Ph.D. DISSERTATION

Demonstration of Unsupervised Learning With Spike-Timing-Dependent Plasticity Using NOR-Type Nonvolatile Memory Arrays

NOR-형 비휘발성 메모리 어레이를 이용한 스파이크 시점 의존 가소성 기반 비지도 학습의 구현

BY

CHUL-HEUNG KIM

February 2019

DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Demonstration of Unsupervised Learning With Spike-Timing-Dependent Plasticity Using NOR-Type Nonvolatile Memory Arrays

NOR-형 비휘발성 메모리 어레이를 이용한 스파이크 시점 의존 가소성 기반 비지도 학습의 구현

지도교수 이종호

이 논문을 공학박사 학위논문으로 제출함

2019 년 2 월

서울대학교 대학원

전기컴퓨터공학부

김철흥

김철흥의 공학박사 학위논문을 인준함

2019 년 2 월

위원장 : 박병국 (인)

부위원장 : 이종호 (인)

위원 : 김재하 (인)

위원 : 유승주 (인)

위원 : 권혁인 (인)

Demonstration of Unsupervised Learning With Spike-Timing-Dependent Plasticity Using NOR-Type Nonvolatile Memory Arrays

by

Chul-Heung Kim

Advisor: Jong-Ho Lee

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering and Computer Science)
in Seoul National University
February 2019

Doctoral Committee:

Professor Byung-Gook Park, Chair
Professor Jong-Ho Lee, Vice-Chair
Associate Professor Jaeha Kim
Professor Sungjoo Yoo
Associate Professor Hyuck-In Kwon

ABSTRACT

Conventional von Neumann computing architecture is at a disadvantage in terms of speed and power consumption in high-level cognitive applications. Therefore, a new architecture to overcome this problem, the neuromorphic system, is attracting attention as the next generation computing system.

In this dissertation, two types of NOR-type nonvolatile memory arrays are proposed for use as synaptic device array in the neuromorphic system. The SONOS gated-diode memory is proposed as the first candidate for the synaptic device. The learning process of MNIST digit patterns is presented by simulation. First, spike-timing-dependent plasticity (STDP) learning in single-neuron string (784 × 1) is demonstrated. Then, STDP learning in multi-neuron array (784 × 3) with lateral inhibition function is demonstrated. Meanwhile, the key factors of STDP unsupervised learning such as input noise density (ρ_{noise}), synaptic weight margin (W_{margin}), and lateral inhibition factor [%] are investigated for the proper learning.

Next, the TFT-type NOR flash memory synaptic device with a half-covered floating gate (FG) that overcomes the disadvantages of the SONOS gated-diode memory is proposed. The long-term potentiation (LTP) and long-term depression (LTD) required for STDP behavior are implemented using the proposed pulse scheme. Unsupervised online learning is successfully demonstrated with STDP learning rule through software simulation reflecting the LTP / LTD characteristics of the fabricated synaptic device. The learning and recognition process of 28 × 28 MNIST handwritten digit patterns are presented.

As a result, an approach is suggested to use hardware-based spiking neural network implemented by synaptic device array using conventional CMOS technology for visual pattern recognition system.

Keywords: neuromorphic system, synaptic device, unsupervised learning, spike-timing-dependent plasticity (STDP), gated-diode memory, NOR flash memory.

Student number: 2013-20779

CONTENTS

Chapter 1

Chapter 2

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Neuromorphic computing

Neuromorphic computing, proposed to overcome the limitations of von Neumann architecture, has received much attention in recent years. In a different way, machine learning has attracted great interest in the IT industry and has been developing rapidly with the performance improvement of graphics processing unit (GPU)-based hardware accelerators. There are various algorithms for machine learning, but the deep neural network (DNN) technology based on the back-propagation (BP) algorithm exerts excellent performance in many fields such as image, speech recognition, translation, and human cognitive ability [1-6]. The most advanced architectures of DNN include convolutional neural networks (CNNs) and recurrent neural networks (RNNs). However, there are important issues concerning power consumption, area occupied by the hardware platform, and training time. Therefore, there is a need to realize a small-area neuromorphic artificial neural network (ANN) with low power [7, 8]. Table 1.1 summarizes the different types of neuromorphic ANNs and their respective features [9]. The human brain described on the left side of the table is very powerful to recognize the real world problem with very low power but the learning mechanism and structure of the human brain are not clearly defined yet. The deep learning shown on the right is based on a software-based learning algorithms and Von Neumann computer architectures. It is very efficient for recognition task, but the power consumption is extremely high. In ANNs using neuromorphic technology, learning algorithms can be categorized into two categories: bio-inspired learning algorithm and software-based learning algorithm [10-12]. Learning algorithms based on biology such as spike-timing dependent-plasticity (STDP) and spike-rate-dependent plasticity (SRDP) have implemented a model of biological neuron cell behavior [13]. For bio-inspired learning algorithms, there are two subcategories: supervised learning and unsupervised learning. Research on ANNs using bio-inspired learning algorithms is biased towards the use of unsupervised learning. However, supervised learning is efficient in a certain field. These bio-inspired learning algorithms have the advantages of local learning and inferencing within large arrays, making them suitable for low power neuromorphic computing. In memory technology for ANN implementation using these algorithms, it is very important to have their own synaptic weight update scheme to learn without the help of external computing systems. In addition, weight updating methods and endurance are also important for reducing time and power consumption in the learning process. Also, in order to prevent errors in the inferencing process of learned synaptic arrays, retention characteristics should also be considered. In contrast, the BP of the neural network is a typical software-based learning algorithm [14]. The weight update is represented by the conductance change of a synaptic device by a BP algorithm based on the error between the target value and the output value. These algorithms allow hardware-based neural networks (HNNs) to provide high speed and low power operation over von Neumann-based platforms, especially GPU-based platforms.

Table 1.1. Different types of neural networks and their respective features [9].

<table>
<tr><th></th><th>Human brain</th><th colspan="2">Neuromorphic</th><th>Deep learning</th></tr>
<tr><td rowspan="2">Target</td><td rowspan="2">Biology</td><td>Spiking neural networks (SNNs)</td><td colspan="2">Deep neural networks (DNNs)
Convolutional neural networks (CNNs)
Recurrent neural networks (RNNs)</td></tr>
<tr><td colspan="2">HW-based</td><td>SW-based</td></tr>
<tr><td rowspan="2">Components</td><td>Neuron array</td><td>Neuron array (Integrate & fire)</td><td>Neuron array (Activation function, Integrate & fire)</td><td rowspan="2">von Neumann architecture (GPU, TPU, etc.)</td></tr>
<tr><td>Synapse array</td><td>Synaptic device array</td><td>Synaptic device array</td></tr>
<tr><td>Learning algorithm</td><td>STDP, SRDP, etc.</td><td>STDP, SRDP (Bio-inspired)</td><td colspan="2">Back-propagation (Software-based)</td></tr>
<tr><td>Power consumption</td><td>Extremely low</td><td>Low</td><td>Intermediate</td><td>High</td></tr>
<tr><td>Maturity</td><td>Extremely high</td><td>Low</td><td>Intermediate</td><td>High</td></tr>
</table>

1.2 Spike-timing-dependent plasticity (STDP)

This chapter introduces the STDP learning algorithm and the features of the ANN implemented using this algorithm. There are two representative bio-inspired learning algorithms, STDP and SRDP. These algorithms are learning methods developed from the learning mechanisms observed in the biological brain. STDP is a learning mechanism that changes the synapse weight by the time difference between the signal from the presynaptic neuron and the signal from the postsynaptic neuron. Figure 1.1 shows the change of synaptic weight between presynaptic and postsynaptic neurons in response to their differences in firing time [15]. Another learning algorithm SRDP determines the synaptic weight change by the frequency of the signal from the presynaptic neuron applied to the synapse. Figure 1.2 shows the weight change in response to the frequency of postsynaptic neuron [16]. Here, we classify ANN using STDP learning algorithm as supervised / unsupervised learning. Next, we describe the requirements of synaptic devices that are used to implement ANN by applying live emotion learning algorithm.

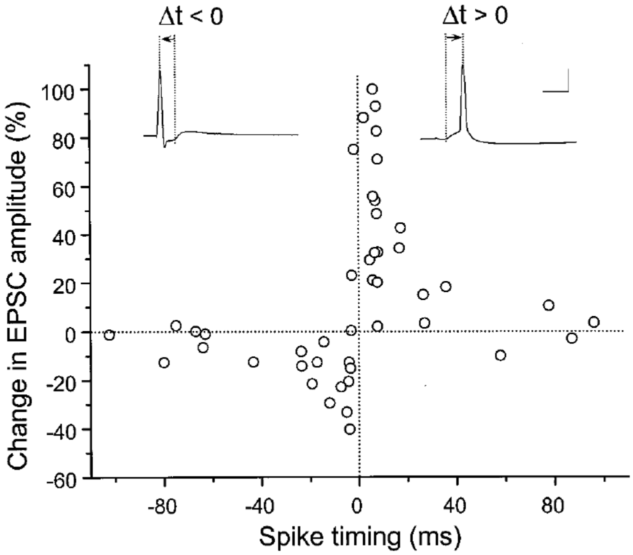


Figure 1.1. Critical window for the induction of synaptic potentiation and depression (learning curve for STDP) [15].

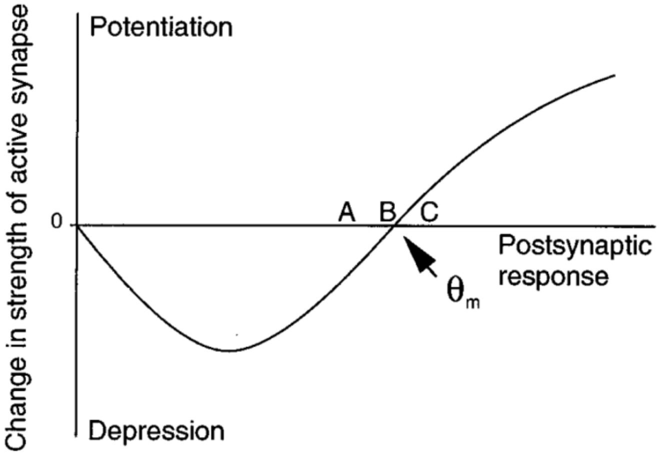


Figure 1.2. Function controlling synaptic plasticity at the Cooper synapse (learning curve for SRDP) [16].

1.2.1 Supervised learning

In the case of bio-inspired learning, unsupervised learning is being studied more widely. However, recent research has shown and demonstrated several advantages of supervised learning. Kim et al explained that supervised learning is significantly more efficient than unsupervised learning with the same number of output neurons and synapses [17]. Adjustment of synapse weight is controlled by feedback spikes from integration & fire (I&F) circuit in unsupervised learning, but feedback from external system was used in the supervised manner. Querlioz et al proposed another application of supervised learning to improve network performance [18]. In conventional networks trained by unsupervised learning, results presented by neurons cannot be distinguished. An additional labeling process is needed to identify the results and supervised learning can play this role on the next layer. Figure 1.3 shows the architecture combining unsupervised and supervised crossbar and figure 1.4 shows the recognition rate of unsupervised / supervised layer. However, since unsupervised learning requires peripheral circuits, it may put a burden on area and power.

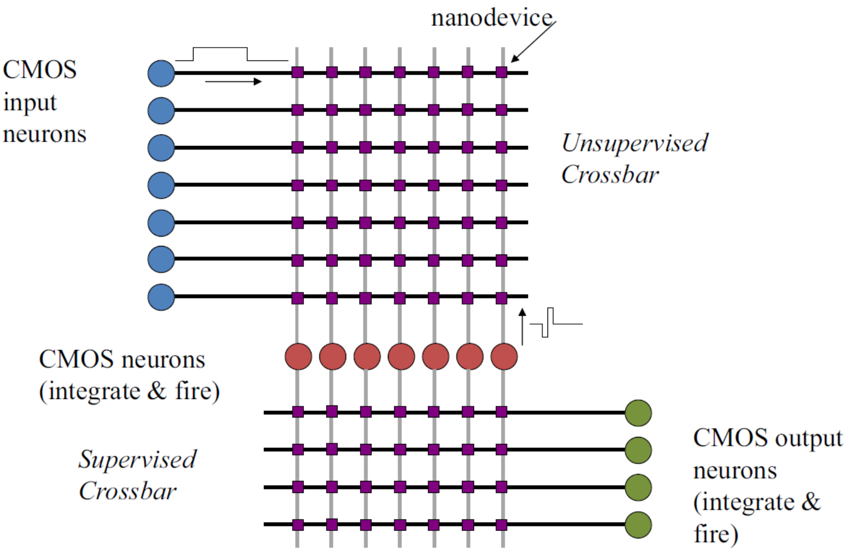

nanodevice
CMOS input neurons
Unsupervised Crossbar
CMOS neurons (integrate & fire)
Supervised Crossbar
CMOS output neurons (integrate & fire)

Figure 1.3. Architecture combining unsupervised and supervised crossbar [18].

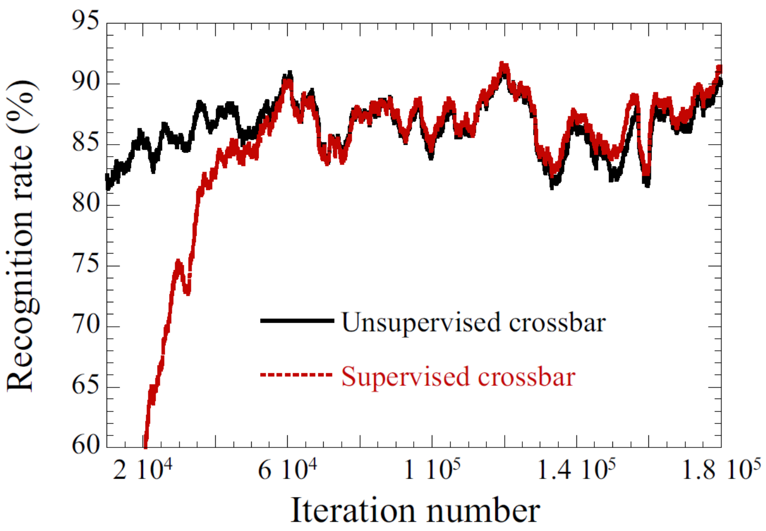


Figure 1.4. Result of recognition rate with two layer [18].

1.2.2 Unsupervised learning

STDP-based unsupervised learning can be applied efficiently to distinguish unlabeled data or unstructured data, which is advantageous for real-time data processing [19]. Diehl et al proposed a biologically plausible unsupervised learning mechanism that included lateral inhibition and adaptive threshold [20]. Using this SNN based on the STDP algorithm, the 95% classification performance of MNIST dataset was demonstrated in two-layer system with 6400 postsynaptic neurons. Although remarkable classification results of unsupervised learning based on STDP have been carried out, demonstration require additional circuits to fine-tune model parameters that are not suitable for processing various types of data. Querlioz et al introduced simplified STDP rule for pattern learning in an unsupervised manner [21, 22]. In these works, simplified STDP scheme by overlapping pre- and post-synaptic signal using simple pulse generation was used. To check SNN robustness, the same group examined the effects of device variability, including memristive synaptic device and CMOS neuron variability, along with system-level simulations on SNN [21]. Improved immunity to device variation is the result of neurons' homeostasis.

Figure 1.5 shows the impact of the neurons’ threshold X_{th} variability, with and without homeostasis on the recognition rate. This biologically plausible property, along with the WTA topology of lateral inhibition, plays an important role in regulating the response of neurons equally to prevent lower threshold neurons from being fired mainly in the network. Several different input encoding methods, learning methods and system structures were presented to enhance SNN's performance. Ambrogio et al proposed an input pulse scheme that uses input noise to suppress background synaptic devices [23]. Figure 1.6 shows the schematic illustration of the neuromorphic network with a 1T1R synapse. This configuration led to increased pattern synaptic device and the depression of background synaptic device for selective learning in an unsupervised manner. However, this configuration requires additional circuits for generating random noise, and it is difficult to optimize the input parameters of noise to handle different types of data.

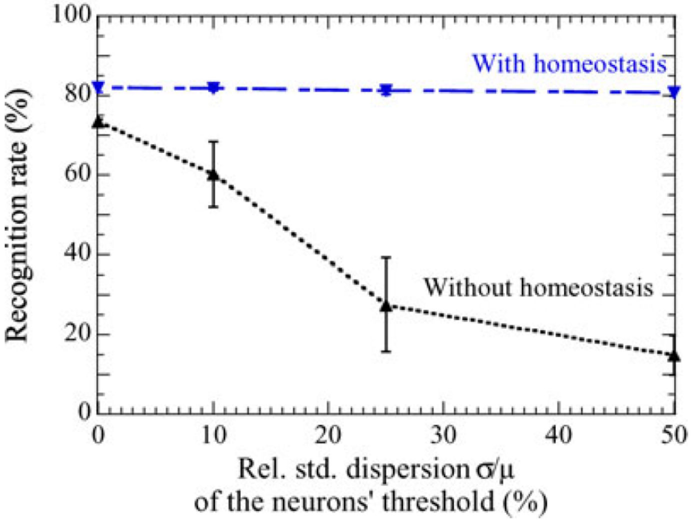


Figure 1.5. Impact of the neurons' threshold X_{th} variability, with and without homeostasis on the recognition rate [21].

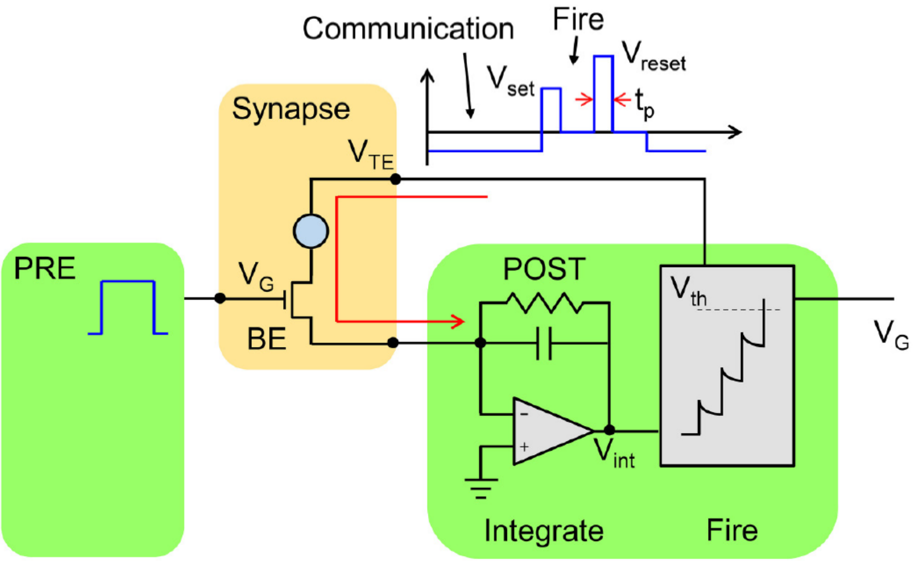


Figure 1.6. Schematic illustration of the neuromorphic network with 1T1R synapse [23].

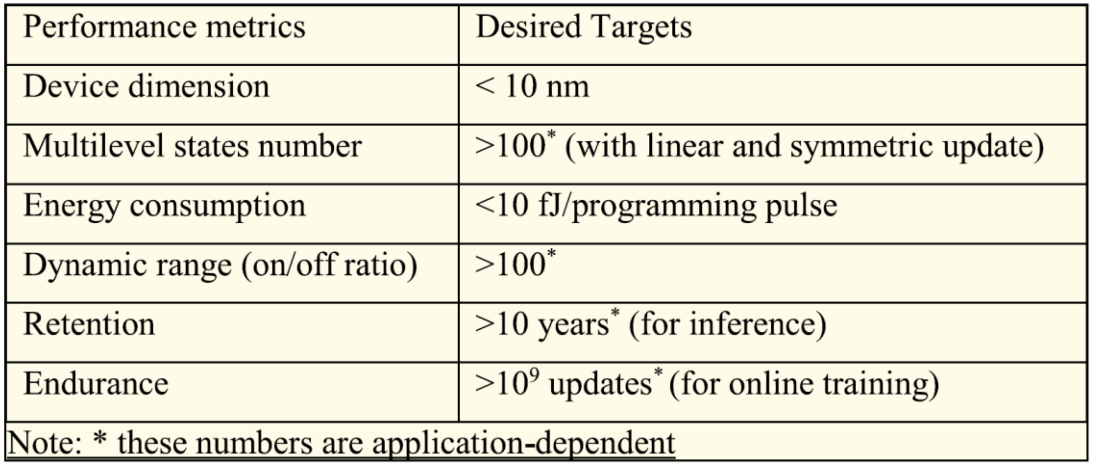
1.2.3 Requirements of synaptic device

Devices used for bio-inspired learning include resistive memory (RRAM), conductive-bridge memory (CBRAM), phase change memory (PCM), spin-based memory, and FET-based memory. By default, device array density is required to perform complex, large-scale tasks. In general, most research groups use crossbar arrays by default to build large-scale parallel computing neural networks. Although the two-terminal devices draw a lot of attention with the ease of implementation of the crossbar array. In fact, the two-terminal device requires a choice to eliminate the sneak path that occurs in the crossbar array configuration. Furthermore, the goal is not to implement an array dedicated to synaptic devices, but to implement large-scale neural network systems, so the CMOS compatibility of synaptic device is important. Therefore, synaptic device must be compatible with CMOS technology for system implementation.

The energy efficiency of the weight learning and inference process of a synaptic device array should also be carefully considered and evaluated differently depending on the application. For synaptic device array used in applications that perform continuous learning in real time, it is important to reduce the power used to update weight. In the case of synapses, which are used primarily in the inference process, power should be reduced in weighted sum operation.

The purpose of the neuromorphic synaptic array is to effectively combine the multiplication results of the input signal with the weights of the memory devices having the analog weight. A number of studies have been conducted to implement the analog memory characteristics [24]. In [24], Yu summarized the desired performance metrics for synaptic devices as shown in Table 1.2. If the unique characteristics of the device make it difficult to achieve gradual conductance changes, the gradual change can be implemented by controlling the pulse shape and adding additional devices (resistors or FETs) [25]. The difference in conductance between a high conductance state and a low conductance state can have a significant impact on the performance of the neural network. However, this is closely related to the size of the neural network according to the application used. It should be noted that the upper limit of the conductance value may increase the overall system power consumption dramatically. In implementing gradual conductivity, many research groups have mainly analyzed the linearity and symmetry of conductance change. In [26], Kim et al reported that the nonlinearity in the conductance change of the synaptic device is not critical to the pattern recognition rate of the system, as shown in figure 1.7. Using STDP and SRDP algorithms, widely used in bio-inspired learning, it is more important to avoid abrupt depression. For supervised learning involving external interventions, the side effects of abrupt depression can be slightly mitigated, but for unsupervised learning without external control, if the conductance of the synaptic devices drops dramatically, the learned synaptic weights may disappear in a moment. Efforts are needed to improve device structure and conductive change mechanisms to prevent abrupt depression. Neural networks are generally known to have some tolerance for device variation [21, 27]. However, the variation between synaptic devices should be as small as possible because it negatively affect the power consumption and speed of the learning process. Recently, research have been carried out on the HNN implementation using proven flash memory technology due to the immaturity of new memory technology. [17, 28, 29].

Table 1.2. Summary of the desirable performance metrics for synaptic devices [24].



