PAPER

Two-step write–verify scheme and impact of the read noise in multilevel RRAM-based inference engine

IOP Publishing Semiconductor Science and Technology
Semicond. Sci. Technol. **35** (2020) 115026 (7pp) https://doi.org/10.1088/1361-6641/abb842

Two-step write–verify scheme and impact of the read noise in multilevel RRAM-based inference engine

Wonbo Shim[1], **Jae-sun Seo**[2] **and Shimeng Yu**[1,3]

[1] Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, United States of America
[2] School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ, United States of America

E-mail: shimeng.yu@ece.gatech.edu

Received 14 August 2020, revised 10 September 2020
Accepted for publication 14 September 2020
Published 9 October 2020

Abstract

Accurate cell conductance tuning is critical to realizing multilevel resistive random access memory (RRAM)-based compute-in-memory inference engines. To tighten the distribution of the cells of each state, we developed a two-step write–verify scheme within a limited number of iterations, which was tested on a test vehicle based on HfO_2 RRAM array to realize 2 bits per cell. The conductance of the cells is gathered in the targeted range within 10 loops of set and reset processes for each step. Moreover, the read noise of the RRAM cells is statistically measured and its impact on the upper bound of analog-to-digital converter (ADC) resolution is predicted. The result shows that the intermediate state cells under relatively high read voltage (e.g. 0.2 V) are vulnerable to the read noise. Fortunately, the aggregated read noise along the column will not disturb the output of a 5 bit ADC that is required for a 128 $\times$ 128 array with 2 bits per cell.

Keywords: multilevel resistive random access memory, write-verify, read noise, compute-in-memory, deep learning inference engine

Some figures may appear in colour only in the online journal

1. Introduction

Nowadays, deep learning is one of the spotlighted fields to hardware and software researchers. To realize an accurate and efficient learning process, deep neural network (DNN) has emerged as a powerful solution in various tasks such as image classification, speech recognition, and language translation. To achieve high accuracy in the DNNs, recently proposed works demand larger network size and deeper network depth. Inevitably, the growing size of the DNNs increases the amount of computation significantly.

Hardware based on Von-Neumann architecture (e.g. CPU and GPU) is commonly used for DNN computation today. However, the huge amount of data movement between the processor and the main memory causes bandwidth limitation and lowers the energy efficiency. Several CMOS-based application-specific integrated circuits (ASIC) accelerators, such as Google TPU [1], are proposed as an alternative. However, the memory wall still exists where the parameters are stored in global buffers, and computation is performed in the digital multiply-and-accumulation (MAC) arrays. Frequent DRAM access is still required because of the limited global buffer capacity.

Compute-in-memory (CIM) has been studied as an alternative paradigm owing to its high throughput and energy efficiency [2]. For the realization of CIM in the non-volatile memory (NVM) arrays, the conductance of the memory cell is utilized to represent the weight, and MAC operation is conducted by activating multiple rows simultaneously and then reading out the analog current summed up along the column. This analog computing method enables high parallelism

[3] Author to whom any correspondence should be addressed.

since the dense NVM array, with millions of memory cells, performs computation in parallel. In addition, the computation performed within the memory array reduces the energy consumption caused by data movement between processor and memory.

Various type of NVMs have been investigated as a synaptic device for CIM application, such as resistive random access memory (RRAM) [3–7], phase change memory (PCM) [8, 9], flash memory [10–13], and ferroelectric field effect transistor (FeFET) [14]. Among them, RRAM has been regarded as a strong candidate for a synaptic device due to its logic compatibility and low integration cost, which has been demonstrated at industrial-grade 22 nm platform [15]. Previous works [16–19] have already proven the capability of the multilevel state characteristics of RRAM. However, more tightened cell conductance distribution is required for multilevel synaptic devices than for the conventional multilevel cell (MLC) data storage applications. This is because a shift of any cell's conductance value may affect the result of the analog-to-digital conversion (ADC) quantization since the current is summed up along the column in an analog manner, while only the overlapping cells (e.g. tail bits) with significant conductance deviation over the neighboring states induce errors in the MLC storage. Moreover, spacing each state linearly in the conductance regime is more challenging compared to the conventional MLC application, which has exponentially spaced conductance states between each level.

