Ph.D. DISSERTATION

Hybrid Training Method for Neuromorphic Hardware Based on Analog AND-Flash Arrays

아날로그 AND 플래시 어레이 기반의 뉴로모픽 하드웨어를 위한 하이브리드 학습 방법

by

DONGSEOK KWON

February 2023

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

Hybrid Training Method for Neuromorphic Hardware Based on Analog AND-Flash Arrays

아날로그 AND 플래시 어레이 기반의 뉴로모픽 하드웨어를 위한 하이브리드 학습 방법

지도교수 김 재 준

이 논문을 공학박사 학위논문으로 제출함

2023 년 2 월

서울대학교 대학원

전기정보공학부

권 동 석

권동석의 공학박사 학위논문을 인준함

2023 년 2 월

위 원 장 : 최 우 영 (인)

부위원장 : 김 재 준 (인)

위 원 : 김 재 하 (인)

위 원 : 이 종 호 (인)

위 원 : 배 종 호 (인)

Hybrid Training Method for Neuromorphic Hardware Based on Analog AND-Flash Arrays

by

Dongseok Kwon

Advisor: Jae-Joon Kim

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 2023

Doctoral Committee:

Professor Woo Young Choi, Chair
Professor Jae-Joon Kim, Vice-Chair
Professor Jaeha Kim
Professor Jong-Ho Lee
Assistant Professor Jong-Ho Bae

ABSTRACT

Neuromorphic computing systems using nonvolatile memory cells advance computational capability by performing large-scale vector-matrix multiplication operations in an analog manner. In addition, neuromorphic computing systems can reduce the inference time and energy consumption of neural network operations, thereby attracting much attention in various fields. Despite the advantages of neuromorphic systems, the conventional training methods show lower accuracy because of the nonideal characteristics of analog synaptic devices. In this work, we propose a new hybrid training method that trains the neuromorphic hardware very efficiently and accurately. The proposed training method does not use conductance tuning processes to accurately update the weight changes to the conductance of synaptic devices, significantly reducing the costs of online training in the hardware. We then experimentally show the high accuracy of the proposed method on the fabricated neuromorphic hardware: AND-type charge-trapping flash array. The AND-type flash array boosts a large-scale vector-matrix multiplication operation

using Kirchhoff's current law. Furthermore, the fabricated array has a nonvolatile memory function with a charge trapping layer ($SiO_2/Si_3N_4/SiO_2$), maintaining the multi-bit weight in a single synaptic device semi-permanently. We show that the accuracy of neuromorphic systems increases to that of the software-based neural network after 1-epoch hybrid training in the fabricated synaptic array. Moreover, the high performance of the proposed method was experimentally verified under various device nonideality conditions, indicating the proposed method can be generally applied to other types of synaptic devices. Our results show that neuromorphic systems using analog nonvolatile memory cells become a more promising platform for future artificial intelligence hardware.

Keywords: Hardware-based neural network, flash memory array, AND type crossbar array, online training, offline training, hybrid training

Student number: 2017-22213

CONTENTS

List of Figures

List of Tables

.

Chapter 1
Introduction

1.1 Neuromorphic computing

Recently, artificial neural networks (ANNs) have advanced human lives in various applications, including image processing [1-5], natural language processing [6-10], and autonomous driving [11-12]. Convolutional neural networks (CNNs), of which structure is greatly influenced by biological vision systems [13, 14], have achieved human-level or superior accuracy in vision applications. However, state-of-the-art training algorithm techniques for ANNs have been developed to enlarge the network size, significantly increasing the computational complexity in network operations [15]. From this perspective, 'von Neumann bottleneck' between the off-chip memory and processing units decreases the computational efficiency of conventional von Neumann computing systems. In order to address the severe issues of conventional computing systems, neuromorphic systems have been proposed to exhibit low energy consumption, parallel computing, and low system

latency [16-22].

1.2 Synaptic Devices

Neuromorphic computing systems mainly consist of synaptic devices that represent weights in neural networks. In particular, nonvolatile synaptic devices can store multi-bit weights with long-term memory functionality, increasing the density of devices compared to digital memory. Moreover, according to Kirchhoff's current law, vector-matrix multiplication (VMM) operations are performed with high parallelism by the sum of the synaptic currents of the devices, reducing the data movement between the memory and processing units. In this regard, there are many studies on synaptic devices such as resistive random-access memory (RRAM) [23-30], phase-change RAM (PCRAM) [31-34], 3-terminal ferroelectric field-effect-transistors (FeFETs) [35-36], charge-trap flash [37-40]. Among them, the charge-trap flash memory device has drawn much attention because of the compatibility of the CMOS fabrication process, good reliability, and massive production capability. The feasibility of 3-dimensional stacking is also an important advantage of the

charge-trap flash memory cells.

The array architectures of the flash memory cells can be configured depending on the purpose of the applications. The array architectures of flash memory cells are mainly categorized as NAND- [41, 42], NOR- [40, 43], and AND-type array architectures [44, 45]. In NAND-type array architecture, it is difficult to perform large-scale VMM operations in parallel because of the cell string structure, thus the array architecture is appropriate in high-density memory applications. On the other hand, NOR- and AND-type array architectures have the advantage of parallel computing with the form of the crossbar of word lines and bit lines. The difference between the NOR- and AND-type array architectures is the configuration of source lines and bit lines (Fig. 1.1), resulting in the difference in the selective write operations. In particular, the selective write operations can be performed in AND-type array architecture using Fowler Nordheim (FN) tunneling operations, in which a small tunneling current flows compared to the on-current of the flash cell. Therefore, for online training that requires a large number of write operations, AND-type array architecture can save energy consumption during the training

operations compared to the NOR-type array architecture.

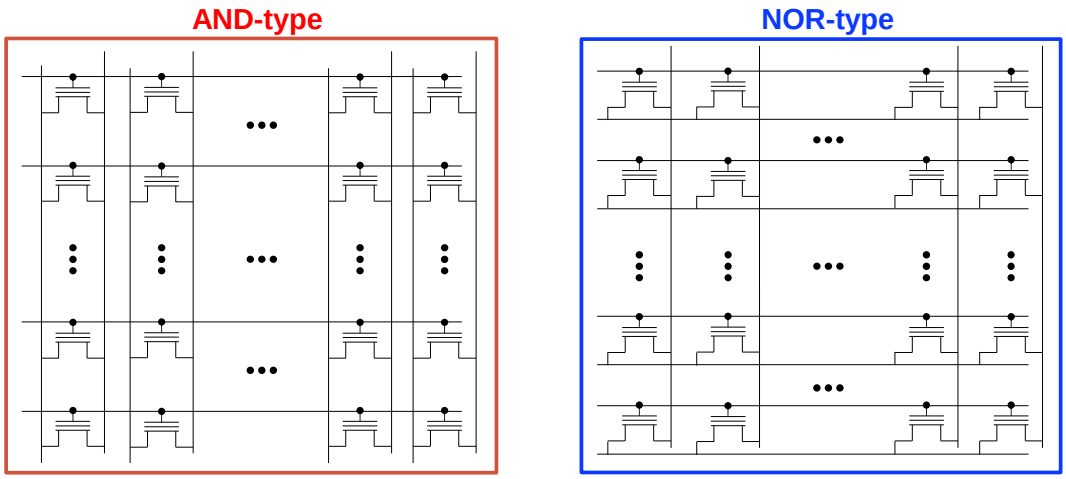


Fig. 1.1. Schematic diagrams of AND- and NOR-type flash array architectures.

1.3 Training Algorithms for Neuromorphic Systems

From the perspective of the training algorithm, the majority of neuromorphic hardware research uses offline (*ex-situ*) training with conductance tuning processes, in which the conductance of synaptic devices is iteratively adjusted to the target value [46-47]. Because the conductance of synaptic devices can represent the offline-trained weights by the tuning process, time-static nonidealities of analog synaptic devices (ex: nonlinearity, device-to-device variation, limited on/off ratio)

can be greatly addressed. However, the offline training methods need to verify that the conductance of the devices reaches the target value, which requires substantial energy and time consumption. Additionally, time-varying device nonidealities (ex: read fluctuation, conductance drift, retention) cannot be addressed in the offline training methods. In contrast, online (*in-situ*) training methods can address time-varying nonidealities because the neuromorphic systems can be trained in real time using in-situ training data. However, the online training methods show poor accuracy because the time-static nonidealities cannot be effectively addressed without the conductance tuning process [17, 48-51]. More importantly, although there are many studies on online training methods for neuromorphic computing systems, they lack experimental demonstration in implemented hardware owing to high training costs to account for weight changes in the synaptic devices. Therefore, a novel training method is necessary to efficiently train the neuromorphic computing systems with high accuracy.

1.4 Purpose of research

This dissertation proposes a new hybrid training method for neuromorphic systems that efficiently perform large-scale VMM operations with analog nonvolatile memory cells. First, we fabricated the AND-type array architecture of the flash memory cells, which exhibits the capability of parallel computing and energy-efficient write operations. We then characterized the synaptic characteristics of the AND-type flash array, including nonlinearity, device-to-device variation, endurance, retention, and dynamic range. Additionally, the synaptic characteristics of the array were optimized for high accuracy of online training.

Next, we experimentally demonstrated the proposed hybrid training method in the fabricated AND-type flash array. Since the hybrid training method adopts both hardware (in-situ) and software (ex-situ) training, the neuromorphic system adjusts the conductance of synaptic devices automatically on the chip while showing the performance of software-based neural networks. More importantly, the proposed method does not use the conductance tuning process that requires substantial communication cost and time to update all weights in the neuromorphic systems.

This property of the proposed method reduces the training cost and enhances the training efficiency of the neuromorphic systems. Finally, we evaluated the accuracy of the proposed method under various nonideal conditions of synaptic devices, verifying that the proposed method can be applied to other neuromorphic systems with various types of synaptic devices. Our successful results will significantly advance the neuromorphic computing systems into a promising hardware platform for artificial intelligence.

1.5 Dissertation outline

The dissertation outline is as follows. Chapter 1 provides an overview of neuromorphic systems, synaptic devices, and training methods. It also covers the contents of the synaptic devices composing the synaptic array based on recent research trends. Chapter 2 describes the AND flash memory array architecture and the measurement results. This chapter includes the device structure, fabrication process steps, analysis, and optimization process of the array operations. Chapter 3 deals with the proposed hybrid training method. This chapter also includes the

measurement results of the training method in the fabricated synaptic array and the comparison of other reported training methods. Finally, chapter 4 concludes this dissertation with a summary.

Chapter 2
AND-type Flash Array Architecture

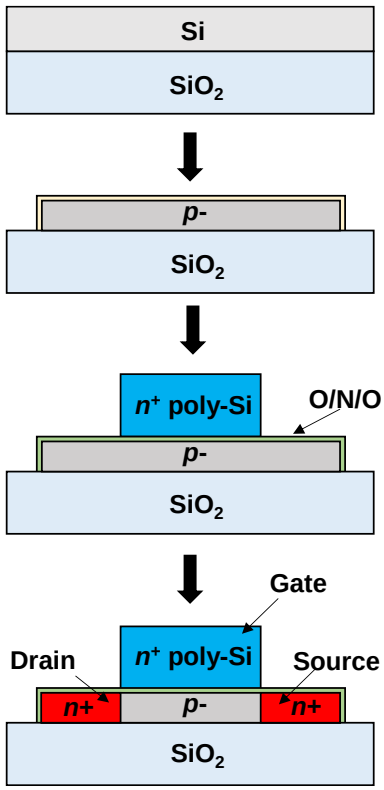
2.1 Array Fabrication and Operation

In the AND-type array architecture, the source-lines (SLs) and bit-lines (BLs) are configured in parallel. Owing to the parallel configuration of SLs and BLs, the channel potential of flash memory cells can be easily modulated without on-current flowing. Therefore, the energy consumption in the selective program/erase (PGM/ERS) operations is significantly decreased. The fabrication process of the AND-type flash array with 3-terminal flash memory cells is shown in Fig. 2.1. The entire fabrication process was conducted on a 6-inch silicon-on-insulator (SOI) wafer with CMOS fabrication process technology. First, Si active layer was

patterned at a thickness of 100 nm, and the implantation process was conducted to form the *p*-body with the boron ions. The doping concentration of the *p*-body was 1×10^{18} cm^{-3}. Subsequently, a gate stack of a tunneling SiO_2 (3 nm), a charge storage layer Si_3N_4 (6 nm), and a blocking SiO_2 (9 nm) was deposited on the Si active layer. Then, a gate of n^+ poly-Si was deposited and patterned. After that, the implantation process was conducted to form a source and drain with the arsenic ions, followed by rapid thermal processing at 1000 °C for 10 s to activate the source and drain. After depositing a passivation oxide with tetraethyl orthosilicate (TEOS) of 300-nm-thickness, contact holes for the gate, source, and drain were defined and etched. Then, a Ti/TiN/Al/TiN (30 nm/30 nm/300 nm/30 nm) stack was deposited for the metal layer.

Fig. 2.2 (a) and (b) show a schematic diagram of the fabricated AND-type flash array and a scanning electron microscopy (SEM) image of the array with the size of 25-word lines (WLs) × 4 BLs (4 SLs). The width/length of the fabricated flash devices in the array is 1 μm/1 μm, respectively. Fig. 2.2 (c) describes the process of the VMM operations in the AND-type array according to Kirchhoff's current law.

Note that the flash device can modulate its conductance by adjusting the threshold voltage of the device with PGM or ERS pulses. When PGM (ERS) pulse is applied to the gate at $V_D = V_S = 0$ V, the electrons (holes) from the channel (body) are injected into the nitride layer and trapped. Depending on the stored charge in the nitride layer, the vertical electric field is modulated, leading to changes in the threshold voltages of the flash devices. In neuromorphic computing systems, the conductance of the flash device represents a weight. Assuming each flash device in the array has its own conductance depending on the trained weights, the voltage inputs are applied to the WLs of the array. Then, the current flows through each flash device depending on its conductance, and the currents are summed along the SLs and BLs according to Kirchhoff's current law. The sum of the currents represents a weighted sum in neuromorphic systems. As such, the VMM operations are efficiently performed in the memory domain in an analog manner.



- **6 inch SOI wafer**
- **Si active patterning**
- **n^+ poly-Si deposition**
- **Gate patterning**
- **O/N/O stack formation**
- **Source/drain implantation & activation**
- **Back-end process**

Fig. 2.1. Key fabrication process step of the 3-terminal flash device on a 6-inch SOI wafer.

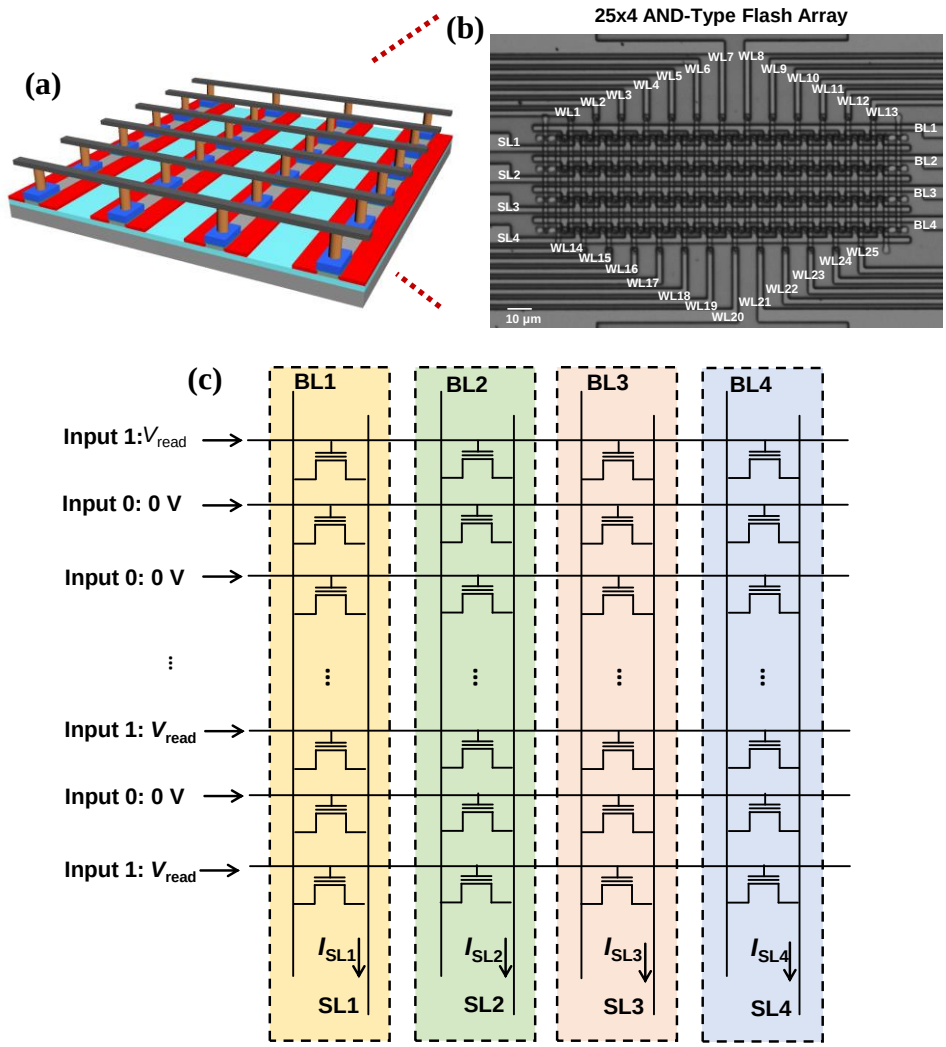


Fig. 2.2. (a) Schematic diagram of the fabricated AND-type flash array architecture. (b) Scanning electron microscopy (SEM) images of the fabricated AND-type flash array (25×4). (c) Schematic diagram of VMM operation in the AND-type array architecture.

