Received 18 February 2020; revised 17 March 2020; accepted 4 May 2020. Date of publication 12 May 2020; date of current version 18 June 2020.
The review of this article was arranged by Editor M. Liu.

Digital Object Identifier 10.1109/JEDS.2020.2993859

Impact of the Crystal Phase of ZrO_2 on Charge Trapping Memtransistor as Synaptic Device for Neural Network Application

YU-CHE CHOU, CHIEN-WEI TSAI, CHIN-YA YI, WAN-HSUAN CHUNG, AND CHAO-HSIN CHIEN

Institute of Electronics, National Chiao Tung University, Hsinchu 30010, Taiwan

CORRESPONDING AUTHOR: C.-H. CHIEN (e-mail: chchien@faculty.nctu.edu.tw)

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 107-2221-E-009-095-MY3, and in part by the Center for the Semiconductor Technology Research under Grant MOST-108-3017-F-009-003.

ABSTRACT In this work, we investigated the effects of the crystal phase of ZrO_2 on charge trapping memtransistors (CTMTs) as synaptic devices for neural network applications. The ZrO_2 deposited through thermal (t-ZrO_2) atomic layer deposition (ALD) and plasma (p-ZrO_2) ALD were analyzed using an X-ray diffractometer, which indicated that the t-ZrO_2 consisted of pure cubic phase, whereas p-ZrO_2 consisted of both cubic and tetragonal phases. Through X-ray photoelectron spectroscopy analysis, we then constructed the energy band diagram of the gate stacks. The ΔE_C of t- and p-ZrO_2 with respect to tunneling and blocking Al_2O_3 were 1.84 and 1.19 eV respectively. Because of the relatively large ΔE_C of t-ZrO_2, the window of the flat band voltage (V_{FB}) shift extracted from charge trapping capacitors was enlarged by 591.9 mV more than the one using p-ZrO_2 as the charge trapping layer. Retention was also improved by 10.4% after 10^5 s in the t-ZrO_2 case. Finally, we fabricated the CTMTs with the gate stack of the t-ZrO_2 case and demonstrated their characteristics as synaptic devices. With the optimization of pulse schemes, we reduced the nonlinear factors of depression (α_d) and potentiation (α_p) from -6.72 and 6.47 to 0.03 and 0.01 respectively, enlarged the ON/OFF ratio from 15.6 to 70.4 and increased the recognition accuracy from 27.6% to 86.5% simultaneously.

INDEX TERMS Germanium, high-κ dielectrics, multilayer perceptron, neural network hardware, synaptic device, zirconium oxide.

I. INTRODUCTION

With the recent development of artificial intelligence (AI), implementing AI in end-point devices has become an extremely active research topic in both academia and industry. AI uses neural network (NN) implementation that involves numerous matrix product operations and neuron storage content transfers [1]. For off-the-shelf hardware platforms, NNs are primarily implemented in von Neumann architecture such as graphic processing units [2], field-programmable gate arrays [3], or neural processing units [4] to boost the operation of their neural networks. However, the problem with NN implementation in von Neumann architecture is more in data movement between arithmetic units and memory [5], in which energy and bandwidth are largely consumed by movement of data from the memory to the arithmetic unit or vice versa, than in computation [6]. Hence, the appreciable improvement in energy consumption and performance is stringent [7]. Nowadays, in-memory computing is a fundamental solution and has attracted much attention, in which complementary metal–oxide–semiconductor transistors are replaced with "synaptic devices," such as flash memory [8], phase change memory [9], resistive random access memory [10]–[13], magnetic random access memory [14], and ferroelectric field-effect transistor [15]. These have exhibited great potential in NN applications. However, the recognition accuracy problem caused by the nonlinearity of weight-to-pulse rate and preneuron-to-postneuron relations has become a prominent topic in recent studies [15]–[17].

