From Lab to Fab– Emerging Materials for the AI-Era Memory Hierarchy

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

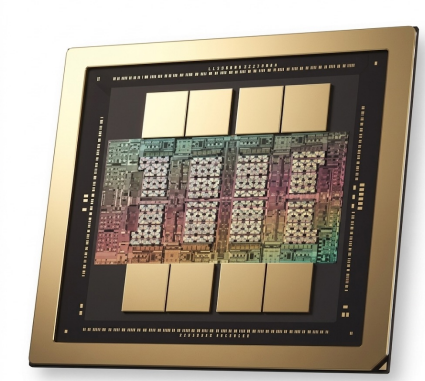
IEEE International Conference on IC Design and Technology (ICICDT 2026)
Dresden, Germany | 23 June 2026

Georgia Tech

Foundational Memory & Storage Technologies

1

DRAM and
High Bandwidth Memory



NVIDIA Rubin

CAPACITY

288 GB

BANDWIDTH

22 TB/s

2

NAND Flash
Storage

CAPACITY

30 – 245 TB

BANDWIDTH

~14 GB/s

← faster, smaller | larger, slower →

Every AI system rides on this hierarchy.

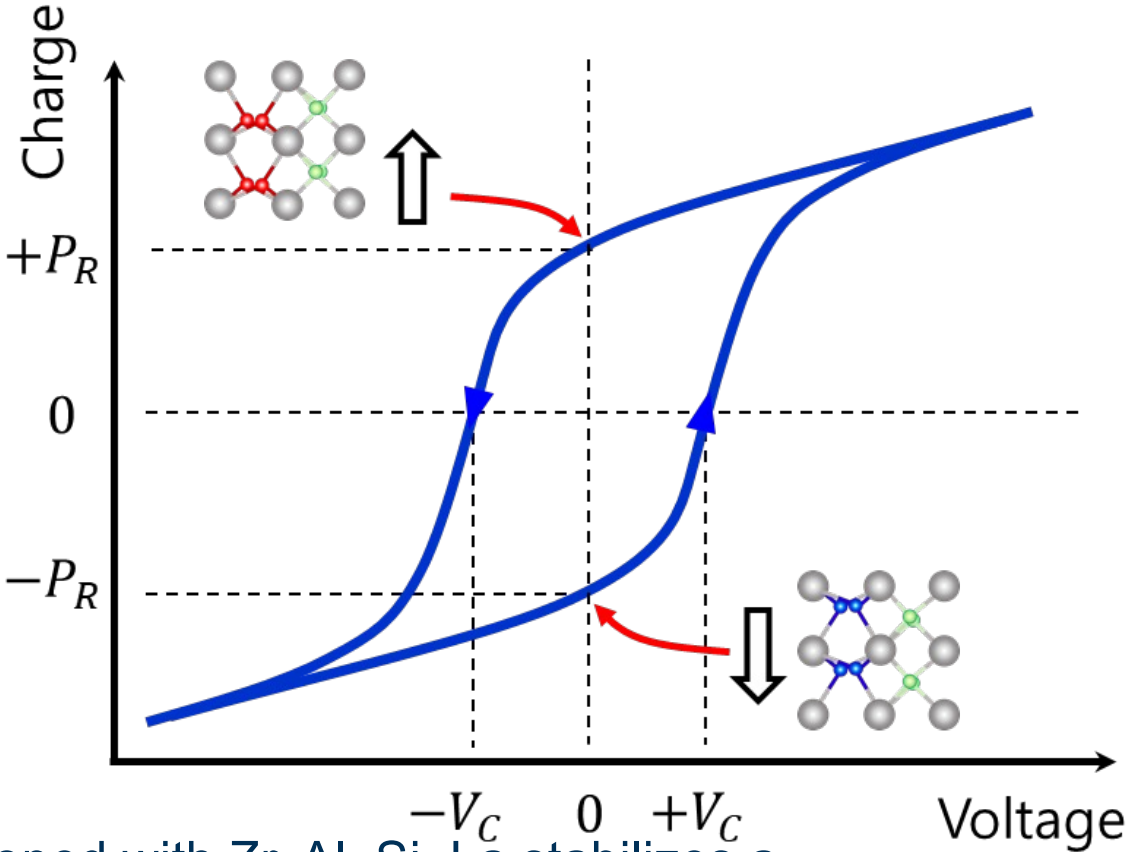
Georgia Tech

Ferroelectricity in already fab-native HfO_2 and alloys

Same **ALD HfO_2** that's been the high-k gate dielectric since the 45 nm HKMG node (2007) Already the high-k in **DRAM capacitors** — a qualified, high-volume material in every logic & memory fab

Embedded **FeRAM / FeFET** in foundry — **GlobalFoundries 22FDX**

Micron 32 Gb 3D FeRAM demonstrated



Controlled annealing of ALD HfO_2 doped with Zr, Al, Si, La stabilizes a ferroelectric orthorhombic phase — **a fab-native material with no new precursor supply chain.**

Ferroelectricity: a strategic-materials platform, not a device

1

ONE MATERIAL SYSTEM

HfO2

CMOS-COMPATIBLE

- Fluorite-structure HfO2 / HZO
- ALD — already in fabs
- No exotic elements

The shared platform

2

PERSISTENT MEMORY

Fe-RAM

DRAM-CLASS, NON-VOLATILE

- Ferroelectric capacitor
- DDR & CXL persistent tier
- Zero refresh, low idle power

Vector A

3

SCALABLE STORAGE

Fe-NAND

BEYOND 1000 LAYERS

- FeFET-based 3D NAND
- Fits existing integration
- Low-voltage, field-driven

Vector B

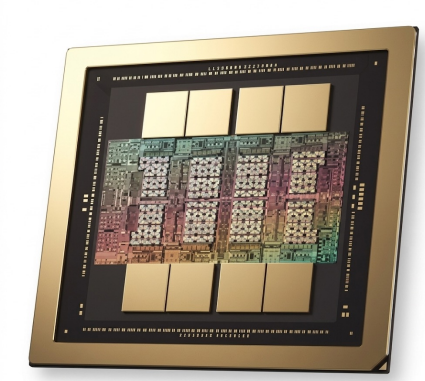
One HfO_2 platform across the hierarchy
HBM · DDR · CXL · NAND

Georgia Tech

Foundational Memory & Storage Technologies

1

DRAM and
High Bandwidth Memory



NVIDIA Rubin

CAPACITY

288 GB

BANDWIDTH

22 TB/s

2

NAND Flash
Storage

CAPACITY

30 – 245 TB

BANDWIDTH

~14 GB/s

← faster, smaller | larger, slower →

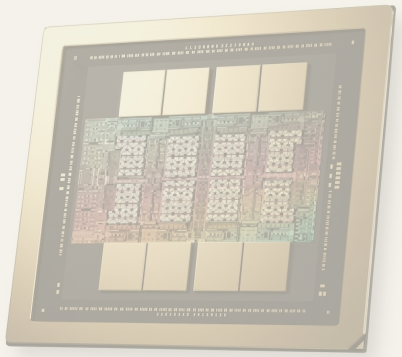
Every AI system rides on this hierarchy.

Georgia Tech

Foundational Memory & Storage Technologies

1

DRAM and
High Bandwidth Memory



NVIDIA Rubin

CAPACITY

288 GB

BANDWIDTH

22 TB/s

2

NAND Flash
Storage

CAPACITY

30 – 245 TB

BANDWIDTH

~14 GB/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

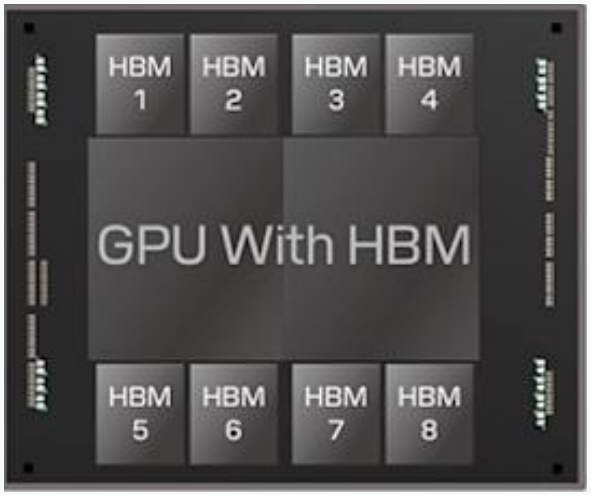
NAND is no longer where the data sleeps.
It's increasingly part of where the AI thinks.