Performance metrics	Desired Targets
Device dimension	< 10 nm
Multilevel states number	>100* (with linear and symmetric update)
Energy consumption	<10 fJ/programming pulse
Dynamic range (on/off ratio)	>100*
Retention	>10 years* (for inference)
Endurance	>10^9 updates* (for online training)

Note: * these numbers are application-dependent

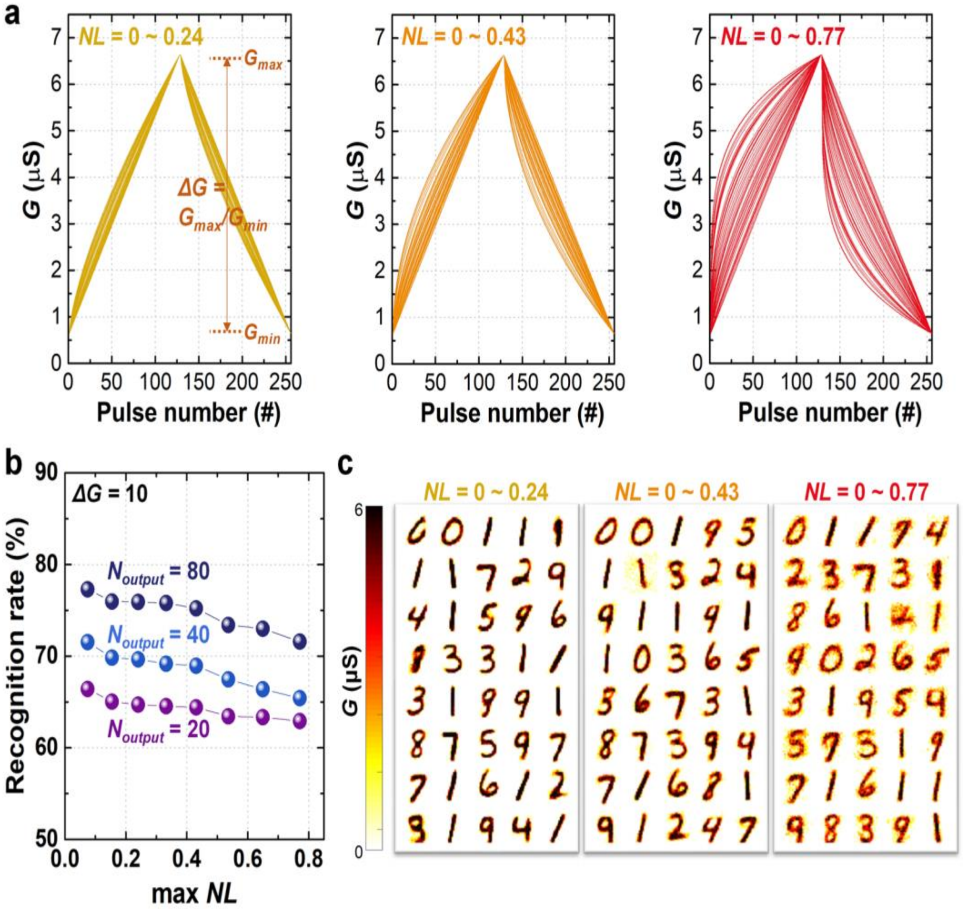


Figure 1.7. (a) The synaptic device conductance (*G*) as a function of applied pulse number with randomly assigned NL values. (b) The simulated recognition rate as a function of maximum NL value after 60000 times of training epochs. (c) The synaptic weights between the input to output neurons with 40 output neurons, when NL ranges are 0 ~ 0.24 and 0 ~ 0.77 [26].

1.3 Conventional technologies

Recently, studies on the neuromorphic system have been conducted in an effort to overcome the limitations of the von Neumann computing system [7]. Given that the existing von Neumann architecture is very limited in terms of speed and power consumption for high-level recognition applications, neuromorphic technology research and development have been active area for solving these problems [8]. In the software field, studies of the deep DNN using BP algorithm [11] were emphasized as an excellent cognitive capability, and efforts were made to apply the results to HNN. Another aspect of implementing such HNN is the use of STDP algorithm, one of several learning algorithms that mimics the behavior of the biological brain [6]. So far there have been many reports on pattern recognition systems that work through supervised learning based on DNN [11] [30]. However, there are many applications in which brain-inspired, unsupervised learning can also be used in actual machine learning [13].

To implement HNN using the STDP algorithm, it is important to replicate long-term potentiation (LTP) / long-term depression (LTD) functions as electrical elements in accordance with the spike firing sequence. Many studies have attempted to reproduce electronic synaptic devices and synaptic plasticity through the CMOS VLSI circuits [31] [32]. In recent years, studies have been actively conducted on the composition of synapse array using the memristor crossbar array. [23], [33]-[36]. Figure 1.8 shows a neural network composed of memristor crossbar array. However, memristors still have disadvantages in device characteristics fluctuations and reliability when configured with large-scale crossbar arrays [37], [38]. Fluctuation in device characteristics in memristors causes a decrease in recognition rate in the pattern recognition process in a real artificial neural network [22]. Research on electrical synapses based on CMOS Field Effect Transistor (FET) were recently carried out. As a result of these efforts, a number of devices have been introduced, including NOMFET [39] and MemFlash [40] [41]. Figure 1.9 shows the schematic drawing of a two-terminal MemFlash circuitry. However, in one of these studies [39], metal nanoparticles are used in memory functions, causing compatibility problems with the CMOS process. In other study [40], it remains unclear how it is used to create a large synaptic array and how it works.

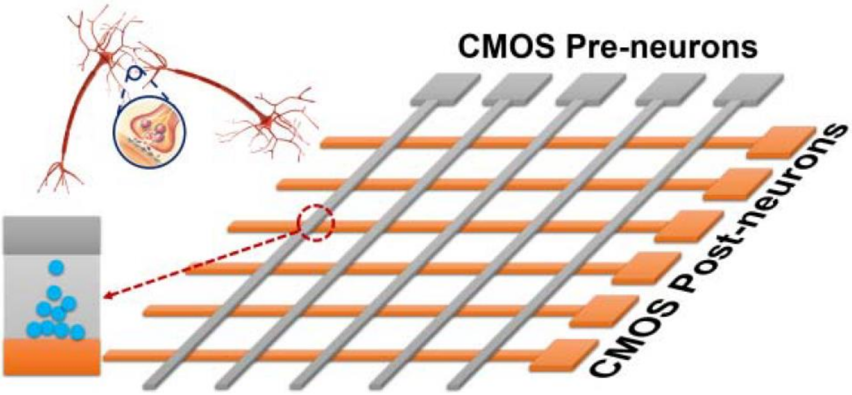


Figure 1.8. Neural network composed of CMOS neurons and HfO_x-based electronic synapses [36].

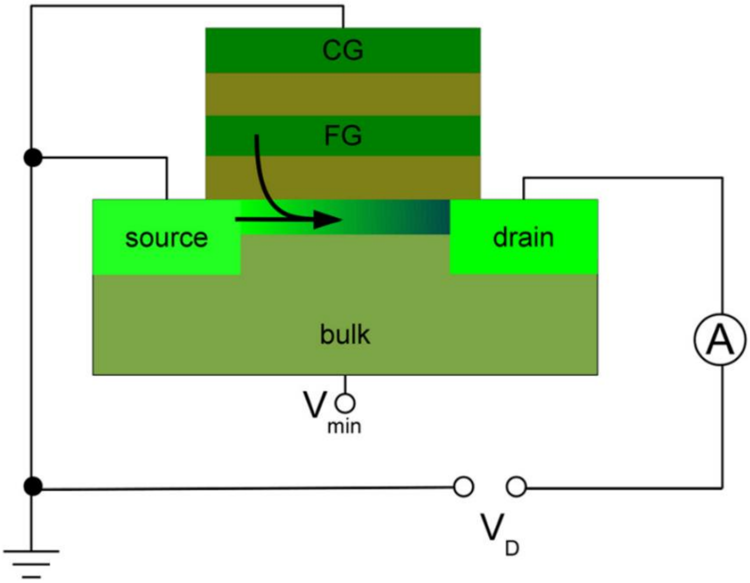


Figure 1.9. Schematic drawing of a two-terminal circuitry (MemFlash) [41].

1.4 Purpose of research

As described above, the need for synaptic device to implement a neural network based on bio-inspired algorithm is emerging. In this work, we propose NOR-type nonvolatile memory arrays to demonstrate unsupervised learning with STDP. First, we investigated SONOS gated-diode that bit-line current of a cell string can be trimmed accurately by controlling the stored charge in each cell. Moreover, we suggest an approach to use the gated-diode memory as a synapse-like neuromorphic hardware. Afterwards, we analyze and review issues that arise when we try to implement a neural network using the gated-diode memory array. To overcome these problems, a new TFT-type NOR flash memory array is proposed. We fabricate a TFT-type NOR flash memory array using the conventional CMOS fabrication process and suggest an approach to use TFT-type NOR flash memory as synaptic device. We also looked at the advantages of this device as synaptic device. Then, we report simulation results of unsupervised learning using STDP in our TFT-type NOR flash memory array. Main point in this work is to demonstrate unsupervised learning with STDP using NOR-type nonvolatile memory array.

1.5 Dissertation outline

This dissertation is composed as follows. Chapter 1 contains an overview of the neuromorphic technology and STDP algorithms. Then, current research trends of synaptic devices are described. The purpose of research and the outline of dissertation are also presented. Chapter 2 describes the structure, characteristics, and pattern learning simulation results of the SONOS gated-diode memory synaptic devices. After that, the discussion of the issues that arise when using the memory array for the HNN system. In Chapter 3, the structure, fabrication process, and measurement results of the TFT-type NOR flash memory are presented. Then, circuit / pulse schemes to utilize the device array for the neural network are presented. The last part of this chapter analyzes the pattern recognition performance and suggest ways to improve it. Finally, the conclusion is delivered in Chapter 4.

The main content of this dissertation has already been published in referred journals [9], [28], and [42]. With the publishers' permission to reuse the article in this dissertation, the major parts of the present thesis were reproduced from [9], [28], and [42] following the publishers' guidelines, respectively.

Chapter 2

SONOS gated-diode memory array

2.1 Device structure

Figure 2.1 shows a 3-D schematic view of a single SONOS gated-diode memory and figure 2.2 shows a 3-D schematic view of SONOS gated-diode memory array consisting of multiple word-lines (WLs) and bit-lines (BLs), respectively. The *n*+ region forms in the upper area of the *p*-type silicone fin and is connected to the BL via MOSFET as a select device of a cell strings. The gate dielectric stack covers the top and both sides of the fin where the *n*+-*p* junction is formed. Here, the thicknesses of the ONO stacks are 3 nm, 6 nm and 9 nm. The gate electrode is formed on the gate stack.

In this operation, the configuration of the diode-type cell string is significantly different from that of the existing FET-type NAND flash memory cell string. The *n*+ diffusion area formed in the upper area, as shown in figure 2.1, proceeds along the fin body. Therefore, when reading the cells selected in this operation, all WLs, except the cells selected in the diode-type cell string, may be floated or bias to a low voltage depending on the resistance of the *n*+ area. However, the *n*+ region of the FET-type cell string is typically formed only in the fin space between adjacent WLs, not in the channel. So, all WLs, except the cells (pass cells) selected in the cell string, must be biased to a large bias (>5 V) that turn on the pass cells on completely. Under certain bias conditions, gate-induced drain leakage (GIDL) current is generated near the *n*+ area surface and becomes BL current (I_{BL}). The GIDL current flows from the BL to the *p*-type area of the *n*+-*p* diode. The *p*-type area of the diode type cell string is connected to each other between the cells. GIDL current in all cells can be added to the BL current. This is equivalent to the total current added by all memristors connected to the neuron network circuit [43]. The detailed fabrication process of a device and a device array can be found in previous work [44].

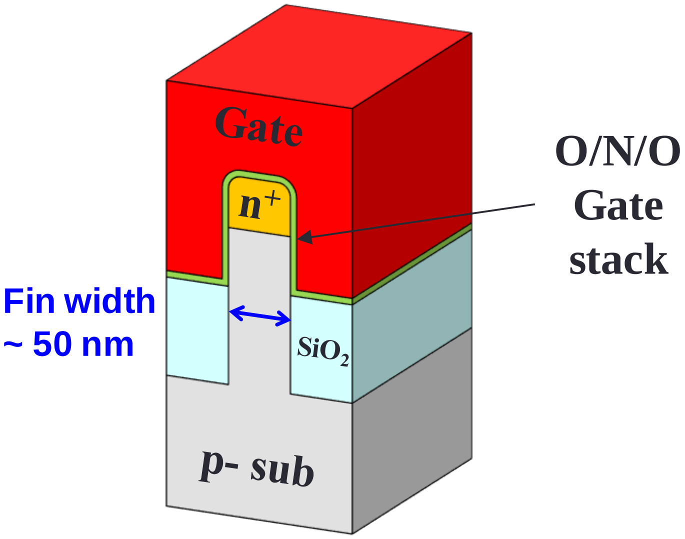


Figure 2.1. 3-D schematic view of a single SONOS gated-diode memory.

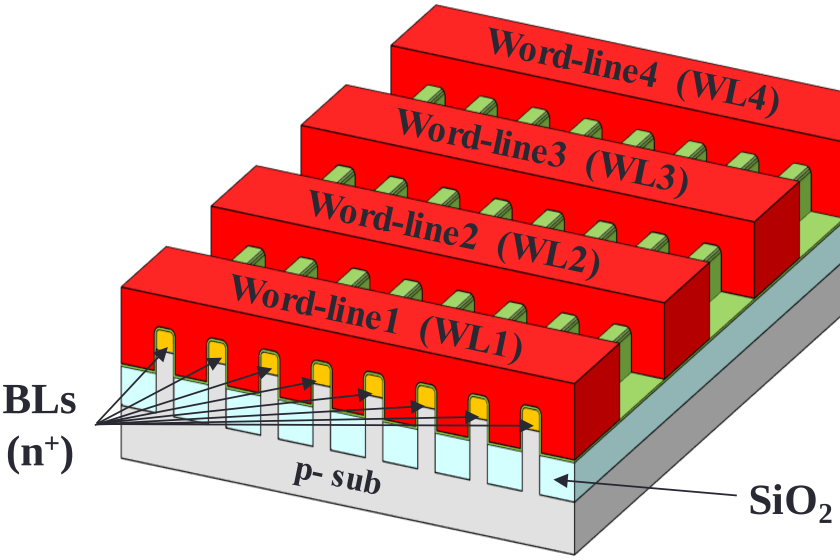


Figure 2.2. 3-D schematic view of SONOS gated-diode memory array consisting of multiple word-lines (WLs) and bit-lines (BLs)

2.2 Device characteristics

In this chapter, we examine the electrical characteristics of the SONOS gated-diode cell. Figure 2.3 shows the I_{BL} (BL current)-V_{WL} (WL bias) curves as a parameter of the V_{BL} (BL bias) [42]. As V_{WL} increases, the I_{BL} increases significantly, due to the increasing the GIDL current. At a negative V_{WL}, the energy band near the surface of the *n*+ region is bent up and the electron/hole pairs are created through a band-to-band tunnel. The electrons flow through the BL and the holes flow into the *p*-region. Therefore, the GIDL current flows from the BL to the *p*-region. As the V_{BL} increases, the band bending increases, significantly increasing the I_{BL}.

Programming (PGM) in the gate-diode memory is performed by FN tunneling. After the programming operation, the I_{BL} read from the programmed cell increases because of the increased band-to-band tunneling (BTBT) caused by electrons stored in the nitride. On the other hand, erasing (ERS) is performed by injecting BTBT hot holes as a result of GIDL generation. The band structures for these PGM/ERS states are shown in figure 2.4 (a), (b), respectively.

Therefore, I_{BL} reading from erased cells decreases due to holes stored in the nitride layer. The I_{BL} detected in the programmed cell is approximately 10^3 times greater than the I_{BL} value of the cell erased under the read bias conditions shown in Table 2.1. Therefore, memory performance can be achieved by detecting the difference in current between these two states. Figure 2.5 shows I_{BL}-V_{WL} curves as the state of the charge storage layer (PGM or ERS).

The I_{BL} can also be increased incrementally by incremental step pulse programming (ISPP) as shown in figure 2.6 [44] [45]. In this scheme, The gate voltage of the program pulse V_{pp} can be increased to a constant value after each program phase. As PGM voltage (V_{pp}) increases, the number of trapped electrons increases, increasing cell current. In figure 2.6, the circle symbol represents the I_{BL}-V_{WL} characteristic after the cell has been programmed to a V_{pp} value of 15 V, where the hexagon symbol shows I_{BL}-V_{WL} curve after the cell is programmed at V_{pp} value of 21 V. Here, between the two curves, the pulse step ΔV_{pp} is 1V.

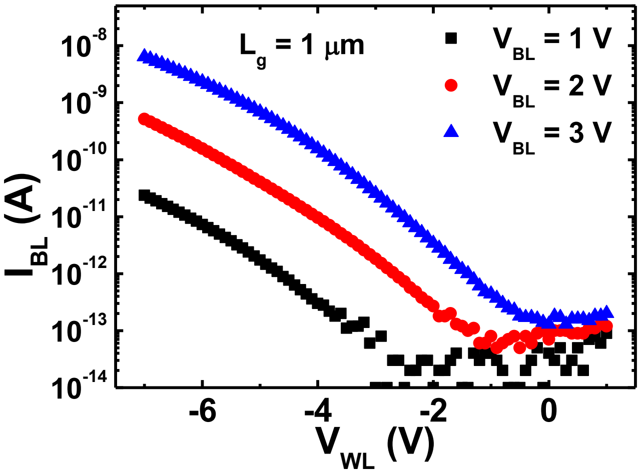


Figure 2.3. Bit-line (BL) current versus word-line (WL) bias for a gated-diode memory cell as a parameter of V_{BL} [42].

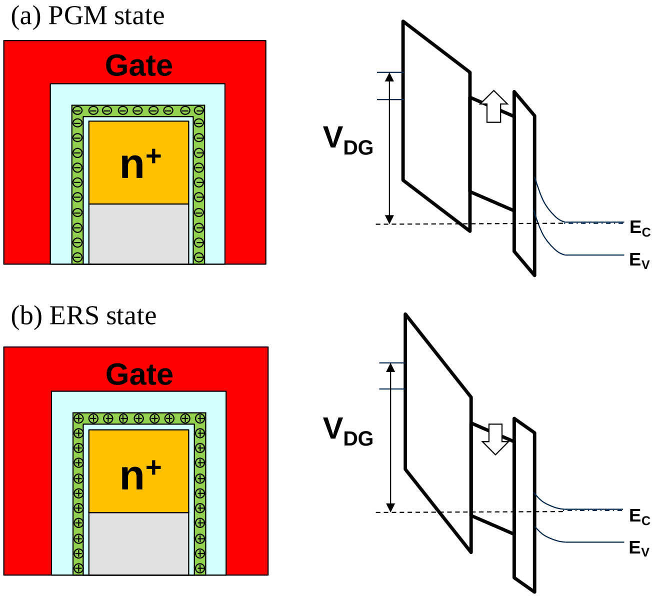


Figure 2.4. Schematic view, which shows the (a) programmed (PGM) and (b) erased (ERS) state with the stored charge at nitride charge trapping layer [44].

Table 2.1. Bias conditions for the programming (PGM) and erasing (ERS) of cells in a cell string [42].

		PGM (FN tunneling)	ERS (BTBT HH)	Read
V_{WL}		20V	-9V	-6V
V_{BL}	Selected	0V	6V	2V
	Unselected	6V	0V	
Time		100μs	10ms	

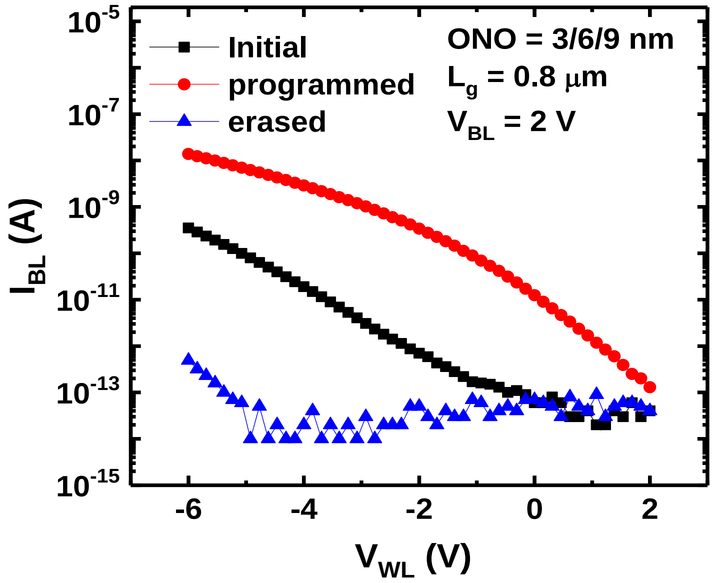


Figure 2.5. I_{BL}-V_{WL} curves as the state of the charge storage layer (PGM or ERS).

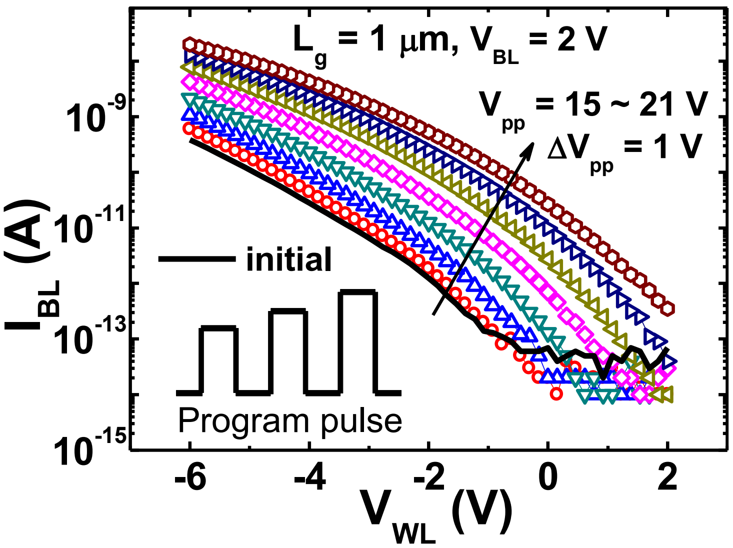


Figure 2.6. Incremental step pulse programming (ISPP) characteristic of the SONOS gated-diode memory cell [42] [44].

2.3 Device measurement results as a synaptic device

2.3.1 Implementation of neural network

Figure 2.7 shows the neural network topology for STDP unsupervised learning [46]. The main purpose of the neural network is to learn and recognize the unlabeled binary MNIST handwritten dataset. Each number pattern is input to the neural network via 784 (28 × 28) input neurons (PRE), which will be passed to the second neuron layer (POST). Each POST output neuron is linked to each other via inhibitory synapse.

3-D conceptual diagram of the synaptic array using SONOS gated-diode memory is shown in figure 2.8. Input signals from PRE neuron are transmitted by the WLs of the memory array and the weighted sum results of synaptic devices are combined into BLs and delivered to POST neuron.

Figure 2.9 shows the schematic circuit diagram of the neural network when implemented on a SONOS gated-diode memory array. The pattern transmitted from the input neuron is input through the gate of each single gated-diode device. The input signal to the gate is converted into a current reflecting the weight stored in the synapse, and is added to the BL of the array. The combined current through the BL is connected to the POST neuron circuitry outside the array, causing the neuron to fire. Each POST neuron is connected via FET type synapse that performs inhibitory action, thereby suppressing the action of neurons other than oneself.

