Adaptive Oxygen Vacancy Diffusion Compensation in MLC Intermediate States for over 10-year Data-retention of TaO_X ReRAM Analog CiM Array

Yusuke Hirata, Kenshin Yamauchi, Naoko Misawa, Chihiro Matsui, and Ken Takeuchi

Dept. of Electrical Eng. and Information Systems, The University of Tokyo, Japan, yusuke.hirata@co-design.t.u-tokyo.ac.jp

Abstract—**This paper proposes compensation methods for data-retention degradation due to oxygen vacancy diffusion in ReRAM, especially for intermediate states. The proposed techniques enable high inference accuracy to be maintained in MLC TaO_X-based analog ReRAM CiM systems, even after 10 years of array-level data-retention. Measurement of 10-year data-retention across multi-level cells is also demonstrated. During data-retention, ReRAM cell current shift caused by oxygen vacancy diffusion in ReRAM devices degrades the MAC values of neural networks. The newly introduced compensation factor α is determined by the mean cell current of the replica cell circuit and adaptively compensates for the MAC value shift in both past and future directions of data-retention time and write variations. In addition, by using DR tolerant system of ELU activation function and Retention Aware Training, the proposed ReRAM CiM systems suppresses the inference accuracy drop by 1.5% over 10-year DR at 85degC.**

Keywords—ReRAM, CiM, data-retention, oxygen vacancy

I. Introduction

In Multi-level Cell (MLC) ReRAM analog Computation-in-Memory (CiM), data-retention (DR) errors of ReRAM reduce the inference accuracy of neural networks (NNs) [1-3]. NN weights stored in ReRAM cells change during DR. It has been reported that changes in multiply-accumulate value (MACV) caused by PRAM drift, which occurs at high temperatures [4], can be corrected by changing the reference voltage [5, 6]. On the other hand, this paper discusses advanced MACV compensation method for ReRAM, not only because of current shift, but also because of retention characteristic difference among the resistance states and large current variation in the intermediate states. This paper proposes a retention-tolerant MLC ReRAM CiM System Technology Co-Optimization (STCO) that adaptively compensates for MACV degradation caused by DR shift as well as write variations, and maintains high inference accuracy over 10-years DR at the chip-level (Fig. 1(a)). In *Prop. 1 (Tech)*, the distorted MACV after DR is corrected by multiplying a newly introduced compensation factor *α*. *Prop. 2 (System)* adopts the Exponential Linear Unit (ELU) function as activations in NN models, which is robust to DR errors. The CiM output is the MACV, which is then input to the ELU activation function. *Prop. 3 (Tech&Cir)* predicts the compensation factor *α* by regression of the mean cell current of replica ReRAM cells. This allows for the adaptive compensation of the MACV in both past and future (bidirectional) directions of the DR time. In *Prop. 4 (System&Tech)*, during NN training, Retention Aware Training (RAT) injects the measured ReRAM DR errors and write variations into NN weights to improve the inference accuracy even after DR (Fig. 1(b)). In the proposed STCO (Fig. 1(c)), after RAT, trained NN weights are written to the ReRAM cells. During inference, MACV is compensated by the factor *α*. When writing to the ReRAM cells, NN weights are quantized to 5 bits and clipped (Fig. 1(d)). In the proposed MLC & SLC Hybrid ReRAM cells, MSB and LSB are written to SLC and MLC cells respectively (Fig. 1(e)).

II. Measured 10-year DR of 4M Analog ReRAM Chip

By using a 4Mbit test chip [7, 8] (Figs. 2(a)(b)), DR of 4Kbits of 40nm TaO_X-based analog ReRAM is measured. Compared to previous works [2, 9-13], this paper presents multi-level and long period measurement. Considering the activation energy of oxygen vacancy (V_O) diffusion, 1.2eV [14], to evaluate 10-year DR, the ReRAM chip is measured at accelerated 150 and 190degC. Figs. 2(c)(d) show the

