Characterization and Mitigation of Relaxation Effects on Multi-level RRAM based In-Memory Computing

Wangxin He[1], Wonbo Shim[2], Shihui Yin[1], Xiaoyu Sun[2], Deliang Fan[1], Shimeng Yu[2], and Jae-sun Seo[1]
[1]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ, USA
[2]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA
Email: jaesun.seo@asu.edu

2021 IEEE International Reliability Physics Symposium (IRPS) | 978-1-7281-6893-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/IRPS46558.2021.9405228

Abstract— **In this paper, we investigate the relaxation effects on multi-level resistive random access memory (RRAM) based in-memory computing (IMC) for deep neural network (DNN) inference. We characterized 2-bit-per-cell RRAM IMC prototypes and measured the relaxation effects over 100 hours on multiple 8 kb test chips, where the relaxation is found to be most severe in the two intermediate states. We incorporated the experimental data into SPICE simulation and software DNN inference, showing DNN accuracy for CIFAR-10 dataset could degrade from 87.35% to 11.58% after 144 hours. To recover the largely degraded accuracy, mitigation schemes are proposed: 1) at the circuit level, the reference voltage for RRAM IMC could be calibrated after 80 hours when the relaxation is saturated. 2) At the algorithm level, the weights are trained with lower percentages to be quantized to the two intermediate states. With both schemes applied, the accuracy could be recovered to 87.32% for long-term stability.**

Index Terms—**RRAM, in-memory computing, multi-level cell, relaxation effect, deep neural network**

I. Introduction

DNNs have been successful in many computer vision and speech recognition applications. While state-of-the-art DNN algorithms continue to achieve higher accuracy with less number of parameters, the most compact models still require >3 million weights to achieve >70% top-1 ImageNet accuracy [1]. This leads to an insatiable demand for high-density memories such as multi-level RRAM. On the other hand, the DNN computations are dominated by multiply-and-accumulate (MAC) operations, but the overall energy consumption of DNN inference hardware has been dominated by memory access and data communication [2], due to the separation of conventional memories with row-by-row access and dedicated MAC engines. To improve the energy-efficiency of DNN inference, in-memory computing (IMC) has emerged as a promising technique, which turns on multiple rows and performs analog MAC computations along the bitline inside the memory.

Recent array-level demonstrations have presented RRAM's potential towards IMC for area-/energy-efficient DNN inference [3-7], but most RRAM based IMC prototypes today feature only single-level cell design [3-5]. Device-level programming of multi-level RRAM has been reported but was limited to row-by-row read-out [6]. Only a few works reported IMC with four-level RRAM devices [7-8], while [7] only demonstrated a simple two-layer multi-layer perceptron for a low 94.4% accuracy for MNIST dataset. More importantly, *most of the prior prototype designs just reported the basic functionality of IMC, while the reliability aspect of RRAM at array-level and during actual IMC operations is largely unexplored, although it can considerably affect the DNN inference accuracy.*

Relaxation occurs as a rapid drift of conductance right after initial programming but gradually saturates in the long term. For HfO_2 RRAM, its relaxation effects at device-level were reported in [5, 9], and read disturb induced RRAM conductance drift behavior was investigated in [10]. In our prior work [8], to maintain relatively stable DNN inference accuracy with RRAM relaxation effects over time, the peripheral circuits (e.g. reference voltages to the sense amplifiers) needed to be recalibrated every once in a while.

In this work, we comprehensively characterized the relaxation behavior with and without IMC operations on array-level with multiple 8 kb test chips [8], and analyzed its impact on DNN inference accuracy over time. The experiments are based on relaxation measurements of 2-bit-per-cell HfO_2 RRAM cells over 1,047 hours (over one month) accumulatively collected from three test chips, which were designed for RRAM based IMC with CMOS peripheral circuits. We present two mitigation schemes to recover the degraded DNN accuracy due to the relaxation effects. First, at the circuit-level, we calibrate the reference voltages after the relaxation saturates. Second, at the algorithm-level, we present a relaxation-aware DNN training technique to maintain high accuracy over time without frequent peripheral circuit calibration.

II. RRAM IMC Bitcell and Chip Design

A. 2-bit-per-cell RRAM IMC

As shown in Fig. 1(a), we use two vertically-adjacent cells and differential wordlines (WLs) to represent one 2-bit weight [8]. Four conductance values with equidistant conductance levels from G_{HIGH} to G_{LOW} corresponds to +3, +1, -1 and -3 weight values. As shown in Fig. 1(b), the element-wise multiplication between binary activation (+1, -1) and four-level weights results in four different pull-down strengths governed by the effective conductance, corresponding to four MAC partial results of +3, +1, -1, and -3.

Our RRAM macro design exhibits a 128×64 array, and with the vertically differential cell structure, each column stores 64 distinct weights. With all cells in the same column conducting in parallel, the sum of multiplication or MAC computation of 64 inputs will be between -192 to +192. Each possible MAC

Authorized licensed use limited to: ASU Library. Downloaded on January 03,2022 at 21:17:48 UTC from IEEE Xplore. Restrictions apply.

