

2019

Short Course 2



# international ELECTRON DEVICES meeting

Sunday, December 8, 2019

Organizer: Ali Keshavarzi, Stanford University

## Short Course 2: Technologies for Memory-Centric Computing

Memory Devices and Selectors for High-Density Memory Technology

3D-Stacked DRAM Technology & Function-in-Memory Solution

Novel Memory Technologies for Advanced CMOS Nodes

Emerging Technologies for Memory-centric and Low-power Architectures

Towards Memory-centric Autonomous Systems: A Technology and Device Perspective

3D NAND Challenges and Potentials



# Memory Devices and Selectors for High-Density Memory Technology

Alessandro Calderoni

*Micron Technology Inc.*

# Outline

- Introduction
- Technology Scaling and Computing Systems Trends
- Memory Scaling Challenges
- Emerging Memory and Selectors
- Conclusions

# Outline

- **Introduction**
- Technology Scaling and Computing Systems Trends
- Memory Scaling Challenges
- Emerging Memory and Selectors
- Conclusions

# Introduction

- Technology Scaling enabled Computing Devices to be everywhere



# Data Is Everywhere

- Memory and Storage Technologies span all market segments
- 300 Billion Gigabytes of semiconductor Memory produced in 2018
  - 40GB for every person on the planet



**Data Center**



**Automotive**



**Mobile & Client**



**Internet of Things**

**85 GB/month/person  
Internet Traffic**

**50GB smart car  
per day**

**77 EB/month  
Mobile Traffic**

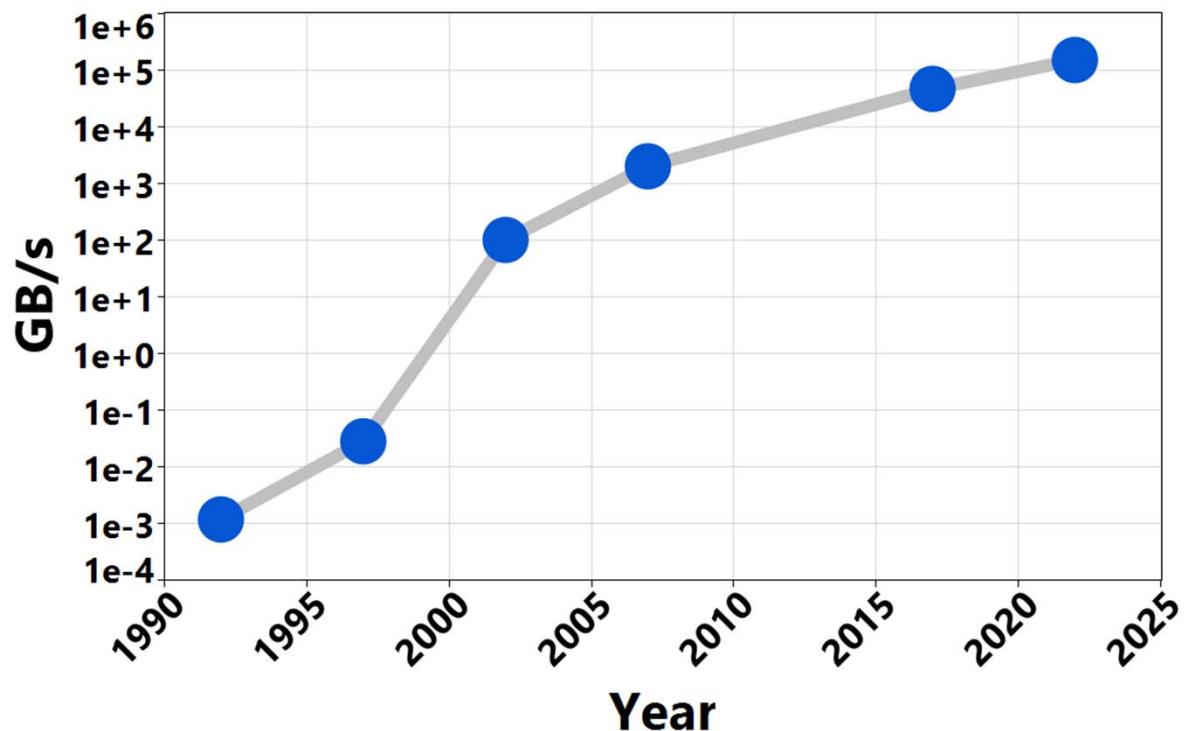
**3.6 connected  
devices per person**

*Reference [1,2 – [www.cisco.com](http://www.cisco.com)] – Projections, by 2020.*

# Data Is Everywhere

- Total Internet traffic has experienced dramatic growth in the past two decades:

Year	Global Internet Traffic
1992	100 GB per day
1997	100 GB per hour
2002	100 GB per second
2007	2,000 GB per second
2017	46,600 GB per second
2022	150,700 GB per second

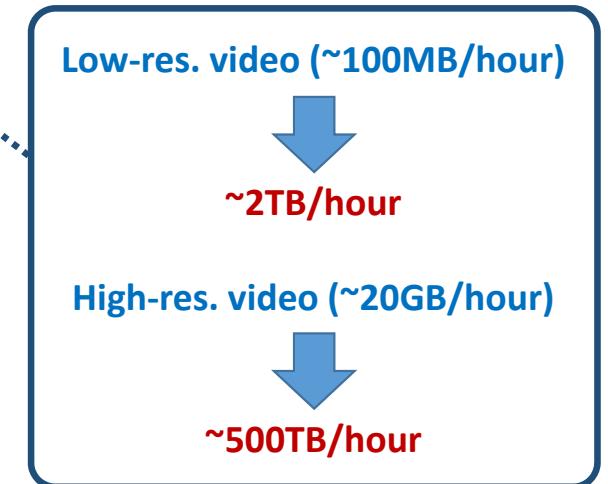
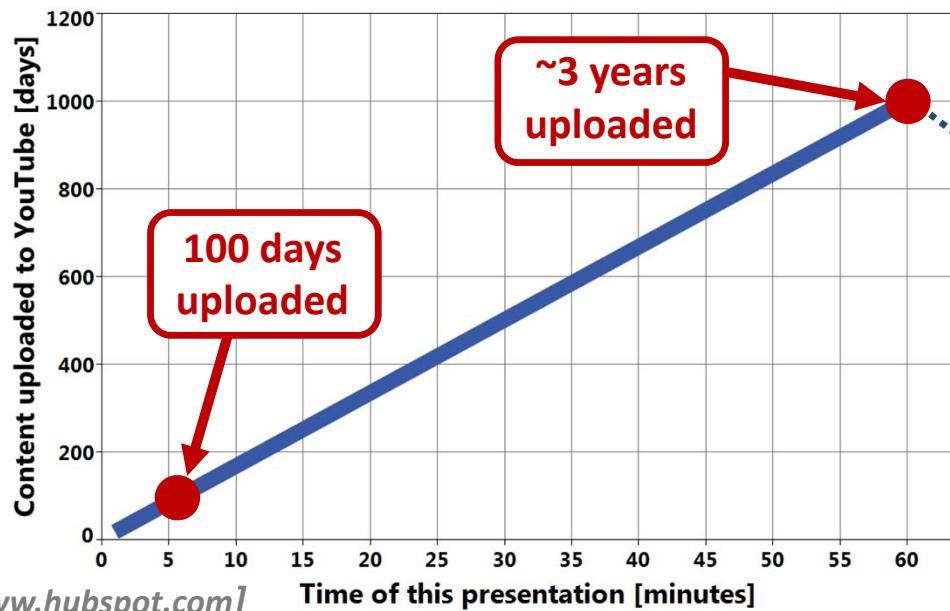


Reference [3 – [www.cisco.com](http://www.cisco.com)]

# Data Is Everywhere

- YouTube\*:
  - 1.9+ billion logged-in users each month
  - Over 1 billion hours watched every day
    - 70% of YouTube watch time comes from Mobile Devices
  - Over 400 hours uploaded every minute (2015, likely much higher now...)

*\*) as of June 2019*

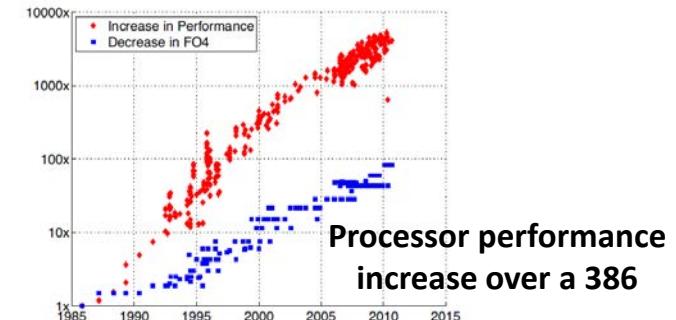
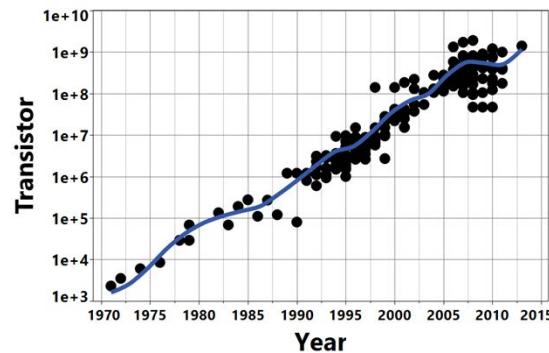
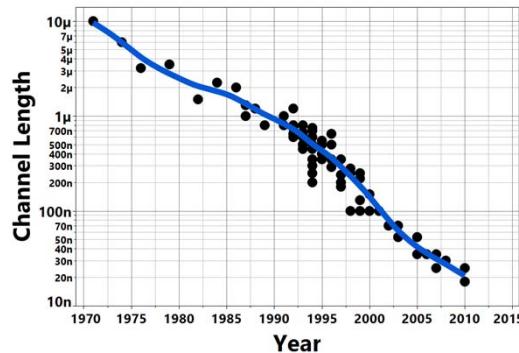


# Outline

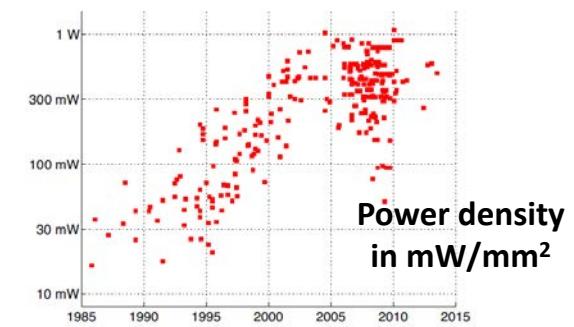
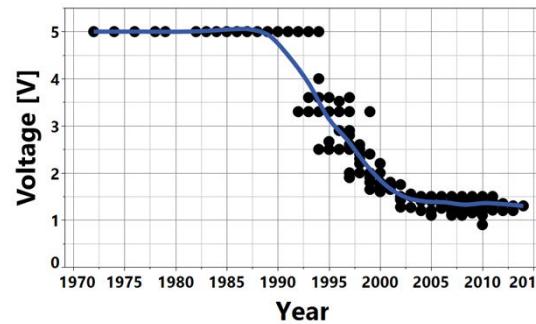
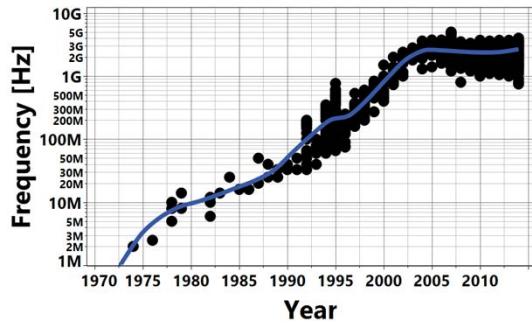
- Introduction
- **Technology Scaling and Computing Systems Trends**
- Memory Scaling Challenges
- Emerging Memory and Selectors
- Conclusions

# CMOS Scaling

- Technology Scaling has followed Moore's exponential law for decades



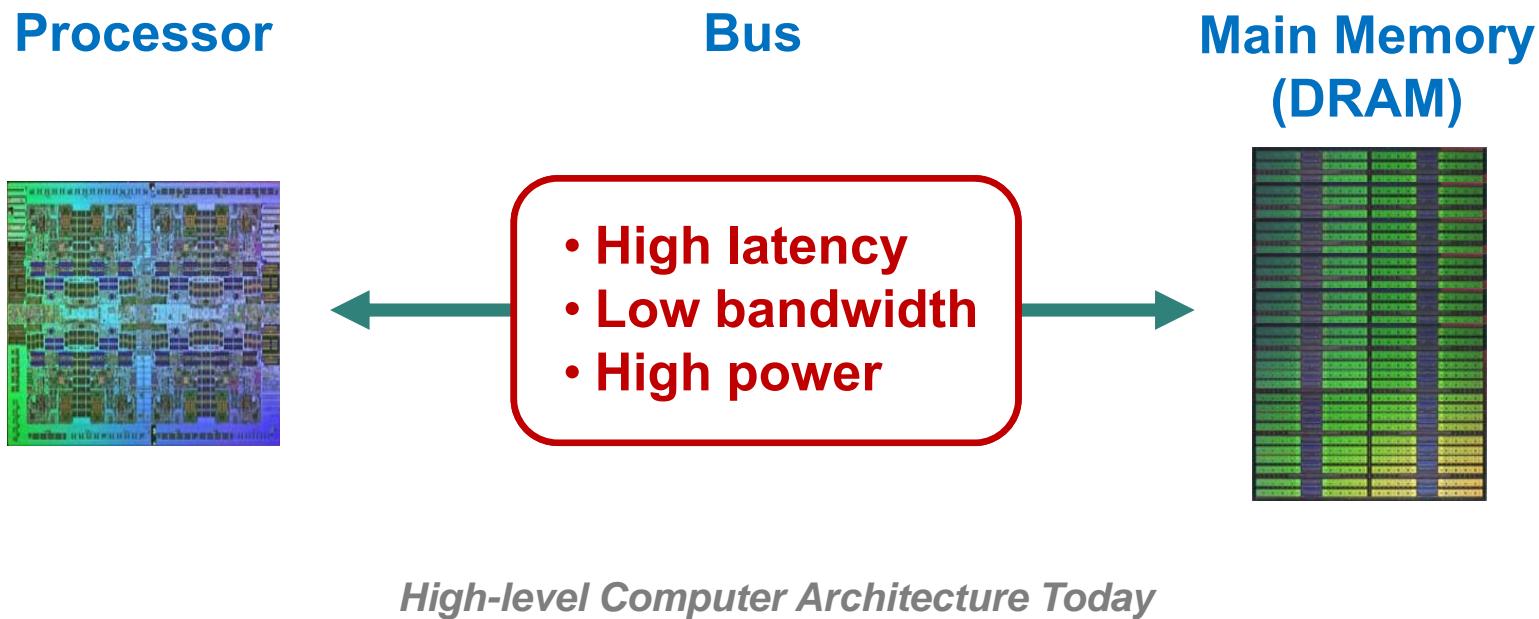
- But over the last decade it has faced new challenges:



Reference [6 - M. Horowitz, "Computing's energy problem (and what we can do about it)"]

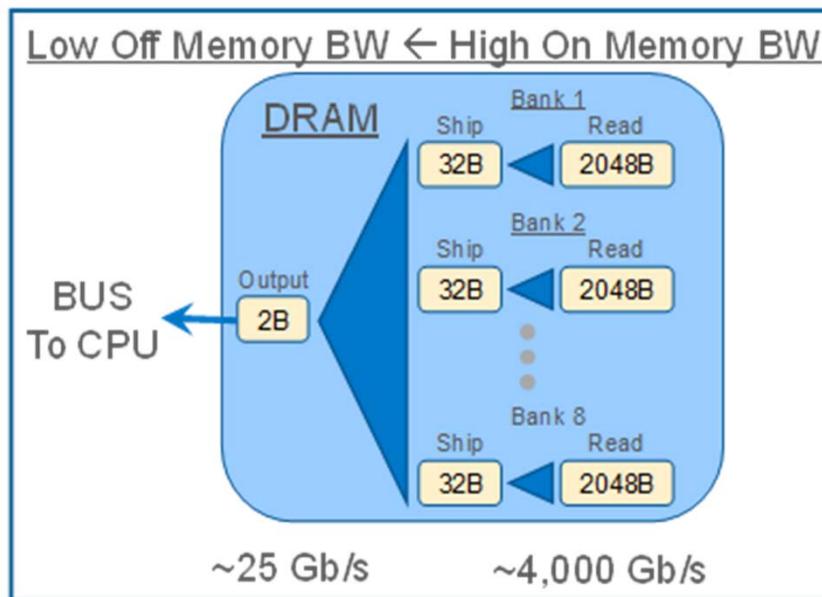
# The Memory Wall

- The Memory Sub-System has become a fundamental bottle-neck for all Computing Systems



# Memory Sub-System – Low Bandwidth

- The gap between the internal memory and the bus bandwidth can be higher than 100x

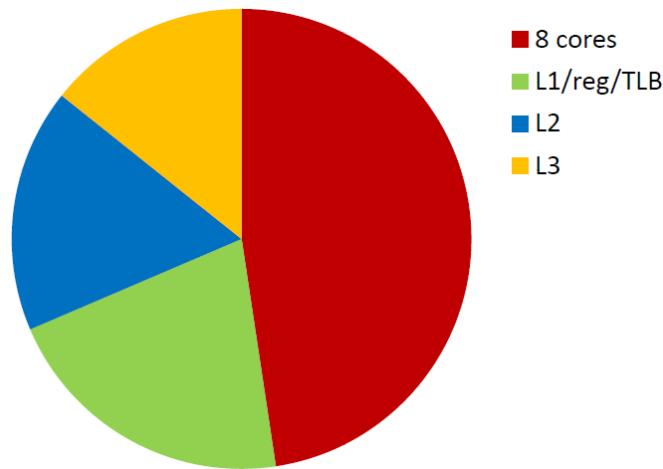


Reference [7 - S. DeBoer, "Memory Technology: The Core to Enable Future Computing Systems"]

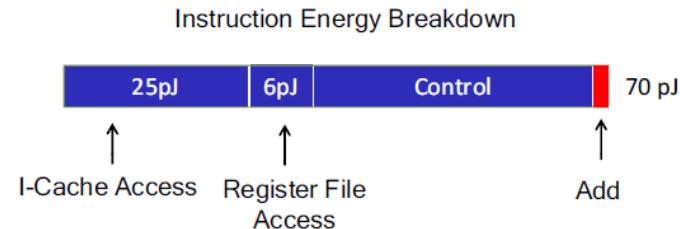
# Memory Sub-System – High Power

- Data movement can consume up to 50% of the system power
- Cache hierarchy consumes a significant amount of energy but it drastically reduces the main memory accesses → reduces overall system energy

Power Breakdown of a 8-core Processor in 40nm Technology



Rough energy cost for various operations in 45nm Technology



Simple Operation ~ 0.1-5pJ

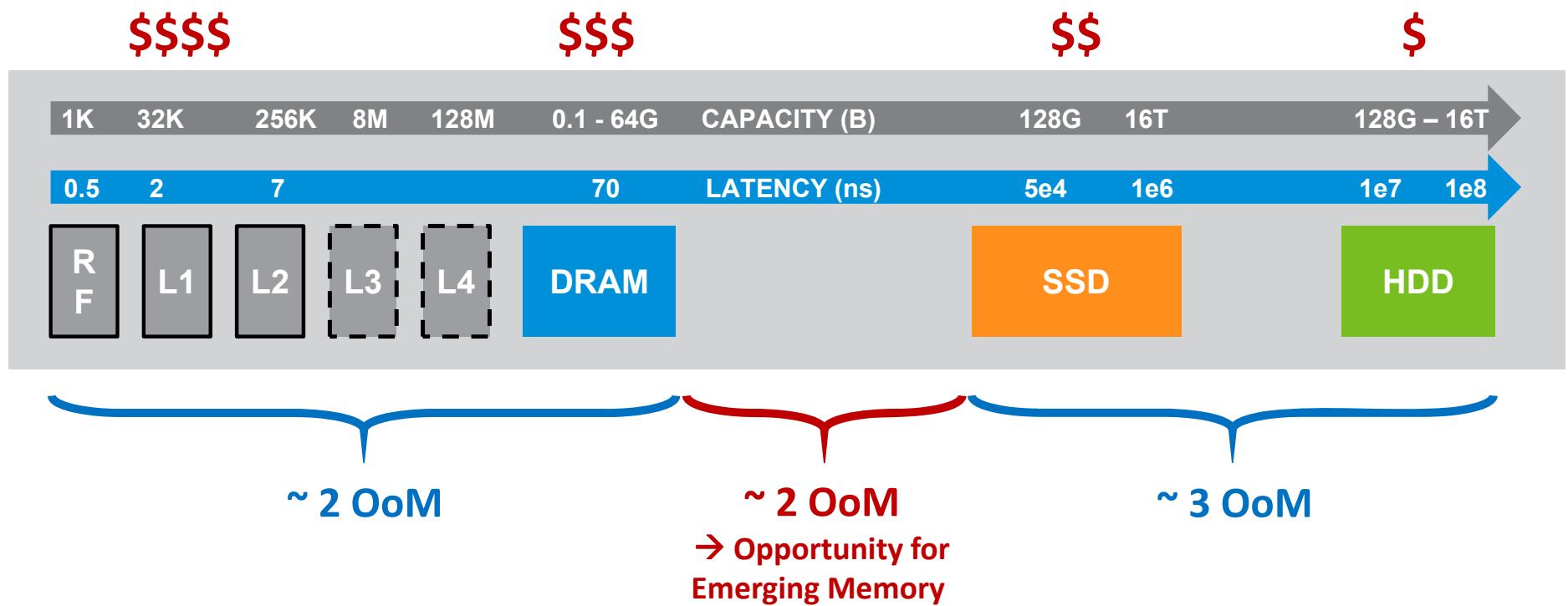
L1 Cache Access ~ 10-100pJ

DRAM Access ~ 1000-3000pJ

Reference [6,8 - M. Horowitz, "Computing's energy problem (and what we can do about it)" D. Pandiyan, et al., "Quantifying the energy cost of data movement for emerging smart phone workloads on mobile platforms"]

# Memory Sub-System – High Latency

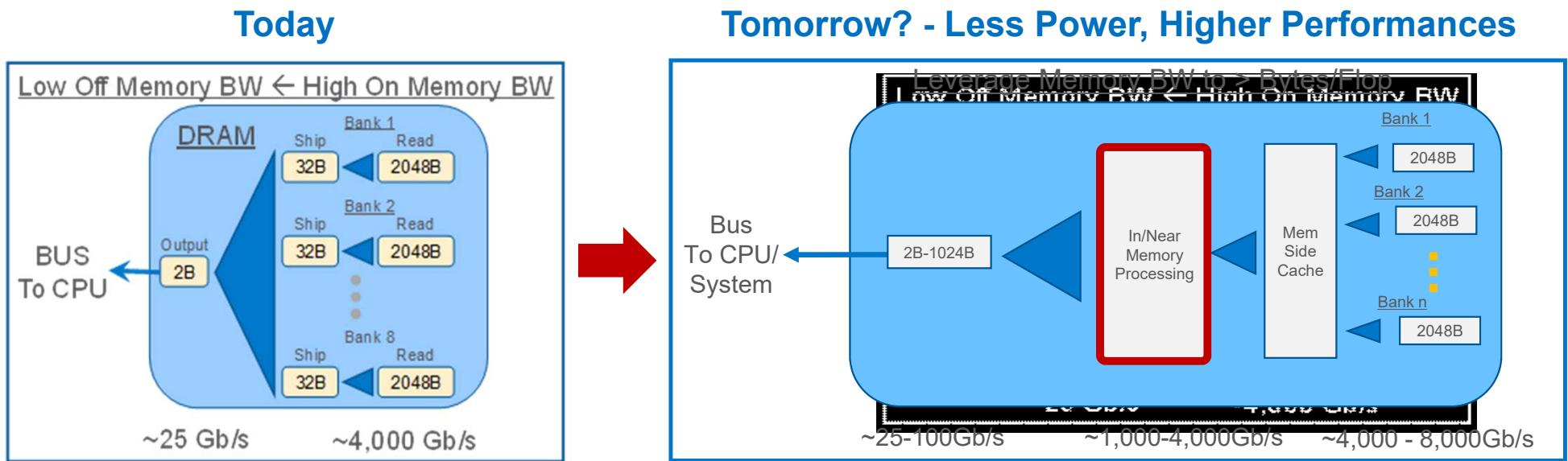
- The entire Memory Hierarchy in today's Computing Systems is dictated by its latency



Reference [7 - S. DeBoer, "Memory Technology: The Core to Enable Future Computing Systems"]

# Transition to Memory-Centric Computing

- To improve system performance and power efficiency  
→ **MOVE compute to where the data is stored**



Reference [7 - S. DeBoer, "Memory Technology: The Core to Enable Future Computing Systems"]

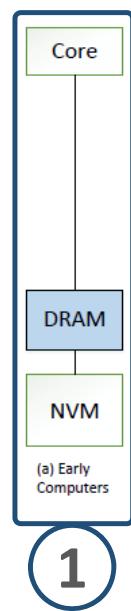
# In-Memory Computing - Examples

- Computing Systems evolution towards in-memory computation

## Classification of Computing Systems based on working set location

■ Working set location

1. Early architectures: Data – Long Bus – Core



Reference [9 - G. Singh, et al., "A Review of Near-Memory Computing Architectures: Opportunities and Challenges"]

# Outline

- Introduction
- Technology Scaling and Computing Systems Trends
- **Memory Scaling Challenges**
- Emerging Memory and Selectors
- Conclusions

# DRAM Scaling Challenges

- **Technology Specific:**

- *Capacitor, Access Device, WL/BL RC, Cost*

- **Process Integration:**

- *Lithography, Alignment and Overlay*

- **Architecture/System-Level:**

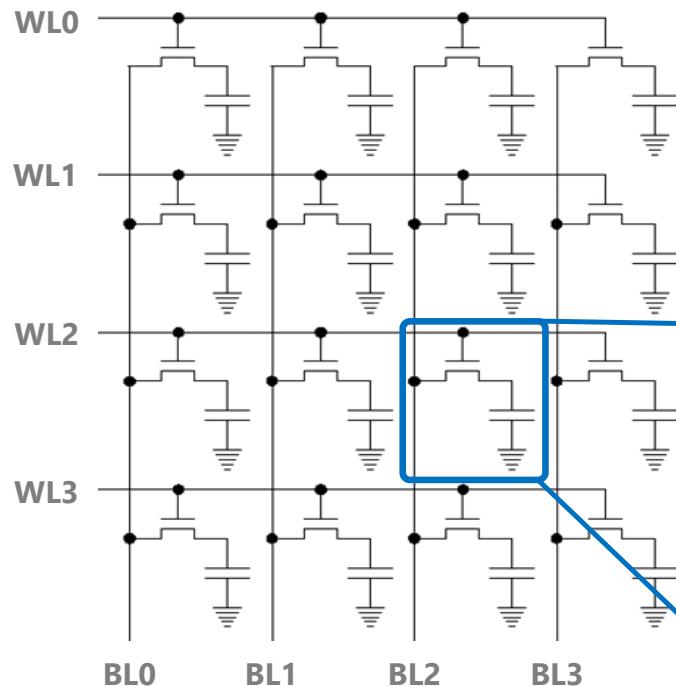
- *PPAC - Power, Performance, Area, Cost*

# DRAM Basics

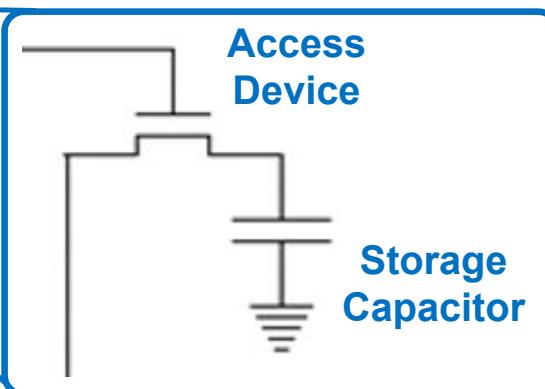
- Capacitor's charge state indicates stored value

Reference [10 - N. Ramaswamy, et al., "Metal Gate Recessed Access Device (RAD) for DRAM Scaling"]

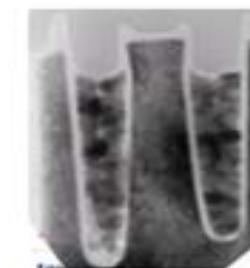
## DRAM Array Architecture



- Cell loses charge when read → **Needs write-back**
- Cell loses charge over time → **Needs refresh**



Access  
Device



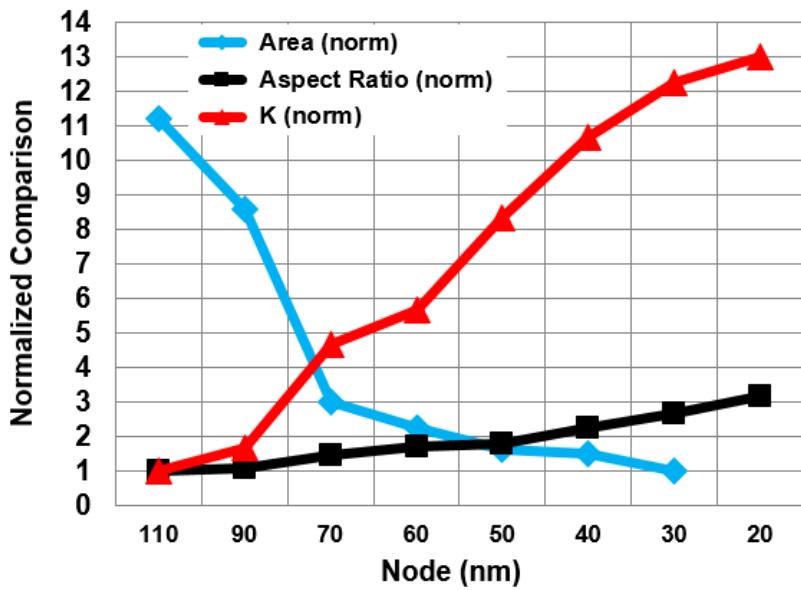
Storage  
Capacitor



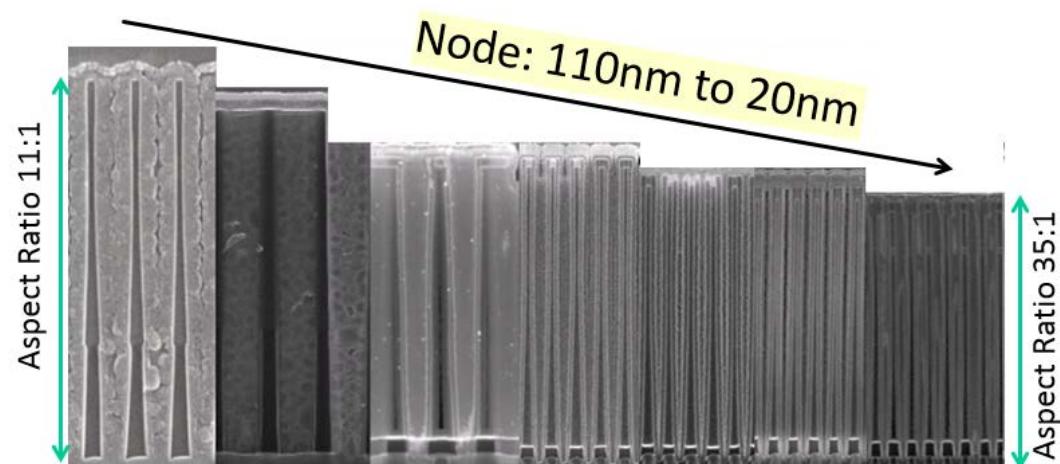
# Technology - Container

- Area has reduced more than 10x - While Capacitor requirement for sensing has stayed constant ~10fF
- Dielectric Constant (k) has increased more than 10x - to compensate for area reduction
- Aspect Ratio (AR) has increased more than 3x - with AR > 35:1

Container Scaling vs. Technology Node

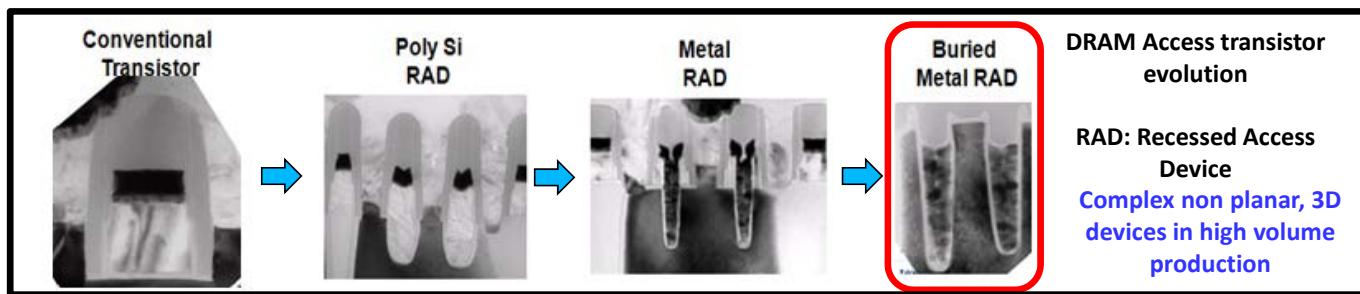


Container Scaling vs. Technology Node



# Technology – Access Device

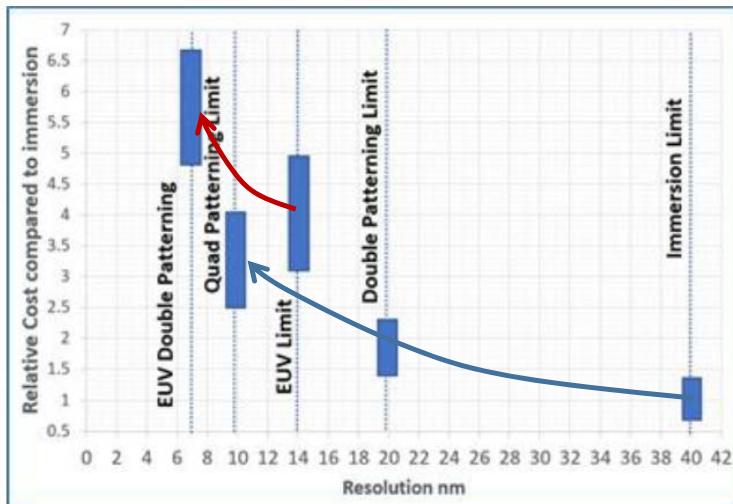
- Highly optimized three dimensional device → running out of space
- On current requirement  $\sim >1\mu\text{A}$  ( $I=C*V/t=10\text{fF}*0.5\text{V}/5\text{ns}$ )
  - To quickly read and write the Cell → specs have not relaxed with scaling
- Off current requirement  $\sim <1\text{fA}$  ( $I=C*V*0.1/t=10\text{fF}*0.05\text{V}/100\text{ms}$ )
  - To minimize the refresh time → refresh commands are consuming a large fraction of the available bandwidth
- Variability and Noise have to be taken into account
  - I.e. the refresh time is set according to the requirement of full distributions



# Process Integration – Lithography and alignment

- DRAM's scaling path requires an aggressive lithography roadmap for both resolution and alignment
- The cost for the lithography step increases with large error bars on the relative costs depending on both the pace of cost maturity for extreme EUV as well as the methodology used for pattern multiplication

## Lithography technology breakpoints and relative costs



## Self-Aligned Quadruple Patterning (SAQP)

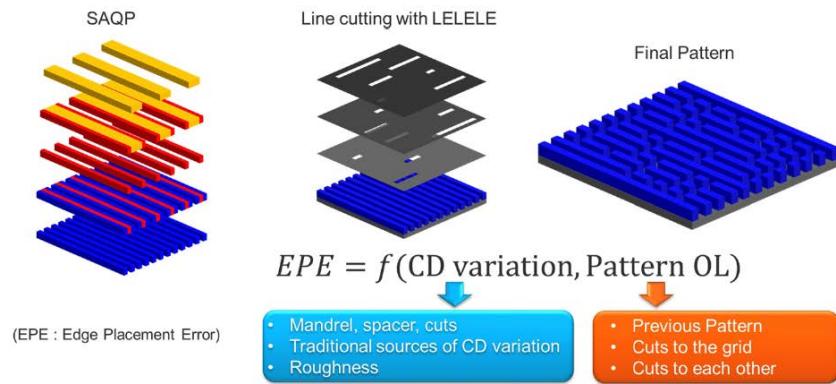


FIG. 2. Illustration of a self-aligned quadruple patterning (SAQP)<sup>8</sup> flow with unidirectional lines and multiple cut masks. Dense lines and space patterns are initially formed by SAQP, and then multiple cuts or trim masks with very fine isolated features are used to cut the lines into useful device features or wiring.<sup>9,10</sup> The EPE is a function of the critical dimension (CD) variation as well as the pattern overlay (OL) error for each mask.<sup>11</sup>

Reference [7 - S. DeBoer, "Memory Technology: The Core to Enable Future Computing Systems"

R. Clark, et al., "Perspective: New process technologies required for future devices and scaling"]

# Architecture/System-Level Optimizations

- Specialized DRAM to optimize Performance/Power/Area/Cost

## LP-DRAM

- Low-Power
- Mobile, Embedded, Automotive



## GDDRx

- High-bandwidth
- High-performance computing (HPC), gaming, automotive and networking



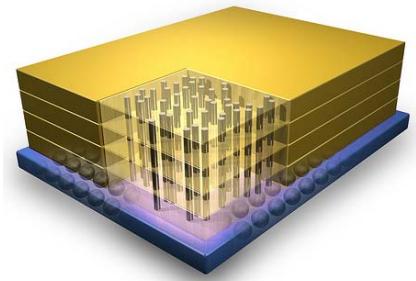
## RL-DRAM

- Reduced-Latency
- Networking, image processing, cache applications

	RLDRAM 3	RLDRAM 2
Data Rate	400–2133 Mb/s	350–1066 Mb/s
t <sub>RC</sub>	8ns	15–20ns
Density	576Mb, 1Gb	288Mb, 576Mb
Voltages	1.35V core; 1.2V IO	1.8V core; 1.5–1.8V IO
Configurations	x18, x36	x9, x18, x36
Burst Length	2, 4, 8	2, 4, 8
Multibank Write	Enables 1.0ns READ t <sub>RC</sub>	Not Applicable
Internal Banks	16	8
t <sub>FAW</sub> Delay	No delay	No delay
Bus Turnaround Delay (RD-WR-RD)	1–2 clock cycles	1–2 clock cycles
RESET Pin	Available	Not Available

## Stacked

- Ultra-high bandwidth
- Reduced power per bit
- Brings data and compute closer!

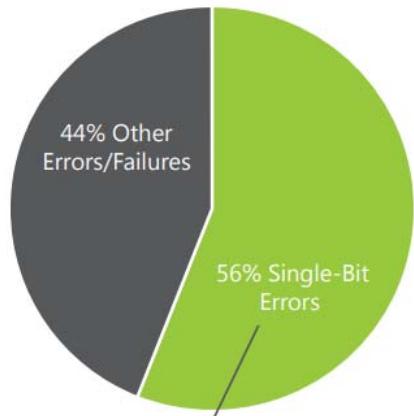


Reference [13,14,15 – [www.micron.com](http://www.micron.com)]

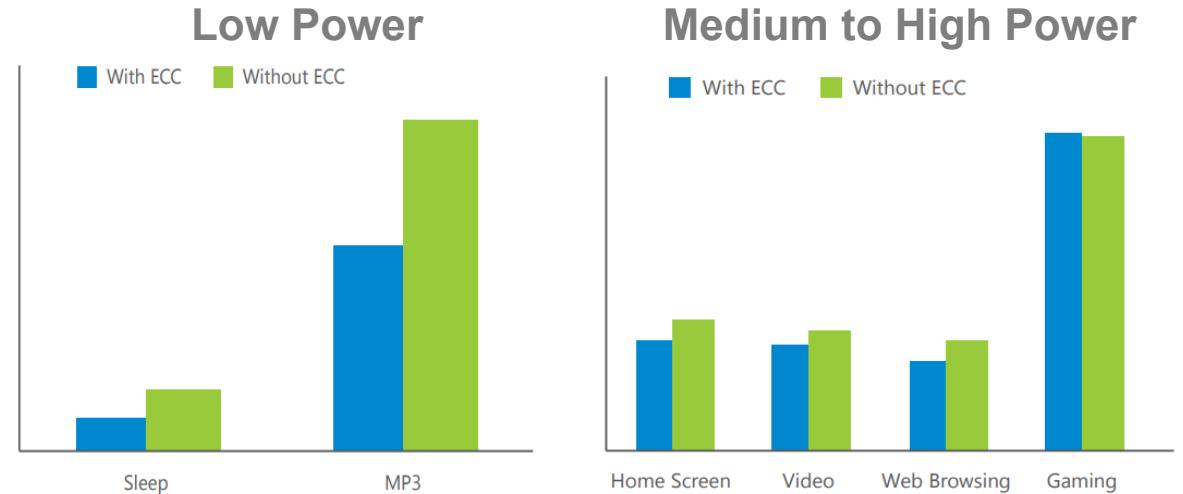
# DRAM Reliability and ECC

- DRAM yields are primarily limited by single-bit errors:
  - “hard” bits (stuck at 0 or 1 state) → repaired using redundant elements
  - “marginal” bits → OK if refreshed more often or written for a longer time
- Error Correcting Codes (ECC) used to improve power and reliability challenges
  - I.e. Hamming codes with 8b of parity to detect and correct 1b in a 128b codeword

LPDRAM field failures  
with known root cause



LPDDR4 Power Consumption Use Cases



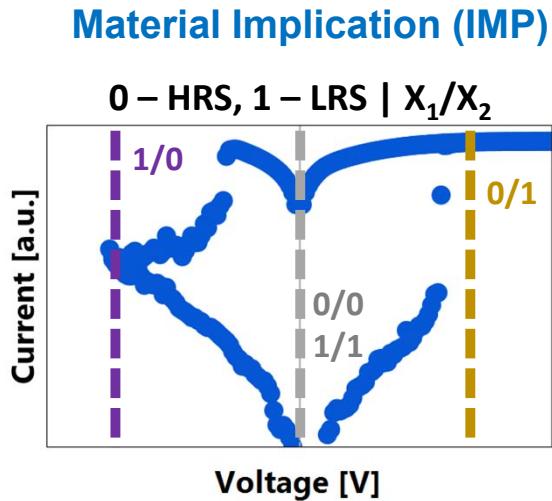
# Outline

- Introduction
- Technology Scaling and Computing Systems Trends
- Memory Scaling Challenges
- **Emerging Memory** and Selectors
- Conclusions

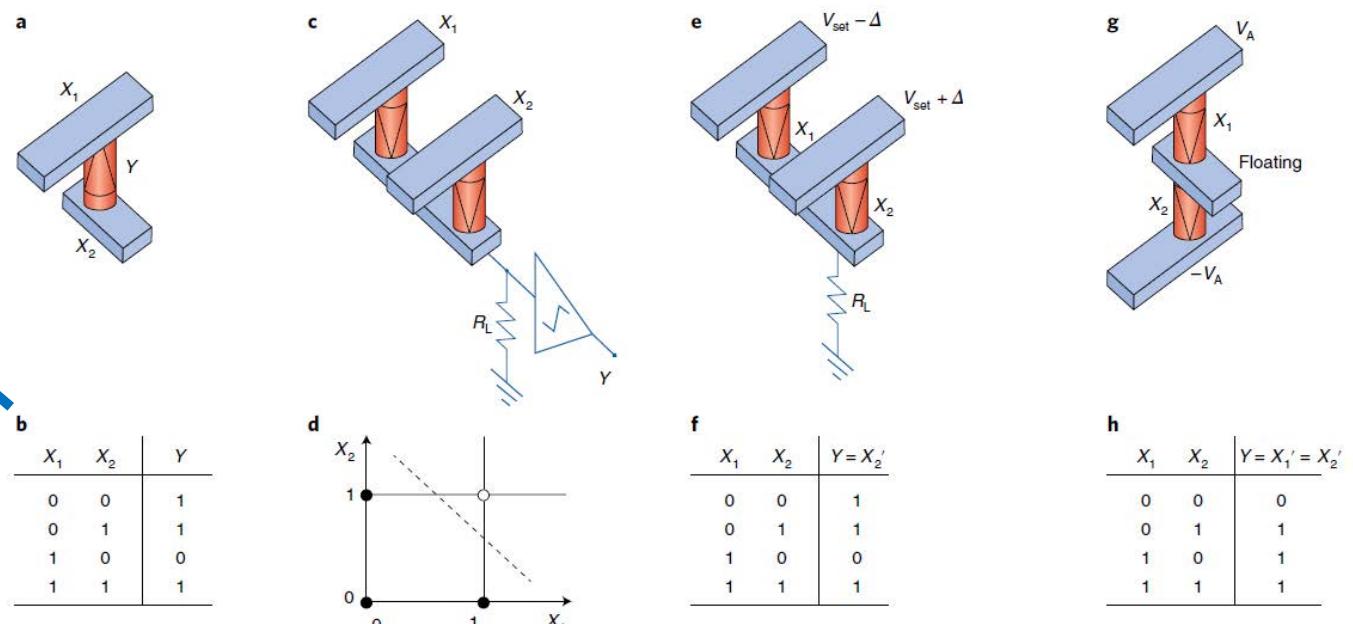
# Logic Gates for In-Memory Computing (IMC)

- Emerging Memory can be used to perform near or in memory computing → different functionally complete logic are possible

Examples of logic gates for near- or in-memory computing



Reference [18]



# Ideal Device Characteristics for In-Memory Computing

## High Speed

- Switching speed <100ns
- Preferably <10ns

## Low Power

- Enable high-parallelism
- Low current/voltage to limit IR-drops in high density array

## Scalable

- Two terminal devices
- Dimensions ~10nm and below

## Low Noise

- Between consecutive
  - reads
  - writes

## High Endurance

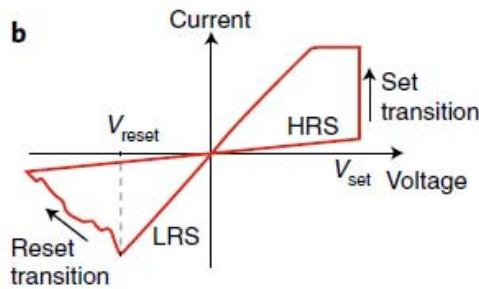
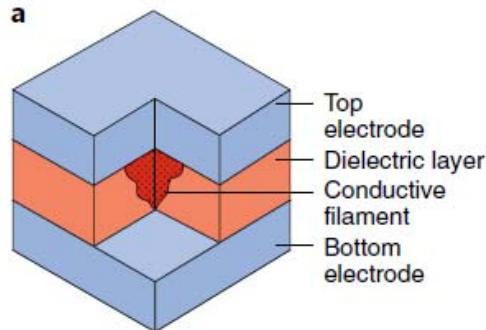
- Ideally, DRAM-like “infinite” reliability  
→ ~1E15 cycles

## Long Retention

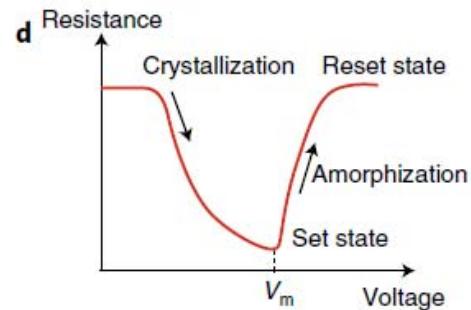
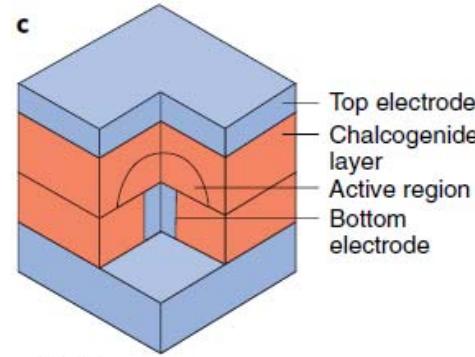
- Main memory that is also non-volatile
- Few years at 70C

# Emerging Memory for IMC

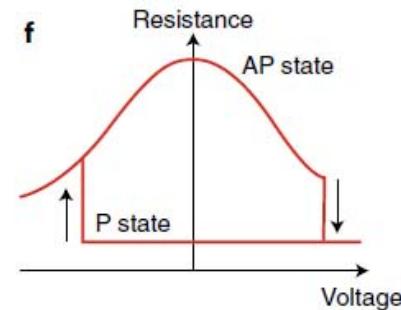
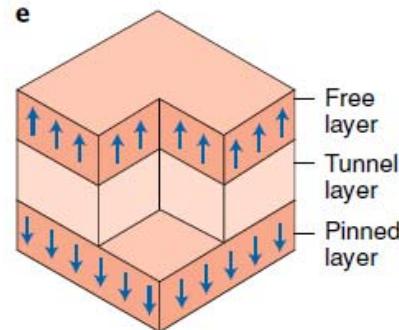
**RRAM**



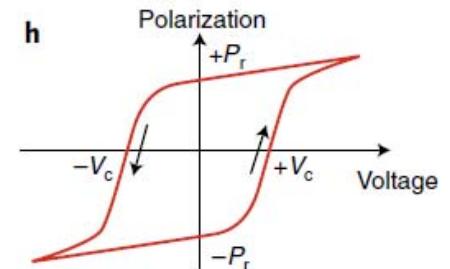
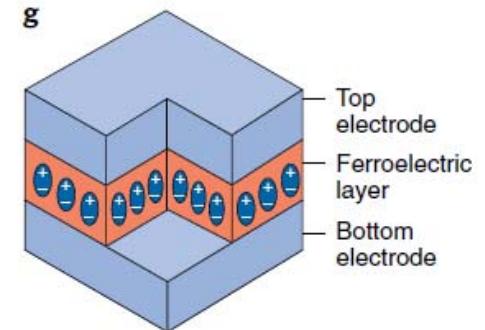
**PCM**



**MRAM**



**FeRAM**

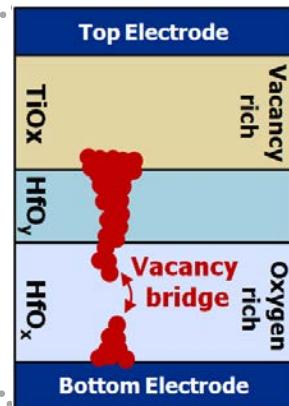
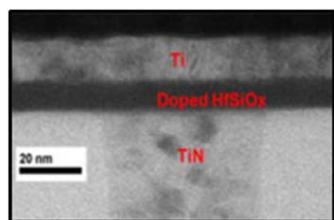


Reference [18 - D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices"]

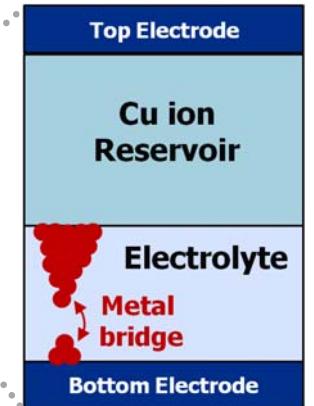
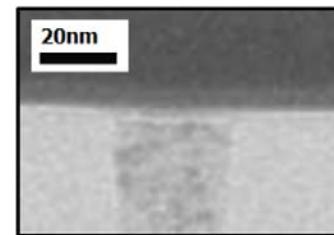
# RRAM - Options

- Broad landscape → everything switches! Egg albumen, banana peels, ...
- Two main categories based on switching mechanism: O-ReRAM, M-ReRAM

O-ReRAM cell stack
Reactive Ti Top Electrode
Amorphous ALD HfSiO <sub>x</sub> Dielectric Layer
TiN Bottom Electrode



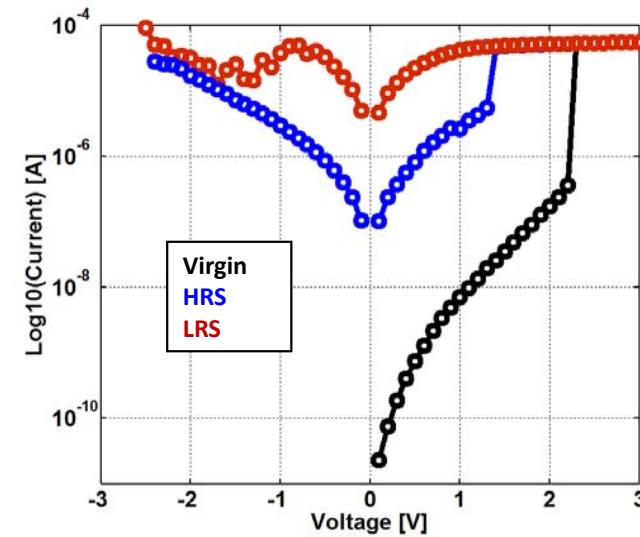
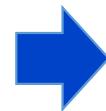
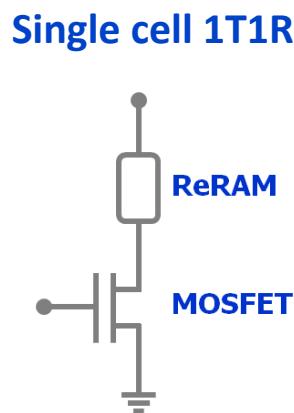
M-ReRAM cell stack
Cu ion reservoir
Electrolyte Layer
TiN Bottom Electrode



Reference [19 - A. Calderoni, et al., "Performance Comparison of O-based and Cu-based ReRAM for High-Density Applications"<sup>28</sup>]

# RRAM – Single Device Characterization

- Since the conductive filament is believed to approach few atoms in size, both O-ReRAM and M-ReRAM are expected to be susceptible to variability, read and program noise
- Single cell data is not sufficient to evaluate and compare cell performance:

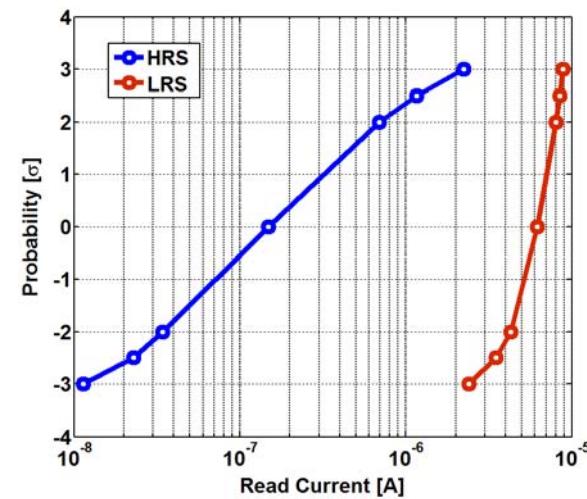
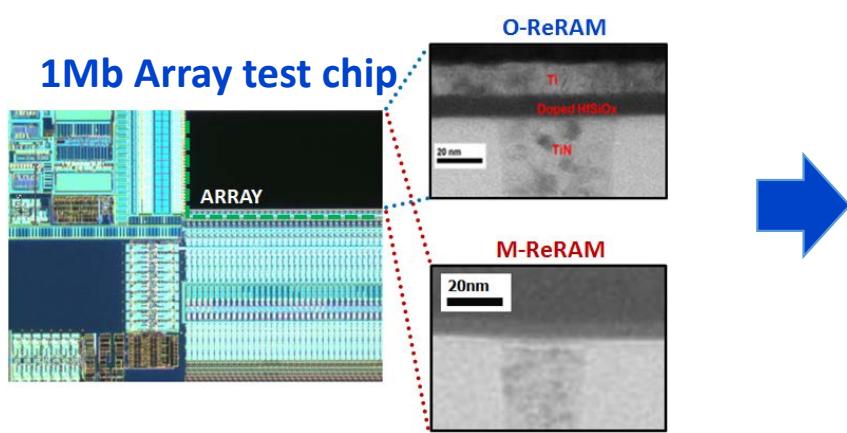


Reference [19 - A. Calderoni, et al., "Performance Comparison of O-based and Cu-based ReRAM for High-Density Applications<sup>29</sup>"]

# RRAM – Variability

- Since the conductive filament is believed to approach few atoms in size, both O-ReRAM and M-ReRAM are expected to be susceptible to variability, read and program noise
- Single cell data is not sufficient to evaluate and compare cell performance:

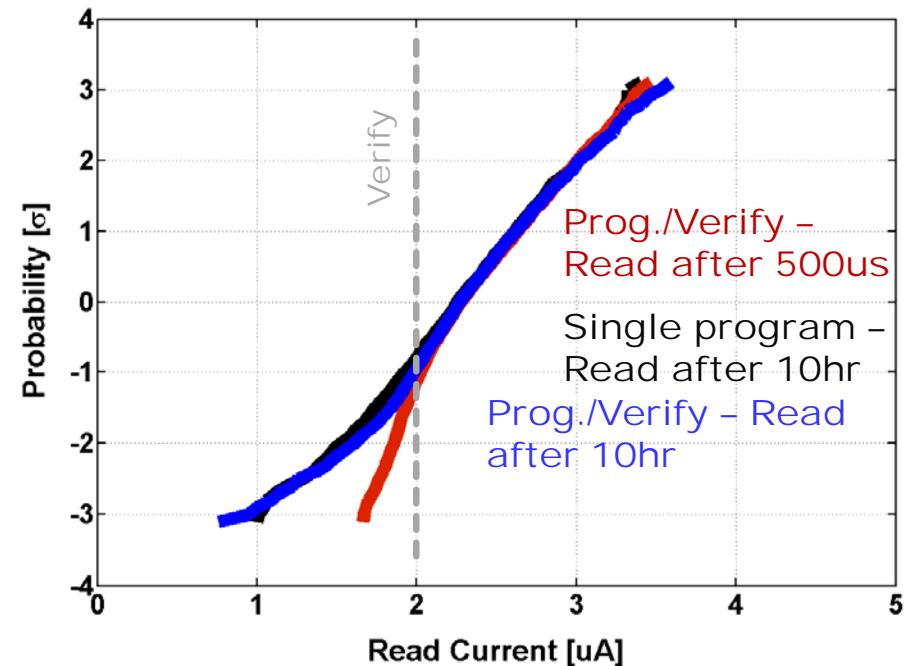
**Need to account for Variability and noise by studying distributions**



Reference [19 - A. Calderoni, et al., "Performance Comparison of O-based and Cu-based ReRAM for High-Density Applications<sup>30</sup>"]

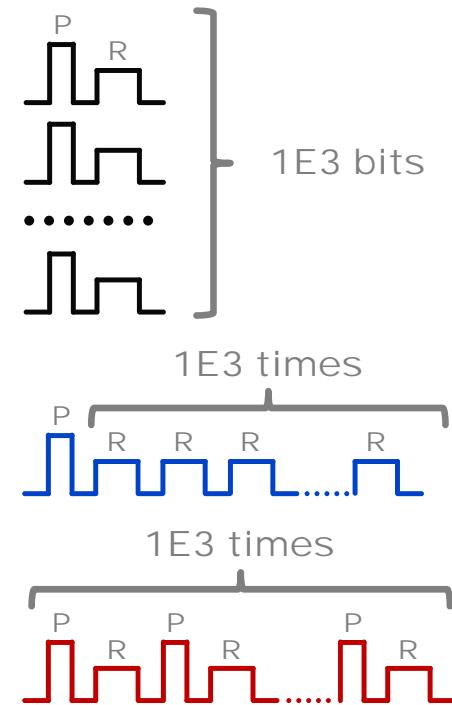
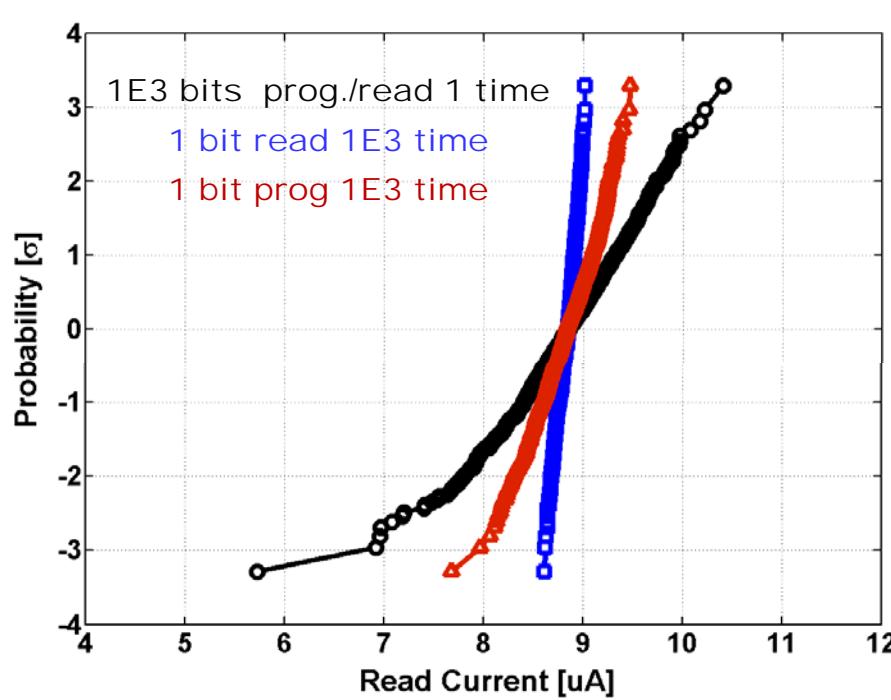
# RRAM – Variability: intrinsic component?

- Can programming algorithms be used to get better window?
- Not really, since we have to deal with intrinsic noise
- Intrinsic noise cannot be overcome (yet) → new material systems with deterministic filaments?
- “Verified” distribution relaxes to single pulse programmed distribution



# RRAM – Read vs. Program Noise

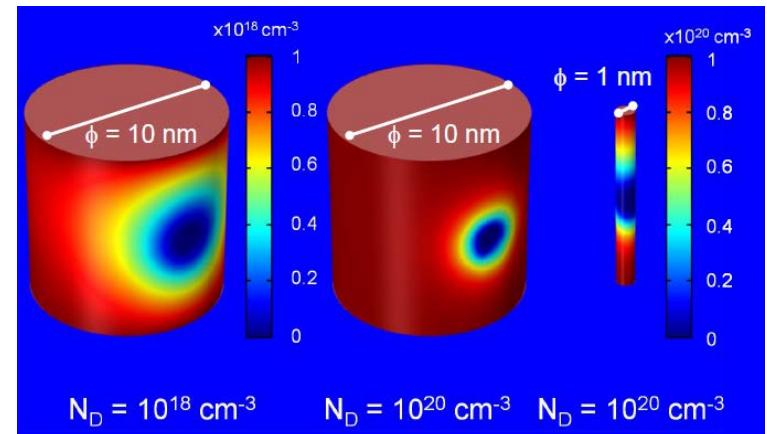
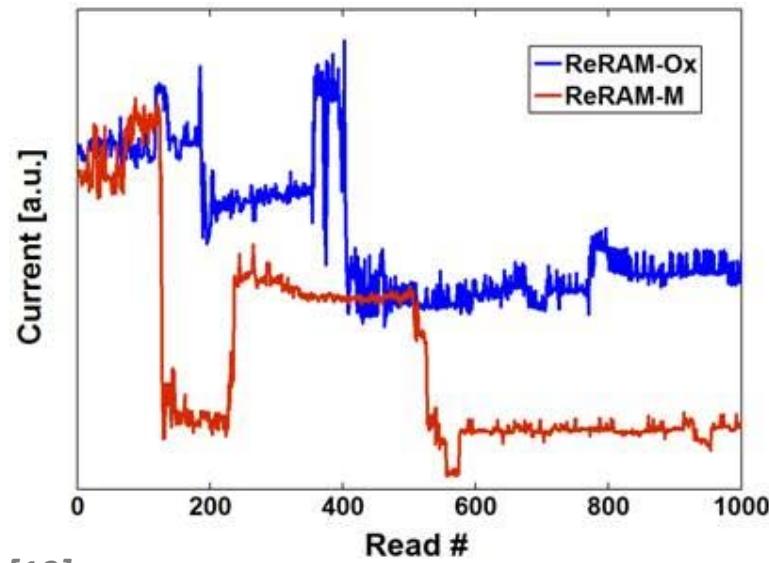
- Program noise results in non-deterministic bit placement
- Read noise results in smearing of the distribution after placement



Reference [20 - A. Calderoni, et. al., "Engineering ReRAM for High-Density Applications"]

# RRAM – Random Telegraph Noise

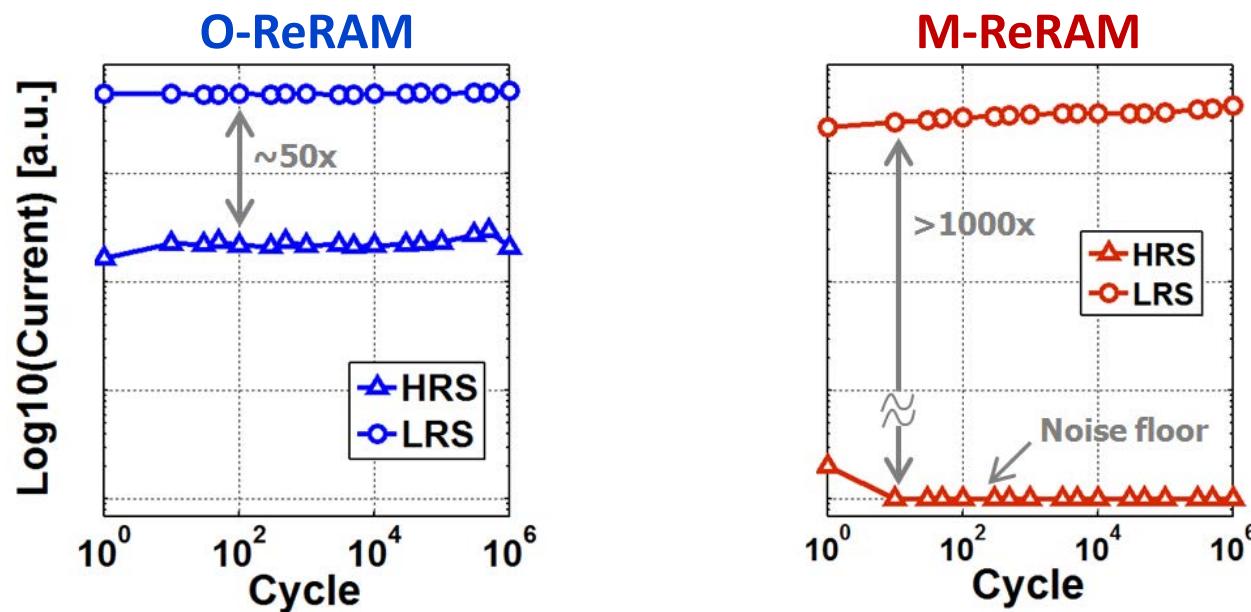
- Both O-ReRAM and M-ReRAM show Random Telegraph Noise
- RTN in O-ReRAM has been ascribed to trap-induced depletion regions in the conductive filament
- A “stronger”, large filament is less susceptible to noise



Reference [19]

# RRAM – Endurance

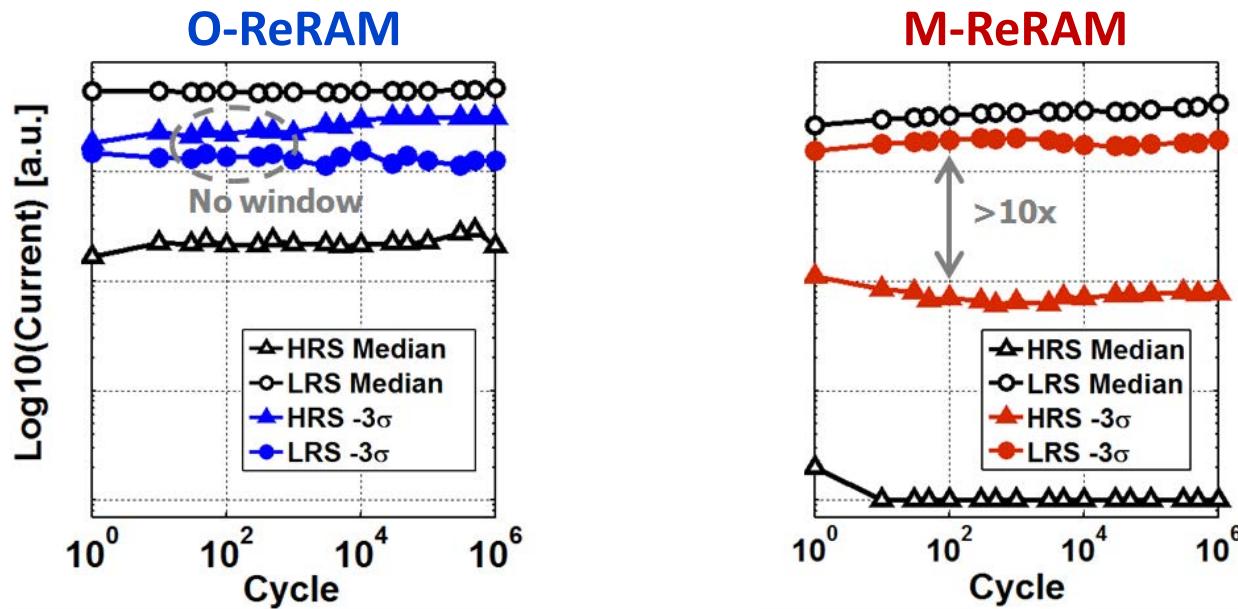
- Read ratio ( $I_{LRS}/I_{HRS}$ ) at the same  $I_{CC}$  is different:
  - ~50x for O-ReRAM and >1000x for M-ReRAM
- Median endurance is > 1E6 for both systems



Reference [19]

# RRAM – Endurance + Variability

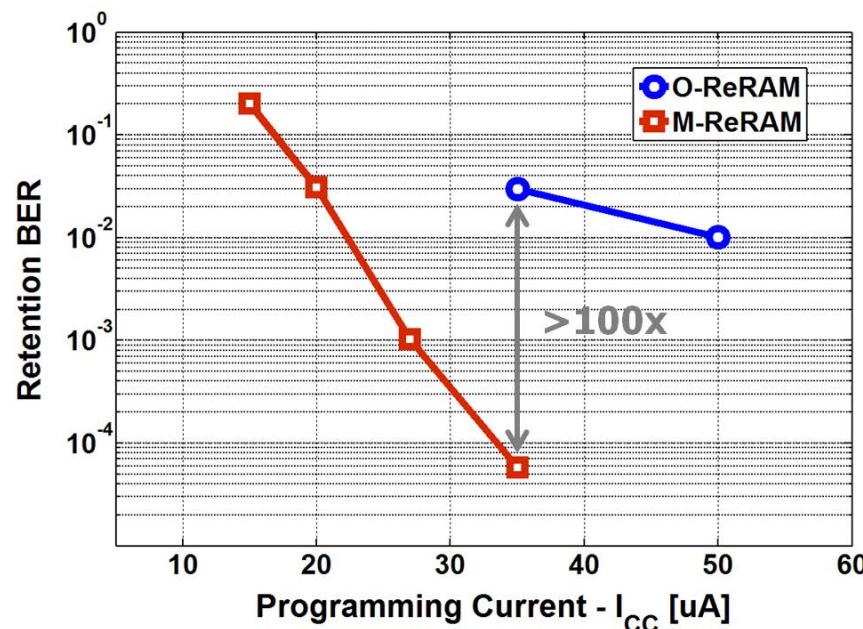
- Cells were cycled to 1E6 with 300ns single pulses with  $I_{CC} \sim 40\mu A$
- At  $3\sigma$  level, **O-ReRAM** has negative read margin through cycling
- At  $3\sigma$  level, **M-ReRAM** has constant positive read margin of  $\sim 2\mu A$



Reference [19]

# RRAM – Retention

- Retention failure was accelerated by 1h bake at 150C
- The resulting BER is a function of  $I_{CC}$
- M-ReRAM shows 100x improvement in BER over O-ReRAM

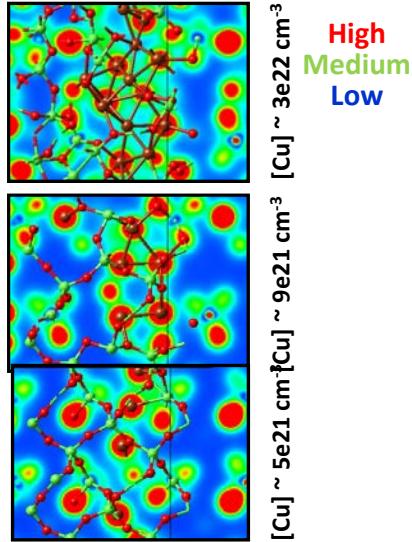


Reference [19]

# RRAM – Retention vs. Write Current

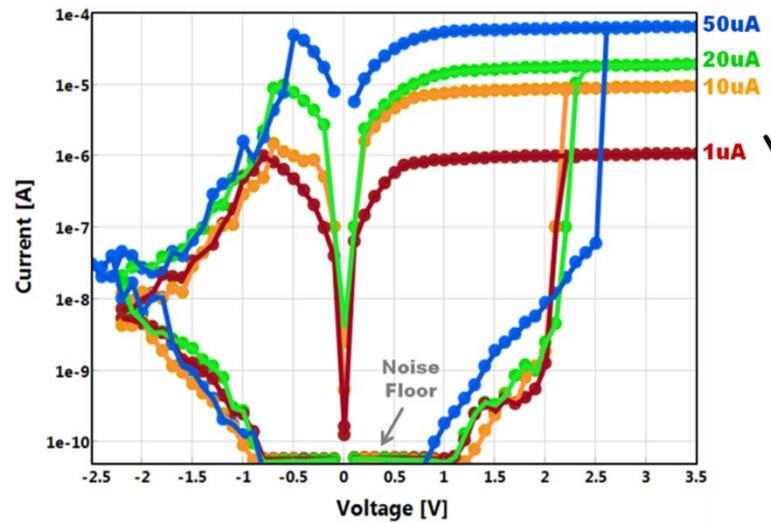
- Is  $I_{CC}=1\mu A$  cell possible?
- At lower  $I_{CC}$ : weaker filament, larger LRS tails, lower read margin

LRS vs. Cu density

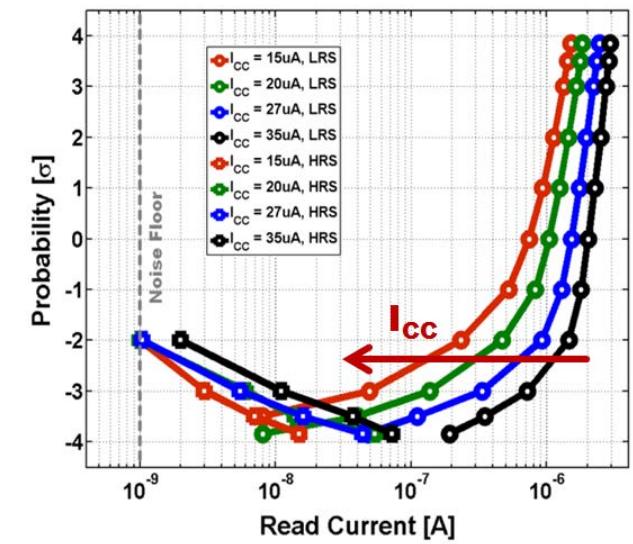


S. C. Pandey et al., JAP, 2015

Single Cell I-V vs.  $I_{CC}$



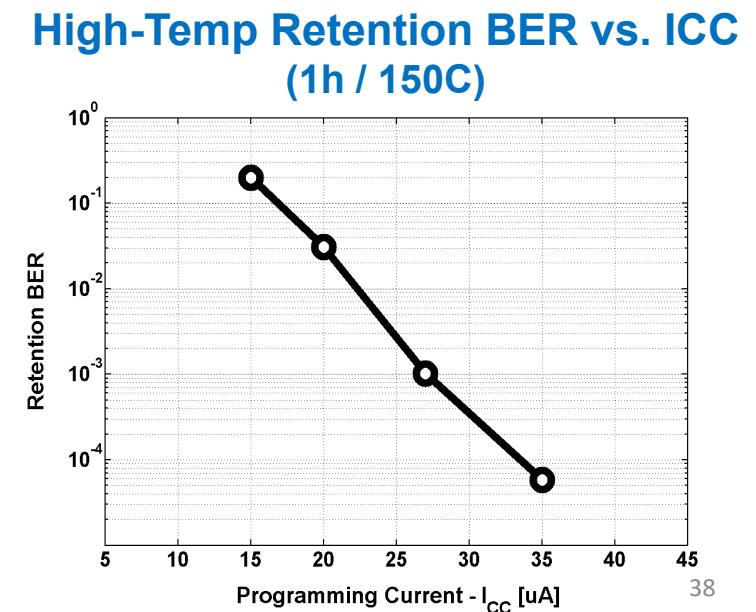
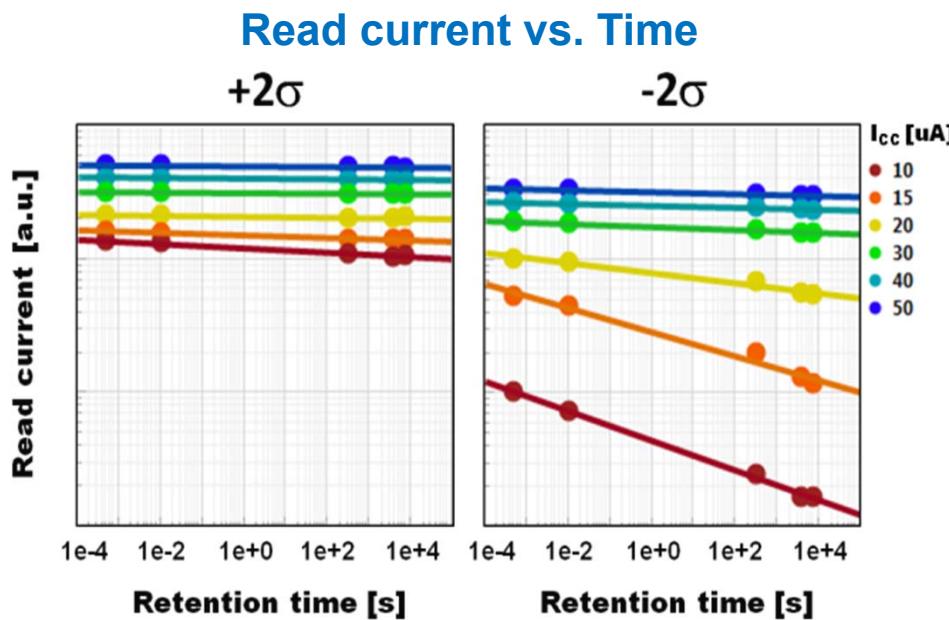
Current distributions vs.  $I_{CC}$



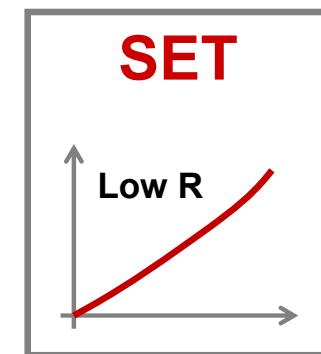
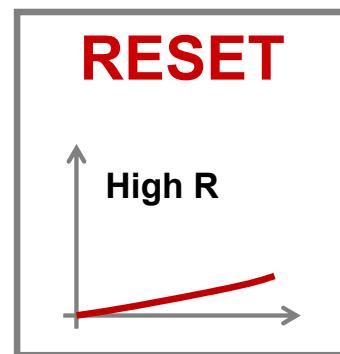
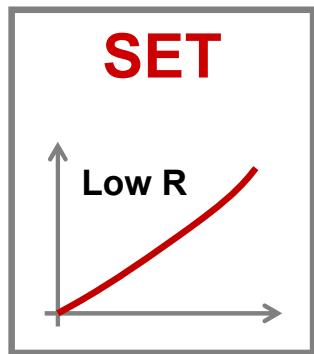
Reference [20 - A. Calderoni, et. al., "Engineering ReRAM for High-Density Applications"]

# RRAM – Retention + Variability

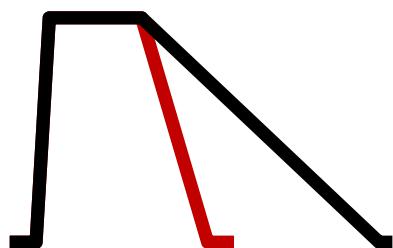
- At lower  $I_{CC}$ :
  - upper portion of the distribution ( $+2\sigma$ ) is stable over time
  - lower portion of the distribution ( $-2\sigma$ ) shows instability
  - Retention BER is a strong function of  $I_{CC}$



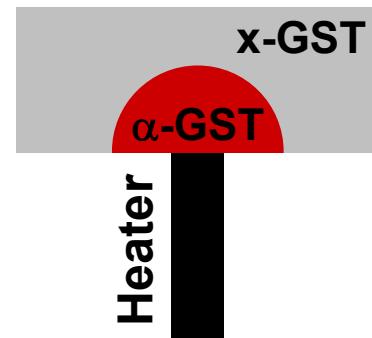
# PCM – Working Principles



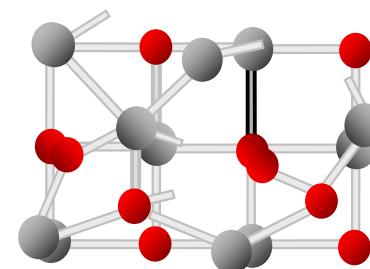
Pulse



Material



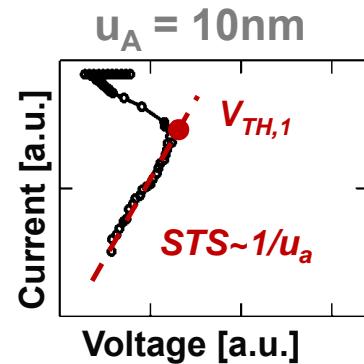
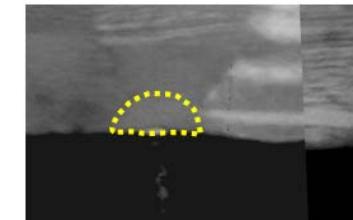
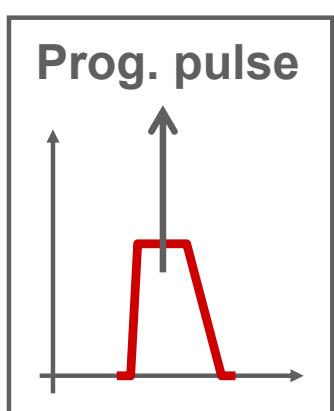
Structure



# PCM – Read Current on Single Cell

- The Reset state is controlled by the thickness of the amorphous region and the activation energy for conduction of the drift phenomenon:

$$I_{READ}(V) = J_{00} \cdot A \cdot e^{-\frac{E_A}{kT} \left(1 - \frac{T}{T_{MN}}\right)} \exp \left[ \frac{q}{kT} \sqrt{\frac{q \cdot V_A}{\pi \epsilon_{GST} u_a}} \left(1 - \frac{T}{T_{MN}}\right) \right]$$

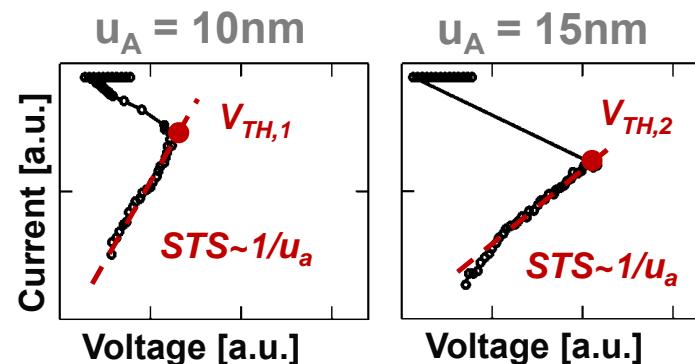
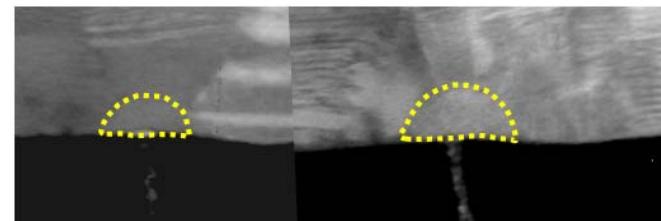
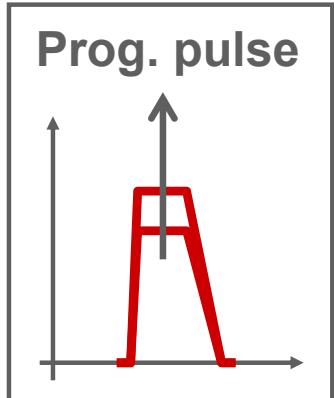


Reference [21 - A. Calderoni, et al.,  
“Physical Modeling and Control of  
Switching Statistics in PCM Arrays”]

# PCM – Read Current on Single Cell

- The Reset state is controlled by the thickness of the amorphous region and the activation energy for conduction of the drift phenomenon:

$$I_{READ}(V) = J_{00} \cdot A \cdot e^{-\frac{E_A}{kT}\left(1 - \frac{T}{T_{MN}}\right)} \exp\left[\frac{q}{kT} \sqrt{\frac{q \cdot V_A}{\pi \epsilon_{GST} u_a}} \left(1 - \frac{T}{T_{MN}}\right)\right]$$

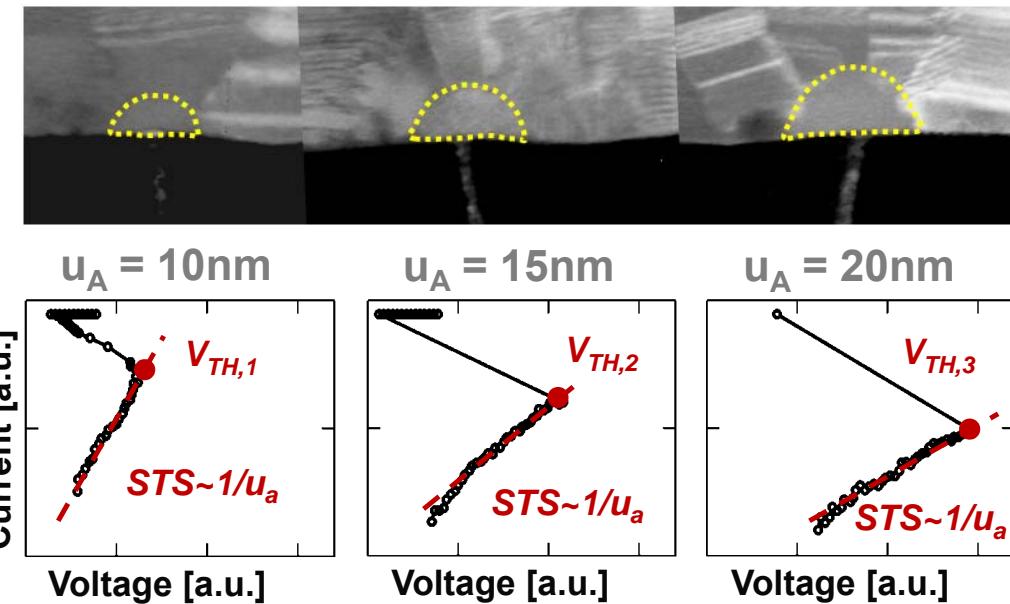
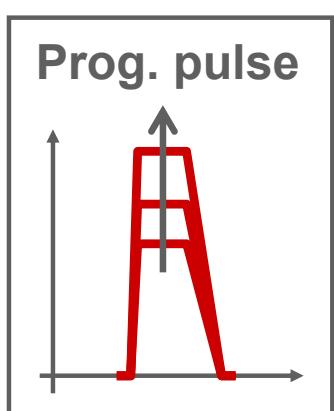


Reference [21 - A. Calderoni, et al.,  
“Physical Modeling and Control of  
Switching Statistics in PCM Arrays”]

# PCM – Read Current on Single Cell

- The Reset state is controlled by the thickness of the amorphous region and the activation energy for conduction of the drift phenomenon:

$$I_{READ}(V) = J_{00} \cdot A \cdot e^{-\frac{E_A}{kT}\left(1-\frac{T}{T_{MN}}\right)} \exp\left[\frac{q}{kT} \sqrt{\frac{q \cdot V_A}{\pi \epsilon_{GST} u_a}} \left(1 - \frac{T}{T_{MN}}\right)\right]$$

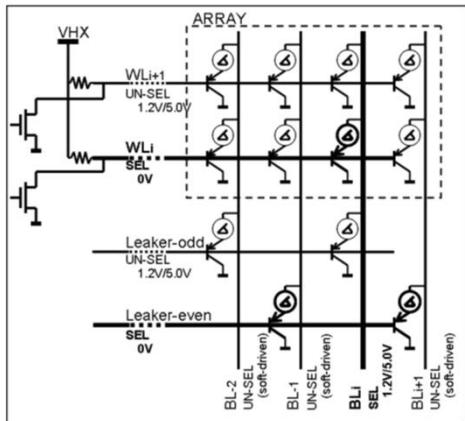


Reference [21 - A. Calderoni, et al.,  
“Physical Modeling and Control of  
Switching Statistics in PCM Arrays”]

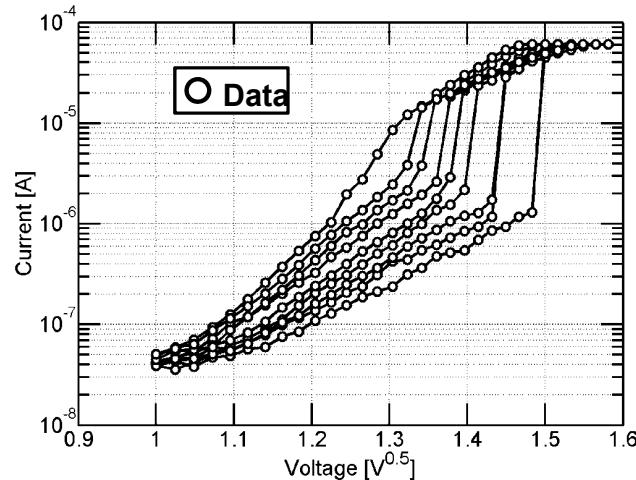
# PCM – Read Current Variability

- Single cell data is not sufficient to evaluate array performances
- Process, parasitic and cell variability have to be taken into account when evaluating device performances

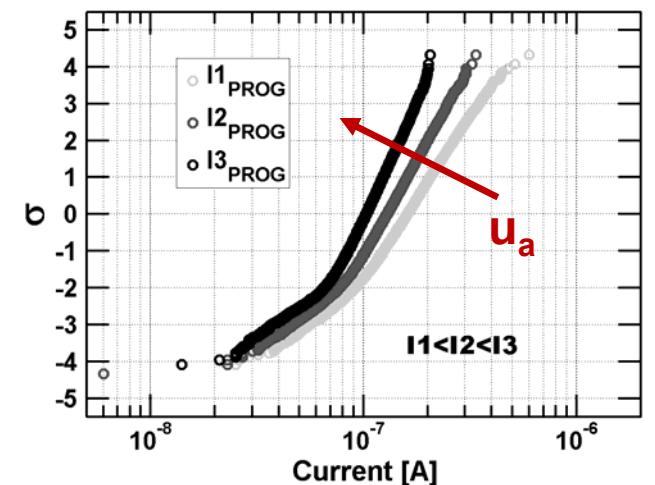
Array Architecture



Array Characterization



Current Distributions vs.  $I_{PROG}$

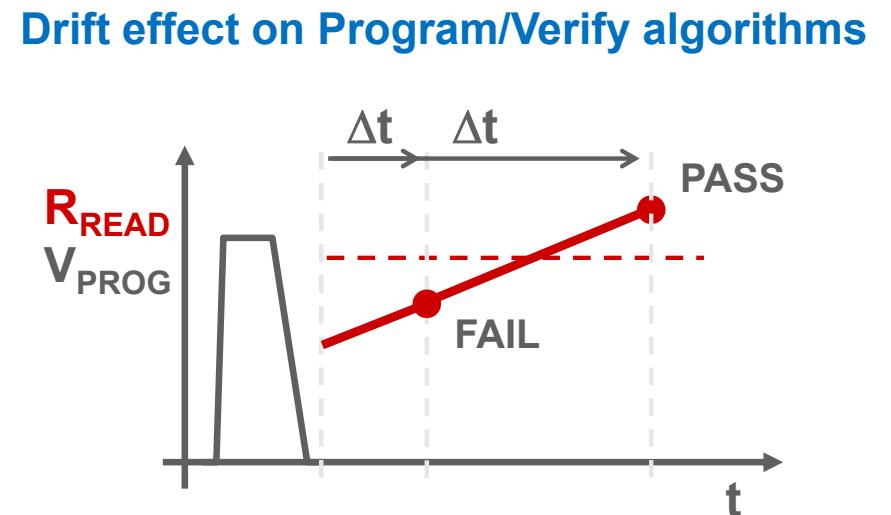
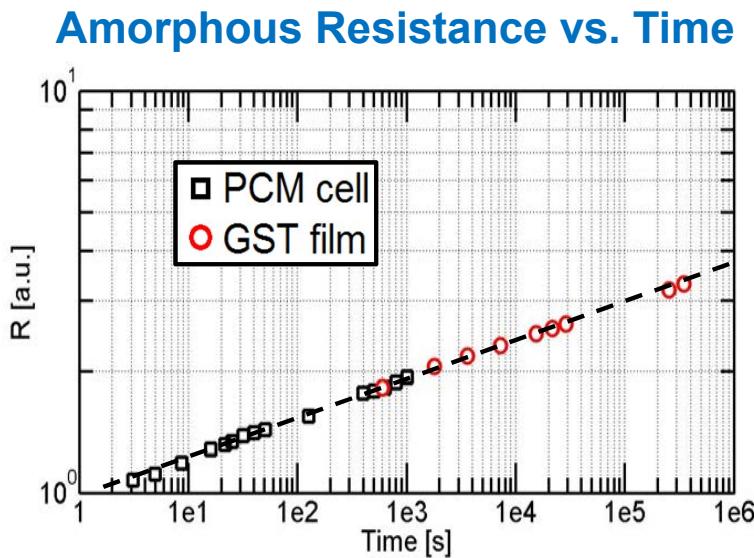


Reference [21 - A. Calderoni, et al., "Physical Modeling and Control of Switching Statistics in PCM Arrays"]

# PCM – Drift on Single Cell

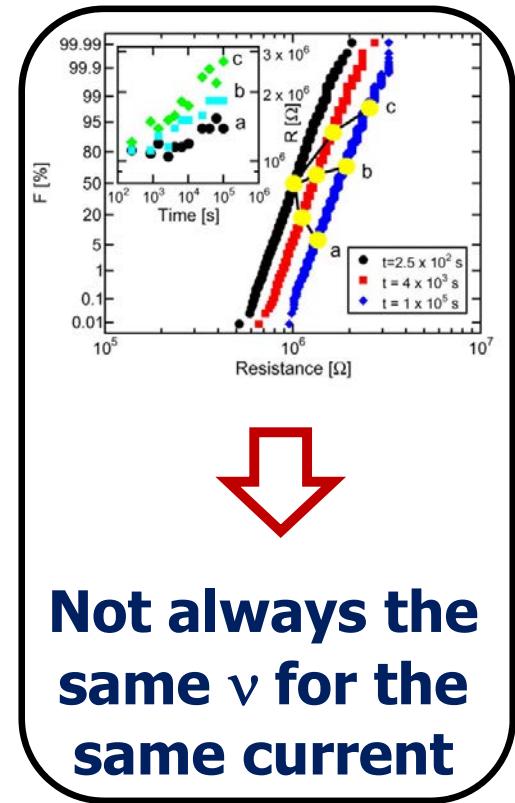
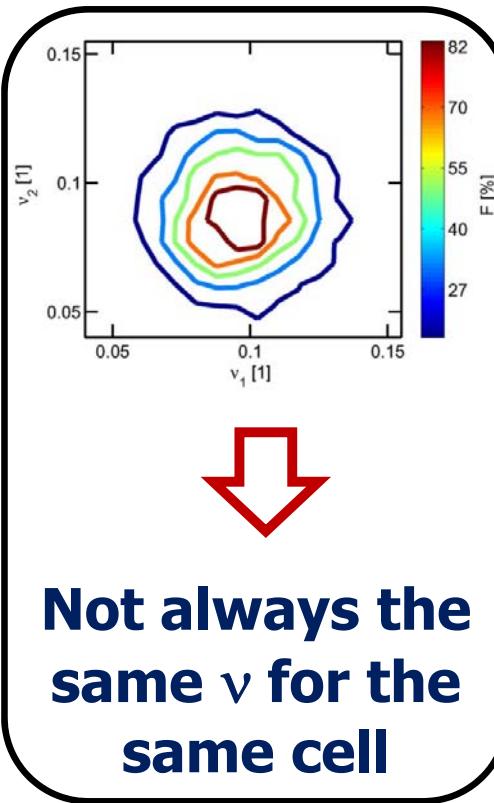
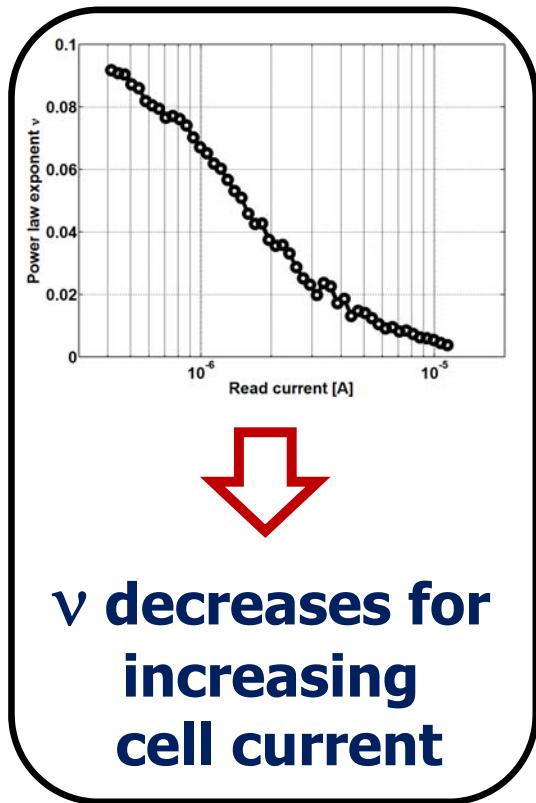
- In the Reset state, the resistance drifts towards higher values with time, following a power-law:
- Problem for Program/Verify algorithms and MLC

$$R(t) = R_0 \left( \frac{t}{t_0} \right)^\nu, \quad \nu \approx 0.1$$



# PCM – Drift Variability

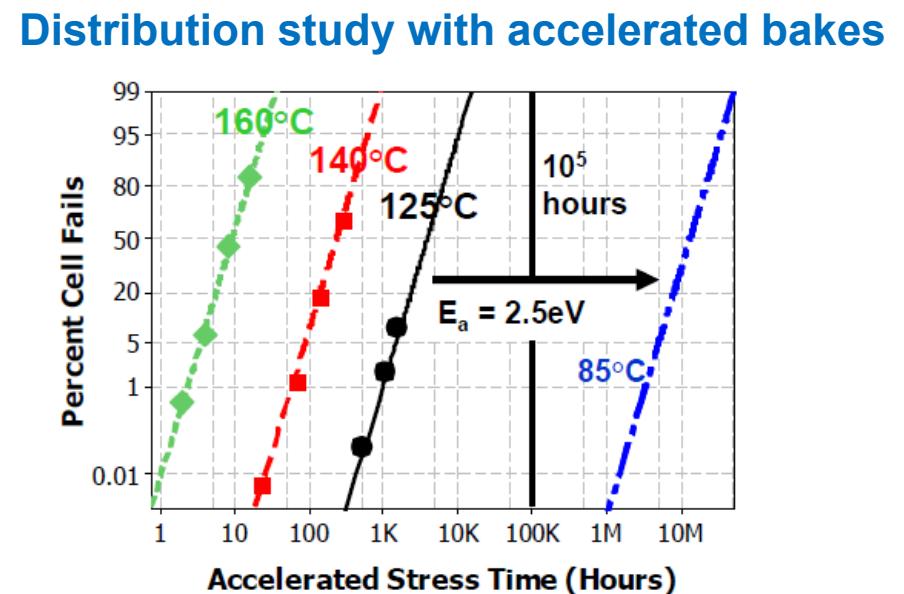
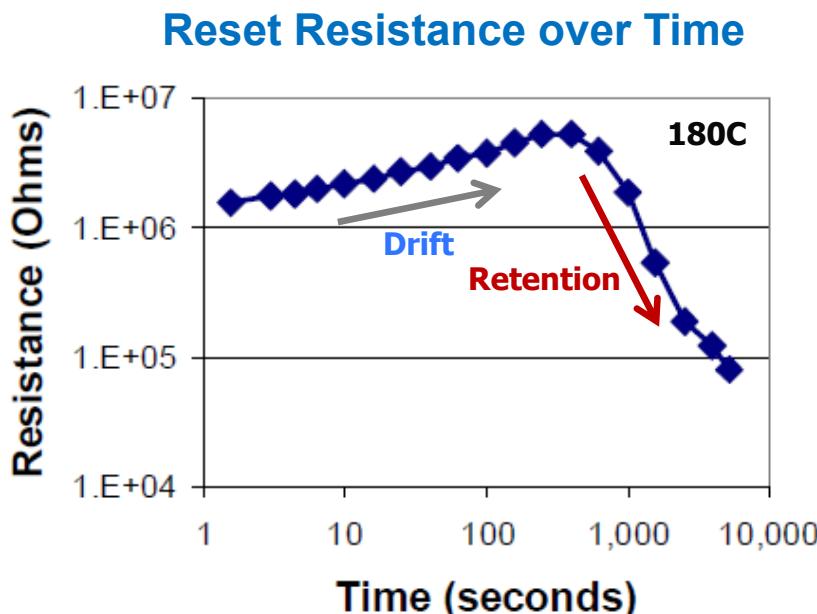
- Reset current distributions are affected by drift variability



Reference [22,23 - M. Boniardi, et al., "Physical origin of the resistance drift exponent in amorphous phase change materials"<sup>45</sup>  
M. Boniardi, et al., "Statistics of Resistance Drift Due to Structural Relaxation in Phase-Change Memory Arrays"]

# PCM - Retention

- The crystalline state (set) is stable
- The amorphous state (reset) initially drift to higher resistance and then data is lost due to a premature crystallization.

