Computationally efficient compact model for ferroelectric field-effect transistors to simulate the online training of neural networks

Darsen Duane Lu\textsuperscript{1,3}, Sourav De\textsuperscript{1}, Mohammed Aftab Baig\textsuperscript{1}, Bo-Han Qiu\textsuperscript{1} and Yao-Jen Lee\textsuperscript{2,3}

\textsuperscript{1} Institute of Microelectronics, National Cheng Kung University (NCKU), Tainan, Taiwan, Republic of China
\textsuperscript{2} Taiwan Semiconductor Research Institute, Hsinchu, Taiwan, Republic of China

E-mail: darsenlu@mail.ncku.edu.tw

Received 6 April 2020, revised 26 May 2020
Accepted for publication 11 June 2020
Published 23 July 2020

Abstract

In this paper, a compact drain current formulation that is simple and adequately computationally efficient for the simulation of neural network online training was developed for the ferroelectric memory transistor. Tri-gate ferroelectric field-effect transistors (FETs) with Hf\textsubscript{0.5}Zr\textsubscript{0.5}O\textsubscript{2} gate insulators were fabricated with a gate-first high-k metal gate CMOS process. Ferroelectric switching was confirmed with double sweep and pulse programming and erasure measurements. Novel characterization scheme for drain current was proposed with minimal alteration of ferroelectric state in subthreshold for accurate threshold voltage measurements. The resultant threshold voltage exhibited highly linear and symmetric across multilevel states. The proposed compact formulation accurately captured the FET gate-bias dependence by considering the effects of series resistance, Coulomb scattering, and vertical field dependent mobility degradation.

Keywords: ferroelectric field-effect transistor, nonvolatile memory, compact modeling, neuromorphic computing

(Some figures may appear in colour only in the online journal)

1. Introduction

Ferroelectric (Fe) field-effect transistors (FETs) with hafnium-based dielectrics have been heavily researched as a promising device for logic, memory, and neuromorphic applications since the proposal of negative capacitance FETs (NCFETs) for low–power electronics \cite{1} and the discovery of CMOS–compatible and highly–scalable HfZrO (HZO) Fe material \cite{2}. In particular, FeFETs are promising for deep neural network (NN) applications because of their low write latency, low cycle-to-cycle variability, high endurance, and highly symmetric and linear multilevel switching characteristics \cite{3–8}. To apply FeFET technology to macro–and system–level benchmarking \cite{9–11} and to design/synthesis tools \cite{12}, an analytical model is required. A conventional SPICE-oriented compact model can accurately model device output behavior. New compact models for FeFETs (including NCFETs) were developed \cite{13–15}. However, the simulation of on-chip NN training requires millions of sample data to pass through hundreds of millions of synaptic devices. Full SPICE simulation with a traditional compact model is too time consuming. Therefore, neuromorphic simulators such as that in \cite{16} adopted and implemented a simple formulation for the potentiation and depression of resistive memory (RRAM). In \cite{3}, the formulation developed for RRAM was utilized for FeFET online NN training. Unfortunately, this approach fails.
to model FeFET-specific properties such as gate bias dependence. In this paper, we present a computationally efficient compact model for NN that is sufficiently simple for evaluation during a large-scale neural network training simulation. In addition, it captures FeFET’s gate bias dependence and is valid from the subthreshold to the strong inversion region.

2. Device fabrication and measurements

Tri-gate Fe FinFETs were fabricated on a silicon-on-insulator wafer with a gate-first process [5]. First, the silicon layer was thinned down by oxidation to approximately 30 nm, which was approximately equal to the final fin height. The active region (including the silicon fin, source, and drain pads) was patterned through e-beam lithography and etching. Subsequently, the gate stack was formed through dry oxidation for a thin (approximately 8 Å) interfacial SiO2 layer, atomic layer deposition for the 10 nm thick HZO layer (HfO2.5Zr0.5O2), and physical vapor deposition for the TiN gate layer. Post deposition rapid thermal annealing for 30 s was performed to foster the formation of orthorhombic crystal phase in HZO for ferroelectricity. The gate was patterned, again through e-beam lithography, with the shortest target gate length of 40 nm. Ion implantation for source and drain formation was performed separately for n-type (nFET) and p-type (pFET) devices. Dopant activation was accomplished through microwave annealing, which, according to our previous study, does not affect Fe properties [17]. The final device cross-sectional transmission electron microscopy image is displayed in figure 1.

