Investigation of Read Disturb and Bipolar Read Scheme on Multilevel RRAM-Based Deep Learning Inference Engine

Wonbo Shim, Yandong Luo, Jae-Sun Seo, *Senior Member, IEEE*,
and Shimeng Yu, *Senior Member, IEEE*

***Abstract*— The multilevel resistive random access memory (RRAM)-based synaptic array can enable parallel computations of vector–matrix multiplication for machine learning inference acceleration; however, any conductance drift of the cell may induce an inference accuracy drop because the analog current is summed up along the column. In this article, the read disturb-induced conductance drift characteristic is statistically measured on a test vehicle based on 2-bit HfO_2 RRAM array. The drift behavior of four states is empirically modeled by a vertical and lateral filament growth mechanism. Furthermore, a bipolar read scheme is proposed and tested to enhance the resilience against the read disturb. The modeled read disturb and proposed compensation scheme are incorporated into a VGG-like convolutional neural network for CIFAR-10 data set inference.**

***Index Terms*— In-memory computing, multilevel resistive random access memory (RRAM), neural network, read disturb, reliability.**

I. Introduction

DEEP learning is the hottest topic in recent years for academia and industry. Deep neural networks (DNNs) have achieved significant success in various tasks such as image classification, speech recognition, and object detection. State-of-the-art deep learning algorithms are aggressively increasing the depth and size of the network to achieve accuracy enhancement, which demands a tremendous amount of computation. Consequently, the data movement between the microprocessor and off-chip memory suffers from excessive power consumption and memory bandwidth limitation. To overcome these challenges, in-memory computing has emerged as an alternative paradigm owing to its high throughput and energy efficiency [1]. Several types of nonvolatile memories (NVMs) have been investigated, such as resistive random access memory (RRAM) [1]–[6], phase change random access memory (PCRAM) [7], [8], flash memory [9]–[12], as a synaptic device for vector–matrix multiplication or weighted sum computation.

RRAM has already proven its multilevel state characteristics [13]–[16] for in-memory computing. However, the reliability requirement of RRAM cell is much more stringent for the multilevel synaptic devices than for the conventional multilevel cell (MLC) data storage. This is because any incremental drift in the conductance of the cell may affect the weighted sum, since the current is summed up along the column, while only the overlapping cells with a significant shift over the neighboring states induce error bits in the MLC storage. Moreover, owing to the analog read-out nature, the digital error correction code (ECC) cannot be applied to the in-memory computing. Prior studies [17], [18] have investigated the endurance and the data retention of the multilevel RRAM cells for in-memory computing, but the read disturb effect has not been addressed in detail. For the deep learning inference engine with read-intensive workloads, the read disturb is a more serious problem than the endurance characteristics. The intermediate states deserve a more comprehensive characterization.

In this article, we statistically measured a 2-bit HfO_2 RRAM based 1-transistor-1-resistor (1T1R) array fabricated at Winbond's 90-nm process [19], and we modeled the resistance drift behavior induced by read disturb with a vertical and lateral filament growth mechanism. This article is an extension of our conference paper [20]. To enhance the resilience against the read disturb, additional work has been performed by a bipolar read scheme validated with experimental data. The peripheral circuit modification is also discussed to support this bipolar read scheme. The modeled conductance drift and the effect of new scheme are incorporated in the Tensorflow software simulation to evaluate the inference accuracy degradation of the VGG-8 network [21] with CIFAR-10 data set.

II. Read Disturb Measurement and Modeling

Fig. 1(a) shows a simple neural network structure with vector–matrix multiplication and Fig. 1(b) shows a RRAM-based 1T1R synaptic array. Conductance matrices of the RRAM cell array correspond to the weight matrices of the neural network layer, and the current summed up along each bitline (BL) is the weighted sum value. Fig. 1(c) shows the die