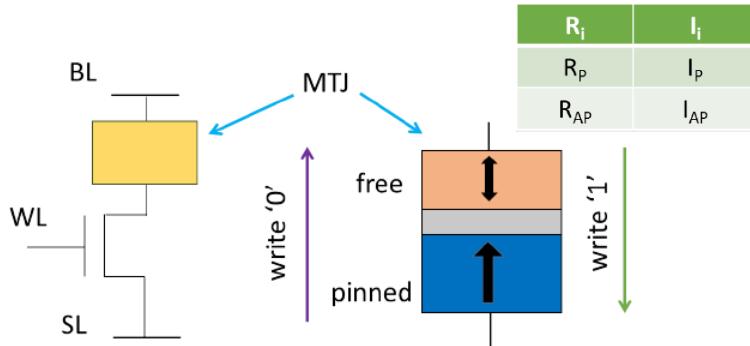


Reference [22,23 - M. Boniardi, et al., "Physical origin of the resistance drift exponent in amorphous phase change materials"<sup>46</sup>  
M. Boniardi, et al., "Statistics of Resistance Drift Due to Structural Relaxation in Phase-Change Memory Arrays"]

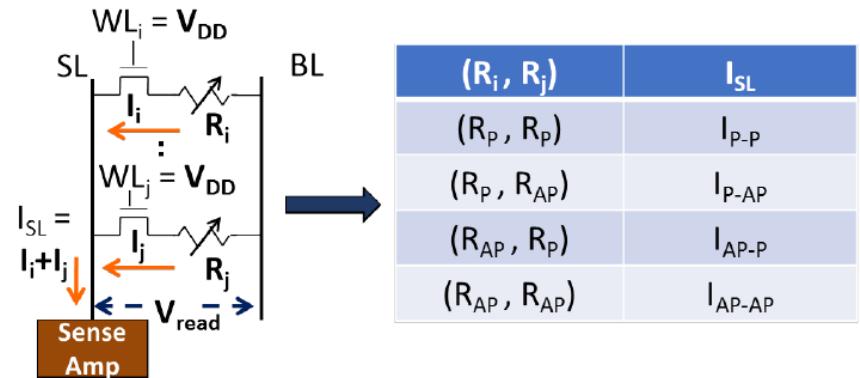
# MRAM – A Compute-In-Memory Example

- Example of how to improve system performance and energy using a Standard Spin-Torque Transfer MRAM (STT-MRAM) array with no changes to the bit cells, but
  - Modifying periphery/decoders/SAs to perform arithmetic and vector operations
  - Using ECC to address reliability and process variations
  - Using extended instruction set and on-chip bus

**STT-MRAM bit cell operation**



**Compute-In-Memory Overview**

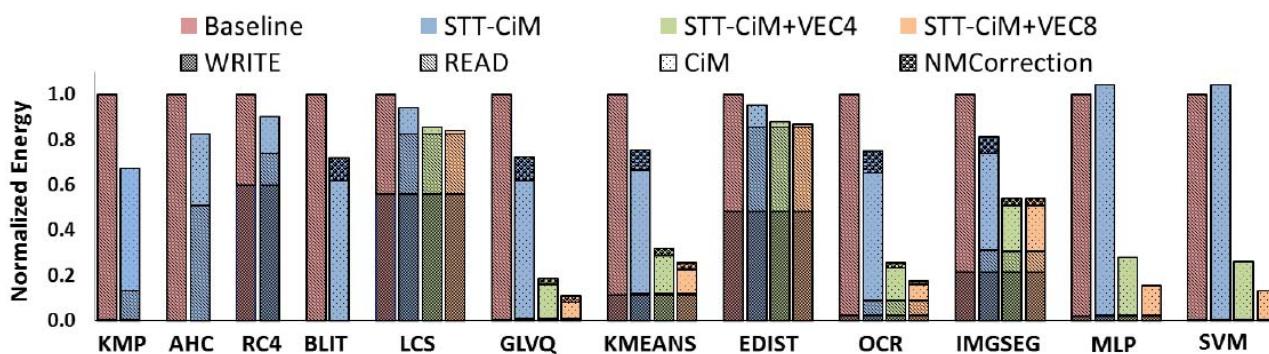


Reference [24 - S. Jain, et al., "Computing in Memory with Spin-Transfer Torque Magnetic RAM"]

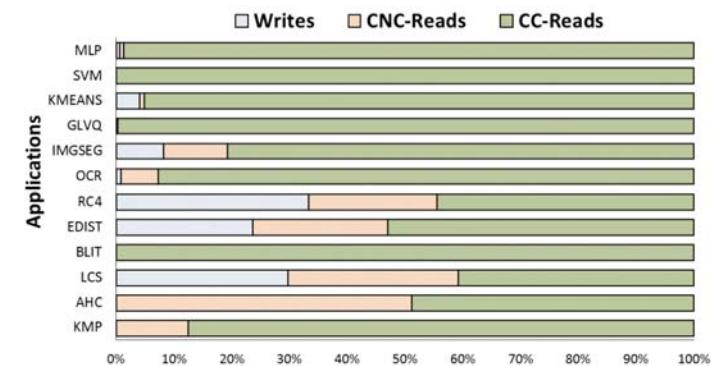
# MRAM – A Compute-In-Memory Example

- System-level energy and performance simulation results of STT-CiM used as a scratchpad in the memory hierarchy of a programmable processor
  - 1.26x / 2.77x / 3.83x average improvement in energy for STT-CiM for different design
  - 1.07x to 1.36x average improvement in speed without vector operations and up to 3.25x and 3.93x for vector lengths of 4 and 8

Application-level Memory Energy



Memory Access Breakdown



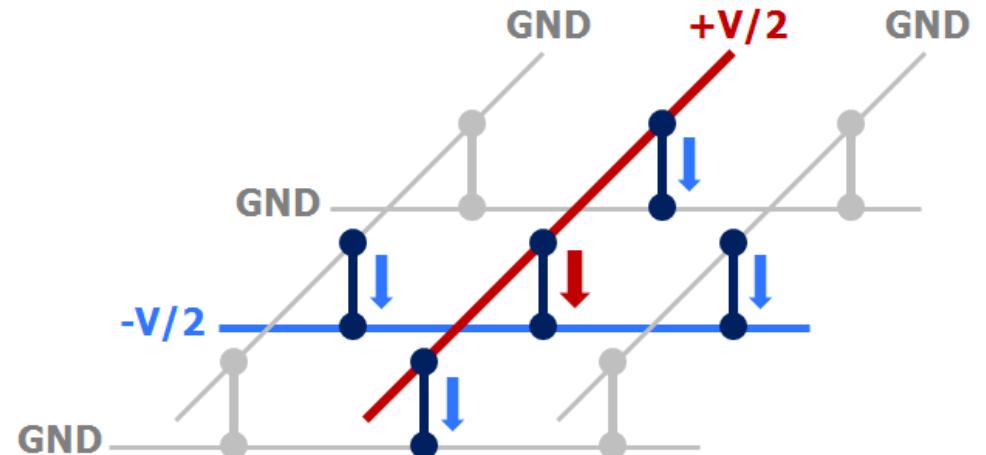
Reference [24 - S. Jain, et al., "Computing in Memory with Spin-Transfer Torque Magnetic RAM"]

# Outline

- Introduction
- Technology Scaling and Computing Systems Trends
- Memory Scaling Challenges
- Emerging Memory and **Selectors**
- Conclusions

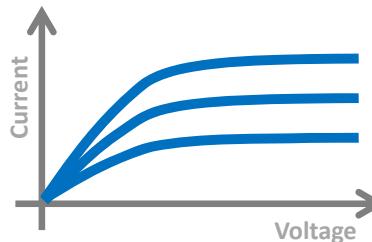
# Why do we need a selector?

- To achieve high-density, cross-point array configuration is required
- Unselected cells along the selected DL and WL will leak during read and write operations → **IR drops**
- IR drops will cause:
  - Higher power
  - Lower effective Cell bias
  - Systematic variations (i.e. near/away from the driver)
  - Limited control of  $I_{CELL}/V_{CELL}$



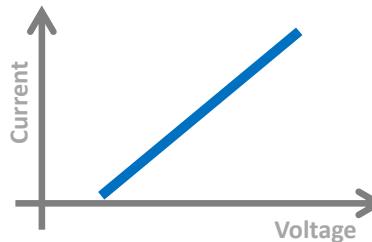
# Selector types

## Transistor 3-Terminal



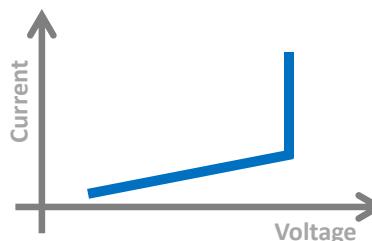
- High ON current to program
- Low OFF current to suppress sneak paths
- Need a third terminal! Usually is  $6F^2$ .
- High performance can be achieved

## Non-Thresholding 2-Terminal



- Achieving High ON current and simultaneously achieving low off current is difficult with these selectors
- Moderate performance is expected

## Thresholding 2-Terminal

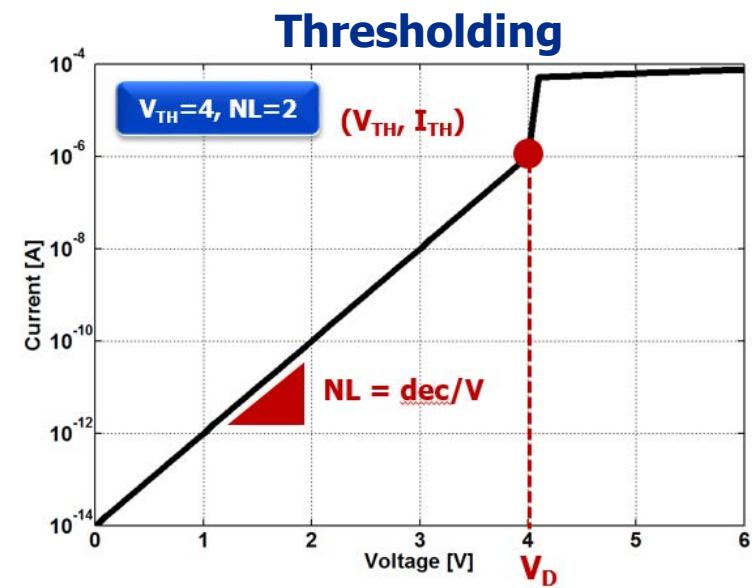
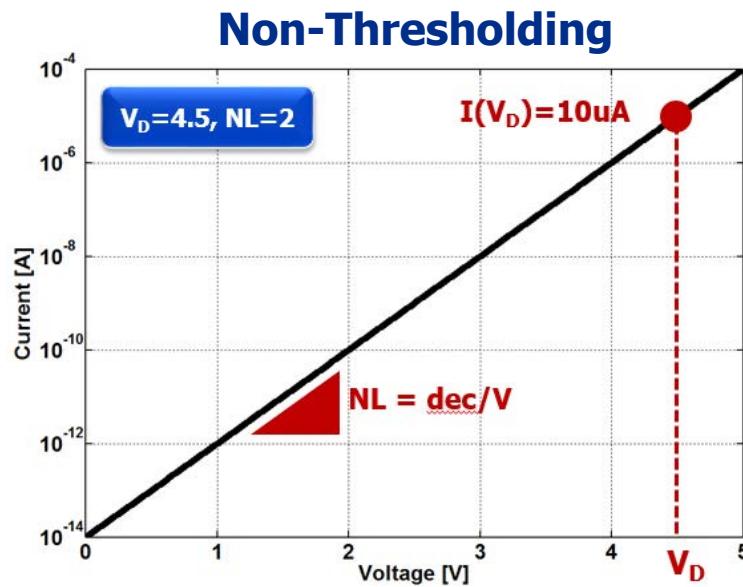


- Minimum foot-print ( $4F^2$ )
- Both high ON current and low OFF current can be achieved with relatively low non-linearity
- High performance can be achieved

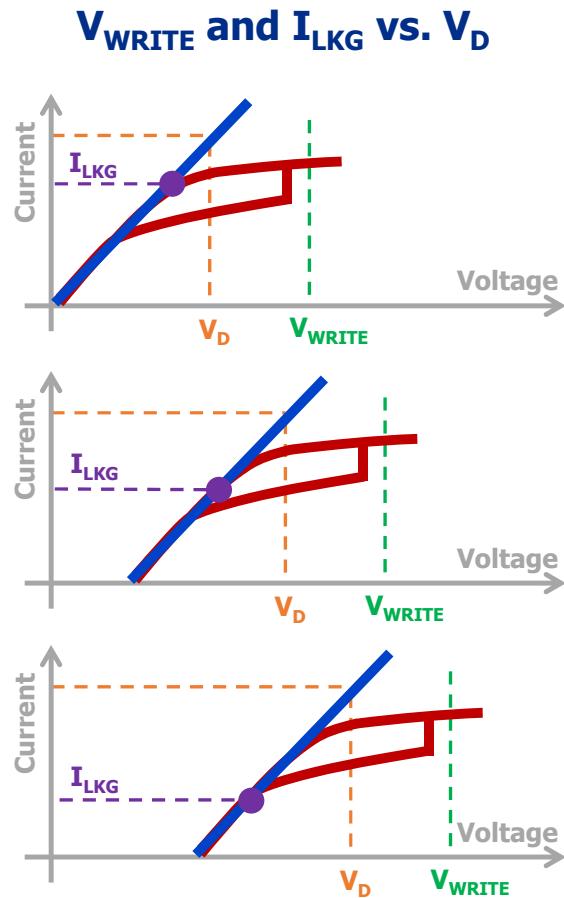
# Selectors Parameters

Reference [25 - N. Ramaswamy, et al., "3D ReRAM: Crosspoint Memory Technologies"]

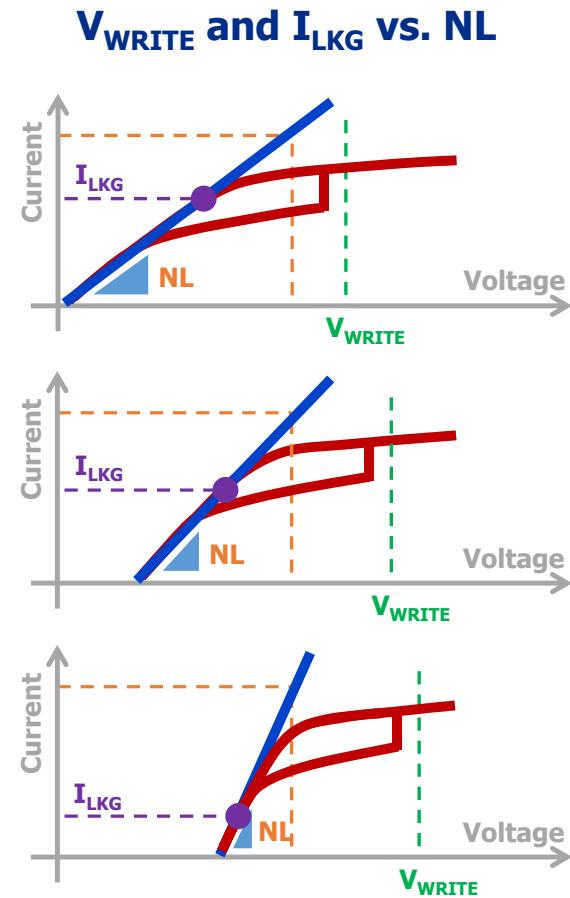
- $NL$  is the non-linearity factor and it's a measure of the 'steepness' of the selector
- $V_D$  is the voltage at a fixed current



# Effect of $V_D$ and NL

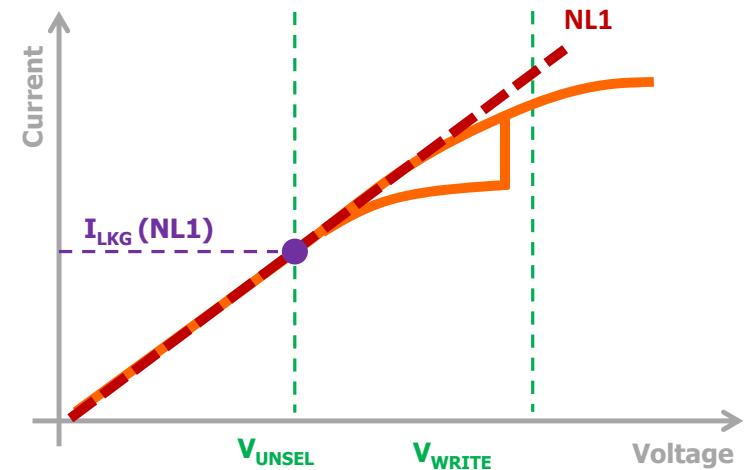


LEAKAGE



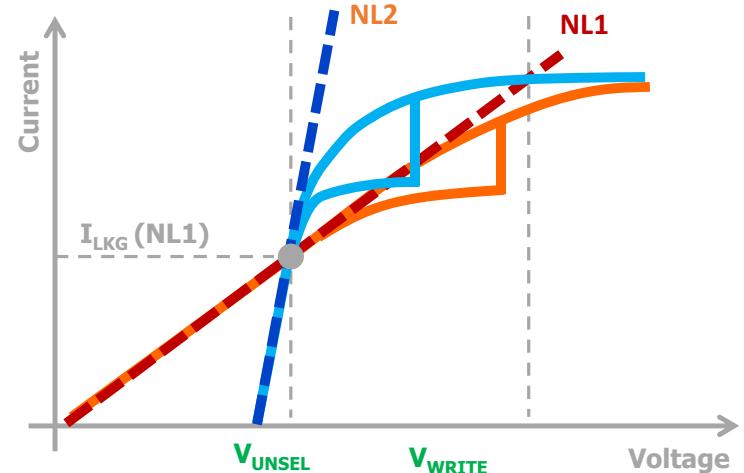
# Selector Evaluation: constant $I_{LKG}$ example

- For a given  $V_D$  and NL, the minimum between  $V_{WRITE}$  and  $V_{READ}$  determines  $V_{UNSEL}$  hence  $I_{LKG}$



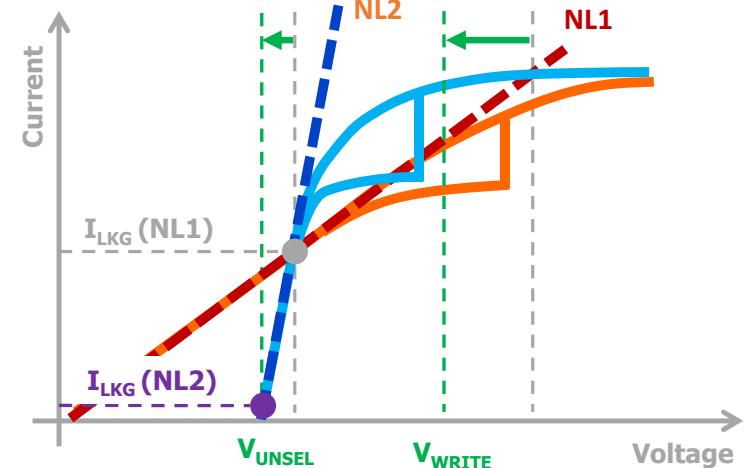
# Selector Evaluation: constant $I_{LKG}$ example

- For a given  $V_D$  and NL, the minimum between  $V_{WRITE}$  and  $V_{READ}$  determines  $V_{UNSEL}$  hence  $I_{LKG}$
- A selector with a higher NL and lower  $V_D$  can yield the same  $I_{LKG}$



# Selector Evaluation: constant $I_{LKG}$ example

- For a given  $V_D$  and NL, the minimum between  $V_{WRITE}$  and  $V_{READ}$  determines  $V_{UNSEL}$  hence  $I_{LKG}$
- A selector with a higher NL and lower  $V_D$  can yield the same  $I_{LKG}$
- In this case,  $V_{WRITE}$  and  $V_{UNSEL}$  can actually be lowered thus yielding a lower  $I_{LKG}$  and better power performances

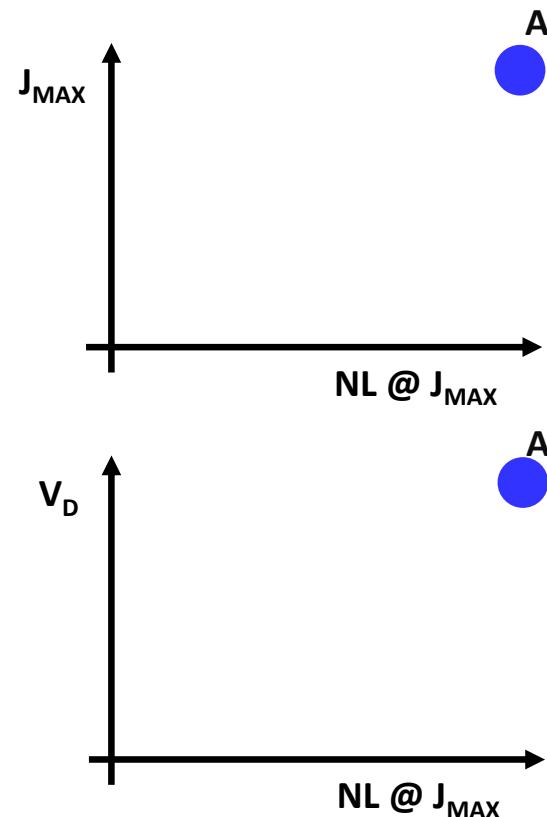
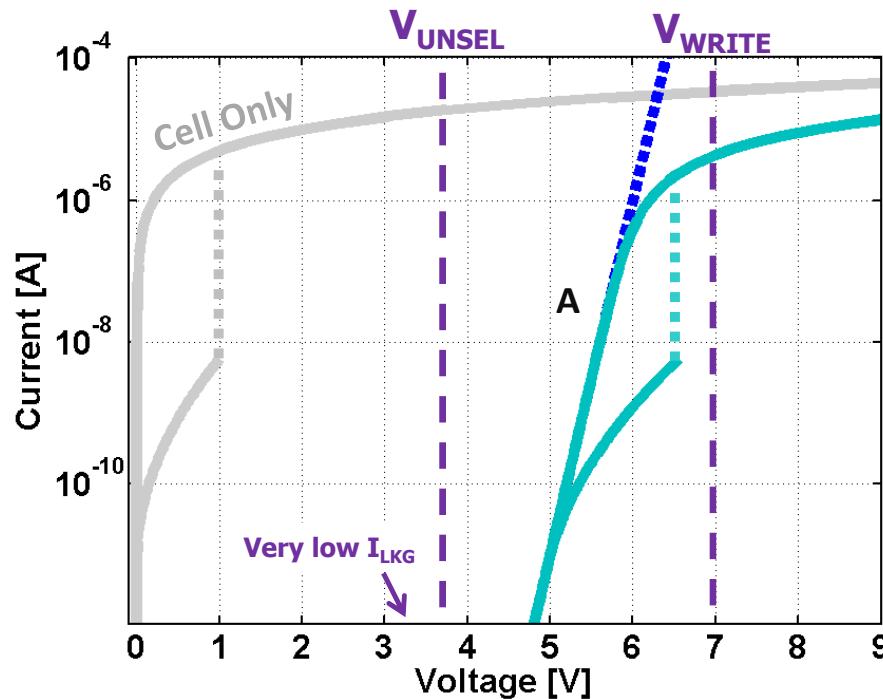


High NL → Lower Leakage  
and Lower Power

Low NL → High V Diode →  
High Voltage CMOS → Higher Power

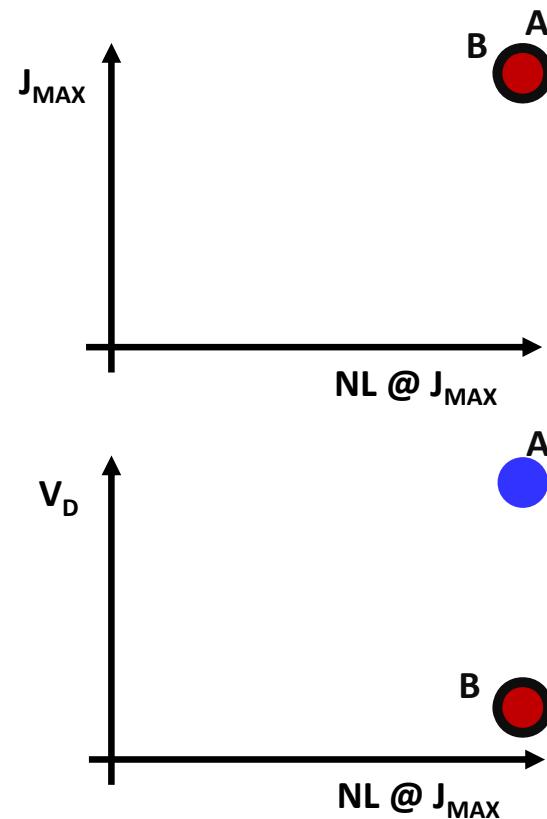
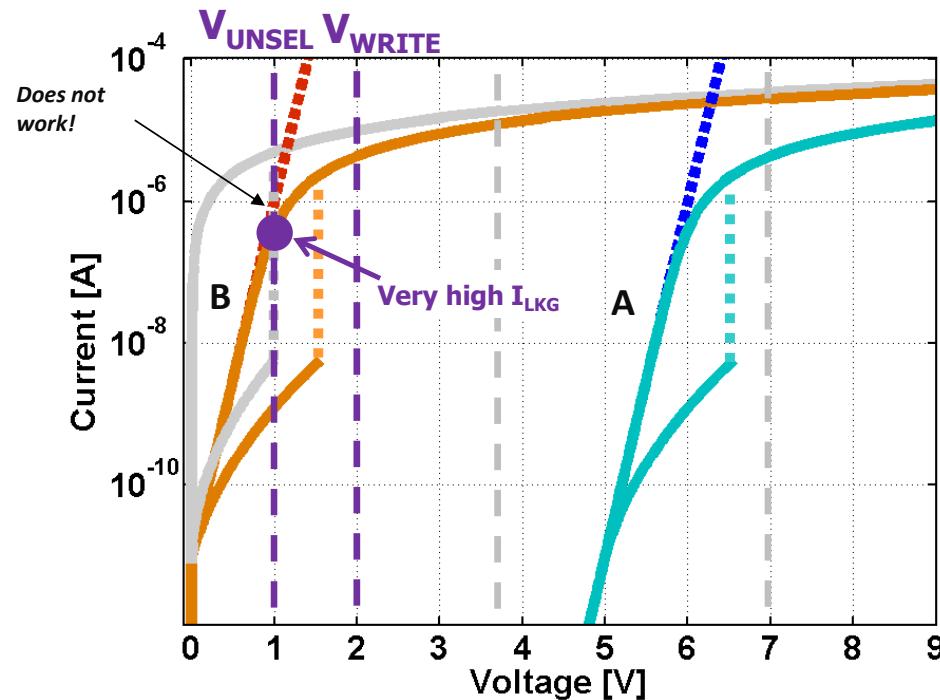
# Does higher NL always work?

- Selector A results in a working cross-point
- Selector B does not work!
  - Need higher  $V_D$  for this selector



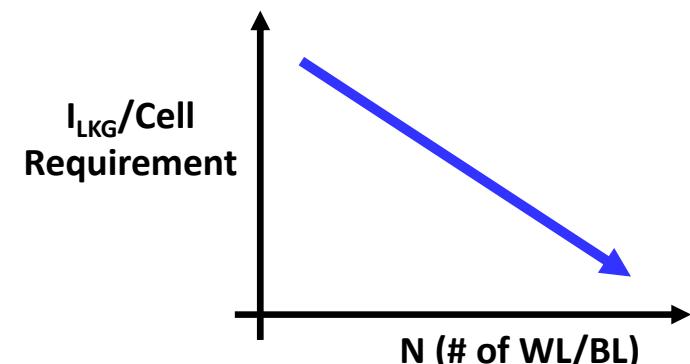
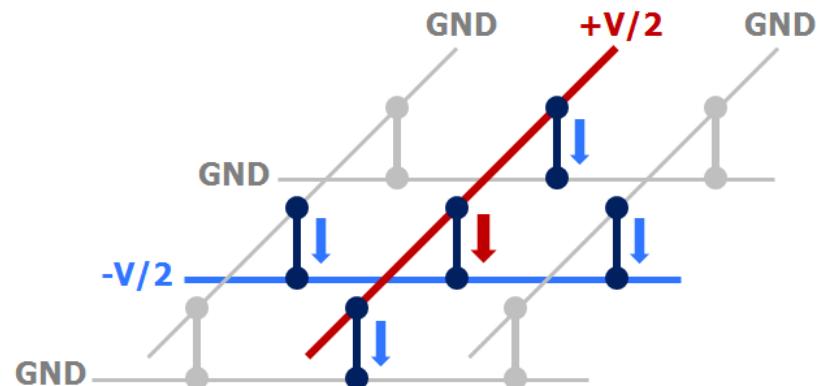
# Does higher NL always work?

- Selector A results in a working cross-point
- Selector B does not work!
  - Need higher  $V_D$  for this selector



# Leakage Requirements vs. Array Size

- The total leakage from unselected cells sets the maximum array size
- To the first order, and not accounting for any variability:
  - During a **WRITE** operation:
    - $I_{LKG} \times N < 10\% \text{ of } I_{CC}$
  - During a **READ** operation:
    - $I_{LKG} \times N < 10\% \text{ of } I_{READ}$
- Cell, Selector and Array variability require a much lower  $I_{LKG}$



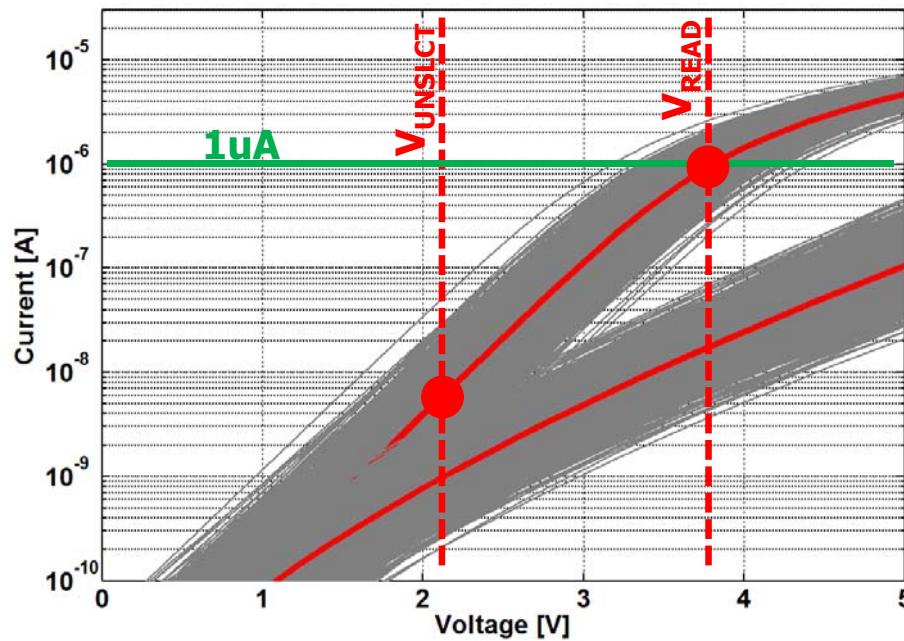
# Non-Thresholding Selector: Ideal Requirements

- $V_{DIODE}$  (V):
  - Upper boundary defined by maximum write bias defined by CMOS:
    - $V_{DIODE} (@ 10\mu A) < 7V$  to have  $V_{MAX} < 10V$  (example)
  - Lower boundary defined by maximum leakage during program and by the read window; also a function of NL:
    - $V_{DIODE} > 4$  (with  $NL > 4$ ) to have  $I_{LKG} < 10\mu A$  and 10x window (example)
- NL (dec/V):
  - Lower boundary defined by maximum leakage during program or read (window margin).
  - $NL > 4$  to have balanced lkg during write and read (example)
- These are ideal requirements to have a working Selector + Cell ‘on paper’ assuming **NO VARIABILITY**
- **How does variability change these ideal selector requirements?**

# Non-Thresholding Selector: Variability

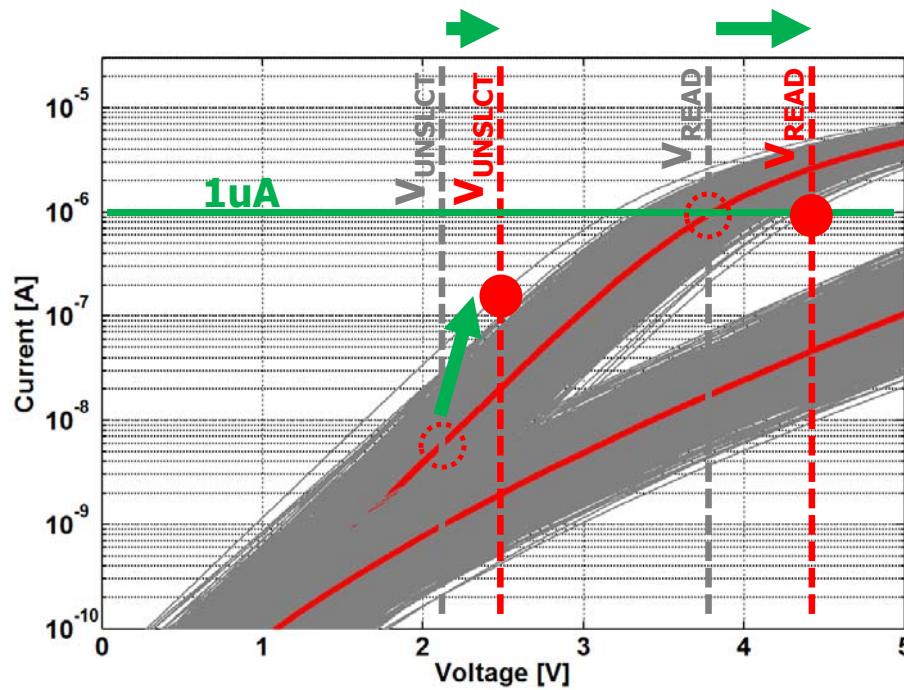
- Without variability:  $V_{\text{READ}}$  set to read 1uA

→ at  $V_{\text{UNSELECT}} = V_{\text{READ}}/2$ ,  $I_{\text{LKG}} \sim 6\text{E-}9$



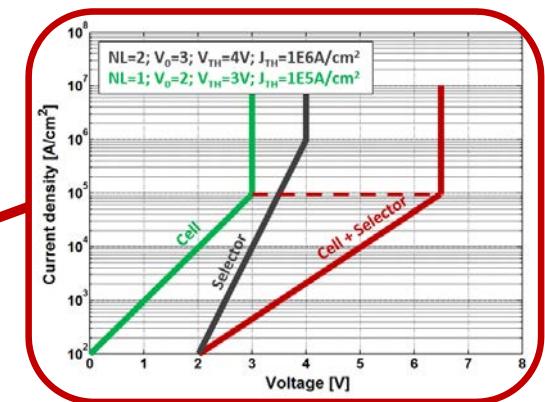
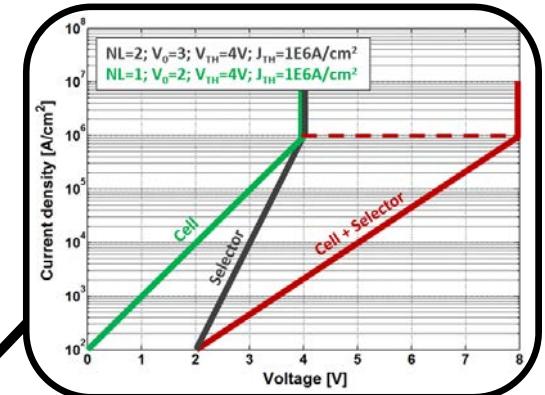
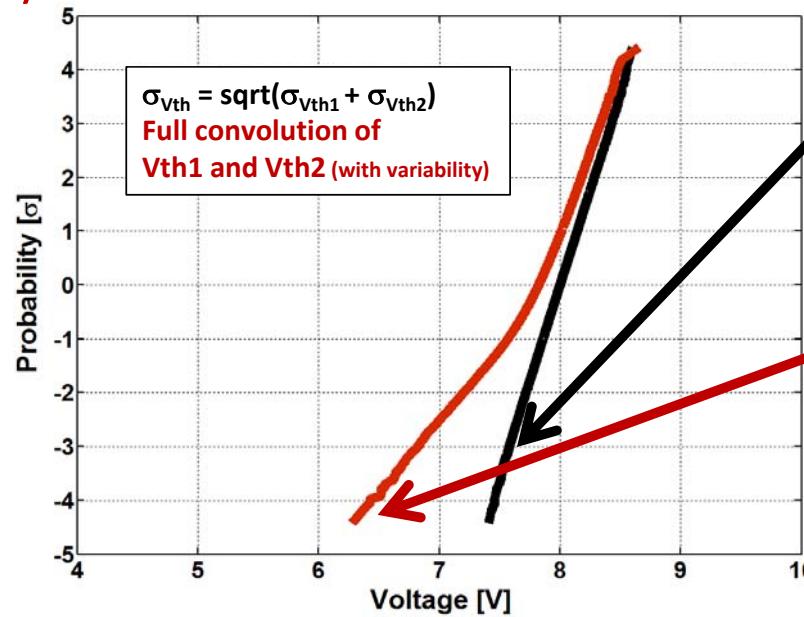
# Non-Thresholding Selector: Variability

- Without variability:  $V_{\text{READ}}$  set to read 1uA  
→ at  $V_{\text{UNSELECT}} = V_{\text{READ}}/2$ ,  $I_{\text{LKG}} \sim 6\text{E-}9$
- With variability:  $V_{\text{READ}}$  set to read 1uA  
→ at  $V_{\text{UNSELECT}} = V_{\text{READ}}/2$ ,  $I_{\text{LKG}} \sim 2\text{E-}7$   
( $>10x$  increase)



# Thresholding Selector: Variability

- A simple convolution of Cell and Selector  $V_{TH}$  underestimate the final variability
- Components such as  $I_{TH}$  mismatch between Cell and Selector have to be properly accounted for to get the correct final variability

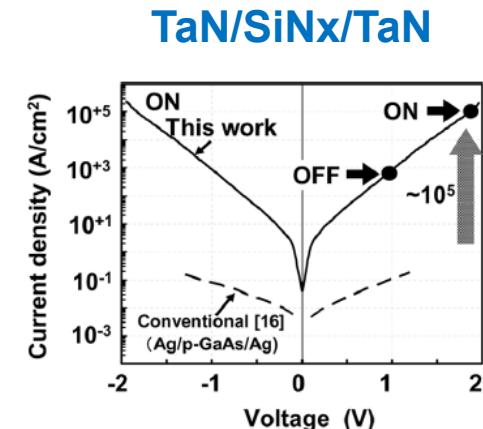
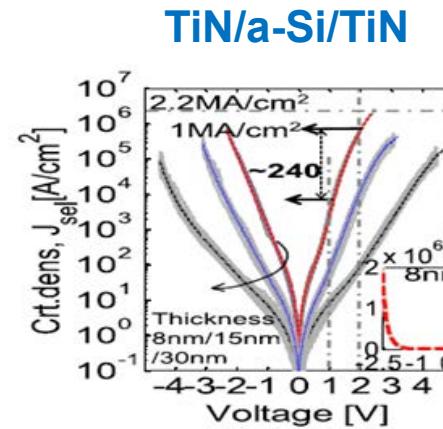
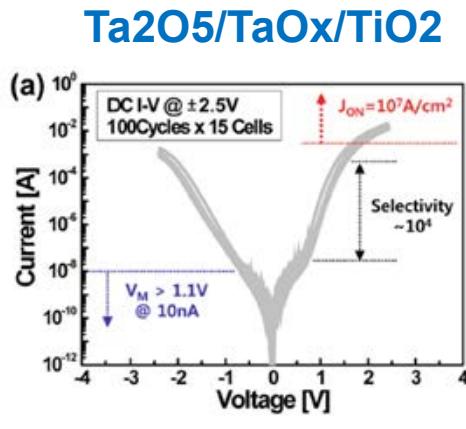
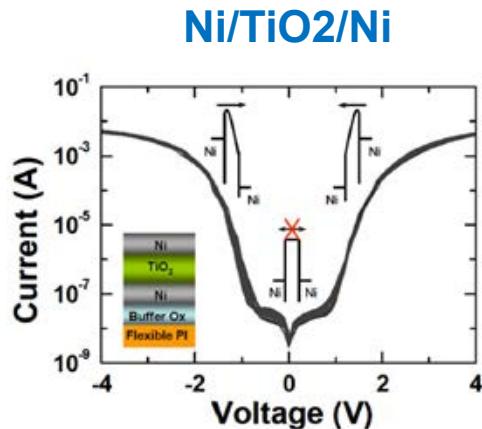


# 2-Terminal – Selector Types

- Barrier Engineering
  - Metal-Semiconductor-Metals, Metal-Insulator-Metals etc.,
- Thresholding
  - Chalcogenide based selectors
  - Thresholding oxides (metal – insulator transitions)
- Volatile Switches
  - Metal Ion motion

# Barrier Engineered Selectors

- Band engineered systems have been demonstrated with several material sets
- Typical issues are: low drive and worse endurance at higher drive currents for scaled selectors

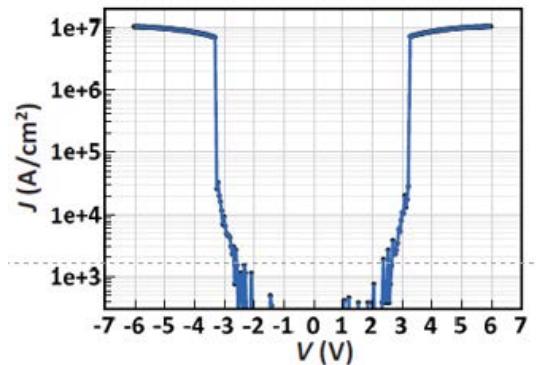


Reference [26,27,28,29 - J.-J. Huang, et al., "Bipolar nonlinear Ni/TiO<sub>2</sub>/Ni selector for 1S1R crossbar array applications" - J. Woo, et al., "Multi-layer tunnel barrier (Ta<sub>2</sub>O<sub>5</sub>/TaO<sub>x</sub>/TiO<sub>2</sub>) engineering for bipolar RRAM selector applications" - L. Zhang, et.al, "Ultrathin Metal/Amorphous-Silicon/Metal Diode for Bipolar RRAM Selector Applications" - A. Kawahara, et al., "An 8 Mb Multi-Layered Cross-Point ReRAM Macro With 443 MB/s Write Throughput"]

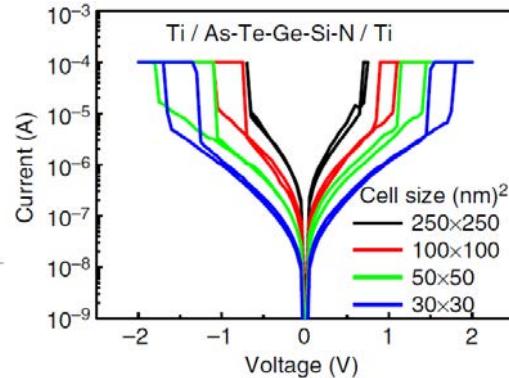
# Thresholding Selectors

- Chalcogenide selectors as well as some oxide based selectors exhibit threshold switching
- Typical issues are: low  $V_{DIODE}$ ,  $V_{TH}/I_{TH}$  variability and turn on/off times

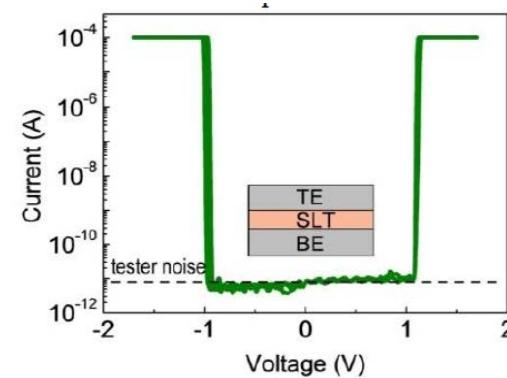
**BC –based Selector**



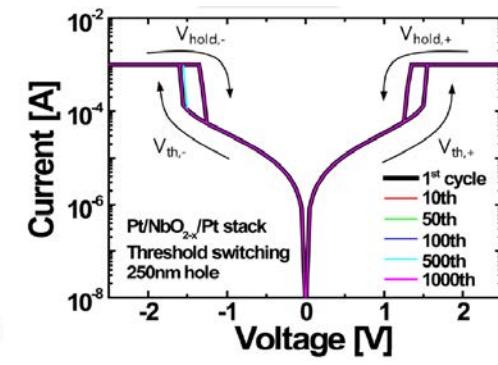
**As-Te-Ge-Si-N**



**“FAST” Selector**



**Pt/NbO<sub>2-x</sub>/Pt**



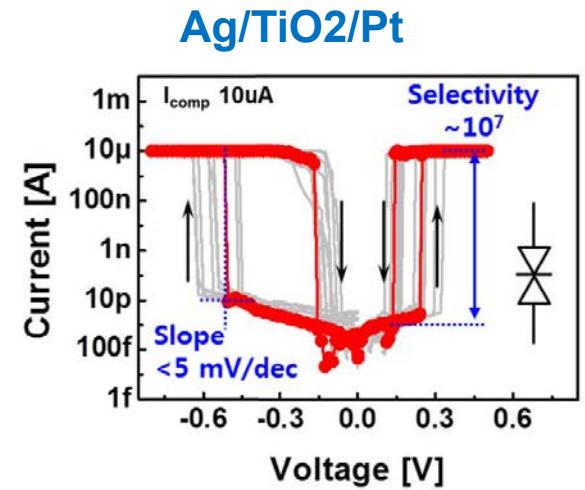
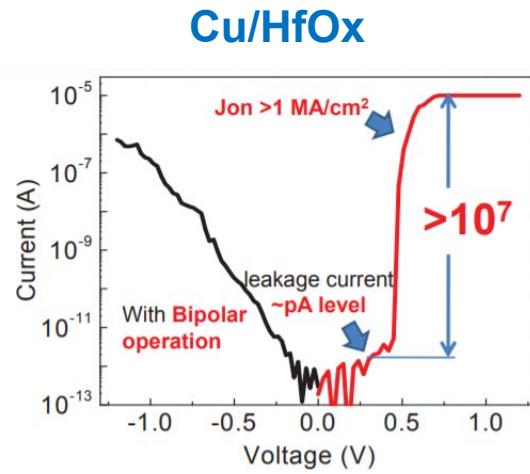
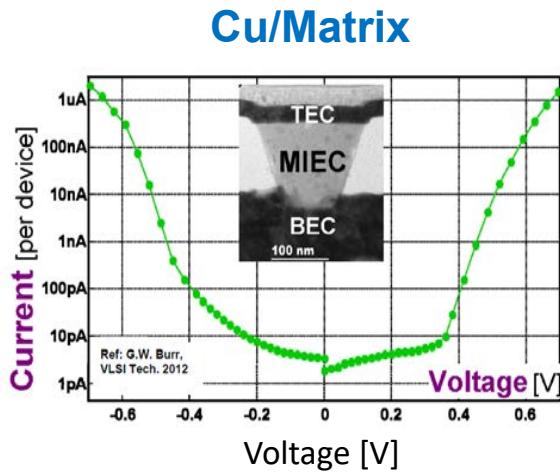
Reference [30,31,32,33] - S. Yasuda, et al., “A Cross Point Cu-ReRAM with a Novel OTS Selector for Storage Class Memory Applications” -

M.-J. Lee, “A plasma-treated chalcogenide switch device for stackable scalable 3D nanoscale memory” - S. H. Jo, et al., “3D-stackable crossbar resistive memory based on field assisted superlinear threshold (FAST) selector” - S. Kim, et al., “Ultrathin (<10nm) Nb<sub>2</sub>O<sub>5</sub>/NbO<sub>2</sub> Hybrid Memory with Both Memory and Selector Characteristics for

High Density 3D Vertically Stackable RRAM Applications”]

# Volatile Switches

- Volatile switches with metallic ion motion can enable steep slopes (NL)
- Typical issues are: low  $V_{DIODE}$ ,  $V_{TH}/I_{TH}$  variability and turn on/off times



Reference [34,35,36 - R. S. Shenoy, "Challenges for selector devices" - Q. Luo, et al., "Cu BEOL compatible selector with high selectivity (>10<sup>7</sup>), Extremely low Off current (~pA) and High Endurance (>10<sup>10</sup>)" - J. Song, et al., "Threshold Selector With High Selectivity and Steep Slope for Cross-Point Memory Array"]

# Outline

- Introduction
- Technology Scaling and Computing Systems Trends
- Memory Scaling Challenges
- Emerging Memory and Selectors
- **Conclusions**

# Conclusions

- Memory has become central to modern systems, from mobile devices to servers, often defining the system cost and performance
- Scaling of DRAM continues as innovations push out the perceived scaling wall, but fundamental economic-driven limitations are being approached
- Emerging Memories are being explored to supplement DRAM and NAND:
  - these memories begin to close the gaps in the memory hierarchy, but they do not challenge NAND cost or DRAM performance
- Modern system architectures are not exploiting the full performance and energy capabilities of the memory:
  - performance and energy are being constrained by the narrow link between the memory and the processor
- Addressing this issue through better integration of compute and memory is a significant opportunity

# References [1/2]

- [1] <https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.html>
- [2] <https://www.cisco.com/c/en/us/solutions/service-provider/vni-network-traffic-forecast/vni-forecast-info.html>
- [3] <https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html>
- [4] <https://www.youtube.com/intl/en-GB/yt/about/press/>
- [5] <https://blog.hubspot.com/marketing/youtube-stats>
- [6] M. Horowitz, "Computing's energy problem (and what we can do about it)", DOI: [10.1109/ISSCC.2014.6757323](https://doi.org/10.1109/ISSCC.2014.6757323)
- [7] S. DeBoer, "Memory Technology: The Core to Enable Future Computing Systems", DOI: [10.1109/VLSIT.2018.8510707](https://doi.org/10.1109/VLSIT.2018.8510707)
- [8] D. Pandiyan, et al., "Quantifying the energy cost of data movement for emerging smart phone workloads on mobile platforms", DOI: [10.1109/IISWC.2014.6983056](https://doi.org/10.1109/IISWC.2014.6983056)
- [9] G. Singh, et al., "A Review of Near-Memory Computing Architectures: Opportunities and Challenges", DOI: [10.1109/DSD.2018.00106](https://doi.org/10.1109/DSD.2018.00106)
- [10] N. Ramaswamy, et al., "Metal Gate Recessed Access Device (RAD) for DRAM Scaling", DOI: [10.1109/WMED.2007.368056](https://doi.org/10.1109/WMED.2007.368056)
- [11] VLSI
- [12] R. Clark, et al., "Perspective: New process technologies required for future devices and scaling", DOI: [10.1063/1.5026805](https://doi.org/10.1063/1.5026805)
- [13] [https://www.micron.com/-/media/client/global/documents/products/product-flyer/flyer\\_lpdram\\_mobile\\_embedded.pdf](https://www.micron.com/-/media/client/global/documents/products/product-flyer/flyer_lpdram_mobile_embedded.pdf)
- [14] [https://www.micron.com/-/media/client/global/documents/products/product-flyer/gddr6\\_product\\_flyer.pdf](https://www.micron.com/-/media/client/global/documents/products/product-flyer/gddr6_product_flyer.pdf)
- [15] [https://www.micron.com/-/media/client/global/documents/products/product-flyer/rldram\\_flyer.pdf](https://www.micron.com/-/media/client/global/documents/products/product-flyer/rldram_flyer.pdf)
- [17] <https://www.micron.com/about/blog/2017/february/the-advantage-of-ecc-dram-in-smartphones>
- [18] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices", DOI: [10.1038/s41928-018-0092-2](https://doi.org/10.1038/s41928-018-0092-2)
- [19] A. Calderoni, et al., "Performance Comparison of O-based and Cu-based ReRAM for High-Density Applications", DOI: [10.1109/IMW.2014.6849351](https://doi.org/10.1109/IMW.2014.6849351)
- [20] A. Calderoni, et. al., "Engineering ReRAM for High-Density Applications", DOI: [10.1016/j.mee.2015.04.044](https://doi.org/10.1016/j.mee.2015.04.044)
- [21] A. Calderoni, et al., "Physical Modeling and Control of Switching Statistics in PCM Arrays", DOI: [10.1109/IMW.2011.5873230](https://doi.org/10.1109/IMW.2011.5873230)
- [22] M. Boniardi, et al., "Physical origin of the resistance drift exponent in amorphous phase change materials", DOI: [10.1063/1.3599559](https://doi.org/10.1063/1.3599559)
- [23] M. Boniardi, et al., "Statistics of Resistance Drift Due to Structural Relaxation in Phase-Change Memory Arrays", DOI: [10.1109/TED.2010.2058771](https://doi.org/10.1109/TED.2010.2058771)
- [23] B. Gleixner, et al., "Data Retention Characterization of Phase-Change Memory Arrays", DOI: [10.1109/RELPHY.2007.369948](https://doi.org/10.1109/RELPHY.2007.369948)
- [24] S. Jain, et al., "Computing in Memory with Spin-Transfer Torque Magnetic RAM", DOI: [10.1109/TVLSI.2017.2776954](https://doi.org/10.1109/TVLSI.2017.2776954)
- [25] N. Ramaswamy, et al., "3D ReRAM: Crosspoint Memory Technologies", IEDM 2017, Short Courses

# References [2/2]

- [26] J.-J. Huang, et al., “Bipolar nonlinear Ni/TiO<sub>2</sub>/Ni selector for 1S1R crossbar array applications”, DOI: [10.1109/LED.2011.2161601](https://doi.org/10.1109/LED.2011.2161601)
- [27] J. Woo, et al., “Multi-layer tunnel barrier (Ta<sub>2</sub>O<sub>5</sub>/TaO<sub>x</sub>/TiO<sub>2</sub>) engineering for bipolar RRAM selector applications”, VLSI, 2013
- [28] L. Zhang, et.al, “Ultrathin Metal/Amorphous-Silicon/Metal Diode for Bipolar RRAM Selector Applications”, DOI: [10.1109/LED.2013.2293591](https://doi.org/10.1109/LED.2013.2293591)
- [29] A. Kawahara, et al., “An 8 Mb Multi-Layered Cross-Point ReRAM Macro With 443 MB/s Write Throughput”, DOI: [10.1109/JSSC.2012.2215121](https://doi.org/10.1109/JSSC.2012.2215121)
- [30] S. Yasuda, et al., “A Cross Point Cu-ReRAM with a Novel OTS Selector for Storage Class Memory Applications”, DOI: [10.23919/VLSIT.2017.7998189](https://doi.org/10.23919/VLSIT.2017.7998189)
- [31] M.-J. Lee, “A plasma-treated chalcogenide switch device for stackable scalable 3D nanoscale memory”, DOI: [10.1038/ncomms3629](https://doi.org/10.1038/ncomms3629)
- [32] S. H. Jo, et al., “3D-stackable crossbar resistive memory based on field assisted superlinear threshold (FAST) selector”, DOI: [10.1109/IEDM.2014.7046999](https://doi.org/10.1109/IEDM.2014.7046999)
- [33] S. Kim, et al., “Ultrathin (<10nm) Nb<sub>2</sub>O<sub>5</sub>/NbO<sub>2</sub> Hybrid Memory with Both Memory and Selector Characteristics for High Density 3D Vertically Stackable RRAM Applications”, DOI: [10.1109/VLSIT.2012.6242508](https://doi.org/10.1109/VLSIT.2012.6242508)
- [34] R. S. Shenoy, “Challenges for selector devices”, IMW 2013 Tutorial
- [35] Q. Luo, et al., “Cu BEOL compatible selector with high selectivity (>10<sup>7</sup>), Extremely low Off current (~pA) and High Endurance (>10<sup>10</sup>)” DOI: [10.1109/IEDM.2015.7409669](https://doi.org/10.1109/IEDM.2015.7409669)
- [36] J. Song, et al., “Threshold Selector With High Selectivity and Steep Slope for Cross-Point Memory Array”, DOI: [10.1109/LED.2015.2430332](https://doi.org/10.1109/LED.2015.2430332)

# 3D-Stacked DRAM Technology & Function-in-Memory Solution

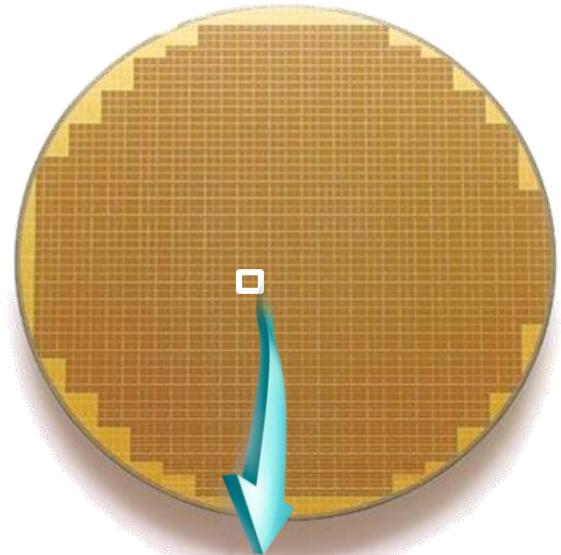
Kyomin Sohn, Ph.D.

*Samsung Electronics*

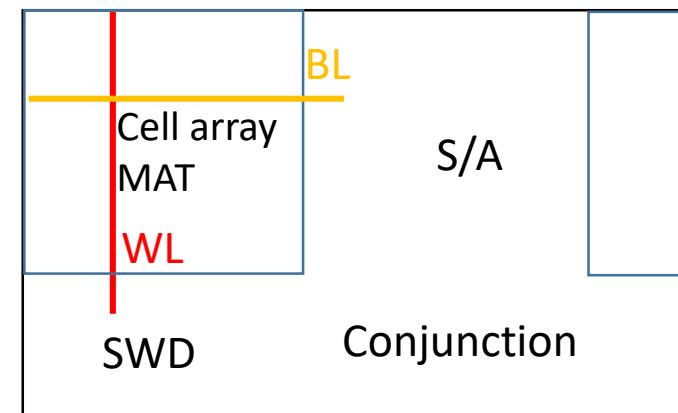
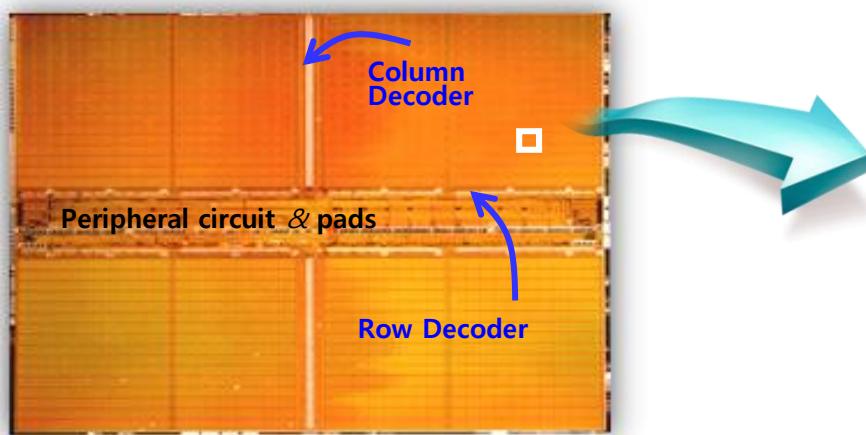
# Outline

- DRAM Technology and Scaling
- 3D-Stacked DRAM Technology
  - Introduction of TSV and 2.5D process
  - Thermal and SI/PI Challenge
- HBM (high bandwidth memory) DRAM
  - Introduction & Architecture
  - Difficulties and Solutions
- Function-in-Memory Solution for AI Application
  - Traditional Memory Solutions
  - FIM (Function-in-Memory) using HBM
- Summary

# DRAM Chip Architecture

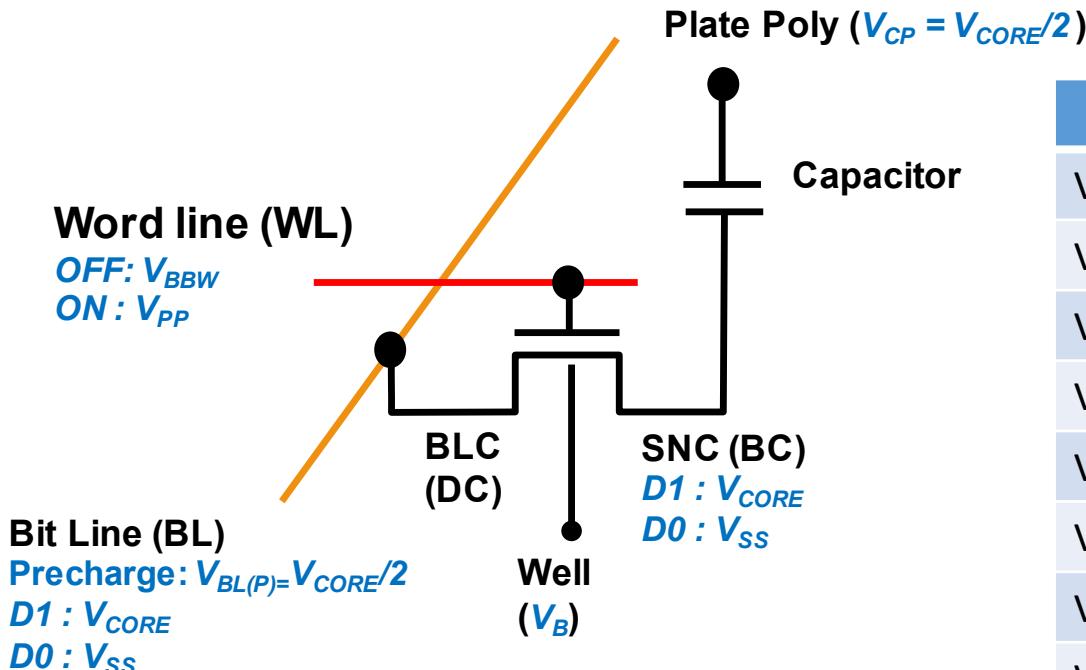


Area	Role	Area Ratio
Cell	<ul style="list-style-type: none"><li>• Data Storage</li></ul>	50~55%
Core : (Row,Col) Decoder, SWD, S/A Conjunction	<ul style="list-style-type: none"><li>• Decoding</li><li>• Data Read/Write</li><li>• Data Restoring</li></ul>	25~30%
Peripheral	<ul style="list-style-type: none"><li>• Control-logic</li><li>• In-out interface</li></ul>	~20%



# DRAM Cell

- VDD is external voltage only supplied from outside of DRAM chip
- Other voltages are internally generated

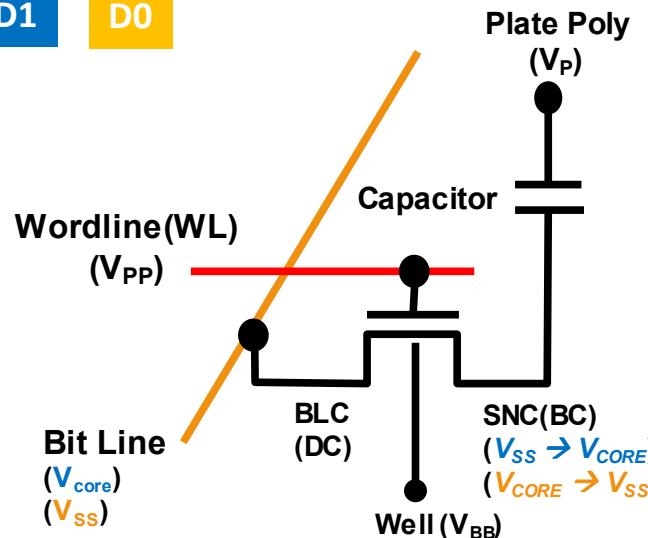


		Level
VDD	External Voltage	1.2V @DDR4
VPP	Word Line Voltage	~3.0V
Vcore	Storage Node Voltage	< 1.2V
Vcp	Capacitor Plate Voltage	$\frac{1}{2}$ Vcore
VBLP	Bit Line Pre-charge Voltage	$\frac{1}{2}$ Vcore
VB	Cell Tr Body Voltage	-0.5~0.8V
VBBW	Negate Word Line Voltage	-0.1~0.4V
VSS	Ground Voltage	0V

[Source: S. Cha , SK Hynix, IEDM 2011]

# Basic Operation : Write

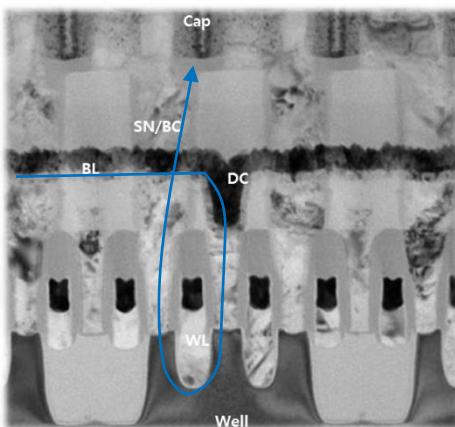
D1      D0



- Cell Transistor should transfer  $V_{CORE}$  to storage node

$$V_{PP} \geq V_{CORE} + V_T + V_{CORE} \times \gamma$$

( $V_T$  : Threshold voltage,  $\gamma$ : body effect,  $V_{CORE}$  : voltage for Data "1" )

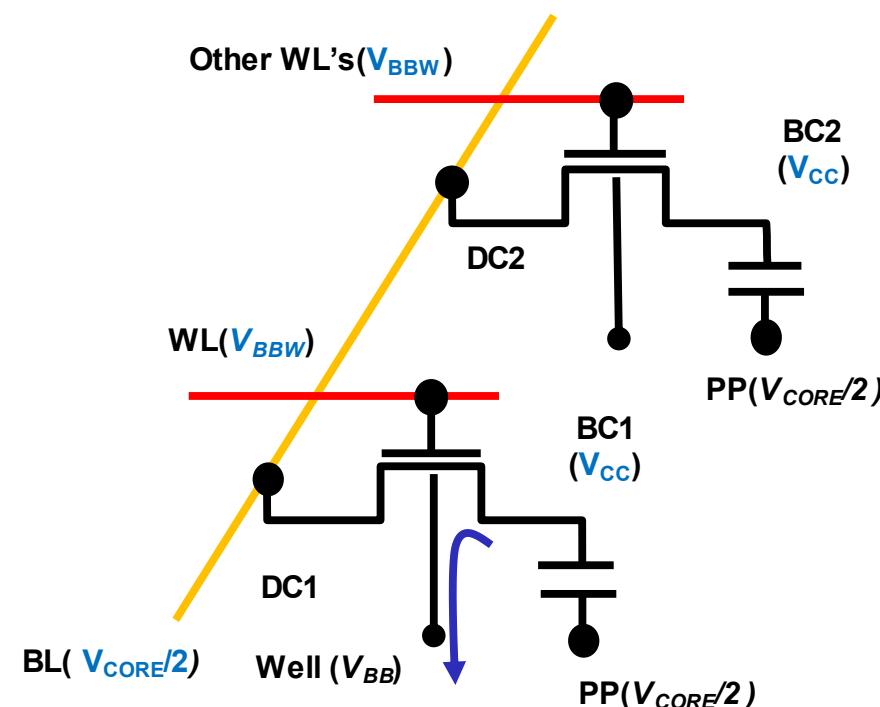


- For faster write;  
increase of  $V_{PP}$ ,  
decrease of  $V_T$ ,  $\gamma$ ,  $V_{CORE}$   
reduction of BL resistance( $R_{BL}$ ), SN  
resistance( $R_{BC}$ ) storage capacitance ( $C_S$ )

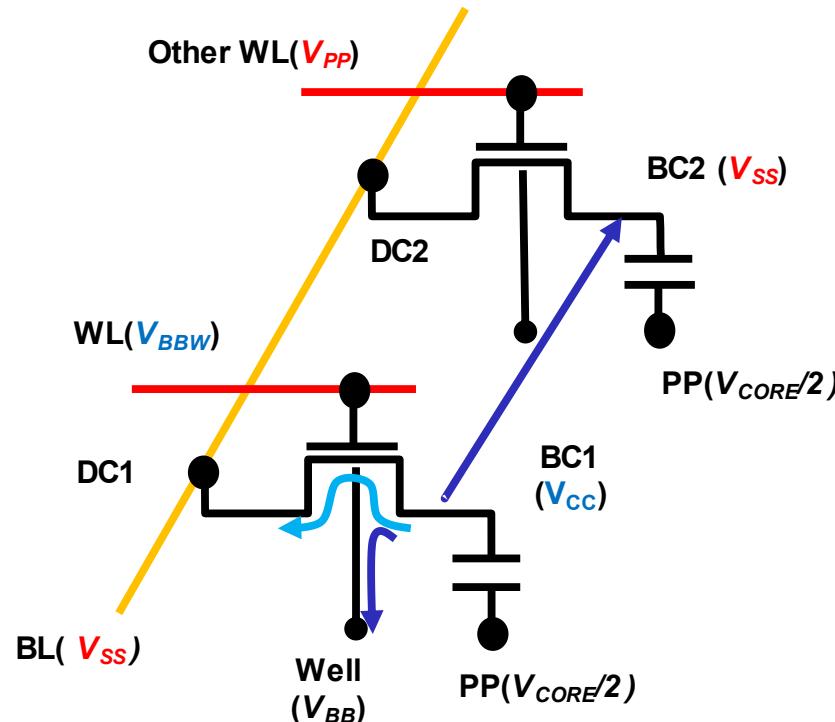
# Basic Operation : Retention

- Retention means how long D1 can be maintained without(static)/with(dynamic) neighbor cell operation

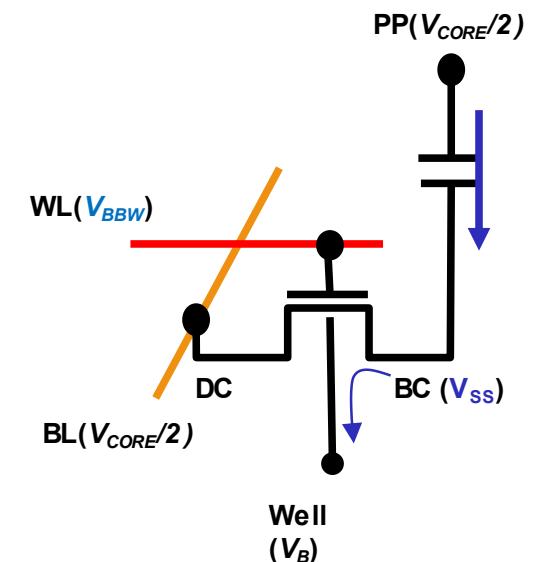
Static retention : Junction + GIDL



Dynamic retention : Junction+ $I_{OFF}$  +Isolation



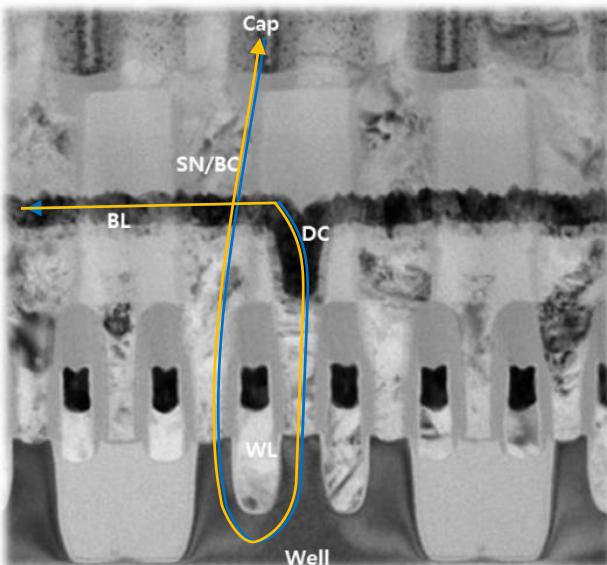
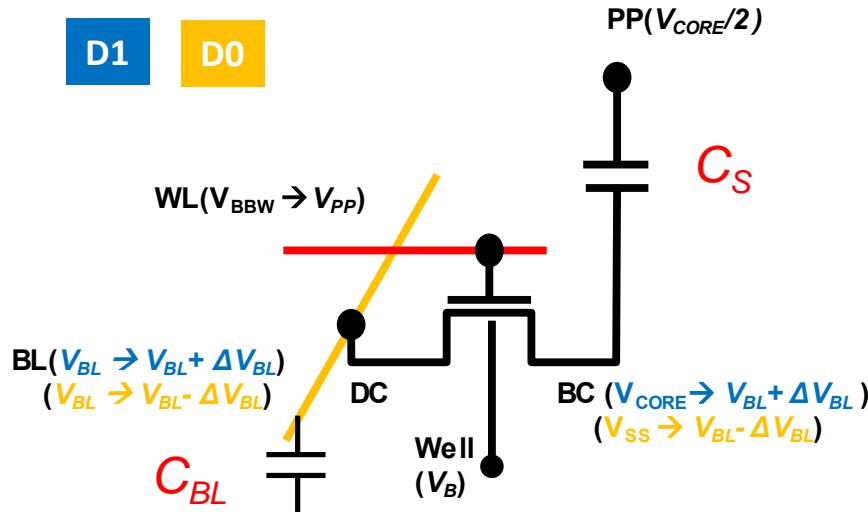
D0 retention :: Cap leakage



# Basic Operation : Read

D1

D0



- During hold, bit-line is pre-charged( $V_{BLP}$ ) and floated
- During *charge sharing*, charges in capacitor flow into bit-line until voltages at both nodes be the same

$$V_{BL} C_{BL} + C_s \cdot V_{CORE} = (C_{BL} + C_s)(V_{BL} + \Delta V_{BL})$$

Pre-charge

$$= (C_{BL} + C_s)(V_{BL} + \Delta V_{BL})$$

D1 hold

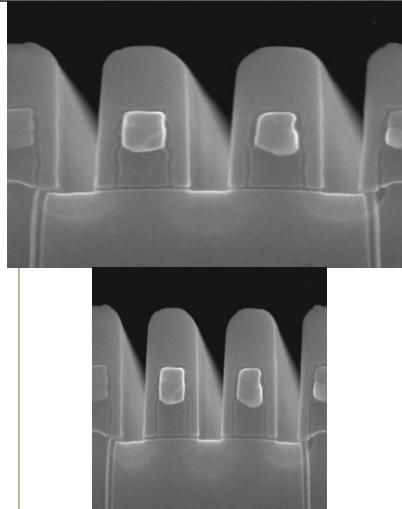
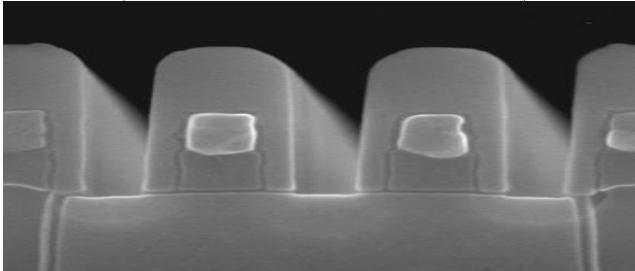
$$= (C_{BL} + C_s)(V_{BL} + \Delta V_{BL})$$

After charge sharing

$$\Delta V_{BL} = \frac{(V_{CORE} - V_{BLP})}{1 + \frac{C_{BL}}{C_s}}$$

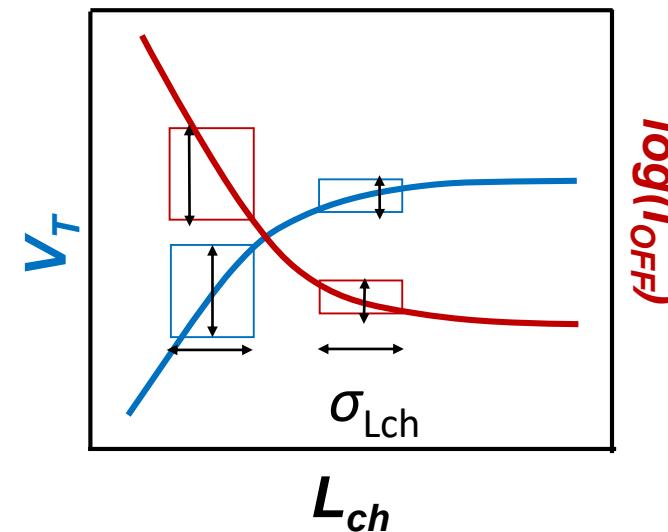
- Storage( $C_s$ ) and bit-line capacitance( $C_{BL}$ ) are major parameters for  $\Delta V_{BL}$

# Cell Transistor Requirements



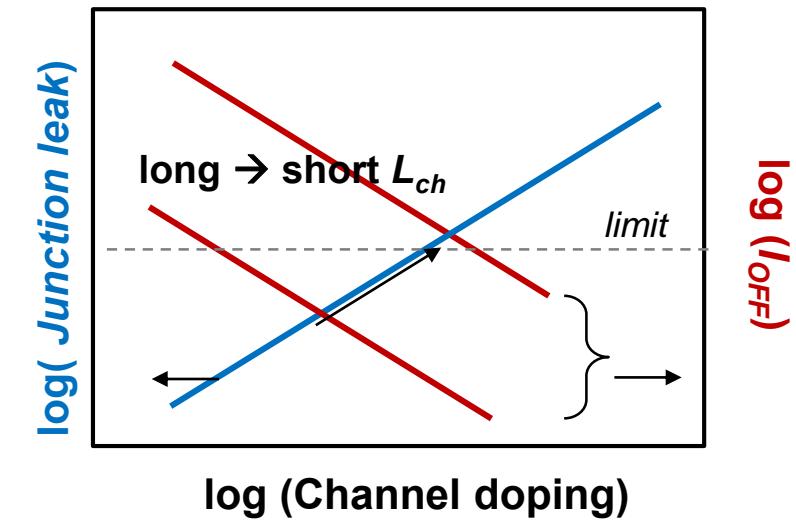
DRAM cell  
scale  
down

Overcome short channel effect to manage  $I_{OFF}$

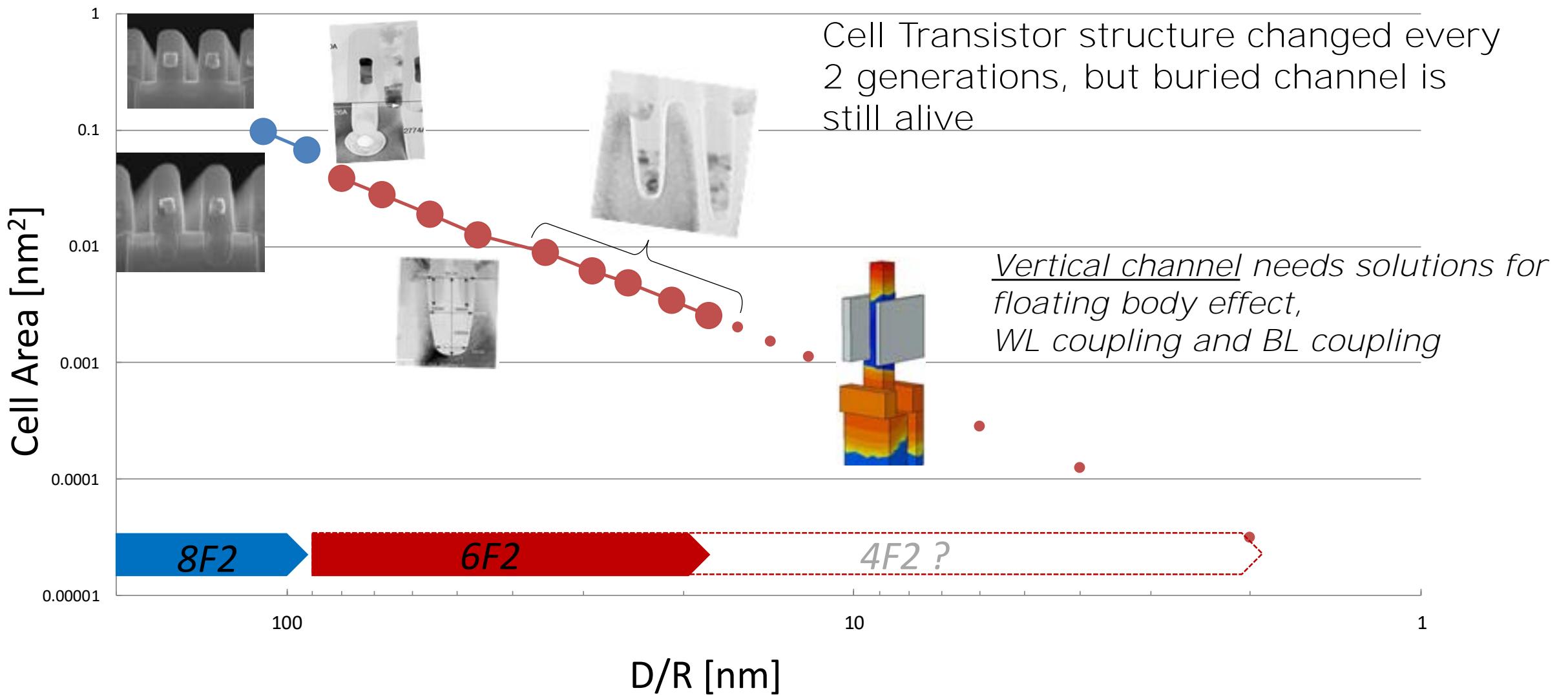


Cell Transistor should be *long channel !!*

Thinner gate oxide → TDDB  
Shallow junction depth and high channel doping → Junction leakage

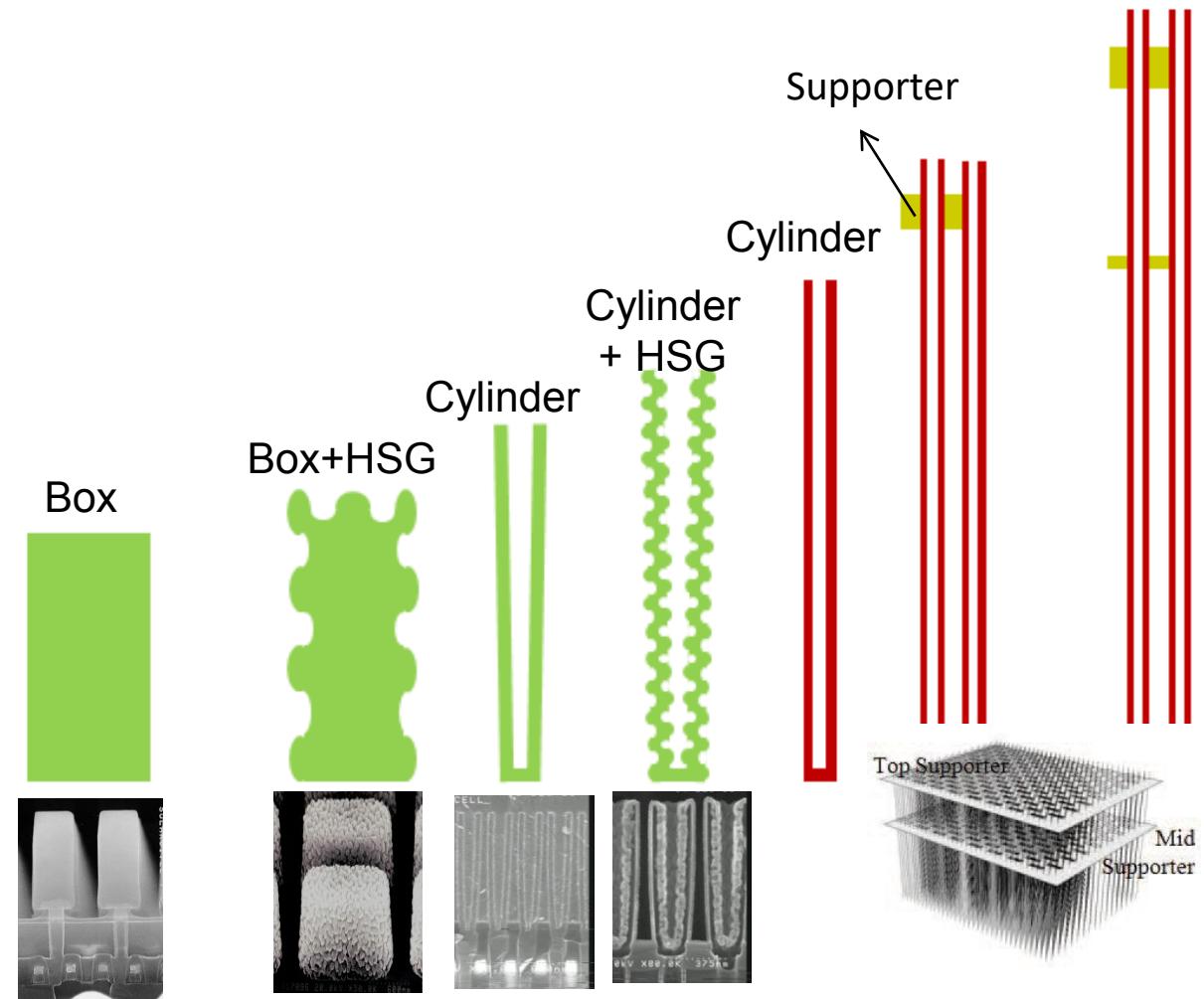
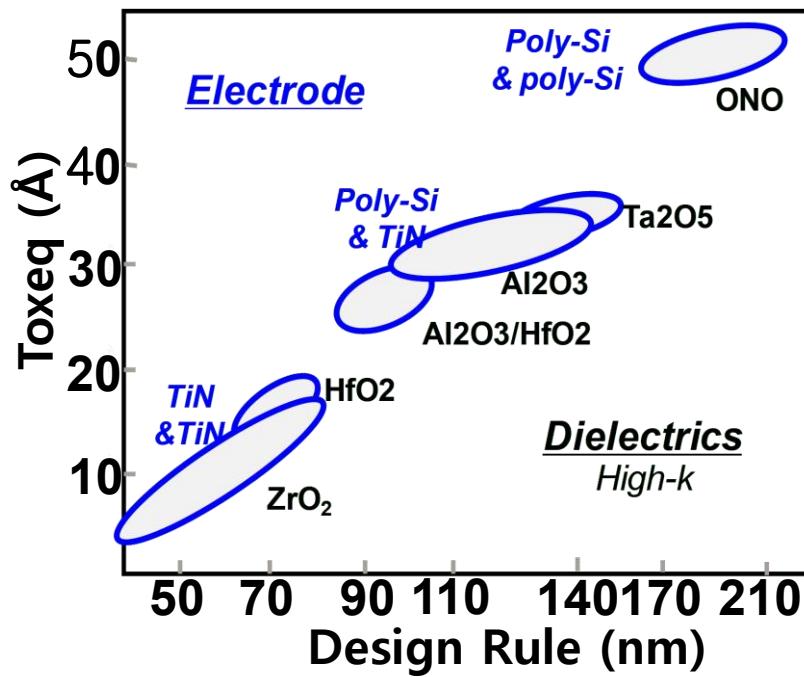


# Cell Transistor Trend



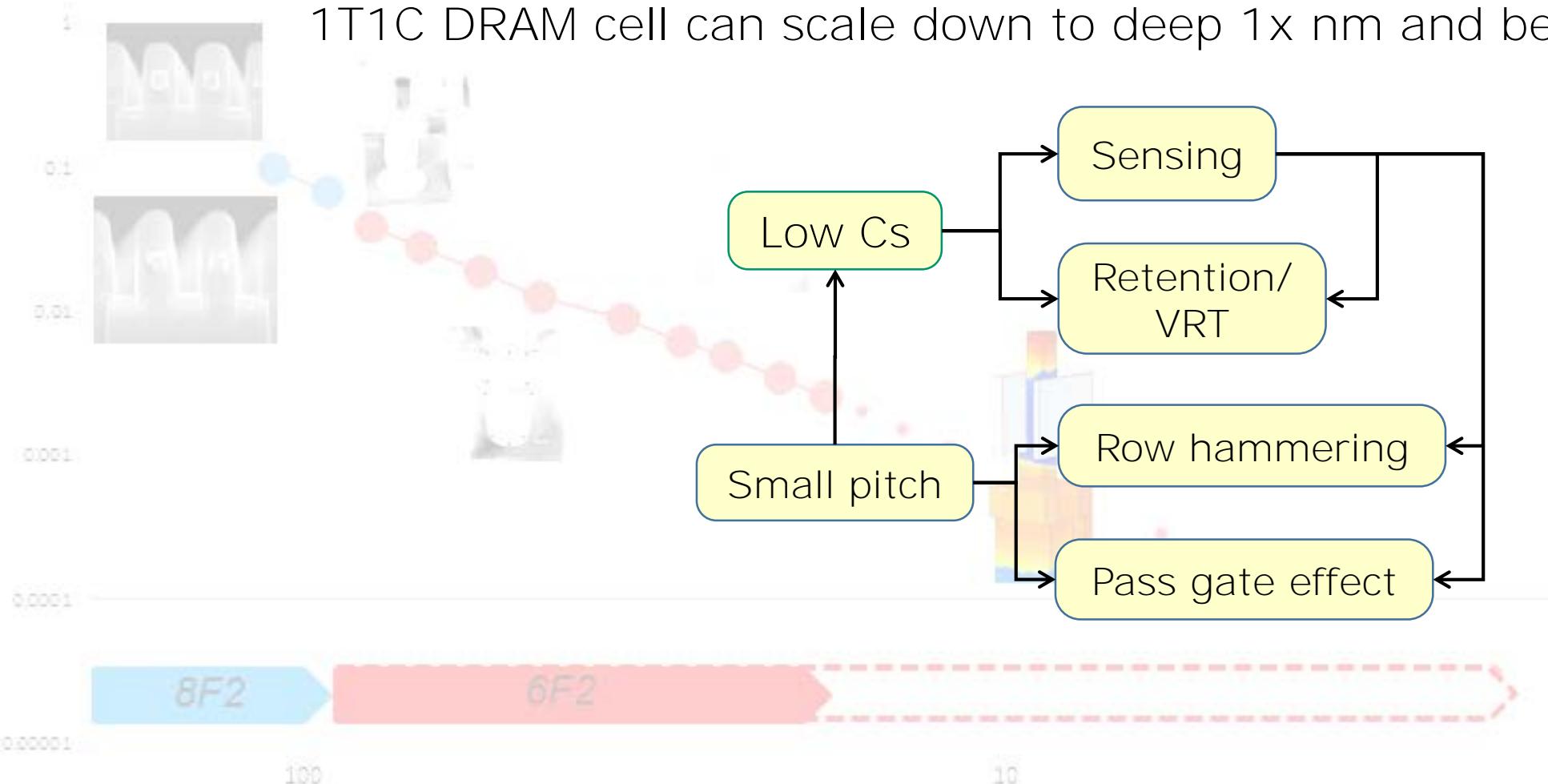
# Cell Capacitor Technology Trend

$$C_S = k \frac{A}{T_{ox.eq}}$$



# Scaling Challenges

1T1C DRAM cell can scale down to deep 1x nm and beyond 10nm ?



