

A Current-Mode Nondestructive Circuit Readout Scheme for High-Density 3-D Ferroelectric Memory

CHAOHAN WANG¹, WAGDY HUSSEIN², YANG XIANG^{1,2}, SHANKHA MUKHERJEE², ISMAIL CEVIK^{1,2}, FERNANDO GARCÍA-REDONDO¹, ANNACHIARA SPAGNOLO², HYUNGROCK OH², MARIE GARCIA BARDON², NICK VAN HELLEPUTTE² (Member, IEEE), JULIEN RYCKAERT², GEERT HELLINGS², DWAIPAYAN BISWAS² (Member, IEEE), AND JAN VAN HOUDT^{2,3} (Fellow, IEEE)

¹imec, Cambridge CB1 2JD, U.K

²imec, Leuven 3001, Belgium

³Department of Physics and Astronomy, KU Leuven, Leuven 3001, Belgium

Corresponding authors: W. HUSSEIN (Wagdy.Hussein@imec.be); Y. XIANG (Yang.Xiang@imec.be.)

This work was supported by the Imec's Industry Affiliation Program (IIAP).

Chaohan Wang and Wagdy Hussein contributed equally to this work.

ABSTRACT The AI era is demanding ever-increasing memory capacity and bandwidth, where 3-D-stacked ferroelectric memory (FeRAM) utilizing nondestructive readout (NDRO) ferroelectric capacitors (FeCAPs) stand as a promising candidate. This work addresses NDRO sensing signal challenges in deeply scaled 3-D FeRAMs using a current-mode sensing scheme, co-designed with the device architecture to overcome signal degradation due to capacitive charge sharing. By exploring gm cell design and time-differential bipolar readout, we demonstrate the feasibility of 3-D stacking up to 128 FeCAPs while maintaining sub- μ s access, transcending the conventional charge-sharing-defined stacking limit of 64 and achieving an effective bit density of ~ 200 Mb/mm². Using a two-stage Miller-compensated high-GBW operational transconductance amplifier (OTA) implementation, we additionally underscore the robustness of our sensing scheme against variability in CMOS circuits, but also in bit-cell FeCAPs, such as FeCAP area variation ($\sim 80\%$), memory window fluctuation (~ 15 aF), and work function asymmetry (~ 200 mV). The proposed architecture further provides a flexible design tradeoff between read speed and energy consumption, where a single-stage low-power OTA design allows read energy scaling down to ~ 10 pJ/bit at μ s-level access time. The FeCAP model is calibrated to experimental data, ensuring realistic simulation fidelity. These results offer instrumental insights into the design-technology co-optimization (DTCO) of 3-D FeRAM-based storage-class memories (SCMs).

INDEX TERMS 3-D stacking, ferroelectric capacitor (FeCAP), ferroelectric memory (FeRAM), nondestructive readout (NDRO), storage-class memory (SCM).

I. INTRODUCTION

THE rapid growth of large language models (LLMs) has triggered an unprecedented surge in data movement, with model sizes reaching billions to trillions of parameters [1]. AI inference workloads, in particular, are increasingly memory-bound [1], demanding high bandwidth and capacity. While high-bandwidth memory (HBM) offers performance, it remains costly and limited in scalability.

General-purpose DDRs and SSDs, on the other hand, suffer from bandwidth and endurance constraints [2].

Recently, 3-D-stacked ferroelectric (FE) capacitor (FeCAP)-based memories [3], [4], [5] have emerged as promising candidates for near DRAM buffer memory, offering high density, nonvolatility, low write voltage, and high endurance (Fig. 1) [6]. The nondestructive readout (NDRO) variants of FeCAP [7], [8], in particular, decouple

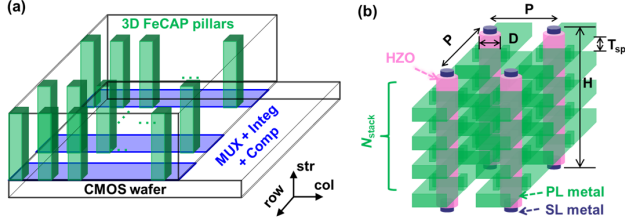


FIGURE 1. (a) Schematic of 3-D FeRAM. FeCAPs are vertically stacked in pillars that sit on top of CMOS sensing circuitry shared per column. (b) Zoomed-in view of N_{stack} FeCAP pillar array assuming a certain pitch (P), height (H), and diameter (D). FeCAPs are formed at the intersection between vertical SL metal, $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ (HZO) thin film, and horizontal PL metal, separated by a vertical spacing of T_{sp} .

read endurance from the fatigue-limited write endurance [7], making them especially suitable for read-intensive AI inference workloads.