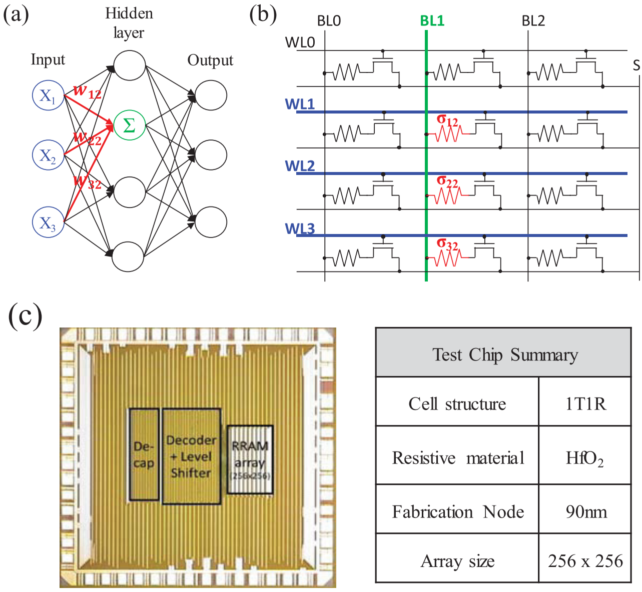
Manuscript received February 18, 2020; revised March 19, 2020; accepted March 30, 2020. Date of publication April 17, 2020; date of current version May 21, 2020. This work was supported in part by ASCENT through the NSF-CCF-1903951, in part by the Semiconductor Research Corporation/Defense Advanced Research Projects Agency (SRC/DARPA) Joint University Microelectronics Program (JUMP) Centers, and in part by the NSF/SRC E2CDA Program. The review of this article was arranged by Editor J. Kang. *(Corresponding author: Shimeng Yu.)*

Wonbo Shim, Yandong Luo, and Shimeng Yu are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: wshim30@gatech.edu; shimeng.yu@ece.gatech.edu).

Jae-Sun Seo is with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281 USA.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2020.2985013



Test Chip Summary	
Cell structure	1T1R
Resistive material	HfO_2
Fabrication Node	90nm
Array size	256 x 256

Fig. 1. (a) Weight matrix between the DNN layers. (b) 1T1R array configuration. (c) Die photograph of tested chip and summary.

photograph of our test vehicle (256 × 256 array) fabricated in Winbond's 90-nm node process.

We realized the 2-bit per cell distribution with the tested RRAM chip, and corresponded each state to the 2-bit weight value. Fig. 2 shows the cumulative probability distribution of four states. The 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 four states where state 1 is the high resistance state (HRS) representing weight −1, and states 2/3/4 are linearly spaced in conductance in the low resistance state (LRS) representing weights −0.5, 0, and 0.5, respectively. In this article, our focus is to characterize the read-disturb; thus, the write-verify scheme is not employed here to tighten the distribution. Ideally, the write-verify protocol is required for programming the conductance to the exact level needed for in-memory computing as demonstrated in our previous work [22]. Chip testing in this article was done at room temperature.

A. Read Disturb on HRS

The resistance change of state 1 (HRS) is modeled by the vertical filament growth, as shown in Fig. 3(a), where the tunneling gap growth rate (*dg/dt*) is exponential to BL voltage under the read stress. Due to the positive feedback mechanism [23], it induces a sudden gap g decrease at some point, and makes the abrupt resistance decrease from HRS to LRS.

The resistance drift of state 1 due to the read stress time up to 500 ms is measured at different BL voltages, as shown in Fig. 4. The total time 500 ms is equivalent to the 5 × 10^7 read cycles if we assume the sense amplifier circuit is designed to require 10 ns per one time read. Resistance of the cells abruptly drops at various stress times during the 500 ms stress. 10% of the cells with low initial resistance are affected by 0.4 V with 500-ms stress, but 10% of the cells with high initial resistance do not change the resistance even at 0.7 V with 500-ms stress. It indicates that the lower the initial resistance is, the faster the resistance drops to the LRS, which implies

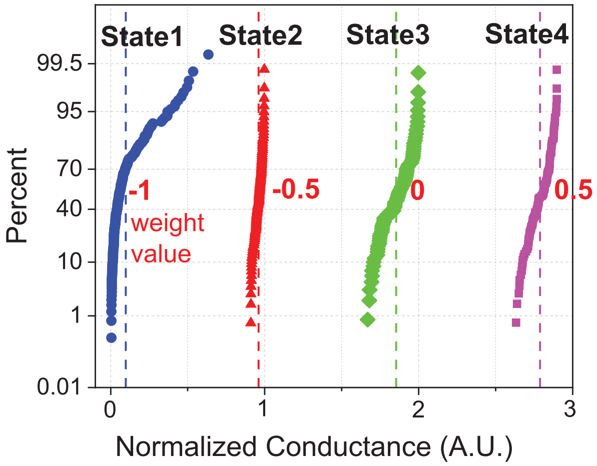


Fig. 2. Cumulative distribution function of initial conductance of 2-bit RRAM. State 2/3/4 was achieved with different SET bias conditions.

