Impact of Multilevel Retention Characteristics on RRAM based DNN Inference Engine

Wonbo Shim[1], Jian Meng[2], Xiaochen Peng[1], Jae-sun Seo[2], Shimeng Yu[1]
[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[2]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ, USA
Email: shimeng.yu@ece.gatech.edu

2021 IEEE International Reliability Physics Symposium (IRPS) | 978-1-7281-6893-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/IRPS46558.2021.9405210

***Abstract*—In this work, the retention characteristics of multilevel HfO_2 resistive random access memory (RRAM) based synaptic array was statistically measured from a 90 nm test chip and modeled at different temperatures. We found that not only the average conductance (especially at the intermediate states) drifts but also the variance of conductance exacerbates at elevated temperatures. To investigate the impact of the synaptic weight drift on deep neural network, the experimental data are modeled into the ResNet-18 simulation with 1-4 weight bit precisions. The result shows that the inference accuracy drops significantly at 55ºC or above, which implies further engineering on RRAM retention or circuit/algorithmic techniques are yet to be applied.**

***Index Terms*--multilevel RRAM, neural network, data retention.**

I. INTRODUCTION

Deep neural networks (DNNs) have achieved significant success to 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 the accuracy enhancement, which demands tremendous amount of computation. Consequently, the data movement between the microprocessor and off-chip memory suffers from excessive power consumption and memory bandwidth limitation in conventional von-Neumann computing architecture such as CPU and GPU. Several CMOS-based application specific integrated circuits (ASIC) accelerators such as Google TPU [1] are proposed as an alternative. However, the memory wall still becomes the bottleneck, where the weight parameters are stored in global buffer and computation is performed in separate digital multiply-and-accumulate (MAC) arrays. Frequent DRAM access is still required because of the limited global buffer capacity.

To overcome these challenges, compute-in-memory (CIM) is proposed as a promising paradigm where the weights are stored in the memory cells and the MAC operation is embedded in memory itself by the weighted sum of analog current along the columns.

Various type of NVMs have been investigated as the synaptic device for CIM application, such as resistive random access memory (RRAM) [2-3], phase change memory (PCM) [4-5], Flash memory [6-7] and ferroelectric field effect transistor (FeFET) [8]. RRAM has attracted great interest to represent the multilevel synaptic weights for accelerating DNNs [9]. Furthermore, multilevel RRAM enables larger MAC throughput with higher memory density [10]. However, multilevel RRAM based CIM for inference applications suffer from the non-ideal characteristics such as read disturb [11]. It should be noted that the requirement on the retention of synaptic weight memory for CIM is more stringent than the conventional multilevel cell (MLC) storage, because any conductance drift of the devices is summed up along the column so that the error bits can be induced at the analog-to-digital converter (ADC) quantized result, as shown in Fig. 1.

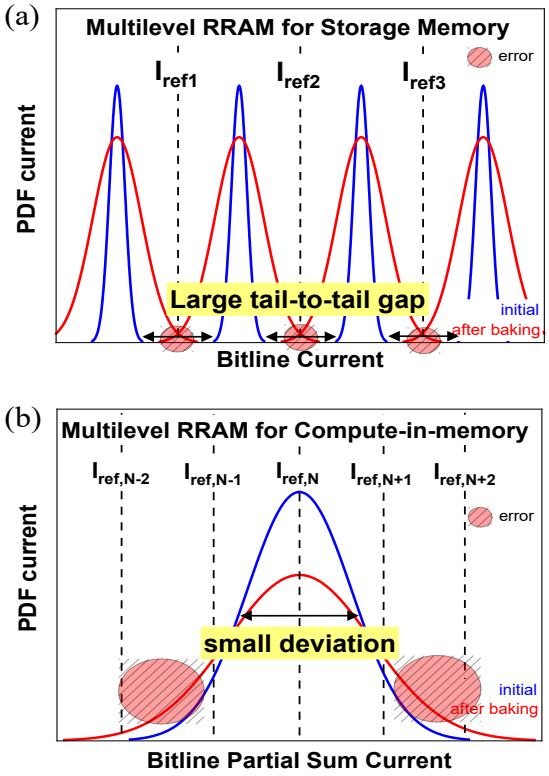

Figure 1. The requirement for multilevel RRAM cells for (a) storage memory and (b) compute-in-memory applications.