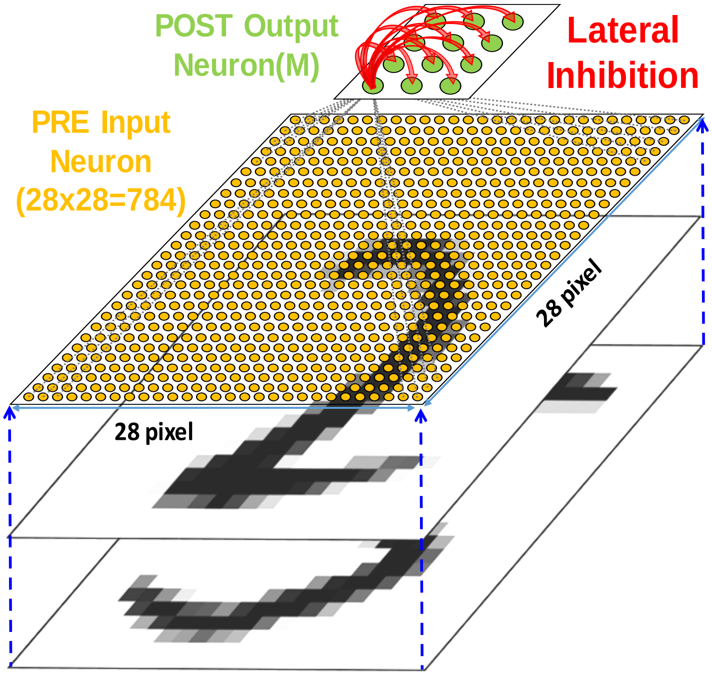


Figure 2.7. Topology of spiking neural network (SNN) using unlabeled binary MNIST handwritten patterns [46].

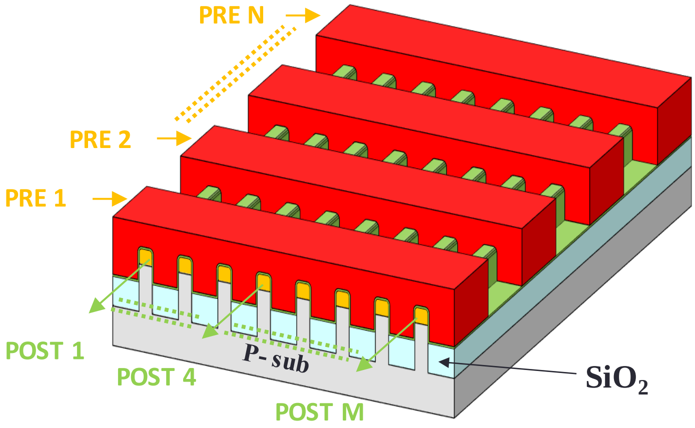


Figure 2.8. 3-D conceptual diagram of the synaptic array using SONOS gated-diode memory array.

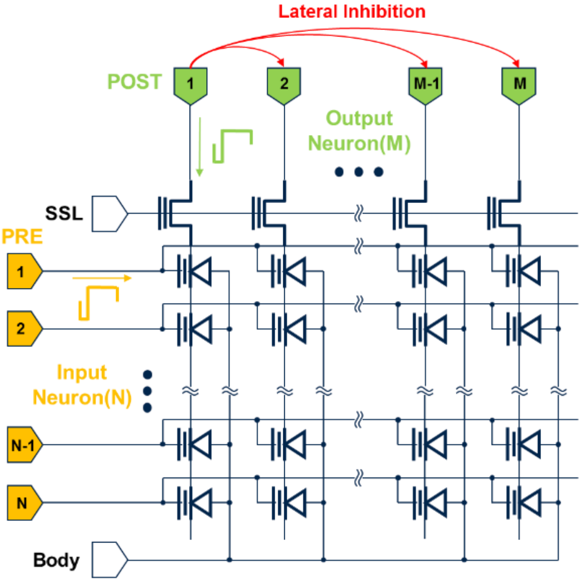
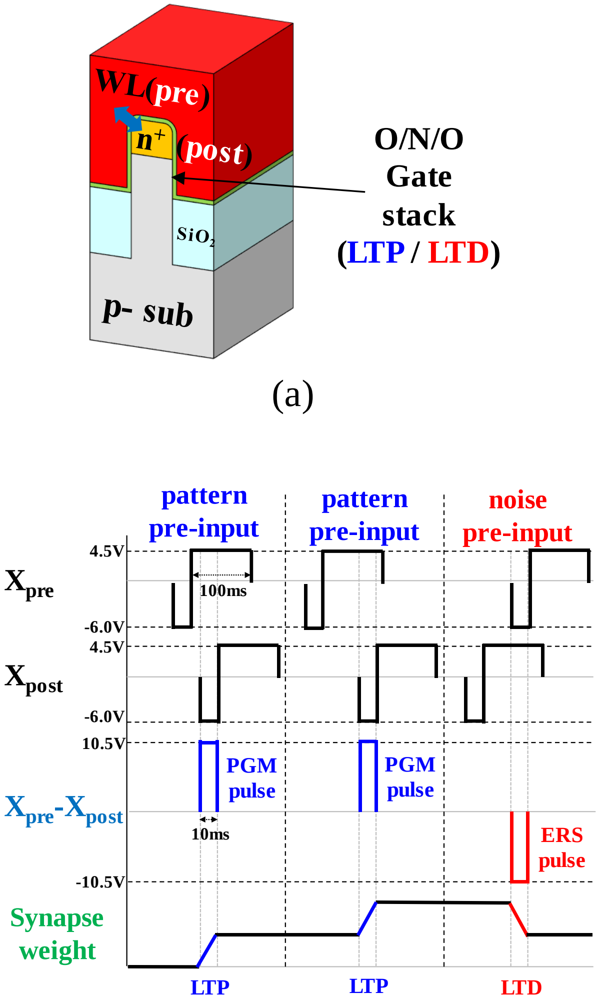


Figure 2.9. Schematic circuit diagram of the neural network for STDP unsupervised learning when implemented on a SONOS gated-diode memory array.

2.3.2 Pulse scheme for STDP weight update

For STDP operation, the synapse cells can be potentiated or depressed selectively using the pulse scheme shown in figure 2.10. As shown in figure 2.10 (a), input signal from PRE neuron and POST feedback signals through an integrated fire circuit are applied to the WL and BL respectively to vary the weight of the synapse cell.

The program and erase operations of charges stored in the nitride storage trapping layer depend on the WL and BL voltage states. When a PRE input pulse is applied and the POST neuron is fired, the tail part of the input pulse overlaps with the head portion of the feedback pulse. Then, a pulse of 10.5 V magnitude and a time of 10 ms is applied to the WL reference to perform program operation at the nitride layer, which simulates synapse LTP operation. On the other hand, if the PRE input signal is applied after POST neuron firing, the tail portion of the feedback pulse overlaps the head portion of the input pulse. Then a pulse having a magnitude of -10.5 V and a time of 10 ms is applied to the WL basis, which causes erase operation in the nitride layer, resulting in the same result as the synapse LTD operation. Using this pulse scheme, the desired pattern image can be trained on the synaptic array by repeatedly inputting the target MNIST image and noise image [23]. If noise input is transmitted after the pattern image, the weights of synapse cells overlapping noise pattern are depressed to increase the weight difference between the learned synapse cell and the un-learned synapse cell. Therefore, by repeating the cross-entry, each synapse string can be trained with the desired pattern. Table 2.2 summarizes the pulse scheme for these weight updates for STDP operation and the pulse scheme for the weight read operation. The read pulses for reading and summing the weights of the synapses are applied to the WL for -6 V and the BL for 2 V magnitude and 10 ms time, respectively.



(a)

(b)

Figure 2.10. (a) Schematic diagram of PRE (input) and POST (feedback) pulses that cause weight updates of SONOS gated-diode synaptic device. (b) Pulse scheme of PRE and POST neurons to the SONOS gated-diode synaptic array that causes a LTP and LTD.

Table 2.2. Bias conditions for the weight update (LTP/LTD) and weight read operation of cells in a SONOS gated-diode synapse array.

	LTP (pattern pre-input)	**LTD (noise pre-input)**	**Weight read**
V_{WL}	4.5 V	-6.0 V	-6 V
V_{BL}	-6.0 V	4.5 V	2 V
V_{B}	0 V	0 V	0 V
Time	10ms	10ms	10ms

2.3.3 LTP/LTD characteristics

The pulse scheme for the synapse weight update described above is applied to actual devices and the results are shown in figure 2.11. Sequential 20 repetitive LTP pulses followed by 20 repetitive LTD pulses are applied to the WL and BL of the synaptic device, and then the weight of the device is read. In this case, the pulse scheme used in each operation is the same as the scheme in Table 2.2. The repeated increase of the synapse weight is confirmed by the repeated application of the LTP pulse, which is dependent on the amount of charge stored in the nitride storage layer. Therefore, when the LTD pulse is applied to fully potentiated cells, the synapse weight rapidly decreases. Figure 2.12 shows the STDP curves derived when the pulse scheme is used. Because of the use of rectangular pulse instead of increasing or decreasing pulse over time, simplified STDP curve is derived.

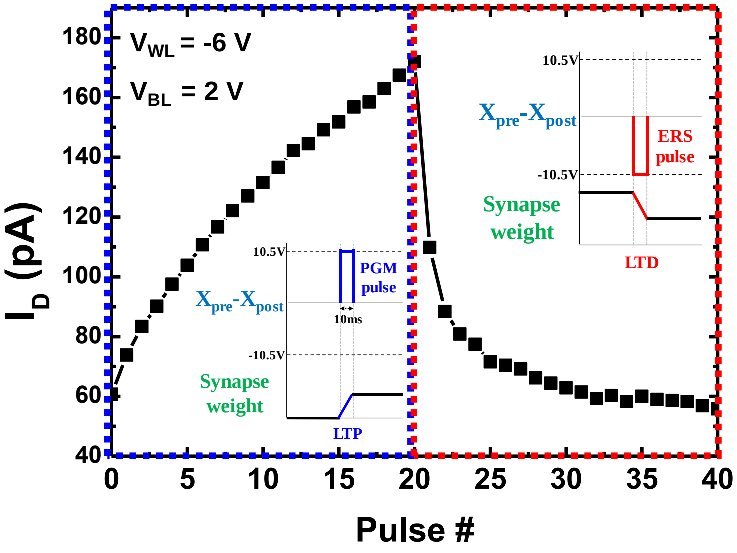


Figure 2.11. The LTP/LTD repetition characteristics of the SONOS gated-diode synaptic device measured using the pulse scheme of figure 2.10.

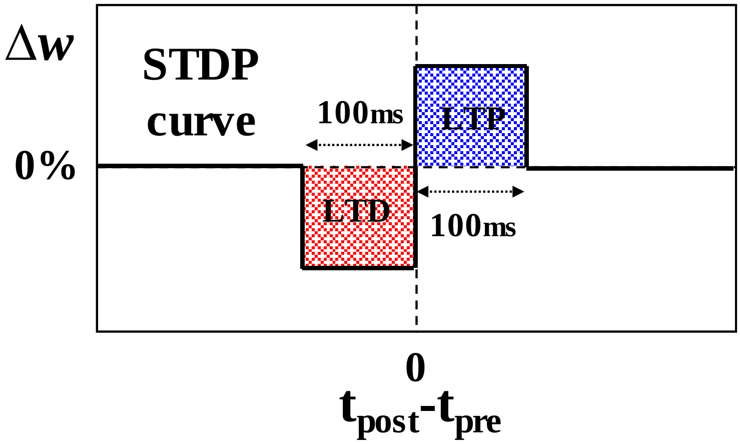


Figure 2.12. STDP curve derived from the pulse scheme of figure 2.10.

2.4 Simulation results of pattern learning

2.4.1 Single-neuron learning

Figure 2.13 shows a flowchart of the overall pattern learning process used in the simulation. The simulation was performed with software MATLAB, and the operating characteristics of synapse reflected the measured characteristics of the SONOS gated-diode memory cell. Output neurons are assumed to be ideal capacitors and comparators. First, reset the synapses by randomizing the weights of all synapses. In the PRE target image for learning, only the part where the input value of a pixel exists in that image triggers the X_{pre} pulse to WL of figure 2.10 (a), which may result in firing of the POST neuron via I&F circuit. This postsynaptic spike is sent to other neurons and inhibits charges accumulated by the neuron's integrate capacitor. This process allows each neuron to learn its own image pattern to implement pattern classification. The neurons also transmit the feedback spike to the BLs of the synapses cells connected to it, which immediately update the synapse weights. Figure 2.14 shows ten 28 × 28 MNIST handwritten target digits and noise pattern used in this learning simulation.

To identify the learning and updating capabilities of the synaptic devices in single-neuron, patterns represent the numbers "2" and "5" were used sequentially. These input patterns are input repeatedly with or without noise pattern for STDP learning. The initial synaptic weights are randomly distributed between the minimum to maximum weights of the SONOS gated-diode memory cells. Figure 2.15 (a) shows the pattern learning process without noise input pattern in a single-neuron containing 784 (28 × 28) synaptic devices (i.e. 784 × 1). The change in the weight map of synapse array is shown when each image is entered sequentially 30 times. As mentioned previously, in the proposed pulse scheme, it appears that the pattern does not update to another pattern because LTD of synapse weights is not possible when noise input is not used. Similarly, the non-pattern background area is also checked that depression does not occur. On the other hand, the result of single-neuron pattern learning with noise input pattern is shown in figure 2.15 (b). The synaptic weights of the array have been correctly updated based on the STDP behavior when each image is displayed 30 times sequentially. The weight learning for pattern “2” is completed through 30 epochs, and after 30 consecutive epochs, learning of the weights corresponding to pattern “5” has been achieved. This result indicates that the desired input pattern and pattern update have been performed successfully.

In the single-neuron learning, weight margin (W_{margin}) between targeted synapse and background synapse has an important influence on pattern recognition rate. This is because the weight margin must be guaranteed above a certain level to distinguish the desired pattern from the other pattern in the recognition process. Single-pattern learning in this single-neuron can be done well without any additional input. However, in case of changing the learned pattern in single-neuron or multi-pattern learning in multi-neuron, noise input patterns are needed. These noise input patterns depress the weights of the background synapses that are not the desired number pattern. These noise input patterns are input through a certain number of synapses randomly selected from total of 784 PRE synapses. The noise density (ρ_{noise}) is determined by the number of randomly selected input synapses. Figure 2.16 shows the influence of such input noise density on pattern learning. Pattern learning requires a certain density of noise. If this noise density is too high, the recognition rate of the pattern is degraded, and the learning reversal phenomenon in which the noise pattern is learned occurs. Therefore, the noise input pattern density used for learning should be carefully considered to optimize learning efficiency.

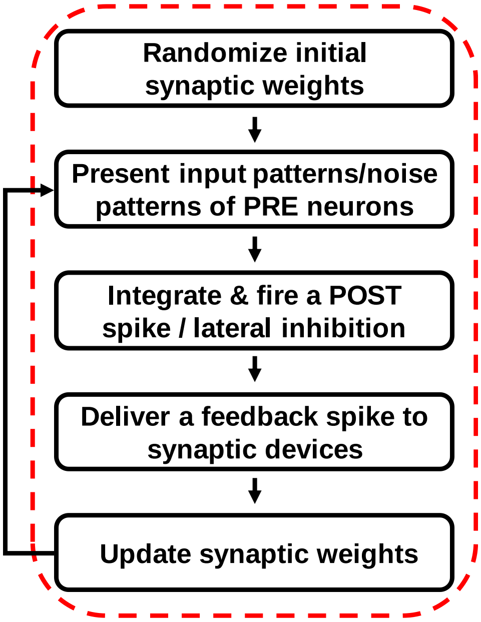

Randomize initial synaptic weights
Present input patterns/noise patterns of PRE neurons
Integrate & fire a POST spike / lateral inhibition
Deliver a feedback spike to synaptic devices
Update synaptic weights

Figure 2.13. Flowchart of the overall pattern learning process used in the simulation.

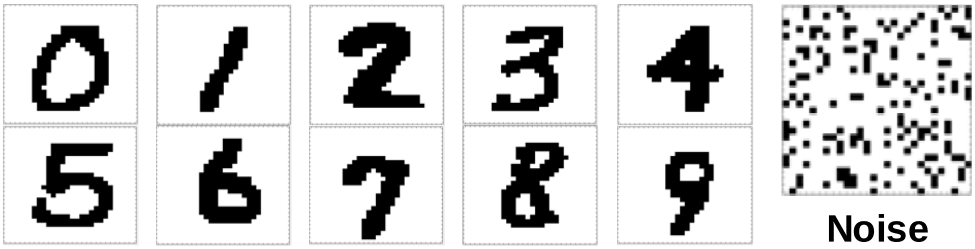


Figure 2.14. The ten 28 × 28 MNIST handwritten target digits and noise pattern used in the learning simulation.

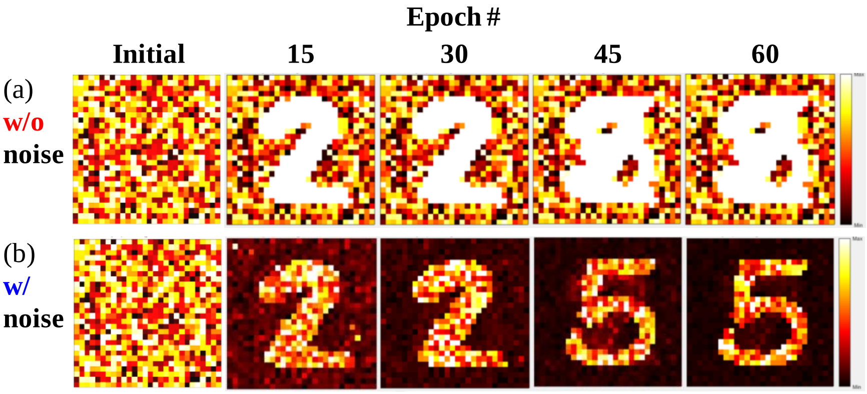


Figure 2.15. Unsupervised pattern learning and updating results (a) without input noise pattern, and (b) with input noise pattern in a single neuron when the first pattern “2” and the second pattern “5” in figure 2.14 were learned 30 times in order.

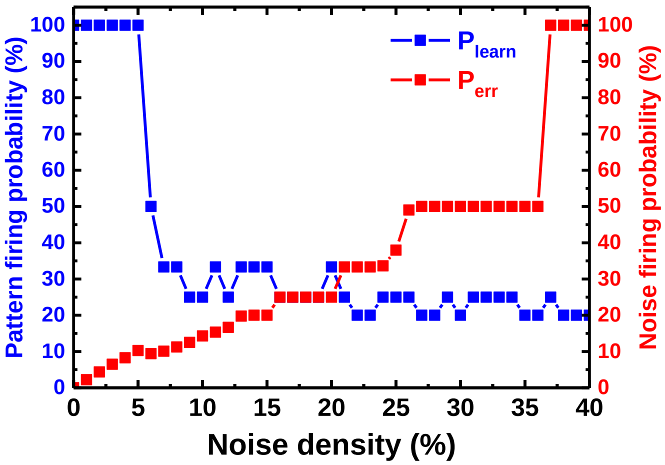


Figure 2.16. Recognizing probability of input and noise pattern as a function of input noise density.

2.4.2 Multi-neuron learning

Further simulation works were conducted by increasing the number of POST output neurons for multi-neuron learning while maintaining the method of adding noise input pattern from the previous chapter. In multi-neuron learning, the main goal is to learn different patterns for each neuron. To identify these characteristics, simulations were performed by selecting the number "3", "6" and "7" from the MNIST handwritten dataset of figure 2.14. In order to check the learning progress in unsupervised manner, the input method was selected along with the noise input on each randomly selected number pattern. One example of this pattern input method can be found in figure 2.17. Figure 2.17 (a) shows an example of an input train in which random noise inputs are applied between and number patterns are entered in random order. Figure 2.17 (b) shows a sequence of input patterns to help understand the random repetitive input of image and noise patterns.

Figure 2.18 shows the pattern learning results for a multi-neuron array (784 × 3) composed of 784 PRE input neurons and three POST neurons. Figure 2.18 (a) shows the changing aspect of synaptic weights when repeatedly applying inputs of figure 2.17 to a multi-neuron array (784 × 3). As can be seen in the figure, each neuron does not learn its own patterns but progresses as if it were lumped together by three patterns. In order to solve this problem and make each neuron learn different patterns, we adopted the concept of lateral inhibition based on biological theory [47] [48]. In multi-neuron learning, an inhibitory synapses are used to lower the membrane potential of neurons other than the firing neurons themselves. An inhibitory factor, which determines how low the membrane potential of the other neurons, should be considered carefully. If the inhibitory factor is too high, only a small number of neurons will fire repeatedly to interfere the learning of other neurons, while if it is too low, it will be difficult to distinguish the neuron's own learning pattern. Figure 2.18 (b) shows the progress of multi-neuron learning when the lateral inhibition function is used. This shows the process of changing the weight states of the synapses corresponding to each neuron at representative epoch numbers. After a certain number of epochs, the weights of the synapses belonging to each neuron are gradually tuned according to a different pattern. These results confirm the possibility of multi-neuron learning.

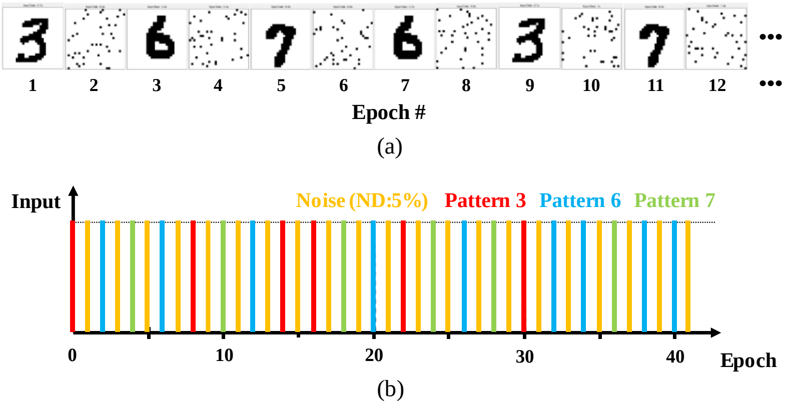


Figure 2.17. (a) An example of input train. (b) A sequence of input patterns of the random repetitive input of image and noise patterns.

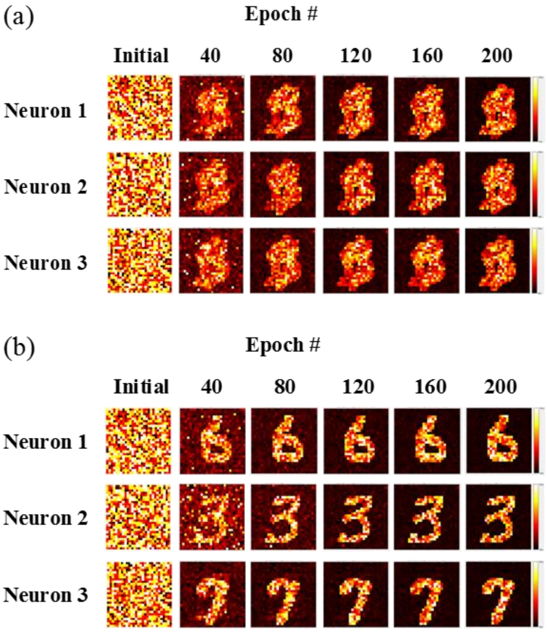


Figure 2.18. Synaptic weight change corresponding to each neuron when pattern train in figure 2.17 were randomly presented 200 times (a) without, and (b) with lateral inhibition function.

2.5 Issues

Previous chapters presented learning capabilities of the SONOS gated-diode memory array as a synaptic device array using the measured characteristics of the devices. However, several critical issues have been identified when trying to actually use these devices as synapses.

The first problem is the unwanted increase in power consumption due to feedback pulses occurring in the learning process. Figure 2.19 shows the previously proposed device structure and feedback pulse (X_{post}) shape. As shown in the figure, the feedback pulse is applied to the BL (n^+-region) of the device array, which contains both positive and negative bias. Therefore, when the feedback pulse is applied to the BL, there is a section where the n^+-p diode between the BL and the p-substrate is turned on, so a large amount of current flows through the learning process. Moreover, in a fabricated device array, p-substrates are tied together throughout the whole wafer, so it cannot be individually controlled during the synaptic weight learning process. This unwanted power consumption can be pointed out as a fatal problem, as it is repeated countless times during repetitive learning processes.

