Understanding the Memory Window of Ferroelectric FET and Demonstration of 4.8-V Memory Window With 20-nm HfO_2

Yixin Qin, *Graduate Student Member, IEEE*, Zijian Zhao, *Graduate Student Member, IEEE*, Suhwan Lim, Kijoon Kim, Kwangsoo Kim, Wanki Kim, Daewon Ha, Vijaykrishnan Narayanan, *Fellow, IEEE*, and Kai Ni, *Member, IEEE*

***Abstract*— In this work, with the goal of developing a large memory window (MW) ferroelectric field-effect transistor (FeFET) for high-density stand-alone storage applications, we provide a deep look into the MW of a FeFET and clarify the definition on the MW through combined theoretical and experimental studies. We show that the following hold: 1) conventionally thought maximum MW of a FeFET (i.e., $2E_C t_{FE}$) is accurate only for dc sweep and may not be accurate for pulsed memory operation as it neglects the contribution from other layers in the gate-stack; 2) for intrinsic FeFET operation that only depends on polarization switching, adding a dielectric layer into a FeFET gate-stack at best keeps the same MW as the one without the dielectric layer if assuming the same polarization is switched in both cases or worse as most likely less polarization can be switched; 3) by integrating FeFETs with or without a top dielectric Al_2O_3 layer, we demonstrate from experiment that adding a dielectric layer reduces the MW; and 4) with a 20-nm-thick HfO_2, a large MW of 4.8 V is demonstrated, which can accommodate tightly distributed multilevel cell (MLC) and triple-level cell (TLC), showing the potential of FeFET as high-density storage technology.**

***Index Terms*— Ferroelectric field-effect transistor (FeFET), HZO, memory window (MW), multilevel cell (MLC), triple-level cell (TLC).**

I. Introduction

NOWADAYS, with the ubiquitous deployment of smart devices, data are being generated at an unprecedented rate, calling for tremendous amount of data storage and data processing. This surge in data generation has created an urgent need for advanced data storage solutions that can keep pace with the relentless accumulation of digital information. In response to this challenge, vertical NAND flash technology has emerged as the backbone of modern data storage systems [1], [2]. With its exceptional performance and continued scaling in the capacity, NAND flash has revolutionized the storage landscape. However, the insatiable demand for data storage capacity continues to grow unabated, prompting the need for further innovations in storage technology. Luckily, the highly scalable nature of vertical NAND flash memory offers a ray of hope in this data storage conundrum. Through a combination of geometrical scaling and functional scaling, NAND flash technology has shown that it can evolve to meet our ever-expanding storage needs in the near future [3], [4]. Geometrical scaling involves stacking more word lines (WLs) vertically, while functional scaling is achieved by packing more bits into a single memory cell [5]. These advancements, although promising, are not without their challenges. One of the most significant hurdles faced by NAND flash memory technology is its suboptimal write performance. This inefficiency stems from the write mechanism itself, where electrons must tunnel through a barrier to enter the charge storage medium. Such a process is highly inefficient. As a result, the required write pulse voltage and width are excessive, leading to issues of reliability and power consumption. For example, the need for high write voltage and longer write pulses not only poses concerns for the durability and lifespan of NAND flash-based storage but also affects the energy efficiency of data centers and portable devices [6], [7], [8]. In addition, recent aggressive vertical scaling significantly reduces the WL pitch, such that a high write voltage can also degrade the WL isolation dielectrics [9], [10].

However, an alternative solution emerges on the horizon in the form of ferroelectric field-effect transistors (FeFETs). Unlike NAND flash memory, FeFETs leverage a highly efficient write mechanism based on polarization switching. This process can be initiated by applying an electric field to switch the polarization, offering a novel approach to data storage that holds great promise for overcoming the limitations of conventional NAND flash technology [11]. The discovery

Manuscript received 23 April 2024; revised 8 June 2024; accepted 10 June 2024. Date of publication 3 July 2024; date of current version 25 July 2024. This work was supported in part by the SUPREME and PRISM, two of the Semiconductor Research Corporation (SRC) Joint University Microelectronics Program 2.0 (JUMP 2.0) Centers; and in part by NSF under Grant 2344819 and Grant 2235366. The review of this article was arranged by Editor P.-Y. Du. *(Corresponding author: Yixin Qin.)*

Yixin Qin, Zijian Zhao, and Kai Ni are with the Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: yqin4@nd.edu).

Suhwan Lim, Kijoon Kim, Kwangsoo Kim, Wanki Kim, and Daewon Ha are with Samsung Electronics Company Ltd., Hwaseong-si, Gyeonggi-do 18448, Republic of Korea.

Vijaykrishnan Narayanan is with the Department of Computer Science and Engineering, Pennsylvania State University, State College, PA 16802 USA.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2024.3418942.

Digital Object Identifier 10.1109/TED.2024.3418942

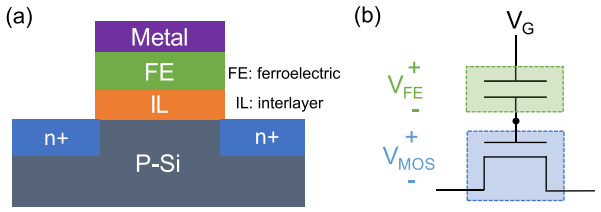

Fig. 1. (a) Device structure of conventional MFIS FeFET. (b) Equivalent capacitance circuit in the MFIS structure.

of ferroelectricity in doped HfO_2 [12] has been a pivotal development, motivating extensive exploration of HfO_2-based FeFETs for various applications. These applications include embedded nonvolatile memory (NVM) [13], [14], [15] and high-density data storage [16], [17], [18], where FeFETs are poised to shine. In the context of high-density data storage, the vertical NAND architecture, proven effective in NAND flash memory, is also applicable to FeFETs, making it a compelling solution for addressing the escalating demands for data storage capacity. To achieve a competitive number of bits per cell, analogous to flash transistors, it is imperative to engineer FeFETs with a large memory window (MW). The MW determines the ability to distinguish between multiple memory states within a single cell, directly impacting the capacity and efficiency of high-density storage solutions. However, the critical evaluation of the FeFET MW is an area of ongoing research and discussion, characterized by nuances and sometimes inconsistencies in theoretical understanding. In the following sections, we delve into a comprehensive analysis of the FeFET MW, considering its theoretical underpinnings and practical implications. We aim to provide a design guideline that enables FeFETs to unlock their full potential in high-density data storage, offering a viable and efficient alternative to the challenges faced by NAND flash technology.

To engineer a large MW FeFET, the most straightforward approach is to increase the ferroelectric layer thickness [19], as typically the MW is correlated with the thickness. However, to compete with the highly scaled vertical NAND flash, the ferroelectric thickness is not expected to be much larger than 20 nm to have a string diameter less than 100 nm [20], [21]. Another approach may be to increase the equivalent oxide thickness (EOT) of the dielectric stack by inserting a dielectric layer into the gate-stack. In this work, within the context of understanding of the MW of a FeFET, the impact of additional dielectric layers on the FeFET MW is theoretically and experimentally evaluated. Note that all the studies in this work are theoretical, demonstrating basic FeFET operation, without considering the extrinsic effects, such as charge trapping and possible ferroelectric property change with its bottom or top interfaces. Such extrinsic effects could also have a huge impact on the MW and have been examined in prior literature [22], [23], [24], [25], [26].

