Ph.D. DISSERTATION

Improvement of RRAM Synaptic Array Using Silicon Nano-Tip Bottom Electrodes and Weight Rearrangement for Neuromorphic Applications

실리콘 나노팁 하부전극과 가중치 재배열을 이용한 뉴로모픽향 저항 변화 메모리 시냅스 어레이의 동작 성능 개선

BY

SUHYUN BANG

August 2022

DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Improvement of RRAM Synaptic Array Using Silicon Nano-Tip Bottom Electrodes and Weight Rearrangement for Neuromorphic Applications

실리콘 나노팁 하부전극과 가중치 재배열을 이용한 뉴로모픽향 저항 변화 메모리 시냅스 어레이의 동작 성능 개선

지도교수 최 우 영

이 논문을 공학박사 학위논문으로 제출함

2022 년 8 월

서울대학교 대학원

전기정보공학부

방 수 현

방수현의 공학박사 학위논문을 인준함

2022 년 8 월

위 원 장 : 김 재 준 (인)

부위원장 : 최 우 영 (인)

위 원 : 이 종 호 (인)

위 원 : 조 성 재 (인)

위 원 : 김 성 준 (인)

Abstract

Resistive-switching random access memory (RRAM) has been considered one of the most promising synaptic device candidates for neuromorphic systems due to its high memory capacity by using simple structures and multi-level storage. In particular, many studies have been conducted on the operation of an array having multiple RRAM synapses in off-chip driving. As research accumulates, large-scale integration and expansion of the number of driving synapses will inevitably occur, and it is necessary to focus on power consumption and wire resistance issues that take place in highly integrated array operation.

In this study, two research directions were suggested. First, for lower power consumption, we fabricated a low-current device by scaling down the switching area. Area reduction is achieved by the anisotropic wet etching of the Si bottom electrodes (BEs). It is observed that the fabricated RRAM with the Si nano-tip BEs show 100~1000x lower current and ~10x lower current than conventional planar RRAM in a single device and array level, respectively. Second, to compensate for

current distortion caused by wire resistance in a large-scale synaptic array, a weight rearrangement method is proposed for the improvement of inference accuracy. The accuracy improvement is evaluated by simulation considering the wire resistance and conductance values of a large-scale RRAM array. In the case of the multi-layer fully connected neural network for pattern recognition, ~8.62% average accuracy improvement at a critical level was achieved by weight rearrangement.

Finally, the RRAM synaptic array operation is demonstrated experimentally to confirm the advantages of the two proposed ideas: silicon nano-tip BEs and weight rearrangement. By using 16 × 2 subarrays, which are parts of a 16 × 16 array, it was confirmed that the inference current becomes lower by the first idea and that the error between the calculated and measured current sum is reduced by the second idea. It is expected that higher inference accuracy achieved in this work will contribute to the implementation of a high-density large-scale RRAM synaptic array.

Keywords: resistive-switching random access memory, neuromorphic system, inference accuracy, synaptic device, Si nano-tip structure, low current, weight rearrangement, wire resistance

Student number: 2015-20931

Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Background

Resistive-switching random-access memory (RRAM) recently has been studied as a potential candidate for synaptic devices for implementing neuromorphic systems because of its behavior similar to the biological synapse and the capability of massive crossbar array configuration [1-6].

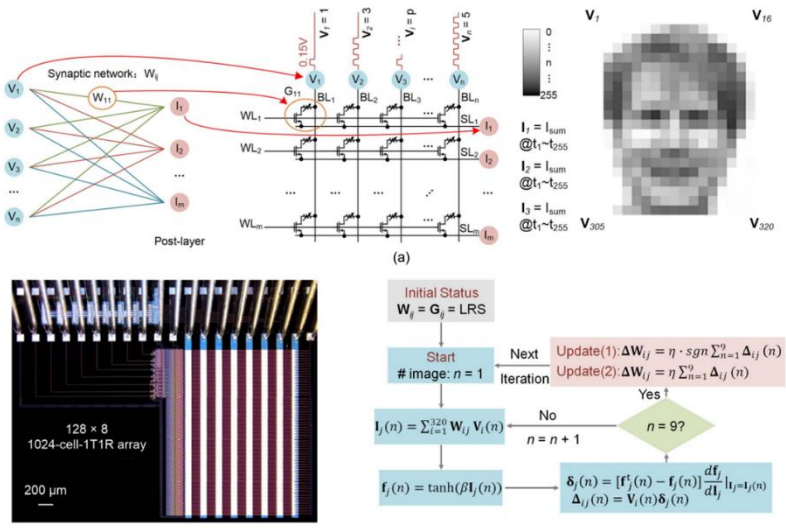

Fig. 1.1. Configuration of the fully-connected neural network and the implementation as RRAM synapse array [2].

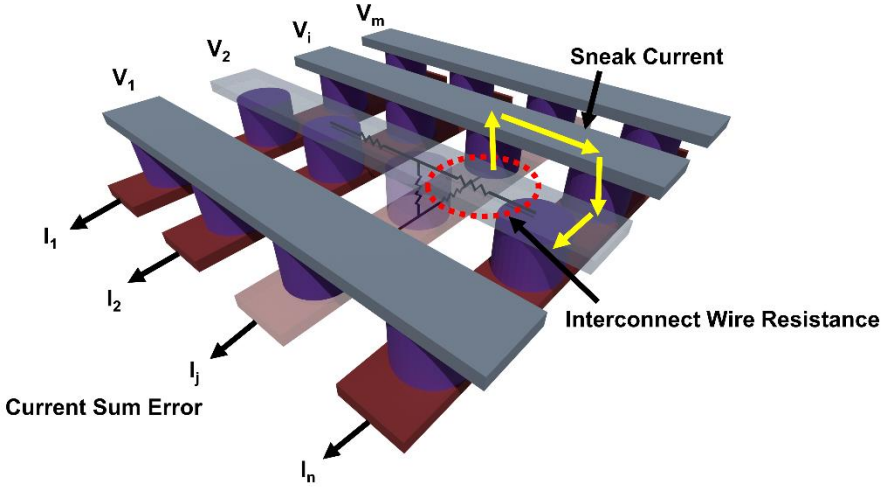

Fig. 1.2. Schematic of the wire resistance effect in RRAM synapse array.

The most widely used method to construct a neuromorphic system using an RRAM synapse is to program the array cells' conductance as the pre-trained weight data obtained from a software-based artificial neural network (ANN) and to perform inference by feed-forward operation of the array with voltage input data, as shown in Fig. 1.2 [7-12]. In general, RRAM synapse is widely used for the exact transfer of trained weight to the selected cell, and there are sufficient achievements in the transfer phase due to the great selectivity of off-state transistors of unselected cells. However, usually, in the inference phase, all transistors are in on-state for performing parallel vector-matrix multiplication (VMM), resulting in the inevitable sneak current problem of large-scale resistor network as shown in Fig. 1.2 [7]. To solve this problem, two directions can be proposed. One is to fabricate a low-current RRAM with greatly increased resistance, and the other is to propose an inference scheme that is less affected by sneak current.

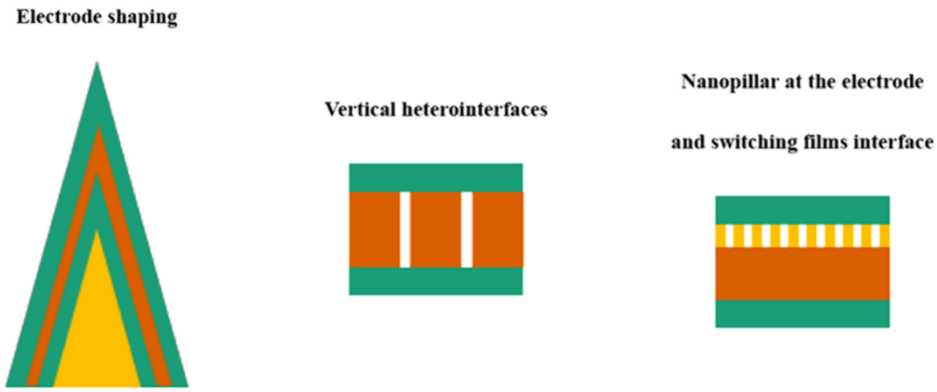

Fig. 1.3. Schematic of various structural engineering of RRAM [16].

First, there have been various studies to fabricate low-current RRAM. If these attempts are divided into two categories, first, there is material engineering that seeks to reduce the current through various engineering such as the insertion of tunneling barriers after understanding the properties of the switching layer (SL) of RRAM [14-17]. Second, structural engineering to scale the operating area by focusing on the device structure itself rather than the SL material was also attempted as shown in Fig. 1.3 [18-23]. In the case of material engineering, since many materials can be used as RRAM, various studies are

possible, but there are too many variables to consider such as material combination, thickness, stoichiometry, and process conditions. On the other hand, structural engineering has the advantage of being easy to implement because it starts with the simple idea of limiting the area through which current flows, and there is a constraint that the operating current level of the used SL material must have an area dependency.

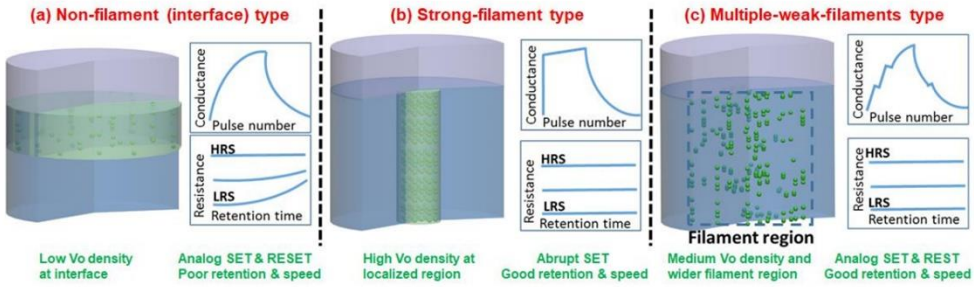

Fig. 1.4. Classification of the filament type in switching layer. (a) non-filament, (b) strong filament, and (c) weak filament type [24].

Several studies have reported that the SL of RRAM is classified

according to its filament characteristics: a strong filament type and a weak filament type [24-30] as shown in Fig. 1.4. In the case of strong filament type, the area dependency of the current level is relatively small because one strong filament is formed and involved in the SET and RESET operations, but the weak filament operates in such a way that several weak filaments are formed and vanished, showing area dependency of the operating current level [28-30]. In particular, the weak filament type device is suitable for use as a synaptic device because there are many previous reports of elaborate gradual switching [24, 27-30]. Therefore, if an appropriate SL material of weak filament type is found and the operating area is scaled, it is expected that low current characteristics suitable for synaptic array operation can be obtained.

Fabricating a low-current RRAM is an intuitive but trivial solution to reduce degradation by wire resistance in array operation. The second solution mentioned above is to propose a VMM scheme in consideration of this wire resistance-induced degradation. Many previous works about VMM operation

have focused on the inference degradation due to device variation [13], *I-V* nonlinearity [31-33], faulty devices [34, 35], endurance [36], retention [37], and the noise in read operation [38-40]. These studies have investigated only the effects of the device characteristics, rather than the effects of the wire resistance in the array. There have been studies about degradation by wire resistance [7, 42-44]. However, those eventually analyzed the effect and suggested design guidance, not clearly proposing a solution to directly compensate for the degradation. Furthermore, alternative read schemes considering the wire resistance effects to implement actual hardware array have been proposed: using reduced input images [4, 8, 40], sequential test and calculating equivalent current [8], post-correction of output current using MATLAB [9] and hardware implementation [45]. These methods, however, inevitably cause other problems such as limited application of input data, running time, operation complexity, or additional hardware configuration. Consequently, a simpler solution to reduce the effect of wire resistance should be analyzed.

1.2 Purpose of the research

In this research, we focus on the inference accuracy degradation depending on wire resistance in RRAM synaptic array and propose a two-way solution. That is, the fabrication of low-current RRAM with structural engineering for area scaling, and the proposal of the novel inference scheme for the improvement of degradation: weight rearrangement method.

First, in terms of structural engineering for area scaling, RRAM of the metal-insulator-semiconductor (MIS) structure was used instead of the metal-insulator-metal (MIM) structure that is commonly used in conventional RRAM. There is also a simple scaling method in which the bottom electrode (BE) is exposed by etching the narrowly patterned hole and the switching material is filled therein. To obtain scalability above a certain level in this way, a complex process for nano-patterning is required. On the other hand, when BE is used as Si, a small active area can be easily created by anisotropic etching of Si, and

various methods have been proposed [46-48]. In this study, switching area scaling is performed by anisotropic etching of Si, and the current reduction effect on weak filament type SL is tested.

The second research goal is to suggest an inference scheme that can improve inference degradation and verify it through simulation. Inferencing with a typical crossbar array, intuitively wire resistance is the smaller the better, and synapse cell resistance will be vice versa. However, reducing the wire resistance meets a trade-off in the aspect of device scalability. In this respect, we conduct a non-ideal array simulation with reasonable values of the wire resistance range suggested by previous studies. In addition, a novel weight rearrange method is applied to the simulation and verified to be generally effective for the improvement of the degradation.

Chapter 2
Device Fabrication and Electrical Characteristics

In this chapter, we compared the operation characteristics of a planar RRAM device with Si BE and an RRAM device with a switching area scaling using a Si nano-tip structure. In the case of the bilayer structure of titanium oxide and aluminum oxide, the SL used in this study, weak filament type operation has been reported in the previous work [28]. Therefore, it can be seen that there will be an effect of switching area scaling on current level reduction, and the verifying experiments and measurements were carried out.

2.1. Fabrication of the single device

2.1.1. Fabrication of the Si BE planar RRAM

We fabricated a single cell RRAM with a planar bottom electrode (BE) structure to confirm the basic operating characteristics of the proposed TiN/TiO_x/AlO_y/p^+-Si RRAM and optimize the SL.

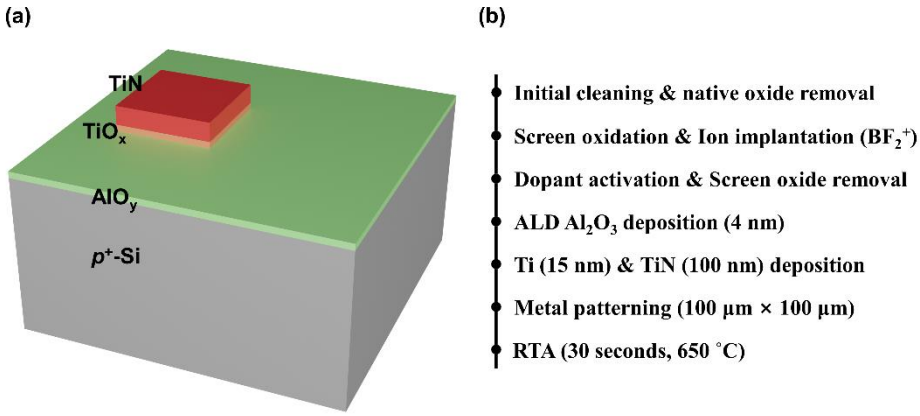

Fig. 2.1. (a) Schematic of the fabricated Si BE planar RRAM structure, and (b) the process flow.

Figure 2.1(a) shows the structure of the proposed planar RRAM device with a Si bottom electrode (BE). The initial device stack consists of TiN, Ti, Al_2O_3, and p^+-Si in order from the top. The process flow of the device is as follows (Fig. 2.1(b)). Initial wafer cleaning was performed including RCA cleaning and native oxide removal using 1% diluted HF solution (DHF). Then, 10-nm-thick screen oxide was deposited by the dry oxidation process, and BF_2^+ ion implantation was performed with 40-keV acceleration energy and $5\times10^{15}cm^{-2}$ dose. After dopant activation with N_2 anneal at 1000 ˚C and removing the screen oxide by BHF solution, atomic layer deposition (ALD) was performed to deposit the Al_2O_3 SL. Target thickness and performed ALD cycles were 40 Å, and 50 cycles respectively. Then, metal layers of Ti, and TiN were deposited sequentially at the thickness of 15 nm, and 100 nm respectively by sputtering. Photolithography was performed over the metal layer with the pad pattern size of 100 μm × 100 μm, and metal dry etching and photoresist ashing and strip were performed as a final process. Additionally, rapid thermal annealing (RTA)

process of 30 seconds under 650 ˚C was performed to optimize the SL.

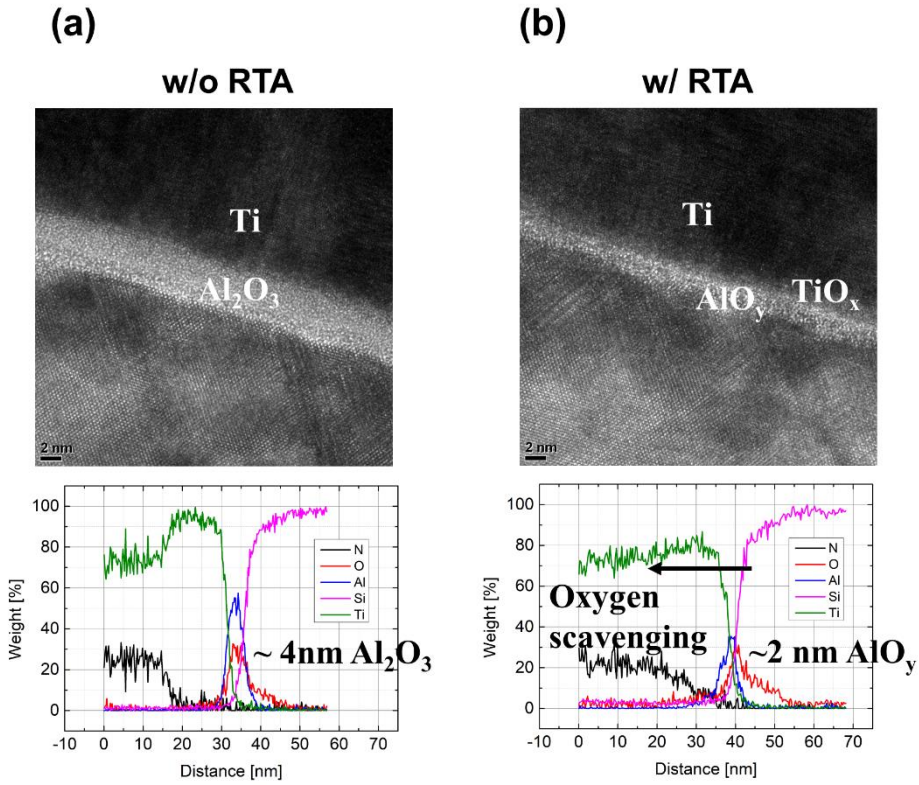

Fig. 2.2. (a) Fabricated SL structure and the EDS analysis without RTA, and (b) optimized SL structure and the EDS analysis with RTA.

Fig. 2.2(a) shows the as-deposited SL structure without RTA. According

to the transmission electron microscope (TEM) image, it is confirmed that the Al_2O_3 layer was formed to a thickness of about 4 nm, and the interface with Ti was formed clear. In this case, the Al_2O_3 SL is too thick, so a hard breakdown occurs during the RRAM switching operation. Consequently, the RESET operation and additional switching do not occur. Fig. 2.2(b) shows the SL structure optimized through the RTA process. The thermal treatment over 425 ˚C causes oxygen scavenging between the Al_2O_3 and Ti interface and the formation of titanium oxide [49]. As a result, the oxygen component contained in Al_2O_3 diffuses toward the Ti layer to form titanium oxide, causing thinning of the Al_2O_3 layer to about 2 nm and the formation of TiO_x/AlO_y bilayer switching material. The optimized structure operates similarly to the previously reported TiO_x/AlO_y SLs [28, 50, 51], and exhibits additional RESET operation and stable resistive switching characteristics.

2.1.2. Fabrication of the Si nano-tip RRAM

To induce the current reduction effect through switching area scaling, we fabricated RRAM with Si nano-tip structure applied to BE. The key idea for fabricating Si nano-tip is an anisotropic wet etching of silicon using a tetramethylammonium hydroxide (TMAH) solution.

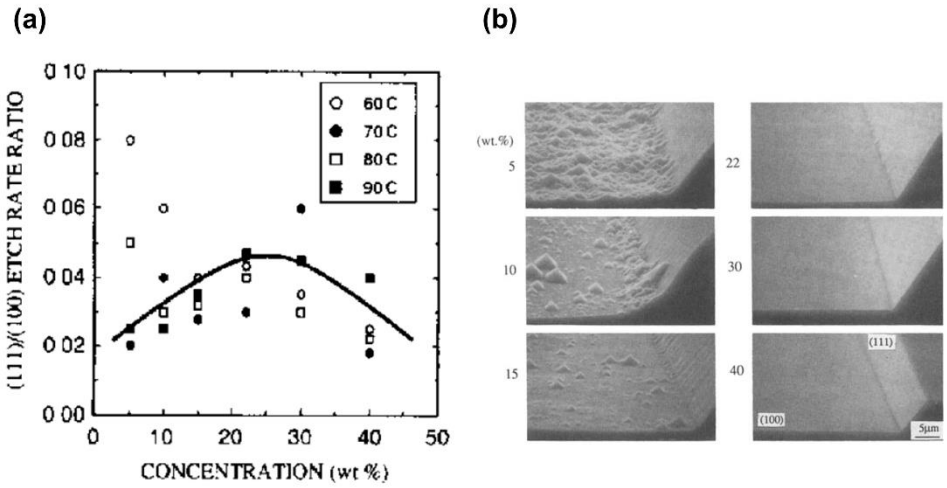

Fig. 2.3. (a) Etch ratio between the Si (111) and (100) plane, and (b) anisotropic etch results of Si under TMAH solution [46].

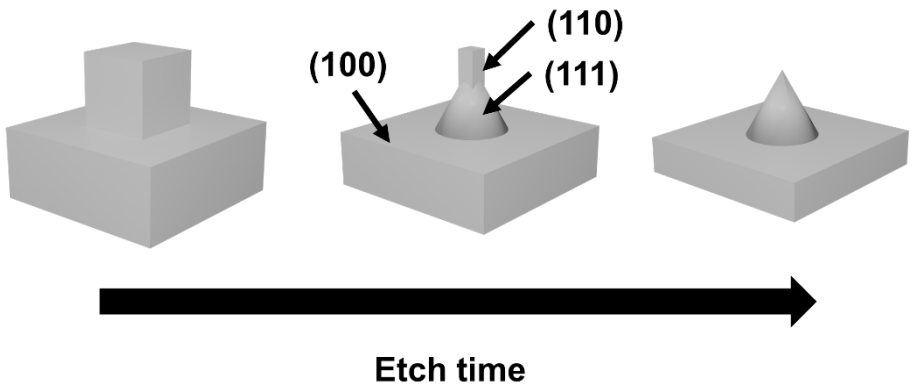

Fig. 2.4. Schematic of Si anisotropic wet etching process to form the Si nano-tip structure.