A second issue is the increase in standby power consumption caused by increased weights of synaptic devices. Figure 2.20 identifies this problem. Indicates when the learning is repeated to increase the weight of the synapses at read bias, in which case the leakage current of the synapse device at the standby voltage of 0 V can also be seen. This means that when learning is over, power consumption occurs even when there is no input from the system, due to the continuous increase in the weight of synapses with unsupervised manner. The increase in power consumption due to this leakage current is also a problem, but this quiescent current has become so large that it can be misjudged as a weight information from synapses in the recognition process that it may cause errors in recognition phase.

The next problem is that the PGM/ERS pulse used in learning phase is too long. Bias conditions used for the weigh update (LTP/LTD) are shown in table 2.2. Here, the width of PGM/ERS pulse for LTP/LTD is 10 ms. This level of weight update time is significantly higher than other non-volatile memory used as synaptic devices [49]. This could lead to serious delay in neuromorphic system with numerous weight updates.

The final issue is the use of noise input signal in the learning process. As previously described, this system uses noise pattern for depression of the background part and pattern update in the learning process. Generating a noise input pattern that has certain density across the whole pixel itself would be a huge burden on the circuit system, and applying it crosswise with image pattern is also disadvantage. The increase in learning time is also one of the side effects that cannot be ignored because the pattern and noise must be applied repeatedly.

So far, we have presented device analysis and learning simulation works to utilize the fabricated SONOS gated-diode memory array as the synaptic array in the neuromorphic system. Then, we followed up with discussions of the critical problems that arise when we actually try to use the device array for the neuromorphic system. In the subsequent chapters, we will propose a new neuromorphic device to overcome these problems and proceed with the fabrication and analysis of the device for learning/recognition task.

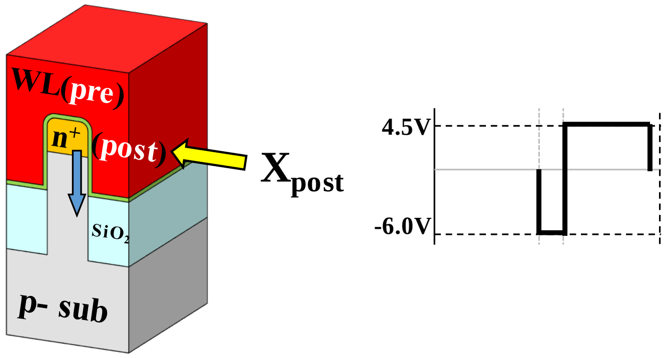


Figure 2.19. Diode current flow caused by pulse of X_{post} used for the weight update (LTP/LTD).

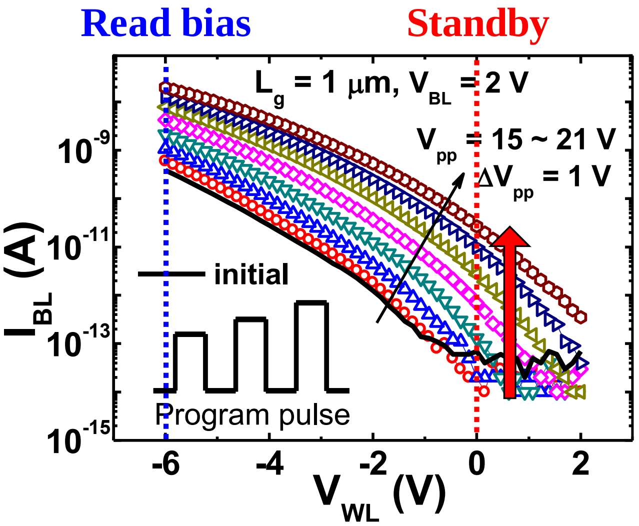


Figure 2.20. Standby power consumption increasing due to the repetitive potentiation of the synaptic weight.

Chapter 3
TFT-type NOR flash memory array

3.1 Device structure

Figure 3.1 shows the 3-D schematic view of the TFT-type NOR flash memory device. Cross-sectional views cut along the directions of A-A' and B-B' are shown in figure 3.2 (a) and (b), respectively. As shown in figure 3.1, each WL and BL (n^+ drain) intersect in the form of a crossbar to simplify scaling memory array on a large scale. In figure 3.2 (a), the drain and source of each unit device are connected via a poly-Si channel that is half covered by an n^+ poly-Si floating gate (FG) through an inter-poly dielectric (IPD) material. All devices connected to a single WL are controlled simultaneously through a single WL.

The TFT-type memory array also has a structure in which the *p*-substrate and the device are isolated by the SiO_2 insulation layer. This structure resolves leakage current issue arising from LTP/LTD operation, which was the first issue in the SONOS gate-diode synaptic array of previous chapter. A half-covered FG is located between the WL and the source, and if voltage is applied between the two electrodes, the PGM and ERS memory operations are performed. Because the FG covers only half the channel, the V_T does not drop below zero in the full ERS state of the synaptic device, which prevents leakage current during the system operation. This will solve the standby power-increasing problem caused by the excessive LTP process of the synaptic device when using unsupervised learning method, which was pointed out as the second issue in the SONOS gated-diode synaptic array.

Figure 3.2 (b) shows the structure of the device in the direction of BL. The FGs of adjacent devices are isolated from each other and configured to perform their own memory operations. However, the placement of source and drain running side by side in the BL direction is common between n memory cells under n WLs to allow current sum from n NOR flash memory cells. Therefore, each memory cell can send its own memory information to the common BL in the form of summed current. This is similar to the configuration of biological synapses, each of which reflects its weight information and combines it with signal sent to the next neuron.

Figure 3.3 shows bird’s eye view of a TFT-type NOR flash memory array.

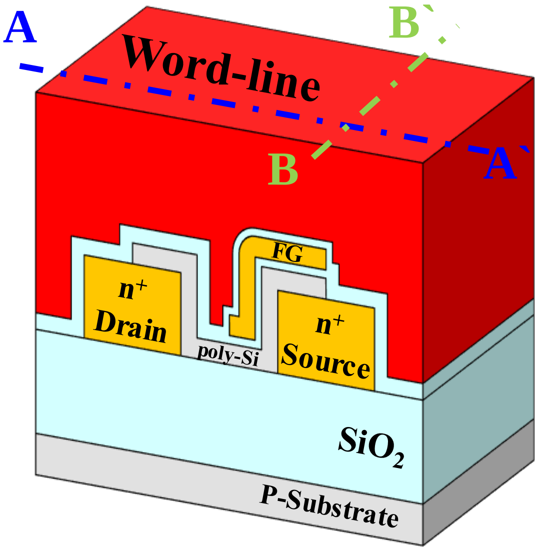


Figure 3.1. The 3-D schematic view of the TFT-type NOR flash memory device

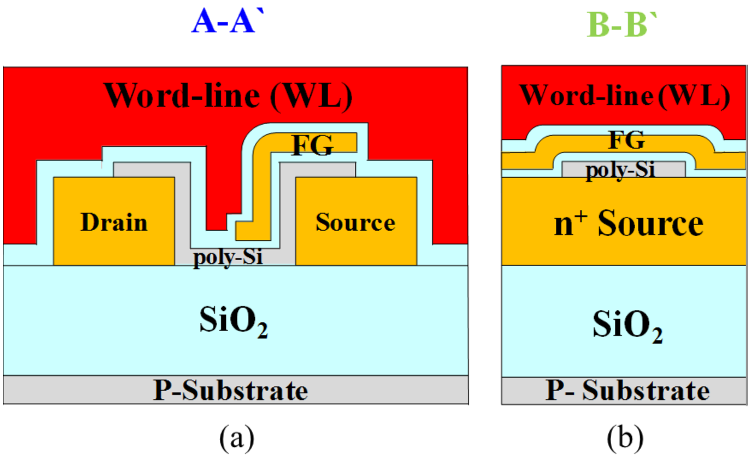


Figure 3.2. Cross-sectional views cut in the (a) WL direction and (b) BL direction of the device.

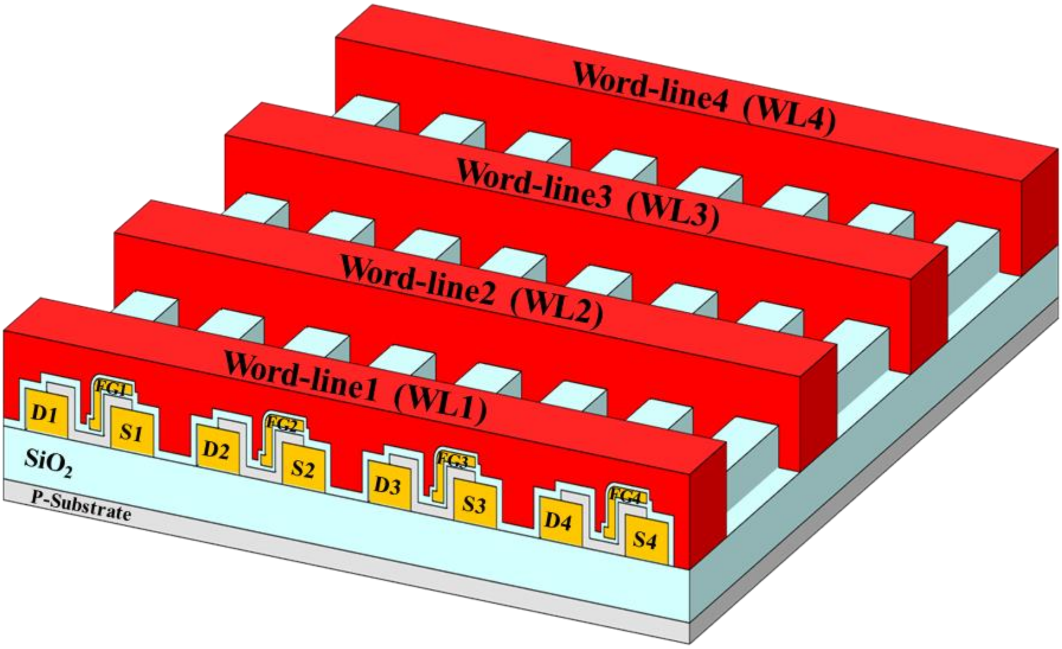


Figure 3.3. Bird’s eye view of a TFT-type NOR flash memory array.

3.2 Device fabrication

The TFT-type NOR flash memory arrays are fabricated on a 6-inch Si wafer with 6 masks and conventional CMOS process technology. The used masks are source/drain formation (1st), poly-Si channel define (2nd), FG formation (3rd), CG formation (4th), contact hole (5th), and metal line formation (6th).

The main fabrication process diagrams and detailed steps are shown in the figures 3.4, and 3.5, respectively [28]. Figure 3.4 shows the schematic cross-sectional views of the key fabrication process steps, and figure 3.5 shows the process flow of the fabrication of TFT-type NOR flash memory.

After cleaning process, which include sulfuric peroxide mixture (SPM), ammonium hydroxide-hydrogen peroxide mixture (APM), hydrochloric acid-hydrogen peroxide-water mixture (HPM), and diluted hydrogen fluoride (DHF), a 300-nm-thick layer of SiO_2 insulator was formed on top on the 6-inch Si wafer by wet oxidation process. Then, a layer of *in situ* n^+-doped poly-Si was formed on an insulator layer. After the doped poly-Si layer was patterned (first mask) by a SS03A9 photoresist (PR), a 20-nm-thick amorphous Si active layer was deposited by a low-pressure chemical vapor deposition (LPCVD), poly-crystalized by annealing, and then patterned (second mask). Figure 3.6 (a) and (b) show SEM images after this fabrication step. A 7-nm-thick layer of SiO_2 was then deposited as a tunneling oxide (T_{ox}) layer by a LPCVD process at 780 °C, after which a layer of n^+-doped poly-Si was formed and patterned as a FG (third mask). To separate the FGs, exposed FG is isotropically etched by reactive-ion-etching (RIE) process with SF_6 gas as shown in figure 3.6 (c). SiO_2 was then deposited at a thickness of 15 nm as a blocking oxide (B_{ox}) layer. The n^+-doped poly-Si was formed and patterned above the B_{ox} as control gate (CG) (fourth mask). After tetraethyl orthosilicate (TEOS) deposition, contact holes for the CGs, sources, drains were formed (fifth mask) by RIE process. Subsequently, Ti/TiN/Aluminum (Al)/TiN electrodes were formed by sputtering and were then patterned (sixth mask) by photolithography. Then, hydrogen (H_2) annealing at 450 °C for 30 min was performed to improve the contact and interface property.

SEM images of a fabricated device are shown in figure 3.6 [28]. Figure 3.6 (a) shows a SEM image of the step corresponding to figure 3.4 (b). Figure 3.6 (b) is a bird's eye view of the same step. Figures 3.6 (c) and (d) show SEM images of fabrication steps corresponding to figures 3.4 (d) and (f), respectively. For the fabricated cell devices in the array, the width of the control gate (W_{CG}) is 2 μm and the length between the source and drain (L_{CG}) is 0.5 μm. One memory cell can be scaled down to 8 F^2 if the W_{CG} is scaled to the minimum feature size (F).

Most of the processes were carried out using the equipment in Inter-University Semiconductor Research Center (ISRC) located in Seoul National University (SNU), Seoul, Korea, and *in situ* n^+-doped poly-Si layer was deposited by using the equipment of National NanoFab Center (NNFC) located in Daejeon, Korea.

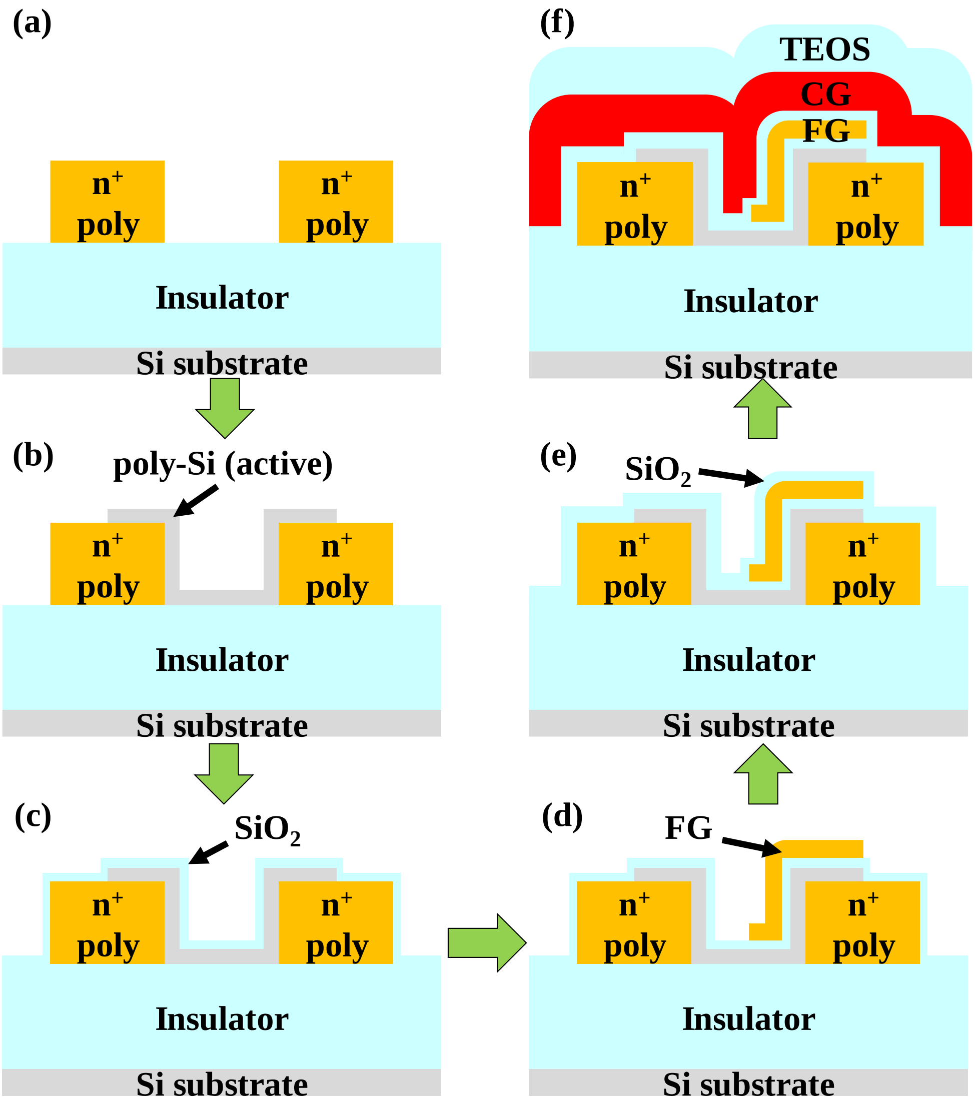


Figure 3.4. (a)-(f) Schematic cross-sectional views of the key fabrication process steps [28].

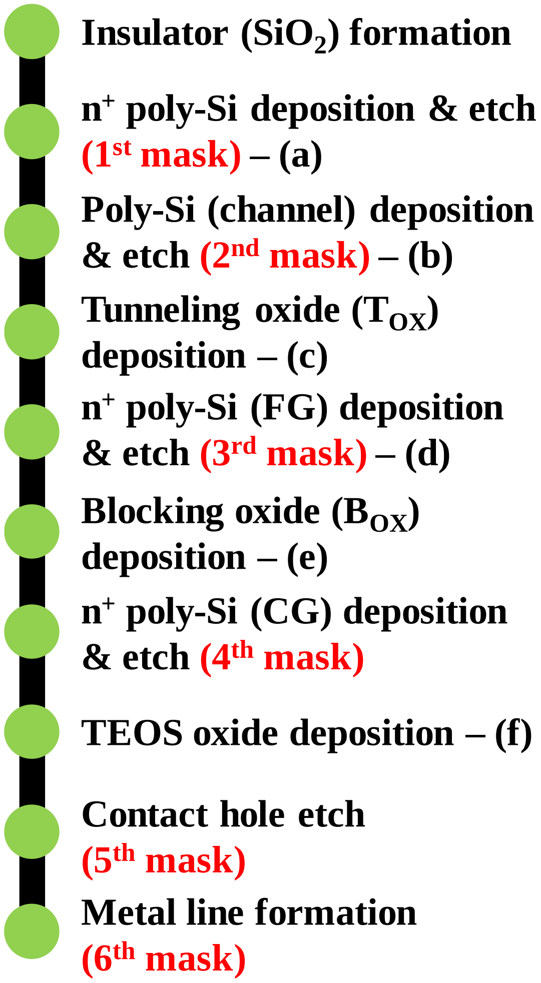

Insulator (SiO_2) formation
n+ poly-Si deposition & etch (1st mask) – (a)
Poly-Si (channel) deposition & etch (2nd mask) – (b)
Tunneling oxide (T_{OX}) deposition – (c)
n+ poly-Si (FG) deposition & etch (3rd mask) – (d)
Blocking oxide (B_{OX}) deposition – (e)
n+ poly-Si (CG) deposition & etch (4th mask)
TEOS oxide deposition – (f)
Contact hole etch (5th mask)
Metal line formation (6th mask)

Figure 3.5. Process flow of the fabrication of TFT-type NOR flash memory [28].

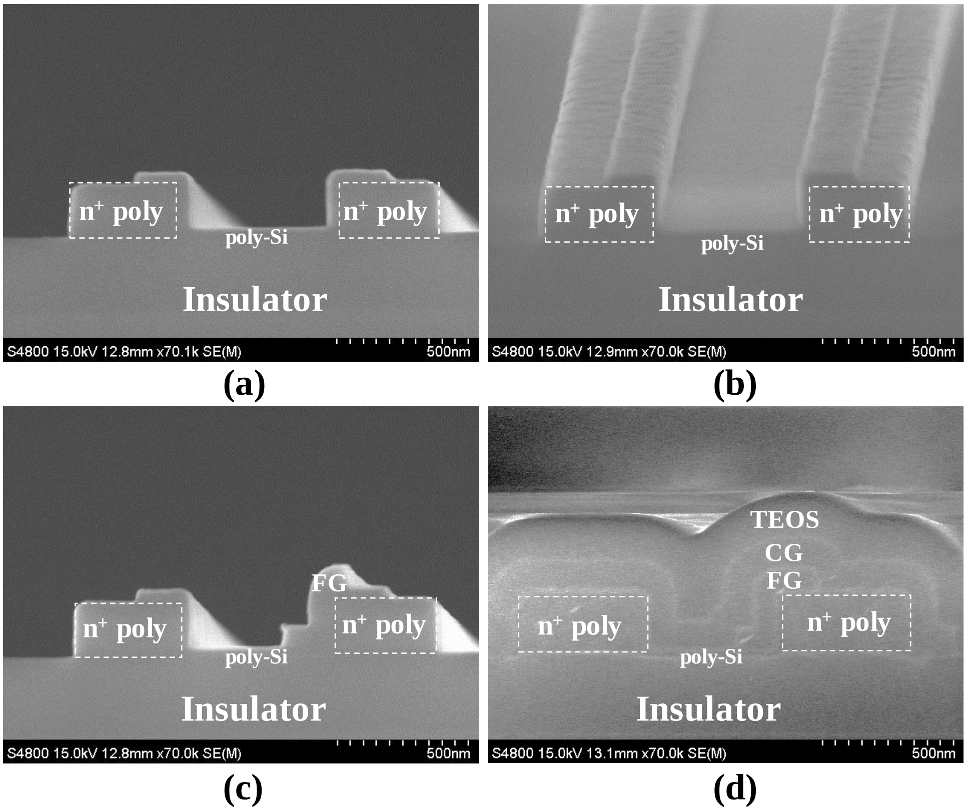


Figure 3.6. SEM cross-sectional images of fabricated structures, (a)-(b) SEM images corresponding to the step shown in figure 3.4 (b), (c) SEM image corresponding to the step shown in figure 3.4 (d), and (d) SEM image corresponding to the step shown in figure 3.4 (f) [28].

3.3 Device measurement results

3.3.1 Current-voltage (*I-V*) characteristics

The direct current (DC) *I-V* characteristics of the fabricated reference TFT and TFT-type NOR flash memory were measured by using semiconductor parameter analyzer (B1500A, Keysight) and cascade probe station.

The I_D-V_{CG} characteristics of a reference TFT and a TFT-type NOR flash memory cell as a parameter of V_D (1, 2, and 3 V) are shown in figures 3.7 and 3.8, respectively. Here, gate width (W_{CG}) and gate length (L_{CG}) of reference TFT and TFT-type NOR flash memory are 2 ㎛ and 0.5 ㎛, respectively. Because oxide layer between CG and FG increase the effective gate oxide thickness, memory devices having FG have larger threshold swing (SS) and lower on-current value than the reference FET. These measurement results show that the fabricated TFT-type NOR flash memory device works well.

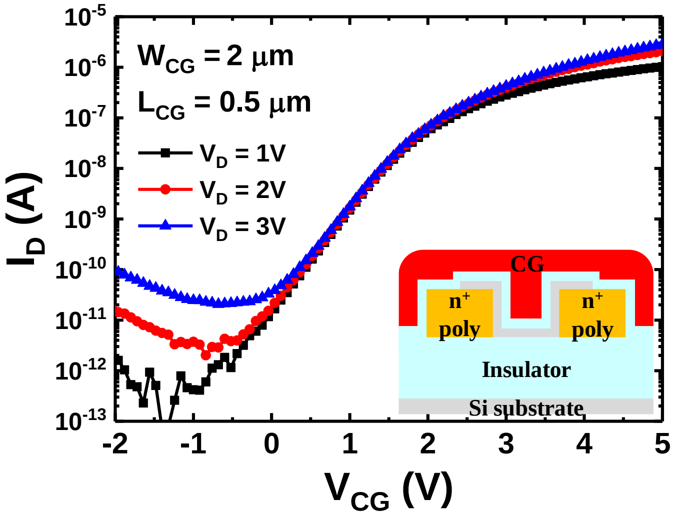


Figure 3.7. Drain current versus control-gate (CG) bias of fabricated reference TFT (w/o FG) as a parameter of the drain voltage (V_D) [28].