II. Theoretical Understanding of FeFET MW

Fig. 1 presents the device structure and an equivalent capacitance circuit of a conventional FeFET. This device consists of an unintentionally grown interlayer and a ferroelectric layer positioned beneath the top gate metal. The fundamental principle at play here is the law of charge conservation in series capacitance. In this context, it dictates that the total charge within the ferroelectric layer, which includes contributions from both polarization and capacitance charge, must be equivalent to the charge within the MOS capacitor. This principle serves as a foundational concept for understanding the operation and behavior of FeFET devices

$$P_{\mathrm{FE}} + C_{\mathrm{FE}} V_{\mathrm{FE}} = C_{\mathrm{MOS}} V_{\mathrm{MOS}} \tag{1}$$

where P_{FE} is the polarization charge, C_{FE} and C_{MOS} are the capacitances of ferroelectric layer and MOS capacitor (i.e., including the substrate and the interlayer), and V_{FE} and V_{MOS} are the voltage drops of ferroelectric layer and MOS capacitor, respectively. Though C_{MOS} is a nonlinear capacitor, in this work, for the convenience of derivation, it is used here, rather than using complicated integral. For precise modeling, numerical simulations need to be considered.

Before engineering the MW of a FeFET, it is important to understand what determines the MW. In the literature, there have been two different formulas for the MW, and the differences between the two have not been scrutinized. Therefore, it is worthwhile to have a clear definition of the MW and understanding their corresponding application scenarios. Fig. 2 summarizes the two different definitions of the MW. One is the hysteresis window that is typically observed during dc I_D–V_G sweep, where each voltage step is applied for a long enough time that a steady state is assumed to be settled [24]. Therefore, the polarization switches along with the dc sweep, and a Q_{FE}–V_{FE} hysteresis loop (e.g., either a minor loop or saturation loop) is traversed. The MW is then defined as the voltage separation between the two V_{TH}.

It is typically claimed that the maximum MW of a FeFET is

$$\mathrm{MW}_{\mathrm{MAX}} = 2 * E_{\mathrm{C}} * t_{\mathrm{FE}} \tag{2}$$

where the E_{C} is the ferroelectric coercive field and t_{FE} is the ferroelectric thickness. This comes from the observation that the maximum ferroelectric Q_{FE}–V_{FE} loop opening is $2E_{\mathrm{C}}t_{\mathrm{FE}}$. This is indeed the maximum MW that can be achieved under the dc I_D–V_G sweep. For a typical FeFET, as shown in Fig. 2, its threshold voltage (V_{TH}) is

$$\begin{aligned} V_{\mathrm{TH}} &= V_{\mathrm{FB}} + V_{\mathrm{SUB}} + V_{\mathrm{IL}} + V_{\mathrm{FE}} \\ &= V_{\mathrm{FB}} + 2\phi_B + \frac{\sqrt{2q\epsilon_{\mathrm{Si}} N_{\mathrm{A}} 2\phi_B}}{C_{\mathrm{IL}}} + V_{\mathrm{FE}} \end{aligned} \tag{3}$$

where the V_{FB}, V_{SUB}, V_{IL}, and V_{FE} are the flatband voltage and voltage drop across the substrate, interlayer, and ferroelectric layer at the V_{TH}, respectively. In addition, ϕ_{B}, C_{IL}, N_{A}, and ϵ_{Si} are the body potential, interlayer capacitance, the substrate doping, and silicon dielectric constant, respectively. Since at the threshold, irrespective of the low-V_{TH} or high-V_{TH} states, the voltage drop across the substrate and the interlayer will be the same, as they are determined by the $2\phi_{\mathrm{B}}$ voltage drop in the substrate at onset of inversion. Therefore, the first three terms in (3) are fixed. Then, the low-V_{TH} ($V_{\mathrm{TH},L}$) or high-V_{TH}

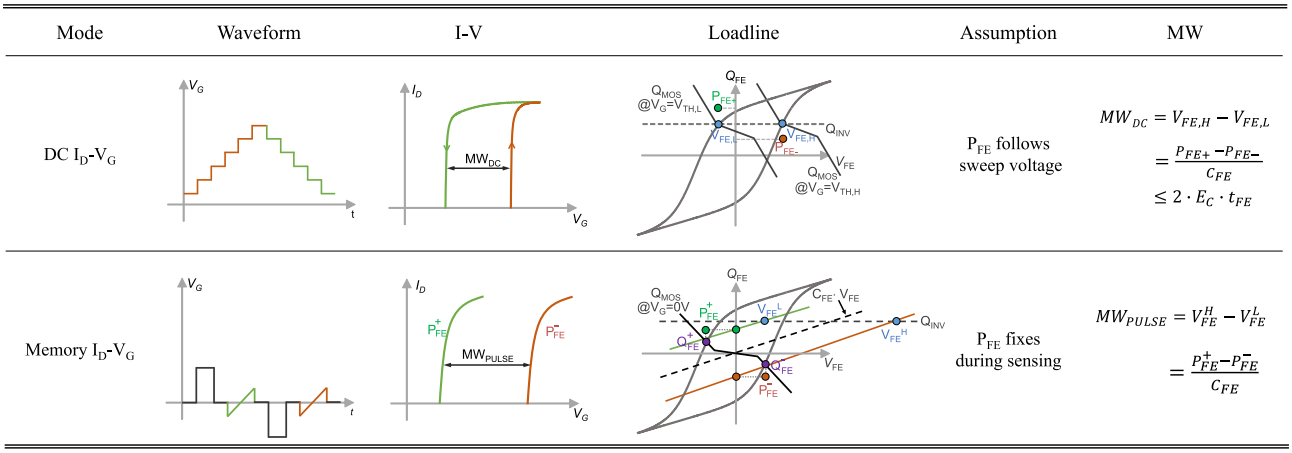

Fig. 2. Comparison of FeFET performance in dc mode and pulse mode.

$(V_{\mathrm{TH},H})$ during dc sweep is

$$\begin{aligned} V_{\mathrm{TH},H} &= V_{\mathrm{FE},H} + \text{const} \\ V_{\mathrm{TH},L} &= V_{\mathrm{FE},L} + \text{const} \\ Q_{\mathrm{INV}} &= P_{\mathrm{FE+}} + C_{\mathrm{FE}} V_{\mathrm{FE},L} \\ &= P_{\mathrm{FE\text{-}}} + C_{\mathrm{FE}} V_{\mathrm{FE},H} \\ \mathrm{MW_{dc}} &= V_{\mathrm{TH},H} - V_{\mathrm{TH},L} \\ &= V_{\mathrm{FE},H} - V_{\mathrm{FE},L} \\ &= \frac{P_{\mathrm{FE+}} - P_{\mathrm{FE\text{-}}}}{C_{\mathrm{FE}}} \\ &\leq 2 * E_{\mathrm{C}} * t_{\mathrm{FE}} \end{aligned} \tag{4}$$

where the $P_{\mathrm{FE+}}/P_{\mathrm{FE-}}$ is the polarization at inversion for the low-V_{TH}/high-V_{TH} state, respectively. Thus, the maximum MW is determined by the maximum separation in the V_{FE} at the same charge, which is $2E_C t_{\mathrm{FE}}$. Graphically, as shown in the loadline analysis shown in Fig. 2, the MW during dc sweep is the separation of the V_{FE} at the onset of condition where the total ferroelectric charge, Q_{FE}, is equal to the inversion charge, i.e., Q_{INV}.

The dc sweep analyzed above is just one type of characterizations that can be done for FeFET, and it is also less relevant for a memory device. For a typical memory, it operates as shown in Fig. 2, and write pulses are typically applied to set/reset the memory states. Especially, in the ideal cases, small read pulses that are small enough such that the memory state is not disturbed will be applied to sense the memory state. Also, the disturbed pulsed read will be between the dc and ideal pulse case. This ideal read can be achieved with either a small enough sensing voltage that the device allows or through fast enough read pulse, such that polarization cannot follow, while source/drain carrier supply can react. Therefore, for the two memory states, there can be two transiently sensed I_D–V_G curves, and the voltage separation of the two is the memory MW, as shown in Fig. 2.

