Memristor crossbar array for binarized neural networks

Cite as: AIP Advances **9**, 045131 (2019); https://doi.org/10.1063/1.5092177
Submitted: 08 February 2019 . Accepted: 18 April 2019 . Published Online: 30 April 2019

Yong Kim, Won Hee Jeong, Son Bao Tran, Hyo Cheon Woo, Jihun Kim, Cheol Seong Hwang, Kyeong-Sik Min, and Byung Joon Choi

Memristor crossbar array for binarized neural networks

Cite as: AIP Advances 9, 045131 (2019); doi: 10.1063/1.5092177
Submitted: 8 February 2019 • Accepted: 18 April 2019 •
Published Online: 30 April 2019

Yong Kim,[1] Won Hee Jeong,[1] Son Bao Tran,[2] Hyo Cheon Woo,[3] Jihun Kim,[3] Cheol Seong Hwang,[3] Kyeong-Sik Min,[2] and Byung Joon Choi[1,a)]

AFFILIATIONS

[1]Department of Materials Science and Engineering, Seoul National University of Science and Technology, Seoul 01811, Korea
[2]School of Electrical Engineering, Kookmin University, Seoul 02707, Korea
[3]Department of Materials Science and Engineering and Inter-University Semiconductor Research Center, College of Engineering, Seoul National University, Seoul 08826, Korea

[a)]**Corresponding Author E-mail:** bjchoi@seoultech.ac.kr

ABSTRACT

Memristor crossbar arrays were fabricated based on a Ti/HfO_2/Ti stack that exhibited electroforming-free behavior and low device variability in a 10 x 10 array size. The binary states of high-resistance-state and low-resistance-state in the bipolar memristor device were used for the synaptic weight representation of a binarized neural network. The electroforming-free memristor was confirmed as being suitable as a binary synaptic device because of its higher device yield, lower variability, and less severe malfunction (for example, hard break-down) than the electroformed memristors based on a Ti/HfO_2/Pt structure. The feasibly working binarized neural network adopting the electroforming-free binary memristors was demonstrated through simulation.

Memristors have been extensively studied as a next-generation nonvolatile memory candidate, known as resistive random access memory (ReRAM) by virtue of their excellent features, such as ultrafast speed (<100ps), long endurance (up to 10^{12} write-erase cycles), and scalability (scale down to 2nm contact dimension) with a simple structure.[1–4] Recently, there have been many studies focused on the possible application of memristors to newly developing fields, such as neuromorphic computing, pattern recognition, sensory device, and security applications.[3,5–20] Such applications are mainly enabled by the analog behavior of memristors due to their working principle – the formation and annihilation of a conducting filament (CF) or channel (conduction path for electrons) composed of mobile ions or vacancies. The resistance of the memristor device could be reconfigured by changing the internal distribution of (defective) ions or vacancies in a CF by using an electric field aided by Joule heating. This ion-species-related reconfiguration process resembles the analog behavior of the biological synapse in the human's neural system.[17] In this regard, memristive switching in an analog manner has been used for synaptic devices in various neural network hardware. However, multilevel switching of a memristor to mimic the analog synapse still remains challenging, because a memristor controlled by internal ion distribution is inherently confronted with variability issues.[3,21] Many of the variability issues are related with the electroforming (EF) step, which changes the highly insulating pristine memristor to a switchable state by either voltage- or current-sweep with appropriate compliance values.[22,23]

By contrast, the digital-like operation of the memristor, i.e., having two resistance states of high-resistance state (HRS) and low-resistance state (LRS), suffers less from variability issues, and could be employed for binarized neural networks (BNNs), which is one of the simplest versions of a neural network (NN).[21,24] BNN only uses binary synaptic weights without requiring the calculation of any high-precision numbers such as 24-, 32-, or 64-bit, which a deep neural network (DNN) normally uses for image recognition or other machine learning applications. The advantage of BNN compared to other NNs is that the binarized operation can be performed simply by bitwise logic rather than complicated multiply-accumulate (MAC) circuits, which saves substantial energy and time for the calculation.[21] Although the low precision and low

area-efficiency of BNN could be disadvantageous, the recognition rate loss of BNN is sufficiently small to make it useful in the application of edge device or Internet of Things (IoTs) possessing low-precision functions of machine learning.[21] Even for the BNN, nonetheless, the presence of reliable memristors with low error rate is necessary.