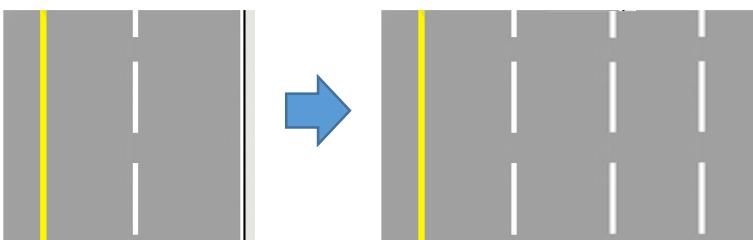
# Outline

- DRAM Technology and Scaling
- 3D-Stacked DRAM Technology
  - Introduction of TSV and 2.5D process
  - Thermal and SI/PI Challenge
- HBM (high bandwidth memory) DRAM
  - Introduction & Architecture
  - Difficulties and Solutions
- Function-in-Memory Solution for AI Application
  - Traditional Memory Solutions
  - FIM (Function-in-Memory) using HBM
- Summary

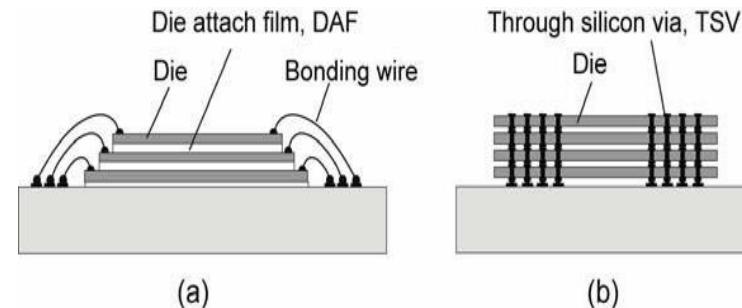
# TSV?

- Interconnection Solution for Wide I/O

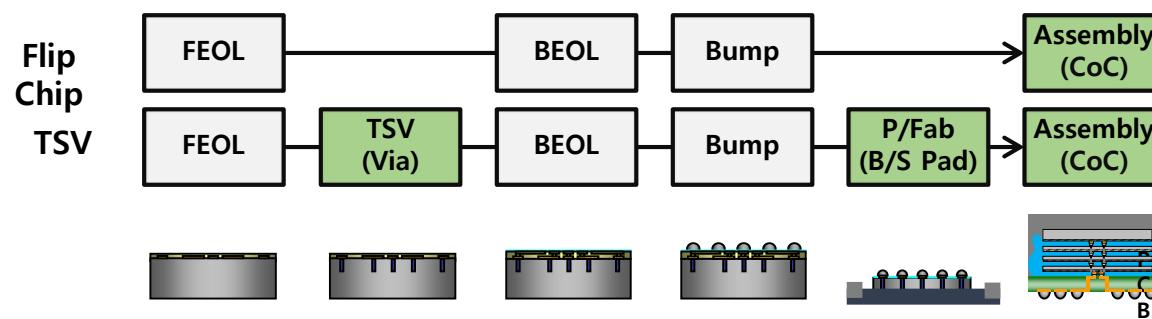
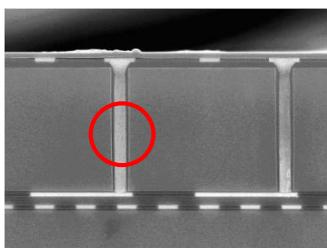
- Requirements for High Bandwidth



- Wide I/O & High Speed using TSV

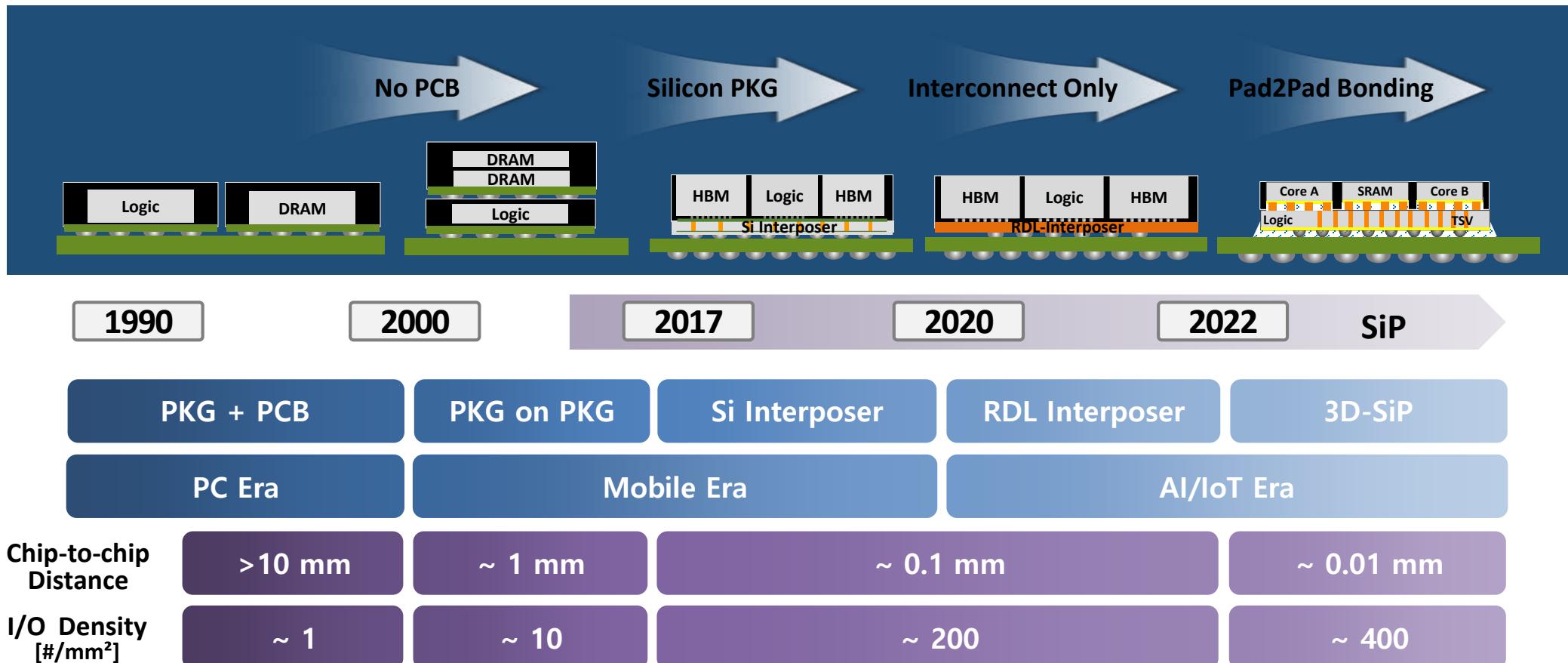


- TSV (Through Silicon Via) Process



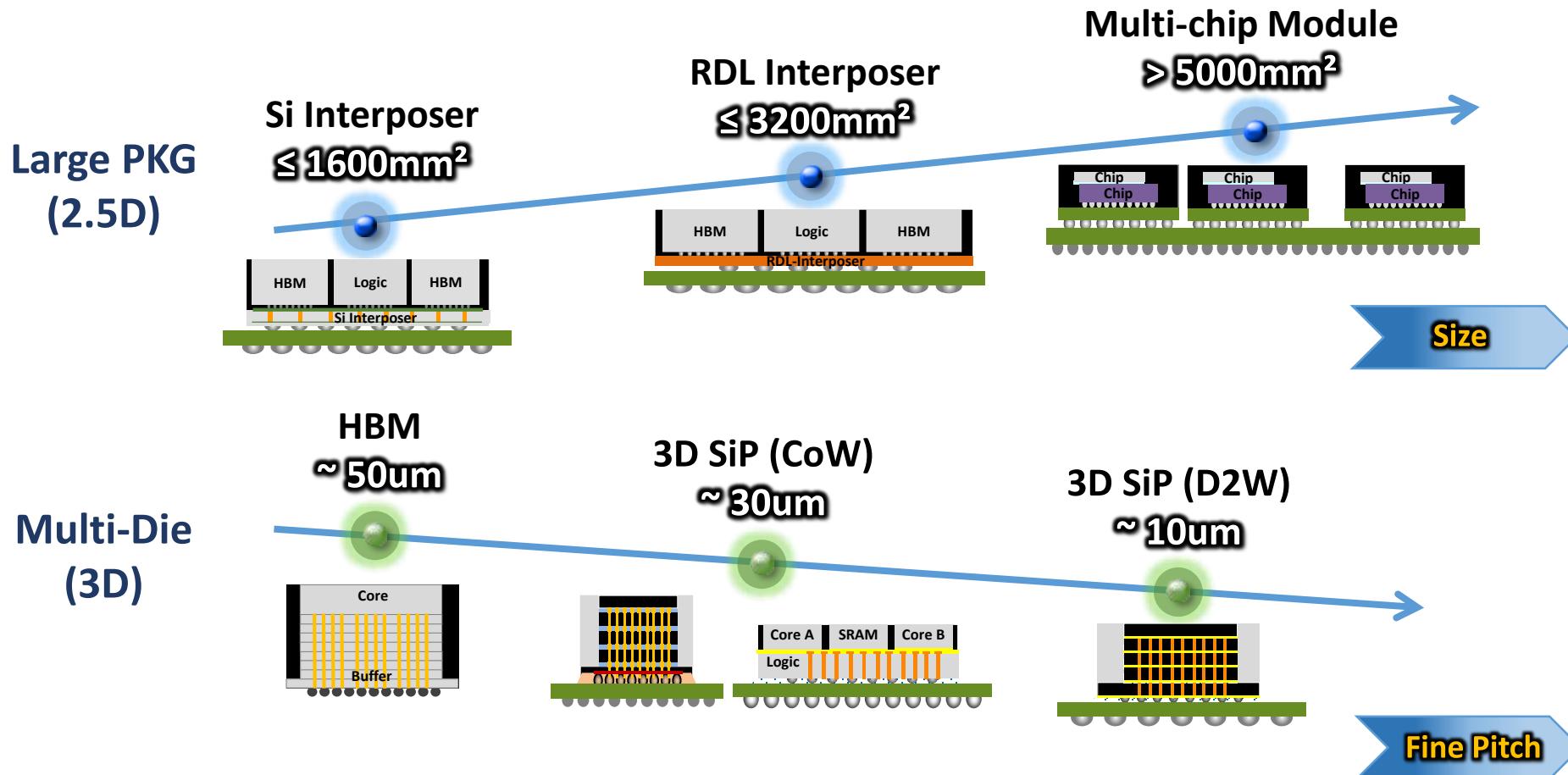
# Package Paradigm Shift

- System integration solution leads the packaging technology



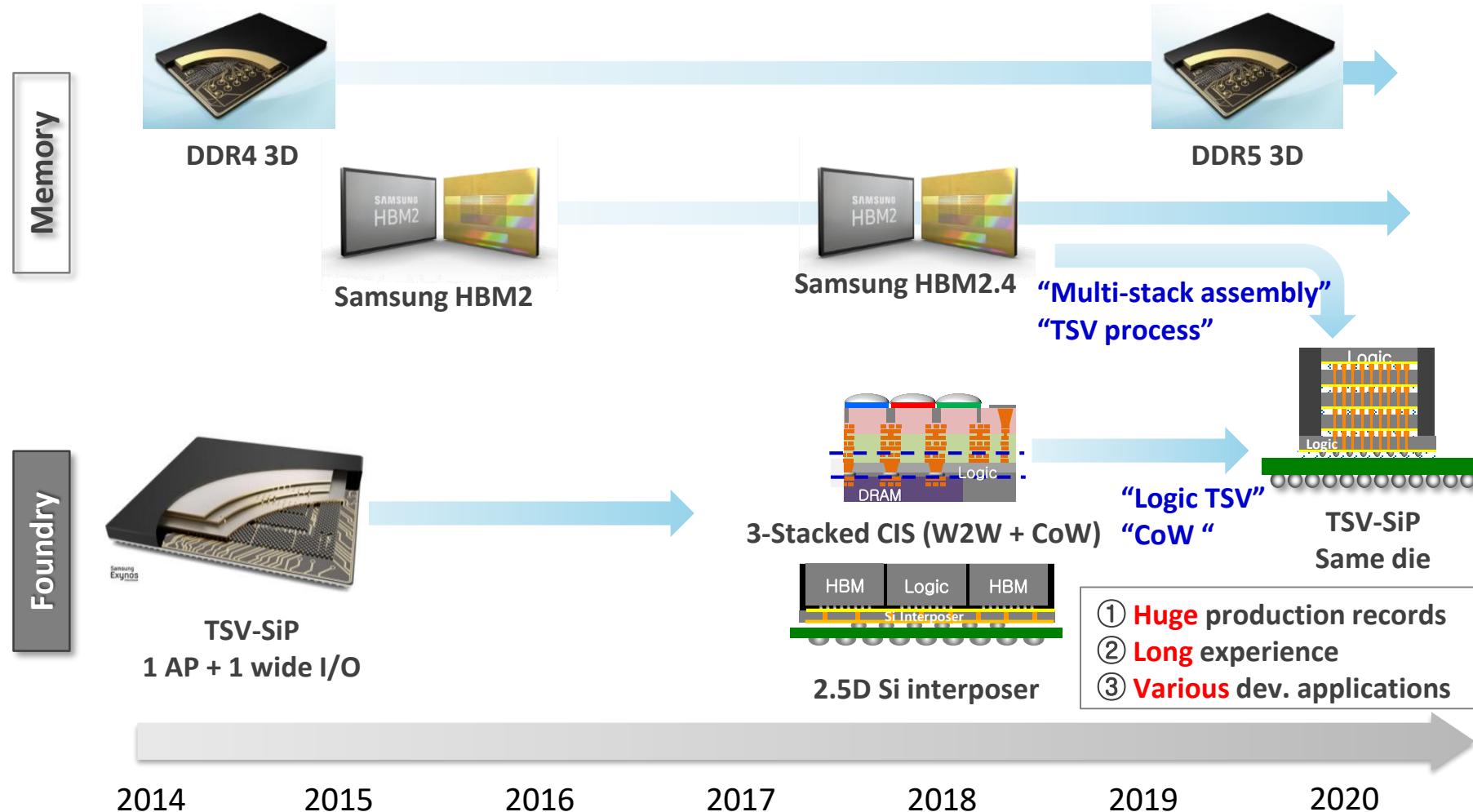
# AI/Server/HPC Platform

- Large package and more integration



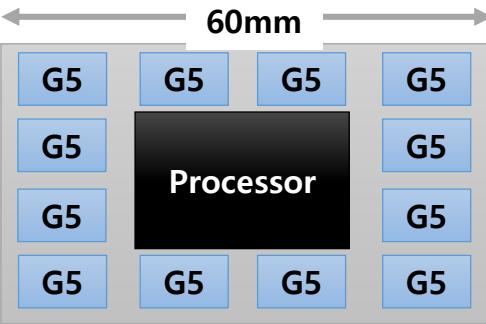
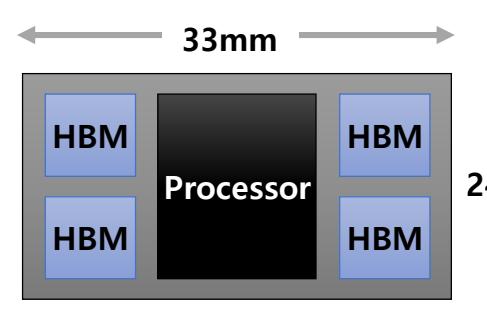
# 3D TSV Productions in Samsung

- DDR4 / DDR5 / HBM / TSV-SiP AP / 3-stack CIS



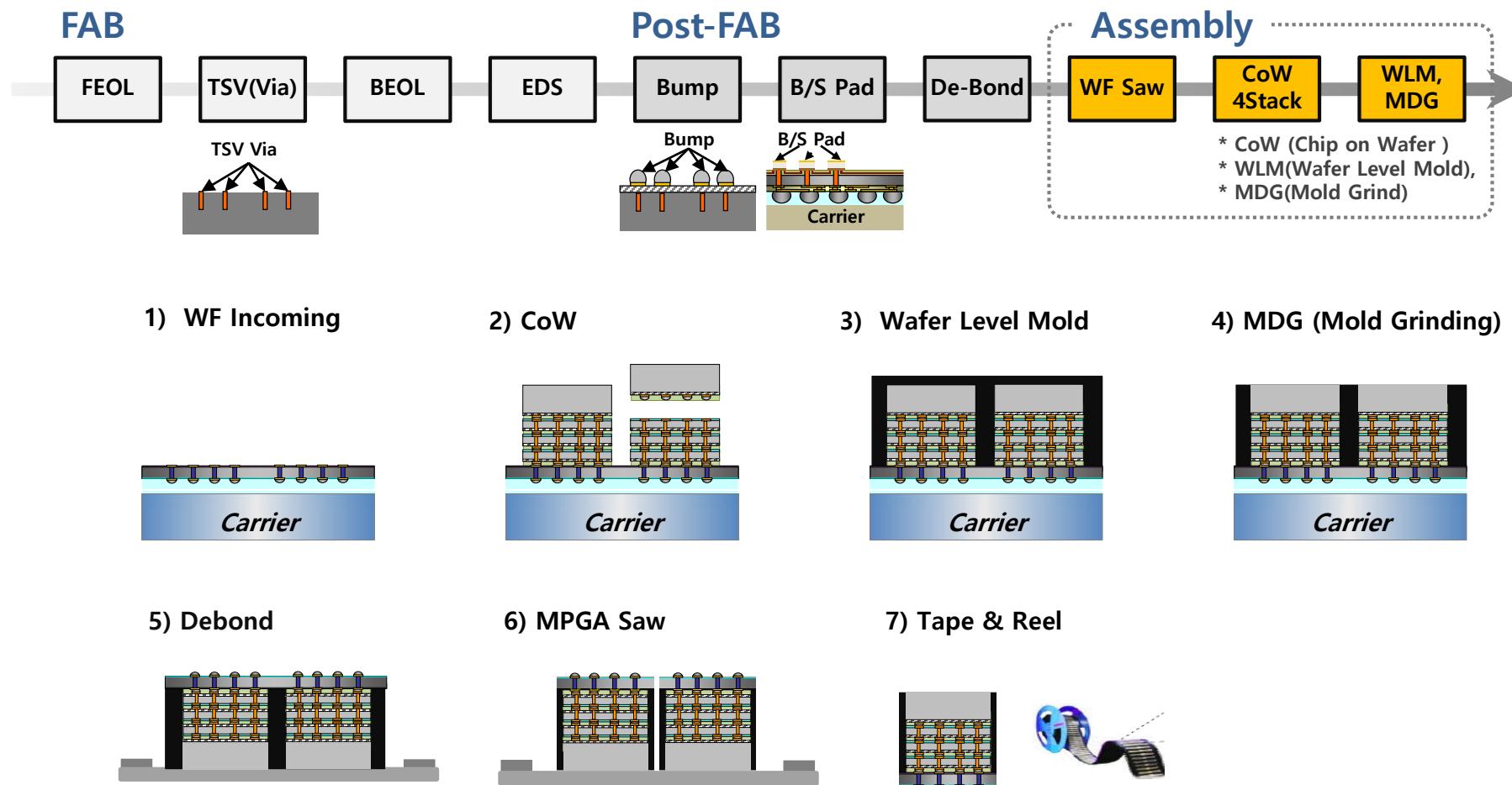
# Advantage of HBM Product

- GDDR5 vs HBM for High Bandwidth Application

ITEM	GDDR5	HBM (High B/W Memory)
System 구조		
DRAM	8Gb GDDR5 12ea	4GB HBM 4ea
Size	3120 mm <sup>2</sup>	-75% 792 mm <sup>2</sup>
Density	12GB	1.3x 16GB
Bandwidth	384GB/s	3.6x 1024GB/s
Power	18.3W (1.5W X GDDR5 12ea)	+18% 9.1W (2.3W X HBM 4ea)
Pin (Ball)	Speed	8 Gbps
	# I/O	32 per chip (Total 384)
		1024 per cube (Total 4096)

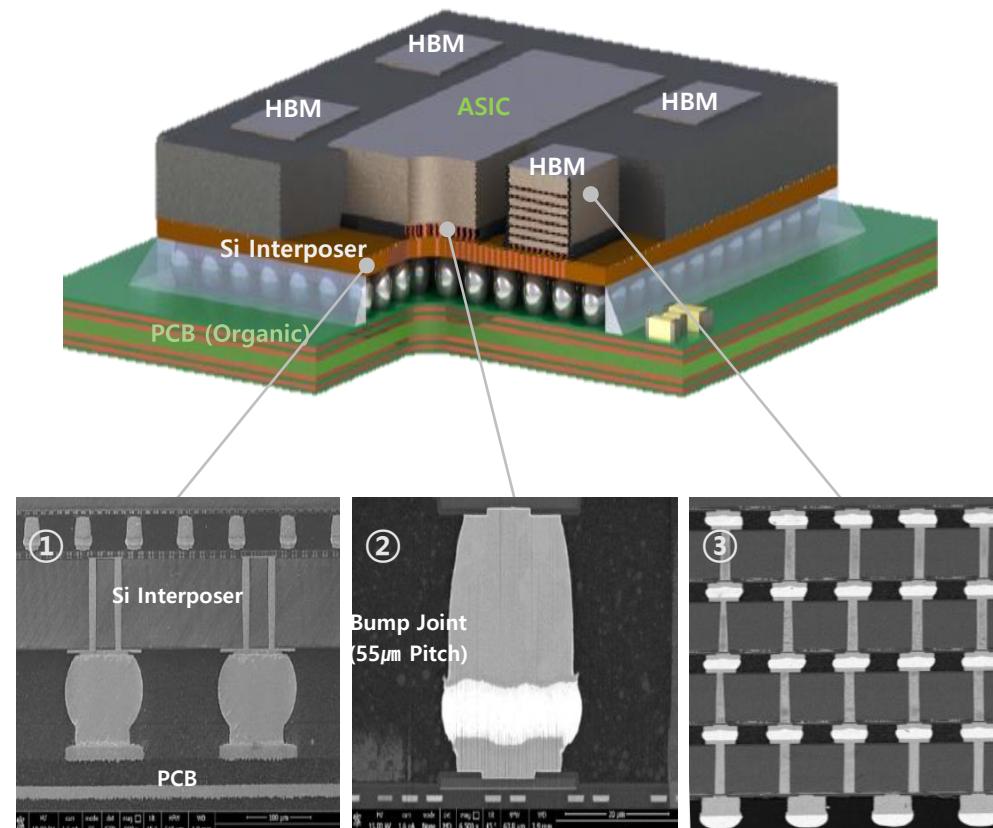
# HBM Process

- Process: FAB – Post-FAB - Assembly

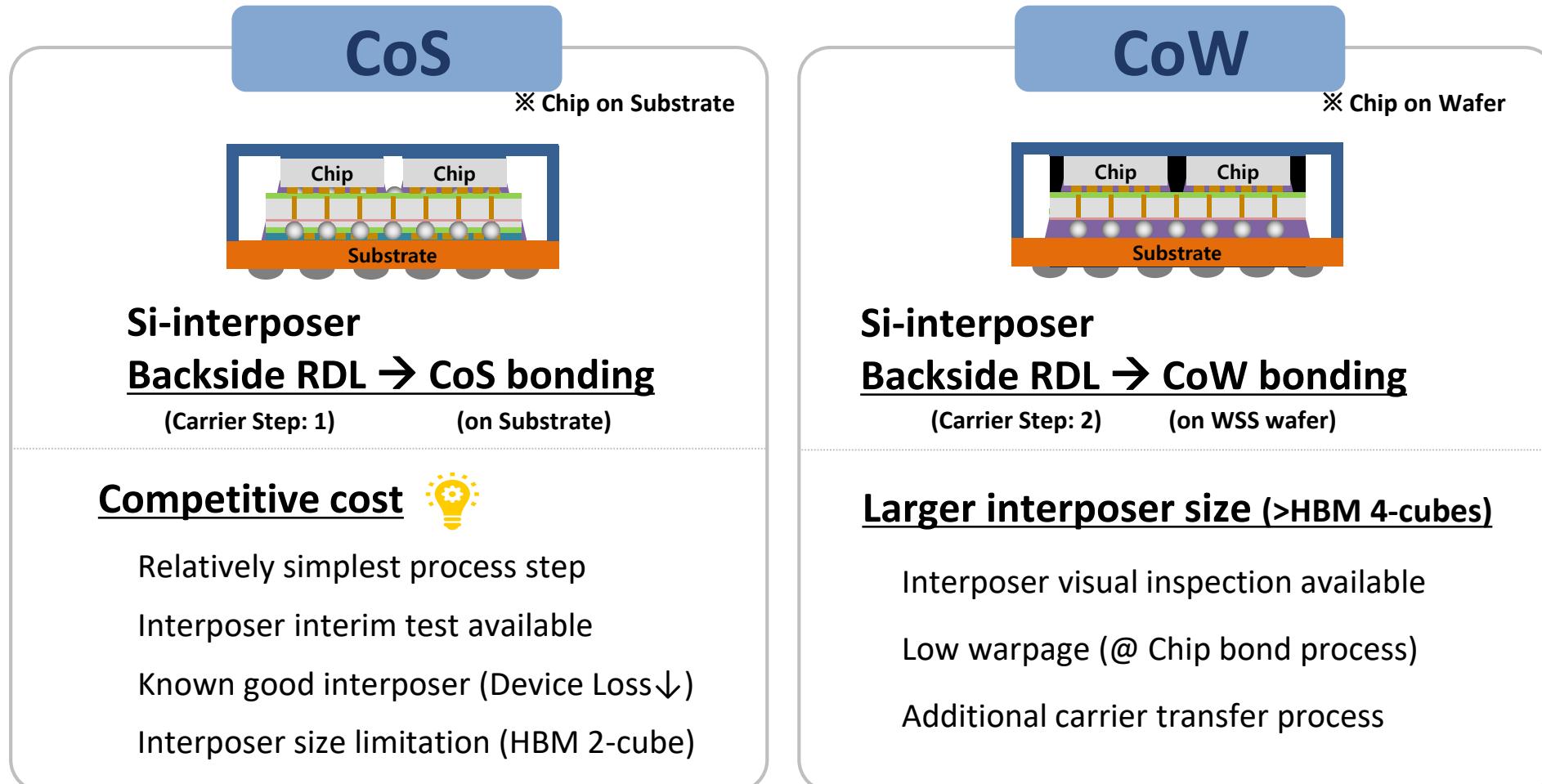


# 2.5D Product

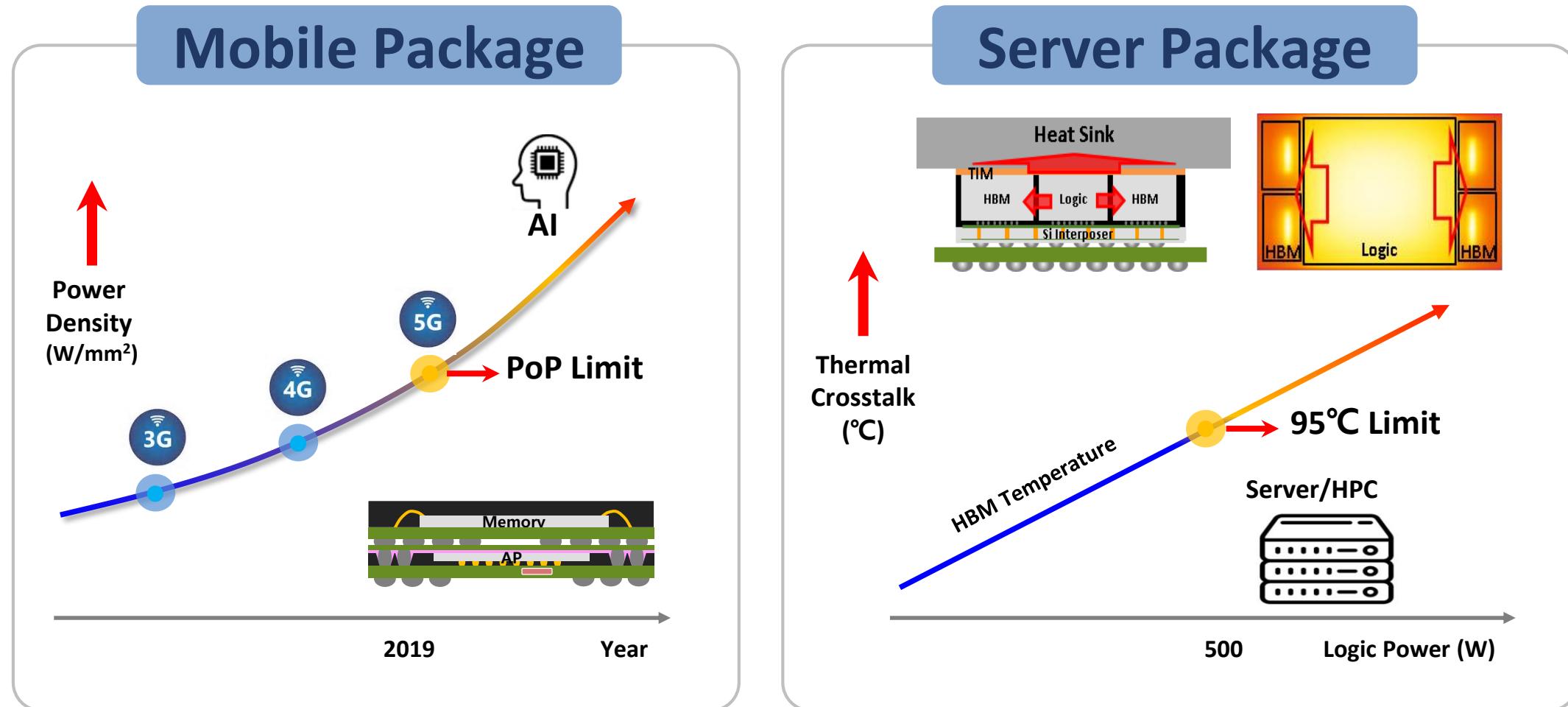
- HBM @ DRAM vendor → SiP(2.5D) @ OSAT → Customer



# 2.5D Assembly Process: CoS vs CoW

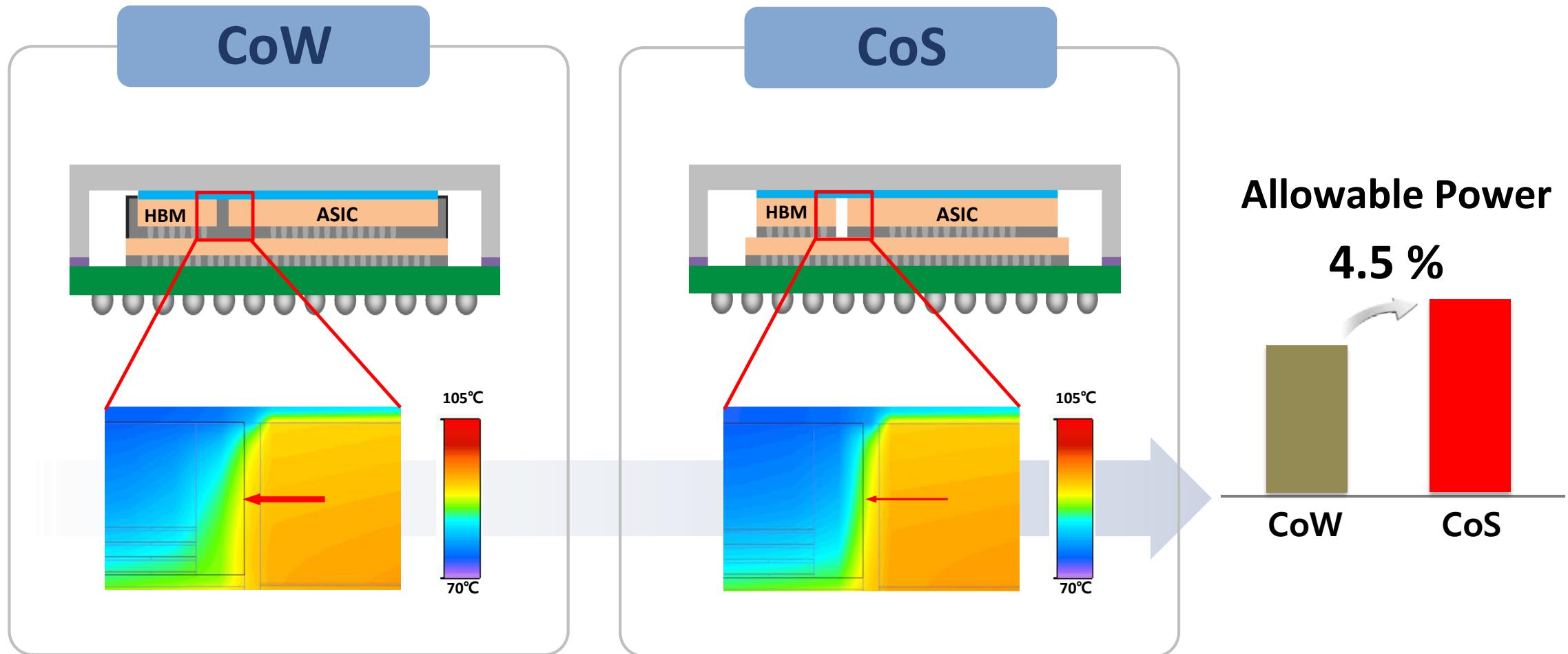


# Package Thermal Challenges



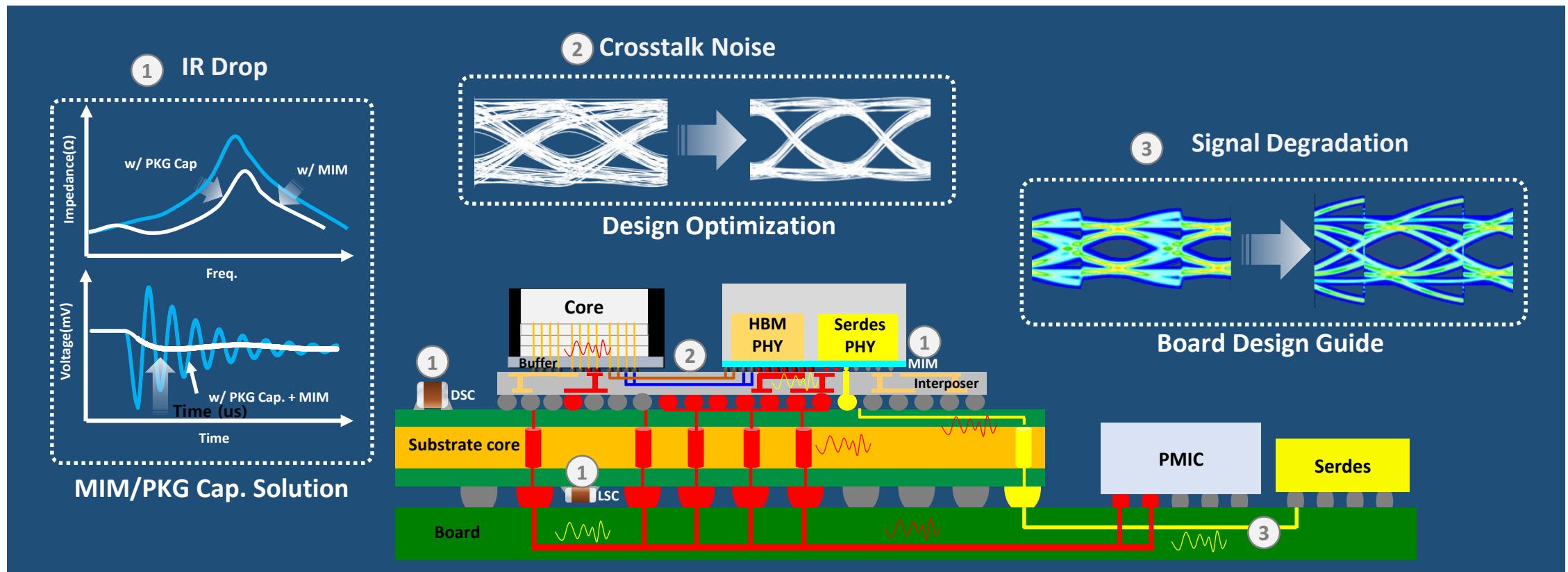
# 2.5D Package Thermal Crosstalk

- Air gap between ASIC and HBM of CoS decreases thermal crosstalk



# 2.5D Package SI/PI Solution

- Complex co-design is crucial

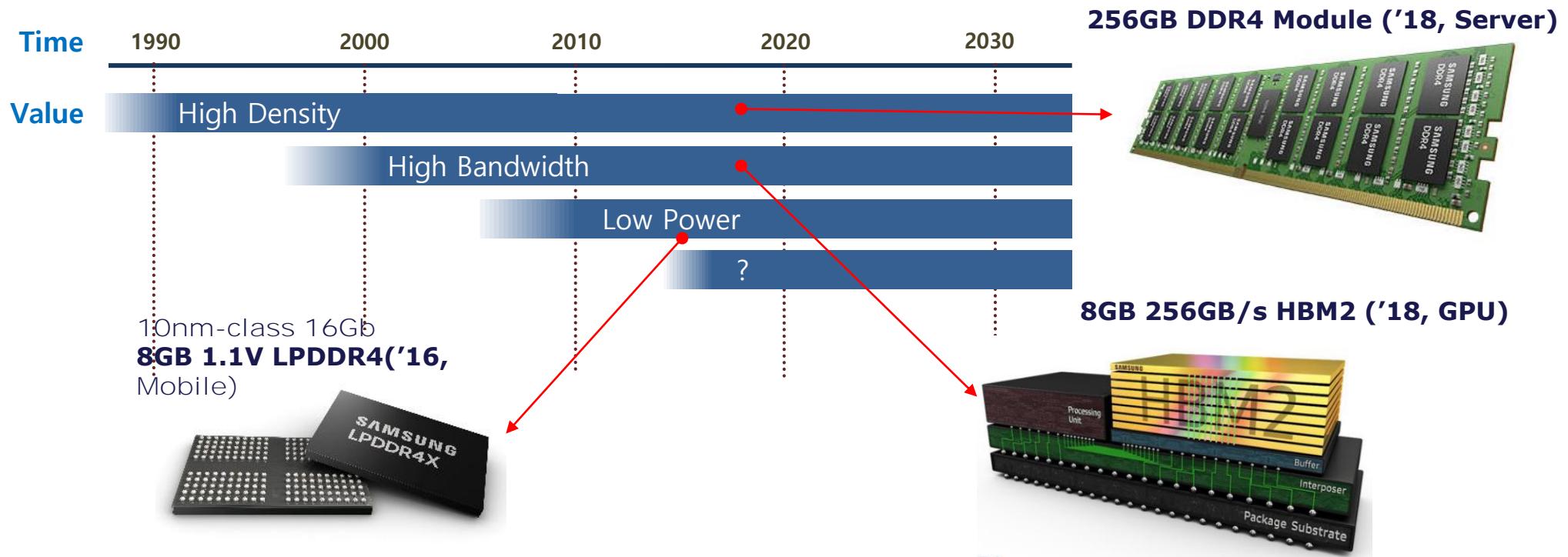


# Outline

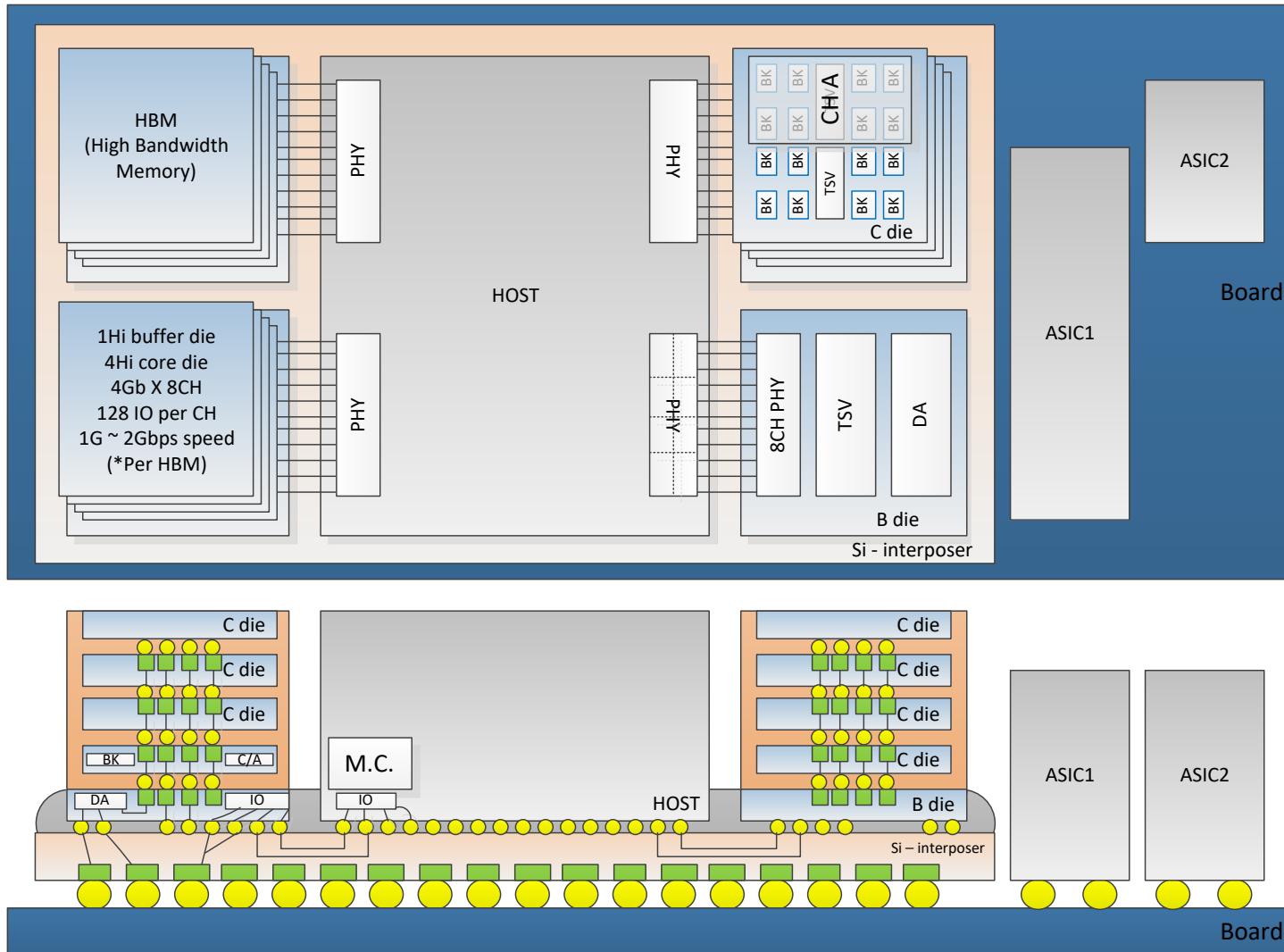
- DRAM Technology and Scaling
- 3D-Stacked DRAM Technology
  - Introduction of TSV and 2.5D process
  - Thermal and SI/PI Challenge
- HBM (high bandwidth memory) DRAM
  - Introduction & Architecture
  - Difficulties and Solutions
- Function-in-Memory Solution for AI Application
  - Traditional Memory Solutions
  - FIM (Function-in-Memory) using HBM
- Summary

# Memory Trend

- Driving force: high density, high bandwidth, and low power
- Now DRAM is preparing for the fourth wave of a new value

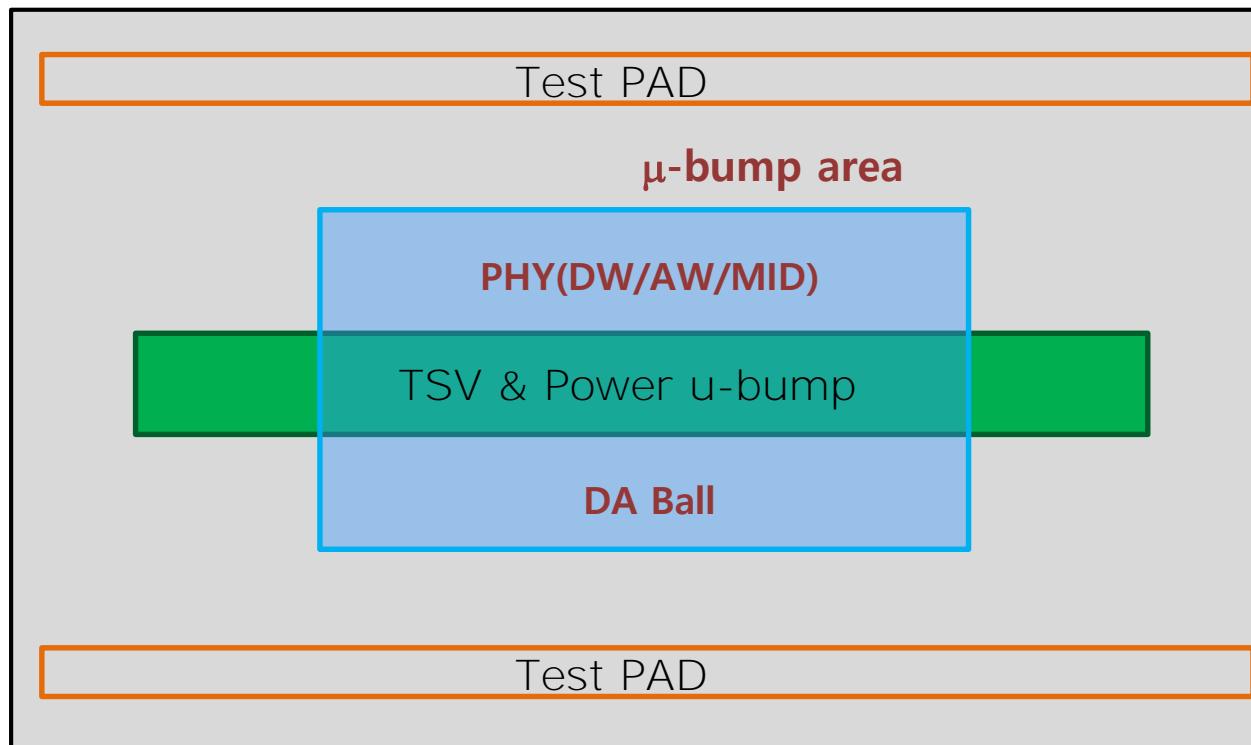


# SiP Structure using HBM



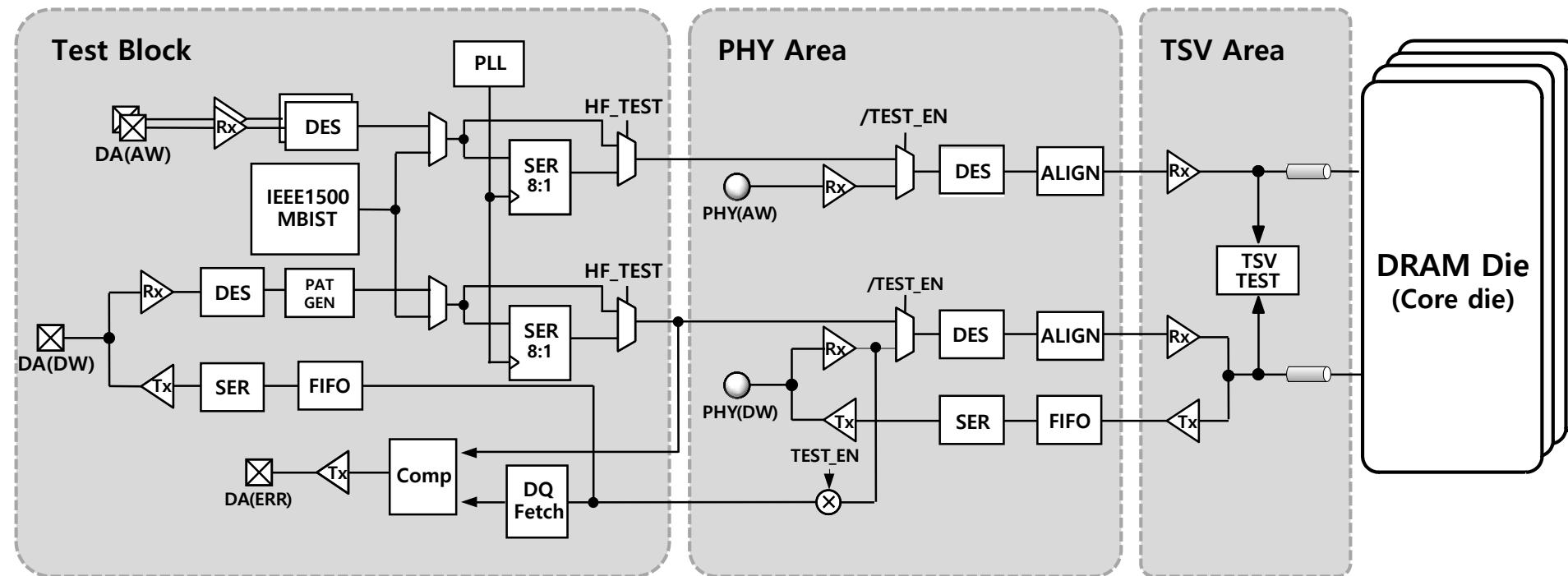
# HBM Architecture – Buffer die (1)

- Buffer die architecture
  - Routing from Core dies to external interface
  - Testability for mass production and failure analysis



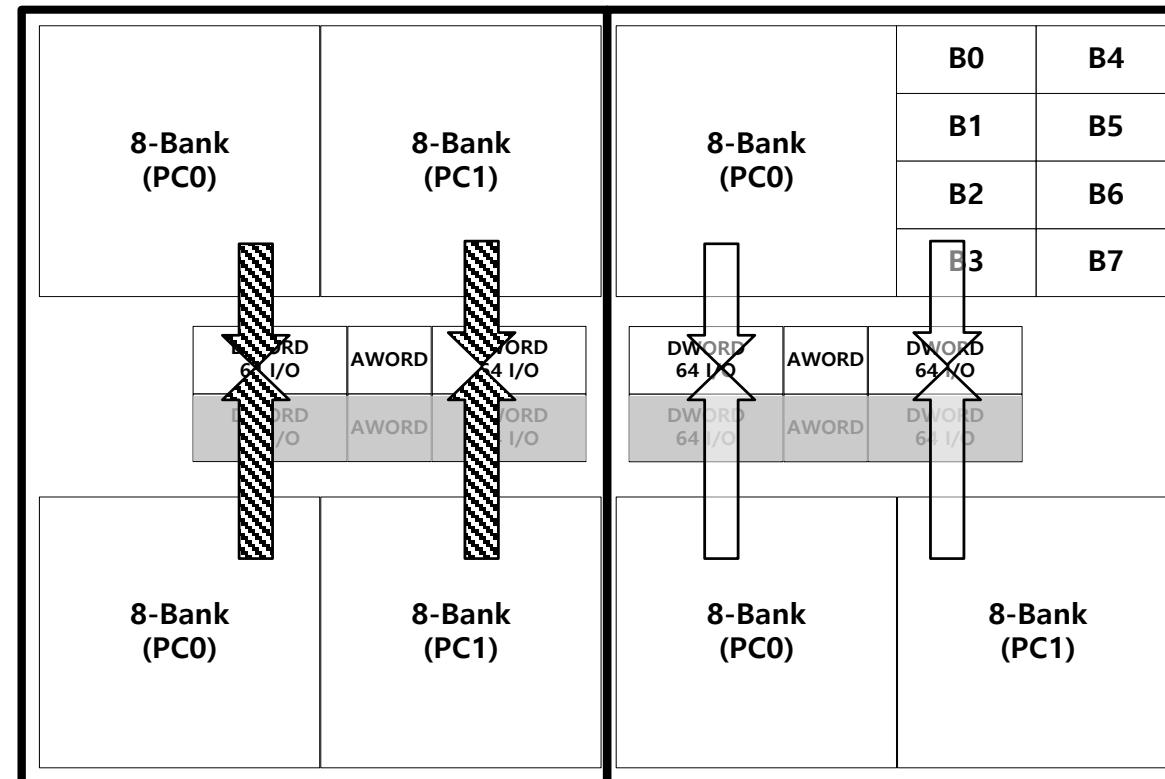
# HBM Architecture – Buffer die (2)

- Full test coverage of normal path (mission mode)
  - 8:1 SERDES using PLL is used for high frequency test
  - TSV test scheme is also implemented



# HBM Architecture – Core die

- Core die is a DRAM die for HBM
  - 2 channels per die, 2 pseudo channels per 1 channel
  - 16 banks: 4 bank groups, 4 banks per 1 bank group
  - Total 64 banks per each chip

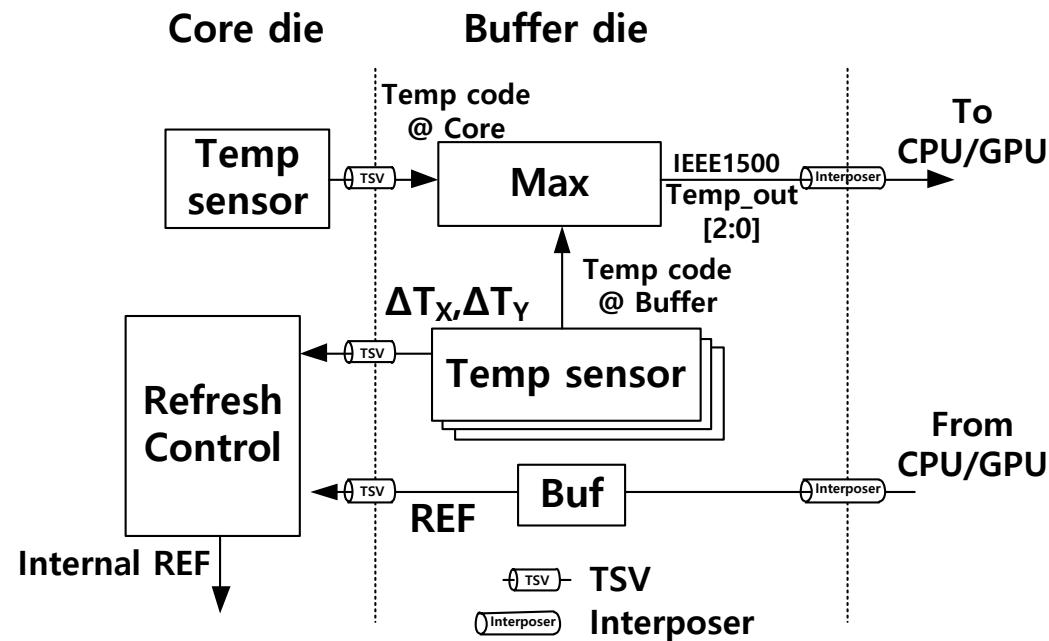
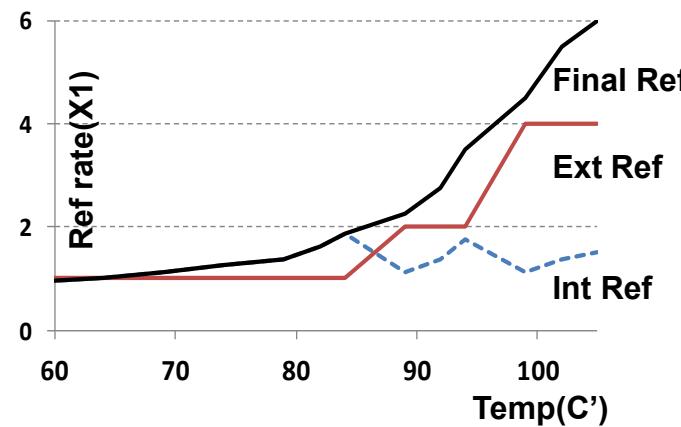


# Key Technologies

- High Power Density and Thermal Effect
  - ~10W power consumption in a small cube
  - Retention characteristics are sensitive to hot temperature
- Reliable TSV Connection
  - Thousands of TSVs between a buffer die and core dies
  - Test, repair and more
- Hard to probe m-bump itself during test stage
  - Test equipment to probe m-bump directly is not ready yet
  - How to guarantee qualified operations

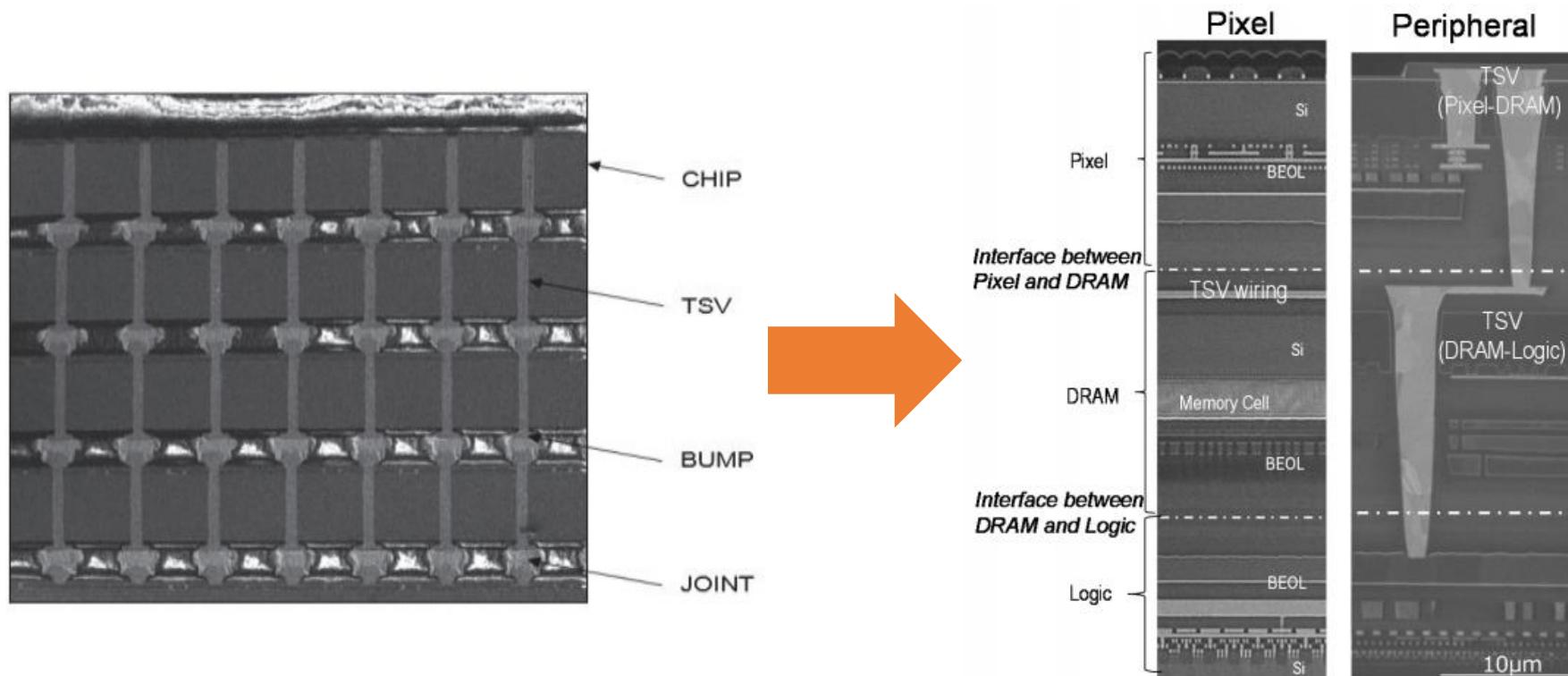
# Refresh Solution for Thermal Issue

- Adaptive Refresh considering Temperature distribution
  - Refresh rate calculation from multiple temperature sensors
  - In-die distribution comes from buffer die sensors
  - Refresh rate is determined by external and internal refresh



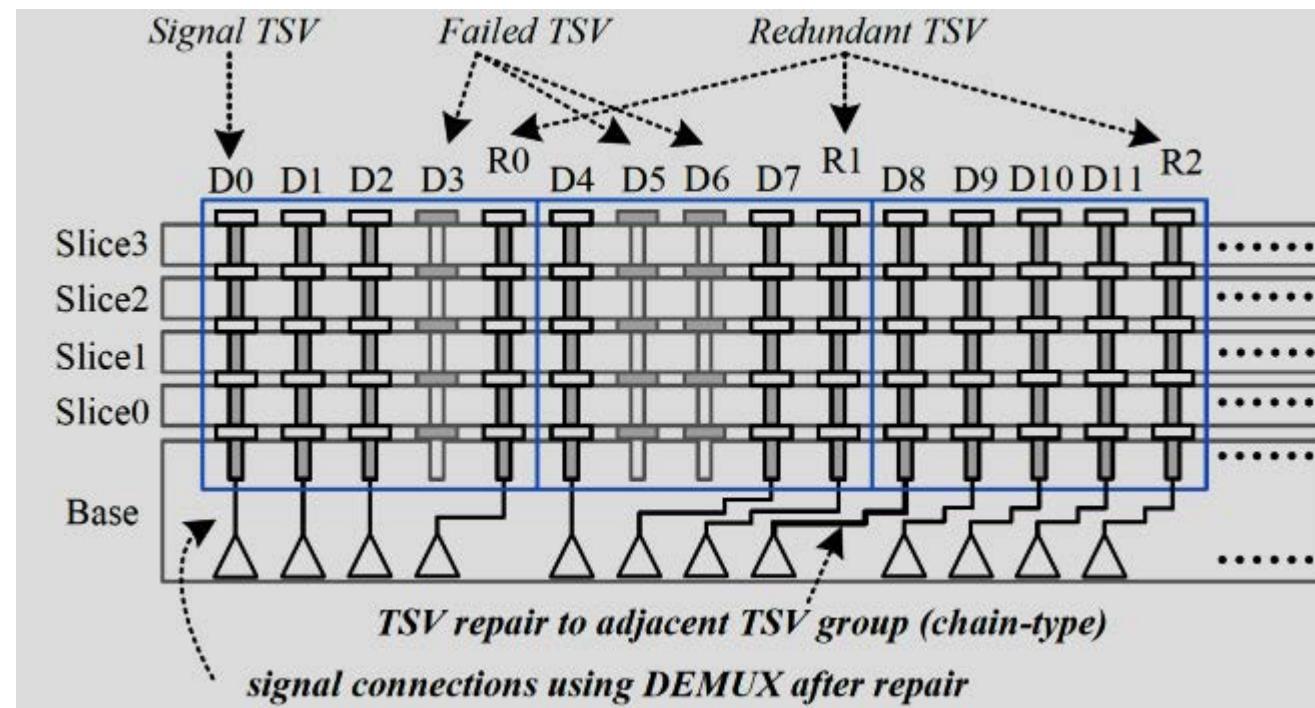
# Package Solution for Thermal Issue

- Bump-less stack solution can reduce thermal resistance hugely
- Need to develop packaging and design breakthrough



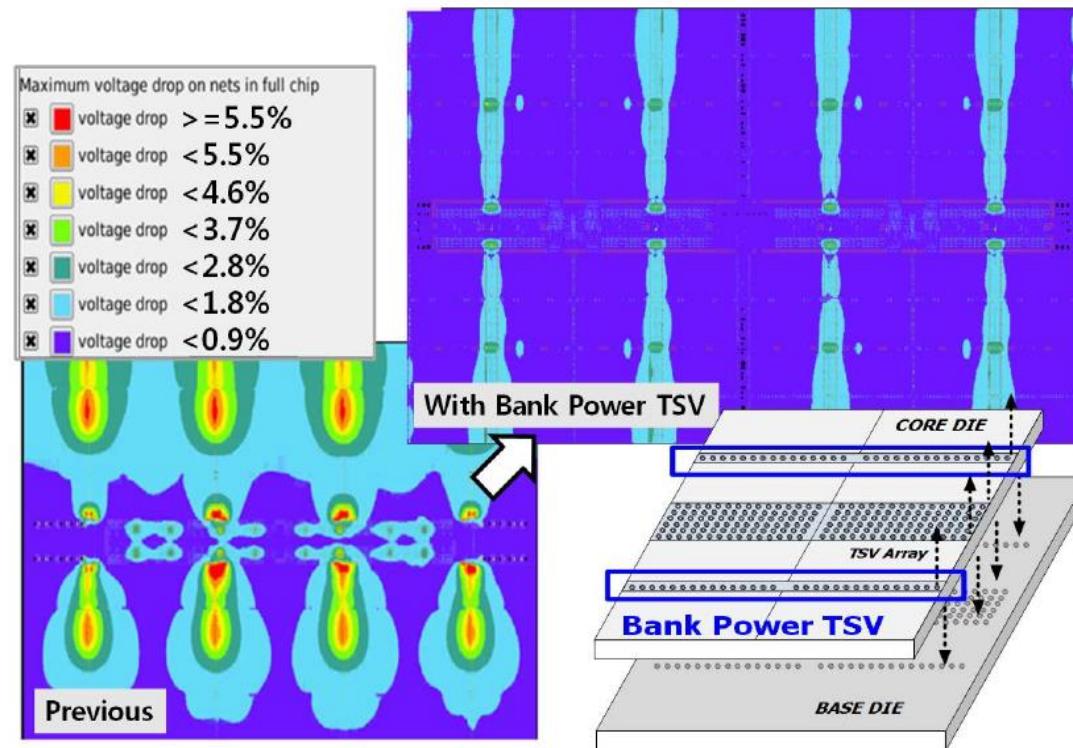
# Mass TSV needs Management

- Thousands of TSV per each core die are necessary
- TSV test and repair scheme are essential



# Power TSV to Improve PDN

- Additional power TSVs can reduce internal voltage drop effectively
- Trade-off between chip size/circuit and power-distribution network

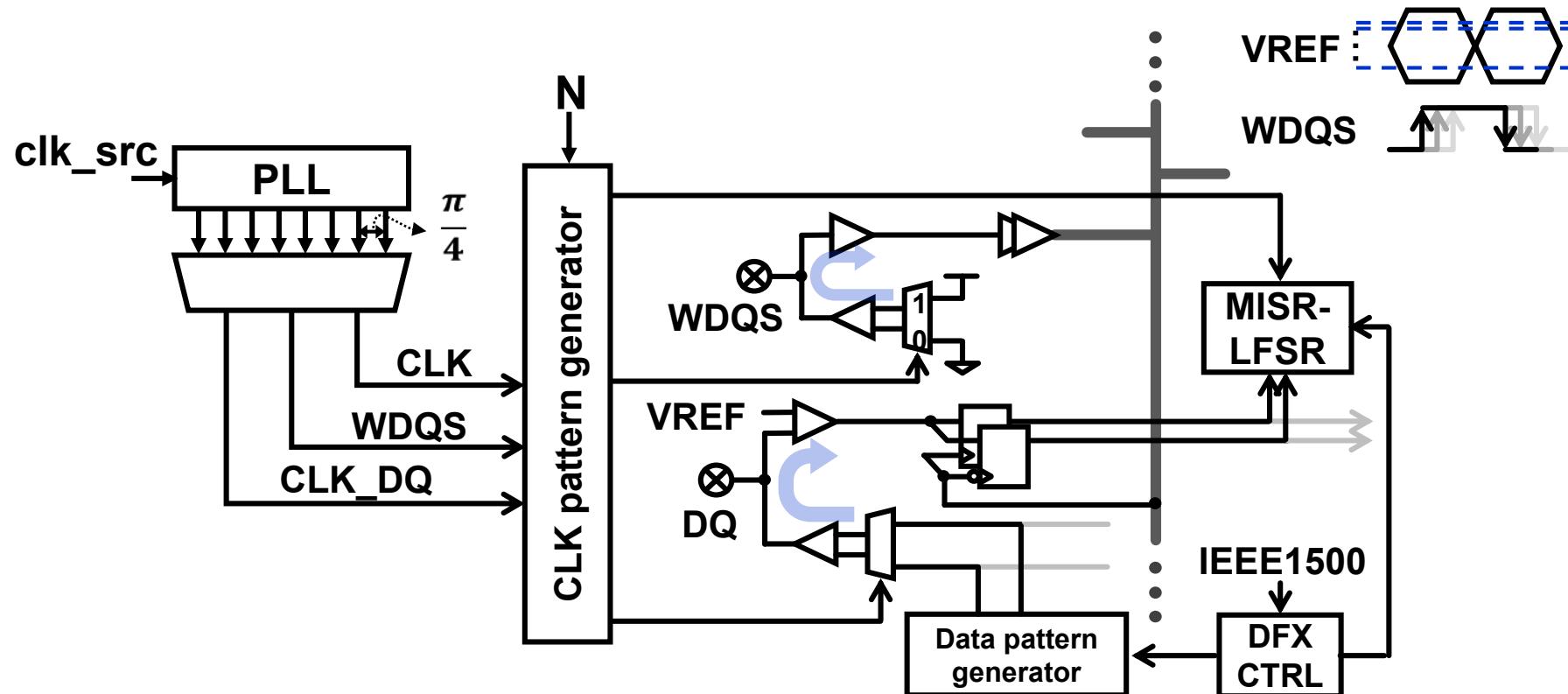


# IO DFx (Design for Excellence)

- Testability of m-bumps
  - Direct probing can hurt m-bump physically: Developing new tester
  - Too many m-bumps to be tested: Efficient test method
- IO parameters
  - Check tISH, tDSH, tDQSS inside DRAM
  - Tools: MISR write, LFSR compare, LFSR read
- IO DFx
  - Output (Serializer, Driver) by PLL clock
  - Input (Buffer, Deserializer, Sampler) by PLL clock
  - Pass/Fail check, Loopback function signature

# Example: MISR Write with Toggle Pattern

- tDSH, MISR write, toggle pattern read



# Next Generation of HBM?

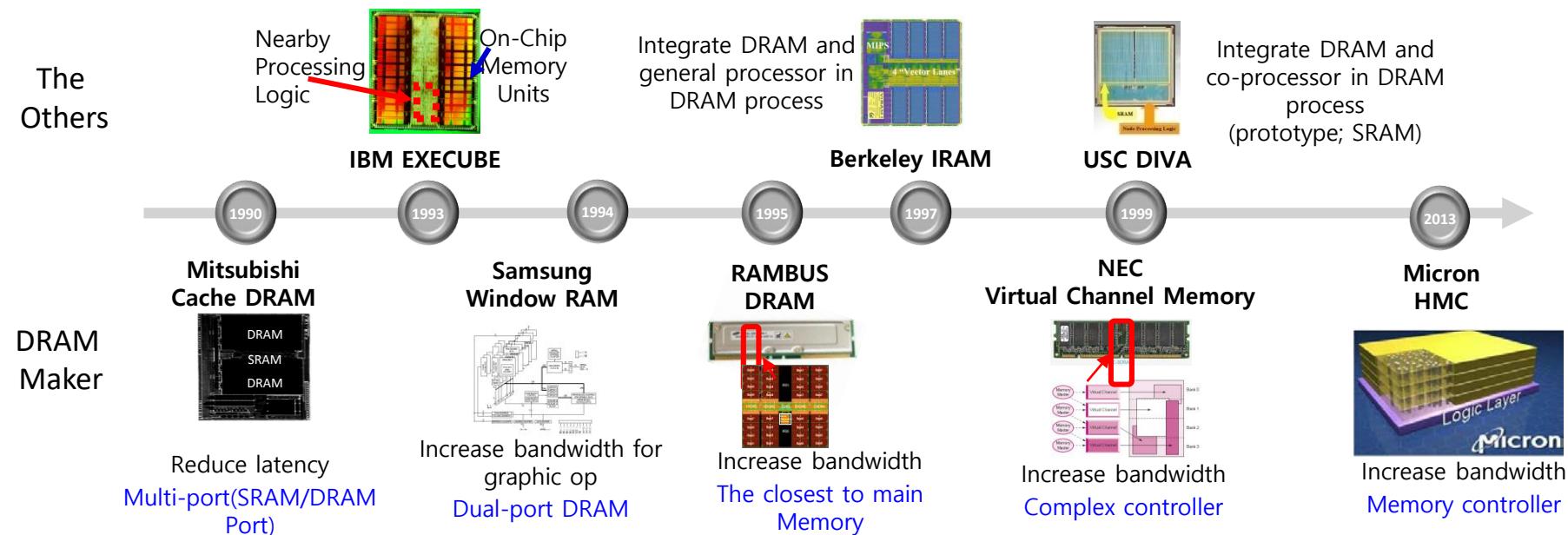
- Higher Bandwidth
  - Gen2: 256GB/s (2Gbps/pin) → 358GB/s (2.8Gbps/pin) → 460GB/s (3.6Gbps/pin)
  - Gen3: 512GB/s~ (4Gbps/pin)
- Higher Capacity
  - 4GB(4H) → 8GB (8H) → 16GB (8H) → 24GB (8H or 12H) → 32GB ~
- Power & Thermal
  - Higher BW with same power: improved power efficiency
  - Low thermal resistance: New package solution
- Enforced DFT & RAS feature
  - Better testability and reliability, DFT solutions for failure analysis
- and Additional Functions for new applications (AI, Automotive, ...)

# Outline

- DRAM Technology and Scaling
- 3D-Stacked DRAM Technology
  - Introduction of TSV and 2.5D process
  - Thermal and SI/PI Challenge
- HBM (high bandwidth memory) DRAM
  - Introduction & Architecture
  - Difficulties and Solutions
- Function-in-Memory Solution for AI Application
  - Traditional Memory Solutions
  - FIM (Function-in-Memory) using HBM
- Summary

# History of Revolutionary DRAMs

- DRAM Makers
  - Focused on increasing bandwidth and reduce latency  
→ not easy to scale in terms of density
- The others
  - Focused on PIM and/or general processing



# Why PIM not successful in early attempts?

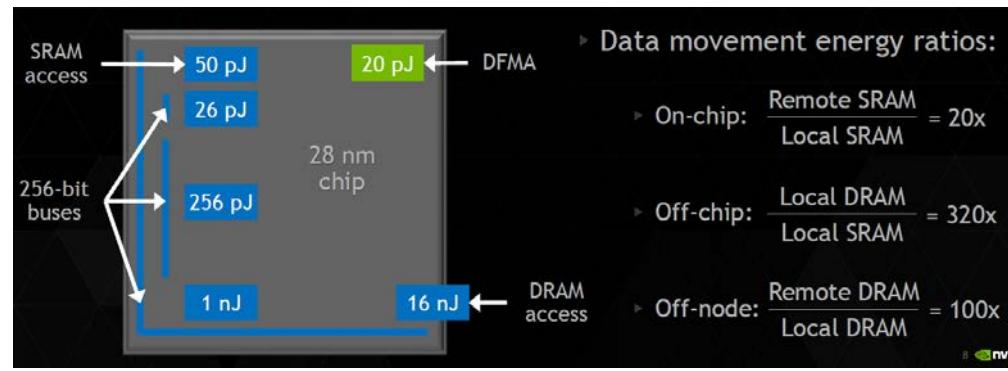
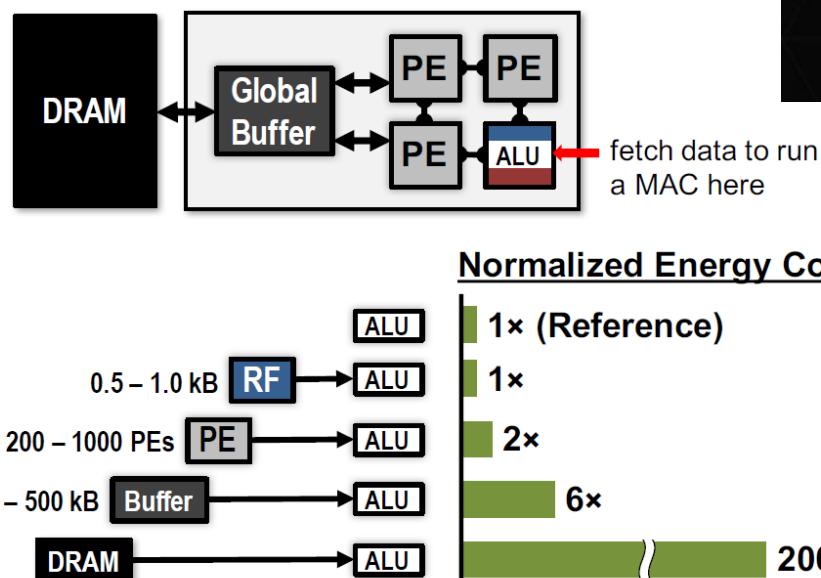
- High cost in comparison with a commodity product in terms of area, process technology of merging DRAM with logic process
- The burden of software stack change for PIM
- Easily replaceable solution by an enhanced existing solutions in terms of bandwidth, latency, and power

# Why PIM is Rising Again?

- Trend: CPU-centric → Memory- & Data-centric
  - Past: CPU did everything
  - Now: Heterogeneous computing, D/C boom
- Application: General Purpose → Specific Application
  - Deep Learning, Neural Network: Data intensive, Pre-defined, Repetitive
- DRAM Change: Convention DRAM(DDR3/4) → HBM
  - Separated logic die: Provides Space for implementation
  - Higher performance peripheral circuits to make a function (ex. MAC)

# Renewed Motivation for PIM

- High energy cost of data movement
  - DRAM access costs orders magnitude of more energy than data processing



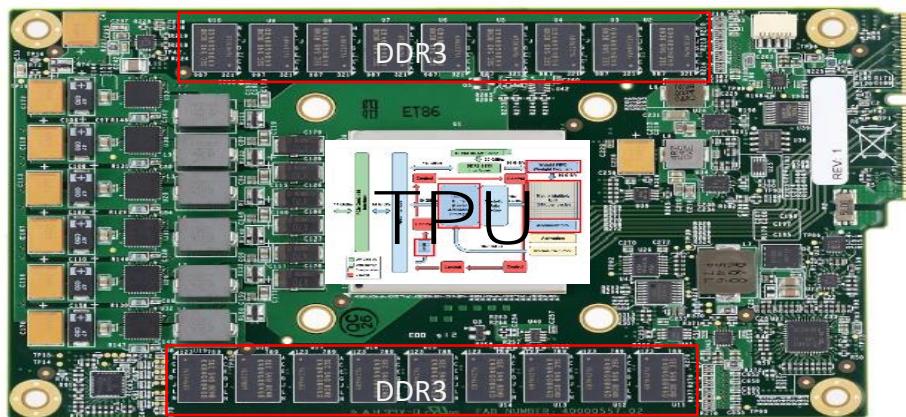
*How can reduce remote and/or off-chip DRAM accesses?*

source: \* Y.-H. Eyeriss, 2016 ISCA

\*\* NVIDIA's Vision for Exascale, NVIDIA Forum

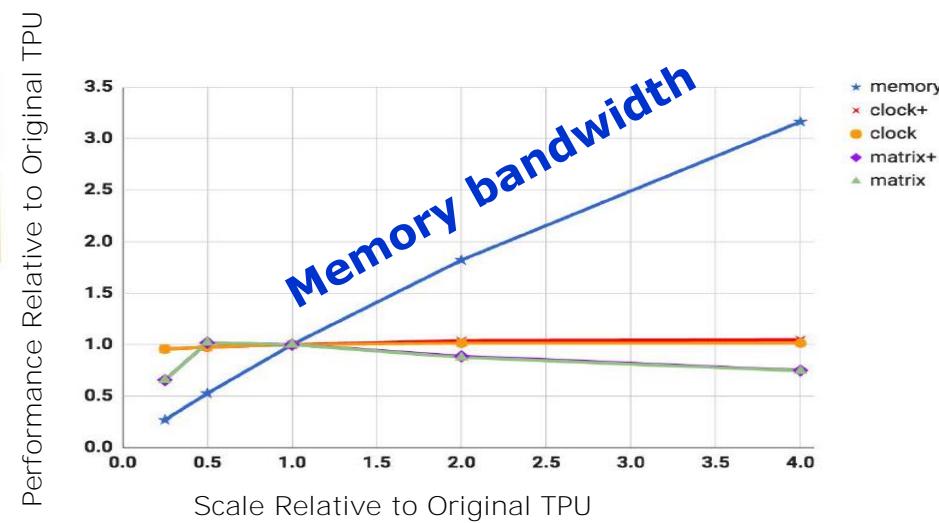
# Renewed Motivation for PIM - AI

- Custom Neural Processor(CNP) & Memory Bandwidth
  - CNPs are memory-bandwidth-starved processors



Google TPU Board

Memory Bandwidth: 30GiB  
4Gb/DDR3/x8/2133Mbps = 18 ea

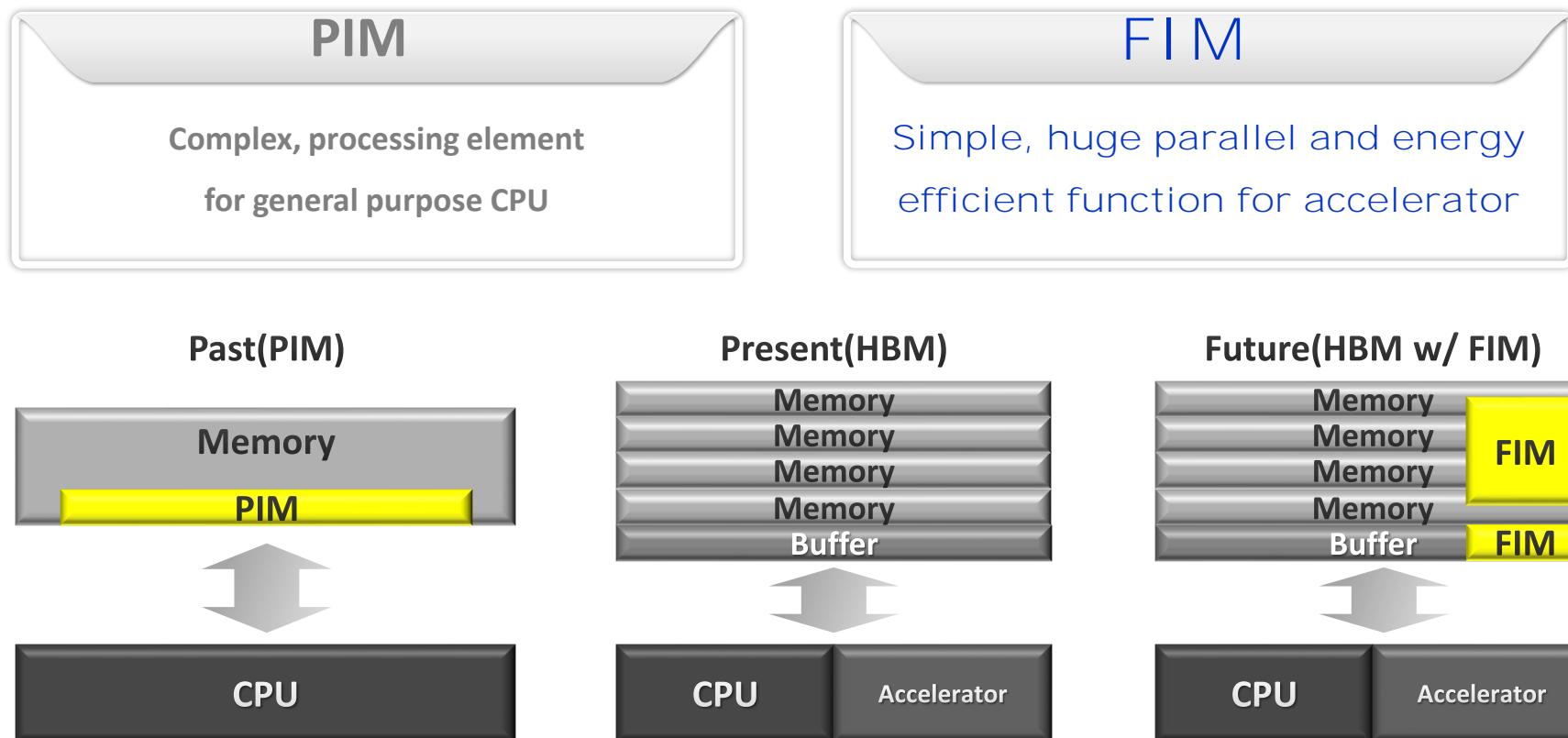


The performance of TPU system shows a high dependency on memory bandwidth

→ Need a PIM-like solution !!

