공학석사 학위논문

뉴로모픽 컴퓨팅을 위한 아날로그 시냅스 소자로써의 금속 산화물 기반의 전기화학 메모리 구조 최적화

Structure optimization of metal oxide based electrochemical memory as an analog synaptic device for neuromorphic computing

2022 년 2 월

서울대학교 대학원

기계공학부

류 다 길

뉴로모픽 컴퓨팅을 위한 아날로그 시냅스 소자로써의 금속 산화물 기반의 전기화학 메모리 구조 최적화

Structure optimization of metal oxide based electrochemical memory as an analog synaptic device for neuromorphic computing

지도 교수 이 윤 석

이 논문을 공학석사 학위논문으로 제출함

2021 년 10 월

서울대학교 대학원

기계공학부

류 다 길

류 다 길의 공학석사 학위논문을 인준함

2021 년 12 월

위 원 장 최 만 수 (인)

부위원장 이 윤 석 (인)

위 원 신 용 대 (인)

Abstract

Despite the success of artificial intelligence(AI) technology, training deep neural networks(DNNs) through computation intensive algorithms is time consuming and high energy consuming. To achieve massive parallel vector-matrix-multiplication(VMM) calculation and energy efficient DNN learning, cross point array with nonvolatile memory(NVM) analog synaptic device has been studied. However, these devices have non-ideal synaptic characteristics due to the limitation of the operating mechanism or material properties. CMOS compatible metal oxide based analog synaptic device with ideal analog synaptic characteristics has been studied and operating mechanism has been reported. In this paper, through dimension and structure change based on reported operating mechanism of the metal oxide based analog synaptic device, nonvolatile and artificial synaptic characteristics are investigated and optimized. Slow ion diffusion through depth changes the average channel conductance, so thinner channel is required for symmetric programming. Programming is occurred under the gated region and field-driven migration acts speed limiting factor, so vertical structure is required for small programming energy. The effect of channel thickness and structure on performance was studied.

Keyword : Nonvolatile memory, Neuromorphic computing, Electrochemical memory, Ionic conduction, Transition metal oxide, Structure optimization
Student Number : 2019-28331

Table of Contents

Chapter 1. Introduction

1.1. Study Background

Recently, artificial intelligence(AI) technology innovation is taking place through an increase in the amount of data that can be collected online and a large-scale demonstration on the cloud, and accordingly, interest in AI is rapidly increasing.[1] There are many AI learning algorithms, and in particular, learning deep neural networks(DNNs) through back-propagation algorithms has achieved great success in fields such as image recognition and natural language processing.[2]

However, DNNs consist of fully connected neurons of multi layers(Figure 1). The back-propagation algorithm for training DNNs consists of numerous vector matrix multiplication(VMM) operations, and the von-Neumann structure such as the existing GPU-based hardware accelerator is a structure in which the memory unit and the processing unit are separated, so a bottleneck occurs during operation. Learning is time consuming and requires a lot of computing resources and computing power(Figure 2 a, b).[3]

Therefore, various methods for accelerating the computation speed have been studied. Structure that mimics the brain has been studied that cross-point array structure where synaptic devices are at the cross point(Figure 2 c). Among them, the use of nonvolatile memory(NVM) devices has been studied. Because NVM has a multi-level conductance state, creating a cross-point array with NVM and learn the conductance value of the NVM corresponding to the weight value of the DNN, processing and store are performed in one device, and parallel weight update is possible.[4-10] As a result, time and power consumption are reduced. As NVM candidates, various devices such as resistive memory(RRAM)[11-14], phase change memory(PCM)[15-17], conductive bridge memory(CBRAM)[11, 18], ferroelectric-based memory[19-21], and field-effect transistor(FET)-based memory[22,

[23] were studied. These devices, however, have degrade DNN learning accuracy because of non-ideal synaptic device characteristics such as non-linear and asymmetry conductance response states, small step number of conductance levels, limited endurance and device to device variation of the NVM device.

In order to overcome these limitations, electrochemical memory(ECRAM) has been proposed as a synaptic device candidate in a cross-point array. ECRAM is a redox transistor device with a three-terminal structure, and 'read' and 'write' operations are decoupled. The conductance of the channel is read by applying voltage to source and drain and measuring current and changed through electrochemical reaction by applying voltage to the gate. Decoupled nature enables low energy switching and better endurance while maintaining non-volatility. Furthermore, the electrochemically driven redox reaction can be precisely and reversibly controlled through the amount of charge through the gate, resulting symmetric switching conductance state.

Previously studied ECRAM devices especially cation(i.e. proton and lithium ion) based have shown CMOS non-compatible and unstable operations.[24,25] Although lithium ion based ECRAM devices are widely studied owing to well-known material and operating mechanism of battery, CMOS non-compatible, the formation of lithium dendrites and open circuit voltage are critical challenges to make neuromorphic system on a chip. In contrast, metal oxide based ECRAMs are CMOS compatible and can be mass produced in the semiconductor production line.[26-29]

In previous study, the mechanism of metal oxide based ECRAM has been investigated.[29] The ionic current under gated region causes oxygen vacancy to move through the electrolyte layer, changing the stoichiometry of the channel material. The programming dynamics are modeled that electric field driven ion migration through electrolyte and free ion diffusion within the channel. In addition, ion diffusion within the channel is not lateral, but in the depth direction. Field-driven migration works as a speed limiting factor in programming.

Here, structure optimization was carried out based on the working principle found in previous studies. Slow ion diffusion through depth changes the average channel conductance, so thinner channel is required for symmetric programming. Programming is occurred under the gated region and field-driven migration acts speed limiting factor, so vertical structure is required for small programming energy. Vertical structure has advantages in scaling.