However, the sense margin during NDRO of FeCAPs reported so far remains limited by their relatively small capacitance window (capacitive memory window (CMW) or, CMW of $\sim 10 \text{ fF}/\mu\text{m}^2$ in metal-FE-metal (MFM) FeCAPs [7]), corresponding to only $\sim 15\%$ of the DRAM ($\sim 70 \text{ fF}/\mu\text{m}^2$ [9]). Moreover, the common string line (SL) electrode in the FeCAP pillar [SL, Fig. 1(b)] inadvertently creates unwanted charge sharing among all stacked FeCAPs, degrading the signal margin inversely proportional to N_{stack} , with 64 shown as a practical upper limit [4]. This inevitably raises concerns about the read signal integrity in 3-D ferroelectric memory (FeRAM) vertical scaling.

Therefore, this article works toward overcoming limited NDRO signal in 3-D FeRAM using a circuit-device co-designed current-mode sensing architecture. We develop a robust sensing solution for deeply scaled NDRO 3-D FeRAM, targeting low latency, resilience to variability, and scalability. Specifically, the proposed current-mode sensing circuit can accurately distinguish FeCAP states across varying stack densities and device conditions.

This article is organized as follows. Section II recapitulates the device assumptions for 3-D FeRAM NDRO and implications on circuit sensing solutions. Section III presents the current-mode read sensing scheme and the simulation results, which is followed by discussion on its power-performance scalability in Section IV and conclusion in Section V.

II. 3-D FeRAM BIT DENSITY SCALING VERSUS READABILITY

Fig. 1 illustrates the 3-D FeRAM array architecture, where N -stacked FeCAPs separated by an isolating spacer thickness T_{sp} occupy an area defined by the FeCAP pillar pitch P^2 . As each FeCAP stores one bit information [i.e., “0” and “1”; see Fig. 2(b)], the normalized area per bit stored in the FeRAM array evaluates to P^2/N_{stack} , the inverse of which (N_{stack}/P^2) gives the “intrinsic bit density” (BD_{int}) [the effective bit density (BD_{eff}) however could vary as

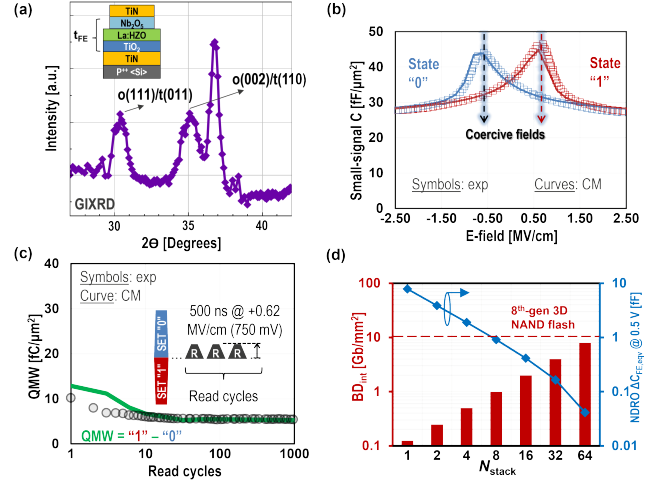


FIGURE 2. (a) GIXRD data of an in-house fabricated HZO-based FeCAP exhibiting crystallization of the HZO thin film in a mixture of FE orthorhombic (o) and/or non-FE tetragonal (t) phases. Experimentally measured (b) quasi-static CMW from small-signal C-V measurements and (c) transient charge memory window (QMW) characteristics of FeCAP [7] with CM calibration. The total thickness of the FE stack (t_{FE}) is 12 nm (9 nm HZO, 2 nm Nb_2O_5 , and 1 nm TiO_2) and (d) intrinsic bitcell density and CM-simulated equivalent NDRO ΔC_{FE} [transient-derived from (c)] assuming $P \approx 90 \text{ nm}$, $D \approx 72 \text{ nm}$, and $T_{sp} = 45 \text{ nm}$ [symbols in Fig. 1(b)]. The t_{FE} is extrapolated to 9 nm at a read voltage of 500 mV. An example of 3-D NAND flash bit density is given per [10].

a function of CMOS circuitry footprint underneath which will be addressed later]. Obviously, piling up more FeCAPs in each pillar maximizes BD_{int} , which is particularly desired for expanding memory capacity in AI inference. Nevertheless, this is accompanied by the unwanted reduction in FeCAP area (A_{FE}), i.e., the lateral area of the cylinder at the cross-point between the plate line (PL) and SL [SL; Fig. 1(b)], and accordingly changes the capacitance. Knowing that the FeCAP pillar height is constrained by etch aspect ratio ($AR \leq 50$) in the fabrication process, i.e., $H \leq AR \times D$, the A_{FE} is defined by

$$A_{FE} \approx \pi \cdot D \cdot \left(\frac{H}{N_{stack}} - T_{sp} \right) < \pi \cdot D \cdot \left(\frac{AR \cdot D}{BD_{int} \cdot P^2} - T_{sp} \right) \quad (1)$$

where A_{FE} decreases even faster than $1/BD_{int}$.

