Influence of characteristic variation of oxide semiconductor and comparison of the activation function in neuromorphic hardware

Hiroya Ikeda\(^1\)\(^a\), Hiroki Yamane\(^2\)\(^b\), Yuta Takishita\(^1\)\(^c\), Mutsumi Kimura\(^1,3\)\(^d\), and Yasuhiko Nakashima\(^1\)\(^e\)

\(^1\) Graduate School of Science and Technology, Nara Institute of Science and Technology
8916-5, Takayama Ikoma, 630–0192, Japan

\(^2\) Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama Ikoma, 630–0192, Japan

\(^3\) Department of Electrics and Informations, Ryukoku University
1-5, Seta Otsu, 520–2194, Japan

\(^a\) ikeda.hiroya.if6@is.naist.jp
\(^b\) yamane.hiroki.yf0@is.naist.jp
\(^c\) takishita.yuta.tt6@is.naist.jp
\(^d\) mutsu@rins.ryukoku.ac.jp
\(^e\) nakashim@is.naist.jp

Received July 31, 2019; Revised November 18, 2019; Published April 1, 2020

Abstract: As the amount of data that people handle increases, the conventional Neumann-type computer architecture is reaching its limits. Therefore, research on hardware implementation of machine learning systems is being actively conducted. In this paper, we have implemented and evaluated neuromorphic hardware that realizes human brain neurons and synapses using oxide semiconductor of amorphous In-Ga-Zn-O (a-IGZO) and a cellular neural network. It was confirmed how variations of initial resistance and deterioration rate of the oxide semiconductor affect operation accuracy of the neuromorphic hardware. Furthermore, we clarified that an activation function suitable for the hardware implementation is a ReLU function.

Key Words: neuromorphic hardware, oxide semiconductor, cellular neural network, letter correction, characteristic variation, activation function, ReLU function
1. Introduction
1.1 Background
In recent years, with the spread of information terminals, the amount of information data handled by people has been explosively increasing [1]. In addition to the data growth, how to process and use big data including unstructured and missed data is a social issue because it cannot be managed [2]. The conventional computers are built on the Neumann-type architecture, which store data and programs to be processed in main memory beforehand [3]. The central processing unit (CPU) calls and executes program instructions and data from the main memory in order. The configurations of the CPU and main memory are physically separate, and they are connected through a transmission path through which each electronic circuit exchanges data. The versatility of the computer can be improved with these configurations by checking the program and results step by step, because the operation instruction and data to be processed are read from the memory at each step, and the operation result is also written one by one. However, because there is a limit to the amount of data that can be transmitted through this transmission path, performance is degraded when data increases so much, and the conventional computers that process data in a fixed procedure cannot handle big data very well [4].

Moreover, progress of semiconductor manufacturing technology have improved computational speed and efficiency of the sequential computing models [5]. However, growth has been slowed owing to the limits of semiconductor miniaturization and rising manufacturing costs. Therefore, there is a growing demand for refinements of the Neumann-type architecture and creation of a non-Neumann-type one [6].

1.2 Neumann-type computing
In spite of the disadvantages of the Neumann-type computers, there are many researches, because they are general-purpose computational models that can be applied to a variety of problems. However, the application of artificial intelligence (AI) technology is expanding, and calculations such as machine learning and combinatorial optimization using big data require a large amount of resources in terms of calculation amount and power consumption. Machine learning is supported by a large number of parameters, and calculating these parameters requires a unit with high computational power. In current machine learning, a graphical processing unit (GPU) known as a processor specialized for image processing and field programmable gate array (FPGA) capable of programming logic circuits suitable for data and processing content to be handled are used as hardware platforms [7, 8]. However, it is itself still a serious problem that a large amount of resources in terms of calculation amount and power consumption are required as abovementioned.

1.3 Non-Neumann-type computing
Instead of the Neumann-type computing, non-Neumann-type computing is considered to be effective to solve the abovementioned problems. Neumann computing explained in Sections 1.1 and 1.2 is an architecture that describes a procedure for solving a problem as a program and executes the program sequentially, but other architectures exist. These are collectively referred to as non-Neumann computing. Some non-Neumann computing has an architecture that essentially does not have a von Neumann bottleneck. For example, it is possible to realize machine learning by providing primitive processing capability to memory devices that were simply used to temporarily store data, and combining them in large quantity, or brain-type computing that mimics a neural network [9]. In these non-Neumann computing memory and processing units are composed near without physical distance, and there is no need to communicate between the logic unit and memory. Therefore, it is not necessary to exchange instructions and data between the processor and memory, which is a problem of the Neumann-type computer, and it is possible to respond to big data in a timely manner.

Among non-Neumann-type computing, brain-type computing is expected as a new technology that solves problems which cannot be handled by existing computers. Brain-type computing has an architecture that mimics a neural network. The neural networks and neuromorphic computing are attracting attention. A neural network is a computational model that uses artificial neurons simpli-
fying the operation of biological neurons as nodes and changes the connection strength among the nodes by learning [10]. This is a model based on statistical methods that mimics only brain function. A neural network is composed of an input layer, an output layer, and a hidden layer, and a synaptic weight that determines the connection strength between neurons existing between the layers. The input data is weighted and the sum of these values is input to the activation function to obtain the output of this neuron. It propagates output data from the input layer to the output layer. Neural networks are attracting attention to improve the performance of machine learning because the accuracy of pattern recognition has been greatly improved by deep learning using a multi layered neural network with many nodes. Large amount of input informations are distributed to and stored in the circuit elements, and can be processed simultaneously by interaction between them. The processing time of brain-type computing is equal to the response time of the circuit elements, and the brain-type computing can perform the processing that requires several tens of steps with the CPU of a conventional digital computer in a few steps [11]. Therefore, it can be said that the brain-type computing is an approach that is very promising as an excellent computing except the Neumann-type computing.