# HBM solution for next generation AI

- Extremely parallel processing using internal bandwidth
- Minimize data transaction for energy efficiency



# HBM-based FIM Enabling Challenges

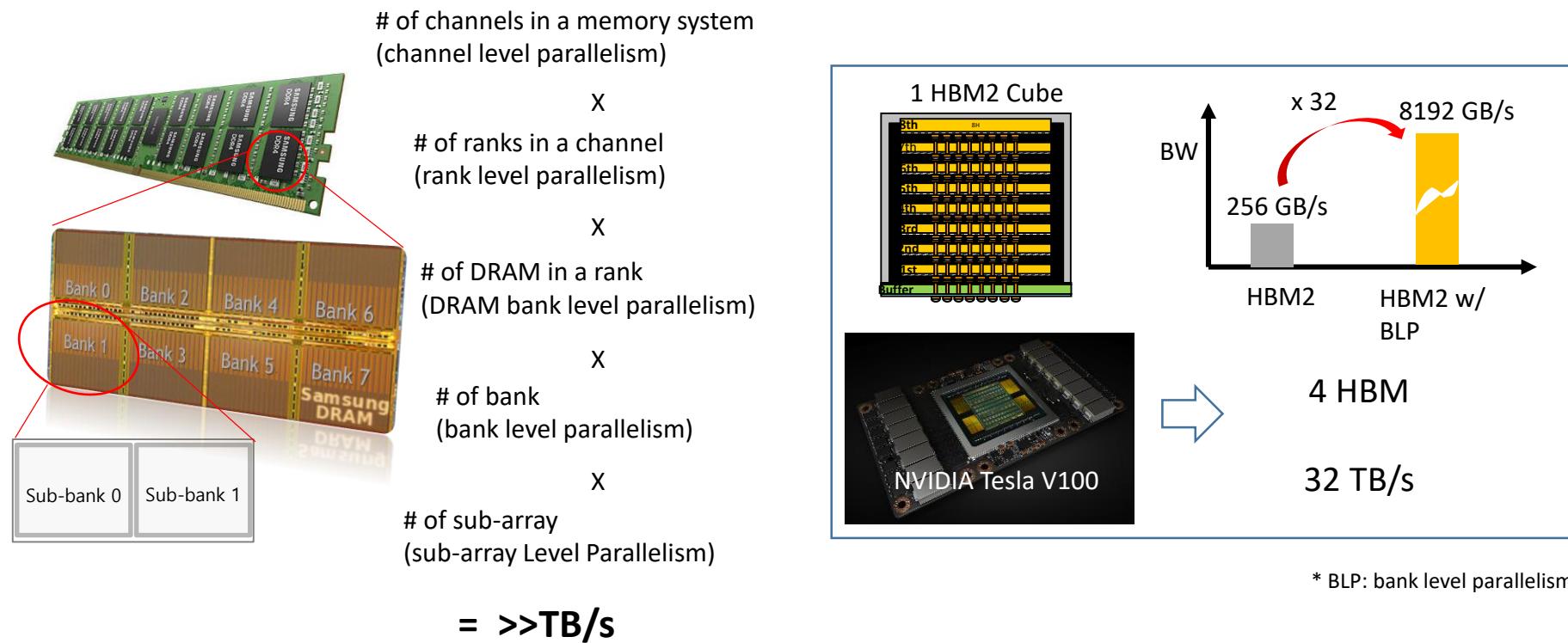
- Exposing **higher bandwidth** for FIM in current HBM
- **Compatibility** with existing application and programming model
  - enable FIM without change of existing application and programming model, or create a programming model for more performance gain
- **Cache coherence and memory consistency**
  - modifications to a particular location (coherence) and different location (consistency) seen in order
- **Virtual memory**
  - virtual(seen by software stack) to physical address translation
- **Power & Thermal overhead**
  - Additional overhead from functions (ALU, data handling, ...) to HBM

# Some Ideas about FIM Challenges

- Internal bandwidth limitation for performance gain
  - Bank/array parallelism of DRAM VS manufacturing cost from area overhead
- Compatibility with existing application and programming model
  - Just add one more library (ex. GPU or NPU) in software stack
- Cache coherence, memory consistency and virtual memory
  - Fully controlled by memory controller of host side
- Power & Thermal constraint
  - internal power from additional function VS interface and data movement power

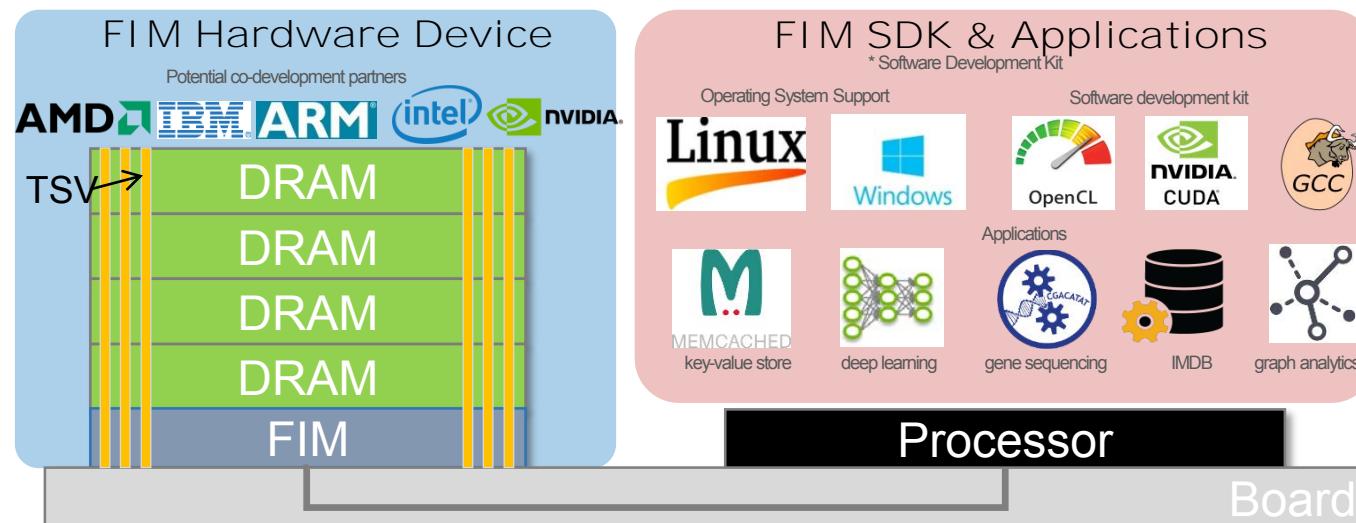
# FIM Exploiting Parallelism in Memory

- Potential parallelism resources
  - sub-array/bank/rank/channel–level-parallel processing
  - A new dimension of parallelism in processing



# Open Collaboration for FIM

- Open collaboration to elevate AI system performance
  - In terms of performance/Watt/area
- One of the candidates
  - FIM to accommodate performance/Watt/area efficient function elements
  - in memory to process big data nearby



# Summary

- Memory System is a Key component for AI application
  - 3D-Stacked DRAM Technology and 2.5D Process are powerful tool
- HBM DRAMs are using for providing huge bandwidth for AI
- Difficulties and current solutions of HBM development
  - Thermal, Power and Testability
- HBM-based FIM is a good candidate to overcome memory bottleneck
  - Circuit researchers must work w/ computer architects
  - System software and application developers

**Title**

*Novel Memory Technologies for Advanced CMOS Nodes*

**Abstract**

*Recent years witnessed rapid acceleration in the development of new memory technologies. Several of them have reached performance and yield levels suitable for high volume manufacturing and commercialization. This work covers the device physics and array-level characterization for STT-MRAM and RRAM, and discusses the suitability of these technologies for solving embedded memory needs of advanced CMOS nodes in Non Volatile and High-Bandwidth-High-Endurance application space.*

**Bio**

*Dr. Oleg Golonzka is a Principal Engineer at Intel Corporation. He received his BA/MS from Moscow Institute of Physics and Technology and completed his Doctoral and Post-Doctoral studies in Physics at The Pennsylvania State University and Massachusetts Institute of Technology. Oleg Golonzka joined Portland Technology Development Division of Intel Corporation in 2001 and worked on device and process integration of Intel's 65nm, 32nm, and 14nm CMOS nodes. Most recently he led the device development and process integration of Intel's STT-MRAM and RRAM-based embedded Non Volatile Memory technologies.*

# **Novel Memory Technologies for Advanced CMOS Nodes**

Instructor: Oleg Golonzka, Intel Corporation

- Embedded memory landscape
- MRAM and RRAM cells and integration into logic technology
- MRAM physics and technology development details
- RRAM physics and endurance challenge
- Embedded non-volatile memory: MRAM vs RRAM
- MRAM: beyond nonvolatile memory applications



# Novel Memory Technologies for Advanced CMOS Nodes

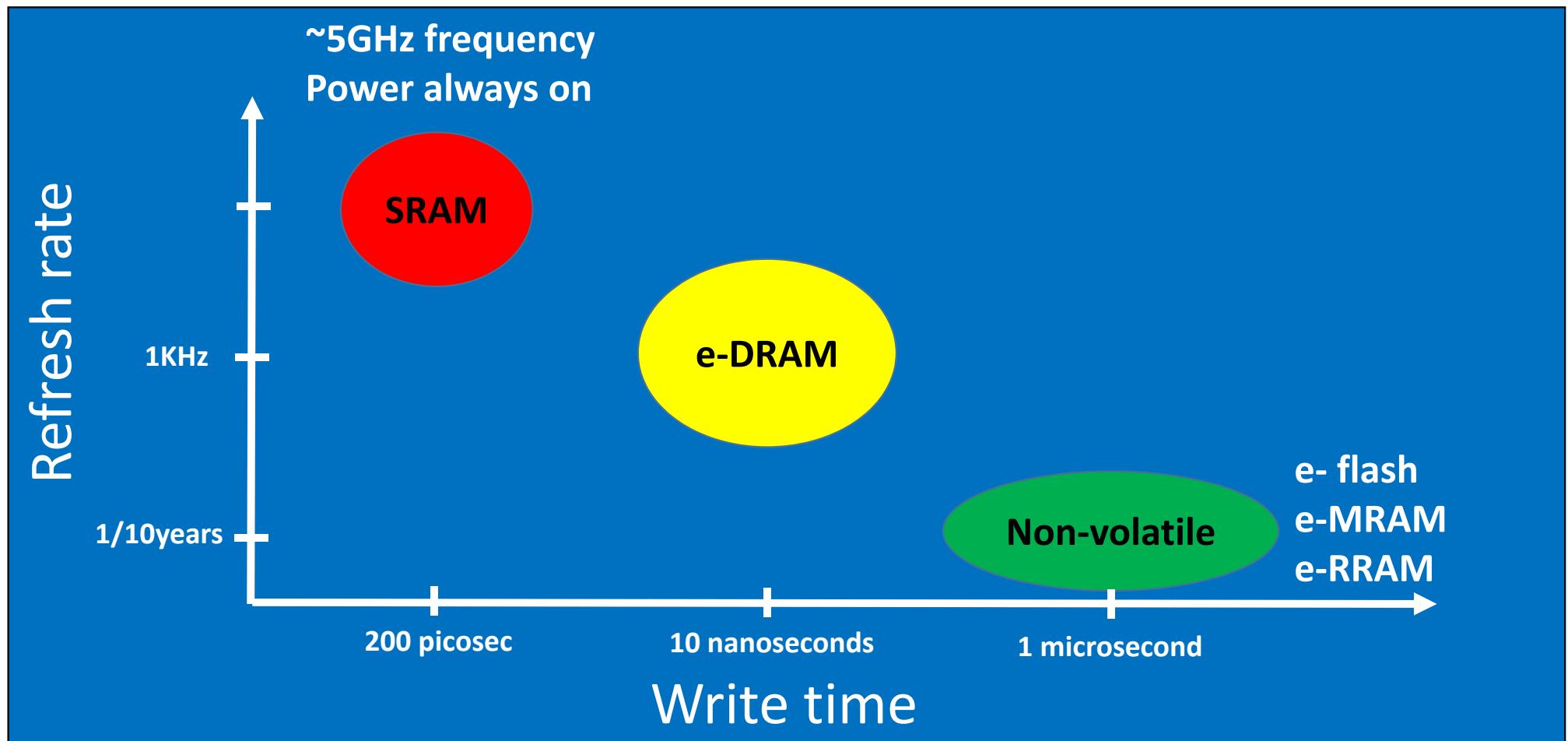
Oleg Golonzka

*Intel Corporation*

# Outline of Presentation

- Embedded memory landscape
- MRAM and RRAM cells and integration into logic technology
- MRAM physics and technology development details
- RRAM physics and endurance challenge
- Embedded non-volatile memory: MRAM vs RRAM
- MRAM: beyond nonvolatile memory applications

# Embedded Memory landscape

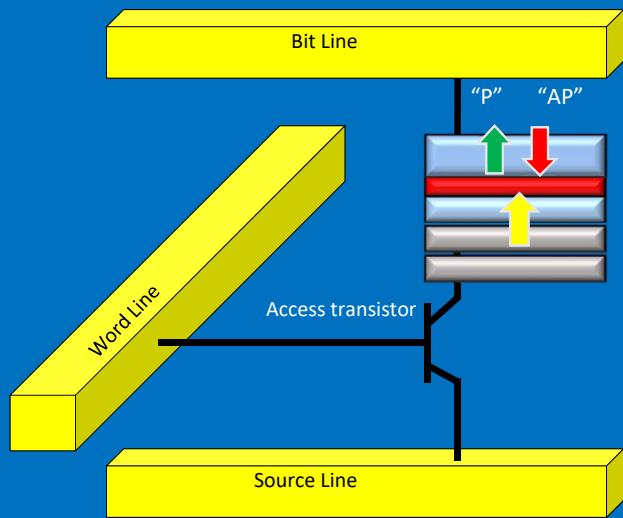


# Outline of Presentation

- Embedded memory landscape
- **MRAM and RRAM cells and integration into logic technology**
- MRAM physics and technology development details
- RRAM physics and endurance challenge
- Embedded non-volatile memory: MRAM vs RRAM
- MRAM: beyond nonvolatile memory applications

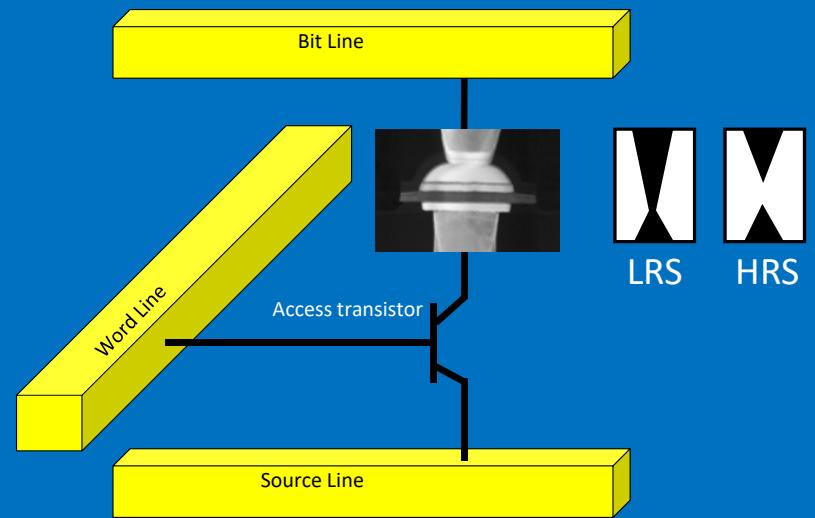
# Cell Layout, 1T-1R

MRAM



Parallel "P" = Low resistance state  
Anti-Parallel "AP" = High resistance state

RRAM

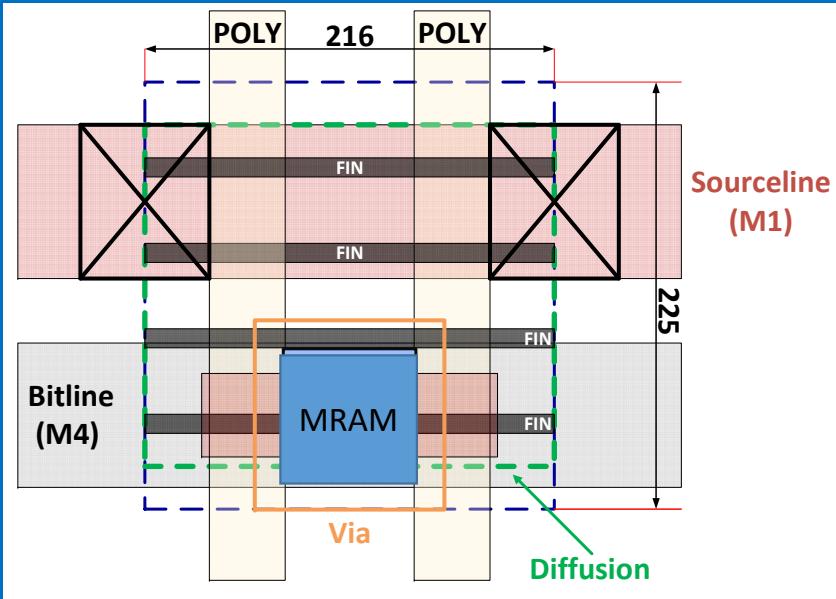


"LRS" = Low resistance state  
"HRS" = High resistance state

# Cell Layout

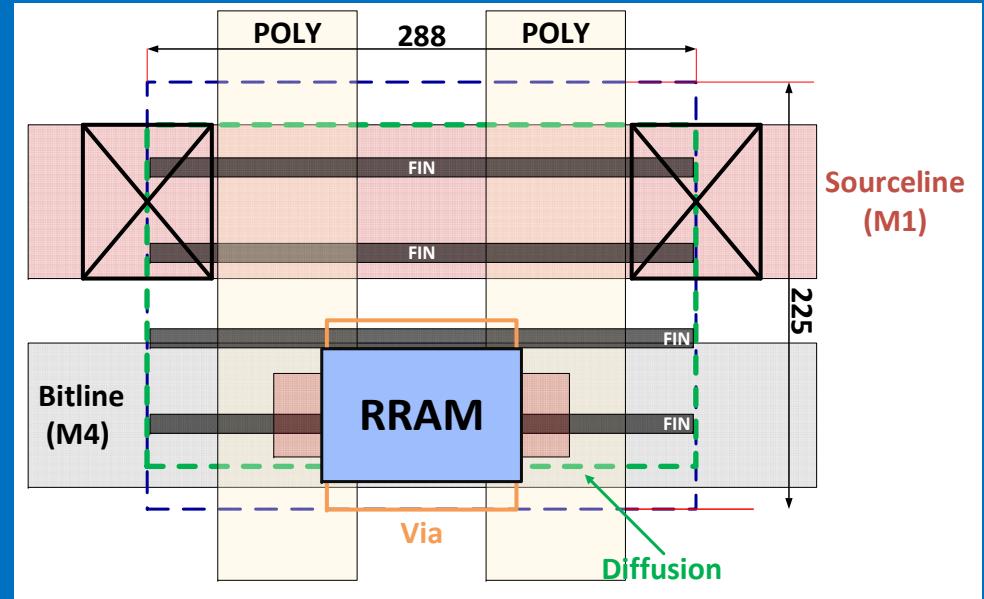
MRAM

216nm x 225nm cell

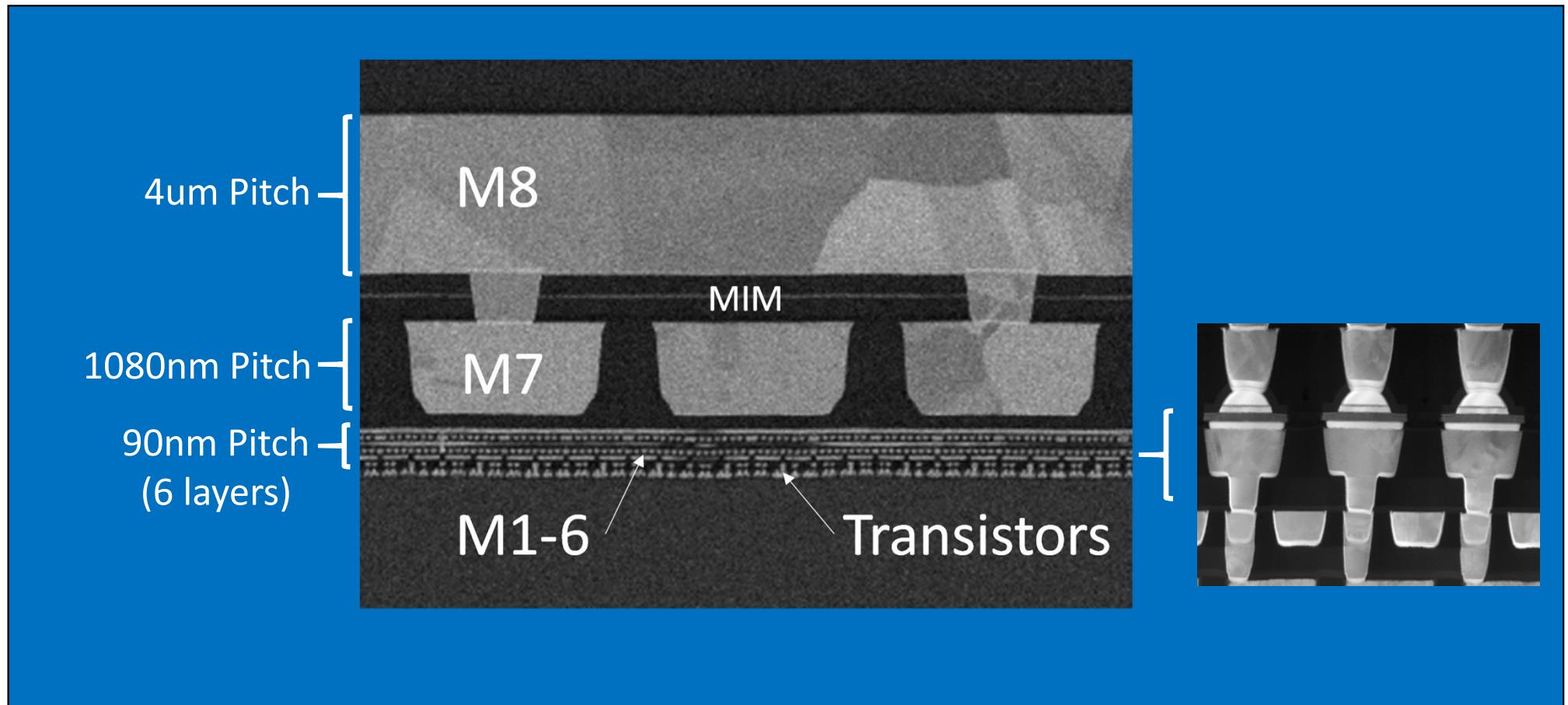


RRAM

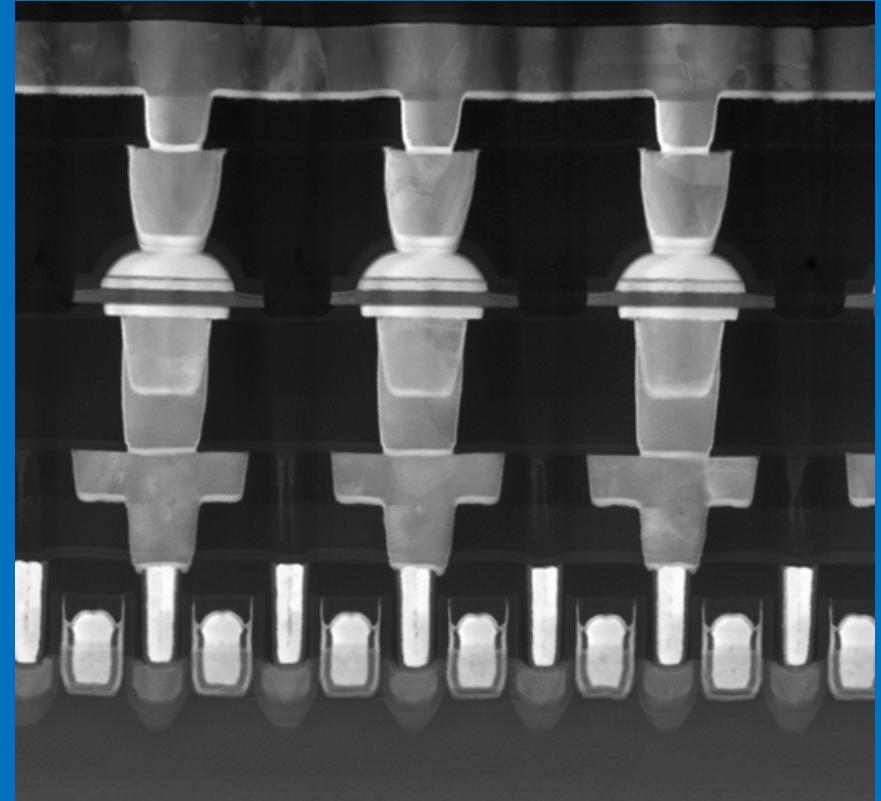
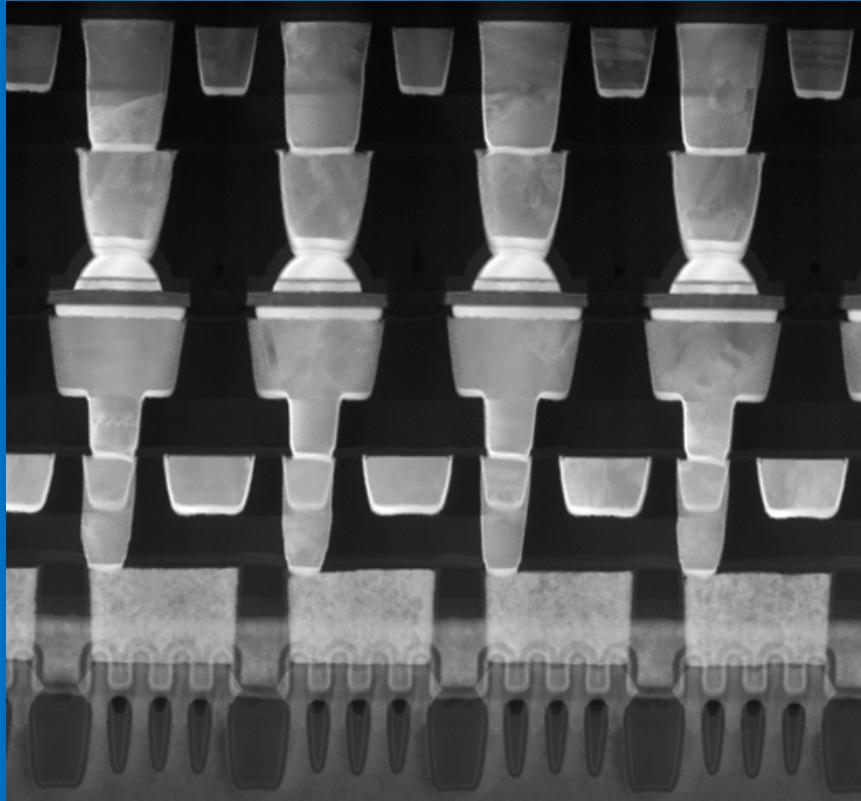
288nm x 225nm cell



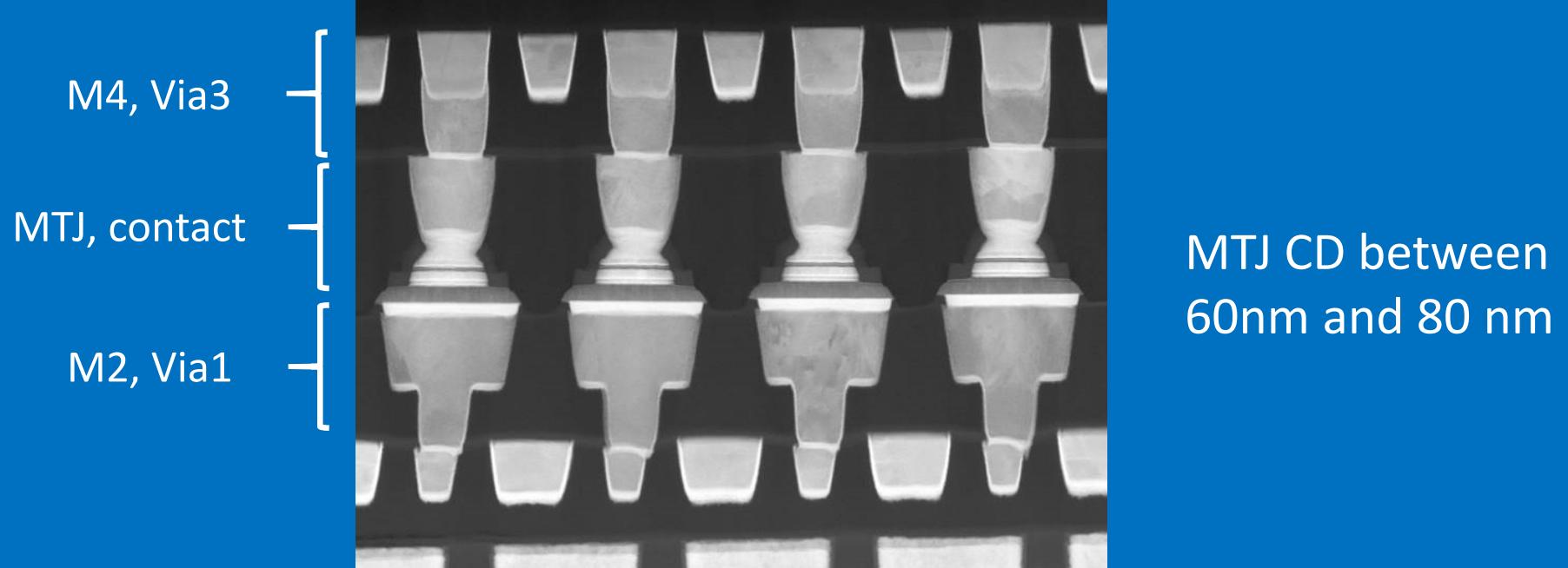
# 22nm Logic



## RRAM: 288nm x 225nm cell



# MRAM Array, embedded between M2 and M4



MTJ CD between  
60nm and 80 nm

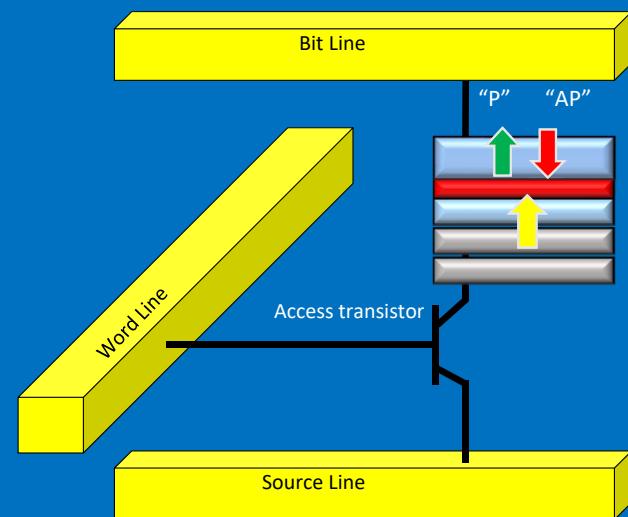
Embedding of MRAM does not affect transistor or interconnect performance

*O. Golonzka et al., 2018 IEEE International Electron Devices Meeting (IEDM)*

# Outline of Presentation

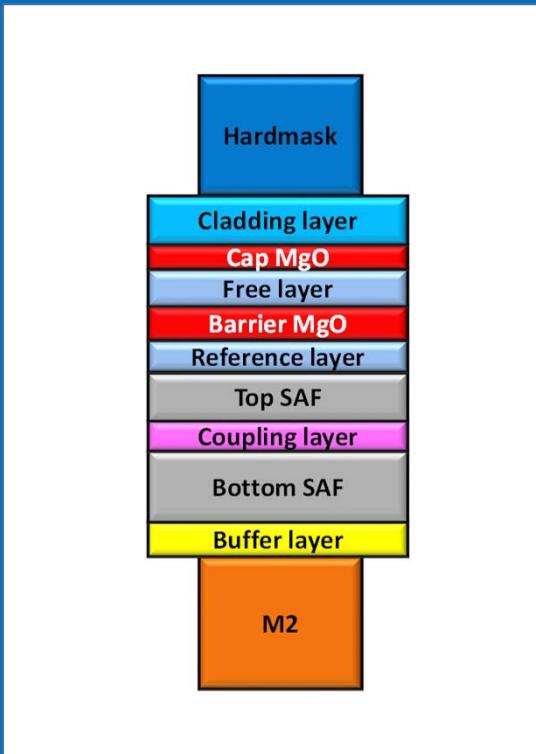
- Embedded memory landscape
- MRAM and RRAM cells and integration into logic technology
- **MRAM physics and technology development details**
- RRAM physics and endurance challenge
- Embedded non-volatile memory: MRAM vs RRAM
- MRAM: beyond nonvolatile memory applications

# MTJ: memory element for MRAM



Parallel "P" = Low resistance state  
Anti-Parallel "AP" = High resistance state

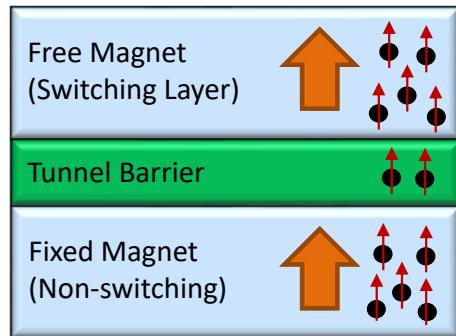
Dual MgO stack



# Basic MRAM operation

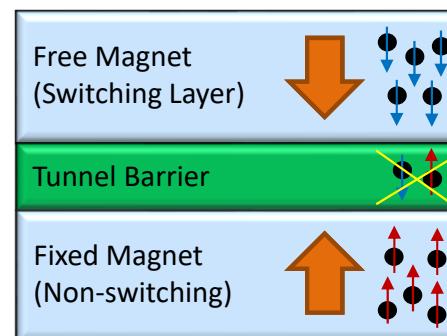
## State “0”

- **Low resistance,  $R_p$**
- Free and Fixed layers are **parallel**.
- Electrons can easily tunnel.



## State “1”

- **High resistance,  $R_{AP}$**
- Free and Fixed layers are **anti-parallel**.
- Tunneling current is reduced.

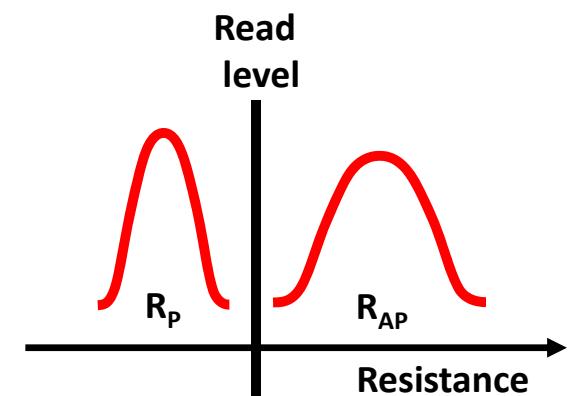


TMR = Tunneling Magnetoresistance

$$TMR = \frac{R_{ap} - R_p}{R_p}$$

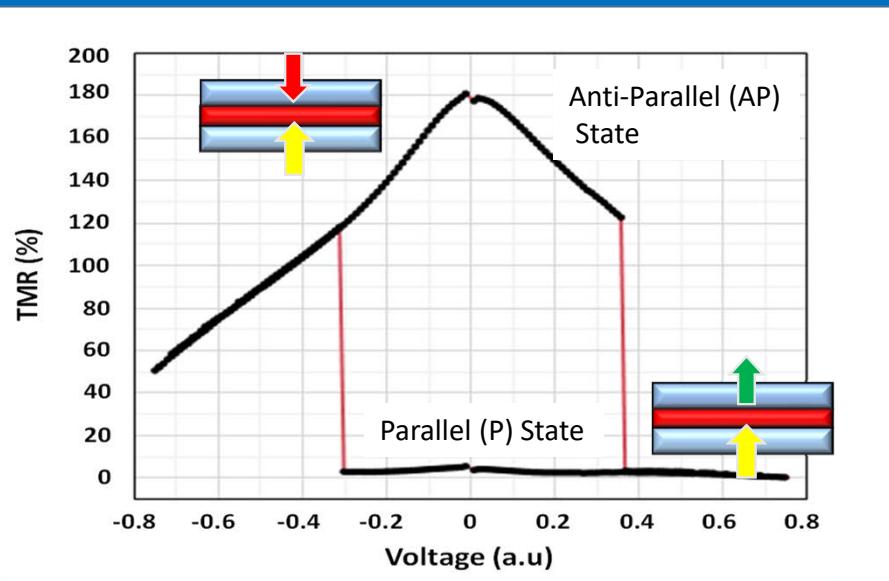
100% TMR = 2x change in Resistance

200% TMR = 3x change in Resistance



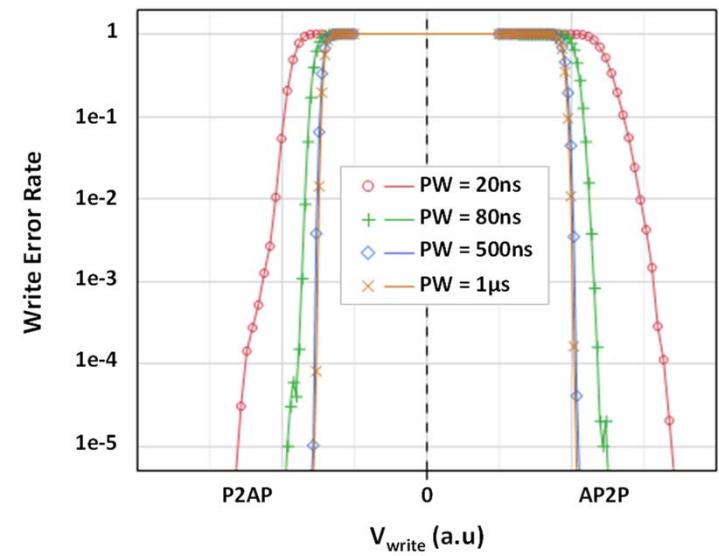
# MTJ: Typical Device Characteristics

R-V response



TMR = 180%

Write error rates

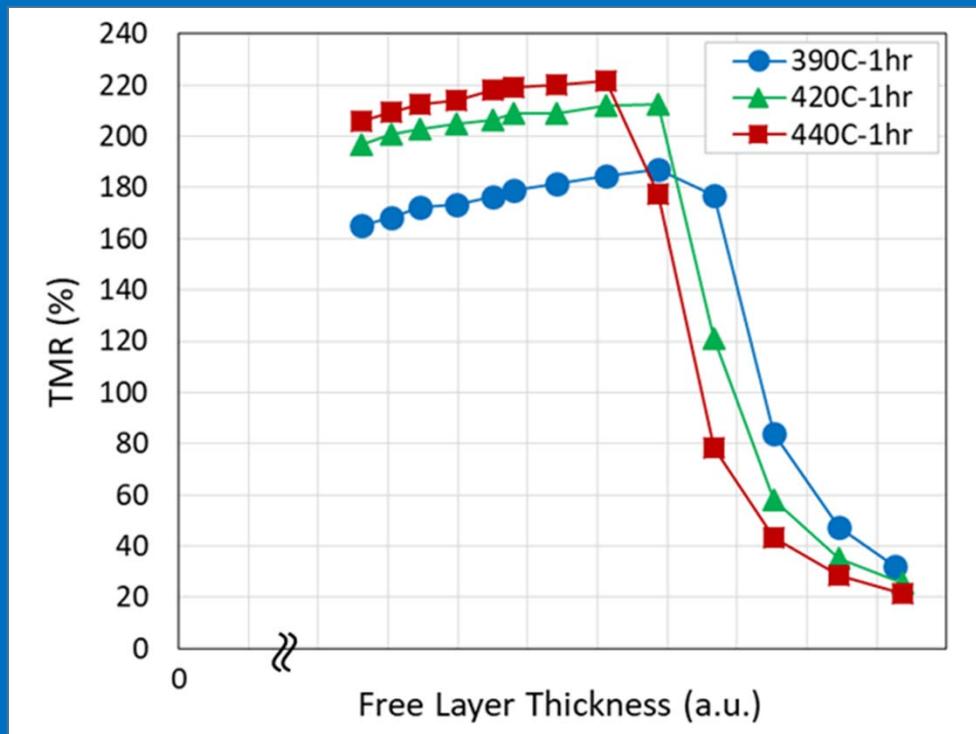


Shorter pulses  $\rightarrow$  higher switching voltage

O. Golonzka et al., 2018 IEEE International Electron Devices Meeting (IEDM)

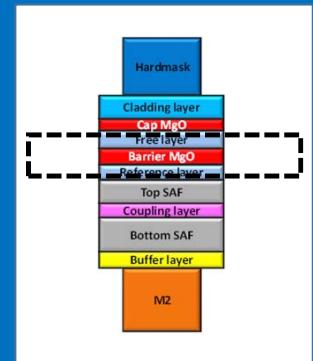
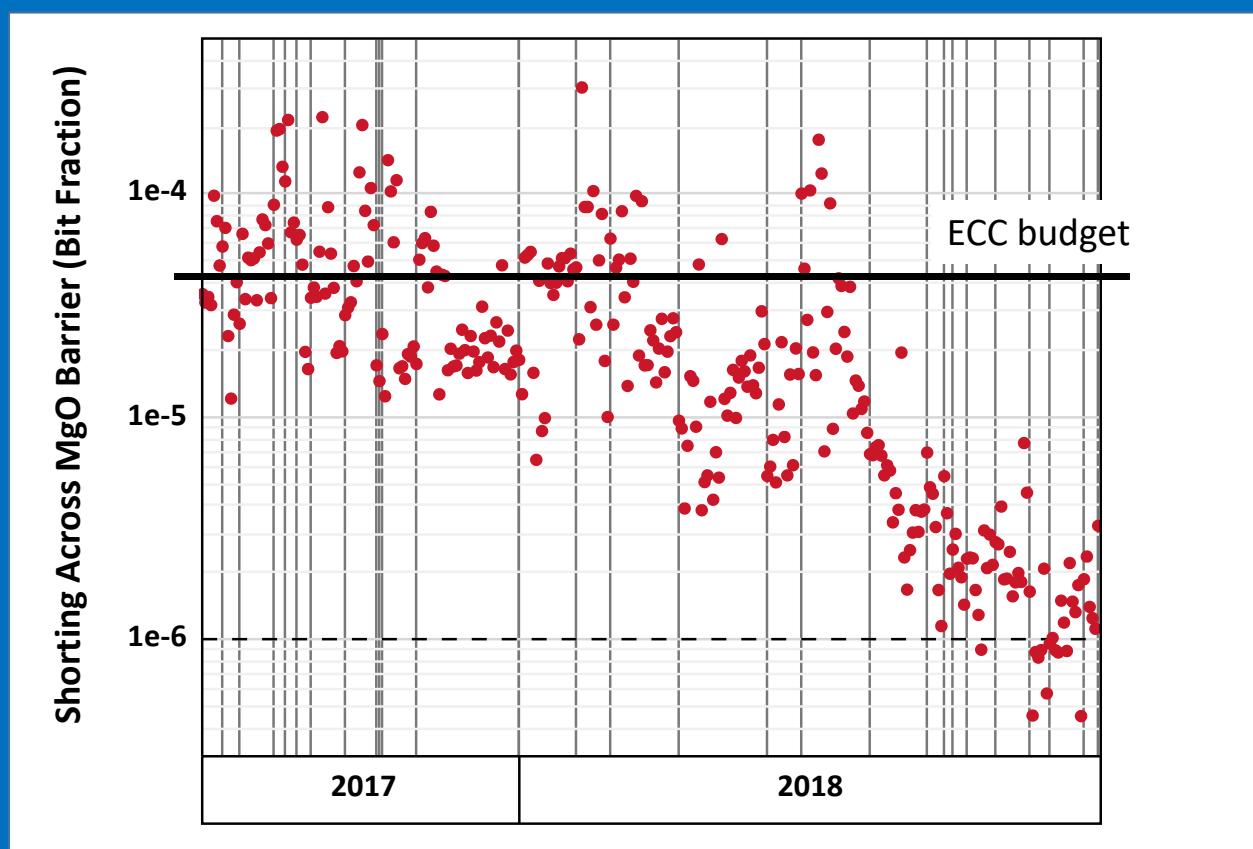
# MTJ: Achieving high thermal stability of the stack

CIPT (unpatterned wafer) data with high temperature anneals



O. Golonzka et al., 2018 IEEE International Electron Devices Meeting (IEDM)

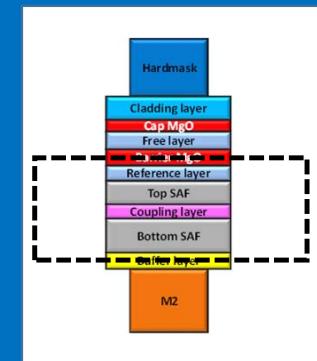
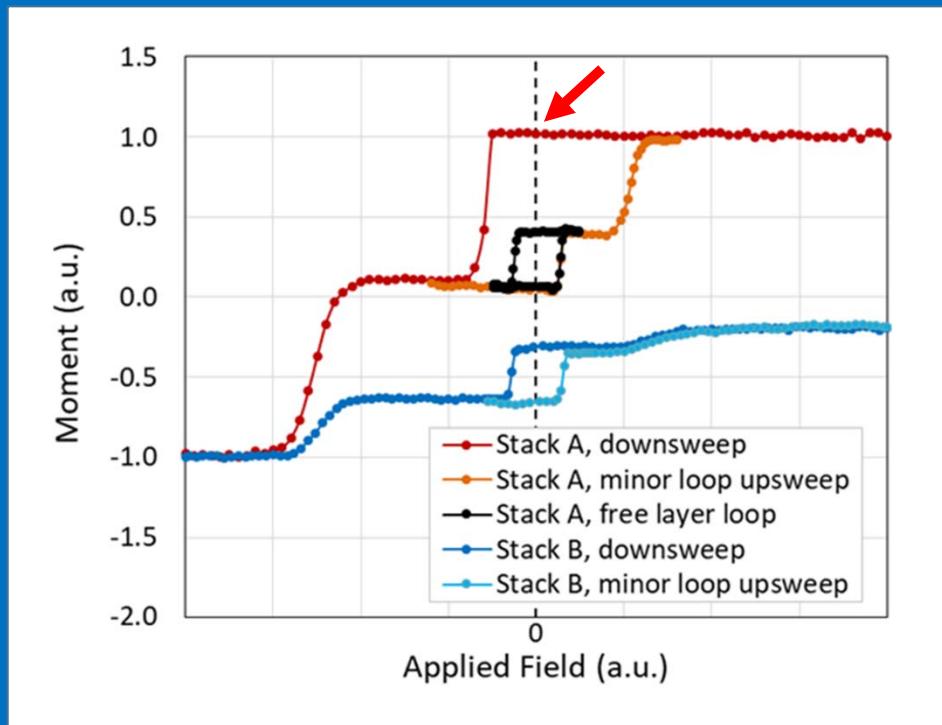
# Eliminating shorting across the MgO barrier



# MTJ: Eliminating SAF locking

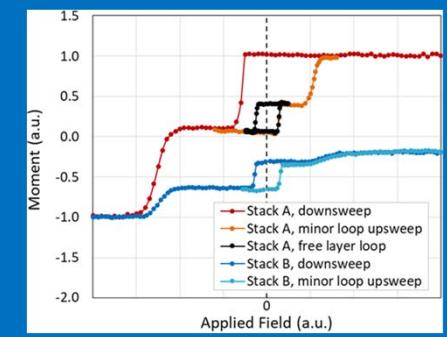
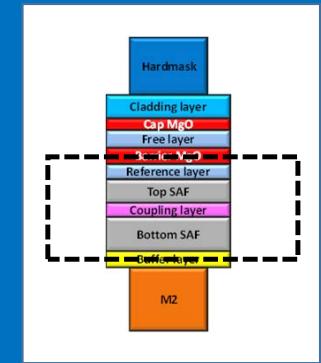
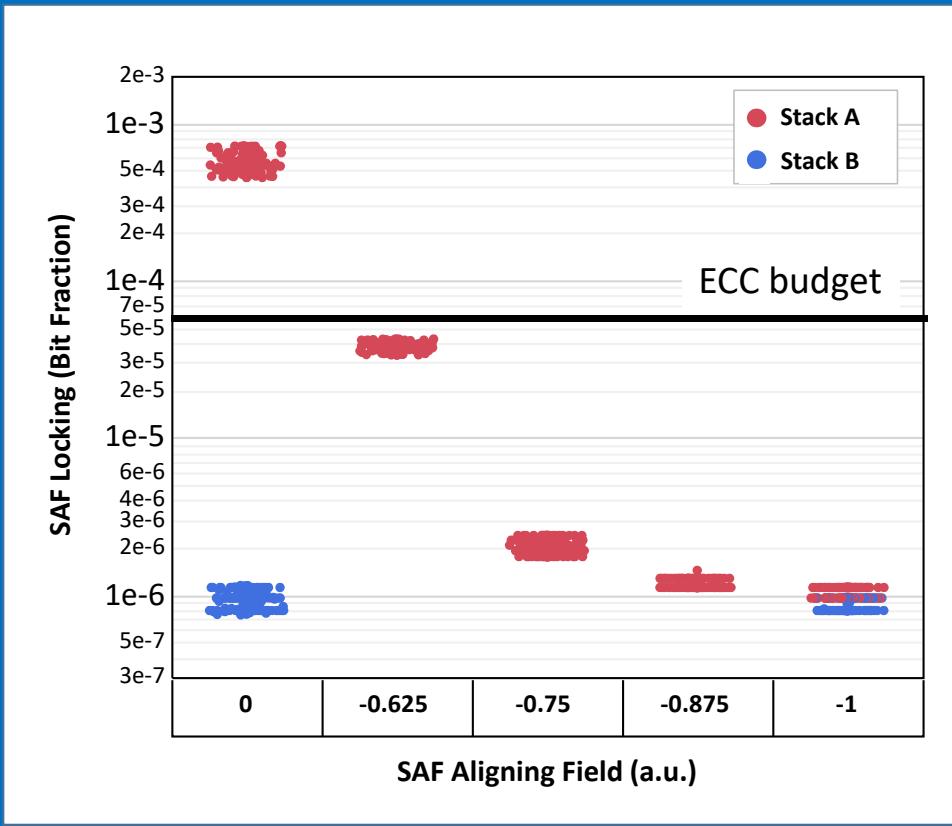
## M-H curves of patterned MTJ arrays

- 3 stable states at “No Field” for a typical stack
- 2 stable states for Advanced Stack B at “No Field”



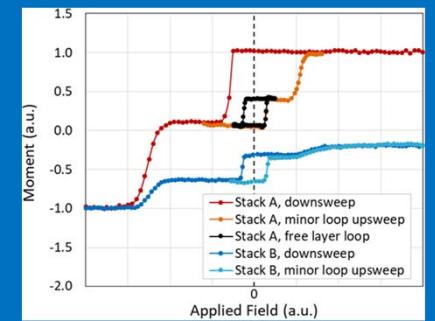
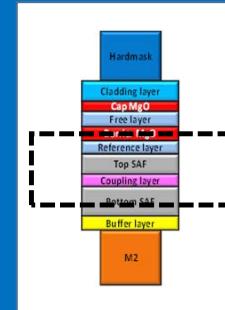
## 7.2 Mbit array: Eliminating SAF locking

- SAF Aligning Field exposure is needed for a typical stack
- No SAF aligning is needed for Advanced Stack B

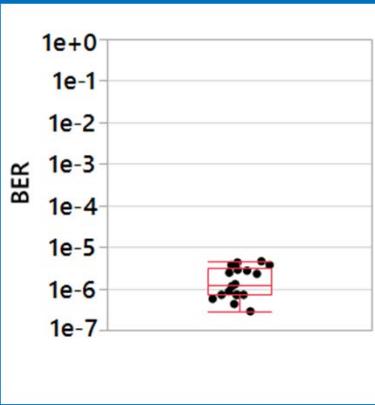


**High magnetic field can erase the data, but memory remains fully functional**

## Stack B: data erase experiment



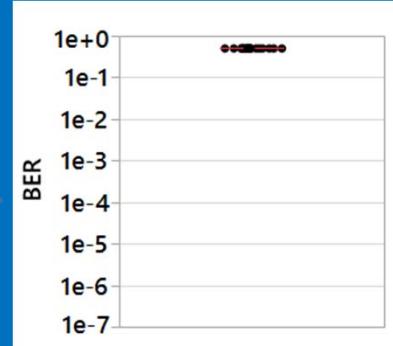
Write + Read,  
Low bit error rate



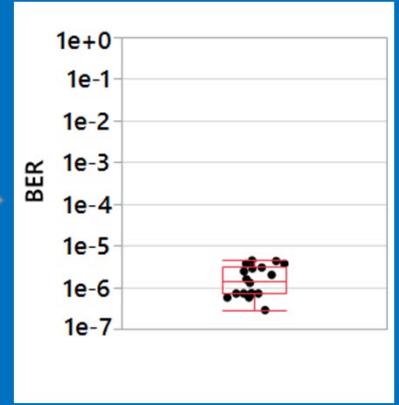
Erase with 1 Tesla  
magnet



Read,  
Data has been erased

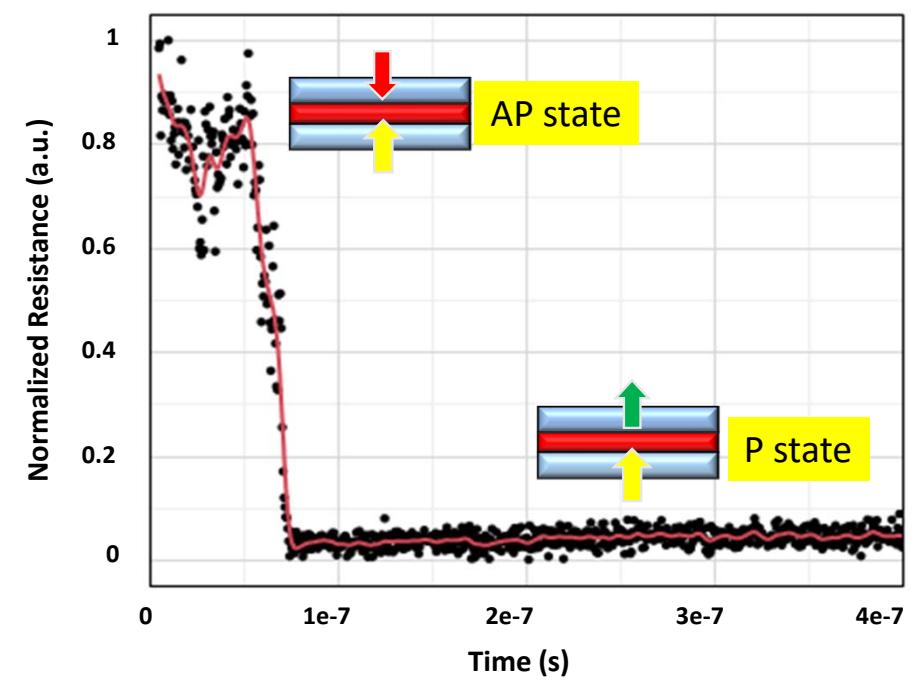


Write + Read,  
Low bit error rate

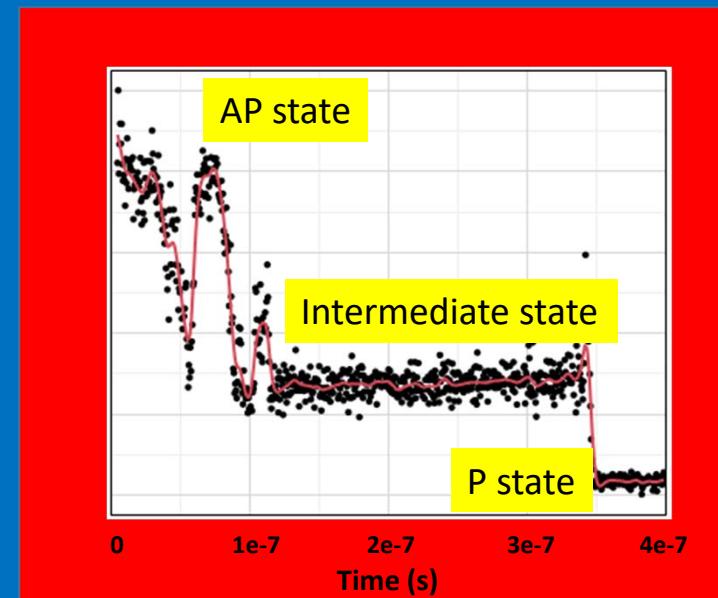


# Time resolved resistance traces: intermediate states

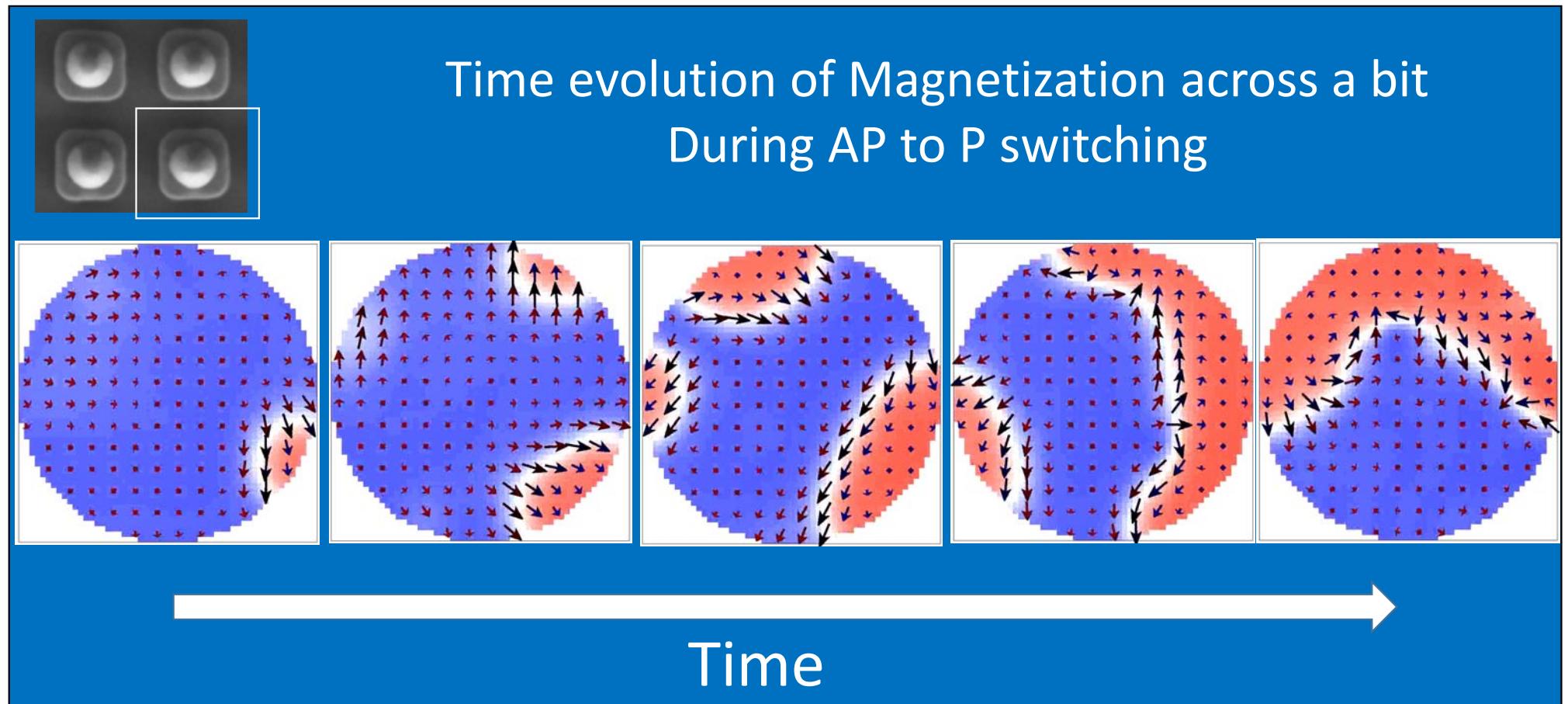
Typical bit



Bit exhibiting an intermediate state

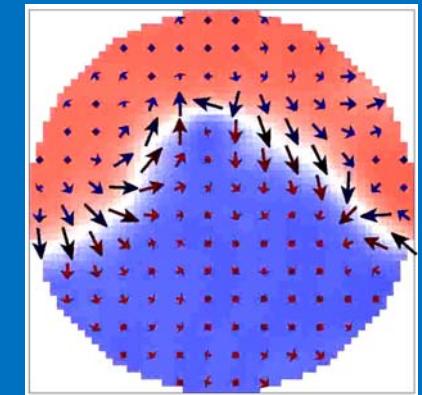
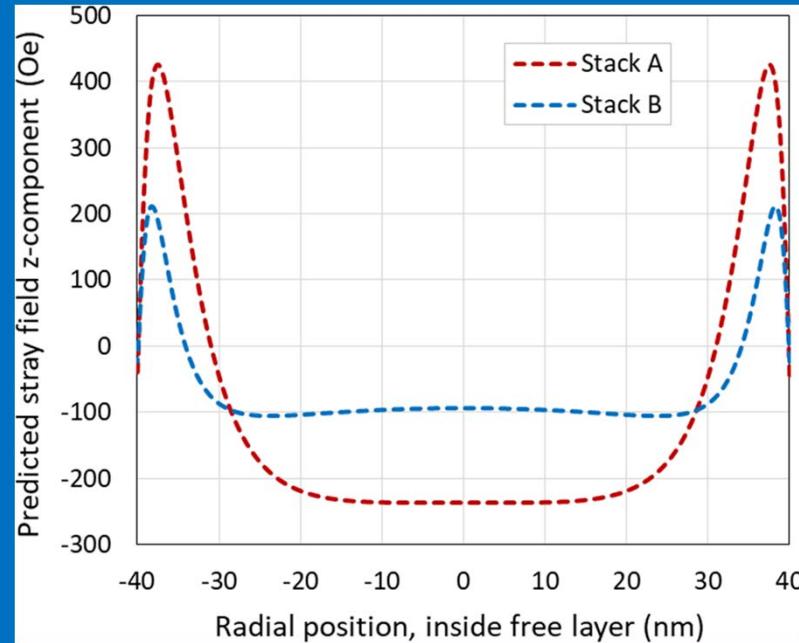
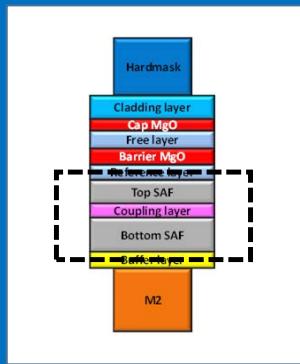


# Simulation: origin of the intermediate states



# Origin of the intermediate states: multi-domain structure

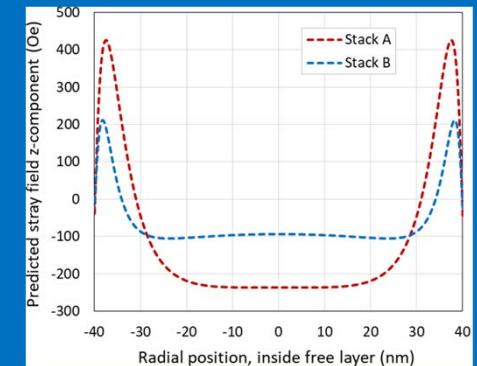
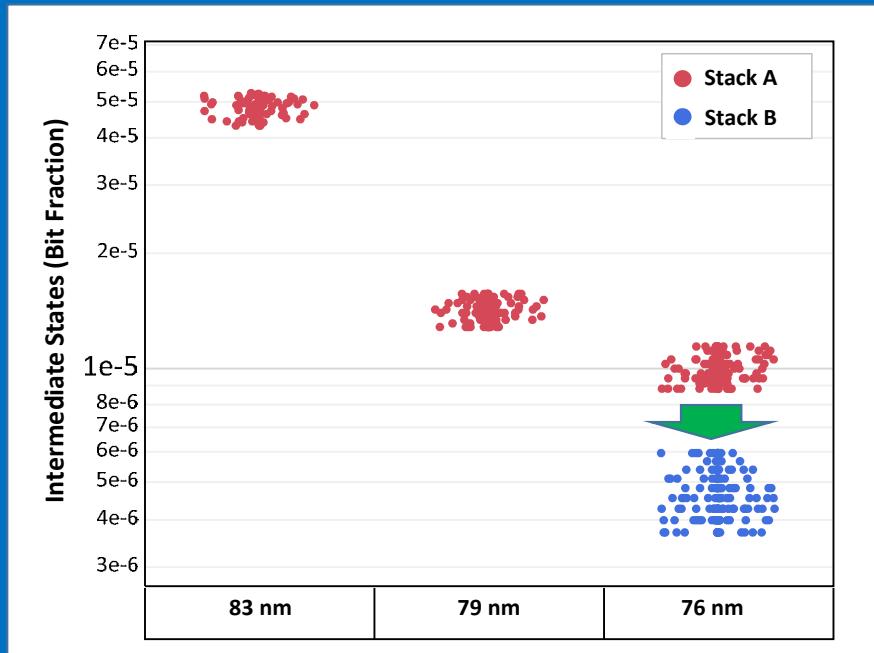
SAF-induced stray field across the device



Substantially improved within-bit stray field non-uniformity for Advanced Stack B

## 7.2 Mbit array: Eliminating Intermediate States

### Intermediate states vs device CD



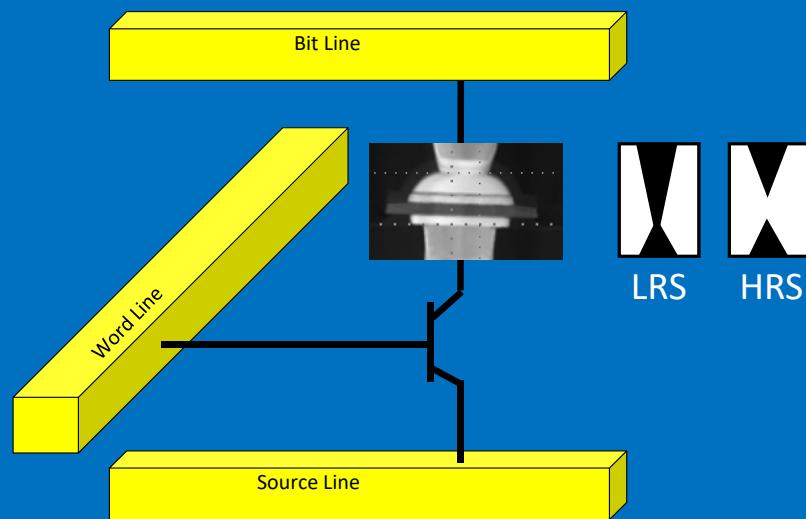
Significant reduction in Intermediate States for Advanced Stack B

*O. Golonzka et al., 2018 IEEE International Electron Devices Meeting (IEDM)*

# Outline of Presentation

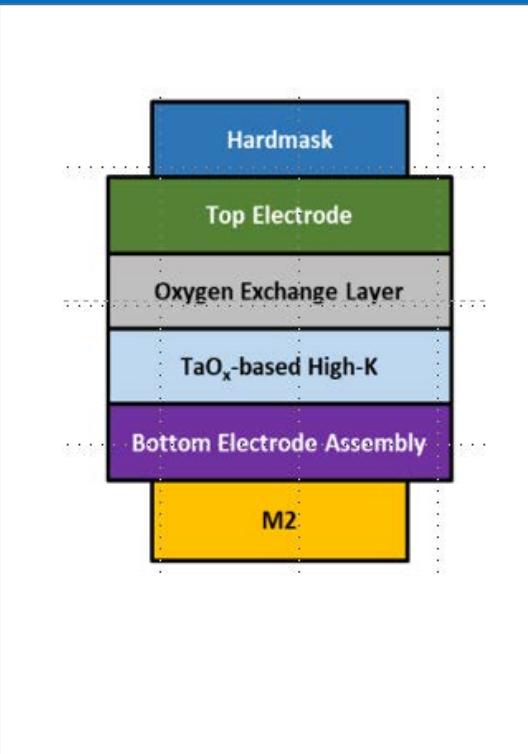
- Embedded memory landscape
- MRAM and RRAM cells and integration into logic technology
- MRAM physics and technology development details
- **RRAM physics and endurance challenge**
- Embedded non-volatile memory: MRAM vs RRAM
- MRAM: beyond nonvolatile memory applications

# RRAM cell memory element

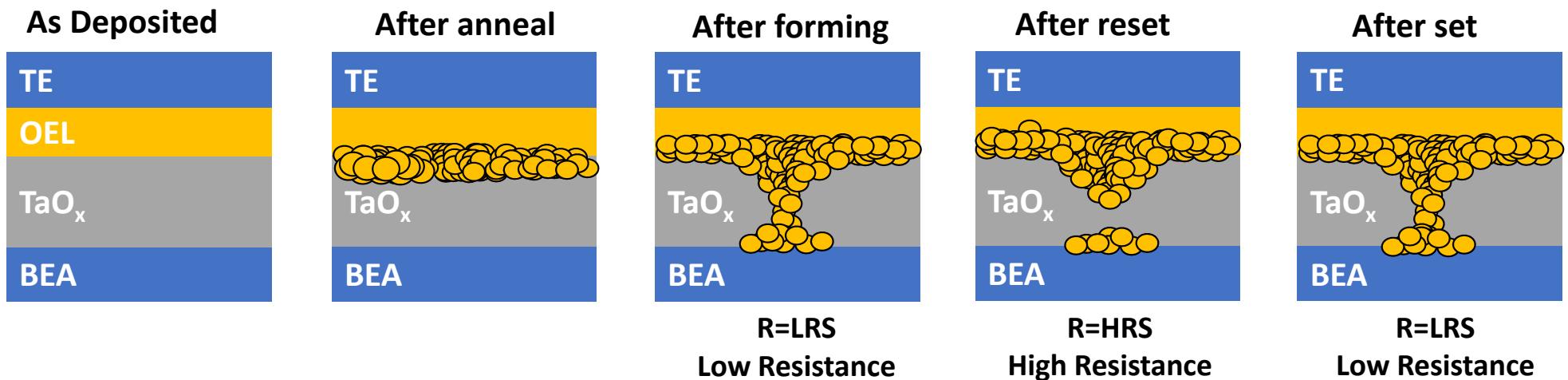


“LRS” = Low resistance state  
“HRS” = High resistance state

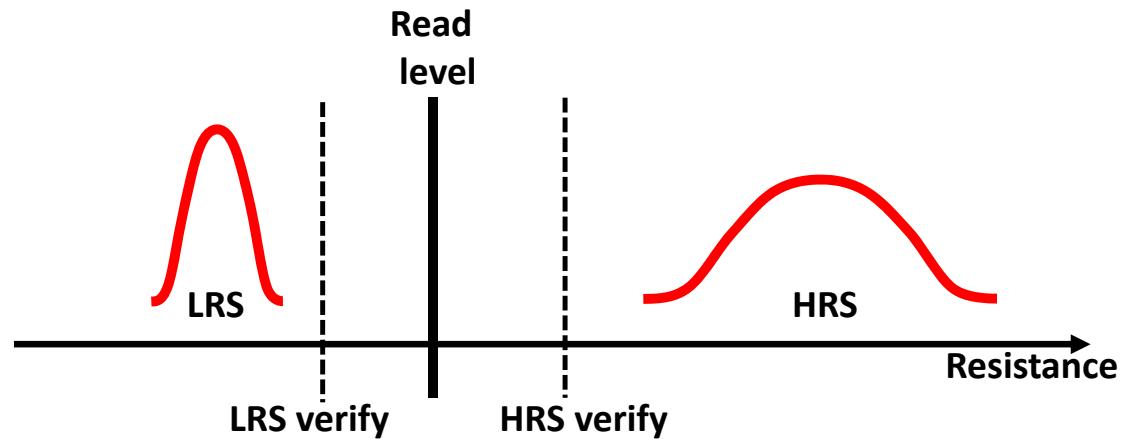
TaO<sub>x</sub> - based stack



# Basic RRAM operation

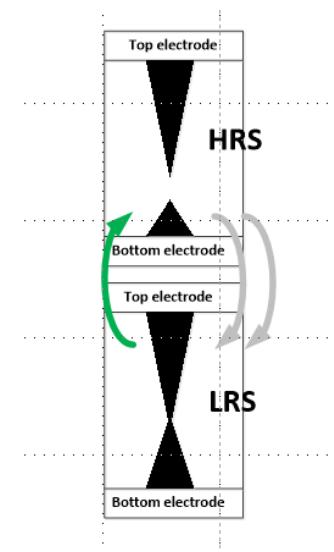
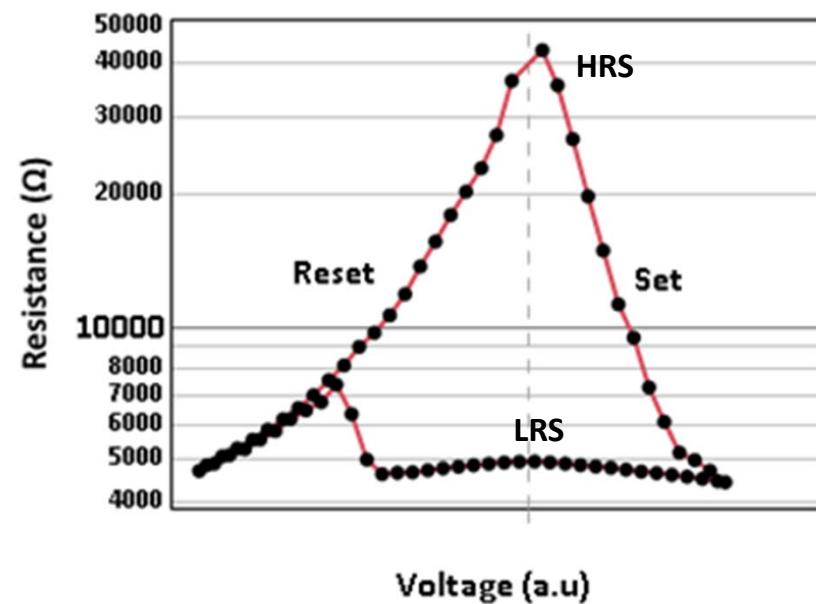


- Write with Write-Verify-Write scheme
- Read at 100mV to avoid read disturb



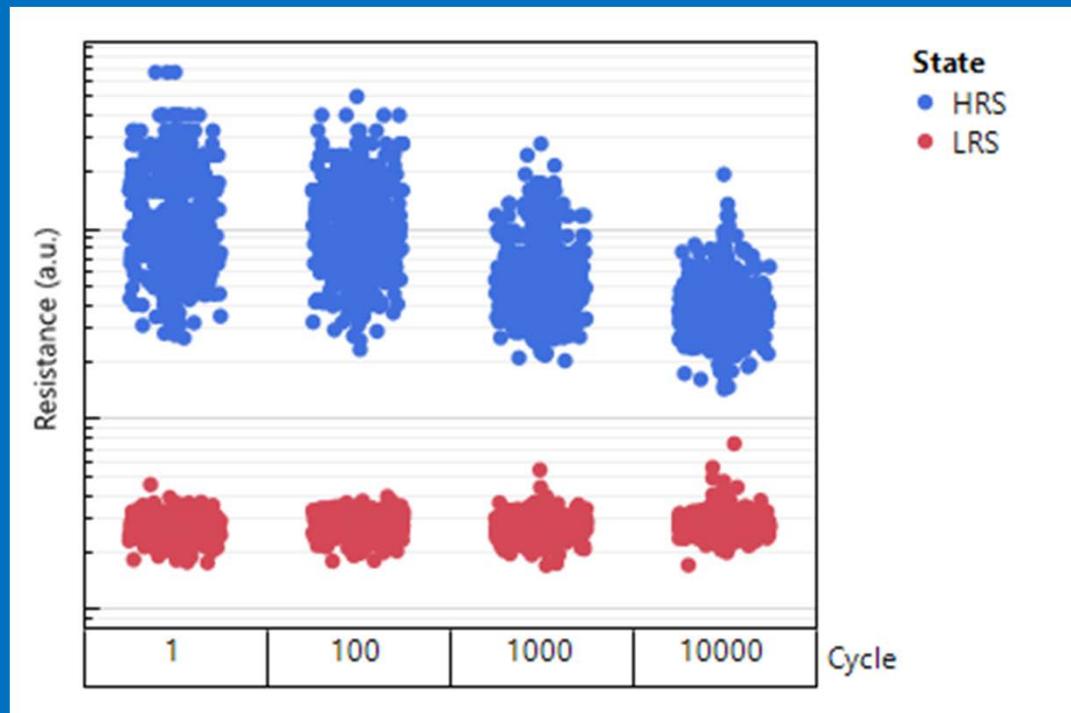
# RRAM: Typical Device Characteristics

## R-V response



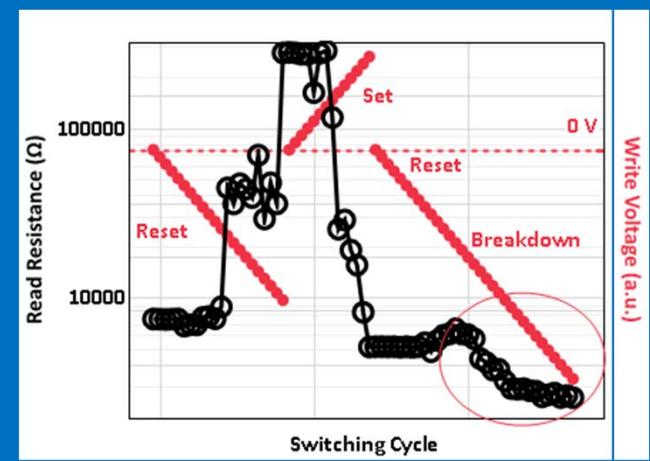
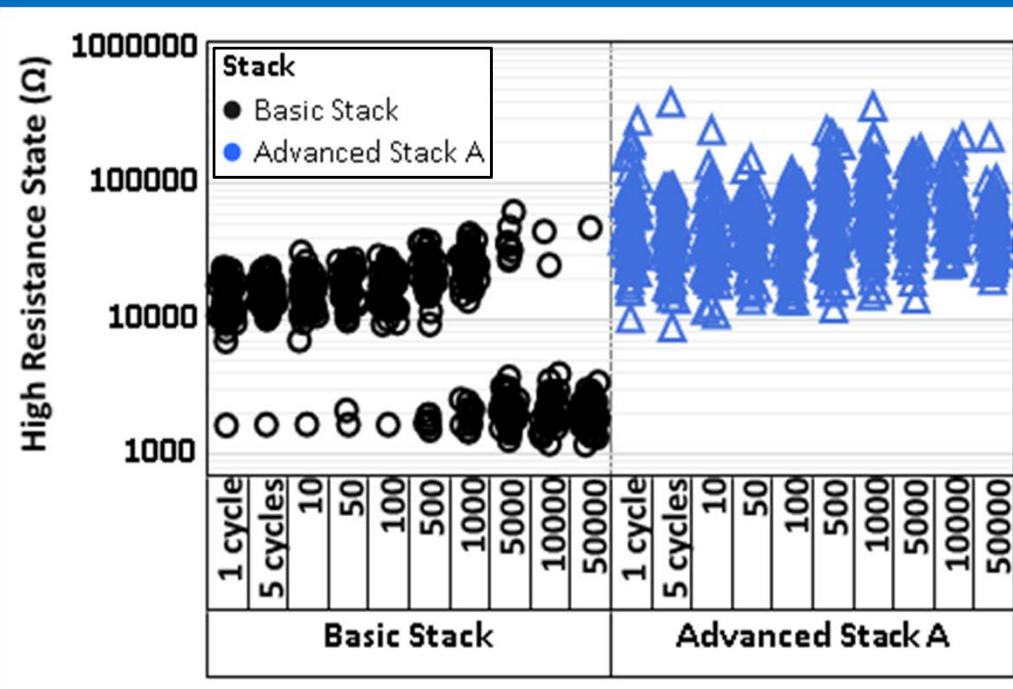
O. Golonzka et al., 2019 Symposium on VLSI Technology

# RRAM challenge: Cycling between LRS and HRS modifies device characteristics

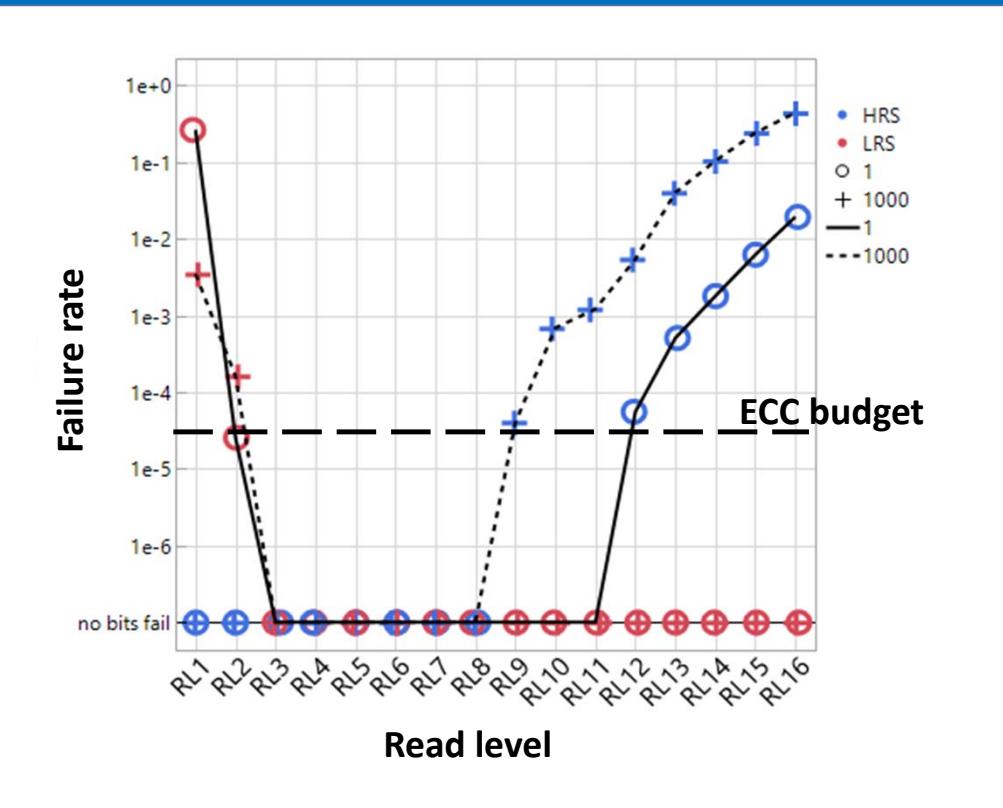


O. Golonzka et al., 2019 Symposium on VLSI Technology

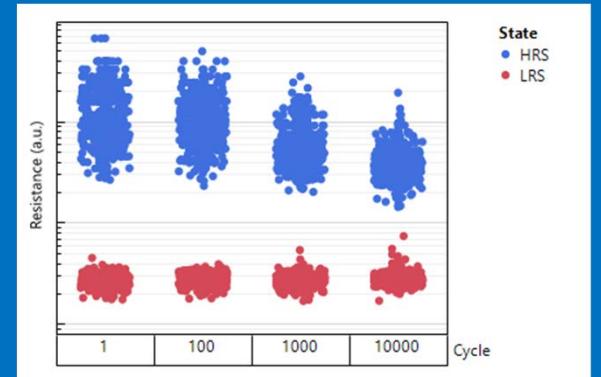
# Endurance failure due to vacancy accumulation at the bottom interface



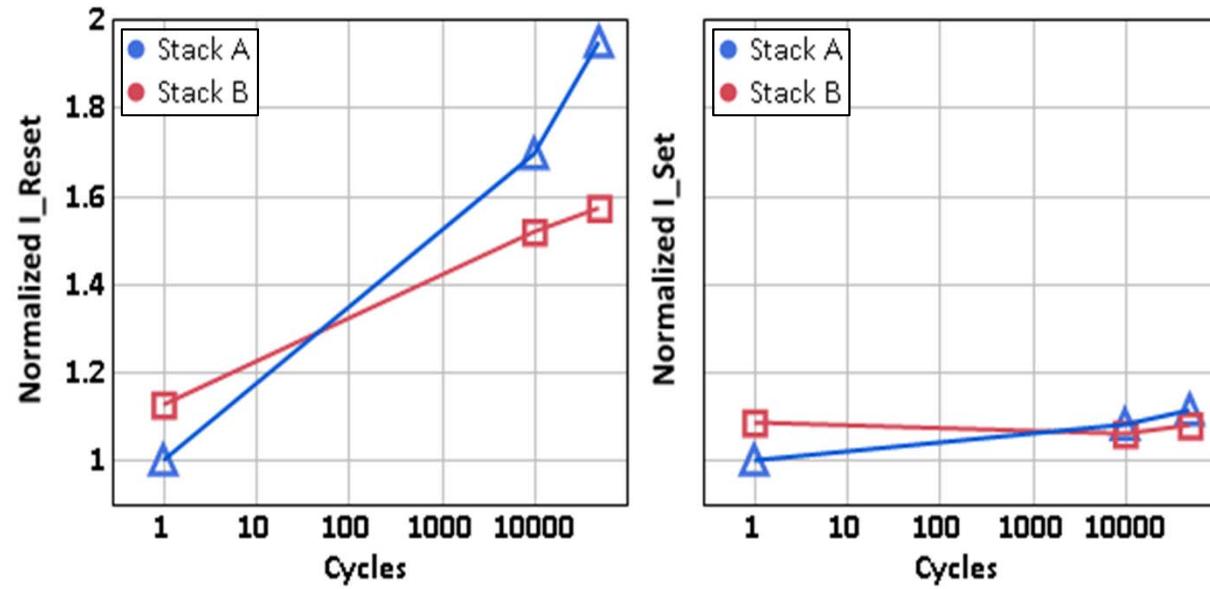
# Cycling RRAM bits closes the resistance window between LRS and HRS states



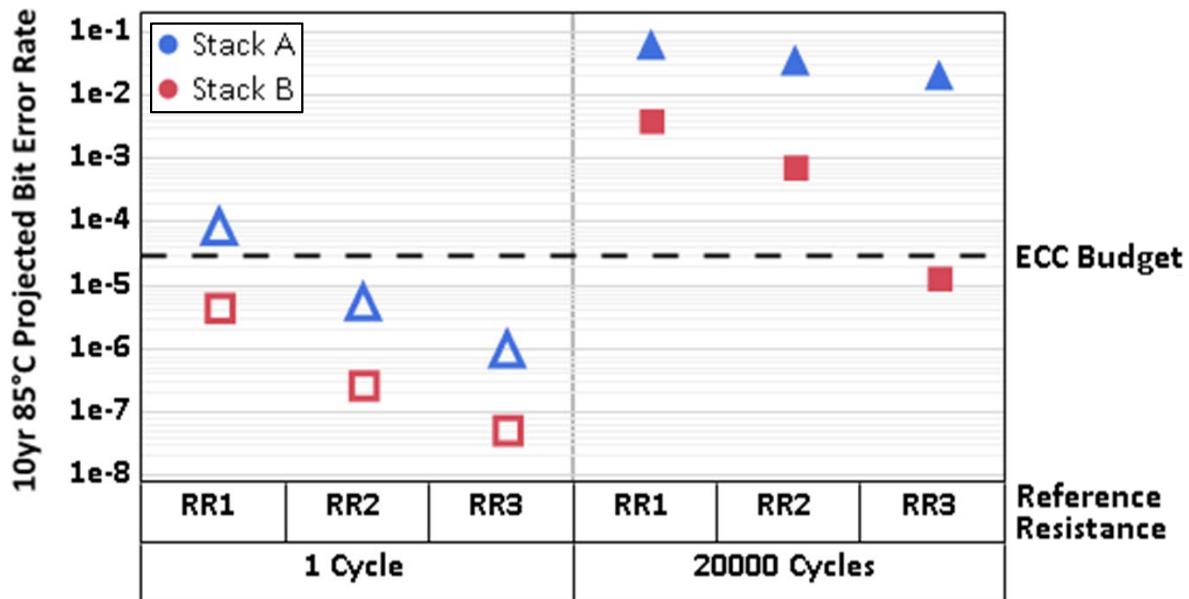
Endurance is limited by  
the degradation of the  
HRS State



# A better stack has less degradation of the Reset Current with cycling

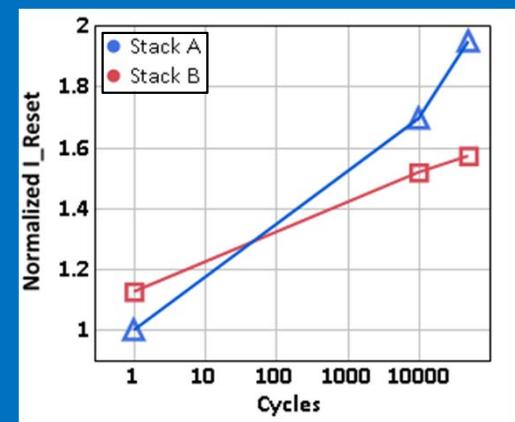


# Retention: fresh bits vs cycled bits: Cycled bits show dramatically worse retention



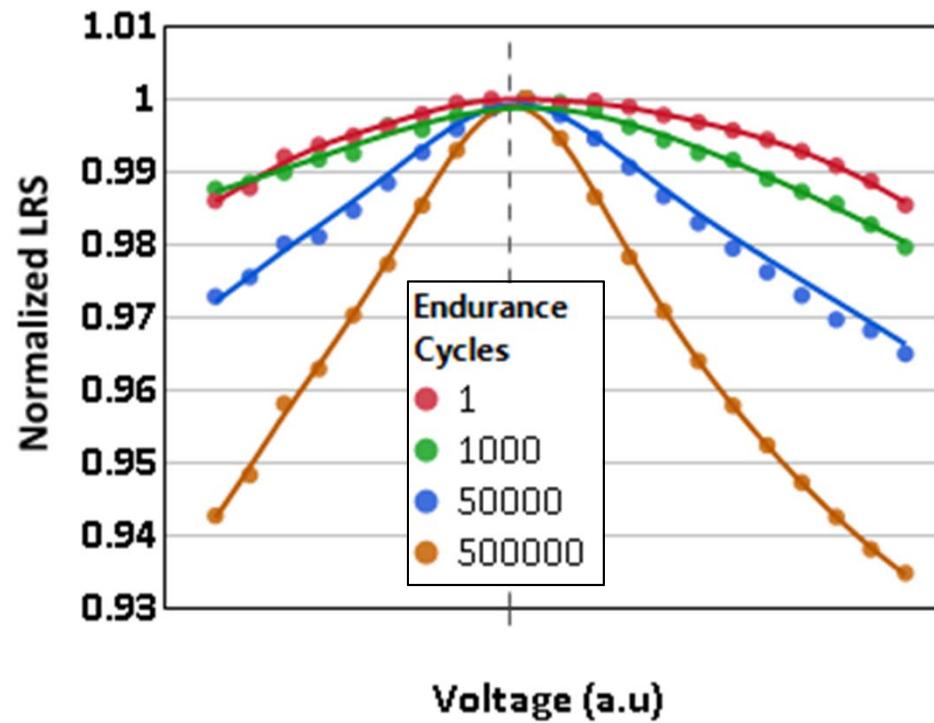
Data from High Density Array

Retention is limited by the degradation of the LRS state



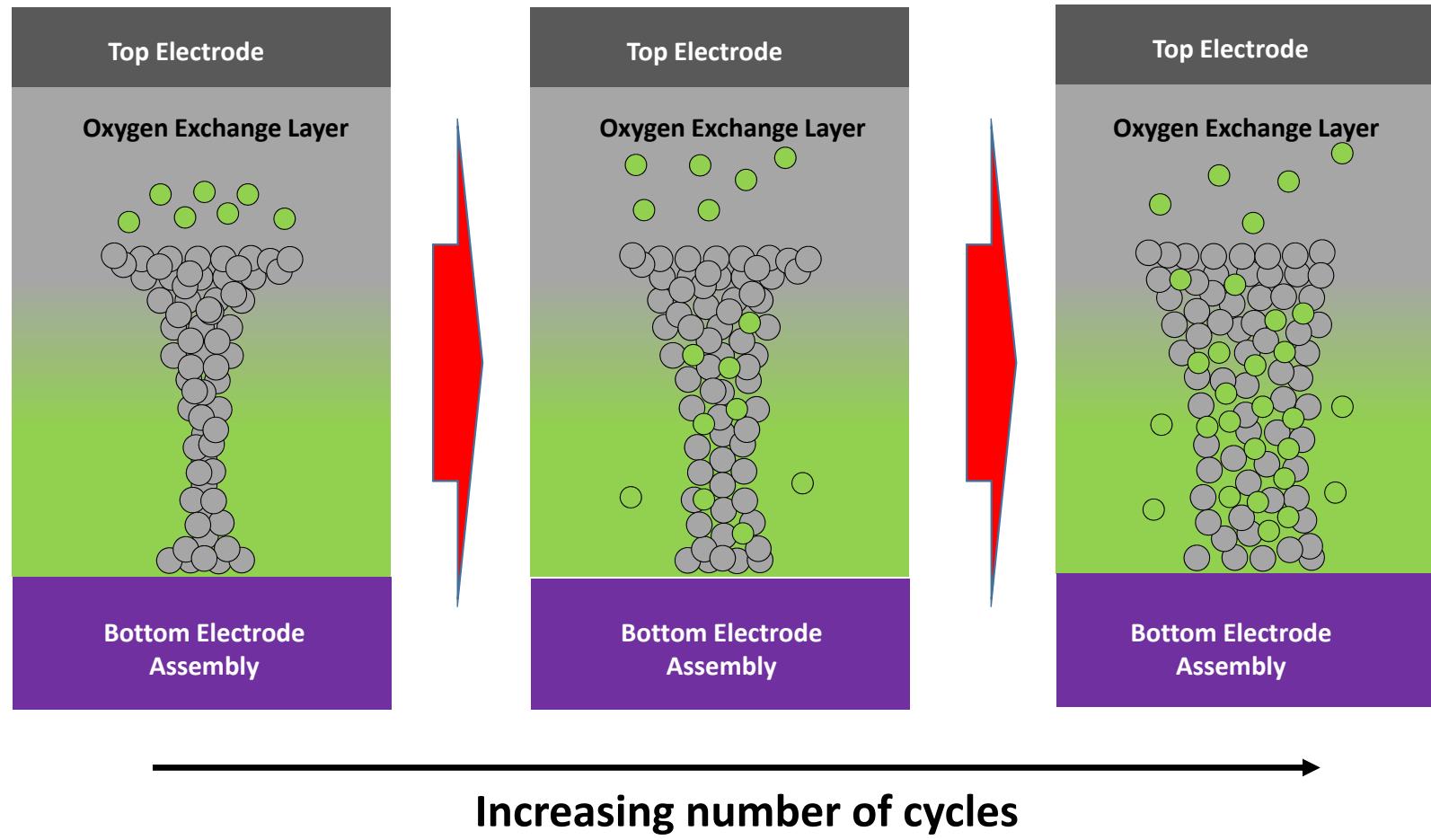
O. Golonzka et al., 2019 Symposium on VLSI Technology

# Filament R-V characteristics of cycled bits



O. Golonzka et al., 2019 Symposium on VLSI Technology

# What is happening to the filament during cycling?

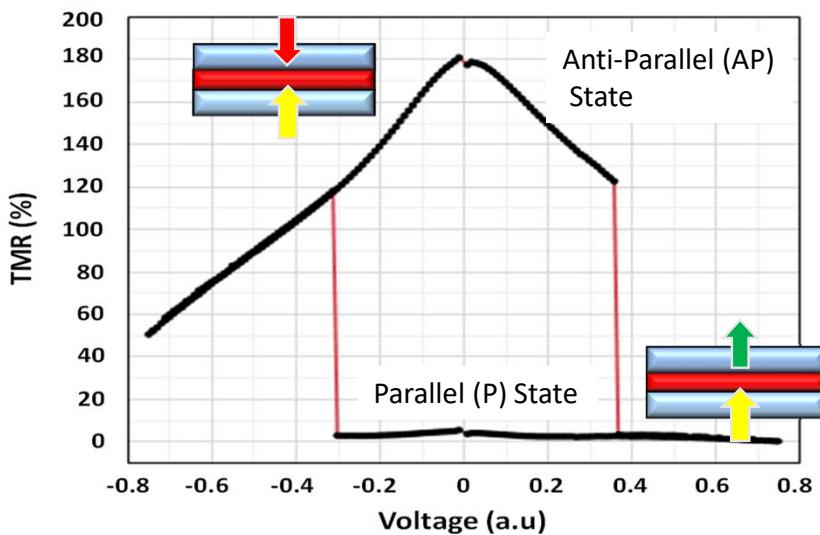


# Outline of Presentation

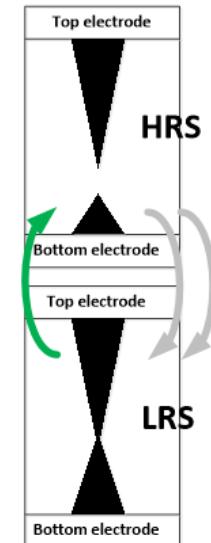
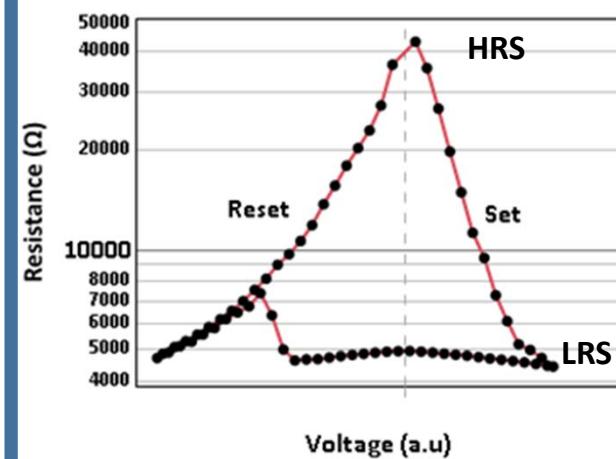
- Embedded memory landscape
- MRAM and RRAM cells and integration into logic technology
- MRAM physics and technology development details
- RRAM physics and endurance challenge
- Embedded non-volatile memory: MRAM vs RRAM
  - Package reflow capability
  - Endurance
  - Retention
- MRAM: beyond nonvolatile memory applications

# Typical Device Characteristics: R-V response

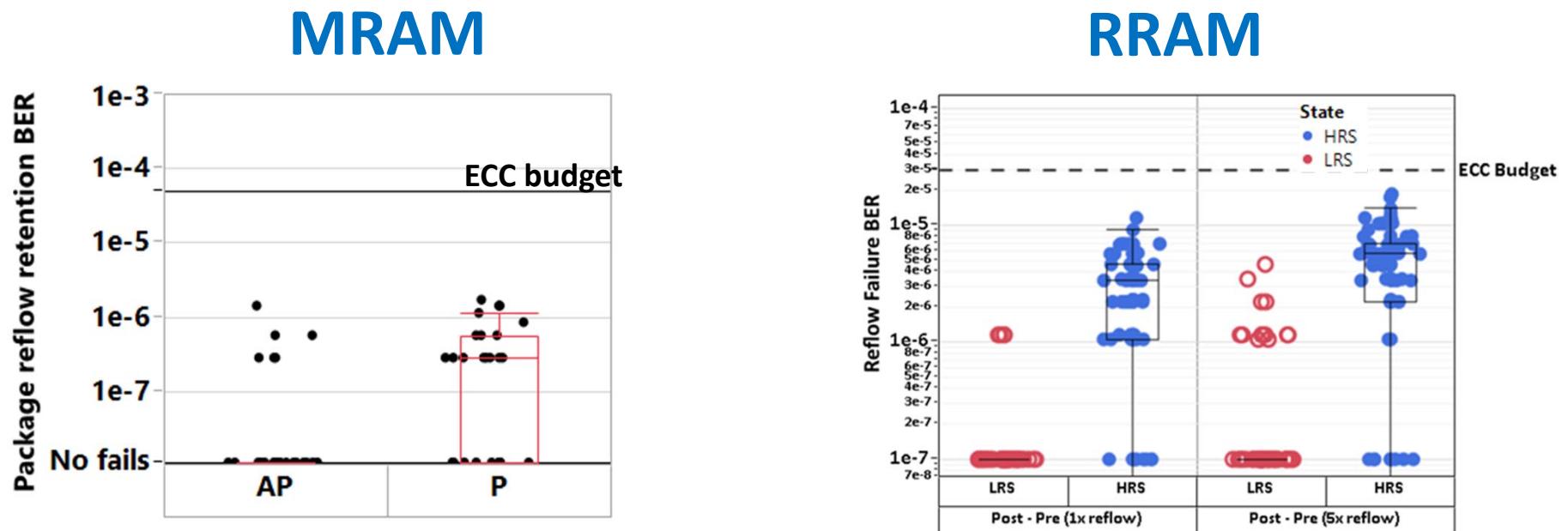
## MRAM



## RRAM



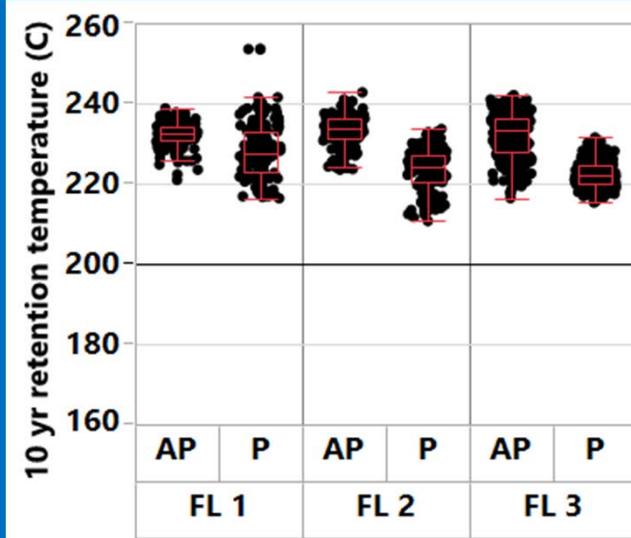
# Package Reflow Capability



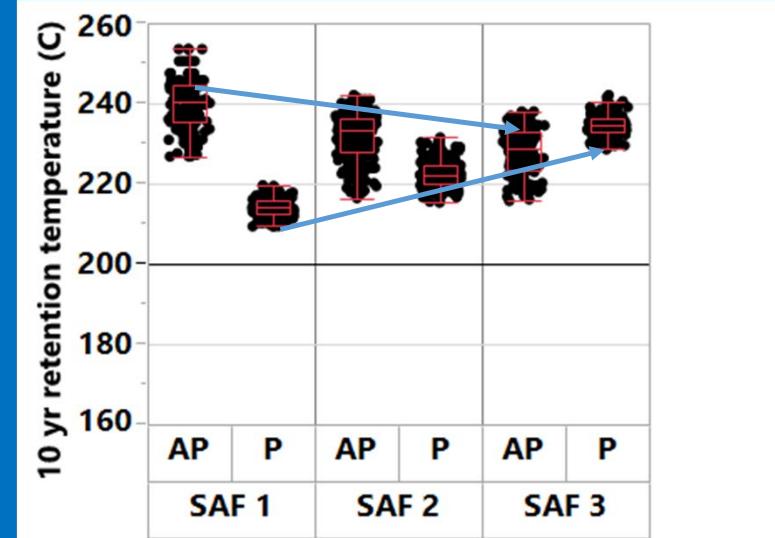
Write / Read / Send through package reflow process / Read

