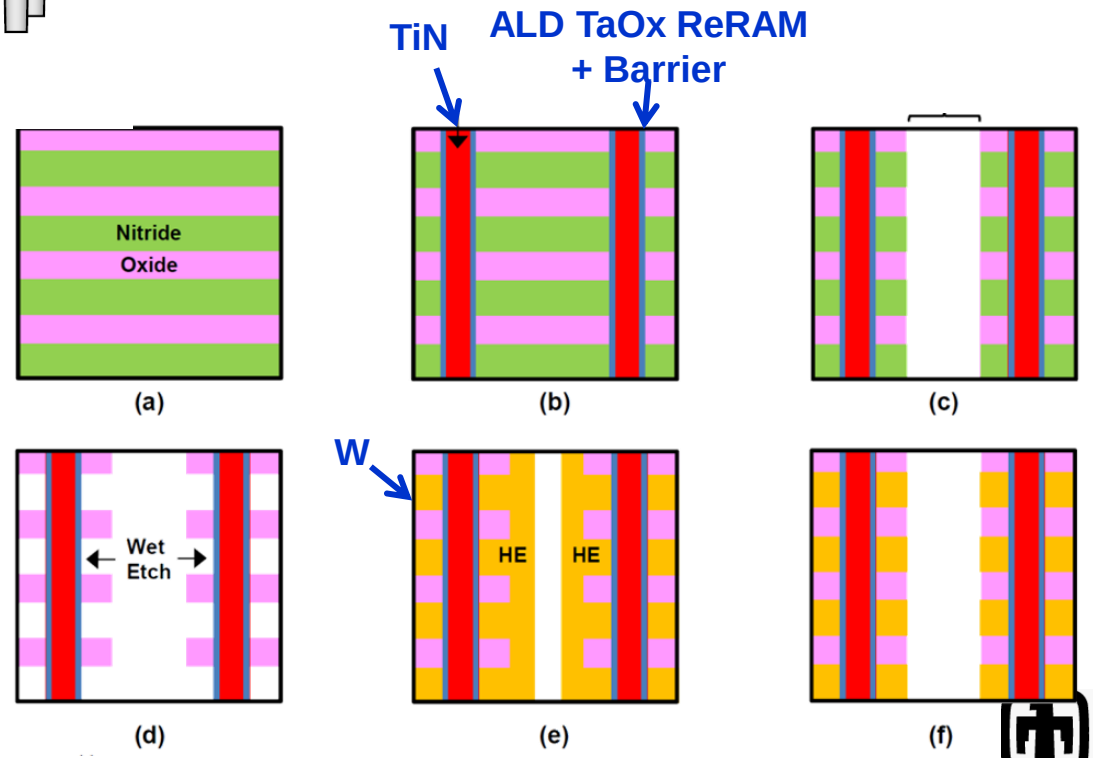
(c) HE Cut Dry Etch

(d) Wet chemical Etch of Nitride

(e) HE Depo.

(f) Metal Node Separation

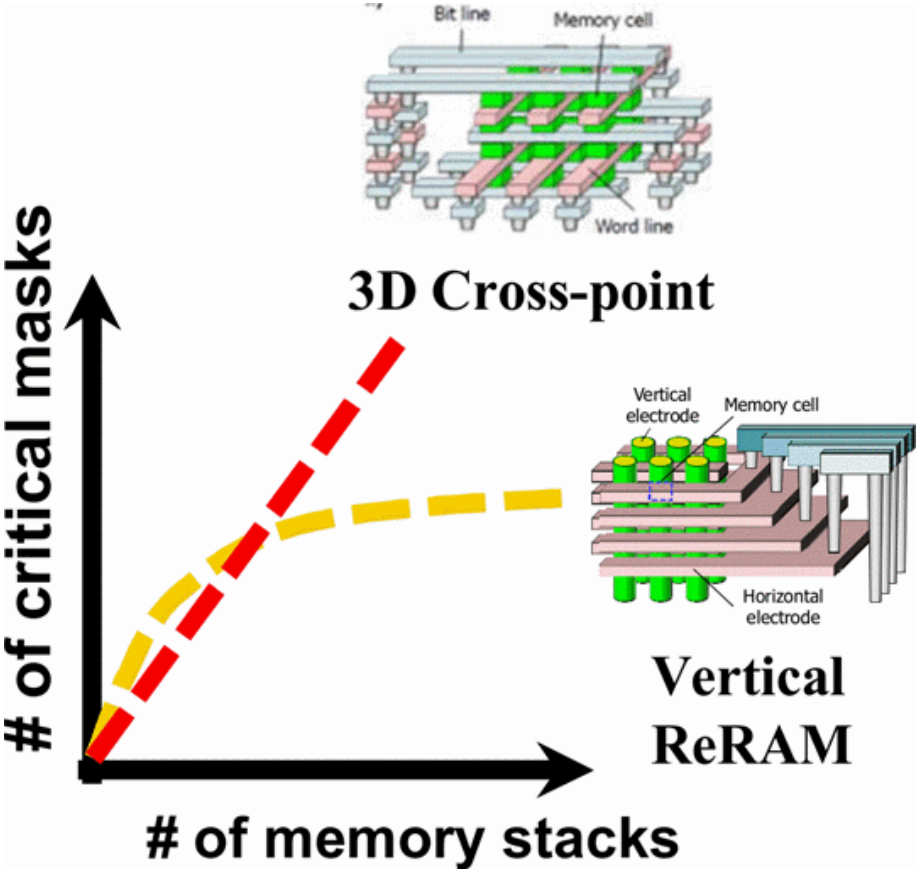
Does this process look Familiar?



IG Baek et al (Samsung), IEDM 2011

VRRAM Enables >8 Layer Stack

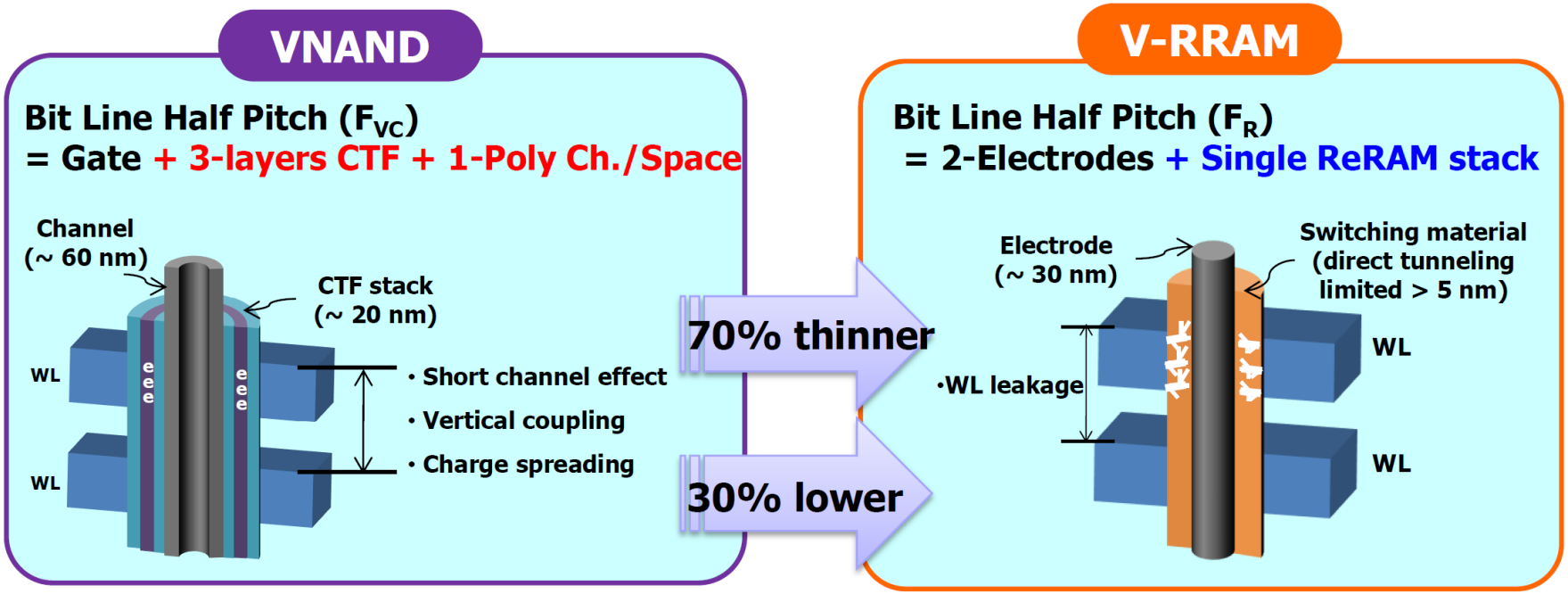
- Critical masks now do not scale with number of layers
- Without VRRAM, >~8 layers is cost prohibitive



SG Park et al (Samsung), IEDM 2012

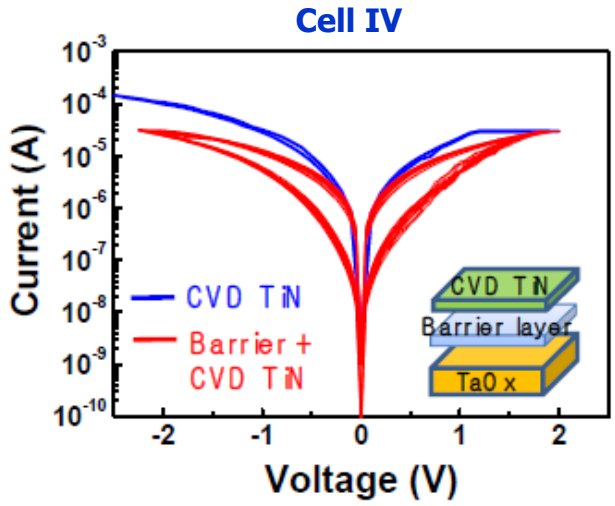
VNAND versus VRRAM

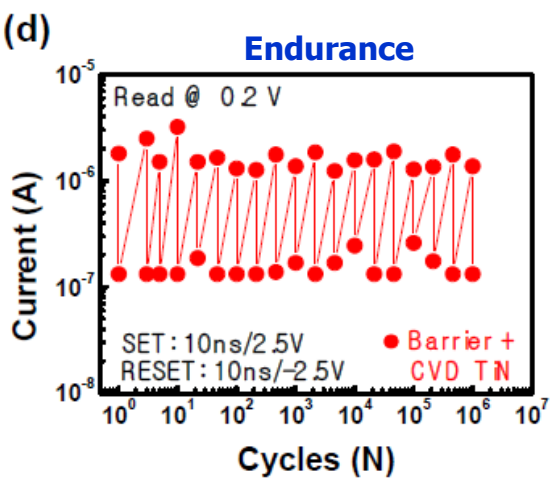
- Compare Samsung's vertical NAND and ReRAM:
  - ReRAM has a scaling advantage
  - ReRAM is typically faster, lower write energy
  - At <20nm, ReRAM typically has better endurance and retention than equivalent scale NAND Flash cell

