Impact of Read Disturb on Multilevel RRAM based Inference Engine: Experiments and Model Prediction

Wonbo Shim[1], Yandong Luo[1], Jae-sun Seo[2] and Shimeng Yu[1]
[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, USA
[2]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, USA
E-mail: shimeng.yu@ece.gatech.edu

***Abstract*—Different from the multilevel cell (MLC) memory where the crossover between tail bits matters, any drift of the conductance of the synaptic device induced by read disturb may aggregate, as the analog current is summed up along the column. In this work, we experimentally measured the conductance drift on 2-bit HfO_2 RRAM array based on 1-transsitor-1-resistor (1T1R) test vehicle. The drift behavior of different states is modeled by vertical and lateral filament growth and saturation. The device model is incorporated into a VGG-like convolutional neural network algorithm for CIFAR-10 dataset. Read voltage should be minimized to 0.3V or below to maintain the inference accuracy.**

***Index Terms*-- Multilevel RRAM, read disturb, neural network inference, in-memory computing.**

I. Introduction

In-memory computing is a promising paradigm that can overcome the challenge of the excessive data transfer in deep neural network (DNN). The multiply-and-accumulate (MAC) operation is embedded in memory itself by the analog weighted current summation along the column. Several emerging memories such as RRAM [1]-[6] and PCRAM [7]-[8], as well as mainstream memories such as SRAM [9] and FLASH [10]-[12] have been investigated for in-memory computing applications. Among them, emerging memories have been proposed as a strong candidate to implement in-memory computing in edge devices because of the non-volatility and logic compatibility. Up to 1Mb binary RRAM macro has been modified to support inference operation of DNN [13].

Multilevel RRAM [14]-[15] can achieve higher density and larger MAC throughput. However, the reliability effects of multilevel RRAM needs to be revisited. It is unlike the MLC memory application, where the error bit is determined by the crossover of the tail bits between adjacent levels that can be corrected by error correction code (ECC). Here any drift of the RRAM conductance may aggregate as inaccurate weighted sum value when read-out from the column in the analog manner, which may adversely impact the inference accuracy. Prior work [16]-[17] have investigated the endurance and the data retention at elevated temperature of multilevel RRAM devices for in-memory computing. The read disturb effect is not well explored, and the prior analysis of the neural network is based on a simple multilayer perceptron (MLP) for MNIST dataset. Compared to the conventional memory applications, the hardware for inference requires more read disturb immunity. This paper is thus focusing on the read disturb analysis and the inference accuracy degradation for a deeper neural network.

In this work, we tested the Winbond's HfO_2 based 1T1R array fabricated at 90nm [18]. A 64kb test vehicle was used for statistical measurement using NI's PXIe system. The 2-bit RRAM resistance change is measured as a function of the read voltage and the stress time or equivalent number of read cycles. The drift behavior is modeled and incorporated in a VGG-8 network [19] simulation for a more complex CIFAR-10 dataset.

II. Read Disturb Measurement and Modeling

Fig. 1 shows the die photo of the 256×256 1T1R HfO_2 RRAM test chip. Fig. 2 shows the cumulative probability distribution of 4 states. Conductance of each state is controlled by V_G at SET operation to limit the SET current. For inference operation, the weight is proportional to the conductance, thus we designed 4 states where state 1 is the high resistance state (HRS), and state 2/3/4 is linearly spaced in conductance in the low resistance state (LRS) regime. No write-verify is employed here to tighten the distribution here. Ideally the write-verify protocol is required for programming the conductance to the exact level needed for in-memory computing as demonstrated in our prior work [20].

Fig. 3 shows the resistance change model in (a) state 1 and (b) state 2/3/4. The resistance change of state 1 is explained by the vertical filament growth, where the tunneling gap decay rate (*dg/dt*) is exponential to bitline (BL) voltage under the read stress. Due to the positive feedback mechanism [21], it induces

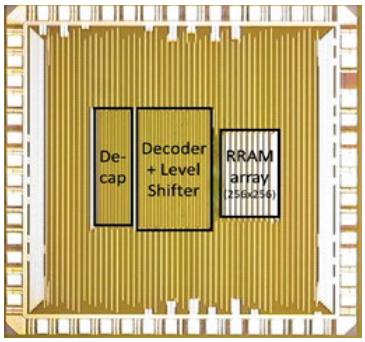

Figure 1. Die photo of the measured 64kb HfO_2 based 1T1R RRAM chip. The stress testing was done by using PGU and SMU units of the NI's PXIe system.

