IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 20, 2022, accepted March 2, 2022, date of publication March 8, 2022, date of current version April 11, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3157333

A Fast Weight Transfer Method for Real-Time Online Learning in RRAM-Based Neuromorphic System

MIN-HWI KIM[1,2], (Member, IEEE), SIN-HYUNG LEE[3], SUNGJUN KIM[4], (Member, IEEE), AND BYUNG-GOOK PARK[1], (Fellow, IEEE)

[1]Inter-University of Semiconductor Research Center (ISRC) and Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, South Korea
[2]Memory Business Division, Samsung Electronics Company, Hwaseong-si, Gyeonggi-do 18448, South Korea
[3]School of Electronics Engineering, School of Electronic, and School of Electronic and Electrical Engineering, Kyungpook National University, Daegu 702-701, South Korea
[4]Division of Electronics and Electrical Engineering, Dongguk University, Seoul 04620, South Korea

Corresponding authors: Sungjun Kim (sungjun@dongguk.edu) and Byung-Gook Park (bgpark@snu.ac.kr)

This work was supported by the Korea Government (Ministry of Science and ICT) through the Institute of Information & Communications Technology Planning & Evaluation (IITP) Grant (2020-0-01294).

ABSTRACT In this work, a synaptic weight transfer method for a neuromorphic system based on resistive-switching random-access memory (RRAM) is proposed and validated. To implement the on-chip trainable neuromorphic system which utilizes large-scale hardware synapse units, a fast and reliable write scheme needs to be established. Based on the experimental results, it is confirmed that the gradual set and full reset operation is the most suitable operation scheme for fast programming due to the fundamental reliability characteristics of the resistive-switching memory cell. Also, the superiority of this programming method using the proposed RRAM compact model is demonstrated. In addition, a one weight/one synaptic device structure is newly adopted for realizing high-density synapse arrays by using a nonnegative weight constraint in supervised learning. Finally, the pattern recognition accuracies obtained at the software and hardware levels are compared.

INDEX TERMS Neuromorphic, hardware-driven artificial intelligence, synaptic device, weight transfer, resistive-switching random-access memory (RRAM), artificial neural network (ANN), cross-point array architecture.

I. INTRODUCTION

Numerous studies have been conducted in academia and industry to imitate the limitless cognitive abilities of the human brain to learn, remember, infer, and forget in an incredibly energy-efficient and natural way [1]. AlphaGo once again opened a door to a new stage of artificial intelligence by introducing an elaborately and systematically trained artificial neural network (ANN) model [2]. The latest deep neural network concept and its learning algorithm should be highly evaluated not only for their learning and inference ability, but also for reproducibility. Recently, Intel announced a spike-event-based neuromorphic chip called Loihi2, which doubles the synapse density and operates 5,000 times faster than the biological neuron [3]. Furthermore, Samsung, which is one of the leading companies in the semiconductor memory business, presented the concept of delivering the structure and function of a human brain into a high-density memory chip [4].

A schematic diagram of a biological neural network is shown in Fig. 1(a). Biological neurons that operate based on the integrate-and-fire mechanism to transmit weighted signals through the synapse region and also the synaptic connections and their long-/short-term plasticity are known to play the most important role in the learning and memory functions of a human brain and various studies which implement those functionalities into electronic systems have been reported [5]–[11]. The conceptual structure of a simple ANN (Fig. 1(b)) is deeply inspired by the biological neural network, and this structure has been the basis of most of the well-known neural networks [12]. Various nonvolatile memory array structures can be considered to

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

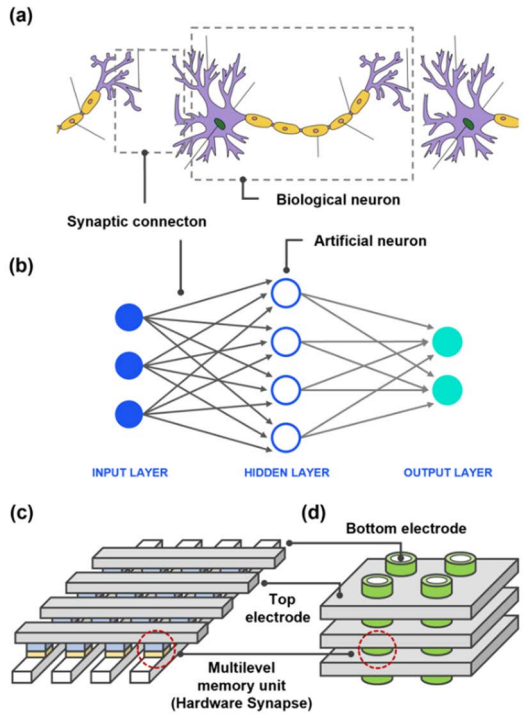

FIGURE 1. Schematic diagrams of (a) a biological neural network and (b) an artificial neural network. (c) A 2D cross-point array structure and (d) a 3D vertical array structure of a nonvolatile resistive memory.

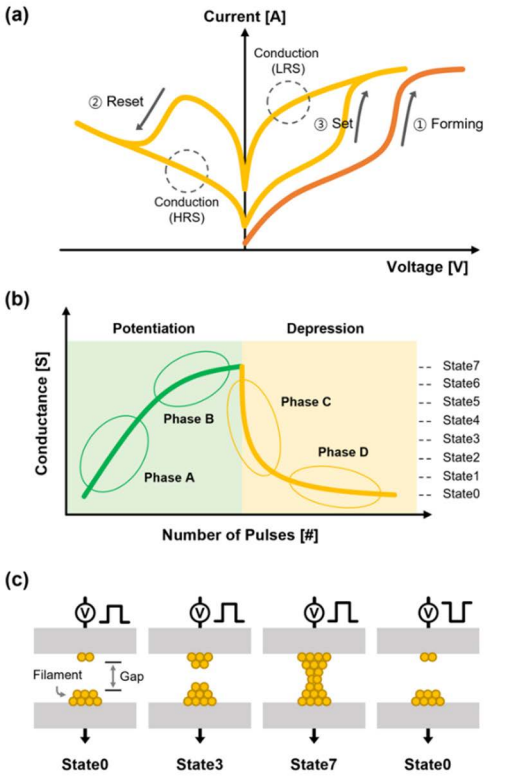

FIGURE 2. (a) Typical *I–V* curves of a multilevel resistive switching memory. From the initial HRS, a memory cell switches to LRS and shows different conduction characteristics based on the materials and resistance levels. (b) In pulse operation, the conductivity is increased by consecutive set pulses and is decreased by consecutive reset pulses. (c) Conductive filaments generally consist of atomic vacancy regions or electron traps. The gap between two electrodes is effectively shortened by set pulses and is ruptured by reset pulses.

realize the connectivity and synaptic plasticity of neural networks at the hardware level, but the cross-point array structure adopting the two-terminal resistive-switching random access memory (RRAM) has been the most actively discussed.

