Ferroelectric NAND Flash: Scaling Storage for the AI era

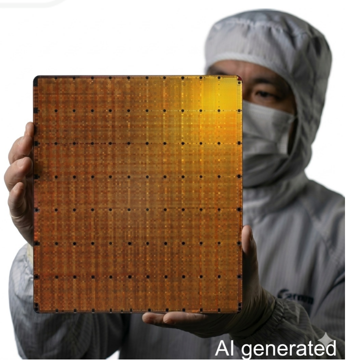
Asif Khan
Associate Professor
School of Electrical and Computer Engineering
School of Materials Science and Engineering
Georgia Institute of Technology
akhan40@gatech.edu

IEEE International Conference on IC Design and Technology (ICICDT 2026)
Dresden, Germany | 23 June, 2026

Georgia Tech

Foundational Memory & Storage Technologies

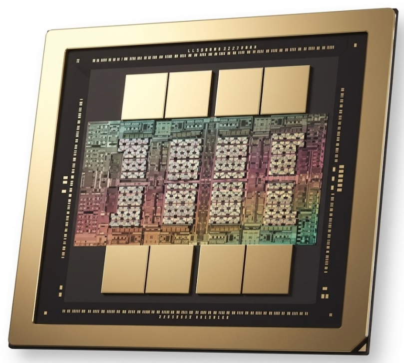
1 SRAM



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.

Georgia Tech

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

~50×

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

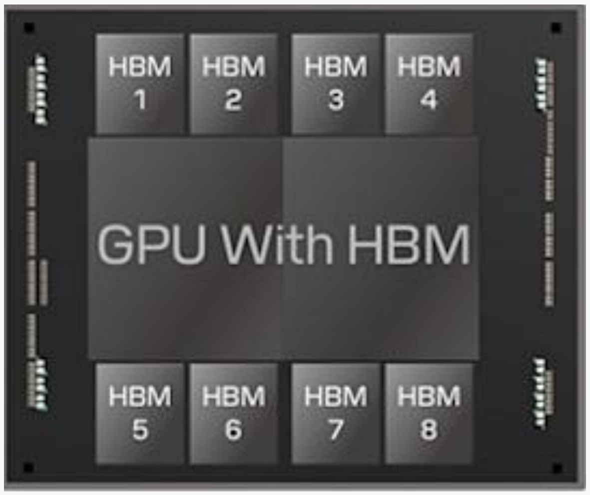
NAND is no longer where the data sleeps.
It's increasingly part of where the AI thinks.

Georgia Tech

What if NAND moved on Package?

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

GPU + HBM



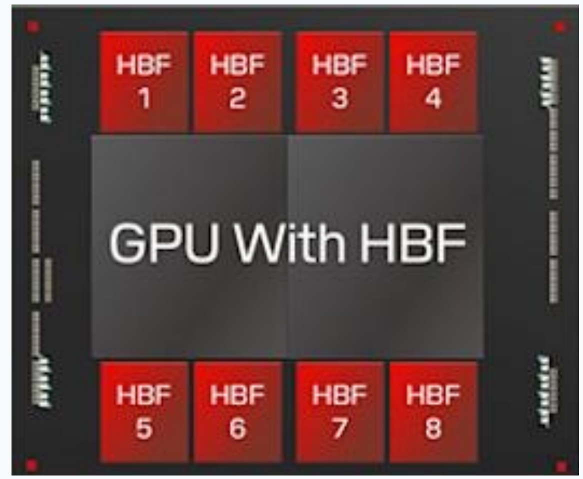
192 GB

TOTAL ON-PACKAGE MEMORY

~8 TB/s · Shipping today (B200)

~21×

GPU + HBF



4,096 GB

TOTAL ON-PACKAGE MEMORY

TB/s class · In development

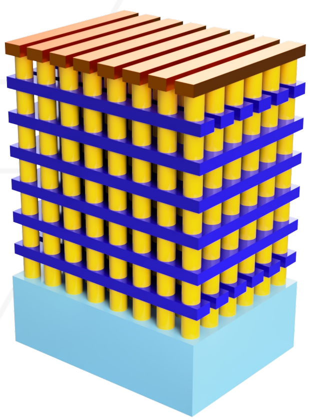
Concept and figure inspired by SanDisk High Bandwidth Flash (HBF). KV caches, mixture-of-experts weights, long-context attention, large embedding tables

21× more memory, on package, HBM-class bandwidth, NAND NVMe-like capacity

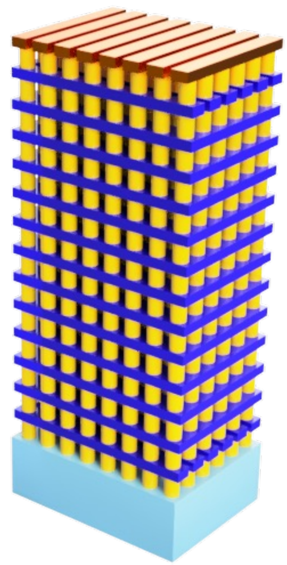
Georgia Tech

https://documents.sandisk.com/content/dam/asset-library/en_us/assets/public/sandisk/collateral/company/Sandisk-HBF-Fact-Sheet.pdf#page=1.00

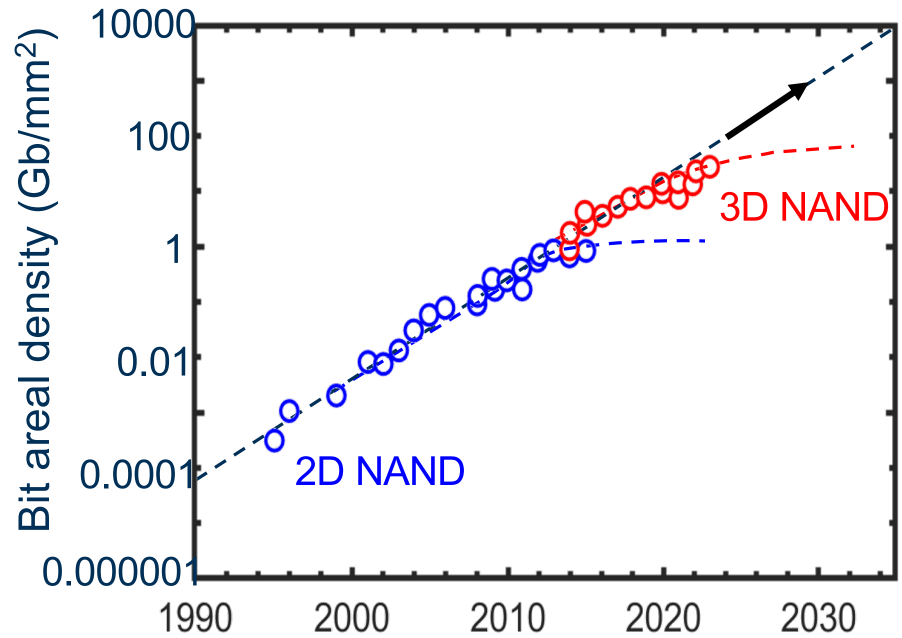
Vertical NAND Scaling



2014
First 3D NAND, Samsung
24 Layers, MLC



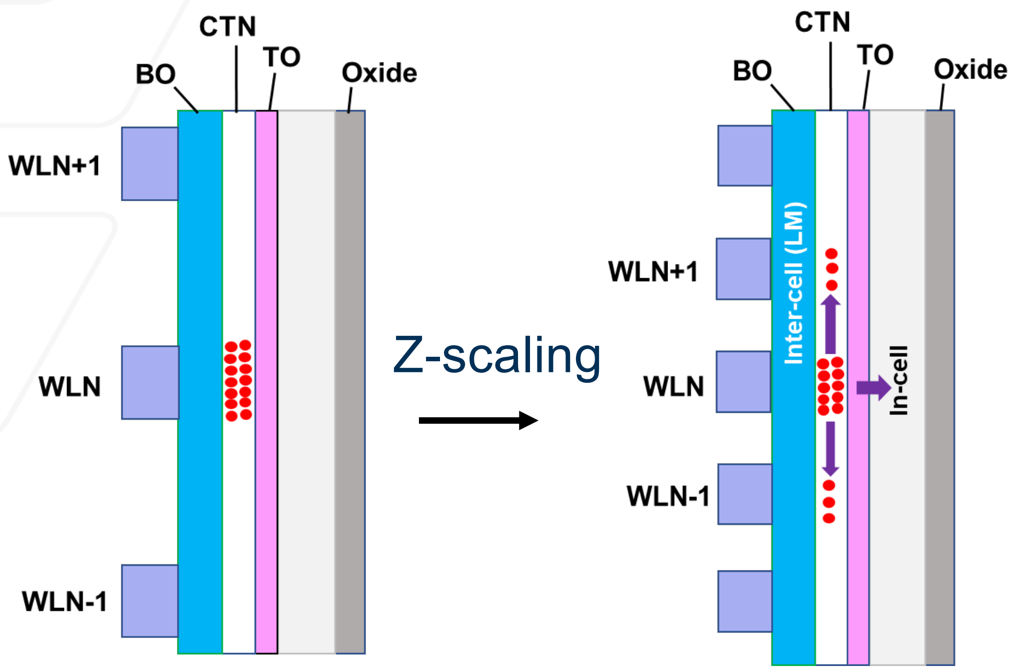
2022-2026
Production
Samsung 236L TLC
SK hynix 238L TLC → 321L (ramping 1H25)
Micron 276L TLC
Prototype / R&D demos:
SK hynix 321L sample (FMS 2023)
SK hynix 300+L demo (ISSCC 2023)
Kioxia / WD >300L demo (VLSI 2023)
Kioxia / Sandisk 1000L using Multi-Stacked Cell Array – CMOS Bonded to Array (2026)