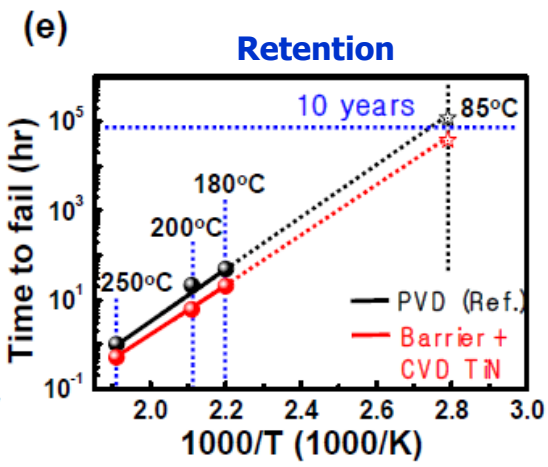


Samsung VRRAM Device Properties

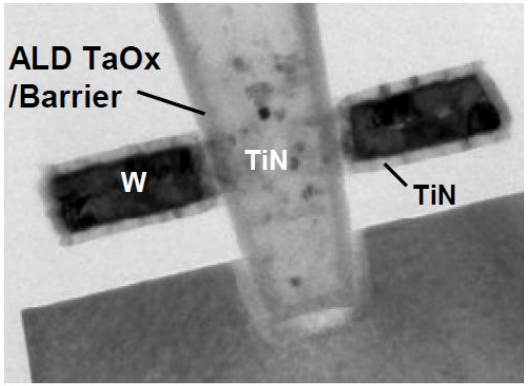
"Dash" Control Devices

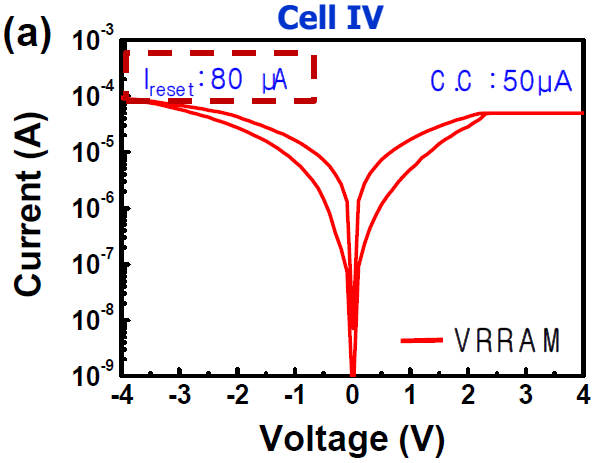


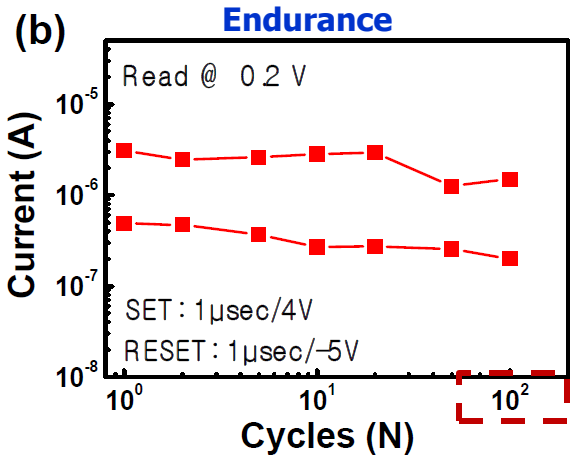




VRRAM

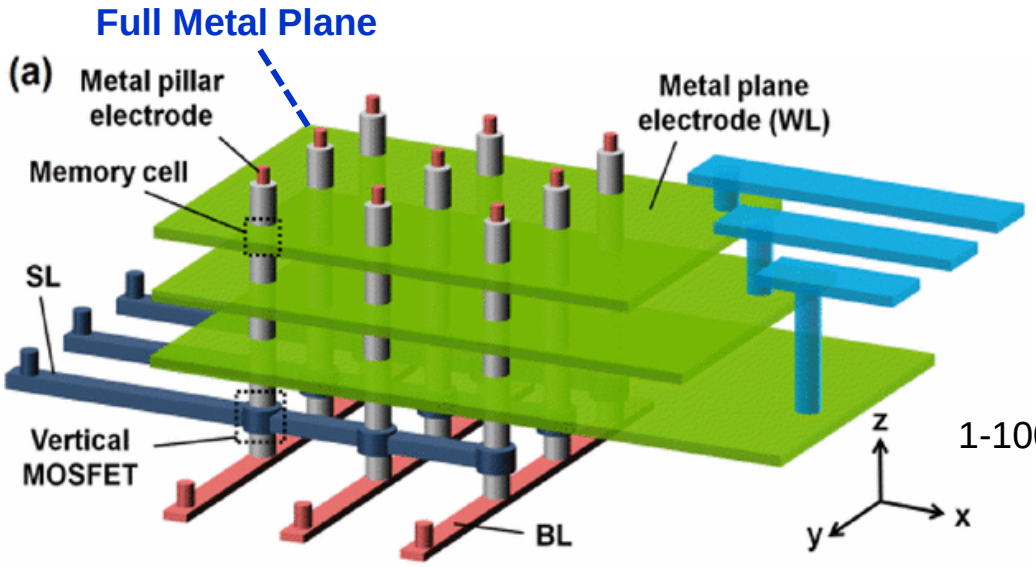


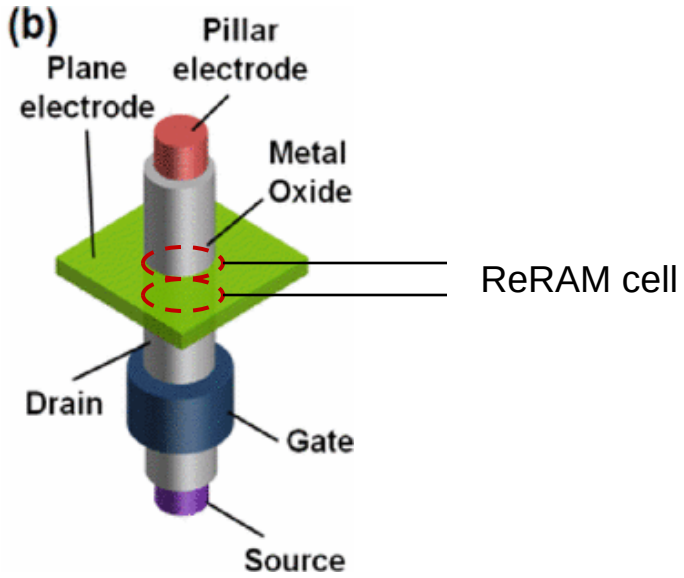


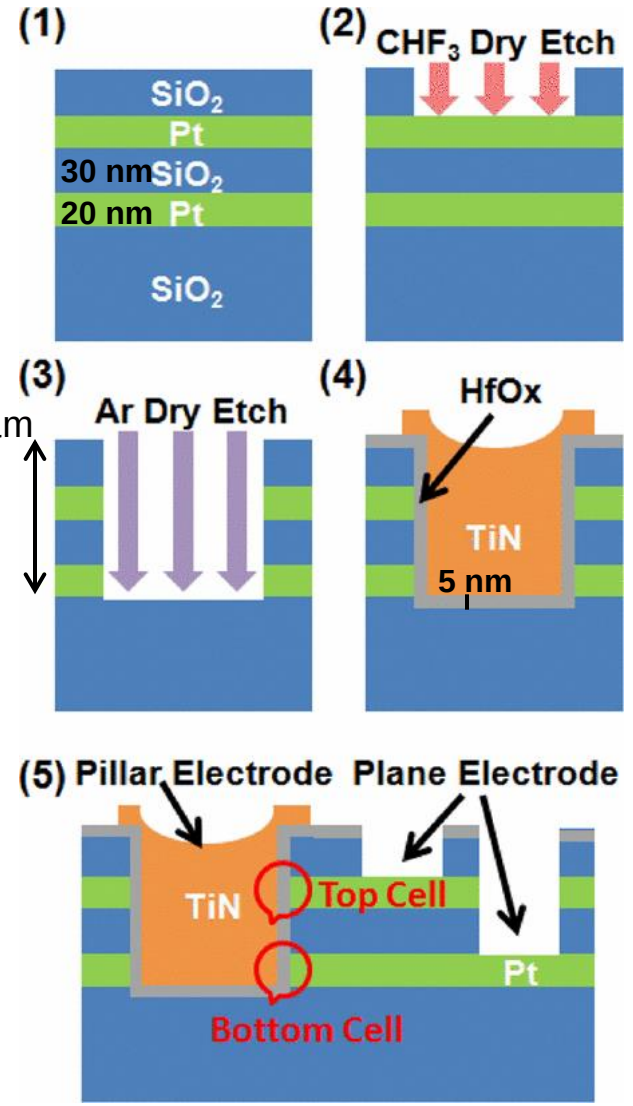


IG Baek et al (Samsung), IEDM 2011

Stanford VRRAM Architecture





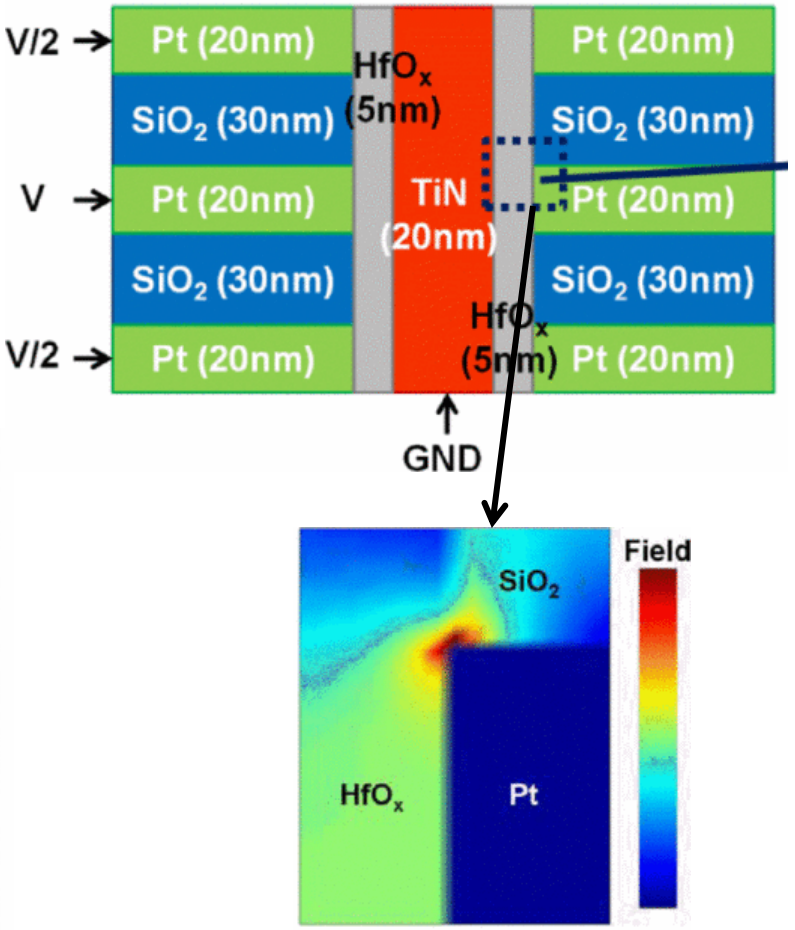


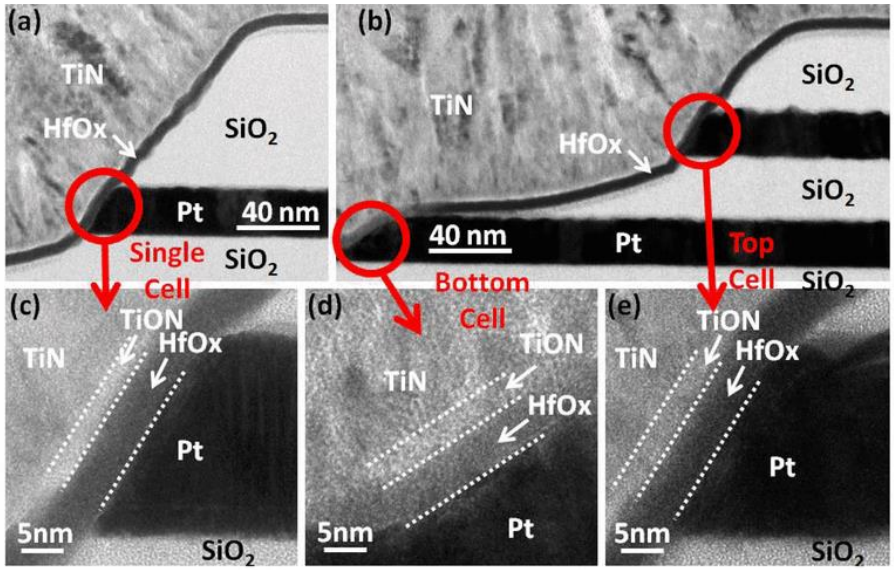
HY Chen et al (Stanford, Peking), IEDM 2012