This work is supported by NSF-CCF-1903951, ASCENT, one of the SRC/DARPA JUMP Centers and the NSF/SRC E2CDA program.

978-1-7281-3199-3/20/$31.00 ©2020 IEEE

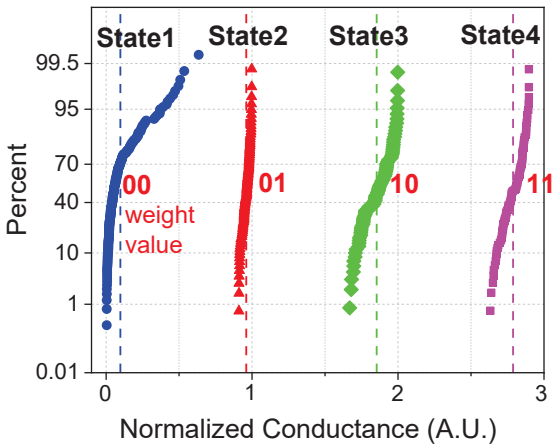

Figure 2. Culmulative distribution function of initial conductance of 2-bit RRAM. State 2/3/4 was achieved with different SET bias conditions.

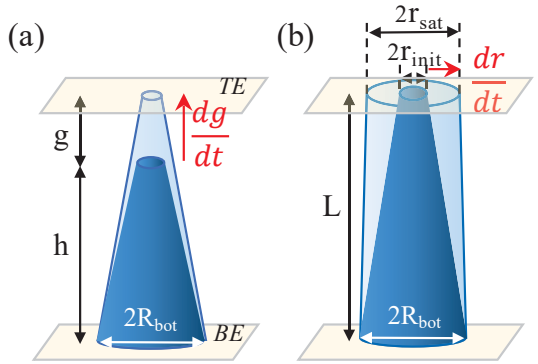

Figure 3. Filament growth mechanism of (a) state 1 (b) state 2/3/4 during the read stress. Vertical filament growth (gap decrease) induces abrupt resistance change in state 1, while lateral growth induces gradual resistance decrease at other states.

sudden gap g decrease at some point, and makes the abrupt resistance decrease from HRS to LRS.

The resistance change of state 2/3/4 is modeled with lateral filament growth. The resistance of LRS state is dominated by the radius r, and the growth of radius r of the filament is saturated when it approaches to the physical limit. This saturation radius and initial radius of filament affect to the rate of filament growth (dr/dt).

Fig. 4 shows the measured resistance of state 1 due to the stress time up to 500ms (equivalent to 5×10^7 read cycles assuming each read time is designed to 10ns) at different bit line voltage. Measured state 1 resistance abruptly drops at various read times during the 500ms stress. It can be seen that the lower the initial resistance is, the faster resistance drop to the LRS that implies positive feedback occurring point is closely related to the initial resistance of HRS. Therefore, the resistance of the HRS state must be designed so high with the consideration not only the conductance itself, but also immune to the read disturb.