A. Goda, IEDM 2024

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

Challenges of vertical NAND flash scaling



❶ Lateral charge migration

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

❷ High write voltage

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

❸ Coupling and disturb

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

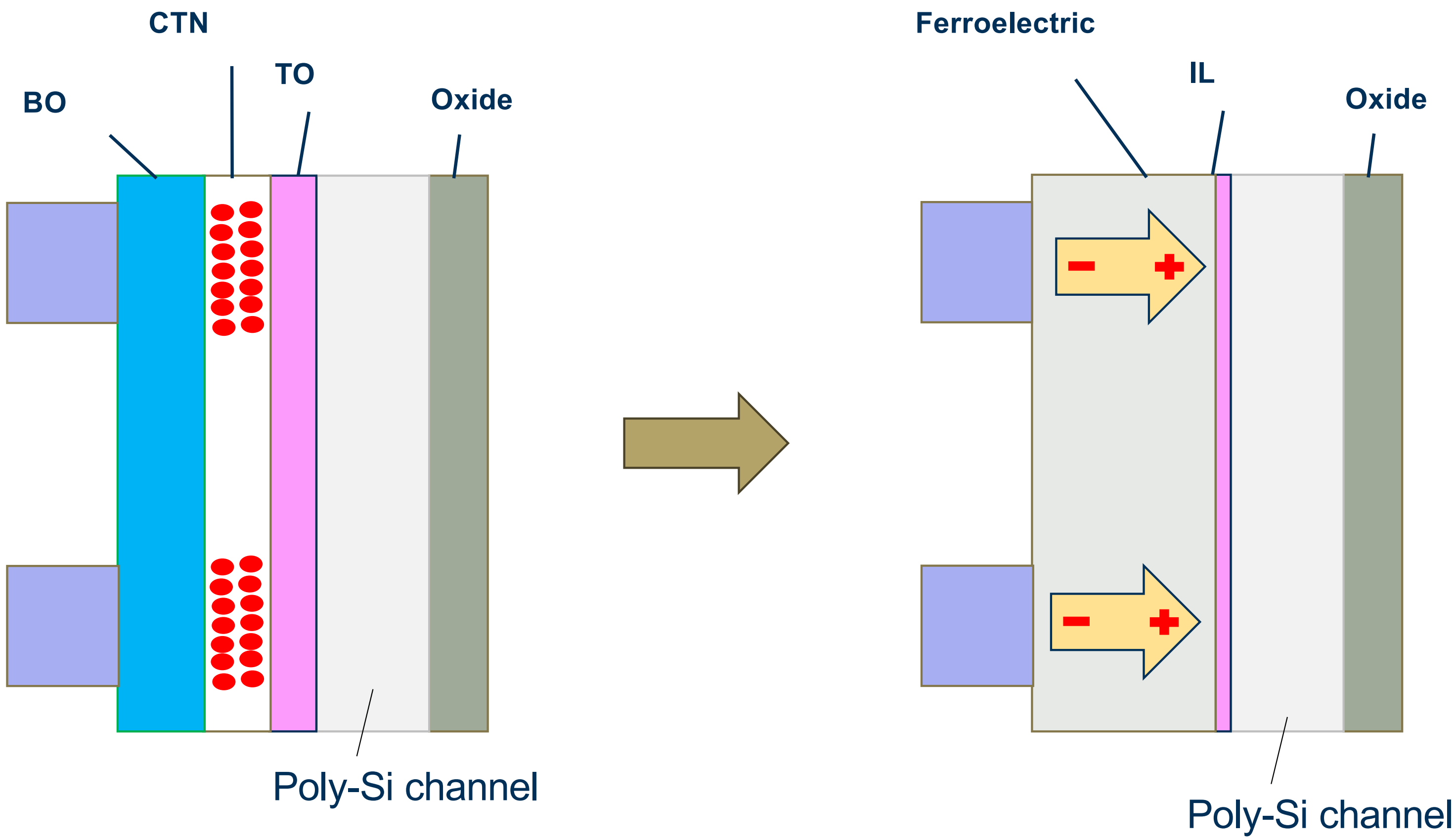
❹ Variability and read margin

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

Georgia Tech

6

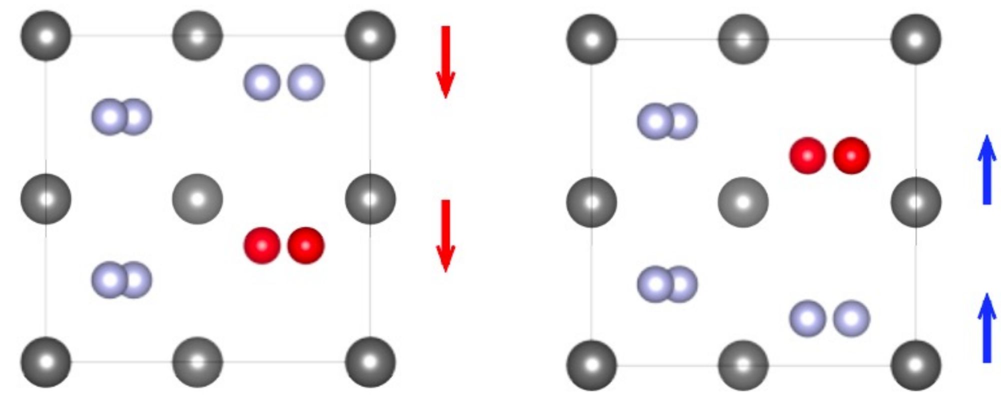
Can ferroelectricity enable next generation NAND flash?

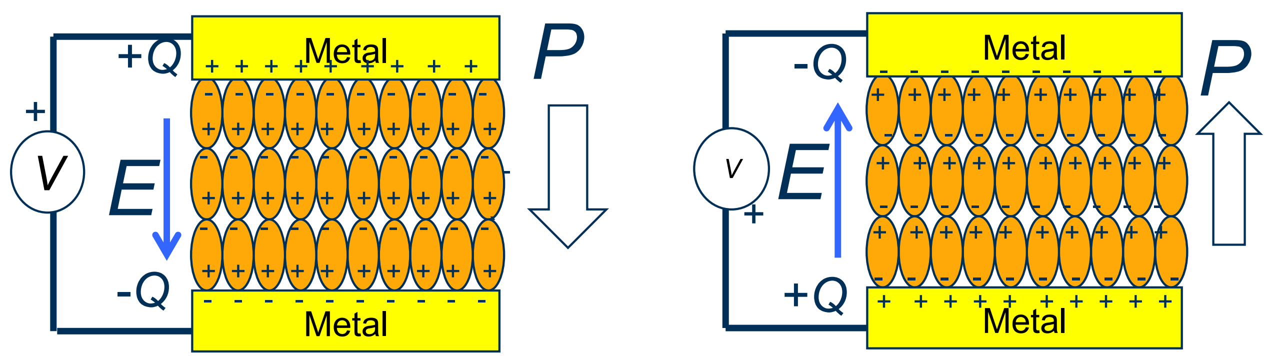


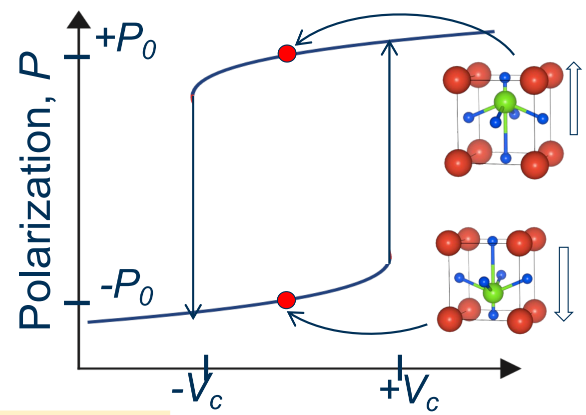
Ferroelectricity

CMOS compatible HfO_2-based fluorite-structure ferroelectric

Orthorhombic *Pca2₁* structure

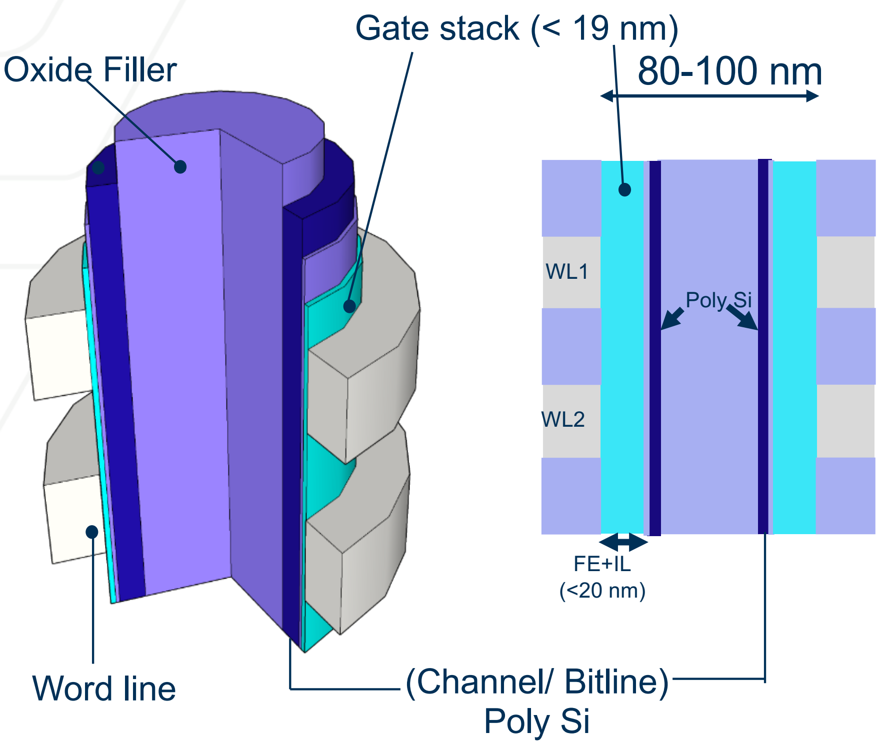


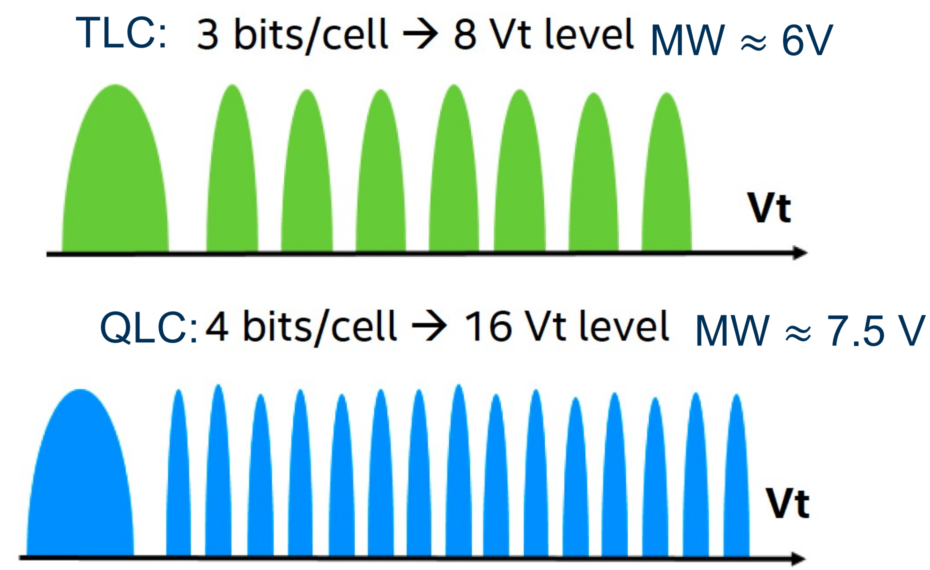




A ferroelectric is an insulator with a spontaneous polarization, which can be switched by an above coercive voltage .

Target metrics for ferroelectric NAND flash





To accommodate the core-shell structure within the hole diameter of 100 nm of the vertical NAND string, the 9 thickness of the gate stack is limited to ~20 nm

