

Ferroelectric NAND Flash: Scaling Storage Beyond 1000 Layer for AI era

Asif Khan

Associate Professor

School of Electrical and Computer Engineering

School of Materials Science and Engineering

Georgia Institute of Technology

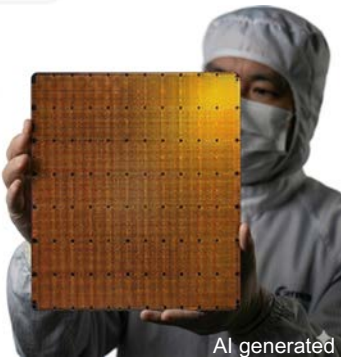
akhan40@gatech.edu

IEEE EDS Webinar, 05/06/2026

Foundational Memory & Storage Technologies

1

SRAM



Cerebras WSE-3

CAPACITY

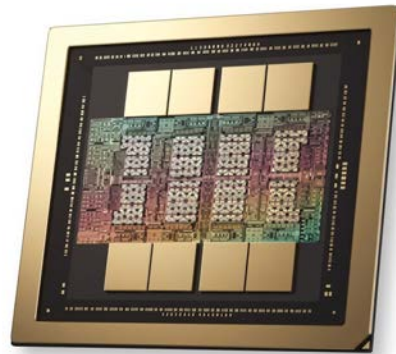
44 GB

BANDWIDTH

21 PB/s

2

DRAM and
High Bandwidth Memory



NVIDIA Rubin

CAPACITY

288 GB

BANDWIDTH

22 TB/s

3

NAND Flash
Storage



CAPACITY

30 – 245 TB

BANDWIDTH

~14 GB/s

4

Magnetic Hard Drives



CAPACITY

24 – 44 TB

BANDWIDTH

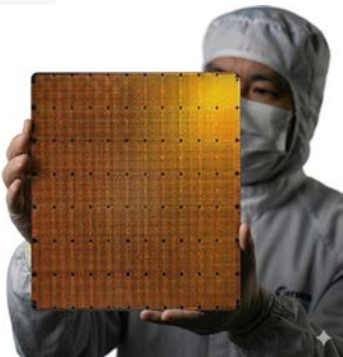
~300 MB/s

← faster, smaller | larger, slower →
Every AI system rides on this hierarchy.

Foundational Memory & Storage Technologies

1

SRAM



Cerebras WSE-3

CAPACITY

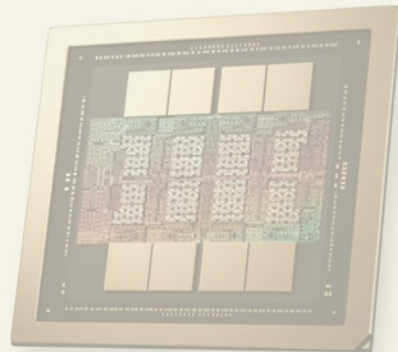
44 GB

BANDWIDTH

21 PB/s

2

DRAM and
High Bandwidth Memory



NVIDIA Rubin

CAPACITY

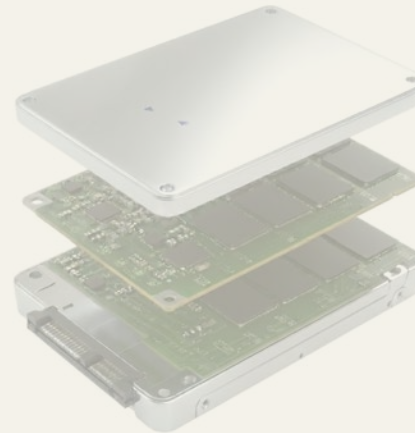
288 GB

BANDWIDTH

22 TB/s

3

NAND Flash
Storage



CAPACITY

30 – 245 TB

BANDWIDTH

~14 GB/s

4

Magnetic Hard Drives



CAPACITY

24 – 44 TB

BANDWIDTH

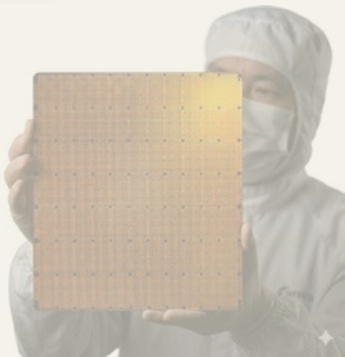
~300 MB/s

← faster, smaller | larger, slower →

Foundational Memory & Storage Technologies

1

SRAM



Cerebras WSE-3

CAPACITY

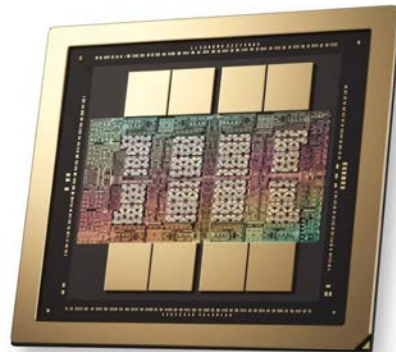
44 GB

BANDWIDTH

21 PB/s

2

DRAM and
High Bandwidth Memory



NVIDIA Rubin

CAPACITY

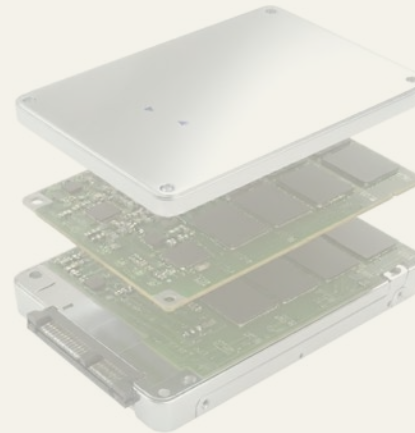
288 GB

BANDWIDTH

22 TB/s

3

NAND Flash
Storage



CAPACITY

30 – 245 TB

BANDWIDTH

~14 GB/s

4

Magnetic Hard Drives



CAPACITY

24 – 44 TB

BANDWIDTH

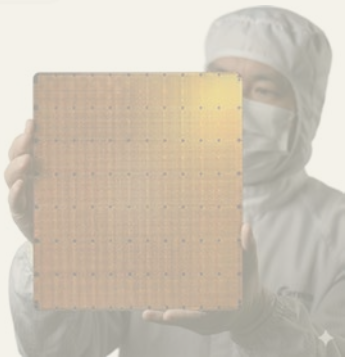
~300 MB/s

← faster, smaller | larger, slower →

Foundational Memory Technologies

1

SRAM



Cerebras WSE-3

CAPACITY

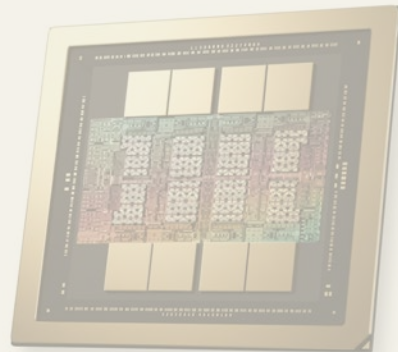
44 GB

BANDWIDTH

21 PB/s

2

DRAM and
High Bandwidth Memory



NVIDIA Rubin

CAPACITY

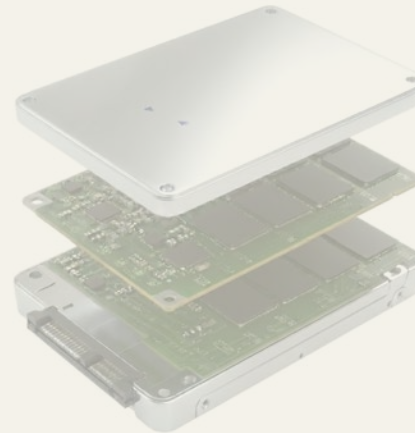
288 GB

BANDWIDTH

22 TB/s

3

NAND Flash
Storage



CAPACITY

30 – 245 TB

BANDWIDTH

~14 GB/s

4

Magnetic Hard Drives



CAPACITY

24 – 44 TB

BANDWIDTH

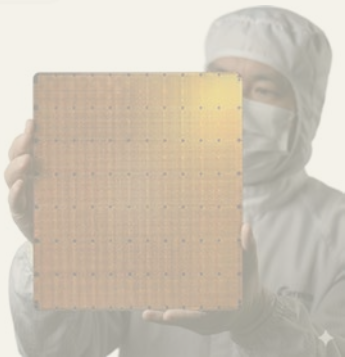
~300 MB/s

← faster, smaller | larger, slower →

Foundational Memory & Storage Technologies

1

SRAM



Cerebras WSE-3

CAPACITY

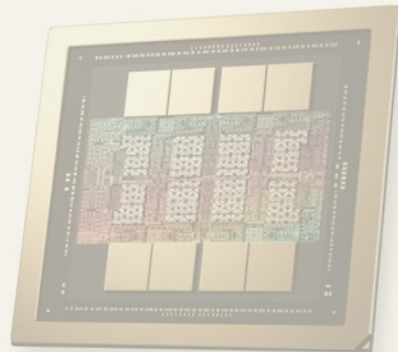
44 GB

BANDWIDTH

21 PB/s

2

DRAM and
High Bandwidth Memory



NVIDIA Rubin

CAPACITY

288 GB

BANDWIDTH

22 TB/s

3

NAND Flash
Storage



CAPACITY

30 – 245 TB

BANDWIDTH

~14 GB/s

4

Magnetic Hard Drives



CAPACITY

24 – 44 TB

BANDWIDTH

~300 MB/s

← faster, smaller | larger, slower →

Why NAND Flash Matters for AI

1

AI's appetite for data is exploding

10s of PB

TRAINING DATA

- Training datasets: PB-scale per frontier model
- Checkpoints: TBs written every few hours
- RAG vector databases: billions of embeddings

Storage demand ~doubling per year

2

NAND has overtaken HDD on every metric

~50x

BANDWIDTH vs HDD

- Capacity: 245 TB SSD vs 44 TB HDD
- Bandwidth: 14 GB/s vs 300 MB/s
- Density: 2.5 PB/2U vs ~30 TB/2U

One SSD replaces eight HDDs

3

NAND is climbing the memory hierarchy

Active

MEMORY TIER

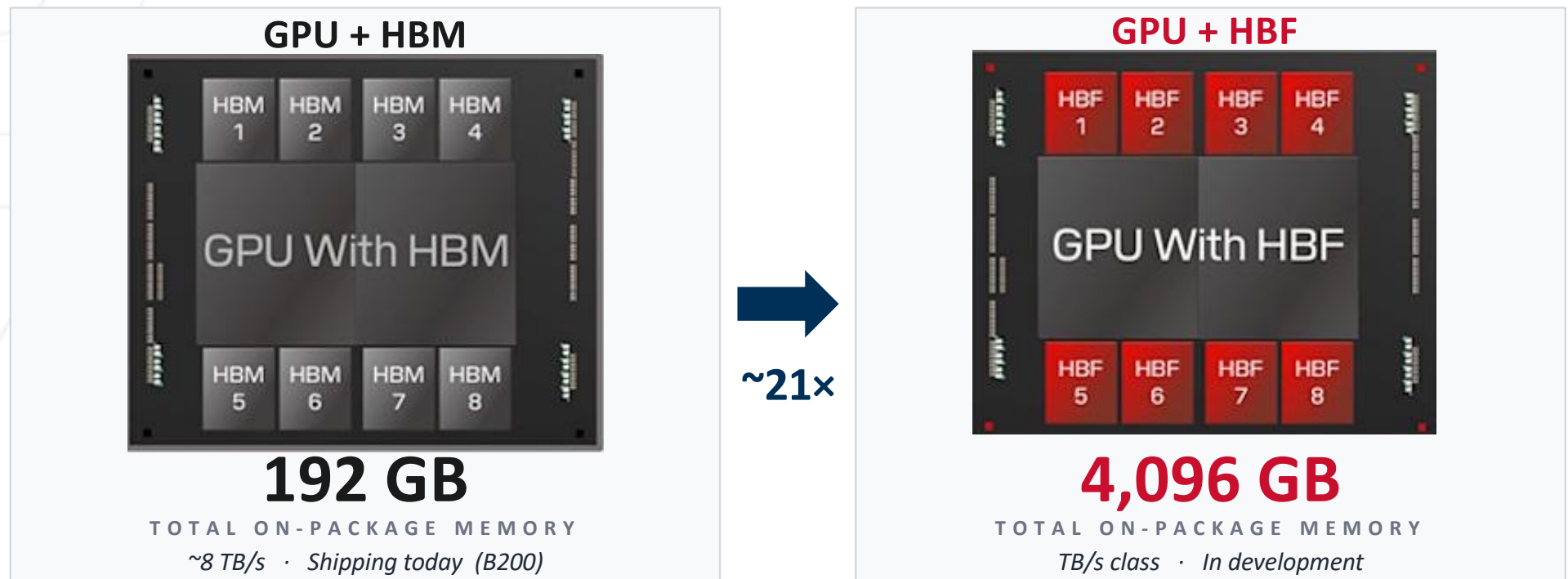
- KV cache offload for long-context inference
- Parameter streaming during training
- Disaggregated memory tiers via CXL: CXL-attached SSD pools sit between HBM and bulk storage, creating an entirely new memory tier

From cold storage to working memory

**NAND is no longer where the data sleeps.
It's increasingly part of where the AI thinks.**

What if NAND moved on Package?

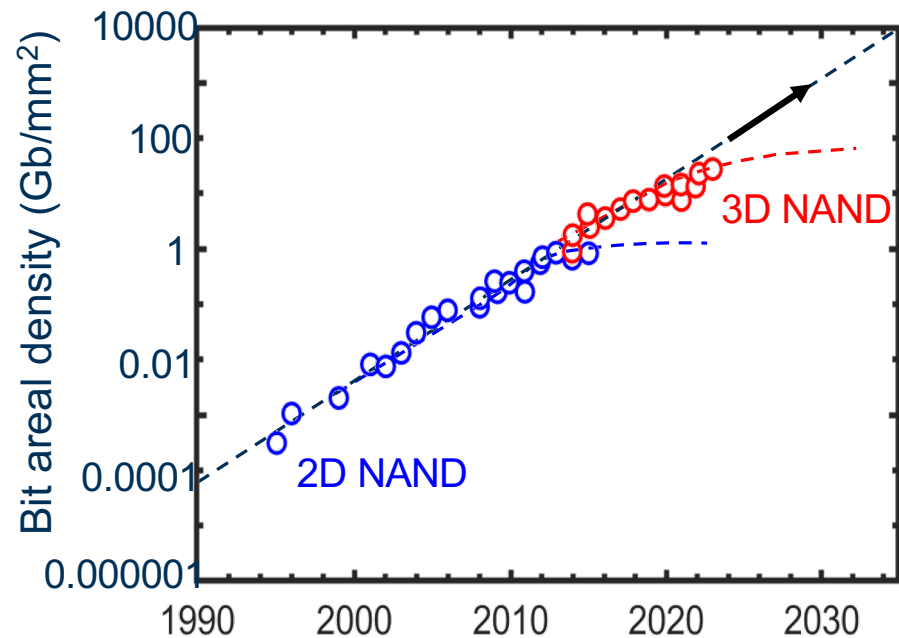
High Bandwidth Flash (HBF) stacks NAND in the HBM form factor — bulk capacity, right next to compute



Concept and figure inspired by SanDisk High Bandwidth Flash (HBF). KV caches, mixture-of-experts weights, long-context attention, large embedding tables

21× more memory, on package, at HBM-class bandwidth — if NAND can scale to deliver it

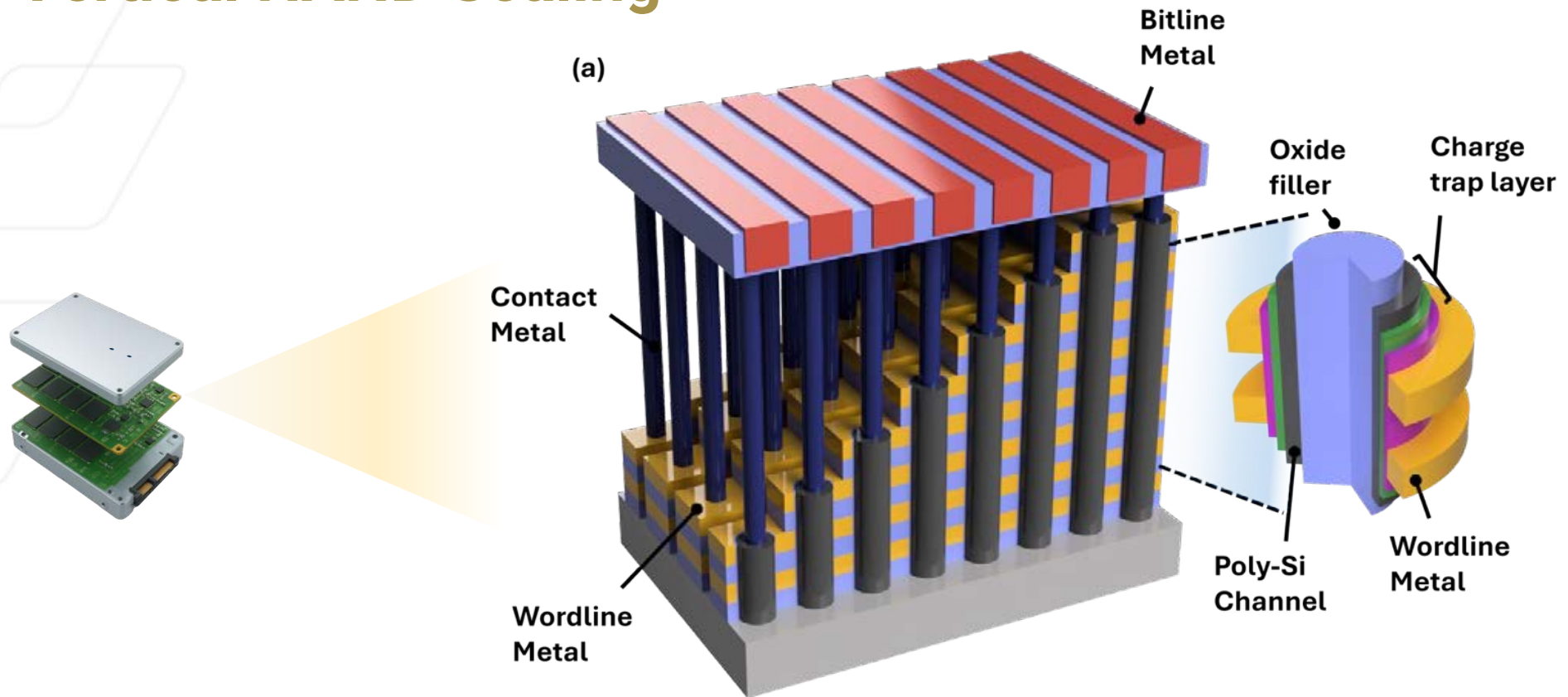
Vertical NAND Scaling



A. Goda, IEDM 2024

Vertical scaling has delivered 50× density per decade — and the climb continues

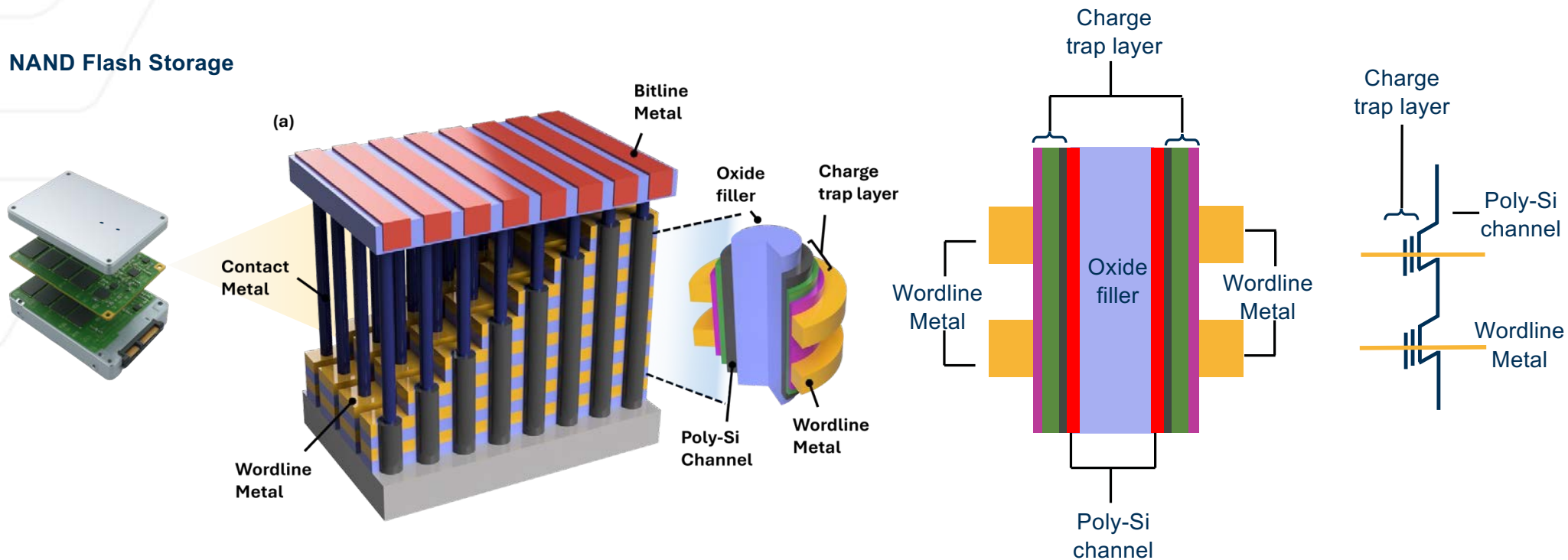
Vertical NAND Scaling



Vertical scaling has delivered 50× density per decade — and the climb continues

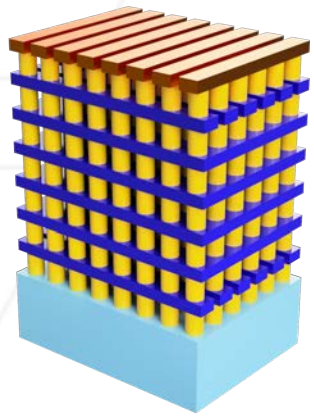
Vertical NAND Scaling

NAND Flash Storage

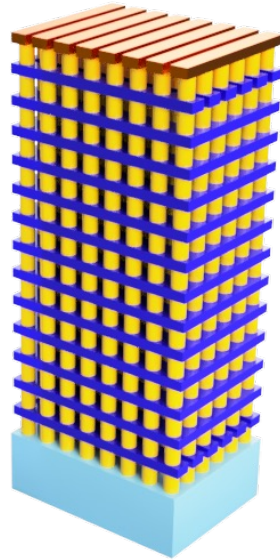


Relentless innovations have led to a 50x improvement in bit arial density per decade.

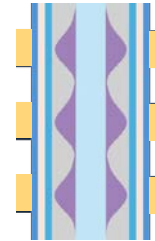
Vertical NAND Scaling



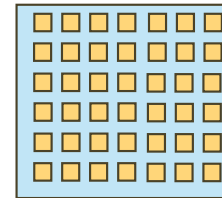
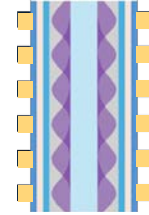
2014
First 3D NAND, Samsung
24 Layers, MLC



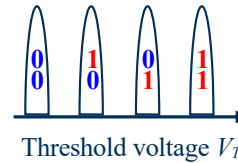
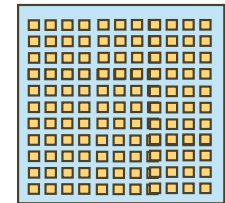
2022-2026
Production
Samsung 236L TLC
SK hynix 238L TLC → 321L (ramping 1H25)
Micron 276L TLC
Prototype / R&D demos:
SK hynix 321L sample (FMS 2023)
SK hynix 300+L demo (ISSCC 2023)
Kioxia / WD >300L demo (VLSI 2023)
Kioxia / Sandisk 1000L using wafer-to-wafer
Cu direct bonding (2026)



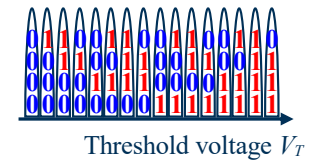
Z-scaling



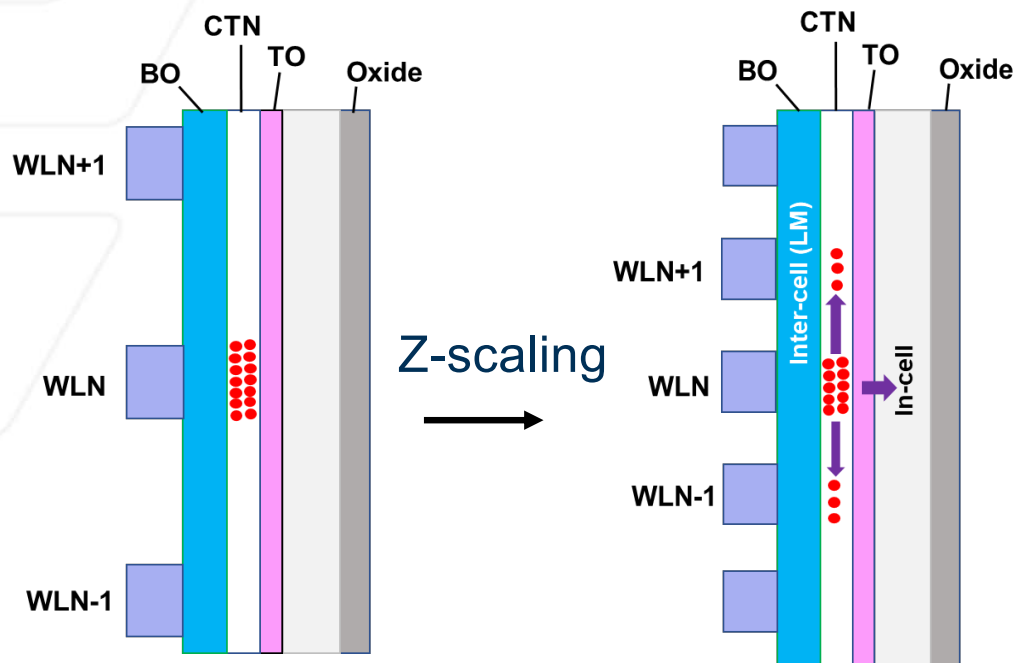
XY-scaling



Logical
scaling



Challenges of vertical NAND flash scaling



1 Lateral charge migration

Continuous SiN trap layer lets stored charge migrate between cells via diffusion and trap-to-trap transport.

