Ph.D. DISSERTATION

Demonstration of Hardware-based Spiking Neural Network Using an AND-type Flash Memory Array Architecture

AND-형 플래시 메모리 어레이를 활용한 하드웨어 기반 스파이킹 뉴럴 네트워크 구현

by

WON-MOOK KANG

February 2022

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

Demonstration of Hardware-based Spiking Neural Network Using an AND-type Flash Memory Array Architecture

AND-형 플래시 메모리 어레이를 활용한 하드웨어 기반 스파이킹 뉴럴 네트워크 구현

지도교수 이 종 호

이 논문을 공학박사 학위논문으로 제출함

2022 년 2 월

서울대학교 대학원

전기정보공학부

강 원 묵

강원묵의 공학박사 학위논문을 인준함

2022 년 2 월

위 원 장 : 박 병 국 (인)

부위원장 : 이 종 호 (인)

위 원 : 김 재 하 (인)

위 원 : 김 재 준 (인)

위 원 : 조 성 재 (인)

Demonstration of Hardware-based Spiking Neural Network Using an AND-type Flash Memory Array Architecture

by

Won-Mook Kang

Advisor: Jong-Ho Lee

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy
(Electrical and Computer Engineering)
in Seoul National University
February 2022

Doctoral Committee:

Professor Byung-Gook Park, Chair
Professor Jong-Ho Lee, Vice-Chair
Professor Jaeha Kim
Professor Jae-Jun Kim
Associate Professor Seongjae Cho

ABSTRACT

Neuromorphic engineering aims to implement a brain-inspired computing architecture as an alternative paradigm to the von Neumann processor. In this work, hardware-based neural networks that enable on-chip training using a thin-film transistor-type AND flash memory array architecture are designed. The synaptic device constituting the array is characterized by a doped *p*-type body, a gate insulator stack composed of SiO_2 / Si_3N_4 / Al_2O_3, and a partially curved poly-Si channel. The *p*-body reduces the circuit burden on the high voltage driver required for both the source and drain lines when changing the synaptic weights. The high-κ material included in the gate insulator stack helps to lower the operating voltage of the device. As the device scales down, the structural characteristics of the device have the potential to increase the efficiency of the memory operation and the immunity to the voltage drop effect that occurs in the bit-lines of the array. In the AND array architecture using fabricated synaptic devices, a pulse scheme for selective memory operation is proposed and verified experimentally. Based on the

measured characteristics of the fabricated synaptic devices and arrays, we design two types of hardware-based spiking neural networks (SNNs) according to the learning purpose. First, we propose a hardware-based SNN for unsupervised learning with spiking-timing-dependent plasticity (STDP) learning rule. The designed network does not use the pulses generated by the external circuitry, but the necessary pulses are generated in each spike neuron circuit. In this architecture, the STDP rule is implemented by the effective pulse scheme for using poly-silicon AND arrays. With the proposed pulse scheme and SNN, 91.63% of recognition accuracy is obtained in MNIST handwritten digit pattern learning using 200 output neurons. Second, we propose a hardware-based SNN for supervised learning with a direct feedback alignment (DFA) learning rule. Due to the DFA algorithm, which does not need to have the same synaptic weight in the forward path and backward path, the AND array architecture can be utilized in designing an efficient on-chip training neural network. Pulse schemes suitable for the proposed AND array architecture are also devised to implement the DFA algorithm in neural networks. In the system-level simulation, the recognition accuracy of up to 97.01% is obtained

in the MNIST pattern learning task based on the proposed pulse scheme and computing architecture. In addition, we propose and verify the integration fabrication method of the proposed synaptic array and complementary metal-oxide-semiconductor (CMOS) circuits. Here, the CMOS circuits include either an integrate-and-fire circuit or a circuit that can change the width or amplitude of the spike signal. The proposed integration fabrication method has the advantage of reducing the number of masks and steps due to the shared process of the synaptic array and CMOS circuit. The proposed integration fabrication method is significant because it presents a methodology for efficient implantation of hardware-based neural networks as well as verification of excellent compatibility of the proposed synaptic device with CMOS.

Keywords: hardware-based spiking neural network, flash memory synaptic device, AND-type array, on-chip training, unsupervised learning, supervised learning, neuron circuit.

Student number: 2015-20881

CONTENTS

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Neuromorphic computing

An artificial neural network (ANN) is a computational model inspired by the brain that excels at complex tasks like recognition, prediction, and classification. Unlike conventional von Neumann processors, the human brain that the ANN attempts to imitate is known as a very energy-efficient computing machine using massively parallel computation and is designed to adapt to the external environment. For this reason, ANN researches are underway in various ways to overcome the limits of traditional computing architectures in specific applications. Various ANN models have been continuously studied to date depending on the application, naturally leading to studies for designing efficient computing architectures to utilize neural networks. Software-based deep neural networks (SW-DNNs) such as convolutional neural network (CNN) and recurrent neural networks (RNN) are representative computing models, which use well-defined mathematical and

analytic algorithms [1-6]. These advanced computing models can offer excellent learning performance, but it is difficult to obtain a highly efficient solution in designing computing architectures for real-time applications. In particular, the vector-by-matrix multiplication (VMM) for the main operation occupies a significant part of the SW-DNNs' computational tasks, which leads to enormous power consumption. For this reason, many studies on how to efficiently perform a large amount of VMM have been recently carried out, and one of the most popular methods is to utilize a crossbar array of an electronic synaptic device [7-9]. As shown in Fig. 1.1, in the crossbar array of the electronic synaptic device, an output signal can be expressed as the current flowing in a predetermined direction of the array when an input signal and a weight are represented by an input voltage and a conductance (G) of the electronic synaptic device, respectively. Since utilizing the crossbar array composed of electronic synaptic devices can reduce power consumption and improve computational speed, the need for hardware-based deep neural networks (HW-DNNs) has emerged. However, the efficient implementation of full HW-DNNs, which involves complex computations such as derivatives, is

still challenging simply by configuring arrays of electronic synaptic devices [10]. This is because if other complex computations, such as weight updates, require the support of software, there are still bottlenecks in power consumption and computational speed, which can greatly reduce the advantages of HW-DNNs. In addition, in representing the synaptic weight as the conductance of the electronic synaptic device, the degradation of system performance caused by the inaccuracy in weight transformation is difficult to solve completely [11-12]. Therefore, electronic circuits that are compatible with guaranteed electronic synaptic devices are essential, and appropriate learning rules for enabling efficient hardware implementation should be supported.

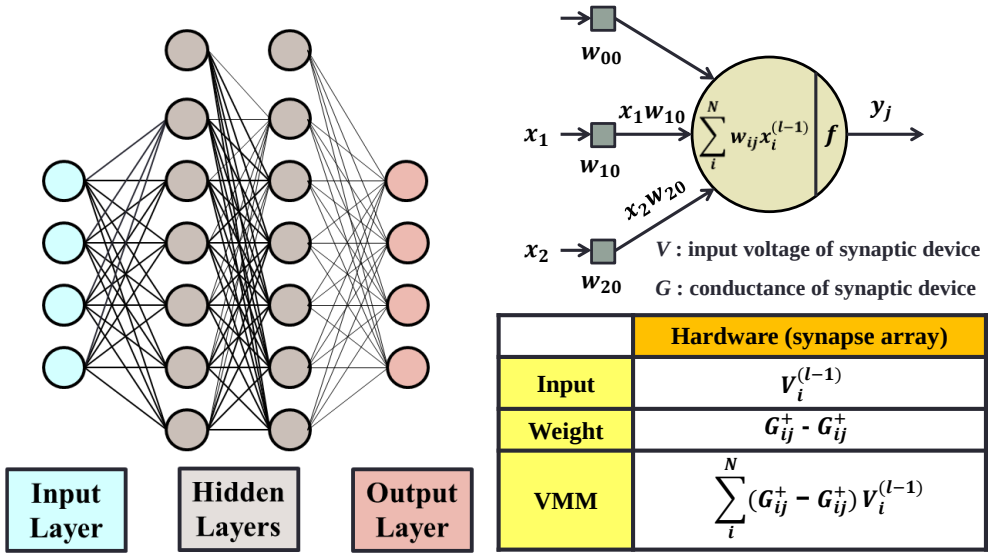


Fig. 1.1. Utilization of crossbar arrays composed of electronic synaptic devices for efficient vector-by-matrix multiplication.

1.2 Hardware-based spiking neural network

Recently, a computing model called spiking neural network (SNN) with integrated neuron circuits and electronic synaptic devices has emerged as a candidate for another efficient ANN model [13-19]. SNN aims to approach biological energy efficiency using learning protocols implemented by synapses and neurons, which are components of the human brain. In SNN, data is encoded as spike patterns and transferred between neurons, relying on learning rules for higher computational efficiency than conventional methods. Neuromorphic computing performed in this kind of network has local data computing characteristics to improve the power and computational efficiency of the computing architecture.

Meanwhile, in-memory computing has raised expectations for the realization of the hardware-based SNN by enabling VMM while utilizing the physical properties of a crossbar synaptic array [20-22]. For this reason, many studies have proposed various types of non-volatile electronic synaptic devices that can constitute the crossbar array as shown in Fig. 1.2 [23-26]. An electronic synaptic device aims to implement synaptic plasticity, a rule of change in synaptic weight

that allows learning by experience as in the nervous system. Among the many candidates for electronic synaptic devices, two-terminal electronic synaptic devices such as phase-change random access memory (PCRAM) [27-29], resistive random-access memory (RRAM) [30-33], spin-based memory [34-37], and ferroelectric memristor [38] have attracted considerable attention from researchers in recent years. These kinds of two-terminal electronic synaptic devices represented by memristors have exhibited problems, such as sneak path currents, device characteristic variation, and poor reliability, despite advantages of scalability and fast operation. On the other hand, the field-effect transistor (FET)-based electronic synaptic device utilizing a charge trap layer is an attractive candidate with many advantages, such as low synaptic current, good reliability, high integration density, a large-conductance window, and process compatibility with CMOS [39-42]. Also, by operating the FET-type electronic synaptic device in the saturation region, it is possible to minimize the change in the synaptic current due to the voltage change applied to the device. Thus, any undesired voltage drop that can occur on the wiring in the array has little effect on the synaptic current. In the case of the three-terminal

electronic synaptic device, unlike the two-terminal, the degree of integration and how the array works depend on the configuration of the array. Therefore, it is important to determine the operation scheme for parallel computation while designing an appropriate array configuration of the electronic synaptic devices according to the type of neural network.

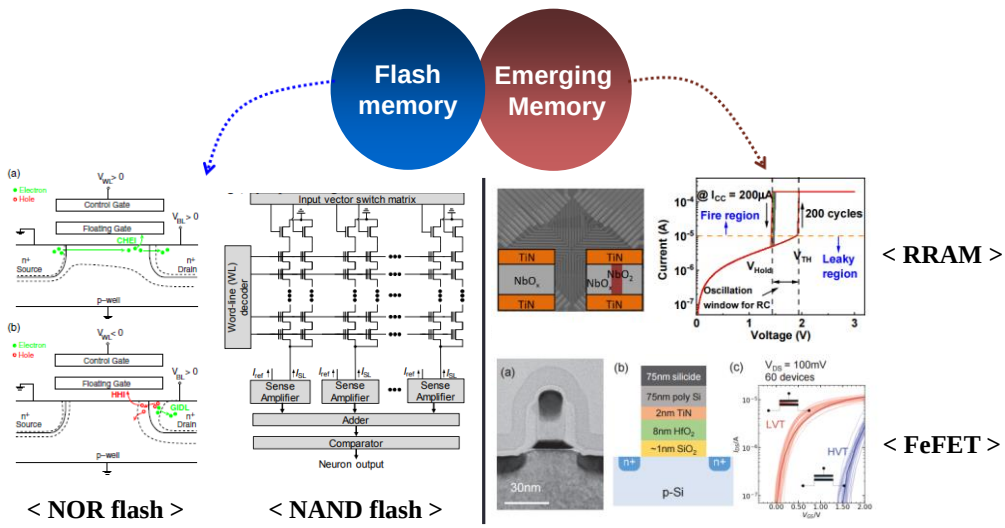


Fig. 1.2. Several candidates for electronic synaptic device that can form crossbar arrays.

1.3 Purpose of research

The learning algorithm used to implement the neuromorphic computing technology varies depending on the application but can be broadly classified into two types according to the data received. Firstly, unsupervised learning, which has purposes such as clustering and association, proceeds learning based only on input data. Among the various learning algorithms for unsupervised learning available in SNNs, spike-timing-dependent plasticity (STDP) and spike-rate-dependent plasticity (SRDP) are representative examples of intuitive synaptic weight change rules [43-46]. On the other hand, supervised learning, which has the purpose of regression or classification, learns by receiving input data and output label information corresponding to the correct answer of the input data as input. A representative learning algorithm for supervised learning is a backpropagation (BP) algorithm. Recently, several researchers have shown the possibility that a BP [47], which exhibits outstanding performance in conventional DNNs [48-50], can also be utilized in SNNs while achieving excellent accuracy [51-53]. However, although methods using the BP algorithm are very effective means on the von Neumann

architecture, they may become inefficient due to weight transport problems when extended to large networks [54, 55]. This includes requirements for symmetric synaptic connections between for- and back-ward paths throughout the network. To overcome the limitations of this training paradigm, various new training methods that relax the synaptic weight symmetry constraints have been recently reported.

In this work, we design hardware-based SNNs to implement unsupervised learning and supervised learning using an AND flash memory synaptic array. First, a thin-film transistor (TFT)-type synaptic device based on flash memory technology is introduced, and its memory characteristics are analyzed. The proposed device has structural differences to improve memory operation and current saturation characteristics compared to the conventional TFT devices. And then, a representative pulse scheme for realizing the selective memory operation of the proposed synaptic devices in the AND array is described and experimentally implemented. Using the verified results, we devise efficient pulse schemes for unsupervised learning and supervised learning implementation in the designed hardware-based SNNs. The designed SNNs aim to implement on-chip training.

Also, system-level simulations are performed based on the characteristics of the fabricated synaptic array and the operation of the SNNs designed for each purpose. Lastly, to fabricate a neuromorphic chip for SNN implementation, we propose an efficient integration fabrication method of the proposed synaptic array and peripheral complementary metal-oxide-semiconductor (CMOS) circuits. The proposed fabrication method aims to integrate the synaptic array and CMOS circuit on a single wafer while reducing the number of masks required. And then, we verify the feasibility of the proposed fabrication method by analyzing the operation of the synaptic array and CMOS circuit and required to implement the hardware-based SNNs.

1.4 Dissertation outline

The following is the structure of this dissertation. Chapter 1 provides an overview of neuromorphic computing and hardware-based SNNs. Chapter 2 describes the proposed TFT-type AND flash memory array. This chapter includes the structure, fabrication process, and characteristics of the synaptic device. Moreover, the measurement results as a synaptic device and array are also presented. Chapter 3 deals with the hardware-based SNN designed to implement the unsupervised learning, the pulse scheme to utilize the proposed synaptic array, and the pattern recognition result reflecting the fabricated device characteristics. Likewise, chapter 4 described the hardware-based SNN designed to implement the supervised learning, the pulse scheme to utilize the proposed synaptic array, and the pattern recognition result reflecting the device characteristics. Chapter 5 present the integration fabrication of the proposed synaptic array and CMOS circuit required for hardware-based SNN implementation. This includes detailed key fabrication steps and experimentally verification of the operation of the synaptic array and CMOS circuit. Finally, chapter 6 concludes this dissertation with a summary.

Chapter 2
TFT-type AND flash memory array

2.1 Device structure and fabrication

Fig. 2.1 (a) and (b) show a bird's eye view of a TFT-type AND flash memory array and a schematic cross-sectional view cut in the word-line (WL) direction, respectively. A scanning electron microscope (SEM) image of the fabrication synaptic device is also shown in Fig. 2.1 (c). In the fabricated AND memory array, the source-line (SL) and drain-line (DL), as bit-line (BL), are formed in parallel, and the WLs are perpendicular to both of them. As shown in Fig. 2.1 (a) and (b), a poly-Si body (*p*-body) line (PL) is formed and electrically separated between the SL and DL. The PL is formed by implanting boron ions and has a doping concentration of 2×10^{17} cm^{-3}. Also, the heights of the source (S) and drain (D) electrodes are lower than those of the SiO_2 spacers on both sides of the *p*-body. So, the channel formed on the partial SiO_2 spacers has a curved shape. Moreover, a Al_2O_3 / Si_3N_4 / SiO_2 (A/N/O) gate insulator stack between the TiN gate and the

polysilicon channel serves to store the synaptic weights. The thickness of each layer included in the gate insulator stack is represented in a transmission electron microscope (TEM) image in Fig. 2.1 (d). Note that the channel width (W) and *p*-body length (L_p) are both 0.5 μm. If the W can be scaled to the minimum feature size (F), the device in the memory array can be scaled down to 8 F^2.

Fig. 2.2 represents the schematic cross-sectional views of the key fabrication steps, and overall process flow of the proposed TFT-type synaptic device. The devices were fabricated on a 6-in Si wafer using CMOS process technology. First, a marker pattern is formed by patterning for the photolithography step of the subsequent process (first mask). And then, a 150-nm thick poly-Si layer was formed on a 350-nm-thick SiO_2 layer, which was grown thermally via a wet oxidation process. After boron implantation was performed, the poly-Si layer was patterned (second mask). A 30-nm-thick SiO_2 film was deposited and anisotropically etched to form SiO_2 film spacers on both sides of the patterned poly-Si layer. Then, the deposition of an n^+-doped poly-Si was followed by a chemical mechanical polishing (CMP) process. Additionally, through chemical dry etching (CDE), the thickness of

the n^+-doped poly-Si was lowered so that the side of the SiO_2 film spacer was partially exposed. Then, a 12-nm-thick amorphous Si layer was deposited as a channel material and poly-crystalized by annealing at 600 ℃ for 24 hours. After channel patterning (third mask), the boron ion was implanted in the contact area of the *p*-body (fourth mask), which was followed by rapid temperature processing (RTP) for activation. Then, device isolation patterning (fifth mask) was performed, and the A/N/O gate insulator stack was deposited. A 27-nm-thick layer of TiN was then deposited as a gate and patterned (sixth mask). After the deposition of a 200-nm-thick tetraethyl orthosilicate (TEOS), contact holes for the gate, S, D, and *p*-body were formed (seventh mask). Finally, Ti/TiN/Al/TiN metal wires were formed through sputtering and patterned (eighth mask).

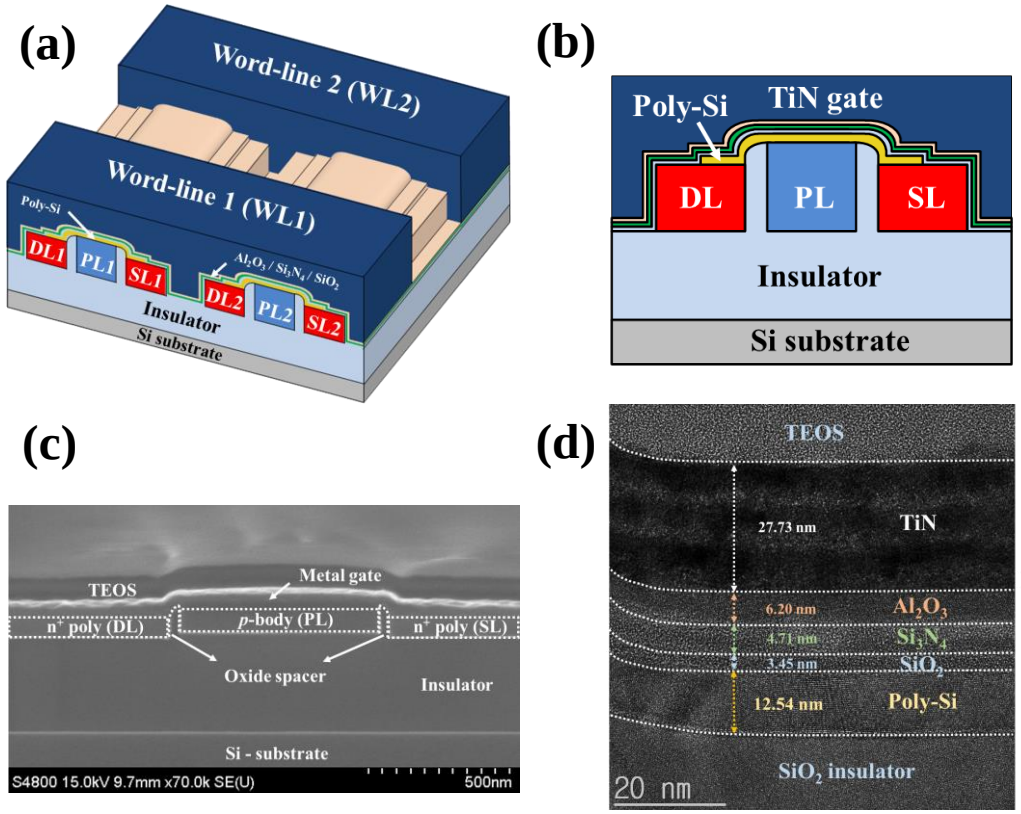


Fig. 2.1. (a) A bird's eye view of a TFT-type AND flash memory array. (b) A schematic cross-sectional view and (c) a SEM image of a fabricated TFT-type single synaptic device. (d) A TEM image including the gate insulator stack (A/N/O) in the fabricated synaptic device.

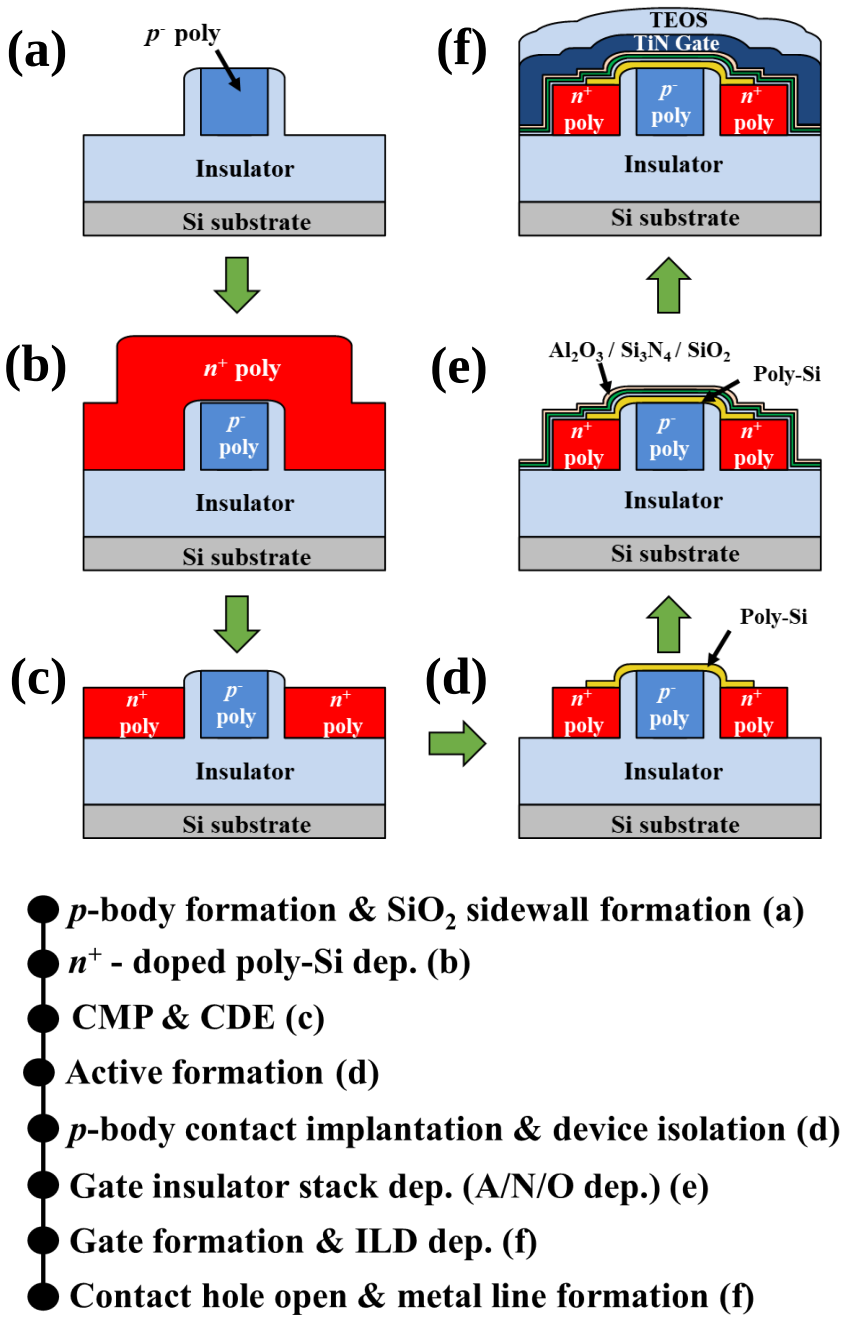


Fig. 2.2. Schematic cross-sectional views of the key fabrication steps, and overall process flow of the proposed synaptic device.

2.2 Characteristics of the device

The proposed TFT-type flash memory device has several advantages making it usable as a synaptic device. The low mobility of the poly-Si channel provides the possibility to reduce power consumption by reducing the current in the synaptic device. The Al_2O_3 used in the gate insulator stack is a material with high-κ characteristics, which can lower the voltage required for memory operation as well as the read operation of the synaptic device. Also, the *p*-body isolated between the S and D electrodes is responsible for providing holes during the erase (ERS) operation of the synaptic device and enables hole injection into the charge trap layer through direct biasing. This method is differentiated from the ERS operation using gate-induced drain leakage current (GIDL) through band-to-band tunneling in conventional 3-terminal devices without a *p*-body, which reduces the burden of designing the high voltage drivers required for both the SL and DL. The structural characteristics of the proposed synaptic device with a curved channel in the part of the side SiO_2 spacers can maximize its merits in various ways, especially when the device is scaled down.

Fig. 2.3 shows the simulated drain current (I_D) – gate voltage (V_{GS}) characteristics each time a device is programmed and erased, sequentially. Fig. 2.3 (a) and (b) represent the simulated results of a device with only a flat channel and a device with a partially curved channel, respectively. In the program (PGM) operation, a PGM pulse with an amplitude of 11 V and a width of 100 μs is applied to the gate of the device in the initial state, and the biases of the S, D, and *p*-body maintain 0 V. In contrast, in the ERS operation, an ERS pulse with 12 V amplitude and 10 ms width is applied to the *p*-body of the device in the stated programmed state. During the ERS operation, the gate maintains 0 V, and the S and D are floated. The simulation results indicate that the two devices have a difference in the memory window, even though the same memory operation is performed. This result is caused by an increase in memory efficiency due to the fact that the electric field from the gate to the channel direction can be concentrated in a partially curved channel. That is, in the charge trap layer located on the curved channel portion, a larger amount of electrons or holes can be stored under the same PGM/ERS conditions when compared to the case of the flat channel. For the same reason, a

partially curved channel in the device improves the current saturation characteristics of the output curve. Fig. 2.4 (a) and (b) show the output resistances (r_o) extracted from the simulated I_D – drain voltage (V_{DS}) characteristics of the device with only a flat channel and the device with a partially curved channel, respectively. The device with the partially curved channel has more immunity to the short channel effect, which occurs as the length of the device is scaled-down, compared to the device with only the flat channel. This is because the gate bias concentrated in the curved channels located near the S and D can effectively suppresses the electric field penetration from the D in the scaled down device. This increases r_o and reduces the current change due to the drain voltage change. Therefore, the proposed device structure significantly reduces the change in drain current due to an unwanted voltage drop across the parasitic resistance in the BL that occurs when the synaptic devices are configured as an array.

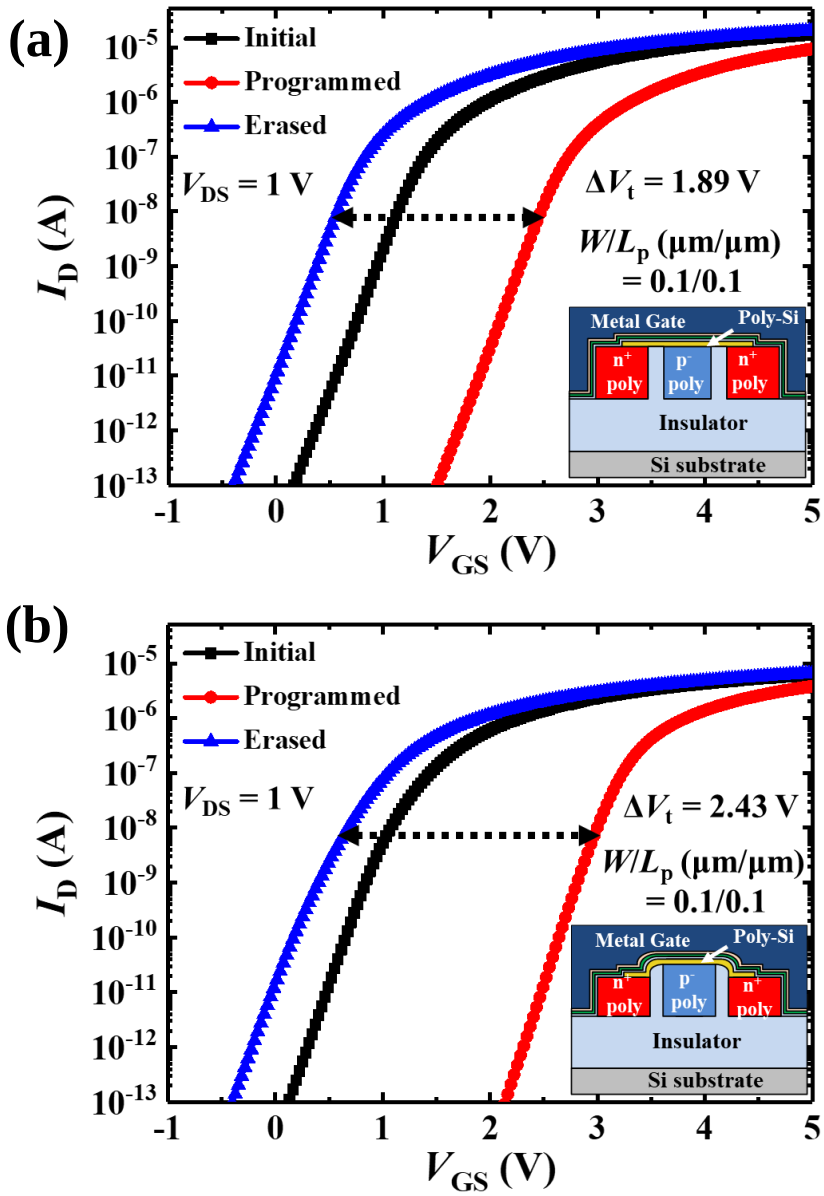


Fig. 2.3. Simulated I_D – V_{GS} characteristics each time (a) a device with only a flat channel and (b) a device with a partially curved channel are programmed and erased sequentially under the same memory conditions.

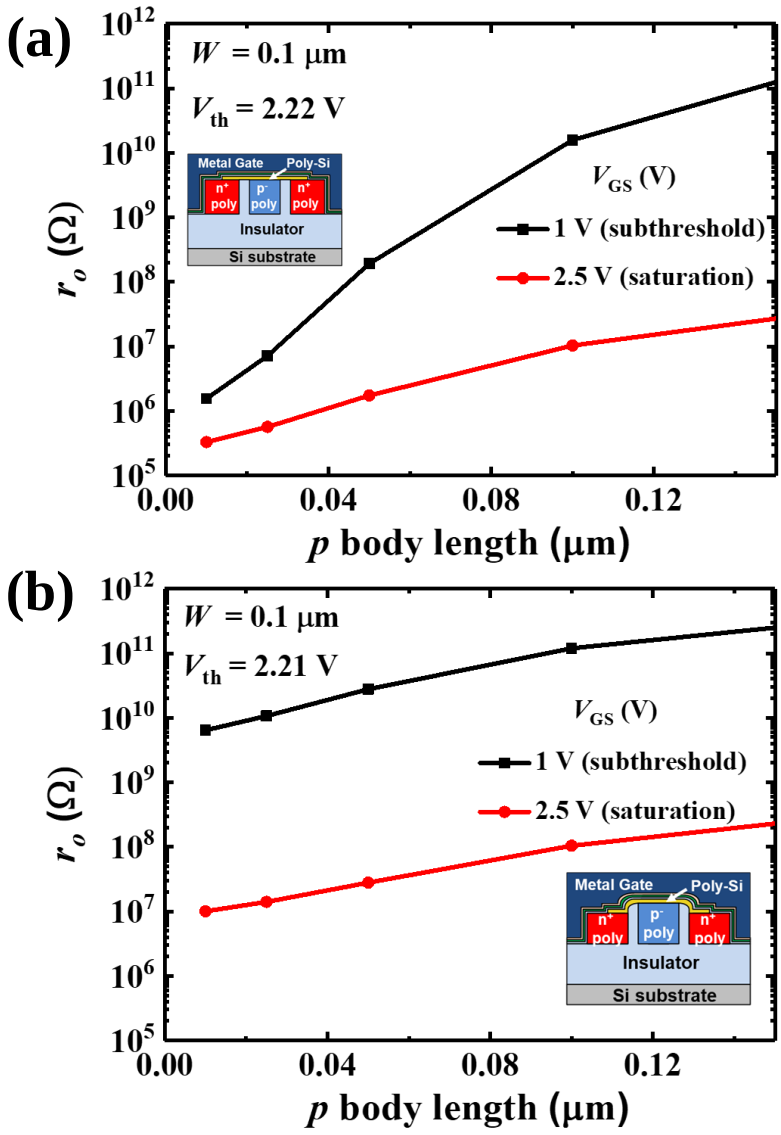


Fig. 2.4. Output impedances extracted from the simulated I_D – V_{DS} characteristics of (a) a device with only a flat channel and (b) a device with a partially curved channel.

2.3 Measurement results as a synaptic device

The measured $I_D - V_{GS}$ characteristics of a specific device in a fabricated TFT-type AND flash memory array as a variable of the V_{DS} are shown in Fig. 2.5. The inset figure shows the measured diode characteristics between the *p*-body and BL electrodes of the device. Fig. 2.6 shows the I_D–V_{DS} characteristics of the fabricated device as a parameter of V_{GS}. Through the I_D–V_{DS} characteristics of the device, the range of the BL voltage (V_{BL}), enabling the saturation region operation of the device in the array, can be determined by considering the specific read voltage (V_{read}) of the synaptic device. If the V_{read} of the synaptic device in the array is selected to be 2 V, the minimum V_{BL} needs to be 1.2 V, at which current saturation occurs. The output impedances (r_o) under several V_{GS} conditions are specified in the inset table. Fig. 2.7 (a) and (b) show the $I_D - V_{GS}$ characteristics measured when an identical PGM or ERS pulse under specific conditions is applied several times to the fabricated synaptic device. The bias conditions for the program and erase are described in the inset. The figures indicate that the threshold voltage (V_{th}) of the device can be shifted by trapping the electrons or holes in the Si_3N_4 layer of the gate

insulator stack. Here, the V_{th} shift indicates that the conductance of the device, which can be expressed as the synaptic weight, can be adjusted by the amount and the polarity of the charges stored in the charge trap layer. In addition, the fact that the device can have various conductance values depending on the number of pulses of the specific polarity applied means that multi-state synaptic weights are possible in on-chip training.

Fig. 2.8 represents the long-term potentiation (LTP) and long-term depression (LTD) characteristics obtained from the measured conductance each time the identical ERS and PGM pulses are applied to the fabricated synaptic device, respectively. The inset figure shows the memory window of the fabricated device used to obtain the LTP and LTD curves. The LTP characteristics are obtained by measuring the conductance of the synaptic device whenever 54 identical pulses (V_{ERS} = 10 V, t_{ERS} = 10 ms) are applied to the p-body under the ERS condition (V_{GS} = 0 V, V_S & V_{DS} = floating). Similarly, the LTD characteristics are obtained by measuring the conductance of the synaptic device whenever 54 identical pulses (V_{PGM} = 7 V, t_{PGM} = 100 μs) are applied to the gate under the PGM condition (V_S &

V_{DS} & $V_{p\text{-body}}$ = 0 V). Note that V_{GS} is 2.2 V and V_{DS} is 2 V in the read operation. As shown in Fig. 2.8, the LTP curve has a relatively linear characteristic compared to the LTD curve, the reason for which can be explained as follows. During the potentiation of the device, the amount of holes stored in the charge trap layer increases logarithmically with the number of potentiation pulses. This means that the effective value of V_{GS} increases logarithmically with the number of potentiation pulses. If the read voltage is set in a specific range in which the current exponentially increases with the gate voltage, the LTP curve can have a linear characteristic. Unfortunately, the LTD curve has a non-linear characteristic because it cannot take the effect of the logarithmic and exponential functions canceling each other out. For this reason, not only the pulses used for the PGM and ERS operations but also the bias condition of the read operation can change the LTP and LTD characteristics. The linearity of the LTP and LTD curves affects the inference accuracy of the neural network because it is related to the degree of the synaptic weight change. In chapter 3 and 4, the LTP/LTD characteristics, as specified in Fig. 2.8, are reflected in the system-level simulation of the proposed HNNs. Fig. 2.9

shows the LTP/LTD characteristics obtained under various read conditions in a specific memory window of the synaptic device. As stated above, it can be seen that the linearity of the LTP/LTD curve varies depending on the read condition of the synaptic device. In other words, the read condition of the device can determine the dynamic range of the synaptic weight that can be utilized under the same memory condition. When the read voltage of the device is set to a voltage lower than the linear region, a relatively wide range of synaptic weights can be used under the same memory pulse condition. However, if the read voltage of the device is included in the subthreshold region, the conductance of the device changes exponentially according to the read voltage. This means that it may be vulnerable to variations caused by the external environment such as temperature. On the other hand, the linear region of the device can have immunity to these kinds of variations. However, to utilize a relatively wide range of synaptic weights while using the read voltage in the linear region, the amplitude or width of the pulses required for memory operation should be large, or the number of pulses should be used more. Fig. 2.10 represents measured cycle-to-cycle variation characteristics in the

LTP/LTD of the fabricated synaptic device. The LTP characteristics are obtained by measuring the conductance of the synaptic device whenever 35 identical pulses (V_{ERS} = 10 V, t_{ERS} = 10 ms) are applied to the *p*-body under the ERS condition (V_{GS} = 0 V, V_S & V_{DS} = floating). Similarly, the LTD characteristics are obtained by measuring the conductance of the synaptic device whenever 35 identical pulses (V_{PGM} = 8 V, t_{PGM} = 100 μs) are applied to the gate under the PGM condition (V_S & V_{DS} & $V_{p\text{-body}}$ = 0 V). Note that V_{GS} and V_{DS} are both 2 V in the read operation. Fig. 2.10 (a) shows the results obtained by continuously measuring the LTP/LTD characteristics 10 times under the conditions specified above. Fig. 2.10 (b) represents the overlapping LTP/LTD characteristics of 10 cycles. The value of $(\sigma/\mu)_{max}$ is analyzed to be 0.062 and 0.097 in LTP and LTD operation, respectively. The analyzed values demonstrate the excellent cycle-to-cycle variation characteristics of the LTP and LTD in the fabricated device.

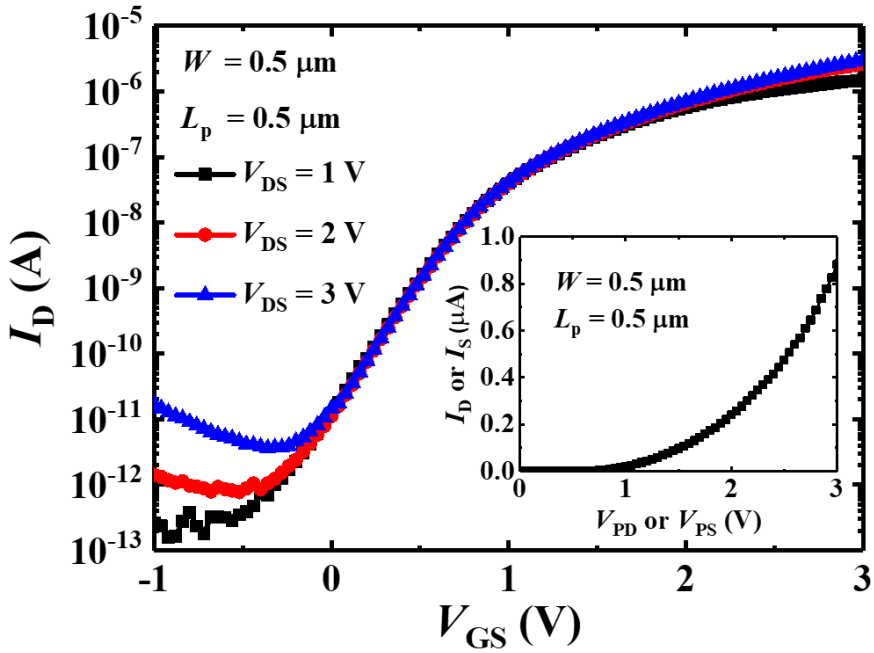


Fig. 2.5. Measured $I_D - V_{GS}$ characteristics of a specific device in a fabricated AND flash memory array as a parameter of V_{DS}. The measured diode characteristic between *p*-body and BL is represented in the inset figure.

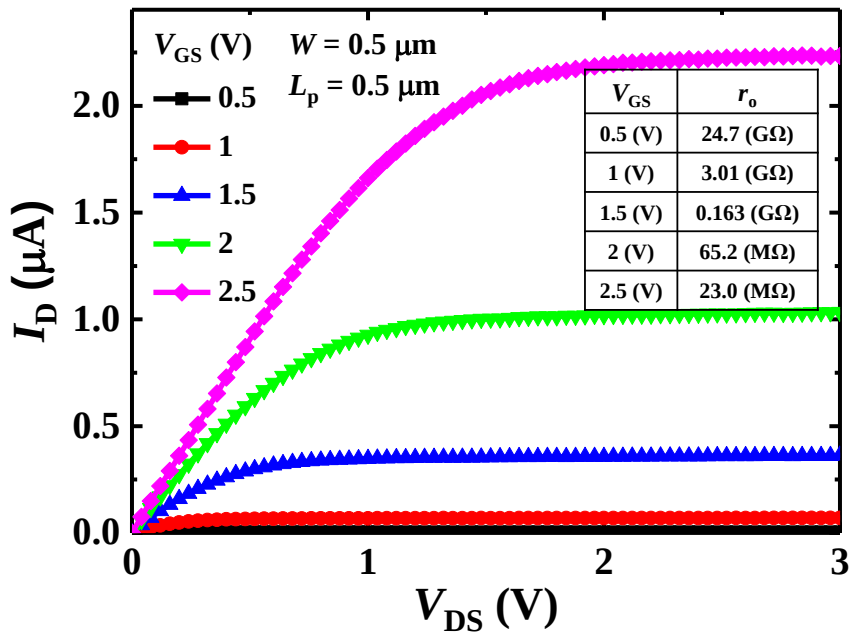


Fig. 2.6. Measured I_D – V_{DS} characteristics of the fabricated device as a parameter of V_{GS}. The measured output impedance as a parameter of V_{GS} is shown in the inset table.