Therefore, many studies have been conducted to precisely tune the conductance of multilevel RRAM cells. Prior works [20, 21] have proposed the write–verify method for synaptic device applications, and suggested set and reset bias schemes with incremental bitline (BL) voltage and fixed gate voltage (V_G) [20], fixed BL voltage and incremental V_G [21], or a full reset if the cell is overset [22], which may require a large number of iteration loops. In this work, we tested a novel two-step write–verify scheme on an RRAM-based 1-transistor-1-resistor (1T1R) array fabricated at 90 nm process [23] to achieve a tightened 2 bit per cell conductance distribution with a limited number of iteration loops.

In addition, we investigated the effect of the read noise on the RRAM-based CIM application. The read noise of the individual RRAM cell current from multiple activated cells merges on the summed current along the BL. This may cause an error in ADC of the merged current and accuracy degradation of the DNN model as well. Going one step further from the prior works [24–26], we measured the read noise of individual cells with multilevel states on our test chip, and investigated its impact on the upper bound of ADC resolution.

2. Two-step write–verify scheme

Figure 1(a) shows the die photo of our test vehicle, which has 64 kb (256 × 256) RRAM cell array and peripheral circuits. Figure 1(b) shows the 1T1R array configuration of the test vehicle and an example of the synaptic array operation where multiple WLs are activated simultaneously and the cell currents of the activated cells are summed up along the BL.

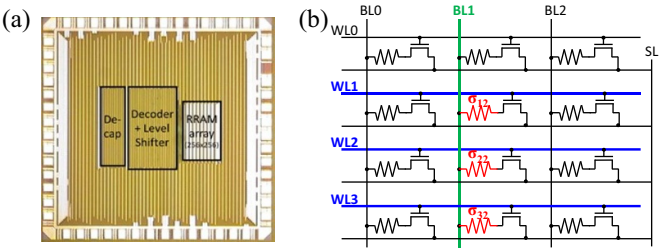


Figure 1. (a) Die photo of the tested chip. (b) 1T1R array configuration of RRAM cell array.

Chip-level statistical measurements were performed using our PXIe testing system from National Instruments.

To realize the multilevel RRAM with tightened cell distributions, we used a novel two-step write–verify scheme as described in figure 2(a). Before starting the first step of write–verify, the reset pulse is applied to all of the cells. The first step of write–verify is conducted with respective set/reset BL voltage, gate voltage (V_G), gate voltage step ($V_{G,STEP}$), and maximum loop. At the first loop of the set operation, all the cells are read and marked as the program (PGM) cells and followed by the first set pulses. From the second loop, the cells having higher conductance than lower-bound conductance (G_{lower}) are marked as inhibited cells, while the cells having lower conductance are marked as PGM cells at the following set operation loop. The set pulse with incremental gate voltage is applied at the following loop only to the PGM marked cells. Even if some of the cells are marked as inhibited in the previous loop, all the cells are read again in every loop to prevent the case in which any kind of conductance variation opportunistically reads the cells within the target range and the cells are passed at the next loop. When the set operation loop reaches the predefined maximum loop (ML1), the bias conditions change to the reset operation mode. During the first loop of the reset operation mode, all the cells are read and followed by the reset pulse. Opposite to the verify operation in set mode, the cells having lower conductance than upper bound conductance (G_{upper}) are marked as inhibited cells for the following reset operation, while the cells having higher conductance are marked as PGM cells. When the reset operation loop reaches ML1, the second step of write–verify is conducted with different set/reset BL voltage, V_G and $V_{G,STEP}$, and maximum loop. The process flow of the second step is identical to the first step while the whole process finishes when the reset operation loop of the second step reaches the ML2.

The advantage of the two-step write–verify scheme comes from its capability to precisely tune the conductance of each cell. Our novel scheme applies relatively high voltage (1.2–2 V) to the BL to coarsely approach the conductance to the target at the first step of write–verify, while only 0.5–0.7 V is applied to the BL at the second step of write–verify to fine-tune the conductance. As can be seen in figure 2(b), lowered BL voltage at the second step helps more gradual conductance control. This enables higher controllability of the conductance than the conventional write–verify processes, which merely use fixed BL voltage or fixed V_G for the whole verify sequence. Figure 2(c) shows the average conductance change per each