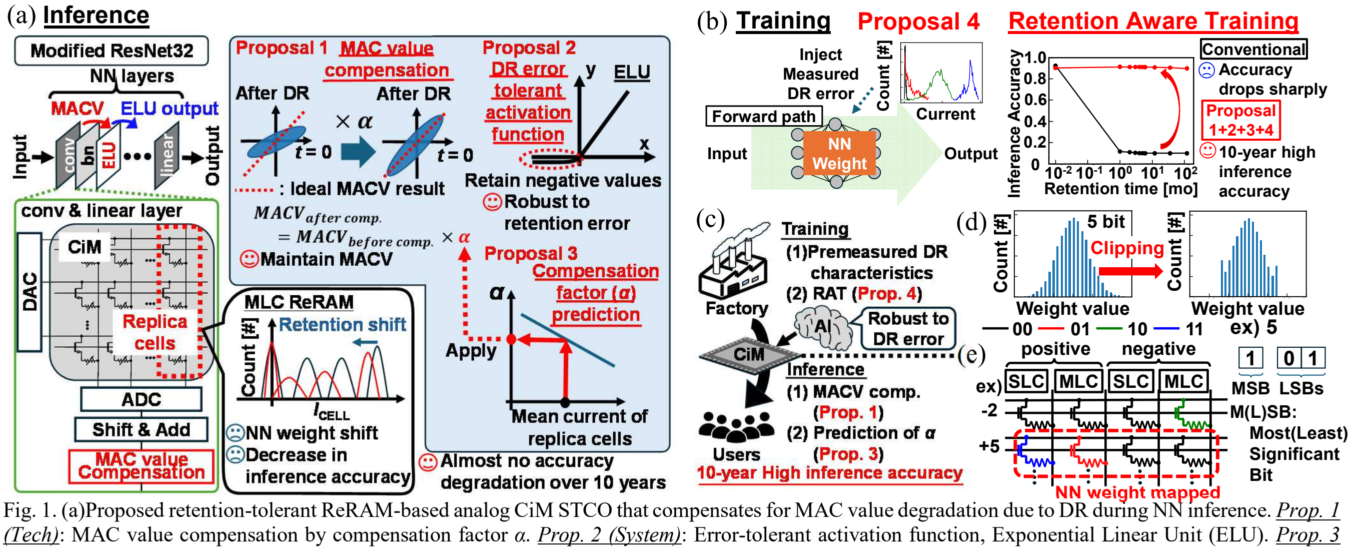


Fig. 1. (a)Proposed retention-tolerant ReRAM-based analog CiM STCO that compensates for MAC value degradation due to DR during NN inference. *Prop. 1 (Tech)*: MAC value compensation by compensation factor *α*. *Prop. 2 (System)*: Error-tolerant activation function, Exponential Linear Unit (ELU). *Prop. 3 (Tech&Cir)*: Compensation factor *α* prediction by replica cell circuit. (b) *Prop. 4 (System&Tech)*: Retention Aware Training (RAT) during NN training. (c) Overall architecture of proposed retention-tolerant ReRAM-based analog CiM STCO. (d) NN Weight expression with quantization and clipping [15]. (e) SLC & MLC Hybrid ReRAM CiM mapping.

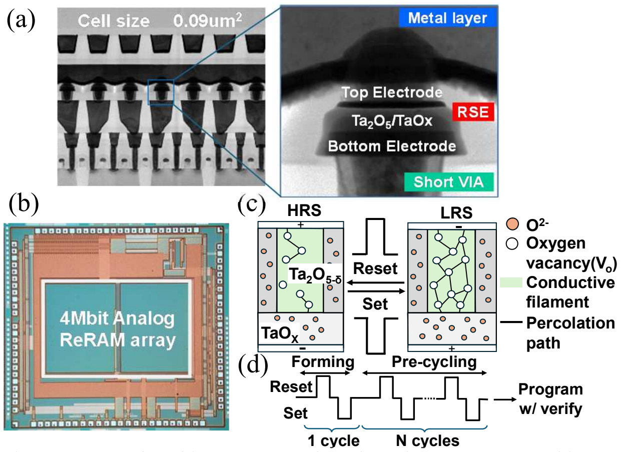


Fig. 2. Measured 4Mbit 40nm TaO$_X$-based Analog ReRAM test chip. (a) TEM cross-section image of 40nm ReRAM cell. (b) Chip die photograph. (c) Switching mechanism of ReRAM [16]. (d) Cell writing protocol with forming, pre-cycling, and verify.