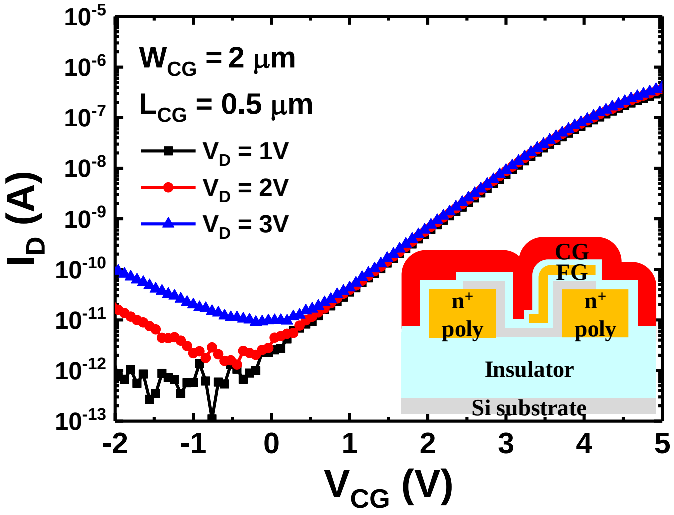
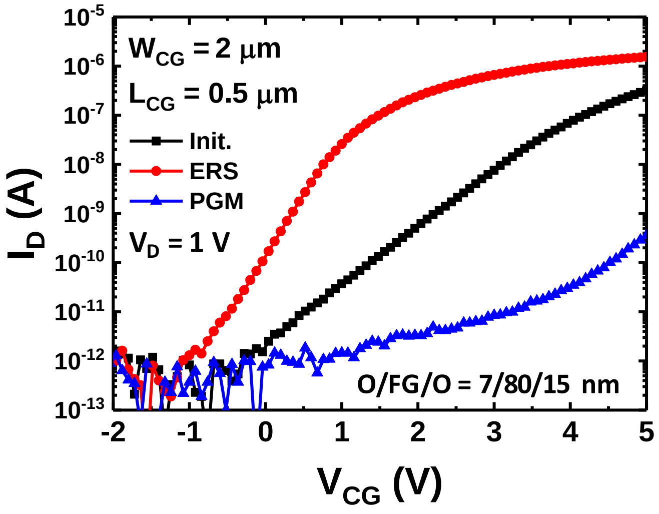


Figure 3.8. Drain current versus control-gate (CG) bias of fabricated TFT-type NOR flash memory (with FG) as a parameter of the drain voltage (V_D) [28].

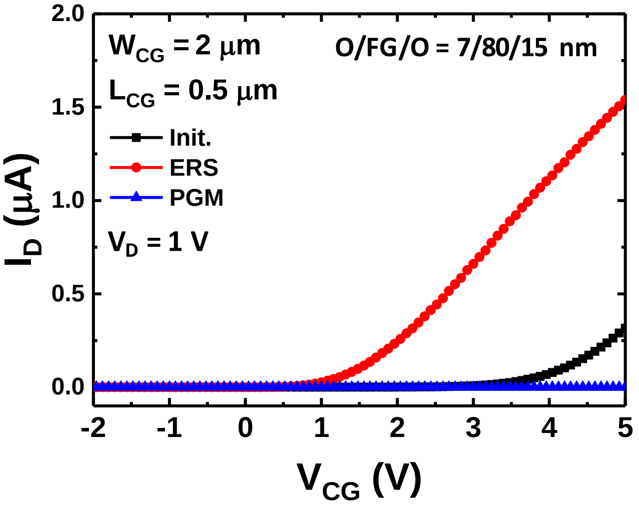
3.3.2 PGM/ERS characteristics of flash memory

All memory cells fabricated in an array represent similar I_D-V_{CG} characteristics in the initial state. The charge stored in the FG of each memory cell is reflected in the on-current (I_D/I_{BL}) of the device. This current flows to the common BL of the array and has the same effect as the weighted sum in the biological synapse array. Figure 3.9 (a) shows the drain current versus control-gate (CG) bias of fabricated TFT-type NOR flash memory (with FG) as a parameter of the memory state (initial, ERS, and PGM) in log scale, and figure 3.9 (b) shows the same result in linear scale. Figure 3.9 (a) confirms that the memory operation of the flash device works well, and this characteristic enables the device to implement the synaptic weight by the memory function. Figure 3.9 (b) represents this characteristic in linear scale. As shown in the figure, the turn-on voltage (V_T) does not fall below zero even under full ERS condition of the device, as mentioned in previous chapter.

The retention characteristics of a TFT-type NOR flash memory cell measured at 300 K is shown in figure 3.10. The I_D difference between the PGM and ERS states is kept beyond two orders of magnitude after 10^4 sec and are extrapolated to remain larger than two orders of magnitude even after 10^6 sec. In the retention characteristics shown in figure 3.10, I_D after ERS decreases slightly with time, and I_D during the PGM state doubles at 10^6 sec. Thus, reliable memory characteristics of the TFT-type NOR flash memory cell are obtained. Here, control-gate width (W_{CG}) and control-gate length (L_{CG}) of measured TFT-type NOR flash memory are 2 ㎛ and 0.5 ㎛, respectively. Among the requirements for electronic devices for use as synapses, the retention characteristic is quite important. Because the memory state of the device represents the trained weight of the synapse, keeping it for a long time is significant for the recognition capability of the synaptic device array.



(a)



(b)

Figure 3.9. Drain current versus control-gate (CG) bias of fabricated TFT-type NOR flash memory (with FG) as a parameter of the memory state (initial, ERS, and PGM) in (a) log scale and (b) linear scale.

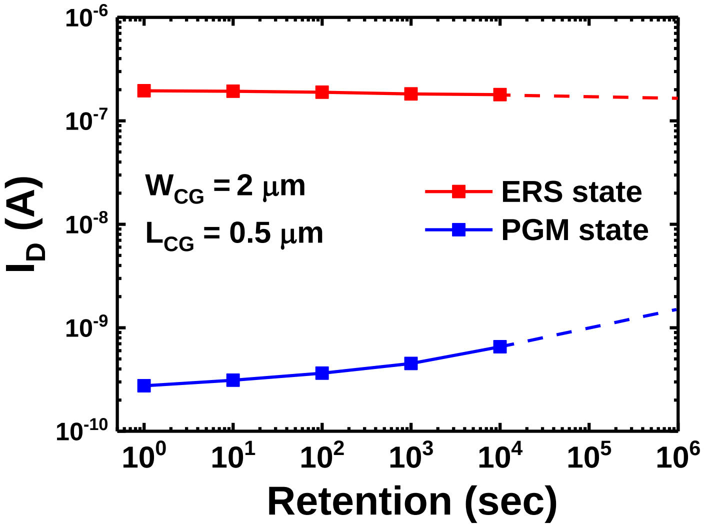


Figure 3.10. Retention characteristics of a fabricated memory device. Here, control-gate and drain biases are 3 V and 1 V, respectively, to read the I_{D}.

3.4 Device measurement results as a synaptic device

3.4.1 Circuit diagram of neuromorphic network

In order to utilize the proposed TFT-type NOR flash memory array as a synaptic device array, additional circuit system is required. Figure 3.11 shows the circuit topology of a neural network system when implemented with a TFT-type NOR flash memory array. Pattern images sent from PRE neurons are entered into the WLs of the memory array. Signals input to WLs reflect the weights stored in the synapses and are converted into current and then added to the common drain line (CDL) of the array. The current in the CDL flows through the current mirror circuit to the POST neuron circuit outside the array. Current is deposited in the membrane capacitor of the POST neuron, which fires when the membrane potential exceeds a certain threshold. Each POST neuron is connected via a FET-type inhibitory synapse which acts as a suppression, which in turn suppresses the firing of the neuron other than itself. The firing signal from the POST neuron triggers a switch between the common source line (CSL) and the ground, allowing the CSL to be connected to the spike-generated circuit. The firing signal is also sent to the spike-generation circuit, which generates a feedback spike pulse to the CSL on its own array, and the output spike pulse is output to the WLs of the synapse array in next layer. The weight update methods and circuit configuration described above enable self-learning of synapses and neurons arrays without intervention of external circuits and computation.

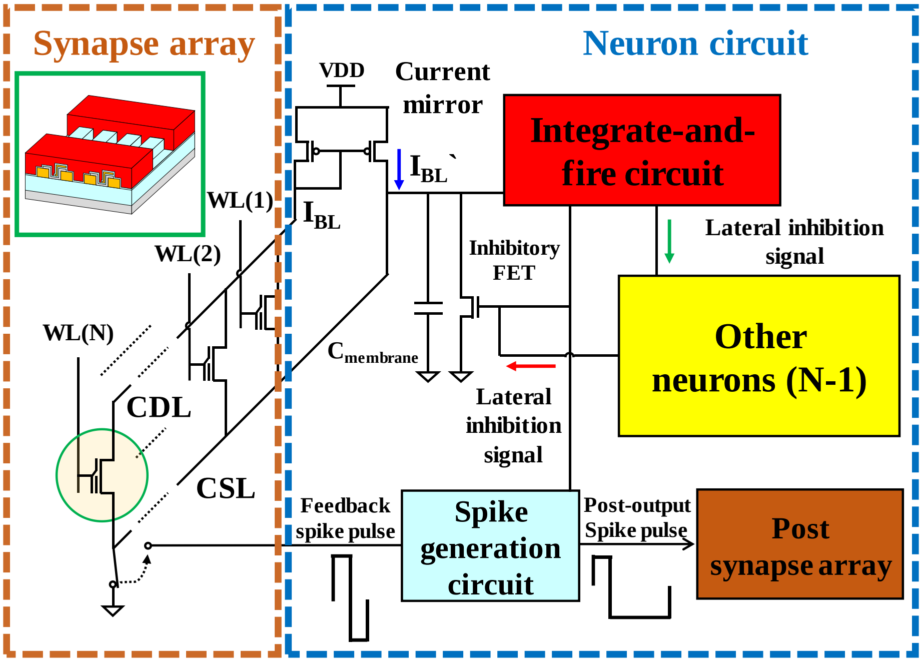


Figure 3.11. Schematic circuit diagram of an unsupervised neuromorphic network with a TFT-type NOR flash memory array and a neuron circuit [28].

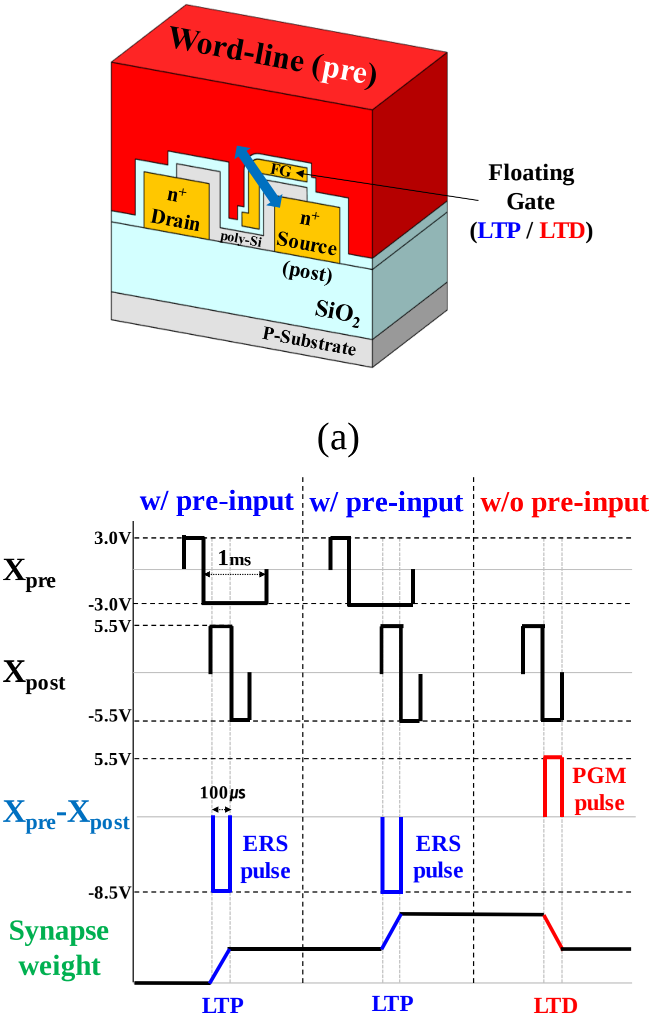
3.4.2 Pulse scheme for STDP weight update

In this chapter, we describes a new pulse scheme for using the proposed TFT-type NOR flash memory as a synaptic device. Here, we will address the third and fourth issues identified in the previous chapter 2.5 when using the SONOS gated-diode memory as synaptic device array. The first issue is that the time length of the pulse used for LTP / LTD operation is too long, and the second issue is that additional noise input pattern must be used to train pattern images. To overcome these two problems, a new pulse scheme for STDP weight update is proposed.

For STDP operation, the synapse cells can be potentiated or depressed selectively using the pulse scheme shown in figure 3.12. The basic principles of synapse cell LTP / LTD operations are as follows. When a certain neurons is fired, the weights of synapses that contribute to the neuron's firing are potentiated. These characteristics are designed to be automatically performed by overlapping PRE input pulse and the feedback pulse generated by the POST output signal. On the other hand, for synapses where no input signal is entered, the feedback signal causes the LTD process. As shown in figure 3.12 (a), input signal from PRE neuron and feedback signal from spike-generation circuits are applied to WL and n^+-source respectively to change the weight of synapse cell. The PGM and ERS operation of charges stored in the FG depend on the voltage state of the WL and the n^+-source connected to the CSL. If the input pulse is applied and then a neuron is fired, the tail portion of the input pulse overlapped by the head portion of the feedback pulse, as represented by the LTP operation shown in figure 3.12 (b). As a result, a -8.5 V pulse, expressed by X_{pre}-X_{post} in figure 3.12 (b), is applied to the WL for 100 μs to perform an ERS operation in the FG similar to an LTP operation in biological synapse. Conversely, if there is no input signal from the PRE input, only the feedback pulse is applied to the memory cell source. This is the same as applying a pulse with a magnitude of 5.5 V and a width of 100 μs to the WL, which stores electrons in FG (PGM Operation) and has the same effect as the LTD of synapses.

Table 3.1 summarizes the pulse scheme for these weight updates and weight reading operations. Reduce power consumption by preventing leakage current during unit PGM and ERS with floated CDL electrode during weight updates. Reading pulses for reading and weighted-sum operation of the synapses are applied to the WL with a magnitude of 3 V and a width of 100 μs.

So far we have proposed a new pulse scheme used to utilize TFT-type NOR flash memory array as a synaptic array. The new pulse scheme can solve the previously mentioned problems. First, as table 3.1 summarizes, the width of pulse used to train the synaptic weight is 100 μs. This is three orders shorter than the 10 ms width of the learning pulse used in the SONOS gated-diode memory. This allows for the implementation of a neuromorphic system with faster learning time. Noise input pattern, which was pointed out as another problem, is not used in the newly proposed pulse scheme at all. This is because a new proposed pulse scheme could depress the weights of the background and the unwanted part without the noise input pattern. This has the advantage of reducing the burden of additional circuits required to produce noise input pattern and reducing learning time compared to the previous pulse scheme that required repeated entering of two inputs.



(a)

(b)

Figure 3.12. (a) Schematic diagram of PRE (input) and POST (feedback) pulses that cause a weight update of TFT-type NOR flash synaptic device. (b) Pulse scheme of PRE and POST neurons to the TFT-type NOR flash synaptic array that causes a LTP and LTD [28].

Table 3.1. Bias conditions for the weight update (LTP/LTD) and weight read operation of cells in a TFT-type NOR flash synapse array [28].

	LTP **(w/ pre-input)**	**LTD** **(w/o pre-input)**	**Weight** **read**
V_{WL}	-3 V	0 V	3 V
V_{CSL}	5.5 V	-5.5 V	0 V
V_{CDL}	floating	floating	1 V
Time	100μs	100μs	100μs

3.4.3 LTP/LTD characteristics

In this chapter, we present the measurement results when the proposed new pulse scheme is applied to the actual synaptic device, and the modelling process for applying the measured characteristics to the pattern recognition simulation.

First, the pulse scheme for the synapse weight update described in previous chapter was applied to the actual devices; these results are shown in figure 3.13. Following 20 iterative LTP pulses, 20 repeated LTD pulses are applied to the WL and the source electrodes of the synaptic device, and the weight of the device is measured between them. In this case, the pulse scheme used for each operation is the same as in table 3.1. The repetitive increase in the weight of the synapse is determined by the repeated application of the LTP pulse depending on the amount of charge stored in the FG. Similarly, the amount of weight change when LTD pulse is applied is affected by the amount of charges stored in the FG. As a result, the degree of weight change depends on the state of the synaptic weight, and needs to be modeled to perform system-level simulation. Therefore, we have modeled the LTP/LTD behavior characteristics of the synapses obtained through measurements.

The information for the numerical modeling is as follows. We have proceeded the numerical modeling of the measurement data. The behavioral modeling results and the parameters used are summarized in table 3.2. Figure 3.14 shows the result of comparing the fitting result with the measurement data, and confirms that the two results match well. In addition, we can estimate the weight change behavior of synapse with various synapse weight state by using this numerical fitting result, which can be confirmed by figure 3.15. Figure 3.15 shows the fitting results for each weight state states, and in this case we extracted the LTP / LTD change ratio for each weight state. This modeling produces individual STDP behaviors based on the weight state of the synapse. Figure 3.16 shows the STDP curves for the three representative weight states derived from the modeling results of figure 3.15.

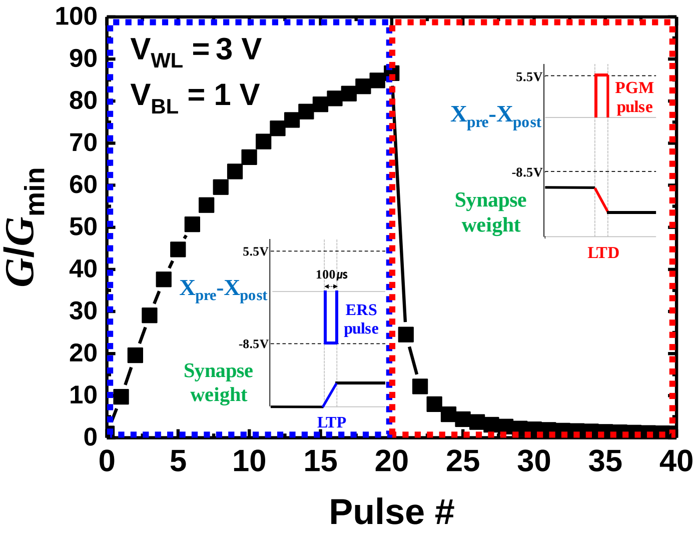


Figure 3.13. The LTP / LTD repetition characteristics of a TFT-type NOR flash memory device measured using the pulse scheme of table 3.1 [28].

Table 3.2. Fitting parameter values of model equations for the simulation [28].

	LTP	LTD	
Model equation	$\delta G = e^{(a+bG+cG^2)}$	$\delta G = -(A_0 + A_1 G + A_2 G^2 + A_3 G^3 + A_4 G^4)$	
Parameter	$a = -19.56$	$A_0 = -4.263 \times 10^{-11}$	$A_3 = -2.811 \times 10^{15}$
	$b = 2.11 \times 10^7$	$A_1 = 0.1186$	$A_4 = 4.1064 \times 10^{22}$
	$c = -2.94 \times 10^{15}$	$A_2 = 6.7244 \times 10^7$	

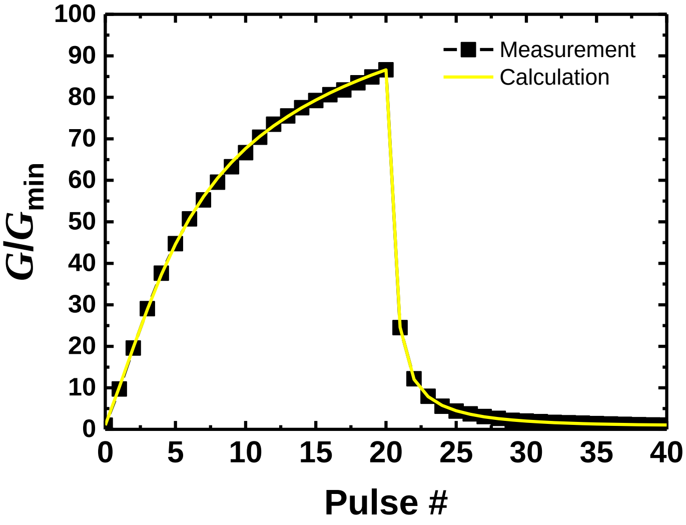


Figure 3.14. Comparison of fitting calculation result with the measurement data.

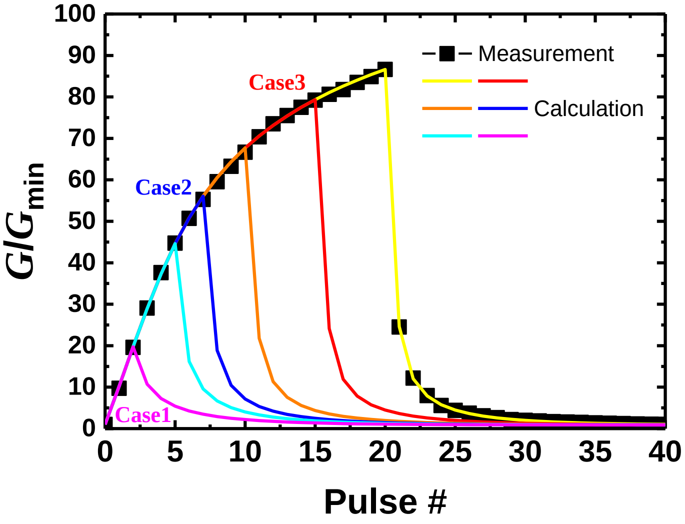


Figure 3.15. The LTP / LTD behavior modeling results in various weight states using calculated fitting results.

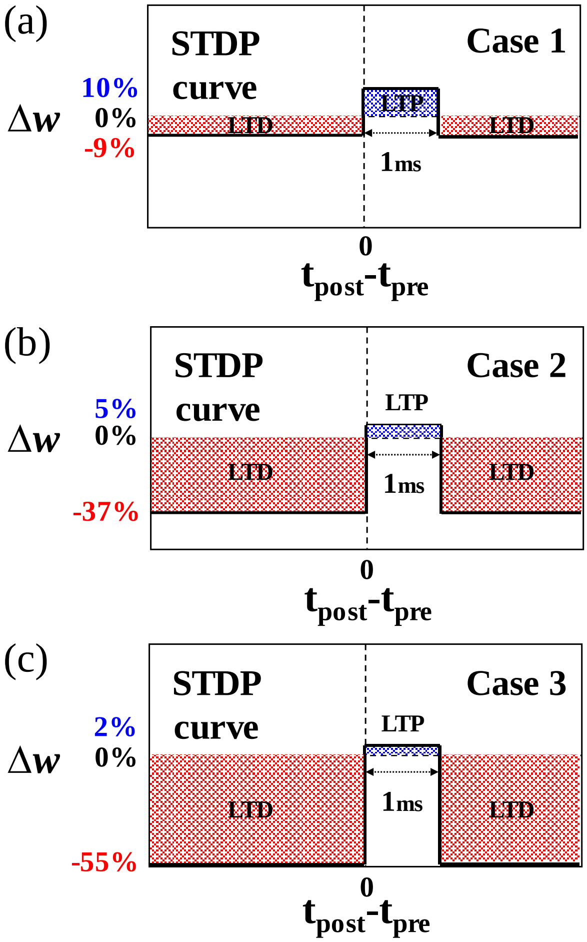


Figure 3.16. STDP behavior depending on the current weight in a synaptic device when the current weight is low (case 1, low G/G_{min}: 19.7) (a), moderate (case 2, moderate G/G_{min}: 56.1) (b), and high (case 3, high G/G_{min}: 79.3) (c) [28].