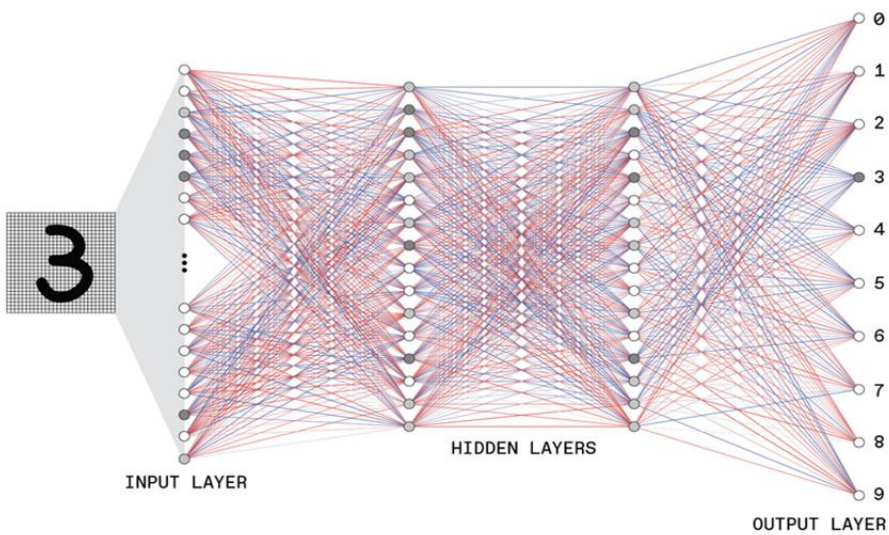

Figure 1. Deep neural network with MNIST dataset inputs, two hidden layers, and 10 outputs.[30]

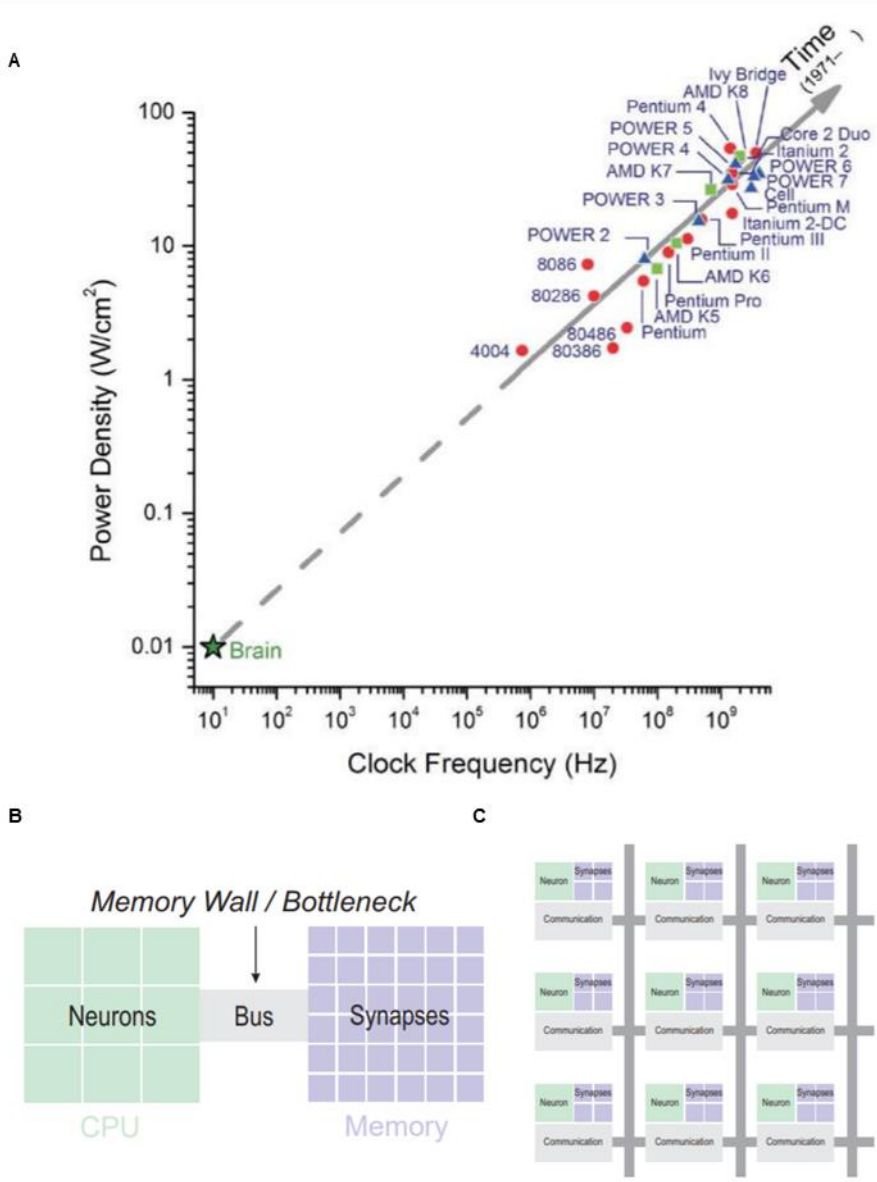

Figure 2. (a)Comparison of brain and von Neumann architecture conventional memories, (b)Schematic of conventional von Neumann architecture that structure that memory unit and processing unit are separated, (c)Schematic of in-memory neuromorphic computing that cross-point array structure where synaptic devices are at the cross point.[31]

1.2. DNN learning algorithm

The artificial neural network consists of an input layer that receives signals, a hidden layer through which data from the input layer flows, and an output layer that outputs results. The hidden layer may be composed of one or multiple layers, and the one composed of multiple hidden layers is called a deep neural network.

In artificial neural networks, different weigh values are used to pass data to the next layer with different weights. The goal of learning neural network is to reduce error values by optimizing weight values. Therefore, the neural network is updated by modifying the weight values in the direction of reducing the error values. The process of learning DNNs is as follows: 1) feed forward update and 2) back propagation.

In feed forward update, it passes through each layer from the input layer, repeats VMMs that multiply and sum the weights, and the process of passing through the activation function, as follows:

$$y_n^l = \sum_m w_{nm}^l x_m^{l-1} + b_n^l \tag{1}$$

$$x_n^l = f(y_n^1) \tag{2}$$

where, y_n^l, w_{nm}^l, x_m^{l-1}, b_n^l, and f are output neuron(before activated), weight, input neuron, bias value of $(l-1)^{th}$ layer and l^{th} layer of DNN, and activation function, respectively. The activation function f can be sigmoid, hyperbolic tangent, softmax, ReLU, etc.

Error values are calculated at the end of feed forward update. The errors between the output value and target value are calculated by the error function, as follows:

$$E_{total}(w,b) = \sum_N \frac{1}{2}(target - output)^2 \tag{3}$$

where w, b, N, target and output are the vector of weights, the vector of bias values, total number of layers, the vector of desired outputs and the vector of outputs of the network, respectively.

Next, in the back propagation step, based on the calculated error function, the update is performed in the direction of reducing the errors. A simple algorithm that gradient descent has been performed to find out weight and bias values minimizing the errors. It finds the slope of the function, moving it continuously toward the lower absolute value of the slope, and repeating it until it reaches the minimum value. By using gradient descent algorithm to back propagation step, the error function reaches to the minimum value and weights and biases are updated as followed:

$$w_{nm} \leftarrow w_{nm} - \eta \frac{\partial E}{\partial w_{nm}} \tag{4}$$

$$b_n \leftarrow b_n - \eta \frac{\partial E}{\partial b_n} \tag{5}$$

Where η is learning rate. Since the updates are performed using the chain rule, it is sequentially reversed when updating. When updating, the weight between the outermost output and the hidden layer is changed in order, so it is called back propagation.