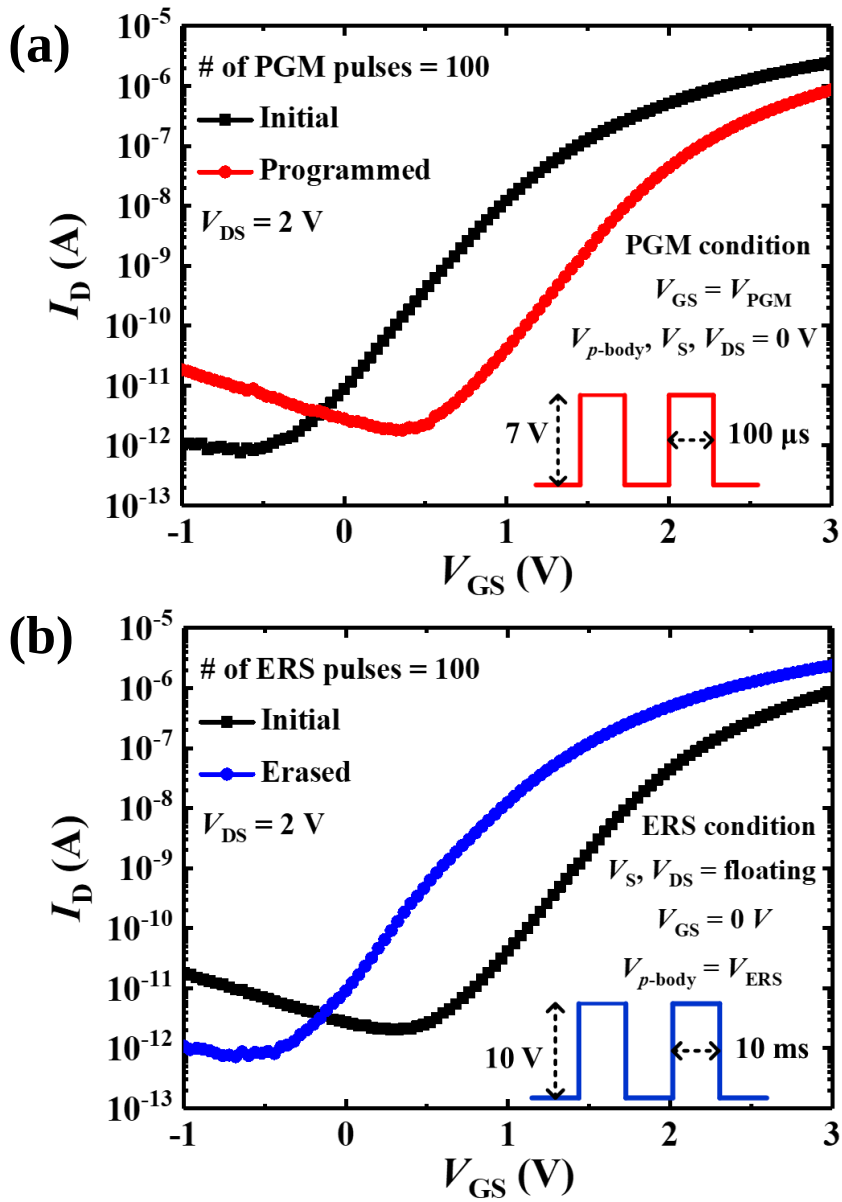


Fig. 2.7. Measured I_D – V_{GS} characteristics when identical (a) PGM or (b) ERS pulse is applied several times to the fabricated synaptic device.

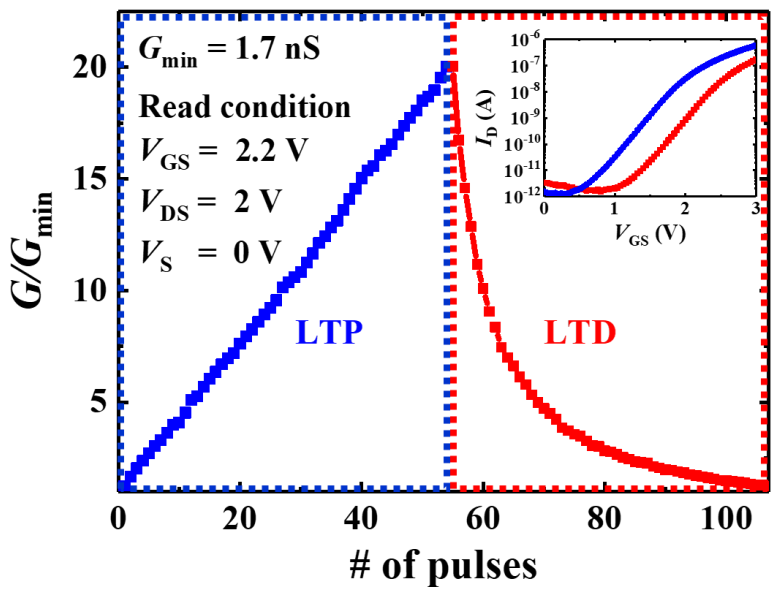


Fig. 2.8. Measured LTP and LTD characteristics of the fabricated synaptic device by applying identical ERS and PGM pulses, respectively. The memory window used to obtain the LTP/LTD curves is specified in the inset figure.

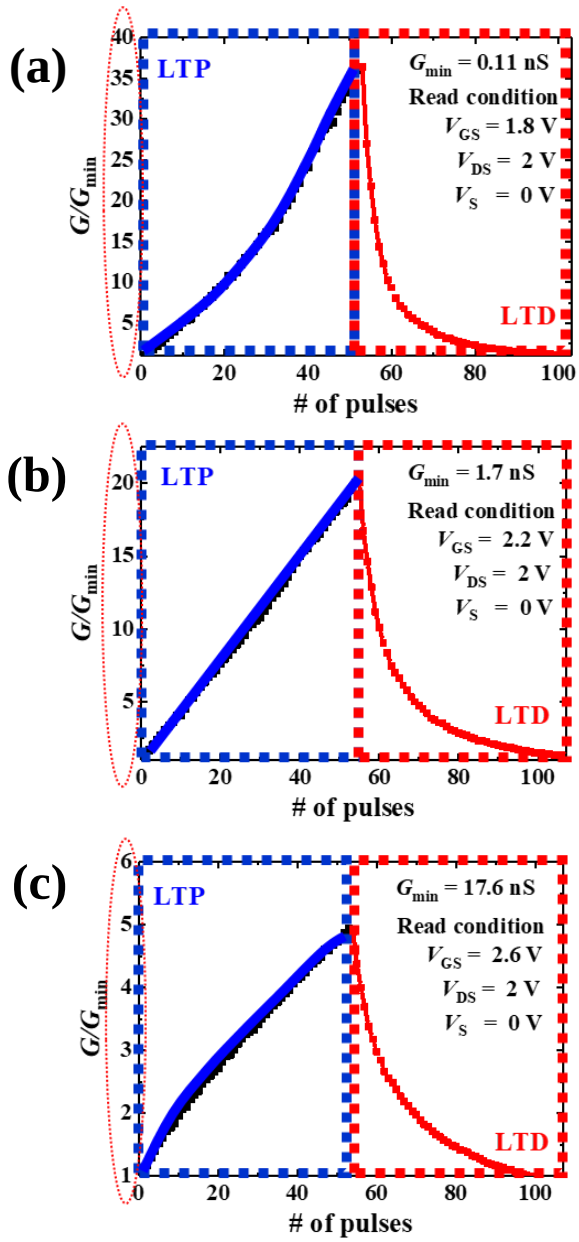


Fig. 2.9. Measured LTP and LTD characteristics of the fabricated synaptic device obtained under various read conditions in a specific memory window.

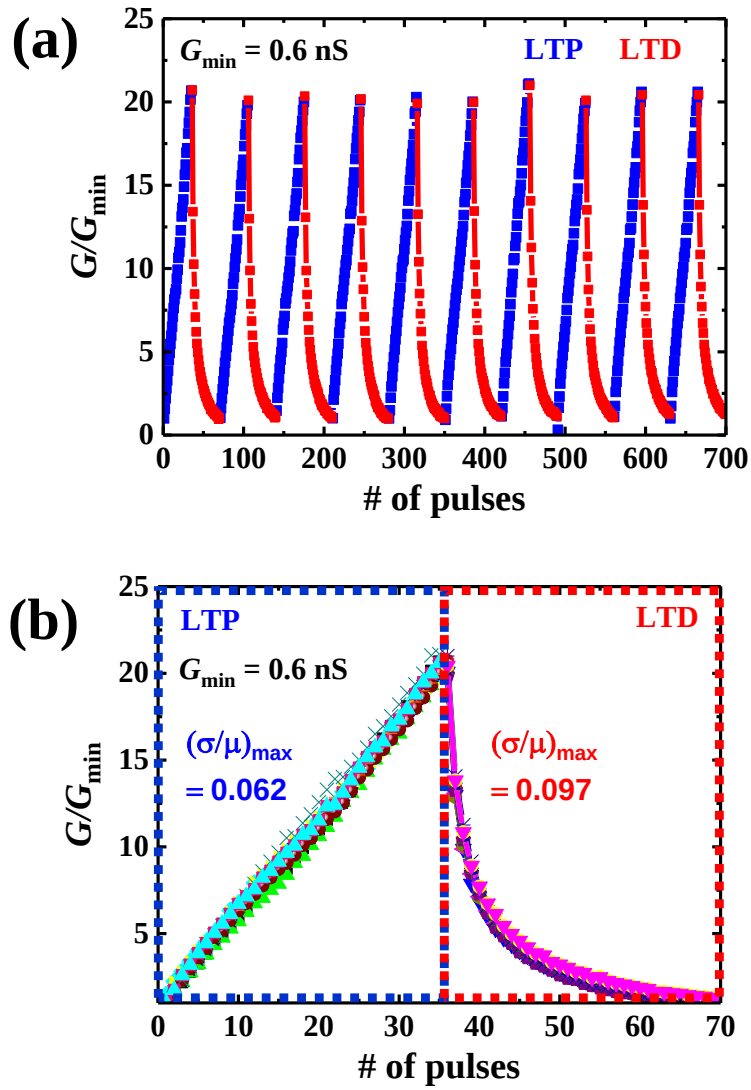


Fig. 2.10. (a) Measured cycle-to-cycle variation characteristics in the LTP/LTD of the fabricated synaptic device. (b) The LTP/LTD characteristics of the fabricated synaptic device obtained by cycling measurement.

2.4 Measurement results as a synaptic array

Fig. 2.11 (a) and (b) represent the SEM image and schematic diagram of the fabricated 2 × 2 AND flash memory array. The device described in the previous section is arranged as a single cell that constitutes the memory array. This type of array architecture, consisting of *n* independent memory cells controlled by *n* WLs, makes it possible to sum the synaptic current of each synaptic device while keeping a relatively small footprint. Also, due to the configuration of the array, Fowler-Nordheim (FN) tunneling can be used in the memory operations required to change the synaptic weights. Compared to the NOR memory array, which has to use the channel hot electron (CHE) method for selective PGM operation in the array, the FN tunneling method used in the AND memory consumes relatively lower power. Therefore, in designing a neural network aimed at on-chip training, the AND memory array architecture can have the advantage of lower power consumption than the NOR memory array. Table 2.1 represents the bias condition of each node required during the read operation and the selective PGM or ERS operation of a specific cell (Cell A in Fig. 2.11) in the 2 × 2 AND flash memory array. Note that

only positive voltage pulses are used to enable selective memory operation. In the read operation, the input signals from the pre-synaptic neurons are transmitted to the WLs as the V_{read}, and the current from the synaptic device with a specific synaptic weight flows through the BL to the post-synaptic neurons. During the PGM operation of Cell A, a PGM pulse with a positive V_{PGM} is applied to WL1, and the bias of DL1, SL1, and *p*-body line 1 (PL1) is maintained at 0 V. During this period, an inhibition (INH) pulse with an amplitude of V_{INH} is applied to PL2 to prevent Cell B crossing WL1 and BL2 from being programmed. In the array composed of a synaptic device with only the S, D, and gate, the INH pulses should be applied to both SL2 and DL2 to prevent voltage coupling with neighboring cells. However, in the proposed array, composed of synaptic devices with the PL, even if SL2 and DL2 are floated, the selective PGM operation is possible. This is because the voltage coupling effect due to the PL is more dominant than that of the neighboring BLs. For this reason, it is not necessary to apply the INH pulses with a relatively large amplitude to the SL and DL. As a result, this PGM operation scheme can offer a great advantage in circuit design, along with the characteristic that a high voltage

driver is not required for both the SL and DL during the ERS operation described in the device's characteristics. Also, during the PGM operation of Cell A, the bias of WL2 should be maintained at 0 V so that Cells C and D are not affected. Conversely, during the ERS operation of the selected cell, an ERS pulse with a positive V_{ERS} is applied to PL1, and the bias of WL1 is maintained at 0 V. During this period, SL1 and DL1 are floated. In order to prevent Cell C crossing PL1 and WL2 from being erased, an INH pulse with an amplitude of V_{INH} is applied to WL2. Here, the value of V_{INH} should be set not only to prevent the ERS operation of Cell C but to also not cause the PGM operation of Cell D. Also, the bias of PL2 should be maintained at 0 V to prevent the remaining cells from being affected. Note that SL2 and DL2 are floated. Fig. 2.12 represents a bias condition for measurement in a selective program operation of a specific cell (Cell A) in the fabricated 2 × 2 AND flash memory array. Fig. 2.13 (a) shows the measured $I_D - V_{GS}$ characteristics of all cells in the fabricated array when a specific cell (Cell A) is programmed in the initial state. As shown in Fig. 2.13 (a), in the initial state, four cells in the array show similar I_D–V_{GS} characteristics. The selective PGM operation of Cell A is performed

by applying 10 PGM pulses with V_{PGM} = 7 V and t_{PGM} = 100 μs to WL1. During the PGM operation, the potentials of the remaining nodes are as follows: V_{WL2} = 0 V, V_{DL1} & V_{SL1} = 0 V, V_{PL1} = 0 V, and V_{PL2} = 3.5 V. Note that the unspecified nodes are floated. As shown in Fig. 2.13 (b), after the selective PGM operation, the V_{th} of Cell A is increased by 0.45 V, while the maximum ΔV_{th} of the remaining cells is 0.03 V. Assuming that the read voltage is 2 V, the current of cell A is reduced by 154 nA, while the absolute value of the maximum change of the remaining cells is 4.52 nA, and the relative current change is less than 2.7%, proving that the selective PGM operation is successfully performed. And then, the selective ERS operation in the array is verified after the PGM operation stated above. Fig. 2.14 represents a bias condition for measurement in a selective erase operation of a specific cell (Cell A) in the fabricated 2 × 2 AND flash memory array, and Fig. 2.15 (a) shows the measured $I_D - V_{GS}$ characteristics of all cells in the fabricated array when a specific cell (Cell A) is erased in the selective programmed state. In the ERS operation, 100 ERS pulses with V_{ERS} = 10 V and t_{ERS} = 1 ms are applied to PL1. During the ERS operation, the potentials of the remaining nodes are as follows: V_{WL1} = 0 V, V_{WL2} =

3.5 V, and V_{PL2} = 0 V. During this period, all DLs and SLs are floated. As shown in Fig. 2.15 (b), The V_{th} of Cell A decreases by 0.4 V, while the amount of ΔV_{th} of the remaining cells is less than 0.02 V. Assuming that the read voltage is 2 V, the current of cell A is increased by 98.5 nA, while the absolute value of the maximum change of the remaining cells is 2.81 nA, and the relative current change is less than 2.8%, proving that the selective ERS operation performs well.

Fig. 2.16 shows a top SEM image of the fabricated 10 × 2 AND flash memory array. Fig. 2.17 represents the measurement result of VMM in the fabricated 10 × 2 AND flash memory array. Fig. 2.17 (a) shows the raw data of the measurement result of performing current summation for 10 devices in the array, and Fig. 2.17 (b) describes the result of analyzing the current summation value when the read voltage is assumed to be 2.2 V. The value obtained by measuring 10 devices individually and mathematically summed up is 7.204 μA, and the value measured by applying the input to 10 WLs at the same time is 7.193 μA, resulting in a current loss of 0.16%. Since the proposed device can secure good characteristics of output resistance due to its structural feature, it is possible to operate the device in the

saturation region. Nevertheless, long BLs in large-scale implementation can lead to loss of current summation. This can be solved by reducing the resistance of n^+ poly-Si used in DLs and SLs, or to design additional metal lines. Fig. 2.18 represents the measurement result of quantization of the synaptic weight in the fabricated synaptic array. The synaptic device fabricated in this work uses Si_3N_4 as a charge trap layer. And there can be variations for each device depending on the initial amount of stored charge in the charge trap layer. However, as previously experimentally verified, since selective memory operation in the synaptic array is possible, synaptic devices can be set to a specific conductance level. Fig. 2.18 (a) shows the measurement result of weight quantization for 4 states (conductance levels = 500, 250, 100, 50 nA), and Fig. 2.18 (b) represents 4 states quantization distribution of synaptic weights in the fabricated 10 × 2 AND flash memory array. In each state, σ/μ is 0.02, demonstrating that the weight quantization is performed relatively evenly.

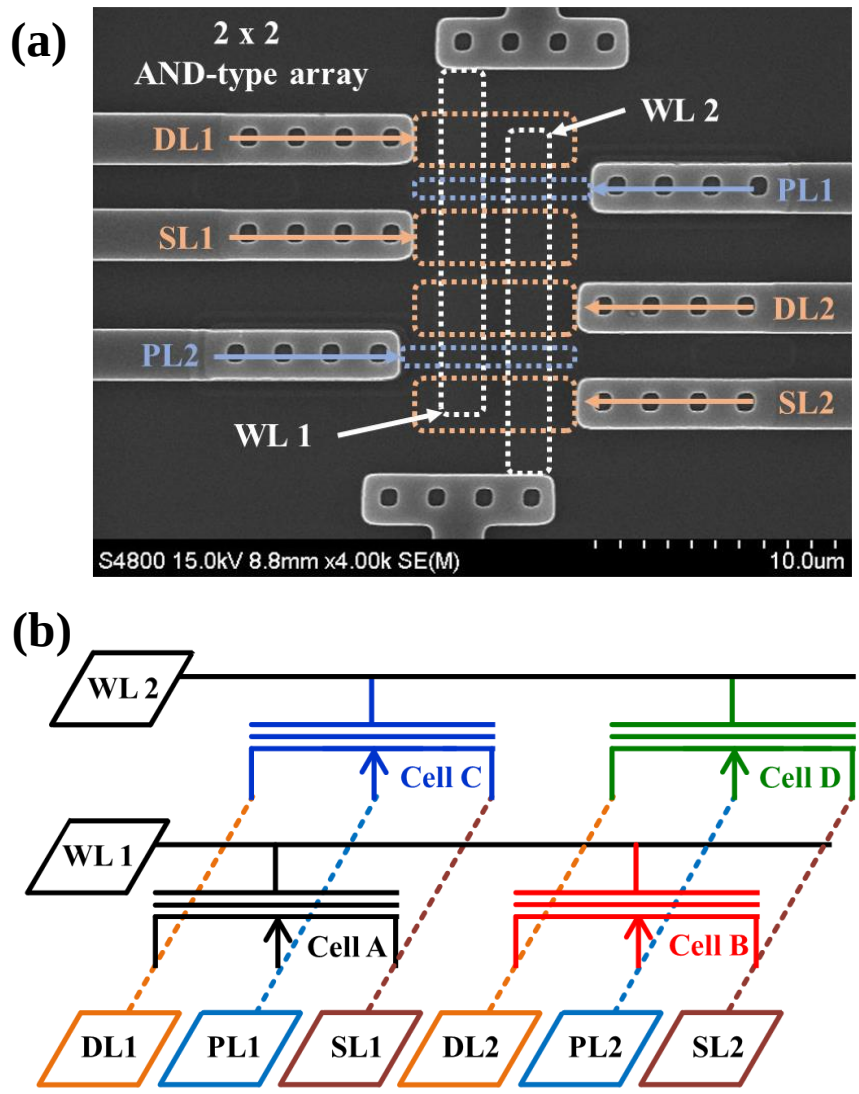


Fig. 2.11. (a) A SEM image and (b) a schematic diagram of the fabricated 2 × 2 AND flash memory array.

Node	Read	PGM	ERS
WL1	V_{read}	V_{PGM}	**0 (V)**
WL2	**0 (V)**	**0 (V)**	V_{INH}
DL1 & SL1	V_{DD} **& 0 (V)**	**0 (V)**	**Floating**
DL2 & SL2	V_{DD} **& 0 (V)**	**Floating**	**Floating**
PL1	**0 (V)**	**0 (V)**	V_{ERS}
PL2	**0 (V)**	V_{INH}	**0 (V)**

Table. 2.1. Bias condition required for read operation of a specific cell (Cell A in Fig. 2.11) and selective program and erase operation of a specific cell (Cell A in Fig. 2.11) in a 2 × 2 AND flash memory array.

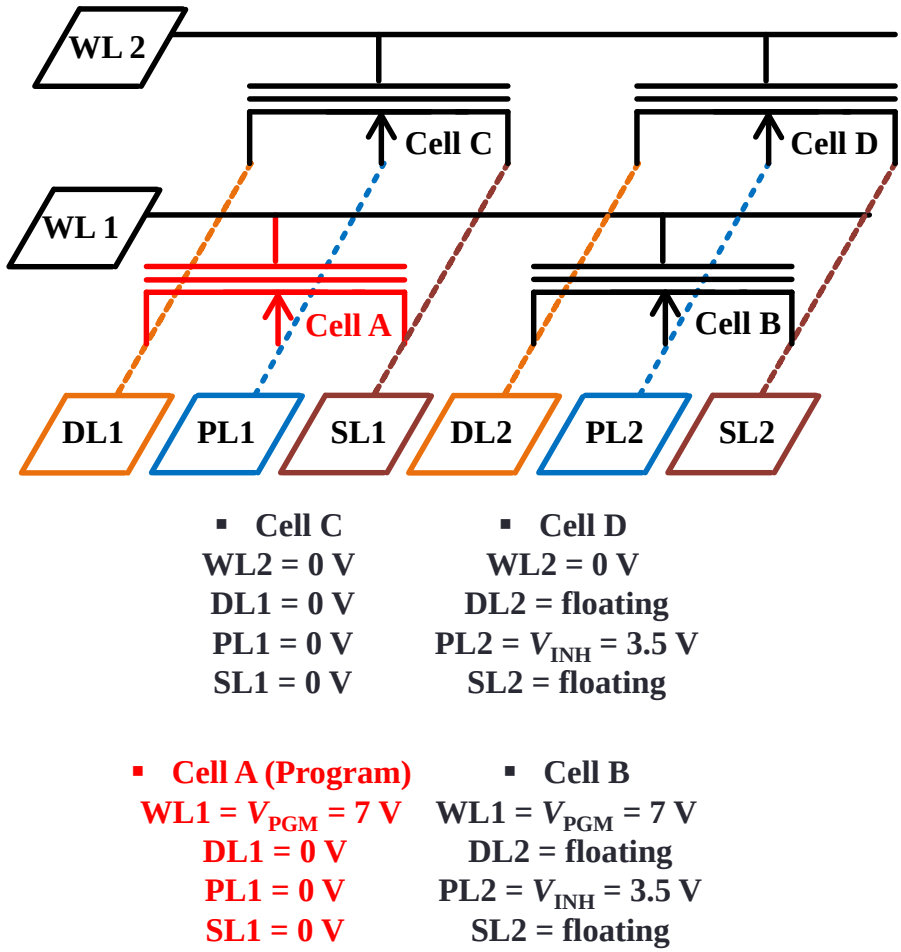


Fig. 2.12. Bias condition for measurement in a selective program operation of a specific cell (Cell A) in the 2 × 2 AND flash memory array.

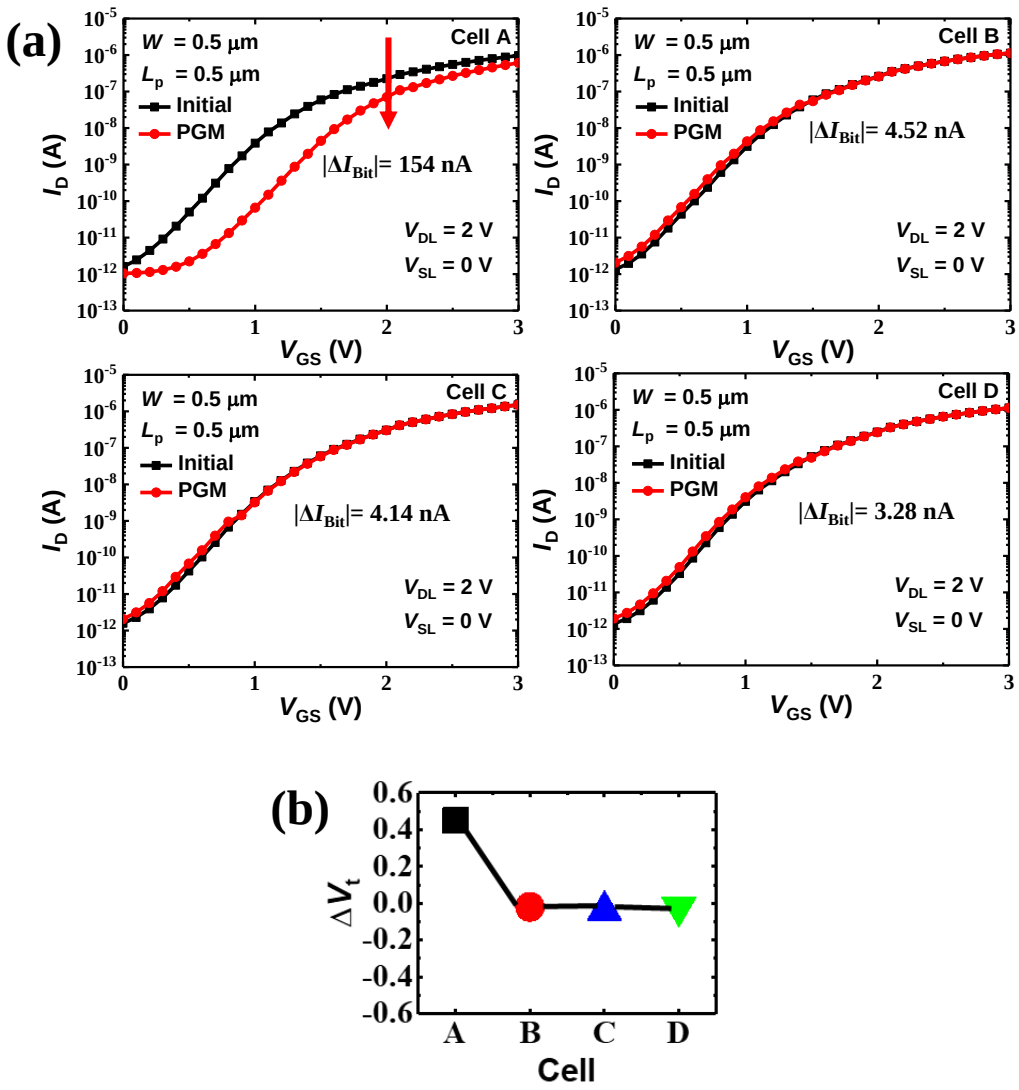


Fig. 2.13. Measured (a) I_D – V_{GS} characteristics and (b) threshold voltages of all cells in the fabricated 2 × 2 AND flash memory array when a specific cell (Cell A) is programmed in the initial state.

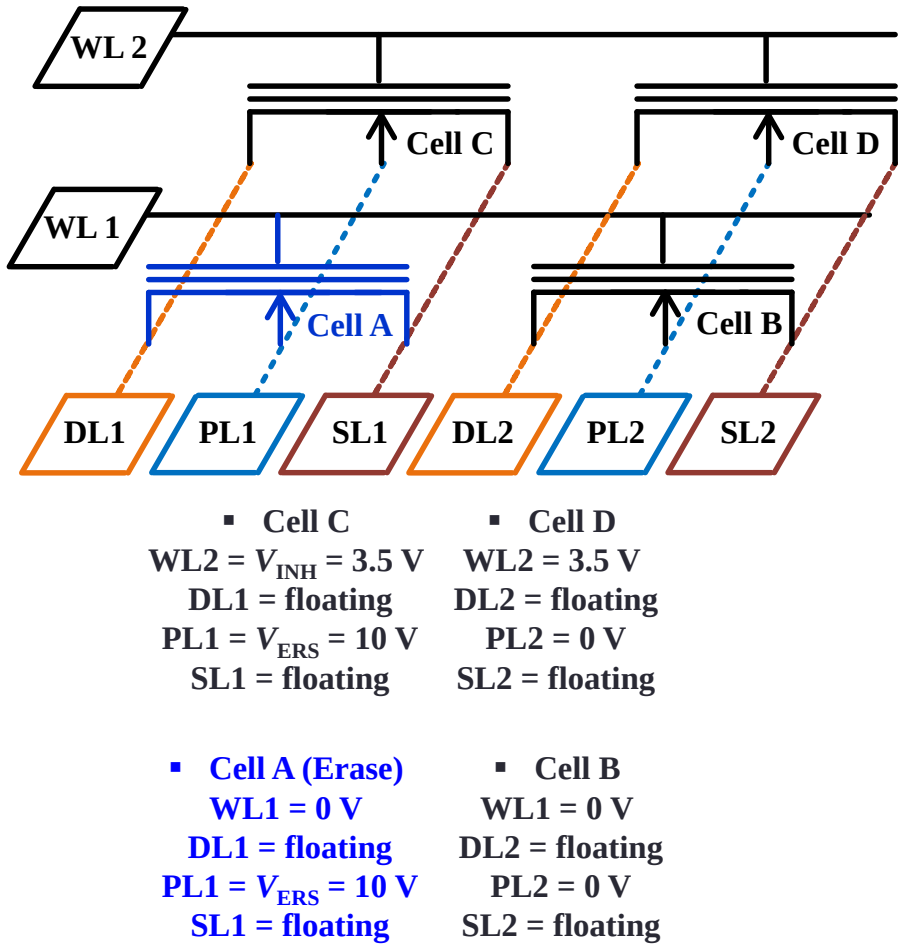


Fig. 2.14. Bias condition for measurement in a selective erase operation of a specific cell (Cell A) in the fabricated 2 × 2 AND flash memory array.

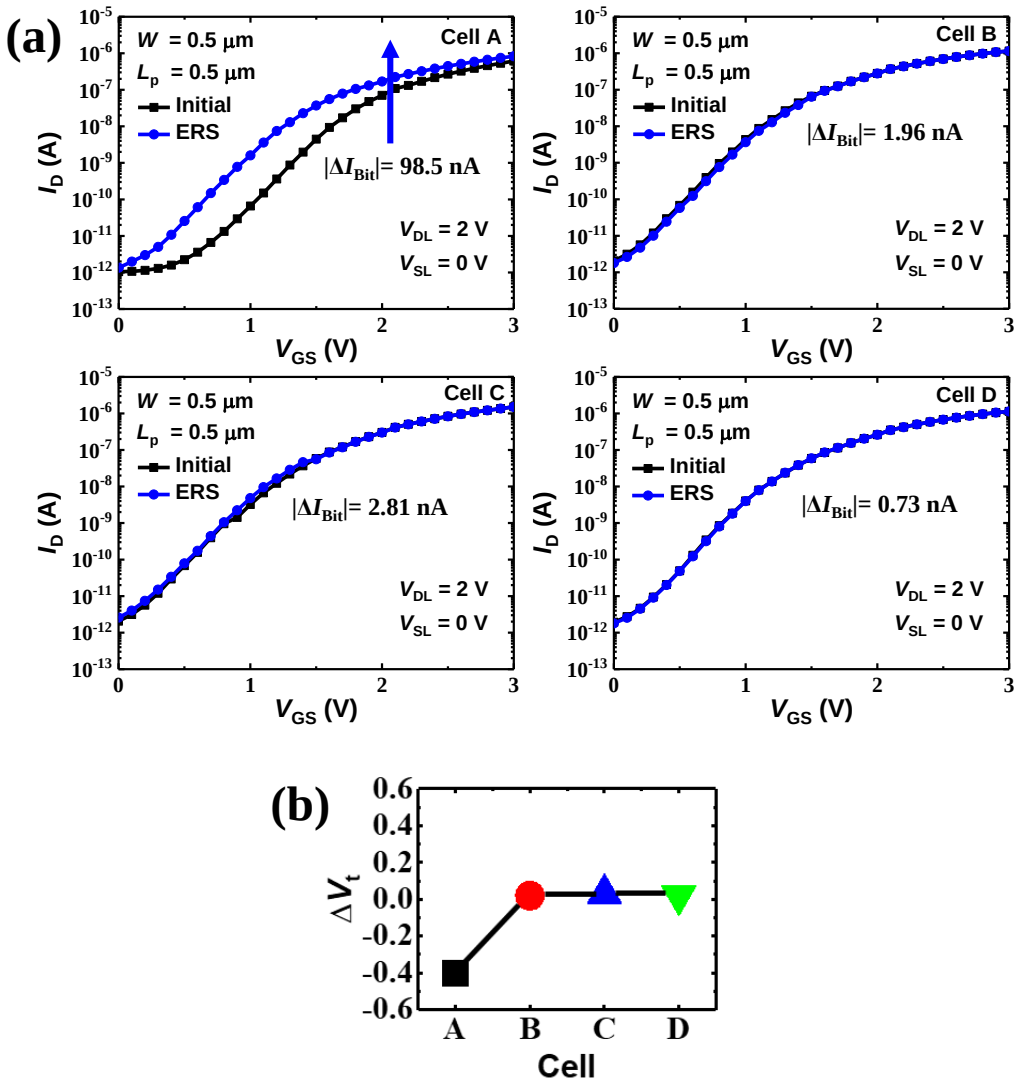


Fig. 2.15. Measured (a) I_D – V_{GS} characteristics and (b) threshold voltages of all cells in the fabricated 2 × 2 AND flash memory array when a specific cell (Cell A) is erased in the selective programmed state.

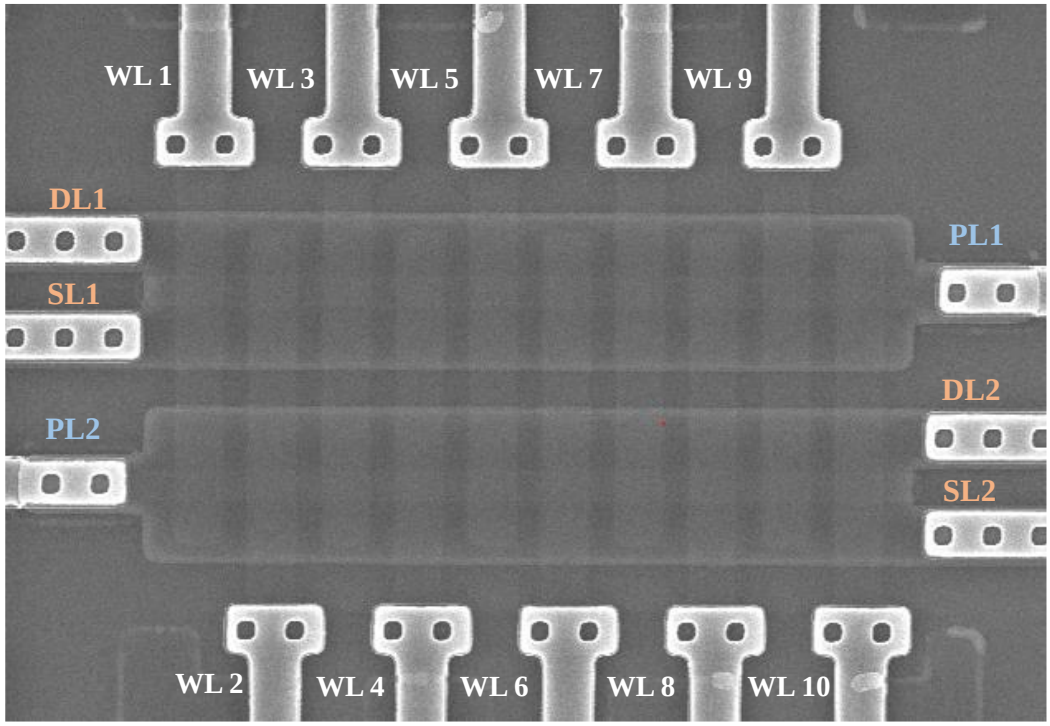


Fig. 2.16. A top SEM image of the fabricated 10 × 2 AND flash memory array.

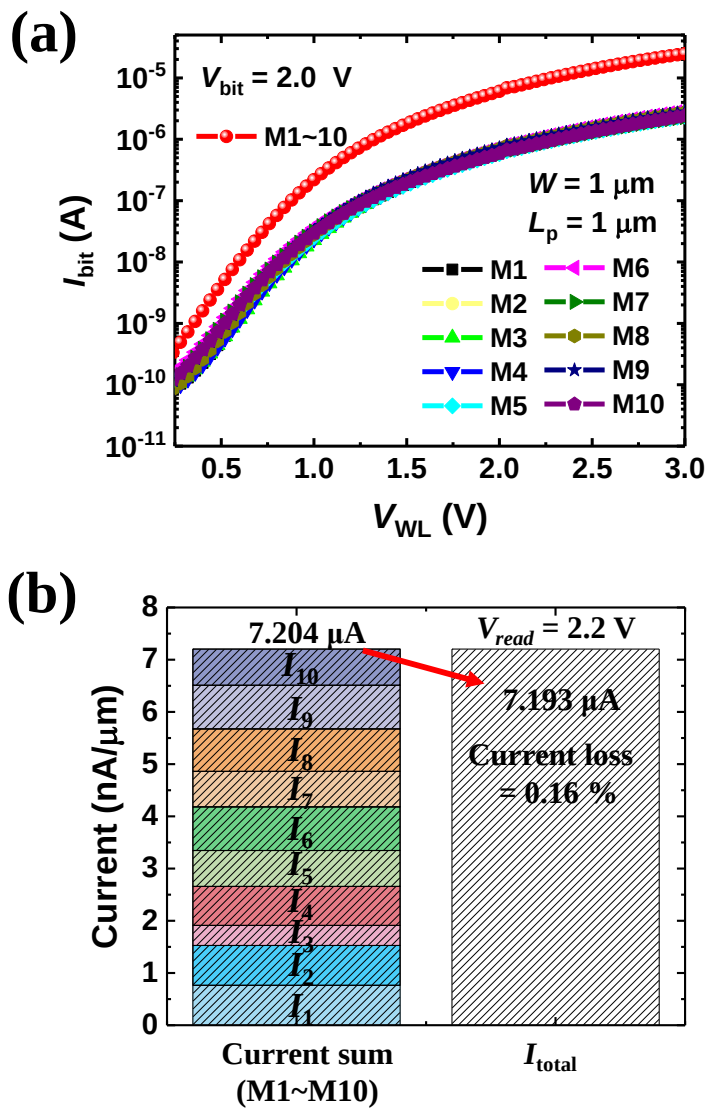


Fig. 2.17. (a) Measurement result of VMM in the fabricated the fabricated 10 × 2 AND flash memory array. (b) Analyzed result of VMM when the read voltage is assumed to be 2.2 V.

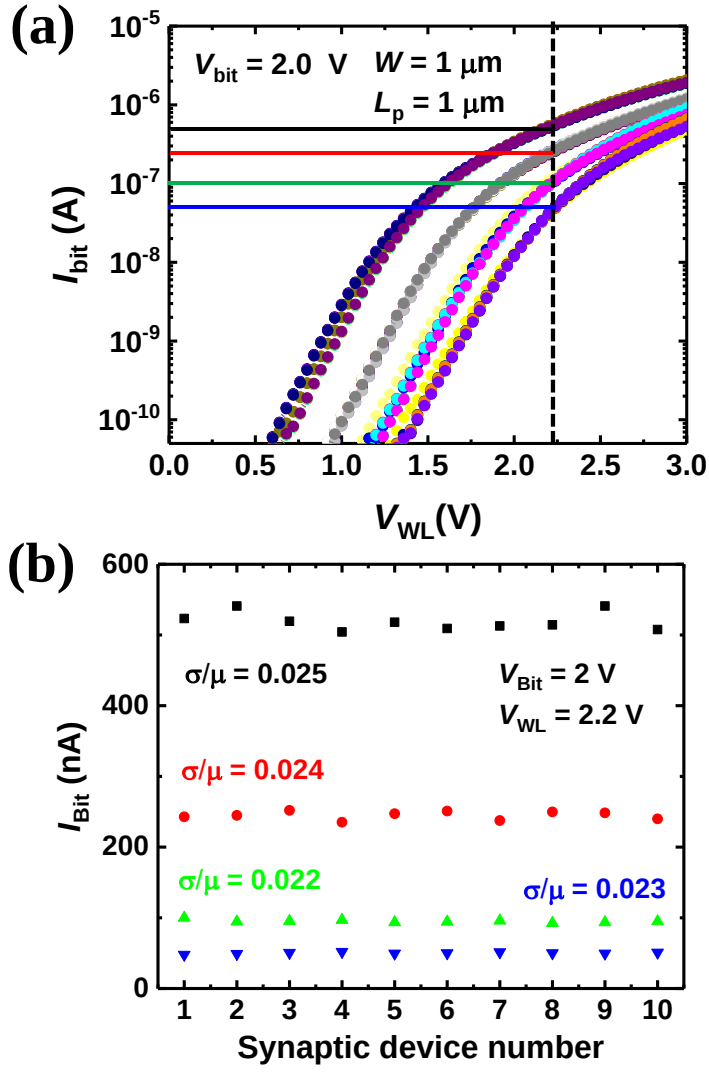


Fig. 2.18. (a) Measurement result of weight quantization for 4 states (conductance levels = 500, 250, 100, 50 nA), and (b) 4 states quantization distribution of synaptic weights in the fabricated 10 × 2 AND flash memory array.