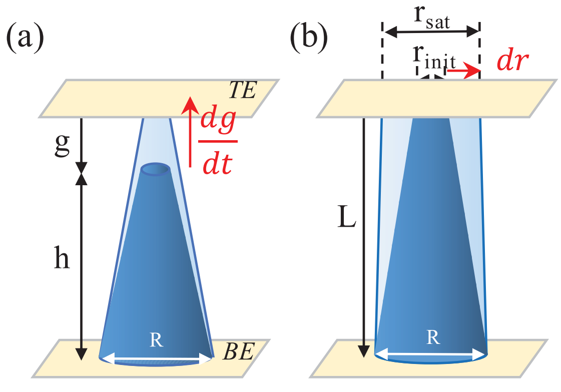


Fig. 3. Filament growth mechanism of (a) state 1 (b) state 2/3/4 during the read stress. Vertical filament growth (gap decrease) induces abrupt resistance change in state 1, while lateral growth induces gradual resistance decrease at other states.

positive feedback occurring time is closely related to the tunneling gap and initial resistance of HRS cell. Therefore, the lower limit of resistance of the HRS must be designed with the consideration not only for the low OFF-current for weighted sum computation, but also for good immunity to the read disturb.

By modifying the compact model of the RRAM device [23], the resistance change during the vertical filament growth can be expressed as the following equations:

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

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

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

where R is the resistance of the filament, *dg/dt* is the gap growth rate, the gap g has the 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., 50 mV throughout this article. γ is g-dependent local field enhancement factor representing voltage dependence to filament height growth, γ_0 is fitting

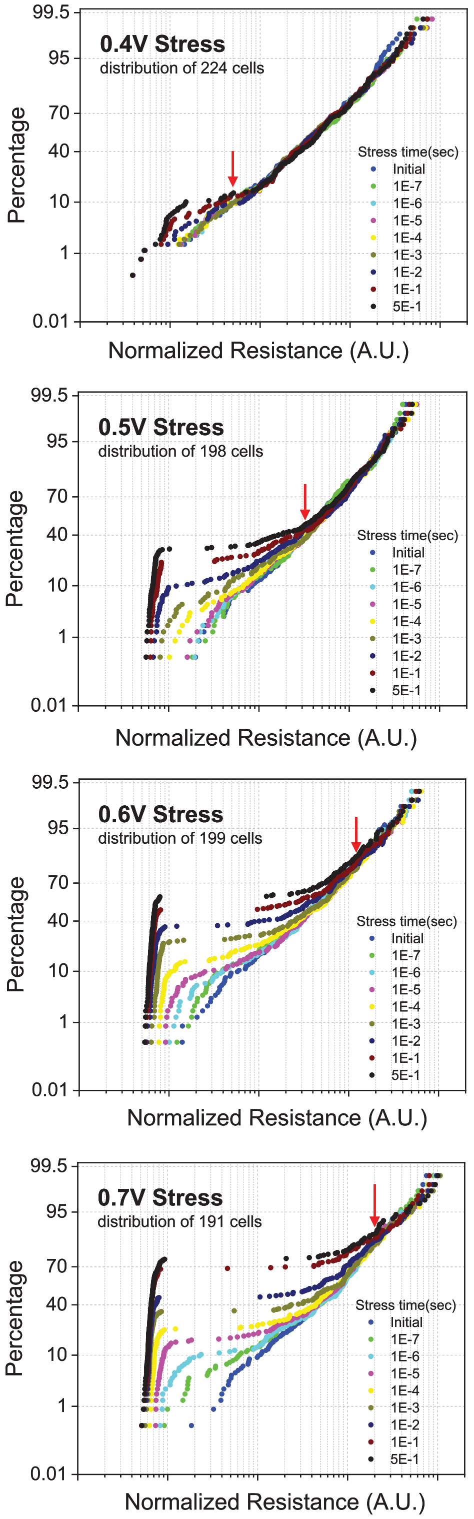


Fig. 4. BL voltage induced resistance change of state 1 RRAM cells during the read stress testing. The cells with low initial resistance are vulnerable to read disturb. The critical resistance (red arrows) to read disturb before 500-ms read stress (equivalent to 5 × 10^7 read cycles assuming each read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

parameter, and a_0 is atomic hopping distance. Two or three representing cells per BL voltage which have substantially equal initial resistance are fitted to the equation as shown in Fig. 5 (line) with the fitting parameters as shown in Table I.

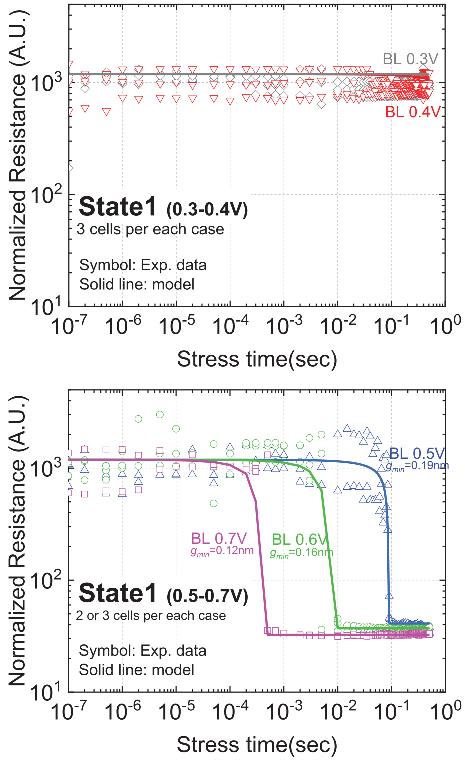


Fig. 5. Read disturb of state 1 cells as a function of stress time.