Sandia National Laboratories

VRRAM Switching Channel Physics

- Switching channel formation altered by vertical structure
- TiON forms between HfOx and TiN
  - Increases cell resistance



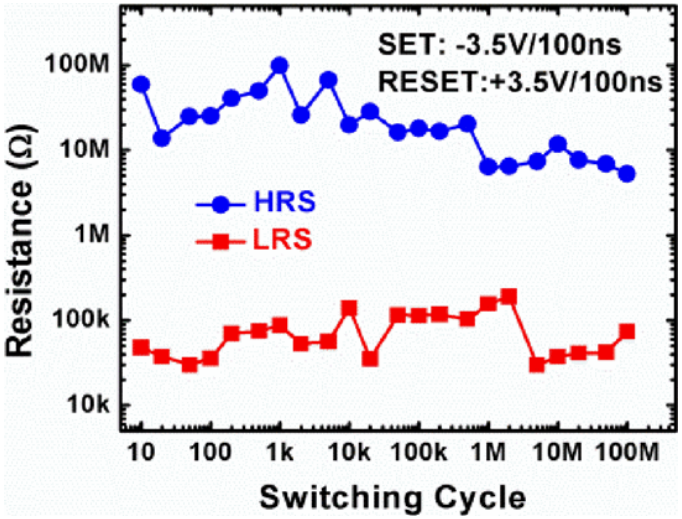


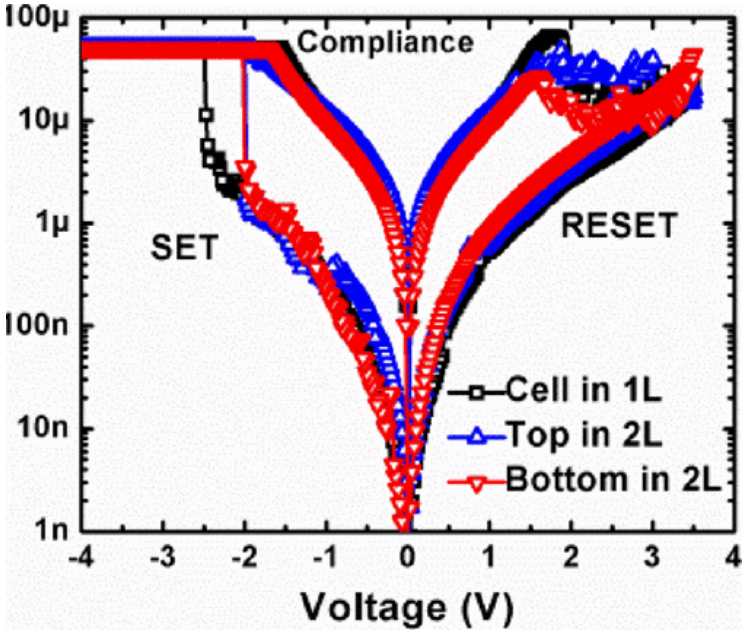
Switching channel forms on corner

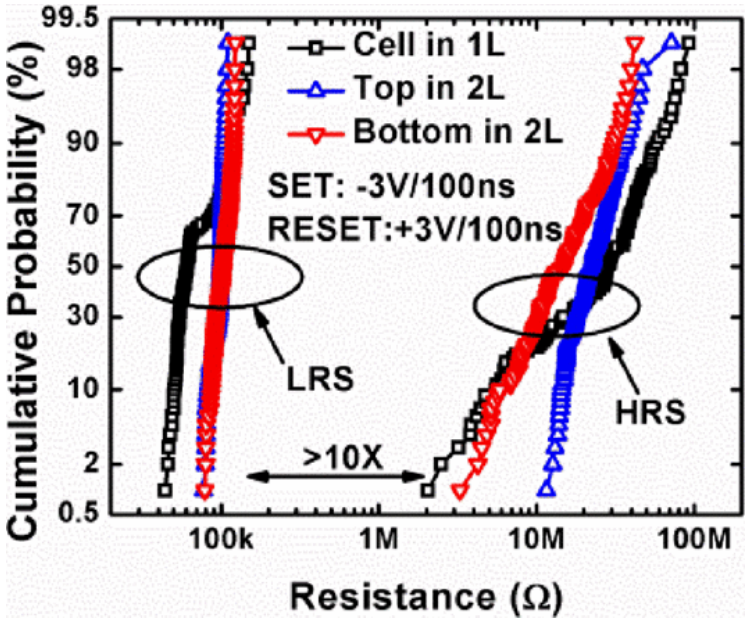
HY Chen et al, IEDM 2012

Stanford VRRAM Device Properties

- One and two layer prototypes
  - Similar properties for both layers
- 10^8 cycles endurance
- R_{OFF}/R_{ON} > 10 (with spread)
- I_{RESET} ~ 50 µA: improved but want lower
- Retention 28 hrs @ 125°C

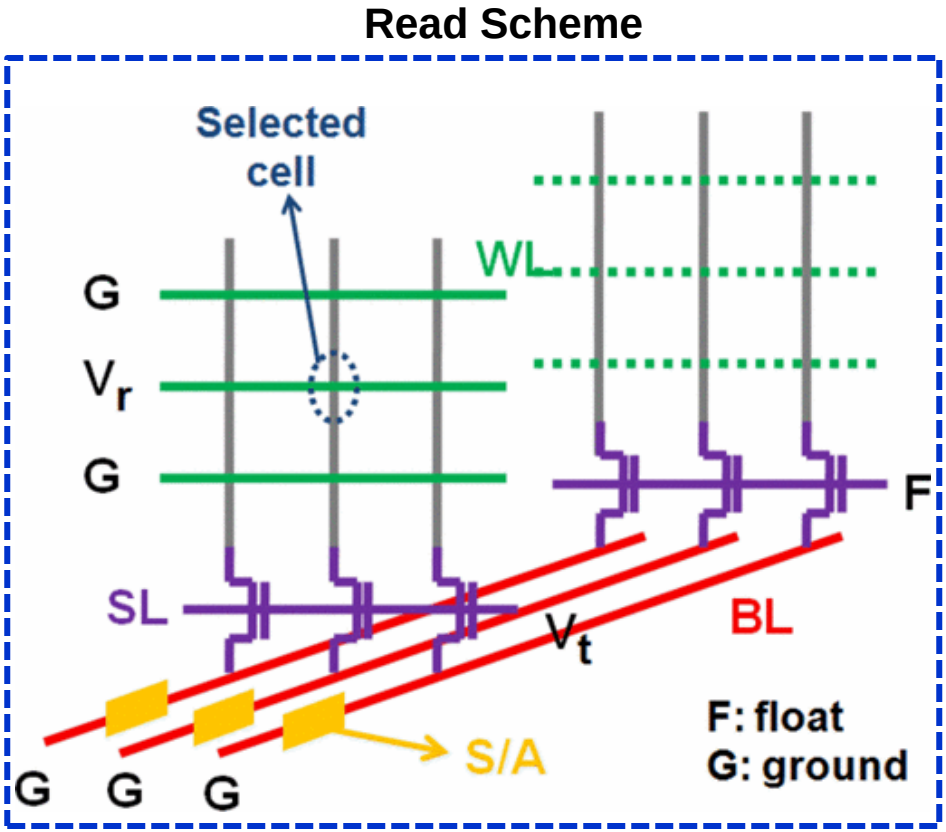


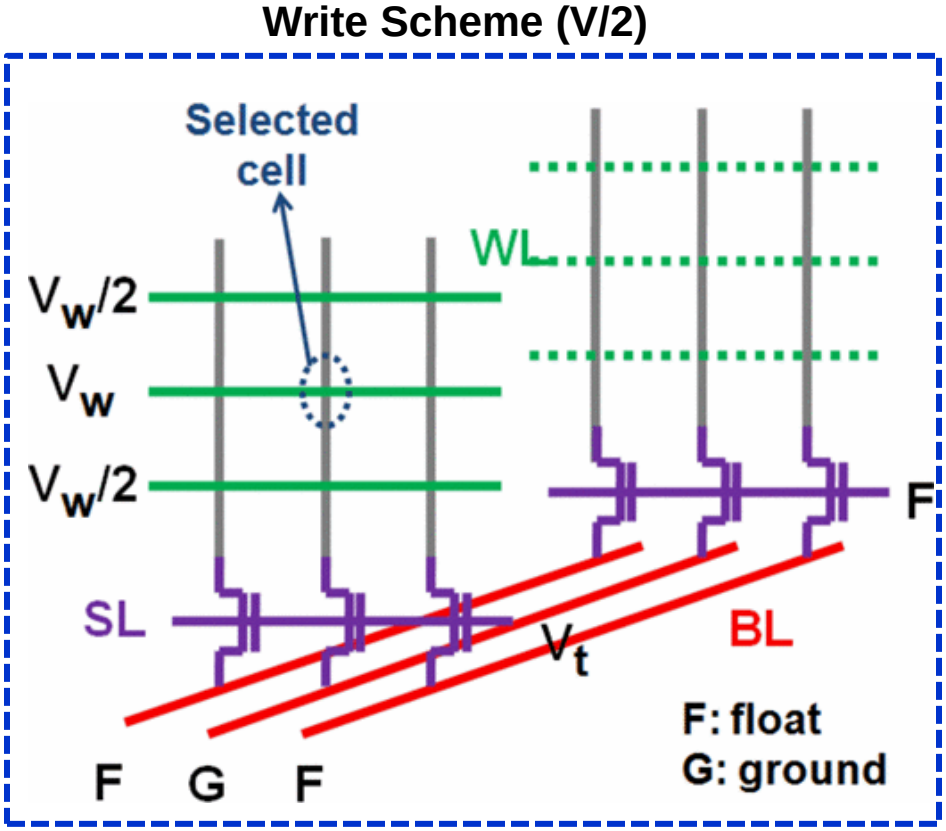




HY Chen et al (Stanford, Peking), IEDM 2012

VRRAM Array Read/Write





HY Chen et al (Stanford, Peking), IEDM 2012

Outline

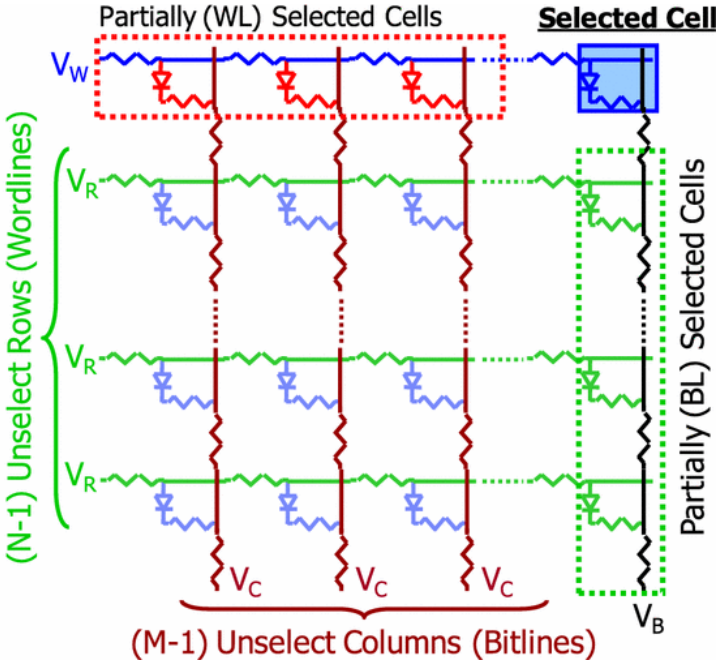
- Intro to ReRAM Device Technology
- 2D Resistive Crossbar Memories
- 3D ReRAM Technology
- 3D ReRAM Challenges
- 3D ReRAM Applications
- Summary and Future Outlook