The characterization of the drain current ($I_d$) while sweeping the gate voltage ($V_{gs}$) over a wide range is convenient for compact model development. However, accurate characterization of the multilevel $I_d-V_{gs}$ during potentiation (programming) and depression (erasure) of the FeFET is not straightforward because the Fe state may be altered during direct current (DC) $I_d-V_{gs}$ sweeps.

For these reasons, we applied a special pulsing scheme to measure $I_d-V_{gs}$ characteristics (figure 2). A 1 µs reset pulse ($V_{reset}$) ensured that the FeFET returned to its erased (high threshold voltage [$V_{th}$]) state. Another 1 µs program pulse ($V_{pgm}$) set the device to the target polarization state. Subsequently, $V_{gs}$ was swept from $V_{start}$ to $V_{stop}$ to measure the transfer characteristics. The measured device has 10 fins, fin height of 30 nm, and fin width of 50 nm.

Selection of $V_{start}$ is crucial. When $V_{start}$ lies in the subthreshold region, the electric field is, in principle, nearly zero for the fully-depleted FinFET, with little chance of Fe switching. We thus tried to keep $V_{start} < V_{th} - 1.0$ V to ensure that $I_d-V_{gs}$ covered the subthreshold region over a 1 V range and $V_{th}$ measurement was accurate. As a sanity check, for a multifin device with $L_g = 70$ nm, we applied the waveform as shown in figure 2 without $V_{pgm}$/$V_{ers}$ pulses, fixed $V_{reset}$ to $-5.0$ V and varied $V_{start}$ from $-0.5$ V to $-4.5$ V in steps of $-0.5$ V while keeping $V_{stop}$ at $+4.0$ V. $V_{th}$ was identically 0.4 V, and a negligible change in $I_d-V_{gs}$ characteristics as function of $V_{start}$ was seen. This experiment shows that gate voltage in the range of $-4.5$ V to $-0.5$ V does not alter Fe polarization state. However, it is still possible that Fe polarization state of HZO changes during the portion of $I_d-V_{gs}$ sweep above $-0.5$ V. When $V_{start}$ was set to $-5.0$ V or less, we begin to observe change in $I_d-V_{gs}$, which is erasure behavior.

Figure 3(a) shows the $I_d-V_{gs}$ series measurements as we varied $V_{pgm}$ in 0.2 V steps. For clarity we only show $I_d-V_{gs}$ curves in 1.0 V steps (4 out of 5 curves are not displayed). The curves merged together at approximately 4.0 V. We further plot the extracted $V_{th}$ as function of $V_{pgm}$ (figure 3(b)). It can be observed that $V_{th}$ was nearly a linear function of $V_{pgm}$. Furthermore, to understand the evolution of $I_d-V_{gs}$ as a function of $V_{pgm}$, we plot $\Delta V_{gs}$, the horizontal shift of the $I_d-V_{gs}$ curve, by using the shift between $I_{gs}(5)$ ($I_d$ of the 5th sweep, with $V_{pgm} = 3.6$ V) and $I_{gs}(20)$ ($I_d$ of the 20th sweep, with $V_{pgm} = 6.6$ V) as an example (figure 4). More precisely,
followed a linear relationship so precisely. We therefore reject this hypothesis.

The characterization technique proposed here is similar to the measurement method in [19], where $I_d-V_{gs}$ sweep with lower magnitude follows program/erase pulse application, except that we have added a reset pulse to force the Fe material back to a polarized state. In both cases, special care must be taken to account for the possible $V_{th}$ shift during $I_d-V_{gs}$ sweep.