Fig. 2.3 shows the etch rate ratio and the etch profiles of the crystalline Si in the TMAH solution according to the orientation. In the case of 25% TMAH solution, since the (100) plane has an etch rate 25 times faster than the (111) plane [46], the slower etched (111) plane is revealed as the etch proceeds. Fig. 2.4 shows the process of the Si anisotropic wet etching experiment we conducted. When etching of silicon patterned in a hexahedron is performed

using 25% TMAH solution, the (111) plane is exposed and the (110) plane of the pattern wall is attached and disappears. Finally, the (100) plane and the (111) plane form an angle of 54.7° to form a cone shape with a sharp nano-tip at the top.

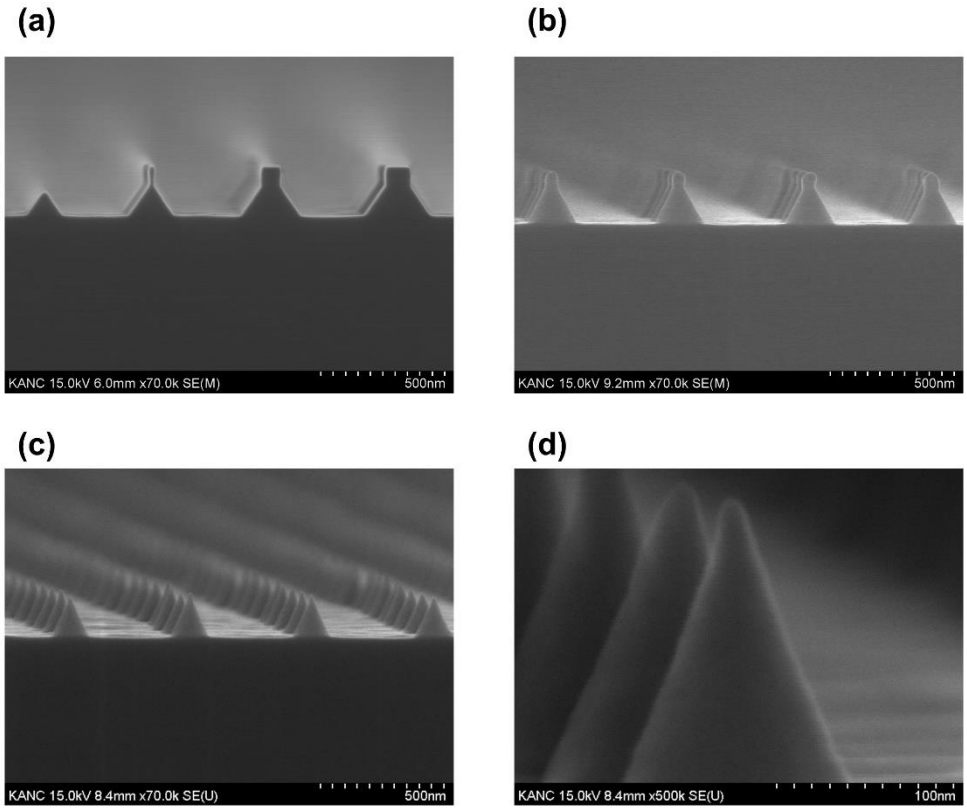

Fig. 2.5. Result SEM images of the Si anisotropic wet etching. (a) Test line

patterns with different start widths are etched forming a wedge structure. (b) Cone structures with insufficient etch time, and (c), (d) optimized etch time.

Figure 2.5 shows the scanning electron microscope (SEM) images of the wet etching test process. To grasp the etch profile according to the etch time, a wedge structure was formed by etching thin test line patterns with different start widths for the same etch time (Fig. 2.5(a)). The etch time optimization is carried out based on the test result, and it was confirmed that the etching time of more than 300 seconds at room temperature with a start pattern width of 300 nm yields an ideal cone shape with a sharp tip as shown in Figs. 2.5(c) and 2.5(d). If the etching time is insufficient, it could be confirmed that the tip was less advanced as shown in Fig. 2.5(b), resulting in a blunt shape. The dimensions of the optimized Si cone shape were confirmed to be a height of 200 nm or less and a width of 300 nm or less.

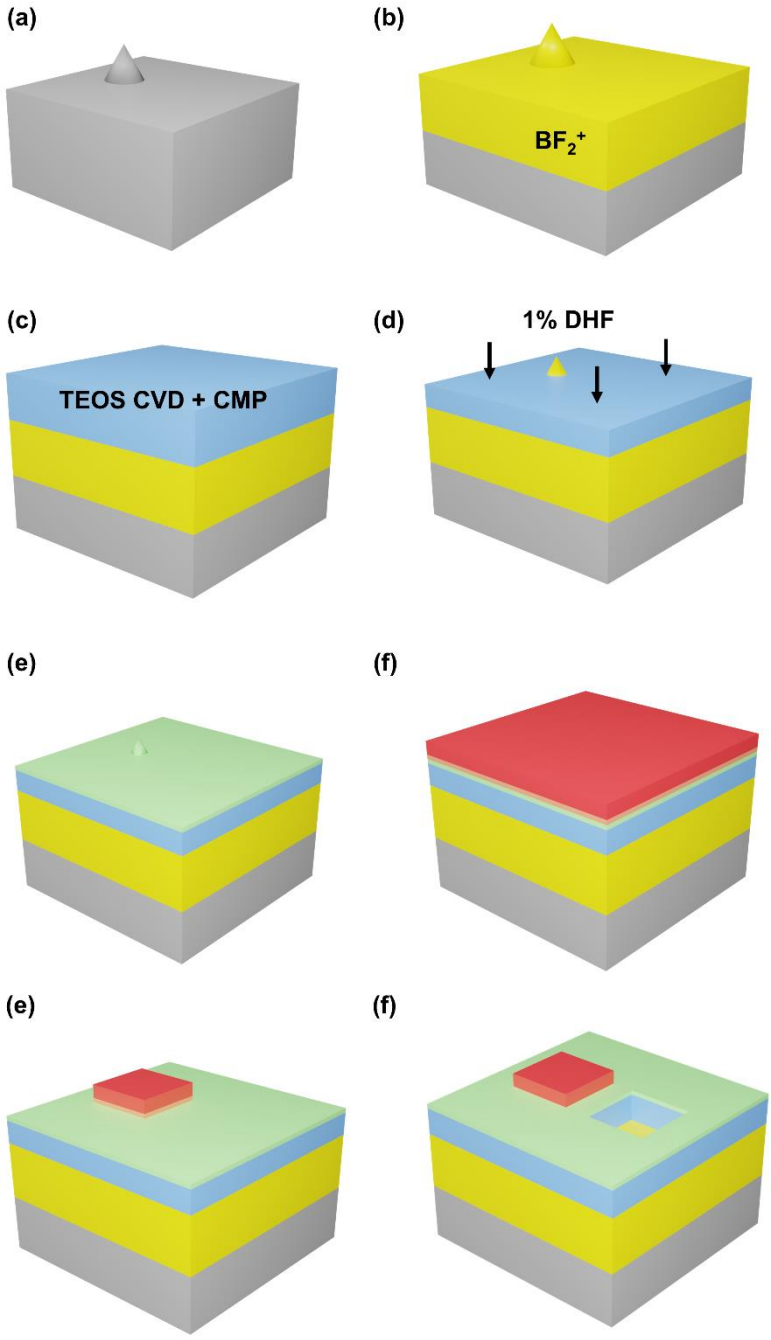

Fig. 2.6. Schematic of the process flow of Si nano-tip single RRAM device.

By applying the optimized Si nano-tip formation process, we fabricated a single cell RRAM device. The stack is applied as the same TiN/TiO_x/AlO_y/p^+-Si structure as the planar device. The fabrication process as shown in Fig. 2.6 is as follows. Initial wafer cleaning was performed the same as the planar device process. The second step was the patterning of 300-nm-width Si hexahedron and the formation of cone shape with TMAH wet etching as shown in Fig. 2.6(a). Then BF_2^+ ion implantation was performed in the same way as the planar device (Fig. 2.6(b)). Plasma-enhanced chemical vapor deposition (PECVD) of tetraethyl orthosilicate (TEOS) SiO_2 of 500-nm-thick over the structure was performed, and chemical-mechanical polishing (CMP) of the TEOS layer proceeded for planarization of SiO_2 (Fig. 2.6(c)). Then using 1% DHF solution, additional SiO_2 trimming was carried out to insulate the rest of the cone shape and expose only the pointed nano-tip at the top as shown in Fig. 2.6(d). Through this process, the actual switching area of RRAM is limited to the exposed nano-tip. As the same as the planar process, deposition of the SL and the top electrode

(TE) metal of 4-nm-thick ALD Al_2O_3 (Fig. 2.6(e)), 15-nm-thick Ti, and 100-nm-thick TiN (Fig. 2.6(f)) was performed sequentially, and metal patterning proceeded (Fig 2.6(g)). After the thermal treatment through RTA at 650 C for 30 seconds, finally, contact patterning was performed to open the Si contact for measurement (Fig. 2.6(h)).

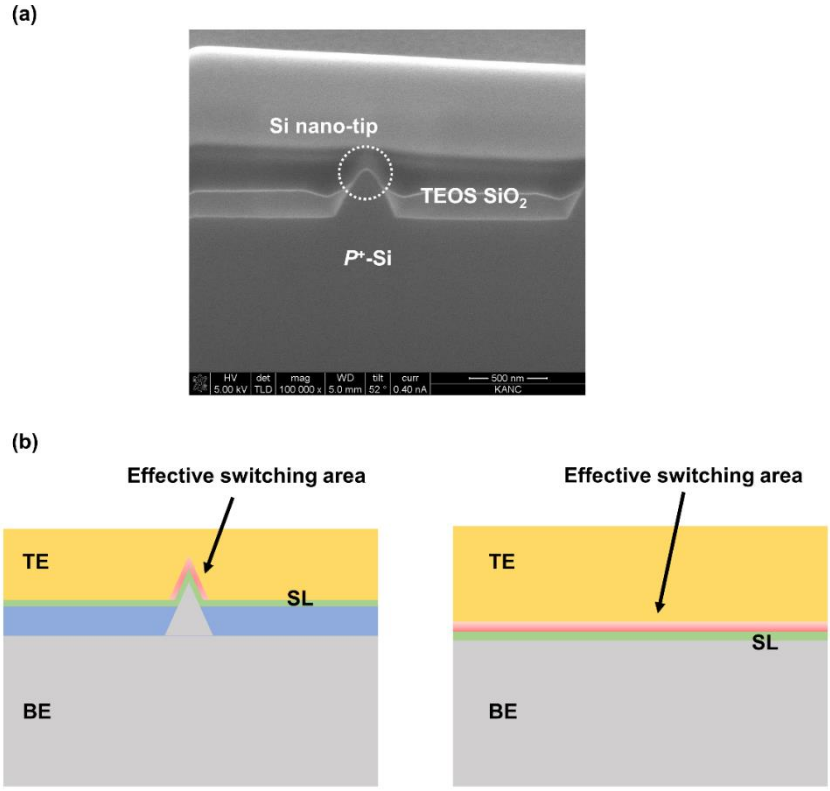

Fig. 2.7. (a) The result SEM image of SiO_2 trimming for the nano-tip exposure,

and (b) schematic of effective switching area reduction of Si nano-tip device compared to the planar device.

Fig. 2.7(a) shows a cross-section of SiO_2 trimmed with 1% DHF solution after the CMP process and embedded cone-shaped Si. The width of the exposed cross-section is about 50 nm, and it is possible to obtain a switching area reduction of more than 10^6 times compared to a planar device with a switching area of 100 μm × 100 μm as depicted in Fig. 2.7(b). This leads to the expectation that effective current reduction can be obtained when applied to the proposed SL stack of titanium oxide and aluminum oxide, which is reported to be the weak filament RRAM.

2.2. Measurement results and discussion of the single device

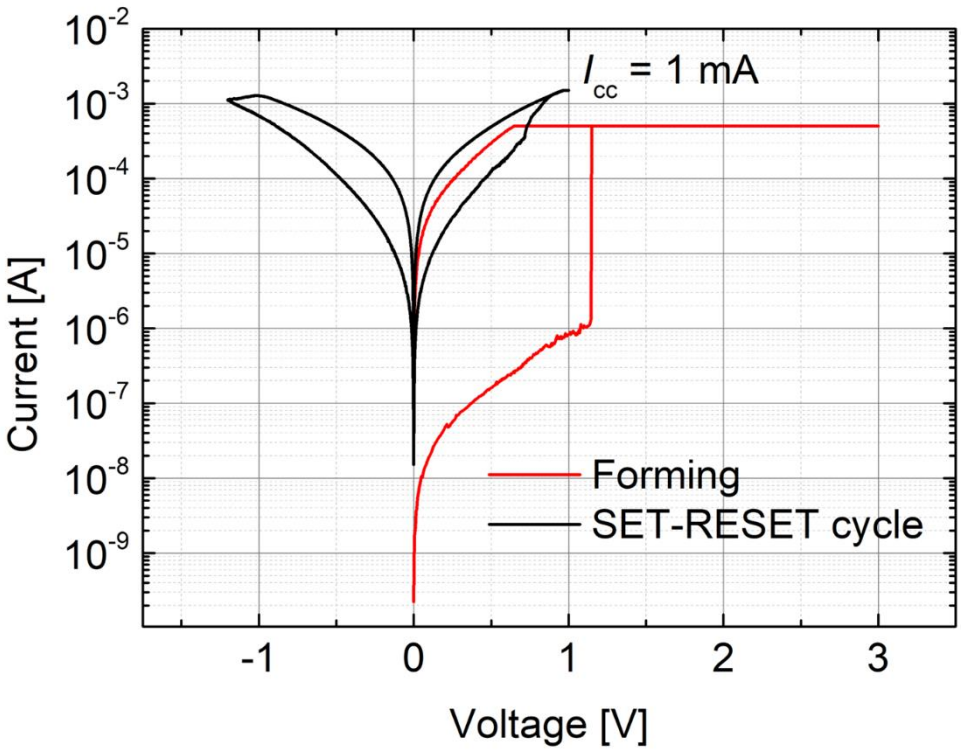

Fig. 2.8. Typical DC *I-V* curve of the planar RRAM.

To compare the fabricated single planar and nano-tip RRAM, basic DC *I-V* measurements were performed. Fig. 2.8 shows a typical DC *I-V* curve of a

planar RRAM. The compliance current required for initial forming was optimized to 500 μA based on the measured cells. In contrast, the compliance current of 1 mA was required for the stable SET-RESET switching cycles after the initial RESET operation. Overall low forming and switching voltage are confirmed. SET voltage was verified to be from 0.8 V to 1.2 V, and RESET voltage was verified to be between -1 V and -1.3 V. The typical operating level of a high resistance state (HRS) current is confirmed to be over 10 μA, and a low resistance state (LRS) current is over 100 μA at a read voltage of 0.2 V.

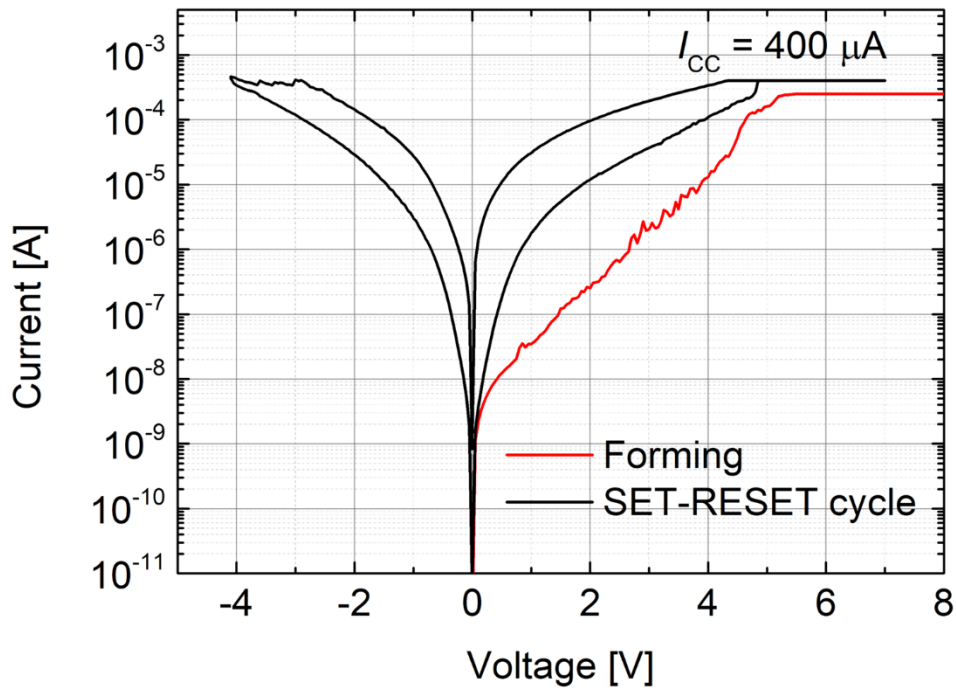

Fig. 2.9. Typical DC *I-V* curve of the nano-tip RRAM.

Fig. 2.9 shows a typical DC *I-V* measurement result of the fabricated nano-tip RRAM. Forming compliance current was optimized at 250 μA, and stable operation was shown at a compliance current of 400 μA in SET-RESET switching after forming. Unlike the case of the planar device, the filament is well formed and operated even at a relatively low compliance current. The SET

and RESET voltages were found to be 4.5 V and -3.5 V, respectively, which were larger than those of the planar device. In the case of a planar device, there is a high probability that there is a relatively weak part within the switching area, and it can be inferred that filaments are formed and operated even at a weak voltage in that part. On the other hand, in the case of the nano-tip device, since the switching area is limited, there is less possibility of a weak part, and it seems that a high voltage over a certain level is required to form a filament. In the case of a nano-tip device, the typical operating level of high resistance state (HRS) current is confirmed to be over 10 nA, and a low resistance state (LRS) current is over 1 μA at a read voltage of 0.2 V.

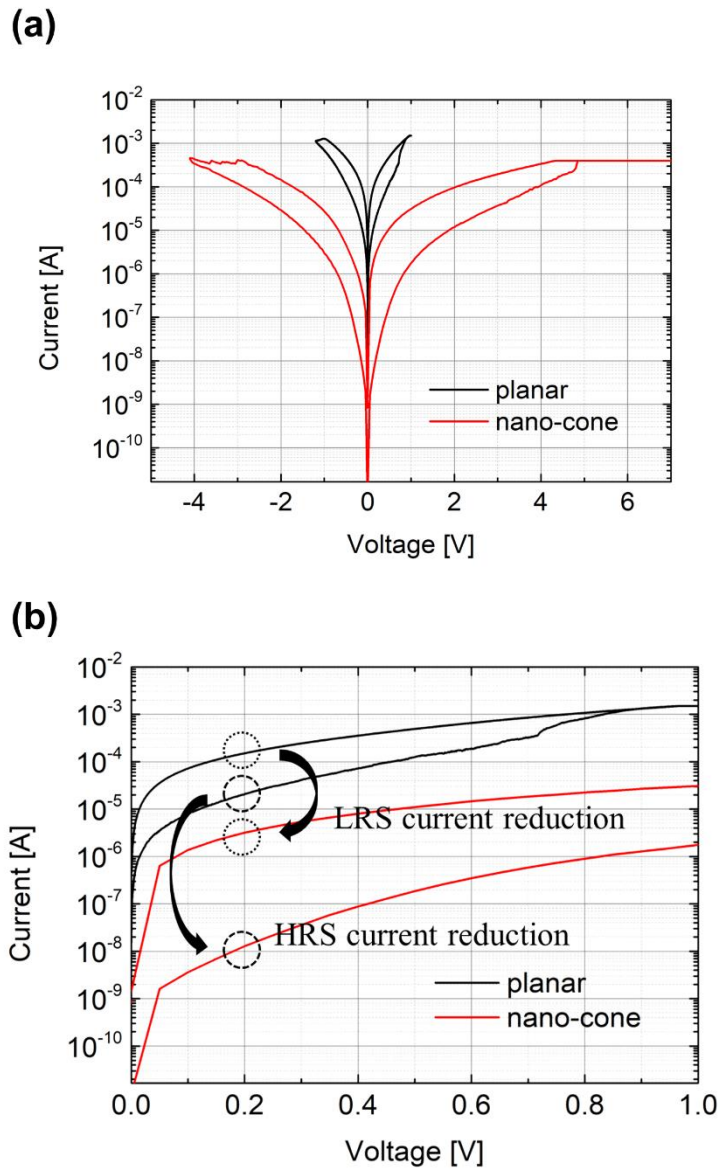

Fig. 2.10. (a) Comparison of typical DC *I-V* curves of the planar and nano-tip device, and (b) comparison of each current level at V_{read} = 0.2 V.

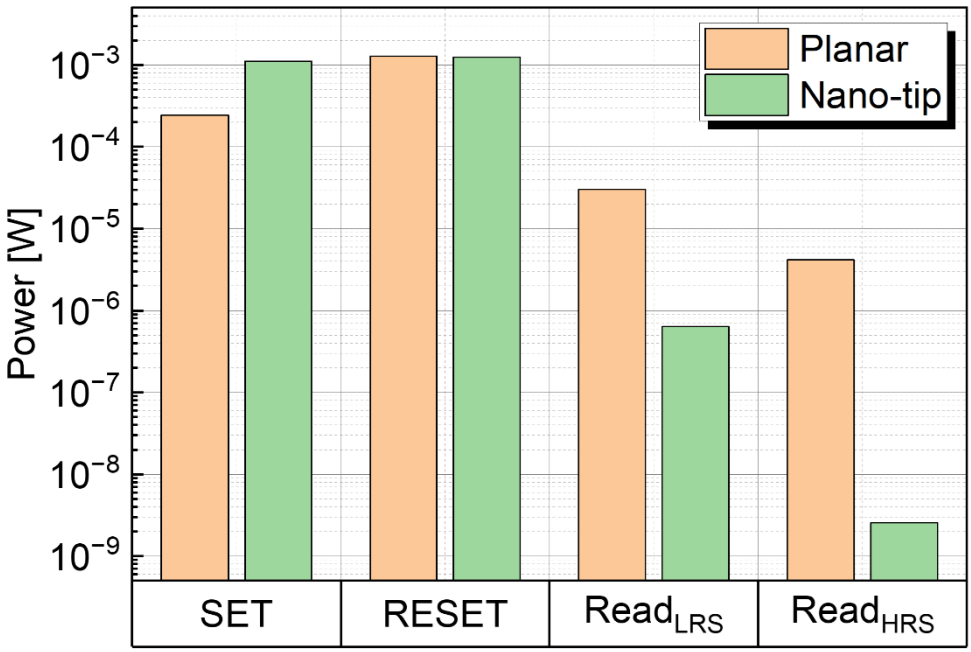

Fig. 2.11. Comparison of power consumption between the planar and nano-tip single device.