On the other hand, for a FeCAP such as an MFM stack with a 9-nm HZO as the FE layer, 1-nm TiO_2 as the seed layer, and 2-nm Nb_2O_5 as the cap layer (refer Fig. 2(a) for grazing incidence X-ray diffraction (GIXRD) data), NDRO relies on the CMW between the states “0” and “1” [Fig. 2(b)], which is currently inadequate [7], [9]. Moreover, to prevent inadvertent switching of FeCAP polarization state during NDRO, one is obliged to suppress the read field (bias) to below the coercive field (voltage), which in turn renders a limited transient charge memory window (QMW; normalized) of only $\sim 6 \text{ fC}/\mu\text{m}^2$ (Fig. 2(c); [7]).

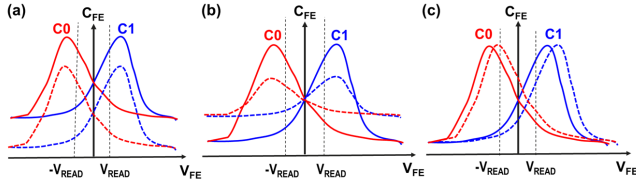


FIGURE 3. Sources of FeCAP variation. (a) Area variation related to process nonuniformity. (b) Memory window fluctuation due to polycrystalline HZO FE material. (c) Shift in effective voltage from asymmetric FeCAP electrodes.

The compound effect of A_{FE} reduction and NDRO nature leads to an inevitable tradeoff between 3-D FeRAM BD_{int} and equivalent ΔC_{FE} in NDRO. In Fig. 2(d), we calculate the equivalent ΔC_{FE} for different N_{stack} with A_{FE} per (1) assuming a 90-nm pillar pitch, while the CMW is estimated using compact model (CM) extrapolation to 6-nm HZO layer (plus 2-nm Nb_2O_5 and 1-nm TiO_2 [7]), at a read voltage of 500 mV (field 0.56 MV/cm); note that the same assumptions will be used henceforth consistently. The CM used in this analysis is calibrated to experimental FeCAP data from [7], ensuring realistic simulation fidelity. Remarkably, according to Fig. 2(d), a 1 Gb/mm² BD_{int} at eight-stacked FeCAPs would allow no more than 1 fF ΔC_{FE} , while an attempt to reach typical 3-D NAND flash density (~ 10 Gb/mm²; [10]) with 64-stacked FeCAPs would land below <0.1 fF. The capability of distinguishing sub-fF ΔC_{FE} is therefore the foremost challenge to NDRO sensing.

On top of the magnitude of ΔC_{FE} , another major challenge to NDRO in scaled 3-D FeRAM lies in the variability of FeCAPs (Fig. 3) that broadly comes in three types: 1) area variation related to process nonuniformity; 2) ΔC_{FE} fluctuation in polycrystalline HZO layer; and 3) shift in effective voltage across the FE layer (V_{FE}) from asymmetric electrodes. Possible mismatch among C_{FE} for FeCAPs in proximity, on the other hand, is left for future work.

III. CURRENT-MODE SENSING SCHEME

A. CIRCUIT ARCHITECTURE

The (sub)-fF-level C_{FE} in deeply scaled 3-D FeRAMs significantly complicates conventional, charge-based switched-capacitor (“SW-cap”) integrator-based sensing schemes [11], which would necessitate multiple charge-transfer iterations in order to accumulate a sufficient signal amplitude, risking excessive power consumption and poor scalability. Addressing these limitations, we instead propose a current-domain readout architecture, as shown in Fig. 4(a), tailored to NDRO in 3-D-stacked FeRAM. The core concept is to employ a transconductance (gm) cell to convert the small voltage difference induced by ΔC_{FE} into a current signal, which is subsequently integrated on a capacitor (C_{int}). The gm cell amplifies the minute voltage signal and can potentially isolate parasitic effects associated with FeCAPs and routing interconnects ($C_{\text{p-cell}}$ in Fig. 4); this will be elaborated on in Section III-B. During read

