Piezoelectric Strain FET (PeFET)-Based Nonvolatile Memories

Niharika Thakuria\textsuperscript{©}, Reena Elangovan, Anand Raghunathan, and Sumeet K. Gupta\textsuperscript{©}, Senior Member, IEEE

Abstract—We propose nonvolatile memory (NVM) designs based on piezoelectric strain FET (PeFET) utilizing piezoelectric/ferroelectric material (PE/FE such as PZT) coupled with 2-D transition metal dichalcogenide (2-D-TMD such as MoS\textsubscript{2})-based transistor. The proposed NVMs store bit information in the form of polarization (\(P\)) of FE/PE, use electric-field driven \(P\)-switching for write, and employ piezoelectricity-induced dynamic bandgap modulation of 2-D-TMD channel for bit sensing. We analyze PeFET with COMSOL-based 3-D modeling, showing that circuit-driven optimization of PeFET geometry is essential to achieve effective strain transduction and adequate bandgap modulation for NVM read. We achieve distinguishability of up to \(11\times\) in binary states of PeFETs. We propose various flavors of PeFET NVMs, namely: 1) high-density (HD) NVM featuring a compact access-transistor-less bit-cell; 2) 1T-1PeFET NVM with segmented architecture, targeted for optimized write energy and latency; and 3) cross-coupled (CC) NVM offering a tradeoff between area and latency. PeFET NVMs offer up to \(7\times\) smaller cell area, 66\% lower write energy, 87\% lower read energy, and 44\% faster read compared to 2-D-FET static random access memory (SRAM). This comes at the cost of high write latency in PeFET NVMs, which can be minimized by virtue of optimized PE geometry.

Index Terms—Ferroelectric (FE), nonvolatile memory, piezoelectric (PE), strain, transition metal dichalcogenides (TMDs).

I. INTRODUCTION

In this era of extreme data processing and storage needs, multicores processors comprising of CMOS static random access memory (SRAM)-based cache are limited in scaling capability and burdened by high static leakage [1]. A memory that is fast, energy-efficient, compact, and has zero standby leakage is recommended for the evolving needs. Emerging nonvolatile memories (NVMs) [2], [3], [4], [5], [6] address concerns of standby leakage. However, most of them [2] rely on current-driven write, which is slow and energy-hungry. An exception to this is FE-based NVMs [3], [4], [5], [6] that employ low-power electric-field (\(E\)) driven write.

FERAM has a compact 1 transistor-1 capacitor cell configuration and high retention but destructive read [3]. FEFET’s offering separate read-write paths overcome the issue of destructive read [4], [6]. FEFET without an interlayer metal (ILM) between FE and the underlying needs high write voltage and suffers from variability, endurance, and retention concerns [4]. FEFET with ILM (FEMFET) reduces write voltage and trap-related issues [5]. However, the floating ILM makes this device vulnerable to gate leakage causing severe bit-sensing challenges [6]. Also, NVM designed with FEMFET requires one or more access transistors [7] hurting its integration density.

Hence, a natural question is whether an NVM can be designed to retain the useful features of FE (e.g., \(E\)-driven write) while alleviating the said concerns. To that effect, we propose a piezoelectric strain FET (PeFET)-based NVM [8]. PeFET features a nondestructive read, and the absence of floating metal in this device eliminates gate leakage concerns [8]. Furthermore, the piezoelectric (PE) layer is controlled by metals on both its sides, alleviating the issues of traps, variability, and retention [8]. Due to its unique structure and operational mechanism, the proposed PeFET NVM can potentially achieve: 1) retention as high as FERAMs (the highest among all FE NVMs); 2) nondestructive read; and 3) elimination of the gate leakage concerns due to the absence of any floating metal layers. However, the benefits of PeFETs are accompanied by certain device limitations [8], which we tackle at the circuit/array level by pursuing a device-circuit co-design approach. With this in view, we propose multiple PeFET-based NVMs. Our contributions to this article are as follows.

1) We present modeling and device characteristics of PeFET and PeFET-based NVMs, namely, high density (HD), 1T-1PeFET, and cross-coupled (CC) cell. HD is targeted for HD memories, while the latter two are for low power write.
2) We conduct an extensive device-circuit co-design to establish the design requirements and area-energy-latency tradeoffs of PeFET NVMs with respect to 2-D-FET/Si SRAMs.

II. BACKGROUND OF PE-BASED DEVICES

Strain-engineered transistors can overcome the performance roadblock of scaled transistors with enhanced current...
(\(I_{\text{ON}}\)), however, at the cost of OFF current (\(I_{\text{OFF}}\)). PE-based strain transistors [9], [13] display enhanced \(I_{\text{ON}}\) at iso-\(I_{\text{OFF}}\) by triggering bandgap modulation when only a nonzero voltage is available to the PE. Jana [9] proposed WiggleFETs that exhibit steep switching based on negative capacitance utilizing dual electrostrictive-PE effect in its gate insulator. In \(\pi\)-FET, a piezolayer in the gate-stack enforces a large vertical compressive strain (which is a limitation) controlled by the gate voltage [10] to achieve a high \(I_{\text{ON}}/I_{\text{OFF}}\) ratio. In piezoelectronic transistor (PET) [11], the resistance of a piezoresistive material is modulated by means of a vertical stress. Electrostrictive field effect transistor (EFET) [12] and molybdenum disulfide (MoS\(_2\))-based piezoelectric FET [13], [17] modulate bandgap with voltage-actuated vertical stress leveraging sensitivity of transition-metal dichalcogenides (2-D-TMDs) to the latter. We extend the idea of piezoelectricity-driven bandgap modulation of 2-D-TMDs to the latter. We extend the idea of piezoelectricity-driven bandgap modulation of 2-D-TMDs to the latter. We extend the idea of piezoelectricity-driven bandgap modulation of 2-D-TMDs to the latter.

Both ferroelectric (FE) and PE properties of PE material are utilized in PeFET. PZT-5H is a PE-suitable material as it possesses: 1) large strain for achieving effective bandgap modulation in 2-D-TMD and 2) sufficiently wide hysteresis of polarization-voltage response for nonvolatility. The latter adds to design complexity of the cited steep-switching-related works of PE-based FETs. There, the primary requirement for PE material (poled in a particular direction) is a large PE coefficient, which can tune the resistance of piezoresistive material [11] or bandgap of Si/2-D-TMD channel [9], [12], [13], [17]. Lead magnesium niobate-lead titanate (PMN-PT) is ideal for such applications, which, however, does not qualify as a preferred PE material for PeFET NVM (that undergoes polarization switching) due to its low coercive field.

