Ferroelectrics for future 3D NAND storage technology (Invited)

Prasanna Venkatesan[1] and Asif Khan[1,2]
[1]*School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA*
[2]*School of Material Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA*

Abstract
Band-engineered ferroelectric field-effect transistors with dielectric inserts have emerged as a potential solution to continued z-scaling in 3D NAND devices. Here, we present a comprehensive optimization of these ferroelectric gate stacks in NAND devices for in-storage compute applications. This involves (1) exploring the design space to optimize the memory window (MW) and (2) band engineering for robust retention, and (3) implementing a novel disturb mitigation scheme to reduce pass disturb. The optimized device is then utilized to demonstrate a high-density in-storage compute solution for protein identification using open modification search.

Introduction
3D NAND based in-storage compute solutions have emerged as a viable alternative for implementing large AI models. However, reliability challenges linked with low z-pitch and high write voltages hinder further z-scaling, thereby limiting data densities. Replacing the charge trap layer in conventional 3D NAND with a ferroelectric layer has been proposed to mitigate these reliability concerns and enable 3D NAND with over 1000 layers [1-4]. Recently, QLC compatible operation (MW > 7.5 V) at low write voltage (< 15 V) in FEFETs has been achieved by either laminating a dielectric layer in the middle of the ferroelectric gate stack (Tunnel dielectric layer, TDL) or placing it next to the gate to act as a gate blocking layer (GBL). In this work, we explore different dielectrics and geometries to optimize for large MW enhancement, study retention of FEFETs with TDL and GBL gate stacks, and characterize disturb in the FEFET with the best retention. In the last section, we benchmark these FE-NAND devices against the incumbent solutions for open modification search for protein identification [5].

Results and Discussions
In order to study the role of different dielectrics and the effect of their position in the ferroelectric gate stack, two-terminal FE-MOSCAPs were fabricated using the process outlined in [6-7]. The MWs of these gate stacks are measured from the C-V curves and summarized in Fig. 2 [8]. It is identified that Al_2O_3 as a TDL and SiO_2 as a GBL exhibit the largest MWs. A hybrid gate stack with Al_2O_3 TDL and SiO_2 GBL exhibits further MW enhancement, enabling a MW as high as 11 V with the same gate stack thickness, a 4.5x improvement over the reference 18 nm HZO gate stack. The origin of MW enhancement caused by these dielectric insert in the ferroelectric gate stacks have been described in detail in [3]. FEFETs with reference (19 nm HZO), TDL (8/3(Al_2O_3)/8) and GBL (14/4(SiO_2)) gate stacks are fabricated following the process flow shown in [9-10] for the retention and disturb characterization.

Retention was characterized in the TDL and GBL FEFETs using the pulse scheme shown in Fig. 3. The evolution of the threshold voltages over time for the TDL FEFETs show less than 1% retention loss. However, the GBL FEFETs show significant retention loss resulting from detrapping of the MW enhancing charges trapped at the FE-GBL interface through the GBL [11]. While MW enhancement is achieved irrespective of the position of the dielectric, robust retention is achieved only when the dielectric is laminated in the middle of the ferroelectric gate stack. The retention loss mechanisms have been explored in detail in [17].
The evolution of the threshold voltage of the TDL FEFETs with pass disturb cycles was measured as shown in Fig. 4. Significant shift in V_T (28%) is observed with pass disturb pulses of $V_{pass} = V_T + 2$ V. It is hypothesized that the positive V_T shift is due electron trapping. In order to mitigate electron trapping, we proposed a mitigation scheme where a periodic refresh is applied every M pass disturb pulses to detrap the electrons and reduce V_T shift. The efficacy of the mitigation scheme in reducing pass disturb from 28% down to 16% and 4% with M = 1000 and M = 10, respectively, indicates that the disturb was caused predominantly by electron trapping rather than polarization reversal [12].
System-level benchmarking is performed for the implementation of open-modification search for protein identification to quantify the advantages of in-storage computing in high-density FE-NAND over other alternatives (Fig. 5). FE-NAND is found to be more energy efficient and faster than conventional 3D NAND and other solutions while offering increased data density.