Neuromorphic computing attracts attention among the brain-type computing, which is a computational model that mimics the structure of biological brain neural circuits and the mechanism of neuronal firing. In contrast, in neuromorphic computing, the hardware itself mimics the brain neural circuit, including the mechanism of living neurons firing [12]. Its basic components are neurons and synapses. A neuron performs non-linear processing and is realized by a CMOS digital arithmetic circuit or an analog arithmetic circuit. The synapse is responsible for determining the strength of the connection between adjacent neurons and storing the strength of the connection with analog or digital memory.

A memory chip in neuromorphic computing has a structure in which a memory cell array is arranged on a matrix similar to a brain neural network. In general, it often has built-in multiply-accumulating functions using Ohm’s law and Kirchhoff’s law. In the previous researches, research was conducted to realize virtual neural networks that are effective for pattern recognition and image processing using software [10, 12–14]. However, software-implemented neural networks have the problem of being large in size and power consumption. Therefore, researches to realize neuromorphic computing built on actual hardware are actively being carried out [15, 16]. It is possible to achieve high performance, large scale integration, low power consumption, etc., using the neuromorphic computing because the biological brain can achieve them [17–19]. In such a system, it is necessary to simplify the configuration of the neuron and the synapse and make it inexpensive.

Research on brain-type computing is being carried out in conjunction with the evolution of brain science that analyzes the mechanisms and functions of the brain through the establishment of machine learning methods in recent deep learning, and the progress of computer science including electronics. In this paper, we propose the realization of neuromorphic hardware with neuromorphic computing aimed at imitating human brain neural network more than brain-type computing.
1.4 This study

In this study, neuromorphic hardware using oxide semiconductor synapse and cellular neural network was developed and evaluated by using logic simulation. In particular, by using the oxide semiconductor synapses of amorphous In-Ga-Zn-O (a-IGZO), which can be manufactured at low temperature and highly integrated in three dimensions, it is possible to construct a larger scale neural network. In order to realize neuromorphic hardware, electronic elements corresponding to neurons and synapses are required. The synapse that connects neuron circuits with neuromorphic hardware is usually composed of resistive elements. The resistance value needs to be rewritten according to learning. Therefore, there is a research report in which a device works as a nonvolatile resistance change memory device such as memristor or ReRAM (Resistive Random Access Memory) that is used for a synapse circuit. The memristor operates as a non-volatile memory in which the resistance value changes stepwise each time the voltage is applied, and each intermediate resistance value is retained even when the voltage is cut off. ReRAM is also referred to as a resistance change type memory, and is a non-volatile memory that stores, as data, the resistance value of an element changed by applying a voltage to a metal oxide sandwiched between electrodes.

Currently, there are several research reports on the development of synaptic devices using memristor and ReRAM structures [20–23]. Examples thereof include titanium oxide (TiO$_2$), hafnium oxide (HfO$_2$), and tantalum oxide (TaOx). Firstly, titanium oxide and hafnium oxide are used in neuromorphic hardware as synaptic devices with a memristor structure [20, 21]. Titanium oxide enables monolithic three-dimensional stacking by sputtering at 300$^\circ$C or less during film formation. However, for a-IGZO, the sputtering temperature during film formation can be about room temperature, and it enables monolithic three-dimensional stacking. In addition, the titanium oxide-based memristor used in the former research has a more complicated structure than a-IGZO because it uses multiple material. a-IGZO functions as a variable resistance element with a single material. It does not have a complicated structure and can be easily manufactured. Hafnium oxide is used for neuromorphic hardware as an analog variable resistance element, like titanium oxide. The hafnium oxide-based memristor used in the latter research has a stacked structure of titanium nitride (TiN) and titanium (Ti) in addition to the oxide hafnium dioxide. Compared to a-IGZO, the material is expensive. Next, We explain researches on tantalum oxide as a device with ReRAM structure. Tantalum oxide has a high relative dielectric constant and excellent insulating properties. It is highly stable in resistance change and has been developed as an analog resistance change element for neuromorphic applications. Previous researches have included tantalum oxide-based ReRAMs that use Ta$_2$O$_5$ / TaOx as the oxide materials [22, 23]. Compared to a-IGZO, the structure is complicated, and it is important to adjust the ratio of the material composition in order to obtain stable operation. In order to realize a neural network at the hardware level, it is necessary to select an optimal device and structure.