2 High write voltage

Fowler–Nordheim programming demands ~ 20 V every cycle — large on-die charge pumps and accelerated tunnel-oxide wear-out. Increased power,

3 Coupling and disturb

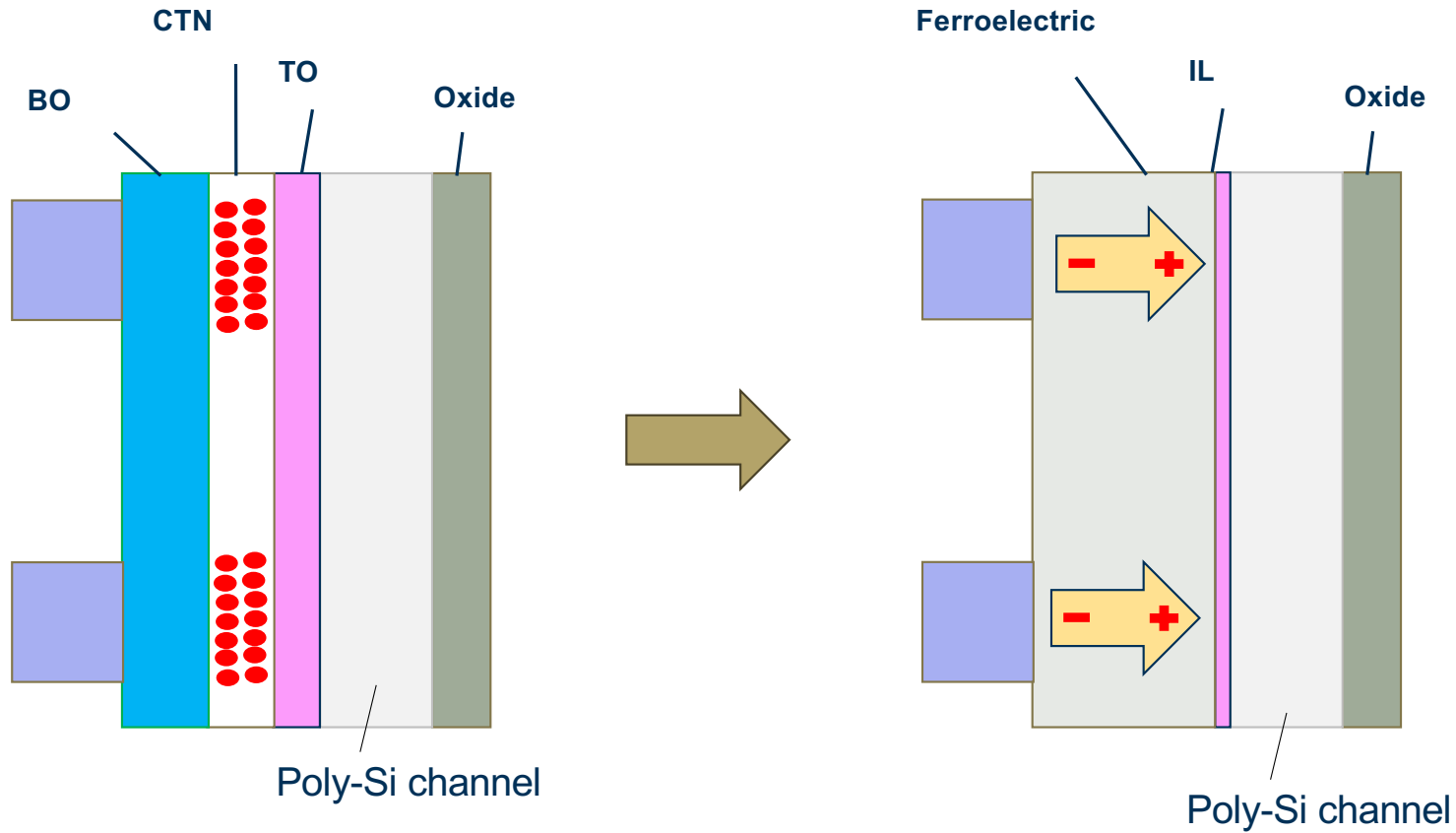
Electrostatic fringing fields perturb adjacent cells; pass-voltage stress accumulates with each read.

4 Variability and read margin

RTN, poly-Si grain variability, and channel taper compress threshold-voltage state distributions. Poly-Si channel resistance grows with cell count; grain boundaries limit mobility and uniformity.

Read sensing margin shrinks

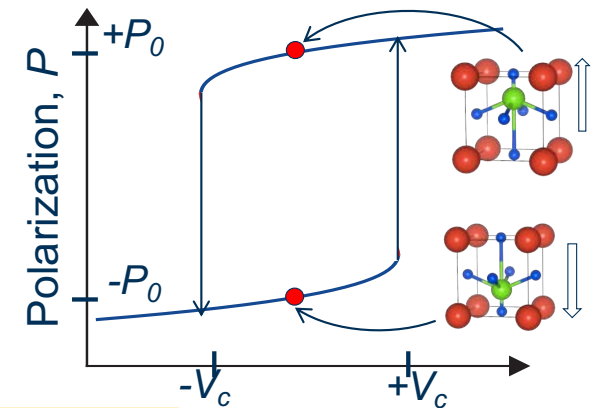
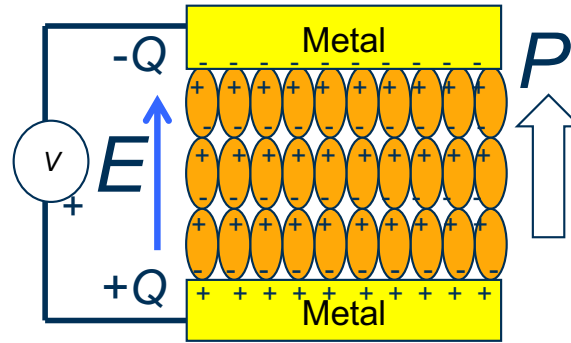
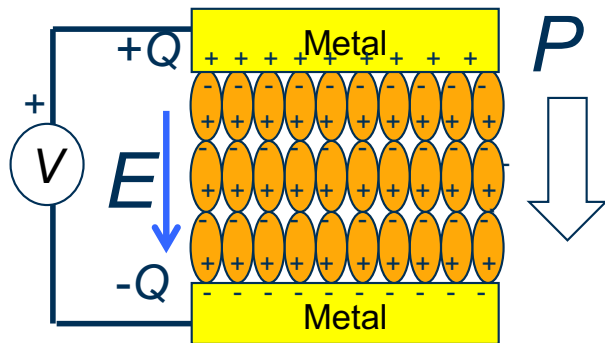
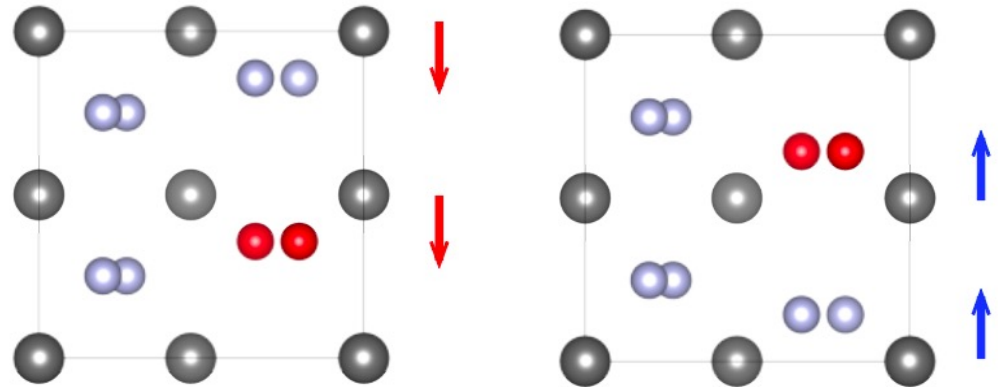
Can ferroelectricity enable next generation NAND flash?



Ferroelectricity

CMOS compatible HfO_2 -based
fluorite-structure ferroelectric

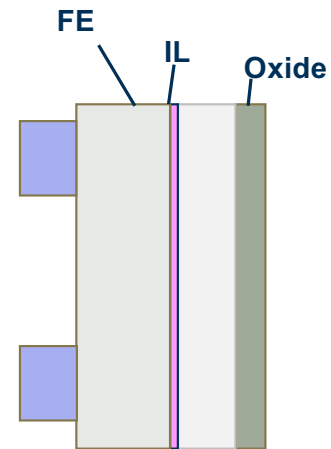
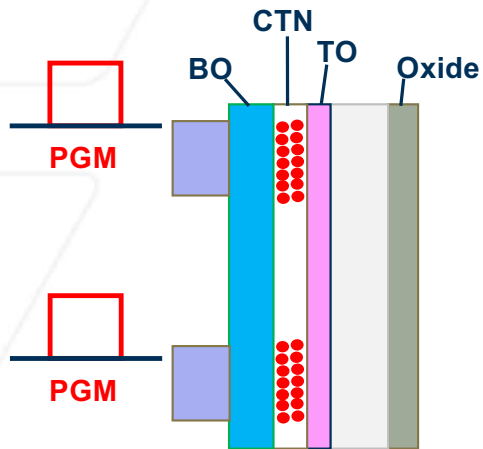
Orthorhombic $Pca2_1$ structure



A ferroelectric is an insulator with a spontaneous polarization, which can be switched by an above coercive voltage .

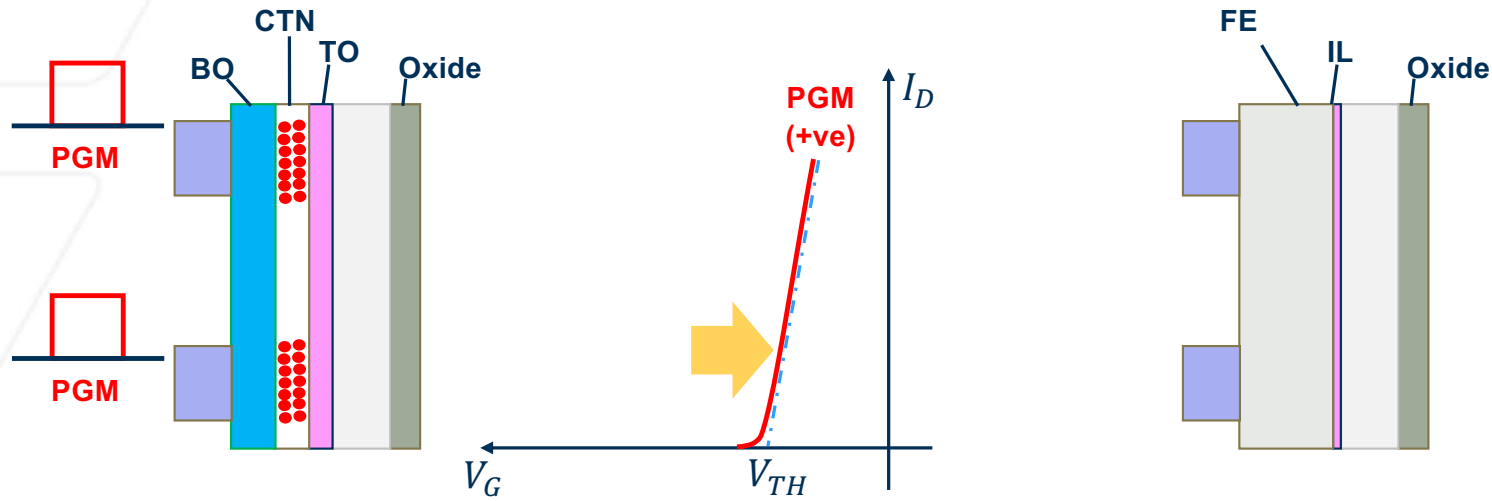
Voltage, V

Charge trap flash vs. ferroelectric flash



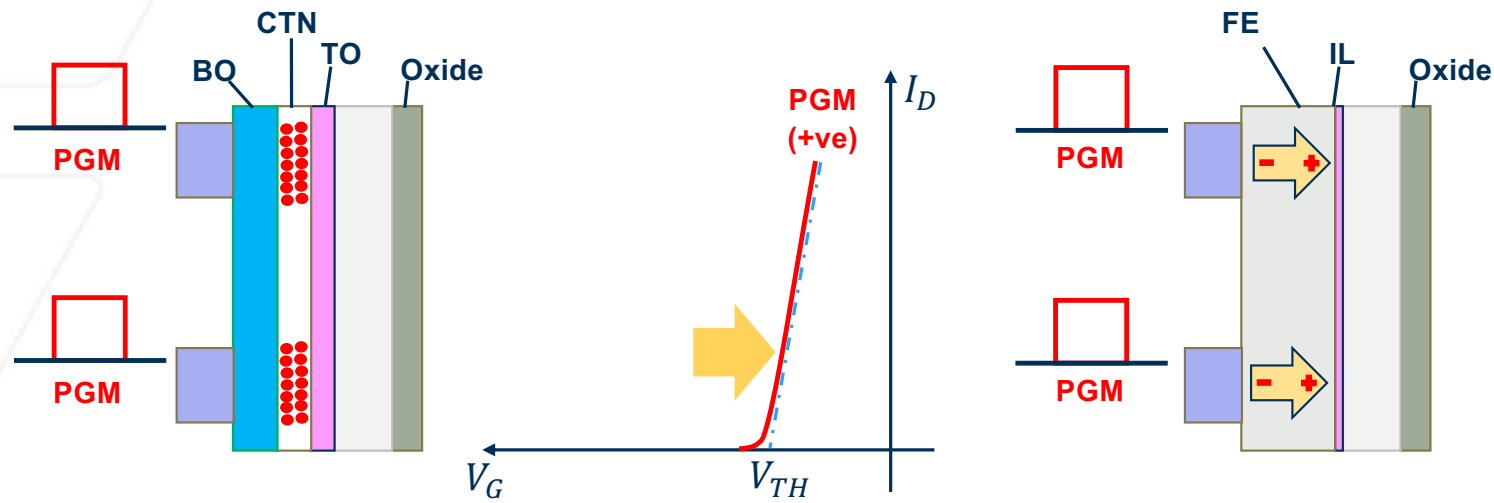
CTF that stores data in the form of charge.

Charge trap flash vs. ferroelectric flash



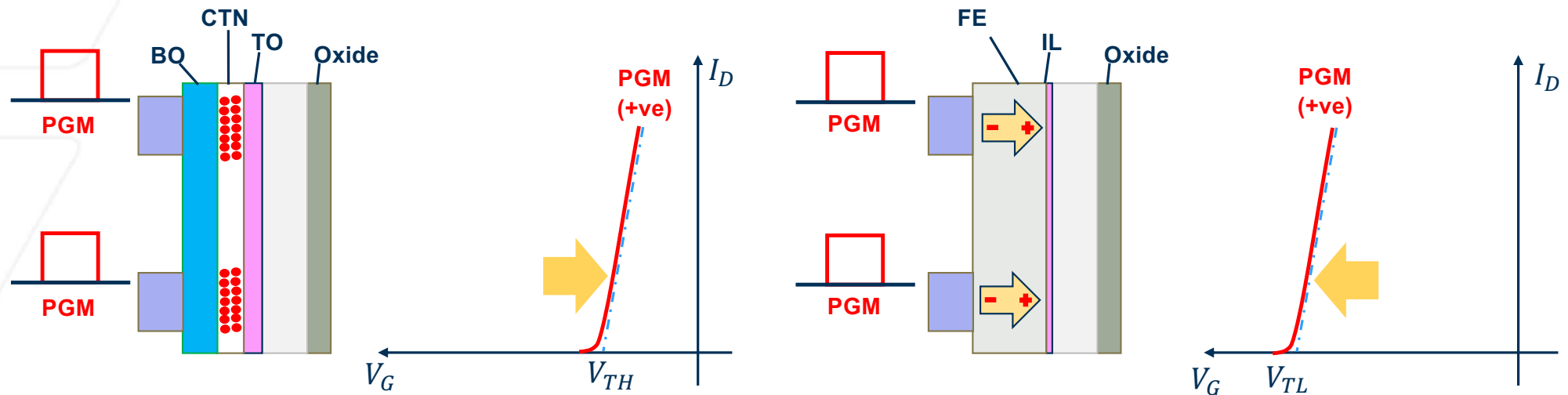
A Positive PGM pulse sets the V_t to high V_t state in CTF NAND.

Charge trap flash vs. ferroelectric flash



A Positive PGM pulse sets the V_t to high V_t state in CTF NAND.

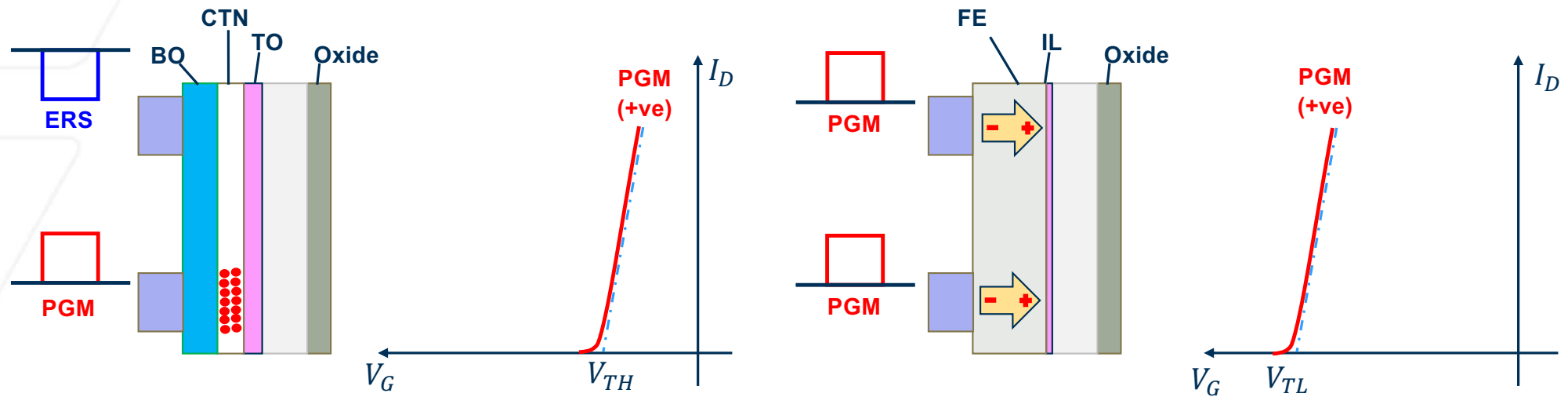
Charge trap flash vs. ferroelectric flash



A Positive PGM pulse sets the V_t to high V_t state in CTF NAND.

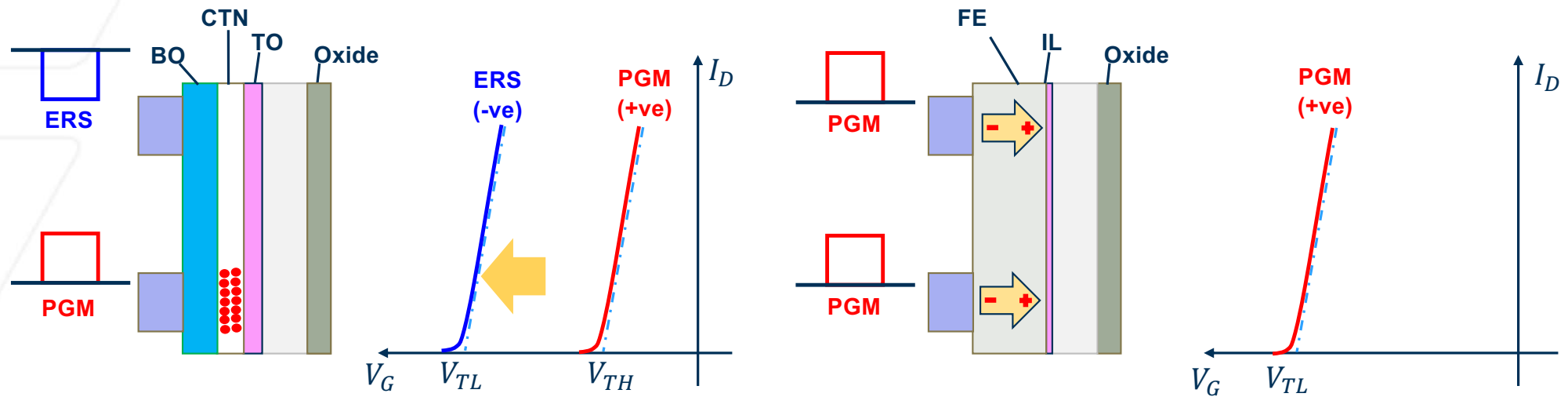
On the contrary, a Positive PGM pulse sets the V_t to low V_t state in a FEFET.

Charge trap flash vs. ferroelectric flash



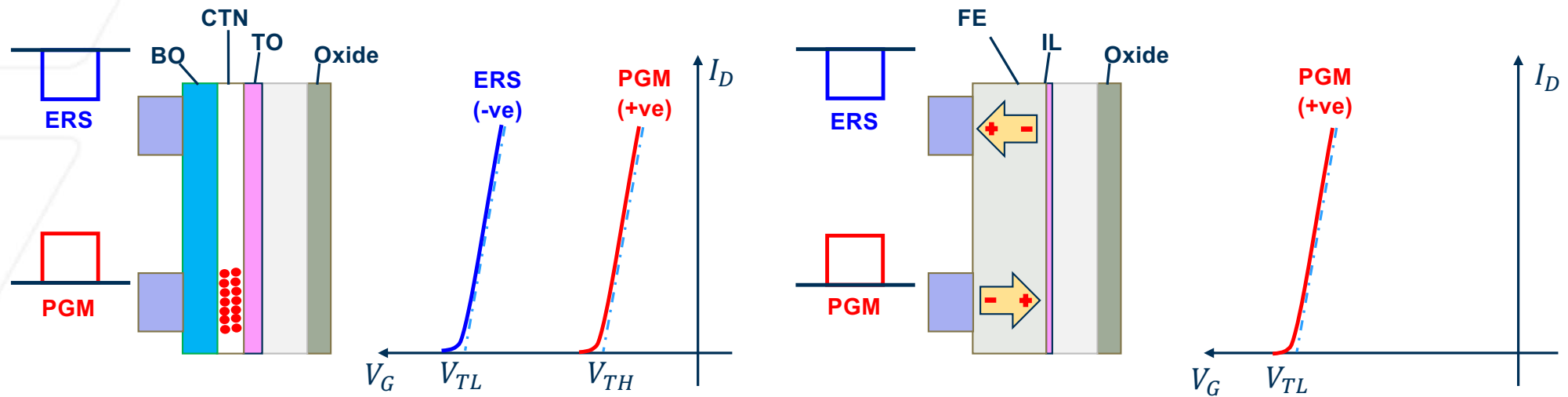
A Negative ERS pulse sets the V_t to low V_t state in CTF NAND.

Charge trap flash vs. ferroelectric flash



A Negative ERS pulse sets the V_t to low V_t state in CTF NAND.

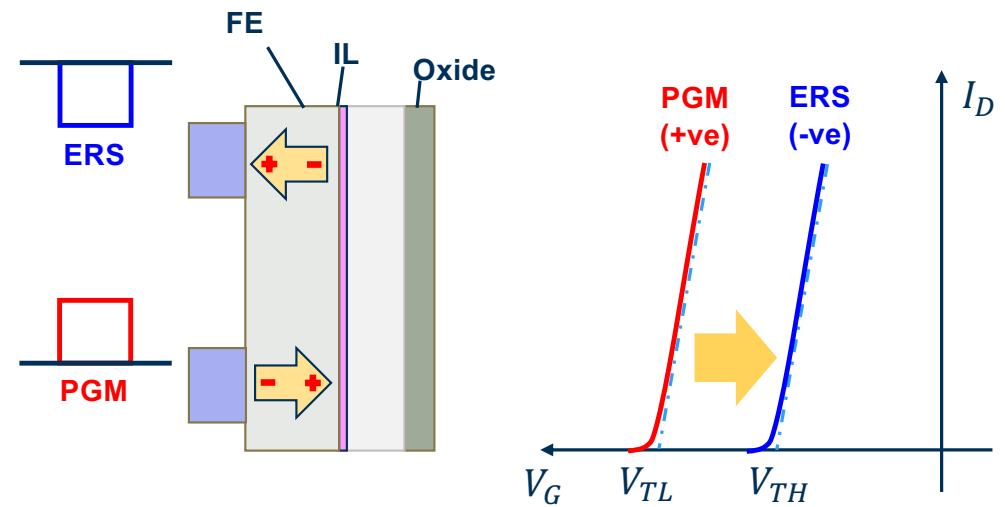
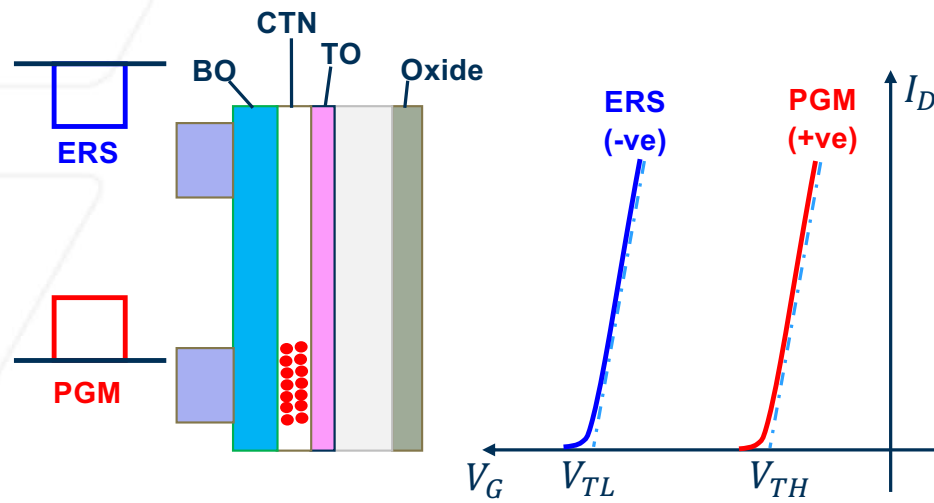
Charge trap flash vs. ferroelectric flash



A Negative ERS pulse sets the V_t to low V_t state in CTF.

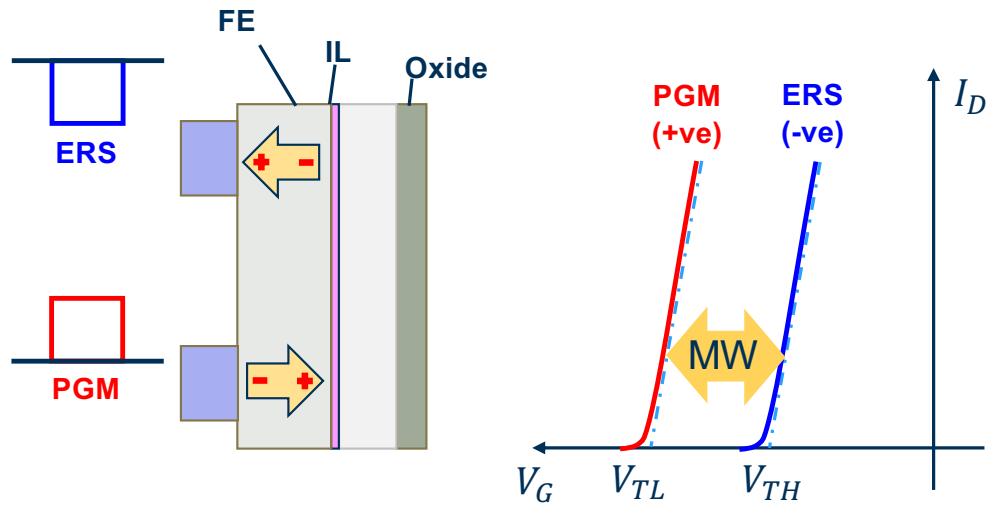
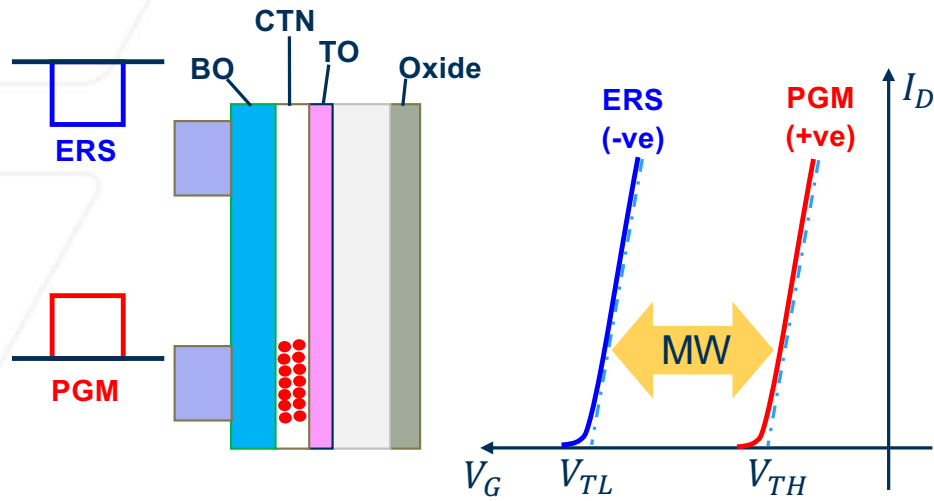
On the contrary, a Negative ERS pulse sets the V_t to high V_t state in FEFET.

