High-Density Multiple Bits-per-Cell 1T4R RRAM Array with Gradual SET/RESET and its Effectiveness for Deep Learning

E.R. Hsieh[1,2,*], M. Giordano[1], B. Hodson[1], A. Levy[1], S.K. Osekowsky[1], R.M. Radway[1], Y.C. Shih[1,3], W. Wan[1], T. F. Wu[1], X. Zheng[1], M. Nelson[6], B.Q. Le[1,4], H.-S.P. Wong[1], S. Mitra[1,5] and S. Wong[1]
[1]Dept. of Electrical Engineering, [5]Dept. of Computer Science, Stanford University, Stanford, CA, USA
[2]Institue of Electronics, National Chiao Tung University, Hsinchu, Taiwan
[3]Dept. of Materials Science and Engineering, National Tsing Hua University, Hsinchu, Taiwan
[4]Dept. of Electrical Engineering, San Jose State University, San Jose, CA, USA, [6]Skywater Technology Foundry, MN, USA
E-mail: *eray0206@gmail.com

Abstract- We present the first demonstration of 1T4R Resistive RAM (RRAM) array storing two bits per RRAM cell. Our HfO_2-based RRAM is built using a logic foundry technology that is fully compatible with the CMOS back-end process. We present a new approach to program RRAM cells using gradual SET/RESET pulses while minimizing disturbances on adjacent cells (belonging to the same 1T4R RRAM structure) – this new approach makes our multiple-bits-per-cell 1T4R RRAM array demonstration possible. We report over 10^6 cycles of endurance and a projected 10-year retention at 120°C. Using measured data from our 2 bits-per-cell 1T4R RRAM array, we analyze multiple deep learning applications and demonstrate high degrees of inference accuracy (within 0.01% of ideal values).

1. INTRODUCTION

Machine learning applications, e.g., deep neural networks (*NNs*), require large amounts of on-chip memory to produce highly accurate results. Resistive RAM (*RRAM*) is a promising memory technology for such applications [1]. NN implementations using on-chip 1T1R RRAM, including those where each 1T1R RRAM cell stores multiple bits, have been demonstrated (e.g., [2, 3]). However, the size of a 1T1R RRAM cell is constrained by the control transistor (1T), which limits the overall RRAM density. 1T4R RRAM (Fig. 1, [4]) overcomes this challenge by effectively reducing the cell size to about $7F^2$ (F is the feature size of the technology). When multiple RRAM layers are stacked (Fig. 1c, [5, 6]), the effective cell size can be further reduced. We demonstrate, for the first time, how to store multiple bits in each cell of such 1T4R RRAM array to further enhance overall on-chip memory density. Our new 1T4R RRAM programming approach makes such demonstration possible. The major contributions of this paper are:

(a) We present the first demonstration of 1T4R RRAM array where each RRAM cell can store 2 bits.

(b) We describe a new approach to 1T4R RRAM programming (FORM, SET, RESET) while minimizing disturbances on neighboring cells (on the same 1T4R structure).

(c) We present, for the first time, a new approach to finely tune the resistance of each RRAM cell using gradual SET/RESET pulses, which plays a crucial role in achieving multiple bits per cell in the RRAM array.

(d) We report good endurance and long retention of 1T4R RRAM cell programmed using our approach.

(e) Our analysis of multiple deep learning applications (using our measured 2 bits-per-cell 1T4R RRAM array data) shows high degrees of inference accuracy (within 0.01% of ideal) and demonstrates the effectiveness of our approach.

2. 1T4R RRAM DETAILS

Our 1T4R RRAM array chip has been manufactured using a 130nm logic CMOS platform with deposition of MIM RRAM including ALD HfO_2 in the BEOL (Fig. 2). The material system of MIM RRAM is fab-friendly and fully compatible with the CMOS backend [7]. A 1-Mbit 1T4R RRAM array (Fig. 3a) and various test structures are designed to characterize the operations. The unit cell of 1T4R (Fig. 3b) is composed of 1 control transistor with its drain terminal connected to the bottom electrodes of 4 parallel RRAM cells. The top electrode of each RRAM is connected to a separate bit-line. The control transistor is a core device. The gate of the control transistor is connected to the word line. The source of the control transistor is connected to a source line which runs parallel with the bit lines. There are separate decoders for the word, bit, and source lines.

