Two- and three-terminal HfO₂-based multilevel resistive memories for neuromorphic analog synaptic elements

Heebum Kang¹, Jinah Park², Dokyung Lee³, Hyun Wook Kim², Sol Jin², Minjoon Ahn³ and Jiyong Woo¹,², ∗¹

¹ School of Electronics and Electrical Engineering, Kyungpook National University, Daegu 41566, Republic of Korea
² School of Electronics Engineering, Kyungpook National University, Daegu 41566, Republic of Korea
∗ Author to whom any correspondence should be addressed.
E-mail: jiyong.woo@knu.ac.kr

Keywords: neuromorphic computing, synaptic device, resistive switching memory, electrochemical random-access memory

Abstract
Synaptic elements based on memory devices play an important role in boosting neuromorphic system performance. Here, we show two types of fab-friendly HfO₂ material-based resistive memories categorized by configuration and an operating principle for a suitable analog synaptic device aimed at inference and training of neural networks. Since the inference task is mainly related to the number of states from a recognition accuracy perspective, we first demonstrate multilevel cell (MLC) properties of compact two-terminal resistive random-access memory (RRAM). The resistance state can be finely subdivided into an MLC by precisely controlling the evolution of conductive filament constructed by the local movement of oxygen vacancies. Specifically, we investigate how the thickness of the HfO₂-switching layer is related to an MLC, which is understood by performing physics-based modeling in MATLAB from a microscopic view. Meanwhile, synaptic devices driven by an interfacial switching mechanism instead of local filamentary dynamics are preferred for training accelerated neuromorphic systems, where the analogous transition of each state ensures high accuracy. Thus, we introduce three-terminal electrochemical random-access memory that facilitates mobile ions across the entire HfO₂ switching area uniformly, resulting in highly controllable and gradually tuned current proportional to the amount of migrated ions.

1. Introduction
As data in various formats are rapidly generated worldwide, modern computing systems based on von Neumann architecture become power-hungry to handle explosive data frequently. Neuromorphic architecture inspired by the biological brain structure has been introduced to enhance computing performance [1]. Artificial neural networks emulating massively cross-linked neurons by numerous synapses enable data execution in a fully parallel fashion, significantly improving latency and power efficiency. To implement this architecture on semiconductor-based hardware chips, synapses and neurons are fundamental building blocks for storing and computing data, which must be demonstrated with electronic devices [2, 3]. The role of the neurons is to collect the data transferred from connected synapses with respect to signals. As a result, silicon transistor-based peripheral circuitries have been built [4]. For the electronic synapse, static random-access memory (SRAM) has been used, where a unit transistor that can be aggressively scaled down to 2 nm [5] is cross coupled [6].

The power consumption of the SRAM-implemented neuromorphic chips was still considerable at the expense of the mature silicon device technology because additional power needed to be continuously supplied to the synaptic elements due to the inherently volatile nature of the SRAM. Besides, the small memory capacity (i.e., tens of MB) of SRAM is challenging, as higher synaptic density is required to run more complex algorithms for extended applications or general purposes beyond handwritten digits. These drawbacks from power and density perspectives highlighted the importance of exploring and developing alternative synaptic components using resistive random-access memory (RRAM) devices, typically demonstrated in simple two-terminal structures. Thus, nonvolatile RRAM implemented in a highly scaled region of approximately 10 nm allows dense synaptic arrays employing crossbar architecture [7], reducing the occupied area. It was found that...
Figure 1. Schematic illustration of the synaptic device configuration options for each purpose.

a computing execution metric in the form of tera operations per second per watt used to describe performance-power efficiency could be improved dramatically based on recent studies examining prototype RRAM chips [8, 9] and systematic end-to-end simulation analysis [10]. In addition to the structural advantages of the compact RRAM cell, a multilevel cell (MLC) characteristic paves the path for versatile classifications instead of digital binary storage of ‘0’ and ‘1’. In the inference phase during recognition, detected patterns decoded as input signals were fed into the word lines of the synaptic RRAM array in parallel. Simple multiplication of the input signal and stored data resulted in the output in the form of a current. The unknown patterns can, thus, be inferred by comparing the computed output currents from each bit line. When multiple states of the synaptic devices are assigned to the array, the combination of the output currents increases exponentially, resulting in highly accurate recognition accuracy.