2.2 Device and Array Characterization

In this section, we characterized the fabricated AND-type flash array for synaptic devices. We investigated the on-current of the devices in the array, as shown in Fig. 2.3 (a). We confirmed that there are no devices at the stuck-off condition, and device-to-device variation of the on-current exists in the array. Besides, there is no remarkable line dependency in the on-current of devices. The drain current (I_D)-gate voltage (V_G) characteristics of the devices in the array were investigated, as shown in Fig. 2.3. (b). The device-to-device variation in the threshold voltages of the flash devices is exhibited in the fabricated flash array.

The measurement results of the long-term potentiation (LTP) and depression (LTD) curves of the devices are shown in Fig. 2.4 (a) and (b), respectively. When the PGM pulses are applied to the gate, the electrons are trapped in the Si_3N_4 layer from the channel, reducing the conductance of the device. In contrast, when the ERS pulses are applied to the gate, the holes are trapped in the Si_3N_4 layer from the body, increasing the conductance of the device. The PGM and ERS pulse conditions are V_{GS} of 8 V, 10 μs and -8 V, 10 ms, respectively, at a V_{DS} of 0 V. As shown in Fig.

2.4, the device-to-device variation is also shown in LTP and LTD curves. In addition, the LTP and LTD curves are nonlinear with respect to the same ERS and PGM pulses. This is because the stored charge in the Si_3N_4 layer can reduce the electric field at the PGM or ERS operations and reduce the tunneling current in the PGM and ERS pulses. Note that the nonlinearity of the LTP and LTD curves causes errors in the VMM operations and weight updates, degrading the overall accuracy of the online training. In order to mitigate the nonlinearity in the LTP and LTD curves, conductance tuning processes were reported, where the conductance is iteratively tuned by PGM and ERS pulses to reach the target value [46, 47]. However, the processes require substantial energy and time consumption to accurately adjust the conductance, thereby degrading the training efficiency and speed. From this perspective, it is necessary to develop a new training algorithm with high accuracy even if the nonlinear LTP and LTD curves are used.

Fig. 2.4 (c) and (d) show the device-to-device variation in the LTP and LTD curves, respectively, particularly the max/min ratio of I_{DS}s and nonlinearity. The measured σ/μ is about 40%. Compared to the device-to-device variation, the

variation in the single curve is relatively small. Fig. 2.4 (e) shows the sequentially measured five times with the same PGM and ERS pulses to investigate the cycle-to-cycle variation in a single flash device. We confirmed that almost the same LTP and LTD curves are exhibited in the repetition, meaning that the fabricated flash device shows low cycle-to-cycle variation.

We investigated the LTP and LTD curves as parameters of PGM and ERS pulse amplitude, as shown in Fig. 2.5 (a) and (b). The LTP and LTD curves were normalized to compare the nonlinearity of the responses. Both LTP and LTD curves become more nonlinear as the pulse amplitude increases. The fitting results of the LTP and LTD curves of flash devices are shown in Fig. 2.6 using the logarithmic conductance response model [52, 53]. Fig. 2.7 verifies selective PGM/ERS operations in the fabricated AND-type flash array. Fig. 2.7 (a) shows the schematic diagram of the selective PGM/ERS operations in the AND-type array architecture. In this scheme, the conductance of cell 1 should be updated, and others should be inhibited. V_{sel} and 0 V are applied to the WL of the selected (cell 1) and inhibited cells (cell 2-4), respectively. 0 V is applied to the BLs and SLs of the selected cells,

and V_{inh} is applied to the BLs and SLs of the inhibited cells. In this scheme, the drain voltage and the source voltage are the same to cut off the on-current in the cells. As shown in Fig. 2.7 (b) and (c), only the conductance of cell 1 is updated by the PGM and ERS pulses, and the conductance of others is inhibited successfully. Note that in online training where the weights are updated in neuromorphic systems, the selective write operations without on-current flowing are essential for low-power operations.

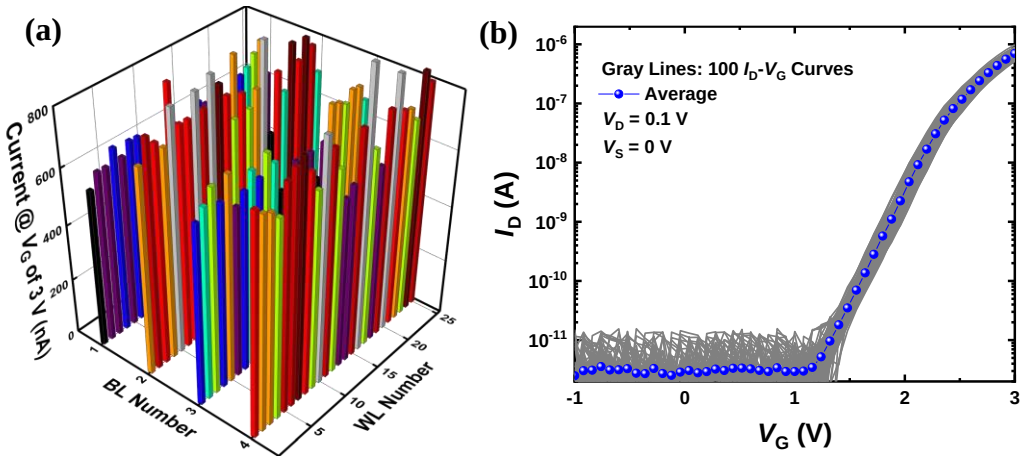


Fig. 2.3. (a) Drain currents (I_Ds) of all devices in the array at a gate voltage (V_G) of 3 V and drain voltage (V_D) of 0.1 V. (b) I_D-V_G curves of all devices in the array and the averaged curve at a V_D of 0.1 V. The device-to-device variation exists in the fabricated synaptic array.

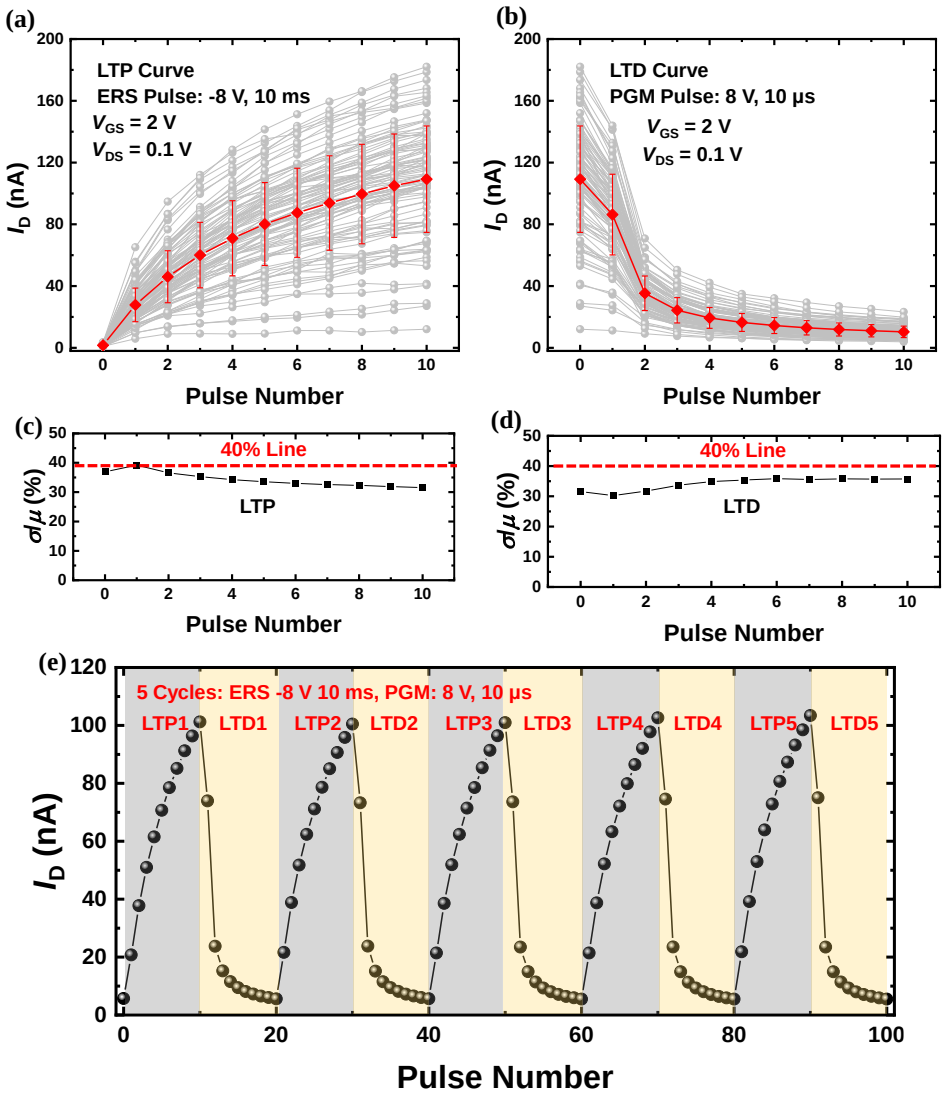


Fig. 2.4. Measured conductance responses of 100 devices in the array for (a) LTP and (b) LTD. ERS pulses (-8 V, 10 ms) and PGM pulses (8 V, 10 µs) are applied to the gate of devices for the LTP and LTD curves, respectively, at a $V_S = V_D = 0$ V. A read bias (V_{GS}= 2 V and V_{DS}= 0.1 V) is applied to the device immediately after

every ERS (or PGM) pulse. The gray lines represent LTP and LTD curves for each device, and the red line represents the averaged values of 100 devices. The error bar indicates the standard deviation value of each point. Standard deviation over mean (σ/μ) values of the (c) 100 LTP and (d) 100 LTD curves. The maximum value of σ/μ is ~40 %. (e) Repeated conductance responses at the ERS/PGM conditions.

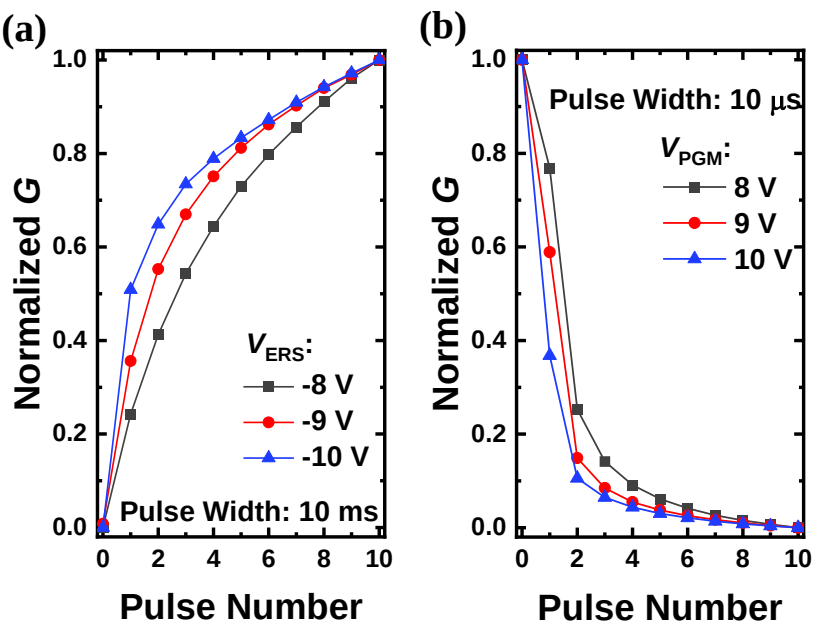


Fig. 2.5. Conductance responses as a parameter of (a) ERS (for LTP) and (b) PGM (for LTD) pulse amplitude. The pulse widths for the ERS and PGM are 10 ms and 10 μs, respectively.

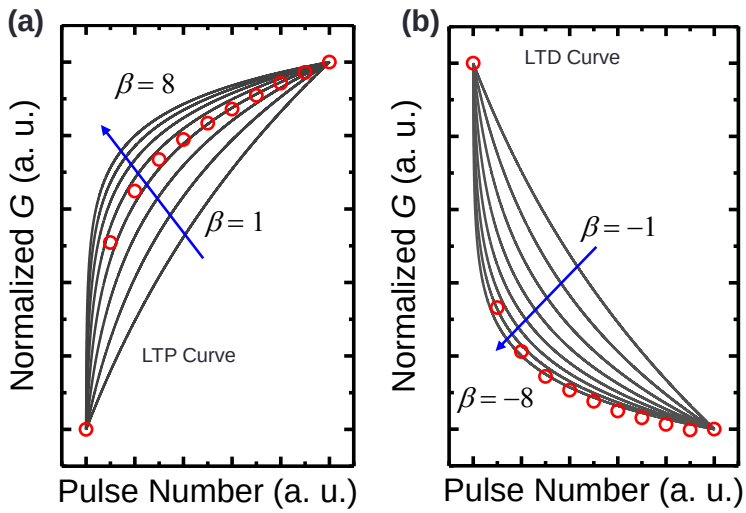


Fig. 2.6. Fitting results of the measured (a) LTP and (b) LTD curves using the logarithmic conductance response model.

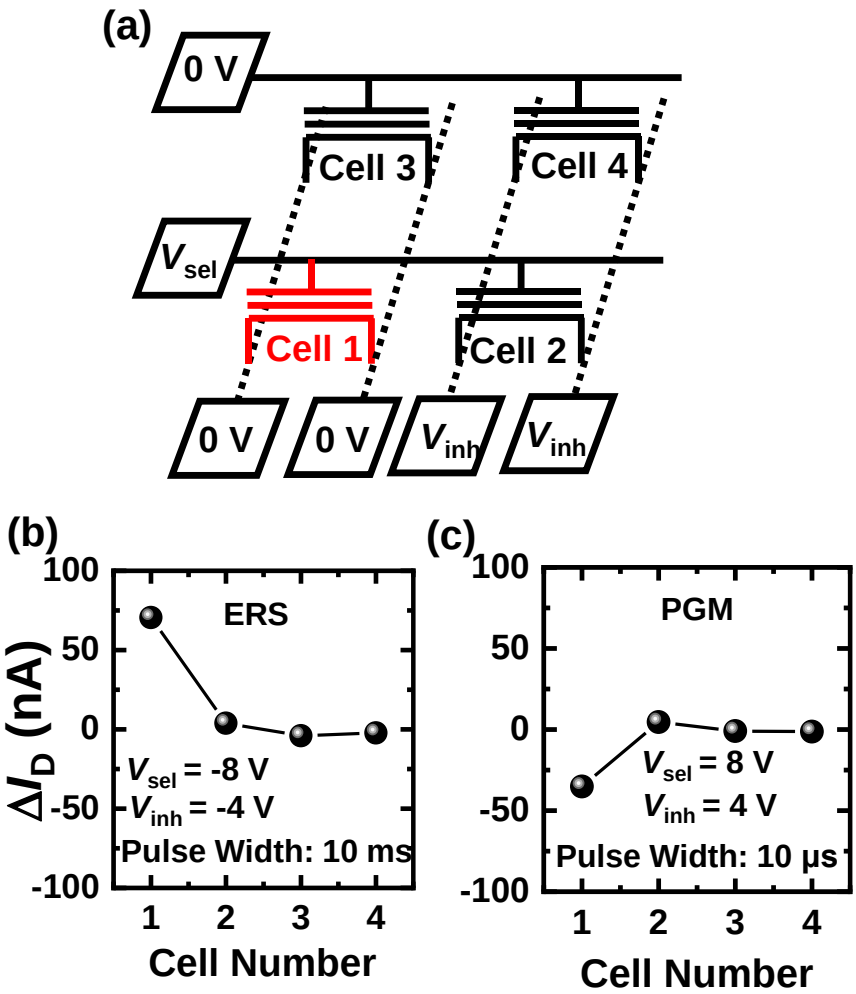


Fig. 2.7. (a) Selective conductance update scheme in the AND-type flash array architecture. Cell 1 is the target device to selectively update its conductance, and the other cells should be inhibited. Measured I_D changes in the array at the given (b) ERS and (c) PGM conditions. The I_D of Cell 1 is updated, and the I_Ds of other cells are not updated.