This MW can also be easily derived by recognizing that at the onset of inversion for both memory states

$$P_{\mathrm{FE}}^{+} + C_{\mathrm{FE}} V_{\mathrm{FE}}^{L} = P_{\mathrm{FE}}^{-} + C_{\mathrm{FE}} V_{\mathrm{FE}}^{H} = Q_{\mathrm{inv}} \tag{5}$$

where P_{FE}^{+} and P_{FE}^{-} are the polarization in ferroelectric layer after positive write voltage and negative write voltage, respectively, and V_{FE}^{L} and V_{FE}^{H} are the ferroelectric voltage drop at the onset of inversion for the low-V_{TH} or high-V_{TH} states, respectively. Then, from the expression of V_{TH} shown in (3), the corresponding V_{TH} can be expressed as follows:

$$\begin{aligned} V_{\mathrm{TH}}^{L} &= V_{\mathrm{FB}} + 2\phi_B + \frac{\sqrt{2q\epsilon_{\mathrm{Si}} N_{\mathrm{A}} 2\phi_B}}{C_{\mathrm{OX}}} - \frac{P_{\mathrm{FE}}^{+}}{C_{\mathrm{FE}}} \\ V_{\mathrm{TH}}^{H} &= V_{\mathrm{FB}} + 2\phi_B + \frac{\sqrt{2q\epsilon_{\mathrm{Si}} N_{\mathrm{A}} 2\phi_B}}{C_{\mathrm{OX}}} - \frac{P_{\mathrm{FE}}^{-}}{C_{\mathrm{FE}}} \end{aligned} \tag{6}$$

where $1/C_{\mathrm{OX}} = 1/C_{\mathrm{FE}} + 1/C_{\mathrm{IL}}$. Therefore, the MW is

$$\begin{aligned} \mathrm{MW_{dc}} &= V_{\mathrm{TH}}^{H} - V_{\mathrm{TH}}^{L} \\ &= \frac{P_{\mathrm{FE}}^{+} - P_{\mathrm{FE}}^{-}}{C_{\mathrm{FE}}}. \end{aligned} \tag{7}$$

Since it is assumed that the polarization is not disturbed in the ideal scenario during the sensing, while the V_{FE} can be varied via the linear dielectric contribution of the ferroelectric, the MW is determined by the polarization of the two memory states. This MW can also understood graphically from the loadline analysis, as shown in Fig. 2. The memory states at $V_{\mathrm{G}} = 0$ V are the intersection points (i.e., Q_{FE}^{+} and Q_{FE}^{-}) between the MOS loadline and the ferroelectric Q_{FE}–V_{FE} hysteresis loop. During memory sensing, for the ferroelectric, only the linear C_{FE} is active, and the polarization states (i.e., P_{FE}^{+} and P_{FE}^{-}) simply shift the linear Q_{FE}–V_{FE} curve. Once P_{FE} is determined, the sensing is done by raising the ferroelectric total charge to the Q_{INV} while keeping the polarization intact. The intersection points between the shifted curves and the inversion charge Q_{INV} are the corresponding V_{FE} at inversion, and their separation is the MW. Note that in this case, it is assumed that the sensing can be performed without impacting the polarization states; therefore, the V_{FE}^{H} and V_{FE}^{L} can be greater than the coercive voltage. If, in practical cases, disturb to polarization states can occur during sensing, then the corresponding V_{FE} for the high-V_{TH} state will be between the two extreme cases, i.e., $V_{\mathrm{FE},H}$, where polarization fully follows sweeping and V_{FE}^{H}, where polarization remains unchanged. Similar arguments also apply for the low-V_{TH} state.

So what is the relationship between the two definitions in (4) and (7)? Here, we provide additional analysis on this. After the memory write operation, there is a finite voltage drop, thus leading to so-called depolarization field in the ferroelectric (i.e., $V_{\rm FE}^{+}$ and $V_{\rm FE}^{-}$), and the voltages are determined by the polarization (i.e., $P_{\rm FE}^{+}$ and $P_{\rm FE}^{-}$) and the capacitance in the gate-stack. In particular,

$$V_{\rm FE}^{+} = \frac{-P_{\rm FE}^{+}}{C_{\rm FE}+C_{\rm MOS}}$$
$$V_{\rm FE}^{-} = \frac{-P_{\rm FE}^{-}}{C_{\rm FE}+C_{\rm MOS}}. \tag{8}$$

At $V_{\rm G}=0$ V, the $V_{\rm FE}$ difference for the two memory states ($\Delta V_{\rm FE}$) is also equal to the underling MOS voltage difference ($\Delta V_{\rm MOS}$)

$$\Delta V_{\rm FE} = \Delta V_{\rm MOS} = \frac{P_{\rm FE}^{+}-P_{\rm FE}^{-}}{C_{\rm FE}+C_{\rm MOS}} \le 2*E_{\rm C}*t_{\rm FE} \tag{9}$$

where the last expression is due to the constraints that the polarization states are bounded by the $Q_{\rm FE}$–$V_{\rm FE}$ hysteresis loop and charges are different $Q_{\rm FE}^{+}\ge 0$ and $Q_{\rm FE}^{-}\le 0$. The separation of $V_{\rm FE}$ is similar to the dc sweep, and their maximum is the same. However, for the pulsed memory operation, unlike dc mode operation, the separation of $V_{\rm FE}$ is not the actual MW. The MW in this case can be defined as the measured gate voltage difference to nullify the $\Delta V_{\rm MOS}$ gap as at the onset of inversion, $V_{\rm MOS}$ is the same irrespective of the memory state. Then, the MW is the value that can compensate for the $\Delta V_{\rm MOS}$, i.e.,

$$\mathrm{MW_{PULSE}} = \Delta V_{\rm G} = \Delta V_{\rm MOS}\left(1+\frac{C_{\rm MOS}}{C_{\rm FE}}\right) = \frac{P_{\rm FE}^{+}-P_{\rm FE}^{-}}{C_{\rm FE}}. \tag{10}$$

The reason why the MW during pulsed sensing is not just the $\Delta V_{\rm MOS}$ is because there are additional voltage drop across the ferroelectric need to be included. That is, why an additional factor $1+C_{\rm MOS}/C_{\rm FE}$ is included. Note that though, at the final expression, the MW does not contain an explicit dependence on the $C_{\rm MOS}$, an implicit dependency does exist. Such a dependency shows up during the memory write process and later relaxation process when write pulse is removed. Together with $C_{\rm FE}$, they jointly determine the maximum polarization charge that can be held without flipping the polarization spontaneously. In general, the higher the $C_{\rm MOS}$, the more charge the underling MOS structure can supply with the same voltage, which means more polarization charge can be supported, hence a larger MW. Based on these analysis, it can be concluded that for FeFET working as a memory, the MW defined in (7) is more appropriate.