3. RESULTS

A. FORMing for 1T4R RRAM Array

The operation of a 1T4R RRAM array is much more complex (vs. a 1T1R array) since interactions between multiple cells (in the same 1T4R structure) must be carefully controlled. Fig. 4 presents our forming approach for 1T4R RRAM array. The conventional approach for 1T1R FORMs a cell (i.e., induces LRS in that cell) and then proceeds to the next cell. However, if this conventional scheme is directly applied to 1T4R (Fig. 4a), the FORMed cell, which is already in the low-resistance state (*LRS*), will experience additional SET (*over-SET*) as another cell inside the same 4R structure is being FORMed. After experiencing multiple over-SETs, the resistance of a FROMed cell may fall to an ultra-low value, and it may not be possible to RESET that cell anymore. Our FORMing scheme in Fig. 4b overcomes this challenge by RESETting a cell (to the high-resistance state or *HRS*) immediately after it is FORMed. Therefore, when a cell is being FORMed, the adjacent FROMed but RESET cells (in the same 1T4R structure) are no longer over-SET. Occasionally, an adjacent RESET cell may be SET accidentally. It is necessary to check and RESET all cells in the 1T4R structure before the next cell is FORMed. Using this strategy, the FORMing yield is 99%, and all FORMed cells can be RESET to HRS (Fig. 5).

B. Gradual SET/RESET

Gradual SET/RESET procedures, where an RRAM cell is finely tuned (through small changes in its resistance values) by applying small changes in voltage amplitudes of SET/RESET pulses, are crucial for storing multiple bits in each RRAM cell. Some prior

publications rely on special fabrication process steps (e.g., [8]) or high-temperature operation (e.g., [9]) for such gradual tuning. Here, we demonstrate, for the first time, procedures to achieve fine tuning of resistance values using gradual SET/RESET pulses for RRAM built using standard process steps and at normal temperature.

Proper FORMing is essential for gradual SET/RESET. Figs 7-9 characterize the gradual change in RRAM conductance (resistance^{-1}) as a function of the *starting LRS* (i.e., the LRS value right after the RRAM is formed). The results in Figs. 7-9 are generated using the procedure in Fig. 6. As shown in Fig. 7, for low starting LRS of 2KΩ, abrupt switching events between LRS and HRS lead to gaps in conductance values thereby preventing fine-tuning for multiple bits per cell. Such a scenario may be more suited for binary storage. When the starting LRS is increased to 5KΩ (Fig. 8), the gaps become narrower, and switching between LRS/HRS becomes more gradual and symmetric. Moreover, the memory window **(**between the highest HRS and the lowest LRS) is also enlarged from 68x (Fig. 7) to 185x (Fig. 8). If the starting LRS is further increased to 10KΩ (Fig. 9), the gaps disappear, but the memory window reduces significantly to only 28X. Hence, the choice of starting LRS is crucial – with appropriate starting LRS, gradual SET/RESET for multi-bits is possible.

Fig. 10 provides a possible explanation for the results in Figs. 7-9. We can divide the transition curve of conductance into 4 stages. During FORMing, in Stage 0, a major filament with a wide radius is created in HfO_2. In addition, small filaments extend from the tip of the major filament to the top electrode. During RESET, in Stage 1, the major filament recesses resulting in abrupt drop in conductance. In Stage 2, the small filaments slowly recess resulting in gradual reduction of conductance. During SET, in Stage 3, the major filament re-connects and the conductance jumps. In Stage 4, the small filaments re-connect and the conductance slowly increases. Therefore, if one can manipulate the radius of the major filament in HfO_2 through the starting LRS (after FORMing), it will be possible to finely tune the conductance using the small filaments. If the major filament is too wide and conductive, it will dominate the switching (Fig. 7). If the major filament is too narrow and resistive, the memory window will be reduced (Fig. 9).