For ultrahigh-density applications, a 3D and vertically stacked structure that is similar to the recent 3D NAND Flash memory may also be considered as shown in Fig. 1(c) [13]. Multilevel conductivity states and long-/short-term memory characteristics that can be realized within a highly scalable cell structure have made RRAM favored by many researchers. However, the switching operation based on the soft breakdown of a switching layer and the read operation that relies on direct charge flows through the switching layer are always a concern for this memory device, which result in reliability issues such as uniformity and endurance [14]–[17]. Although several approaches considering switching layer engineering, pulse operation scheme, and unit memory structure (1R, 1S1R, 1T1R, etc.) have been proposed and investigated to improve the reliability, they still do not reach the industrial requirements level. At the same time, a thorough study on the write method of a synapse array considering the device characteristics such as multilevel state and reliability, as well as the memory architecture, is required for a high-density and high-speed neuromorphic system. A fully parallel write method in a resistive synapse array can be a suitable option due to its as speed performance and parallelism, but it has some limitations [18].

First, it is necessary to check the state of preneuron and postneuron nodes when performing quantitative weight update before each write operation complicating the write operation itself. In addition, it has difficulty utilizing various techniques such as the incremental-step-pulse programming scheme and verify/inhibit operation because it uses pulses with a fixed voltage amplitude and is also based on the assumption that the conductivity change is proportional to the overlapped pulse width. In this work, a fast write method for hardware transfer of ideally trained synaptic weight from an ANN is proposed and validated. For the validation, an RRAM compact model is adopted into a cross-point array and adjusted to fit the device characteristics. For RRAM cells showing voltage-dependent switching characteristics, a sequence of write operation utilizing gradual set and full reset (GSFR) is proposed and verified through SPICE simulations. Furthermore, a one weight/one synaptic device (1W1S) implementation is adopted using a nonnegative weight constraint in software training to realize a high-density synapse array. Finally, the pattern recognition accuracy of the multilevel conductance synaptic array is compared with that of a software-based ANN.

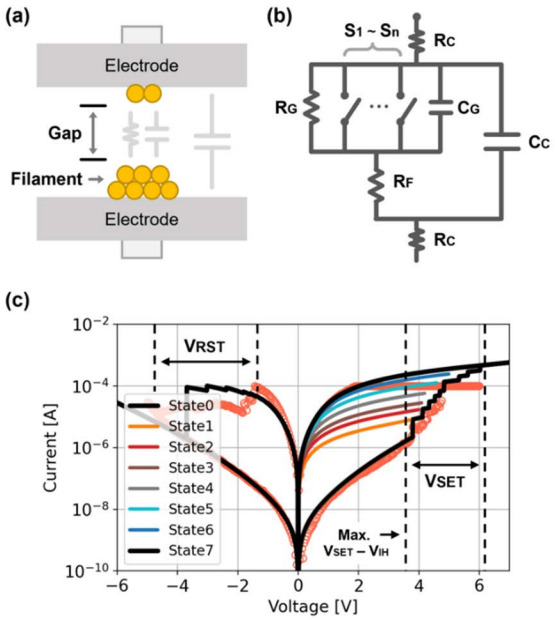

FIGURE 3. (a) Schematic diagram of a resistive memory cell structure. The yellow circles are electron traps mostly formed close to electrodes. (b) Circuit model which consists of several circuit elements such as resistors, capacitors, and voltage-controlled switches. (c) DC characteristics obtained from measurement (orange circle) and our circuit model (solid lines). By controlling the stop voltage from 3.8 to 5.0 V, the conductivity can be determined within a certain range.

TABLE 1. Circuit element and their value and description for unit cell modeling.

Element	Value	Description
S_1~S_n (Voltage-controlled switches)	V_T (Threshold voltage): 1.0 ~ 1.74 V V_H (Hysteresis voltage): 1.8 ~ 3.7 V R_{ON} (On resistance): 30 ~ 500 kΩ R_{OFF} (Off resistance): 10^{12} Ω	Gap resistance (HRS to LRS)
R_G (Nonlinear resistor)	$5 \cdot 10^4 \cdot \exp[-(\text{abs}(V_{GAP}) - 0.8) / 0.8]$ kΩ	Gap Resistance (HRS)
R_F (Nonlinear resistor)	$10^1 \cdot \exp[-(\text{abs}(V_{FIL}) - 1.0) / 0.3]$ kΩ	Filament Resistance
R_C (Linear resistor)	100 Ω	Contact resistance
C_C (Capacitor)	0.1 fF	Cell Capacitance
C_G (Capacitor)	0.1 fF	Gap Capacitance

II. SPICE COMPACT MODEL FOR MULTILEVEL RRAM

A. RRAM CELL CHARACTERISTICS

An RRAM typically shows the bipolar switching phenomenon, which is the transition of the cell resistance state from high resistance state (HRS) to low resistance state (LRS) and vice versa under the opposite voltage polarity (Fig. 2(a)). The switching voltage may vary depending on the switching layer material, the thickness, and the combination with the electrode materials [19], [20]. Some devices exhibit gradual set or reset switching over a specific voltage range, and the characteristics can be used to obtain multilevel resistance states using pulse width/amplitude modulation to store multibit data. Each resistance state has its own conduction mechanism and represents different I-V behaviors depending on the switching material and the presence of the conducting filament. The conductivity of the RRAM cell can be gradually modulated by applying appropriate pulses [21]. In studies on memory-based neuromorphic systems, the phenomena of conductivity increase and decrease are commonly called long-term potentiation and depression, respectively, which are named after the biological synaptic plasticity characteristics. In phase A in Fig. 2(b), the conductivity is increased uniformly by positive set pulses, and in phase B, the amount of change begins to decrease due to degradation of the switching efficiency. By adopting the incremental-pulse scheme, the phase-A-like behavior can be extended but will have limitations in the end. Phase C represents the initial stage of the conductivity decrease by reset pulses. Similarly, when entering phase D, the amount of change also begins to decrease. It is known that the conductivity change behavior by an identical pulse can be asymmetric in potentiation and depression, and various studies to modulate the characteristics at the device level have been reported [11]–[23]. The shape of the filament corresponding to each resistance state (states 0–7) can be described as shown in Fig. 2(c). All traps or vacancy regions involved in the conduction expand further under a certain condition, eventually connecting both electrodes. Also, under a certain condition, the connected dense filament regions can be ruptured, which is known to be mainly related to Joule heating [15]–[17].