To better understand the MW defined in (7), we can consider two extreme cases. In one case, if $C_{\rm MOS}$ is very small, such that $Q_{\rm FE}^{+}=Q_{\rm FE}^{-}=0$, i.e., the loadline intersects the $Q_{\rm FE}$–$V_{\rm FE}$ loop at the coercive field points. In such a case

$$-C_{\rm FE}E_{\rm FE}t_{\rm FE}+P_{\rm FE}^{+}=C_{\rm FE}E_{\rm FE}t_{\rm FE}+P_{\rm FE}^{-}=0. \tag{11}$$

Therefore, the MW in this case is simply $2E_{\rm C}t_{\rm FE}$. In another extreme case where $C_{\rm MOS}$ is almost infinite, such as a conducting metal, the intersection points will be close to the y-axis, such that the $Q_{\rm FE}^{+}=-Q_{\rm FE}^{-}=P_r$. Then, the MW can reach $2P_r/C_{\rm FE}$. However, this can rarely be achieved as an almost infinite $C_{\rm MOS}$ means that it is impossible to modulate the carrier concentration and deplete the carriers. In practical semiconductors, due to limited charge supplying capability, the intersection points are not far from 0 μC/cm^2, therefore giving an MW close to $2E_{\rm C}t_{\rm FE}$.

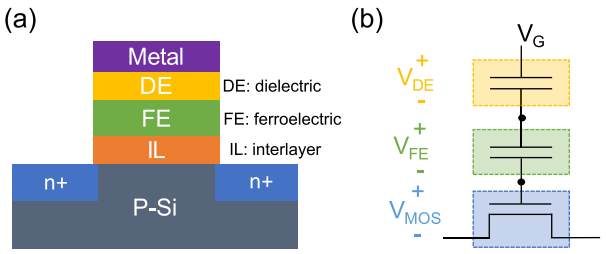

Fig. 3. (a) Device structure of FeFET with additional DE layer. (b) Equivalent capacitance circuit in the MFIS structure.

III. Impact of Dielectric Layer on FeFET MW

Next, we will examine theoretically the impact of a dielectric layer above the ferroelectric layer on the MW of a FeFET. The device structure and equivalent capacitance circuit are shown in Fig. 3. Since $V_{\rm TH}$ increases with the EOT of the stack, one may think that the MW also increases with EOT. In this work, we perform combined theoretical and experimental studies to clarify both points and show the impact of an additional dielectric layer on the FeFET MW. These insights will be useful for the engineering of large MW FeFET for high-density storage.

Similarly, with the above-discussed MFIS structure, three capacitors in series should follow the charge conservation law, where total charge of each layer in the gate-stack is equal

$$P_{\rm FE}+C_{\rm FE}V_{\rm FE}=C_{\rm MOS}V_{\rm MOS}=C_{\rm DE}V_{\rm DE} \tag{12}$$

where $C_{\rm DE}$ and $V_{\rm DE}$ are the capacitance and voltage drop of the inserted dielectric layer. Also, the voltage distribution of each layer in the series capacitors needs to satisfy

$$V_{\rm G}=V_{\rm FE}+V_{\rm MOS}+V_{\rm DE}. \tag{13}$$

Then, for pulsed memory operation, the corresponding $V_{\rm TH}$ after applying positive or negative gate write pulses will be

$$V_{\rm TH}^{L} = V_{\rm FB}+2\phi_B+\frac{\sqrt{2q\epsilon_{\rm Si}N_{\rm A}2\phi_B}}{C'_{\rm OX}}-\frac{P_{\rm FE}^{+}}{C_{\rm FE}}$$
$$V_{\rm TH}^{H} = V_{\rm FB}+2\phi_B+\frac{\sqrt{2q\epsilon_{\rm Si}N_{\rm A}2\phi_B}}{C'_{\rm OX}}-\frac{P_{\rm FE}^{-}}{C_{\rm FE}} \tag{14}$$

where $1/C'_{\rm OX}=1/C_{\rm FE}+1/C_{\rm IL}+1/C_{\rm DE}$. The resulting MW is the same as that in (7). Therefore, the increased EOT does not increase the MW as $V_{\rm TH}$ of both states is shifted by the same amount. The analysis shows that if the same polarization is switched as the case where the dielectric is not added, then the MW remains the same. Note that all these analyses are relying on the assumption that only polarization switching is present, and no other $V_{\rm TH}$ modulation mechanisms are introduced, whose studies will be left for future work.

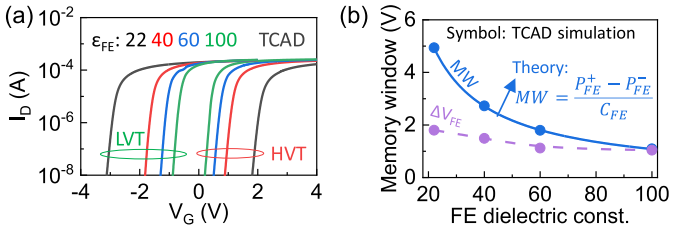

Fig. 4. (a) I_D–V_G curves shift under different FE dielectric constants in TCAD simulation. (b) Comparison of simulated MW and the theory, and MW is larger than ΔV_{FE}.

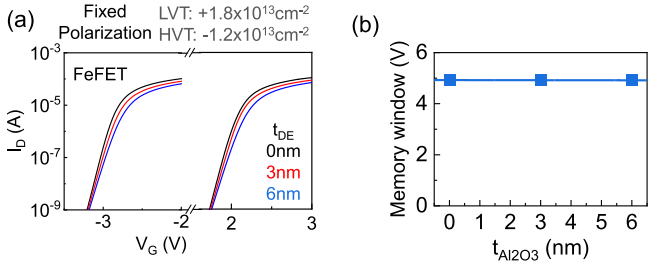

Fig. 5. (a) I_D–V_G curves of different dielectric thicknesses in FeFET under fixed polarization. (b) Extracted MW with different dielectric thicknesses.

IV. Modeling of FeFET MW

Next modeling of the FeFET will be performed to validate the MW for pulsed memory mode. For this study, TCAD simulations are performed. As, during the process of memory sensing, the applied read voltage does not induce a change in the polarization charge status, therefore, the equivalence between the Preisach model and the presence of fixed charges for both the low-V_{TH} and high-V_{TH} states in our previous work [27], [28] is leveraged to directly emulate the polarization charge through the introduction of a fixed interfacial charge denoted as Q_{fix}. With this experimental setup, we focused on studying the impact of the C_{FE} on the MW by altering the dielectric constant of the ferroelectric layer while maintaining a constant interfacial charge. The resulting I_D–V_G curves for the two distinct memory states are presented in Fig. 4(a). It becomes evident that as the dielectric constant, i.e., C_{FE}, of the ferroelectric layer increases, the high-V_{TH} decreases, while the low-V_{TH} increases. This shift in V_{TH} ultimately results in a reduction of the MW of the FeFET. In Fig. 4(b), all the simulated MW values and voltage differences in the ferroelectric layer were compiled and plotted. It is confirmed that the extracted MW values align closely with the theoretical formula shown in (7) discussed earlier and are greater than the voltage drop in the ferroelectric layer ΔV_{FE}. This confirmation reinforces the validity of the MW expression.

Next, the impact of additional dielectric layer above ferroelectric is studied. V_{TH} of the MOSFET tends to increase as the dielectric layer thickness grows. This phenomenon can be readily understood by considering that a thicker dielectric layer necessitates a larger gate voltage distribution across the gate-stack. Also, the V_{TH} depends linearly on the dielectric layer thickness due to the impact of EOT on V_{TH} in MOSFETs. Based on this understanding, similar simulations were carried

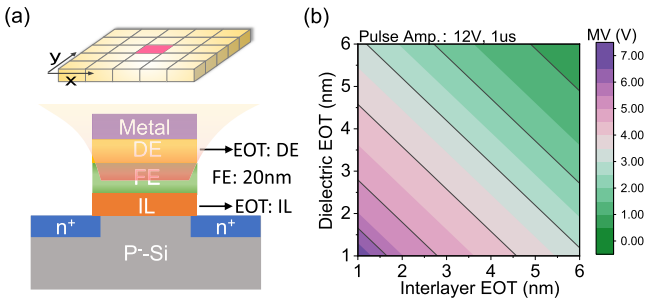

Fig. 6. (a) FeFET model for switching study is adapted from our prior work [27]. (b) MW dependence on the dielectric layer and interlayer EOT under a fixed write voltage.