Another design aspect of PeFET is that 2-D-TMD in the channel should allow sufficient piezoelectricity-induced bandgap modulation. Various experiments have demonstrated monotonic bandgap reduction due to out-of-plane pressure in 2-D-TMD, e.g., MoS\(_2\). The bandgap coefficient of multilayer MoS\(_2\) for out-of-plane uniaxial stress is \(\approx800\) meV/GPa [14]. EFET [12] hinges on bandgap reduction of multilayer-MoS\(_2\) with piezoelectricity-induced out-of-plane pressure. Peña-Alvarez et al. [15] report a high bandgap coefficient (\(\approx800\) meV/GPa) in monolayer MoS\(_2\). For that reason, we prefer monolayer MoS\(_2\) for PeFET design.

### III. PeFET Design and Modeling

#### A. Device Structure and Operation

PeFET consists of drain (\(D\)), gate (\(G\)), source (\(S\)), and back (\(B\)) terminals [see Fig. 1(a) and (b)]. A PE material (which also exhibits ferroelectricity, e.g., PZT-5H) is deposited between \(G\) and \(B\). The channel is made of 2-D TMD, e.g., monolayer MoS\(_2\). The process flow in PeFET is similar to that in [29].

1. **Write**: Polarization (+/−\(P\)) in PE represents bit 1/0 stored in PeFET. The application of a voltage across \(G\) and \(B\) (\(V_{GB} \)) enables writing of 1/0 to PE [see Fig. 1(b)]. \(V_{GB} > V_C\) (coercive voltage of PE) leads to \(-P \rightarrow +P\) switching in PE, while \(V_{GB} < V_C\) yields \(+P \rightarrow -P\) storage [shown by arrows in Fig. 1(c)]. We present the calibration of \(P - E\) response from simulated PeFET model with that from experiments of PZT-5H [16] for \(V_{GB} = [0.7\ V] \geq |V_C|\).

2. **Read**: In addition to write, we use voltage at \(G\) to stimulate \(P\)-dependent PE behavior in PZT and electrostatically control the 2-D-TMD channel. Read voltage (\(0 < V_R < 0.35\ V < V_C\) applied to \(G\) (keeping \(B = 0\)) generates strain in PE, which is transduced to 2-D TMD for bandgap modulation (see Fig. 2). For \(V_R = 0.35\ V\) in PeFET with \(+P\), positive strain is achieved \(S_{PE} = (\Delta \alpha_{PE}/\rho_{\text{PE}}) > 0\) as shown by experimental strain-electric field butterfly response of PZT-5H [16] in Fig. 2(a). In \(-P\) state, \(V_R = 0.35\ V\) causes \(\Delta S_{PE} < 0\) [see Fig. 2(b)]. Stress develops in PE \(\sigma_{PE}\) due to \(\Delta S_{PE}\), which transduces as pressure in TMD \(\sigma_{TMD}\), the latter leading to dynamic modulation in bandgap \(\Delta \alpha_{E \text{TMD}}\) in (1) of TMD [15]. If \(\sigma_{TMD} > 0\) (due to \(\Delta S_{PE} > 0\) in \(+P\)), \(\Delta \alpha_{E \text{TMD}} < 0\) or bandgap decreases (1). Otherwise, when \(\sigma_{TMD} < 0\) (for \(\sigma_{PE} < 0\) in \(-P\)), \(\Delta \alpha_{E \text{TMD}} > 0\), and bandgap is higher (1). \(\Delta \alpha_{E \text{TMD}} < 0\) in \(+P\) results in drain current \(I_{\text{DS}} = I_{\text{LRS}}\) (low resistance state), while \(\Delta \alpha_{E \text{TMD}} > 0\) in \(-P\) causes \(I_{\text{DS}} = I_{\text{HRS}}\) (high resistance state).

\[
E = E_0 \pm \Delta \alpha_{E \text{TMD}} = E_0 \pm \sigma_{TMD} \rho_{\text{TMD}}
\]

where \(E_0\) is the intrinsic bandgap. Note that electron mobility increases due to positive uniaxial strain on MoS\(_2\) [17], which in this work occurs for \(+P\), enhancing \(I_{\text{LRS}}\).
negative strain reduces mobility, degrading $I_{\text{HRS}}$. With the said improvement of $I_{\text{HRS}}$ and decrease of $I_{\text{LRS}}$, distinguishability ($I_{\text{HRS}}/I_{\text{LRS}}$) improves. Hence, PeFET characteristics presented here with typical mobility ($\mu = 90 \text{ cm}^2/\text{Vs}$ [18]) is a conservative analysis.

**B. Modeling**

We develop a simulation framework for our proposed PeFET (see Fig. 3). We model the $P-E$ response of PZT-5H using Presisch model [19] [see Fig. 3(a)] and calibrate it with experimental data in [16] [see Fig. 1(c)]. $P$-switching delay, $t_{\text{PE}} = 1.8$ ns for PZT-5H [20], is incorporated using resistor–capacitor network [19].

