The Origin of CBRAM With High Linearity, ON/OFF Ratio, and State Number for Neuromorphic Computing

Yanming Liu, Jialu Gao, Fan Wu, He Tian, *Senior Member, IEEE*, and Tian-Ling Ren, *Senior Member, IEEE*

***Abstract*— Conductive bridge random access memory (CBRAM) has been concentrated recently for its ultrasmall size, low power consumption, synaptic characteristics, and application in neuromorphic computing. However, the accuracy of the CBRAM array-based neural network is not high enough due to the low linearity, limited ON/OFF ratio, and the number of states. The original illustration and the optimization methods are still paucity. In this work, the origin of the characteristics of CBRAM has been revealed from the filament distribution of the devices, which inspires us to design an inserted graphene structure of CBRAM and preset seeds leading to high linearity (0.995), ON/OFF ratio (26.4), and the number of states (63). The Monte Carlo simulation results reveal that the CBRAM with more seeds can promote a larger number of potential advantage path (PAP) conducing better characteristics. Moreover, the PAP can be modulated by the number of preset seeds. Finally, a handwritten recognition neural network has been realized by using a 1T-1R array, and high recognition accuracy (92%) has been obtained, which shows that devices with higher PAP can eventually promote higher recognition accuracy.**

***Index Terms*— Conductive bridge random access memory (CBRAM), high linearity, number of states, ON/OFF ratio.**

I. Introduction

With the rapid development of Artificial Intelligence (AI) technology, problems, such as high power dissipation and high transmission consumption, are limited by conventional computing hardware [1]. In order to solve these problems, neuromorphic computing that is based on synaptic devices has been advanced breaking the restriction of the von Neumann architecture. As an integrated system, the characteristics of bottom devices as memories and transistors would highly determine the accuracy of the neuromorphic computing structure. The amount of previous work has concentrated on enhancing the accuracy of the neural network based on synaptic devices to promote the development of neuromorphic computing [2]–[4]. Their works focused on the composition of the circuits and established the connection of the neural network algorithm and the bottom devices. Choi *et al.* [5] emulated a handwritten recognition based on the SiGe epitaxial memory. Zhou *et al.* [6] simulated a letter recognition system based on the optoelectronic resistive random access memory (RRAM). These studies indicated that achieving high accuracy requires devices optimization. However, the device design for high linearity, high ON/OFF ratio, and enough number of states is still elusive. In this work, a device design is provided about how to realize high linearity, high ON/OFF ratio, and enough number of states. Moreover, the particle distribution perspective inside the device is also provided.

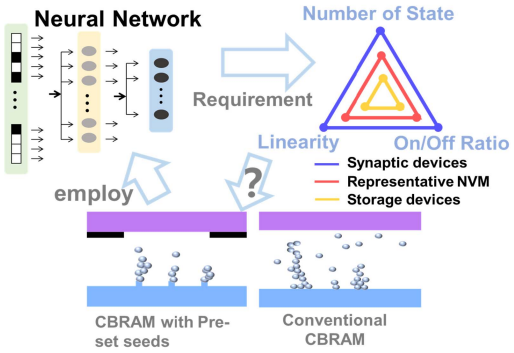

Fig. 1. Schematic of the relation about neuromorphic computing from the most bottom to the top applications.

In this work, the origin of high-performance conductive bridge random access memory (CBRAM) has been explored to understand the device physics and to know the strategy to improve the accuracy of the neural network. Fig. 1 shows the flowchart of this work. For satisfying the requirements of the neural network hardware, a device model has been built to enable CBRAM with high linearity, high ON/OFF ratio, and enough number of states. A structure of CBRAM with preset seeds and inserted nanohole graphene has been proposed. The graphene can be used to confine the drift of the

Manuscript received January 7, 2021; revised February 12, 2021; accepted March 4, 2021. Date of publication March 23, 2021; date of current version April 22, 2021. This work was supported in part by the National Key Research and Development Program under Grant 2016YFA0200400; in part by the National Natural Science Foundation of China under Grant 62022047, Grant 61874065, and Grant 51861145202; and in part by the Beijing Innovation Center for Future Chips, Tsinghua University and the Independent Research Program of Tsinghua University under Grant 20193080047. The work of He Tian was supported in part by the Young Elite Scientists Sponsorship Program by the China Association for Science and Technology (CAST) under Grant 2018QNRC001 and in part by the Fok Ying-Tong Education Foundation under Grant 171051. The review of this brief was arranged by Editor P. Narayanan. *(Corresponding authors: He Tian; Tian-Ling Ren.)*

Yanming Liu is with the School of Aerospace Engineering, Tsinghua University, Beijing 100084, China, and also with Institute of Microelectronics, Tsinghua University, Beijing 100084, China.