Key Challenges for 3D ReRAM

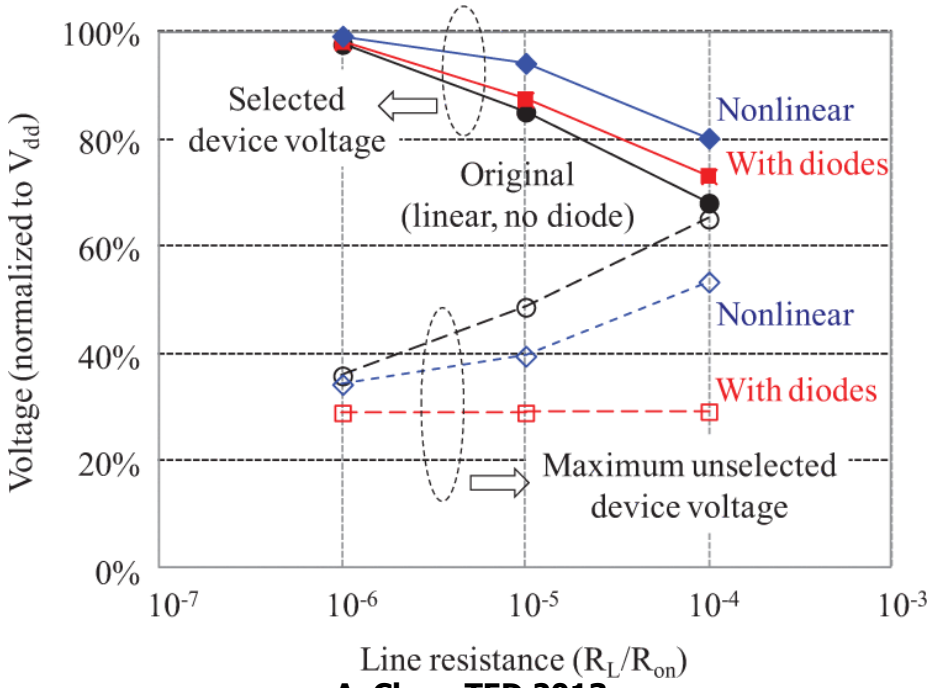
- Sneak paths and array parasitics
  - Both can limit the number of devices per row/col
  - High nonlinearity of I-V to avoid sneak paths
  - Low current (<10 μA) / high resistance to maintain read/write margin
  - Parasitic capacitance increases rise time
- Must maintain basic cell performance
  - Endurance: 10^3 to >10^8 (depends on application)
  - Retention: > 10 years at 85°C
- Device parametric variability
- Yield challenges

Sneak Paths

- Similar problem in 2D and 3D crossbars:
  - Density limited if single transistor per cell required
  - Unselected and partially selected allow current flow
  - Interferes with read and write operations
- Solution: inline select device or built-in cell IV nonlinearity



P Narayanan, G Burr et al, JEDS Sept 2015

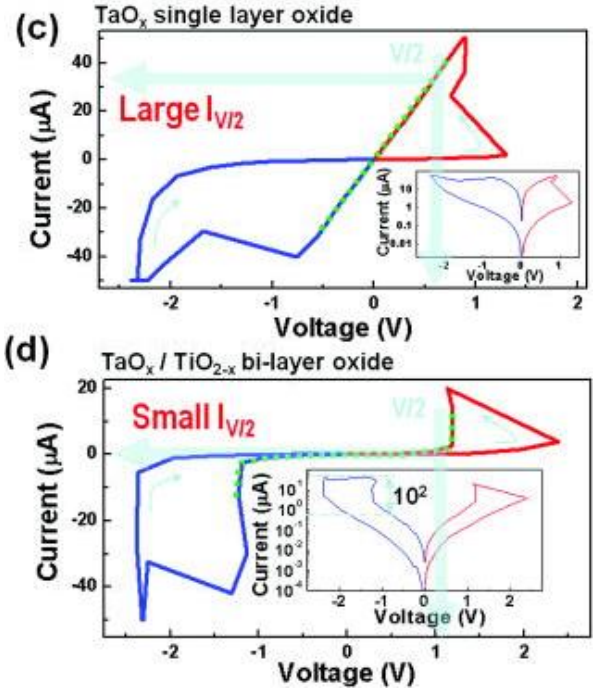


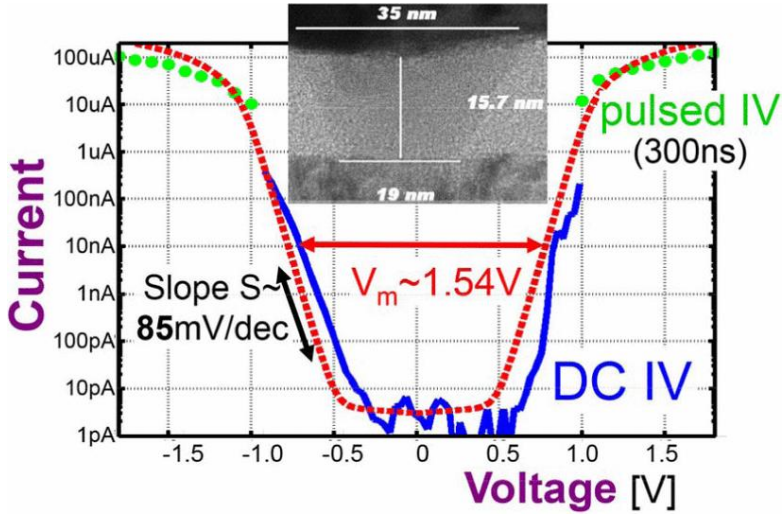
A. Chen, TED 2013

Select Device

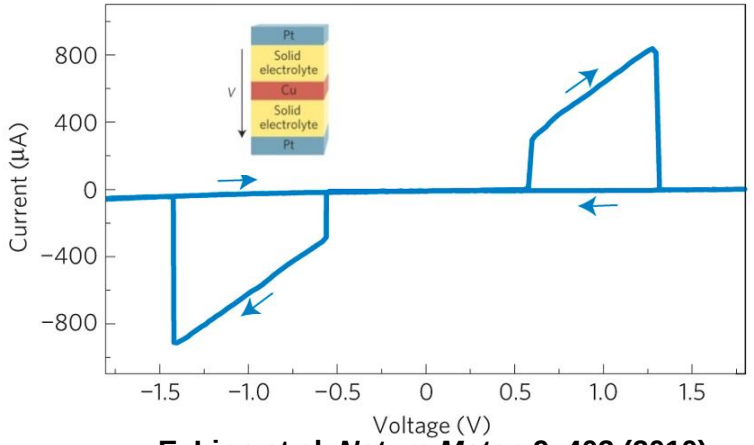
- In-line Schottky and pn diodes
- In-line threshold switch; MIT switch
- Complementary resistive switch (CRS)
- Mixed Ionic Electronic Conductor (MIEC)
- Nonlinear ReRAM cell (may use extra layer)
  - Most compatible with 3D VRRAM

J. Yang et al, APL 100, 113501, 2012.





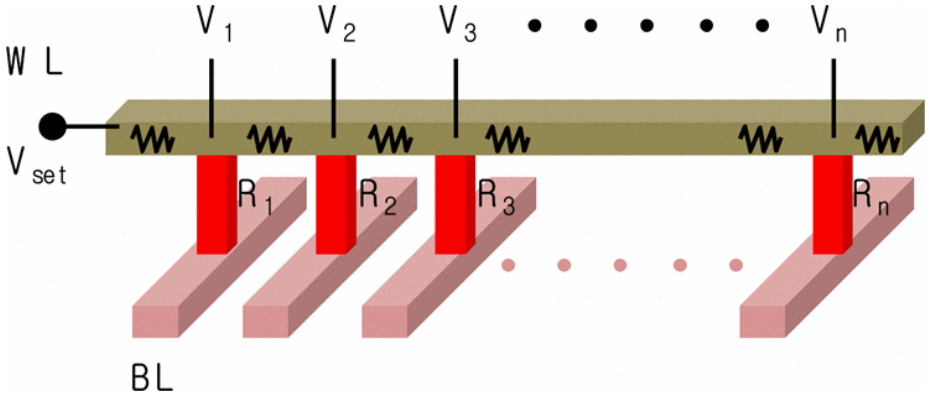
P Narayanan, G Burr et al, JEDS Sept 2015

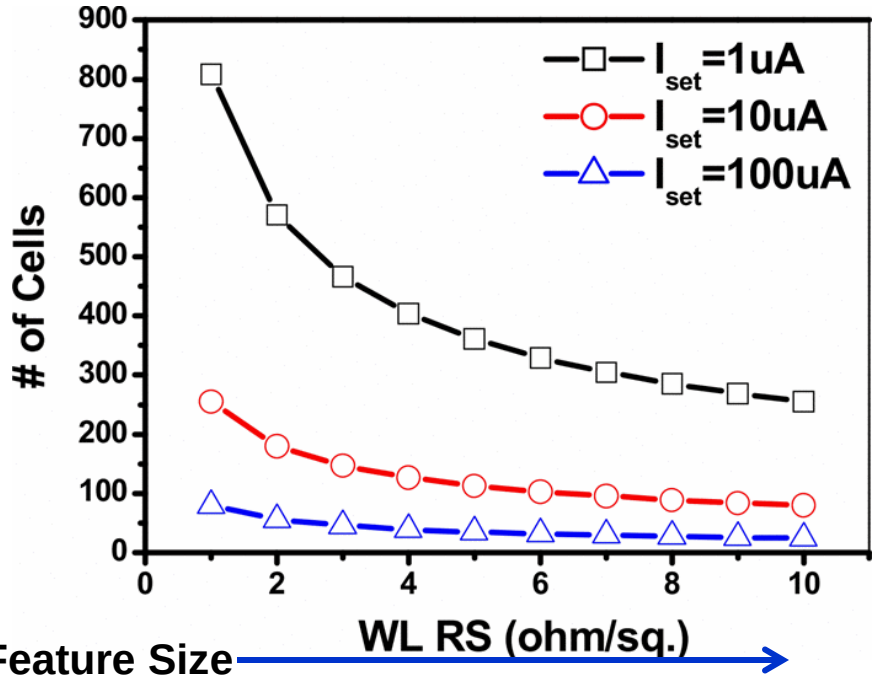


E. Linn et al, *Nature Mater.* 9, 403 (2010)

Why is Low Cell Current Needed?

