저작자표시-비영리-변경금지 2.0 대한민국

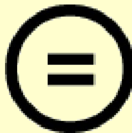
이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

**저작자표시.** 귀하는 원저작자를 표시하여야 합니다.

**비영리.** 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.


**변경금지.** 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

**저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.**

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

Ph.D. DISSERTATION

# $SiO_2$ Fin-based AND-type Flash Synaptic Array for Hardware-based Neural Networks

# 하드웨어 기반 신경망을 위한 $SiO_2$ 핀 기반 AND-형 플래시 시냅스 어레이

by

SOOCHANG LEE

August 2022

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

# $SiO_2$ Fin-based AND-type Flash Synaptic Array for Hardware-based Neural Networks

# 하드웨어 기반 신경망을 위한 $SiO_2$ 핀 기반 AND-형 플래시 시냅스 어레이

지도교수 최 우 영

이 논문을 공학박사 학위논문으로 제출함

2022 년 8 월


서울대학교 대학원

전기정보공학부

이 수 창


이수창의 공학박사 학위논문을 인준함

2022 년 8 월

| | | |
|---|---|---|
| 위 원 장 : | 김 재 하 | (인) |
| 부위원장 : | 최 우 영 | (인) |
| 위 원 : | 이 종 호 | (인) |
| 위 원 : | 김 윤 | (인) |
| 위 원 : | 배 종 호 | (인) |

# $SiO_2$ Fin-based AND-type Flash Synaptic Array for Hardware-based Neural Networks

by

Soochang Lee

Advisor: Woo Young Choi

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

Doctoral Committee:

Professor Jaeha Kim, Chair
Professor Woo Young Choi, Vice-Chair
Professor Jong-Ho Lee
Professor Yoon Kim
Professor Jong-Ho Bae

# ABSTRACT

Neuromorphic computing systems have emerged as a novel artificial intelligence paradigm to overcome the von Neumann bottleneck by mimicking the biological nervous system. Synaptic devices for hardware-based neural networks (HNNs) in neuromorphic computing systems require parallel computability, high scalability, low-power operation, and selective write operation. In this work, a $SiO_2$ fin-based AND flash memory synaptic device for a HNN is proposed. The proposed device having a round-shaped channel structure with a 6 nm-wide thin oxide fin improves program performance compared to a flash synaptic device with planar-type channel by locally enhancing electric fields. The AND flash cell shows a high on/off current ratio over $10^5$, a low sup-pA off-current, and a high dynamic range of synaptic weights over $10^3$ with a low program voltage below 9 V. Selective write operation is performed using program and erase inhibition pulse schemes in the fabricated AND array based on $SiO_2$ fin, and weighted sum operation is experimentally verified. In addition, a 3D AND flash synaptic array with round-

shaped poly-Si channel is designed and fabricated to improve scalability. Key fabrication steps are proposed to address misalignment issues. The proposed 3D AND array performs selective write operation using program and erase inhibition pulse schemes.

A novel synaptic architecture with two AND flash memory cells for off-chip learning is proposed. The novel synapse structure based on AND flash cells is used to perform parallel XNOR operation and bit-counting for binary neural networks (BNNs). Proposed BNN based on the AND flash array structure exhibits a classification accuracy of 89.9% on CIFAR-10 dataset, comparable to that of an ideal software-based BNN. Furthermore, differential synaptic architecture using AND flash array is proposed to improve robustness against on-current retention loss.

**Keywords**: AND flash memory, synaptic device, fin-type flash device, 3D flash memory, hardware-based neural network, binary neural network, neuromorphic system.

**Student Number**: 2016-23288

# CONTENTS

**Abstract.................................................................i**

**Contents.................................................................iii**

**List of Figures.........................................................vi**

**List of Tables.........................................................xiii**

**Chapter 1**

**Introduction.............................................................1**

1.1 Neuromorphic computing.................................................1

1.2 Synaptic devices...........................................................3

1.3 Purpose of research.......................................................5

1.4 Dissertation outline.......................................................7

**Chapter 2**

**$SiO_2$ fin-based AND flash synaptic array..........................8**

2.1 Device structure........................................................................8

2.2 Fabrication process....................................................................10

2.3 Cell characteristics....................................................................16

2.4 Array characteristics..................................................................25

**Chapter 3**

**3D AND flash synaptic array with rounded channel.......34**

3.1 Device structure........................................................................34

3.2 Fabrication process....................................................................37

    3.2.1 Cell process steps................................................................39

    3.2.2 WL contact pad process steps..............................................54

3.3 Cell characteristics....................................................................57

3.4 Array characteristics..................................................................62

**Chapter 4**

**Off-chip learning based on AND flash synaptic**

**Array.......................................................................72**

4.1 Binary neural networks based on AND flash synaptic array.........72

4.1.1 AND flash synaptic architecture.........................................72

4.1.2 Differential synaptic architecture........................................80

4.2 Quantized neural networks based on AND flash synaptic array...84

**Chapter 5**

**Conclusion...................................................................86**

**Bibliography...............................................................89**

**Abstract in Korean......................................................96**

# List of Figures

Figure 1.1. Schematics of (a) von Neumann architecture and (b) deep neural networks....2

Figure 1.2. Schematics of (a) NOR-type flash memory array and (b) AND-type flash memory array....4

Figure 2.1. (a) 3D schematic of proposed oxide fin-based AND array. (b) Cross-sectional schematic of the single cell....9

Figure 2.2. (a)-(i) Cross-sectional and 3D schematics of the main fabrication steps, and (j) device fabrication flow....14

Figure 2.3. (a) Cross-sectional TEM image of the fin-based synaptic device fabricated and (b) magnified cross-sectional TEM image of a red dotted region in (a)....15

Figure 2.4. (a) Measured program and (b) erase properties of the fabricated $SiO_2$ fin-based AND flash memory....21

Figure 2.5. (a) ISPP characteristics of a fin-type synaptic device. (b) Comparison of ISPP properties between a fin-type device and a planar-type device. (c) Electric

field's distribution of fin-type and planar-type flash cells at $V_G$ = 8 V. (d) ISPP results of the fin-type synaptic device with varying oxide fin widths......................22

Figure 2.6. (a), (b) Program and erase characteristics of a synaptic cell obtained by identical write pulses. (c) Potentiation and depression characteristics of the device.....................................................................................................23

Figure 2.7. Fig. 2.7. (a), (b) RT and 85℃ retention properties of the $SiO_2$ fin-based AND flash synaptic cell with different synaptic weight ranges. (c) Cycling characteristics of the fabricated $SiO_2$ fin-based flash memory................................24

Figure 2.8. (a) Top SEM image of the fabricated 2×2 AND flash synaptic array. (b) 3D schematic view of the fabricated $SiO_2$ fin-based AND flash synaptic array. (c) Schematic of the 2×2 AND flash synaptic array..................................................29

Figure 2.9. (a) Bias conditions for a single-cell (cell A) selective program operation. (b) Cell program and program inhibition properties in the AND synaptic array......30

Figure 2.10. (a) Bias conditions for a single-cell (cell A) selective erase operation. (b) Cell erase and erase inhibition properties in the AND synaptic array................31

Figure 2.11. (a) Top SEM image of the fabricated 10×10 AND flash synaptic array. (b) Weighted sum current and each cell current along a BL (BL 3) of the 10×10

synaptic array. (c) Comparison between weighted sum current and the sum of each cell current along BL 3 of the 10×10 synaptic array at $V_{read}$ =2 V...........................32

Figure 2.12. (a) Top SEM image of the fabricated 24×8 AND flash synaptic array. (b) Comparison between weighted sum current and the sum of each cell current along BL 3 of 24×8 synaptic array at $V_{read}$ =2 V. (c) Current sum error as a function of the number of inputs..........................................................................................33

Figure 3.1 (a) 3D schematic of 3D AND flash synaptic array and cell structure. (b) Cross-sectional schematics of the 3D AND flash cell. (c) Unit cell area of the proposed 3D AND flash device.........................................................................36

Figure 3.2. Misalignment improvements using one-shot patterning compared to three-shot patterning...............................................................................................38

Figure 3.3. Key fabrication steps of the proposed 3D AND array with rounded channel................................................................................................................39

Figure 3.4. Schematics of multi-layer ON stack and filled oxide in WL cut area.......................................................................................................................45

Figure 3.5. Schematics of (a) via holes and trench patterning, and (b) dummy poly-Si filling in holes and trench patterned...................................................................46

Figure 3.6. (a) Schematics of partial nitride wet etching. (b) X-cut and (b) Y-cut cross-sectional SEM image after partial nitride wet etch process............................47

Figure 3.7. Schematics of (a) separated channel formation by layer and (b) following passivation process. (c) Y-cut cross-sectional SEM image after the passivation process..........................................................................................48

Figure 3.8. (a) Schematics of dummy poly-Si etching inside BL/SL holes and following nitride partial etching process. (b) X-cut cross-sectional SEM image after the $Si_3N_4$ partial etching process...........................................................................49

Figure 3.9. (a) Schematics of $n^+$-doped poly-Si deposition and poly-Si etch-back process. (b) X-cut cross-sectional SEM image after the etch-back process.............50

Figure 3.10. Schematics of dummy poly-Si etching inside WL trench and partial nitride wet-etching inside WL trench......................................................................51

Figure 3.11. Schematics of O/N/A gate insulator stack deposition and WL formation.................................................................................................................52

Figure 3.12. (a) Schematics of CMP process. (b) Y-cut cross-sectional SEM image after the CMP process...................................................................................53

Figure 3.13. WL contact pad fabrication steps.......................................................55

Figure 3.14. Schematic plane views of WL of each floor and cross-sectional view of the contact pad area........................................................................................56

Figure 3.15. (a) Cross-sectional TEM image of the proposed 3D AND flash device. (b) Magnified cross-sectional TEM image of a red dashed box in (a)....................59

Figure 3.16. (a) ISPP and (b) ISPE characteristics of the fabricated 3D AND flash device with round-shaped channel........................................................................60

Figure 3.17. (a), (b) Program and erase characteristics of 3D AND synaptic device obtained by identical write pulses. (c) Potentiation and depression characteristics of the 3D AND device......................................................................................61

Figure 3.18. (a) 3D schematic diagram and (b) top SEM image of the fabricated 2 × 1 × 3 AND flash array. Bias condition for selective (c) erase and (d) program operations..................................................................................................66

Figure 3.19. (a) Measured selective erase properties of a 3D AND synaptic array in Z-direction. (b) Change of threshold voltages of cells in 2 × 1 × 3 AND flash array when selective erase is carried out........................................................................67

Figure 3.20. (a) Measured selective program properties of a 3D AND synaptic array in Z-direction. (b) Change of threshold voltages of cells in 2 × 1 × 3 AND flash array when selective program is carried out.........................................................68

Figure 3.21. (a) Top SEM image of the 4 × 2 × 3 AND array. (b) 3D schematic of the 4 × 2 × 3 AND array....69

Figure 3.22. (a) Bias condition for selective erase operation in the 4 × 2 × 3 AND array. (b) Selective erase properties of the 3D AND array in the XY-plane. (c) Change of threshold voltages of cells in the 3D AND array in selective erase operation....70

Figure 3.23. (a) Bias condition for selective program operation in the 4 × 2 × 3 AND array. (b) Selective program properties of the 3D AND array in the XY-plane. (c) Change of threshold voltages of cells in the 3D AND array in selective program operation....71

Figure 4.1. (a) Schematic diagram of synapse architecture based on two flash devices in an AND array. (b) Truth table of XNOR operation using an AND flash-based synaptic architecture....76

Figure 4.2. (a) Schematic of AND array structure for BNNs. (b) A current-latch based CSA in the proposed BNNs....77

Figure 4.3. (a) Structure of a VGG-9 based on AND flash arrays. (b) Effect of dynamic range on recognition rate of CIFAR-10. (c) Effect of retention loss of cells on recognition accuracy of CIFAR-10....78

Figure 4.4 (a) Differential synaptic architecture using two AND flash cells. (b) A DSA comparing $I_{BL, ODD}$ and $I_{BL, EVEN}$. (c) Schematic of AND flash array using the DSA for BNNs........................................................................................82

Figure 4.5. (a) Effect of dynamic range on recognition accuracy of CIFAR-10 using the differential synaptic architecture. (b) Effect of retention loss of cells on classification accuracy of CIFAR-10. (c) Effect of device variation in BNNs on recognition accuracy on CIFAR-10..........................................................83

Figure 4.6. (a) Recognition accuracy on CIFAR-10 using a QNN based on fin-type AND synaptic array. (b) Effect of device variation in QNNs on classification rates of CIFAR-10........................................................................................85

# List of Tables

Table. 4.1. VGG-9 for CIFAR-10.........................................................................79

# Chapter 1
# Introduction

## 1.1 Neuromorphic computing

In the era of exponential data growth, current von Neumann based information processing systems are dramatically challenged by issues of speed and power consumption [1], [2]. In order to address the limitation of the classical computing architecture as shown in Fig. 1.1(a), brain-inspired neuromorphic computing system has been proposed to achieve low-power operation [3]–[8]. Neuro-inspired computing architecture has shown outstanding performance in computationally demanding tasks such as image or speech recognition and classification based on deep learning algorithms (Fig. 1.1(b)) [9]–[13]. In this non-von Neumann architecture, in-memory computing is performed, exhibiting fast processing speed and low-power computing required for edge devices. With the help of software, neuromorphic systems based on deep neural networks (DNNs) using the backpropagation algorithm have been highlighted for its excellent computational

capability and simplicity of the simulation [15], [16]. The in-memory computing is desirable to carry out real-time big data processing, which takes advantage of massive parallel computations in DNNs. The basic computing in DNNs is the vector-by-multiplication (VMM). In order to reduce power consumption of neuromorphic systems, hardware-based crossbar arrays are exploited to perform VMM operations by Ohm's law and Kirchhoff's current law.

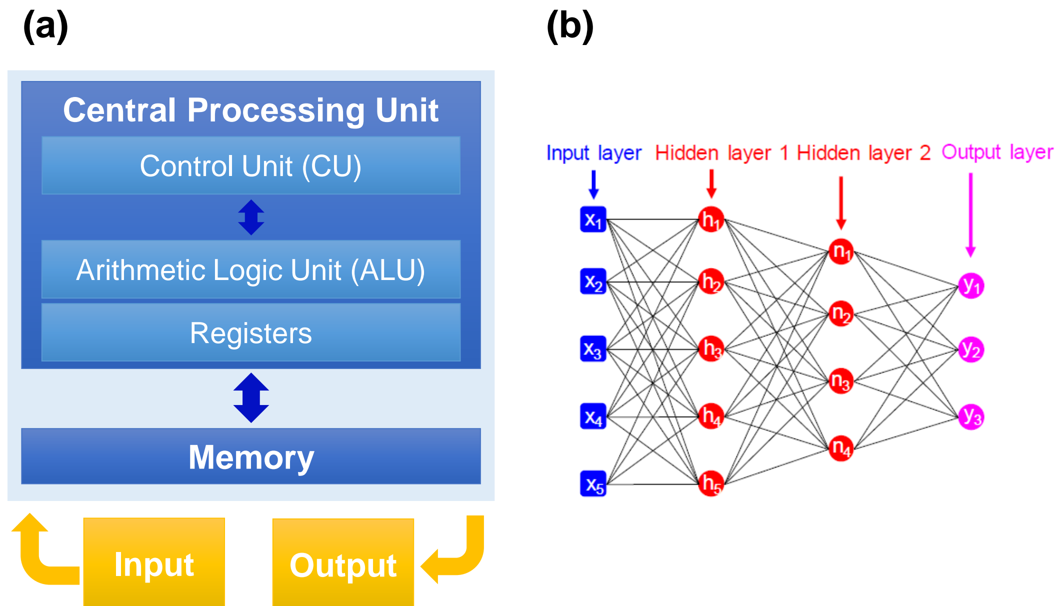



Fig. 1.1. Schematics of (a) von Neumann architecture and (b) deep neural networks.

## 1.2 Synaptic devices

Recently, hardware-based neural networks (HNNs) have been proposed to support parallel computing using deep learning algorithms [17], [18]. In order to implement HNN systems, emerging synaptic devices including resistive switching random access memory (RRAM), phase change memory (PCM) have been suggested to form energy efficient and scalable crossbar arrays which can be simply adopted for parallel computations [19]–[32]. However, the crossbar array structure of the 2-terminal passive devices suffers from a sneak-path current problem leading to incorrect output results. Although 1-transistor/1-resistor structure or additional selectors are used to address this problem, scalability or process complexity can be other bottlenecks of the solutions. In addition these devices still have reliability issues including device characteristic variation, which should be alleviated to create massive synaptic arrays.

On the other hand, flash synaptic devices based on highly mature and CMOS compatible technology [33]–[43], such as NOR and AND flash memory devices, have been proposed to perform parallel processing. Although NOR flash-based

synaptic arrays can easily achieve a fast and selective program operation using a hot carrier injection at the drain side, their high bit line (BL) current during program operations is not desirable to implement a low power hardware neuromorphic system. Unlike a NOR flash synaptic array showing a program operation using a channel-hot-electron (CHE) injection, an AND flash memory synaptic array uses Fowler-Nordheim (FN) tunneling-based programming, which performs a highly energy-efficient program operation.

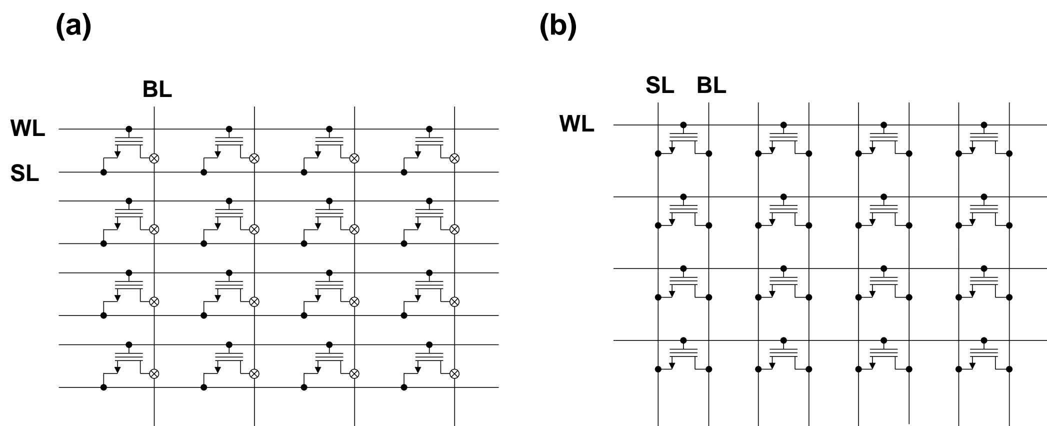



Fig. 1.2. Schematics of (a) NOR-type flash memory array and (b) AND-type flash memory array.

## 1.3 Purpose of research

As aforementioned, an AND-type flash array has several advantages including low power selective program/erase characteristics and parallel computation. AND flash devices based on highly mature and CMOS compatible technology have been suggested as reliable synaptic devices to form energy efficient and scalable synaptic arrays in massive neural networks. In particular, 3D AND array models have been recently reported to achieve high scalability for HNNs.

In this work, we propose and fabricate a high-density $SiO_2$ fin-based AND flash memory array. Compared to a planar-type flash cell, the proposed AND flash cell with a fin-type round channel is shown to have a larger dynamic range and lower program voltage using FN tunneling. Then selective program and erase schemes in the $SiO_2$ fin-based AND flash array proposed are described and experimentally carried out.

Additionally, a novel 3D stackable AND flash synaptic array with round-shaped channel is proposed to improve scalability. We suggests efficient fabrication method to alleviate misalignment issues. Memory characteristics of 3D AND array is also

investigated to verify synaptic properties. As in 2D AND array, selective program and erase scheme is introduced and experimentally implemented.

Then a novel synapse design using two adjacent flash cells in AND synaptic arrays is proposed and analyzed to implement XNOR behaviors and bit-counting operation in a parallel fashion for BNNs. BNNs based on the proposed AND flash cells with large dynamic range pave the way for low-power off-chip neuromorphic devices that can achieve high classification performance comparable to the software baseline. Furthermore, differential synaptic architecture for BNNs is proposed to achieve stable off-chip learning performance against retention loss. Classification performance of QNN based on AND array is investigated to verify applicability of analog synaptic properties.

## 1.4 Dissertation outline

The structure of this dissertation is as follows: Chapter 1 provides an overview of neuromorphic computing and related synaptic devices. The purpose of the research and outline of this dissertation are introduced. Chapter 2 describes the proposed $SiO_2$ fin-based AND flash memory array. In this chapter, the device structure, fabrication process, and characteristics of the synaptic device are presented. Moreover, cell measurement results and array measurement results are also presented. Chapter 3 describes the proposed 3D AND flash memory array with round-shaped channel. This chapter includes the device structure, fabrication process, and characteristics of the synaptic device. 3D AND array measurements results are also presented in this chapter. Chapter 4 describes the hardware-based BNN introducing novel two cell-based AND-type synaptic arrays. In addition, differential synaptic architecture using AND flash array is proposed to show advanced classification performance. Lastly, chapter 5 summarizes the dissertation.

# Chapter 2

# $SiO_2$ fin-based AND flash synaptic array

## 2.1 Device structure

A $SiO_2$ fin-based flash synaptic device in an AND array is designed to have a round-shaped channel structure with high curvature by using a thin oxide fin with a thickness ranging from a few to several tens of nanometers. Fig. 2.1 shows a 3D schematic of proposed $SiO_2$ fin-based AND synaptic array and a schematic cross-sectional view of a $SiO_2$ fin-based AND flash memory cell. A $SiO_2$ fin is used to separate a drain line and a source line (SL), resulting in a fin-type round poly-Si channel that crosses over between them. The major merits of the proposed synaptic device are the following: (1) the rounded shape of the channel makes it possible to do programming with low voltage due to its gate-all-around (GAA)-like structure; the detailed analysis is shown in the next section. (2) The round-shaped poly-Si channel device can exhibit a low synaptic current with increasing the oxide fin height. (3) The $SiO_2$ fin-based AND flash memory device shows a smaller effective

cell size (~$6F^2$) compared to a NOR flash device (~$10F^2$) with the help of a small footprint of the oxide fin in it. A bit-line (BL) linking drain nodes and a SL connecting source nodes are laid out in parallel to form the AND flash array. (4) Selective program/erase properties can be readily obtained by using FN tunneling due to its parallel array design. Moreover, a $SiO_2$/ $Si_3N_4$/$Al_2O_3$ (O/N/A) gate stack is located between a poly-silicon channel and a TiN metal gate to store synaptic weights and improve programming/erasing window.

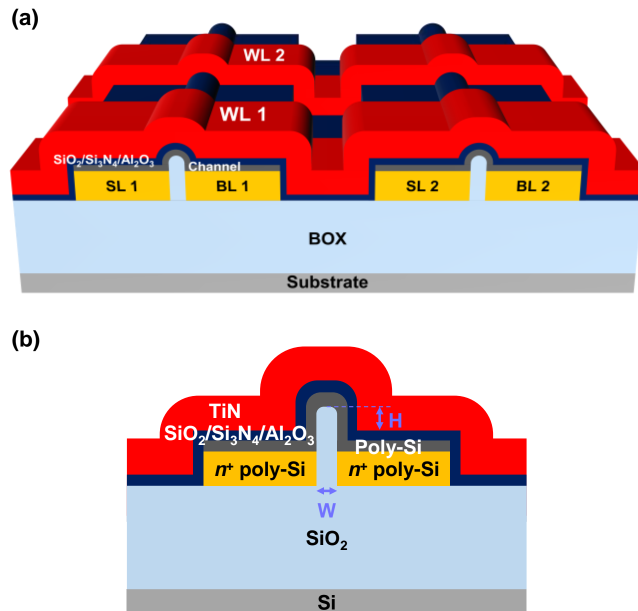


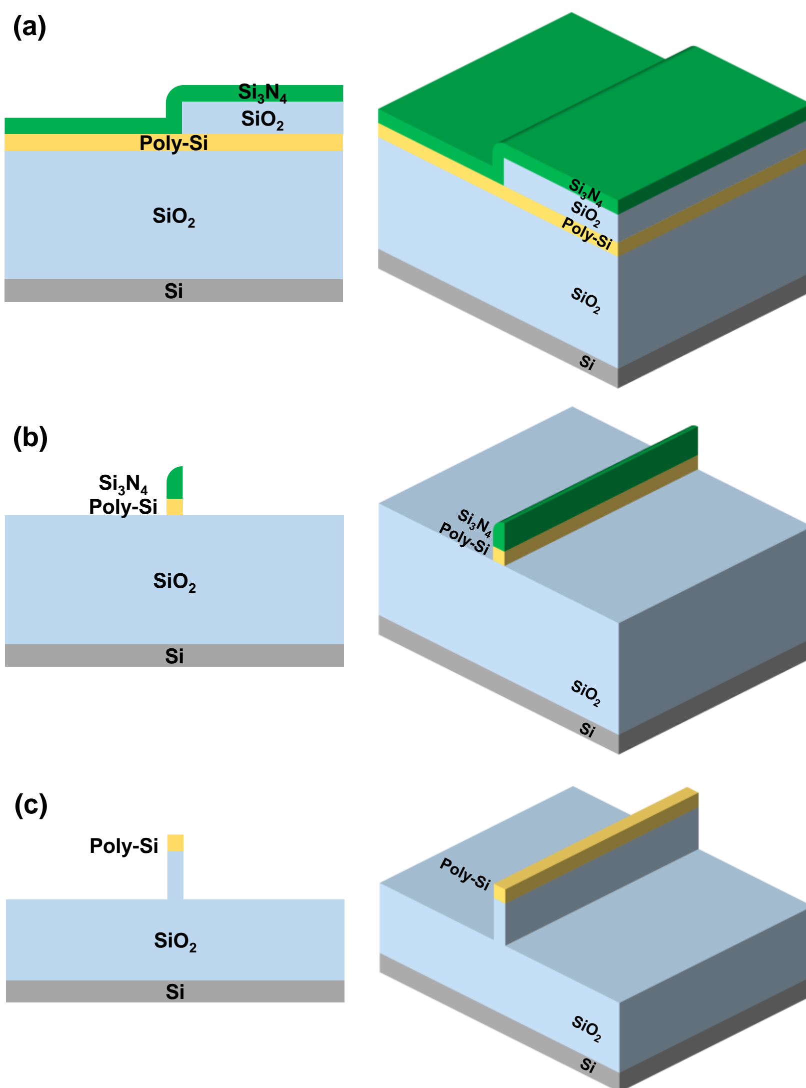

Fig. 2.1. (a) 3D schematic of proposed oxide fin-based AND array. (b) Cross-sectional schematic of the single cell.