As shown in Fig. 2.10(a), it was confirmed that the nano-tip device successfully reduced the current compared to the planar device. According to Fig. 2.10(b), the LRS current decreased by 10^2 times and the HRS current decreased by 10^3 times. Fig. 2.11 shows the comparison of power consumption

between the planar and nano-tip single device, containing SET, RESET switching power and LRS, HRS read power based on DC *I-V* sweep curve. Although the switching voltage of the nano-tip device is increased compared to the planar device, the read operation occurs much more than the switching during the synaptic RRAM operation for driving the neuromorphic system such as data inference. Therefore, in terms of overall power reduction, the net effect due to the decrease in the operating current is greater than the adverse effect due to the increase in the switching voltage.

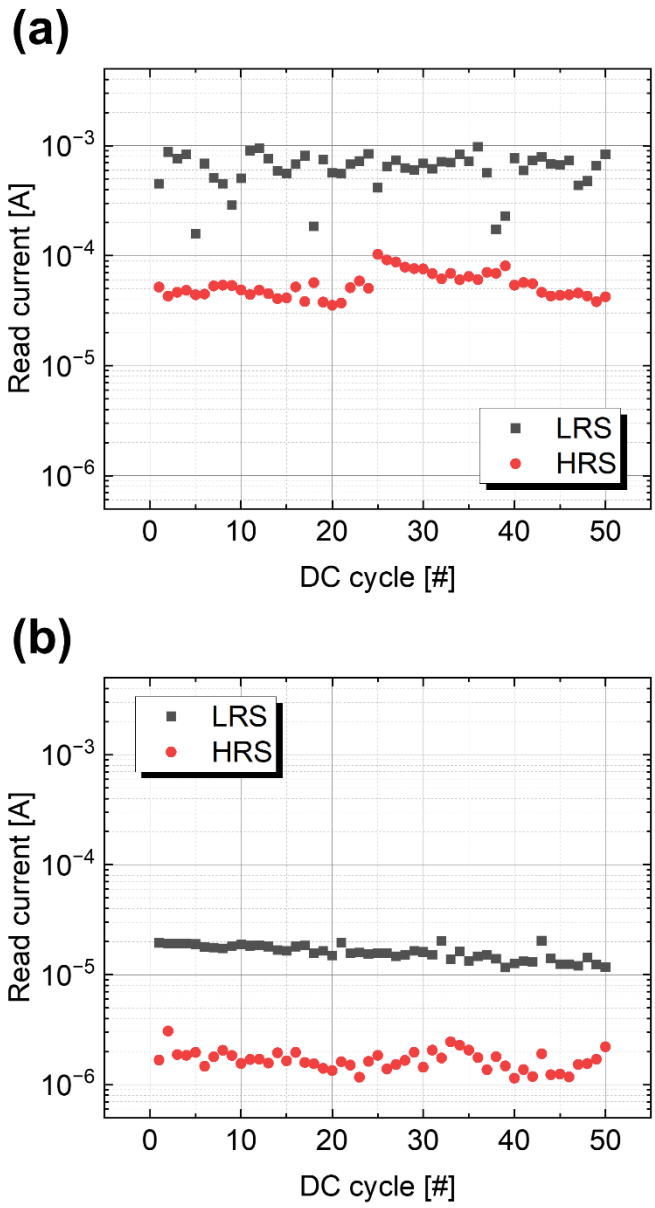

Fig. 2.12. DC on-off 50 cycles operation of (a) planar, and (b) nano-tip device.

Figure 2.12 shows the results of measuring DC cycle-to-cycle endurance and variation of the planar and nano-tip device. The cycle was measured while applying the optimized SET and RESET voltages and compliance current to each device, and the read voltage was 0.2 V. It was confirmed that both devices endured the LRS and HRS current levels well during 50 cycles of DC stress. As confirmed in Fig. 2.12(a), the planar device maintained the LRS and HRS current levels, but the SET operation was not enough in the middle of cycles, causing a variation problem in which the relatively low LRS current was read. As shown in Fig. 2.12(b), it was confirmed that the LRS or HRS current level of the nano-tip device was kept constant during the cycle. What can be inferred from this result is that, in the case of a planar device, the position of the filament caused by the SET operation can be easily changed to another location during the next RESET-SET switching, resulting in the cycle-dependent current level. On the other hand, if the effective switching area is limited to the form of a nano-tip, the possibility that the

filaments are formed in several places is reduced. Therefore, the filament regenerates in the place where it was formed in the previous cycle during the next SET operation after RESET, maintaining a constant operating current.

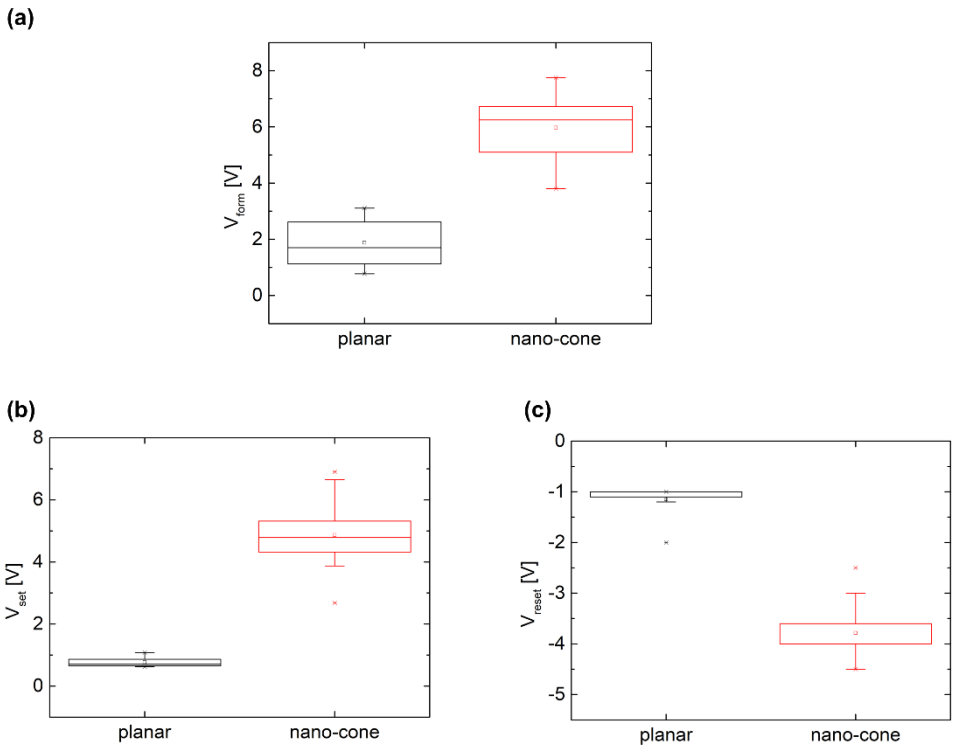

Fig. 2.13. (a) Forming, (b) SET, and (c) RESET voltage distribution of 20 devices of planar and nano-tip RRAM.

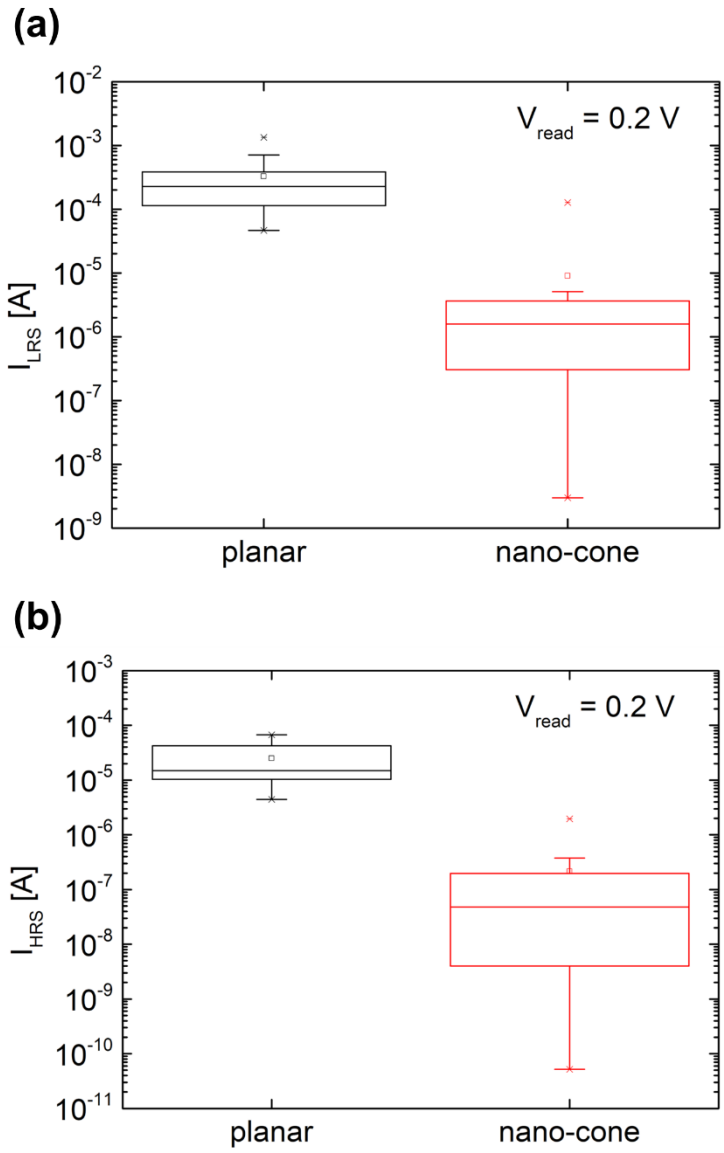

Fig. 2.14. (a) LRS, and (b) HRS current distribution of 20 devices of planar and nano-tip RRAM, at V_{read} = 0.2 V.

Fig. 2.13 shows the distribution of forming, SET, and RESET voltages of 20 planar and nano-tip devices, respectively. As was confirmed by DC measurement, the device-to-device operation voltages also increased in the nano-tip device compared to the planar device. Similarly, in Fig. 2.14, the distribution of LRS and HRS currents measured at a read voltage of 0.2 V for each of 20 devices was confirmed. As verified by DC measurement, the current level was also statistically confirmed to decrease by 10^2 times in the case of LRS and 10^3 times in the case of HRS. It is confirmed that the device-to-device variation of nano-tip RRAM in both voltage and current distribution is slightly large because the process variation is more likely to intervene compared to the planar device. In particular, in the SiO_2 trimming process after CMP, since the variation of CMP is very large, the variation in the SiO_2 thickness after trimming increases, so the variation in the exposed area of the nano-tip also increases. Nevertheless, the current reduction effect is clear so it can be expected that the best performance is obtained when the variation is reduced through the process

optimization.

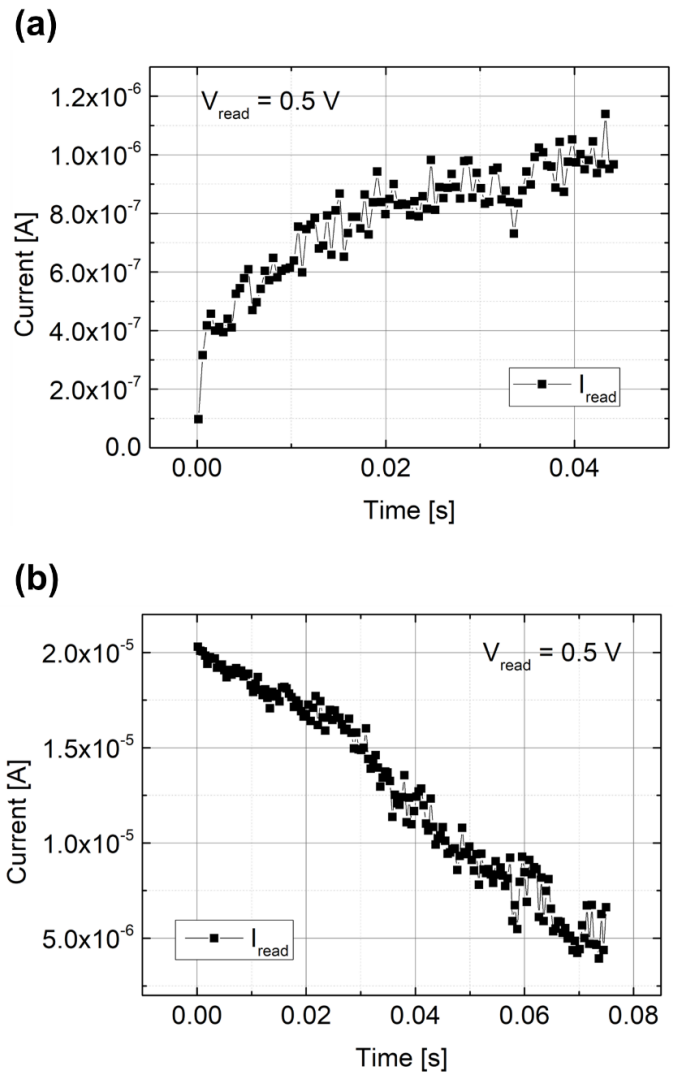

Fig. 2.15. Gradual switching characteristics of nano-tip RRAM device, with ISPP (a) SET, and (b) RESET operation.

Fig. 2.15 shows the gradual switching characteristics of nano-tip RRAM. It was measured by applying the incremental step pulse programming (ISPP) method, and the SET pulse consists of a pulse width of 100 μs, a start voltage of 3.5 V, and a voltage increment of 10 mV. The RESET pulse was constructed with a pulse width of 100 μs, a start voltage of -2.8 V, and a voltage increment of -5 mV. As a result of the measurement, gradual switching characteristics were confirmed in both the SET and the RESET operation, and in particular, more uniform switching was observed in the RESET operation. The gradual switching characteristic of the weak filament type RRAM is made by several small filaments involved in the operation [28]. When the effective switching area becomes too small, the possibility of multiple filaments formation is reduced. Then the gradual switching characteristic is inferred to be weakened. In the case of the nano-tip device, the switching area did not become extremely small to limit the gradual switching characteristics, and it is inferred that the device has the ability to form multiple filaments.

2.3. Fabrication of the Si nano-tip RRAM array

Based on the fabrication, measurement, and optimization of a single device, a synaptic RRAM array with a Si nano-tip BE structure was fabricated. In the case of the SL structure, a TiN/TiO_x/AlO_y/p^+-Si stack was adopted like the fabricated single device.

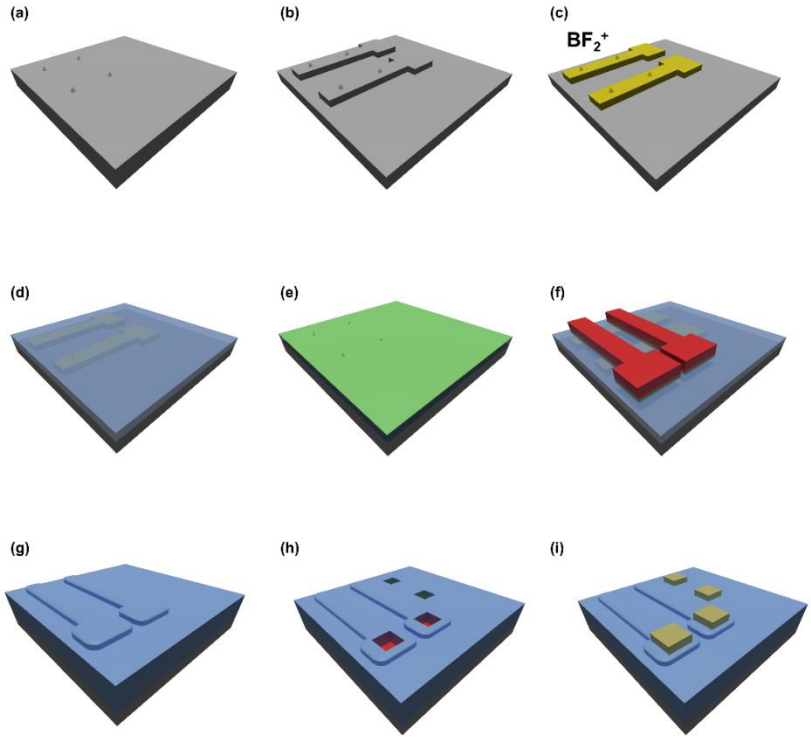

Fig.2.16. Schematic of the process flow of the Si nano-tip RRAM array.

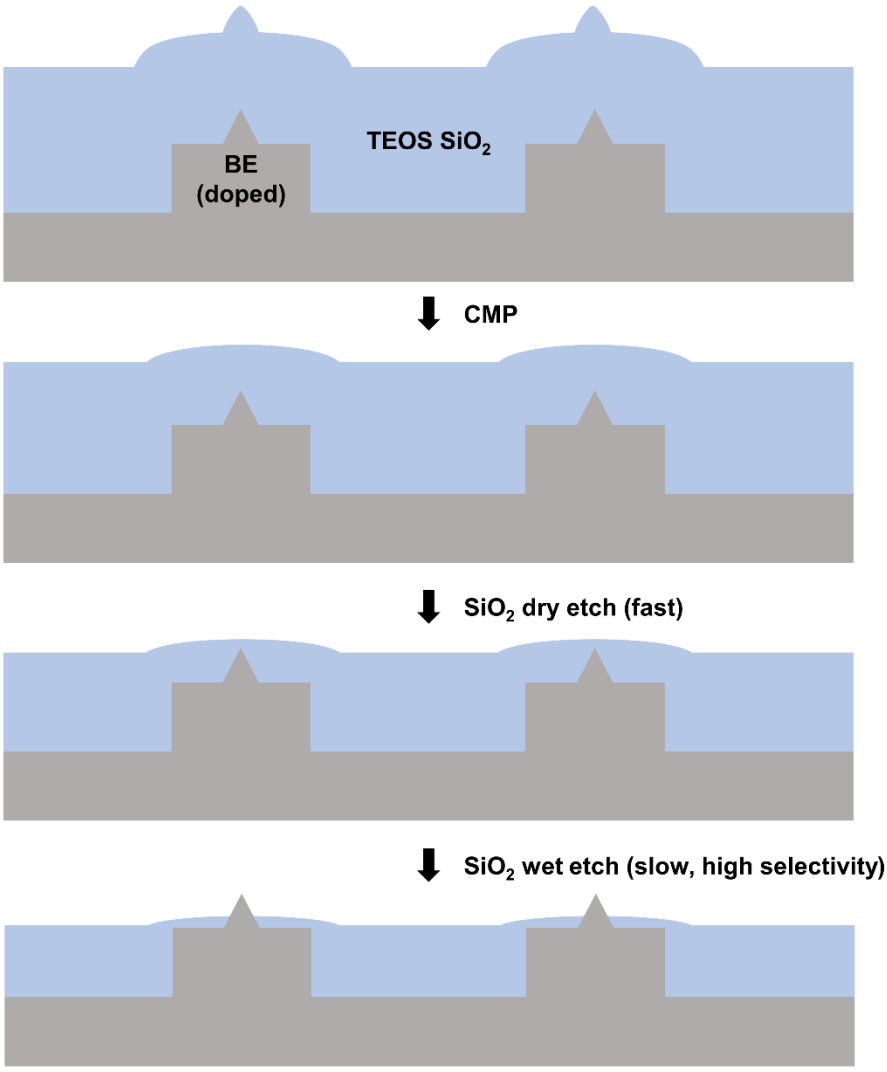

Fig. 2.17. Schematic of the two-step TEOS SiO_2 trimming step after CMP process.

Fig. 2.16 briefly shows the process of fabricating the Si nano-tip array. The process is as follows. Initial wafer cleaning was performed, and the patterning of 300-nm-width Si hexahedron and the formation of cone shape with TMAH wet etching was proceeded as shown in Fig. 2.16(a). Then, BE patterning was performed to be aligned with the cone patterns (Fig. 2.16(b)). BF_2^+ ion implantation with 40 keV energy and 5×10^{15} cm-^{2} dose was performed after the screening oxide and photoresist implant mask was developed. After the screening oxide and implant mask were removed, RTA under 1000 ˚C for 10 seconds to activate dopants was carried out (Fig. 2.16(c)). After 700-nm-thick TEOS deposition and CMP, TEOS SiO_2 trimming step (Fig. 2.16(d)) was divided into two steps consisting of dry and wet etch (Fig. 2.17). Two-step trimming is developed to minimize the CMP time for the protection of cone above BE patterns, and because of too slow remaining oxide etch rate of 1% DHF solution. The SL was deposited as the same as the single device process (Fig. 2.16e)), and sequential Ti-TiN deposition and patterning were performed

(Fig. 2.16(f)). Then, RTA under 650 ˚C for 30 seconds was performed for SL optimization, and inter-layer dielectric (ILD) 300-nm-thick TEOS was deposited to isolate the array from the pad metals (Fig. 2.16(g)). Contact hole patterning with pre-metal cleaning was carried out (Fig. 2.16(h)), and finally, the pad metal of sequential 30-nm-thick Ti, 30-nm-thick TiN, 400-nm-thick Al, and 30-nm-thick TiN was patterned (Fig. 2.16(i)).