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

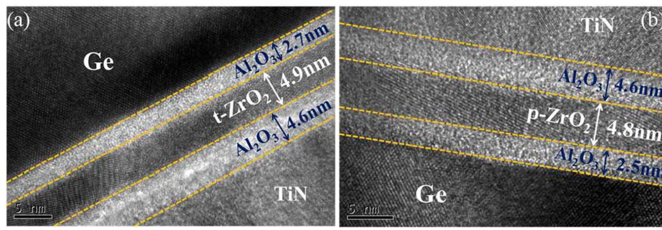

FIGURE 1. TEM images of the gate stacks of (a) t-ZrO_2 and (b) p-ZrO_2 cases. The physical thickness of the tunneling, charge trapping, and blocking layers are (a) 2.7, 4.9, and 4.6 nm respectively, and (b) 2.5, 4.8, and 4.6 nm respectively.

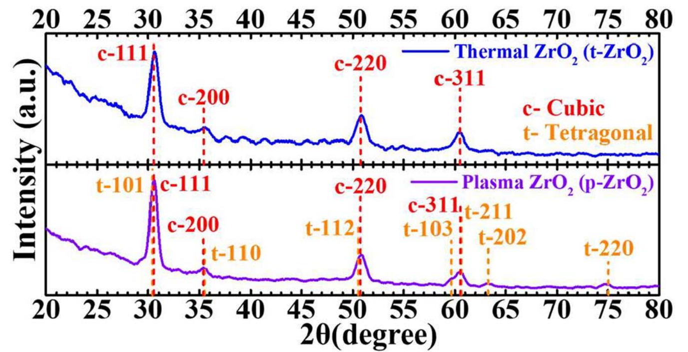

FIGURE 2. XRD patterns of t-ZrO_2 and p-ZrO_2 thin film. The dash lines and three digits mark the diffraction peaks and orientation of cubic or tetragonal phases of ZrO_2, and the numbers represent the 2θ angles with the prefix indicating its crystal phase. The p-ZrO_2 thin film has extra peaks at t-103, t-202, and t-220, which indicate the composition of the tetragonal phase.

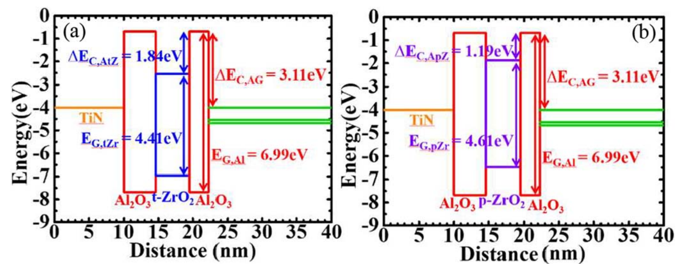

FIGURE 3. Energy band diagram of (a) t-ZrO_2 and (b) p-ZrO_2 charge trapping capacitor gate stacks extracted from XPS analysis with flat band bias. The E_G of t- and p-ZrO_2 are 4.41 and 4.61 eV, respectively. The ΔE_C of t- and p-ZrO_2 regarding Al_2O_3 are 1.84 and 1.19 eV, respectively.

In this study, we fabricated charge trapping memtransistors (CTMTs) with ZrO_2 as the charge trapping layer (CTL). We discovered that ZrO_2 deposited with two types of atomic layer deposition (ALD) systems depicted different crystallography through X-ray diffractometry (XRD) analysis, and then revealed distinct band alignments of Al_2O_3 through X-ray photoelectron spectroscopy (XPS). This affected the charge trapping capability of ZrO_2 for CTMTs. The influences were apparently demonstrated with the characteristics of the CTMTs in the neural network application.

II. FABRICATION AND PHYSICAL CHARACTERIZATION OF CHARGE TRAPPING CAPACITOR

First, the (100)-oriented p-type Ge substrate was cleaned with diluted hydrofluoric acid. A GeO_2 IL and high-κ gate stacks were then deposited using a plasma-enhanced ALD system in situ. We chose 30-cycle Al_2O_3 as the tunneling layer and 60-cycle Al_2O_3 as the blocking layer because of its high bandgap (E_G), conduction band offset (ΔE_C), and capability of stopping the diffusion of Ge into the high-κ dielectric, which would lead to higher interfacial trap density [18]. A 60-cycle ZrO_2 was deposited as the CTL by both thermal (t-ZrO_2) and plasma (p-ZrO_2) ALD for comparison. Afterwards, a 400°C 60s rapid thermal annealing was performed as postdeposition annealing for improving the quality of the high-κ dielectric, and then a 50 nm of TiN was deposited using physical vapor deposition (PVD) and patterned as gate metal. Finally, a 10 nm of Ti and a 300 nm of Al were deposited with PVD as the backside contact. Fig. 1 presents the transmission electron microscope images of the two gate stacks with the aforementioned physical thickness of each layer.