Charge trap flash vs. ferroelectric flash



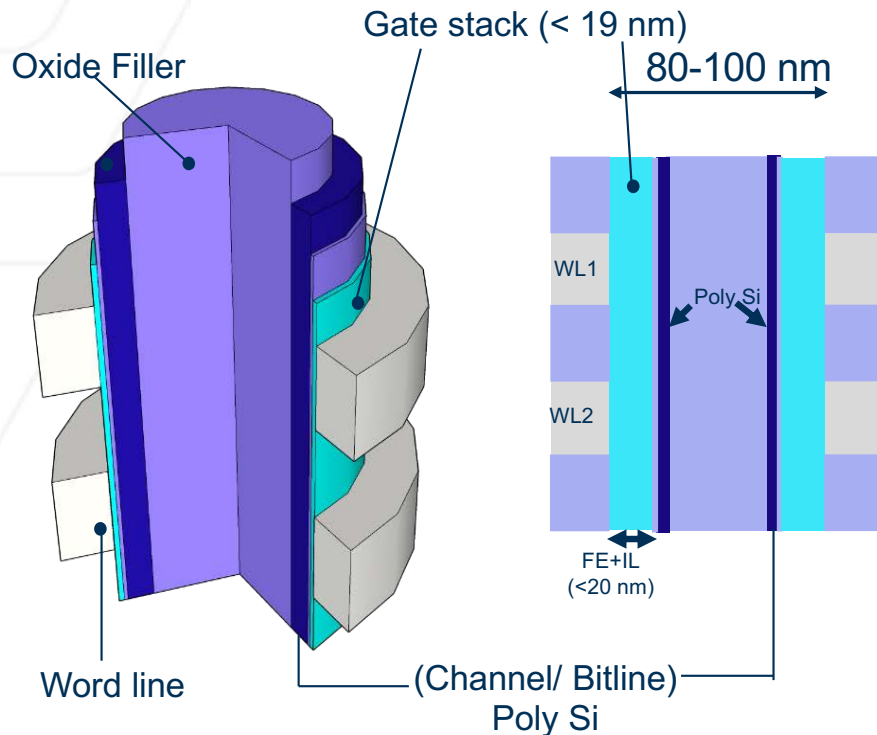
Negative ERS pulse sets the V_t to high V_t state in FEFET

Charge trap flash vs. ferroelectric flash



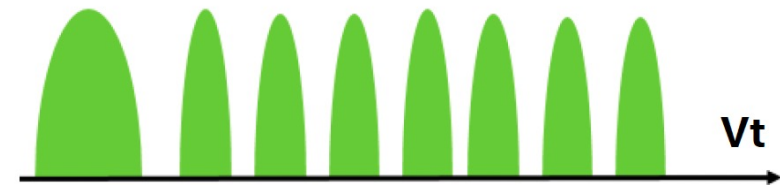
$$MW = V_{TH} - V_{TL}$$

Target metrics for ferroelectric NAND flash

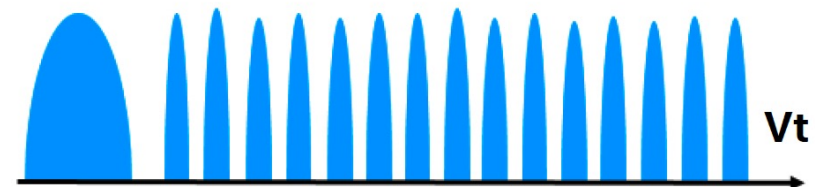


To accommodate the core-shell structure within the hole diameter of 100 nm of the vertical NAND string, the²⁵ thickness of the gate stack is limited to ~20 nm

TLC: 3 bits/cell $\rightarrow$ 8 Vt level MW $\approx$ 6V



QLC: 4 bits/cell $\rightarrow$ 16 Vt level MW $\approx$ 7.5 V



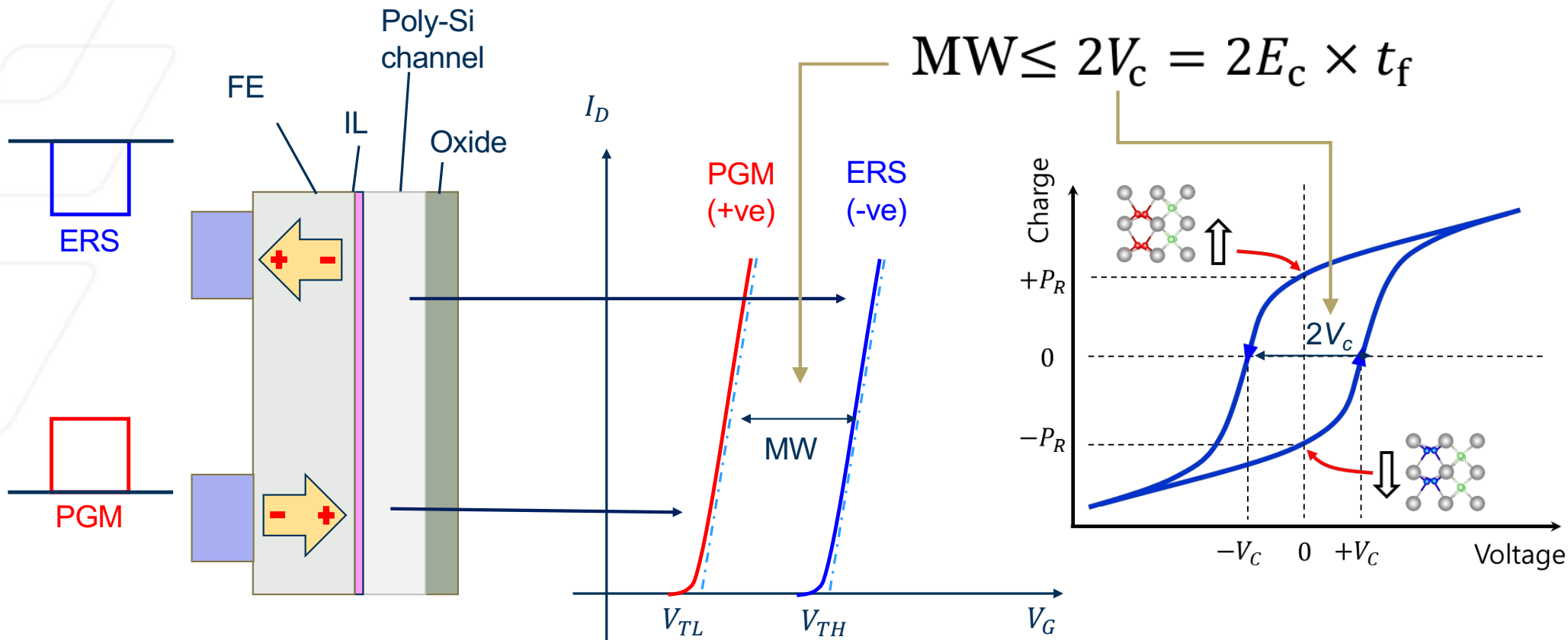
Design constraints

Gate stack thickness, $t \leq 19$ nm

Write voltage ≤ 15 V

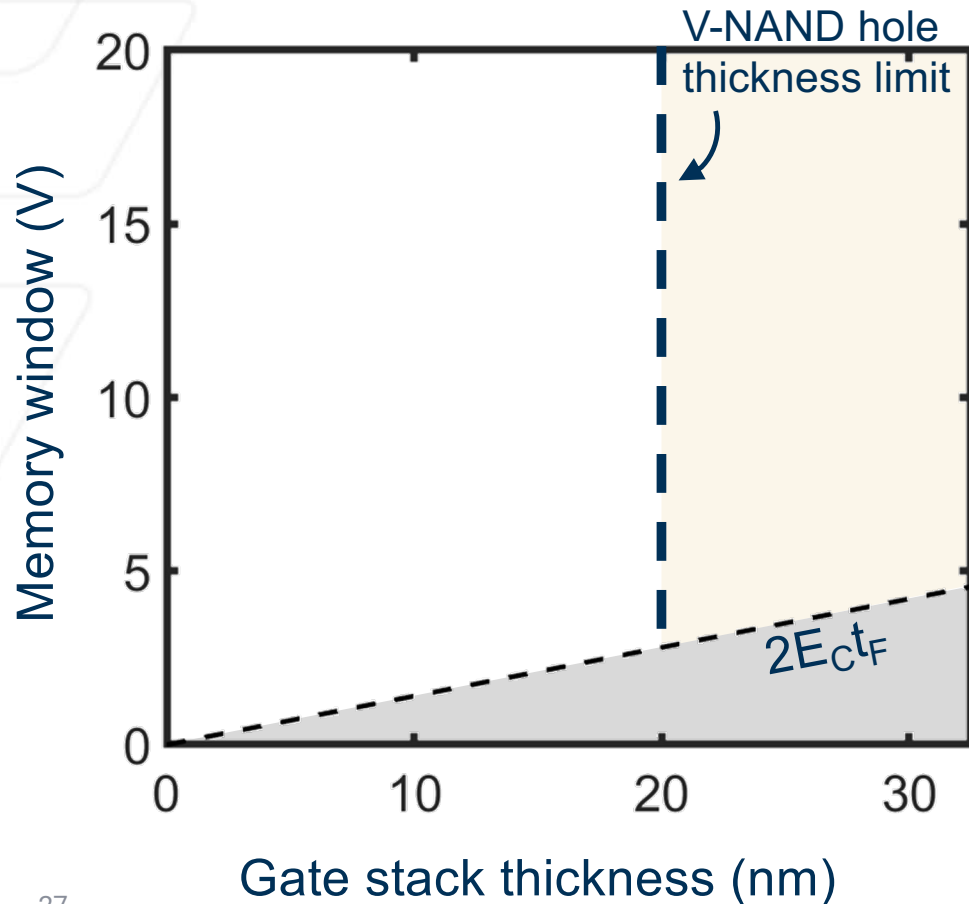
Memory window, MW ≥ 6.5 V

Ferroelectric flash



The maximum memory window of a FeFET is theoretically limited by the memory window of the ferroelectric layer.

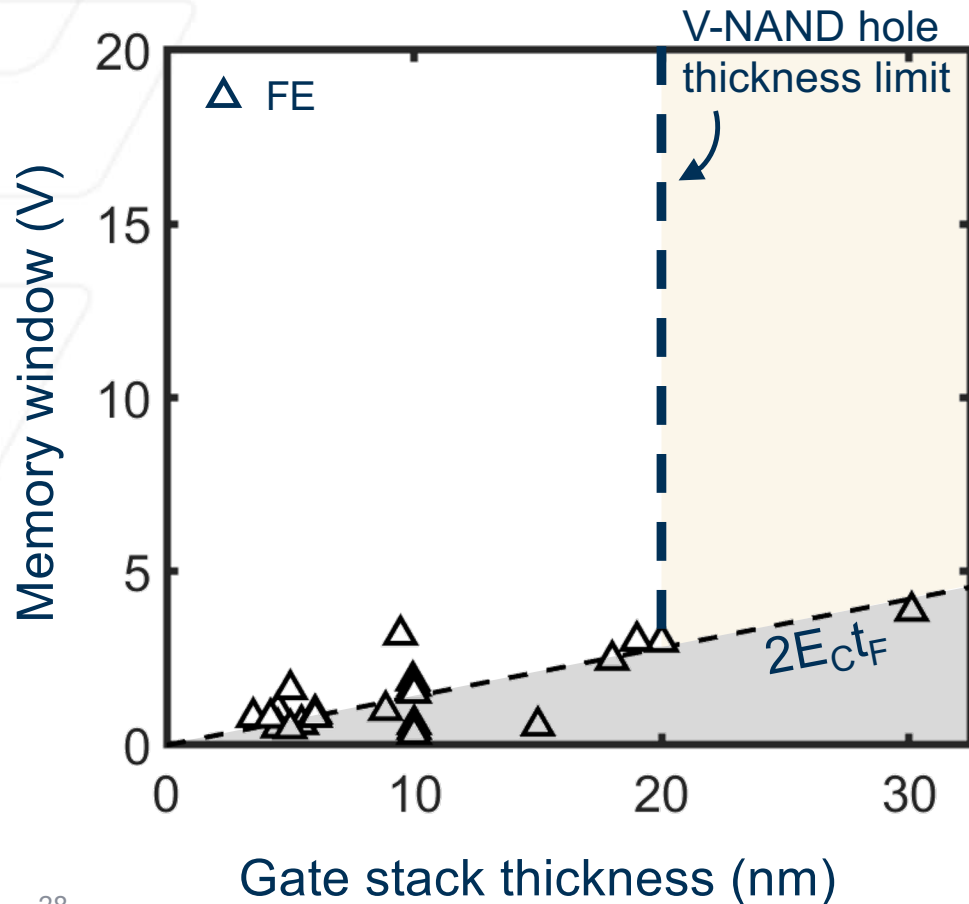
Memory window of FEFETs is limited by V_c



$$MW \leq 2V_c = 2E_c \times t_f$$

The maximum memory window of an FeFET is theoretically limited by the memory window of the ferroelectric layer.

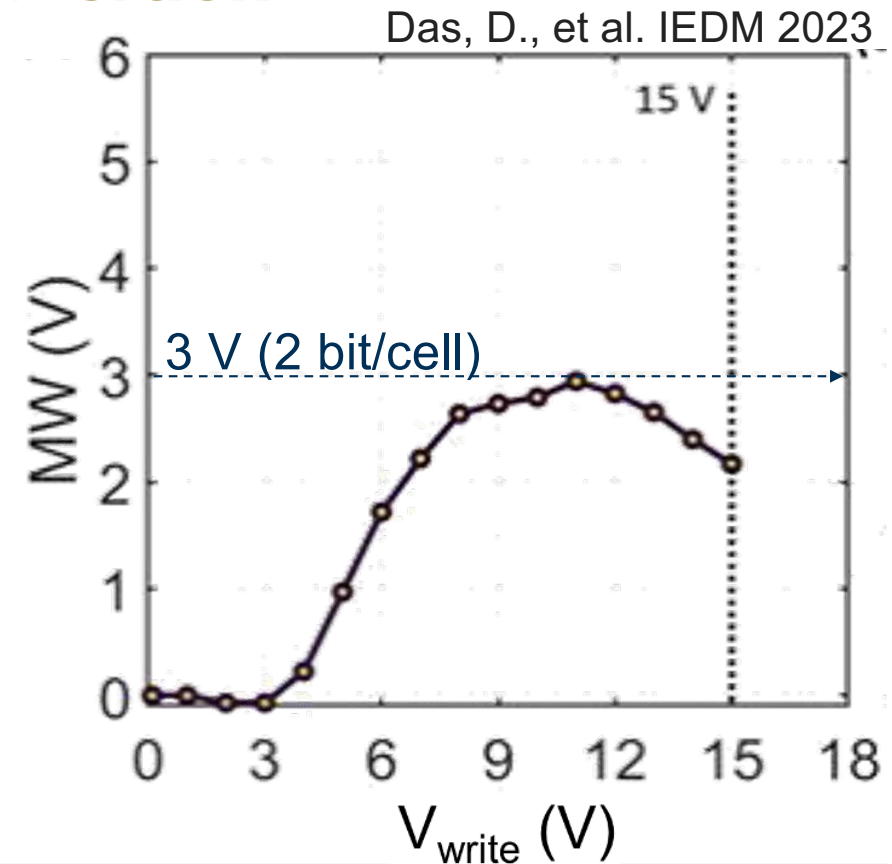
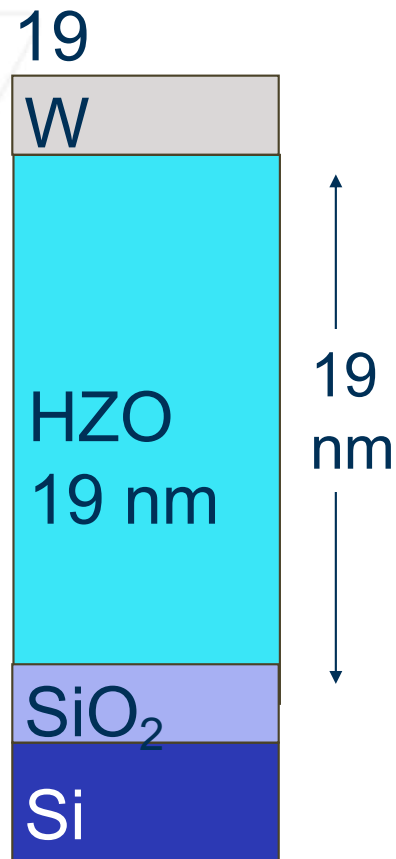
Memory window of FEFETs is limited by V_c



$$MW \leq 2V_c = 2E_c \times t_f$$

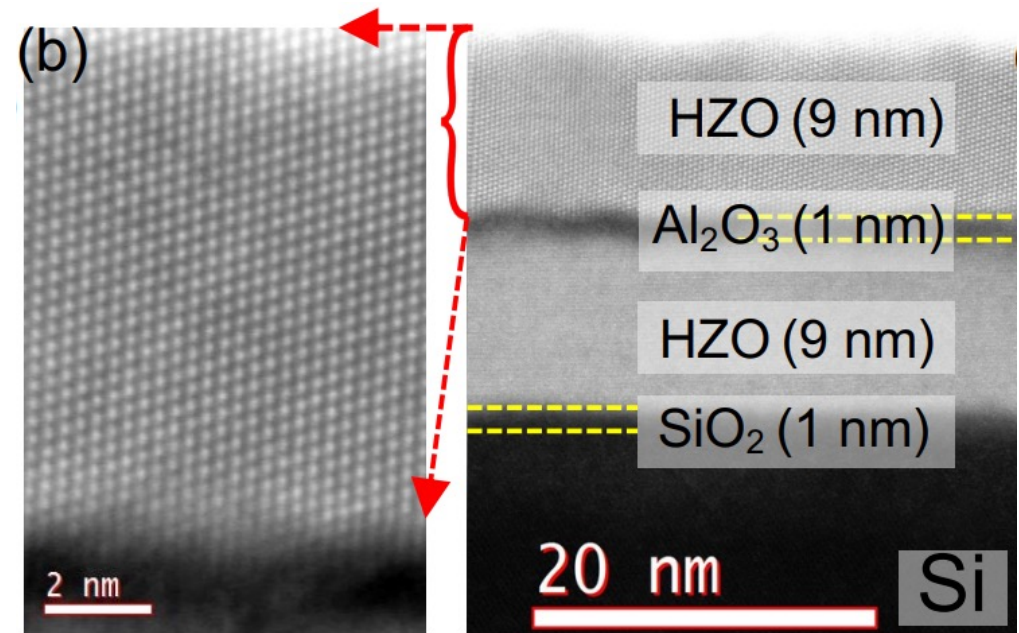
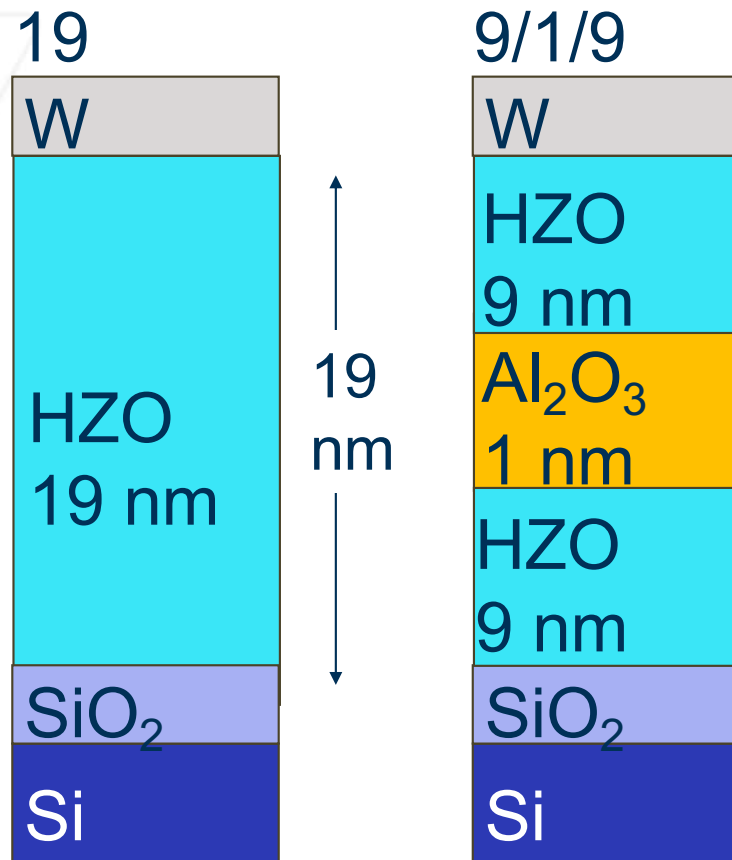
The maximum memory window of an FeFET is theoretically limited by the memory window of the ferroelectric layer.

Increasing the MW by dielectric insertion

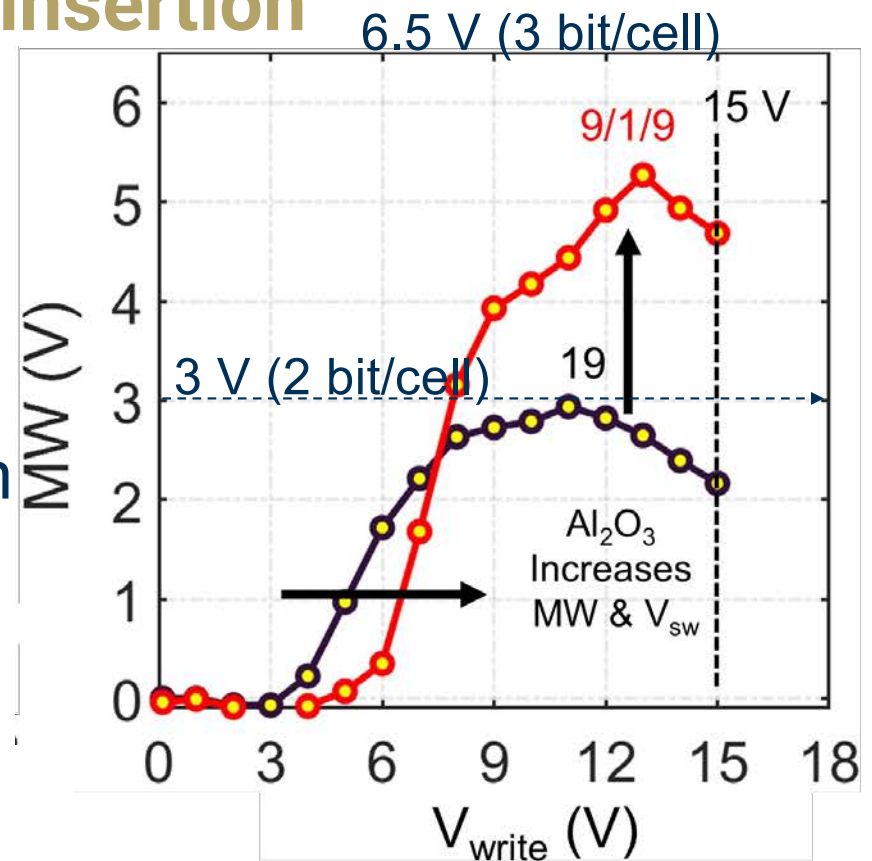
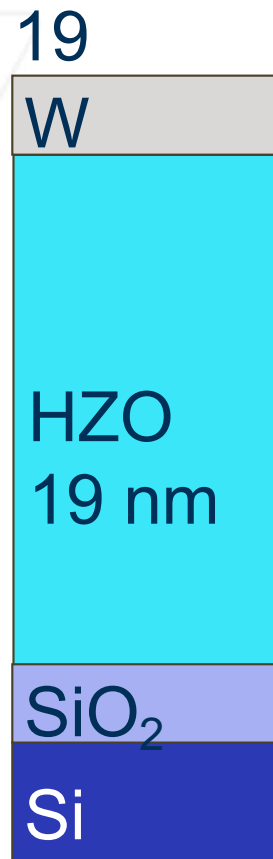


10 nm HZO layer leads to maximum memory window of 2.9 V.

