Emerging Non-Volatile Memories-Applications, Challenges and Solutions

Swaroop Ghosh

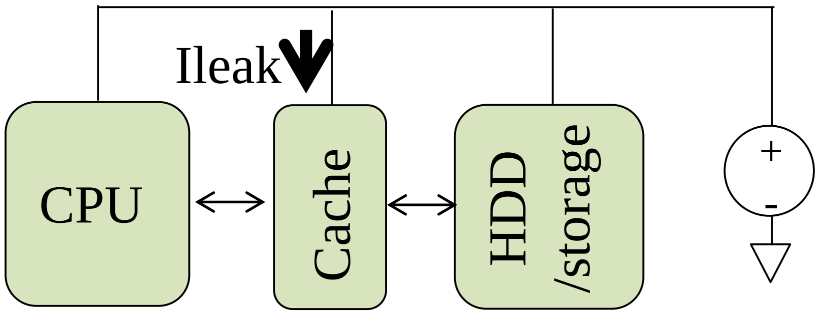
School of EECS, The Pennsylvania State University

szg212@psu.edu

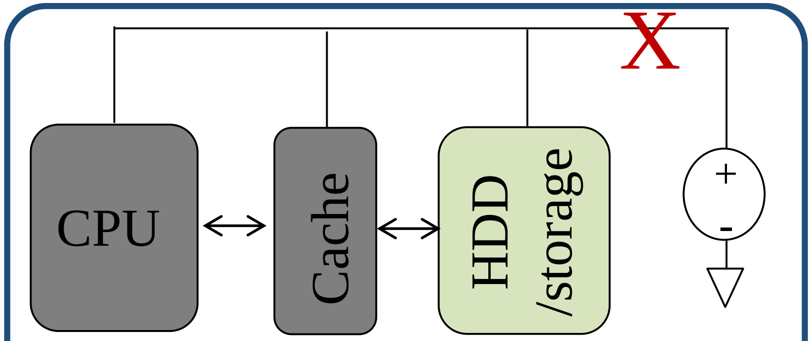
Sponsors

Why Emerging NVM?

Conventional

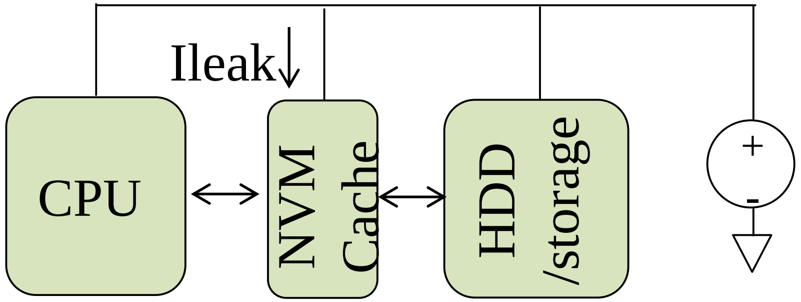


Suffers from leakage, bandwidth limited by cache size

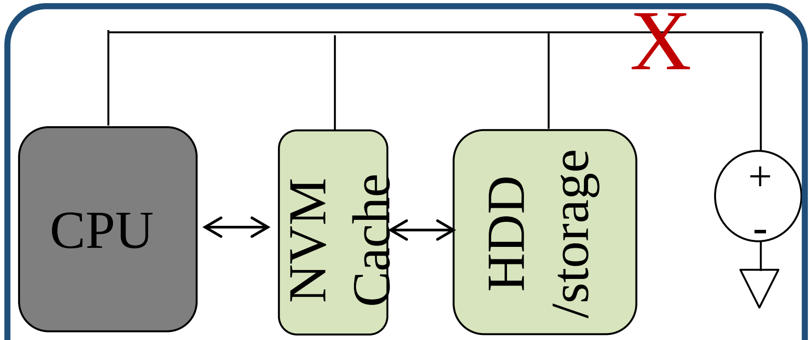


Looses data if supply is cut, requires power even at idle mode

Emerging NVM



Less leakage, small bitcell footprint, high bandwidth due to large cache

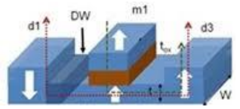


Supply can be disconnected at idle, saves power, maintains instant ON

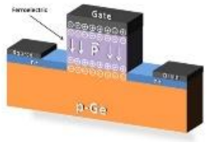
PCRAM

Electrode
GST
TiN
SiO₂
Electrode

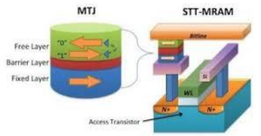
DWM



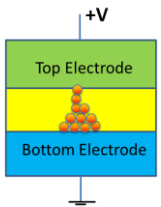
FeRAM



STTRAM

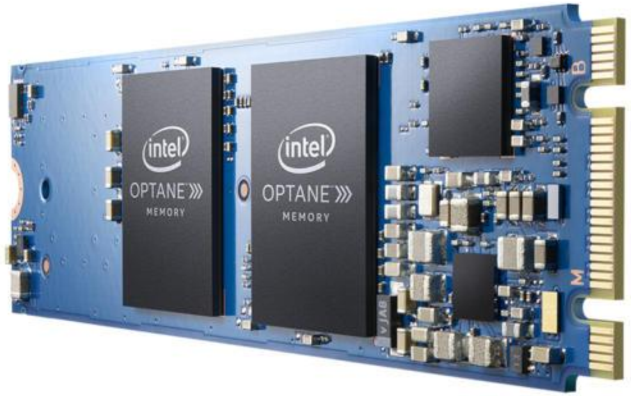


ReRAM



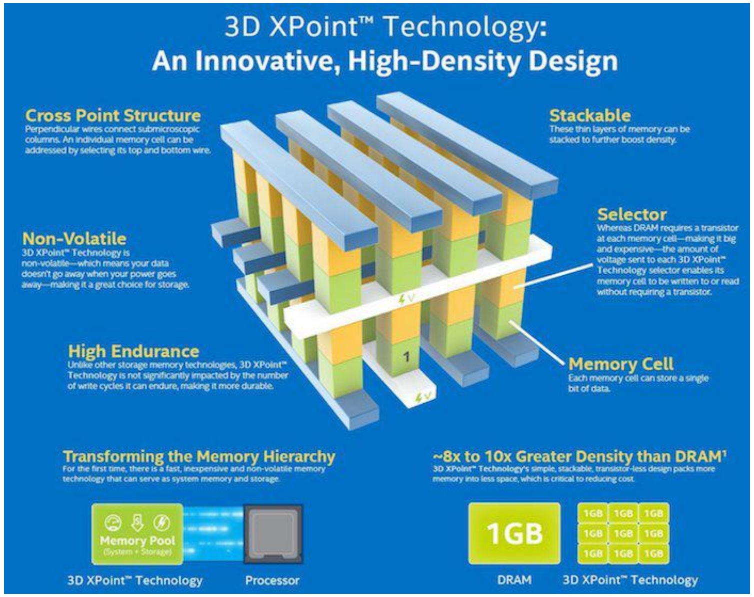
Recent Commercialization of Emerging NVMs

Phase Change RAM*



Intel unveils its Optane hyperfast memory

Intel released few key details around its new non-volatile memory



3D XPoint™ Technology:
An Innovative, High-Density Design

Cross Point Structure
Perpendicular wires connect submicroscopic columns. An individual memory cell can be addressed by selecting its top and bottom wire.

Non-Volatile
3D XPoint™ Technology is non-volatile—which means your data doesn't go away when your power goes away—making it a great choice for storage.

High Endurance
Unlike other storage memory technologies, 3D XPoint™ Technology is not significantly impacted by the number of write cycles it can endure, making it more durable.

Stackable
These thin layers of memory can be stacked to further boost density.

Selector
Whereas DRAM requires a transistor at each memory cell—making it big and expensive—the amount of voltage sent to each 3D XPoint™ Technology selector enables its memory cell to be written to or read without requiring a transistor.

Memory Cell
Each memory cell can store a single bit of data.

Transforming the Memory Hierarchy
For the first time, there is a fast, inexpensive and non-volatile memory technology that can serve as system memory and storage.

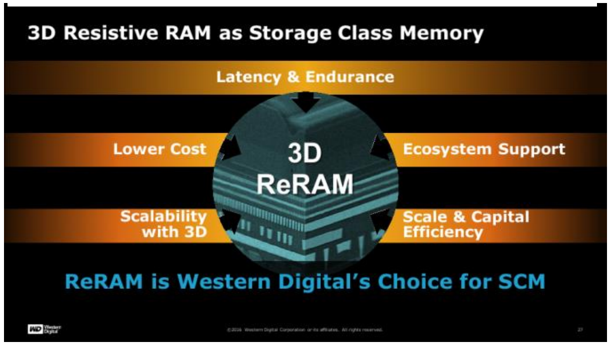
~8x to 10x Greater Density than DRAM[1]
3D XPoint™ Technology's simple, stackable, transistor-less design packs more memory into less space, which is critical to reducing cost.

STT- MRAM

Published: *March 9, 2017*

Everspin unveils a new low latency, PCIe NVMe card based on Spin Torque MRAM

ReRAM



3D Resistive RAM as Storage Class Memory

Latency & Endurance
Lower Cost
3D
ReRAM
Ecosystem Support
Scalability with 3D
Scale & Capital Efficiency

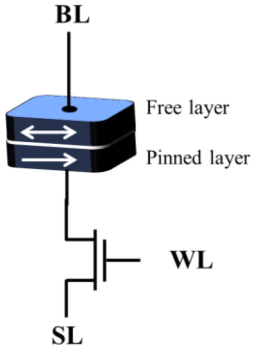
ReRAM is Western Digital's Choice for SCM

Western Digital to Use 3D ReRAM as Storage Class Memory for Special-Purpose SSDs

by Anton Shilov on August 12, 2016 8:00 AM EST

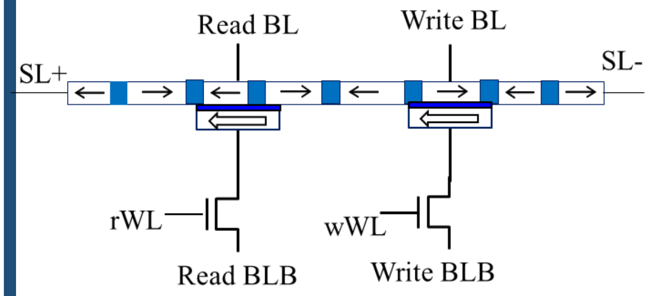
Emerging Technologies

STTRAM



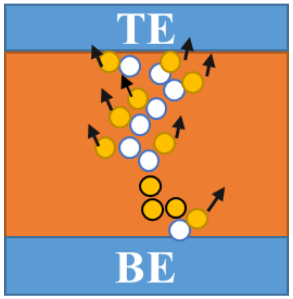
- Single bit/cell
- Footprint = ~12-40F^2
- Random access
- Read/write latency
  - Read/write of MTJ
- Read: Sensing MTJ resistance
- Write: Flip free layer

DWM



- Multiple bits/cell
- Footprint = 2.5F^2
- Tailored for serial access
- Read/write latency
  - Shift + read/write of MTJ
- Read: Sensing MTJ resistance
- Write: Flip NW domain

ReRAM

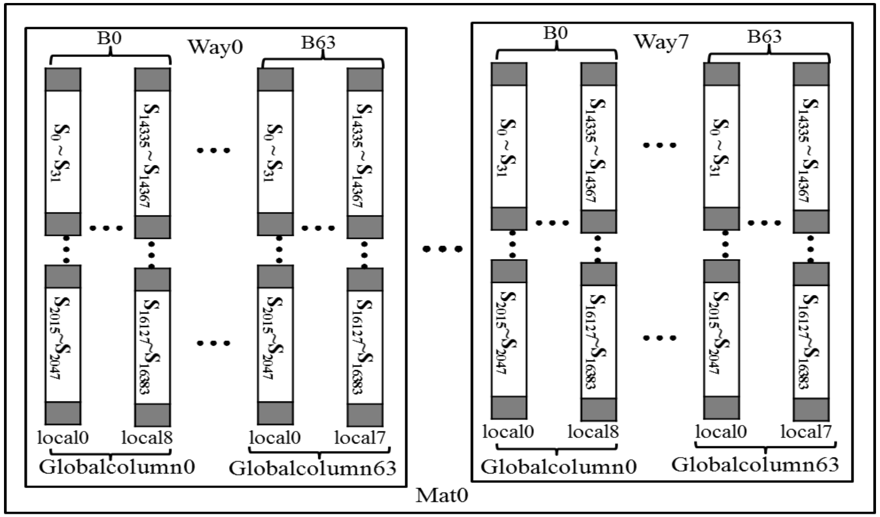


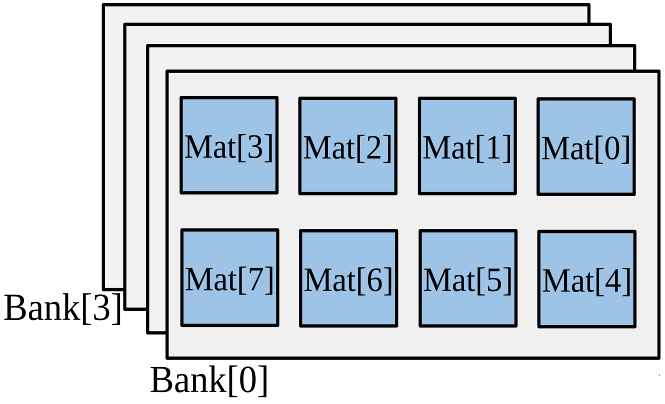
- Multiple bit/cell
- Footprint = ~4F^2
- Random access
- Read/write latency
  - Read/write of ReRAM
- Read: Sensing ReRAM resistance
- Write: break or make conductive path

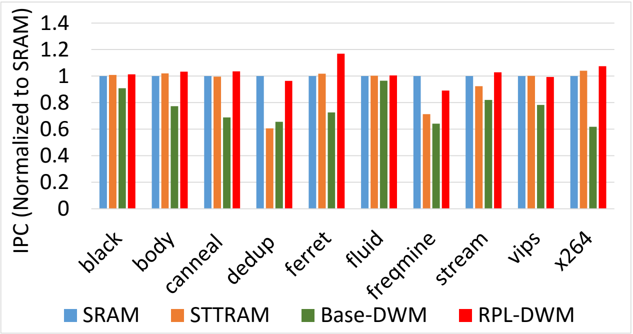
Outline

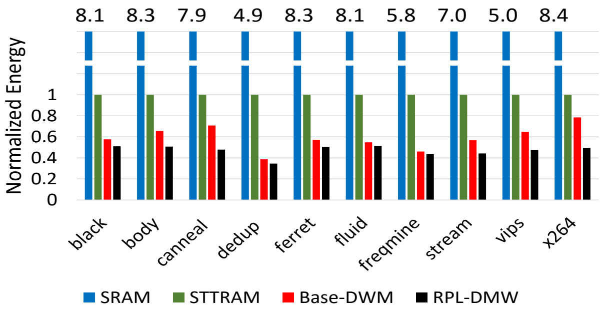
- Introduction and motivation
- Applications
  - Last level cache
  - Energy efficient computing
  - Security
- Challenges
  - Retention test
  - Long read/write latency
  - High asymmetric read/write current
- Solutions
  - Retention compression
  - Circuit to system synergistic design
  - Attack sensors and prevention
  - Sensing circuit
- Summary