Design constraints

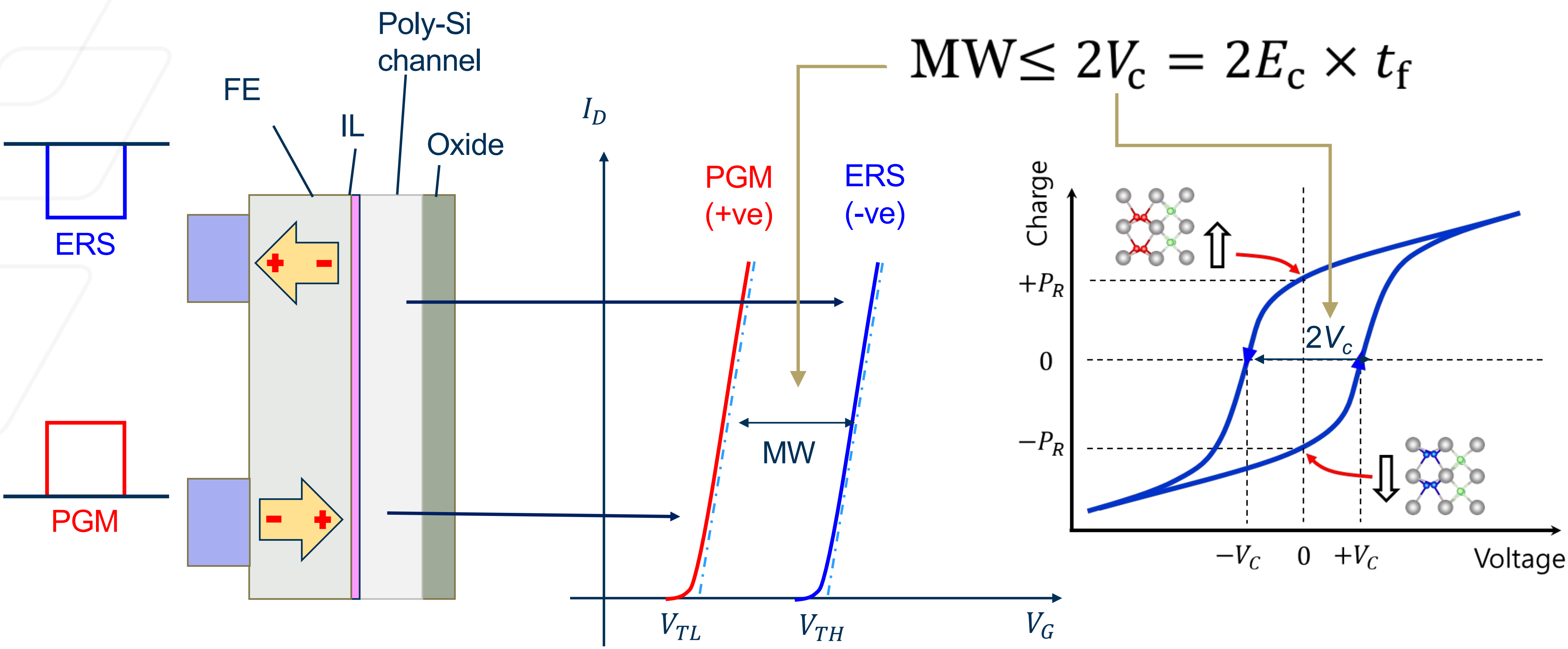
Gate stack thickness, $t \leq 19$ nm

Write voltage ≤ 15 V

Memory window, MW ≥ 6.5 V

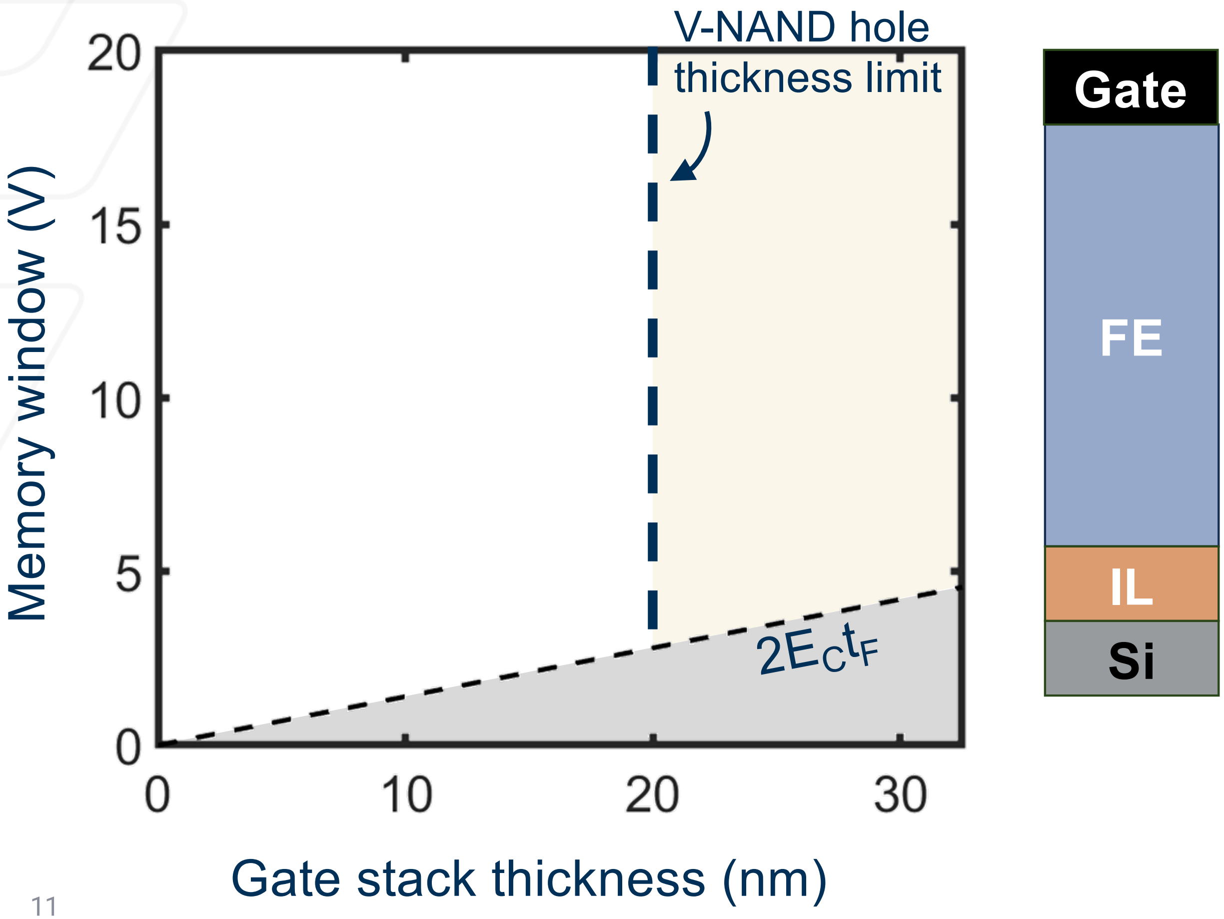
Ferroelectric flash

$$\mathrm{MW} \leq 2V_c = 2E_c \times t_f$$



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

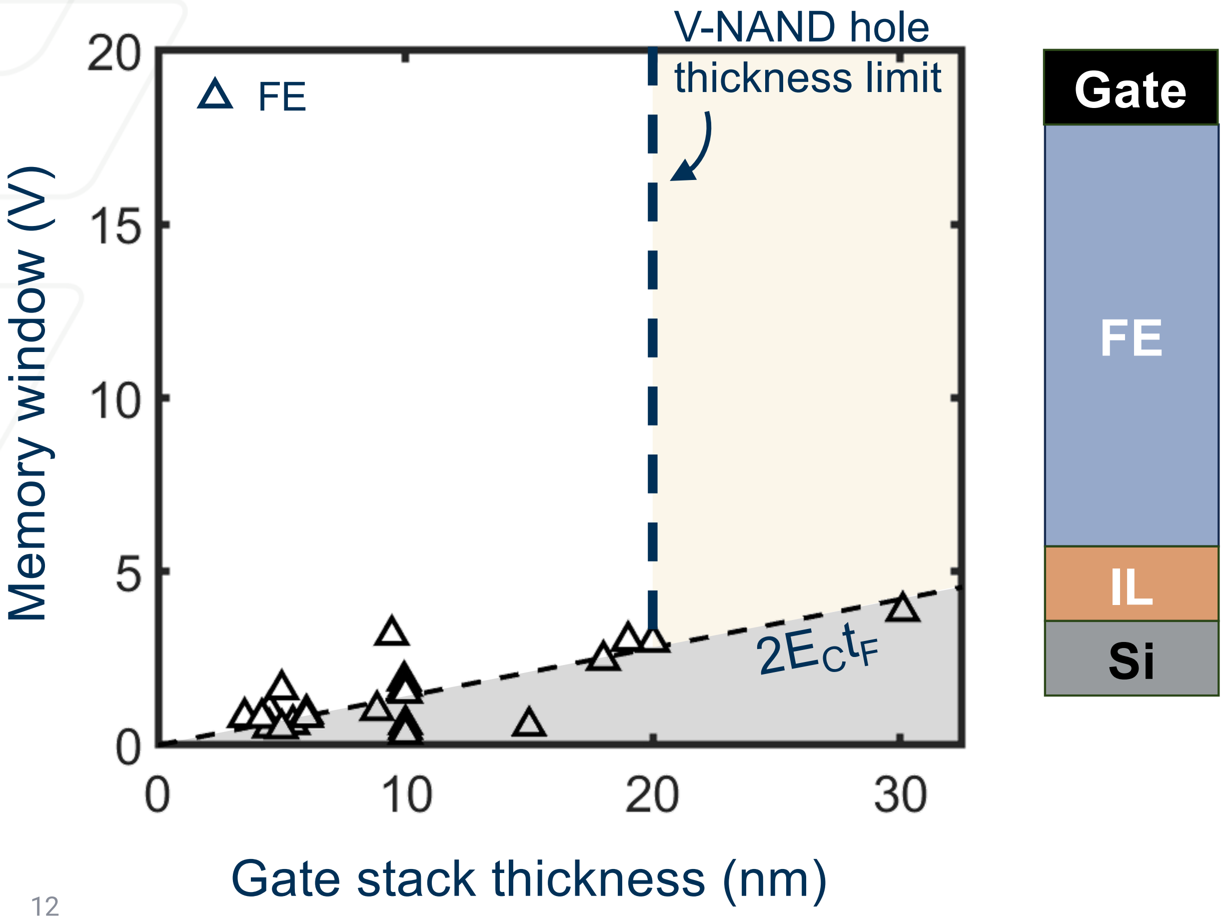
Memory window of FEFETs is limited by V_C



$$\mathrm{MW} \leq 2V_c = 2E_c \times t_f$$

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

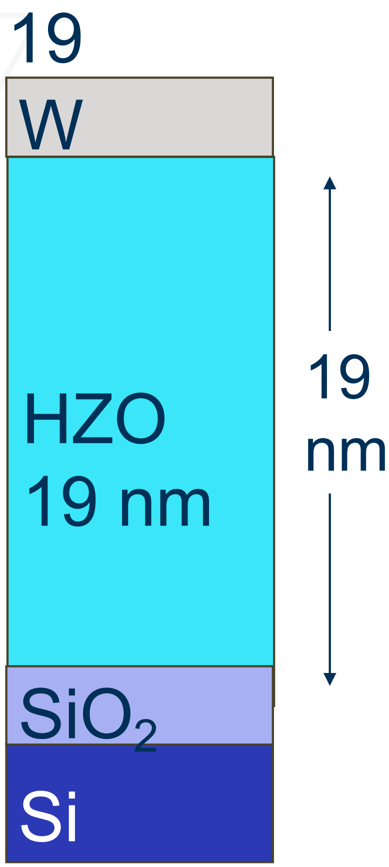
Memory window of FEFETs is limited by V_C