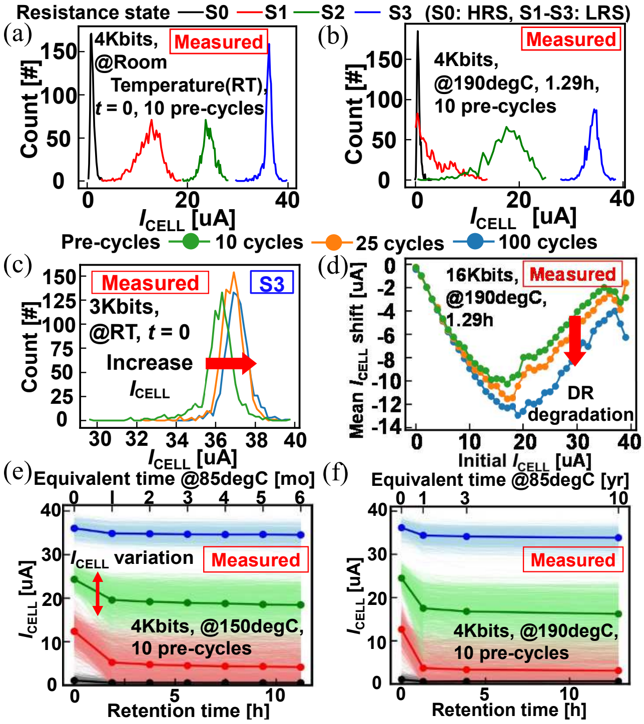


Fig. 3. Measured cell current (I_{CELL}) of DR. # of cell count of (a) DR time t = 0 and (b) after DR of S0-S3 states. Pre-cycling trade-off; measured I_{CELL} shift by number of pre-cycles (c) at t = 0 and (d) after DR (16Kbits=1Kbits×16states). Measured I_{CELL} shift by DR at (e) 150degC and (f) 190degC with average I_{CELL} (thick line) and I_{CELL} variations (thin lines) of 4Kbits.

switching mechanism and writing protocol of ReRAM cells. After conductive filament (CF) formation and pre-cycling, cells are written to S0-S3. S0 is the reset-state (HRS) and S1-S3 are set-states (LRS).

The measured ReRAM cell current (I_{CELL}) distribution is shown at DR time t = 0 (just after write) at RT (Fig. 3(a)) and after 1.29-hour DR at 190degC (Fig. 3(b)). I_{CELL} decreases after DR. It is observed that chip-level DR degrades S1 and S2 more significantly than S3. Also, for the first time, this paper observes measured pre-cycling trade-off. The more precycles increase I_{CELL} at t = 0 (larger I_{CELL} range in Fig. 3(c)), but accelerates DR degradation (Fig. 3(d)). After DR at 150 and 190degC in Figs. 3(e)(f), the average I_{CELL} (thick line) decreases, and the 4Kbits I_{CELL} variations (thin lines) increase for S0-S3. Additionally, in intermediate S1 and S2, I_{CELL} decreases significantly with DR. 6-month DR at 85degC is equivalent to 11.2 hour at 150degC, and 10-year DR at 85degC is equivalent to 12.9 hour at 190degC, as derived from the Arrhenius equation. 10-year DR shifts NN weight values and MACV from their ideal values (Figs. 4(a)(b)).

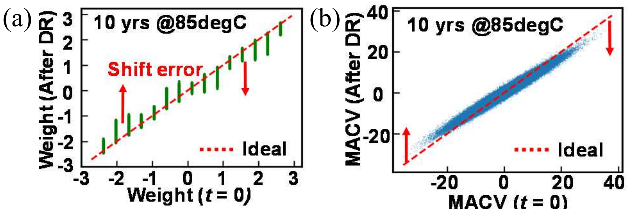


Fig. 4. (a) Weight value shift and (b) simulated MAC value shift from ideal value after 10-year DR.