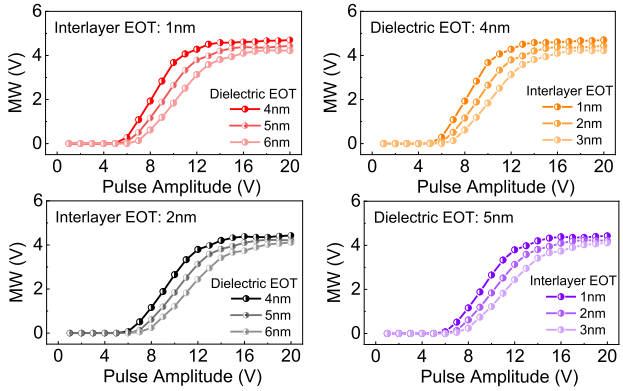

Fig. 7. Simulation results of MW dependence with fixed EOT of interlayer, EOT of dielectric layer, and different pulse amplitudes. Thicker dielectric reduces MW.

out in a FeFET with different dielectric layers, as depicted in Fig. 5(a). Various dielectric layer thicknesses were explored in FeFET. Consequently, MW extracted in Fig. 5(b) demonstrates that the MW remains constant under varying dielectric layer thickness conditions. Based on these studies, it confirms the theoretical picture presented in Fig. 2.

The aforementioned studies have operated under the assumption of a constant polarization for different dielectric thicknesses; however, this assumption may not hold in practical FeFET devices. Insertion of a top dielectric could have multiple effects, such as changing the ferroelectricity of the film, creating more interface defects, and modulating the switchable polarization due to more depolarization field. In this theoretical study, the ferroelectric properties are considered to be the same when introducing different dielectric layers, and only the impact of the dielectric layer on the switching is considered. For the switching dynamics study, our previously developed multidomain Monte Carlo FeFET model [27] is modified by incorporating the dielectric layer electrically, as illustrated in Fig. 6(a). The nucleation-limited polarization switching kinetics and multiple domains with a distribution of switching activation field are modeled. This model has been calibrated using experimentally measured switching dynamics in a FeFET in our previous work, and the depolarization field effect was also automatically considered [27]. Subsequently, the effects of changes in EOT were explored for both the interlayer and dielectric layer while keeping a fixed write

pulse (±12 V, 1 μs). The results are presented in Fig. 6(b), showcasing a remarkably linear relationship between the MW and the combined thickness of these two layers, as the total thickness determines the EOT. This also suggests that it does not matter where the dielectric layers are located as electrically their roles are the same.

In Fig. 7, it is delved deeper into this investigation by varying the interlayer EOT from 1 to 3 nm and the dielectric layer EOT from 4 to 6 nm. The graph illustrates the MW as a function of write pulse amplitude for different EOT values with a fixed pulsewidth of 1 μs. These findings reinforce the notion that an increase in the dielectric EOT leads to a reduction in the switched polarization at a given pulse amplitude due to stronger depolarization field. In essence, from these results, it can be concluded that, for pure FeFET operation that relies on the polarization switching physics, the addition of extra layers does not contribute to an increase in the MW, shedding light on a crucial aspect of FeFET behavior [29], [30].

V. Experimental Investigation of FeFET MW

For a comprehensive study of FeFET MW, we also conducted experimental investigations by integrating a 20-nm-thick $Hf_{0.5}Zr_{0.5}O_2$ FeFET. To avoid the formation of significant monoclinic dielectric phase in the thick film, a thin Al_2O_3 layer (1 nm) is inserted after 10-nm $Hf_{0.5}Zr_{0.5}O_2$ growth [31]. The process flow for the fabricated FeFET is depicted in Fig. 8(a). The fabrication is carried out on a p-type silicon substrate. After phosphorus ion implantation and activation, the isolation oxide in the gate area is removed, and the gate area is thoroughly cleaned. The gate dielectric layers $Hf_{0.5}Zr_{0.5}O_2$ and Al_2O_3 are deposited through atomic layer deposition (ALD) at the temperatures of 250 °C and 200 °C, respectively. Source/drain via is opened by reactive-ion etching (RIE) and buffered oxide etching (BOE). A 90-nm-thick tungsten (W) layer is sputtered on the wafer to serve as source, drain, and gate metal. The final step involves annealing the device in forming gas (N_2 + H_2, 350 °C 1 min) to passivate the surface and reduce the transistor subthreshold swing, and N_2 (500 °C 20 s) to facilitate the crystallization of the ferroelectric material. To study the impact of top dielectric layer, a sample with 3-nm top Al_2O_3 is deposited on $Hf_{0.5}Zr_{0.5}O_2$, while the control does not have the Al_2O_3. The cross-sectional schematics of a single FeFET and a top-view scanning electron microscopy (SEM) image are also included in Fig. 8(a). Fig. 8(b) and (c) shows the transmission electron microscopy (TEM) cross sections of the gate-stack of FeFETs without and with top Al_2O_3, respectively. The energy-dispersive X-ray spectroscopy (EDX) line scans are shown in Fig. 9(a) and (c), and electron energy loss spectroscopy (EELS) elemental mapping is shown in Fig. 9(b) and (d) for FeFETs without/with the top Al_2O_3 layer, respectively. These elemental profiles of both devices confirm the intended design.

Fig. 10(a) and (b) shows the measured V_{TH} for both the low-V_{TH} and high-V_{TH} states with varying write pulse amplitudes at a fixed 1-μs pulsewidth for FeFETs without and with top Al_2O_3, respectively. To verify device-to-device variations in switching dynamics, ten devices for each structure are measured. These measurements were taken at identical device areas and write voltages ranging from 5 to 11 V with a step size of 0.1 V. As the write voltage amplitude increases, a decrease in the low-V_{TH} state is observed, while the increase in high-V_{TH} state quickly saturates, accompanied by an expansion of the MW. The MW reached saturated with an 11-V write voltage and cannot further increase with larger applied voltages. In FeFETs without the top Al_2O_3, the maximum MW is achieved at approximately 4.8 V, while in FeFETs with the top Al_2O_3, the maximum MW is approximately 4.3 V, as illustrated in Fig. 10(a) and (b). Furthermore, Fig. 10(c) and (d) presents the retention results for both FeFET structures at several different write amplitudes. The MW is larger in the FeFETs without top Al_2O_3 dielectric than the ones with top Al_2O_3, which is consistent with the theoretical investigation in previous sections. In addition, Fig. 11 demonstrates the endurance performance of FeFET

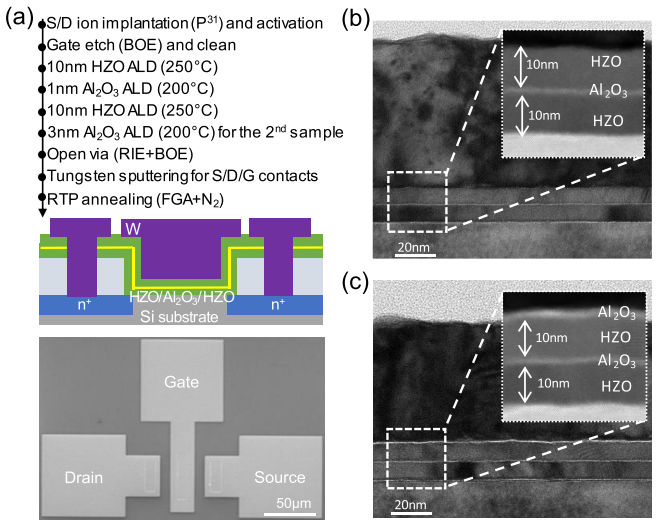

Fig. 8. (a) Key fabrication process flow, cross-sectional schematic, and SEM top view of FeFET. Gate-stack TEM of (b) FeFET without top Al_2O_3 layer and (c) FeFET with top Al_2O_3 layer.