2.3 Optimization of Device Nonidealities

We analyze and optimize the PGM and ERS conditions in the fabricated flash array for better synaptic characteristics. Fig. 2.8 (a) and (b) show the variation in the threshold voltages (V_{th}s) and I_{DS}s of 100 devices in the flash array, respectively. The variation in the V_{th}s of flash devices increases as the ERS pulse amplitude increases. On the other hand, the variation in the I_{DS}s decreases as the ERS pulse amplitude increases. The variation in the V_{th}s is 0.0366 V at a -8 V pulse, which is relatively small compared to other ERS pulse amplitude. However, the operation regions of the devices at a -8 V ERS pulse are mainly around the subthreshold region, in which the I_{DS}s are changed exponentially by the V_{th} variations. Thus, the relative variation in the I_{DS}s becomes larger than other ERS pulse amplitudes. In terms of linearity, the ERS pulse amplitude of -8 V is more advantageous than other pulse amplitudes. As shown in Fig. 2.9, the LTP curves at a -8 V ERS pulse are more linear than those at -9 V and -10 V ERS pulses. The amount of charge stored in the nitride layer logarithmically increases with the number of pulses [54, 55]. Thus, the exponential relationship between I_{DS} in the subthreshold region and the

V_{th} changes can be canceled out at an ERS of -8 V, resulting in more linear LTP curves. Note that the device-to-device variation can be mitigated in online training, but the nonlinearity significantly degrades the accuracy of online training [48-51]. Furthermore, the low currents in the LTP curves at a -8 V pulse are advantageous for low-power neuromorphic systems.

The retention characteristics of the fabricated flash device are shown in Fig. 2.9 as a parameter of pulse amplitude. The measurements were conducted at room temperature. The V_{th} changes ($|\Delta V_{th}|$) after a retention time of 10^4s are less than 0.2 V, even at the PGM and ERS pulse amplitudes of 10 V. These characteristics indicate that the nonvolatile memory functionality is successfully implemented in the fabricated flash devices with the charge trap layer (Si_3N_4). In particular, when -8 V and 8 V of ERS and PGM pulses are used, the $|\Delta V_{th}|$ is less than 0.02 V after a retention time of 10^4 s. The endurance characteristics of the flash device are shown in Fig. 2.10 as a parameter of PGM and ERS pulse amplitude. Since only a small part of the full memory window is used in the given PGM and ERS pulse conditions, the device degradation is not significantly exhibited in the PGM and ERS cycling

test. This result means that the fabricated flash devices are advantageous for online training of neuromorphic systems, in which many conductance changes are required to update weights.

The comprehensive synaptic characteristics of the fabricated flash device are shown in Table 2.1 as a parameter of PGM and ERS pulse amplitude. The comparison was performed in terms of variation, nonlinearity, retention, endurance, and dynamic range. Compared to other types of synaptic devices, our flash device shows superior synaptic characteristics. In particular, our flash device shows the optimized synaptic characteristics under PGM and ERS pulse amplitudes 8 V and -8 V of condition.

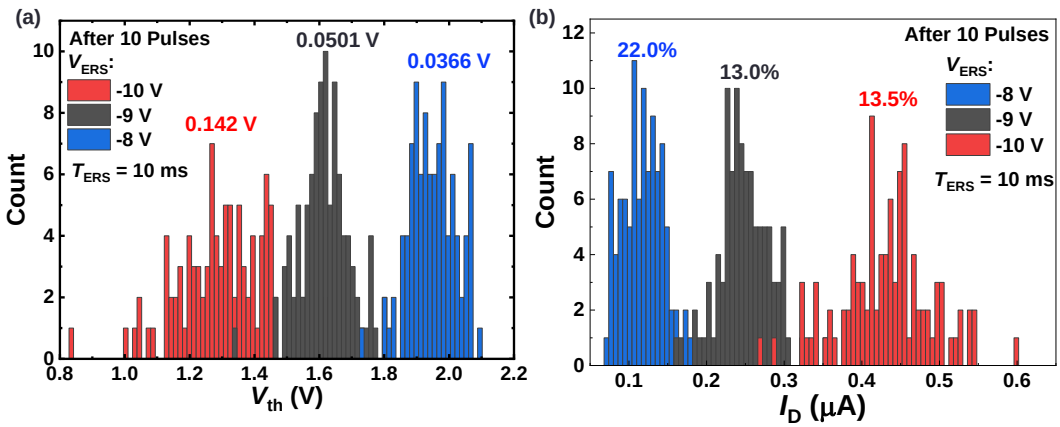


Fig. 2.8. Variation in (a) V_{th}s and (b) I_{DS}s of 100 devices in the fabricated AND-type flash array after ten ERS pulses of 10 ms pulse width. The IDs are measured at V_{GS} of 2.0 V and V_{DS} of 0.1 V. The inner numbers indicate the standard deviation of each distribution.

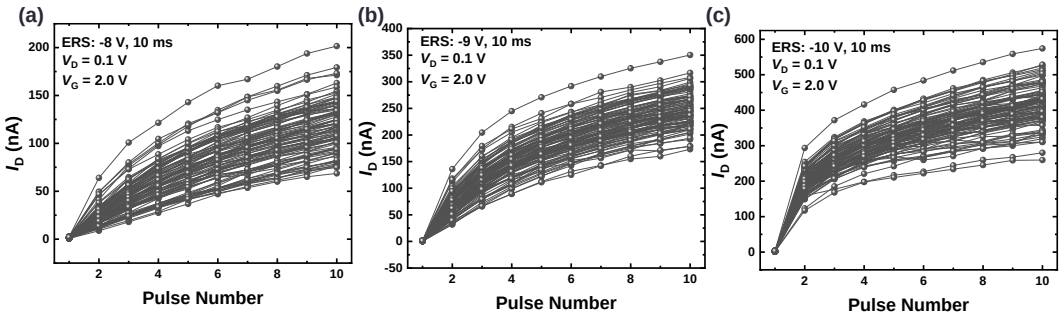


Fig. 2.9. Conductance responses to the ERS pulses with amplitudes of (a) -8 V, (b) -9 V, and (c) -10 V. The LTP curves are measured at V_{GS} of 2.0 V and V_{DS} of 0.1 V.

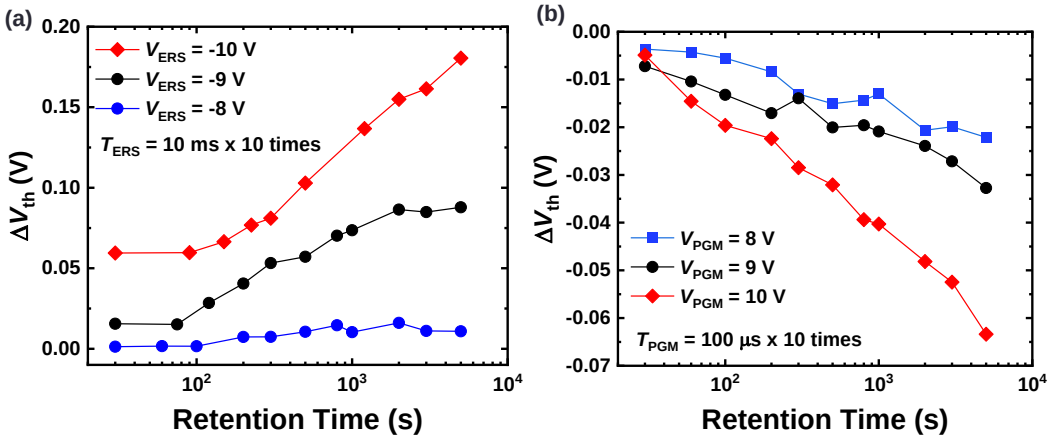


Fig. 2.10. Retention characteristics of the fabricated flash device in (a) ERS and (b) PGM state. Each retention characteristic was measured after ten pulses were applied to the device.

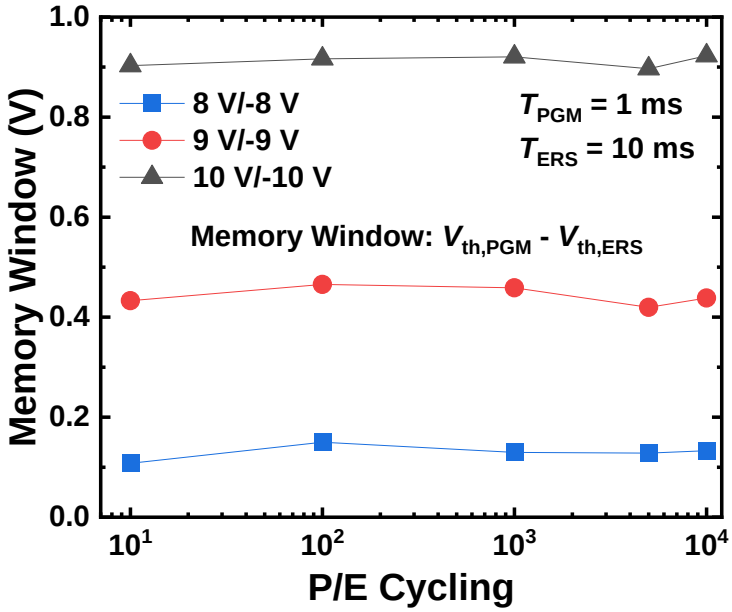


Fig. 2.11. Endurance characteristics of the fabricated flash device as a parameter of PGM and ERS pulse amplitude.

TABLE 2.1
COMPARISON OF SYNAPTIC CHARACTERISTICS WITH OTHER TYPES OF SYNAPTIC DEVICES

Other works	Nonlinearity	Variation	Retention (s)	Endurance (Cycles)	Dynamic Range
S. H. Jo et al. [56]	2.4	X	X	X	12.5
L. Gao et al. [57]	1.85	X	X	X	2
S. Park et al. [58]	3.68	X	X	X	6.84
J. Woo et al. [59]	1.94	X	X	X	4.43
W. Chung et al. [60]	1.22	X	X	$>10^6$	>100
S. Suh et al. [61]	0.96	9.76 %	$>10^3$	$>10^6$	>10
M.-K. Kim et al. [62]	0.80	3.93%	> X	$>10^4$	>10
This work (10 V)	4.54	13.5%	$>10^4$	$>10^5$	190
This work (9 V)	2.52	13.0%	$>10^6$	$>10^5$	280
This work (8 V)	1.33	22.5%	$>10^8$	$>10^5$	82

Chapter 3
Hybrid Training Method

3.1 Offline Training

Many neuromorphic systems using analog nonvolatile synaptic devices have adopted quantized neural networks (QNNs) because the bit precision of weights and activations can be significantly reduced [63-67]. In the training process of the QNNs, high-precision weights (w_cs) are updated first by the activation value (a) and delta value (δ), which are obtained with the quantized weights (w_qs). The relationship between w_cs and w_qs is non-differential; thus, a straight-through estimator (STE) is used for QNN training, which approximates the non-differential quantizing function as a differential function. Then, the updated w_qs are obtained with the updated w_cs. Through this weight update rule, QNNs achieve very high accuracy compared to software-based neural networks with full-precision weights and activations, although the inference of QNNs is performed with low-precision weights and activations. The training process of QNNs with a linear quantization

function is represented in Algorithm 1. The weight updates are performed with ADAM optimizer in the Pytorch framework. Adding a batch normalization layer can improve the training efficiency of QNNs, but we do not use the techniques to concentrate on the effects of the proposed training method.

Algorithm 1. Offline training process. L-layer network, quantization function Q, initialized high-precision weight W_c, quantized weight W_q, quantized activation function f_q, gradient g.

Requirements: a minibatch of inputs and targets (a^0, y), learning rate γ, initialized W_c.

for l = 1 **to** L **do** // Forward propagation

$W_q^l \leftarrow Q(W_c^l)$

$s^l \leftarrow W_q^l a^{l-1}$

if $l < L$ **then**

$a^l \leftarrow f_q(s^l)$

end if

end for

Compute the gradient in layer L, $g(a^L)$, knowing a^L and y.

for l = L **to** 1 **do** //Backward propagation

$g(a^{l-1}) \leftarrow g(s^l)W_q^l$

$g(W_q^l) \leftarrow g(s^l)^T a^{l-1}$

end for

for l = 1 **to** L **do**

$g(W_c^l) \leftarrow g(W_q^l)\frac{\partial w_q^l}{\partial w_c^l}$ // STE, $\frac{\partial w_q^l}{\partial w_c^l} = 1$

$W_c^l \leftarrow Update(W_c^l, g(W_c^l), \gamma)$

end for

3.2 Online Training

Motivated by the QNN training method, we propose an online training method using the w_cs and STE, in which the weight updates are performed in neuromorphic systems. Thanks to the use of w_cs and STE, the QNN training method shows high accuracy even with limited network conditions, such as 1-bit weights and 1-bit activations in binary neural networks. From this point of view, we expect that the use of the w_cs and STE in online training also significantly mitigates the accuracy degradation by nonidealities of the analog devices. The detailed process of the proposed online training is explained in Algorithm 2.

The weights in the synaptic array (w_{array}s) are represented with the conductance of synaptic devices as $w_{array} = \alpha G - 0.5$, where α is the normalization value to normalize the dynamic range of the synaptic devices within a range of [0, 1], and G is the conductance of the synaptic device. The value of 0.5 is subtracted to represent a negative weight [34]. At the beginning of the online training, we assume that w_cs are initialized and transferred to w_{array}s by modulating the number of ERS pulses. In forward propagation, VMM operations are performed in the synaptic array by

the current sum, and the activation function (a linearly quantized ReLU function) is applied to the results of the current sum. Then, the activation values are applied to the following synaptic array until the last layer that classifies the images. In backward propagation, VMM operations are performed with the readout w_{array}s in the software, and the gradient of w_c is calculated with the gradient of w_{array} and STE. Subsequently, the w_cs are updated with the gradient of w_c by the ADAM optimizer. Then, the pulse number (*PN*) matrix, in which w_cs are rounded to have *n* levels (*n*: the number of conductance of synaptic devices), is calculated with the updated w_cs. The number of levels can be modulated with the number of conductance levels in the LTP and LTD curves. After that, a single PGM or ERS pulse is applied to the synaptic device, whose corresponding *PN* is updated, resulting in the conductance changes in the synaptic array (i.e., the changes in the w_{array}). Note that the proposed online training does not adopt the conductance tuning process to mitigate the nonlinear weight updates in synaptic devices. The tuning process requires increased training costs and peripheral circuits, increasing the hardware burden for online training.

Algorithm 2. Online training process using synapse array. L-layer network, high-precision weight W_c, weight in array W_{array}, quantized activation function f_q, gradient g, applying write pulses to the synaptic array *ApplyingPulse*, rounding function R.