- V=IR drops across line resistors
  - R ↑ with scaled lines
- Required write and read voltages set by cell
- Lower cell current→ Higher fraction of read/write voltage across selected cell→ larger array possible



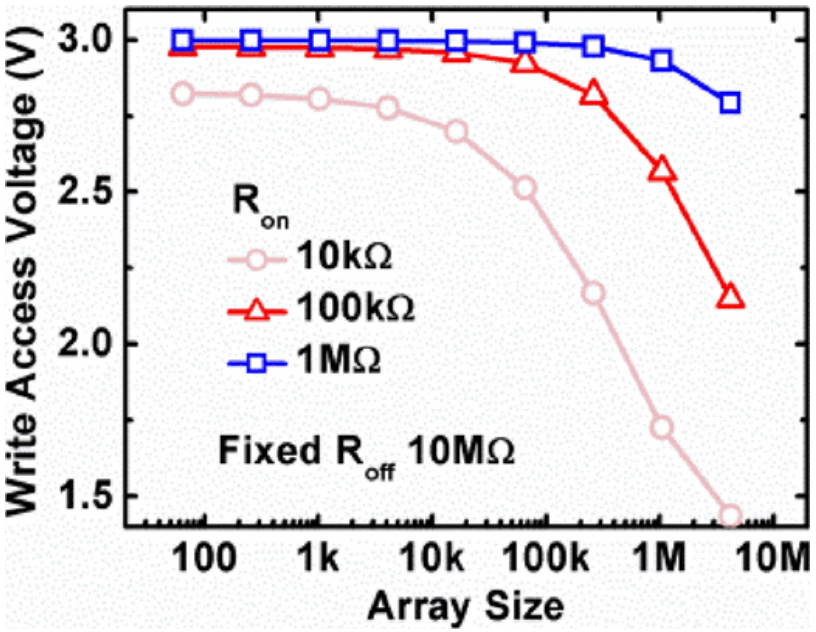


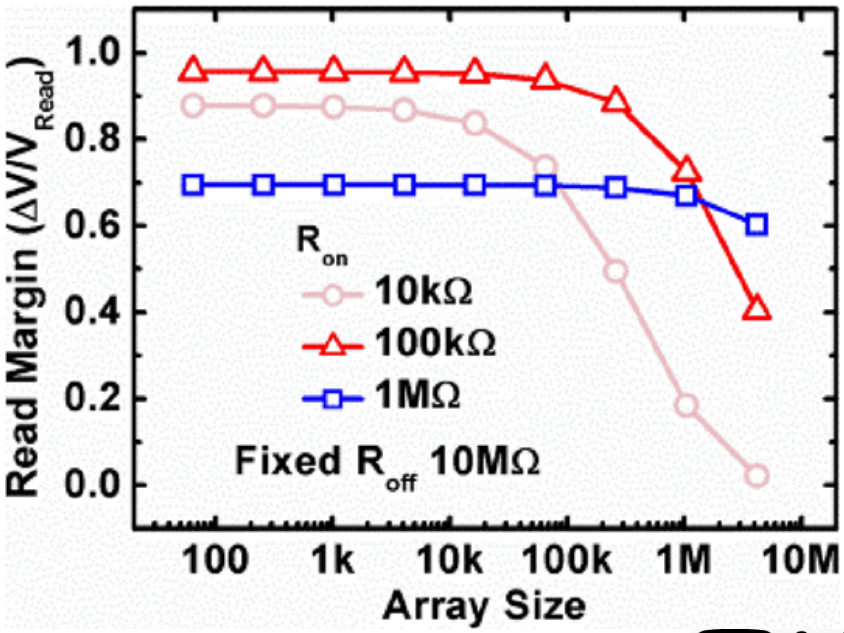
Shrinking Feature Size →

SG Park et al (Samsung), IEDM 2012

Cell Resistance

- Increasing resistance helps read and write:
  - Maximize write access voltage (voltage across cell)
  - Maintain high read margin
- Maximize array size to avoid additional overhead circuitry

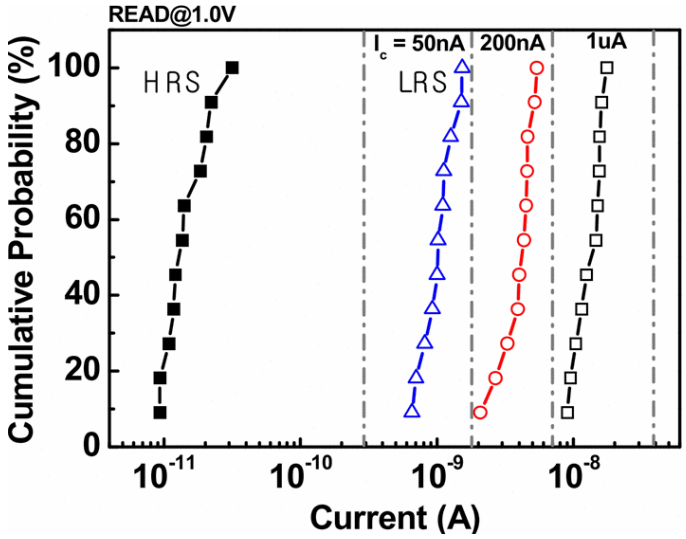


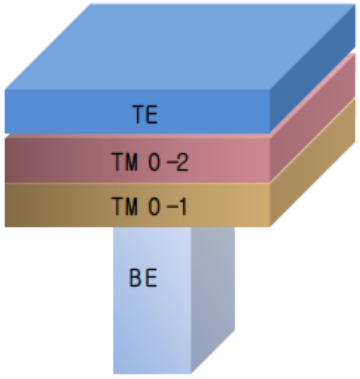


HY Chen et al, IEDM 2012

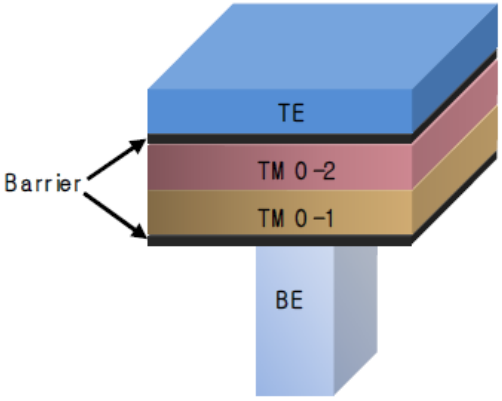
Low Current ReRAM

- SG Park et al (Samsung) have proposed a barrier layer inline with the ReRAM cell
- Acts as tunnel barrier
- Sub 1 μA MLC operation possible!
- Now need to integrate with VRRAM

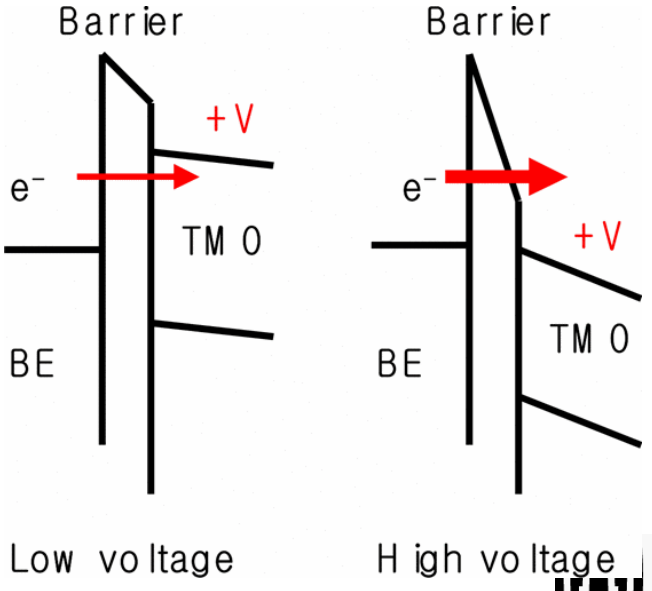




(a) High current cell



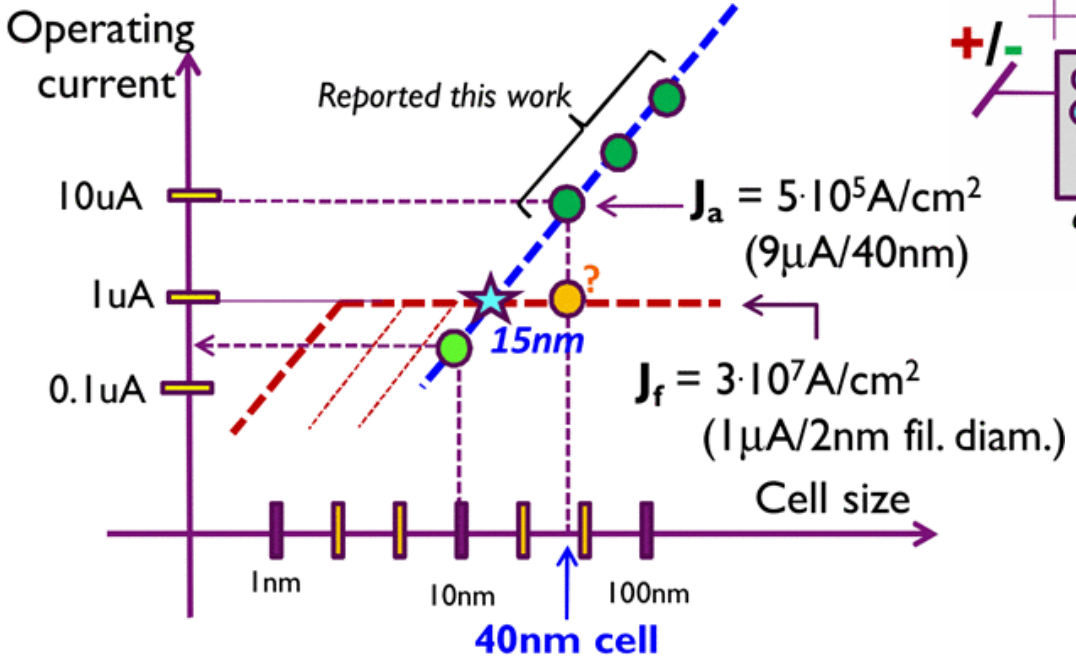
(b) Low current cell

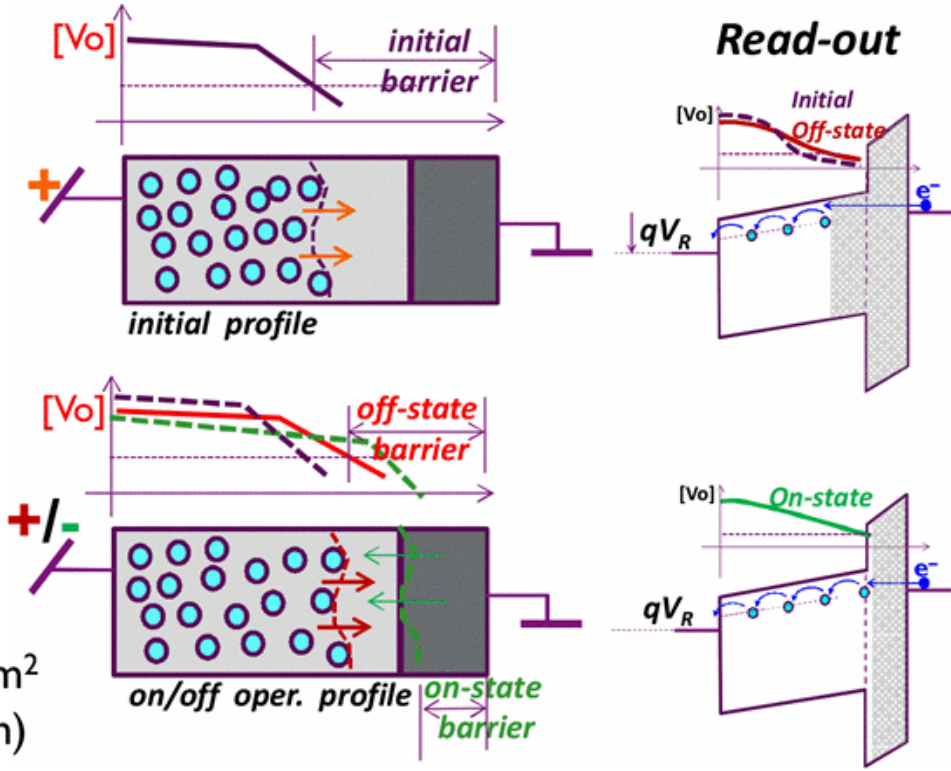


SG Park et al (Samsung), IEDM 2012

Low Current ReRAM

- Non filamentary switching
  - Vacancy modulated conducting oxide (VMCO)
- Current scales linearly with device area
  - Sub 1µA in 10 nm cell