The weights and bias of the same layer are updated as a single vector−matrix multiplication. For convenience of calculation, let consider bias as the last term of weights. The error at the output layer, δ_n^N, and the error at the l^{th} layer, δ_n^l, are as followed:

$$\delta_n^N = \frac{\partial E}{\partial x_n} f'(y_n) \tag{6}$$

$$\delta_n^l = \sum_m w_{nm}^{l+1} \delta_m^{l+1} f'(y_n^l) \tag{7}$$

The error for any layer can be computed by combining equation (6) and (7). From these error values, the gradient of error function can be obtained as:

$$\frac{\partial E}{\partial w_{nm}^l} = x_m^{l-1} \delta_n^l \tag{8}$$

Then, the weights are updated with learning rate η as follows:

$$w_{nm}^{l} \leftarrow w_{nm}^{l} - \eta x_{m}^{l-1} \delta_{n}^{l} \tag{9}$$

The update of weights can be expressed as vector−vector outer product operation:

$$w^{l} \leftarrow w^{l} - \eta x^{l-1} \times \delta^{l} \tag{10}$$

In back propagation step, in each layer, the total number of multiplications are $m \times n$, with neglecting derivative calculation process. The m and n is the number of neurons in $(l-1)^{th}$ layer and the number of neurons in l^{th} layer. In DNNs, it is computationally intensive task and consumes lots of computing power and time consuming.

When the use of a crossbar array with synaptic device at the cross point can dramatically reduce computation time and power consumption. The weights are stored as the conductance of the device, the VMM is accelerated by Kirchhoff's law, and the OPU is performed according to a stochastic parallel update scheme (Figure 3 a−c). Since the operation is done in memory, it is energy−efficient and parallel operation is possible with one order of time complexity.

$$\vec{y} = W\vec{x} \tag{11}$$

$$\vec{I} = W\vec{V} \tag{12}$$

There are several important synaptic device characteristics that number of states, linearity, symmetry, weight update variation and programming power.

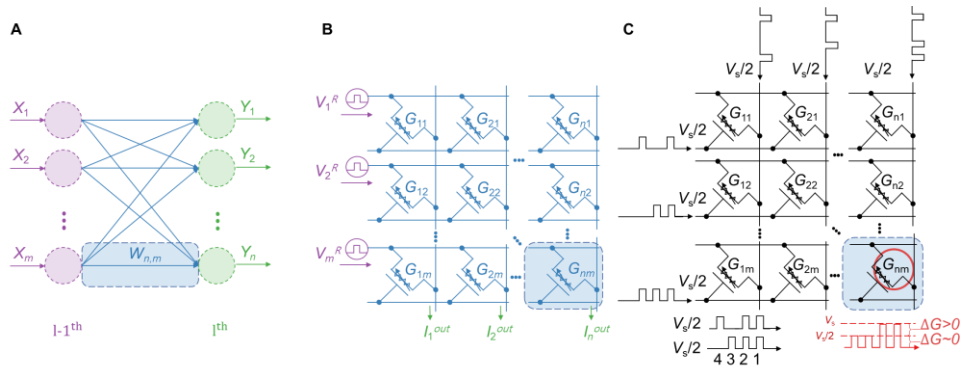

Figure 3. Weight matrix in neural network corresponds to the conductance matrix of cross point array with synaptic devices (a)Part of the DNN where $(l-1)^{th}$ layer and l^{th} layer, each consists of m neurons and n neurons, (b)VMM can be conducted by applying voltage pulses to the rows and reading the current outputs in the columns, (c)Half voltage selection scheme for massive parallel programming[29]

1.3. Purpose of study

Here, we proceed with structural optimization to improve properties based on the investigated mechanism of operation of the metal oxide-based ECRAM. Specifically, the asymmetry property occurs because free ion diffusion within the channel occurs slowly in the depth direction and changes the channel conductance. To improve this, we propose a device with a thin channel. It has been demonstrated by simulation that the symmetry property improves as the channel thickness decreases. Also, the programming voltage is very large with potentiation 8V and depression −6V, and since the source and drain are on the same plane, there is a problem that the cell size becomes large when it is made into an array. To solve this, a device with a new structure is required. It has been experimentally proven to have advantages in terms of scaling and energy efficiency by fabricating device with a new structure, vertical structure.

Chapter 2. Experiment

2.1. Deposition of metal oxide

Metal oxide based ECRAM consists of three metal oxide layers that channel, electrolyte, and reservoir, respectively. As a channel material, transition metal oxide is mainly used whose conductance changes according to the cation oxidation number. Tungsten oxide was selected as the channel material as it is dramatic conductance changing n-type semiconductor.

Since the device operates at room temperature, material with good room temperature ionic conductivity is used as the electrolyte. HfO_2 was selected as electrolyte material whose intrinsic defects determine programming characteristics.

2.1.1. Deposition of channel

The WOx films were deposited by reactive sputtering with varying Ar to O_2 ratio using 3" pure metal tungsten(99.99%) target. Reactive sputtering was conducted under fixed 5 x 10^{-6} Torr base pressure, 20 sccm Ar flow rate, 3.3 sccm O_2 flow rate, 10mTorr working pressure and 150 W RF power. Pre-sputtering before deposition was performed for deposition under constant conditions by removing target surface contamination including oxide.

2.1.2. Deposition of electrolyte

The HfO_2 films were deposited by ALD with TEMAHf Hf-precursor and H_2O oxidant. The use of H_2O reactant, lowing processing chamber temperature to 150oC, is effective in improving oxygen ion conductivity due to the oxygen defect stoichiometry compared to ozone reactant.

2.2. Electrical measurement of thin films

2.2.1. Thickness of thin films

Thickness of thin films were measured by SEM or Bruker Dektak XT-A surface profiler. Thin films were deposited on Si substrate.