Fig. 5(a) shows a ReRAM physical model for DR and pre-cycling. During DR, $\mathrm{V_O}$ diffuses and $\mathrm{V_O}$ density in conductive filament (CF) decreases. Also, more pre-cycles expands the CF diameter and decreases $\mathrm{V_O}$ density in CF. Thus, at large DR and high pre-cycling numbers, $\mathrm{V_O}$ in CF becomes sparse and conduction paths in CF are more likely to be cut off. As a result, I_{CELL} decreases. Based on the physical model [17], the cell current of low-resistance states (I_{LR}) after DR time t is theoretically formulated.

$$I_{\mathrm{LR}} = \frac{A}{\sqrt{t}}\left(1-\frac{B}{t}\right)+C \tag{1}$$

Parameter A is proportional to the square of jumping distance of electrons inside CF, d (Fig. 5(b)). By fitting A, B, and C with the least-squares method, the physical model-based I_{LR} of S1-S3 states fits well with the measured I_{CELL} after DR (Fig. 5(c)). Fig. 5(d) shows the parameter A and d for each state. Fig. 5(e) illustrates the physical model for S1-S3 states. In the intermediate S1 and S2 states, larger d, i.e., larger A, results in sparser $\mathrm{V_O}$ inside the CF, leading to a decrease in I_{CELL} during DR.

III. Proposed $\mathrm{V_O}$ Diffusion Compensation for Inference

In the following experiments, using quantized and clipped weights stored in SLC & MLC cells (Fig. 1(e)), inference accuracy with weight value degradation is estimated based on the measured DR data (Figs. 3(e)(f)). During training, QAVAT [18] ensures robustness to ReRAM write variations.

A. MAC Value Compensation

Fig. 6(a) shows the proposed MAC value compensation. This compensation is given by $MACV_{\mathrm{after\_comp}}$ =

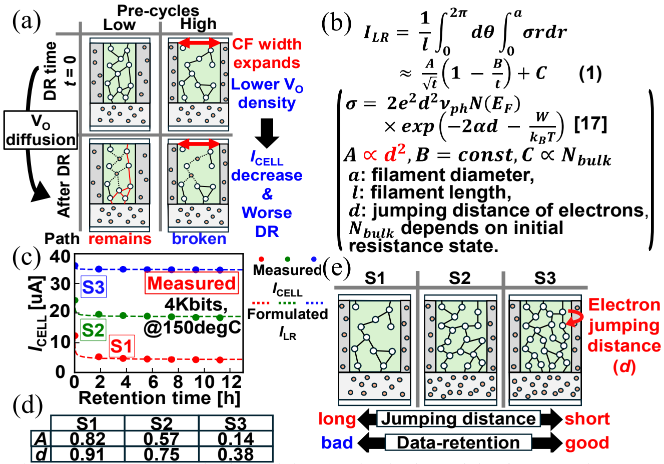

(a)
Pre-cycles
Low
High
DR time t = 0
CF width expands
Lower V$_O$ density
V$_O$ diffusion
After DR
I_{CELL} decrease & Worse DR
Path remains
broken
(b)
$I_{LR} = \frac{1}{l}\int_0^{2\pi} d\theta \int_0^{a} \sigma r dr$
$\approx \frac{A}{\sqrt{t}}\left(1-\frac{B}{t}\right)+C$ (1)
$\sigma = 2e^2 d^2 v_{ph} N(E_F) \times exp\left(-2\alpha d - \frac{W}{k_B T}\right)$ [17]
$A \propto d^2, B = const, C \propto N_{bulk}$
a: filament diameter,
l: filament length,
d: jumping distance of electrons,
N_{bulk} depends on initial resistance state.
(c)
I_{CELL} [uA]
40 30 20 10 0
S3
S2
S1
Measured
4Kbits, @150degC
0 2 4 6 8 10 12
Retention time [h]
Measured I_{CELL}
Formulated I_{LR}
(d)
S1 S2 S3
A 0.82 0.57 0.14
d 0.91 0.75 0.38
(e)
S1
S2
S3
Electron jumping distance (d)
long
Jumping distance
short
bad
Data-retention
good

Fig. 5. ReRAM physical model. (a) Physical model of DR and pre-cycling. (b) Formulated I_{LR} of S1-S3 with parameters A, B, and C. (c) Plots of formulated I_{LR} of S1-S3 (d) Comparison of parameters A and d for S1-S3. (e) Physical model of S1-S3.

$\alpha \times MACV_{\text{before\_comp}}$, where α is the newly introduced compensation factor. Based on the reference time t_{ref}, Backward compensation (comp.) corrects MACV after DR based on the earlier time ($t_{\text{ref}} < t_{\text{cur}}$), and thus $\alpha > 1$. Forward comp. ($t_{\text{ref}} > t_{\text{cur}}$, $\alpha < 1$) adjusts MACV in the opposite time direction of Backward comp.. Bidirectional comp. adjusts the MAC of both earlier and later times based on the MACV of the intermediate t_{ref} time. In Fig. 6(b), t_{ref} = 1-month DR is optimal to correct MACV of both t_{cur} = 0 and 10-year DR because of the intermediate I_{CELL} drop at 1-month DR and subsequent gradual shifts (Figs. 3(e)(f)).