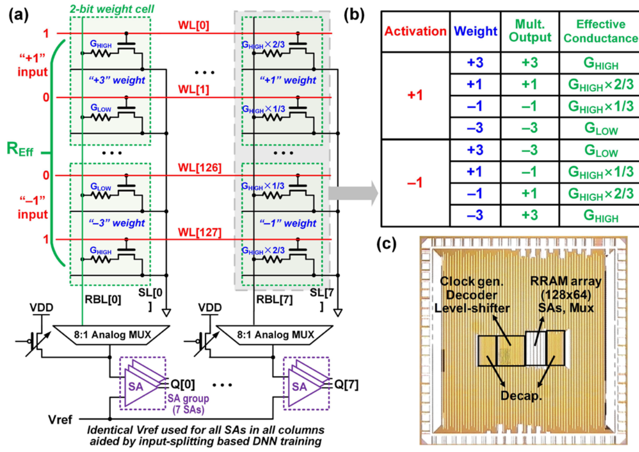


Fig. 1. (a) 2-bit-per-cell RRAM schematic; (b) Conductance representation of multiplication with 2-bit weights; (c) Chip micrograph. Adapted from [8].

value is represented by the RBL voltage (V_{RBL}), which is formed by the resistive divider between a controllable PMOS header and the parallel RRAM cells plus the analog multiplexer.

B. RRAM IMC Macro Periphery and Chip Design

V_{RBL} is compared with a reference voltage (V_{ref}) using voltage-mode sense amplifiers (SAs). One SA group consists of seven SAs, which can work in two different modes. With seven different V_{ref} voltages, the seven SAs can operate as a 3-bit (8-level) flash ADC [3]. Alternatively, we can use the seven SAs with identical V_{ref} to vote majority and obtain the binary output for the input-splitting algorithm [8]. For higher array-efficiency and density, each SA group is shared among every eight columns of RRAM array. The SAs convert the analog V_{RBL} voltage that represents the partial MAC results into digital values, which will be further analyzed for the DNN inference accuracy.

The RRAM macro can operate in two modes. First, the row decoder generates one-hot WL signals for cell-level RRAM programming and resistance read-out. Second, the row decoder asserts all differential wordline (WL) signals of the 128×64

Table I. Relaxation setup information for eight experiments.

Experiment #	Chip #	Total Hours	# of array-level IMC executed during total hour
A1	#1	94	0
A2	#2	94	0
B1	#3	108	7
B2	#1	144	10
B3	#1	156	9
B4	#2	144	9
B5	#1	154	49
B6	#1	153	12

simultaneously for IMC operation and performs the partial MAC computation.

The prototype chips (Fig. 1(c)) were fabricated in 90nm CMOS technology [11] that monolithically integrates 128×64 HfO_2 RRAM array (between M1 and M2) with SAs, column multiplexers, clock generator, row/column decoder, level-shifters, decoupling capacitors, etc.

III. Experiment Results

A. Relaxation and Experiment Setup

We measured the relaxation effects of four-level RRAM device/array over time for different operating conditions across three different chips. During RRAM programming, we tighten the conductance distribution using a write-verify programming protocol [8] so that the initial conductance is within 5% of the target state that ideally maps the four weight values.

Table I describes the eight different relaxation experiments that we conducted on three test chips (#1-#3) to monitor the effect of RRAM relaxation. For A1 and A2 experiments, we focus on the RRAM array resistance change without IMC for 94 hours. On the other hand, for B1, B2, B3, and B4 experiments, we performed array-level IMC in hardware a

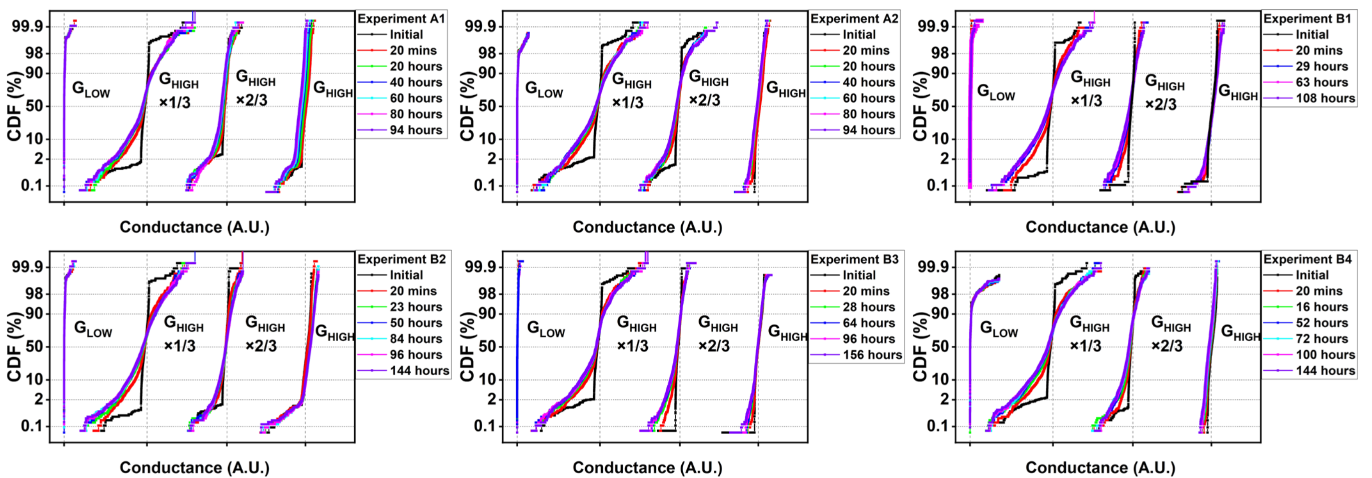


Fig. 2. Four-level RRAM relaxation effect over time for six experiments A1/A2 (without IMC) and B1/B2/B3/B4 (with IMC). For the six experiments, most relaxation occurs right after the initial programming, and subsequently saturates over time.

