4-bit Multilevel Operation in Overshoot Suppressed Al$_2$O$_3$/TiO$_x$ Resistive Random-Access Memory Crossbar Array

Sungjoon Kim, Jinwoo Park, Tae-Hyeon Kim, Kyungho Hong, Yeongjin Hwang, Byung-Gook Park,* and Hyungjin Kim*

To apply resistive random-access memory (RRAM) to the neuromorphic system and improve performance, each cell in the array should be able to operate independently by reducing device variation. In addition, it is necessary to lower the operating current of the RRAM cell and enable gradual switching characteristics to mimic the low-energy operations of biological. In most filamentary RRAMs, however, overshoot current occurs in the forming stage, and the RRAM shows large device variation, high operating current, and abrupt set and reset switching characteristics. Herein, the shortcomings occurring in the forming stage are overcome by introducing and optimizing an overshoot suppression layer. Consequently, the RRAM exhibits gradual switching characteristics both in the set and reset regions, thereby enabling implementation of 4-bit multilevel operation. In addition, the forming step can be easily performed in a $16 \times 16$ crossbar array owing to its self-compliance characteristics without disturbing neighboring cells in the array. The tuning and vector–matrix multiplication (VMM) operations are also experimentally verified in the array. Finally, classification performance with off-chip training is compared in terms of accuracy and robustness to tuning tolerance depending on the number of bits of the implemented multiconductance levels.

1. Introduction

The importance of artificial intelligence (AI) research based on neuromorphic systems has increased because of the limitations of serial computing systems known as von Neumann architectures. In particular, to implement AI in an existing computing system, a huge amount of serial computation is required, and computation time increases sharply, which is a major limitation to the expansion of AI functions.[1-10] Computation time is more problematic when iterative weight sum calculations are required, such as in deep neural networks.[11-13] Therefore, it is necessary to create synaptic devices that can be connected to one neuron in a massively parallel manner as can be seen in a human neural network, and it must be able to implement a weighted sum for continuously incoming signals. To fabricate such an electronic device, nonvolatile memory characteristics are essential, and various weight values are required for programming signals. Recently, various devices, such as resistive random-access memory (RRAM), phase-change random-access memory, ferroelectric tunnel junction, and charge-trap-based flash devices, have been studied as synapses to simultaneously implement nonvolatile and weight change characteristics.[14-51] Among them, RRAM can be made into a 4F$^2$ crossbar array with a simple structure of two terminals and has the advantage of high operating speed; therefore, many studies have been conducted on RRAM-based synaptic devices. However, an RRAM device that satisfies all desirable characteristics such as low power, multilevel, retention, endurance, and low parameter variation has not been developed yet.[14-17] For low power operation and suppression of line resistance effect, an RRAM should have a low operating current, which is quite difficult to obtain due to overshoot current during the forming process. Moreover, the overshoot current must be suppressed to prevent interference and destruction of neighboring elements.[38-41] A device capable of overcoming the loss of vector–matrix multiplication (VMM) accuracy is also required because of the current–voltage nonlinearity of RRAM.[42-44]

In this study, to approach the conditions required as an ideal synapse of a neuromorphic system, a structure suppressing the overshoot current is investigated to lower the operating current and improve the device variation. A 4-bit multilevel was also implemented using gradual switching and pulse-induced tuning algorithm, and the operation in a $16 \times 16$ crossbar array was experimentally verified. The 4-bit multilevel states were clearly defined as target conductance values equally spaced within the
device switching range, and each state was finely programmed by writing and verifying algorithms in the entire array. In our previous study, we confirmed that breakdown damage caused by overshoot current was minimized in the forming stage in an overshoot suppression layer (OSL) inserted device. Therefore, neither external compliance current (CC) nor an additional transistor is required. In addition, it can be used in large-scale array operation because the difference between the forming and set voltages is small. If the forming voltage is considerably larger than the set voltage, it causes a set or reset operation in nearby cells, or, in severe cases, produces a hard breakdown. However, in the case of the OSL RRAM device optimized in this study, because the forming and the set voltages are similar, the neighboring RRAM devices are rarely disturbed during the full array operations. Thanks to the self-compliance characteristics of the OSL device, all the electrical characteristics including forming, switching, and tuning operations of multilevel states can be experimentally conducted in the passively integrated crossbar array instead of single devices without selecting device such as diode or transistor. Finally, multilayer perceptron (MLP) neural network is trained with floating-point weight values and then transferred for modified national institute of standards and technology (MNIST) handwritten digit classification by off-chip training with multilevel conductance values of memristor devices and tuning algorithm.