Conclusion
Ferroelectric NAND devices can be optimized to enable ultra-high density in-storage compute in scales previously unprecedented, potentially allowing for petabyte scale memories. Such parallelization and in-storage operation coupled with the low write energies and latencies compared to CTF NAND devices, make FE-NAND devices as a suitable candidate for large dataset processing.

Acknowledgements
This work was supported by Samsung Electronics and SUPREME, one of the seven SRC-DARPA JUMP 2.0 centers. Fab was done at the IEN, supported by the NSF-NNCI program (ECCS-1542174).

References
[1] Han et al., IEDM (2023). [2] Lim et al., IEDM (2023). [3] Das et al., IEDM (2023). [4] Kim et al., VLSI (2024). [5] Kang et al., PACT (2022). [6] Fernandes et al., EDL (2024). [7] Fernandes et al., TED (2024). [8] Das et al., EDTM (2024). [9] Tasneem et al., IEDM (2021). [10] Park et al., JEDS (2024). [11] Venkatesan et al., EDL (2025). [12] Venkatesan et al., EDL (2024). [13] Kang et al., Bioinformatics (2023). [14] Kang et al., TCAD (2024). [15] Fan et al., DAC (2024). [16] Hsu et al., MEMSYS (2023). [17] Venkatesan et al., IRPS (2025).

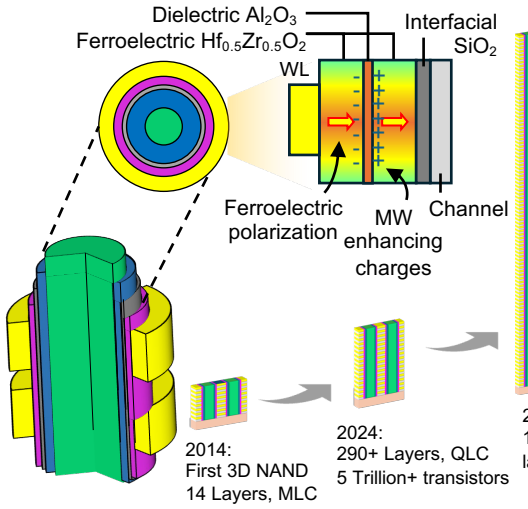

Fig. 1: Ferroelectrics have emerged as an alternative for the charge trap flash layer in 3D-NAND to enable continue z-scaling.

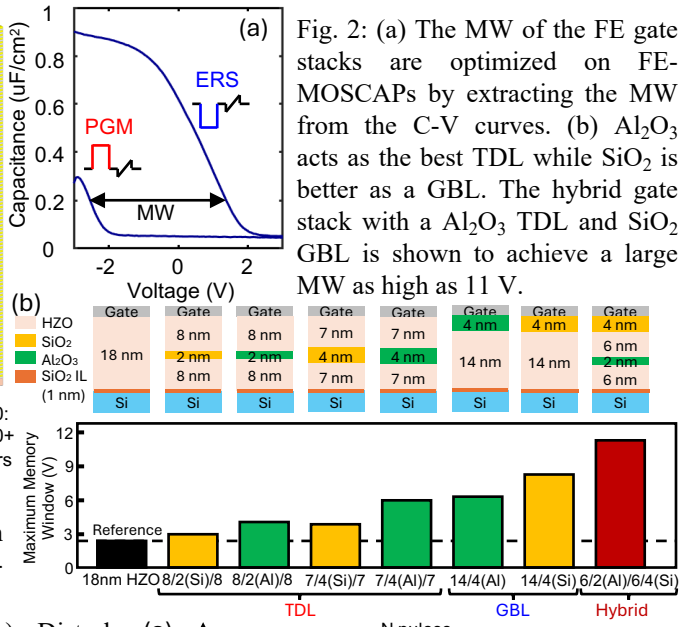

Fig. 2: (a) The MW of the FE gate stacks are optimized on FE-MOSCAPs by extracting the MW from the C-V curves. (b) Al_2O_3 acts as the best TDL while SiO_2 is better as a GBL. The hybrid gate stack with a Al_2O_3 TDL and SiO_2 GBL is shown to achieve a large MW as high as 11 V.