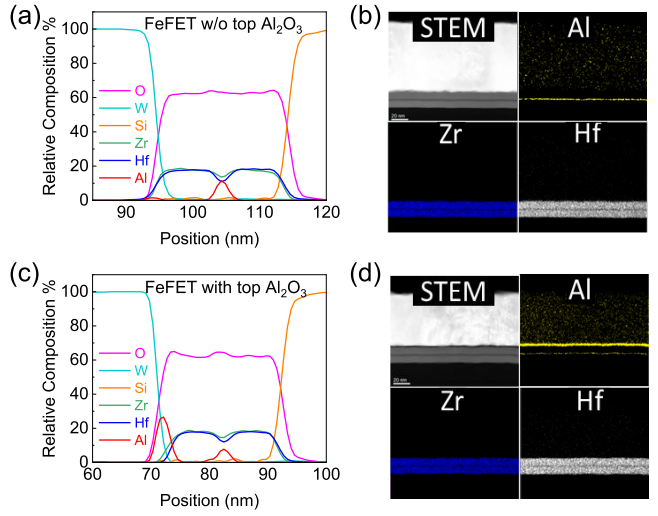

Fig. 9. EDX line scanning analysis of FeFET (a) without top Al_2O_3 layer and (c) with top Al_2O_3 layer. Electron energy loss spectroscopy elemental mapping of FeFET (b) without top Al_2O_3 layer and (d) with top Al_2O_3 layer.

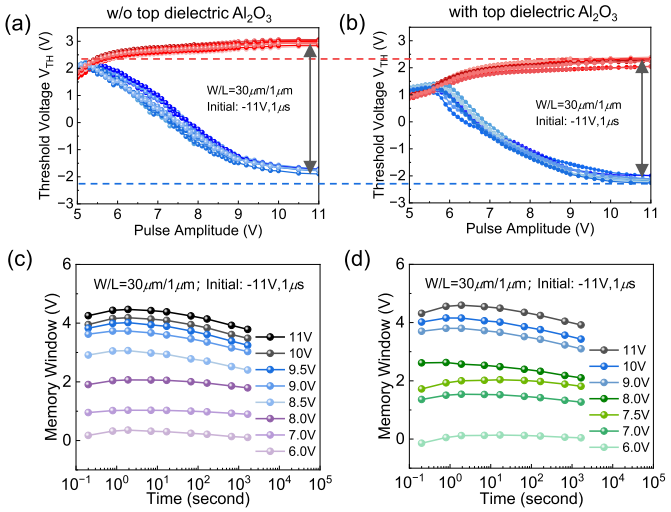

Fig. 10. Statistical measurement of V_{TH} shift under different write pulse amplitudes in FeFET (a) without dielectric Al_2O_3 layer and (b) with dielectric Al_2O_3 layer. Retention measurement of FeFET (c) with dielectric and (d) without dielectric layer.

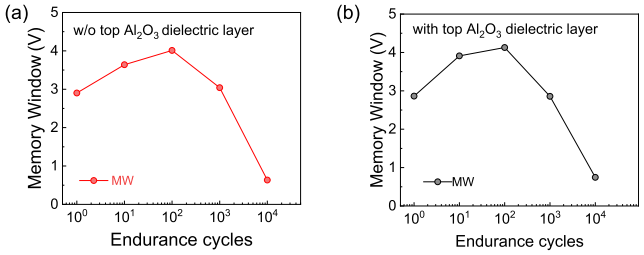

Fig. 11. Endurance measurement of FeFET (a) without top dielectric Al_2O_3 layer and (b) with top dielectric Al_2O_3 layer under 11-V pulse.

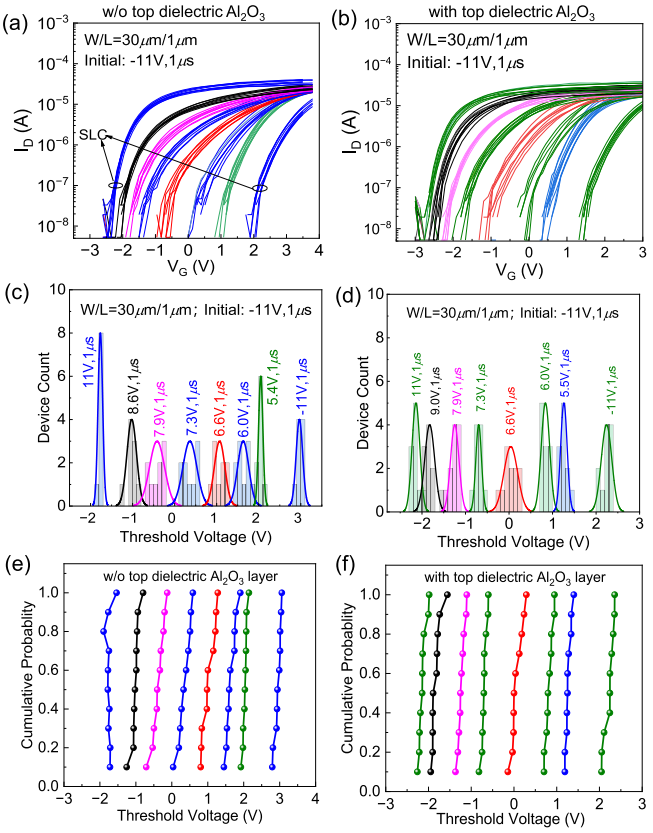

Fig. 12. TLC demonstration. I_D–V_G of FeFET (a) without top Al_2O_3 layer and (b) with top Al_2O_3 layer at eight write voltages. The 3-bit levels of V_{TH} distribution (c) without top Al_2O_3 layer and (d) with top Al_2O_3 layer. Cumulative probability of 3-bit V_{TH} (e) without top Al_2O_3 layer and (f) with top Al_2O_3 layer.

with and without top Al_2O_3 dielectric layer. Both devices can only show around $10^3/10^4$ cycles with MW convergence. The endurance of these FeFET devices is similar to typical FeFET [32] as well. To enhance the reliability performance, it is required to study the physical mechanism behind the endurance degradation.

The ability of our developed FeFETs to enable multilevel cell (MLC) operation is confirmed, especially with the presence of a large MW. In the case of the FeFET without the top Al_2O_3 layer, blue color lines in Fig. 12(a) present the I_D–V_G curves under four different write voltages (−11.0, 6.0, 7.3, and 11.0 V) for ten devices being tested. The extracted MLC V_{TH} distribution is displayed in Fig. 12(c), demonstrating tightly distributed V_{TH} levels with negligible overlap, indicating promising MLC operation. The cumulative probability distribution of V_{TH} shown in Fig. 12(e) further corroborates the feasibility of MLC in the FeFET without the top Al_2O_3 layer. Similarly, the MLC capability is also evident in the FeFET structure with the Al_2O_3 dielectric layer, as shown in green lines in Fig. 12(b), (d), and (f). Even though the MW is slightly smaller compared with the structure without the dielectric layer, it still demonstrates the potential for MLC operation.