3.5 Simulation results of pattern recognition

3.5.1 Overall flow of pattern learning and recognition

From this chapter, we will verify the pattern learning and recognition capabilities of the proposed TFT-type NOR flash memory when applied to the neuromorphic system. Figures 3.17 (a) and (b) show the flowchart of the overall pattern learning and recognition process used in the simulations, respectively. The simulation was performed with software MATLAB, and the operating characteristics of the synapses are determined by the measured characteristics of the TFT-type NOR flash memory cell. For the simulation work, neuron circuits were assumed to consist of ideal capacitors and comparators. The pattern learning process is described in figure 3.17 (a).

First, reset the synapses by randomizing the weight of all synapses. In the PRE target image for learning, only the part where the input value of the pixel exists in that image triggers the X_{pre} pulse of figure 3.12 (b) to the WLs. The X_{pre} signals make each synapse device send weighted current to the CDL, which can cause the firing of the POST neuron through I&F circuit. The resulting postsynaptic spike is sent to other neurons and inhibits them by discharging accumulated charges in their integrate capacitor. This process allows each neuron to learn its own image pattern for implementing pattern classification. The fired neuron also sends a feedback spike to the CSL of the synapse cells connected to it, which immediately updates the synaptic weights in that neuron. Figure 3.17 (b) illustrates the recognition process of neurons connected to trained synapses. Additional circuits such as that for lateral inhibition and feedback spike delivery are not required during the recognition process. Once the target image is entered, the resulting POST neurons can be observed through the I&F circuits. At this point, the recognition rate of the trained neurons can be checked by comparing it with that of the other neurons. Until now, we have a full overview of the simulations for pattern learning and recognition, and then we will look at actual pattern learning, classification, and recognition results using this simulation structure.

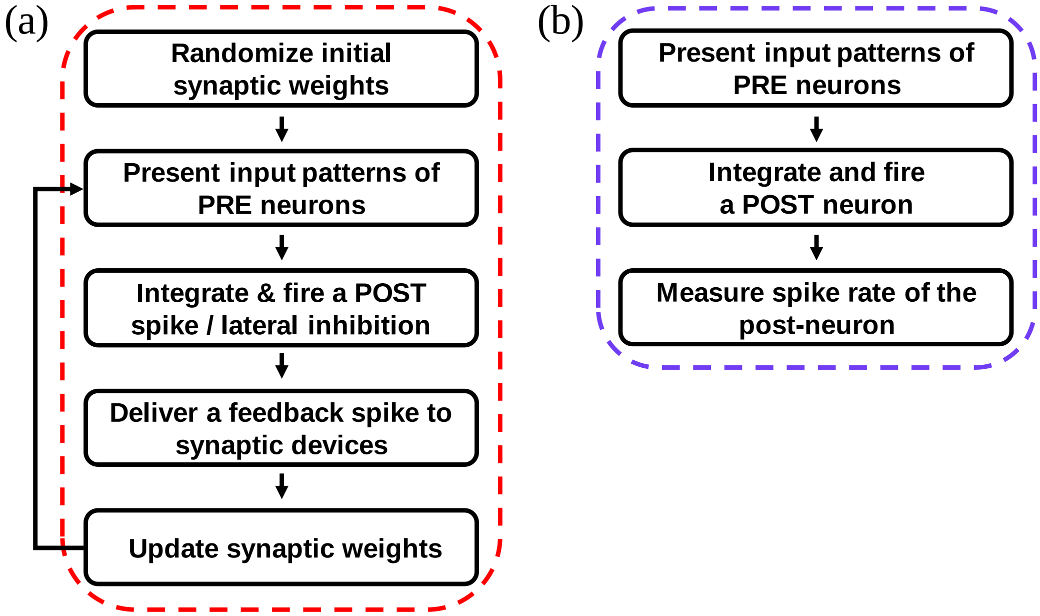

(a)
Randomize initial synaptic weights
Present input patterns of PRE neurons
Integrate & fire a POST spike / lateral inhibition
Deliver a feedback spike to synaptic devices
Update synaptic weights
(b)
Present input patterns of PRE neurons
Integrate and fire a POST neuron
Measure spike rate of the post-neuron

Figure 3.17. Flowchart of pattern (a) learning and (b) recognition [28].

3.5.2 Dot-pattern learning and classification

In this chapter, we have identified learning and classification capabilities of dot-patterns using the characteristics of the synaptic device array and MATLAB simulation structure that have been presented. Figure 3.18 shows the four 3 × 3 target input images used in this simulation.

Figure 3.19 shows the pattern learning process in a single neuron containing nine synaptic devices (9 × 1). It is confirmed that the synaptic weights of the array update correctly based on the STDP action in the synapse array when each image is presented 70 times sequentially. First, when pattern 1 is learned for 70 epochs, the red lines indicate the average weight of the synapses corresponding to pattern 1 (solid symbols) and the average weight of the synapses not corresponding to pattern 1 (open symbols). In order to confirm the updating of the learned pattern when the other pattern is applied to the learned synapse array, pattern 2 was presented for 70 epochs after the pattern 1 learning step. During the pattern 2 learning period, the blue lines show the potentiation of the synapses corresponding to pattern 2 (solid symbols) and the depression process of synapses not corresponding to pattern 2 (open symbols) are performed. The inset figures show the weight of each synapse in the array when each input pattern is applied. The initial synaptic weights are randomly distributed between the minimum to maximum weights of the proposed memory cells. It can be confirmed that the weight learning for pattern 1 is completed through 70 epochs, and after the subsequent 70 epochs, the learning of weights corresponding to pattern 2 is achieved. This result shows that the learning of the desired input pattern and pattern updating are performed successfully. For the proposed unsupervised pattern learning, there are no additional input signals, which consume more power and make the learning process more complex.

Figure 3.20 shows the pattern learning and recognition results for a multi-neuron array (9 × 4) composed of nine (3 × 3) PRE-input neurons and four POST neurons. It is necessary to take advantage of the lateral inhibition function of each of the four neurons to ensure the learning of their own unique patterns. To implement the lateral inhibition function, inhibitory synapses are used to lower the membrane potential of neurons other than the fired neuron. POST neurons are connected to each other via an inhibitory synapse. An inhibitory factor (in this case

30%), which determines the amount of reduction in the surrounding neurons, should be considered carefully. If the inhibitory factor is too high, only a small number of neurons will fire repeatedly, interfering with the learning of other neurons, while if it is too low, it will be difficult to distinguish each neuron's own learning pattern. Fig. 11 (a) shows the progress of multi-neuron learning when the input patterns are presented repeatedly through the PRE input of the synaptic array. The average value of the pattern (solid symbols) and the background (open symbols) weight of neurons 1-4 over epochs are shown. In the early stages of learning, there are oscillations of the pattern weights, but after a certain number of epochs, the weights of the synapses belonging to each neuron are gradually tuned according to a different pattern. The classification ability of patterns using the synapse array that has undergone this multi-pattern learning process is shown in figure 3.20 (b). This figure indicates that a specific POST neuron fires in response to each input pattern when four input patterns are applied in a random order. This system is thus able to distinguish four distinct patterns with the multi-neuron array effectively.

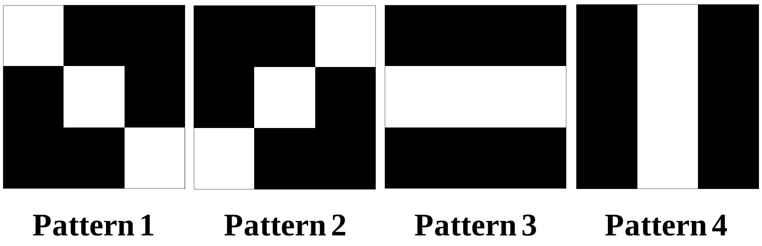


Figure 3.18. The four 3 × 3 image patterns used in the simulation.

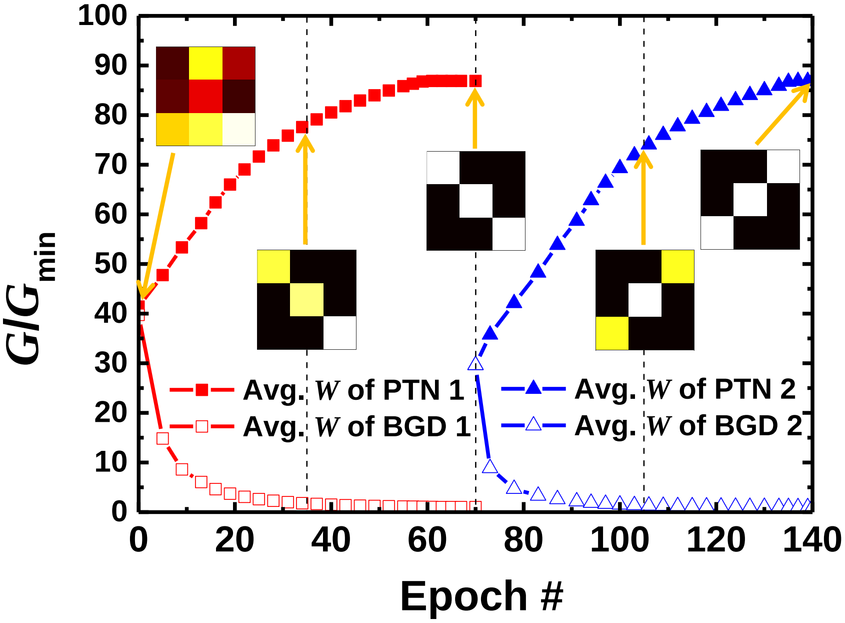


Figure 3.19. Unsupervised pattern learning and updating results with a single neuron. Average weights of the targeted pattern synapses (PTN, solid symbols) and the background synapses (BGD, open symbols) when the first pattern 1 and the second pattern 2 in figure 3.18 were learned 70 times in order. Inset images represent the weight map of the synapse array at the time of each epoch.

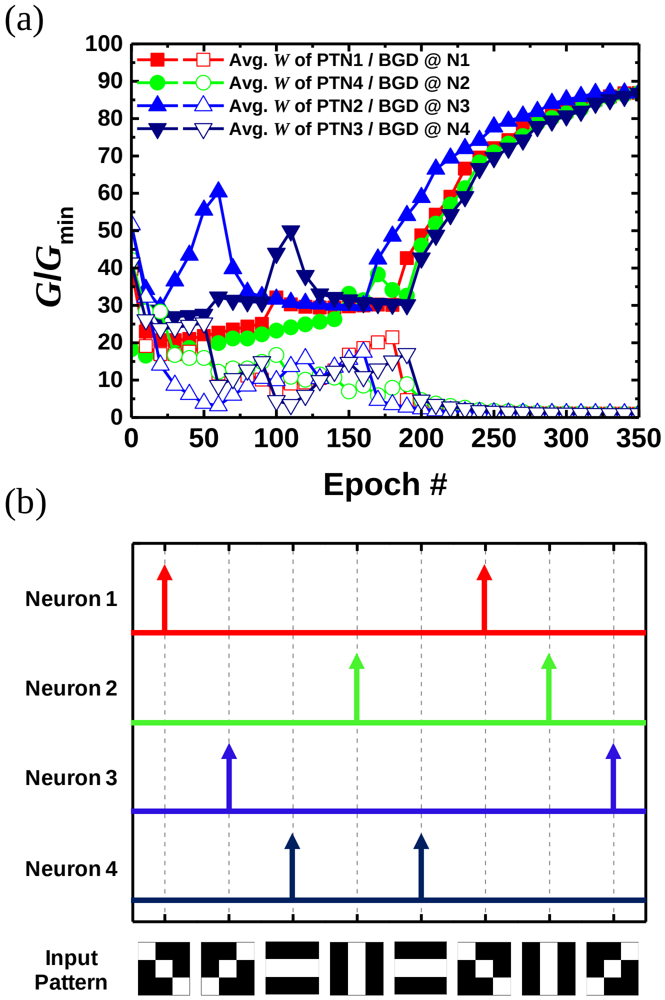


Figure 3.20. (a) Result of unsupervised multi-pattern learning and recognition. Average weights of the targeted pattern synapses (solid symbols) and the background synapses (open symbols) are shown when patterns 1-4 in figure 3.18 were sequentially presented 350 times. (b) Classification behavior of neurons when random image patterns are applied after the multi-pattern learning process.

3.5.3 MNIST pattern learning and classification

In the previous chapter, pattern learning and classification processes were demonstrated using simple dot-patterns. However, these examples are not enough to apply to real-world pattern recognition problem because image patterns are too simple (only 9 pixels) and only have four distinct types. Therefore, we have extended the sample patterns to the MNIST handwritten dataset and discussed the results of the learning simulations. Figure 3.21 shows the ten 28 × 28 MNIST handwritten target input images used in this simulation.

Figure 3.22 shows the pattern learning process of a single neuron, which includes 784 (28 × 28) synaptic devices (i.e. 784 × 1). Verify that the synaptic weights of the array are updated correctly based on the STDP behavior of the synaptic device array when each image is displayed 80 times sequentially. The insertion figures show the weight maps of each snapshot in the array when each input pattern is applied. The initial synaptic weights are randomly distributed between the minimum and maximum weights of the proposed memory cell. The weight learning for pattern 2 is completed in the 80 epochs, and after the 80 consecutive epochs, weight learning for pattern 5 is completed, and after 80 epochs, weight learning for pattern 9 has been completed. This result indicates that the desired input pattern and pattern update have been performed successfully. These results suggest that in single-neuron learning using the proposed synaptic device array, it is possible to learn the desired pattern and to change the learned weight array into a different pattern.

Figure 3.23 shows pattern learning and classification results for multiple neuron array (784 × 10), consisting of 784 PRE input neurons and 10 POST neurons. In order to implement lateral inhibition, inhibitory synapses are used to reduce the potential of membrane in neurons other than those fired. These inhibitory synapses are implemented by inhibitory FETs connected to membrane capacitors in each neuron as shown is figure 3.11. Each POST neuron is connected to each other via these inhibitory FETs as shown in figure 3.11. An inhibition factor (in this case 47%) determining the potential reduction of membrane in neurons other than those fired. Figure 3.23 (a) shows the progress of multiple neuron learning when the input digit patterns are repeatedly displayed through PRE input neurons of the synaptic device array. This figure shows the process of changing the weight states of the synapses corresponding to each neuron at each epoch count. In the early stages of learning there are oscillations of pattern weights, but after a certain number of epochs, the weights of synapses belonging to each neuron are gradually adjusted according to different patterns. The ability to classify patterns using the proposed synaptic device array with multiple pattern learning sequence is shown in figure 3.23 (b). This figure shows how POST neurons fire in response to each input pattern when 10-digit patterns are applied in random order. The digit patterns can be classified by comparing the POST neuron's firing rate of a neuron that has learned the digit pattern with the POST neuron's firing rate of other neurons. This effectively identifies 10 distinct patterns in multiple Neuron arrays. As a result, 10 distinct digit patterns are effectively distinguished in the proposed multi-neuron array.

In Addition, we analyzed the effect of the synaptic device variation in the classification task using hardware based neural network. We have evaluated the accuracy of learning based on device variation in the proposed neuromorphic system. The results are shown in the figure 3.24. We have considered device-to-device variations in the potentiation, depression, and minimum/maximum values of the device conductance. The parameters of the synaptic devices are determined to have a Gaussian distribution. The ratio of the standard deviation to the mean represents the device-to-device variability. Although the characteristic of neural network does not show a significant change in the final learning results even if the variation of the device varied by 30%, the epochs for learning unique pattern are increasing. It is also possible to confirm the difference in sharpness in the weight map even after the learning is completed. If the variation exceeds 30%, it can be confirmed that the learning is not performed smoothly. This result can be changed according to the size of the network and the type of the pattern, but it can be seen that the variation of the device results in degradation of learning accuracy. Therefore, it is important to design the artificial neural network as to minimize the variations of the devices used.

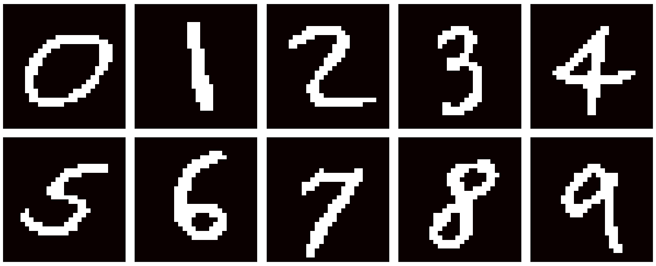


Figure 3.21. The ten 28 × 28 MNIST handwritten digits used in the simulation [28].

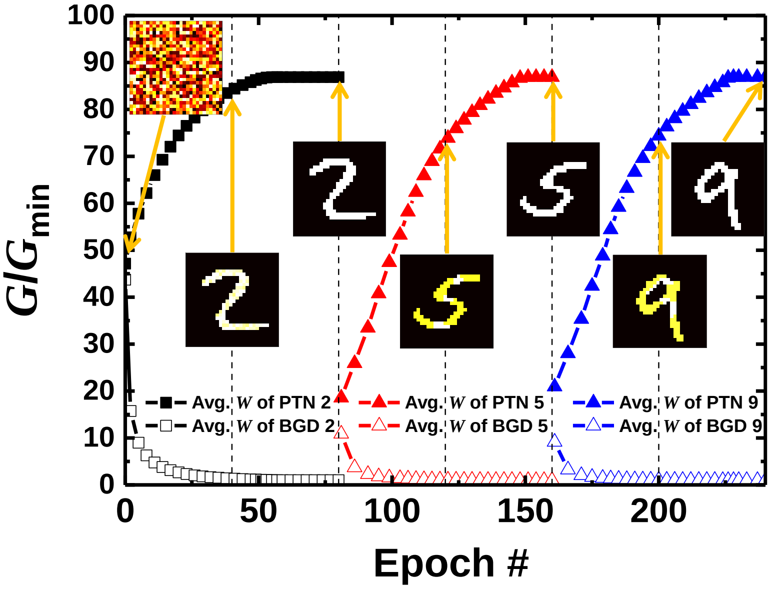


Figure 3.22. Unsupervised pattern learning and updating results with a single neuron. Average weights of the targeted pattern synapses (PTN, solid symbols) and the background synapses (BGD, open symbols) when the first pattern 2, the second pattern 5, and the third pattern 9 in figure 3.21 were learned 80 times in order. Inset images represent the weight map of the synapse array at the time of each epoch [28].

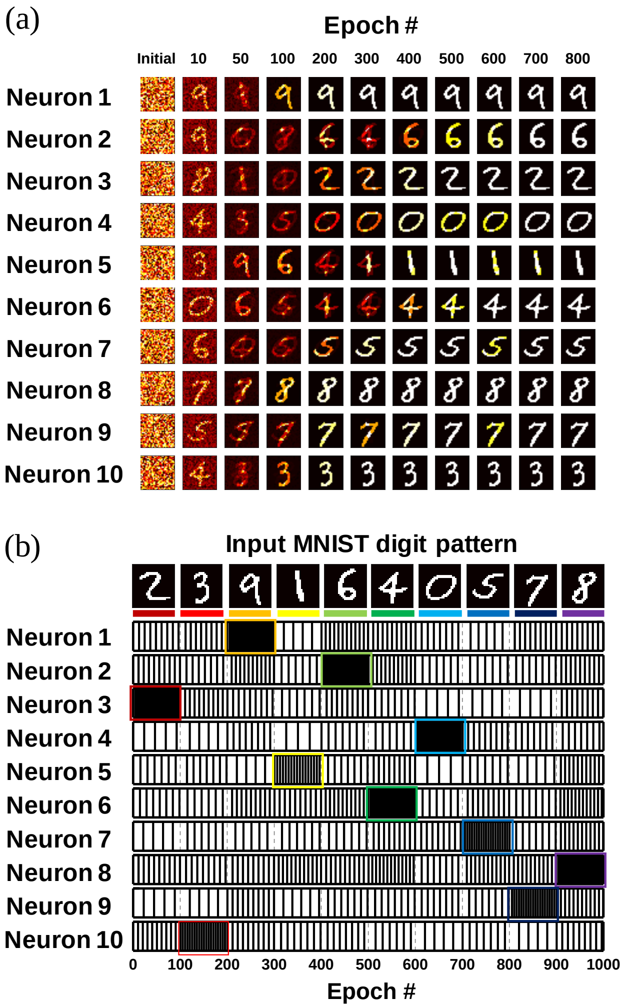


Figure 3.23. Result of unsupervised multi-pattern learning and recognition with the multi-neuron array. (a) The process of changing the weights of the synapses corresponding to each neuron are shown when patterns 0-9 in figure 3.21 were randomly presented 800 times. (b) Classification behavior of neurons when random digit patterns are applied after the multi-pattern learning process [28].

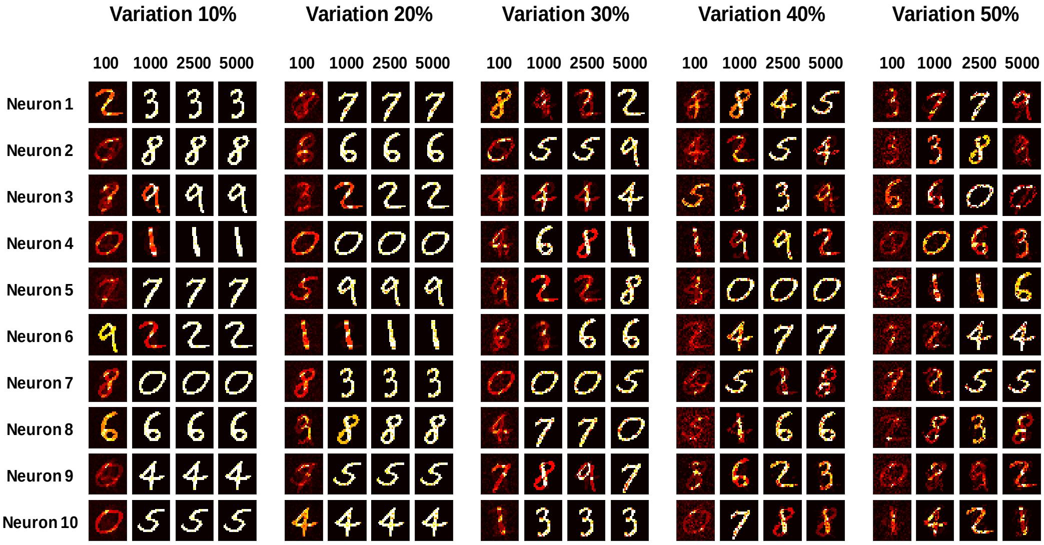


Figure 3.24. Result of unsupervised multi-pattern learning progress according to device variation.

3.5.4 Homeostatic property for high cognitive performance

In previous chapters, we have discussed the learning and classification capabilities of simple dot-patterns and MNIST handwritten digit patterns on a neural network using the proposed synaptic device array. However, so far, simulations have been conducted using only a limited number of selected patterns, which in fact must be learned and recognized based on a large database of various forms. Therefore, we have expanded learning task to use 60,000 training examples of full binary MNIST dataset. Then, 10,000 test examples of MNIST dataset were used to check the recognition accuracy of the neural network [51].