3. Neuromorphic-oriented compact model

With solid understanding of characterization results, we developed a model for FeFET memory device with simple yet physically sound expressions for use in a neuromorphic simulator [9-11]. We adopted the basic drain current formulation for the industry standard Berkeley Short-Channel Insulated-Gate Field-Effect Transistor Model 4 (BSIM4) compact model [19] in the linear region as follows:

$$I_d = \frac{K \cdot V_{gs eff} \cdot V_{ds}}{1 + UA \cdot (V_{gs eff} + \text{DELTA})^2 + Ud \frac{V_{ds}}{V_{gs eff} + V_{th}} + RDS \cdot K \cdot V_{gs eff}}$$

where the term with parameters UA, EU and DELTA accounts for vertical field dependent mobility degradation; that with parameter UD accounts for Coulomb scattering in the moderate and weak inversion regions (mobMod = 3 option in BSIM4); and that with RDS accounts for the source and drain series resistance. Note that the original 2 $V_{th}$ term in the effective vertical field calculation is absorbed into DELTA for simplicity.

In addition, to cover both weak and strong inversion regions, we adopted the effective gate overdrive function from BSIM4 [18]:

$$\Delta V_g = V_{id(5)}(I_d(20)) - V_g$$

where $V_{id(5)}(I_n)$ function finds the gate voltage at which drain current $I_d(5)$ is equal to $I_n$. An excellent agreement to a linear model in the on state (strong inversion) was identified.

The likely explanation for the linear relationship in figure 4 is as follows. For the 5th $I_d-V_{gs}$ curve ($I_n(5)$), as $V_g$ increased beyond $V_{th}$, finite electric field across the HZO insulator caused gradual switching of ferroelectric dipoles, and thus the reduction of $V_{th}$. Such reduction in $V_{th}$ translated to a decrease in $\Delta V_g$, the horizontal shift between $I_d(5)$ and $I_d(20)$. The relationship was linear because gate overdrive, electric field across HZO, Fe polarization change, $V_{th}$ shift, and $\Delta V_g$ were all proportional to one another.

An alternative explanation for the convergence of curves at high gate voltage is the degradation of mobility due to surface electric field from ferroelectric dipoles. However, in that case the relationship between $V_g$ and $\Delta V_g$ would not have
RMS Error is defined as:

\[ \text{RMS Error} = \sqrt{\frac{\sum (y_m - y_s)^2}{N}} \]

where \( y_m \) and \( y_s \) are measured and simulated data, respectively, and \( N \) is the total number of samples. SPICE model is more accurate in the moderate to weak inversion region, as shown in figure 5. Such advantage of SPICE only reduced the RMS error slightly but will likely result in better static leakage prediction.

Note that equations (2) and (3) models only one \( I_d-V_{gs} \) curve. To model the entire series of curves at different Fe states, we adopted linear threshold shifting based on our understanding of linear behavior (figure 4). The effective gate voltage was expressed as:

\[ V_{eff} = V_{gs} + \left[ (FA \cdot P + FB) - (FC \cdot P - FD) \right] \cdot V_{th} \]

where \( n \) is the sub-threshold ideality factor, and \( V_{OFF} \) is a parameter for the \( V_{th} \) shift in the weak inversion region only.

Table 1. Calibrated FeFET Model Parameters.

| Parameter | Value |
|-----------|-------|
| \( K \)   | \( 8.27 \times 10^{-4} \) |
| \( UA \)  | 1.1   |
| \( DELTA \) | 0.604 |
| \( EU \)  | 1.26  |
| \( NFACTOR \) | 4.02 |
| \( VOFF \) | -0.41 |

Table 2. Comparison of FeFET Technologies.