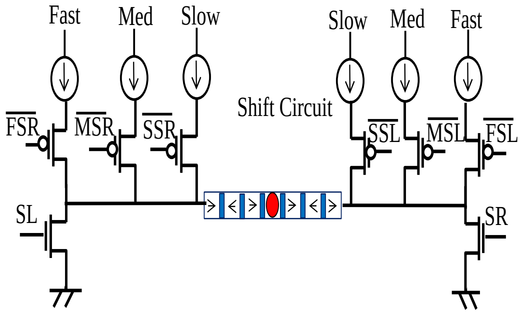
Last Level Cache





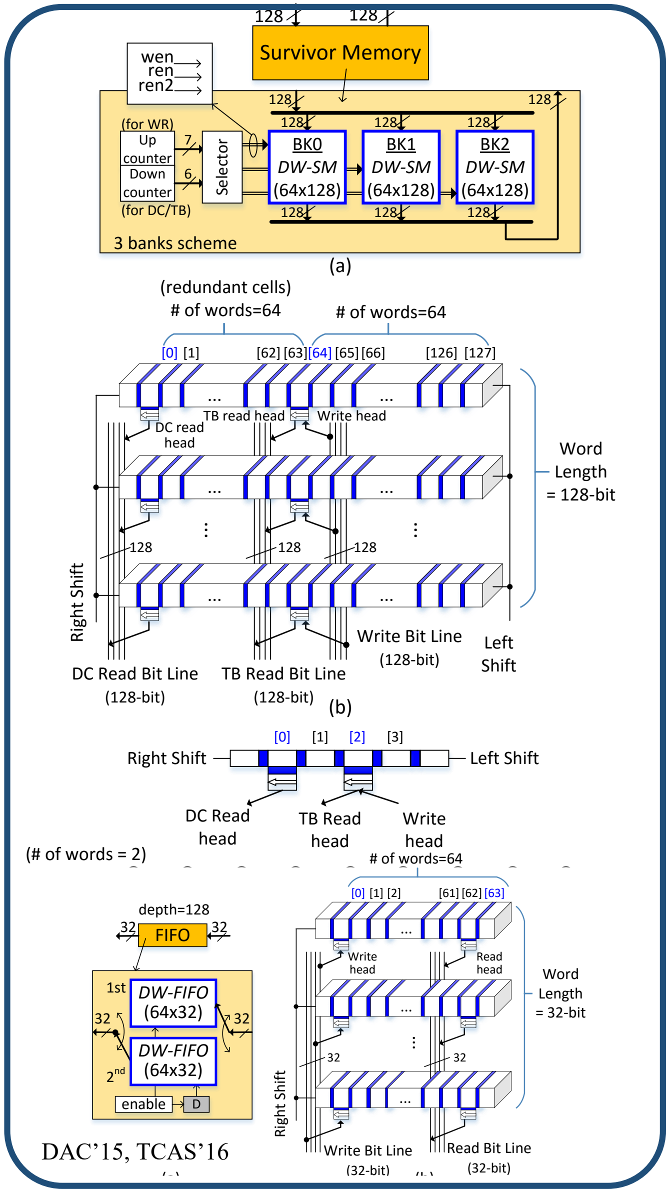




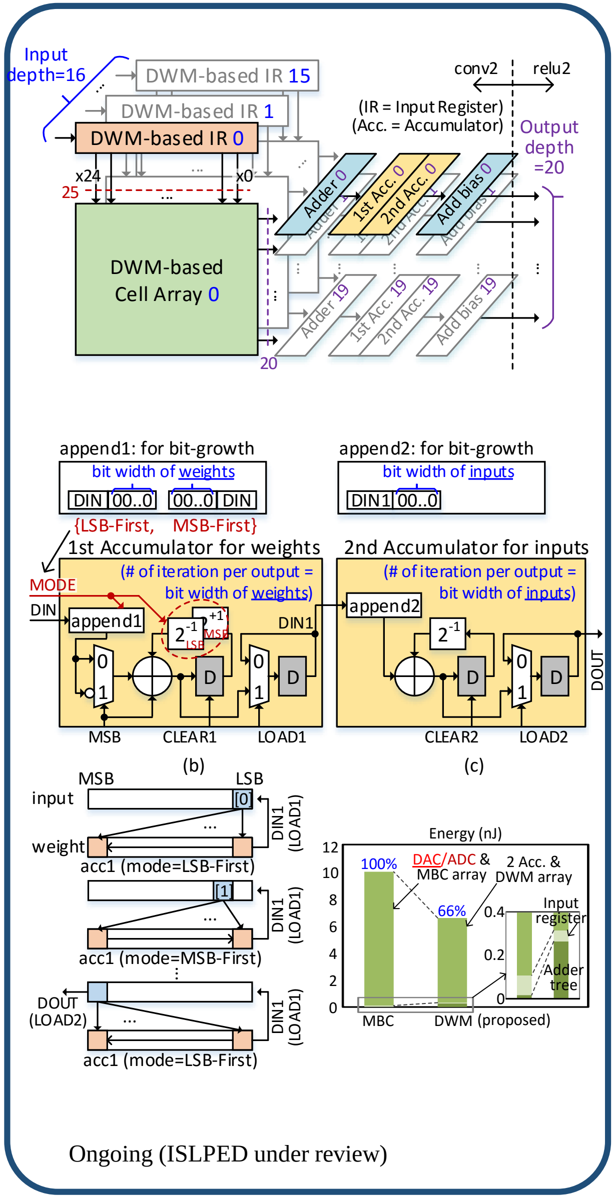


- Performance improvement: 3-33%
- Power reduction: 1.2X-14.4X

Digital Signal Processing and Neuro-Inspired Computing



DAC’15, TCAS’16

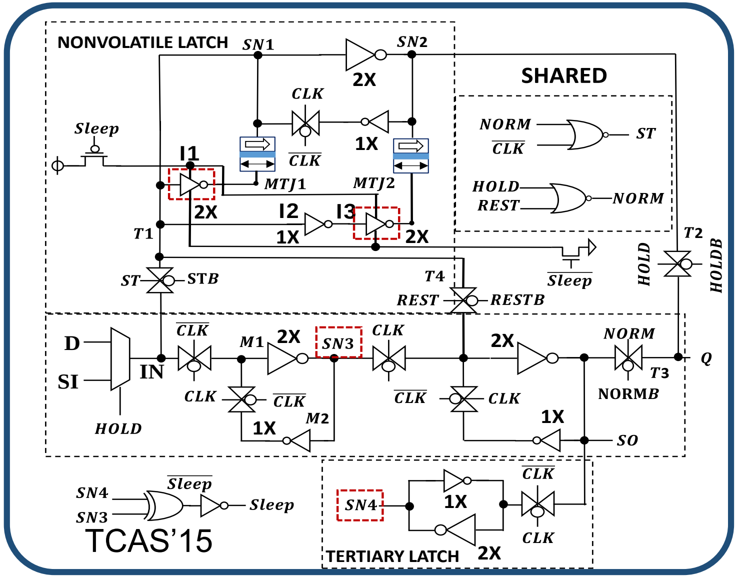


Ongoing (ISLPED under review)

- DWM based Viterbi decoder
  - 66.4 % area and 59.6 % power savings
- DWM based 8K point FFT processor
  - 60.6 % area and 60.3 % power savings
- Neuro-inspired computing
  - 34% energy savings compared to memristor based computing
  - Bit-width extendibility

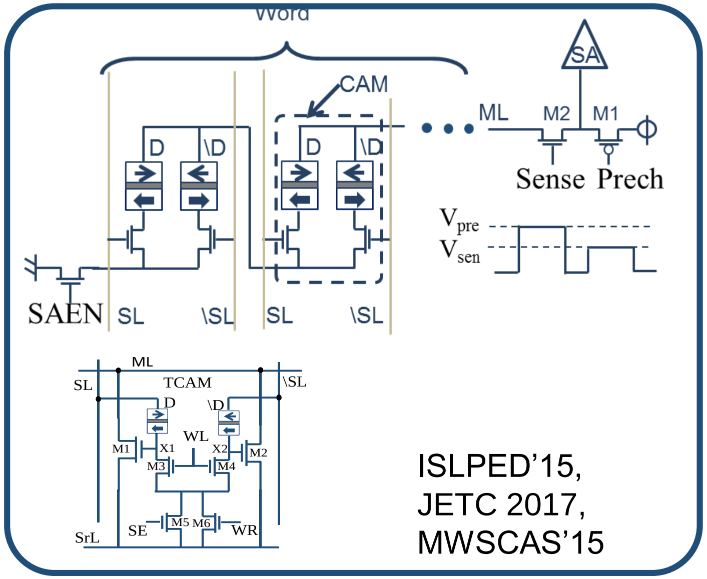
Energy-Efficient Memory Design

State retentive sequentials



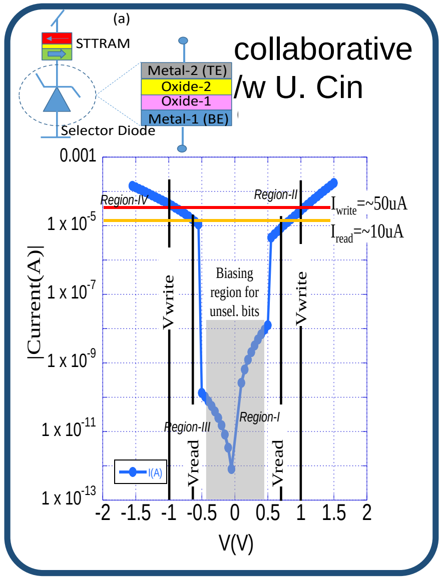
TCAS'15

Non volatile CAMs



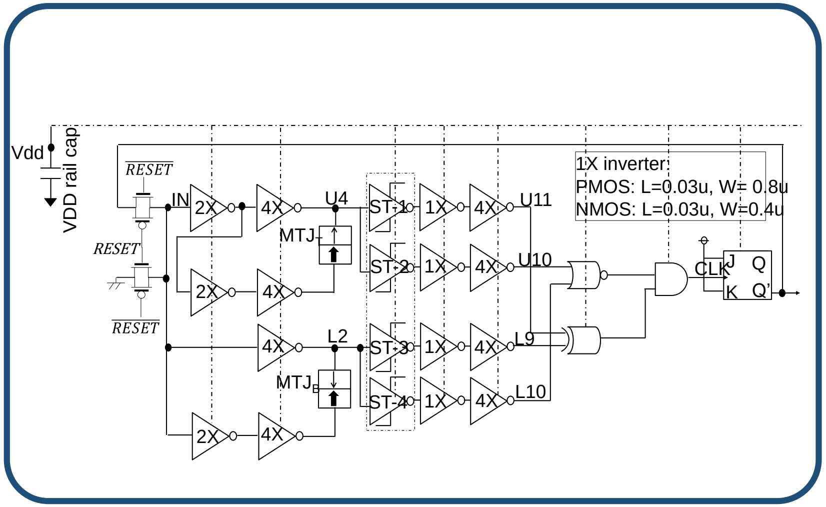
ISLPED'15, JETC 2017, MWSCAS'15

Selector for STTRAM



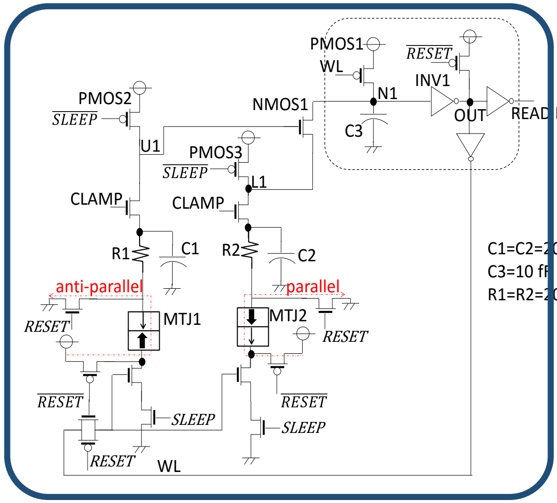
collaborative /w U. Cin

Write latency sensor



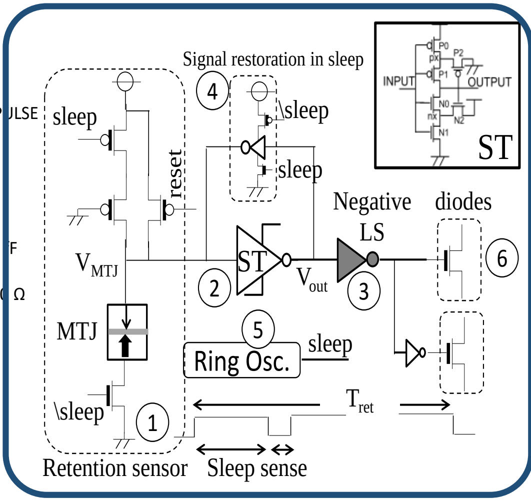
1X inverter:
PMOS: L=0.03u, W= 0.8u
NMOS: L=0.03u, W=0.4u

Read latency sensor



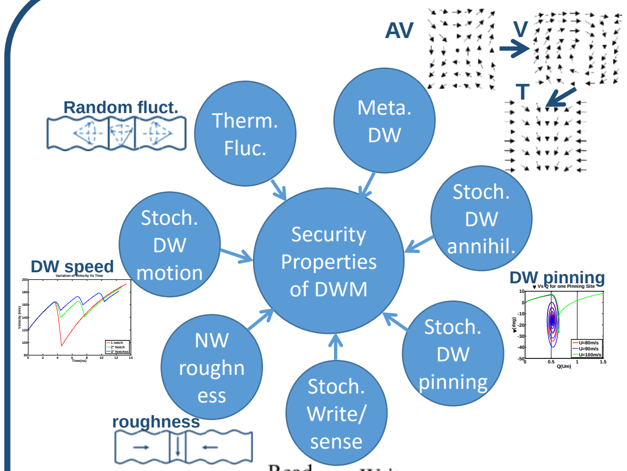
C1=C2=20 fF
C3=10 fF
R1=R2=200 Ω

Retention sensor



Copyright: Swaroop Ghosh

Spintronics for Security


Security Properties of DWM

Therm. Fluc.