Authorized licensed use limited to: ASU Library. Downloaded on January 03,2022 at 21:17:48 UTC from IEEE Xplore. Restrictions apply.

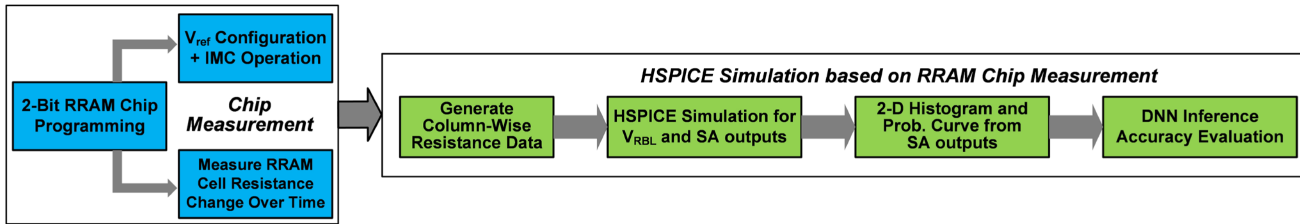


Fig. 3. RRAM chip measurement and simulation framework of this work.

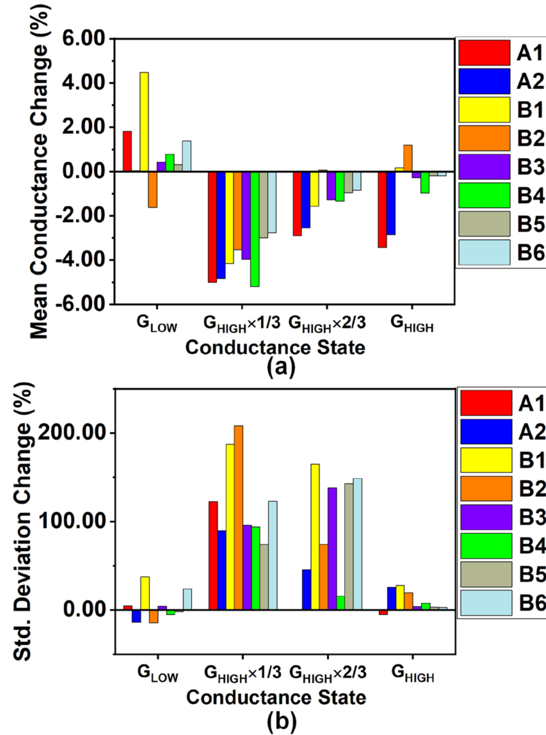


Fig. 4. (a) Mean of conductance and (b) standard deviation of conductance changes in 100 hours are shown for the eight experiments.

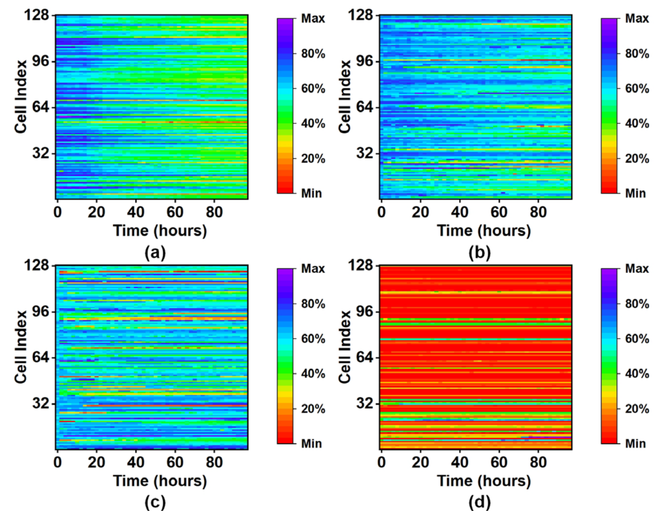


Fig. 5. Four-level conductance color-map. (a) G_{HIGH}; (b) G_{HIGH} ×2/3; (c) G_{HIGH} ×1/3; (d) G_{LOW}. For each level, 128 cells' conductance values are from A1 experiment. Warmer color represents a lower conductance value.

different number of times during the total experiment hours (up to 156 hours). During B5 and B6 experiments, more IMC operations are executed to verify the mitigation between the read-disturb induced RRAM drift effect [10] and the RRAM relaxation effect.

For six experiments from Table I, Fig. 2 shows the measured conductance results of the four-level RRAM devices starting from the programming time up to 156 hours. The conductance values for the four-level RRAMs shown in Fig. 2 are in the range of a few μS to a few hundreds of μS.