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



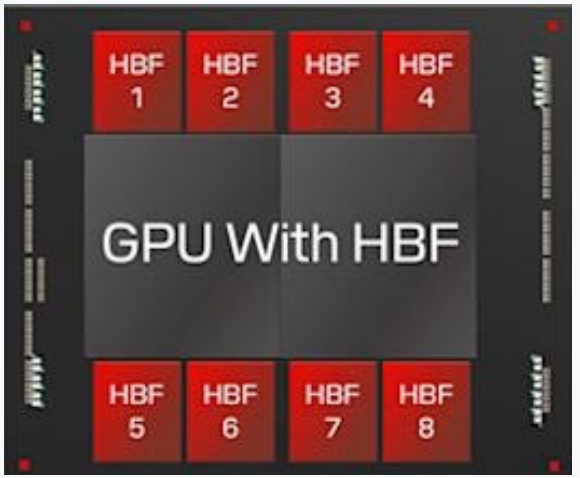
192 GB

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

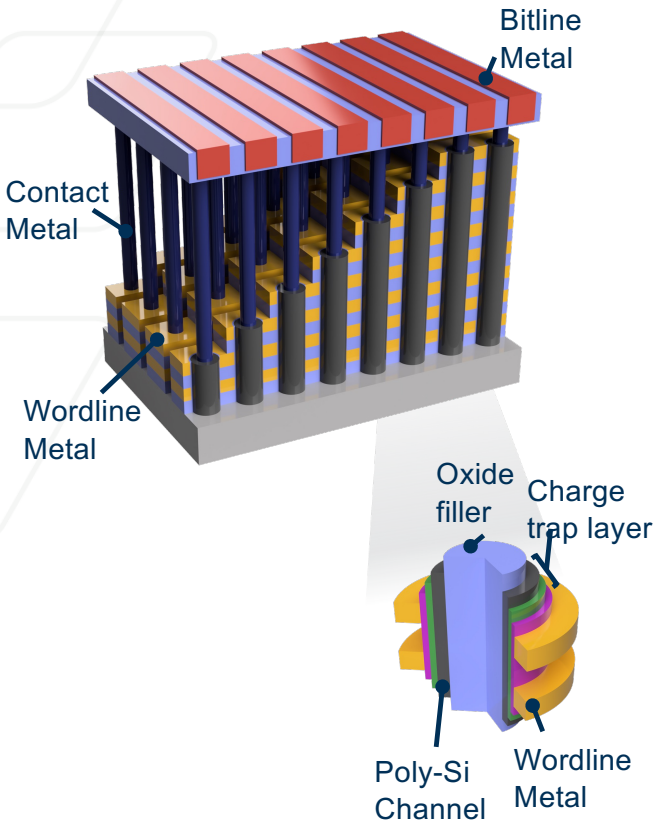
Concept and figure inspired by SanDisk High Bandwidth Flash (HBF). KV caches, mixture-of-experts weights, long-context attention, large embedding tables

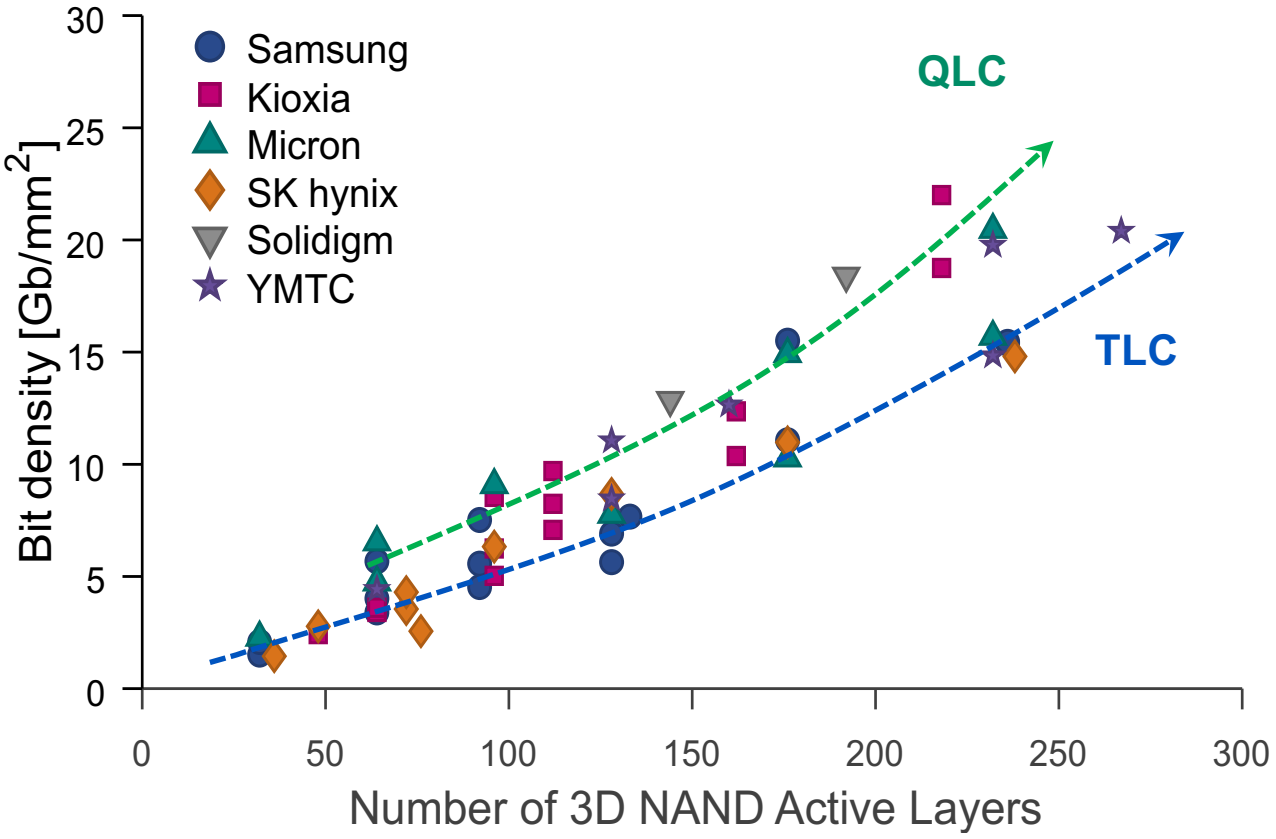
21× more memory, on package, HBM-class bandwidth, NAND NVMe-like capacity

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



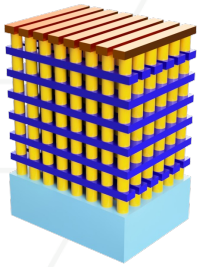


Venkatesan, Song, Thareja et al. (under review in Nature)

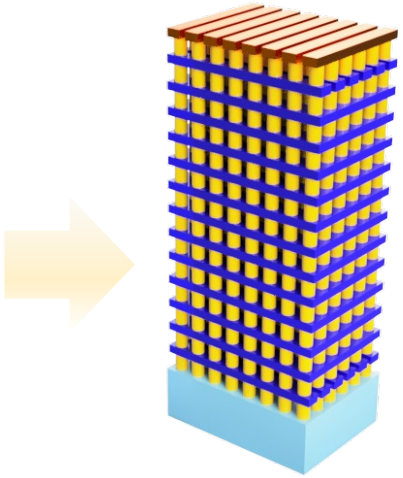
A 50x improvement in bit arial density per decade.

Georgia Tech

Vertical NAND Scaling

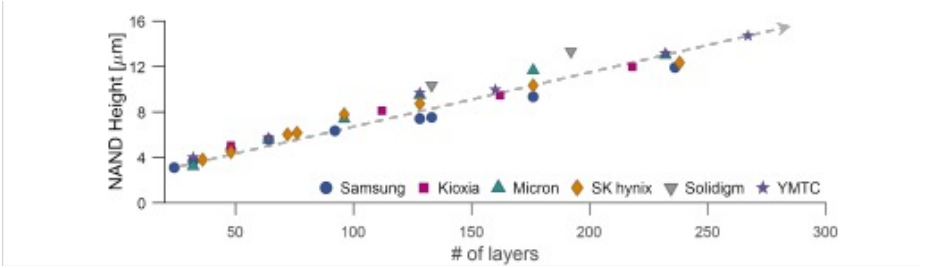


2014
First 3D NAND, Samsung
24 Layers, MLC

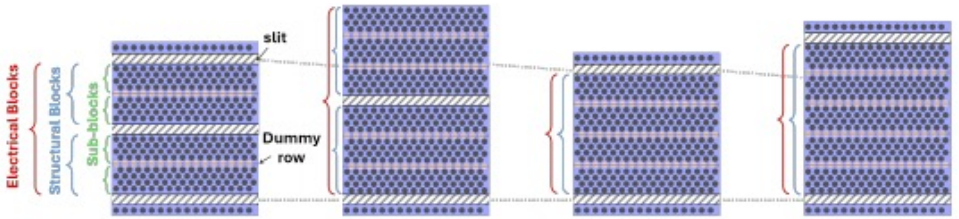


2022-2026
Production
Samsung 236L TLC
SK hynix 321L (ramping 1H25)
Micron 276L TLC
Prototype / R&D demos:
SK hynix 300+L demo (ISSCC 2023)
Kioxia / WD >300L demo (VLSI 2023)
Kioxia / Sandisk 1000L using Multi-Stacked Cell Array – CMOS Bonded to Array (2026)