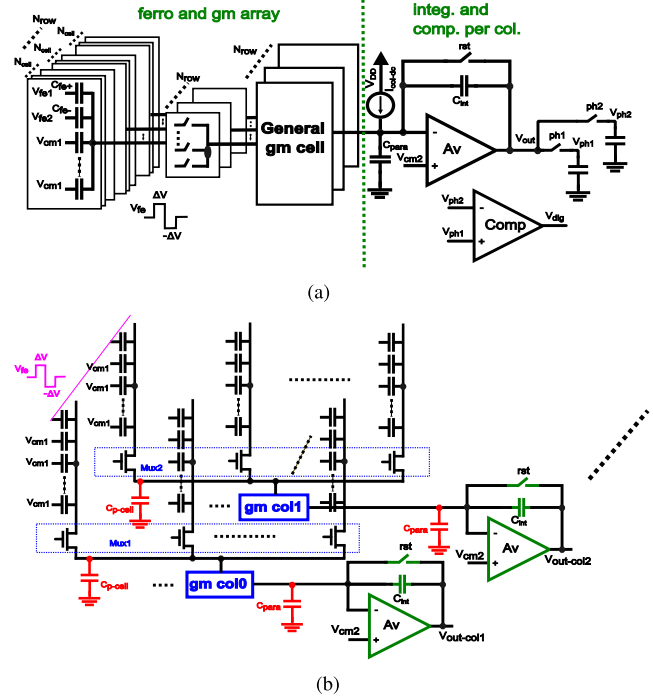


FIGURE 4. (a) Proposed current-mode sensing scheme with gm cell. A current integrator generates different output voltages (V_{out}). A current subtractor removes the DC component of the gm current. (b) Array readout of selected word bytes with multiplexing.

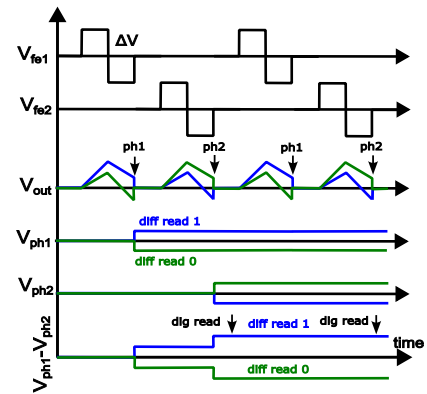


FIGURE 5. Timing diagram of the proposed time-differential bipolar readout.

operation, a read voltage V_{fe} is applied across the target FeCAP [Fig. 4(a) and (b)], forming a voltage divider at the gate of the gm cell. The resulting C_{FE} -dependent current is then integrated on the capacitor C_{int} , generating V_{out} at the output node. The following equation shows the dependency of V_{out} and C_{FE} :

$$V_{\text{out},0/1} = V_{\text{fe}} \frac{C_{\text{FE},0/1}}{C_{\text{FE},0/1} + C_{\text{p-cell}}} \times g_m \frac{T_{\text{int}}}{C_{\text{int}}}. \quad (2)$$

Fig. 5 shows the timing diagram of the current-mode system featuring *time-differential bipolar readout*, where one bit information is stored in two complementary FeCAPs

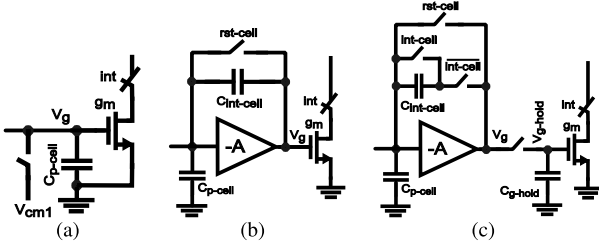


FIGURE 6. (a) Basic gm cell. (b) gm cell with simple preamplification. (c) gm cell with an in-cell SW-cap integrator.

[C_{fe+} and C_{fe-} in Fig. 4(a)] with opposite states (i.e., “01” and “10” used as one binary bit). During the read operation, the selected complementary FeCAP pair [Fig. 4(a)] is sensed in two consecutive phases. In the first phase, one of the two FeCAPs is driven with a bipolar V_{fe1} pulse, and the resulting FeCAP state-dependent gate voltage is converted by the gm cell into a modulated current that is integrated on C_{int} , producing the first sampled output V_{ph1} (Fig. 5). In the second phase, the complementary FeCAP (which sits in spatial proximity) is driven with a V_{fe2} pulse of identical time and amplitude to V_{fe1} and sensed through the same gm-cell and current-integrator path, producing the second sampled output V_{ph2} (Fig. 5). This generates two V_{out} voltages (V_{ph1} and V_{ph2}) in two consecutive clock cycles, sampled at ph1 and ph2, respectively. The binary state stored in the complementary FeCAP pair is then determined by their difference $V_{ph1} - V_{ph2}$ through its sign and magnitude.