Notably, the neuromorphic systems not only perform robust classification iteratively based on the predefined dataset, but also enable learning. The reversible shift between acquired states plays a more crucial role in this training phase than the number of states achieved [11]. That is, when an incorrect output occurred, the system started to adjust the state up or down to calculate the actual expected output based on back-propagation learning algorithms. For this update process, an identical pulse train technique, in which pulses of a fixed width and amplitude are repeatedly addressed, is preferred to alleviate the design complexity and burden from a driving circuit standpoint. Thus far, most RRAM devices have shown MLC, but nonuniformly and abruptly updated behavior is challenging, although various engineering methodologies have been attempted [12, 13]. It seems to have been overcome by employing electrochemical random-access memory (ECRAM) solely dedicated to the analog synaptic device at the expense of footprint loss [14]. The three-terminal ECRAM structure resembles a conventional silicon transistor, but the movement of mobile ions in the solid electrolyte driven by gate voltage into the channel region makes the device nonvolatile. Simultaneously, the ions progressively moved back and forth into the channel layer depending on gate voltage polarity, resulting in a wide range of channel currents.

Therefore, we demonstrated two- and three-terminal resistive memory devices for synaptic elements suitable for neuromorphic computing tasks. Fab-friendly HfO2 material was served as the switching layer for both devices, considering complementary metal-oxide semiconductor (CMOS) compatibility. In the first part of the discussion, we addressed the MLC characteristics of the two-terminal HfO2-based RRAM devices enabled by locally formed conductive filament for inference-accelerated systems, as shown in figure 1. Next, we demonstrated three-terminal HfO2-based ECRAM devices actuated by Cu ion migration through the entire area, favorable for achieving analogously tuned multiple states aimed at training-related systems.

2. Methods

Both RRAM and ECRAM were fabricated with fully CMOS compatible materials and processes. For the RRAM, a 6 nm-thick HfO2 layer was deposited by atomic layer deposition on the TiN bottom electrode. A 15 nm-thick Ti metal electrode was then used for scavenging to generate oxygen vacancies by absorbing the oxygen ions from the HfO2 due to the lower Gibbs free energy, resulting in a Ti/HfO2/TiN RRAM stack (from top to bottom) [12]. In this simple and compact two-terminal structure, the top (or bottom) electrode was connected to the voltage source (or ground) for measurement. For the three-terminal ECRAM, the ion migration was induced by introducing additional gate. A WO3 channel with a width and length of about μm each was first formed on the SiO2 substrate. Two W metal pads for source and drain were deposited at the edge of the WO3 channel to read the channel current. Then, HfO2 and Cu were subsequently deposited by sputtering for electrolyte and mobile ion-supplied gate electrode, respectively. Unlike the two-terminal structure, the voltage application
methods for programming and read operations were different, indicating that the current paths were decoupled. By applying a voltage to the gate and grounding the source, mobile ions were migrated vertically during programming. The changed state due to the ion migration can be identified by applying the voltage to the drain instead of the gate.

3. Results and discussion

3.1. HfO2 RRAM-based inference-accelerated systems

Figure 2(a) displays current–voltage (I–V) characteristics of 50 cycles for the HfO2 RRAM. Bipolar resistive switching behavior was visible after a forming process that led to the creation and clustering of oxygen vacancies at approximately 5 V. The flow of electrons through the formed oxygen vacancy-based conductive filament resulted in a low resistance state (LRS), corresponding to a set operation. The sudden increase in the current was limited by a compliance current (I_{cc}) to prevent permanent breakdown. When a negative voltage was applied to the RRAM, the LRS was converted to a high resistance state (HRS) equal to the reset operation. Consequently, the oxygen vacancies in the filament were dissolved, disconnecting the path between two electrodes. The memory window between LRS and HRS was greater than two orders of magnitude at a read voltage (\text{V}_{\text{read}}) of 0.1 V during the repeated cycle, as shown in figure 2(b).