2.2.2. Conductivity measurement

The measurement thin film samples were deposited on SiO_2 surface. The sheet resistance was measured at 4-point probe. Conductivity can be calculated from the measured sheet resistance as follows:

$$\sigma = \frac{1}{\rho} = \frac{1}{R_s \cdot t} \tag{13}$$

2.2.3. Ionic conductivity measurement

The ionic conductivity was measured by Electrochemical impedance spectroscopy(EIS) measurement. EIS measurement sample was fabricated with vertically stacked Au / electrolyte / Au on SiO_2 substrate. Trough-plane EIS measurement was conducted on the hot plate in the dry room. The frequency range of applied AC voltage was from 7 MHz to 10 mHz and the temperature range was RT, 100, 150, 200, 250 °C. The equivalent circuit model to fit the measured data was shown in Figure 4. The fitting process is conducted by Z Fit of Bio Logic EC LAB.

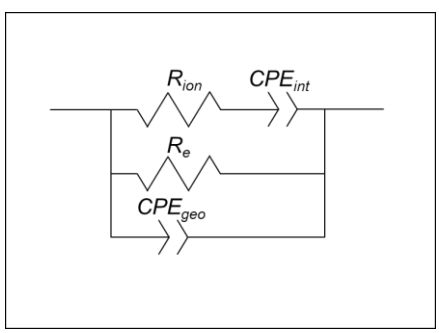

Figure 4. The equivalent circuit model with ionic and electronic resistance.[32]

2.2.4. Surface uniformity

The measurement thin film samples were deposited on Si surface. Atomic force microscope(AFM) was used to measure the surface uniformity of the thin film sample.

2.3. Hole fabrication

In the vertical structure device, the process of hole etch and deposition on the hole sidewall was added. Establishing an optimized hole fabrication process was needed because processes with poor step coverage such as sputtering and evaporation were used when fabrication devices.

2.3.1. Hole etch

SF6 was selected as the etchant gas to etch the metal line and the inter layer dielectric at the same time. Hole etch process was conducted under 750 W ICP power, 100 W RF power, 20 sccm SF_6 flow rate, 20 sccm N_2 flow rate and 10 mTorr working pressure.

2.3.2. Hole sidewall deposition

Due to the linearity deposition nature of sputtering and evaporation, deposition is less on the sidewall than on the surface. It is necessary to check the thickness to establish the vertical structure device fabrication process

2.4. Device fabrication

Devices were fabricated by conventional CMOS-compatible processes including photolithography, sputtering, evaporation, ALD, PECVD, dry etch and wet etch. Channel, all of metal lines and reservoir layers were patterned by conventional photolithography and lift off process. Especially, lift off process performed with

LOR/AZ5214 photoresist bilayer to adjusting photoresist shape to form an undercut.

2.4.1. Thin channel device

Figure 5 shows the whole fabrication process flow of thin channel device. First, WO_x channel was deposited by reactive sputtering on the Si substrate with 90 nm thickness thermal grown silicon oxide layer. The thickness of channel was splited by 5, 10, 15nm and the dimension was 100 x 100 um2. Followed by the deposition of metal line was sputtered tungsten. ALD was used to deposit 10.5 nm thickness HfO_2 layer. Evaporation was used to deposit 60 nm thickness MoO_3 layer, 20 nm thickness Ti layer and 30 nm thickness Au layer. Gate geometry was non-overlapped with source and drain. The dimension of gate was 90 um x 100, 50 and 25 um. To contact with source and drain, HfO_2 layer on contact pad area was wet etched.

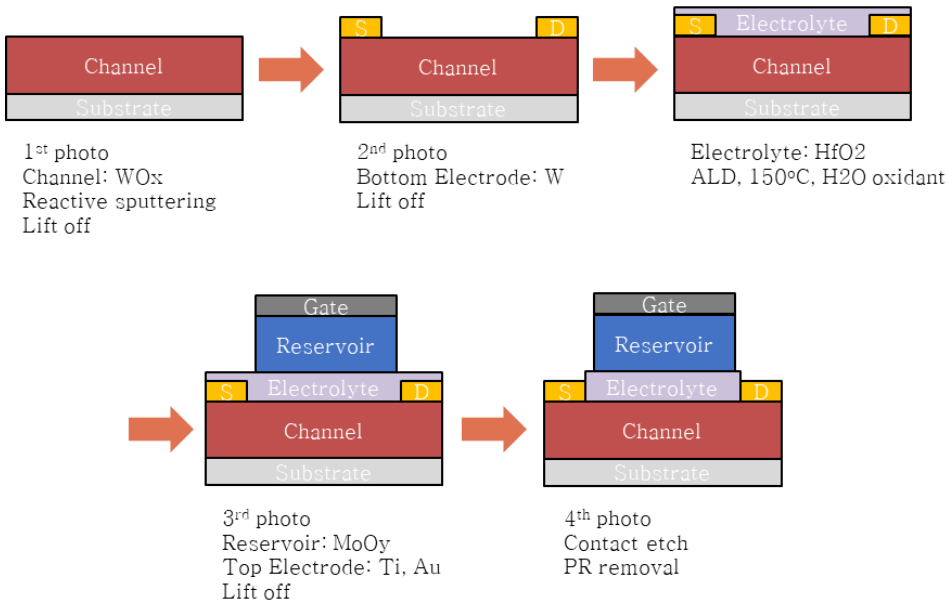

Figure 5. Fabrication process of thin film device

2.4.2. Vertical structure device

Figure 6 shows the whole fabrication process flow of vertical structure device. First, 1 um SiO_2 insulating layer was deposited by

PECVD on the Si substrate. W metal line as source deposition was sputtered followed by the deposition of 100 nm silicon oxide inter layer dielectric layer was deposited by PECVD. W metal line as drain deposition was sputtered followed by the deposition of 100 nm silicon oxide inter layer dielectric layer was deposited by PECVD again. Hole etch was performed to define area for post-processing. After hole etch, WO_x channel was deposited by reactive sputtering on hole sidewall. HfO_2 was deposited by ALD. Evaporation was used to deposit MoO_3 and Au. The adhesion layer Ti between MoO_3 and Au was deposited by sputtering. HfO_2 layer on contact pad was wet etched to contact with source and drain.