Chapter 3
Hardware-based SNN for unsupervised learning

3.1 SNN using spike-timing-dependent plasticity (STDP)

A spike-timing-dependent plasticity (STDP) is one of the representative examples of the synaptic weight change rules that enable unsupervised learning in SNN [56-58]. The network trained with STDP reinforces spike patterns associated with previously occurring stimulation and suppresses meaningless spiking activity, depending on the timing of pre-synaptic and post-synaptic neurons' firings. The STDP learning rule as a simple, intuitive, and efficient method has significant advantages in constructing an event-driven computing architecture. To devise a hardware-based neural network that can effectively utilize the AND flash memory array, a fully-connected (FC) SNN using the STDP algorithm is designed. The proposed SNN architectures are designed so that the synaptic weight increases and decreases according to the timing of the neuron firing. Fig. 3.1 represents LTP/LTD of electronic synaptic devices in a crossbar array depending on firings of neurons

in SNN architecture using the STDP learning rule. In the crossbar architecture, input (pre-) and output (post-) neurons are connected through the electronic synaptic devices. The basic principles of LTP/LTD processes satisfying STDP in electronic synaptic devices are as follows. When the post-neuron fires after signals from pre-neurons pass through synapses to a post-neuron, the electronic synaptic devices that contribute to the neuron's firing undergo the LTP process. On the other hand, when the post-neuron fires before the signals from the pre-neurons pass through the synapses to the post-neuron, electronic synaptic devices connected to the post-neuron undergo the LTD process. Pulses for LTP and LTD processes are generated from the input and output neurons and applied to the electronic synaptic devices. In more detail, the input pulses generated by the input neurons are applied to the WLs of the electronic synaptic array, and the feedback pulses generated by the fired output neurons are applied to SL, DL, and PLs.

To describe the systematic operation of the input and output neurons, Fig. 3.2 shows a conceptual diagram of the designed SNN using the STDP learning rule. In the proposed SNN architectures, there are global pulse generator modules for

systematically operating each neuron in a single neuron layer. Neurons within a single neuron layer share a global pulse generator, which is responsible for generating the various pulses required for neuron operation. And, each neuron is selectively assigned the necessary pulses depending on whether the neuron fires or not. Global pulse generators use signals generated by the neurons, eliminating the need for an additional external controller in the architecture to operate the neurons. In other words, the global pulse generator receives the spike signal generated by a fired neuron to generate the necessary pulses. Note that if one neuron fires in one layer of neurons, the other neuron is inhibited. In order to selectively apply various pulses generated by the global pulse generator to each neuron, each neuron includes switches for a specific purpose and switch control units for adjusting the switches. Most of the modules in the global pulse generator consist of circuits required for increasing the width and amplitude of the pulse. This is because the width and amplitude of the pulses to utilize electronic synaptic devices based on flash memory technology are very large compared to those of the spike signal from neurons. In addition, the width and amplitude of the pulses needed to control the operation

required for each neuron during the LTP/LTD processes should also be similar to those of the pulses applied to the electronic synaptic devices. In general, stable design condition for circuits required for increasing the amplitude of the pulse requires a fairly large area, which leads to an increase in the area of the computing architecture when this type of circuit is included in each neuron circuit. For this reason, as the number of neurons in a single neuron layer and the complexity of the required pulses increase, this computing architecture including the global pulse generators has an effective advantage in terms of system area as well as the systematic operation of the neurons.

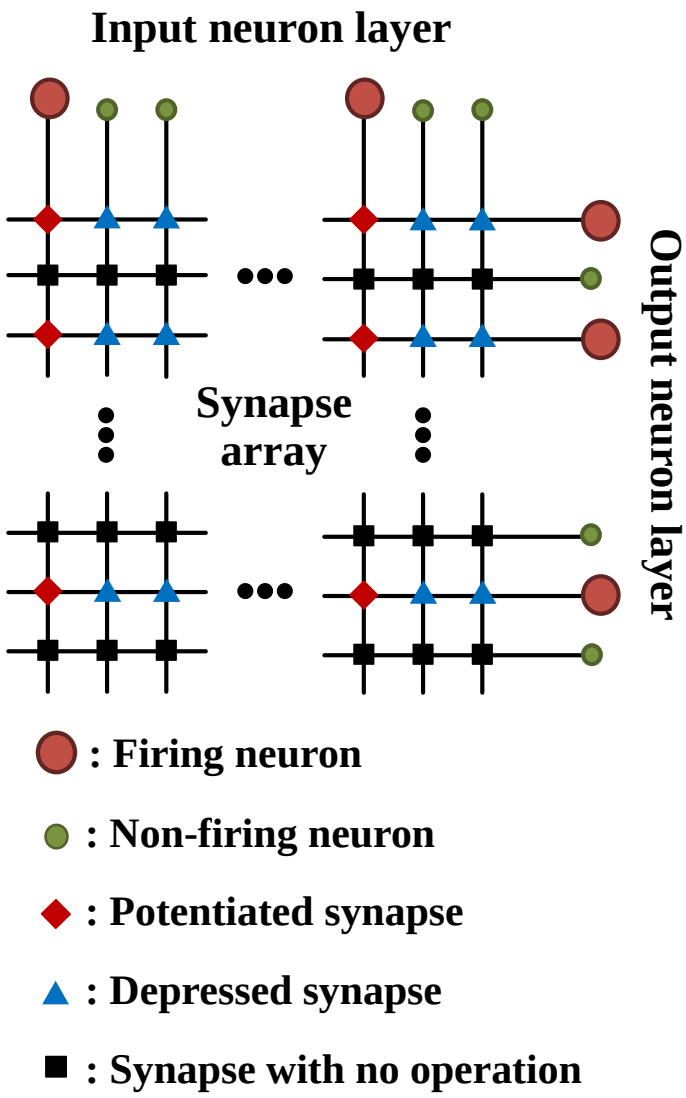


Fig. 3.1. LTP/LTD characteristics of electronic synaptic devices in crossbar array depending on firings of neurons in SNN using the STDP learning rule.

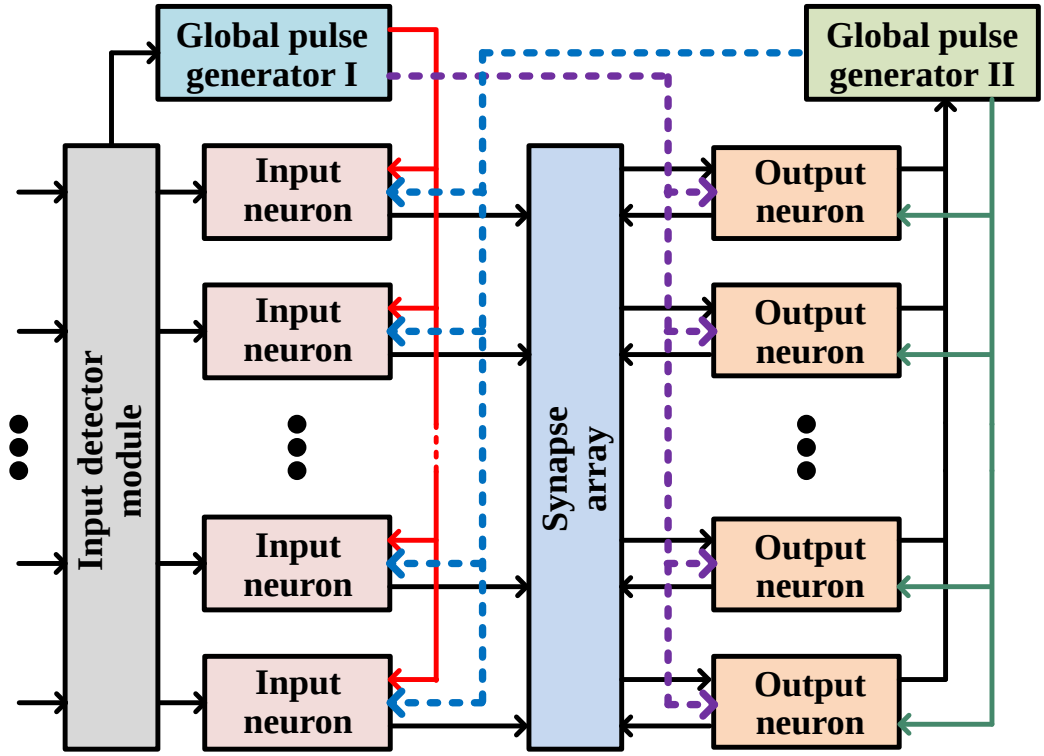


Fig. 3.2. A conceptual diagram of the designed SNN using the STDP learning rule.

3.2 Pulse scheme for STDP learning rule

Many studies using the STDP learning rule utilize the overlapping of pulses to implement the selective memory operation of the electronic synaptic devices [40, 42, 59, 60]. Fig. 3.3 represents an example of a pulse scheme that enables selective LTP/LTD of electronic synaptic devices in neural networks using the STDP learning rule and the AND-type array. This pulse scheme does not include the inhibition (INH) pulse and utilizes the overlapping of pulses. Fig. 3.3 (a) shows the pulses applied to each terminal of the electronic synaptic devices corresponding to the diamonds in Fig. 3.1. These devices, connected to both the fired output neurons and the input neurons that receive input signals have increased synaptic weight by the LTP process. Input pulses that consist of a positive voltage held during the read operation time and a constant negative voltage maintained for the sufficient time required for ERS operation are generated from the input neuron and applied to the WLs of these devices. At this time, except for the read operation time, the drain terminals become floating nodes to prevent the system error due to unnecessary leakage current. In addition, feedback pulses generated from output neurons are

applied to PLs of these devices at the time when output neurons fire. This feedback pulse has a form that maintains the positive voltage for the sufficient time required for the ERS operation and has the constant negative voltage corresponding to $-V_{PGM}$ for the time required for the PGM operation. And then, while the time at which the negative voltage is maintained in the WLs and the time at which the positive voltage is maintained in the S and D lines overlaps, the devices are erased by the LTP process, increasing the weight. At this point, the end of the tail portion of the input pulse is set later than that of the feedback pulse to prevent unwanted PGM operation in the devices. Fig. 3.3 (b) shows the pulses applied to each terminal of the electronic synaptic devices corresponding to the triangles in Fig. 3.1. These devices, connected to both the fired output neurons and the input neurons that don't receive input signals, undergo the LTD process. The WLs of these devices are maintained at 0 V because no input is present, and the feedback pulses are applied to the PLs because they are connected to the fired output neuron. These devices undergo PGM operation by the tail portion of the feedback signal with $-V_{PGM}$, which causes the weight to be decreased by the LTD process. Fig. 3.3 (c) and (d) represent the pulses

applied to each terminal of the electronic synaptic devices connected to non-fired output neurons. The PLs of these devices maintain 0 V for the duration of the feedback pulse so that the amount of stored charge affecting the weight does not change. This method is very intuitive and simple since there is no need to generate additional INH pulses. However, as suggested above, the methods to utilize the overlapping of pulses have the disadvantages of requiring pulses with negative voltage values. This means that the circuit responsible for generating pulses with negative voltage values and supplying them stably to a large number of electronic synaptic devices can place a heavy burden on the design of the hardware-based computing architecture. The pulse scheme proposed in this study consists of pulses that use only positive voltage values without using signal superposition.

Fig. 3.4 illustrates a proposed pulse scheme that enables selective LTP/LTD of electronic synaptic devices in neural networks using the STDP learning rule and the AND-type array. The proposed pulse scheme has the same purpose as the pulse scheme described above but utilizes INH pulses to enable selective LTP/LTD processes in the electronic synaptic array. Fig. 3.4 (a) shows pulses applied to each

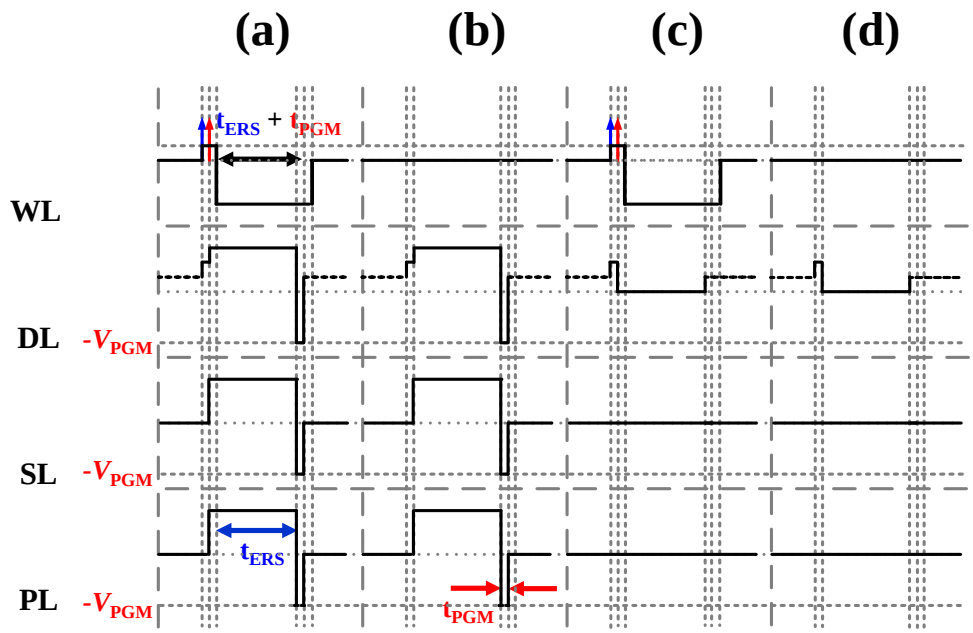
terminal of the electronic synaptic devices undergoing the LTP process. These devices undergo the ERS operation by the applied feedback pulse to the PL, which is generated from the output neuron after the read operation. For the devices undergoing the LTD process represented in Fig. 3.4 (b), the PL of these devices is affected by the feedback pulse because it is connected to the same output neuron as in Fig. 3.4 (a). At this time, the PGM pulse that has a V_{PGM} amplitude and a width longer by t_{PGM} than the width of the feedback pulse is applied to the WLs of these devices. This is to prevent unnecessary ERS operation during the feedback pulse and to cause the LTD processes due to the pulses on the WLs of these devices. And then, as shown in Fig. 3.4 (c) and (d), the pulses having the V_{inh} amplitude and the same width as the pulse applied to WLs in Fig. 3.4 (b) are applied to the PLs of the devices that are not connected to the fired output neuron. These INH pulses can act as the suppressor for LTD operation that can be caused by the PGM pulses.

Arranging the electronic synaptic devices in the array in the form of a crossbar is a very efficient way in terms of system area and power consumption in designing the neural computing architecture for parallel computing. And, it is also important

to efficiently design the neuron circuits connected to an array of electronic synaptic devices and peripheral circuits for synaptic weight change. Furthermore, the peripheral circuit used to change the synaptic weight should be designed considering the type of electronic synaptic device. In particular, when using electronic synaptic devices that require pulses with a relatively high voltage and a wide width, the configuration of the peripheral circuits and the neural computing architecture can vary significantly depending on the type of pulse used.

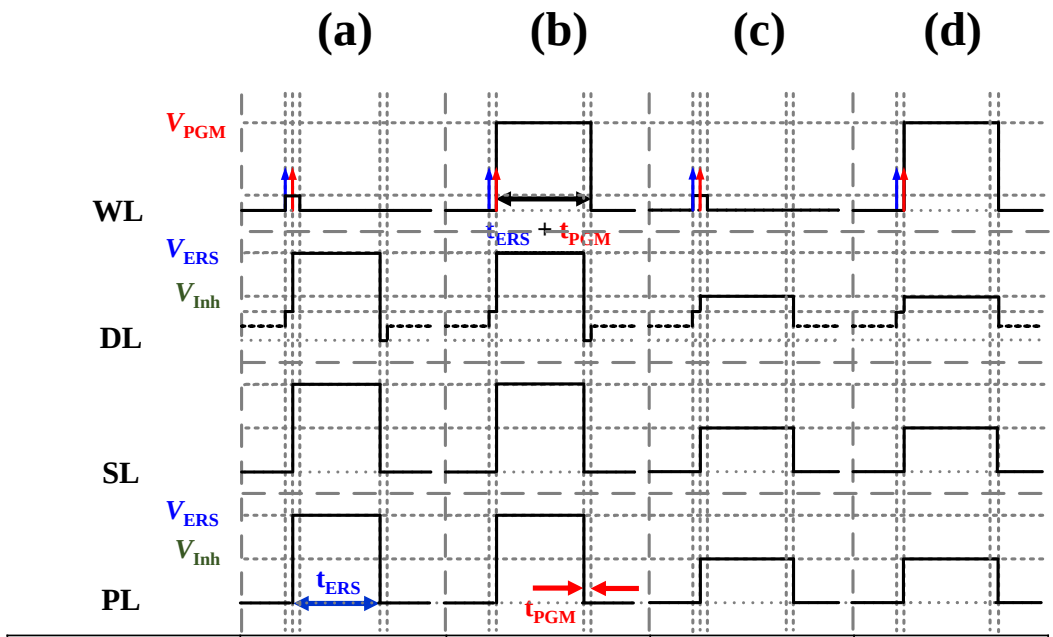
Although the neuron circuits should generate the additional inhibition pulse for the proposed pulse scheme, they are more efficient in terms of area and energy consumption than the neuron circuits using the overlapping of pulses to implement STDP. This is because a fairly large capacitor is required to generate and control the pulses having negative voltage values in the pulse scheme using the overlapping of pulses. Even, a larger capacitor is needed to generate a negative voltage simply by using the charging and discharging method without using a negative power supply that can be a burden on the circuit design. Since the capacitance of this capacitor, which should be included in each neuron, increases as the number of

electronic synaptic devices connected to the neuron increase, the area burden of hardware can be increased. In addition, in order to generate a desired negative voltage value stably using a capacitor having a large capacitance, the current required for charging and discharging should be increased, which inevitably increases the energy consumption of the neuron circuit. For example, in a fully connected 2-layer SNN architecture including 784 input neurons and 200 output neurons using the overlapping pulse scheme, each input and output neuron should additionally contain large capacitors corresponding to tens of ~pF. To generate a large value of negative voltage required for memory operation using corresponding values of such capacitors, the energy consumption of several ~nJ is required, which implies several times more energy consumption than the circuit proposed in this study.



Input Neuron	Firing	Non-firing	Firing	Non-firing
Output Neuron	Firing	Firing	Non-firing	Non-firing

Fig. 3.3. An example of a pulse scheme without using INH pulses for realizing selective LTP/LTD of electronic synaptic devices using the STDP learning rule and the AND-type array.



Input Neuron	Firing	Non-firing	Firing	Non-firing
Output Neuron	Firing	Firing	Non-firing	Non-firing

Fig. 3.4. A proposed pulse scheme using INH pulses for realizing selective LTP/LTD of electronic synaptic devices using the STDP learning rule and the AND-type array.

3.3 MNIST pattern learning and classification

The system-level pattern learning simulation is performed by the software Python to evaluate the proposed hardware-based SNN architecture. Fig. 3.5 illustrates the flow chart of the learning and recognition process in the hardware-based SNNs using the STDP algorithm. The detailed description of the MNIST handwritten digit pattern learning method is as follows. First, the training dataset as input is a total of 60000, and the order of all data is set randomly. And the training dataset is encoded in binary rate coding. This means that if the brightness of the dataset is 0.5 or less, it is maintained at 0 during the time set. Conversely, when the brightness is 0.5 or more, it is maintained at 1 during the time step. Then, as 28 × 28 input patterns are repeatedly applied to the synaptic array, changes in synaptic weights occur with feedback signals generated by fired neurons according to the STDP learning rule based on the proposed pulse scheme. In this case, the integrate-and-fire model is used as the neuron. If the first fired output neuron is named A as the time step increases by one, the membrane potential of the remaining output neurons other than neuron A is set to 0 to take the winner-take-all (WTA) method.

Then, among the 784 synapses connected to neuron A, the conductance of the synapse with input is increased (LTP), whereas the conductance of the synapse with no input is decreased (LTD). The LTP/LTD characteristics reflect the characteristics of the electronic synaptic device discussed in Chapter 2.3, and a total of 784 synaptic weight updates are performed. Fig. 3.6 shows training curves of the hardware-based SNN based on the STDP algorithm for the MNIST test set classification as a parameter of the number of output neurons. After the training process, we can obtain 91.63 % of recognition accuracy, and it can be seen that a weight mapping image corresponding to 784 × 200 synapses becomes clear as the number of training sets increases as shown in Fig. 3.7.

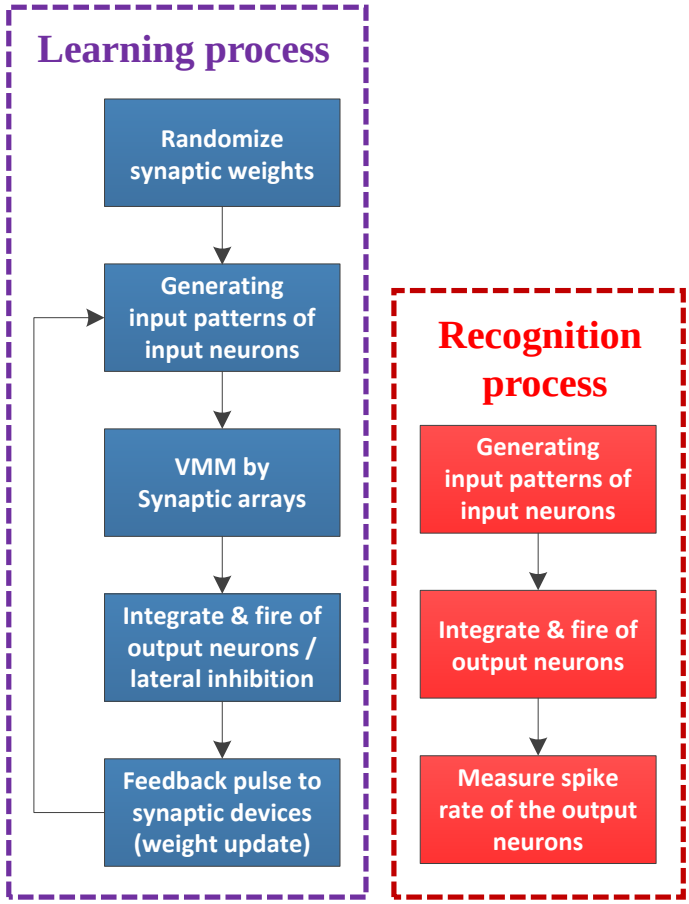


Fig. 3.5. Flow chart of the learning and recognition process in the hardware-based SNNs using the STDP algorithm.

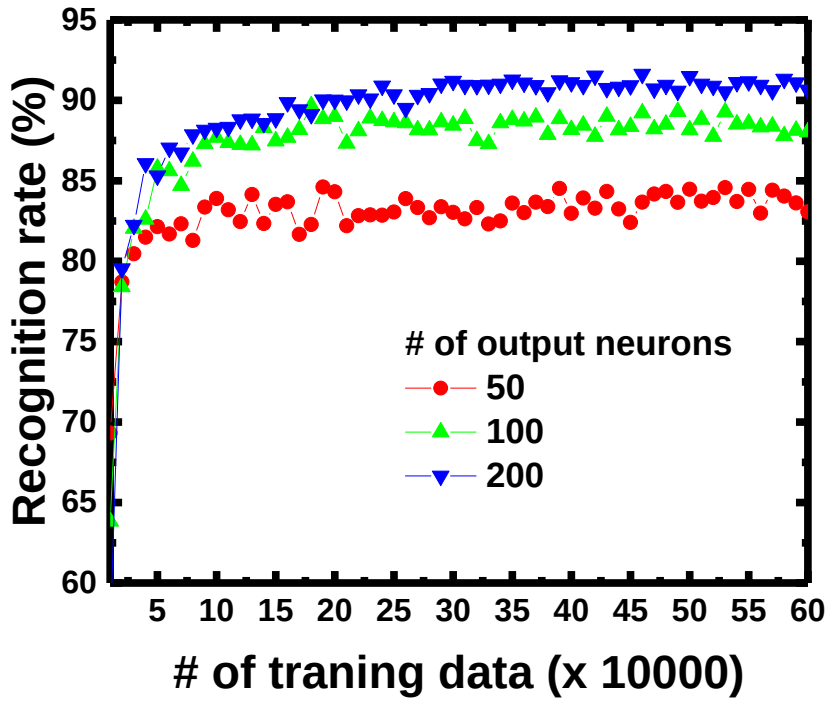


Fig. 3.6. Training curves of the hardware-based SNN based on the STDP algorithm for the MNIST test set classification as a parameter of the number of output neurons.

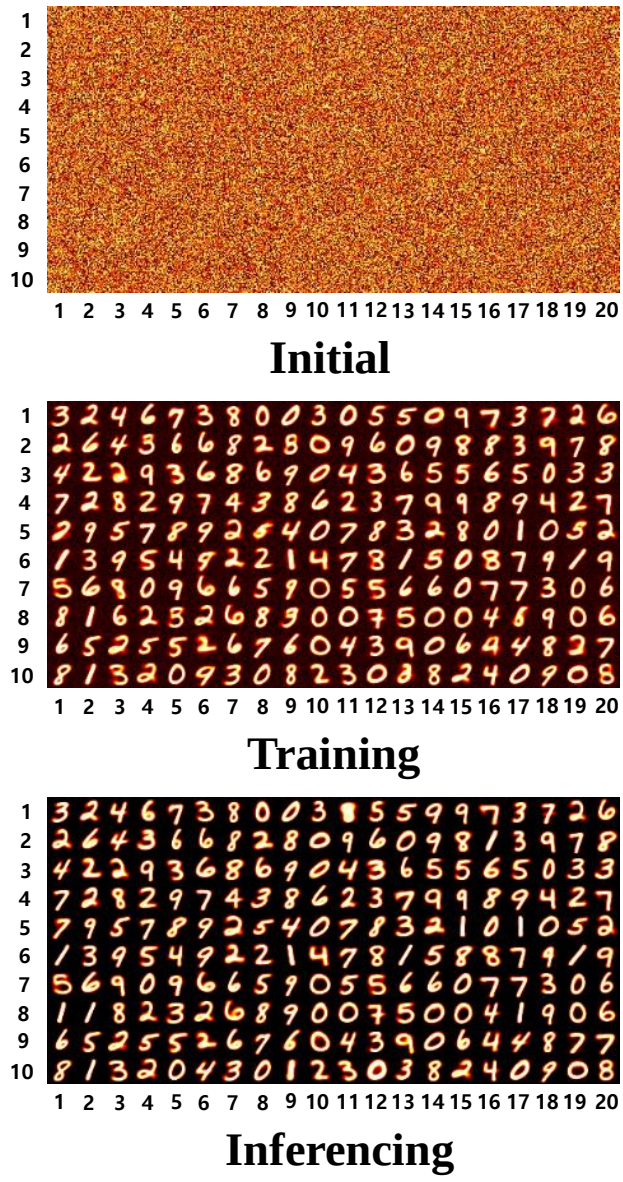


Fig. 3.7. Weight mapping images of the synaptic weights after training process of the MNIST pattern utilizing the fully connected 784 input neurons and 200 output neurons.

Chapter 4
Hardware-based SNN for supervised learning

4.1 SNN using direct feedback alignment (DFA)

A feedback alignment (FA) algorithm leverages the existing structure of the BP algorithm but replaces the backward synaptic weights with random connections to solve the weight transport problem [61]. For this reason, in a computing architecture using the FA algorithm, synaptic weights located in the shallow layer do not require information about the synaptic weights of the deep layers. As a method derived from the FA algorithm, a direct feedback alignment (DFA) algorithm uses a method in which the error information from the output layer is directly propagated to all upstream layers, and it can achieve performance competitive with backpropagation in fully-connected layers [62-64]. This learning method enables parallel processing in the propagation of errors, reducing the backward phase latency and the number of computations required for the synaptic weight update in the network. Furthermore, the use of this learning algorithm also provides an opportunity to

alleviate the complex design requirements of a hardware-based neural network (HNN) for on-chip training. Fig. 4.1 (a), (b), and (c) show the error transportation configuration in HNN utilizing BP, FA, and DFA, respectively.

In order to devise a hardware-based neural network that can effectively utilize the AND flash memory array, a fully-connected (FC) three-layer SNN using a DFA algorithm is designed. In the BP algorithm, the synaptic weights in the forward path and backward path are identical as shown in Fig. 4.2 (a). Therefore, the synaptic arrays used in the forward path should be used again in the backward path, while the current summation of the backward path is performed in the same array orthogonally to that of the forward path. Unfortunately, as shown in Fig. 4.2 (b), and (c), since the SLs and DLs in the AND flash memory array are formed in parallel, the current summation is only performed in one direction. Thus, this array architecture is not suitable for the backpropagation algorithm. However, since the DFA algorithm uses random backward connections to solve the weight transport problem, the DFA algorithm does not need to have identical synaptic weights in the forward and backward paths. In other words, the current summation in the forward

and backward paths is performed in separated synaptic arrays. This enables the HNN using the AND flash memory array to update the synaptic weights on the chip and achieve superior performance close to that of the backpropagation algorithm while taking advantage of the array architecture.

Fig. 4.3 represents the schematic illustration of a designed FC SNN for DFA. The specified architecture has 28 × 28 input neurons, 256 hidden neurons, and 10 output neurons. The internal synaptic arrays in the form of the AND architecture are located between adjacent neurons. The error information required for the external layers is provided by the difference between the output layer value and the teaching layer value. Note that, in DFA, the synaptic weights in the external synaptic arrays are not updated although the training is performed. Since the external synaptic array is directly connected between the output layer and the multiple hidden layers, which implies “direct” feedback alignment, the increase in the area due to the external synaptic arrays is relatively small compared to the overall area of the SNN. Here, two 256 × 10 excitatory and inhibitory arrays are required as the external synaptic arrays. The SNN using DFA consists of three

phases: Forward-propagation Phase (FP), Backward-propagation Phase (BP), and Update Phase (Update).

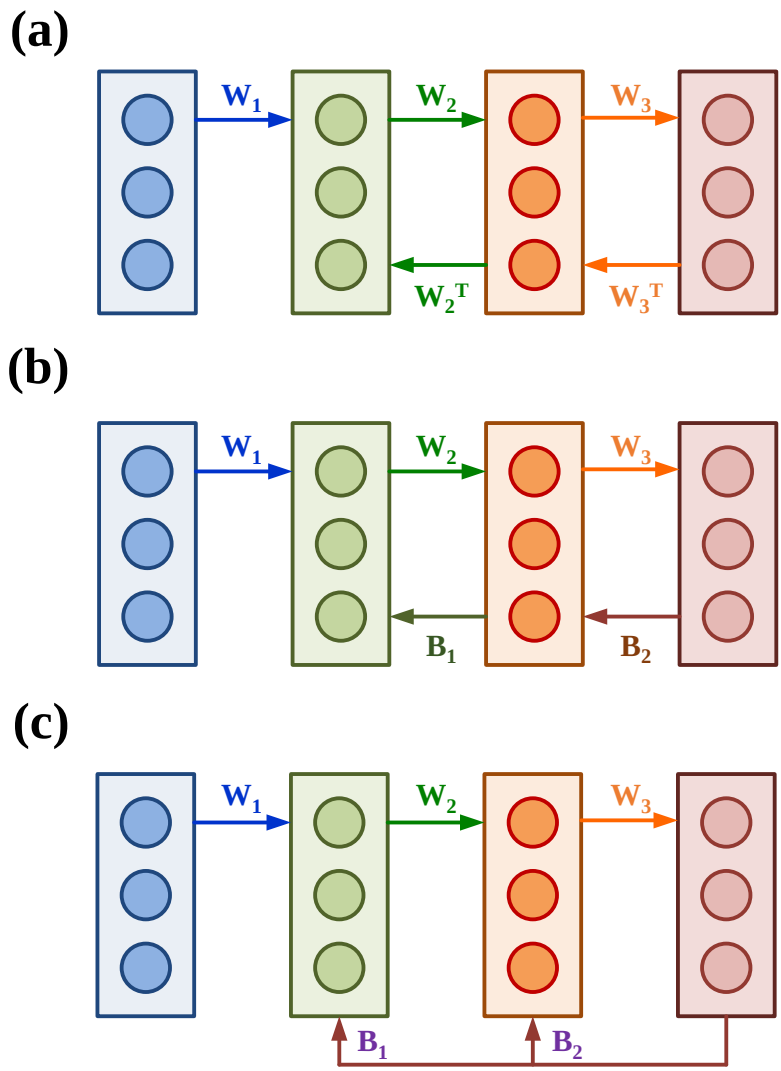


Fig. 4.1. Error transportation configuration in HNN utilizing (a) BP, (b) FA, and (c) DFA.

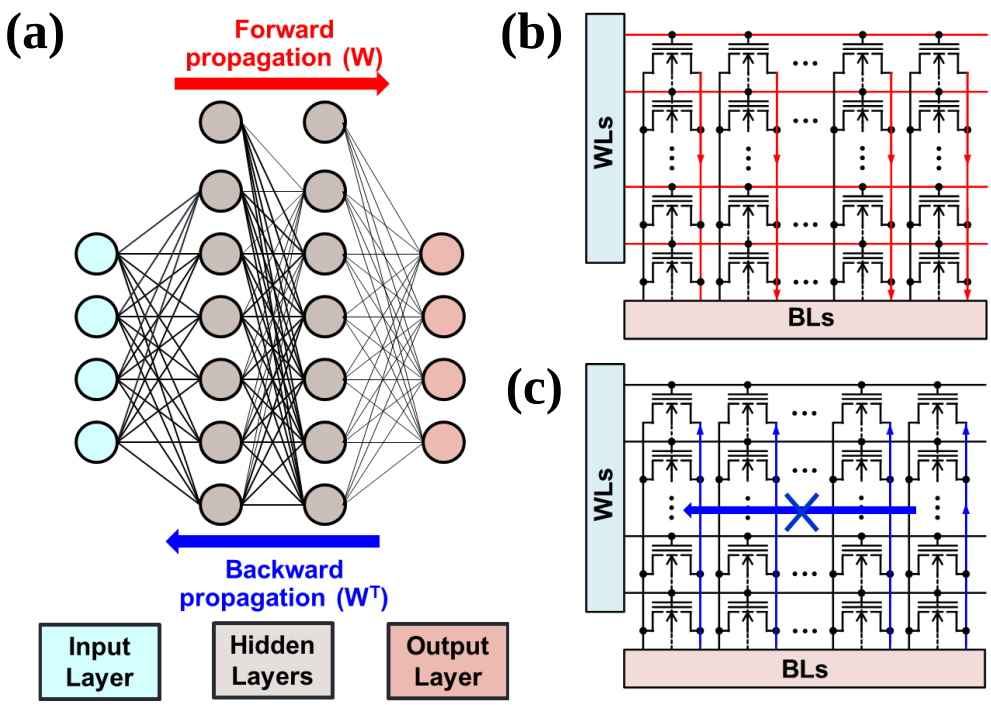


Fig. 4.2. (a) Symmetry of synaptic weights in HNN using BP algorithm. (b) Signal flow of forward path in synaptic array utilizing AND flash memory array. (c) Limitations of realization of backward path in synaptic array utilizing AND flash memory array.

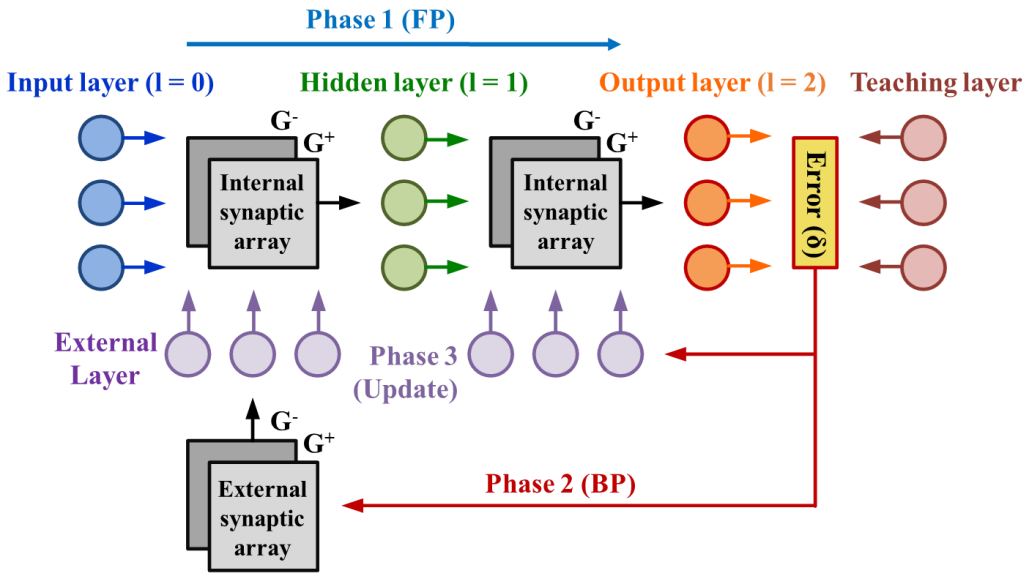


Fig. 4.3. Schematic illustration of a FC SNN for DFA.

4.2 Pulse scheme for DFA learning rule

The proposed pulse scheme for the DFA algorithm is represented in Fig. 4.4 and Fig. 4.5. Fig. 4.4 shows examples of pulses generated from each internal layer over time in the FP and BP. In the FP, for the inference, the read operation is performed in the internal synaptic array. A Poisson-distributed spike train (S^0) is generated at the input layer in the form of a voltage pulse with an amplitude and width of V_{read} and t_{w}, respectively. Then, the voltage spikes are applied to the WLs of the internal synaptic arrays, and the current summation is performed along the DLs. In the hidden layer l ($l \in \{1,\ldots, L\}$), the current summed in the DL is integrated into the membrane capacitor of the integrate and fire (IF) neuron to determine the membrane voltage as follows:

$$V_{\text{mem}}^{l}(t_s) = V_{\text{mem}}^{l}(t_s - 1) + \frac{\sum S^{l-1}(G_{+,int}^{l} - G_{-,int}^{l})t_w}{C_{mem}} \quad (1)$$

where V_{mem} is the membrane voltage, t_{s} is a time step ($t_{\text{s}} \in \{1, \cdots, T\}$), and C_{mem} is the membrane capacitance. Here, a synaptic weight is represented by $W = G_{+,\text{int}} - G_{-,\text{int}}$, where $G_{+,\text{int}}$ and $G_{-,\text{int}}$ are the conductance of the excitatory and inhibitory synaptic devices in the internal synaptic array, respectively. This means that two

devices are used to represent one synaptic weight. If the membrane voltage exceeds the specific V_{th} of the IF neuron, the IF neuron generates a spike as follows:

$$if\ V_{mem}^{l}(t_s) > V_{th}:\ \begin{cases} S^{l}(t_s) = V_{read} \\ V_{mem}^{l}(t_s) = V_{mem}^{l}(t_s) - V_{th} \\ g^{l} = 1 \end{cases} \quad (2)$$

$$else:\quad S^{l}(t_s) = 0 \quad (3)$$

where g is an approximated derivative of the neuron's activation function. When the FP begins for a given input signal, the g of all neurons' activation function are initialized to 0. When the IF neuron generates a spike, the corresponding g is set to 1. Note that the membrane potential in the fired neuron is reduced by the value of the selected V_{th} to prevent information loss. The generated spike from the IF neuron is the same type of spike as the input voltage spike, and it is applied to the WLs of the next internal synaptic array. This signal transmission process in the forward direction is similarly performed between the hidden and output layers. For this reason, the hidden and output layers can be configured with the same IF neuron circuit [65]. In the last layer ($l = L$), the number of spikes from the output neuron indicates the prediction value for the corresponding neuron. In the teaching layer,

the teaching spike is generated every t_s for the correct label, and the teaching spike is not generated for the wrong label. The number of teaching spikes and the number of output spikes from the output layer are compared to obtain the *delta* value in the output layer (δ^L).

In the BP, the error pulse whose width is modulated by the δ^L is applied to the WLs of the synaptic devices in the external synaptic arrays. The pulse-width modulation (PWM) circuit can easily modulate the width of the error pulse with the δ [66]. The width of the error pulse is set so that not only the error value but also the error sign can be expressed. For example, when the error value is 0, a reference pulse having a certain width can be set. And then, if the error value is positive, the error pulse with a width that is longer than the width of the reference pulse by the error value is generated. Conversely, if the error value is negative, an error pulse with a width shorter than the width of the reference pulse by the error value is generated. The delta value in the hidden layer (δ^l) is obtained by the current summation along the BLs in the external synaptic array and is stored in the capacitor in the neuron of the external layer as follows:

$$\delta^l = \frac{\sum \lambda \delta^L (G_{+,ext}^{l+1} - G_{-,ext}^{l+1})}{C_{BP}} \mathrm{g}^l \qquad (4)$$

where C_{BP} is the capacitance responsible for charge integration in the neuron in the external layer. Note that λ is a constant converting the delta value into the width of the voltage pulse with an amplitude maintained at V_{read}. The neuron in the external layer is used for receiving the summed current from the external synaptic array during the BP, and it stores the δ^l in the capacitor. After all δ^ls of the neurons are obtained in the BP, the synaptic weights in the internal array are updated during Update.

Fig. 4.5 (a) and (b) show examples of pulses applied to the WLs and PLs of the internal synaptic array during Update. During Update, the synaptic weights change simultaneously in a pair of internal synaptic arrays representing $G_{+,int}$ and $G_{-,int}$. Each time step during Update consists of two separate time steps. During the first period of each time step, the selective PGM operation is performed in the internal synaptic arrays, and depression of the internal synaptic devices occurs. On the other hand, the second period of each time step is for selective ERS operation in the synaptic arrays, and the potentiation of the internal synaptic devices occurs.

Similar to the backpropagation algorithm, the amount of synaptic weight change is obtained as $\Delta W^l \propto x^l \delta^{l+1}$, where x is the activated value of the pre-synaptic neuron. In the proposed hardware-based SNN, x is transformed into the input update pulse that correlates with the total number of spikes generated from the pre-synaptic neuron, and δ is converted into the delta update pulse with a width determined by the value of δ. First, the input update pulses from the pre-synaptic neurons are applied to the WLs of the internal synaptic arrays. The number of input update pulses with an amplitude of V_{PGM} in the first period of each time step is proportional to the number of spikes generated from the pre-synaptic neuron. In the second period of the time step, when no pulse is applied to the first period, an input pulse with an amplitude of V_{INH} is applied. Meanwhile, the delta update pulses are generated from the external layer and applied to the PLs of the internal synaptic arrays. The width of the delta update pulse applied to the synaptic device in the internal synaptic array is obtained as follows:

$$\Delta t_+^l = \lambda_+ |\delta^l|, \Delta t_-^l = \lambda_- |\delta^l| \quad (5)$$

where Δt_+^l and Δt_-^l are the widths of the delta update pulse for LTP and LTD,

respectively. Note that λ_+ and λ_- are related to the rate of the training. Also, different delta update pulses are applied to each of the internal synaptic arrays, representing $G_{+,\mathrm{int}}$ and $G_{-,\mathrm{int}}$ according to the sign of δ. As shown in Fig. 4.5 (a) and (b), for example, the delta update pulse corresponding to δ^+ applied to the internal synaptic array for $G_{+,\mathrm{int}}$ has the same form as that corresponding to δ^- applied to the internal synaptic array for $G_{-,\mathrm{int}}$. The delta update pulse is applied T times to the internal synaptic array during Update. As a result, in the internal synaptic array, the synaptic weights are updated by the delta update pulses applied to the PLs and the input update pulses applied to the WLs.