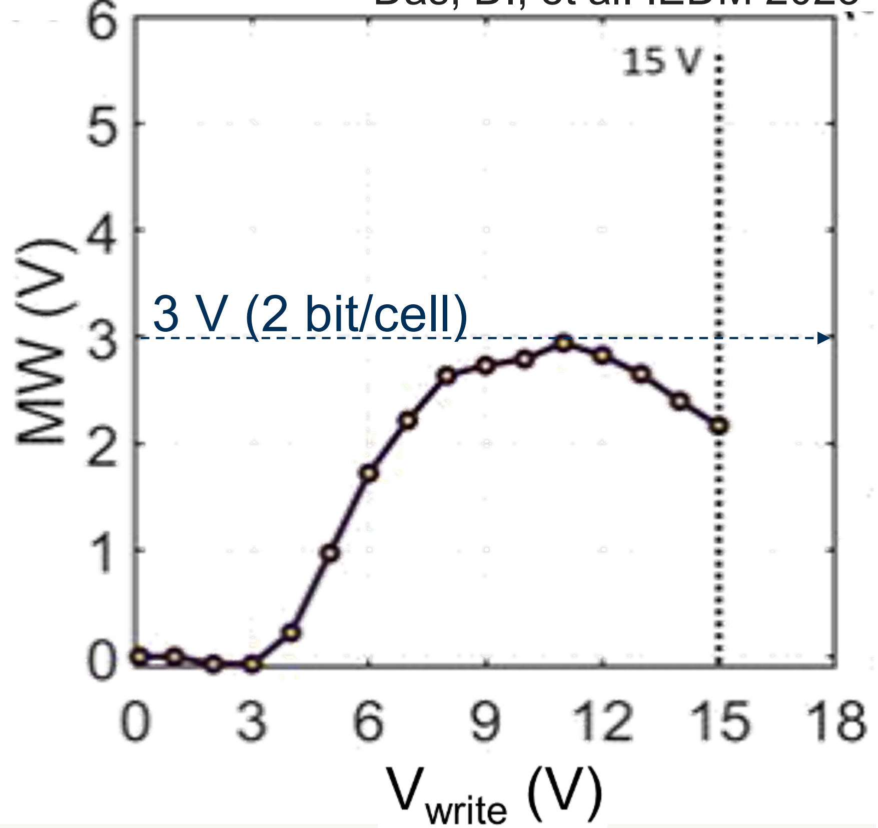
$$MW \leq 2V_c = 2E_c \times t_f$$

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

Increasing the MW by dielectric insertion



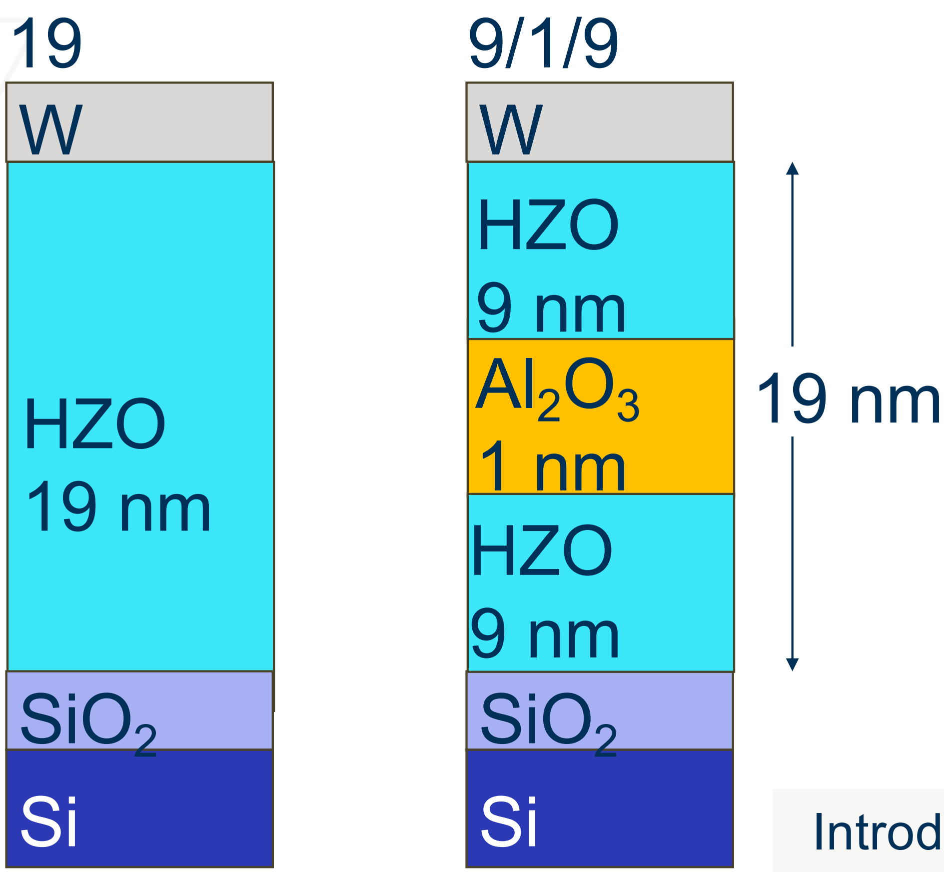
Das, D., et al. IEDM 2023

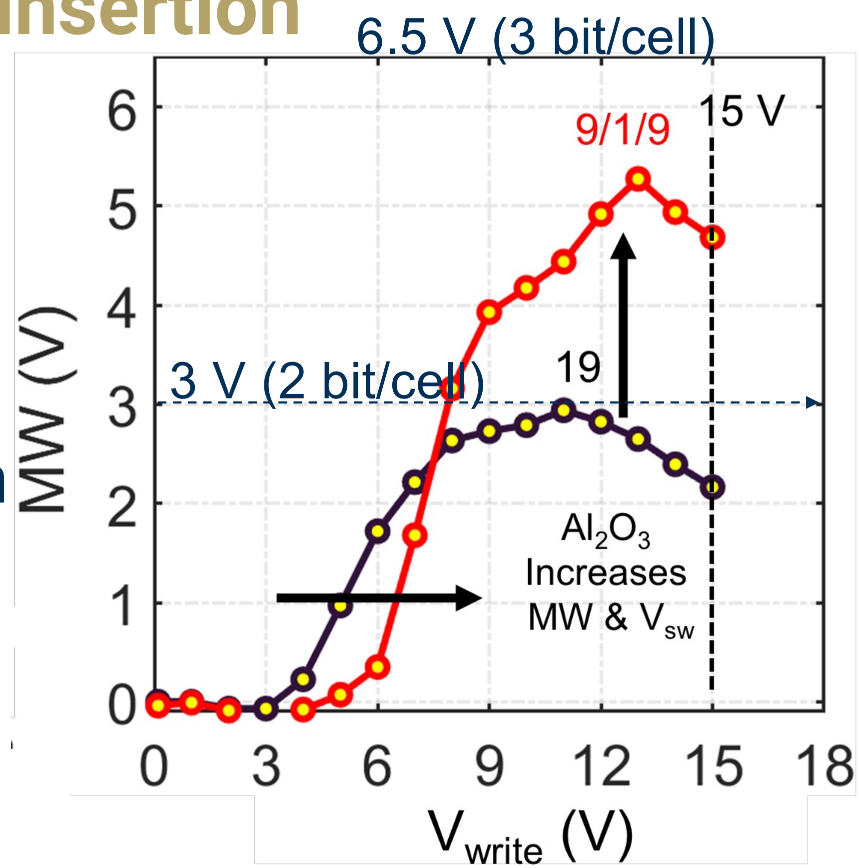


10 nm HZO layer leads to maximum memory window of 2.9 V.

Das, Khan, et al. IEDM 2023 (highlight paper): 10.1109/IEDM45741.2023.10413697

Increasing the MW by dielectric insertion

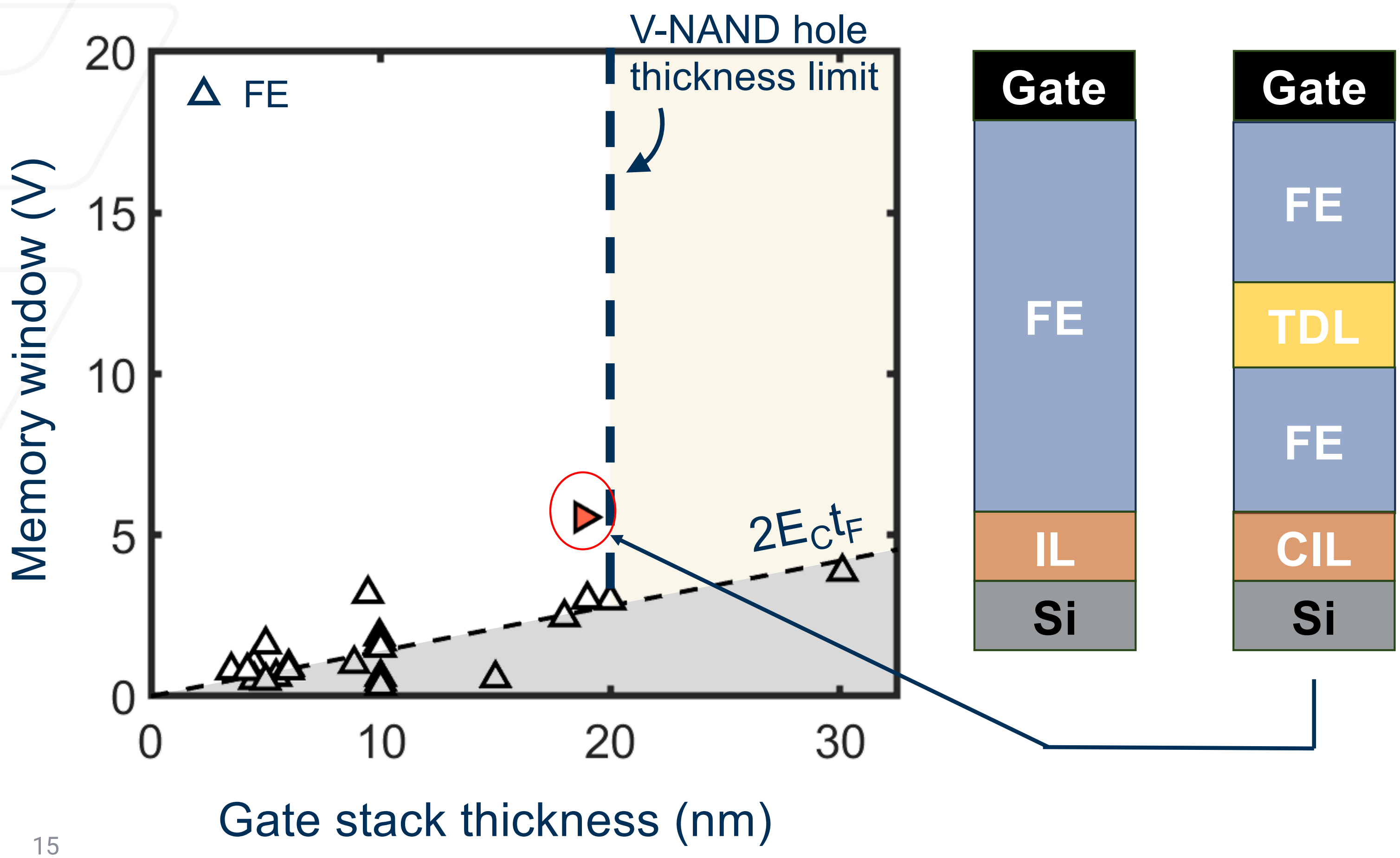




Introducing 1 nm Al_2O_3 intermediate layer dramatically memory window to 5.3 V.