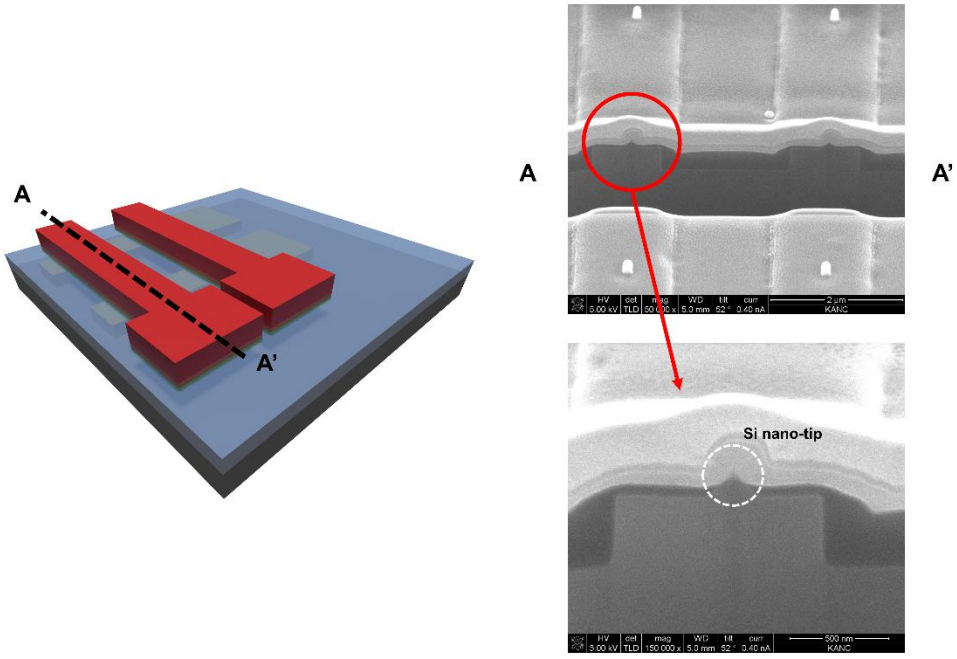

Fig. 2.18. SEM images of the cross-section along the TE of fabricated array.

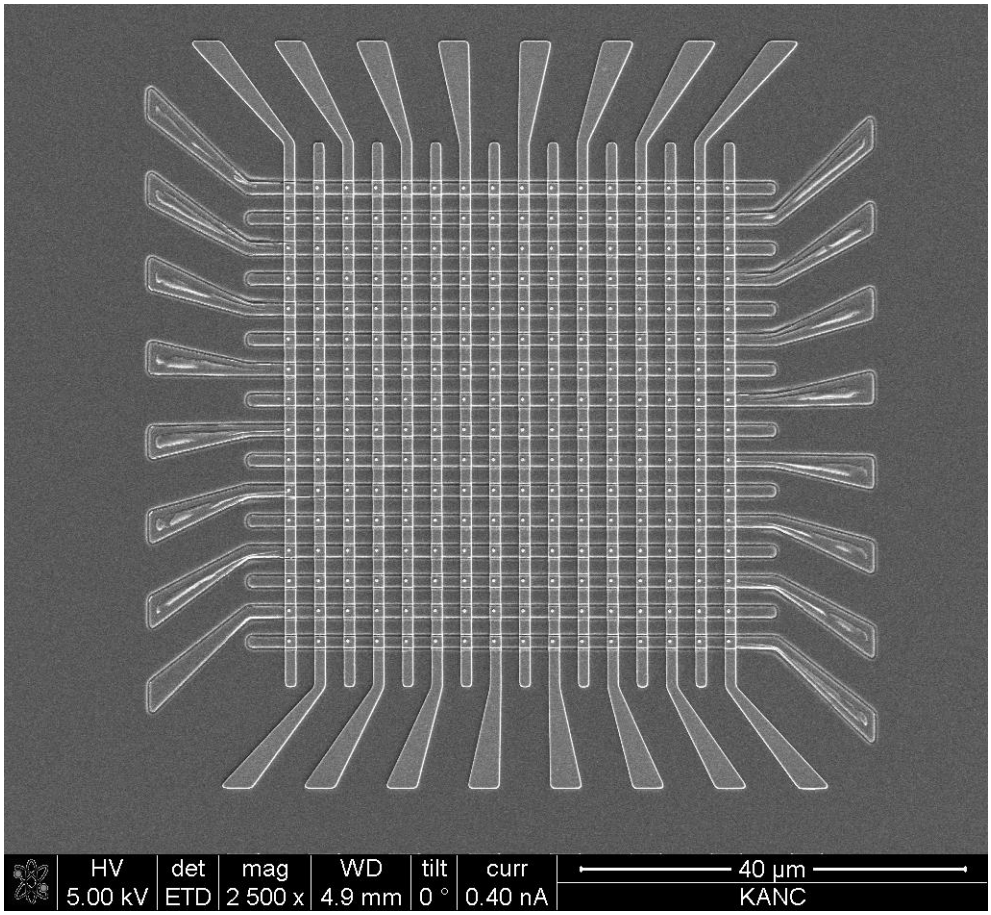

Fig. 2.19. Top view SEM image of the fabricated 16 × 16 nano-tip array before the backend process.

The most important and difficult part of the array process was to protect the cone pattern above the BE from the CMP process and trim the TEOS SiO_2

to an appropriate thickness so that only the sharp vertex of the cone is exposed. Appropriate process conditions were secured through optimized SiO_2 thickness monitoring. As a result of checking the cross-section along the TE, it was confirmed that the cone part was well buried in the surrounding SiO_2 and only the tip was exposed as shown in Fig. 2.18. Fig. 2.19 shows a top view SEM image of a 16 × 16 nano-tip array after TE patterning is completed. Patterned TE and BE equally have 1-μm-width and 2-μm-pitch. Verifying the small dots inside the cross-point where TE and BE intersect, it can be confirmed that the cone pattern and nano-tip were well formed in 256 cells without exception.

A planar array was also fabricated as a control device. Over etch was performed to reveal not only the cone but also the BE during the SiO_2 trimming step. In this case, since the entire area where BE and TE overlap becomes a switching area, it operates as a planar array.

2.4. Measurement results and discussion of the Si nano-tip RRAM array

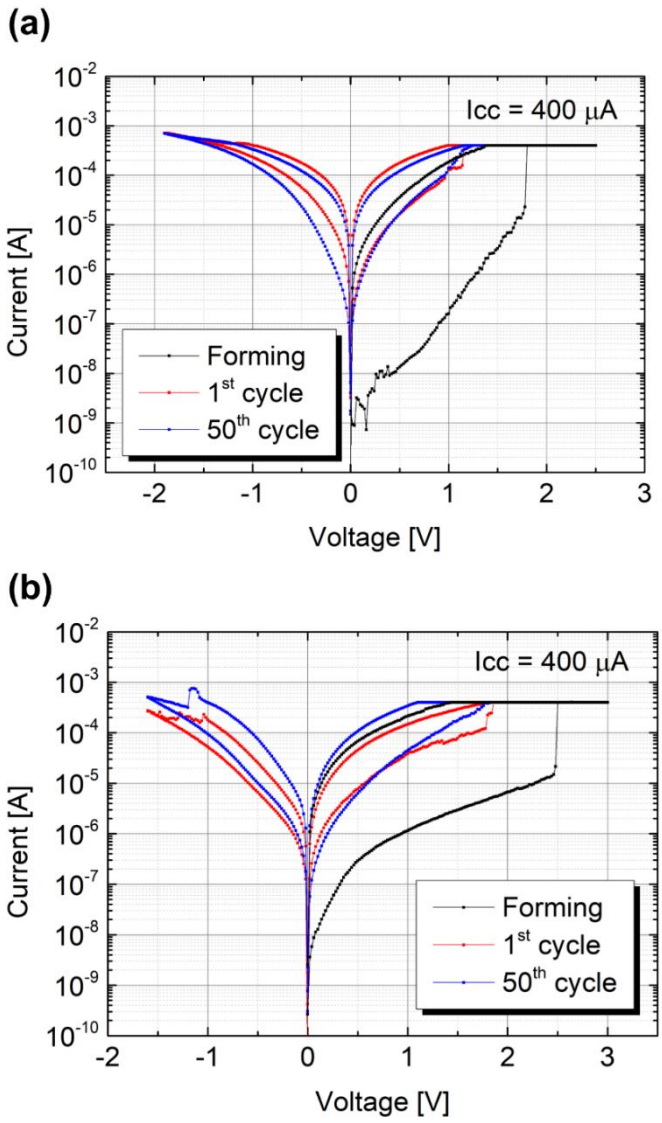

Fig. 2.20. Typical DC I-V curves of (a) planar, and (b) nano-tip array.

Fig. 2.20 shows typical DC operation characteristics of the planar array and nano-tip array, respectively. As measured in the single device, endurance was confirmed through DC stress of 50 cycles. Both devices maintain their operating curves well during stress. Unlike the case of a single device, the compliance current of the planar array and the nano-tip array was not significantly different. In the case of SET voltage, about 1.7 V in the nano-tip array was slightly larger than 1.2 V in the planar array, and there was no significant difference in the RESET voltage. In the case of the current level, both LRS and HRS current showed a difference within 10 times respectively. Compared to the case of the single device, the difference in overall operating characteristics was reduced. This is because the difference in the switching area between the single planar device and the single nano-tip device was theoretically more than 10^6 times, whereas the planar array and the nano-tip array were reduced to theoretically 10^2 times.

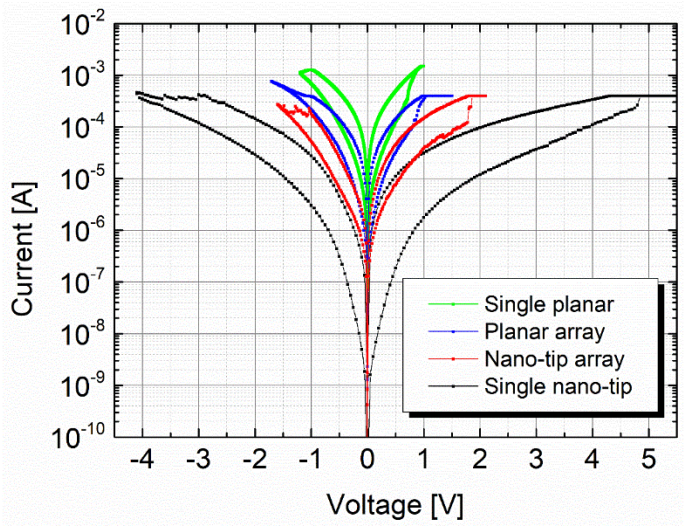

Fig. 2.21. Comparison of DC *I-V* curves of the four different devices.

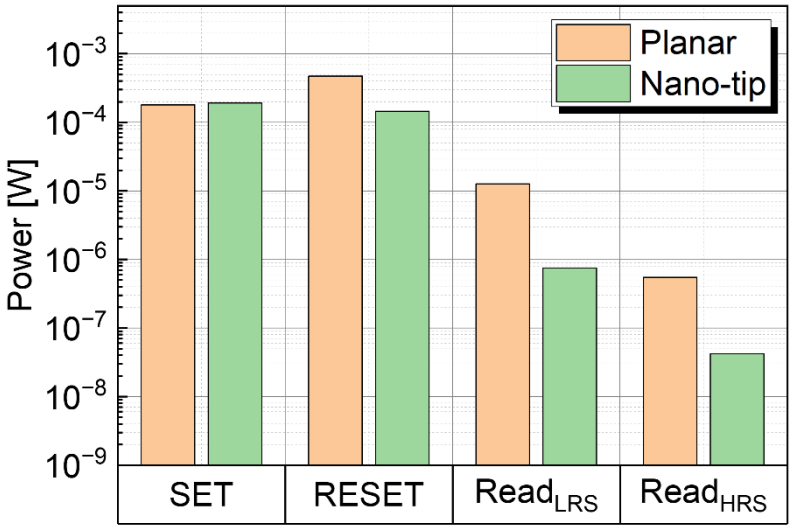

Fig. 2.22. Comparison of power consumption between the planar and nano-tip array device.

Fig. 2.21 shows the comparison of DC switching curves of four types of devices including single devices at once. In the order of single planar, planar array, nano-tip array, and single nano-tip, the decrease in operating current and increase in operating voltage can be confirmed. Fig. 2.22 shows the comparison of power consumption between the planar and nano-tip array device, same as Fig. 2. 11. The nano-tip array has an operating current level larger than that of the single nano-tip but smaller than that of the planar array, so the difference of read power between two array devices is decreased. This is because it is more difficult to control the thickness left of trimmed SiO_2 in the process of nano-tip array than when manufacturing a single nano-tip device. Thus, A larger area of the nano-tip part was exposed in the array than the single device and acted as a larger switching area. If the thickness of SiO_2 above the BE can be accurately monitored, it is possible to expose the nano-tip precisely as much as desired, so performance improvement can be expected.

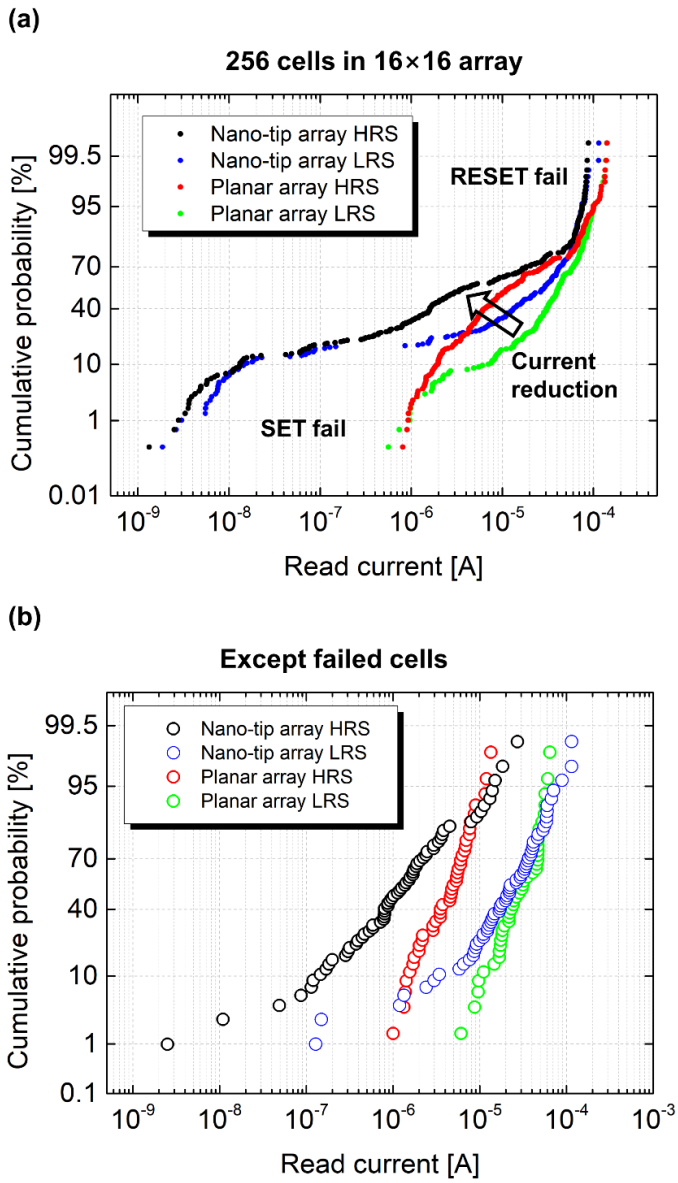

Fig. 2.23. DC on-off read current cumulative probability plots of planar and nano-tip array at V_{read} = 0.2 V for (a) whole cells in 16 × 16 array, and (b) with exception of failure cells.

Fig. 2.23 shows the operating current distribution of the planar array and nano-tip array with 16 × 16 size. From the pristine state, the cells were read one by one while operating one cycle DC sweep under fixed operating voltage for all 256 cells, and LRS, HRS current were measured in each SET-RESET cycle at V_{read} = 0.2 V. As a result, it was statistically confirmed that the operating current decreased in the nano-tip array. Excluding the cells that failed to switch under the fixed DC cycle condition, the LRS current region quite overlaps, but the HRS current region has significantly lower in the nano-tip array.

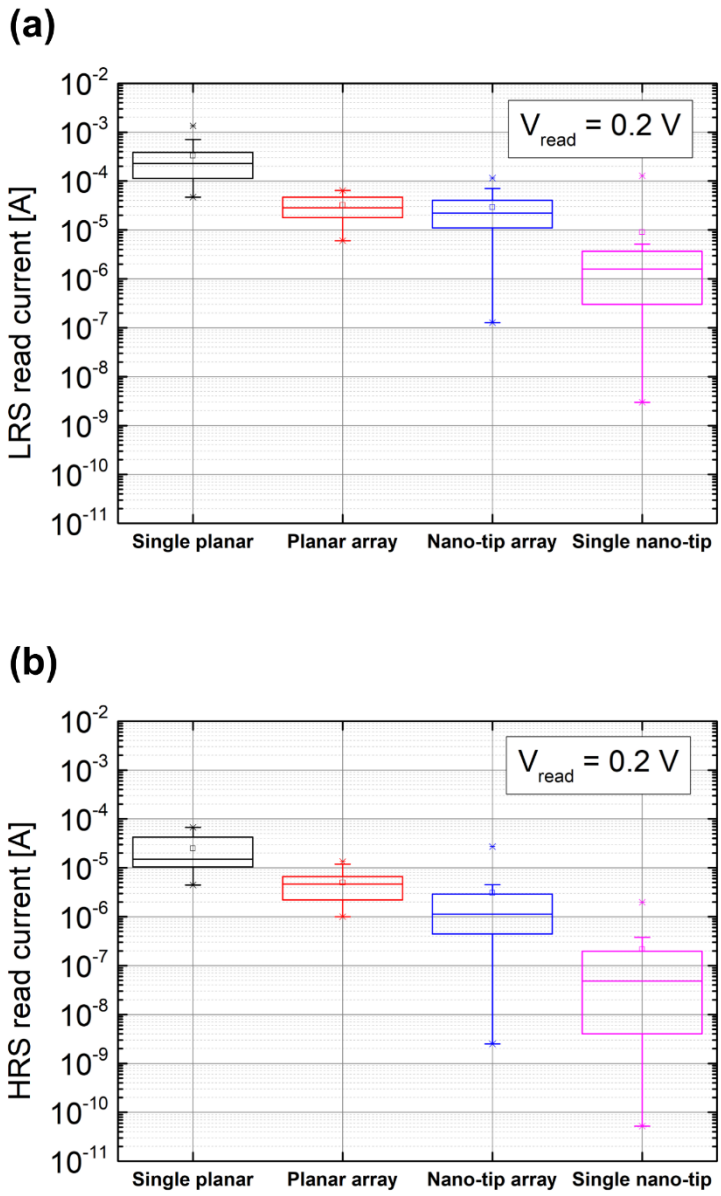

Fig. 2.24. Comparison of operation current distribution of the four different devices. (a) LRS, and (b) HRS read current at V_{read} = 0.2 V.

Fig. 2.24 statistically shows the current distribution of the four manufactured devices. In both LRS current and HRS current, the effect of reducing the current according to the area reduction was confirmed, and it was confirmed that the nano-tip structure was advantageous for the low current operation in both single devices and arrays compared to the planar structure. In the case of the device-to-device variation problem occurring in the nano-tip, as mentioned above, it appears that the process variation caused by the SiO_2 trimming process played a large role.

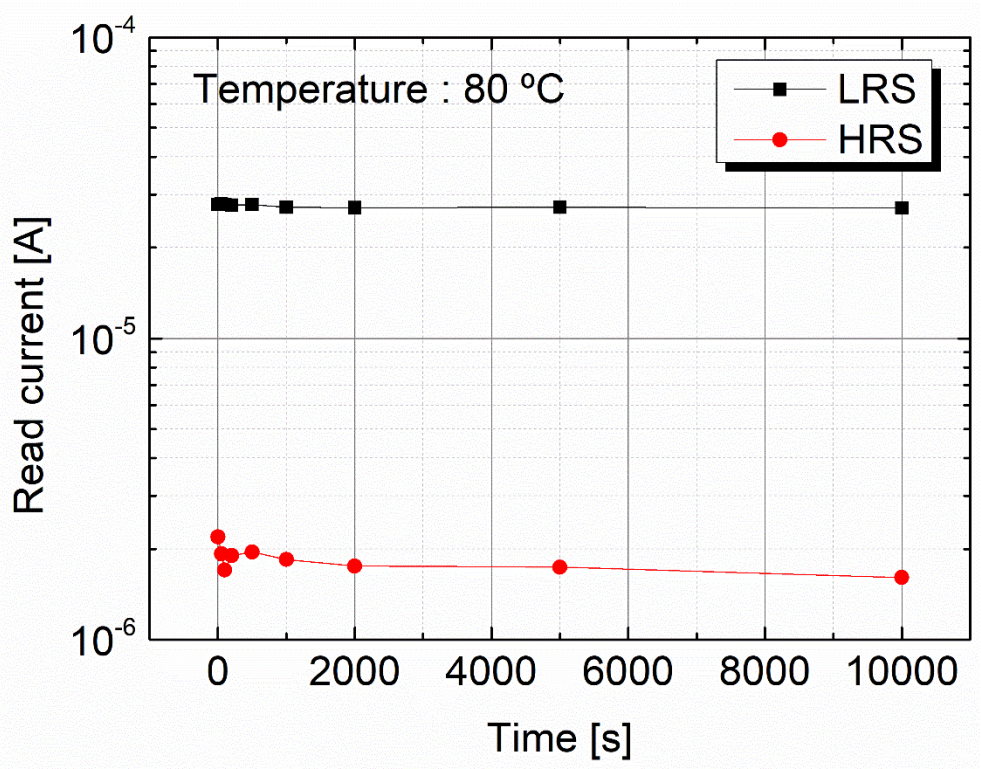

Fig. 2.25. On-off retention of nano-tip array for 10,000 seconds at 80 ˚C, and V_{read} = 0.2 V.

Fig. 2.25 shows the on-off state retention of the nano-tip array. Two cells programmed with LRS and HRS, respectively, were measured for 10,000 seconds at a temperature of 80 ˚C. It was confirmed that the window is not changed during the measurement time and each state was maintained well.

Chapter 3

Neural Network Simulation on the RRAM Array

In the previous chapters, we investigated the structural and electrical properties of RRAM synaptic devices and arrays using Si nano-tip BE. Based on the obtained results, we conducted a simulation to predict the performance of the proposed device when operating a large-scale neural network. Performing the parallel vector-matrix multiplication (VMM) operation for the data inference with a large-scale resistor network, by necessity, inference accuracy degradation may occur by the sneak current by IR drop generated between the nodes inside the array.

In this chapter, we conducted the synapse array simulation utilizing Python scripts and a carefully designed resistor network. In addition, we proposed the novel accuracy boosting software pre-processing method: weight rearrangement. It was confirmed that the weight rearrangement method effectively alleviates inference accuracy degradation caused by wire resistance through a relatively simple algorithm. Through the simulation, we verified the following: the effect of wire resistance inside large-scale synaptic resistor networks, and the effect of the weight rearrangement under various array conditions.