Therefore, in this work, the "electroforming-free" ("EF-free") memristor crossbar array (CBA) was fabricated, and it was applied to the BNN for simple 3-bit pattern recognition. The EF-free binary memristor CBA is considered to be suitable as a BNN hardware, since binary operation could be performed in parallel while the positive features of memristors remained effective. Technologically, the EF process poses several issues, such as the current overshoot, non-uniform electroforming bias distribution,[23] and an over-sized peripheral circuit for the high-voltage drive.[22] The EF-free memristor can mitigate such issues, which therefore greatly releases a burden for the realization of a memristor CBA. As the EF is an electrical stress process that changes the highly insulating pristine oxide layer to a relatively defective and electrically switchable layer, the switchable state can be alternatively acquired by a chemical mean. In this work, the inert Pt electrode of the EF-necessary Ti/HfO_2/Pt memristor was changed to a chemically reactive Ti electrode, resulting in the Ti/HfO_2/Ti structure. Such a seemingly symmetrical structure may seem inappropriate for use in bipolar resistance switching (BRS) mode, but as shown in this work, it demonstrated highly feasible BRS characteristics with EF-free performance.

Memristor CBAs were fabricated having a square junction with widths varying from 4 to 20 μm. Figure 1(a) shows the optical microscopic image of the 10×10 CBAs (total 100 cells included) with a 5 μm memristor junction width. Electron-beam evaporated Ti(20nm)/Pt(40nm) or Pt(40nm) metal lines were used as bottom electrode, patterned by the lift-off process with photolithography. Approximately 3nm-thick HfO_2 thin films were deposited by thermal atomic layer deposition (ALD) using tetrakis-ethylmethylamino hafnium (TEMA-Hf) and deionized water vapor as a metal organic precursor and an oxygen source, respectively, at a wafer temperature of 250°C. A Pt(40nm)/Ti(20nm) top electrode was deposited by electron-beam evaporation through photolithographic patterning in order to form a crossbar junction device. Figure 1(b) shows a schematic diagram of the device stack. The two-terminal current - voltage (I - V) characteristics of the devices were measured using a semiconductor parameter analyzer (HP-4155A), as shown at the bottom of Fig. 1(b). A quasi-DC voltage sweep (sweep rate from 10^0 to 10^2V/sec) was applied to the top electrode with the bottom contact grounded at ambient temperature in all of the electrical measurements. Figure 1(c) shows DC I-V curves in log I – linear V scale as measured from the ten randomly chosen devices in the CBA, which is composed of the Ti/HfO_2/Ti memristors. The EF process was not necessary for the initial resistive switching from the virgin state. Current compliance (500 μA in this curves) was used for the set transition from the HRS to the LRS. Figure 1(d) shows the resistance values of LRS and HRS, read at +0.3V from three-times I-V measurement in the ten devices. Lower variability is observed in the resistance values of LRS than in those of HRS, which could be beneficial in the BNNs. It should be noted that the CBA does not adopt any selector, so there could be a rather significant sneak current problem involved. Nonetheless, all the randomly selected