In this work, we tested the HfO_2 based 1T1R 64kb array fabricated at 90 nm process [12]. The retention characteristics of 2-bit RRAM cells are statistically measured and modeled.

W. Shim and J. Meng contributed equally to this work. This work is supported by ASCENT and C-BRIC, two of the SRC/DARPA JUMP Centers and NSF/SRC E2CDA program.

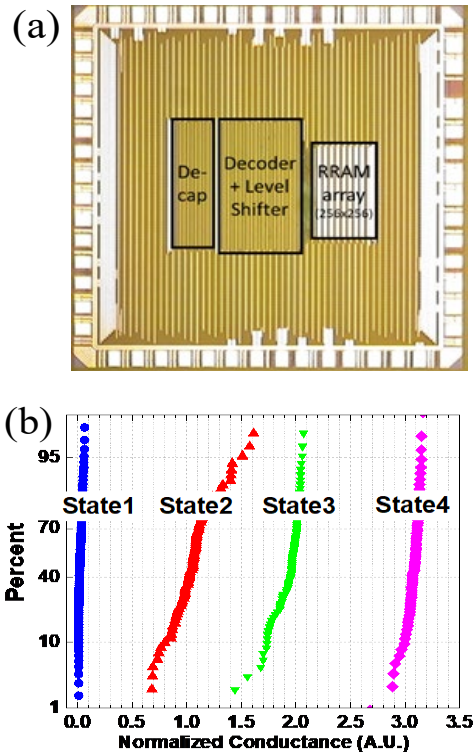

Figure 2. (a) Die photo of the measured 64kb HfO2 based 1T1R RRAM chip. (b) Initial multilevel cell conductance distribution after write verify.

The retention model is incorporated into the inference accuracy simulation of ResNet-18 model [13] on CIFAR-10 dataset.

II. Retention Measurement and Modeling

Fig. 2(a) shows the die photo of the 256×256 1T1R HfO_2 based RRAM test chip with peripheral circuits. We realized the 2-bit per cell distribution with the RRAM test chip measurements. For inference operation, the weight is proportional to the conductance, thus we designed four states as follows. Fig. 2(b) shows the cumulative probability distribution of the initial conductance of multilevel states measured at room temperature (25ºC). The resistances of the State 1 cells are in high resistance state (HRS), while the resistances of the State 2/3/4 cells are in low resistance state (LRS) where each of the states' conductances are linearly spaced to represent the 2-bit weight. The initial conductance distributions were tightened with the two-step write-verify scheme [14]. The conductance of each state was controlled by SET and RESET current during the iterative SET and RESET loops. The bias conditions (V_G and V_D) are optimized respectively for each state.

As shown in Fig. 3, the conductance drift of the RRAM cells in the test chip are measured up to 80,000 seconds at the temperature from 25ºC to 120ºC. The average conductance values are displayed for every state. The average conductance of the cells decreases over the baking time where the conductance drift rate is significant in State 2 and State 3. The State 2 and State 3 cells are the intermediate states which have relatively low stability in the viewpoint of weak filament.

The conductance not only drifts but also fluctuates over time. The ratio of standard deviation (σ) over the average conductance (μ) of the State 2 cells also increases significantly