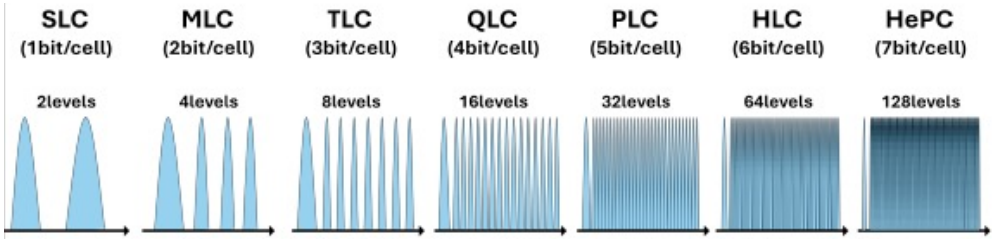
Layer scaling



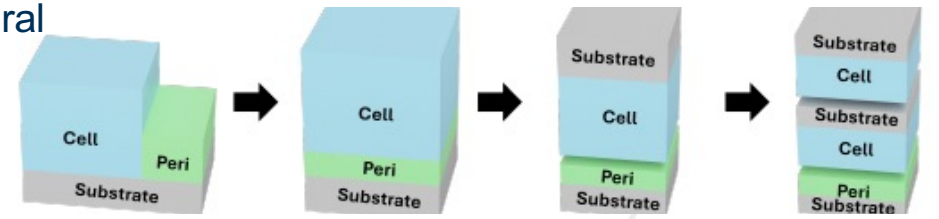
Lateral scaling



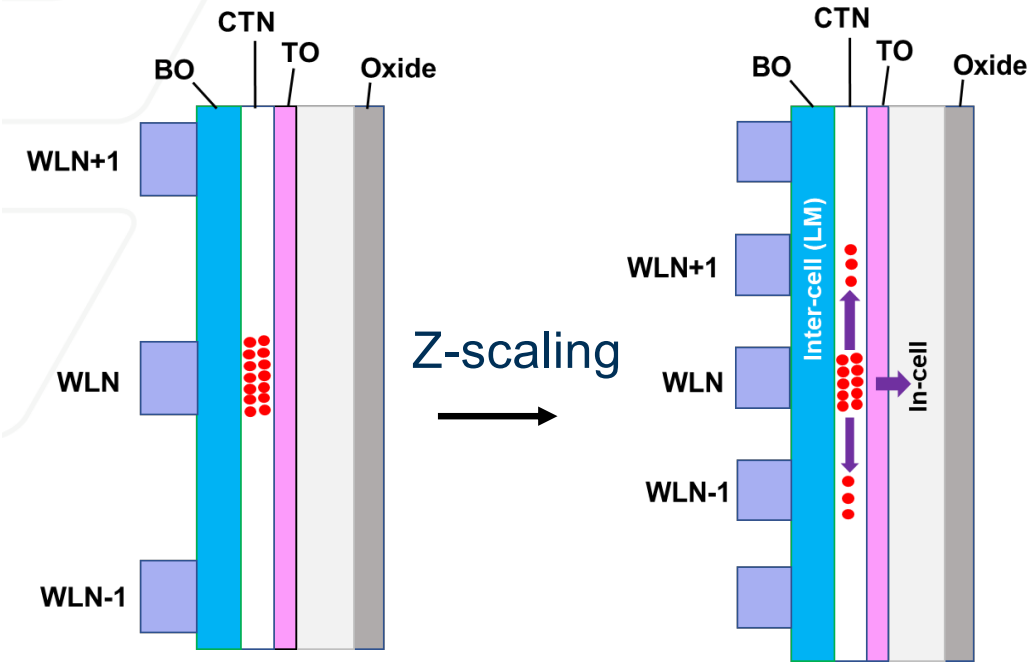
Logical Scaling



Architectural 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

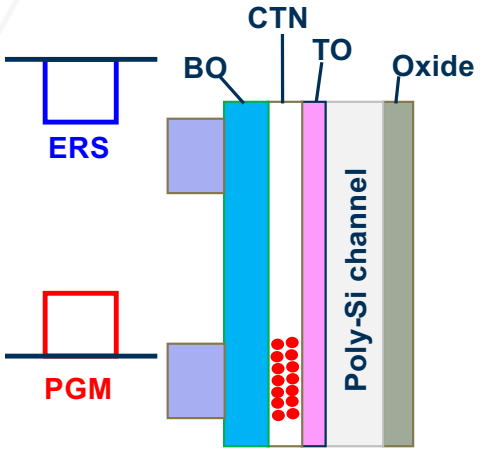
RTN, poly-Si grain variability, and channel taper compress threshold-voltage state distributions. Poly-Si channel resistance grows with cell count; grain boundaries limit mobility and uniformity.

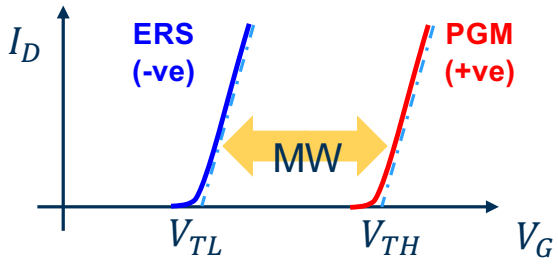
Read sensing margin shrinks

11

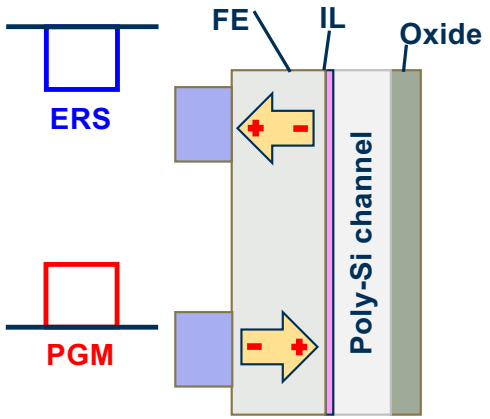
Charge trap flash vs. ferroelectric flash

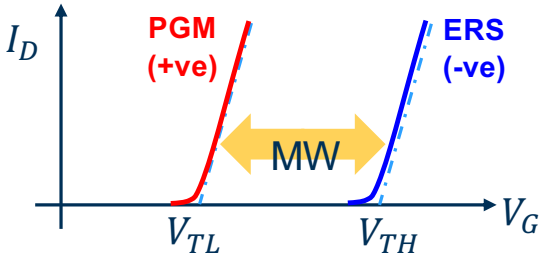
Conventional charge trap flash





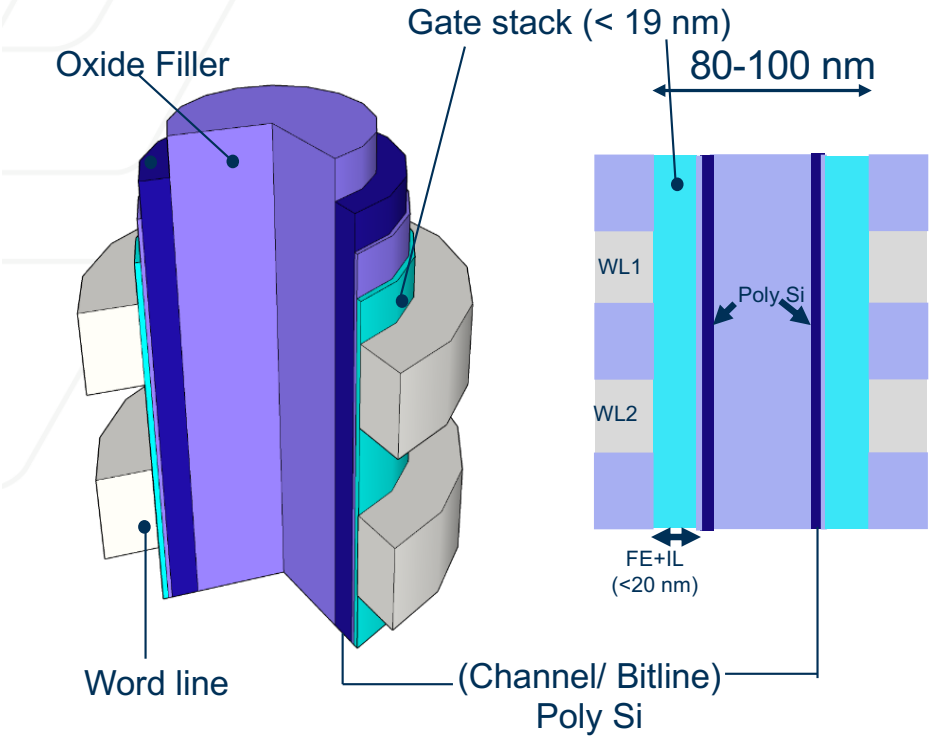
Ferroelectric NAND flash



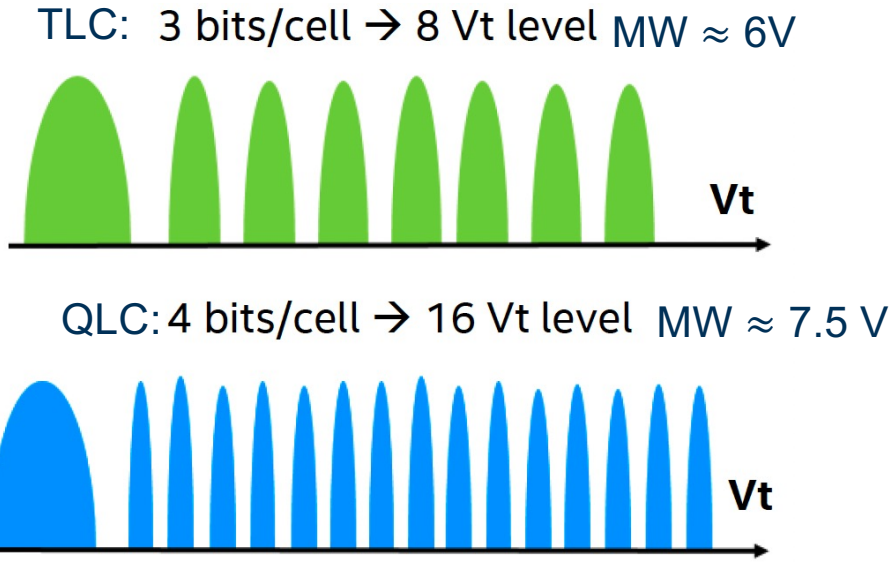


Georgia Tech

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 thickness of the gate stack is limited to ~20 nm



Design constraints

Gate stack thickness, $t \leq 19$ nm

Write voltage ≤ 15 V

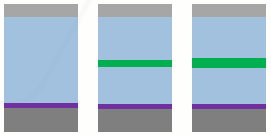
Memory window, MW ≥ 6.5 V

Recent progress ferroelectric NAND flash

Results from Georgia Tech

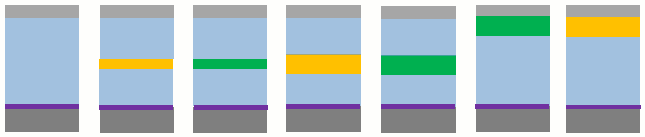
Batch 1: Tunnel Dielectric Layer

Das et al., IEDM 2023:



Batch 2: Choice and position of dielectric insert

Fernandes et al., EDL 2024:

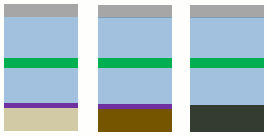


Batch 3: Hybrid structures

Fernandes et al., TED 2024:

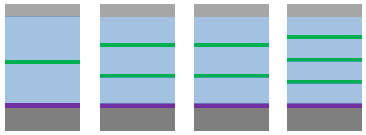
Batch 4: Channel materials

Fernandes et al., IMW 2025:



0.5 nm TDL inserts

Lee et al., TED 2024:



AOS channels

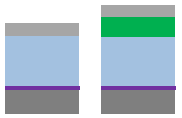
Yoo et al., VLSI 2024:

Joh et al., IEDM 2024:

Gate blocking layer

Lim et al., IEDM 2023:

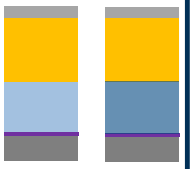
Hu et al., EDL 2024:



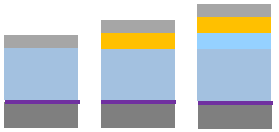
Yang et al., ICSIST 2024:

Hu et al., EDL 2024:

Chu et al., EDL 2024:

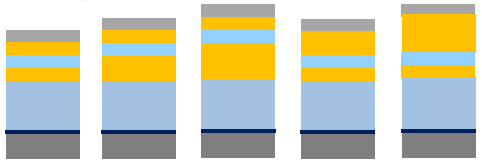


Kuk et al., IEDM 2024:

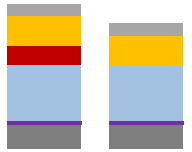


Qin et al., IEDM 2024:

Han et al., IRPS 2025:



G. Kim et al., IEDM 2024:



Ferroelectrics:

- HZO
- HAO

Dielectrics:

- SiO_2
- Al_2O_3
- Si_3N_4
- HfO_2
- TiO_2

Channel IL:

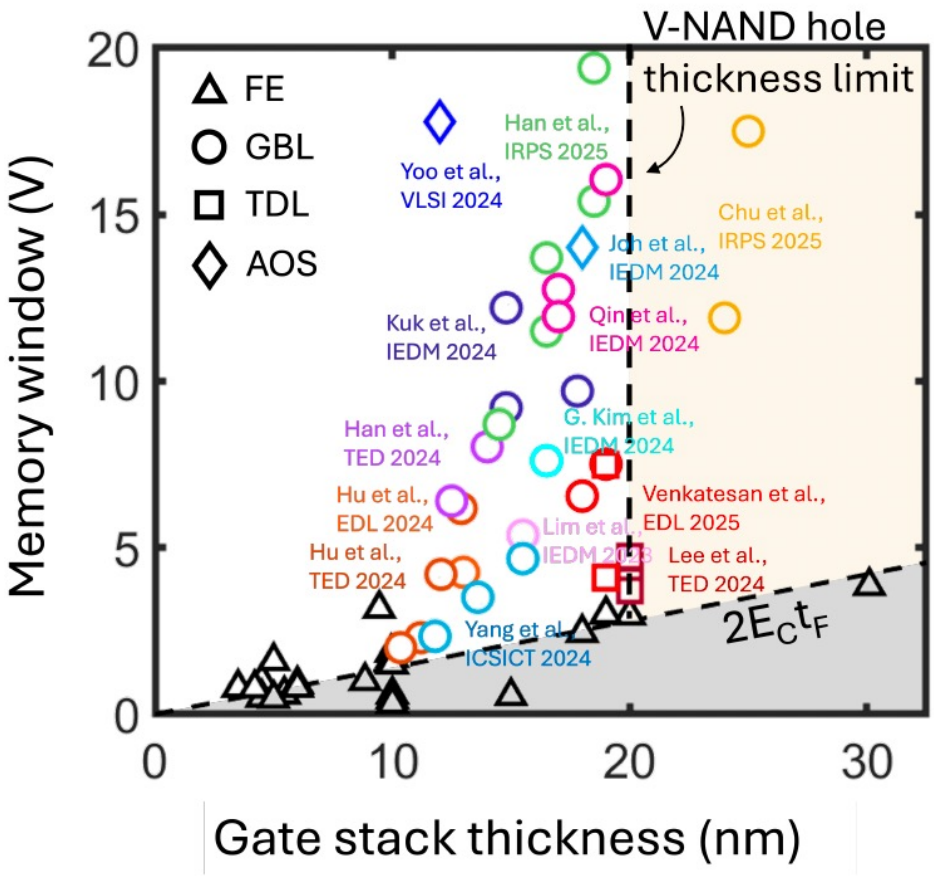
- SiO_2 IL (1 nm)
- Al_2O_3/AlON IL (1 nm)

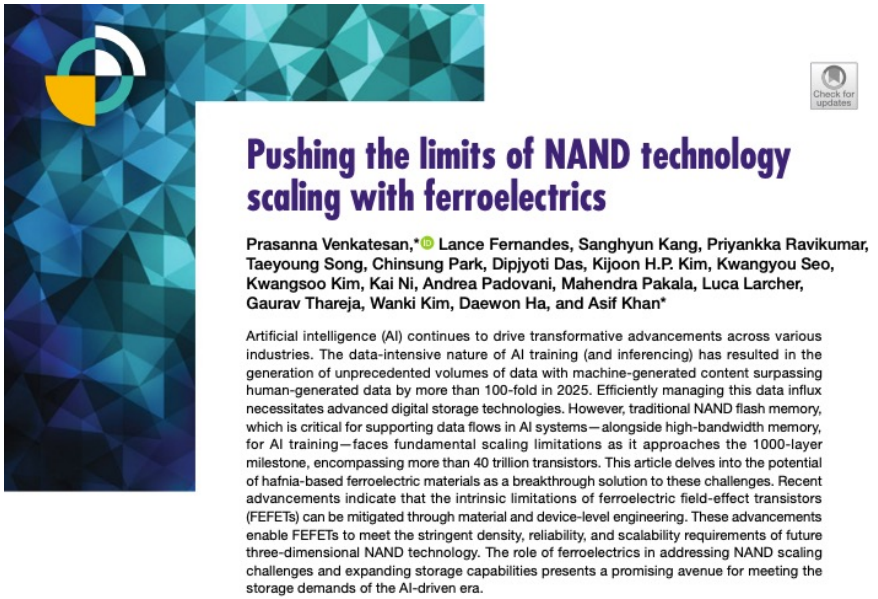
Semiconductor:

- Si
- AOS
- SOI
- Poly-Si

Georgia Tech

Increasing the MW by dielectric insertion





Pushing the limits of NAND technology scaling with ferroelectrics

Prasanna Venkatesan,* Lance Fernandes, Sanghyun Kang, Priyankka Ravikumar, Taeyoung Song, Chinsung Park, Dipjyoti Das, Kijoon H.P. Kim, Kwangyou Seo, Kwangsoo Kim, Kai Ni, Andrea Padovani, Mahendra Pakala, Luca Larcher, Gaurav Thareja, Wanki Kim, Daewon Ha, and Asif Khan*

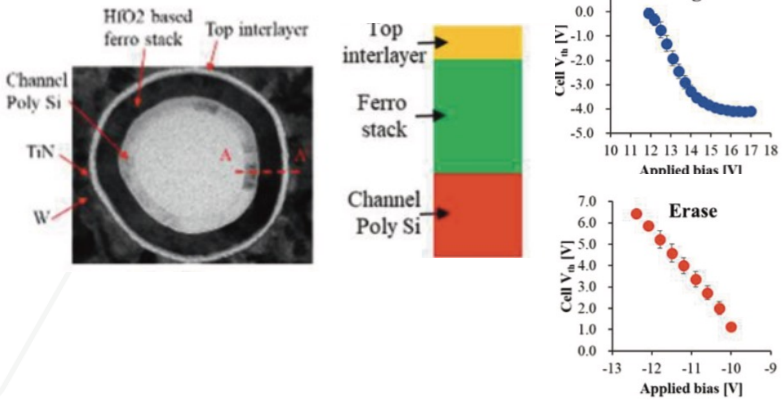
Artificial intelligence (AI) continues to drive transformative advancements across various industries. The data-intensive nature of AI training (and inferencing) has resulted in the generation of unprecedented volumes of data with machine-generated content surpassing human-generated data by more than 100-fold in 2025. Efficiently managing this data influx necessitates advanced digital storage technologies. However, traditional NAND flash memory, which is critical for supporting data flows in AI systems—alongside high-bandwidth memory, for AI training—faces fundamental scaling limitations as it approaches the 1000-layer milestone, encompassing more than 40 trillion transistors. This article delves into the potential of hafnia-based ferroelectric materials as a breakthrough solution to these challenges. Recent advancements indicate that the intrinsic limitations of ferroelectric field-effect transistors (FEFETs) can be mitigated through material and device-level engineering. These advancements enable FEFETs to meet the stringent density, reliability, and scalability requirements of future three-dimensional NAND technology. The role of ferroelectrics in addressing NAND scaling challenges and expanding storage capabilities presents a promising avenue for meeting the storage demands of the AI-driven era.