Particularly in this paper, in order to produce neuromorphic hardware, it is necessary to evaluate characteristic variation of the device used for synapses and select an activation function adapted to the neuron circuit. What is considered in this research is the variation in the initial resistance value of the oxide semiconductor and the variation in the degradation rate of them. In addition, activation functions to be evaluated in this research are a step function, sigmoid function [24], and ReLU function [25], which are widely known as nonlinear functions and often used in neural networks. Influence of the characteristic variation in oxide semiconductor as synapses in the neuromorphic hardware and the optimal activation function for the neuron circuit will be evaluated by the logic simulation and discussed. In order to confirm whether the fabricated neuromorphic hardware works as expected, we evaluated it by simulating it as a letter correction system. The letter correction system which has been simulated this time is an algorithm that learns letters of 12 $\times$ 12 pixels and then corrects slightly different shapes to the learned letters [26]. At that time, the input letter pattern is memorized in the synaptic connection strength among pixels that compose the letter. Therefore, the entered letters are memorized over the entire network. In this paper, it is not a neuron circuit that memorizes information. In the paper [26], neuromorphic hardware is fabricated using cellular neural network and oxide semiconductor a-IGZO. The paper examines the effect of a-IGZO’s variation in deterioration rate of neuromorphic hardware using a letter correction algorithm. In this paper,
the effect of variation in the “initial resistance value” of a-IGZO on the neuromorphic hardware in addition to the deterioration rate is verified using a letter correction algorithm. Furthermore, the configuration of the neuron circuit is changed to verify which of the three “activation functions” of the Step function, Sigmoid function, and ReLU function can obtain the highest letter correction rate. Through this verification, the optimal activation function of neuromorphic hardware is determined.

2. Related research

Neuromorphic computing is a hardware-implemented approach that mimics the neuron and synapse functions and firing mechanisms of the brain \[27, 28\]. It is different from a technical approach to improve the calculation performance of product-sum operation using the GPU and FPGA based on the conventional Neumann type computer. The neuromorphic computing has the same advantages as the brains of a living bodies because it mimics the neural network of the brains at the hardware level. First, the machine size is very small: The machine size of Watson, which is famous for the conventional neural networks, is known to be about 10 refrigerators. On the other hand, neuromorphic computing that mimics the brain can be realized with a chip size as large as that in a personal computer \[29\]. Next, there is the advantage of low power consumption: Although Watson consumes about 100 kW, neuromorphic computing is expected to operate at only 20 W \[30\]. Finally, it has the advantage of being robust. The human brain is known to lose 100,000 neurons every day, but it can still maintain the necessary functionality. Given the above advantages, neuromorphic computing certainly seems to have better performance than traditional neural networks. Some neuromorphic computing is well known, such as Qualcomm’s Zeroth processor \[32\], TeraDeep’s nn-X \[33\], IBM’s TrueNorth \[34\], and Kyushu Institute of Technology’s integrated brain system \[35\], but these large scale integrated (LSI) chips are hybrid systems of the Neumann-type computers and neuromorphic computing. For example, only a few neuron elements are prepared and many neuron elements are virtually emulated using time sharing algorithm, discrete values are stored in digital memory and continuous values like analog memory are approximated, etc. Therefore, the advantages of neuromorphic computing can be obtained only in limited ways. The neuromorphic chip “Loihi” reported in the paper \[29\] is made of digital circuits, with 128 neuromorphic cores arranged in a mesh, 130,000 neurons and 130 million synapses. Moreover, it is equipped with a synaptic load updating function of a single chip. The neuromorphic chip “Neurogrid” reported in the paper \[31\] has a structure in which digital and analog circuits are mixed. Neurogrid implements all circuits other than the axon axis with analog circuits, which makes it extremely energetically efficient and can simulate millions of neurons in real time using several watts. In addition, since the conventional semiconductor manufacturing technology is used, only a two dimensional structure can be obtained in principle whereas three dimensional structures are used in the brains, many processing elements are prepared, and the connection structure is simple.

As an example of a neuromorphic system, we will explain IBM’s TrueNorth \[34\]. TrueNorth chip is the largest of IBM’s chips at least at that time with 5.4 billion transistors, 1 million programmable neurons, and 256 million programmable synapses and runs with as little as the power of 70 mW. However, it is still far from a human brain with 100 billion neurons and 100 to 150 trillion synapses. If implementing with TrueNorth is considered, the 100 trillion synapses in a human brain is about 400,000 times of the 256 million synapses in TrueNorth. That is, in order to reproduce a human brain, the 400,000 TrueNorth chips and power consumption of 70 mW x 400,000 = 28 kW are required. Therefore, although it is possible to achieve it using a huge server, it can be understood that embedding into an independent device is not realistic.

Our proposed neuromorphic hardware is a complete analog computer made of using oxide semiconductor a-IGZO. The synaptic weight value is recorded as an analog value on a single chip, and the synaptic load is updated. Consequently, higher energy efficiency can be expected. In this proposal, synapses can be stacked three-dimensionally by using oxide semiconductors, and in the future, it can be expected to develop neuromorphic hardware with high integration.

3. Proposed method

In this section, we propose a method to configure neuromorphic hardware using some novel elemental...
technologies. An oxide semiconductor is used to configure the neural network that is easy to implement in hardware, and then how the network learns is shown.

3.1 Oxide semiconductor synapse

Oxide semiconductors exhibit excellent performance even when deposited in the amorphous phase at low temperatures [36], and in addition, oxide semiconductors can be potentially formed using printing methods, which facilitate achieving a three-dimensional layer structure for neuromorphic computing. a-IGZO is a typical example among these [37], and recently, it is used as thin film transistors (TFTs) for flat panel display (FPDs) such as high resolution liquid crystal displays (LCDs) and organic light emitting diode (OLEDs) [38]. Figure 3 shows the a-IGZO device [41]. An a-IGZO thin film is deposited on a silicon wafer with metal electrodes using radio frequency (RF) magnetron sputtering, where the sputtering target is IGZO ceramic whose composition is In : Ga : Zn = 1 : 1 : 1 and the sputtering gas is Ar, and the thickness of the a-IGZO thin film is 70 nm. These manufacturing conditions are the same as those for the TFTs with the highest performance.