# MRAM: Retention, 10 years at >210C

Temperature at which error rate will reach 1E-5 after 10 years

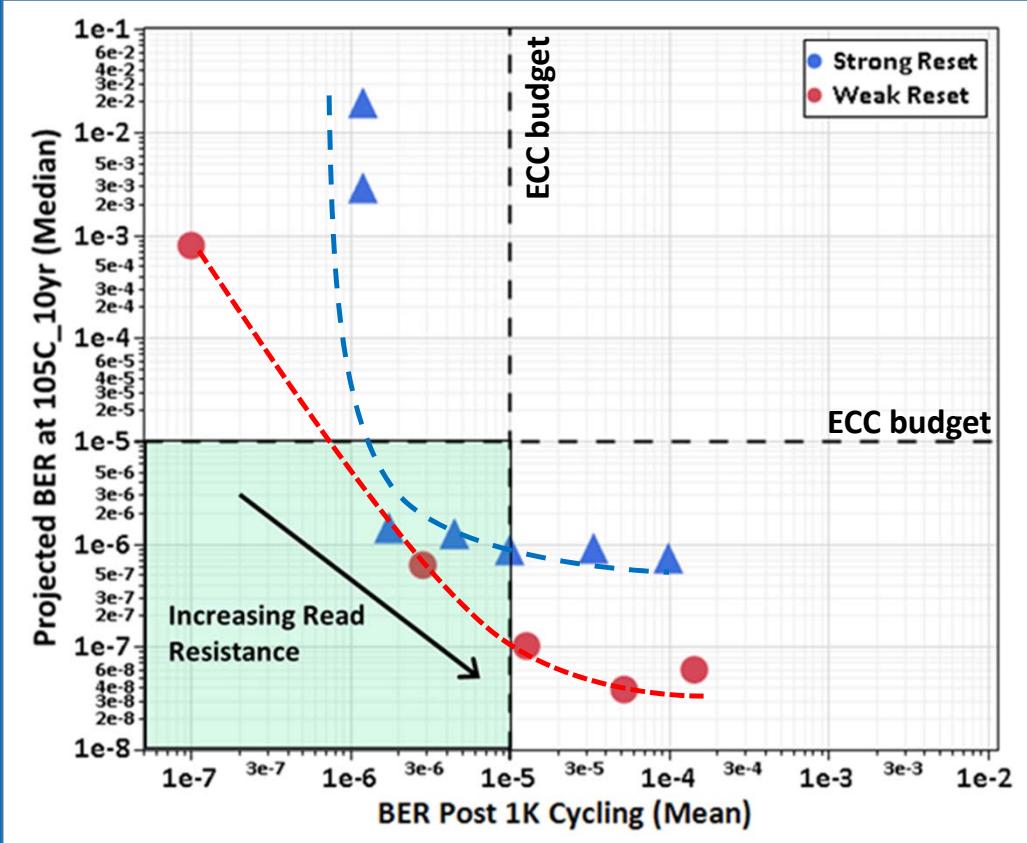


Retention for a Free Layer thickness skew



Retention for a SAF skew

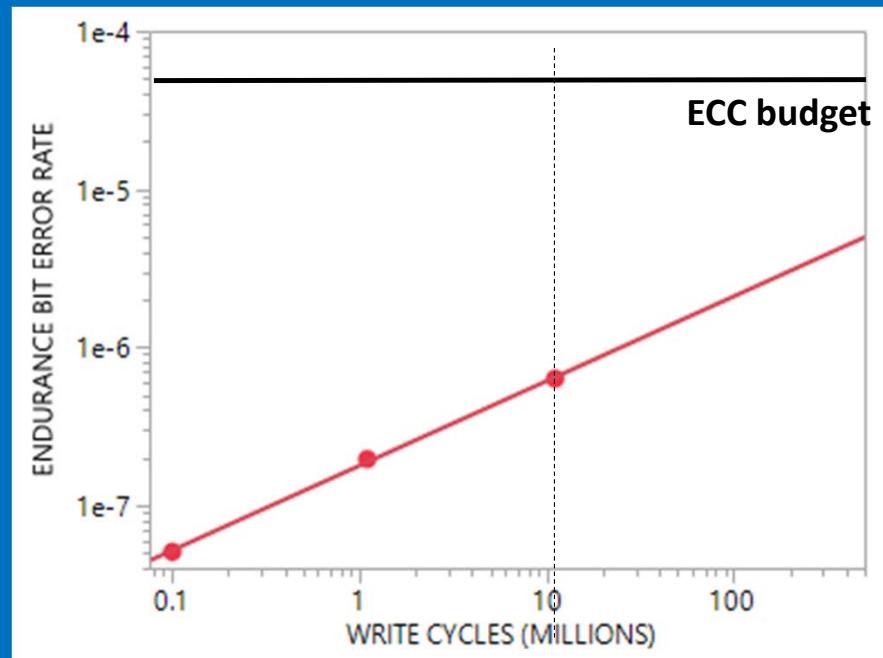
# RRAM array after 1K cycling: Endurance vs Retention



Data from  
High Retention Array

# NVM-MRAM: Endurance well exceeding $10^7$ cycles

Wafer median bit error rate vs number of cycles

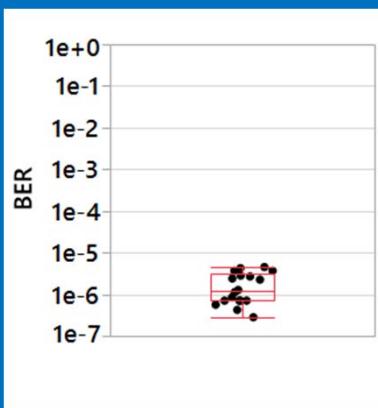


Write / Read / Cycle  $10^5$  times / Read / Cycle  $10^6$  times / Read / Cycle  $10^7$  times / Read

**MRAM: High magnetic field can erase the data,  
but memory remains fully functional**

## Data erase experiment

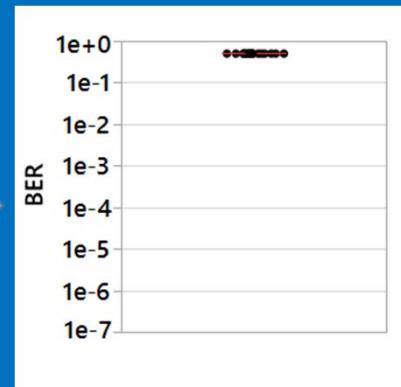
Write + Read,  
Low bit error rate



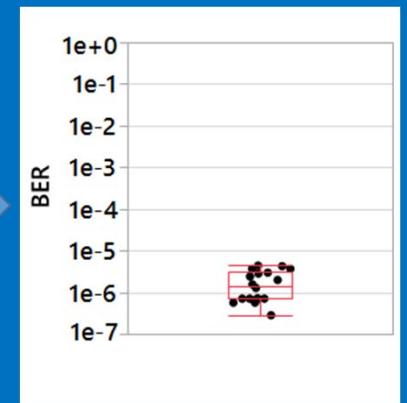
Erase with 1 Tesla  
magnet



Read,  
Data has been erased



Write + Read,  
Low bit error rate



# e-NVM: Key array performance characteristics

## MRAM

- Solder reflow capability
- 125°C operating temperature
- >210°C 10-year retention
- $>10^7$  cycle endurance

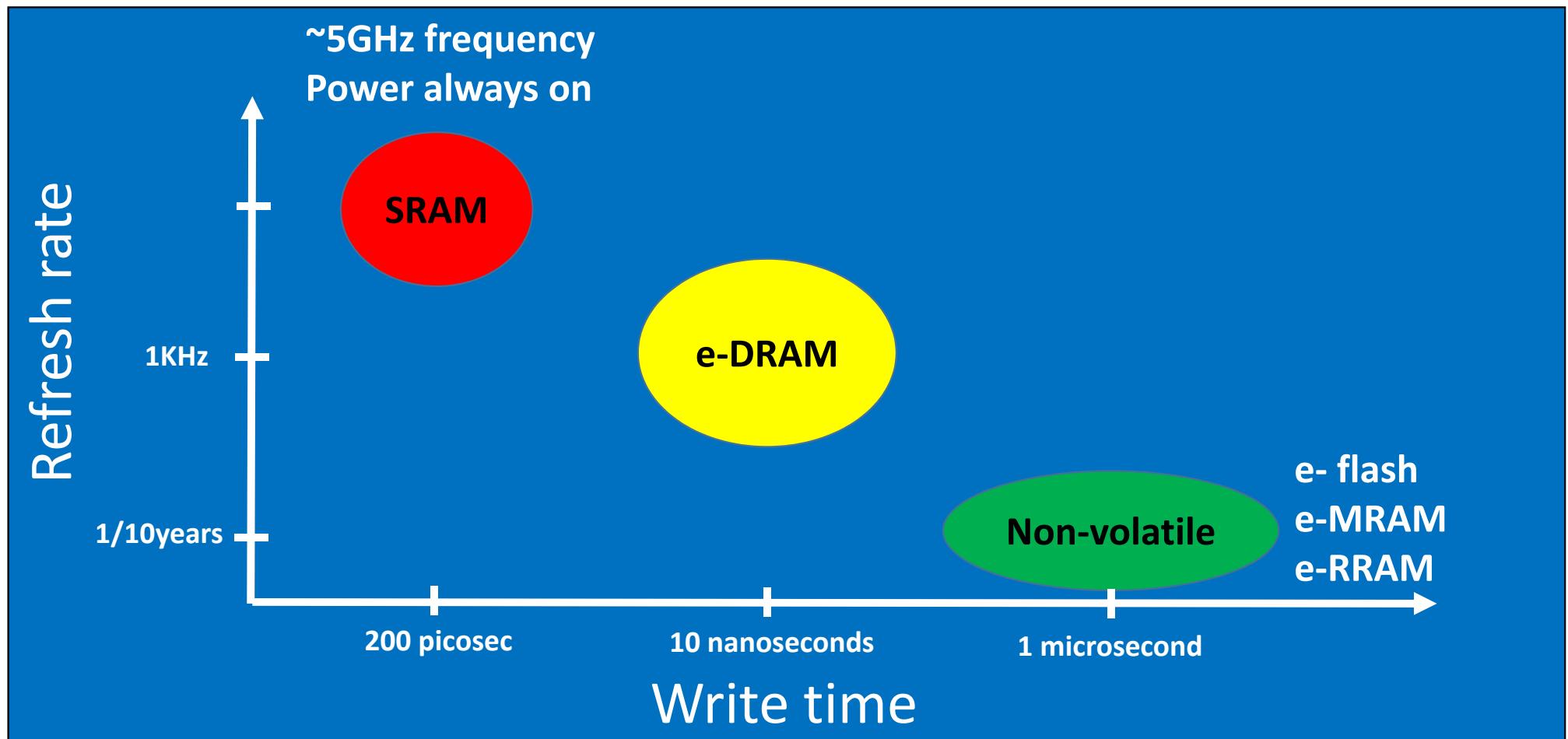
## RRAM

- Solder reflow capability
- 105°C operating temperature
- >105°C 10-year retention
- $>10^3$  cycle endurance
- Magnetic attack immunity

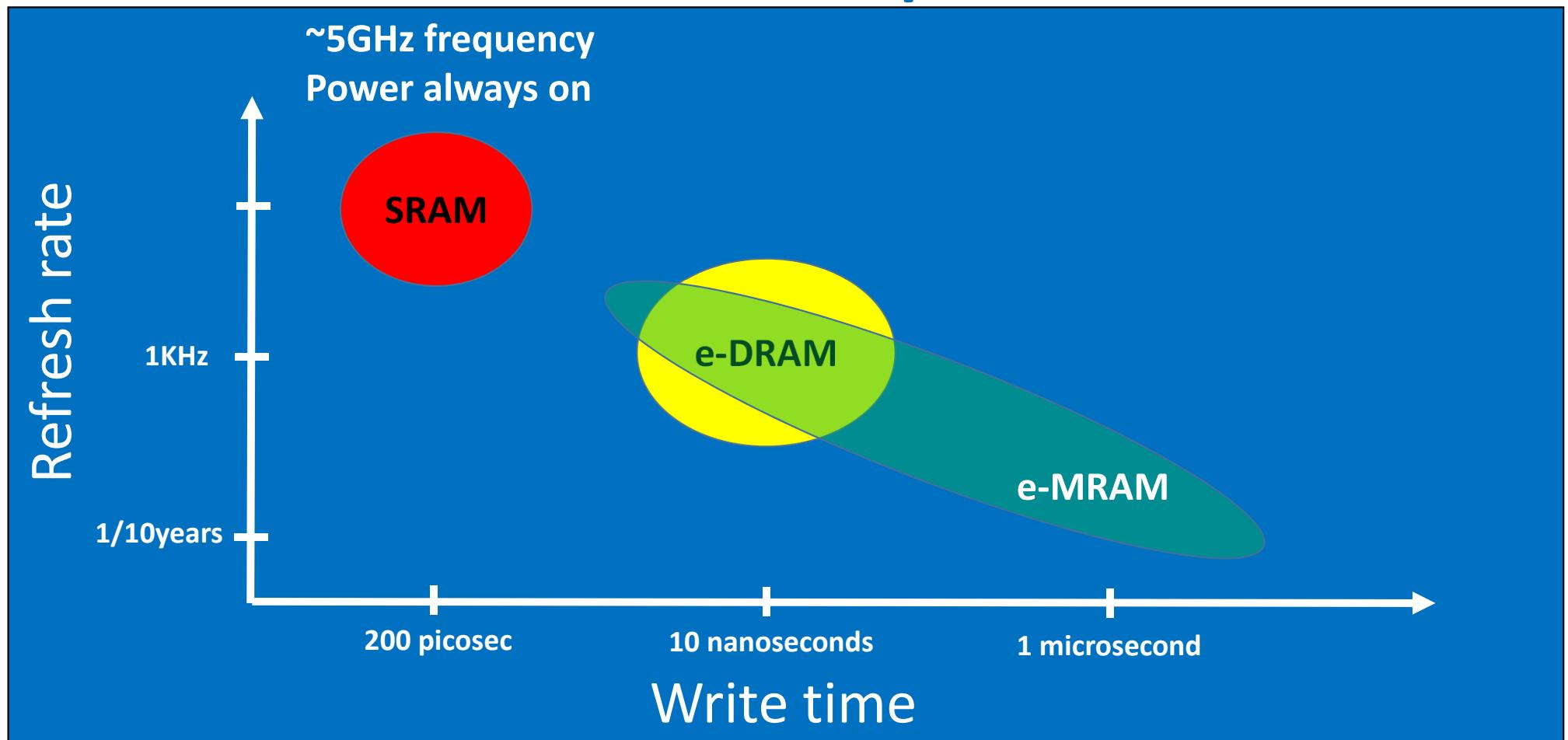
# Outline of Presentation

- Embedded memory landscape
- MRAM and RRAM cells and integration into logic technology
- MRAM physics and technology development details
- RRAM physics and endurance challenge
- Embedded non-volatile memory: MRAM vs RRAM
- **MRAM: beyond nonvolatile memory applications**

# Embedded Memory landscape



# STT-MRAM on Embedded Memory landscape

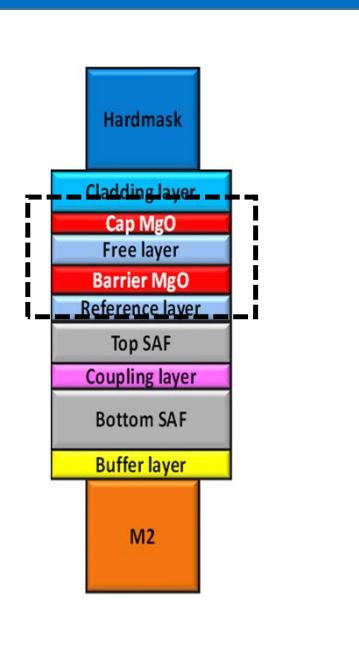
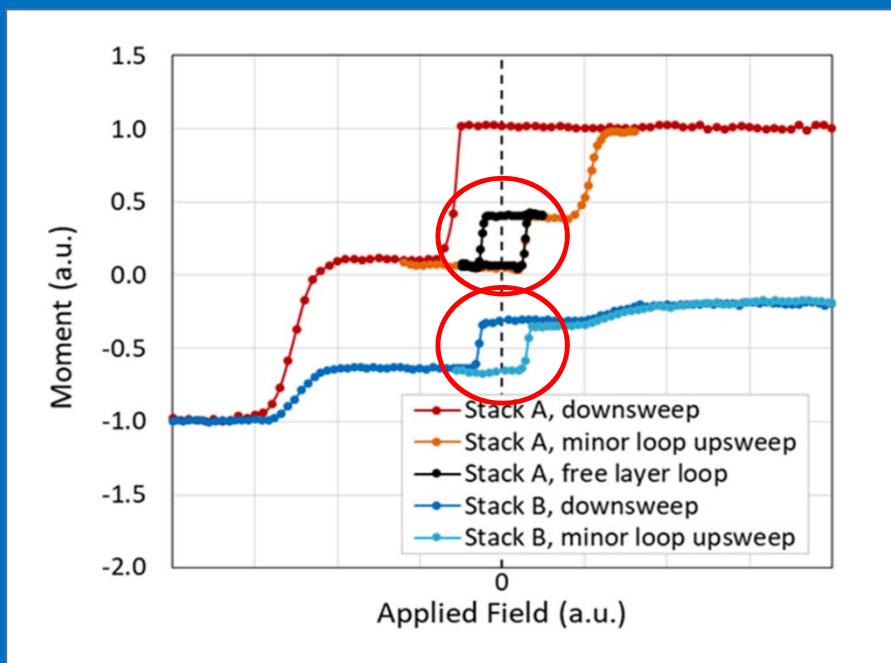


# **Extending MRAM capabilities into e-DRAM space**

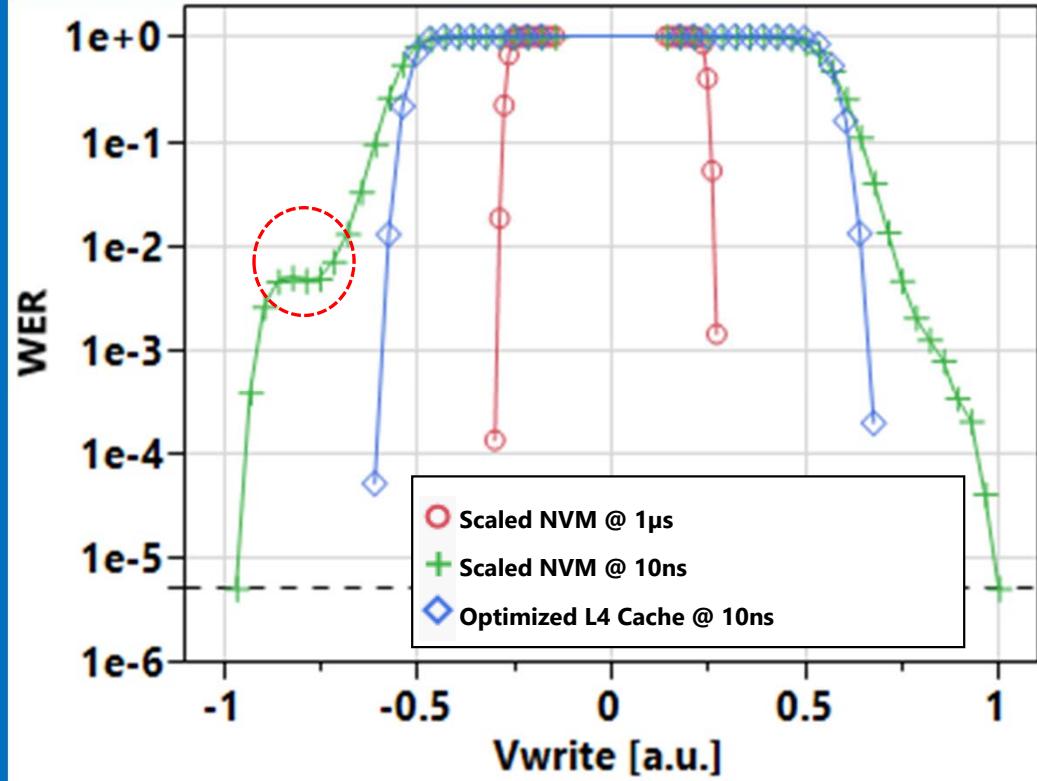
- 1) What changes are needed to reduce MRAM bit write time?
- 2) Can we write MRAM bits with nanosecond pulses?
- 3) Is MRAM capable of endurances needed for L4 cache applications?
- 4) Is MRAM cell competitive from density perspective?

# MTJ: what determines the switching energy?

M-H curves of patterned MTJ arrays

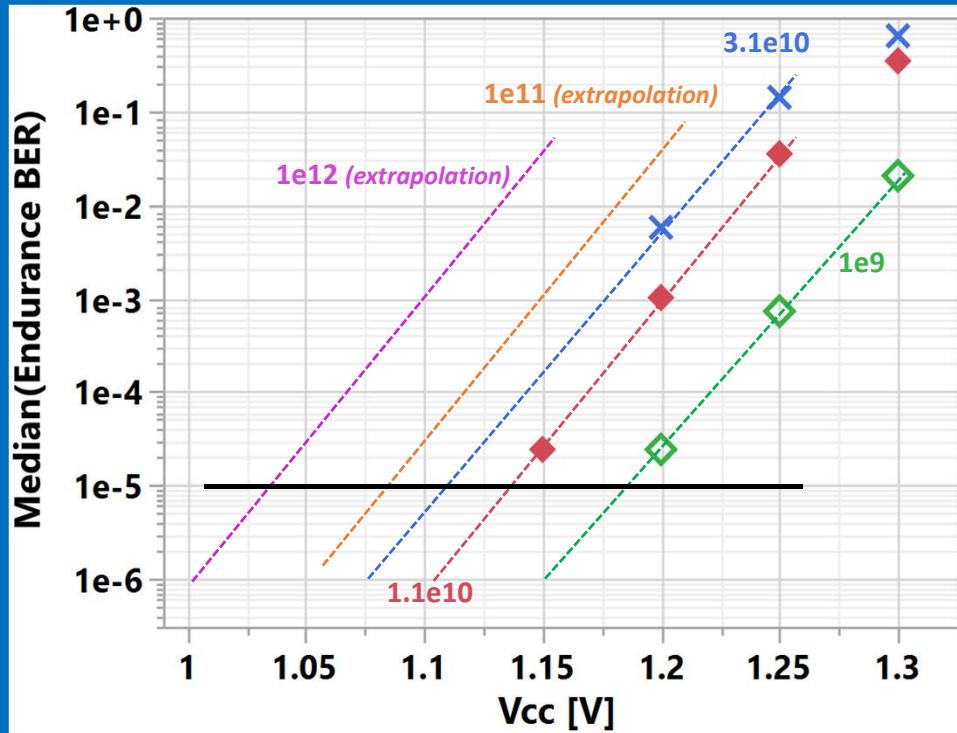


# MRAM n-sec switching, write error rate



J.G. Alzate *et al.*, to be published, 2019 IEEE International Electron Devices Meeting (IEDM)

# MRAM 10nsec switching endurance, array data

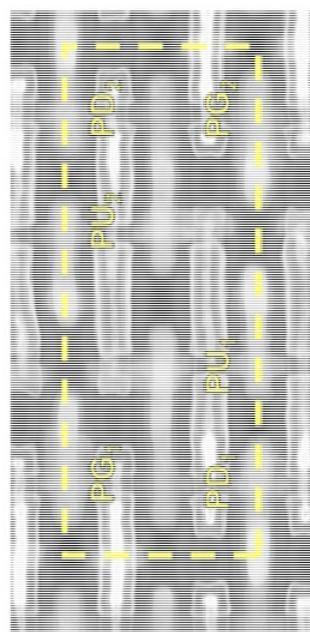


<1E-5 error rate demonstrated for 1E12 cycles at VCC of 1V

J.G. Alzate et al., to be published, 2019 IEEE International Electron Devices Meeting (IEDM)

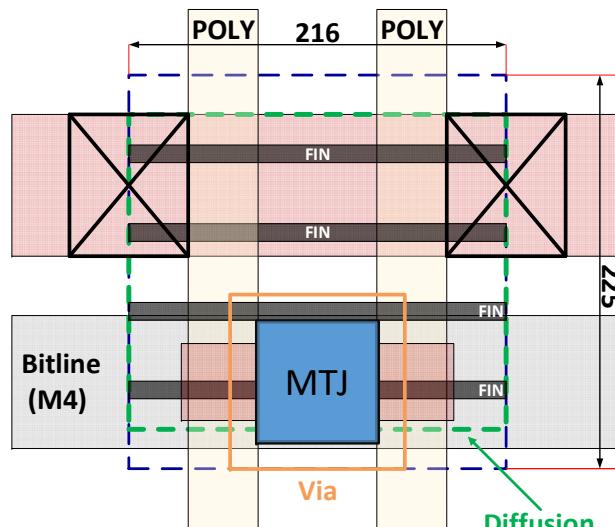
# SRAM vs MRAM cell comparison

Intel 10nm FinFET technology  
SRAM, 6 transistors



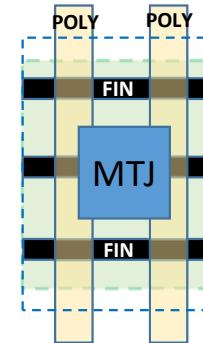
$0.0312 \mu\text{m}^2$

Intel 22FFL FinFET technology  
e-NVM MRAM, 1T-1R



$0.0486 \mu\text{m}^2$

Hypothetical MRAM cell  
for L4 cache applications  
2 Poly, 3 fins, 1T-1R



$<0.015 \mu\text{m}^2$

# Summary

- MRAM and RRAM based embedded nonvolatile memories are here, high yielding and integrated into FinFET logic
- RRAM possesses immunity to magnetic field attack and is the right choice for a variety of product form factors
- MRAM possesses excellent endurance characteristics and easy process tunability between nonvolatile and fast switching regimes
- Nano-second switching, high endurance, ease of integration into CMOS and small foot print make MRAM very competitive in the L4 cache replacement space



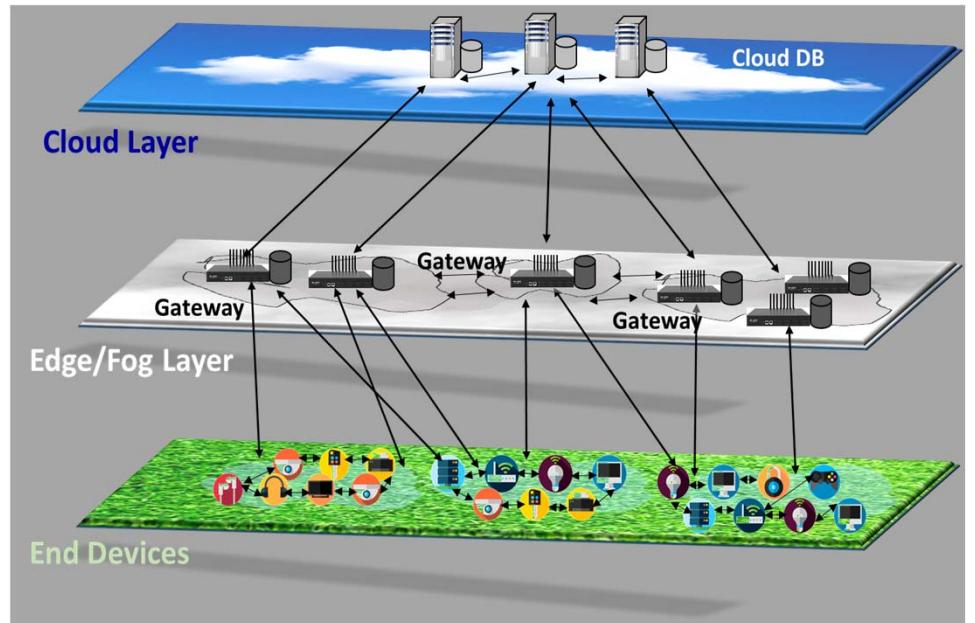
# **Emerging Technologies for Memory-centric and Low-power Architectures**

Edith Beigne, Silicon Research

**facebook**

# Low Power circuits challenges

- Embedded circuits and systems: power challenges
  - Autonomy
  - Thermal issues
- Many variations limiting circuits performances
- Performances optimizations for **high energy efficiency**



[B. De Salvo, Plenary talk , ISSCC 2018]

# AR/VR challenges



SENSING &  
RECONSTRUCTING  
REALITY



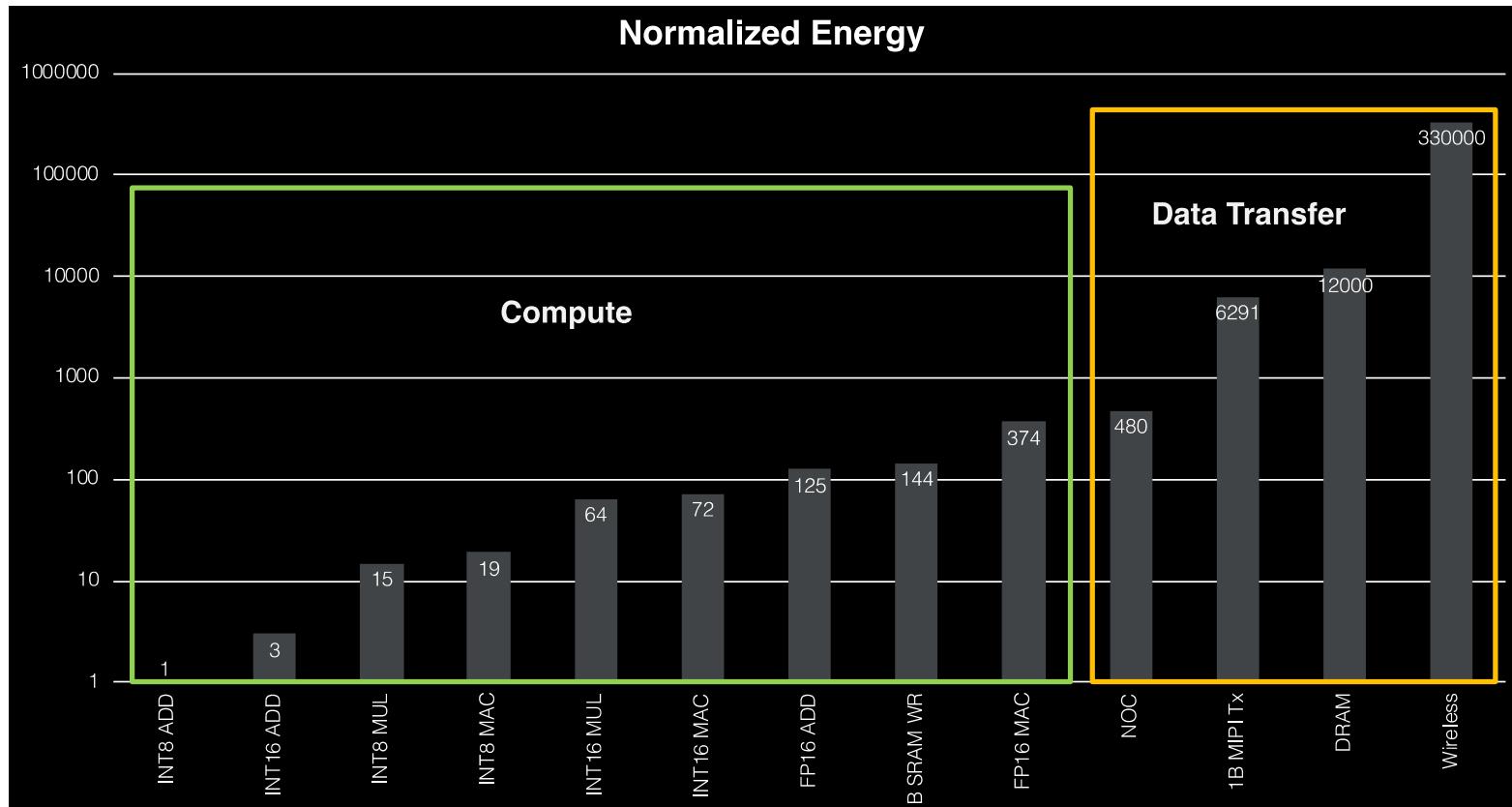
DRIVING  
THE PERCEPTUAL  
SYSTEM



INTERACTION

[S. Rabii, Plenary talk , VLSI 2019]

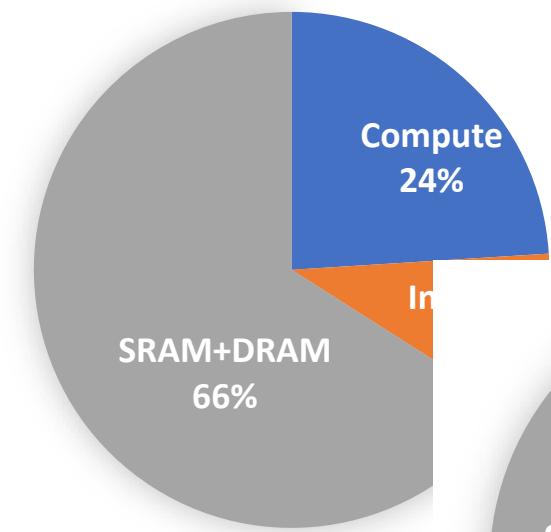
# Cost of Memories and Data Movement



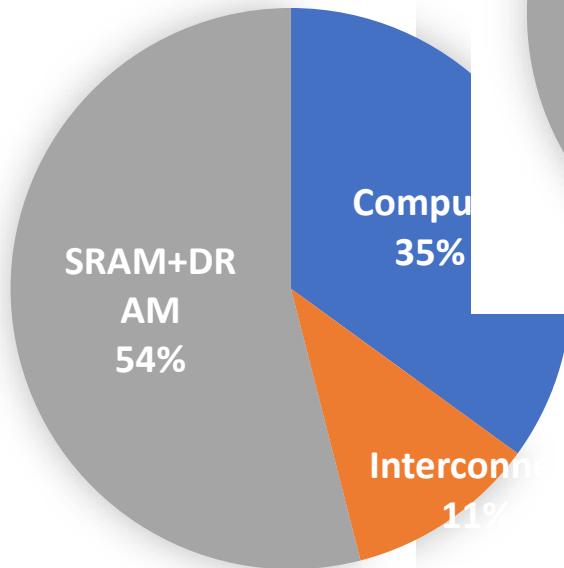
[S. Rabii, Plenary talk , VLSI 2019]

# Cost of Memories and Data Movement

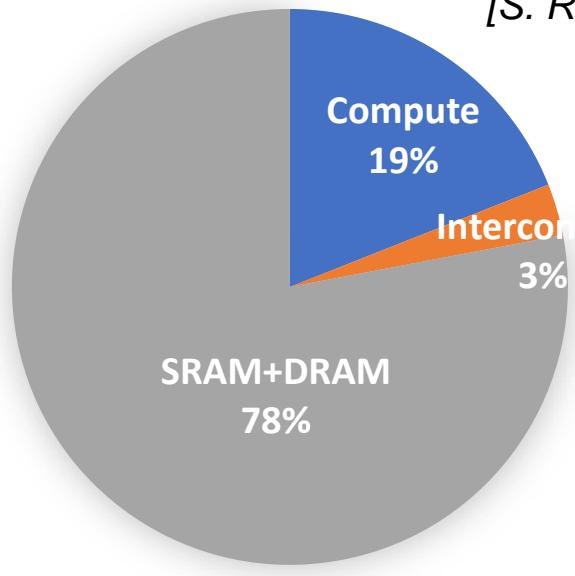
[S. Rabii, Plenary talk , VLSI 2019]



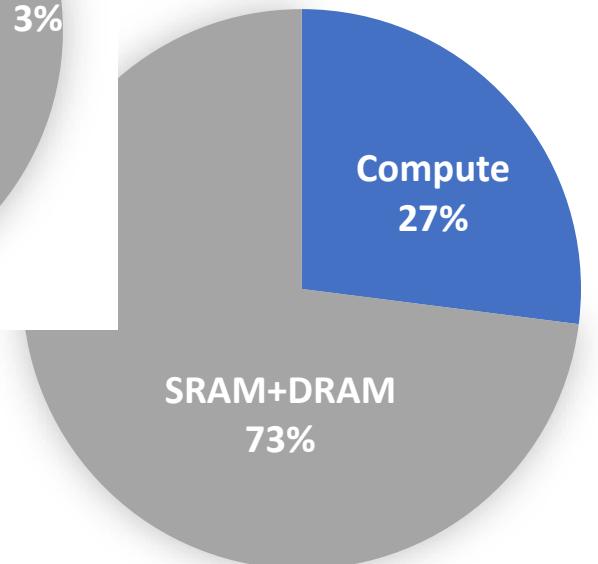
Audio



Hand Tracking

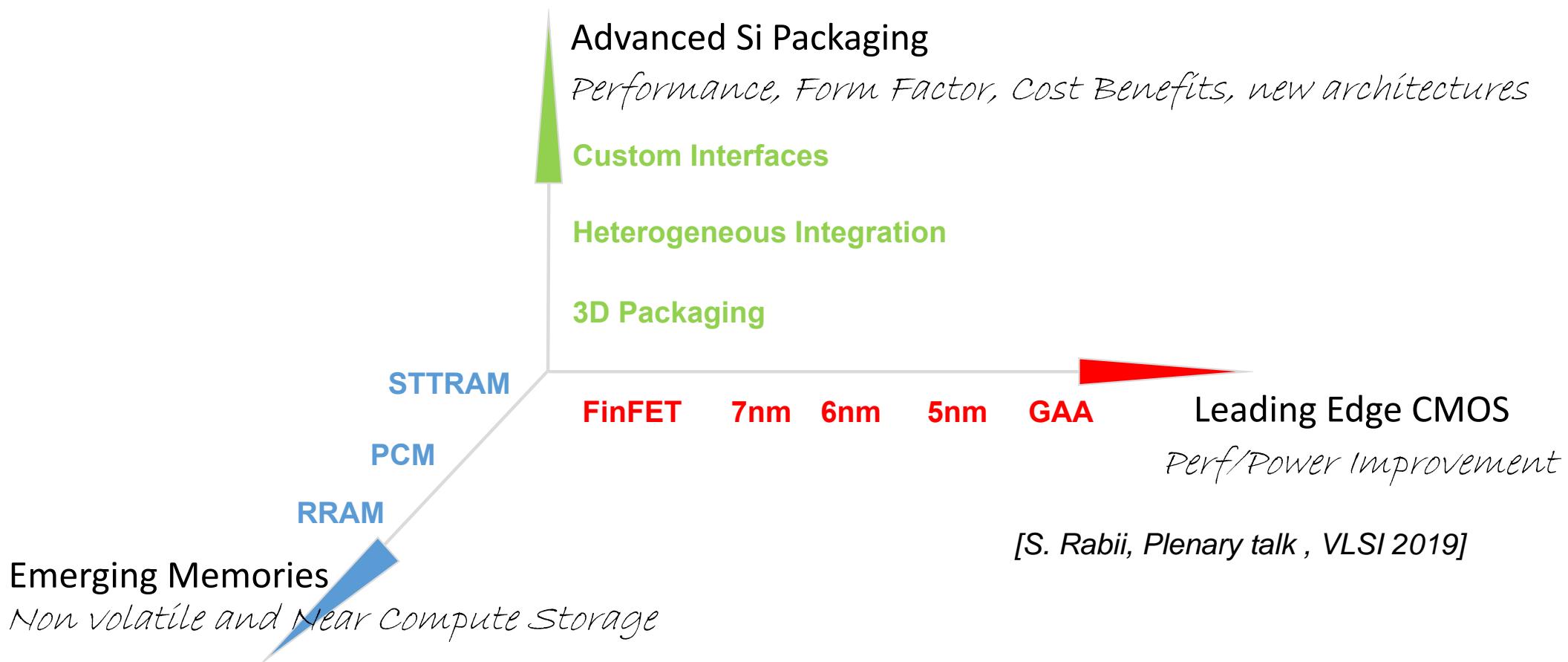


SLAM



Depth

# Technology options for Memory-centric and Low-power Architectures

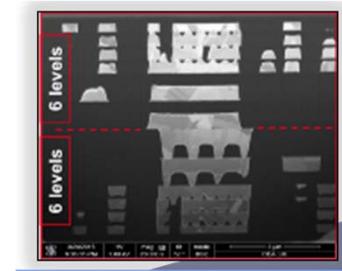
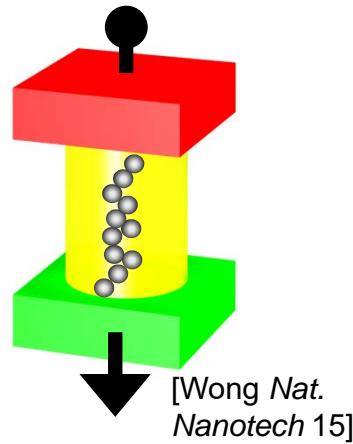
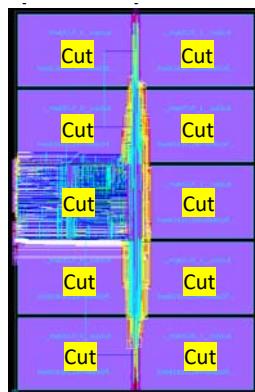


# Memory power trade-offs – new architectures, design and technologies

**SRAM**  
*Power reduction  
and Error  
resilience*

**Non-volatile  
memories**  
*RRAM Application  
to AI*

**3D integration**  
*New architectures,  
TSVs and Hybrid  
bonding*



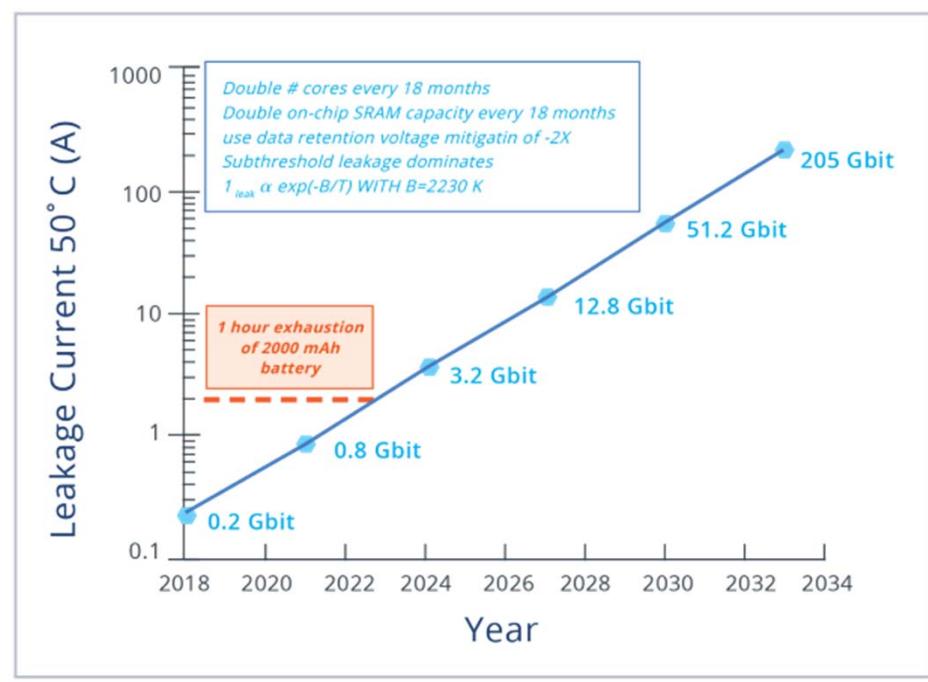
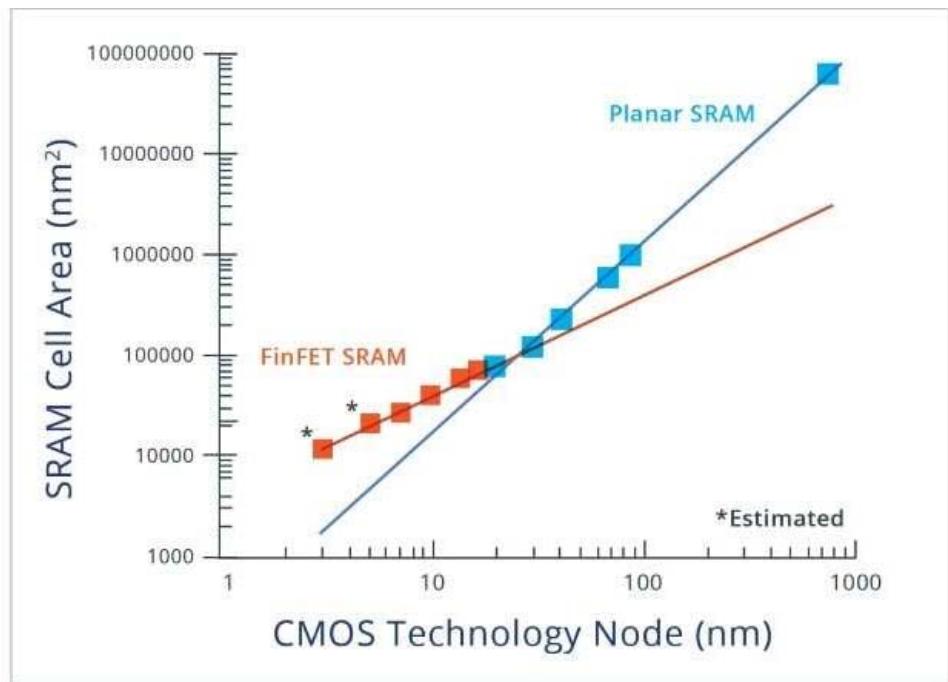
[Vivet, DATE 2019]

# Memory power trade-offs – new architectures, design and technologies

SRAM

*Power reduction  
and Error  
resilience*

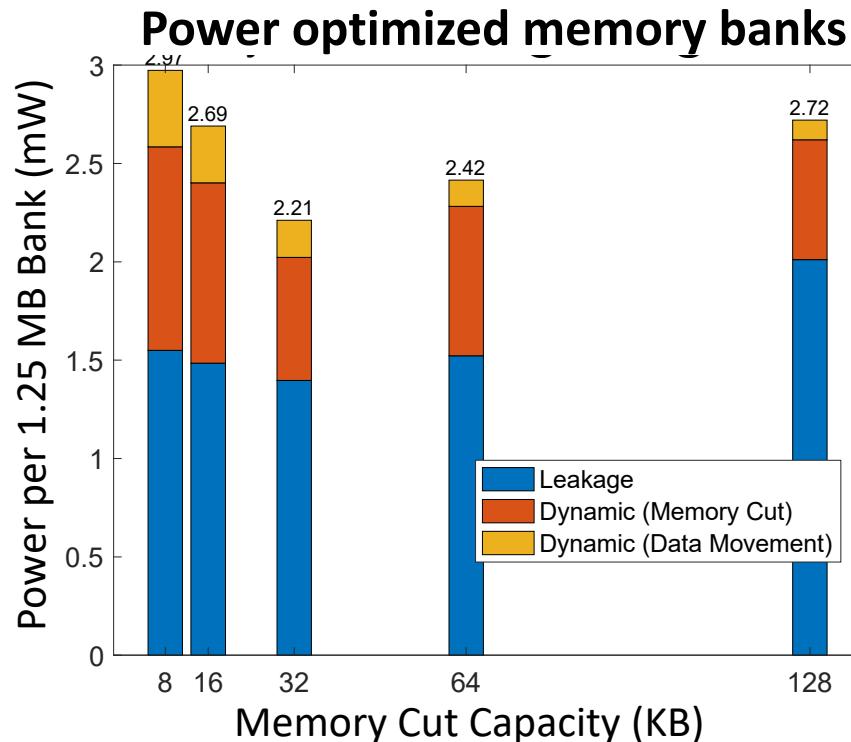
# SRAM issues in modern applications



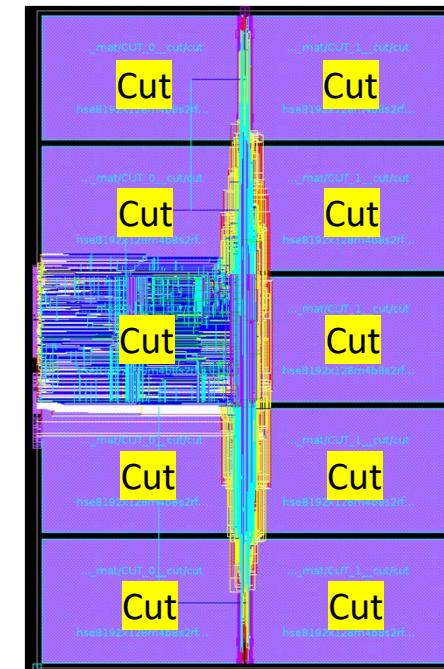
[A. Walker, *The Trouble with SRAM*, EETimes 2018]

# SRAM Memory benchmark: large capacity

- SRAM design exploration using compiler generated memory cuts



**1 of 8 (1.25 MB) SRAM Banks**

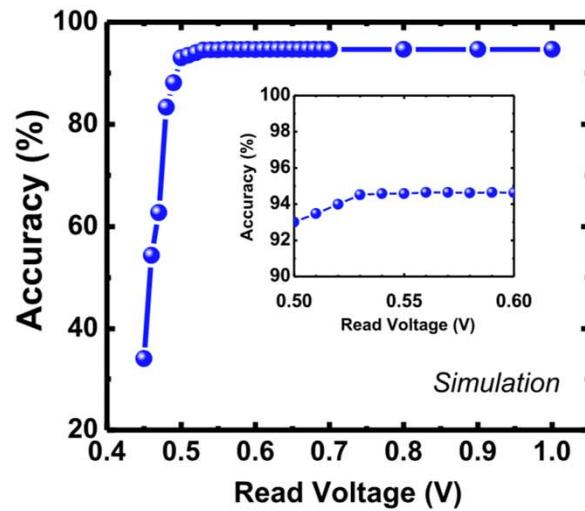


\*500 MHz clock, 0.8V, Compiler generated memory cuts 14nm technology

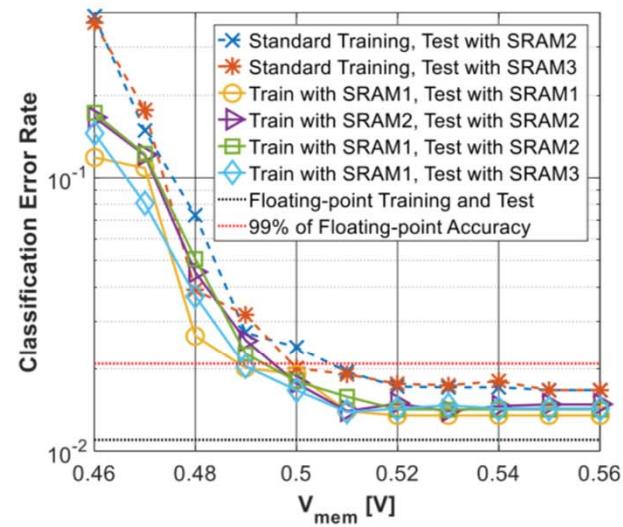
# **SRAM integration trade-offs**

- Activate efficiently SRAM power states
  - Embedded control, data-driven
- Lower voltage supply to its minimum
  - Write assist techniques
- Explore error resilience
  - Allows aggressive voltage scaling – energy savings
  - Live with higher BER (probability to flip 0 <->1)

# Neural Networks are resilient to errors



X. Sun et al., "Low-VDD Operation of SRAM Synaptic Array for Implementing Ternary Neural Network," in IEEE Transactions on VLSI Systems, vol. 25, no. 10, pp. 2962-2965, Oct. 2017



L. Yang and B. Murmann, "SRAM voltage scaling for energy-efficient convolutional neural networks," 2017 18th International Symposium on Quality Electronic Design (ISQED), Santa Clara, CA, 2017, pp. 7-12.

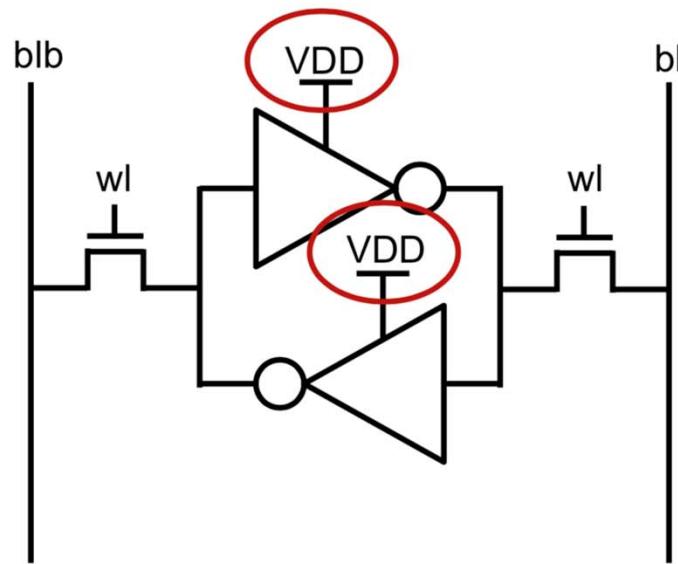
# **SRAM Failures mechanisms**

- SRAM cell failures
- Sense Amplifier failures
- Address decoder failures

# SRAM Failures mechanisms

- Bitcell current decrease in discharging the bitcell through the access transistor during read operation (*access failure*)
- Increased disturbance to the cell content during read which flips the cell value (*flipping read failure*)
- Unsuccessful write due to the deviation of the strength of the access transistor and cross coupled inverter (*write failure*)
- Instability of a cell in holding its content due to excessive mismatch (especially when the supply voltage is lowered) (*hold failure*)

# Power reduction & BER



## Supply Voltage Reduction

**Dynamic Power**



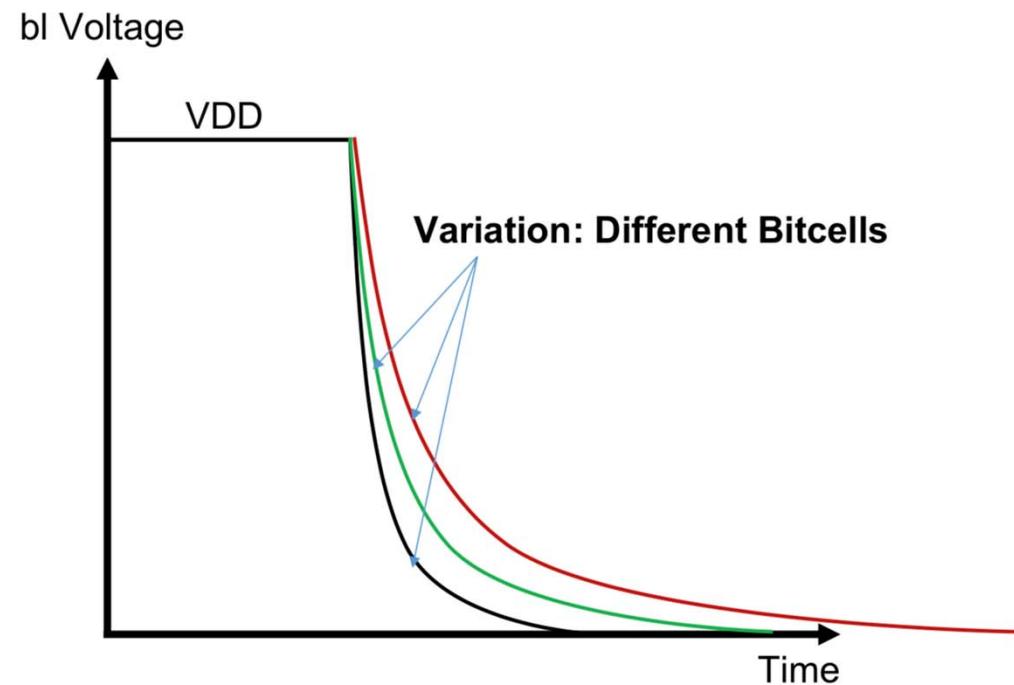
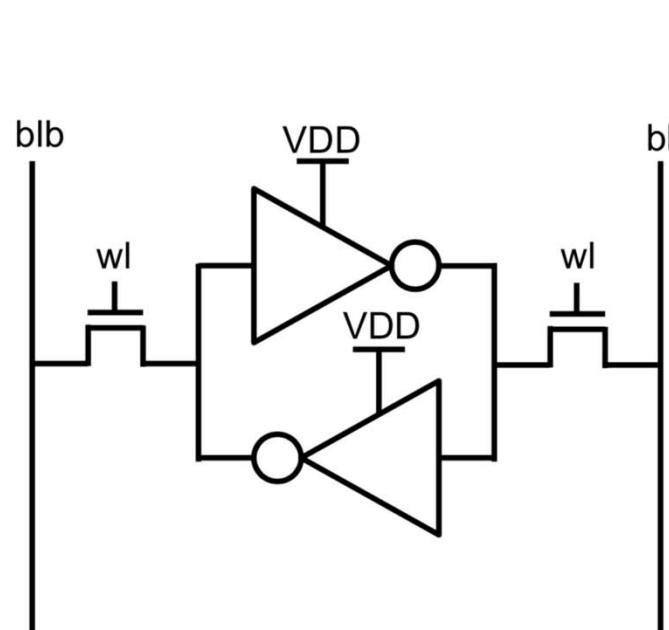
**Static Power**



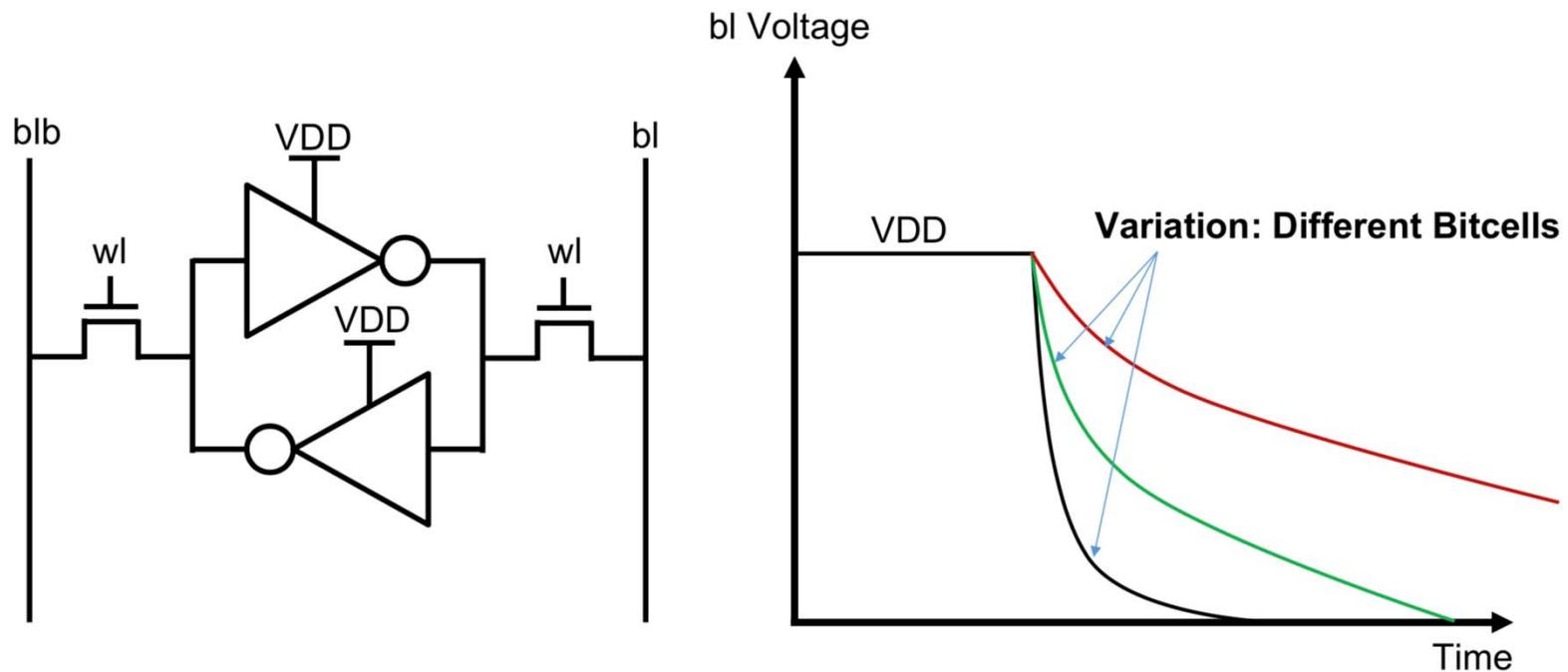
**Error Rate**



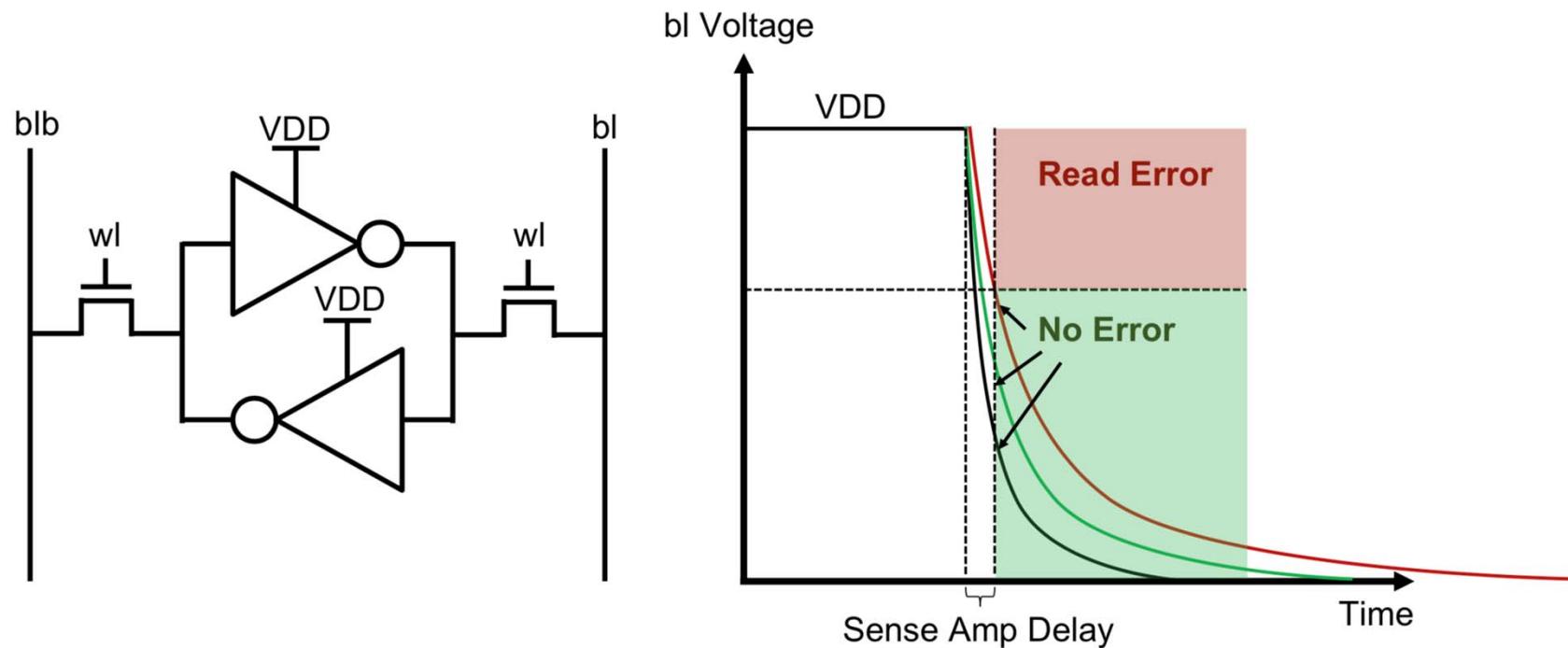
# Impact of variations at low voltages



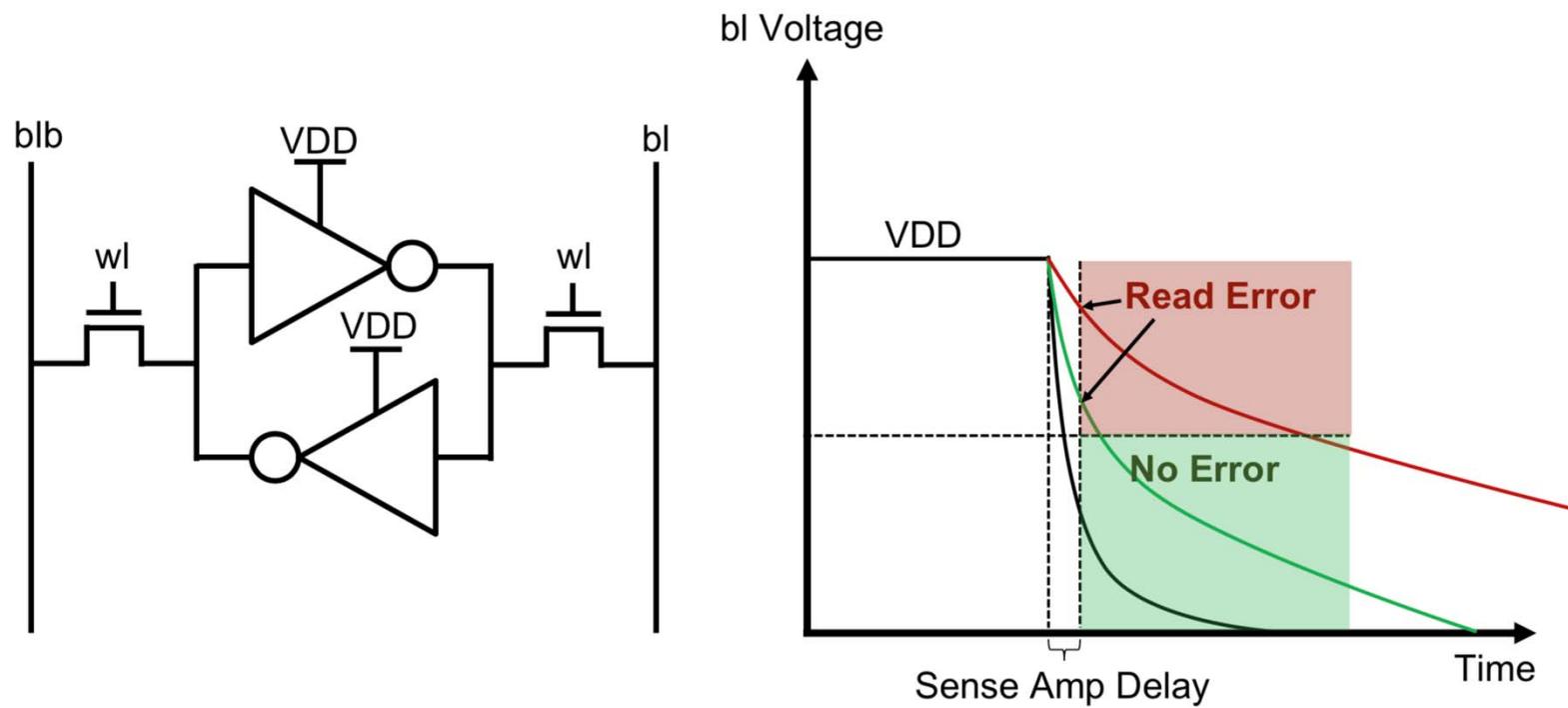
# Impact of variations at low voltages



# Impact of variations at low voltages – Error rate

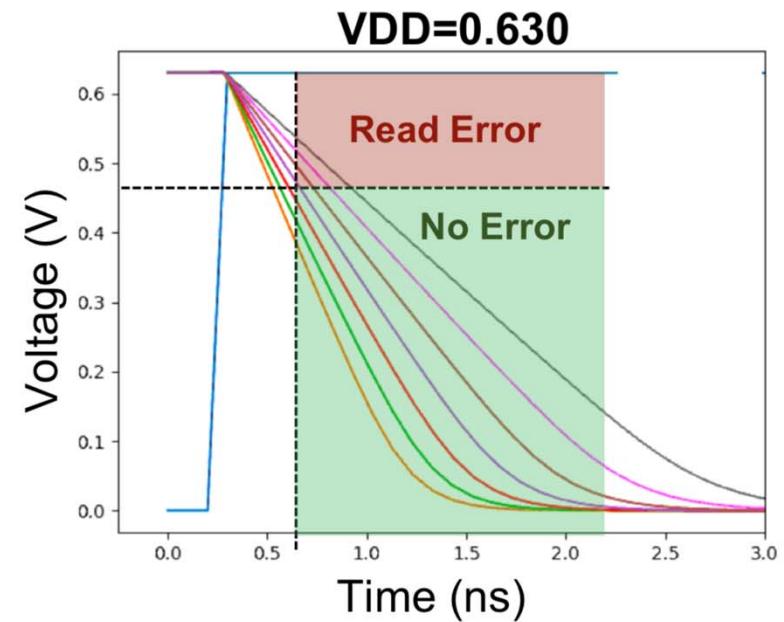
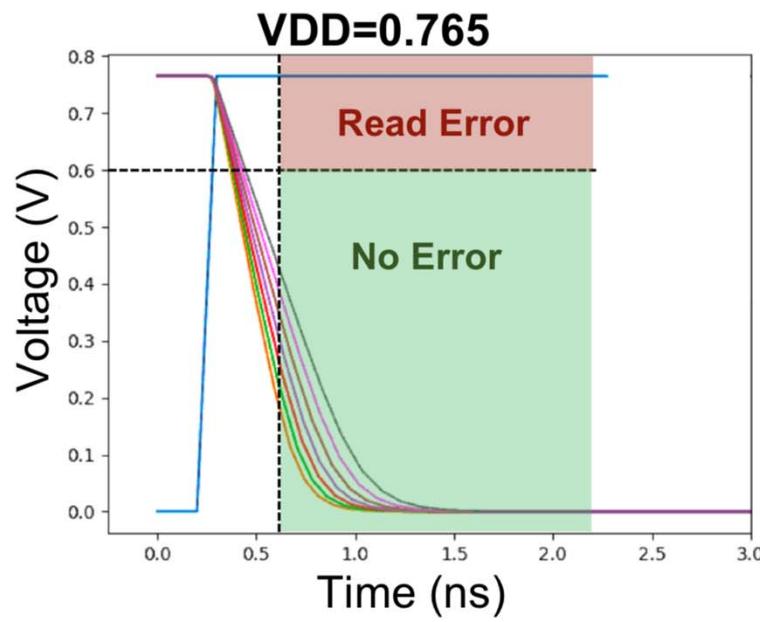


# Impact of variations at low voltages – Error rate

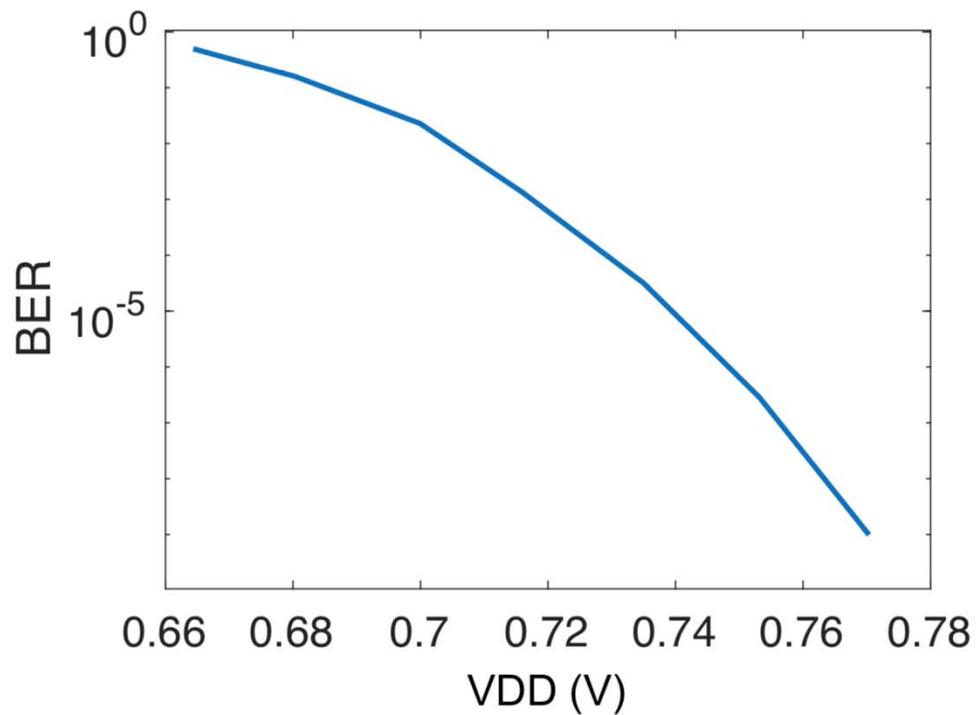


# Extracted BER from simulation (PVT & Vdd)

Extract BER for each VDD

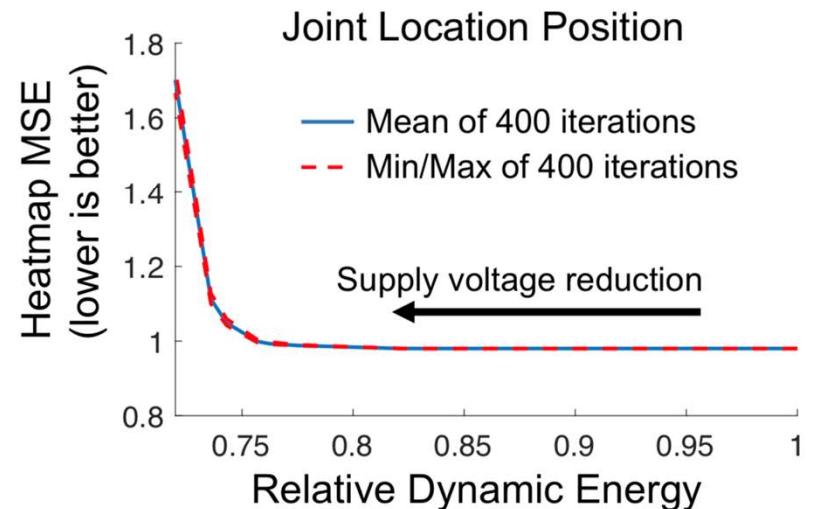
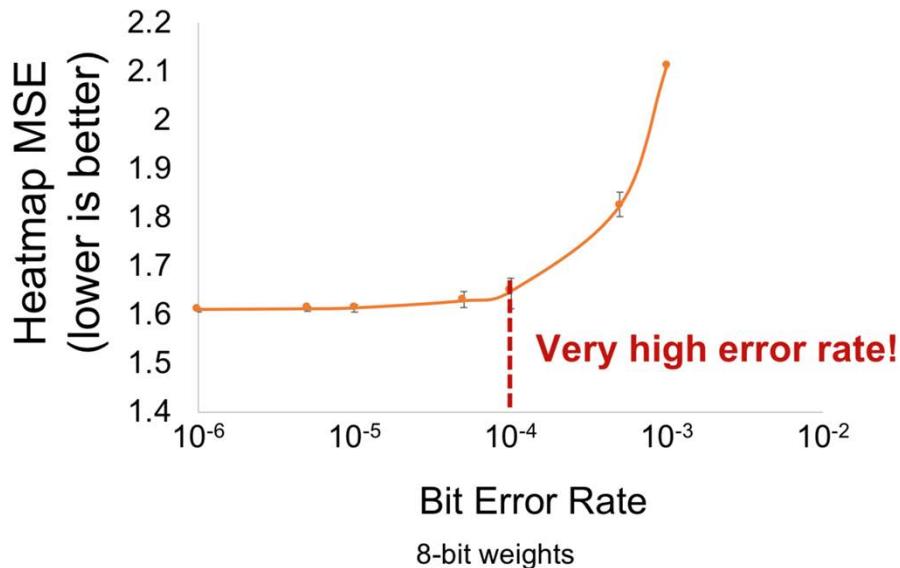


# Extracted BER from simulation (PVT & Vdd)



# Application to NN and power savings

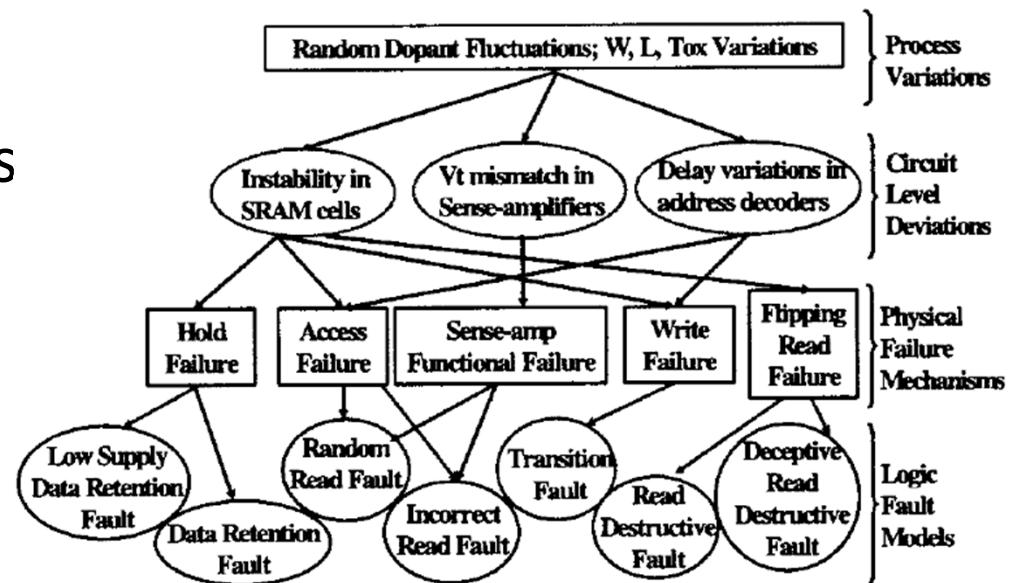
- 25% Dynamic Energy Savings w/o loss in accuracy



[S. Rabii, Plenary talk , VLSI 2019]

# SRAM Failures mechanisms

- SRAM cell failures
- Sense Amplifier failures
- Address decoder failures



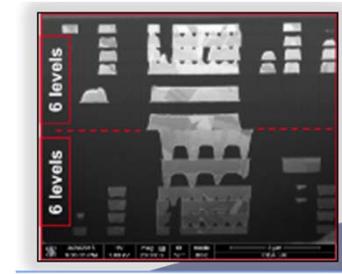
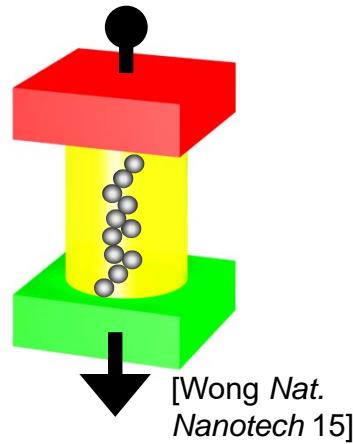
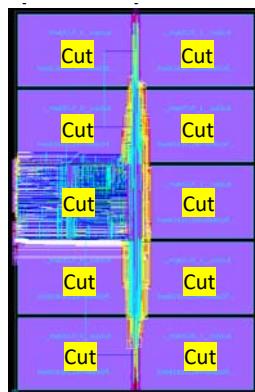
[Q.Chen, VTS 2005]

# Memory power trade-offs – new architectures, design and technologies

**SRAM**  
*Power reduction  
and Error  
resilience*

**Non-volatile  
memories**  
*RRAM Application  
to AI*

**3D integration**  
*New architectures,  
TSVs and Hybrid  
bonding*



[Vivet, DATE 2019]

# Memory power trade-offs – new architectures, design and technologies

Non-volatile  
memories

*RRAM Application  
to AI*



# Edge Applications

## Applications

Machine Learning  
(ML)

Control

Security

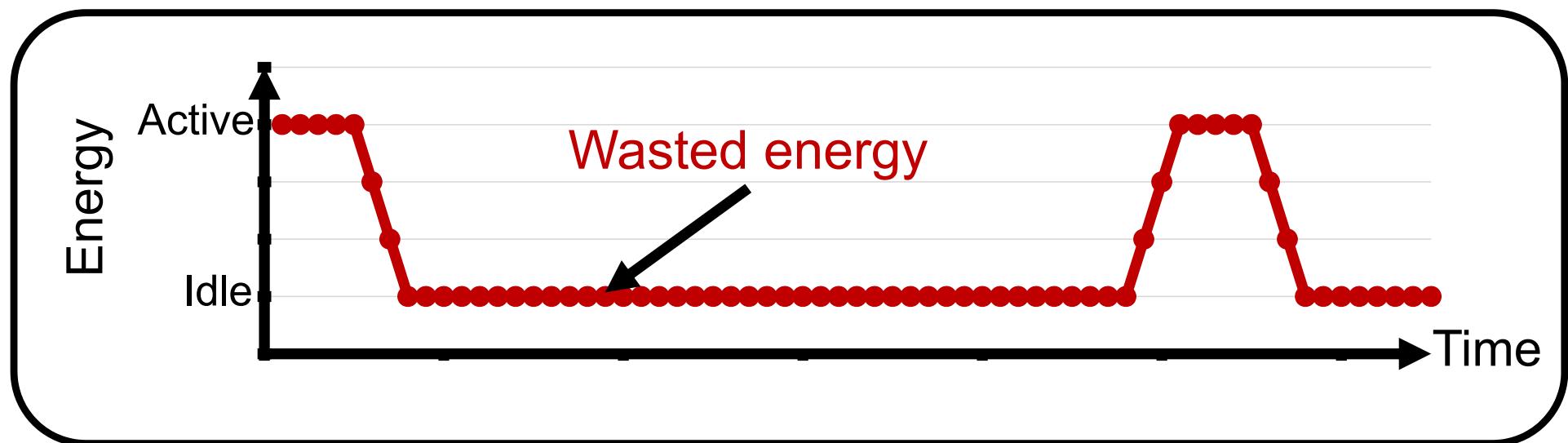
...