An RRAM device is considered as one of the plausible candidates for a synaptic device in a neuromorphic hardware system because of its scalability and simple process. Gradual resistance changes of RRAM cells can be used to simply express the connection strength of biological synapses in hardware, and it is also necessary to implement an on-chip learning method such as spike-timing-dependent plasticity (STDP) in a unit device [24]. At the same time, low energy consumption in an RRAM-based synaptic device is necessary because a large number of synaptic devices operate simultaneously and in parallel even if an ideal 1W1S structure is configured. Few studies have been reported to overcome this problem by adding a thin dielectric film to suppress the operating current [25], [26]. Similarly, the RRAM device used in this work was able to suppress the operating current to μA-level by using a thin SiO_2 as a tunnel barrier layer. Detailed explanations of the process flow and the operating principles can be found in our previous work [27].

B. RRAM CELL MODELING FOR SPICE SIMULATION

It is necessary to implement the switching and conduction characteristics of RRAM cells at the circuit and system levels for the development and verification process. Several studies have been conducted and reported to realize the general characteristics of RRAM cells [28]. A SPICE compact model was also proposed, and it has been successfully applied to

various RRAM devices in a cross-point array [19]. Fig. 3(a) represents the RRAM cell structure. In general, the switching layer of a filamentary-type RRAM cell is divided into a filament and other region, and the filament region is connected or broken depending on its resistance state. It has been reported that the filament region may be an oxygen vacancy for metal oxide materials or an electron trap for silicon nitride materials, which plays a major role in charge conduction. Fig. 3(b) shows the proposed RRAM cell model for SPICE circuit simulation. Linear and nonlinear resistors are used to realize the conduction behavior of the memory device. Also, several voltage-dependent switches are used to obtain the voltage-dependent switching characteristics. The I–V characteristics of our device and the model are represented together in Fig. 3(c). The parameters are elaborately adjusted to implement the voltage-dependent switching characteristics in the range of –6.0 to 6.0 V, which are importantly utilized in the write scheme proposed in this work. For example, the values of a nonlinear resistor R_G and the voltage-controlled switches S_1–S_n are determined to describe more accurately the conduction behavior through the rupture and the connection of conduction paths. Also, a nonlinear resistor R_F is placed in series and adjusted to describe the LRS conduction behavior after the initial forming process. Although it is necessary to determine a linear resistance R_C considering the magnitude and variation of the contact and line resistances in the actual array configuration, it was not considered in this study. In addition, capacitive elements such as the cell (C_C) and gap (C_G) capacitances need to be elaborately determined, but they are beyond the scope of this study and does not affect the experimental results. Table 1 summarizes the circuit components and parameters and their values in the model.

III. FAST WRITE SCHEME FOR REAL-TIME LEARNING OF HARDWARE NEURAL NETWORK

Each layer of a neural network trained at the software level may contain a lot of continuous or quantized weight values. To ensure the software-level inference accuracy, it is necessary to accurately transfer the trained weight matrix into the hardware synapse array while suppressing nonideal phenomena such as voltage drop by line resistances, read disturb, and leakage current. In this part, a write method that transfers the weight matrix into synapse arrays for high density and high-speed applications is proposed and verified.

To obtain multilevel resistance states and reach the target state, the following schemes can be considered: gradual set and full reset (GSFR), full set and gradual reset (FSGR), and gradual set and gradual reset (GSGR). Although the GSGR method, which takes full advantage of the RRAM's bidirectional gradual switching characteristics, seems the most attractive, but it needs to be reconsidered from two perspectives. First, negative voltages that are repeatedly applied can cause serious reliability problems. For example, Fig. 4(a) shows that reset switching failure can easily occur in pulse operation by repeated applied negative voltage pulses. Distributions of conductance change in the gradual set and reset

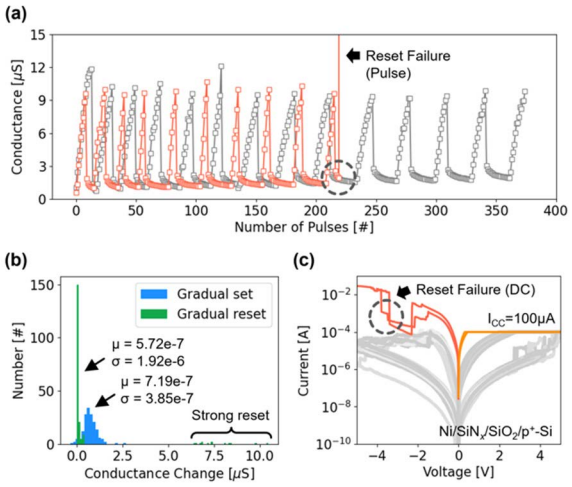

FIGURE 4. (a) Pulse operations showing that RRAM cells are vulnerable to reset failure when repetitive negative voltage stress is applied. (b) Distributions of conductivity change in gradual set (blue) and reset (green) operation. (c) DC cycles showing reset failures under negative stress and high current condition.