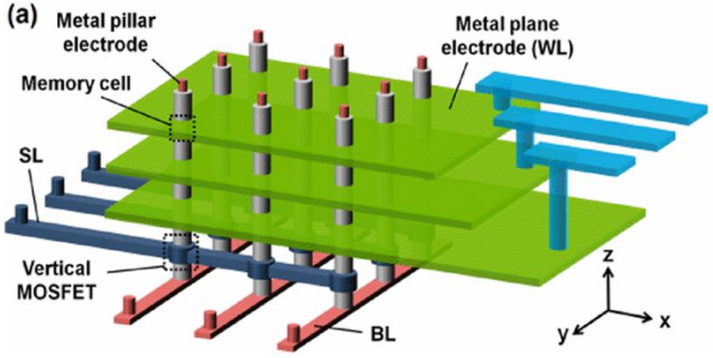


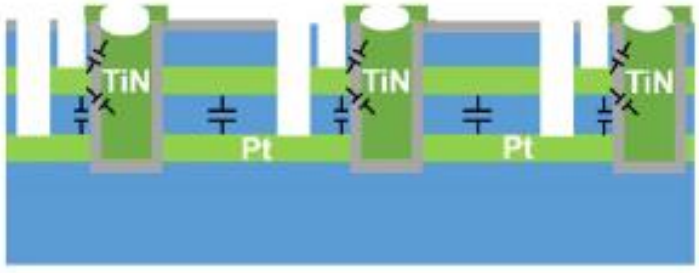
B Govoreanu et al (Imec), IEDM 2013

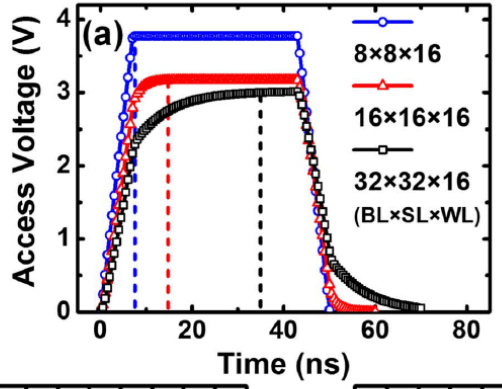
Sandia National Laboratories

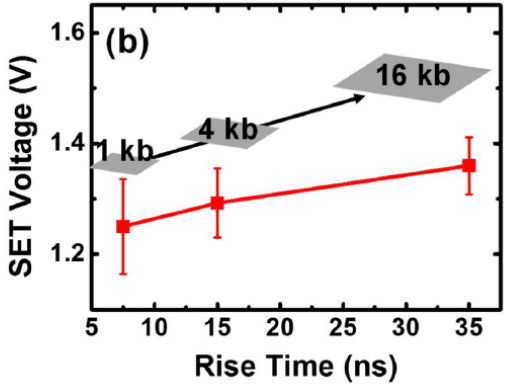
VRRAM Array Capacitance

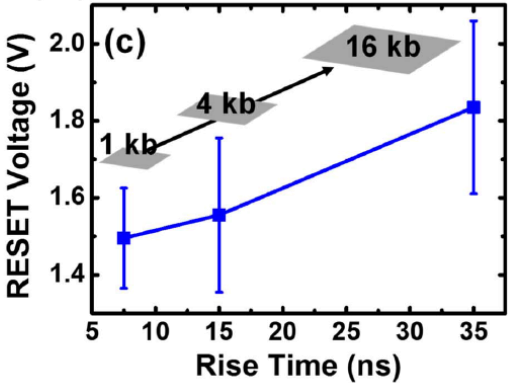
- VRRAM architecture adds additional parasitic capacitances
- Parasitic capacitance:
  - Sets max array dimensions
  - Increases energy per read/write
  - Sets lower time limit for read/write operation











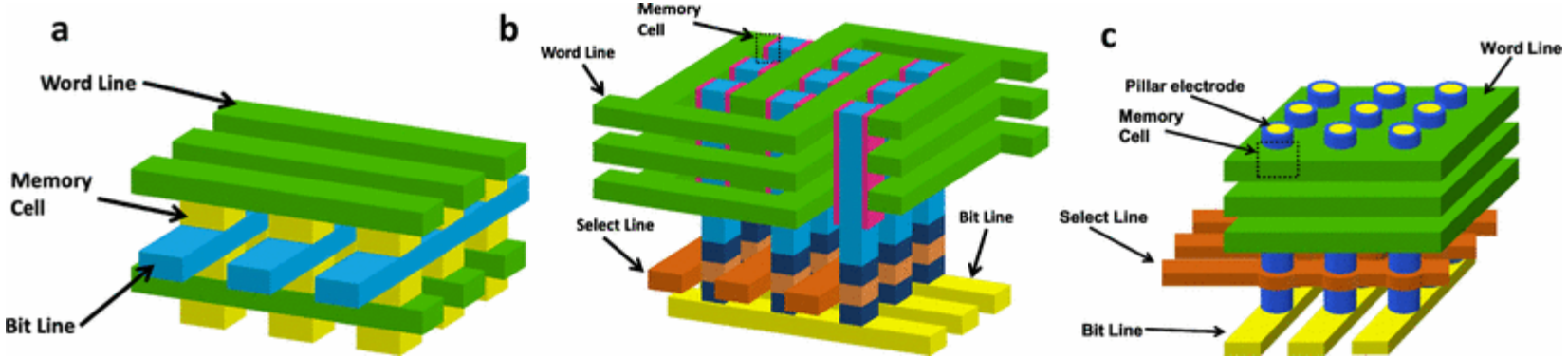
H Li et al (Stanford), TED 2015

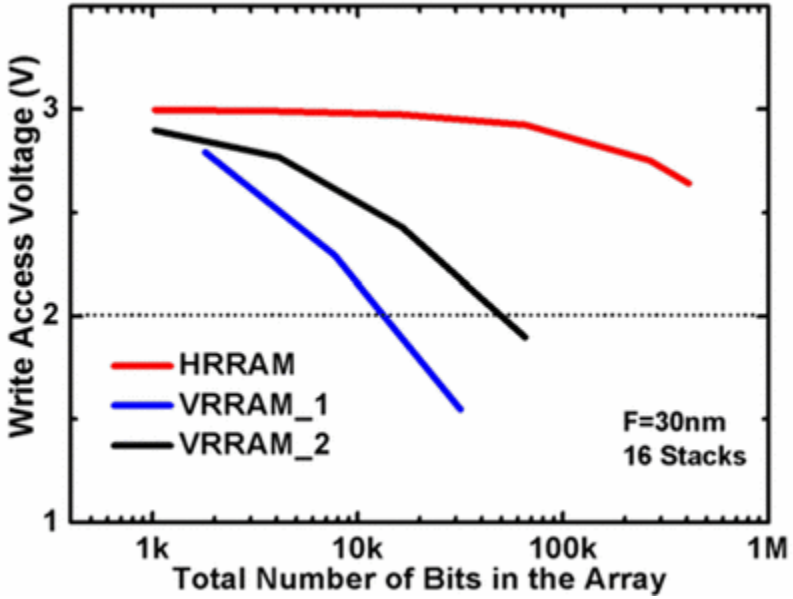
Effect of Array Parasitics: Comparison of Different 3D Architectures

3D Crosspoint "HRRAM"

Imec: Etched between pillars "VRRAM_1"

Stanford: Full Metal Planes "VRRAM_2"





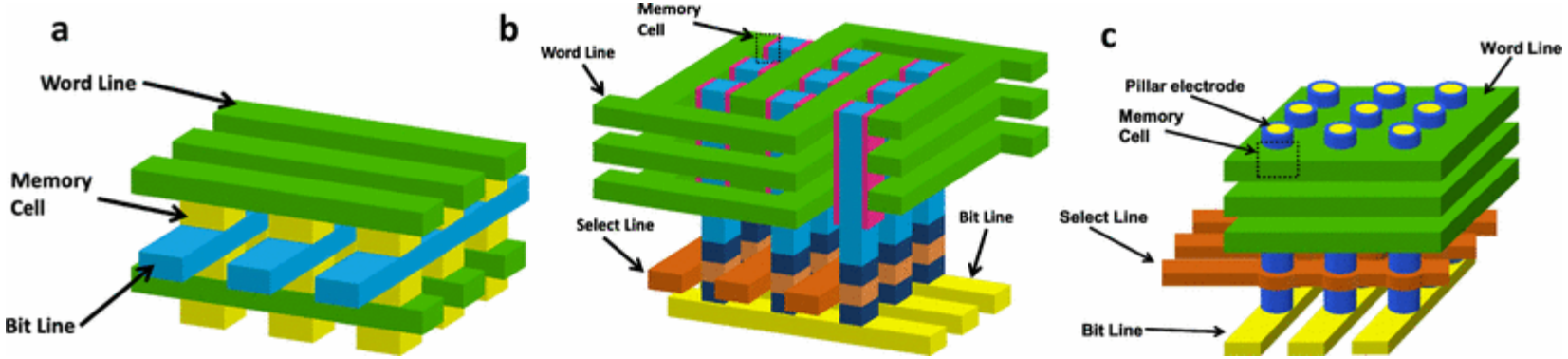
S Yu et al, ISCAS 2014

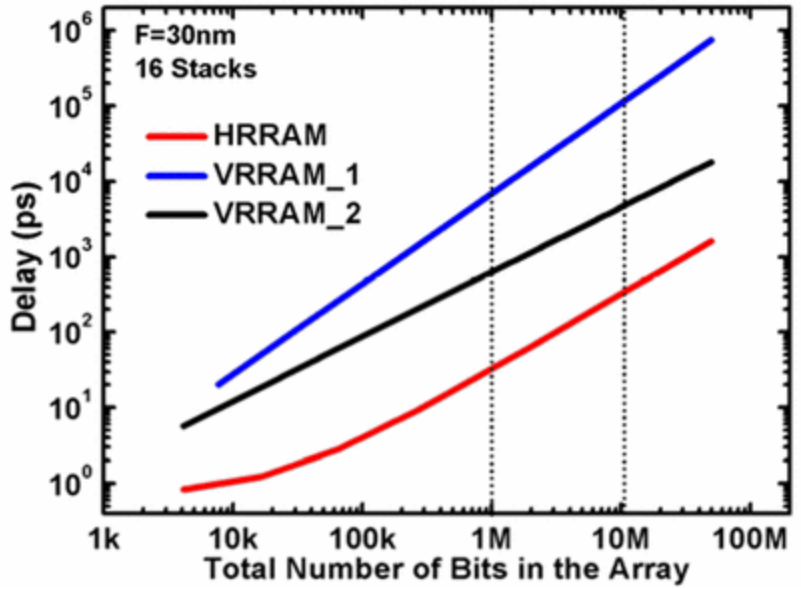
Effect of Array Parasitics: Comparison of Different 3D Architectures