Furthermore, we model pressure transduced from PE to 2-D-TMD ($\sigma_{\text{TMD}}$) using the 3-D structure of PeFET (including hammer and nail effect) in COMSOL multiphysics suite. Our model integrates solid mechanics, electrostatics, and their coupling using the strain-charge form of the constitutive equations for PE [see Fig. 3(a)] to obtain $\sigma_{\text{PE}}$ and $\sigma_{\text{TMD}}$ required for $\Delta E_G$ (1). To efficiently transduce $\sigma_{\text{PE}}$ to $\sigma_{\text{TMD}}$, we utilize hammer and nail effect [11], signified by $\kappa = (A_{\text{TMD}}/A_{\text{PE}}) = (L_{\text{TMD}} W_{\text{TMD}})/L_{\text{PE}} W_{\text{PE}} < 1$, wherein the area of 2-D-TMD above the gate ($A_{\text{TMD}}$) acts as the nail, while PE serves as the hammer [see Fig. 3(b)]. We design $A_{\text{TMD}}$ to be smaller than the area of PE ($A_{\text{PE}}$) by choosing appropriate PE dimensions. $A_{\text{TMD}}$ is fixed by the feature size ($F = L_{\text{TMD}}$) and width ($W_{\text{TMD}}$) of PeFET. We choose $F = 20$ nm and $W_{\text{TMD}} = 30$ nm (minimum width for high integration density). We select metals with high stiffness for the nail, PE, and source/drain contacts (see Fig. 3) and stiff encapsulant (e.g., Al$_2$O$_3$) surrounding the device. The encapsulant provides favorable boundary conditions for stress ($\sigma_{\text{TMD}}$) to be maximized at the TMD channel. We ensure zero displacements in all directions on the top surface of the 2-D TMD, the source/drain contacts of the PeFET, and bottom surface of the back contact. The geometric entity between the bottom contact and TMD is now fully constrained ensuring net zero displacement of the whole structure while allowing piezoelectricity-induced displacement (or strain) in PE to be localized toward the TMD channel. $\sigma_{\text{TMD}}$ is converted to $\Delta E_G$ using the reported bandgap coefficient ($\omega_{\text{TMD}}$) in Table I.

Finally, $\Delta E_G$ is self-consistently coupled with Verilog-A-based 2-D-TMD FET model [21] that reflects dynamically changing bandgap on charge/potential of the 2-D-TMD channel. We use a capacitive network-based model like Stanford 2-D Semiconductor (S2DS) transistor model [21], modifying it for a back-gated 2-D FET. $I_{\text{DS}}$ reflects not only the effect of electrostatics but also that of bandgap modulation in PeFET device characteristics. The 2-D-TMD FET model provides $P$-dependent $I_{\text{DS}}$ of PeFET. The parameters used in the model are based on experiments or reported literature (see Table I).

**C. Device Characteristics**

Our results from COMSOL simulations [Fig. 3(b)] highlight $\sim 12 \times$ increase in $\sigma_{\text{TMD}}$ compared to $\sigma_{\text{PE}}$ at $\kappa = 0.04$ due to the hammer and nail effect (details in next section). To obtain polarization-dependent transfer characteristics ($I_{\text{DS}} - V_{\text{GS}}$) of PeFET [see Fig. 4(a)] with $\kappa = 0.04$ and to avoid $P$-switching, we apply gate voltage ($V_G$) $< V_C$ of PZT-5H ($= 0.54$ V at PE thickness of 600 nm), while keeping back voltage ($V_B$) $= 0$. For comparison, we also simulate a baseline PeFET device that behaves like a MoS$_2$ FET [with $V_{G_B} = 0$ in Fig. 4(a)], which is electrostatically controlled and devoid of PE effect. Our results show at $V_{\text{GS}} = 0.35$ V, $+P$ state in
the PeFET yields $\sigma_{TMD} = 0.64$ GPa and $\Delta E_G = 51$ meV [see Fig. 4(b)] causing 2.3x higher $I_{DS}$ ($I_{HRS}$). While $-P$ results in $3.4x$ lower $I_{DS}$ ($I_{HRS}$) at $V_{GS} = 0.35$ V compared to baseline. Based on $I_{DS} - V_{GS}$ characteristics, we identify that $0.3$ V < $V_{GS}$ < $0.4$ V provides optimal distinguishability ($I_{LRS}/I_{HRS}$), necessary current for read operation, and ample read disturbs margin ($V_{C} - V_{GS} \sim 200$ mV). We choose $V_{GS} = V_{R} = 0.35$ V for $I_{LRS}/I_{HRS} = 8\times$ [see Fig. 4(c)] at $\kappa = 0.04$.

D. $\kappa$ Analysis

Lowering $\kappa$ by tuning PE width ($W_{PE}$), hence $A_{PE}$ [see Fig. 3(b)] boosts stress in TMD (and $\Delta E_G$) due to the hammer and nail effect. By increasing $W_{PE}$ from 90 to 180 nm, $\kappa$ decreases from 0.07 to 0.03, leading to 1.78x increase in $\sigma_{TMD}$ and $\Delta E_G$ [see Fig. 4(b)]. Hence, for $+P$, more aggressive bandgap reduction is observed at small $\kappa$, increasing $I_{LRS}$. Whereas in $-P$, bandgap increases with a decrease in $\kappa$, which weakens $I_{HRS}$. Strong $I_{LRS}$ combined with weak $I_{HRS}$ increases $I_{LRS}/I_{HRS}$ and hence distinguishability from 3x to 11 as $\kappa$ is reduced from 0.07 to 0.03 [see Fig. 4(c)].

IV. PEFET-BASED NVM DESIGNS

We now propose three flavors of PeFET-based NVMs, viz., HD, 1T-1PeFET, and CC cells.

A. HD NVM (see Fig. 5)

HD NVM is designed with 1-PeFET, without an access transistor [see Fig. 5(a) and (b)]. $G$ of PeFET is connected to word line (WL) that is shared along a row. $B$ and $D$ are connected to write bitline (WBL) and read bitline (RBL), respectively [see Fig. 5(c)], and are shared along a column. $B$, being shared by adjacent cells in a column [evident from cell boundary in Fig. 5(b)], helps in minimizing area. Next, we discuss write and read operations.

1) Write: WL and WBL form a cross-point-like write port with $C_{PE}$ between them [see Fig. 5(c)]. To write, WBL and WL of the accessed cells are asserted, such that $|V_{GB}|$ across PE is $V_{DD} > V_{C}$. We implement a two-phase ($\phi_{12}$) operation [see Fig. 5(d)]. To write $+1$/$+P$, we drive WBL to $-V_{DD}/2$. Then, WL is driven to $V_{DD}/2$ in $\Phi_1$ and switch to $-V_{DD}/2$ in $\Phi_2$. Consequently, $V_{GB} = V_{DD} = 0.7$ V in $\Phi_1$ causes PE to switch to $+P$ (“1”), with no effect on $P$ in $\Phi_2$ since $V_{GB} = 0$ [see Fig. 5(d)]. Similarly, “0”/$-P$ is written by driving WBL to $V_{DD}/2$ and switching WL to $V_{DD}/2$ in $\phi_1$ and $-V_{DD}/2$ in $\phi_2$. $V_{GB} = -V_{DD} = -0.7$ V triggers $-P$ switching in $\Phi_2$. RBL is kept at 0 V during write.