Requirements: a minibatch of inputs and targets (a^0, y), learning rate γ

for l = 1 **to** L **do** // Forward propagation

$s^l \leftarrow W_{\text{array}}^l a^{l-1}$ // VMM using weights in array

if $l < L$ **then**

$a^l \leftarrow f_q(s^l)$

end if

end for

Compute the gradient in layer L, $g(a^L)$, knowing a^L and y.

for l = L **to** 1 **do** //Backward propagation

$g(a^{l-1}) \leftarrow g(s^l) W_{array}^l$ //VMM using weights in array

$g(W_{array}^l) \leftarrow g(s^l)^T a^{l-1}$

end for

for l = 1 **to** L **do**

$g(W_c^l) \leftarrow g\left(W_{\text{array}}^l\right) \frac{\partial W_{\text{array}}^l}{\partial W_c^l}$ // STE, $\frac{\partial W_{\text{array}}^l}{\partial W_c^l} = 1$

$PN_0^l \leftarrow R(W_c^l)$

$W_c^l \leftarrow Update(W_c^l, g(W_c^l), \gamma)$

$PN_1^l \leftarrow R(W_c^l)$

$W_{array}^l \leftarrow ApplyingPulse(PN_1^l, PN_0^l)$ //Applying PGM or ERS pulse to array

end for

3.3 Hybrid Training

We propose a hybrid training method that combines the offline and online training methods. The above offline and online training methods use the w_cs and STE during the training processes, and we connect the w_cs between the training processes. In other words, the w_cs in the online training at the beginning are replaced by the offline trained w_cs that show high accuracy. Then, the trained w_cs are transferred to the conductance of synaptic devices in neuromorphic systems. The online training is additionally performed to train the systems using *in-situ* VMM data, as represented in Algorithm 2. As a result, the neuromorphic systems can achieve high accuracy faster than when only the online training is conducted from the beginning. Moreover, the number of weight updates can be significantly decreased because the trained weights are transferred, reducing the cost of the online training in the neuromorphic systems.

The process of the proposed hybrid training method is as follows: pre-training (offline training), weight transfer, and online training. First, a neural network is designed (for example, a 5-layer CNN with two convolutional layers and three

fully-connected layers for the MNIST image classification task (Fig. 3.1 (a)). In the pre-training step, the neural network is offline trained by the QNN training method (Algorithm 1) in the PyTorch framework using the cross-entropy loss function. The number of conductance levels in synaptic devices determines the weight precision in the training process. In the weight transfer step, the conductance of all synaptic devices is initialized to the minimum conductance by applying a long PGM pulse to the synaptic array. Then, a *PN* matrix is calculated with the pre-trained w_cs with the *n* levels, which is the same as the number of conductance in LTP and LTD curves of flash synaptic devices (Fig. 3.1 (b)). Subsequently, ERS pulses are applied column by column to the synaptic devices, as many as the corresponding *PN*, as represented in Fig. 3.1 (c) and (d). Note that no conductance tuning process is used. As a result of applying ERS pulses to the devices, the pre-trained w_cs are transferred to the conductance of synaptic devices, and w_{array}s are obtained with the conductance. In the online training step, the weight updates are conducted in the neuromorphic systems with the in-*situ* training data, as represented in Algorithm 2. After the weight transfer step, the w_{array}s are distributed around the w_cs because

analog synaptic devices have device nonidealities such as nonlinearity and device variation, which degrades the accuracy of neuromorphic systems. In this case, the additional online training can improve the accuracy of the systems with the *in-situ* training data, which reflects the device nonidealities (Fig. 3.1 (e)). Using the proposed hybrid training method, neuromorphic systems can achieve high accuracy compared to the accuracy of software-based neural networks while reflecting the device nonidealities as well as the hardware imperfections (ex: wire resistance, parasitic capacitance, external noise, etc.).

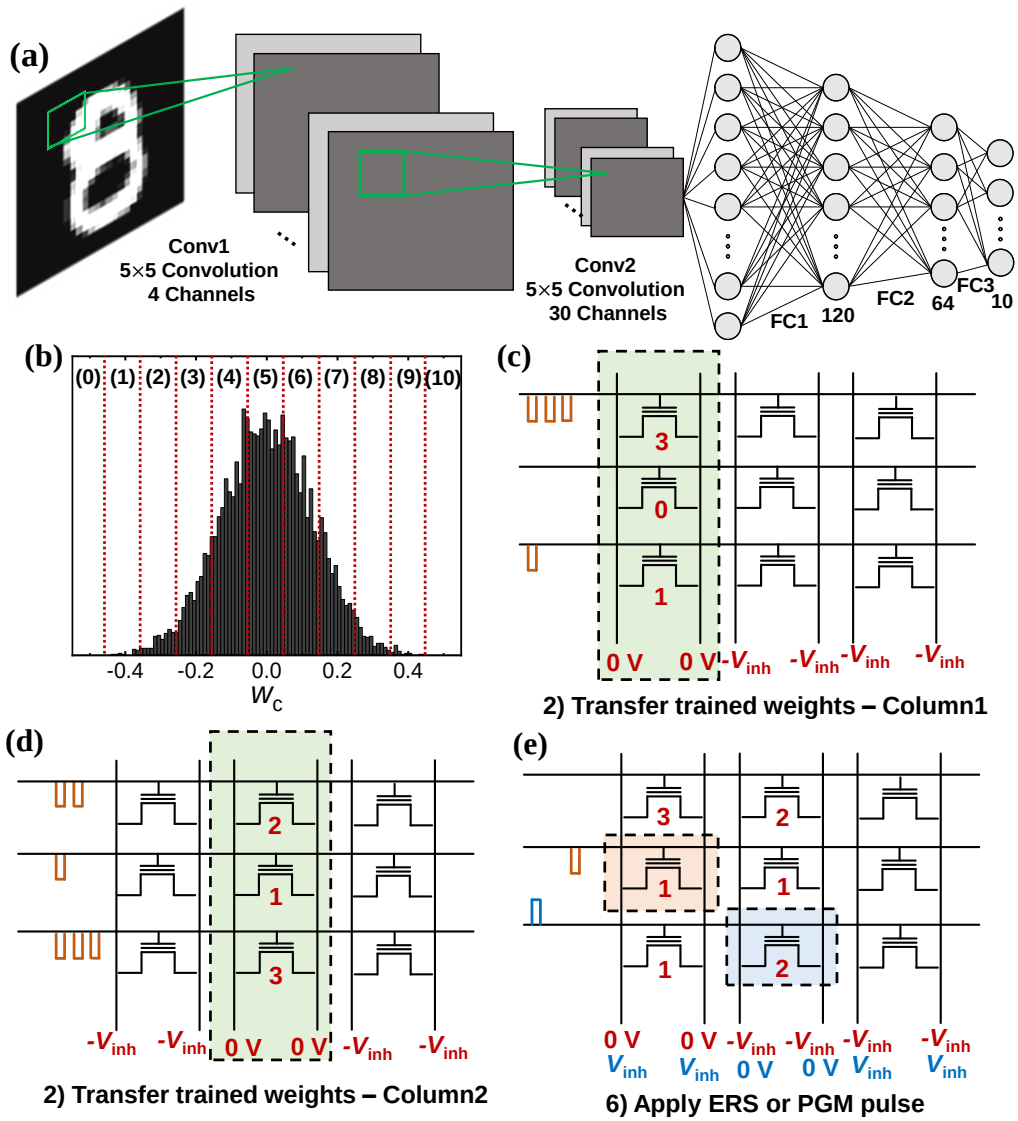


Fig. 3.1. (a) Structure of a CNN with two convolutional layers and three fully connected layers. (b) Example of the distribution of w_cs pre-trained by QNN training method. The w_c distribution is divided into the number of conductance levels, and the w_c is transferred to the conductance of the device in the array by

applying the ERS pulses. (c, d) Schematic diagrams of the weight transfer step. The weight transfer process is performed column-by-column. The voltage of 0 V is applied to SL and BL of the selected column, and $-V_{inh}$ (voltage for inhibited line) is applied to the SLs and BLs of the unselected column. The bias of V_{inh} is half of the V_{sel} (voltage for selected line). (e) Schematic diagram of applying update pulses in online training. ERS pulses are applied to the devices in which the pulse number increases during the training. PGM pulses are applied to the devices in which the pulse number decreases.

3.4 Demonstration of Hybrid Training in Hardware

We apply the proposed hybrid training method to the fabricated AND-type flash array. The 5-layer CNN is designed for MNIST image classification, as shown in Fig. 3.1 (a). The input image is binarized, and the activation function of the hidden layers is the ReLU function linearly quantized to 8-bit. Fig. 3.2 shows the classification accuracy of the CNN as a parameter of weight level. The CNN is trained in the software using the QNN training method (Algorithm 1). The accuracy

of the CNN increases as the weight level increases because more precise weights can extract the features of the input images better. The offline trained CNN achieves an accuracy of 99.0 % with 11-level weights.

In the weight transfer step, the weights in convolution layer 1 (Conv1 in Fig. 2.3 (a), 5×5 kernels, 4 channels) are transferred to the conductance of the fabricated synaptic array (25 WLs × 4 BLs). The w_{array}s in the other layers are modeled in the software as w_qs with a Gaussian distribution function for variation in the array (σ/μ = 40 %). In the online training step, the flattened input images are applied to the WLs of the array, and the VMM operations in the Conv1 layer are performed using the current sum in the fabricated synaptic array. The VMM operations in other layers are performed in software using the modeled w_{array}s to which the device variation is applied. Then, the backward propagation is performed in the software using the w_{array}s of the fabricated synaptic array and the modeled w_{array}s to calculate Δw_c and update the *PN*. According to the updates in the *PN*, the PGM and ERS pulses are applied to the corresponding synaptic device. In this work, only the Conv1 layer is trained to focus on the effects of the hybrid training method on the

implemented synaptic array, and the w_{array}s in other layers have fixed values with the variation during the online training process. If larger synaptic arrays are implemented to cover all layers of the network, the accuracy of the network can be further improved.

Table 3.1 compares the accuracy of MNIST image classification before the online training step is performed. If the offline trained weights in the Conv1 layer are used as they are, an accuracy of 97.8% is obtained even with the variation (σ/μ = 40 %) applied to other layers. On the other hand, after the weight transfer step of the Conv1 layer is performed to the fabricated AND-type synaptic array, the accuracy is significantly decreased. This accuracy degradation is caused by the nonlinearity and the device variation in the synaptic array, which affect the VMM operations of the Conv1 layer. Note that the features of the input images are directly extracted in the Conv1 layer with a small size of weights. Thus, the accuracy is affected by the weight changes in the Conv1 layer [68]. In contrast, the effective training of the Conv1 layer can raise the accuracy of neuromorphic systems to that of CNNs in which the pre-trained weights in the Conv1 layer are used as they are.

This approach is significantly efficient for demonstrating the high performance of the training method in the implemented synaptic array.

The experimental results of the hybrid training in the fabricated AND-type flash array are shown in Fig. 3.3. The experiment results are compared with the simulation results, in which the updates in w_{array}s of the Conv1 layer are calculated by the LTP and LTD model of the synaptic devices [52, 53]. In the simulation, the weights are also fixed during the online training process, except for the Conv1 layer. Fig. 3.3 (a) represents the loss value during the online training, where the black line indicates the experiment loss value, and the red line indicates the simulation loss value. The PGM and ERS conditions for updating the conductance are 8 V, 10 μs, and -8 V, 10 ms, respectively. At the start point of the training iteration, the loss value is relatively high, indicating that the nonidealities of the AND-type flash array cause errors in the weight transfer step. However, the loss value in the experiment is rapidly reduced as the training iteration increases. The average values of the loss in the experiment and simulation results are ~0.051 and ~0.035, respectively, for the last 100 training iterations. The small difference between the loss values can be

caused by the read fluctuation, parasitic resistance, or conductance drift over time, which are not calculated in the simulation.

The classification accuracy of the neuromorphic system is evaluated in the simulation and the experiment, as shown in Fig. 3.3 (b). An accuracy gap is exhibited between the pre-trained network (99.0 %) and the neuromorphic system just after the weight transfer step (82.5%). However, the accuracy in the experiment also rapidly increases as the training iteration increases. In particular, the accuracy is recovered to 98.2 % after 1-epoch online training of the Conv1 layer in the fabricated synaptic array. In this experiment, the LTP and LTD curves of the fabricated synaptic devices are very nonlinear, which can cause significant accuracy degradation in the reported online training methods. However, in the proposed training method, the w_cs are trained first in the software using the STE in online training step; thus, the linear and symmetric weight updates can be performed computationally. The updates in w_{array}s are performed if the *PN* of the device is changed, thereby minimizing the nonlinear weight updates and increasing the accuracy in the proposed training method.

In order to analyze the weight changes over the training iterations, we trace the I_{DS}s of randomly selected devices, as shown in Fig. 3.3 (c). At the abrupt I_D changes during the training iterations, the PGM or ERS pulse is applied to the device. We confirmed that the increase and decrease in I_{DS}s are asymmetric to each other because of the LTP and LTD curves of the flash devices. Furthermore, the device nonidealities, including noise, I_D drift, and read fluctuation, are exhibited in the training process. However, by applying the proposed hybrid training, the neuromorphic systems achieve high accuracy compared to the baseline accuracy. These results indicate that the proposed method is significantly effective in improving the accuracy of neuromorphic systems in which various device nonidealities exist. Additionally, the experimental demonstration of the training performance in the fabricated synaptic array is one of the major contributions of this work compared to other papers on online training methods demonstrated only in software.

Measurement results of the I_{DS}s of all devices in the Conv1 layer are shown in Fig. 3.4 (a) during the 1-epoch training. This figure indicates the whole training

process and the conductance changes in the fabricated synaptic array. Fig. 3.4 (b) compares the measured w_{array}s in the flash array with w_cs in the software after 1-epoch training. The measured w_{array} distribution does not match the w_c distribution, which mainly results from the device nonidealities, as shown in Fig. 3.3 (c). It is worth noting that the measured w_{array}s in the Conv1 layer are trained while reflecting the device nonidealities in the synaptic array. The CNN with the w_{array}s achieves very high classification accuracy (98.2%), which is close to that of pre-trained CNNs in the software (99.0%). Fig. 3.5 shows the w_{array} distributions in the other layers except for the Conv1 layer before and after the training process. The distributions in each layer are exactly the same before and after the training process, validating that the weights in the layers except for the Conv1 layer are fixed during the training process. It also indicates that the accuracy improvement of the CNN is achieved by the online training of the Conv1 layer after the weight transfer step. Fig. 3.6 shows the experimental results of the hybrid training in ten different synaptic arrays. After 1-epoch online training of the Conv1 layer, the CNNs with

the ten different arrays achieve high accuracy (average accuracy: 97.49%).

TABLE 3.1

ACCURACY BEFORE ONLINE TRAINING

	Conv1: a* **Others: a***	**Conv1: a*** **Others: b***	**Conv1: c*** **Others: b***
Accuracy (%)	99.0	97.8	82.5

a*: offline trained $w_{q,}$ b*: offline trained w_q w/ 40% variation

c*: w_{array} in the fabricated array after weight transfer step

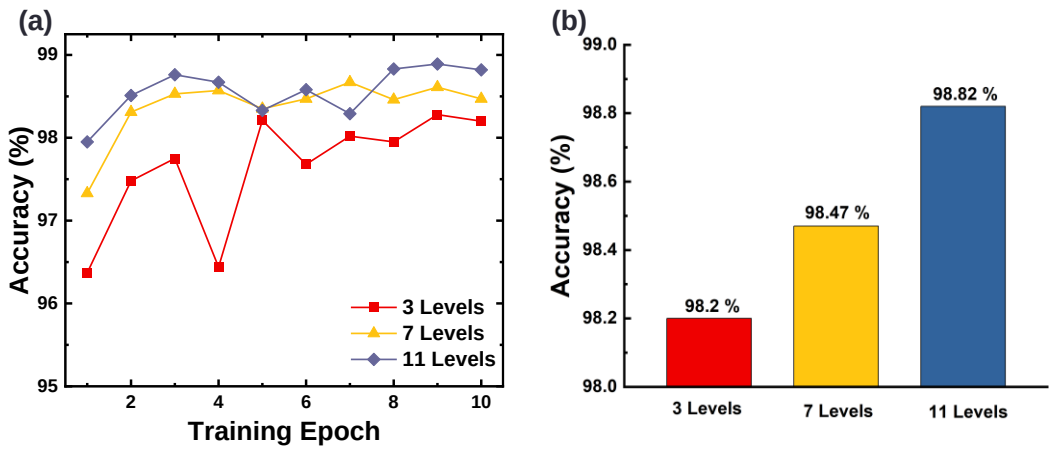
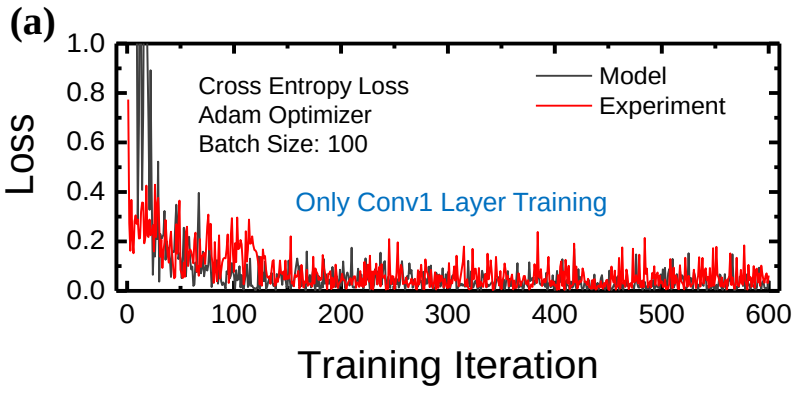
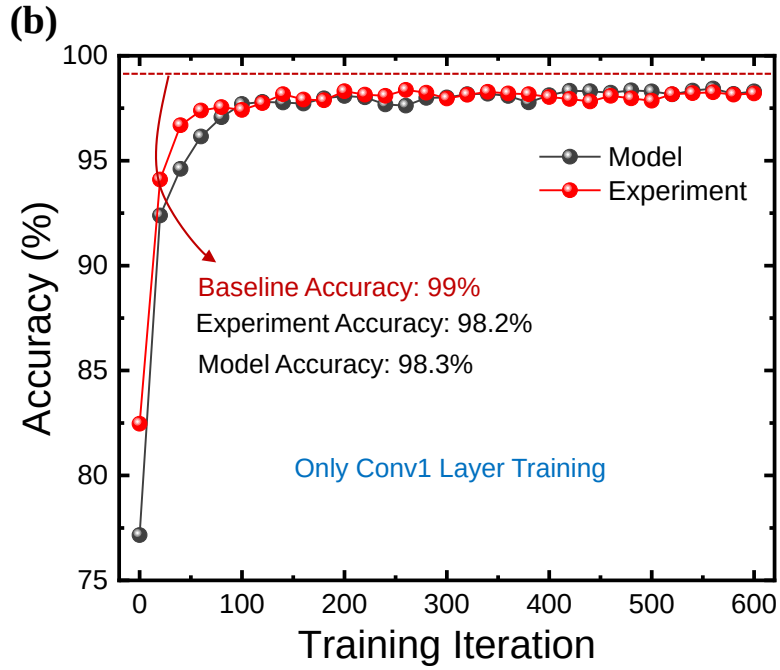


Fig. 3.2. (a) Training curves of the 5-layer CNN trained by the QNN training method as a parameter of weight levels. (b) Accuracy of the CNN at 10-epoch.


