FeFET-Based Synaptic Cross-Bar Arrays for Deep Neural Networks: Impact of Ferroelectric Thickness on Device-Circuit Non-Idealities and System Accuracy

C. Wang, J. Victor, A. K. Saha, X. Chen, M. Si, T. Sharma, K. Roy, P. D. Ye, and S. K. Gupta
Purdue University, West Lafayette, IN, USA, Email: wang4015@purdue.edu / Phone: (765) 409-9583

Introduction: Ferroelectric transistors (FeFETs) are amongst the most promising candidates for synaptic cross-bar array designs used for in-memory computation (IMC) of matrix-vector multiplications (MVMs) (Fig. 1) in deep neural networks (DNNs). However, FeFETs have several non-ideal attributes such as non-linearities, variations and faults. These, along with circuit non-idealities in the cross-bar array [1], produce output currents that may deviate from the expected (ideal) currents (Fig. 1). This can lead to computation errors, impairing system accuracy. Therefore, FeFETs need to be judiciously designed to minimize the impact of the device-circuit non-idealities on DNN accuracy. While several works have explored the implications of FeFETs in DNNs [2-3], analysis on the impact of FE thickness (T_{FE}) on system accuracy, accounting for the device-circuit non-idealities is lacking. *To that end, we analyze the impact of T_{FE} on the characteristics of FeFET-based synaptic devices based on physical models calibrated to experiments (Fig. 2). We present how T_{FE} scaling affects the IMC of MVMs and DNN accuracy, considering device-circuit non-idealities including variations and faults in FeFETs and wire/sink/driver resistances in the cross-bar array.*

Modeling and Experiments: First, we fabricate and characterize metal-ferroelectric-insulator-metal (MFIM) stacks (details in [4]) comprised of HZO and Al_2O_3 for three different T_{FE} (=10nm, 7nm and 5nm). Based on the experiments, we calibrate our in-house phase-field models [4], which utilize time-dependent Ginzburg Landau and Poisson's equations to model MFIM and FeFETs. Fig. 3 illustrates the effective background permittivity (ε_r) and coercive voltage (V_C) of HZO versus T_{FE}, showing a close match between experiments and phase-field model. It can be observed that ε_r increases as T_{FE} is scaled. This is due to the transformation of in-plane to out-of-plane electric fields near the domain walls on the application of voltage [4]. As the number of domain walls increases with T_{FE} scaling, this effect is enhanced [4]. We also observe a non-linear decrease in V_C with T_{FE} scaling due to multi-domain effects. We capture these trends in FeFET models by obtaining ε_r and V_C versus T_{FE} for the gate stack of FeFET (10/7/5nm HZO + interfacial SiO_2) (Fig. 3). We utilize these parameters as a function of T_{FE} in our compact model of FeFET based on Preisach equations [5]. The compact model is calibrated to experiments on FeFETs in [6], (Fig. 4).

The Impact of T_{FE} on FeFET-Synapse: Fig. 5 shows the transfer characteristics of FeFETs for different T_{FE}, with 2-bits (4-levels) stored per FeFET. The reset state encodes synaptic weight=0, while the three set states encode weight=1, 2 and 3. To compute the product of the weight matrix ($[w_{ij}]$) with the input vector ($[In_i]$) which is = $\Sigma w_{ij}In_i$, input voltage is applied on the gate of FEFET (V_{GS}), with V_{GS}=0 and V_{READ} encoding input of 0 and 1, respectively. Each FeFET produces a current which corresponds to the scalar multiplication $w_{ij}In_i$. Fig. 5 shows two key effects of T_{FE} scaling on the current for scalar product = 0 (I_{OUT0}): (1) I_{OUT0} for weight=1/2/3 and input=0 decreases as T_{FE} scales. This is due to reduction in short channel effects. (2) I_{OUT0} for weight=0 and input=1 increases as T_{FE} scales, which is because of the shrinking memory window. Note, in our analysis, the first effect captures the dependence of ε_r on T_{FE} (Fig. 3(a)) and the second effect accounts for the non-linear dependence of V_C on T_{FE} (Fig. 3(b)). Since I_{OUT0} should be as small as possible for minimum effect of non-idealities, T_{FE} scaling leads to two opposing effects in the context of IMC robustness. To understand their effect on the overall system accuracy, we perform system analysis next.