2) Read: $V_{R}$ ($= V_{DD}/2 = 0.35 < V_{C}$) on WL, keeping WBL at 0 V. Moreover, $V_{DD}$ is supplied to the RBL of accessed cells, and read current ($I_{LRS}/I_{HRS}$) is sensed on RBL in a single phase, depending on the polarization state ($+P/-P$). Half-accessed rows have WL at 0 V, while in half-accessed columns, RBL voltage is 0 V. Hence, current flow through half-accessed/unaccessed cells is negligible or zero.

The lack of AX in HD cells and sharing of back contact in the layout result in a compact design. However, the absence of AX exposes the PE capacitance ($C_{PE}$) of half-accessed cells directly on WBL. Since $C_{PE}$ is typically large (dielectric constant $\sim 4000\epsilon_0$ [16]), write energy efficiency is adversely affected. Based on our analysis, we attribute that 78% of write energy originated from charging/discharging of $C_{PE}$ followed by 12% due to metal capacitance switching of BLs/WL and 10% due to $P$-switching. To minimize the effect of $C_{PE}$ on write energy, we propose PeFET NVMs with AX that isolates WBL from $C_{PE}$.

B. 1T-1PeFET NVM (see Fig. 6)

1T-1PeFET consists of n-type 2-D FET connected to back contact ($B$) of PeFET [see Fig. 6(a)]. Two-dimensional FET AX compactly abuts PeFET than with a CMOS AX, the latter leading to potential area penalty due to additional design rules. We adopt a CMOS-2-D materials-based hybrid architecture

Biasing WBLs/WLs of accessed cells with $\pm V_{DD}/2$ for write allows WBLs/WLs of half-accessed/unaccessed cells to be kept at 0 V [see Fig. 5(e)]. This is because half-accessed cells still experience $|V_{GB}| \leq V_{DD}/2 < V_{C}$ and unaccessed cells have $V_{GB} = 0$, averting write disturbs. Having WBL/WL of half-accessed/unaccessed at 0 V cuts down write energy that may be present in other cross-point architectures based on $V_{DD}/2$ or $V_{DD}/3$ biasing for half/unaccessed cells [23]. In the absence of access transistors (AX) in PeFET, the use of negative voltages for write does not incur large energy overheads, whereas designs with AX do [7] since in the latter, WLs of all unaccessed rows are driven to a negative voltage to avert turning them ON. For the same reason, in our other PeFET designs with AX, we constrain the write biasing voltage to be positive (details later).

2) Read: We apply $V_{R}$ ($= V_{DD}/2 = 0.35 < V_{C}$) on WL, keeping WBL at 0 V. Moreover, $V_{DD}$ is supplied to the RBL of accessed cells, and read current ($I_{LRS}/I_{HRS}$) is sensed on RBL in a single phase, depending on the polarization state ($+P/-P$). Half-accessed rows have WL at 0 V, while in half-accessed columns, RBL voltage is 0 V. Hence, current flow through half-accessed/unaccessed cells is negligible or zero.

The lack of AX in HD cells and sharing of back contact in the layout result in a compact design. However, the absence of AX exposes the PE capacitance ($C_{PE}$) of half-accessed cells directly on WBL. Since $C_{PE}$ is typically large (dielectric constant $\sim 4000\epsilon_0$ [16]), write energy efficiency is adversely affected. Based on our analysis, we attribute that 78% of write energy originated from charging/discharging of $C_{PE}$, followed by 12% due to metal capacitance switching of BLs/WL and 10% due to $P$-switching. To minimize the effect of $C_{PE}$ on write energy, we propose PeFET NVMs with AX that isolates WBL from $C_{PE}$.
(also in CC NVM array), where peripherals are designed with Si-based CMOS and core memory array (PeFETs and AX) with 2-D material. Unlike in HD NVM, large $C_{PE}$ on WBL is replaced by smaller drain capacitance from AX in a column. We explore two variants of layout for this cell: 1) tall layout for area optimization [see Fig. 6(b)] and 2) wide layout where cell height is minimized [see Fig. 6(c)]. The latter is motivated by our understanding that bitline capacitance is dominating contributor for delay and energy in 1T-1PeFET arrays. For 1T-1PeFET array that is devoid of segmentation (segmentation is discussed next), 79% of total write energy is attributed to $|V_{DD/2}|$ switching of WBLs in half-accessed cells. WBLs of half-accessed cells must be driven to $V_{DD/2}$ to ensure $|V_{GB}| < V_{C}$ across their PEs to prevent $P$-switching, which causes a large overhead in write energy. To mitigate this, we present segmented 1T-1PeFET NVM.

### C. Segmented Architecture of 1T-1PeFET NVM

1T-1PeFET segmented architecture [see Fig. 6(d)] that eliminates energy overheads from half-accessed/unaccessed cells is motivated by FERAM architecture [24]. The 1T-1PeFET segmented array is divided into multiple segments each comprising of 64 columns that correspond to the size of a word ($N_W$) and $N_R$ rows. If the entire array is constituted of $N_C$ columns and $N_R$ rows, there are $N (= N_C/N_W)$ segments, each of size $N_W \times N_R$. Moreover, global plate line (GPL) is responsible for the said segmentation. The array contains $N = N_C/N_W$ GPLs. GPLs provide input to $N_R$ buffers in each segment. These buffers act as gates for a segment and their output drives a local plate line (LPL). LPL connects the 64 ($= N_W$) cells. Note that LPL is loaded with a smaller net $C_{PE}$ from 64 cells instead of an entire row (as in a standard architecture). The WL, which runs through a row, provides supply voltage to the buffers and activates the AXs of 1T-1PeFET NVMs in the accessed row. In addition, each column in a segment has its own WBL and RBL. Each segment comprises of: 1) GPL; 2) $N_R$ LPLs; 3) $N_R$ WLs shared by other segments; 4) $N_W$ WBLs; and 5) $N_W$ RBLs.

