Next Generation Nonvolatile Memory

Its Impact on Computer System

Dec.04.2013

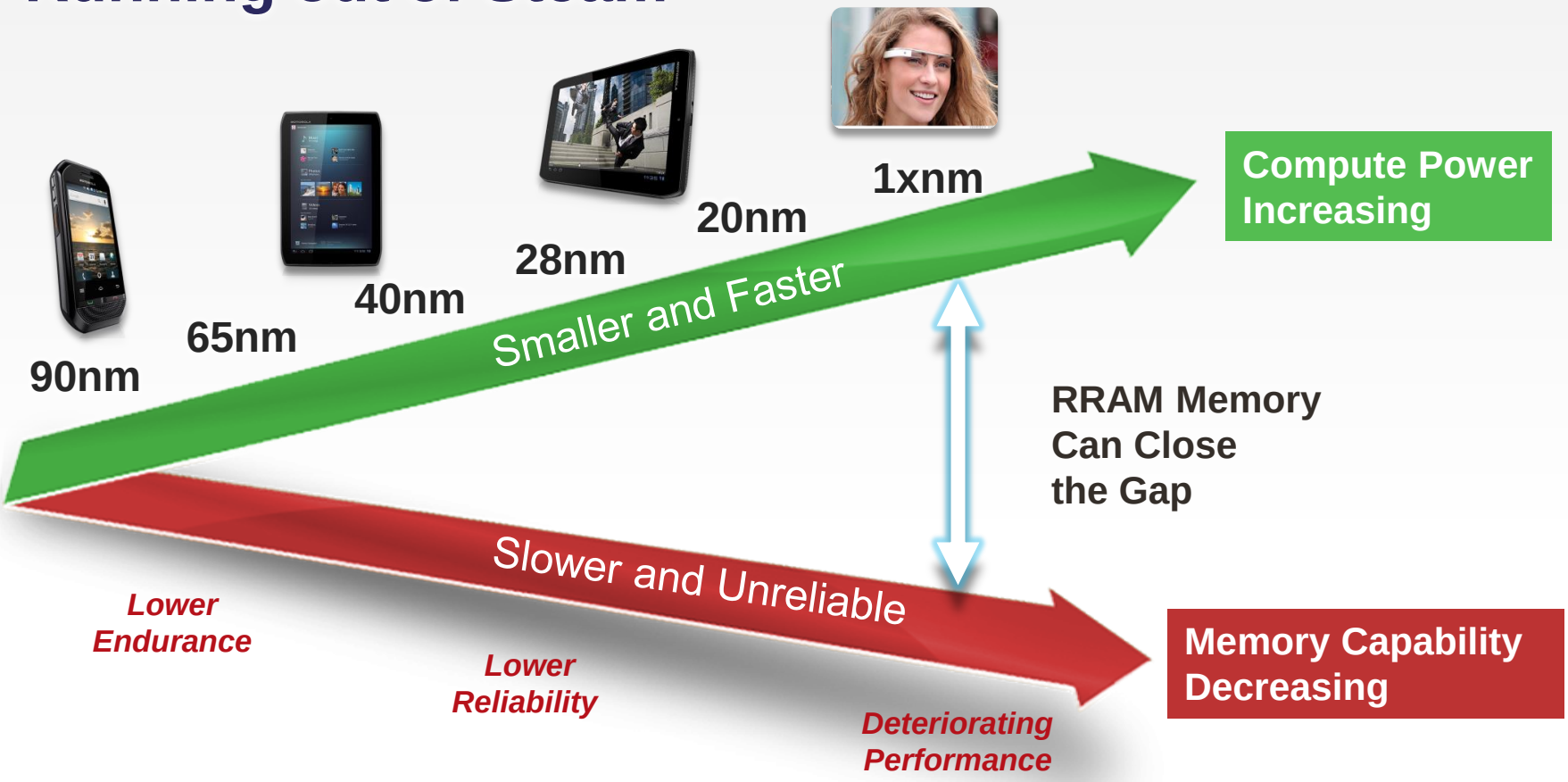
Sung Hyun Jo and Hagop Nazarian

Crossbar

Next Generation Nonvolatile Memory
Its Impact on Computer System

- Challenges of current NVM technology
- Requirements for next generation memory
- Next Generation Memory Developments
- Operation Mechanism of various RRAMs
- Advanced RRAM Technology
- Design & Architectural attributes
- System Benefits
- Comparison with current NVM technology

Traditional Non-Volatile Memory Technology is Running out of Steam

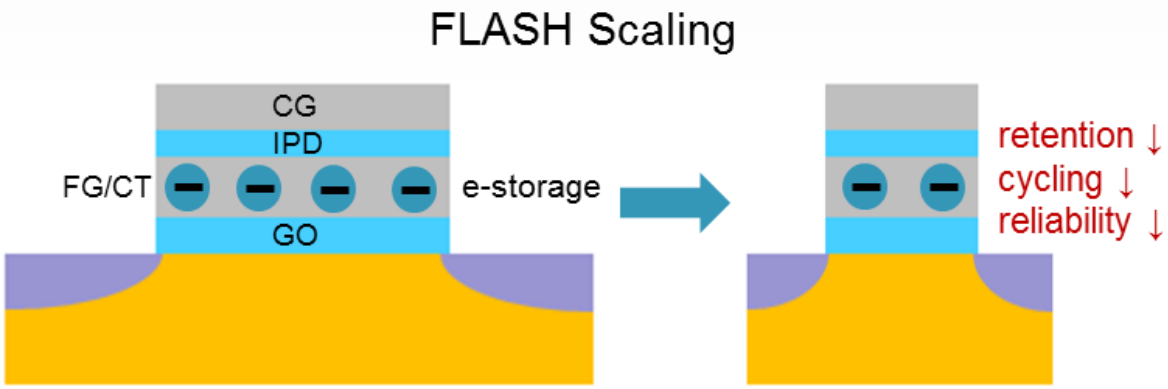


"For several years now, companies have focused on developing a next generation memory technology that will lead to significant improvements in reliability, performance, low power operation and scalability compared to existing non-volatile memories. ***Forward Insights believes that RRAM, including Crossbar's approach***, has the potential to succeed NAND flash memory due to its scalability and manufacturability." – Greg Wong, Forward Insights, August 2013

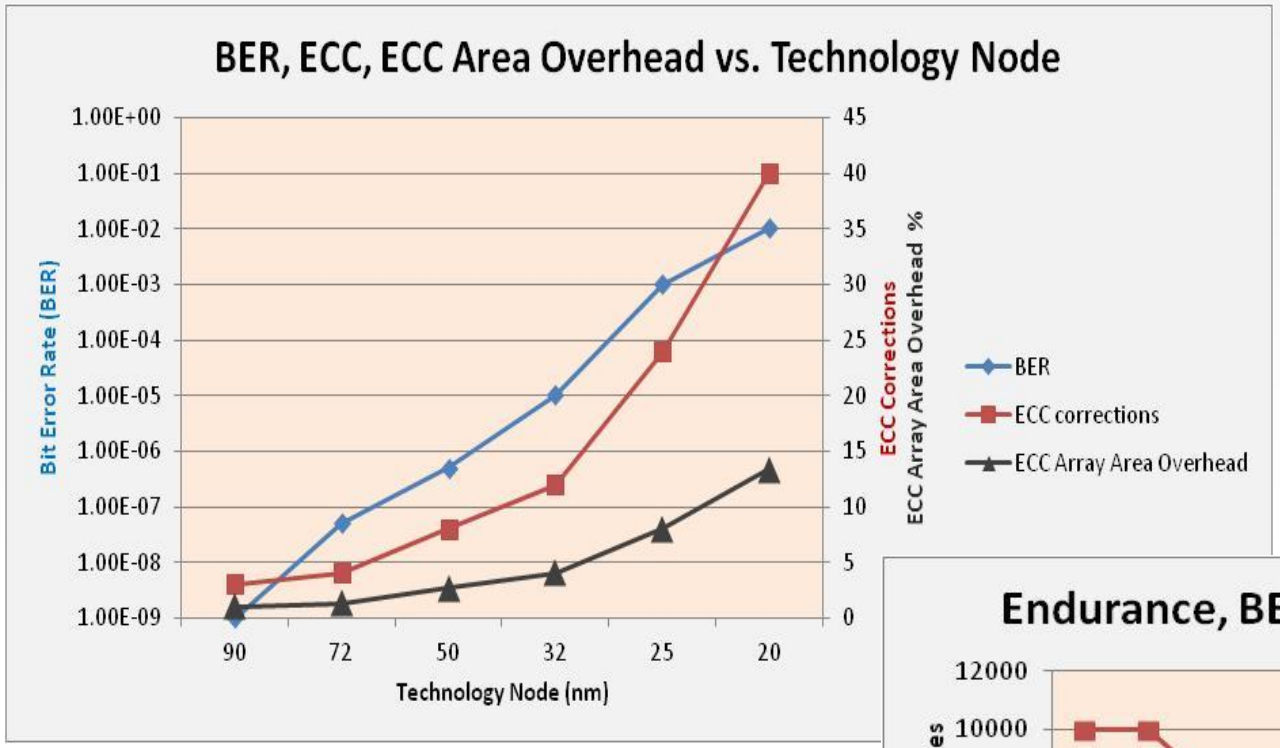
"The current storage medium, planar NAND, is seeing challenges as it reaches the lower lithographies, pushing against physical and engineering limits. ***The next generation non-volatile memory, such as Crossbar's RRAM, would bypass those limits***, and provide the performance and capacity necessary to become the replacement memory solution." - Michael Yang, IHS, August 2013

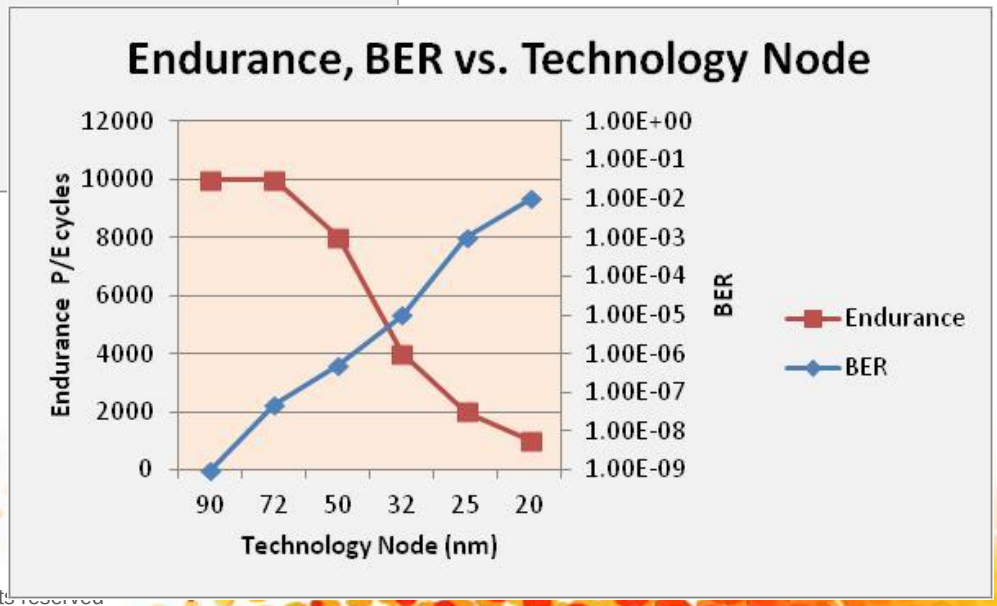
Flash Memory Scaling Challenge

- Information storage in Flash is based on charge density (C/cm^2)
- At 20nm, ~100 electrons are stored in the FG ($\Delta Vt = 1V$)
- Losing a few electrons can cause severe reliability issues
- Scaling = exponentially increasing BER, reduced data retention and cycling



Scaling challenges on BER, and Endurance





System Requirements For Next Generation Memory

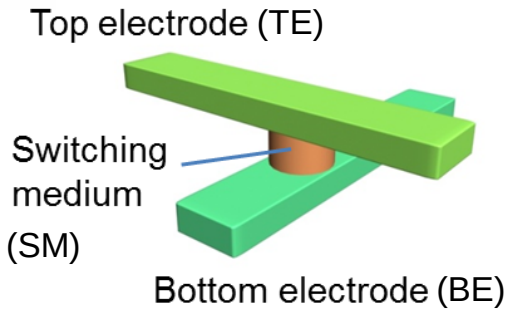
- Reduce Latency
- Lower Power Consumption
- Improved Reliability and Higher P/E Cycles
- Scalable to several generations
- Embeds in advanced CMOS technology nodes
- Cost effective

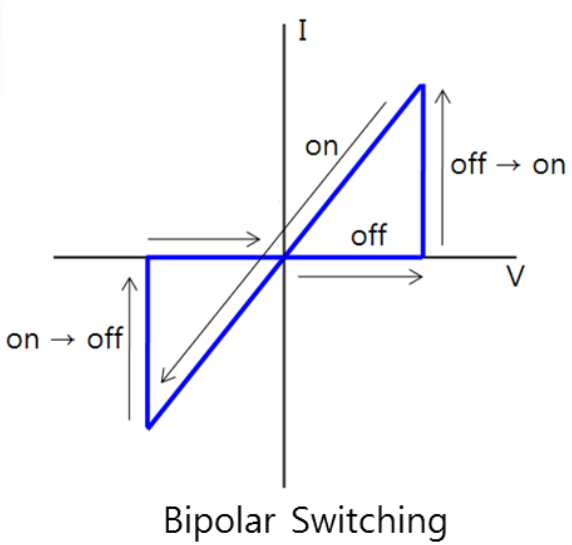
- RRAM is the emerging technology with impressive characteristics. It will meet the demands for next generation systems

Introduction to RRAM Technology

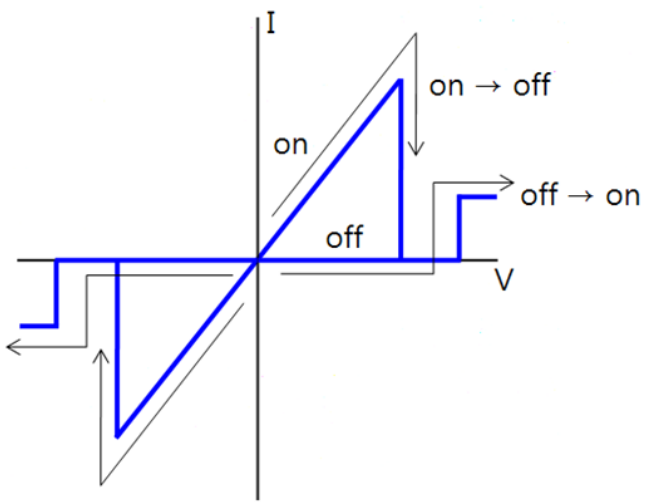
Resistive Random Access Memory (RRAM)

- Non-charge based emerging nonvolatile memory technology
- Typically two terminal structure
- Information storage based on multiple electrical resistance states
  - Resistance switching by voltage or current signal
- Either bipolar and/or unipolar switching





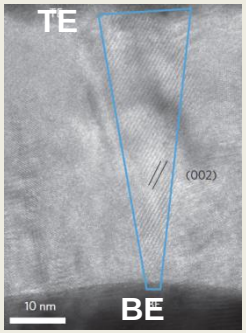
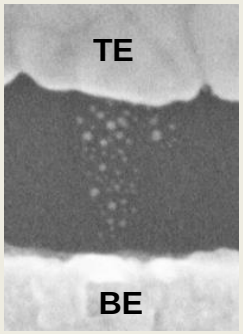
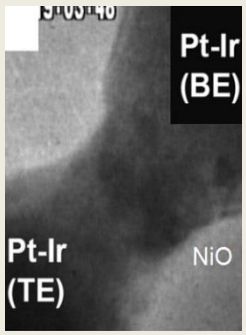
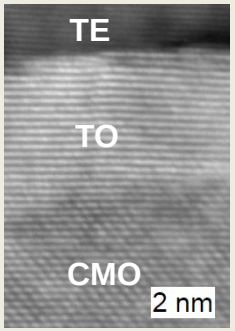
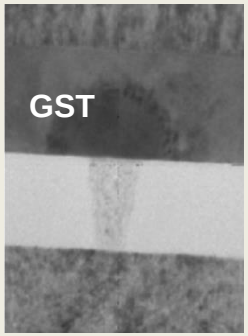
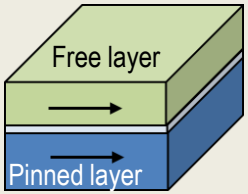
Bipolar Switching



Unipolar Switching
(Non-polar)