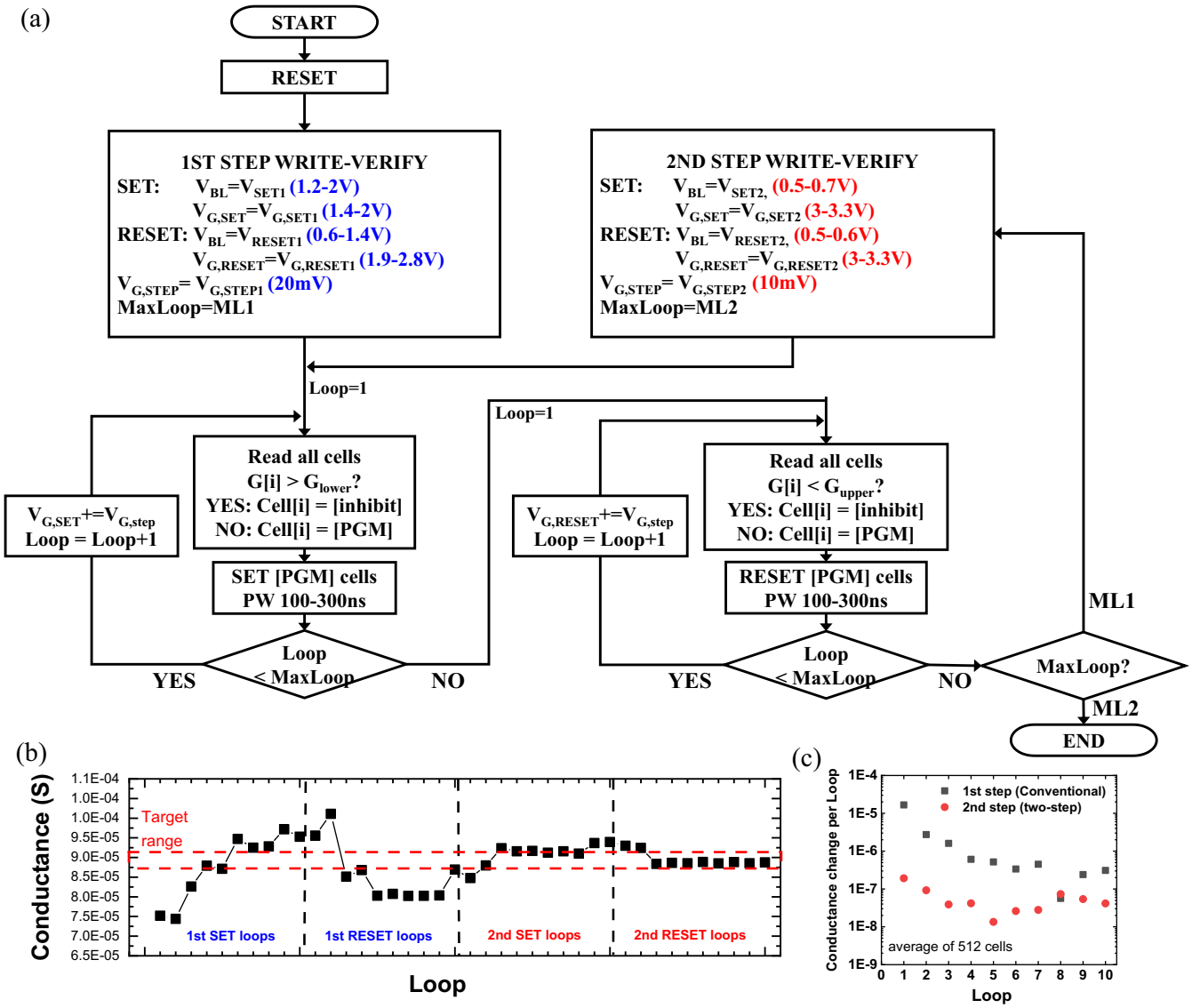


Figure 2. (a) The two-step write–verify scheme for multilevel RRAM. The sequence for one state of the multilevel RRAM is shown here, and it is repeated for other states with different bias conditions. (b) Example of a loop-by-loop conductance tuning process into the target range. (c) Conductance change per each loop during the first and second step write–verify.

loop during the first and second step write–verify. The second step write–verify with lowered BL voltage enables 30–80 times slower conductance change in the beginning loops than the first step write–verify, which can be considered as same as conventional write–verify process (fixed BL voltage throughout the whole process). The complex optimization process and the longer programming time for each state are acceptable for the inference engine application.