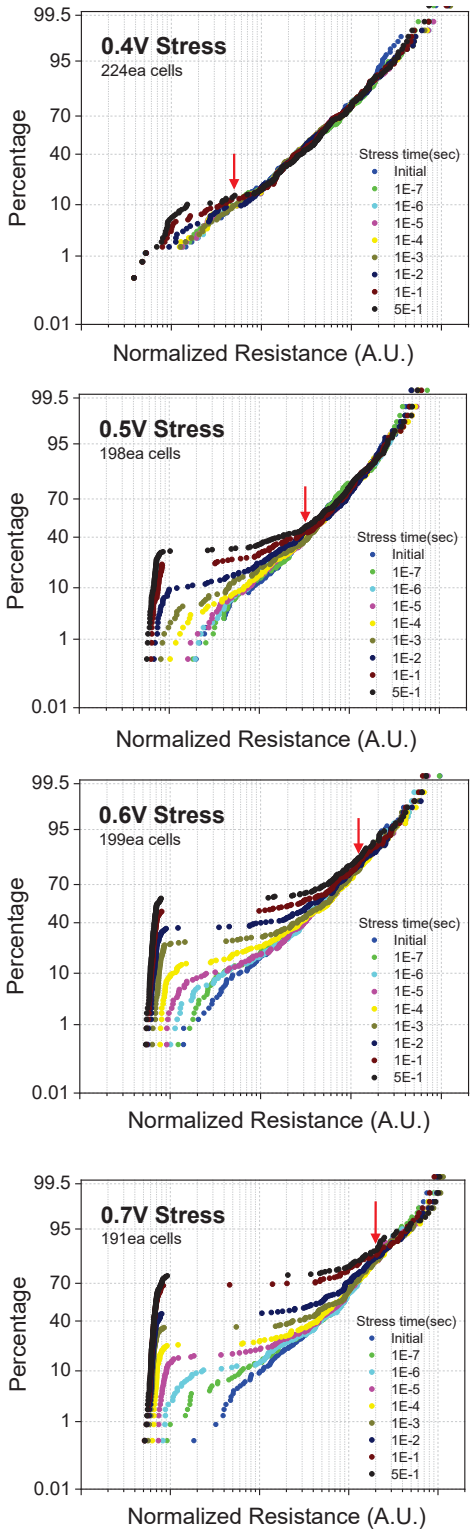

Figure 4. BL voltage induced resistance change of state 1 RRAM cells during the read stress testing. The cells with low initial resistance are vulnerable to read disturb. The critical resistance(red arrows) to read disturb before 500ms read stress (equivalent to 5×10^7 read cycles assuming each read is 10ns) increases with the BL voltage. Each stress is not a single pulse, but a series of accumulated pulses.

Modifying the compact model of RRAM device [21], the resistance change during the vertical filament growth can be expressed as following equations (1)-(3).

$$R = V_{Rout} \,/\, (I_0 e^{-\frac{g}{g_1}} sinh\left(\frac{V_{Rout}}{V_0}\right)) \quad (1)$$

$$\frac{dg}{dt} = -v_0 e^{-\frac{E_a}{kT}} \sinh\left(\frac{\gamma a_0}{L} \cdot \frac{qV}{kT}\right) \quad (2)$$

$$\gamma = \gamma_0 - \beta \cdot (\frac{g}{g_0})^3 \quad (3)$$

where R is the resistance of the filament, dg/dt is the gap growth rate, the gap g has lower and upper limit of g_{min} and g_{max}. V is the read stress voltage applied to BL, and V_{Rout} is BL voltage applied during the resistance read-out measurement, i.e. 50mV throughout this work. γ is g-dependent local field enhancement factor representing voltage dependence to filament height growth, γ_0 is fitting parameter, and a_0 is atomic hopping distance. By fitting the parameters as shown in Table 1, the models are finely fitted to the measured data for various stress voltages as shown in Fig. 5.

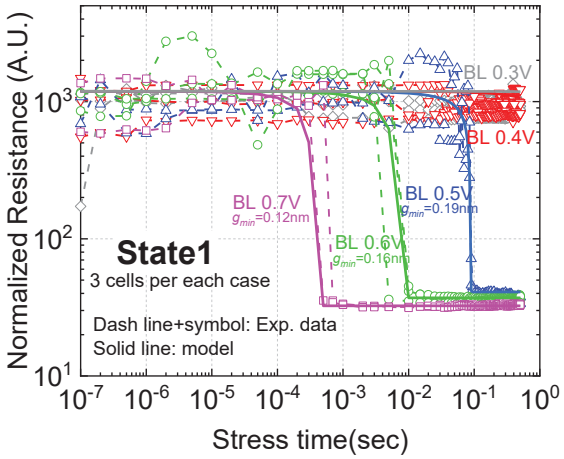

Figure 5. Read disturb of state 1 cells as a function of stress time.

Fig. 6 shows the resistance change of state 2/3/4. Measured cells have initial conductance corresponding to weight 01, 10 and 11 as in Fig.2. We could derive several physical models from such measured characteristics. First, after the 500ms read stress, resistance decrease saturates at similar resistance irrespectively to the BL voltage or initial resistance which implies saturation radius is always similar. Second, initial resistance affects the rate of resistance decrease during the radius growth phase. Third, the time at which saturation occurs is related only to the BL voltage regardless of the initial resistance. Lateral filament growth phase is following the compact model in Eq. 4 [22] and saturation phase are modeled empirically here as Eq. 5.