## 2.2 Fabrication process

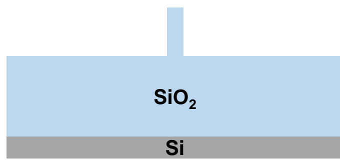
Fig. 2.2 depicts main fabrication steps for an AND flash memory cell based on a $SiO_2$ fin. The AND-type synaptic cell and array are fabricated using six masks. First, a 430 nm-thick $SiO_2$ buried oxide is grown on the Si substrate by wet oxidation at 1000 °C for 80 min, followed by sequential formation of poly-Si and $SiO_2$ layers. The 150 nm-poly-Si on the buried oxide layer is deposited by low pressure chemical vapor deposition (LPCVD), and the 120 nm-$SiO_2$ layer is grown on the poly-Si layer via wet oxidation at 1000 °C for 15 min. The $SiO_2$ layer on the poly-Si layer is patterned and then a 60 nm-$Si_3N_4$ layer is formed by LPCVD (Fig. 2.2(a)). The $Si_3N_4$ layer is anisotropically etched to form a sidewall as a hard mask, followed by an selective wet etching of remained $SiO_2$ on the poly-Si film using buffered oxide etchant ($NH_4F$:HF = 7:1). After the poly-Si patterning utilizing the $Si_3N_4$ hard mask (Fig. 2.2(b)), the patterned poly-Si is used as a hard mask to etch a section of the $SiO_2$ film, producing an oxide fin. (Fig. 2.2(c)). Isotropic etching is then used to remove the poly-Si hard mask as shown in Fig. 2.2(d). A layer of *in situ* $n^+$-doped poly-Si, 150 nm in thickness, is deposited for SLs and BLs by LPCVD

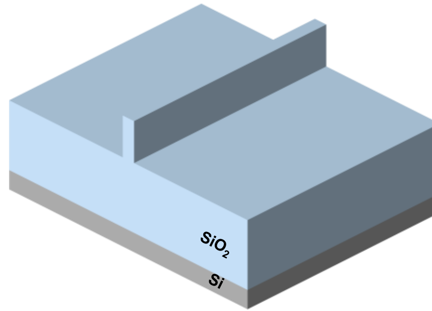
(Fig. 2.2(e)). Then chemical mechanical polishing (CMP) is performed, as shown in Fig. 2.2(f), then a chemical dry etch of the poly-Si surface is used to extrude a fin-oxide onto the planarized $n^+$-doped poly-Si surface (Fig. 2.2(g)). After an amorphous silicon (*a*-Si) layer has been deposited by the LPCVD method, it is re-crystallized at 600 °C for 24 hours to create an undoped poly-Si channel. A $SiO_2$ /$Si_3N_4$/$Al_2O_3$ (O/N/A: 2.8/4.5/6.0 nm) gate dielectric stack is deposited after the poly-Si active layer is patterned. The tunneling oxide $SiO_2$ and the charge trapping layer $Si_3N_4$ are deposited by LPCVD process and the blocking oxide of alumina is formed by atomic layer deposition (ALD) process. A TiN metal gate is patterned by isotropic etching of the conformal TiN layer deposited using metal-organic chemical vapor deposition (MOCVD) process (Fig. 2.2(i)). Lastly, the back end of line (BEOL) process is performed. Fig. 2.3 shows transmission electron microscope (TEM) cross-sectional images of the fabricated $SiO_2$ fin-based device. The fin oxide width is 6 nm, channel thickness is 8 nm, and effective channel length is around 80 nm. Note that a further wet etching of the oxide fin with diluted HF (DHF, DIW:HF = 100:1) can be used to modify the oxide fin's width before the poly-Si channel

formation.

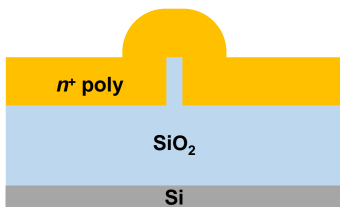


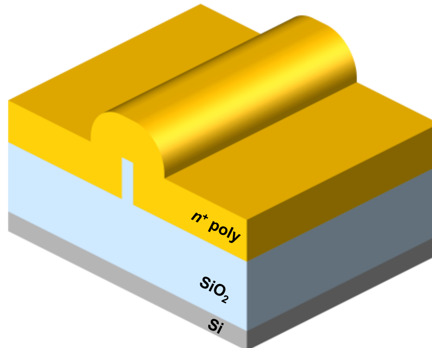
(d)



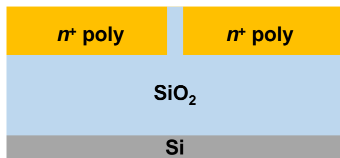



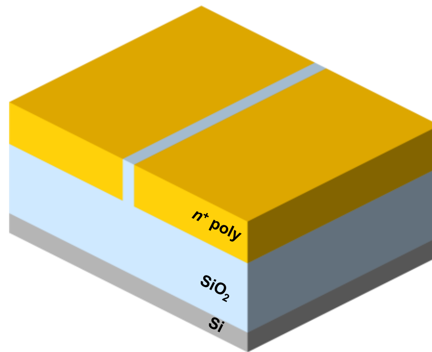

(e)



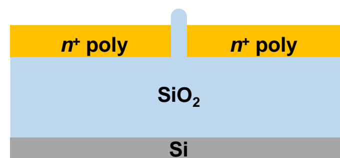



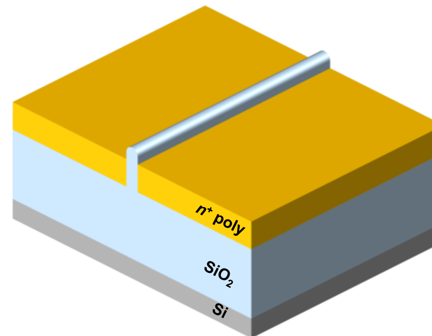

(f)




(g)



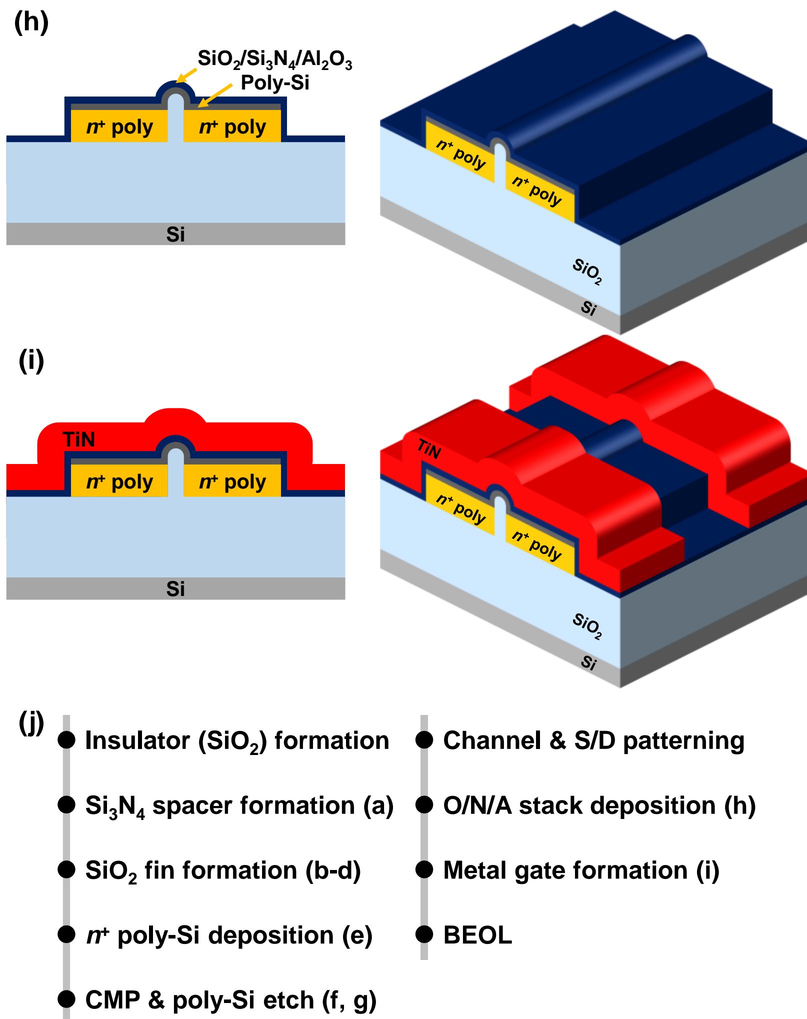



Fig. 2.2. (a)-(i) Cross-sectional and 3D schematics of the main fabrication steps, and (j) device fabrication flow.

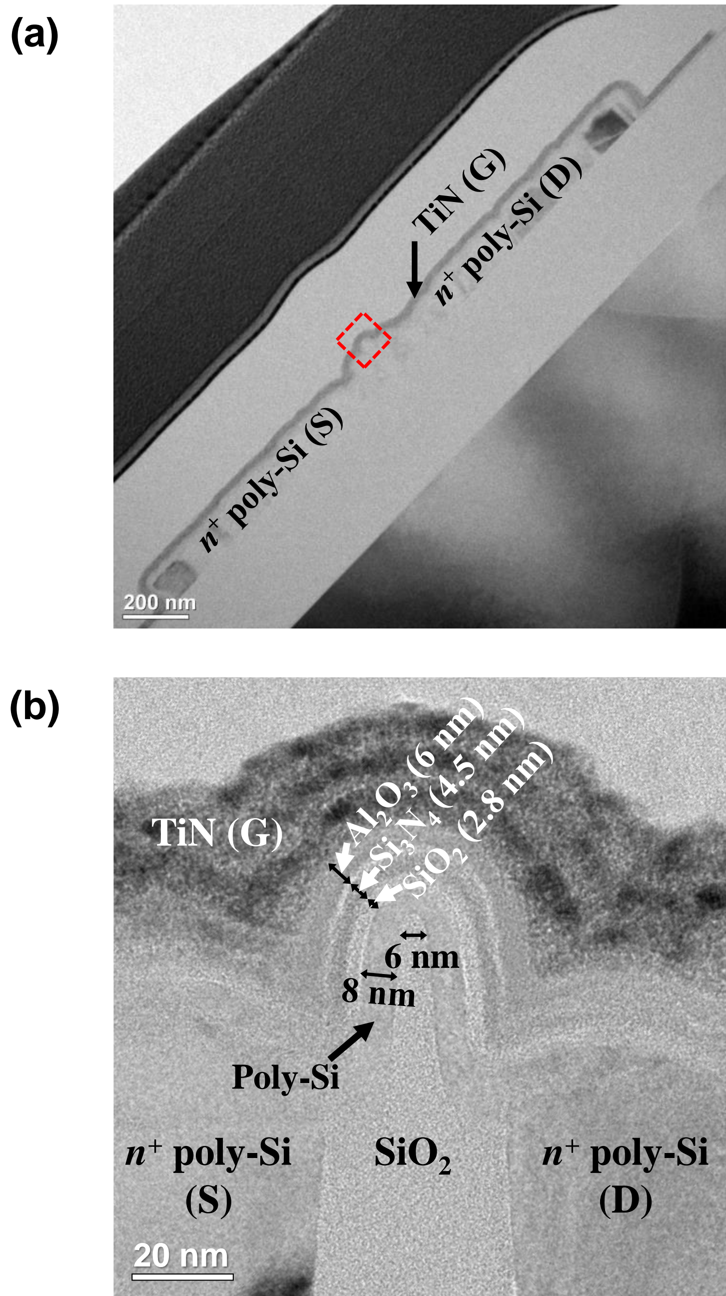



Fig. 2.3. (a) Cross-sectional TEM image of the fin-based synaptic device fabricated and (b) magnified cross-sectional TEM image of a red dotted region in (a).

## 2.3 Cell characteristics

In order to assess the synaptic properties of the suggested $SiO_2$ fin-based AND synaptic cell, the program/erase properties of the manufactured cell are measured and analyzed.

Fig. 2.4 shows measured $I_D$-$V_G$ curves of a single cell in program ($V_G$ = 5~8 V, $V_S$ = $V_D$ =0 V, $t$ = 100 μs) and erase operation ($V_S$ = $V_D$ = 7 V, $V_G$ =0 V, $t$ = 10 ms). Due to a parallel BLs and SLs of AND-type arrays, FN-programming and erasing are carried out by applying same P/E voltage to a source and a drain while applying high or low voltage to a gate node. Note that only positive program or erase voltages are applied to the gate or the source and drain in program and erase operation, respectively. The fabricated fin-type flash synaptic device shows a high on-off current ratio over $10^5$ as well as sub-pA off-current. As shown in Fig. 2.4(a), a high maximum-minimum synaptic conductance ratio over $10^3$ is also obtained by using a low incremental-step-pulse programming (ISPP) voltage under 9 V. High on-off current ratio over $10^5$ is provided, and a low program and erase voltages below 7 V can be exploited to control analog synaptic weights using the proposed synaptic

device.

Fig. 2.5 (a) illustrates that simulated $I_D$-$V_G$ curves of a proposed device in ISPP operations ($V_G$ = 5~8 V, $V_S$ = $V_D$ = 0 V, $t$ = 100 μs) are well fitted with measurement results using Sentaurus TCAD simulation tool. In order to consider the grain boundary properties in the poly-Si channel, simulation parameters such as the mobility, lifetime of the electron and hole in the poly-Si channel are calibrated to 40 $cm^2V^{-1}s^{-1}$ and 25 $cm^2V^{-1}s^{-1}$, 1 ns and 0.3 ns, respectively [44]-[46]. Work function of the TiN metal gate, effective tunneling mass of electron and hole in $SiO_2$ tunneling oxide layer are also calibrated to 4.11 eV, 0.35 $m_0$ and 0.38 $m_0$ [47]. The poly-Si/$SiO_2$ interface trap density is also calibrated to $4.5\times10^{12}$ $cm^2$ [44].

Using the simulation parameters of the device fabricated, the programming characteristics of a planar-type device are compared to those of a fin-type device. Both simulated and measured program characteristics in Fig. 2.5(b) indicate that the fin-type flash device shows higher programming efficiency than the planar-type flash device. Therefore, the proposed fin-based flash device efficiently reduces a required program voltage to obtain the necessary memory window. This decrease

in program voltage derives from the GAA structure's local electric field enhancement effect, as illustrated in Fig. 2.5(c) [48]. Note that a lower local electric field across a blocking oxide in the fin-type device as compared to the planar-type device reduces electron back-tunneling in an initial erase operation even more. Fig. 2.5(d) exhibits that reducing the oxide fin width improves the programming efficiency of the proposed device as a result of the increased local field across the tunneling oxide. As shown in Fig. 2.3(b), in order to maximize program efficiency, a thin oxide fin with a thickness of 6 nm was used to incorporate process conditions into the fabricated device.

Fig. 2.6 shows measured synaptic characteristics of the fabricated single synaptic cell obtained by applying identical erase and program pulses one hundred times respectively. By checking the number of applied program or erase pulses, multi-level analog synaptic conductance can be obtained. >1 order of magnitude of maximum-minimum synaptic conductance ratio is obtained at <7 V low program ($V_{PGM}$ = 6.5 V, $t$ = 100 μs) and erase voltages ($V_{ERS}$ = 7 V, $t$ = 10 ms).

Furthermore, retention characteristics of the proposed fin-based AND flash

have been investigated to evaluate fin oxide-based flash as a synaptic device for off-chip learning. In off-chip training, only the read operation occurs in the inference process after weight update of synaptic devices, the synaptic weight's maintenance feature is the most essential characteristic. Fig. 2.7 (a) and (b) show room temperature (RT, 25°C) and 85°C retention performances of the oxide fin-based AND flash synaptic cell with different dynamic ranges. Long RT retention time over 10k seconds in both erased and programmed memory state and >2k conductance ratio over the retention time at room temperature are obtained [44]. Even at high temperatures, independent of the dynamic range, the current level in the erase state was well preserved, but the retention characteristic in the program state was not maintained for more than 1k seconds. This degradation can be attributed to the fact that the tunneling oxide of the AND memory device proposed is composed of a rather thin LPCVD medium temperature oxide (MTO) of less than 3 nm. Therefore, the retention characteristics can be improved by employing a thicker tunneling oxide than 3 nm grown by dry or $N_2O$ oxidation [49]. Cycling characteristics were also measured to evaluate memory performances. Using low

program and erase bias, memory window is well maintained over $10^4$ programming cycles.

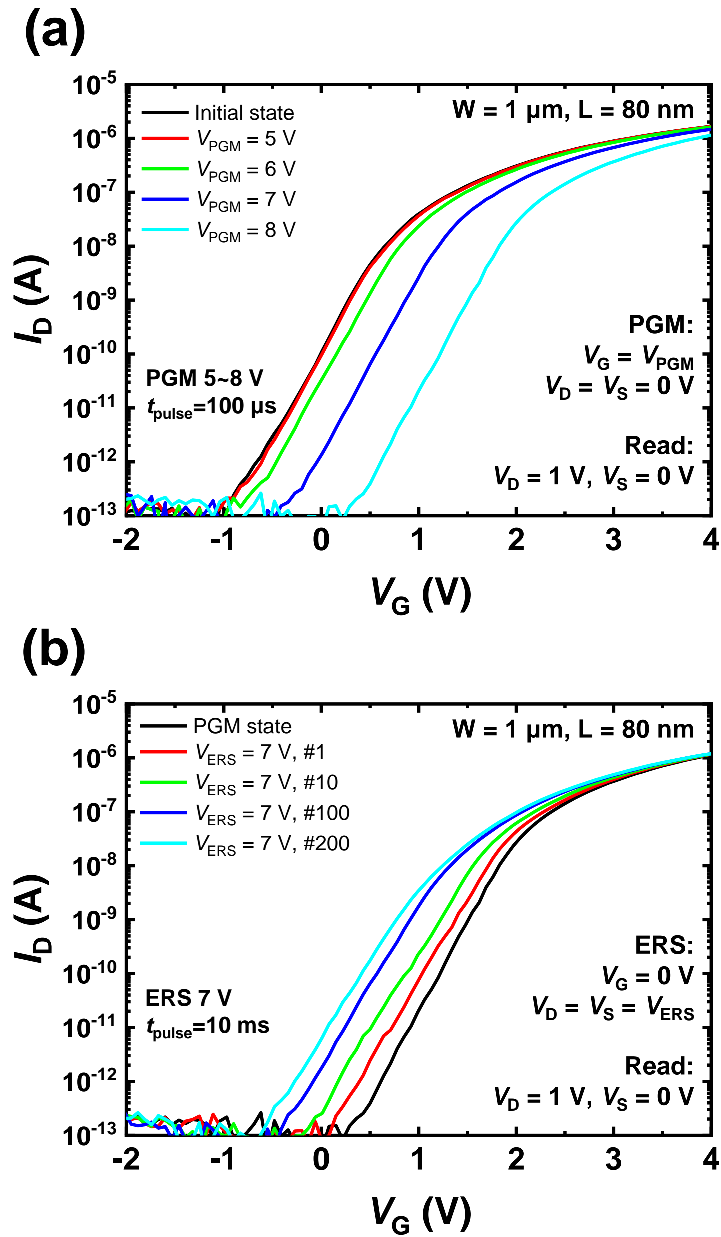



Fig. 2.4. (a) Measured program and (b) erase properties of the fabricated $SiO_2$ fin-based AND flash memory.

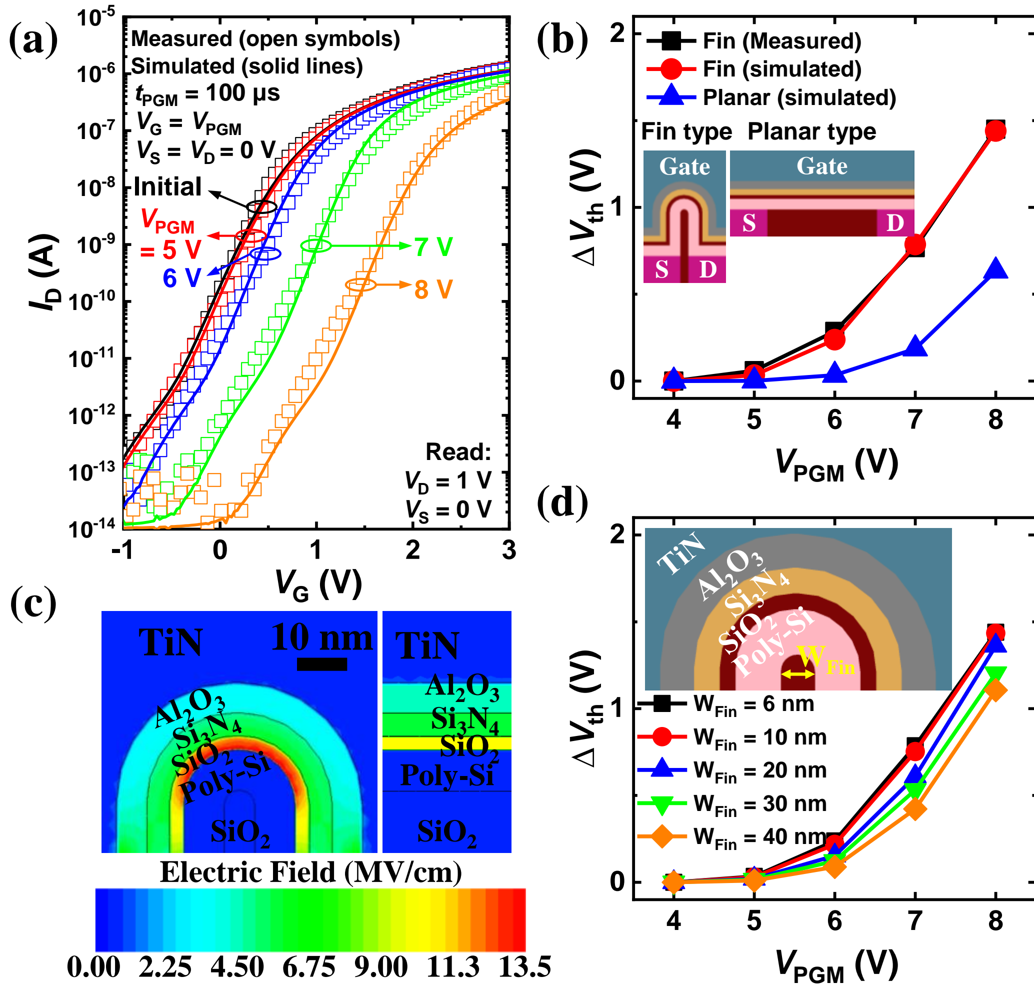



Fig. 2.5. (a) ISPP characteristics of a fin-type synaptic device. (b) Comparison of ISPP properties between a fin-type device and a planar-type device. (c) Electric field's distribution of fin-type and planar-type flash cells at $V_G$ = 8 V. (d) ISPP results of the fin-type synaptic device with varying oxide fin widths.

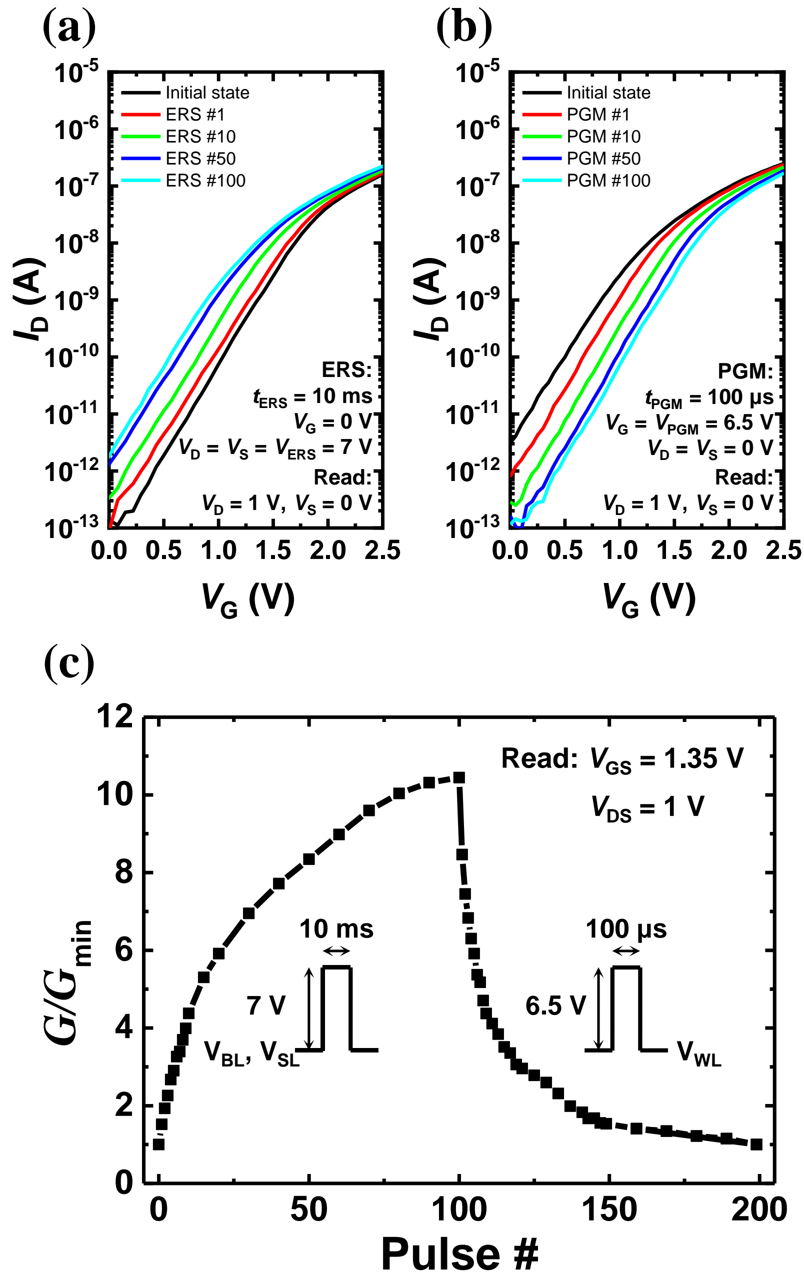



Fig. 2.6. (a), (b) Program and erase characteristics of a synaptic cell obtained by identical write pulses. (c) Potentiation and depression characteristics of the device.