3.1 Configuration of the neural network based on the RRAM array

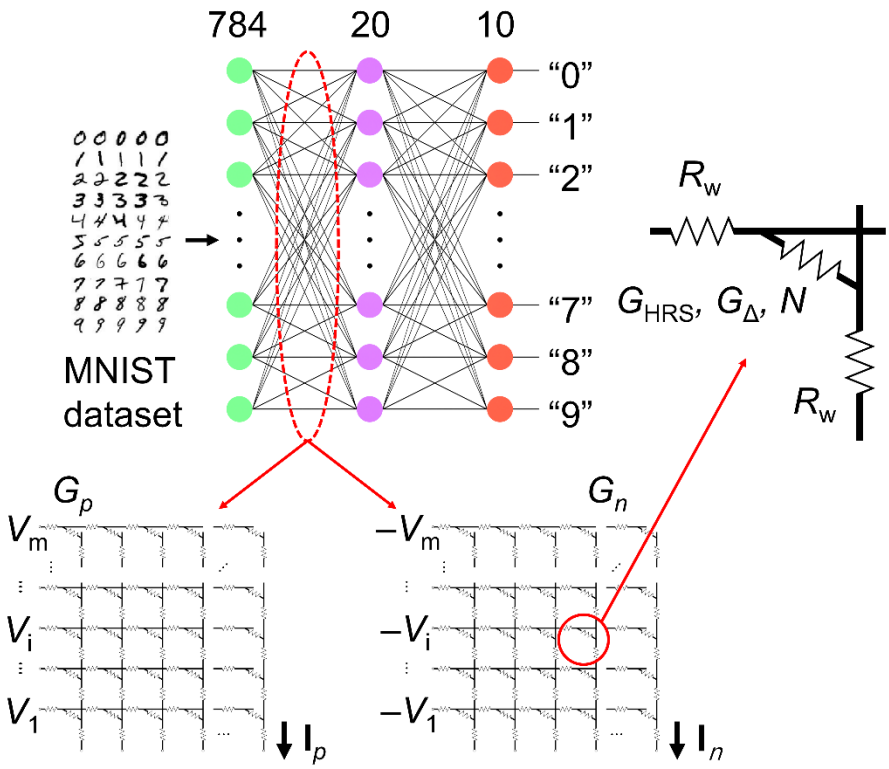

Fig. 3.1. Schematic of the configured fully-connected 2-layer neural network and its implementation as complementary synaptic RRAM arrays.

Fig. 3.1 shows the fully-connected neural network structure in our

simulation. The network was constructed for MNIST pattern recognition, and it consists of a hidden layer containing 20 neurons, 784 input neurons, and 10 output neurons. The synapse layer between each neuron layer consists of two resistive networks containing RRAM synaptic elements: an array with positive weight values (G_p) and an array with negative weight values (G_n). Looking at the internal elements of each array, wire resistance (R_w) exists between the nodes, and three variables represent the state of the RRAM cell: the lowest conductance value (G_{HRS}) that the RRAM can represent, and the required dynamic conductance window (G_w) for enough expression of the number of weight levels (N).

The simulation parameters are summarized in Table I. First, the wire resistance (R_w) was appropriately set in the range of 0.1 Ω to 10 Ω referring to the results of previous studies [7, 8, 41, 42, 52, 53]. The lowest conductance value (G_{HRS}) that RRAM can express was determined based on the device measurement results of previous chapters, ranging from 1 μS to 100 μS. For

simplicity, the rest G_w and N values were fixed to 10, respectively.

Parameter	Range and unit	Description
R_w	0.1 Ω ~ 10 Ω	Varying wire resistance value between adjacent nodes in the array
G_{HRS}	1 μS ~ 100 μS	Varying minimum conductance value of RRAMs
N	10	Number of expressible uniform conductance states of RRAMs
G_w	10	Conductance window of RRAMs enough to express the states N

Table 3.1. Array parameters used in the simulation.

After training the given fully-connected neural network in software, the process of obtaining the recognition accuracy was preceded. The training was performed by the general gradient descent method, and the software inference accuracy was 96.41% for 10,000 MNIST datasets. The trained weight values were again converted through appropriate scaling and quantization processes to

be the same as the conductance range of RRAM given in Table I. The same data inference process was performed through a non-ideal synapse array considering the converted RRAM conductance and wire resistance.

In the weight converting process, we used a general method of dividing the weights into complementary positive (G_p) and negative (G_n) conductance arrays [54, 55]. A positive voltage signal and a negative voltage signal having the same magnitude are input to each array, respectively. Then, the total conductance (G^l) of each l-th synaptic layer can be expressed as a combination of quantized N-state conductance array matrix $G_p{}^l$ and $G_n{}^l$, as the equation below.

$$G^l = G_p^l - G_n^l \tag{3.1}$$

In addition, the cell conductance values that $G_p{}^l$ and $G_n{}^l$ can have are shown using the array internal variables G_{HRS}, G_w, and N as follows.

$$G_{\mathrm{HRS}} G_w = G_{\mathrm{HRS}} + (N-1)\alpha \tag{3.2}$$

Here, α means the gap between each conductance level. Consequently, the quantized conductance values of an array are as follows.

$$G_{\mathrm{HRS}},\ G_{\mathrm{HRS}}+\alpha,\ G_{\mathrm{HRS}}+2\alpha,\ \cdots,\ G_{\mathrm{HRS}}+(N-1)\alpha$$

As a result, a synapse layer can express up to $2N$-1 conductance levels between $G_{pmin}^{l}-G_{nmax}^{l}=-(N-1)\alpha$ and $G_{pmax}^{l}-G_{nmin}^{l}=(N-1)\alpha$ including zero weight. For example, if complementary G_p and G_n array with G_{HRS} = 1 μS, G_w = 10, and N = 10 constitute one synapse layer, 19 conductance levels from -99 μS to 99 μS can exist in the layer.

In the presence of wire resistance, the VMM operation becomes quite complex and cannot be expressed as a simple product of inputs and weights. To solve this non-ideal VMM, we constructed a matrix equation using Kirchhoff's current law and array parameters. The matrix equation representing the system is as follows.

$$\mathbf{G}_{eq}\mathbf{v}_u=\mathbf{v}_i \tag{3.3}$$

$\mathbf{v}_i$ and $\mathbf{G}_{eq}$ are a vector and a matrix defined as input and array parameters, respectively. $\mathbf{v}_u$ is a non-ideal array node voltage vector, containing the values to find out. Solving the linear system $\mathbf{v}_u = \mathbf{G}_{eq}^{-1}\mathbf{v}_i$, all the node voltage

values in the array are found where wire resistance exists. If the array node voltage is known, the non-ideal output current vector **I** can be obtained through the following equation.

$$\mathbf{I} = G_w \mathbf{v}_0 \tag{3.4}$$

$\mathbf{v}_0$ is a sub-vector of $\mathbf{v}_u$ and contains the node voltage values of BL crossing the first WL from the current output terminal. G_w is the reciprocal of R_w, and the output terminal voltage is assumed to be a virtual ground [56]. The sensing resistor is omitted for simplicity assuming that the wire segment (R_w) takes the role, and the output current is assumed to be properly scaled before being fed to the next layer.

When using the input vector of the first synapse array ($\mathbf{x}^1$) for each MNIST dataset, non-ideal inference operation through the l-th synapse array is expressed as follows.

$$\mathrm{ReLU(x)} = \begin{cases} x \ (x \geq 0) \\ 0 \ (x < 0) \end{cases} \tag{3.5}$$

$$\left.\begin{aligned} \mathbf{v}_p^l &= \mathbf{x}^l \\ \mathbf{v}_n^l &= -\mathbf{x}^l \\ \mathbf{I}_p^l &= G_w \mathbf{v}_{p0}^l \\ \mathbf{I}_n^l &= G_w \mathbf{v}_{n0}^l \\ \mathbf{x}^{l+1} &= \mathrm{ReLU}\left(\mathbf{I}_p^l + \mathbf{I}_n^l\right) \end{aligned}\right\} l = 1,2 \tag{3.6}$$

$\mathbf{v}_p{}^l$ and $\mathbf{v}_n{}^l$ are voltage input signals applied to the positive and negative conductance arrays, respectively. The non-ideal output current from each positive ($\mathbf{I}_p{}^l$) and negative ($\mathbf{I}_n{}^l$) conductance array is obtained from each BL voltage vector ($\mathbf{v}_{p0}{}^l$, $\mathbf{v}_{n0}{}^l$) according to equation (3.4). The obtained current vectors are added, converted into a voltage signal through TIA, and transmitted as the input of the next neuron layer as shown in equation (3.6).

For simplicity, it is assumed that the negative components of the added current are rectified to zero and the remaining components are converted into voltage signals. Besides, it fits well with ReLU activation of equation (3.5), the most usual feed-forward method used in ANNs. Through these processes, inference through a non-ideal array was performed, and the accuracy was calculated by comparing 10,000 answer labels with the final output label having

the highest current value in each test data.

3.2 Weight rearrangement method for inference accuracy improvement

Inside the array, the larger sneak current is generated as the resistance of each synaptic RRAM cell is low or the wire resistance is high. The best way is to consider these factors in the design stage of the array. However, the inference accuracy of the already fabricated array can be additionally improved by adding a simple software preprocessing process before transferring the weights, and this method is the weight rearrangement (WR) that we newly proposed.

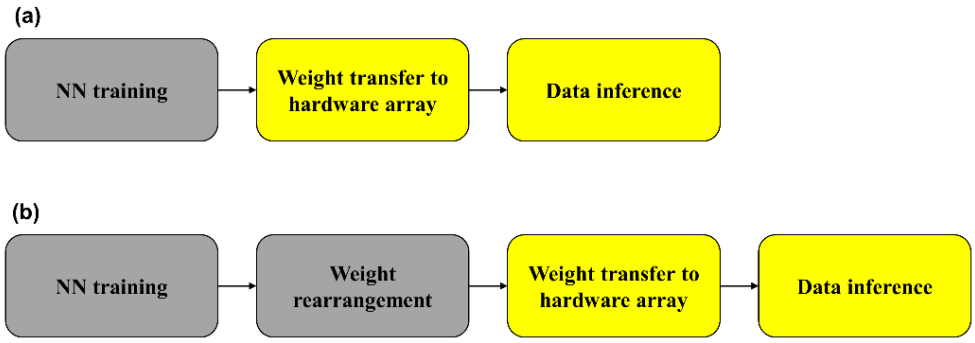

Fig. 3.2. Flow chart of the inference process (a) without WR, and (b) with WR. Gray boxes indicate the software processing, and yellow boxes indicate the hardware processing.

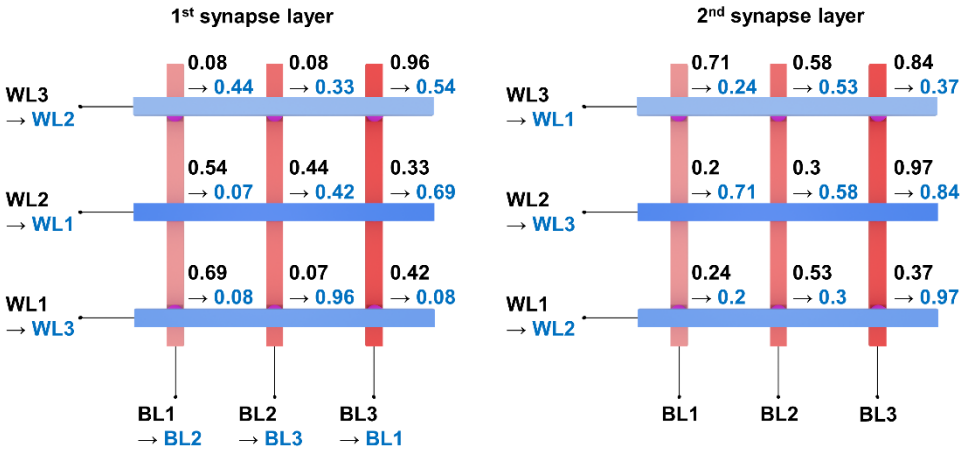

Fig. 3.3. Schematic of the simple example of WR algorithm. The WL containing the larger maximum weight is rearranged toward the BL terminals. To maintain

the synapse layer connection, the BL of the front layer was rearranged according to the WL order of the rear layer.

Fig. 3.2(a) shows the general process when performing off-chip neural network operation using a synaptic array. In this process, data inference is performed by transferring the pre-trained weight values to the array and feeding an input signal. As shown in Fig. 3.2(b), the WR process is performed in the previous stage of weight transfer, as a software preprocessing.

During the read operation of the array, the farther the cell with a small resistance value is from the terminal, the larger sneak current occurs, because it weakly prevents the voltage drop along the electrode line [57]. Therefore, if the cell with a small resistance value (large weight) is relocated close to the terminal, the effect of reducing sneak current and improvement of the inference accuracy is expected. The algorithm of weight rearrangement is shown in Fig. 3.3. It is implemented by the rule: sequencing the WLs containing larger maximum

weights to be relocated closer to the BL output terminals. For example, referencing the second synapse layer in Fig. 3.3, there are maximum weights of 0.53, 0.97, and 0.84 when looking at WL1, WL2, and WL3 respectively. The descending order of those values is 0.97, 0.84, and 0.53, thus the corresponding rearranged WL sequence is WL2, WL3, and WL1 from the output terminals. As rearranging the second layer's WLs by the rule, the first layer's BLs are also rearranged accordingly to maintain connectivity between the layers. Despite the first layer's BLs being permutated, the maximum weights in the first layer's WLs are not changed. Then the same rule is applied again, so the first layer's WLs are permutated in the same manner, bringing rearrangement of relevant weights of the whole arrays.

Generally, in the structure of multi-layer NN, WR is implemented like back-propagation from the last synapse layer to the first synapse layer. Consequently, the weight-rearranged array yields the ideally same output current compared to the original array when the rearranged input signal is used.

The method is realized by applying the sorting algorithm to the pre-trained weights and transferring the rearranged weights to the hardware array instead of the original weights. The general descending sort algorithm has a time complexity of O($n$log$n$), thus WR is implemented in a relatively short time even when applying to the large-scale array.

3.3 Simulation results

Fig. 3.4 shows the weight distribution of the G_p of the first synapse layer when G_{HRS} = 1 μS, with or without WR for example. It is shown that high conductance values are concentrated at the output terminal, as a result of the rearrangement. Consequently, it is easily expected to appear mentioned effect of mitigating unwanted IR drop and sneak current due to large weights far from BL terminals.

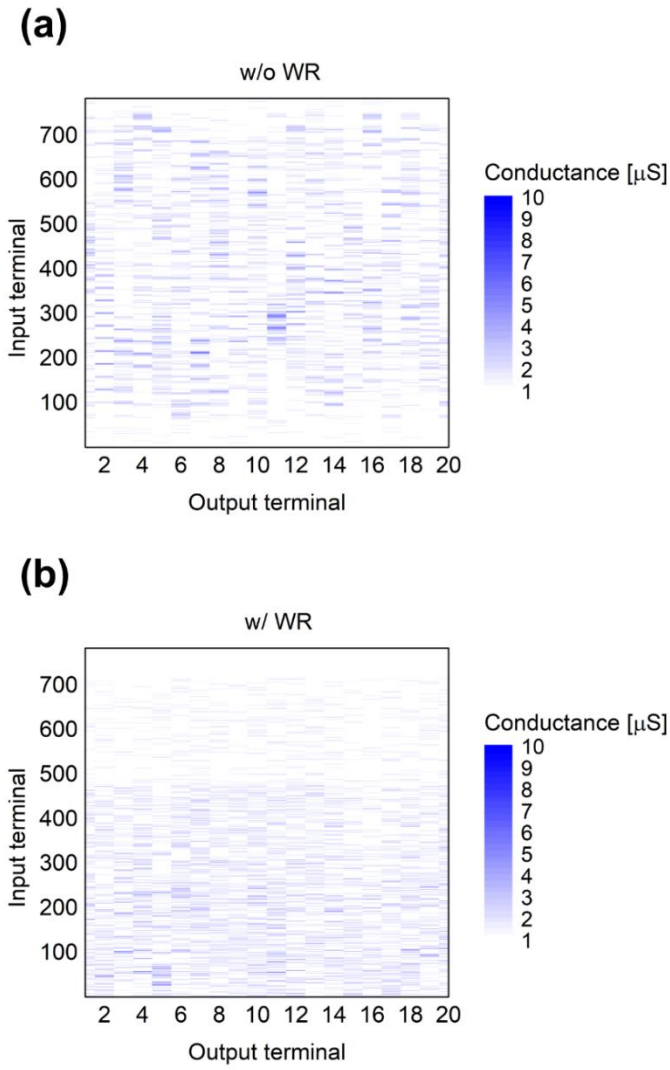

Fig. 3.4. Transferred weight maps of the first G_p synapse arrays, (a) without or (b) with WR, when G_{HRS} = 1 μS.

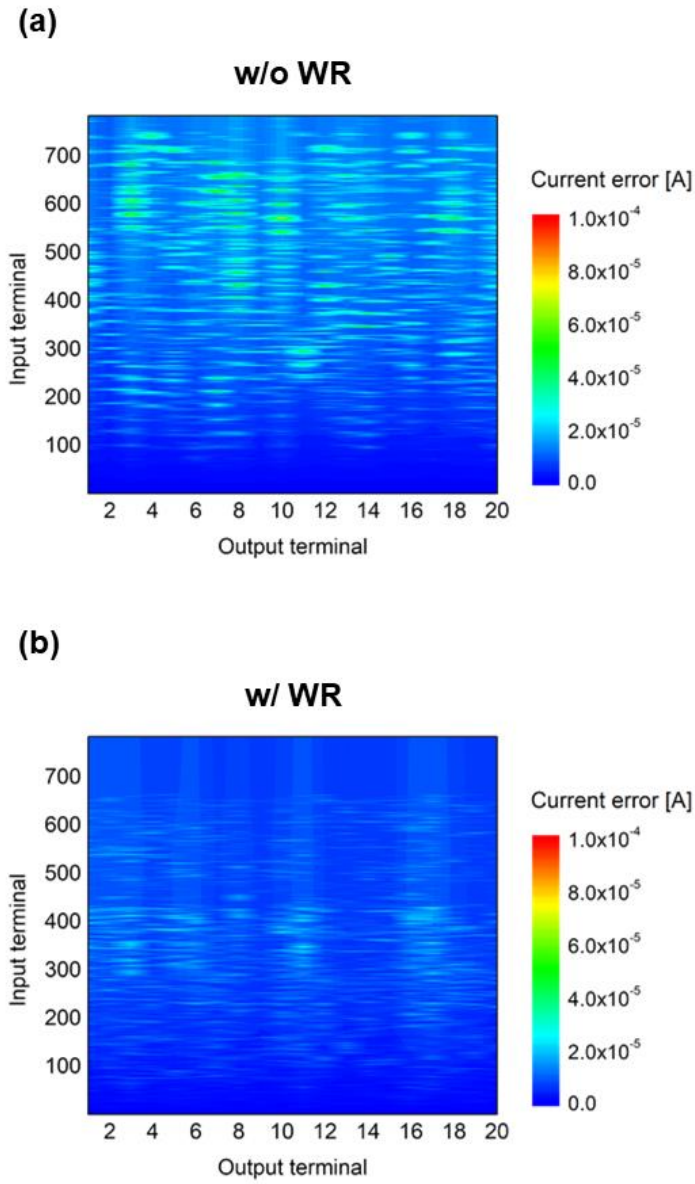

Fig. 3.5. Current error maps of the G_p array of the first synapse layer, when (a) without using WR, and (b) with WR. Inputs of an MNIST data sample and the array conditions of G_{HRS} = 1 μS, R_w = 1 Ω are used.

To confirm the WR effect more directly, current error maps of the G_p array of the first synapse layer are plotted (Fig. 3.5). An MNIST data sample and the array conditions of G_{HRS} = 1 μS, R_w = 1 Ω are used for the test. In the error map, each contour color level represents the cell current difference between ideal and non-ideal cases. As predicted, when not using WR, it is confirmed that the large current error mainly occurs in the area far from BL terminals in the array (Fig. 3.5(a)). As shown in Fig. 3.5(b), When using WR, it is verified that the current error of the synapse array is visibly mitigated due to the rearranged small weights far from BL terminals.

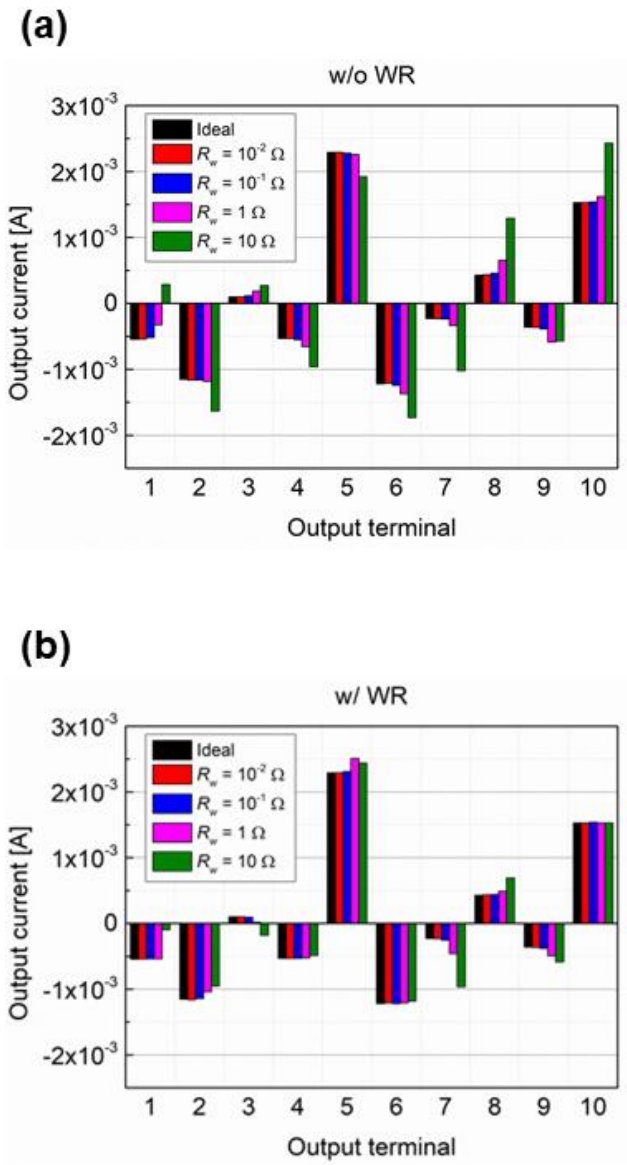

Fig. 3.6. (a) An example of flipped second synapse output priority due to the current error induced by wire resistance when not applying WR. (b) Applying WR maintains the priority. An MNIST data sample is used for the test.