The Impact of T_{FE} on DNN Accuracy: Our evaluation methodology is based on GENIEx [1], which allows us to obtain the non-ideal output accounting for the interactions of FeFET characteristics with the circuit non-idealities in a cross-bar array (parameters in Fig. 6). Fig. 7 compares the accuracy of ResNet-20 for CIFAR-10 dataset for different T_{FE} (10nm/7nm/5nm). For a nominal design (without variations and faults), accuracy of T_{FE}=7nm and 10nm is close to software (ideal) accuracy while T_{FE}=5nm has the lowest accuracy. This is due to high I_{OUT0} (input=1, weight=0) for T_{FE}=5nm, leading to the largest impact of non-idealities. When we consider random variations in synaptic conductance and stuck-at faults in FeFETs, T_{FE}=5nm shows the largest degradation in accuracy compared to the nominal case. The accuracy for T_{FE}=7nm is slightly more than T_{FE}=10nm when variations are considered. This is because T_{FE}=7nm has a lower I_{OUT0} (input=0 weight=1/2/3) compared to T_{FE} = 10nm (Fig. 5), while still maintaining small I_{OUT0} (input=1 weight=0) due to reasonably large memory window. This suggests that in our analysis, T_{FE} = 7nm achieves a balanced trade-off between the two opposing effects of T_{FE} scaling, leading to a high resilience to non-idealities.

Summary: We analyzed the effect of T_{FE} on the characteristics of FeFET synaptic devices and system accuracy considering device-circuit non-idealities. We showed that T_{FE} scaling leads to increase in ε_r, a non-linear increase in V_C, and reduction in leakage and memory window of FeFETs. This has two opposing effects on I_{OUT0}, leading to non-monotonic effect of T_{FE} on system accuracy (especially when variations are considered).

References: [1] Chakraborty *et al*, *DAC* 2020 [2] Jerry *et al*, *IEDM* 2017 [3] Saito *et al*, *IEDM* 2021 [4] Saha *et al*, *IEDM* 2020 [5] Saha *et al*, *DRC* 2018 [6] Ni *et al*, *IEDM* 2018 [7] Moon *et al*, *Intel Tech.* 2008 [8] Mistry *et al*, *IEDM* 2007.

Acknowledgements: This work was supported by SRC/DARPA-funded C-BRIC Center.

2023 Device Research Conference (DRC) | 979-8-3503-2310-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/DRC58590.2023.10187042

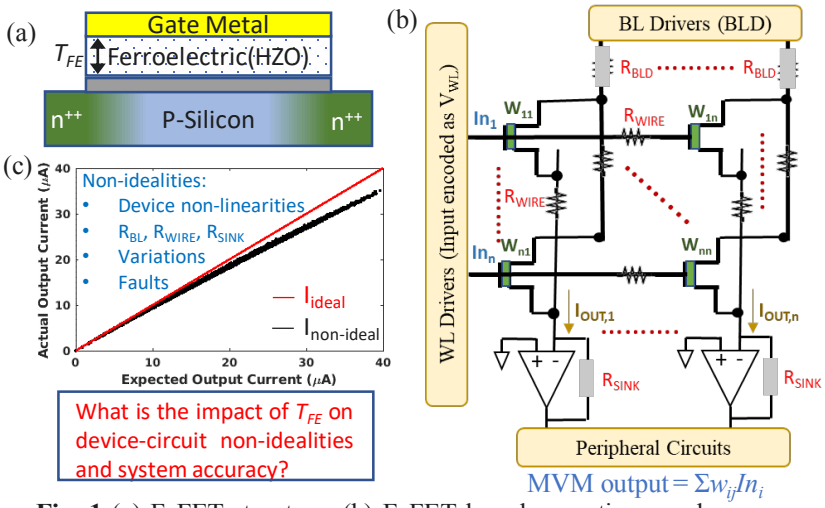

Fig. 1 (a) FeFET structure. (b) FeFET-based synaptic cross-bar array for DNN accelerators. (c) Actual output currents ($I_{non\text{-}ideal}$) deviate from expected output currents (I_{ideal}).

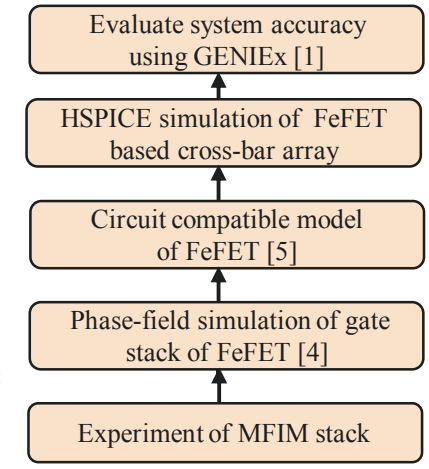

Fig. 2 Overview of simulation flow from evaluation of FeFET-based synapses, cross-bar arrays and DNNs.