![Fig. 3. a-IGZO device.](image)

Figure 4 shows a current flowing through the a-IGZO device. Here, 1.8 V was applied to the left and right metal electrodes, and the current flowing through the a-IGZO element was measured. It was found that the conductance decreases continuously when the current flows, and the decrease in conductance is irreversible but relatively slow. This characteristic can be used to make use of the modified Hebbian learning rules described below. The function required for the synapses in the learning phase is that the connection strength properly changes depending on the state of two connected neurons, and maintains. The conductance can be regarded as the connection strength because how easily the signal is transmitted is dependent on it. Moreover, when the two neurons connected to the device are in different states, there is a voltage deference, current flows, deteriorating occurs, and the conductance decrease, which corresponds to that the connection strength decreases. On the other hand, when the two neurons connected to the device are in the same states, there is no voltage deference, no current flows, no deteriorating occurs, and no conductance decrease, which corresponds to that the connection strength maintains. This learning method is different from Hebb’s learning rule [42], which raises the connection strength of synapse when the neighboring neurons are both in firing states. In the literature [43], this learning method is called the modified Hebb’s learning rule, where it was confirmed that AND, OR, and XOR circuits are remembered based on this learning rule using poly-Si TFTs. We use the oxide semiconductor a-IGZO as synapses, which is a device with
a characteristic that the conductance deteriorates when current flows [38].

Here, we describe the size and power consumption when a-IGZO thin films are used as synapses and the number of a-IGZO is equal to the number of synapses in the human brain. When a-IGZO is used in the synaptic element of our proposed neuromorphic hardware, the size of the a-IGZO thin film is roughly 1 $\mu$m$^3$. If the number of a-IGZO is equal to the number of synaptic elements in the human brain, 100 trillion a-IGZO synapses are required. If we do not consider the size of the neuron element here, the total size of the neuromorphic hardware is thought to be as small as 100 cm$^3$. This is because the number of neuronal elements is much smaller than the number of synaptic elements as shown in previous studies. In addition, since the power consumption of one a-IGZO thin film is 1 pW, the power consumption of the entire neuromorphic hardware is thought to be as small as 100 W. This power consumption is only about 5 times that of the human brain. Besides, a learning function is naturally included because the connection strength of the network can be changed independently. This is done along the modified Hebbian learning rule described above. We compare the specs of the TrueNorth chip in the previous research with that of the neuromorphic hardware which we propose under the assumption of the same number of synapses as the human brain. In terms of size, the is 172 m$^2$, and our proposed neuromorphic hardware is 100 cm$^3$. Our proposal does not take into account of neurons size, so it cannot be clearly compared. However, since a-IGZO can be stacked three-dimensionally, it has a high degree of integration. In terms of power consumption, TrueNorth is 28 kW, which is 14,000 times the power consumption of humans, whereas our proposal is 100 W, which is only five times the power consumption of humans. Finally, our proposed neuromorphic hardware has a learning function because the synaptic connection strength can be changed independently by changing the resistance value of the synapse. This is a major difference in the conventional neuromorphic hardware, which does not have its own learning function. Thus, our neuromorphic hardware has the potential for solving the size and power consumption problems of the conventional neuromorphic hardware and achieving the research objectives.

3.2 Cellular neural network

Figure 5 shows a configuration of a neuron circuit, where one inverter is composed of two transistors. In this case, the total number of required transistors is only four, and it is suitable for high integration. The neuron circuit of this configuration exhibits characteristics similar to a sigmoid function, and has the simple property of maintaining the state of voltage and firing or changing stably in accordance with a certain threshold. This neuron circuit will be optimized in the following section.

The configuration of the entire cellular neural network we produce is shown in Fig. 6. In the cellular neural network, adjacent neurons are connected by synapses, and each neuron is connected to top, bottom, left, right and diagonal neurons [39, 40]. The cellular neural network is suitable for pattern recognition and image processing because its structure is similar to that of retinal cells. Furthermore, since the connection of neurons through synapses is not complicated, it can be easily scaled up when
the hardware is created. However, there are concerns that the number of synapses is small and the efficiency of learning is low, and it is important to confirm the operation on simulations or real machines.

Figure 7 shows the synapse connection between the neurons, where the synapse connection includes a concordant synapse and discordant synapse. The concordant synapse connects the positive output of the previous neuron and the input of the next neuron and operates to make both neurons in the same state. On the other hand, the discordant synapse connects the negative output of the previous neuron and the input of the next neuron and operates to make both neurons into different states. These two synapses can be used to increase or decrease the net connecting strength like an actual synapse. For example, deterioration of the concordant synapse relatively increases the net connecting strength.
Figure 8 shows the connection of a neuron and eight adjacent neurons. The red arrow indicates the direction of current flow, and the size of the arrow roughly indicates the amount of current flow. The currents flowing from the output of the adjacent previous eight neurons are weighted through two types of synapses, concordant synapse and discordant synapse, according to the connecting strength of each synapse, and all the currents are summed and input to the next neuron. Depending on the input current, the next neuron becomes either firing state or stable state, and the state is maintained by the neuron circuit. Furthermore, in this cellular neural network, since the synapse is connected not only in the direction from the previous neuron to the next neuron but also interactively, there are four synapses between each neurons. Therefore, it is expected that the number of the synapses will increase and the accuracy of the neural network will be improved.