| Reference | Linearity (\( \alpha_p/\alpha_d \)) | On-state resistance (\( \Omega \)) | # levels | Active Area (\( \mu m^2 \)) |
|-----------|----------------------------------|---------------------------------|----------|--------------------------|
| [3]       | +1.75/ +1.46                     | 559 kΩ                          | 32       | 1190                     |
| [4]       | +1.73/ +1.86                     | 50 kΩ                           | 32       | –                        |
| [6]       | +0.44/ –0.38                     | 3.5 kΩ                          | 35       | –                        |
| [7]       | +0.12/ –0.09                     | 0.22 kΩ                         | 32       | –                        |
| [8]       | +1.22/ –1.75                     | 12.5 kΩ                         | 256      | 0.0034                   |
| This work | +0.67/ –1.13                     | 11.5 kΩ                         | 27       | 0.055                    |

\[ V_{gs} = \frac{nqT}{\eta} \ln \left[ 1 + \exp \left( \frac{V_{gs} - V_{th}}{2nkT/q} \right) \right] \]

\[ V_{vsteff} = \frac{nkT}{q} \cdot \ln \left\{ \frac{1 + \exp \left( \frac{V_{gs} - V_{th}}{2nkT/q} \right)}{1 + n \cdot \exp \left( -\frac{V_{gs} - 2V_{OFF}}{2nkT/q} \right)} \right\} \]

\[ V_{eff} = V_{gs} + \left[ (FA \cdot P + FB) - (FC \cdot P - FD) \right] \cdot V_{th} \]

Figure 5. Comparison of SPICE model and analytical expression (equation (2)). SPICE model shows better accuracy, particularly in the weak to moderate inversion regions.

4. NN applications

Symmetric and linear potentiation/depression characteristics are crucial for the online training of NNs [9]. By selecting suitable starting pulse amplitudes (+2.6 V for potentiation; and −0.8 V for erase), excellent symmetry and linearity can be achieved (figure 6). In figure 6, the modeled conductance with parameters given in table 1 is also plotted. There is discrepancy at high and low conductance states. Such discrepancy can be corrected if we specifically optimize for \( V_{gs} = 1.0 \) V. Alternatively, the model in [16] can also fit the measured curve well in strong inversion region and can be used for extracting the non-linearity factor for potentiation (\( \alpha_p \)) and depression (\( \alpha_d \)) [3].

In table 2, we compare this work with other FeFET technologies. The linearity of the FeFET synaptic device is quantified using the non-linearity parameter, \( \alpha \) [3]. In general, FeFETs present better linearity than does RRAM, with a lower \( \alpha \) [9]. However, such a high linearity is only possible when the program/erase voltage varies with pulse number. This conclusion is consistent with findings in the literature [3–6].
Figure 6. Normalized channel conductance when program/erase pulses are applied. A total of 28 program pulses and 24 erase pulses were applied, which is equivalent to 4 or 5 bits. ($W_{eff} = 1.1 \mu m$, $L_g = 50 nm$) The model based on parameters in table 1 exhibits slight discrepancy when compared to measured data. Such discrepancy is removed by re-fitting at $V_{gs} = 1.0 V$. The data is also fitted with the formula in [16] (in strong inversion only) to evaluate $\alpha_p/\alpha_n$.

5. Conclusions

We have proposed a new method to characterize ferroelectric FETs to ensure the device threshold voltage itself is not altered during $I_{d}$–$V_{gs}$ sweep. Tri-gate ferroelectric FETs are fabricated and characterized with the new method and exhibit excellent symmetry and linearity during multi-level operation, ferroelectricity-dominated switching as opposed to charge trapping, making the device very promising for neuromorphic applications. In order to evaluate the device performance in neural networks, a computationally efficient FeFET-specific analytical model is developed based on the measurement results. With only slight degradation in accuracy compared to SPICE, but significantly simpler, the model can be readily adopted to simulate online training of deep neural networks.

Acknowledgments

This work was jointly supported by the Ministry of Science and Technology (Taiwan) grant MOST–108–2634–F–006–08 and is part of research work by MOST’s AI Biomedical Research Center. We are grateful to the Taiwan Semiconductor Research Institute for nanofabrication facilities and services and to Dr. Wen-Jay Lee and Nan-Yow Chen of the National Center for High-Performance Computing for helpful suggestions on AI computation. This manuscript was edited by Wallace Academic Editing.