Resistance Switching Classification

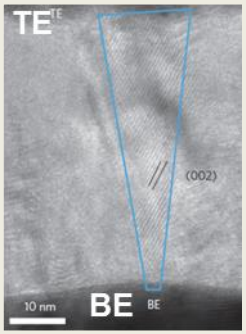
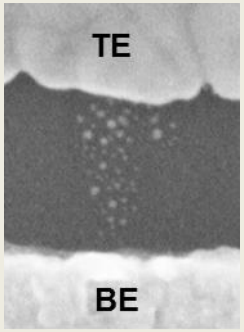
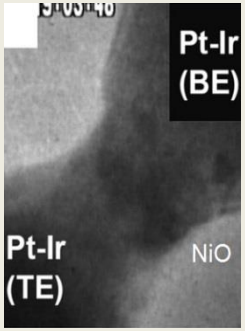
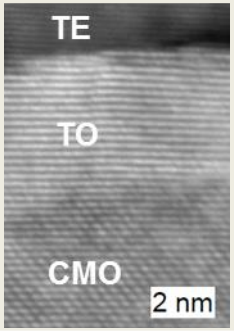
- RRAM utilizes 1D or 2D effect → ultimate scaling potential

Resistance Switching						
Switching Mechanism	Valence change	Electro-chemical metallization	Thermo-chemical	Uniform V_o exchange	Thermal	Magneto resistance
Switching Polarity	Bipolar	Bipolar	Unipolar	Bipolar	Unipolar	Bipolar
Device Example						
Physical Effect	**1D** Filament			**2D** Interface	**3D** Bulk	
NVM Category	RRAM				PCRAM	MRAM

Images from -
1. Kwon et al., Nat. Nanotech. (2010)
2. Lu et al., Nat. Commun. (2012) (Crossbar)
3. Fujii et al., J. Appl. Phys. (2013)
4. Sanchez et al., NCCAVS (2009)
5. Sebastian et al., J. Appl. Phys. (2011)

RRAM

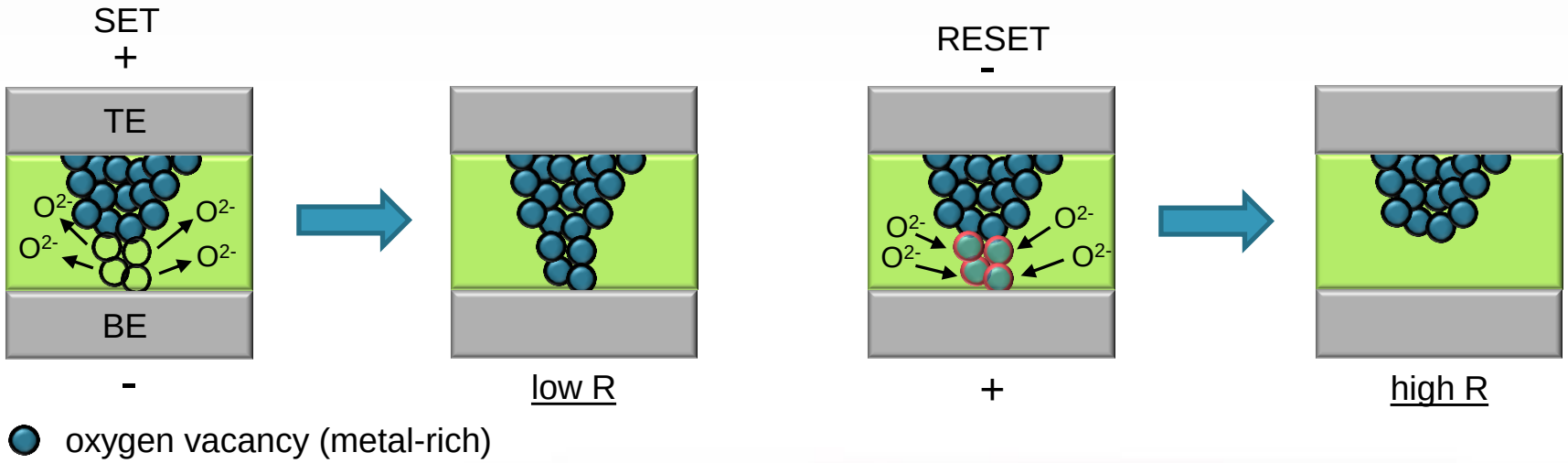
- Discrete 1D filament allows low power, high density & reliable RRAM

RRAM			
Valence change	Electro-chemical metallization	Thermo-chemical	Uniform V_o exchange
Bipolar	Bipolar	Unipolar	Bipolar
			
Continuous CF	Discrete CF / Continuous CF	Continuous CF	2D Interface
1D Conducting Filament (CF)			

Images from -
1. Kwon et al., Nat. Nanotech. (2010)
2. Lu et al., Nat. Commun. (2012) (Crossbar)
3. Fujii et al., J. Appl. Phys. (2013)
4. Sanchez et al., NCCAVS (2009)
5. Sebastian et al., J. Appl. Phys. (2011)

Valence Change RRAM

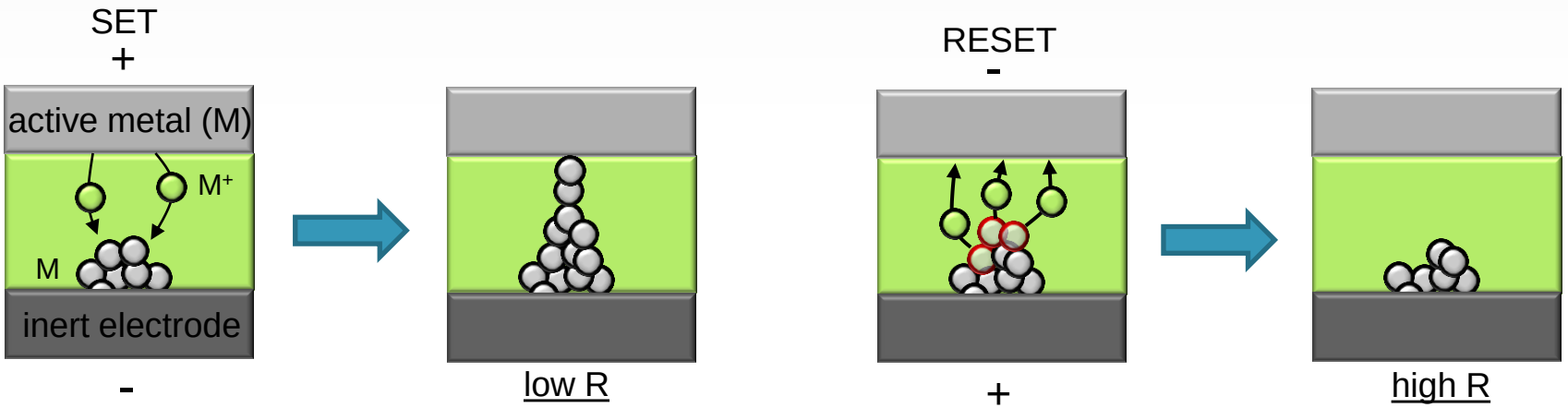
- Bipolar switching by *the migration of oxygen* under electric field
- Switching medium – typically transition metal oxide (e.g. TaO_x, HfO_x, TiO_x)
- Electrode – typically inert metal (e.g. Pd, Pt)
- SET – generation of oxygen vacancies and formation of a filament(s)
- RESET – oxidation of the filament(s)



*Actual filament(s) growth direction (e.g. BE → TE, TE → BE) depends on several factors such as switching layer material (e.g. oxygen deficient vs. metal deficient, electron affinity) and bias scheme

Electrochemical Metallization RRAM

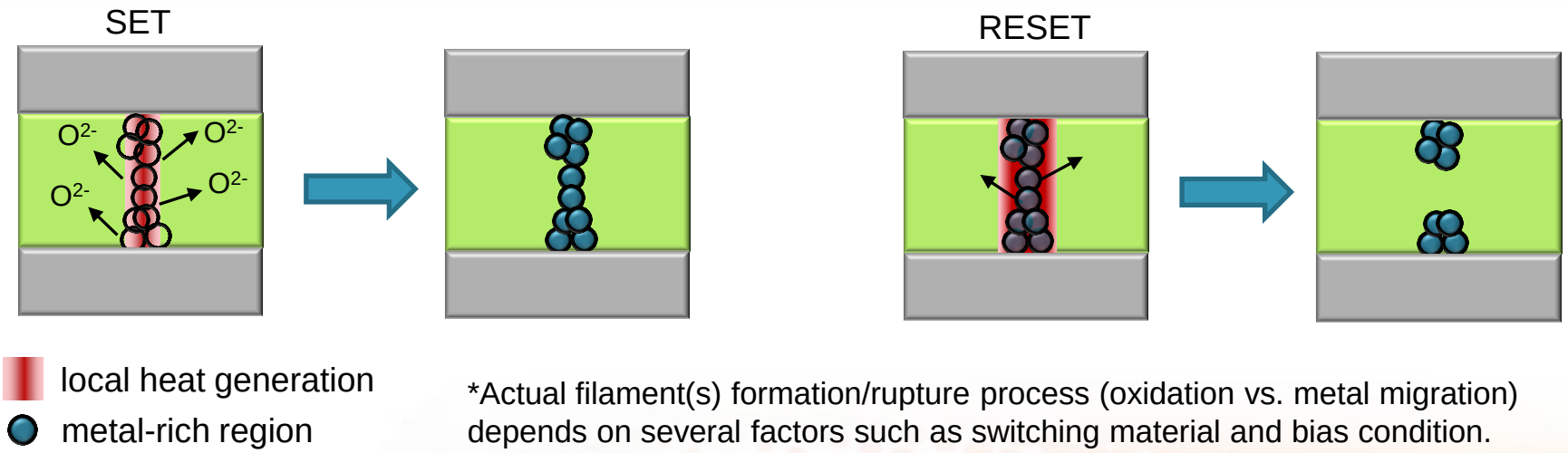
- Bipolar switching by *the migration of metal ions* under electric field
- Electrode – active metal (e.g. Ag, Cu,..)
- Various switching materials such as chalcogenide, amorphous silicon,…
- SET – anodic dissolution of active metal and formation of a filament(s)
- RESET – electrochemical dissolution of the filament(s)



*Actual filament(s) growth direction (e.g. BE → TE, TE → BE) depends on several factors such as metal ion mobility in the switching medium, ion trap density, leakage current density, and bias scheme

Thermochemical RRAM

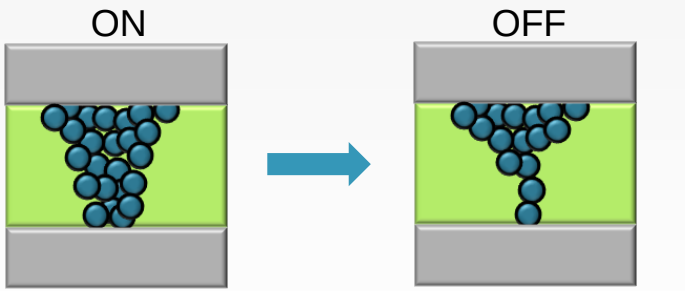
- Unipolar switching (fuse – antifuse) triggered by Joule heating
  - Local dielectric breakdown → heating → local structural modification (local redox reaction)
- Switching medium – some transition metal oxides (e.g. NiO)
- SET – local heating-induced V_o generation or electrode metal diffusion with current compliance
- RESET – thermal dissolution (rupture) of the filament with higher current (larger heating)



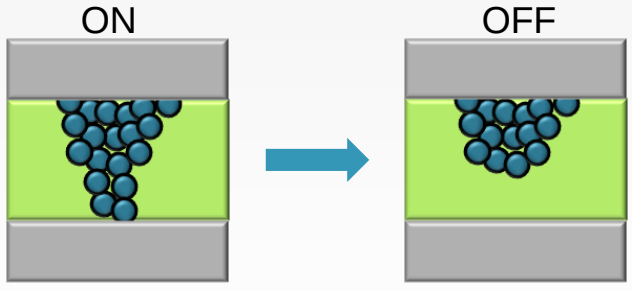
*Actual filament(s) formation/rupture process (oxidation vs. metal migration) depends on several factors such as switching material and bias condition.

Signature of 1D Filamentary Switching

- ON – area independent
- OFF – depends on switching materials and bias conditions

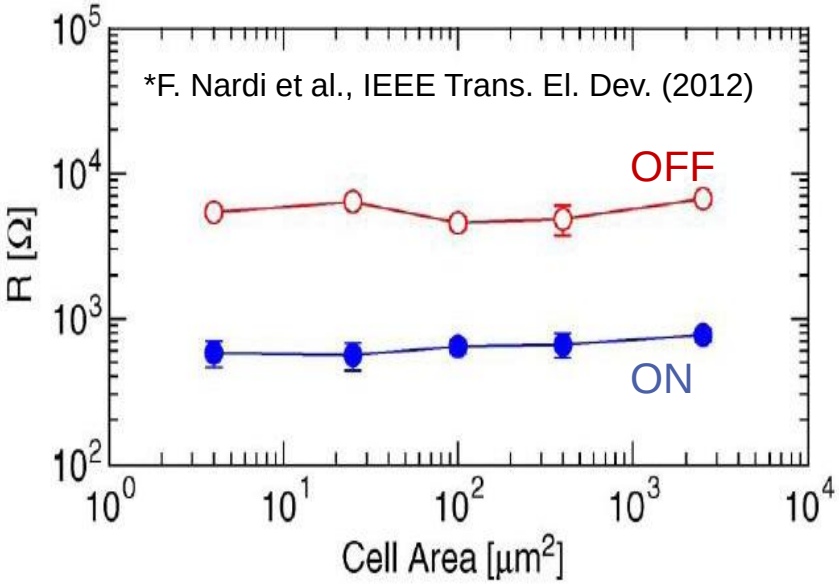


Quantum point contact

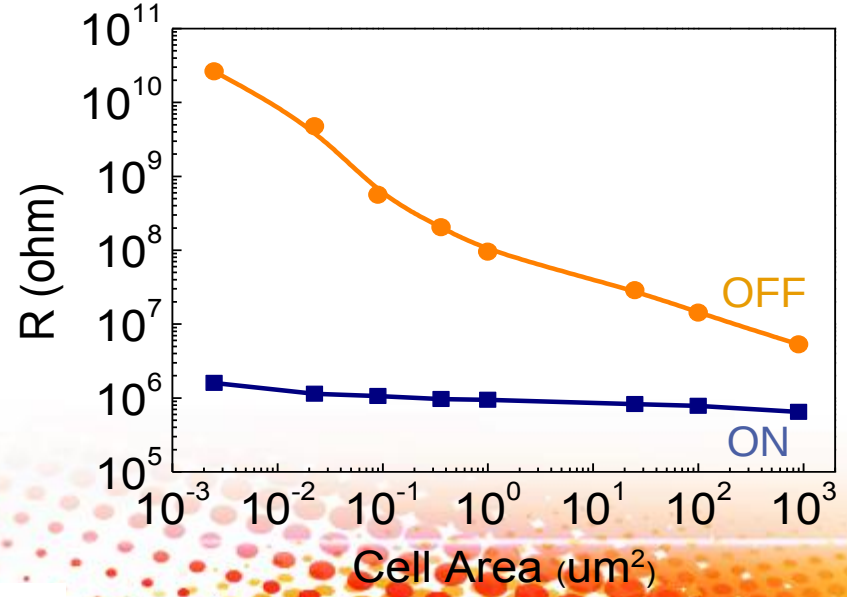


Tunneling/Schottky contact

Switching medium examples - HfOx, TaOx

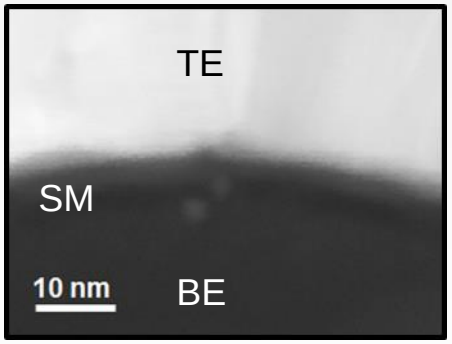


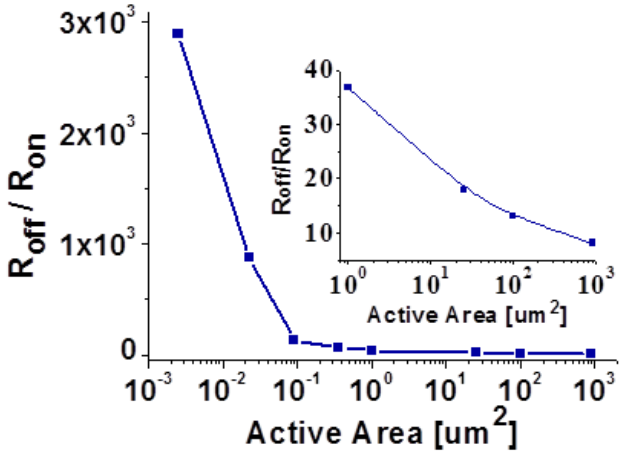
Switching medium examples – a-Si, TiOx



RRAM Scaling

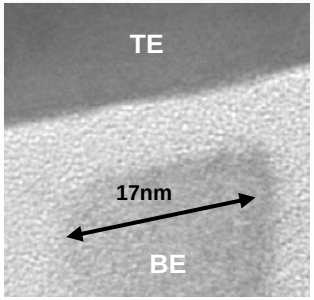
- Area independent conducting filament
- ON/OFF ratio improves as device size decreases
  - Higher sensing margin (faster read speed)
  - Larger array possible
- Sub 10nm scaling potential