TABLE I
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	0.25nm		(State4) 17.8nm
E_a	0.8eV	$t_{ch}(V)$	6500×exp(-38V+0.7) sec
r_0	25	$t_{sat}(V)$	$10^{-14.7V+6.7}$ sec
β	1	c_{sat}	2
g_0	0.6nm		

B. Read Disturb on LRS

The resistance change of states 2/3/4 is modeled with lateral filament growth as shown in Fig. 3(b). Assuming that the cone-shaped filament has a fixed bottom radius (R_{bot}) at any time, the resistance of the filament of LRS is dominated by the top radius r. Since the mechanism of lateral filament growth is analogous to the vertical filament growth, the speed of lateral filament growth is also exponential to the BL voltage. But the growth of the radius r of the filament gradually saturates when it approaches the limit, i.e., when filament growth is confined by the size of the RRAM cell itself. The saturation top radius and the initial top radius of filament affect the speed of filament growth (dr/dt).

Fig. 6 shows the measured (symbols) resistance change of states 2/3/4. The measured cells have initial conductance corresponding to weights −0.5, 0, and 0.5 as in Fig. 2. Irrespective of the initial resistance and read stress voltage, the resistance

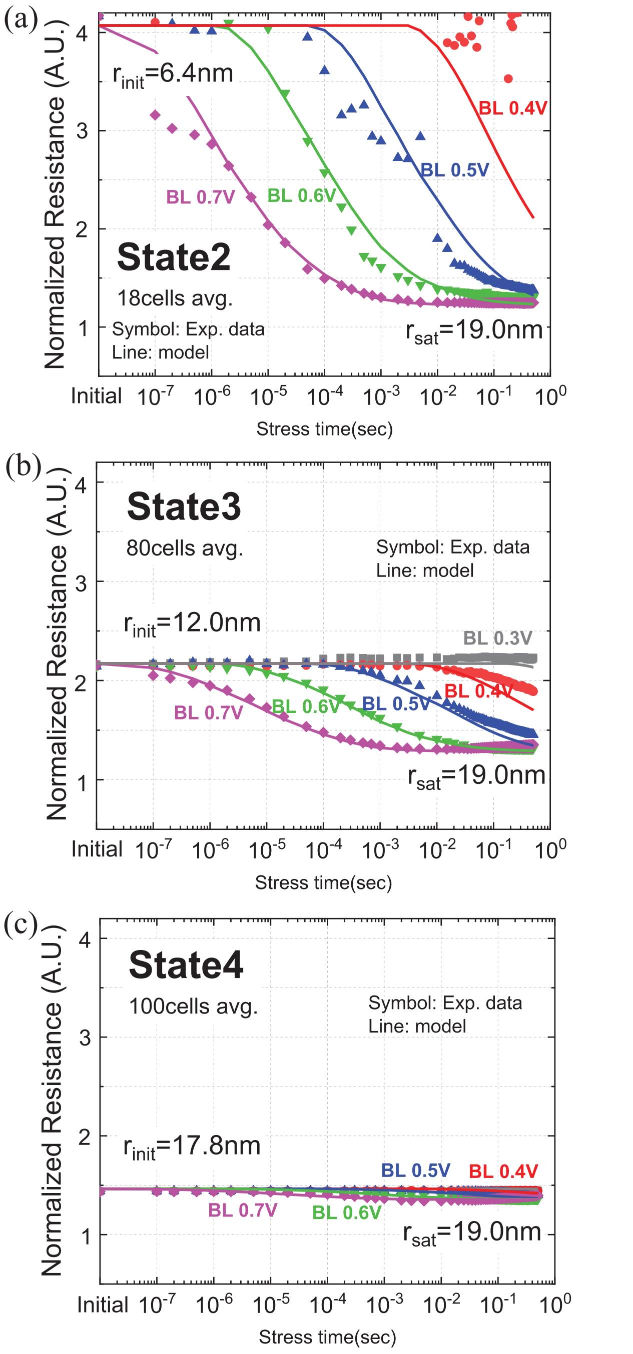


Fig. 6. Read disturb of (a) state 2, (b) state 3, and (c) state 4. Regardless of the initial resistance, all the states have similar saturation resistance and saturated eventually. The model fits the experimental data well using the same set of parameters.

drift always saturates at the same resistance, which indicates that the saturation top radius is only determined by the physical size of the RRAM cell.