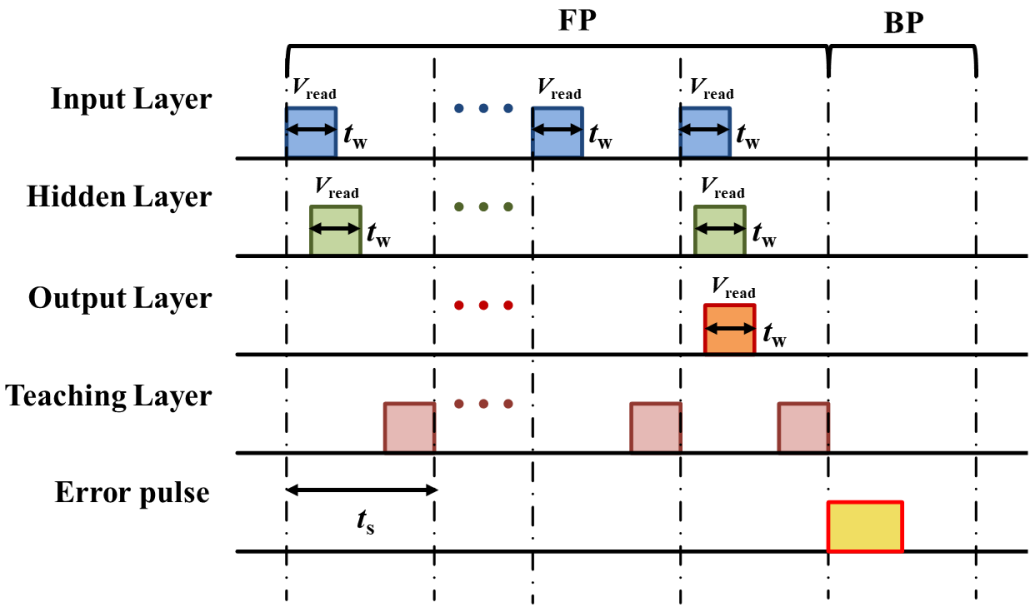


Fig. 4.4. Examples of pulses generated from each internal layer over time in the FP and BP.

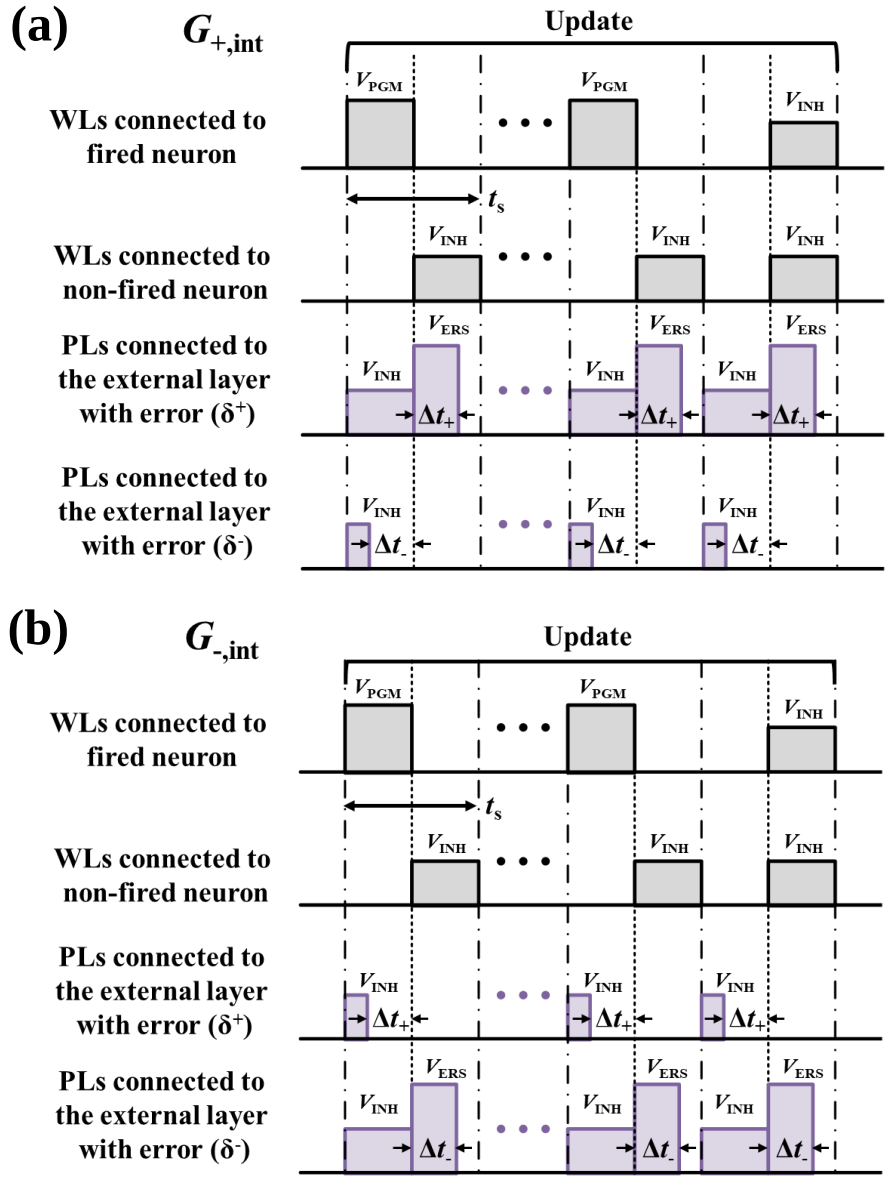


Fig. 4.5. Examples of pulses applied to the WLs connected to the internal layers and the PLs connected to the external layers in the Update. (a) and (b) show the excitatory and inhibitory synaptic devices in the internal synaptic array, respectively.

4.3 MNIST pattern learning and classification

Fig. 4.6 shows a flow chart representing the overall training process using the DFA in the designed network. The blue, red, and purple boxes indicates the FP, BP, and Update, respectively. A 1-hidden layer network (784-256-10) is designed, as described in Fig. 4.3. Fig. 4.7 shows the training curves of the designed hardware-based SNNs based on the DFA algorithm for the MNIST test set classification. The training is performed for 20 epochs, and the batch size of the training is 1 to minimize memory usage. As shown in Fig. 4.7, the SNN using ideal synaptic devices, which has a linear conductance response and no variation, achieves 97.7% for MNIST classification when $T = 10$ (total number of time steps). This accuracy is slightly lower than the accuracy of the ANNs in our previous work (98.2 %) [67], meaning that the DFA algorithm can show comparable training performance to the backpropagation algorithm. In addition, the accuracy of the SNNs using the AND flash memory array is evaluated. The nonlinearity and dynamic range of the synaptic device are set to the values when V_{read} and V_{DS} are 2.2 V and 2 V, respectively, as described in Fig. 2.8. Although the LTP curve of the device is near-

linear with respect to the number of erase pulses, the nonlinear LTD curve of the device degrades the accuracy of the SNNs (97.01%). This is because the expected weight updates cannot be exactly reflected in the conductance updates due to the nonlinear response. The inset shows the accuracy of the SNNs using the AND flash memory array depending on *T*. *T* is a key factor in improving the performance of the SNNs, since *T* determines the precision of the neuron's activation function. The IF neuron represents the intensity of the input current in the form of the number of spikes. Therefore, as *T* increases, the IF neuron represents the intensity more precisely within the extended *T*. Fig. 4.8 (a) and (b) show the recognition results of the 10 output neurons in the SNNs based on the AND flash memory array at the initial state and the end (epoch = 20) of the training, respectively. The output neuron number corresponds to the MNIST image number, and the image can be classified within 10 time steps. As shown in Fig. 4.8 (b), only the output neuron that has learned a particular image generates frequent spikes, while the other neurons rarely generate spikes.

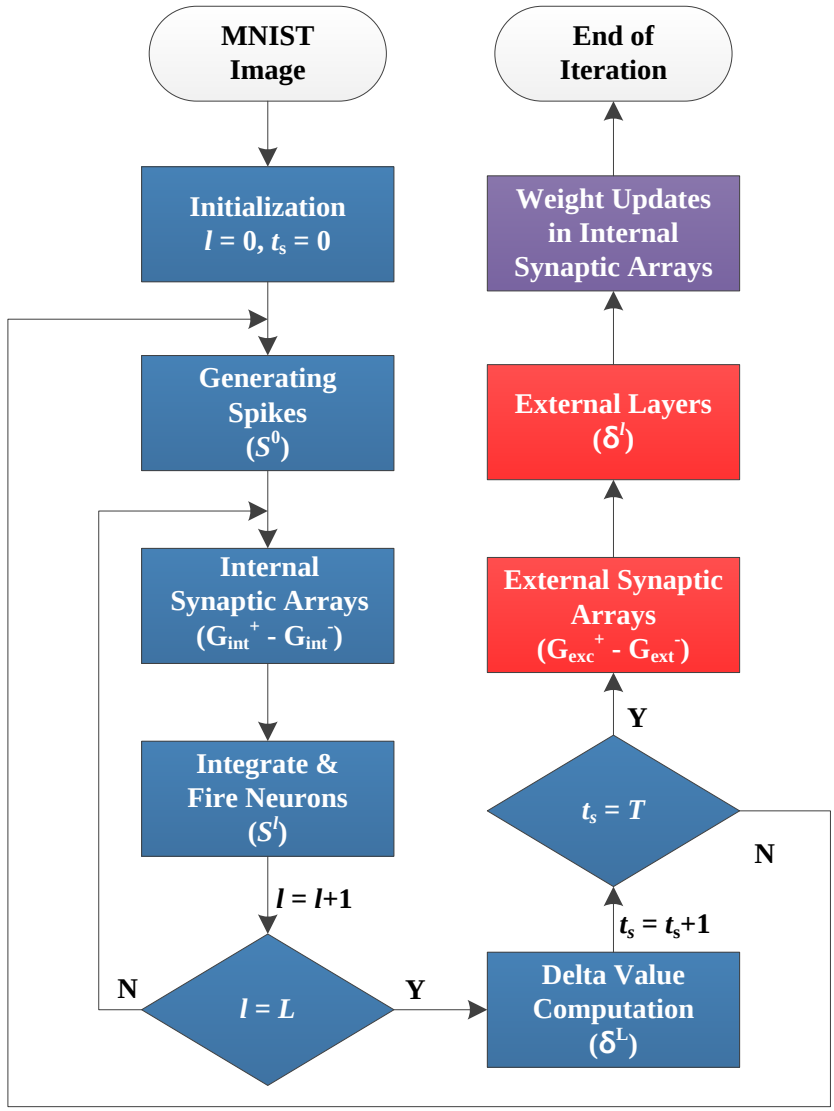


Fig. 4.6. Flow chart of the training process in the hardware-based SNNs using DFA algorithm.

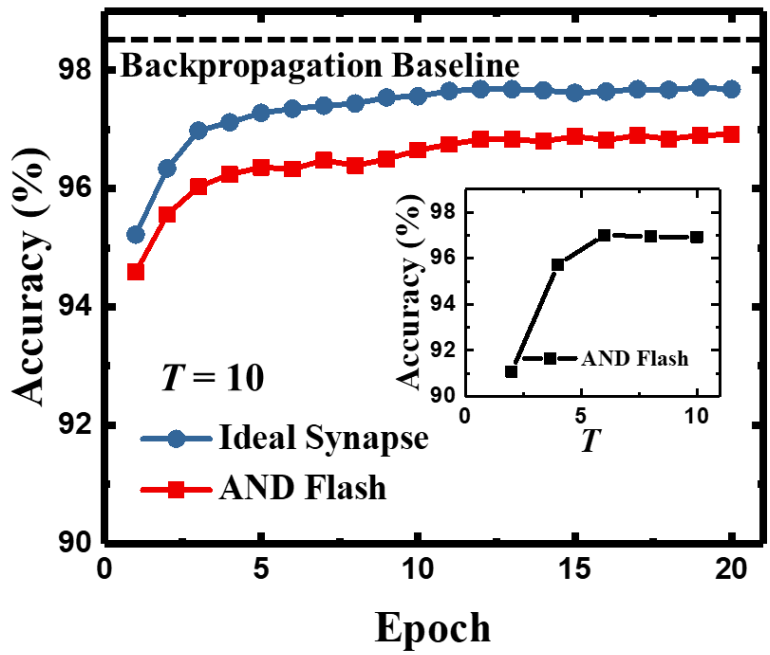


Fig. 4.7. Training curves of the hardware-based SNN based on the DFA algorithm for the MNIST test set classification. The inset represents the accuracy of the SNNs using the AND flash memory array depending on T.

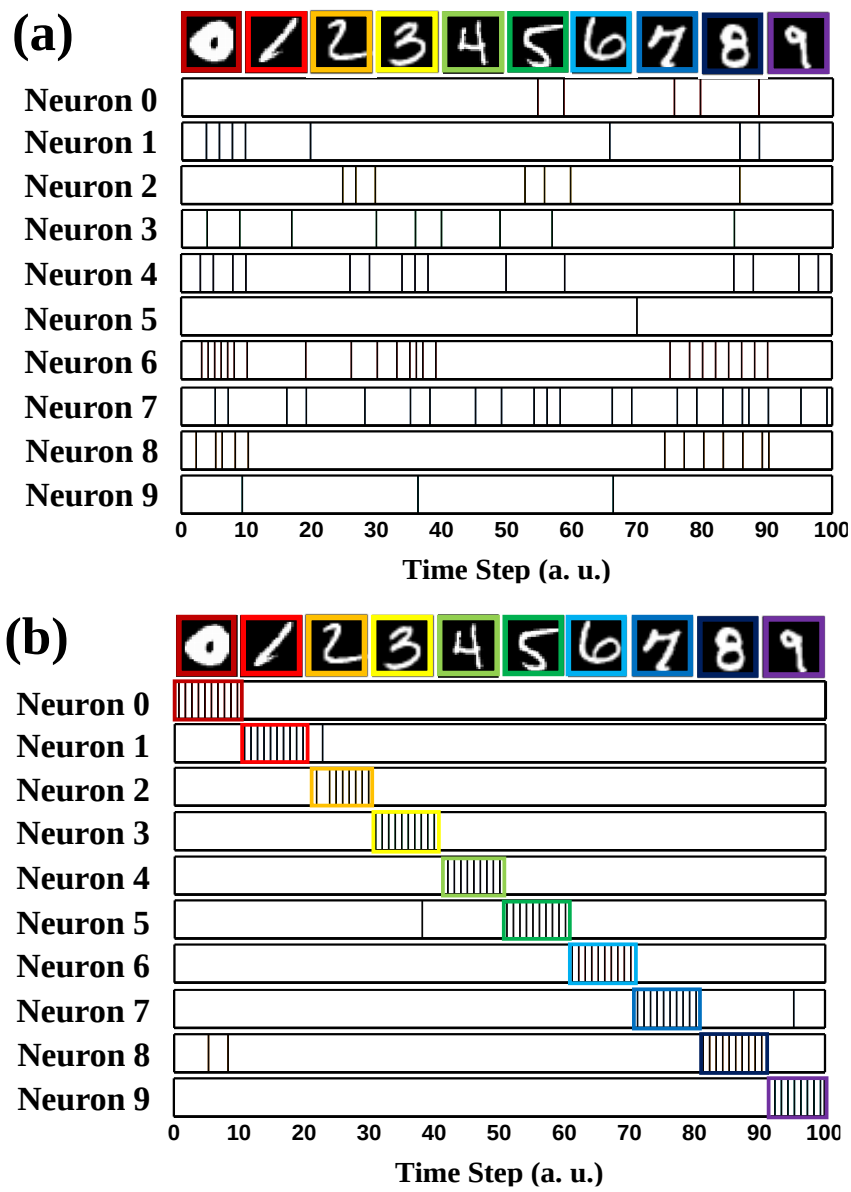


Fig. 4.8. Recognition results of the 10 output neurons in the designed SNN using DFA based on the AND flash memory array at (a) the initial state and (b) the end of the training.

Chapter 5
Hardware implementation of neural networks

5.1 Integration of a synaptic array and CMOS circuits

To implement the designed hardware-based neural networks, not only the synaptic array but also the peripheral circuits should be supported. For example, an integrate-and-fire (I&F) circuit required for the neuron module, or various types of circuits that convert the output pulse from the I&F circuit into the input pulse of the synaptic array should be included in the chip design. Therefore, to target a neuromorphic chip for implementing the neural network, research on the effective integration fabrication of the synaptic array and CMOS circuits should be preceded. In this chapter, the integration fabrication method of the proposed synaptic array and CMOS circuits is proposed, and the results of experimental verification of the operation of the synaptic array and several CMOS circuits are presented.

The proposed AND-type synaptic array [68] and CMOS circuit were fabricated on a 6-inch *p*-type single-crystalline Si wafer with the (100) orientation using 12

masks and conventional CMOS process technology. The used masks are marker formation (1st), *n*-well / *p*MOS channel implantation (2nd), active (CMOS) define (3rd), *n*MOS field / *n*MOS channel implantation (4th), *p*-body (synaptic device) formation (5th), gate (CMOS) & S/D (synaptic device) formation (6th), *p*MOS S/D & *p*-body contact (synaptic device) implantation (7th), *n*MOS S/D implantation (8th), channel (synaptic device) define (9th), gate (synaptic device) formation (10th), contact hole (11th), and metal line formation (12th). The 2nd and 4th masks are used separately in the specified fabrication steps, respectively, and the 6th and 7th masks are used once each to share the process steps of the synaptic array and CMOS circuit. Full fabrication processes were carried out in Inter-University Semiconductor Research Center located in Seoul National University.

Fig. 5.1 and Fig. 5.2 show the schematic cross-sectional views of the key fabrication steps and overall process flow of the proposed synaptic array and CMOS circuit integration, respectively. First, after a standard cleaning process including sulfuric peroxide mixture (SPM), ammonium hydroxide-hydrogen peroxide mixture (APM), hydrochloric acid-hydrogen peroxide-water mixture (HPM), and

diluted hydrogen fluoride (DHF), the area to be the reference marker for all fabrication is patterned (1st mask). This is to compensate for the loss of the reference marker due to the CMP in the overall fabrication. A 10-nm-thick SiO_2 layer was deposited as a sacrificial oxide by a low-pressure chemical vapor deposition (LPCVD). And, a phosphorus ion implantation with a dose of 3.0×10^{12} cm^{-2} and energy of 120 keV for *n*-well doping was performed (2nd mask). A diffusion process by annealing at a temperature of 1100 ℃ for 11 hours was followed by removing the sacrificial oxide by wet etching in 100:1 DHF. After a 10-nm-thick SiO_2 layer was formed via a dry oxidation process at 950 ºC, a 150-nm-thick Si_3N_4 layer was deposited by the LPCVD process. Then, a layer of Si_3N_4/SiO_2 was patterned by a reactive ion etching (RIE) to define the regions that will be the active area of the CMOS (3rd mask). And, *n*MOS field implantation was performed by a boron ion implantation with a dose of 1.6×10^{13} cm^{-2} and energy of 40 keV (4th mask). Field implantation is for the isolation of *n*MOS in the CMOS circuit operation, and compensation for the boron concentration that can be lost in the process of the field oxide (FOX) formation. A field oxide was grown thermally by the wet oxidation

process at 1000 ℃, and the oxynitride, nitride, and pad oxide were stripped in sequence. After a 30-nm-thick SiO_2 layer was formed by dry oxidation at 950 ℃, a 10-nm-thick sacrificial oxide was formed by wet etching in 100:1 DHF. This step is to solve the problem called "white ribbon" that occurs along the edge of the active area in the local oxidation of silicon (LOCOS) process. Then, a boron implantation ion implantation with a dose of 4.0×10^{12} cm^{-2} and energy of 28 keV for *n*MOS channel doping was performed (4^{th} mask). The mask used in this fabrication step is the same as the mask used in the *n*MOS field implantation in the previous step. And, a *p*MOS channel implantation by BF_2^+ ions with a dose of 2.7×10^{12} cm^{-2} and energy of 25 keV was followed by a *p*MOS punch-through implantation by P^+ ions with a dose of 1.1×10^{12} cm^{-2} and energy of 110 keV (2^{nd} mask). The mask used in this fabrication step is the same as the mask used in the *n*-well implantation in the previous step. The channel implantation of the CMOS is to control the V_{th} of the *n*MOS and the *p*MOS. In particular, the punch-through implantation of the *p*MOS was included in the fabrication process because it uses n^+-doped poly-Si as a gate material and is designed to use a buried channel. After a 250-nm-thick poly-Si layer

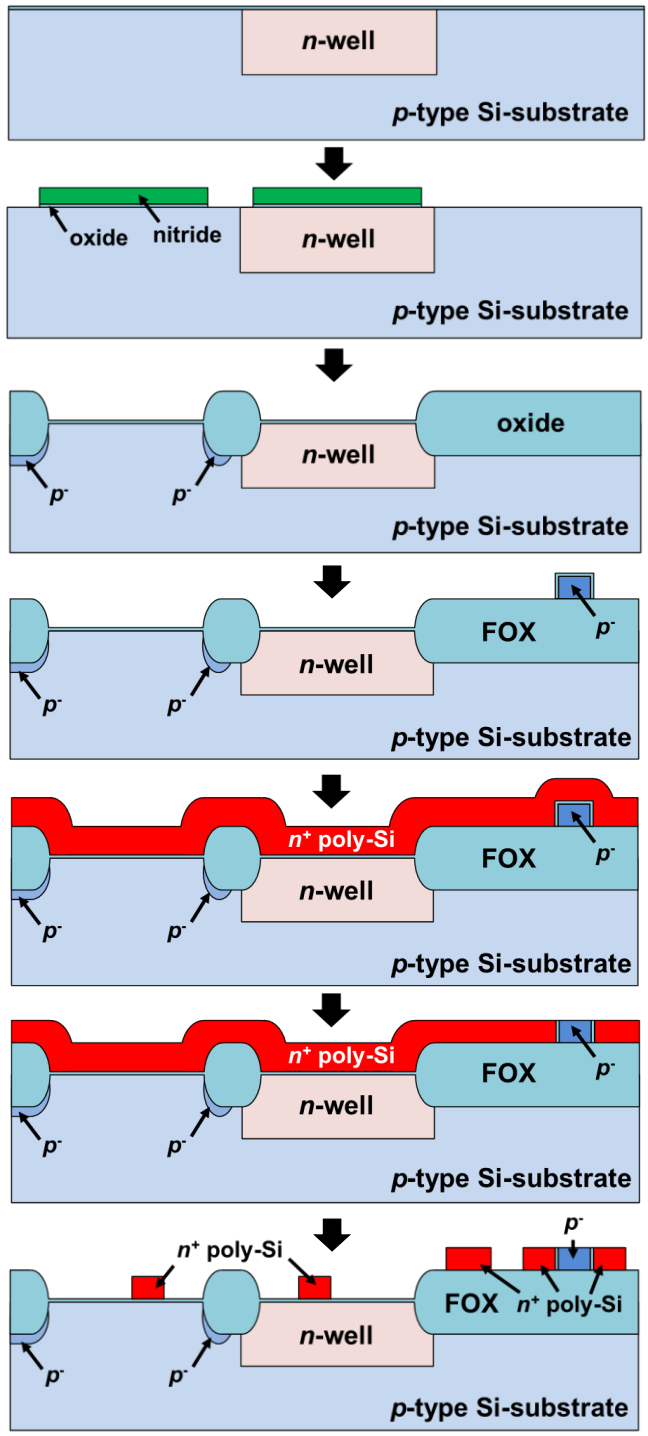
was formed by the LPCVD process at 625 ℃, a 10-nm-thick sacrificial SiO_2 layer was deposited. After BF_2^+ implantation with a dose of 5.0×10^{12} cm^{-2} and energy of 150 keV was performed, the sacrificial oxide was removed by wet etching in 100:1 DHF. Then, the poly-Si layer was patterned by the RIE process to form the *p*-body of the synaptic device on the FOX (5^{th} mask). After removing the sacrificial oxide in the active area of the CMOS, a 10-nm-thick SiO_2 layer was formed via dry oxidation at 850 ℃. At this time, a 12-nm-thick SiO_2 film was also formed on both sides and the upper surface of the patterned poly-Si layer on the FOX. This oxide is for the separation of the *p*-body and the S/D of the synaptic device. Then, the deposition of a 300-nm-thick *in situ* n^+-doped poly-Si was deposited, and the CMP process was performed. While the n^+ poly-Si layer, which will become the S/D of the synaptic device, is flattened through the polishing process, the n^+ poly-Si layer, which will become the gate of the CMOS, is hardly affected as shown in Fig. 5.3. Removing the oxide exposed on the top of the p-body of the synaptic device by wet etching in 100:1 DHF was followed by the CDE for lowering the thickness of the n^+-doped poly-Si layer. After patterning the n^+-doped poly-Si layer by the RIE

process to form the gate of the CMOS and the S/D of the synaptic device (6^{th} mask), a phosphorus ion implantation with a dose of 5.0×10^{13} cm^{-2} and energy of 10 keV for the lightly doped drain (LDD) implantation of the *n*MOS was performed (8^{th} mask). The mask used in LDD implantation is the same as the mask that will be used for the S/D implantation of the *n*MOS in the later step. A 60-nm-thick SiO_2 film was deposited and anisotropically etched by the RIE process to form SiO_2 film spacers on both sides of the patterned n^+-doped poly-Si layer. And, the S/D implantation of the *p*MOS by BF_2^+ ions with a dose of 2.0×10^{15} cm^{-2} and energy of 25 keV (7^{th} mask) was followed by the S/D implantation of the *n*MOS by As^+ ions with a dose of 2.0×10^{15} cm^{-2} and energy of 40 keV (8^{th} mask). During the S/D implantation of the *p*MOS, the *p*-body contact of the synaptic device on the FOX is also implanted at the same time. The implanted ions are then activated and diffused by rapid thermal annealing (RTA) at a temperature of 1050 ℃ for 5 sec. Then, a 13-nm-thick amorphous Si layer was deposited by the LPCVD at 550 ℃ as a channel material of the synaptic device and poly-crystalized by annealing at 600 ℃ for 24 hours. For isotropic etching of the channel of the synaptic device, a 20-nm-thick

SiO_2 layer was deposited and patterned by wet etching through DHF (9^{th} mask). This was followed by the CDE process for patterning of the channel of the synaptic device. A gate insulator stack of SiO_2 / Si_3N_4 / Al_2O_3 as the tunneling oxide layer/charge storage layer/blocking oxide layer was deposited. Al_2O_3 film was formed through the atomic layer deposition (ALD) process, and the remaining layers were deposited by the LPCVD process at 780 ℃. Similar to the channel etching process of the synaptic device, the deposition of a 30-nm-thick layer of TiN by metal-organic chemical vapor deposition (MOCVD) was followed by the deposition of a 300-nm-thick SiO_2 layer. After patterning the SiO_2 film by wet etching using buffered oxide etchant (BHF), the TiN layer was also patterned by wet etching using diluted hydrogen peroxide (10^{th} mask). After the deposition of TEOS by a plasma-enhanced CVD (PECVD) process, contact holes were formed by the RIE process (11^{th} mask). And then, Ti (30 nm) / TiN (30 nm) / Al (300 nm) / TiN (30 nm) metal wires were formed through sputtering and patterned (12^{th} mask). Finally, H_2 annealing at 400 ℃ for 10 min was performed was performed to improve the interface characteristics.

A bird's eye view of the synaptic array and CMOS circuit integration is represented in Fig 5.4. This fabrication method aims to efficiently integrate the synaptic array and the CMOS circuit on a single wafer. As explained in the detailed fabrication steps, this fabrication method has the advantage of reducing the number of masks and steps due to the shared process of the synaptic array and CMOS circuit. First, the gate of the CMOS and the S/D of the synaptic device use the same material as n^+ poly-Si. Moreover, contact doping of the *p*-body of the synaptic device can be formed in the S/D implantation process of the *p*MOS. In conclusion, it is possible to reduce the total number of masks by four, including the formation process of contact holes and electrodes. Additionally, the design of passive devices such as electrical capacitors and resistors is also possible in the proposed integration fabrication method. The electrical capacitors are formed through the gate insulator between n^+-doped poly-Si and TiN material. The electrical resistors can be also formed through the *p*-body line used when designing the synaptic array. A relatively lightly-doped *p*-body line has the advantage of efficiently designing an area occupied by the electrical resistors. As a result, passive devices can be designed in

the same process without using additional masks. In the stated fabrication method, there is a total of one metal line, but even if the fabrication process is designed with several metal layers, the advantages of the fabrication method are the same. In addition, although the CMOS circuit is formed through the LOCOS process in the proposed fabrication method, the advantages of the fabrication method can be maintained even when the isolation oxide is formed through the shallow trench isolation (STI) process.


n-well
p-type Si-substrate
oxide
nitride
n-well
p-type Si-substrate
oxide
n-well
p-
p-
p-type Si-substrate
FOX
p-
n-well
p-
p-
p-type Si-substrate
n+ poly-Si
FOX
p-
n-well
p-
p-
p-type Si-substrate
n+ poly-Si
FOX
p-
n-well
p-
p-
p-type Si-substrate
n+ poly-Si
p-
FOX
n+ poly-Si
n-well
p-
p-
p-type Si-substrate

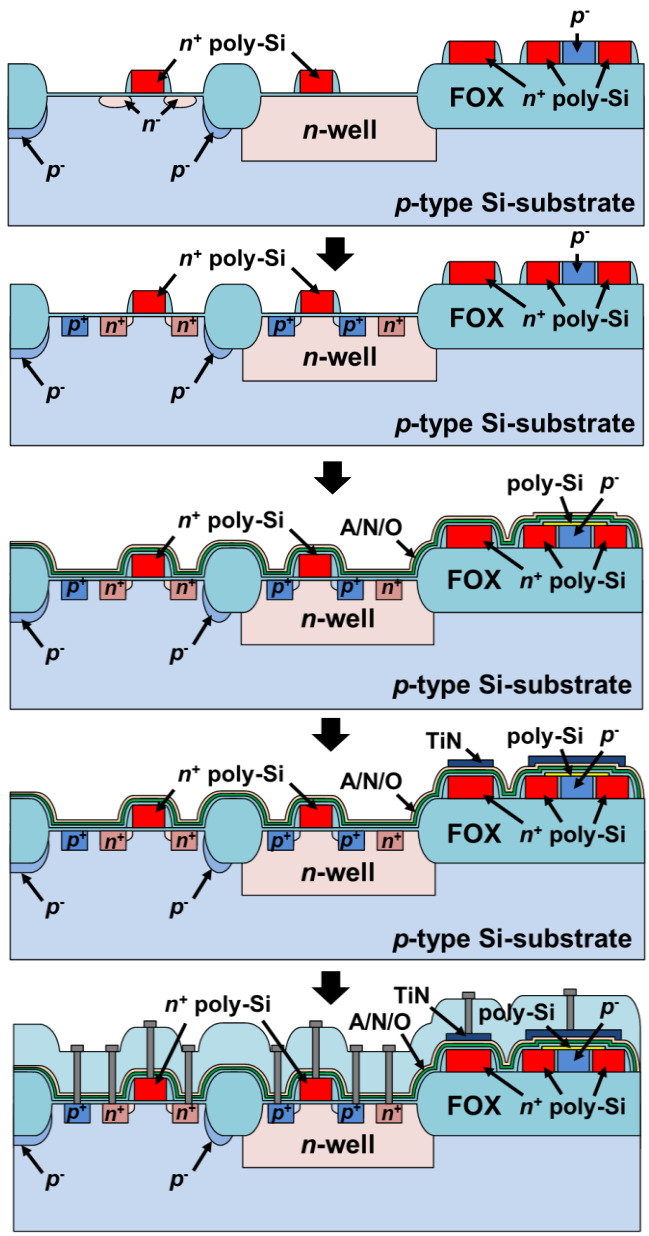


Fig. 5.1. Schematic cross-sectional views in the key fabrication steps of the integration of the proposed synaptic array and CMOS circuit.

- **6-inch (100) *p*-type Si wafer**
- **Marker photo (1st mask) and patterning**
- **Oxide deposition (t = 100 Å), *n*-well photo (2nd mask)**
- ***n*-well implant (P^+ 120 keV 3e12)**
- **Drive-in (1100 ºC, 11 hour), and oxide strip**
- **Dry oxidation (t = 100 Å), nitride deposition (t = 1500 Å)**
- **Active photo (3rd mask) and patterning**
- ***n*MOS field implant photo (4th mask)**
- ***n*MOS field implant (B^+ 40 keV 1.6e13)**
- **Wet oxidation (field oxide) (t=5500 Å)**
- **Oxynitride strip, nitride strip, pad oxide strip**
- **Dry oxidation (t=300 Å), oxide etch back (t=100 Å)**
- ***n*MOS channel implant photo (4th mask)**
- **nMOS channel implant (B^+ 28 keV 4e12)**
- ***p*MOS channel implant photo (2nd mask)**
- ***p*MOS channel implant (BF_2^+ 25 keV 2.7e12)**
- **Punch-through implant (P^+ 110 keV 1.1e12)**
- **Poly-Si deposition (t=2500 Å), oxide deposition (t=100 Å)**
- ***p*-body (synapse) implant (BF_2^+ 150 keV 5e12)**
- ***p*-body photo (5th mask) and patterning**
- **Oxide strip**
- **Dry oxidation**
  (formation of $t_{gate\text{-}oxide(CMOS)}$ and $t_{separate\text{-}oxide(Synaptic\ device)}$)
- **n^+ poly-Si deposition (t=3000 Å)**
- **Chemical mechanical polishing (CMP)**
- **DHF and chemical dry etching (CDE)**
- **n^+ poly-Si gate (CMOS) & S/D (synapse) photo (6th mask)**
- **n^+ poly-Si gate (CMOS) & S/D (synapse) patterning**

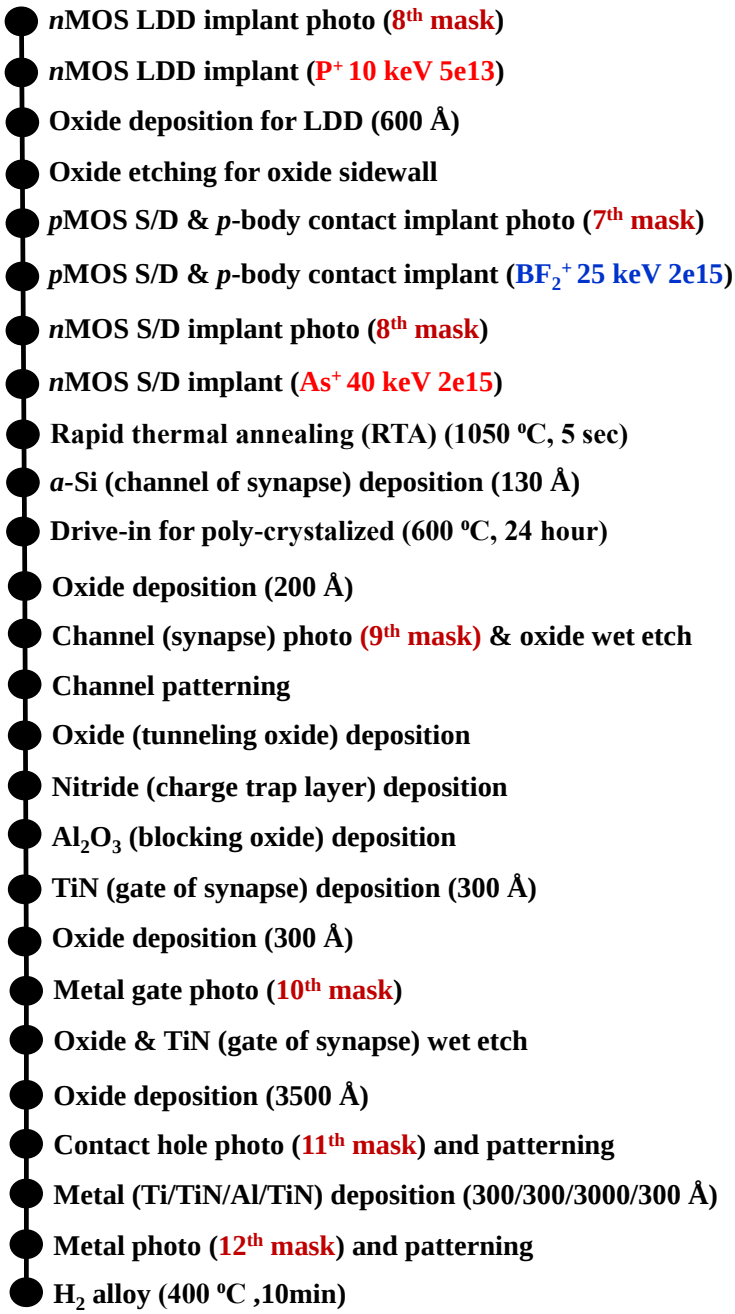


Fig. 5.2. Overall process flow of the integration of the proposed synaptic array and CMOS circuit.

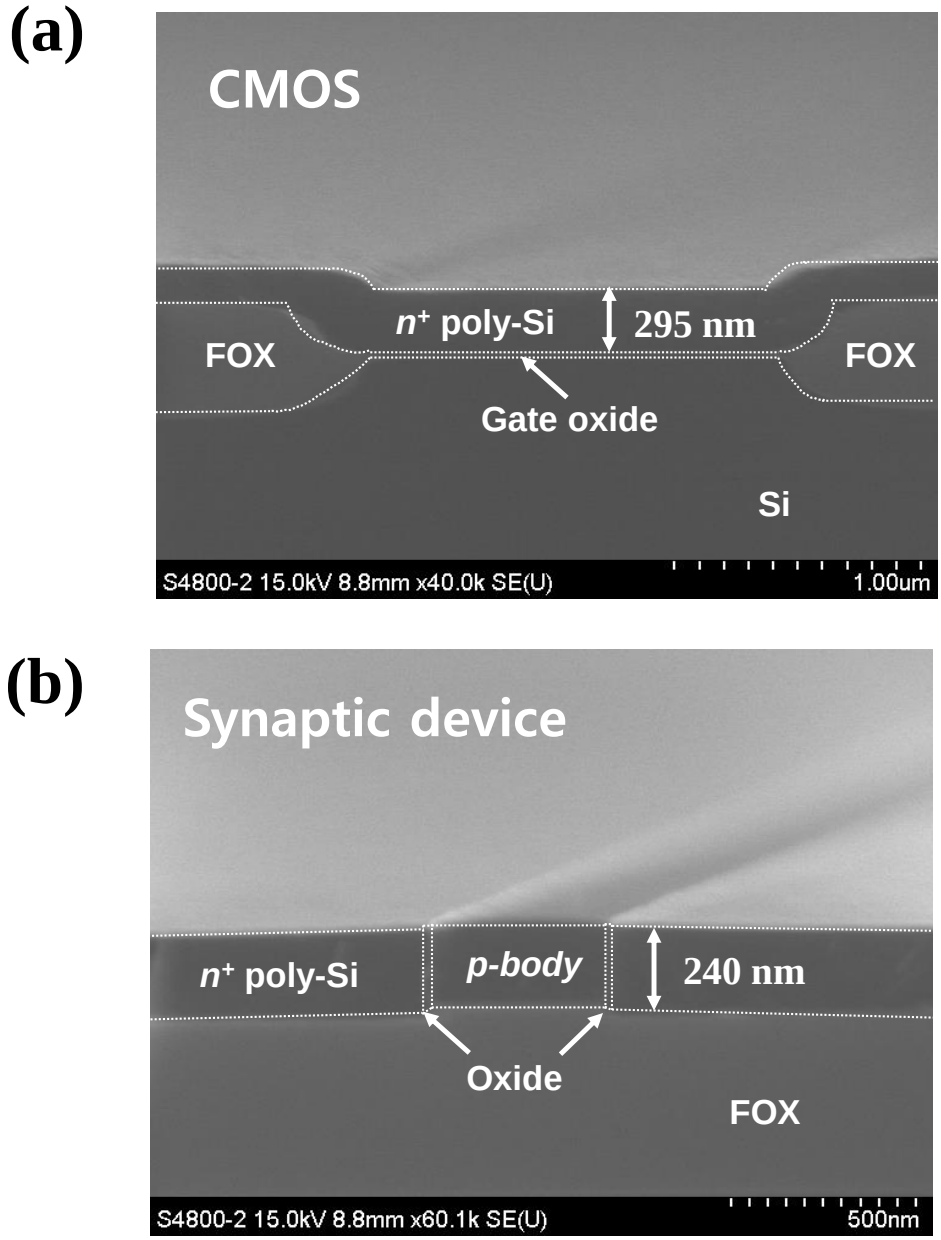


Fig. 5.3. A SEM image of the fabricated (a) MOSFET and (b) TFT-type single synaptic device after the CMP process.

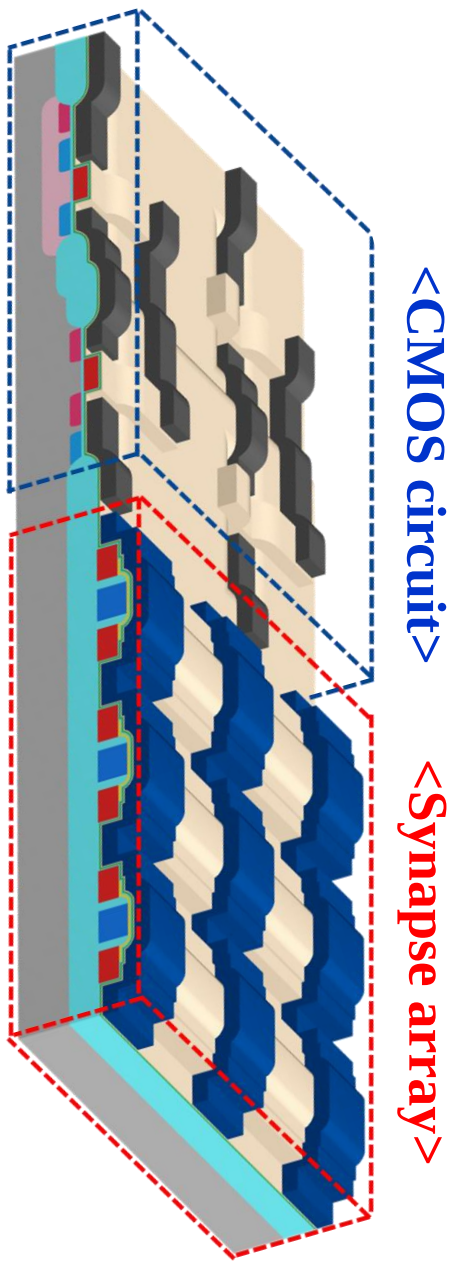


Fig. 5.4. A bird's eye view of the synaptic array and CMOS circuit integration.