2. Results and Discussions

2.1. Device Optimization

In our previous study, oxygen-rich TiO$_y$ was inserted as an OSL ($y$ value of approximately 1.81) to prevent overshoot current during the forming step, and its characteristics including the thickness effect of each layer were described. The device exhibits significantly abrupt switching characteristics and limiting compliance current is necessary without the OSL (Figure S1a, Supporting Information). In this study, we have optimized the OSL RRAM to reduce device parameter variation, enabling array operation and suitability for multilevel implementation. Although the characteristics of overshoot suppression were verified, the variation in the voltage range of overshoot suppression disturbs full array operations. One of the reasons for the large variation was that the OSL thickness is not uniform due to the scavenging effect of titanium, which is the adhesion layer of the top electrode and a part of oxygen reservoir TiO$_x$ layer. In the region where the OSL thickness is thinned by scavenging effect, filament formation can occur abruptly during the set switching process, and device-to-device switching uniformity can be degraded. Therefore, the titanium thickness target was reduced from 10 to 2 nm and the oxygen ratio in TiO$_x$ was raised from 1.54 to 1.72 ($x$ values in TiO$_x$) as shown in Figure 1. TiO$_2$ was deposited by reactive sputtering using a mixed gas of Ar and O$_2$, and the oxygen flow rate was increased by 20% in the optimized OSL device to reduce the proportion of Ti unbound with oxygen and Ti suboxide. To examine in detail how Ti and oxygen bonds are formed, X-ray photoelectron spectroscopy (XPS) analysis is performed as shown in Figure 1e. In the Ti 2$p_{3/2}$ graph, the binding energy of Ti$^0$, Ti$^{2+}$, Ti$^{3+}$, and Ti$^{4+}$ is measured to be 454.89, 455.93, 457.25, and 458.91 eV, respectively. From the result of deconvolution of Ti 2$p_{3/2}$ peaks, it can be observed that the proportion of Ti$^0$ is significantly reduced (from 19.11% to 14.03%) as the oxygen flow rate increases. The overshoot suppression effect is not uniform and stable when thickness fluctuation increases due to Ti scavenging as shown in Figure 1c, which

![Figure 1](https://www.advancedsciencenews.com/)

Figure 1. a,b) Schematic diagram of the OSL and optimized OSL RRAM structures. c,d) Transmission electron microscopy images of the OSL and optimized OSL RRAM devices. e) XPS analysis of TiO$_x$ (Ti 2p) according to different oxygen composition.
results in relatively abrupt and nonuniform switching characteristics (Figure S1b, Supporting Information). In contrast, when both Ti thickness and the oxygen ratio are optimized, the OSL thickness remains constant, and overshoot suppression properties are maintained in the overall devices as shown in Figure 1d, which leads to gradual and uniform switching characteristics (Figure S1c, Supporting Information). In addition, the overshoot suppression can be further enhanced with the lowered forming voltage by reducing the Al₂O₃ thickness from 2 nm (20 cycles) to 1.5 nm (15 cycles) in the atomic layer deposition (ALD) process. Switching characteristics cannot be observed due to extremely high tunneling leakage current if the Al₂O₃ thickness is reduced further below. Reduction in the Al₂O₃ thickness is also expected to improve the variation in the voltage range thanks to more symmetrical oxygen scavenging inside Al₂O₃ by TiOₓ.[48] Based on this optimized OSL device structure, 16/16 crossbar arrays are fabricated with the active region of 2.5 × 2.5 μm² defined by contact aligner lithography.