Importantly, the use of *time-differential bipolar readout* is crucial to overcoming CMOS mismatch in circuitry, as the two *complementary FeCAPs* will experience the same mismatch that cancel each other out, while the *bipolar V_{fe}* traverses both positive and negative voltages across the FeCAP, making it immune to FeCAP electrode asymmetry [Fig. 3(c)]. To enable such cancellation, the two complementary FeCAPs encoding one bit are intentionally placed above or under each other within the same vertically stacked pillar, so that dominant process-induced FeCAP variations are spatially correlated and move together; residual uncorrelated FeCAP-to-FeCAP mismatch is not fully canceled and remains part of the device-variability margin. This time-differential read approach effectively relaxes the gm cell size and supports a greater N_{stack} even if it halves the BD_{eff} and doubles the read time. To recover BD_{eff} , each gm cell in the array is shared by multiple FeCAP pillars through multiplexing [Fig. 4(a)] which reduces its layout overhead; meanwhile, multiple gm cells operating on the same row define the word width [Fig. 4(b)].

B. EXAMPLE IMPLEMENTATION IN 22-nm CMOS TECHNOLOGY

Three gm-cell variants are designed with increasing design complexity (Fig. 6), aiming at resolving shrinking ΔC_{FE} with N_{stack} (Table 1). Notably, variant (c) incorporates a *preamplifier* (-A) and an *in-cell SW-cap integrator* in the

TABLE 1. Simulated Comparison of gm-Cell Architectures.

Architecture	Basic g_m	In-cell amp g_m	In-cell sw-cap integ g_m
C_{FE0}/C_{FE1} (fF)	1.62 / 2.03	0.65 / 0.81	0.063 / 0.079
ΔC_{FE} (aF)	410	160	16
Max. supported N_{stack}	16	32	128
MUX ratio	16-to-1	64-to-1	64-to-1
FeRAM BD_{eff} (Mbit/mm ²)	26	50	200
Read time (μ s)	0.16	0.20	0.82
Energy per bit (pJ, 27°C)	153	223	563
Success read (%) under noise	>99.87	>99.87	>97.72

TABLE 2. Sim. Read Across gm Cells Versus FeCAP var.

Variation	Basic gm	In-cell amp.	SW-cap integ.
A_{FE} nonuniformity	$\pm 80\%$	$\pm 80\%$	$\pm 80\%$
ΔC_{FE} fluctuation	180 aF	77 aF	15 aF
V_{FE} shift	± 200 mV	± 200 mV	-60 to +200 mV

¹ N_{stack} and MUX ratio consistent with Table 1

gm cell; the *preamplifier* forces the SL voltage in the pillar [Fig. 1(b)] at a constant virtual ground level regardless of N_{stack} , while the *in-cell SW-cap integrator* transfers C_{FE} -dependent charges on the cell integrator $C_{int,cell}$, accumulating voltage signal on the gm gate (V_g) over integration cycles. Importantly, it is via the feedback provided by the *preamplifier* (gain ~ 200) that our design successfully decouples the output sensing signal from the “self-parasitic” ($C_{p,cell}$) of stacked FeCAPs, i.e., $N_{stack}C_{FE}$ [4], enabling vertical stacking up to 128 where the ΔC_{FE} is less than 20 aF.

Meanwhile, the introduction of in-cell *preamplifier* in Fig. 6(b) and (c) also makes its input-referred noise the dominant noise contribution rather than that of the column operational transconductance amplifier (OTA) or from other sources. This *preamplifier* noise appears at V_g and is subsequently amplified through the gm cell and current integrator. To ensure output signal integrity, the relevant noise metric needs to be evaluated at the comparator input by comparing the time-differential output signal ($V_{ph2} - V_{ph1}$) [Fig. 4(a)] with its simulated output noise level.

The proposed sensing scheme is simulated on a 22-nm CMOS technology platform [12] as a proof of concept, using thick-oxide transistors (1.8 V V_{DD}). The integrator employs a two-stage OTA with Miller compensation, supporting a GBW of 200 MHz. In the implemented design, the OTA achieves a simulated DC gain (A_0) of 83.6 and 3-dB bandwidth (ω_0) of 83.85 krad/s, and a phase margin of 71° at the typical 27 °C corner. For the column gm current integrator [Fig. 4(a)], the OTA DC open-loop gain determines the steady-state virtual-ground accuracy at the OTA input that sees the column bitline parasitic capacitance [C_{para} ; Fig. 4(a)], while the GBW determines how fast the input gm current settles onto the feedback integration capacitor C_{int} . The gm current settling time (τ_{col}) is jointly determined by the OTA GBW and by the C_{para} at the

TABLE 3. Read Simulations: “Basic” and “In-Cell SW-Cap Integrator” gm Cells in Fig. 6 Versus CMOS Variabilities.