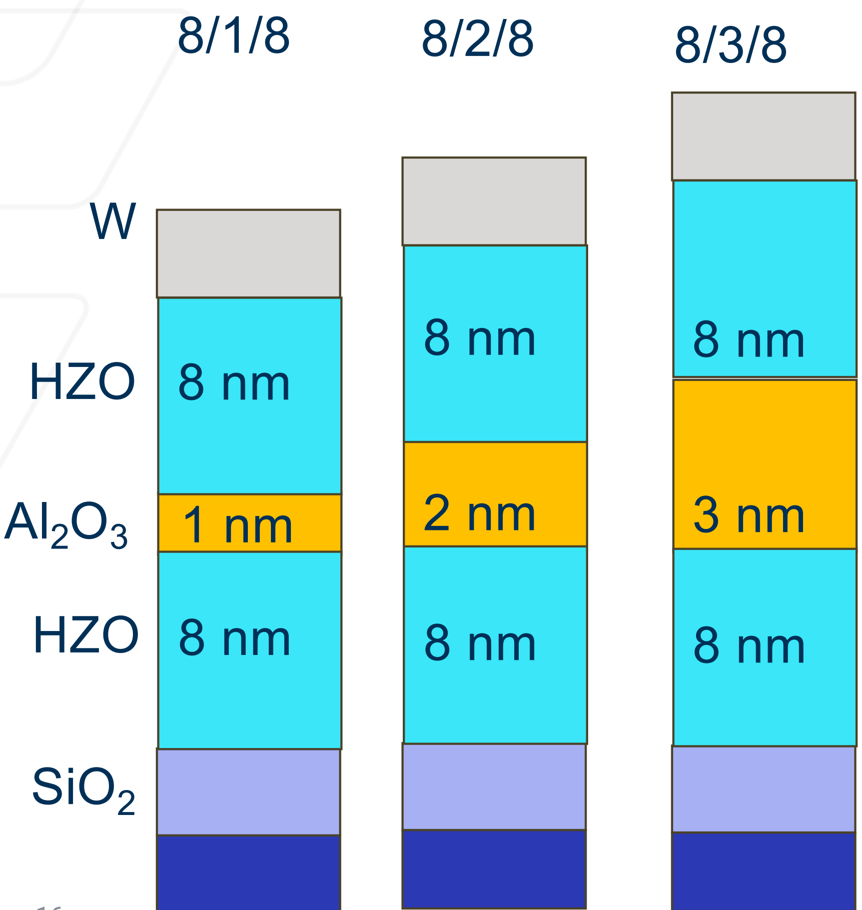
Das, Khan, et al. IEDM 2023 (highlight paper): 10.1109/IEDM45741.2023.10413697

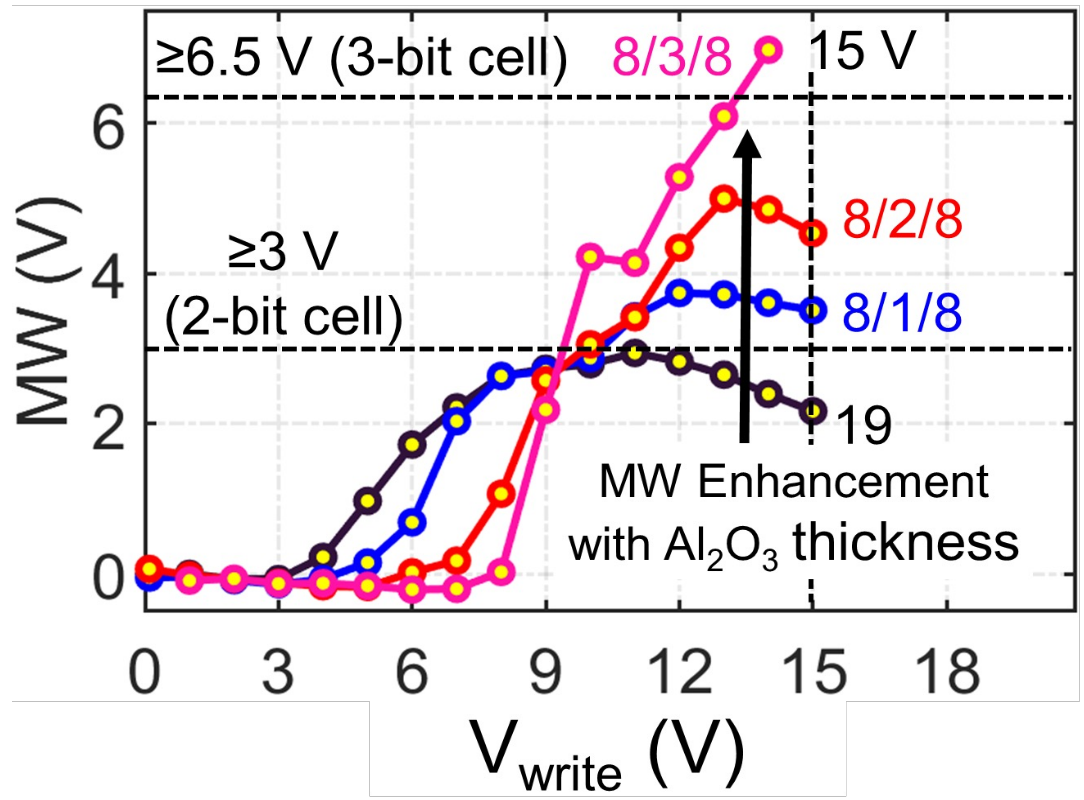
Memory window of FEFETs is limited by V_C



Georgia Tech

Increasing the MW by dielectric insertion

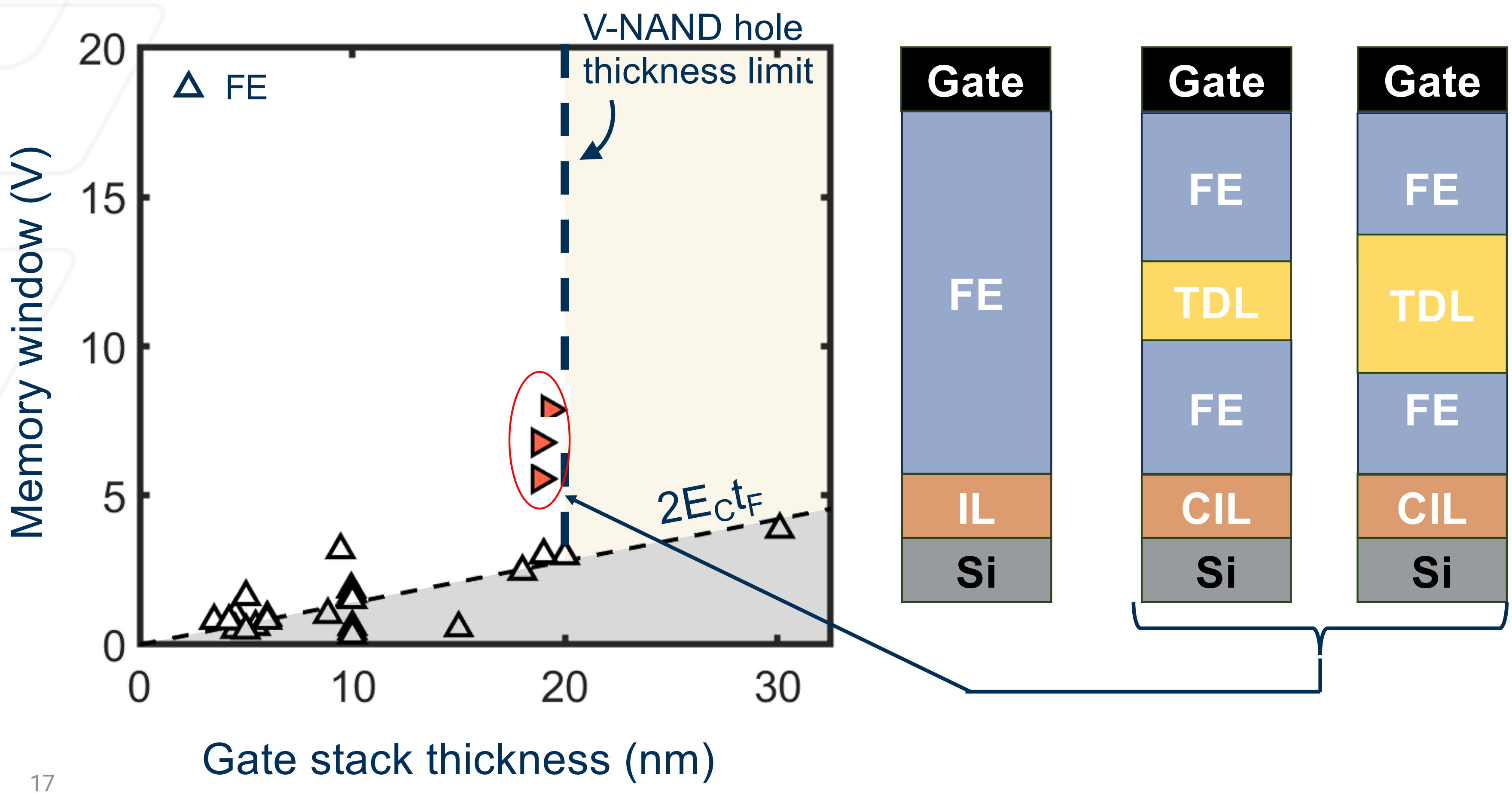




FE 8 nm – AlO 3 nm – FE 8 nm leads to a 7.3 V maximum MW for a 14 V write voltage.

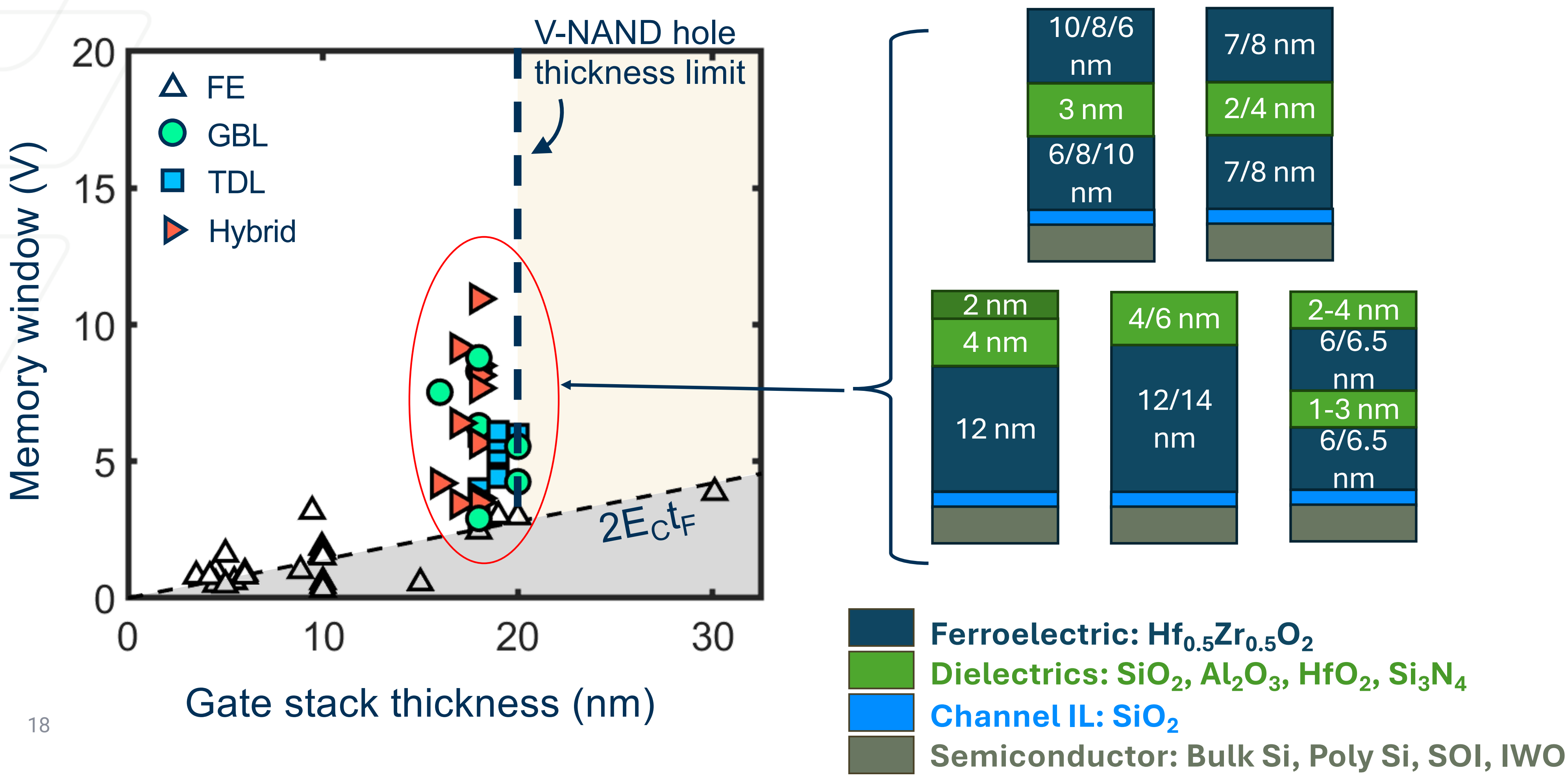
Das, Khan, et al. IEDM 2023 (highlight paper): 10.1109/IEDM45741.2023.10413697