Jialu Gao is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.

Fan Wu, He Tian, and Tian-Ling Ren are with the Institute of Microelectronics, Tsinghua University, Beijing 100084, China, and also with the Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing 100084, China (e-mail: tianhe88@tsinghua.edu.cn; rentl@tsinghua.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2021.3065013.

Digital Object Identifier 10.1109/TED.2021.3065013

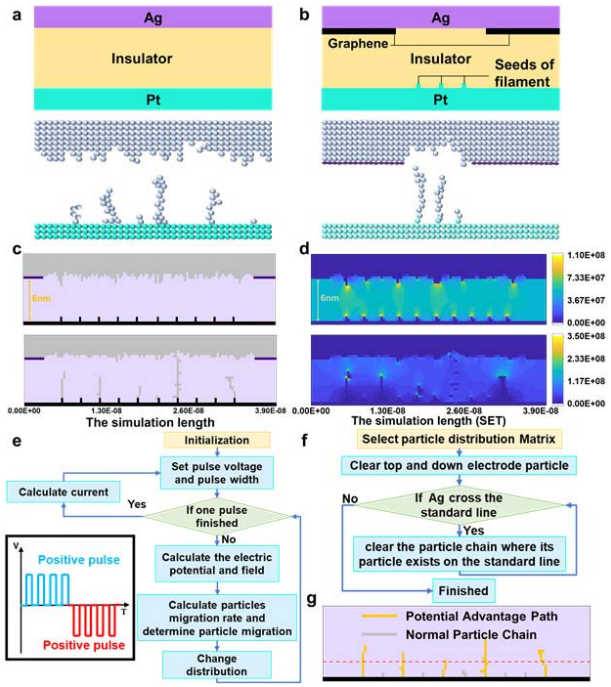

Fig. 2. (a) Basic structure of CBRAM. (b) CBRAM with inserted graphene and preset seeds. (c) Particle distribution of the initial state of CBRAM (preset seeds = 10) and the state after the SET process. (d) Corresponding electrical field distribution. (e) Flowchart of the simulation. (f) Flowchart of the method to judge PAP. (g) Definition of the PAP.

active metal particle, which promotes the low power of the devices and modulates the growth of filaments. The seeds are the protuberances on the surface of the bottom electrode, which can further precisely manipulate the formation of the filament. The purpose of modulating the growth of filaments can be achieved by controlling the seeds, finally leading to the requirement characteristic. For verifying the model, devices are used in the handwritten recognition algorithm and perform high recognition accuracy. This work has indicated a method of optimization of the characteristic of CBRAM promoting the development of neuromorphic computing, which establishes the link from the most bottom to the top applications.

II. Device Design for Neuromorphic Computing

A. Structure of CBRAM

The typically CBRAM top electrode material is Ag or Cu [7], [8]. We choose Ag top electrode as an example through the whole simulation, and the Cu top electrode can also be controlled by spatial confinement [9], [10]. The filaments in the conventional CBRAM [see Fig. 2(a)] are uncontrollable and stochastic leading to low linearity and unsuccessful to satisfy the requirement from neuromorphic computing. In order to solve the above problems, we proposed to use a graphene nanohole layer. The inserting of graphene at the CBRAM interface has two functions. One is controlling the size of filaments, and another is low-power operation. The inserted graphene nanohole layer can confine the growth of the filament by screening the particle movement in the graphene-covered region [7]. The inserting of graphene can also reduce the power consumption of devices due to the larger interface resistance between the insulator and the top electrode [11]. Fig. 2(b) shows the proposed device structure with graphene nanohole layer and preset seeds, which can control the distribution of the filaments. The graphene nanohole can be created via experiment by atomic force microscopy tip [8] or focused ion beam [12], and the preset seeds can be formed by a series of small prepulse. The number of seeds can be modulated and regarded as the derivation of the filaments. The resistance is highly influenced by filament distribution. Fig. 2(c) shows the particle distribution of CBRAM with ten seeds before and after the SET process, which shows the disparity between the initial distribution and the final distribution. The corresponding electrical field distribution has been shown in Fig. 2(d). Comparing Fig. 2(c) and (d), it is found that most of the drift particles that accumulate on top of the preset seeds have a strong electrical field, which is easier to attract the cation to deoxygenize on it.

