Extreme Dimensional Scaling of Ferroelectric Tunnel Junctions toward Stochastic Probabilistic Computing

Laeyong Jung, and Hyunsang Hwang*
Department of Materials Science and Engineering, Pohang University of Science and Technology (POSTECH), Pohang 37673, Korea
*Corresponding E-mail: hwanghs@postech.ac.kr

***Abstract*— We demonstrate the extreme dimensional scaling of $Hf_xZr_{1-x}O_2$ (HZO)-based ferroelectric tunnel junctions (FTJs) with an ultra-thin 0.9 nm HZO layer down to 10 nm × 10 nm. Aggressive area scaling suppresses premature dielectric breakdown, enabling high electric fields that promote orthorhombic phase formation and secure a sufficient tunneling electroresistance ratio (>10). While large-area devices exhibit negligible noise due to averaging over numerous domains, extreme area scaling reveals random telegraph noise arising from electron capture and emission at oxygen vacancy traps at the HZO and TiON interface. The stochastic switching probability is continuously tunable from 0.07 to 0.87 by modulating the write pulse amplitude, which controls the remnant polarization and thereby trap occupancy, following a sigmoid-like response. Using the measured sigmoid response as a probabilistic-bit (p-bit) activation function, a weighted max-cut optimization achieves the global minimum in 82.4% of runs versus 54.2% for deterministic optimization, establishing extreme area scaling as a promising pathway toward hardware p-bit computing.**

Keywords—Ferroelectric tunnel junction memory, scaling, probabilistic computing, Zr-doped hafnium oxide

I. INTRODUCTION

Ferroelectric tunnel junctions (FTJs) based on $Hf_xZr_{1-x}O_2$ (HZO) have emerged as promising candidates for next-generation non-volatile memory, offering ultra-fast switching speed, high integration density, low operating voltage, and excellent CMOS compatibility [1, 2]. However, achieving sufficient tunneling current density requires aggressive thickness scaling, which simultaneously promotes the non-ferroelectric tetragonal phase at the expense of the ferroelectric orthorhombic phase, resulting in degraded remnant polarization and a reduced on/off ratio [3, 4]. This thickness-performance trade-off represents a fundamental bottleneck in FTJ scaling.

Beyond thickness, large device area introduces additional limitations. The high defect density inherent to large-area ultra-thin HZO films increases the probability of premature dielectric breakdown, preventing reliable high-field operation [5]. Moreover, the macroscopic response of large-area devices reflects statistical averaging over numerous ferroelectric domains, obscuring the intrinsic stochastic behavior at the nanoscale that is essential for emerging probabilistic-bit (p-bit) computing paradigms [6].

In this work, we address these challenges through extreme dimensional scaling of HZO-based FTJs. By combining a microwave annealing (MWA) process with aggressive area scaling down to 10 nm × 10 nm, we reduce the total defect count and enhance the breakdown field, enabling orthorhombic phase formation and an on/off ratio exceeding 10 in a 0.9 nm HZO FTJ. Furthermore, extreme area scaling unveils pronounced RTN that is entirely suppressed in large-area counterparts, with switching probability continuously tunable via write pulse amplitude through polarization-controlled trap occupancy at the HZO and TiON interface. The sigmoid-like probability response demonstrates the feasibility of dimensionally scaled HZO FTJs as hardware p-bits, establishing a viable pathway from deterministic memory operation toward probabilistic computing.

This work was supported by the National Research Foundation of Korea (NRF) under Grant 2018R1A3B1052693.