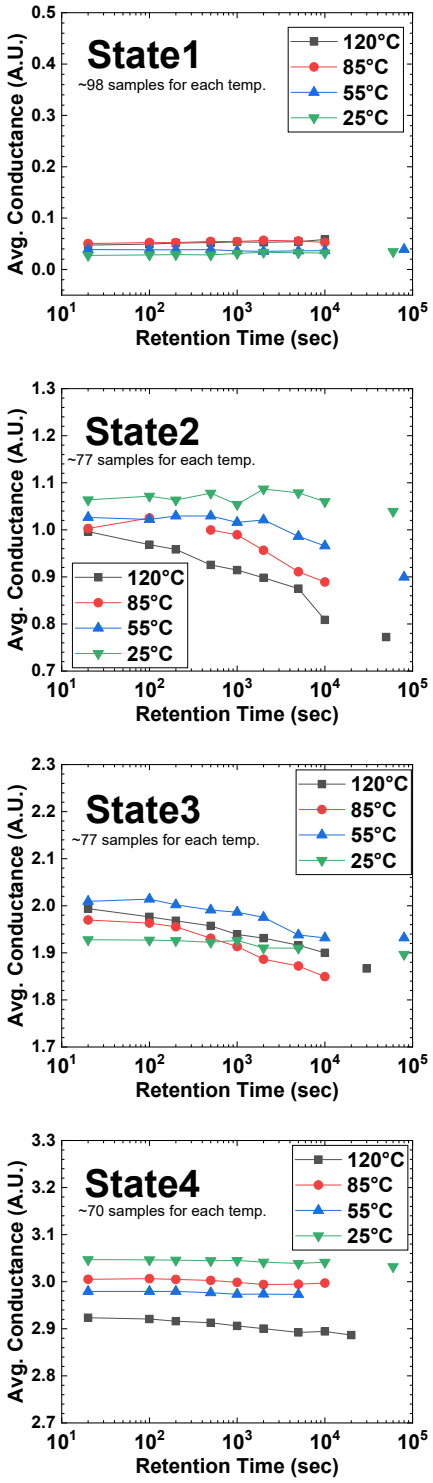

Figure 3. Measured average of conductance over time for each state at different temperatures.

over time and is accelerated by high temperature as shown in Fig. 4. Although σ/μ of the State 1 cells are very high and fluctuate over time, the conductance of the State 1 cells is several tens of times lower than that of the cells of other states.

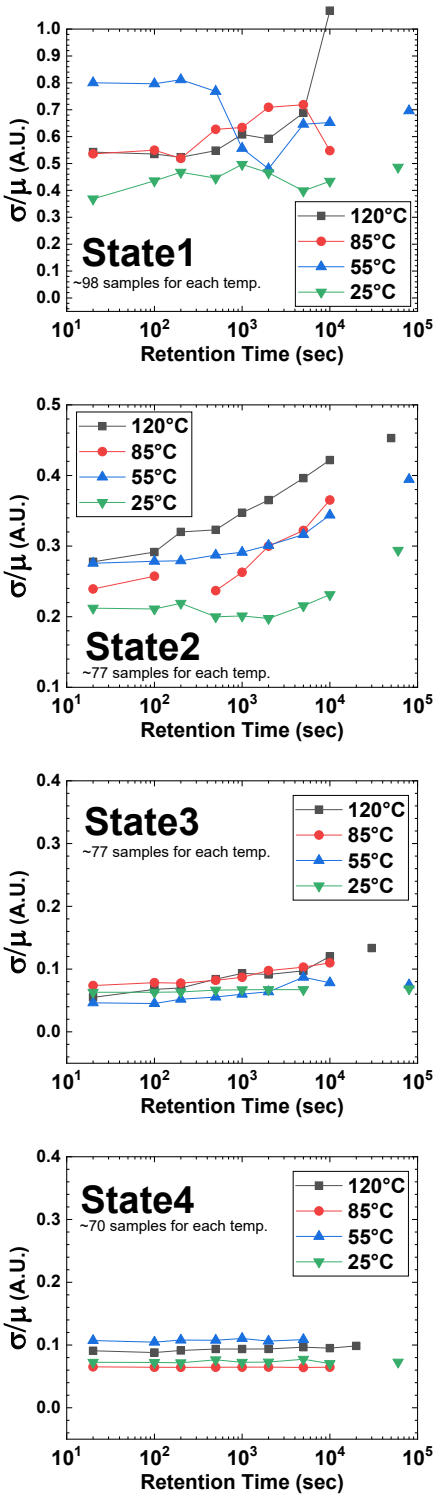

Figure 4. Measured sigma (σ) over average (μ) of conduction over time for each state at different temperatures.

Therefore, the high σ/μ ratio of State 1 cells does not significantly affect the total weighted sum current.

We modeled the measured retention characteristics in the following equations. The Δμ is fitted linearly to the logarithmic time as (1), where A_{avg} is the average conductance drift rate that

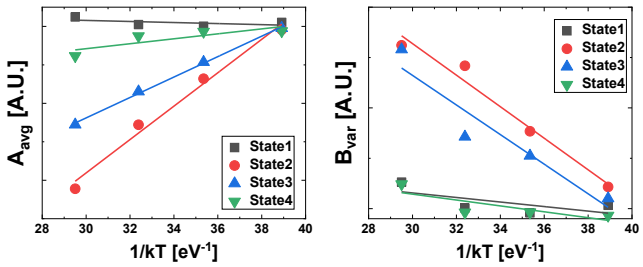

Figure 5. Temperature dependency of the conductance drift rate and fitting result on 1/kT plot for each state.