5.2 Measurement results of a synaptic array

Fig. 5.5 (a) shows a TEM image of a fabricated TFT-type single synaptic device in the integration fabrication of the synaptic array and CMOS circuit. The gate insulator stack (A/N/O) in the fabricated synaptic device is represented in Fig. 5.5 (b). There are two differences from the synaptic device discussed in Chapter 2. First, the SiO_2 film between the *p*-body and S or D is simultaneously formed via dry oxidation when forming the gate oxide of CMOS. By forming a thinner thickness than the same SiO_2 film of the synaptic device discussed in Chapter 2, it is possible to increase the voltage coupling effect during selective PGM or ERS operation in the synaptic array. Second, H_2 annealing included in the process of the CMOS circuit is additionally performed, thereby improving the interface characteristics between the poly-Si and the gate insulator stack. The measured $I_D - V_{GS}$ characteristic of a fabricated TFT-type single synaptic device through the integration fabrication of the synaptic array and CMOS circuit is shown in Fig. 5.6. Compared to the synaptic device discussed in Chapter 2, improved *SS* and on-current characteristics can be seen.

Fig. 5.7 (a) and (b) show the I_D – V_{GS} characteristics measured when an identical PGM or ERS pulse under specific conditions is applied several times to the fabricated synaptic device through the integration fabrication of the synaptic array and CMOS circuit. In the inset, the bias conditions for the PGM and ERS are specified. It can be seen that the voltage required for the same ERS operation performed in chapter 2.3 is significantly reduced due to the reduction in the thickness of the tunneling oxide included in the gate insulator stack. This is the same as meaning that the width of the pulse required for the same ERS operation can be reduced. Fig. 5.8 represents the measured I_D – V_{GS} characteristics when identical PGM or ERS pulse with relatively shorter width is applied several times to the fabricated synaptic device through the integration fabrication of the synaptic array and CMOS circuit. The fact that memory operation is possible with pulses having a relatively short width and low voltage means that the burden on the circuit generating the pulses required for memory operation can be reduced. In addition, the power consumed in the circuit generating the pulse required for the memory operation can be reduced although power consumption does not occur in the

memory operation on the AND-type array architecture.

Fig. 5.9 shows a SEM image of the fabricated 15 x 3 AND flash memory array through the integration fabrication of the synaptic array and CMOS circuit. To verify the selective memory operation of the synaptic devices in the synaptic array, specific number patterns are learned. Fig. 5.10 represents a bias condition for measurement in a selective PGM operation of specific cells in the 15 × 3 AND flash memory array. During the PGM operation of a specific cell, a PGM pulse with a positive V_{PGM} is applied to WL connected to the specific cell, and the bias of DL, SL, and PL connected to the specific cell is maintained at 0 V. During this period, the WLs of cells that do not want PGM operation while located in the same BL maintains 0 V. In addition, an INH pulse with an amplitude of V_{INH1} is applied to other PLs to prevent the PGM operation of cells crossing the WL to which the PGM pulse is applied. Note that the other SL and PL are floated. On contrary, Fig. 5.11 represents a bias condition for measurement in a selective ERS operation of specific cells in the 15 × 3 AND flash memory array. During the ERS operation of the selected cell, an ERS pulse with a positive V_{ERS} is applied to PL connected to the

selected cell, and the bias of WL is maintained at 0 V. During this period, an INH pulse with an amplitude of V_{INH2} is applied to the WLs of cells that are in the same BL and do not want ERS. Note that the both SLs and PLs in the array are floated during the selective ERS operation. Fig. 5.12 shows the measurement results of learning specific number patterns in the fabricated flash memory array. It can be seen that the number patterns of 2 in BL1, 4 in BL2, and 7 in BL3 are learned, respectively. The current values of the synaptic devices specified in the number patterns are the values measured when the read voltage is 2 V. Fig. 5.13 represents the measured $I_D - V_{GS}$ characteristics of each cell in the fabricated flash memory array after learning the specific number patterns. The synaptic devices for which each number pattern is to be learned are in the ERS state, and the synaptic devices corresponding to the area without a number pattern are in the PGM state. In addition, Fig. 5.14 shows the measurement result of each I_{BL} according to the input pattern in the fabricated flash memory array after learning the specific number patterns. When an input such as a learned number pattern is applied to WL, it can be confirmed that the learned BL has the largest current value.

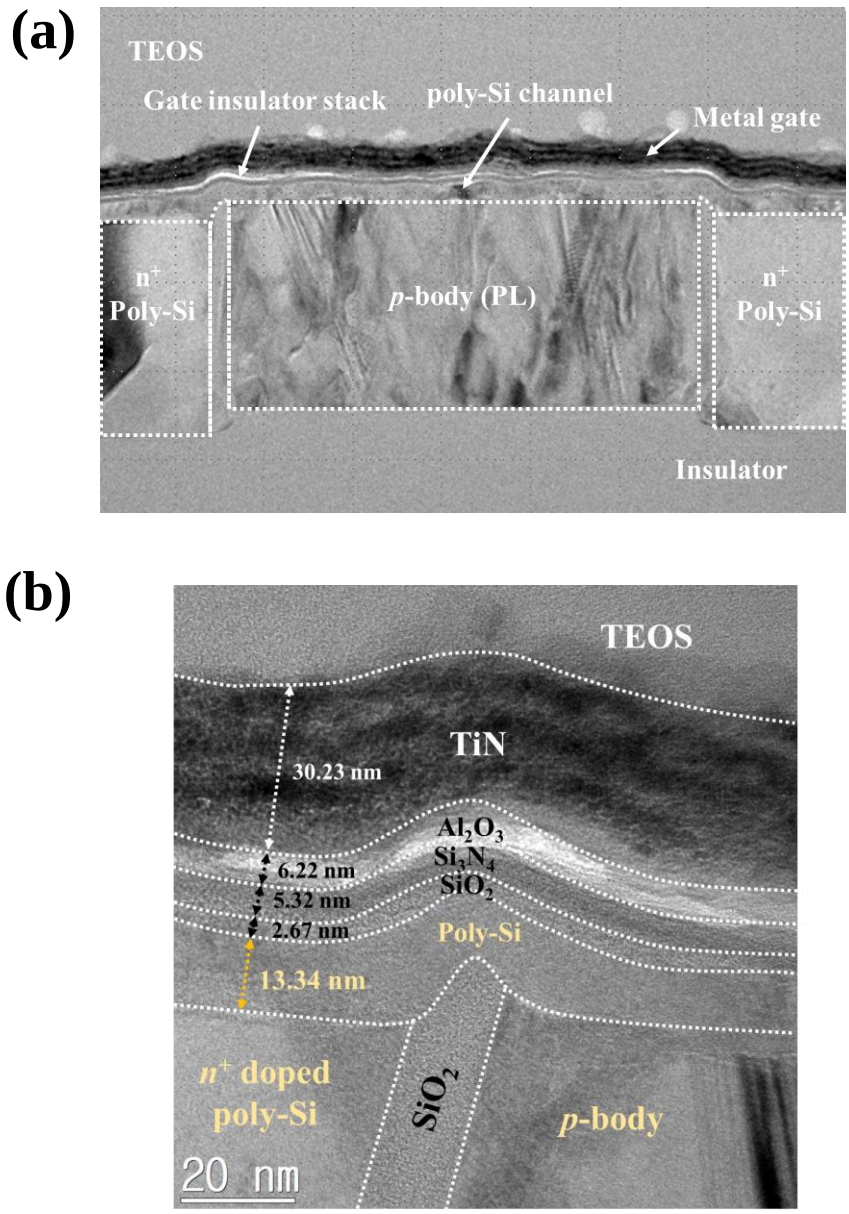


Fig. 5.5. A cross-sectional TEM image of (a) a fabricated TFT-type single synaptic device through the integration fabrication of the synaptic array and CMOS circuit. (b) The gate insulator stack (A/N/O) through the fabricated synaptic device.

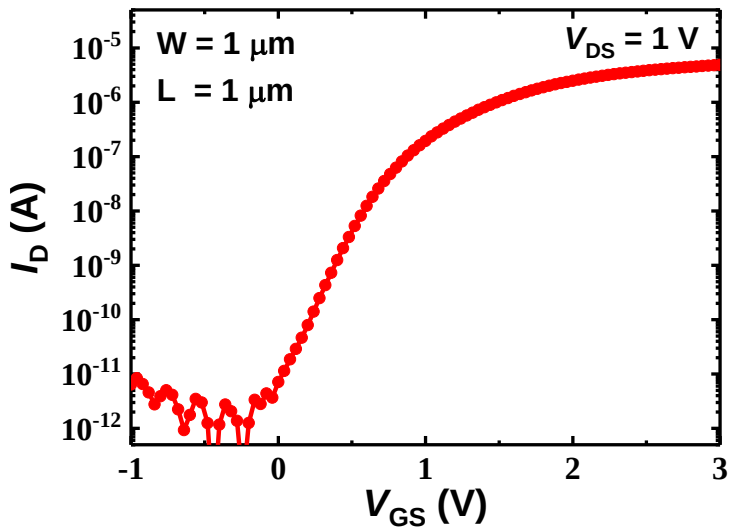


Fig. 5.6. Measured I_D – V_{GS} characteristic of the fabricated TFT-type single synaptic device through the integration fabrication of the synaptic array and CMOS circuit.

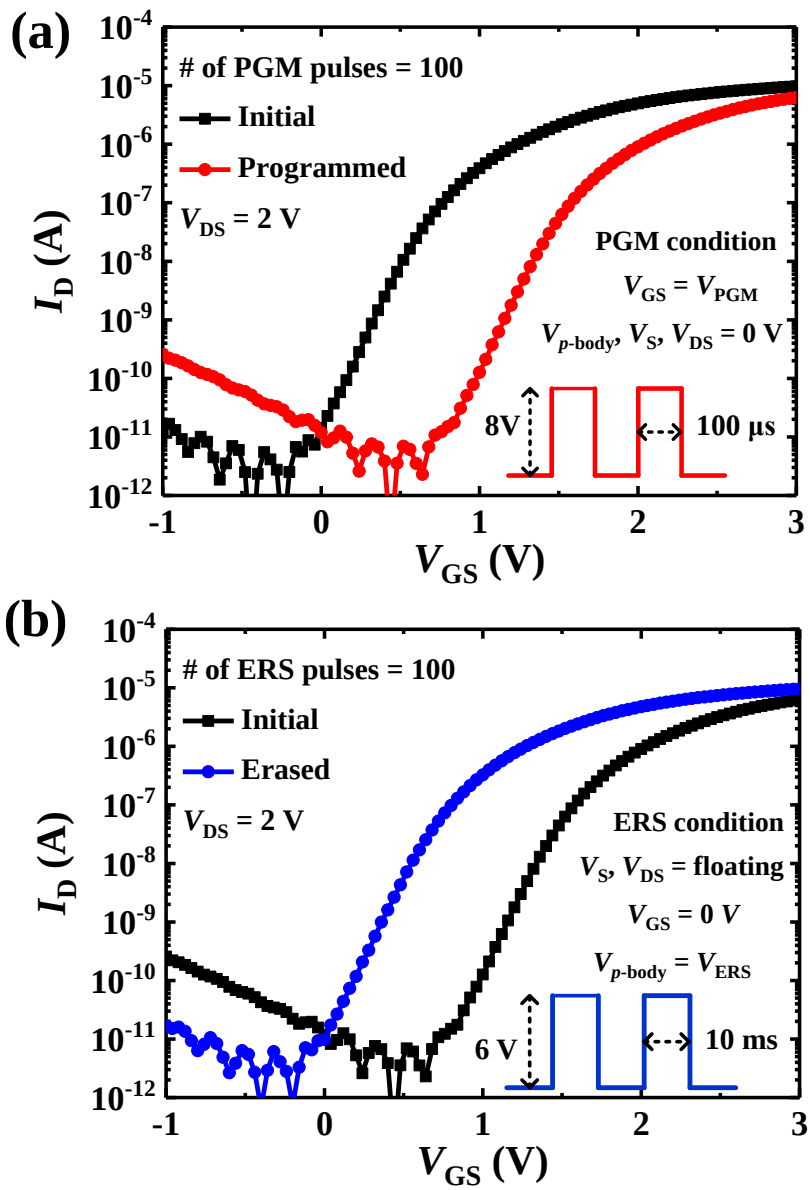


Fig. 5.7. Measured I_D – V_{GS} characteristics when identical (a) PGM or (b) ERS pulse (with relatively longer width) is applied several times to the fabricated synaptic device through the integration fabrication of the synaptic array and CMOS circuit.

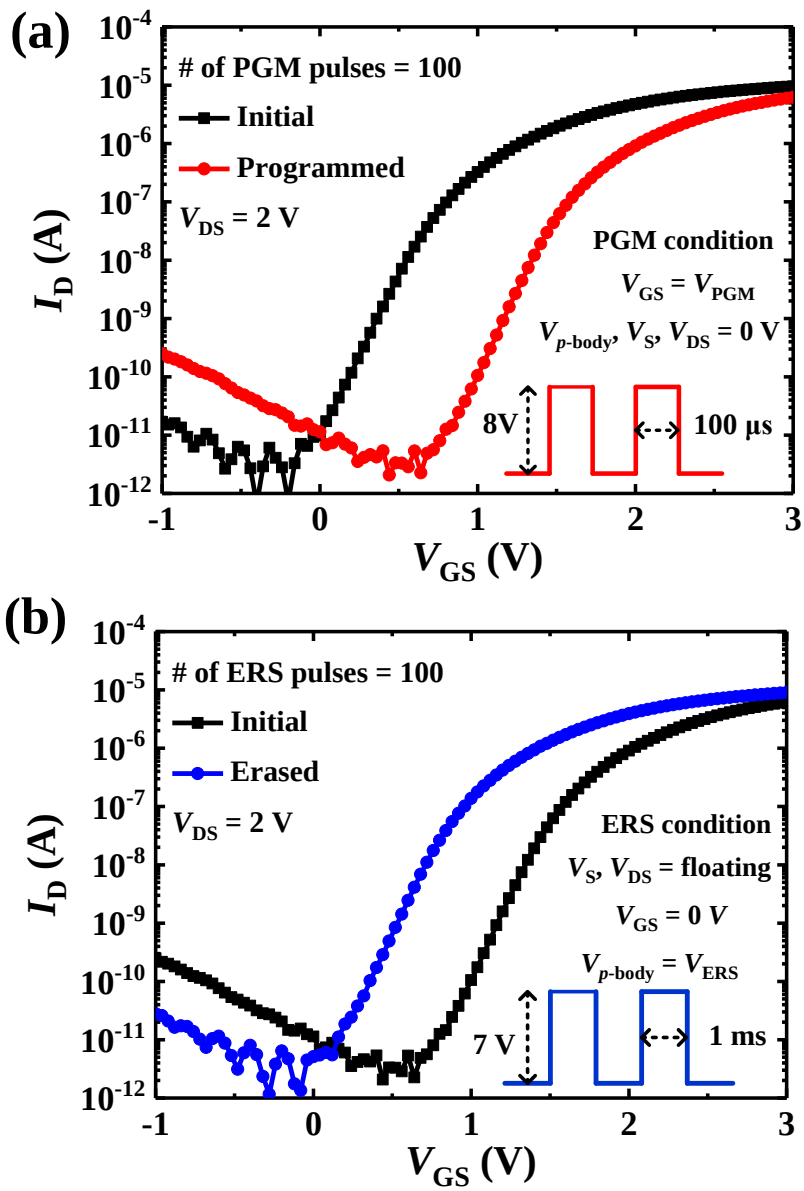


Fig. 5.8. Measured I_D – V_{GS} characteristics when identical (a) PGM or (b) ERS pulse (with relatively shorter width) is applied several times to the fabricated synaptic device through the integration fabrication of the synaptic array and CMOS circuit.

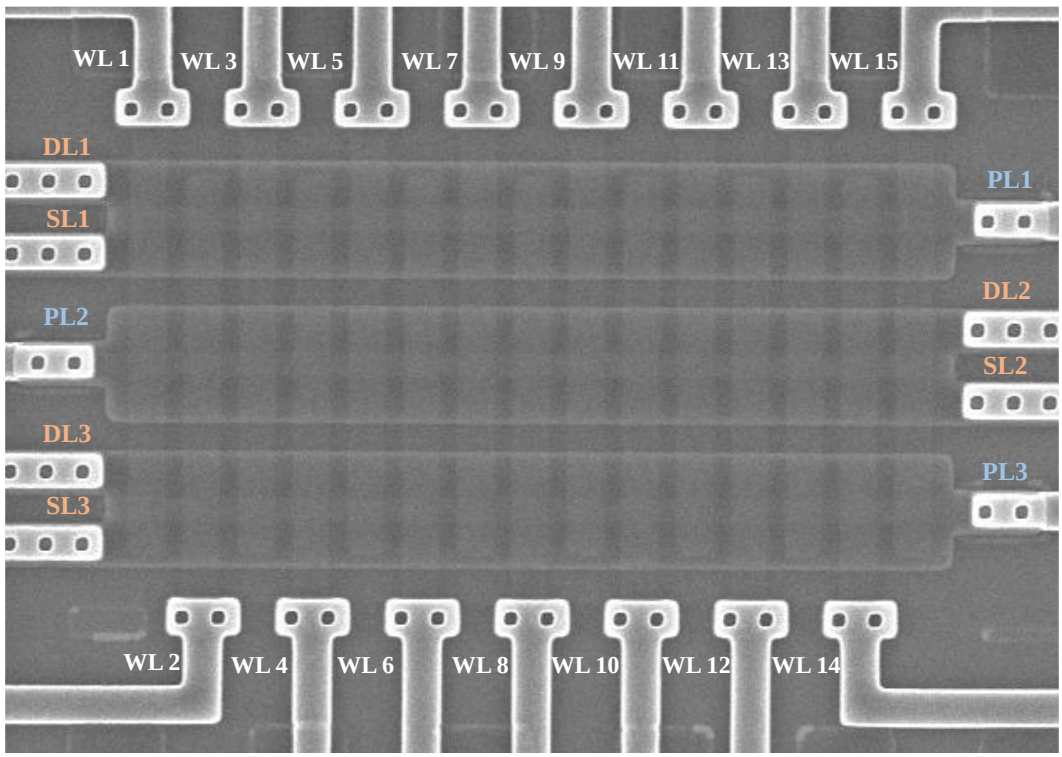
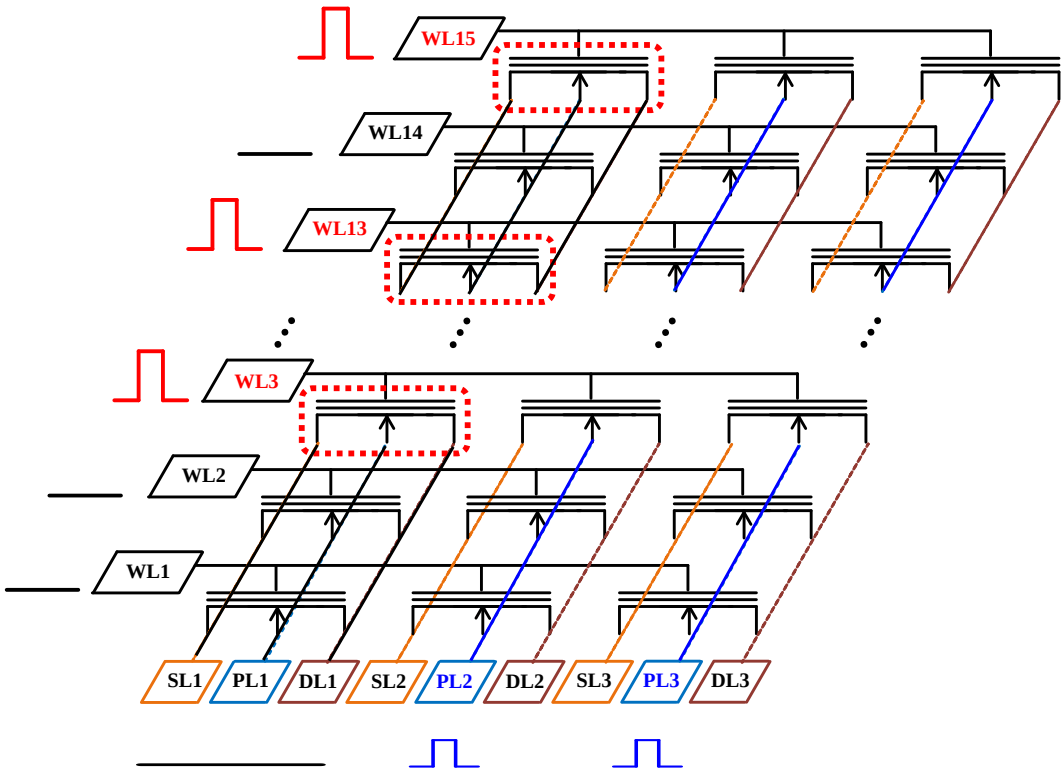


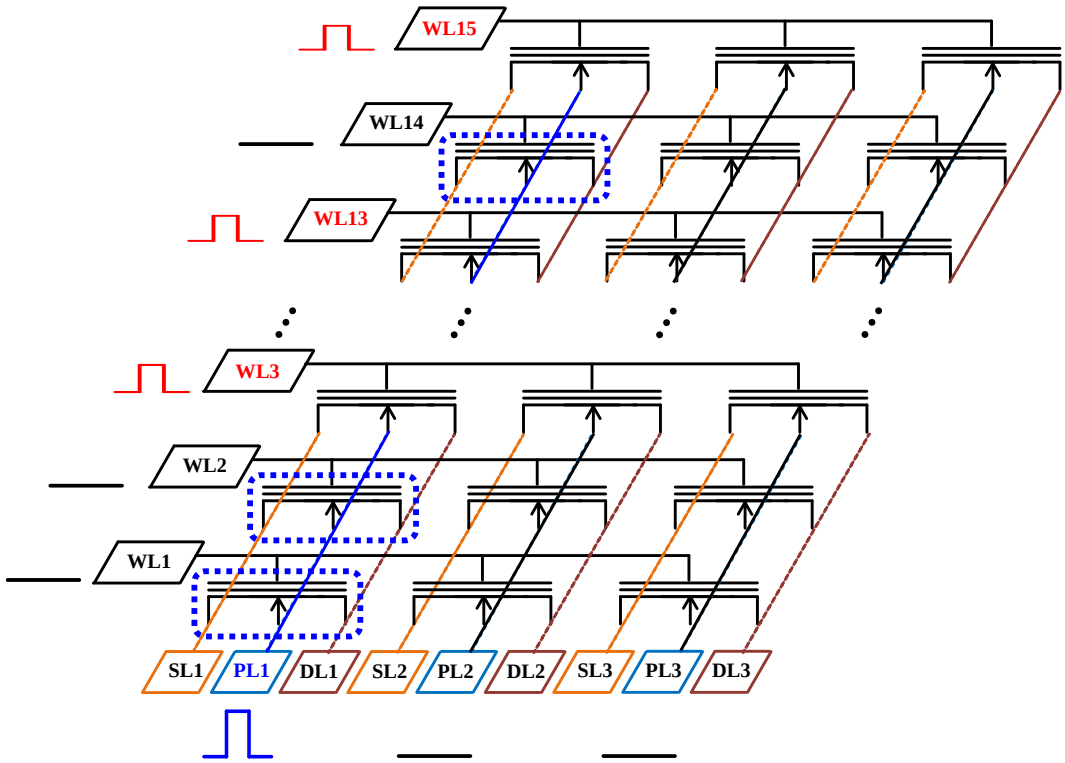
Fig. 5.9. A top SEM image of the fabricated 15 x 3 AND flash memory array through the integration fabrication of the synaptic array and CMOS circuit.



PGM : V_{WL} = V_{PGM}, V_{DL} = V_{SL} = V_{PL} = 0 V

Inhibition : V_{WL} = 0 V or V_{DL} = V_{SL} = floating, V_{PL} = V_{INH1}

Fig. 5.10. A bias condition for measurement in a selective PGM operation of specific cells in the 15 × 3 AND flash memory array.



ERS : V_{WL} **= 0 V,** V_{DL} **=** V_{SL} **= floating,** V_{PL} **=** V_{ERS}

Inhibition : V_{WL} **=** V_{INH2} **or** V_{DL} **=** V_{SL} **= floating,** V_{PL} **= 0 V**

Fig. 5.11. A bias condition for measurement in a selective ERS operation of specific cells in the 15 × 3 AND flash memory array.

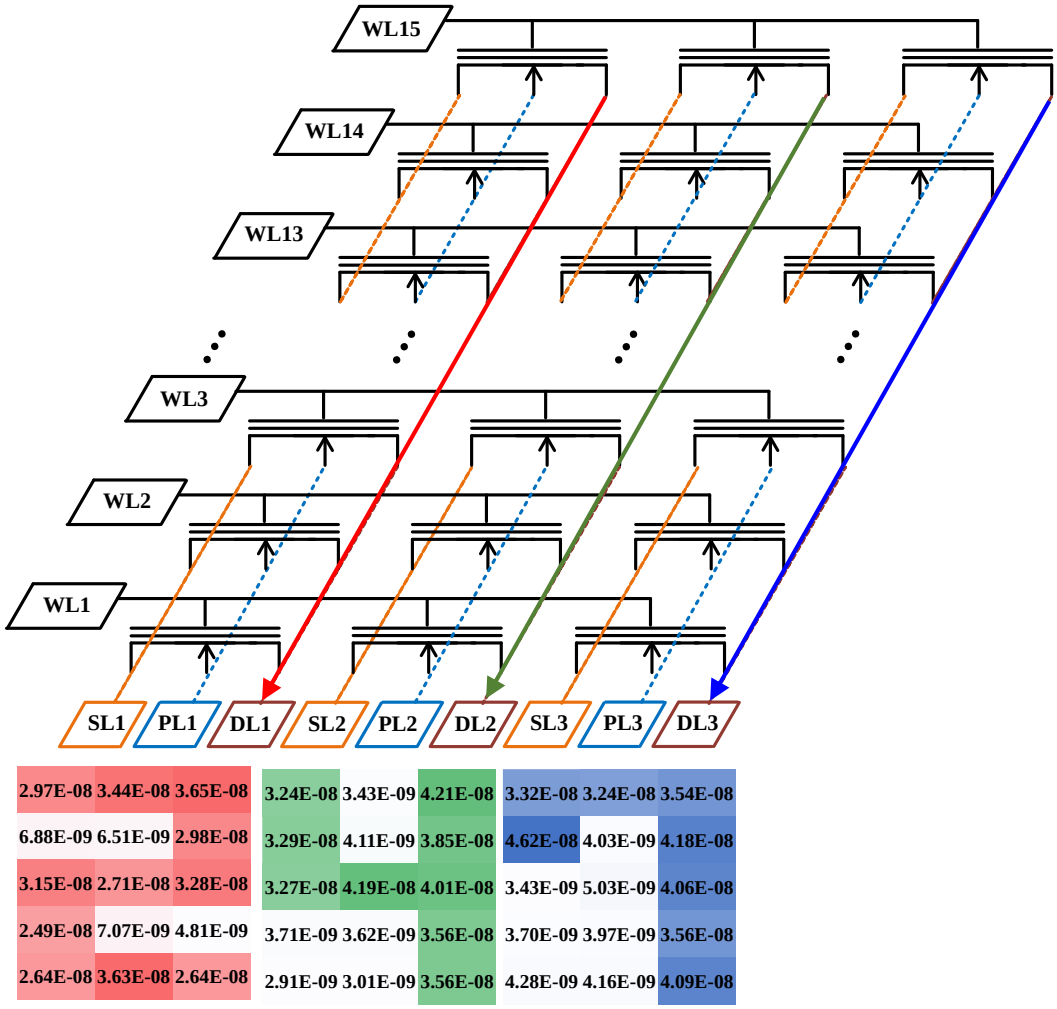


Fig. 5.12. Measurement results of learning specific number patterns in the fabricated flash memory array.

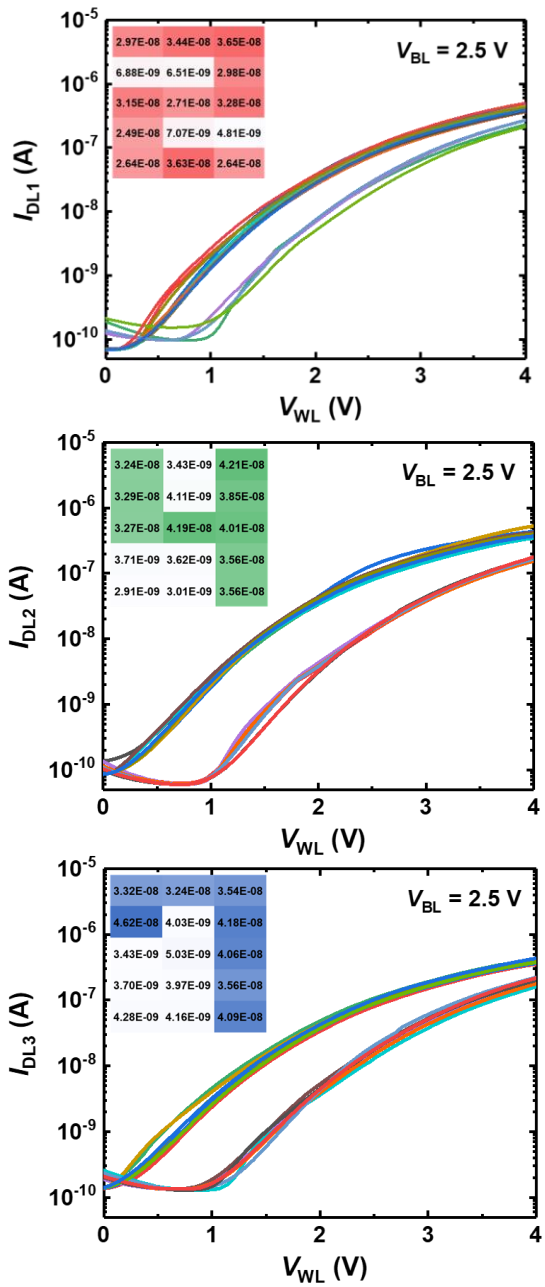


Fig. 5.13. Measured I_D – V_{GS} characteristics of each cell in the fabricated flash memory array after learning the specific number patterns.

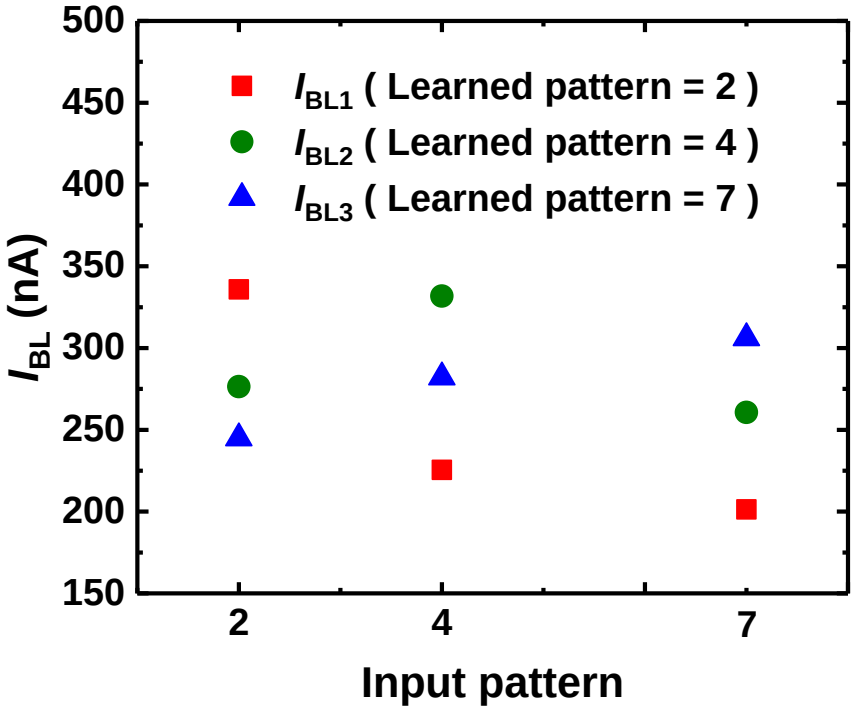


Fig. 5.14. The measurement result of each I_{BL} according to the input pattern in the fabricated flash memory array after learning the specific number patterns.

5.3 Measurement results of CMOS circuits

Fig. 5.15 shows a cross-sectional SEM image of a fabricated MOSFET through the integration fabrication of the synaptic array and CMOS circuit. Fig. 5.16 represents the measured I_D – V_{GS} characteristics as a parameter of V_{DS} of the fabricated CMOS without LDD implantation. In the *n*MOS, as shown in Fig. 5.16 (a), the curves show *n*-type behaviors with a *SS* of 90 mV/decade, a V_{th} of 0.52 V at V_{DS} = 1 V. In the *p*MOS, as shown in Fig. 5.16 (b), the curves show *p*-type behaviors with a *SS* of 87 mV/decade, a V_{th} of -0.87 V at V_{DS} = 1 V. The V_{th} is calculated by fitting through the constant current method at the point where I_D = 100 nA. In the *n*MOS without LDD implantation, the drain-induced barrier lowering (DIBL) phenomenon is almost non-existent. On the other hand, in the *p*MOS without LDD implantation, the value of DIBL is extracted to 67 mV/V. This is caused by the effect of shortening the channel length due to the diffusion of boron ion, which is used as the S/D of *p*MOS. Fig. 5.17 represents the measured I_D – V_{GS} characteristics as a parameter of V_{DS} of the fabricated CMOS with LDD implantation. In the *n*MOS, as shown in Fig. 5.17 (a), the curves show *n*-type

behaviors with a *SS* of 92 mV/decade, a V_{th} of 0.55 V at V_{DS} = 1 V. In the *p*MOS, as shown in Fig. 5.17 (b), the curves show *p*-type behaviors with a *SS* of 90 mV/decade, a V_{th} of -0.94 V at V_{DS} = 1 V. Compared to the case without LDD implantation, the value of DIBL in *p*MOS decreased to 44 mV/V. Fig. 5.18 (a) and (b) show output characteristics (I_D-V_{DS}) of *n*MOS and *p*MOS, respectively. Fig. 5.19 to 5.21 show the measurement results for various types of breakdown voltages (BVs) of the fabricated CMOS devices. These BVs provide important information for designing CMOS circuits using various voltages. The values of all BVs are extracted only for devices with both *W* and *L* of 0.5 μm among the fabricated CMOS devices. Fig. 5.19 represents the measured junction BV of the fabricated CMOS devices. The junction BV is extracted by applying a reverse bias between the body and the source or drain of the MOSFET, and it is measured to be around 20 V for both the *n*MOS and *p*MOS. Fig. 5.20 shows the measured gate oxide BV of the fabricated CMOS devices. The gate oxide BV can be extracted with two values according to the sign of V_{GB}, and it is measured by applying the voltage between the gate and the body of the MOSFET. In both *n*MOS and *p*MOS, when the value of V_{GB} is positive, the

gate oxide BV is measured to be ~30 V. On the other hand, when the value of V_{GB} is negative, it is measured to be ~15 V. Finally, Fig. 5.21 represents the measured drain to source breakdown voltage (BVDSS) of the fabricated CMOS devices. The electrical parameter BVDSS value is extracted by applying the bias between the source and drain while the gate is floating state, and it is measured to be around 10 V for both the *n*MOS and *p*MOS.

The *n*MOS and *p*MOS described above are connected in series to compose a CMOS logic inverter. Fig. 5.22 shows a SEM image of the fabricated CMOS logic inverter consisting of *n*MOS and *p*MOS through the integration fabrication of the synaptic array and CMOS circuit. The corresponding voltage transfer characteristics (VTC) of the fabricated CMOS logic inverter are shown in Fig. 5.23 (a) as a parameter of V_{DD}. The VTC displays good inverting performance with a full logic swing, abrupt transition, and almost zero current in the static condition. The logic transition occurs at a V_{IN} closer to zero than V_{DD} since the *n*MOS has a relatively lower V_{th} than the *p*MOS. The voltage gains of the fabricated CMOS logic inverter are higher than 50 for given V_{DD}s from 1 V to 3 V. The transition region

widths, which can be defined as the region where the gain is larger than 1, under different V_{DD}s are plotted in Fig. 5.23 (b). The transition regions for all measured V_{DD}s are less than 0.35 V. Another key component in influencing sensitivity and tolerance to signal interference in logic gates is the noise margin. Noise margin can be calculated by extracting ideal logic-high voltage V_{OH}, ideal logic-low voltage V_{OL}, maximum low input voltage in transition region V_{IL}, and minimum high input voltage in transition region V_{IH}. Therefore, noise margins for logic state 1 (NM_H) and logic state 0 (NM_L) are defined as $NM_H = V_{OH} - V_{IH}$ and $NM_L = V_{IL} - V_{OL}$. In this work, the normalized total noise margins to V_{DD}, which can be defined as $(NM_L+NM_H)/V_{DD}$, are calculated at different V_{DD}s.

Fig. 5.24 (a) and (b) represent a circuit diagram and a SEM image of the fabricated I&F circuit through the integration fabrication of the synaptic array and CMOS circuit, respectively. The C1 named the membrane capacitor (C_{mem}) plays a role in the integration of the current signal from the synapse array. The value of the C_{mem} can vary depending on the amount of current from the synapse array and the frequency of firing required by the system. The capacitance of C_{mem} determines the

L and W of M6 for resetting the neuron circuit. The operation scheme of the I&F circuit is as follows. When current signals from the synapse array are repeatedly transmitted to the neuron, the C_{mem} described above performs the integration function. And then, the membrane potential (V_{mem}) by an accumulation of charge in C_{mem} increases within the range of the threshold voltage (V_{th}) of M1. At the very moment V_{mem} becomes higher than V_{th} of M1, the M1 turns on and node 1 in the high state becomes low. Then, the output node of inverter 1 changes from the low state to the high state, and this initialized the state of V_{mem} using M6. In addition, the output node of inverter 2 operates M5 after the delay time by C2 and returns node 1 to its original state. M2, M3, and M4 exist to compensate for the stabilization of each node state in the neuron circuit. The C2 determines the width of a spike generated at the output node, which can be designed as a minimum value as long as a certain width is ensured. Fig. 5.25 and Fig. 5.26 represent the measurement results of the operation of the fabricated I&F circuit. The input of the fabricated I&F circuit is connected to a MOSFET, which can be assumed a virtual synaptic device, and the current signal of the MOSFET is transmitted to the I&F circuit through the

current mirror. As shown in Fig. 5.25, it can be seen that the frequency of firing in the I&F circuit increases as the voltage of the input pulse signal increases. Fig. 5.26 shows the firing frequency of the I&F circuit according to the period of the input pulse. For an input pulse with 10 μs, the I&F circuit fires once for every 5 inputs, whereas for an input pulse with 25 μs, it fires once for every 2 inputs.

The I&F circuit plays an important role in the main operation of the neuron circuit, but an additional circuit is required to convert the output signal of the I&F circuit into a pulse to be applied to the synaptic device. This is because the period and amplitude that the pulse as input of the synaptic device should have are different from those of the output pulse of the I&F circuit. For this reason, a pulse width extension circuit and a voltage shifter circuit responsible for changing the period and amplitude of the pulse, respectively, should be included in the neuron circuit. Fig. 5.27 (a) and (b) show a circuit diagram and a SEM image of the fabricated pulse width extension circuit through the integration fabrication of the synaptic array and CMOS circuit, respectively. The operation scheme of the pulse width extension circuit is as follows. In the stable state, the input is the low state so that

the output from the NOR gate is the high state. In this period, the M7 acting as a resistor is connected to the V_{DD}, and the input potential of the last inverter is equal to this voltage. As a result, the output from the last inverter will be in the low state, which means zero output. When an input pulse is triggered, the output of the NOR gate changes to the low state resulting in an output of the inverter equal to the high state. And then, the inverter maintains this unstable state until the capacitor charges up through the M7. When the M7 reaches the threshold voltage of the inverter, the output of the circuit changes to the low state. As a result, the width of the generated pulse is determined by the resistance of M7, and the values of C3. Here, the W and L of M7 are 1 μm and 5 μm, respectively. Fig. 5.28 represents the measurement result of the operation of the fabricated pulse width extension circuit. The inset figure shows the input pulse of the pulse width extension circuit with a period of 10 ns, and it is confirmed that the period of the output increases as the value of V_a, which determines the resistance value of M7, increases. In addition, the reproducibility of the circuit operation for repeated input pulse at a specific V_a value is also verfied. Fig. 5.29 (a) and (b) represent a circuit diagram and a SEM image

of the fabricated voltage level shifter circuit through the integration fabrication of the synaptic array and CMOS circuit, respectively. The operation scheme of the voltage level shifter circuit is as follows. When the input signal is in the low state, the M11 pulls down the output node to the ground. In the opposite case, when the input signal is in the high state, the M10 turns on, and the M9 pulls up the output node to V_{DD1}. The design consideration of the voltage level shifter circuit is to ensure that the W/L ratios of M8 and M9 are less than those of M10 and M11 to reduce the current spikes generated when the circuit changes state. The parameters used in the fabricated voltage level shifter circuit are as follows: $W_{M8,\,M9}$ = 1 μm, $L_{M8,\,M9}$ = 10 μm, $W_{M10,\,M11}$ = 10 μm, $L_{M10,\,M11}$ = 1 μm. Fig. 5.30 shows the measurement result of the operation of the fabricated voltage level shifter circuit. From the result, it can be seen that the input pulse with an amplitude of 1.5 V is changed into the output pulse of various amplitudes according to the V_{DD1}.