II. DEVICE FABRICATION AND PROCESS OPTIMIZATION

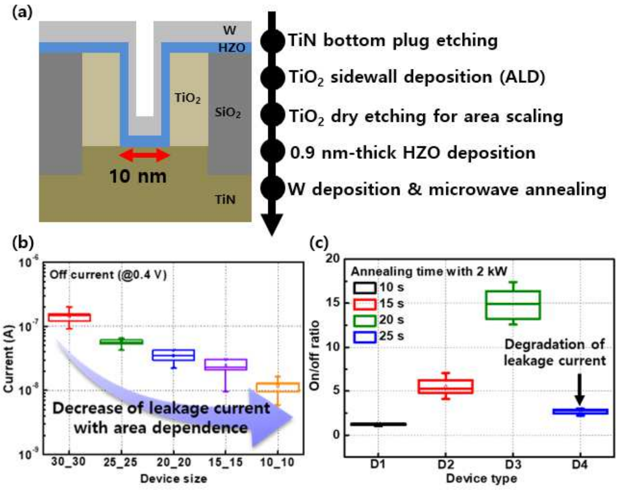

Fig. 1. (a) Schematic of the device structure and fabrication process of the HZO-based FTJ. (b) Off-state leakage current as a function of device area, showing a decrease with reducing device size down to 10 nm x 10 nm. (c) On/off ratio as a function of MWA conditions at 2 kW, exhibiting that an optimal annealing time yields the highest on/off ratio.

Fig. 1(a) illustrates the device structure and fabrication process of the HZO-based FTJs. The fabrication begins with TiN bottom plug etching, followed by TiO_2 sidewall deposition via atomic layer deposition (ALD). TiO_2 dry etching is then performed to define the device area, enabling aggressive area scaling down to 10 nm × 10 nm. Subsequently, a 0.9 nm-thick HZO layer is deposited, followed by W top electrode deposition and microwave annealing (MWA) to crystallize the ferroelectric orthorhombic phase. As shown in Fig. 1(b), the off-state leakage current measured at 0.4 V exhibits a clear area dependence, decreasing systematically as the device area is reduced. This trend confirms that the total defect count within the HZO layer scales with device area, and that smaller devices contain fewer defects, effectively suppressing leakage current. This area-dependent leakage reduction validates the successful fabrication of extremely scaled 10 nm × 10 nm FTJ devices and supports reliable high-field operation without premature dielectric breakdown. Fig. 1(c) presents the on/off ratio as a function of MWA conditions. Among the annealing times investigated at 2 kW, the 20 s condition yields the highest on/off ratio, attributed to optimal orthorhombic phase formation under this thermal budget. Annealing beyond 20 s, however, results in degraded on/off ratio due to excessive leakage current induced by over-annealing, which compromises the tunneling barrier integrity. These results establish 20 s MWA as the optimal process condition, balancing ferroelectric phase formation and leakage suppression for reliable FTJ operation.

III. Ultrathin FTJ Memory Operation with Area Scaling

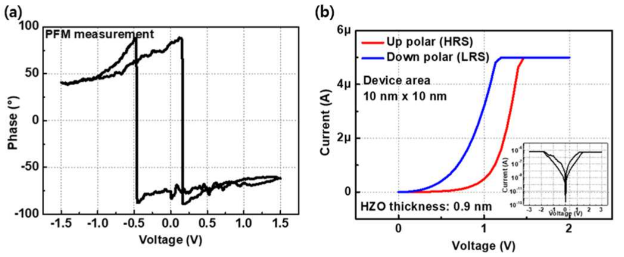

Fig. 2. (a) PFM phase-voltage hysteresis loop measured on the 0.9-nm HZO test structure, confirming robust intrinsic ferroelectricity. (b) Current-voltage characteristics of the 10 nm × 10 nm FTJ with 0.9 nm HZO, showing distinct resistance states at a read voltage of 0.4 V.

Fig. 2(a) shows the piezoresponse force microscopy (PFM) phase-voltage hysteresis loop of the fabricated device, confirming the ferroelectric nature of the 0.9 nm HZO layer with a clear phase contrast of approximately 180°. Building on this ferroelectric characteristic, Fig. 2(b) presents the current-voltage characteristics of the 10 nm × 10 nm FTJ with 0.9 nm HZO. The device exhibits distinct up-polar (OFF) and down-polar (ON) states, securing an on/off ratio exceeding 10 at a read voltage of 0.4 V. The inset further confirms bipolar resistive switching behavior, demonstrating that reliable FTJ memory operation is achievable at this ultra-thin HZO thickness.

The key enabler for memory operation at this thickness is extreme area scaling, which significantly shifts the breakdown field distribution as shown in Fig. 3(a). The Weibull analysis of breakdown field reveals that as the device area is reduced from ~4×10^6 nm² to 100 nm², the breakdown field increases significantly with a slope of approximately 41 [7]. This trend is attributed to the reduction in the total number of defects within the HZO layer at smaller device areas, which directly lowers the probability of defect-assisted premature breakdown [8]. Consequently, the extremely scaled devices can withstand significantly higher electric fields without breakdown, enabling the application of large write pulse amplitudes that are otherwise inaccessible in large-area counterparts.