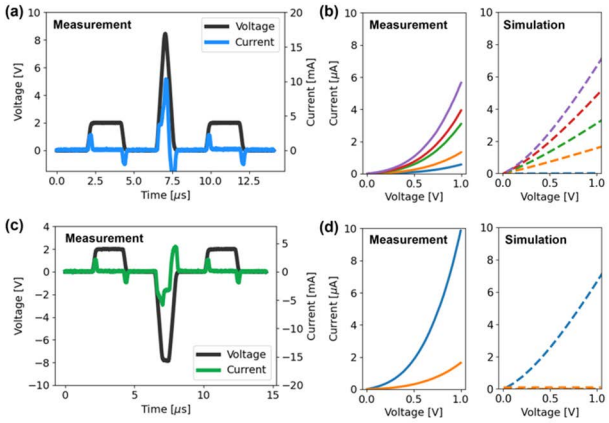

FIGURE 5. (a) Pulse operation for gradual set switching. (b) DC sweeps to check the transition of resistance state by set pulses. (c) Pulse operation for abrupt reset switching. (d) DC sweeps to check the transition of resistance state by reset pulses.

are shown in Fig. 4(b). It is confirmed that the gradual set operation is more reliable and uniform from the fact that the standard deviation of conductance change by one pulse is as small as 20% compared to the case of the gradual reset operation. DC cycles in Fig. 4(c) also show that reset operation has more difficulty preventing switching failure compared to the set operation, which can be easily managed by limiting the current flow using current compliance (I_C). In fact, reset operations have a condition that can easily cause breakdown because the RRAM device must withstand high current and temperature under negative bias conditions in order to trigger reset switching without the current limit. On the other hand, set failure, which can be defined as failure of proper switching operation from HRS to LRS, has been considered a relatively minor issue and rarely reported, and it never occurred in our experiment. In addition, the high resistance tail of the set-state distribution, known to be caused by this set failure [29],

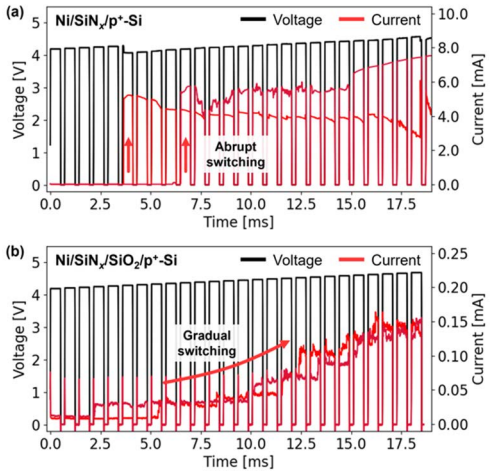

FIGURE 6. Voltage and current waveforms of (a) Ni/SiN_x/p^+-Si and (b) Ni/SiN_x/SiO_2/p^+-Si RRAM cells when applying the incremental step pulse scheme in two different cycles. By inserting a thin SiO_2 layer, highly reliable gradual switching characteristics can be achieved and current overshoot can be suppressed by itself.

is thought to be mostly resolved using the current compliance and the incremental-pulse scheme proposed in this study. Second, the bidirectional gradual switching operation may lead to longer time by complicating the write scheme and may also increase the burden of the peripheral circuitry by requiring a bidirectional sense-out operation. Recently, GSGR-based write methods have been proposed [30], [31]. Compared to the method proposed in this work, they are identical in terms of adopting iterative loops of the program-verify operation and incremental-step-pulse technique to reach the target conductivity of gradual RRAM. The main difference between those methods and the proposed method is whether gradual set or reset is determined based on the comparison with the target state and the current cell state. Such a decision process before each write loop and a bidirectional switching process may not be beneficial in terms of overall speed and device reliability.

Pulse operation and its effect on device conductivity are shown in Fig. 5(a)-(d). With a positive set pulse, the conductivity level gradually increases (Fig. 5(a) and (b)). Similarly, with a negative reset pulse, the conductivity level abruptly decreases (Fig. 5(c) and (d)). As described in Fig. 2(b), there is a fundamental difference in the pattern of the conductivity change due to positive-/negative-pulse application especially in the initial stage. Using these asymmetric switching behaviors, the concept of a high-speed write method is proposed and validated. It can be easily understood that the slight differences between the measurement and simulation data shown in Fig. 5(b) and (d) have little effect on the overall speed of the write method because it cooperatively utilizes the incremental-pulse scheme, multiple set/read loops, and verify/inhibit techniques. The voltage and current waveforms showing the effect of the inserted thin (∼1.5 nm) oxide layer on the switching characteristics are shown in Fig. 6.

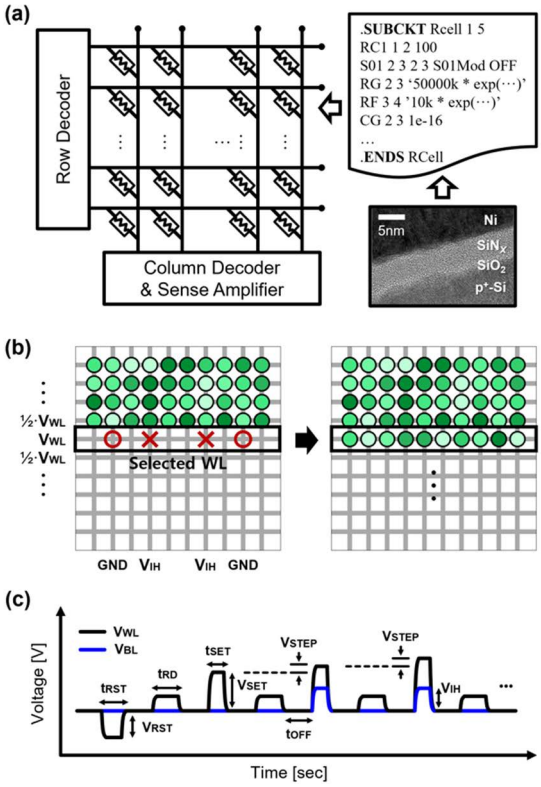

FIGURE 7. (a) Resistive-switching memory array and the subcircuit unit for array-level SPICE simulation. (b) Write target cells and inhibit cells of the selected WL in the proposed write operation. (c) Fast write scheme using the GSFR characteristics of multilevel resistive memory cells.

As incremental step pulses from 4.2 to 4.6 V are applied in two different cycles, abrupt switching occurs and a high current of several mA flows in the absence of SiO_2, which may seriously affect the device reliability. On the other hand, when SiO_2 is inserted, it can be seen that a reliable gradual switching that guarantees multilevel resistance states occurs consistently and that operating current is reduced by more than 10 times due to the SiO_2 tunnel barrier. The reduction in operating current has an important meaning because it is directly related to the reduction in energy consumption and the maximum number of synaptic devices that operate in parallel and simultaneously.