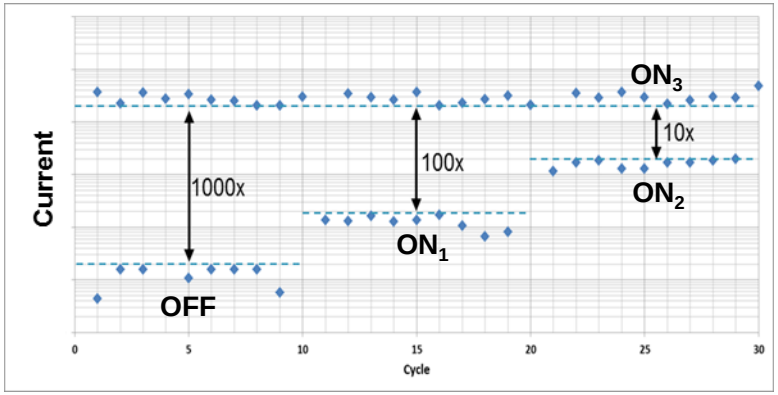


Sub-20nm Crossbar RRAM

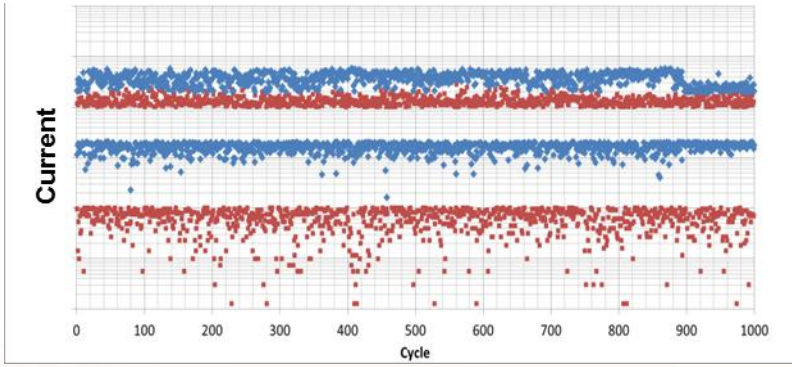
- Superior performance still maintained in sub-20nm devices
- Large ON/OFF ratio allows ≥ 2bits/cell on the same physical bit



Multi-level cell (MLC) Demo (sub-20nm device)

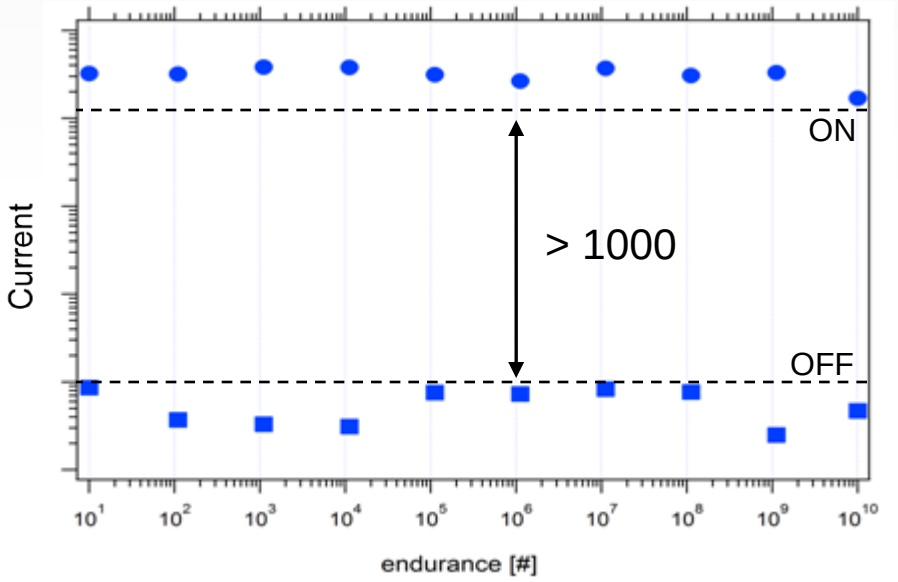


MLC Cycling (sub-20nm device)

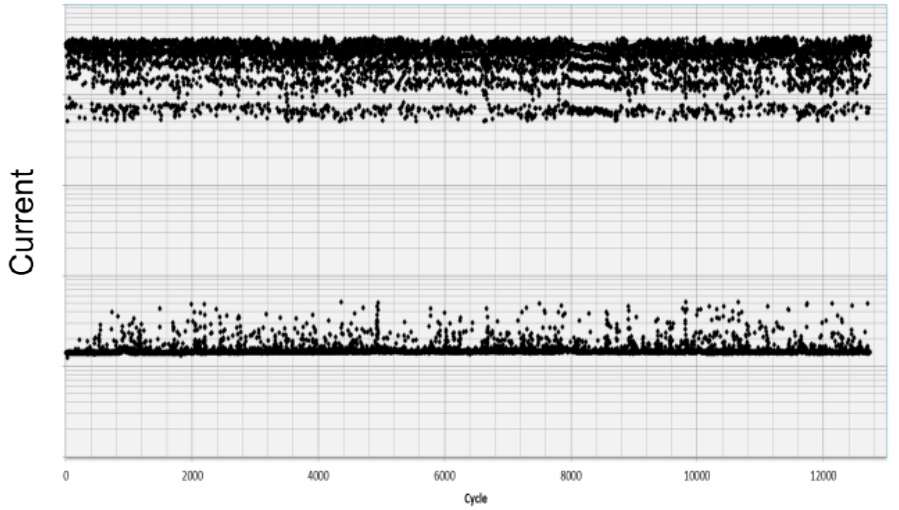


Large endurance > 10^{10} P/E cycles

- Crossbar cell has demonstrated endurance >10^{10} cycles
- ON/OFF ratio of >100X is maintained

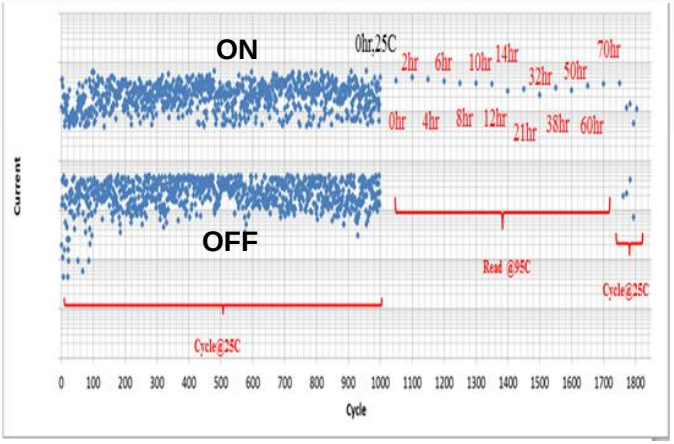


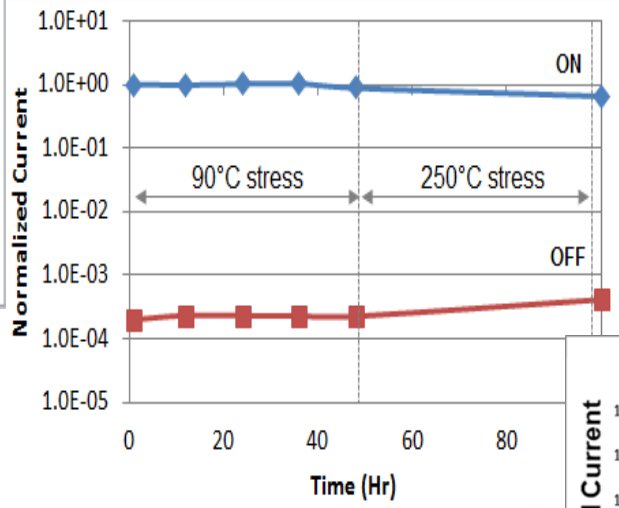
MLC cycling

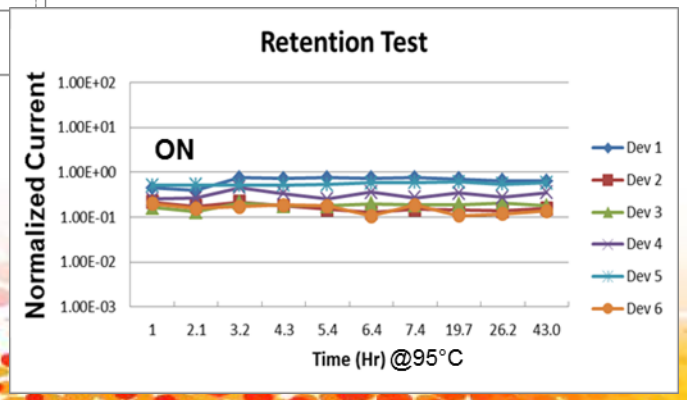


Retention > 10yr @85°C

- Large ON/OFF ratio maintained @85°C for 10yrs
- Multiple devices measured under the same conditions, show very similar retention characteristics

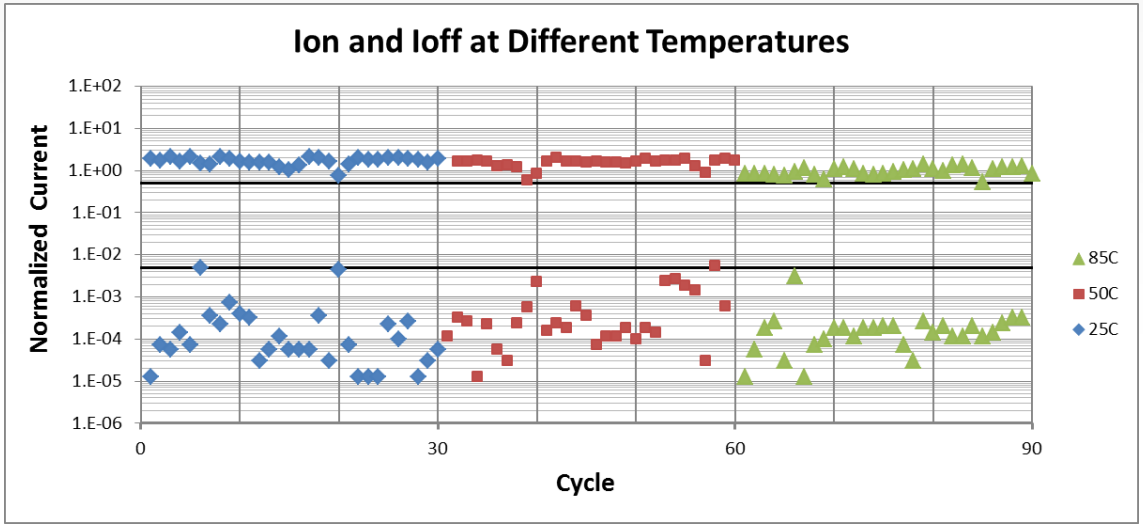


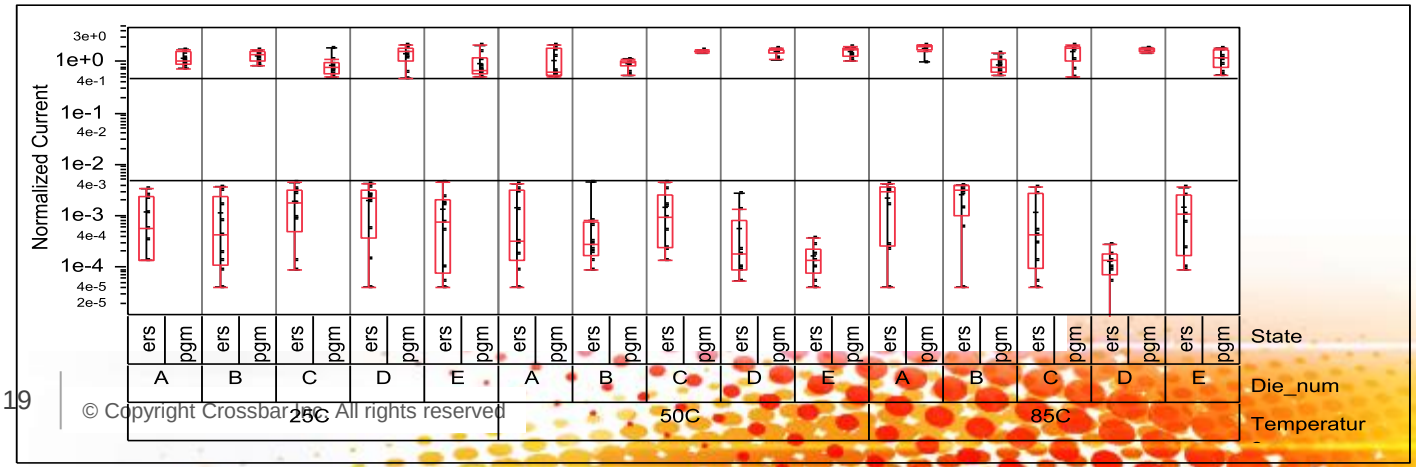




Good Thermal Stability

- Cycling parameters show no dependence on temperature dependence
- 100X ON/OFF ratio is maintained across the whole temperature range

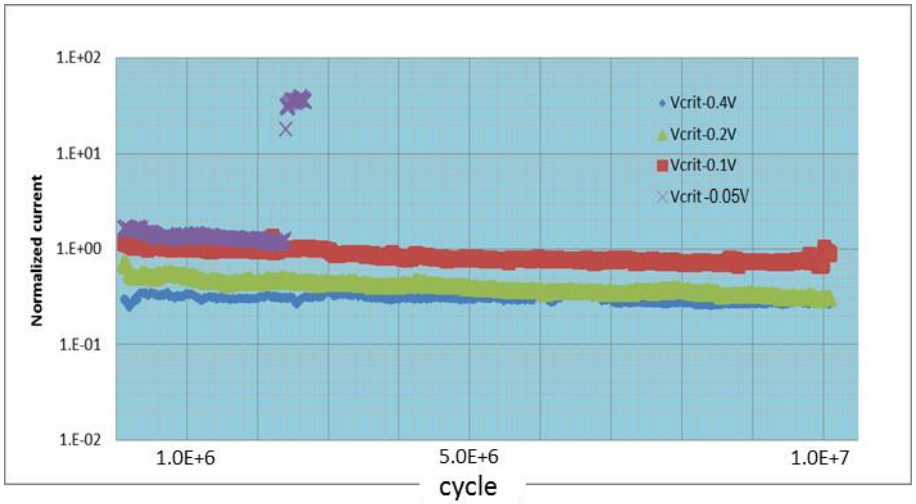




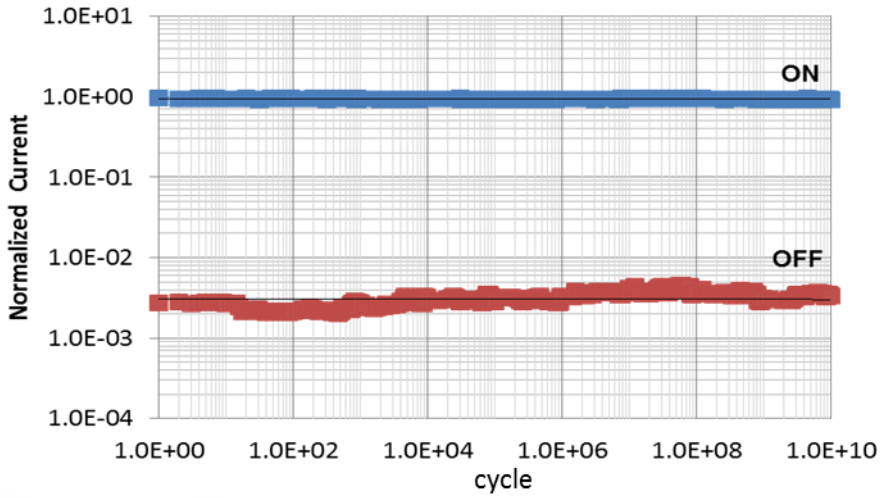
Immune to Program and Read Disturb

- No program disturb observed at voltages lower than the programming voltage
- No change in either the program state or the erased state after >10B read cycles
- Immunity to read disturb is maintained at 85°C