B. *Retention-error Tolerant Activation Function*

Retention-error tolerant ELU activation function replaces all ReLU activations in ResNet32. ReLU is weak against DR errors because ReLU outputs are zero for negative inputs. DR error can flip the MACV between positive and negative values potentially causing the ReLU output to vanish (Fig. 7(a)). Due to the non-zero nature of the ELU activation outputs for negative inputs, ELU is robust against DR errors and MACV shifts (Fig. 7(b)). As a result, proposed ELU successfully compensates MACV of any layers in ResNet32 (Fig. 7(c)). The introduced overhead due to the ELU function calculating non-zero outputs for negative inputs is negligibly small compared to the gains of massively-parallel MAC.

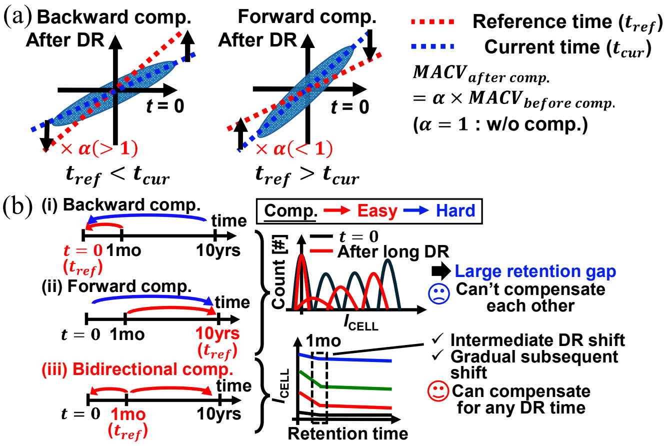


Fig. 6. *Prop. 1 (Tech)*: MAC value compensation by compensation factor α. (a) Backward compensation (comp.) and Forward comp with respect to reference time t_{ref}. (b) Three methods of MAC compensation: (i) Backward with t_{ref} = 0, (ii) Forward with t_{ref} = 10 years, and (iii) Bidirectional comp. with t_{ref} = 1 month, for example.

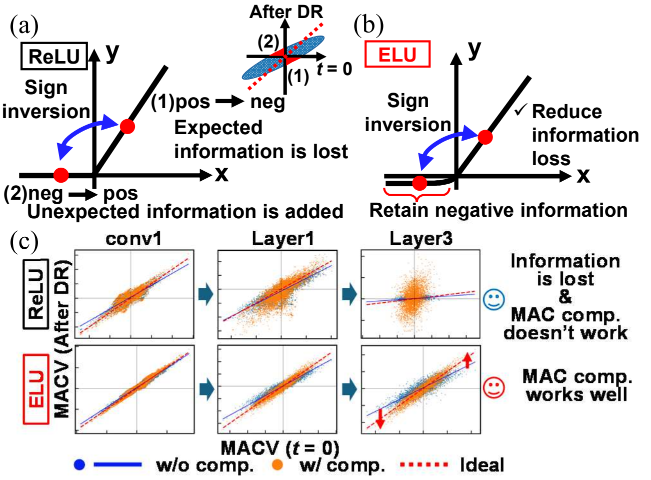


Fig. 7. *Prop. 2 (System&Tech)*: Error-tolerant activation function, ELU. (a) ReLU's vulnerability and (b) ELU's robustness to DR error. (c) Comparison of compensation effectiveness by ReLU and ELU.