## NVM Requirements

- Integrated on-chip
- Low power
- Low latency
- High density
- High endurance

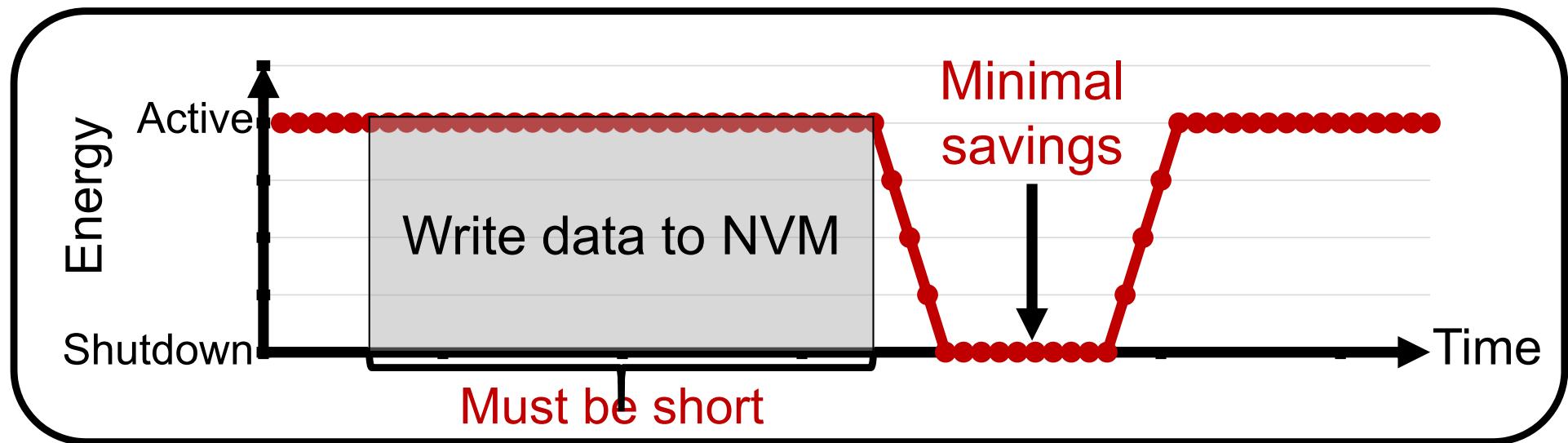
[T. Wu, ISSCC 2019]

# NVM Energy Saving Opportunity



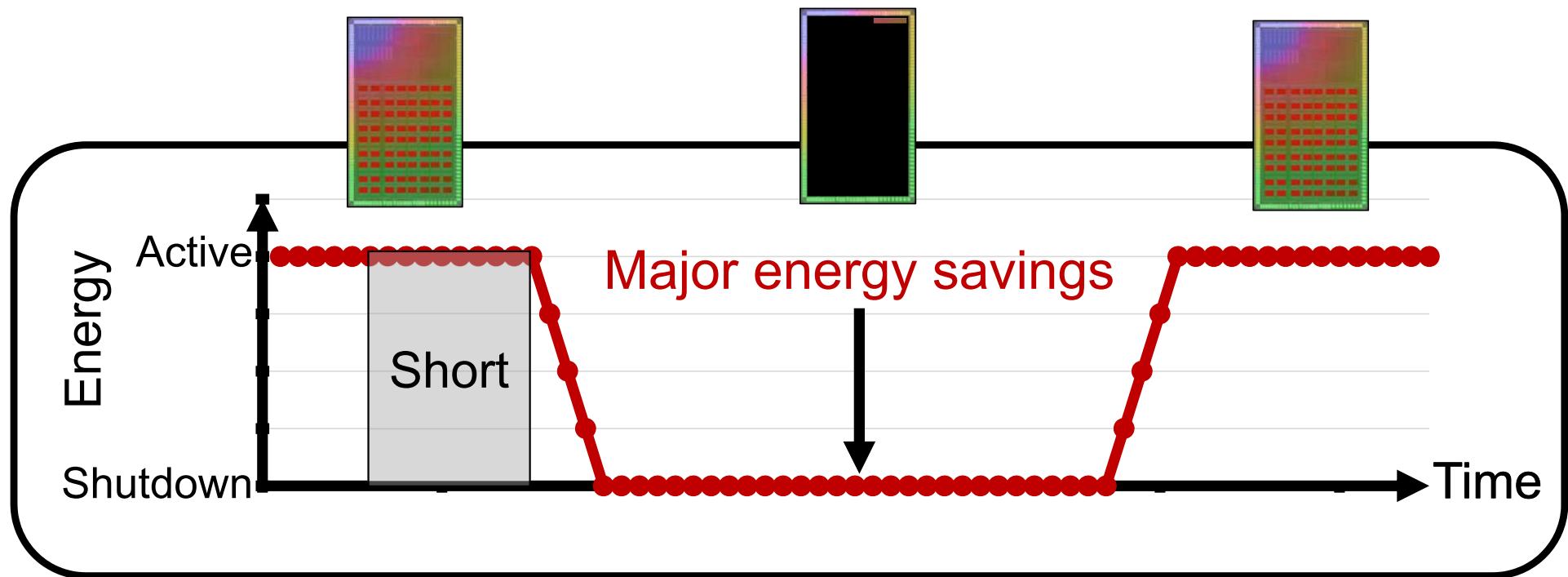
[T. Wu, ISSCC 2019]

# NVM Energy Saving Opportunity



[T. Wu, ISSCC 2019]

# NVM Energy Saving Opportunity

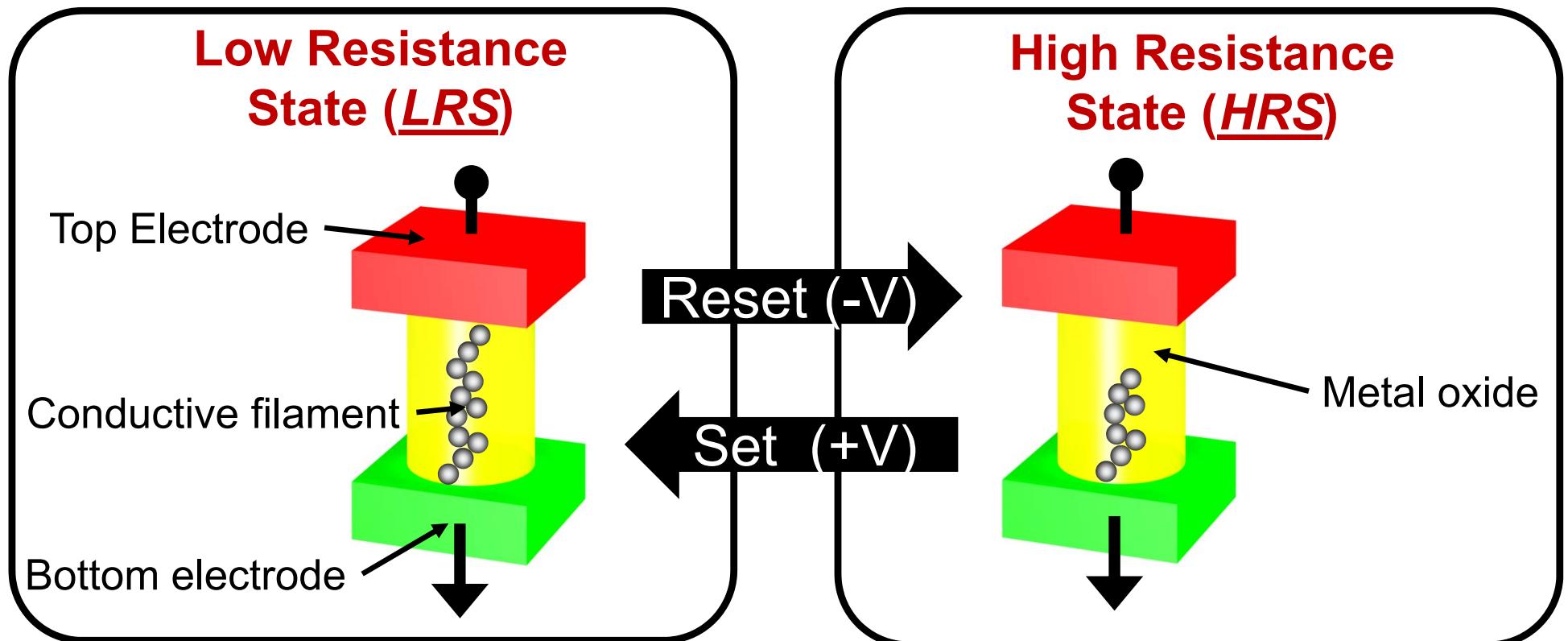


Opportunity: New, Faster NVM

[T. Wu, ISSCC 2019]

# Resistive RAM (OxRAM)

*Voltage controlled resistance*



[Wong Nat. Nanotech 15]

[T. Wu, ISSCC 2019]

# Micro- controller and Endurer: a use-case example

## Data scratchpad

- SRAM buffer for new endurance

## Processor core

- Apples-to-apples comparison

## Memory controller & RRAM repair

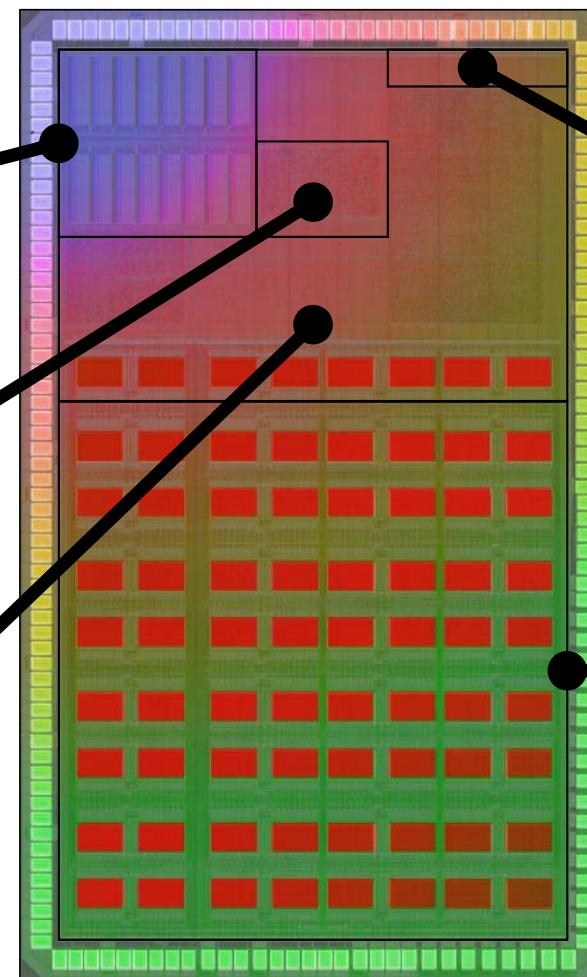
- 100% failed bits recovered

## Scheduler

- 5,000× quicker shutdown vs. Flash

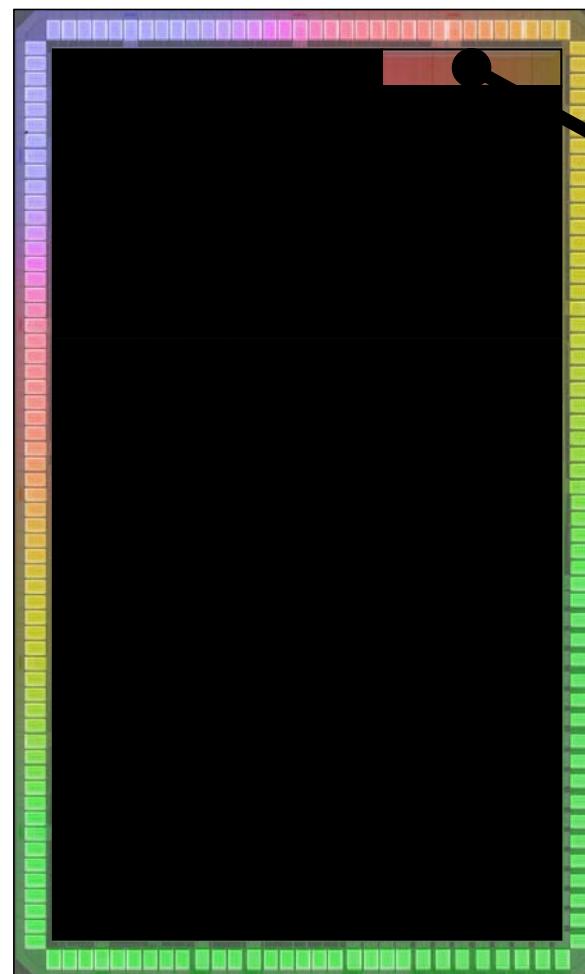
## RRAM on-chip

- 10× lower energy vs. Flash
- Multiple bits/cell



[Wu /SSCC 2019]

# Shutdown Mode



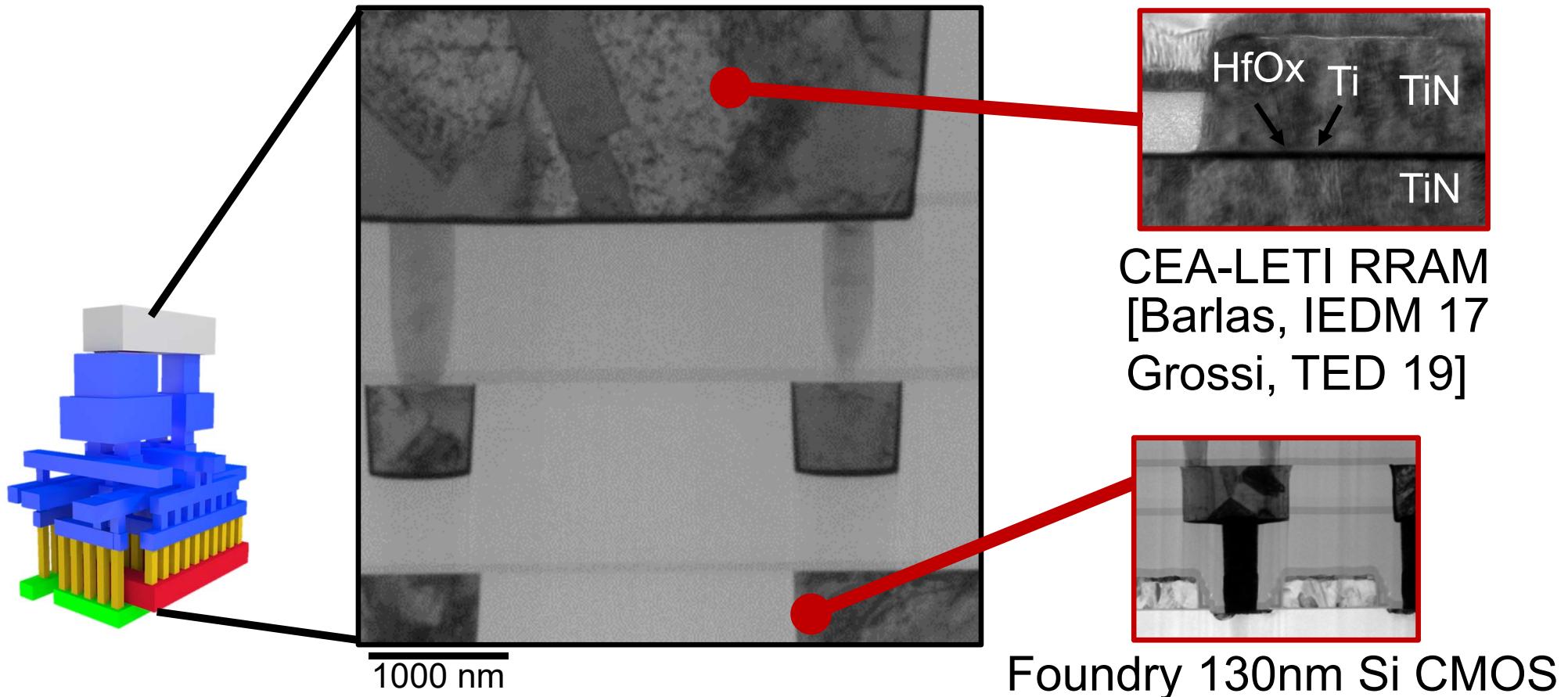
## Scheduler

- Waits for wakeup

[T. Wu, ISSCC 2019]

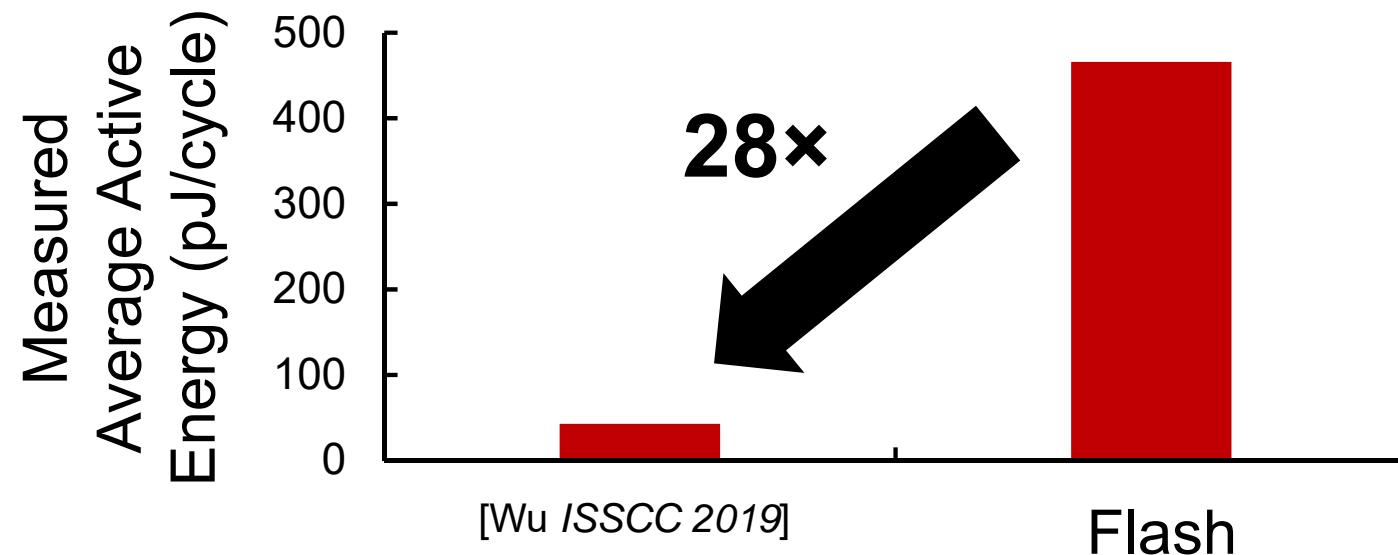
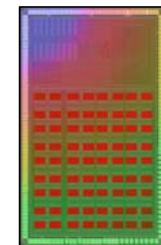
[Wu /SSCC 2019]

# RRAM Integrated on Silicon CMOS



# Active Mode

- **11× - 28×** lower active energy vs. Flash
  - Application dependent

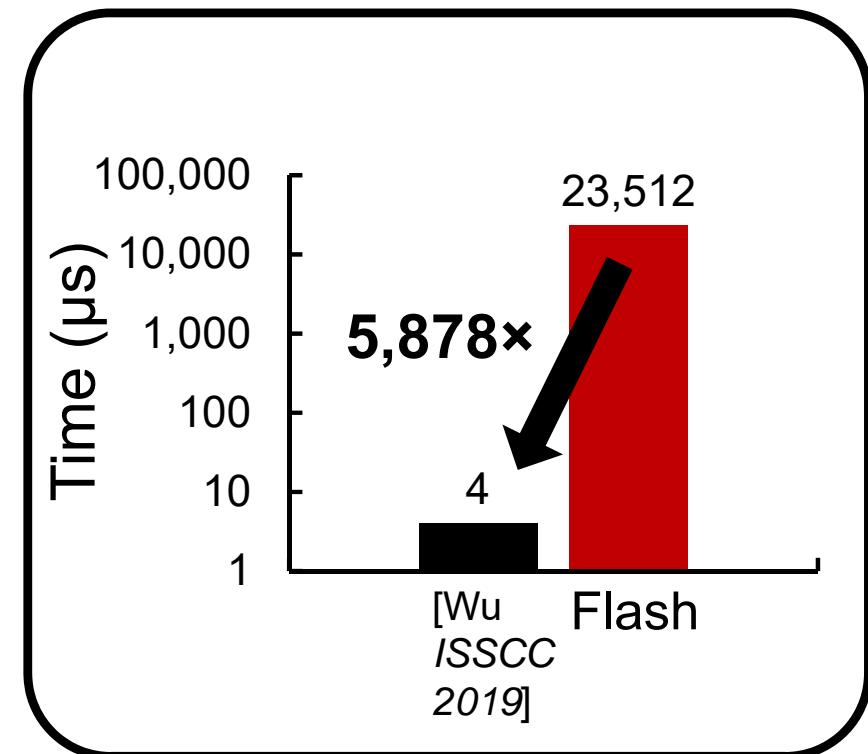
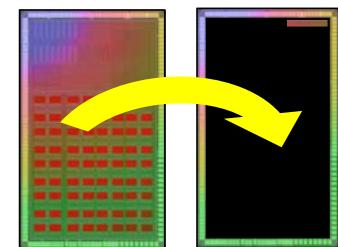


[Wu /SSCC 2019]

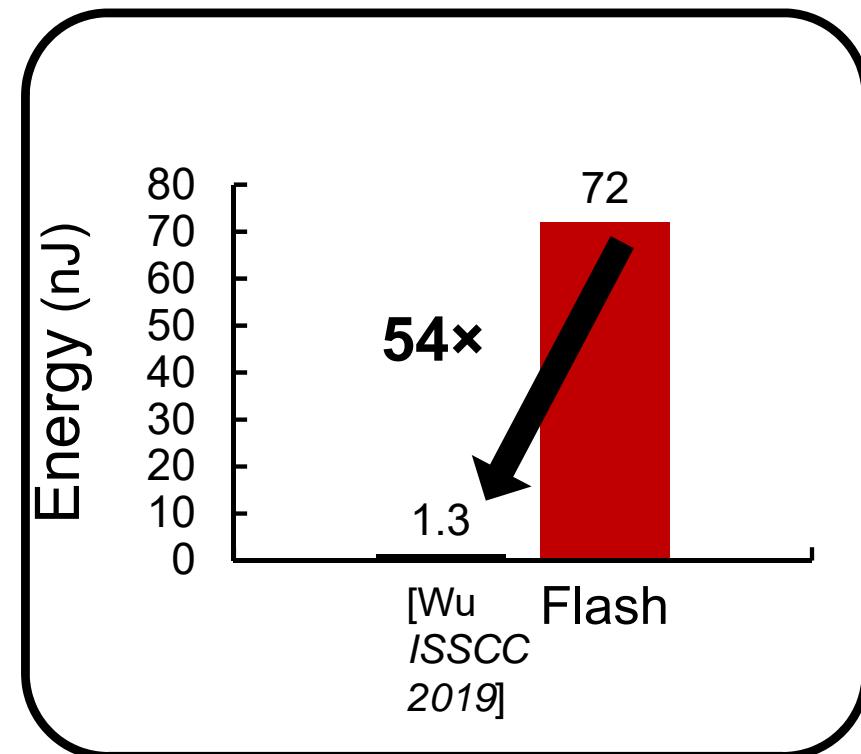
Application: Kalman Filter (control)

# Active to Shutdown

- Fine-grained temporal power gating



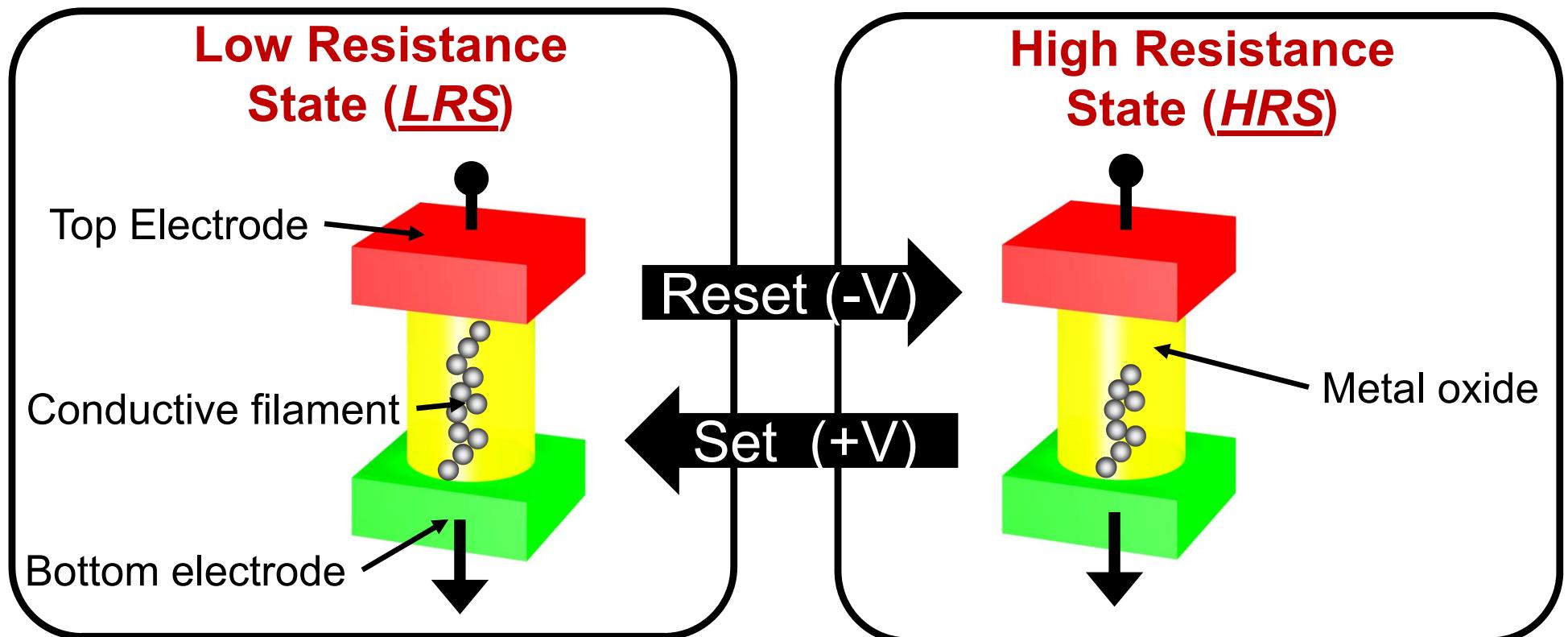
[Wu /ISSCC 2019]



Application: Kalman Filter (control)

# RRAM Endurance

**Maximum Set-Reset cycles before failure**



[Wong Nat. Nanotech 15]

# RRAM: Multiple Failure Sources

Addressed in this work

✓ Fabrication yield

✓ Temporary write failures (TWFs)

✓ Permanent write failures (PWFs)

# RRAM: Multiple Failure Sources

Addressed in this work

✓ Fabrication yield

✓ Temporary write failures (TWFs)

✓ Permanent write failures (PWFs)

**ENDURER**

# Endurance Resiliency using Random Remap

- Random address remap
  - Periodic: e.g., every 30 mins.
- SRAM buffer
  - Filter frequent writes to same RRAM address
- **No extra RRAM**, 16 Bytes extra SRAM, 1% extra RRAM writes

[Wu /SSCC 2019] [Aly, Proc IEEE 2019]

# Random Address Remap

- $\delta$ : random number generated periodically
- $M$ : total RRAM words

**Remap ( $\delta$ )**  
periodically

For each address  $A$ :

Move RRAM [ $A$ ] to  
RRAM  $[(A + \delta) \bmod M]$

**Access (RRAM [A])**  
normal operation

Address redirect

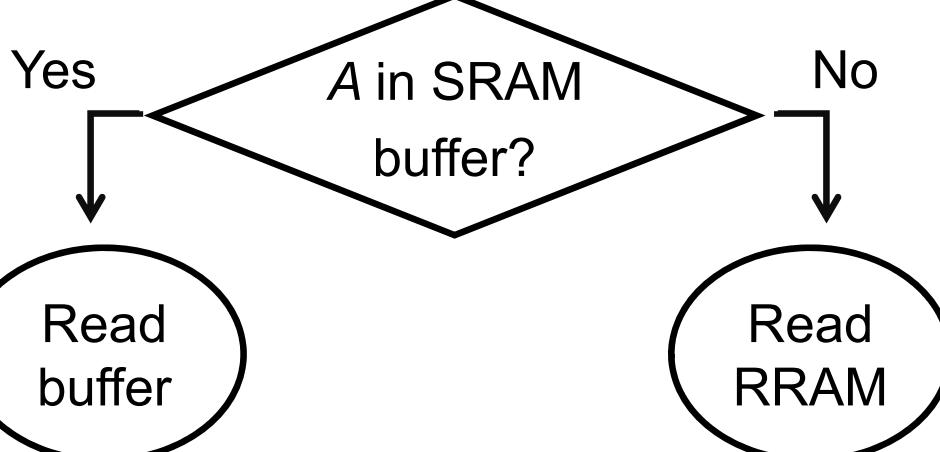
Access  
RRAM  $[(A + \delta) \bmod M]$

# SRAM Buffer

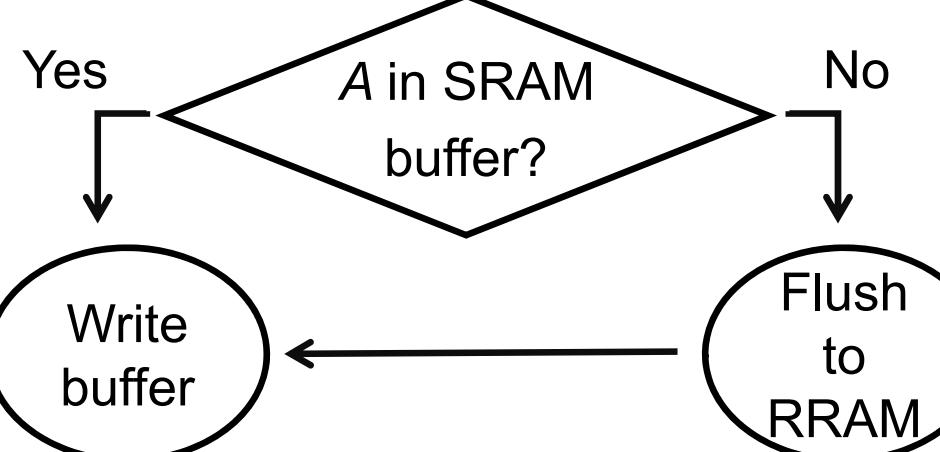
- Fully associative SRAM
  - 16 Bytes in [Wu /SSCC 2019]

[Aly, Proc. IEEE 19]

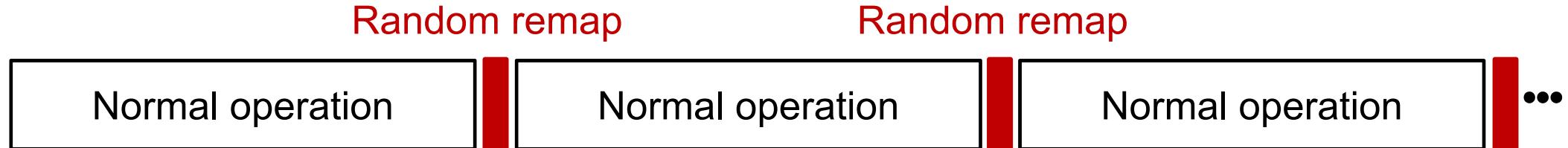
## Read (RRAM [A])



## Write (RRAM [A])



# ENDURER



## Max Set-Reset cycles / RRAM word

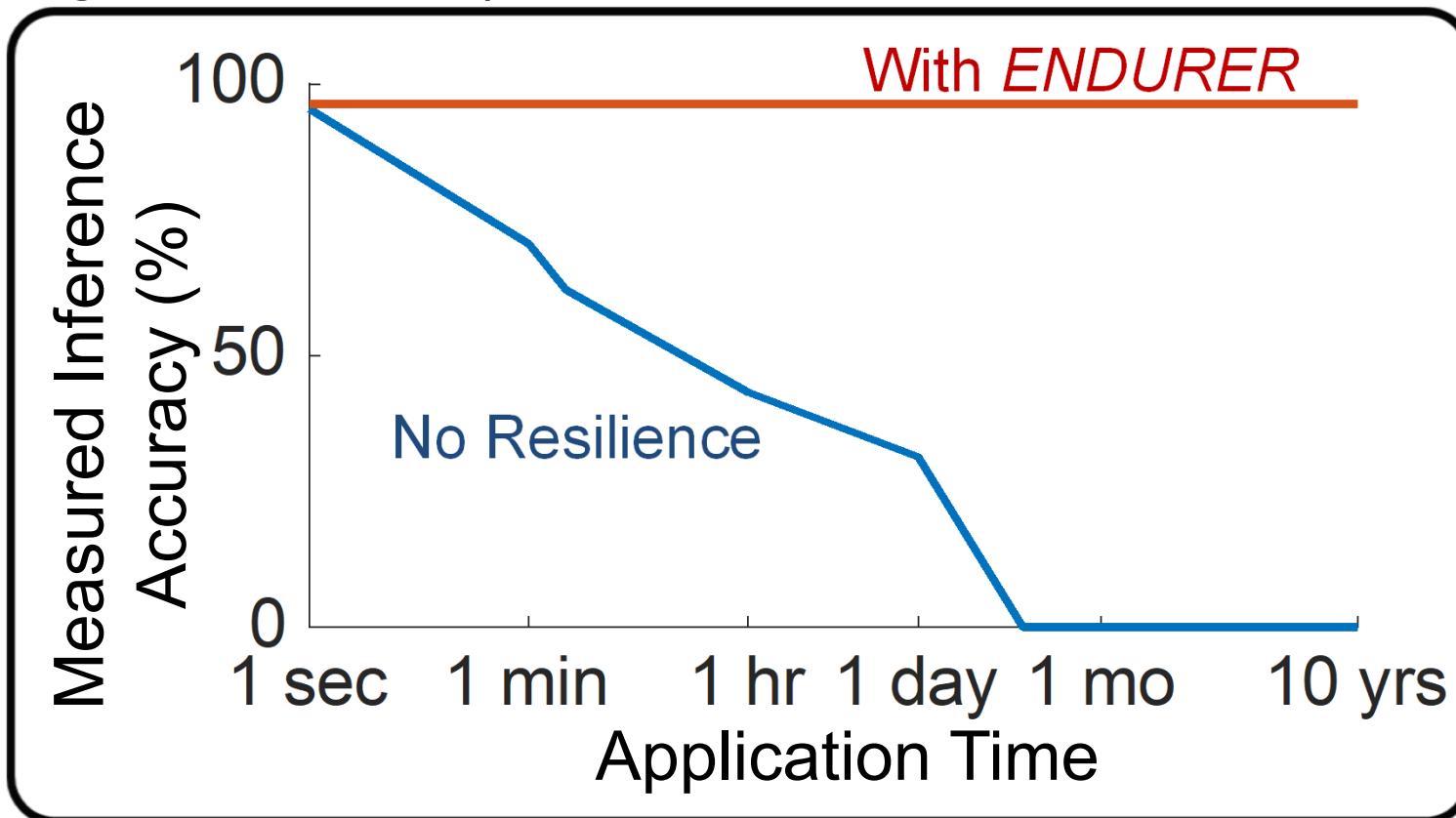
	No ENDURER	Remap	Buffer	Remap + Buffer
MNIST Inference	$2.4 \times 10^{10}$	$10^7$	$4.8 \times 10^8$	$7 \times 10^5$

10 years continuous operation

[Wu ISSCC 2019]  
[Aly, Proc. IEEE 19]

# ENDURER Case Study: Neural Network

- 10 year lifetime (continuous inference in hardware)



Application: CNN (MNIST dataset) running inference continuously

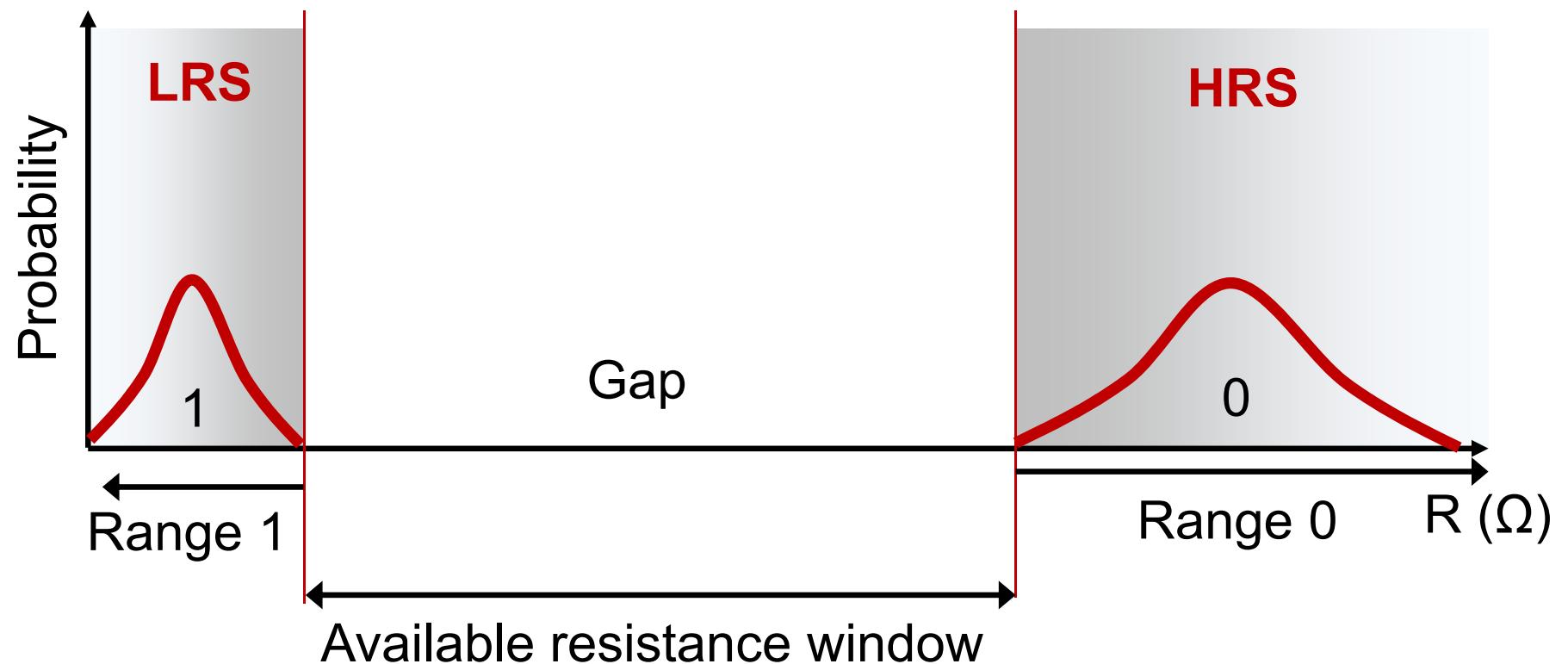
[T. Wu, ISSCC 2019]

# Existing Multi-bit/cell RRAM

	Array level	Bits per cell	Cells measured
[Chien, IEDM 11]	No	3	Standalone single cell
[Prakash, EDL 15]	No	3	Standalone single cell
[Stathopoulos, Scientific Reports 17]	No	6.5	Standalone single cell
[Sheu, Symp. VLSI Circuits 09]	Yes	2	40

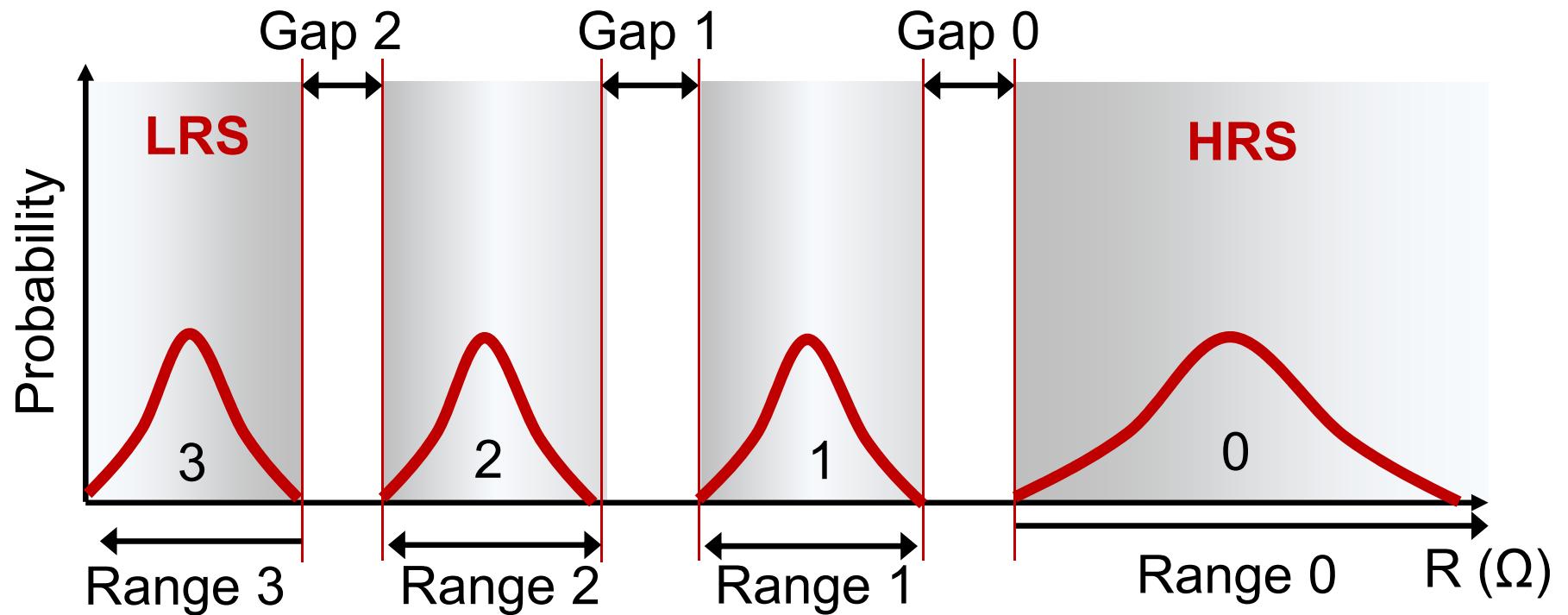
[Le, IEEE TED 19]

# Single-bit/cell RRAM



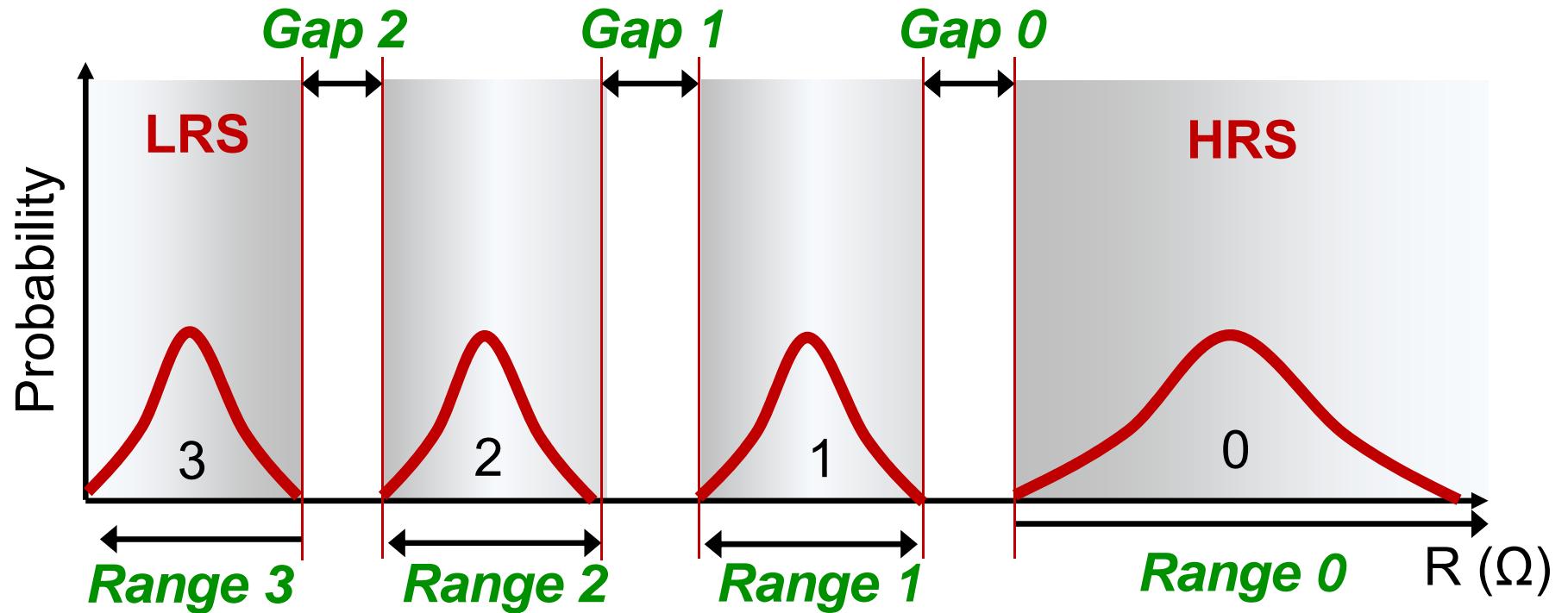
[Le, IEEE TED 19]

# 2-bits/cell RRAM



[Le, IEEE TED 19]

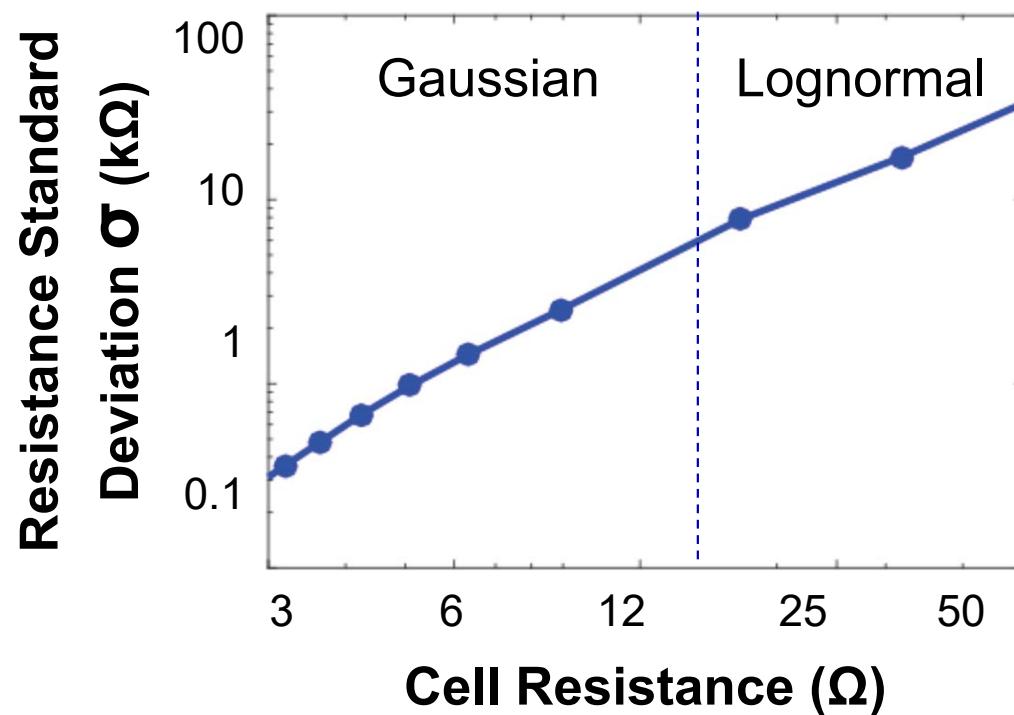
# Gaps and Ranges: How?



[Le, IEEE TED 19]

# RRAM Variations

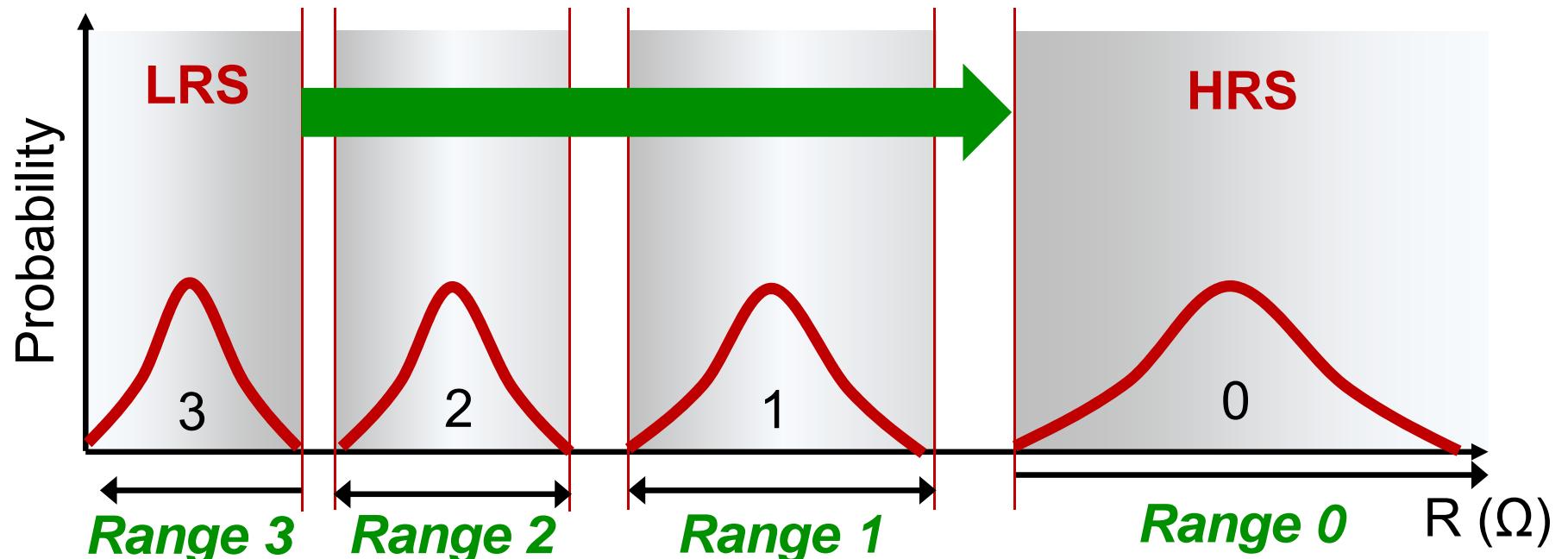
*Higher resistance, more variation*



[Grossi, IEDM 16]

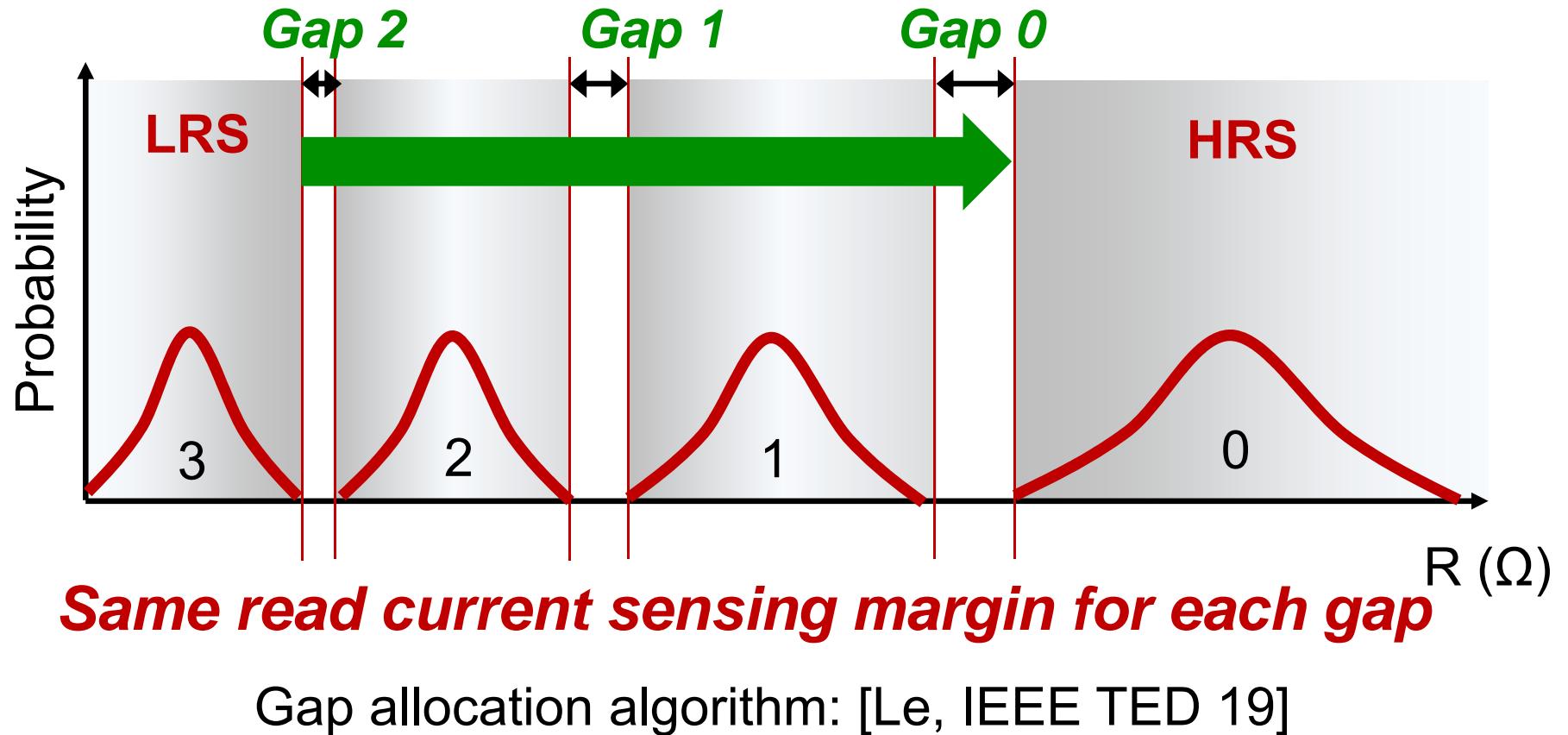
# Variation-aware Range Allocation

*Higher resistance, wider range*



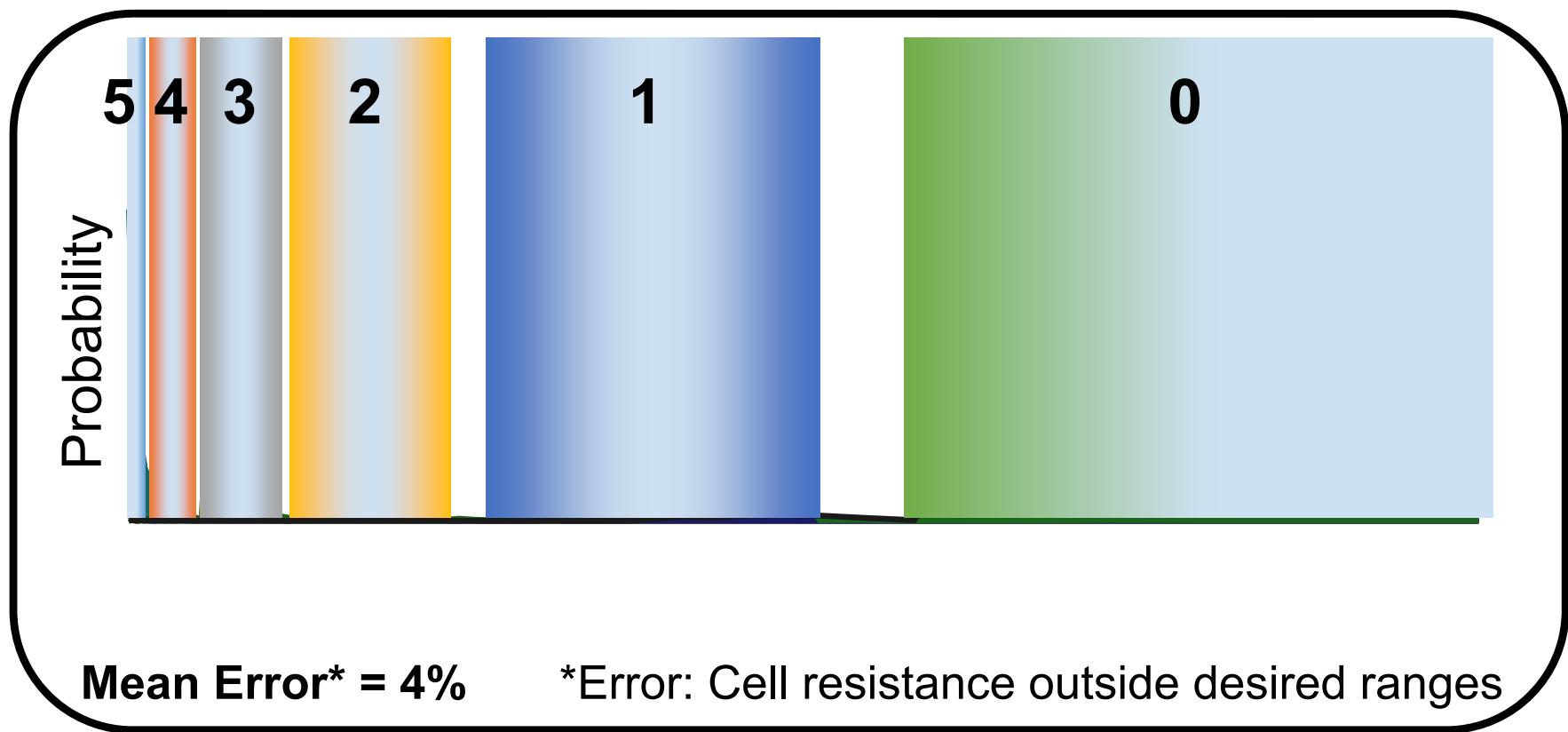
Range allocation algorithm: [Le, IEEE TED 19]

# Larger Gaps for Higher Resistances



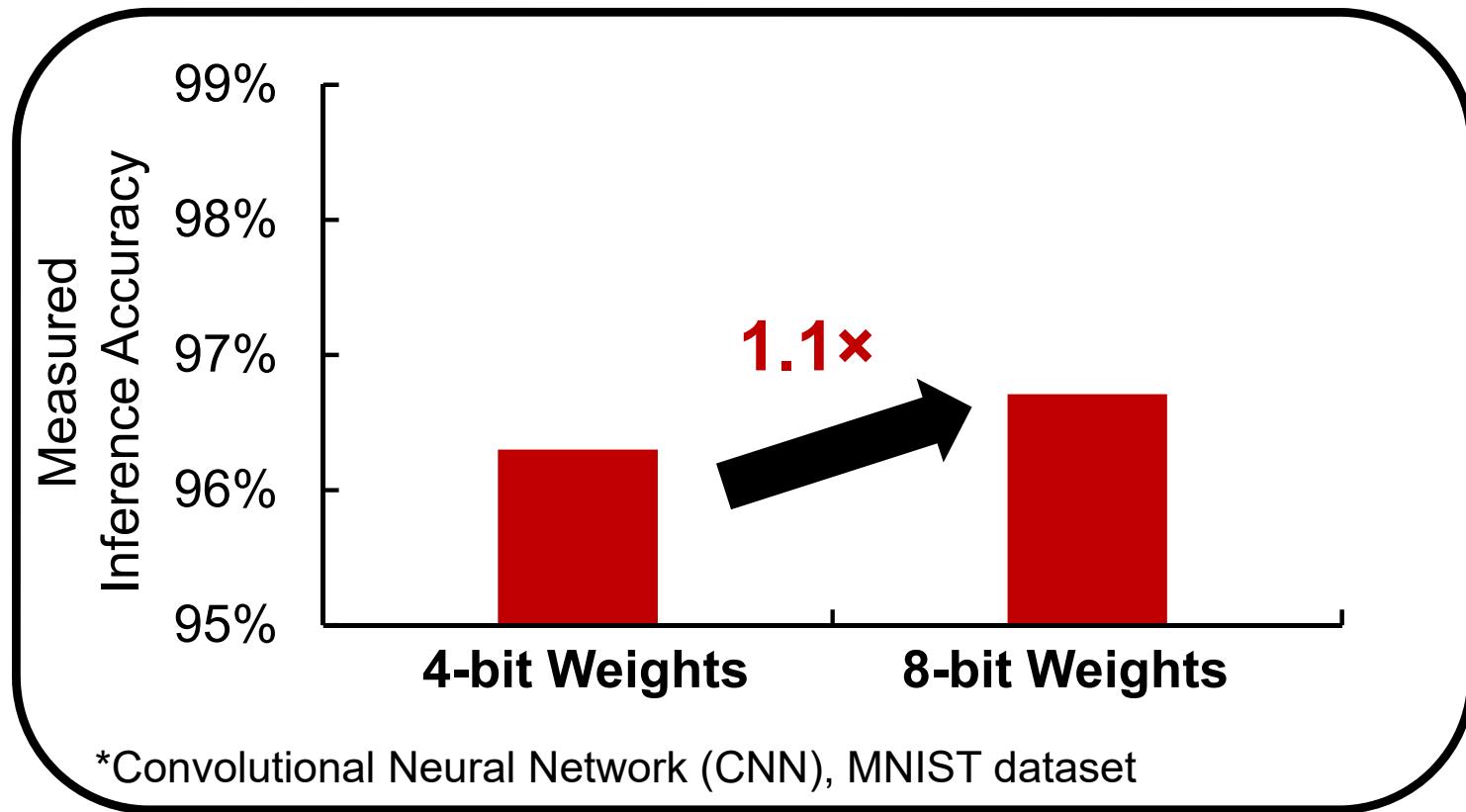
# Multiple bits/cell

**6 levels (2.6 bits) / cell achieved**



# Exploit Increased Capacity

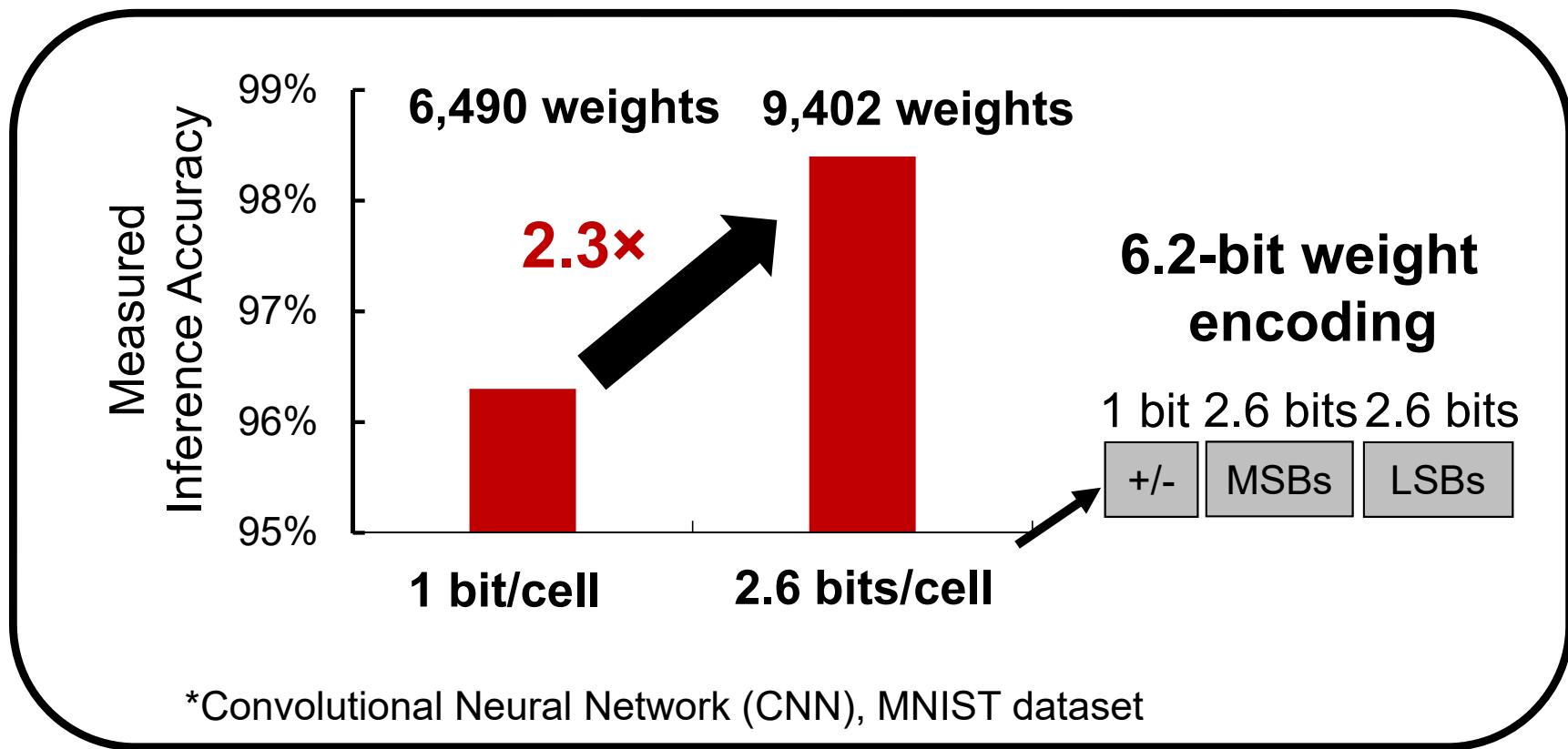
More precise weights?



[Wu /SSCC 2019]

# Exploit Increased Capacity

Larger Network?



[Wu /SSCC 2019]

# RRAM SoA Systems: Comparison Table

WU, ISSCC 19	Liu, ISSCC 16	Su, Symp. VLSI Circuits 17	Chen, ISSCC 18
--------------	------------------	-------------------------------	-------------------

Multiple bits/cell in RRAM system	<b>Yes, 2.6 bits/cell 1<sup>st</sup> System demo</b>	No	
-----------------------------------	----------------------------------------------------------	----	--

Applications	<b>6 (ML, control, security)</b>	1		
Active Power	<b>0.24-0.48 mW (10 MHz)</b>	3.3 mW (100 MHz)	22 mW (20 MHz)	Not reported (64 MHz)
Active to Shutdown	<b>0.3–2.69 nJ / 0.5 – 8 µs</b>	400 nJ / 4µs	5,000 nJ / 100µs	
Shutdown to Active	<b>152 pJ / 200 ns</b>	450 pJ / 130ns	510 pJ / 50 ns	

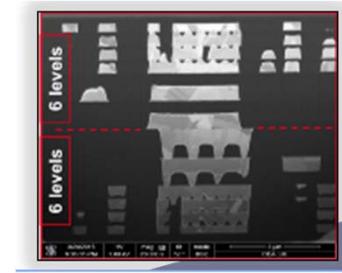
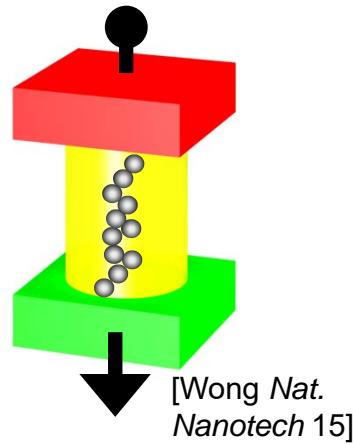
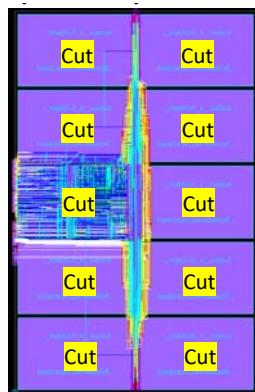
RRAM Resilience	<b>ENDURER + Yield recovery + TWFs</b>	Not reported
Chip Lifetime	<b>10 years (ML inference) with ENDURER</b>	

# Memory power trade-offs – new architectures, design and technologies

**SRAM**  
*Power reduction  
and Error  
resilience*

**Non-volatile  
memories**  
*RRAM Application  
to AI*

**3D integration**  
*New architectures,  
TSVs and Hybrid  
bonding*

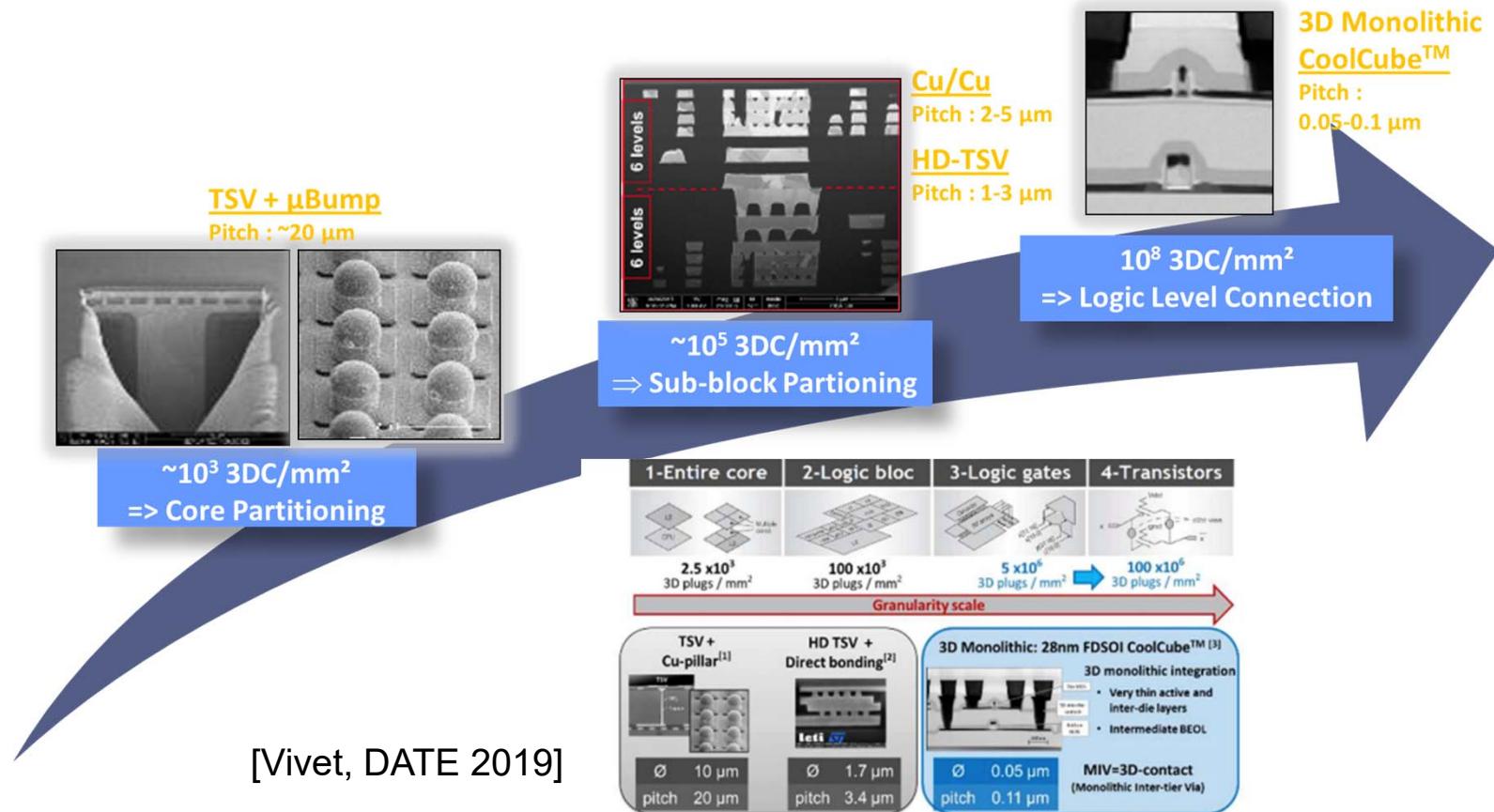


[Vivet, DATE 2019]

# Memory power trade-offs – new architectures, design and technologies

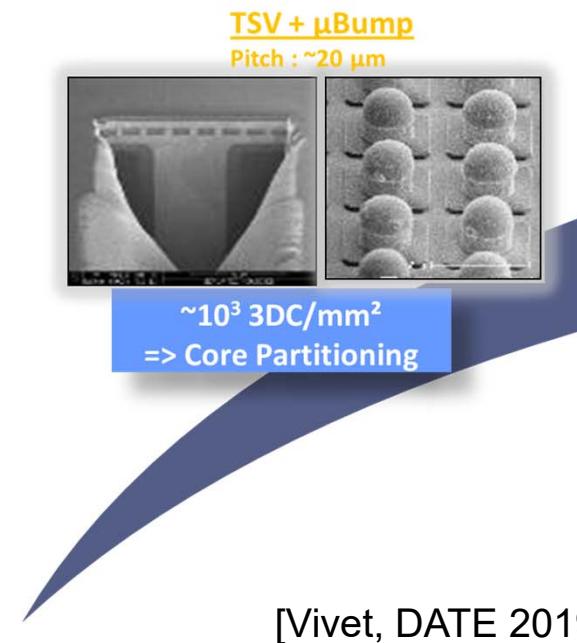
3D integration  
*New architectures,  
TSVs and Hybrid  
bonding*

# 3D stack technologies



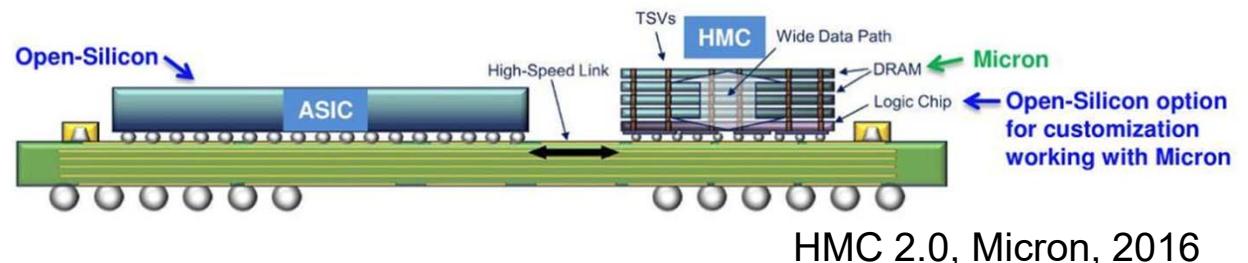
[Vivet, DATE 2019]

# 3D stack technologies

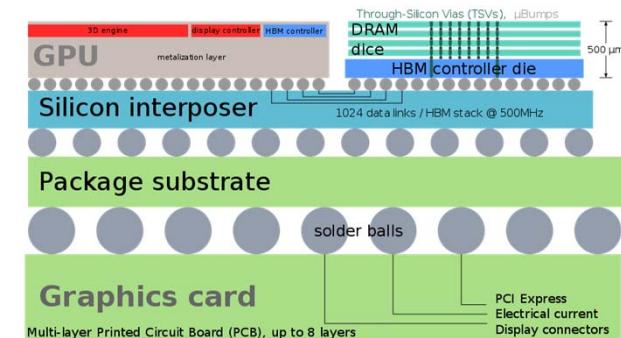


[Vivet, DATE 2019]

# HBM and HMC architectures : stacking DRAMs

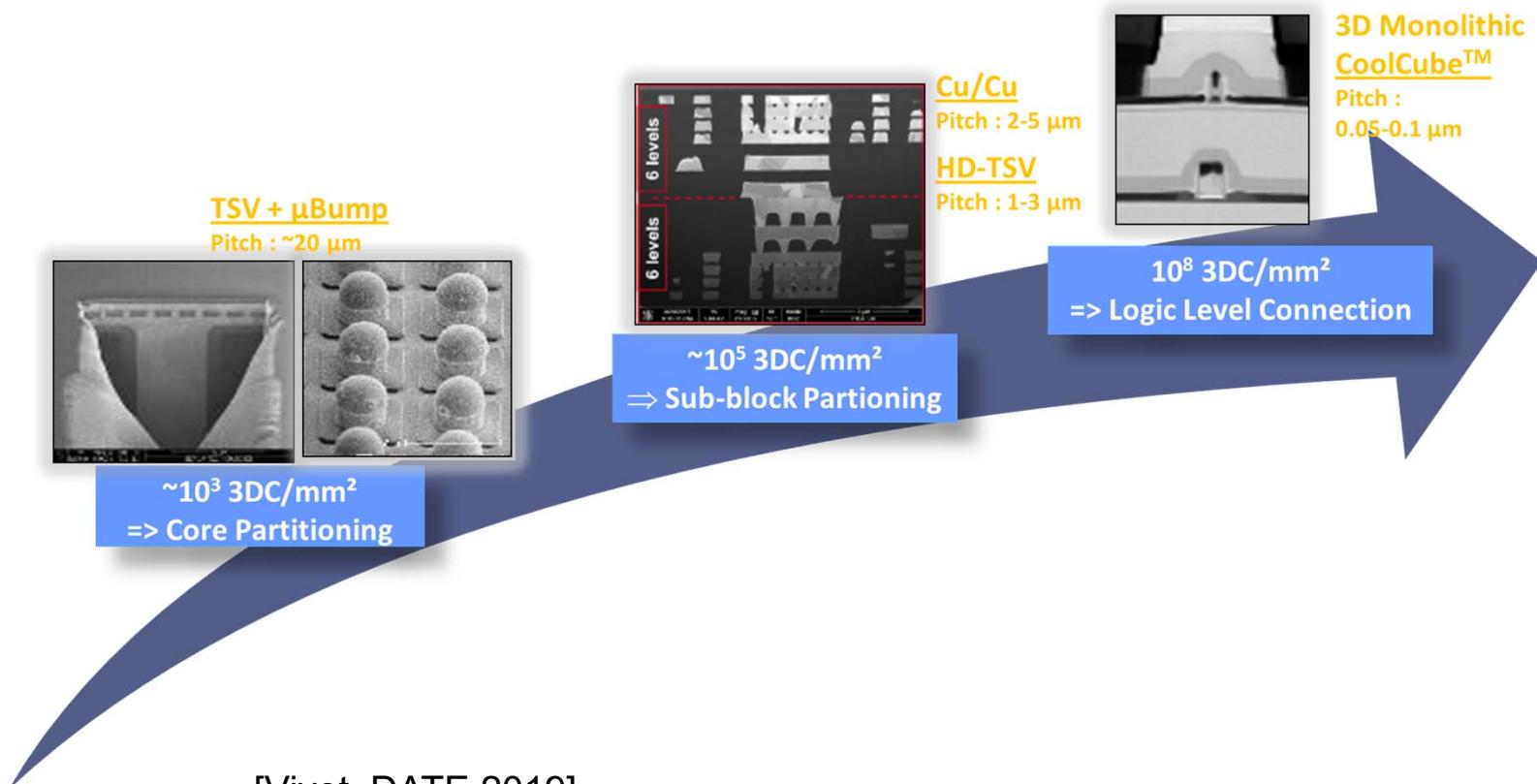


- High performance RAM interfaces for 3D-stacked SDRAM
- Higher bandwidth achieved by stacking DRAM dies
- HBM includes an optional base die with a memory controller, which are interconnected by through-silicon vias (TSVs) and microbumps

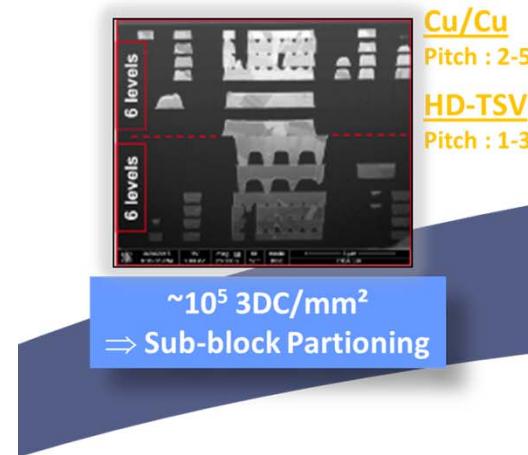


JEDEC, industry standard

# 3D stack technologies



# 3D stack technologies

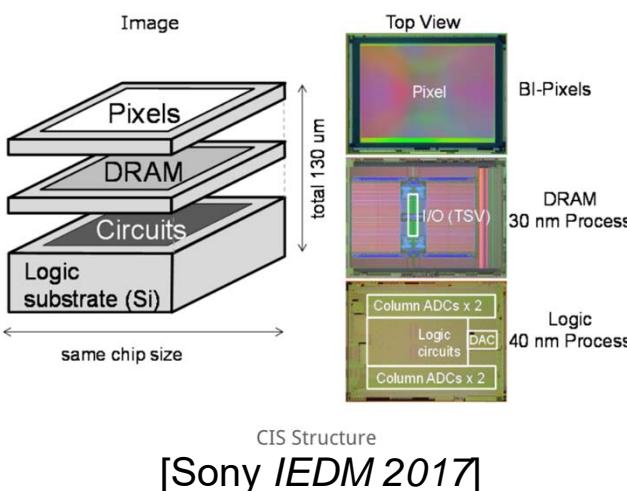


[Vivet, DATE 2019]

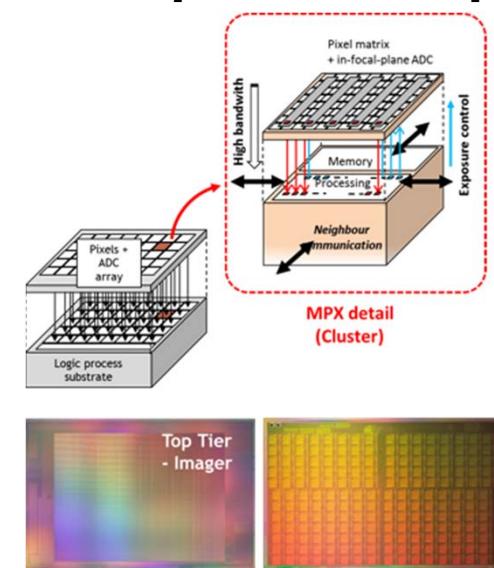
Slide 61

# Hybrid bonding: 3D stacked vision systems

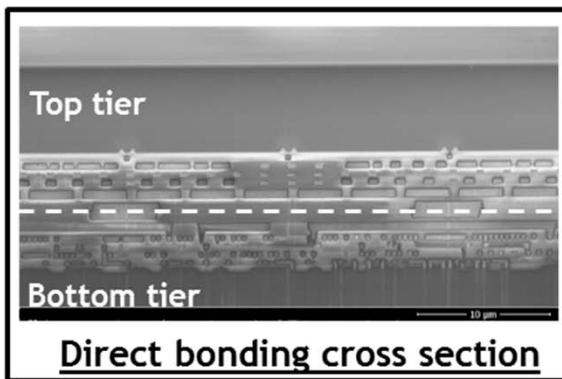
- Image sensors are memory-centric architectures
- 2 to 3 layers 3D stacked BSI vision chip
- Parallel computing by exploiting in-focal-plane pixel readout circuits
- Higher frame rates, without reducing ADC resolution
- Large capacity memories



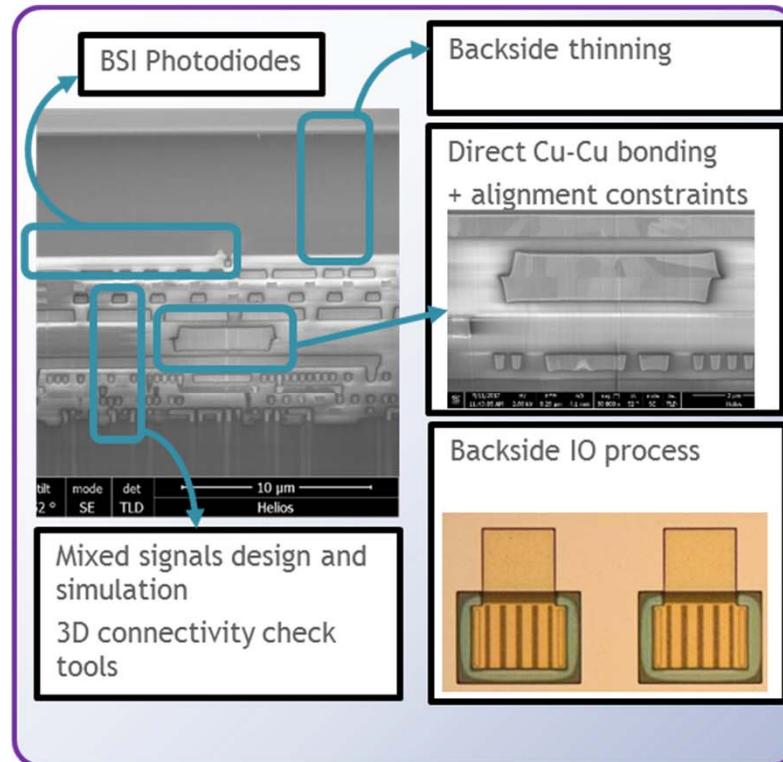
[Millet VLSI 2018]



# 3D stack process for backside imager

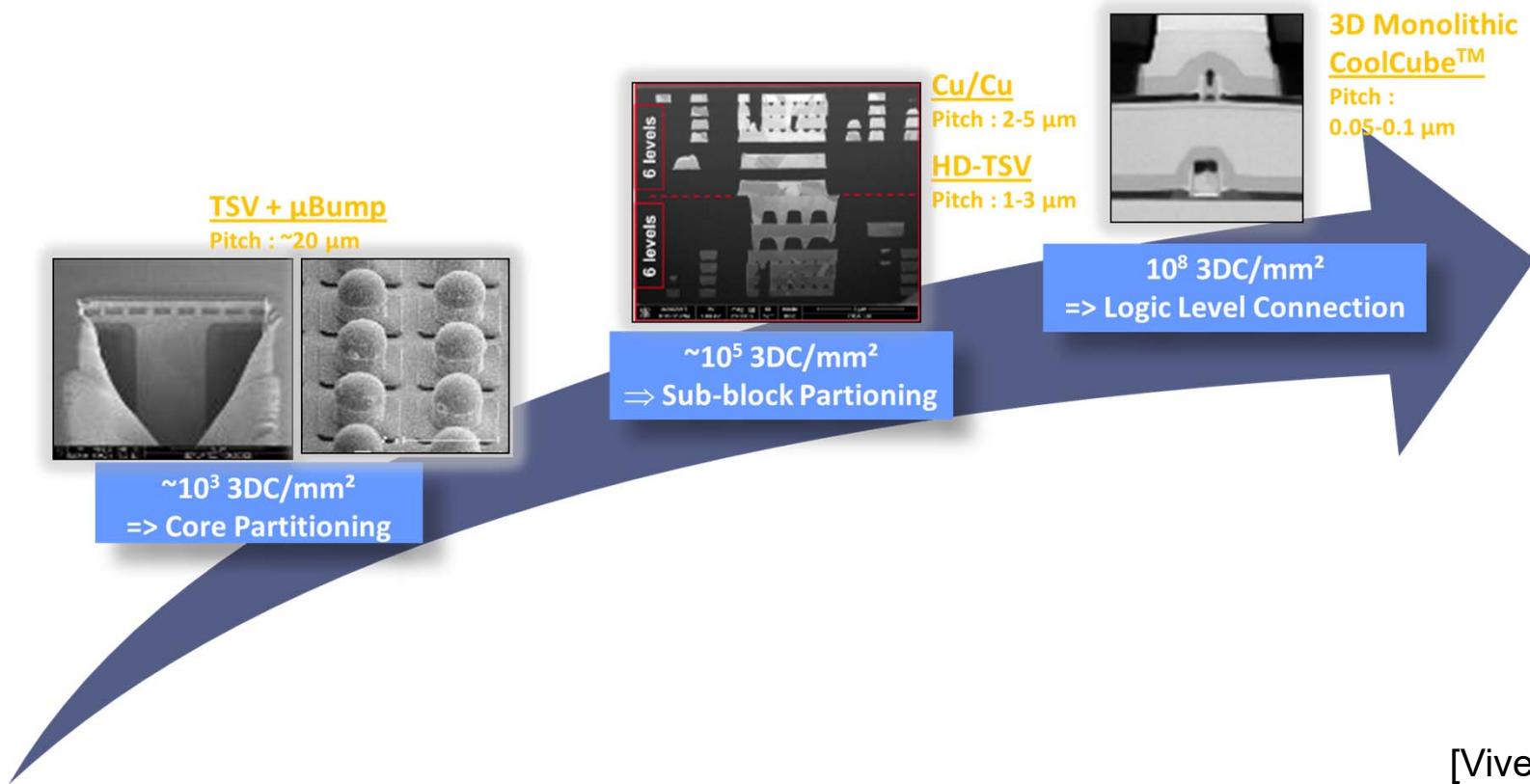


technology	130nm 1P7M / 130nm 1P7M
chip area	187mm <sup>2</sup>
sensor area	113mm <sup>2</sup>
Pixel size	12μm x 12μm
Pixel fill factor	75%
3D bonding	Cu-Cu direct bonding
Nb of 3D contacts	6528 core signals 670 power lines 11.3k pad connections
PE	3072



[Millet VLSI 2018]

# 3D stack technologies



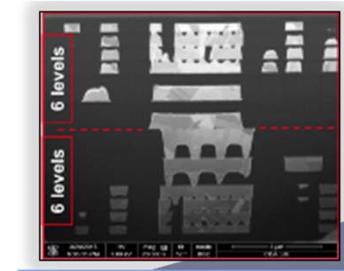
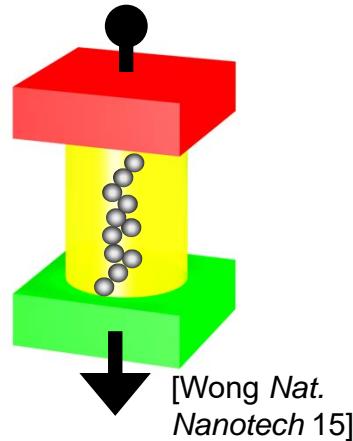
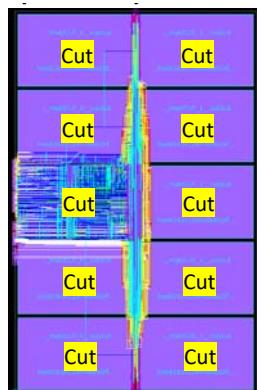
[Vivet, DATE 2019]

# Conclusions

**SRAM**  
*Power reduction  
and Error  
resilience*

**Non-volatile  
memories**  
*RRAM Application  
to AI*

**3D integration**  
*New architectures,  
TSVs and Hybrid  
bonding*



[Vivet, DATE 2019]



# Towards Memory-centric Autonomous Systems: *A Technology and Device Perspective*

Arijit Raychowdhury

School of Electrical and Computer Engineering  
Georgia Institute of Technology

*arijit.raychowdhury@ece.gatech.edu*

# Outline

- Merged Memory and Logic
- Emerging Memory Technologies and Outlook
- Towards Data-Centric Near/In-Memory Systems
  - Data-flow Architectures for ML
  - Compute-Communicate-Iterate for Optimizations
  - Real-Time Learning Systems
- Outlook

# Outline

- Merged Memory and Logic
- Emerging Memory Technologies and Outlook
- Towards Data-Centric Near/In-Memory Systems
  - Data-flow Architectures for ML
  - Compute-Communicate-Iterate for Optimizations
  - Real-Time Learning Systems
- Outlook

# Exploding Model Sizes in AI

- Rapid increase in model sizes leads to memory capacity and bandwidth demand

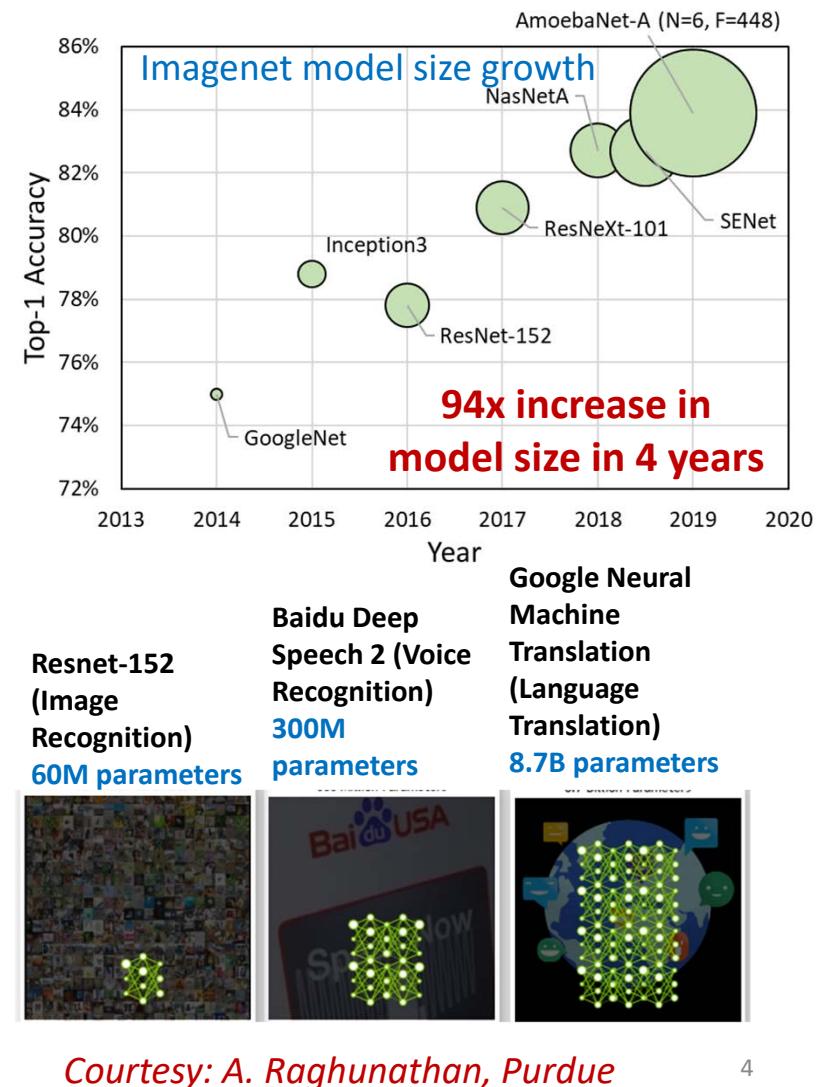
OUTRAGEOUSLY LARGE NEURAL NETWORKS:  
THE SPARSELY-GATED MIXTURE-OF-EXPERTS LAYER

Noam Shazeer<sup>1</sup>, Azalia Mirhoseini<sup>\*1</sup>, Krzysztof Maziarz<sup>\*2</sup>, Andy Davis<sup>1</sup>, Quoc Le<sup>1</sup>, Geoffrey Hinton<sup>1</sup> and Jeff Dean<sup>1</sup>

<sup>1</sup>Google Brain, {noam,azalia,andydavis,qvl,geoffhinton,jeff}@google.com  
<sup>2</sup>Jagiellonian University, Cracow, krzysztof.maziarz@student.uj.edu.pl

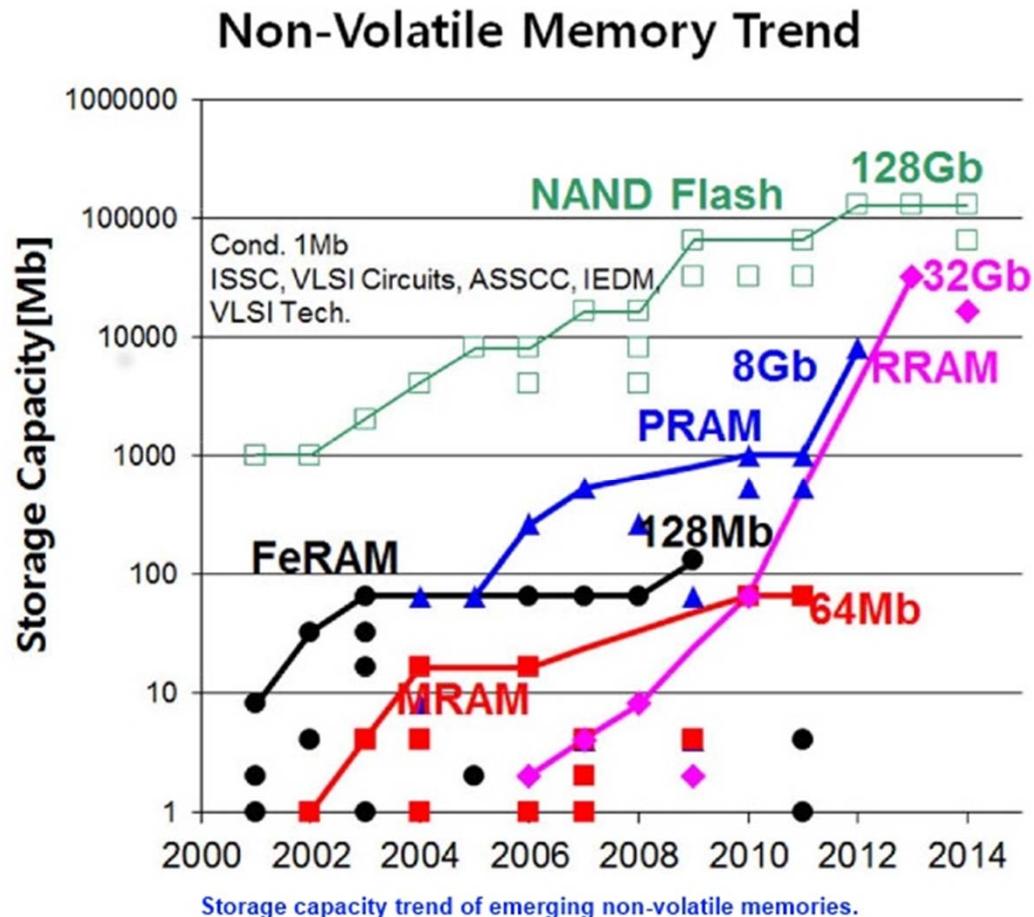
ABSTRACT

The capacity of a neural network to absorb information is limited by its number of parameters. Conditional computation, where parts of the network are active on a per-example basis, has been proposed in theory as a way of dramatically increasing model capacity without a proportional increase in computation. In practice, however, there are significant algorithmic and performance challenges. In this work, we address these challenges and finally realize the promise of conditional computation, achieving greater than 1000x improvements in model capacity with only minor losses in computational efficiency on modern GPU clusters. We introduce a Sparsely-Gated Mixture-of-Experts layer (MoE), consisting of up to thousands of feed-forward sub-networks. A trainable gating network determines a sparse combination of these experts to use for each example. We apply the MoE to the tasks of language modeling and machine translation, where model capacity is critical for absorbing the vast quantities of knowledge available in the training corpora. We show that our models achieve state-of-the-art performance with only 60M parameters! On large language models, our models achieve significantly better performance with 137B parameters!



# Exploding Model Sizes in AI

- Rapid growth to memory bandwidth

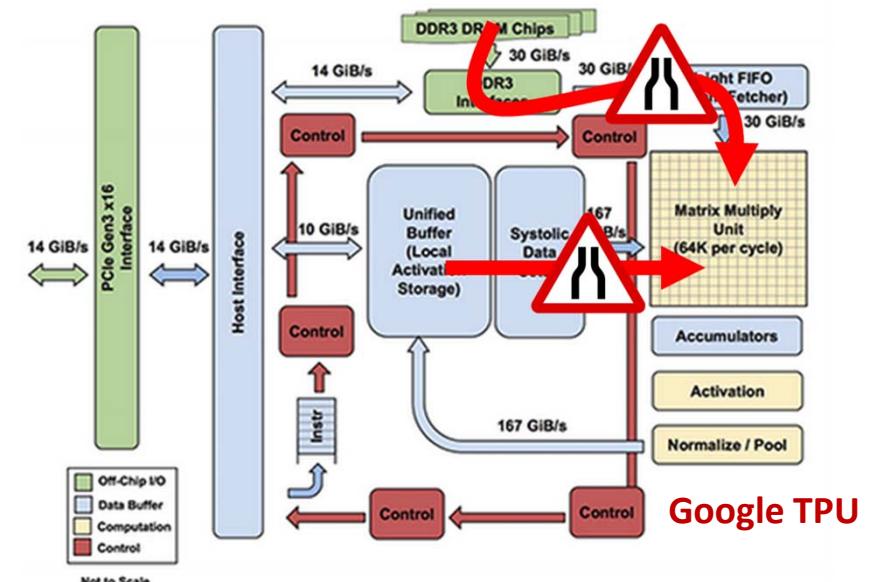
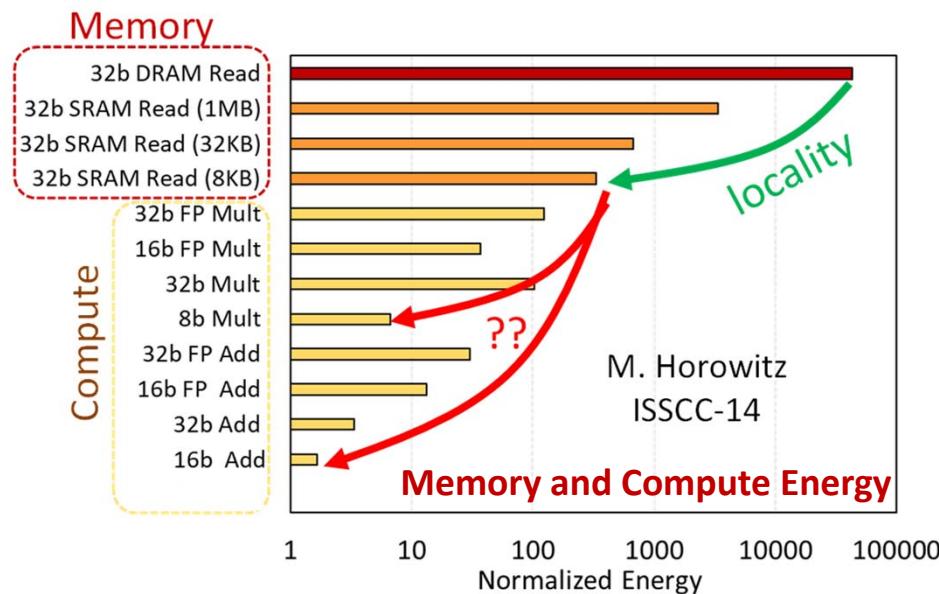


## Can NVMs help?

- + High density
- + Ultra-low leakage
- + Resistive and analog states
- + Support for In-Memory compute

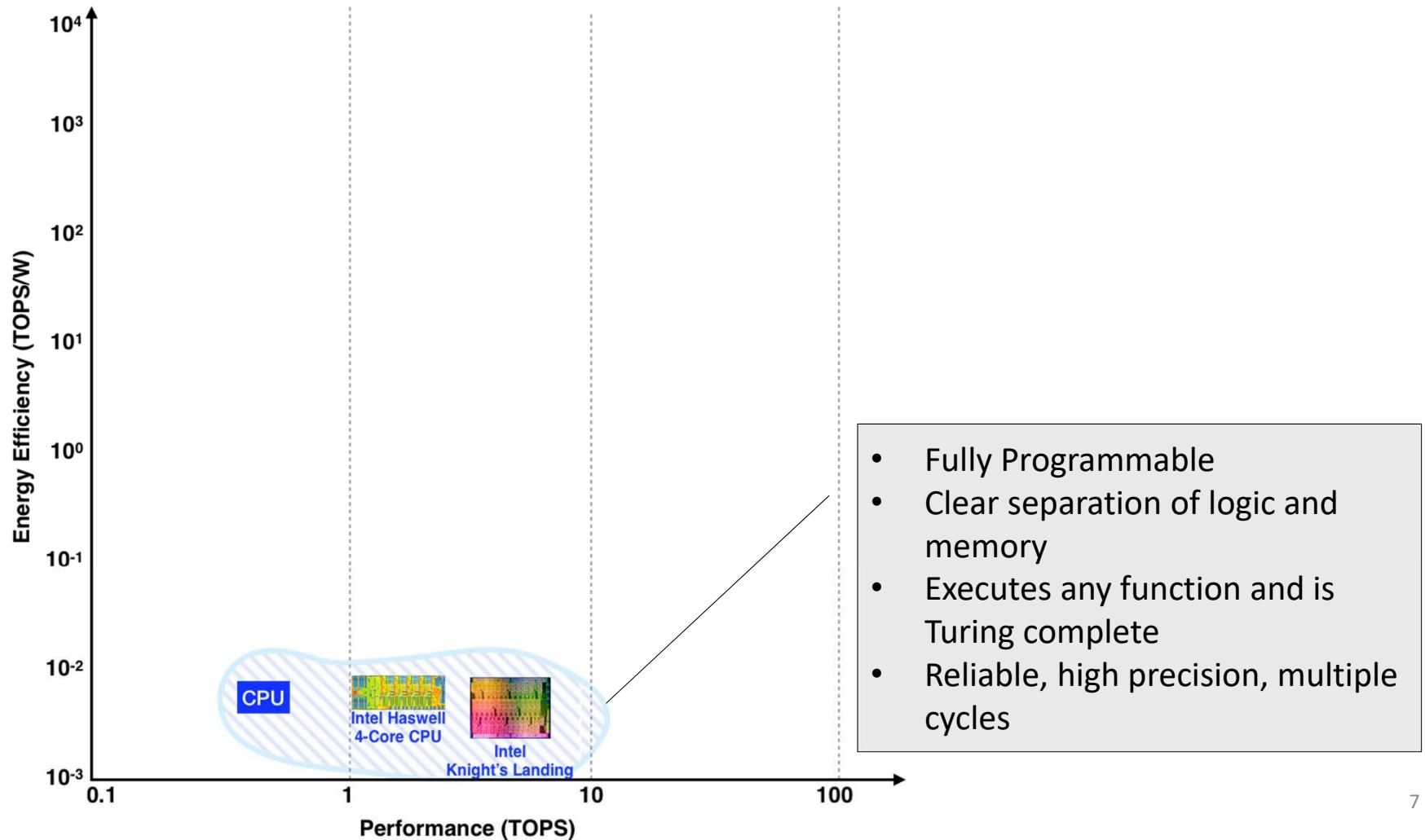
- High write energy
- High write latency
- Limited endurance

# Challenge: Memory Access Bottleneck

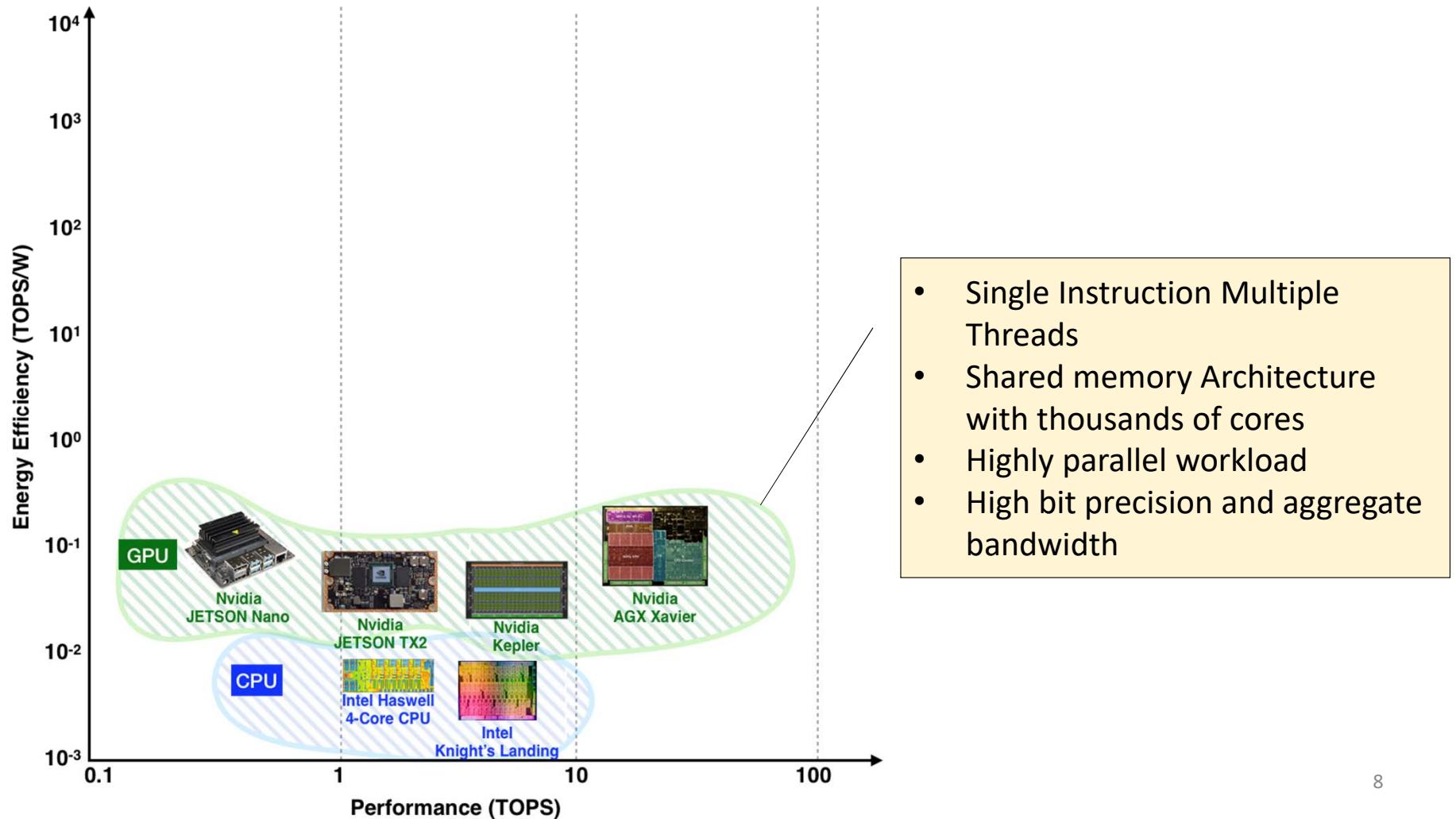


- Memory access energy 1-3 orders of magnitude higher than compute
- Accelerators help by exploiting locality and near-memory computing
- However... super-linear gains in compute energy with precision scaling make the memory bottleneck worse!