Considering the filamentary dynamics of the RRAM, the manner to achieve MLC is categorized as either adjusting the I_{cc} [15] or reset voltage (\text{V}_{\text{reset}}) [16], as shown in figure 3. The higher I_{cc} induced more oxygen vacancies toward the filament, thickening its size. The LRS was lowered by increasing the I_{cc} value setting from 200 to 750 \text{\mu}A at a given \text{V}_{\text{reset}} of −1.2 V (figure 3(a)). However, the multiple states demonstrated in the LRS regime caused the huge output current resulting from summed total activated LRSs, worsening power consumption. On the other hand, the same amount of states was obtained by modulating \text{V}_{\text{reset}} from −0.9 to −1.15 V with a finely divided \text{V}_{\text{reset}} step of −0.05 V at the given I_{cc} of 1 mA, as shown in figure 3(b). The I–V characteristics assigned to each condition were measured repeatedly 10 times and then \text{V}_{\text{reset}} was gradually increased in the same device. The negative \text{V}_{\text{reset}} caused the vacancies to disperse from the filament and the connected filament started to be dissolved at about −0.5 V, increasing the resistance of the RRAM. As larger \text{V}_{\text{reset}} amplitude was applied, a switching gap between the electrode and ruptured filament was widened. The multilevel states were thus positioned in the range of \mu\text{A}, which was approximately 100 times lower than the LRS. The state uniformity was vulnerable to cycling at the expense of realizing an MLC at low current levels [17] because the ruptured filament with the switching gap tended to be easily disturbed by residual vacancies, inducing state fluctuation.

The switching variability of the RRAM can be improved by constricting the switching place [18], which was realized by geometrical scaling of the switching layer thickness [19]. In general, the chemical reaction occurs at the interface by absorbing the oxygen ions from the oxide, creating additional mixed oxide layer. This means that the oxygen ions should be first migrated from the HfO2 toward the reactive Ti electrode, and then the reaction is performed due to the contact of oxygen ions with Ti atoms. That is, how much the reactant corresponding to the Ti atom is exposed to the oxygen ions becomes important [18]. Therefore, the amount of oxygen vacancies in the HfO2 created by the chemical reaction is expected to be the same when the Ti scavenging layer of the same thickness is applied. The impact of the vacancies on the evolution of the filament was
Figure 3. The MLC I–V curves are achieved by elaborately tuning either (a) $I_{cc}$ or (b) $V_{\text{reset}}$. Ten cycles were measured for each condition.

Figure 4. (a) The memory window as a function of the HfO2 switching layer thickness. The trade-off relationship between the median memory window written in figure and switching uniformity. (b) The read current was evaluated by steadily increasing $V_{\text{reset}}$ and the (c) $V_{\text{set}}$–$V_{\text{reset}}$ graph with respect to the HfO2 thickness from 4 to 6 nm.

thus pronounced in the thinner HfO2 layer. Since the preexisting vacancies were relatively sufficient to establish the filament, the probability of generating additional vacancies was reduced, thereby suppressing randomly formed paths. As shown in figure 4(a), the distribution of each current level read at 0.1 V extracted from the I–V traces of 50 cycles for the RRAMs operated by the same voltage conditions was tighter in the switching layer thinned to 4 nm. While the uniformity was improved, the memory window was narrowed, which must be significant to achieve an MLC, resulting in a trade-off relationship. The $V_{\text{reset}}$ was varied for all RRAM devices to identify the root cause of the small memory window. As presented in figure 4(b), average read currents of 10 cycles in the 5 and 6 nm-thick HfO2 RRAMs were well-controlled by $V_{\text{reset}}$. Conversely, the current became flat without a noticeable change in the entire $V_{\text{reset}}$ range in the 4 nm-thick HfO2 RRAM, making it difficult to distinguish the states. In addition, as the switching layer decreased in thickness, the set voltage ($V_{\text{set}}$) seemed to be independent of $V_{\text{reset}}$, as shown in figure 4(c). In general, the $V_{\text{set}}$ followed a proportional relationship with $V_{\text{reset}}$ because the vacancies needed to travel according to the extent of filament rupture by $V_{\text{reset}}$, so the required $V_{\text{set}}$, which is the driving force to lead the vacancies, was also more significant. As expected, the $V_{\text{set}}$ and $V_{\text{reset}}$ closely followed the linear relationship for the 6 nm HfO2 RRAM exhibited. For the 5 nm-thick HfO2 film, the linear line shifted toward the low $V_{\text{set}}$ and eventually exhibited a nearly constant $V_{\text{set}}$ with respect to the $V_{\text{reset}}$ for the 4 nm-thick HfO2 layer.