With an appropriate starting LRS (Fig. 8), Fig. 11 shows our gradual SET/RESET approach. Instead of programming a target resistance directly from HRS (e.g., in [3]), we reach the target resistance through a fine-grained combination of gradual SET/RESET pulses. Fig 12 shows that by using gradual SET pulses to tune values near LRS, and gradual RESET pulses to tune values near HRS, switching gaps can be avoided – thus, resistance values can be tuned between LRS and HRS in a fine-grained manner. Fig. 13 uses our scheme in Fig. 11 to demonstrate that, at a single-cell level, the conductance can be tuned to 128 values (i.e., 7 bits). Of course, the number of bits per cell achievable at the array level will be much less due to cell-to-cell variations. We apply our Fig. 11 strategy (with appropriate FORMing condition) to demonstrate multiple bits-per-cell operation for a large 1T4R array.

C. 2 Bits-per-Cell 1T4R Array Demonstration

Multiple bits-per-cell operation of 1T4R RRAM requires more precise control (vs. 1T1R) since disturbances between adjacent cells (in the same 1T4R structure) will be more serious. For example, Fig. 14 shows that when a cell in 1T4R is selected to be SET, the other 3 (unselected) cells can experience a small RESET, which can disturb the values stored in those (unselected) cells. Therefore, we need additional compensating SET operations to restore the values in those (unselected) cells. By applying the compensating SETs, the disturbances in the unselected cells can be well alleviated, as shown in the insert of Fig. 14. The few low-voltage compensating SET pulses do not affect the other cells.

Fig. 15 illustrates the statistical result of programming 2 bits (i.e., 4 levels or states) in each cell of our 1T4R array. The 4 levels are distributed evenly in an overall memory window of 285X. Furthermore, to demonstrate the capability of programming arbitrary patterns into the 1T4R array, Fig. 16 shows a checkerboard of 00/01/10/11 patterns programmed on a 1 Kbit (32x32) 1T4R RRAM array with a bit-error-rate (*BER*) of 1.56%. The BER is expected to be further reduced with process control and yield enhancement.

Fig. 17 shows the results of single-cell endurance tests. The 4 levels, 00/01/10/11, can be cycled more than 10^6 times. Fig. 18 shows the projected retention time for the 4 levels using high-temperature accelerated testing. The retention time is determined by the HRS and is projected to be more than 10 years at 120^{o}C.

D. Effectiveness for Deep Learning Inference

We trained multiple NNs (Fig. 19) with 2-bit quantized weights using conventional training approaches (e.g., [10]). These NNs are well-known to achieve high inference accuracy despite 2-bit weights [11]. The various datasets (MNIST, MNIST-fashion, CIFAR-10, STI10), used in Fig. 19 (with an AlexNet CNN architecture), represent various degrees of complexity. A similar approach is also applicable for NNs that require more than 2 bits for weights. Using error probabilities (Fig. 19a) extracted from our measured data in Fig. 16, we simulated the above inference tasks to evaluate the resulting inference accuracies (when errors are induced in the weight storage due to the use of 2 bits-per-cell 1T4R RRAM array). Using a join counts permutation test (null hypothesis: bit errors spatially random), we cannot disprove the null hypothesis with statistical significance [12]. Thus, using i.i.d. (identical, independent distribution) error probabilities across the array, Fig. 19b shows that we achieve inference accuracy (i.e., with error model in Fig. 19b) within 0.01% of ideal (i.e., without error model in Fig. 19b) – this confirms the effectiveness of our 2 bits-per-cell 1T4R RRAM approach for these deep learning applications.