2.2. Switching Characteristics

In Figure 2a, the initial direct current (DC) sweep switching characteristics of the device structures without the OSL, with the OSL, and with the optimized OSL are compared. To verify the overshoot suppressed region, the voltage is increased until hard breakdown for all three devices. The device without the OSL suffers a hard breakdown near 2 V due to overshoot current and a reset switching operation could not be conducted. In the optimized OSL device, the voltage at which the filament starts to form is decreased as the Al₂O₃ thickness is decreased; therefore, the overshoot current flowing in the Al₂O₃ layer by the lowered forming voltage could be more easily suppressed by the OSL, and the overshoot suppressed region for self-compliance is increased. Figure 2b shows that the optimized OSL device has an overshoot suppression region of 1.7–2.9 V and the forming occurs in this region. The forming voltage of the optimized OSL device is uniform as 1.98 ± 0.13 V, while that of the nonoptimized OSL device is 2.76 ± 0.50 V with large variation as shown in Figure S2, Supporting Information, which make it difficult for full array operations. Although the initial current level is increased owing to the thin Al₂O₃ thickness of the optimized OSL device, the variation in the forming voltage becomes small and the controllability in gradual set switching is enhanced (Figure S1c, Supporting Information), which implies that the optimized OSL device exhibits more suitable characteristics for analog computing applications.

In addition, the set and reset sweep curves are compared to verify the electrical characteristics of the resistive switching properties for different RRAM structures as shown in Figure 3a. In the case of the reference device without the OSL, an external CC of 100 μA is used during the forming process. However, the
current level increases rapidly in the set and reset operations compared with the initial level because the overshoot current cannot be completely suppressed. In contrast, the OSL prevents the formation of an excessively large filament, and the set and reset current levels are reduced by one to two orders of magnitude compared with those of the device without the OSL. This is because the conductive filament in the OSL device can be oxidized more deeply during the reset process because the absolute voltage at which reset switching occurs in the OSL device is higher due to the voltage divide between the OSL.

![Figure 4](image_url)

Figure 4. a) Set and reset sweep characteristics of the optimized OSL device with changing sweep voltage by 20 mV. b) Read current results at 0.2 V according to the DC sweep count.

![Figure 5](image_url)

Figure 5. Conductive-AFM current scan images of the fabricated RRAM devices. a) RRAM device without OSL has a few strong filaments at LRS and b) OSL RRAM has multiple weak filaments at LRS. c) Schematic diagram of c-AFM measurement on the OSL RRAM. d) Comparison of LRS and HRS current depending on the active area of the devices with and without OSL.
and filament in the switching layer. As the OSL acts as an inherent compliance component, a transistor or an external component for compliance current is not required for the OSL device, which helps the full array operation of passive crossbar structure. Figure 3b shows the results of the forming, first set, and reset curves of 40 optimized OSL devices; here, the operating ranges of the forming, set, and reset voltages between devices are not significantly different (<0.5 V). This is because the thickness variation of the OSL is decreased as the thickness of the titanium adhesion layer that causes oxygen scavenging becomes thinner (2 nm), and the forming voltage is reduced by optimizing the Al2O3 thickness. Thus, the filament size could be small, and the current level remains similar before and after the forming step.

To examine the gradual switching characteristics according to the switching voltage, the DC switching characteristics of the optimized OSL device are measured with changing the sweep voltage by 20 mV as shown in Figure 4a. It is confirmed that the set operation occurs gradually from approximately 1.6 V. Following the set operation, the reset operation is also conducted while increasing the voltage from ~0.5 V, and it is confirmed that the current level gradually decreases from ~2 V. The reset operation no longer occurs when it exceeds ~2.7 V. DC sweep proceeds in set–reset order and read is performed at 0.2 V after one sweep is completed. Figure 4b shows the read current results according to the DC sweep order, and it can be seen that the set and reset operations start from 1.6 and ~2 V, respectively. After reaching the resistive switching voltage, the current value is gradually modulated according to the voltage change.