A schematic diagram of an RRAM cross-point array and the description of our cell model are represented in Fig. 7(a). The device characteristics can be implemented using the SPICE subcircuits, and this model can be simply embedded in the netlist of the cross-point array. The concept of the WL-by-WL write scheme is described in Fig. 7(b). V_{WL} is applied to the selected WL, and GND or V_{IH} is applied to the BL to determine whether it is written or inhibited. In addition, $1/2V_{WL}$ is applied to the unselected WL, thereby suppressing the conductivity change (write disturb) of the cells in the unselected WLs. The voltage waveforms in Fig. 7(c) are sequences that perform fast multilevel write operations. For

TABLE 2. Simulation parameters.

Parameter	Value	Description
t_{SET}	1 µs	SET pulse width
t_{RST}	2 µs	RESET pulse width
t_{RD}	1 µs	READ pulse width
t_{OFF}	200 ns	Interval between two pulses
V_{SET}	3.8 ~ 5.0 V	SET pulse amplitude
V_{RST}	−5 V	RESET pulse amplitude
V_{STEP}	0.1 ~ 0.4 V	Incremental voltage step

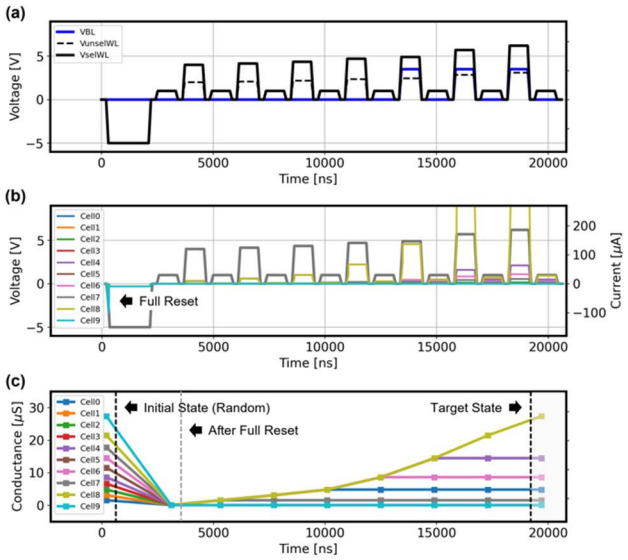

FIGURE 8. (a) Voltage waveforms of the BL and selected/unselected WL. (b) Voltage and current waveforms when fast GSFR write scheme is used. (c) Transition of each cell conductivity when the proposed write scheme is adopted.

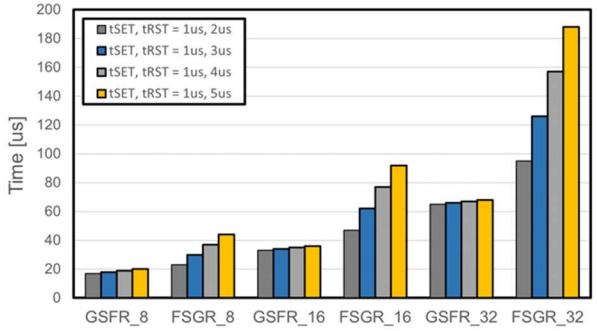

FIGURE 9. Comparision of the total time required to write 8, 16, and 32 conductivity states in a 784 × 10 hardware synapse array. The GSFR operation is always superior compared to the FSGR operation in terms of reducing the total write time.

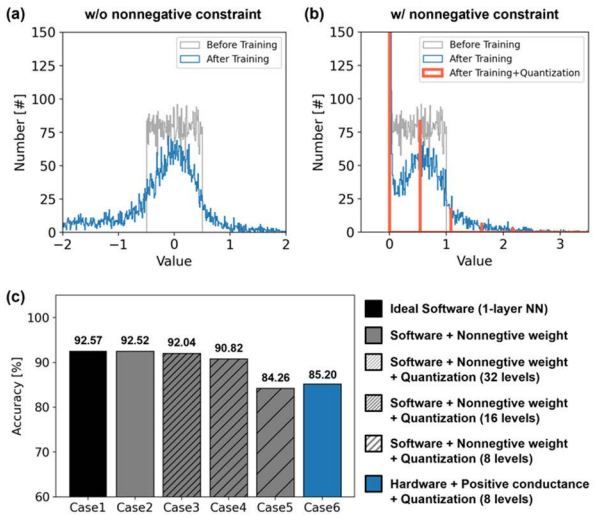

FIGURE 10. Weight distributions of an artificial neural network (a) when there is no initial and in-training constraint, and (b) when there is a nonnegative weight constraint. (c) Comparision of pattern recognition accuracies.

the first step, all RRAM cells in the selected WL are prepared to have the lowest conductivity using the full reset operation. Then, a small set voltage V_{SET} is applied to obtain a relatively low conductivity state, and the amplitude is increased by V_{STEP}. A read voltage V_{RD} is applied between each switching pulse to sense out the conductivity state, and whether or not the cell is inhibited is determined depending on whether or not the target conductivity is reached. The inhibit voltage V_{IH} has a margin that $(V_{SET} - V_{IH})$ does not cause set switching operation and that $(1/2V_{SET} - V_{IH})$ does not cause reset switching operation as summarized in the following equations.

$$(V_{SET} - V_{IH}) < V_{SET,Min} \tag{1}$$

$$|V_{RST,Min}| > |(1/2V_{SET} - V_{IH}) \tag{2}$$

Table 2 summarizes the parameters used in the circuit simulations.

IV. RESULTS AND DISCUSSION

The voltage waveforms of the BL and selected/unselected WL are shown in Fig. 8(a). V_{WL} from 4.0 to 6.2 V was applied to the selected WL, and $1/2V_{WL}$ was applied to unselected WL. BL is grounded at the beginning of the cycle, and V_{IH} was applied to the BL when the cell conductivity became equal to or greater than the target conductivity. Fig. 8(b) and (c) shows the current waveforms and conductivity state of each cell in the selected WL. After the full reset operation, all cells in the WL are prepared to have the lowest conductivity state (state 0). In addition, it is confirmed that the conductivity of each cell is increased for each cycle, and when the target conductivity is reached, the state is maintained by the inhibit operation. According to the experimental results, it was confirmed that full reset operation consumes 87.76 pJ and gradual set operations consume 30.58 pJ to 2.52 nJ, respectively. Because the energy consumption of the write

operation increases with the cell conductance, it is important to achieve a precise control of the multilevel states in the low conductivity region. More specifically, the insertion of a thin oxide layer can be an appropriate option to reduce the energy consumption as confirmed in Fig. 6. It should be noted that unexpected or nonideal phenomena such as variations from the target conductivity and initial conductivity change after the write operation were not been considered in the circuit simulation. The total time required for the overall write operation of the entire array when adopting the proposed sequence and the GSFR/FSGR operation can be expressed as follows:

$$t_{\text{total, GSFR}} = n_{\text{WL}} \times [1 \times (t_{\text{RST}} + t_{\text{RD}}) + \alpha \cdot (n_{\text{state}} - 1) \times (t_{\text{SET}} + t_{\text{RD}})] \quad (3)$$

$$t_{\text{total, FSGR}} = n_{\text{WL}} \times [1 \times (t_{\text{SET}} + t_{\text{RD}}) + \alpha \cdot (n_{\text{state}} - 1) \times (t_{\text{RST}} + t_{\text{RD}})] \quad (4)$$