We tested and optimized the bias condition of the two-step write–verify scheme with our HfO_2-based RRAM chip as shown in figure 3. We designed four states, where state 1 is the high resistance state (HRS) and states 2/3/4 are linearly spaced in conductance in the low resistance state (LRS) regime. Figure 3(a) shows the conductance distributions of a targeted intermediate state after each verify sequences. The cell distribution after the first set operation (10 loops) indicates that even though the incremental V_G pulse programming was used with 20 mV step, the cell distribution is still too wide to fit into the narrow target conductance range (±3% of target value). The first step reset process lowers the conductance of the cells having higher conductance than the upper bound of the target range. However, >50% of the cells have lower conductance than the lower bound of the target range after the first step reset process. Simply repeating these set/reset processes with the same BL voltage and V_G (first loop) makes it challenging to reduce the ratio of the cells out of the target range. Accordingly, we reduced the BL voltage and increased the V_G, which are optimized for the second step set/reset operations. Since a lower BL voltage facilitates more gradual conductance change of each cell, a greater number of cells have conductance in the target range, even with the same number of maximum loops (ML2 = 10).

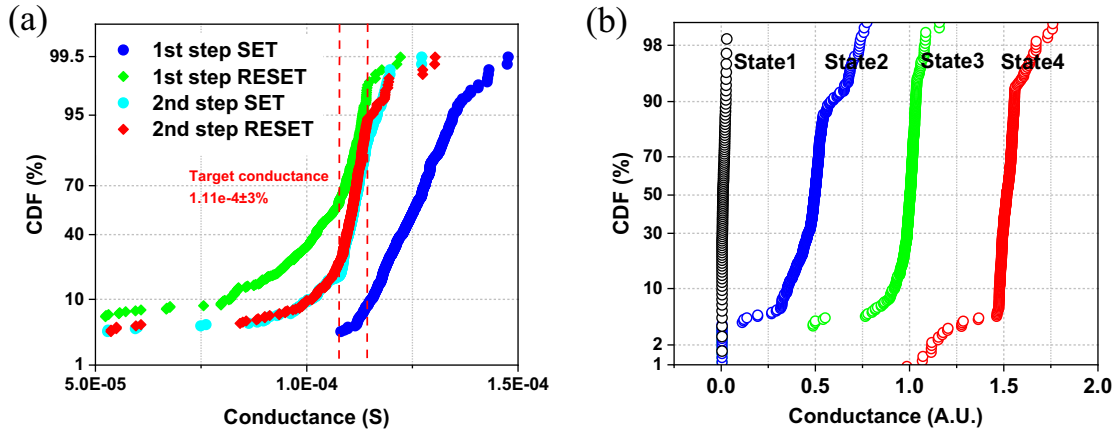


Figure 3. (a) Conductance distribution of measured 512 RRAM cells after first and second step set and reset operation with ML1 = 10, ML2 = 10, $V_{G,STEP1}$ = 0.02 V, $V_{G,STEP2}$ = 0.01 V. (b) Conductance distribution of four states (512 cells in each state) with two-step write–verify scheme.

Figure 3(b) shows the cumulative probability distribution of four states. We realized the tightly distributed states 2 and 3 (512 cells per state) with two-step write–verify scheme for 2 bits per cell. The bias conditions were optimized separately with respect to each state. The distributions of states 1 and 4 were achieved only with one-step write–verify, since the conductance of the LRS and HRS cells drift more gradually relative to the intermediate states, so the two-step verify sequences were not essential.

3. Impact of the read noise of RRAM cell current on weighted sum computation

For the analysis of the noise effect on weighted sum computation in the RRAM array, we tested the read noise on RRAM cell by repeating the read operation 1000 times first. Figure 4 shows the measured conductance data from three different samples. Each read takes 5 ms to calculate the conductance using the source measuring unit (SMU) in our measurement system. The result shows that the conductance of the displayed cells oscillates ±2% due to the random telegraph noise (RTN) within the RRAM cell.