B. Simulation of CBRAM

Due to the stochastic of the filament growth, a model to simulate the behavior of the CBRAM under the pulse mode has been established by the Monte Carlo method [see Fig. 2(e)]. In this model, the positive pulses have been applied until the filament contacts the top electrode, and then, the voltage reverses. The effect of the filament drift for the current calculation has been considered, which improves the resistance model in the literature [7]. The major formulation can be expressed as

$$R = \left(\frac{n_1+\alpha}{n_0}\right)^{\beta} \times R_1 + \left(1 - \left(\frac{n_1+\alpha}{n_0}\right)^{\beta}\right) \times R_2 \quad (1)$$

where R is the representative resistance of device; R_1 and R_2 are metal and insulator unit resistances, respectively; n_1 and n_0 are the numbers of the drift metal and insulator particle, respectively; α is a parameter that is determined by the number of seeds; and β is a nonlinear coefficient. Moreover, we optimize the tunnel current model. The tunnel currents that are generated by multiple filaments have been considered. The initial tunnel current can be expressed as

$$I_2 = A\frac{4\pi q^2 m}{h_0^3 \alpha^2 \phi_0}\left(\frac{V}{d}\right)^2 e^{-\frac{2ad\sqrt{q\phi_0^3}}{3V}}, \quad \alpha = \frac{4\pi\sqrt{2\pi}}{h_0}, \quad V > \phi_0 \quad (2)$$

where d is the distance between the top electrode and the peak of the filament, h_0 is the Planck constant, ϕ_0 is the barrier height with zero bias, m is the effective mass of the electron, and A is the area of section filament. We introduce the regulation method to recalculate I_2 instead of only calculating the maximum d.

C. Description of Device Behavior

The potential advantage path (PAP) has been defined to evaluate the behavior of the devices. Fig. 2(g) shows the schematic of the PAP, which has to be treated as the length of the particle chain larger than 30% height of the insulator layer. The connected domain that intersects the red line is defined as PAP. The selected particle distribution must be the state after the SET process guarantying the filament has sufficient growth, as it can exhibit more information about

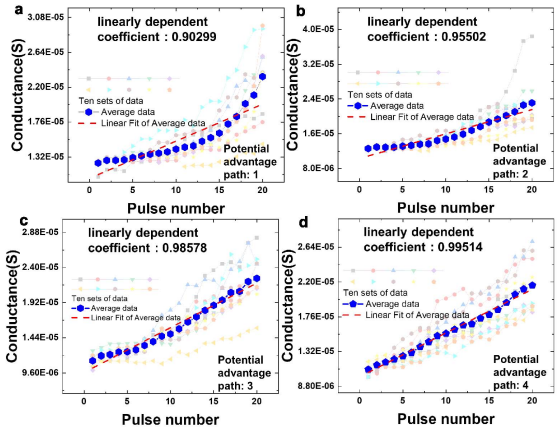

Fig. 3. Conductance variation under applied pulse with different numbers of PAP shows the diverse linearity. The numbers of PAP in (a)–(d) are 1, 2, 3, and 4, respectively. The linearity of the *C–P* curves enhances as the PAP increases.

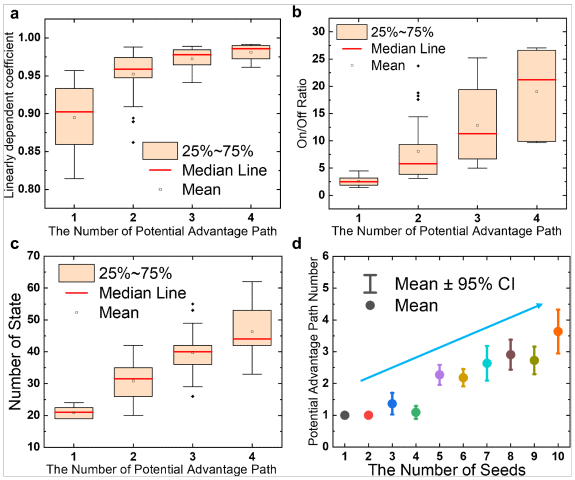

Fig. 4. (a) Deviation of the linearity-dependent coefficient, (b) ON/OFF ratio, and (c) number of states with different number of PAP, respectively. (d) Number of PAP will gradually increase as the number of preset seeds increases.

the distribution of filament. The schematic of the flowchart of judging PAP has been shown in Fig. 2(f). The PAP is used to evaluate the behavior of the devices, which can reveal the origin of the linearity, ON/OFF ratio, and the number of states.

