M.S. DISSERTATION

Synaptic Devices Based on 3-D AND Flash Memory Architecture for Neuromorphic Computing

뉴로모픽 컴퓨팅을 위한 3D 및 플래시 메모리 아키텍처를 기반으로 한 시냅스 모방소자

by

YOO-HYUN NOH

August 2019

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

Synaptic Devices Based on 3-D AND Flash Memory Architecture for Neuromorphic Computing

뉴로모픽 컴퓨팅을 위한 3D 및 플래시 메모리 아키텍처를 기반으로 한 시냅스 모방소자

지도교수 이 종 호

이 논문을 공학석사 학위논문으로 제출함

2019 년 8 월

서울대학교 대학원

전기컴퓨터공학부

노 유 현

노유현의 공학석사 학위논문을 인준함

2019 년 8 월

위 원 장 : 박 병 국 (인)

부위원장 : 이 종 호 (인)

위 원 : 김 재 하 (인)

Synaptic Devices Based on 3-D AND Flash Memory Architecture for Neuromorphic Computing

by

Yoo-Hyun Noh

Advisor: Jong-Ho Lee

Confirming the master thesis written by

YOO-HYUN NOH

(Electrical Engineering and Computer Science)

in Seoul National University

August 2019

Professor Byung-Gook Park, Chair

Professor Jong-Ho Lee, Vice-Chair

Associate Professor Jaeha Kim

ABSTRACT

Conventional “Von Neumann architecture” performs computation on a CPU, which is connected to a memory device (DRAM) via buses. As a result, power and speed are the biggest limitations of today’s high-capacity computing. On the other hand, neuromorphic computing, which imitates the computation of brain, is highly desired due to the ability to operate memory and logic in parallel.

A neuromorphic computing is accomplished by the modification of connection strength of neurons and synapses. Synapses have a memory function similar to that used in conventional computing. For a successful computation, a synaptic device should have following characteristics: high scalability, low power, high speed, high reliability, and non-volatility. Unfortunately, there are no standardized devices that satisfy all these conditions. Even though various memristor-based devices have been reported as synaptic devices, the devices are still being studied for reliability and variation issues. NOR flash is much more mature, but it is less scalable because of the large area of one memory cell (~$10F^2$). In addition, high-density NAND flash

memory is difficult to implement the current summation required for neuromorphic computing because the cells are connected in series in a string.

In this dissertation, I propose a new synaptic device that can effectively perform neuromorphic computing in a 3-D stack structure. A new 3-D synaptic device with stackable AND-type Rounded Dual Channel (RDC) flash memory architecture operates at low power by using the FN program/erase method, utilizes a high speed by a parallel read operation, and performs in a high density with multi-layer stacking. Key fabrication steps are explained and the successful operation of the device in 3-D stacked structure is verified by device simulation. In addition, devices are fabricated by stacking three layers, and their operation is confirmed. The proposed 3-D stacked AND architecture and device structure proposed in this work is expected to be a promising candidate for highly integrated synaptic devices.

Keywords: AI, Neuromorphic, Synaptic device, 3D stackable, AND flash, NOR flash.

Student number: 2017-27245

CONTENTS

List of Figures

Chapter 1

Introduction

1.1 Study background

Recently, the machine learning has attracted a great deal of attention in the IT industry and is being developed rapidly with the performance enhancement of the GPU-based hardware accelerator. the deep neural network technology based on the back-propagation (BP) algorithm has shown excellent performance in many areas including image, speech recognition, and translation, even sometimes outperforms human cognitive abilities [1]. However, there are important challenges about power consumption, speed, occupied area of the hardware platform and training times. Therefore, the need for implementing neuromorphic artificial neural networks (ANNs) with low power, high speed and small area has been emerging. A neuromorphic computing, which imitates the computation of brain, is highly desired due to the ability to operate memory and logic in parallel [2] (Figure 1.1). Since the neuromorphic computing consists of only circuits necessary for neural

network operation, it can benefit hundreds of times more in terms of power, area and speed than using CPU and GPU to compute neural network