Meta. DW

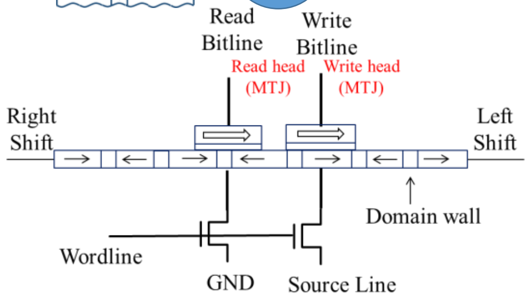
Stoch. DW annihil.

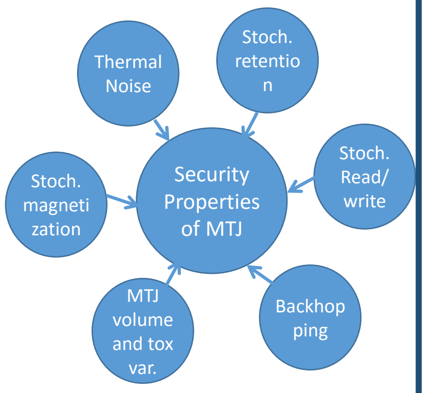
Stoch. DW pinning

Stoch. Write/ sense

NW roughness

Stoch. DW motion




Security Properties of MTJ

Thermal Noise

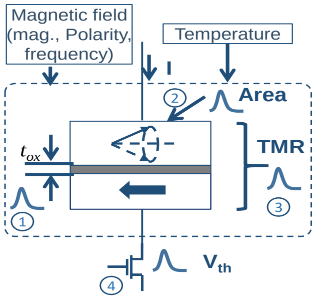
Stoch. retention

Stoch. Read/ write

Backhopping

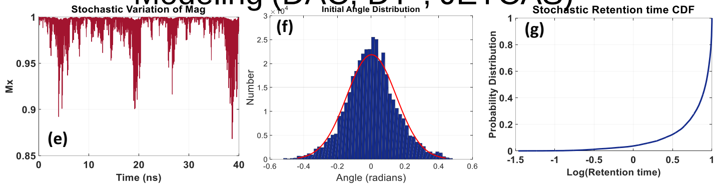
MTJ volume and tox var.

Stoch. magnetization

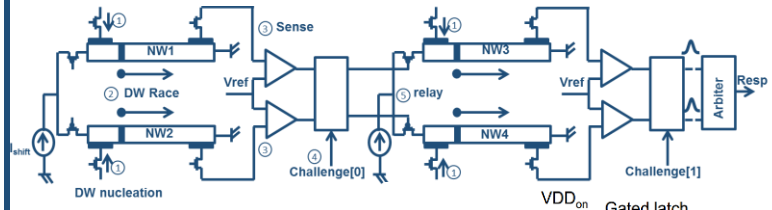


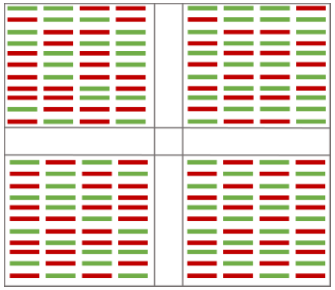
Collaboration with Univ. of Cincinnati

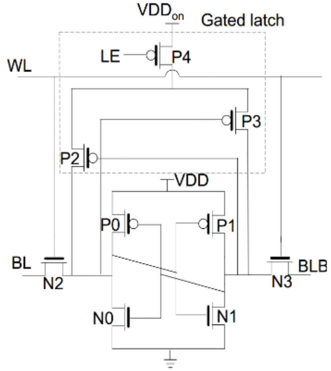
Modeling (DAC, DT*, JETCAS)



PUF (JETC, ISQED, HOST)





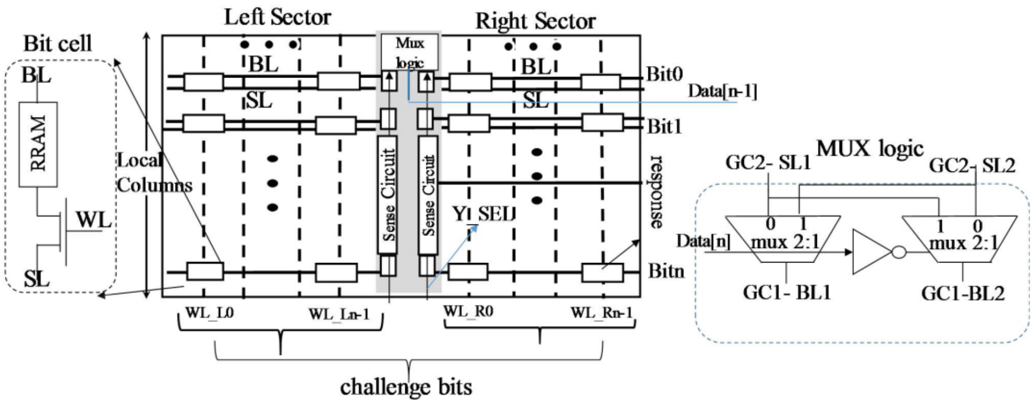


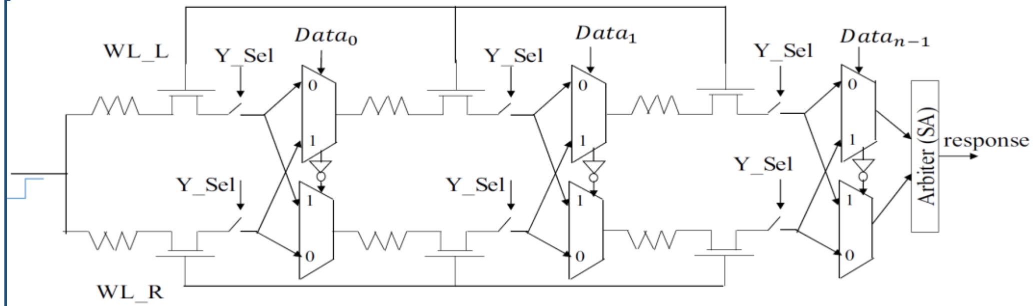
Collaboration with iastate

http://ieeexplore-spotlight.ieee.org/article/domain-wall-memory-could-be-the-next-hardware-security-hack/

ReRAM for Security

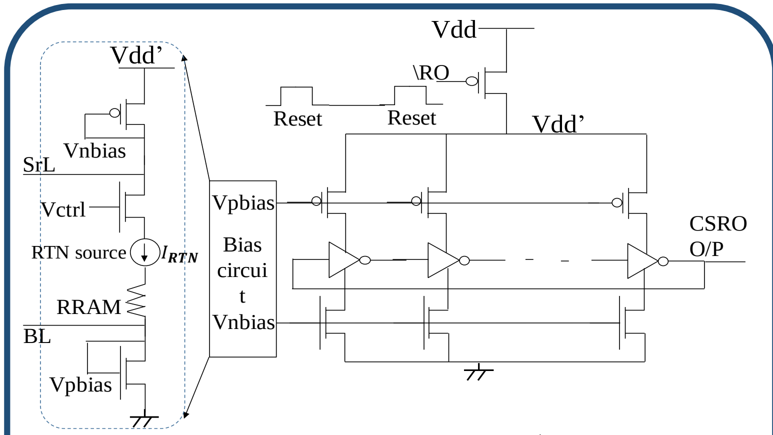
PUF

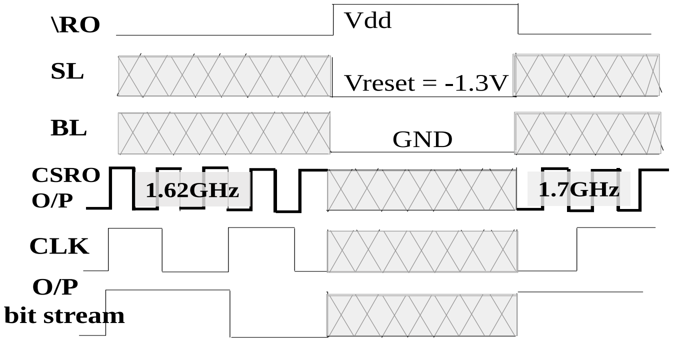


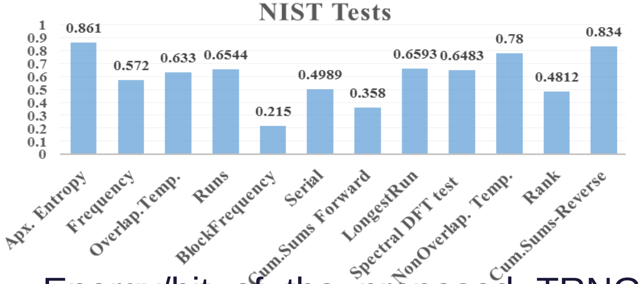


- Sense circuits in conventional memory architecture employed as arbiter
- Number of CRPs increase exponentially with array size
- Minimally invasive
- 0.13% intra HD and 51.3% inter HD with sufficient response randomness

TRNG







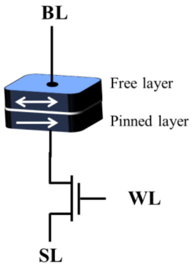
- Energy/bit of the proposed TRNG is 22.8fJ

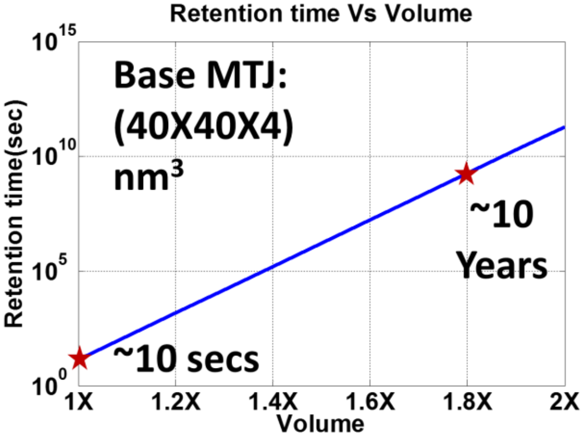
R. Govindaraj, ICCD, 2016

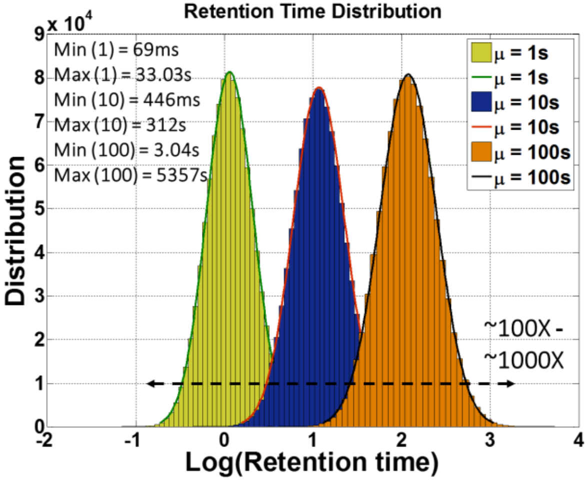
Outline

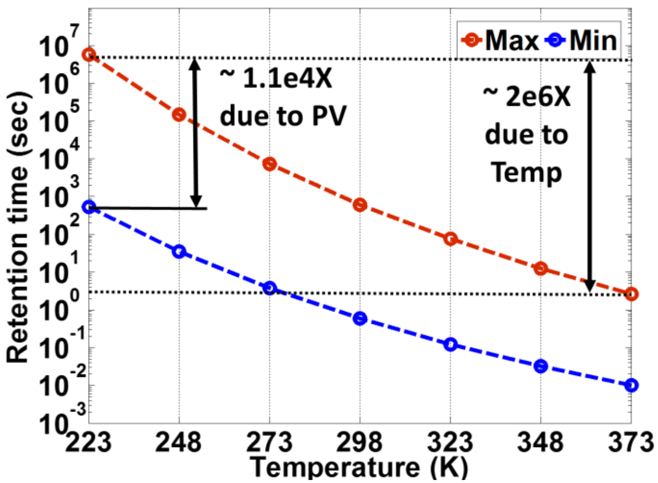
- Introduction and motivation
- Applications
  - Last level cache
  - Energy efficient computing
  - Security
- Challenges
  - Retention test
  - Long read/write latency
  - High asymmetric read/write current
- Solutions
  - Retention compression
  - Circuit to system synergistic design
  - Attack sensors and prevention
  - Sensing circuit
- Summary

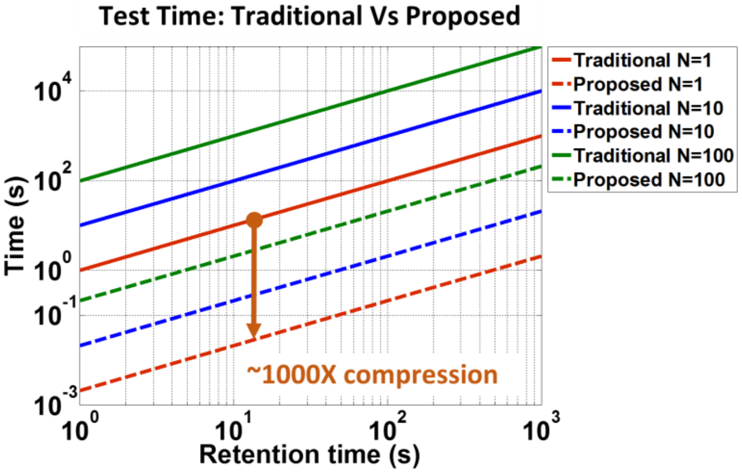
NVM Characteristics-High Retention Time









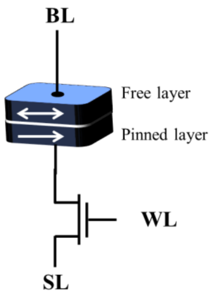


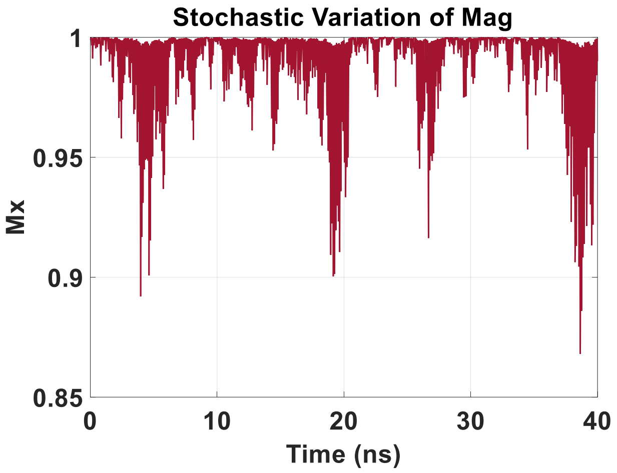
- High test time due to high retention time of bitcell
  - Cannot waive retention characterization