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

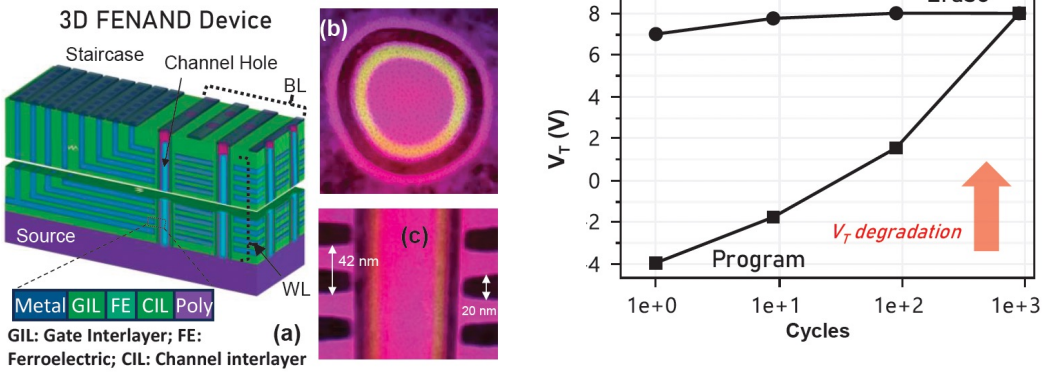
Georgia Tech

Ferroelectric NAND from the industry

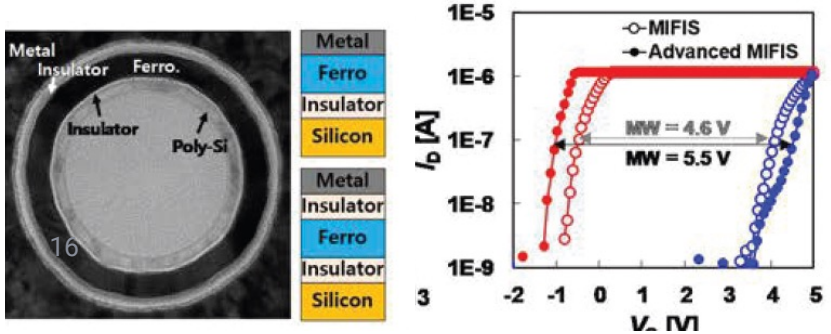
SK Hynix
Yoon et al. VLSI 2023



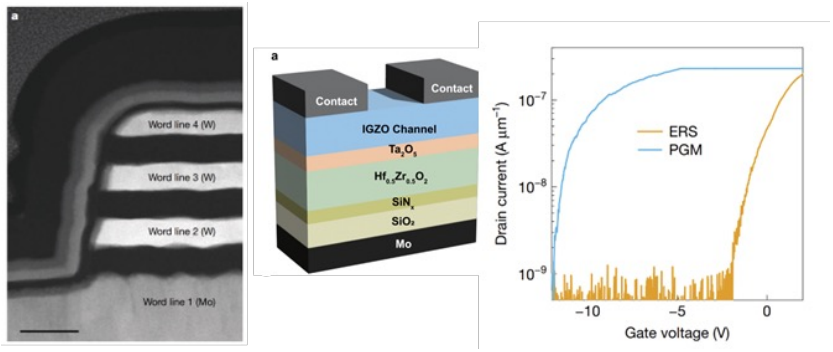
Micron
Sharma et al. VLSI 2025



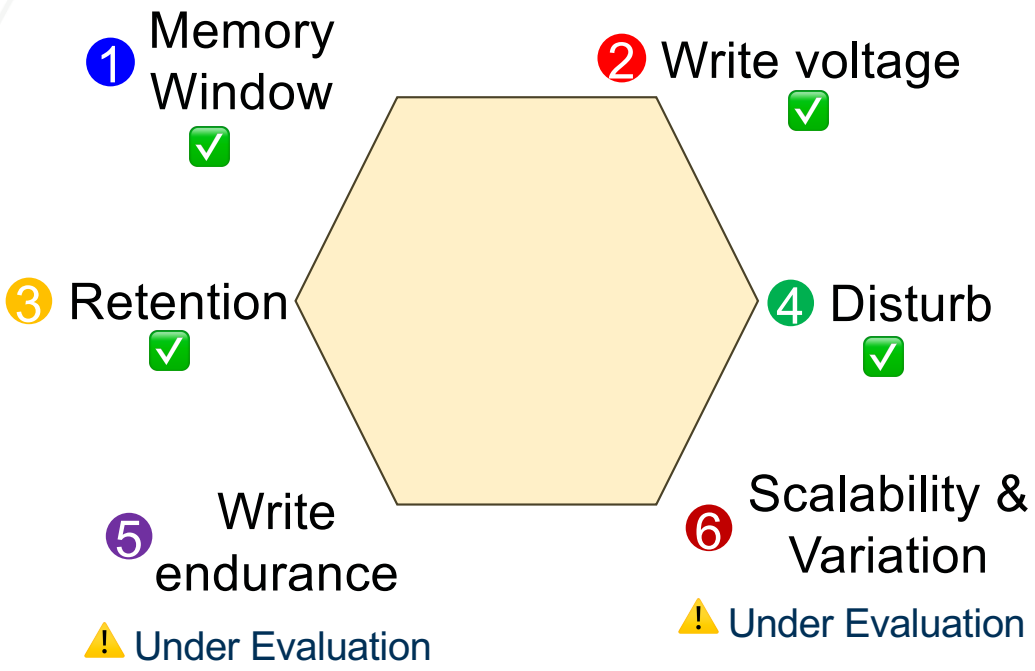
Samsung
Lim et al. IEDM 2023



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.

Georgia Tech

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.

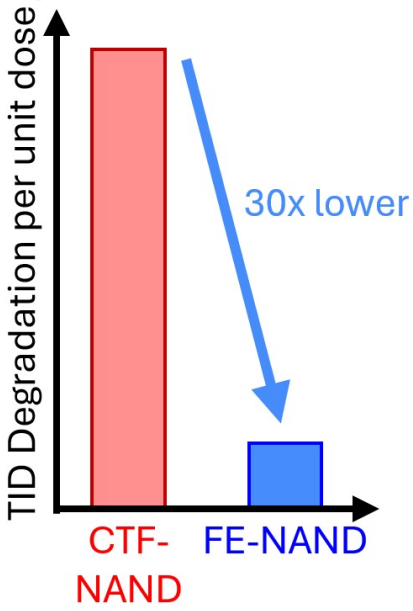
- **Magnetic storage remains impractical for many spacecraft systems** because of mass, volume, power, and mechanical constraints.

Ferroelectric NAND flash may provide a solution.

Georgia Tech

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.



NANO LETTERS

Open Access

This article is licensed under CC-BY 4.0

pubs.acs.org/NanoLett

Letter

Enabling Radiation Hardness in Solid-State NAND Storage Utilizing a Laminated Ferroelectric Stack

Lance Fernandes,* Stuart Wodzro, Prasanna Venkatesan, Priyankka Ravikumar, Ming-Yen Lee, Minji Shon, Dyutimoy Chakraborty, Taeyoung Song, Sanghyun Kang, Salma Soliman, Mengkun Tian, Jason Yeager, Jackson Adler, Jiayi Chen, Zekai Wang, Douglas Wolfe, Shimeng Yu, Andrea Padovani, Suman Datta, Biswajit Ray, and Asif Khan*

Cite This: *Nano Lett.* 2026, 26, 3390–3397

Read Online

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

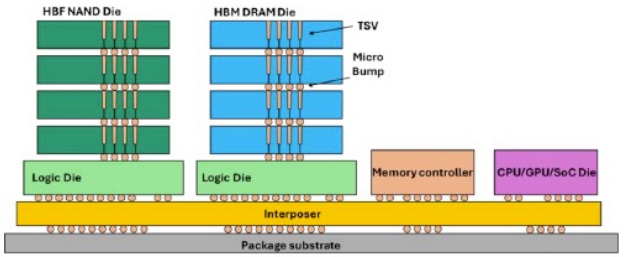
Future work: SEE, displacement damage, retention/endurance after irradiation, disturb behavior, and thermal recovery

Georgia Tech

Where NAND is Going Next

1

ACTIVE MEMORY TIER



Workload-tuned NAND

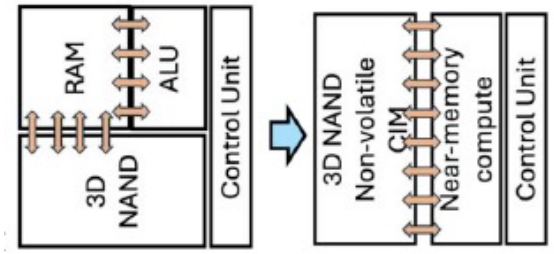
- QLC for cold archival
- SLC-like for inference buffers for AI

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

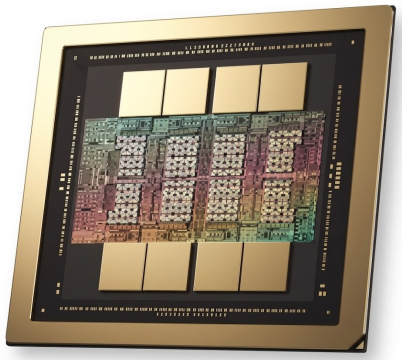
From a single technology to a portfolio — NAND, tuned to the workload

