IEICE TRANSACTIONS

on Electronics

DOI:10.1587/transele.2023CTP0002

Publicized:2024/04/09

This advance publication article will be replaced by the finalized version after proofreading.

A PUBLICATION OF THE ELECTRONICS SOCIETY

The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3chome, Minato-ku, TOKYO, 105-0011 JAPAN

INPUT THE TYPE OF MANUSCRIPT

Comprehensive Analysis of Read Fluctuations in ReRAM CiM by Using Fluctuation Pattern Classifier

Ayumu Yamada*, Zhiyuan Huang, Naoko Misawa, Chihiro Matsui, *non-members,* and Ken Takeuchi, *member*

SUMMARY In this work, fluctuation patterns of ReRAM current are classified automatically by proposed fluctuation pattern classifier (FPC). FPC is trained with artificially created dataset to overcome the difficulties of measured current signals, including the annotation cost and imbalanced data amount. Using FPC, fluctuation occurrence under different write conditions is analyzed for both HRS and LRS current. Based on the measurement and classification results, physical models of fluctuations are established.

key words: *ReRAM, Computation-in-Memory (CiM), Fluctuation, RTN, Oxygen Vacancy.*

1. Introduction

In recent years, conventional Neumann-type computer architecture faces issues of energy consumption and large latency by transporting data from memory to the processor. To tackle this problem, Computation-in-Memory (CiM) has been proposed [1-16]. It computes multiply-accumulate (MAC) operations in memory array without data transportation.

As a memory of CiM, several types of memories have been considered, i.e., static random access memory (SRAM) [1], resistive RAM (ReRAM) [2-12], phase change memory (PCM) [13], and ferroelectric field effect transistor (FeFET) [14-15]. Although volatile memory-based CiM has advantages in some points such as low error rate and process cost, non-volatile memory-based CiM (nvCiM) is a good option for edge application because of its analog characteristics and energy-efficiency. Especially, ReRAM CiM has been studied for its small cell area, CMOS compatibility and so on.

Because the MAC operation is a main computational operation in neural network, CiM is also used as an accelerator for simulated annealing [9] and hyperdimensional computing [15], but its application with respect to neural networks is well studied especially. To implement neural network on ReRAM CiM, its weights are mapped as the conductance of ReRAM cells. Thus, memory non-idealities, including write variation and read-disturb [17] are the large issues in ReRAM CiM [16]. In addition, ReRAM has been suffered from its conductance fluctuation [2-7, 18-26]. The main fluctuation pattern has been considered the random telegraph noise (RTN) conventionally, but there are still other fluctuation patterns are reported [2, 18-22, 26]. G. González-Cordero et al. (2021) adopted self-organization map (SOM) to fluctuating signal analysis [26], but it does not work properly on our measured signals. In addition, such unsupervised method requires the interpretation of the obtained results.

In this work, the fluctuation patterns of ReRAM are investigated in both high resistance state (HRS) and low resistance state (LRS). To analyze complicated fluctuation patterns automatically, fluctuation pattern classifier (FPC) is constructed. FPC is trained on artificial dataset [2].

The achievements of this work are as follows:

- In section 2, characteristics of ReRAM readout current is presented and the impact of conductance fluctuation on CiM-based neural network accelerator is investigated.
- In section 3, CNN-based Fluctuation Pattern Classifier (FPC) is trained with artificial dataset, created by Markov model-based Synthetic Data Generator (MM-SDG). Moreover, FPC trained on artificial data is applied to measured signals, and Fluctuation Reduction Write (FRW) is proposed based on the results obtained

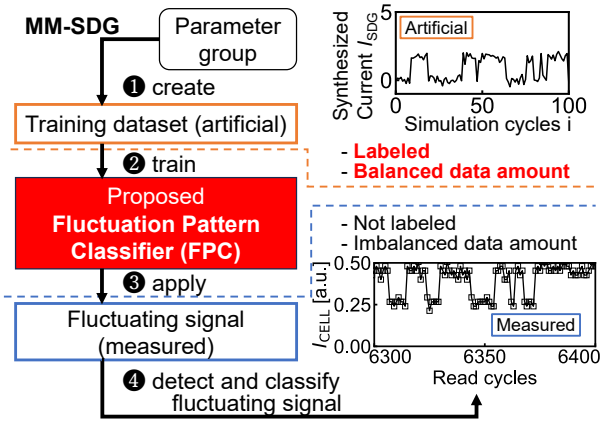

Fig. 1 Overview of this work. Proposed fluctuation pattern classifier (FPC) is trained on artificial training dataset, created by Markov model-based synthetic data generator (MM-SDG) under assumed parameters. Trained FPC is applied to measured fluctuating signals and it detects fluctuations and classify their patterns.

The authors are with Department of Electrical Engineering and Information Systems, The University of Tokyo, Bunkyo-ku, Tokyo, 113-8656 Japan.
*e-mail: yamada@co-design.t.u-tokyo.ac.jp

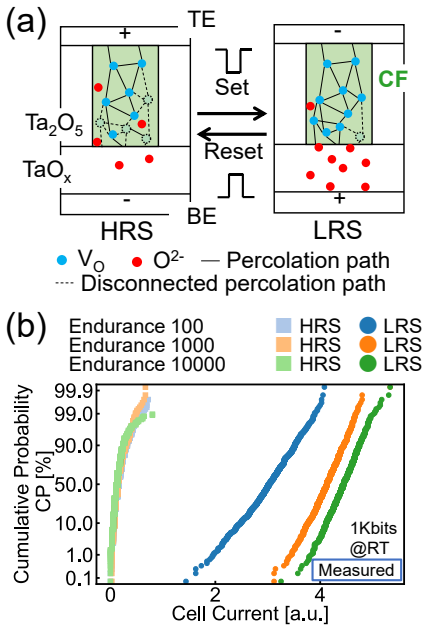

Fig. 2 (a) Structure and switching model of TaOx-based ReRAM. Set and Reset voltage moves Oxygen ions and the resistance of conductive filament (CF) changes (© 2023 IEEE [2]). (b) Measured current distribution under 100, 1000, and 10000 endurance cycles.

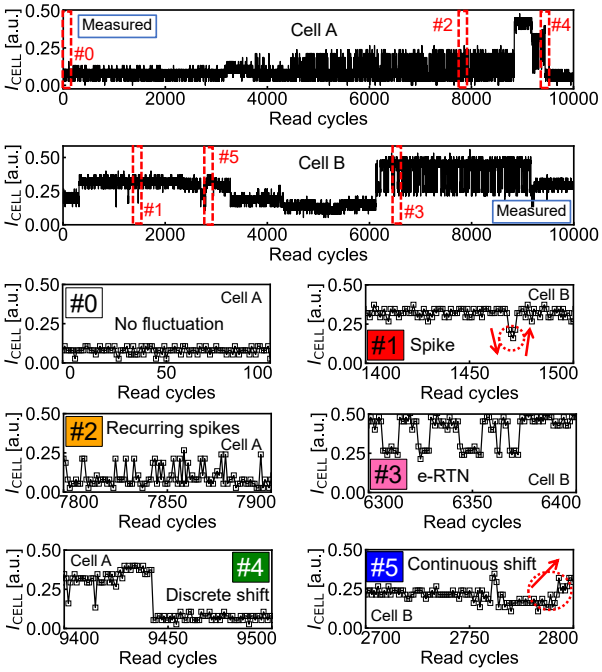

Fig. 3 Measured ReRAM current series. 5 patterns (#1-#5) of fluctuation and #0 ('no fluctuation') are assumed in this work (© 2023 IEEE [2]).

by FPC.