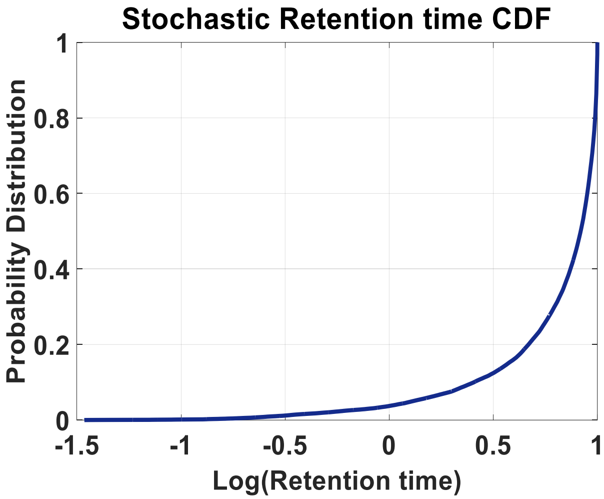
- Test question
  - How to reduce test time of NVMs?
  - How to characterize retention in presence of variations

A. Iyengar, Swaroop Ghosh, S. Srinivasana, "Retention testing of STTRAM", IEEE Design and Test (D&C), 2016

NVM Characteristics-High Retention Time (Stochastic Variation)



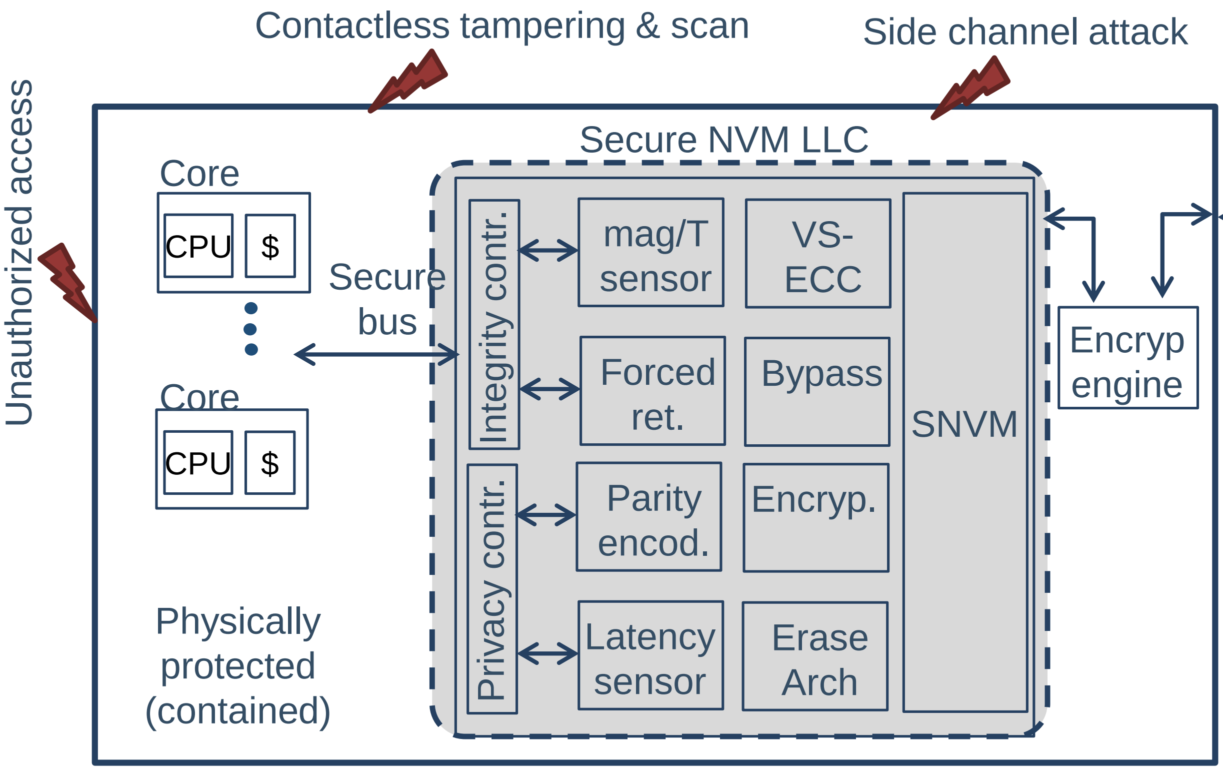


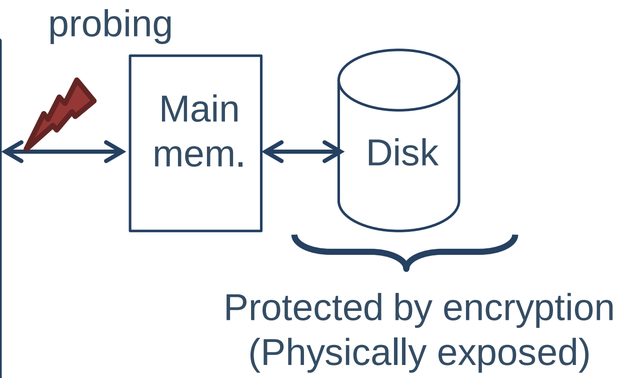


- Random variation in magnetization
  - Retention time of the same bit changes over time
  - Require multiple execution of test to guarantee retention time
- Test question
  - How to identify the worst case retention time?

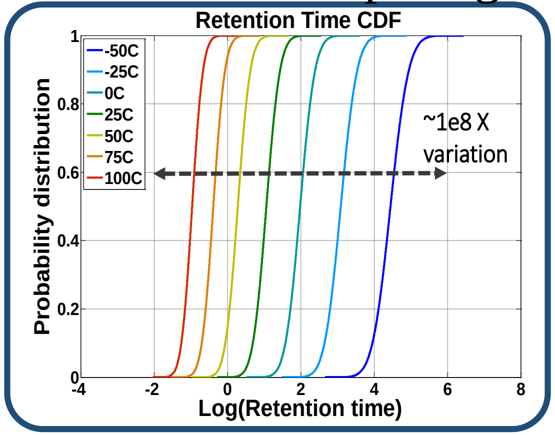
A. Iyengar, Swaroop Ghosh, S. Srinivasana, “Retention testing of STTRAM”, IEEE Design and Test (D&C), 2016

Data Privacy Issues in NVM





Thermal tampering



- Persistent data can be accessed between power cycle
- Short retention bitcells can be used to auto-erase the data in clear
  - Freezing of chip can modulate the retention
  - Encryption is latency sensitive
- New features are needed to secure the data

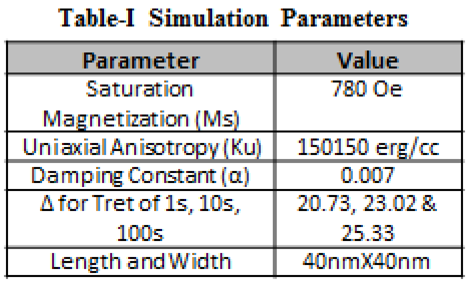
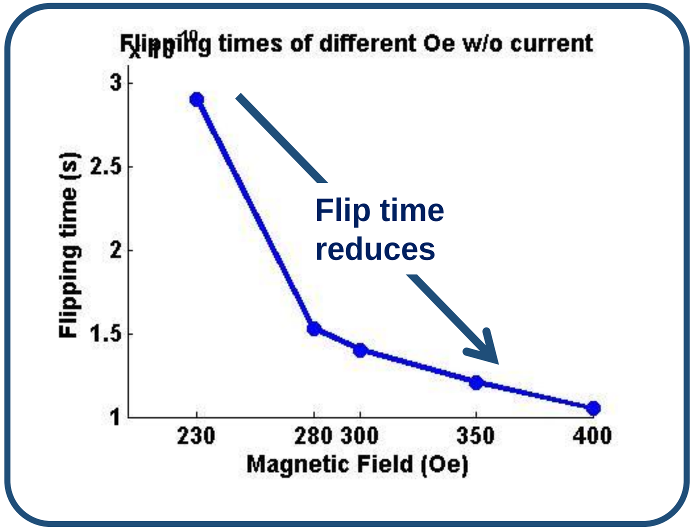


Table-I Simulation Parameters

Parameter	Value
Saturation Magnetization (Ms)	780 Oe
Uniaxial Anisotropy (Ku)	150150 erg/cc
Damping Constant (α)	0.007
Δ for Tret of 1s, 10s, 100s	20.73, 23.02 & 25.33
Length and Width	40nmX40nm

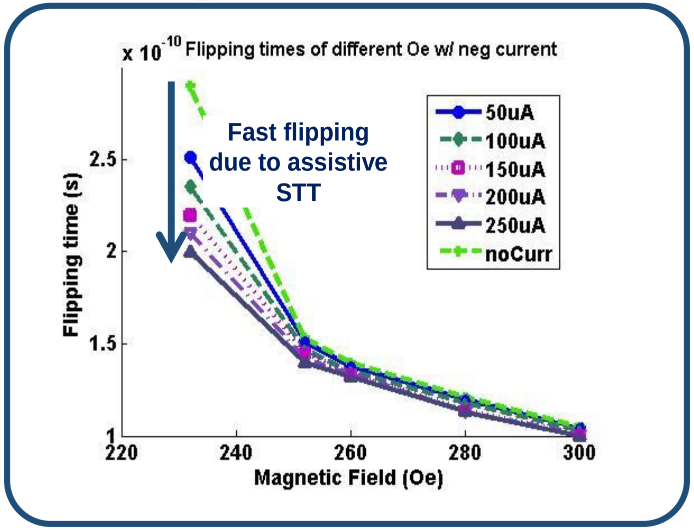
NVM Characteristics- Sensitivity to Ambient Parameters

Impact of DC field



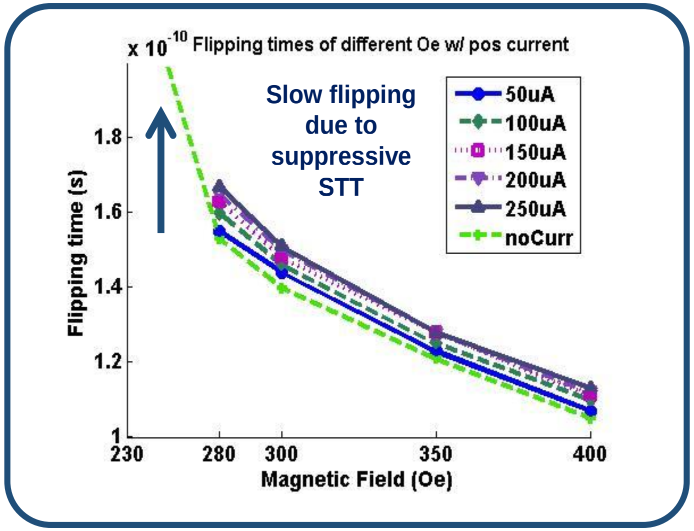
Flip time reduces

Impact of DC field



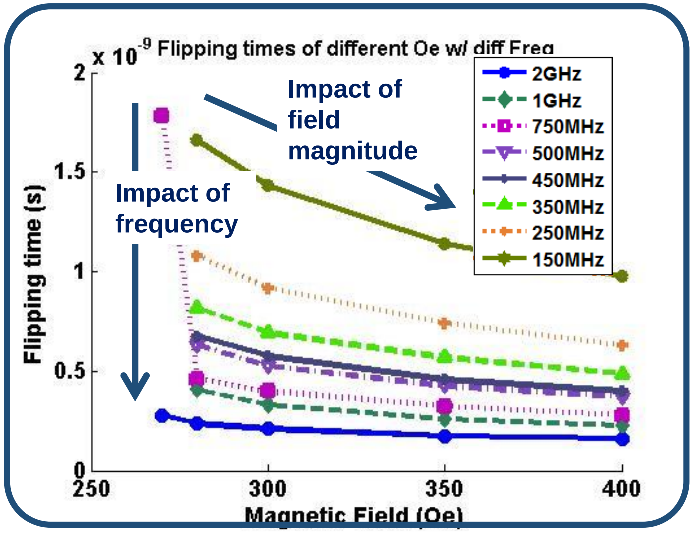
Fast flipping due to assistive STT

Impact of DC field



Slow flipping due to suppressive STT

Impact of AC field



Impact of field magnitude

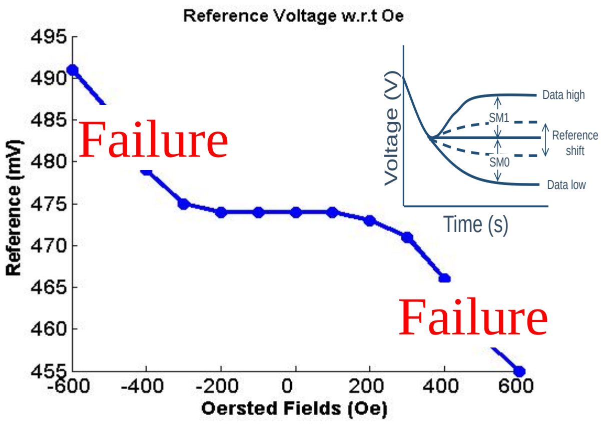
Impact of frequency

Jae-won Jang, Jongsun Park, Swaroop Ghosh, Swarup Bhunia, "Self-Correcting STTRAM under Magnetic Field Attacks", IEEE Design Automation Conference (DAC), 2015

- Test challenge
  - How to characterize magnetic tolerance

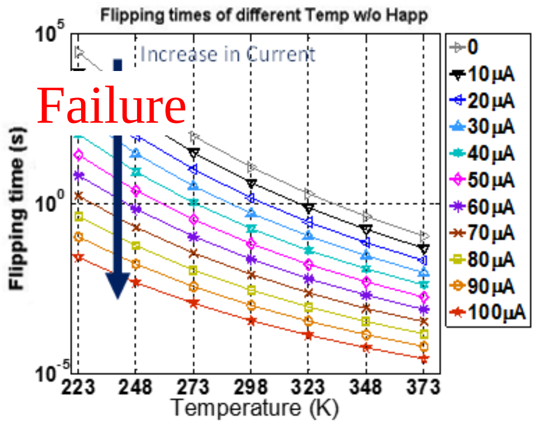
NVM Characteristics- Sensitivity to Ambient Parameters

Reference voltage shift



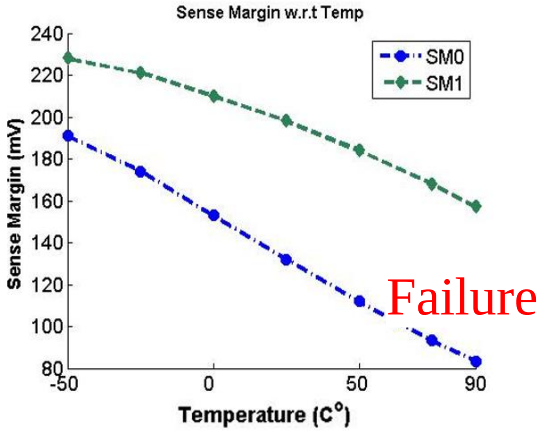
Failure

Failure

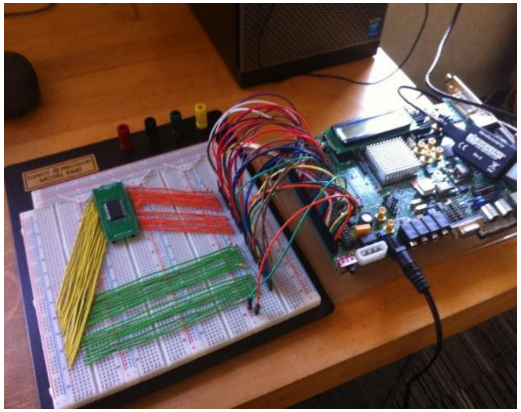


Failure

Experimental validation (HOST'16 demo)