3.3 Learning principle
Hebbian learning rules are typical rules in biological and artificial neural networks [42]. In Hebbian learning rules, the synapse connection strength increases when the states of both neurons connected
to the synapse are the firing states and maintains or decreases otherwise. On the other hand, in the modified Hebbian learning rule [43], the synapse connecting strength is expressed by the relationship between those of the concordant synapse and discordant synapse. The synaptic connection strength changes depending on the state of adjacent neurons, which is similar to Hebbian learning rules. The way of changes in the synaptic connection strength according to the state of adjacent neurons is described below. Figure 9(a) shows the connection of neurons when the neighboring neurons are in the same state. In this case, large amount of current flows from the input of the next neuron through the discordant synapse connected to the negative output of the previous neuron. As a result, the discordant synapses deteriorate, namely, the deterioration of discordant synapse decreases the electrical conductance, and the synaptic connection strength relatively increases. Figure 9(b) shows the connection of neurons when the neighboring neurons are in different states. In this case, large amount of current flows from the positive output of the previous neuron through the concordant synapse connected to the input of the next neuron. As a result, the concordant synapses deteriorate, namely the deterioration of concordant synapse decreases the electrical conductance, and the synaptic connection strength relatively decreases.

(a) Increased synaptic connections  (b) Decreased synaptic connections

**Fig. 9.** Modified Hebbian learning rule.

The advantage of modified Hebbian learning rules is that it is not necessary to add circuits to control synapse connection strength because the synapse connection strength is automatically controlled using local currents and voltages. Therefore, the modified Hebbian learning rules make it easy to implement neural networks at the hardware level.

**4. Simulation method**

In this research, we evaluated a letter correction system using the logic simulation to verify the operation of the neuromorphic hardware. The outline of the letter correction algorithm is described in this section. Figure 10 shows a cellular neural network composed of 25 x 25 neurons. Here, we used 12 x 12 neurons capable of input / output (I/O) and hidden neurons, where one hidden neuron was placed between all the I/O neurons. The learning efficiency is improved by increasing the hidden layer, but if the hidden layer is increased too much, it becomes difficult for the input data to reach the output layer. In the example of Fig. 10, the character “1” is input to the I/O neuron, the neuron in the firing state is shown in red, and the neuron in the stable state is shown in white.

We give initial values to each neuron and synapse for simulation, where a voltage of 0.9 V is input to the neuron as an initial value, which is a half of the voltage source of 1.8 V. Then, the resistance value is input to the synapse in consideration of its characteristic variation that occurs when the oxide semiconductor is actually manufactured. Figure 11 shows a letter pattern to be learned. The letter pattern is input, namely, the corresponding I/O neurons are set in the firing or stable state, where 1.8 V is input to the neuron in firing state, and 0 V is input to the neuron in stable state. After that, the states of all neurons including I/O neurons and hidden neurons are determined according to the connection strength of oxide semiconductor synapses between the neighboring neurons. Next, current flows through the oxide semiconductor synapse according to the states of the neurons, as a result, the synapse is deteriorated, and the synapse connection strength is reduced. After inputting the letter
pattern to be learned in this way, a slightly different letter pattern is input to the I/O neuron, and we check if the letter pattern output matches the learned letter pattern. If they match, the letter correction is successful, whereas if they do not match, the same procedure is repeated.

Figure 12 shows a successful example of the letter correction. When learning is completed and letter correction is to be performed, the letter pattern of Fig. 12(a) is input, and after waiting for a while, the letter pattern of Fig. 12(b) is output, which means that the letter correction is successful.

5. Initial resistance variation
5.1 Initial resistance variation
It is known that the initial resistance of the oxide semiconductor has variations when it is fabricated. Figure 13 shows the results of the measured resistance of 12 a-IGZO samples. The results show that the mean is 3.75 MΩ and the standard deviation is 2.5 MΩ, and although many samples are about 3 MΩ to 4 MΩ, the standard deviation becomes large because two samples exceed 8 MΩ. Therefore, the influence of the initial resistance variations of the oxide semiconductor on the performance of the
Fig. 12. Example of the successful letter correction.

Fig. 13. Initial resistance variation of the oxide semiconductor synapses.

neuromorphic hardware was evaluated by the logic simulation.

In this research, we used the Monte Carlo method of random input for the synapses of the entire network, assuming that the initial resistance variation of the oxide semiconductor follows a standard normal distribution. Figure 14 shows the histogram of standard normal distribution with the mean of 3.75 MΩ and standard deviation of 1 MΩ. Here, the standard deviation was varied from 0 to 4 MΩ, and the letter correction rate was evaluated. The reason why the standard normal distribution is used for the variation in the initial resistance value of a-IGZO is there are more factors that cause the variation in the initial resistance value of a-IGZO than one. Factors that causing variation include the film formation temperature during a-IGZO film formation, the amount of impurities inside the film formation tool, the degree of vacuum, etc. These factors cannot be determined every time during film formation. Thus, we used the standard normal distribution, which is the most common distribution function and a probability distribution applied to various natural phenomena.