Furthermore, triple-level cell (TLC) is also demonstrated in both FeFET structures. For device without dielectric layer, Fig. 12(a) presents eight distinct I_D–V_G curves, showcasing the device's response under a range of write voltages. The corresponding distribution of extracted V_{TH} and its cumulative distribution are depicted in Fig. 12(c) and (e), respectively. In this case, small overlaps between neighboring states can be observed. Since this was achieved using a single-shot write without a write-and-verify process, there is room for further variation reduction. Based on these investigations, TLC operation in a large MW FeFET without a dielectric layer can be demonstrated. Similarly, the TLC characterization of FeFET with top dielectric layer is shown in Fig. 12(b), (d), and (f). Due to the slightly smaller MW in FeFET with top Al_2O_3 layer, eight V_{TH} levels distribution exhibits more overlap than that observed in the FeFET without the top dielectric layer. Both FeFET structures can support TLC, but the device with the smaller MW has more overlap between states. Note that these demonstrations just demonstrate the feasibility. To really demonstrate a technology with the developed devices, more studies and optimizations are necessary, such as scaling for high density, variability especially the tail bit, and reliability.

VI. Discussions

All the studies in this work are focused on the intrinsic FeFET behavior where only the polarization switching is considered without considering extrinsic effects. Also, the experiment data reported in this work are not contradictory to the intrinsic operation picture, especially that the top

Al_2O_3 insertion does not increase the MW. Along this line of thinking, it is interesting to check what kinds of engineering that can be done to enhance the MW. First approach could be to reduce the C_{FE}. The most obvious method is to simply increase ferroelectric thickness. However, to have a comparable or competitive density in the XY plane compared with vertical NAND flash, the ferroelectric thickness need to be no more than 20 nm, thus presenting new constraints and challenges to maximize the MW in this limitation. Another approach will be to reduce the ferroelectric dielectric constant. Though, for HfO_2, using dopants, such as Si or Al, could reduce the dielectric constant, the small concentration does not have a large impact. Looking for other ferroelectric materials that has a lower dielectric constant is one option but not always feasible.

Another approach will be to increase the ΔP_{FE}. This may not always be easy. Even though trapping can increase ΔP_{FE}, it does not increase the window as the trapping will compensate the ΔP_{FE} in MFIS structure, and their net results determine the MW. However, one previously reported design, i.e., metal–ferroelectric–metal–insulator–semiconductor (MFMIS) FeFET, can be analyzed in this aspect [33], [34], [35]. It has been shown that by appropriately reducing the ferroelectric capacitor with respect to underling MOSFET, the MW can be enhanced. At a first glance, it may be ascribed to a smaller C_{FE} due to reduced area. It is actually not, as the absolute charge will also scale with the area. The reason why the MW is enhanced in this case is because for a properly designed stack, there will be more ferroelectric voltage drop, such that more polarization can be switched.

Although this work focuses on the intrinsic polarization behavior of FeFET, extrinsic effects, such as charge trapping, could play an inseparable role during the memory operation. Such effect could be harmful or beneficial, depending on the gate-stack design. For instance, it is very likely to have charge injection from the channel into the ferroelectric layer during memory write operation. In this case, as injected carriers are trapped in the ferroelectric layer or between the ferroelectric layer and the channel layer, the measured MW of FeFET cannot reach to the simulated maximum MW and largely lower the MW [20], [25]. To maximize the MW in limited gate-stack thickness of FeFET, the extrinsic effects, if properly designed, can also be harnessed to work in synergy with polarization switching. One potential approach involves engineering the gate-stack to maximize charge injection from the gate side while minimizing the channel-side injection, such that a net charge injection from the gate side can be induced [25], [26]. One reported work demonstrated a very large MW above 10 V with one novel dielectric-ferroelectric gate-stack [18]. Although the exact information of the ferro stack and top interlayer is unknown, that structure could possibly utilize the efficient charge trapping layer as the top dielectric and laminated doped HfO_2 as the FE stack layer. For instance, silicon nitride as the top interlayer could theoretically greatly enhance the gate-side injection so as to enlarge MW. Furthermore, optimized laminated HZO structure can boost the orthorhombic phase fraction and effective polarization, resulting in larger MW and high reliability [36]. However, significant charge storage with top dielectric Al_2O_3 dielectric layer was not observed in our work, and this might be related to the high quality of Al_2O_3 layer that suppresses gate-side charge injection. Therefore, a dielectric layer with abundant deep defects that can trap carriers could be a potential role to boost MW. With injected electrons trapped close to the gate side under negative write, the polarization-induced positive V_{TH} shifts will add to the trapped electron-induced positive V_{TH} shift, thereby enhancing the MW. Such injected carriers have additional benefit in weakening the depolarization field in the ferroelectric during retention, therefore helps hold the polarization. The effectiveness of this approach is worth further study.

VII. Conclusion

In this study, we have conducted both theoretical and experimental investigations into the MW of FeFETs, aiming to provide guidance for the development of high-density storage devices. We have systematically clarified the assumptions and interpretation of typically measured MW of a FeFET. We also established that the addition of a dielectric layer to the gate-stack does not necessarily improve the MW when the ferroelectric switching is the only mechanism. Experiments have confirmed that a large MW, reaching 4.8 V, is achieved with a 20-nm-thick $Hf_{0.5}Zr_{0.5}O_2$ layer. These results offer valuable insights for the future design of FeFETs aimed at high-density storage applications, providing a foundation for further advancements in this technology.

References

[1] A. Goda, "3-D NAND technology achievements and future scaling perspectives," *IEEE Trans. Electron Devices*, vol. 67, no. 4, pp. 1373–1381, Apr. 2020, doi: 10.1109/TED.2020.2968079.

[2] L. Heineck and J. Liu, "3D NAND flash status and trends," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2022, pp. 1–4, doi: 10.1109/IMW52921.2022.9779282.

[3] R. Meyer, Y. Fukuzumi, and Y. Dong, "3D NAND scaling in the next decade," in *IEDM Tech. Dig.*, Dec. 2022, pp. 26.1.1–26.1.4, doi: 10.1109/IEDM45625.2022.10019570.

[4] A. Goda and K. Parat, "Scaling directions for 2D and 3D NAND cells," in *IEDM Tech. Dig.*, Dec. 2012, pp. 2.1.1–2.1.4, doi: 10.1109/IEDM.2012.6478961.

[5] R. Micheloni, L. Crippa, C. Zambelli, and P. Olivo, "Architectural and integration options for 3D NAND flash memories," *Computers*, vol. 6, no. 3, p. 27, 2017, doi: 10.3390/computers6030027.

[6] A. S. Spinelli, C. M. Compagnoni, and A. L. Lacaita, "Reliability of NAND flash memories: Planar cells and emerging issues in 3D devices," *Computers*, vol. 6, no. 2, p. 16, Apr. 2017, doi: 10.3390/computers6020016.

[7] Y. Li and K. N. Quader, "NAND flash memory: Challenges and opportunities," *Computer*, vol. 46, no. 8, pp. 23–29, Aug. 2013, doi: 10.1109/MC.2013.190.

[8] C. Zhao, C. Zhao, S. Taylor, and P. Chalker, "Review on non-volatile memory with high-*k* dielectrics: Flash for generation beyond 32 nm," *Materials*, vol. 7, no. 7, pp. 5117–5145, Jul. 2014, doi: 10.3390/ma7075117.

[9] S. Rachidi et al., "At the extreme of 3D-NAND scaling: 25 nm Z-pitch with 10 nm word line cells," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2022, pp. 1–4, doi: 10.1109/IMW52921.2022.9779303.