Increasing the MW by dielectric insertion

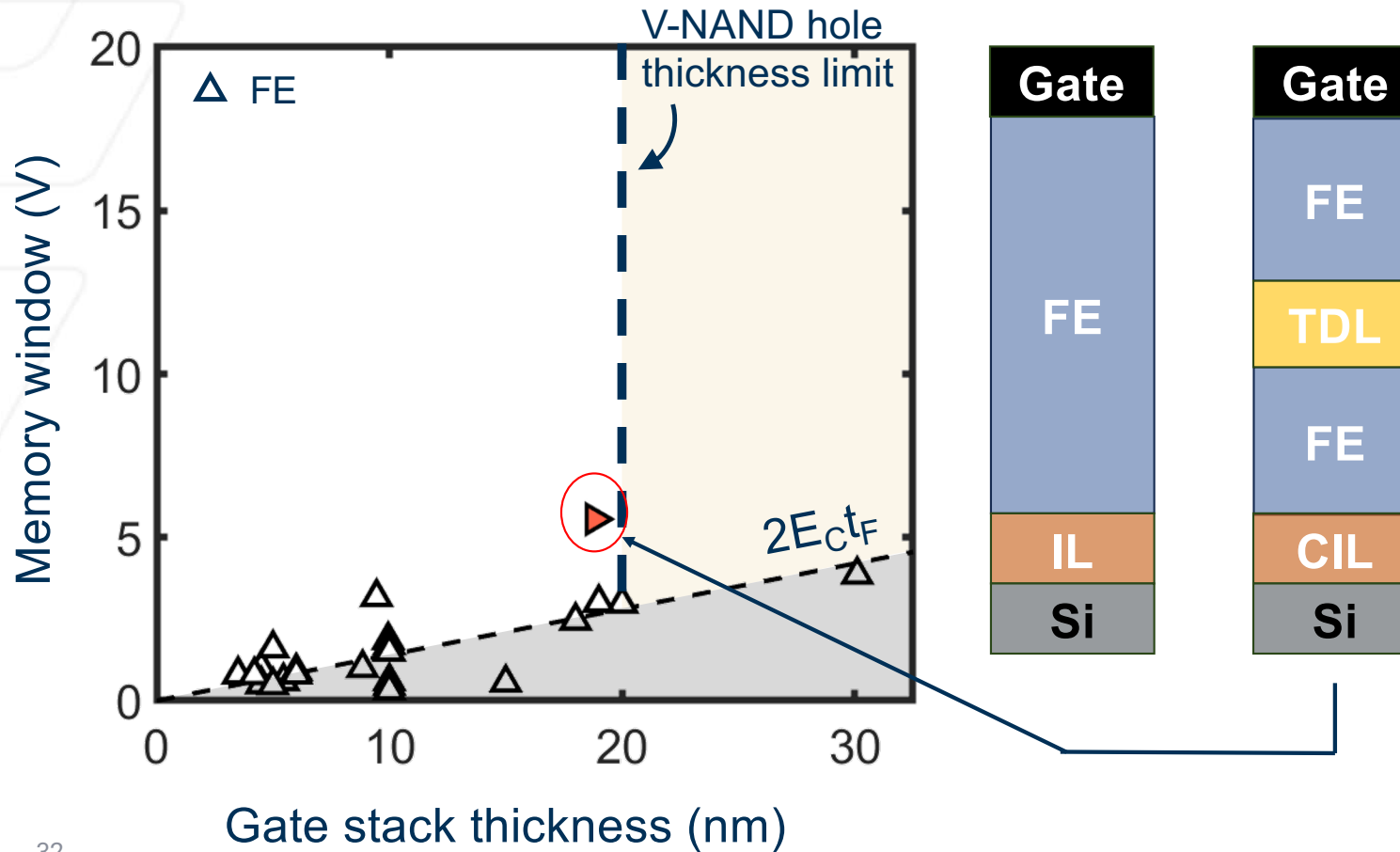


Increasing the MW by dielectric insertion

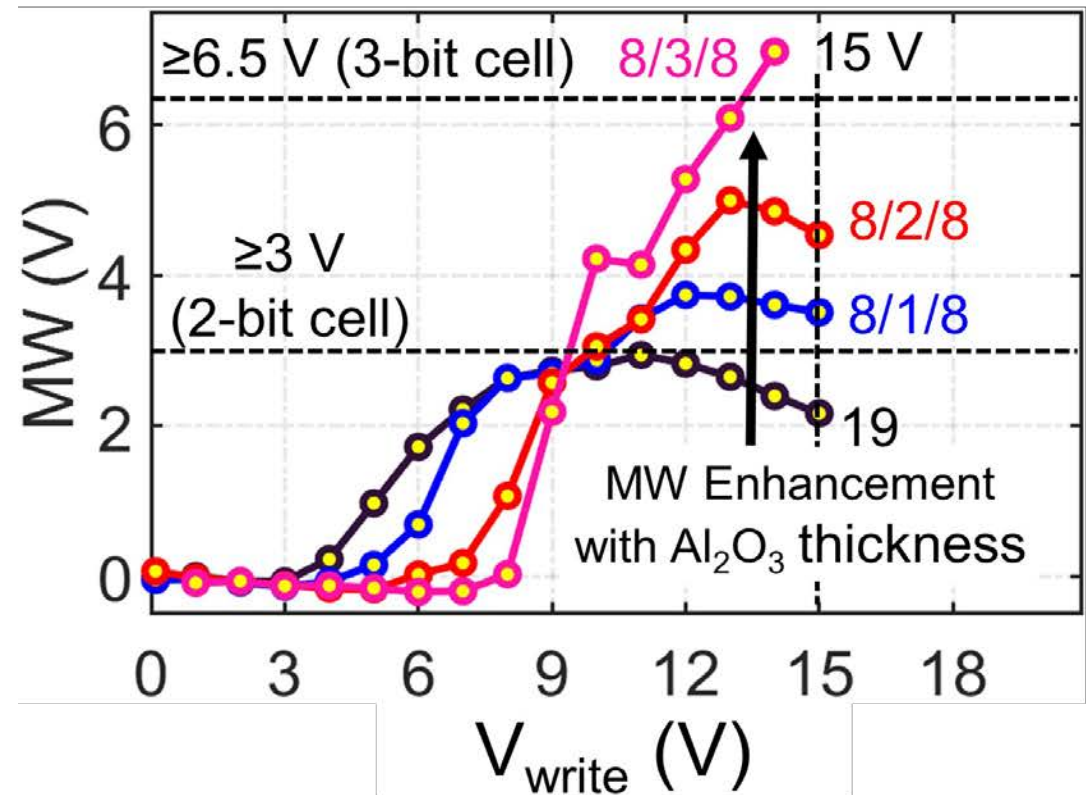
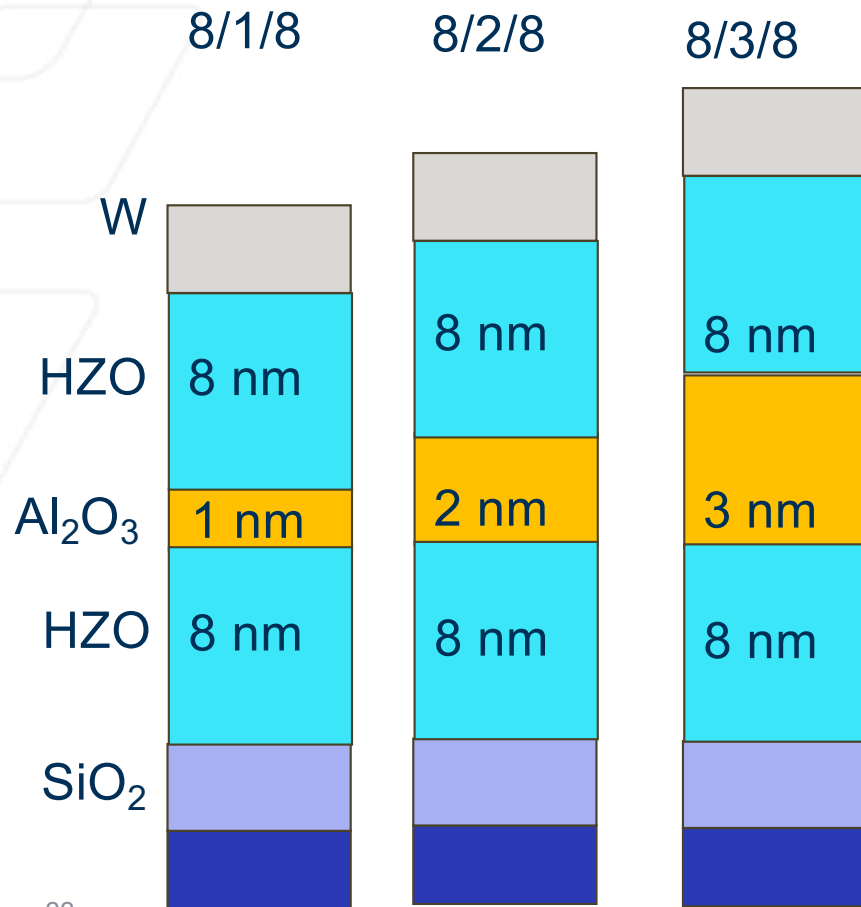


Introducing 1 nm Al₂O₃ intermediate layer dramatically memory window to 5.3 V.

Memory window of FEFETs is limited by V_c

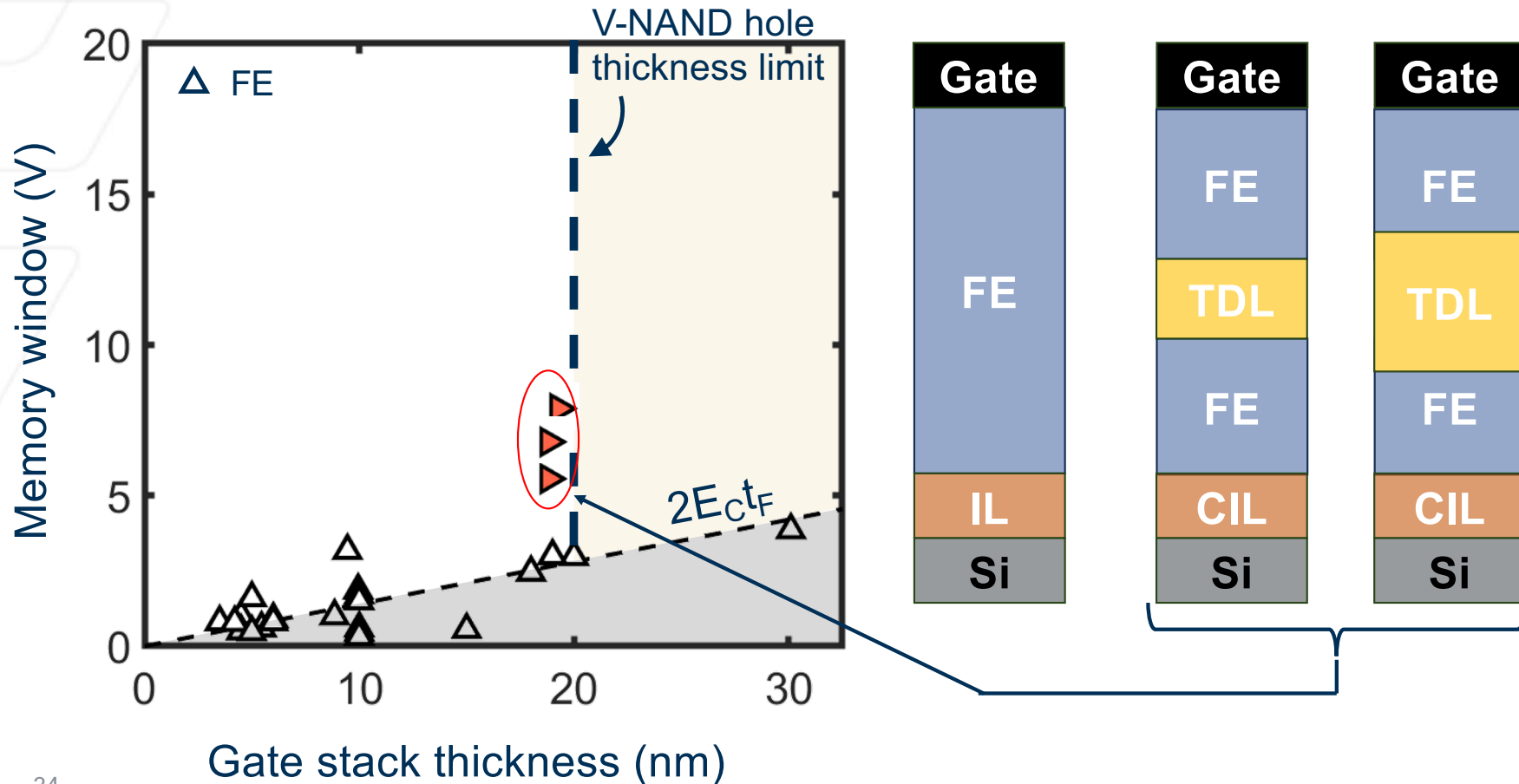


Increasing the MW by dielectric insertion



FE 8 nm – AlO 3 nm – FE 8 nm leads to a 7.3 V maximum MW for a 14 V write voltage.

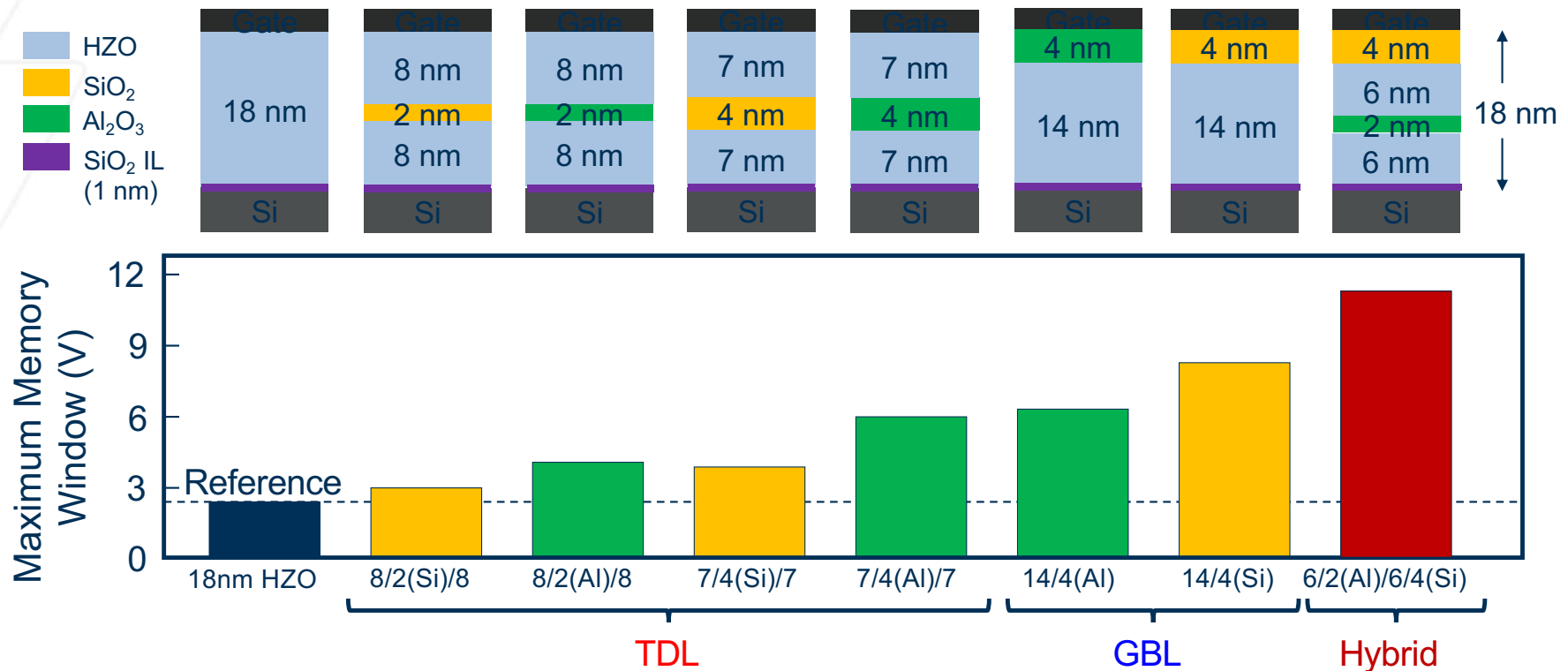
Increasing the MW by dielectric insertion



34

Das, Khan, et al. IEDM 2023 (highlight paper): [10.1109/IEDM45741.2023.10413697](https://doi.org/10.1109/IEDM45741.2023.10413697)

Position and choice of the dielectric

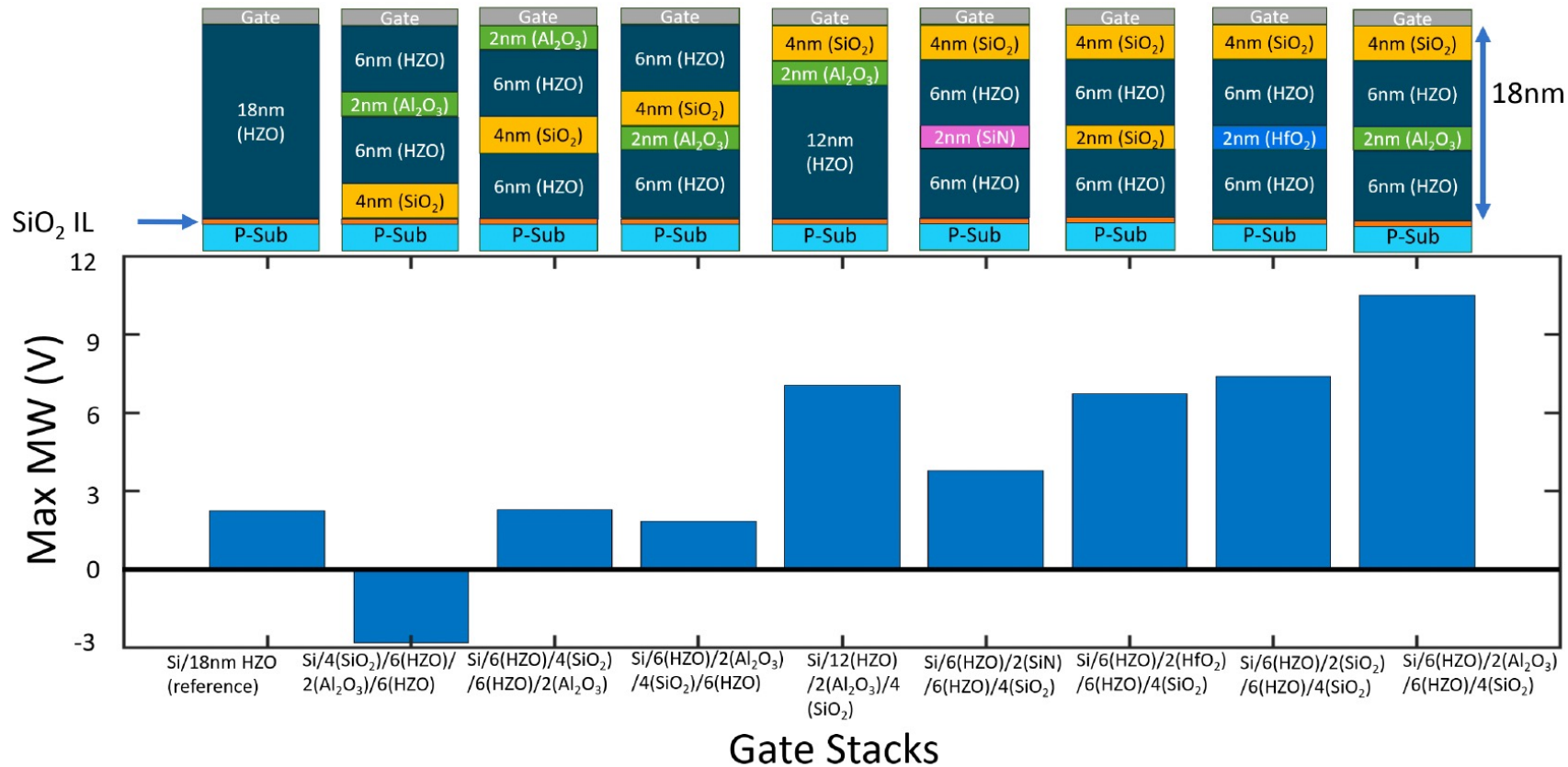


- GBL stacks have higher MW than TDL stacks for similar thickness and material.
- SiO₂ in GBL and Al₂O₃ in TDL gives higher MW for similar dielectric thickness

Fernandes, Khan et al. *Material choices for tunnel dielectric layer and gate blocking layer for ferroelectric NAND applications*. IEEE Electron Device Letter 45, 1776 - 1779 (2024).

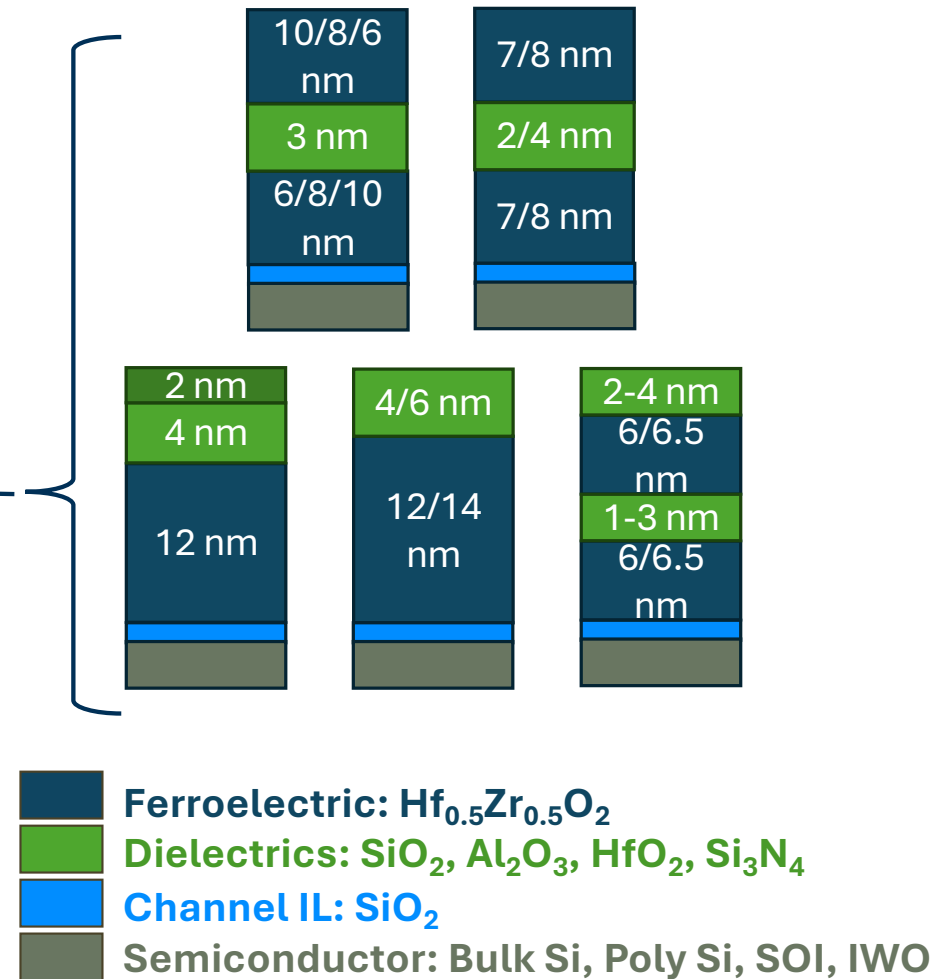
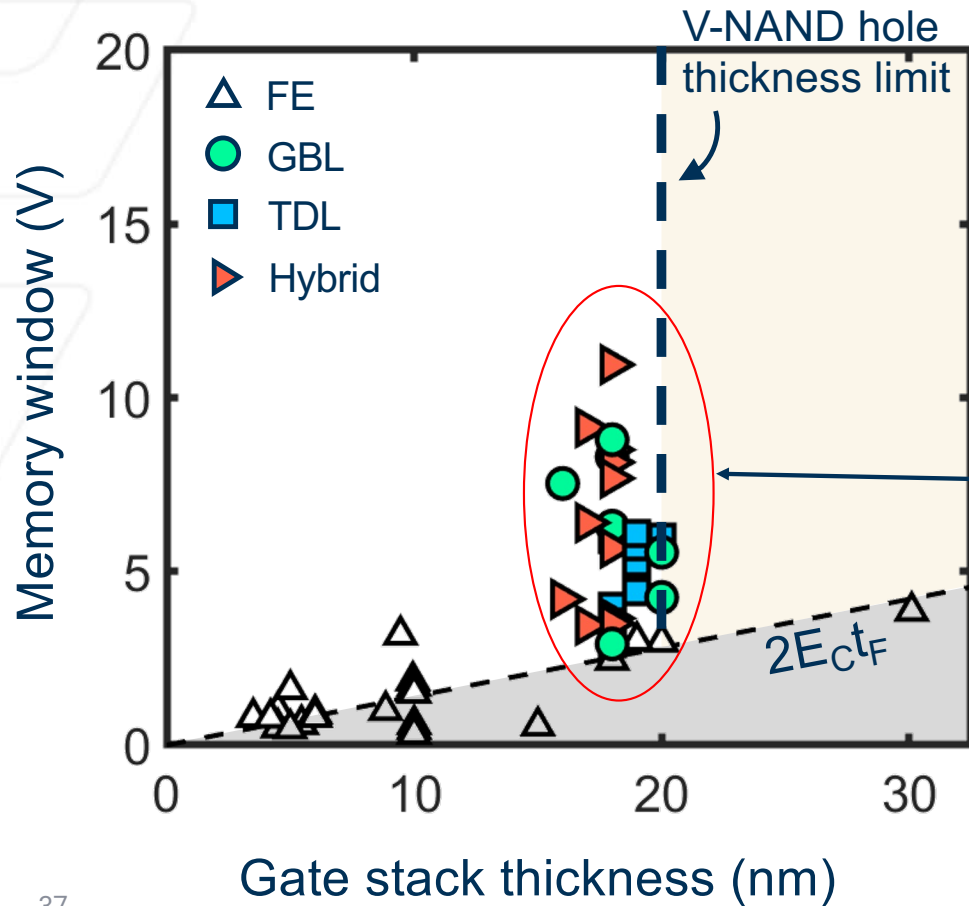
<https://doi.org/10.1109/LED.2024.3437239>

Position and choice of the dielectric



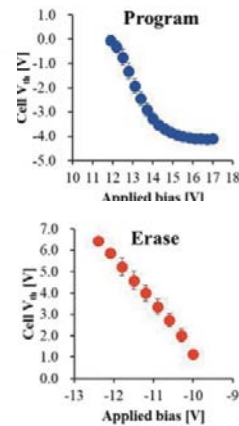
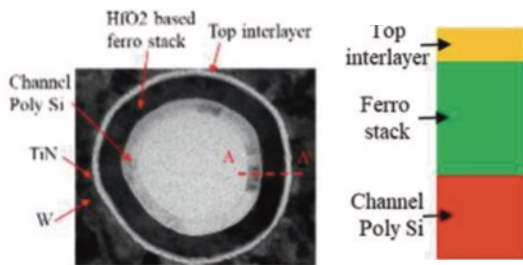
Fernandes, Khan et al. *Optimizing Memory Window for Ferroelectric Nand Applications: An Experimental Study on Dielectric Material Selection and Layer Positioning*. IEEE Transactions on Electron Devices 72, 234 - 239 (2024).
<https://doi.org/10.1109/TED.2024.3504475>

Increasing the MW by dielectric insertion

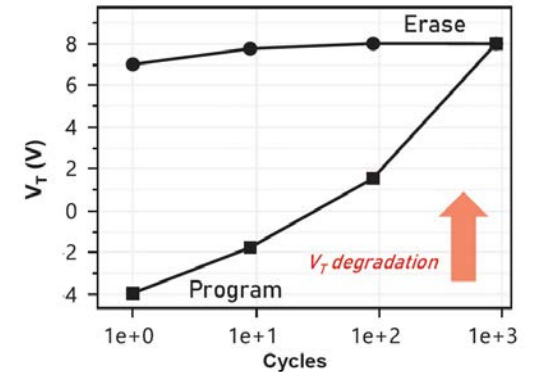
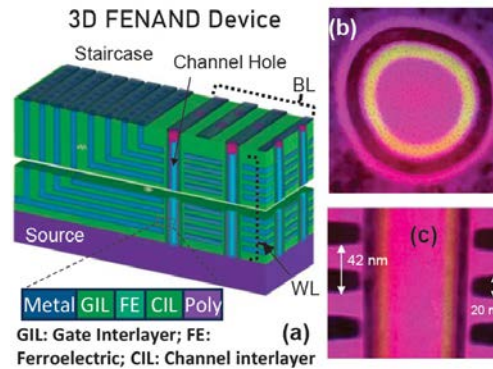


Increasing the MW by dielectric insertion

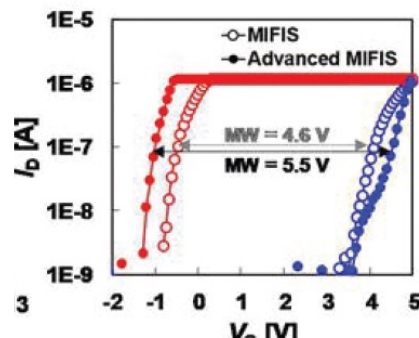
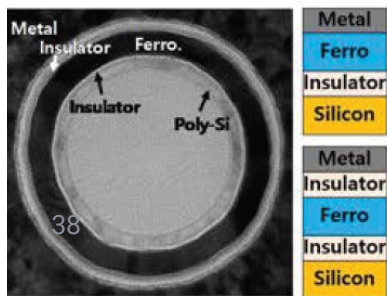
SK Hynix Yoon et al. VLSI 2023