Program disturb test



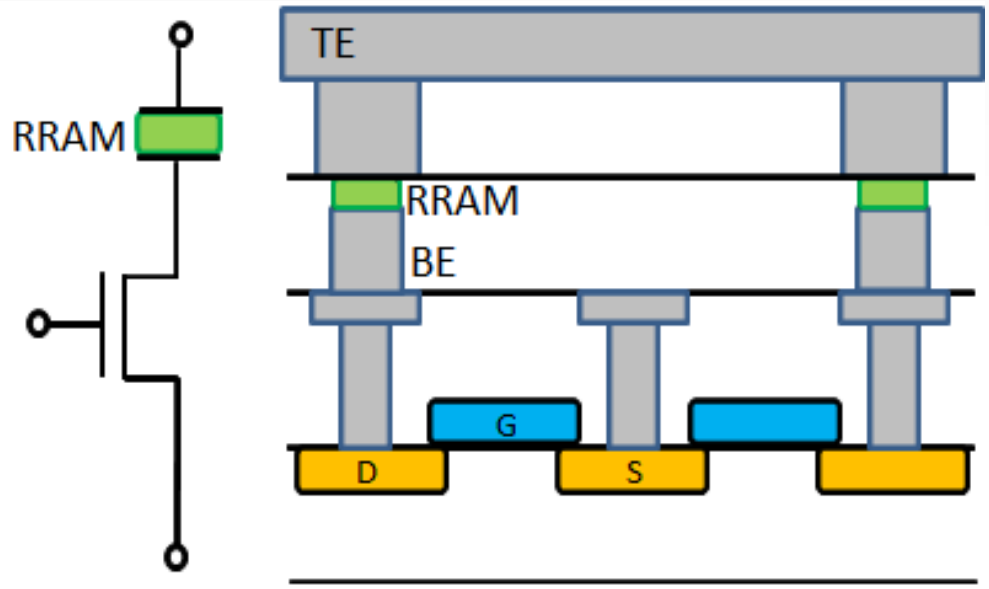
Read disturb test



RRAM Integration – *1* Transistor per *1* RRAM Cell

1T1R

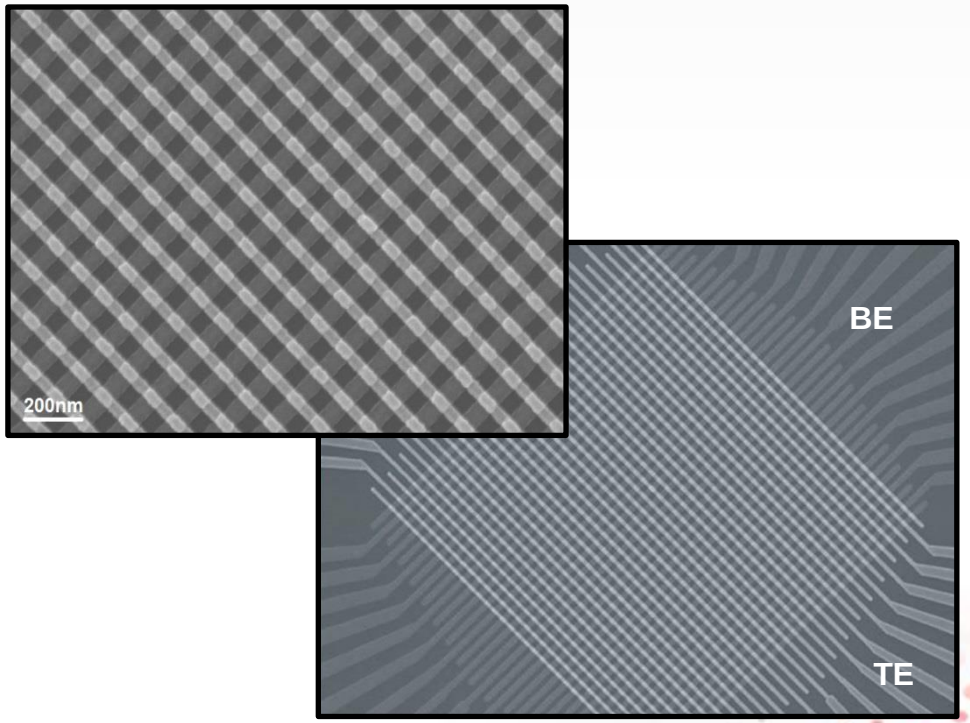
- For high performance (e.g. speed)

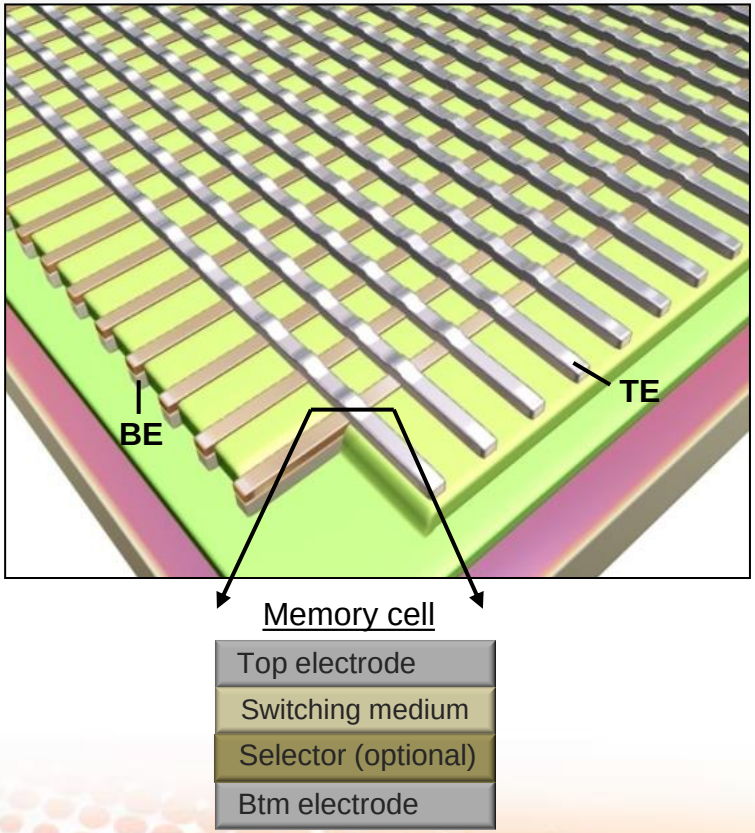


RRAM Integration - *1* Transistor per *n* RRAM Cells

Crossbar (1TnR)

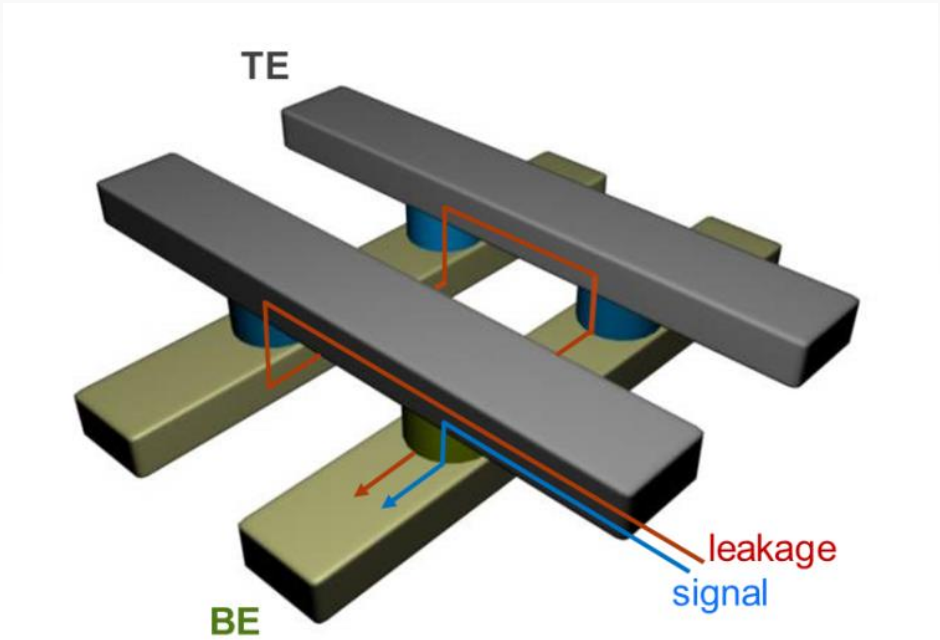
- For high density

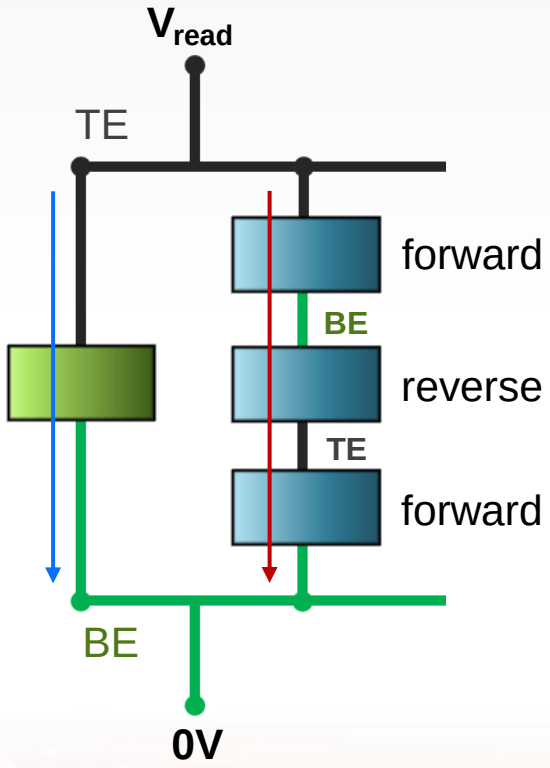




Crossbar Architecture – Leakage Current Control

- Reducing leakage current by
  - Non-linear IV (increased R in small bias)
  - Rectifying IV (increased R in reverse bias)

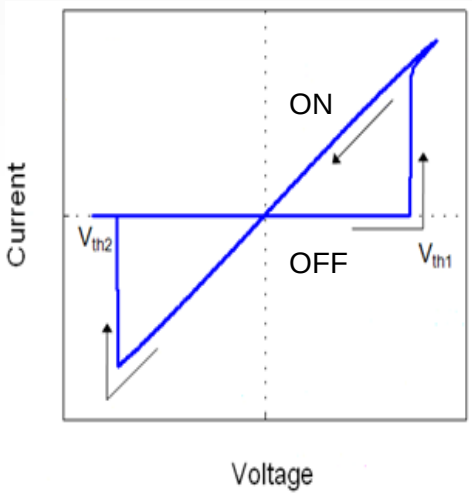




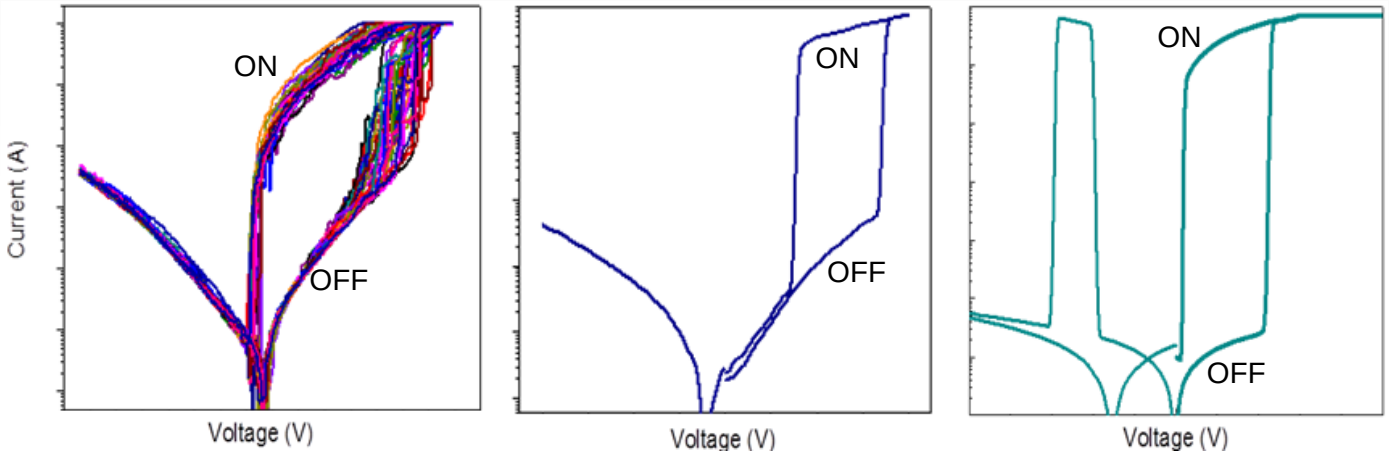
Switching Behavior Modulation

- Both non-linear IV and rectifying switching obtained by switching medium optimization and process control
  - With still high on/off ratio of 10^3 ~ 10^6

Focused on speed

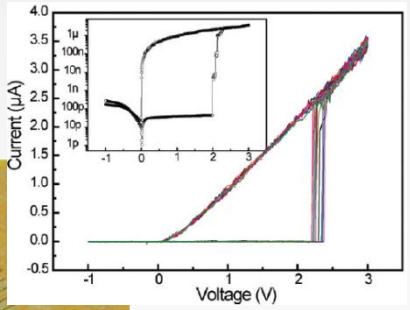


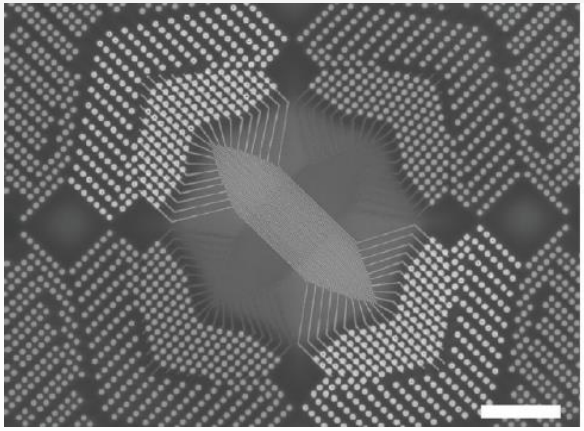
Focused on density & low power

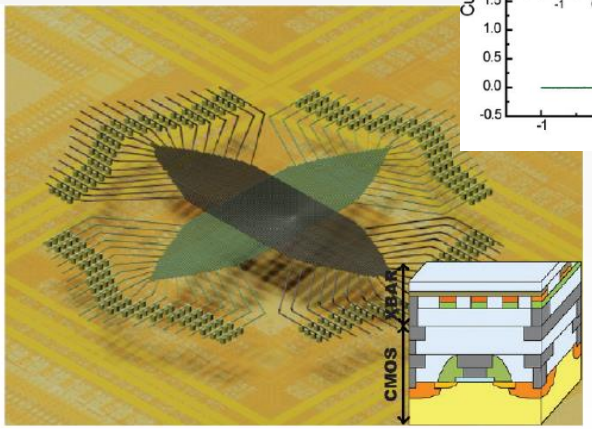


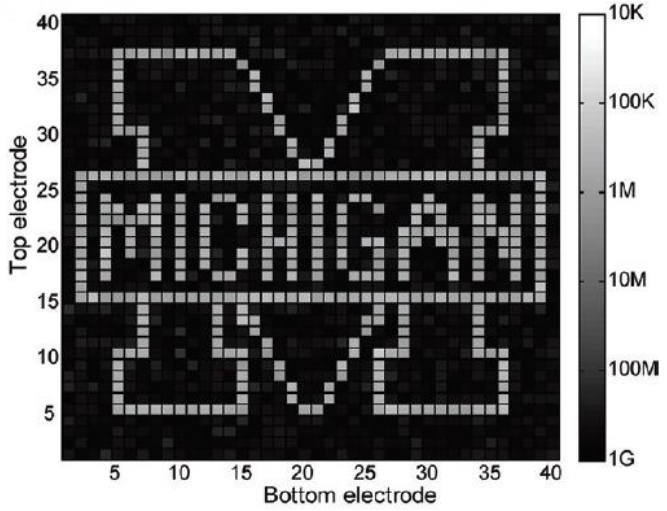
*IV curves obtained from different devices which are designed for different product requirements