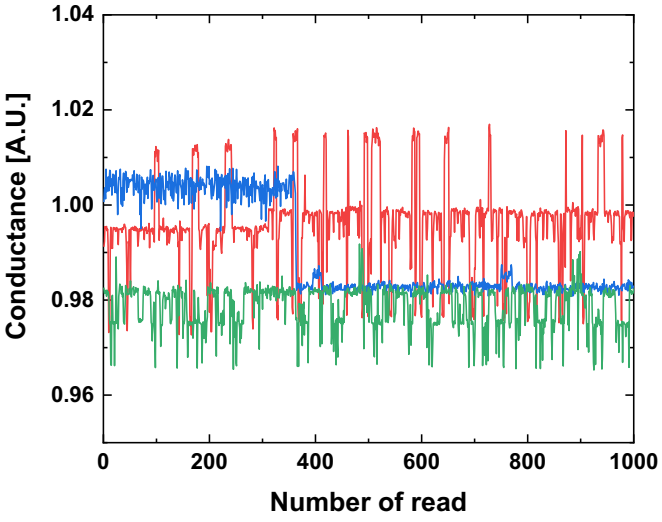


Figure 4. Measured read noise as a normalized conductance for three different cells. Prior to the read noise test, these cells were programmed to be within state 3.

To investigate the effect of noise-induced RRAM conductance variation on weighted sum computation in the synaptic array, we post-processed the measured data of the conductance of 128 cells. We merged the conductance of multiple cells, and then the average and standard deviation of merged conductance are extracted. Assuming the weight sparsity and input sparsity are 50%, integrating the conductance of 8, 16, and 32 cells can respectively represent 32, 64, 128-row weighted sum operation. Figure 5(a) shows that the larger the array size is, the higher the variation of merged conductance due to the accumulated noise along the BL. However, the ratio over the average conductance decreases with larger array sizes, as shown in figure 5(b). This implies that larger arrays may average out the variation of individual cells. In terms of the cell conductance fluctuation, the intermediate state 2 has the largest fluctuation. The intermediate states are generally more vulnerable to the RTN-induced variation. Since the HRS (state 1) cells have negligible effect on the merged current owing to the large on/off ratio of tested RRAM cells, the noise of the HRS is not shown here.

Figure 6 shows the read voltage dependency of noise-induced weighted sum current variation. The read voltage applied to the BL below 0.15 V has less impact on the fluctuation, while 0.2 V read voltage significantly increases the fluctuation. Higher read voltage (>0.25 V) may induce larger RTN-originated noise variation. However, we do not show the result here, because the conductance drift caused by read disturb becomes dominant [27], and the noise-induced variation cannot be distinguished from the read disturb results. In order to reduce the power consumption of the parallel CIM operation on the memory array, it is desirable to use current mode sensing which could use a relatively small (<0.2 V) read voltage