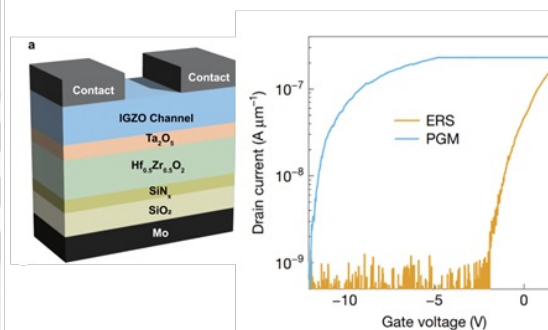
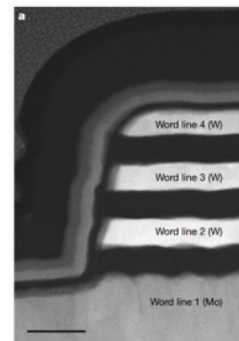
Micron Sharma et al. VLSI 2025



Samsung Lim et al. IEDM 2023



Samsung Sharma et al. Science 2025

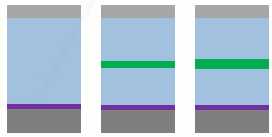


Increasing the MW by dielectric insertion

Our results

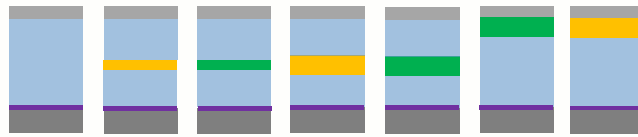
Batch 1: Tunnel Dielectric Layer

Das et al., IEDM 2023:



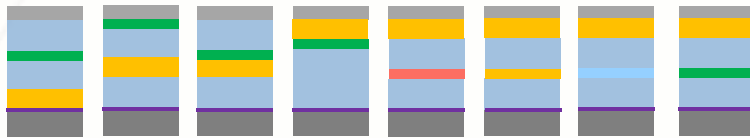
Batch 2: Choice and position of dielectric insert

Fernandes et al., EDL 2024:



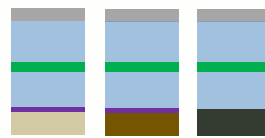
Batch 3: Hybrid structures

Fernandes et al., TED 2024:



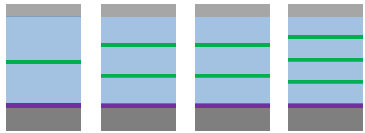
Batch 4: Channel materials

Fernandes et al., IMW 2025:



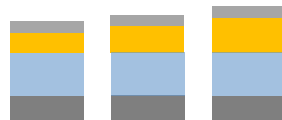
0.5 nm TDL inserts

Lee et al., TED 2024:



AOS channels

Yoo et al., VLSI 2024:



Joh et al., IEDM 2024:

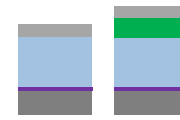


Gate blocking layer

Lim et al., IEDM 2023:



Hu et al., EDL 2024:



Yang et al., ICSIST 2024:



Hu et al., EDL 2024:



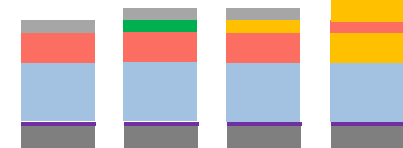
Chu et al., EDL 2024:



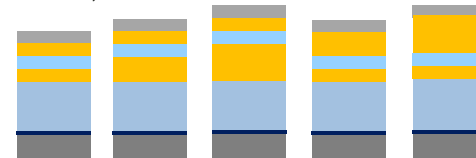
Kuk et al., IEDM 2024:



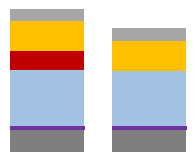
Qin et al., IEDM 2024:



Han et al., IRPS 2025:



G. Kim et al., IEDM 2024:



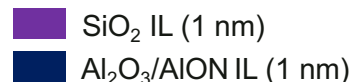
Ferroelectrics:



Dielectrics:



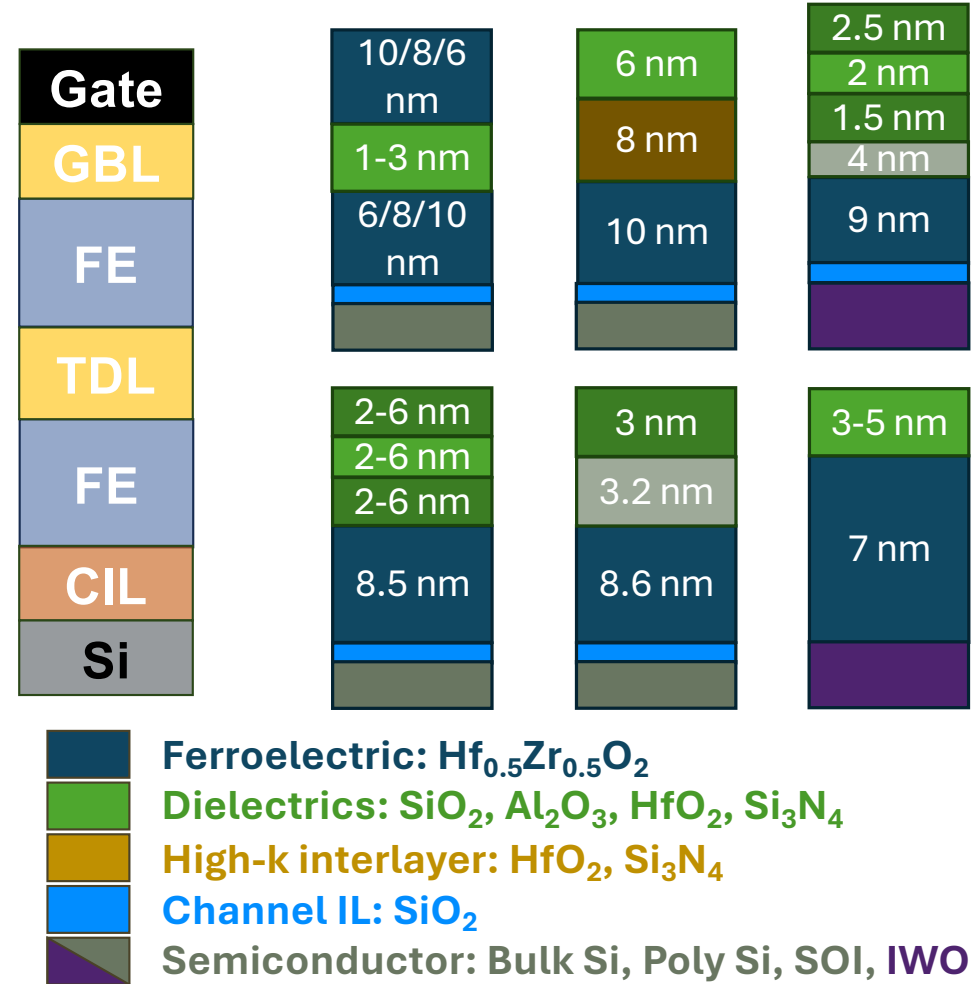
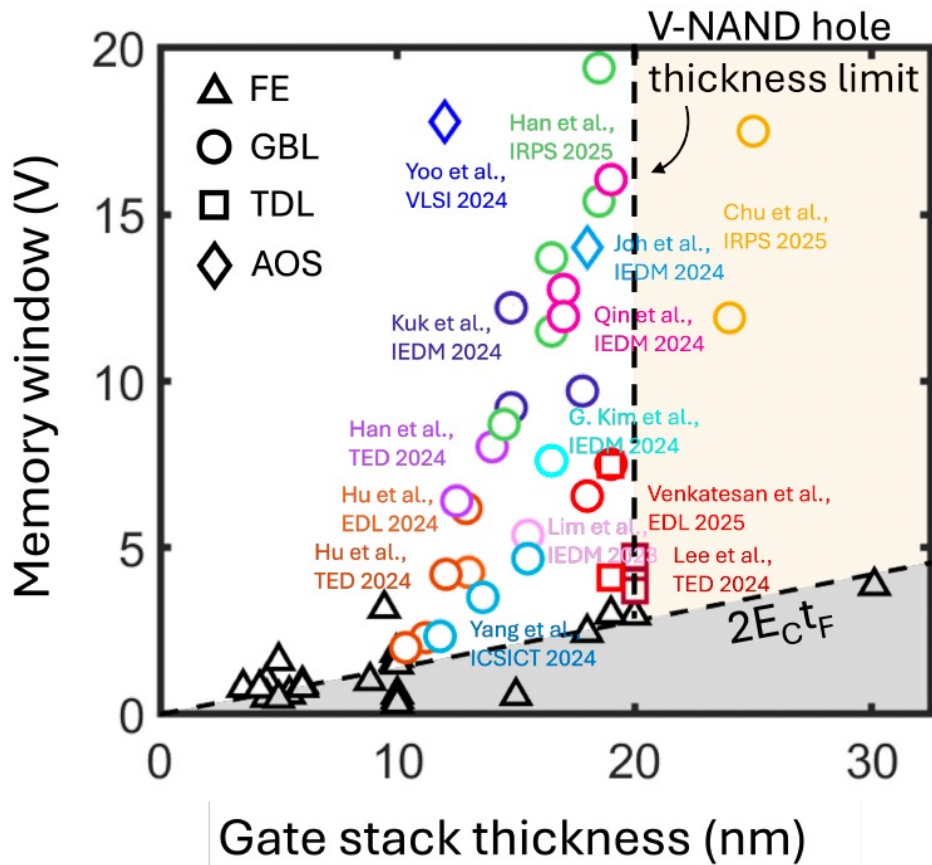
Channel IL:



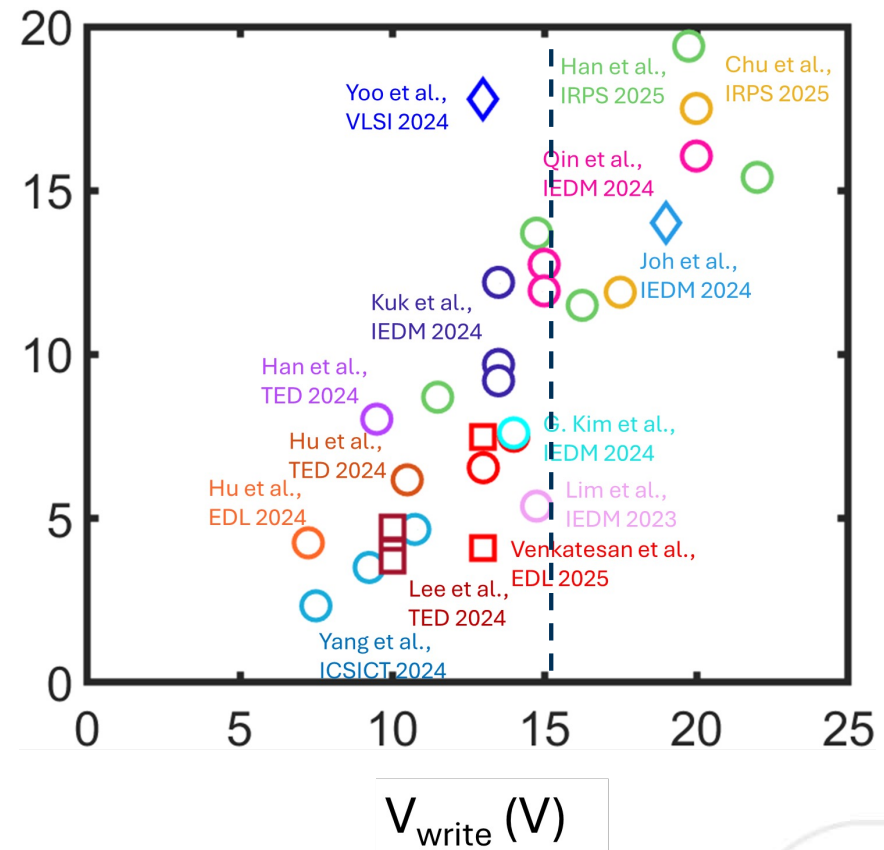
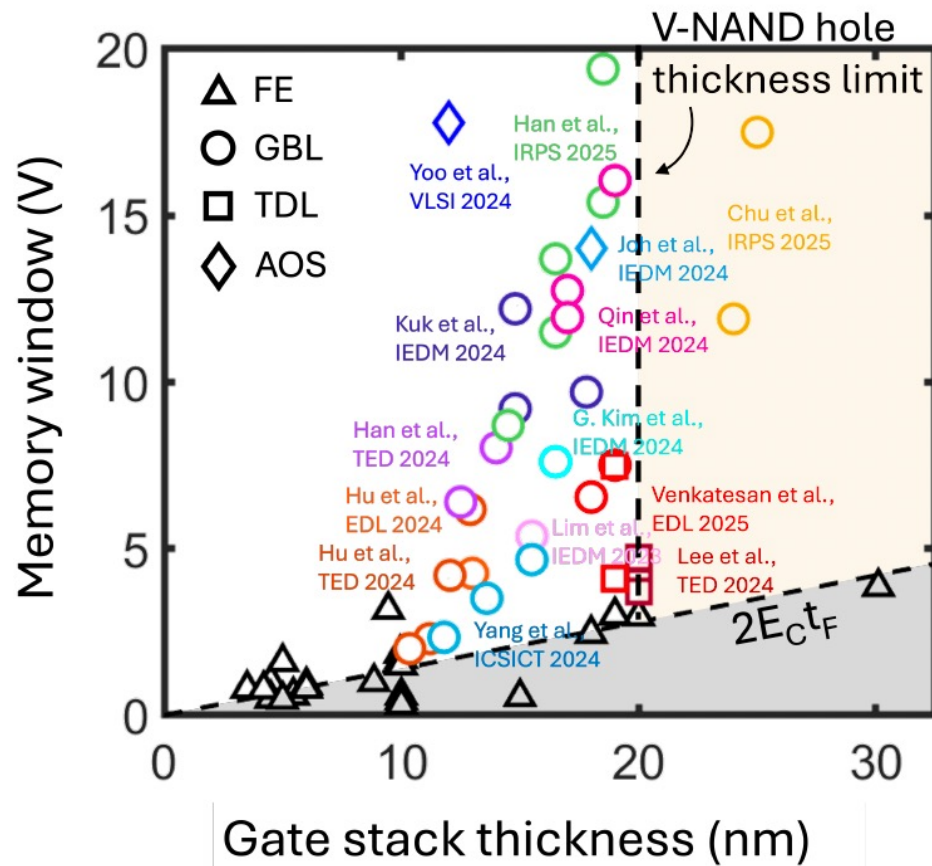
Semiconductor:



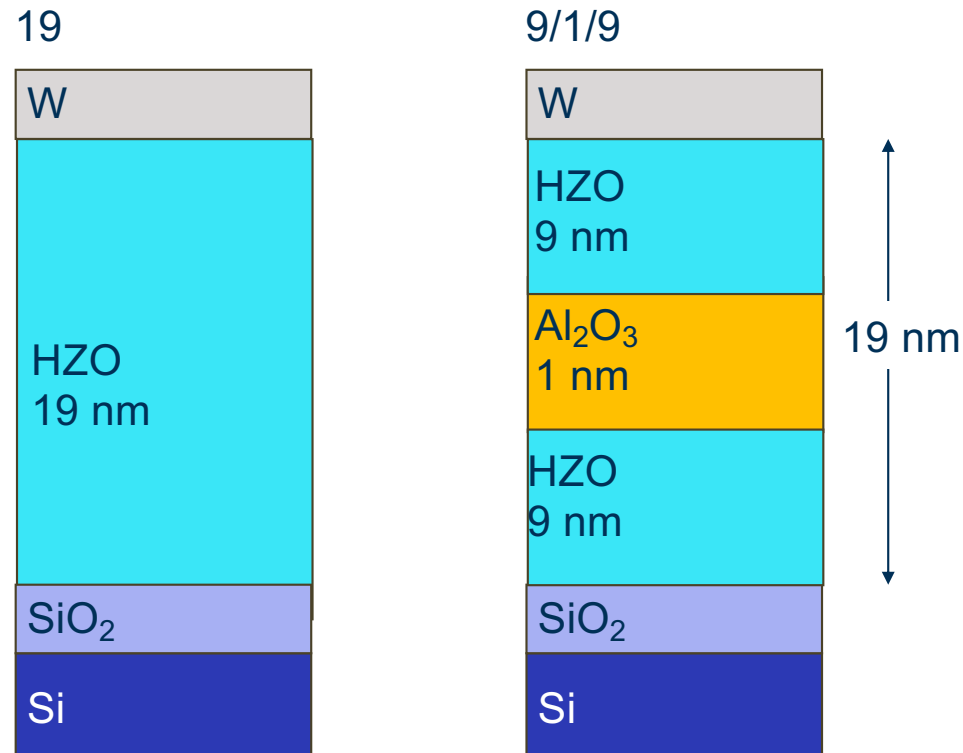
Increasing the MW by dielectric insertion



Increasing the MW by dielectric insertion

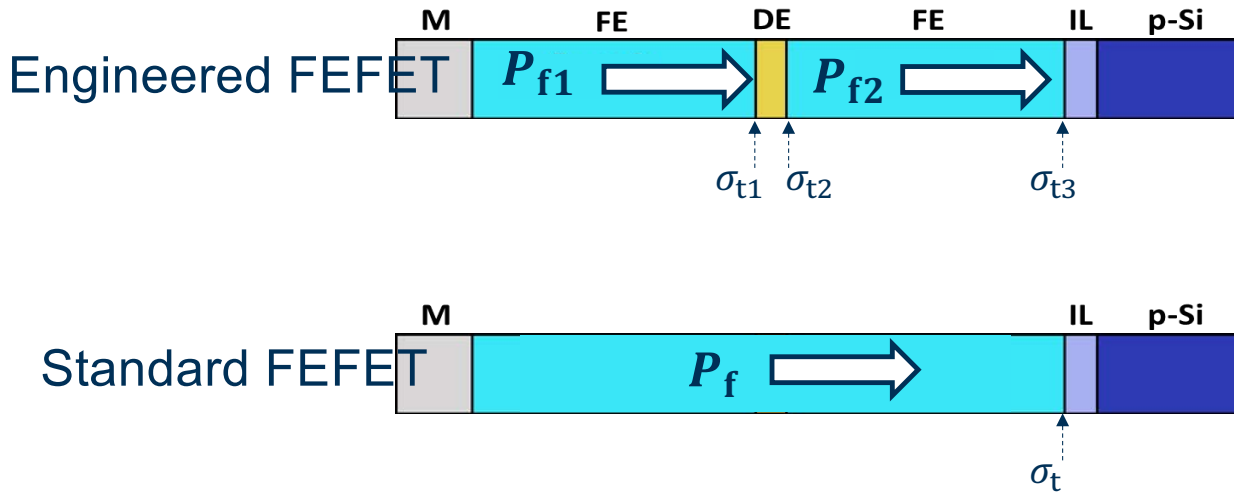


Where is the MW enhancement coming from?



Das, D., et al. IEDM 2023: [10.1109/IEDM45741.2023.10413697](https://doi.org/10.1109/IEDM45741.2023.10413697)

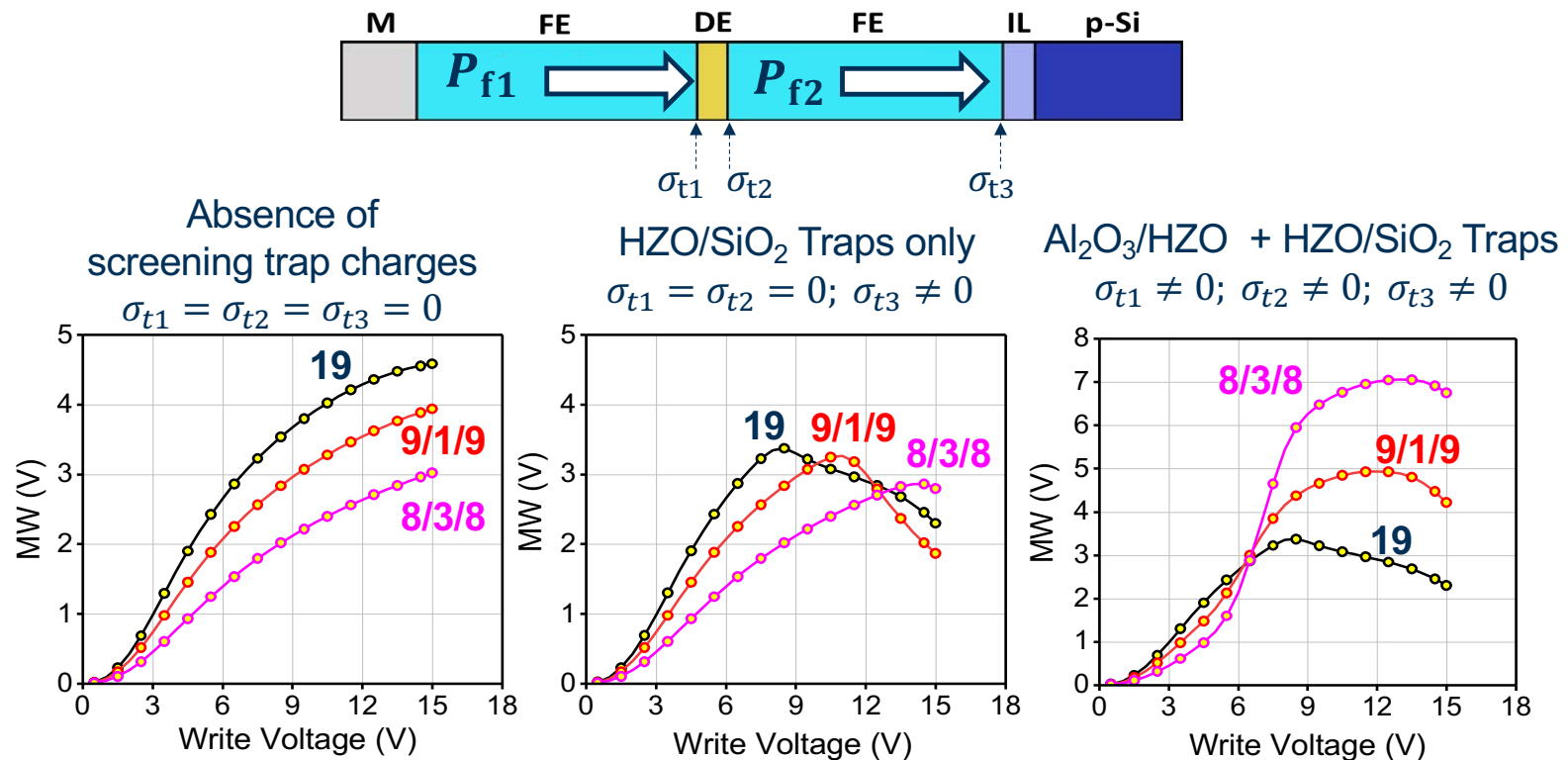
Where is the MW enhancement coming from?



Trapped charges are the key differentiator.

Das, D., et al. IEDM 2023: [10.1109/IEDM45741.2023.10413697](https://doi.org/10.1109/IEDM45741.2023.10413697)

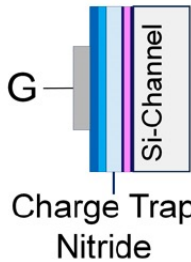
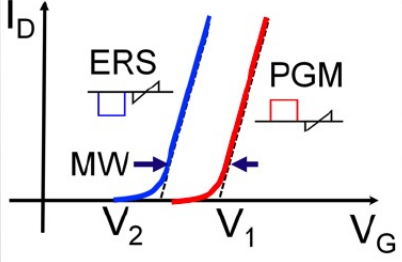
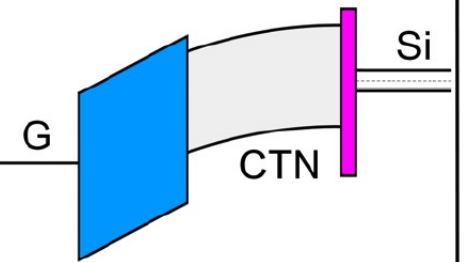
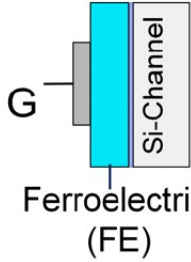
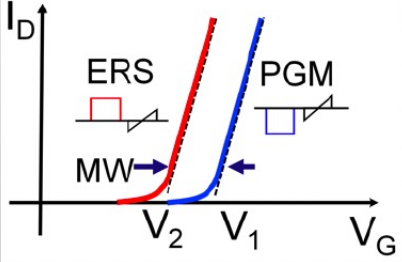
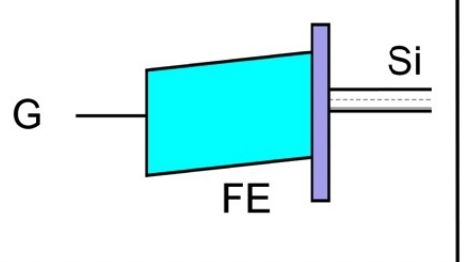
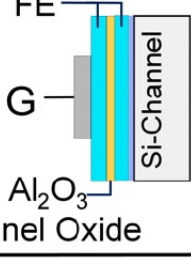
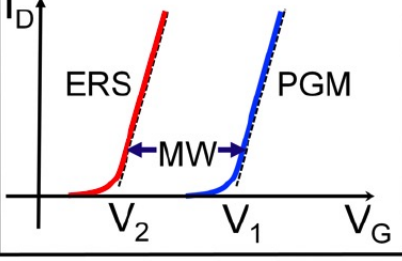
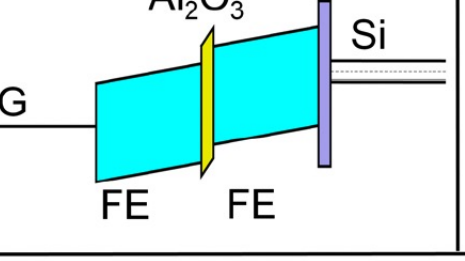
Decoupling the impacts of different trapping mechanisms



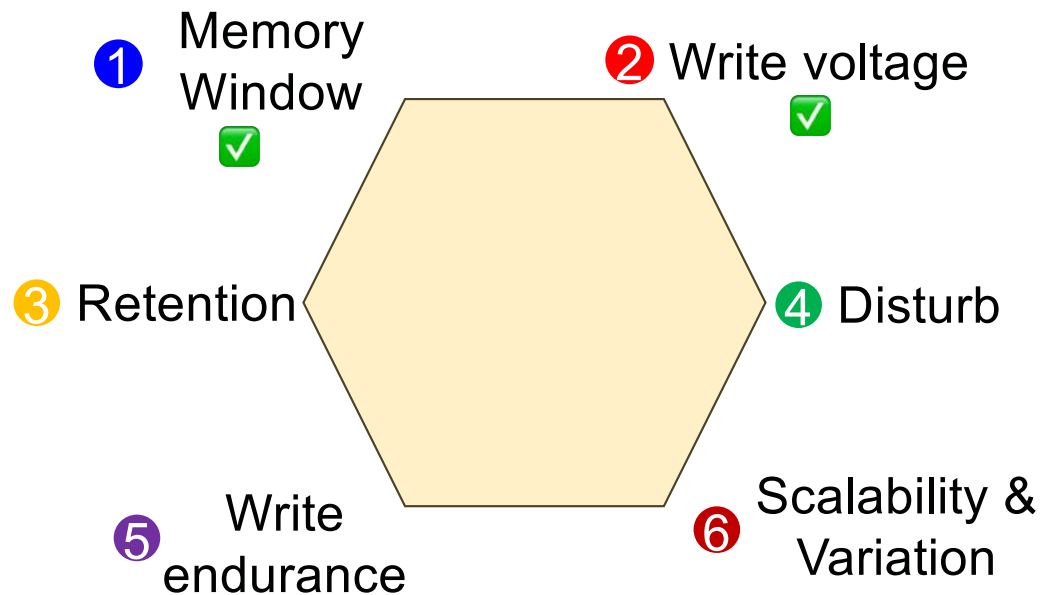
Trap effects are responsible for the observed MW trend
(MW: 8/3/8 > 9/1/9 > 19)

Das, D., et al. IEDM 2023: [10.1109/IEDM45741.2023.10413697](https://doi.org/10.1109/IEDM45741.2023.10413697)

Charge trap flash vs ferroelectric flash

	Structure	Characteristics	Band diagram at FB condition	Attributes
Charge Trap Flash	 Charge Trap Nitride			✓ Large MW × Large Write Voltage Mechanism: Trapping
Conventional FEFET	 Ferroelectric (FE)			× Low MW ✓ Low Write Voltage Mechanism: Polarization Switching
Proposed FEFET	 Al₂O₃ tunnel Oxide			✓ Large MW ✓ Moderate Write Voltage Mechanism: Polarization Switching + Trapping + Tunneling

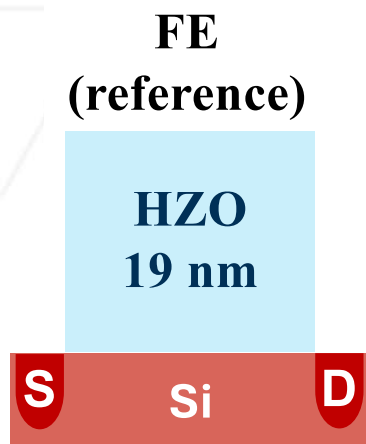
Evaluating ferroelectric NAND Flash



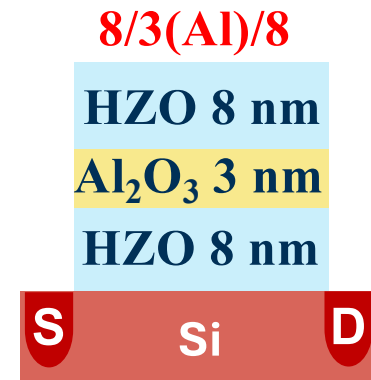
Venkatesan, Khan, et al. "Pushing the limits of NAND technology scaling with ferroelectrics: P. Venkatesan et al." *MRS Bulletin* 50.9 (2025): 1094-1107.