C. *Compensation Factor Prediction*

This paper finds that the measured mean I_{CELL} of S3 in replica cells, I_{replica}, is also linear with the logarithm of DR time (log(t)), and that the optimal α for every instance of DR time is linear against log(t) (Fig. 8(a)). Based on this new observation, the proposed replica cell circuit optimizes α (Fig. 8(b)). By logarithmic regression, optimal α is predicted directly from the measured I_{replica}. Optimal α is defined at the highest inference accuracy for every instance of DR time with Backward comp. (Fig. 8(c)). Note that all 1Kbit replica cells are written to S3. The area overhead and power/latency of the Kbit scale replica cell circuit are negligible considering the Mbit scale CiM. Finally, due to the error-tolerance of the NN, the proposed compensation is robust against I_{replica} variations. From the acceptable inference accuracy drop, the acceptable α variation ($\Delta\alpha$) and acceptable I_{replica} variation ($\Delta I_{\text{replica}}$) are determined (Figs. 9(a)(b)). Fig. 9(c) shows acceptable $\Delta\alpha$ for every instance of the measured DR time. In Fig. 9(d), the acceptable $\Delta I_{\text{replica}}$ is much larger than the measured actual $\Delta I_{\text{replica}}$ (I_{CELL} of S3 in Figs. 3(e)(f)). Thus, the proposed STCO is robust against replica cell variations.

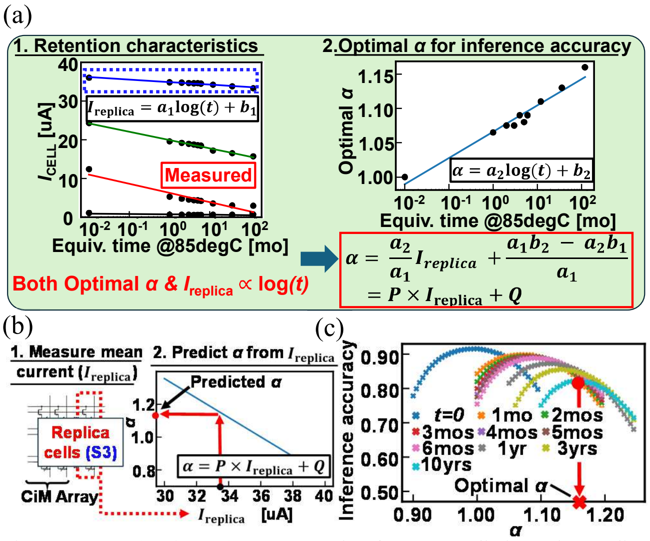
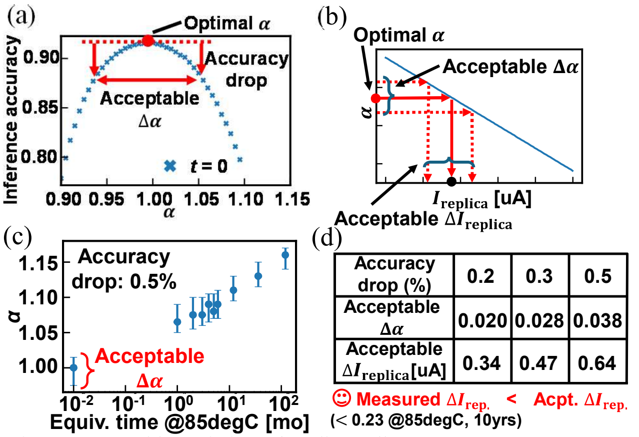


Fig. 8. *Prop. 3 (Tech&Cir)*: Compensation factor α prediction using replica cell circuit. (a) α prediction formula by mean I_{CELL} of S3 and optimal α after DR. (b) α prediction from mean I_{CELL} of S3 of replica cells, I_{replica}. (c) Optimal α that achieves the highest inference accuracy at any instance of DR time by Backward comp. (t_{ref} = 0).



Accuracy drop (%)	0.2	0.3	0.5
Acceptable $\Delta\alpha$	0.020	0.028	0.038
Acceptable $\Delta I_{replica}$[uA]	0.34	0.47	0.64

Fig. 9. Acceptable variation of replica cells. $\Delta\alpha$, ΔI_{CELL}, $\Delta I_{\text{replica}}$ are α, I_{CELL}, I_{replica} variations. (a) Determination of acceptable $\Delta\alpha$ from acceptable inference accuracy drop, for example, initial DR. (b) Determine acceptable $\Delta I_{\text{replica}}$ from acceptable $\Delta\alpha$. (c) Acceptable $\Delta\alpha$ after DR when inference accuracy drop < 0.5%. (d) Acceptable $\Delta\alpha$ and acceptable $\Delta I_{\text{replica}}$ with different inference accuracy drop percentage.