5.2 Evaluation results and discussion

Figure 15 shows the influence of the standard deviation of the initial resistance on the correction rate of the letter correction. Here, the horizontal axis is the standard deviation of the initial resistance, and the vertical axis is the correction rate. It was found that the correction rate start to decrease as the standard deviation increase more than 1.2 MΩ. Therefore, it is necessary to keep the standard
deviation of initial resistance within 1.2 MΩ in order to make the correction factor 90% or more. The standard deviation of 1.2 MΩ is 32% for the mean of 3.75 MΩ, which is large enough for electronic devices. If the equipment and environment for manufacturing the oxide semiconductor is improved, this neuromorphic hardware can operate properly.

6. Deterioration rate variation

6.1 Deterioration rate variation

It is known that the deterioration rate also has variation in the process of flowing current through the oxide semiconductor for deterioration. Therefore, the influence of degradation rate variation on the neuromorphic hardware was evaluated by the logic simulation. Figure 16 shows the deterioration rate obtained from 15 synapses. It can be seen that the deterioration rate variation is about 12%.

Figure 17 shows the histogram of standard normal distribution with the mean of 1 GΩ/A and standard deviation of 0.1 GΩ/A. Here, the standard deviation was varied in the range of 0 to 4%, and the letter correction rate was evaluated. The reason why the standard normal distribution is used for fluctuation of the deterioration rate of a-IGZO is the same as written in section 5.1.
6.2 Evaluation results and discussion

Figure 18 shows the influence of deterioration rate variation on the correction rate. Here, the horizontal axis is the standard deviation of deterioration rate, and the vertical axis is the correction rate. It was found that the letter correction rate decreased when the standard deviation of the deterioration rate exceeded 12% of the mean, which was larger than the influence of the initial resistance variation.

7. Activation function

In the previous section, we evaluated the influence of characteristic variations of the oxide semiconductors on the performance of the neuromorphic hardware. It was clarified that the correction rate decreases when the characteristic variation increases. Therefore, it is necessary to improve the correction rate from the viewpoints of the network architecture. In this section, we show the results of the logic simulation for several activation functions in order to evaluate the influence of the activation functions of the neuron circuits on the letter correction rate.

7.1 Activation function

An activation function is a function that is applied after linear transformation in neural networks. Figure 19 shows the role of activation function. The state of the previous stage neuron is multiplied by the synapse weight, the sum of each is calculated, and the state of the next stage neuron is determined
by the activation function. In this research, a step function, sigmoid function, and ReLU function are considered as the activation function.

### 7.1.1 Step function
The step function is the simplest activation function in which the output is switched at a threshold. It was used when a simple perceptron appeared in 1950 [44]. Equation (1) shows the equation of the step function. Figure 20(a) shows a graph of the step function in which the output value is 0 when the input value is 0 or less, and the output value is 1 when the input value is higher than 0.

\[
f(x) = \begin{cases} 
0 & x < 0 \\
1 & x \geq 0
\end{cases}
\] (1)

### 7.1.2 Sigmoid function
The sigmoid function is an activation function in which the output approaches 1 as the input value increases, and approaches 0 as the input value decreases. It was commonly used when back-propagation appeared in 1986 [24]. Equation (2) shows the equation of the sigmoid function. Unlike the step function, the sigmoid function has a feature that the output changes continuously with changes in the
7.1.3 ReLU function

The ReLU function is an activation function that outputs the same value as the input when the input is positive, and outputs 0 when the input is negative. Equation (3) shows the equation of the ReLU function. It is widely used at present because it can solve the problem that the back-propagation becomes difficult when the neuron layers are deep in the neural networks [25]. Figure 20(c) shows a graph of the ReLU function.

\[ f(x) = \begin{cases} 
0 & x < 0 \\
 x & x \geq 0 
\end{cases} \tag{3} \]

7.2 Implementation of activation function circuit

The Step function and the ReLU circuit can be realized by using the following MOS circuits for neuron circuits. The neuron circuit is made up of two inverters, and each inverter is represented by a MOS circuit as shown in Fig. 21. These are general CMOS inverters, and are composed of nMOS and pMOS transistors. A ReLu circuit can be realized using these inverters by changing the pMOS transistor to an nMOS transistor and connecting the input terminal to a power supply as shown in Fig. 22. The characteristics of each neuron circuit when the input voltage is increased by circuit simulation using HSPICE are shown in Fig. 23 and 24. As a result, the same characteristics as the graphs shown in Figs. 20(a) and (b) were confirmed.

7.3 Evaluation results and discussion

Figure 25 shows the results of difference in correction rate depending on the activation function. The correction rate of sigmoid function falls most rapidly and continues to decline gradually. On the other
Fig. 23. Circuit simulation result of step function.

Fig. 24. Circuit simulation result of ReLU function.

Fig. 25. Change in the letter correction rate by the initial resistance variation for each activation function.
hand, the correction rate for the ReLU function maintains high level even if the initial resistance variation increase to 1.4 MΩ. Therefore, it is considered appropriate to use the ReLU function as an activation function in the neuromorphic hardware proposed in this research.

We will consider the reason why the ReLU function exhibits the excellent performance. As shown in Fig. 26, we will assume the case where the connection strength of a synapse has accidentally become stronger than other synapses due to the characteristic variation. For the case of step function and sigmoid function, even if the change of input is small, output of the neuron sometimes changes from non-firing to firing states drastically. Therefore, it is considered that the influence of characteristic variation seems serious. On the other hand, for the case of the ReLU function, the output does not change drastically and changes gradually. This is considered to be the reason why the ReLU function is the most suitable for the activation function.