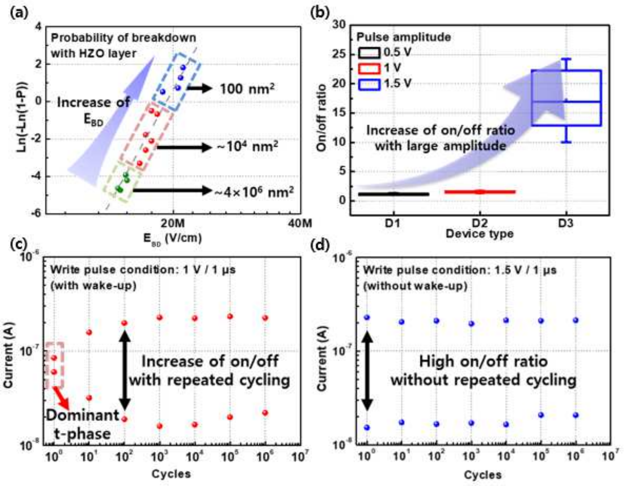

Fig. 3. (a) Weibull analysis of breakdown field for different device areas. The breakdown field increases significantly as the device area is reduced. (b) On/off ratio as a function of device type for various pulse amplitudes. (c) Cycling characteristics under a small write pulse condition and (d) a large write pulse condition.

As demonstrated in Fig. 3(b), applying a small pulse amplitude of 0.5 V or 1 V fails to induce sufficient orthorhombic phase formation, resulting in a negligible on/off ratio. In contrast, a large pulse amplitude of 1.5 V generates a sufficiently strong electric field to promote orthorhombic phase formation [9], yielding an average on/off ratio of approximately 15 across multiple devices. The dependence of cycling behavior on pulse amplitude is further examined in Fig. 3(c) and 3(d). Under a 1 V / 1 μs write pulse condition, the device initially exhibits a low on/off ratio dominated by the tetragonal phase, which gradually improves through repeated cycling due to the wake-up effect. A meaningful on/off ratio of approximately 10 is secured from 100 cycles, reaching its maximum beyond 1000 cycles. In contrast, under a 1.5 V / 1 μs write pulse condition shown in Fig. 3(d), the strong electric field immediately induces sufficient orthorhombic phase formation from the first cycle, securing a high on/off ratio without requiring any wake-up process. These results collectively demonstrate that extreme area scaling is essential for enabling high-field operation in ultra-thin HZO FTJs, which in turn determines the achievable on/off ratio.

IV. Scaling-Induced Stochastic Behavior

As the device area is aggressively scaled down, the number of ferroelectric domains within the HZO layer decreases dramatically, transitioning from a regime of collective domain switching to one governed by only a few domains. Concurrently, the total number of structural defects inherent to the HZO film, including grain boundaries and oxygen vacancies, is significantly reduced with decreasing device area. In large-area devices, the contributions of individual defects are averaged out over numerous domains, rendering their effects statistically negligible and producing a deterministic macroscopic response. However, as shown in Fig. 4(a), extreme area scaling reduces the active volume to the point where the device response becomes highly sensitive to the behavior of individual defects. Under these conditions, the stochastic nature of single defect

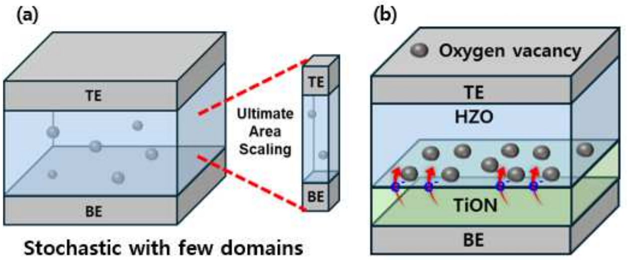

Fig. 4. (a) Schematic illustrations of scaling-induced behavior, resulting in the decrease of defects in smaller area. (b) Oxygen vacancy traps at the HZO/TiON interface acting as active sites for electron capture and emission.

events is no longer screened by ensemble averaging and becomes directly observable in the electrical characteristics.