Georgia Tech

Foundational Memory & Storage Technologies

1

DRAM and
High Bandwidth Memory



NVIDIA Rubin

CAPACITY

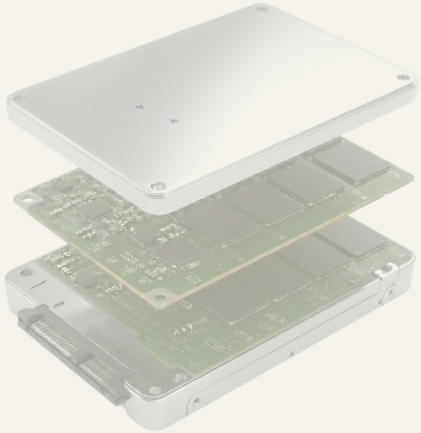
288 GB

BANDWIDTH

22 TB/s

2

NAND Flash
Storage



CAPACITY

30 – 245 TB

BANDWIDTH

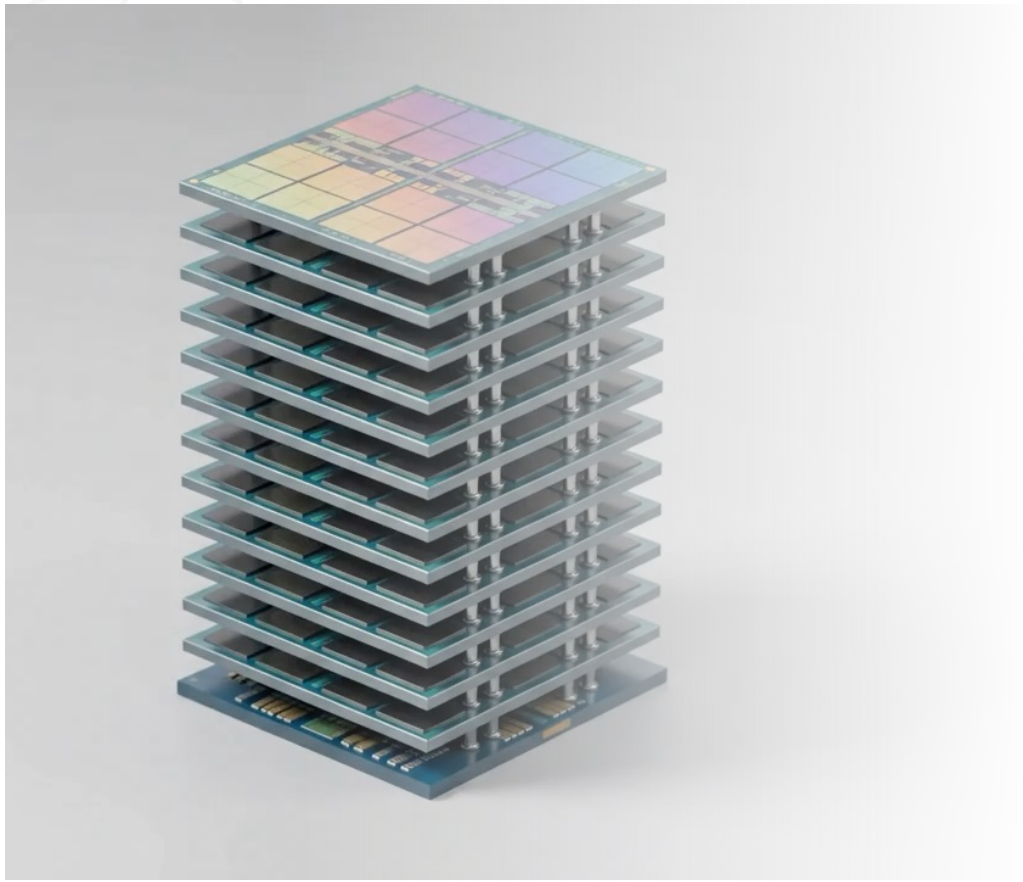
~14 GB/s

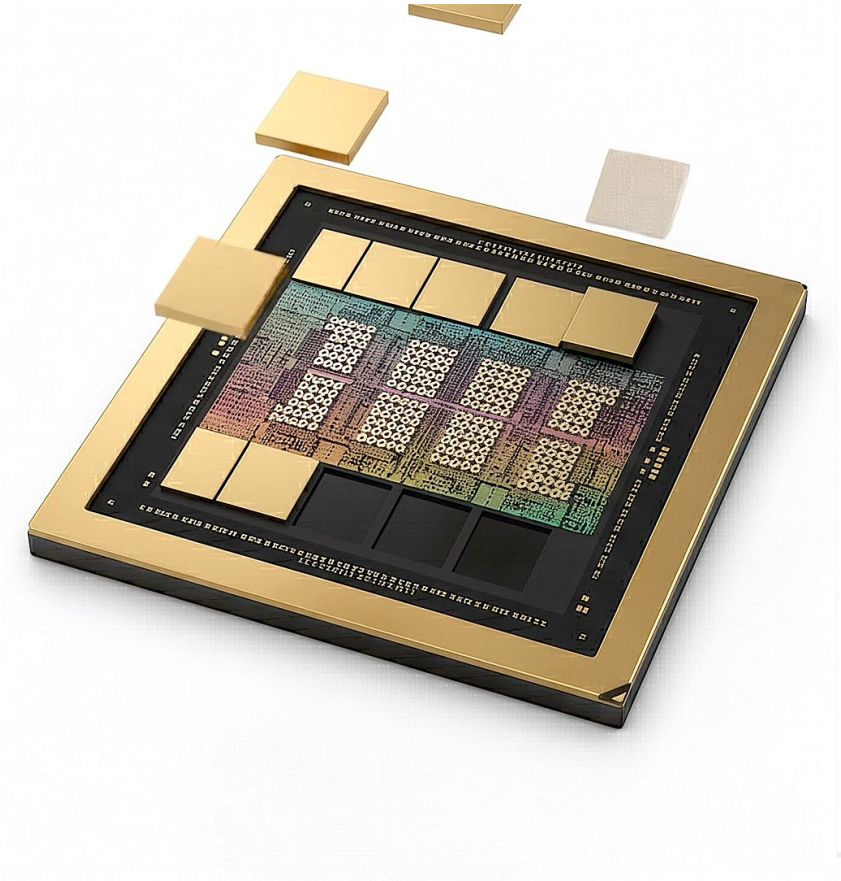
← faster, smaller | larger, slower →

Every AI system rides on this hierarchy.