Condition	Basic Gm Cell ($N_{\text{stack}} = 16$)			SW-Cap Integrator Gm Cell ($N_{\text{stack}} = 128$)		
	$\Delta V_{\text{ph}12}^1$ (mV)	Noise (2σ) (mV)	Mismatch (2σ) (mV)	$\Delta V_{\text{ph}12}^1$ (mV)	Noise (2σ) (mV)	Mismatch (2σ) (mV)
Nominal	79	11.2	–	125	97.5	–
Worst PVT corner	52	8.4	–	68	54	–
Mismatch ¹	–	–	9	–	–	33

¹ $V_{\text{ph}1} - V_{\text{ph}2} = \pm \Delta V_{\text{ph}12}$, hence the swing being $2 \times \Delta V_{\text{ph}12}$

² Mismatch in typical corner, at nominal V_{DD} and 27 °C

integrator input per the following form:

$$\tau_{\text{col}} = \frac{1}{\omega_p} = \frac{1}{\omega_0 \left(1 + \frac{A_0 C_{\text{int}}}{C_{\text{para}} + C_{\text{int}}}\right)}. \quad (3)$$

Using $\text{GBW} = A_0 \omega_0 = 1.27 \text{ Grad/s}$, $A_0 = 15 \text{ kS}$, $C_{\text{int}} = 300 \text{ fF}$, and $C_{\text{para}} = 2 \text{ pF}$, the τ_{col} evaluates to 6 ns.

Utilizing the gm cell and OTA designs above, Table 1 gives the power-performance-area estimate of the FeRAM NDRO operation. It is revealed that the read time is primarily dictated by the current integration time needed to achieve a sufficient sense margin across CMOS PVT corners in the presence of noise sources (Table 3), a process that oftentimes lasts significantly longer than τ_{col} . Indeed, the read time amounts to 160 ns in the case of basic gm cell [Fig. 6(a)], whereas for the *in-cell SW-cap integrator* gm cell [Fig. 6(c)] it takes up to 820 ns before the accumulated signal to noise ratio (SNR) at the gate of gm cell can effectively overcome the noise generated by the *preamplifier*. In both cases, however, the read time is significantly longer than and thus not limited by τ_{col} (6 ns), and by extension not limited by the GBW (200 MHz) of our OTA design. In fact, based on our sensitivity studies, a 75-MHz GBW would suffice to support all three types of gm cell designs presented in Fig. 6 in terms of sensing signal integrity; in other words, the foremost challenge to FeCAP NDRO operation is not about the speed, but rather about the signal amplitude. This suggests that the OTA design space allows additional power-speed tradeoff without compromising NDRO itself (see Section IV).

Notwithstanding the demonstrated high resolution in ΔC_{FE} in deeply scaled 3-D FeRAMs, the growing complexity in gm cells with extra transistors (Fig. 6) does add to silicon area overhead, notably, the *in-cell preamplifier*. Fortunately, this area can be shared via multiplexing (Fig. 4), allowing BD_{eff} to linearly scale with N_{stack} across different gm cells (Table 1). Comparing BD_{eff} and BD_{int} [Fig. 2(d)], though, it is evident that the density scaling in our design is limited by the sensing circuitry rather than by the 3-D FeRAM array, which leaves room for improvement. Further, with enhanced ΔC_{FE} resolution in Fig. 6(b) and (c), the read time and energy inevitably increase with N_{stack} (and hence BD_{eff}) in order to guarantee the desired success read

rate, thus presenting a tradeoff between density scaling and energy/speed.

C. VARIABILITY RESILIENCE

Unless otherwise stated, the robustness simulations reported in Tables 2 and 3 use the high-GBW two-stage OTA implementation. The FeCAP variability simulations further assume that the two complementary FeCAPs used to encode one bit are adjacent within the same vertically stacked pillar, so that dominant process variations are correlated. The proposed architecture demonstrates notable resilience to the FeCAP variability discussed in Fig. 3, with the results summarized in Table 2. Taking the “basic gm cell” as an example [Fig. 6(a)], the system tolerates up to 80% area variability, can reliably read out the signal under a ΔC_{FE} as large as 180 aF and remains functional up to 200 mV V_{FE} shift caused by electrode asymmetry (at a V_{read} of 500 mV).

Regarding CMOS PVT and mismatch variations, Table 3 presents circuit simulation results obtained using both basic gm cell [Fig. 6(a)] and the *in-cell SW-cap integrator* gm cell [Fig. 6(c)] under the same conditions presented in Table 1. PVT simulations include two extreme operating temperatures (−20 °C and 100 °C), a $\pm 10\%$ supply-voltage variation, and all five process corners. In the worst PVT corner, $\Delta V_{\text{ph}12} = |V_{\text{ph}1} - V_{\text{ph}2}|$ still holds 52 and 68 mV for the basic and *in-cell SW-cap integrator* gm cells, respectively, offering a 43-mV (basic) and 14-mV (*in-cell SW-cap*) margin above the 2σ output noise level. In the meantime, the CMOS mismatch proves to have fairly limited impact ($\leq 33 \text{ mV}$ at Monte Carlo 2σ) thanks to our time-differential read technique (Fig. 5).