III. Device Behaviors Results and Discussion

The definition of PAP shows the trend in describing the behavior of the devices. Fig. 3 shows that the linearity of the conductance–pulse curve (*C–P* curve) would increase with the number of PAP increasing. The devices with the number of PAP from 1 to 4 have been emulated. For each case, ten sets of data have been selected to depict. The average lines have been calculated and plotted as a blue line. The linearly dependent coefficients of the average of each set are 0.903, 0.955, 0.986, and 0.995, respectively, showing the increasing trend. In addition, Fig. 4(a) shows the plots of the deviation of the linearly dependent coefficient of each set of data, which demonstrates that, with the PAP increasing, the deviations have gradually decreased and the linearities have increased. The gradually decreasing deviation manifests the increasing device-to-device uniformity. Moreover, with the increasing

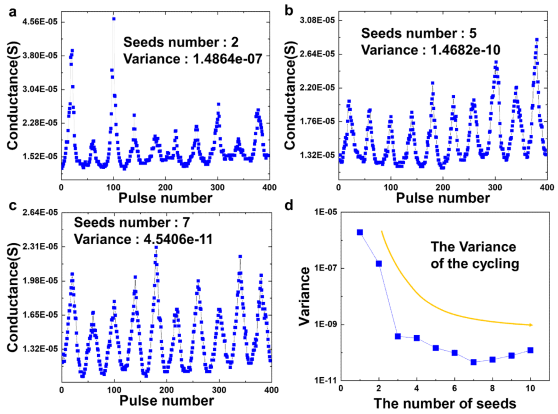

Fig. 5. (a) Preset seeds = 2. (b) Preset seeds = 5. (c) Preset seeds = 7. Gradual switching in CBRAM using constant voltage pulses applied to the top electrode. (d) Schematic of the variance of the cycling. With the increasing number of seeds, the variance of the cycling has gradually decreased.

number of PAP, the ON/OFF ratio and the number of states have a similar growth trend with the linearity. The results show that the ON/OFF ratio as PAP = 4 is eight times larger than that as PAP = 1 [see Fig. 4(b)]. Besides, the increment of the state's number as PAP increasing also satisfies the analogical trend [see Fig. 4(c)]. The results indicate that the higher number of PAP will bring better characteristics. In order to find a way to enhance the number of the PAP, the seeds from 1 to 10 have been emulated and been depicted the corresponding number of the PAP in Fig. 4(d). The simulation result asserts that increment the number of the seeds can gradually increase the number of the PAP.

Besides, the cycle-to-cycle stability of our devices has also been studied. Fig. 5(a) shows *C–P* curves of ten cycles under 20 positive pulses and 20 negative pulses applied with CBRAM (preset seeds = 2). Fig. 5(b) and (c) shows similar curves with CBRAM (preset seeds = 5.7), respectively. The variances of cycling curves are shown in Fig. 5(d), which demonstrates that CBRAM with more preset seeds (implying more PAP) possess lower variance and also means increasing cycle-to-cycle stability.

Finally, a benchmark has been shown in Table I. Previous literature shows limited linearity, ON/OFF ratio, and the number of states, which is similar to the behaviors of CBRAM with a low number of PAP. This indicates that the devices in literature may have a low number of PAP, which leads to limited linearity, ON/OFF ratio, and the number of states. Moreover, the device with a high number of seeds implying high PAP displays that the linearity, ON/OFF ratio, and the number of states can reach 0.995, 26.4, and 63, respectively, which is higher than the devices in the previous literature.