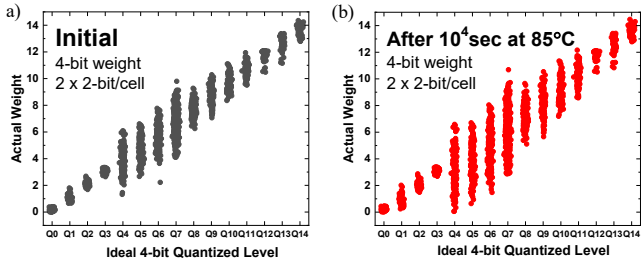

Figure 6. Weight distribution dispersion with retention model at (a) initial condition and (b) after 10^4 sec at 85ºC for 4-bit weight precision.

depends on states and temperatures. The Δσ is also fitted as (2) with sigma conductance drift rate (B_{var}).

$$\Delta\mu = \mu(t) - \mu_{init} = A_{avg} \times log\ t \quad (1)$$

$$\Delta\sigma = \sigma(t) - \sigma_{init} = B_{var} \times log\ t \quad (2)$$

Fig. 5 shows the temperature dependency of the average and sigma of conductance drift rate (A_{avg} and B_{var}). Because the cells of State 2 and State 3 are the intermediate states which have relatively low stability in the viewpoint of a weak filament, the activation energy on filament deformation is low.

III. Inference Accuracy Simulation

We incorporated such 2-bit RRAM retention model into the ResNet-18 network to simulate the DNN inference accuracy on CIFAR-10 dataset. We adopted the PACT quantizer [15] for 2-bit/4-bit training and the BNN training [16] method to generate the pre-trained models in this work.

To implement the 4-bit weight into the RRAM array, we mapped two 2-bit per cell RRAM cells to one weight. The actual 4-bit weight has non-ideally quantized distribution as shown in Fig. 6, where the retention baking especially degrades the distribution of the quantized levels mapped with State 2 and 3. Fig. 7(a) implies that the inference accuracy drops significantly even when only the average drift model is considered. Considering the degradation of variation aggravates the accuracy further as shown in Fig.7(b).

We also simulated ResNet-18 network with 2-bit/1-bit weights as shown in Fig. 8. DNNs with lower weight precision show higher robustness and alleviate the accuracy loss, while trading off lower initial accuracy. It should be noted that the inference accuracy simulation results heavily depend on the training algorithm techniques including the neural network

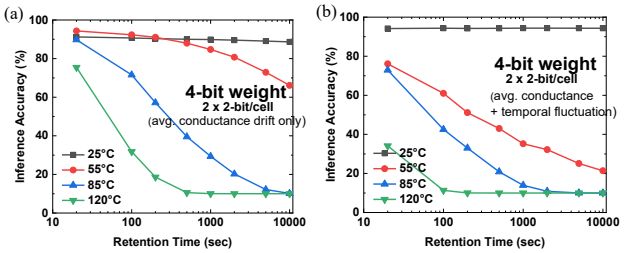

Figure 7. Simulated inference accuracy of ResNet-18 with (a) average conductance drift model only and (b) after adding temporal fluctuation model incorporated for 4-bit weight (two 2-bit per RRAM cells).

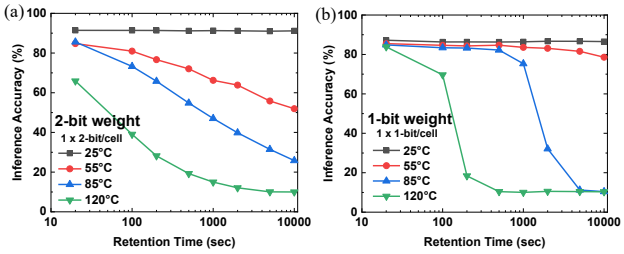

Figure 8. Simulated inference accuracy on ResNet-18 with (a) 2-bit weight (one 2-bit per cell) and (b) binary weight (one 1-bit per cell using only State 1 and 4).

topology. Different quantization schemes will also impact the distributions of the conductance states.