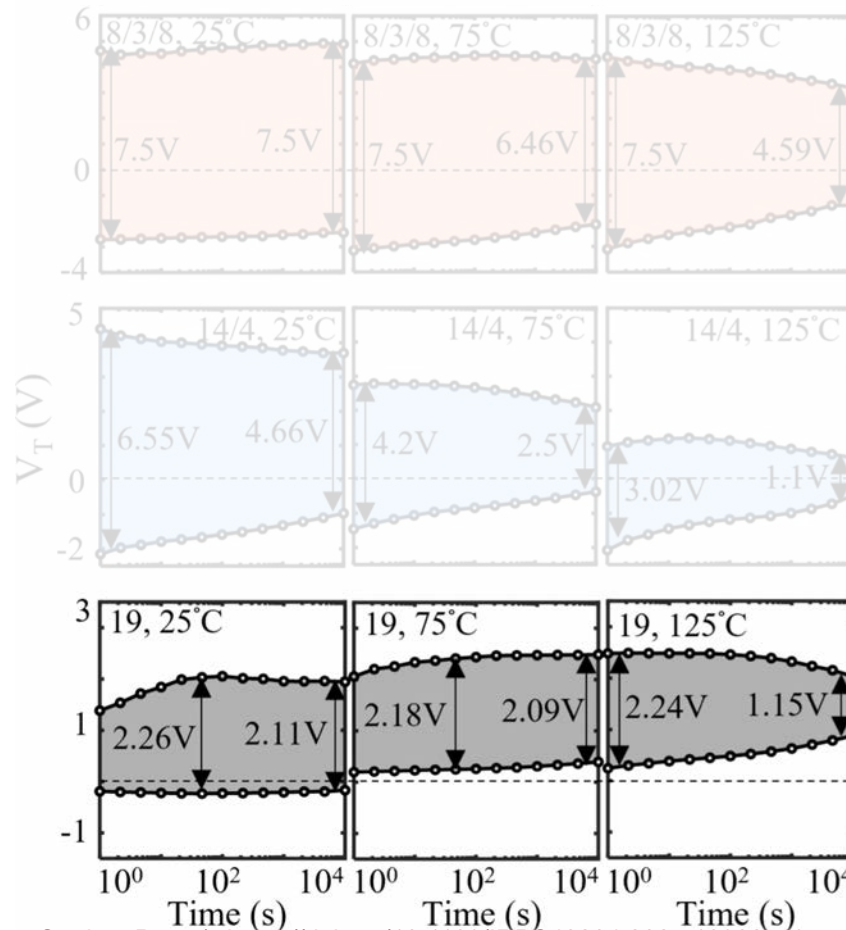
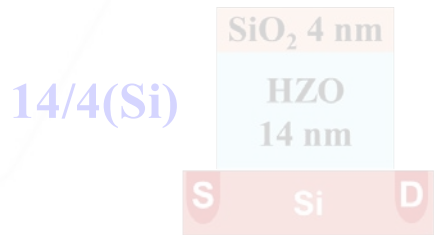
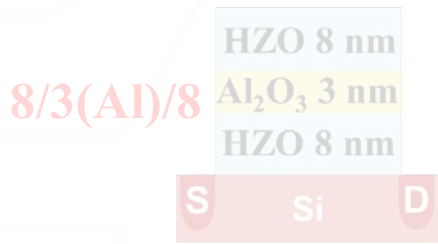
Ferroelectric NAND gate stacks



Gate blocking layer (GBL) **Tunnel dielectric layer (TDL)**

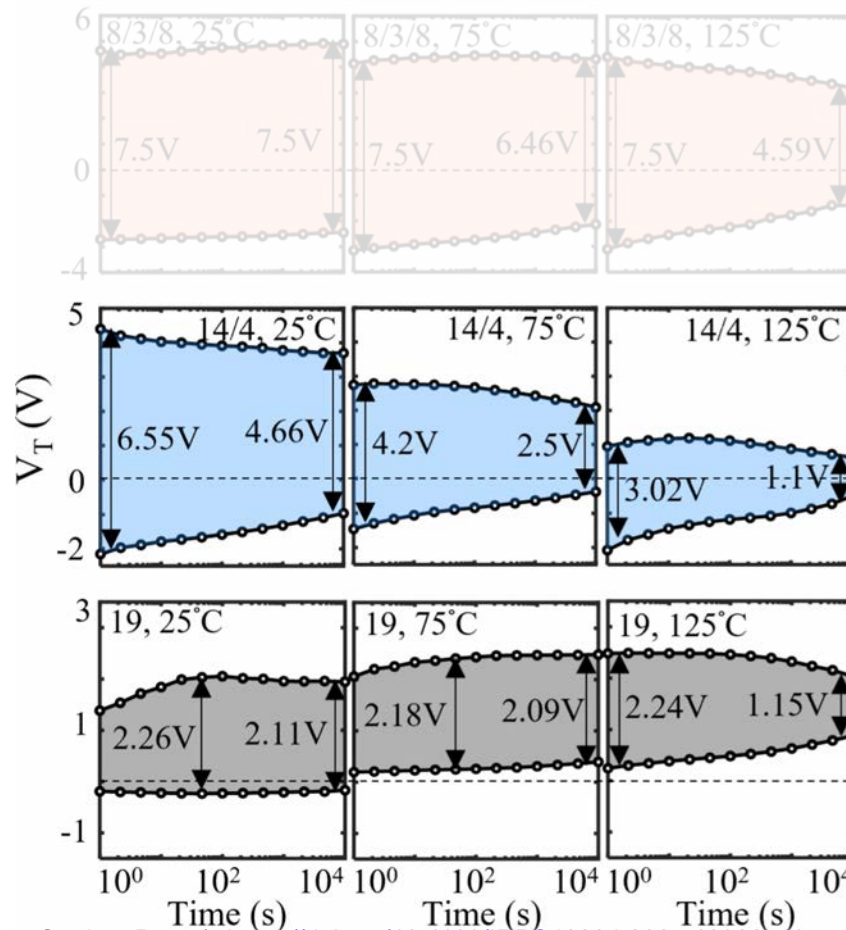
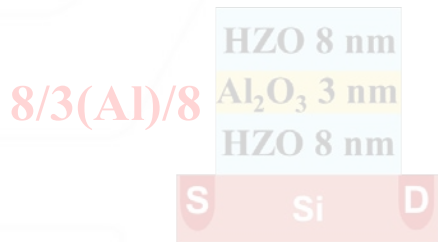


Temperature Dependent Retention for Fe-NAND stacks



Reference stack shows slight degradation at Higher temperature

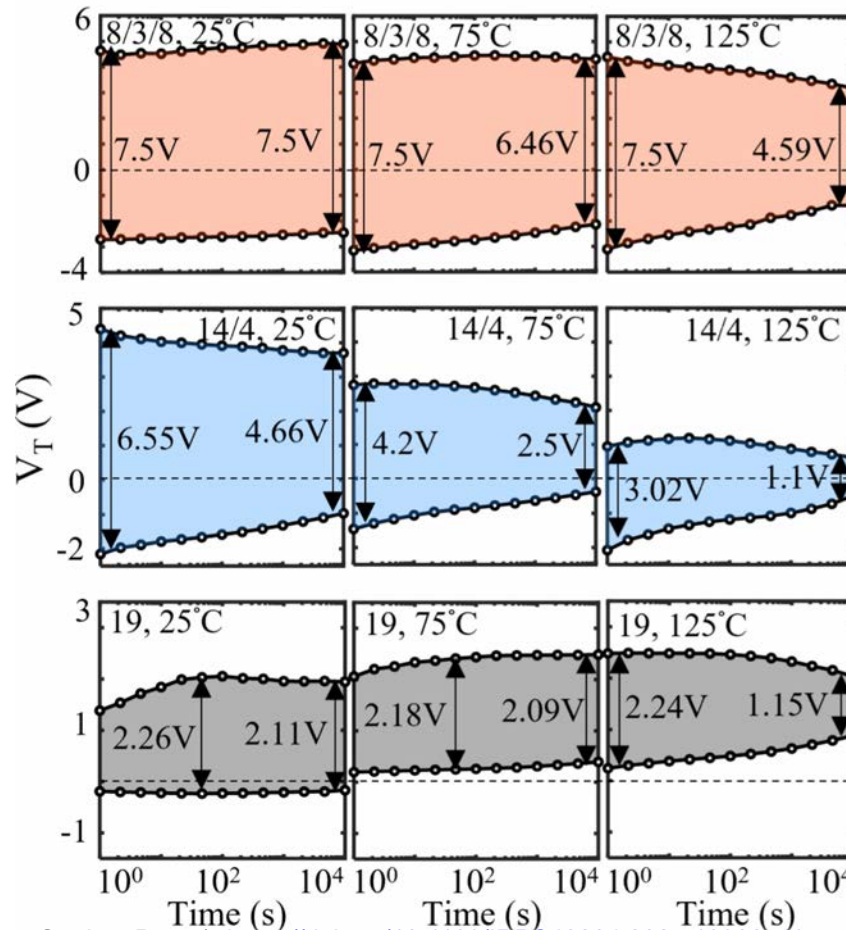
Temperature Dependent Retention for Fe-NAND stacks



GBL stack has lower initial MW at high temperature, but MW loss remain the same

Reference stack shows slight degradation at Higher temperature

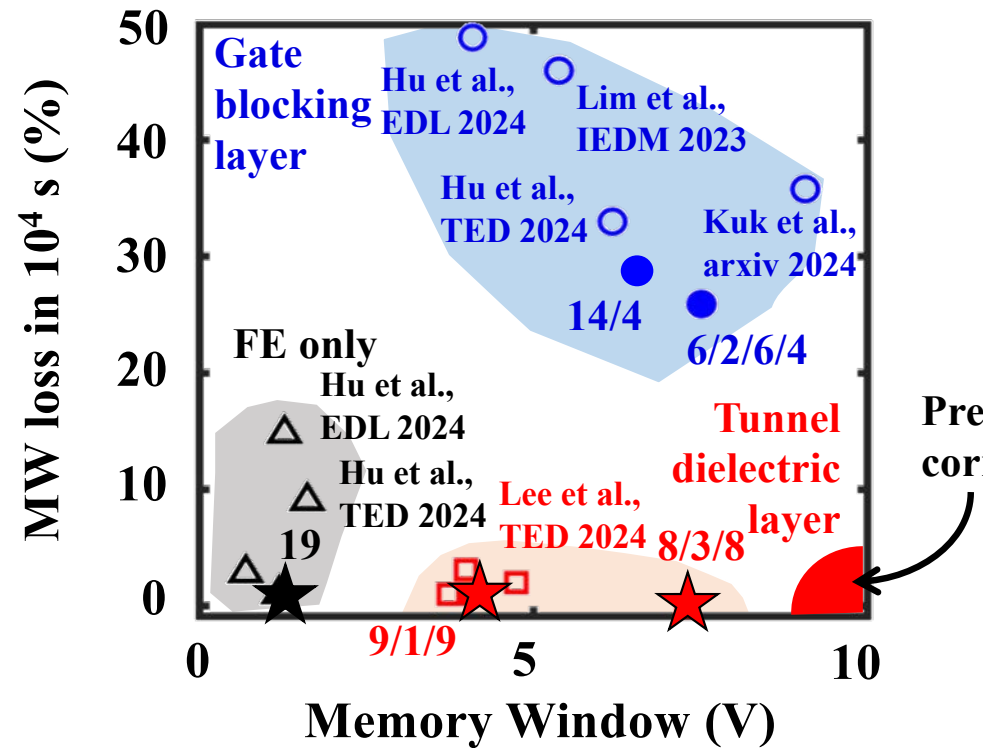
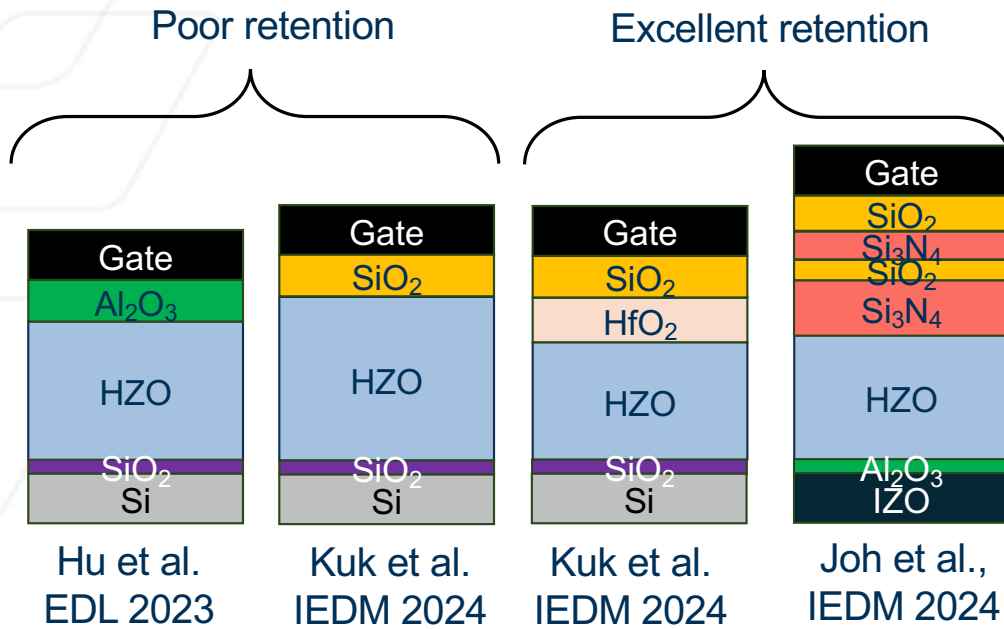
Temperature Dependent Retention for Fe-NAND stacks



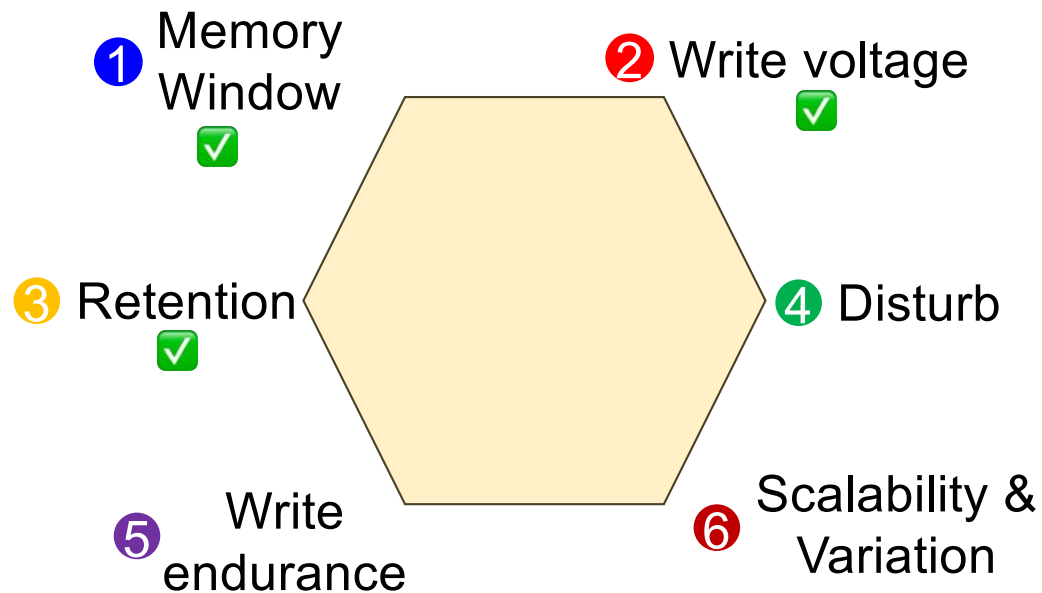
GBL stack has lower initial MW at high temperature, but MW loss remain the same

Reference stack shows slight degradation at Higher temperature

Retention benchmark



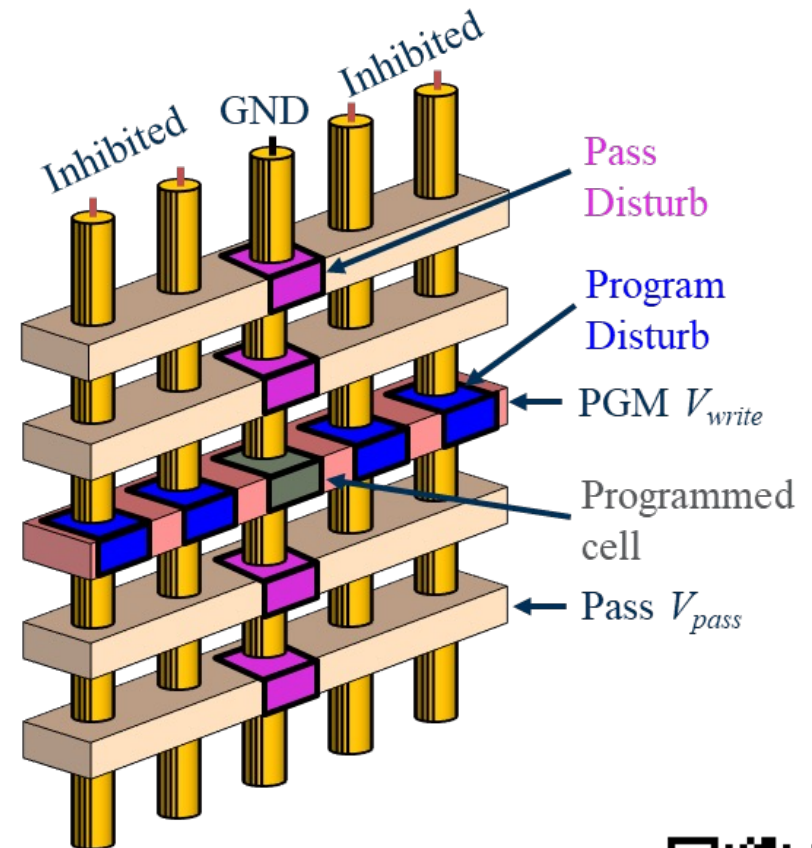
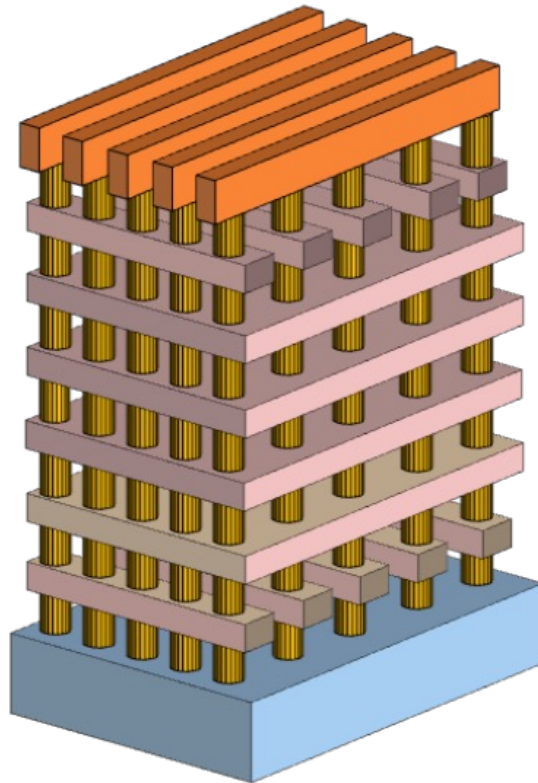
Evaluating ferroelectric NAND Flash



Venkatesan, Khan, et al. "Pushing the limits of NAND technology scaling with ferroelectrics: P. Venkatesan et al." *MRS Bulletin* 50.9 (2025): 1094-1107.



Disturb in Ferroelectric NAND

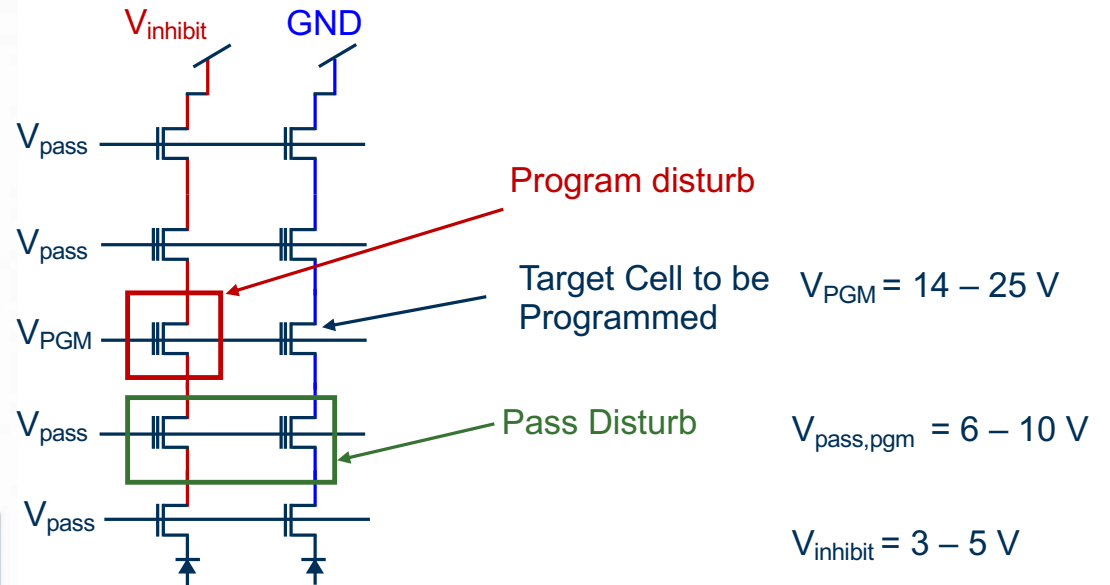
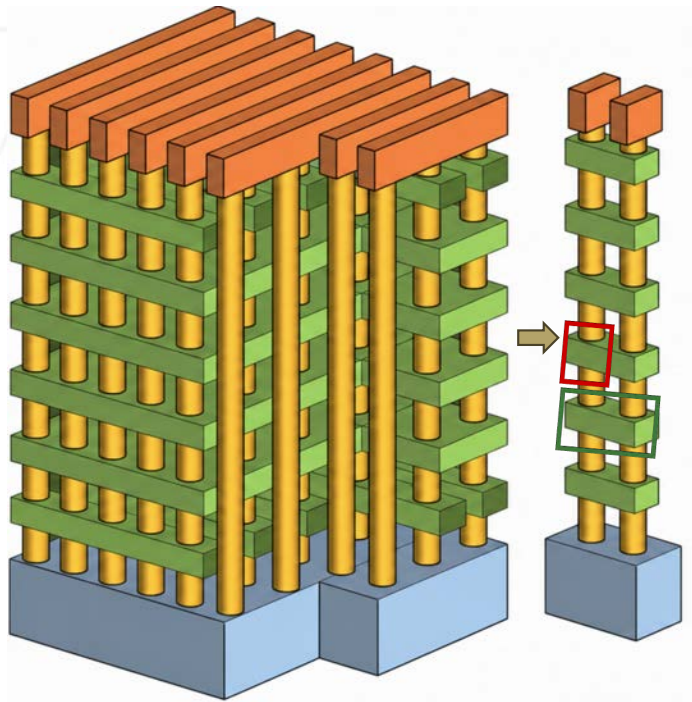


A pass voltage is applied to unaddressed cells while accessing a different cell along the bit line. This pass voltage could disturb the V_T of the cell either due to electron trapping or polarization back switching.