The overall workflow of the relaxation measurement, simulation and DNN accuracy evaluation is shown in Fig. 3. During the chip measurement process, the 2-bit RRAM chips are programmed with the subsets of the DNN with the input-splitting algorithm [8]. With reference voltage configuration, IMC operation, and relaxation effect, the RRAM array cell resistance changes are monitored over time. To better characterize the relaxation impact on effective resistance of the column (R_{eff}), V_{RBL} and DNN accuracy, we performed HSPICE simulation on the RRAM array column with the peripheral circuits, where we used the individual RRAM resistance measurements from the eight experiments in Table I. With the HSPICE simulation results, further data processing (2-D histogram and probability table generation) similar to hardware data processing is available, and we can generate DNN inference accuracy for RRAM array performance analysis, under different relaxation effect and operation stress condition.

While we can measure IMC results directly from the RRAM chips, we performed HSPICE simulation with RRAM device measurement results, to separate the relaxation effect over time with the IMC operation. In fact, as we discuss in Section III.D, read disturb drift effects by IMC operations can partially mitigate the relaxation effects of RRAM devices.

B. Relaxation Measurements and DNN Inference Accuracy

For the relaxation measurements, we monitored the cell-level resistance changes over time for the 128×64 RRAM array across eight experiments, and Fig. 4 shows the results. The two intermediate states of G_{HIGH}×1/3 and G_{HIGH}×2/3 experienced more decrease in average conductance and more increase in standard deviation over time (Fig. 4), indicating that they are less stable than G_{HIGH} and G_{LOW}.

As shown in the cumulative distribution function (CDF) of Fig. 2 and four-level conductance color map in Fig. 5, we

Authorized licensed use limited to: ASU Library. Downloaded on January 03,2022 at 21:17:48 UTC from IEEE Xplore. Restrictions apply.

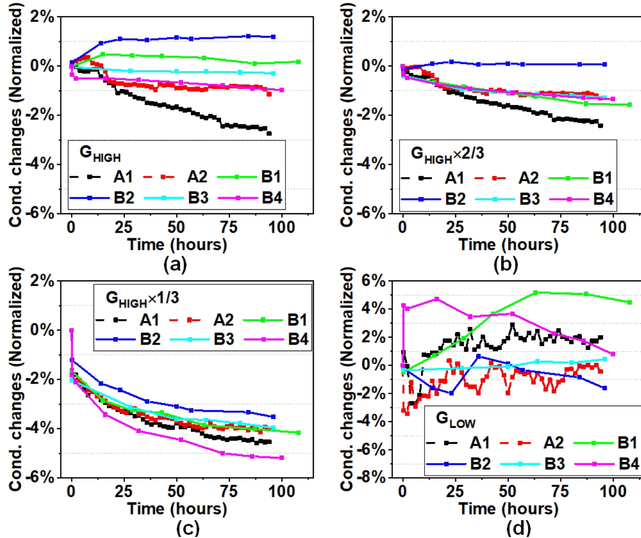


Fig. 6. Average conductance change of four-level RRAM devices for six experiments. G_{HIGH} ×1/3 is affected the most by relaxation. While G_{LOW} tends to fluctuate over time, they contribute negligible current for IMC.

observed a noticeable relaxation effect for six experiments (A1-A2, B1-B4) obtained from the measurement of three test chips. It can be seen that a large portion of relaxation occurs right after the programming (Fig. 2), and a similar behavior has also been reported in other prior HfO_2 RRAM works [12]. Low resistance states of the cells tend to reduce their conductance over time, which agrees with what was reported in [9], while high resistance states tend to fluctuate over time (Fig. 6) but they contribute negligible current to the bitline for MAC operation.

During the RRAM measurement based HSPICE simulation, to reflect the time-induced noise or variation of SAs, we added a random variation value for each V_{ref} in the SA group. One input vector applied on 128 rows will perform IMC with 2-bit weights in the 128×64 array, and the output from the SA group indicates the computed results for the partial MAC value within one column. For each partial MAC value, Fig. 7(a) and Fig. 7(b) show the 2-D histograms of IMC measurements from the RRAM chips at 0 hours and 144 hours, respectively. Fig. 7(c) and Fig. 7(d) represent the 2-D histograms of the HSPICE simulations based on individual RRAM device measurements at 0 hours and 144 hours, respectively, which show similar behavior as the IMC measurement results. The 2-D histogram

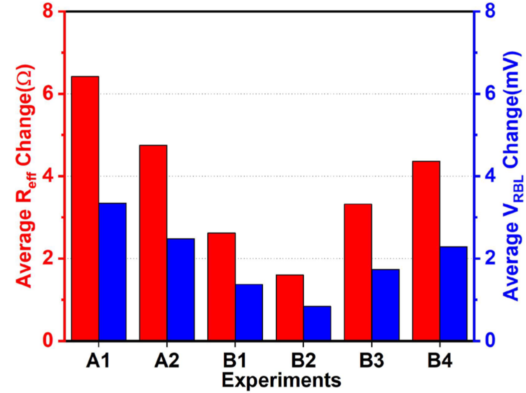


Fig. 8. R_{eff} and V_{RBL} correlation. V_{RBL} will increase with R_{eff} from the relaxation effect.