Acknowledgment

This work was supported in part by the DARPA ERI 3DSoC program. E.R. Hsieh is supported in part by the Research of Excellence program MOST107-2633-E-009-003, and NCTU-Stanford Dragon-gate program, MOST 107-2911-I- 009-514, Ministry of Science and Technology, Taiwan.

References

[1] H.-S.P. Wong & S. Salahuddin, *Nature Nanotech*, 10, pp. 191-194, 2015.
[2] B. Q. Le, et al., *IEEE Transactions on Electron Devices*, vol. 66, no. 1, pp. 641-646, Jan. 2019.
[3] T. F. Wu, et al., *ISSCC*, pp. 226-228, Feb. 17-21 2019.
[4] C.-W. S. Yeh & S. Wong, *JSSC*, vol. 50, pp. 1299-1309, May, 2015.
[5] M. M. Sabry Aly, et al., *Proceedings of the IEEE*, vol. 107, no. 1, pp. 19-48, Jan. 2019.
[6] M. M. Sabry Aly, et al., *IEEE Computer*, vol. 48, no. 12, pp. 24-33, Dec. 2015.
[7] C. C. Chou, et al., *ISSCC*, pp. 478-479, Feb. 5-9, 2017.
[8] W. Wu, et al., *Symp. on VLSI Technology*, pp. T103-T104, June 18-22, 2018.
[9] C. Walczyk, et al., *Trans. on Electron. Dev*ices, vol. 58, pp. 3124-3130, Sept. 2011.
[10] F. Li, et al., *arXiv preprint arXiv:1605.04711*, 2016.
[11] I. Hubara, et. al., *Journal of Machine Learning Research*, vol. 18, pp. 1-30, 2018.
[12] A.D. Cliff & J.K. Ord. *Spatial Processes: Models and Applications.* Pion, London, 1981.

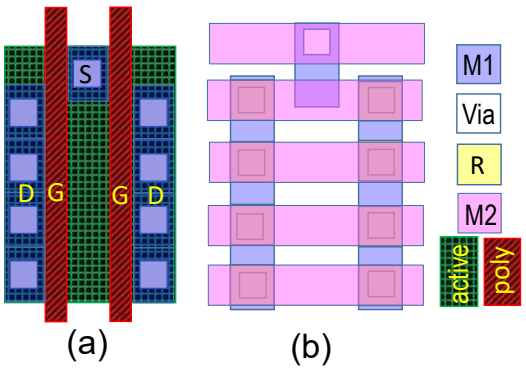
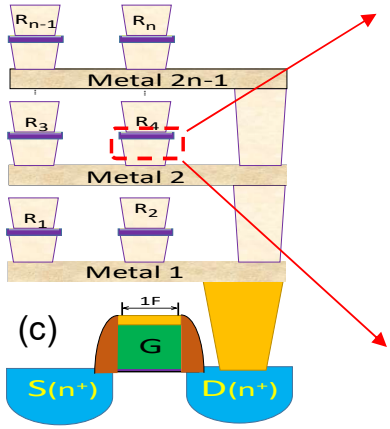


Fig. 1 Layout of 1T4R **(a)** Active, Poly, Contact, and M1; **(b)** M1, Via, RRAM, M2. **(c)** Stacking of multiple RRAM layers to further reduce effective cell size. (BL= Bit Line)



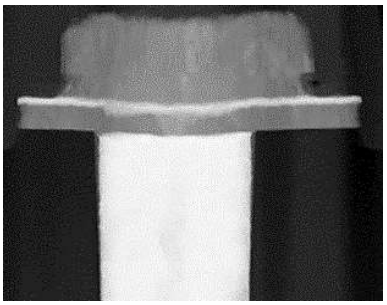


Fig. 2 Cross-sectional TEM of RRAM.