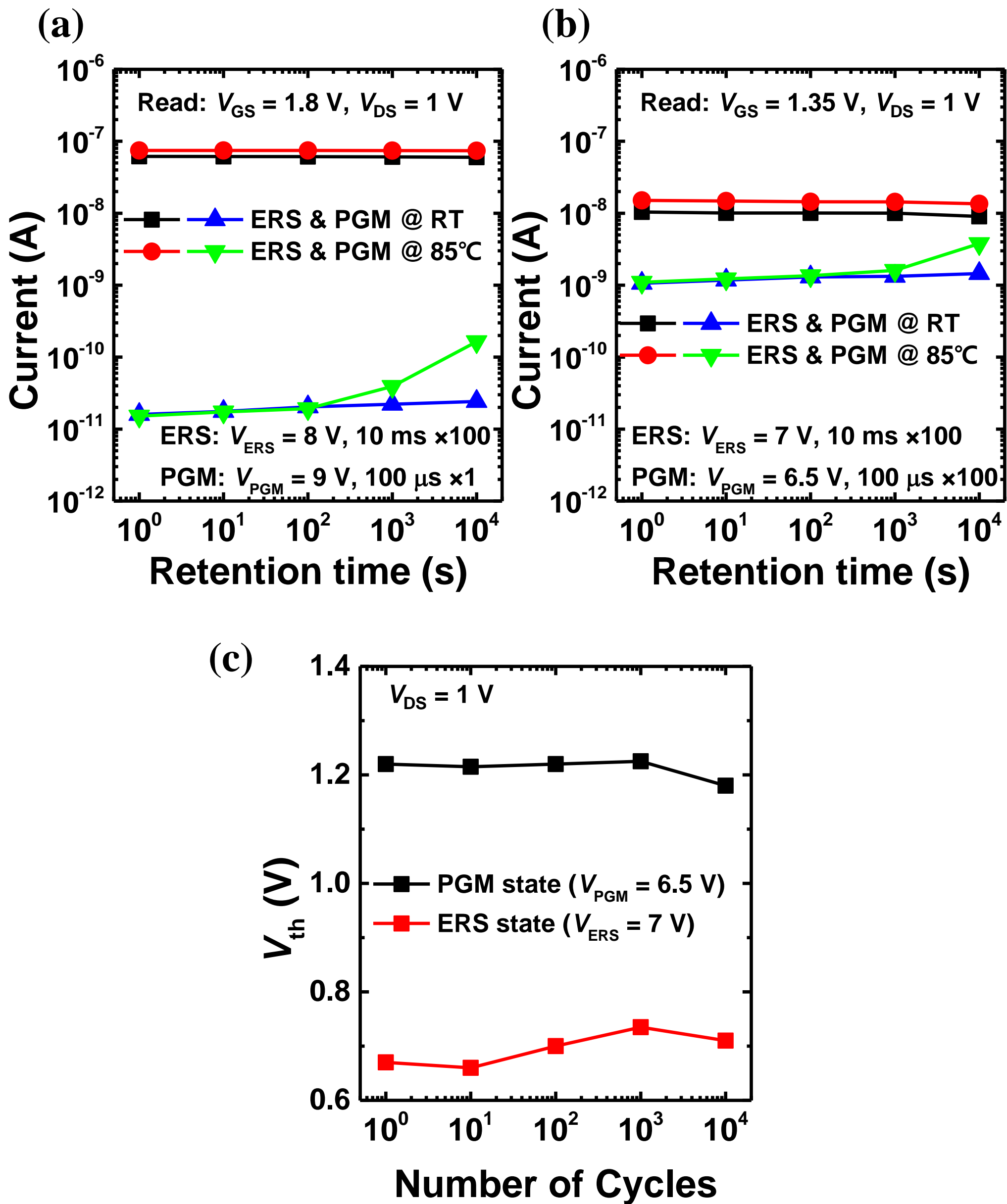



Fig. 2.7. (a), (b) RT and 85℃ retention properties of the $SiO_2$ fin-based AND flash synaptic cell with different synaptic weight ranges. (c) Cycling characteristics of the fabricated $SiO_2$ fin-based flash memory.

## 2.4 Array characteristics

To implement hardware-based deep neural networks in which a significant amount of program or erase operations are carried out during training, selective program and erase characteristics of massive parallel synaptic arrays with low disturbance are required. As aforementioned, an AND flash memory array exhibits selective program and erase operations, which provides various directions to update synaptic weights. Inputs are applied to WLs in the AND array architecture, and each BL current indicates weighted sum of each output neuron, which represents a VMM result. Fig. 2.8 shows a fabricated 2×2 $SiO_2$ fin-based AND flash synaptic array. A single cell in the proposed AND-type array occupies only $6F^2$, which is 40% smaller compared to that in a NOR-type array. FN tunneling-based selective program and erase are conducted in AND arrays. With the help of parallel crossbar AND-type array configuration, program and erase inhibit operation can be easily conducted by applying appropriate program and erase inhibit voltages to other unselected cells, respectively.

As depicted in Fig. 2.9, program inhibition bias schemes are designed to

achieve selectable program performance for parallel in-memory computing. For cell programming, a program voltage ($V_{PGM}$) is applied to WL of selected cell (cell A), while other unselected neighbor cells are inhibited by applying a program inhibit voltage ($V_{INH}$) to unselected BLs and SLs. Adopting the program inhibition scheme in Fig. 2.9(a), a positive program voltage ($V_{PGM}$ = 7 V, 100 μs) is applied one hundred times to a WL of selected cell A and program inhibition voltage of 3.5 V is applied to BL and SL of unselected cells as the program voltage is applied. As a result of programming using half program voltage for program inhibition, the current flowing in cell A decreased by 131 nA, while currents flowing in other cells changed by less than 5 nA as shown in Fig.2.9(b). Note that the cell to which the program inhibition voltage is applied is not erased while preventing FN-programming of neighbor cells sharing the WL of the selected cell. For selective cell erasing, a positive erase voltage ($V_{ERS}$) is applied to BL and SL of selected cell (cell A) after programming, while other unselected neighbor cells are inhibited by applying an erase inhibit voltage ($V_{INH}$) to unselected WLs. Adopting the erase inhibition scheme in Fig. 2.10(a), a positive erase voltage ($V_{ERS}$ = 7 V, 10 ms) is

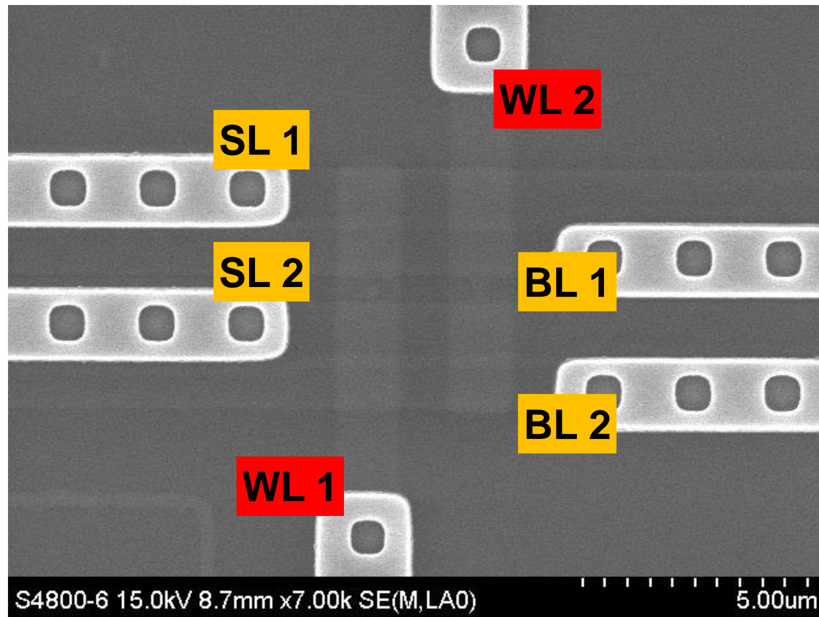
applied one hundred times to BL and SL of selected cell A and erase inhibit voltage of 3.5 V is applied to WLs of unselected cells as the erase voltage is applied. As a result of erasing using half erase voltage for erase inhibition, the current flowing in cell A increased by 29.8 nA at a read voltage of 2.15 V, while currents flowing in the other cells changed by less than 0.7 nA at the same read voltage as shown in Fig. 2.10(b). Note that the cell to which the erase inhibition voltage is applied is not programmed while preventing FN-erasing of neighbor cells sharing the BL or SL of the selected cell.

As illustrated in Fig. 2.11(a), a 10×10 AND flash synaptic array is designed and has been fabricated to investigate synaptic characteristics in the array. Fig. 2.11(b) shows the weighted sum current along a BL as a parameter of input WL gate bias. To measure the current of each cell connected to the BL, read bias is applied to the WL of the cell to be measured, and the other cells are turned off. Note that the weighted sum current flowing along BL 3 is measured 0.7 % lower than the sum of cell currents in each of the ten at read voltage of 2 V.

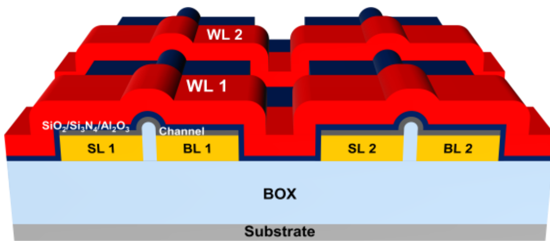
In order to investigate current sum accuracy in more massive array, a 24×8

AND flash synaptic array was fabricated as shown in Fig. 2.12(a). Fig. 2.12(b) shows weighted sum current and the sum of each cell current along BL 3 of 24×8 synaptic array at $V_{read}$ =2 V. The measured current sum error reached about 5%. Fig. 2.12(c) shows current sum error as a function of the number of inputs. The larger the input size, the greater current sum error. This tendency is due to the IR drop along the BL or SL, which can be further improved by reducing BL/SL resistance.

(a)




(b)




(c)

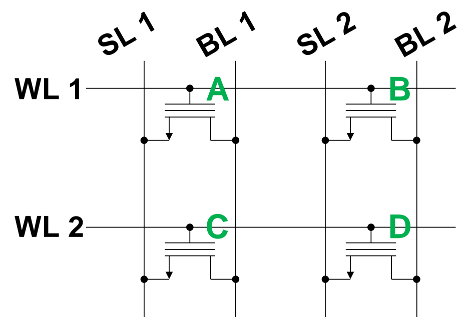
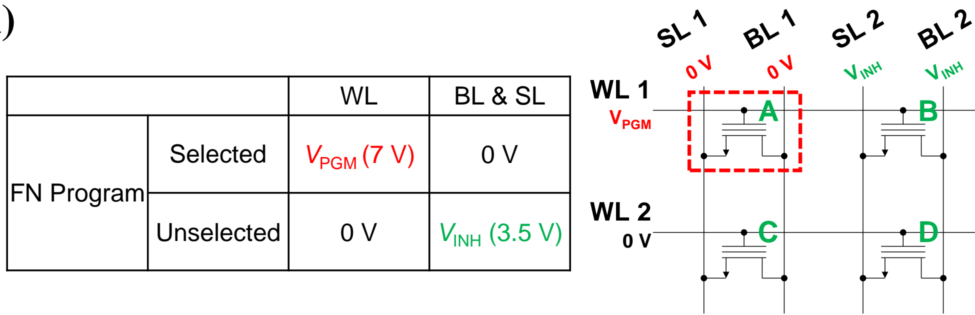



Fig. 2.8. (a) Top SEM image of the fabricated 2×2 AND flash synaptic array. (b) 3D schematic view of the fabricated $SiO_2$ fin-based AND flash synaptic array. (c) Schematic of the 2×2 AND flash synaptic array.

(a)



| | | WL | BL & SL |
|---|---|---|---|
| FN Program | Selected | $V_{PGM}$ (7 V) | 0 V |
| | Unselected | 0 V | $V_{INH}$ (3.5 V) |



(b)

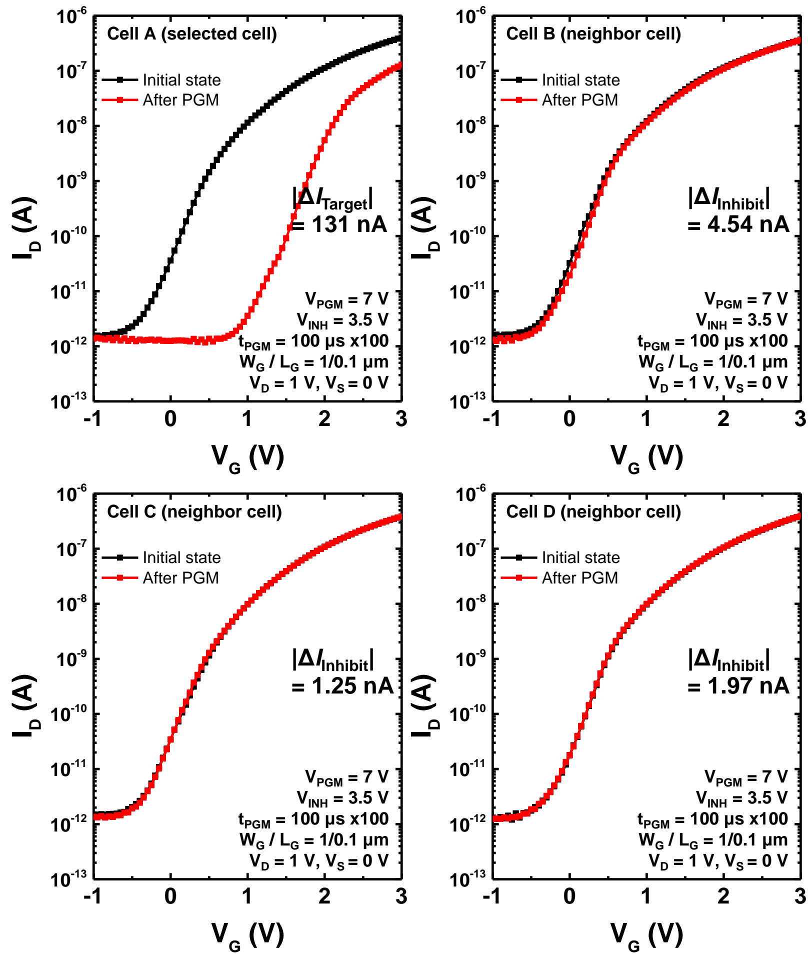


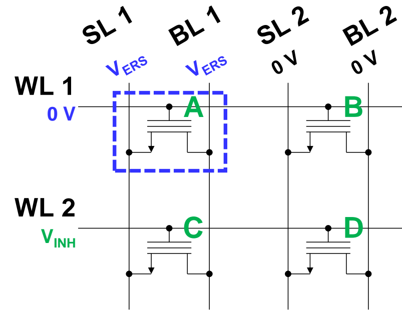


Fig. 2.9. (a) Bias conditions for a single-cell (cell A) selective program operation.

(b) Cell program and program inhibition properties in the AND synaptic array.

(a)

| | | WL | BL & SL |
|---|---|---|---|
| FN erase | Selected | 0 V | $V_{ERS}$ (7 V) |
| | Unselected | $V_{INH}$ (3.5 V) | 0 V |




(b)

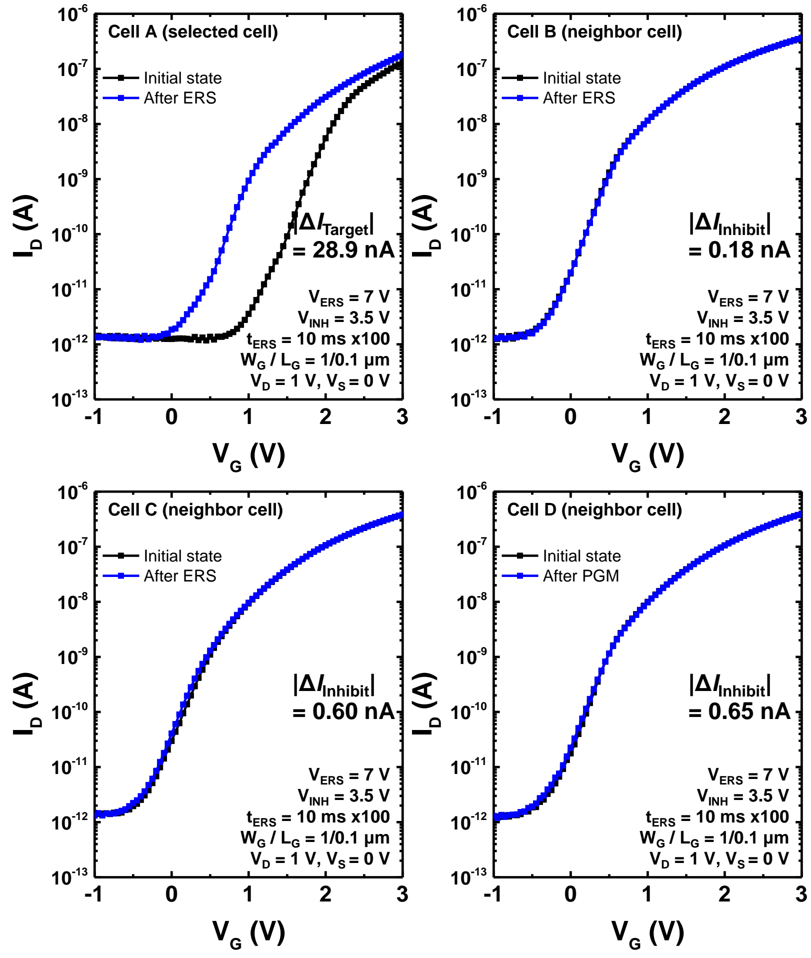



Fig. 2.10. (a) Bias conditions for a single-cell (cell A) selective erase operation. (b) Cell erase and erase inhibition properties in the AND synaptic array.

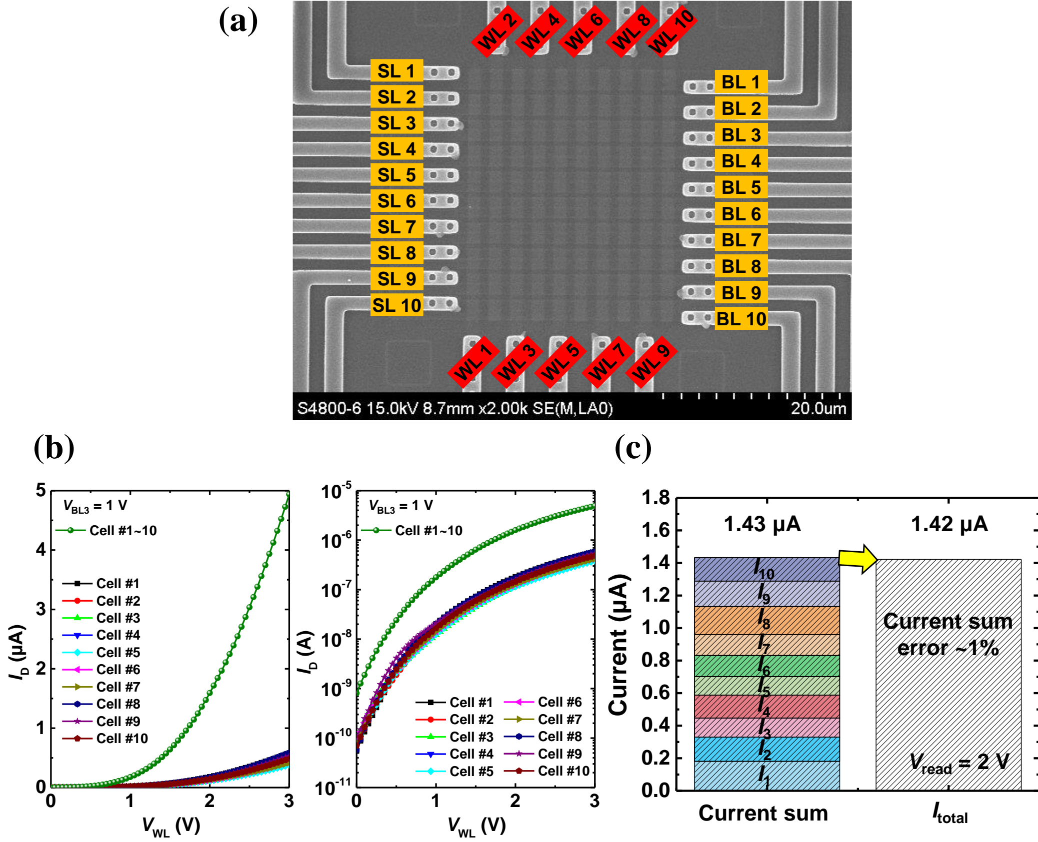



Fig. 2.11. (a) Top SEM image of the fabricated 10×10 AND flash synaptic array. (b) Weighted sum current and each cell current along a BL (BL 3) of the 10×10 synaptic array. (c) Comparison between weighted sum current and the sum of each cell current along BL 3 of the 10×10 synaptic array at $V_{read}$ =2 V.

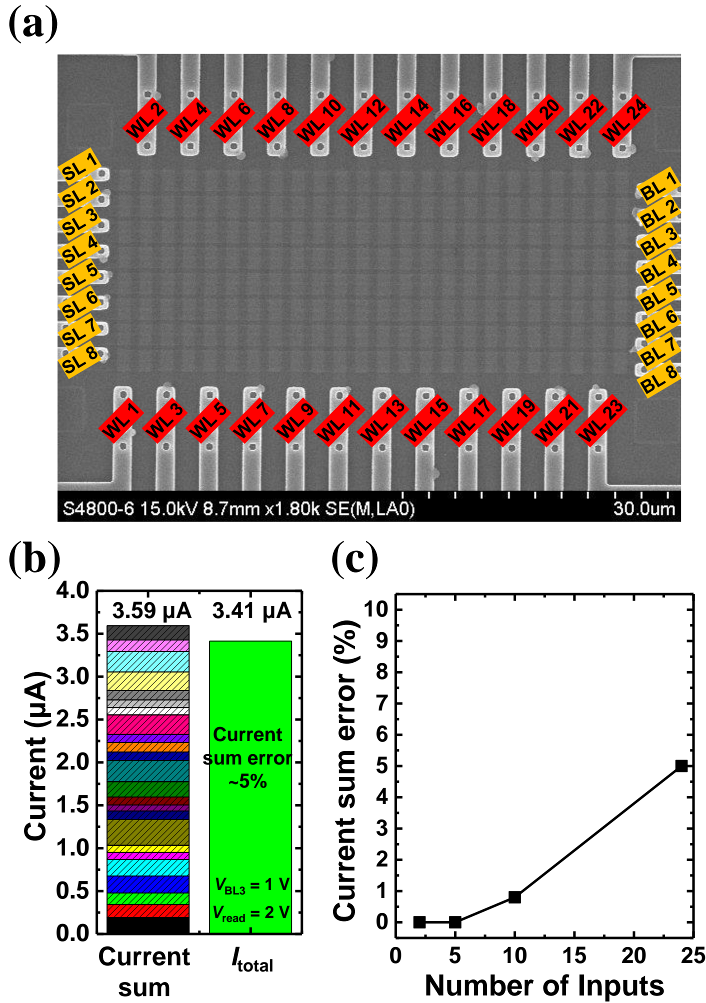


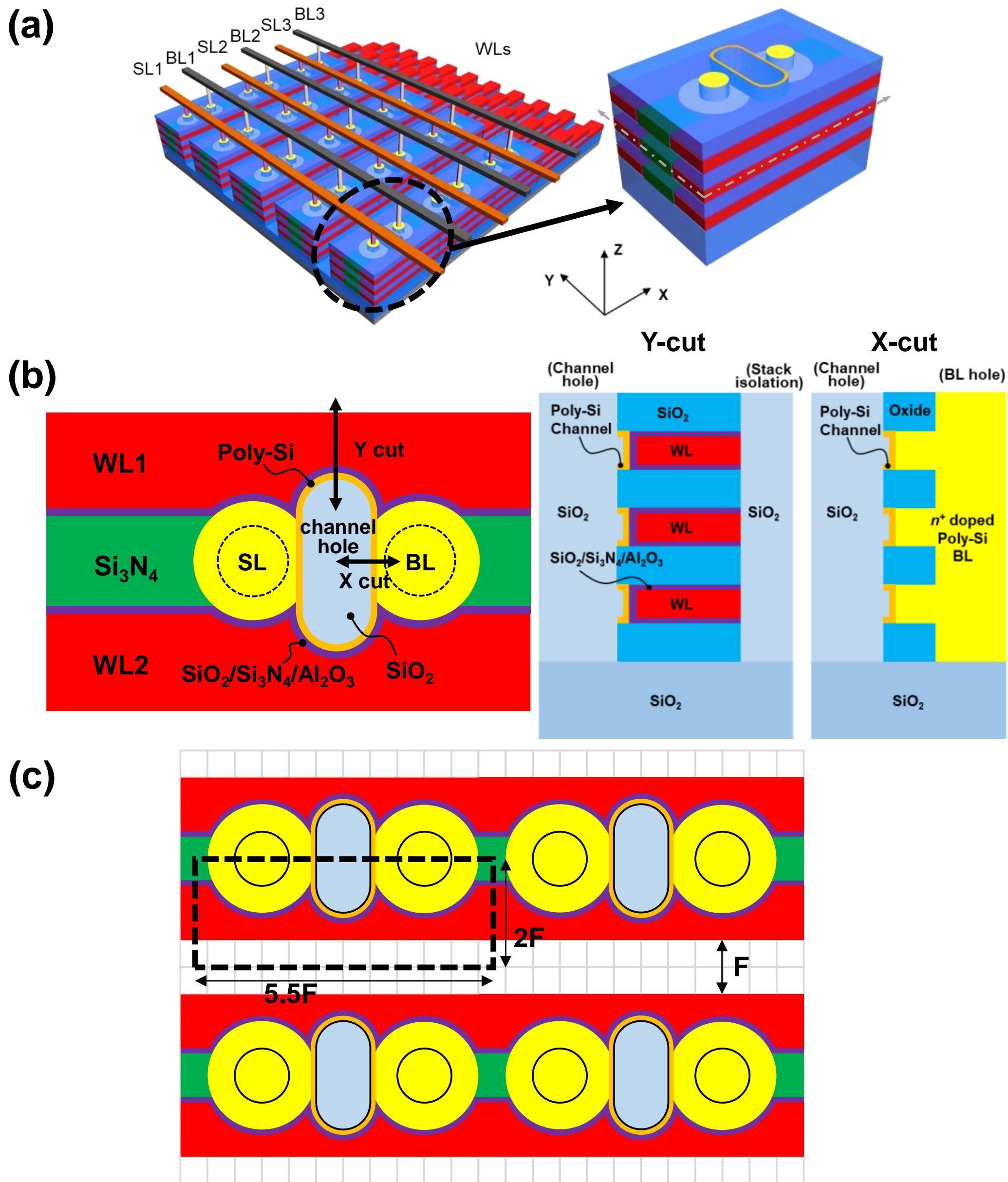

Fig. 2.12. (a) Top SEM image of the fabricated 24×8 AND flash synaptic array. (b) Comparison between weighted sum current and the sum of each cell current along BL 3 of 24×8 synaptic array at $V_{read}$ =2 V. (c) Current sum error as a function of the number of inputs.