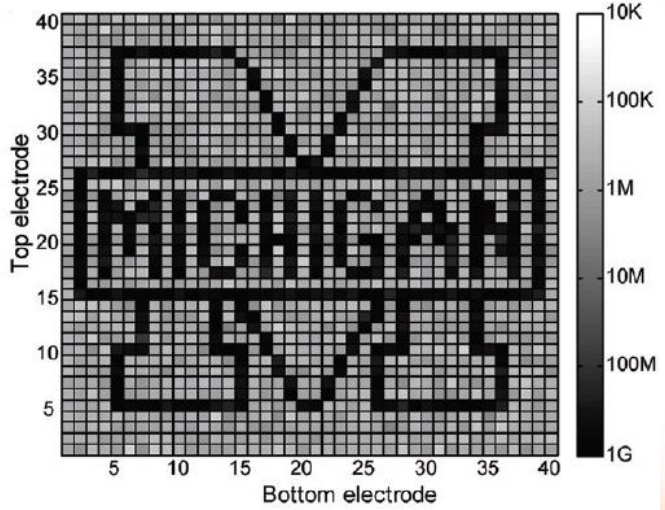
Information Storage in Passive Crossbar







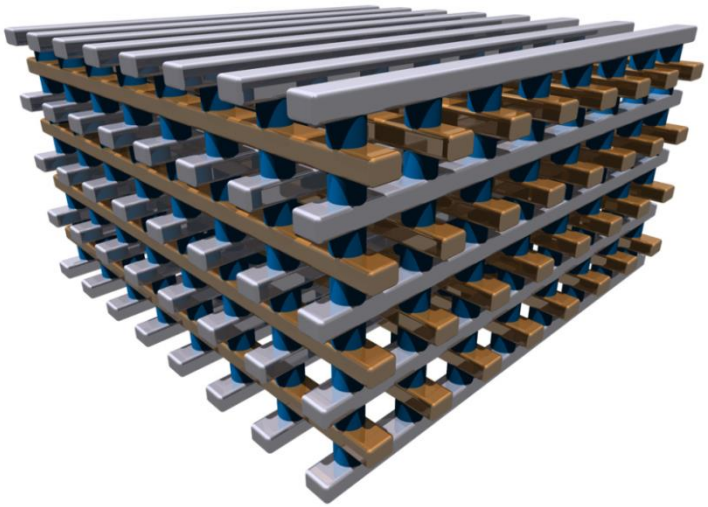


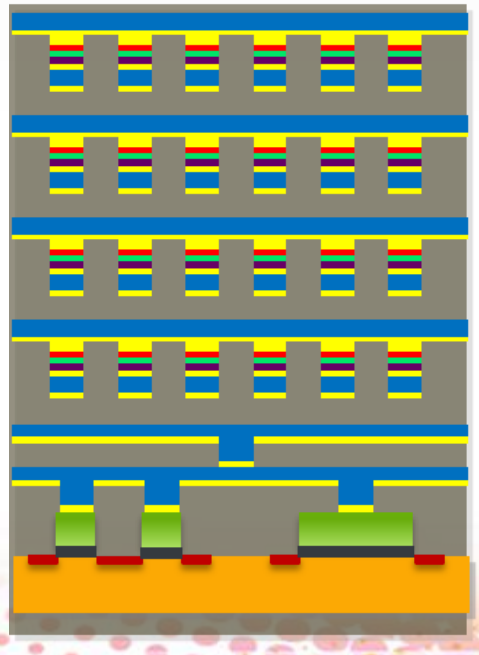


Lu et al., Nano Lett. (2012)

Stackable 3D Memory Array

- Simple materials and structure
- Low temperature fabrication process
- Easy integration with standard CMOS logic
- → 3D stackable memory architecture





Crossbar RRAM memory layers

CMOS logic

RRAM for Neuromorphic System

Modern Computer System – Complex & Inefficient

- Computer systems consume several orders of magnitude higher energy than the animal's brain for complex (multiple inputs) tasks

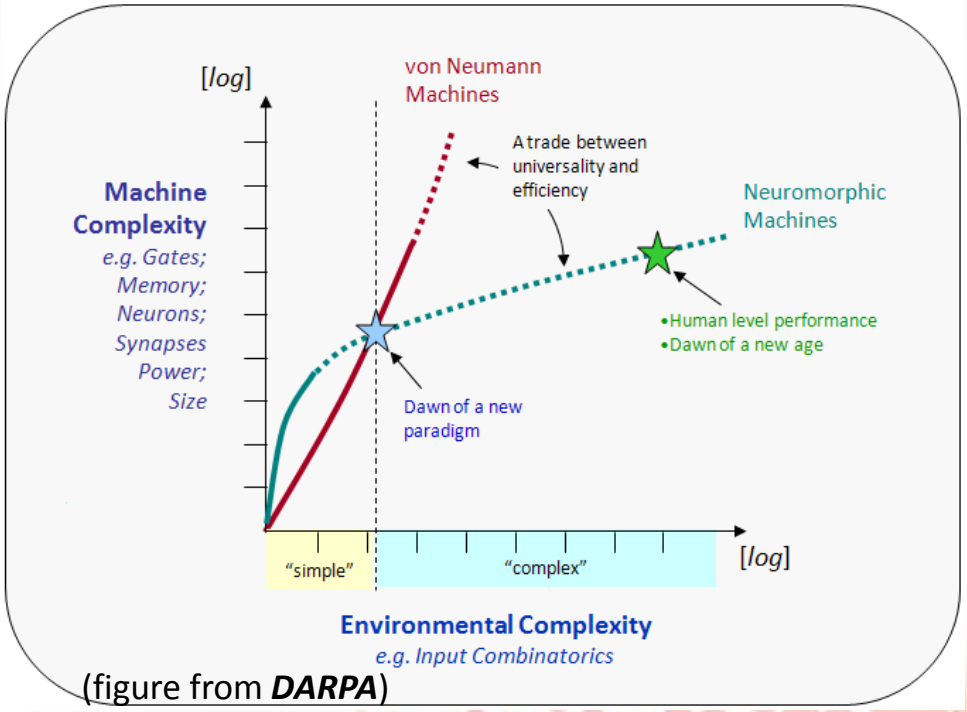
(http://www.photocat.co.uk)

IBM Blue Gene/P supercomputer
Capable of cat's cortical level simulation at 83 times slower than the real

(http://en.wikipedia.org/wiki/Blue_Gene)

Highly Parallel Computing for Improved Efficiency

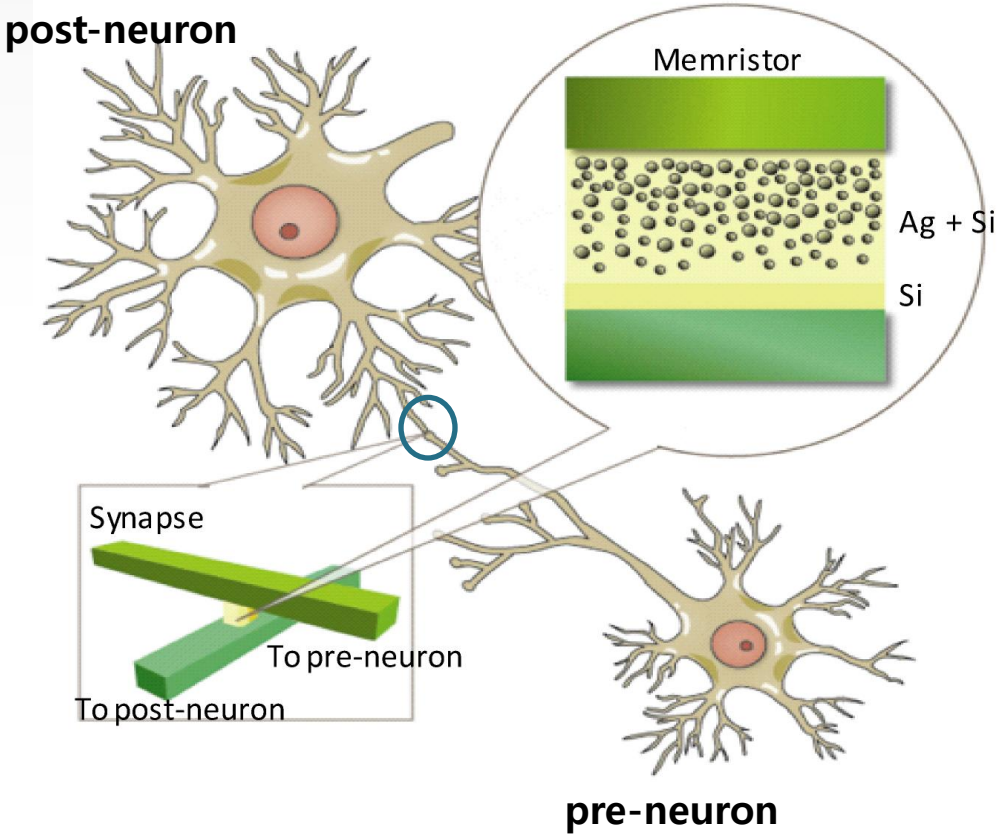
- *Sequential processing nature of computers* (inefficient and complex system architecture) ↔ *Highly parallel nature of the neural system* (highly efficient system)
- Key to the high efficiency of bio-systems is the large connectivity between neurons

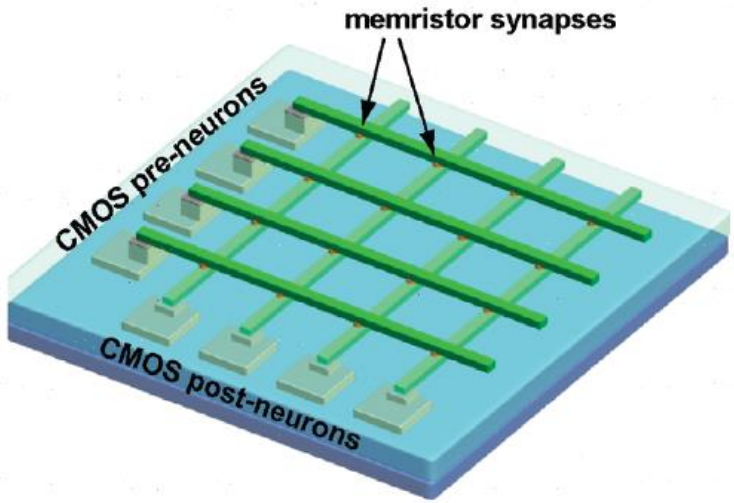


(figure from ***DARPA***)

RRAM Synapse for Neuromorphic System

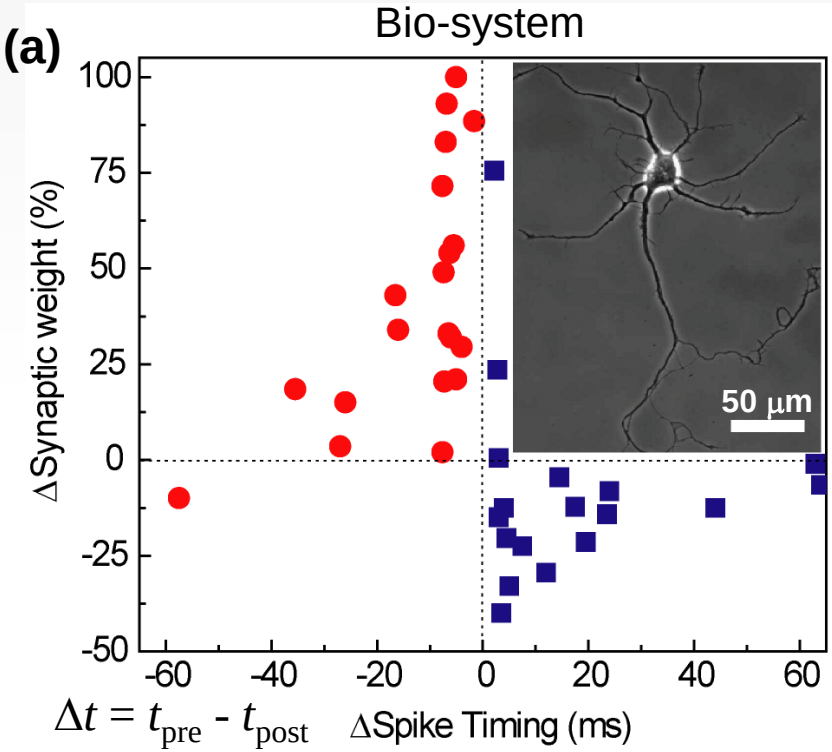
- CMOS neurons + RRAM synapses in a neuromorphic system
- Crossbar structure for the neural network

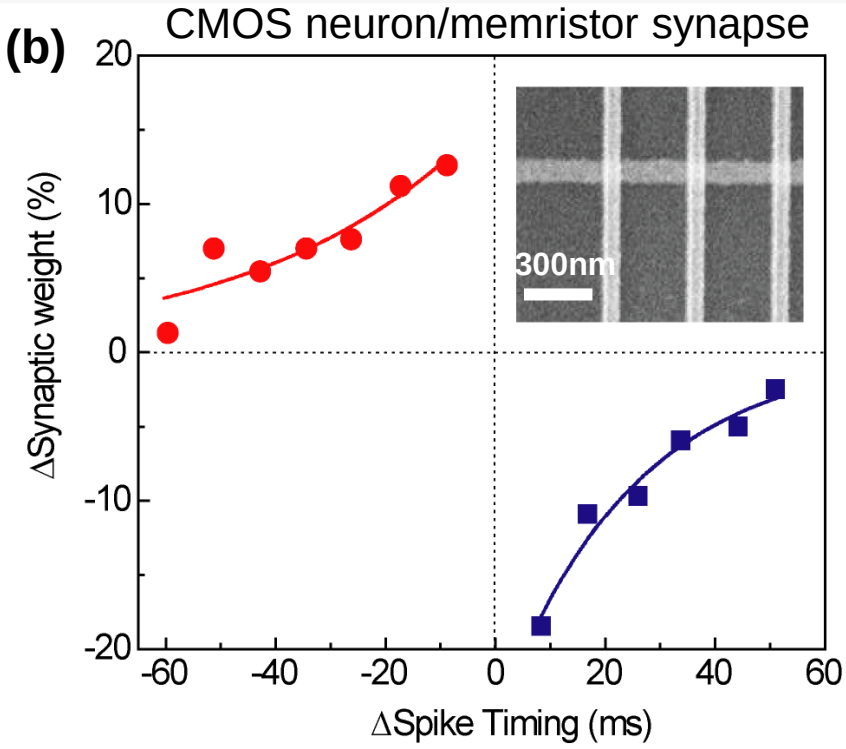




Synaptic Function Demonstration by RRAM Synapse

- STDP (Spike Timing Dependent Plasticity) implemented by a hybrid CMOS neuron/memristive device (RRAM) synapse system





- support important synaptic functions
- frame work for neuromorphic systems

Jo et al., Nano Lett. (2010)

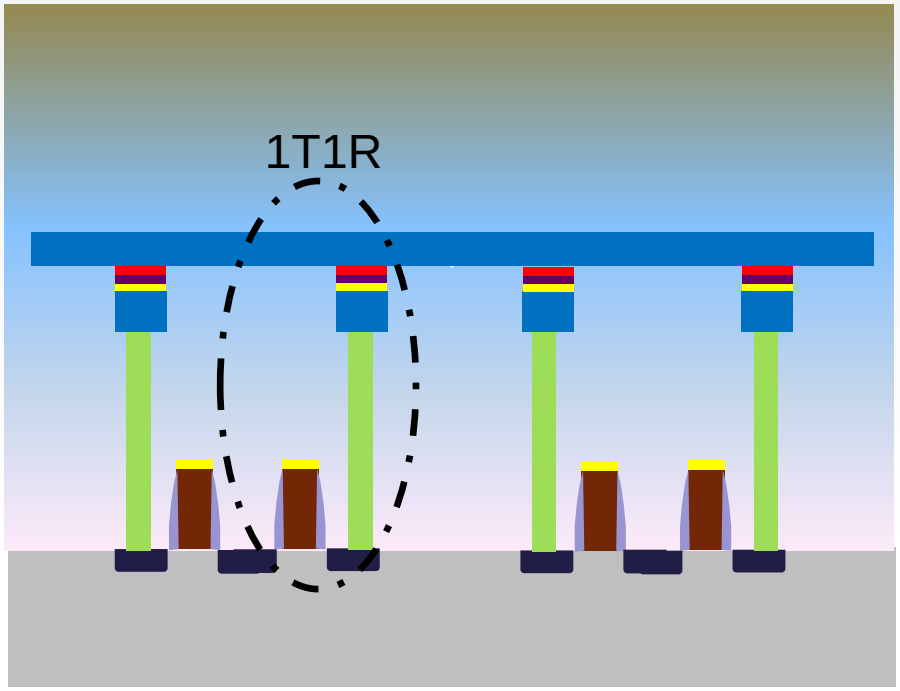
Product specifications
Data, Code, and Embedded