$$R = \rho_{ON} L/\pi \cdot r \cdot R_{bot} \quad (4)$$

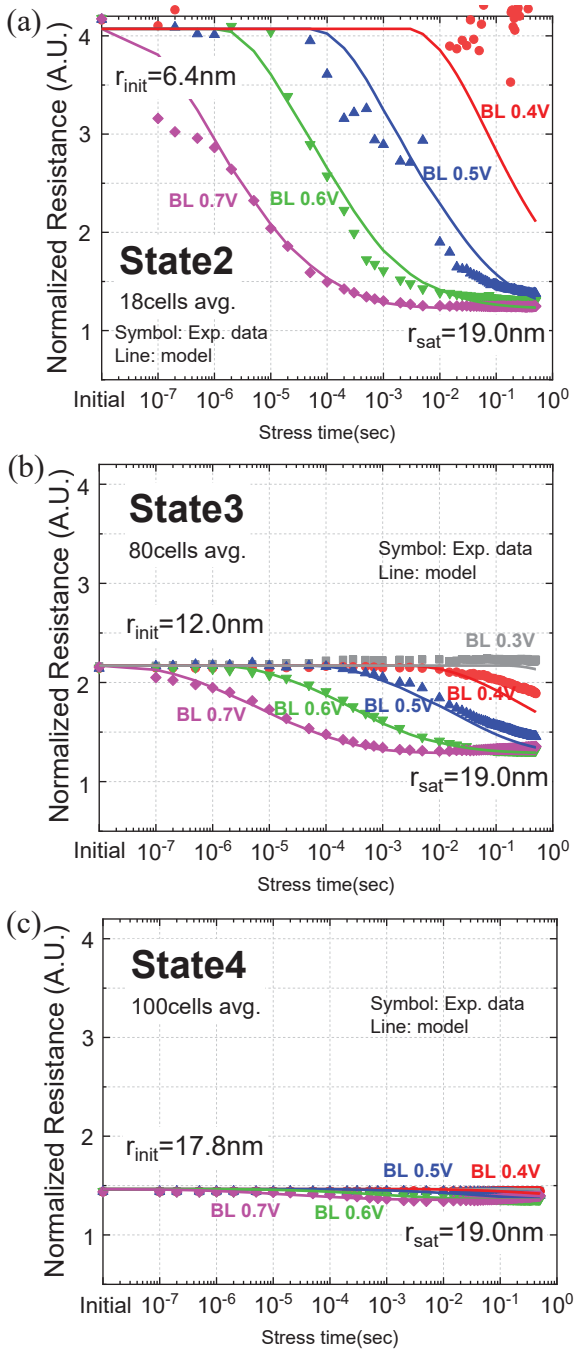

Figure 6. Read disturb of (a) state 2 (b) state 3 (c) state 4. Regardless of the initial resistance, all the states have similar saturation resistance and saturated eventually. The model fits the experimental data well using the same set of parameters.

$$\frac{dr}{dt} = \frac{\alpha(r_{sat} - r_{init})}{t_{stress}} \log\frac{t_{stress}}{t_{ch}(V)} \times 1/(1 + e^{-c_{sat}\log\frac{t_{sat}(V)}{t_{stress}}}) \quad (5)$$

where dr/dt is the radius growth rate, r_{sat} and r_{init} are saturation top radius and initial top radius of filament, respectively. R_{bot} is the saturation bottom radius which is assumed unchanged. First term of equation (5) represent the rate of radius growth is linearly proportional to the difference of r_{sat} and r_{init}, and last term represent the time which saturation is the function of stress

voltage only. Measured data are well fitted to this model in all the stress conditions using the same set of parameters as shown in Table 1.

TABLE 1. MODEL PARAMETERS

Vertical Filament Growth		Lateral Filament Growth	
Symbol	Value	Symbol	Value
g_1	0.3nm	α	0.09
V_0	47mV	r_{sat}	19.0nm
v_0	0.1m/s	r_{init}	(State2) 6.4nm
L	6nm		(State3) 12.0nm
a_0	2.5E-10m		(State4) 17.8nm
E_a	0.8eV	$t_{ch}(V)$	$6500\times\exp(-38V+0.7)$ sec
r_0	25		
β	1	c_{sat}	2
g_0	0.6nm	$t_{sat}(V)$	$10^{-14.7V+6.7}$ sec

III. INFERENCE ACCURACY SIMULATION

VGG-8 network [19] (as shown in Fig. 7(a)) on CIFAR-10 dataset is used to simulate the inference accuracy degradation due to the drift of multilevel RRAM states. 2-bit weight and 4-bit activations are used for the inference of this network, and the maximum accuracy that can be achieved by software simulation is 91.72%. Each 2-bit weight was mapped to one RRAM cell. Modeled conductance drift ratio compared to the initial conductance was reflected to the change of weight value in the simulation.