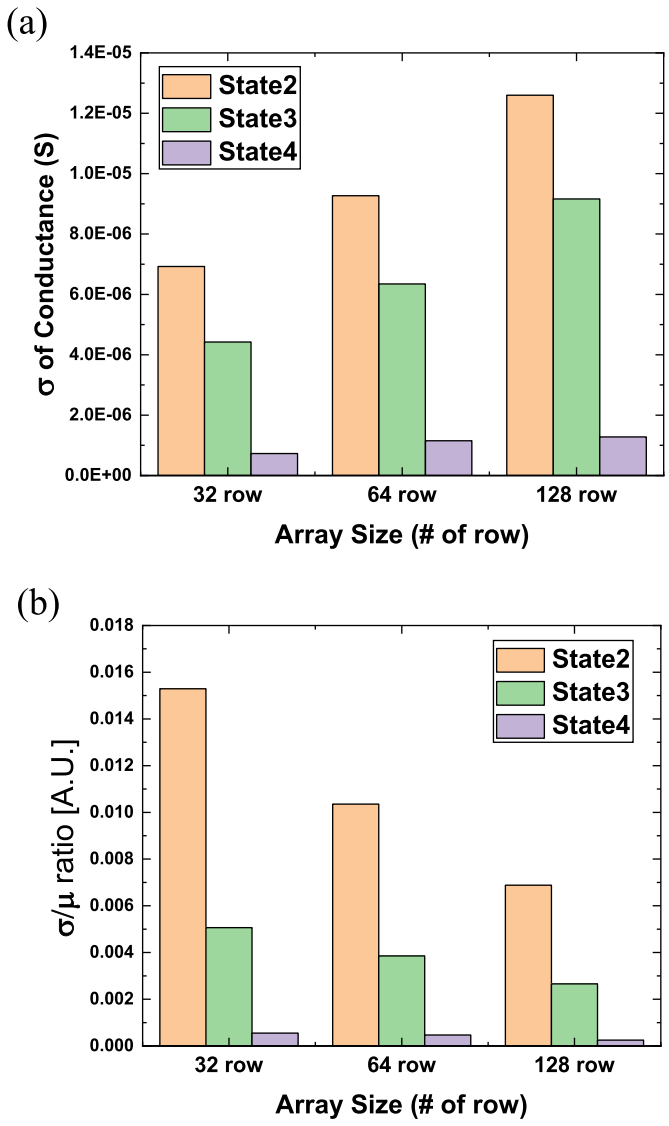


Figure 5. (a) Standard deviation of merged conductance and (b) standard deviation over average merged conductance for various synaptic array sizes. The read voltage applied to the BL is 0.2 V.

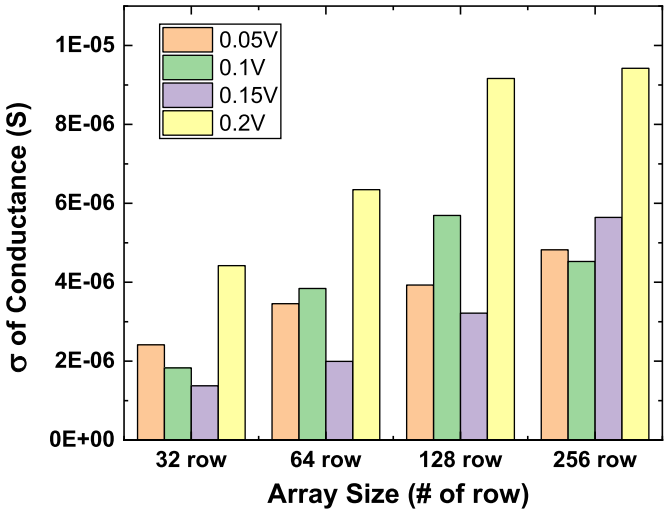
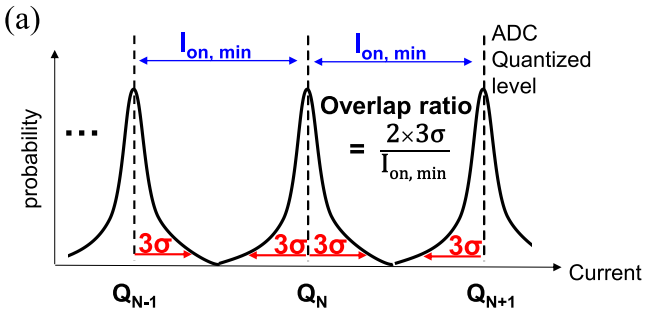


Figure 6. Standard deviation of merged conductance with various read voltages applied to the BL. The conductance target of the measured cells is state 3.



(b)

		Overlap ratio		
Level	Read voltage (V)	32 row (7-bit ADC)	64 row (8-bit ADC)	128 row (9-bit ADC)
state2	0.05	16.2%	24.8%	60.1%
	0.1	18.8%	48.4%	73.8%
	0.15	60.2%	96.2%	147.8%
	0.2	74.8%	100.1%	136.1%
state3	0.05	26.1%	37.3%	42.4%
	0.1	19.7%	41.5%	61.5%
	0.15	14.9%	21.5%	34.7%
	0.2	47.7%	68.5%	99.0%
state4	0.05	4.9%	7.7%	11.5%
	0.1	6.9%	10.8%	11.9%
	0.15	4.6%	8.2%	8.4%
	0.2	7.9%	12.5%	13.8%

Figure 7. (a) Conceptual schematic of overlap ratio with the ADC quantized level and read-noise-induced distribution. (b) Calculated overlap ratio with various read voltages, array sizes, and conductance levels.

that is clamped, where the aggregated read noise becomes a concern for the ADC resolution.