IV. Application for Neuromorphic Computing

In order to verify the practical performance of our devices, a handwritten recognition neural network has been set up based on the preset CBRAM. The Modified National Institute of Standards and Technology (MNIST) database has been used to train the network. The bottom circuit of the connection between different layers of the neural network has been shown in Fig. 6(a). This circuit has two modes as writing and reading

TABLE I
PERFORMANCE COMPARISON

Structure	On/Off ratio	Linearity	Number of states	Reference
Ag/Graphene/HfO_2/Pt (10 seeds)	26.4	0.991	63	This work
Ag/Graphene/HfO_2/Pt (1 seeds)	3.05	0.932	23	This work
Cu/AlO_x/a-CO_x/TiN_xO_y/TiN	2.88	0.977	≈37	[13]
Ag/Ag-Si mixed/Ag	14.1	0.968	≈51	[14]
ZrTe/Al_2O_3/Ta	2.75	0.978	≈29	[15]
AlO_x/TiN_x/PCMO	3.58	0.962	≈50	[16]
Ti/HfO_x/HfO_2/AlO_x	2.43	0.959	≈25	[17]
Cu/pV3D3/Al/PES	6.10	0.939	≈38	[18]
Ag-Cu/SiO_2/Au	3.68	0.914	≈18	[19]

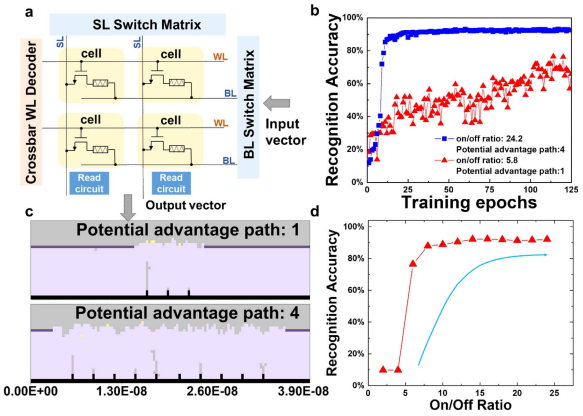

Fig. 6. (a) Circuit design of the handwritten recognition. (b) Recognition accuracy variation with the increasing epochs of devices with dissimilar ON/OFF ratios for different numbers of PAP. (c) Particle distribution with PAP = 1 and PAP = 4. (d) Recognition accuracy variation by the devices with the increasing on/off ratio under 125 epochs.

modes that are controlled by the word line (WL), achieving computing and storage on the same devices. Two sets of parameters from the devices with different seeds have been used to train the neural network, as shown in Fig. 6(b). The device with a high ON/OFF ratio shows better accuracy with the corresponding trend results from the literature [2]. Fig. 6(c) shows the particle distribution after the SET process with different PAPs, leading to distinct ON/OFF ratio and, finally, reflecting on the recognition accuracy. Fig. 6(d) depicts the effect of the recognition accuracy by ON/OFF ratio. Results show that devices with higher PAP can eventually promote higher recognition accuracy.

V. CONCLUSION

In conclusion, we propose a structure of CBRAM with preset seeds, which leads to extremely high linearity, high ON/OFF ratio, and enough state of devices. PAP has been defined to describe the behavior of the devices and show the trend of modulating the linearity, ON/OFF ratio, and the number of states. Moreover, the relation between the number of seeds and PAP has been discussed, which indicates a way to enhance the performance. The cycle behavior of devices has also been studied. These results indicate that the increasing number of PAPs can lead to high linearity, high ON/OFF ratio, a high number of states, and low cycling variance. Finally, devices with higher PAP have been used to compose a neuromorphic computing application, which shows high recognition accuracy. This work reveals the origin of high linearity, ON/OFF ratio, and the number of states in the device level, which can promote the neuromorphic applications from the most bottom.

REFERENCES

[1] H.-J. Yoo, "1.2 Intelligence on silicon: From deep-neural-network accelerators to brain mimicking AI-SoCs," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 20–26, doi: 10.1109/ISSCC.2019.8662469.

[2] S. Yu, P.-Y. Chen, Y. Cao, L. Xia, Y. Wang, and H. Wu, "Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect," in *IEDM Tech. Dig.*, Dec. 2015, pp. 17.13.11–17.13.14, doi: 10.1109/IEDM.2015.7409718.

[3] M. Zhao *et al.*, "Investigation of statistical retention of filamentary analog RRAM for neuromophic computing," in *IEDM Tech. Dig.*, Dec. 2017, pp. 39.34.31–39.34.34, doi: 10.1109/IEDM.2017.8268522.