In addition, conductive-AFM (c-AFM) analysis (Park systems NX10, tip radius <25 nm) was performed to experimentally prove the conduction mechanism of the OSL RRAM structure for the programmed high resistance state (HRS) and low resistance state (LRS) as shown in Figure 5. The RRAM devices with and without the OSL were compared to verify the switching and self-compliance mechanisms of the OSL device. To switch each device to LRS, a 250 × 250 nm² area was scanned by applying 2.5 V to the tip and then scanned again with a read voltage of 0.5 V to check where the conductive filaments were formed as illustrated in Figure 5c. As shown in Figure 5a, the reference device without the OSL has several strong filaments exceeding the current limitation (10 nA) of the c-AFM tool at LRS, which implies that the hard breakdown can occur during forming or set operations in the reference device without compliance current. In addition, the reference device has the little dependency of the device area on the operating current because only a small part of the device area is used for conduction, which means that it is hard to scale down the device current by the device area. On the other hand, relatively weak multiple filaments (<10 nA) are formed and widely distributed all over the device area at the LRS of the OSL device as shown in Figure 5b. These results lead us to believe that the operation current can be controlled by the device area, and the gradual switching behaviors of the OSL devices are originated from the fact that multiple weak filaments

---

**Figure 6. a–d) Gradual switching characteristics according to the pulse width of 100 ns to 100 µs in ISPP (read pulse width 100 µs).**
distributed over the device area are gradually formed and then ruptured according to set and reset voltage changes. In addition, the relatively abrupt and ununiform switching characteristics of the unoptimized OLS device in Figure S1b, Supporting Information, can also be understood with the c-AFM results. When the OSL thickness fluctuates significantly, relatively strong filaments can be formed stochastically at the area with thin OSL, and irregularly abrupt switching characteristics can occur (Figure S3e–h, Supporting Information). It is also experimentally confirmed with varying the device area from 6.25 to 100 μm² that the OSL device has a strong area dependency on the operating current, while the reference device without the OSL is rarely affected by the device area due to their own switching mechanisms as shown in Figure 5d. It is preferred that the operational current of synaptic device can be scaled down according to the device area because a large-size neural network would cause extremely high weighted current sum at output node if not.

In short, the gradual switching mechanism of OSL RRAM has been verified as multiple weak filaments through three methods. First, the gradual curve in the overshoot suppressed region of the optimized OSL RRAM shown in Figure 2b implies that multiple weak filaments can be formed stochastically at the area with thin OSL, and irregularly abrupt switching characteristics can occur (Figure S3e–h, Supporting Information). These results lead us to believe that the multiple weak filaments are distributed over the device area, enabling the gradual switching properties of the OSL device.

2.3. Multilevel Operations

To increase the accuracy of the neuromorphic system, the RRAM synaptic device must have a weight of 3-bit or more. However, in the case of RRAM with abrupt characteristics of set and reset, it can only be used as a binary of 1 or 0, and at least three RRAM devices are required to implement 3-bit operation. Therefore, if the number of weights that one device can have increased, the inference accuracy can be enhanced accordingly. The optimized OSL RRAM device can have multiple current levels by adjusting the set and reset voltages, as described above, and the number of weights that can be represented in the pulse operation is confirmed (hereinafter, optimized OSL devices are referred to as OSL devices). First, we verify the gradual switching

![Figure 7](image-url)

**Figure 7.** a) 4-bit multilevel characteristics of the OSL devices measured with ISPVA scheme. b) Number of pulses used to move one conductance state. c,d) ISPVA operation characteristics of pristine to 1 μA and 500 nA.
characteristics according to the switching speed of the device while varying the set and reset pulse widths from 100 ns to 100 μs (Figure 6a–d); here, an incremental step pulse programming scheme (ISPP) is used for resistive switching. The increments of set and reset are set to 20 and 10 mV, respectively, and the read voltage is set to 0.2 V. This facilitates the gradual set and reset operations even with a short pulse width of 100 ns, although a large current variation occurs due to randomly distributed oxygen vacancies in the switching layer between each pulse step.[50]

Even a large current variation could be adjusted to the desired conductance level by using an incremental step pulse with a verification algorithm (ISPVA), but the number of iteration cycles would increase. In the pulse width from 1 to 100 μs, the current variation between pulses is relatively small. However, for pulse lengths longer than 100 μs, the current tends to decrease rapidly as the absolute reset voltage increases. Therefore, these values are not suitable for programming various current levels precisely and the pulse width should be set to values higher than 100 ns and lower than 10 μs to implement multiple levels through the OSL device.[51,52]