we used the XRD to investigate the composition of the crystal phase within t-ZrO_2 and p-ZrO_2 thin films. Fig. 2 displays the XRD result, which indicates that t-ZrO_2 was constituted purely of the cubic phase, whereas p-ZrO_2 contained both cubic and tetragonal phases; this is attributed to the fact that there were more reactive oxygen radicals created by O_2 plasma in the p-ZrO_2 case [19], [20]. We next used the XPS to determine the energy band diagrams [21] of two gate stacks which are shown in Fig. 3. The E_G of t- and p-ZrO_2 are 4.41 and 4.61 eV respectively. The ΔE_C of t- and p-ZrO_2 with respect to Al_2O_3 are 1.84 and 1.19 eV respectively. Our results are consistent with previous research [22]. By combining the results using XPS and XRD, we could conclude that the band structure difference between t-ZrO_2 and p-ZrO_2 was caused by the incorporation of the tetragonal crystal phase into p-ZrO_2, which was not observed in t-ZrO_2.

III. ELECTRICAL CHARACTERIZATION OF CHARGE TRAPPING CAPACITOR

After the fabrication and physical analysis, we characterized the charge trapping capability of the charge trapping capacitors (CTCs). Fig. 4(a) presents the relationship between flat band voltage (V_{FB}) shift and depression/potentiation pulse width, which indicated that t-ZrO_2 had a larger window and higher charge trapping efficiency for any depression/potentiation pulse width. Using t-ZrO_2 as a CTL enlarged the V_{FB} shift window by 591.9 mV and generated a faster depression/potentiation speed than did using p-ZrO_2. The retention of the two cases is displayed in Fig. 4(b). Using t-ZrO_2 as the CTL resulted in 10.4% more V_{FB} shift than did using p-ZrO_2 after 10^5 s. These results are consistent with previous research [22], which indicated that cubic phase ZrO_2 had a higher trap density well as that predicted

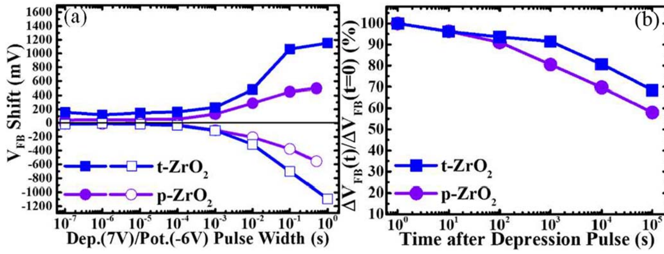

FIGURE 4. (a) V_{FB} shift-to-depression/potentiation pulse width curve of t- and p-ZrO_2 charge trapping capacitors. t-ZrO_2 charge trapping capacitor indicated a larger V_{FB} shift window and faster depression/potentiation. (b) Retention of t- and p-ZrO_2 charge trapping capacitors. The t-ZrO_2 charge trapping capacitor registered 10.4% more V_{FB} shift than did p-ZrO_2 CTC after 10^5 s.

by our extracted band diagrams shown in Fig. 3 signifying that t-ZrO_2 had larger ΔE_C with Al_2O_3.

IV. CHARGE TRAPPING MEMTRANSISTOR AS SYNAPTIC DEVICE

We integrated the high-κ stack of t-ZrO_2 CTC into CTMTs as their gate stacks and benchmarked their performances as synaptic devices. Fig. 5(a) shows the operation of neural networks using multilayer perceptron (MLP) with one hidden layer for Modified National Institute of Standards and Technology (MNIST) database recognition. The synaptic array was composed of the synapses that connected one layer of neurons to the next. Fig. 5(b) reveals the architecture of the CTMT synaptic array on MLP hardware acceleration derived from Fig. 5(a). Each CTMT represented a synapse, and their drains/sources represented pre/post-neurons. Similar to the manipulation of the V_{FB} of CTCs, we could manipulate the threshold voltage (V_T) of CTMTs through applying positive/negative pulses on the gate of the CTMTs to depress/potentiate the synapses, which shifted the V_T of the CTMTs negatively or positively. This was called depression or potentiation and is illustrated in Fig. 5(c). The drain voltage and current of the CTMTs represented the value of preneurons and postneurons, respectively, whose relations could be simplified as (1). Through adjusting V_T, we could manipulate the channel conductance of the CTMTs, as presented in Fig. 5(d), which represented the weight (W_{ji}) of the synapses.