- In section 4, physical models of ReRAM fluctuation for both HRS and LRS are established based on the observation. Differences between the characteristics of fluctuation occurrence under different endurance/verification conditions are explained by the physical models.

2. Resistive Random Access Memory (ReRAM)

2.1 Principle

Resistive Random Access Memory (ReRAM) is one of the non-volatile memories. Structure of TaO_X-based ReRAM and its switching model is shown in Fig. 2(a). Each ReRAM cell has a layered metal-insulator-metal structure, and stores information according to the resistance of the insulator part. According to the hopping percolation model, which is usually used to explain ReRAM conduction [27-30], electrons conduct in the form of moving from one hopping site to another in an insulator layer. Hopping sites are due to Oxygen vacancies (V_O). Hence, the resistance of ReRAM cell is controlled by applying voltage externally. As shown in Fig. 2(a), set voltage (V_{SET}) ejects oxygen ions (O^{2-}) from the Ta_2O_5 layer to the reservoir layer (TaO_X) and then the cell becomes low resistance state (LRS). In the contrary, oxygen ions move from reservoir layer to Ta_2O_5 by reset voltage (V_{RESET}) and the cell becomes high resistance state (HRS). Read voltage (V_{READ}) is applied as well as set and reset, but the voltage is low compared with set and reset voltage and thus the resistance does not change largely.

Fig. 2(b) shows the measured current distribution of 1K cells at room temperature. The current distributions are divided between HRS and LRS, and the LRS current value increases with increasing number of endurance (set/reset) cycles.

2.2 Fluctuation

Despite the low voltage, the actual observed current value varies with the readout. For ReRAM, this is mainly considered to be caused by random telegraph noise (RTN). Conventionally, RTN is explained as the discrete current level change by the electron trapping/de-trapping to a trap site. However, a number of patterns are observed in the changes of measured current value that are different from the discrete changes between several levels, such as RTN, and these cannot be explained by the simple RTN model.

2.3 Impact of fluctuation on ReRAM CiM

Several research studies have reported that the RTN has significant impact on the inference accuracy of ReRAM CiM-based neural network accelerator [3-7]. In these works, RTN is simulated by physical parameters or sampled from measured distributions.

In this work, a 3-layer multilayer perceptron (MLP) in Fig. 4(a) is used to estimate the impact of fluctuations on ReRAM CiM-based neural network accelerator. The number of nodes in the hidden layer is 100. MNIST dataset is used to train/test the MLP. Weight and input/output quantization are 4-bit. Each weight of the MLP is

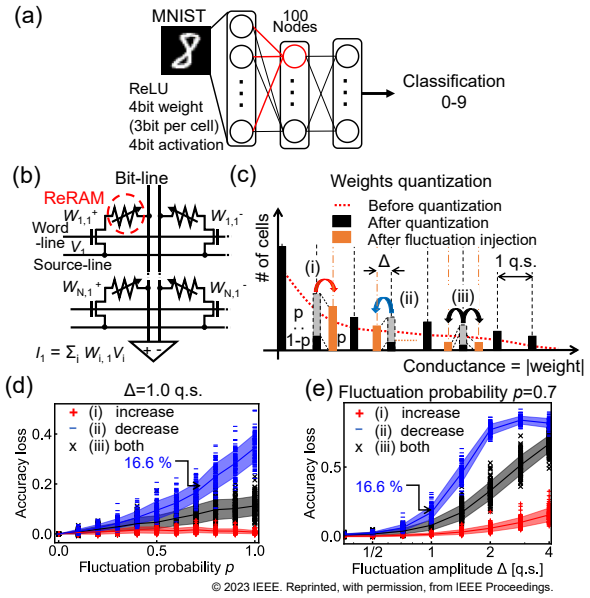

Fig. 4 (a) 3-layer multilayer perceptron (MLP) model used to investigate the impact of fluctuation. (b) ReRAM CiM architecture. (c) Simplified fluctuation model with fluctuation probability p and amplitude Δ. Impact of fluctuation with (d) probability p under Δ=1.0 q.s. and (e) amplitude Δ under p=0.7. (© 2023 IEEE [2])

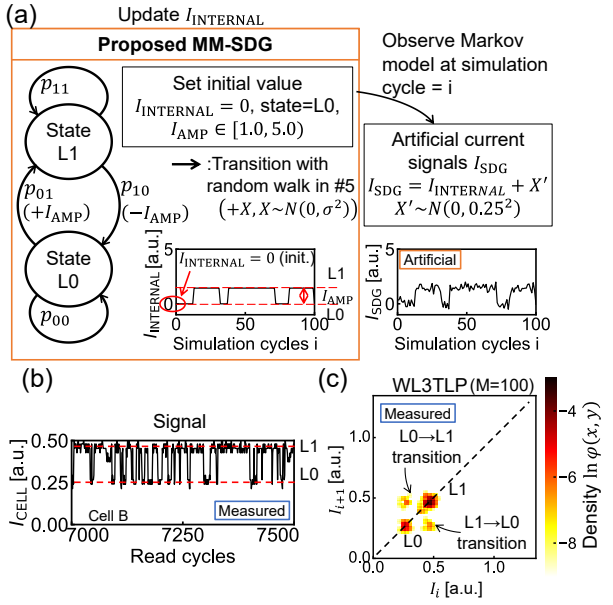

Fig. 5 (a) Proposed Markov model-based Synthetic Data Generator (MM-SDG). Synthesized signal I_{SDG} is obtained by updating $I_{INTERNAL}$ and observing the state of Markov model for 100 cycles. State transition between 2 state (L0, L1) random walk happens with probabilities in Table I. (b) A sample of measured current signals with 2 levels (L0, L1). (c) Converted WL3TLP with M×M (M=100) pixels [24]. Clusters on diagonal represent no transition and others represent level transitions. (© 2023 IEEE [2])

represented by a pair of ReRAM cells, as shown in Fig. 4(b). Because the fluctuation patterns considered in this study are not confined to the RTN, simplified fluctuation model with two-level is introduced in the simulation of the MLP. Instead, the probability and magnitude of these fluctuations are adjusted arbitrarily to represent general fluctuations. This simplified fluctuation model, depicted in Fig. 4(c), introduces two key parameters: fluctuation probability p and fluctuation amplitude Δ. Moreover, the fluctuation direction is varied, including (i) increase, (ii) decrease, and (iii) both directions. The proportion of weight changes is defined as the fluctuation probability p, and the magnitude of weight change is determined by the fluctuation amplitude Δ. Δ is normalized by the quantization step (q.s.). When the fluctuation direction is (i) increase or (ii) decrease, the conductances of ReRAM cells either increase or decrease by Δ with a probability of p. In case of (iii) both directions, the conductances of ReRAM cells both increase and decrease by Δ with a probability of p/2 in each direction.

Figs. 2(d) and 2(e) show the degradation in inference accuracy resulting from the simplified fluctuation model. Both an increase in fluctuation probability p and an increase in amplitude Δ lead to reduced accuracy. When the fluctuation direction is (ii) decrease, with parameters set at p=0.7 and Δ=1.0 q.s., the inference accuracy decreases by 16.6% on average.