1) **Write**: Two phase signals, i.e., $V_{DD} (\Phi_1) \rightarrow 0 (\Phi_2)$, are applied to GPL [see Fig. 6(e)] of the accessed segment during write. Let the first word of first segment [accessed word in segment [0] of Fig. 6(d)] is to be written to. WL is asserted with a boosted voltage, $V_{bst} = V_{DD} + V_{TH}$, to mitigate $V_{TH}$ loss in AX. WL provides supply voltage to activate buffer [0], which allows LPL of the accessed word (LPL[0]) to latch to $V_{bst}$ in $\Phi_1$ and 0 in $\Phi_2$ corresponding to $GPL = V_{DD} (\Phi_1) \rightarrow 0 (\Phi_2)$. Write input data $V_{DD}$ or 0 V is provided to the WBLs of $N_W$ cells to write $P$ or $P$ respectively [see Fig. 6(e)]. For polarization switching, we ensure $V_{GB} > V_C$ and $V_{GB} = V_{LPL} - V_{WBL}$. With this background, we now elaborate on $P \rightarrow P$ switching. For that, WBL is provided $V_{DD}$. This results in $V_{GB} = -V_{DD}$ in $\Phi_2$ [see Fig. 6(e)], with the polarization kept intact by $V_{GB} < V_C$ in $\Phi_1$. Hence, $P$ state (logic 1) is written in $\Phi_2$. Similarly, if $WBL = 0$ V, $V_{GB} = -V_{bst}$ in $\Phi_1$ and +$P$ state is being written [see Fig. 6(e)]. We use $V_{DD} = 0.7$ V and $V_{bst} = 1$ V [see Fig. 6(e)]. All RBLs along with WBLs and GPLs of unaccessed words/segments (e.g., GPL [1] to GPL [N − 1]) are kept at 0 V. Hence, their LPLs are also at 0 V. Thus, energy expended by these segments is minimized in this architecture.

2) **Read**: We accomplish read by driving GPL of accessed segment and WL to $V_R = 0.35$ V while WBLs = 0 V. This results in $V_R$ on LPL [e.g., LPL[0] [0] in Fig. 6(d)]. Read current is sensed with RBL voltage at $V_{DD}$. The voltages of GPL, WBL, and RBL of unaccessed sections and WL of unaccessed rows are 0 V.

### D. CC PeFET NVM (Fig. 7)

By applying segmentation in 1T-1PeFET, we focused and minimized the interaction of $C_{PE}$. However, a limitation of 1T-1PeFET is the area overhead from AX. Also, unlike in an HD cell, B is an internal node that cannot be shared, constraining its compactness as per design rules. Moreover, although reduced considerably, $C_{PE}$ still needs to be switching during read. To alleviate $C_{PE}$ during read while also targeting cell area reduction compared to 1T-1PeFET cells, we propose CC PeFET NVM cells [Fig. 7(a) and (b)]. The technique of cross-coupling two PeFETs in this cell allows us to share the back contact (along a row), conserving cell area. Our layout analysis [see Fig. 7(b)] suggests that area per bit in CC is up to $1.66 \times$ lower than 1T-1PeFET (wide cell) at $k = 0.04$. Moreover, the back contact (hence $C_{PE}$) need not be switched in CC NVM during read [$LPL = 0$ V in Fig. 7(c)], improving read latency/energy.

We design CC cells [see Fig. 7(a)] with two PeFETs (A and B) that are cross coupled by connecting $D$ of A ($D_A$) to $G$ of B ($G_B$) and vice versa. $D_A$ and $D_B$ are connected to bitlines BL$_1$ and BL$_2$, respectively, through AX$_1$ and AX$_2$. $G$ of AX$_1$/AX$_2$ is controlled by WL. The back contacts of A and B are driven by PL. Bits in A and B can be stored, read, and written individually. Hence, each CC cell is equivalent to a 2-bit NVM.

### E. Segmented Architecture of CC NVM

We propose segmentation in the CC NVM array [see Fig. 7(c)] like a 1T-1PeFET array, so that PL has reduced $C_{PE}$ load during write. Each segment comprises of: 1) $N_R$ rows; 2) 32 columns of CC cells that effectively act as 64 bits of a word ($N_W$); 3) a GPL that provides input to $N_R$ buffers, that are in turn activated by a WL; and 4) $N_R$ LPLs driven by the output of buffers [see Fig. 7(c)]. Notice that PL (back contact) is replaced by LPL, which is shared horizontally by 32 CC cells within a segment. One key difference between 1T-1PeFET and CC segmented array is that LPL needs to be switched to $V_R$ during read in the former, while it is 0 V in CC NVM. This eliminates the effect of $C_{PE}$ during read for improved read latency and energy efficiency. Next, we explain the write and read operations in a segmented CC array.

1) **Write**: Here, we explain various cases of writing to A and B.

a) $(+P, -P)$ or $(-P, +P)$: When we intend to store complementary states to A and B, i.e., $+P$ and $-P$ to A and $-P$ or $+P$ to B, BL$_1$ and BL$_2$ are driven to 0 V (or $V_{DD}$) and $V_{DD}$ (or 0 V), respectively [see Fig. 7(d)].
(φ₂) is applied to the GPL of a selected segment and WL is asserted with V_bias = V_DD + V_TH. LPL receives 0 → V_DD as an output. Moreover, the voltage at BL₁ and BL₂ is passed to G₂ and G₄, respectively, by CC NVM’s AXs after WL assertion. For PeFET A, V_GB of A (V_GB,A) = V_BL₂ − V_LPL. Similarly, V_GB,B is V_BL₁ − V_LPL. Say, V_BL₁ = 0, V_BL₂ = V_DD, and LPL is signaled with 0 (φ₁) → V_DD (φ₂); then, we get V_GB,A = V_BL₂ − V_LPL = V_DD in φ₁; +P is being written to A. For B, V_GB,B = V_BL₁ − V_PL = −V_DD during φ₂ which is when −P is written.

b) (−P, −P): To store −P in both A and B, BL₁ and BL₂ are maintained at 0 V. After WL is asserted and 0 → V_DD appears on LPL (like before), V_GB,A/V_GB,B = −V_DD is obtained during φ₂ switching both PEs to −P [see Fig. 7(d)].

c) (+P, +P): To write +P to both A and B, BL₁ and BL₂ are driven to V_DD. In this context, an important aspect needs to be highlighted. When WL is asserted, voltage from BL₁ and BL₂ (which are at V_DD) charges Dₐ (and Gₙ) and Dₜ (and G_A). As a result, PeFETs A and B are turned on, while write ensues. This causes Dₐ and Dₜ (and connected Gₙ and G_A) to drop to V_DD − ΔV [see Fig. 7(d)]. Reduced gate voltage implies lower V_GB across PEs of both PeFETs, which ultimately increase write latency.