Failure

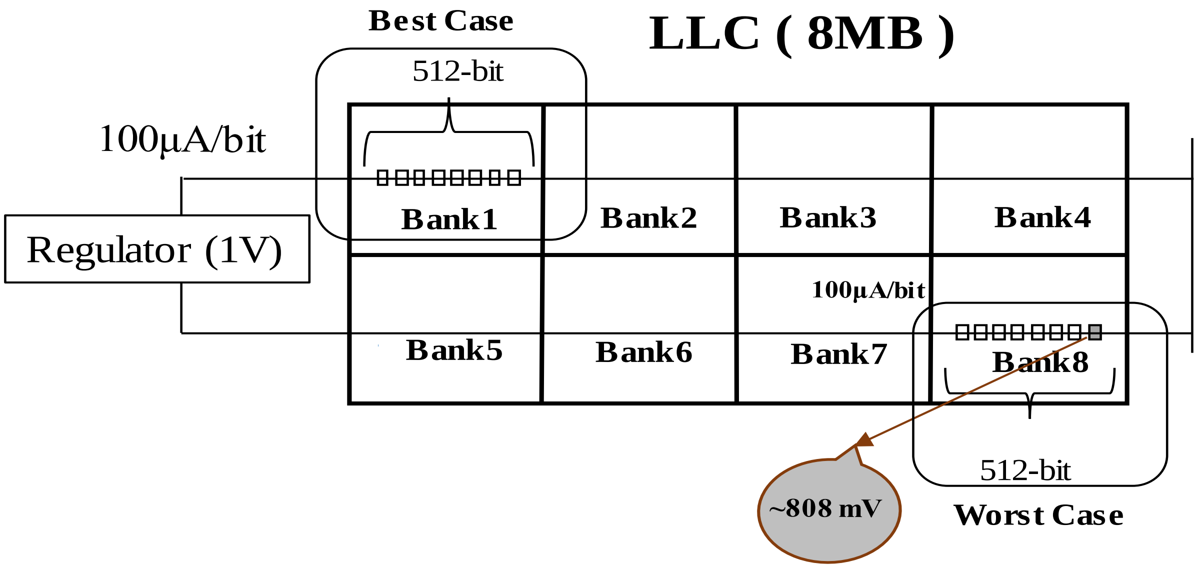


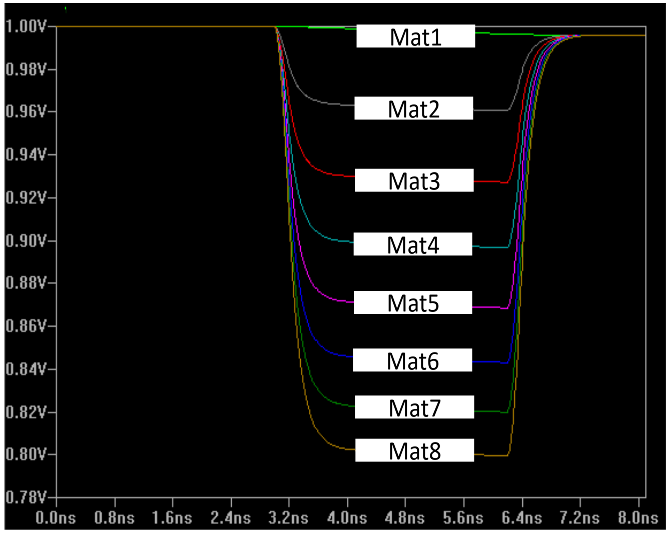
http://www.hostsymposium.org/hardware-demo-list.php

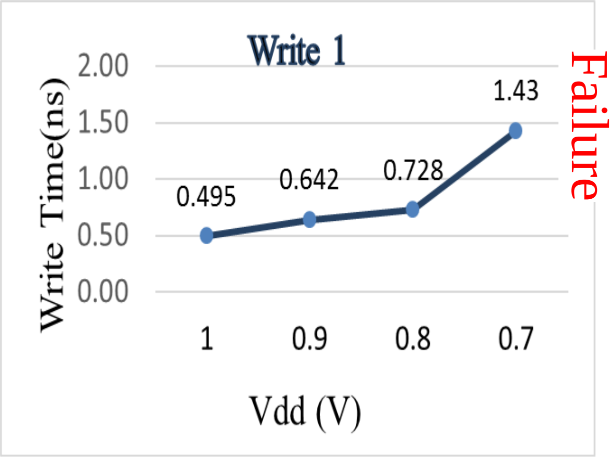
- Test question
  - How to characterize NVM under sensitivities
  - Can we detect security attacks?

Jae-won Jang, Swaroop Ghosh, ISLPED, 2016

NVM Characteristics-High and Asymmetric Write and Read Current





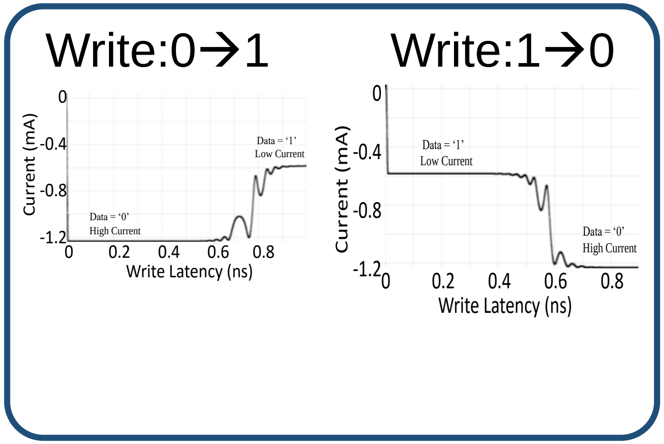


- High write current triggers droop
  - Depends on pattern
- Test question
  - Identifying test pattern to validate worst case droop

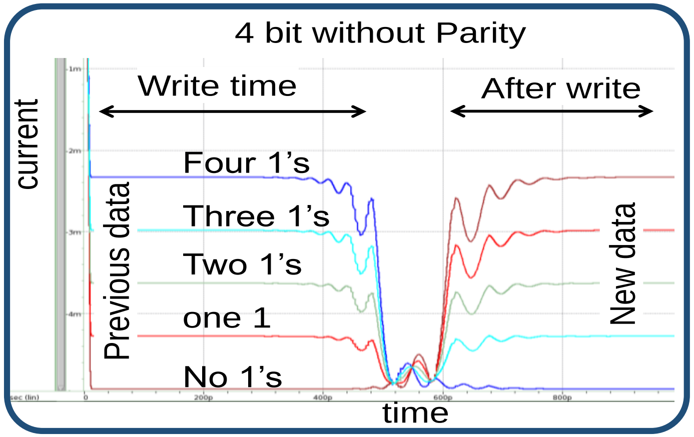
R. Aluru, Swaroop Ghosh, "Droop mitigating last level STTRAM cache", DATE, 2017

Security Implications- Privacy

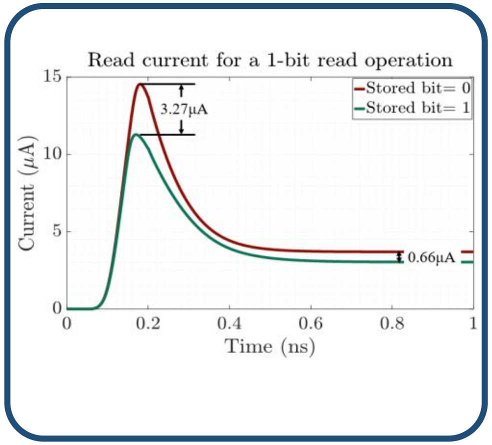
Asymmetric WR current/latency

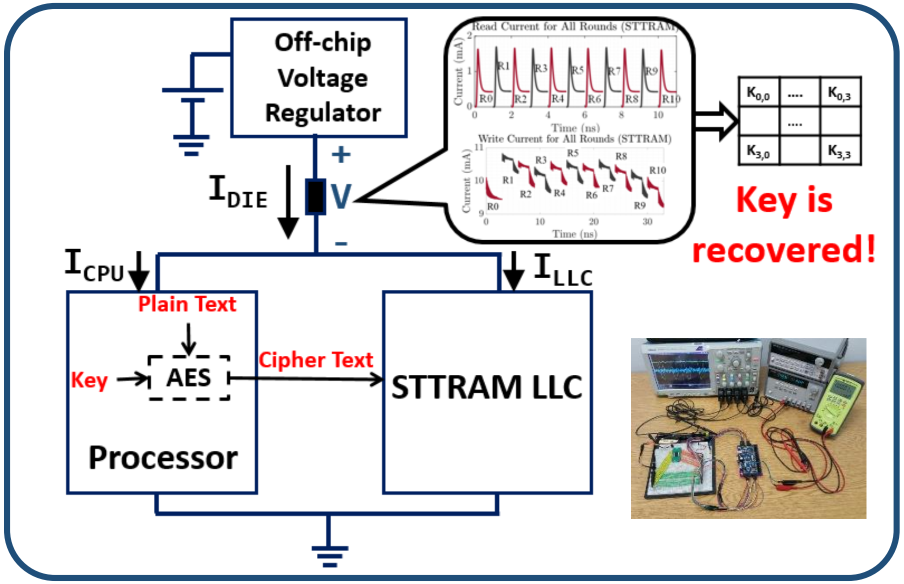


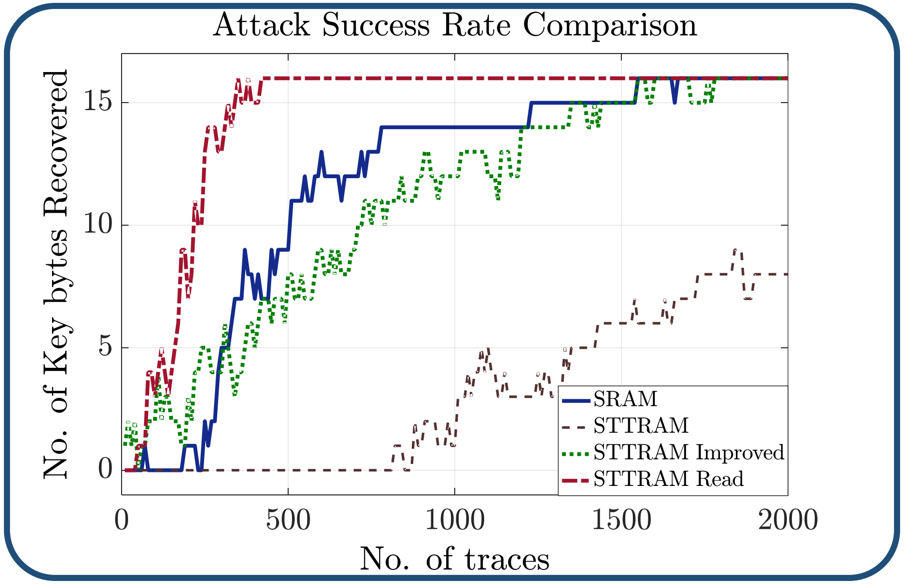
4-bit Write



Asymmetric RD current





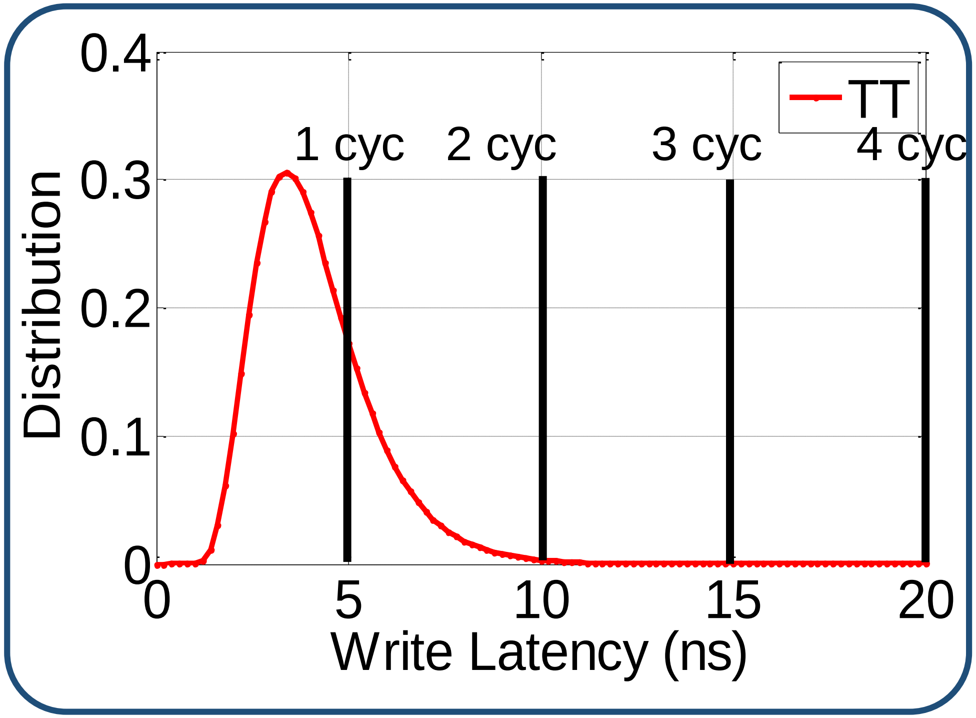


- Test question
  - Characterize asymmetric write current

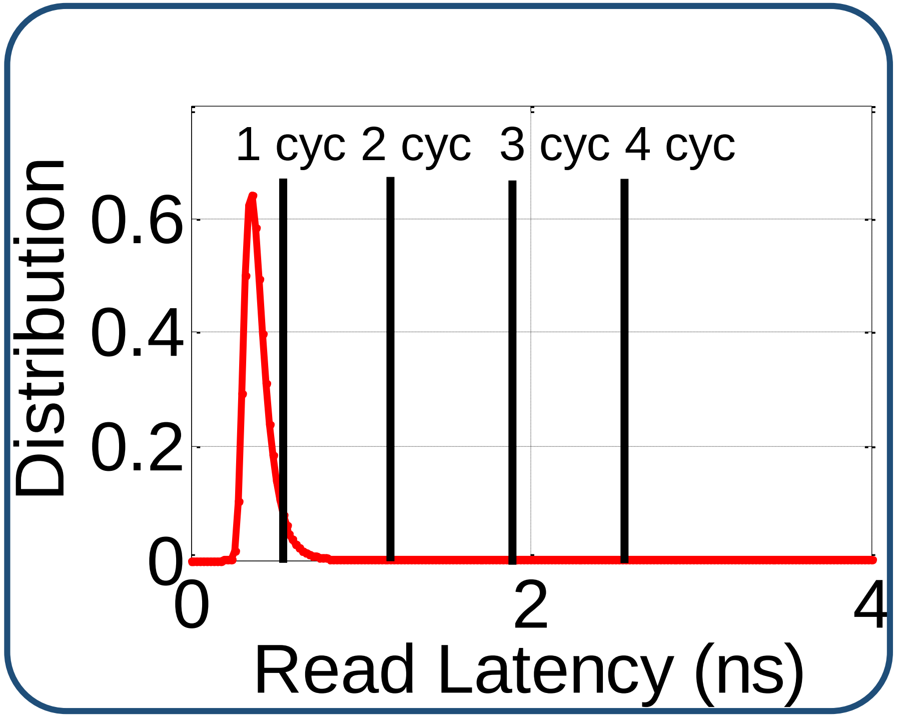
N. Rathi, S Ghosh, H. Naeimi, "Side Channel Attacks on STTRAM and Low-Overhead Countermeasures ", DFT 2016