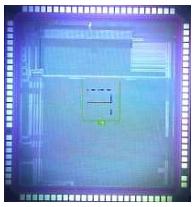


Fig. 3 (a)

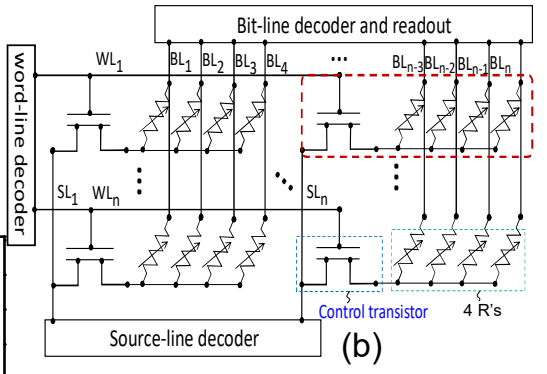


Fig. 3 (a) 1-Mbit 1T4R array. **(b)** The schematic of 1T4R array. BL = Bit Line, WL = Word Line, SL = Source Line

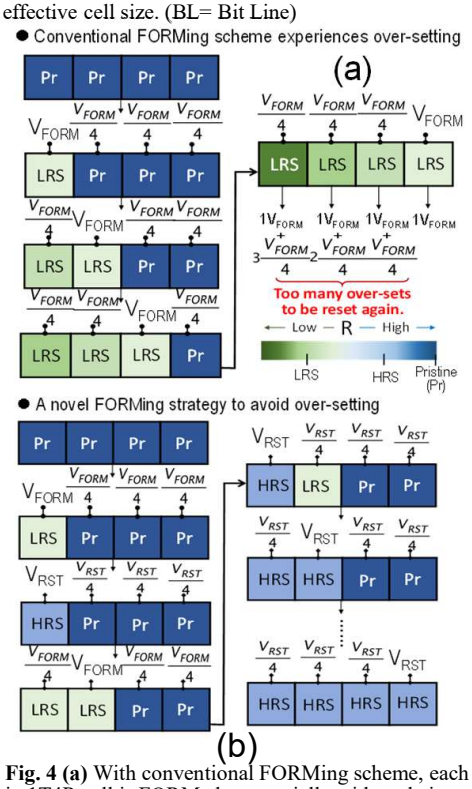


Fig. 4 (a) With conventional FORMing scheme, each R in 1T4R cell is FORMed sequentially without being RESET. The FORMed R experiences over-SET and cannot not be RESET. **(b)** In our FORMing scheme for 1T4R, each FORMed R is immediately RESET before FORMing the next R, to prevent FORMed R's from being over-SET.

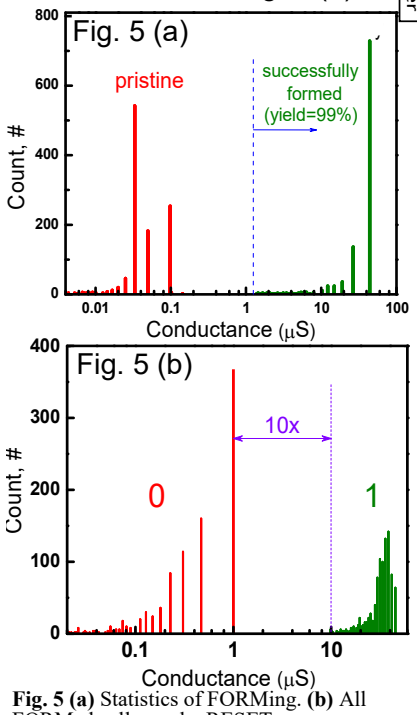


Fig. 5 (a) Statistics of FORMing. **(b)** All FORMed cells can be RESET.

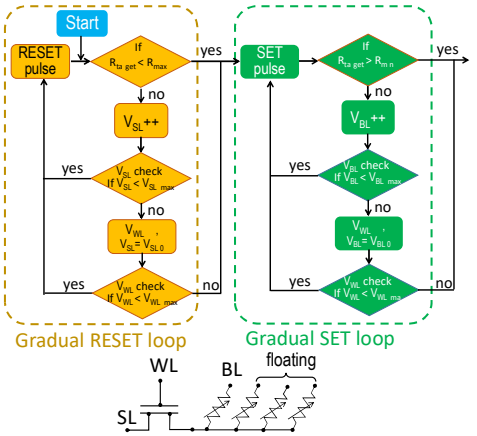


Fig. 6 Procedure to generate data in Figs. 7-9. VSL = source line voltage, VBL = bit line voltage, ++ = slight increase.