First, the $I_D - V_G$ characteristic of the CTMT is shown in Fig. 6. The channel width of the CTMT is 100 μm as well as the channel length. The subthreshold swing is 292 mV/dec and the window between potentiation and depression states is 1390 mV.

Second, the pulse number–weight relationships are shown in Fig. 7. With three types of pulse scheme, the nonlinearity factor and ON/OFF ratio were improved simultaneously. The stepping pulse scheme could improve the nonlinear factors of depression (α_d) and potentiation (α_p) from -6.72 and 6.47 to -2.83 and -0.12 respectively, and the ON/OFF ratio from 15.6 to 24.4. With the optimization of the pulse scheme, α_d and α_p could be further improved to 0.03 and 0.01, and the

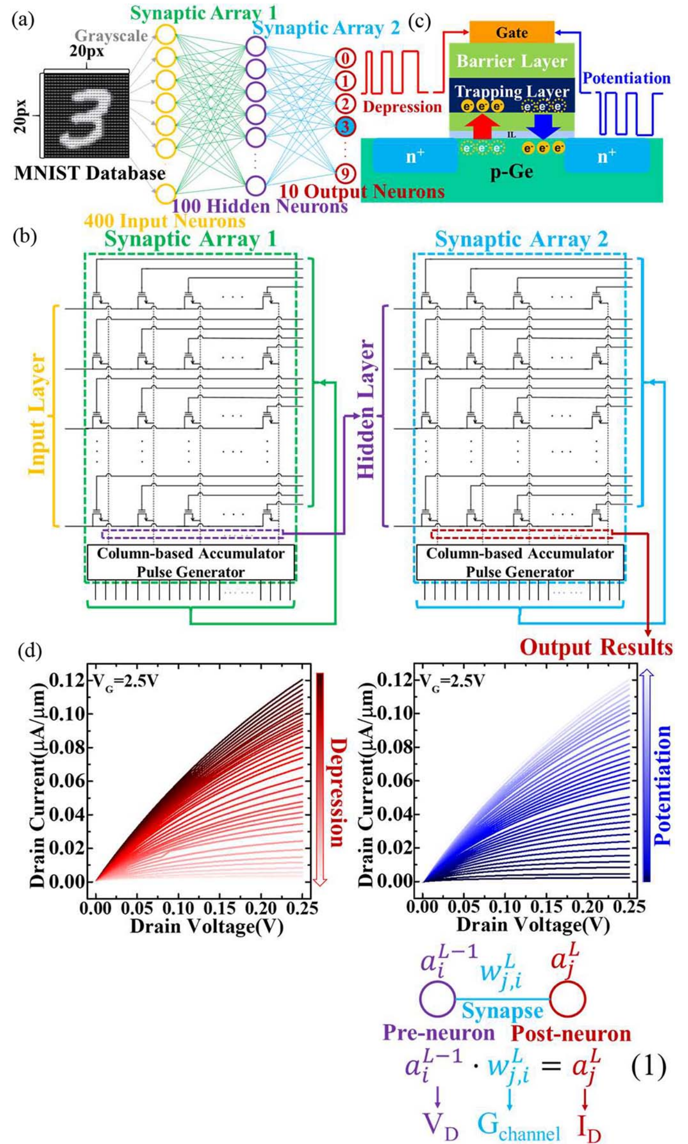

FIGURE 5. (a) Illustration of neural network architecture using MLP with one hidden layer for MNIST database recognition. (b) Architecture of CTMT synaptic array for MLP hardware acceleration derived from (a). (c) Operation principle of the CTMTs as synaptic devices. Applying positive/negative pulses to depress/potentiate the synapse, which shifted the V_T of the CTMTs negatively/positively. (d) $I_D - V_D$ curves of the CTMTs under a fixed V_G bias during depression/potentiation fit (1) well for weighted calculation.