IV. SCALABILITY OF CURRENT-MODE SENSING

Incidentally, the read energy consumption in the read scheme discussed above appears noticeably high ($\sim 10^2 \text{ pJ/bit}$; Table 1) even compared to 3-D triple-level cell (TLC) 3-D NAND flash NAND flash ($\sim 10 \text{ pJ/bit}$; [13]). The root cause is identified as the OTA in the integrator [Fig. 4(a)] accounting for $>75\%$ of the total power consumption, which so far has been optimized for GBW by adopting a two-stage Miller OTA (Fig. 7). Replacing it with a low-GBW design (e.g., single-stage 5T OTA; Fig. 7) curtails energy consumption by removing extra biasing current while only slightly increasing the

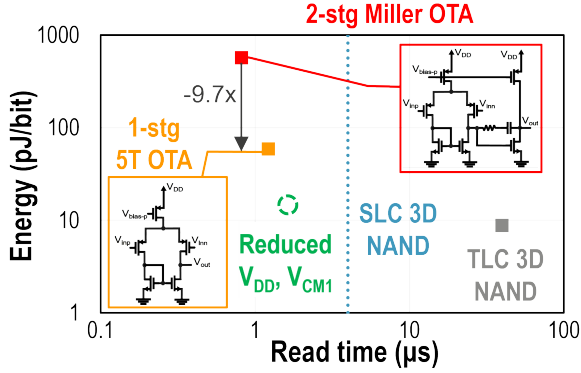


FIGURE 7. Simulated read speed-energy in $N_{stack} = 128$ (MUX 64-to-1) using gm cell design in Fig. 6(c), with energy-speed tradeoff achieved by simplified OTA design and biasing conditions. TLC 3-D NAND flash read energy and speed are based on [13], while SLC read speed on [14].

integration time (from 0.82 to 1.2 μ s; Fig. 7) required to reach the desired sensing margin.

It should be noted that the robustness results reported in Tables 2 and 3 correspond to the high-GBW two-stage OTA implementation, whereas the single-stage low-power OTA in Fig. 7 is only used to illustrate energy-speed tradeoff. Nevertheless, since the OTA GBW and consequently the τ_{col} are hardly limiting the current integration time—as discussed in Section III-B—such tradeoff is not expected to majorly impact the read speed. Indeed, in the 128-stacked example in Fig. 7, the read time only increases by less than 50% once we switch to a single-stage 5T OTA, while the read energy immediately drops by $\sim 10\times$. Further scaling of read energy to ~ 10 pJ/bit level at μ s-scale speed can be accomplished by lowering the full circuit V_{DD} and by decreasing the gm bias voltage V_{cm1} [Fig. 4(a)]. While a lower V_{DD} would understandably affect the virtual-ground accuracy by lowering the A_0 , we estimate the consequences to be minor across a broad range of $A_0 \geq 46$ dB (based on sensitivity studies), and as such the current integration time for reading to be well within 2 μ s (Fig. 7). Still, the projected μ s-level speed would be significantly faster than the 40- μ s read latency reported for TLC 3-D NAND flash [13]. Meanwhile, although single-level cell (SLC) 3-D NAND flash can achieve a comparable read latency (e.g., ~ 4 μ s for SLC [14]; see Fig. 7) and thus remains relevant for read-speed comparison, the inherent Fowler–Nordheim program operation still relies on high voltages (e.g., 12 V; [14]), which inevitably limits its write endurance ($\sim 10^5$; [2]). In contrast, 3-D NDRO FeRAM, thanks to its polarization switching physics for writing, offers low-voltage write operations (sub-3 V in this work) and therefore high endurance (10^9 ; [7]), which positions it more favorably for SCM applications (up to 10^8 ; [2]). Our scheme hence promises scalability in power-performance by properly targeting the energy and speed specifications based on application-driven memory bandwidth requirements.

V. CONCLUSION

This work proposed a class of current-mode-integrator circuit sensing solutions to NDRO 3-D FeRAM-based storage-class memories. Exploiting current-mode gm cell designs, we demonstrated sub- μ s readout in 128-stacked FeCAP arrays achieving an effective bit density of 200 Mb/mm², highlighting the feasibility of overcoming charge-sharing-defined 3-D stacking limit by employing circuit feedback techniques. We further emphasize our unique *time-differential bipolar readout* that guarantees NDRO signal integrity against multiple variability sources, both in CMOS and in FeCAP. The proposed sensing scheme, in addition, offers flexibility in tailoring read speed-energy tradeoff based on application demands and hence scalability for 3-D FeRAM-based SCMs.