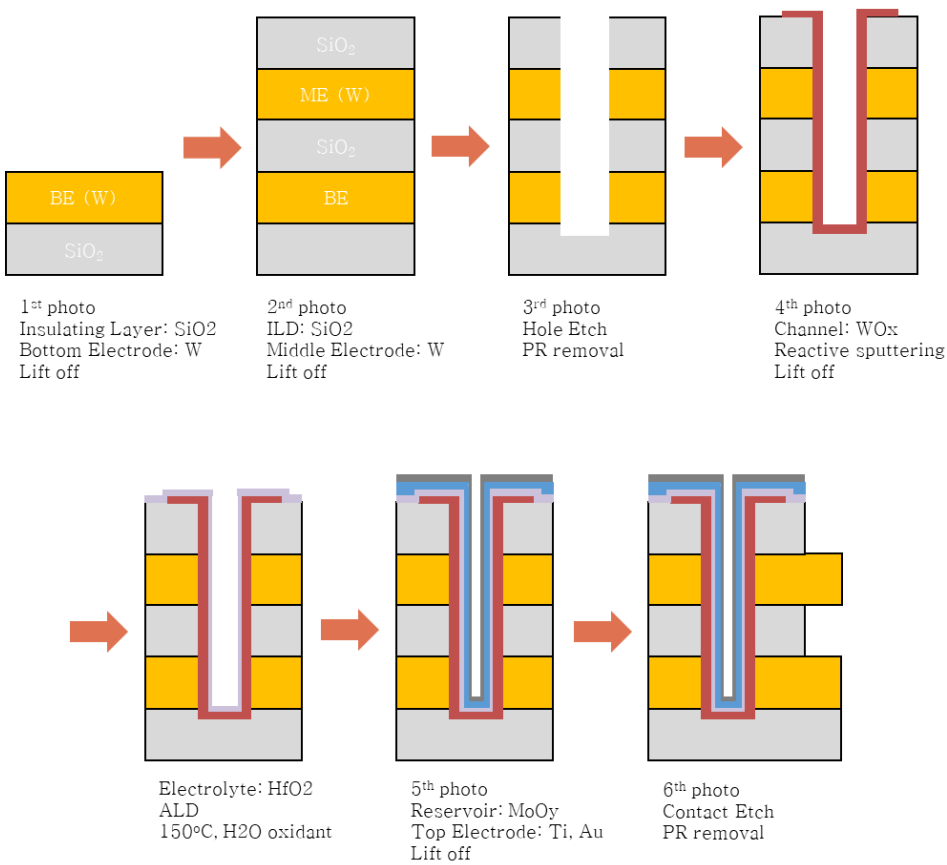

Figure 6. Fabrication process of vertical structure device

2.5. Electrical measurement of device

All electrical measurements were conducted by Keithley 4200A-SCS parameter analyzer or Keithley 2450 SourceMeter.

Chapter 3. Results and Discussion

3.1. Material properties of metal oxide

3.1.1. Tungsten oxide films by reactive sputtering

Since the tungsten oxide films of 15nm or less were deposited by sputtering, AFM was performed to check surface uniformity. The root mean square values of surface roughness of 5 nm, 10nm, and 15 nm thin films are 140.994, 179.054, and 137.716 pm each.

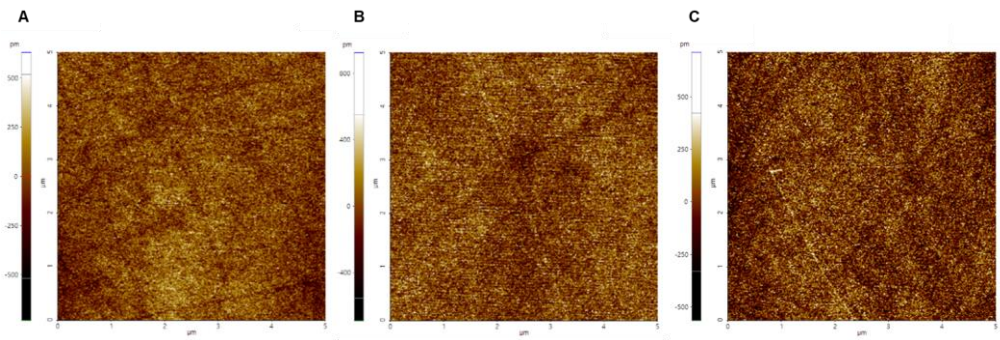

Figure 7. Surface roughness of (a)5 nm thin film, (b)10 nm thin film, and (c)15 nm thin film

3.1.2. Hafnia film by ALD

HfO2 deposited with H_2O oxidant has an activation energy of 0.46 eV and has an 1.63 x 10^{-13} Scm^{-1} room temperature ionic conductivity value when obtained by extrapolating EIS measurement(Figure 8). Compared to the HfO_2 with ozone oxidant(activation energy of 0.66 eV and has an 1.64x10^{-14} Scm^{-1} room temperature ionic conductivity), the ionic conductivity is improved 10 times.

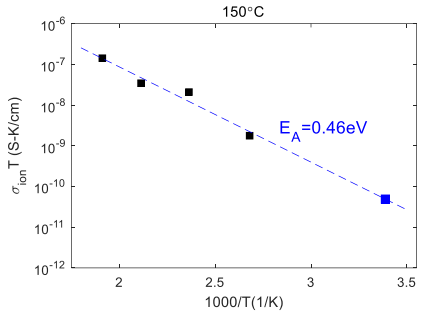

Figure 8. EIS measurements of the HfO_2 film. From four temperature points, the activation energy and ionic conductivity at room temperature can be estimated.

3.2. Hole structure

When 50, 100, 200 nm films were deposited by sputtering on horizontal surface, less than 20 nm, 25 nm, 85 nm were deposited on the sidewall surface each(Figure 9).