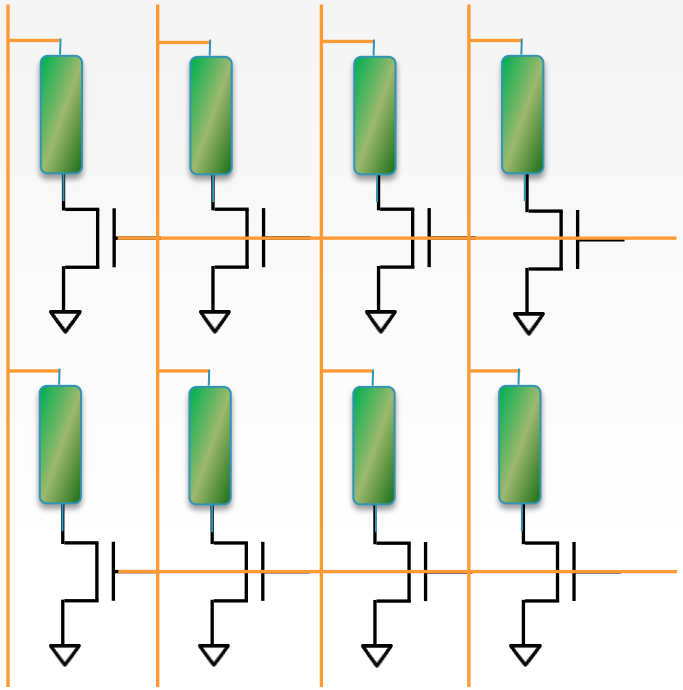
Crossbar Offers Compelling Technical Advantages

	Embedded		Code Storage		Data Storage	
Applications	**MCU Config bit FPGA** 		**Printers, Small density OTP, MTP, SPI, STB, DTV, Phones, Large density** 		**Solid State Drive SLC/MLC Server Memory** 	
	eFLASH	**Crossbar™**	**CODE FLASH**	**Crossbar™ CODE**	**Data Storage NAND Flash**	**Crossbar**
Density	**256K-4Mbit**	**256K-16Mbit**	**512K-8G**	**512K-64Gb**	128Gbit	256Gbit
Technology	**90nm**	**<10nm**	**45nm**	**<10nm**	20nm	<10nm
Cell Size	$18F^2$ - $42F^2$	$5.4F^2$ - $18F^2$	$6\text{-}12F^2$	$5.4F^2$	$5.4F^2$	$4.5F^2$
Program byte Program page	10us -	2us -	10us - 300us 700us -1.4ms	2us 256us	Not Capable 1.2ms	**2us** 16us
Erase byte Erase page Erase block	Not Capable Not Capable 25ms	2us 256us 4ms	**Not Capable Not Capable** 25ms-60ms	**2us 256us** 4ms	**Not Capable Not Capable** 3ms(2MB)	**2us 67us(8KB)** 2ms(2MB)
Read Latency	30ns-100ns	30ns-100ns	100ns	100ns	**50us**	**1us**
Endurance Retention	1 million 10Yr@125C	1 Million 10Yr@125C	100K 20Yr@55C	100K 20Yr@55C	**<1K 1yr @40C**	**10K 10yr @40C**

Design & Architectural Attributes

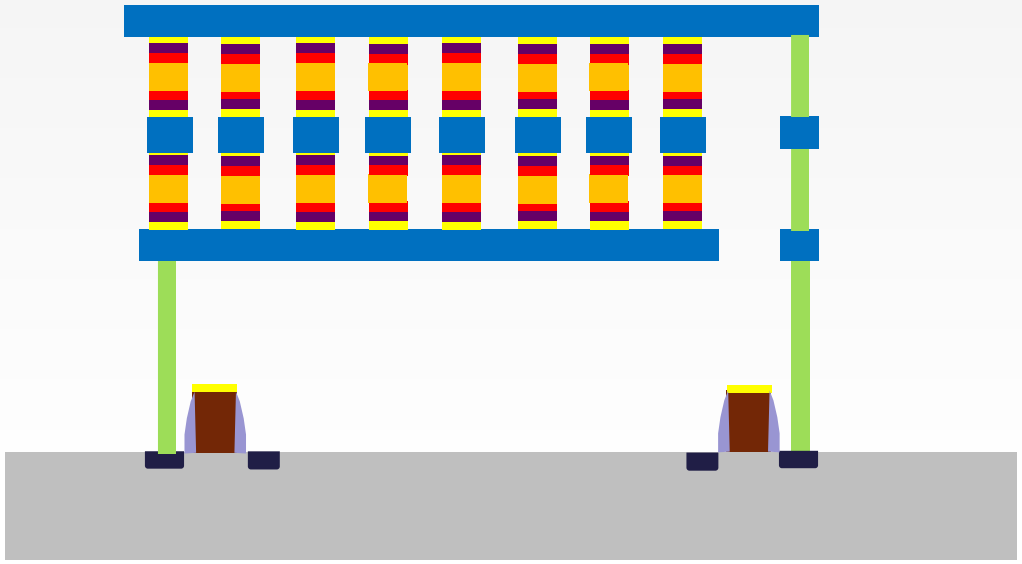
RRAM Design Suited for Embedded Memory

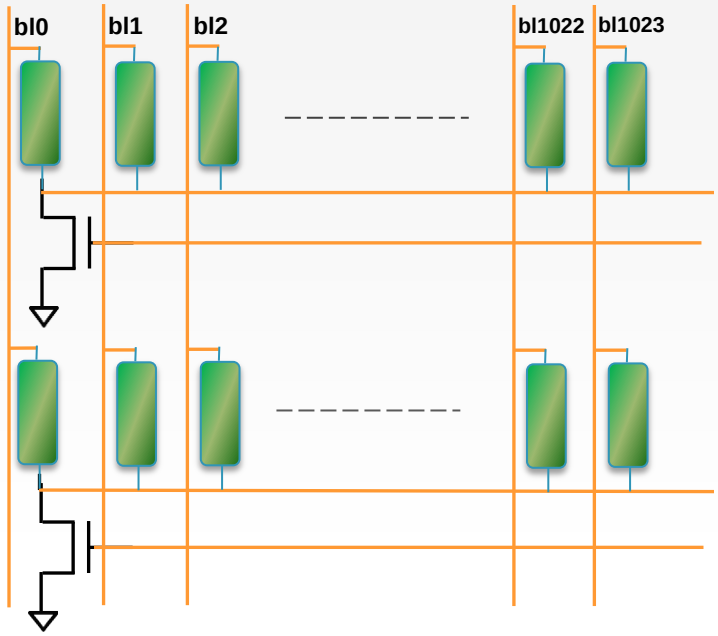




- Suited for high speed embedded memory operation
- Backend process integration. Easier to integrate and less expensive than eFlash

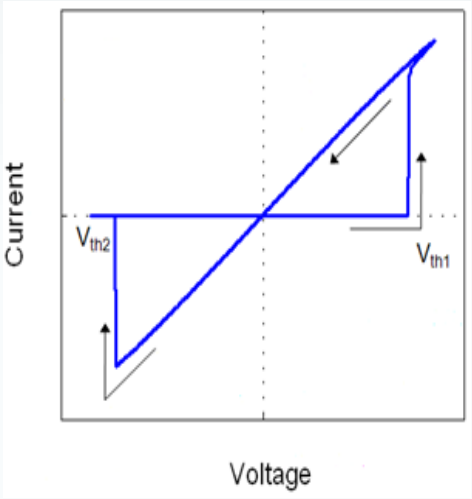
Crossbar array architecture
Suited for high density memory NOR/NAND



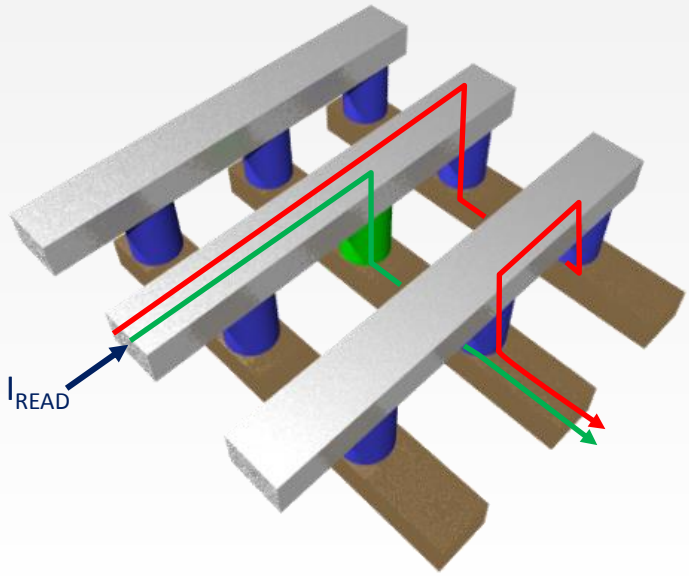


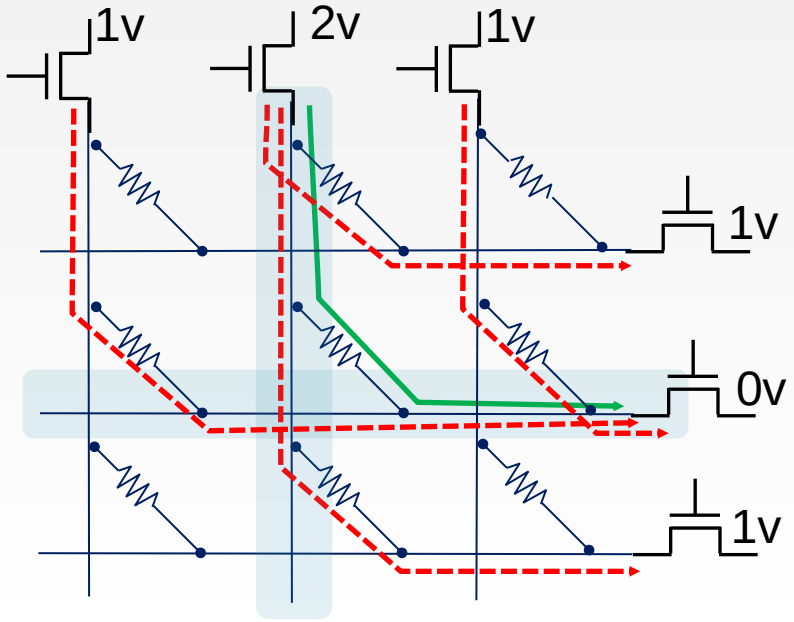
- One transistor selects many RRAMs
- Stackable architecture - Effective cell $4F^2/L$ - L is the number of stacks - $1F^2$ with 4 stacks
- The transistor sizes is not the cell size limiter – No need to down scale the transistors
- Area under the array could be utilized for peripheral circuits – Provides high array efficiency
- Competitive with NOR, NAND, and next generation 3D NAND architectures
- Backend process Integration

RRAM array with linear resistance characteristics or without select device

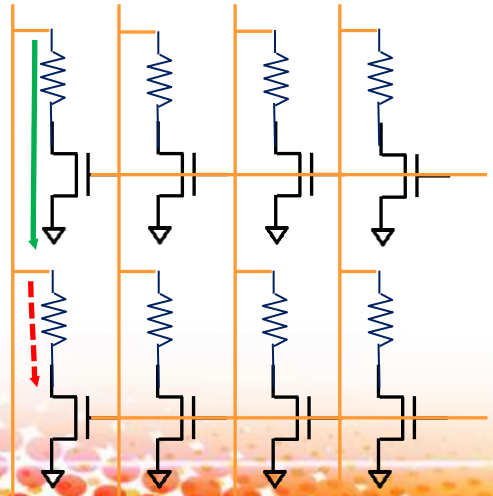


Linear resistance RRAM in a cross-point array

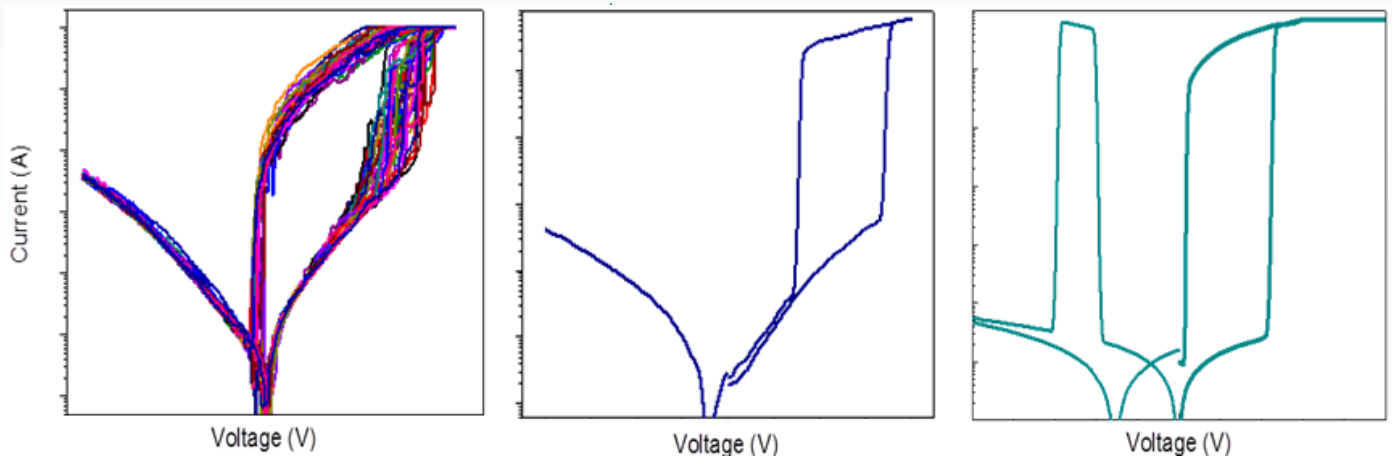




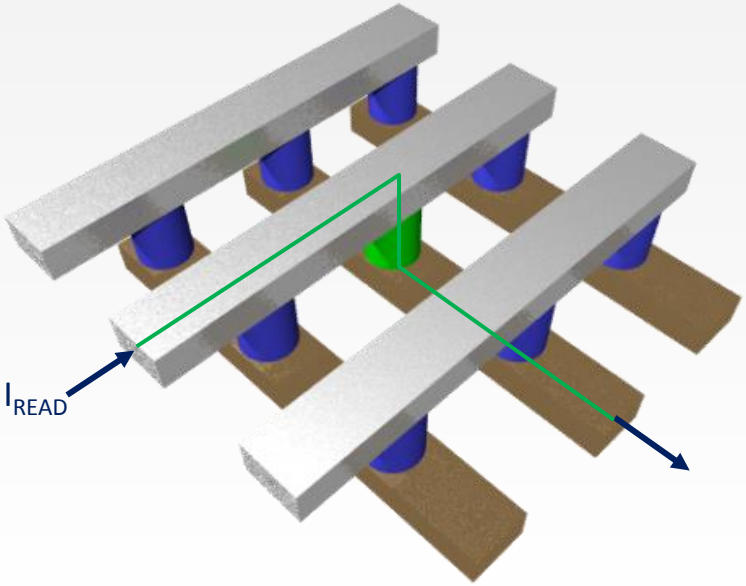
- Making cross-point 1TnR arrays with linear resistance RRAM cells generates sneak paths (dotted red lines) significantly reducing sensing margin, Increasing power, and limiting sector size
- Biasing is very challenging - Any small potential difference between unselected BL & WL generates will generate very large current consumption
- Therefore, Linear resistance RRAM utilize 1T1R architecture