REFERENCES

- [1] A. Gholami, Z. Yao, S. Kim, C. Hooper, M. W. Mahoney, and K. Keutzer, “AI and memory wall,” *IEEE Micro*, vol. 44, no. 3, pp. 33–39, May 2024, doi: [10.1109/MM.2024.3373763](https://doi.org/10.1109/MM.2024.3373763).
- [2] S. Legtchenko et al., “Managed-retention memory: A new class of memory for the AI era,” 2025, *arXiv:2501.09605*.
- [3] S. Deng et al., “Comparative advantages of 2T-nC FeRAM in empowering high density 3D ferroelectric capacitor memory,” in *IEDM Tech. Dig.*, Dec. 2023, pp. 1–4, doi: [10.1109/IEDM45741.2023.10413855](https://doi.org/10.1109/IEDM45741.2023.10413855).
- [4] J. Duan et al., “Towards scalable 3D integration of 2T-nC FeRAM with hundreds of layer stacking,” in *IEDM Tech. Dig.*, Dec. 2025, pp. 1–4, doi: [10.1109/IEDM50572.2025.11353513](https://doi.org/10.1109/IEDM50572.2025.11353513).
- [5] M. M. Dahan et al., “Novel ultrafast non-destructive readout of FeRAM by low-voltage transient current,” in *Proc. IEEE Int. Memory Workshop (IMW)*, Monterey, CA, USA, May 2025, pp. 1–4, doi: [10.1109/IMW61990.2025.11026930](https://doi.org/10.1109/IMW61990.2025.11026930).
- [6] N. Ramaswamy et al., “NVDram: A 32Gb dual layer 3D stacked non-volatile ferroelectric memory with near-DRAM performance for demanding AI workloads,” in *IEDM Tech. Dig.*, Dec. 2023, pp. 1–4, doi: [10.1109/IEDM45741.2023.10413848](https://doi.org/10.1109/IEDM45741.2023.10413848).
- [7] S. Mukherjee et al., “Enhanced capacitive memory window by improving remnant polarization in ferroelectric capacitors for non-destructive read,” *IEEE Electron Device Lett.*, vol. 46, no. 4, pp. 576–579, Apr. 2025, doi: [10.1109/LED.2025.3539834](https://doi.org/10.1109/LED.2025.3539834).
- [8] S. A. Siddiqui et al., “FeRAM capacitor with novel low-power, non-destructive and high endurance read operation for high-density embedded memory,” in *Proc. Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Kyoto, Japan, Jun. 2025, pp. 1–3, doi: [10.23919/VLSITECHNOLOGYANDCIR65189.2025.11075099](https://doi.org/10.23919/VLSITECHNOLOGYANDCIR65189.2025.11075099).
- [9] Y. Xiang, S. Mukherjee, G. Hellings, H. Oh, M. Garcia Bardon, and J. Van Houdt, “Compact modeling and design exploration of non-destructive read-out 1T1C FeRAM,” *IEEE Trans. Electron Devices*, vol. 71, no. 8, pp. 4685–4691, Aug. 2024, doi: [10.1109/TED.2024.3418304](https://doi.org/10.1109/TED.2024.3418304).
- [10] M. Kim et al., “A 1Tb 3b/Cell 8th-generation 3D-NAND flash memory with 164MB/s write throughput and a 2.4Gb/s interface,” in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, vol. 65, CA, CA, USA, Feb. 2022, pp. 136–137, doi: [10.1109/ISSCC42614.2022.9731640](https://doi.org/10.1109/ISSCC42614.2022.9731640).
- [11] B. Razavi, “The switched-capacitor integrator [A circuit for all seasons],” *IEEE Solid State Circuits Mag.*, vol. 9, no. 1, pp. 9–11, Winter. 2017, doi: [10.1109/MSSC.2016.2624178](https://doi.org/10.1109/MSSC.2016.2624178).
- [12] GlobalFoundries. *FDX FD-SOI Technology*. Accessed: Sep. 7, 2026. [Online]. Available: <https://gf.com/technology-platforms/fdx-fd-soi/>
- [13] K. Yanagidaira et al., “A 1Tb 3b/cell 3D-flash memory with a 29%-improved-energy-efficiency read operation and 4.8Gb/s power-isolated low-tapped-termination I/Os,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2025, pp. 1–3, doi: [10.1109/ISSCC49661.2025.10904509](https://doi.org/10.1109/ISSCC49661.2025.10904509).
- [14] T. Kouchi et al., “A 128Gb 1-bit/cell 96-word-line-layer 3D flash memory to improve the random read latency with tProg = 75 μ s and tR = 4 μ s,” *IEEE J. Solid-State Circuits*, vol. 56, no. 1, pp. 225–234, Jan. 2021, doi: [10.1109/JSSC.2020.3028393](https://doi.org/10.1109/JSSC.2020.3028393).