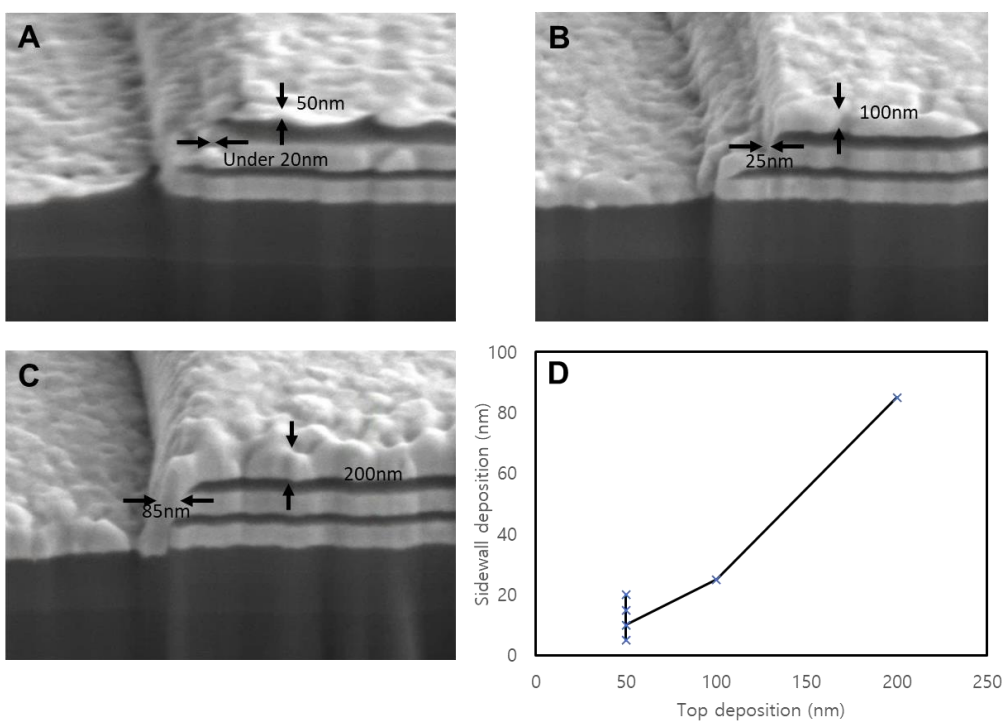

Figure 9. The hole sidewall deposition with horizontal deposition

thickness of (a)50 nm, (b)100 nm and (c)200nm, (d)The relation between sidewall deposition with horizontal deposition.

3.3. Thin channel device

3.3.1. Diffusion simulation of channel thickness

In metal oxide based ECRAM, oxygen ions that have passed through the electrolyte layer move through free ion diffusion within the channel. At this time, ion diffusion occurs in the channel region below the gate region, and an oxygen ion concentration gradient occurs along the depth direction of the channel.

These are the results of simulations considering the diffusion in the channel and stoichiometry distribution according to the channel depth over time when a constant voltage bias is applied to the gate. At this time, in the case of a 15 nm channel device, it was confirmed that when the conductance was calculated by integrating the conductivity with the thickness, the hysteresis in the G-Q graph was shown, that is, the same as the actual measured value. The reason for hysteresis can be inferred by looking at the simulation results. From the very beginning of potentiation, a saturated conductive layer exists at the interface with the electrolyte, and it propagates deeper as time goes on. In the depression, the interface becomes sharply insulating, but the conductance decreases slowly because the depth of the channel is still conductive. This is a source of hysteresis and is expected to cause asymmetry in programming.[29]

Intra-channel diffusion simulations and G-Q graphs were calculated for devices with 10 nm and 5 nm channels, which are thinner than the conventional 15 nm(Figure 10). As a result, as the channel thickness decreased, the hysteresis in the G-Q graph decreased(Figure 11). The hysteresis in the G-Q graph means that even with the same ion charge, different conductance values can be obtained due to the vertical distribution of ions, which greatly affects the asymmetry. Therefore, it can be inferred from the

simulation results that the asymmetry decreases with a thinner channel.

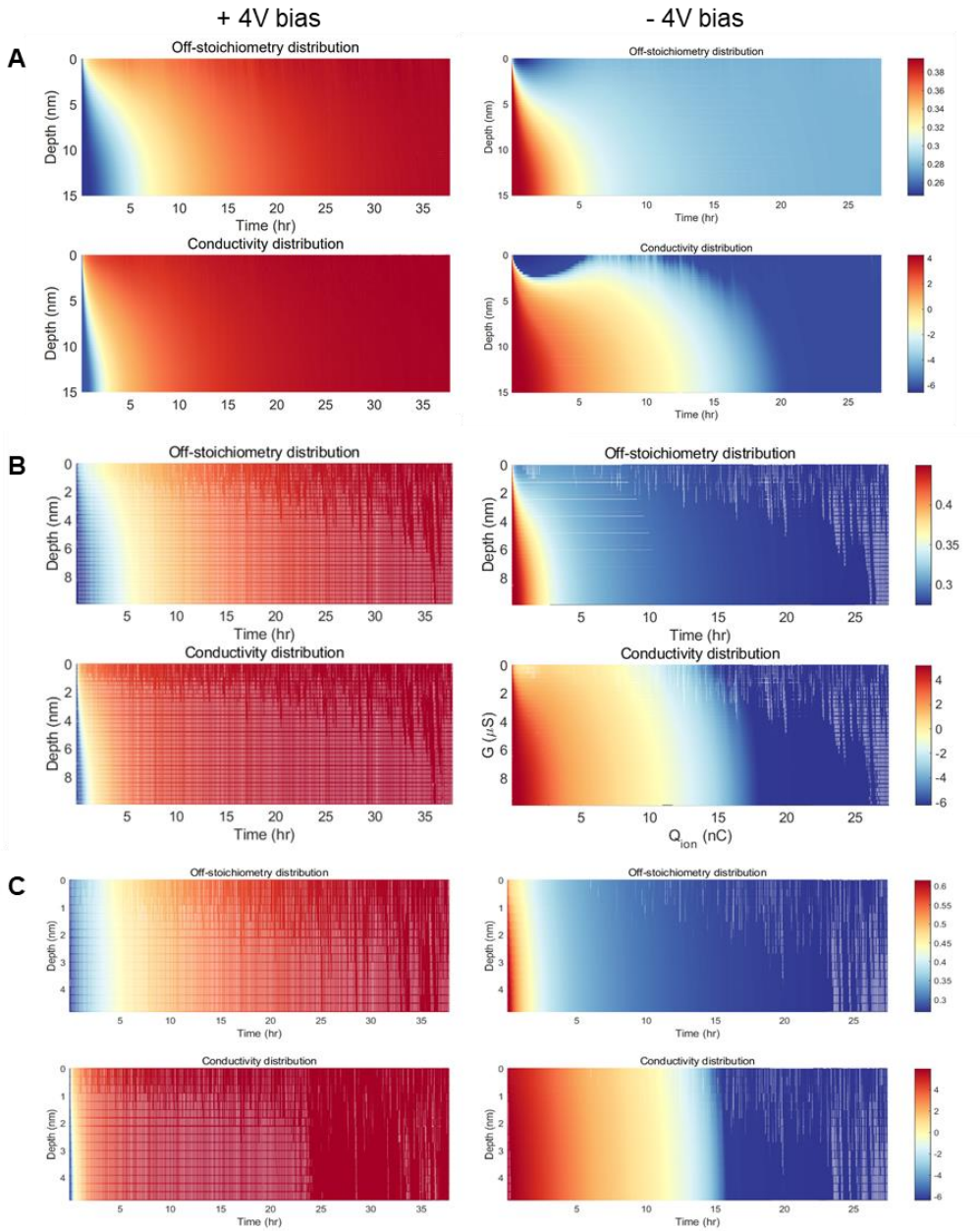

Figure 10. Diffusion simulation of oxygen ion diffusion along the channel depth of (a)15 nm channel device, (b)10 nm channel device, (c)5 nm channel device with diffusivity, $D = 5 \times 10^{-21}$ cm^2/s.