Since the weights are represented by a pair of HRS and LRS cells (or 2 HRS cells if the weight value is 0) in the assumed two-cell differential pair weight mapping scheme, more than half of ReRAM cells are set to HRS. In addition, the weights of a neural network are basically sparse matrices, the number of HRS cells is overwhelmingly large. Therefore, it is important to control the conductance of the HRS as well as the LRS, which represents the absolute value of the weights.

3. Fluctuation Pattern Classifier

3.1 Fluctuation Pattern Classifier (FPC)

As depicted in Fig. 3, the measured signals exhibit various fluctuation patterns, even when originating from the same ReRAM cell. Additionally, to use these signals as training dataset for supervised learning, they pose challenges for annotation. Even in cases where annotation is feasible, it may involve an unequal distribution of data, such as an overabundance of pattern #0 compared to others, making accurate learning difficult. Consequently, supervised learning with the measured signals is not a suitable approach for this application. Alternatively, unsupervised learning is a viable method, but it necessitates the interpretation of results. In this study, a solution is presented to classify ReRAM fluctuation patterns without the need for interpretation. This solution converts time-series current data into 2D images and classify them with a Convolutional Neural Network (CNN)-based Fluctuation Pattern Classifier (FPC). FPC is trained with a synthetic dataset through supervised learning.

Table 1 Parameters of MM-SDG used in Fig. 5(a)

#	p_{00}	p_{01}	p_{10}	p_{11}	σ	μ	Tran.
0H/L	1.0	0.0	0.0	1.0	-	-	= 0
1H/L	0.99	0.01	0.50	0.50	-	-	≥ 1
2H/L	0.75	0.25	0.95	0.05	-	-	≥ 1
3H/L	0.90	0.10	0.10	0.90	-	-	≥ 1
4H	0.99	0.01	0.01	0.99	-	-	≥ 1
4L	0.99	0.01	0.0	1.0	-	-	≥ 1
5H	1.0	0.0	0.0	1.0	0.05-0.10	0.0	= 0
5L	1.0	0.0	0.0	1.0	0.15-0.20	0.02	=0

3.2 MM-SGD configuration for LRS and HRS

The datasets used for training and testing the FPC are generated artificially using the proposed Markov model-based Synthetic Data Generator (MM-SDG, Fig. 5(a)). MM-SDG has an internal Markov model, where state transits following probabilities (p_{00}, p_{01}, p_{10}, p_{11}). The Markov model is assumed to consist of two states, denoted as L0 and L1, which correspond to the observed current I_{CELL} and transitions between these two levels, as shown in Fig. 5(b).
Each dataset consists of a sequence of synthesized current values (I_{SDG}) derived from 100 updates of the Markov model within MM-SDG. During the dataset creation process, parameters such as transition probabilities, amplitude (I_{AMP}), and properties of random walks (mean and distribution, denoted as μ and σ respectively) are defined, and initial setting for $I_{INTERNAL}$ in MM-SDG is set to 0. I_{AMP} is selected randomly from a uniform distribution in the range [1.0, 5.0). Additionally, the Markov model's state is initially set to L0. Under these conditions, if the state of the Markov model transitions from LS_i to LS_{i+1} (where S_i and S_{i+1} are either 0 or 1) during simulation cycle i to i+1, the value of $I_{INTERNAL}$(i+1) is determined using the following Equation (1):

$$I_{\mathrm{INTERNAL}}(\mathrm{i}+1) = I_{\mathrm{INTERNAL}}(\mathrm{i}) + (S_{\mathrm{i}+1} - S_{\mathrm{i}})I_{\mathrm{AMP}} + X \quad (1)$$

where X is a randomly sampled value from a Gaussian distribution $N(\mu, \sigma^2)$, if σ is not 0.
Artificial current signal I_{SDG} is obtained by the following equation (2) based on $I_{INTERNAL}$:

$$I_{\mathrm{SDG}}(\mathrm{i}) = I_{\mathrm{INTERNAL}}(\mathrm{i}) + X' \quad (2)$$

where X' is a randomly chosen value from a Gaussian distribution to account for fluctuations of sufficiently small amplitude.
This work categorizes fluctuations into five distinct patterns, denoted as #1-#5, along with a "no fluctuation" category represented as #0, as depicted in Fig. 2. To achieve accurate classification, the Fluctuation Pattern Classifier (FPC) is individually trained for both the HRS (HRS model) and LRS (LRS model). Parameters for the MM-SDG for both the HRS and LRS models are listed in Table 1. While parameters #0-#3 are shared between these models,

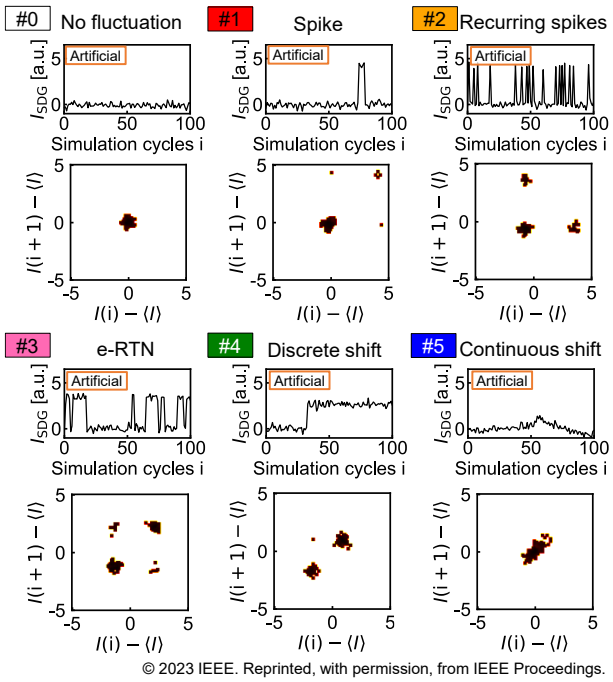

Fig. 6 Samples of each pattern of artificial current signals I_{SDG} and corresponding WL3TLP obtained from MM-SDG with HRS parameters (© 2023 IEEE [2]).

parameters for #4 and #5 differ. A sample of each pattern for the artificial current signal I_{SDG} for the HRS model is shown in Fig. 6. Fluctuations #1 and #2 exhibit spikes within their respective current sequences, with #2 having a greater number of spikes due to higher transition probabilities (p_{01} and p_{10}). Fluctuation #3 emulates electron Random Telegraph Noise (e-RTN) with p_{01}=p_{10}=0.1. Discrete and continuous shifts in ISDG are assumed for fluctuation patterns #4 and #5, respectively. Note that, if no state transition occurs in 100 simulation cycles, I_{SDG} data is not included in the generated dataset for patterns #1-#4.
To train (test) FPC, MM-SDG creates the training (test) dataset with 24,000 (2,400) fluctuation data in total, 2,000 (200) fluctuation data and their sign-reversed data for each pattern.