Increasing the MW by dielectric insertion

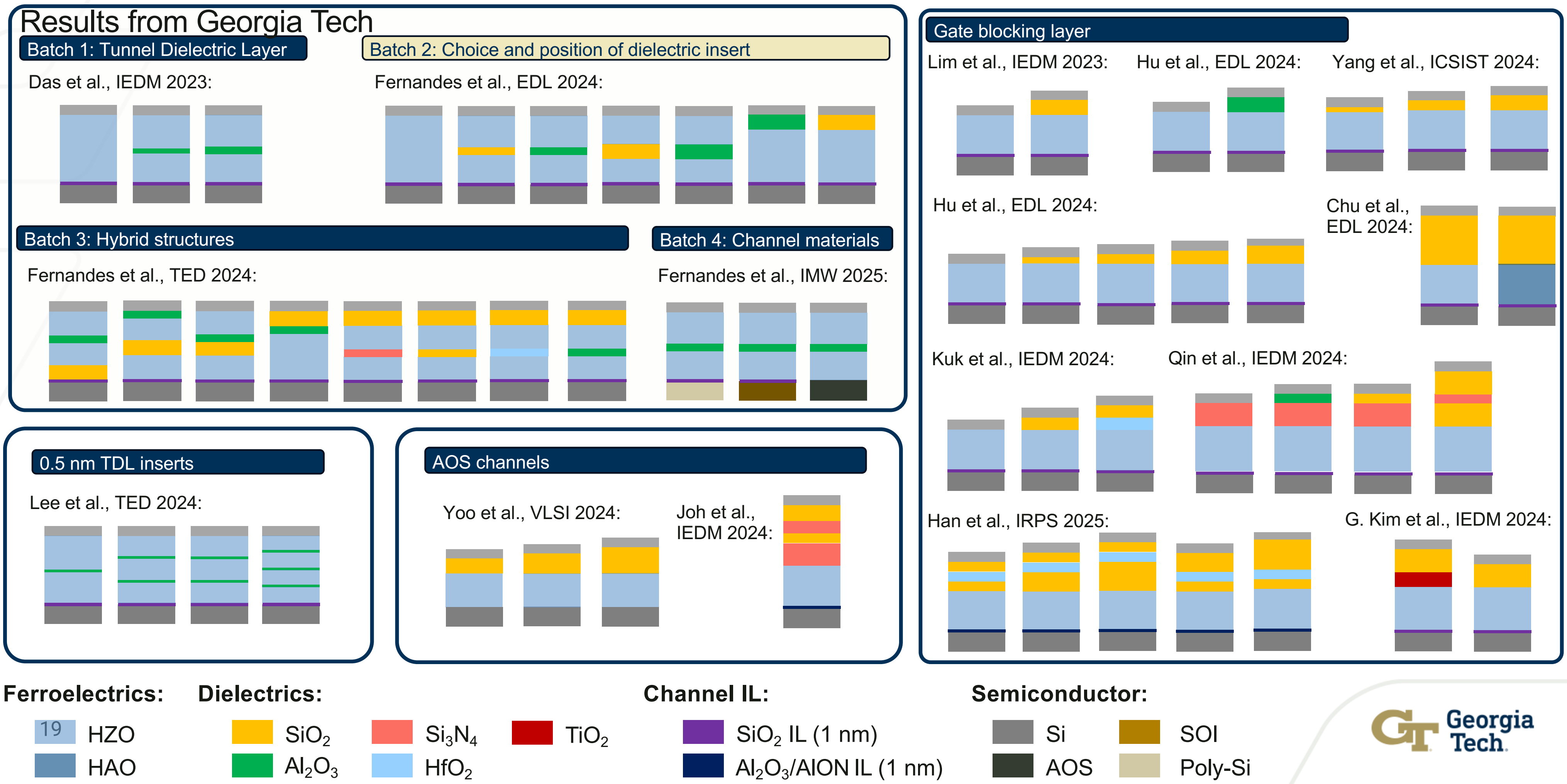


Das, Khan, et al. IEDM 2023 (highlight paper): [10.1109/IEDM45741.2023.10413697](https://doi.org/10.1109/IEDM45741.2023.10413697)