# Chapter 3
# 3D AND flash synaptic array with rounded channel

## 3.1 Device structure

In order to achieve high scalability for massive synaptic array, 3D AND flash synaptic array with rounded channel has been proposed. As aforementioned, round-shaped channel structure wrapped by a metal gate exhibits superior memory performance to reduce programming voltage for low-power synaptic update operation in HNNs. Fig. 3.1 shows proposed 3D AND flash synaptic device structure. It can be seen as a structure in which $SiO_2$ fin-based flash synaptic cells are stacked in a vertical direction. A BL and a SL are arranged in vertical direction to connect memory cells in parallel by forming BL and SL plugs. A thin poly-Si channel of each cell stacked vertically is located around a channel hole in which oxide insulator material is filled. A high-k gate insulator stack, consisting of $SiO_2$, $Si_3N_4$, and $Al_2O_3$ layers, is deposited conformally outside the thin channel by

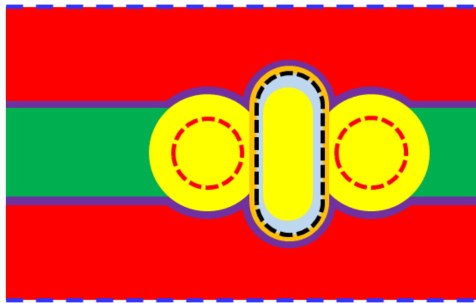
LPCVD and ALD process. The charge stored in charge trap layer of $Si_3N_4$ determines the cell current representing synaptic weight. TiN metal WLs are formed through gate last process. In each trench, two parallel WLs surround both sides of rounded channel in channel hole, showing a GAA-like structure. Note that thin poly-Si round-shaped channels of each cell connected to the BL plug are separated from each other inside the channel hole in vertical direction. The proposed synaptic cell exhibits an effective unit cell area of $11F^2$, as shown in Fig. 3.1(c), which is smaller than the previously reported 3D AND-type GAA-like device [42]. It also shows smaller the radius of curvature to improve programming efficiency based on local field enhancement effects compared to the previous proposed 3D AND-type device with the round channel. In the proposed 3D stackable AND array, BL and SL resistance can be enhanced by thinning the thickness of WL space compared to that in 2D AND array.

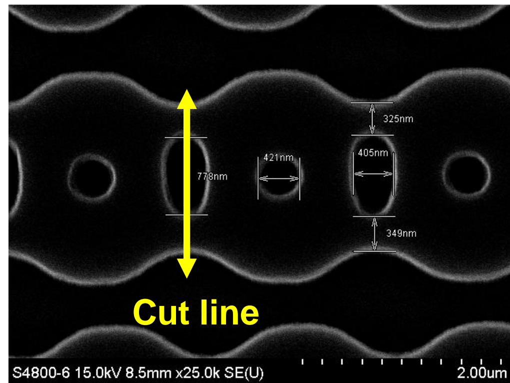


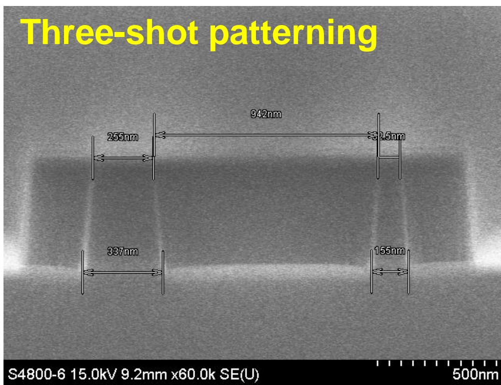

Fib. 3.1 (a) 3D schematic of 3D AND flash synaptic array and cell structure. (b) Cross-sectional schematics of the 3D AND flash cell. (c) Unit cell area of the proposed 3D AND flash device.

## 3.2 Fabrication process

As aforementioned, the proposed 3D AND array with rounded channel can be fabricated by etching BL/SL holes, channel hole and WL trench. When channel holes, SL/BL holes, and WL trenches are all patterned independently, device variation can occur owing to photomask misalignment. Therefore, we propose a process in which three masks are combined into one mask and only the parts that need processing after patterning are opened separately. As can be seen from patterning experiment results in Fig. 3.2, misalignment could be reduced from 100 nm to 40 nm or less by integrating three masks into one. Schematics and key steps of a fabrication process for 3D AND flash array with round-shaped channel are shown in Fig. 3.3. The detailed fabrication steps are described in the following section.



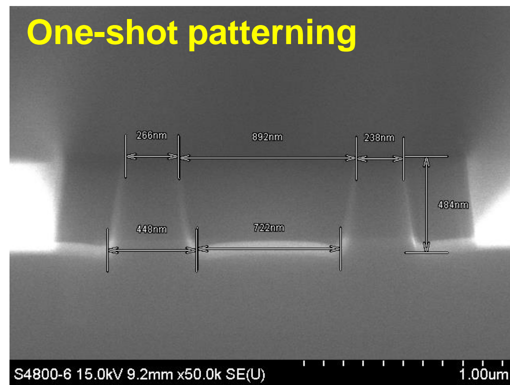




Fig. 3.2. Misalignment improvements using one-shot patterning compared to three-shot patterning.

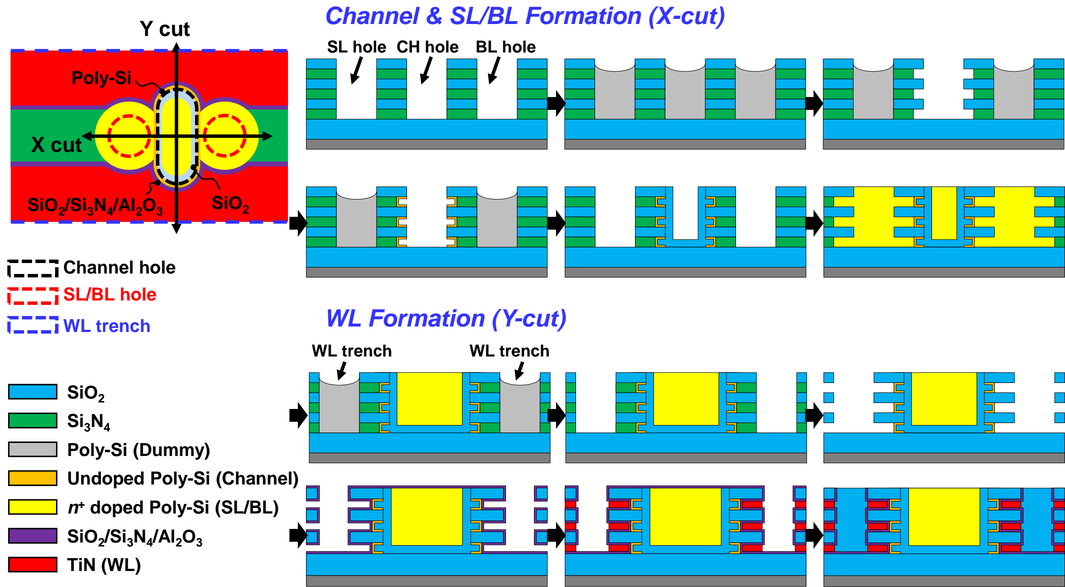



Fig. 3.3. Key fabrication steps of the proposed 3D AND array with rounded channel.

### 3.2.1 Cell process steps

Most of the fabrication steps were carried out using the equipment of Inter-university Semiconductor Research Center (ISRC) located in Seoul National University (SNU), Seoul, Republic of Korea, and ON stack formation by plasma enhanced chemical vapor deposition (PECVD) was implemented by using the equipment of Electronics and Telecommunications Research Institute located in Daejeon, Republic of Korea.

The proposed 3D AND synaptic devices are fabricated with ten masks. First, multi-layer ON (100 nm $SiO_2$/ 60 nm $Si_3N_4$) stack on a 300 nm-thick thermally grown $SiO_2$ layer is deposited by PECVD process as shown in Fig. 3.4. The multi-layer ON stack includes three nitride layers to form 3-layer 3D AND device. The highly selective etching of $Si_3N_4$ over $SiO_2$ is required to fabricate 3D flash device. Using hot phosphoric acid at 160°C, ON stack deposited by PECVD process shows higher wet-etch selectivity between $Si_3N_4$ and $SiO_2$ of 22:1 compared to rather poor etch selectivity of 10:1 with ON stack formed by LPCVD process. Before patterning SL/BL plug, channel plug and WL trench simultaneously, WL cut process is carried out by etching ON stack in a part of the region where the WL is to be formed and filling with 500 nm-thick PECVD oxide. Note that the top oxide thickness is maintained at 200 nm through etch-back process after filling the WL cut region with oxide.

After WL cut process, channel hole, SL/BL hole, and WL trench are etched using second photo mask as illustrated in Fig. 3.5(a). In order to reduce cell variation resulted from via hole etch process, 600 nm-thick ON stack is etched

keeping chuck temperature at 60°C in inductively coupled plasma oxide/nitride etcher to produce steep etch slope. The next step is to fill all open holes and trenches with poly-Si material using LPCVD process. 400 nm-thick poly-Si layer is deposited to fill via holes and etch-back process follows as shown in Fig. 3.5(b). Using dummy poly-Si to fill the inside of holes and WL trench, it is possible to use a process method that open a specific hole using a photo mask and $Si_3N_4$ exposed inside the hole or trench can be selectively etched.

Then dummy poly-Si inside channel hole is removed by isotropic etching using channel hole open mask to form channel poly-Si. The isotropic etching of dummy poly-Si filled inside channel hole is carried out using $SF_6$ etchant gas. After that, nitride layers inside channel hole exposed to the outside are partially etched to form the interlayer channel region using selective nitride wet etching process as shown in Fig. 3.6. Phosphoric acid ($H_3PO_4$, 160°C) is used to etch nitride layer selectively, which has a high etch selectivity between $Si_3N_4$ and $SiO_2$ of 22:1. 30 nm of $Si_3N_4$ layers are partially etched in the process.

After nitride layers are partially inside the channel plug and an *a*-Si layer has

been deposited by the LPCVD method, the *a*-Si layer is re-crystallized at 600 °C for 24 hours to create an undoped poly-Si channel. To separate poly-Si channels by layer, reactive ion etching (RIE) process is carried out using channel hole open mask as shown in Fig. 3.7(a). Then plasma-enhanced tetraethyl-orthosilicate (PE-TEOS) is deposited for passivation of poly-Si channels and etch-back process follows (Fig. 3.7(b)). Fig. 3.7(c) shows a cell structure after channel passivation.

Similar to the case of the channel formation process, dummy poly-Si inside SL/BL holes is removed by isotropic etching using SL/BL plug open mask to form doped poly-Si SL/BL. The isotropic etching of dummy poly-Si filled inside SL/BL holes is carried out using $SF_6$ etchant gas. After that, $Si_3N_4$ layers inside SL/BL plugs exposed to the outside is partially etched using hot $H_3PO_4$ at 160°C to create the space in which SL and BL are formed. Note that selective nitride wet etching process should be carried out until the undoped poly-Si channel is exposed as shown in Fig. 3.8(a). In this process step, 500 nm of $Si_3N_4$ layers are partially etched as shown in Fig. 3.8(b).

Then a layer of *in situ* $n^+$-doped poly-Si, 430 nm in thickness, is deposited for

SLs and BLs by LPCVD. Etch-back process follows to remove doped poly-Si on other region except SL and BL holes by dry etching poly-Si in this work as shown in Fig. 3.9. In order to form flat profile of doped SLs and BLs, poly-Si CMP can be utilized with following dry etching of poly-Si.

After the BL and SL formation, WL pad formation process is carried out to form WL contact pad area. The detailed process steps for WL pad are described the next section.

Then dummy poly-Si inside WL trench is removed by isotropic etching using WL trench open mask to form metal WLs by layer. The isotropic etching of dummy poly-Si filled inside WL trench is carried out using $SF_6$ etchant gas. After that, $Si_3N_4$ layers inside WL trench exposed to the outside is partially etched using hot $H_3PO_4$ at 160°C to create the space in which WLs are formed as shown in Fig. 3.10. Note that selective nitride wet etching process should be carried out until all parts of the channel are revealed so that the WL covers around the entire poly-Si channel as shown in Fig. 3.10.

After WL space is provided, a gate dielectric stack are formed before WL

formation. The gate dielectric stack consisting of $SiO_2$ /$Si_3N_4$/$Al_2O_3$ (O/N/A: 3/5/6 nm) layer is deposited as shown in Fig. 3.11. The tunneling oxide $SiO_2$ and the charge trapping layer $Si_3N_4$ are deposited by LPCVD process and the blocking oxide of alumina is formed by ALD process. Then WL formation is carried out. Conformal TiN metal is deposited to fill WL space by MOCVD process. In order to separate WLs by layer, isotropic wet-etching of TiN layer is performed using diluted hydrogen peroxide solution (DIW:$H_2O_2$ = 3:1, 60°C). Fig. 3.11 shows separated WLs in the cell region.

The next step is to deposit an insulating material to fill within the WL trench and planarize the surface using CMP process as shown in Fig.3.12. A 750 nm-thick PE-TEOS oxide layer is deposited for filling WL trench and insulating staircase shaped WL contact area. CMP is carried out to polish all deposited 750 nm-thick PE-TEOS oxide layer. Stopping point of CMP is shown in Fig. 3.12(b)

Lastly, the back end of line (BEOL) process is executed. A 300 nm-thick PE-TEOS oxide is deposited as an inter-layer dielectric (ILD). Contact hole patterning process is divided into two parts; WL contact holes patterning and SL/BL contact

holes patterning. Because the WL and SL/BL contact holes have different etch amounts due to the ON stack thickness of 400 nm, two masks are used to etch PE-TEOS oxide in the WL and SL/BL contact holes. Then Ti/TiN/AL/TiN layers, deposited by sputtering and MOCVD process, are patterned to form metal lines.

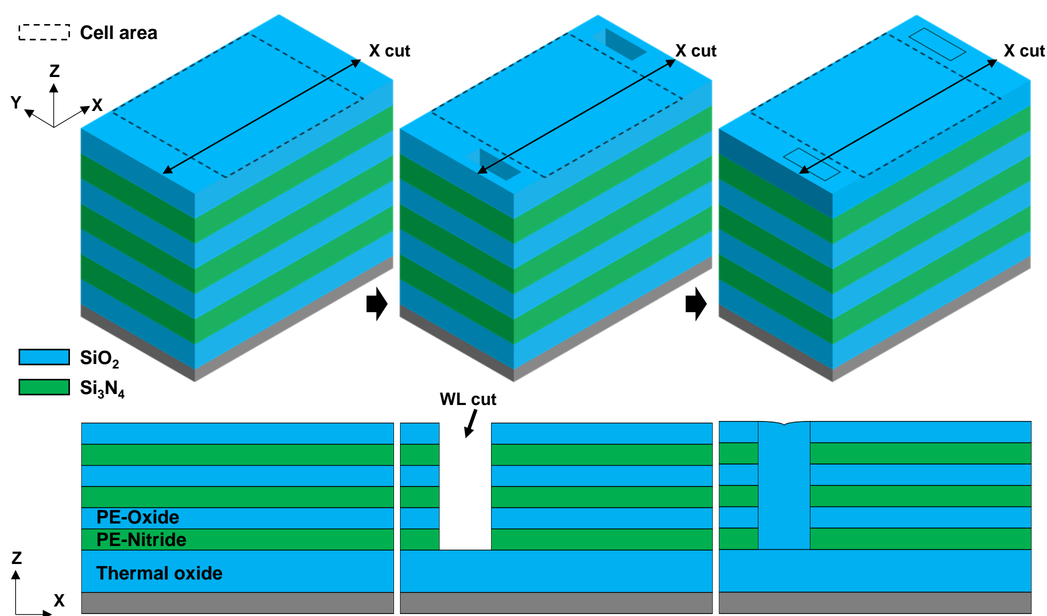



Fig. 3.4. Schematics of multi-layer ON stack and filled oxide in WL cut area.

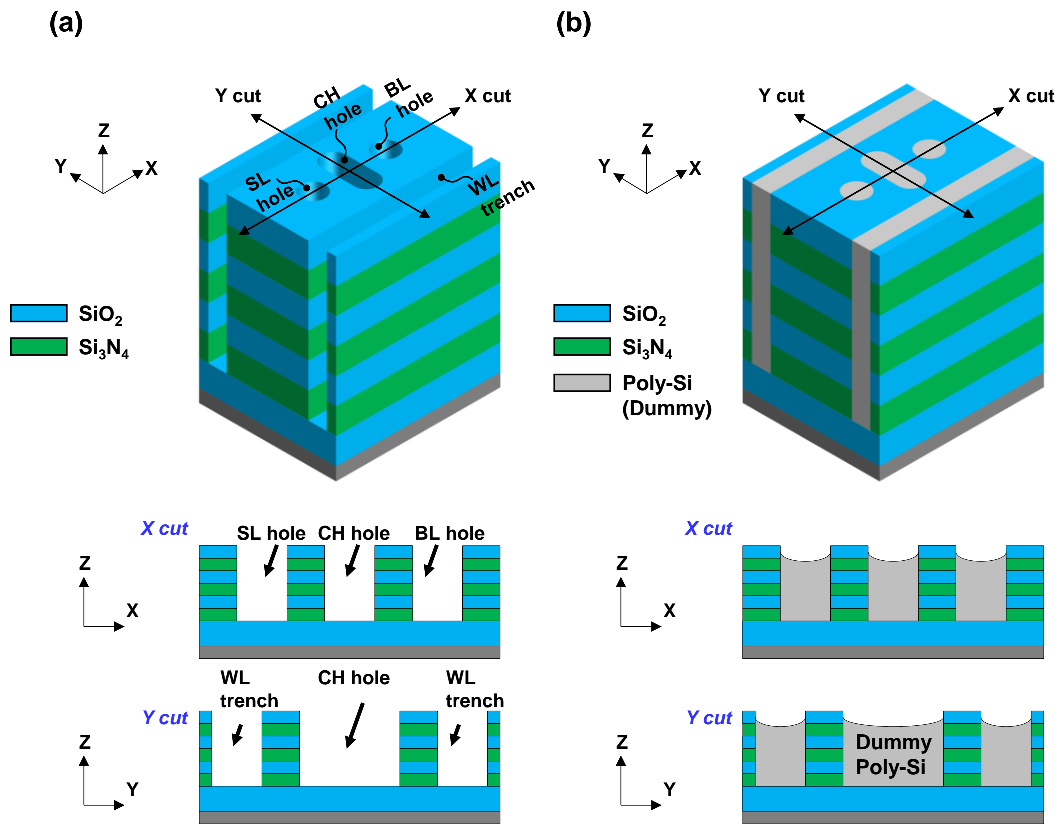
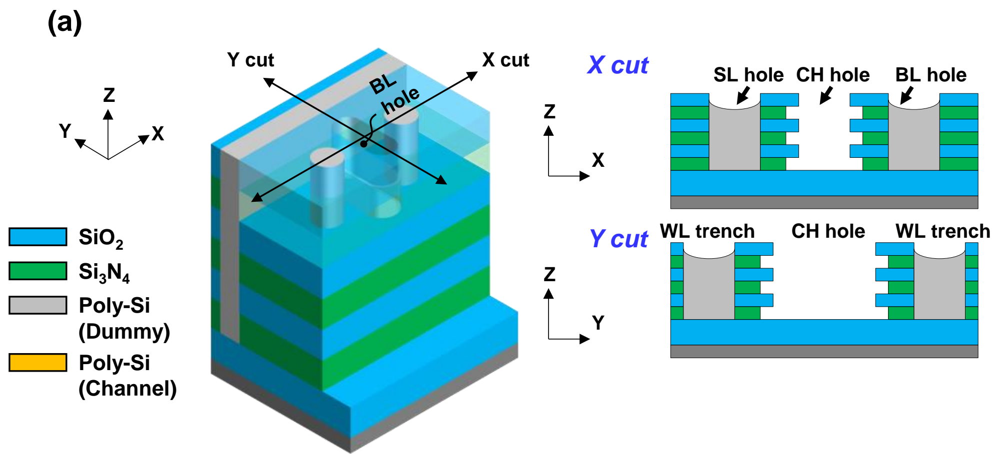
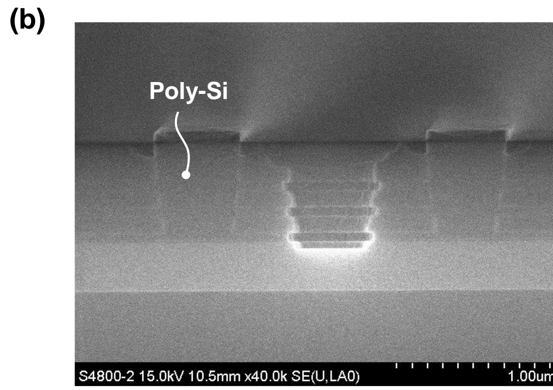



Fig. 3.5. Schematics of (a) via holes and trench patterning, and (b) dummy poly-Si filling in holes and trench patterned.



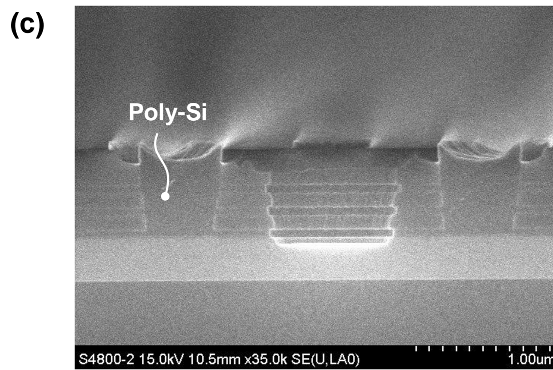




Fig. 3.6. (a) Schematics of partial nitride wet etching. (b) X-cut and (b) Y-cut cross-sectional SEM image after partial nitride wet etch process.

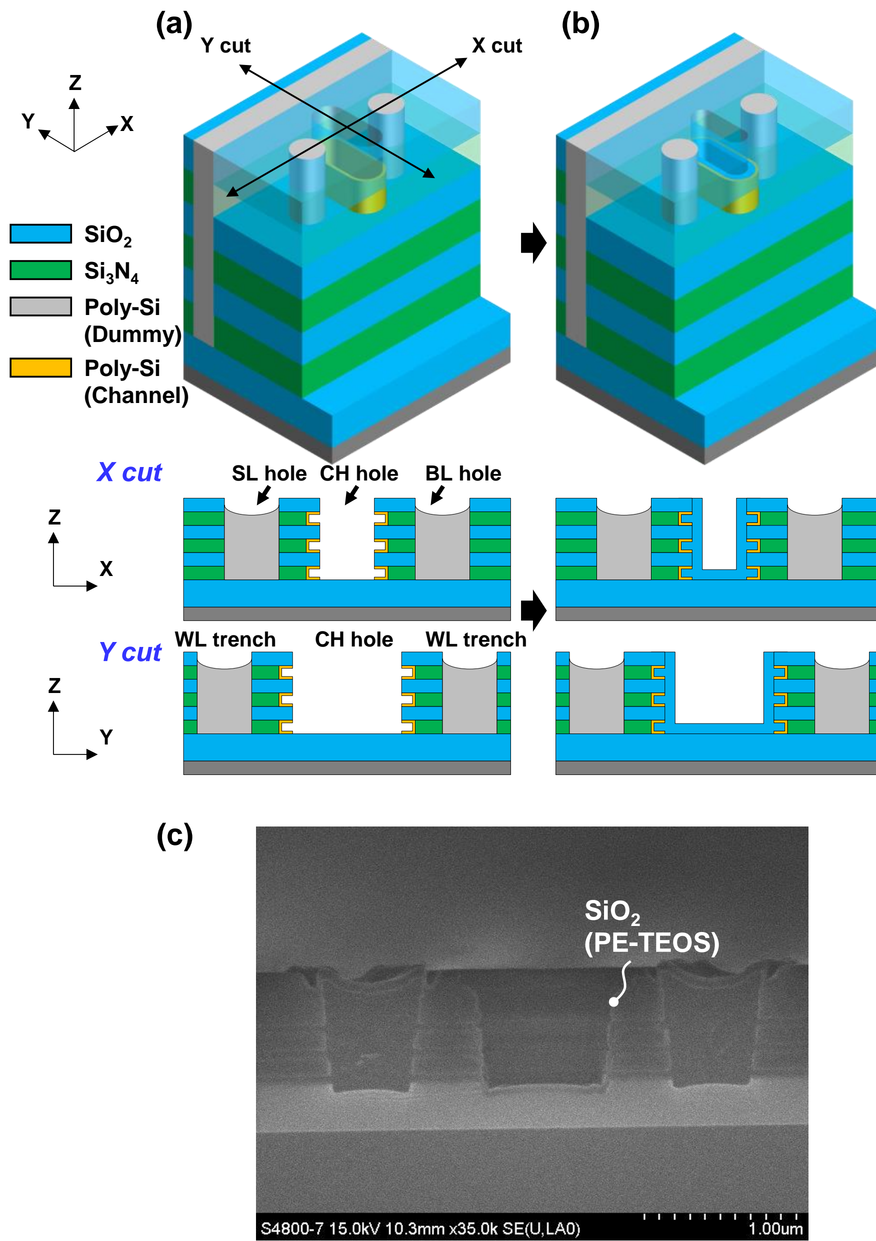



Fig. 3.7. Schematics of (a) separated channel formation by layer and (b) following passivation process. (c) Y-cut cross-sectional SEM image after the passivation process.