ON/OFF ratio to 70.4. Although the optimization of pulse scheme may increase the latency in training phases, it can be simply resolved by the implementation of cloud training with CTMTs as the end-point reference element.

Finally, we benchmarked the CTMTs with the recognition accuracy of MNIST database by NeuroSim [23]. We constructed the MLP model with 400 pre-neurons for 20px*20px hand-written-number patterns, 100 hidden neurons and 10 post-neurons for ten digits. With three types of pulse scheme, the recognition accuracy is improved dramatically. The stepping pulse scheme could improve

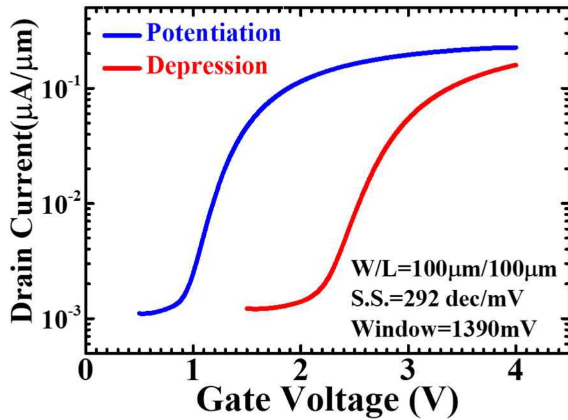

FIGURE 6. $I_D - V_G$ curves of the CTMT under potentiation and depression states. The channel width of the CTMT is 100μm and the channel length is 100 μm. The subthreshold swing is 292 mV/dec. The window between potentiation and depression states is 1390 mV.

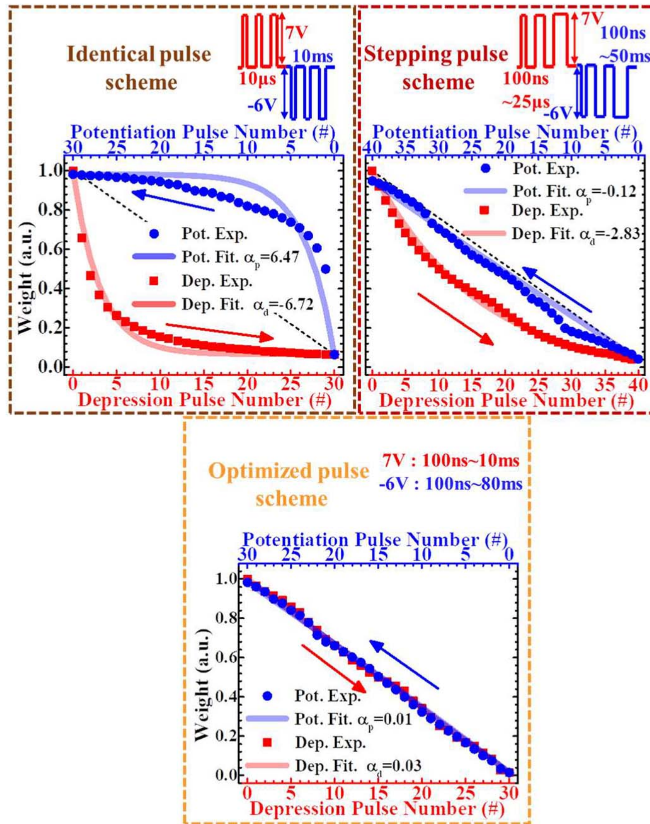

FIGURE 7. Pulse number-to-weight relations under identical, stepping, and optimized pulse scheme.

the recognition accuracy from 27.6% to 82.8%. With the optimization of the pulse scheme, recognition accuracy could be further improved to 86.5%. All the data are presented in Fig. 8.

V. CONCLUSION

We deposited two types of ZrO_2 using thermal and plasma ALD as the CTL within the gate stacks of CTCs.