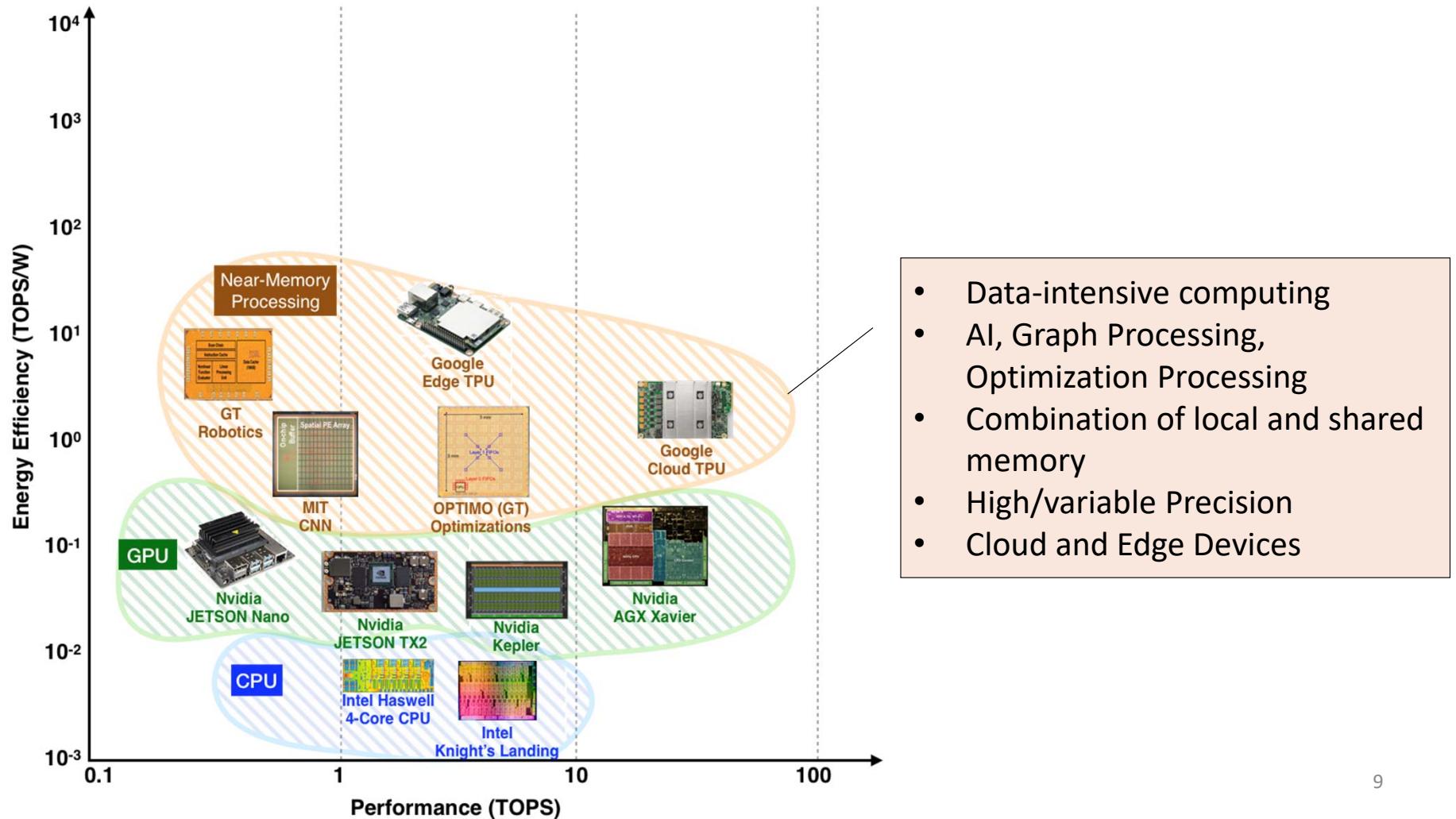
# Computing: Power – Performance Design-Space



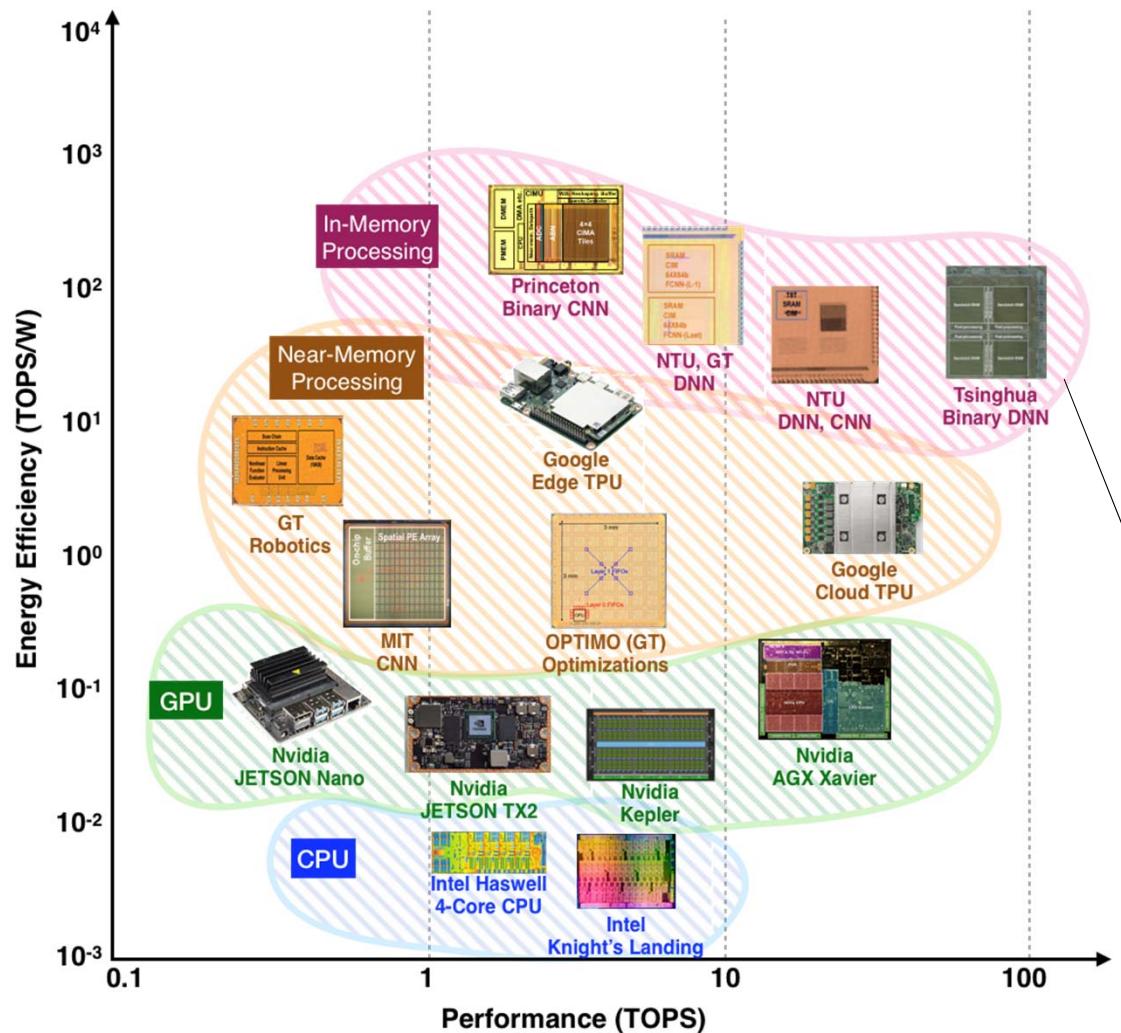
# Computing: Power – Performance Design-Space



# Computing: Power – Performance Design-Space

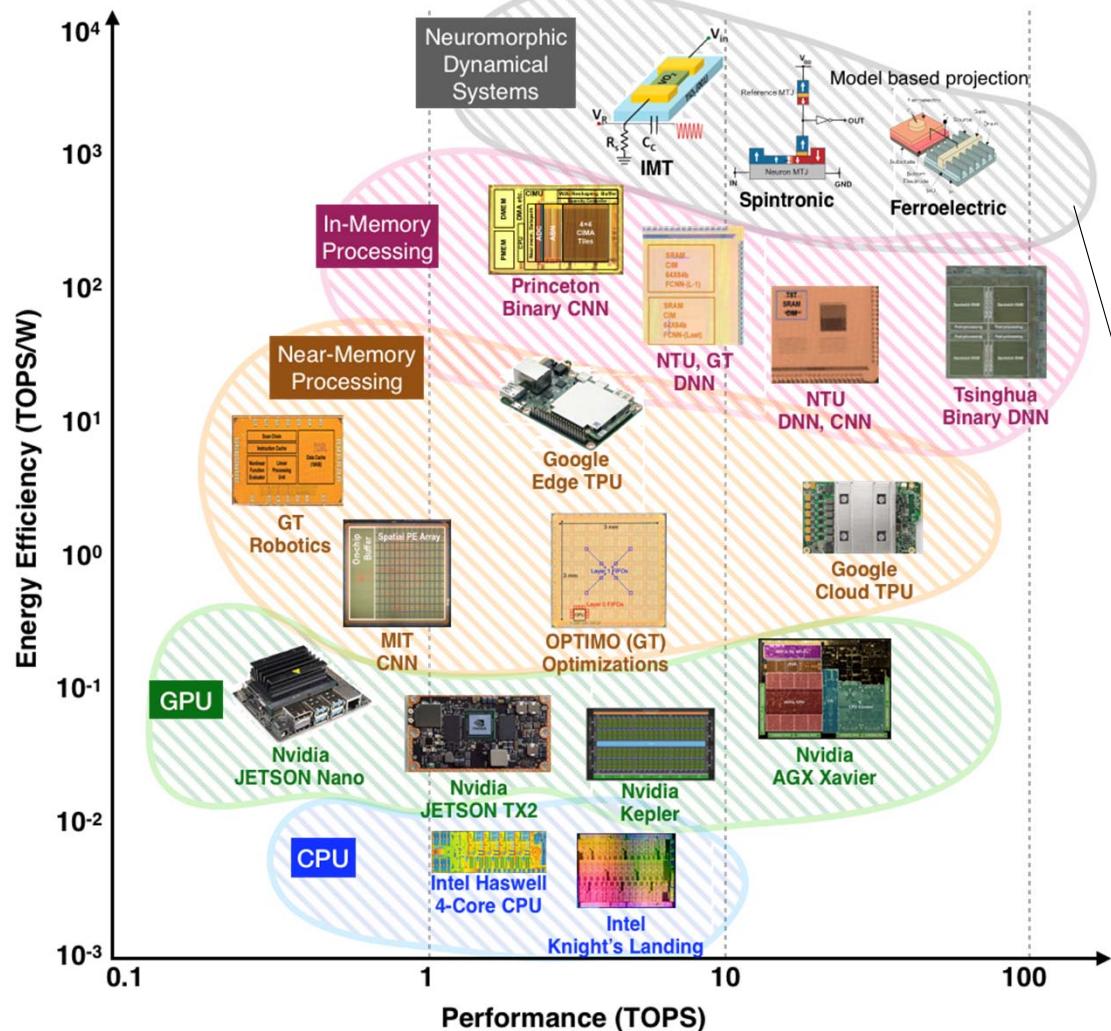


# Computing: Power – Performance Design-Space



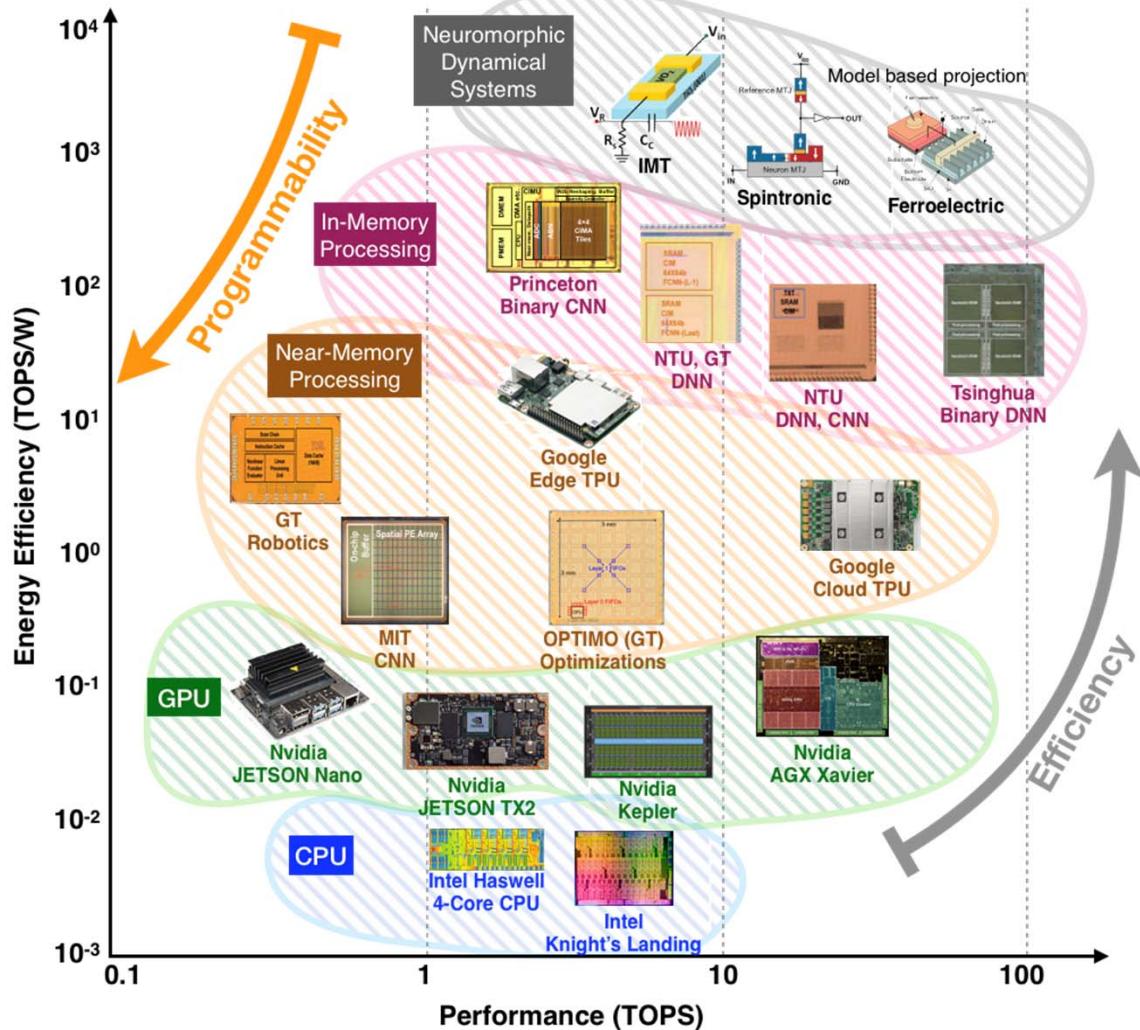
- Vector Processing
- AI with Limited Model Size
- Analog Addition on the Bit-Lines
- Limited Bit Resolution
- Expensive peripherals but O(1) Vector Product
- Mostly Edge Devices

# Computing: Power – Performance Design-Space



- Brain-inspired Continuous-time Dynamics
- State Information Embedded in Voltage/Time/Phase
- Spiking Networks, Hopfield Networks
- Merged Logic and Memory
- General AI, Autonomous Systems. Optimization Solvers

# Computing: Power – Performance Design-Space



Benchmarking Need across the stack

- (1) **Accuracy:** Algorithmic accuracy
- (2) **Convergence:** Accuracy of the solution
- (3) Energy and delay per op (**op** = inference, training, optimizations)

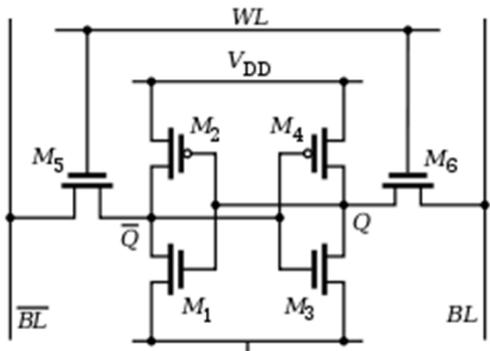
- (1) **Accuracy:** Arithmetic accuracy amidst variations, non-linearity
- (2) Energy and delay per op (**op** = vector MAC, search)

- (1) **Memory bit-cell Read-Write-Endurance**
- (2) Energy and delay per op (**op** = RD, WR)
- (3) Tx energy/op and delay/op

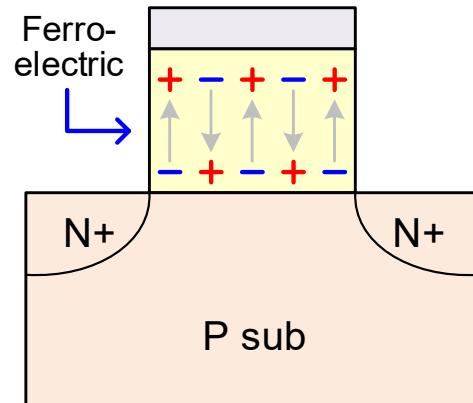
# Outline

- Merged Memory and Logic
- Emerging Memory Technologies and Outlook
- Towards Data-Centric Near/In-Memory Systems
  - Data-flow Architectures for ML
  - Compute-Communicate-Iterate for Optimizations
  - Real-Time Learning Systems
- Outlook

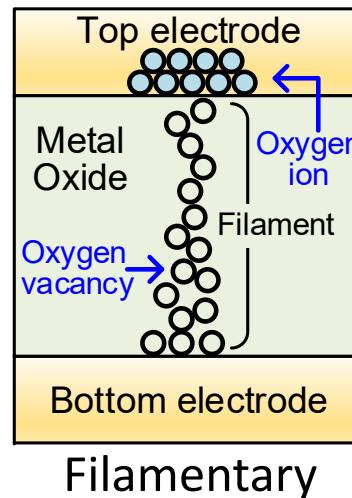
# Embedded Memory Technologies



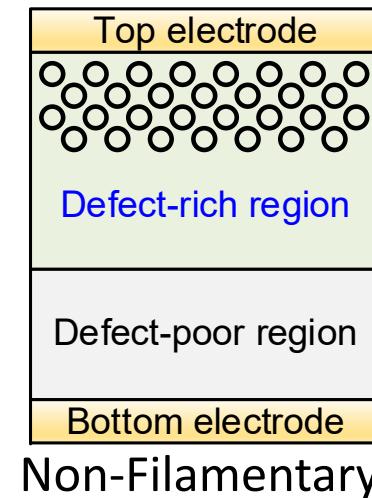
SRAM



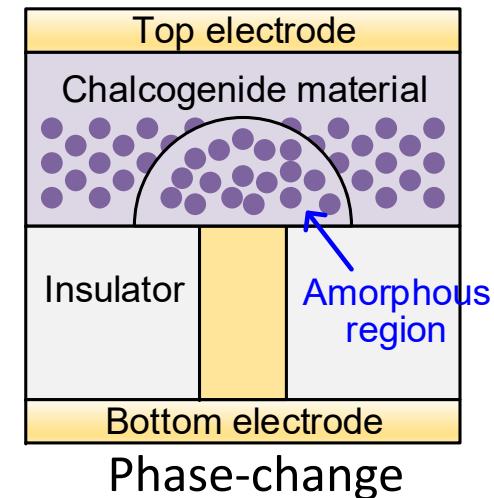
Ferro-Electric



Filamentary



Non-Filamentary



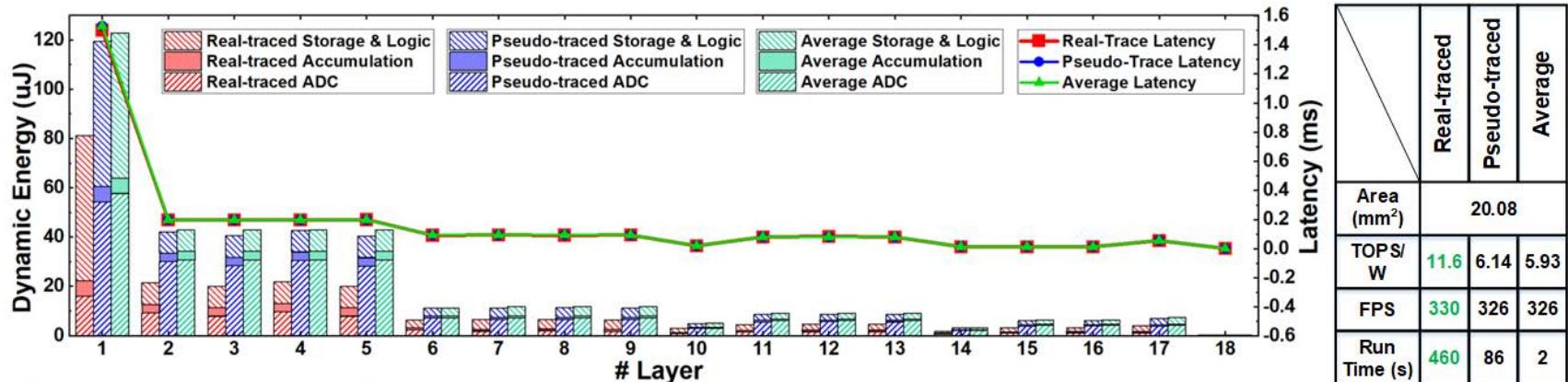
Phase-change

- Embedded memory technologies should retain high density with low-latency.
- Read (and write) latency should be acceptable to avoid compute bottlenecks.
- Process and voltage compatibility is a must.

*Courtesy: S. Yu, Georgia Tech*

# Emerging eNVM Technologies

VGG-8 (8-bit activation; 8-bit weight) on CIFAR10, with Novel Weight Mapping and Dataflow									
Technology node (LSTP)	7 nm		32 nm						
Device	SRAM		SRAM		RRAM (Intel)	TaOx/HfOx (TsingHua)	GST PCM (IBM)	HZO FeFET (NotreDame)	ECRAM (IBM)
ADC precision	Sequential	4-bit	Sequential	4-bit	5-bit	5-bit	5-bit	5-bit	5-bit
Cell Precision	1-bit		1-bit		2-bit	4-bit	4-bit	4-bit	4-bit
Ron ( $\Omega$ )	\	\	\	\	6k	100k	40k	500k	500M
On/Off Ratio	\	\	\	\	17	10	12.5	100	40
Inference Accuracy (%)	92%		92%		91%				
Area (mm <sup>2</sup> )	4.65	4.28	97.83	87.47	86.07	20.45	22.63	19.71	19.71
Memory Utilization (%)	99.29%	99.29%	99.29%	99.29%	98.69%	97.05%	97.05%	97.05%	97.05%
L-by-L Latency (ms)	0.85	0.15	1.61	0.33	19.92	1.16	2.65	0.38	0.28
L-by-L DynamicEnergy (uJ)	13.25	10.63	162.79	76.82	285.96	32.17	38.41	27.69	28.57
L-by-L Leakage power (mW)	104.85	101.69	1.41	1.33	0.22	0.11	0.11	0.11	0.11
Energy Efficiency (TOPS/W)	3.95	14.95	3.70	7.92	2.10	18.97	15.76	22.17	21.51
Throughput (FPS)	1171.15	6875.94	619.53	3001.13	50.16	859.75	378.03	2617.24	3623.67



\* Storage & Logic: buffers, digital logic modules (e.g. decoder, switch matrix, mux), interconnect

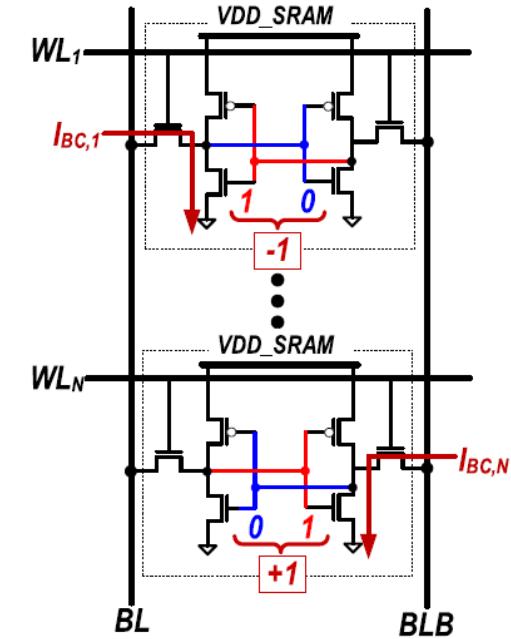
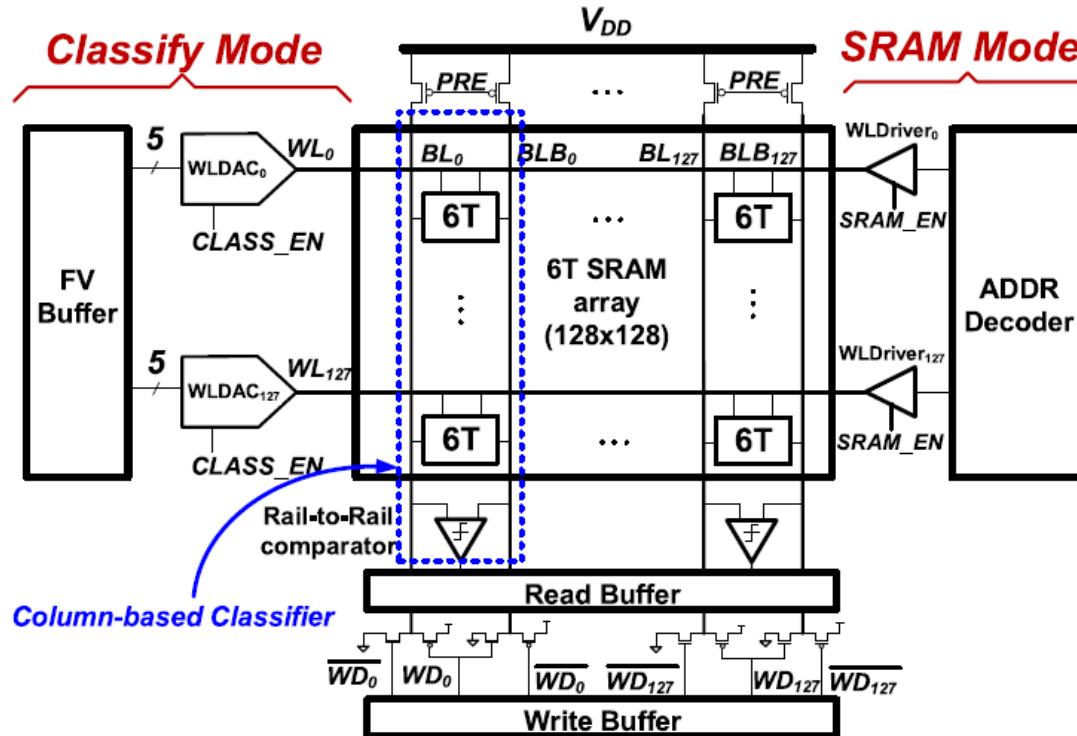
\* Accumulation: adders, shift+adders, adder trees, accumulation units

Courtesy: S. Yu, Georgia Tech

# Outline

- Merged Memory and Logic
- Emerging Memory Technologies and Outlook
- **Towards Data-Centric Near/In-Memory Systems**
  - Data-flow Architectures for ML
  - Compute-Communicate-Iterate for Optimizations
  - Real-Time Learning Systems
- Outlook

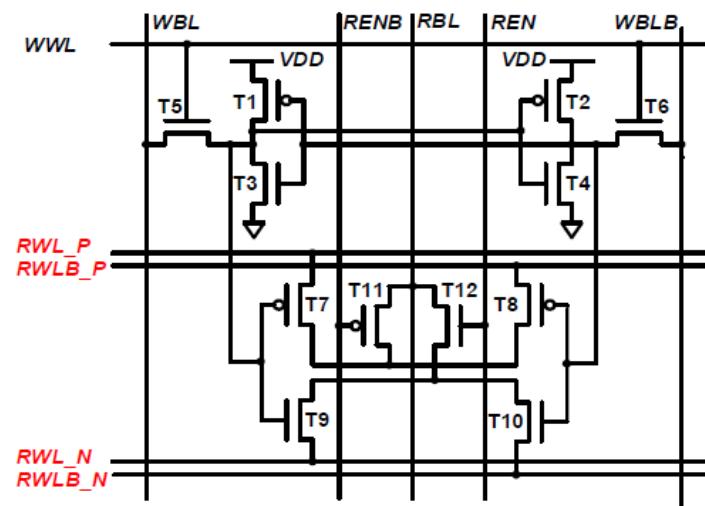
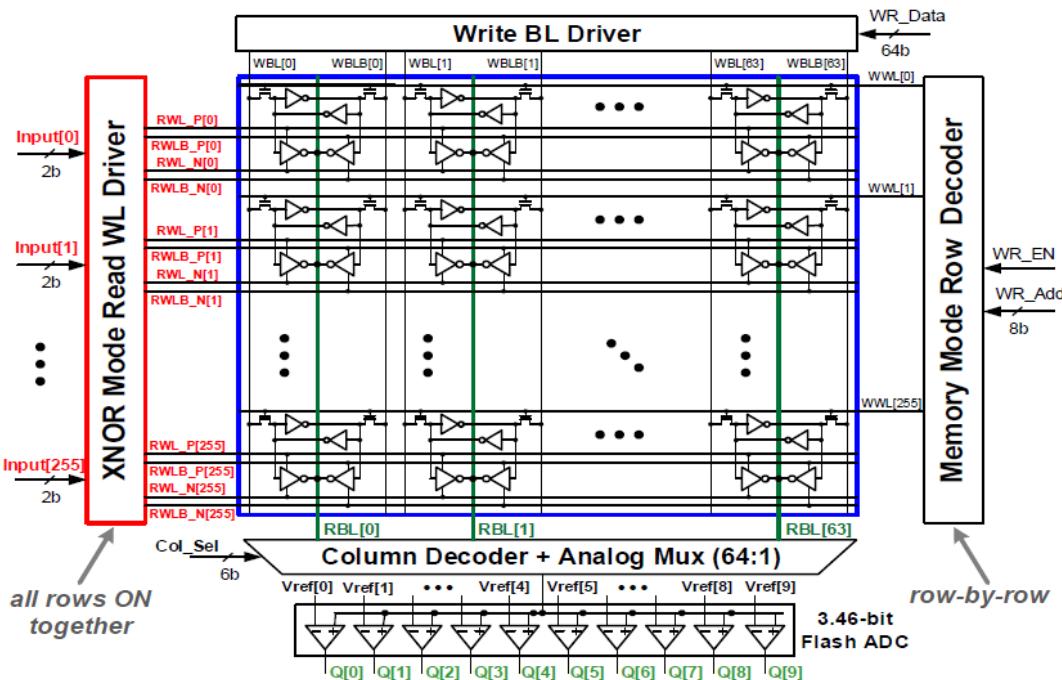
# SRAM based Vector Processing for ML Applications



[J. Zhang, JSSC 2017]

- SRAM-based in-memory computation of ML classifiers
- Key primitive is vector products that can enable a large class of ML applications
- Not all applications require neural networks

# Binary CNN/DNN Computation on SRAM

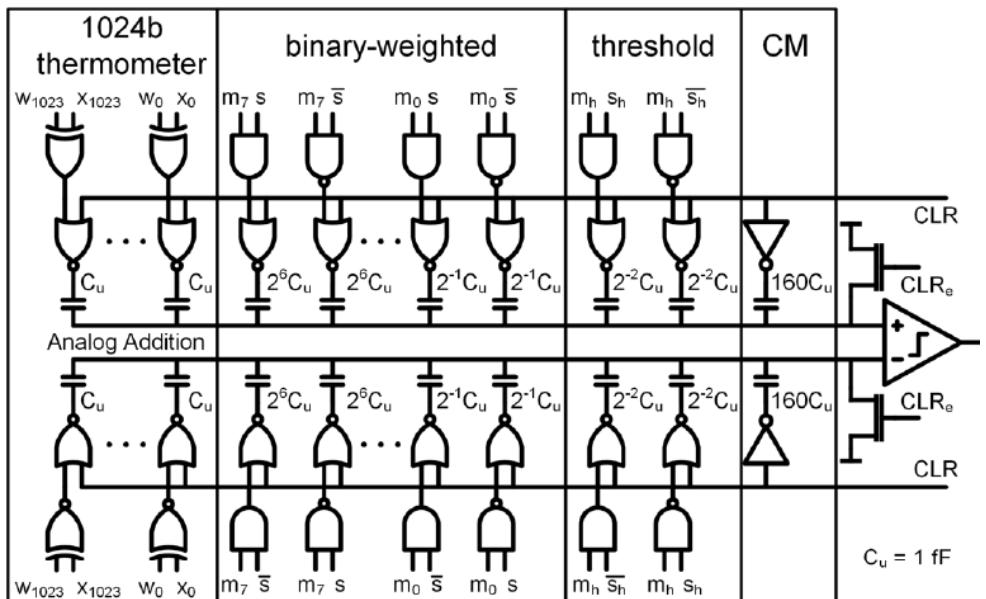
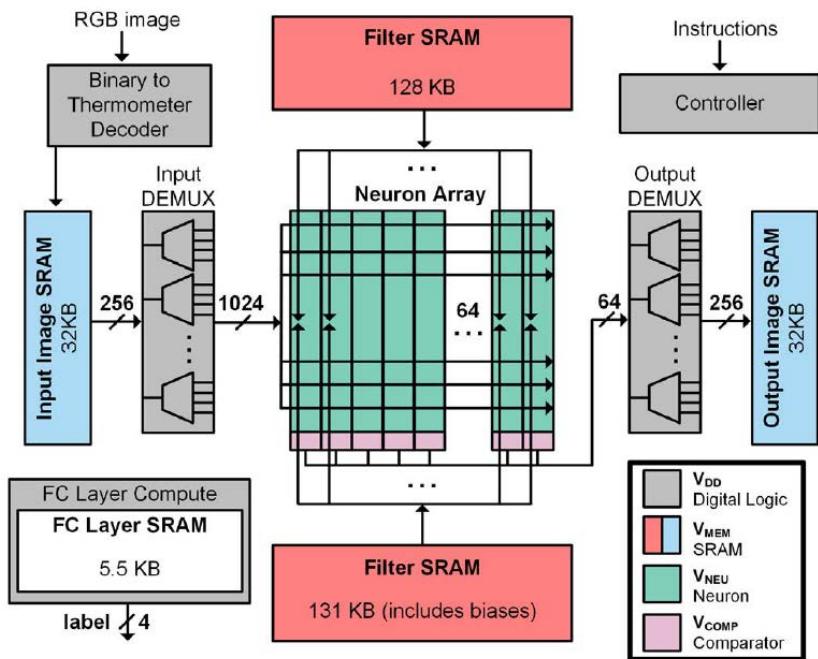


XNOR-SRAM for binary  
DNN/CNN

[Z. Jiang, VLSI 2018]

- BL based processing can enable XNOR operations used in binary neural networks
- Multiple word-lines are simultaneously activated
- Very low BL voltage can cause read disturb

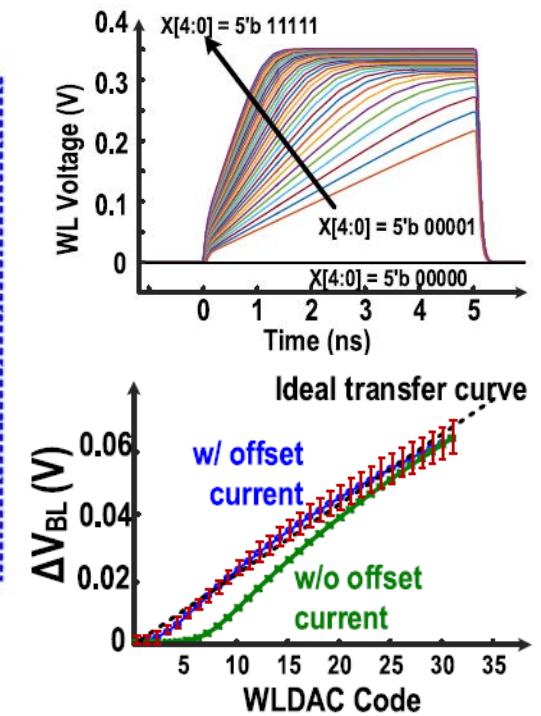
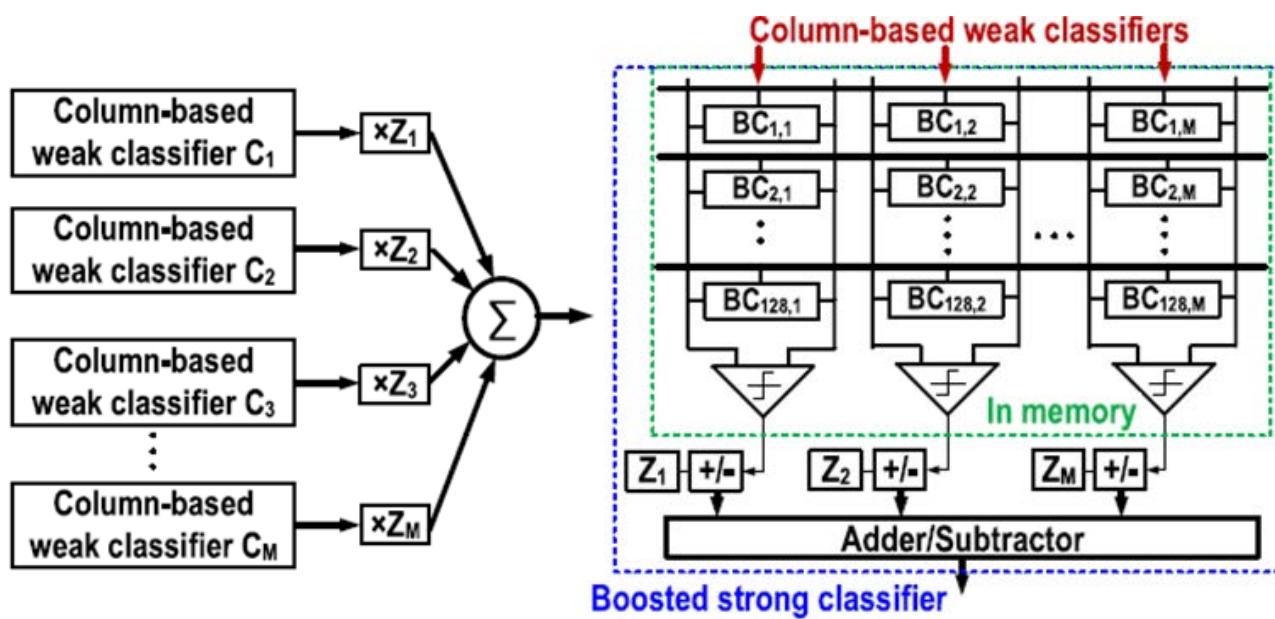
# Binary CNN with Mixed-Signal Primitives



[D. Bankman, ISSCC 2018]

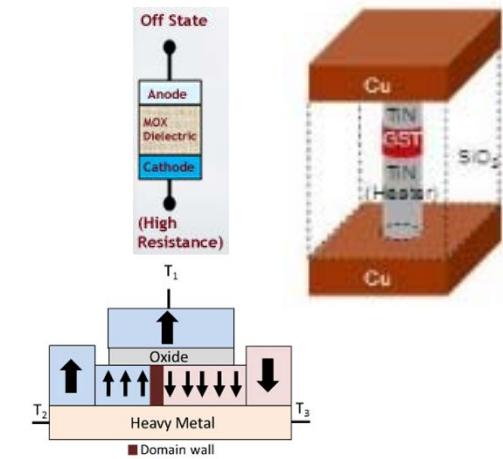
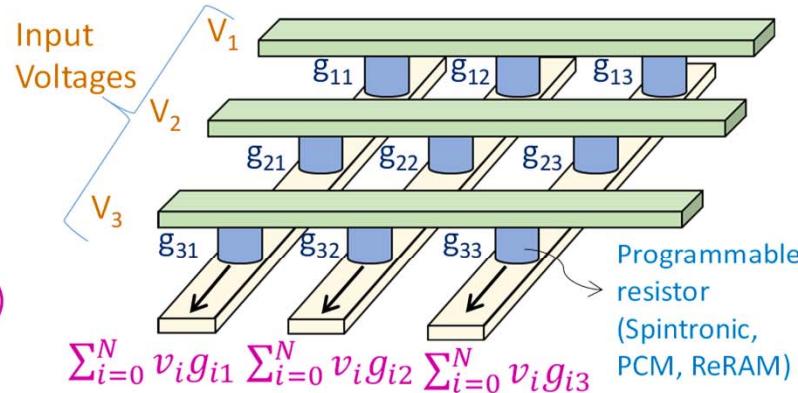
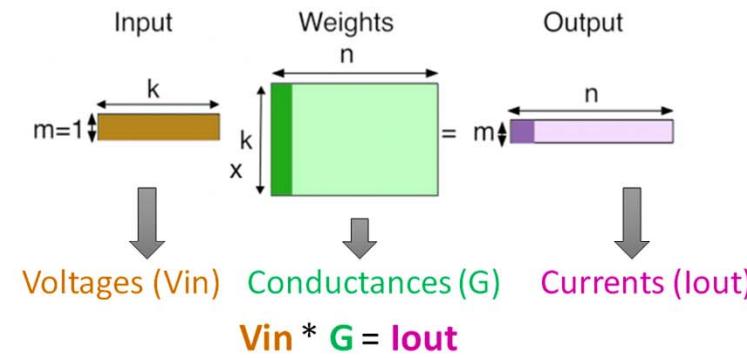
- Mixed-signal binary CNN can enable higher energy-efficiency
- Thermometer coding of data allows easy updates
- Design requires careful design of capacitance matching and off-set cancellation

# Scaling up to Complex ML Classifiers



- Simple column-based classifiers may not be able to solve complex problems
- Weak classifiers can be combined to create more complex and stronger classifiers
- Adaboost is a popular algorithm technique that can enable strong classifiers from linear ones

# Analog Computing on the Bit-line

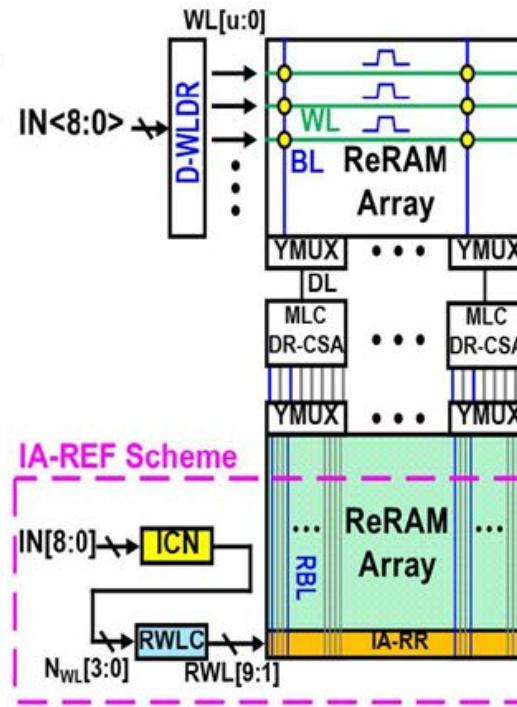
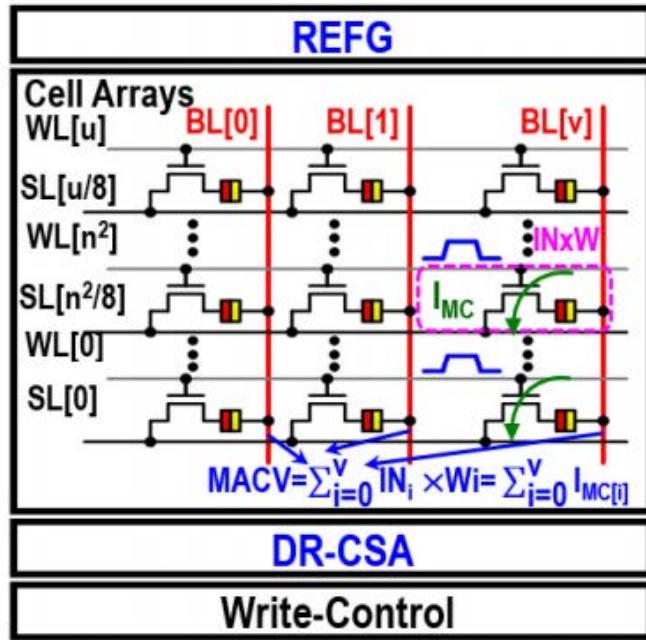


W. Lu et al. NanoLetters'10,  
H.S.P. Wong et al. Proc. IEEE'12  
K. Roy et al. ICCAD'13

Synaptic device Type	Ag:a-Si [S.H. Jo et al.]	TaO <sub>x</sub> /TiO <sub>2</sub> [L Gao et al.]	PCMO [S. Park et al.]	AlO <sub>x</sub> /HfO <sub>2</sub> [J. Woo et al.]	SOT-DWM [A. Sengupta et al.]
# conductive states	97	102	50	40	64
$R_{ON}$	26 MΩ	5 MΩ	23 MΩ	16.9 KΩ	200 KΩ
ON/OFF ratio	12.5	2	6.84	4.43	7

# From SRAMs to eNVM

D-WLDR Time



		nvCIM-P				nvCIM-N			
Input/WL (IN)	Ternary Weight	Weight/MC (W)	Product (INxW)	I <sub>MC</sub>	Weight/MC (W)	Product (INxW)	I <sub>MC</sub>		
0	+1	+1 (LRS)	0	0	0 (HRS)	0	0		
	+1	+1 (LRS)	+1	I <sub>LRS</sub>	0 (HRS)	0	I <sub>HRS</sub>		
0	0	0 (HRS)	0	0	0 (HRS)	0	0		
	0	0 (HRS)	0	I <sub>HRS</sub>	0 (HRS)	0	I <sub>HRS</sub>		
0	-1	0 (HRS)	0	0	-1 (LRS)	0	0		
	-1	0 (HRS)	0	I <sub>HRS</sub>	-1 (LRS)	-1	I <sub>LRS</sub>		

Computing-in-memory RRAM for binary DNN AI edge processor

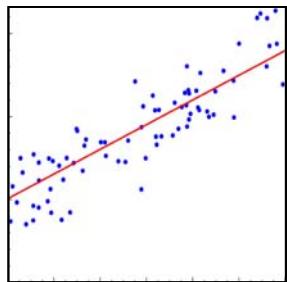
[W. Chen, ISSCC 2018]

- Resistive eNVM provides a natural technology platform for in-memory vector processing
- Current research is geared towards single cycle vector processing with parallel WL firing
- Key challenges lie in the design of peripheral circuits

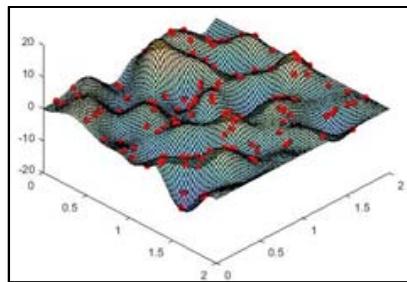
# Outline

- Merged Memory and Logic
- Emerging Memory Technologies and Outlook
- **Towards Data-Centric Near/In-Memory Systems**
  - Data-flow Architectures for ML
  - **Compute-Communicate-Iterate for Optimizations**
  - Real-Time Learning Systems
- Outlook

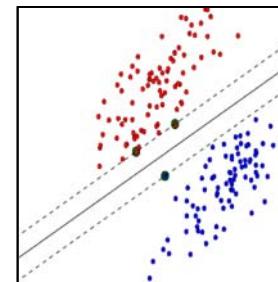
# Optimization ... Everywhere



## Regression



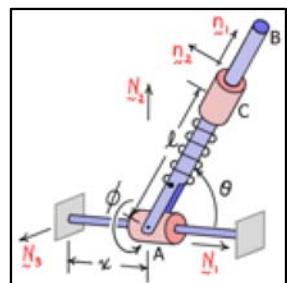
# Image Processing



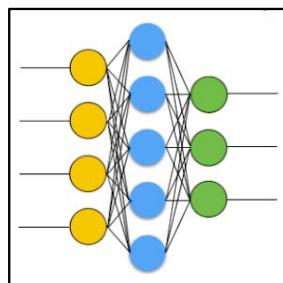
## Classification

# Constrained Optimization

## General Formulation



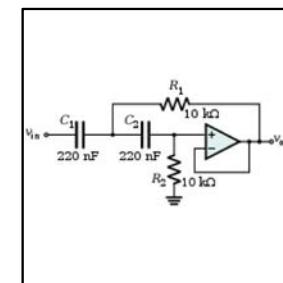
## Mechanics



## Model training



## Economics



## Circuits

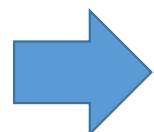
*minimize*  $f(x)$   
*subject to*  $Ax = b$

- Optimizations on large data-sets is a quintessential problem in data analytics
  - Require complex operations on large data with multiple iterative accesses
  - Require large bandwidth to data as well a programmable on-chip data-movement

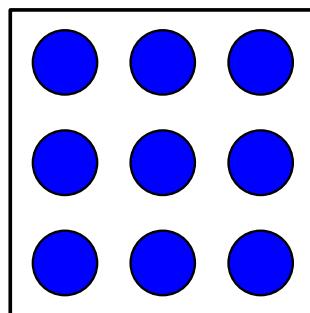
# Near Memory Architectures:

## Alternative Direction Method of Multipliers (ADMM)

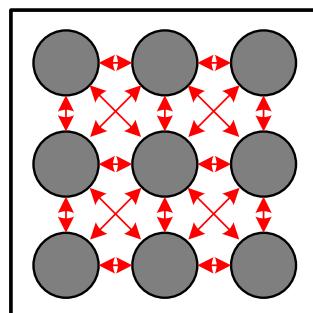
minimize  $f(\mathbf{x}) + g(\mathbf{z})$   
subject to  $A\mathbf{x} + B\mathbf{z} = \mathbf{c}$



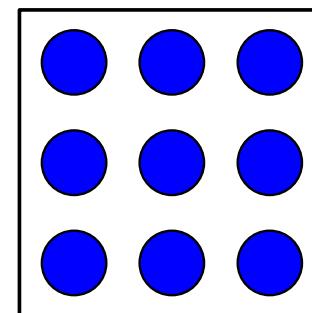
$$x_i^{k+1} := \operatorname{argmin}_{x_i} \left( f(x_i) + \frac{\rho}{2} \|x_i - z^k + u_i^k\|_2^2 \right)$$
$$z^{k+1} := \operatorname{argmin}_z \left( g(z) + \frac{N\rho}{2} \|z - \bar{x}^{k+1} - \bar{u}^k\|_2^2 \right)$$
$$u_i^{k+1} := u_i^k + x_i^{k+1} - z^{k+1}$$



Update  $x$



Update  $z$

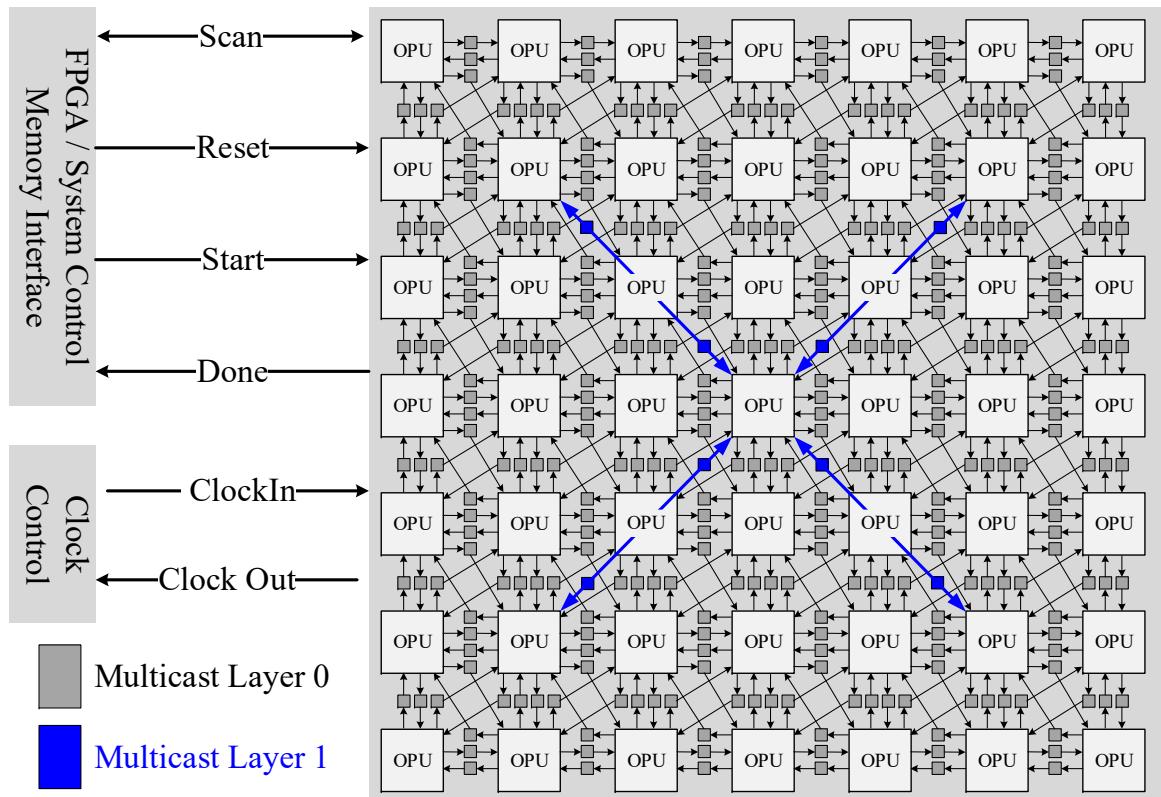


Update  $u$

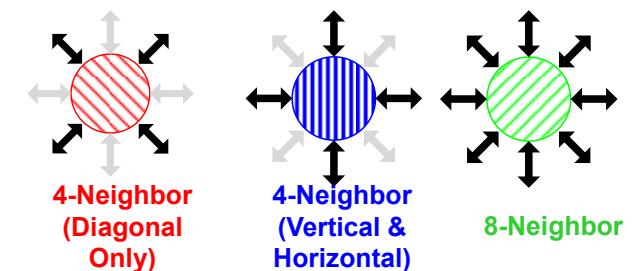
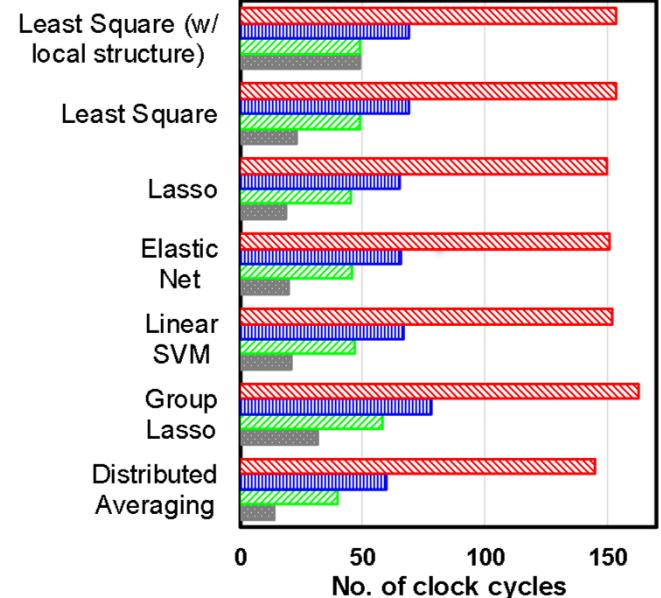


- One optimal way to solve optimizations on distributed data uses a combination of local processing and iterative communication

# Demonstration Vehicle: SRAM on CMOS Platform

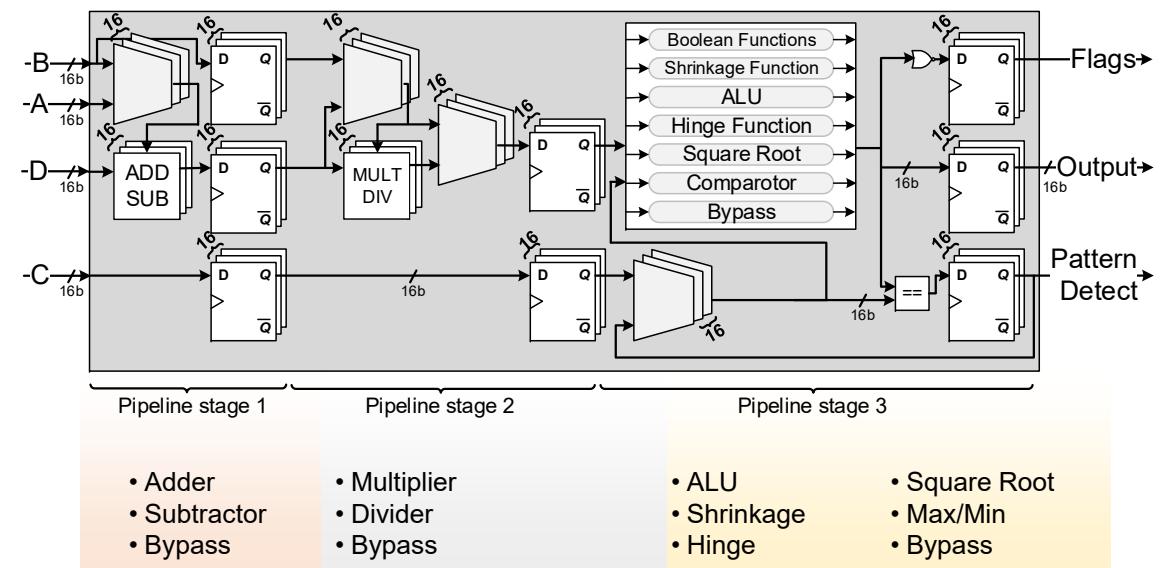
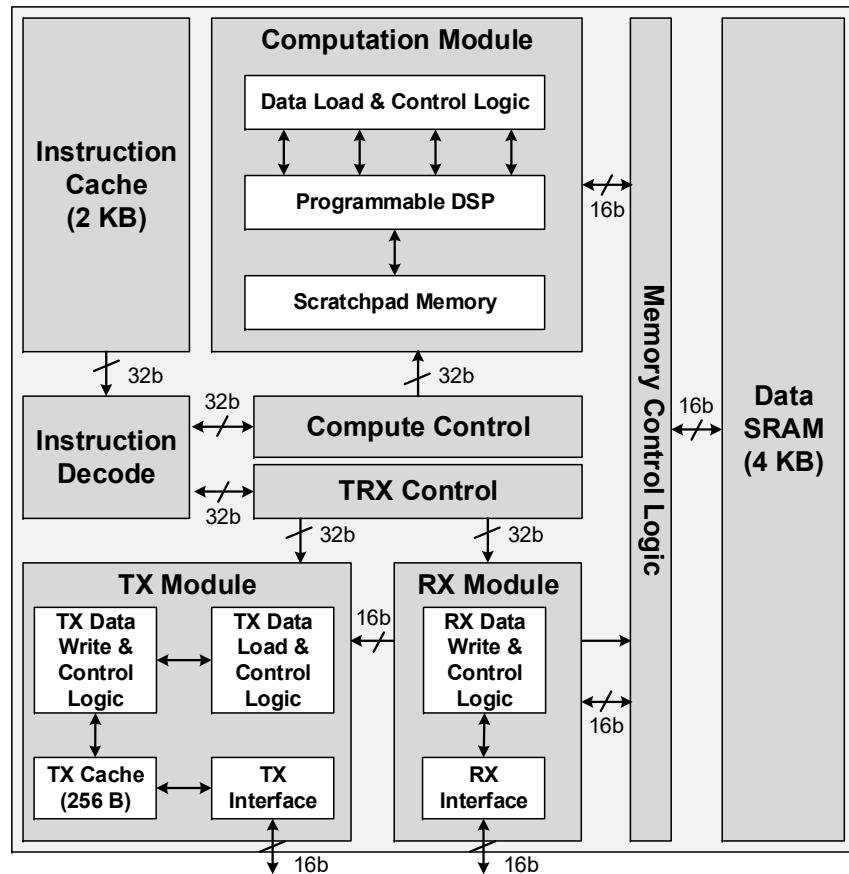


■ 4-Neighbor (Diagonal Only)  
■ 4-Neighbor (Vertical & Horizontal)  
■ 8-Neighbor  
■ 8-Neighbor + Hierarchical Multicast Network (This work)



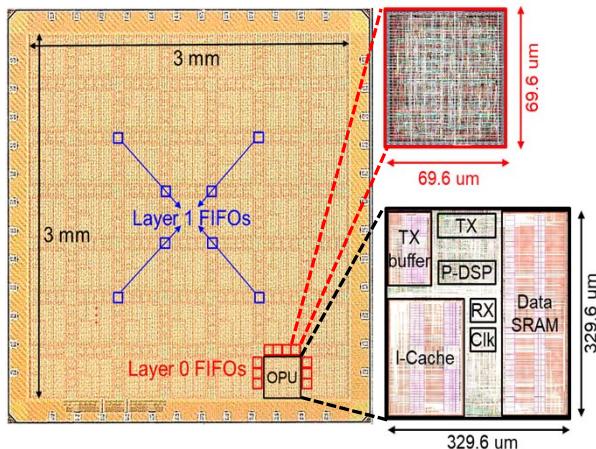
- Algorithm relies on local near-memory computation and iterative communication

# Near Memory (SRAM) Computing

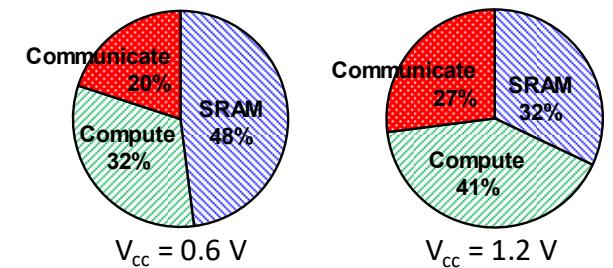
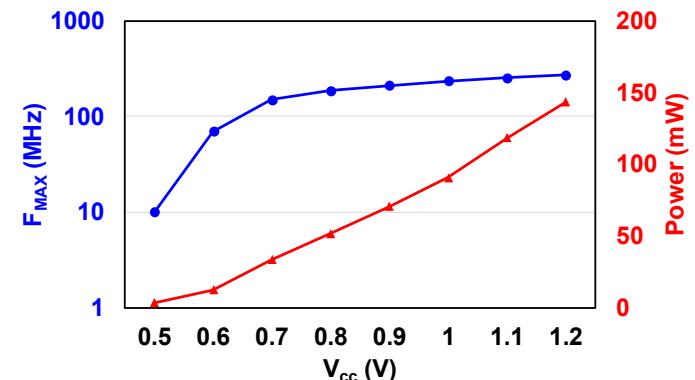


- For harder problems beyond ML, complex near-memory computing are needed
- This requires support for more linear algebraic kernels – other than matrix multiplications

# Test-chip Demonstration

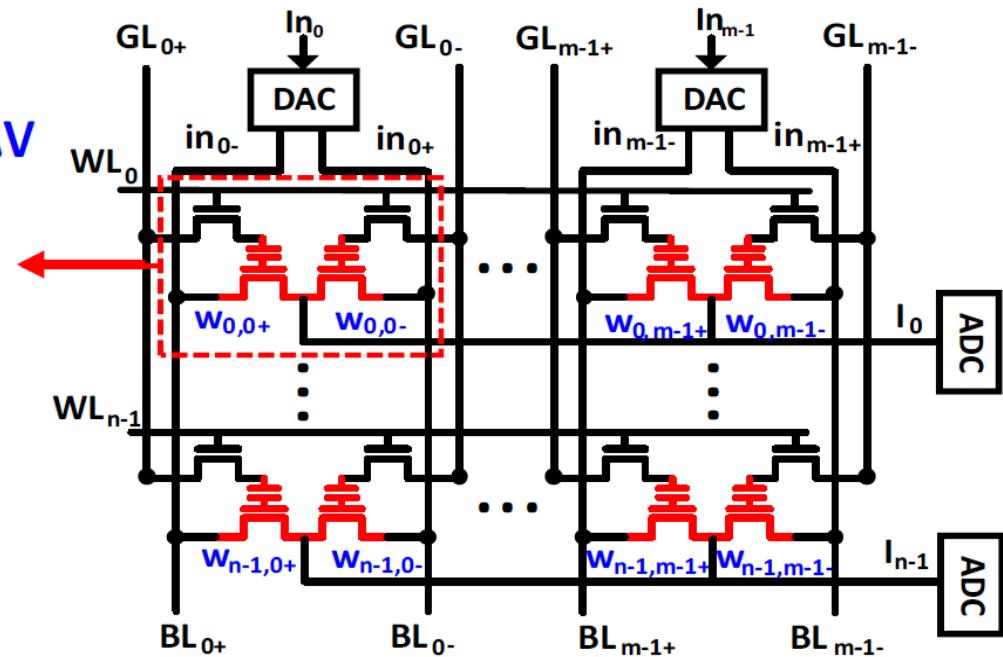
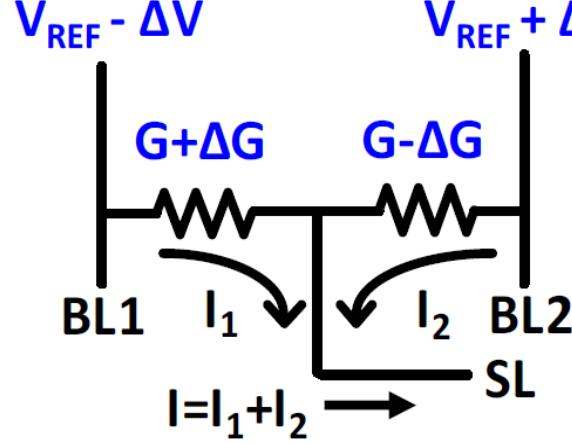


<b>Technology</b>	TSMC 65nm GP 1P9M
<b>Chip Size</b>	3.41 mm x 3.41 mm
<b>Core Area</b>	3 mm x 3 mm
<b>Package</b>	QFN6x6-48
<b>Pin Count</b>	48
<b>Gate Count (logic only)</b>	2725 kGates (NAND2)
<b>On-Chip SRAM</b>	306.25 KB
<b>Number of OPUs</b>	49
<b>No. of pipeline stages in P-DSP</b>	3
<b>Core / IO Supply Voltage</b>	0.5-1.2 V / 2.5 V
<b>Clock Rate</b>	10-270 MHz
<b>Network</b>	Asynchronous & Mesochronous
<b>Peak Energy Efficiency</b>	279 GOPS/W
<b>Arithmetic Precision</b>	16-bit fixed-point



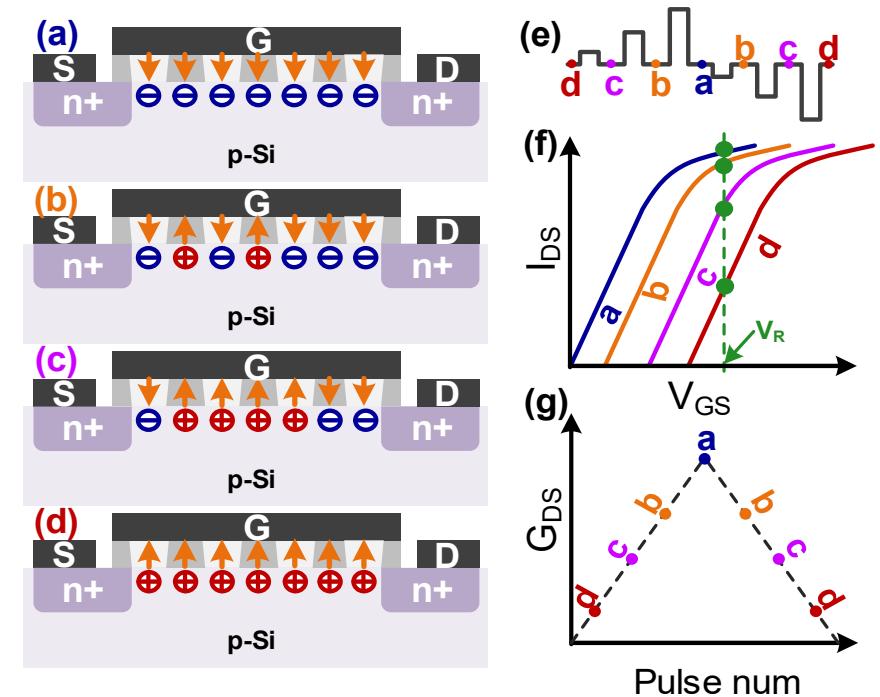
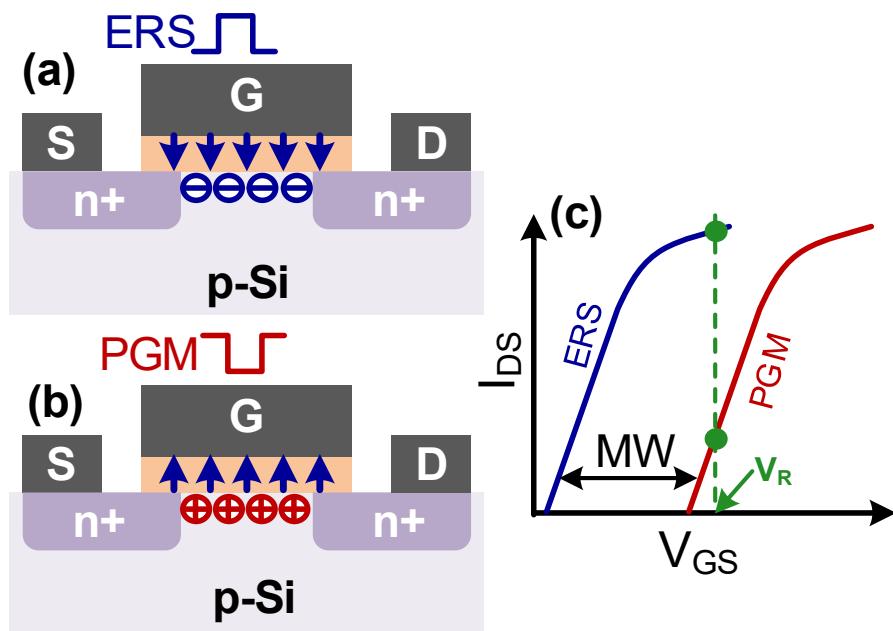
- Characterizing power-performance trade-off and measured energy-efficiency

# System Architecture for Positive and Negative Operands



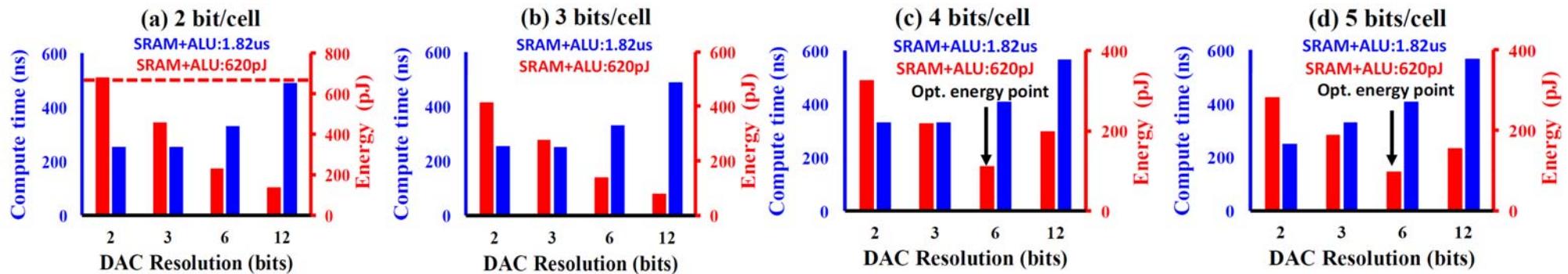
- Multiple devices are needed to emulate positive and negative operands
- Peripheral circuits consume energy and latency – needs to be comprehensively studied

# Ferroelectric FET as an Embedded NVM



- Ferroelectric FETs have promising characteristics and multi-level embedded storage.
- The transistor gain plays an important role in distinguishability of states.

# Solving ADMM on FeFET Arrays



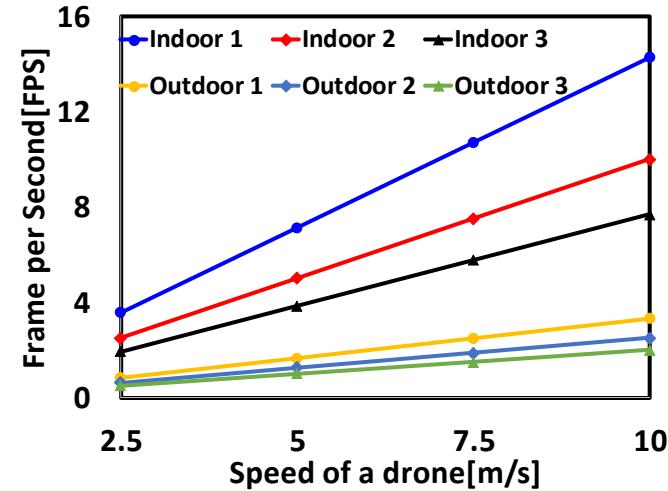
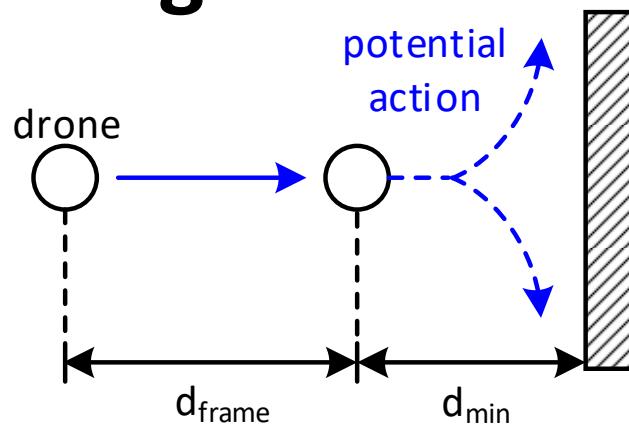
- The figure shows the design-space exploration of ADC, DAC bit resolution and number of bits per FeFET cell in terms of latency and energy dissipation.
- At a system level, this results in 21X (3X) improvement in energy-efficiency (performance) compared to an SRAM+ALU architecture

	Baseline	SRAM PIM	FEFET PIM	PIM Ratios
Compute – Delay (us)	83	1.8	0.6	3X
Energy (uJ)	1360	400	21	19X
EDP (uJ x us)	112880	720	12	60X

# Outline

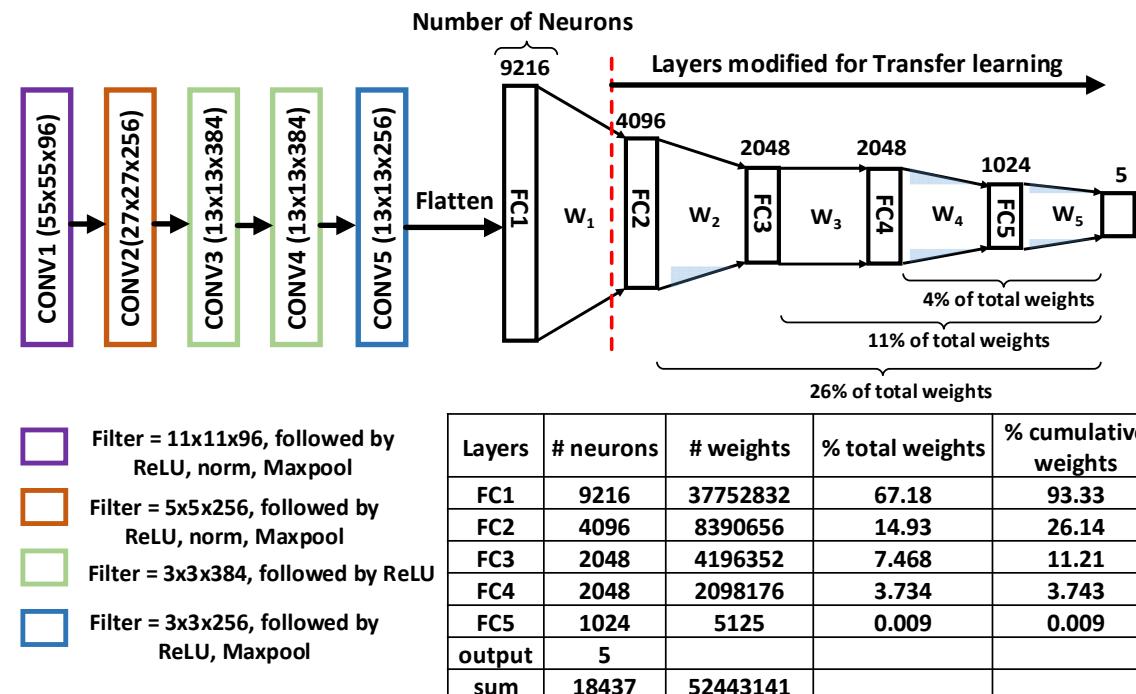
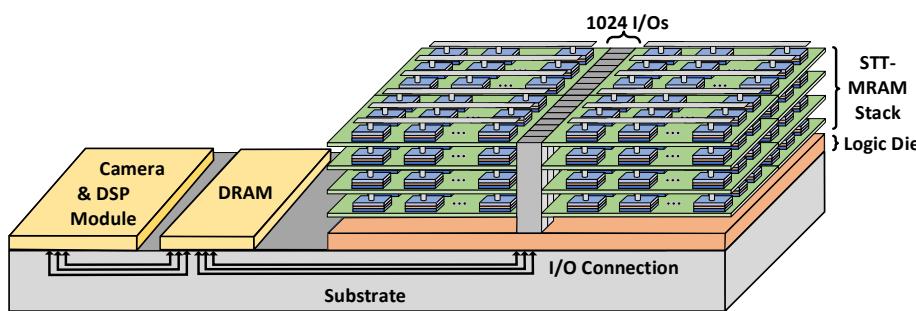
- Merged Memory and Logic
- Emerging Memory Technologies and Outlook
- **Towards Data-Centric Near/In-Memory Systems**
  - Data-flow Architectures for ML
  - Compute-Communicate-Iterate for Optimizations
  - **Real-Time Learning Systems**
- Outlook

# Edge Robotics: Autonomy via Reinforcement Learning



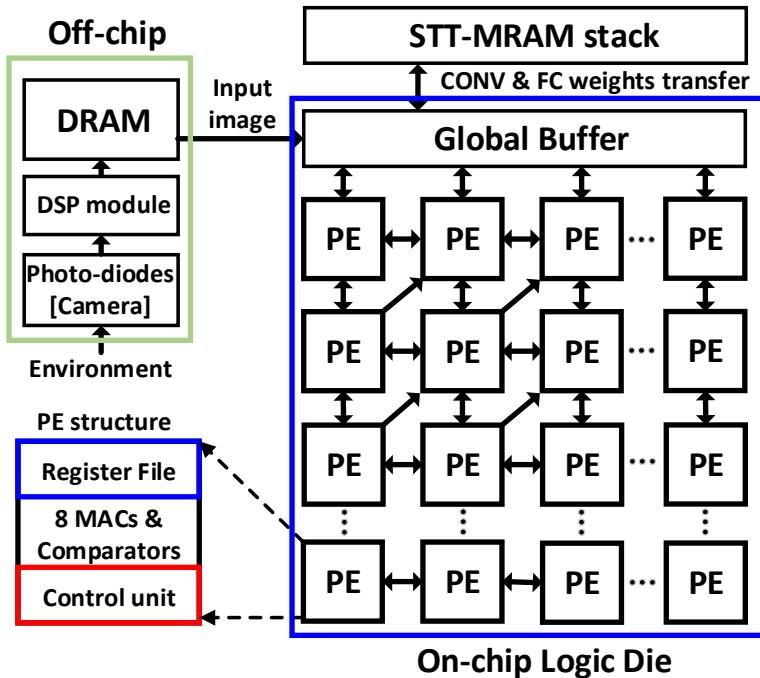
- High performance autonomous agents such as drones need to learn via interactions with the environment.
- This is possible via real-time reinforcement learning where the model weights need to be updated in real-time
- Write speeds of most of the dense memory technologies are insufficient for real-time reinforcement learning

# Template Problem: Autonomous Navigation



- Convolution Neural Network based controller for end-to-end reinforcement learning
- Real time updates with eNVM will not allow the agent to fly at reasonable speeds
- Solution: Hierarchical reinforcement learning to match the memory hierarchy

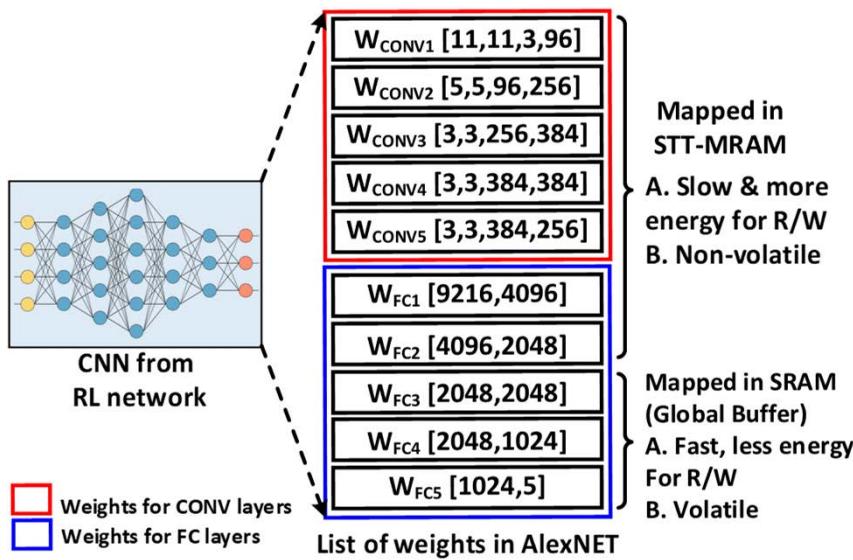
# System Architecture



Technology	NanGate 15nm FreePDK
Number of PEs	1024 (32 row, 32 column)
Global buffer /scratchpad	30MB/4.2MB
Register File per PE	4.5KB
Operation voltage	0.8V
Clock speed	1Ghz
Peak Throughput	1.5TOPS/W
Arithmetic precision	16 bit fixed-point
Bandwidth between PEs	128 bit

- Systolic array of processing element with row-stationary algorithm
- Large Global buffer for storing:
  - Weights used for forward propagation
  - Sum of gradients of weights and biases for last 3 layers of CNN

# Architectural Mapping from System to Stacked STT-MRAM



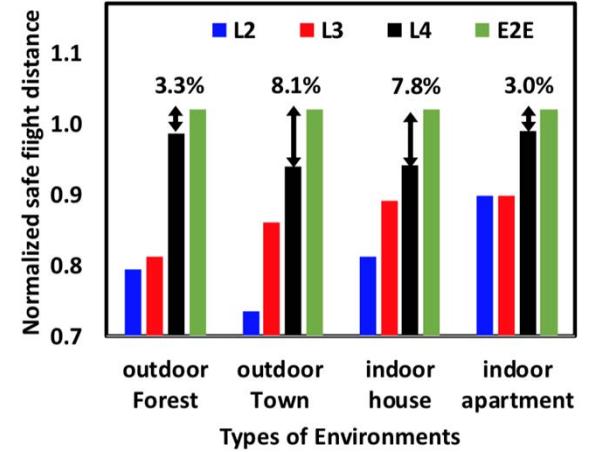
- For cases of training last 2/3 layers, we don't access STT-MRAM
- For a case of training last 4 layers,(FC2, FC3, FC4, FC5), we access STT-MRAM for training of FC2 layer
- For training all layers, we access STT-MRAM for every iteration

# Demonstration of Transfer Learning with Online Reinforcement Learning

- Rich set of virtual worlds include indoor and outdoor environments
- End-to-end infrastructure from Unreal Engine to TF

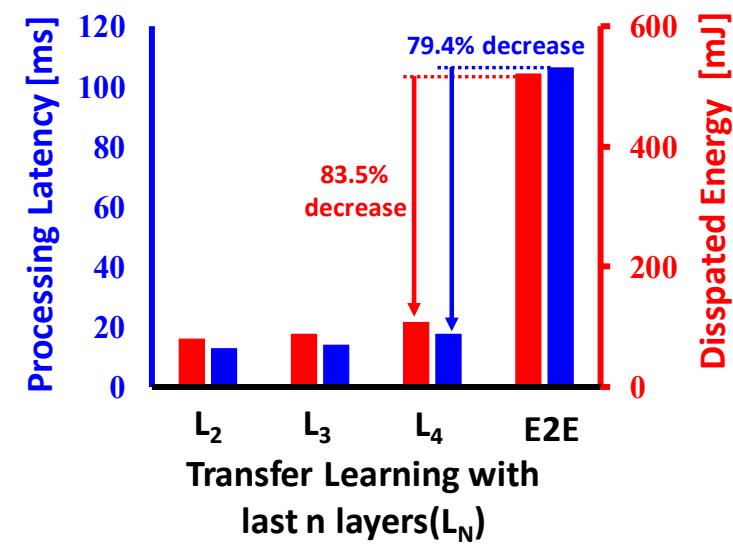
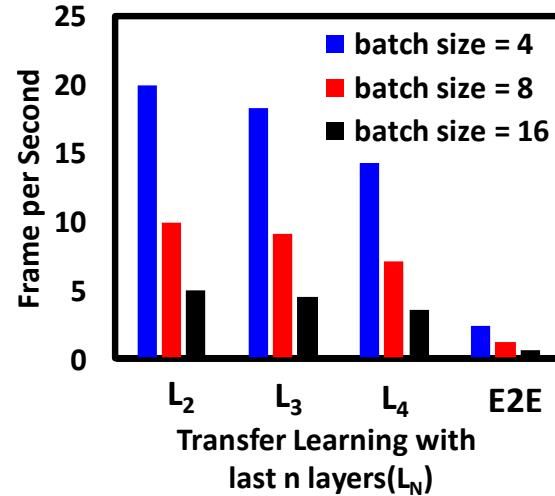
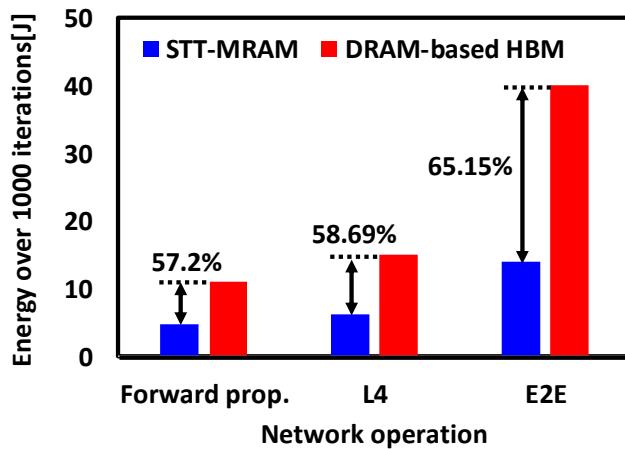
Transfer Learning

- Emulation of STT-RAM + SRAM system with online RL
- Real flight in real, unseen environment
- 3X improvement in drone speed



- A hierarchical memory system can be used for hierarchical RL
- Loss of performance is acceptable compared to E2E learning

# System Results and Outlook



- With batch size of 4, we can support 15~20 FPS with last 2/3/4 layer training
- 79.4% & 83.45% decrease in latency and energy compared to latency and energy of End-to-End learning case.
- eNVM is a significant driver for low overall power but real-time systems need to be designed keeping the write latency/energy in mind

# Outlook

- The demand for larger memory capacity will continue to grow
- Data-centric compute are not only for AI but also for a large class of problems in data-analysis
- Technology scaling exacerbates the memory bottleneck
- Near-memory and in-memory architectures can alleviate some of the problems of the memory bottleneck
- There are design and technology challenges with near-memory architectures that need to be addressed up-front
- A close collaboration between hardware and algorithm design is needed to the next generation of energy-efficient computing for large-scale data-analysis

# Acknowledgements

Students and post-docs:

- Muya Chang
- Insik Yoon
- Aqeel Anwar Malik
- Yan Fang



Collaborators:

- Titash Rakshit (Samsung)
- Rajiv Joshi (IBM)
- Suman Datta (Univ of Notre Dame)
- Shimeng Yu (GT)
- Asif Khan (GT)
- Vivek De (Intel)
- Jim Tschanz (Intel)
- Justin Romberg (GT)



# 3D NAND Challenges and Potentials

Jian Chen

Western Digital Corp.

# Forward-looking Statements

## Safe Harbor | Disclaimers

This presentation contains forward-looking statements that involve risks and uncertainties, including, but not limited to, statements regarding our product and technology portfolio, market positioning, business strategy and growth opportunities, market trends, and data growth and its drivers. Forward-looking statements should not be read as a guarantee of future performance or results, and will not necessarily be accurate indications of the times at, or by, which such performance or results will be achieved, if at all. Forward-looking statements are subject to risks and uncertainties that could cause actual performance or results to differ materially from those expressed in or suggested by the forward-looking statements.

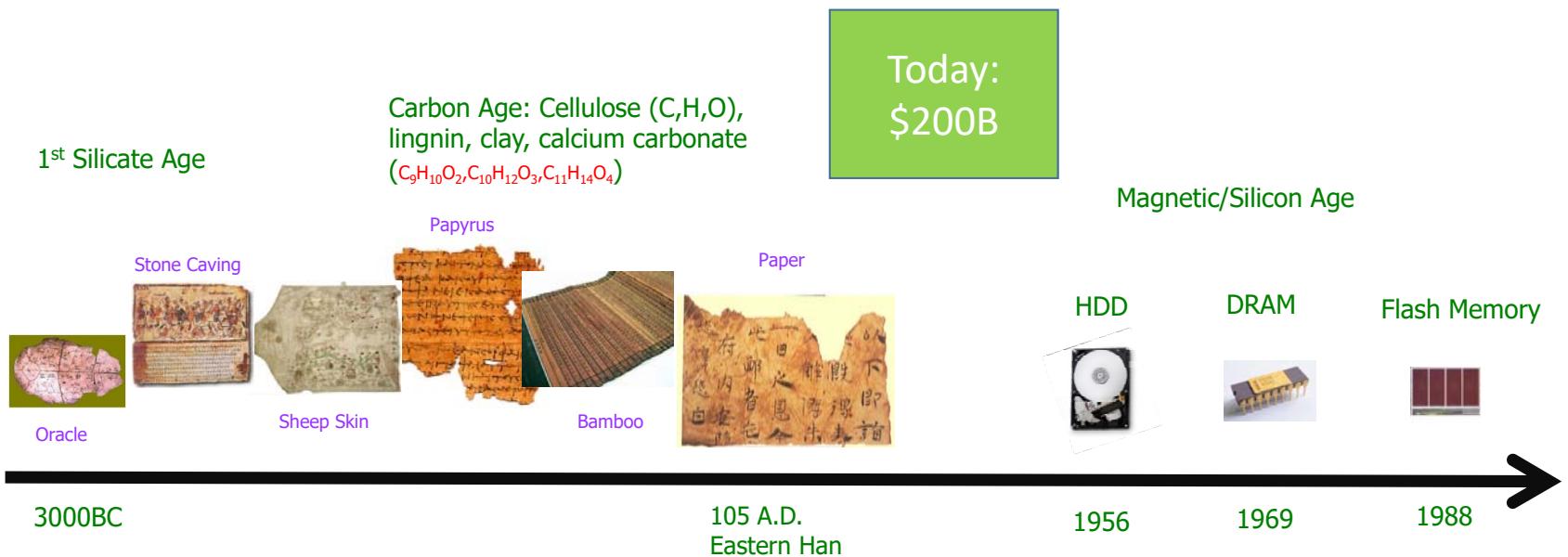
Key risks and uncertainties include, among others: volatility in global economic conditions; actions by competitors; our development and introduction of products based on new technologies and expansion into new data storage markets; unexpected advances in competing technologies; business conditions; growth in our markets; and pricing trends and fluctuations in average selling prices. More information about the risks and uncertainties that could affect our business are listed in our filings with the Securities and Exchange Commission (the “SEC”) and available on the SEC’s website at [www.sec.gov](http://www.sec.gov), including our most recently filed periodic report, to which your attention is directed. We do not undertake any obligation to publicly update or revise any forward-looking statement, whether as a result of new information, future developments or otherwise, except as required by law.