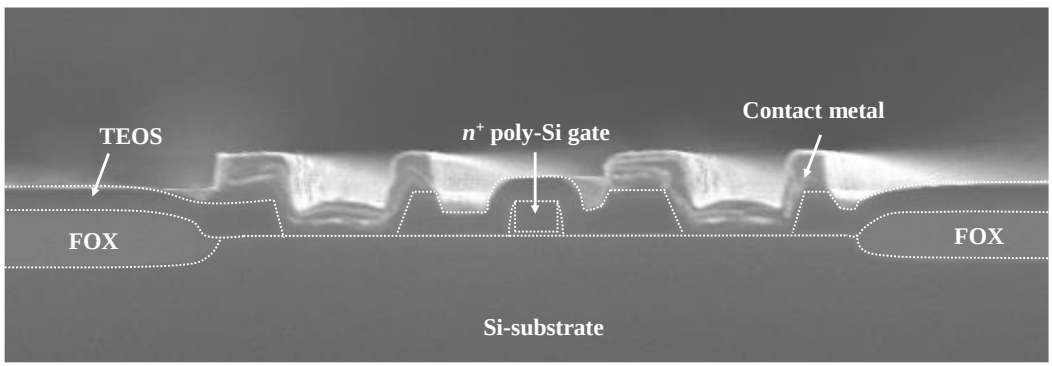


Fig. 5.15. A cross-sectional SEM image of a fabricated MOSFET through the integration fabrication of the synaptic array and CMOS circuit.

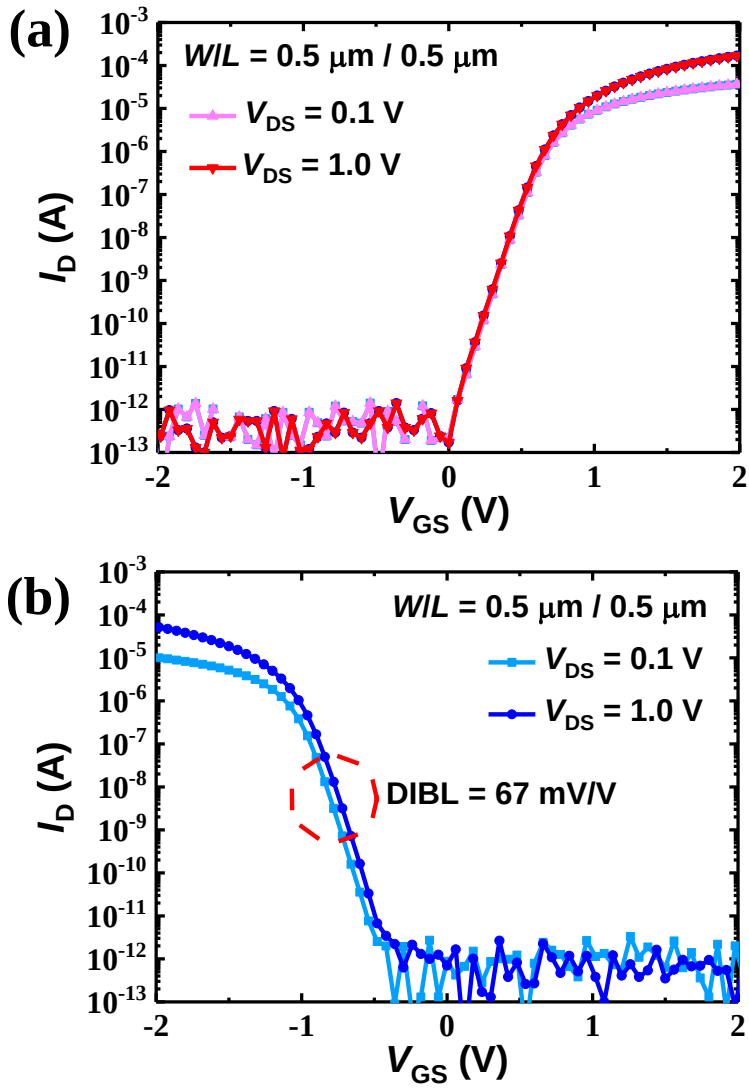


Fig. 5.16. Measured I_D – V_{GS} characteristics as a parameter of V_{DS} of the fabricated (a) *n*MOS and (b) *p*MOS without LDD implantation through the integration fabrication of the synaptic array and CMOS circuit.

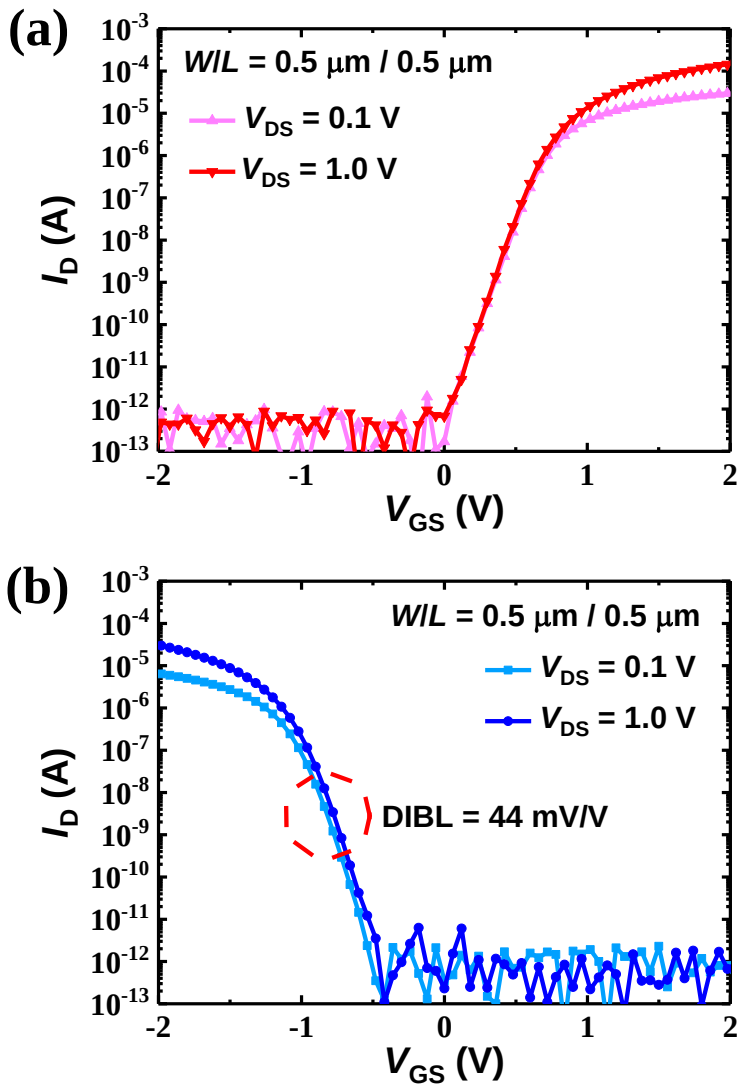


Fig. 5.17. Measured I_D – V_{GS} characteristics as a parameter of V_{DS} of the fabricated (a) *n*MOS and (b) *p*MOS with LDD implantation through the integration fabrication of the synaptic array and CMOS circuit.

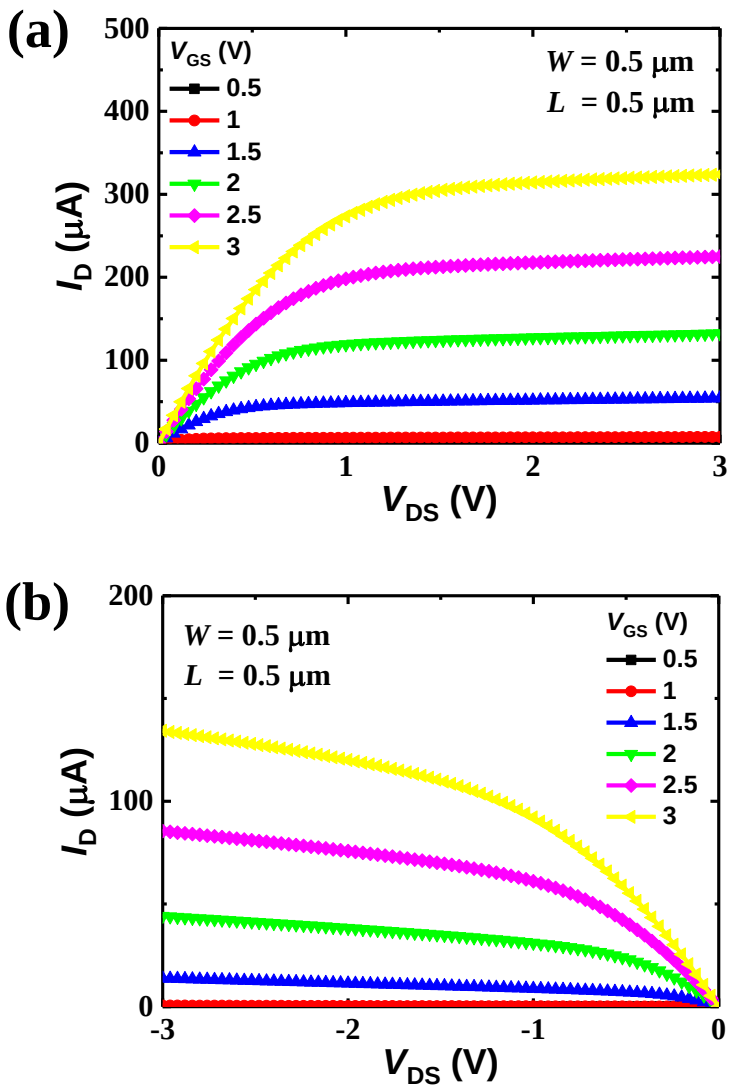


Fig. 5.18. Measured I_D – V_{DS} characteristics as a parameter of V_{GS} of the fabricated (a) *n*MOS and (b) *p*MOS with LDD implantation through the integration fabrication of the synaptic array and CMOS circuit.

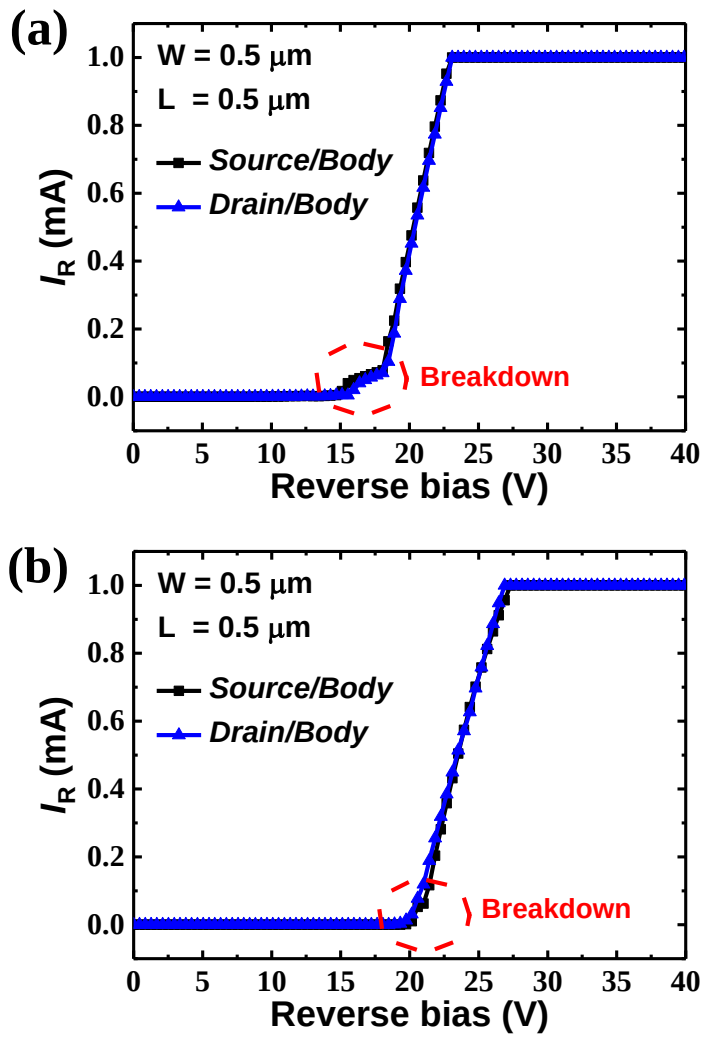


Fig. 5.19. Measured junction BV of the fabricated (a) *n*MOS and (b) *p*MOS through the integration fabrication of the synaptic array and CMOS circuit.

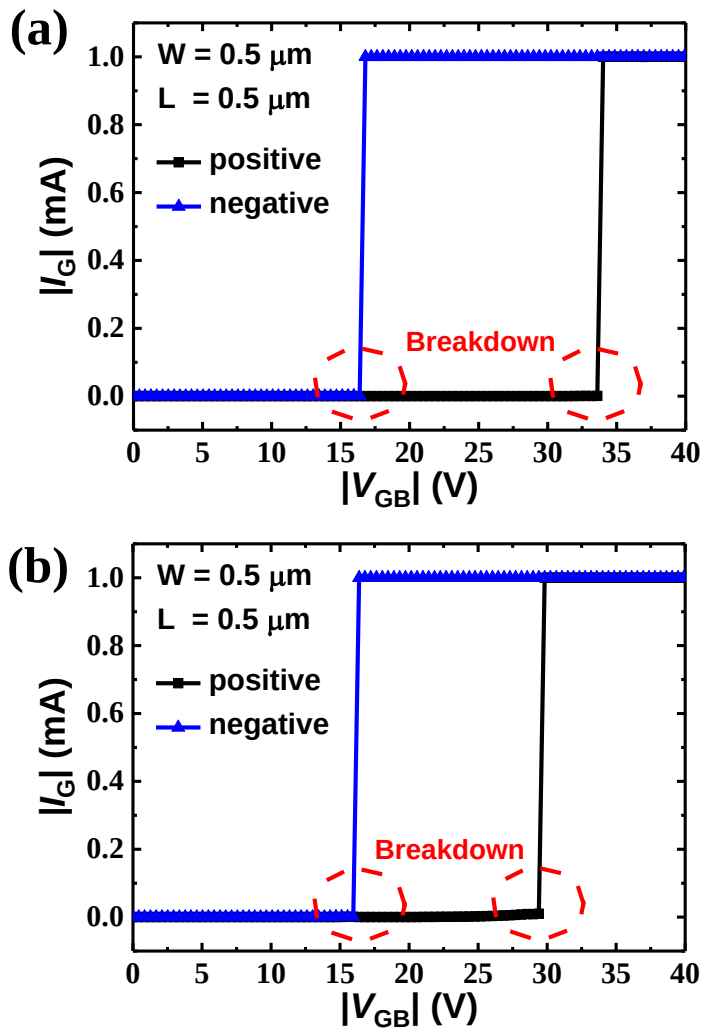


Fig. 5.20. Measured gate oxide BV of the fabricated (a) *n*MOS and (b) *p*MOS through the integration fabrication of the synaptic array and CMOS circuit.

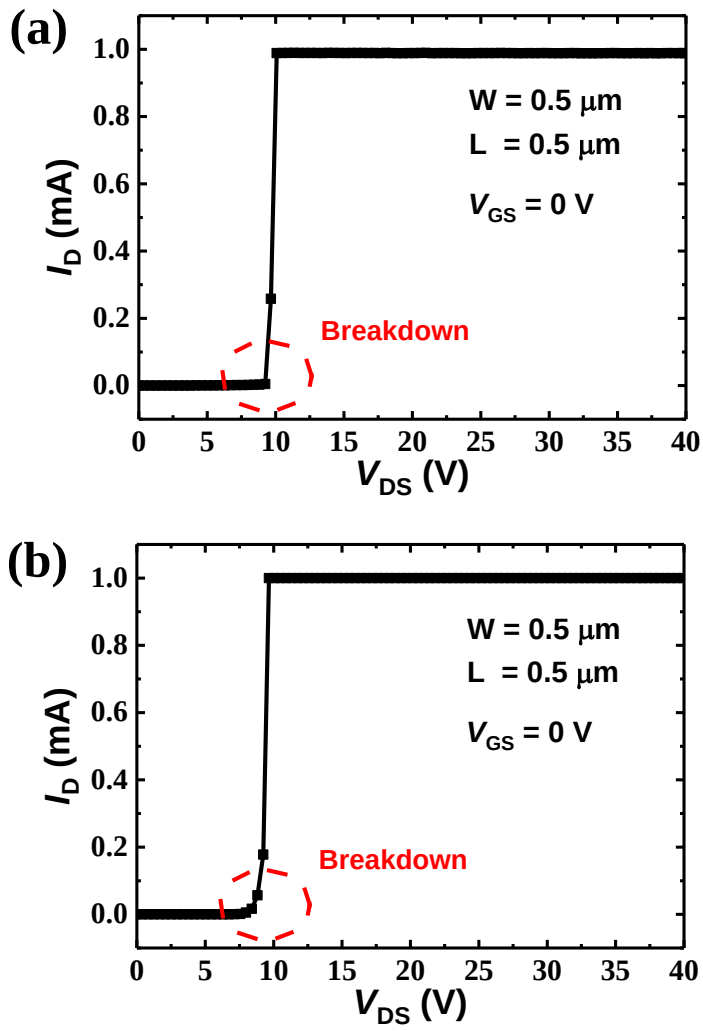


Fig. 5.21. Measured BVDSS of the fabricated (a) *n*MOS and (b) *p*MOS through the integration fabrication of the synaptic array and CMOS circuit.

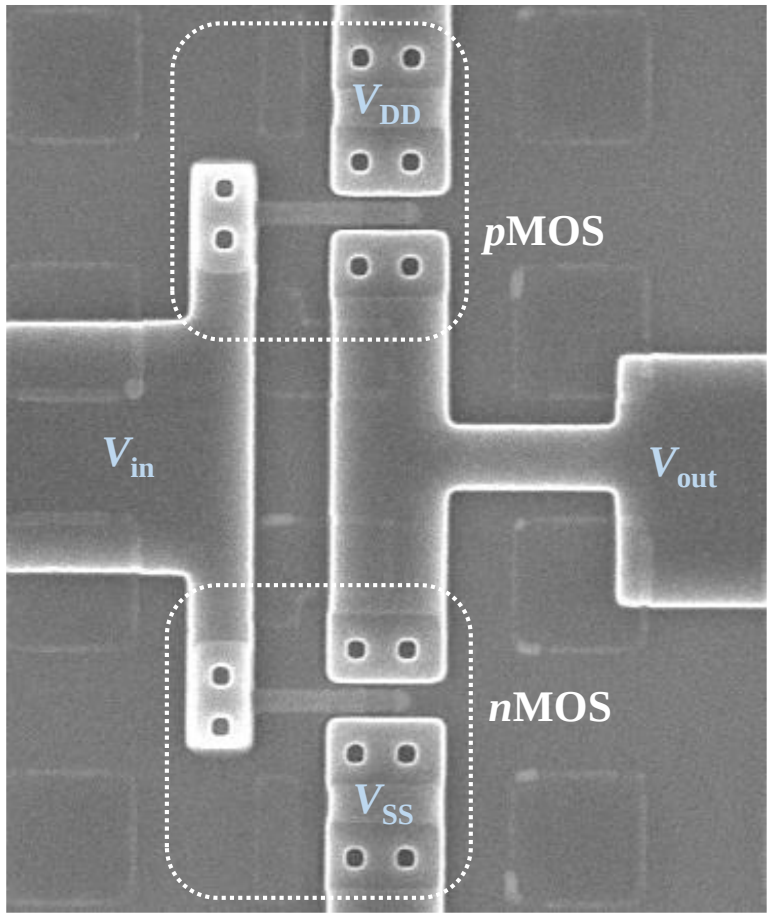


Fig. 5.22. A top SEM image of the fabricated CMOS logic inverter consisted of *n*MOS and *p*MOS through the integration fabrication of the synaptic array and CMOS circuit.

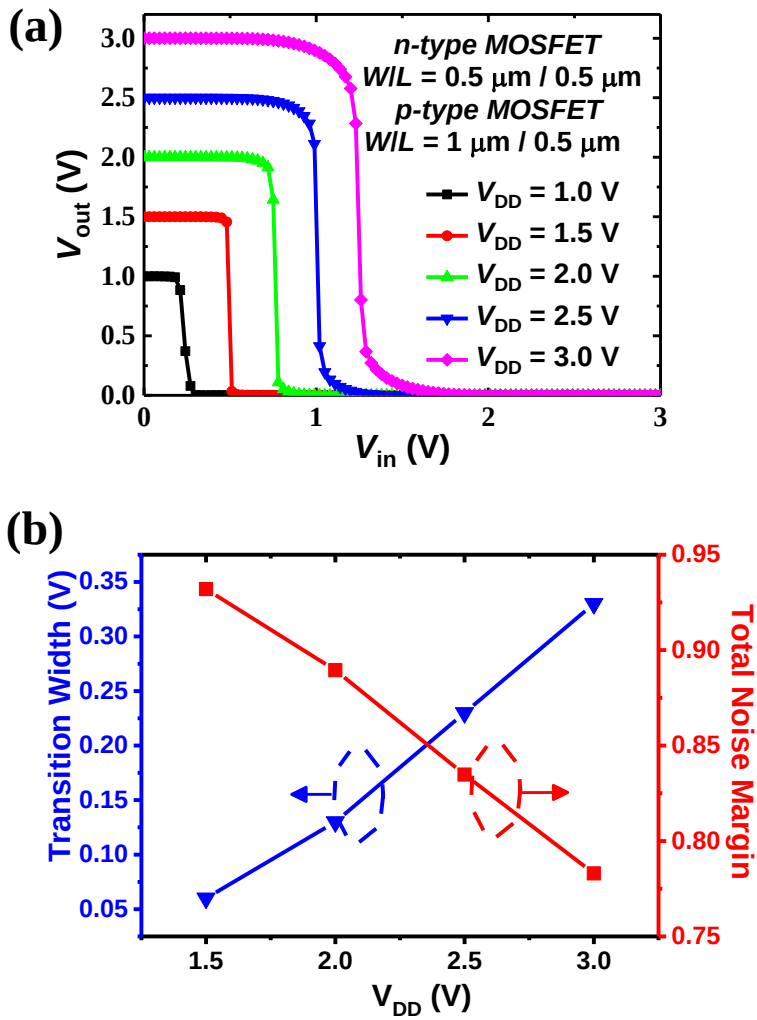


Fig. 5.23. (a) Voltage transfer characteristics (output voltage as a function of the input voltage) of a CMOS logic inverter consisted of *n*MOS and *p*MOS through the integration fabrication of the synaptic array and CMOS circuit. (b) Transition width and the normalized total noise margin to V_{DD} are shown on the left and right y-axes, respectively, as a function of V_{DD}.

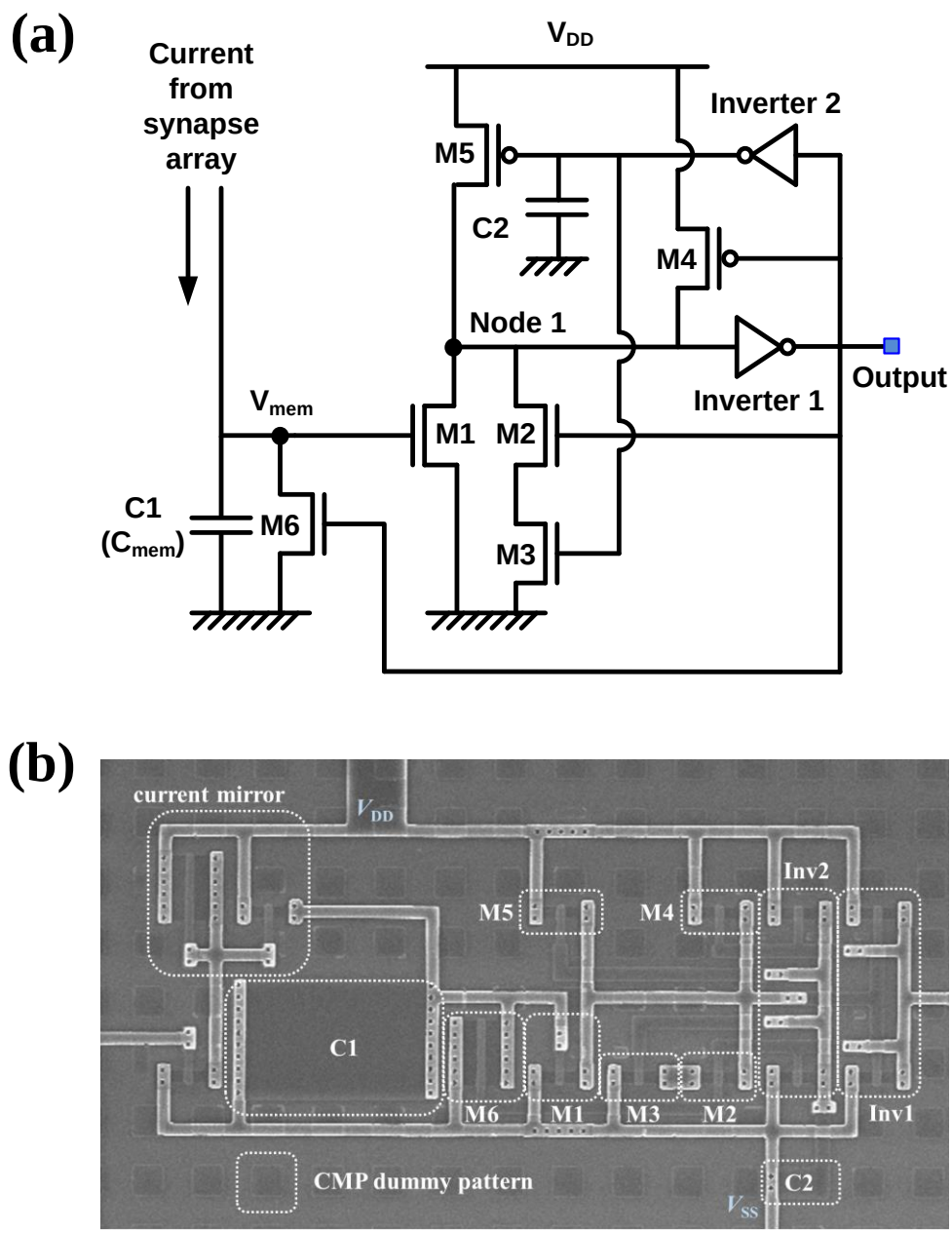


Fig. 5.24. (a) A circuit diagram and (b) a top SEM image of the fabricated I&F circuit through the integration fabrication of the synaptic array and CMOS circuit.

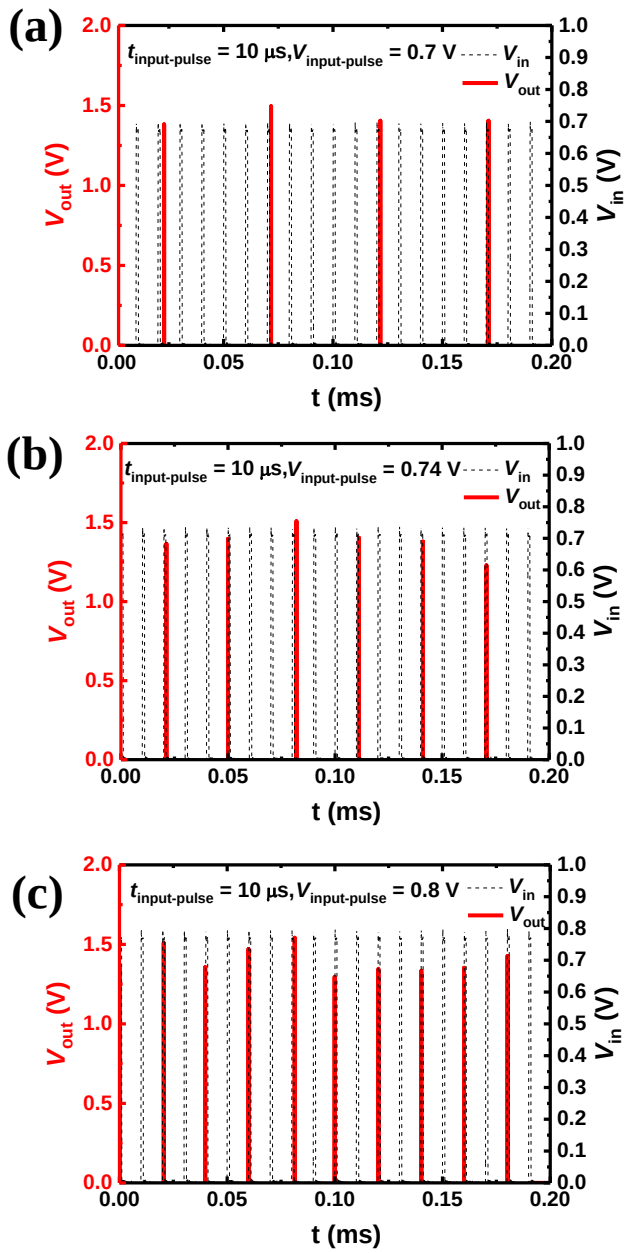


Fig. 5.25. The measurement result of the operation of the fabricated I&F circuit according to the amplitude of the input pulse.

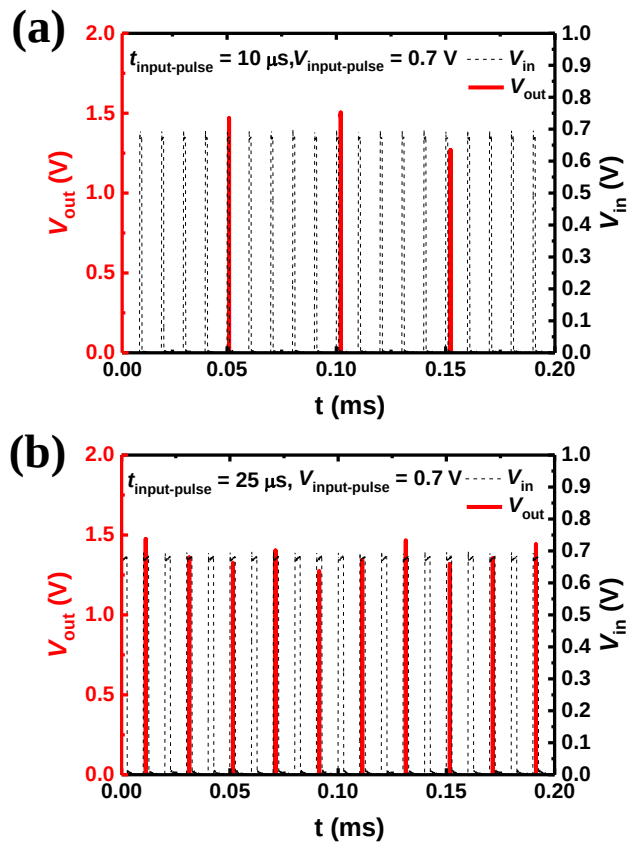


Fig. 5.26. The measurement result of the operation of the fabricated I&F circuit according to the width of the input pulse.

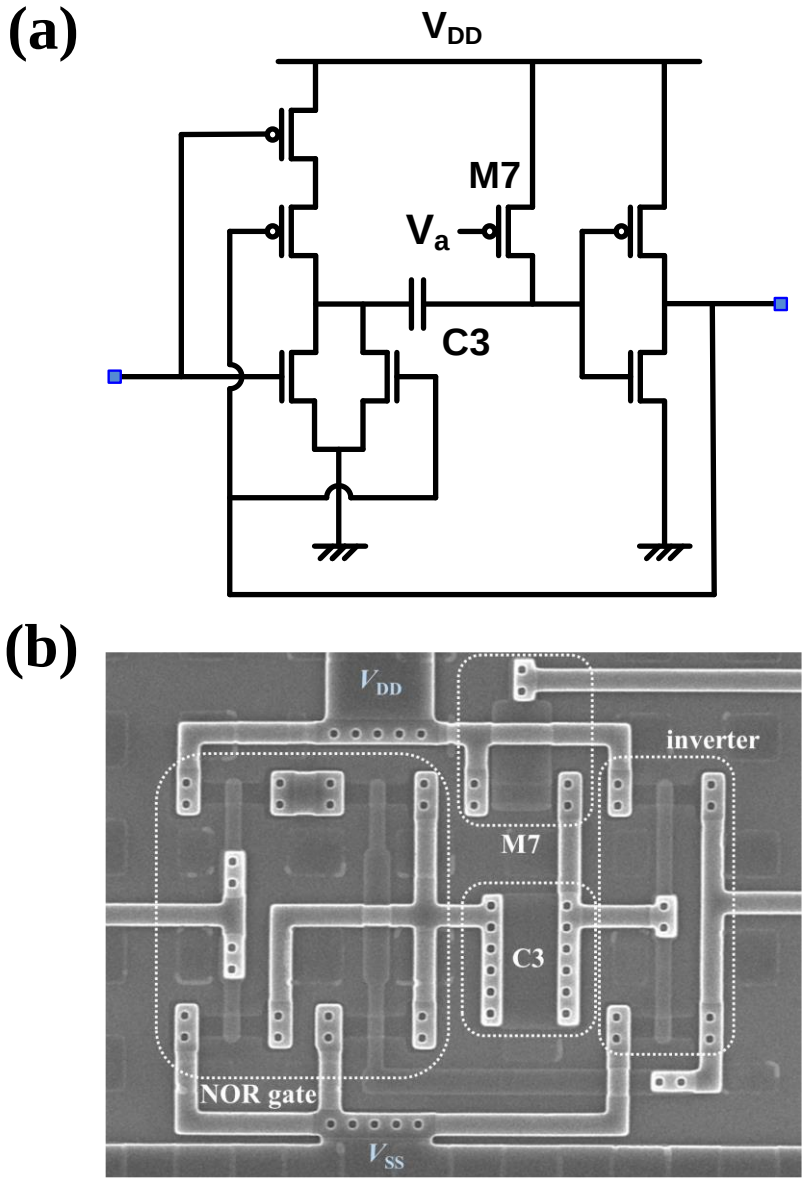


Fig. 5.27. (a) A circuit diagram and (b) a top SEM image of the fabricated pulse width extension circuit through the integration fabrication of the synaptic array and CMOS circuit.

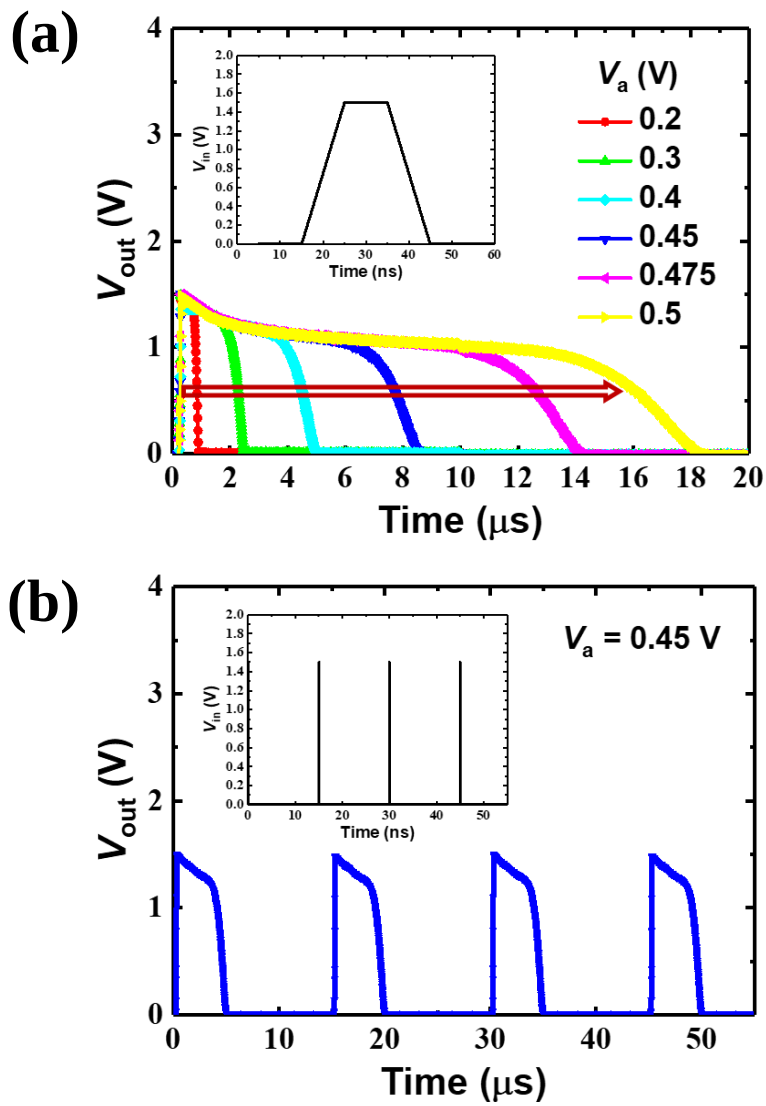


Fig. 5.28. The measurement result of the operation of the fabricated pulse width extension circuit.

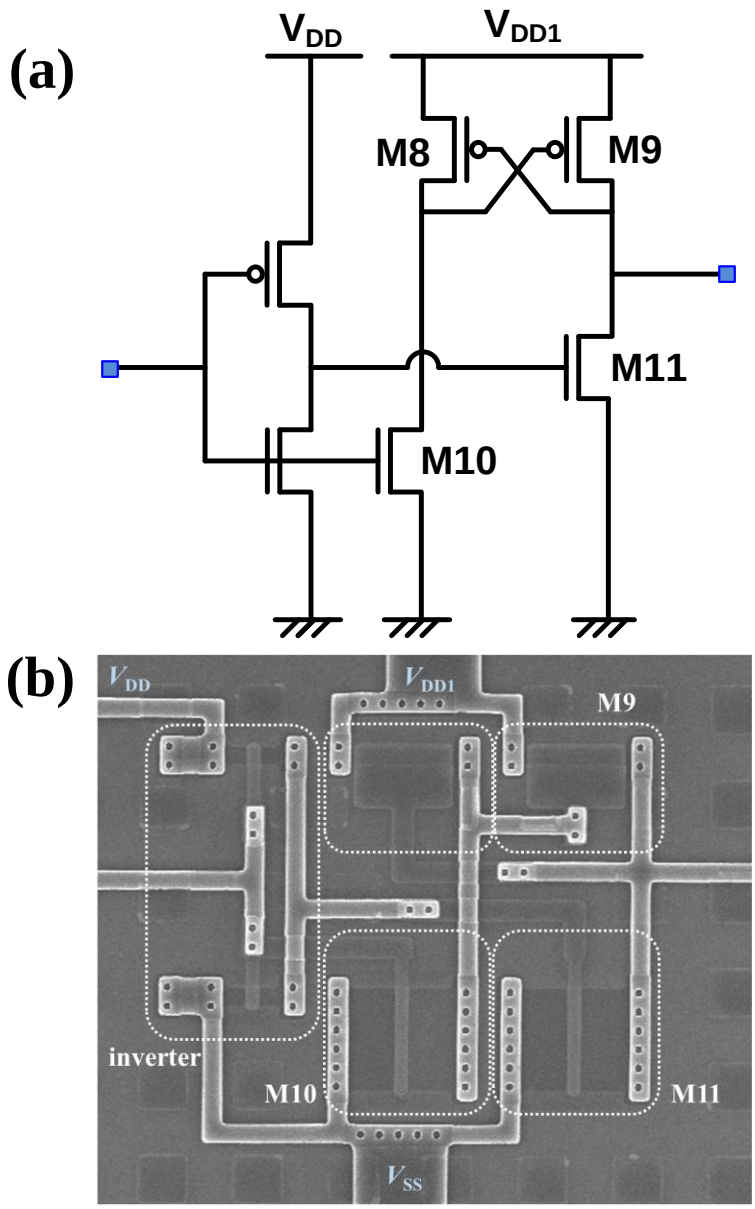


Fig. 5.29. (a) A circuit diagram and (b) a top SEM image of the fabricated voltage level shifter circuit through the integration fabrication of the synaptic array and CMOS circuit.

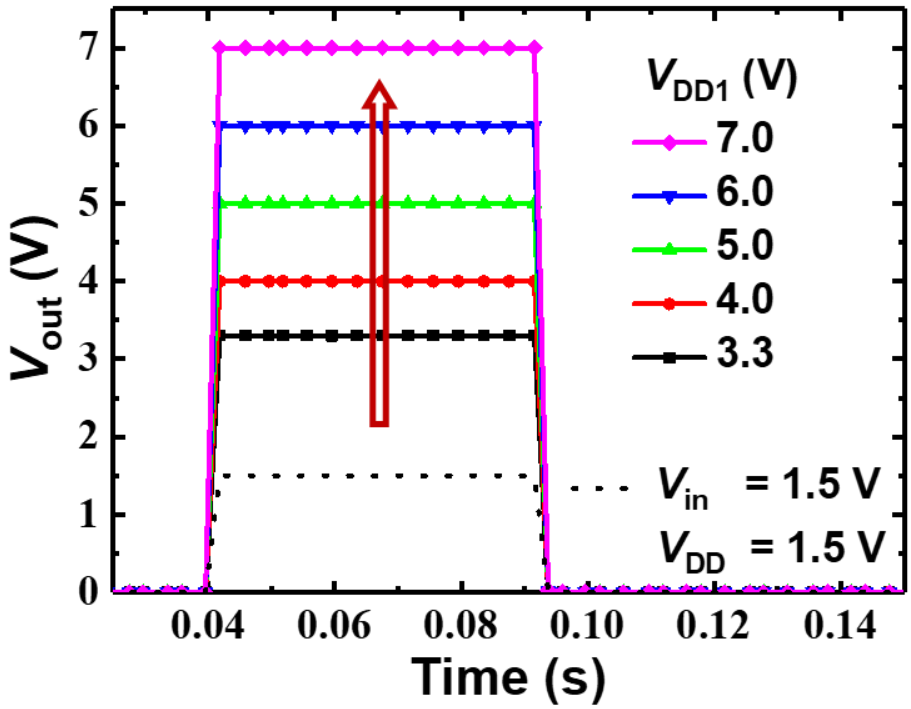


Fig. 5.30. The measurement result of the operation of the fabricated voltage level shifter circuit.