F. Read

Currents through BL₁ and BL₂ are sensed after driving them to V_R. WL is asserted with V_DD, so that both cells get V_G = V_R required for read. GPL and hence LPL of the accessed segment are at 0 V during read. The idea is to have V_GB = ~V_R for both A and B. Depending on whether +P or −P is stored in A (or B), I_RRS₁ or I_HRS₁ is sensed on BL₁ (or BL₂). As in write, cross-coupling has effects during read which we elaborate on next.

1) Let, A/B store −P/−P. Since both cells are in HRS, Dₐ and Dₜ in Fig. 7(c) are at V_R. Let currents sensed on BL₁/BL₂ be I_RRS₀₀.

2) Next, let −P/+P be the stored states of A/B, respectively. As A is in HRS, Dₐ is pulled to V_R as desired. On the other hand, Dₜ is pulled down to V_R − ΔV by the action of LRS in cell B. Gₙ and G_B have similar voltage as Dₜ and Dₐ due to a CC connection. So, cell A gets a V_GB,A = V_R − ΔV < V_R. This lowered gate voltage makes A (storing −P) more resistive than

HRS in the −P/−P state (previous example), and now, we obtain I_RRS₁₁ on BL₁, which is lesser than I_HRS₀₀ in (a). For B, Gₙ is V_R, and we obtain I_LRS₁₁ (with V_GB = V_R and V_DS = V_R − ΔV) on BL₂. If A/B has +P/−P instead, we obtain I_HRS₁₀ on BL₁ and I_LRS₁₀ on BL₂.

3) When +P is stored in both PeFETs (i.e., they are in LRS), their drain and connected gate charge to V_R − ΔV. Hence, V_GB,A/B = V_DS,A/B = V_R − ΔV. Diminished V_GB allows a lower current (let it be called I_LRS₁₁) than I_RRS₁₁ or I_LRS₀₁ on BL₁ and BL₂ in this LRS state. Finally, we conclude that I_RRS₁₁ < I_LRS₁₀ and I_HRS₀₁ > I_HRS₁₀ or I_HRS₀₁. We define distinguishability for CC NVM, considering the worst case, as the ratio of I_LRS₁₁/I_HRS₀₀, which is lower than the other PeFET NVMs. Note that in HD and 1T-1PeFET, distinguishability is identical to I_LRS₁₁/I_HRS₁₀. While CC NVM exhibits overheads and complex read/write due to cross-coupling, its advantage (as noted earlier) is effective contact sharing leading lower bit area than 1T-1PeFET. Next, we quantify the area, performance, and energy efficiency of the proposed PeFET NVMs and compare them to 2-D-FET SRAMs.

V. ANALYSIS OF PEFET-BASED NVM ARRAYS (κ = 0.04)

A. Layout and Area [see Fig. 8(a)]

We use λ (= gate length/2)-based rules and gate/metal pitch defined by Intel 20-nm process [25] for the layout analysis of PeFET NVMs. In HD NVM, back terminal is shared among adjacent cells in a column (WBL) resulting in a cell height (H) of one poly-pitch (PP = 9λ). Cell width (W = 18λ) is controlled by width of PE (dictated by κ as

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
discussed in Section III). Finally, an HD cell’s area is $162\lambda^2$. We present tall [see Fig. 8(a)] and wide [see Fig. 8(a)] layouts for 1T-1PeFET NVM. In the tall variant, we minimize cell area by stacking AX along $H$. WBL contact (or drain of AX) is shared with adjacent cells in the column. However, the sharing of RBL contact is restricted by the internally positioned back terminal. Hence, $H$ of 1T-1PeFET tall cell is equal to 2.5 PP (22.5\lambda) or 2.5x larger footprint than HD. Moreover, we propose a wide layout (1T-1PeFET wide NVM) with decreased $H = 2\text{ PP } (18\lambda)$ to minimize BL capacitance for improved energy/latency. The $\kappa$-dependent width of the back contact and access transistor width including contacts sets $W$. Its area shows an overhead of 3.1x over HD. The area of CC NVM is 1.4x and 1.7x lower than 1T-1PeFET tall and wide NVM, respectively, while it is 1.8x larger than HD cell. The improvement of CC NVM area over 1T-1PeFET is achieved by efficient contact sharing: 1) of back contact across PL (along rows unlike in other cells) and 2) of BL with adjacent cells (compare that with limited contact sharing in 1T-1PeFETs). HD, CC, 1T-1PeFET tall, and wide NVMs are 4.7x, 2.5x, 1.87x, and 1.53x area efficient compared to 2-D-FET SRAM.

B. Distinguishability [see Fig. 8(b)]
HD and 1T-1PeFET exhibit the same distinguishability ($I_{LRS}/I_{HRS} = 8\times$) due to similar strain transduced at a fixed $\kappa$. In CC NVM, lower distinguishability of 3x is observed at the same $\kappa$ due to limiting read condition, i.e., $I_{LRS10}/I_{HRS10}$ [refer Section IV(E2c)].

C. Write Energy and Latency [see Fig. 8(c)]
In this subsection and next (Section V-C and V-D), we present energy/latency of PeFET-based NVMs for a 512 x 512 (32 k拜) array. Note that the results are based solely on the respective memory arrays (excluding peripherals) to highlight the intrinsic merits and drawbacks of the proposed circuits.

HD NVM shows degraded write energy and latency than other designs due to $C_{PE}$ from half-accessed cells adding to WBL and PL (refer to Section IV). In segmented 1T-1PeFET NVM, AX shields WBL from this capacitance leading to energy and latency improvements over HD. The 1T-1PeFET wide design shows minimum latency due to its layout-enabled minimization of BL capacitance. CC NVM combines the benefits of AX (like in 1T-1PeFET) with a smaller area for further write energy optimization. However, it shows higher latency than the 1T-1PeFETs despite its smaller cell area due to $P$-switching occurring at diminished $V_{GS}$ (described in Section IV). Finally, the write energy of HD, segmented 1T-1PeFET tall, wide, and CC NVM is 48%, 56%, 61%, and 65% lower than 2-D-FET SRAM. Note that these benefits are reported considering 30% utilization [26] for L2 cache. The baseline SRAM cache leaks for the remaining 70% utilization adding leakage energy, whereas PeFET NVMs do not. Based on our HSPICE simulation of 2-D-FET-based SRAM, standby leakage energy per SRAM cell is 5.35E-20 J. The large size of SRAM cells is the main contributor for write energy along with nonnegligible leakage energy. We observe 22%, 28%, 31%, and 134% overhead in latency in 1T-1PeFET wide, tall, CC, and HD NVMs (caused by $P$-switching) compared to 2-D-FET SRAM.