IV. Conclusions

The retention characteristics of multilevel resistive random access memory (RRAM) are measured and modeled, then the effects on inference accuracy degradation are investigated. Different from the conventional MLC storage, multi-bit RRAM based inference engine requires more stringent retention characteristics to maintain the inference accuracy. While we assumed the reference voltage of ADCs are fixed in this work, reference voltage generation using dummy columns with additional RRAM cells may compensate the retention induced conductance drift (but cannot fully compensate the temporal fluctuation). Further algorithmic techniques or refresh schemes at circuit-level are also required to mitigate the accuracy drop of the CIM DNN accelerators.

Acknowledgment

The authors thank Winbond Electronics for RRAM chip fabrication support. Conductance is normalized per non-disclosure-agreement (NDA).

References

[1] N. P. Jouppi, C. Young, N. Patil, D. Patterson G. Agrawal, R. Bajwa, S. Bates, et al., "In-datacenter performance analysis of a tensor processing unit," *ACM/IEEE International Symposium on Computer Architecture (ISCA)*, 2017, pp. 1-12. DOI:https://doi.org/10.1145/3079856.3080246.

[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, DOI: 10.1038/nature14441.

[3] 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, DOI: 10.1038/s41928-019-0270-x.

[4] S. Ambrogio, P. Narayanan, , 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*, pp. 60–67, 2018, DOI: 10.1038/s41586-018-0180-5.

[5] W. Kim, R. L. Bruce, T. Masuda, 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, DOI: 10.23919/VLSIT.2019.8776551.

[6] 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, DOI: 10.1109/IEDM.2017.8268341.

[7] H. -T. Lue, P. -K. Hsu, M. -L. Wei, T. -H. Yeh, P. -Y. Du, W. -C. Chen, K. -C. Wang and C. -Y. Lu, "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)," *IEEE International Electron Devices Meeting (IEDM),* San Francisco, CA, 2019, pp. 38.1.1-38.1.4, DOI: 10.1109/IEDM19573.2019.8993652.

[8] K. Ni , B. Grisafe, W. Chakraborty, A. K. Saha, S. Dutta, M. Jerry, J. A. Smith, S. Gupta, and S. Datta, "In-Memory Computing Primitive for Sensor Data Fusion in 28 nm HKMG FeFET Technology," *2018 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2018, pp. 16.1.1-16.1.4, DOI: 10.1109/IEDM.2018.8614527.

[9] C.-X. Xue et al, "A 1Mb multibit ReRAM computing-in-memory macro with 14.6ns parallel MAC computing time for CNN-based AI edge processors," *IEEE International Solid-State Circuits Conference (ISSCC),* San Francisco, CA, USA, 2019, pp. 388-390, doi: 10.1109/ISSCC.2019.8662395.

[10] 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, doi: 10.1002/aelm.201800876.

[11] W. Shim, Y. Luo, J. Seo and S. Yu, "Investigation of read disturb and bipolar read scheme on multilevel RRAM-based deep learning inference engine," *IEEE Transactions on Electron Devices*, vol. 67, no. 6, pp. 2318-2323, June 2020, DOI: 10.1109/TED.2020.2985013.

[12] 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, doi: 10.1109/IEDM.2017.8268314

[13] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 770-778, Jun. 2016, Las Vegas, USA, DOI: 10.1109/CVPR.2016.90.

[14] W. Shim, J.-S. Seo, S. Yu, "Two-step write–verify scheme and impact of the read noise in multilevel RRAM-based inference engine," *Semicond. Sci. Technol.* **35** 115026, 2020, DOI: 10.1088/1361-6641/abb842.

[15] J. Choi, Z. Wang, S. Venkataramani, P. I. Chuang, V. Srinivasan, K. Gopalakrishnan, "PACT: Parameterized Clipping Activation for Quantized Neural Networks", *International Conference on Learning Representations (ICLR)*, Apr. 2018, Vancouver, Canada, arxiv.org/abs/1805.06085v2.

[16] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, Y. Bengio, "Binarized Neural Networks", *30th Conference on Neural Information Processing Systems (NIPS),* 2016, Barcelona, Spain. arxiv.org/abs/1602.02505.