After verifying the pulse width effect, we evaluate whether a 4-bit multilevel conductance could be realized with ISPVA in the OSL RRAM. When using ISPVA, a read pulse is applied between programming pulses to verify device state and compare with the target conductance. A reset pulse is applied if the measured conductance is higher than the target value, while the amplitude of programming pulse is increased and applied again when the measured conductance is still lower than the target conductance within the specified tolerance. The pulse width is set to 10 μs, and the read voltage is set to 0.2 V. Under the DC sweep condition, the OSL RRAM operates within the HRS–LRS range of 30 nA to 2 μA at 0.2 V. For the off-chip training manner, accurate tuning operation is required rather than linear weight-update characteristics because the system performance can be close to

Figure 8. a) Room temperature retention for 4-bit conductance states. b) 85 °C retention for four conductance states. c) Cumulative distribution of 4-bit conductance characteristics of 100 devices. d) Endurance characteristics measured by moving to different conductance states every 1000 cycles. e) Endurance characteristics measured by repeatedly tuning 4-bit from G1 to G18 and from G18 to G1 in sequence.
software-based approach as long as pretrained weight values can be finely transferred into synapse array. Based on the measurement data, equally spaced 18 levels (G1 to G18) with 50 nA intervals are set within the range of 100 nA to 1 μA. The error tolerance in terms of current is set to ±20 nA such that each conductance level does not overlap, and the possibility of multilevel operation is checked even in the repetitive resistive switching situation. Figure 7a shows the value when each conductance level is moved by one level using ISPVA for 70 cycles. Consequently, multilevel operation is possible within the ±20 nA error tolerance in all the conductance states. In Figure 7b, the number of pulses consumed in each cycle is plotted, and a median of 36 pulses (average of 49 pulses) is required when moving from one level to another. The number of pulses required for level movement is not necessarily proportional to the distance between conductance levels. For example, to adjust from G18 to G1, on average, approximately 60 pulses are needed. To realize a neuromorphic system operating at low power, it is necessary to reduce the number of write and read pulses applied to program a specific conductance level, and the current flowing through the RRAM synaptic cell should be low. Therefore, the number of iterations must be reduced by tuning the method of applying a pulse. Figure 7c,d shows the process of programming from the pristine state to the LRS of 1 μA and 500 nA, respectively. The ratio of conductance programming at the set mode is 59.5% for LRS of 1 μA and 54.7% for LRS of 500 nA. Therefore, we can confirm that both the set and reset modes exhibit gradual switching characteristics and the target conductance is reached rapidly.

To implement a reliable multilevel operation, the retention characteristics in 20 states (30 nA to 1 μA) are verified as shown in Figure 8a, and it is observed that all the states are maintained up to 10⁶ s. In addition, the high-temperature retention characteristics of the four states at 85 °C were measured as shown in Figure 8b, which confirms that the states were distinctly maintained for 10⁵ s and each state can be read stably. The retention characteristics were measured at one point every 90 s after adjusting each conductance level with ISPVA. As shown in Figure 8c, the states are tuned using ISPVA to have 4-bit states for 100 cells in the 16 × 16 array, and it is expressed as cumulative probability, which verifies that 4-bit operation with ±20 nA tolerance can be conducted in the whole array. Similarly, endurance properties are confirmed by moving from different states to the LRS for 1000 cycles as shown in Figure 8d. Accordingly, the resistive switching operation to target a specific conductance level could be performed more than 4000 times. Approximately 1.2 × 10⁴ pulses are utilized to reach each state 4000 times. Finally, in order to check whether the repetitive 4-bit multilevel operation is performed, 34 states corresponding to G1 to G18 and G17 to G2 was programmed 150 times in sequence. A total of 5100 conductance state shifts were performed, 5.8 × 10⁴ pulses were applied for conductance state tuning, and 4-bit operation was possible until the last switching as shown in Figure 8e.