Fig. 7(a) shows the maximum resistance drift rate of states 2/3/4 at different BL voltages. This result indicates that the initial state of the cell is the most influential factor of the resistance decrease rate. The time when the resistance drift saturates is shown in Fig. 7(b). It implies that the saturation time is closely correlated with the BL stress voltage, irrespective of the initial state or the resistance drift rate. Therefore, *dr/dt* is the function of the initial radius, and the saturation effect is also included as a function of BL voltage in our lateral filament growth model. As a result, the lateral filament growth phase is following the compact model in (4) [24], and the saturation phase is modeled empirically here as follows:

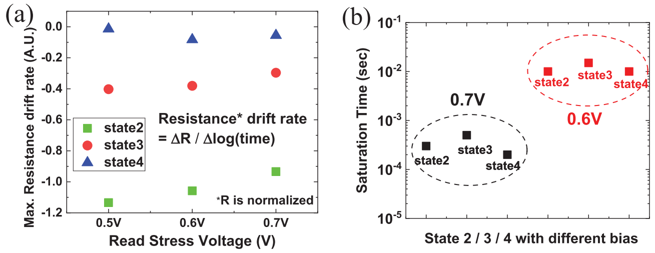


Fig. 7. (a) Maximum resistance drift rate and (b) saturation stress time due to the various condition of read stress.

$$R = \rho_{\mathrm{ON}} L/\pi \cdot r \cdot R_{\mathrm{bot}} \tag{4}$$

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

where *dr/dt* is the radius growth rate, and r_{sat} and r_{init} are the saturation top radius and initial top radius of the filament, respectively. R_{bot} is the saturation bottom radius which is assumed unchanged. The first term of (5) represents the growth phase of the filament radius that is linearly proportional to the difference of r_{sat} and r_{init}, and the last term represents the saturation phase as a function of the stress voltage only. As shown in Fig. 6 (lines), the measured data are well fitted to this model at all the stress conditions with one common set of parameters in Table I.

III. Inference Accuracy Simulation

To quantify the read disturb effect on 2-bit RRAM-based inference engine, we simulated the inference accuracy degradation in the VGG-8 network [21] [as shown in Fig. 8(a)] with CIFAR-10 data set. The 2-bit weight values are pretrained by software simulation, and each of the 2-bit weights is mapped to the four states of one RRAM cell for inference simulation. The 4-bit activations are used, and the maximum accuracy that can be achieved by software simulation is 91.72%.

As discussed in Section II, we assumed that each of read cycles takes 10 ns. The conductance drift at different read cycles and different BL voltages (V_{read}) is achieved through the modeled equations (1)–(5), then the ratio of change over the initial conductance is incorporated into the pretrained weight to simulate the effect on the inference accuracy.

The inference accuracy simulation up to 5 × 10^7 read cycles is shown in Fig. 8(b). If the BL voltage is 0.7 V, the inference accuracy abruptly drops to the 10% within 20 read cycles which is the same accuracy as random selection for CIFAR-10 data set. If the BL voltage is 0.3 V, the initial inference accuracy (91.72%) can be sustained up to 2 × 10^7 read cycles. Accuracy degradation is mainly caused by the cells with state 2 and state 3 because of a relatively weak disturb immunity and significant conductance increase than that of state 1 and state 4. Even a smaller read voltage (e.g., BL $<$ 0.3 V) could potentially be less susceptible to read disturb. However, the reduced sense margin by smaller read voltage will impose challenges to the peripheral circuits design.

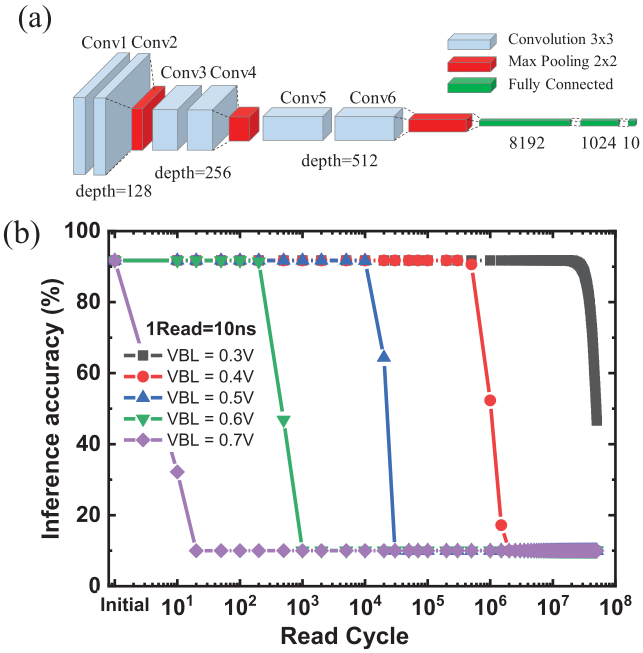


Fig. 8. (a) Simulated VGG-8 network architecture and (b) CIFAR-10 data set inference accuracy simulation result due to read disturb.