Fig. 3.6 shows a more specific example that implies accuracy degradation and compensation effect of WR. An MNIST sample has the prior three output terminals of 5, 10, and 8 (4, 9, and 7 for the MNIST label) in sequence ideally. As wire resistance increases, current outputs are distorted, and current error at the prior terminals increases. When the wire resistance is finally high enough (R_w = 10 Ω) and not applying WR, the output from the second terminal flips the winner of the ideal case, then the sequence of the prior three terminals changes to 10, 5, and 8 (9, 4, and 7 for MNIST label) deriving incorrect answers. When applying WR, it is confirmed that the output priority maintains even under high R_w, implying the improvement of accuracy.

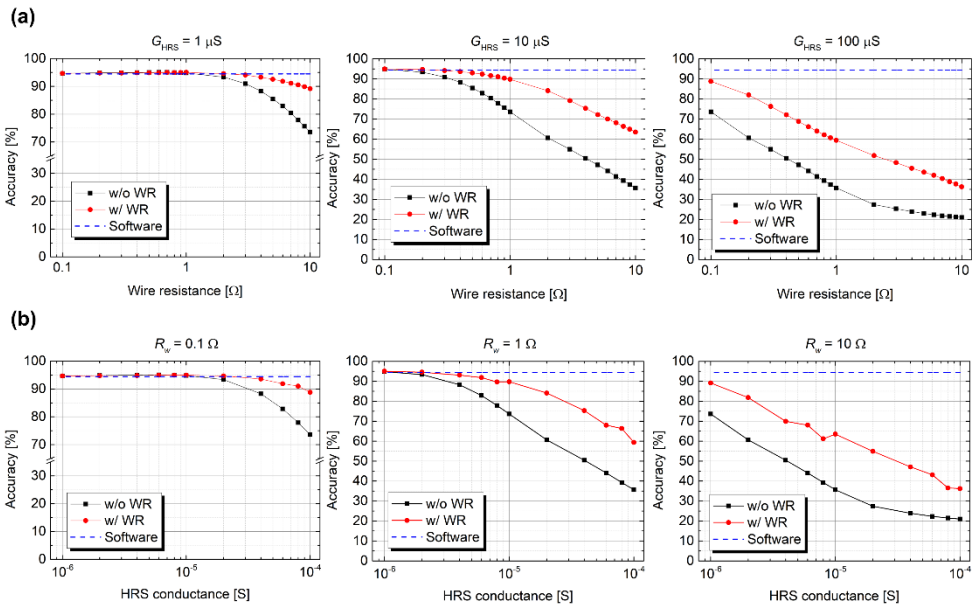

Fig. 3.7. Inference accuracy degradation without WR and improvement with WR, (a) as a function of wire resistance under specific G_{HRS} conditions, and (b) as a function of HRS conductance under specific R_w conditions.

Fig. 3.7(a) shows the change and improvement of accuracy with WR as a function of wire resistance under specific G_{HRS} conditions, and Fig. 3.7(b) shows the same accuracy trend but as a function of HRS conductance under specific R_w conditions. In both cases, the amount of improvement by WR

increased as the degradation is more severe. In addition, if the degradation becomes worse above a certain level, the amount of improvement decreases again. It is confirmed that the maximum amount of improvement obtained under all tested conditions in Fig. 3.7 was 27.86%, boosting the inference accuracy from 35.67% to 63.53%. In addition, it is possible to predict the difference in accuracy between the manufactured nano-tip and planar device through the graphs. Based on the measurement result, the HRS of the nano-tip and planar array is 10 μS and 100 μS, respectively. As an example, if applying this to the case of wire resistance of 1 Ω, it is confirmed that the inference accuracy increases by 40% from 35% to 75%. As a result, it is verified that the higher HRS value of the nano-tip has the effect of preventing sneak current flowing inside the array and improving the accuracy.

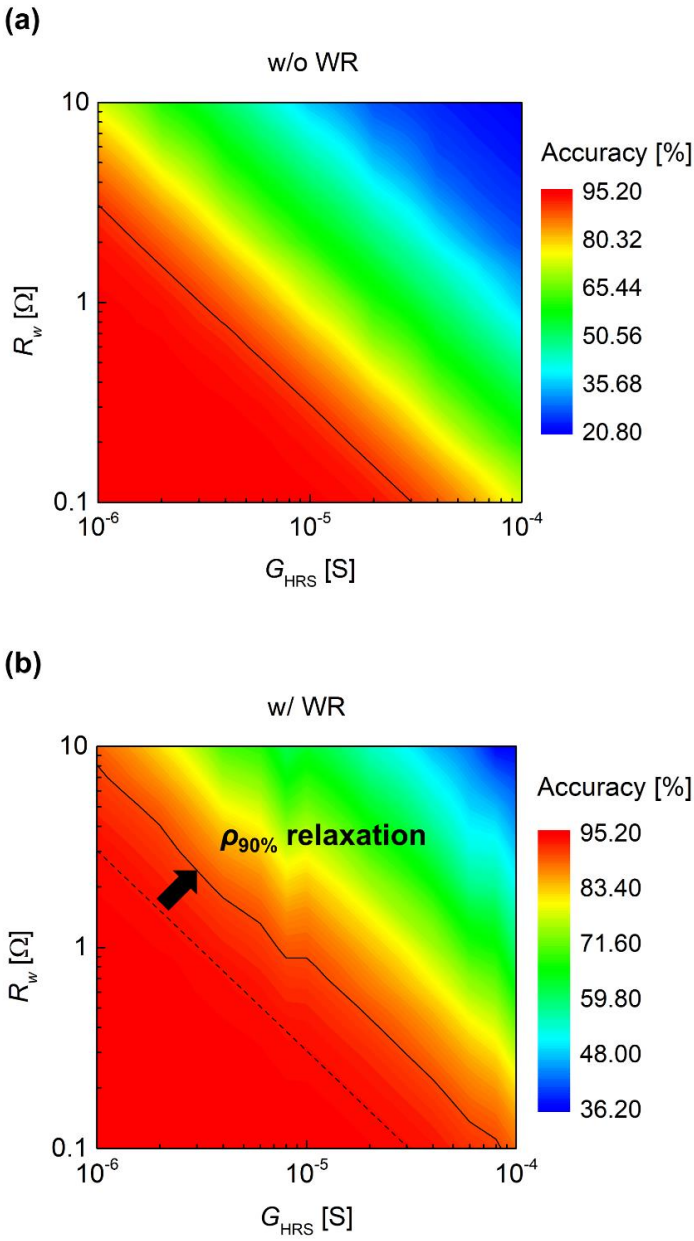

Fig. 3.8. Inference accuracy plot depending on the cell and wire resistance with or without WR. (a) A solid line represents the constraint for 90% accuracy

without WR. (b) The solid line represents the constraint for 90% accuracy with WR and a dashed line which is the case without WR is represented to show the relaxation of the constraint.

Fig. 3.8 shows the contour plot of non-ideal inference accuracy over 10,000 MNIST test data, under varying RRAM cell and wire resistance within the range suggested in the previous section. The inference accuracies under certain condition of R_w or G_{HRS} with or without WR is predicted from the plot. For example, the plot shows that the inference degradation possibly occurs up to under 20% accuracy without WR when both R_w and G_{HRS} are simultaneously bad, being unusable. Applying WR, it is confirmed that the high accuracy area has broadened and the accuracy improved under the same array conditions.

In addition, it is found in the plot that contour lines of the same accuracy are almost linear in a double logarithmic plot, so it is assumed that a product

constant for achieving certain accuracy exists. The constant is defined as follows.

$$\rho = G_{\mathrm{HRS}} R_w \tag{3.7}$$

For example, if the constant for achieving 90% accuracy is defined as $\rho_{90\%}$, the value of $\rho_{90\%} = 3.2 \times 10^{-6}$ is obtained from the contour plot of 3.9(a). In other words, the design constraint of the suggested fully-connected NN for achieving over 90% inference accuracy is predicted as the cell-wire resistance ratio of 3.2×10^{-6} without WR. Referencing the wire resistance value (R_w = 1 Ω) of a reasonable technology node [42], cell conductance of G_{HRS} = 3.2 μS must satisfy to meet the practical accuracy while facing the challenge of HRS variation and large operation voltage. With WR, one expects to mitigate those constraints by software pre-processing without much effort. The constraint for 90% accuracy ($\rho_{90\%}$) is represented as a solid and dashed line in Fig. 3.8(b), with or without WR respectively. From the plot we verified that $\rho_{90\%}$ is mitigated from 3.2×10^{-6} to roughly 9×10^{-6} with WR, resulting in a more relaxed design

constraint. Consequently, about 2.81 times of array scalability or HRS limitation is obtained for the same accuracy using WR.

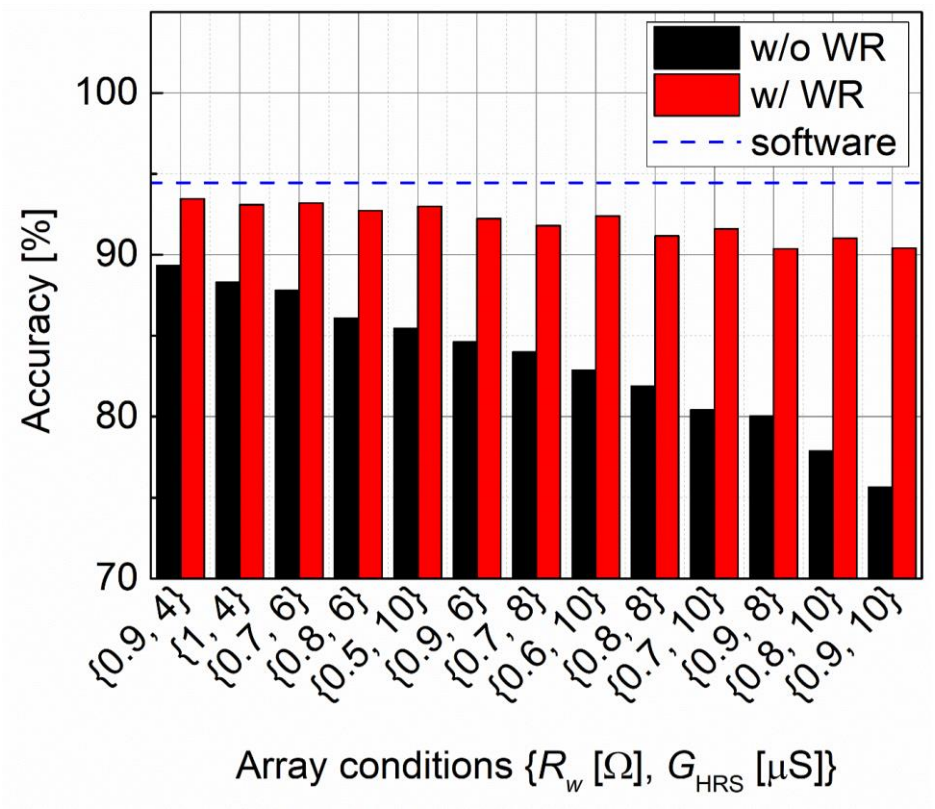

Fig. 3.9. Accuracy compensation with WR when testing 13 arrays having different R_w and G_{HRS}. Each array condition value is expressed in a bracket.

The most important merit of the WR method we concluded is that it raises certain insufficient array accuracies over a meaningful level, without hardware change. Examples of the accuracy compensation effect of WR from under 90% to over 90% are shown in Fig. 3.9. 13 arbitrary arrays that have different G_{HRS} and R_w are used. In the first arbitrary array (G_{HRS} = 4 μS, R_w = 0.9 Ω) the degradation occurs up to 89.34%, and WR improves the accuracy up to 93.45% which is close to software. Another arbitrary array (G_{HRS} = 10 μS, R_w = 0.9 Ω) demonstrates the unusable performance of 75.64% accuracy and using WR the accuracy is enhanced to a practical level (90.4%) with a 14.76% increase. In rest-tested conditions, there are several unpractical arrays whose accuracy goes under 90%, but WR improves whole that degradation to over 90%. Consequently, WR is proven to be capable of putting the degraded arrays to practical use.

In conclusion, it is confirmed that the degraded accuracy improved averagely by 8.62% in tested array conditions. It means that one can operate the

already fabricated array with higher performance when applying WR. An additional notable thing is the proposed method requires only the simple sorting algorithm, so can be generally extended to other types of inference networks not only for MNIST.

Chapter 4

Hardware Implementation of the Si nano-tip RRAM Array for Neuromorphic Application

In this chapter, we examine the operation of the fabricated Si nano-tip RRAM array as a synaptic device. We measured the characteristics of a basic synaptic operation of the proposed array. Demonstration of the weight rearrangement method simulated in the previous chapter is also covered.

4.1. Measurement results of the synaptic operation of the Si nano-tip array

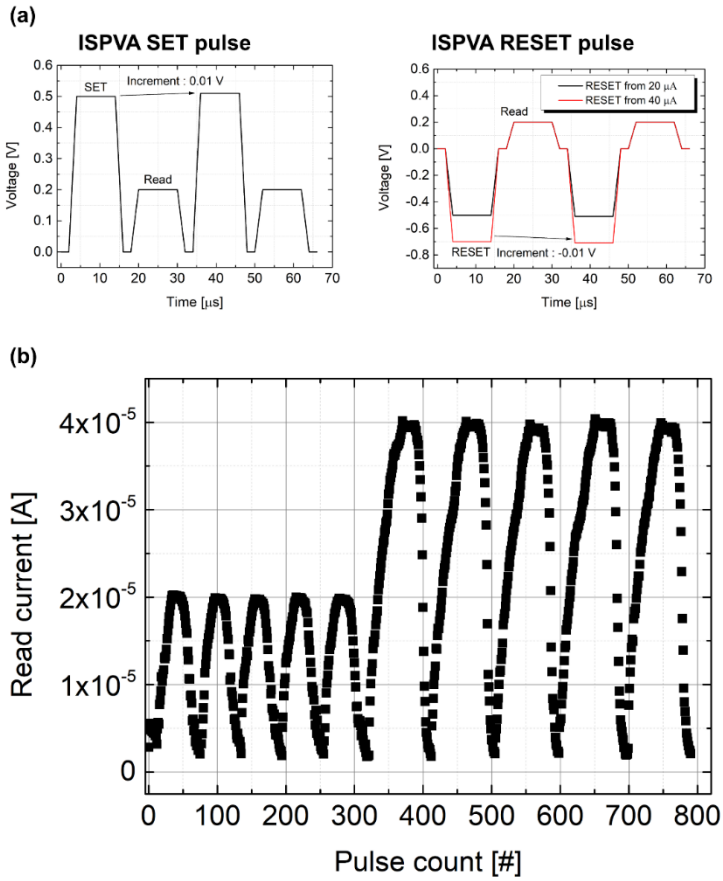

Fig. 4.1. Gradual switching characteristics of nano-tip array device for ISPVA method. (a) ISPVA SET and RESET pulse waveforms, and (b) the switching result as pulse count.

Fig. 4.1 shows the results of incremental step pulse with verify algorithm (ISPVA) programming in the nano-tip array. the SET pulse consists of a pulse width of 10 μs, a start voltage of 0.5 V, and a voltage increment of 10 mV. The RESET pulse was constructed with a pulse width of 10 μs, a start voltage of -0.5 V for RESET from 20 μA read current, -0.7 V for RESET from 40 μA read current, and a voltage increment of -10 mV. As a result, the nano-tip array device shows the capability of elaborate gradual switching in the window from 2 μA to 40 μA at Vread = 0.2 V. In addition, endurance with about 800 ISPVA pulses was also verified. Based on the result, weight transfer to the nano-tip array using ISPVA was conducted.

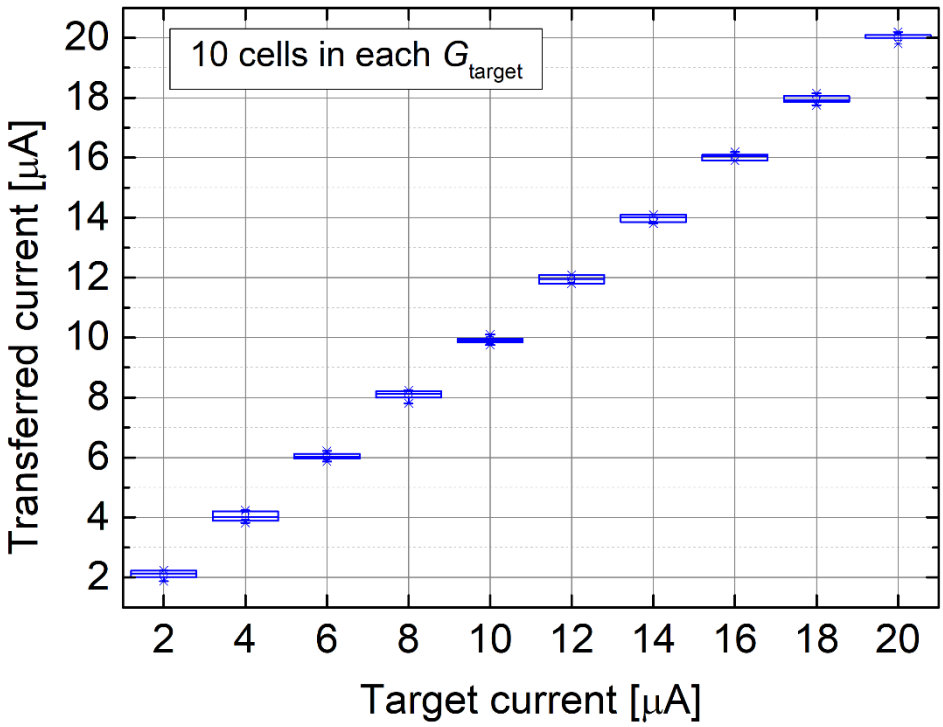

Fig. 4.2. Weight transfer result in a 10 × 10 subarray in a 16 × 16 nano-tip array. At the read voltage of 0.2 V, 10 level conductance were defined from 2 μA to 20 μA, and 10 cells in each conductance are programmed.

Fig. 4.2 shows the results of weight transfer by defining 10 × 10 subarrays in a 16 × 16 nano-tip array. Target conductance was set at 10 levels at 2 μA intervals from 2 μA to 20 μA at a read voltage of 0.2 V. As a result of the

transfer using ISPVA, it can be confirmed that the transfer was performed with a very small variation for 10 cells per each weight.

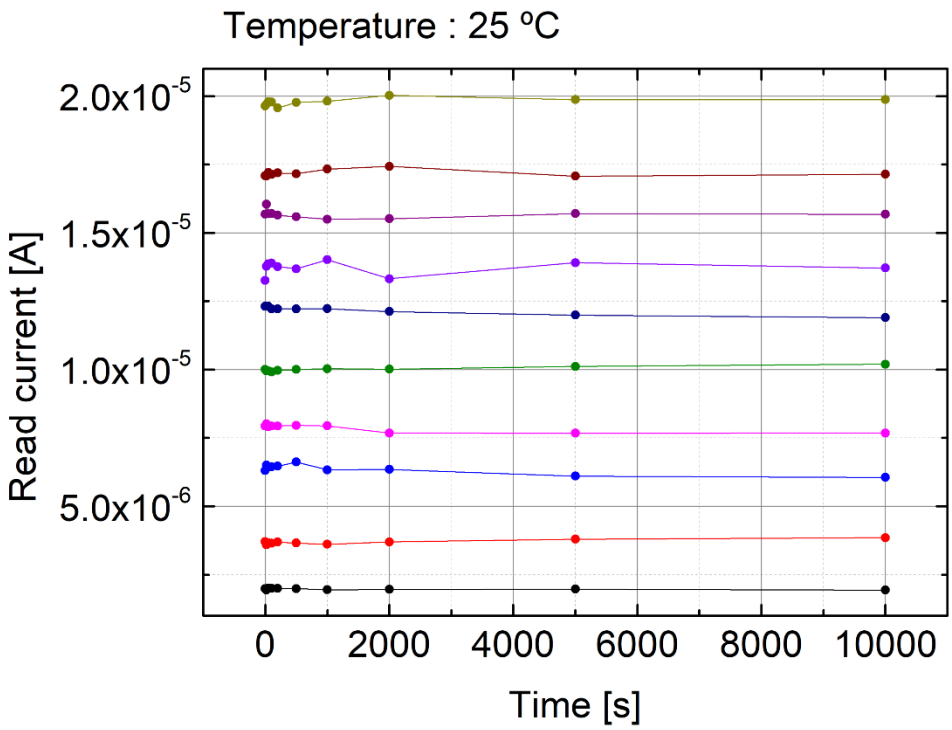

Fig. 4.3. 10 level retention of weights from 2 μA to 20 μA read current, under room temperature.

Fig. 4.3 shows the results of measuring retention for 10,000 seconds for 10-level weights that have been transferred. It was carried out at room temperature, and as a result, it was confirmed that the weights were well maintained without significant disturbance during the measurement time.

4.2. Demonstration of the weight rearrangement with the Si nano-tip RRAM array

To confirm the effect of the weight rearrangement method we proposed in the previous chapter in the actual array, verification measurement was performed using the fabricated nano-tip array.

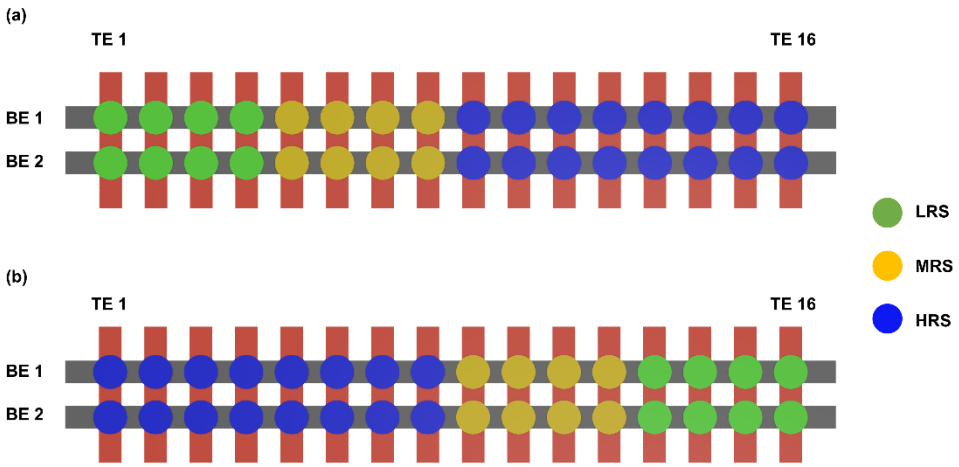

Fig. 4.4 Schematic of the measurement for weight rearrangement. (a) the best case, and (b) the worst case of weight arrangement. The middle resistance state (MRS) was defined and transferred.