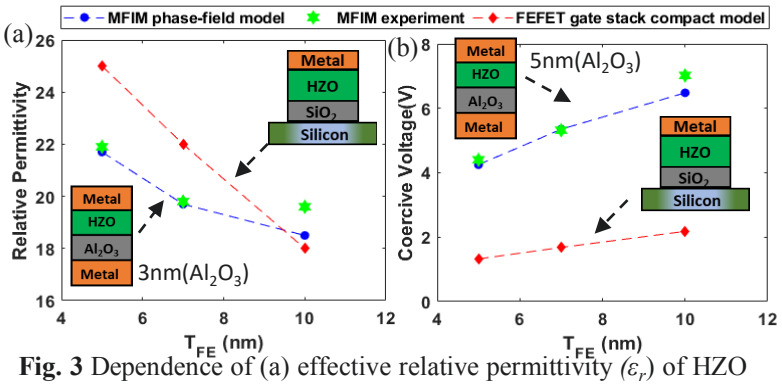

Fig. 3 Dependence of (a) effective relative permittivity (ε_r) of HZO and (b) coercive voltage (V_C) on T_{FE} for MFIM stack (phase-field simulation matches with experiment) and gate stack of FeFET (for compact model). ε_r increases and V_C decreases as T_{FE} decreases.

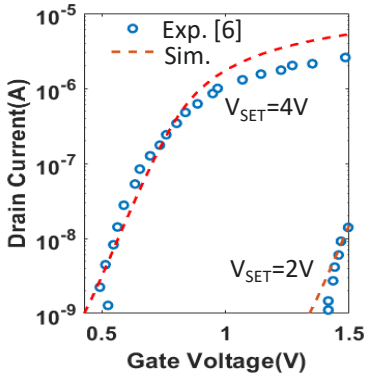

Fig. 4 Simulated FeFET I_{DS}-V_{GS} showing good match with experiment (T_{FE}=10nm, L=450nm, W=450nm, P_S=30μC/cm², P_R=30μC/cm²).

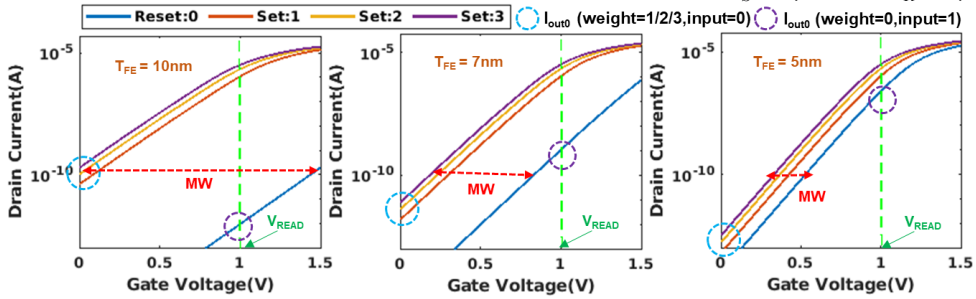

Fig. 5 FeFET I_{DS}-V_{GS} characteristics for different T_{FE} (10nm/7nm/5nm) and weight (0/1/2/3) at V_{DS}=0.25V showing memory window and leakage reduce. This reduces I_{out0} for input=0 and weight=1/2/3 but increases I_{out0} for input=1 and weight=0 as T_{FE} is scaled (L=45nm, W=67.5nm).

Technology	45nm	Metal Pitch	160nm [8]
Array Size	64*64	Gate Pitch	160nm [8]
R_{BL}	500Ω	V_{BL}	0.25V
R_{SINK}	100Ω	V_{WL}	1V
R_{WIRE}	3.3Ω/μm [7]	R_{VIA}	2.85Ω
Bits/input signal	1b	Bits/device	2b

Fig. 6 Parameters for FeFET cross-bar array simulation.

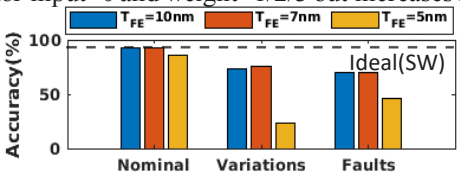

Fig. 7 Accuracy for different T_{FE} considering variations (σ/μ=10%) and faults (0.25% stuck-at 0, 0.25% stuck-at 1) for CIFAR-10 dataset on ResNet-20. T_{FE} =7nm leads to minimum impact on device-circuit non-idealities.

979-8-3503-2310-8/23/$31.00 ©2023 IEEE