3.3 Preprocessing

Because measured I_{CELL} and generated I_{SGD} signals are time-series data, they are required to be transformed into 2D images to input to the CNN-based FPC. There exist some techniques for converting signals into 2D images, with one example being the Time-lag plot (TLP).
TLP essentially forms a scatter plot based on pairs of data points, (I(i), I(i+1)). To address noisy signals, an extended variant called Weighted TLP (WTLP) has been introduced [23]. However, it's worth noting that WTLP demands substantial computational resources. To achieving both noise robustness and reduced computational complexity compared to WTLP, a method known as Locally Weighted TLP (LWTLP) was proposed. Especially, WL3TLP

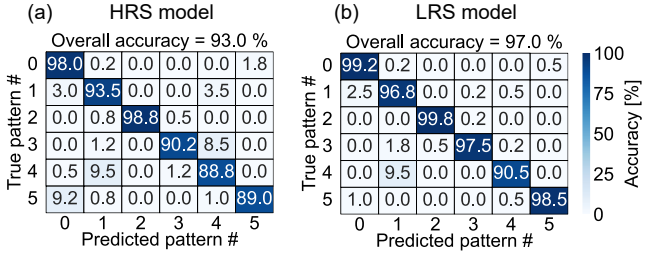

Fig. 7 Confusion matrix for the FPC on synthetic test dataset. (a) HRS model achieves 93.0% (© 2023 IEEE [2]) and (b) LRS model achieves 97.0% accuracy in total.

represents a special case of LWTLP [24]. In this work, a WL3TLP with a grid of M×M pixels (M=100), is adopted for converting sequences of I_{CELL} or I_{SGD} into images. As seen in Figs. 5(b) and 5(c), a WL3TLP applied to a series of current signals featuring two distinct levels exhibits two clusters along the diagonal of the WL3TLP plot. These clusters correspond to the current levels, as the diagonal segments indicate the absence of transitions between current levels (I(i)=I(i+1)), while other clusters signify level transitions.

To prevent the CNN-based FPC from making determinations based on the position of the current levels in the WL3TLP, namely the absolute current values, the mean of the current series (denoted as $\langle I \rangle$) is subtracted when converting to WL3TLP. The range of the WL3TLP is set within -5 uA (a.u.) to +5 uA (a.u.) for I_{CELL} (I_{SGD}). A sample of each assumed fluctuation pattern generated through MM-SDG and its corresponding WL3TLP representation is illustrated in Fig. 6.

FPCs (named as HRS and LRS models) achieved prediction accuracies of 93.0% and 97.0% on the HRS and LRS test datasets, respectively. The confusion matrices detailing the FPC's performance on the test dataset can be found in Figs. 7(a) and 7(b).

3.4 Application to measured data

Fig. 8(a) (8(b)) shows the prediction outcomes of the CNN-based FPC for measured HRS (LRS) current signals featuring various fluctuation patterns. These predictions were made using the FPC trained on artificially created datasets. Fluctuation predictions are performed on segments consisting of 100 read cycles, and the FPCs offer reasonable predictions even for the measured current data.

Furthermore, the FPCs are applied to classify measured signals under various Verify/Endurance conditions before read cycles. For each condition, measurements and analyses were conducted on 1K cells. Fig. 9 illustrates the number of cells predicted for each fluctuation pattern under the following conditions: V_{SET}=2.7V, V_{RESET}=2.0V, Endurance=100, w/o Verify. V_{READ} during read cycles is set to 0.4V.

In this case, fluctuation patterns #4 and #5 are the primary factors causing fluctuation. The fluctuation rate (FR), which represents the proportion of cells predicted to exhibit fluctuations at least once within 500 read cycles immediately after the conventional write, stands at 41.8% (Fig. 9, left) but decreases to 10.7% (Fig. 9, right) after 500 pre-read cycles. Here, because fluctuation pattern #1 has limited impact on the ReRAM CiM-based neural network accelerator due to the rapid return of the current to the dominant value after a fluctuation, patterns #4 and #5 assume greater significance. Note that, the cells shown in Figs. 3 and 8 represent samples with different fluctuation patterns within the same cell readout and do not indicate an overall trend of fluctuation occurrence.

Fig. 10 shows the number of cells predicted to be fluctuated as each pattern under varying Verify/Endurance conditions before read cycles for HRS current. The number of maximum Verify cycles ranges from 0 (no Verify) to 10 and 100, with Endurance cycles fixed at 100 (upper part of Fig. 10). Additionally, the number of Endurance cycles varies from 100, 1000, to 10000, without Verify (lower part of Fig. 10). Voltage conditions remain consistent with those in Fig. 9.

The occurrence of fluctuation patterns #4 and #5 diminishes as the number of Endurance cycles before read cycles increases. Note that, the number of Verify cycles does not appear to significantly impact the occurrence of fluctuations. Fig. 11 shows the number of cells predicted to be fluctuating under different conditions for LRS current. The dominant fluctuation pattern is #5 (continuous shift). Fluctuation

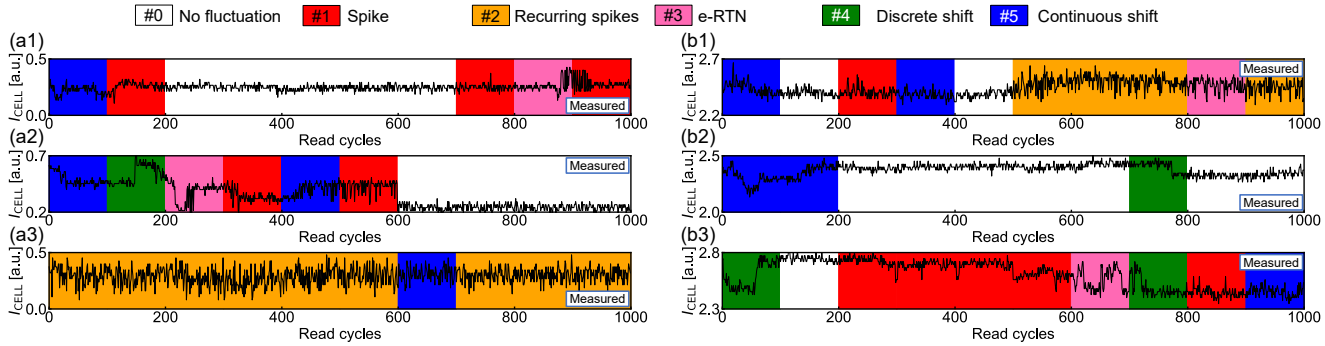

Fig. 8 Results of CNN-based FPC trained on synthetic data applied to the measured current signals I_{CELL} of 3 cells for (a) HRS and (b) LRS. Each color corresponds to samples in Fig. 2. Prediction is successfully achieved for each section of 100 read cycles. In all graphs in Figure 8, the range of the current on the vertical axis is aligned at 0.5 a.u.. Note that these cells are the samples of those containing different fluctuation patterns in the same cell readout and do not represent an overall trend of declining the fluctuation occurrence shown in Fig. 9.

occurrence decreases as the cycles increases as well as HRS current. In addition, Verify and higher endurance also improves fluctuation occurrence. Unlike HRS current, LRS fluctuation improves by Verify, but high-Endurance is more remarkable compared with Verify.

Hence, Fluctuation Reduction Write (FRW) to reduce fluctuation occurrence is introduced based on the observation, involving a combination of high-Endurance and extended pre-read cycles. The protocol for FRW can be found in Fig. 12.

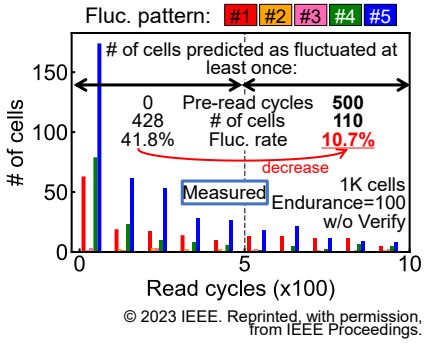

Fig. 9 Number of cells predicted to be fluctuating. Fluctuation rate reduces to 10.7% with 500 pre-read cycles. 1K cells are measured with V_{SET}=2.7V, V_{RESET} =2.0V, V_{READ} =0.4V (© 2023 IEEE [2]).