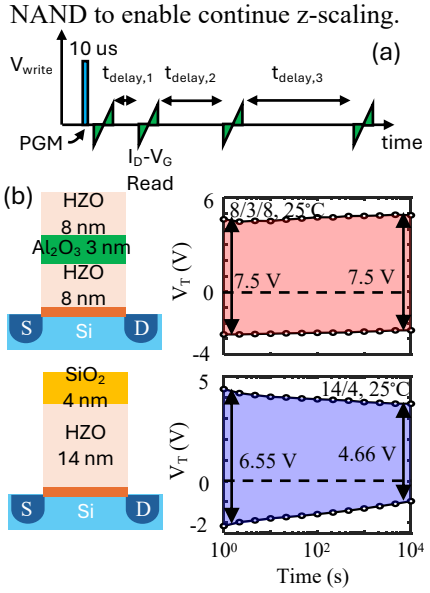

Fig. 3: (a) Pulse scheme for retention characterization. (b) Retention at RT in the TDL and GBL FEFETs shows that TDL FEFETs demonstrate robust retention while GBL FEFETs show a 29% retention loss after 1e4 s in line with other GBL FEFETs which exhibit between 25 and 50% retention loss.

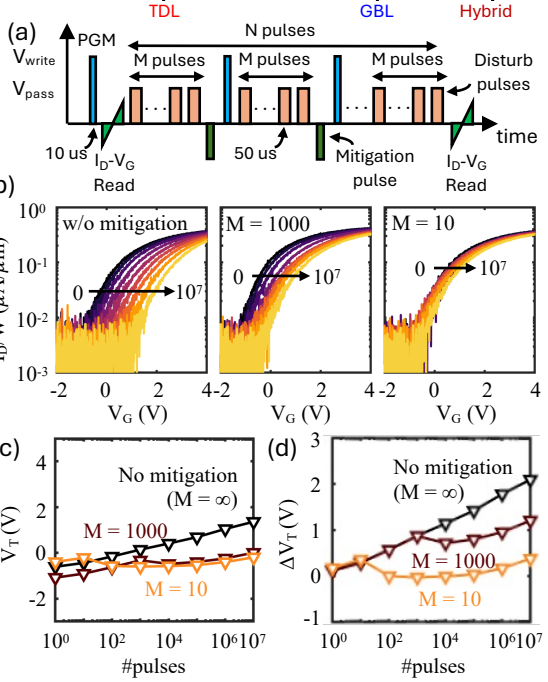

Fig. 4: (a) Disturb mitigation scheme for reducing pass disturb to acceptable levels. (b) Evolution of I_D-V_G of the PGM state in the 8/3(Al)/8 FEFET after increasing number of pass disturb pulses (V_{pass} = V_T + 2 V) with no mitigation and mitigation pulses applied every 1000 and 10 cycles. (c&d) V_T and ΔV_T shows that applying the mitigation pulse reduces pass disturb from 28% down to 4%.

Fig. 5: As a benchmarking standard, we assume a repository with one billion reference HVs and 15k query HVs. The QLC FE-NAND devices retain the advantages arising from the parallelism and read energy efficiency of CTF NAND while achieving significantly higher data densities.

Architecture	GPU [13]	DRAM [14]	MLC ReRAM [15]	3D NAND [16]	FeNAND (this work)
Compute type	-	Near-memory	Near-memory	In-memory	In-memory
Algorithm	HOMS-TC	HyperOMS-PIM-DRAM	HyperOMS-PIM-ReRAM	HyperOMS-3DNAND	HyperOMS-FeNAND
Technology Node	RTX 4090 (5 nm)	22 nm DRAM (DDR4) 28 nm node-compute	130 nm RRAM with 3M cells	NAND: 14 nm ASIC: 7 nm FinFET	FeNAND: 14 nm ASIC:7 nm FinFET
Speed	1x (23min)	2.43x	1.71x	423x	737x
Energy Efficiency	1x(454kJ)	101x	516x	7230x	22146x
Capacity Limit(per module)	24GB	128GB	3M cells	16TB	>100TB