ORCID iDs

Darsen Duane Lu https://orcid.org/0000-0002-1956-6093
Sourav De https://orcid.org/0000-0002-1930-8799
Mohammed Aftab Baig https://orcid.org/0000-0002-4727-1006
Bo-Han Qiu https://orcid.org/0000-0003-3015-0811
Yao-Jen Lee https://orcid.org/0000-0001-9643-111X

References

[1] Salahuddin S and Datta S 2008 Use of negative capacitance to provide voltage amplification for low power nanoscale devices Nano Lett. 8 405–10
[2] Müller J, Böscke T S, Schröder U, Mueller S, Bräuhau D, Böttger U, Frey L and Mikolajick T 2012 Ferroelectricity in simple binary ZrO$_2$ and HfO$_2$ Nano Lett. 12 4318–23
[3] Jerry M, Chen P-Y, Zhang J, Sharma P, Ni K, Yu S and Datta S, 2017 Ferroelectric FET analog synapse for acceleration of deep neural network training Proc. IEEE Int. Electron Devices Meeting (IEDM) (San Francisco, CA, USA) pp 139–42
[4] Oh S, Kim T, Kwak M, Song J, Woo J, Jeon S, Yoo I K and Hwang H 2017 HfZrO$_x$-based ferroelectric synapse device with 32 levels of conductance states for neuromorphic applications IEEE Electron Device Lett. 38 732–5
[5] De S, Baig M A, Qiu B-H, Lu D, Sung P-J, Hsueh F-K, Lee Y-J and Su C-J 2020 Tri-gate ferroelectric FET characterization and modelling for online training of neural networks at room temperature and 233K Device Research Conf. (Columbus, OH, USA) pp 1–2

Figure 7. DC double-sweep measurement of an $L_g = 70$ nm FeFET. Counter-clockwise current-voltage curves for nFET indicated Fe switching. Inset shows device conductance with program and erase pulse applications. (For clarity, only $I_d$ during the 2V read pulse is shown).

In addition, unlike RRAM, FeFETs’ synaptic properties are highly dependent on $V_{gs}$. The compact model presented here is useful for evaluating the figure of merits such as power consumption, circuit delay, and training accuracy as a function of $V_{gs}$, and is ultimately used for neuromorphic circuit design.

To rule-out memory operations due to charge-trapping, double-sweep nFET $I_{d}$–$V_{gs}$ measurements were performed. Counter-clockwise $I_{d}$–$V_{gs}$ indicated Fe switching (figure 7) [23]. We confirmed that $V_{th}$ decreased/increased with program/erase pulse applications, respectively (figure 7 inset).
[6] Hu V P-H, Lin -H-H, Zheng Z-A, Lin Z-T, Lu Y-C, Ho L-Y, Lee Y-W, Su C-W and Su C-J 2019 Split-gate FeFET (SG-FeFET) with dynamic memory window modulation for non-volatile memory and neuromorphic applications Symp. on VLSI Technology (Kyoto, Japan) pp 134–5

[7] Peng Y et al 2019 Nanocrystal-embedded-insulator (NEI) ferroelectric field-effect transistor featuring low operating voltages and improved synaptic behavior IEEE Electron Devices Lett. 40 1933–6

[8] Chung W, Si M and Ye P D 2018 First demonstration of Ge ferroelectric nanowire FET as synaptic device for online learning in neural network with high number of conductance state and \( G_{\text{max}} / G_{\text{min}} \) Proc. IEEE Int. Electron Devices Meeting (IEDM) (San Francisco, CA, USA) pp 344–7

[9] Chen P-Y, Peng X and Yu S 2017 NeuroSim+: an integrated device to algorithm framework for benchmarking synaptic devices and array architectures Proc. IEEE Int. Electron Devices Meeting (IEDM) (San Francisco, CA, USA) pp 135–8