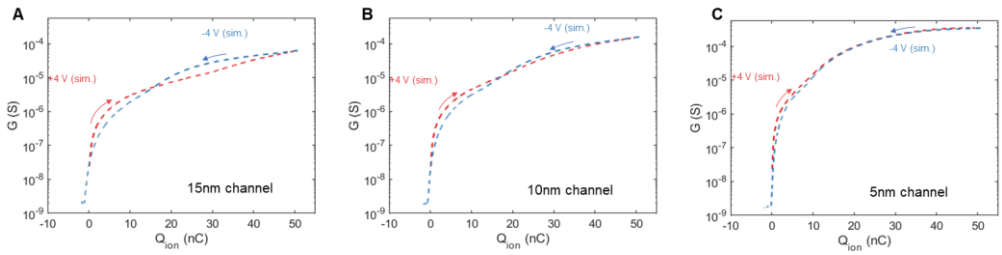

Figure 11. Channel conductance and ionic charge relation from the diffusion simulation (a)15 nm channel device, (b)10 nm channel device, (c)5 nm channel device

3.4. Vertical structure device

3.4.1. Programming characteristics

Since the vertical structure device has separate source and drain layers, the cell size can be reduced from the existing 12F2 to 4F2. In addition, channel volume scaling can be easily performed because the channel length is determined by the ILD thickness, not the photolithography resolution. A decrease in channel volume means a decrease in Q_{ion}, and thus a decrease in programming energy. Also, since it has the same structure as VNAND, it is possible to stack one layer and then stack it again. Cell density can be easily increased through stacking. Also, since the gate is in the form of a shell located in the center, there is an advantage in terms of programming energy. Unlike the planar type, the electric field concentration occurs in the central part. This has the effect of reducing the programming voltage. As a result, programming energy is reduced. Figure 12 shows low voltage programmable. It had achieved +1 V, −1 V programming voltage. Also it had 500 potentiation, depression conductance states and achieved endurance 105 times.

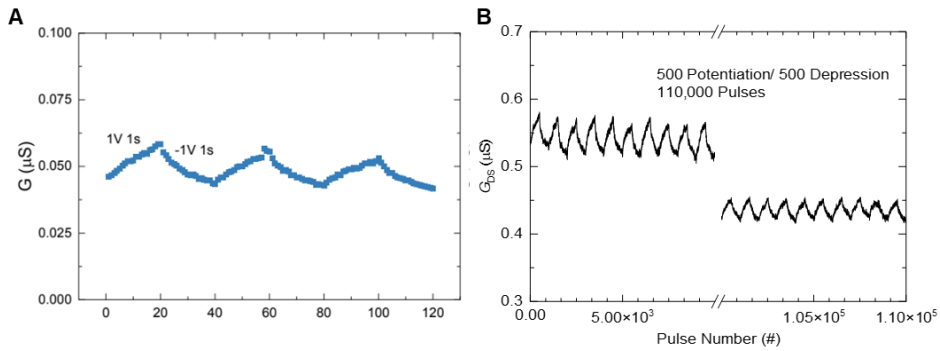

Figure 12. Programming characteristic of vertical structure device. (a)Low voltage programming, (b)Endurance of 10^5 pulses.

Chapter 4. Conclusion

In this paper, two structural approaches have been studied for performance improvement through device structure optimization. It is a method to take advantage in terms of symmetry by reducing the ion diffusion effect in the depth direction in the channel by making the channel thin, and a method to take advantage in terms of scaling and energy efficiency by making the channel vertical. To fabricate a thin channel device, the surface uniformity of the thin film was checked, and to fabricate a vertical device, the hole structure and sidewall deposition were checked. As a result of device channel diffusion simulation, it was confirmed that the hysteresis decreased as the channel became thinner, and it was confirmed that it has an advantage in terms of scaling as a result of manufacturing and measuring vertical structure devices and in terms of energy efficiency through low voltage programming.

Bibliography

(1) Haefner, N., Wincent, J., Parida, V., & Gassmann, O. (2021). Artificial intelligence and innovation management: A review, framework, and research agenda. Technological Forecasting and Social Change, 162, 120392.

(2) LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

(3) Xu, X., Ding, Y., Hu, S. X., Niemier, M., Cong, J., Hu, Y., & Shi, Y. (2018). Scaling for edge inference of deep neural networks. Nature Electronics, 1(4), 216-222.

(4) Dong, X., Xu, C., Xie, Y., & Jouppi, N. P. (2012). Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 31(7), 994-1007.

(5) Zhou, Y., Han, S. T., Chen, X., Wang, F., Tang, Y. B., & Roy, V. A. L. (2014). An upconverted photonic nonvolatile memory. Nature communications, 5(1), 1-8.

(6) Narayanan, P., Fumarola, A., Sanches, L. L., Hosokawa, K., Lewis, S. C., Shelby, R. M., & Burr, G. W. (2017). Toward on-chip acceleration of the backpropagation algorithm using nonvolatile memory. IBM Journal of Research and Development, 61(4/5), 11-1.

(7) Meena, J. S., Sze, S. M., Chand, U., & Tseng, T. Y. (2014). Overview of emerging nonvolatile memory technologies. Nanoscale research letters, 9(1), 1-33.

(8) Derhacobian, N., Hollmer, S. C., Gilbert, N., & Kozicki, M. N. (2010). Power and energy perspectives of nonvolatile memory technologies. Proceedings of the IEEE, 98(2), 283-298.

(9) Xiao, T. P., Bennett, C. H., Feinberg, B., Agarwal, S., & Marinella, M. J. (2020). Analog architectures for neural network acceleration based on non-volatile memory. Applied Physics Reviews, 7(3), 031301.

(10) Ambrogio, S., Narayanan, P., Tsai, H., Shelby, R. M., Boybat, I., Di Nolfo, C., ... & Burr, G. W. (2018). Equivalent-accuracy accelerated neural-network training using analogue memory.