Fluctuations degrades inference accuracy when they occur during the inference phase of the ReRAM CiM-based neural network accelerator. By implementing a higher number of endurance cycles and incorporating the early phase of readout (pre-read), which is susceptible to fluctuations, into the write operation, the impact of fluctuations on the neural network, along with the fluctuation rate, can be significantly reduced. If necessary, Verify operations can also be performed.

Table 2 provides insights into the fluctuation rate (FR) under various combinations of Endurance and pre-read cycle conditions. For HRS current, adopting high-Endurance (10000) or extending pre-read cycles (500) alone reduces the fluctuation rate by 11.5 points and 31.1 points, respectively. The lowest fluctuation rate is achieved when both high-Endurance and long pre-read cycles (FRW) are employed, resulting in a 35.2 points reduction in ReRAM fluctuation rate.

For LRS current, 67.4 points and 37.9 points improvement in fluctuation rate is achieved by high-endurance and long pre-read. FRW improves fluctuation rate by 74.7 points.

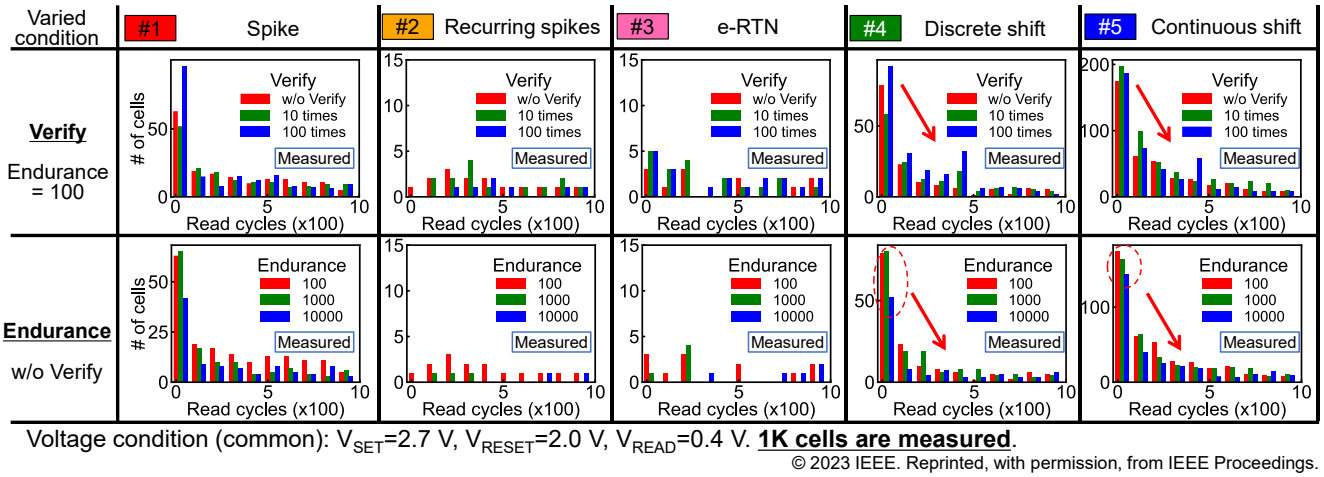

Fig. 10 Number of cells predicted as fluctuating with different Verify (upper) and Endurance (lower) before read cycles for HRS current. Dominant fluctuation patterns are #1, #4 and #5. By increasing read cycles, occurrence of fluctuation patterns #4 and #5 decrease (© 2023 IEEE [2]).

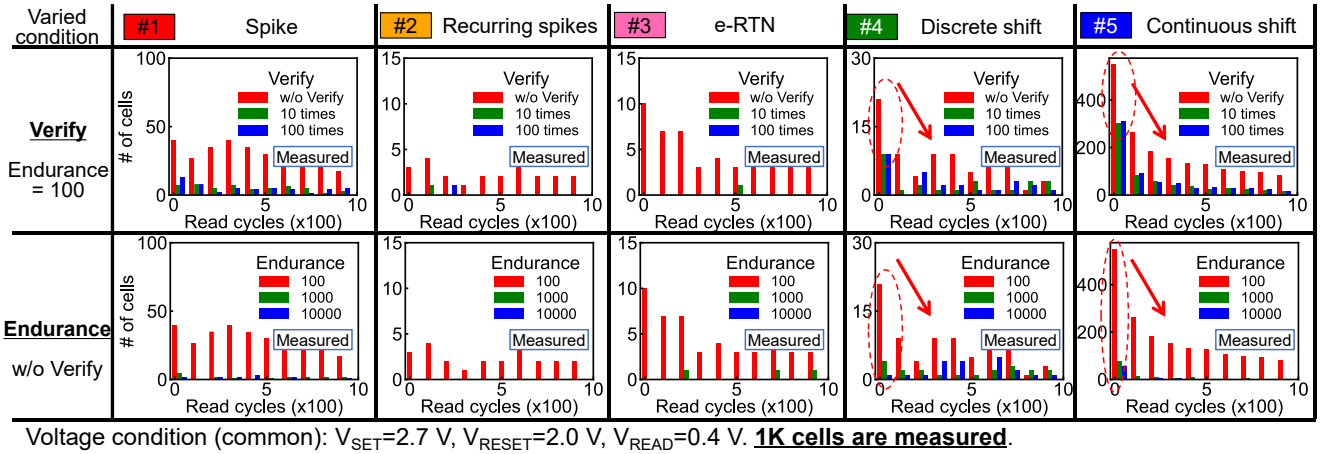

Fig. 11 Number of cells predicted as fluctuating with different Verify (upper) and Endurance (lower) before read cycles for LRS current. Dominant fluctuation pattern is #5. By increasing read cycles, occurrence of fluctuation patterns #5 decrease.

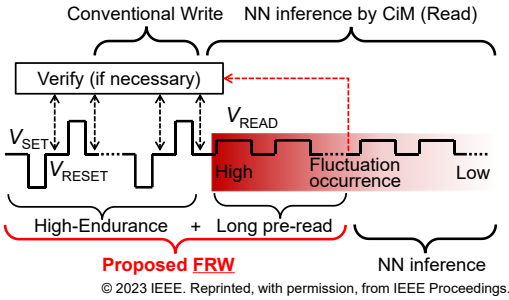

Fig. 12 Protocol of proposed FRW. Because the fluctuation rate just after conventional write is higher than that of after pre-read cycles, FRW with both high-Endurance and long pre-read cycles is proposed (© 2023 IEEE [2]).

Table 2 Fluctuation rate (FR) under different conditions and FRW

Endurance	Verify	FR of HRS [%]		FR of LRS [%]	
		Pre-read 0	Pre-read 500	Pre-read 0	Pre-read 500
100	0	41.8	10.7	76.8	38.9
100	10	45.5	12.7	43.8	12.0
100	100	48.7	7.9	44.6	12.8
1000	0	39.3	8.6	12.1	2.9
10000	0	30.3	6.6	9.4	2.1

If Endurance cycles increased more, the fluctuation rate could be reduced further. However, because ultra-high Endurance cycles (e.g. 10^5 or 10^6) will increase bit-error rate [9], 10^4 is selected as Endurance of FRW in this work.