In these equations, n_{WL} is the number of word lines, n_{state} is the number of conductivity states, and α is the number of pulses to increase on the conductivity state. Fig. 9 shows the total time required to implement 8, 16, and 32 conductivity states in a 784 $\times$ 10 hardware synapse array. In this work, it is assumed that only one pulse is required to move between states ($\alpha = 1$). It is confirmed that $t_{\text{total, GSFR}}$ is 35.3% to 176.5% smaller than $t_{\text{total, FSGR}}$ in the case of $t_{\text{SET}} = 1$ μs and $t_{\text{RST}} = 2, 3, 4, 5$ μs. In addition, $t_{\text{total, GSFR}}$ increases from 80.0% to 282.4% when the number of states increases from 8 to 16 and 32. Therefore, under conditions that require a longer reset pulse compared to the set pulse [21], the GSFR operation is always superior in terms of reducing the total write time. Implementing multilevel states is fundamentally different from implementing a binary state, and always requires longer write time. It should be noted that this comparison is limited to the performance of the WL-by-WL write method using gradual switching characteristics. Also, it can be concluded that it is advantageous to reduce the number of states if possible provided that the performance degradation such as inference accuracy is minimized. The inference accuracy of a hardware synapse array after weight transfer from an ANN was obtained by SPICE circuit simulations and compared to those of different software cases. The weight distributions of a one-layer ANN before and after supervised training are represented in Fig. 10(a). When trained for 50 epochs without any constraints, it can be seen that most of the weight values are distributed in the range of -2 to 2. Fig. 10(b) shows the weight distributions before and after supervised learning, and after quantization when there is an initial/in-training nonnegative weight constraint. The nonnegative weight constraint is set before training process starts and is applied each time to adjust the weight values. With the nonnegative constraint, the initial and final weight values are distributed in the range of zero or more. After training, the weight values can be quantized to suit the hardware implementation. The red bars in Fig. 10(b) indicate the position and distribution of the quantized weights. In Fig. 10(c), Case1 shows the pattern recognition accuracy when the Modified National Institute of Standards and Technology (MNIST) dataset is used in an ideal software level. It can be seen that the nonnegative weight constraint has little effect on the accuracy (Case2). According to the experimental results in Fig. 10(c), it can be confirmed that the negative weight value is not always essential if there are more than 32 weight levels and to solve problems such as simple pattern recognition. This result is also consistent with the results of a recently reported study [11]. At the same time, weight quantization can affect the accuracy depending on the number of states (Case3$-$Case5). In particular, when the number of states is smaller than 16, the accuracy decreases to less than 90% (Case5). From the relationship between weight quantization and inference accuracy, it can be seen that 16 levels (4 bits) or higher weight resolution are required to achieve acceptable accuracy (e.g., 90%). This would be a fundamentally necessary feature for synapse weight to learn or classify the characteristics of image patterns with analog levels. These synaptic properties related to quantization are generally consistent with those reported in the literature although it may be slightly affected by the neural network structure, the number of parameters and the complexity of the image dataset [32], [33]. The weight values from software training can be converted into conductivity of synaptic devices. It is assumed that the conductivity ranges from 0.02 to 25.2 μS, and can be quantized to more than eight levels, as shown in Fig. 3(c). Case6 shows that the accuracy of the inference operation performed at the hardware level is equivalent to that at the software level. A subtle difference in the accuracies between software and hardware is related to the number of test datasets. In our experiment, 10,000 test images in the MNIST dataset were used to check the accuracy of the software, and 500 test images were used in the circuit simulation to check the accuracy of the hardware. From the results in Figs. 9 and 10, it can be concluded that there is a trade-off between the inference accuracy and the total write time, with both mainly determined by the number of conductivity states, i.e., n_{state}.

V. CONCLUSION

In this work, a fast and reliable write scheme that fully utilizes the GSFR operation of RRAM is proposed and validated. To realize the voltage-dependent resistance switching behavior of a CMOS-compatible SiN_x-based RRAM cell, a SPICE compact model is introduced and adjusted based on the device characteristics. Considering that set pulses generally requires relatively shorter time than reset pulses and that reset failure is a more vulnerable and difficult-to-handle phenomenon, it is confirmed that the GSFR-based WL-by-WL scheme is superior in terms of speed and reliability compared to the FSGR-based operation. Finally, the inference accuracy of the synaptic memory array with the nonnegative constraint and weight quantization, is quantitatively compared with that of an ideal software algorithm.

REFERENCES

[1] C. Mead, "Neuromorphic electronic systems," *Proc. IEEE*, vol. 78, no. 10, pp. 1629–1636, Oct. 1990, doi: 10.1109/5.58356.

[2] D. Silver, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016, doi: 10.1038/nature16961.

[3] M. Davies, *Taking Neuromorphic Computing to the Next Level with Loihi 2*. Mountain View, CA, USA: Intel, 2021. [Online]. Available: https://download.intel.com/newsroom/2021/new-technologies/neuromorphic-computing-loihi-2-brief.pdf

[4] D. Ham, H. Park, S. Hwang, and K. Kim, "Neuromorphic electronics based on copying and pasting the brain," *Nature Electron.*, vol. 4, no. 9, pp. 635–644, Sep. 2021, doi: 10.1038/s41928-021-00646-1.

[5] M. F. Bear, B. W. Connors, and M. A. Paraadiso, "The changing brain," in *Neuroscience: Exploring Brain*, vol. 23, E. Lupash, ed. Philadelphia, PA, USA: LWW, 2006, pp. 689–723.