IV. Proposed Retention Aware Training with Measured Data-Retention Data

Retention Aware Training (RAT) with Bidirectional comp. is proposed, where measured 1-month DR data (Figs. 3(e)(f)) are injected into NN weights during training. Optimal α achieves high inference accuracy at every instance of DR time (Fig. 10(a)). Proposed Bidirectional comp. shows better accuracy than Backward/Forward comp. (Fig. 10(b)). As for the pre-cycling trade-off, 10 pre-cycles is always optimal even when I_{CELL} shifts by DR (Fig. 10(c)). Finally, Prop. 1+2+3+4 achieves high inference accuracy after 10-year DR (Fig. 10(d)). The entire result is shown in TABLE I.

V. Conclusion

This work proposes the approaches for ReRAM CiM that compensate for chip-level DR error and write variation. MLC data-retention measurements with longer term compared to the previous works (TABLE II) are also demonstrated. It is found that DR degradation is especially large in the intermediate states of ReRAM. By optimizing the number of pre-cycles during write operations in addition to proposed methods for inference and training, the degradation of inference accuracy is suppressed to a maximum of 1.5% over a 10-year period at 85degC with 40nm MLC $\mathrm{TaO_X}$ analog ReRAM CiM array.

Acknowledgement

This paper is based on results obtained from a project, JPNP23015, subsidized by the New Energy and Industrial Technology Development Organization (NEDO). The authors thank S. Yoneda, S. Ito, S. Awamura, M. Morimoto, and S. Muraoka of NTCJ.

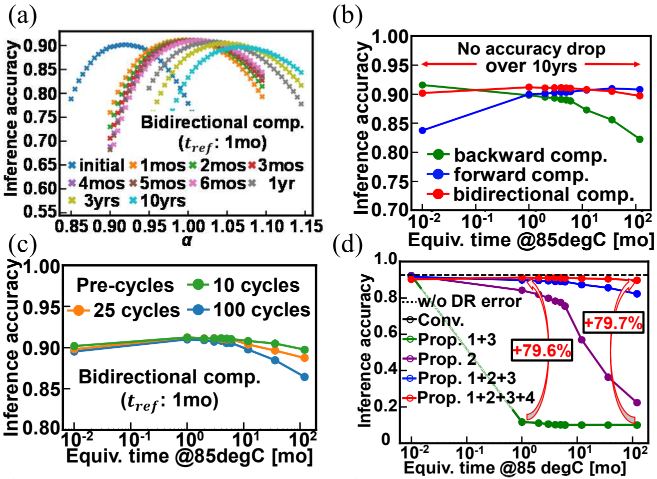


Fig. 10. *Prop. 4 (System&Tech)*: Retention Aware Training (RAT). (a) α dependent inference accuracy at different DR. (b) Inference accuracy comparison with three MACV compensation methods. (c) Inference accuracy comparison at different pre-cycling numbers w/ Bidirectional comp. and RAT. (d) Inference accuracy by proposed methods.

TABLE I. Inference accuracy improvement by proposed retention-tolerant ReRAM-based analog CiM

Accuracy [%]	Conv.	Prop. 1+3	Prop. 2	Prop. 1+2+3	Prop. 1+2+3+4
DR time $t = 0$	92.3	92.4	91.5	91.6	90.2
1 month	11.6	11.8	+79.6% 84.2	89.8	91.2
1 year	10.0	10.0	56.9	87.3	90.8 -1.5%
3 years	10.0	10.0	36.3	85.6	90.5
10 years	10.0	10.0	+79.7% 22.3	82.2	89.7

TABLE II. Comparison of chip-level prior data-retention measurements

	IRPS2022 [2] (Fig. 2)	IEDM2023 [9] (Fig. 19)	IRPS2024 [10] (Fig. 7)	This work (Fig. 3)
Technology node	90nm	40nm	40nm	40nm
Capacity	64Kbit	1Mbit	1Mbit	4Mbit
Device	$\mathrm{HfO_X}$	$\mathrm{TaO_X}$	$\mathrm{TaO_X}$	$\mathrm{TaO_X}$
MLC	2 bits/cell	3.3 bits/cell	4 bits/cell	2 bits/cell
Data-retention measurement	2 bits/cell 10^4 seconds 85degC	1 bit/cell 3 hours 563K	4 bits/cell 10^4 seconds 150degC	2 bits/cell 12.9 hours 190degC
Equivalent time	N.A.	10years 150degC	N.A.	10years 85degC

References

[1] Y.-H. Lin *et al.*, "Performance Impacts of Analog ReRAM Non-ideality on Neuromorphic Computing," *IEEE Transactions on Electron Devices*, 2019, vol. 66, no. 3, pp. 1289-1295.