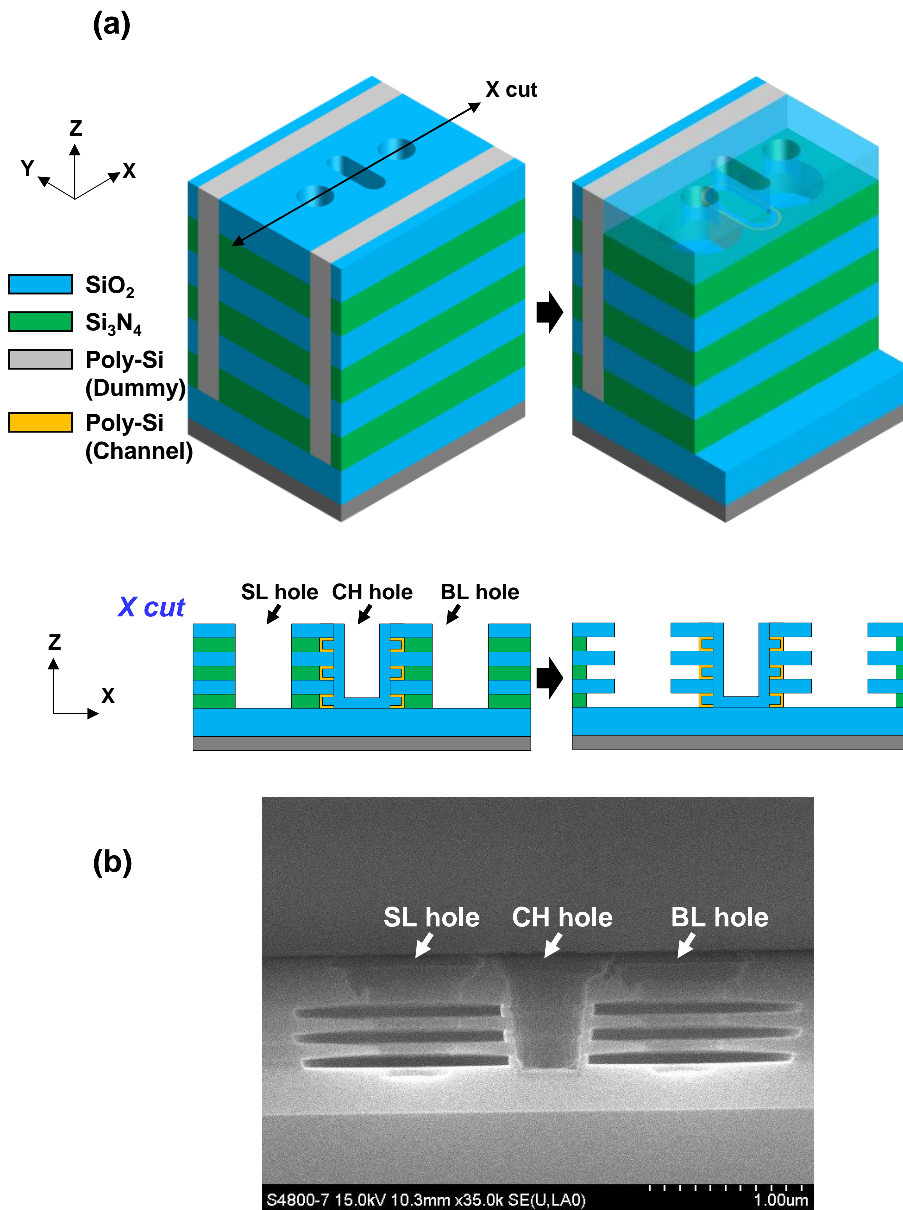
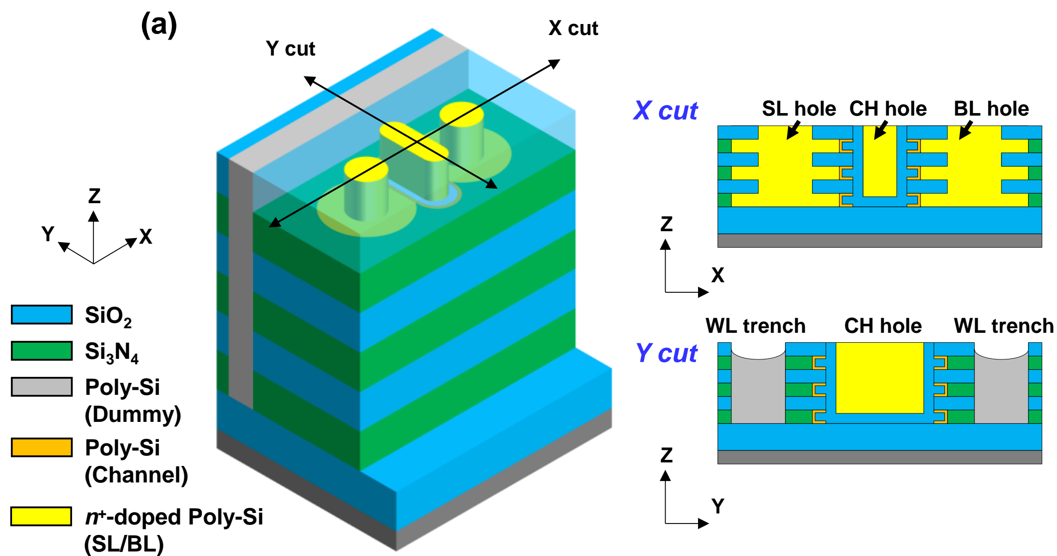



Fig. 3.8. (a) Schematics of dummy poly-Si etching inside BL/SL holes and following nitride partial etching process. (b) X-cut cross-sectional SEM image after the $Si_3N_4$ partial etching process.



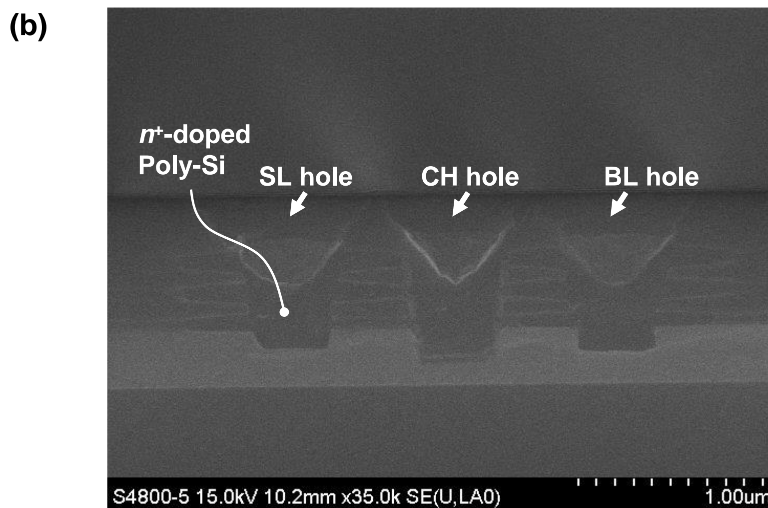




Fig. 3.9. (a) Schematics of $n^{+}$-doped poly-Si deposition and poly-Si etch-back process. (b) X-cut cross-sectional SEM image after the etch-back process.

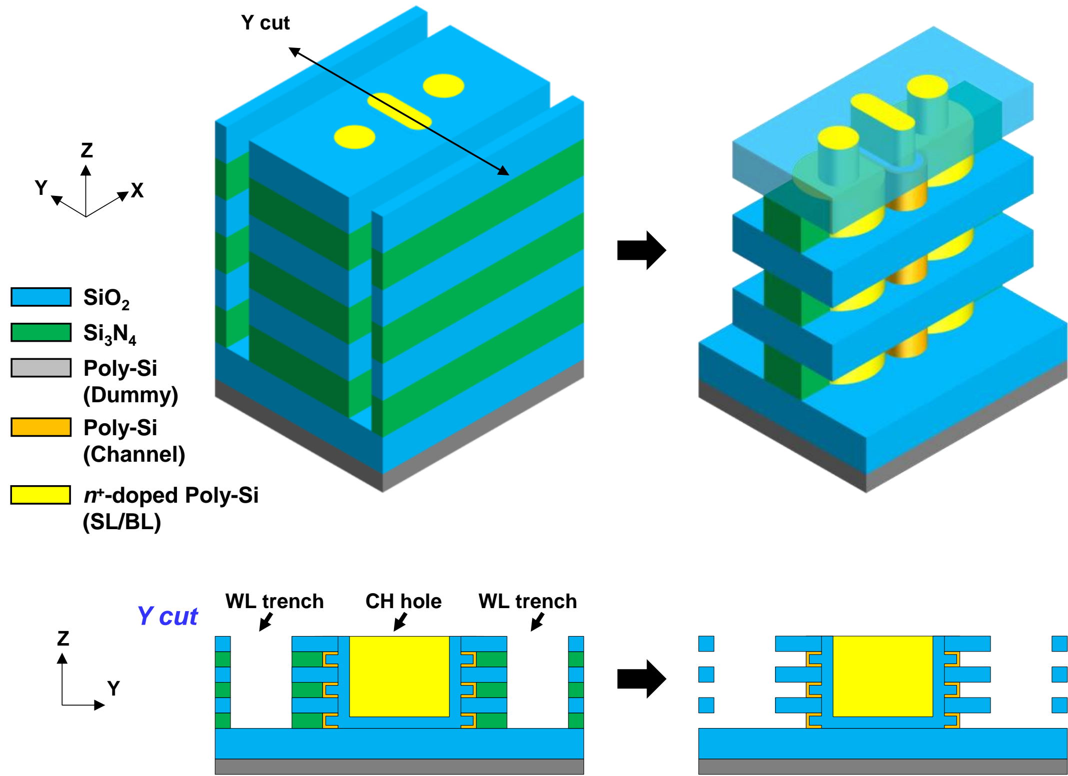
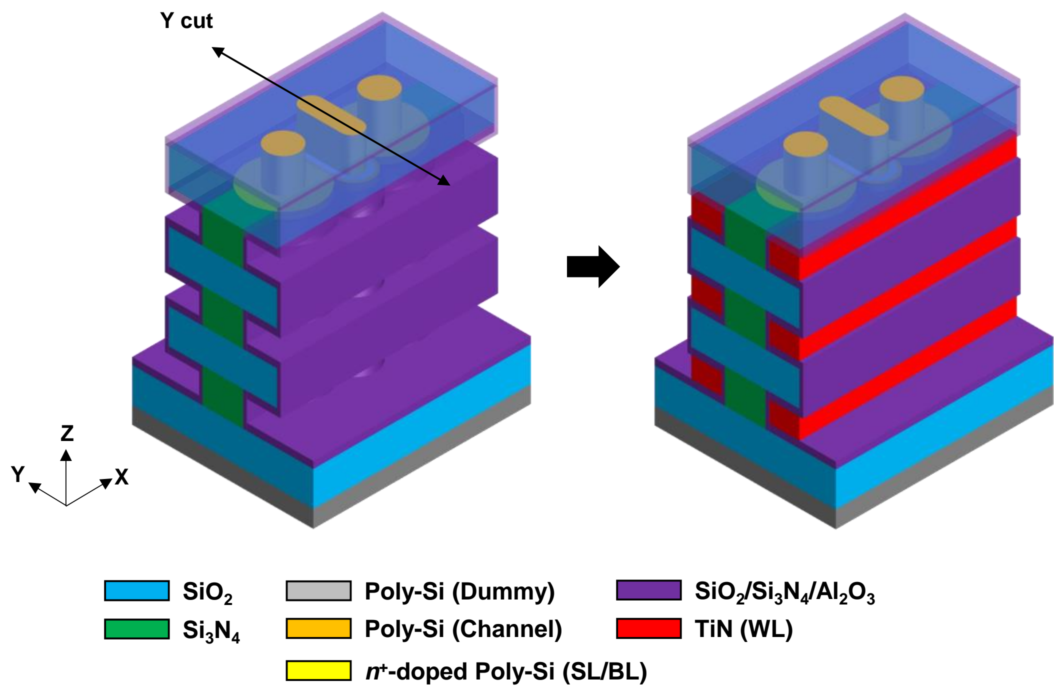



Fig. 3.10. Schematics of dummy poly-Si etching inside WL trench and partial nitride wet-etching inside WL trench.



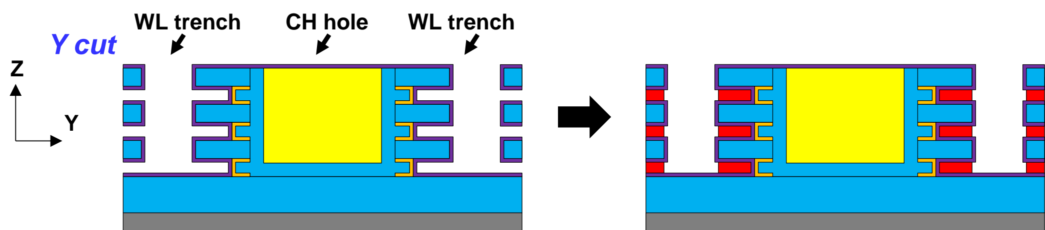
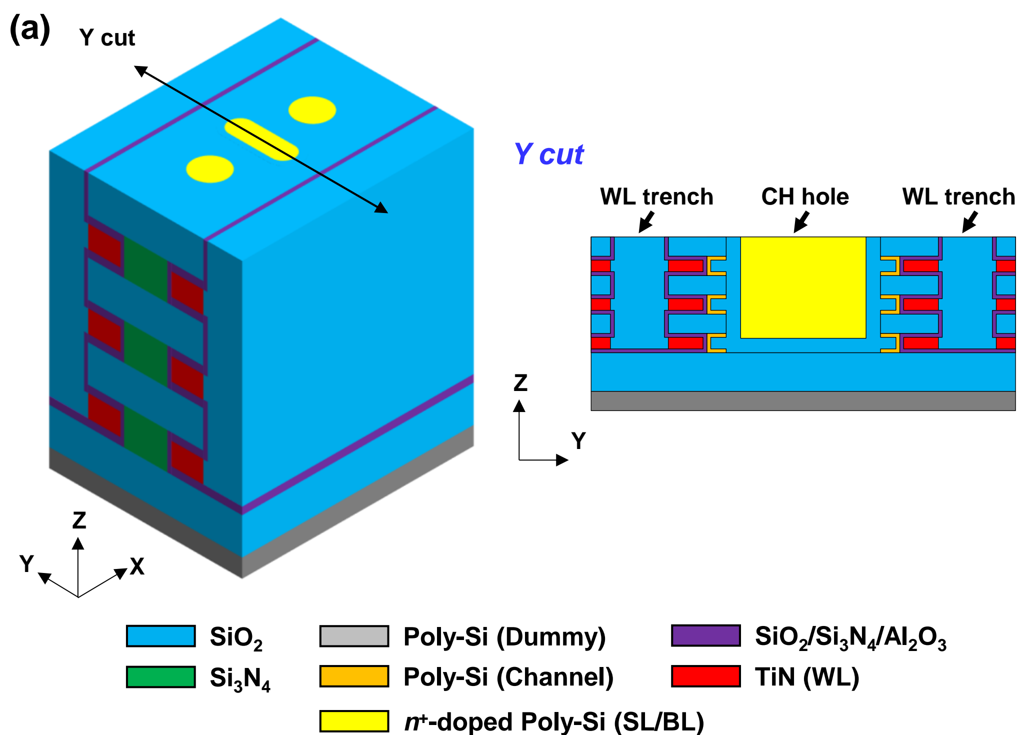




Fig. 3.11. Schematics of O/N/A gate insulator stack deposition and WL formation.



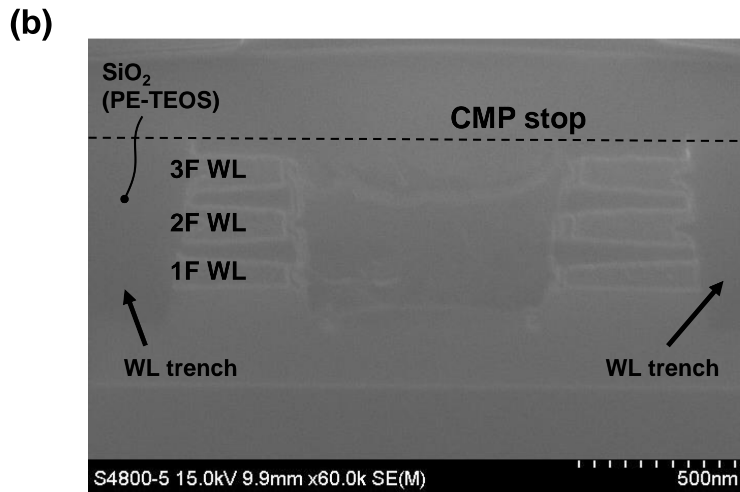




Fig. 3.12. (a) Schematics of CMP process. (b) Y-cut cross-sectional SEM image after the CMP process.

### 3.4.2 WL contact pad process steps

Fig. 3.13 shows WL contact pad process flow. Contact pads to the WLs are created using controlled edges that produce the distinctive staircase-shaped structure. The staircase-shaped structure is defined using three masks including the WL trench open mask. As aforementioned, before removal of dummy poly-Si inside WL trench, 3F WL edge is patterned using a 3F open mask. After 3F WL edge is defined, 2F WL edge is patterned using a 2F open mask by etch stopping until all second nitride layer is etched. Fig. 3.14 shows the plane and cross-sectional views of contact pad region after the back-end process is done. As aforementioned, the oxide via in WL cut area separates WLs into two parallel WLs surrounding both sides of rounded channel in channel hole.

However, in this process, the ON etch amount was insufficient when forming the 2F WL edge, resulting in the connection of 1F and 2F WL through the 1F contact hole. In a follow-up study, the etch amount control should be fine-tuned.

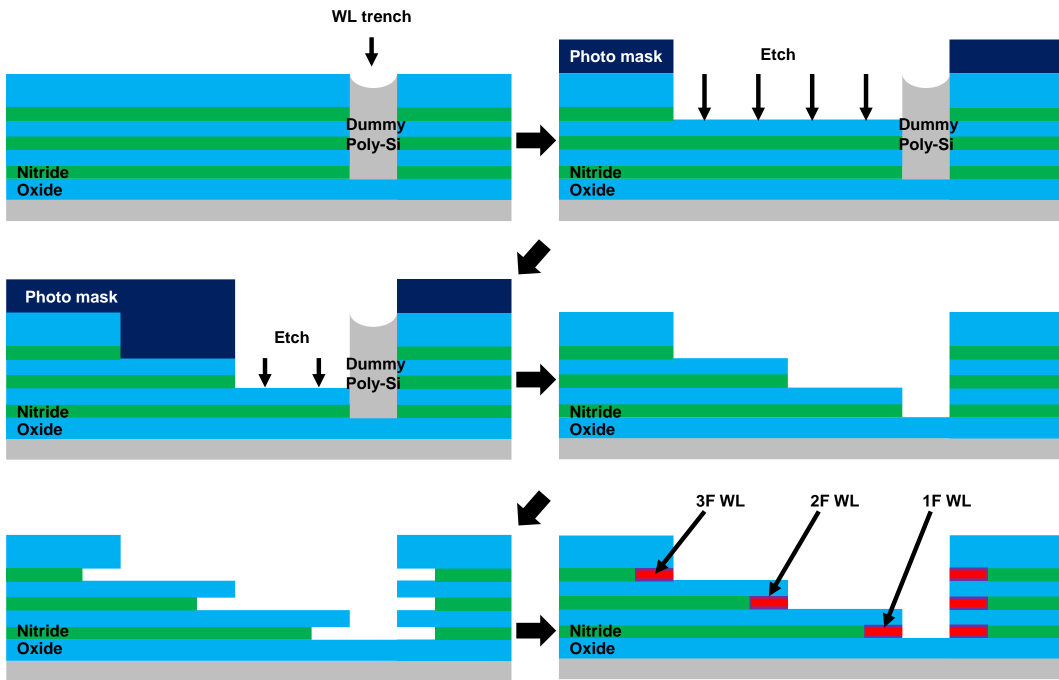



Fig. 3.13. WL contact pad fabrication steps.

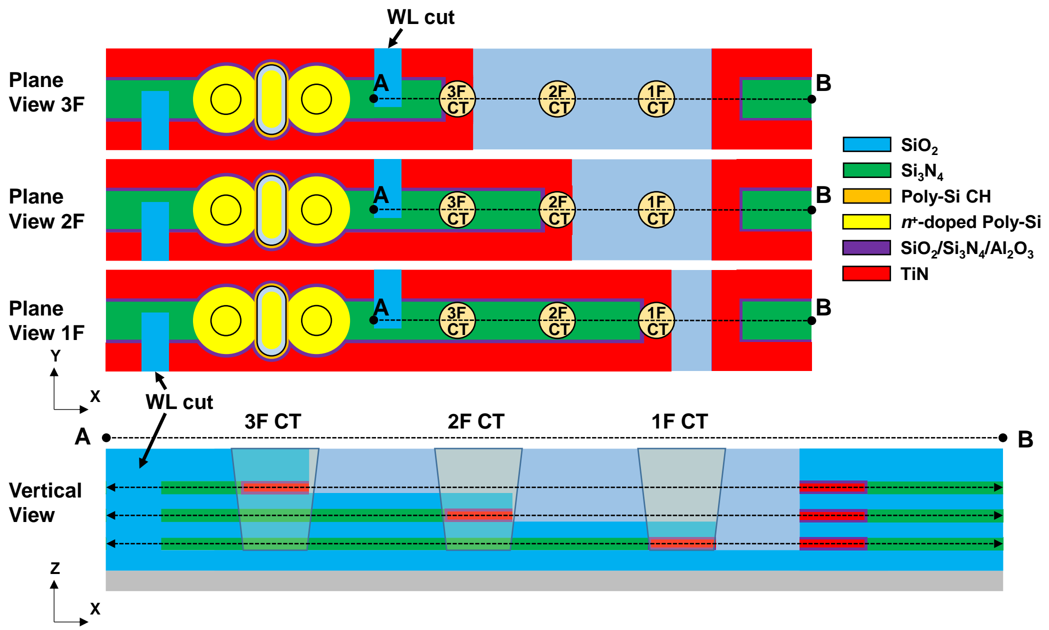



Fig. 3.14. Schematic plane views of WL of each floor and cross-sectional view of the contact pad area.

## 3.3 Cell characteristics

TEM images of a fabricated 3D AND flash cell with round-shaped channel shows the 3-layer 3D AND device as depicted in Fig. 3.15(a). The gate insulator stack consisting of $SiO_2$ /$Si_3N_4$/$Al_2O_3$ and a 10 nm-thick poly-Si channel are well identified in magnified TEM image of Fig. 3.15(b). The channel width and length of a stackable AND flash cell are 60 nm and 800 nm, respectively. However, some WLs are shown to be finely connected to one another, implying that the amount of wet etch required to properly separate the WLs was inadequate in the WL separation procedure. Therefore, in this chapter, cell characteristic analysis is performed by grouping six cells connected to one BL.

In order to evaluate synaptic characteristics on the fabricated 3D synaptic cell, programming and erasing characteristics of the fabricated stacked cell were measured and analyzed. Fig. 3.16 shows measured $I_D$-$V_G$ transfer curves of a 3D AND synaptic device in program ($V_G$ = 5~8 V, $V_S$ = $V_D$ =0 V, $t$ = 100 μs) and erase operation ($V_S$ = $V_D$ = 5~8 V, $V_G$ =0 V, $t$ = 10 ms). The fabricated 3D AND device shows a high on-off current ratio ($>10^4$) as well as sub-pA off current. ISPP and

incremental-step-pulse erasing (ISPE) using 5~8 V program and erase bias were carried out to exhibit over 1 V memory window. Due to parallel and vertical BLs and SLs in 3D AND-type arrays, FN-programming and erasing are carried out by applying same voltage to SL/BL while applying high or low voltage to the WL. Here, only positive program or erase voltages are applied to grouping WLs or the SL/BL in program and erase operation, respectively.

Fig. 3.17 shows synaptic properties of the fabricated 3D AND synaptic cell. Multi-level analog synaptic conductance can be obtained by using erase and program pulses eighty times respectively, showing >1 order of magnitude of maximum-minimum synaptic conductance ratio as shown in Fig. 3.17(b). These synaptic characteristics are measured at 7 V program ($t$ = 100 μs) and erase voltages ($t$ = 1 ms).