To quantify the impact of noise-induced variation on ADC accuracy, we defined the overlap ratio as a metric indicating the degree of ADC error as shown in figure 7. The 3 sigma of the probability distributions are calculated from the standard deviations (sigma), which are extracted from the measured data with various conductance levels, read voltages, and array sizes. The minimum space between the adjacent ADC quantized levels is same as one-step current (i.e. state 2's current) of a single RRAM cell in the case of maximum ADC resolution. We defined the overlap ratio by 2 times 3 sigma over minimum on cell current. With RRAM cells in the intermediate states, as the array size becomes larger, the overlap ratio of neighboring quantized levels increases and the error probability of digitized ADC output rises. Our results suggest that 9-bit ADC for full accumulation resolution of a 128 × 128 array with 2-bit per cell array is infeasible due to overlapping ratio of >100% in state 2.

4. Conclusions

In this paper, two-step write–verify for multilevel RRAM is proposed. Low BL voltage in the second step of write–verify enables more precise conductance control to achieve

tight cell distribution. The effect of read-noise-induced conductance variation on ADC accuracy becomes worse in larger arrays and higher BL voltages. However, the required ADC resolution could be lowered depending on the DNN model and dataset. For example, 5-bit ADC is found to be sufficient for 128 × 128 array with 2-bit/cell array for CIFAR-10 dataset [28], and therefore the impact of the noise is still acceptable.

Acknowledgments

We acknowledge the RRAM chip fabrication process provided by Winbond Electronics, Taiwan. This work is supported by NSF-CCF-1903951, ASCENT/C-BRIC, two of the SRC/DARPA JUMP Centers, and the NSF/SRC E2CDA program.