4. Physical model

Figs. 13(a) and (b) show the physical models for HRS and LRS, respectively. In fluctuation patterns #1, #2, and #3, electron trapping (de-trapping) at hopping sites (Oxygen vacancy, V_O) leads to the disconnection (connection) of percolation paths with varying capture/emission time constants for the trap sites (Figs. 13(a1) and 13(b1)). This model aligns with the conventional Random Telegraph Noise (RTN) model.

Fluctuation #4 is a result of O^{2-} ion trapping/de-trapping at the V_O, although the number of remaining O^{2-} ions decreases with an increase in read cycles due to the bonding between O^{2-} ions and V_O. In these fluctuation patterns, percolation paths are either connected or disconnected, leading to discrete current levels determined by the state of the percolation paths.

Continuous current shifts (fluctuation #5) are brought about by the deformation of the lattice structure of β-Ta_2O_5 (see Fig. 13(c)) [31, 32]. Set/Reset voltages strain the lattice structure by extracting lattice oxygen (O_L) atoms during Set (Fig. 13(c2)) or injecting O^{2-} ions during Reset (Fig. 13(c4)) into/from the lattice. The quasi-thermal equilibrium of the lattice with O_L differs from that without O_L. Consequently, lattice relaxation and geometry optimization occur due to the electric field or Joule heating from V_{READ} during the pre-read process of FRW [31]. Through relaxation and geometry optimization, the lattice transitions to a quasi-thermal equilibrium state, resulting in changes in the potential energy of V_O, corresponding to V_O movement (as depicted in Figs. 13(a2) and 13(b2). Since the transition of the lattice to the stable state is continuous, the current value also undergoes a continuous shift, ultimately reaching stability when the lattice reaches quasi-thermal equilibrium.

As the number of Endurance cycles increases, the size of the conductive filament (CF) expands [8, 27]. In the case of HRS, the fluctuation rate decreases because the expansion of CF results in an increased number of hopping sites not involved in conduction (Fig. 13(a3)).

Meanwhile, for LRS, the electron conduction become ohmic as the number of V_O is large, and the trap sites causing electron trapping/de-trapping are isolated (red circle in Fig. 13(b)) from CF [18, 25]. Thus electron trapping/de-trapping to partial isolated sites have a few impacts on the cell current. In addition, Verify also shows a decreasing trend in fluctuation rate for LRS. This corresponds to increasing the number of Set/Reset cycles, especially for ReRAM cells that do not meet the Verify reference current in LRS (i.e., when the number of V_O is small and conductive filament is not sufficiently expanded). Thus, the fluctuation rate decreases due to Verify is explained in the same way as Endurance. Another characteristic point about the number of fluctuations in the LRS is the sharp decrease in fluctuation pattern #5 (continuous shift) with increasing endurance.

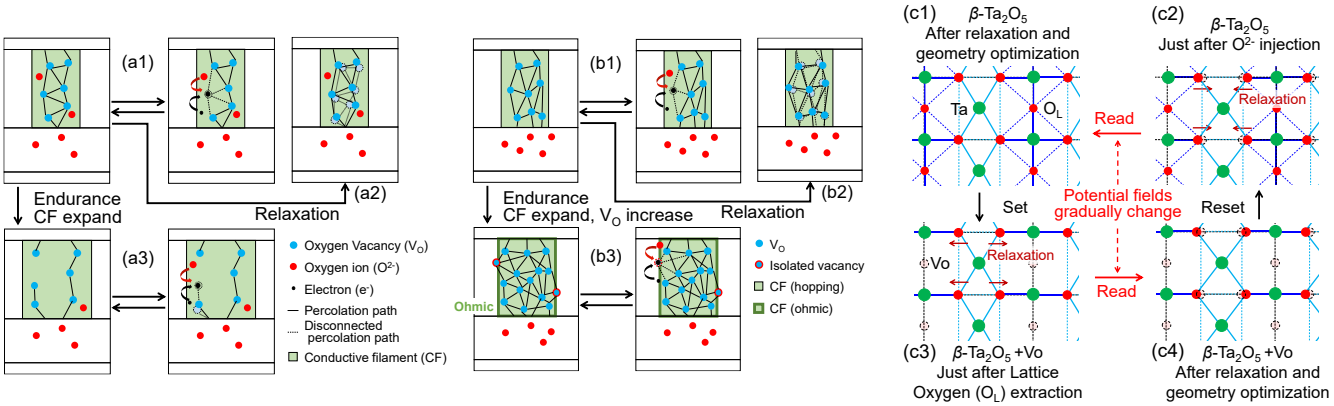

Fig. 13 Physical models of ReRAM for (a) HRS fluctuation, (b) LRS fluctuation and (c) lattice relaxation of β-Ta_2O_5 ((a), (c) © 2023 IEEE [2]).

Although a decrease in pattern #5 with increasing endurance is also observed in the HRS, the decrease is less severe than in the LRS. When CF is sufficiently developed and the conduction characteristics in CF is ohmic, the effects of lattice relaxation are averaged out and it is unlikely to cause significant changes to the measured current values. As a result, the number of continuous shift detected in LRS current is reduced.

5. Conclusion

Fluctuation on ReRAM CiM-based neural network accelerator causes large inference accuracy loss. To estimate the impact of fluctuation on ReRAM CiM and design high-performance ReRAM CiM-based NN accelerator, analytical methods of ReRAM fluctuating signals are required.
Measured ReRAM fluctuating signals such as RTN and V_O movement are classified by the proposed CNN-based FPC for both HRS and LRS. Artificially created labeled dataset created by the proposed MM-SDG enables supervised learning of FPC. Effectiveness of proposed FRW with high Endurance and long pre-read cycles to decline the occurrence of ReRAM fluctuation is demonstrated by analysis of the measured I_{CELL} signals of ReRAM under different conditions by applying FPC. Higher-Endurance of FRW expands CF of ReRAM, and longer pre-read cycles cause relaxation and geometry optimization of Ta_2O_5 lattice structure due to the electric field and Joule-heating by V_{READ}. Proposed FRW improves fluctuation rate of ReRAM by 35.2 points for HRS and 74.7 points for LRS current. However, proposed FRW develops CF and the conductance increases for LRS. If, to represent mid-range weight values, CFs are not fully developed and the cells in intermediate resistance states, fluctuations are more likely to occur, and the inference accuracy will degrade.
Proposed analysis method using neural networks trained on artificial data is effective not only for ReRAM fluctuations, but also for applications where it is difficult to annotate training data or where the amount of training data is imbalanced for each label.

Acknowledgments

The authors thank H. Akinaga, H. Shima, and Y. Naitoh of National Institute of Advanced Industrial Science and Technology (AIST) and S. Yoneda, S. Ito, S. Muraoka, and K. Kawai of Nuvoton Technology Corporation Japan (NTCJ). This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

[1] K. Lee, J. Kim and J. Park, "Low-cost 7T-SRAM Compute-In-Memory design based on bit-line charge-sharing based analog-to-digital conversion," 2022 IEEE/ACM International Conference on Computer Aided Design (ICCAD), San Diego, CA, USA, 2022, pp. 1-8.

[2] A. Yamada, N. Misawa, C. Matsui and K. Takeuchi, "ReRAM CiM Fluctuation Pattern Classification by CNN Trained on Artificially Created Dataset," 2023 IEEE International Reliability Physics Symposium (IRPS), Monterey, CA, USA, 2023, pp. 1-6. DOI: 10.1109/IRPS48203.2023.10118305.