Figure 9. a) Photomicrograph of the fabricated 16 × 16 crossbar array. b) Schematic diagram of reading out all cells connected to each BE at once. c) Visualization result of read current after switching a specific cell to the LRS in the 16 × 16 array. d) Comparison of read results at the current sum of each cell connected to BE and the line current of I_B1 to I_B16.
The RRAM synapses are generally integrated into a crossbar-type array with the advantages of a small area of $4F^2$ and a simple process. A $16 \times 16$ crossbar array with the optimized OSL devices is fabricated as shown in Figure 9a and examined to verify whether all the devices in the whole array can adequately perform forming, reset, and set operations. In addition, to properly implement VMM from the input stage, it is necessary to check the line resistance effect, which can affect the RRAM cell at the read step. If the difference between the cell and line resistances is not sufficiently large, a read error can occur between the current sum of each cell and the line current. When comparing these values, the error should be small to enhance the recognition judgment error in neuromorphic systems. As shown in Figure 9b, the read voltage of 0.2 V is applied to top electrodes (TEs) 1–16, and 0 V is applied to each bottom electrode (BE) to read each BE line at once while a separate GND is connected to the other BEs. After forming all the cells in the array, the set and reset operations are performed (Figure S4, Supporting Information). In the forming stage, the suppressed overshoot features have an important role. In the case of devices that require compliance in the forming stage, the suppressed overshoot features have an important role.

In the case of the OSL devices, there is a slight difference in forming voltage (1.98 ± 0.13 V) and set voltage (1.80 ± 0.18 V) with the suppressed overshoot current (Figure S2, Supporting Information), and other cells in the $16 \times 16$ array are not destroyed during the forming process. It is verified that the yield of cells which can be switched for the set and reset operations is 99.6% after the forming procedure (Figure S5, Supporting Information). When all the cells are in the HRS and only specific cells are set, as shown in Figure 9c, it is possible to write only the desired cells without significantly affecting the surrounding cells. The line current values from each line of $I_{B1}$ to $I_{B16}$ with the programmed states as in Figure 9c are compared with the current sum of each cell as shown in Figure 9d, and it is confirmed that the $R^2$ value of the trend line is 0.9995 and a precise VMM operation can be realized in the array with the OSL devices. However, the line current should not be neglected when the array size is considerably larger; therefore, it is necessary to further reduce the current level flowing through the RRAM cell and increase the difference from the line resistance.

Finally, the off-chip learning by the tuning operation of multilevel memristor devices is verified with $28 \times 28$ MNIST handwritten digit classification. A fully connected neural network ($784 \times 512 \times 10$) was trained first by software-based manner having the training and test accuracy as 99.99% and 98.18%, respectively, after 100 epochs of training as shown in Figure 10a. Then, the pretrained weight values are quantized and transferred by assuming that one synapse consists of two memristors to represent both positive and negative weight values as shown in Figure 10b. The classification accuracy assuming zero tuning
error is summarized depending on the quantized weight levels to 6-bit operations as shown in Figure 10c. It is found that the accuracy drops significantly to 91.95% when using 3-bit quantization, while it rarely drops to 98.04%, 98.06%, and 98.13% when using 4-bit, 5-bit, and 6-bit quantization, respectively. It confirms that 4-bit tuning operation of the memristive weight values (Figure S6, Supporting Information) is enough to suppress weight distortion and minimize the classification accuracy drop. Even though the accuracy can be slightly increased when 5-bit and more, a lower tuning tolerance (error) is required in return due to the narrowed window between each conductance state, leading to the fact that the number of pulses required to reach a target conductance is increased (Figure S7, Supporting Information). The tuning tolerance is further considered in the classification accuracy with varying ±10 to ±50 nA as shown in Figure 10d. When the tuning tolerance (±10 nA) is lower than the weight margin between neighboring weight states (50 nA), the average accuracy remains still high with an extremely small variation (97.80 ± 0.10%). However, the average accuracy is degraded significantly with a large variation (85.16 ± 4.79%) when the tuning tolerance (±50 nA) exceeds the weight margin because the weight transition to adject state can occur and the pretrained neural network can be distorted seriously. It is suggested that ±20 nA of tuning tolerance can be a compromised allowance considering the trade-off between tuning accuracy and the required number of pulses for tuning operation.