NVM Characteristics-High and Asymmetric Write and Read Latency

High Write and Read current



Long Write and Read Latency

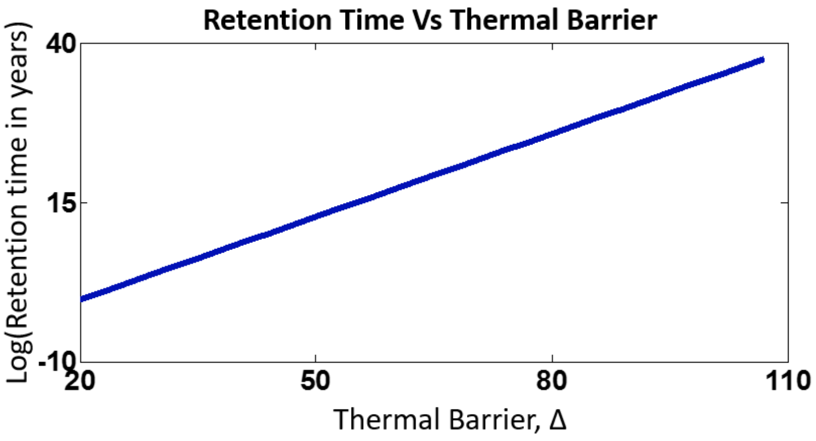


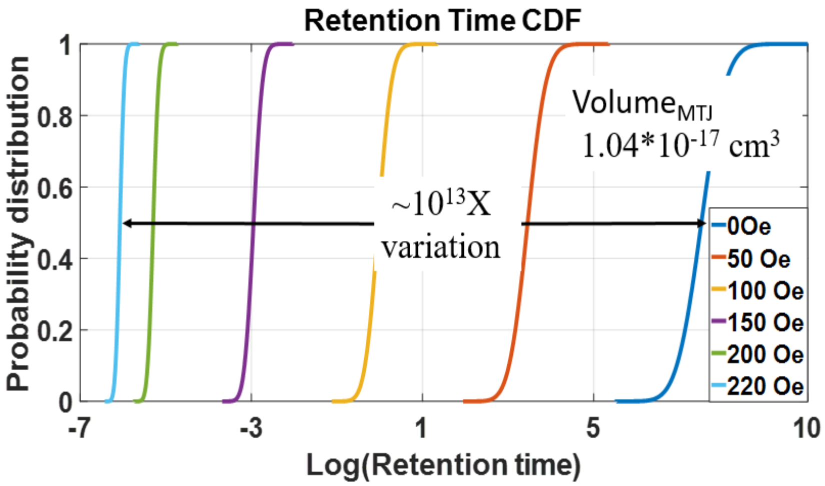
- Long tail of read and write latency
- Test question
  - How to characterize write and read latency at fast test time?

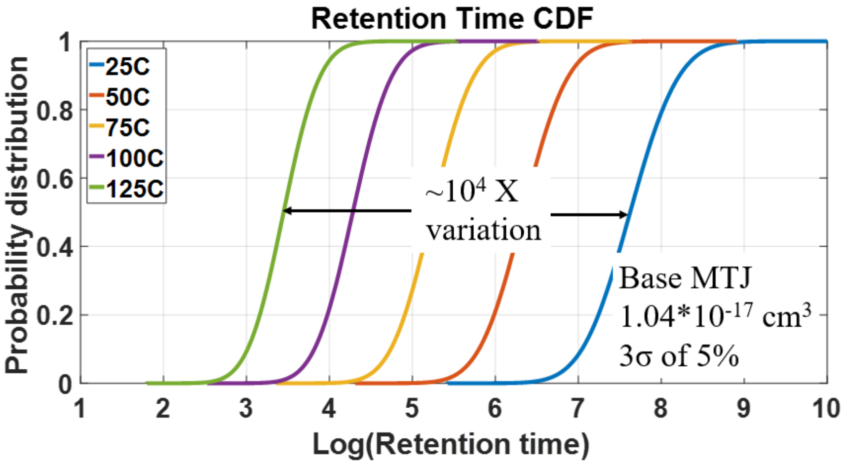
Outline

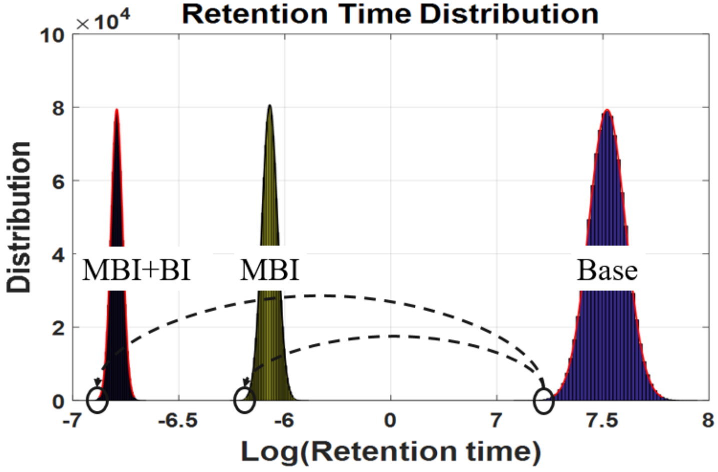
- Introduction and motivation
- Applications
  - Last level cache
  - Energy efficient computing
  - Security
- Challenges
  - Retention test
  - Long read/write latency
  - High asymmetric read/write current
- Solutions
  - Retention compression
  - Circuit to system synergistic design
  - Attack sensors and prevention
  - Sensing circuit
- Summary

Retention Testing using Test Time Compression





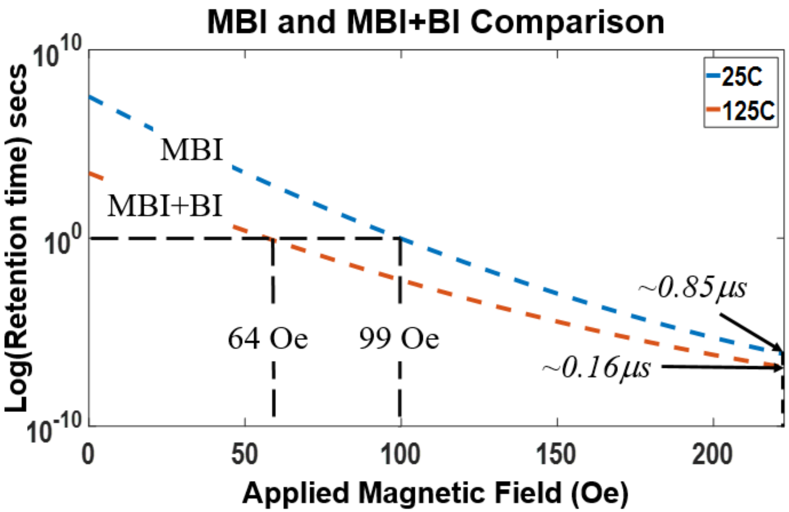


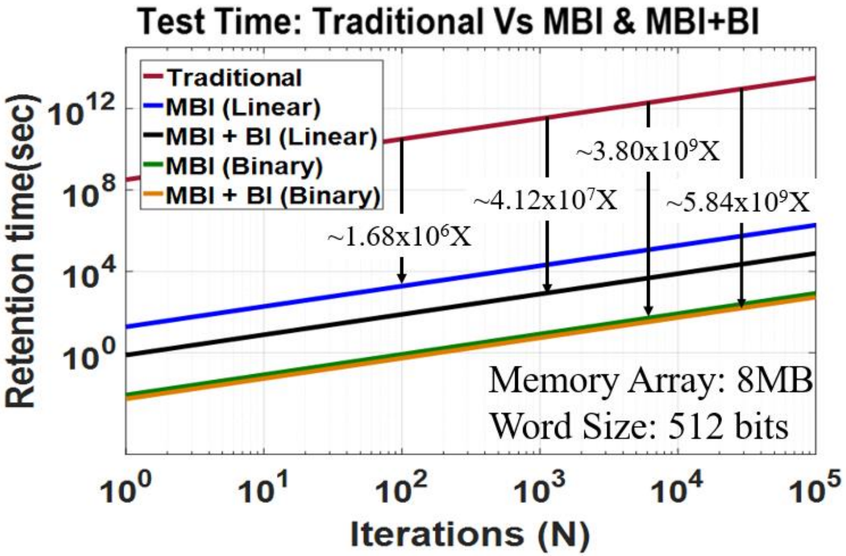


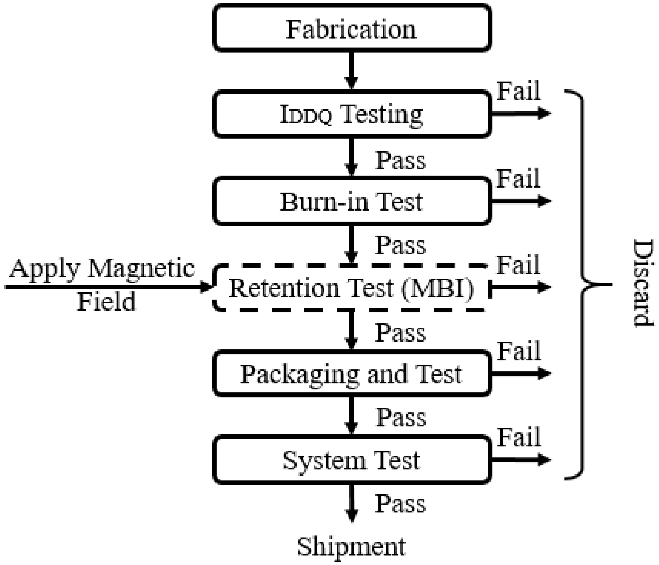
- Exploit STTRAM sensitivity to compress retention time
- Test time with lower retention is low

M. N.,I. Khan A. Iyengar, Swaroop Ghosh, “Magnetic burn-in for STTRAM retention testing”, DATE, 2017

Retention Testing using Test Time Compression






Fabrication

IDDQ Testing

Fail

Pass

Burn-in Test

Fail

Pass

Apply Magnetic Field

Retention Test (MBI)

Fail

Pass

Packaging and Test

Fail

Pass

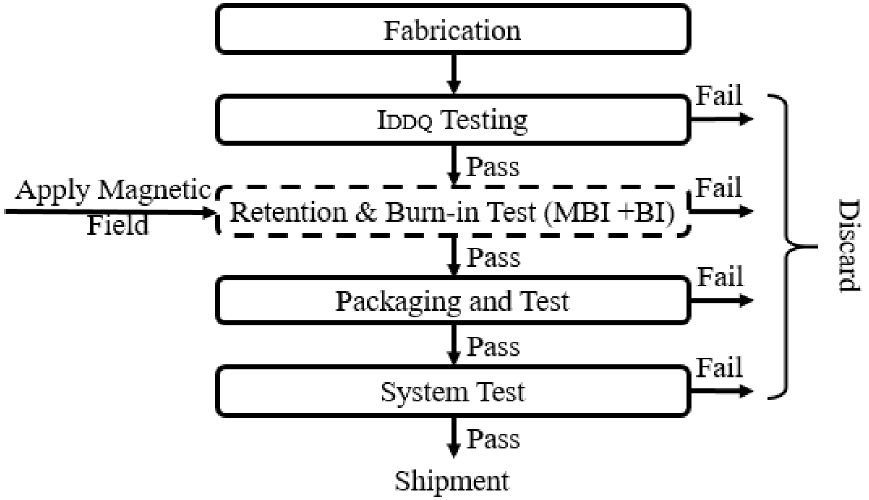
System Test

Fail

Pass

Shipment

Discard


Fabrication

IDDQ Testing

Fail

Pass

Apply Magnetic Field

Retention & Burn-in Test (MBI +BI)