[6] G. W. Burr, R. M. Shelby, S. Sidler, C. di Nolfo, J. Jang, I. Boybat, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. N. Kurdi, and H. Hwang, "Experimental demonstration and tolerancing of a large-scale neural network (165 000 Synapses) using phase-change memory as the synaptic weight element," *IEEE Trans. Electron Devices*, vol. 62, no. 11, pp. 3498–3507, Nov. 2015, doi: 10.1109/TED.2015.2439635.

[7] M.-W. Kwon, K. Park, M.-H. Baek, J. Lee, and B.-G. Park, "A low-energy high-density capacitor-less I&F neuron circuit using feedback FET co-integrated with CMOS," *IEEE J. Electron Devices Soc.*, vol. 7, pp. 1080–1084, 2019, doi: 10.1109/JEDS.2019.2941917.

[8] M.-H. Kim, S. Hwang, S. Bang, T.-H. Kim, D. K. Lee, M. H. R. Ansari, S. Cho, and B.-G. Park, "A more hardware-oriented spiking neural network based on leading memory technology and its application with reinforcement learning," *IEEE Trans. Electron Devices*, vol. 68, no. 9, pp. 4411–4417, Sep. 2021, doi: 10.1109/TED.2021.3099769.

[9] H.-L. Park, M.-H. Kim, M.-H. Kim, and S.-H. Lee, "Reliable organic memristors for neuromorphic computing by predefining a localized ion-migration path in crosslinkable polymer," *Nanoscale*, vol. 12, no. 44, pp. 22502–22510, Nov. 2020, doi: 10.1039/D0NR06964G.

[10] H. Park, M. Kim, H. Kim, and S. Lee, "Self-selective organic memristor by engineered conductive nanofilament diffusion for realization of practical neuromorphic system," *Adv. Electron. Mater.*, vol. 7, no. 8, Aug. 2021, Art. no. 2100299, doi: 10.1002/aelm.202100299.

[11] M.-H. Kim, H.-L. Park, M.-H. Kim, J. Jang, J.-H. Bae, I. M. Kang, and S.-H. Lee, "Fluoropolymer-based organic memristor with multifunctionality for flexible neural network system," *NPJ Flexible Electron.*, vol. 5, no. 1, pp. 1–8, Dec. 2021, doi: 10.1038/s41528-021-00132-w.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2009.

[13] H. Tanaka, M. Kido, K. Yahashi, M. Oomura, R. Katsumata, M. Kito, Y. Fukuzumi, M. Sato, Y. Nagata, Y. Matsuoka, Y. Iwata, H. Aochi, and A. Nitayama, "Bit cost scalable technology with punch and plug process for ultra high density flash memory," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2007, pp. 14–15, doi: 10.1109/VLSIT.2007.4339708.

[14] C. Wang, H. Wu, B. Gao, T. Zhang, Y. Yang, and H. Qian, "Conduction mechanisms, dynamics and stability in ReRAMs," *Microelectron. Eng.*, vols. 187–188, pp. 121–133, Feb. 2018, doi: 10.1016/j.mee.2017.11.003.

[15] K.-C. Chang, T.-C. Chang, T.-M. Tsai, R. Zhang, Y.-C. Hung, Y.-E. Syu, Y.-F. Chang, M.-C. Chen, T.-J. Chu, H.-L. Chen, C.-H. Pan, C.-C. Shih, J.-C. Zheng, and S. M. Sze, "Physical and chemical mechanisms in oxide-based resistance random access memory," *Nanosc. Res. Lett.*, vol. 10, no. 1, p. 120, Mar. 2015, doi: 10.1186/s11671-015-0740-7.

[16] M. Lanza, "A review on resistive switching in high-K dielectrics: A nanoscale point of view using conductive atomic force microscope," *Materials*, vol. 7, no. 3, pp. 2155–2182, Mar. 2014, doi: 10.3390/ma7032155.

[17] D. Carta, I. Salaoru, A. Khiat, A. Regoutz, C. Mitterbauer, N. M. Harrison, and T. Prodromakis, "Investigation of the switching mechanism in TiO_2-based RRAM: A two-dimensional EDX approach," *ACS Appl. Mater. Interface*, vol. 8, no. 30, pp. 19605–19611, Jul. 2016, doi: 10.1021/acsami.6b04919.

[18] L. Gao, I.-T. Wang, P.-Y. Chen, S. Vrudhula, J.-S. Seo, Y. Cao, T.-H. Hou, and S. Yu, "Fully parallel write/read in resistive synaptic array for accelerating on-chip learning," *Nanotechnology*, vol. 26, Nov. 2015, Art. no. 455204, doi: 10.1088/0957-4484/26/45/455204.

[19] M.-H. Kim, S. Kim, K.-C. Ryoo, S. Cho, and B.-G. Park, "Circuit-level simulation of resistive-switching random-access memory cross-point array based on a highly reliable compact model," *J. Comput. Electron.*, vol. 17, no. 1, pp. 273–278, Mar. 2018, doi: 10.1007/s10825-017-1116-2.

[20] M.-H. Kim, S. Cho, and B.-G. Park, "Nanoscale wedge resistive-switching synaptic device and experimental verification of vector-matrix multiplication for hardware neuromorphic application," *Jpn. J. Appl. Phys.*, vol. 60, no. 5, May 2021, Art. no. 050905, doi: 10.35848/1347-4065/abf4a0.

[21] M.-H. Kim, S. Kim, S. Bang, T.-H. Kim, D. K. Lee, S. Cho, J.-H. Lee, and B.-G. Park, "Pulse area dependent gradual resistance switching characteristics of CMOS compatible SiNx-based resistive memory," *Solid-State Electron.*, vol. 132, pp. 109–114, Jun. 2017, doi: 10.1016/j.sse.2017.03.015.

[22] I.-T. Wang, C.-C. Chang, L.-W. Chiu, T. Chou, and T.-H. Hou, "3D Ta/TaO_x/TiO_2/Ti synaptic array and linearity tuning of weight update for hardware neural network applications," *Nanotechnology*, vol. 27, no. 36, Sep. 2016, Art. no. 365204, doi: 10.1088/0957-4484/27/36/365204.