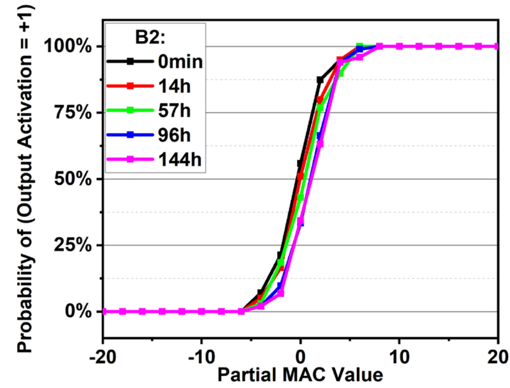


Fig. 9. Probability curve shifting is observed over time from B2 experiment.

is used to generate a probability curve (Fig. 9), which states the probability for each MAC value output to be quantized to "+1" for the binary output ("+1" or "-1") of the RRAM array, based on the input-splitting scheme [8].

Based on the RRAM conductance changes measured from six experiments, we compared the V_{RBL}, R_{eff}, and DNN accuracy change over time. First, we discovered a strong correlation in R_{eff} and V_{ref} changes over time. We simulated the average R_{eff} and V_{ref} changes between 0 and ~100 hours for six experiments (Fig. 8). A larger increase in R_{eff} corresponds to a larger V_{RBL} increase and leads to SA output difference over time. Second, we simulated the MAC operation starting from the time of 0

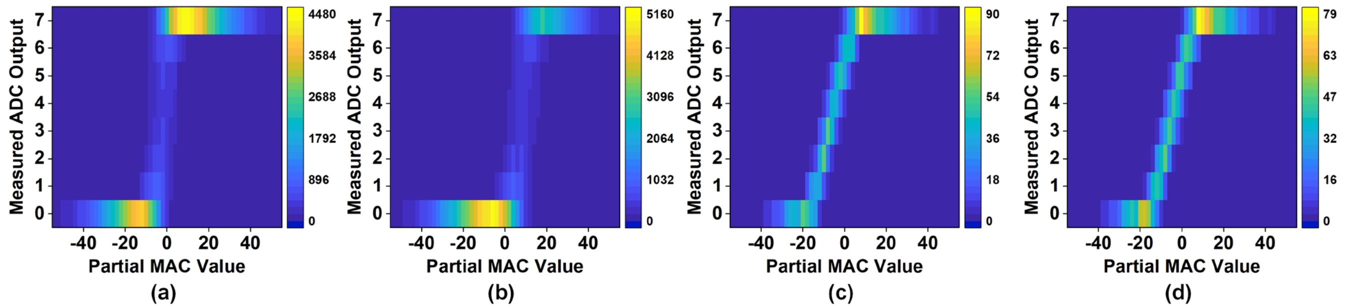


Fig. 7. (a)-(b) 2-D histogram from IMC measurements from RRAM chips at 0 and 144 hours for B2 experiment. (c)-(d) 2-D histogram from HSPICE simulation using individual RRAM device measurements from 0 to 144 hours for B2 experiment.

Authorized licensed use limited to: ASU Library. Downloaded on January 03,2022 at 21:17:48 UTC from IEEE Xplore. Restrictions apply.

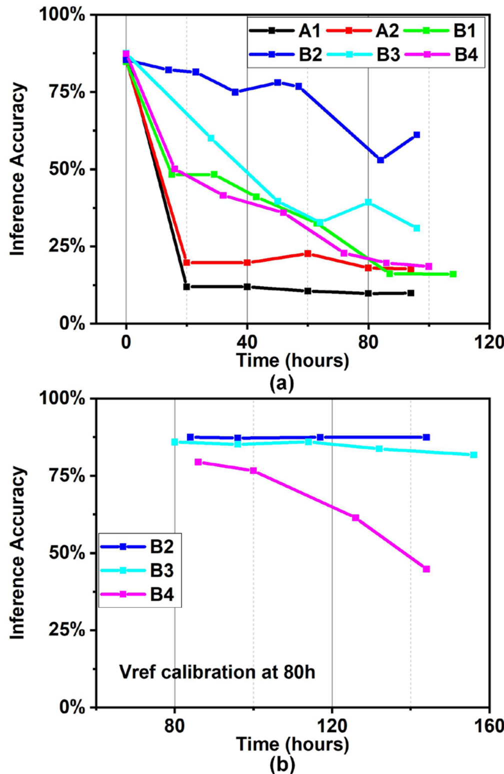


Fig. 10. (a) DNN accuracy drops from 0 to ~100 hours for 6 experiments. (b) DNN accuracy drops with V_{ref} calibrated at ~80 hours for B2/B3/B4.