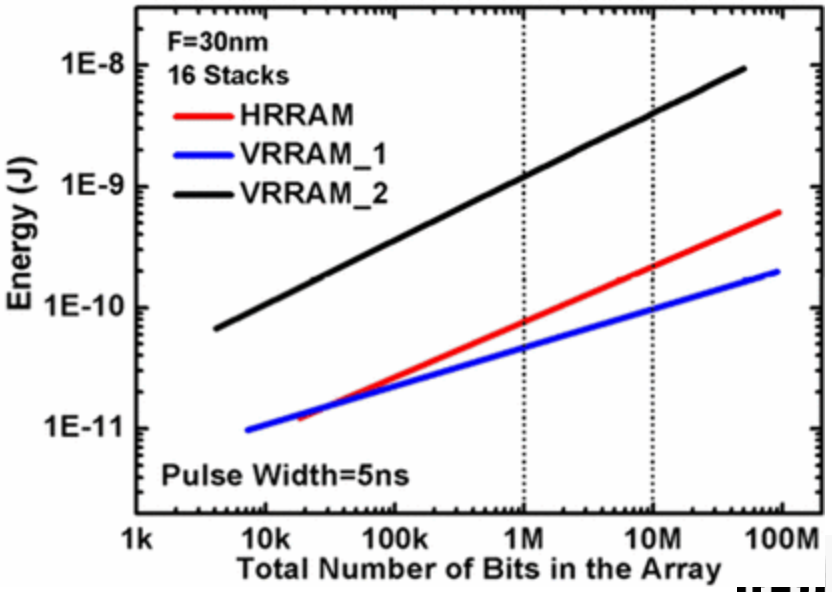
3D Crosspoint "HRRAM"

Imec: Etched between pillars "VRRAM_1"

Stanford: Full Metal Planes "VRRAM_2"







S Yu et al, ISCAS 2014

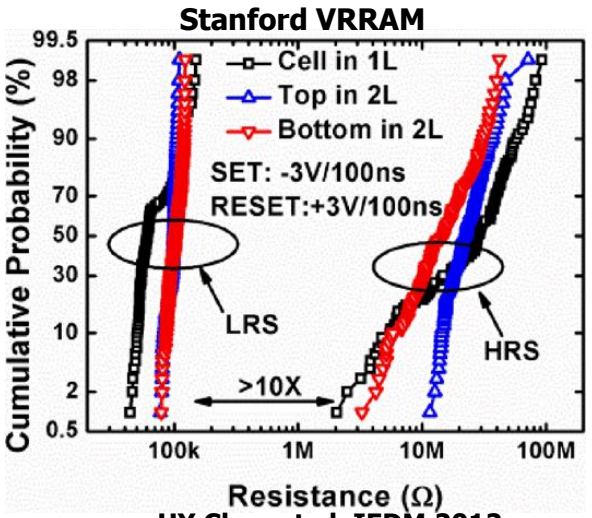
Sandia National Laboratories

Summary of 3D Vertical ReRAM Design Tradeoffs

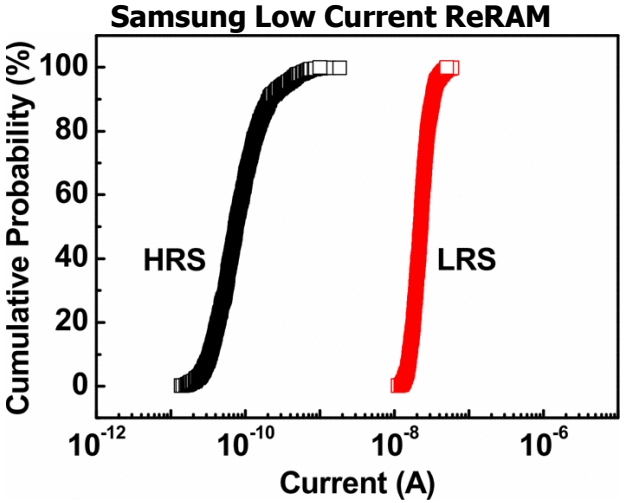
- **Maximize Bits per Array**
  - Decrease line resistance (increase V_W across cell) – Note: scaling works against this
  - Increase cell resistance (removes parasitic current)
  - Increase cell or select device nonlinearity (reduce sneak current)
- **Maximize Speed**
  - *Decrease* cell resistance
  - Decrease line RC
  - Reduce ReRAM cell dimensions
- **Minimize Write Energy**
  - Increase cell resistance
  - Decrease line RC
  - Reduce ReRAM cell dimensions
  - Increase cell nonlinearity
- Each affected by write scheme (e.g. V/2, V/3, etc)

Parametric Variations

- HRS and LRS distributions should not overlap
- Affected by write stochasticity and read noise
- Have been improved by
  - High uniformity film deposition process as ALD
  - Write verify – but cost additional system resources
- Bottom right: HRS/LRS CPD over full 300 mm wafer
  - Good single bit operation
  - Difficult to achieve MLC



HY Chen et al, IEDM 2012



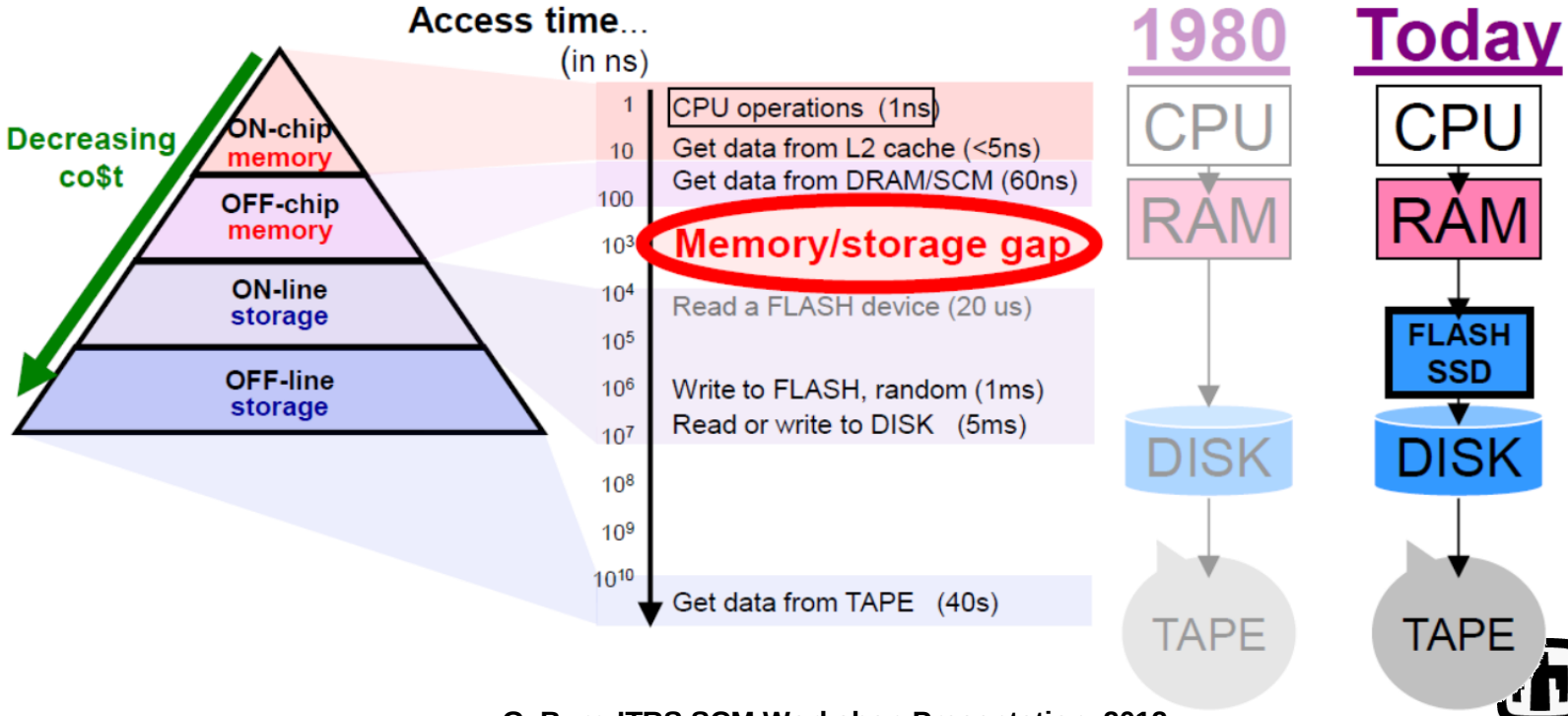
SG Park et al (Samsung), IEDM 2012

Outline

- Intro to ReRAM Device Technology
- 2D Resistive Crossbar Memories
- 3D ReRAM Technology
- 3D ReRAM Challenges
- 3D ReRAM Applications
- Summary and Future Outlook

Storage Class Memory

- **Key Observation: DRAM/Main memory 4 orders of magnitude faster than NAND SSD read (6 orders magnitude for HDD)**
- **Major bottleneck in modern computing – 1000x improvement in datacenter power and footprint possible**
- **Solve with "storage class memory" that is closer to DRAM**
- **Can 3D ReRAM solve this?**