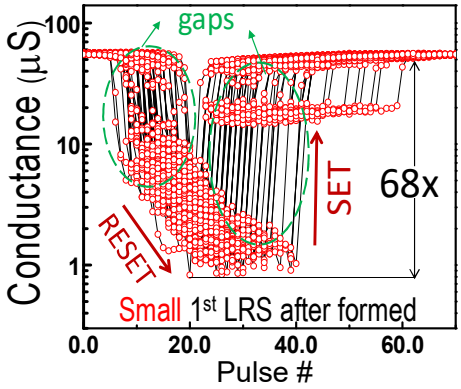


Fig. 7 With low starting LRS of 2KΩ (after FORMing), there exists wide gaps between two resistance states, which cannot support fine tuning using gradual SET/RESET pulses.

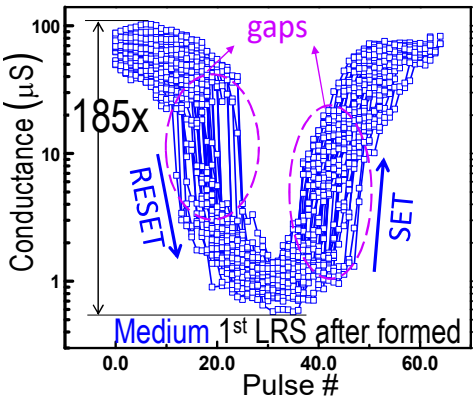


Fig. 8 With intermediate starting LRS of 5KΩ (after FORMing), gaps narrow. This supports fine tuning using gradual SET/RESET pulses. Tuning window expands to 185x.

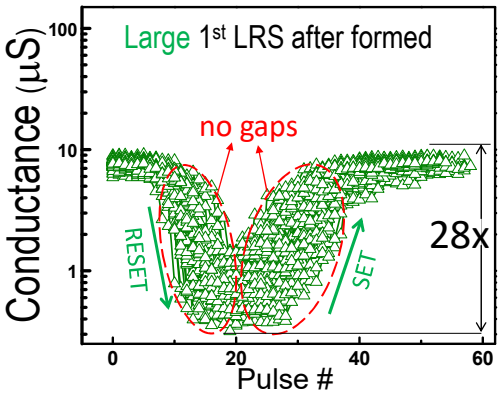


Fig. 9 For high starting LRS of 10KΩ, gaps disappear but the window shrinks to 28x.

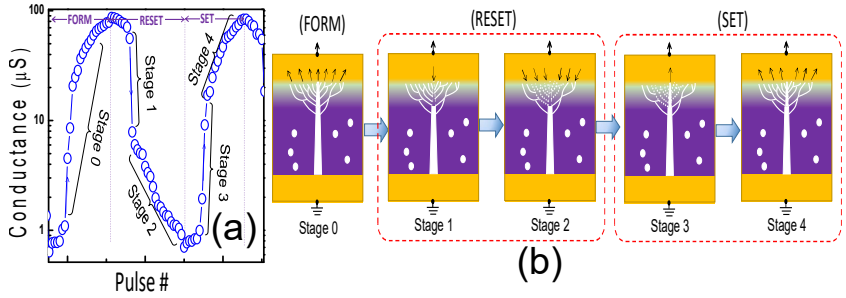


Fig. 10 After FORMing, the transition mechanism during RESET/SET can be divided into 4 stages. During FORMing, in Stage 0, a major filament with a wide radius and small filaments have developed. During RESET, in Stage 1, the major filament recesses and the conductance drops abruptly; in Stage 2, the small filaments slowly recess and the conductance drops gradually. During SET, in Stage 3. the major filament re-connects and the conductance jumps; in Stage 4, the small filaments re-connect and the conductance slowly increases.