[23] J. Woo, K. Moon, J. Song, M. Kwak, J. Park, and H. Hwang, "Optimized programming scheme enabling linear potentiation in filamentary HfO_2 RRAM synapse for neuromorphic systems," *IEEE Trans. Electron Devices*, vol. 63, no. 12, pp. 5064–5067, Dec. 2016, doi: 10.1109/TED.2016.2615648.

[24] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano Lett.*, vol. 10, no. 4, pp. 1297–1301, 2010, doi: 10.1021/nl904092h.

[25] V. Gupta, G. Lucarelli, S. Castro-Hermosa, T. Brown, and M. Ottavi, "Characterisation & modelling of perovskite-based synaptic memristor device," *Microelectron. Rel.*, vol. 111, Aug. 2020, Art. no. 113708, doi: 10.1016/j.microrel.2020.113708.

[26] V. Gupta, G. Lucarelli, S. Castro, T. Brown, and M. Ottavi, "Perovskite based low power synaptic memristor device for neuromorphic application," in *Proc. 14th Int. Conf. Design Technol. Integr. Syst. Nanosc. Era (DTIS)*, Apr. 2019, pp. 1–6, doi: 10.1109/DTIS.2019.8734983.

[27] S. Kim, S. Jung, M.-H. Kim, S. Cho, and B.-G. Park, "Resistive switching characteristics of Si3N4-based resistive-switching random-access memory cell with tunnel barrier for high density integration and low-power applications," *Appl. Phys. Lett.*, vol. 106, no. 21, May 2015, Art. no. 212106, doi: 10.1063/1.4921926.

[28] D. Panda, P. P. Sahu, and T. Y. Tseng, "A collective study on modeling and simulation of resistive random access memory," *Nanosc. Res. Lett.*, vol. 13, no. 1, pp. 1–48, Dec. 2018, doi: 10.1186/s11671-017-2419-8.

[29] S. Balatti, S. Ambrogio, D. C. Gilmer, and D. Ielmini, "Set variability and failure induced by complementary switching in bipolar RRAM," *IEEE Electron Device Lett.*, vol. 34, no. 7, pp. 861–863, Jul. 2013, doi: 10.1109/LED.2013.2261451.

[30] E. Hsieh, M. Giordano, and B. Hodson, "High-density multiple bits-per-cell 1T4R RRAM array with gradual SET/RESET and its effectiveness for deep learning," in *IEDM Tech. Dig.*, Feb. 2019, pp. 843–846, doi: 10.1109/IEDM19573.2019.8993514.

[31] J. Chen, H. Wu, B. Gao, J. Tang, X. S. Hu, and H. Qian, "A parallel multibit programing scheme with high precision for RRAM-based neuromorphic systems," *IEEE Trans. Electron Devices*, vol. 67, no. 5, pp. 2213–2217, May 2020, doi: 10.1109/TED.2020.2979606.

[32] C.-C. Chang, P.-C. Chen, T. Chou, I.-T. Wang, B. Hudec, C.-C. Chang, C.-M. Tsai, T.-S. Chang, and T.-H. Hou, "Mitigating asymmetric nonlinear weight update effects in hardware neural network based on analog resistive synapse," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 8, no. 1, pp. 116–124, Mar. 2018, doi: 10.1109/JETCAS.2017.2771529.

[33] G. Charan, A. Mohanty, X. Du, G. Krishnan, R. V. Joshi, and Y. Cao, "Accurate inference with inaccurate RRAM devices: A joint algorithm-design solution," *IEEE J. Explor. Solid-State Comput. Devices Circuits*, vol. 6, no. 1, pp. 27–35, Jun. 2020, doi: 10.1109/JXCDC.2020.2987605.

MIN-HWI KIM (Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from Seoul National University (SNU), in 2013 and 2020, respectively. In 2020, he joined Samsung Electronics, Hwaseong-si, South Korea, where he has been working on the design of NAND flash memories.

SIN-HYUNG LEE received the B.S. and Ph.D. degrees in electrical engineering from Seoul National University, South Korea, in 2013 and 2019, respectively. He is currently an Assistant Professor with the School of Electronics Engineering, Kyungpook National University, Republic of Korea. His research interests include neuromorphic electronics, artificial synapse, memristors, and organic electronics.

SUNGJUN KIM (Member, IEEE) received the Ph.D. degree in electrical engineering from Seoul National University, Seoul, South Korea, in 2017. From 2017 to 2018, he was a Senior Engineer with Samsung Electronics Company Ltd., South Korea. In 2018, he joined Chungbuk National University, South Korea, as an Assistant Professor. In 2020, he also joined Dongguk University, Seoul, as an Assistant Professor.

BYUNG-GOOK PARK (Fellow, IEEE) received the B.S. and M.S. degrees in electronics engineering from Seoul National University (SNU), in 1982 and 1984, respectively, and the Ph.D. degree in electrical engineering from Stanford University, in 1990. From 1990 to 1993, he worked with the AT&T Bell Laboratories, where he contributed to the development of 0.1 micron CMOS and its characterization. From 1993 to 1994, he was with Texas Instruments, developing 0.25 micron CMOS. In 1994, he joined SNU as an Assistant Professor at the School of Electrical Engineering (SoEE), where he is currently a Professor. In 2002, he worked with Stanford University as a Visiting Professor, on his sabbatical leave from SNU. From 2008 to 2010, he led the Inter-University Semiconductor Research Center (ISRC), SNU, as the Director. He has authored and coauthored over 1000 research articles in journals and conferences. His current research interests include the design and fabrication of nanoscale CMOS, flash memories, silicon quantum devices, and organic thin-film transistors. He has served as a Committee Member for several international conferences, including the Microprocesses and Nanotechnology, the IEEE International Electron Devices Meeting, the International Conference on Solid State Devices and Materials, and the IEEE Silicon Nanoelectronics Workshop. He received the Best Teacher Award from SoEEm, in 1997; the Doyeon Award for Creative Research from ISRC, in 2003; the Haedong Paper Award from the Institute of Electronic Engineers of Korea (IEEK), in 2005; the Educational Award from the College of Engineering, SNU, in 2006; the Haedong Research Award from IEEK, in 2008; the Nano Research Innovation Award from the Ministry of Science, ICT and Future Planning of Korea, in 2013; and Research Excellence Award from Seoul National University, in 2015. He has served as an Editor for IEEE Electron Device Letters.