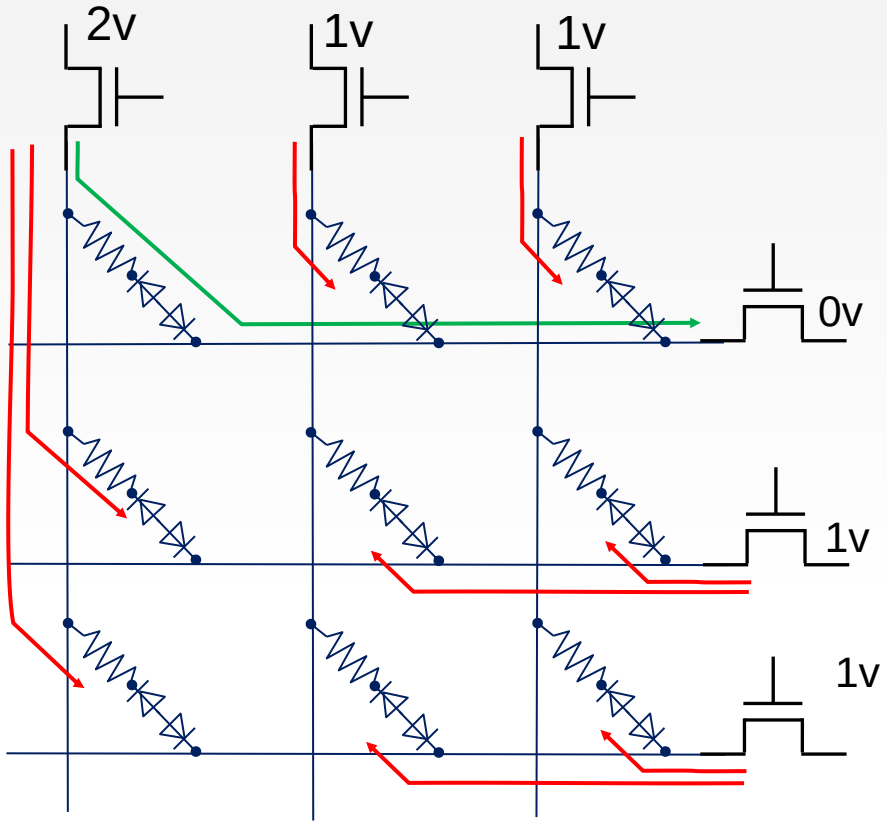


RRAM array with Non-Linear Hysteric IV



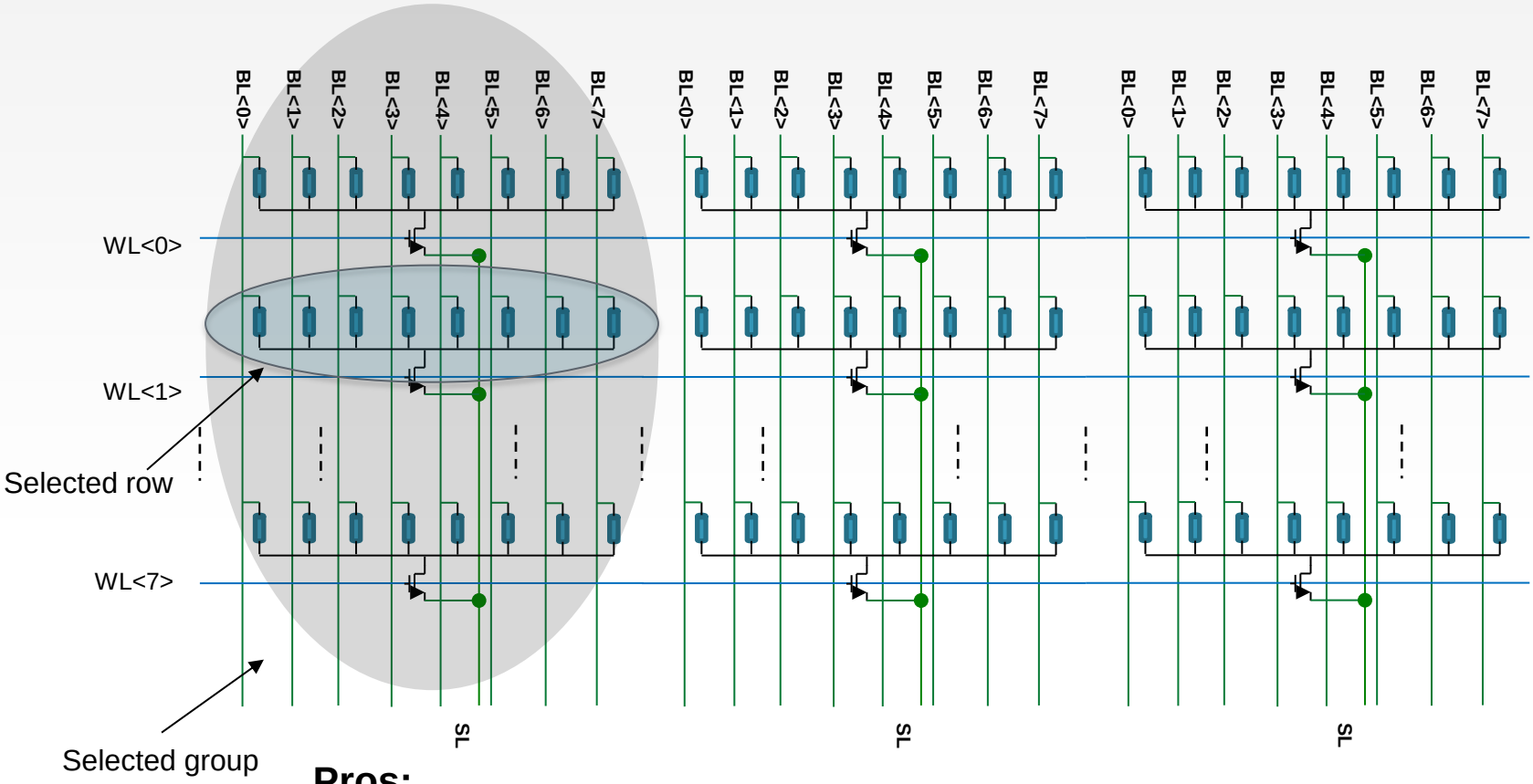
Non-linear RRAM in a cross-point array





- RRAM with nonlinear complementary barrier characteristics will:
  - Mitigate the sneak path problem.
  - Yield larger arrays and larger sensing margins, higher programming throughput, and larger array efficiency

Word based Crossbar RRAM Array – Power optimized



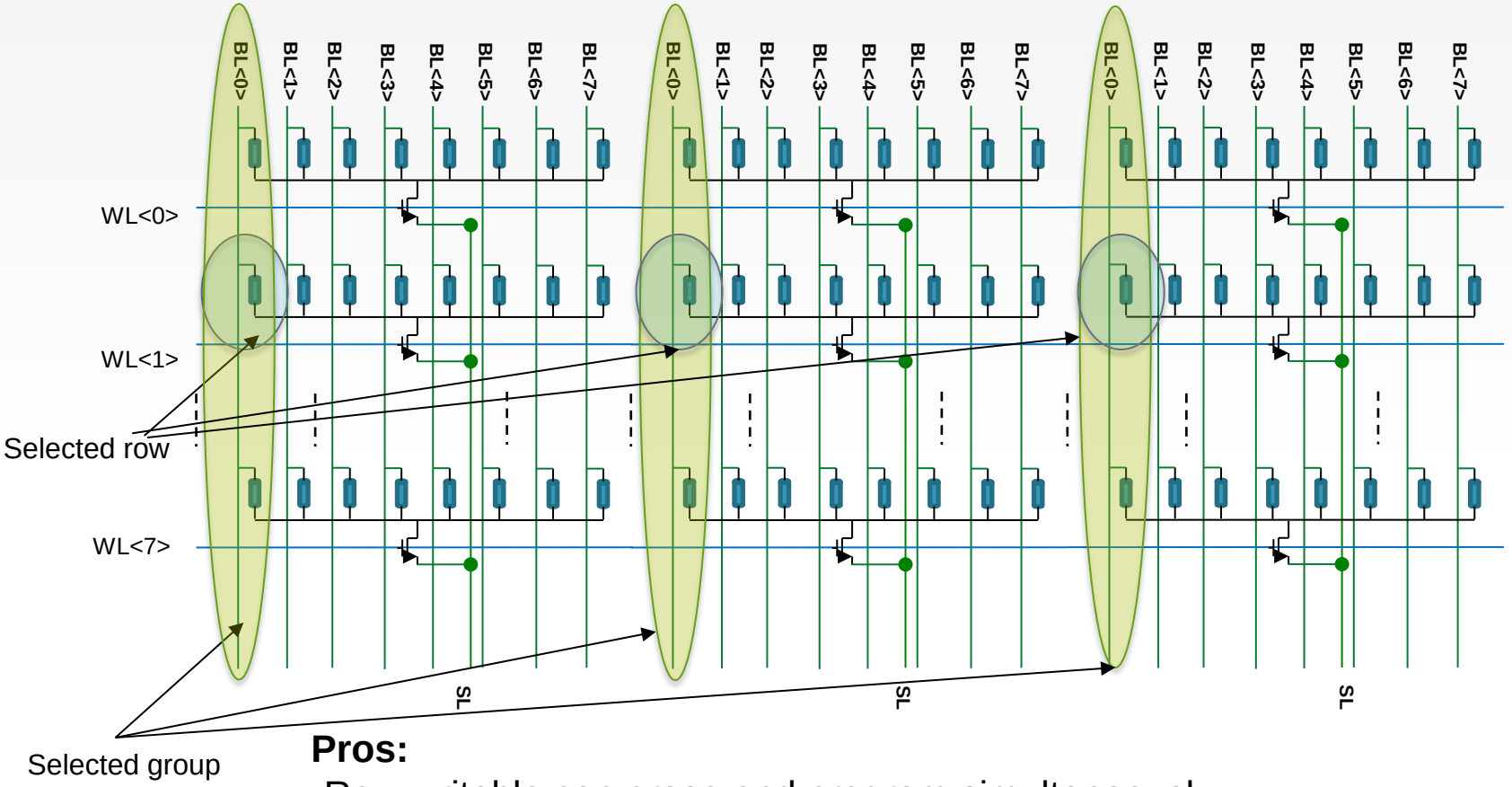
Pros:

- Row Alterable for program and erase
- Lower power consumptions – precharge/activate one bank for a byte
- Potentially better immunity to disturb conditions

Cons:

- Slower sensing and pattern sensitive

I/O based Crossbar RRAM Array architecture – Write/ Read Speed optimized



Pros:

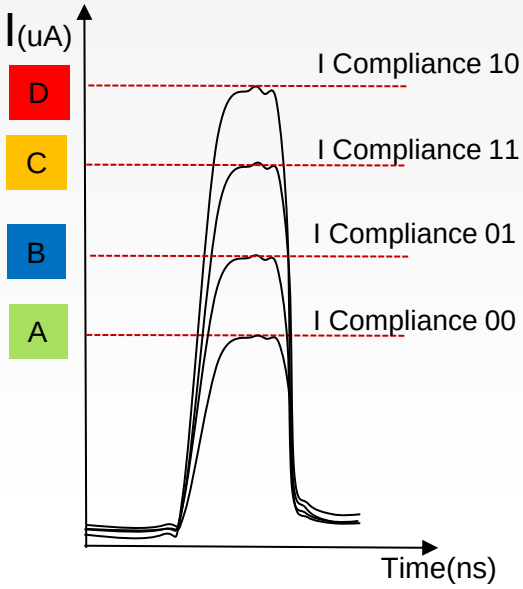
-Row writable can erase and program simultaneously

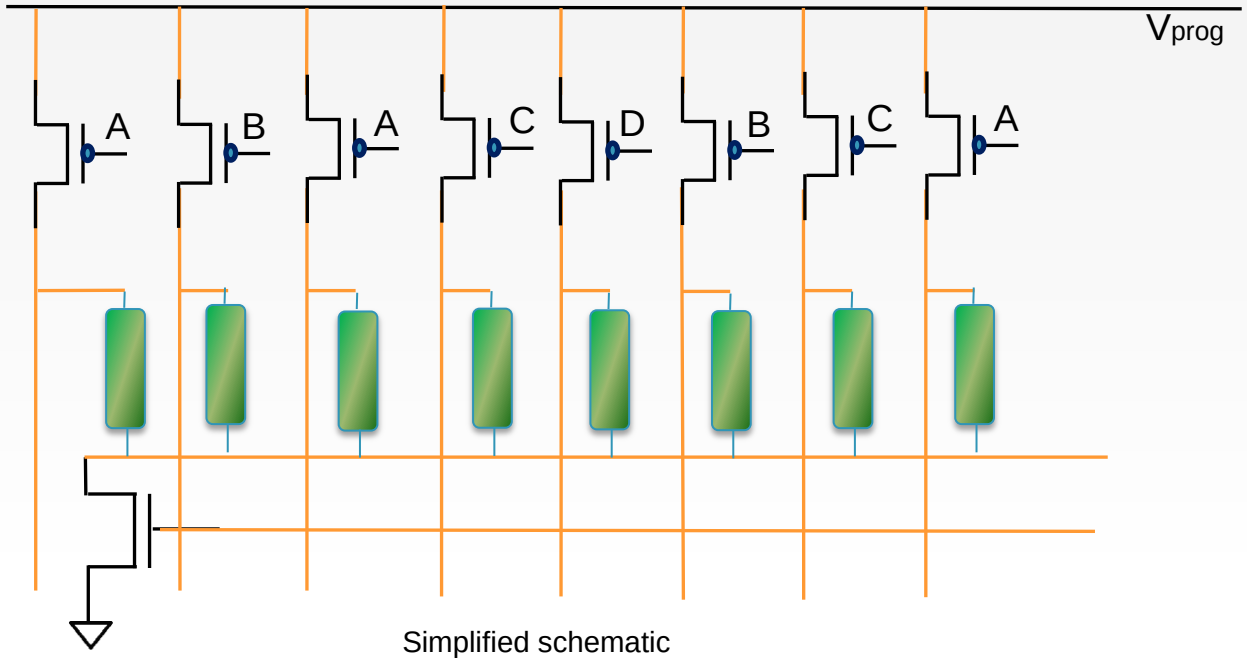
-Faster sensing speed and less pattern sensitive

Cons:

- Higher power consumption – Precharges 8 banks for a byte

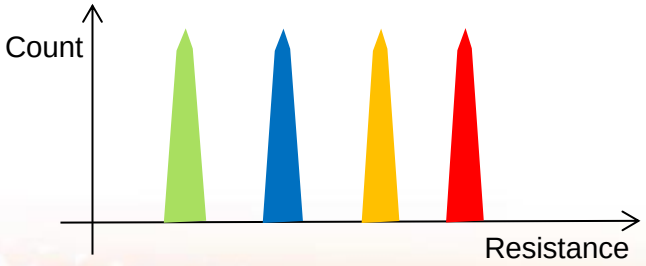
MLC 2bit Programming – With Current compliance





Simplified schematic

- Each RRAM cell is MLC programmed into different resistance values by limiting the current flowing in each cell during program operation

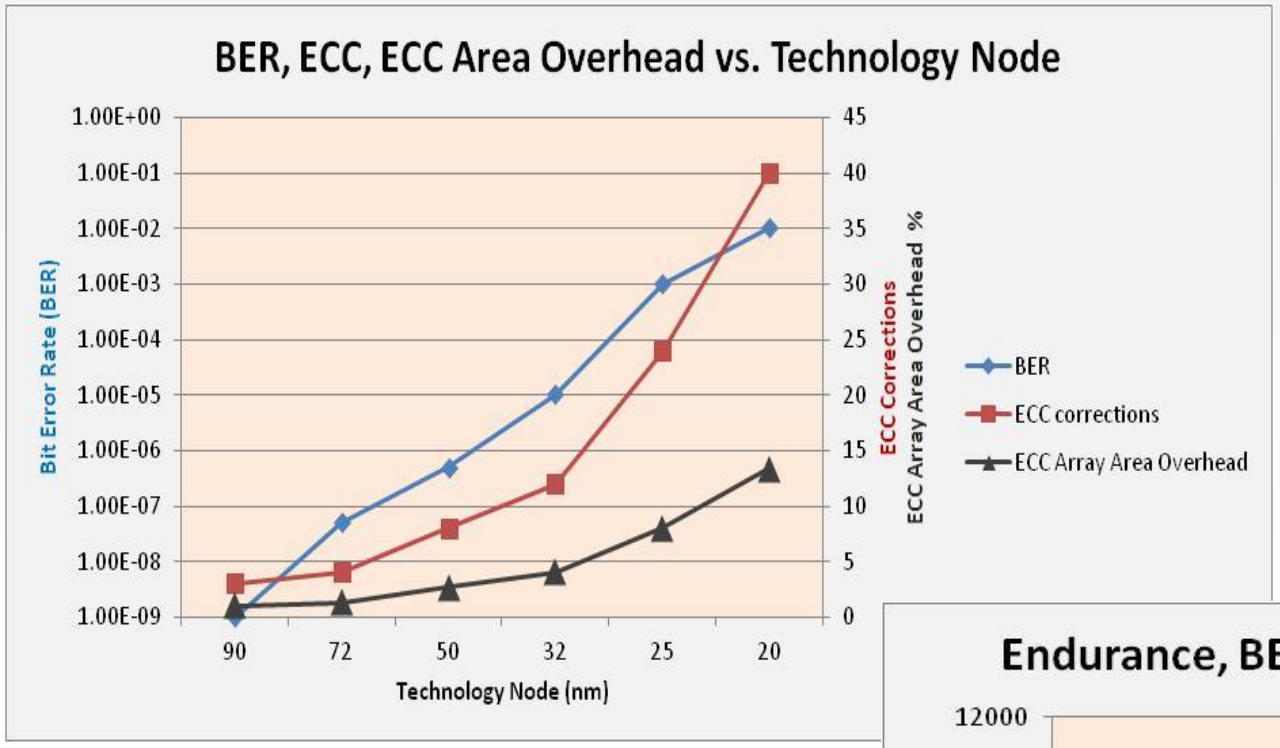


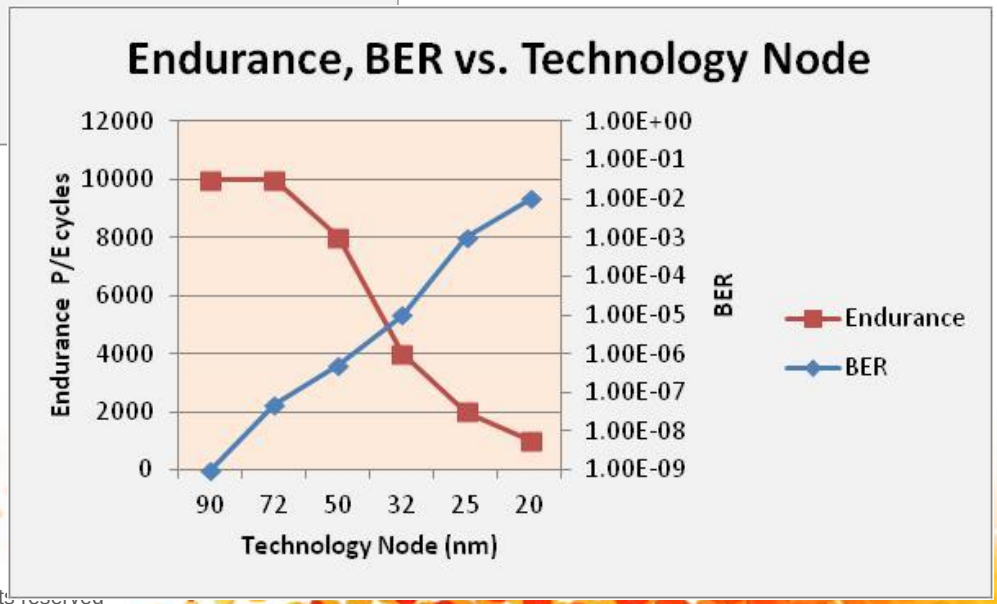
Crossbar RRAM and Its Impact on System Performance