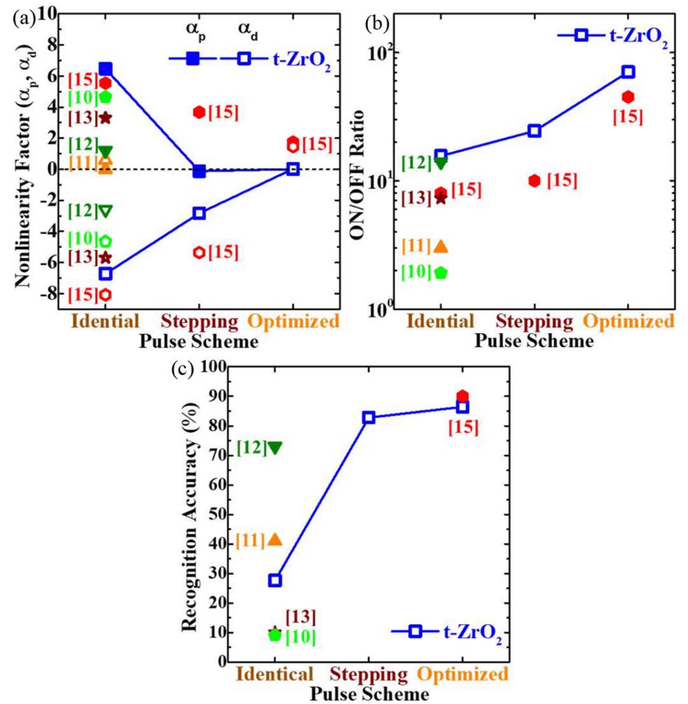

FIGURE 8. Benchmark of (a) nonlinearity (α_d, α_p), (b) ON/OFF ratio, and (c) recognition accuracy. The nonlinearity (α_d, α_p) of identical, stepping, and optimized pulse schemes are (−6.72, 6.47), (−2.83, −0.12), and (0.03, 0.01). The ON/OFF ratio of identical, stepping, and optimized pulse schemes were 15.6, 24.4, and 70.4. The recognition accuracy was 27.6%, 82.8%, and 86.5% respectively.

Through the XRD analysis, we observed that t-ZrO_2 was composed of the pure cubic phase, whereas p-ZrO_2 was composed of both cubic and tetragonal phases. We then used the XPS to determine the band structure of the gate stacks. The ΔE_C of t- and p-ZrO_2 with respect to tunneling and blocking Al_2O_3 were 1.84 eV and 1.19 eV respectively. With larger ΔE_C, t-ZrO_2 used as the CTL in CTCs caused a 591.9-mV higher V_{FB} shift and 10.4% better retention than did that of p-ZrO_2. We then used t-ZrO_2 as the CTL within the gate stack of the CTMTs. With the optimization of pulse schemes, we reduced α_d and α_p from −6.72 and 6.47 to 0.03 and 0.01, respectively, increased the ON/OFF ratio from 15.6 to 70.4 and the recognition accuracy from 27.6% to 86.5% simultaneously.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," presented at the Proc. 25th Int. Conf. Neural Inf. Process. Syst., vol. 1, 2012.

[2] S. Chetlur *et al.*, "cuDNN: Efficient primitives for deep learning," 2014. [Online]. Available: arXiv:1410.0759.

[3] G. Lacey, G. W. Taylor, and S. Areibi. *Deep Learning on FPGAs: Past, Present, and Future*. [Online]. Available: https://arxiv.org/abs/1602.04283

[4] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proc. ACM/IEEE 44th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2017, pp. 1–12, doi: 10.1145/3079856.3080246.

[5] J. Partzsch and R. Schuffny, "Analyzing the scaling of connectivity in neuromorphic hardware and in models of neural networks," *IEEE Trans. Neural Netw.*, vol. 22, no. 6, pp. 919–935, Jun. 2011, doi: 10.1109/TNN.2011.2134109.

[6] G. Desoli *et al.*, "14.1 A 2.9TOPS/W deep convolutional neural network SoC in FD-SOI 28 nm for intelligent embedded systems," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Feb. 2017, pp. 238–239, doi: 10.1109/ISSCC.2017.7870349.

[7] Y. Du *et al.*, "An analog neural network computing engine using CMOS-compatible charge-trap-transistor (CTT)," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 38, no. 10, pp. 1811–1819, Oct. 2019, doi: 10.1109/TCAD.2018.2859237.