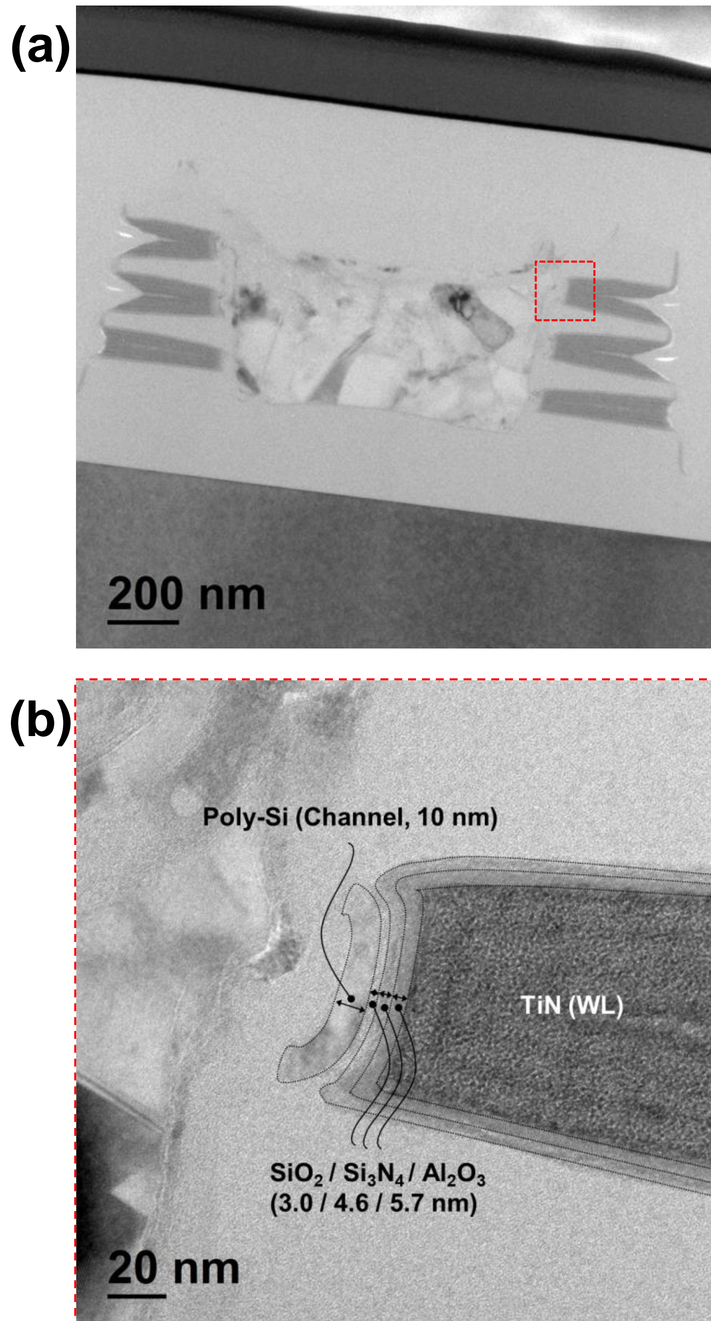



Fig. 3.15. (a) Cross-sectional TEM image of the proposed 3D AND flash device. (b) Magnified cross-sectional TEM image of a red dashed box in (a).

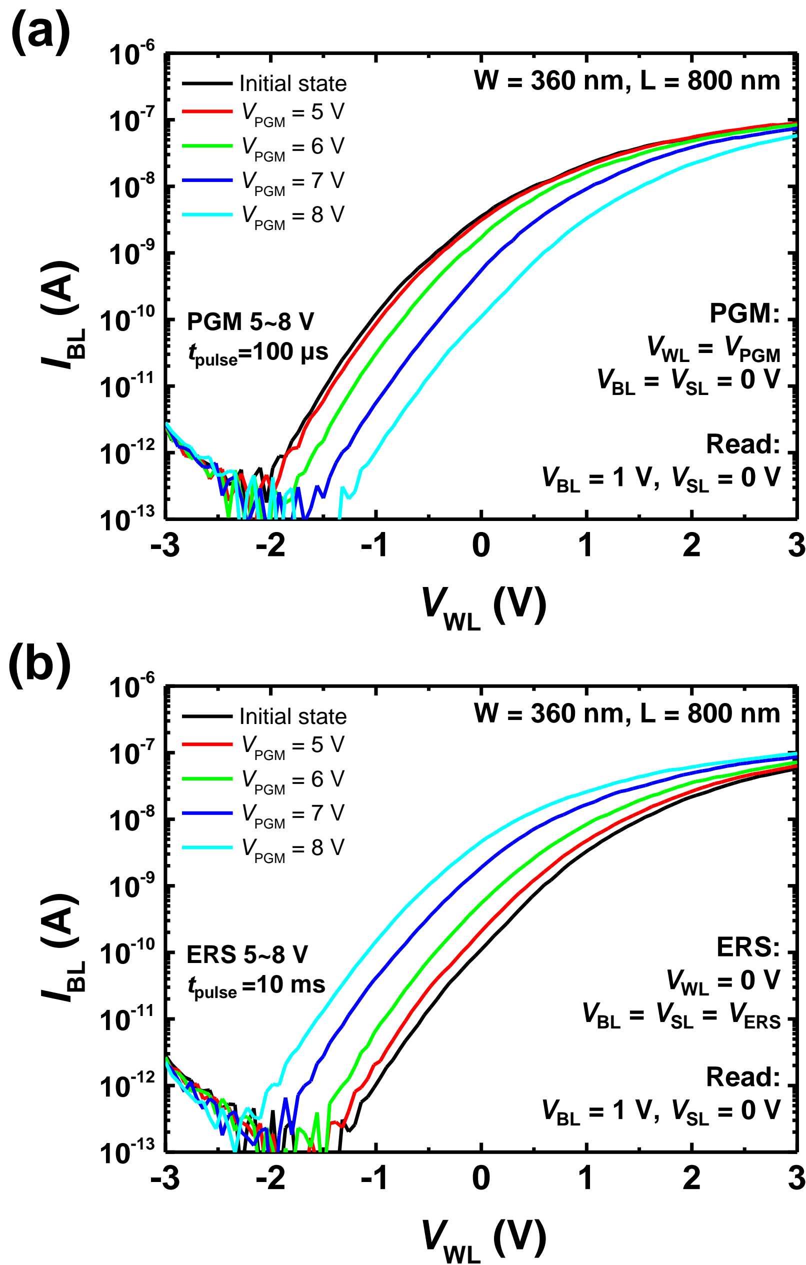



Fig. 3.16. (a) ISPP and (b) ISPE characteristics of the fabricated 3D AND flash device with round-shaped channel.

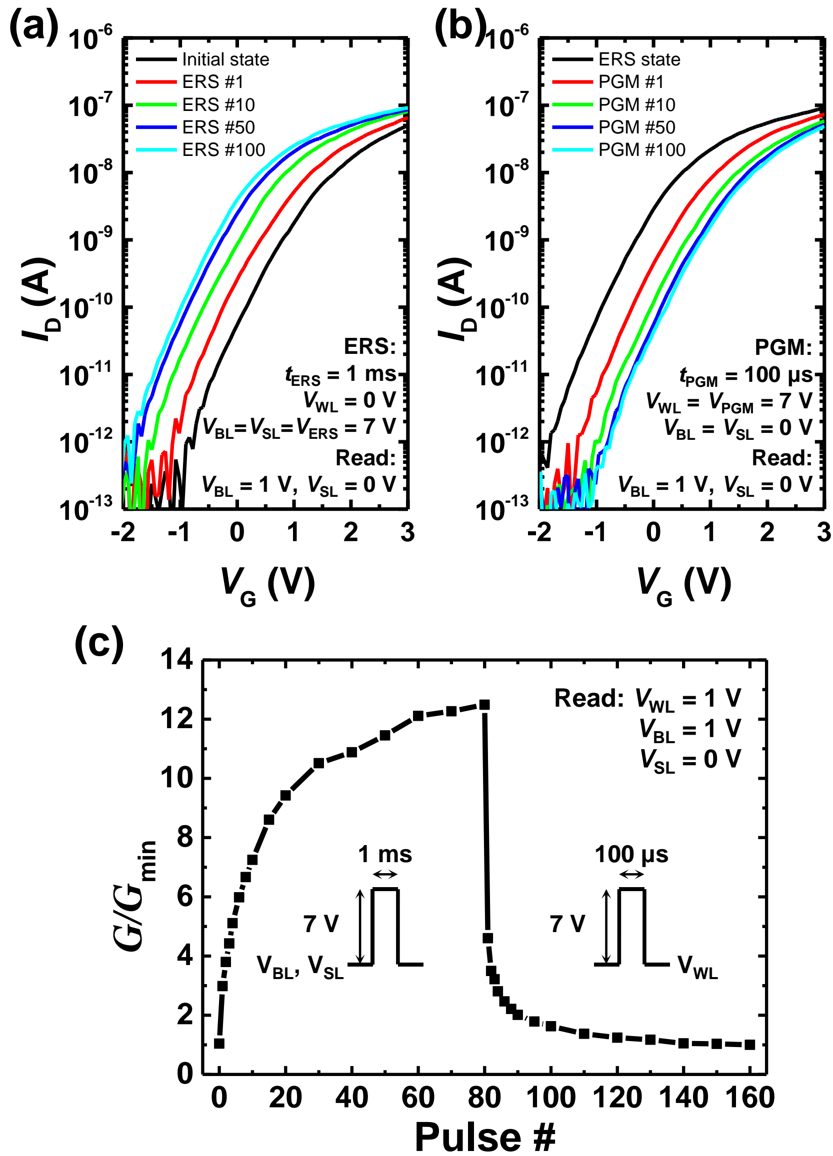



Fig. 3.17. (a), (b) Program and erase characteristics of 3D AND synaptic device obtained by identical write pulses. (c) Potentiation and depression characteristics of the 3D AND device.

## 3.4 Array characteristics

In order to evaluate memory operations in a 3D AND array, selective program and erase characteristics based on FN-tunneling are analyzed. In a 3D AND device, unlike a 2D AND device, neighboring cells are spread out in three different directions, and the characteristics of selective synaptic updates are analyzed based on the directions.

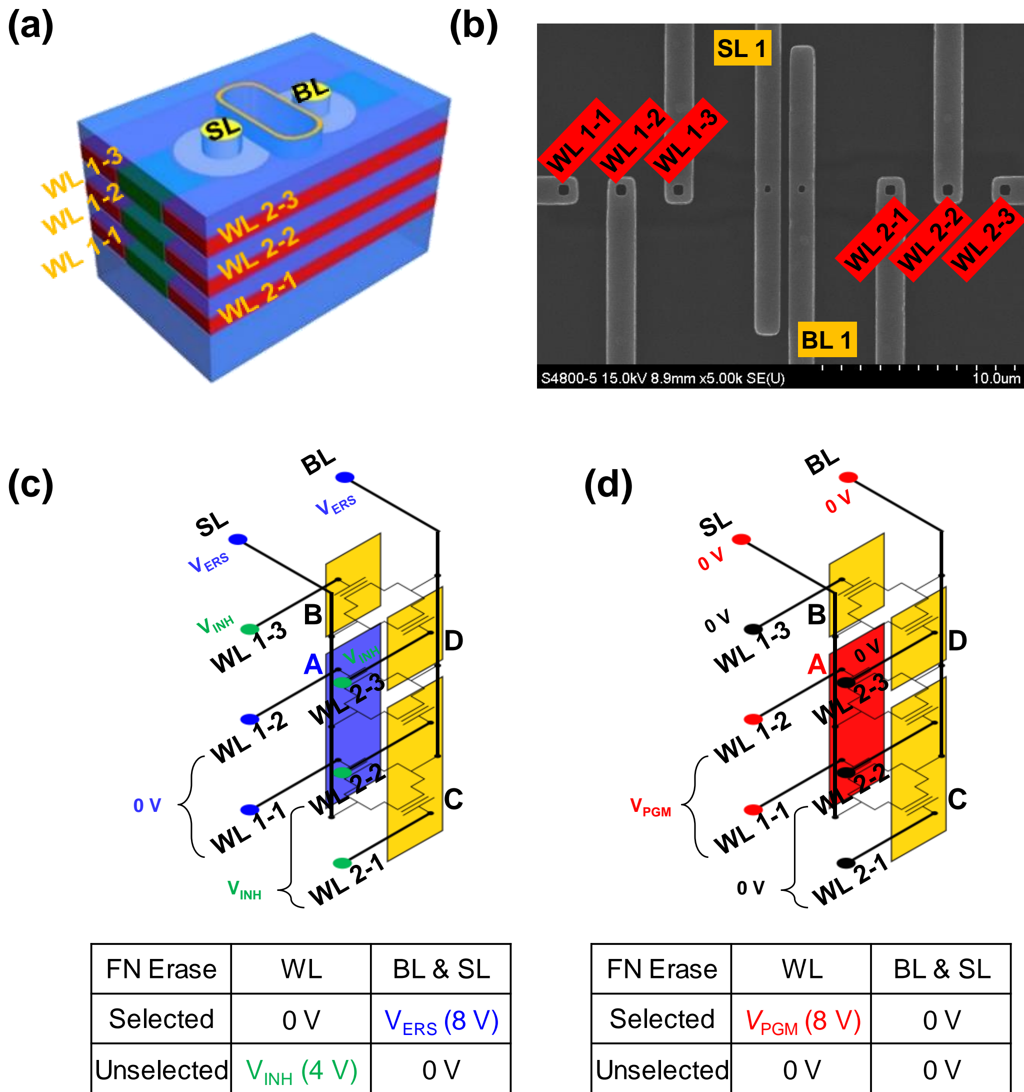
Fig. 3.18(a) and (b) show a 3D schematic diagram and top SEM image of the 2 × 1 × 3 AND array. As aforementioned, because the WL of the first and second layer were connected together due to the WL pad formation issue, the two WLs were expressed as one terminal. Therefore, it can be seen that there are four measurable cells in the 2 × 1 × 3 AND flash array. As depicted in Fig. 3.18(c) and (d), erase and program inhibition bias schemes are designed to achieve selectable erase/program performance in Z-direction in the 3D array. For selective cell erasing, a positive erase voltage ($V_{ERS}$) is applied to BL and SL of selected cell (cell A) after programming, while other unselected neighbor cells are inhibited by applying a erase inhibit voltage ($V_{INH}$) to unselected WLs. Adopting the erase inhibition

scheme in Fig. 3.18(c), a positive erase voltage ($V_{ERS}$ = 8 V, 10 ms) is applied to BL and SL of selected cell A and erase inhibit voltage of 4 V is applied to WLs of unselected cells as the erase voltage is applied. As a result of erasing using half erase voltage for erase inhibition, the current flowing in cell A increased by 6.49 nA at a read voltage of 1 V, while currents flowing in the other cells changed by less than 250 pA at the same read voltage as shown in Fig. 3.19(a). Note that all cells to which the erase inhibition voltage is applied are program-inhibited as shown in Fig. 3.19(b). For cell programming, a positive program voltage ($V_{PGM}$) is applied only to a WL of selected cell (cell A), while other WLs are set to ground. Adopting the program inhibition scheme in Fig. 3.18(d), a positive program voltage ($V_{PGM}$ = 8 V, 100 μs) is applied to the WL of selected cell A. The current flowing in cell A decreased by 5.57 nA at a read voltage of 1 V, while currents flowing in other cells changed by less than 30.1 pA at the same read voltage as shown in Fig. 3.20(a).

As depicted in Fig. 3.21(a), a 4 × 2 × 3 AND flash array is designed and has been fabricated to evaluate selective erase/program characteristics in XY-plane in the 3D array. Fig. 3.21(b) shows a 3D schematic of the 4 × 2 × 3 AND array. Note

that six cells connected to one vertical BL are grouped to measure *IV* characteristics. As shown in Fig. 3. 22(a), erase inhibition bias scheme is designed to exhibit selective program performance in the XY-plane in the 3D array. For selective cell erasing as in a 2D AND array, a positive erase voltage ($V_{ERS}$) is applied to BL and SL of selected cell (cell A) after all cells are programmed, while other unselected neighbor cells are inhibited by applying an erase inhibit voltage ($V_{INH}$) to unselected WLs. Adopting the erase inhibition scheme in Fig. 3. 22(a), a positive erase voltage ($V_{ERS}$ = 8 V, 10 ms) is applied to BL and SL of selected cell A and erase inhibit voltage of 4 V is applied to WLs of unselected cells as the erase voltage is applied. As a result of erasing using half erase voltage for erase inhibition, the current flowing in cell A increased by 24.4 nA, while currents flowing in the other cells changed by less than 16.3 pA as shown in Fig. 3. 22(c). Note that cells to which the erase inhibition voltage is applied are not programmed while preventing erase of cells sharing the BL or SL of the selected cell as shown in Fig. 3. 22(b). For selective cell programming, a program inhibition bias scheme is designed in the 3D AND array as in a 2D AND array (Fig. 3.23(a)). A program voltage ($V_{PGM}$) is applied to

WL of selected cell (cell A), while other unselected neighbor cells are inhibited by applying a program inhibit voltage ($V_{INH}$) to unselected BLs and SLs. Adopting the program inhibition scheme in Fig. 3.23(a), a program voltage of 8 V (100 μs) and a program inhibition voltage of 4 V are used. As a result of programming using half program voltage for program inhibition, the current flowing in cell A decreased by 23.2 nA, while currents flowing in other cells changed by less than 0.67 nA as shown in Fig. 3.23(c). Note that the cell to which the program inhibition voltage is applied is erase-inhibited while preventing program of cells sharing the WL of the selected cell as depicted in Fig. 3.23(b).




| FN Erase | WL | BL & SL |
| --- | --- | --- |
| Selected | 0 V | $V_{ERS}$ (8 V) |
| Unselected | $V_{INH}$ (4 V) | 0 V |

| FN Erase | WL | BL & SL |
| --- | --- | --- |
| Selected | $V_{PGM}$ (8 V) | 0 V |
| Unselected | 0 V | 0 V |

Fig. 3.18. (a) 3D schematic diagram and (b) top SEM image of the fabricated 2 × 1 × 3 AND flash array. Bias condition for selective (c) erase and (d) program operations.

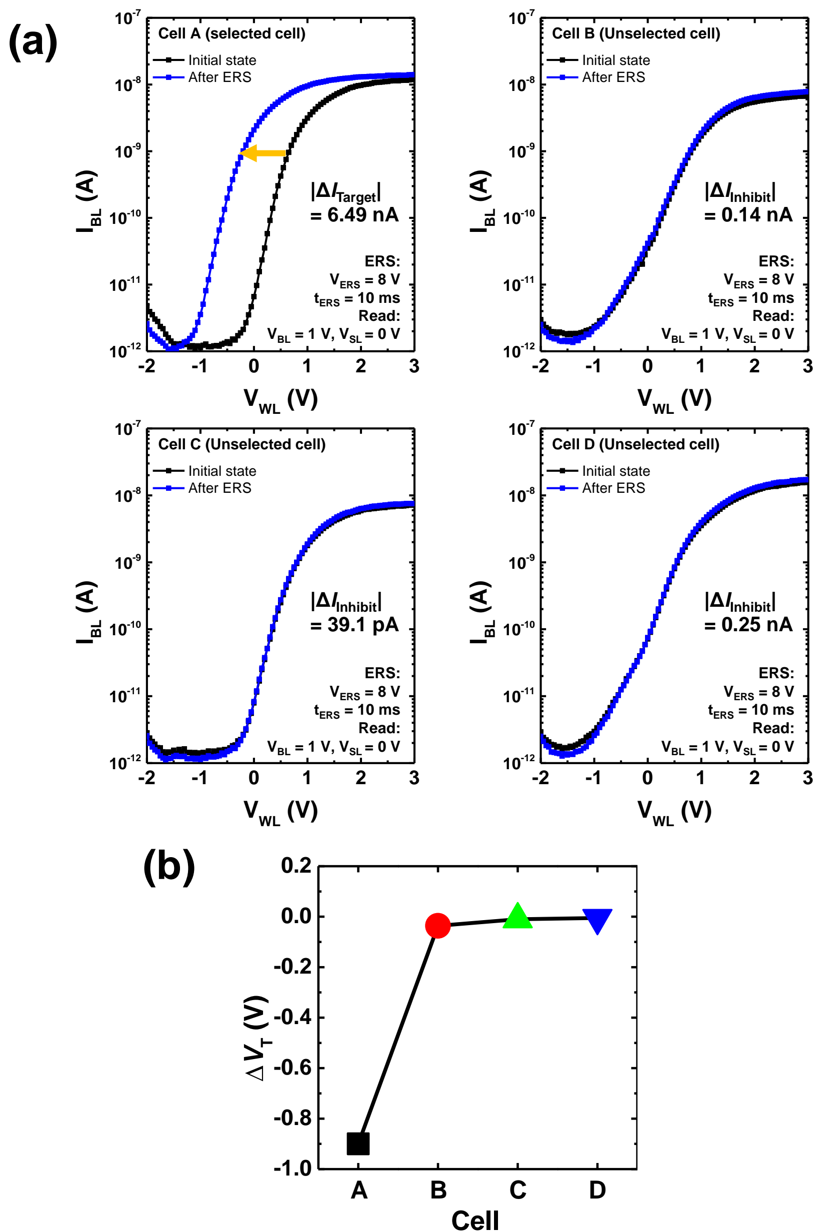



Fig. 3.19. (a) Measured selective erase properties of a 3D AND synaptic array in Z-direction. (b) Change of threshold voltages of cells in 2 × 1 × 3 AND flash array when selective erase is carried out.

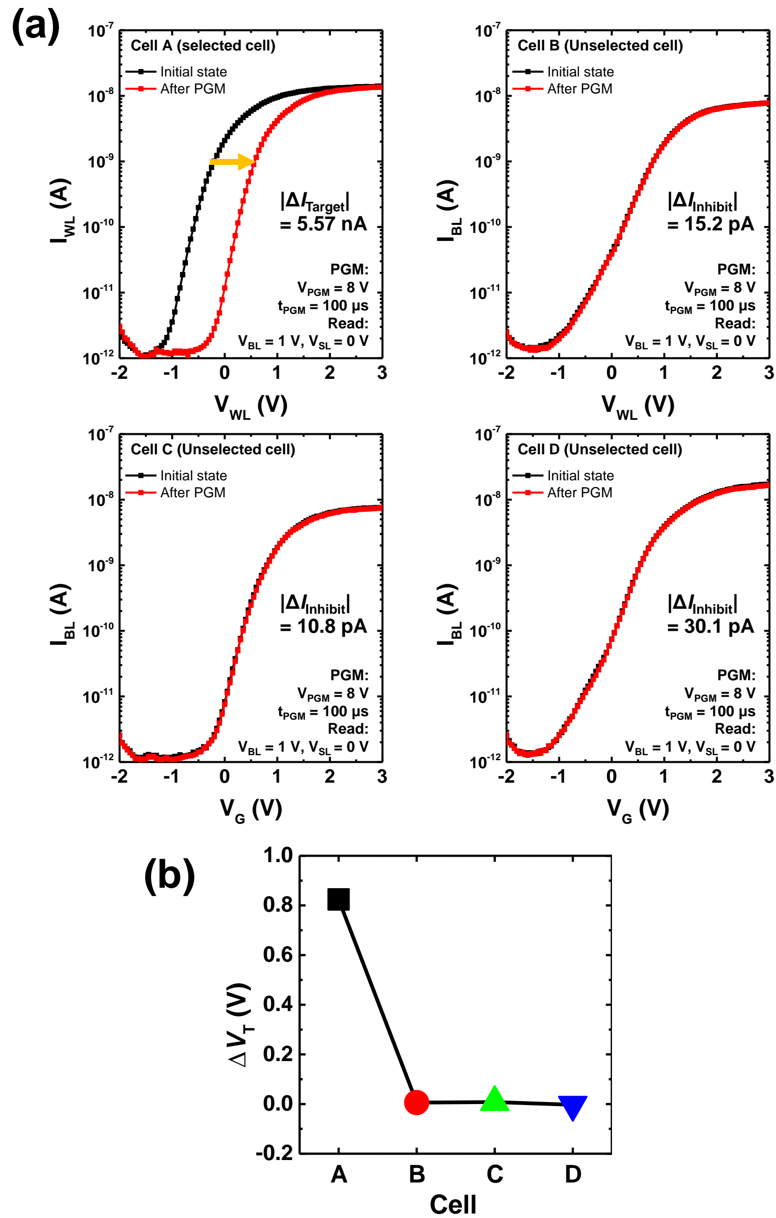
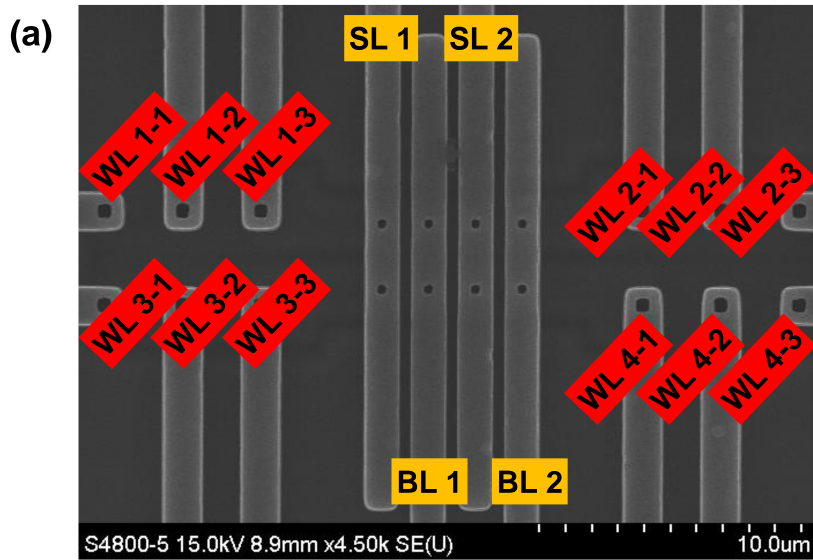



Fig. 3.20. (a) Measured selective program properties of a 3D AND synaptic array in Z-direction. (b) Change of threshold voltages of cells in 2 × 1 × 3 AND flash array when selective program is carried out.