Figure 4.4 shows a schematic of weight rearrangement verification measurement. A 16 × 2 array, which is a subarray of a 16 × 16 nano-tip array, was defined to be tested. The best case was determined when large weights are transferred near the BE sensing terminal as shown in Fig. 4.4(a). Conversely, the worst case (Fig. 4.4(b)) was defined as the case of transferring a large weight to the far side from the BE sensing terminal. After applying the read voltage 0.2

V to each crossbar cell once, the added cell current values become the ideal VMM. Besides, the value read from the BE sensing node applied simultaneously from all TE input terminals becomes the measured VMM. If the weight rearrangement is effective, the difference between the ideal VMM and the actual VMM should be smaller in the best case than in the worst case.

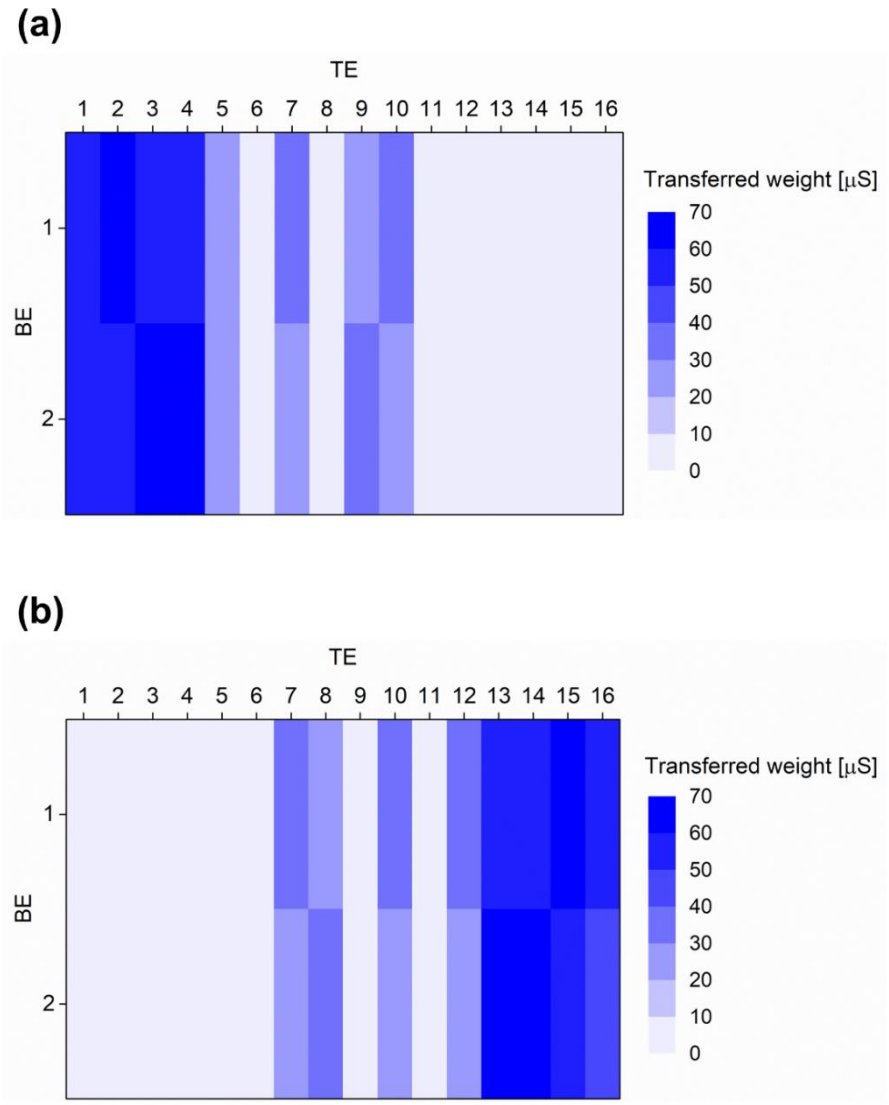

Fig. 4.5. The map of weights actually transferred to a 16 × 2 array (a) for the best case, and (b) worst case.

Fig. 4.5 is the result of executing ISPVA weight transfer on a 16 × 2 nano-tip subarray. Fig. 4.5(a) is the best case, and Fig. 4.5(b) is the worst case. When executing the best case, SET failure occurred in TE numbers 6 and 8. Instead, transfer was performed using TE numbers 9 and 10. In the worst case, the weight arrangement was reversed and transferred. Target conductance was set at 10 μA for LRS, 5 μA for MRS, and 1 μA for HRS at a read voltage of 0.2 V, and when converted into Siemens units, it is 50 μS, 25 μS, and 5 μS, respectively. As a result of the transfer, it was confirmed that the target conductance was well reached in both the best case and the worst case.

(a)

Best case of the weight arrangement

BE	Ideal VMM [μA]	Measured VMM [μA]	Error [%]
1	76.4	70.8	**7.33**
2	76.2	71.4	**6.3**

(b)

Worst case of the weight arrangement

BE	Ideal VMM [μA]	Measured VMM [μA]	Error [%]
1	79.2	71.6	**9.6**
2	74.1	68.8	**7.15**

Fig. 4.6. Results of weight rearrangement measurement.

Figure 4.6 is the result of measuring weight rearrangement. In the case of the best case (Fig. 4.6(a)), the BE numbers 1 and 2 had ideal VMM values of

76.4 μA and 76.2 μA, respectively, and the measured VMM values were 70.8 μA and 71.4 μA, respectively. The error values calculated from each BE were 7.33% and 6.3% in the best case. In the worst case (Fig. 4.6(b)), ideal VMM values of 79.2 μA and 74.1 μA were obtained for BE numbers 1 and 2, respectively. The measured VMM values were 71.6 μA and 68.8 μA, respectively, with 9.6% and 7.15% errors. In both BE numbers 1 and 2, the error of the current sum was measured to be smaller in the best case than in the worst case and improved by 2.27% in BE number 1 and 0.85% in BE number 2.

Chapter 5

Conclusion

In this thesis dissertation, we fabricated a single RRAM device and RRAM array using a Si nano-tip structure and investigated the electrical characteristics. The difference in operating characteristics with the planar single RRAM and array fabricated as a control device was also confirmed, and a meaningful decrease in the current level is verified as the most critical difference. Therefore, we also verified that the operating current of the SL we used has an area dependency. In addition, we found trade-offs in which the operating voltage or device-to-device variation increases, but the variation has

room for improvement through process optimization, and the operating voltage is not large enough to be highlighted as a disadvantage in the total power consumption during inference. Gradual switching capability required for operation as a synaptic device and weight transfer capability through ISPVA were also confirmed. In addition, it was confirmed that reliability characteristics such as weight retention and switching endurance are sufficient to be used as synaptic devices.

As a second research goal, a novel inference scheme of weight rearrangement was proposed, and a non-ideal synapse array simulation was performed to verify the method. A resistor network including synaptic RRAM was constructed by referring to previously reported wire resistance values. The MNIST data inference was performed, and the accuracy change depending on the array condition was predicted. Also, it was confirmed through simulation that the accuracy degradation improved when weight rearrangement is applied. Simulation results showed that the scheme is effective in improving general

inference accuracy degradation by wire resistance. Finally, a test demonstration of weight rearrangement was performed in the fabricated array, and it was confirmed that the effect was effective even in the actual operation of the array.

Bibliography

[1] K. Moon *et al.*, "RRAM-based synapse devices for neuromorphic systems," *Faraday Discussions,* vol. 213, pp. 589-589, Feb. 1, 2019.

[2] Y. Xi *et al.*, "In-memory Learning with Analog Resistive Switching Memory: A Review and Perspective," *Proceedings of the IEEE,* vol. 109, no. 1, pp. 14-42, Jan. 2021.

[3] Z. J. Shen *et al.*, "Advances of RRAM Devices: Resistive Switching Mechanisms, Materials and Bionic Synaptic Application," *Nanomaterials,* vol. 10, no. 8, Aug. 2020.

[4] V. Milo *et al.*, "Multilevel HfO_2-based RRAM devices for low-power neuromorphic networks," *Apl Materials,* vol. 7, no. 8, Aug. 2019.

[5] D. Ielmini, "Brain-inspired computing with resistive switching memory (RRAM): Devices, synapses and neural networks," *Microelectronic Engineering,* vol. 190, pp. 44-53, Apr. 15, 2018.

[6] Y. Z. Zhang, P. Huang, B. Gao, J. F. Kang, and H. Q. Wu, "Oxide-based filamentary RRAM for deep learning," *Journal of Physics D-Applied Physics,* vol. 54, no. 8, Feb. 25, 2021.

[7] Y. Liao *et al.*, "A Compact Model of Analog RRAM With Device and Array Nonideal Effects for Neuromorphic Systems," *IEEE Transactions on Electron Devices,* vol. 67, no. 4, pp. 1593-1599, Apr. 2020.

[8] C. Li *et al.*, "Efficient and self-adaptive in-situ learning in multilayer memristor neural networks," *Nature Communications,* vol. 8, Jun 19, 2018.

[9] C. Li *et al.*, "Analogue signal and image processing with large memristor crossbars," *Nature Electronics,* vol. 1, no. 1, pp. 52-59, Jan. 2018.

[10] P. Yao *et al.*, "Face classification using electronic synapses," *Nature Communications,* vol. 8, May 12, 2017.

[11] A. Mohanty, X. Du, P.-Y. Chen, J.-s. Seo, S. Yu, and Y. Cao, "Random

sparse adaptation for accurate inference with inaccurate multi-level RRAM arrays," in *2017 IEEE International Electron Devices Meeting (IEDM)*, 2017: IEEE, pp. 6.3. 1-6.3. 4.

[12] A. Mehonic, D. Joksas, W. H. Ng, M. Buckwell, and A. J. Kenyon, "Simulation of Inference Accuracy Using Realistic RRAM Devices," *Frontiers in Neuroscience,* vol. 13, June 12, 2019.

[13] Y. Z. Zhang *et al.*, "An Improved RRAM-Based Binarized Neural Network With High Variation-Tolerated Forward/Backward Propagation Module," *IEEE Transactions on Electron Devices,* vol. 67, no. 2, pp. 469-473, Feb. 2020.

[14] J. Woo *et al.*, "Multi-layer tunnel barrier ($Ta_2O_5/TaO_x/TiO_2$) engineering for bipolar RRAM selector applications," in *2013 Symposium on VLSI Technology*, 2013: IEEE, pp. T168-T169.

[15] Y.-S. Chen *et al.*, "Novel Defects-Trapping TaO_x/HfO_x RRAM With Reliable Self-Compliance, High Nonlinearity, and Ultra-Low Current," *IEEE Electron Device Letters,* vol. 35, no. 2, pp. 202-204, 2013.

[16] U. Chand, K.-C. Huang, C.-Y. Huang, and T.-Y. Tseng, "Mechanism of nonlinear switching in HfO_2-based crossbar RRAM with inserting large bandgap tunneling barrier layer," *IEEE Transactions on Electron Devices,* vol. 62, no. 11, pp. 3665-3670, 2015.

[17] S. Swathi and S. Angappane, "Low power multilevel resistive switching in titanium oxide-based RRAM devices by interface engineering," *Journal of Science: Advanced Materials and Devices,* vol. 6, no. 4, pp. 601-610, 2021.

[18] Q. Xue *et al.*, "Ultralow Set Voltage and Enhanced Switching Reliability for Resistive Random-Access Memory Enabled by an Electrodeposited Nanocone Array," *ACS Applied Materials & Interfaces,* 2022.

[19] Z.-J. Liu, J.-C. Chou, S.-Y. Wei, J.-Y. Gan, and T.-R. Yew, "Improved resistive switching of textured ZnO thin films grown on Ru electrodes," *IEEE Electron Device Letters,* vol. 32, no. 12, pp. 1728-1730, 2011.

[20] H. J. Kim *et al.*, "Fabrication of a Cu-cone-shaped cation source inserted

conductive bridge random access memory and its improved switching reliability," *Advanced Functional Materials,* vol. 29, no. 8, p. 1806278, 2019.

[21] Y. Sun *et al.*, "Guiding the growth of a conductive filament by nanoindentation to improve resistive switching," *ACS Applied Materials & Interfaces,* vol. 9, no. 39, pp. 34064-34070, 2017.

[22] W. Xiao, W. Song, Y. P. Feng, D. Gao, Y. Zhu, and J. Ding, "Electrode-controlled confinement of conductive filaments in a nanocolumn embedded symmetric–asymmetric RRAM structure," *Journal of Materials Chemistry C,* vol. 8, no. 5, pp. 1577-1582, 2020.

[23] Y.-C. Shih *et al.*, "Rational Design on Controllable Cation Injection with Improved Conductive-Bridge Random Access Memory by Glancing Angle Deposition Technology toward Neuromorphic Application," *ACS Applied Materials & Interfaces,* vol. 13, no. 46, pp. 55470-55480, 2021.

[24] B. Gao *et al.*, "Modeling disorder effect of the oxygen vacancy distribution in filamentary analog RRAM for neuromorphic computing," in *2017 IEEE International Electron Devices Meeting (IEDM)*, 2017: IEEE, pp. 4.4. 1-4.4. 4.

[25] B. Gao *et al.*, "Identification and application of current compliance failure phenomenon in RRAM device," in *Proceedings of 2010 International Symposium on VLSI Technology, System and Application*, 2010: IEEE, pp. 144-145.

[26] Y. Xi *et al.*, "Impact and quantization of short-term relaxation effect in analog RRAM," in *2020 4th IEEE Electron Devices Technology & Manufacturing Conference (EDTM)*, 2020: IEEE, pp. 1-4.

[27] Z. Jiang *et al.*, "Bidirectional analog conductance modulation for RRAM-based neural networks," *IEEE Transactions on Electron Devices,* vol. 67, no. 11, pp. 4904-4910, 2020.

[28] S. Kim *et al.*, "4-bit Multilevel Operation in Overshoot Suppressed Al_2O_3/TiO_x Resistive Random-Access Memory Crossbar Array," *Advanced*

Intelligent Systems, p. 2100273, 2022.

[29] S. Park *et al.*, "RRAM-based synapse for neuromorphic system with pattern recognition function," in *2012 international electron devices meeting*, 2012: IEEE, pp. 10.2. 1-10.2. 4.

[30] K. Kim *et al.*, "Enhanced analog synaptic behavior of SiN_x/a-Si bilayer memristors through Ge implantation," *Npg Asia Materials,* vol. 12, no. 1, pp. 1-13, 2020.

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

[32] T. Kim, H. Kim, J. Kim, and J. J. Kim, "Input Voltage Mapping Optimized for Resistive Memory-Based Deep Neural Network Hardware," *IEEE Electron Device Letters,* vol. 38, no. 9, pp. 1228-1231, Sept. 2017.

[33] H. Kim, T. Kim, J. Kim, and J. J. Kim, "Deep Neural Network Optimized to Resistive Memory with Nonlinear Current-Voltage Characteristics," *Acm Journal on Emerging Technologies in Computing Systems,* vol. 14, no. 2, July 2018.

[34] L. X. Xia *et al.*, "Stuck-at Fault Tolerance in RRAM Computing Systems," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems,* vol. 8, no. 1, pp. 102-115, Mar. 2018.

[35] J. Woo *et al.*, "Exploiting defective RRAM array as synapses of HTM spatial pooler with boost-factor adjustment scheme for defect-tolerant neuromorphic systems," *Scientific Reports,* vol. 10, no. 1, July 16, 2020.

[36] J. Y. Woo and S. M. Yu, "Two-Step Read Scheme in One-Selector and One-RRAM Crossbar-Based Neural Network for Improved Inference Robustness," *IEEE Transactions on Electron Devices,* vol. 65, no. 12, pp. 5549-5553, Dec. 2018.

[37] Y. C. Xiang *et al.*, "Impacts of State Instability and Retention Failure of Filamentary Analog RRAM on the Performance of Deep Neural Network," *IEEE Transactions on Electron Devices,* vol. 66, no. 11, pp. 4517-4522,

Nov. 2019.

[38] Y. D. Du *et al.*, "Exploring the Impact of Random Telegraph Noise-Induced Accuracy Loss on Resistive RAM-Based Deep Neural Network," *IEEE Transactions on Electron Devices,* vol. 67, no. 8, pp. 3335-3340, Aug. 2020.

[39] W. Shim, Y. D. Luo, J. S. Seo, and S. M. 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.

[40] Y. Lin *et al.*, "Demonstration of generative adversarial network by intrinsic random noises of analog RRAM devices," in *2018 IEEE International Electron Devices Meeting (IEDM)*, 2018: IEEE, pp. 3.4. 1-3.4. 4.

[41] W. Choi *et al.*, "WO_x-Based Synapse Device With Excellent Conductance Uniformity for Hardware Neural Networks," *IEEE Transactions on Nanotechnology,* vol. 19, pp. 594-600, 2020.

[42] J. Woo and S. M. Yu, "Impact of Selector Devices in Analog RRAM-Based Crossbar Arrays for Inference and Training of Neuromorphic System," *IEEE Transactions on Very Large Scale Integration (Vlsi) Systems,* vol. 27, no. 9, pp. 2205-2212, Sept. 2019.

[43] L. X. Xia *et al.*, "Technological Exploration of RRAM Crossbar Array for Matrix-Vector Multiplication," *Journal of Computer Science and Technology,* vol. 31, no. 1, pp. 3-19, Jan. 2016.

[44] Z. Sun, G. Pedretti, E. Ambrosi, A. Bricalli, W. Wang, and D. Ielmini, "Solving matrix equations in one step with cross-point resistive arrays," *Proceedings of the National Academy of Sciences,* vol. 116, no. 10, pp. 4123-4128, 2019.

[45] S. N. Truong, "Compensating Circuit to Reduce the Impact of Wire Resistance in a Memristor Crossbar-Based Perceptron Neural Network," *Micromachines,* vol. 10, no. 10, Oct. 2019.

[46] O. Tabata, R. Asahi, H. Funabashi, K. Shimaoka, and S. Sugiyama, "Anisotropic etching of silicon in TMAH solutions," *Sensors and*

Actuators A: Physical, vol. 34, no. 1, pp. 51-57, 1992.

[47] I. Zubel, "Silicon anisotropic etching in alkaline solutions III: On the possibility of spatial structures forming in the course of Si (100) anisotropic etching in KOH and KOH+IPA solutions," *Sensors and Actuators A: Physical,* vol. 84, no. 1-2, pp. 116-125, 2000.

[48] I. Barycka and I. Zubel, "Silicon anisotropic etching in KOH-isopropanol etchant," *Sensors and Actuators A: Physical,* vol. 48, no. 3, pp. 229-238, 1995.

[49] A. Zalar, B. Baretzky, F. Dettenwanger, M. Rühle, and P. Panjan, "Interdiffusion at the Al2O3/Ti interface studied in thin-film structures," *Surface and Interface Analysis: An International Journal devoted to the development and application of techniques for the analysis of surfaces, interfaces and thin films,* vol. 26, no. 11, pp. 861-867, 1998.

[50] F. Wu *et al.*, "Memristor based on TiO_x /Al_2O_3 bilayer as flexible artificial synapse for neuromorphic electronics," *IEEE Transactions on Electron Devices,* vol. 69, no. 1, pp. 375-379, 2021.

[51] L. Alekseeva, T. Nabatame, T. Chikyow, and A. Petrov, "Resistive switching characteristics in memristors with Al_2O_3/TiO_2 and TiO_2/Al_2O_3 bilayers," *Japanese Journal of Applied Physics,* vol. 55, no. 8S2, p. 08PB02, 2016.

[52] A. Chen, "A Comprehensive Crossbar Array Model With Solutions for Line Resistance and Nonlinear Device Characteristics," *IEEE Transactions on Electron Devices,* vol. 60, no. 4, pp. 1318-1326, Apr. 2013.

[53] X. C. Peng, R. Madler, P. Y. Chen, and S. M. Yu, "Cross-point memory design challenges and survey of selector device characteristics," *Journal of Computational Electronics,* vol. 16, no. 4, pp. 1167-1174, Dec. 2017.

[54] G. W. Burr *et al.*, "Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element," *IEEE Transactions on Electron Devices,* vol. 62, no. 11, pp. 3498-3507, 2015.

[55] J. Park, M. Kwak, K. Moon, J. Woo, D. Lee, and H. Hwang, "TiO_x-Based RRAM Synapse With 64-Levels of Conductance and Symmetric Conductance Change by Adopting a Hybrid Pulse Scheme for Neuromorphic Computing," *IEEE Electron Device Letters,* vol. 37, no. 12, pp. 1559-1562, Dec. 2016.

[56] Z. Sun, G. Pedretti, and D. Ielmini, "Fast solution of linear systems with analog resistive switching memory (RRAM)," in *2019 IEEE International Conference on Rebooting Computing (ICRC)*, 2019: IEEE, pp. 1-5.

[57] Y. Zhu, X. Zhao, and K. Qiu, "Insights and Optimizations on IR-drop Induced Sneak-Path for RRAM Crossbar-based Convolutions," in *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2020: IEEE, pp. 506-511.

초 록

저항 변화 메모리 소자는 단순한 구조로 인한 고집적 가능성, 멀티레벨 동작 등의 특성으로 인해 뉴로모픽 시스템에서 시냅스 소자 및 어레이로써의 활용 가능성을 인정받아 관련 연구가 활발히 진행되고 있다. 특히 오프칩 구동에 있어 다수의 저항 변화 메모리 시냅스를 가지는 어레이의 동작에 대한 연구도 많이 진행되었다. 연구가 축적되면서 이러한 대규모 집적과 구동 시냅스 개수의 확장은 필연적으로 일어나게 될 것이고, 고집적된 어레이 동작에서 발생할 수 있는 전력 소모, 라인 저항 문제들에 대해서도 초점을 맞출 필요가 있다.

본 연구에서는 두 가지의 목표가 제시되었다. 첫째로 저전력 동작을 위해, 우리는 실리콘 하부 전극의 비등방성 식각 특성을 활용하여 스위칭 영역의 미세화를 통한 저전류 소자를 제작하였다. 제작된 나노팁 하부전극 구조의 소자는 일반적인 평판 구조의 소자 대비 단일소자는 약 100배에서 1000배 정도, 어레이의 경우 약 10배 정도의 전류 감소 효과가 확인되었다.

둘째로 대규모 시냅스 어레이에서의 라인저항에 의한 동작 왜곡 보완을 위해서, 가중치 재배열을 통한 추론 정확도 향상 방법을 제시하였다. 저항 변화 메모리 소자의 전도도 값과 라인저항 값을 활용하여 시뮬레이션으로 대규모 어레이를 구현하였고, 일반적인 상황과 가중치 재배열 적용 상황에서 추론 정확도를 도출하였다. 결과적으로 시뮬레이션에 사용된 패턴인식 네트워크에서는 유의미한 레벨에서 평균 8.62%의 정확도 향상이 웨이트 재배열에 의해 이루어졌다.