[10] Agarwal S, Gedrim R B J, Hsia A H, Hughart D R, Fuller E J, Talin A A, James C D, Plimpton S J and Marinella M J 2017 Achieving ideal accuracies in analog neuromorphic computing using periodic carry Symp. on VLSI Technology (Kyoto, Japan) pp 174–5

[11] Le H-H, Hong W-C, Du J-W, Lin T-H, Hong Y-X, Chen I-H, Lee W-J, Chen N-Y and Lu D D 2020 Ultralow power neuromorphic accelerator for deep learning using Ni/HfO2/TiN resistive random access memory IEEE Electron Devices Technology and Manufacturing (EDTM) (Penang, Malaysia) pp 1–4

[12] Lin W-F, Tsai D-Y, Tang L, Hsieh C-T, Chou C-Y, Chang P-H and Hsu L 2019 ONNC: A compilation framework connecting ONNX to proprietary deep learning accelerators IEEE Int. Conf. on Artificial Intelligence Circuits and Systems (AICAS) (Hsinchu, Taiwan) pp 214–8

[13] Duarte J P, Khandelwal S, Khan A I, Sachid A, Lin Y-K, Chang H-L, Salahuddin S and Hu C 2016 Compact models of negative-capacitance FinFETs: lumped and distributed charge models Proc. IEEE Int. Electron Devices Meeting (IEDM) (San Francisco, CA, USA) pp 754–7

[14] Pahwa G, Dutta T, Agarwal A and Chauhan Y S 2017 Compact model for ferroelectric negative capacitance transistor with MFS structure IEEE Trans. Electron Devices 64 1366–74

[15] Ni K, Jerry M, Smith J A and Datta S 2020 A circuit compatible accurate compact model for ferroelectric-FETs Symp. on VLSI Technology (Honolulu, HI, USA) pp 131–2

[16] Yu S, Chen P-Y, Cao Y, Xia L, Wang Y and Wu H 2015 Scaling-up resistive synaptic arrays for neuro-inspired architecture: challenges and prospect Int. Electron Devices Meeting (IEDM) (San Francisco, CA, USA) pp 451–4

[17] Sung P-J et al 2019 Fabrication of Ω-gated negative capacitance finFETs and SRAM Int. Symp. on VLSI Technology, Systems and Applications (VLSI-TSA) (Hsinchu, Taiwan) pp 106–7

[18] Dabhi C K, Parihar S S, Agrawal H, Paydavosi N, Morshed T H, Lu D D, Yang W, Dunga M V, Xi X, He J, Liu W, Cao K, Jin X, Ou J J, Chen M, Chauhan Y S, Niknejad A M and Hu C 2017 BSIM4 4.8.1 user’s manual (available online at: http://bsim.berkeley.edu/models/bsim4/)

[19] Breyer E T, Mulaosmanovic H, Nikolajick T and Slesazeck S 2017 Reconfigurable NAND/NOR logic gates in 28 nm FD-SOI FeFET technology Int. Electron Devices Meeting (IEDM) (San Francisco, CA, USA) pp 669–72

[20] Lu D, De S and Qiu B-H 2020 Ferroelectric FET compact model for neuromorphic IEEE Dataport (http://dx.doi.org/10.21227/86qq-5n18)

[21] Dunga M V, Lin C-H, Lu D D, Xiong W, Clevelin C R, Patruno P, Hwang J-R, Yang F-L, Niknejad A M and Hu C 2007 BSIM-MG: a versatile multi-gate fet model for mixed-signal design Symp. on VLSI Technology (Kyoto, Japan) pp 60–61

[22] ProPlus Electronics Co., Ltd. 2019 BSIMProPlus™ Model Extractor Device Modeling Guide, Version 3.19.1

[23] Alam M N K, Kaczer B, Ragnarsson L-A, Popovici M, Rzepa G, Horiguchi N, Heyns M and van Houdt J 2019 On the characterization and separation of trapping and ferroelectric behavior in HfZrO FET IEEE J. Electron Devices Soc. 7 855–62