[8] O. Fujita and Y. Amemiya, "A floating-gate analog memory device for neural networks," *IEEE Trans. Electron Devices*, vol. 40, no. 11, pp. 2029–2035, Nov. 1993, doi: 10.1109/16.239745.

[9] S. B. Eryilmaz *et al.*, "Experimental demonstration of array-level learning with phase change synaptic devices," in *Proc. IEEE Int. Electron Devices Meeting*, Dec. 2013, pp. 25.5.1–25.5.4, doi: 10.1109/IEDM.2013.6724691.

[10] I.-T. Wang, Y.-C. Lin, Y.-F. Wang, C.-W. Hsu, and T.-H. Hou, "3D synaptic architecture with ultralow sub-10 fJ energy per spike for neuromorphic computation," in *Proc. IEEE Int. Electron Devices Meeting*, Dec. 2014, pp. 28.5.1–28.5.4, doi: 10.1109/IEDM.2014.7047127.

[11] J. Woo *et al.*, "Improved synaptic behavior under identical pulses using AlOx/HfO_2 bilayer RRAM array for neuromorphic systems," *IEEE Electron Device Lett.*, vol. 37, no. 8, pp. 994–997, Aug. 2016, doi: 10.1109/LED.2016.2582859.

[12] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano Lett.*, vol. 10, no. 4, pp. 1297–1301, Mar. 2010, doi: 10.1021/nl904092h.

[13] S. Park *et al.*, "Neuromorphic speech systems using advanced ReRAM-based synapse," in *Proc. IEEE Int. Electron Devices Meeting*, Dec. 2013, pp. 25.6.1–25.6.4, doi: 10.1109/IEDM.2013.6724692.

[14] D. Fan and S. Angizi, "Energy efficient in-memory binary deep neural network accelerator with dual-mode SOT-MRAM," in *Proc. IEEE Int. Conf. Comput. Design (ICCD)*, Nov. 2017, pp. 609–612, doi: 10.1109/ICCD.2017.107.

[15] M. Jerry *et al.*, "Ferroelectric FET analog synapse for acceleration of deep neural network training," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2017, pp. 6.2.1–6.2.4, doi: 10.1109/IEDM.2017.8268338.

[16] S. Yu, P. Chen, Y. Cao, L. Xia, Y. Wang, and H. Wu, "Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2015, pp. 17.3.1–17.3.4, doi: 10.1109/IEDM.2015.7409718.

[17] P.-Y. Chen *et al.*, "Mitigating effects of non-ideal synaptic device characteristics for on-chip learning," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2015, pp. 194–199, doi: 10.1109/ICCAD.2015.7372570.

[18] Y.-H. Tsai *et al.*, "Improving thermal stability and interface state density of high-κ stacks by incorporating Hf into an interfacial layer on *p*-Germanium," *IEEE Electron Device Lett.*, vol. 37, no. 11, pp. 1379–1382, Nov. 2016, doi: 10.1109/LED.2016.2613999.

[19] K. Kukli *et al.*, "Influence of thickness and growth temperature on the properties of zirconium oxide films grown by atomic layer deposition on silicon," *Thin Solid Films*, vol. 410, no. 1, pp. 53–60, Jan. 2002. [Online]. Available: https://doi.org/10.1016/S0040-6090(2)00272-9

[20] D. Panda and T.-Y. Tseng, "Growth, dielectric properties, and memory device applications of ZrO_2 thin films," *Thin Solid Films*, vol. 531, pp. 1–20, Mar. 2013. [Online]. Available: https://doi.org/10.1016/j.tsf.2013.01.004

[21] M. Perego, G. Seguini, and M. Fanciulli, "Energy band alignment of HfO_2 on Ge," *J. Appl. Phys.*, vol. 100, no. 9, 2006, Art. no. 093718, doi: 10.1063/1.2360388.

[22] T. V. Perevalov and D. R. Islamov, "Oxygen polyvacancies as conductive filament in Zirconia: First principle simulation," *ECS Trans.*, vol. 80, no. 1, pp. 357–362, Aug. 2017, doi: 10.1149/08001.0357ecst.

[23] P. Chen, X. Peng, and S. Yu, "NeuroSim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 37, no. 12, pp. 3067–3080, Dec. 2018, doi: 10.1109/TCAD.2018.2789723.