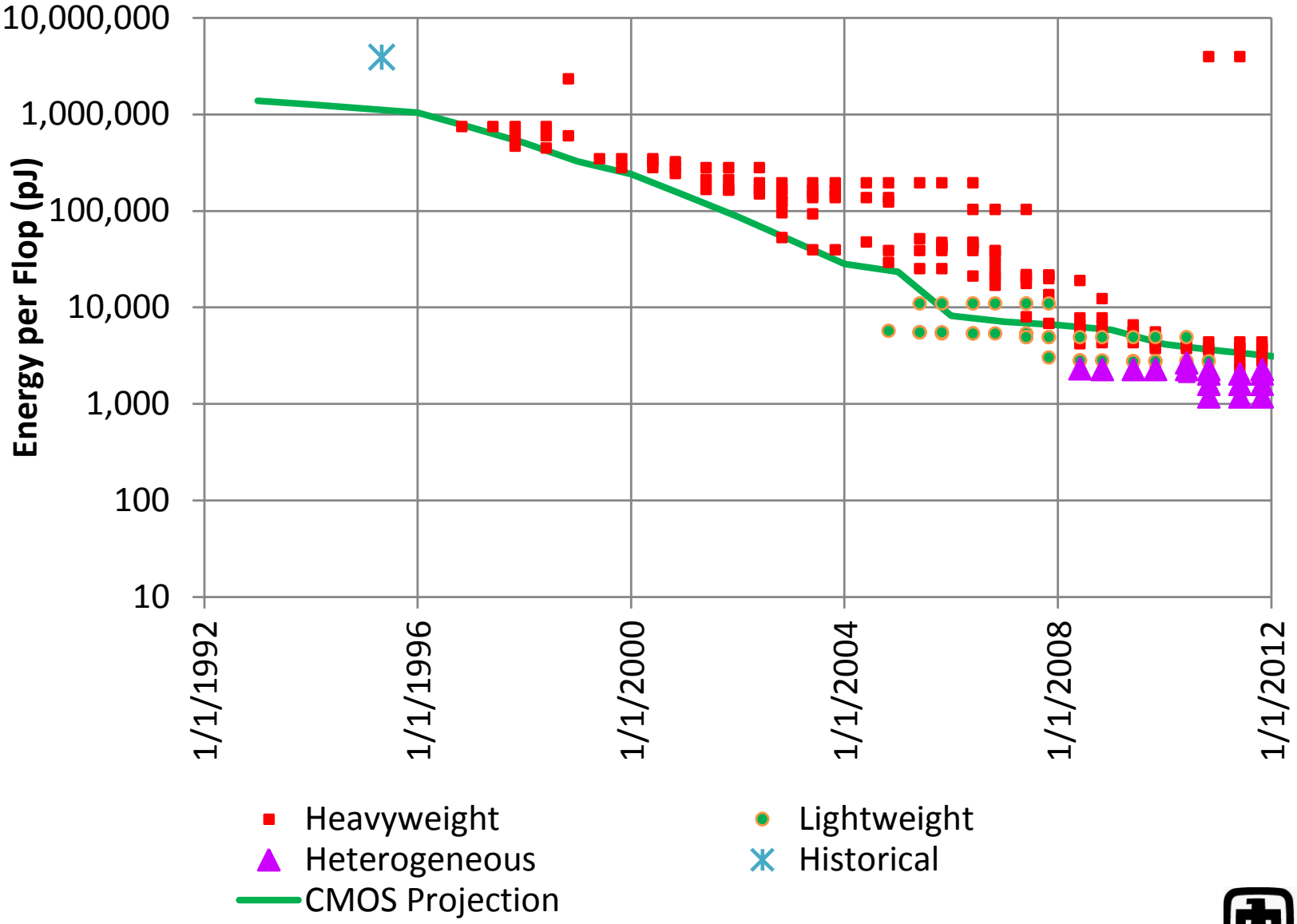
G. Burr, ITRS SCM Workshop Presentation, 2012

Exascale Computing Challenge

- Tianhe-2 (China)
  - Power: 17 MW
  - ~50 PFLOP/s
  - P/FLOPs = 500 pJ
- Titan (US – Oak Ridge)
  - Power: 8 MW
- X pJ per operation = X MW per 10^{18} operations/sec (Exaflop)
- Typical Coal Fired Power Plant
  - Power: 500 MW = 1 Exaflop with today's technology!
- Palo Verde Nuclear Generating Station
  - Power: 3 GW
- 1 MW = $1,000,000/year power bill

Will 6 ExaFLOP need dedicated Nuclear Power Plant?

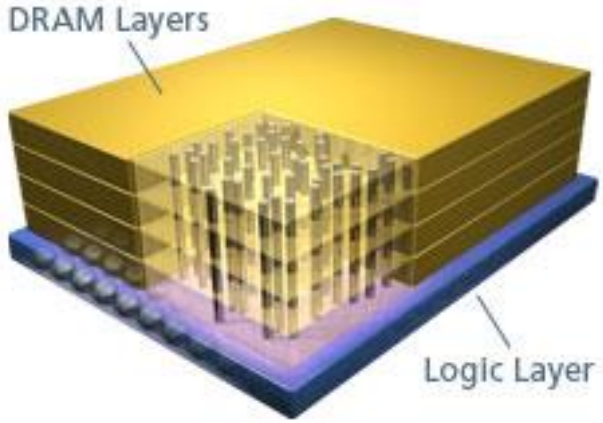
Exascale Computing Challenge

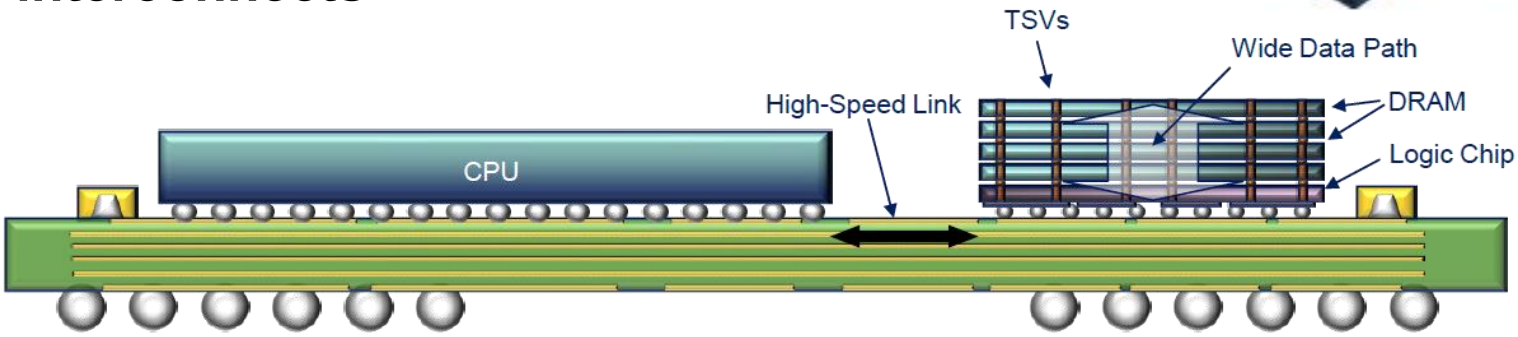


Courtesy Peter Kogge

Present Day Solution: DRAM Die Stack

- Micron/Intel Hybrid Memory Cube
- DRAM die stacked on logic
- Connected via through-silicon-via
- Combine with on-chip optical interconnects





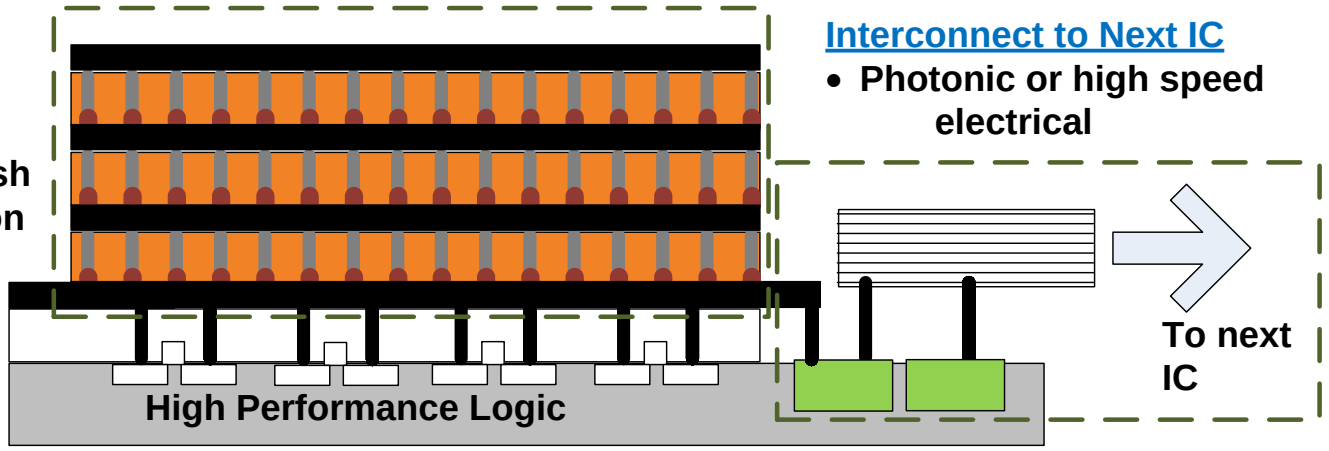
Technology	VDD	IDD	BW GB/s	Power (W)	mW/GB/s	pj/bit	real pJ/bit
SDRAM PC133 1GB Module	3.3	1.50	1.06	4.96	4664.97	583.12	762
DDR-333 1GB Module	2.5	2.19	2.66	5.48	2057.06	257.13	245
DDRII-667 2GB Module	1.8	2.88	5.34	5.18	971.51	121.44	139
DDR3-1333 2GB Module	1.5	3.68	10.66	5.52	517.63	64.70	52
DDR4-2667 4GB Module	1.2	5.50	21.34	6.60	309.34	38.67	39
HMC, 4 DRAM w/ Logic	1.2	9.23	128.00	11.08	86.53	10.82	13.7

Micron, Hotchips 2011

Beyond One ExaFLOP HPC

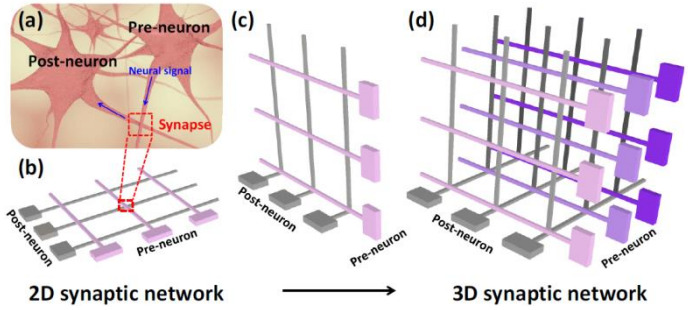
- TSV stacking should enable 1 exaflop @ 20 MW
  - 1 exaFLOP requires 20 pJ per FLOP
- What is next?
- 10 Exaflop will require 2 pJ per FLOP including memory
  - Likely will need integrated memory technologies

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



Next Generation of Neural Algorithm Computing

- **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!**
- **Significant power savings using a ReRAM based HW accelerator**
  - **Reports of as high as 10^{8}x energy efficiency over CPU-based system**
- **Also: 3D layering of memory considered "biology inspired"**



Wang et al (NTCU), IEDM 2014

Outline

- Intro to ReRAM Device Technology
- 2D Resistive Crossbar Memories
- 3D ReRAM Technology
- 3D ReRAM Challenges
- 3D ReRAM Applications
- Summary and Future Outlook

Summary and Future Outlook

- ReRAM is a rapidly emerging memory device technology
  - Especially oxide-based bipolar filamentary and conducting bridge categories
- 3D ReRAM will enable scaling to >terabit per chip densities
- VRRAM allows many layers without adding critical masks
- Challenge: Create a 3D ReRAM that meets all requirements:
  - Minimize effect of sneak paths and array parasitics
  - Need <10μA write current (<1μA better)
  - Maintain excellent endurance, retention, yield
  - Excellent uniformity across die and wafer
- Intense research topic in the device community
  - Large scale 3D crosspoint product announced in 2015
- Key Future Applications:
  - 3D ReRAM Storage Class Memory (5-10 yrs)
  - Beyond Exascale computing; neural computing

Microsystems
Science, Technology & Components

Questions??

Sandia National Laboratories

Contact Info

Matthew Marinella
Advanced Semiconductor Device R&D
Sandia National Laboratories
505-844-7848
matthew.marinella@sandia.gov

Acknowledgements

- SNL ReRAM team
- This work was partially funded by Sandia's Laboratory Directed Research and Development (LDRD) Program