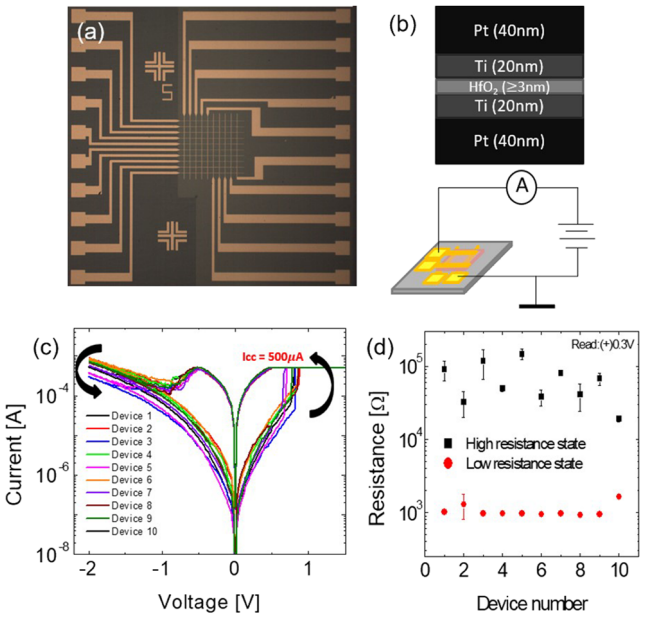

FIG. 1. Fabricated forming-free memristor crossbar array. (a) Optical microscopic image of 10x10 crossbar array with 5μm junction width. The ten contact pads on the left are the bottom electrodes while the other ten on the right are the top electrodes. (b) Schematic diagram of the device structure and measurement set-up. The memristors with a material stack of Pt (40nm)/Ti (20nm)/HfO_2 (≥3nm)/Ti (20nm)/Pt (40nm). (c) I-V curves for ten randomly selected devices. (d) Average and standard deviation of resistance value for low/high resistance state at a read voltage of (+) 0.3V for three-times switching cycles for ten devices.

cells perform quite nicely, which would not be the case if the devices required EF.

Since the synaptic weights of BNNs are either +1 or -1 with no intermediate values, the two conductance values (inverse of resistance value) can be used as the synaptic weights. Using the conductance, as opposed to resistance, in BNN is beneficial for mitigating the relatively high variability problem of the HRS. This is because the conductance values of the HRS are much lower than those of the LRS (by ~ two orders of magnitude), the degree of variation also becomes smaller. The variabilities in the set and reset voltages were much less severe than those of resistance. (Fig. S1 of online supplementary material).

The impact of EF-free memristors (Ti/HfO_2/Ti) in CBAs was investigated in 2×2 CBAs. Since there is no additional selector device or rectifying function in the memristor device itself, the sneak current into the defective neighboring device in the CBA could induce larger variability in the resistance state and programming voltage. Figure 2 shows I-V loops of ten sets of 2×2 CBAs confirming the yield of the device and its ability to effectively work with defective devices. Normal devices showed typical I-V loops, while two types of defective devices – those working with smaller resistance ratio (Defective (I)) and LRS-stuck (Defective (II)) – showed malfunctioning I-V loops. The device yield was about 70% (28 normal devices observed among 40 total devices). Although several CBAs possessed more than two