Georgia Tech

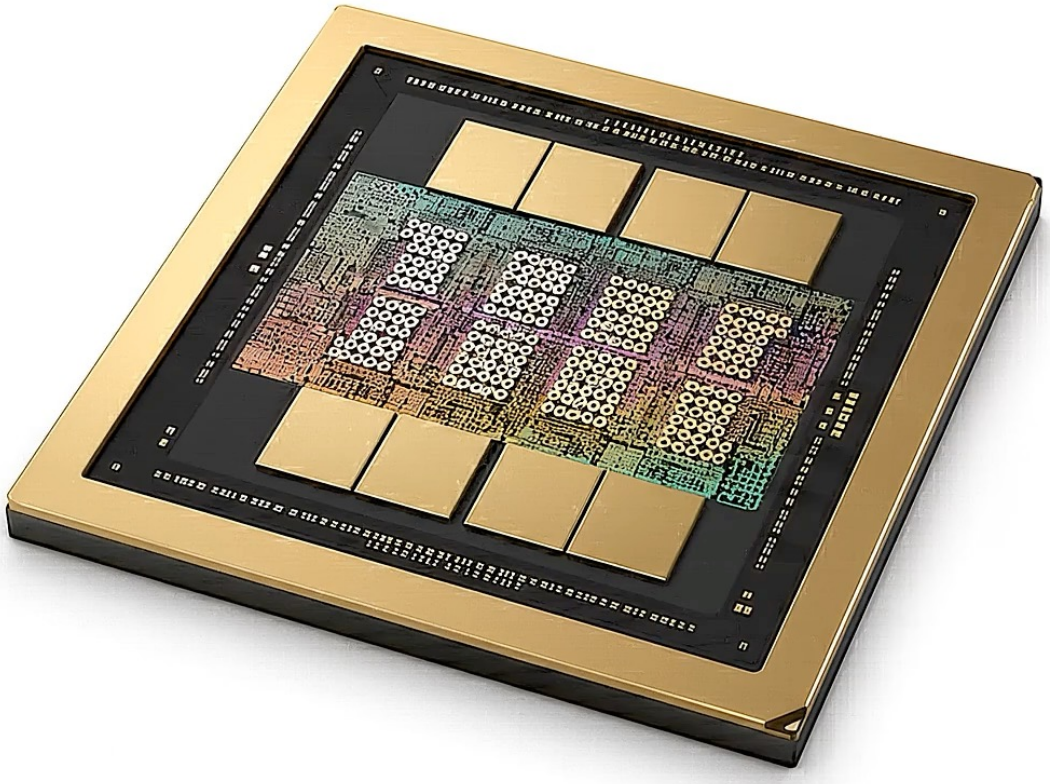
High Bandwidth Memory





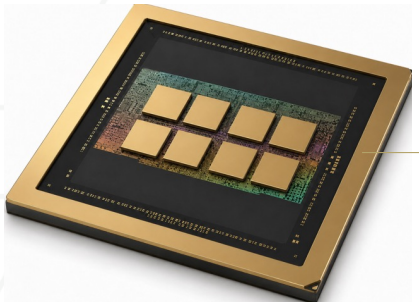
Georgia Tech

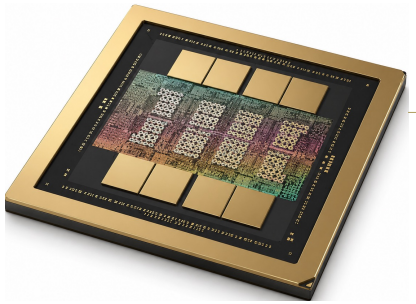
High Bandwidth Memory stacked on GPUs

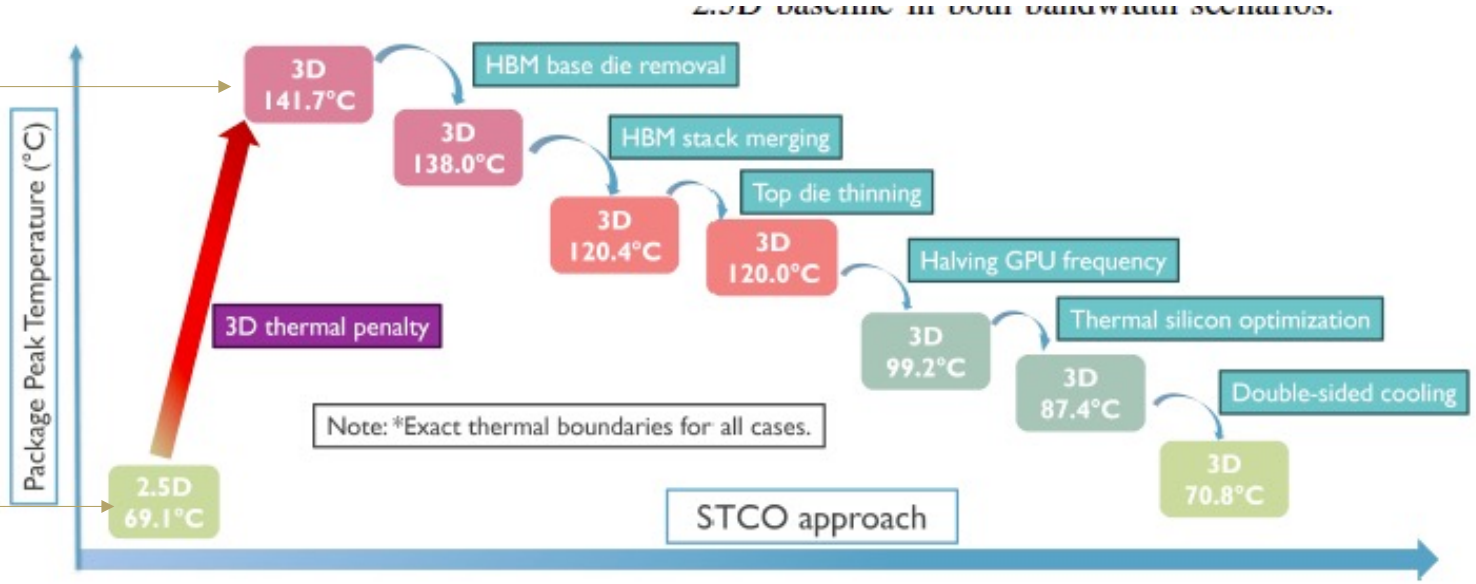


Georgia Tech

3D HBM-on-GPU hits a thermal wall





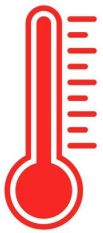


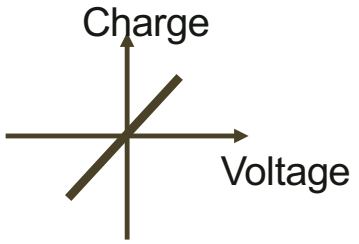
Chen, Yukai, et al. "Breaking thermal bottleneck in 3D HBM-on-GPU integration via system-technology co-optimization." *2025 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2025.

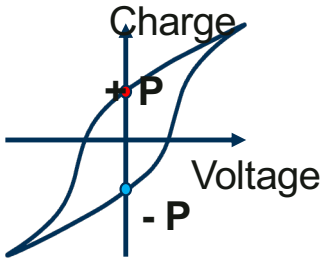
Georgia Tech

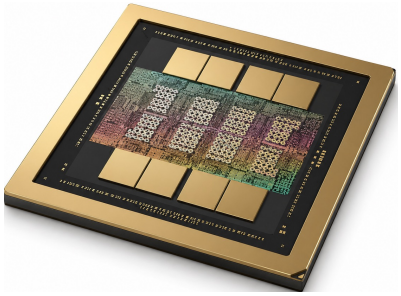
3D HBM-on-GPU hits a thermal wall

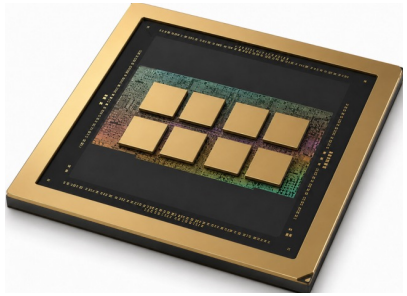
$$I_{\text{off}}(T) \propto T^2 \exp\left(-\frac{q\,V_{th}(T)}{n\,k\,T}\right)$$

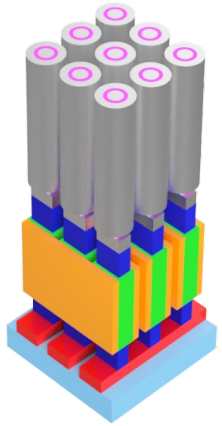


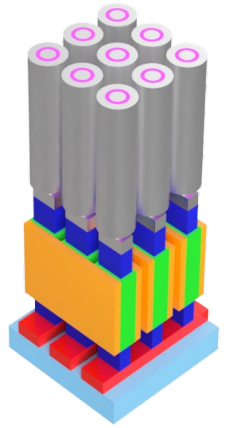




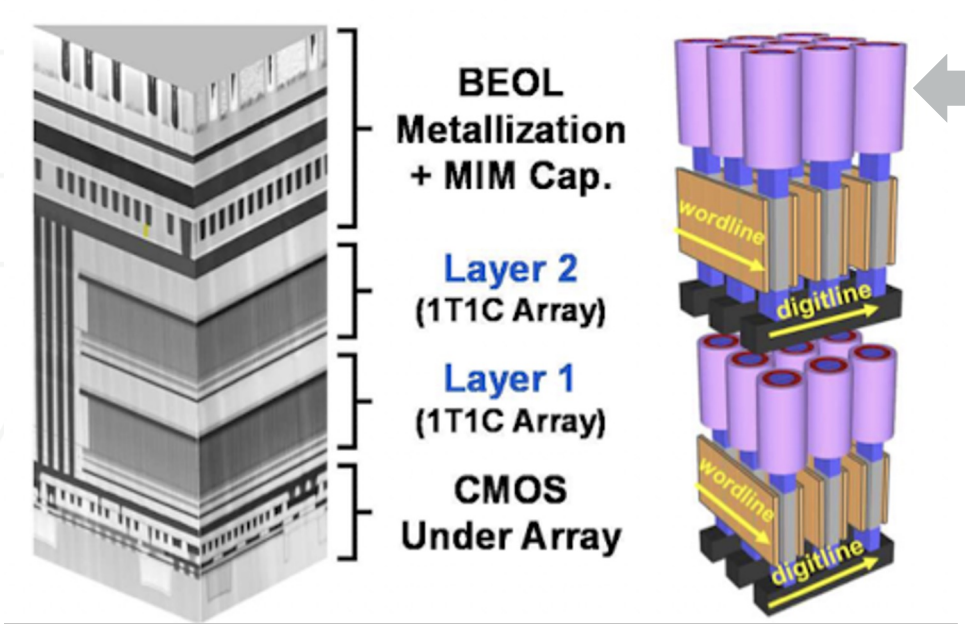


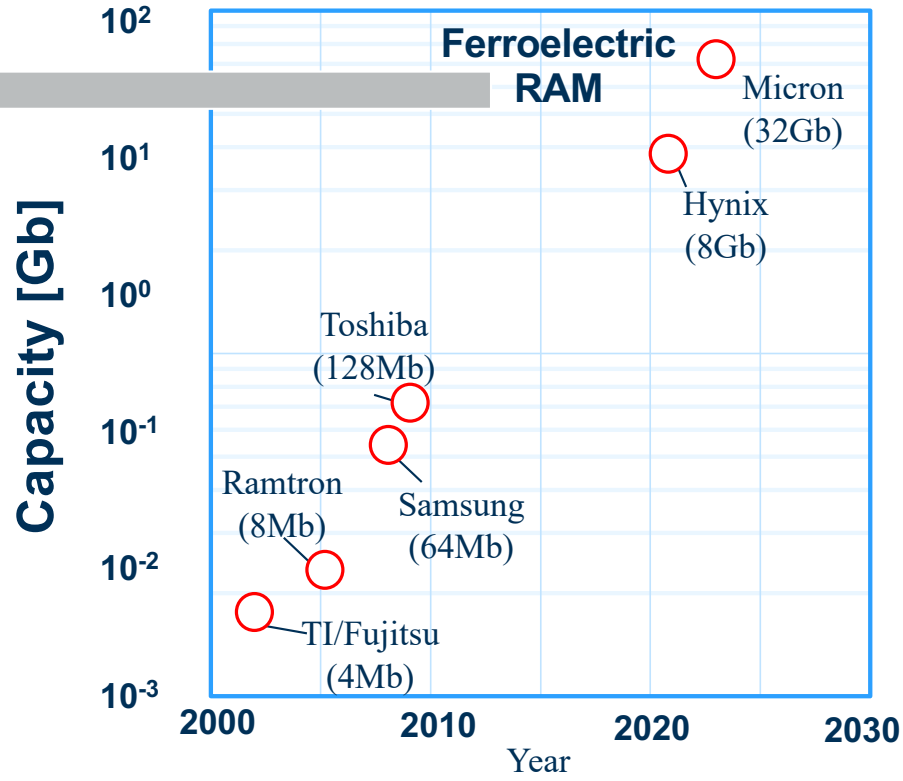






Ferroelectric RAM





Latency : < 100 ns
Capacity: 32 Gb Array
Density : 0.45 Gb/mm^2

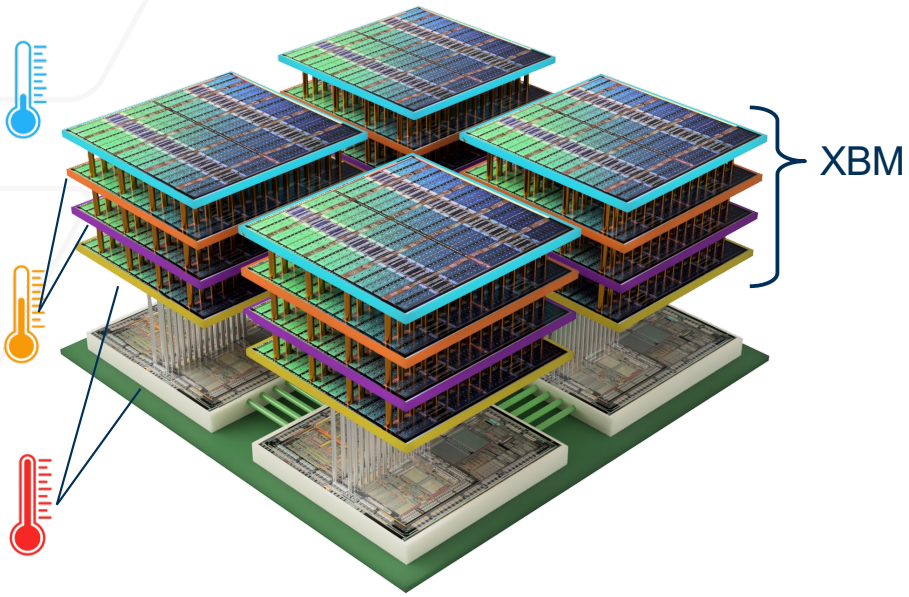
Volatility : No Refresh required

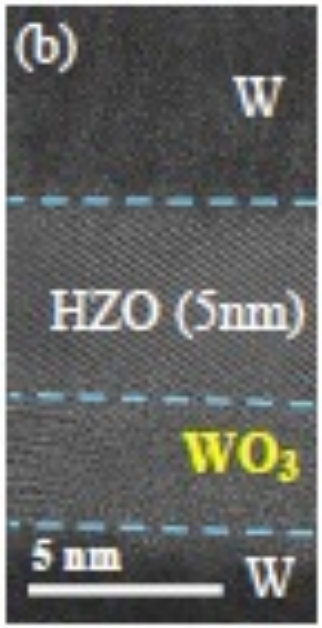
Compatible with LPDDR5 RFM.