IV. Bipolar Read Scheme

As discussed in Section III, the conductance drift in the RRAM-based inference engine induces a significant accuracy drop, which infers that the hardware lifetime is quite limited for read-intensive inference workloads. Therefore, we proposed a bipolar read scheme, and implemented the peripheral circuit modification as shown in Fig. 9. The conductance of the cells is sustained in the acceptable range (e.g., ±1% of initial resistance) under the 0.4 V stress by switching the BL and sourceline (SL) voltage after the predetermined number of read operation. Both BL and SL of every cell are connected to the MUX, and the current sense amplifier can sense the weighted sum current from either BL or SL at the read operation by configuring the MUX setup with a read cycle counter.

Fig. 10 shows the measured maximum time for positive (t_{POS}) and negative (t_{NEG}) stress to constrain the resistance into ±1% of the initial resistance for each LRS. t_{POS}, t_{NEG}, and the ratio of positive and negative time ($t_{\mathrm{POS}}/t_{\mathrm{NEG}}$) are different for each state. Because the cells of state 2 and state 3 are the intermediate states which have relatively low stability in the viewpoint of the a weak filament, the rates of both positive and negative resistance change are so fast, i.e., the shorter time is allowed to confine the resistance in desired range.

Although the ideal value and ratio of maximum t_{POS} and t_{NEG} is different for all states, t_{POS} and t_{NEG} cannot be chosen differently for each state because multiple cells with different states are connected to one BL, and they are activated at the same time during the weighted sum operation. So we chose t_{POS} and t_{NEG} values which are optimized for state 2 because the cells of state 2 will be degraded faster than other states and may affect the accuracy drop most.

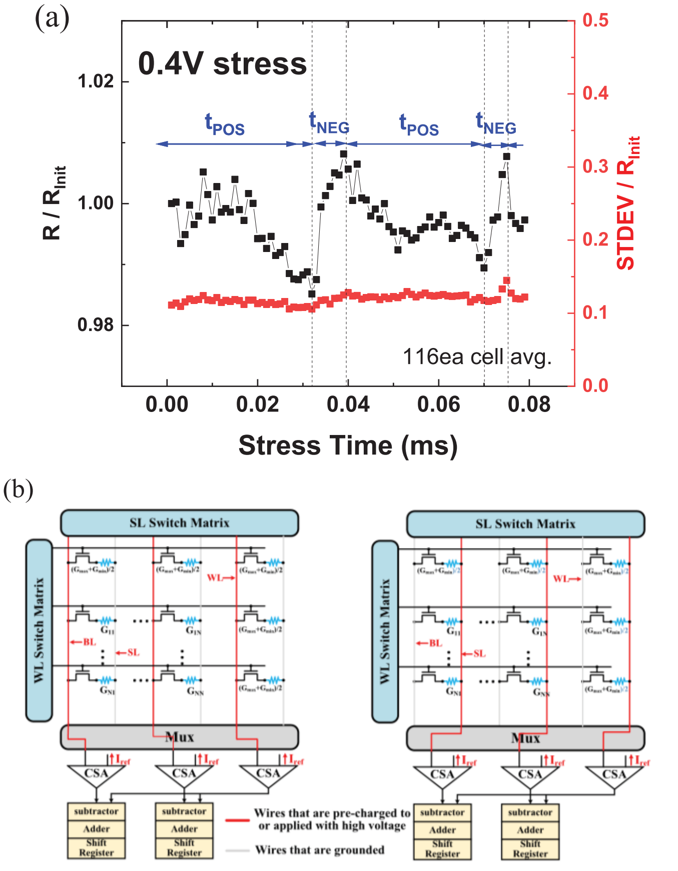


Fig. 9. (a) Resistance constraint effect (±1%) by bipolar read scheme for state 3 at 0.4 V stress and the standard deviation of 116 cell are shown and (b) peripheral circuit configuration for BL and SL voltage switching.