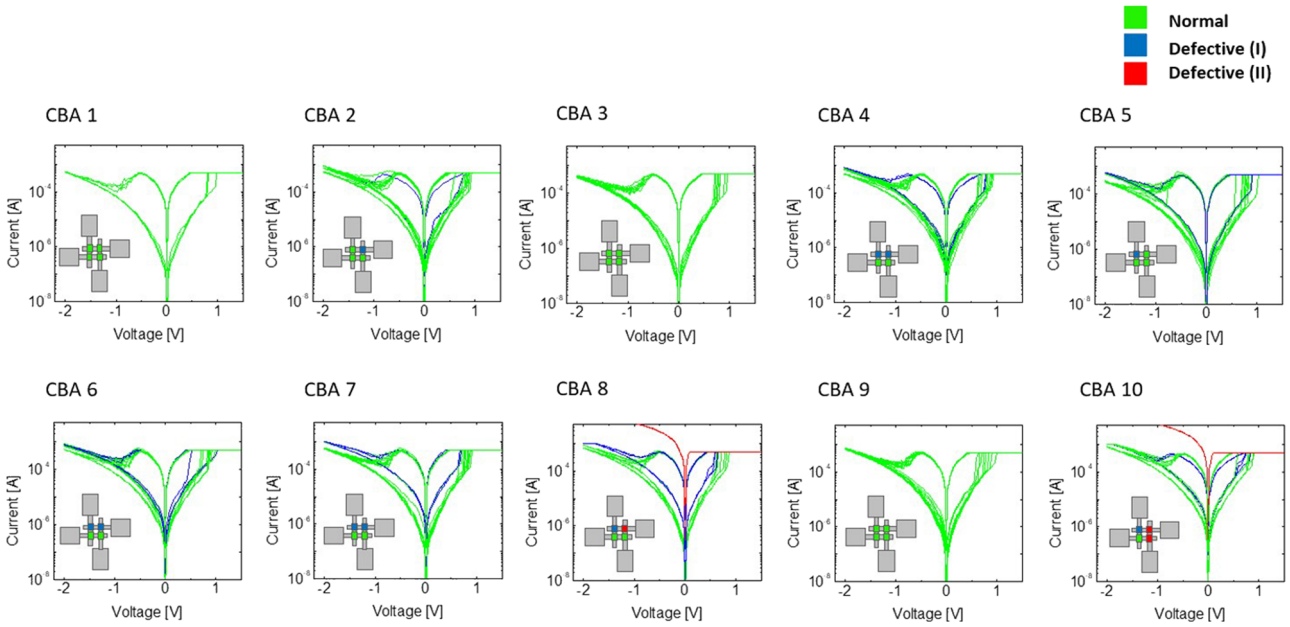

FIG. 2. Effect of defective devices in CBAs on electrical characteristics by measuring the switching curves of the Ti/HfO_2/Ti devices in 2x2 CBAs. The inset is a schematic illustration of 2x2 CBA showing that the four devices were working. Each device was marked by green, yellow, and red based on the status of the resistance switching properties as working normally (Normal), working with small resistance ratio (Defective (I)), and broken-down (Defective (II)), respectively.

defective devices, normal devices still worked effectively. By contrast, 2×2 CBAs with EF memristor (Ti/HfO_2/Pt) showed much larger device variability with defective devices. The EF process and typical I-V loops measured from EF memristor (Ti/HfO_2/Pt) are presented in Fig. S2 of supplementary material. The low variability and higher device yield in a CBA are crucial factors for application to the BNNs.

It is believed that the EF process could affect device variability because a different nature (position, size, composition, etc.) of CF is generated in the individual device during the EF process. Generally, the EF process demands a higher voltage than that of resistive switching in a functioning device. A potential issue of the EF process in a CBA is that the high voltage and long stress time (owing to the long integration time and multiple steps of DC sweep measurement protocol) may induce a large discharging current – from the capacitor-like memristors and parasitic capacitive elements in a CBA – passing through the electroformed device during the EF process.[25] From a material perspective, the EF-free device initially possesses a large amount of oxygen-related defects, such as oxygen vacancies, in the interface of the reactive metal (Ti) and insulator (HfO_2) of the top and bottom electrodes. The fact that the ALD process slightly oxidized the underlying metal (Ti) layer could lead to an asymmetric relationship between the top and bottom electrodes for the BRS and also provide necessary oxygen vacancies.[26,27] In addition, the lower barrier height of Ti/HfO_2 could also induce a fluent electron injection from the electrode to the insulator, which resulted in the facile formation of CF without the aid of the high electric field-induced EF process in EF memristor (Ti/HfO_2/Pt).