Fail

Pass

Packaging and Test

Fail

Pass

System Test

Fail

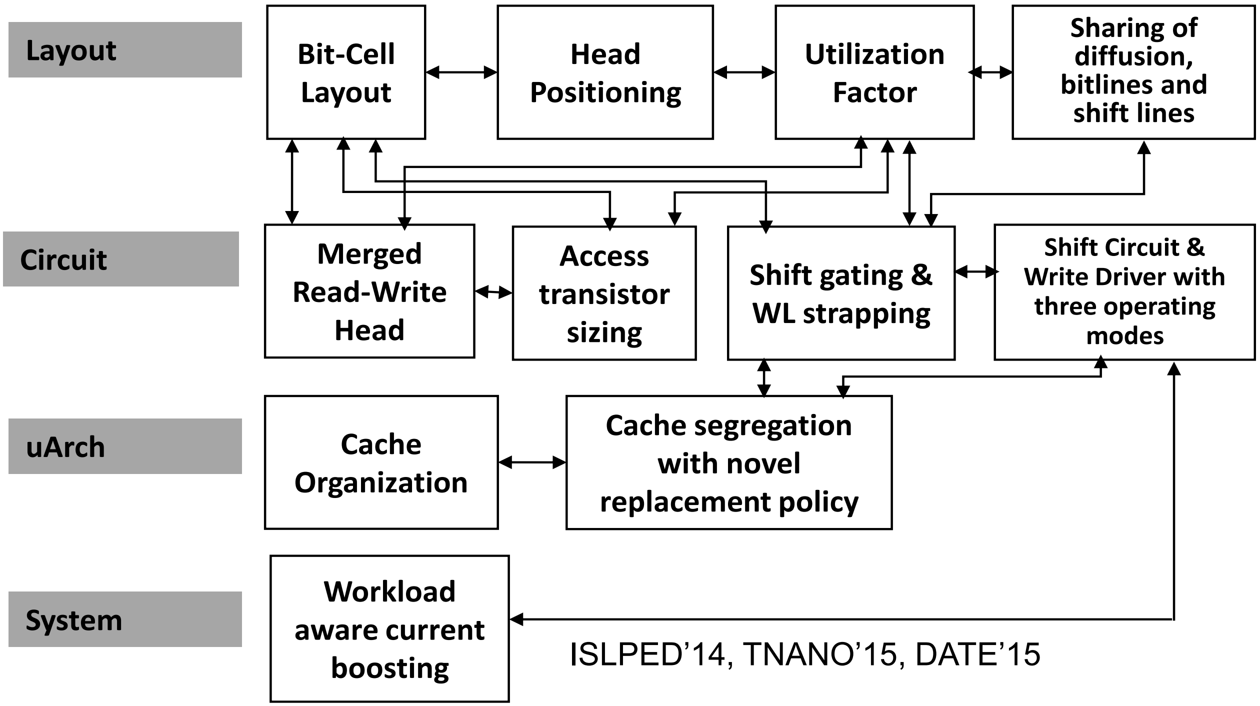
Pass

Shipment

Discard

M. N.,I. Khan A. Iyengar, Swaroop Ghosh, “Magnetic burn-in for STTRAM retention testing”, DATE, 2017

Energy-Efficient Memory Design



Layout

Circuit

uArch

System

Bit-Cell Layout

Head Positioning

Utilization Factor

Sharing of diffusion, bitlines and shift lines

Merged Read-Write Head

Access transistor sizing

Shift gating & WL strapping

Shift Circuit & Write Driver with three operating modes

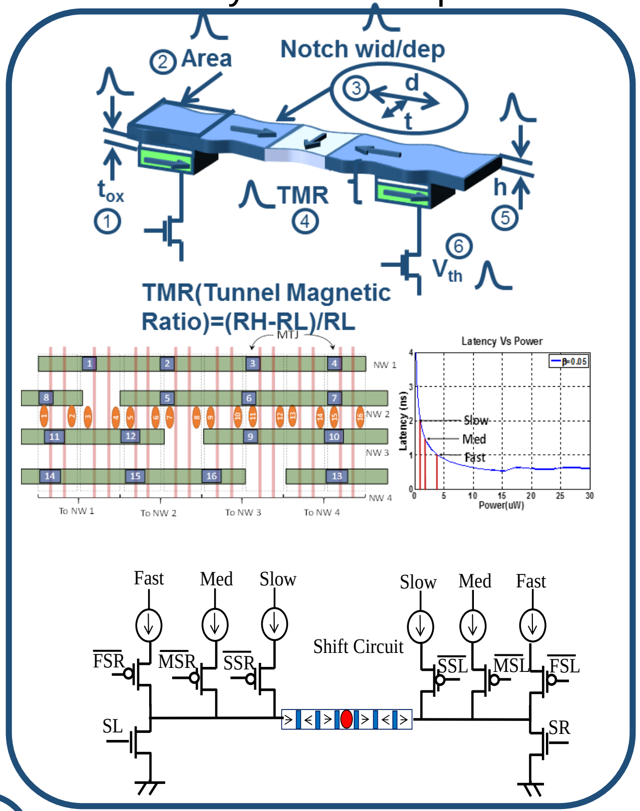
Cache Organization

Cache segregation with novel replacement policy

Workload aware current boosting

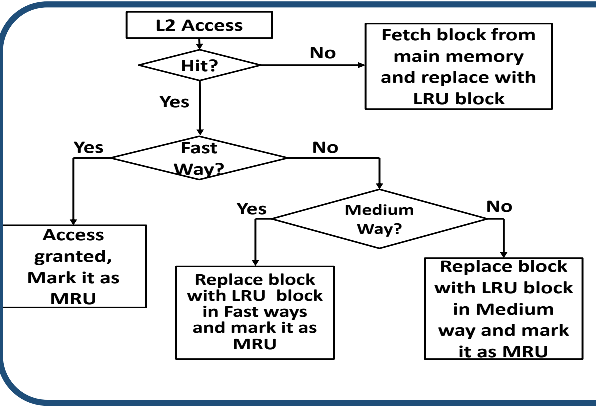
ISLPED’14, TNANO’15, DATE’15

Layout/circuit opt



TMR(Tunnel Magnetic Ratio)=(RH-RL)/RL

Cache segregation



L2 Access

Hit?

No

Fetch block from main memory and replace with LRU block

Yes

Fast Way?

Yes

No

Access granted, Mark it as MRU

Medium Way?

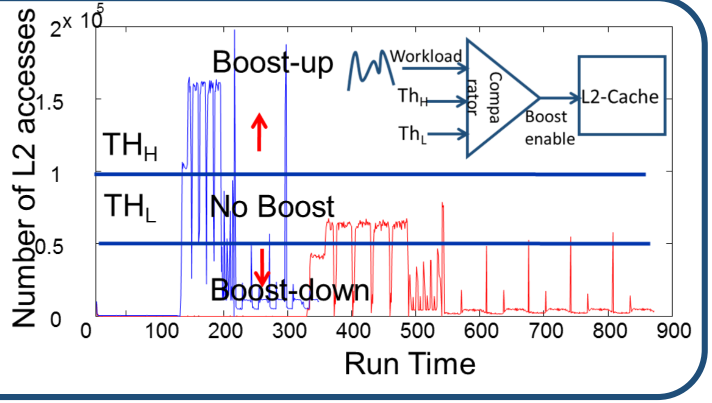
Yes

No

Replace block with LRU block in Fast ways and mark it as MRU

Replace block with LRU block in Medium way and mark it as MRU

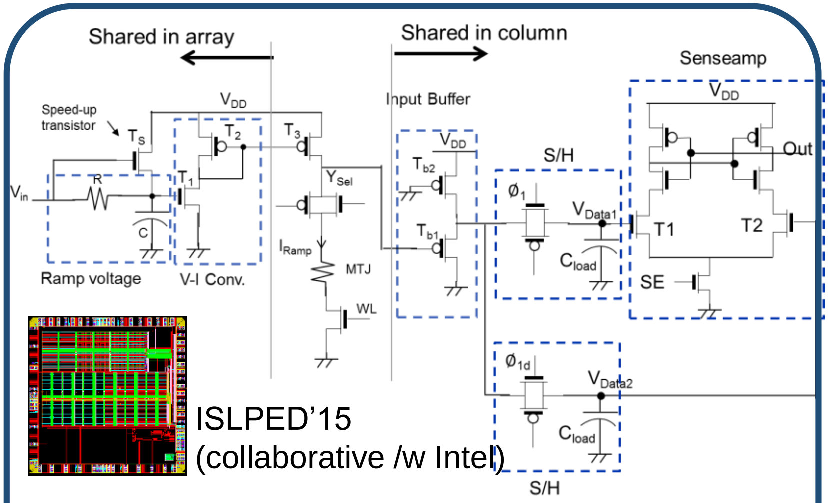
Workload adaptive modulation



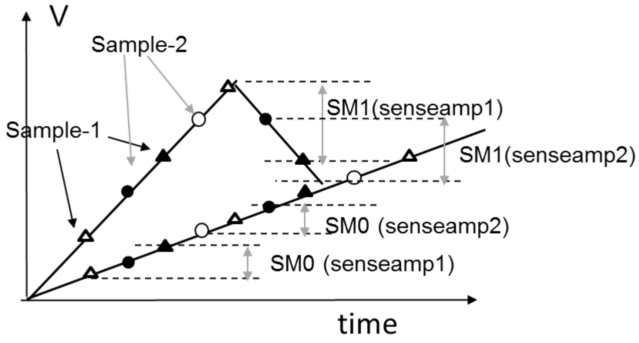
- ~30% perf improvement
- >10X power saving

Energy-Efficient Memory Design

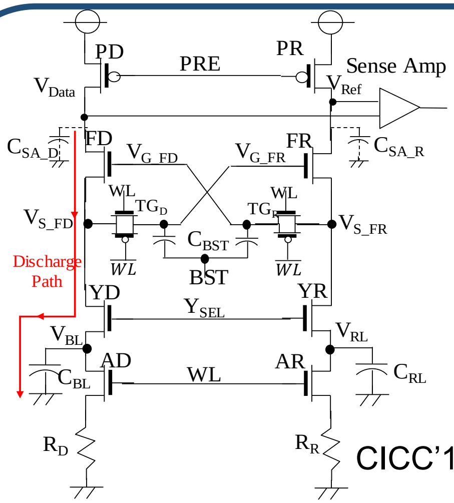
Slope sensing circuit

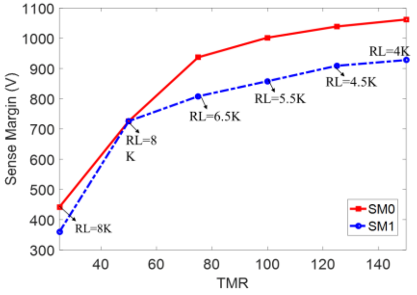


ISLPED’15
(collaborative /w Intel)

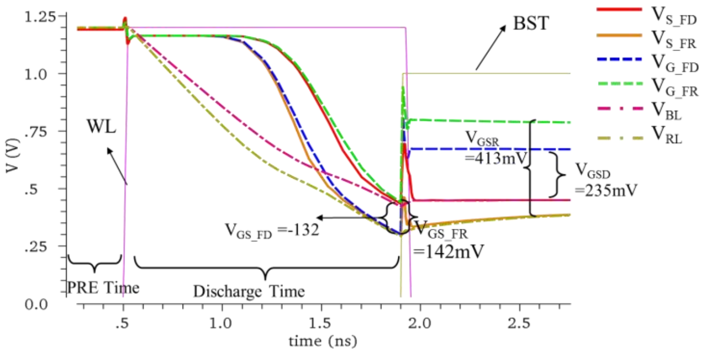


Sensing with column voltage boosting

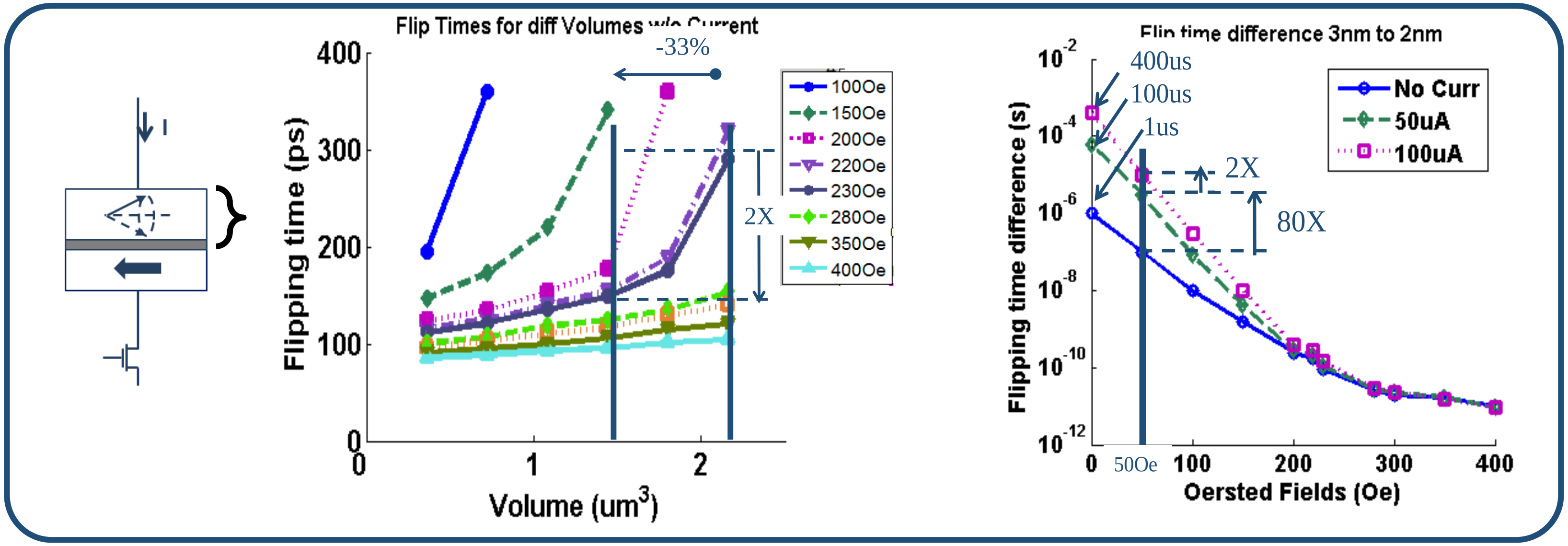




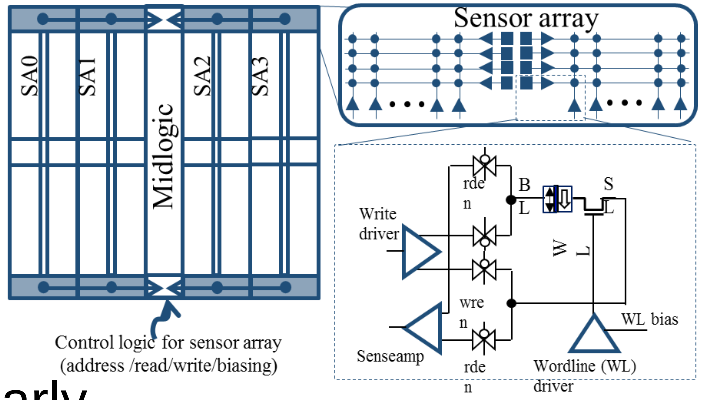
CICC’17 (under review, collaborative /w Intel)



Magnetic Field Sensor

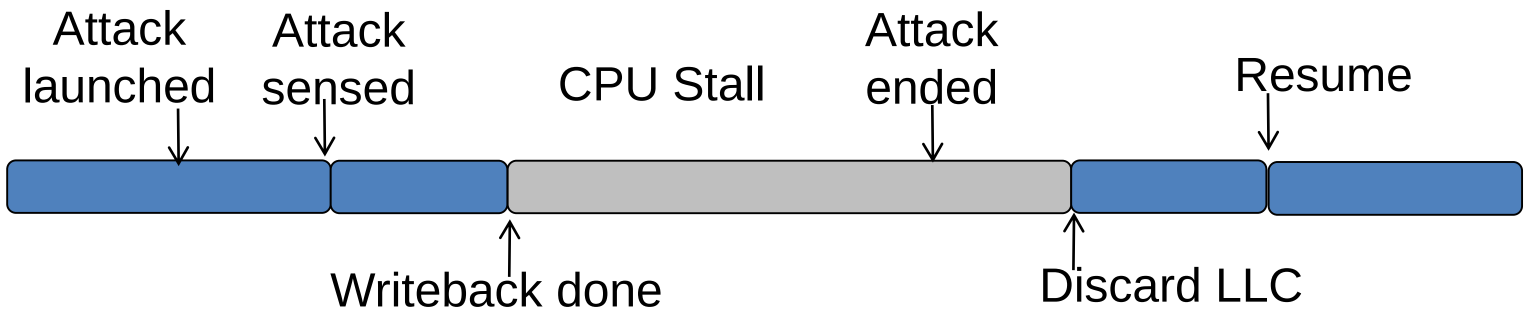


- Key requirements
  - Proactive sensing
  - Sense magnitude and polarity
- Sensor design
  - Small volume for early sensing
  - Weak write of sensor array to fail early
- Challenges
  - Identifying false alarms
  - Power consumption in sensor