8. Conclusions

For the cellular neural networks using the oxide semiconductors, the influence of characteristic variations in the oxide semiconductor on the performance of the neuromorphic hardware was evaluated by the logic simulation. First, the influence of the initial resistance variation of the oxide semiconductor was evaluated. As a result, it was found that when the standard deviation of the variation exceeded 32%, the letter correction rate dropped sharply. For electronic devices, it can be said that the standard deviation of 32% is sufficiently large. Therefore, it is considered that the letter correction system can operate sufficiently if the environment for manufacturing oxide semiconductors and uniformity of the device are improved. Next, the influence of the deterioration rate variation of the oxide semiconductor was also evaluated. As a result, it was found that letter correction rate decreased when the standard deviation of the deterioration rate variations exceeded 12%. Moreover, it turned out that the influence of the deterioration rate variation is larger than that of the initial resistance variation. Furthermore, three activation functions were applied to the neuron circuit to obtain the letter correction rate. The ReLU function has been found to work best in our proposed neuromorphic hardware. As future prospects, it is suggested to devise network structures that are not sensitive to the characteristic variation of the oxide semiconductors. Furthermore, if the characteristic variation...
can be suppressed in the manufacturing process of the oxide semiconductors, it is possible to create more accurate neuromorphic hardware.

Acknowledgments

This work is partially supported by KAKENHI (C) 16K06733, VLSI Design and Education Center (VDEC), Yazaki Memorial Foundation for Science and Technology, Support Center for Advanced Telecommunications Technology Research (SCAT), Research Grants in the Natural Sciences from the Mitsubishi Foundation, the Telecommunications Advancement Foundation, RIEC Nation-wide Cooperative Research Projects, collaborative research with ROHM Semiconductor, collaborative research with KOA Corporation, and Innovative Materials and Processing Research Center in Ryukoku University. Lastly, We would like to thank Ms.Bich in Nara Institute Science and Technology for the improvement of this manuscript.

References

[1] Ministry of Internal Affairs and Communications, “The arrival of the big data use year,” White Paper Information and Communications, pp. 52–62, 2017.
[2] A. Nagamatsu, “Current Issues Related to Big Data Utilization in Japan,” Journal of information and management, vol. 36, no. 4, pp. 3–11, 2016.
[3] J. von Neumann, “First Draft of a Report on the EDVAC,” Philadelphia: University of Pennsylvania, 1945.
[4] FUJITSU, Development Process of Electronic Computer, vol. 16, no. 1, pp. 95–105, 1965.
[5] TOKYO OHKA KOGYO CO.LTD., “TOK’s microfabrication processing technology and the future of computers,” CSR Report 2017, pp. 13–18, 2017.
[6] G.E. Moore, “Cramming More Components Onto Integrated Circuits,” Proceedings of the IEEE, vol. 86, no. 1, pp. 82–85, 1998.
[7] J. Nickolls and W.J. Dally, “The GPU Computing Era,” IEEE Micro, vol. 30, no. 2, pp. 56–69, 2010.
[8] Altera Corp, “Arria 10 Core Fabric and General Purpose I/Os Handbook,” Arria 10 Handbook, 2016.
[9] IBM, “IBM Scientists Demonstrate In-memory Computing with 1 Million Devices for Applications in AI,” https://www.ibm.com/blog...2017/10/ibm-scientists-demonstrate-memory-computing-1-million-devices-applications-ai/
[10] J.E. Dayhoff, “Neural Network Architectures, An Introduction,” Van Nostrand Reinhold, 1990.
[11] G. Matsumoto, Y. Shigematsu, and M. Ichikawa, “Information science from brain research,” Journal of Information Processing Society of Japan, vol. 38, no. 5, pp. 369–375, 1997.
[12] R. Hecht-Nielsen, “Neurocomputing,” Addison-Wesley Reading, 1990.
[13] S. Becker, “Self-Organizing Neural Network that discovers Surfaces in Random-Dot Stereograms,” Nature, no. 355, pp. 161–163, 1992.
[14] J.V. Stone, “Neural-Network Models: Predicting Spontaneous Recovery of Memory,” Nature, no. 414, pp. 167–168, 2001.
[15] Z. Du, D.D.B.-D. Rubin, Y. Chen, L. He, T. Chen, L. Zhang, C. Wu, and O. Temam, “Neuromorphic Accelerators: A Comparison Between Neuroscience and Machine-learning Approaches,” Proceeding of the 48th International Symposium on Microarchitecture, pp. 494–507, 2015.
[16] J. Schemmel et al., “Live demonstration: A scaled-down version of the brainscales wafer-scale neuromorphic system,” Proceedings of the 2012 IEEE International Symposium on Circuits and Systems (ISCAS), p. 702, 2012.
[17] T. Kasakawa, “An Artificial Neural Network at Device Level using Simplified Architecture and Thin-Film Transistors,” IEEE Trans. Electron Devices, vol. 57, no. 10, pp. 2744–2750, 2010.
[18] M. Kimura, “Apoptotic Self-Organized Electronic Device using Thin-Film Functions,” Jpn. J. Appl. Phys., 54, 03CB02, 2015.
[19] M. Kimura, “Cellular Neural Network formed by Simplified Processing Elements composed of Thin-Film Transistors,” Neurocomputing, vol. 248, pp. 112–119, 2017.