Chapter 6
Conclusion

In this work, we have investigated hardware-based SNNs based on the AND flash memory synaptic array. Since flash memory technology has already been proven in terms of nonvolatile characteristics, reliability, and mass production, it can be said to have a suitable aspect to be used as the electronic synaptic device. The proposed TFT-type synaptic device has a layer of poly-Si as a channel material and a high-κ material included in the gate insulator stack. The structure of the synaptic device with a boron-doped body reduces the circuit burden for the memory operation, and the partially curved channel can maximize the merit of the memory characteristics and output impedance when the device is scaled down. An effective pulse scheme for selective memory operation in the fabricated AND array was proposed and verified experimentally. The AND array can have the advantage in on-chip learning since the power consumption is determined by a relatively low tunneling current flowing through the device when updating synaptic weights. For

unsupervised learning, we have proposed a SNN that supports the STDP pulse scheme available for AND type array. The designed computing architecture does not generate pulses from external circuits but makes the necessary pulses using output signals from the neuron circuits in the architecture. System-level simulation based on pulse scheme and proposed SNN architecture was demonstrated, and an accuracy of 91.63% was obtained for the MNIST dataset. Although the proposed pulse scheme utilizes the inhibition pulses for selective memory operation in the synaptic array, it is more efficient in terms of system area and energy consumption of the computing architecture than the pulse schemes that use overlapping pulses. For supervised learning, a FC three-layer SNN for utilizing the DFA algorithm was designed based on the characteristics of the synaptic array and pulse scheme. Due to the asymmetry of the synaptic array in the DFA, it was designed by adding relatively small external arrays compared to the overall area of the SNN. An efficient pulse scheme for on-chip training was also proposed, and a system-level simulation was performed. The simulation result obtained from the SNN consisting of 28 × 28 input neurons, 256 hidden neurons, and 10 output neurons shows

superior accuracy (97.01%). This is a recognition rate comparable to that of a software-based network and shows that DFA can be used for on-chip learning in hardware SNNs employing AND flash memory arrays. To implement the designed hardware-based neural networks, not only the synaptic array but also the peripheral additional circuits should be supported. In addition, to target a neuromorphic chip for implementing a neural network, an efficient integration fabrication of a synaptic array and a CMOS circuit is required. For this reason, we have proposed and verified the fabrication method in which the proposed synaptic array and CMOS circuit can be efficiently integrated. The proposed fabrication method integrates synaptic arrays and CMOS circuits on the same wafer and has the advantage of reducing the number of masks and process steps. In the fabricated synaptic array, selective memory operation was verified by learning the specific number pattern. In addition, circuits such as CMOS inverters, I&F circuit, pulse width extension circuit, and voltage level shifter that can be included in the neuron circuit have been verified. The designed fabrication method is significant in that it presents a methodology that can efficiently implement a hardware-based neural network.

Appendix A

Neuron circuits to implement a hardware-based neural network using the STDP learning algorithm and the pulse scheme not including the inhibition pulses

Fig. A.1 (a) and (b) represent conceptual diagrams of blocks including signal flows in the input neuron and the global pulse generator I, respectively, to implement the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme not including the inhibition pulses. To generate the input pulse of the form described in Fig. 3.3, we design the circuit block that does not use a negative power source. In the input neuron, three pulses with different amplitudes and widths are used to generate the input pulse with a 3-level amplitude. Since two of the three pulses have the same width, two pulse extension modules are included in the input neuron to convert a short spike received from the input detector module into pulses with two different widths. And then, the pulses extended in the

pulse extension module are assigned to the pulse selection module to make each pulse with a specific amplitude. Since the pulse selection module generates the desired pulses depending on whether or not the signal is received from the pulse extension module, the input pulse generator module generates the input pulse only when the input neuron receives the input signal. At this time, the pulses required for the pulse selection module are provided by the global pulse generator I. While, among the pulses generated by the global pulse generator I, the pulse that does not participate in the input pulse generator module is involved in maintaining the input pulse with a positive voltage. In addition, switches should also be included in the input neuron so that the pulse with a positive voltage and the pulse from the input generator module are not affected by each other. The control pulses for these switches are also given from the global pulse generator I. In Fig. A.1 (b), the pulse entering the output neurons allows the output neuron to receive current only during the read operation time of in the electronic synaptic array, thereby preventing the system error due to the leakage current. Note that the dotted arrows in the figure indicate two or more signals. Fig. A.2 shows the input neuron circuit for generating

the input pulse used in the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme not including the inhibition pulses. First, the pulse width generated by the pulse extension module including A.M1 and A.C1 is determined by the resistance of A.M1 and the capacitance of A.C1. The same principle applies to pulse extension modules that include A.M2 and A.C2. The operation of the pulse selection module can be explained by the potential of node 1 (A.N1). In the absence of the input signal, the potential of A.N1 remains high and the front node of A.C3 remains low because the output terminals of the pulse extension module and global pulse generator remain at 0 V. However, when the input detector module detects the input signal, the potential of A.N1 can be changed depending on whether the input signal is received in the input neuron. At this time, if there is the input signal in the input neuron, the potential of A.N1 has a low state because the output terminal of the pulse extension module and the global pulse generator I change to the high state. This means that the output terminal of the pulse selection module has a high state. On the other hand, when there is no input signal in the input neuron, the output terminal of the pulse extension module is kept low

state, and the output terminal of the global pulse generator is kept high so that A.N1 remains in the high state. The input pulse generator module consists of two capacitors (A.C3 and A.C4) and two MOSFETs (A.M3 and A.M4). The A.C3 and A.M3 are responsible for generating negative voltage through the charging and discharging of the capacitor, and A.C4 and A.M4 determine the length of time that the negative voltage value is maintained. The A.S1 and A.S2 switches help ensure that the positive and negative voltage pulses are not affected by each other when they are applied to the electronic synaptic array, respectively. Fig. A.3 (a)-(c) show examples of input signals applying to several input neurons included in the input neuron layer. Fig. A.3 (d) represents input signals of the global pulse generator I when input signals are recognized by the input detector module. Fig. A.3 (e)-(g) show input pulses generated by the input neuron circuit in each representative neuron. In each represenatative neuron, the desired form of input pulses for the pulse scheme are generated at the time the input signal comes in the input neuron. Fig. A.3 (h) shows an enlarged view of the part of the input pulse that is involved in the read operation. The read operation time depends on the conductance range of

the electronic synaptic device and the component that accepts the current in the output neuron. If the read operation time is too long, it can be disadvantageous in terms of the energy consumption of the system. If it is too short, the ability to distinguish neurons can be lost. For this reason, it is necessary to determine the read operation time by considering the characteristics of the electronic synaptic device, the number of the synaptic devices in the array, and the potential condition of the output neuron determined by the current flowing through the electronic synaptic array.

Fig. A.4 (a) and (b) represent conceptual diagrams of blocks including signal flows in the output neuron and the global pulse generator II, respectively, to implement the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme not including the inhibition pulses. Since the form of the feedback pulse to be generated from the output neuron is similar to that of the input pulse, the components of the output neuron and global pulse generator II are almost the same as those of the input neuron and global pulse generator I. However, the output neuron includes an integrate-and-fire module that accepts

current from the electronic synaptic array as input and determines the timing of firing depending on the potential of a specific component in the output neuron. Fig. A.5 shows the output neuron circuit for generating the input pulse used in the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme not including the inhibition pulses. In the integrate-and-fire module, the A.C9 named the membrane capacitor serves to integrate the current from the electronic synaptic array. The capacitance value of A.C9 can be determined by the amount of current received from the electronic synaptic array during a particular period, and the frequency of output neuron's firing required for the computing architecture. Furthermore, the capacitance value of A.C9 determines the specification of A.M10, which resets the membrane potential (V_{mem}) of the output neuron. The operation of the output neuron circuit is described as follows. The V_{mem} due to charge accumulation in C_{mem} can increase within the range of the V_{th} of A.M9, connected to one terminal of C_{mem}. If V_{mem} becomes higher than V_{th} of A.M9, the A.M9 turns on and the state of node 2 (A.N2) in the high becomes low. Then, the output node of the integrate-and-fire module changes from the low state to the high

state by the inverter, and A.M10 can initialize the state of the V_{mem}. Next, after a certain delay time, A.N2 returns to its original state by A.M11. A.M12-14 compensate for the stabilization of each node state in the integrate-and-fire module circuit. A.C10 determines the width of the output spike of the integrate-and-fire module and can be designed to the minimum as long as operation in the connected circuit is guaranteed. The output spike from the integrate-and-fire module is converted into the feedback pulses with positive and negative voltages values through circuits similar to the input neuron. However, in the output neuron, the specifications of the pulse extension modules differ from those of the input neuron. This is because the input pulse and the feedback pulse have different times of maintaining the positive and negative voltages. The switches A.S3 and A.S4 are connected to the DL of the electronic synaptic array. They allow the current flowing through the electronic synaptic devices to affect the V_{mem} of the output neuron only when there is the input signal. In addition, they allow the potential of DL in the electronic synaptic array to be properly determined by the feedback pulse from the output neuron during the memory operation. Similarly, A.S5-7 switches serve to

determine the potential of the PL and SL in the electronic synaptic array. The PL and SL of the electronic synaptic array should maintain 0 V during the read operation time and have the same potential as the feedback pulse when the feedback pulse is generated. Fig. A.6 shows the behavior of each representative neuron circuit in the single output neuron layer. Fig. A.6 (a)-(c) represent the V_{mem} changes in each output neuron, and Fig. A.6 (d)-(f) shows the firing spikes from the integrate-and-fire module caused by the V_{mem} change in each output neuron. As shown in Fig. A.6 (a)-(f), when the V_{mem} of a certain output neuron exceeds the V_{th} of M9, it can be seen that the integrate-and-fire module in the corresponding neuron generates a firing spike. In addition, it can be confirmed that all neurons have a refractory period during a time when the feedback pulse is generated due to the firing of the output neuron. Fig. A.7 (a)-(c), (d)-(f), and (g)-(i) show the potential of the PLs, the SLs, and the DLs, respectively, of the electronic synaptic devices in the AND-type array connected to representative output neurons in the single output neuron layer. Simulation results for the potential of the PL, the SL, and the DL of the electronic synaptic devices due to the feedback pulse show that the pulse scheme described in

Fig. 3.3 is effectively implemented by the designed output neuron circuit. Fig. A.7 (j) and (k) show an enlarged view of the tail portion of the feedback pulse and the drain potential in the read operation, respectively. As shown in Fig. A.7 (k), the drain potential of the electronic synaptic device has the required read voltage only during the read operation time.

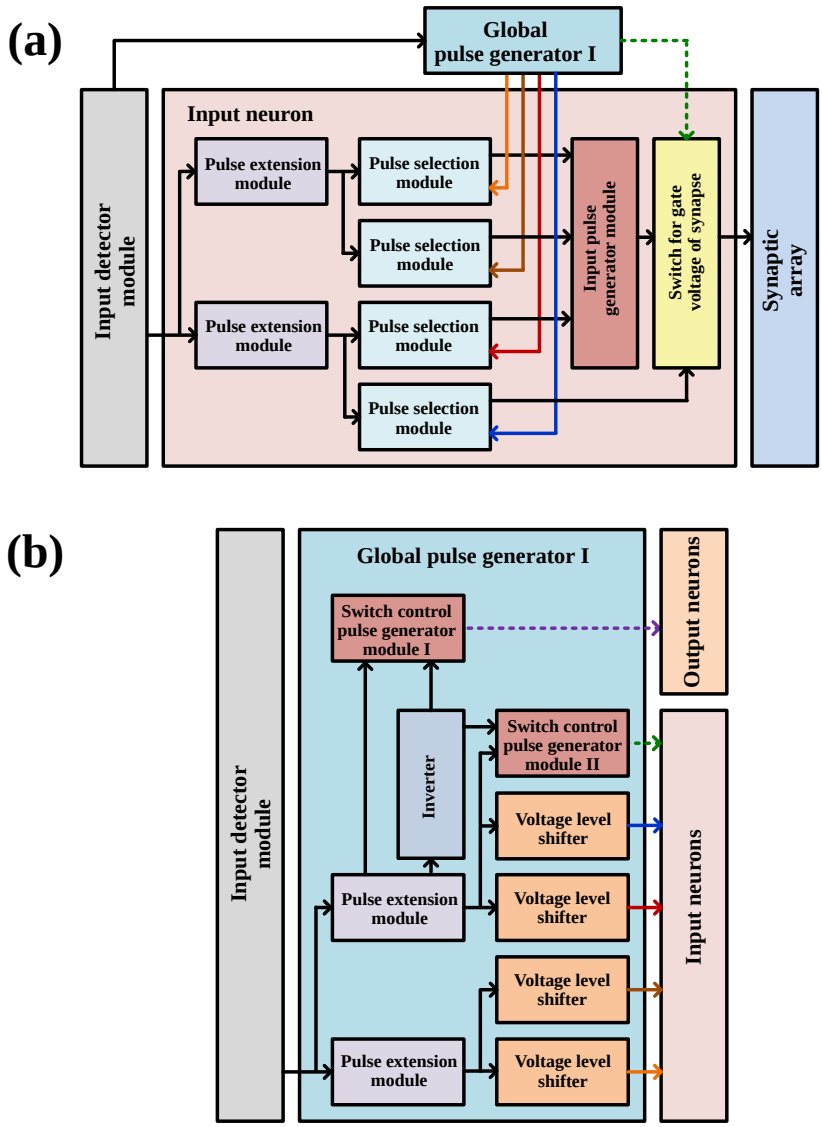


Fig. A.1. Conceptual diagrams of blocks including signal flows in the (a) input neuron and the (b) global pulse generator I to implement the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme not including the inhibition pulses.

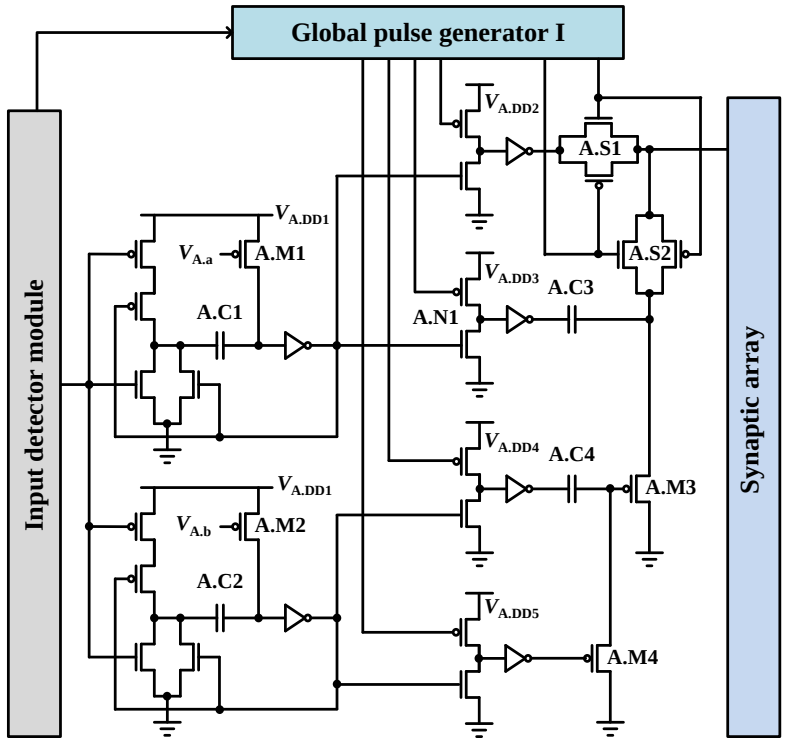


Fig. A.2. An input neuron circuit for generating the input pulse used in the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme not including the inhibition pulses.

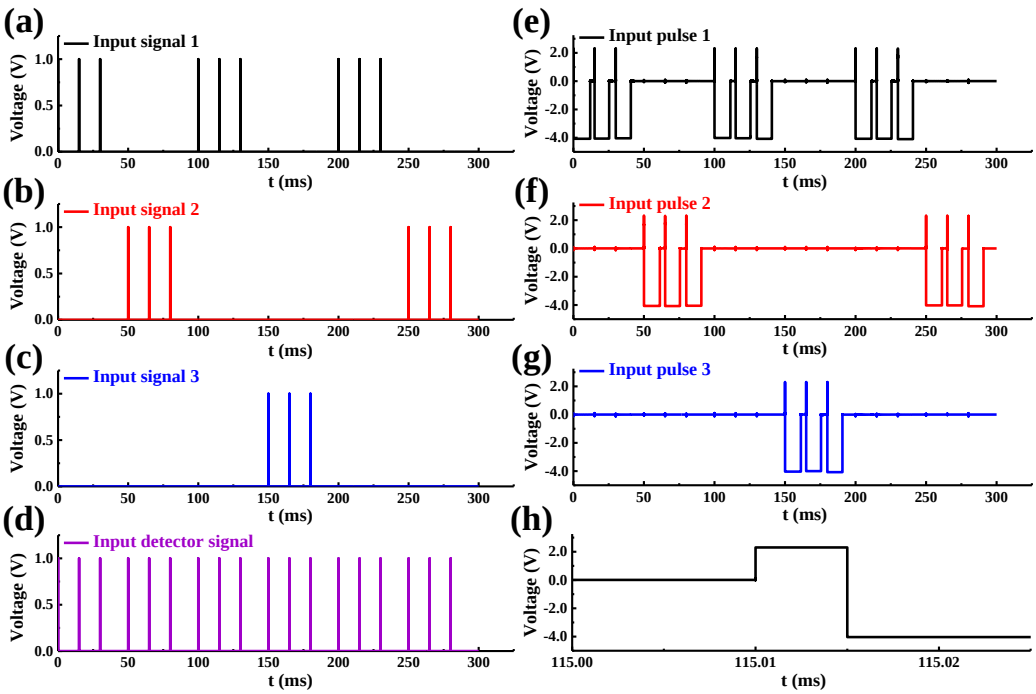


Fig. A.3. Operation of input neuron circuits to implement the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme not including the inhibition pulses. (a)-(c) Input signals received by representative input neurons included in the single input neuron layer. (d) Input signals of the global pulse generator I when input signals are recognized by the input detector module. (e)-(g) Input pulses generated by the input neuron circuit in each representative neuron. (h) An enlarged view of the part of the input pulse that is involved in the read operation.

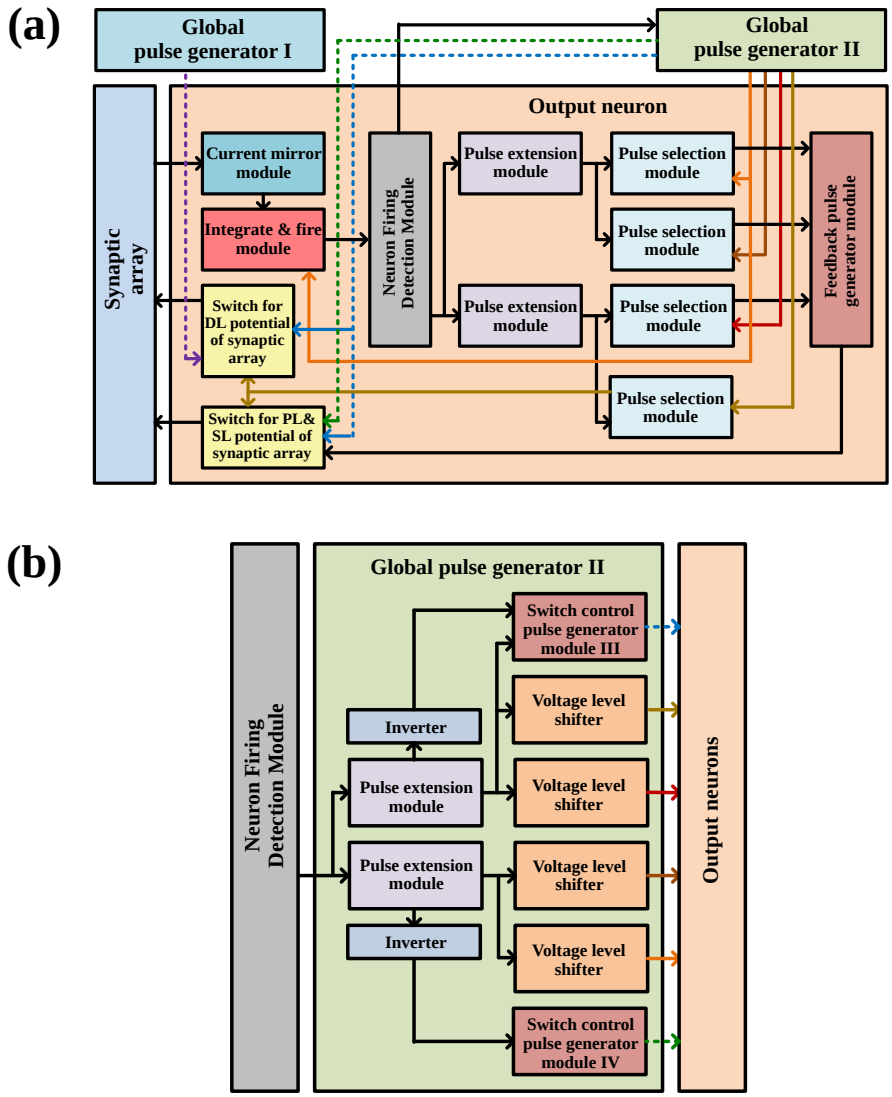


Fig. A.4. Conceptual diagrams of blocks including signal flows in the (a) output neuron and the (b) global pulse generator II to implement the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme not including the inhibition pulses.

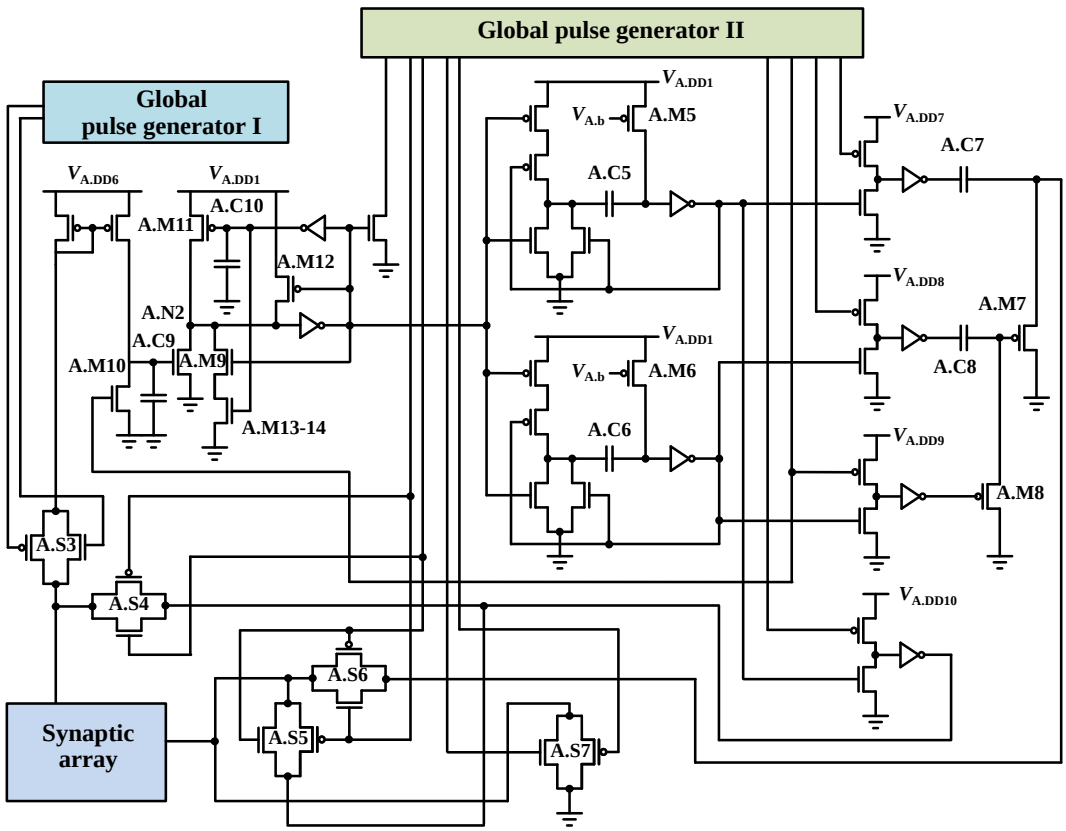


Fig. A.5. An output neuron circuit for generating the input pulse used in the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme not including the inhibition pulses.

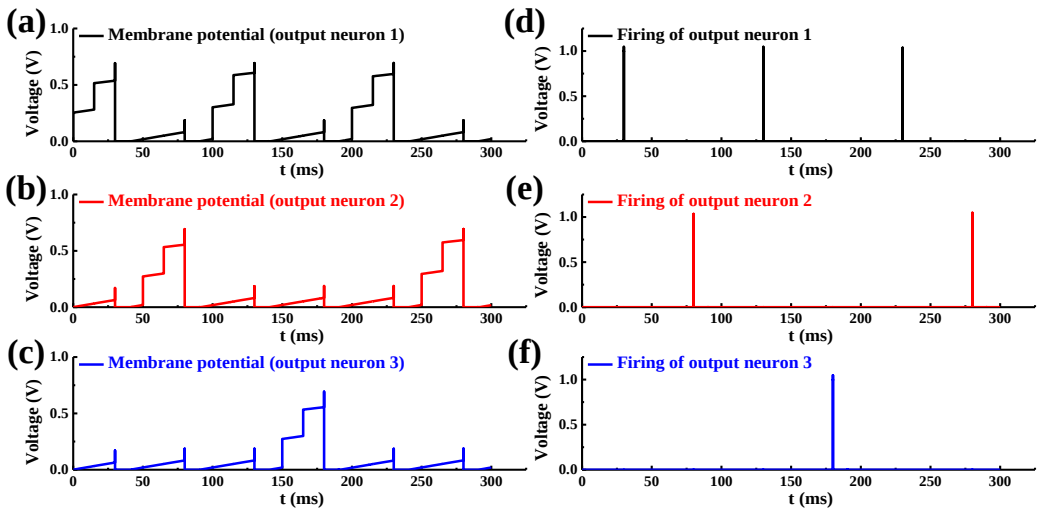


Fig. A.6. Operation of integrate-and-fire circuits to implement the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme not including the inhibition pulses. (a)-(c) Membrane potentials of representative output neurons included in the single output neuron layer. (d)-(f) Input signals of the global pulse generator II when output signals recognized by the neuron firing detection module.

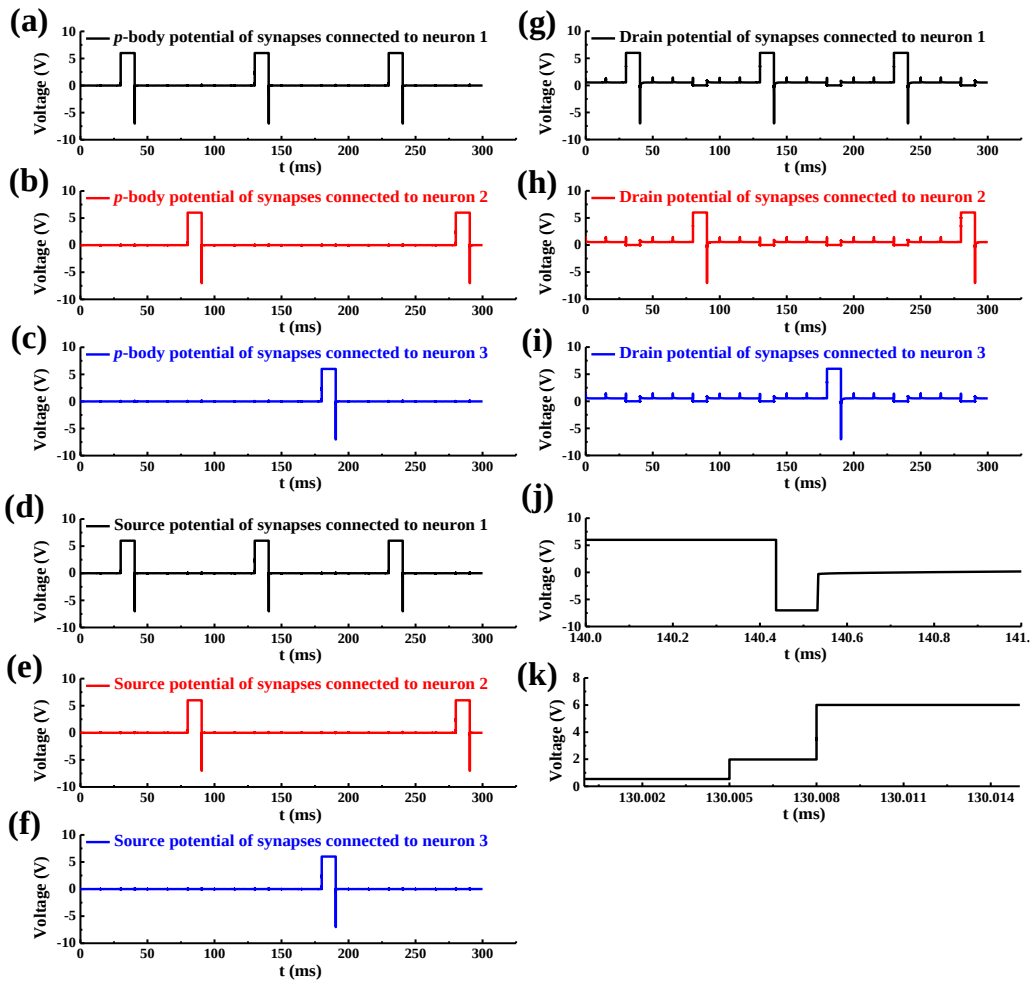


Fig. A.7. Potentials of (a)-(c) the PLs, (d)-(f) the SLs, and (g)-(i) the DLs of the electronic synaptic devices in the AND-type array connected to representative output neurons in the single output neuron layer. (j) and (k) show an enlarged view of the tail portion of the feedback pulse and the drain potential in the read operation, respectively.

Appendix B

Neuron circuits to implement a hardware-based neural network using the STDP learning algorithm and the pulse scheme including the inhibition pulses

Fig. B.1 (a) and (b) represent conceptual diagrams of blocks including signal flows in the input neuron and the global pulse generator I, respectively, to implement the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme including the inhibition pulses. In the pulse scheme as illustrated in Fig. 3.4, a pulse having a width required for the read operation is applied to the WLs of electronic synaptic devices connected to the fired output neuron while having the input signal. Also, the pulse with the constant amplitude and width should be applied to the WLs of the electronic synaptic devices even when no input signal is applied. For this reason, the input neuron uses logic circuits to determine the input pulse depending on whether there is an input signal or not

and whether an output neuron fires. A detailed description of the operation in the input neuron is as follows. First, in the case of the input neuron with the input signal, the input signal is converted to the pulse width for the read operation through the pulse extension module. At this time, when the output neuron fires during the read operation, the output of the pulse extension module connected to the output of the AND gate has the same width as the pulse generated by the fired output neuron in the global pulse generator II. On the contrary, if no output neuron fires during the read operation, the output of the pulse extension module connected to the output of the AND gate remains in the low state, and the input of the XOR gate from the global pulse generator II also remains the low state. In other words, in the presence of the input signal such that the input pulse is involved in the read operation, the output of the XOR gate remains low state regardless of whether or not the output neuron fires. As a result, in the input neuron that accepts the input signal, the input pulse that affects the read operation can be applied to the WL of the electronic synaptic array. On the other hand, in the input neuron that does not receive the input signal, the output of the pulse extension module connected to the output of the AND

gate remains low by the same principle. If there is the fired output neuron at this time, the output of the XOR gate changes to the high state, allowing a certain pulse to be applied to the WL of the electronic synaptic array as shown in Fig. 3.4. The dotted arrow generated by the global pulse generator I represent the switch control pulses that allow the output neuron to receive current only during the read operation of the electronic synaptic array. Fig. B.2 shows the input neuron circuit for generating the input pulse used in the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme including the inhibition pulses. The operation principles of the pulse extension module and pulse selection module included in the input neuron circuit were described in Appendix A. The A.S8 and A.S9 switches allow one of the pulses involved in the read operation or PGM operation to be assigned to the electronic synaptic array according to the operation of the input neuron. Fig. B.3 shows the operation of input neuron circuits to implement the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme including the inhibition pulses. The input signals received by the several input neurons as shown in Fig. B.3 (a)-(c) produce the inputs

of the global pulse generator I as shown in Fig. B.3 (d). Fig. B.3 (e)-(g) show the pulses that are generated through the input neuron circuit and applied to the WL of the electronic synaptic array connected to each input neuron. In each neuron, the input pulse is generated that is involved in the read operation when the input signal is sent to the input neuron. In addition, when output neuron fires, neurons that don't do not receive an input signal generate input pulses that are involved in LTD process. Fig. B.3 (h) shows an enlarged view of the part of the input pulse that is involved in the read operation.

Fig. B.4 (a) and (b) represent conceptual diagrams of blocks including signal flows in the output neuron and the global pulse generator II, respectively, to implement the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme including the inhibition pulses. The integrate-and-fire module, the current receiving part of the output neuron, is identical to that of the output neuron described in Appendix A. Since the form of the feedback pulse to be generated from the output neuron is similar to that of the input pulse, the components of the output neuron and global pulse generator II are almost the same

as those of the input neuron and global pulse generator I. In the pulse scheme using the INH pulses, the output neuron should select one of the pulses for the ERS operation of the electronic synaptic array and the INH pulse depending on whether the output neuron fires or not. A detailed description of the operation in the output neuron is as follows. The pulse extension module connected to the input terminal of the XOR gate determines the width of the INH pulse. When the output neuron fires, the output of the XOR gate remains low state because the two pulses received by the XOR gate are high state. On the contrary, when the output neuron does not fire, only the pulse received by the XOR gate from the global pulse generator II becomes high, and the output of the XOR gate changes to the high state. This operation occurs, of course, when one of the output neurons included in the single output neuron layer fires. This logic circuit allows the INH pulse to be given to the PLs of the electronic synaptic array only when the output neuron does not fire as shown in Fig. 3.4. The output pulse from the pulse selection module connected to the pulse extension module that determines the width of the INH pulse acts as a control pulse to select one of the pulses for the ERS operation and the INH pulse.

Fig. B.5 shows the output neuron circuit for generating the input pulse used in the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme including the inhibition pulses. The operation principles of the major modules including the integrate-and-fire module, pulse extension module, and pulse selection module were covered in Appendix A. The A.S10 and A.S11 switches help select the ERS Pulse or INH pulse as the feedback pulse. In addition, A.S12 and A.S13 switches determine the potential of the DL of the electronic synaptic array. Fig. B.6 shows the behavior of each representative neuron circuit in the single output neuron layer. Similar to the operation of the integrate-and-fire module stated in Appendix A, it can be confirmed that the integrate-and-fire module of the fired output neuron emits the spike signal when the V_{mem} of a specific output neuron exceeds a certain potential. Fig. B.7 (a)-(c), (d)-(f), and (g)-(i) show the potential of the PLs, the SLs, and the DLs, respectively, of the electronic synaptic devices in the AND-type array connected to representative output neurons in the single output neuron layer. Simulation results for the potential of the PL, the SL, and the DL of the electronic synaptic devices due to the feedback pulse show that

the pulse scheme described in Fig. 3.4 is effectively implemented by the designed output neuron circuit. That is, it is confirmed that the feedback pulse for the LTP process of the electronic synaptic device is selectively applied to the PL of the array connected to the fired output neuron. Also, it is confirmed that the INH pulses are applied to the PLs of the electronic synaptic array connected to the non-fired output neurons. The potential of the DL is maintained at a certain read voltage only during the read operation time of the electronic synaptic array and becomes the floating node during the rest of the time. Fig. B.7 (j) and (k) show an enlarged view of the tail portion of the feedback pulse and the drain potential in the read operation, respectively. As shown in Fig. B.7 (j), a synaptic device that is connected to the firing neuron and receives no input is subjected to LTD by the input pulse with V_{PGM} amplitude applied to the WL while the feedback pulse is maintained at 0 V.

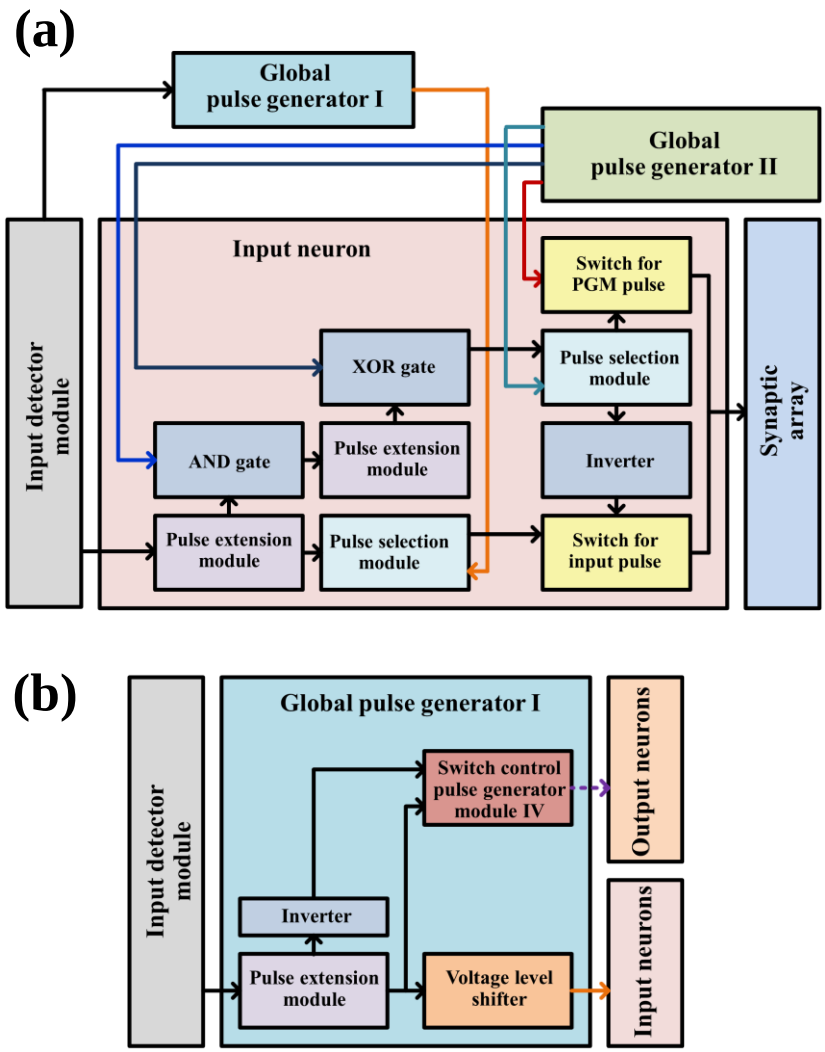


Fig. B.1. Conceptual diagrams of blocks including signal flows in the (a) input neuron and the (b) global pulse generator I to implement the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme including the inhibition pulses.

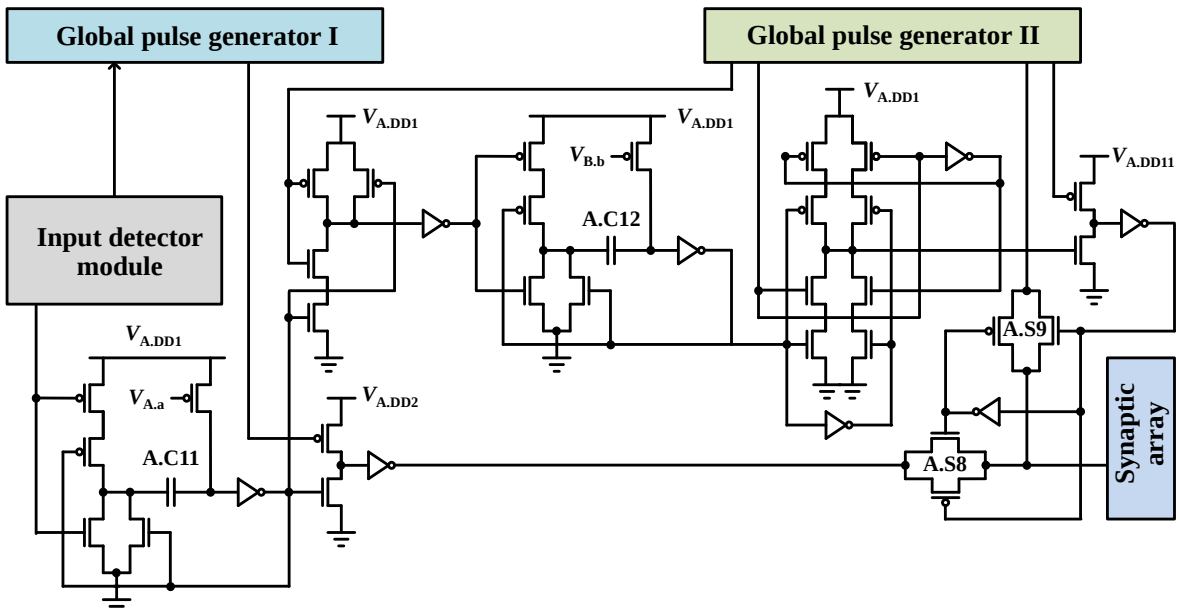


Fig. B.2. An input neuron circuit for generating the input pulse used in the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme including the inhibition pulses.

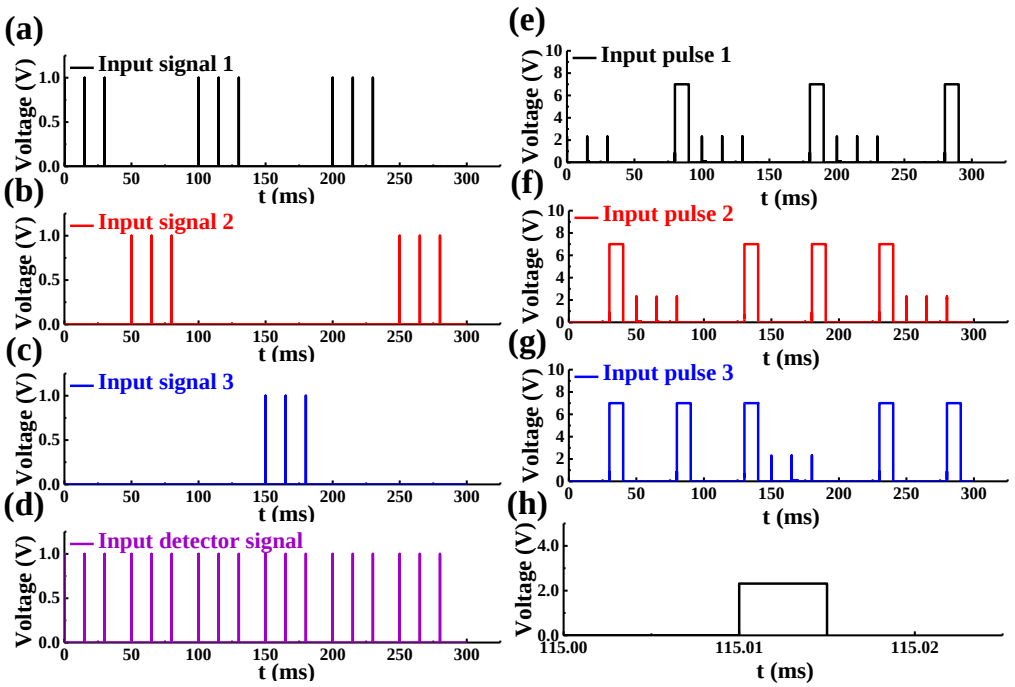


Fig. B.3. Operation of input neuron circuits to implement the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme including the inhibition pulses. (a)-(c) Input signals received by representative input neurons included in the single input neuron layer. (d) Input signals of the global pulse generator I when input signals are recognized by the input detector module. (e)-(g) Input pulses generated by the input neuron circuit in each representative neuron. (h) An enlarged view of the part of the input pulse that is involved in the read operation.