In particular, as depicted in Fig. 4(b), oxygen vacancies present at the interface between the HZO layer and the TiON interfacial layer on the bottom electrode act as active trap sites. In the extremely scaled device, electron capture and emission events at these individual oxygen vacancy traps directly modulate the current, giving rise to detectable random telegraph noise (RTN) [10]. Unlike large-area devices where such trap-mediated fluctuations remain buried beneath the averaged response, the extremely scaled FTJ provides a unique platform for resolving single-trap dynamics. This scaling-induced stochasticity, rooted in the discrete nature of interfacial oxygen vacancy traps, opens a new avenue for exploiting the inherent randomness of nanoscale ferroelectric devices in emerging stochastic computing systems such as p-bit computing [11].

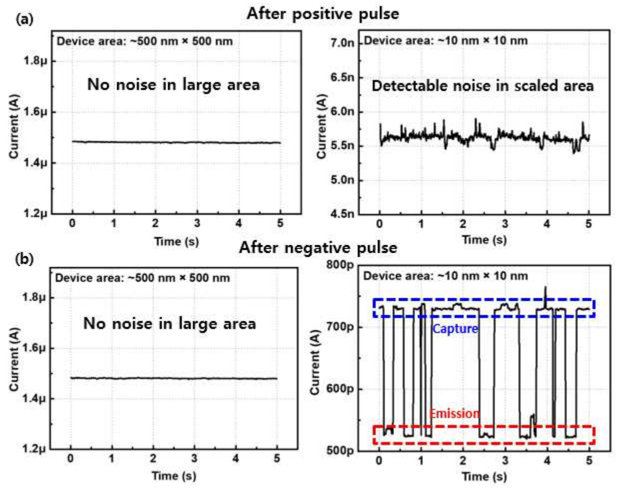

Fig. 5. (a) Time-domain current characteristics of large-area and extremely scaled FTJ devices after applying positive write pulse and (b) negative write pulse. RTN is absent in large-area devices due to domain averaging while pronounced stochastic fluctuations are observed in the scaled device.

Fig. 5 compares the time-domain current characteristics of large-area (~500 nm × 500 nm) and extremely scaled (~10 nm × 10 nm) FTJ devices under positive and negative pulse conditions. As shown in Fig. 5(a), after applying a positive pulse, the large-area device exhibits a stable and noise-free current trace, with no detectable stochastic fluctuation. Similarly, Fig. 5(b) shows that the large-area device remains equally featureless after a negative pulse. These results confirm that RTN remains undetectable in large-area devices regardless of the polarization state, consistent with the domain averaging effect discussed above.

In contrast, the extremely scaled 10 nm × 10 nm device reveals distinct current fluctuations. After a positive pulse, as shown in Fig. 5(a), a relatively weak noise signal is observed. This is attributed to the polarization direction formed by the positive pulse, which drives oxygen vacancies away from the HZO/TiON interface, reducing the available trap density at the interface and thereby limiting electron capture events. After a negative pulse, however, the opposite polarization direction attracts oxygen vacancies toward the HZO/TiON interface, significantly increasing the interfacial trap density. Under this condition, as shown in Fig. 5(b), pronounced two-level RTN is clearly observed, with discrete switching between capture and emission states. The discrete nature of the current transitions confirms that individual oxygen vacancy traps at the HZO/TiON interface are responsible for the observed stochastic behavior. These results demonstrate that polarization direction serves as an effective means to control interfacial trap occupancy, and that extreme area scaling is a prerequisite for resolving such single-trap dynamics in HZO-based FTJs.

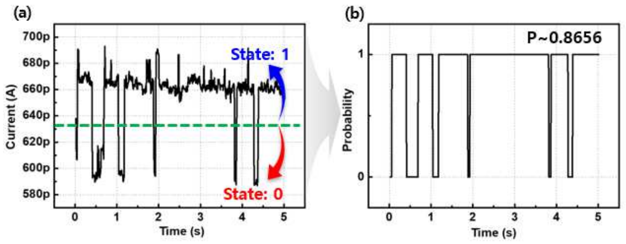

Fig. 6. (a) Time-domain current trace of the scaled FTJ measured after applying a negative write pulse, exhibiting discrete two-level fluctuation. (b) Corresponding binary probability signal converted from the measured current trace.