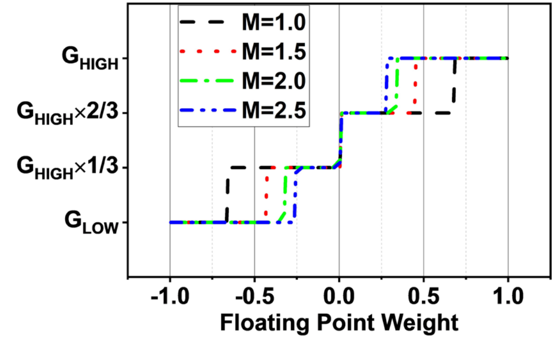
hours to ~100 hours and use the V_{ref} setting at 0 hours throughout the entire experiment duration. The results indicate that a worse relaxation will lead to a larger change in average R_{eff}, and correspondingly results in a larger V_{RBL} change. These V_{RBL} changes cause the difference in SA group output, 2-D histogram, and shifting behavior in probability curves (Fig. 9), and lead to a considerable DNN inference accuracy loss, e.g. from 87.35 % to 11.58% for B4 (Fig. 10(a)).

As we observe that a large portion of the RRAM relaxation effect occurs soon after the RRAM programming, one approach to avoid huge loss from it is to wait until most relaxation saturates, then calibrate the V_{ref} for SAs, and use the calibrated V_{ref} for the ensuing time. We selected ~80 hours as the time to perform V_{ref} re-calibration for two reasons. First, while large conductance changes are observed right after programming (in 20 minutes from Fig. 2), non-negligible relaxation further occurs from 20 minutes to ~80 hours, and the relaxation largely saturates after ~80 hours. Second, our aim is that the V_{ref} re-calibration method can mitigate the long-term relaxation effect without additional write operations (e.g. throughout hundreds of hours, while our experiments in this paper are up to ~150 hours due to the time limits). Compared to re-calibrating V_{ref} in 20 minutes or a few hours after programming, performing V_{ref} re-calibration in ~80 hours could enable a longer operational time for the RRAM chips.

Based on the B2/B3/B4 experiments, we characterized the MAC operation starting from the time of around 80 hours to 144/156 hours and used the V_{ref} setting obtained at time of around 80 hours through the remaining time of experiments (Fig. 10(b)). Compared to Fig. 10(a), a relatively smaller accuracy drop occurs over time in Fig. 10(b), due to the saturated RRAM relaxation behavior. This circuit-level technique improves the long-term stability for IMC, although B4 experiment needs further improvement.

Table II. Weight distribution change for DNN training with different magnification factor (M) for VGG-like CNN for CIFAR-10 dataset.

Mag. Factor for Training	G_{HIGH} (%)	$G_{HIGH}\times2/3$ (%)	$G_{HIGH}\times1/3$ (%)	G_{LOW} (%)	Baseline DNN Accuracy
1.0	24.603	25.578	25.482	24.338	87.60%
1.5	32.948	17.157	17.128	32.767	87.39%
2.0	36.766	13.189	13.198	36.847	87.24%
2.5	39.389	10.755	10.716	39.141	87.60%



W=3*[(round(clip(weight_array*M, -1, +1)*1.5+1.5)-1.5)/1.5]

Fig. 11. DNN training with higher magnification factor (M) pushes more weights to +3 (G_{HIGH}) and -3 (G_{LOW}).

C. Relaxation-aware DNN Training and Improvement

As seen in Fig. 3, among the four levels of 2-bit-per-cell RRAM, two intermediate states suffer more relaxation, due to the weak filament in RRAM cell. If the DNN is aware of this drawback on the two intermediate states and reduces the overall percentage of these states during the training procedure, the inference accuracy should have less impact from the relaxation effect. Therefore, we introduce a magnification factor (*M*) during the training of 2-bit-weight VGG-like DNN for CIFAR-10 dataset (Fig. 11). By increasing the *M* during training, it pushes more weights within each layer to the highest and lowest resistance stages, with almost no initial accuracy change compared to the baseline DNN (Table II).

Subsequently, using the measured RRAM conductance values from B2-B4 experiments, we evaluated the DNN accuracy over time for different *M* values from 1.0 to 2.5 (Table II), using V_{ref} settings from the time at 0 hours and around 80 hours (Fig. 12). Compared with the results in Fig. 10(a), the re-trained DNNs with higher *M* achieves much higher DNN accuracy over time for both V_{ref} settings. By combining both schemes, DNN accuracy of >87.2% for CIFAR-10 is achieved for B4 over 144 hours with a single V_{ref} re-calibration at 86 hours. This indicates that the relaxation-aware training scheme largely alleviated the impact of RRAM relaxation.

Authorized licensed use limited to: ASU Library. Downloaded on January 03,2022 at 21:17:48 UTC from IEEE Xplore. Restrictions apply.