When the BL read voltage is 0.3V, 91.72% inference accuracy can be sustained up to 2×10^7 read cycle as shown in Fig. 7(b). Accuracy degradation is mainly caused by state 2 and 3, because of a relatively weak disturb immunity and significant conductance increase than other states.

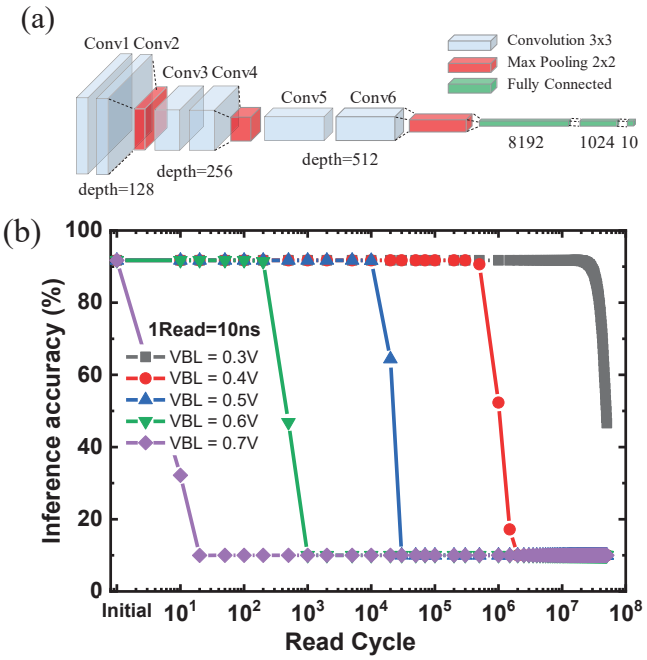

Figure 7. (a) Simulated VGG-8 network architecture (b) CIFAR-10 dataset inference accuracy simulation result due to read disturb.

IV. CONCLUSIONS

Read disturb induced conductance drift of multilevel RRAM was measured in 64kb test chip. Behavior of 4 states are modeled by filament growth mechanism and incorporated in inference simulation. Conductance drift degraded inference accuracy of VGG-8 if read voltage is >0.3V, inferring that the voltage-mode analog-to-digital converter (ADC) sense margin is limited. Further device engineering and circuit design techniques are to be developed to enhance the inference lifetime for the RRAM based inference engine.

ACKNOWLEDGMENT

We acknowledge the RRAM chip fabrication process provided by Winbond Electronics, Taiwan.

REFERENCES

[1] S. Yu, "Neuro-inspired computing with emerging non-volatile memory," *Proc. IEEE*, vol. 106, no. 2, pp. 260-285, 2018.

[2] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature* 521, pp. 61–64, May. 2015.

[3] J. Woo, K. Moon, J. Song, S. Lee, M. Kwak, J. Park, and H. Hwang, "Improved synaptic behavior under identical pulses using AlO_x/HfO_2 bilayer RRAM array for neuromorphic systems," *IEEE Electron Device Letters*, VOL. 37, NO. 8, pp. 994-997, Aug. 2016.

[4] F. Cai, J. M. Correll, S. H. Lee, Y. Lim, V. Bothra, Z. Zhang, M. P. Flynn, W.D. Lu M.A. Zidan, J. P. Strachan, and W. D. Lu, "A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations," *Nature Electronics* **2** (7), 290-299.

[5] W. Wu, H. Wu, B. Gao, P. Yao, X. Zhang, X. Peng, S. Yu, H. Qian, "A methodology to improve linearity of analog RRAM for neuromorphic computing," *Symposium on VLSI Technology*, June, 2018, art. No. 8510690, pp. 103-104.

[6] T. Gokmen, Y. Vlasov, "Acceleration of deep neural network training with resistive cross-point devices: design considerations," *Frontiers in Neuroscience*, 10, 333, 2016.

[7] S. Ambrogio, Narayanan, P., H. Tsai, R. M. Shelby, I. Boybat, C. Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. P. Farinha, B. Killeen, C. Cheng, Y. Jaoudi and G. W. Burr, "Equivalent-accuracy accelerated neural-network training using analogue memory," *Nature* **558**, 60–67, 2018.