(a)
Loss
1.0
0.8
0.6
0.4
0.2
0.0
Cross Entropy Loss
Adam Optimizer
Batch Size: 100
Model
Experiment
Only Conv1 Layer Training
0
100
200
300
400
500
600
Training Iteration


(b)
Accuracy (%)
100
95
90
85
80
75
Model
Experiment
Baseline Accuracy: 99%
Experiment Accuracy: 98.2%
Model Accuracy: 98.3%
Only Conv1 Layer Training
0
100
200
300
400
500
600
Training Iteration

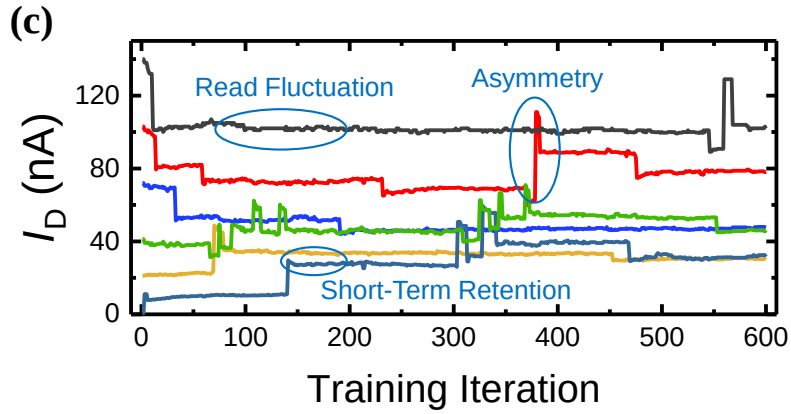

(c)
I_D (nA)
120
80
40
0
Read Fluctuation
Asymmetry
Short-Term Retention
0
100
200
300
400
500
600
Training Iteration

Fig. 3.3. (a) Cross entropy loss value with respect to the training iteration during the online training step. (b) Accuracy curves of the neuromorphic systems for MNIST test set images. The baseline accuracy from the QNN pre-training is 99.0 %. (c) I_D changes over the online training iteration in six flash devices. Various device nonidealities are shown in the I_D changes.

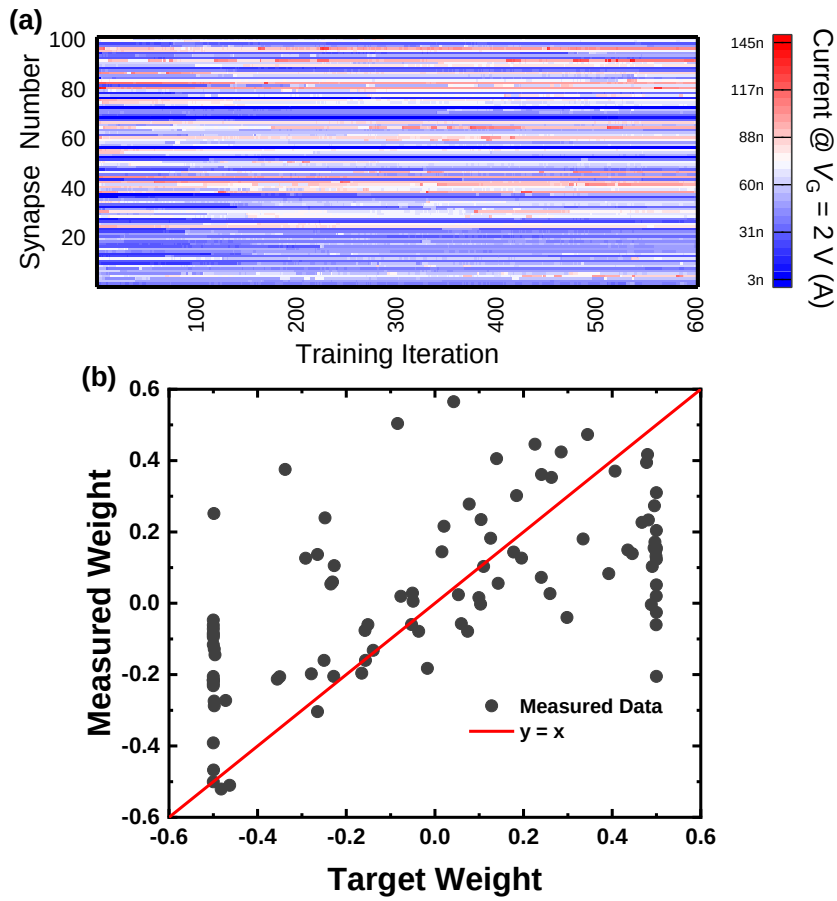


Fig. 3.4. (a) I_D changes of 100 devices in the fabricated synaptic array representing weights in the Conv1 layer over online training iteration. This result represents the overall training process. (b) Measured w_{array}s in the fabricated synaptic array versus w_c in software after 1-epoch online training.

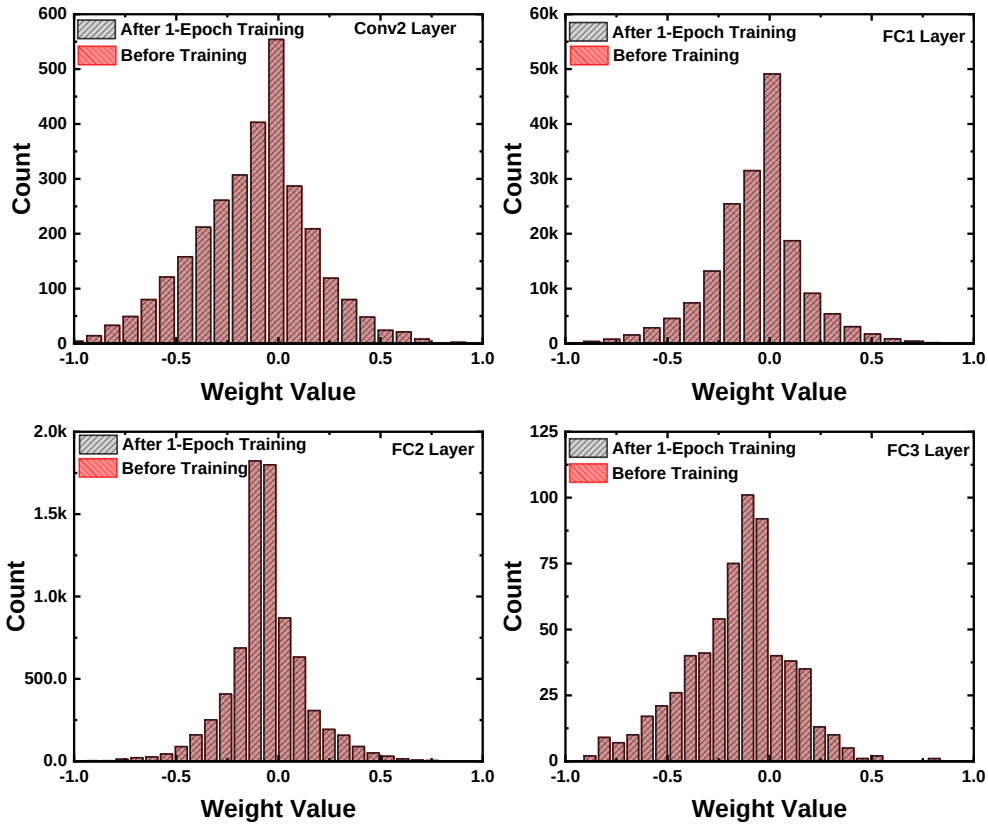


Fig. 3.5. Distribution of weights before and after online training in the layers except for the Conv1 layer.

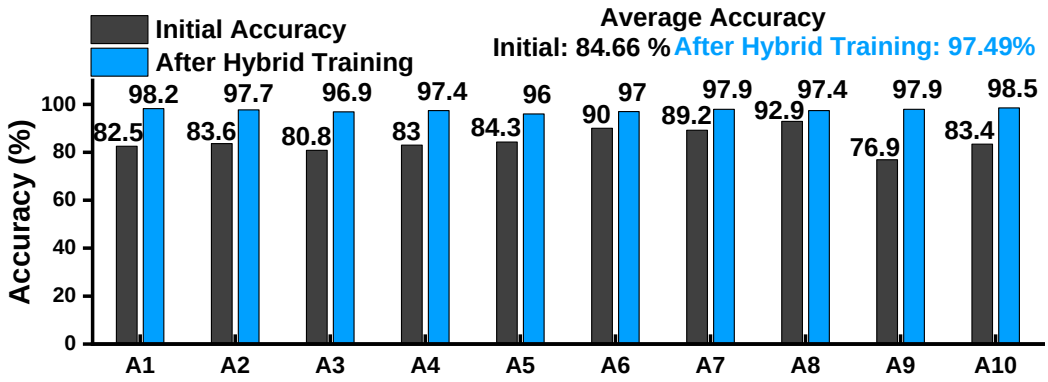


Fig. 3.6. Accuracy evaluation in ten fabricated synaptic arrays using the proposed hybrid training (A1~A10). Only the Conv1 layer in the CNN was trained..

3.3 Evaluation of Hybrid Training for Device Nonidealities

Fig. 3.7 (a) and (b) show the measurement results of the hybrid training as a parameter of PGM and ERS pulse amplitude. The nonlinearity of LTP and LTD curves in the flash devices is modulated by the PGM and ERS pulse amplitude, as shown in Fig. 2.5. In particular, the LTP and LTD curves are significantly nonlinear at a pulse amplitude of 10 V. The abrupt I_D changes in the nonlinear curves cause the abrupt changes in weights. Thus, the decreasing speed in the training loss at a pulse amplitude of 10 V is slower than that at other pulse amplitudes. In addition, the more fluctuated loss curve is exhibited at a larger pulse amplitude. These features in the loss curves are also shown in the accuracy curves in Fig. 3.7 (b). The increasing speed in the accuracy is slow at a pulse amplitude of 10 V with more significant fluctuations. Although the maximum classification accuracy of ~97% is achieved at a pulse amplitude of 10 V, the highly nonlinear conductance responses lower the training performance of the proposed method. However, the nonlinearity effects can be mitigated by decreasing the learning rate in the online training step. The decreased learning rate reduces the number of weight updates in the synaptic

devices and the number of abrupt weight updates. Thus, a more stable accuracy curve can be obtained by reducing the learning rate, resulting in high classification accuracy (97.76%), as shown in Fig. 3.8.

Fig. 3.9 (a) shows the I_D changes of 100 flash devices in the fabricated array over the retention time after the 1-epoch training was conducted. The training was conducted with a pulse amplitude of 8 V and -8 V for PGM and ERS, respectively. The fabricated flash devices show superior retention characteristics with the charge trap layer, as shown in Fig. 2.9. The small V_{th} changes over time result in the trained I_{DS}s being maintained over three days. Fig. 3.9 (b) shows the accuracy degradation after 8 hours in 3 different flash arrays. Due to the nonvolatile memory characteristics of flash devices, the CNNs with the arrays maintain high accuracy after 8 hours. The endurance characteristics of flash devices under the PGM and ERS conditions are shown in Fig. 3.10. (a). The V_{th} difference in the PGM and ERS states of the flash device is maintained until the cycles of 10^5. Because only a small part of the entire memory window in the flash device is used in the cycling tests, the device degradation is not extensively exhibited until 10^5 cycles. It is worth

noting that the total number of PGM and ERS pulses to the flash synaptic array is ~500 during the training process, as shown in Fig. 3.10 (b). The average number of PGM and ERS pulses to one device is ~5, which is significantly lower than 10^5. The number of PGM and ERS pulses can be additionally decreased by reducing the learning rate. Therefore, the device degradation by the PGM and ERS pulses during the training process is insignificant.

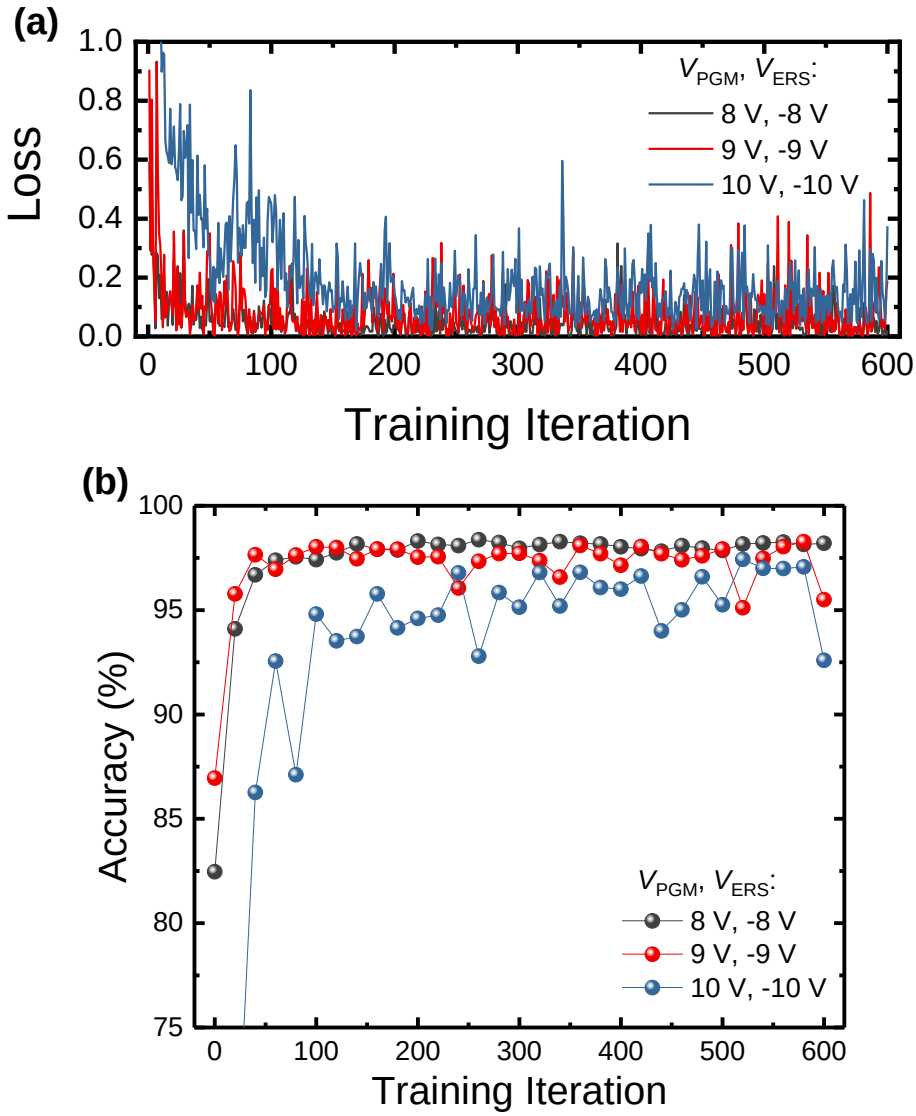


Fig. 3.7. (a) Cross entropy loss value with respect to the online training iteration as a parameter of PGM and ERS pulse amplitude. Pulse widths for the PGM and ERS operations are 10 μs and 10 ms, respectively. (b) Accuracy curves of the neuromorphic systems as a parameter of the PGM and ERS pulse amplitude.