[This presentation contains financial measures defined as non-GAAP. The non-GAAP measures are used by the company’s management to forecast, evaluate and review the financial results of the company. Management believes these non-GAAP financial measures are useful because they provide meaningful comparisons to prior periods and exclude certain items that may not be indicative of the underlying performance of the company’s business. These non-GAAP financial measures should be used in addition to, and in conjunction with, results presented in accordance with GAAP to better understand the company’s financial performance. Non-GAAP measures are not in accordance with, or an alternative for, measures prepared in accordance with GAAP and may be different from non-GAAP measures used by other companies.]<sup>1</sup>

# **Outline**

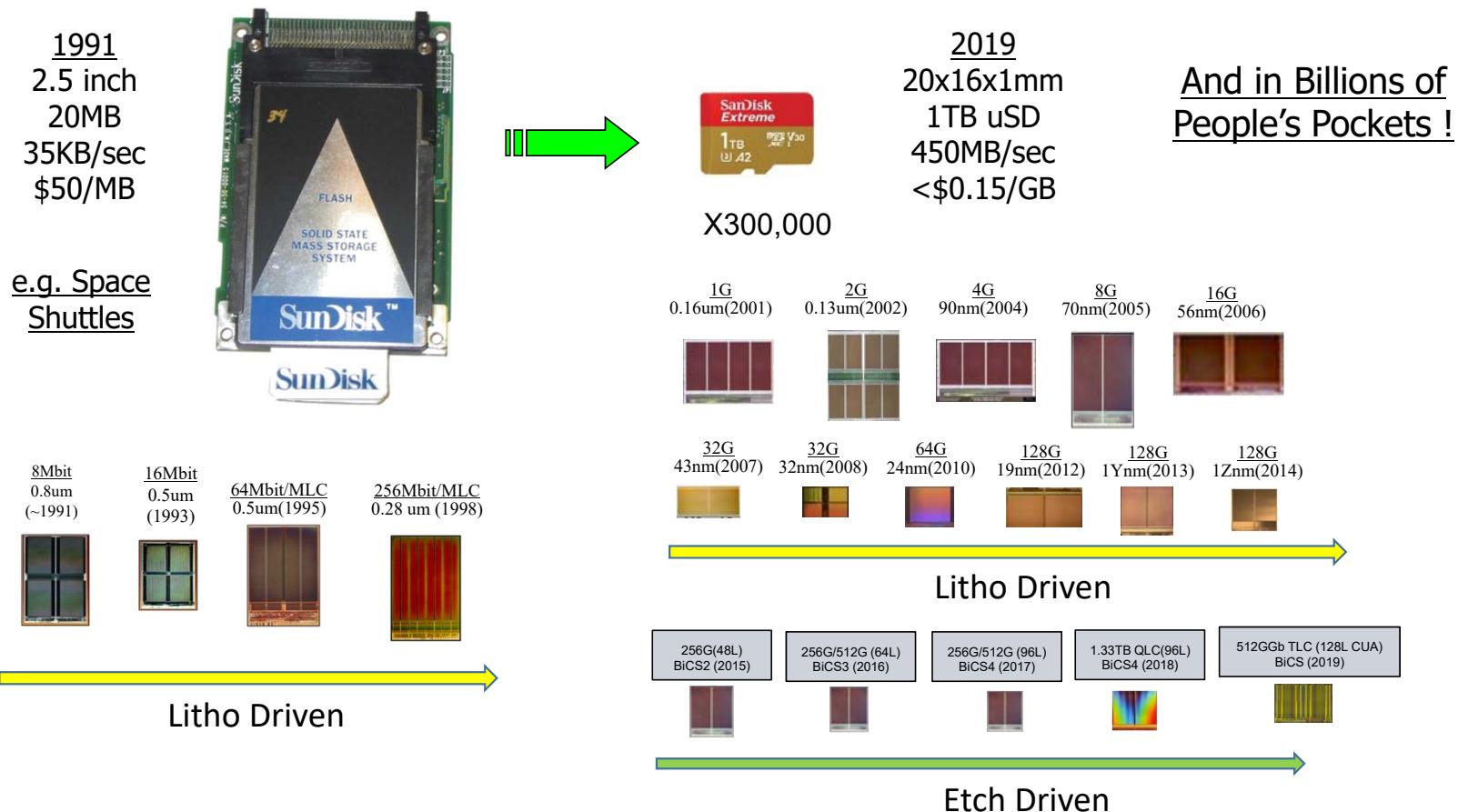
- 1. Introduction to 3D NAND**
- 2. Scaling Challenges and Opportunities**
  - Cost Scaling and CapEx
  - Architecture Options
  - Device/Process Scaling and Challenges
- 3. Versatile NAND and System Solution Opportunities**
- 4. Conclusions**

# Global Storage Technology Roadmap

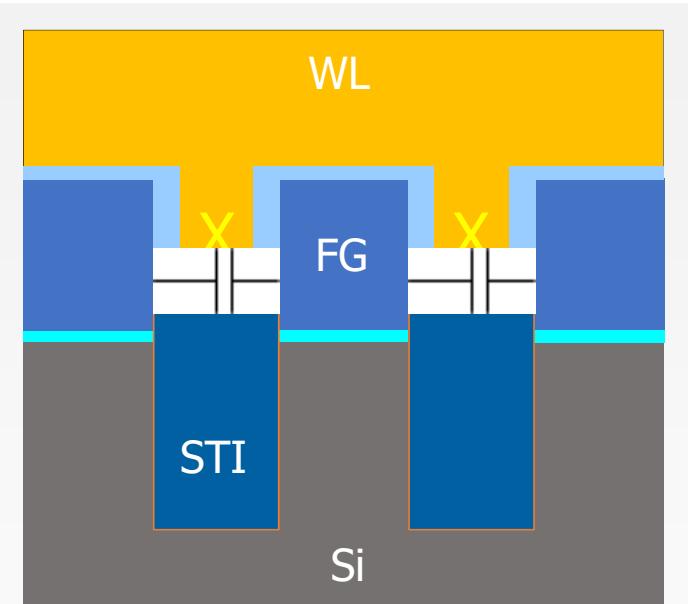


Data and storage are forever

# The Last 31 Years ...



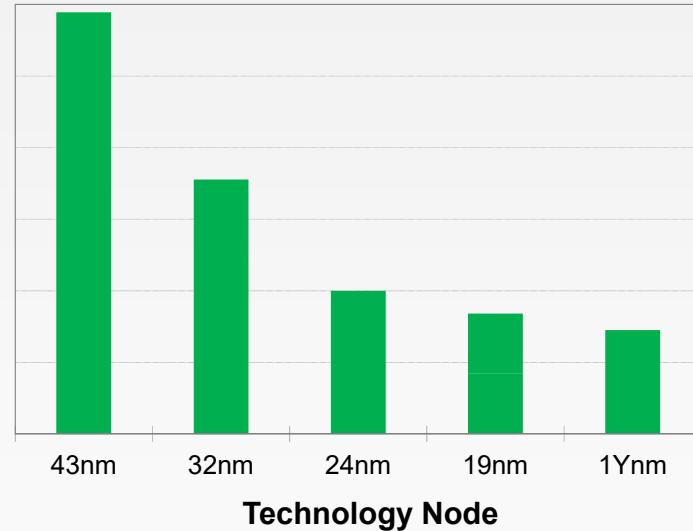
# NAND Device Scaling Challenges



## Physical limitations

- Cell to cell coupling
- IPD (Inter-Poly Dielectric) thickness

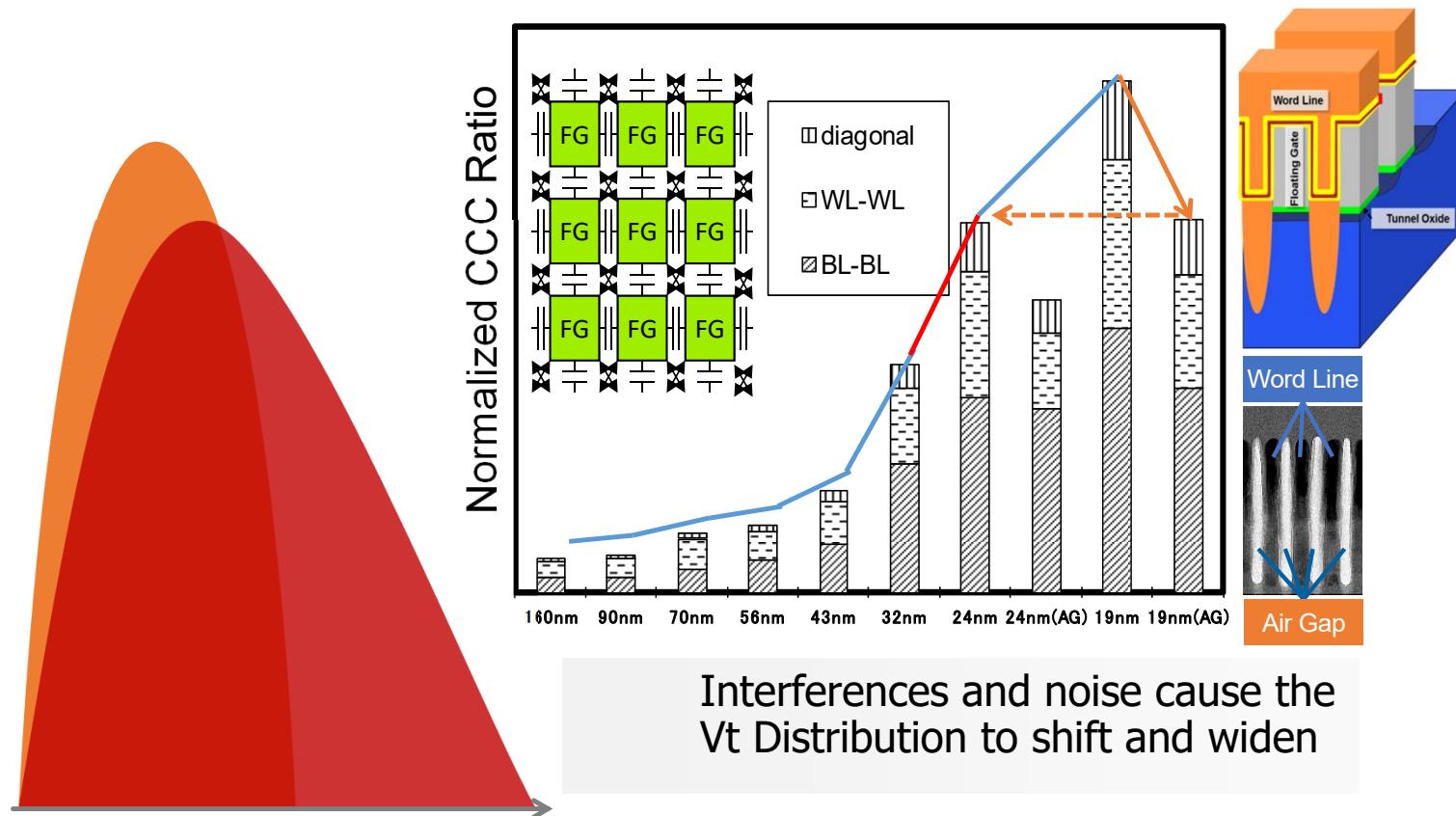
Reduction in # of Electrons With Scaling



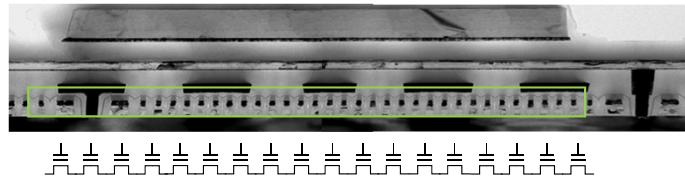
## Electrical limitations

- Reduction in # of electrons in the cell
- The cell Vt shift by 1 electron

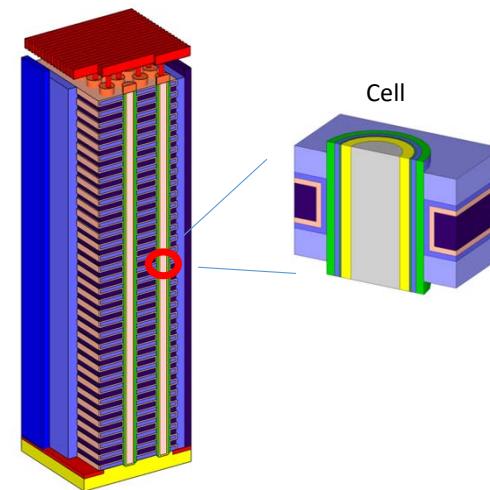
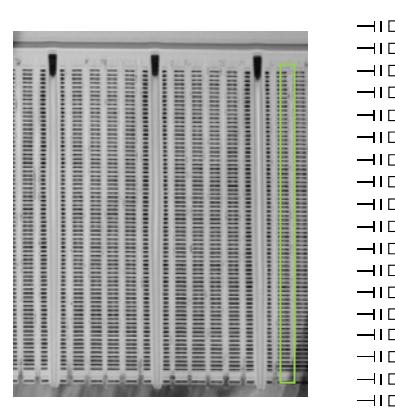
# The Effect of Cell to Cell Interferences



# 2D NAND, 3D NAND and Scaling



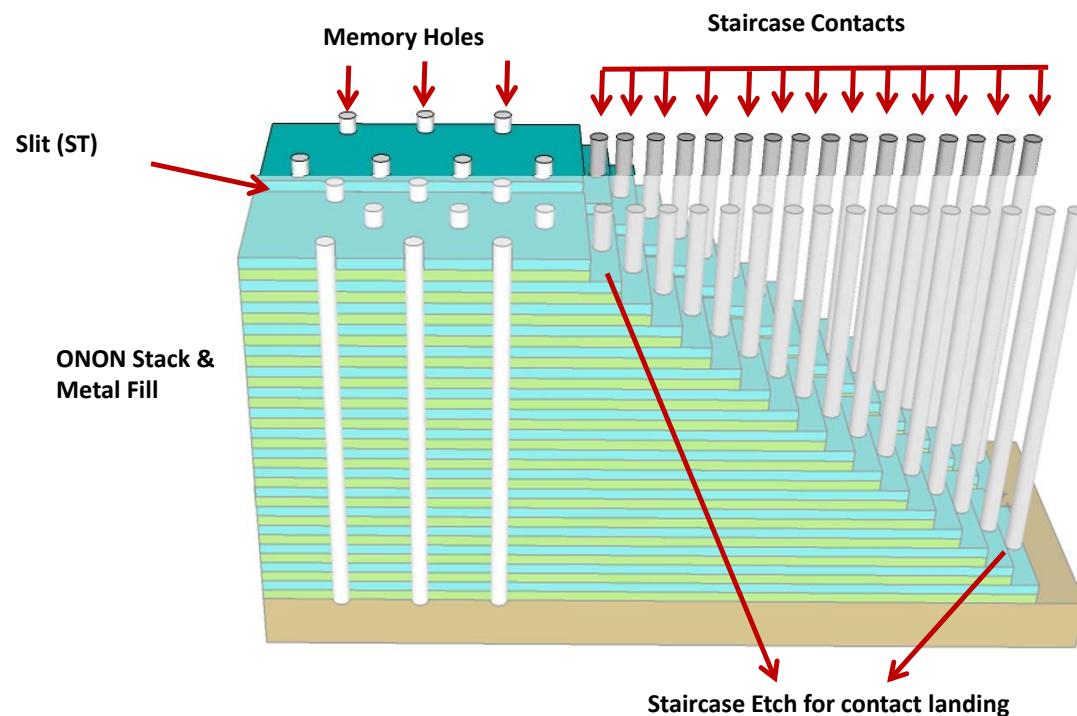
←2D NAND:  
X-Y Scaling (Moore's Law)  
Logical-scaling (2bits, 3bits/cell)



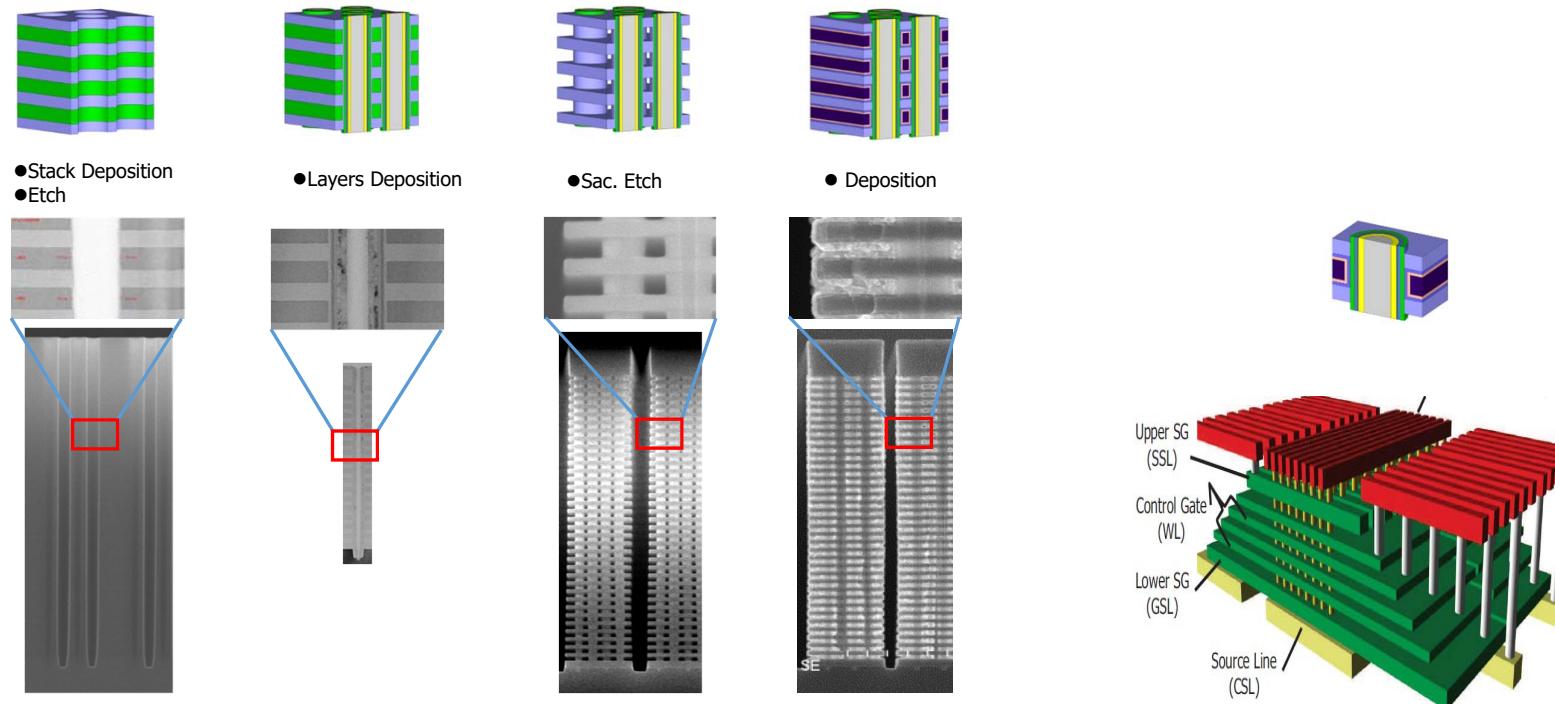
←3D NAND  
Z-scaling (up)

XY → XY+L → XYL+Z → XYLZ+M(Material)+S(System)

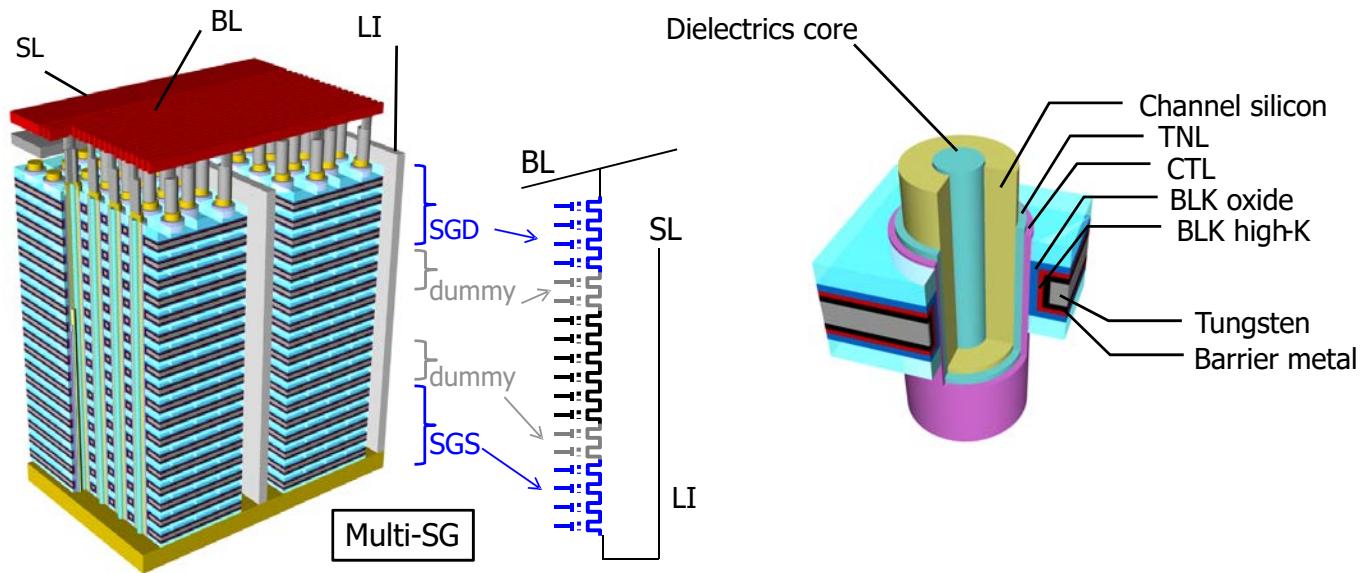
# Basic 3D NAND Schematic



# 3D NAND Critical Process Steps (Punch & Plug)

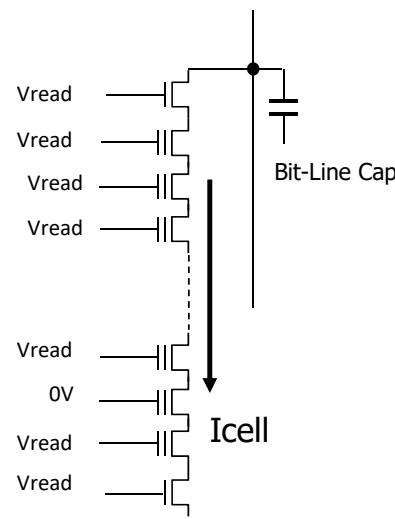


# 3D NAND Architecture



- Cylinder cell, shielding bit line to bit line coupling
- 3D block consists of multiple 2D blocks
- Multiple strings connected to 1BL (Only 1 string to 1BL in 2D NAND)
- Ever larger block sizes ( 32KB in 0.16um 1Gbit in 2000, >18MB in 96L in 2018)

# Basic NAND Read/Sense Operation



Read Sensing: Bit-line pre-charge,  
then discharge by  $I_{cell}$ .

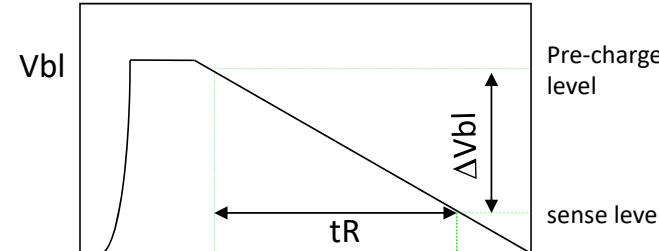
$$I_{cell} (Vt_0, Vt_1, \dots, Vt_{15}, V_{read}) > C_b * \Delta V_{bl} / t_R$$

$C_b$ : bit-line cap

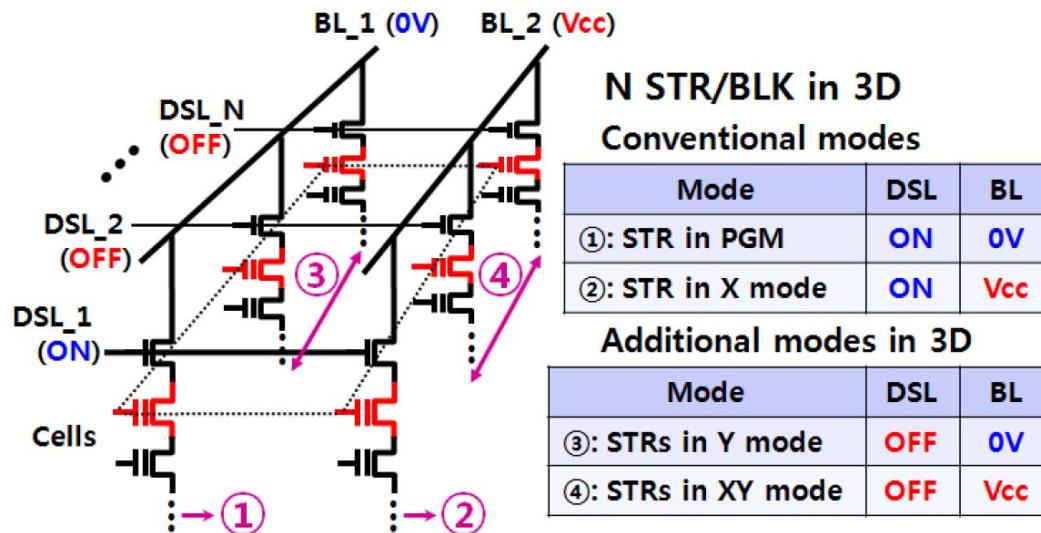
$\Delta V_{bl}$ : bit-line swing

$t_R$ : Time for read

For "1" cell,  $I_{cell} > \sim 50\text{nA}$  (e.g.)



# 3D NAND Program Disturb



- Three program disturb modes (X, Y, XY)

# Outline

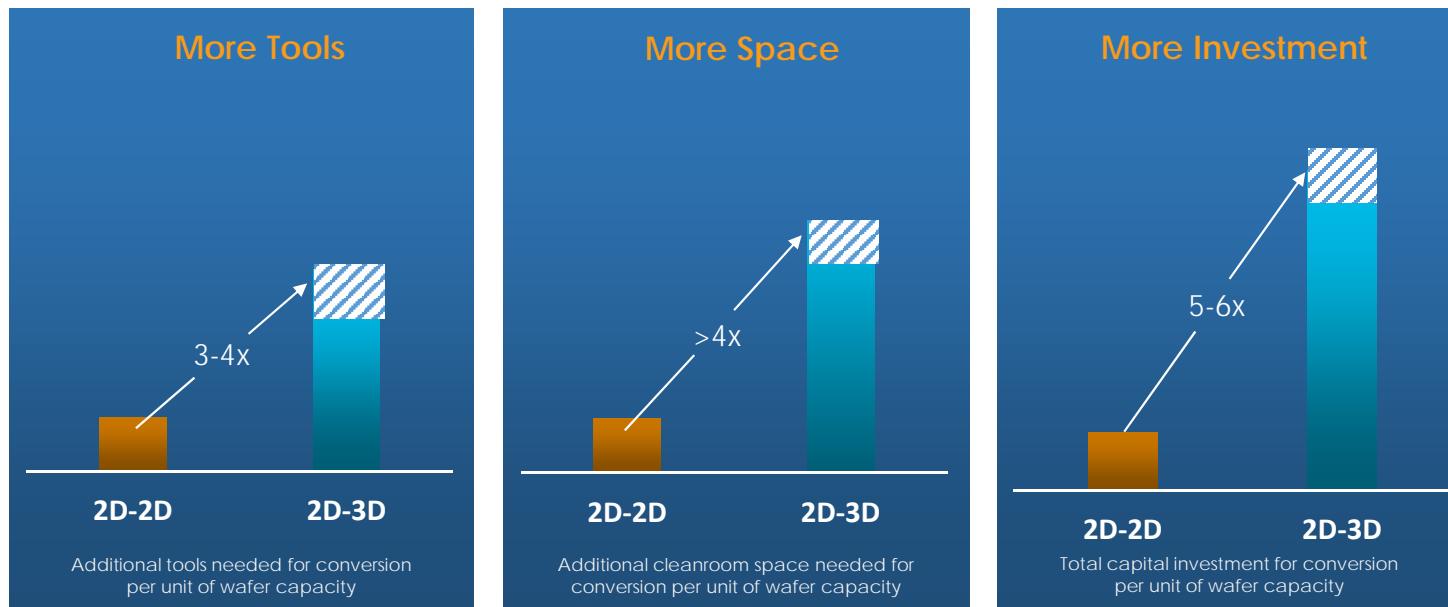
1. Introduction to 3D NAND
2. Scaling Challenges and Opportunities
  - Cost Scaling and CapEx
  - Architecture Options
  - Device/Process Scaling and Challenges
3. Versatile NAND and System Solution Opportunities
4. Conclusions

# Scaling: It's About COST, COST, COST

Many inter-connected factors affect the cost in different ways.  
Heaven/hell for total engineering/product optimization.

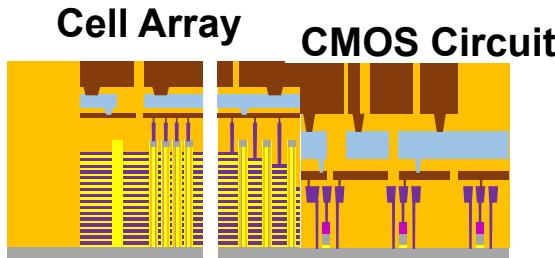
- *CapEx (Depreciation), Tools Availability, Indirect Material (IDM), Director Material*
  - *Efficiency in converting from existing tools*
  - *Speed of ramp and yield bring-up and qualification*
  - *Die Size (different architectures, # of tiers, placement of CMOS)*
  - *Process complexity*
  - *Yields, Reliability*
  - *# of data worldline Layers*
  - *ON (or OP) pitch, total memory hole stack height*
  - *Array density (pack how many memory hole/mm<sup>2</sup>)*
  - *Read and write performance, IO, block size, multiple plane random read perf.*
  - *How many logical bits (SLC/MLC/TLC/QLC/PLC...)*
- ... Many different optimizations. No one size fits all.  
It's not just # of layers, or die size.

# CapEx: 2D→3D and 3D→3D NAND Conversions Are More Complex than 2D→2D Transitions



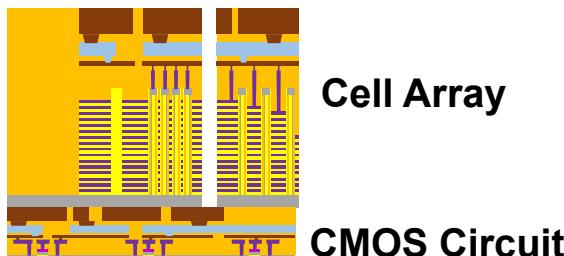
- CapEx is the most significant portion of the cost.
- Longer process flow means larger investments from node-to-node compared to 2D.
- More space and time needed to convert, more technology cost reduction pressure.

# Architecture Choices: Where to place the CMOS



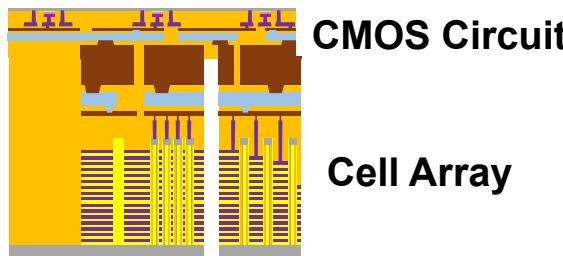
CMOS Next to Array  
aka: next door parking

Lowest process/CapEx cost.  
Simple, traditional.  
But larger die area, and less parallelism in program.



CMOS Under the Array  
aka: underground parking

Small die size.  
But more complex process and larger CapEx.  
More parallelism in program.



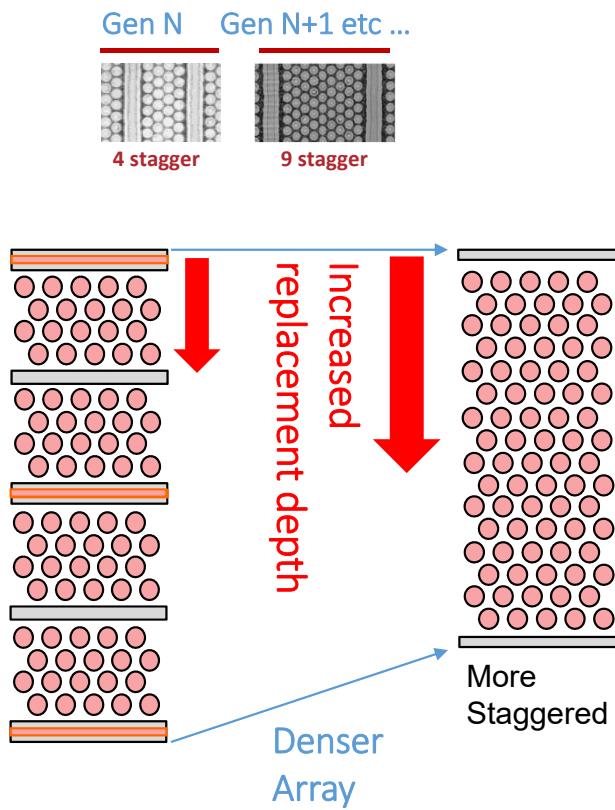
CMOS Above the Array  
aka: rooftop parking

Small die size.  
But more process steps, larger CapEx and extra wafer cost.  
More parallelism in program.  
Can be shorter cycle time with more tools.

# Some Scaling Strategies

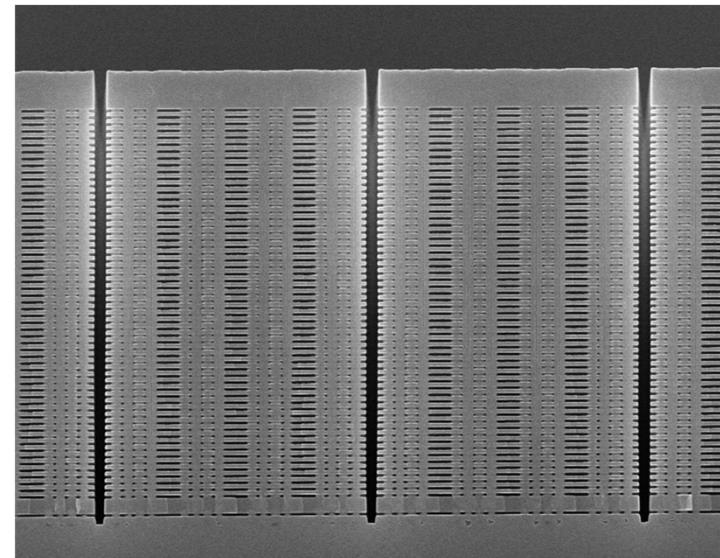
- XY Scaling: Smaller MH pitch
- XY Scaling: Divide the cell
- XY Scaling: Larger blocks
- XY Scaling: Better staircase design
- Z Scaling: More layers
- Z Scaling: Reduce each layer thickness
- Z Scaling: Stack more tiers
- Z Scaling: Improve the channel
- Z Scaling: Improve WL RC
- Z Scaling: Placement of CMOS circuits
- Z Scaling: Simpler process, contact merges
- L Scaling: TLC-QLC-PLC...
- Tool throughputs
- ...XY Scaling: Denser array

# Denser Array: Replacement Depth

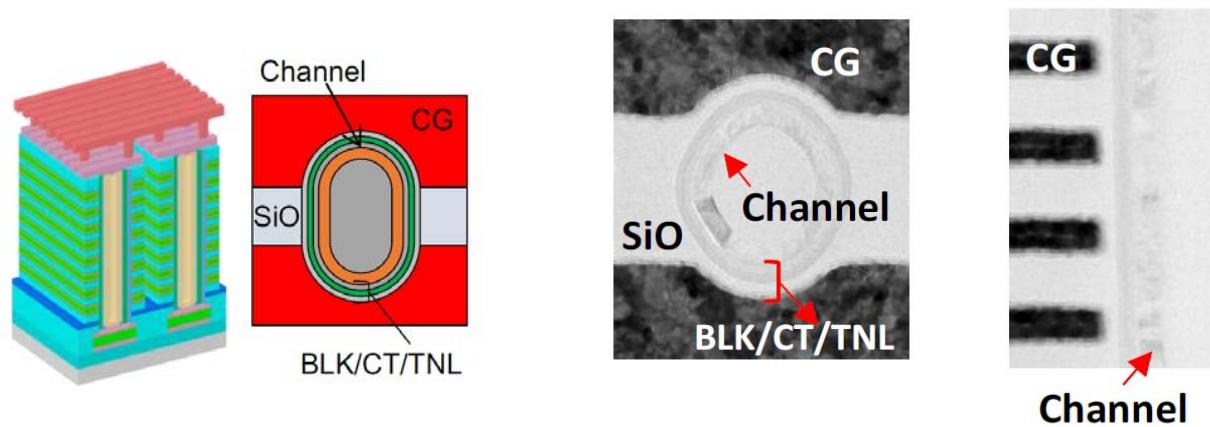


Increased replacement depth requires:

- selective Nitride strip
- improved metal fill
- low resistive metal

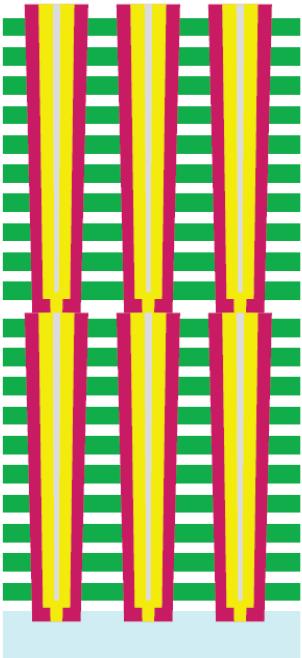


# Device Scaling: Split Gate Semicircular Cell



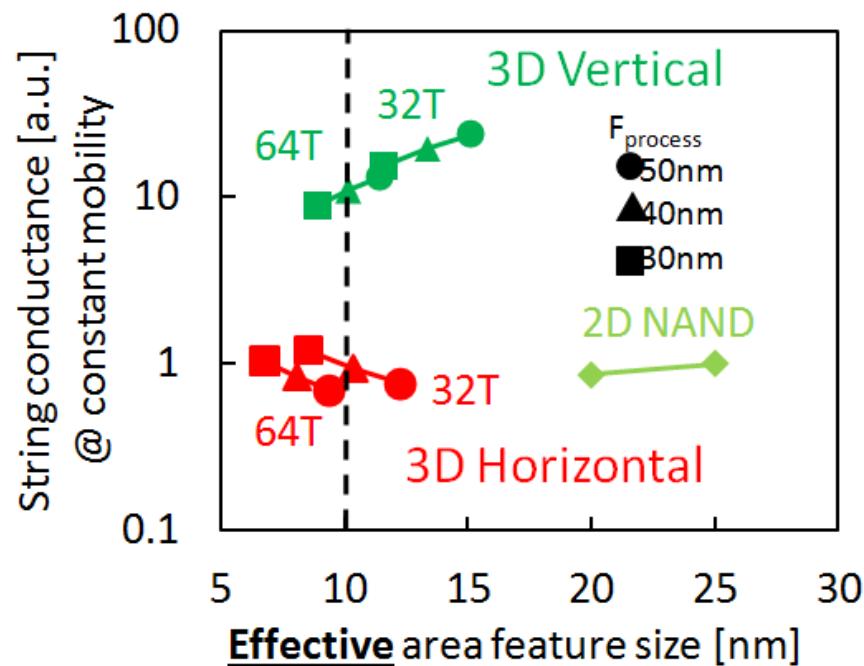
By splitting the control gate, traditional memory cell is divided into two, thus increasing memory density.

# Multi-Tiers Stacking



- Allows more layers with certain memory hole etch capability
- Allows smaller memory hole pitch and denser array
- But more process steps needed
- Alignment of two tiers can be challenging
- Already used by multiple companies

# Channel mobility requirement for 3D NAND



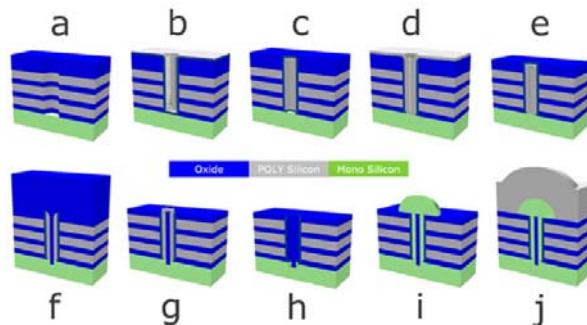
3D NAND	Mobility required (relative to crystalline Si)
Vertical	$\sim 1/10x$
Horizontal	$\sim 1x$ (no degradation)

A. Goda and K. Parat, IEDM2012

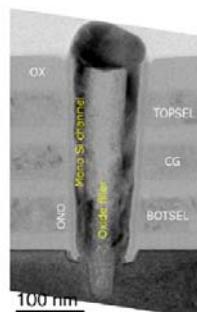
- Poly Si channel used in 3D NAND has poor mobility concern.
- 3D NAND has  $\sim 10x$  device advantage for cell current due to the wide channel width.
- Channel mobility has to be engineered to maximize this benefit.

A. Goda, 2014 IMW 3D NAND

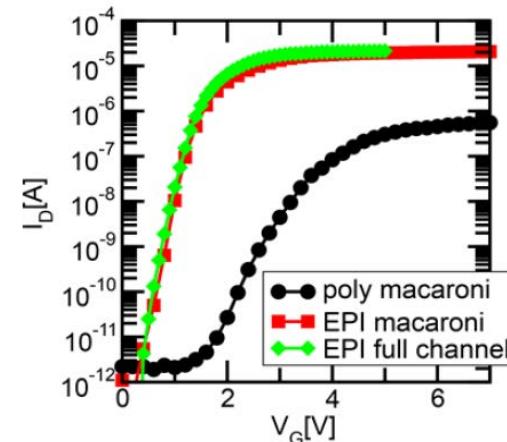
# Monocrystalline Si Macaroni Memory Channel



Replacement channel process scheme.



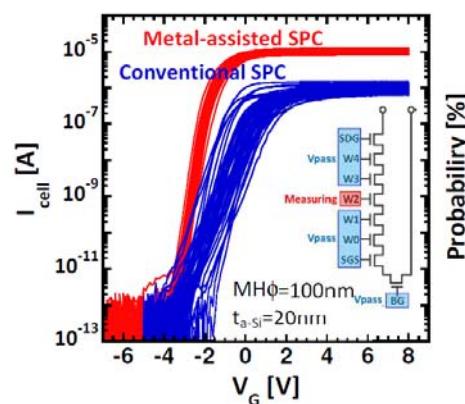
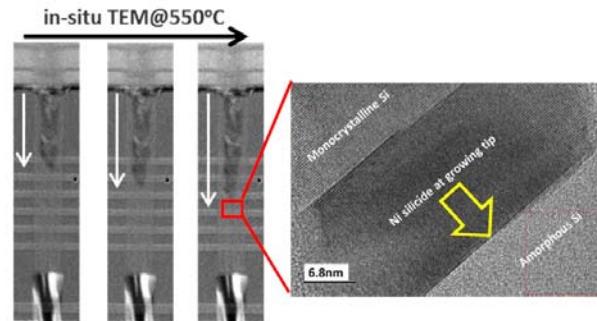
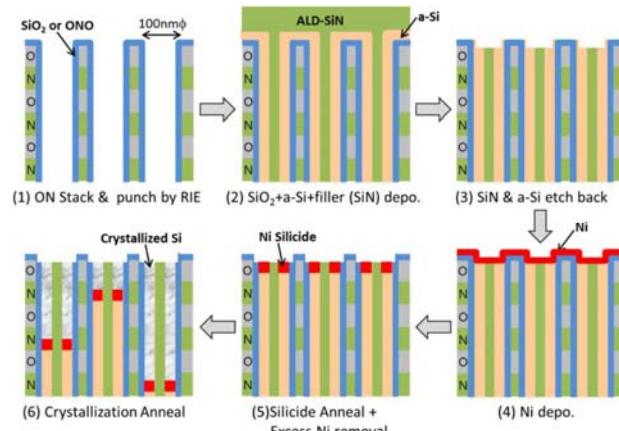
TEM Cross section after Si channel epi growth showing high quality Bottom interface.



R. Delhougne, 2018 Symposium on VLSI Technology

# Memory Channel Engineering: Metal Assisted SPC

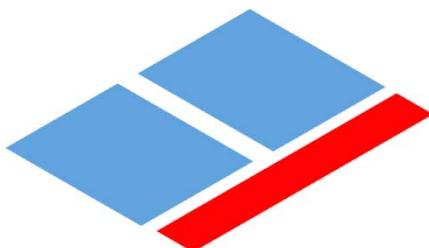
(SPC: Solid Phase Crystallization)



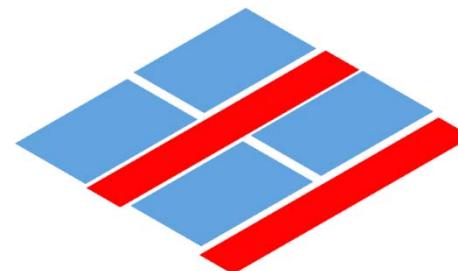
By using metal assisted SPC,  
monocrystalline Si growth in  
vertical channel.

# Costs and Performance

- Cost/GB is most critical for memory, but perf. also important.
- Performance gain is value.
- System solutions that enable user of lower perf. raw die is value.
- Simple rule of thumb:
- One the same technology node, typically 4 Plane die is ~15% bigger than 2 Plane die. But write performance doubles.
- Read perf. is also function of WL RC and cost.



2 Plane



4 Plane

# Process Challenges for 3D NAND

- Almost all scaling options post process challenges!
- Etch & Film intensive! **Etch is the new litho!**
- Thin-Film depositions in challenging geometries
  - Metals, barrier layers, di-electrics, epi etc
- RIE of Thin Film stacks
  - OPOP, ONON....
  - High aspect ratio hole etch at high TPUTs
  - Staircase etch with good CDU at high TPUTs
- Wet etch/cleans
  - Nitride recess
  - MH bottom clean
- Stress handling, planarization

# Key Thin Film Processes in 3D NAND

## Planar Films

- ONON Stack (PECVD)
- Carbon Hard-Mask Deposition (PECVD) → For MH, ST RIE & other masking steps

*Challenges: Uniformity, Defectivity, TPUT*

## 2D Features

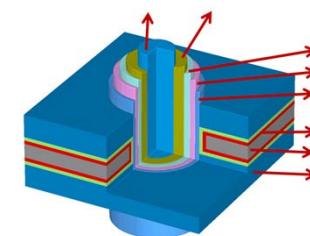
- MH Films: BLK Ox, CTL SiN, ONO TNL, Si Channel – Mostly Thermal CVD, ALD
- Metal Contacts: LI Fill – Thermal CVD
- Epi (LPCVD)

*Challenges: Uniformity, Step Coverage*

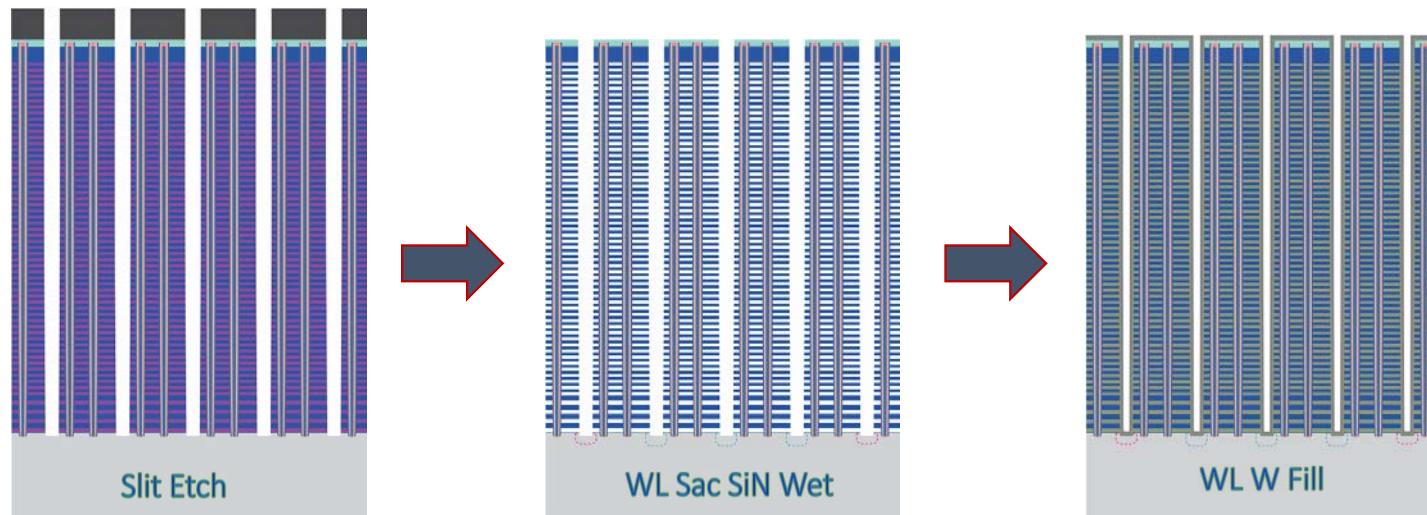
## 3D Features

- AlOx BLK – Thermal CVD
- WL TiN Barrier Layer – LPCVD
- WL-W – CVD (Gen2), ALD

*Challenges: Uniformity, Step Coverage, Gap-Fill, Low Residuals*

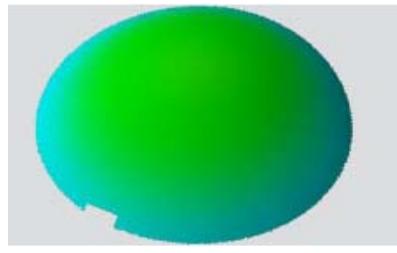


# A Key 3D NAND Challenge: Gate Replacement

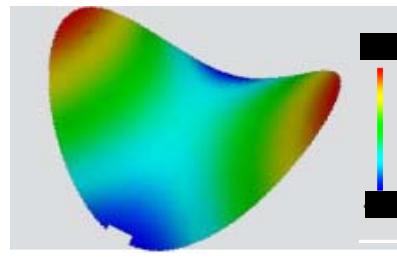


- Tungsten(W): Metal of choice due to its low  $R_s$  and thermal stability
- Several key issues for WL metallization:
  - Increasing complexity for gap fill
  - High wafer stress & High  $R_s$
  - Fluorine residuals

# To Manage the Stress

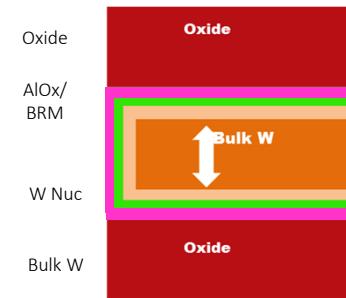
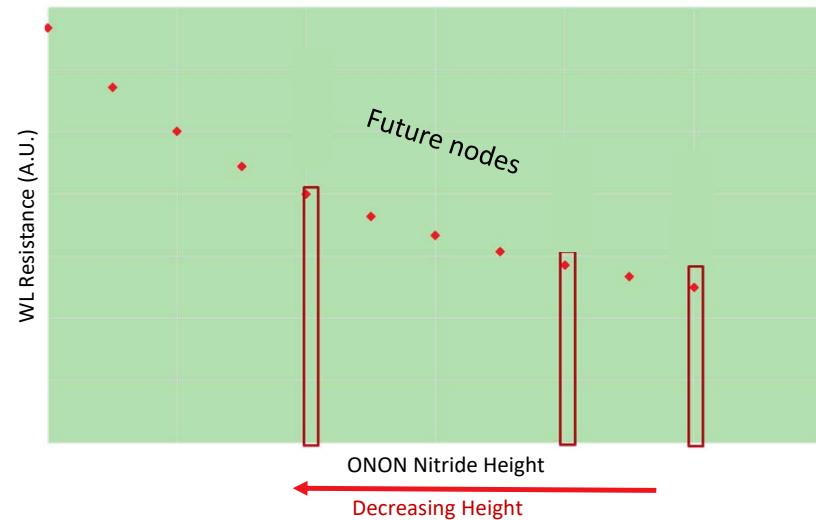


After CVD  
W-Fill



- CVD W-fill induces localized stress that contributes to the overall deformation of the wafer
- Problem becomes worse for higher # of layers

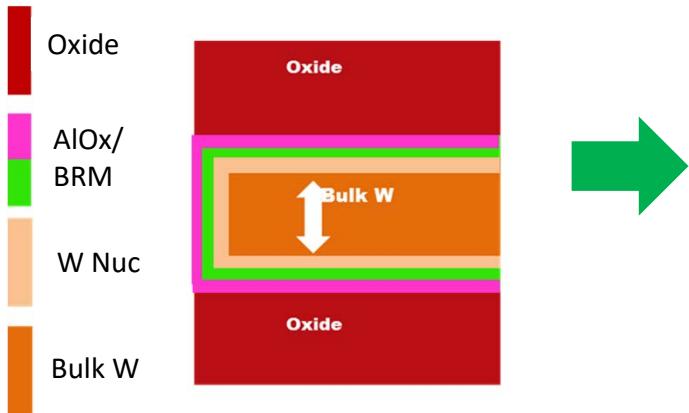
# WL Resistance Scaling & Challenges



- WL Rs expected to increase in future due to nitride thickness scaling
- Low WL resistivity solutions (thinner barriers, nucleation layer, lower bulk Rs, alternative metals) have to be developed.

# Alternate Metals Candidates

Tungsten gapfill limits ON scaling.



- W deposited from WF<sub>6</sub> => F incorporation
- TiN ( $\rho \sim 300 \mu\Omega\text{-cm}$ ) layer needed
- W needs a nucleation layer
  - Occupies valuable real estate
  - Very high resistance ( $\rho \sim 200 \mu\Omega\text{-cm}$ )
  - Highest F content
- W bulk fill ( $\rho \sim 20 \mu\Omega\text{-cm}$ ) => Adds stress

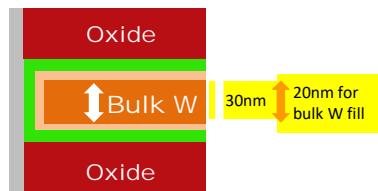
22 <b>Ti</b> Titanium 47.88	23 <b>V</b> Vanadium 50.942	24 <b>Cr</b> Chromium 51.996	25 <b>Mn</b> Manganese 54.938	26 <b>Fe</b> Iron 55.933	27 <b>Co</b> Cobalt 58.933	28 <b>Ni</b> Nickel 58.693	29 <b>Cu</b> Copper 63.546	30 <b>Zn</b> Zinc 65.39
40 <b>Zr</b> Zirconium 91.224	41 <b>Nb</b> Niobium 92.906	42 <b>Mo</b> Molybdenum 95.95	43 <b>Tc</b> Technetium 98.907	44 <b>Ru</b> Ruthenium 101.07	45 <b>Rh</b> Rhodium 102.906	46 <b>Pd</b> Palladium 106.42	47 <b>Ag</b> Silver 107.868	48 <b>Cd</b> Cadmium 112.411
72 <b>Hf</b> Hafnium 178.49	73 <b>Ta</b> Tantalum 180.948	74 <b>W</b> Tungsten 183.85	75 <b>Re</b> Rhenium 186.207	76 <b>Os</b> Osmium 190.23	77 <b>Ir</b> Iridium 192.22	78 <b>Pt</b> Platinum 195.08	79 <b>Au</b> Gold 196.967	80 <b>Hg</b> Mercury 200.59

- Gap fill
- Low resistivity & stress
- Low cost

# Comparison of W, Co & Ru

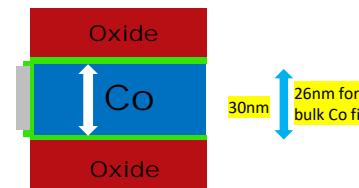
Metal	Bulk Modulus [Gpa]	T Melting [°C]	$\rho$ [ $\mu\Omega\text{-cm}$ ]
W	310	3422	1X
Co	180	1495	1.2X
Ru	220	2334	1.4X

CVD/ALD Tungsten



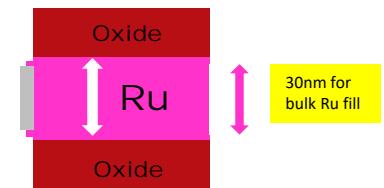
Less real estate  
High fluorine content  
Overall high resistivity  
Severe stress

CVD/ALD Cobalt



More real estate  
No fluorine  
Lower resistivity  
Lower stress

CVD/ALD Ruthenium



More real estate  
No fluorine  
Slightly higher resistivity  
(offset by 'no barrier')

# Etch Is the New Litho

## Holes



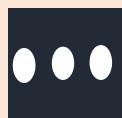
Memory Hole (**MH**) → To enable Memory films deposition



Supporting Hole → To enable supporting structure during replacement process



Stair Contact (**CC**) → To enable metal connection to each Word line (WL)

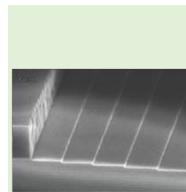


CMOS Contact (**CS**) → To enable metal connection to CMOS

## Trenches



Slit (**ST**) → To enable W replacement



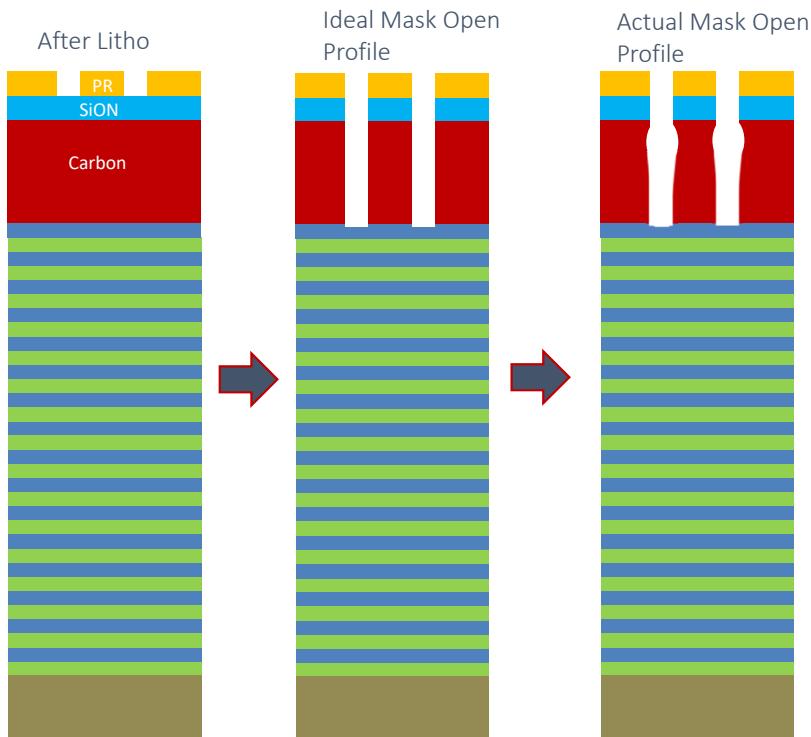
### Staircase

Staircase Etch → To enable contacts for each Word line (WL)

### Memory Hole Bottom RIE

- Make Bottom Punch in MH to enable connection to bottom Si

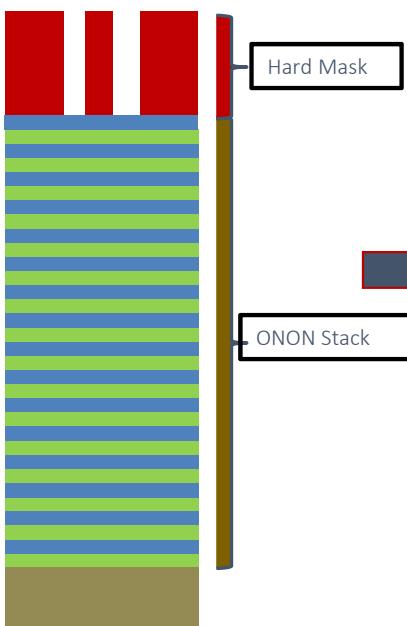
# Hard Mask for Dry Etching



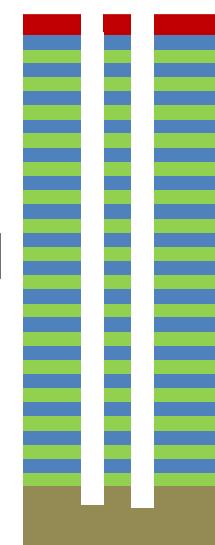
- Carbon (APF) is hard mask (HM) material of choice for high aspect ratio etching
- Typical HM aspect ratio (AR) is  $\geq 15$ .
- HM open profile control along with circularity (for holes) and LER/LWR (for trenches) is very challenging
- HM open profile plays a significant role in final stack etch profile
- Advanced RIE tools with complex recipes (multiple gas chemistries & high RF powers) used to achieve target specs

# Memory Hole Etch: Heart and Soul of 3D NAND

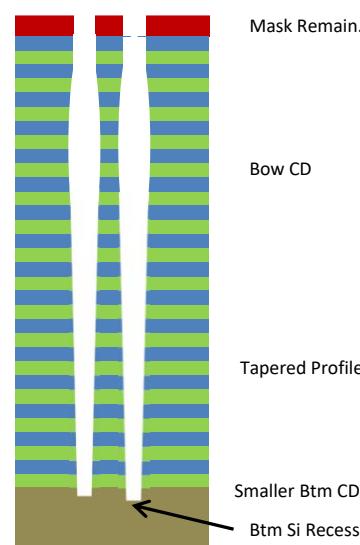
MH : Pre Etch



Post Etch : Ideal Profile



Post Etch : Actual Profile

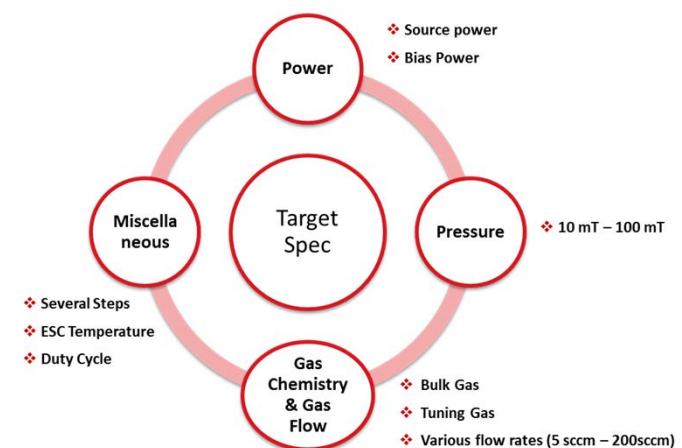


Mask Remain.

Bow CD

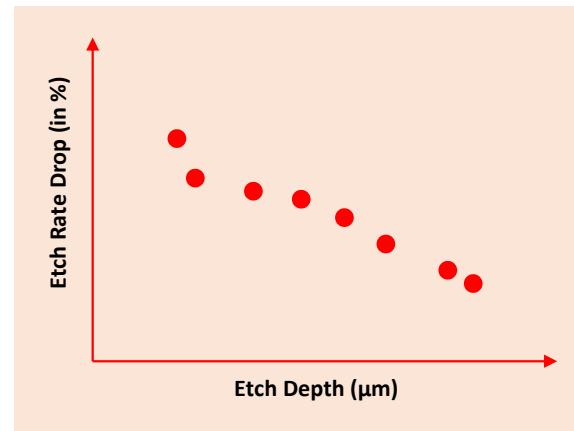
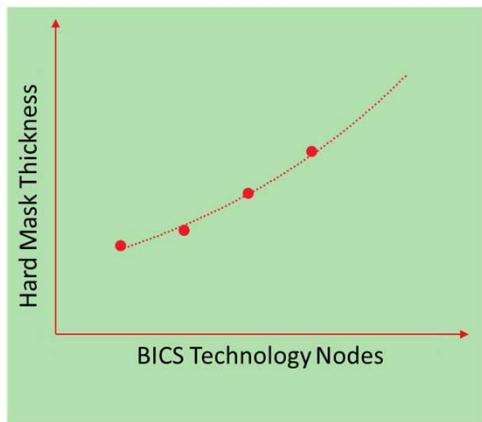
Tapered Profile

Smaller Btm CD



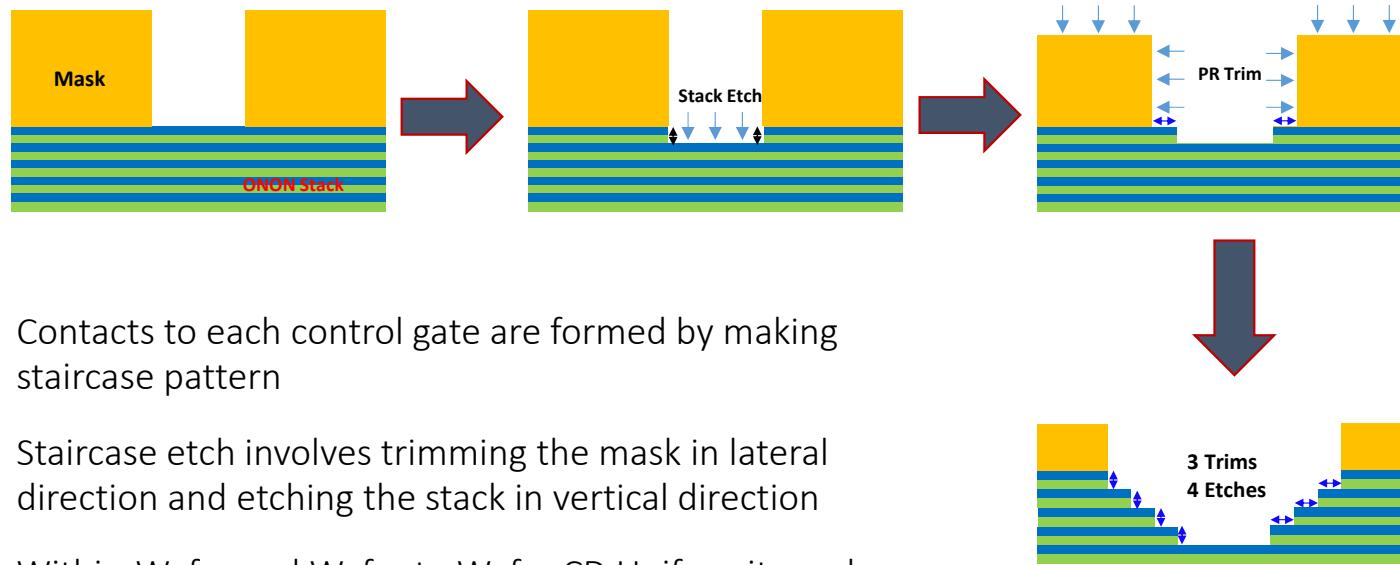
# Scaling Challenges for MH RIE

- Z-scaling needs more layers and ON stack height increases. XY scaling demands smaller MH CD.
- This results in >>40:1 Aspect Ratio (AR) openings



- Future scaling requires thicker conventional hard mask to etch high aspect ratio (HAR) holes
- Etch Rate drops significantly while etching deeper ON/OP films
- Advanced HAR etching will be key enabler for future 3D NAND

# Staircase Etch for Contact Formation



- Contacts to each control gate are formed by making staircase pattern
- Staircase etch involves trimming the mask in lateral direction and etching the stack in vertical direction
- Within-Wafer and Wafer to Wafer CD Uniformity and Productivity are critical

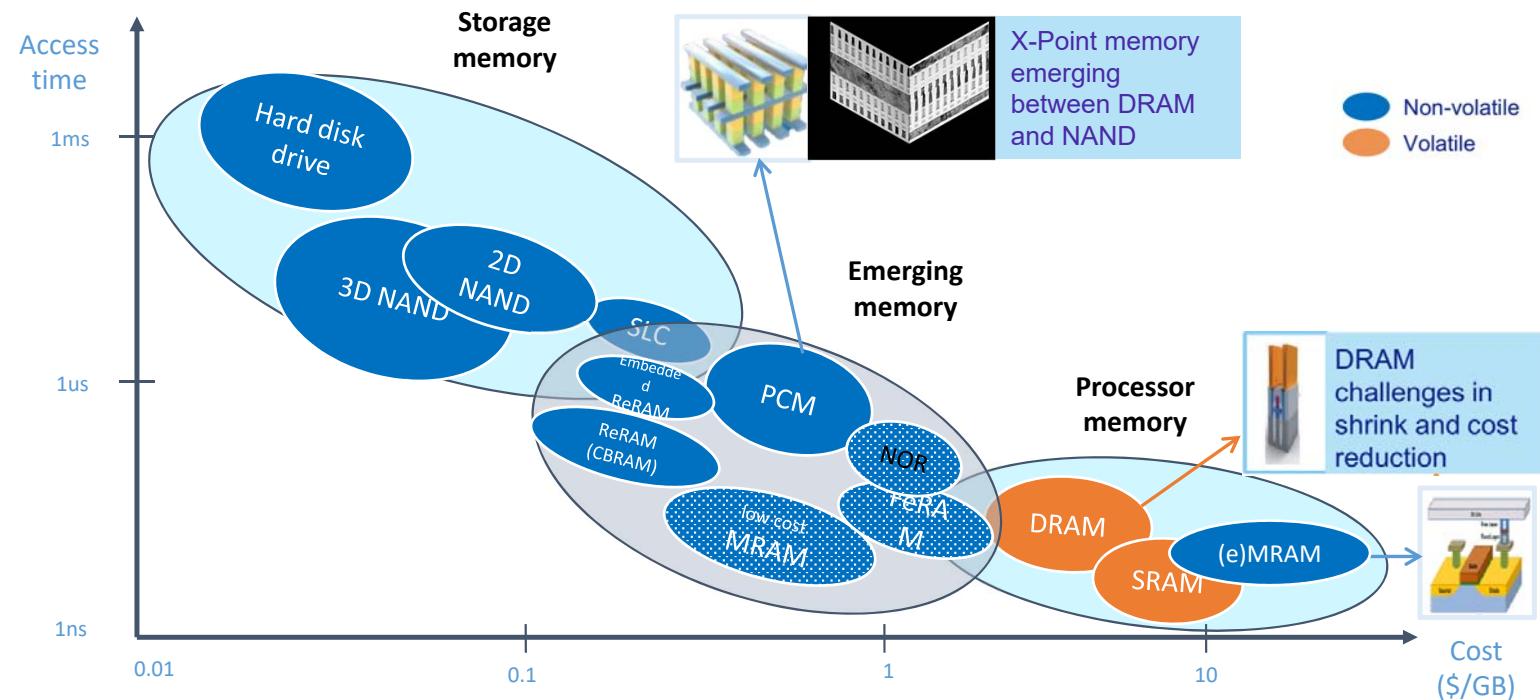
# New Opportunities: Tools and Materials

- New tool architecture designed for 3D
- New chemistry and tool for high aspect ratio hole etch
- Higher selectivity hard mask
- Low resistivity low stress wordline materials
- Improved channel interface and mobility for cell current
- Improved uniformity control
- Better fundamental understanding
- Process control, chamber matching
- .....
- **Need the industry to work together to build the future**

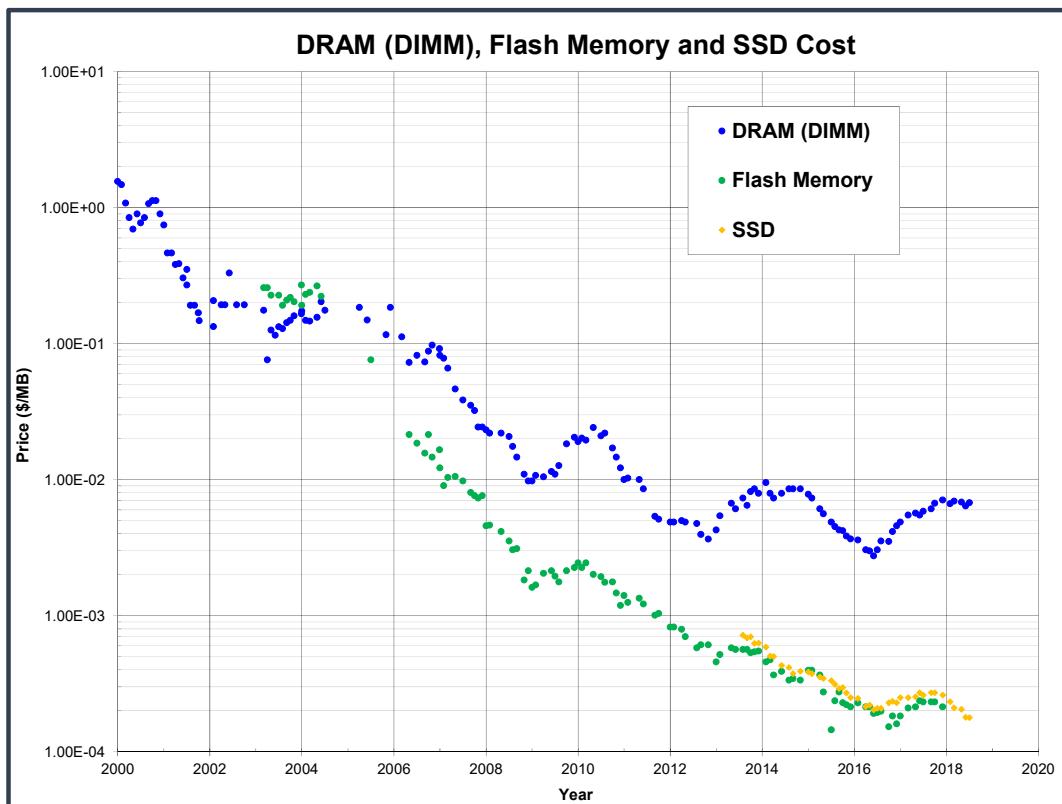
# Outline

- 1. Introduction to 3D NAND**
- 2. Scaling Challenges and Opportunities**
  - Cost Scaling and CapEx
  - Architecture Options
  - Device/Process Scaling and Challenges
- 3. Versatile NAND and System Solution Opportunities**
- 4. Conclusions**

# Promise and Reality of Various Memory Technologies



# Widening Cost Gap Between DRAM and NAND

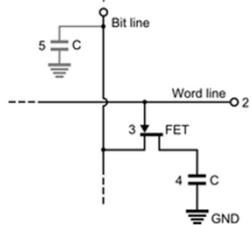


- DRAM scaling slowing significantly
- Many options available to reduce the needed DRAM usage in various systems

<https://jcmit.net/flashprice.htm>

# Reality of Today's Tiers of Memory/Storage

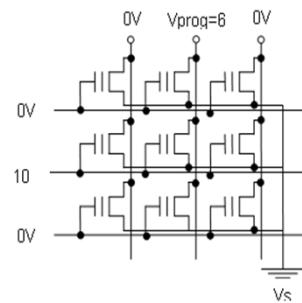
DRAM



House with own plane  
And access to runway

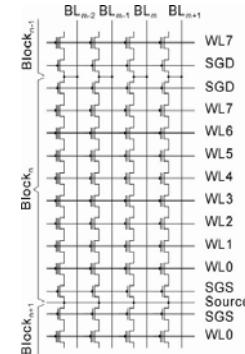
~\$70B

NOR Flash



Expensive Hotel

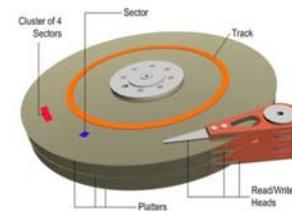
NAND Flash



No Hall Way Hotel  
Disturb problems

~\$60B

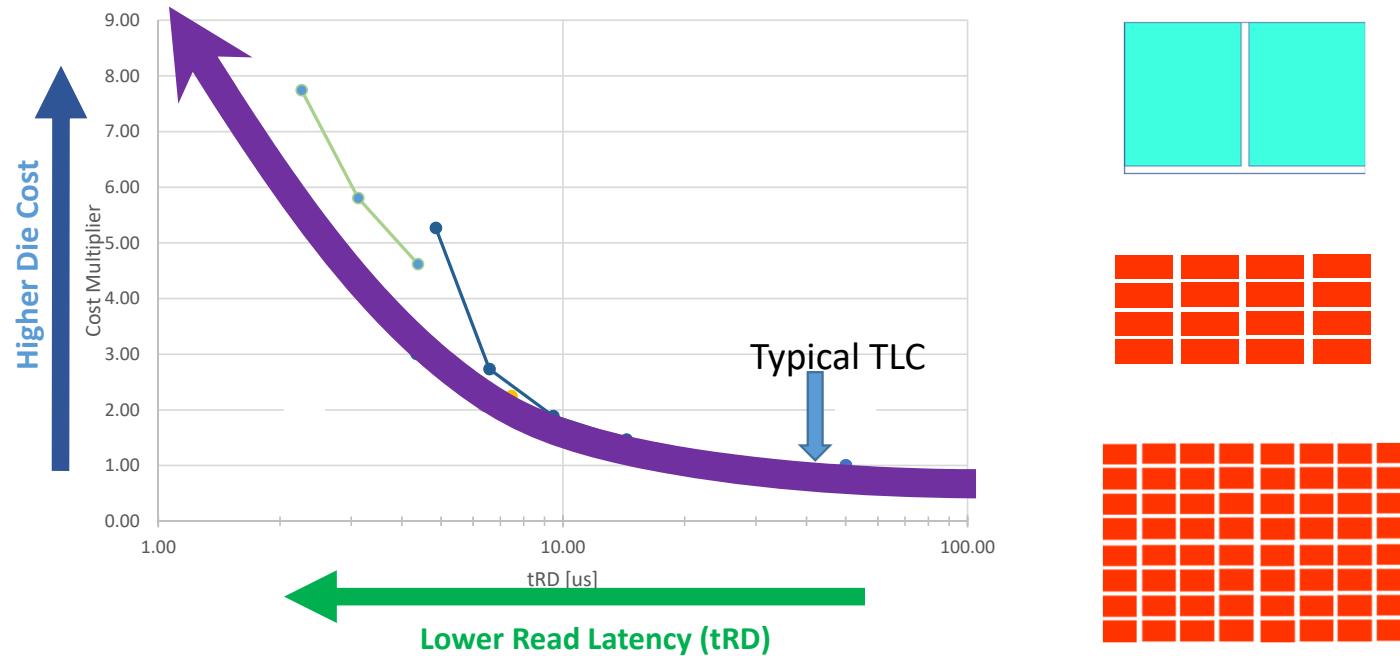
HDD



Chopper to drop guest

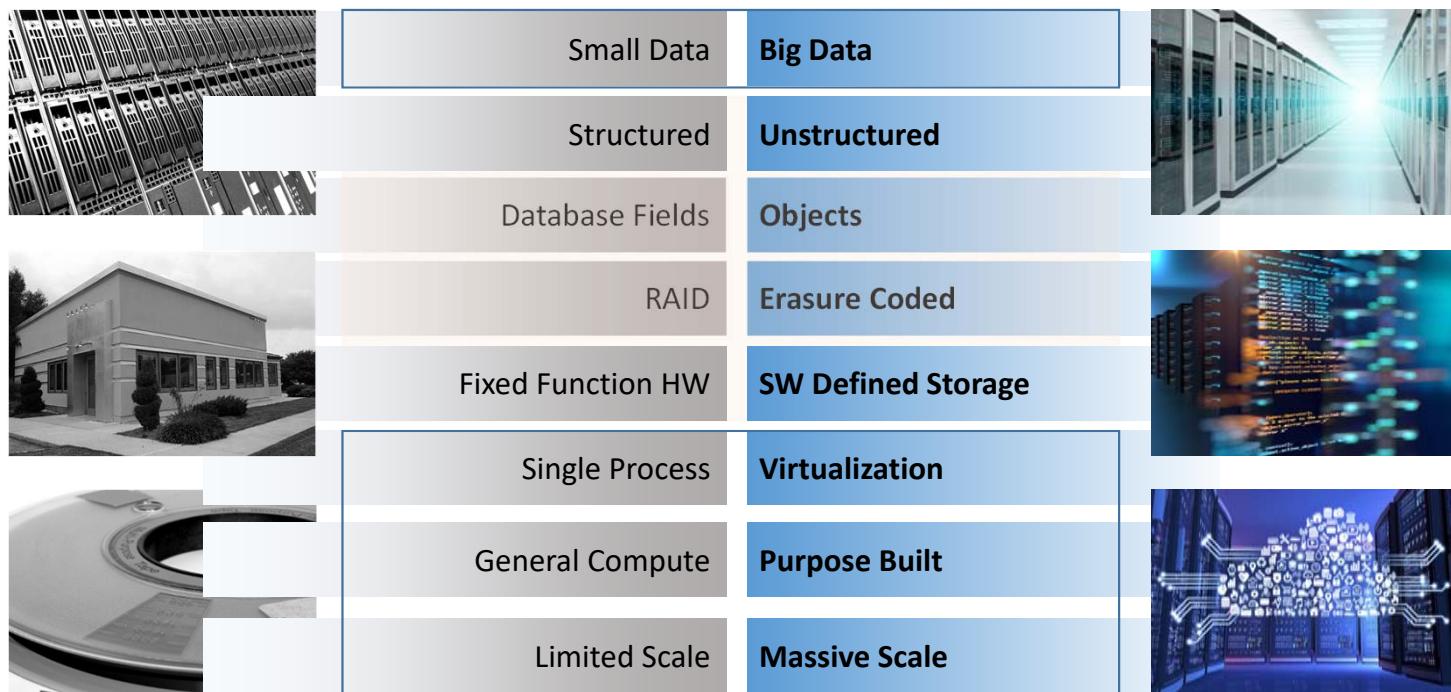
~\$20B

# NAND: Large Range of Perf. Options Available



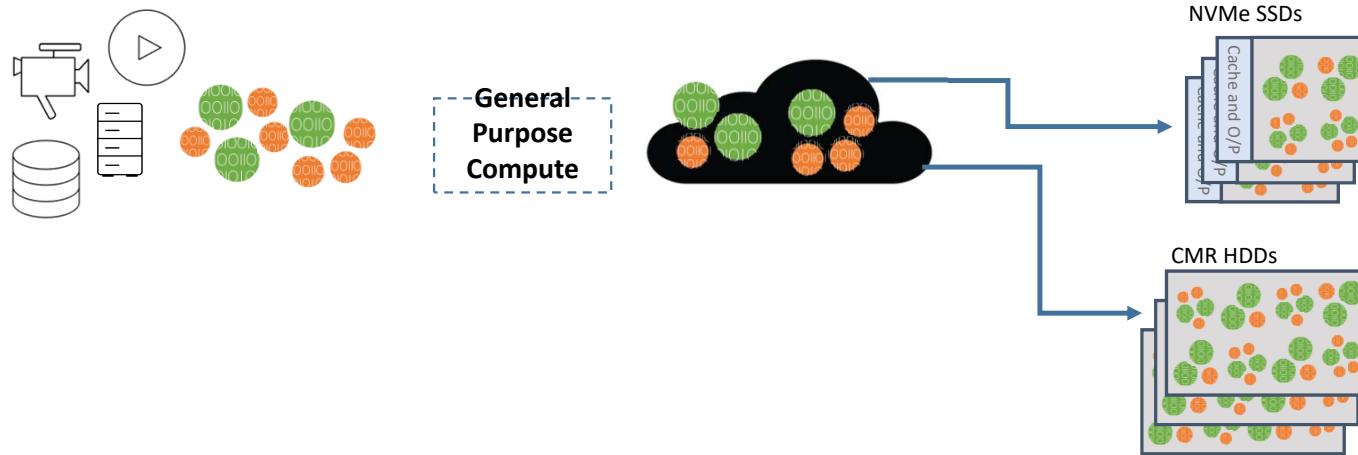
- With proper trade-off, can achieve read access time tR approaching 1usec
- Can achieve write perf. of 1GB/sec
- Endurance >500K or >1M cycles

# Fundamental Shifts in the Data Center



# Inefficiency in Data Centers Today

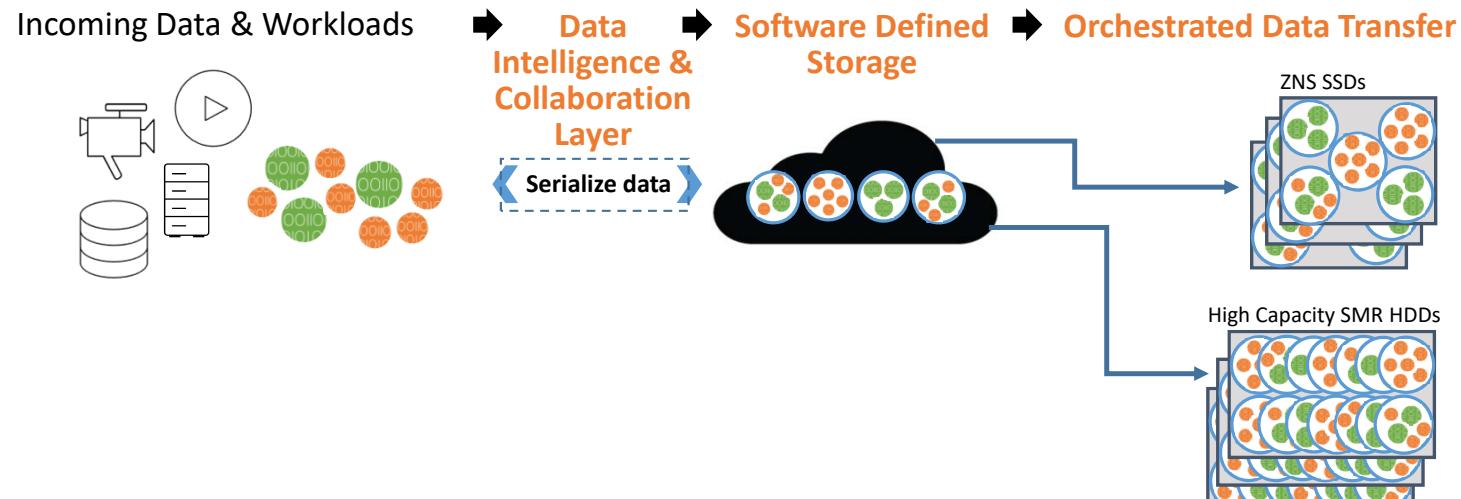
Incoming Data & Workloads → Compute → LBA0, LBA1, LBAXx → Pass-through Data Transfer



**General purpose storage must allocate unused capacity and/or DRAM to manage any and all workloads**

# Re-architecting for the Zettabyte Age

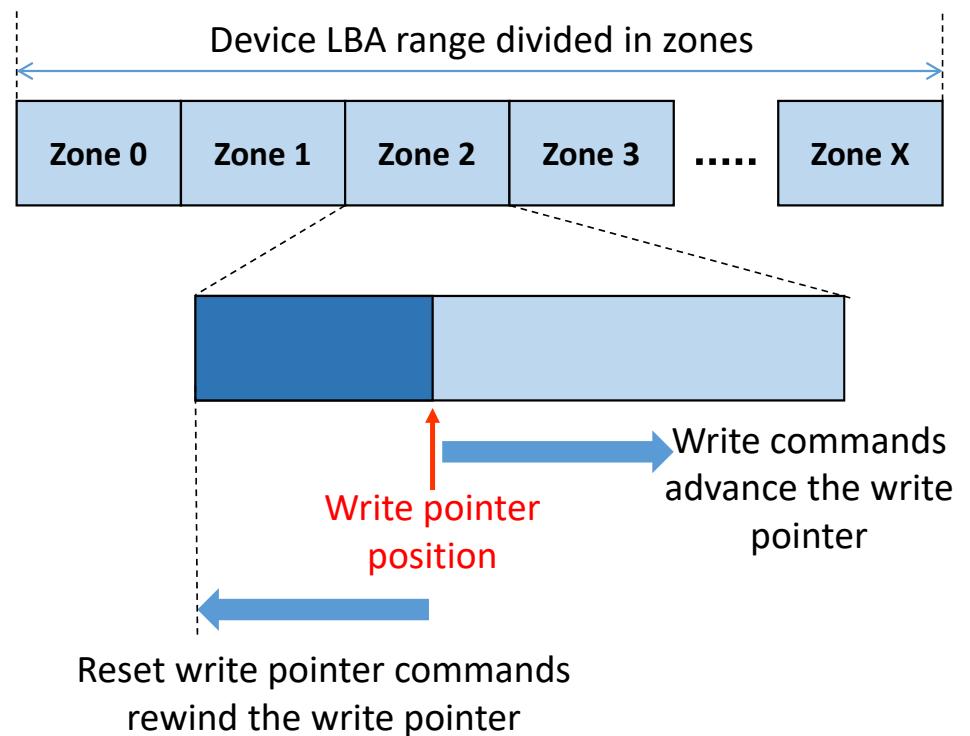
Workload-optimized storage with leveraged application development



**Serialized data streams enable greater efficiency as the storage device manages data placement.**

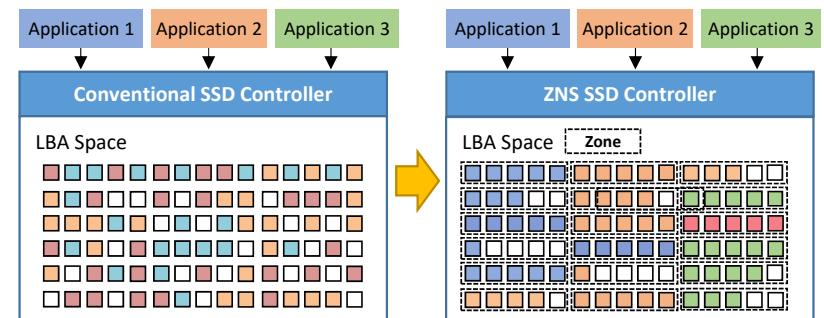
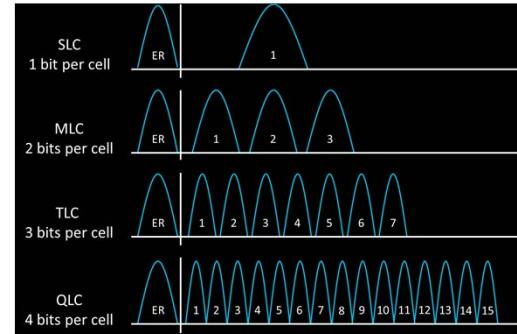
# What are Zoned Block Devices?

- The storage device logical block addresses are divided into ranges of zones.
- Writes within a zone must be sequential.
- The zone must be erased before it can be rewritten.



# Zoned Block Devices for SSDs

- TLC & QLC increases capacity but at cost of
  - Less endurance
  - Lower performance
  - More DRAM to map higher capacity
- With Zoned Block Access, host/SSD cooperate on optimal data placement
  - Reduces write amplification and internal data movement
- Lower write amplification and data movement advantageous to TLC & QLC
  - Reduces wear
  - Improves latency outliers and throughput
  - Reduces DRAM in SSD (smaller L2P)
  - Also happens to reduce OP

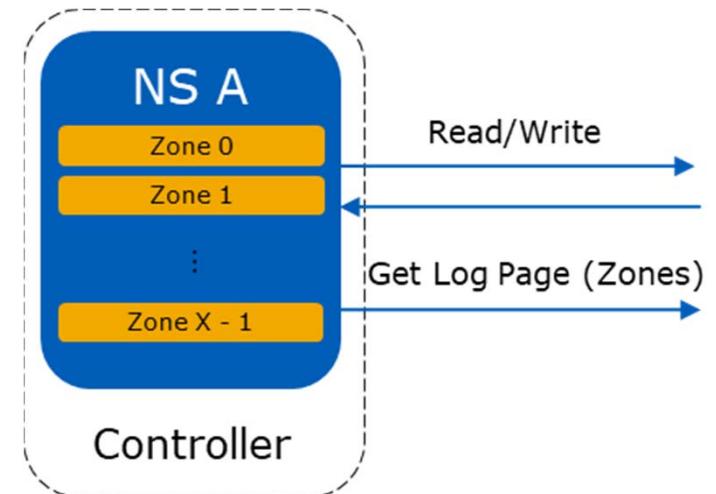


# Zoned Name Spaces (ZNS)

- Ongoing Technical Proposal in the NVMe™ working group
- New Zoned Command Set – Inherits the NVM Command Set and adds zone support.
- Aligns to the existing host-managed models defined in the ZAC/ZBC specifications.
  - Note that it does not map 1:1. Beware of the details.
- Optimized for Solid State Drives
  - Zone Capacity
  - Zone Attributes introduced to optimize for SSD characteristics
  - Zone Append
  - Zone Descriptors

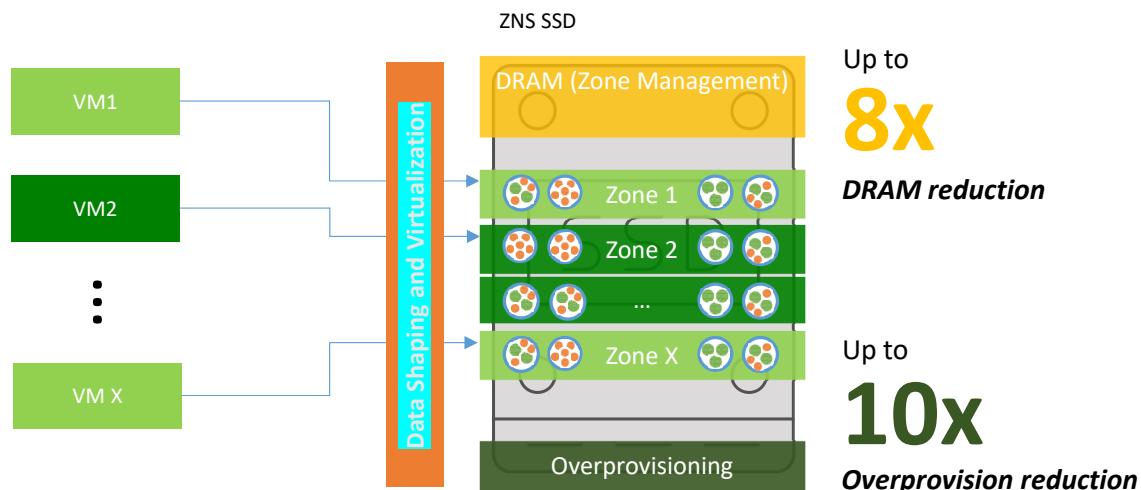


Under review



# ZNS Enables Efficient, Purpose-Built Storage (ZNS: Zoned Named Spaces)

Provides host with system-level intelligence for data placement



**ZNS simplifies firmware and data placement**

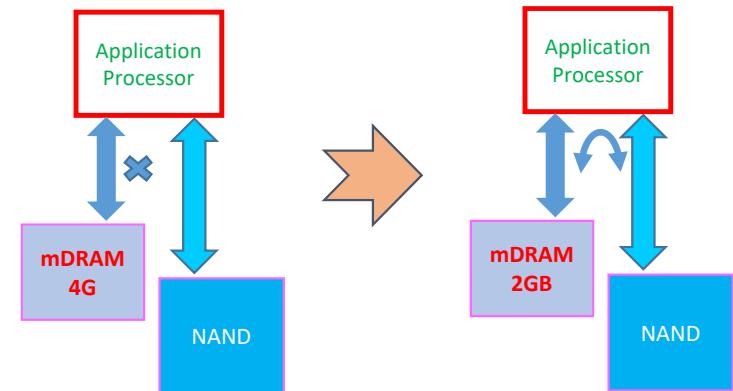
Source: Western Digital internal modeling data, April 2019

50

# Mobile System Today

## Current status in the eco systems:

- Android System kills Apps to free RAM space instead of swapping out relevant pages when memory is low
- Enablement of efficient swap can reduce DRAM needs
- Many hurdles and concerns: performances, IO bandwidth, Flash memory endurance, user experience etc.
- By solving IO and flash endurance/perf. bottleneck, can enable efficient swap solution that can reduce DRAM needs and enlarge virtual memory



# Conclusions

- 3D NAND just entering Teenage Year, long road ahead
- Significant cost scaling and CapEx challenges
- Novel architectures, materials and processes in development
- Ample innovations opportunities for tool and material
- # of Layer is not the issue, it's the economy!
- Solutions in system/eco-system to utilize the full potential of 3D NAND