Fig. 6(a) shows the time-domain current trace measured at a read voltage of 0.3 V after applying a negative write pulse to the 10 nm × 10 nm FTJ device. The read voltage of 0.3 V ensures no polarization switching occurs during the RTN measurement window. The current fluctuates discretely between two distinct levels, corresponding to the capture and emission of electrons at individual oxygen vacancy traps at the HZO/TiON interface. To quantify the stochastic switching, a threshold current defined as the midpoint between the two current levels is used to classify the device state. Current values above the threshold are assigned as State 1, while those below are assigned as State 0.

Based on this state classification, the time-domain current trace is converted into a binary probability signal as shown in Fig. 6(b). The switching probability is then extracted by calculating the fraction of the total measurement time during which the device resides in State 1. For the representative trace shown, the extracted probability indicates that the device spends the majority of the measurement window in State 1 under this pulse condition, as further analyzed in the following section. This probability extraction method provides a straightforward and reliable means of quantifying the stochastic characteristics of the scaled FTJ, forming the basis for the polarization-controlled probability modulation [12].

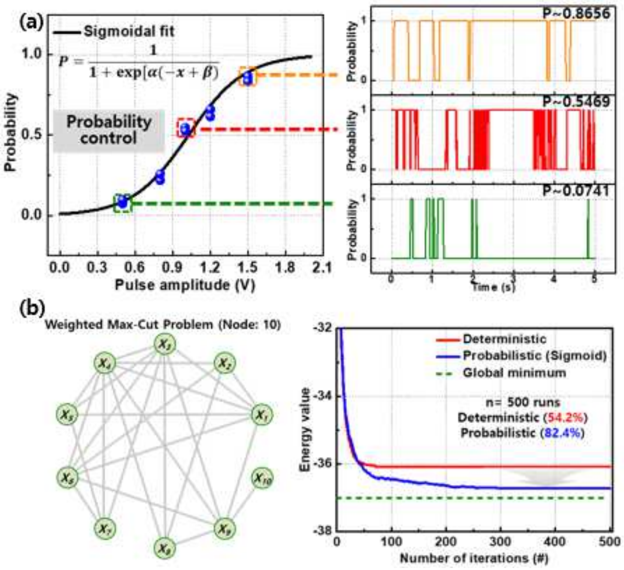

Fig. 7. (a) Hardware p-bit switching probability precisely controlled as a function of the negative write pulse amplitude, showing excellent agreement with a continuous sigmoidal fit. (b) Schematic of a 10-node weighted max-cut problem and its energy convergence traces, validating that the FTJ-calibrated probabilistic approach successfully outperforms deterministic optimization in finding the global minimum.

To investigate the tunability of switching probability through polarization control, the negative write pulse amplitude is systematically varied from 0.5 V to 1.5 V while the read voltage is fixed at 0.3 V. As shown in Fig. 7(a), the extracted probability increases monotonically with increasing pulse amplitude, demonstrating clear and continuous probability control. Specifically, at 0.5 V, the probability is as low as 0.0741, rising to 0.5469 at 1.0 V and further to 0.8656 at 1.5 V where the device predominantly occupies State 1.

The physical origin of this probability modulation is rooted in the polarization-dependent redistribution of oxygen vacancies at the HZO/TiON interface. As the negative pulse amplitude increases, the remnant polarization magnitude grows, driving more positively charged oxygen vacancies toward the interface. The increased oxygen vacancy concentration raises the trap site density, facilitating electron capture, while also providing additional hopping pathways for electron tunneling across the HZO barrier. This polarization-mediated control thus provides a direct and continuous handle on the stochastic switching probability of the scaled FTJ device.

The relationship between pulse amplitude and switching probability is well described by a sigmoid function $P = 1 / [1 + \exp\{\alpha(-x + \beta)\}]$ with fitted parameters $\alpha = 4.47$ and $\beta = 1.03$, in close agreement with the experimentally measured probability of 0.5469 at 1.0 V. This behavior is directly analogous to the sigmoid activation function required for p-bit operation. To validate this feasibility, a weighted max-cut problem is simulated on a randomly generated 10-node graph (edge probability = 0.5, weights = 1~3), as shown in Fig. 7(b). At each iteration, the probabilistic solver samples each node state using the measured sigmoid as the p-bit activation function, while the deterministic baseline employs a greedy algorithm that selects the locally optimal state. Over 500 independent runs, the probabilistic approach achieves the global minimum in 82.4% of cases compared to 54.2% for the deterministic approach, demonstrating that the polarization-controlled sigmoid response of the scaled HZO FTJ enables effective hardware p-bit operation for combinatorial optimization applications.