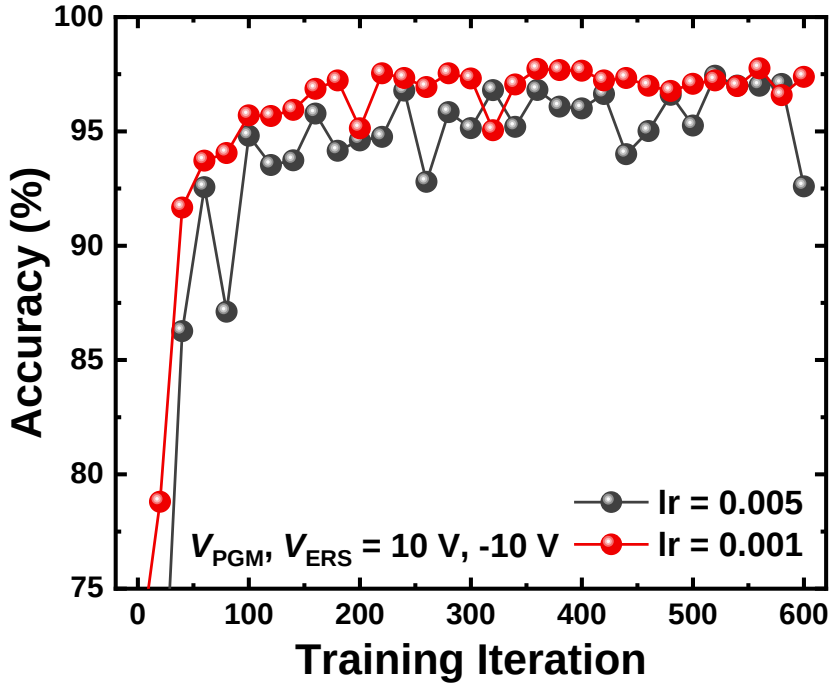


Fig. 3.8. Accuracy curves of the neuromorphic systems during the online training step as a parameter of learning rate (*lr*). The pulse amplitudes of the PGM and ERS operations are 10 V and -10 V, respectively.

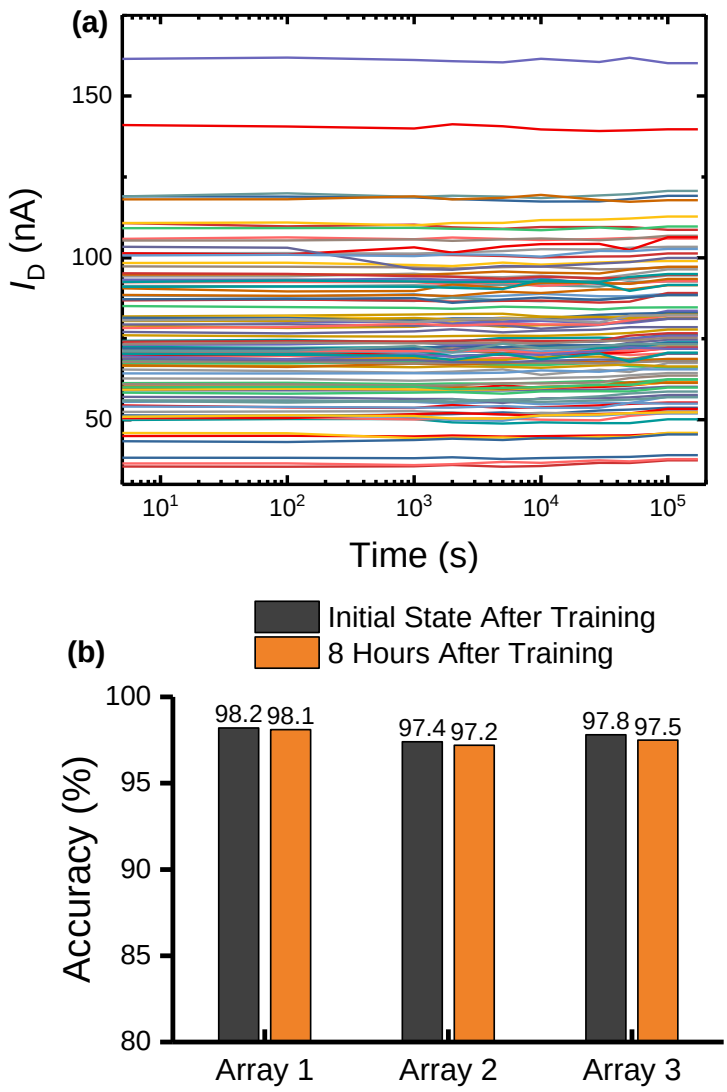


Fig. 3.9. (a) I_{DS} of 100 devices over the retention time. (b) Accuracy comparison in three different arrays after 8 hours.

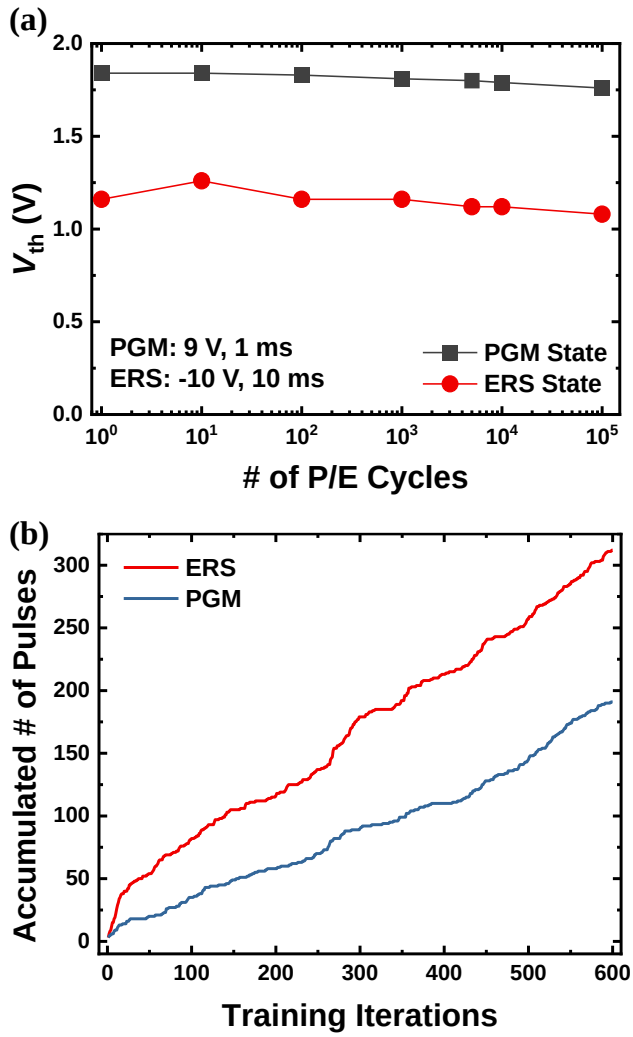


Fig. 3.10. (a) Threshold voltage (V_{th}) of the flash device with respect to the number of PGM and ERS cycles. The pulse amplitude and width are 9 V and 1 ms for the PGM operation, and -10 V and 10 ms for the ERS operation. (b) Measured number of PGM and ERS pulses applied to 100 flash devices in the online training step.

3.4 Comparison with Online Training and Other Works

The main advantage of the proposed hybrid training method is that the training efficiency is significantly improved compared to the online training conducted from the beginning. Fig. 3.11 shows the accuracy comparison of the hybrid training and online training for MNIST image classification. The accuracy curve of the hybrid training was obtained in the fabricated synaptic array, and that of the online training was obtained in the software, where the nonlinearity in LTP and LTD curves with a pulse amplitude of 8 V and device-to-device variation were reflected. The hybrid training was conducted on the Conv1 layer, and the online training was conducted on all layers. As shown in Fig. 3.11, the accuracy of the online training is ~98.3% after 10-epoch training, even with the nonlinear conductance response, which is very close to the accuracy of offline trained CNN. This training result means the online training itself significantly enhances the accuracy of neuromorphic systems with hardware nonidealities. However, the proposed hybrid training improves the accuracy of neuromorphic systems much faster than the online training conducted from the beginning. In the hybrid training, the neuromorphic system achieves an

accuracy of ~98.3 % in less than one epoch training using only the Conv1 layer. This advantage dramatically enhances the training efficiency in neuromorphic systems. The total number of PGM and ERS pulses between the hybrid training and the online training are compared in Table 3.2. In the hybrid training, the total number of PGM and ERS pulses is ~500 times for an accuracy of 98.3%, whereas it is ~2.4×10^6 times in the online training. The difference in the number of PGM and ERS pulses is because the online training conducted from the beginning requires all-layer training with longer training iterations to achieve high accuracy. Given that the training cost in neuromorphic systems increases as the number of weight updates increases, the low number of weight updates in hybrid training is one of the main advantages.

The comparison of this work and other reported online training algorithms for neuromorphic systems is shown in Table 3.3. There are many studies on neuromorphic systems using offline training methods, which adopt the conductance tuning process to transfer the offline trained weights to the conductance of synaptic devices. In offline training methods, high accuracy of software-based neural

networks can be achieved if the weights are accurately transferred. However, errors in the weight transfer step exist in neuromorphic systems, degrading the accuracy. In addition, the time-varying device nonidealities can degrade the performance of neuromorphic systems with offline training. Many studies on neuromorphic systems use online training methods in which weight training is performed in the systems. However, most online training methods are verified in the software simulation, although the training method is proposed for neuromorphic hardware. Therefore, the online training methods cannot fully calculate the device nonidealities in the software simulation, meaning that the reported online training methods cannot fully reflect the hardware imperfections. Some works on online training methods verified in the hardware; however, they adopt the closed-loop conductance tuning process to update the weights, significantly increasing the training cost. On the other hand, the hybrid training method proposed in this work exhibits superior accuracy improvements in neuromorphic systems without the conductance tuning process. This result shows that the proposed method has immunity to the time-static nonidealities of synaptic devices without the tuning

process. Furthermore, the hybrid training method can reduce the training cost and increase training efficiency with 1-epoch training for the Conv1 layer. Notably, the high performance of the proposed hybrid training method is experimentally verified with various device nonideality conditions in the fabricated flash synapse array. Therefore, the proposed hybrid training method can be generally applied to neuromorphic hardware with various analog synaptic devices.

TABLE 3.2
NUMBER OF PGM/ERS PULSES FOR HIGH ACCURACY THROUGHOUT TRAINING

	Hybrid Training: a*	**Online Training: b***
# of PGM Pulses	$\mathbf{1.91\times10^{2}}$	1.18×10^{6}
# of ERS Pulses	$\mathbf{3.12\times10^{2}}$	1.17×10^{6}

a*: Conducted on Conv1 layer, 1-epoch training for accuracy of 98.3%

b*: Conducted on all layers, 10-epoch training for accuracy of 98.3%

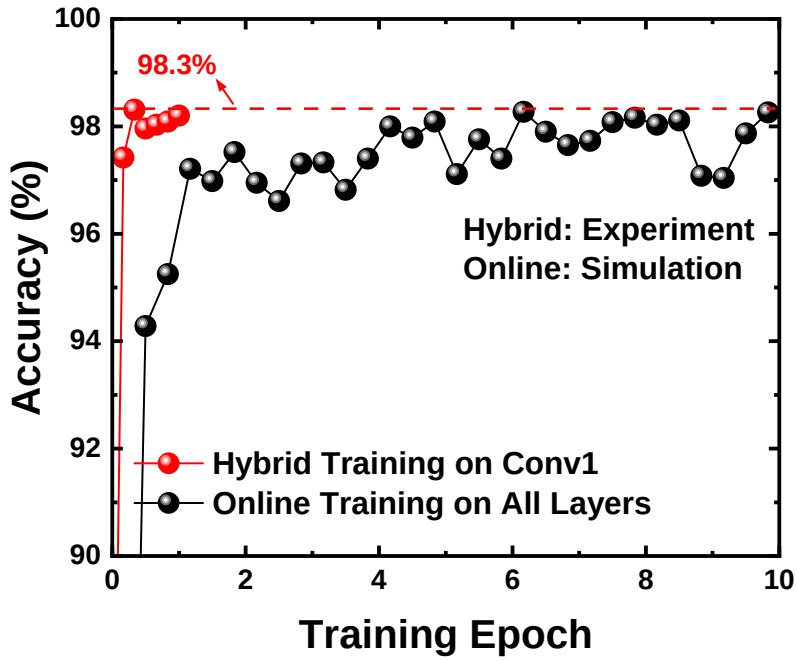


Fig. 3.11. Accuracy curves of the CNN with different training methods. The training curve of hybrid training was evaluated in the experiment, and that of online training was evaluated in the simulation.

TABLE 3.3

COMPARISON WITH OTHER WORKS

Other works	Training method	Evaluation Site	Training Epoch	ΔG Tuning	Accuracy
Zhang et al. [49]	Online Sign BP	Software Simulation	20	X	94.50%
Chang et al. [51]	Online BP	Software Simulation	> 10	X	97.93%
Fu et al. [69]	Online BP	Software Simulation	125	X	95.55%
Lim et al. [48]	Online Manhattan Learning Rule	Software Simulation	1	X	95.36%
D. Kwon et al. [17]	Online BP	Software Simulation	20	X	97.83%
Ambrogio et al. [70]	Online BP	*Hardware PCM*	20	Closed Loop	97.94%
C. Li et al. [71]	Online BP	*Hardware RRAM*	> 1	ISPP-Based	91.71%
P. Yao et al. [72]	*Hybrid BP*	*Hardware RRAM*	1	Closed Loop	96.19% (Only Last Layer)
This work	*Hybrid BP*	*Hardware AND Flash*	*≤ 1*	*X*	*98.2% (Only Conv1)*

Chapter 4
Conclusion

In this work, we have fabricated a flash-type synaptic device with high reliability, scalability, and CMOS process compatibility. The fabricated flash device has a charge trap layer of Si_3N_4, exhibiting nonvolatile memory functionality and the capability of multi-bit weight storage. We have also fabricated an AND-type array architecture with flash devices. Due to the parallel SLs and BLs, the fabricated AND-type array has the advantage of high efficiency in VMM operations and selective PGM and ERS operations. Furthermore, to utilize the flash device as an artificial synaptic device, the synaptic characteristics of the array have been systematically analyzed and optimized in terms of device variation, nonlinearity, and reliability. These optimizing results indicate that the fabricated flash array is outstanding as a synaptic array for low-power and highly reliable neuromorphic systems.

Besides, we have proposed a novel hybrid training method, which combines the offline training and online training for neuromorphic systems. The performance

of the proposed training method was experimentally demonstrated in the fabricated AND-type flash array. After the weight transfer step, the neuromorphic system exhibits a degraded accuracy (82.5%) for MNIST image classification, which is lower than the accuracy of offline trained CNN. However, the accuracy of the neuromorphic system rapidly increases to 98.2% by using the hybrid training on only the Conv1 layer for one epoch. Furthermore, the proposed method was experimentally verified to achieve high accuracy under various device nonideality conditions. These results indicate that the hybrid training method can be generally applied to neuromorphic systems using other types of synaptic devices. Consequently, the proposed hybrid training method provides a highly efficient training solution for neuromorphic systems using analog synaptic devices.