D. Read Energy and Latency [see Fig. 8(c)]
Among PeFET designs, CC shows minimum read energy as it is compact, and PL connected to $C_{PE}$ is not activated during read. HD NVM shows improved read energy than 1T-1PeFET NVMs due to smaller RBL capacitance (due to lower layout height) compared to the other two designs. Since RBL switching energy is a dominant component of read energy, its reduction affects read energy similarly. We observe lower read energy in PeFET designs than 2-D-FET SRAM attributed to: 1) negligible leakage and 2) capability to selectively turn off RBL discharge of half-accessed cells during read in PeFET arrays. Note that we access 64 bits during read/write operations for both PeFET and SRAM-based arrays for consistency of comparison. The read energy of CC, HD, 1T-1PeFET wide, and tall shows 87%, 85%, 77%, and 74% improvement, respectively, compared to 2-D-FET SRAM at 30% cache utilization. The impact of $C_{PE}$ during read causes latency to increase in HD, 1T-1PeFET tall, and 1T-1PeFET wide by 7%, 93%, and 100% compared to 2-D-FET SRAM. Contrarily, the CC cell shows 30% improvement since $C_{PE}$ is bypassed during read (by keeping LPL to 0 V).

VI. $\kappa$ Analysis of PeFET-Based NVM Arrays
We have established in Section III that small $\kappa$ improves strain transduction to 2-D-TMD channel at the cost of device footprint. Here, we discuss the effect of $\kappa$ in NVM array design.
A. Area [see Fig. 9(a)]

Increasing $\kappa$ leads to a decrease in the cell area of PeFET NVMs. Area improvement for PeFET NVM HD over 2-D-FET SRAM ranges from $4\times$ to $7\times$ at $\kappa = 0.03$–0.07. The size of CC cell is $2.15\times -3.5\times$ lower than 2-D-FET SRAM for $\kappa = 0.03$–0.07. The benefit from 1T-1PeFET (tall) is $1.76\times -2.8\times$ for $\kappa = 0.03$–0.07.

B. Distinguishability [see Fig. 9(b)]

The efficiency of strain transduction increases at a lower $\kappa$, boosting distinguishability. For $\kappa = 0.03/0.07$, it is $11\times/3\times$, respectively, in HD, 1T-1PeFET tall, and wide designs. In CC, it lies in the range of $1.5\times -5\times$ for $\kappa = 0.07$–0.03.

C. Write Energy and Latency [see Fig. 9(c) and (d)]

Write energy and latency of PeFET NVMs improves at higher $\kappa$ (lower PeFET footprint). Large $\kappa$ decreases: 1) WL metal capacitance and 2) FE/PE switching capacitance in all PeFET NVMs. In HD, it also reduces $C_{PE}$ of half-accessed cells, which adds to WBL/WL. Hence, HD exhibits a steeper write energy and latency compared to 1T-1PeFET NVMs. In HD, it also reduces metal capacitance and 2) FE/PE switching capacitance in all PeFET NVMs. Area improvement for PeFET NVM HD over 2-D-FET SRAM ranges from $4\times$–$7\times$.

D. Read Energy and Latency [see Fig. 9(e) and (f)]

The read energy of all PeFET designs is nearly insensitive to $\kappa$. This is because read energy is dominated by BL capacitance switching energy, which is unaffected by $\kappa$ [see Fig. 9(e)]. Note by modifying $\kappa$ (or horizontal dimension), we do not alter layout height or RBL capacitance. Read latency decreases as $\kappa$ increases due to smaller cell size and reduced $C_{PE}$ associated with WL/PL in HD and 1T-1PeFET designs. Improvement in HD and CC is $40\%$ and $44\%$, respectively, for $\kappa = 0.07$, whereas the overhead in 1T-1PeFET designs decreases to $87\%$.

VII. BENCHMARKING OF PEFET NVMs

In Table II, we benchmark our results with 6T-SRAM simulated using 22-nm planar Si-FET and data from literature for 7-nm FinFET-based 6T-SRAM [27]. To be consistent with [27], we include contribution of peripherals in Table II. The peripherals that are active during write are row decoder, row driver, column decoder, and column driver with additional precharge circuitry and sense-amplifier being active during read. Note that we excluded contribution of peripherals in Sections V and VII to bring out intrinsic merits and drawbacks of PeFET-based NVMs. PeFET NVMs show improvements in cell area, read/write energy compared to Si/2-D-FET SRAMs. Improvement in write/read energy is achieved due to low leakage in PeFETs, with tradeoff in latency due to interaction of PE capacitance in PeFETs during read/write. Therefore, our PeFET designs are attractive for systems with low energy/power budget compared to Si/2-D-FET SRAMs.

VIII. FUTURE OUTLOOK

We propose PeFET NVMs focusing on stand-alone memory process flow. We envision further optimization of PeFET-based NVMs with other CMOS compatible PE/FE materials that enable integration of PeFETs into CMOS flow. Si-doped HfO$_2$ is a promising example as experiments have demonstrated reasonable PE effects in them [28]. Besides good CMOS compatibility, Si-doped HfO$_2$ ensures improved PE thickness scalability, enhanced performance, and energy efficiency of PeFET NVMs. However, the information of certain parameters such as elastic tensor components of Si-HfO$_2$ is currently limited, which requires further characterization. Although this work leverages 20-nm node, PeFETs can be further scaled due to excellent gate length and effective oxide thickness (EOT) scaling of TMDs [22]. However, requirement for desirable $\kappa$ ($= \lambda_{TMD}/A_{PE}$) needs to be kept in mind while scaling device width. Despite this constraint, PeFET NVMs are up to $4.3\times$ compact than a Si-based 6T-SRAM. Three-dimensional-stacked PeFET architectures [30] provide further direction for scaling.