[8] W. Kim, R.L. Bruce, T. Masuda3, G.W. Fraczak, N. Gong, P. Adusumilli, S. Ambrogio, H. Tsai, J. Bruley, J.-P. Han, M. Longstreet, F. Carta, K. Suu and M. BrightSky, "Confined PCM-based analog synaptic devices offering low resistance-drift and 1000 programmable states for deep learning, " *Symposium on VLSI Technology*, June, 2019, pp. T66-67.

[9] M. Kang, S. K. Gonugondla, A. Patil and N. R. Shanbhag, "A multi-functional in-memory inference processor using a standard 6T SRAM array," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 2, pp. 642-655, Feb. 2018.

[10] X. Guo, F. Merrikh Bayat, M. Bavandpour, M. Klachko, M. R. Mahmoodi, M. Prezioso, K. K. Likharev, D.B. Strukov, "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology," *IEEE International Electron Devices Meeting (IEDM),* San Francisco, CA, 2017, pp. 6.5.1-6.5.4.

[11] M. Kim, M. Liu, L. Everson, G. Park, Y. Jeon, S. Kim, S. Lee, S. Song, and C. H. Kim, "A 3D NAND Flash ready 8-Bit convolutional neural network core demonstrated in a standard logic process," *IEEE International Electron Devices Meeting (IEDM),* San Francisco, CA, 2019, pp. 38.3.1-38.3.4.

[12] J. Lee, B.-G. Park, Y. Kim, "Implementation of boolean logic functions in charge trap flash for in-memory computing,", *IEEE Electron Device Letters*, vol. 40, no. 9, pp. 1358-1361, Sept. 2019.

[13] C.-X. Xue, W.-H. Chen, J.-S. Liu, J.-F. Li, W.-Y. Lin *et al,* "A 1Mb multibit ReRAM computing-in-memory macro with 14.6ns parallel MAC computing time for CNN-based AI edge processors," *2019 IEEE International Solid- State Circuits Conference (ISSCC*), San Francisco, CA, USA, 2019, pp. 388-390.

[14] X. Sheng, C.E. Graves, S. Kumar, X. Li, B. Buchanan, L. Zheng, S. Lam, C. Li, J.P. Strachan, "Low conductance and multilevel CMOS integrated nanoscale oxide memristors," *Advanced Electronic Materials,* Vol.5, Sep.2019.

[15] C. Li, D. Belkin, Y. Li, P. Yan, M. Hu, N. Ge, H. Jiang, E. Montgomery, P. Lin, Z. Wang, W. Song, J. P. Strachan, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, and Q. Xia, "Efficient and self-adaptive in-situ learning in multilayer memristor neural networks," *Nature Communications* **9**, 2385(2018).

[16] M. Zhao, H. Wu, B. Gao, X. Sun, Y. Liu, P. Yao, Y. Xi, X. Li, Q. Zhang, K. Wang, S. Yu, H. Qian, "Characterizing endurance degradation of incremental switching in analog RRAM for neuromorphic systems," *2018 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2018, pp. 20.2.1-20.2.4.

[17] M. Zhao, H. Wu, B. Gao, Q. Zhang, W. Wu, S. Wang, Y. Xi, D. Wu, N. Deng, S. Yu, H. Chen and H. Qian, "Investigation of statistical retention of filamentary analog RRAM for neuromophic computing," *2017 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2017, pp. 39.4.1-39.4.4.

[18] C. Ho, S.-C. Chang, C.-Y. Huang, Y.-C. Chuang, S.-F. Lim, M.-H. Hsieh, S.-C. Chang, H.-H. Liao, "Integrated HfO_2-RRAM to achieve highly reliable, greener, faster, cost-effective, and scaled devices," *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2017, pp.2.6.1-2.6.4.

[19] K Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition". arXiv 1409.1556. 2014.

[20] S. Yin, Y. Luo, Y. Kim, W. He, X. Han, X. Sun, H. Barnaby, J.-J. Kim, S. Yu, J.-S. Seo, "Monolithically integrated RRAM and CMOS based in-memory computing optimizations for efficient deep learning," *IEEE Micro*, vol. 39, no. 6, pp. 54-63, 2019.

[21] P.-Y. Chen, S. Yu, "Compact modeling of RRAM devices and its applications in 1T1R and 1S1R array design," *IEEE Trans. Electron Devices*, vol. 62, no. 12, pp. 4022-4028, 2015.

[22] S. Yu, and H. -S. P. Wong, "Compact modeling of conducting bridge random access memory (CBRAM)," *IEEE Transactions on Electron Devices*, vol. 58, no. 5, pp. 1352-1360, 2011.