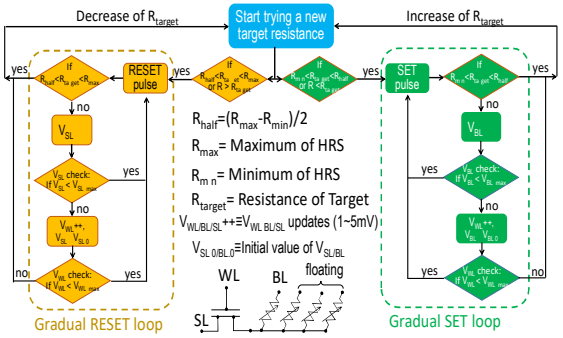


Fig. 11 Tuning by a combination of gradual SET and gradual RESET pulses. When the target resistance is closer to the minimum of LRS, gradual SET is performed. Gradual RESET is performed if the target is closer to the maximum HRS.

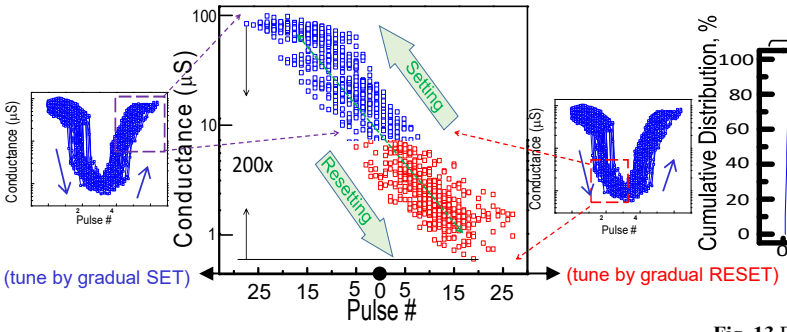


Fig. 12 Based on the algorithm described in Fig. 11, the results show that resistance values can be finely tuned in a very wide memory window.

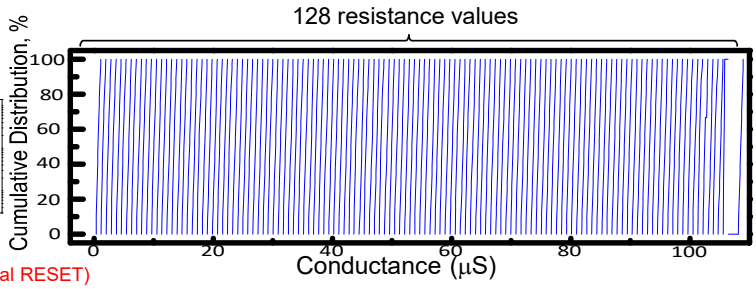


Fig. 13 By using the combination of gradual SET and RESET pulses in Fig. 11 (with an appropriate forming condition), a single cell can be tuned to 128 resistance values in a wide memory window (200x).

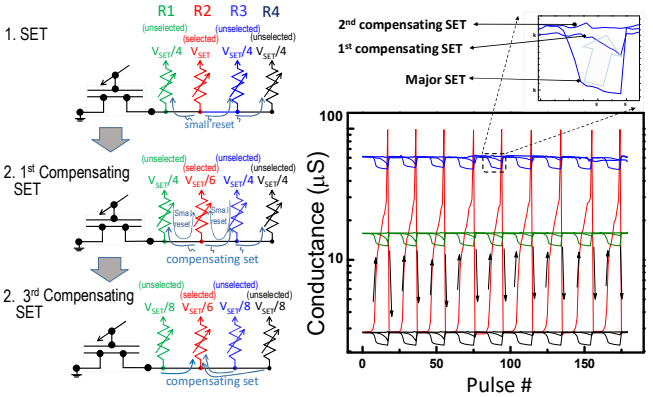


Fig. 14 (a) During SET of a selected cell, the unselected cells can be disturbed by a small RESET. Compensating SET is performed to restore values of the disturbed cells. **(b)** As illustrated in the insert, the conductance of unselected cell is gradually restored to the original level.