However, if the multi-pattern learning method presented previously is applied, each neuron does not learn its individual number pattern as desired. Only a few neurons are trained exclusively, and these neurons are repeatedly fired. These misguided learning results can be found in figure 3.25 (a). This is the result of identifying the firing frequency of 30 POST neurons by entering 10000 test datasets after learning 60000 training datasets. As shown in the figure, only a few neurons are fired intensively, and the others are rarely responding to any digit pattern input.

To address this imbalance problem, homeostatic properties, one of the operational principles of biological neurons, were introduced. Figure 3.26 shows the operating principles of the biological homeostatic property [50]. If the activity level is too low (left), the calcium concentration falls below the target. As calcium concentrations decrease, membrane current can return the neuron's activity to the target level. If the firing rate and calcium concentration are too high (right), adjust the membrane and synaptic current in the opposite direction to reset the target activity level [50]. Therefore, we adopted the homeostatic property of these biological neurons and applied them to the neuron circuit system used for pattern learning simulation. As with the biological characteristics, the membrane threshold voltage of the neuron was adjusted according to the activity of the neuron. If the neuron's activity is too high, it is set to increase the membrane threshold and vice versa. Figure 3.25 (b) shows the activity of POST neurons after applying homeostatic properties. It is confirmed that neuron’s firing is much more balanced than when homeostatic properties are not applied.

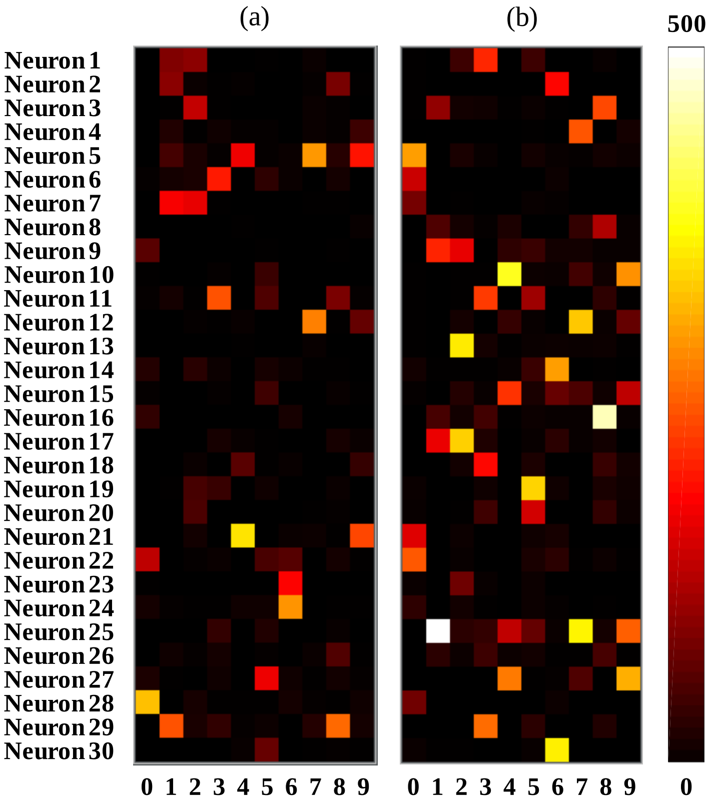


Figure 3.25. Result of unsupervised multi-pattern learning progress using 30 POST neurons (a) without, and (b) with homeostatic property.

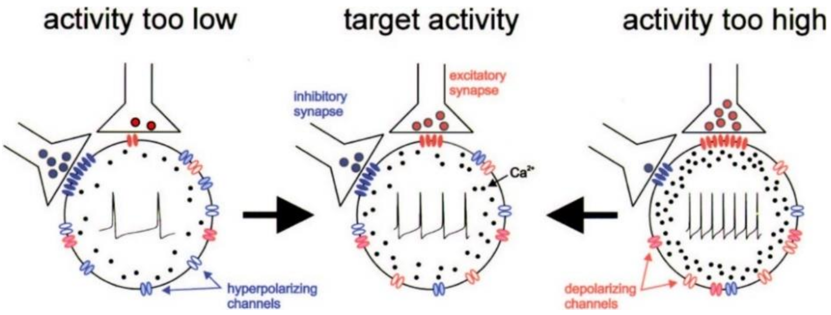


Figure 3.26. Basic mechanisms of activity-dependent homeostatic regulation in model neurons [50].

3.5.5 Pulse scheme optimization

In the previous chapter, the homeostatic property was adopted to balance firing rate between POST neurons. However, introducing the homeostatic property did not complete the learning and recognition process of the entire MNIST dataset, which is due to the characteristics of abrupt LTD phenomenon of the device. Figure 3.27 (a) shows the results of the device measurement showing the abrupt LTD characteristic used in the previous simulations. As shown in figure 3.27 (b), involvement in the LTD of a synaptic device is the tail portion of X_{post} pulse. Therefore, it is possible to analyze that the pulse amplitude of the tail section of the X_{post} can be adjusted to resolve the abrupt LTD phenomenon. Figure 3.28 shows the change in the conductance of the measured synaptic device by applying the LTD pulse in three additional cases other than the -5.5 V used in the previous measurement. The depression processes with LTD pulses amplitude of -5.2 V for case 1, -5 V for case 2, and -4.8 V for case 3 are shown, respectively. After confirming that the amplitude of the LTD pulse can be adjusted to improve the abrupt depression phenomenon, the measurement results were applied to the simulations. Figure 3.29 shows the results of learning and recognition when applying the measurement results in these three cases. From case 1 to 3, the slower the LTD conductance changes, the more consistent the input pattern and the desired output result. This result led to the completion of a new pulse scheme optimized through case 3, which greatly improved the entire MNIST dataset learning / recognition.

Finally, we have proceeded for obtaining a higher recognition rate by expanding the number of POST neurons. Figure 3.30 shows the recognition rates of proposed spiking neural network for unsupervised online learning as a parameter of the number of POST neurons. In the figure, we can see an increase in recognition rate by increasing the number of POST neurons from 10 to 100. Finally, the recognition rate reached the 82% when 100 POST neurons were used.

Figure 3.31 (a)-(c) identifies a confusion matrix when the number of POST neurons are 30, 50, and 100. Figure 3.31 (d) shows that when the number of POST neurons is 100, the synapses belonging to each neuron have been trained well for the specific numbers.

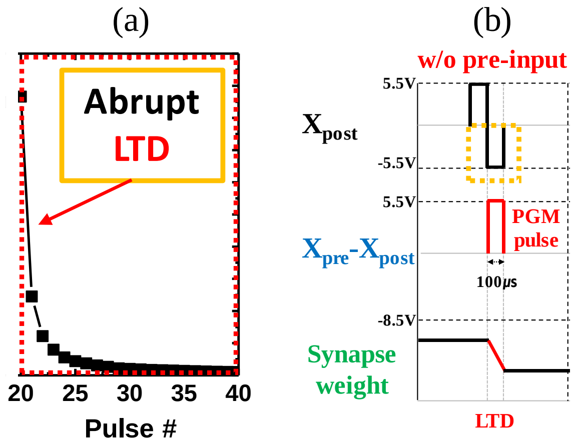


Figure 3.27. (a) The abrupt LTD phenomenon as a result of previous measurement. (b) The portion of pulse scheme causing the LTD.

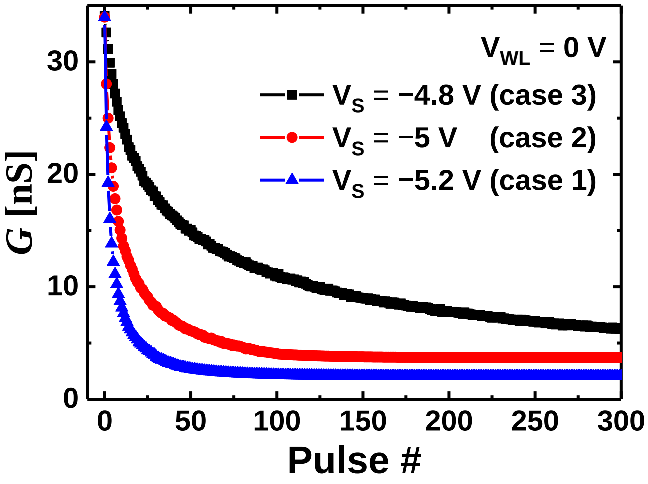


Figure 3.28. The LTD characteristics using three different LTD pulse amplitudes (Case 1: -5.2 V, Case 2: -5 V, and Case 3: -4.8 V) [51].

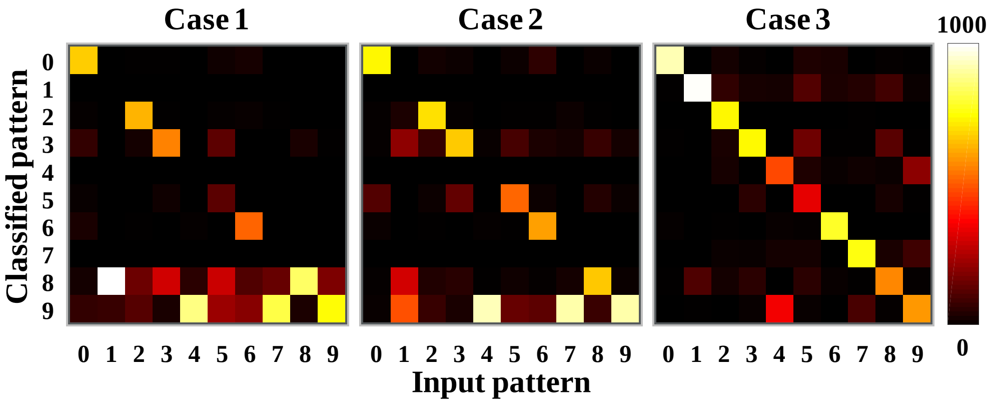


Figure 3.29. Confusion matrix of pattern recognition results based on three different cases of pulse conditions (Case 1, 2, and 3).

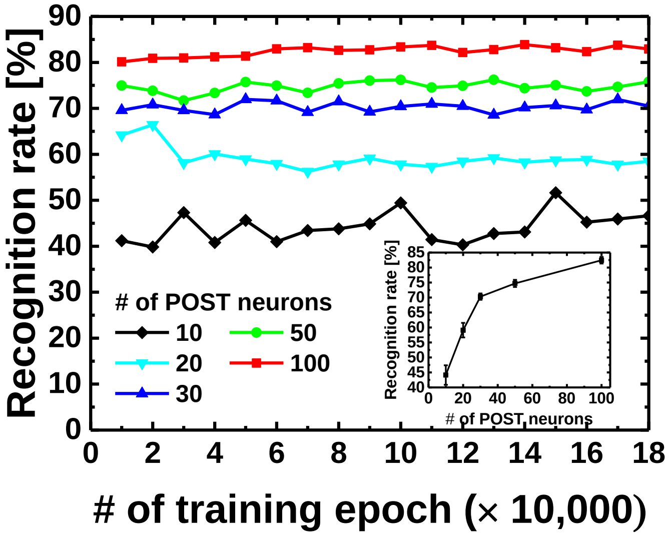


Figure 3.30. Recognition rates of proposed neural network for unsupervised online learning as a parameter of the number of POST neurons.

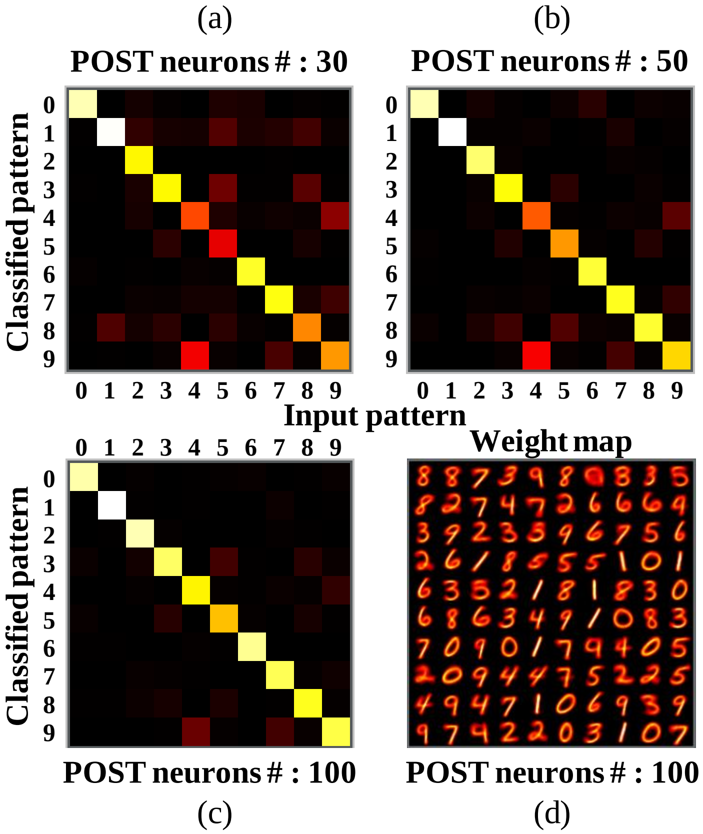


Figure 3.31. Confusion matrix of the full binary MNIST pattern classification results with (a) 30 POST neurons, (b) 50 POST neurons, (c) 100 POST neurons. (d) Learning result of synaptic weights after unsupervised online learning with 100 POST neurons.

Chapter 4
Conclusion

In this dissertation, we have demonstrated unsupervised online learning with STDP learning algorithm using two types of NOR-type nonvolatile memory arrays.

As the first candidate, we have presented a SONOS gated-diode memory array and investigated current behavior as memory performance in the cell array. Then, we have proposed a neural network configuration and LTP / LTD pulse scheme for using the memory array as a synaptic device array, and we have analyzed the measurement results of applying LTP / LTD pulse scheme to the actual device. Afterwards, pattern learning capabilities of MNIST digit patterns were identified in single- (784 × 1) and multi-neuron (784 × 3) using software MATLAB simulations reflecting the measurement results of the device. The simulation results were analyzed based on the key factors of STDP unsupervised learning, input noise density (ρ_{noise}), synaptic weight margin (W_{margin}), and lateral inhibition factor [%]. The chapter then completed with a discussion of the critical issues that arise when using the SONOS gated-diode memory array as hardware-based neural network.

In the next chapter, we have proposed a TFT-type NOR flash memory array to address the problems that occurred earlier, and analyzed the structures of the device array. We have fabricated successfully a TFT-type NOR flash memory to be used in a synaptic device array for unsupervised online learning using the STDP learning algorithm. The fabrication process of the TFT-type NOR flash memory device was explained and its characteristics as an electronic synaptic device were analyzed. Because the device structure enables PGM / ERS operations to be performed by the gate and source voltages, it is possible to implement the STDP behavior of a synapse without any additional circuit configuration. In addition, WLs and BLs are configured as the crossbar types, enabling excellent scalability for large-scale synaptic arrays. Moreover, unsupervised learning and recognition with the STDP learning rule were demonstrated using the proposed memory array. Through MATLAB simulation, the learning of simple 3 × 3 dot-patterns and 28 × 28 MNIST handwritten digit patterns was done based on the STDP characteristics of the devices, and the pattern classification performance was investigated. It was confirmed that learning and classification are possible in single- (784 × 1) and multi-neuron (784 × 10) arrays. In addition, to improve the pattern recognition capability of the proposed synaptic device array, the homeostatic property was adopted to enhance the learning ability. To perform a high-level recognition task, the proposed LTD pulse scheme was also optimized, and the learning and inferencing capabilities of entire MNIST handwritten digit patterns were verified.

In summary, we have presented the feasibility of implementing a scalable hardware-based spiking neural network for unsupervised pattern recognition task using a TFT-type NOR flash memory array through system-level software MATLAB simulation.

Appendix A

Current-steering digital-to-analog conversion utilizing GIDL current in SONOS gated-diode memory string

The memristors have been actively studied for a variety of potential applications, such as neuromorphic computing chips [52]-[54], programmable analog circuits [55]-[58], and non-volatile memory [59]. This is because the memristors are made up to two terminals, which makes it advantageous to implement a crossbar array. However, when integrated with CMOS technology, there are several issues that arise, including poor reliability and increased process complexity. Also, select devices are always required for crossbar array operations. The SONOS gate-diode array, built on the Si wafer, can take advantage of these benefits while maintaining good reliability. Memory operation is performed by charging and discharging the carrier in the nitride layer of the oxide / nitride / oxide (O/N/O) gate dielectric stack. By controlling the charges stored in each cell with cell strings in an array, which is similar to controlling the memory state in memristors, we can modify the current flowing through the n^+ region to realize the current-steering digital-to-analog converter (DAC).

In this chapter, we examine the non-volatile memory properties of the SONOS gate-diode by characterizing the current from the cell string while programming or erasing the cell. It will also be shown that the current can be trimmed accurately by controlling the stored charge in each cell. Moreover, we suggest an approach which utilizes the SONOS gated-diode memory as a programmable analog circuit element. As an example, we demonstrate a current-steering DAC with binary-weighted programmed cells in a cell string.

As shown in figure 2.6, the I_{BL} of the SONOS gate-diode memory can be gradually increased through incremental step pulse programming (ISPP) [45]. When using FN tunneling as a programming mechanism, the gate voltage of the program pulse V_{pp} increases to a constant value after each program step. As programming voltage (V_{pp}) increases, the number of trapped electrons increases, increasing cell current. This allows the current of each cell can be set to the desired value accurately. Using these characteristics, we demonstrate a 6-bit binary weighted DAC implemented in the 1×6 SONOS gated-diode memory string. The I_{BL} for each cell in the string can be set to exact value using the ISPP method. The I_{BL} in each cell increases as V_{pp} increases. The six cells in the string are programmed with ISPPs to have I_{BL}s of 0.1, 0.2, 0.2, 0.4, 0.8, 1.6, and 3.2 nA, respectively, under read bias conditions, as shown in figure A.1.

Figure A.2 shows the measured analog output versus digital input characteristics of a configured 6-bit DAC. The current change step of 0.1 nA is clearly visible in the insert figure. Differential nonlinearity (DNL) and integral nonlinearity (INL) are extracted from the measured transfer characteristics of a 6-bit DAC and are illustrated in figure A.3. We also implemented a 4-bit DAC with the same string. The measured signal-to-noise and distortion ratios (SNDR) for 4-bit and 6-bit DACs implemented are 25.65 dB and 37.76 dB respectively, corresponding to the effective number of bits (ENOB) of 3.97 and 5.98 bits. The great linearity characteristic is due to the superior manageability and retention characteristics of I_{BL} and the good retention characteristics.

In conclusion, by programming SONOS gated-diode memory array using an incremental step pulse programming method, an area-efficient, 6-bit current-steering DAC was realized with near-ideal INL and DNL characteristics.

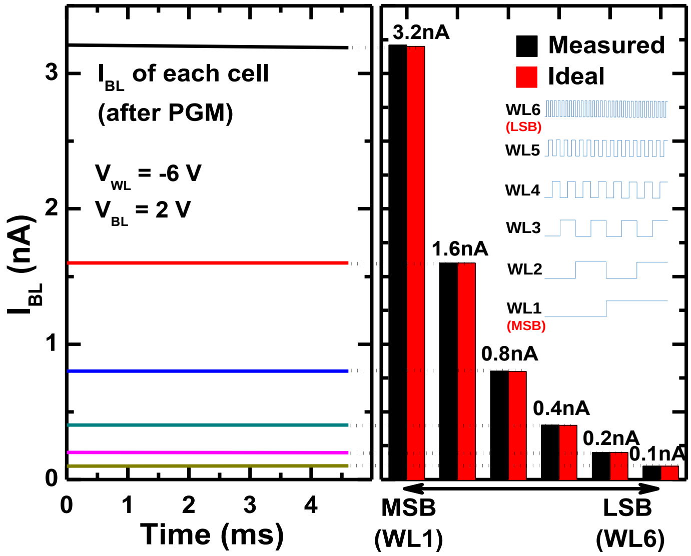


Figure A.1. Tuning the six-cells to have binary-weighted cell currents (0.1, 0.2, 0.4, 0.8, 1.6 and 3.2 nA) with the ISPP. [42].

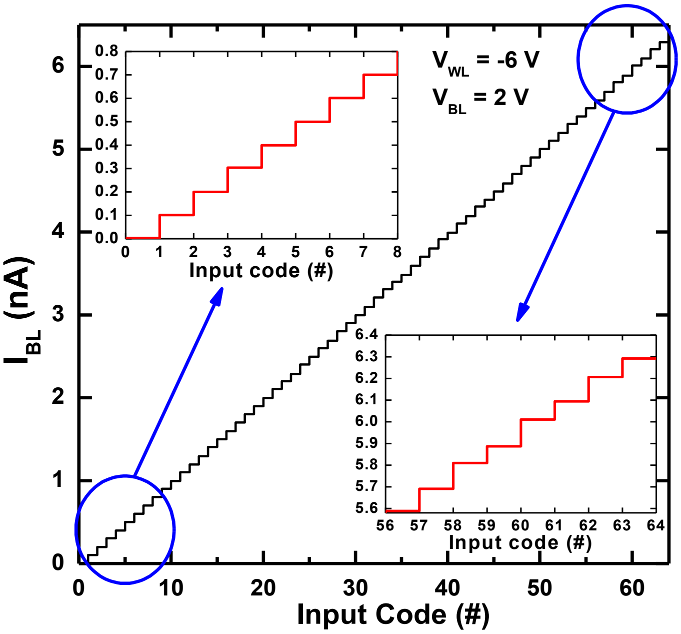


Figure A.2. Measured analog output versus input code for the 6-bit DAC [42].

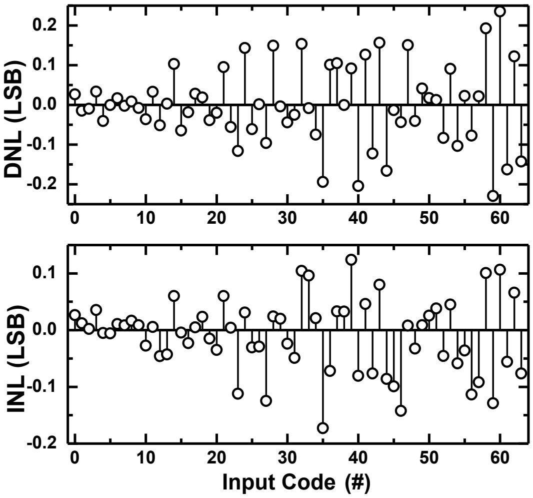


Figure A.3. The differential nonlinearity (DNL) and integral nonlinearity (INL) measured in least significant bit (LSB) as a function of the input code for a 6-bit DAC [42].

Bibliography

[1] G. E. Hinton, and R. R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks,” *Science*, vol. 313, pp. 504-507, 2006.

[2] T. Mikolov, M. Karafiát, L Burget, J Cernocký, and S. Khudanpur, “Recurrent neural network based language model,” *Interspeech*, vol. 2, pp. 45-48, 2010.

[3] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, pp. 1929-1958, 2014.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Proc. Adv. Neural Information Processing Systems (NIPS)*, pp. 1097-1105, 2012.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going Deeper with Convolutions,” *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-9, 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, Dec. 2016.

[7] G. Indiveri, and S.-C. Liu, “Memory and Information Processing in Neuromorphic Systems,” *Proc. IEEE.*, vol. 103, pp. 1379-1397, 2015.

[8] C.-S. Poon, and K. Zhou, “Neuromorphic silicon neurons and large-scale neural networks: challenges and opportunities,” *Front. Neurosci.*, vol. 22, p. 108, 2011.

[9] C.-H. Kim, S. Lim, S. Y. Woo, W.-M. Kang, Y.-T. Seo, S. T. Lee, S. Lee, D. Kwon, S. Oh, Y. Noh, H. Kim, J. Kim, J.-H. Bae and J.-H. Lee, “Emerging memory

technologies for neuromorphic computing," *Nanotechnology*, vol. 30, p. 032001, 2018.