[10] Y.-H. Hsiao, H.-T. Lue, T.-H. Hsu, K.-Y. Hsieh, and C.-Y. Lu, "A critical examination of 3D stackable NAND flash memory architectures by simulation study of the scaling capability," in *Proc. IEEE Int. Memory Workshop*, May 2010, pp. 1–4, doi: 10.1109/IMW.2010.5488390.

[11] S. Yoon et al., "Highly stackable 3D ferroelectric NAND devices: Beyond the charge trap based memory," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2022, pp. 1–4, doi: 10.1109/IMW52921.2022.9779278.

[12] T. S. Böscke, J. Müller, D. Bräuhaus, U. Schröder, and U. Böttger, "Ferroelectricity in hafnium oxide thin films," *Appl. Phys. Lett.*, vol. 99, no. 10, Sep. 2011, Art. no. 102903, doi: 10.1063/1.3634052.

[13] S. George et al., "Nonvolatile memory design based on ferroelectric FETs," in *Proc. 53rd Annu. Design Autom. Conf.*, Jun. 2016, pp. 1–6, doi: 10.1145/2897937.2898050.

[14] M. Trentzsch et al., "A 28 nm HKMG super low power embedded NVM technology based on ferroelectric FETs," in *IEDM Tech. Dig.*, Dec. 2016, pp. 11.5.1–11.5.4, doi: 10.1109/IEDM.2016.7838397.

[15] K. Ni et al., "A novel ferroelectric superlattice based multi-level cell non-volatile memory," in *IEDM Tech. Dig.*, Dec. 2019, pp. 8–28, doi: 10.1109/IEDM19573.2019.8993670.

[16] T. Ali et al., "A multilevel FeFET memory device based on laminated HSO and HZO ferroelectric layers for high-density storage," in *IEDM Tech. Dig.*, Dec. 2019, pp. 28.7.1–28.7.4, doi: 10.1109/IEDM19573.2019.8993642.

[17] H. W. Park, J. Lee, and C. S. Hwang, "Review of ferroelectric field-effect transistors for three-dimensional storage applications," *Nano Select*, vol. 2, no. 6, pp. 1187–1207, Jun. 2021, doi: 10.1002/nano.202000281.

[18] S. Yoon et al., "QLC programmable 3D ferroelectric NAND flash memory by memory window expansion using cell stack engineering," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2023, pp. 1–2, doi: 10.23919/vlsitechnologyandcir57934.2023.10185294.

[19] T. Ali et al., "Silicon doped hafnium oxide (HSO) and hafnium zirconium oxide (HZO) based FeFET: A material relation to device physics," *Appl. Phys. Lett.*, vol. 112, no. 22, May 2018, Art. no. 222903, doi: 10.1063/1.5029324.

[20] S. Lim et al., "Comprehensive design guidelines of gate stack for QLC and highly reliable ferroelectric VNAND," in *IEDM Tech. Dig.*, Dec. 2023, pp. 1–4, doi: 10.1109/iedm45741.2023.10413820.

[21] D. Das et al., "Experimental demonstration and modeling of a ferroelectric gate stack with a tunnel dielectric insert for NAND applications," in *IEDM Tech. Dig.*, Dec. 2023, pp. 1–4, doi: 10.1109/iedm45741.2023.10413697.

[22] S. Deng et al., "Examination of the interplay between polarization switching and charge trapping in ferroelectric FET," in *IEDM Tech. Dig.*, Dec. 2020, pp. 4.4.1–4.4.4, doi: 10.1109/IEDM13553.2020.9371999.

[23] S. Deng et al., "Unraveling the dynamics of charge trapping and detrapping in ferroelectric FETs," *IEEE Trans. Electron Devices*, vol. 69, no. 3, pp. 1503–1511, Mar. 2022, doi: 10.1109/TED.2022.3143485.

[24] K. Toprasertpong, M. Takenaka, and S. Takagi, "Memory window in ferroelectric field-effect transistors: Analytical approach," *IEEE Trans. Electron Devices*, vol. 69, no. 12, pp. 7113–7119, Dec. 2022, doi: 10.1109/TED.2022.3215667.

[25] S. Yoo et al., "An analytical interpretation of the memory window in ferroelectric field-effect transistors," *Appl. Phys. Lett.*, vol. 123, no. 22, Nov. 2023, Art. no. 222902, doi: 10.1063/5.0168515.

[26] S. R. Rajwade, K. Auluck, J. B. Phelps, K. G. Lyon, J. T. Shaw, and E. C. Kan, "A ferroelectric and charge hybrid nonvolatile memory—Part I: Device concept and modeling," *IEEE Trans. Electron Devices*, vol. 59, no. 2, pp. 441–449, Feb. 2012, doi: 10.1109/TED.2011.2175396.

[27] S. Deng et al., "A comprehensive model for ferroelectric FET capturing the key behaviors: Scalability, variation, stochasticity, and accumulation," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2020, pp. 1–2, doi: 10.1109/VLSITechnology18217.2020.9265014.

[28] K. Ni, S. Thomann, O. Prakash, Z. Zhao, S. Deng, and H. Amrouch, "On the channel percolation in ferroelectric FET towards proper analog states engineering," in *IEDM Tech. Dig.*, Dec. 2021, pp. 15.3.1–15.3.4, doi: 10.1109/IEDM19574.2021.9720631.

[29] N. Tasneem et al., "The impacts of ferroelectric and interfacial layer thicknesses on ferroelectric FET design," *IEEE Electron Device Lett.*, vol. 42, no. 8, pp. 1156–1159, Aug. 2021, doi: 10.1109/LED.2021.3088388.

[30] M. Liao et al., "Impact of saturated spontaneous polarization on the endurance fatigue of Si FeFET with metal/ferroelectric/interlayer/Si gate structure," *IEEE Trans. Electron Devices*, vol. 70, no. 8, pp. 4055–4061, Aug. 2023, doi: 10.1109/TED.2023.3285715.

[31] H. J. Kim et al., "Grain size engineering for ferroelectric $Hf_{0.5}Zr_{0.5}O_2$ films by an insertion of Al_2O_3 interlayer," *Appl. Phys. Lett.*, vol. 105, no. 19, Nov. 2014, Art. no. 192903, doi: 10.1063/1.4902072.

[32] Y. Raffel et al., "Endurance improvements and defect characterization in ferroelectric FETs through interface fluorination," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2022, pp. 1–4, doi: 10.1109/IMW52921.2022.9779277.

[33] Z. Zheng et al., "Boosting the memory window of the BEOL-compatible MFMIS ferroelectric/anti-ferroelectric FETs by charge injection," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2022, pp. 389–390, doi: 10.1109/VLSITechnologyandCir46769.2022.9830466.

[34] X. Wang et al., "Deep insights into the interplay of polarization switching, charge trapping, and soft breakdown in metal-ferroelectric-metal-insulator-semiconductor structure: Experiment and modeling," in *IEDM Tech. Dig.*, Dec. 2022, pp. 13.3.1–13.3.4, doi: 10.1109/IEDM45625.2022.10019390.

[35] S. Woo, G. Choe, A. I. Khan, S. Datta, and S. Yu, "Design of ferroelectric-metal field-effect transistor for multi-level-cell 3D NAND flash," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2023, pp. 1–4, doi: 10.1109/IMW56887.2023.10145961.

[36] H. J. Lee et al., "Laminated ferroelectric FET with large memory window and high reliability," *IEEE Trans. Electron Devices*, vol. 71, no. 4, pp. 2411–2416, Apr. 2024, doi: 10.1109/TED.2024.3371945.