We performed physics-based modeling in MATLAB to understand these observations at a microscopic view. The switching gap representing $g$ was designed to be varied over time based on the field-induced ion migration through generation and recombination of the vacancies [20, 21]. The numerical equation to describe the process is as follow:

$$\frac{dg}{dt} = -v_0 \cdot \exp\left(\frac{E_a}{kT}\right) \cdot \sinh\left(\frac{\gamma a_0 q V}{LkT}\right)$$

(1)

where $v_0$ is the velocity containing attempt-to-escape frequency, $E_a$ is the activation energy for vacancy migration, $\gamma$ is the field enhancement factor, $a_0$ is the hopping site distance, and $L$ is the switching layer’s thickness. $\gamma$ can be extracted with the following relation:

$$\gamma = \gamma_0 - \beta\left(\frac{g}{g_1}\right)^3$$

(2)
Table 1. The parameters used in the simulation.

| Parameter | Value
|-----------|--------|
| L (nm)    | 4      |
|           | 5      |
|           | 6      |
| Gap (nm)  | 0.35   |
|           | 0.4    |
|           | 0.65   |
| HRS       | 0.85–1.00 |
|           | 1.05–1.50 |
|           | 1.16–1.92 |
| $E_a$ (eV) | 1.07  |
|           | 1.17  |
|           | 1.46  |
| $I_0$ (mA) | 2.456 |
| $g_0$ (nm) | 0.275 |
| $g_1$ (nm) | 1.5   |
| $V_0$ (V)  | 0.3   |
|           | 0.27  |
|           | 0.19  |
| $\gamma_0$ | 16.5 for set |
|           | 13 for reset |
| $\beta$   | 8 for set |
|           | 0.5 for reset |
| $\nu_0$ (ms^{-1}) | 0.1 |
| $a_0$ (nm) | 0.25 |

Figure 5. The simulated $I–V$ curves of the RRAMs matched with the experimental data.

where all the parameters are fitting parameters. Then, the current directly related to the extent of the growth of the filament is described as

$$I = I_0 \cdot \exp\left(-\frac{g}{g_0}\right) \cdot \sinh\left(\frac{V}{V_0}\right)$$  (3)

where $I_0$ is the constant, and $g_0$ and $V_0$ are the gap and voltage coefficient, respectively.

The parameters used in the simulation are listed in table 1.

The specific gap and range for each HfO$_2$ thickness were empirically used. The experimental traces (symbols) were fitted with the simulated MLC $I–V$ curves (lines) achieved by expansion and contraction of the gap, as shown in figure 5. More importantly, we identified that lowering $E_a$ from 1.46 and 1.17 to 1.07 eV at the same given parameters primarily led the $V_{\text{set}}$ to be invariant, accompanying a narrowed interval of multiple HRS as a function of HfO$_2$ thickness. As discussed above, a relatively larger number of vacancies actively participated in the filament evolution in the thinner HfO$_2$, where the vacancy migration was easier due to lowered $E_a$ [22]. Although the filament was disconnected at approximately $-0.5$ V during the reset operation for the 4 nm HfO$_2$ RRAM, numerous vacancies remained around the filament, making it difficult to lower the HRS level. The larger $V_{\text{reset}}$ helped reduce the HRS by pushing the vacancies further away from the filament. However, the reset breakdown phenomenon indicating an irreversible set at negative voltage instead of lowering the HRS was observed (not shown here). It limited the maximum allowable $V_{\text{reset}}$ to $-1.2$ V for the 4 nm-thick HfO$_2$, where the tolerant $V_{\text{reset}}$ extent was $-1.9$ V. Therefore, our findings revealed that the optimal switching layer thickness considering the MLC and reliability such as uniformity and failure needed to be selected.