Figure 3(a) and (b) show the changes in the I-V loops of EF-free and EF memristor devices in 2×2 CBAs with increasing numbers of defective cells. Even one normal EF-free device with three defective cells (mixed with type I and II defective cells) shows almost similar I-V loops, as shown in Fig. 3(a). By contrast, an EF memristor device exhibited memristive I-V loops with large variation with one defective neighboring cell, as shown in Fig. 3(b). However, normally working device showed leaky I-V curves when it was surrounded by three defective cells in a CBA. It appeared that the normal device was misled to an LRS-stuck state, which means that no bias was applied to the selected device in the HRS since the short circuit was created by three neighboring cells in a broken-down state (type II). HSPICE modeling was performed in order to confirm the effect of neighboring cells in defective states (type I and II). Here, the resistance was set to 141 × 10^3 and 990Ω for HRS and LRS in a normal cell, respectively, 18.6 and 1.6KΩ for HRS and LRS in a type I defective cell, respectively, and 180Ω for a type II defective cell. Fig. 3(c) shows that a normal cell can work with a lower resistance ratio, even when it is surrounded by one type I and two type II defective cells. However, three type II cells can prohibit the working and reading of normal cells due to the formation of a short circuit through the sneak path, as shown in Fig. 3(d). Therefore, it was considered that the sneak path current issue could be mitigated in a CBA consisting of EF-free memristors because of the higher device yield, lower variability, and less severe malfunction (e. g. hard break-down).

An EF-free memristor was applied to the BNN for simple pattern recognition. Fig. 4(a) shows a schematic diagram of BNN with double-column CBA to calculate binary synaptic weights

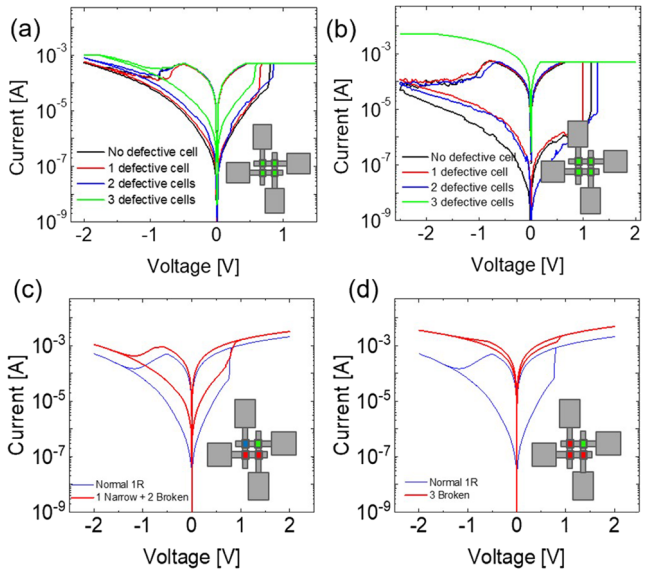

FIG. 3. Variation of I-V loops with an increasing number of defective devices in 2x2 CBAs. (a) I-V loops for forming-free devices (Ti/HfO_2/Ti). (b) I-V loops for electro-formed devices (Ti/HfO_2/Pt). HSPICE simulation result of I-V loops for a normal EF device surrounded by (c) two defective (I) cells and one defective (II) cell, and (d) three defective (II) cells.

of +1 and -1.[21] In this simulation, a single EF-free memristor device was used for two 3×3 CBAs to form a BNN hardware. Memristors can be programmed to either HRS or LRS corresponding to the conductance of g_{HRS} or g_{LRS}, respectively. Such a conductance could be the synaptic weight between the input and out patterns. The binary synaptic weights (HRS or LRS) are stored on the CBA in BNN. In this work, BNN is composed of a double-column CBA for simulation. In Fig. 4(a), the inverted synaptic weights for the left column in the M+ CBA are stored in the right column in the M- CBA. The input pattern represents the input neurons delivering voltage pulses to the CBA. The input pattern as a vector, [x_1, x_2, x_3], can be multiplied to the corresponding synaptic weights, [g_1, g_2, g_3] in the M+ and M- CBAs, resulting in the summation of current in the same columns corresponding to the two CBAs. Then, the column current is converted to a voltage according to the activation function, such as ReLU (rectified linear unit) or Sigmoid function. As such, the output vector, [y_1, y_2, y_3], could be used as an output neuron or a hidden neuron for the next layer in a BNN. In the general form, the jth output neuron, y_j, can be formulated with the following equation;

$$y_j = f\{\sum_0^n (x_i \cdot g_{i,j+} - x_i \cdot g_{i,j-})\} = f\{\sum_0^n x_i \cdot (g_{i,j+} - g_{i,j-})\} \quad (1)$$