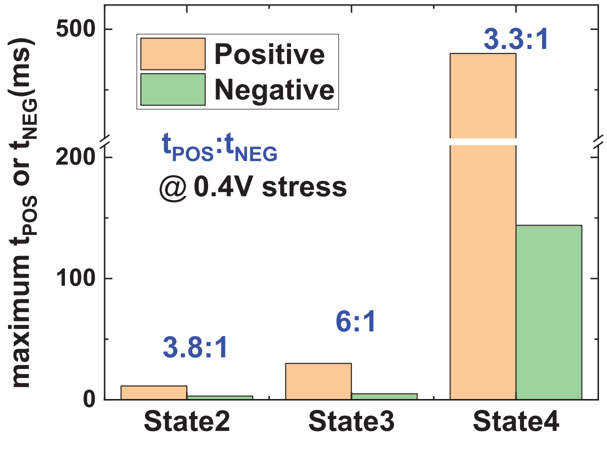


Fig. 10. Maximum time for positive and negative stress to constrain the resistance into ±1% of initial resistance under 0.4 V stress for each LRS.

Fig. 11 shows the simulated accuracy degradation result with optimized t_{POS} and t_{NEG} for state 2. BL and SL voltages switch each other when the accumulated read time reaches the 5.4 ms for positive direction and 1.4 ms for negative direction. The bipolar scheme can maintain the inference accuracy above 80% up to 1.36 M read cycle that is 2.5 times longer than the normal read at 0.4-V BL voltage. However, the conductance of the cells with state 3 oscillates out of acceptable range and induces inference accuracy oscillation. To fully exploit the advantage of bipolar read scheme, we can use the WAGE [25] mapping method to be states 1, 2, and 3 mapped to the

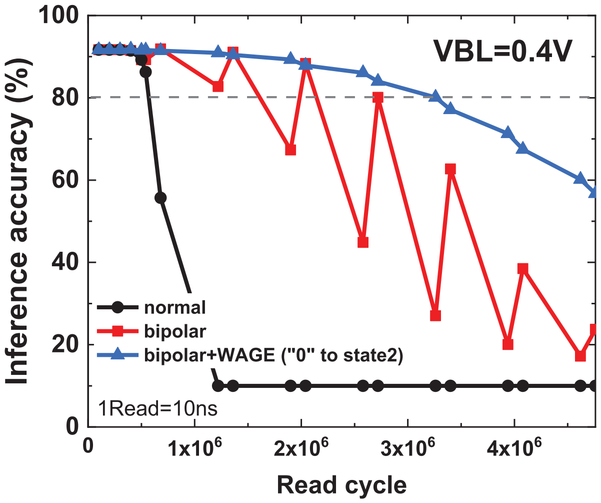


Fig. 11. Inference accuracy enhancement by bipolar read scheme. $t_{POS} = 5.4$ ms, $t_{NEG} = 1.4$ ms.

weight −0.5, 0, and 0.5, respectively. Because the weight 0 has the largest impact on the inference accuracy, WAGE mapping can reduce the accuracy oscillation with the >80% accuracy up to 3.26 M read cycle at this bipolar read condition, where the conductance of the cell with state 2 is very stable.

V. Conclusion

Read disturb-induced conductance drift in multilevel RRAM was measured in a 64-kb test chip. Degradation behavior of four states are modeled by filament vertical and lateral growth mechanism and incorporated in deep learning inference simulation. Conductance drift degraded inference accuracy notably when read voltage is >0.3 V, inferring that read voltage should be minimized with a possible tradeoff of the sense margin. Bipolar read scheme is proposed to alleviate the conductance degradation, so that the inference accuracy can be maintained six times longer against read disturb. Further, device engineering and circuit design techniques are to be developed to enhance the inference lifetime for the RRAM-based inference engine.

Acknowledgment

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

References

[1] S. Yu, "Neuro-inspired computing with emerging nonvolatile memorys," *Proc. IEEE*, vol. 106, no. 2, pp. 260–285, Feb. 2018, doi: 10.1109/JPROC.2018.2790840.

[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*, vol. 521, no. 7550, pp. 61–64, May 2015, doi: 10.1038/nature14441.

[3] J. Woo *et al.*, "Improved synaptic behavior under identical pulses using AlO_x/HfO_2bilayer RRAM array for neuromorphic systems," *IEEE Electron Device Lett.*, vol. 37, no. 8, pp. 994–997, Aug. 2016, doi: 10.1109/LED.2016.2582859.

[4] F. Cai *et al.*, "A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations," *Nature Electron.*, vol. 2, no. 7, pp. 290–299, Jul. 2019, doi: 10.1038/s41928-019-0270-x.

[5] W. Wu *et al.*, "A methodology to improve linearity of analog RRAM for neuromorphic computing," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2018, pp. 103–104, doi: 10.1109/VLSIT.2018.8510690.