(a)




(b)

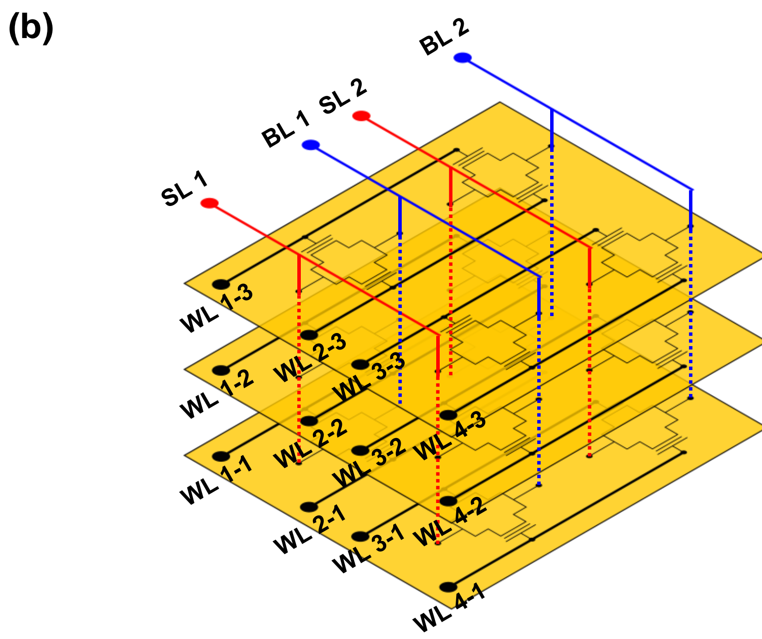
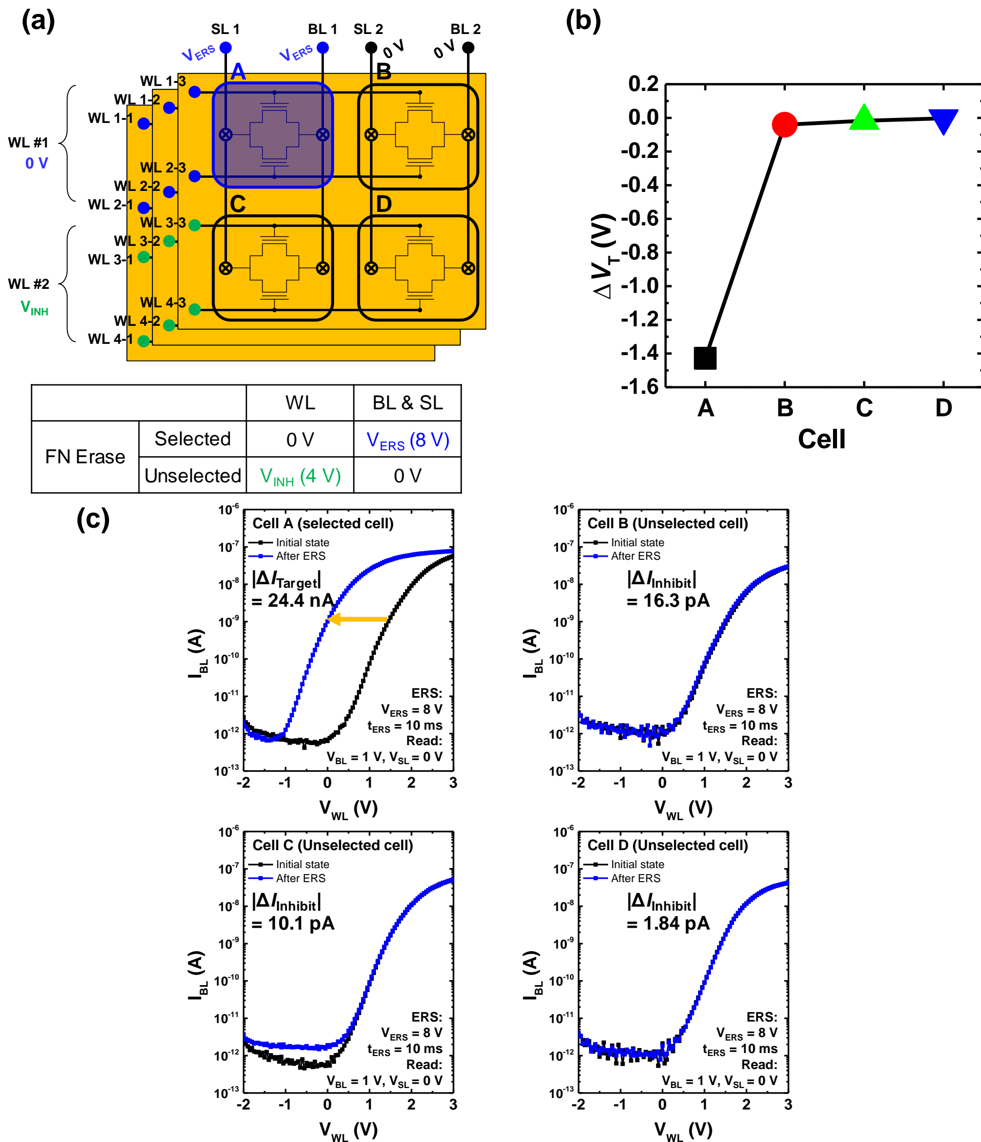
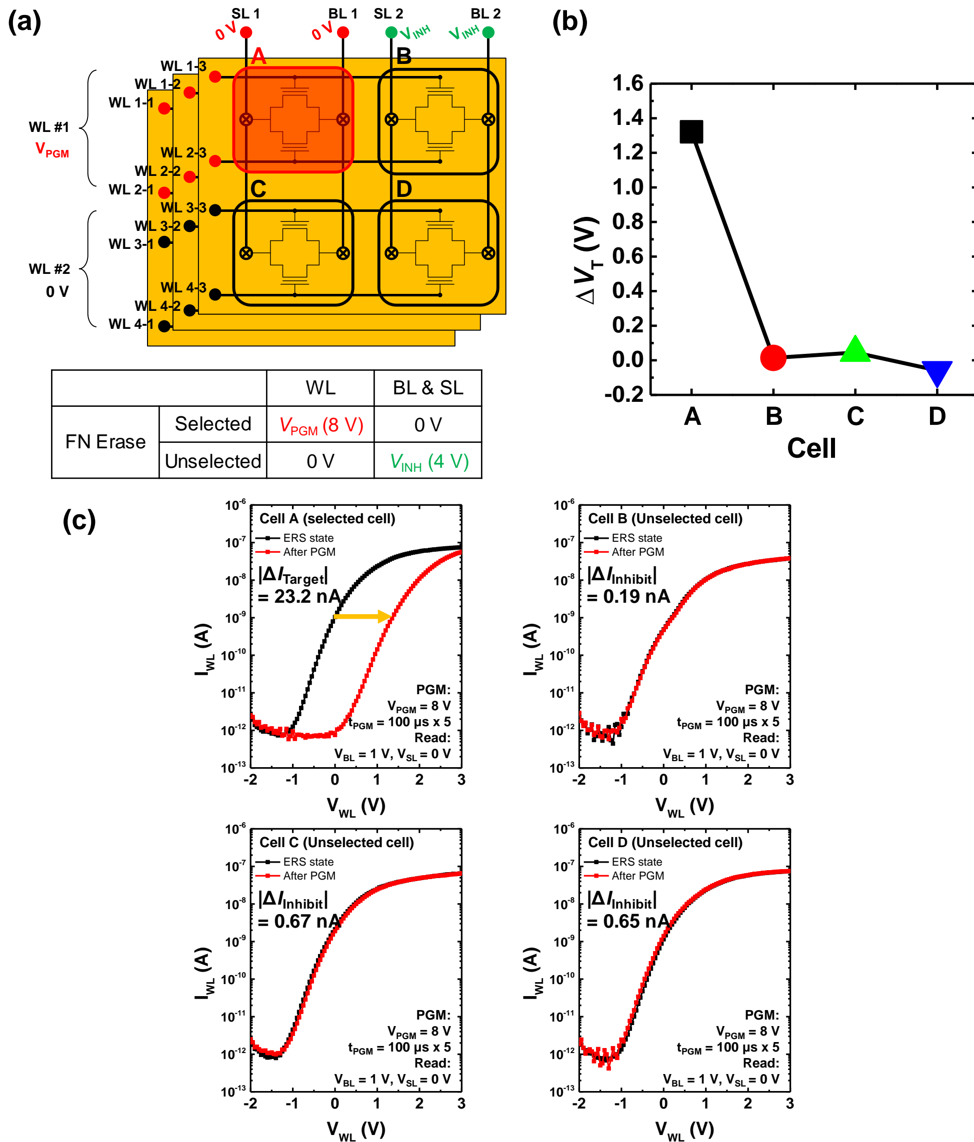



Fig. 3.21. (a) Top SEM image of the 4 × 2 × 3 AND array. (b) 3D schematic of the 4 × 2 × 3 AND array.




| | | WL | BL & SL |
|---|---|---|---|
| FN Erase | Selected | 0 V | $V_{ERS}$ (8 V) |
| | Unselected | $V_{INH}$ (4 V) | 0 V |

Fig. 3.22. (a) Bias condition for selective erase operation in the 4 × 2 × 3 AND array. (b) Selective erase properties of the 3D AND array in the XY-plane. (c) Change of threshold voltages of cells in the 3D AND array in selective erase operation.




| | | WL | BL & SL |
|---|---|---|---|
| FN Erase | Selected | $V_{PGM}$ (8 V) | 0 V |
| | Unselected | 0 V | $V_{INH}$ (4 V) |

Fig. 3.23. (a) Bias condition for selective program operation in the 4 × 2 × 3 AND array. (b) Selective program properties of the 3D AND array in the XY-plane. (c) Change of threshold voltages of cells in the 3D AND array in selective program operation.

# Chapter 4
# Off-chip learning based on AND flash synaptic Array

## 4.1 Binary neural networks based on AND flash synaptic array

### 4.1.1 AND flash synaptic architecture

In order to utilize the proposed $SiO_2$ fin-based AND array as a synaptic array, a novel synapse structure with two AND flash memory cells for BNNs is proposed to perform parallel XNOR operations and bit-counting [44]. AND-type flash-based synaptic arrays provide parallel processing thanks to their crossbar topology where each BL current is utilized to represent VMM. To realize XNOR operations in BNNs, a novel synaptic architecture with two AND flash memory cells is shown in Fig. 4.1. A synapse is made up of two neighboring AND flash cells that decide memory states in a complimentary manner; the top and bottom cells in the bit-cell design are in the erase and program states, respectively, representing a synaptic

weight of +1, and vice versa, representing a synaptic weight of −1. The AND flash-based synapse's WLs receive input voltages in its complementary configuration to represent the input value of ±1 as illustrated in Fig. 4.1(a). Note that when a read bias ($V_{read}$) is sent to the WL of an erased cell only then does the large BL current of the bit-cell flow, allowing the XNOR operation to be carried out as shown in Fig. 4.1(a) and Fib. 4.1(b). Not only are inputs transmitted concurrently into all WLs in the proposed AND flash synaptic array, but also a source-select line (SSL) and a bit-select line (BSL) are provided to minimize leakage current as shown in Fig. 4.2(a). Therefore, both the idle power consumption and the current sensing errors are diminished, allowing for fast event-driven parallel processing. By setting the reference current value of the current sensing amplifier (CSA, Fig. 4.2(b)) to half of the maximum BL current when maximal synaptic current flows to all bit-cells in the column line, it is possible to implement the 1-bit activation of each output neuron.

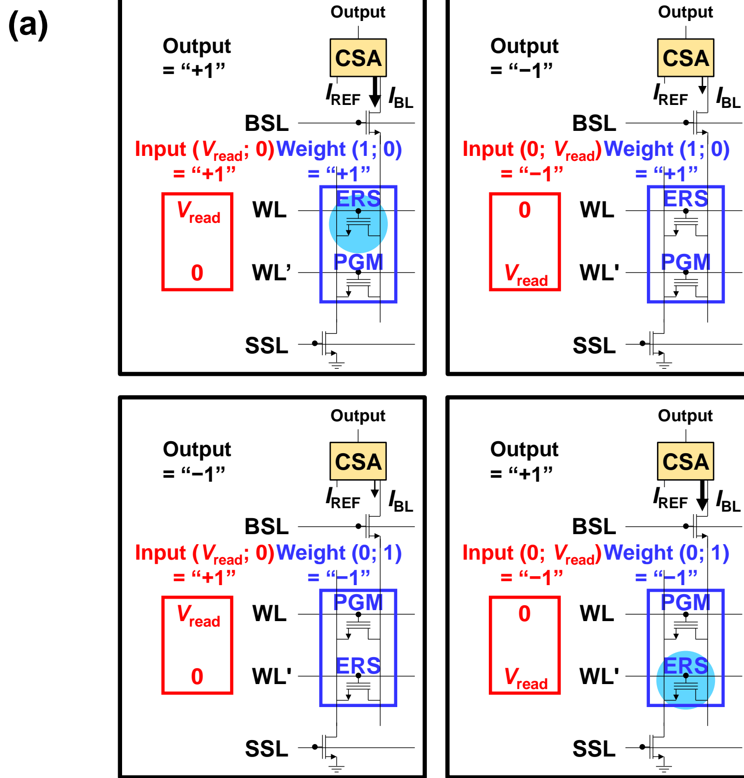
A VGG-9 binary convolutional neural network (CNN) based on binary synaptic properties of the measured AND flash memory is proposed to classify the CIFAR-

10 datasets in an effort to examine and assess the feasibility of the AND flash-based BNNs proposed as shown in Fig. 4.3(a). The VGG-9 binary CNN is used for off-chip learning; it has six convolution layers (CONV) and three fully connected layers (FC). The number of input and output channels needed for the weighted sum operation determines the size of the corresponding synaptic array in each layer. For example, a 27×128 synaptic array is used to perform a single convolution computation in the CONV 1 layer with 3×3 kernels. The FC 1 layer, which has 8192 inputs and 1024 outputs, also exhibits an 8192×1024 synaptic array, indicating that it has the largest input size of the proposed binary CNN. The details for each layer of the binary CNN are shown in Table 4.1, which can also be obtained from Fig. 4.3(a). The size of each layer's synaptic array is determined by how many input and output channels are needed for the weighted sum processing. Synaptic weights determined in a software-based binary CNN are then transmitted to the cells in synaptic arrays, taking into consideration the dynamic range and sub-pA off leakage current properties of the cells. After input activations are sent to memory array WLs, the CSA of each output neuron compares the reference current to the BL current to

determine the binary activation outcome during the inference simulation.

The influence of dynamic range on the classification rate of CIFAR-10 pictures is shown in Fig. 4.3(b). In order for BNNs to attain the software baseline accuracy, a synaptic conductance ratio of three orders of magnitude or greater is necessary in the hardware-based BNNs using CSA with reference current. In comparison to the baseline accuracy, the proposed BNN employing the $SiO_2$ fin-based synapse model with observed dynamic range and retention property degrades classification accuracy by just 0.5%. However, it is analyzed that the recognition rate decreases significantly when the 1% on-current retention loss of the proposed fin-type device is utilized as shown in Fig. 4.3(c).

(a)




(b)

| Input | WL | WL' |
|---|---|---|
| +1 | $V_{read}$ | 0 V |
| −1 | 0 V | $V_{read}$ |

| Weight | Top cell | Bottom cell |
|---|---|---|
| +1 | $V_{th,\ ERS}$ | $V_{th,\ PGM}$ |
| −1 | $V_{th,\ PGM}$ | $V_{th,\ ERS}$ |

| Input | Weight | $I_{BL}$ (output) |
|---|---|---|
| +1 | +1 | $I_{on,\ ERS}$ (+1) |
| −1 | +1 | $I_{off}$ (−1) |
| +1 | −1 | $I_{off}$ (−1) |
| −1 | −1 | $I_{on,\ ERS}$ (+1) |

Fig. 4.1. (a) Schematic diagram of synapse architecture based on two flash devices in an AND array. (b) Truth table of XNOR operation using an AND flash-based synaptic architecture.

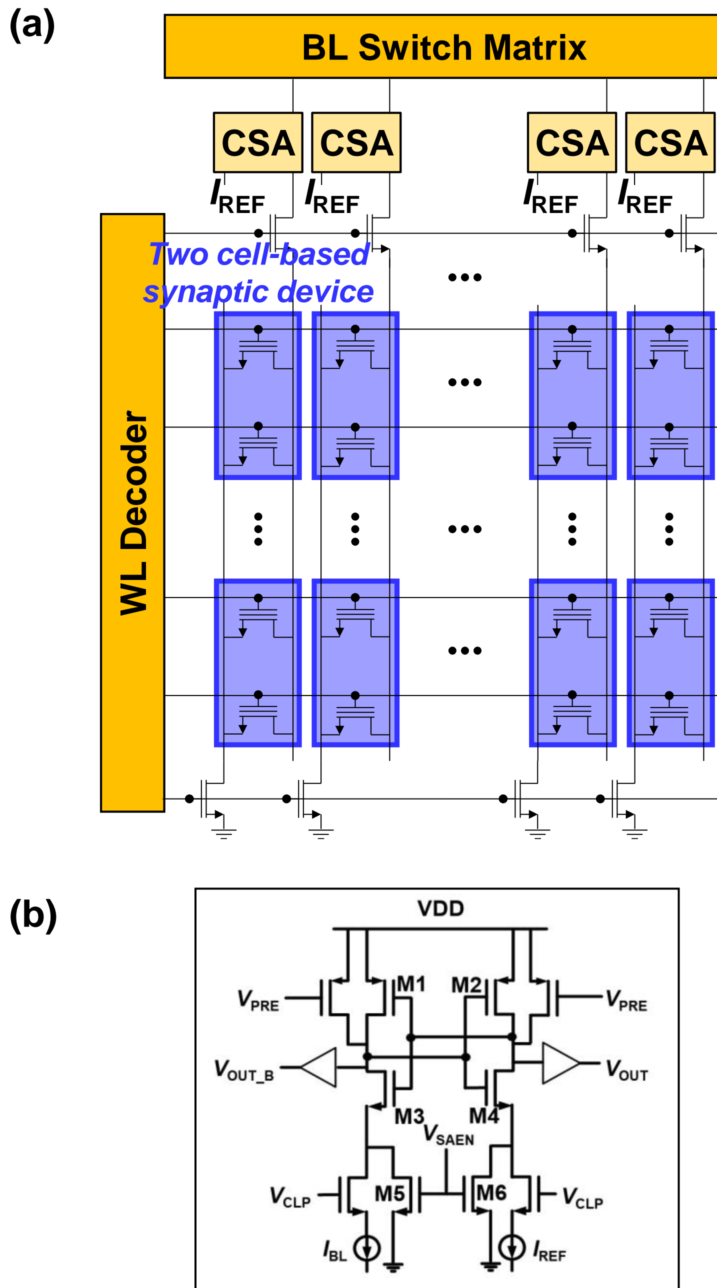
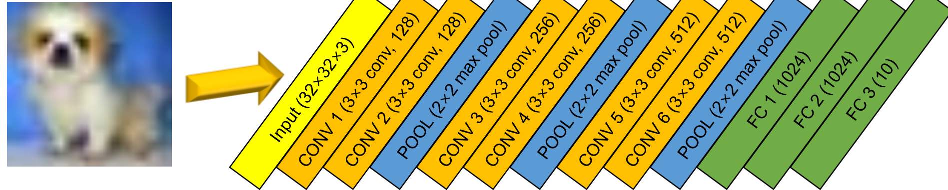
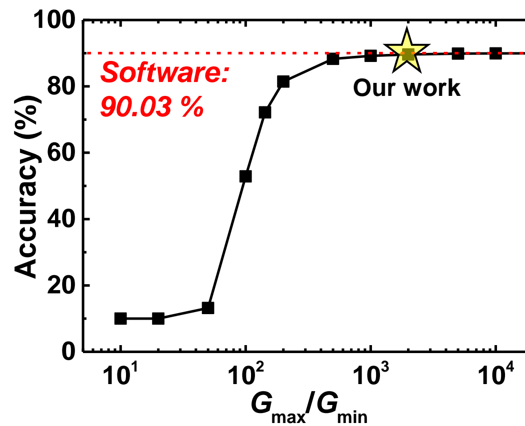



Fig. 4.2. (a) Schematic of AND array structure for BNNs. (b) A current-latch based CSA in the proposed BNNs.

(a)




(b)




(c)

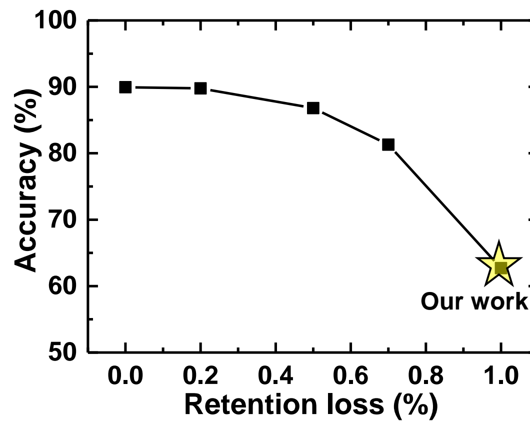



Fig. 4.3. (a) Structure of a VGG-9 based on AND flash arrays. (b) Effect of dynamic range on recognition rate of CIFAR-10. (c) Effect of retention loss of cells on recognition accuracy of CIFAR-10.

| Layer type | Kernel size | Input size | Output size | Memory array size |
|---|---|---|---|---|
| CONV 1 | 3×3 | 3×32×32 | 128×32×32 | 27×128 |
| CONV 2 | 3×3 | 128×32×32 | 128×32×32 | 1152×128 |
| Max Pool | 2×2 | 128×32×32 | 128×16×16 | - |
| CONV 3 | 3×3 | 128×16×16 | 256×16×16 | 1152×256 |
| CONV 4 | 3×3 | 256×16×16 | 256×16×16 | 2304×256 |
| Max Pool | 2×2 | 256×16×16 | 256×8×8 | - |
| CONV 5 | 3×3 | 256×8×8 | 512×8×8 | 2304×256 |
| CONV 6 | 3×3 | 512×8×8 | 512×8×8 | 4608×256 |
| Max Pool | 2×2 | 512×8×8 | 512×4×4 | - |
| FC 1 | 1×1 | 1×8192 | 1×1024 | 8192×1024 |
| FC 2 | 1×1 | 1×1024 | 1×1024 | 1024×1024 |
| FC 3 | 1×1 | 1×1024 | 1×10 | 1024×10 |

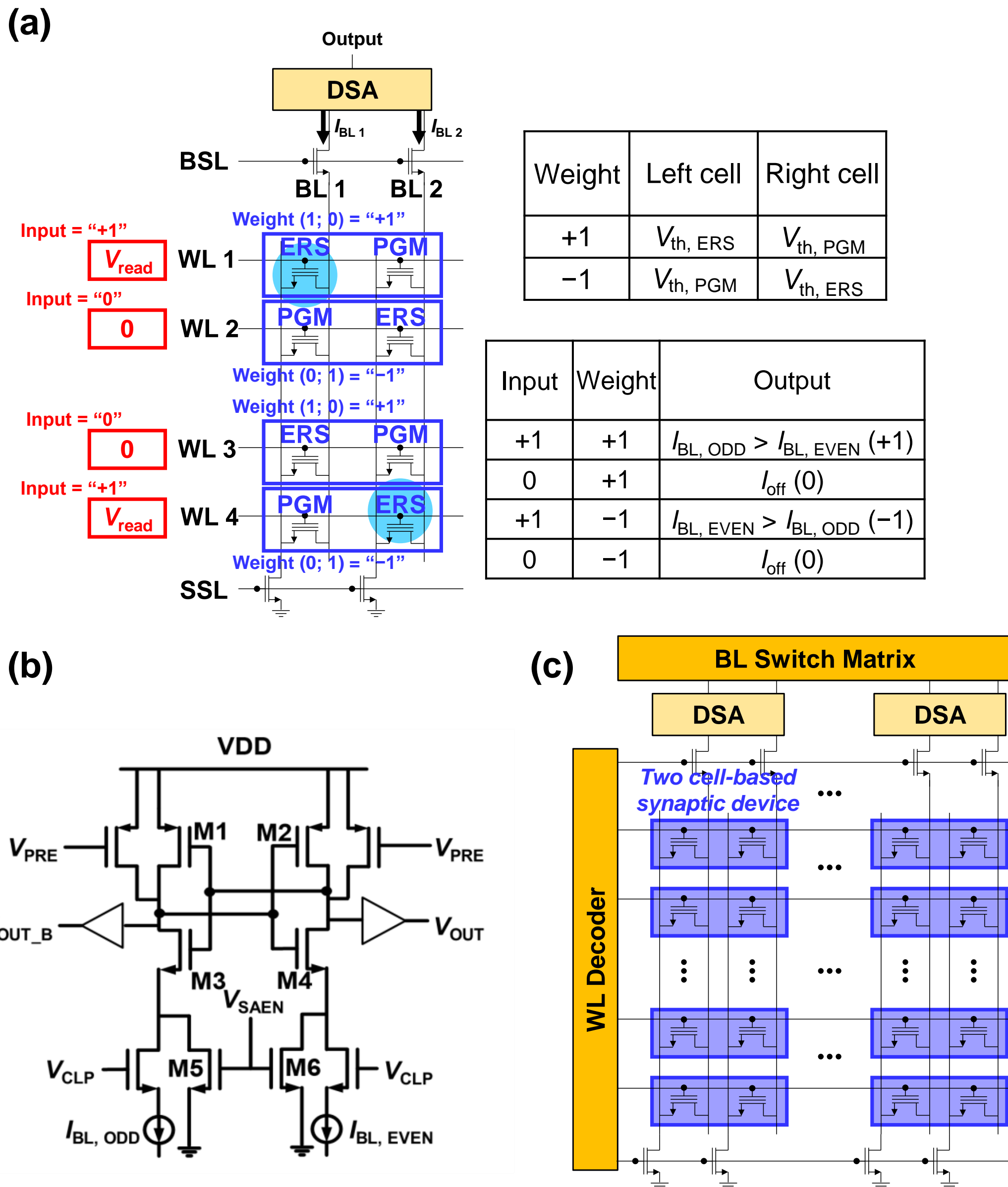
Table. 4.1. VGG-9 for CIFAR-10.