Bibliography

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolu tional neural networks," *In Proc. Adv. Neural Inf. Process. Syst*., pp. 1097-1105, 2012.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *In Proc. IEEE conf. Comput. Vis. Pattern Recognit. (CVPR),* pp. 770-778, 2016.

[3] R. Girshick, "Fast R-CNN," *in Proc. IEEE Int. Conf. Comput. Vis. (ICCV),* p, 1440, 2015.

[4] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 38, no. 2, pp. 295-307, 2016.

[5] Q. V. Le, "Building high-level features using large scale unsupervised learning," *In Proc. Int. Conf. Mach. Learn. (ICML),* pp. 8595-8598, 2012.

[6] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," *In Proc. Conf. Empirical Methods Natural Lang. Process*., pp. 1422-1432, 2015.

[7] Yin, W. et al., "Comparative Study of CNN and RNN for Natural Language Processing," Preprint at https://arxiv.org/pdf/1702.01923, 2017.

[8] Conneau, A. et al., "Very Deep Convolutional Networks for Natural

Language Processing," Preprint at https://arxiv.org/abs/1606.01781, 2016.

[9] Y. Goldberg, "A Primer on Neural Network Models for Natural Language Processing," *Journal of Artificial Intelligence Research*, vol. 57, pp. 345-420, 2016.

[10] A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 10, pp. 4291-4308, 2021.

[11] Wu, B. et al., "SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving," *In Proc. IEEE conf. Comput. Vis. Pattern Recognit. (CVPR*), pp. 129-137, 2017.

[12] A. E. Sallab, M. Abdou, E. Perot, and S. Yogamani, "Deep reinforcement learning framework for autonomous driving," *Electronic Imaging*, vol. 19, pp. 70-76, 2017.

[13] K. Fukushima et al., "Neocognitron: a hierarchical neural network capable of visual pattern recognition," *Neural Netw.,* vol. 1, pp. 119-130, 1998.

[14] M. Riesenhuber, and T. Poggio, "Hierarchical models of object recognition in cortex," *Nat. Neurosci*., vol. 2, 1999.

[15] S. Yu, "Neuro-Inspired Computing With Emerging Nonvolatile Memory," *Proceedings of IEEE*, vol. 106, pp. 260-285, 2018.

[16] D. Kwon, S. Y. Woo, J.-H. Bae, S. Lim, B.-G. Park, and J.-H. Lee, "Hardware-based Spiking Neural Networks Using Capacitor-Less Positive

Feedback Neuron Devices," *IEEE Transactions on Electron Devices*, vol. 68, no. 9, pp. 4766-4772, 2021.

[17] D. Kwon et al, "On-chip Training Spiking Neural Networks Using Approximated Backpropagation With Analog Synaptic Device," *Front. Neurosci*., 14.423, 2021.

[18] P. U. Diehl et al., "Fast-Classifying, High-Accuracy Spiking Deep Networks Through Weight and Threshold Balancing," *In 2015 IEEE International Joint Conference on Neural Networks (IJCNN),* 2015.

[19] H. Kim, M. R. Mahmoodi, H. Nili, and D. B. Strukov, "4K-memristor analog-grade passive crossbar circuit," *Nat. Commun*., vol. 12, p. 5198, 2021.

[20] Fuller, E. J. et al., "Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing," *Science,* vol. 364, pp. 570–574, 2019.

[21] D. Kwon, G. Jung, W. Shin, Y. Jeong, S. Hong, S. Oh, J.-H. Bae, B.-G. Park, J.-H. Lee, "Low-power and reliable gas sensing system based on recurrent neural networks," *Sensors and Actuators B: Chemical*, vol. 340, p. 129258, 2021.

[22] D. Kwon, G. Jung, W. Shin, Y. Jeong, S. Hong, S. Oh, J. Kim, J.-H. Bae, B.-G. Park, and J.-H. Lee, "Efficient fusion of spiking neural networks and FET-type gas sensors for a fast and reliable artificial olfactory system," *Sensors and Actuators B: Chemical,* vol. 345, p. 130419, 2021.

[23] Li, C. et al., "Efficient and self-adaptive in-situ learning in multilayer memristor neural networks," *Nat. Commun.*, vol. 9, p. 2385, 2018.

[24] K. Moon, M. Kwak, J. Park, D. Lee, and H. Hwang, “Improved conductance linearity and conductance ratio of 1T2R synapse device for neuromorphic systems,” *IEEE Electron Device Lett*., vol. 38, no. 8, pp. 1023-1026, 2017.

[25] Prezioso, M. et al., “Training and operation of an integrated neuromorphic network based on metal-oxide memristors,” *Nature*, vol. 521, pp. 61–64, 2015.

[26] G. C. Adam,B. D. Hoskins, M. Prezioso, F. M.-Bayat, B. Chakrabarti, and D. B. Strukov, “3-D memristor crossbars for analog and neuromorphic computing applications,” *IEEE Trans. Electron Devices*, vol. 64, no. 1, pp. 312–318, 2017.

[27] Merrikh Bayat, F. et al., “Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits,” *Nat. Commun*., vol. 9, p. 2331, 2018.

[28] Wang, Z. et al., “Fully memristive neural networks for pattern classification with unsupervised learning,” *Nat. Electron*., vol. 1, pp. 137-145, 2018.

[29] Y. H. Jang, W. Kim, J. Kim, K. S. Woo, H. J. Lee, J. W. Jeon, S. K. Shim, J Han, and C. S. Hwang, “Time-varying data processing with nonvolatile memristor-based temporal kernel,” *Nat. Commun*., vol. 12, p. 5727, 2021.

[30] T. Dalgaty, N. Castellani, C. Turck, K.-E. Harabi, D. Querlioz, and E. Vianello, “In situ learning using intrinsic memristor variability via Markov chain Monte Carlo sampling,” *Nat. Electron.,* vol. 4, pp. 151-161, 2021.

[31] I. Boybat, M. L. Gallo, S. R. Nandakumar, T. moraitis, T. Parnell, T. Tuma,

B. Rajendran, Y. Leblebici, A. Sebastian, and E. Eleftheriou, “Neuromorphic computing with multi-memristive synapses,” *Nat. Commun.,* vol. 9, p. 2514, 2018.

[32] I. Boybat, et al., “Stochastic weight updates in phase-change memory-based synapses and their influence on artificial neural networks,” *In Proc. IEEE 13th Conf. Ph.D. Res. Microelectron. Electron (PRIME),* pp. 13-16, 2017.

[33] S. Kariyappa, et al., “Noise-Resilient DNN: Tolerating Noise in PCM-Based AI Accelerators via Noise-Aware Training,” *IEEE Transactions on Electron Devices*, vol. 68, no. 9, pp. 4356-4361, 2021.

[34] Burr, G. W. et al., “Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165000 Synapses) Using Phase-Change Memory as the Synaptic Weight Element,” *IEEE Transactions on Electron Devices*, vol. 62, pp. 3498-3507, 2015.

[35] Y. Kaneko, Y. Nishitani, and M. Ueda, “Ferroelectric artificial synapses for recognition of a multishaded image,” *IEEE Trans. Electron Devices*, vol. 61, pp. 2827–2833, 2014.

[36] S. Boyn et al., “Learning through ferroelectric domain dynamics in solid-state synapses,” *Nat. Commun*., 8, 14736, 2017.

[37] 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 Transactions on Electron Devices*, vol. 65, no. 5, pp. 1774-1780, 2018.

[38] P. Wang, F. Xu, B. Wang, B. Gao, H. Qian, and S. Yu, “Three-Dimensional NAND Flash for Vector-Matrix Multiplication,” *IEEE Trans. Very Large Scale Integration Systems*, vol. 27, no. 4, pp. 988-991, 2019.

[39] D. Kwon, W. Shin, J.-H. Bae, S. Lim, B.-G. Park, and J.-H. Lee, “Investigation of low-frequency noise characteristics in gated Schottky diodes,” *IEEE Electron Device Letters*, vol. 42, no. 3, pp. 442-445, 2021.

[40] M.-B. Farnood, X. Guo, M. Klachoko, M. Prezioso, K. K. Likharev, and D. B. Strukov, “High-performance mixed-signal neurocomputing with nanoscale floating-gate memory cell arrays,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 4782-4790, 2017.

[41] M. Kim, M. Liu, L. Everson, G. Park, Y. Jeon, S. Kim, S. Lee, S. Song, and C. H. Kim, “A 3D NAND Flash Ready 8-Bit Convolutional Neural Network Core Demonstrated in a Standard Logic Process,” *2019 IEEE International Electron Devices Meeting.*

[42] S.-T. Lee, H. Kim, J.-H. Bae, H. Yoo, N. Y. Choi, D. Kwon, S. Lim, B.-G. Park, 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),* 2019.

[43] Y. Xiang, P. Huang, R. Han, C. Li, K. Wang, X. Liu, and J. Kang, “Efficient and robust spike-driven deep convolutional neural networks based on NOR flash computing array,” *IEEE Transactions on Electron Devices*, vol. 67, no. 6, 2020.

[44] Y.-T. Seo, D. Kwon, Y. Noh, S. Lee, N.-K. Park, S. Y. Woo, B.-G. Park, and J.-H. Lee, "3-D AND-Type Flash Memory Architecture With High-k Gate Dielectric for High-Density Synapse Devices," *IEEE Transactions on Electron Devices,* vol. 68, no. 8, pp. 3801-3806, 2021.

[45] W.-M. Kang, et al., "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.

[46] F. Alibart, L. Gao, B. D. Hoskins, and D. B. Strukov, "High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm," *Nanotechnology*, vol. 23, no. 7, 075201, 2012.

[47] L. Gao, P.-Y. Chen, and S. Yu, "Programming protocol optimization for analog weight tuning in resistive memories," *IEEE Electron Device Lett*., vol. 36, no. 11, pp. 1157-1159, 2015.

[48] 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, 2019.

[49] Zhang, H. Wu, P. Yao, W. Zhang, B. Gao, N. Deng, and H. Qian, "Sign backpropagation: An on-chip learning algorithm for analog RRAM neuromorphic computing systems," *Neural Networks*, vol. 108, pp. 217-223, 2018.

[50] X. Peng, S. Huang, H. Jiang, A. Lu, and S. Yu, "DNN+NeuroSim V2.0: AN

End-to-End Benchmarking Framework for Compute-in-Memory Accelerators for On-Chip Training," *IEEE Transactions on Computer-Aided Design of Integrated Circuits AND Systems*, vol. 40, no.11, pp. 2306-2319, 2021.

[51] C.-C. Chang et al., "Mitigating asymmetric nonlinear weight update effects in hardware neural network based on analog resistive synapse, *IEEE J. Emerg. Select. Top. Circ. Syst*., vol. 8, pp. 116-124, 2017.

[52] D. Querlioz et al., "Learning with memristive devices: How should we model their behavior?," In Proc. IEEE/ACM Int. Symp. NANOARCH., pp. 150-156, 2011.

[53] C.-H. Kim et al., "Emerging memory technologies for neuromorphic computing," *Nanotechnology,* vol. 30, no. 3, 2018.

[54] F. R. Libsch and M. H. White, "Charge transport and storage of low programming voltage SONOS/MONOS memory devices," *Solid-State Electron*., vol. 33, no. 1, pp. 105–126, 1990.

[55] J.-H. Bae, S. Lim, B.-G. Park, and J.-H. Lee, "High-Density and Near-Linear Synaptic Device Based on a Reconfigurable Gated Schottky Diode," *IEEE Electron Device Letters,* vol. 38, no. 8, 2017.

[56] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano letters*, vol. 10, no. 4, 1297-1301. 2010.

[57] L. Gao *et al.,* "Fully parallel write/read in resistive synaptic array for accelerating on-chip learning," *Nanotechnology,* vol. 26, no. 45, 2015.

[58] S. Park *et al.,* "Neuromorphic speech systems using advanced ReRAM-based synapse," *IEEE International Electron Device Meeting (IEDM)*, 2013.

[59] J. Woo *et al.,* "Improved synaptic behavior under identical pulses using AlOx/HfO2 bilayer RRAM array for neuromorphic systems," *IEEE Electron Device Letters,* vol. 37, no. 8, 2016.

[60] W. Chung, M. Si, and P. D. Ye, "First Demonstration of Ge Ferroelectric Nanowire FET as Synaptic Device for Online Learning in Neural Network with High Number of Conductance State and Gmax/Gmin," *2018 IEEE International Electron Device Meeting (IEDM),* 2018.

[61] S. Seo, B. Kim, D. Kim, S. Park, T. R. Kim, J. Park, H. Jeong, S. O. Park, T. Park, H. Shin, M. S. Kim, Y. K. Choi, and S. Choi, "The gate injection-based field-effect synapse transistor with linear conductance update for online training," *Nature Communications*, vol. 13, no. 1, 2022.

[62] M.-K. Kim, and J.-S. Lee, "Ferroelectric Analog Synaptic Transistors," *Nano Letters,* vol 19, 2019.

[63] S. Zhou et al., "Dorefa-net: Taining low bitwidth convolutional neural networks with low bitwidth gradients," Preprint at https://arxiv.org/pdf/1606.06160, 2018.

[64] I. Hubara, M. Courbariux, D. Soudry, R. E.-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," *The Journal of Machine Learning Research*, vol. 18, pp. 6869-

6898, 2017.

[65] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, “Xnor-net: Imagenet classification using binary convolutional neural networks,” *European conference on computer vision,* pp. 525-542, 2016.

[66] M. Courbariaux, et al., “Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1,” Preprint at https://arxiv.org/pdf/1602.02830, 2016.

[67] S.-T. Lee, D. Kwon, H. Kim, H. Yoo, and J.-H. Lee, “NAND flash based novel synaptic architecture for highly robust and high-density quantized neural networks with binary neuron activation of (1, 0),” *IEEE Access,* vol. 8, pp. 114330-114339, 2020.

[68] H. Kim, J.-H. Bae, S. Lim, S.-T. Lee, Y.-T. Seo, D. Kwon, B.-G. Park, and J.-H. Lee, “Efficient precise weight tuning protocol considering variation of the synaptic devices and target accuracy,” *Neurocomputing*, vol. 378, pp. 189-196, 2020.

[69] J. Fu et al., “Mitigating Nonlinear Effect of Memristive Synaptic Device for Neuromorphic Computing,” *IEEE Journal of Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 2, 2019.

[70] S. Ambrogio et al., “Equivalent-accuracy accelerated neural-network training using analogue memory,” *Nature*, vol. 558, pp. 60-67, 2018.

[71] Li, C. et al., “Efficient and self-adaptive in-situ learning in multilayer

memristor neural networks," *Nat. Commun*., vol. 9, p. 2385, 2018.

[72] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and H. Qian, "Fully hardware-implemented memristor convolutional neural network," *Nature*, vol. 577, pp. 641-646, 2020.

초 록

아날로그 비휘발성 메모리 셀을 시냅스 장치로 사용하는 뉴로모픽 기술은 대규모 벡터 행렬 곱셈 연산을 수행하기 위해 시간과 에너지 소비를 줄일 수 있어, 폰 노이만 구조의 컴퓨팅 아키텍처를 대체할 수 있는 기술로서 매우 유망하다. 그러나 뉴로모픽 하드웨어에 관한 보고된 학습 방법은 아날로그 장치의 비이상적인 특성으로 인해 정확도가 현저히 낮고, 비용이 상당히 높은 전도도 조정 프로세스가 필요하다. 따라서 이 논문은 비휘발성 아날로그 메모리 셀을 활용하여 뉴로모픽 하드웨어를 효율적으로 훈련시키는 새로운 하이브리드 훈련 방법을 제안하고, 구현된 하드웨어에 실험적으로 시연하여 제안하는 방법의 고성능을 보여준다. 제안하는 훈련 방법은 아날로그 시냅스 장치 전도도에 가중치의 변화를 반영하기 위한 전도도 조정 프로토콜에 의존하지 않아 온라인 훈련의 비용이 상당히 줄어든다. 또한 이 논문에서는 대규모 벡터 행렬 곱셈 연산을 가속화하는 뉴로모픽 하드웨어로 AND형 플래시 메모리 어레이를 구현한다. 제작된 뉴로모픽 하드웨어는 전하 포획층(SiO2/Si3N4/SiO2)과 함께 비휘발성 메모리 기능을 가지고 있어 반영구적으로 가중치를 유지한다. 그런 다음 제안된 훈련 방법을 제작한 시냅스 어레이에 적용하여, 첫 번째 시냅스 층에 1-epoch 학습으로도 소프트웨어 기반 신경망의 정확도에 근접함을 보여준다. 또한 제안하는 하이브리드 훈련 방법은 가중치 업데이트 특성이 극도로 비선형적인 다양한 유형의 시냅스 장치를 포함한 저전력 뉴로모픽 하드웨어에 효율적으로

적용될 수 있는 것을 검증한다. 제작된 하드웨어에서 제안된 방법의 성공적인 시연은 비휘발성 아날로그 메모리 셀을 사용하는 뉴로모픽 하드웨어가 미래 인공 지능을 위한 보다 유망한 플랫폼이 될 수 있음을 보여준다.