[10] T. Masquelier, and S. J. Thorpe, "Unsupervised learning of visual features through spike timing dependent plasticity," *PLoS Comput. Biol*, vol. 3, pp. 247-257, 2007.

[11] G. W. Burr *et al.*, "Experimental demonstration and tolerancing of a large-scale neural network (165000 synapses) using phase-change memory as the synaptic weight element," *IEEE Trans. Electron Devices*, vol. 62, pp. 3498-3507, 2015.

[12] P. A. Merolla *et al.*, " A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, pp. 668-673, 2014.

[13] V. Milo, G. Pedretti, R. Carboni, A. Calderoni, N. Ramaswamy, S. Ambrogio, and D. Ielmini, "Demonstration of hybrid CMOS/RRAM neural networks with spike time/rate-dependent plasticity," *IEEE Int. Electron Devices Meeting (IEDM)*, 2016.

[14] Y. Lecun *et al*., "Gradient-based learning applied to document recognition," *Proc. IEEE.*, vol. 86, pp. 2278-2324, 1998.

[15] G.-Q. Bi, and M.-M. Poo, "Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type," *The Journal of Neuroscience*, vol. 18, pp. 10464-10472, 1998.

[16] M. F. Bear, "A synaptic basis for memory storage in the cerebral cortex," *Proc. Natl. Acad. Sci.*, vol. 93, pp. 13453-13459, 1996.

[17] H. Kim, S. Hwang, J. Park, and B.-G. Park, "Silicon synaptic transistor for hardware-based spiking neural network and neuromorphic system," *Nanotechnology.*, vol. 28, p. 405202, 2017.

[18] D. Querlioz, W. S. Zhao, P. Dolfus, J. O. Klein, O. Bichler, and C. Gamrat, “Bioinspired networks with nanoscale memristive devices that combine the unsupervised and supervised learning approaches,” *IEEE/ACM Int. Sym. on Nanoscale Architectures (NANOARCH)*, 2012.

[19] A. D. Almasi, S. Wozniak, V. Cristea, Y. Leblebici, and T. Engbersen, “Review of advances in neural networks: Neural design technology stack,” *Neurocomputing*, vol. 174, pp. 31-41, 2016.

[20] P. U. Diehl and M. Cook, “Unsupervised learning of digit recognition using spike-timing-dependent plasticity,” *Front. Comput. Neurosci*, vol. 9, p. 99, 2015.

[21] D. Querlioz, O. Bichler, P. Dollfus, and C. Gamrat, “Immunity to device variations in a spiking neural network with memristive nanodevices,” *IEEE Trans. Nanotechnol.*, vol. 12, pp. 288-295, 2013.

[22] D. Querlioz, O. Bichler, and C. Gamrat, “Simulation of a memristor-based spiking neural network immune to device variations,” *Int. Joint Conf. on Neural Networks (IJCNN)*, 2011.

[23] S. Ambrogio, N. Ciocchini, M. Laudato, V. Milo, A. Pirovano, P. Fantini, and D. Ielmini, “Unsupervised learning by spike timing dependent plasticity in phase change memory (PCM) synapses,” *Front. Neurosci.*, vol. 10, p. 56, 2016.

[24] S. Yu, “Neuro-Inspired Computing With Emerging Nonvolatile Memory,” *Proc. IEEE.*, vol. 106, pp. 260-285, 2018.

[25] S. Yu, Y. Wu, R. Jeyasingh, D. Kuzum, and H.-S. P. Wong, “An Electronic Synapse Device Based on Metal Oxide Resistive Switching Memory for Neuromorphic Computation,” *IEEE Trans. Electron Devices,* vol. 58, pp. 2729-2737, 2011.

[26] S. Kim, M. Lim, Y. Kim, H.-D. Kim, and S.-J. Choi, “Impact of Synaptic Device Variations on Pattern Recognition Accuracy in a Hardware Neural Network,”

Sci. Rep., vol. 8, p. 2638, 2018.

[27] S. Yu, B. Gao, Z. Fang, H. Yu, J. Kang, and H.-S. P. Wong, “A Low Energy Oxide-Based Electronic Synaptic Device for Neuromorphic Visual Systems with Tolerance to Device Variation,” *Adv. Mater.*, vol. 25, pp. 1774-1779, 2013.

[28] C.-H. Kim, S. Lee, S. Y. Woo, W.-M. Kang, S. Lim, J.-H. Bae, J Kim, and J.-H. Lee, “Demonstration of Unsupervised Learning With Spike-Timing-Dependent Plasticity Using a TFT-Type NOR Flash Memory Array,” *IEEE Trans. Electron Devices*, vol. 65, pp. 1774-1780, 2018.

[29] H.-S. Choi, D.-H. Wee, H. Kim, S. Kim, K.-C. Ryoo, B.-G. Park, and Y. Kim, “3-D Floating-Gate Synapse Array With Spike-Time-Dependent Plasticity,” *IEEE Trans. Electron Devices*, vol. 65, pp. 101-107, 2018.

[30] D. Ciregan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3642-3649, 2012.

[31] M. A. C. Maher, S. P. Deweerth, M. A. Mahowald, and C. A. Mead, “Implementing neural architectures using analog VLSI circuits,” *IEEE Trans. circuits and systems*, vol. 36, pp. 643-652, 1989.

[32] M. Mahowald and R. Douglas, “A silicon neuron,” *Nature*, vol. 354, pp. 515-518, 1991.

[33] D. Kuzum, S. Yu, and H.-S. P. Wong, “Synaptic electronics: materials, devices and application,” *Nanotechnology*, vol. 24, p. 382001, 2013.

[34] M. Prezioso, Y. Zhong, D. Gavrilov, F. Merrikh-Bayat, B. Hoskins, G. Adam, K. Likharev, and D. B. Strukov, “Spiking neuromorphic networks with metal-oxide memristors,” *IEEE Int. Symposium on Circuits and Systems (ISCAS)*, 2016.

[35] M. Chu, B. Kim, S. Park, H. Hwang, M. Jeon, B. H. Lee, and B.-G. Lee, “Neuromorphic hardware system for visual pattern recognition with memristor array and CMOS neuron,” *IEEE Tran. Industrial Electronics*, vol. 62, pp. 2410-

2419, 2015.

[36] P. Huang, D. Zhu, S. Chen, Z. Zhou, Z. Chen, B. Gao, L. Liu, X. Liu, and J. Kang, “Compact Model of HfOX-Based Electronic Synaptic Devices for Neuromorphic Computing,” *IEEE Trans. Electron Devices*, vol. 64, pp. 614-621, 2017.

[37] P. Peyman, E. Amat, and A. Rubio, “Reliability challenges in design of memristive memories,” *5th European Workshop on CMOS Variability (VARI)*, pp. 1-6, 2014.

[38] S. Hamdioui, H. Aziza, and G. Ch. Sirakoulis, “Memristor Based Memories: Technology, Design and Test,” *9th International Conference on Design & Technology of Integrated Systems in Nanoscale Era (DTIS)*, pp. 1-7, 2014.

[39] F. Alibart, S. Pleutin, O. Bichler, C. Gamrat, T. Serrano-Gotarredona, B. Linares-Barranco, and D. Vuillaume, “A memristive nanoparticle/organic hybrid synapstor for neuroinspired computing,” *Advanced Functional Materials*, vol. 22, pp. 609-616, 2012.

[40] C. Riggert, M Ziegler, D Schroeder, W. H. Krautschneider, and H. Kohlstedt, “MemFlash device: floating gate transistors as memristive devices for neuromorphic computing,” *Semiconductor Science and Technology*, vol. 29, 2014.

[41] M. Ziegler, M. Oberländer, D. Schroeder, W. H. Krautschneider, and H. Kohlstedt, “Memristive operation mode of floating gate transistors: A two-terminal MemFlash-cell,” *Applied physics letters*, vol. 101, p. 263504, 2012.

[42] C.-H. Kim, J.-W. Lee, J. Kim, J. Kim, and J.-H. Lee, “GIDL Characteristics in Gated Diode Memory String and its Application to Current-Steering Digital-to-Analog Conversion,” *IEEE Trans. Electron Devices*, vol. 62, pp. 3272-3277, 2015.

[43] A. Ahmad., A. Ahmad., F. Raissi, and H. Hajghassem, “Efficient hybrid CMOS-Nano circuit design for spiking neurons and memristive synapses with STDP,” *IEICE transactions on fundamentals of electronics, communications and*

computer sciences, pp. 1670-1677, 2010.

[44] Ju-Wan Lee, “Gated-diode memory cell and array utilizing GIDL current,” Ph.D. dissertation, Seoul national university, 2014.

[45] H.-T. Lue, T.-H. Hsu, S.-Y. Wang, E.-K. Lai, K.-Y. Hsieh, R. Liu, and C.-Y. Lu, “Study of incremental step pulse programming (ISPP) and STI edge effect of BE-SONOS NAND flash,” *IEEE International. Reliability Physics Symposium*, pp. 693-694, 2008.

[46] S.-T. Lee, S. Lim, N. Choi, J.-H. Bae, C.-H. Kim, S. Lee, D. H. Lee, T. Lee, S. Chung, B.-G. Park, and J.-H. Lee, "Neuromorphic Technology Based on Charge Storage Memory Devices," *IEEE Symposia on VLSI Technology and Circuits*, pp. 169-170, 2018.

[47] H. Meinhardt, A. Gierer, "Pattern formation by local self-activation and lateral inhibition," *Bioessays*, vol. 22, pp. 753-760, 2000.

[48] M. Schmuker, T. Pfeil, and M. P. Nawrot, "A neuromorphic network for generic multivariate data classification," *Proceedings of the National Academy of Sciences*, vol. 111, pp. 2081-2086, 2014.

[49] G. W. Burr, R. M. Shelby, A. Sebastian, S. Kim, S. Kim, S. Sidler, K. Virwani, M. Ishii, P. Narayanan, A. Fumarola, L. L. Sanches, I. Boybat, M. L. Gallo, K. Moon, J. Woo, H. Hwang, and Y Leblebici, "Neuromorphic computing using non-volatile memory," *Advances in Physics: X*, vol. 2, pp. 89-124, 2017.

[50] E. Marder, and Astrid A. Prinz, "Modeling stability in neuron and network function: the role of activity in homeostasis," *BioEssays*, vol. 24, pp. 1145-1154, 2002.

[51] S. Lee, C.-H. Kim, S. Oh, B.-G. Park, and J.-H. Lee, "Unsupervised online learning with multiple postsynaptic neurons based on spike-timing-dependent plasticity using a TFT-type NOR flash memory array," arXiv:1811.07115.

[52] J. J. Yang, D. B. Strukov, and D. R. Stewart, "Memristive devices for com

puting," *Nature Nanotechnology*, vol. 8, pp. 13-24, 2012.

[53] D. Kuzum, R. G. Jeyasingh, B. Lee, and H. S. P. Wong, "Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing," *Nano letters*, vol. 12, pp. 2179-2186, 2011.

[54] G. S. Rose, R. Pino, and Q. Wu, "A low-power memristive neuromorphic circuit utilizing a global/local training mechanism," *IEEE international joint conference on Neural networks (IJCNN)*, pp. 2080-2186, 2011.

[55] S. Shin, K. Kim, and S.-M. (Steve) Kang, "Memristor Applications for Programmable Analog ICs," *IEEE Trans. on Nanotechnology*, vol. 10, pp. 266-274, 2011.

[56] L. Gao, F. Merrikh-Bayat, F. Alibart, X. Guo, B. D. Hoskins, K.-T. Cheng, and D. B. Strukov, "Digital-to-Analog and Analog-to-Digital conversion with Metal Oxide Memristors for Ultra-Low Power," *IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, pp. 19-22, 2013.

[57] Y. V. Pershin, and M. D. Ventra, "Practical approach to programmable analog circuits with memristors," *IEEE Transactions on Circuits and Systems I*, vol. 57, pp. 1857-1864, 2010.

[58] W. Robinett, M. Pickett, J. Borghetti, Q. Xia, G. S. Snider, G. Medeiros-Ribeiro, and R. S. Williams, "A memristor-based nonvolatile latch circuit," *Nanotechnology*, vol. 21, p. 235203, 2010.

[59] S. H. Jo, K.-H. Kim, T. Chang, S. Gaba, and W. Lu, "Si memristive devices applied to memory and neuromorphic circuits," *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 13-16, 2010.

초 록

기존의 폰 노이만 컴퓨팅 구조는 높은 수준의 인지 응용분야에서 속도와 전력 소비 측면에서 불리한 구조를 지니고 있다. 따라서 이러한 문제를 해결하기 위해 새롭게 제안된 신경모방 컴퓨팅은 차세대 컴퓨팅 시스템으로 주목을 받고 있다.

본 논문에서는 두 가지 종류의 NOR-형 비휘발성 메모리 어레이를 신경모방 시스템의 시냅스 어레이로 사용하도록 제안한다. 전하 저장 층을 포함하는 게이트를 갖는 다이오드 메모리 어레이가 시냅스 모방 소자의 첫 번째 후보로 제안된다. 시뮬레이션을 통해 MNIST 손글씨 이미지 패턴의 학습 과정을 보여준다. 첫째로, 단일 뉴런 스트링 (784 × 1) 에서 스파이크 시점 의존 가소성 기반 학습이 시연된다. 그런 다음 측면 억제 기능을 사용하여 다중 뉴런 어레이 (784 × 3) 에서 스파이크 시점 의존 가소성에 기반한 학습을 시연한다. 한편, 적절한 학습을 위해 입력 잡음 밀도 (ρ_{noise}), 시냅스 가중치 간극 (W_{margin}), 측면 억제 계수 [%] 와 같은 스파이크 시점 의존 가소성 기반 비지도 학습의 주요 요인들에 대해 조사한다.

다음으로, 전하 저장 층을 포함하는 게이트를 갖는 다이오드 메모리

의 단점들을 극복하는 활성 채널의 절반을 덮는 플로팅 게이트가 포함된 박막 트랜지스터형 NOR 플래시 메모리 시냅스 모방 소자를 제안한다. 제안된 펄스 구동 방식을 활용하여 스파이크 시점 의존 가소성 동작에 필요한 장기 강화 및 약화 기능이 구현된다. 공정 제작된 메모리 어레이의 장기 강화 / 약화 특성을 반영하는 소프트웨어 시뮬레이션을 통해 스파이크 시점 의존 가소성 학습 규칙을 이용한 비지도 실시간 학습이 성공적으로 시연된다. 28 × 28 MNIST 손글씨 숫자 패턴의 학습 및 인식 과정을 제시한다.

결과적으로, 기존 CMOS 기술을 사용하여 제작된 시냅스 모방 소자로 구성된 하드웨어 기반 신경망을 시각 패턴 인식 시스템으로 사용하는 방안이 제안되었다.

주요어: 신경모방 시스템, 시냅스 모방 소자, 비지도 학습, 스파이크 시점 의존 가소성, 게이트를 갖는 다이오드 메모리, NOR-형 플래시 메모리.

학번: 2013-20779

List of Publications

Journals

1. Young-Tak Seo, Myoung-Sun Lee, **Chul-Heung Kim**, Sung Yun Woo, Jong-Ho Bae, Byung-Gook Park, and Jong-Ho Lee, “Si-based FET-type synaptic device with short-term and long-term plasticity using high-k gate stacks," *IEEE Transactions on Electron Devices*, vol. 66, no. 2, pp. 917-923, Jan. 2019.
2. Dongseok Kwon, Suhwan Lim, Jong-Ho Bae, Sung-Tae Lee, Hyeongsu Kim, **Chul-Heung Kim**, Byung-Gook Park, and Jong-Ho Lee, “Adaptive weight quantization method for nonlinear synaptic devices," *IEEE Transactions on Electron Devices*, vol. 66, no. 1, pp. 395-401, Dec. 2018.
3. **Chul-Heung Kim**, Suhwan Lim, Sung Yun Woo, Won-Mook Kang, Young-Tak Seo, Sung Tae Lee, Soochang Lee, Dongseok Kwon, Seongbin Oh, Yoohyun Noh, Hyeongsu Kim, Jangsaeng Kim, Jong-Ho Bae, and Jong-Ho Lee, “Emerging memory technologies for neuromorphic computing," *Nanotechnology*, vol. 30, no. 3, p. 032001, Nov. 2018.
4. Kyu-Bong Choi, Sung Yun Woo, Won-Mook Kang, Soochang Lee, **Chul-Heung Kim**, Jong-Ho Bae, Suhwan Lim, and Jong-Ho Lee, "A split-gate positive feedback device with an integrate-and-fire capability for a high-density low-power neuron circuit," *Frontiers in Neuroscience*, vol. 12, p. 704, Oct. 2018.
5. Suhwan Lim, Jong-Ho Bae, Jai-Ho Eum, Sungtae Lee, **Chul-Heung Kim**, Dongseok Kwon, Byung-Gook Park, and Jong-Ho Lee, "Adaptive learning rule for hardware-based deep neural networks using electronic synapse devices," *Neural Computing and Applications*, pp. 1-16, 2018.
6. **Chul-Heung Kim**, Soochang Lee, Sung Yun Woo, Won-Mook Kang,

Suhwan Lim, Jong-Ho Bae, Jaeha Kim, and Jong-Ho Lee, "Demonstration of unsupervised learning with spike-timing-dependent plasticity using a TFT-type NOR flash memory array," *IEEE Transactions on Electron Devices*, vol. 65, no. 5, pp. 1774-1780, May. 2018.

7. **Chul-Heung Kim**, Ju-Wan Lee, Junseok Kim, Jaeha Kim, and Jong-Ho Lee, "GIDL characteristics in gated-diode memory string and its application to current-steering digital-to-analog conversion," *IEEE Transactions on Electron Devices*, vol. 62, no. 10, pp. 3272-3277, Oct. 2015.

Conferences

1. Jangsaeng Kim, **Chul-Heung Kim**, Sung Yun Woo, Won-Mook Kang, Young-Tak Seo, Soochang Lee, Seongbin Oh, Jong-Ho Bae, Byung-Gook Park, and Jong-Ho Lee, "Initial synaptic weight distribution for fast learning speed and high recognition rate in STDP-based spiking neural network," *The 26th Korean Conference on Semiconductors*, Feb. 2019.
2. Soochang Lee, **Chul-Heung Kim**, Seongbin Oh, Byung-Gook Park, and Jong-Ho Lee, "Excitatory and inhibitory synaptic behavior of analog synapses using TFT-type NOR flash memory cells," *The 26th Korean Conference on Semiconductors*, Feb. 2019.
3. Won-Mook Kang, **Chul-Heung Kim**, Soochang Lee, Sung Yun Woo, Jong-Ho Bae, Byung-Gook Park, and Jong-Ho Lee, "A spiking neural network with a global self-controller for unsupervised learning based on spike-timing-dependent plasticity using flash memory synaptic devices," *The 26th Korean Conference on Semiconductors*, Feb. 2019.
4. Sung Yun Woo, Kyu-Bong Choi, Jangsaeng Kim, Won-Mook Kang, **Chul-Heung Kim**, Young-Tak Seo, Jong-Ho Bae, Byung-Gook Park, and Jong-Ho Lee, "Implementation of homeostasis functionality in neuron circuit

using split-gate device for spiking neural network," *The 26th Korean Conference on Semiconductors*, Feb. 2019.

5. Seongbin Oh, **Chul-Heung Kim**, Soochang Lee, Jangsaeng Kim, Byung-Gook Park, and Jong-Ho Lee, "Effect of pruning on energy-efficient spiking neural network trained by back-propagation," *The 26th Korean Conference on Semiconductors*, Feb. 2019.

6. Sung-Tae Lee, Suhwan Lim, Nagyong Choi, Jong-Ho Bae, **Chul-Heung Kim**, Soochang Lee, Dong Hwan Lee, Tackhwi Lee, Sungyong Chung, Byung-Gook Park, and Jong-Ho Lee, "Neuromorphic technology based on charge storage memory devices," *IEEE Symposia on VLSI Technology and Circuits*, pp. 169-170, Jun. 2018.

7. Suhwan Lim, Jong-Ho Bae, Jai-Ho Eum, Sungtae Lee, **Chul-Heung Kim**, Dongseok Kwon, and Jong-Ho Lee, "Hardware-based neural networks using a gated schottky diode as a synapse device," *IEEE International Symposium on Circuits and Systems (ISCAS)*, May. 2018.

8. Soochang Lee, **Chul-Heung Kim**, Byung-Gook Park, and Jong-Ho Lee, "Unsupervised learning of image patterns using multiple postsynaptic neurons based on spike-timing-dependent plasticity," *The 25th Korean Conference on Semiconductors*, Feb. 2018.

9. Seongbin Oh, **Chul-Heung Kim**, Soochang Lee, Byung-Gook Park, and Jong-Ho Lee, "Classification for grayscale images using supervised spike rate-based learning," *The 25th Korean Conference on Semiconductors*, Feb. 2018.

10. **Chul-Heung Kim**, Soochang Lee, Byung-Gook Park, and Jong-Ho Lee, "Demonstration of unsupervised learning with spike-timing-dependent plasticity using a SONOS gated-diode memory array," *The 25th Korean Conference on Semiconductors*, Feb. 2018.

11. **Chul-Heung Kim**, Suhwan Lim, Sung Yun Woo, Byung-Gook Park, and

Jong-Ho Lee, "Demonstration of unsupervised learning with spike-timing dependent plasticity for neuromorphic system," *The 24th Korean Conference on Semiconductors*, Feb. 2017.

12. Sung Yun Woo, **Chul-Heung Kim**, Kyu-Bong Choi, Suhwan Lim, Jaeha Kim, Byung-Gook Park, and Jong-Ho Lee, "Homeostatic neuron circuit using double-gate device for spiking neural network," *The 24th Korean Conference on Semiconductors*, Feb. 2017.
13. Suhwan Lim, Jong-Ho Bae, Jun-Mo Park, Jai-Ho Eum, Won-Mook Kang, **Chul-Heung Kim**, Myoung-Sun Lee, Sung Yun Woo, Byung-Gook Park, and Jong-Ho Lee, "Synaptic devices based on reconfigurable gated schottky diodes for highly-linear potentiation," *The 24th Korean Conference on Semiconductors*, Feb. 2017.
14. Sung Yun Woo, **Chul-Heung Kim**, and Jong-Ho Lee, "Synaptic device based on gated-diode memory string using GIDL current for neuromorphic system," *The 23rd Korean Conference on Semiconductors*, Feb. 2016.
15. Chang-Hee Kim, **Chul-Heung Kim**, Yoonki Hong, Jongmin Shin, Kyu-Bong Choi, In-Tak Cho, Chul-Ho Won, Do-Kywn Kim, Jung-Hee Lee, and Jong-Ho Lee, “AlGaN/GaN MISFET gas sensor having a horizontal floating gate,” *International Conference on Electronics, Information and Communication (ICEIC)*, Jan. 2015.

Patents

1. Jong-Ho Lee, **Chul-Heung Kim,** and Suhwan Lim, "Vertical neuromorphic devices stacked structure and array of the structure."
   ▪ US Patent US 10,103,162 B2, Oct. 2018
2. Jong-Ho Lee, **Chul-Heung Kim,** and Sung Yun Woo, "Reconfigurable devices, device array for neuromorphic."
   ▪ Korean Patent filed 10-2016-0149727, Nov. 2016
3. Jong-Ho Lee, **Chul-Heung Kim,** and Sung Yun Woo, "Neuromorphic device with excitatory and inhibitory functionalities."
   ▪ US Patent US 9,431,099 B2, Aug. 2016
   ▪ Korean Patent KR 10-1695737, Jan. 2017

Honors

1. Silver Prize, The 23rd Humantech Thesis Contest, Samsung Electronics, Feb. 2017.