3. Conclusion

In this study, it is verified that the OSL prevents the hard breakdown of the switching layer in the forming stage and also serves as a current limiter. Therefore, the OSL RRAM has self-compliance characteristics and can be operated at a low current. The device variation is optimized by reducing the scavenging effect of the titanium adhesion layer used in TE, and the forming and set voltages are maintained almost identical by lowering the Al2O3 switching layer thickness. Therefore, when the forming voltage is applied to the cells inside the array, the OSL reduces overshoot damage and minimizes the destructive effect on the surrounding cells. In addition, because the filament is not formed abruptly, it reaches the desired conductance level in both reset and set operations. To prove this gradual switching mechanism of OSL RRAM, c-AFM was measured, and multiple weak filaments in OSL structure contributed to realizing gradual switching. In addition, as the entire device area contributes to the current conduction through the formation of multiple weak filaments, the operating current can be further reduced by scaling. Consequently, a multi-level implementation of 4-bit or more is possible using ISPVA. The programming operation of a 16 × 16 crossbar array is also demonstrated and the precise VMM operation is confirmed. These results lead us to conclude that the OSL can facilitate the reduction of device variation and enable large-scale array operation using filamentary RRAM with various materials. Furthermore, it is verified that the real weight values which are pretrained in the designed MLP can be used with little classification accuracy drop for MNIST even with 4-bit weight quantization. The classification accuracy drops when the tuning tolerance exceeds the weight margin between neighboring states and there is a trade-off between tuning accuracy and the required number of pulses.

4. Experimental Section

Device Fabrication: First, a 200 nm-thick SiO2 layer was deposited on the 4 inch p-type silicon wafer for isolation between the Si substrate and the bottom electrode of the memristor device. Then, a Ti/Pt (8.5 nm/50 nm) BE layer was deposited by the e-gun evaporator. To pattern the BE and TE, we used the i-line (365 nm) contact aligner and the lift-off process. After the BE process, an Al2O3 (1.5 nm, 15 cycles) layer was deposited by the ALD as the switching layer. Followed by a TiOx layer (17 nm) as the oxygen reservoir, TiO2 (4.5 nm) as the overshoot suppression layer was deposited by the DC reactive sputter using the Ti target under the Ar and O2 atmosphere. The O2 flow rate was controlled to differentiate the oxygen composition in the TiO2 and TiOx layer. Finally, a Ti/Pt (2 nm/80 nm) TE layer was deposited by the e-gun evaporator. The active region in the OSL memristor array was formed as the crossbar structure (2.5 × 2.5 μm2) of the BE and TE.

Electrical Characterization: All measurements were performed with Keysight B1500A, and pulse measurements were performed using the WFGMU module. In addition, to operate the cells in the 16 × 16 array, Keysight’s ES250A switching matrix (low-leakage switch mainframe) was used. A customized C++ code was used to operate an incremental step pulse with a verification algorithm.

Software Algorithm: The software algorithms for learning 28 × 28 MNIST handwritten digit classification is as follows: minibatch stochastic gradient descent (SGD) method, 100 of minibatch size, 0.1 of initial learning rate, and 0.99 of exponential learning rate decay parameter. Minibatch SGD means that a method in which the weights are updated by stochastically selecting loss as much as the minibatch size. During training, 0.1 of the initial learning rate is multiplied by 0.99 for every epoch, and the learning rate reflected in the update is scaled down to induce stable learning. The initial weight distribution is set as a uniform default from −0.3 to 0.3. All these simulations are performed using Python code-based library provided by Pytorch.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This work was supported in part by the NRF grant funded by the Korean government (grant nos. 2020M3H5A1081111, 25%, 2021R1C1C1014530, 25%, and 2021M3J3A2A01037927, 25%), in part by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (grant no. IITP-2021-0-02052, 25%) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), in part by the Brain Korea 21 Four Program, and in part by Samsung Electronics Co., Ltd. (grant no. I0201228-08268-01).

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.
Keywords
crossbar arrays, memristor, multilevel operation, overshoot current suppression, tuning

Received: December 21, 2021
Revised: April 27, 2022
Published online: May 26, 2022