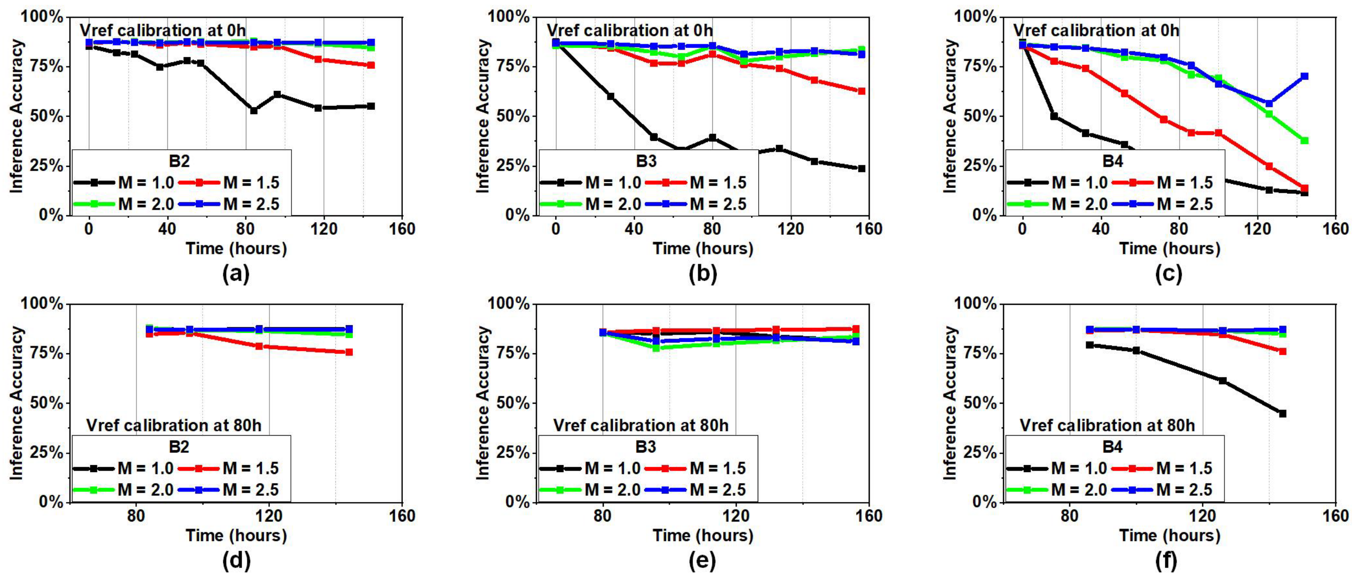


Fig. 12. (a)-(c) For experiment B2/B3/B4, we tested the DNN inference accuracy from 0 to 144/156 hours, for RRAM arrays with weight distribution using different magnification factor (M). Overall, higher percentage of "+3/-3" weights offers improved robustness against relaxation. (d)-(f) DNN accuracy trends with V_{ref} calibrated at ~80 hours for B2/B3/B4 experiments.

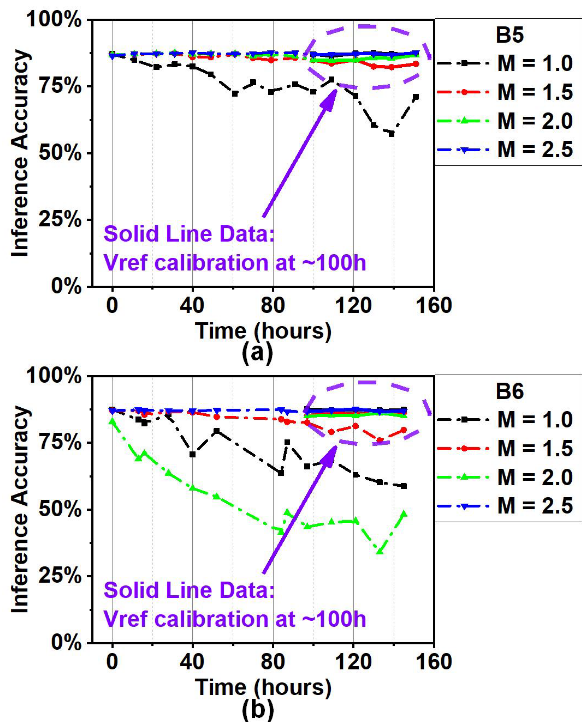


Fig. 13. Simulation results of B5/B6 experiments under high frequency IMC operation stress. B5 and B6 show better accuracy retention after 150 hours.

D. RRAM Read Disturb on Relaxation Effect Mitigation

As discussed in [10], when the read voltage is higher than a certain level, read disturb will occur as conductance drift for both LRS and HRS states. High RBL voltage will tend to increase the conductance of LRS, and reduce that of HRS, which is the opposite trend of RRAM relaxation effect. In Fig. 2(a), compared to the no IMC operation experiments A1/A2, we observed that B1-B4 experiments have less mean conductance change. When IMC operation is applied on the chip, the voltage between the RRAM cells increases from normal read voltage (0.2V) to a relatively high value (0.4-0.6V), thus the read disturb induced conductance drift takes place. This drift mitigates the relaxation effect on LRS, and results in fewer conductance changes over time.

To verify this hypothesis, B5 and B6 experiments are conducted with a greater number of IMC operations and measured for inference accuracy performance comparison. B5 executes IMC operation every three hours, and B6 executes IMC operation every twelve hours, both having a higher IMC operation frequency than B1-B4 experiments. Overall, B5 and B6 experiments have less conductance reduction than B1-B4 at three LRS states (Fig. 3).