In this regard, the 5 nm RRAM was optimal in this work. Based on the achieved MLC (figure 6(a)), the feasibility of inference tasks on the Modified National Institute of Standards and Technology (MNIST) dataset was evaluated. 10 output neuron layers were connected with 784 input neurons layers through two hidden layers with 250 and 125 neurons to determine handwritten digits from 0 to 9, as shown in figure 6(b). The dataset was pretrained with 10 bits, and the synaptic weight was then quantized [23]. Since a relatively simple MNIST dataset was utilized to classify, higher inference accuracy of approximately 94% was ideally displayed even though the synaptic device exhibited binary states, as shown in figure 6(c). As the bit of the RRAM was
increased, an accuracy greater than 95% was ensured. A significant deviation-dependent drop in the accuracy occurred for the 1 bit per RRAM, assuming that the instability of the state was applied for the real case. The MLC of the RRAM noticeably mitigated the accuracy degradation.

From a practical point of view, investigation on reliability of the multiple states is required for hardware implementation of RRAM based neural networks [24–26]. By introducing a verification algorithm, equally spaced multilevels can be demonstrated at the chip-level [27]. State stability is more important when considering inference engines, where the states are rarely programmed. Specifically, the state instability can be accelerated due to the elevated temperature that occurs while the chip is running. In this case, a refresh algorithm can be a solution to mitigate the resistance drift toward a specific direction or random motion [28]. Finally, permanent failure analysis should be carefully managed. As we discussed above, the unexpected reset breakdown of the device, which was one of the failure modes, induced stuck-at-LRS [29]. Since the sum of the total resistances of the RRAMs located in the selected column is interested in the neuromorphic systems, the high current of the faulty single device makes it difficult to classify the output values during the inference. Therefore, a systematic assessment of robust multilevel RRAM should be studied through an understanding of the root causes of possible failure scenarios in the various operating environments.

3.2. HfO2 ECRAM-based training-accelerated systems

In the aforementioned HfO2 RRAM, jumps between each state were solely available when either $I_{cc}$ or $V_{reset}$ was constantly changed, heading upwards or downwards, respectively. The use of continued nonidentical conditions every cycle becomes a massive burden in the driving or peripheral circuitries, eventually degrading system power and latency. Hence, we studied the three-terminal ECRAM device designed to respond to identical pulses for the training-accelerated systems. For the desired analog synaptic operation, mobile ions in the electrolyte actuated by an additional gate terminal reached the channel, usually consisting of an oxide semiconductor [14, 30]. Then, the oxide semiconductor channel became highly conductive due to the injected ions serving as dopants. It resulted in a controllable channel current as a function of the gate input. In the early stage, Li ions and their corresponding ion conductor, such as lithium phosphorus oxynitride, were used for mobile ions and solid electrolytes, respectively, inspired by rechargeable secondary ion batteries [14, 31–34]. Various mobile ions such as Na [35] and H [36] have been explored recently based on their excellent synaptic behavior. Here, by taking into account the CMOS compatibility, we configured the ECRAM device with the Cu/HfO2/WO$_x$ stack. To provide Cu mobile ions, a Cu electrode was used, which can be directly integrated into the back-end-of-line interconnect process. An HfO2 electrolyte was utilized to allow the gate-controlled Cu ion motion. Since the W atom has multiple valence states in the WO$_x$ channel, it can induce a wide range of conductivity changes.