ORCID iD

Wonbo Shim https://orcid.org/0000-0002-9669-7310

References

[1] Jouppi N P *et al* 2017 In-datacenter performance analysis of a tensor processing unit *ACM/IEEE Int. Symp. on Computer Architecture (ISCA)* pp 1–12
[2] Yu S 2018 Neuro-inspired computing with emerging non-volatile memory *Proc. IEEE* **106** 260–85
[3] Prezioso M, Merrikh-Bayat F, Hoskins B D, Adam G C, Likharev K K and Strukov D B 2015 Training and operation of an integrated neuromorphic network based on metal-oxide memristors *Nature* **521** 61–64
[4] Woo J, Moon K, Song J, Lee S, Kwak M, Park J and Hwang H 2016 Improved synaptic behavior under identical pulses using AlOx/HfO2 bilayer RRAM array for neuromorphic systems *IEEE Electron Device Lett.* **37** 994–7
[5] Cai F 2019 *et al* A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations *Nat. Electron.* **2** 290–9
[6] Wu W, Wu H, Gao B, Yao P, Zhang X, Peng X, Yu S and Qian H 2018 A methodology to improve linearity of analog RRAM for neuromorphic computing *IEEE Symp. on VLSI Technology* pp 103–4
[7] Gokmen T and Vlasov Y 2016 Acceleration of deep neural network training with resistive cross-point devices: design considerations *Front. Neurosci.* **10**
[8] Ambrogio S *et al* 2018 Equivalent-accuracy accelerated neural-network training using analogue memory *Nature* **558** 60–67
[9] Kim W *et al* 2019 Confined PCM-based analog synaptic devices offering low resistance-drift and 1000 programmable states for deep learning *IEEESymp. on VLSI Technology*, pp T66–67
[10] Guo X, Merrikh-Bayat F, Bavandpour M, Klachko M, Mahmoodi M R, Prezioso M, Likharev K K and Strukov D B 2017 Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology *IEEE Int. Electron Devices Meeting (IEDM) (San Francisco, CA)*, pp 6.5.1–6.5.4
[11] Wang P, Xu F, Wang B, Gao B, Wu H, Qian H and Yu S 2019 Three-dimensional NAND Flash for vector-matrix multiplication *IEEE Trans. VLSI Syst.* **27** 988–91
[12] Lue H-T, Hsu P-K, Wei M-L, Yeh T-H, Du P-Y, Chen W-C, Wang K-C and Lu C-Y 2019 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 Int. Electron Devices Meeting (IEDM) (San Francisco, CA)* pp 38.1.1–38.1.4
[13] Lee S-T, Kim H, Bae J-H, Yoo H, Choi N Y, Kwon D, Lim S, Park B-G and Lee J-H 2019 High-density and highly-reliable binary neural networks using NAND flash memory cells as synaptic devices *IEEE Int. Electron Devices Meeting (IEDM) (San Francisco, CA, USA)* pp 38.4.1–38.4.4
[14] Ni K, Grisafe B, Chakraborty W, Saha A K, Dutta S, Jerry M, Smith J A, Gupta S and Datta S 2018 In-memory computing primitive for sensor data fusion in 28 nm HKMG FeFET technology *2018 IEEE Int. Electron Devices Meeting (IEDM) (San Francisco, CA)* pp 16.1.1–16.1.4
[15] Jain P *et al* 2019 A 3.6Mb 10.1Mb/mm2 embedded non-volatile ReRAM macro in 22nm FinFET technology with adaptive forming/set/reset schemes yielding down to 0.5V with sensing time of 5ns at 0.7V *IEEE Int. Solid- State Circuits Conf. (ISSCC), (San Francisco, CA, USA)* pp 212–4
[16] Sheng X, Graves C E, Kumar S, Li X, Buchanan B, Zheng L, Lam S, Li C and Strachan J P 2019 Low conductance and multilevel CMOS integrated nanoscale oxide memristors *Adv. Electron. Mater.* **5** 1800876
[17] Li C *et al* 2018 Efficient and self-adaptive *in-situ* learning in multilayer memristor neural networks *Nat. Commun.* **9** 2385
[18] Marinella M J, Agarwal S, Hsia A, Richter I, Jacobs-Gedrim R B, Niroula J, Plimpton S J, Ipek E and James C D 2018 Multiscale co-design analysis of energy, latency, area, and accuracy of a ReRAM analog neural training accelerator *IEEE J. Emerging Sel. Top. Circuits Syst.* **8** 86–101
[19] Milo V, Zambelli C, Olivo P, Pérez E, Mahadevaiah M K, Ossorio O G, Wenger C and Ielmini D 2019 Multilevel HfO2-based RRAM devices for low-power neuromorphic networks *APL Mater.* **7** 081120
[20] Zhao M *et al* 2018 Characterizing endurance degradation of incremental switching in analog RRAM for neuromorphic systems *IEEE Int. Electron Devices Meeting (IEDM) (San Francisco, CA)* pp 20.2.1–20.2.4
[21] Chen J, Wu H, Gao B, Tang J, Hu X S and Qian H 2020 A parallel multibit programing scheme with high precision for RRAM-based neuromorphic systems *IEEE Trans. Electron Devices* **67** 2213–7
[22] Yin S, Seo J-S, Kim Y, Han W, Barnaby H, Yu S, Luo Y, He W, Sun X and Kim -J-J 2019 Monolithically integrated RRAM and CMOS based in-memory computing optimizations for efficient deep learning *IEEE Micro* **39** 54–63
[23] Ho C, Chang S-C, Huang C-Y, Chuang Y-C, Lim S-F, Hsieh M-H, Chang S-C and Liao -H-H 2017 Integrated HfO_2-RRAM to achieve highly reliable, greener, faster, cost-effective, and scaled devices *IEEE Int. Electron Devices Meeting (IEDM) (San Francisco, CA)* pp 2.6.1–2.6.4
[24] Ambrogio S, Balatti S, McCaffrey V, Wang D and Ielmini D 2015 Noise-induced resistance broadening in resistive switching memory–Part II: array statistics *IEEE Trans. Electron Devices* **62** 3812–9
[25] Ielmini D, Nardi F and Cagli C 2010 Resistance-dependent amplitude of random telegraph signal noise in resistive switching memories *Appl. Phys. Lett.* **96** 053503
[26] Ambrogio S, Balatti S, Cubeta A, Calderoni A, Ramaswamy N and Ielmini D 2014 Statistical fluctuations in HfOx resistive-switching memory: part II–random telegraph noise *IEEE Trans. Electron Devices* **61** 2920–7
[27] Shim W, Luo Y, Seo J and Yu S 2020 Investigation of read disturb and bipolar read scheme on multilevel RRAM-based

deep learning inference engine *IEEE Trans. Electron Devices* **67** 2318–23

[28] Peng X, Huang S, Luo Y, Sun X and Yu S2019DNN+NeuroSim: an end-to-end benchmarking framework for compute-in-memory accelerators with versatile device technologies*IEEE Int. Electron Devices Meeting (IEDM)(San Francisco, USA)*: pp 32.5.1–32.5.4