### 4.1.2 Differential synaptic architecture

To improve the robustness of AND flash-based BNNs against on-current retention loss, differential synaptic architecture using AND flash array has been proposed. Fig. 4.4 shows differential structure using two AND flash cells for XNOR operations in BNNs. Synapse consists of two AND flash cells adjacent in a direction parallel to WLs where memory states are determined in a complementary fashion; the left cell and the right cell in the bit-cell structure are in the erase state and the program state, respectively, representing a synaptic weight of +1, and vice versa, representing a synaptic weight of −1 as shown in Fig. 4.4(a). In this BNN scheme, input or output activation are used as unsigned values of 0 or 1 depending on whether or not a read bias is applied to the WL. The main difference from the previous model is that the output is determined using differential current sense amplifier (DSA, Fig. 4.4(b)). When the input read bias is applied to the WL, the odd-numbered or even-numbered BL current flows depending on the weight, and the DSA senses the difference between the two currents to obtain an output result. Fig. 4.4(c) shows AND flash memory-based array using the DSA for BNN. In the

AND flash-based BNN using DSAs as shown in Fig. 4.4(d), once inputs are transmitted to WLs of memory arrays in parallel, a DSA of each output neuron shows the unsigned activation result by comparing the odd-numbered BL current with even-numbered BL current.

Fig. 4.5 shows classification performance of an AND flash-based BNN using DSAs as function of dynamic range, on-current retention loss, and cell variation. Based on measured characteristics of $SiO_2$ fin-based AND flash cells, AND flash-based BNN using DSAs of the same VGG-9 networks in Fig. 4.3 exhibits <0.1 % degradation of classification accuracy compared to the baseline accuracy assuming 1% on-current retention loss of the device and >50 dynamic range. As shown in Fig. 4.5(c), considering 10% device variation analyzed, there is only 0.2 % degradation of recognition rate compared to the baseline accuracy.



| Weight | Left cell | Right cell |
|---|---|---|
| +1 | $V_{th,\ ERS}$ | $V_{th,\ PGM}$ |
| −1 | $V_{th,\ PGM}$ | $V_{th,\ ERS}$ |

| Input | Weight | Output |
|---|---|---|
| +1 | +1 | $I_{BL,\ ODD} > I_{BL,\ EVEN}$ (+1) |
| 0 | +1 | $I_{off}$ (0) |
| +1 | −1 | $I_{BL,\ EVEN} > I_{BL,\ ODD}$ (−1) |
| 0 | −1 | $I_{off}$ (0) |



Fig. 4.4 (a) Differential synaptic architecture using two AND flash cells. (b) A DSA comparing $I_{BL,\ ODD}$ and $I_{BL,\ EVEN}$. (c) Schematic diagram of AND flash memory-based array using the DSA for BNNs.

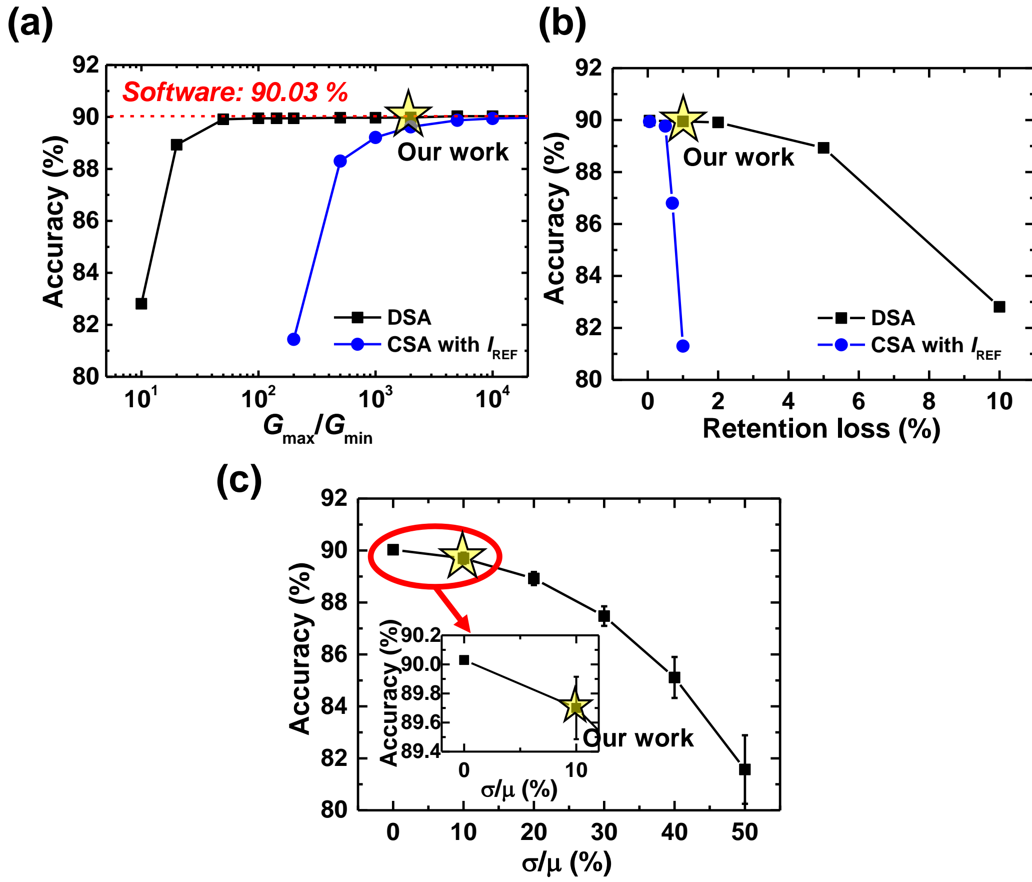



Fig. 4.5. (a) Effect of dynamic range on recognition accuracy of CIFAR-10 using the differential synaptic architecture. (b) Effect of retention loss of cells on classification accuracy of CIFAR-10. (c) Effect of device variation in BNNs on recognition accuracy on CIFAR-10.

## 4.2 Quantized neural networks based on AND flash synaptic array

Since the proposed $SiO_2$ fin-based synaptic device can implement analog weights, it can be utilized to quantized neural networks (QNNs). As shown in Fig. 2.6, over 16-level analog synaptic weights can be implemented by using identical write pulses as updating. The same VGG-9 binary CNN as mentioned in Fig. 4.3 using quantized AND flash synaptic weight behaviors is used to classify the CIFAR-10 image datasets. The ideal ReLU activation function was used, and the recognition rate during CIFAR-10 inference was analyzed with different synaptic weight levels as shown in Fig. 4.6. As a result, it can be confirmed that a high recognition rate of 92% at the 4-weight level or higher. Also, in the case of 8 weight levels which can be implemented, as a result of evaluating the performance according to the weight variation, it was confirmed that the accuracy loss hardly appeared in the measured variation of about 10%.

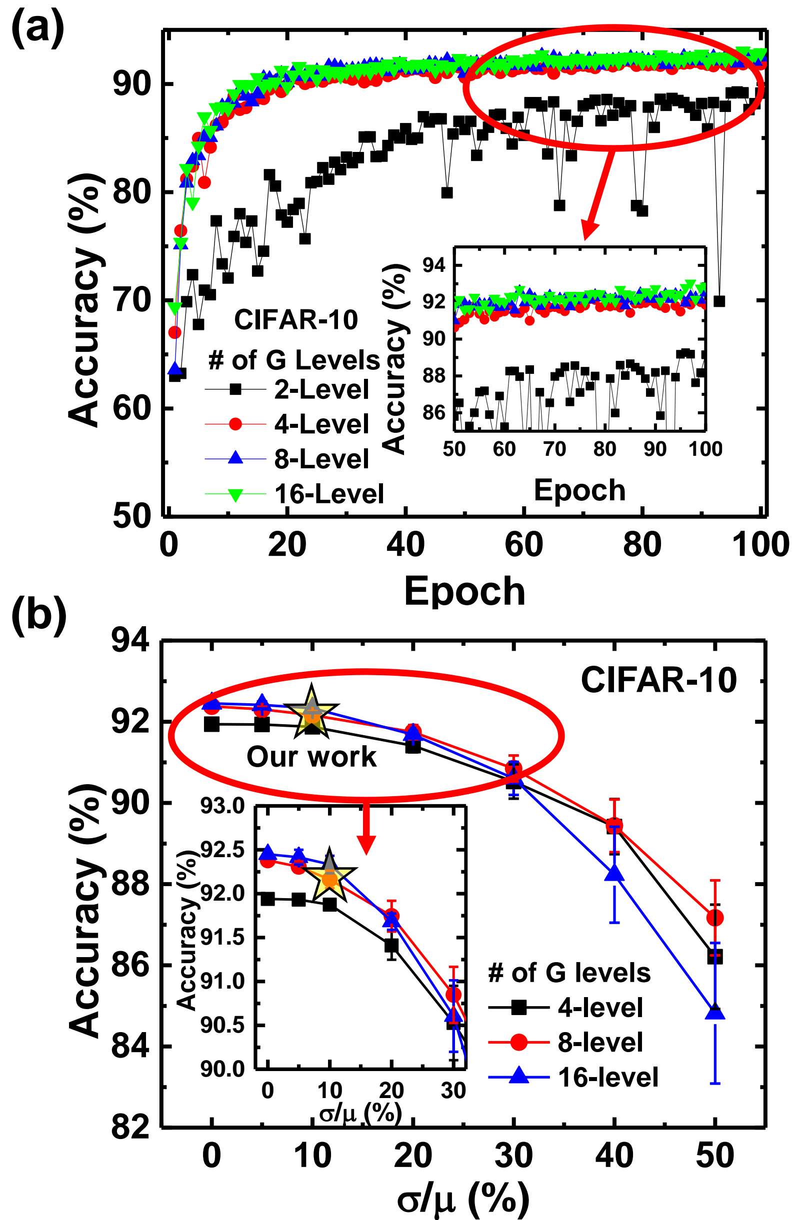



Fig. 4.6. (a) Recognition accuracy on CIFAR-10 using a QNN based on fin-type AND synaptic array. (b) Effect of device variation in QNNs on classification rates of CIFAR-10.

# Chapter 5
# Conclusion

In this work, we have proposed and fabricated a high-density synaptic device using a $SiO_2$ fin-based AND flash memory. The device's rounded-shape channel structure with a thin oxide fin of 6 nm in width allows for a reduction in program voltage caused by local field enhancement effect. The narrow oxide fin can be processed using a spacer patterning technology. The proposed device shows a high dynamic range ($>10^3$) and sub-pA off current with a low program voltage below 9 V, and achieves a larger memory window compared to the planar-type flash memory device. With the help of the introduction of a high-*k* $Al_2O_3$ blocking layer and a TiN metal gate, the proposed array achieves improved memory window, which offers low program/erase voltage to update multi-level synaptic weights. The fabricated AND array based on $SiO_2$ fin efficiently performs selective program and erase by using program and erase inhibition pulse schemes. Compared to the NOR-type flash memory array, FN tunneling can be used for selective program operation,

which serves as low-power programming. Weighted sum operation was also successfully performed by measuring BL currents when input signals are applied to multiple WLs simultaneously. In addition, a 3D AND flash synaptic array with round-shaped poly-Si channel has been designed and fabricated to improve scalability. Memory window of the 3D AND flash device with rounded channel was also ensured using ISPP and ISPE operation with a low program/erase voltage below 9 V. Selective program and erase operation in Z-direction and XY-plane of the fabricated 3-layer AND flash synaptic array was experimentally verified. A hardware-based BNN using novel two cell-based synaptic devices arranged in AND array architecture has been proposed to implement parallel XNOR operation and bit-counting, for the first time. The proposed Hardware-based BNNs require a synaptic conductance ratio greater than three orders of magnitude in order to reach the software baseline error rate. Furthermore, differential synaptic architecture using AND flash array was proposed and showed robustness against on-current retention loss using >50 dynamic range of synaptic weights. Classification performance of QNN based on AND array has also analyzed to verify applicability

of analog synaptic properties.

# Bibliography

[1] J. Backus, “Can programming be liberated from the von Neumann style?: A functional style and its algebra of programs,” *Commun. ACM*, vol. 21, no. 8, pp. 613–641, 1978.

[2] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, “A million spikingneuron integrated circuit with a scalable communication network and interface,” *Science*, vol. 345, no. 6197, pp. 668–673, 2014.

[3] G. Indiveri and S.-C. Liu, “Memory and information processing in neuromorphic systems,” *Proc. IEEE*, vol. 103, no. 8, pp. 1379–1397, 2015.

[4] G. W. Burr, R. M. Shelby, A. Sebastian, S. Kim, S. Kim, S. Sidler, K. Virwani, M. Ishii, P. Narayanan, A. Fumarola, L. L. Sanches, I. Boybat, M. Le Gallo, K. Moon, J. Woo, H. Hwang, and Y. Leblebici, “Neuromorphic computing using nonvolatile memory,” *Adv. Phys. X*, vol. 2, no. 1, pp. 89–124, 2017.

[5] S. Yu, “Neuro-inspired computing with emerging nonvolatile memory,” *Proc. IEEE*, vol. 106, pp. 260-285, 2018.

[6] D. Ielmini and H.-S.-P. Wong, “In-memory computing with resistive switching devices,” *Nature Electron.*, vol. 1, no. 6, pp. 333–343, 2018.

[7] K. K. Parhi and N. K. Unnikrishnan, “Brain-inspired computing: Models and architectures,” *IEEE Open J. Circuits Syst.*, vol. 1, pp. 185–204, 2020.

[8] D. Markovic, A. Mizrahi, D. Querlioz, and J. Grollier, “Physics for neuro-

morphic computing," *Nat. Rev. Phys.*, vol. 2, no. 9, pp. 499−510, 2020.

[9] C. A. Mead, "Neuromorphic electronic systems," *Proc. IEEE*, vol. 78, pp. 1629–1636, 1990.

[10] C. S. Poon and K. Zhou, "Neuromorphic silicon neurons and large-scale neural networks: challenges and opportunities", *Front. Neurosci.*, vol. 5, no. 108, pp. 1-3, 2011.

[11] D. Kuzum, S. Yu, H.S. Wong, "Synaptic electronics: materials, devices and applications," *Nanotechnology*, vol. 24, no. 38, pp. 1–22, 2013.

[12] C. Merkel, R. Hasan, N. Soures, D. Kudithipudi, T. Taha, S. Agarwal, M. Marinella, "Neuromemristive systems: Boosting efficiency through brain-inspired computing," *Computer*, vol. 49, pp. 56-64, 2016.

[13] M. A. Zidan, J. P. Strachan, and W. D. Lu, "The future of electronics based on memristive systems," *Nat. Electron.*, vol. 1, no. 1, pp. 22–29, 2018.

[14] G. Chakma, N. D. Skuda, C. D. Schuman, J. S. Plank, M. E. Dean, and G. S. Rose, "Energy and area efficiency in neuromorphic computing for resource constrained devices," *in Proc. ACM Great Lakes Symp. VLSI*, 2018, pp. 379–383.

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[17] O. Bichler, M. Suri, D. Querlioz, D. Vuillaume, B. DeSalvo, and C. Gamrat, "Visual pattern extraction using energy-efficient "2-PCM Synapse" neuromorphic architecture," *IEEE Trans. Electron Devices*, vol. 59, pp. 2206-2214, 2012.

[18] M. Prezioso, F. Merrikh-Bayat, B.D. Hoskins, G.C. Adam, K.K. Likharev,

D.B. Strukov, “Training and operation of an integrated neuromorphic network based on metal-oxide memristors,” *Nature*, vol. 521, pp. 61-64, 2015.

[19] S. Yu, Y. Wu, R. Jeyasingh, D. Kuzum, and H. -S. P. Wong, "An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation," *IEEE Trans. Electron Devices*, vol. 58, pp. 2729-2737, 2011.

[20] 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,” *in IEDM Tech. Dig.*, 2011, pp. 4.4.1–4.4.4.

[21] D. Kuzum, R. G. D. Jeyasingh, B. Lee, and H.-S. Philip Wong, “Nano-electronic programmable synapses based on phase change materials for brain-inspired computing,” *Nano lett.*, vol. 12, no. 5, pp. 2179-2186, 2012.

[22] S. Yu, B. Gao, Z. Fang, H. Yu, J. Kang, and H.-S. P. Wong, “A neuromorphic visual system using RRAM synaptic devices with sub-pJ energy and tolerance to variability: Experimental characterization and largescale modeling,” *in IEDM Tech. Dig.*, 2012, pp. 10.4.1–10.4.4.

[23] S. B. Eryilmaz, D. Kuzum, R. G. D. Jeyasingh, S. Kim, M. BrightSky, C. Lam, and H.-S. P. Wong, “Experimental demonstration of arraylevel learning with phase change synaptic devices,” *in IEDM Tech. Dig.*, 2013, pp. 25.5.1–25.5.4.

[24] G. W. Burr, R. M. Shelby, S. Sidler, C. di Nolfo, J. Jang, I. Boybat, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. N. Kurdi, and H. Hwang, “Experimental demonstration and tolerancing of a largescale neural network (165000 synapses) using phase-change memory as the synaptic weight element,” *IEEE Trans. Electron Devices*, vol. 62, pp. 3498-3507, 2015.

[25] J.-W. Jang, S. Park, G.W. Burr, H. Hwang, Y.-H. Jeong, “Optimization of conductance change in PrCaMnO based synapse devices for neuromorphic systems,”

*IEEE Electron Device Lett.,* vol. 36, pp. 457-459, 2015.

[26] J. Woo, K. Moon, J. Song, S. Lee, M. Kwak, J. Park, and H. Hwang, “Improved synaptic behavior under identical pulses using $AlO_x/HfO_2$ bilayer RRAM array for neuromorphic systems,” *IEEE Electron Device Lett.*, vol. 37, no. 8, pp. 994–997, 2016.

[27] S. Ambrogio, S. Balatti, V. Milo, R. Carboni, Z.-Q. Wang, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Neuromorphic learning and recognition with one-transistor-one-resistor synapses and bistable metal oxide RRAM," *IEEE Trans. Electron Devices*, vol. 63, pp. 1508-1515, 2016.

[28] 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,” *in IEDM Tech. Dig.*, 2016, pp. 440-443.

[29] B. Gao, H. Wu, J. Kang, H. Yu, and H. Qian, “Oxide-based analog synapse: Physical modeling, experimental characterization, and optimization,” *in IEDM Tech. Dig.*, 2016, pp. 3–7.

[30] X. Sun, S. Yin, X. Peng, R. Liu, J.-S. Seo, and S. Yu, “XNOR-RRAM: A scalable and parallel resistive synaptic architecture for binary neural networks,” *in Proc. Design, Automat. Test Eur. Conf. Exhib. (DATE),* 2018, pp. 1423–1428.

[31] M. Bocquet, T. Hirztlin, J.-O. Klein, E. Nowak, E. Vianello, J.-M. Portal, and D. Querlioz, “In-memory and error-immune differential RRAM implementation of binarized deep neural networks,” *in IEDM Tech. Dig.*, 2018, pp. 20.6.1–20.6.4.

[32] Z. Zhou, P. Huang, Y. C. Xiang, W. S. Shen, Y. D. Zhao, Y. L. Feng, B. Gao, H. Q. Wu, H. Qian, L. F. Liu, X. Zhang, X. Y. Liu, and J. F. Kang, “A new hardware implementation approach of BNNs based on nonlinear 2T2R synaptic cell,” *in IEDM Tech. Dig.,* 2018.

[33] S. Yin, X. Sun, S. Yu, and J.-S. Seo, “High-throughput in-memory computing

for binary deep neural networks with monolithically integrated RRAM and 90-nm CMOS," *IEEE Trans. Electron Devices*, vol. 67, no. 10, pp. 4185–4192, 2020.

[34] J.-H. Bae, S. Lim, B.-G. Park, J.-H. Lee, "High-density and near-linear synapse device based on a reconfigurable gated Schottky diode," *IEEE Electron Device Lett.*, vol. 38, pp. 1153-1156, 2017.

[35] H. Kim, S. Hwang, J. Park, and B.-G. Park, "Silicon synaptic transistor for hardware-based spiking neural network and neuromorphic system," *Nanotechnol.*, vol. 28, no. 40, pp. 405202-1–405202-10, 2017.

[36] 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.

[37] G. Malavena, A. S. Spinelli, and C. M. Compagnoni, "Implementing spike-timing-dependent plasticity and unsupervised learning in a mainstream NOR flash memory array," *in IEDM Tech. Dig.*, 2018, pp. 35–38.

[38] F. Merrikh-Bayat, X. Guo, M. Klachko, M. Prezioso, K. K. Likharev, and D. B. Strukov, "High-performance mixed-signal neurocomputing with nanoscale floating-gate memory cell arrays," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4782–4790, 2018.

[39] H.-T. Lue, W. Chen, H.-S. Chang, K.-C. Wang, and C.-Y. Lu, "A novel 3D AND-type NVM architecture capable of high-density, low-power in-memory sum-of-product computation for artificial intelligence application," *in Proc. IEEE Symp. VLSI Technol.*, 2018.

[40] Y. Noh, Y. Seo, B. Park, and J.-H. Lee, "Synaptic devices based on 3-D AND flash memory architecture for neuromorphic computing," in *Proc. IEEE Int. Memory Workshop*, 2019.

[41] 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,” *in IEDM Tech. Dig.*, 2019.

[42] H.-T. Lue, G.-R. Lee, T.-H. Yeh, T.-H. Hsu, C. Lo, C.-L. Sung, W.-C. Chen, C.-T. Huang, K.-Y. Shen, M.-Y. Wu, P. Tseng, M.-F. Hung, C.-J Chiu, K.-Y. Hsieh, K.-C. Wang, and C.-Y. Lu, "3D AND: a 3D stackable flash memory architecture to realize high-density and fast-read 3D NOR flash and storage-class memory," *in IEDM Tech. Dig.*, 2020, pp. 6.4.1-6.4.4.

[43] W.-M. Kang, D. Kwon, S. Y. Woo, S. Lee, H. Yoo, J. Kim, B.-G. Park, and 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.

[44] S. Lee, H. Kim, S.-T. Lee, B.-G. Park, and J.-H. Lee, "$SiO_2$ fin-based flash synaptic cells in AND array architecture for binary neural networks," *IEEE Electron Device Lett.*, vol. 43, no. 1, pp. 142-145, 2022.

[45] A. Aziz, O. Bonnaud, H. Lhermite, and F. Raoult, “Lateral polysilicon pn diodes: Current-voltage characteristics simulation between 200 K and 400 K using a numerical approach,” *IEEE Trans. Electron Devices*, vol. 41, no. 2, pp. 204–211, 1994.

[46] Y. Morimoto, Y. Jinno, K. Hirai, H. Ogata, T. Yamada, and K. Yoneda, “Influence of the grain boundaries and intragrain defects on the performance of poly-Si thin film transistors,” *J. Electrochem. Soc.*, vol. 144, no. 7, pp. 2495–2501, 1997.

[47] M. Hirose, ‘‘Electron tunneling through ultrathin $SiO_2$,’’ *Mater. Sci. Eng., B*, vol. 41, no. 1, pp. 35–38, 1996.

[48] T.-H. Hsu, H.-T. Lue, E.-K. Lai, J.-Y. Hsieh, S.-Y. Wang, L.-W. Yang, Y.-C.

King, T. Yang, K.-C. Chen, K.-Y. Hsieh, R. Liu, and C.-Y. Lu, “A high-speed BE-SONOS NAND flash utilizing the field-enhancement effect of FinFET,” *in IEDM Tech. Dig.*, 2007, pp. 913-916.

[49] S.-C. Lai, H.-T. Lue, C.-W. Liao, T.-B. Wu, M.-J. Yang, Y.-H. Lue, J.-Y. Hsieh, S.-Y. Wang, G.-L. Luo, C.-H. Chien, K.-Y. Hsieh, R. Liu, and C.-Y. Lu, “Highly reliable MA BE-SONOS (Metal-$Al_2O_3$ bandgap engineered SONOS) using a $SiO_2$ buffer layer,” *in Proc. VLSI-TSA*, 2008, pp. 58–59.

# 초 록

뉴로모픽 컴퓨팅 시스템은 생물학적 신경계를 모방하여 폰 노이만 병목 현상을 극복하는 새로운 인공 지능 패러다임으로 등장하였다. 뉴로모픽 컴퓨팅 시스템의 하드웨어 기반 신경망을 위한 시냅스 소자는 병렬 연산 가능성, 높은 집적도, 저전력 동작, 선택적인 쓰기 동작을 필요로 한다. 본 논문에서는, 하드웨어 기반 신경망을 위한 $SiO_2$ 핀 기반의 AND 플래시 메모리 어레이를 제안한다. 6 nm 폭의 얇은 산화물 핀 기반의 원형 채널 구조를 갖는 제안된 소자는 국부적으로 전계를 강화하여 평면형 채널 구조의 플래시 시냅스 소자 대비 프로그램 성능을 향상시킨다. AND 플래시 셀은 $10^5$ 이상의 높은 온/오프 전류 비율, pA 미만의 오프 전류, 그리고 9 V 이하의 낮은 프로그래밍 전압을 사용하여 $10^3$ 이상의 높은 시냅스 가중치의 동적 범위를 보인다. $SiO_2$ 핀을 기반으로 제작된 AND 어레이에서는 프로그램 및 이레이즈 억제 펄스 방식을 사용하여 선택적 쓰기 동작이 효율적으로 수행되고 가중치 합 동작이 실험적으로 검증된다. 또한, 집적도를 높이기 위해 원형 폴리실리콘 채널을 갖는 3D AND 플래시 시냅틱 어레이가 설계 및 제작된다. 오정렬 문제를 해결하기 위한 주요 공정 단계가 제안된다. 제안된 3차원 AND 어레이는 프로그램 및 이레이즈 억제 펄스 방식을 사용하여 선택적 쓰기 동작을 수행한다.

오프 칩 학습을 위해 두 개의 AND 플래시 메모리 셀을 기반으로 하는 새로운 시냅스 아키텍처를 제안한다. AND 플래시 셀 기반의 새로운

시냅스 구조는 이진 신경망을 위한 병렬 XNOR 연산과 비트-셈을 수행하도록 사용된다. AND 플래시 어레이 기반의 제안된 이진 신경망은 CIFAR-10 데이터에서 이상적인 소프트웨어 기반 이진신경망의 인식 정확도와 유사한 89.9%의 정확도를 보인다. 나아가 우리는 AND 플래시 어레이를 이용한 차동 시냅스 아키텍처을 제안하여 전류 유지 손실에 대한 안정성을 높인다.

주요어: AND 플래시 메모리, 시냅스 소자, 핀형 플래시 소자, 3차원 플래시 메모리, 하드웨어 기반 신경망, 이진 신경방, 뉴로모픽 시스템.

학번: 2016-23288