Here, $g_{i,j+}$ and $g_{i,j-}$ are the conductance representing a synaptic weight of either +1 or -1 in the M+ and M-, respectively. f means the activation function of the neuron circuits. Since g_{LRS} is greater than g_{HRS}, ($g_{i,j+}$-$g_{i,j-}$) can be approximated by +g_{LRS} or −g_{LRS}, which can be interpreted as +1 or -1. By using a double-column scheme, conductance or synaptic weight could be clearly differentiated.

As an exemplary work, three patterns of [LRS-LRS-LRS], [LRS-LRS-HRS], and [HRS-LRS-LRS] were programmed on the first, second, and third columns, respectively, of the M+ CBA. Similarly, their complement patterns were stored on the M- CBA, as shown in Fig. 4(a). Here, the measured conductance values in I-V curves of EF-free memristor in a normal state were used for simulation. Fig. 4(b) shows the measured column current of I_{1+}-I_{1-}, where I_{1+} and I_{1-} represent the first column currents of M+ and M-, respectively. The x-axis represents eight input patterns from [000] to [111]. The measured currents of I_{1+}-I_{1-} for the eight input patterns can be the values calculated with Eq. (1). Similarly, Fig. 4(c) shows the column current of I_{3+}-I_{3-}, as measured from the third columns in both M+ and M-. The measured column currents in Figs. 4(b) and (c) for the eight input patterns clearly show the Eq. (1) can be calculated on the M+ and M- CBAs. Therefore, the input patterns of [111], [110], and [011] are serially identified by the three columns, respectively, in this particular pre-trained weight matrix in CBAs. If the sizes of M+ and M- CBAs are scaled up larger in terms of the number of arrays, it becomes possible to calculate the Eq. (1) using the two CBAs to test more complicated vectors, such as MNIST (Modified National Institute of Standards and Technology). The MNIST input vector is composed of 28×28 pixels. Assuming that there are 1024 hidden neurons, 784 input signals are applied to 784 rows of the expanded M+ and M- CBAs, and accordingly, 1024 column currents for 1024 hidden neurons can be obtained. The number of input, hidden, and output neurons were 784, 1024,

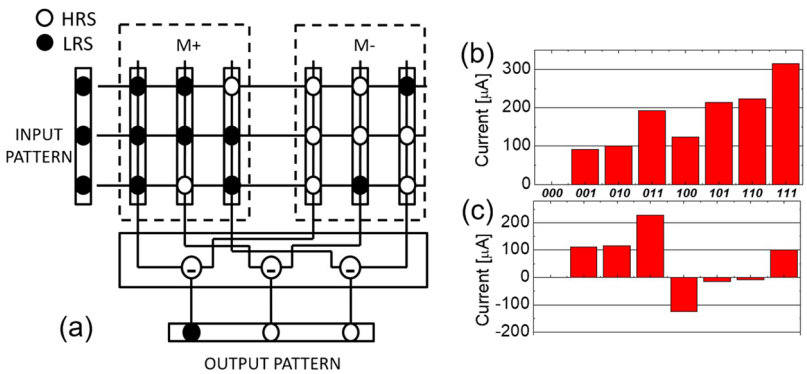

FIG. 4. Forming-free memristor applied to BNN. (a) Schematic diagram of the memristor BNN structure consisting of a double-column 3×3 crossbar array. The open and closed circles represent memristors in HRS and LRS, respectively. Inverted synaptic weights for the left column in the M+ CBA are stored in the right column in the M- CBA. (b) The measured output current for the input voltage ([111]) is consistent with [LRS LRS LRS]. (c) The measured output current for the input voltage ([011]) is consistent with [HRS LRS LRS].