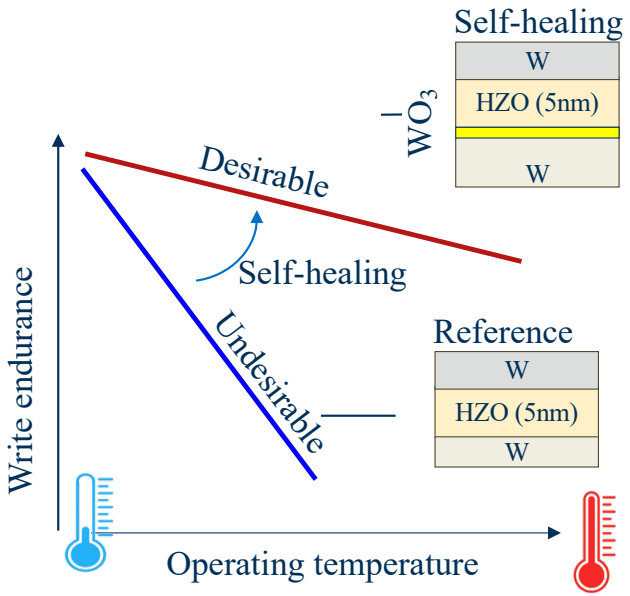
Ramaswamy, Nirmal, et al. "NVDRAM: A 32Gb dual layer 3D stacked non-volatile ferroelectric memory with near-DRAM performance for demanding AI workloads." *2023 International Electron Devices Meeting (IEDM)*. IEEE, 2023.

Georgia Tech

XBM – eXtreme Bandwidth Memory with ferroelectric RAMs



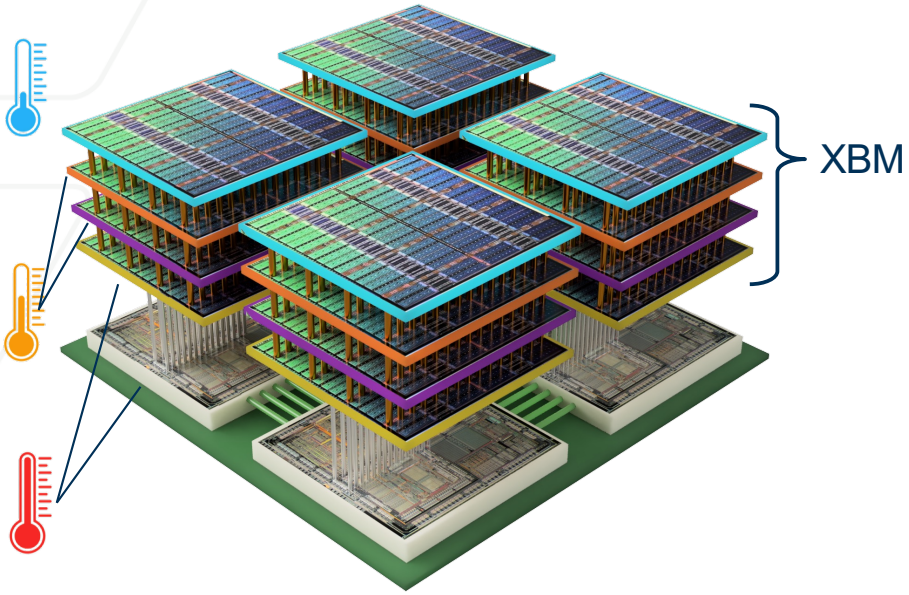




Afroze, Nashrah, et al. "Self-Healing Ferroelectric Capacitors with~ 1000x Endurance Improvement at High Temperatures (85–125° C)." *2024 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2024.

Georgia Tech

XBM – eXtreme Bandwidth Memory with ferroelectric RAMs



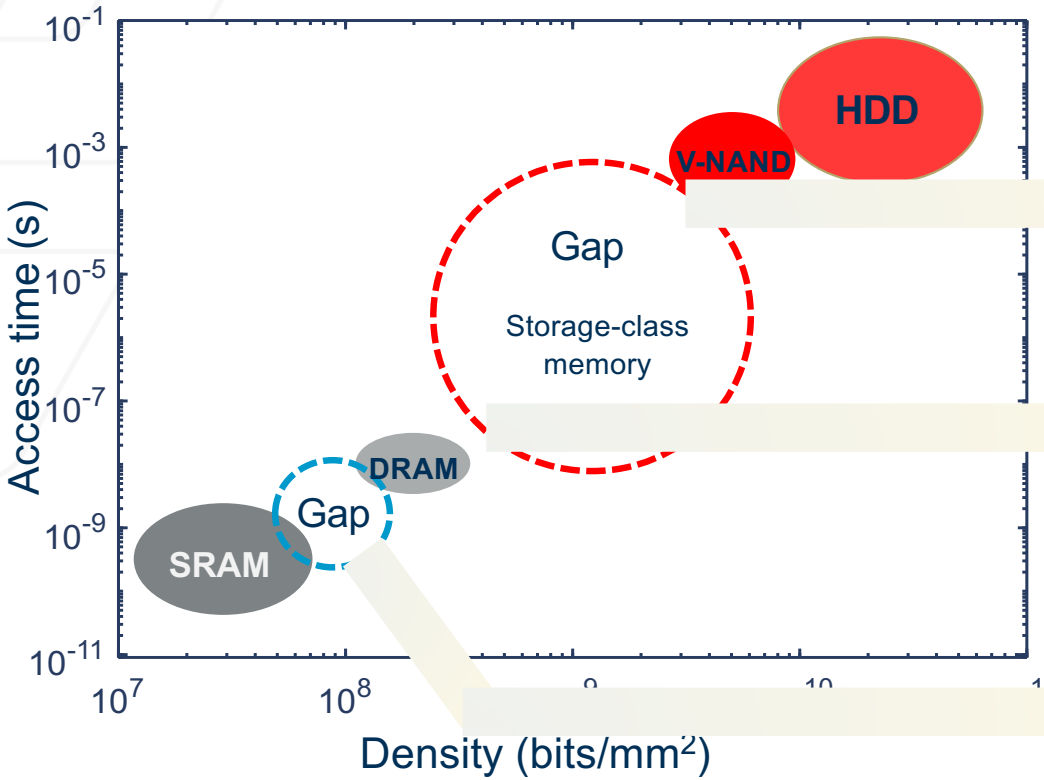
Afroze, Nashrah, et al. "Self-Healing Ferroelectric Capacitors with~ 1000x Endurance Improvement at High Temperatures (85–125° C)." *2024 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2024.

(b)

		IEDM 2023[1]	IEDM 2022[4]	Adv. Eng. Mat. 2023[5]	IEDM 2020[7]	IRPS 2023[8]	This work
Film thickness (nm)		5.7	10	7	10	7	5
Write Voltage (V)		1.5	2.2	2.1	3	2	1.8
Electric field (MV/cm)		2.6	2.2	3	3	2.9	3.6
Full or Partial switching?		Full	Partial	Full	Full	Full	Full
Wakeup		No	Yes	Yes	No	No	No
Cycles to Wake-up at RT		0	10^6	10^4	0	0	0
Pristine $2P_r$ ($\mu C/cm^2$)		60	15.1	16	12	35	38.8
Maximum write endurance (# cycles)	25-35°C	10^{12}	10^{12}*	10^7	10^9	10^8	10^{12}
	75-85°C	—	10^{10}*	10^6	—	10^7	10^9
	100-110°C	—	10^{10}*	10^6	10^8	10^7	10^8
	120-135°C	—	10^{10}*	$5x10^5$	—	$6x10^6$	$5x10^8$
Effective activation energy, $E_{a,eff}$ (eV)		—	0.23	0.12	0.38	0.35	0.23

*=partial switching

Ferroelectrics for memory and storage



Ferroelectrics for flash

- Voltage scaling below 15 V
- Z-scaling

Ferroelectrics for the DRAM/Storage class space

Ramaswamy et al. Micron, IEDM 2024

- Dual Layer 3-D stacked FRAM
- Capacity – 32 Gb (Sk Hynix DRAM –-8 Gb)
- Latency – 180 ns (LPDDR5 – 60 ns)
- Density – 450 Mb/mm^2

Ferroelectrics for embedded memories/cache

- Density
- BEOL compatibility
- Fast

Thank you

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)

Our sponsors

Semiconductor Research Corporation

SAMSUNG ELECTRONICS

3D FeM²

U.S. DEPARTMENT of ENERGY

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

Georgia Tech