NAND Characteristics, Impact, Remedies, and Trade off

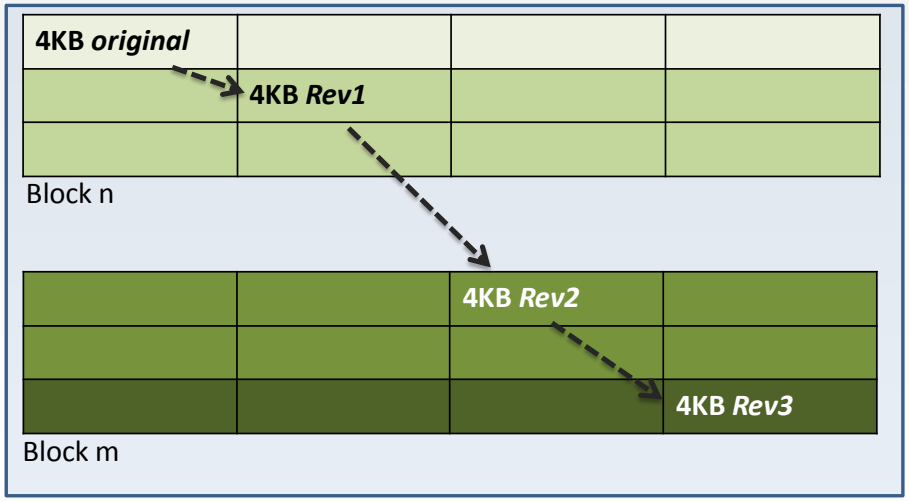
NAND Characteristics	Impact to Storage System	Improved by	Trade off
Low Retention & high BER	Reduces lifetime	ECC (BCH, LDPC)	Controller Overhead & Cost Power consumption
Low P/E Cycles	Reduces lifetime	Wear Leveling	Performance & Controller Overhead & Cost
No ReWrite feature No page alterable No Page erase	Write amplification	Garbage Collection	Performance & Controller Overhead & Cost
Slow page read	Random Read Performance & Latency	None	Performance

Present NAND FLASH Technology Trends



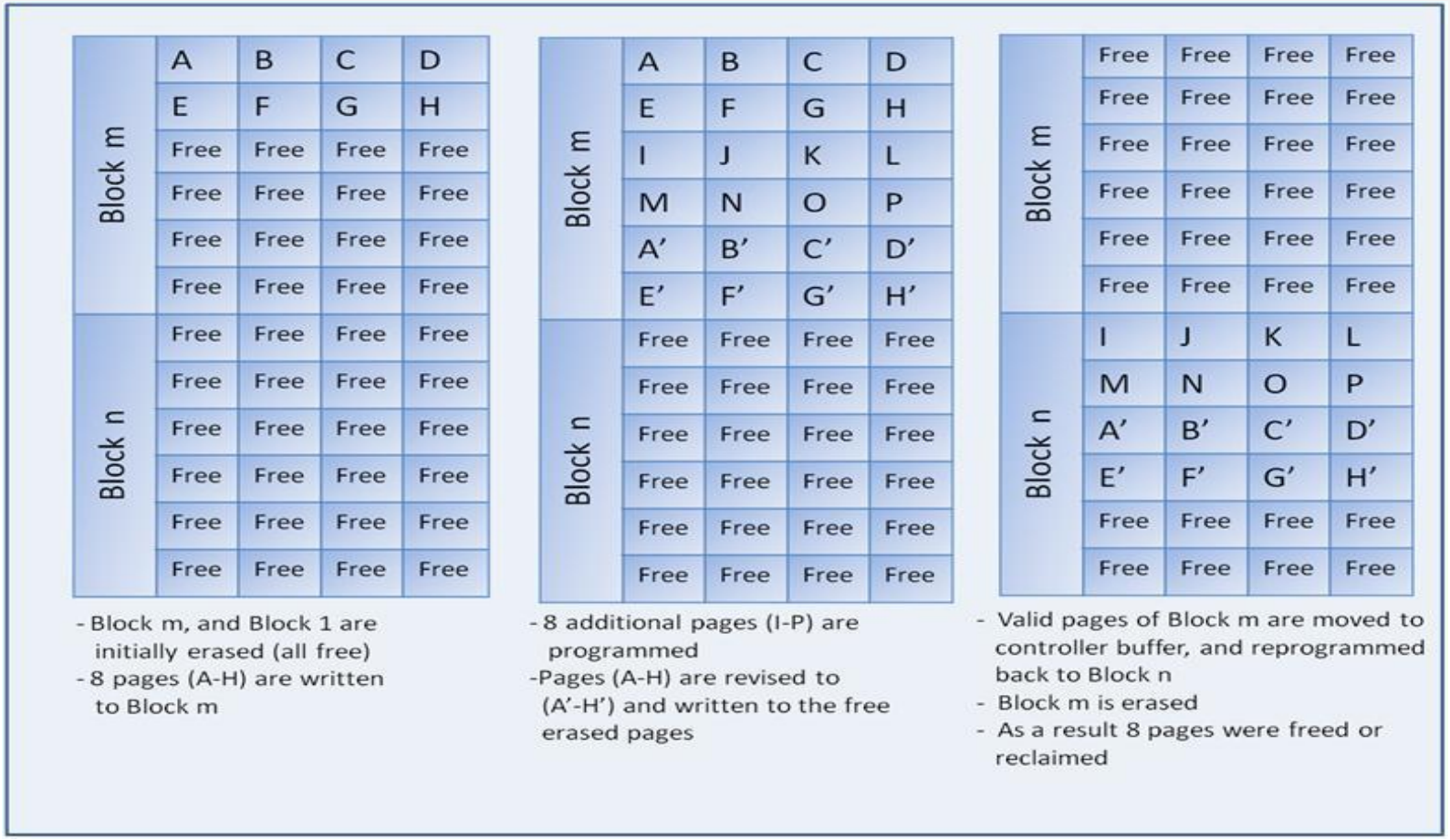


NAND's Re-Write limitation, and Data Revision process



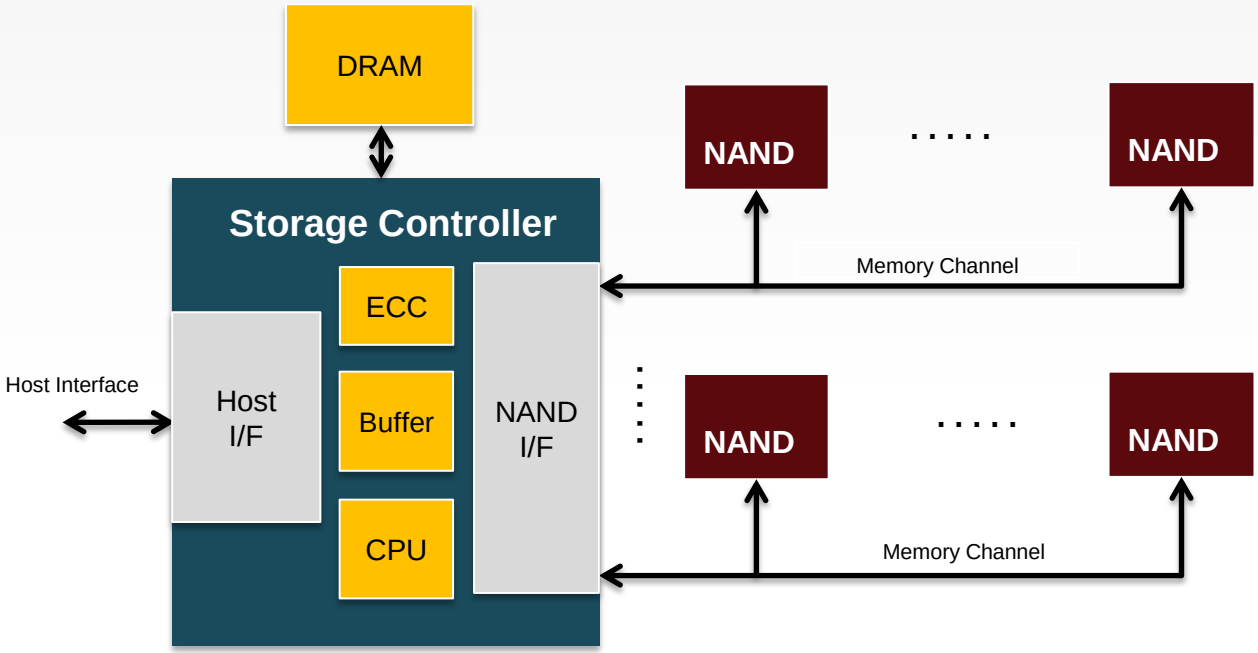
- NAND cannot revise or alter data on a page level – Erase is performed through the bulk substrate which is common to the entire NAND block of the memory cells
- The entire block of memory need to be erased before revising a data. This process will take long time and will accelerate device reliability degradation due to excessive Program Erase (P/E) cycles
- To circumvent excessive P/E cycles revised data is programmed in an erased location
- Logical to Physical mapping (L2P) is also generated and stored in a DRAM location to direct the controller to the address of the revised data
- The controller has to update and maintain this every time data revision is performed

Garbage Collection & Write Amplification



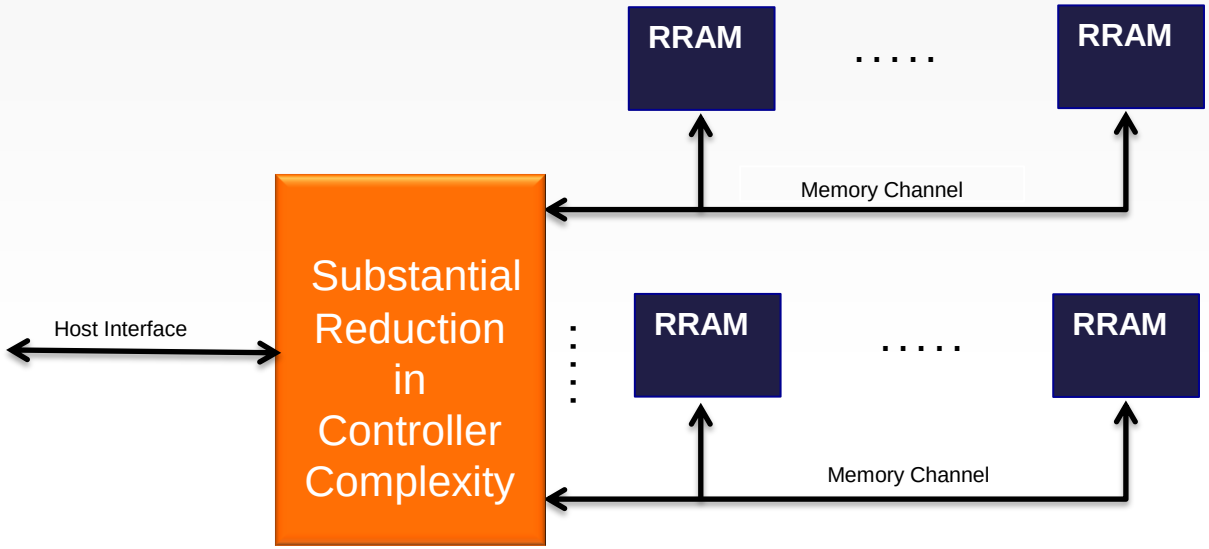
- In this example, 16 pages of data had to be moved from block 0 to block 1 to free up 8 pages that were occupied by stale data. Write Amplification for such a storage device is equal to 24(total pages in a block)/8(freed up pages) = 3.

SSD System NAND-Based



NAND Shortcomings: L2P Mapping, Garbage Collection, Wear Leveling, Bad Block Management, ECC Complexity

SSD System RRAM-Based

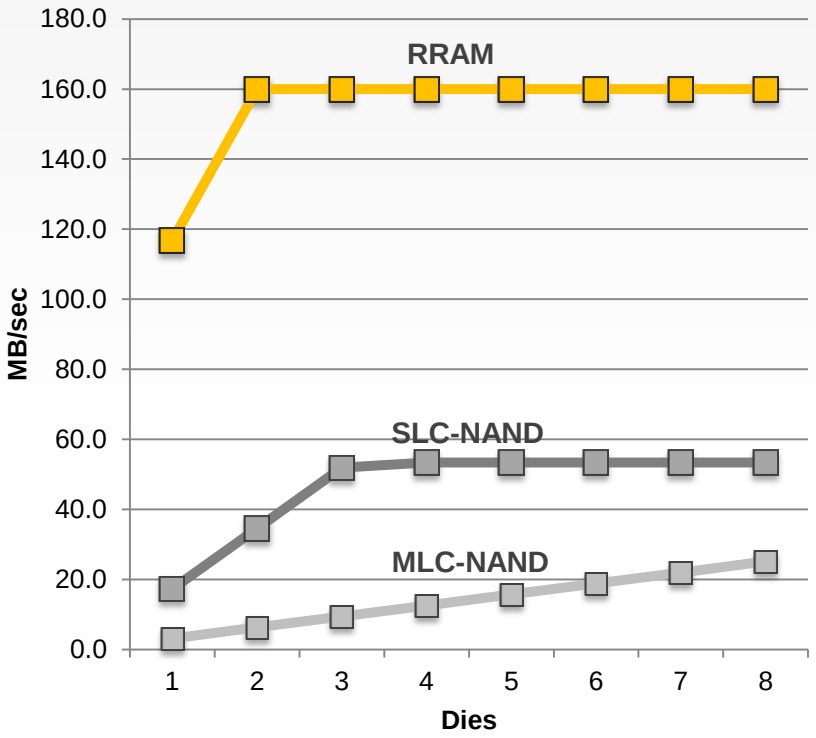


RRAM-Based SSD substantially reduces NAND shortcomings, thus significantly reducing controller complexity

50

SSD System Write Performance with NAND & RRAM

NAND Spec.	MLC	SLC	RRAM
NAND bus freq DDR (MHz)	100	100	100
Bus width (bits)	8	8	8
Page Size (KB)	16	16	4
Shift Time + Overhead (us)	100	100	25
Program Time (ms)	1.5	0.3	.032
Read Latency (us)	50	25	1
Write Amplification	3	2	1
Effective Write xfer rate (MB/s)	32	53	160



- **Maximum utilization of the channel**
- **5X performance improvement**

In Summary

- **RRAM provides future Systems**
  - **Superior performance and lower system power consumption**
  - **Better reliability**
  - **Larger densities with 3D integration**
  - **Embedded memory in advanced CMOS nodes**
  - **Ease of manufacturability with standard CMOS compatible material**
  - **Scalability sub <10nm nodes**
  - **New system architectures**