and 10, accordingly, the memristor array size used for the simulation was 784×1024 and 1024×10, for the first and second layers. Due to the lack of memristor array size, simulation was conducted based on the single memristor device. By using a certain appropriate activation function, it is possible to map the typical neural networks to the CBA architecture. It has been estimated that the MNIST recognition rate can be as high as 96.1% for these memristor-based BNNs, despite its adoption of the simple binay weight values.[21] When the memristors with a gradual change of conductance were adopted in a similar network structure, the MNIST recognition rate can be improved to be slightly better than this memristor-based BNNs. However, such a task demands more energy and highly demanding control of the multi-level conductance of the memristors.

In comparison with the previous work on HfO_2-based ReRAM for CNNs done by Garbin et al. in Ref. 24, there are two main different aspects could be highlighted in this work. First, the device stack of Ti/HfO_2/TiN or Ti/HfO_2/Pt used in Ref. 24 necessarily demanded the EF process showing the inverse relationship of time-voltage dependence. In the case of EF-free Ti/HfO_2/Ti in this work, however, the EF process could be diminished by replacing the inert Pt or TiN bottom electrode into Ti active metal. In addition, device variability and failure could be largely improved, which is also important parameters for the high recognition rate for BNN. Second, although binary approach was stated in Ref. 24, analog synaptic weight was still adopted by combining *n* multiple binary memristor cells operating in parallel for utilizing a single analog synapse. On the other contrary, BNN used in this work only demands binary synaptic weights without requiring the calculation of any high-precision multi-bit numbers.

In summary, EF-free memristor CBAs with Ti/HfO_2/Ti structure were fabricated in micro-scale. As compared to electroformed memristor with Ti/HfO_2/Pt structure, the EF-free memristor in CBA showed better performance, higher device yield, lower variability in resistance and switching voltage, and less severe malfunction, such as hard break-down. Such EF-free memristor CBAs were applied to the BNN hardware using a two-column scheme. Eight input patterns were applied to the CBAs with pre-trained synaptic weights, which resulted in the corresponding pattern recognition by maximizing the current summation. This simple memristor CBA with the absence of the selector device or multi-level operation could significantly reduce the burden for fabrication. It is expected that the EF-free memristor CBA could be used for a pattern recognition system in edge devices or IoT applications.

See supplementary material for the supporting information on the device.

This work was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-IT1701-07.