[3] T. Zanotti, F. M. Puglisi and P. Pavan, "Low-bit precision neural network architecture with high immunity to variability and random telegraph noise based on resistive memories," 2021 IEEE International Reliability Physics Symposium (IRPS), Monterey, CA, USA, 2021, pp. 1-6. DOI: 10.1109/IRPS46558.2021.9405103.

[4] Y. Du, L. Jing, H. Fang, H. Chen, Y. Cai, R. Wang, J. Zhang, and Z. Ji, "Exploring the impact of random telegraph noise-induced accuracy loss on resistive RAM-based deep neural network," in IEEE Transactions on Electron Devices, vol. 67, no. 8, pp. 3335-3340, Aug. 2020.

[5] D. Joksas, P. Freitas, Z. Chai, W. H. Ng, M. Buckwell, C. Li, W. D. Zhang, Q. Xia, A. J. Kenyon and A. Mehonic, "Committee machines—a universal method to deal with non-idealities in memristor-based neural networks," Nature Communications, vol. 11, 4273, 2020. DOI: 10.1038/s41467-020-18098-0.

[6] Z. Chai, P. Freitas, W. Zhang, F. Hatem, J. F. Zhang, J. Marsland, B. Govoreanu, L. Goux, and G. S. Kar, "Impact of RTN on pattern recognition accuracy of RRAM-based synaptic neural network," in IEEE Electron Device Letters, vol. 39, no. 11, pp. 1652-1655, Nov. 2018. DOI: 10.1109/LED.2018.2869072.

[7] J. Kang, Z. Yu, L. Wu, Y. Fang, Z. Wang, Y. Cai, Z. Ji, J. Zhang, R. Wang, Y. Yang and R. Huang, "Time-dependent variability in RRAM-based analog neuromorphic system for pattern recognition," 2017 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2017, pp. 6.4.1-6.4.4. DOI: 10.1109/IEDM.2017.8268340.

[8] M. Arita, A. Tsurumaki-Fukuchi, Y. Takahashi, S. Muraoka, S. Ito and S. Yoneda, "Nanoscale filaments in Ta-O resistive RAM bit array: microscopy analysis and switching property," 2019 IEEE 11th International Memory Workshop (IMW), Monterey, CA, USA, 2019, pp. 1-4. DOI: 10.1109/IMW.2019.8739389.

[9] K. Taoka, N. Misawa, S. Koshino, C. Matsui and K. Takeuchi, "Simulated annealing algorithm & ReRAM device co-optimization for Computation-in-Memory," 2021 IEEE International Memory Workshop (IMW), Dresden, Germany, 2021, pp. 1-4. DOI: 10.1109/IMW51353.2021.9439610.

[10] M. F. Chang, J. M. Hung, P. C. Chen and T. H. Wen, "Reliable computing of ReRAM based Compute-in-Memory circuits for AI edge devices," 2022 IEEE/ACM International Conference On Computer Aided Design (ICCAD), San Diego, CA, USA, 2022, pp. 1-6. DOI: 10.1145/3508352.3561119.

[11] R. Mochida, K. Kouno, Y. Hayata, M. Nakayama, T. Ono, H. Suwa, R. Yasuhara, K. Katayama, T. Mikawa and Y. Gohou, "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, Honolulu, HI, USA, 2018, pp. 175-176. DOI: 10.1109/VLSIT.2018.8510676.

[12] X. Sun and S. Yu, "Impact of non-ideal characteristics of resistive synaptic devices on implementing convolutional neural networks," in IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 9, no. 3, pp. 570-579, Sept. 2019. DOI: 10.1109/JETCAS.2019.2933148.

[13] V. Joshi, M. Le Gallo, S. Haefeli, I. Boybat, S. R. Nandakumar, C. Piveteau, M. Dazzi, B. Rajendran, A. Sebastian and E. Eleftheriou, "Accurate deep neural network inference using computational phase-change memory," Nature Communications, vol. 11, 2473, 2020. DOI: 10.1038/s41467-020-16108-9.

[14] C. Matsui, K. Toprasertpong, S. Takagi and K. Takeuchi, "Energy-efficient reliable HZO FeFET Computation-in-Memory with local multiply & global accumulate array for source-follower & charge-