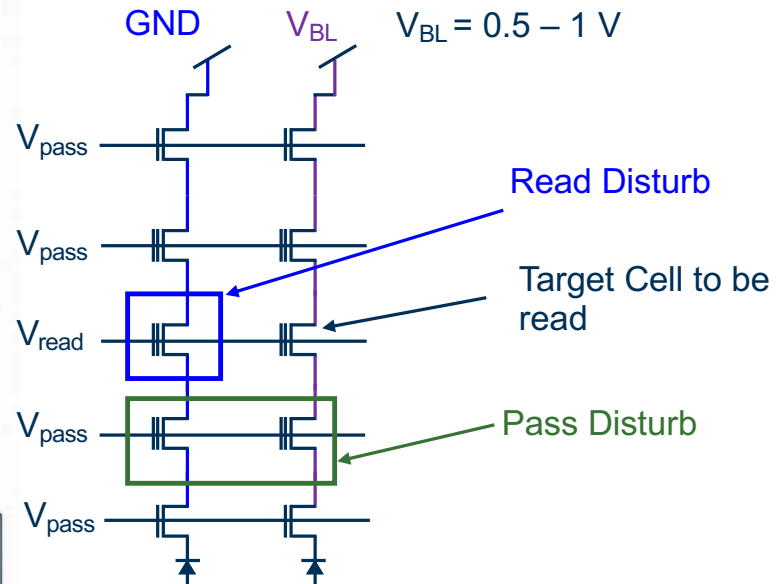
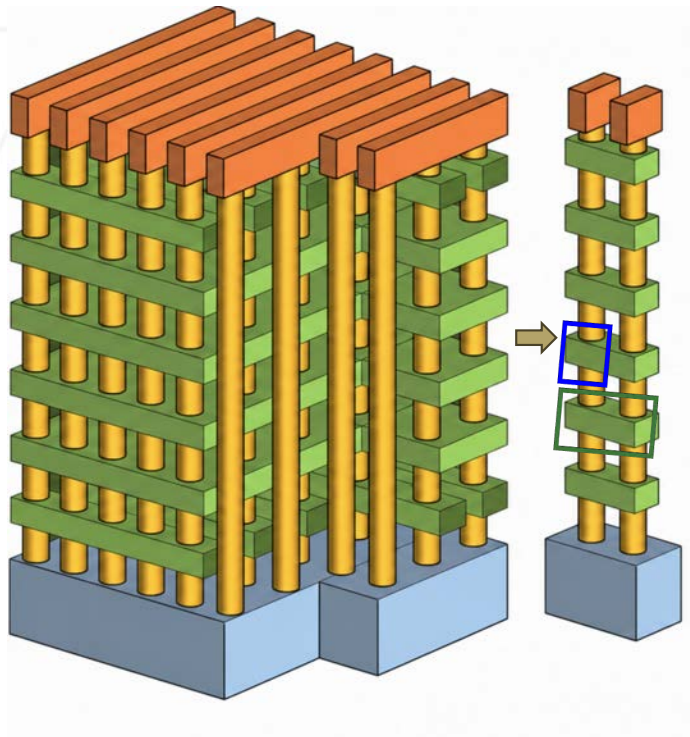
Venkatesan, P., et al. EDL 2024: [10.1109/LED.2024.3467210](https://doi.org/10.1109/LED.2024.3467210)



Program and Pass Disturb



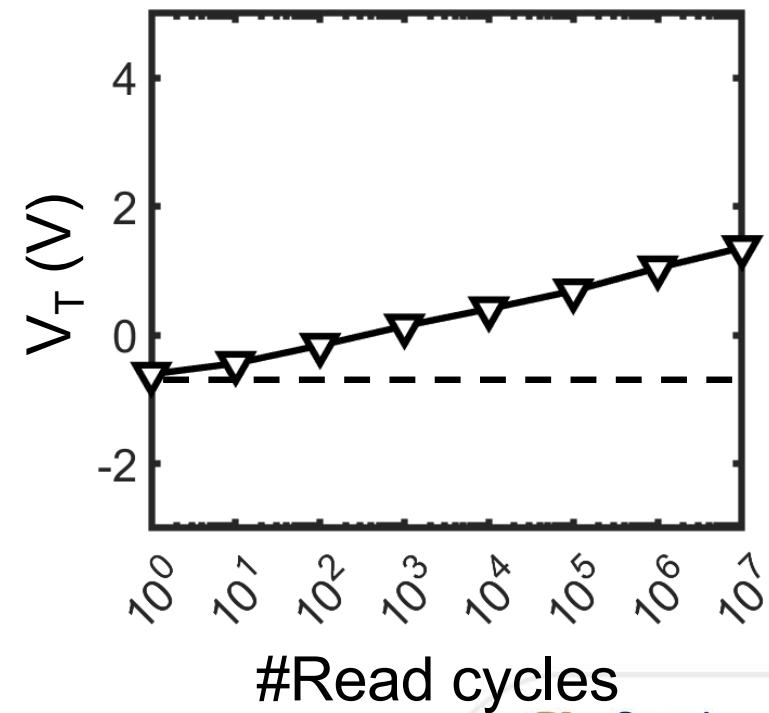
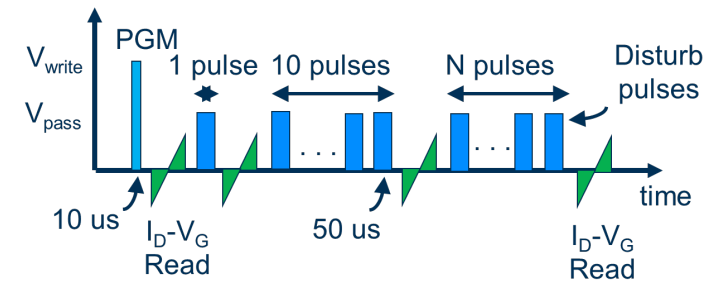
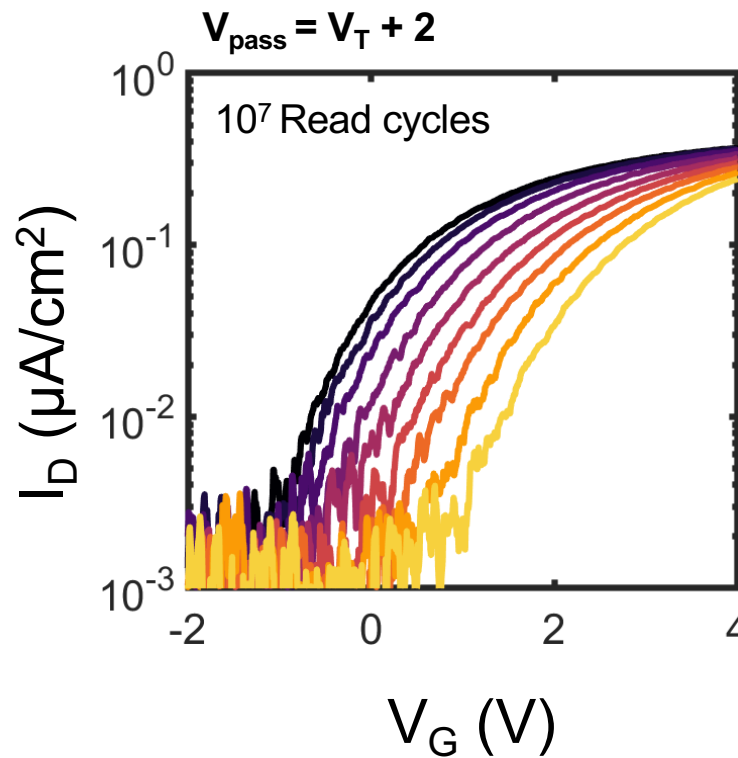
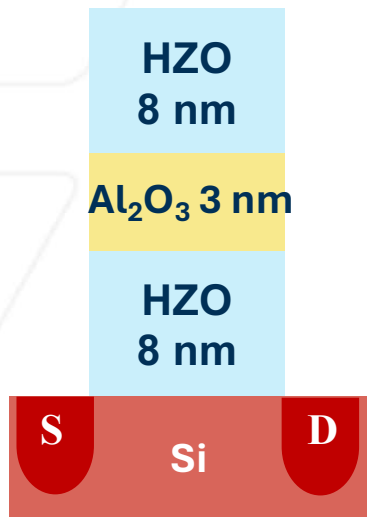
Read and Pass Disturb



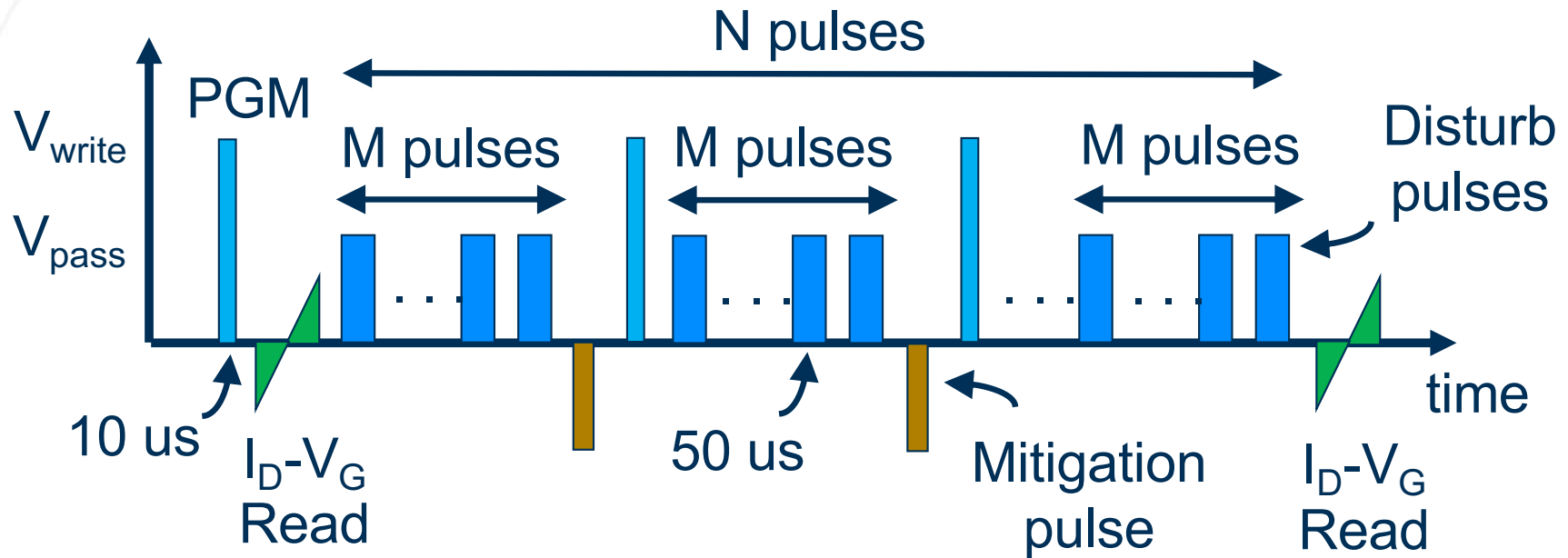
$$V_{read} = 0 - 5 \text{ V}$$

$$V_{pass,read} = 6 - 8 \text{ V}$$

Disturb characterization

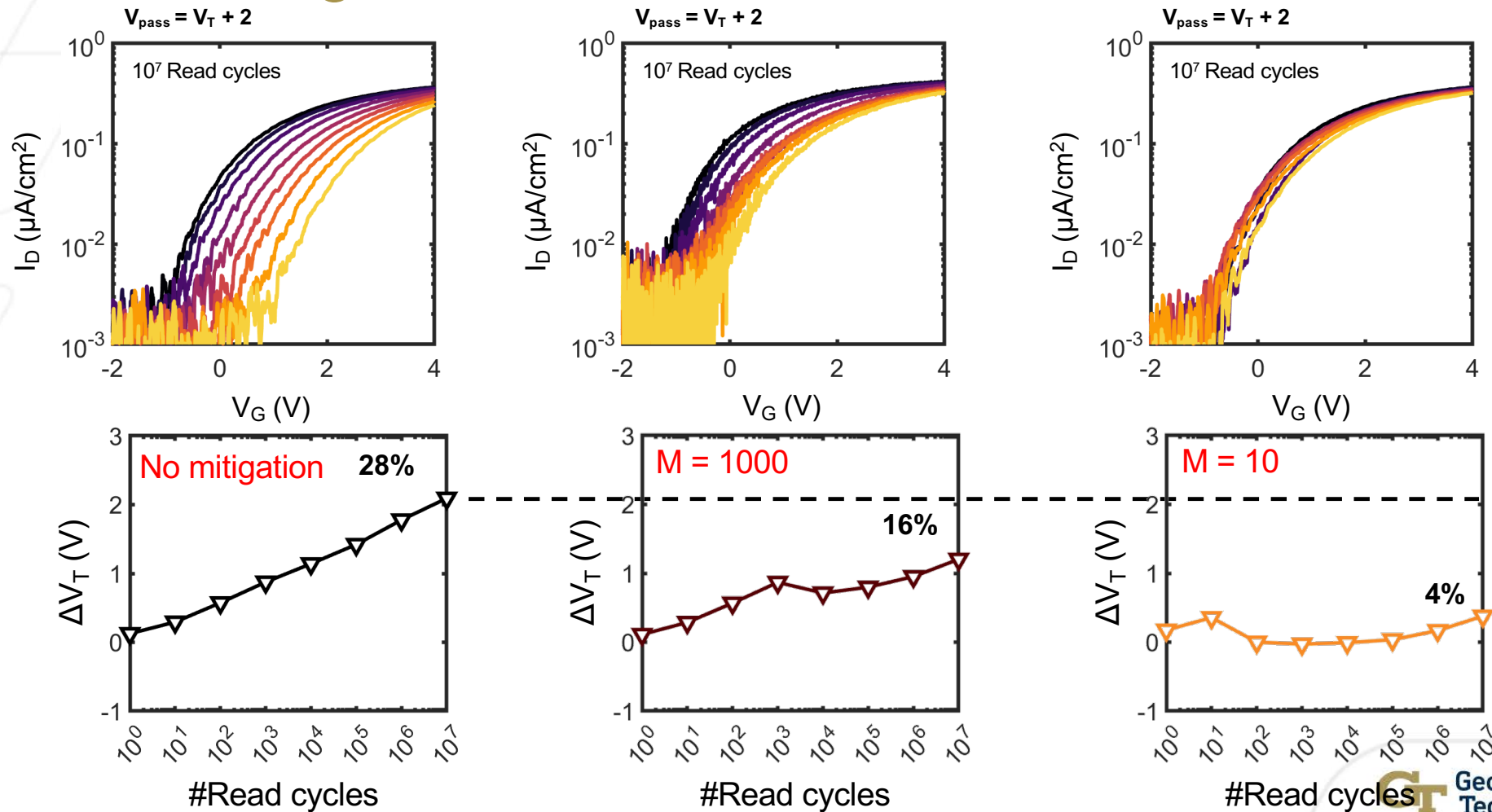


A Disturb Mitigation Scheme for ferroelectric NAND

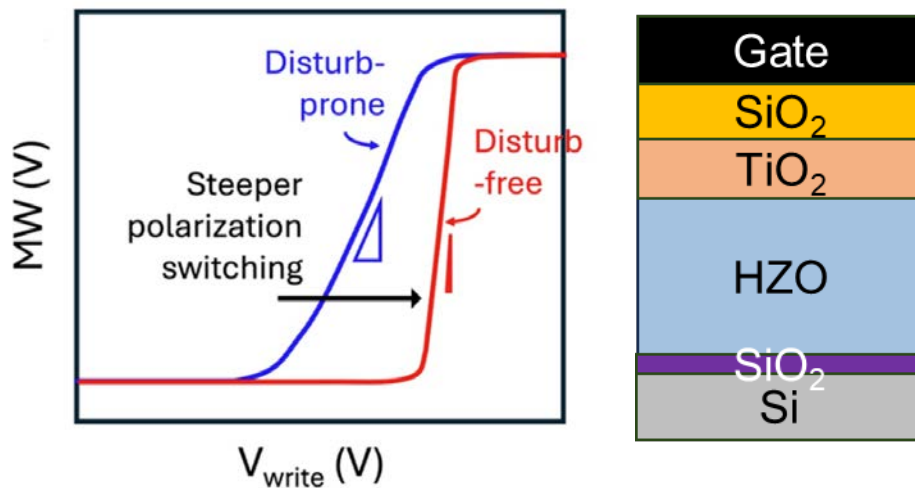


We introduce a mitigation pulse to detrap the electrons trapped due to the disturb pulses

A Disturb Mitigation Scheme for ferroelectric NAND

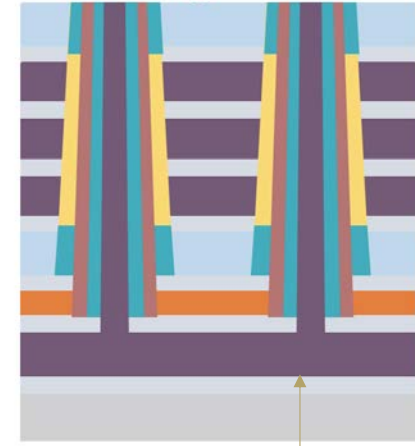


Techniques to mitigate disturb



Inserting a TiO_2 layer in GBL can reduce disturb.

G. Kim, et. al., IEDM 2024,
doi: 10.1109/IEDM50854.2024.10873411

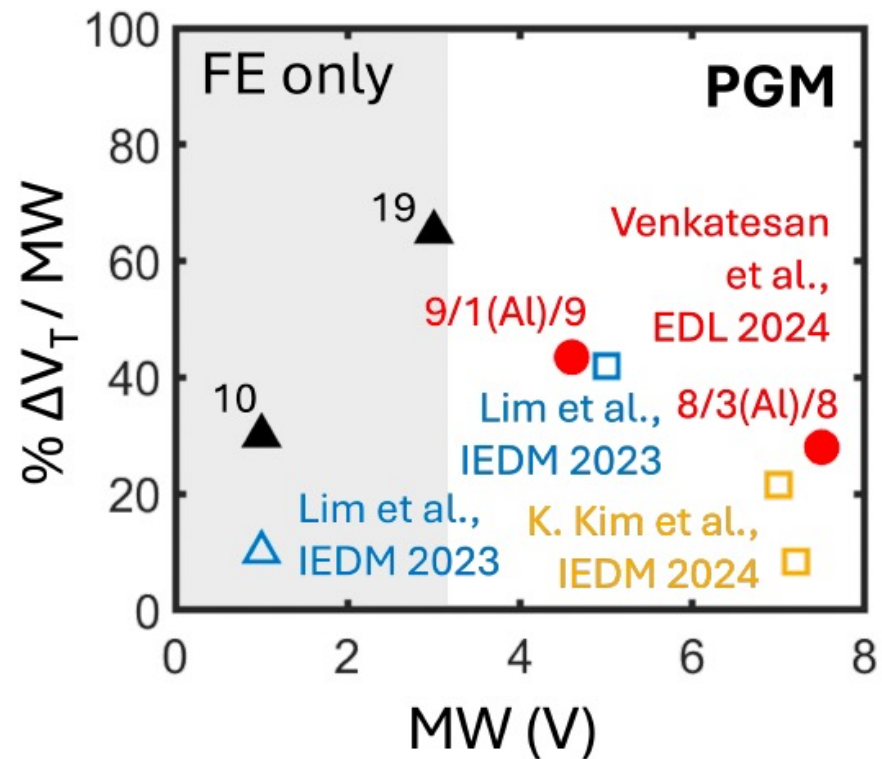
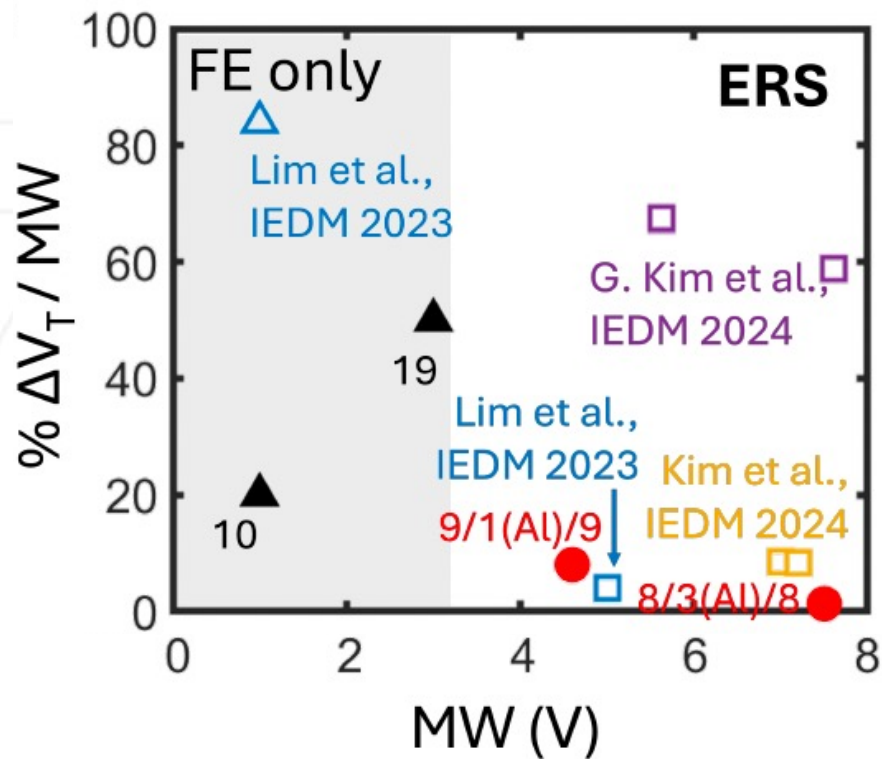


Oxide filler replaced by a metallic filler.

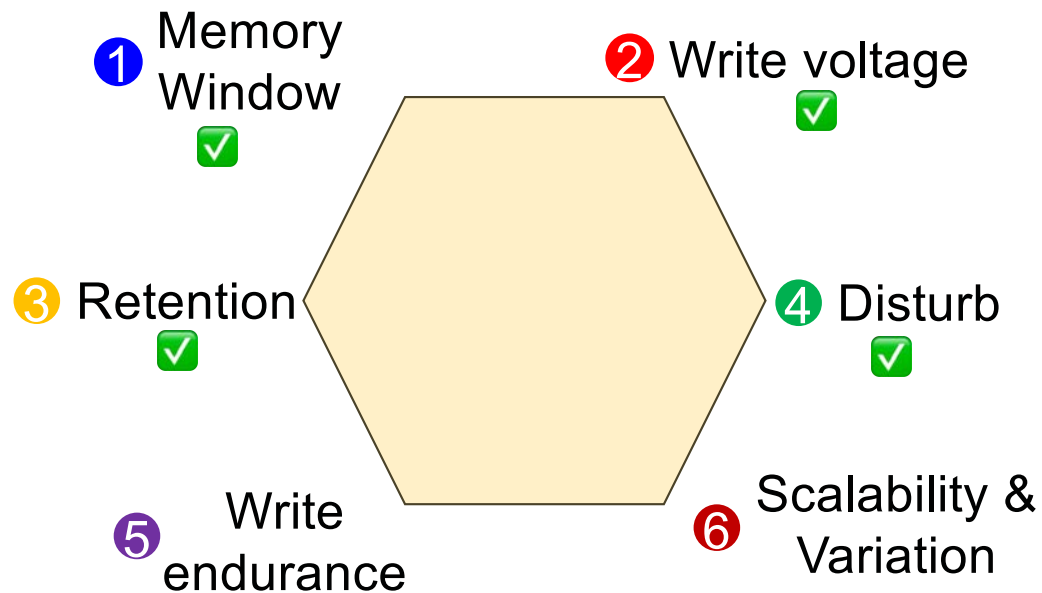
Pass operation is performed by a separate non-ferroelectric gate stack so that the ferroelectric state is not disturbed

X. Niu, et. al., IEDM 2025,
doi: doi: 10.1109/IEDM50572.2025.11353563

Pass disturb



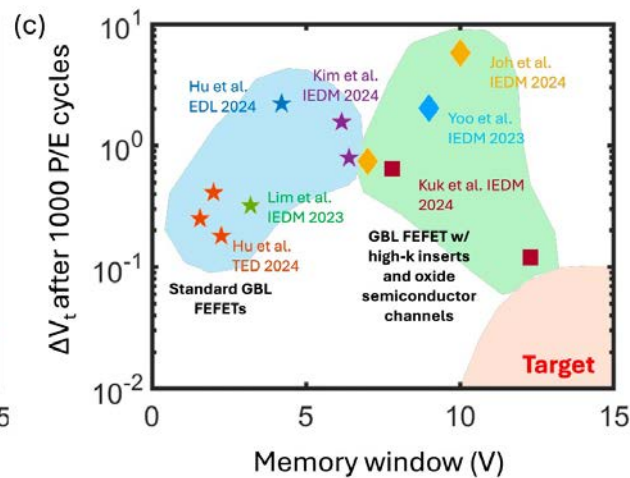
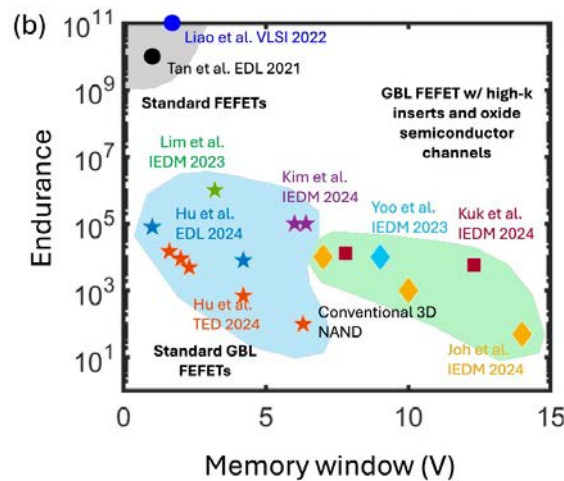
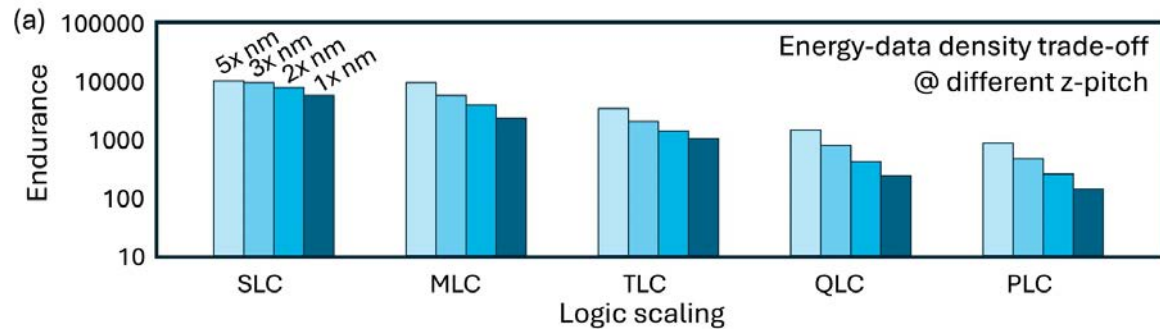
Evaluating ferroelectric NAND Flash



Venkatesan, Khan, et al. "Pushing the limits of NAND technology scaling with ferroelectrics: P. Venkatesan et al." *MRS Bulletin* 50.9 (2025): 1094-1107.