Increasing the MW by dielectric insertion

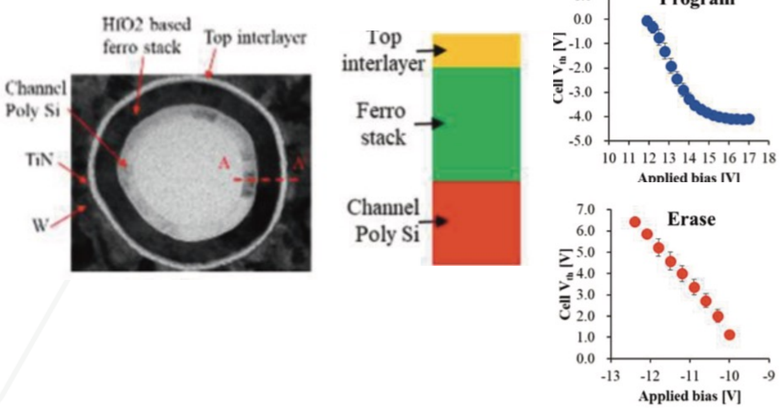


Increasing the MW by dielectric insertion

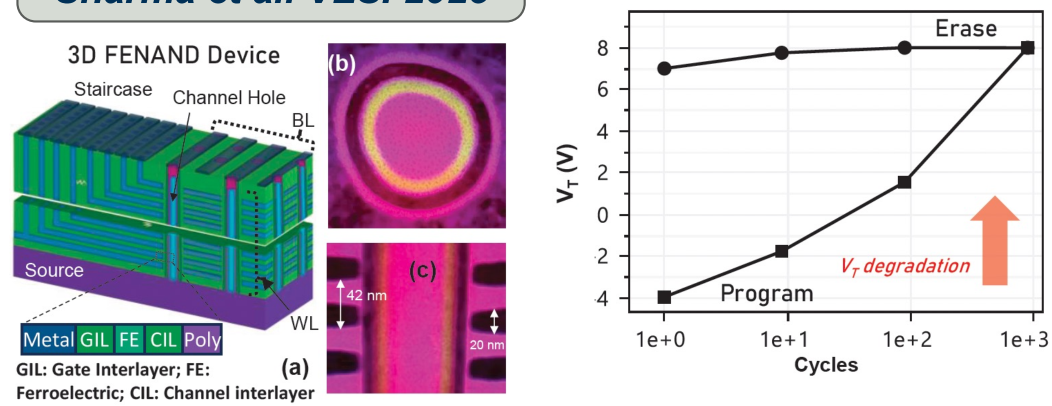


Ferroelectric NAND from the industry

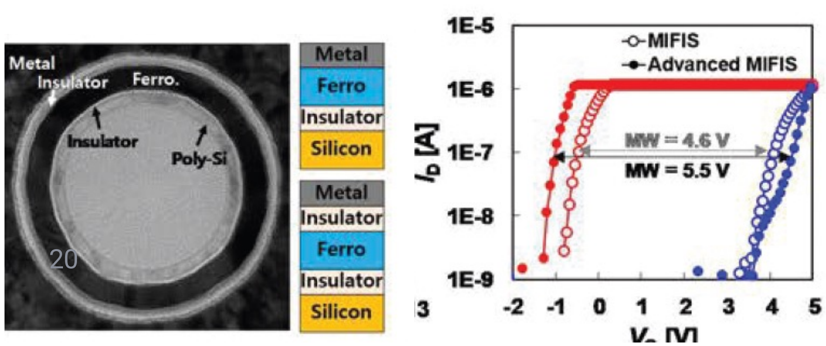
SK Hynix
Yoon et al. VLSI 2023



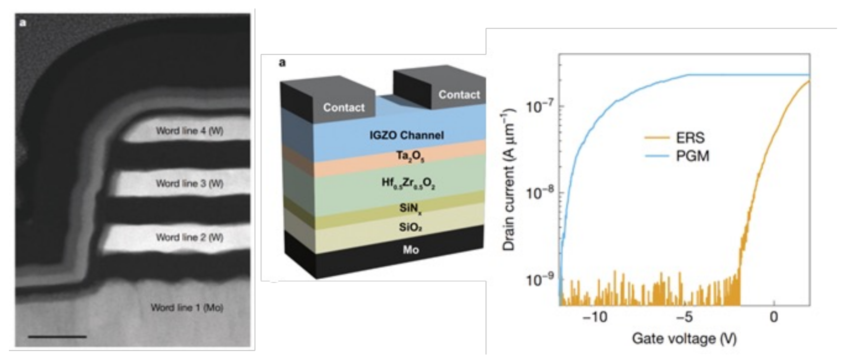
Micron
Sharma et al. VLSI 2025



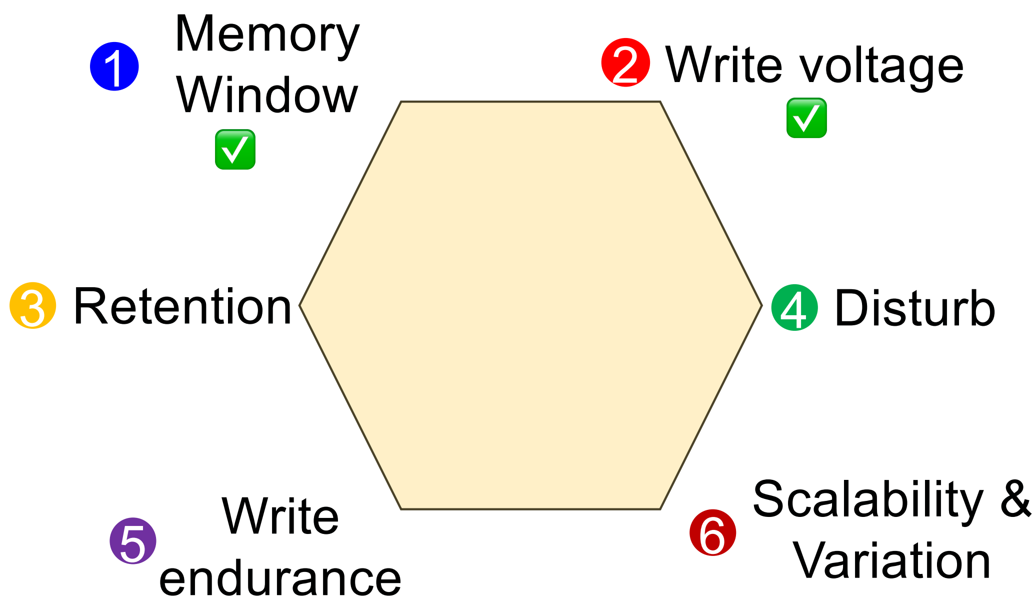
Samsung
Lim et al. IEDM 2023



Samsung
Sharma et al. Science 2025

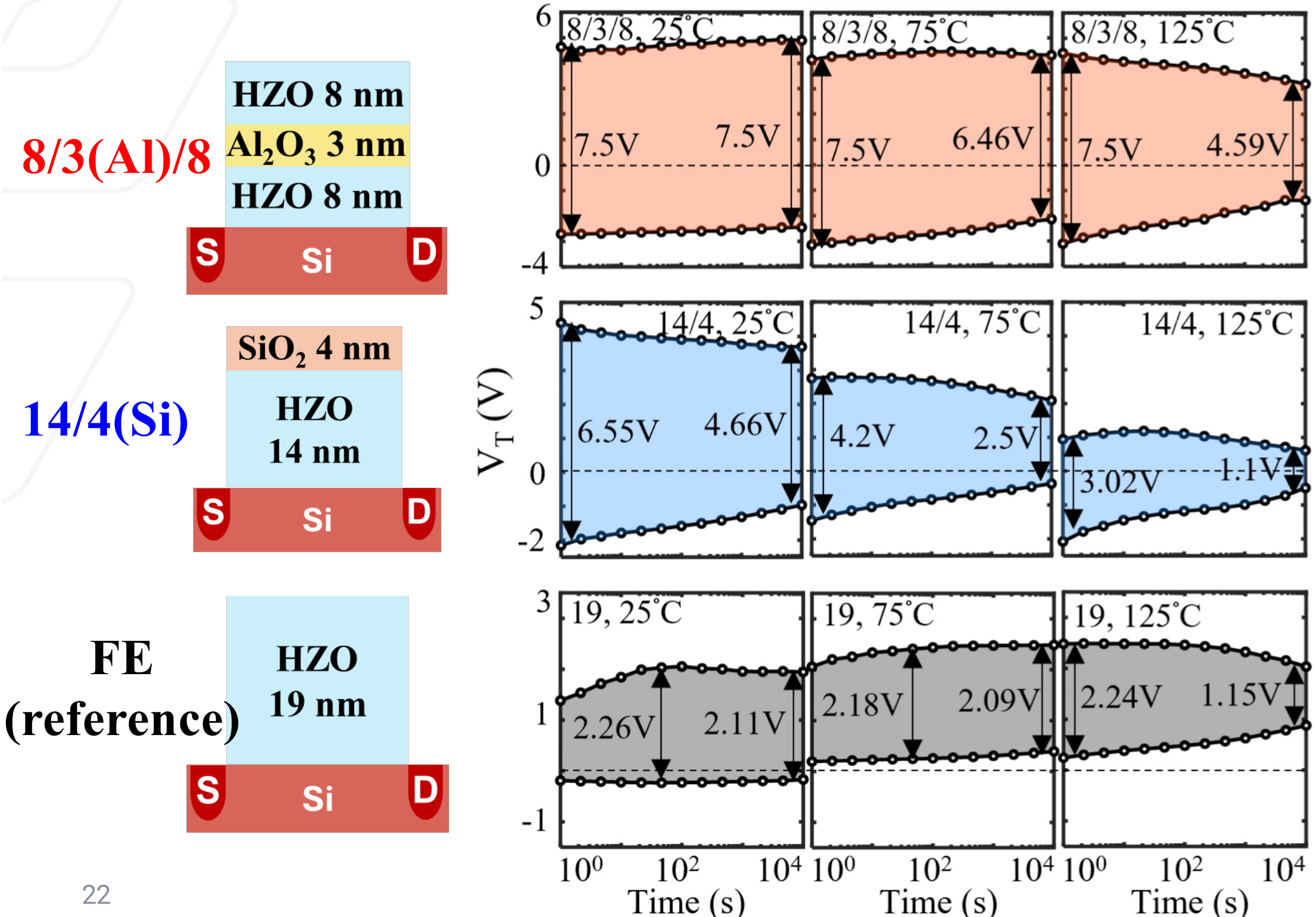


Evaluating ferroelectric NAND Flash



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

Retention



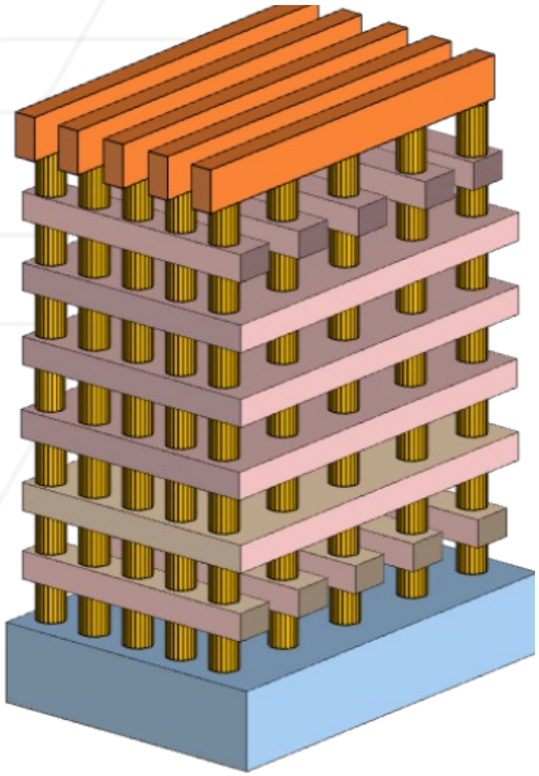
GBL stack has lower initial MW at high temperature, but MW loss remain the same

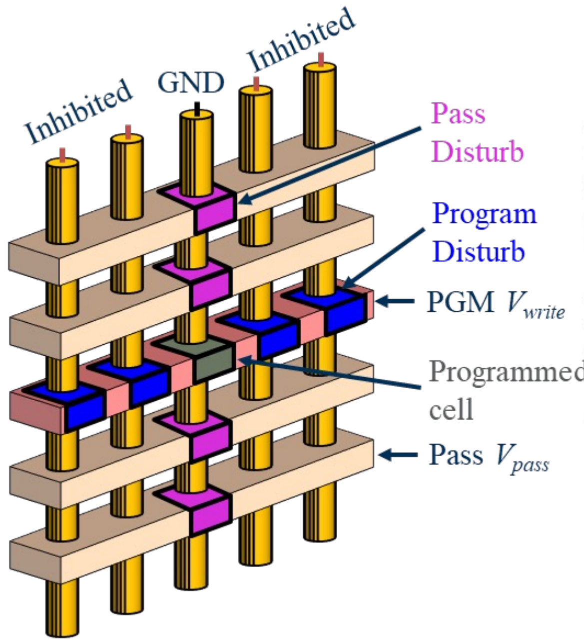
Reference stack shows slight degradation at Higher temperature

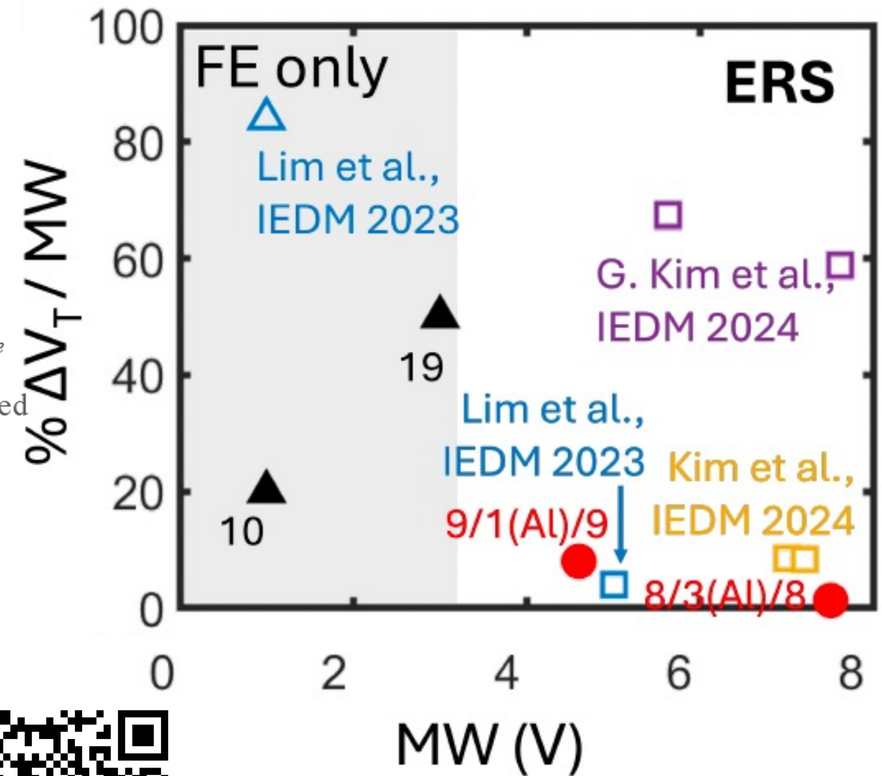
Venkatesan, Khan, et al. IRPS 2025 (Best Student Paper): https://doi.org/10.1109/IRPS48204.2025.10982749
Venkatesan, Khan, et al. EDL 2025 (Editor's Pick): 10.1109/LED.2024.3523235

YouTube Link: https://youtu.be/wYLvvmKorGQ?si=wLm34Tboy9SmMhXQ

Disturb

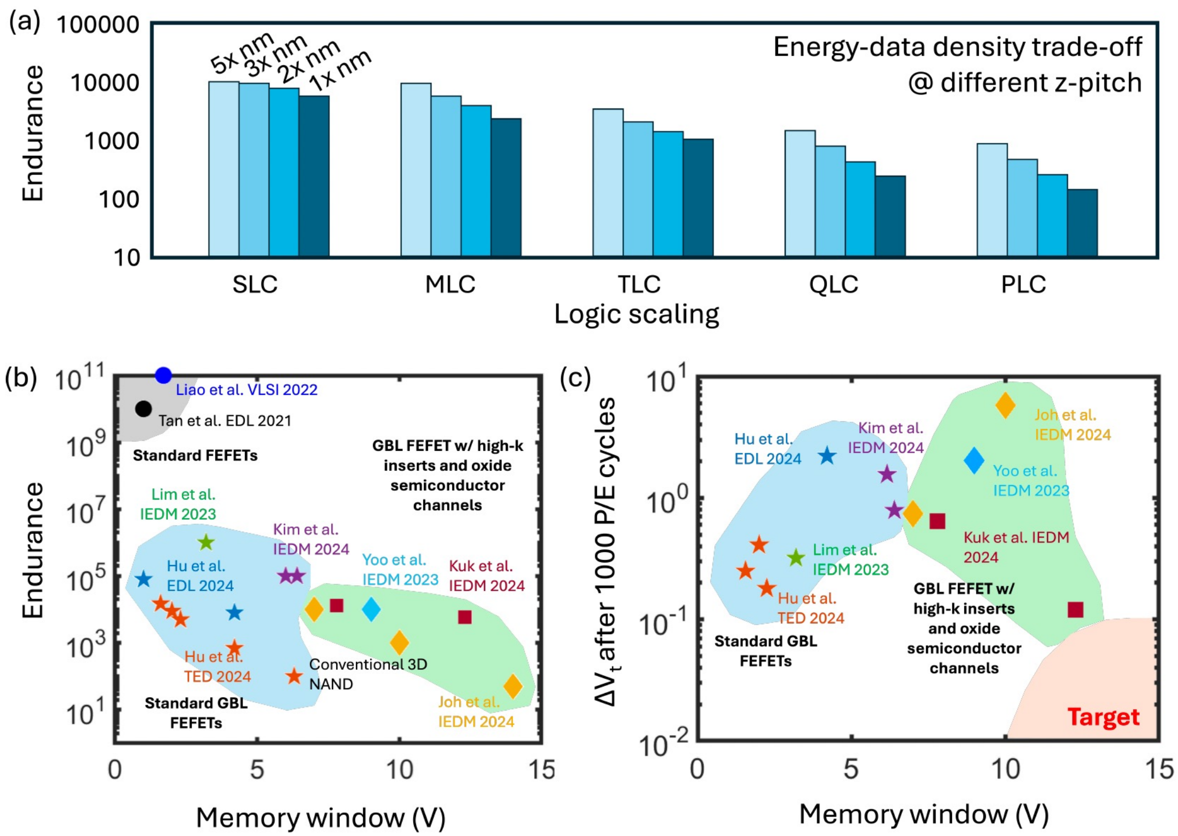






P. Venkatesan *et al.*, "Disturb and its Mitigation in Ferroelectric Field-Effect Transistors With Large Memory Window for NAND Flash Applications," in *IEEE Electron Device Letters*, vol. 45, no. 12, pp. 2367-2370, Dec. 2024, doi: 10.1109/LED.2024.3467210