REFERENCES

[1] A. C. Torrezan, J. P. Strachan, G. Medeiros-Ribeiro, and R. S. Williams, Nanotechnol. **22**, 485203 (2011).
[2] M.-J. Lee, C. B. Lee, D. Lee, S. R. Lee, M. Chang, J. H. Hur, Y. Kim, C. Kim, D. H. Seo, S. Seo, U. Chung, I. Yoo, and K. Kim, Nat. Mater. **10**, 625 (2011).
[3] M. A. Zidan, J. P. Strachan, and W. D. Lu, Nat. Electron. **1**, 22 (2018).
[4] S. Pi, C. Li, H. Jiang, W. Xia, H. Xin, J. J. Yang, and Q. Xia, Nat. Nanotechnol. **14**, 35 (2018).
[5] Y. Yang, B. Chen, and W. D. Lu, Adv. Mater. **27**, 7720 (2015).
[6] J. J. Yang, D. B. Strukov, and D. R. Stewart, Nat. Nanotechnol. **8**, 13 (2013).
[7] C. Li, M. Hu, Y. Li, H. Jiang, N. Ge, E. Montgomery, J. Zhang, W. Song, N. Dávila, C. E. Graves, Z. Li, J. P. Strachan, P. Lin, Z. Wang, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, and Q. Xia, Nat. Electron. **1**, 52 (2017).
[8] D. S. Jeong and C. S. Hwang, Adv. Mater. **30**, 1704729 (2018).
[9] M. Hu, C. E. Graves, C. Li, Y. Li, N. Ge, E. Montgomery, N. Davila, H. Jiang, R. S. Williams, J. J. Yang, Q. Xia, and J. P. Strachan, Adv. Mater. **30**, 1705914 (2018).
[10] Y. Kim, Y. J. Kwon, D. E. Kwon, K. J. Yoon, J. H. Yoon, S. Yoo, H. J. Kim, T. H. Park, J. W. Han, K. M. Kim, and C. S. Hwang, Adv. Mater. **30**, 1704320 (2018).
[11] M. A. Zidan, Y. Jeong, and W. D. Lu, in Proc. IEEE/ACM Int. Symp. Nanoscale Archit. NANOARCH 2017 (2017), pp. 81–82.
[12] C. Du, F. Cai, M. A. Zidan, W. Ma, S. H. Lee, and W. D. Lu, Nat. Comm. **8**, 2204 (2017).
[13] P. Yao, H. Wu, B. Gao, S. B. Eryilmaz, X. Huang, W. Zhang, Q. Zhang, N. Deng, L. Shi, H. S. P. Wong, and H. Qian, Nat. Comm. **8**, 15199 (2017).
[14] Z. Wang, S. Joshi, S. E. Savel'ev, H. Jiang, R. Midya, P. Lin, M. Hu, N. Ge, J. P. Strachan, Z. Li, Q. Wu, M. Barnell, G. L. Li, H. L. Xin, R. S. Williams, Q. Xia, and J. J. Yang, Nat. Mater. **16**, 101 (2017).
[15] N. Ge, J. H. Yoon, M. Hu, E. J. Merced-Grafals, N. Davila, J. P. Strachan, Z. Li, H. Holder, Q. Xia, R. S. Williams, X. Zhou, and J. J. Yang, Sci. Rep. **7**, 40135 (2017).
[16] C. C. Hsieh, A. Roy, Y. F. Chang, D. Shahrjerdi, and S. K. Banerjee, Appl. Phys. Lett. **109**, 223501 (2016).
[17] D. S. Jeong, K. M. Kim, S. Kim, B. J. Choi, and C. S. Hwang, Adv. Electron. Mater. **2**, 1600090 (2016).
[18] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, Nature **521**, 61 (2015).
[19] S. Kim, C. Du, P. Sheridan, W. Ma, S. Choi, and W. D. Lu, Nano Lett. **15**, 2203 (2015).
[20] J. H. Yoon, Z. Wang, K. M. Kim, H. Wu, V. Ravichandran, Q. Xia, C. S. Hwang, and J. J. Yang, Nat. Comm. **9**, 417 (2018).
[21] K. Van Pham, T. Van Nguyen, S. B. Tran, and H. Nam, J. Semicon. Tech. Sci. **18**, 568 (2018).
[22] D. S. Jeong, B. J. Choi, and C. S. Hwang, in *Resist. Switch* (John Wiley & Sons, Ltd, 2016), pp. 289–316.
[23] K. J. Yoon, W. Bae, D. K. Jeong, and C. S. Hwang, Adv. Electron. Mater. **2**, 1600326 (2016).
[24] D. Garbin, E. Vianello, O. Bichler, Q. Rafhay, C. Gamrat, G. Ghibaudo, B. DeSalvo, and L. Perniola, IEEE Trans. Electron Devices **62**, 2494 (2015).
[25] S. J. Song, K. M. Kim, G. H. Kim, M. H. Lee, J. Y. Seok, R. Jung, and C. S. Hwang, Appl. Phys. Lett. **96**, 112904 (2010).
[26] T. H. Park, S. J. Song, H. J. Kim, S. G. Kim, S. Chung, B. Y. Kim, K. J. Lee, K. M. Kim, B. J. Choi, and C. S. Hwang, Sci. Rep. **5**, 15965 (2015).
[27] T. H. Park, S. J. Song, H. J. Kim, S. G. Kim, S. Chung, B. Y. Kim, K. J. Lee, K. M. Kim, B. J. Choi, and C. S. Hwang, Phys. Stat. Sol. **9**, 362 (2015).