마지막으로 제작된 어레이에서 시냅스 동작 특성을 확인하고, 웨이트 재배열 효과를 검증하였다. 제작된 16 × 16 어레이의 부분 어레이인 16 × 2 어레이를 활용하여 측정한 결과 첫 번째 아이디어에 의해 저전류 동작이 되는 것과 두 번째 아이디어에 의해 전류 합 오차가 줄어드는 것을 확인하였다. 본 연구에서 제시된 방법은 앞으로 연구가 진행될 고집적 대규모 시냅스 어레이에 활용하였을 때 효과적인 개선을 불러올 것으로 기대된다.

주요어 : 저항 변화 메모리, 뉴로모픽 시스템, 추론 정확도, 시냅스 소자, 실리콘 나노팁 구조, 저전류, 가중치 재배열, 라인저항

학번 : 2015-20931

List of Publications

Journal

[1] Yeon-Joon Choi, **Suhyun Bang**, Tae-Hyeon Kim, Kyungho Hong, Sungjoon Kim, Sungjun Kim, Seongjae Cho, and Byung-Gook Park, "Analytically and empirically consistent characterization of the resistive switching mechanism in a Ag conducting-bridge random-access memory device through a pseudo-liquid interpretation approach," *Physical Chemistry Chemical Physics*, Vol. 23, Nov. 2021.

[2] Min-Hwi Kim, Sungmin Hwang, **Suhyun Bang**, Tae-Hyeon Kim, Dong Keun Lee, Md. Hasan Raza Ansari, Seongjae Cho, and Byung-Gook Park, "A More Hardware-Oriented Spiking Neural Network Based on Leading Memory Technology and Its Application With Reinforcement Learning," *IEEE Transactions on Electron Devices*, Vol. 68, no. 9, Sept. 2021.

[3] Yeon-Joon Choi, **Suhyun Bang**, Tae-Hyeon Kim, Dong Keun Lee, Kyungho Hong, Sungjun Kim, and Byung-Gook Park, "Effect of Ag source layer thickness on the switching mechanism of TiN/Ag/SiN_x/TiN conductive bridging random access memory observed at sub-uA current," *Semiconductor Science and Technology*, Vol. 36, no. 5, 055014, Apr. 2021.

[4] Dong Keun Lee, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Yeon-Joon Choi, Kyungho Hong, Sungjun Kim, Seongjae Cho, Jong-Ho Lee, and Byung-Gook Park, "Improvement of Resistive Switching Characteristics of Titanium Oxide Based Nanowedge RRAM Through Nickel Silicidation," *IEEE Transactions on Electron Devices*, Vol. 68, no. 1, pp. 438-442, Jan. 2021.

[5] Yeon-Joon Choi, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Dong Keun Lee, Kyungho Hong, Chae Soo Kim, Sungjun Kim, Seongjae Cho, and Byung-Gook Park, "Insertion of Ag Layer in TiN/SiNx/TiN RRAM and Its Effect on Filament Formation Modeled by Monte Carlo

Simulation," *IEEE Access*, pp. 228720-228730, Dec. 2020.

[6] Dong Keun Lee, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Sungjun Kim, Seongjae Cho, and Byung-Gook Park, "Multi-Level Switching Characteristics of Si3N4-Based Nano-Wedge Resistive Switching Memory and Array Simulation for In-Memory Computing Application," *Electronics*, Vol. 9, no. 8, pp. 1-8, July 2020.

[7] Chandreswar Mahata, Changmin Lee, Youngseo An, Min-Hwi Kim, **Suhyun Bang**, Chae Soo Kim, Ji-Ho Ryu, Sungjun Kim, Hyoungsub Kim, and Byung-Gook Park, "Resistive switching and synaptic behaviors of an HfO_2/Al_2O_3 stack on ITO for neuromorphic system," *Journal of Alloys and Compounds*, Vol. 826, p. 154434, June 2020.

[8] Kyungho Hong, Kyung Kyu Min, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Dong Keun Lee, Yeon Joon Choi, Chae Soo Kim, Jae Yoon Lee, Sungjun Kim, Seongjae Cho, and Byung-Gook Park, "Investigation of the Thermal Recovery From Reset Breakdown of a SiN_x-Based RRAM," *IEEE Transactions on Electron Devices*, Vol. 67, no. 4, pp. 1600-1605, Apr. 2020.

[9] Dong Keun Lee, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Yeon-Joon Choi, Sungjun Kim, Seongjae Cho, and Byung-Gook Park, "HfO_x-based Nano-wedge Structured Resistive Switching Memory Device Operating at sub-μA Current for Neuromorphic Computing Application," *Semiconductor Science and Technology*, Vol. 35, no. 5, pp. 055002.1-10, Mar. 2020.

[10] Chandreswar Mahata, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Dong Keun Lee, Yeon-Joon Choi, Sungjun Kim, and Byung-Gook Park, "SiO_2 layer effect on atomic layer deposition Al_2O_3-based resistive switching memory," *Applied Physics Letters*, Vol. 114, pp. 182102, May. 2019.

[11] Dong Keun Lee, Min-Hwi Kim, Tae-Hyeon Kim, **Suhyun Bang**, Yeon-Joon Choi, Sungjun Kim, Seongjae Cho, and Byung-Gook Park,

"Synaptic behaviors of HfO_2 ReRAM by pulse frequency modulation," *Solid-State Electronics*, Vol. 154, pp. 31-35, Apr. 2019.

[12] **Suhyun Bang**, Min-Hwi Kim, Tae-Hyeon Kim, Dong Keun Lee, Sungjun Kim, Seongjae Cho, and Byung-Gook Park, "Gradual switching and self-rectifying characteristics of Cu/α-IGZO/p^+-Si RRAM for synaptic device application," *Solid-State Electronics*, Vol. 150, pp. 60-65, Dec. 2018.

[13] Min-Hwi Kim, Sungjun Kim, **Suhyun Bang**, Tae-Hyeon Kim, Dong Keun Lee, Seongjae Cho, and Byung-Gook Park, "Uniformity Improvement of SiN_x-Based Resistive Switching Memory by Suppressed Internal Overshoot Current," *IEEE Transactions on Nanotechnology*, Vol. 17, No. 4, pp. 824-828, May 2018.

[14] Tae-Hyeon Kim, Sungjun Kim, Hyungjin Kim, Min-Hwi Kim, **Suhyun** Bang, Seongjae Cho, and Byung-Gook Park, "Highly uniform and reliable resistive switching characteristics of a Ni/WO_x/p^+-Si memory device," *Solid-State Electronics*, Vol. 140, pp. 51-54, Feb. 2018.

[15] Sungjun Kim, Yao-Feng Chang, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Ying-Chen Chen, Jong-Ho Lee, Byung-Gook Park, "Ultralow power switching in silicon-rich SiN_y/SiN_x double-layer resistive memory device," *Physical Chemistry Chemical Physics*, Vol. 19, No. 29, pp. 18988-18995, Aug. 2017.

[16] Min-Hwi Kim, Sungjun Kim, **Suhyun Bang**, Tae-Hyeon Kim, Dong Keun Lee, Seongjae Cho, Jong-Ho Lee, and Byung-Gook Park, "Pulse area dependent gradual resistance switching characteristics of CMOS compatible SiNx-based resistive memory," *Solid-State Electronics*, Vol. 132, pp. 109-114, June 2017.

[17] Sungjun Kim, Sunghun Jung, Min-Hwi Kim, Tae-Hyeon Kim, **Suhyun Bang**, Seongjae Cho, and Byung-Gook Park, "Nano-cone resistive memory for ultralow power operation," *Nanotechnology*, Vol. 28, No. 12, pp. 125207-1-125207-7, Mar. 2017.

Conference

[1] **Suhyun Bang**, Tae-Hyeon Kim, Dong Keun Lee, Kyung Kyu Min, Sungjoon Kim, Yeon-Joon Choi, Kyungho Hong, and Byung-Gook Park, "Investigation of current sum error in the spiking neural network of RRAM synapse array," Korean Conference on Semiconductors, Feb. 2021.

[2] Yeon-Joon Choi, **Suhyun Bang**, Tae-Hyeon Kim, Dong Keun Lee, Kyungho Hong, Sungjun Kim, and Byung-Gook Park, "Effect of the Thickness of Ag-Inserted Layer on the Mechanism of TiN/Ag/SiN_x/TiN RRAM," Korean Conference on Semiconductors, Feb. 2021.

[3] Yeon-Joon Choi, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Dong Keun Lee, Chae Soo Kim, Kyungho Hong, Seongjae Cho, and Byung-Gook Park, "Investigation into the Effects of Ag Insertion Layer in TiN/SiN_x/TiN ReRAM through Monte Carlo Simulation," Korean Conference on Semiconductors, Feb. 2020.

[4] Kyungho Hong, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Dong Keun Lee, Kyung Kyu Min, Yeon Joon Choi, Chae Soo Kim, and Byung-Gook Park, "Effect of Post Annealing Process on SiN_x-based RRAM Operation," Korean Conference on Semiconductors, Feb. 2020.

[5] **Suhyun Bang**, Min-Hye Oh, Min-Hwi Kim, Tae-Hyeon Kim, Dong Keun Lee, Yeon-Joon Choi, Chae Soo Kim, Kyungho Hong, Seongjae Cho, Sungjun Kim, and Byung-Gook Park, "Validation of Spiking Neural Networks Using Resistive-Switching Synaptic Device with Spike-Rate-Dependent Plasticity," International Conference on Electronics, Information, and Communication (ICEIC), pp 373-376, Jan. 2020.

[6] Dong Keun Lee, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Yeon-Joon Choi, Sungjun Kim, Seongjae Cho, and Byung-Gook Park, “Effect of tunneling barrier layer insertion on HfO_2-based RRAM,” International

Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), June 2019.

[7] Dong Keun Lee, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Yeon-Joon Choi, Sungjun Kim, Seongjae Cho, and Byung-Gook Park, “Comparison of switching characteristics of HfO_x RRAM device with different switching layer thickness,” IEEE Silicon Nanoelectronics Workshop (SNW), June 2019.

[8] Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Dong Keun Lee, Sungjun Kim, Seongjae Cho, and Byung-Gook Park, “Improved Gradual Reset Phenomenon in SiN_x-based RRAM by Diode-Connected Structure,” IEEE Silicon Nanoelectronics Workshop (SNW), June 2019.

[9] Min-Hye Oh, **Suhyun Bang**, Min-Woo Kwon, and Byung-Gook Park, “A new device characteristic model generation by machine learning,” Electron Devices Technology Manufacturing (EDTM) conference, Mar. 2019.

[10] Yeon-Joon Choi, Min-Hwi Kim, Tae-Hyeon Kim, Dong Keun Lee, **Suhyun Bang**, Kyungho Hong, Chaesoo Kim, Sungjun Kim, and Byung-Gook Park, "Trap and Electron Occupancy Analysis in Si_3N_4 Resistive Switching Layer," International Conference on Electronics, Information, and Communication (ICEIC), pp 173-175, Jan. 2019.

[11] Dong Keun Lee, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Yeon-Joon Choi, Kyungho Hong, Chaesoo Kim, Sungjun Kim, Seongjae Cho, and Byung-Gook Park, "Scaling Effect of Ti/HfO_2/p^+-Si Stacked Resistive Switching device for Neuromorphic Application," 31st International Microprocesses and Nanotechnology Conference, Nov. 2018.

[12] Min-Hwi Kim, Sungmin Hwang, **Suhyun Bang**, Tae-Hyeon Kim, Dong Keun Lee, Sungjun Kim, Seongjae Cho, and Byung-Gook Park, "Sophisticated Conductivity Control of Gradual RRAM Cross-point Array for Reinforcement Learning," 31st International Microprocesses and Nanotechnology Conference, Nov. 2018.

[13] Yeon-Joon Choi, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Dong Keun Lee, Seongjae Choi, and Byung-Gook Park, "Reliability Prediction of a Wedge-Structured CBRAM through Monte Carlo Simulation," Asia-Pacific Workshop on Fundamentals and Applications of Advanced Semiconductor Devices (AWAD), July 2018.

[14] Dong Keun Lee, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Yeon-Joon Choi, Sungjun Kim, Seongjae Cho, and Byung-Gook Park, "Improvement of HfO_2-based ReRAM's Nonlinearity and Reset Current Level by Tunneling Barrier Layer Thickness Modulation," IEEE Silicon Nanoelectronics Workshop (SNW), June 2018.

[15] Dong Keun Lee, Min-Hwi Kim, Tae-Hyeon Kim, **Suhyun Bang**, Yeon-Joon Choi, and Byung-Gook Park, "Multi-Level Switching Characteristics of SiN_x-Based Nano-Wedge Resistive Switching Memory," 2017 Asia-Pacific Workshop on Fundamentals and Applications of Advanced Semiconductor Devices, pp. 333-335, July 2017.

[16] Tae-Hyeon Kim, Sungjun Kim, Sunghun Jung, Min-Hwi Kim, **Suhyun Bang**, Dong Keun Lee, and Byung-Gook Park, "Fabrication of Nano-cone RRAM and Analysis of its Electrical Concentration Effect," ChinaRRAM International Workshop, June 2017.

[17] Min-Hwi Kim, Sungjun Kim, **Suhyun Bang**, Tae-Hyeon Kim, Dong Keun Lee, Seongjae Cho, Jong-Ho Lee, and Byung-Gook Park, "Uniformity Improvement of SiN_x based Resistive Switching Memory by Suppressed Internal Overshoot Current," Silicon Nanoelectronics Workshop 2017, pp. 19-20, June 2017.

[18] Dong Keun Lee, Sungjun Kim, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, and Byung-Gook Park, "Fabrication of Nano-Wedge Resistive Switching Memory and Analysis on Its Switching Characteristics," Silicon Nanoelectronics Workshop 2017, pp. 55-56, June 2017.

[19] **Suhyun Bang**, Sungjun Kim, Min-Hwi Kim, Tae-Hyeon Kim, Dong-Keun Lee, and Byung-Gook Park, "Investigation on the RRAM Overshoot

Current Suppression with Circuit Simulation," Silicon Nanoelectronics Workshop 2017, pp. 81-82, June 2017.

[20] Sungjun Kim, Sunghun Jung, Min-Hwi Kim, Tae-Hyeon Kim, **Suhyun Bang**, Seongjae Cho, and Byung-Gook Park, "Si_3N_4-Based RRAM with Flexibility of Physical Design," 29th International Microprocesses and Nanotechnology Conference, pp. -, Nov. 2016.

[21] Kyung Kyu Min, Kyungho Hong, Chae Soo Kim, Yeon-Joon Choi, Tae-Hyeon Kim, **Suhyun Bang**, Dong Keun Lee, Min-Hwi Kim, Sungjun Kim, and Byung-Gook Park, "Characterization of CMOS Compatible HfO_x RRAM on Metal-Insulator-Semiconductor (MIS) Stacks," Asia-Pacific Workshop on Fundamentals and Applications of Advanced Semiconductor Devices (AWAD), July 2019.

[22] Kyungho Hong, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Kyungkyu Min, Dong Keun Lee, Yeon-Joon Choi, Chae Soo Kim, Sungjun Kim, and Byung-Gook Park, "Thermal Recovery from Reset Breakdown in SiNx-based RRAM," Nano Korea, July 2019.

[23] Tae-Hyeon Kim, Min-Hwi Kim, **Suhyun Bang**, Dong Keun Lee, and Byung-Gook Park, "Effects of pulse amplitude on conductance change of TiO_x-based synaptic devices," 한국반도체학술대회, Feb. 2019.

[24] Chae Soo Kim, **Suhyun Bang**, Tae-Hyeon Kim, Kyungkyu Min, Min-Hwi Kim, Dong Keun Lee, Yeon-Joon Choi, Kyungho Hong, Sungjun Kim, and Byung-Gook Park, "Design and Switching Characteristics Analysis of 3D Vertical W/SiN_x/n^+-Si/p-Si 1D1R unipolar RRAM," 한국반도체학술대회, Feb. 2019.

[25] 김민휘, 김성준, **방수현**, 김태현, 이동근, 조성재, 박병국, "실리콘 질화막 기반 저항 변화 메모리의 전압 의존적 저항 변화 특성 모델링," 하계종합학술대회, June 2018.

[26] Yeon-Joon Choi, Sungjun Kim, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Dong Keun Lee, Seongjae Cho, and Byung-Gook Park, "Close Investigation of Electric Field Concentration Effect in the Wedge Structure

through Numerical Analysis for Nanoscale ReRAM Application," 한국반도체학술대회, Feb. 2018.

[27] Dong Keun Lee, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Yeon-Joon Choi, Seongjae Cho, and Byung-Gook Park, "Neuromorphic Behaviors of HfO_2 ReRAM by Pulse Frequency Modulation," 한국반도체학술대회, Feb. 2018.

[28] Dong Keun Lee, Sungjun Kim, Min-Hwi Kim, Suhyun bang, Tae-Hyeon Kim, and Byung-Gook Park, "Area Dependency of Operating Voltages in Nano-Wedge Resistive Random Access Memory," NANO Korea, July 2017.

[29] Sungjun Kim, Min-Hwi Kim, Tae-Hyeon Kim, **Suhyun Bang**, Dong Keun Lee, Hee-Dong Kim, and Byung-Gook Park, "Scaling effect of device size on resistive switching characteristics in SiN-based RRAM," NANO Korea, July 2017.

[30] Min-Hwi Kim, Sungjun Kim, **Suhyun Bang**, Tae-Hyeon Kim, Dong Keun Lee, Seongjae Cho, and Byung-Gook Park, "Fabrication and Resistive Switching Characteristics of Silicon Nano-wedge Structure RRAM," 한국반도체학술대회, pp. -, Feb. 2017.

[31] **Suhyun Bang**, Sungjun Kim, Hyungjin Kim, Tae-Hyeon Kim, and Byung-Gook Park, "Investigation on Pulse Operation of Cu/IGZO/Si Structured ReRAM," 한국반도체학술대회, pp. -, Feb. 2017.

[32] Tae-Hyeon Kim, Sungjun Kim, Min-Hwi Kim, **Su-Hyun Bang**, Dong Keun Lee, and Byung-Gook Park, "Resistive switching characteristics of a Ni/WO_x/p^+-Si RRAM by pulse analysis," 한국반도체학술대회, pp. -, Feb. 2017.

[33] Tae-Hyeon Kim, Sungjun Kim, Min-Hwi Kim, **Su-Hyun Bang**, Dong Keun Lee, and Byung-Gook Park, "Reset Current Reduction for RRAM with Si Bottom Electrode," 2017 International Conference on Electronics, Information, and Communication, pp. 97-98, Jan. 2017.

[34] **방수현**, 김성준, 박병국, "멀티 레벨 셀 활용을 위한 터널링장벽이 삽입된 질화막 기반의 저항 변화 메모리의 점진적

저항변화 특성," 대한전자공학회 추계학술대회, pp. 110-112, Nov. 2016.

[35] 이동근, 김성준, 김민휘, **방수현**, 김태현, 박병국, "Si_3N_4 및 SiO_2 이중층 절연막 기반 저항 변화 메모리 구동에 compliance 전류가 미치는 영향," 대한전자공학회 추계학술대회, pp. 48-51, Nov. 2016.

[36] **Su-Hyun Bang**, Sungjun Kim, Hyungjin Kim, Tae-Hyeon Kim, and Byung-Gook Park, "Self-rectifying resistive switching Characteristics in Cu/IGZO/Si structure," International Technical Conference on Circuits/Systems, Computers and Communications, pp. 323-326, July 2016.

[37] 김태현, 김성준, 김민휘, **방수현**, 이동근, 박병국, "저전력 동작을 위한 실리콘 나이트라이드 기반의 저항 변화 메모리 소자에서 forming 전압이 미치는 효과," 하계종합학술대회, pp. -, June 2016.

[38] 김민휘, 김성준, **방수현**, 김태현, 이동근, 조성재, 박병국, "CMOS 호환 가능한 Si_3N_4 기반 저항 변화 메모리의 점진적 저항 변화 특성 분석," 하계종합학술대회, pp. 2557-2560, June 2016.

[39] Tae-Hyeon Kim, Hyungjin Kim, Sungjun Kim, **Su-Hyun Bang**, and Byung-Gook Park, "Resistive Switching Characteristics of RRAM with WO_x Switching Layer Prepared by Rapid Thermal Oxidation," 한국반도체학술대회, pp. -, Feb. 2016.

[40] **Su-Hyun Bang**, Sungjun Kim, Hyungjin Kim, Tae-Hyeon Kim, and Byung-Gook Park, "Synaptic Device based on Memristor using Cu/ITO/Si Structure," 한국반도체학술대회, pp. -, Feb. 2016.

[41] Sungjun Kim, Sunghun Jung, Hyungjin Kim, Min-Hwi Kim, **Suhyun Bang**, Tae-Hyeon Kim, Minju Yun, Hee-Dong Kim, Seongjae Cho, and Byung-Gook Park, "실리콘 산화막 기반의 저항변화메모리," 대한전자공학회 추계학술대회, pp. -, Nov. 2015.

Acknowledgements

First of all, I would like to express my sincere gratitude to my advisor Prof. Woo Young Choi. As well as, I would like to express my infinite thanks to Prof. Byung-Gook Park, who has supervised me since I entered graduate school. Their enthusiastic guidance and brilliant inspiration helped me successfully complete my Ph.D. degree, which I would never have been able to achieve on my own. I would also like to appreciate Prof. Sungjae Cho, one of the thesis examination committee, who helped me a lot in writing research papers and gave precious advice. I would also like to thank Prof. Jae-Joon Kim, Prof. Jong-Ho Lee, and Prof. Sungjun Kim for participating in the committee and for providing valuable feedback on my graduation research.

Next, I would like to thank my beloved parents for their generous and consistent support, both mentally and financially during my lengthy graduate school years. In addition, I appreciate my younger brother who silently supported me in starting his career beforehand.

Lastly, I am very grateful to my colleagues as seniors and juniors, and alumni of SMDL who have overcome difficult times together. Especially I would like to express my gratitude to alumni Tae-Hyeon Kim and Dong Keun

Lee, colleagues Yeon-Joon Choi, Sungjoon Kim, and Kyungho Hong in the RRAM team who shared the same research topic and conducted numerous discussions together. As laboratory mates and friends, SMDL members gave me a lot of motivation and helped me finish my degree without a hitch, making my graduate school life more meaningful.