Prevention (1)- Stalling

- Stall the CPU and wait till the attack is over
- For gradually ramping attack
  - Better than shutting down the entire system
  - Will not work for sudden attack since dirty data is corrupted



Attack launched

Attack sensed

CPU Stall

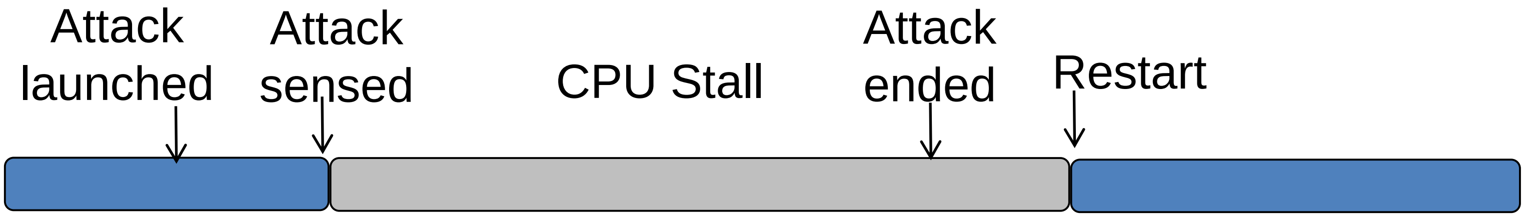
Attack ended

Resume

Writeback done

Discard LLC

- For sudden attack
  - Processor is restarted after the attack
  - Applications can resume from application level checkpointing



Attack launched

Attack sensed

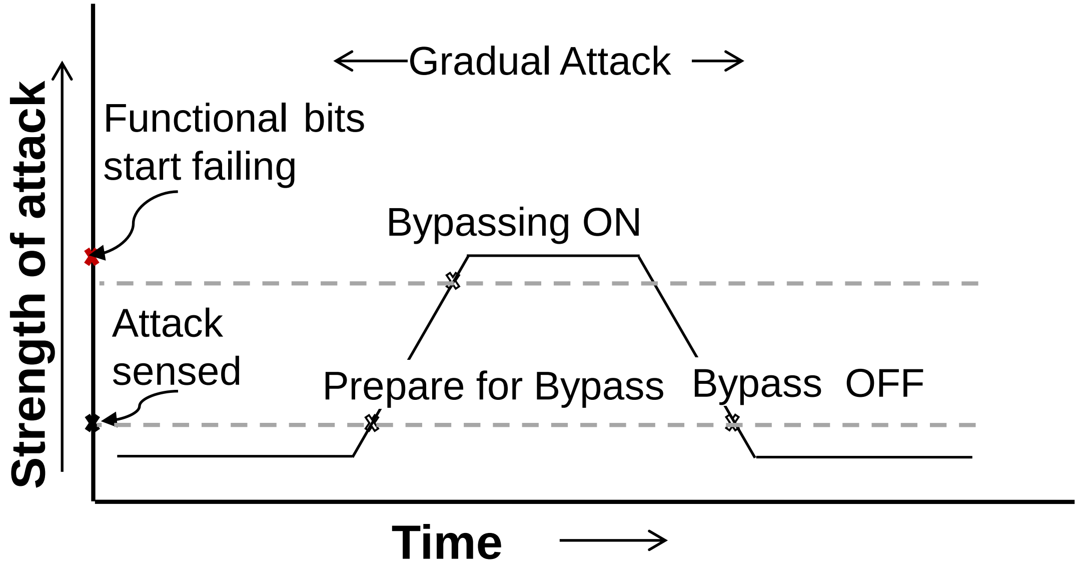
CPU Stall

Attack ended

Restart

- Both approaches disable computations during attack
  - Attacker can exploit these features to drain the battery

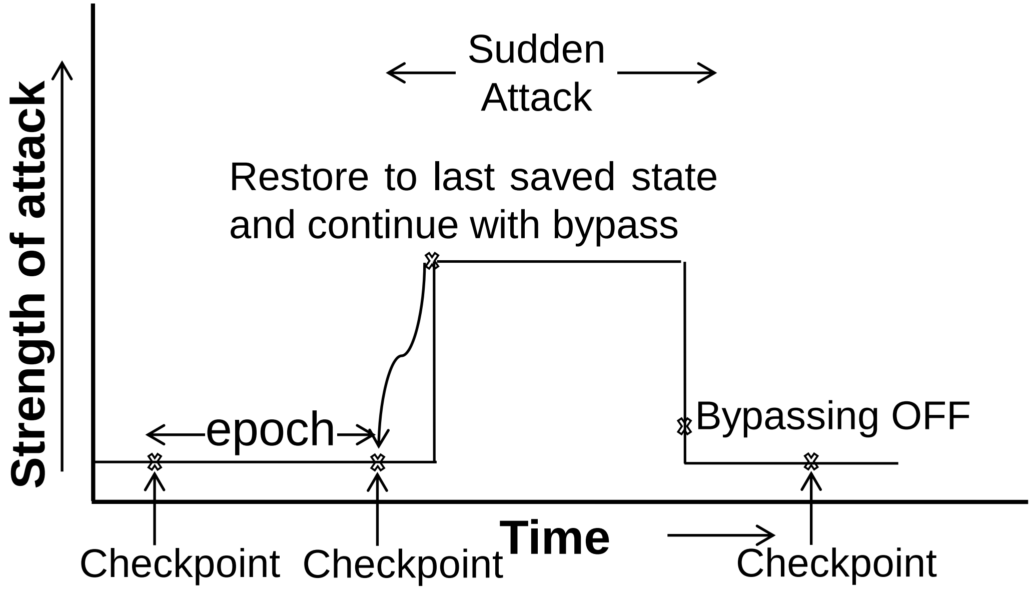
Prevention (2)- Cache Bypassing



- Key ideas
  - Since LLC is under attack, bypass it
  - Perform computation seamlessly without LLC
  - Update the main memory before starting bypass
  - Invalidate LLC before exiting bypass

*Nitin Rathi, Asmit De, *Helia Naeimi and Swaroop Ghosh, “Cache Bypassing and Checkpointing to Circumvent Data Security Attacks on STTRAM”, http://arxiv.org/abs/1603.06227

Prevention (3)- Checkpointing

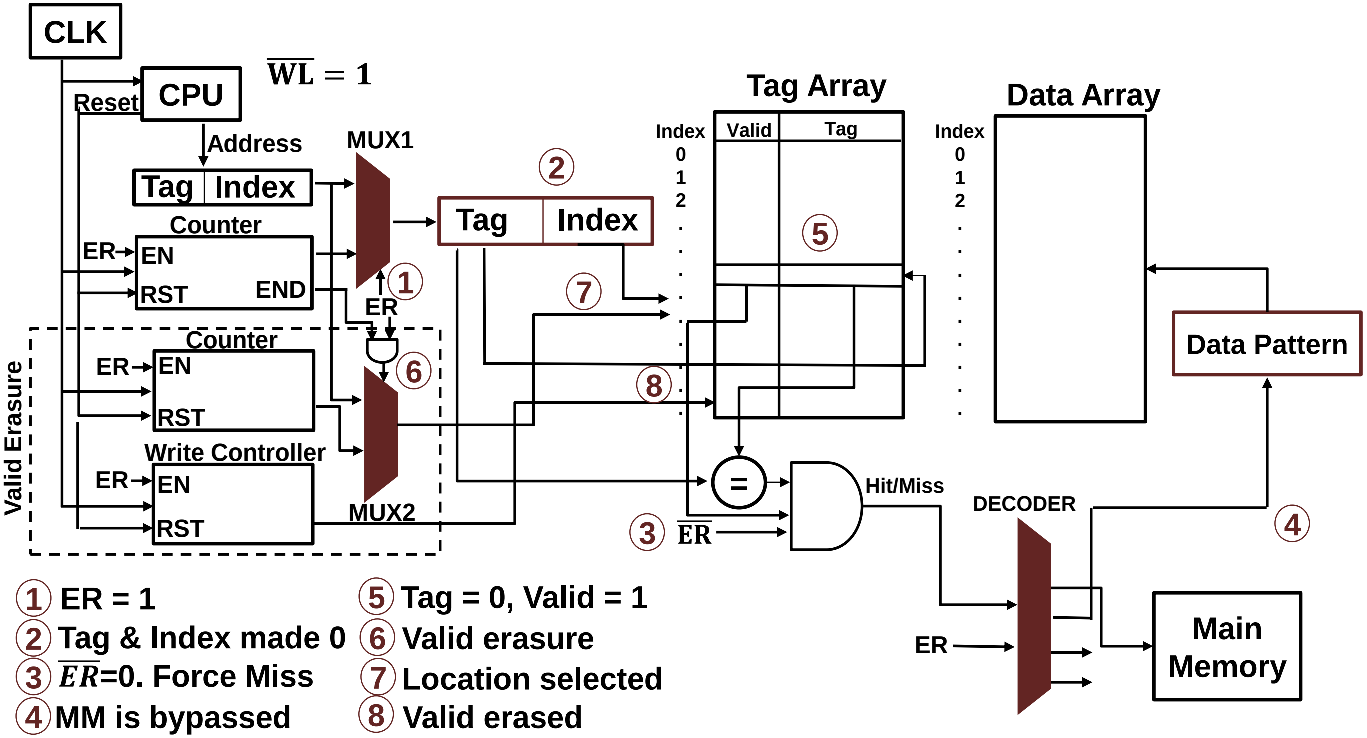


- Key ideas
  - Save processor state and update main memory periodically
  - If attack, go back to last saved state & start with LLC bypass
  - Can handle sudden corruption of memory

- Challenges
  - Need to stop main memory writeback between checkpoints
  - Performance loss due to checkpoint which depends on
    - Epoch
    - LLC full

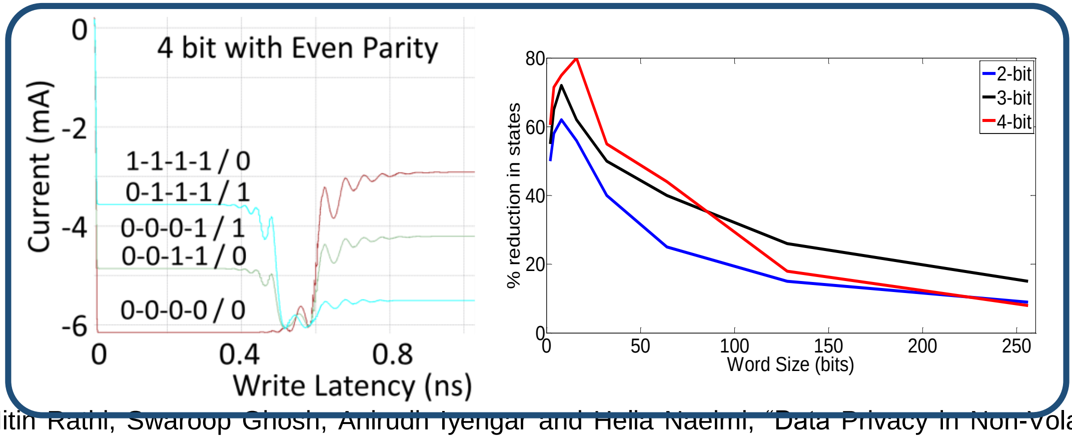
*Nitin Rathi, Asmit De, *Helia Naeimi and Swaroop Ghosh, “Cache Bypassing and Checkpointing to Circumvent Data Security Attacks on STTRAM”, http://arxiv.org/abs/1603.06227

Protecting Data Privacy

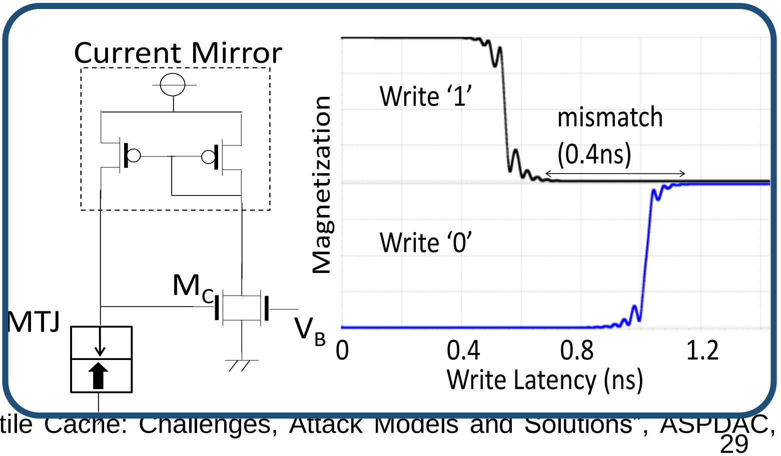


1 ER = 1
2 Tag & Index made 0
3 $\overline{ER}$=0. Force Miss
4 MM is bypassed
5 Tag = 0, Valid = 1
6 Valid erasure
7 Location selected
8 Valid erased

Adding dummy bits



Constant Current Write



Nitin Rathi, Swaroop Ghosh, Anirudh Iyengar and Helia Naeimi, "Data Privacy in Non-Volatile Cache: Challenges, Attack Models and Solutions", ASPDAC, 2016.

Conclusions

- Emerging NVMs are promising for broad range of applications
- NVMs possess unique challenges that could be design and security issues
- We proposed novel techniques to solve the challenges
- Proposed solutions are also applicable to other NVMs

Thank You!

Acknowledgements

LOGICS lab students, collaborators from Intel, Nanyang Tech Univ (NTU), Univ. of Florida, Iowa State, Univ. of Cincinnati and Korea Univ.

Graduate Students

PhD

A. Iyengar(‘13)

P:7, Pat:3

H. Motaman(‘13)

P:6, Pat:1

A. Saki (‘17)

P:0, Pat:0

R. Govindaraj(‘14)

P:3+2*, Pat:1

J. Jang(‘15)

P:4, Pat:1

MS

D. Vontela

P:1+2*, Pat:0

C. Lin

P:3, Pat:2

R. Aluru

P:1*, Pat:0

I. Reddy

P:2, Pat:1

Md. N. Khan (‘16)

P:1*, Pat:0

A. De (‘16)

P:1*, Pat:0

- Graduated
  - 5 MS: Kenneth Ramclam, Jae-won Jang, Radha Aluru, Deepak Vontela, Ithihasa Reddy
  - Published more than 40 IEEE papers in last 5 years

*LOGICS: Lab. Of Green and secure Integrated Circuits and Systems