[20] M. Prezioso, F. Merrikh-Bayat, B.D. Hoskins, G.C. Adam, K.K. Likharev, and D.B. Strukov, “Training and operation of an integrated neuromorphic network based on metal-oxide memristors,” Nature, vol. 521, pp. 61–64, 2015.

[21] E. Covi, S. Brivio, A. Serb, T. Prodromakis, M. Fanciulli, and S. Spiga, “HfO2-based memristors for neuromorphic applications,” In Proceedings of the 2016 IEEE International Symposium on Circuits and Systems (ISCAS), Montreal, QC, Canada, pp. 22–25, 2016.

[22] A. Kawahara, R. Azuma, Y. Ikeda, K. Kawai, Y. Katoh, Y. Hayakawa, K. Tsuji, S. Yoneda, A. Himeño, K. Shimakawa, et al., “An 8 Mb multi-layered cross-point ReRAM macro with 443 MB/s write throughput,” IEEE J. Solid-State Circuits, vol. 48, no. 1, pp. 178–185, 2013.

[23] Z. Wei, Y. Kanzawa, K. Arita, Y. Katoh, K. Kawai, S. Muraoka, S. Mitani, S. Fujii, K. Katayama, M. Iijima, et al, “Highly reliable TaOx ReRAM and direct evidence of redox reaction mechanism,” In Proceedings of the 2008 IEEE International Electron Devices Meeting, pp. 15–17, 2008.

[24] Y. LeCun, L. Bottou, G.B. Orr, and K.R. Muller, “Efficient BackProp,” 1998.

[25] X. Glorot, A. Bordes, and Y. Bengio, “Deep Sparse Rectifier Neural Networks,” Aistats, vol. 15, p. 275, 2011.

[26] H. Ikeda, H. Yamane, M. Kimura, Y. Shibayama, and Y. Nakashima, “Evaluation of Neuromorphic Hardware using Cellular Neural Networks and Oxide Semiconductors,” 2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), pp. 603–608, 2019.

[27] C. Mead, “Analog VLSI and Neural Systems,” Addison-Wesley Publishing, 1989.

[28] Y. Arima, K. Mashido, K. Okada, T. Yamada, A. Maeda, H. Kondoh, and S. Kayano, “A 336-neuron, 28 k-synapse, self-learning neural network chip with branch-neuron-unit architecture,” IEEE Journal of Solid State Circuits, vol. 26, no. 11, pp. 1637–1644, 1991.

[29] Qualcomm, “Introducing Qualcomm Zeroth Processors: Brain inspired computing,” https://www.qualcomm.com/news/onq/2013/10/10/introducing-qualcomm-zeroth-processors-brain-inspired-computing, Accessed in May 2019.

[30] Y. Chen and Q. Wu, “Neuromorphic computing: A SoC scaling path for the next decades,” IEEE International SOC Conference, 2012.

[31] B.V. Benjamin et al., “Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations,” Proc. IEEE, vol. 102, no. 5, pp. 699–716, 2014.

[32] V. Gokhale, J. Jin, A. Dundar, B. Martini, and E. Culurciello, “A 240 G-ops/s Mobile Co-processor for Deep Neural Networks,” Computer Vision and Pattern Recognition Workshops, 2014.

[33] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G. Nam, B. Jackson, and D.S. Modha, “Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip,” IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 34, no. 10, pp. 1537–1557, 2015.

[34] Information processing model by neuron’s spike firing timing and its LSI implementation, http://www.brain.kyutech.ac.jp/~morie/topics/spiking.shtml, Accessed in March 2019.

[35] J.F. Wager, “Transparent electronics,” Science, vol. 300, no. 5623, pp. 1245–1246, 2003.

[36] K. Nomura, H. Ohta, A. Takagi, T. Kamlia, M. Hirano, and H. Hosono, “Room-temperature fabrication of transparent flexible thin-film transistors using amorphous oxide semiconductors,” Nature, vol. 432, pp. 488–492, 2004.

[37] S. Imai and M. Kimura, “Degradation evaluation of a-IGZO TFTs for application to AM-OLEDs,” Electron Device Letters, vol. 31, pp. 963–965, 2010.
[39] L.O. Chua and L. Yang, “Cellular neural networks: theory,” *IEEE Transactions on Circuits and Systems*, vol. 35, no. 10, pp. 1257–1272, 1988.

[40] L.O. Chua and L. Yang, “Cellular neural networks: applications,” *IEEE Transactions on Circuits and Systems*, vol. 35, no. 10, pp. 1273–1290, 1988.

[41] M. Kimura, K. Ikushima, D. Yamakawa, H. Yamane, and Y. Nakashima, “Real Neuromorphic System using LSI Chip and Thin-Film Devices,” *International Conference of Neuromorphic Systems 2019 (ICONS2019)*, 2019.

[42] D.O. Hebb, “The organization of behavior: A neuropsychological theory,” *Science Education*, vol. 34, no. 5, pp. 336–337, 1950.

[43] T. Kasakawa, H. Tabata, R. Onodera, H. Kojima, M. Kimura, H. Hara, and S. Inoue, “An artificial neural network at device level using simplified architecture and thin-film transistors,” *IEEE Transactions on Electron Devices*, vol. 57, no. 10, pp. 2744–2750, 2010.

[44] H.D. Block, “The Perceptron: A Model for Brain Functioning I,” *Reviews of Modern Physics*, vol. 34, no. 1, pp. 123–135, 1962.