Figure 7(a) presents a transmission electron microscopic (TEM) image of the multilayered ECRAM’s gate stack, and each layer was clearly classified by energy-dispersive x-ray (EDX) analysis, as shown in figure 7(b). In addition to the HfO2 layer, we investigated various electrolyte materials, such as SiO2 [36] and MoO3 [37], specifically selected from previously reported literature. Figure 8 demonstrates that the normalized read current due to migrated ions controlled by a gate pulse of 8 V with a pulse width of 1 s was a strong function of the electrolytes used. No change in the MoO3/WO$_x$ stack’s read current was shown regardless of the gate pulse successively applied. Alternatively, the field-driven Cu ions started to be involved in the channel area in both the SiO2 and HfO2 electrolytes. Notably, the channel current only responded when hydrogen was doped in the SiO2 based on the preliminarily performed study. Therefore, it can be inferred that the observed change in the channel current was not due to interface defects, but was mainly derived by additionally incorporated mobile ions in the electrolyte. Since a hydrogen-free undoped SiO2 electrolyte was utilized, we believe that the Cu ion...
migration is essential in adjusting the channel current. Compared with the SiO₂ electrolyte, the HfO₂ electrolyte enabled more sensitive increment under the continuously addressed gate pulse, indicating progressive ion movement in the HfO₂/WOₓ stack.

The synaptic characteristics can be achieved through material optimization of the HfO₂/WOₓ structure, discussed in detail in reference [38], as shown in figure 9(a). A lateral channel current at 0.5 V with respect to the identical gate voltage of ±3 V with a pulse width of 1 s was analogously tuned. The higher gate voltage amplitude enlarged the dynamic range of the current change. Furthermore, the gate-controlled channel property was electrically verified using the examining area scaling analysis, as shown in figure 9(b). As the channel area was reduced, the read current was proportionally lowered, meaning that the entire channel area participated in the resistive switching. We measured the gate current by applying a voltage to the gate and grounding to the source to clarify the dominant switching location. The current was a level below the μA range by sweeping the voltage to 3 V without an abrupt current increase, as shown in figure 9(c). Instead, under the repeated sweeping voltage, the current was smoothly increased. It indicated that no leakage paths due to either oxide breakdown or clustering of the Cu ions were formed locally, and the impact of the gate side on the channel current was, thus, negligible.

The obtained synaptic characteristics were expressed in terms of a linearity factor, α, which is the performance index of the synaptic device [39], where the current increase or decrease was 2.42 or 0.21, respectively, as shown in figure 10. The impact of the linearity on the recognition accuracy was then evaluated using the constructed neural network with back-propagation algorithms in MATLAB. As the error arose in the output neurons during propagation, the states obtained from the ECRAM synaptic devices were updated following the states versus pulse number curve (figure 10(a)). The filamentary RRAM exhibited the nonlinear and asymmetric synaptic response, resulting in low accuracy of approximately 10%. On the other hand, the training accuracy
Figure 9. (a) The channel current as a function of the number of gate pulses. (b) The initial read current measured at 0.5 V was linearly proportional to the channel area. (c) The gate current for repeated voltage sweeps was steadily increased.

Figure 10. (a) Comparison of RRAM and ECRAM devices from synaptic behavior achieved using the identical pulse scheme. (b) The accuracy was significantly improved in the ECRAM due to the gradually tuned current levels as a function of pulse number.

(90.8%) was dramatically improved by gradually adjusting the update operation of the ECRAM, as shown in figure 10(b).

4. Conclusion

We addressed CMOS-compatible HfO2-based RRAM and ECRAM devices for inference and training accelerated neuromorphic systems, respectively. We examined how the HfO2 switching layer thickness was involved in achieving an MLC employing electrical characterization and MATLAB-based modeling. Our measurement results revealed that a trade-off relationship between an MLC and switching uniformity depending on the HfO2 thickness. The appropriate thickness of the main switching layer should be considered when the $V_{\text{reset}}$ modulation technique was used for the MLC in the filamentary switching. Instead of a locally formed filament, whole area switching was preferred for the training-enhanced analog synaptic behavior. The Cu ion-actuated ECRAM with the HfO2/WOx stack has, thus, been exploited, and the gradual synaptic response driven by the number of identical gate voltages was demonstrated. We hope that our findings can provide insight into designing the device structure aimed at a specific target application.