Nature, 558(7708), 60-67.
(11) Xia, L., Gu, P., Li, B., Tang, T., Yin, X., Huangfu, W., ... & Yang, H. (2016). Technological exploration of RRAM crossbar array for matrix-vector multiplication. Journal of Computer Science and Technology, 31(1), 3-19.
(12) Deng, Y., Huang, P., Chen, B., Yang, X., Gao, B., Wang, J., ... & Liu, X. (2012). RRAM crossbar array with cell selection device: A device and circuit interaction study. IEEE transactions on Electron Devices, 60(2), 719-726.
(13) Park, S., Kim, H., Choo, M., Noh, J., Sheri, A., Jung, S., ... & Hwang, H. (2012, December). RRAM-based synapse for neuromorphic system with pattern recognition function. In 2012 international electron devices meeting (pp. 10-2). IEEE.
(14) Hong, X., Loy, D. J., Dananjaya, P. A., Tan, F., Ng, C., & Lew, W. (2018). Oxide-based RRAM materials for neuromorphic computing. Journal of materials science, 53(12), 8720-8746.
(15) Suri, M., Bichler, O., Querlioz, D., Cueto, O., Perniola, L., Sousa, V., ... & DeSalvo, B. (2011, December). Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction. In 2011 International Electron Devices Meeting (pp. 4-4). IEEE.
(16) Kim, S., Ishii, M., Lewis, S., Perri, T., BrightSky, M., Kim, W., ... & Lam, C. (2015, December). NVM neuromorphic core with 64k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous in-situ learning. In 2015 IEEE international electron devices meeting (IEDM) (pp. 17-1). IEEE.
(17) Boybat, I., Le Gallo, M., Nandakumar, S. R., Moraitis, T., Parnell, T., Tuma, T., ... & Eleftheriou, E. (2018). Neuromorphic computing with multi-memristive synapses. Nature communications, 9(1), 1-12.
(18) Cha, J. H., Yang, S. Y., Oh, J., Choi, S., Park, S., Jang, B. C., ... & Choi, S. Y. (2020). Conductive-bridging random-access memories for emerging neuromorphic computing. Nanoscale, 12(27), 14339-14368.

(19) Oh, S., Kim, T., Kwak, M., Song, J., Woo, J., Jeon, S., ... & Hwang, H. (2017). HfZrO x-based ferroelectric synapse device with 32 levels of conductance states for neuromorphic applications. IEEE Electron Device Letters, 38(6), 732-735.
(20) Ryu, H., Wu, H., Rao, F., & Zhu, W. (2019). Ferroelectric tunneling junctions based on aluminum oxide/zirconium-doped hafnium oxide for neuromorphic computing. Scientific reports, 9(1), 1-8.
(21) Seo, M., Kang, M. H., Jeon, S. B., Bae, H., Hur, J., Jang, B. C., ... & Choi, Y. K. (2018). First demonstration of a logic-process compatible junctionless ferroelectric FinFET synapse for neuromorphic applications. IEEE Electron Device Letters, 39(9), 1445-1448.
(22) Sanchez Esqueda, I., Yan, X., Rutherglen, C., Kane, A., Cain, T., Marsh, P., ... & Zhou, C. (2018). Aligned carbon nanotube synaptic transistors for large-scale neuromorphic computing. ACS nano, 12(7), 7352-7361.
(23) Kim, S., Gokmen, T., Lee, H. M., & Haensch, W. E. (2017, August). Analog CMOS-based resistive processing unit for deep neural network training. In 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS) (pp. 422-425). IEEE.
(24) Fuller, Elliot J., et al. "Li-ion synaptic transistor for low power analog computing." Advanced Materials 29.4 (2017): 1604310.
(25) van de Burgt, Yoeri, et al. "A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing." Nature materials 16.4 (2017): 414-418.
(26) Kim, Seyoung, et al. "Metal-oxide based, CMOS-compatible ECRAM for deep learning accelerator." 2019 IEEE International Electron Devices Meeting (IEDM). IEEE, 2019.
(27) Lee, C., Rajput, K. G., Choi, W., Kwak, M., Nikam, R. D., Kim, S., & Hwang, H. (2020). Pr 0.7 Ca 0.3 MnO 3-Based Three-Terminal Synapse for Neuromorphic Computing. IEEE Electron Device Letters, 41(10), 1500-1503.

(28) Lee, Hyerin, et al. "Nonvolatile Memory and Artificial Synaptic Characteristics in Thin-Film Transistors with Atomic Layer Deposited HfOx Gate Insulator and ZnO Channel Layer." Advanced Electronic Materials 6.9 (2020): 2000412.
(29) Jeong, Yangho, et al. "Elucidating Ionic Programming Dynamics of Metal-Oxide Electrochemical Memory for Neuromorphic Computing." Advanced Electronic Materials (2021): 2100185.
(30) Moore, Samuel K., David Schneider, and Eliza Strickland. "How Deep Learning Works: Inside the Neural Networks that Power Today's AI." IEEE Spectrum 58.10 (2021): 32-33.
(31) Merolla, Paul A., et al. "A million spiking-neuron integrated circuit with a scalable communication network and interface." Science 345.6197 (2014): 668-673.
(32) Wang, Shanyu, et al. "Separating electronic and ionic conductivity in mix-conducting layered lithium transition-metal oxides." Journal of Power Sources 393 (2018): 75-82.

초록

인공 지능(AI) 기술의 성공에도 불구하고 연산 집약적 알고리즘을 통한 심층 신경망(DNN) 교육은 시간과 에너지를 많이 소모합니다. 대규모 병렬 벡터 행렬 곱셈(VMM) 계산 및 에너지 효율적인 DNN 학습을 달성하기 위해 비휘발성 메모리(NVM) 아날로그 시냅스 장치를 사용한 교차점 배열이 연구되었습니다. 그러나 이러한 장치는 작동 메커니즘이나 재료 특성의 제한으로 인해 이상적이지 않은 시냅스 특성을 가지고 있습니다. 이상적인 아날로그 시냅스 특성을 가진 CMOS 호환 금속 산화물 기반 아날로그 시냅스 소자가 연구되고 작동 메커니즘이 보고되었습니다. 본 논문에서는 보고된 금속 산화물 기반 아날로그 시냅스 소자의 작동 메커니즘을 기반으로 치수 및 구조 변경을 통해 비휘발성 및 인공 시냅스 특성을 조사하고 최적화합니다. 깊이를 통한 느린 이온 확산은 평균 채널 전도도를 변경하므로 대칭 프로그래밍에는 더 얇은 채널이 필요합니다. 프로그래밍은 게이트가 덮힌 영역 아래에서 발생하고 전기장에 의한 이온 이동은 속도 제한 요인으로 작용하므로 작은 프로그래밍 에너지를 위한 수직 구조가 필요합니다. 성능에 대한 채널 두께 및 구조의 영향을 연구했습니다.

핵심어 : 비휘발성 메모리, 뉴로모픽 컴퓨팅, 전기화학 메모리, 이온 전도, 전이금속 산화물, 구조 최적화

학번 : 2019-28331