[4] P. Yao *et al.*, "Face classification using electronic synapses," *Nature Commun.*, vol. 8, no. 1, p. 15199, Aug. 2017, doi: 10.1038/ncomms15199.

[5] S. Choi *et al.*, "SiGe epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations," *Nature Mater.*, vol. 17, no. 4, pp. 335–340, Apr. 2018, doi: 10.1038/s41563-017-0001-5.

[6] F. Zhou *et al.*, "Optoelectronic resistive random access memory for neuromorphic vision sensors," *Nature Nanotechnol.*, vol. 14, no. 8, pp. 776–782, Aug. 2019, doi: 10.1038/s41565-019-0501-3.

[7] Y. Liu, K. Yang, X. Wang, H. Tian, and T.-L. Ren, "Lower power, better uniformity, and stability CBRAM enabled by graphene nanohole interface engineering," *IEEE Trans. Electron Devices*, vol. 67, no. 3, pp. 984–988, Mar. 2020, doi: 10.1109/TED.2020.2968731.

[8] X. Zhao *et al.*, "Confining cation injection to enhance CBRAM performance by nanopore graphene layer," *Small*, vol. 13, no. 35, Sep. 2017, Art. no. 1603948, doi: 10.1002/smll.201603948.

[9] F. Yuan *et al.*, "Real-time observation of the electrode-size-dependent evolution dynamics of the conducting filaments in a SiO_2 layer," *ACS Nano*, vol. 11, no. 4, pp. 4097–4104, Apr. 2017, doi: 10.1021/acsnano.7b00783.

[10] Y. Zhao *et al.*, "Mass transport mechanism of Cu species at the metal/dielectric interfaces with a graphene barrier," *ACS Nano*, vol. 8, no. 12, pp. 12601–12611, Dec. 2014, doi: 10.1021/nn5054987.

[11] J. Lee, C. Du, K. Sun, E. Kioupakis, and W. D. Lu, "Tuning ionic transport in memristive devices by graphene with engineered nanopores," *ACS Nano*, vol. 10, no. 3, pp. 3571–3579, Mar. 2016, doi: 10.1021/acsnano.5b07943.

[12] S. Garaj, W. Hubbard, A. Reina, J. Kong, D. Branton, and J. A. Golovchenko, "Graphene as a subnanometre trans-electrode membrane," *Nature*, vol. 467, no. 7312, pp. 190–193, Sep. 2010, doi: 10.1038/nature09379.

[13] S. Ginnaram, J. T. Qiu, and S. Maikap, "Controlling cu migration on resistive switching, artificial synapse, and glucose/saliva detection by using an optimized AlO_X interfacial layer in a-CO_X-based conductive bridge random access memory," *ACS Omega*, vol. 5, no. 12, pp. 7032–7043, Mar. 2020, doi: 10.1021/acsomega.0c00795.

[14] 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, Apr. 2010, doi: 10.1021/nl904092h.

[15] Y. Shi *et al.*, "Neuroinspired unsupervised learning and pruning with subquantum CBRAM arrays," *Nature Commun.*, vol. 9, no. 1, p. 5312, Dec. 2018, doi: 10.1038/s41467-018-07682-0.

[16] S. Park *et al.*, "Electronic system with memristive synapses for pattern recognition," *Sci. Rep.*, vol. 5, no. 1, p. 10123, Sep. 2015, doi: 10.1038/srep10123.

[17] J. Woo *et al.*, "Improved synaptic behavior under identical pulses using AlO_X/HfO_2 bilayer RRAM array for neuromorphic systems," *IEEE Electron Device Lett.*, vol. 37, no. 8, pp. 994–997, Jun. 2016, doi: 10.1109/LED.2016.2582859s.

[18] B. C. Jang *et al.*, "Polymer analog memristive synapse with atomic-scale conductive filament for flexible neuromorphic computing system," *Nano Lett.*, vol. 19, no. 2, pp. 839–849, Feb. 2019, doi: 10.1021/acs.nanolett.8b04023.

[19] H. Yeon *et al.*, "Alloying conducting channels for reliable neuromorphic computing," *Nature Nanotechnol.*, vol. 15, no. 7, pp. 574–579, Jul. 2020, doi: 10.1038/s41565-020-0694-5.