Acknowledgments

This work was supported by the National Research Foundation (NRF) Grant funded by the Korean government (MSIT) (NRF-2020M3F3A2A01081775 and NRF-2021R1C1C1003261).
Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Conflict of interest

There are no conflicts of interest to declare.

ORCID iDs

Jiyong Woo @ https://orcid.org/0000-0002-4968-6985

References

[1] Mead C 1990 Neuromorphic electronic systems Proc. IEEE 78 1629–36
[2] Woo J, Kim J H, Im J-F and Moon S E 2020 Recent advancements in emerging neuromorphic device technologies Adv. Intell. Syst. 2 2000111
[3] Ielmini D and Ambrogio S 2020 Emerging neuromorphic devices Nanotechnology 31 092001
[4] Indiveri G et al 2011 Neuromorphic silicon neuron circuits Front. Neurosci 5 75
[5] Shen P-C et al 2021 Ultralow contact resistance between semimetal and monolayer semiconductors Nature 593 211–7
[6] Dong Q, Sinanigil M, Erbagci B, Sun D, Khwa W, Liao H, Wang Y-C and Chang J 2020 15.3 A 351TOPS/W and 372.4GOPS compute-in-memory SRAM macro in 7 nm FinFET CMOS for machine-learning applications IEEE Int. Solid-State Circuits Conf.
[7] Burr G W et al 2017 Neuromorphic computing using non-volatile memory Adv. Phys. X 2 89–124
[8] Mochida R et al 2018 A 4M synapses integrated analog ReRAM based 66.5 TOPS/W neural-network processor with cell current controlled writing and flexible network architecture Symp. on VLSI Technology
[9] Yao P, Wu H, Gao B, Tang J, Zhang Q, Zhang W, Yang J J and Qian H 2020 Fully hardware-implemented memristor convolutional neural network Nature 577 641–6
[10] Peng X, Huang S, Luo Y, Sun X and Yu S 2019 DNN+NeuroSim: an end-to-end benchmarking framework for compute-in-memory accelerators with versatile device technologies IEEE Int. Electron Devices Meeting vol 32
[11] Woo J and Yu S 2018 Resistive memory-based analog synapse: the pursuit for linear and symmetric weight update IEEE Nanotechnol. Mag. 12 36–44
[12] Woo J, Moon K, Song J, Lee S, Kwak M, Park J and Hwang H 2016 Improved synaptic behavior under identical pulses using AIO/HfO2 bilayer RRAM array for neuromorphic systems IEEE Electron Device Lett. 37 994–7
[13] Wu W, Wu H, Gao B, Deng N, Yu S and Qian H 2017 Improving analog switching in HfO2-based resistive memory with a thermal enhanced layer IEEE Electron Device Lett. 38 1019–22
[14] Fuller E J, Ghaly F E, Léonard F, Agarwal S, Plimpton S J, Jacobs-Gedrim R B, James C D, Marinella M J and Talin A A 2017 Li-ion synaptic transistor for low power analog computing Adv. Mater. 29 1604310
[15] Milo V et al 2021 Accurate program/verify schemes of resistive switching memory (RRAM) for in-memory neural network circuits IEEE Trans. Electron Devices 68 3832–7
[16] Yu S, Wu Y, Jeyasingh R, Kuzum D and Wong H S P 2011 An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation IEEE Trans. Electron Devices 58 2729–37
[17] Goux I et al 2013 Understanding of the intrinsic characteristics and memory trade-offs of sub-μA filamentary RRAM operation Symp. on VLSI Technology T162–T163
[18] Lee J et al 2010 Diode-less nano-scale ZrOx/HfO2 RRAM device with excellent switching uniformity and reliability for high-density cross-point memory applications Int. Electron Devices Meeting
[19] Lee J, Park J, Jung S and Hwang H 2011 Scaling effect of device area and film thickness on electrical and reliability characteristics of RRAM IEEE Int. Interconnect Technology Conf.
[20] Chen P-Y and Yu S 2015 Compact modeling of RRAM devices and its applications in 1T1R and 1S1R array design IEEE Trans. Electron Devices 62 4022–8
[21] Jiaing Z, Wu Y, Yu S, Yang L, Song K, Karim Z and Wong H-S P 2016 A compact model for metal-oxide resistive random access memory with experiment verification IEEE Trans. Electron Devices 63 1884–92
[22] Yang J, Strukov D B and Stewart D R 2013 Memristive devices for computing Nat. Nanotechnol. 8 13–24
[23] Choi W, Kwak M, Kim S and Hwang H 2021 Neural network training acceleration with RRAM-based hybrid synapses Front. Neurosci. 15 690418
[24] Lanza M et al 2019 Recommended methods to study resistive switching devices Adv. Electron. Mater. 5 1800143
[25] Tang J et al 2019 Bridging biological and artificial neural networks with emerging neuromorphic devices: fundamentals, progress, and challenges Adv. Mater. 31 1901276
[26] Zheng N and Mazumder P 2019 Learning in Energy-Efficient Neuromorphic Computing: Algorithm and Architecture Co-design (New York: Wiley)
[27] Romore-Zaliz R, Pérez E, Jiménez-Molina F, Wenger C and Roldán J 2021 Study of quantized hardware deep neural networks based on resistive switching devices, conventional versus convolutional approaches Electronics 10 346
[28] Xiang Y, Huang P, Zhao Y, Zhao M, Gao B, Wu H, Qian H, Liu X and Kang J 2019 Impacts of state instability and retention failure of filamentary analog RRAM on the performance of deep neural network IEEE Trans. Electron Devices 66 4517–22
[29] Woo J, Nguyen T V, Kim J H, Im J-P, Im S, Kim Y, Min K-S and Moon S E 2020 Exploiting defective RRAM array as synapses of HTM spatial pooler with boost-factor adjustment scheme for defect-tolerant neuromorphic systems Sci. Rep. 10 11703
[30] Tang J et al 2018 ECRAM as scalable synaptic cell for high-speed, low-power neuromorphic computing Int. Electron Devices Meeting vol 13
[31] Kim M et al 2021 Multinary data processing based on nonlinear synaptic devices J. Electron. Mater. 6 3471–7
[32] Lee C et al 2021 Li memristor-based MOSFET synapse for linear $I-V$ characteristic and processing analog input neuromorphic system Japan J. Appl. Phys. 60 024003