Write Endurance



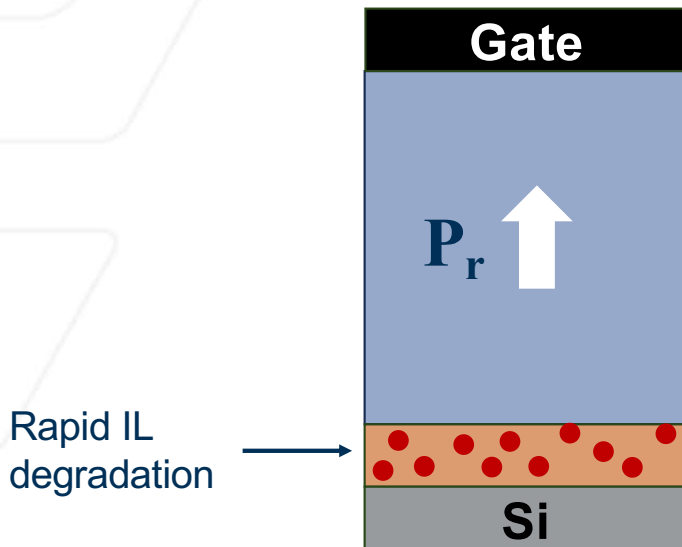
Endurance decreases with increasing MW

IL degradation drives V_t shift in FEFETs

Venkatesan, P., et al. IMW 2025

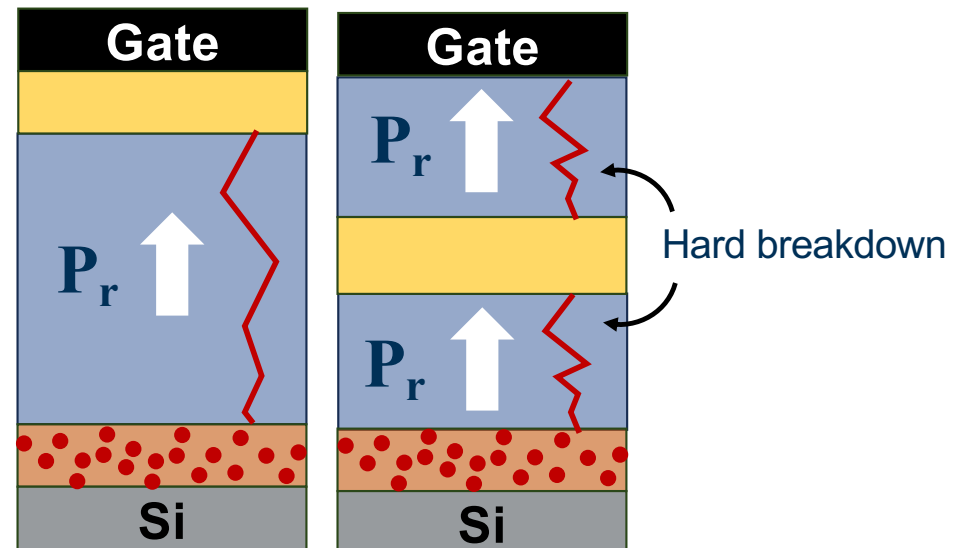
Write Endurance

Standard FEFETs



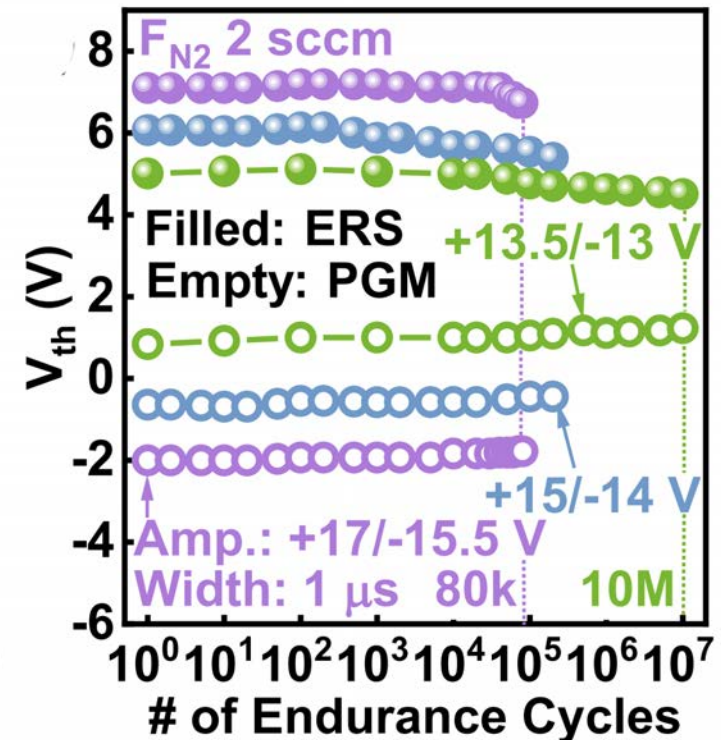
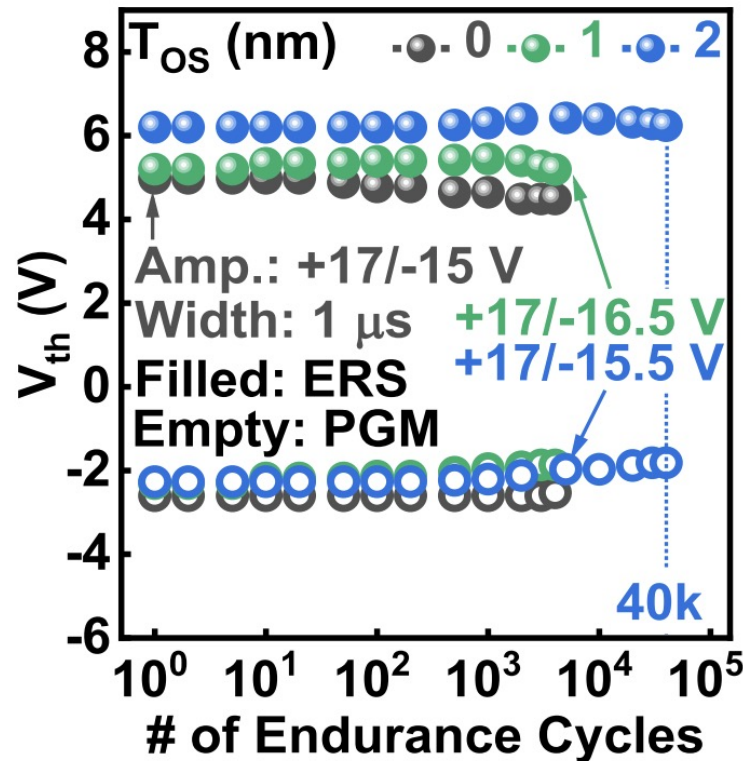
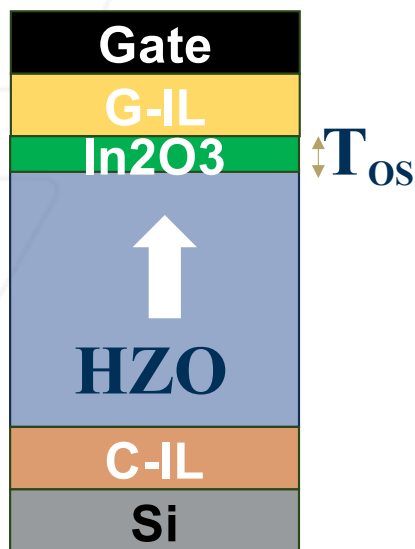
- Small MW
- $V_{\text{write}} / V_{\text{BD}}$ lower than 1
- IL degradation dominates
- MW closure
- Minimal FE degradation

Engineered FEFETs



- Large MW
- $V_{\text{write}} / V_{\text{BD}}$ close to 1
- Rapid IL degradation
- Increased risk of hard breakdown

Insertion of Oxygen Reservoir Layer to improve Fe-NAND endurance

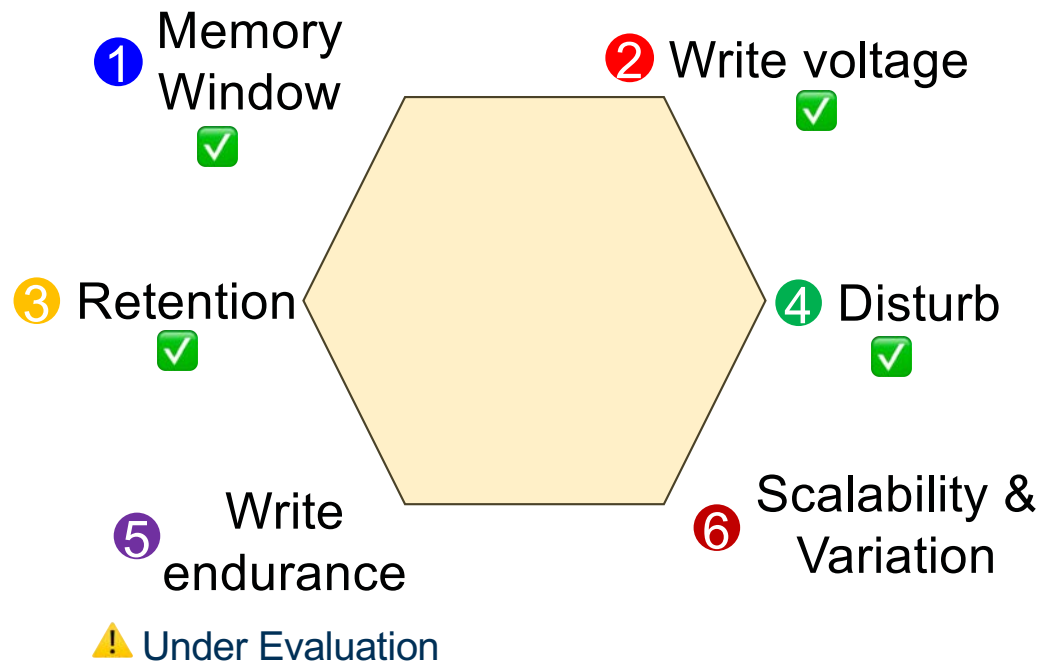


Introducing In2O3 layer acts as oxygen reservoir layer for HZO layer to enable larger MW and high endurance by optimizing the thickness and deposition parameters of the ORL layer

64

H. Kang, et, al., IEDM 2025, TED 2026

Evaluating ferroelectric NAND Flash

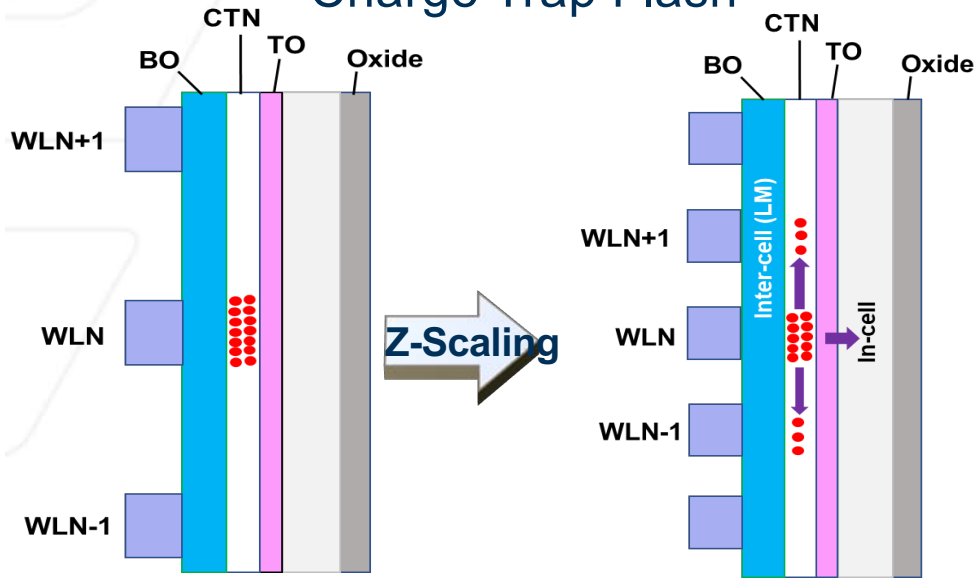


Venkatesan, Khan, et al. "Pushing the limits of NAND technology scaling with ferroelectrics: P. Venkatesan et al." *MRS Bulletin* 50.9 (2025): 1094-1107.



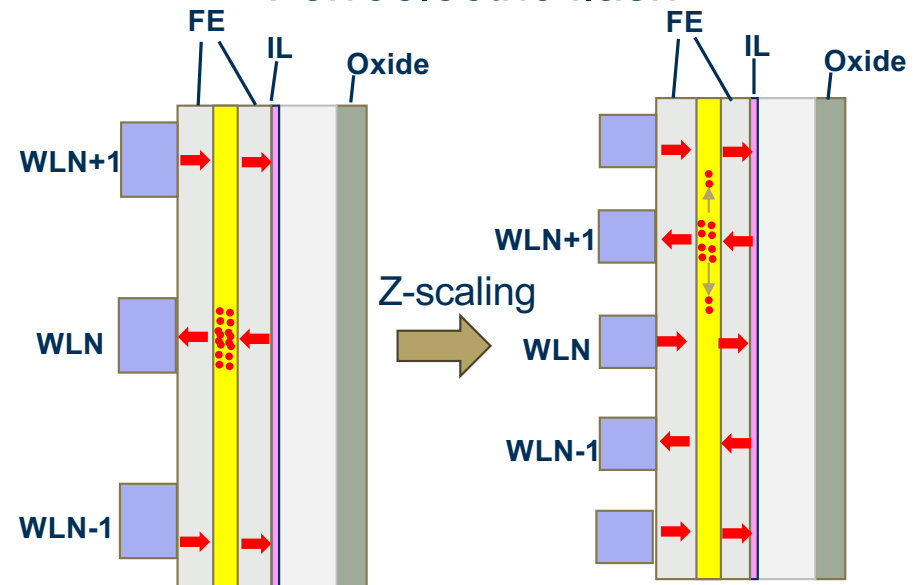
Lateral charge migration

Charge Trap Flash



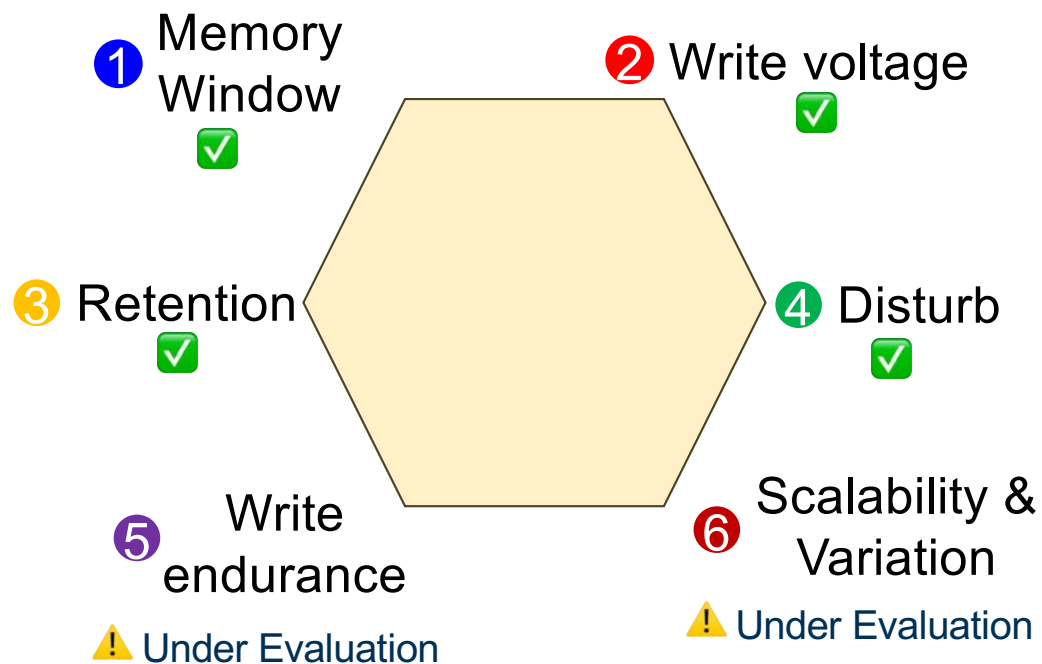
Lateral charge migration prevents aggressive Z-scaling for 3D NAND.

Ferroelectric flash



The nature of trapped charges at the ferroelectric—interlayer interface will determine the scaling potential of ferroelectric 3D NAND technology.

Evaluating ferroelectric NAND Flash



Venkatesan, Khan, et al. "Pushing the limits of NAND technology scaling with ferroelectrics: P. Venkatesan et al." *MRS Bulletin* 50.9 (2025): 1094-1107.



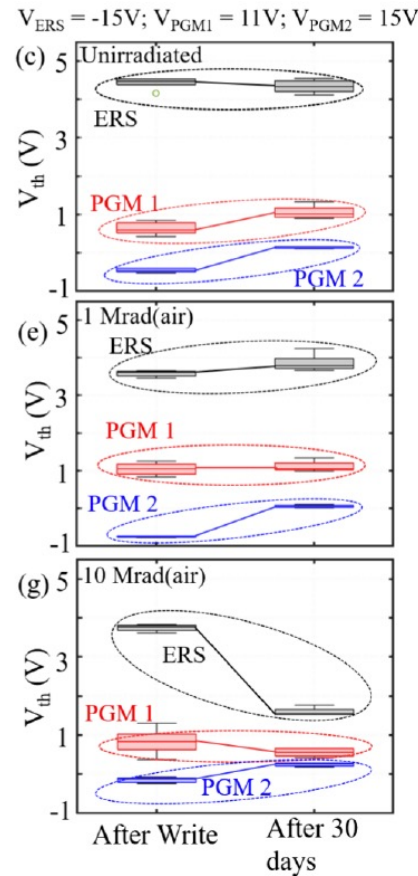
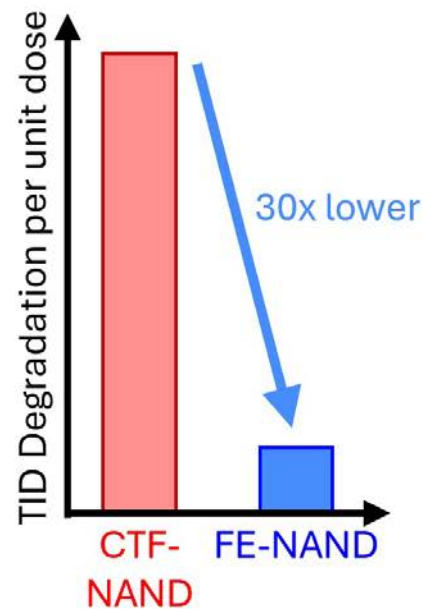
Space applications and ferroelectric NAND flash

- Space is becoming a key infrastructure layer for **AI, sensing, communication, navigation, and defense**.
 - **Spacecraft memories must retain data under extreme radiation.**
 - Radiation dose increases sharply with mission orbit:
LEO satellites: **5–30 krad**
GEO missions: **100–300 krad**
Jupiter/Europa missions: approaching **1 Mrad**
 - **CTF-NAND can show retention loss even at relatively low doses** of approximately **50 krad**, creating risk for long-duration missions.
 - **Magnetic storage remains impractical for many spacecraft systems** because of mass, volume, power, and mechanical constraints.
- 68
- Ferroelectric NAND flash may provide a solution.



Space applications and ferroelectric NAND flash

All properties of ferroelectric NAND transistors, including retention, disturb, and write endurance remain intact up to 1 Mrad TID.



Fernandes, Lance, et al. "Enabling Radiation Hardness in Solid-State NAND Storage Utilizing a Laminated Ferroelectric Stack." *Nano Letters* 26.10 (2026): 3390-3397.

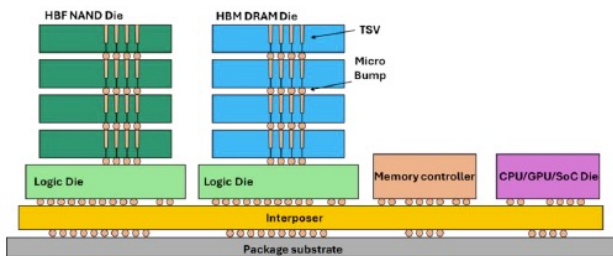


Future work: SEE, displacement damage, retention/endurance after irradiation, disturb behavior, and thermal recovery

Where NAND is Going Next

1

ACTIVE MEMORY TIER



Workload-tuned NAND

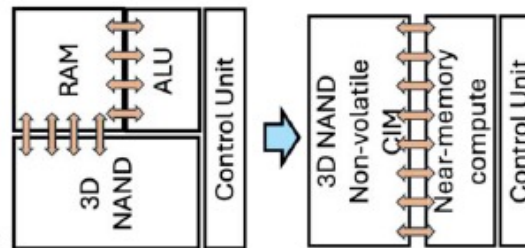
- QLC for cold archival
- SLC-like for inference buffers for AI

One NAND no longer fits all

the right cell for the right job

2

COMPUTE SUBSTRATE



Compute-in-Memory (CIM)

- MAC inside the NAND array
- In-storage vector search for RAG

The array becomes the computer

10–100× lower energy for memory-bound AI

3

RAD-HARD STORAGE



Orbital Infrastructure for AI

- Petabyte storage on satellites
- Deep space missions

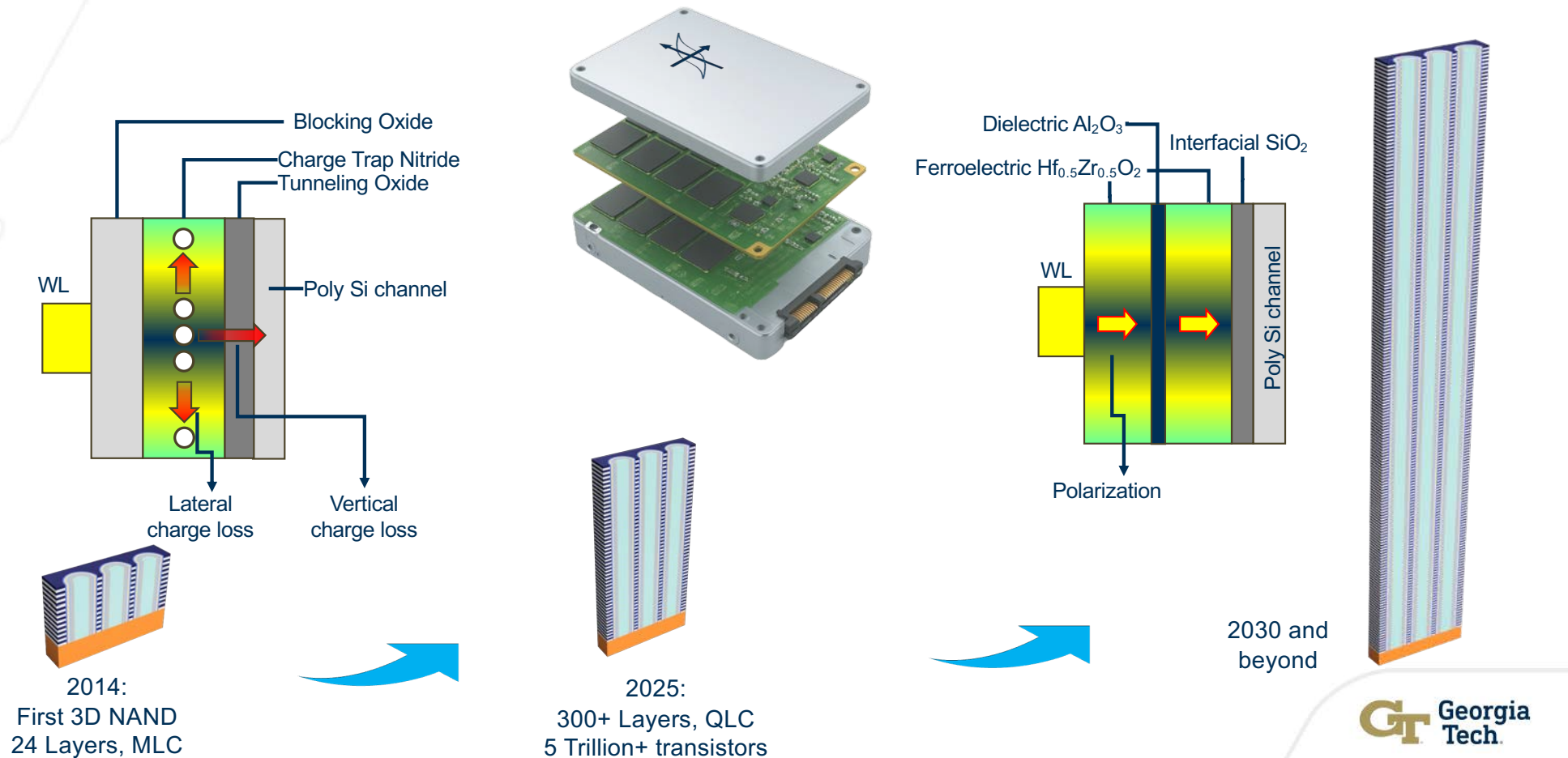
NAND leaves the data center

rad-hardness for orbital AI

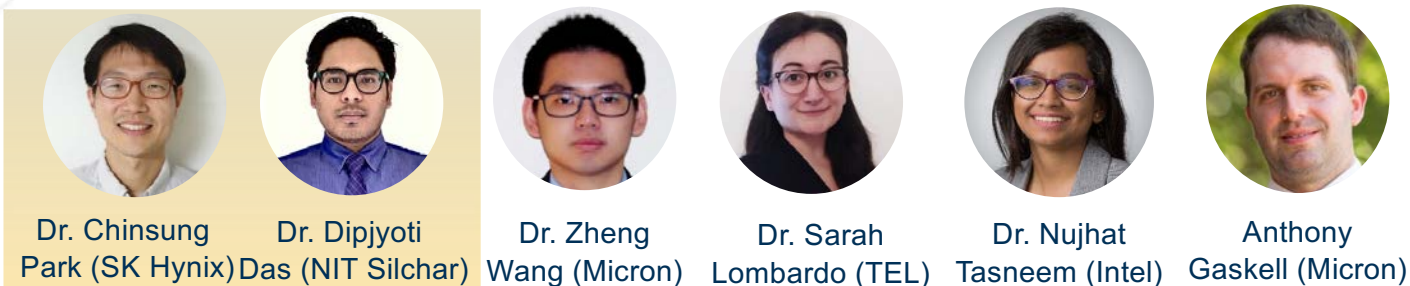
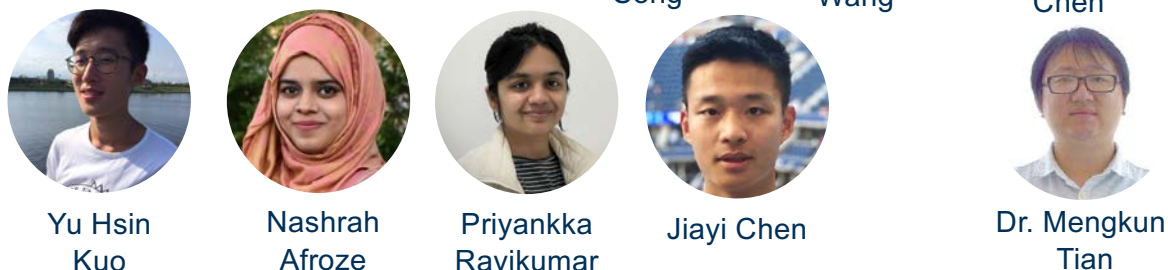
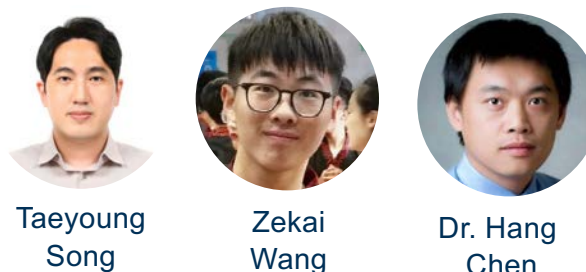
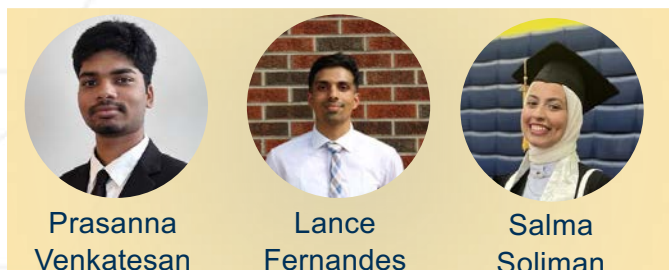
From a single technology to a portfolio — NAND, tuned to the workload

Ferroelectrics for Future NAND Storage Technology

Ferroelectric augmentation can provide enable NAND scaling in future generations



Our team



Collaborators:

Prof. Andrea Padovani, DIEF, UNIMORE
 Dr. Gaurav Thareja, Dr. Luca Larcher, Applied Materials
 Prof. Lauren Garten, Prof. Shimeng Yu, Prof. Suman Datta, Georgia Tech

Our sponsors





midtown

ATLANTA