V. Conclusion

In this work, we demonstrate the extreme dimensional scaling of HZO-based FTJs with a 0.9 nm HZO layer, achieving reliable on/off ratio (>10) through MWA process optimization and aggressive area scaling down to 10 nm × 10 nm. Extreme area scaling suppresses defect density, enhances breakdown field, and enables high-field orthorhombic phase formation without repeated cycling. Furthermore, scaling-induced RTN arising from polarization-controlled electron capture and emission at oxygen vacancy traps at the HZO/TiON interface is unveiled, with switching probability continuously tunable following a sigmoid function. Based on these results, a weighted max-cut simulation yields an 82.4% success rate versus 54.2% for deterministic optimization, demonstrating the potential of extreme dimensional scaling toward hardware p-bit computing.

References

[1] J. Hwang, Y. Goh, and S. Jeon, "Physics, structures, and applications of fluorite-structured ferroelectric tunnel junctions," *Small*, vol. 20, no. 9, p. 2305271, 2024.

[2] L. Jung, J. Lee, S. Oh, and H. Hwang, "Superior retention (>1 year, 85°C) and memory window (~1.8 V) using ultra-thin HZO FTJ with OTS selector for X-point memory applications," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2023, pp. 1–4.

[3] P. D. Lomenzo, S. Slesazeck, M. Hoffmann, T. Mikolajick, U. Schroeder, and B. Max, "Ferroelectric Hf1-xZrxO2 memories: device reliability and depolarization fields," in *Proc. 19th Non-Volatile Memory Technology Symp. (NVMTS)*, 2019, pp. 1–8.

[4] M. H. Park, H. J. Kim, Y. J. Kim, Y. H. Lee, T. Moon, K. D. Kim, S. D. Hyun, and C. S. Hwang, "Study on the size effect in Hf0.5Zr0.5O2 films thinner than 8 nm before and after wake-up field cycling," *Appl. Phys. Lett.*, vol. 107, no. 19, 2015.

[5] T. Francois et al., "Impact of area scaling on the ferroelectric properties of back-end of line compatible Hf0.5Zr0.5O2 and Si:HfO2-based MFM capacitors," *Appl. Phys. Lett.*, vol. 118, no. 6, 2021.

[6] K. Y. Camsari, R. Faria, B. M. Sutton, and S. Datta, "Stochastic p-bits for invertible logic," *Phys. Rev. X*, vol. 7, no. 3, p. 031014, 2017.

[7] C. Sire, S. Blonkowski, M. J. Gordon, and T. Baron, "Statistics of electrical breakdown field in HfO2 and SiO2 films from millimeter to nanometer length scales," *Appl. Phys. Lett.*, vol. 91, no. 24, 2007.

[8] X. Long, H. Tan, F. Sanchez, I. Fina, and J. Fontcuberta, "Ferroelectric electroresistance after a breakdown in epitaxial Hf0.5Zr0.5O2 tunnel junctions," *ACS Appl. Electron. Mater.*, vol. 5, no. 2, pp. 740–747, 2023.

[9] D. Zhou et al., "Wake-up effects in Si-doped hafnium oxide ferroelectric thin films," *Appl. Phys. Lett.*, vol. 103, no. 19, 2013.

[10] C. Han et al., "A monolithic ferroelectric-ionic duality for stochastic-neuromorphic core integration," *Adv. Mater.*, p. e17440, 2026.

[11] R. H. Koo, W. Shin, S. T. Lee, D. Kwon, and J. H. Lee, "Stochastic behavior of random telegraph noise in ferroelectric devices: Impact of downscaling and mitigation strategies for neuromorphic applications," *Chaos Solitons Fractals*, vol. 191, p. 115856, 2025.

[12] S. Luo, Y. He, B. Cai, X. Gong, and G. Liang, "Probabilistic-bits based on ferroelectric field-effect transistors for probabilistic computing," *IEEE Electron Device Lett.*, vol. 44, no. 8, pp. 1356–1359, 2023.