Then, similar relaxation simulations are reproduced with HSPICE simulations and both circuit-level and algorithm-level optimization methods are applied for B5 and B6 experiments. Fig. 13 shows the DNN inference accuracy changes over time under two mitigation methods. Compared to B2-B4 results in Fig. 12, the accuracy degradation without any optimization scheme (M=1.0 case) is lower after 150 hours (87.18% to 71.12% for B5, 87.49% to 58.81% for B6). Regarding the two proposed mitigation schemes together with the purposely induced read-disturb, both methods result in similar or better accuracy retention over time compared with B2-B4 experiments. For B5 and B6 experiments, applying a single scheme is sufficient for DNN accuracy retention against the relaxation effect. V_{ref} recalibration alone (in solid lines in Fig. 13) can recover the accuracy of original DNN cases (M=1.0) to be above 83% after 150 hours. Similarly, relaxation-aware DNN training scheme alone serves similar improvement results as B2-B4 experiments. M=2.5 case in B5 and B6 (blue dash lines) can maintain the accuracy above 83% over 150 hours, without the help of V_{ref} calibration method.

The results above indicate that the read disturb drift effects can also partially cancel out and mitigate the non-ideality from relaxation effect, overall improving the DNN inference accuracy over time.

Authorized licensed use limited to: ASU Library. Downloaded on January 03,2022 at 21:17:48 UTC from IEEE Xplore. Restrictions apply.

IV. Conclusion

In this paper, we comprehensively characterized the relaxation effects of multi-level HfO_2 RRAM at array-level for in-memory computing hardware targeting DNN inference applications. Relaxation effects are noticeable at intermediate states of multi-level RRAM, but can be compensated using circuit-level (e.g. V_{ref} calibration after relaxation saturation) and algorithm-level (e.g. relaxation-aware DNN training for weight re-distribution) techniques that we proposed and demonstrated. Also, the non-ideality from the read disturb induced drift effect can be utilized to mitigate the relaxation effect and could potentially enhance the DNN inference accuracy retention over time.

Acknowledgment

The authors thank Winbond Electronics for RRAM chip fabrication support. This work is partially supported by NSF grants 1652866/1715443/1740225, JUMP CBRIC/ASCENT (SRC program sponsored by DARPA), NSF/SRC E2CDA, and SRC AIHW program.

References

[1] L. Deng, G. Li, S. Han, L. Shi and Y. Xie, "Model compression and hardware acceleration for neural networks: a comprehensive survey," *Proceedings of the IEEE*, vol. 108, no. 4, pp. 485-532, April 2020, doi: 10.1109/JPROC.2020.2976475.

[2] B. Zimmer *et al.*, "A 0.32–128 TOPS, scalable multi-chip-module-based deep neural network inference accelerator with ground-referenced signaling in 16 nm," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 4, pp. 920-932, April 2020, doi: 10.1109/JSSC.2019.2960488.

[3] S. Yin, X. Sun, S. Yu and J. Seo, "High-throughput in-memory computing for binary deep neural networks with monolithically integrated RRAM and 90-nm CMOS," *IEEE Transactions on Electron Devices*, vol. 67, no. 10, pp. 4185-4192, Oct. 2020, doi: 10.1109/TED.2020.3015178.

[4] C. Xue *et al.*, "15.4 A 22nm 2Mb ReRAM compute-in-memory macro with 121-28TOPS/W for multibit MAC computing for tiny AI edge devices," *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, San Francisco, CA, USA, 2020, pp. 244-246, doi: 10.1109/ISSCC19947.2020.9063078

[5] W. Wan *et al.*, "33.1 A 74 TMACS/W CMOS-RRAM neurosynaptic core with dynamically reconfigurable dataflow and in-situ transposable weights for probabilistic graphical models," *IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2020, pp. 498-500, doi: 10.1109/ISSCC19947.2020.9062979.

[6] E. R. Hsieh *et al.*, "High-density multiple bits-per-Cell 1T4R RRAM array with gradual SET/RESET and its effectiveness for deep learning," *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2019, pp. 35.6.1-35.6.4, doi: 10.1109/IEDM19573.2019.8993514.

[7] Q. Liu *et al.*, "33.2 A fully integrated analog ReRAM based 78.4TOPS/W compute-in-memory chip with fully parallel MAC computing," *IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2020, pp. 500-502, doi: 10.1109/ISSCC19947.2020.9062953

[8] W. He *et al.*, "2-bit-per-cell RRAM-based in-memory computing for area-/energy-efficient deep learning," *IEEE Solid-State Circuits Letters*, vol. 3, pp. 194-197, 2020, doi: 10.1109/LSSC.2020.3010795

[9] C. Wang *et al.*, "Relaxation effect in RRAM arrays: demonstration and characteristics," *IEEE Electron Device Letters*, vol. 37, no. 2, pp. 182-185, Feb. 2016, doi: 10.1109/LED.2015.2508034.

[10] 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.

[11] C. Ho *et al.*, "Integrated HfO_2-RRAM to achieve highly reliable, greener, faster, cost-effective, and scaled devices," *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2017, pp. 2.6.1-2.6.4, doi: 10.1109/IEDM.2017.8268314.

[12] R. Degraeve *et al.*, "Quantitative model for post-program instabilities in filamentary RRAM," *IEEE International Reliability Physics Symposium (IRPS)*, Pasadena, CA, 2016, pp. 6C-1-1-6C-1-7, doi: 10.1109/IRPS.2016.7574567.

Authorized licensed use limited to: ASU Library. Downloaded on January 03,2022 at 21:17:48 UTC from IEEE Xplore. Restrictions apply.