[2] J. Meng *et al.*, "Sparse and Robust RRAM-based Efficient In-memory Computing for DNN Inference," *2022 IEEE International Reliability Physics Symposium*, 2022, pp. 3C.1-1-3C.1-6.

[3] Y. Zhang *et al.*, "An RRAM retention prediction framework using a convolutional neural network based on relaxation behavior," *Neuromorphic Computing and Engineering*, 2023, 3.1: 014011.

[4] L. Pistolesi *et al.*, "Drift Compensation in Multilevel PCM for in-Memory Computing Accelerators," *2024 IEEE International Reliability Physics Symposium*, 2024, pp. 1-4.

[5] A. Antolini *et al.*, "An embedded PCM Peripheral Unit adding Analog MAC In-Memory Computing Feature addressing Non-linearity and Time Drift Compensation," *IEEE 48th European Solid State Circuits Conference*, 2022, pp. 109-112.

[6] V. Voillet *et al.*, "Temperature and Drift-Aware High-Level PCMbased Array Model for Reliable Hardware IMC design," *2025 IEEE International Reliability Physics Symposium*, 2025, pp. 1-4.

[7] M. Morimoto *et al.*, "Resistive Switching Element scaling toward 22nm embedded ReRAM and beyond," *2024 International Conference on Solid State Device and Materials*, 2024, pp. 121-122.

[8] R. Mochida *et al.*, "A 4M Synapses integrated Analog ReRAM based 66.5 TOPS/W Neural-Network Processor with Cell Current Controlled Writing and Flexible Network Architecture," *2018 IEEE Symposium on VLSI Technology*, 2018, pp. 175-176.

[9] Q. Wang *et al.*, "A Logic-Process Compatible RRAM with 15.43 Mb/mm^2 Density and 10years@150°C retention using STI-less Dynamic-Gate and Self-Passivation Sidewall," *2023 International Electron Devices Meeting*, 2023, pp. 1-4.

[10] J. Sun et al. "ASAP: An efficient and reliable programming algorithm for multi-level RRAM cell," *2024 IEEE International Reliability Physics Symposium*, 2024, pp. 1-4.

[11] C.-Y. Wu *et al.*, "Emerging Memory RRAM Embedded in 12FFC FinFET Technology for industrial Applications," *2023 International Electron Devices Meeting*, 2023, pp. 1-4.

[12] T. Srimani *et al.*, "Foundry Monolithic 3D BEOL Transistor + Memory Stack: Iso-performance and Iso-footprint BEOL Carbon Nanotube FET+RRAM vs. FEOL Silicon FET+RRAM," *2023 IEEE Symposium on VLSI Technology and Circuits*, 2023, pp. 1-2.

[13] A. Kumar *et al.*, "Filament-Free Bulk RRAM with High Endurance and Long Retention for Neuromorphic Few-Shot Learning On-Chip," *2024 IEEE International Electron Devices Meeting*, 2024, pp. 1-4.

[14] D. Ielmini, "Modeling the Universal Set/Reset Characteristics of Bipolar RRAM by Field- and Temperature-Driven Filament Growth," in *IEEE Transactions on Electron Devices*, 2011, vol. 58, no. 12, pp. 4309-4317.

[15] R. Zhao *et al.*, "Improving neural network quantization without retraining using outlier channel splitting," *International conference on machine learning*. PMLR, 2019, pp. 7543-7552.

[16] T. Ninomiya *et al.*, "Conductive filament expansion in TaOx bipolar resistive random access memory during pulse cycling," *Japanese Journal of Applied Physics* 52.11R, 2013: 114201.

[17] Z. Wei *et al.*, "Retention Model for High-Density ReRAM," *2012 4th IEEE International Memory Workshop*, 2012, pp. 1-4.

[18] Z. Deng and M. Orshansky, "Variability-Aware Training and Self-Tuning of Highly Quantized DNNs for Analog PIM," *2022 Design, Automation & Test in Europe Conference & Exhibition*, 2022, pp. 712-717.