sharing voltage sensing," 2021 Symposium on VLSI Technology, Kyoto, Japan, 2021, pp. 1-2.
[15] C. Matsui, E. Kobayashi, K. Toprasertpong, S. Takagi and K. Takeuchi, "Versatile FeFET voltage-sensing analog CiM for fast & small-area hyperdimensional computing," 2022 IEEE International Symposium on Circuits and Systems (ISCAS), Austin, TX, USA, 2022, pp. 3403-3407. DOI: 10.1109/ISCAS48785.2022.9937237.
[16] K. Higuchi, C. Matsui and K. Takeuchi, "Investigation of memory non-ideality impacts on non-volatile memory based Computation-in-Memory AI inference by comprehensive simulation platform," 2022 IEEE Silicon Nanoelectronics Workshop (SNW), Honolulu, HI, USA, 2022, pp. 1-2, doi: 10.1109/SNW56633.2022.9889067.
[17] S. Fukuyama, A. Hayakawa, R. Yasuhara, S. Matsuda, H. Kinoshita and K. Takeuchi, "Comprehensive analysis of data-retention and endurance trade-off of 40nm TaOx-based ReRAM," 2019 IEEE International Reliability Physics Symposium (IRPS), Monterey, CA, USA, 2019, pp. 1-6. DOI: 10.1109/IRPS.2019.8720436.
[18] F. M. Puglisi, A. Padovani, L. Larcher and P. Pavan, "Random telegraph noise: measurement, data analysis, and interpretation," 2017 IEEE 24th International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA), Chengdu, China, 2017, pp. 1-9. DOI: 10.1109/IPFA.2017.8060057.
[19] S. Vecchi, P. Pavan and F. M. Puglisi, "A unified framework to explain random telegraph noise complexity in MOSFETs and RRAMs," 2023 IEEE International Reliability Physics Symposium (IRPS), Monterey, CA, USA, 2023, pp. 1-6. DOI: 10.1109/IRPS48203.2023.10117832.
[20] L. Reganaz, D. Deleruyelle, Q. Rafhay, J. Minguet Lopez, N. Castellani, J. F. Nodin, A. Bricalli, G. Piccolboni, G. Molas and F. Andrieu, "Investigation of resistance fluctuations in ReRAM: physical origin, temporal dependence and impact on memory reliability," 2023 IEEE International Reliability Physics Symposium (IRPS), Monterey, CA, USA, 2023, pp. 1-6. DOI: 10.1109/IRPS48203.2023.10117882.
[21] C. Wang, H. Wu, B. Gao, T. Zhang, Y. Yang and H. Qian, "Conduction mechanism, dynamics and stability in ReRAMs," Microelectronic Engineering, vols. 187-188, pp. 121-133, 2018. DOI: /10.1016/j.mee.2017.11.003.
[22] S. Ambrogio, S. Balatti, V. McCaffrey, D. Wang and D. Ielmini, "Impact of low-frequency noise on read distributions of resistive switching memory (RRAM)," 2014 IEEE International Electron Devices Meeting, San Francisco, CA, USA, 2014, pp. 14.4.1-14.4.4. DOI: 10.1109/IEDM.2014.7047051.
[23] J. Martin-Martinez, J. Diaz, R. Rodriguez, M. Nafria and X. Aymerich, "New weighted time lag method for the analysis of random telegraph signals," in IEEE Electron Device Letters, vol. 35, no. 4, pp. 479-481, April 2014.
[24] G. González-Cordero, M. B. González, F. Jiménez-Molinos, F. Campabada and J. B. Roldán, "New method to analyze random telegraph signals in resistive random access memories," Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena, vol. 37, no. 1, 012203, 2019.
[25] K. Sugawara, H. Shima, M. Takahashi, Y. Naitoh, H. Suga, H. Akinaga, "Low-frequency-noise spectroscopy of TaOx-based resistive switching memory," Adv. Electron. Mater., vol. 8, no. 8,, 2100758, 2022. DOI: 10.1002/aelm.202100758.
[26] G. González-Cordero, M.B. González, M. Zabala, K. Kalam, A. Tamm, F. Jiménez-Molinos, F. Campabadal, J.B. Roldán, "Study of RTN signals in resistive switching devices based on neural networks," Solid-State Electronics, vol. 183, 108034, 2021. DOI: 10.1016/j.sse.2021.108034.
[27] O. P. Das and S. K. Pandey, "Influence of conducting filament dimension on the performance of ReRAM device in the SET state," 2020 IEEE International Symposium on Smart Electronic Systems (iSES), Chennai, India, 2020, pp. 13-16. DOI: 10.1109/iSES50453.2020.00014.
[28] Y. M. Strelniker, S. Havlin, R. Berkovits and A, Frydman, "Resistance distribution in the hopping percolation model," Phys. Rev. E., vol. 72, 016121, July 2005. DOI: 10.1103/PhysRevE.72.016121.
[29] A. Rodriguez-Fernandez, C. Cagli, J. Suñe and E. Miranda, "Switching voltage and time statistics of filamentary conductive paths in HfO2-Based ReRAM devices," in IEEE Electron Device Letters, vol. 39, no. 5, pp. 656-659, May 2018. DOI: 10.1109/LED.2018.2822047.
[30] Z. Wei, R. Yasuhara, K. Katayama, T. Mikawa, T. Ninomiya and S. Muraoka, "Quantitative method for estimating characteristics of conductive filament in ReRAM," 2014 IEEE International Symposium on Circuits and Systems (ISCAS), Melbourne, VIC, Australia, 2014, pp. 842-845. DOI: 10.1109/ISCAS.2014.6865267.
[31] Y. Zhao, P. Huang, Z. Chen, C. Liu, H. Li, B. Chen, W. Ma, F. Zhang, B. Gao, X. Liu and J. Kang, "Modeling and optimization of bilayered TaOx RRAM based on defect evolution and phase transition effects," in IEEE Transactions on Electron Devices, vol. 63, no. 4, pp. 1524-1532, April 2016. DOI: 10.1109/TED.2016.2532470.
[32] H.-F. Zhang, B.-Y. Ning, T.-C. Weng and X.-J. Ning, "Which phase of Ta2O5 being of the largest dielectric constant," Journal of the American Ceramic Society, vol. 104, no. 12, pp. 6413-6423, 2021.

Ayumu Yamada Received the B.S. degree in Electrical Engineering from the University of Tokyo in 2022.
He is now a master course student in Takeuchi Laboratory in the department of Electrical Engineering and Information Systems, the University of Tokyo.
His current research interests include Computation-in-Memory (CiM), emerging non-volatile memories, neuromorphic computing, and Bayesian machine learning.

Zhiyuan Huang Received the B.S. degree in Electrical Engineering from the School of Information and Communication Engineering, the University of Electronic Science and Technology of China in 2021.
He is now a master course student in Takeuchi Laboratory in the department of Electrical Engineering and Information Systems, the University of Tokyo.
His current research interests include Computation-in-Memory (CiM), emerging non-volatile memories, and machine learning.

Naoko Misawa Received the M.S. degree from Imperial College London in 2012. She is currently an academic staff in Takeuchi Laboratory in the department of Electrical Engineering and Information Systems, Graduate School of The University of Tokyo. Her research interests include emerging non-volatile memories, neuromorphic computing, and Vision Transformer.

Chihiro Matsui is currently a Project Associate Professor in the Department of Electronics Engineering and Information Systems, Graduate School of Engineering of The University of Tokyo. Her research interest includes system, circuit, and device co-design with emerging non-volatile memories for enterprise applications. She earned her B.S. and M.S. degrees in Physics from Ochanomizu University, Tokyo, Japan, in 2003 and 2005, respectively, and her Ph.D. degree in Information Security Sciences from Chuo University, Tokyo, Japan, in 2018. She was a Project Assistant Professor of Research and Development Initiative at Chuo University from 2018 to 2020 and a Project Assistant Professor in the Department of Electronics Engineering and Information Systems, Graduate School of Engineering of The University of Tokyo from 2020 to 2023.

Ken Takeuchi is currently a Professor at Department of Electrical Engineering and Information Systems, Graduate School of Engineering of The University of Tokyo. He is now working on data-centric computing such as computation in memory, approximate computing, datacenter scale computing, AI chip design and brain-inspired memory. He received the B.S. and M.S. degrees in Applied Physics and the Ph.D. degree in Electric Engineering from The University of Tokyo in 1991, 1993 and 2006, respectively. In 2003, he also received the M.B.A. degree from Stanford University. Since he joined Toshiba in 1993, he had been leading Toshiba's NAND flash memory circuit design for fourteen years. He was an Associate Professor at Department of Electrical Engineering and Information Systems, Graduate School of Engineering of The University of Tokyo from 2007 till 2012. He was a Professor at Department of Electrical, Electronic and Communication Engineering, Faculty of Science and Engineering of Chuo University from 2012 till 2020. In 2020, he rejoined The University of Tokyo. He designed six world's highest density NAND flash memory products such as 0.7um 16Mbit, 0.4um 64Mbit, 0.25um 256Mbit, 0.16um 1Gbit, 0.13um 2Gbit and 56nm 8Gbit NAND flash memories. He holds 228 patents worldwide including 124 U.S. patents. Especially, with his invention, “multipage cell architecture”, presented at Symposium on VLSI Circuits in 1997, he successfully commercialized world's first multi-level cell NAND flash memory in 2001. He has authored numerous technical papers, one of which won the Takuo Sugano Award for Outstanding Paper at ISSCC 2007. He is serving as the program chair of Asian Solid-State Circuits Conference (A-SSCC) in 2023. He served as the symposium chair/co-chair of Symposium on VLSI Circuits in 2021/2020. He served as the program chair/co-chair of Symposium on VLSI Circuits in 2019/2018. He has also served on the program committee member of International Solid-State Circuits Conference (ISSCC), Custom Integrated Circuits Conference (CICC), Asian Solid-State Circuits Conference (A-SSCC), International Memory Workshop (IMW), International Conference on Solid State Devices and Materials (SSDM) and Non-Volatile Memory Technology Symposium (NVMTS). He served as a tutorial speaker at ISSCC 2008, forum speaker at ISSCC 2015, SSD forum organizer at ISSCC 2009, 3D-LSI forum organizer at ISSCC 2010, Ultra-low voltage LSI forum organizer at ISSCC 2011 and Robust VLSI System forum organizer at ISSCC 2012.