[6] T. Gokmen and Y. Vlasov, "Acceleration of deep neural network training with resistive cross-point devices: Design considerations," *Frontiers Neurosci.*, vol. 10, p. 333, Jul. 2016, doi: 10.3389/fnins.2016.00333.

[7] S. Ambrogio *et al.*, "Equivalent-accuracy accelerated neural-network training using analogue memory," *Nature*, vol. 558, pp. 60–67, Jun. 2018, doi: 10.1038/s41586-018-0180-5.

[8] W. Kim *et al.*, "Confined PCM-based analog synaptic devices offering low resistance-drift and 1000 programmable states for deep learning," in *Proc. Symp. VLSI Technol.*, Jun. 2019, pp. 66–67, doi: 10.23919/VLSIT.2019.8776551.

[9] X. Guo *et al.*, "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2017, pp. 6.5.1–6.5.4, doi: 10.1109/IEDM.2017.8268341.

[10] P. Wang *et al.*, "Three-dimensional NAND flash for vector–matrix multiplication," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 27, no. 4, pp. 988–991, Apr. 2019, doi: 10.1109/TVLSI.2018.2882194.

[11] H.-T. Lue *et al.*, "Optimal design methods to transform 3D NAND flash into a high-density, high-bandwidth and low-power nonvolatile computing in memory (nvCIM) accelerator for deep-learning neural networks (DNN)," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2019, pp. 38.1.1–38.1.4

[12] X. Gu, Z. Wan, and S. S. Iyer, "Charge-trap transistors for CMOS-only analog memory," *IEEE Trans. Electron Devices*, vol. 66, no. 10, pp. 4183–4187, Oct. 2019, doi: 10.1109/TED.2019.2933484.

[13] X. Sheng *et al.*, "Low-conductance and multilevel CMOS-integrated nanoscale oxide memristors," *Adv. Electron. Mater.*, vol. 5, no. 9, Sep. 2019, Art. no. 1800876, doi: 10.1002/aelm.201800876.

[14] C. Li *et al.*, "Efficient and self-adaptive *in-situ* learning in multilayer memristor neural networks," *Nature Commun.*, vol. 9, no. 1, pp. 1–8, Dec. 2018, doi: 10.1038/s41467-018-04484-2.

[15] M. J. Marinella *et al.*, "Multiscale co-design analysis of energy, latency, area, and accuracy of a ReRAM analog neural training accelerator," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 8, no. 1, pp. 86–101, Mar. 2018, doi: 10.1109/JETCAS.2018.2796379.

[16] V. Milo *et al.*, "Multilevel HfO_2-based RRAM devices for low-power neuromorphic networks," *APL Mater.*, vol. 7, no. 8, Aug. 2019, Art. no. 081120, doi: 10.1063/1.5108650.

[17] M. Zhao *et al.*, "Characterizing endurance degradation of incremental switching in analog RRAM for neuromorphic systems," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2018, pp. 20.2.1–20.2.4, doi: 10.1109/IEDM.2018.8614664.

[18] M. Zhao *et al.*, "Investigation of statistical retention of filamentary analog RRAM for neuromophic computing," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2017, pp. 39.4.1–39.4.4, doi: 10.1109/IEDM.2017.8268522.

[19] C. Ho *et al.*, "Integrated HfO_2-RRAM to achieve highly reliable, greener, faster, cost-effective, and scaled devices," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2017, pp. 2.6.1–2.6.4, doi: 10.1109/IEDM.2017.8268314.

[20] W. Shim, Y. Luo, J.-S. Seo, and S. Yu, "Impact of read disturb on multilevel RRAM based inference engine: Experiments and model prediction," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Dallas, TX, USA, Apr. 2020.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[22] S. Yin *et al.*, "Monolithically integrated RRAM- and CMOS-based in-memory computing optimizations for efficient deep learning," *IEEE Micro*, vol. 39, no. 6, pp. 54–63, Nov. 2019, doi: 10.1109/MM.2019.2943047.

[23] P.-Y. Chen and 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, Dec. 2015, doi: 10.1109/TED.2015.2492421.

[24] S. Yu and H.-S.-P. Wong, "Compact modeling of conducting-bridge random-access memory (CBRAM)," *IEEE Trans. Electron Devices*, vol. 58, no. 5, pp. 1352–1360, May 2011, doi: 10.1109/TED.2011.2116120.

[25] S. Wu, G. Li, F. Chen, and L. Shi, "Training and inference with integers in deep neural networks," 2018, *arXiv:1802.04680*. [Online]. Available: http://arxiv.org/abs/1802.04680