[33] Choi Y, Lee C, Kim M, Song Y, Hwang H and Lee D 2019 Structural engineering of Li-based electronic synapse for high reliability IEEE Electron Device Lett. 40 1992–5

[34] Lee C, Lee J, Kim M, Woo J, Koo S-M, Oh J-M and Lee D 2019 Two-terminal structured synaptic device using ionic electrochemical reaction mechanism for neuromorphic system IEEE Electron Device Lett. 40 546–9

[35] Lee K, Lee J, Nikam R D, Heo S and Hwang H 2020 Sodium-based nano-ionic synaptic transistor with improved retention characteristics Nanotechnology 31 455204

[36] Lee J, Lim S, Kwak M, Song J and Hwang H 2019 Understanding of proton induced synaptic behaviors in three-terminal synapse device for neuromorphic systems Nanotechnology 30 255202

[37] Yang C-S, Shang D-S, Liu N, Fuller E J, Agrawal S, Talin A A, Li Y-Q, Shen B-G and Sun Y 2018 All-solid-state synaptic transistor with ultralow conductance for neuromorphic computing Adv. Funct. Mater. 28 1804170

[38] Kang H and Woo J 2021 Ca-ion-actuated three-terminal neuromorphic synaptic devices based on binary metal-oxide electrolyte and channel Appl. Phys. Lett. 119 072103

[39] Jang J W, Park S, Burr G W, Hwang H and Jeong Y H 2015 Optimization of conductance change in Pr$_{1-x}$Ca$_x$MnO$_3$-based synaptic devices for neuromorphic systems IEEE Electron Device Lett. 36 457–9