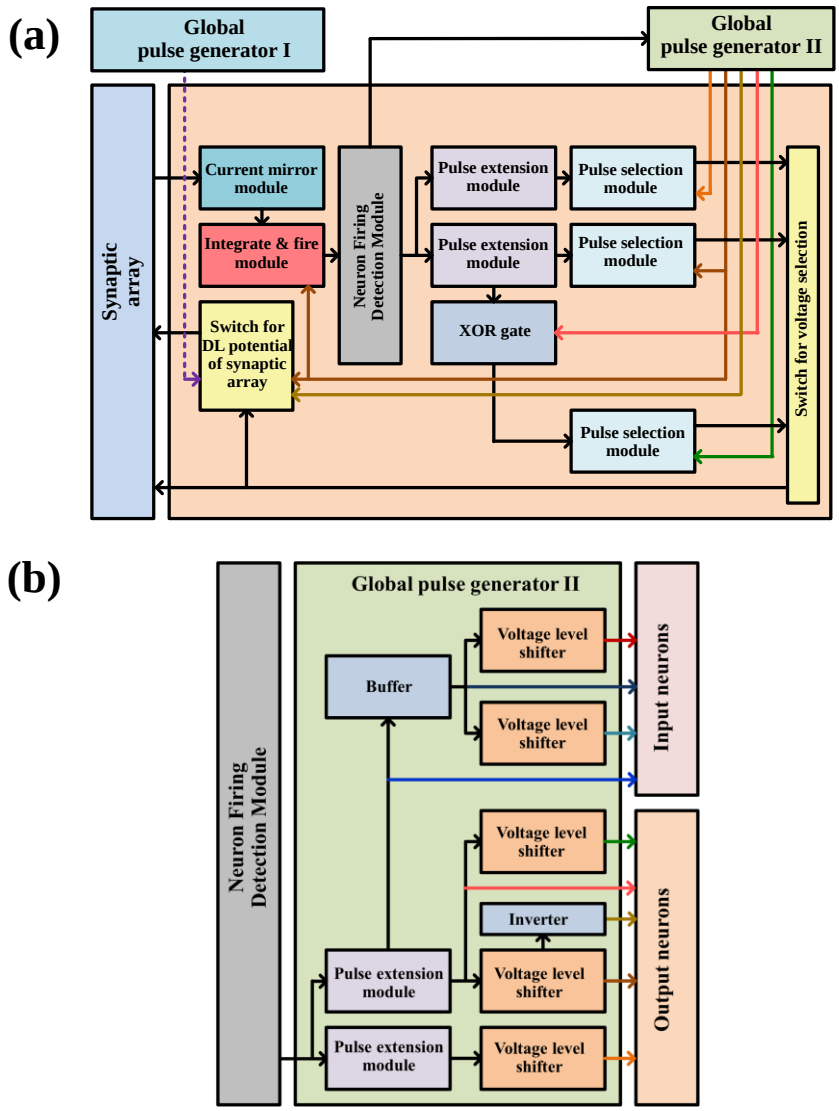


Fig. B.4. Conceptual diagrams of blocks including signal flows in the (a) output neuron and the (b) global pulse generator II to implement the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme including the inhibition pulses.

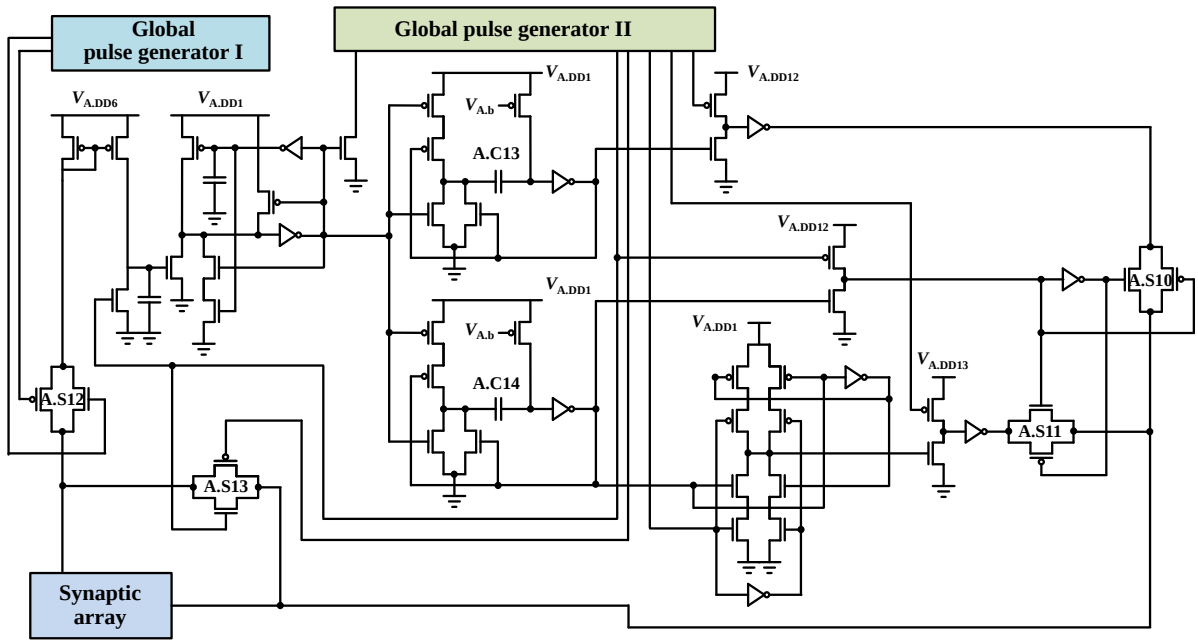


Fig. B.5. An output neuron circuit for generating the input pulse used in the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme including the inhibition pulses.

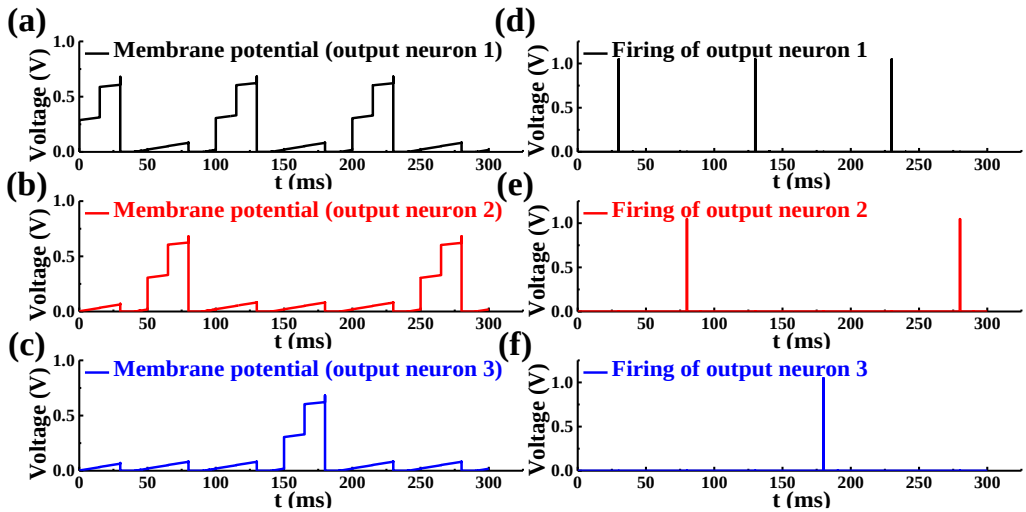


Fig. B.6. Operation of integrate-and-fire circuits to implement the designed hardware-based neural network using the STDP learning algorithm and the pulse scheme including the inhibition pulses. (a)-(c) Membrane potentials of representative output neurons included in the single output neuron layer. (d)-(f) Input signals of the global pulse generator II when output signals recognized by the neuron firing detection module.

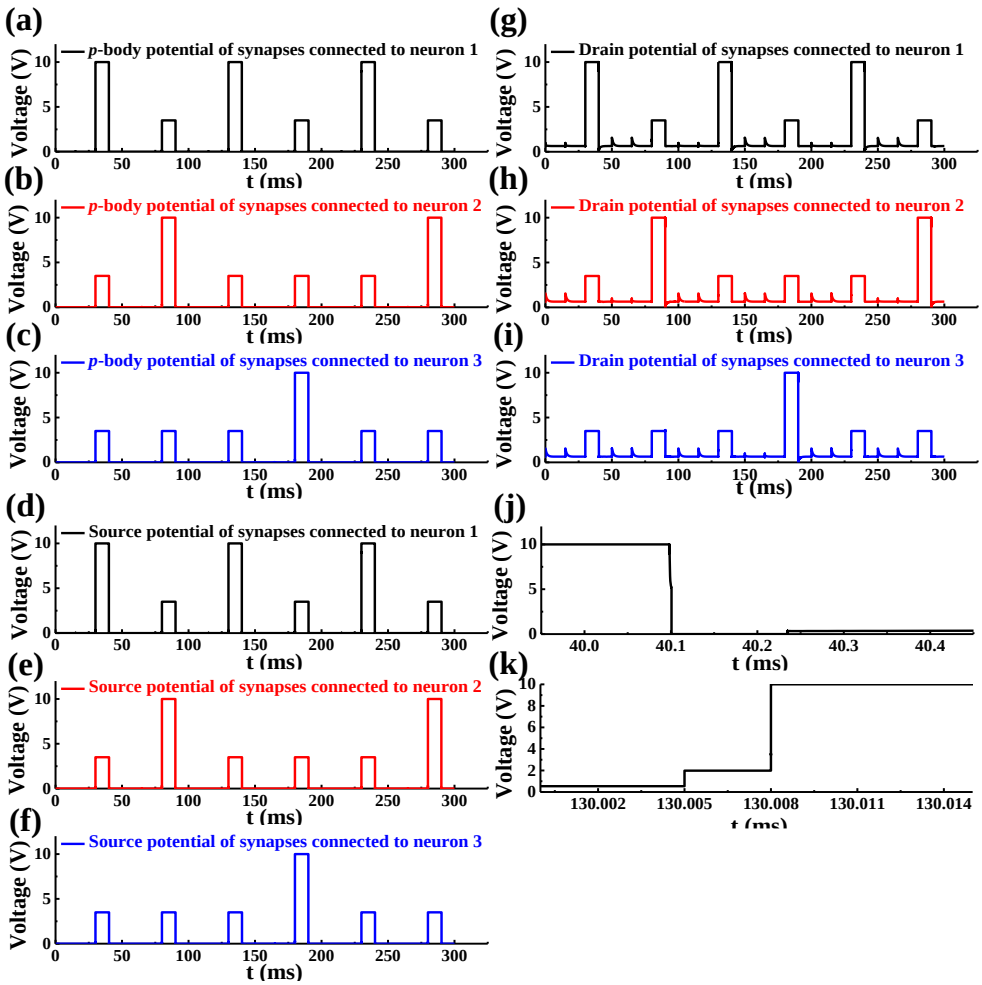


Fig. B.7. Potentials of (a)-(c) the PLs, (d)-(f) the SLs, and (g)-(i) the DLs of the electronic synaptic devices in the AND-type array connected to representative output neurons in the single output neuron layer. (j) and (k) show an enlarged view of the tail portion of the feedback pulse and the drain potential in the read operation, respectively.

Bibliography

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NIPS)*, pp. 1097-1105, 2012.

[2] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A. -R. Mohamed, G. Dahl, and B. Ramabhadran, "Deep Convolutional Neural Networks for Large-scale Speech Tasks," *Neural Networks*, vol. 64, pp. 39-48, 2015.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinnovich, "Going deeper with convolutions," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-9, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016.

[5] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," *IEEE Transactions on Neural Networks Learning Systems*, vol. 28, no. 10, pp. 2222-2232, 2017.

[6] A. Graves, A. -R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6645-6649, 2013.

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

[8] K. -H. Kim, S. Gaba, D. Wheeler, J. M. Cruz-Albrecht, T. Hussain, N. Srinivasa,

and W. Lu, “A Functional Hybrid Memristor Crossbar-Array/CMOS System for Data Storage and Neuromorphic Applications,” *Nano Letter*, vol. 12, no. 1, pp.389-395, 2012.

[9] M. Hu, H. Li, Y. Chen, Q. Wu, G. S. Rose, and R. W. Linderman, “Memristor Crossbar-Based Neuromorphic Computing System: A Case Study,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 10, pp. 1864-1878, 2014.

[10] E. O. Neftci, C. Augustine, S. Paul, and G. Detorakis, “Event-Driven Random Back-Propagation: Enabling Neuromorphic Deep Learning Machines,” *Frontiers in Neuroscience*, vol. 11, article 324, 2017.

[11] S. Lim, J. -H. Bae, J. -H. Eum, S. Lee, C. -H. Kim, D. Kwon, B. -G. Park, and J. -H. Lee, “Adaptive learning rule for hardware-based deep neural networks using electronic synapse devices,” *Neural Computing and Applications*, vol. 31, pp. 8101–8116, 2018.

[12] C. -C. Chang, P. -C. Chen, T. Chou, I. -T. Wang, B. Hudec, C. -C. Chang, C. -M. Tsai, T. -S. Chang, and T. -H. Hou, “Mitigating Asymmetric Nonlinear Weight Update Effects in Hardware Neural Network Based on Analog Resistive Synapse,” *IEEE Journal of Emerging Selected Topics in Circuits and Systems*, vol. 8, no. 1, pp. 116-124, 2018.

[13] D. Querlioz, O. Bichler, P. Dollfus, and C Gamrat, “Immunity to device variations in a spiking neural network with memristive nanodevices,” *IEEE Transactions on Nanotechnology*, vol. 12, issue 3, pp. 288-295, 2013.

[14] A. Basu, S. Shuo, H. Zhou, M. H. Lim, G. -B. Huang, “Silicon spiking neurons for hardware implementation of extreme learning machines,” *Neurocomputing*, vol. 102, pp. 125-134, 2012.

[15] N. K. Kasabov, “NeuCube: A spiking neural network architecture for mapping,

learning and understanding of spatio-temporal brain data," *Neural Networks*, vol. 52, pp. 62-76, 2014.

[16] X. Wu, V. Saxena, and K. Zhu, "Homogeneous Spiking Neuromorphic System for Real-World Pattern Recognition," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol.5, iss, 2, pp. 254-266, 2015.

[17] N. Qiao, H. Mostafa, F. Corradi, M. Osswald, F. Stefanini, D. Sumislawska, and G. Indiveri, "A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses," *Frontiers in Neuroscience*, vol. 9. article 141. 2015.

[18] H. Mostafa, "H Supervised learning based on temporal coding in spiking neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, issue 7, pp. 3227–3235, 2016.

[19] J. H. Le, T. Delbruck, and M. Pfeiffer, "Training deep spiking neural networks using backpropagation." *Frontiers in Neuroscience*, vol. 10, article 508, 2016.

[20] M. Hu, H. Li, Y. Chen, Q. Wu, G. S. Rose, and R. W. Linderman, "Memristor crossbar-based neuromorphic computing system: A case study," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, issue 10, pp. 1864-1878, 2014.

[21] K. -H. Kim, S. Gaba, D. Wheeler, J. M. Cruz-Albrecht, T. Hussain, N. Srinivasa, and W. Lu, "A functional hybrid memristor crossbar-array/CMOS system for data storage and neuromorphic applications," *Nano Letter*, vol. 12, no. 1, pp. 389-395, 2011.

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

[23] G. Malavena, A. S. Spinelli, and C. M. Compagnoni, "Implementing Spike-

Timing-Dependent Plasticity and Unsupervised Learning in a Mainstream NOR Flash Memory Array," *2018 IEEE International Electron Devices Meeting (IEDM)*, pp. 2.3.1-2.3.4, 2018.

[24] S. Lee, H. Kim, J. Bae, H. Yoo, N. Y. Choi, D. Kwon, S. Lim, B. -G. Park, and J. -H. Lee, "High-Density and Highly-Reliable Binary Neural Networks Using NAND Flash Memory Cells as Synaptic Devices," *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 38.4.1-38.4.4, 2019.

[25] X. Zhang et al., "Fully Memristive SNNs with Temporal Coding for Fast and Low-power Edge Computing," *2020 IEEE International Electron Devices Meeting (IEDM)*, pp. 29.6.1-29.6.4, 2020.

[26] T. Soliman *et al*., "Ultra-Low Power Flexible Precision FeFET Based Analog In-Memory Computing," *2020 IEEE International Electron Devices Meeting (IEDM)*, pp. 29.2.1-29.2.4, 2020.

[27] M. Suri, O. Bichler, D. Querlioz, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Phase change memory as synapse for ultra-dense neuromorphic systems: application to complex visual pattern extraction," *IEEE Electron Devices Meeting*, pp. 4.4.1-4.4.4, 2011.

[28] S. B. Eryilmaz, D. Kuzum, G. D. Jeyasingh, S. B. Kim, M. BrightSky, C. Lam, and H. -S. Philip Wong, "Experimental demonstration of array-level learning with phase change synaptic devices," *IEEE International Electron Devices Meeting (IEDM)*, pp. 25.5.1-25.5.4, 2013.

[29] D. Kuzum, R. G. D. Jeyasingh, B. Lee, and H. -S. Philip Wong, "Nanoelectronic Programmable Synapses Based on Phase Change Materials for Brain-inspired Computing," *Nano letters*, vol. 12, no. 5, pp. 2179-2186, 2012.

[30] K. Seo, I. Kim, S. Jung, M. Jo, S. Park, J. Park, J. Shin, K. P. Biju, J. Kong, K. Lee, B. Lee, and H. Hwang, "Analog memory and spike-timing-dependent

plasticity characteristics of a nanoscale titanium oxide bilayer resistive switching device," *Nanotechnology*, vol. 22, no. 25, article 254023, 2011.

[31] M. Zhao, H. Wu, B. Gao, Q. Zhang, W. Wu, S. Wang, Y. Xi, D. Wu, N. Deng, S. Yu, H. Y. Chen, and H. Qian, "Investigation of statistical retention of filamentary analog RRAM for neuromorphic computing," *IEEE International Electron Devices Meeting (IEDM)*, pp.39.4.1-39.4.4, 2017.

[32] T. Ohno, T. Hasegawa, T. Tsuruoka, K. Terabe, J. K. Gimzewski, and M. Aono, "Short-term plasticity and long-term potentiation mimicked in single inorganic synapses," *Nature Materials*, vol. 10, pp. 591-595, 2011.

[33] S. Yu, B. Gao, Z. Fang, H. Yu, J. Kang, and H. -S. Philip Wong, "A neuromorphic visual system using RRAM synaptic devices with Sub-pJ energy and tolerance to variability: Experimental characterization and large-scale modeling," *IEEE Electron Devices Meeting*, pp. 10.4.1-10.4.4, 2012.

[34] S. Lequeux, J. Sampaio, V. Cros, K. Yakushiji, A. Fukushima, R. Matsumoto, H. Kubota, S. Yuasa, and J. Grollier, "A magnetic synapse: multilevel spin-torque memristor with perpendicular anisotropy," *Scientific Reports*, vol. 6, article 31510, 2016.

[35] A. F. Vincent, J. Larroque, W. S. Zhao, N. B. Romdhane, O. Bichler, C. Gamrat, J. -O. Klein, S. Galdin-Retailleau, and D. Querlioz, "Spin-transfer torque magnetic memory as a stochastic memristive synapse," *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1074-1077, 2014.

[36] G. Srinivasan, A. Sengupta, and K. Roy, "Magnetic Tunnel Junction Based Long-Term Short-Term Stochastic Synapse for a Spiking Neural Network with On-Chip STDP Learning," *Scientific Reports*, vol. 6, article 29545, 2016.

[37] W.A. Borders, H. Akima, S. Fukami, S. Moriya, S. Kurihara, Y. Horio, S. Sato, and H. Ohno, "Analogue spin–orbit torque device for artificial-neural-network-

based associative memory operation," *Applied Physics Express*, vol. 10, no. 1, article 013007, 2017.

[38] A. Chanthbouala, V. Garcia, R. O. Cherifi, K. Bouzehouane, S. Fusil, X. Moya, S. Xavier, H. Yamada, C. Deranlot, N. D. Mathur, M. Bibes, A. Barthelemy, and J. Grollier, "A ferroelectric memristor," *Nature Materials*, vol. 11, pp. 860-864, 2012.

[39] 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, no. 10, article 104011, pp. 1-9, 2014.

[40] 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, iss. 5, pp. 1774–1780, 2018.

[41] C. Diorio, P. Hasler, A. Minch, and C. A. Mead, "A single-transistor silicon synapse," *IEEE Transactions on Electron Devices*, vol. 43, iss. 11, pp. 1972-1980, 1996.

[42] G. Malavena, A. S. Spinelli, C. M. Compagnoni, "Implementing spike-timing-dependent plasticity and unsupervised learning in a mainstream NOR Flash Memory Array," *IEEE International Electron Devices Meeting (IEDM)*, pp.2.3.1-2.3.4, 2018.

[43] 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 International Electron Devices Meeting (IEDM)*, pp. 440-443, 2016.

[44] C. Zamarreno-Ramos, L. A. Camunas-Mesa, J. A. Perez-Carrasco, T. Masquelier, T. Serrano-Gotarredona, and B. Linares-Barranco, "On spike-timing-

dependent plasticity, memristive devices, and building a self-learning visual cortex," *Frontiers in Neuroscience*, vol. 5, article 26, pp. 1-22, 2011.

[45] P. U. Diehl, and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Front. Comput. Neurosci.*, vol. 9, article 99, pp. 1-9, 2015.

[46] V. Milo, G. Pedretti, R. Carboni, A. Calderoni, N. Ramaswamy, S. Ambrogio, and D. Ielmini, "A 4-Transistors/1-Resistor hybrid synapse based on resistive switching memory (RRAM) Capable of spike-rate-dependent plasticity (SRDP)," *IEEE Trans. VLSI Systems*, vol. 26, iss. 12, pp. 2806-2815, 2018.

[47] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," *Advances in Neural Information Processing Systems*, pp. 396-404, 1989.

[48] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11-26, 2017.

[49] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A. -R. Mohamed, G. Dahl, and B. Ramabhadran, "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks*, vol. 64, pp. 39-48, 2015.

[50] Y. Guo, Y. Liu, A, Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27-48, 2016.

[51] G. Indiveri, F. Corradi, and N. Qiao, "Neuromorphic architectures for spiking deep neural networks," *IEEE International Electron Devices Meeting (IEDM)*, pp. 4.2.1-4.2.4 , 2015.

[52] Y. Wu, L. Deng, G. Li, J. Zhu and L. Shi, "Spatio-temporal backpropagation for training high-performance spiking neural networks", *Front. Neurosci.*, vol. 12, pp. 1-12, 2018.

[53] J. H. Lee, T. Delbruck, and M. Pfeiffer, “Training deep spiking neural networks using backpropagation,” *Front. Neurosci.*, vol. 10, article 508, pp. 1-13, 2016.

[54] S. Grossberg, “Competitive learning: From interactive activation to adaptive resonance,” *Cognitive science*, vol. 11, iss. 1, pp. 23-63, 1987.

[55] D. Zipser and D. E. Rumelhart, “The neurobiological significance of the new learning models,” *Computational Neuroscience*, pp. 192-200, 1993.

[56] T. Masquelier, and S. J. Thorpe, “Unsupervised Learning of Visual Features through Spike Timing Dependent Plasticity,” *PLoS Computational Biology*, vol. 3, no. 2, pp. 247–257, 2007.

[57] 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,” *Frontiers in Neuroscience*, vol. 10, article 56, 2016.

[58] S. R. Kheradpisheh, M. Ganjtabesh, T. Masquelier, “Bio-inspired unsupervised learning of visual features leads to robust invariant object recognition,” *Neurocomputing*, vol. 205, pp. 382-392, 2016.

[59] D. Querlioz, O. Bichler and C. Gamrat, "Simulation of a memristor-based spiking neural network immune to device variations," *The 2011 International Joint Conference on Neural Networks*, pp. 1775-1781, 2011.

[60] J. Bill and R. Legenstein, “A compound memristive synapse model for statistical learning through STDP in spiking neural networks”, *Frontiers in Neuroscience*, vol. 8, article 412, 2014.

[61] T. P. Lillicrap, D. Cownden, D. B. Tweed, and C. J. Akerman, “Random synaptic feedback weights support error backpropagation for deep learning,” *Nature Comnunications*, vol. 7, article 13276, pp. 1-10, 2016.

[62] A. Nokland, “Direct feedback alignment provides learning in deep neural networks,” *Advances in Neural Information Processing Systems* 29, pp. 1037-1045, 2016.

[63] B. Crafton, A. Parihar, E. Gebhardt, and A. Raychowdhury, “Direct feedback alignment with sparse connections for local learning,” *Front. Neurosci.*, vol. 13, article 525, pp. 1-12, 2019.

[64] E. O. Neftci, C. Augustine, S. Paul, and F. Detorakis, “Event-driven random back-propagation: Enabling neuromorphic deep learning machines,” *Front. Neurosci.*, vol. 11, article 324, 2017.

[65] W. -M. Kang, C. -H. Kim, S. Lee, S. Y. Woo, J. -H. Bae, B. -G. Park, J. -H. Lee, “A Spiking Neural Network with a Global Self-Controller for Unsupervised Learning Based on Spike-Timing-Dependent Plasticity Using Flash Memory Synaptic Devices,” *International Joint Conference on Neural Networks (IJCNN)*, pp. 1-7, 2019.

[66] S. Lim, D. Kwon, J. -H. Eum, S. -T. Lee, J. -H. Bae, H. Kim, C. –H. Kim, B. -G. Park, “Highly reliable inference system of neural networks using gated Schottky diodes,” *IEEE Journal of the Electron Devices Society*, vol. 7, pp. 522-528, 2019.

[67] D. Kwon, S. Lim, J. -H. Bae, S. -T. Lee, H. Kim, C. -H. Kim, B. -G. Park, J. -H. Lee, “Adaptive weight quantization method for nonlinear synaptic devices,” *IEEE Transactions on Electron Devices*, vol. 66, iss. 1, pp. 395-401, 2018.

[68] W. -M. Kang, D. Kwon, S. Y. Woo, S. Lee, H. Yoo, J. Kim, B. -G. Park, J. -H. Lee, “Hardware-Based Spiking Neural Network Using a TFT-Type AND Flash Memory Array Architecture Based on Direct Feedback Alignment,” *IEEE Access*, vol. 9, pp. 73121-73132, 2021.

초 록

뉴로모픽 기술은 폰 노이만 프로세서의 대안으로서 두뇌에서 영감을 받은 컴퓨팅 아키텍처를 구현하는 것을 목표로 한다. 이 논문에서는 박막 트랜지스터형 및 플래시 메모리 어레이 아키텍처를 사용하여 온칩 훈련을 가능하게 하는 하드웨어 기반 신경망을 설계한다. 어레이를 구성하는 시냅스 소자는 도핑된 p형 바디, SiO_2 / Si_3N_4 / Al_2O_3로 구성된 게이트 절연막 스택 및 부분적으로 구부러진 폴리실리콘 채널을 특징으로 한다. 시냅스 소자 구조에 포함된 바디 영역은 시냅스 가중치를 변경할 때 소스 및 드레인 라인 모두에 필요한 고전압 드라이버의 회로 부담을 줄일 수 있다. 또한 게이트 절연막 스택에 포함된 high-κ 물질은 시냅스 소자의 동작 전압을 낮출 수 있다. 시냅스 소자의 크기가 축소됨에 따라 소자의 구조적인 특징은 메모리 동작의 효율성 뿐만 아니라 어레이의 비트 라인에서 발생하는 전압 강하 효과에 대한 내성을 증가시킨다. 우리는 제작된 시냅스 소자를 이용한 AND형 어레이 구조에서 선택적인 메모리 동작을 위한 펄스 방식을 제안하고 실험적으로 검증한다. 이후 제작된 시냅스 소자 및 어레이의 측정된 특성을 기반으로 학습 목적에 따라 2가지 유형의 하드웨어 기반 스파이크 신경망 (SNN)을 설계한다. 먼저 스파이크 시점 의존 가소성 기반 학습 규칙을 이용하여 비지도 학습을 위한 하드웨어 기반 SNN을 제안한다. 설계된 네트워크는 외부 회로에서 펄스를 생성하지 않으며 각 스파이크 뉴런 회로에서 필요한 펄스들이 생성된다. 이러한 네트워크에서 스파이크 시점 의존 가소

성 기반 학습 규칙은 폴리실리콘 AND형 어레이를 사용하기 위한 효과적인 펄스 구동 방식을 통해 구현된다. 제안된 펄스 구동 방식과 SNN을 기반으로 200개의 출력 뉴런을 사용하는 MNIST 필기 숫자 패턴 학습에서 91.63 %의 인식 정확도를 얻을 수 있다. 두 번째로, 우리는 직접 피드백 정렬 학습 규칙을 사용하여 지도 학습을 위한 하드웨어 기반 SNN을 제안한다. 순방향 경로와 역방향 경로에서 동일한 시냅스 가중치를 가질 필요가 없는 직접 피드백 정렬 알고리즘으로 인해 AND형 어레이 아키텍처는 효율적인 온칩 훈련 신경망 설계에 활용될 수 있다. AND형 어레이 아키텍처에 적합한 펄스 구동 방식도 신경망에서 직접 피드백 정렬 알고리즘을 구현하기 위해 고안된다. 시스템 수준 시뮬레이션에서 제안된 펄스 구동 방식과 컴퓨팅 아키텍처를 기반으로 하는 MNIST 패턴 학습에서 최대 97.01%의 인식 정확도를 얻을 수 있다. 또한, 우리는 제안된 시냅스 어레이와 CMOS 회로의 집적 공정 과정을 제안하고 이를 검증한다. 제안하는 집적 공정 방법은 시냅스 어레이와 CMOS 회로의 공정 과정을 공유함으로써 마스크와 공정 수를 줄일 수 있는 장점이 있다. 제안된 집적 공정 방법은 제안하는 시냅스 소자와 CMOS와의 우수한 호환성을 검증할 뿐만 아니라, 하드웨어 기반 신경망을 효율적으로 구현할 수 있는 방법론을 제시한다는 점에서 의의를 갖는다.

주요어 : 하드웨어 기반 스파이크 신경망, 플래시 메모리 시냅스 소자, AND형 어레이, 온칩 학습, 비지도 학습, 지도 학습, 뉴런 회로.
학번 : 2015-20881

List of Publications

Journals

1. ***Won-Mook Kang**, *Dongseok Kwon, Sung Yun Woo, Soochang Lee, Honam Yoo, Jangsaeng Kim, Byung-Gook Park, and Jong-Ho Lee, "Hardware-Based Spiking Neural Network Using a TFT-Type AND Flash Memory Array Architecture Based on Direct Feedback Alignment," *IEEE Access,* vol. 9, pp. 73121-73132, 2021.
2. Jangsaeng Kim, Dongseok Kwon, Sung Yun Woo, **Won-Mook Kang**, Soochang Lee, Seongbin Oh, Chul-Heung Kim, Jong-Ho Bae, Byung-Gook Park, and Jong-Ho Lee, "Hardware-based spiking neural network architecture using simplified backpropagation algorithm and homeostasis functionality," *Neurocomputing,* vol. 428, pp. 153-165, 2021.
3. Jangsaeng Kim, Dongseok Kwon, Sung Yun Woo, **Won-Mook Kang**, Soochang Lee, Seongbin Oh, Chul-Heung Kim, Jong-Ho Bae, Byung-Gook Park, and Jong-Ho Lee, "On-chip trainable hardware-based deep Q-networks approximating a backpropagation algorithm," *Neural computing and applications,* vol. 33, pp. 9391-9402, 2021.
4. *Sung Yun Woo, *Dongseok Kwon, Nagyong Choi, **Won-Mook Kang**, Young-Tak Seo, Min-Kyu Park, Jong-Ho Bae, Byung-Gook Park, and Jong-Ho Lee, "Low-Power and High-Density Neuron Device for Simultaneous Processing of Excitatory and Inhibitory Signals in Neuromorphic Systems," *IEEE Access,* vol. 8, pp. 202639-202647, 2020.
5. 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 double-gate device for spiking neural network", *Solid-State Electronics*, vol. 165, p. 107741, 2020.

6. 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", *Solid-State Electronics*, vol. 165, p. 107742, 2020.

7. **Won-Mook Kang,** In-Tak Cho**,** Jeongkyun Roh, Changhee Lee, and Jong-Ho Lee, "Low-Frequency Noise Characteristics in Multi-Layer WSe_2 Field Effect Transistors with Different Contact Metals," *Journal of Nanoscience and Nanotechnology*, vol. 19, no. 10, pp. 6422-6428, 2019.

8. Sung Yun Woo, Kyu-Bong Choi, Suhwan Lim, Sung-Tae Lee, Chul-Heung Kim, **Won-Mook Kang**, Dongseok Kwon, Jong-Ho Bae, Byung-Gook Park, and Jong-Ho Lee, "Synaptic device Using a Floating Fin-Body MOSFET With Memory Functionality for Neural Network," *Solid-State Electronics*, vol. 156, pp. 23-27, 2019.

9. *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, 2018.

10. 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, article 704, 2018.

11. 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, 2018.

12. **Won-Mook Kang**, SungTae Lee, In-Tak Cho, Tae-Hyung Park, Hyeonwoo Shin, Cheol Seong Hwang, Changhee Lee, Byung-Gook Park, and Jong-Ho Lee, "Multi-layer WSe_2 field effect transistor with improved carrier-injection contact by using oxygen plasma treatment," *Solid-State Electronics*, vol.140, pp. 2-7, 2018.

13. Yoonki Hong, **Won-Mook Kang**, In-Tak Cho, Jongmin Shin, Meile Wu, and Jong-Ho Lee, "Gas-Sensing Characteristics of Exfoliated WSe_2 Field-Effect Transistors", *Journal of Nanoscience and Nanotechnology*, vol. 17, no. 5, pp. 3151-3154, 2017.

14. Jun-Mo Park, In-Tak Cho, **Won-Mook Kang**, Byung-Gook Park, and Jong-Ho Lee, "Method to Eliminate Gate and Drain Bias Stresses in Transfer Curves of WSe_2 Field Effect Transistors with Single Channel Pulsed I–V Measurement," *Journal of Nanoscience and Nanotechnology*, vol. 17, no. 5, pp. 3382-3385, 2017.

15. Sung Tae Lee, In-Tak Cho, **Won-Mook Kang**, Byung-Gook Park, and Jong-Ho Lee, "Accurate extraction of WSe_2 FETs parameters by using pulsed I-V method at various temperatures," *Nano Convergence*, vol. 3, article 31, 2016.

16. **Won-Mook Kang,** In-Tak Cho**,** Jeongkyun Roh, Changhee Lee, and Jong-Ho Lee, "High-gain complementary metal-oxide-semiconductor inverter based on multi-layer WSe_2 field effect transistors without doping," *Semiconductor Science and Technology*, vol. 31, no.10, 2016.

17. Jun-Mo Park, In-Tak Cho, **Won-Mook Kang**, Byung-Gook Park, and Jong-Ho Lee, "Elimination of the gate and drain bias stresses in I–V characteristics of WSe_2 FETs by using dual channel pulse measurement," *Applied Physics Letters*, vol. 109, issue 5, p.053503, 2016.

Conferences

1. Sung Yun Woo, **Won-Mook Kang**, Young-Tak Seo, Soochang Lee, Seongbing Oh and Jong-Ho Lee “Demonstration of Integrate-and-fire Neuron Circuit for Spiking Neural Networks," *The 28th Korean Conference on Semiconductors*, Jan. 2021.
2. Jangsaeng Kim, Sung Yun Woo, **Won-Mook Kang**, Soochang Lee, Seongbin Oh, Gyuho Yeom, Jiseong Im, Joon Hwang, Byung-Gook Park, and Jong-Ho Lee, “Hardware Implementation of Reinforcement Learning with Analog Synaptic Devices,” *The 28th Korean Conference on Semiconductors*, Jan. 2021.
3. Sung Yun Woo, Seongbin Oh, **Won-Mook Kang**, Young-Tak Seo, Soochang Lee, Jangsaeng Kim and Jong-Ho Lee, “A Low-Power Neuron Circuit Using a Positive Feedback Device for Spiking Neural Networks,” *ECS Meeting*, Nov. 2020.
4. **Won-Mook Kang,** Soochang Lee, Jangsaeng Kim, Byung-Gook Park, and Jong-Ho Lee, “Unsupervised Learning Architecture Based on Spike-Timing-Dependent Plasticity Using Flash Memory Synaptic Devices,” *IEEE Silicon Nanoelectronics Workshop (SNW)*, Jun. 2020.
5. Sung Yun Woo, **Won-Mook Kang**, Nagyong Choi, Young-Tak Seo, Soochang Lee, Seongbing Oh, Jangsaeng Kim, Byung-Gook Park, and Jong-Ho Lee “Analysis of Split-gate Positive Feedback Device for Neuron Circuit at Variable Temperatures," *The 27th Korean Conference on Semiconductors*, Feb. 2020.
6. Jangsaeng Kim, Sung Yun Woo, **Won-Mook Kang**, Byung-Gook Park, and Jong-Ho Lee, “Implementation of Homeostasis Functionality Using Active Leaky Path of Membrane Potential in STDP-based Spiking Neural Network,” *The 27th Korean Conference on Semiconductors*, Feb. 2020.

7. Jong-Ho Lee, Sung Yun Woo, Sung-Tae Lee, Suhwan Lim, **Won-Mook Kang**, Young-Tak Seo, Soochang Lee, Dongseok Kwon, Seongbin Oh, Yoohyun Noh, Hyeongsu Kim, Jangsaeng Kim, Jong-Ho Bae, “Review of candidate devices for neuromorphic applications,” *49th European Solid-State Device Research Conference (ESSDERC)*, Sep. 2019.

8. **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,” *International Joint Conference on Neural Networks (IJCNN)*, July. 2019.

9. Sung Yun Woo, **Won-Mook Kang**, Kyu-Bong Choi, Jangsaeng Kim, Chul-Heung Kim, Jong-Ho Bae, Byung-Gook Park, and Jong-Ho Lee, “Analyzation of Positive Feedback device with Steep Subthreshold Swing Characteristics in 14 nm FinFET Technology”, *IEEE Electron Devices Technology and Manufacturing Conference (EDTM)*, Mar. 2019.

10. **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.

11. 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.

12. 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.

13. **Won-Mook Kang**, In-Tak Cho, and Jong-Ho Lee, “Low-Frequency Noise (LFN) Characteristics in Multi-layer WSe_2 Field Effect Transistor with Different Contact Metals,” *The 25th Korean Conference on Semiconductors*, Feb. 2018.

14. **Won-Mook Kang**, SungTae Lee, In-Tak Cho, Tae-Hyung Park, Hyeonwoo Shin, Cheol Seong Hwang, Changhee Lee, Byung-Gook Park, and Jong-Ho Lee, “Multi-layer WSe_2 field effect transistor with improved carrier-injection contact by using oxygen plasma treatment,” *The 24th Korean Conference on Semiconductors*, Feb. 2017.

15. 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.

16. Jong-Ho Bae, Jun-Mo Park, Jai-Ho Eum, **Won-Mook Kang**, Jaeha Kim, Byung-Gook Park, and Jong-Ho Lee, " Reconfigurable Device with Programmable Bottom Gate Array,” *The 24th Korean Conference on Semiconductors*, Feb. 2017.

17. Jai-Ho Eum, Jun-Mo Park, Jong-Ho Bae, **Won-Mook Kang** and Jong-Ho Lee, "Study on Source/Drain Metal Contact in a Poly-Si Reconfigurable Field Effect Transistor Having Double-Gate Structure,” *Asia-Pacific Workshop on Fundamentals and Applications of Advanced Semiconductor Devices (AWAD)*, July. 2016.

18. Jun-Mo Park, In-Tak Cho, **Won-Mook Kang**, Byung-Gook Park, and Jong-Ho Lee, “Method to Eliminate Gate and Drain Bias Stresses in Transfer Curves of WSe_2 Field Effect Transistors with Single Channel Pulsed I–V Measurement,” *The 23th Korean Conference on Semiconductors*, Feb. 2016.

19. Yoonki Hong, **Won-Mook Kang**, In-Tak Cho, Meile Wu, and Jong-Ho Lee, "Gas-Sensing Characteristics of Exfoliated WSe_2 Field-Effect Transistors", *The 23th Korean Conference on Semiconductors*, Feb. 2016.

20. **Won-Mook Kang**, In-Tak Cho, and Jong-Ho Lee, "Extraction of Schottky Barrier Parameters for Pd/WSe_2/Au Vertical Diode," *The 23th Korean Conference on Semiconductors*, Feb. 2016.

21. Jun-Mo Park, In-Tak Cho, **Won-Mook Kang**, Byung-Gook Park, and Jong-Ho Lee, "Comparison of DC, Fast I-V, and Pulsed I-V measurement method in multi-layer WSe2 field effect transistors," *International Conference on Electronics, Information, and Communication (ICEIC)*, Jan. 2016.

22. In-Tak Cho**, Won-Mook Kang,** Jeongkyun Roh, Changhee Lee, and Jong-Ho Lee, "Temperature Effects on Current-Voltage and Low Frequency Noise Characteristics of Multilayer WSe_2 FETs," *International Symposium on the Physical and Failure Analysis of Integrated Circuits* (IPFA) Jun. 2015.

Patents

1. Jong-Ho Lee, **Won-Mook Kang**, "Synaptic devices and array."
   ▪ Korean Patent filed 10-2020-0019056, Feb. 2020

2. Jong-Ho Lee, Sung Yun Woo**, Won-Mook Kang**, "Neuromorphic System."
   ▪ Korean Patent filed 10-2018-0150912, Nov. 2018
   ▪ United States Patent filed US 16 205,478, Nov. 2018

Honors

1. Best Paper Award, The 28th Korean Conference on Semiconductors, Jan. 2021.

2. Gold Prize, The 24rd Humantech Thesis contest, Samsung Electronics, Feb. 2018.