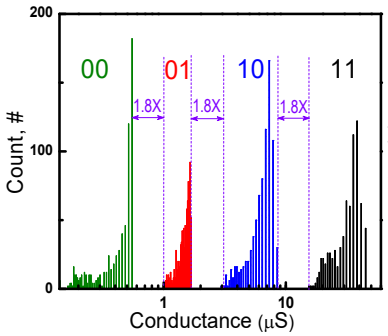


Fig. 15 Distribution of 2-bits-per-cell 1T4R RRAM array (1 Kbit) obtained by using gradual SET/RESET (Fig. 11) and compensating SET (Fig. 14) programming.

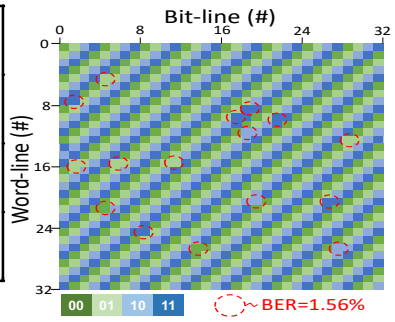


Fig. 16 Checkboard of 00, 01, 10, 11, in a 1Kbit (32x32) 1T4R RRAM array to demonstrate the capability of programming arbitrary patterns.

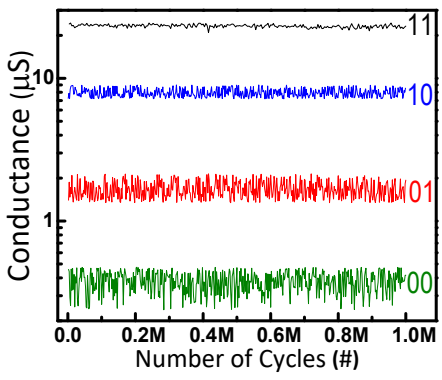


Fig. 17 Endurance of 2-bits-per-cell in 1T4R RRAM array. The 4 levels can be cycled 10^6 times.

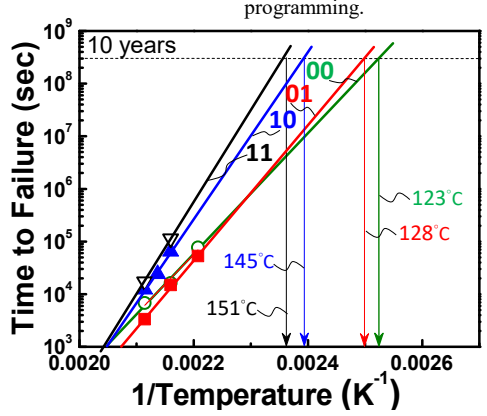


Fig. 18 Data retention of 2bit-per-cell in 1T4R is projected to 10-years at a temperature of about 120°C.

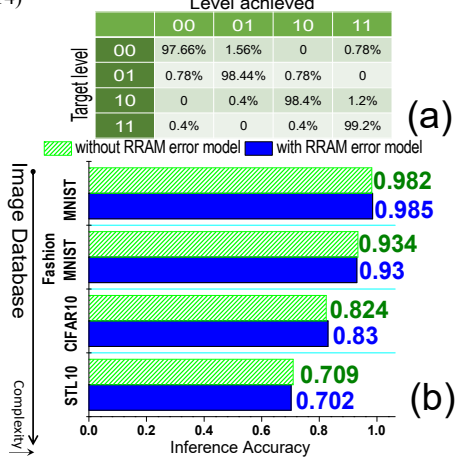


Fig. 19 (a) Error probabilities for a single cell. **(b)** Inference accuracies with and without RRAM errors for various NN applications by using AlexNet.