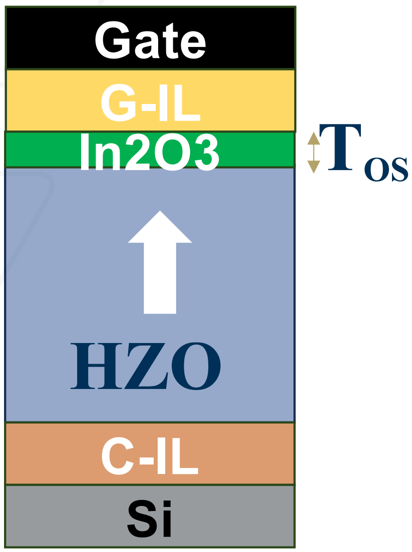
Write Endurance

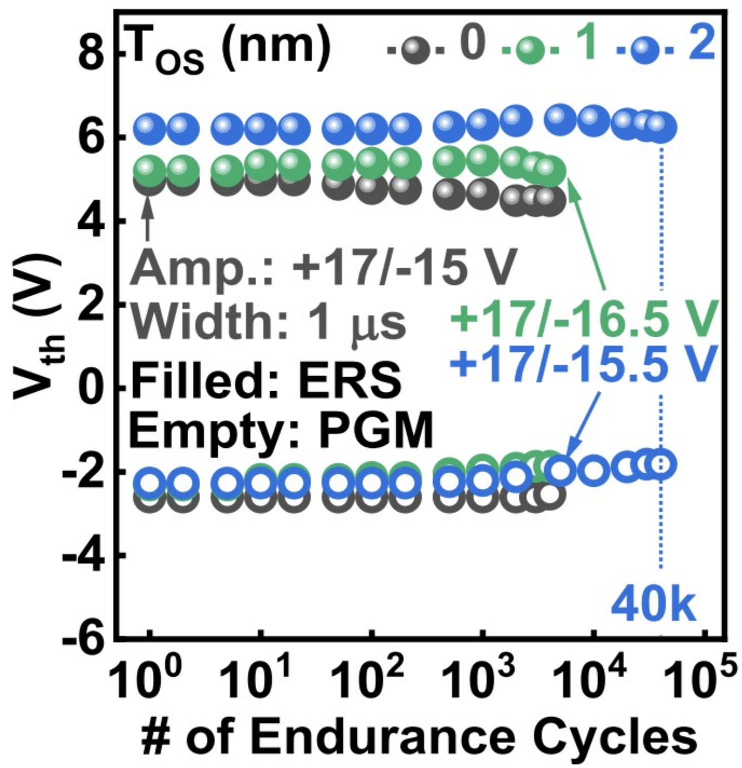


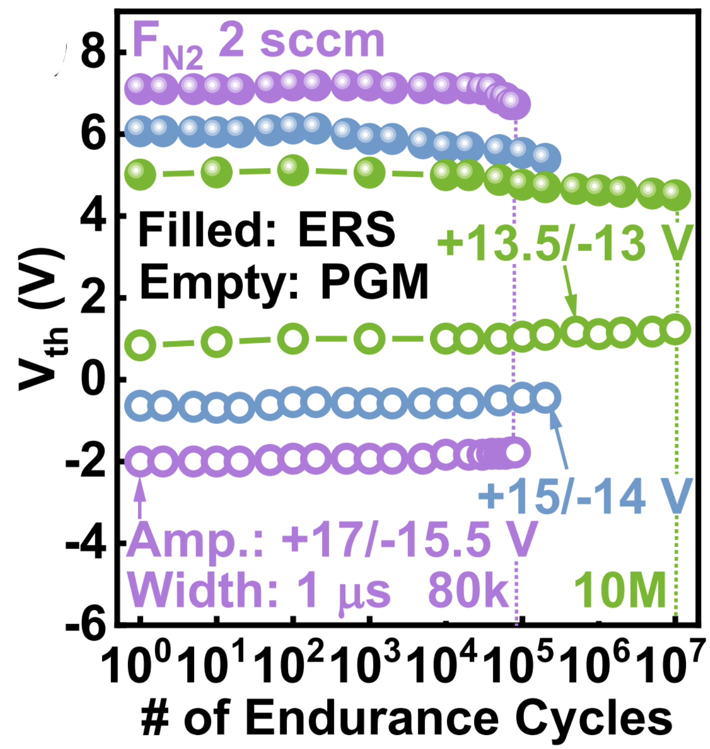
Venkatesan, Khan, et al. "Pushing the limits of NAND technology scaling with ferroelectrics: P. Venkatesan et al." *MRS Bulletin* 50.9 (2025): 1094-1107.

Georgia Tech

Insertion of Oxygen Reservoir Layer to improve Fe-NAND endurance



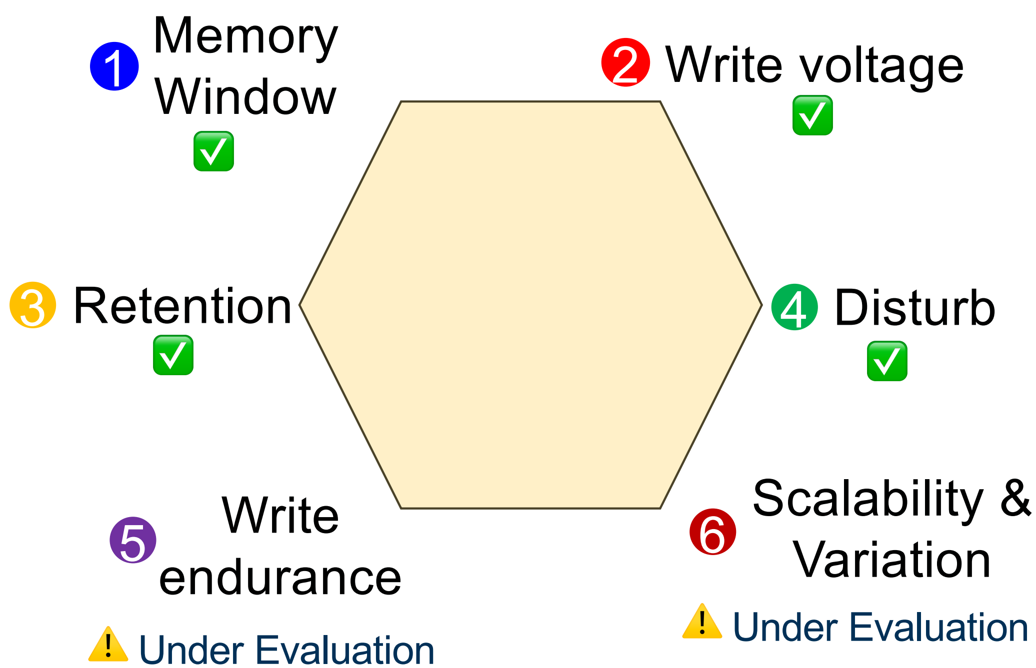




Introducing In2O3 layer acts as oxygen reservoir layer for HZO layer to enable larger MW and high endurance by optimizing the thickness and deposition parameters of the ORL layer

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.

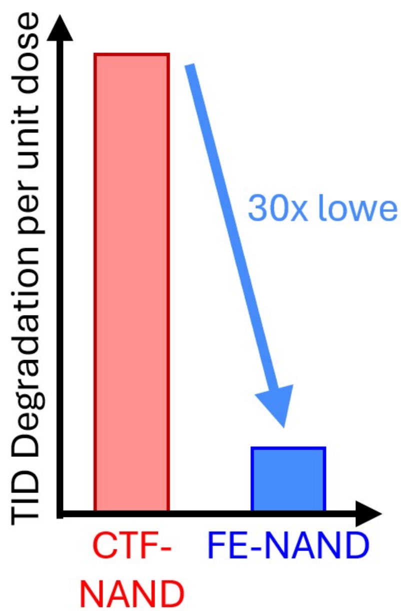
Space applications and ferroelectric NAND flash

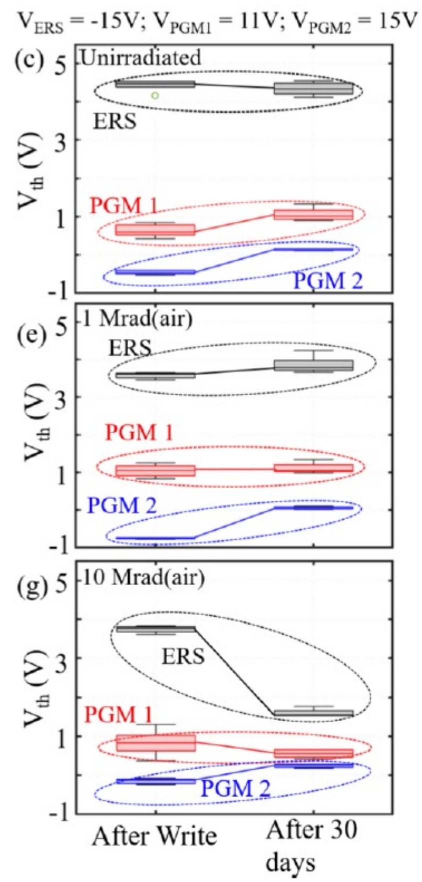
- Space is becoming a key infrastructure layer for **AI, sensing, communication, navigation, and defense**.
- **Spacecraft memories must retain data under extreme radiation.**
  - Radiation dose increases sharply with mission orbit:
    LEO satellites: **5–30 krad**
    GEO missions: **100–300 krad**
    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.

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.

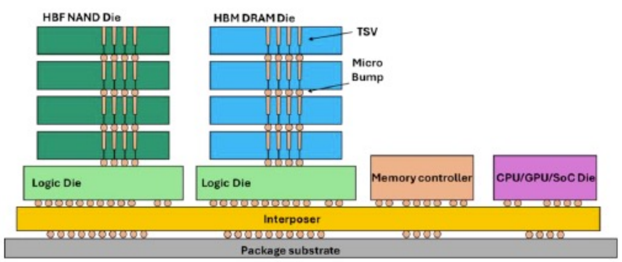
Future work: SEE, displacement damage, retention/endurance after irradiation, disturb behavior, and thermal recovery

Georgia Tech

Where NAND is Going Next

1

ACTIVE MEMORY TIER



Workload-tuned NAND

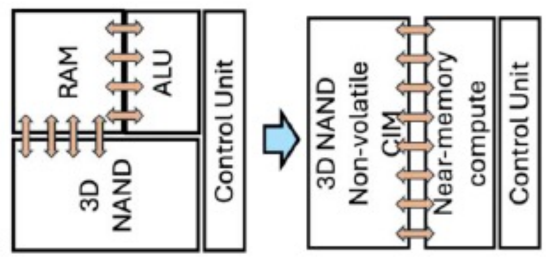
- QLC for cold archival
- SLC-like for inference buffers for AI

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

Georgia Tech

Our team


Prasanna Venkatesan


Lance Fernandes


Salma Soliman


Taeyoung Song


Zekai Wang


Dr. Hang Chen


Yu Hsin Kuo


Nashrah Afroze


Priyankka Ravikumar


Jiayi Chen


Dr. Mengkun Tian


Dr. Chinsung Park (SK Hynix)


Dr. Dipjyoti Das (NIT Silchar)


Dr. Zheng Wang (Micron)


Dr. Sarah Lombardo (TEL)


Dr. Nujhat Tasneem (Intel)


Anthony Gaskell (Micron)

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

SRC® Semiconductor Research Corporation

SAMSUNG ELECTRONICS

3D FeM²

U.S. DEPARTMENT of ENERGY

Georgia Tech