IX. CONCLUSION

We propose PeFET utilizing polarization-based bit storage in FE material, electric field-driven write, and piezoelectricity-induced dynamic bandgap modulation of 2-D-TMD channel for bit sensing. We present HD, 1T-1PeFET tall, wide, and CC NVMs based on PeFETs. We analyze them at the array level applying device and circuit optimizations for improved area-energy-latency tradeoffs. HD surpasses other PeFET NVMs in compactness with up to $7\times$ smaller area than 2-D-FET SRAM. PeFET NVMs with other CMOS compatible PE/FE materials that enable integration of PeFETs into CMOS flow. Si-doped HfO$_2$ is a promising example as experiments have demonstrated reasonable PE effects in them [28]. Besides good CMOS compatibility, Si-doped HfO$_2$ ensures improved PE thickness scalability, enhanced performance, and energy efficiency of PeFET NVMs. However, the information of certain parameters such as elastic tensor components of Si-HfO$_2$ is currently limited, which requires further characterization.
REFERENCES

[1] P. Gepner and M. F. Kowalik, “Multi-core processors: New way to achieve high system performance,” in Proc. Int. Symp. Parallel Comput. Electr. Eng. (PARELEC), 2006, pp. 9–13.

[2] G. Sun, J. Zhao, M. Poremba, C. Xu, and Y. Xie, “Memory that never forgets: Emerging nonvolatile memory and the implication for architecture design,” Nat. Sci. Rev., vol. 5, no. 4, pp. 577–592, Jul. 2018.

[3] J.-H. Kim et al., “A highly reliable FRAM (ferroelectric random access memory),” in Proc. 45th Annu. IEEE Int. Rel. Phys. Symp., Apr. 2007, pp. 554–557.

[4] J. Muñoz et al., “Ferroelectricity in HfO2 enables nonvolatile data storage in 28 nm HKMG,” in Symp. VLSI Technol. Dig. Tech. Papers, Jun. 2012, pp. 25–26.

[5] K. Ni et al., “SoC logic compatible multi-bit FeMFET weight cell for neuromorphic applications,” in IEDM Tech. Dig., Dec. 2018, pp. 13.2.1–13.2.4.

[6] S. K. Thirumal and S. K. Gupta, “Gate leakage in non-volatile ferroelectric transistors: Device-circuit implications,” in Proc. 76th Design Res. Conf. (DRC), Jun. 2018, pp. 1–2.

[7] X. Li et al., “Design of 2T/cell and 3T/cell nonvolatile memories,” IEEE Trans. Electron Devices, vol. 66, no. 10, pp. 4216–4221, Oct. 2020, doi: 10.1109/TED.2020.3022344.

[8] A. Shafaei, H. Afzali-Kusha, and M. Pedram, “Minimizing the energy-delay product of SRAM arrays using a device-circuit-architecture co-optimization framework,” in Proc. 49th Annu. Design Autom. Conf., San Francisco, CA, USA, Jun. 2012, pp. 492–497.

[9] A. Saha and S. K. Gupta, “Modeling and comparative analysis of hysteretic ferroelectric and anti-ferroelectric FETs,” in Proc. 76th Device Res. Conf. (DRC), Santa Barbara, CA, USA, Jun. 2018, pp. 1–2, doi: 10.1109/DRC.2018.8442136.

[10] P. K. Larsen, G. L. M. Kampichler, M. J. E. Ulenaers, G. A. C. M. Spierings, and R. Cuppens, “Nanosecond switching of thin ferroelectric films,” Appl. Phys. Lett., vol. 59, no. 5, pp. 611–613, 1991, doi: 10.1063/1.105402.

[11] S. V. Suryawanshi and E. Pop, “S2DS: Physics-based compact model for circuit simulation of two-dimensional semiconductor devices including non-idealities,” J. Appl. Phys., vol. 120, no. 22, Dec. 2016, Art. no. 224503, doi: 10.1063/1.4971404.

[12] D. S. Schulman, A. J. Arnold, and S. Das, “Contact engineering for 2D materials and devices,” Chem. Sci. Rev., vol. 47, no. 9, pp. 3057–3058, 2018.

[13] X. Peng, R. Madler, P. Y. Chen, and S. Yu, “Cross-point memory design challenges and survey of selector device characteristics,” J. Comput. Electron., vol. 16, no. 4, pp. 1167–1174, 2017, doi: 10.1007/s10825-017-1062-z.

[14] J. T. Rickes et al., “A novel sense-amplifier and plate-line architecture for ferroelectric memories,” Integr. Ferroelectr., vol. 48, no. 1, pp. 109–118, Jan. 2002, doi: 10.1080/713718311.

[15] Accessed: May 5, 2023. [Online]. Available: https://en.wikipedia.org/wiki/7_nm_lithography_process

[16] S. P. Park, S. Gupta, N. Mojumder, A. Raghunathan, and K. Roy, “Future cache design using STT MRAMs for improved energy efficiency: Devices, circuits and architecture,” in Proc. 49th Annu. Design Autom. Conf., San Francisco, CA, USA, Jun. 2012, pp. 492–497.

[17] A. Shafaei, H. Afzali-Kusha, and M. Pedram, “Minimizing the energy-delay product of SRAM arrays using a device-circuit-architecture co-optimization framework,” in Proc. 53rd Annu. Design Autom. Conf. (DAC), Austin, TX, USA, Jun. 2016, pp. 1–6, doi: 10.1145/2897937.2898044.

[18] S. Kirbach, K. Kühnel, and W. Weinreich, “Piezoelectric hafnium oxide thin films for energy-harvesting applications,” in Proc. IEEE 18th Int. Conf. Nanoelectron. (IEEE-NANO), Jul. 2018, pp. 1–4, doi: 10.1109/NANO.2018.8662675.

[19] N. Thakuria, D. Schulman, S. Das, and S. K. Gupta, “2-D strain FET (2D-SFET) based SRAMs—Part I: Device-circuit interactions,” IEEE Trans. Electron Devices, vol. 67, no. 11, pp. 4866–4874, Nov. 2020, doi: 10.1109/TED.2020.3022344.

[20] V. P.-H. Hu et al., “Energy-efficient monolithic 3-D SRAM cell with BEOL MoS2 FETs for SoC scaling,” IEEE Trans. Electron Devices, vol. 67, no. 10, pp. 4216–4221, Oct. 2020, doi: 10.1109/TED.2020.3018099.