주요어 : 하드웨어 기반 신경망, 플래시 메모리 시냅스 어레이, AND형 어레이, 온라인 학습, 오프라인 학습, 하이브리드 학습
학번 : 2017-22213

List of Publications

Journals

1. **D. Kwon***, K.-H. Lee, S. Y. Woo, J. H. Ko, W. Y. Choi, B.-G. Park, and J.-H. Lee, “Highly Linear Analog Spike Processing Block Integrated with an AND-Type Flash Array and CMOS Neuron Circuits,” *IEEE Transactions on Electron Devices, Accepted.* (Co-First Author)

2. W. Shin, R.-H. Koo, S. Hong, **D. Kwon**, J. Hwang, B.-G. Park, and J.-H. Lee, “Highly Efficient Self-Curing Method in MOSFET Using Parasitic Bipolar Junction Transistor,” *IEEE Electron Device Letters,* vol. 43, no. 7, 2022.

3. S.-T. Lee, H. Kim, H. Yoo, **D. Kwon**, and J.-H. Lee, “Novel, parallel and differential synaptic architecture based on NAND flash memory for high-density and highly-reliable binary neural networks,” Neurocomputing, vol. 498, no. 7, 2022.

4. Shin, J.-H. Bae, **D. Kwon**, R.-H. Koo, B.-G. Park, D. Kwon, and J.-H. Lee, “Investigation of Low-Frequency Noise Characteristics of Ferroelectric Tunnel Junction: from Conduction Mechanism and Scaling Perspectives,” *IEEE Electron Device Letters,* vol. 43, no. 6, 2022.

5. **D. Kwon***, S. Y. Woo, and J.-H. Lee, “Review of Analog Neuron Devices for Hardware-based Spiking Neural Networks,” *Journal of Semiconductor Technology and Science,* vol. 22, no. 2, 2022.

6. J. Im, J. Kim, H. Yoo, J.-W. Baek, **D. Kwon**, S. Oh, J. Kim, J. Hwang, B.-G. Park, and J.-H. Lee, "On-Chip Trainable Spiking Neural Networks Using Time-To-First-Spike Encoding," *IEEE Access,* vol. 10, 2022.

7. S. Oh, **D. Kwon**, G. Yeom, W.-M. Kang, S. Lee, S. Y. Woo, J. Kim, J.-H. Lee, "Neuron Circuits for Low-Power Spiking Neural Networks Using Time-To-First-Spike Encoding," *IEEE Access,* vol. 10, 2022.

8. Shin, K.-K. Min, J.-H. Bae, J.-K. Lee, S. Hong, J. Kim, Y. Jeong, **D. Kwon**, R.-H. Koo, G. Jung, C. Han, J. Kim, B.-G. Park, D. Kwon, and J.-H. Lee, "Synergistic improvement of sensing performance in ferroelectric transistor gas sensors using remnant polarization," *Materials Horizons*, vol. 9, 2022.

9. W. Shin, K.-K. Min, J.-H. Bae, J. Yim, **D. Kwon**, Y. Kim, J. Yu, J. Hwang, B.-G. Park, D. Kwon, and J.-H. Lee, "Comprehensive and accurate analysis of the working principle in ferroelectric tunnel junctions using low-frequency noise spectroscopy," *Nanoscale,* vol. 14, no. 6, 2022.

10. M.-K. Park, H.-N. Yoo, J. Hwang, S. Y. Woo, **D. Kwon**, Y.-T. Seo, J.-H. Lee, and J.-H. Bae, "CMOS-Compatible Low-Power Gated Diode Synaptic Device for Hardware-Based Neural Network," *IEEE Transactions on Electron Devices,* vol. 69, no. 2, 2021.

11. W. Shin, J.-H. Bae, S. Kim, K. Lee, D. Kwon, B.-G. Park, and D. Kwon, and J.-H. Lee, "Effects of high-pressure annealing on the low-frequency noise characteristics in ferroelectric FET," *IEEE Electron Device Letters,* vol. 43, no. 1, 2021.

12. **D. Kwon***, G. Jung, W. Shin, Y. Jeong, S. Hong, S. Oh, J. Kim, J.-H. Bae, B.-G. Park, and J.-H. Lee, "Efficient fusion of spiking neural networks and FET-type gas sensors for a fast and reliable artificial olfactory system," *Sensors and Actuators B: Chemical*, vol. 345, 130419, 2021.
13. **D. Kwon***, S. Y. Woo, J.-H. Bae, S. Lim, B.-G. Park, and J.-H. Lee, "Hardware-Based Spiking Neural Networks Using Capacitor-Less Positive Feedback Neuron Devices," *IEEE Transactions on Electron Devices,* vol. 68, no. 9, 2021.

14. H. Kim, J. Hwang, **D. Kwon**, J. Kim, M.-K. Park, J. Im, B.-G. Park, and J.-H. Lee, "Direct Gradient Calculation: Simple and Variation-Tolerant On-Chip Training Method for Neural Networks," *Advanced Intelligent Systems,* vol. 3, no. 8, 2100064, 2021.

15. J. Kim, **D. Kwon**, S. Y. Woo, W.-M. Kang, S. Lee, S. Oh, C.-H. Kim, J.-H. Bae, B.-G. Park, J.-H. Lee, "On-chip trainable hardware-based deep Q-networks approximating a backpropagation algorithm," *Neural Computing and Applications,* vol. 33, no. 15, 2021.

16. **D. Kwon***, G. Jung, W. Shin, Y. Jeong, S. Hong, S. Oh, J.-H. Bae, B.-G. Park, and J.-H. Lee, "Low-power and reliable gas sensing system based on recurrent neural networks," *Sensors and Actuators B: Chemical,* vol. 340, 129258, 2021.

17. **D. Kwon***, Y.-T. Seo, Y. Noh, S. Lee, M.-K. Park, S. Y. Woo, B.-G. Park, and J.-H. Lee, "3-D AND-Type Flash Memory Architecture With High-k Gate Dielectric for High-Density Synaptic Devices," *IEEE Transactions on Electron Device,* vol. 68, no. 8, 2021. (Co-

First Author)

18. S. Oh, S. Lee, S. Y. Woo, **D. Kwon**, J. Im, J. Hwang, J.-H. Bae, B.-G. Park, and J.-H. Lee, "Spiking Neural Networks With Time-to-First-Spike Coding Using TFT-Type Synaptic Device Model," *IEEE Access*, vol. 9, 2021.

19. **D. Kwon***, W.-M. Kang, S. Y. Woo, S. Lee, H. Yoo, J. Kim, B.-G. Park, and J.-H. Lee, "Hardware-based spiking neural networks using a TFT-type AND flash memory array architecture based on direct feedback alignment," *IEEE Access,* vol. 9, 2021. (Co-First Author)

20. **D. Kwon***, W. Shin, J.-H. Bae, S. Lim, B.-G. Park, and J.-H. Lee, "Impacts of Program/Erase Cycling on the Low-Frequency Noise Characteristics of Reconfigurable Gated Schottky Diodes," *IEEE Electron Device Letters,* vol. 42, no. 6. 2021. (Co-First Author)

21. J. Kim, **D. Kwon**, S. Y. Woo, W.-M. Kang, S. Lee, S. Oh, C.-H. Kim, J.-H. Bae, B.-G. Park, and J.-H. Lee, "Hardware-based spiking neural network architecture using simplified backpropagation algorithm and homeostasis functionality," *Neurocomputing,* vol. 428, 2021.

22. **D. Kwon***, W. Shin, J.-H. Bae, S. Lim, B.-G. Park, and J.-H. Lee, "Investigation of low-frequency noise characteristics in gated Schottky diodes," *IEEE Electron Device Letters,* volt. 42, no. 3, 2021.

23. **D. Kwon***, S. Y. Woo, N. Choi, W.-M. Kang, Y.-T. Seo, M.-K. Park, J.-H. Bae, B.-G. Park, and J.-H. Lee, "Low-power and high-density

neuron device for simultaneous processing of excitatory and inhibitory signals in neuromorphic systems," *IEEE Access,* vol. 8, 2020. (Co-First Author)

24. S.-T. Lee, S. Lim, J.-H. Bae, **D. Kwon**, H. Kim, B.-G. Park, and J.-H. Lee, "Pruning for hardware-based deep spiking neural networks using gated Schottky diode as synaptic devices," *Journal of Nanoscience and Nanotechnology,* vol. 20, no. 11, 2020.

25. S.-T. Lee, S. Lim, N. Choi, J.-H. Bae, **D. Kwon**, H. Kim, B.-G. Park, and J.-H. Lee, "Effect of Word-Line Bias on Linearity of Multi-Level Conductance Steps for Multi-Layer Neural Networks Based on NAND Flash Cells," *Journal of Nanoscience and Nanotechnology,* vol. 20, no. 7, 2020.

26. S.-T. Lee, **D. Kwon**, H. Kim, H. Yoo, and J.-H. Lee, "NAND flash based novel synaptic architecture for highly robust and high-density quantized neural networks with binary neuron activation of (1, 0)," *IEEE Access,* vol. 8, 2020.

27. H. Kim, J.-H. Bae, S. Lim, S.-T. Lee, Y.-T. Seo, **D. Kwon**, B.-G. Park, and J.-H. Lee, "Efficient precise weight tuning protocol considering variation of the synaptic devices and target accuracy," *Neurocomputing,* vol. 378, 2020.

28. W. Shin, G. Jung, S. Hong, Y. Jeong, J. Park, D. Kim, D. Jang, **D. Kwon**, J.-H. Bae, B.-G. Park, and J.-H. Lee, "Proposition of deposition and bias conditions for optimal signal-to-noise-ratio in resistor- and FET-type gas sensors," *Nanoscale,* vol. 12, no. 8, 2020.

29. **D. Kwon***, S. Lim, J.-H. Bae, S.-T. Lee, H. Kim, Y.-T. Seo, S. Oh, J. Kim, K. Yeom, B.-G. Park, and J.-H. Lee, "On-chip training spiking neural networks using approximated backpropagation with analog synaptic devices," *Frontiers in Neuron science,* vol. 423, 2020.

30. 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, no. 11, 2019.

31. S.-T. Lee, S. Lim, N.-Y. Choi, J.-H. Bae, **D. Kwon**, B.-G. Park, and J.-H. Lee, "Operation scheme of multi-layer neural networks using NAND flash memory as high-density synaptic devices," *IEEE Journal of Electron Devices Society,* vol. 7, 2019.

32. J.-H. Bae, S. Lim, **D. Kwon**, S.-T. Lee, H. Kim, and J.-H. Lee, "Gated Schottky Diode-Type Synaptic Devices with a Field-Plate Structure to Reduce the Forward Current," *Journal of Nanoscience and Nanotechnology,* vol. 19, no. 10, 2019.

33. S. Y. Woo, K.-B. Choi, S. Lim, S.-T. Lee. C.-H. Kim, W.-M. Kang, **D. Kwon,** J.-H. Bae, B.-G. Park, and J.-H. Lee, "Synaptic device using a floating fin-body MOSFET with memory functionality for neural network," *Solid-State Electronics,* vol. 156, 2019.

34. **D. Kwon***, S. Lim, J.-H. Eum, S.-T. Lee, J.-H. Bae, H. Kim, C.-H. Kim, B.-G. Park, and J.-H. Lee, "Highly Reliable Inference System of Neural Networks Using Gated Schottky Diodes," *IEEE Journal of the Electron Devices Society,* vol. 7, 2019. (Co-First Author)

35. J.-H. Bae, S. Lim, **D. Kwon**, J.-H. Eum, S.-T. Lee, H. Kim, B.-G. Park, and J.-H. Lee, "Near-linear potentiation mechanism of gated Schottky diode as a synaptic device," *IEEE Journal of the Electron Devices Society,* vol. 7, 2019.

36. J.-H. Bae, H. Kim, **D. Kwon**, S. Lim, S.-T. Lee, B.-G. Park, and J.-H. Lee, "Reconfigurable field-effect transistor as a synaptic device for XNOR binary neural network," *IEEE Electron Device Letters,* vol. 40, no. 4, 2019.

37. **D. Kwon***, S. Lim, J.-H. Bae, S.-T. Lee, H. Kim, B.-G. Park, and J.-H. Lee, "Adaptive weight quantization method for nonlinear synaptic devices," *IEEE Transactions on Electron Devices,* vol. 66, no. 1, 2018.

38. C.-H. Kim, S. Lim, S. Y. Woo, W.-M. Kang, Y.-T. Seo, S.-T. Lee, S. Lee, D.Kwon, S. Oh, Y. Noh, H. Kim, J. Kim, J.-H. Bae, and J.-H. Lee, "Emerging memory technologies for neuromorphic computing," *Nanotechnology,* vol. 30, no. 3, 2018.

Conferences

1. **D. Kwon***, S.-H. Park, H.-N. Yoo, J.-W. Back, J. Hwang, Y. Yang, J.-J. Kim, and J.-H. Lee, "Retention Improvement in Vertical NAND Flash Memory Using 1-bit Soft Erase Scheme and its Effects on Neural Networks," *2022 International Electron Devices Meeting (IEDM)* (Co-First Author)

2. H.-N. Yoo, J.-W. Back, N.-H. Kim, **D. Kwon**, B.-G. Park, and J.-H. Lee, "First Demonstration of 1-bit Erase in Vertical NAND Flash Memory," *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI*

Technology and Circuits).

3. W. Shin, S. Hong, Y. Jeong, G. Jung, J. Park, **D. Kwon**, D. Jang, D. Kim, B.-G. Park, and J.-H. Lee, "Efficient Improvement of Sensing Performance Using Charge Storage Engineering in Low Noise FET-type Gas Sensors," *2020 IEEE International Electron Devices Meeting (IEDM).*

4. Y. Jeong, W. Shin, S. Hong, G. Jung, J. Park, D. Jang, D. Kim, **D. Kwon**, B.-G. Park, and J.-H. Lee, "Highly Sensitive Amplifier Circuit Consisting of Complementary *p*FET-type and Resistor-type Gas Sensors," *2020 IEEE International Electron Devices Meeting (IEDM).*

5. S.-T. Lee, H. Kim, J.-H. 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).*

6. **D. Kwon***, S. Lim, J.-H. Bae, S.-T. Lee, H. Kim, Y.-T. Seo, G. Yeom, B.-G. Park, and J.-H. Lee, "Investigation of Adaptive Weight Quantization in Application for Convolutional Neuromorphic System Using Gated Schottky Diode," *2019 International Conference on Solid State Devices and Materials (SSDM).*

7. J.-H. Lee, S. Y. Woo, S.-T. Lee, S. Lim, W.-M. Kang, Y.-T. Seo, S. Lee, **D. Kwon**, S. Oh, Y. Noh, H. Kim, J. Kim, and J.-H. Bae, "Review of candidate devices for neuromorphic applications," *ESSDERC 2019-49th European Solid-State Device Research Conference (ESSDERC).*

8. **D. Kwon*,** S. Lim, S.-T. Lee, H. Kim, J.-H. Bae, and J.-H. Lee, "Investigation of neural networks using synapse arrays based on gated Schottky diodes," *2019 International Joint Conference on Neural*

Networks (IJCNN). (Co-First Author)

9. S.-T. Lee, S. Lim, N. Choi, J.-H. Bae, **D. Kwon**, H. Kim, B.-G. Park, and J.-H. Lee, "Input Voltage Scheme for DOT Product Engine Using NAND Flash Cells," *2019 China Semiconductor Technology International Conference (CSTIC).*

10. S.-T. Lee, S. Lim, J.-H. Bae, **D. Kwon**, H. Kim, B.-G. Park, and J.-H. Lee, "Dot Product Engine Using Gated Schottky Diode with Quantized Weight," *2019 Electron Devices Technology and Manufacturing Conference (EDTM)*

Honors

1. Gold Prize, The 28th Humantech Thesis contest, Samsung Electronics, Feb. 2022.
2. Bronze Prize, IEEE Seoul Section Student Conference, Oct. 2021.