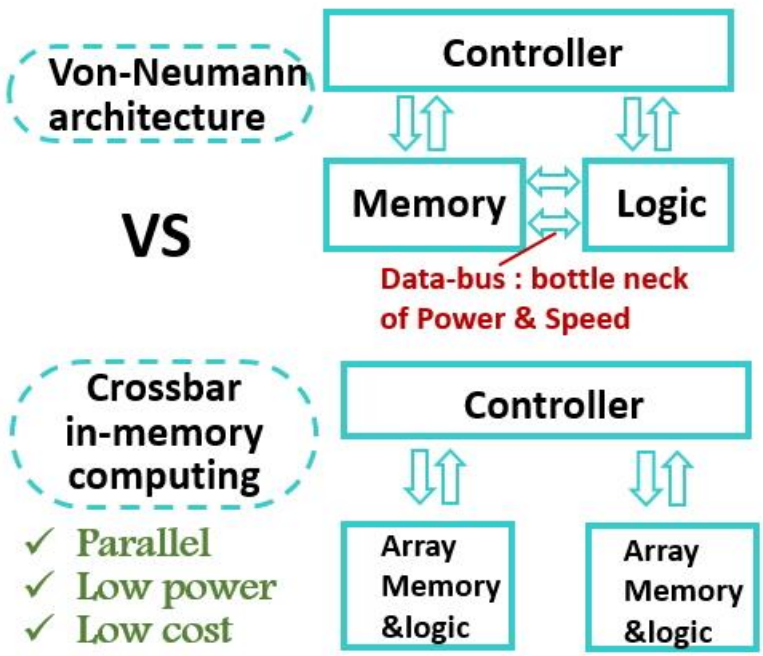


Figure 1.1. Schematics of the Von Neumann bottleneck due to data exchange between memory and arithmetic units, and a solution based on crossbar in-memory computing [2] .

Neuron circuits and synaptic devices are needed to implement neuromorphic computing [3]. Electronic synaptic devices can represent the weight values of neural networks with its multi-level conductance values and perform the massively parallel computations using these conductance values. Recently, various devices such as memristors (RRAM), conductive bridge memory (CBRAM), phase change memory (PCM), spin-based memory and FET-based memory have been reported as a synaptic device.

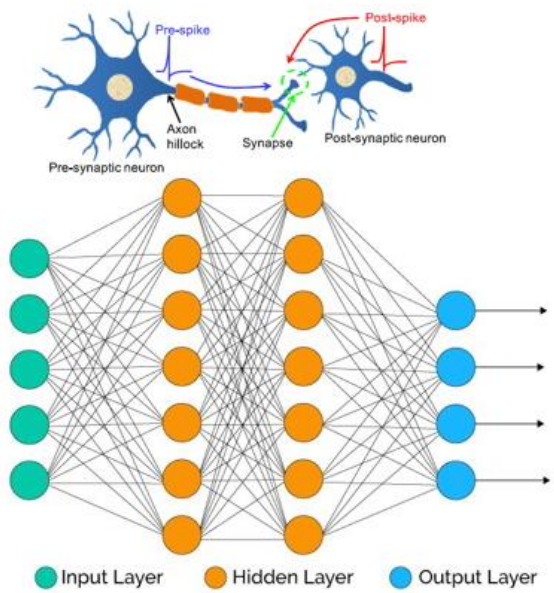


Figure 1.2. Relative timings of neuronal spikes from the pre-synaptic neuron and the post-synaptic neuron determine the weight change in synapse [3].

From a neuromorphic system point of view, RRAM is attractive in terms of simple structure, low power, CMOS compatibility and scalability for high density. Especially, when configured in cross bar structures, an RRAM based computing system can simply implement vector-matrix multiplication of the neural network. However, since the filament formation process is inherently abrupt and difficult to control, challenges exist such as reliability issues and device variation [4]. Also, the crossbar structure is susceptible to undesirable interference currents, so-called "sneak currents" flowing through adjacent cells.

In the case of NOR flash memory, A random access operation is possible. It has a matured process and good reliability characteristics, but it has poor scaling characteristics because it is an isolated transistor type memory (~$10F^2$)[5][6]. In addition, NAND flash memory has a matured fabrication, good reliability characteristics and excellent scalability with 3D stacks. However, it is difficult to implement the current summation required for neuromorphic computing because the cells are connected in series in a string (Figure 1.3).

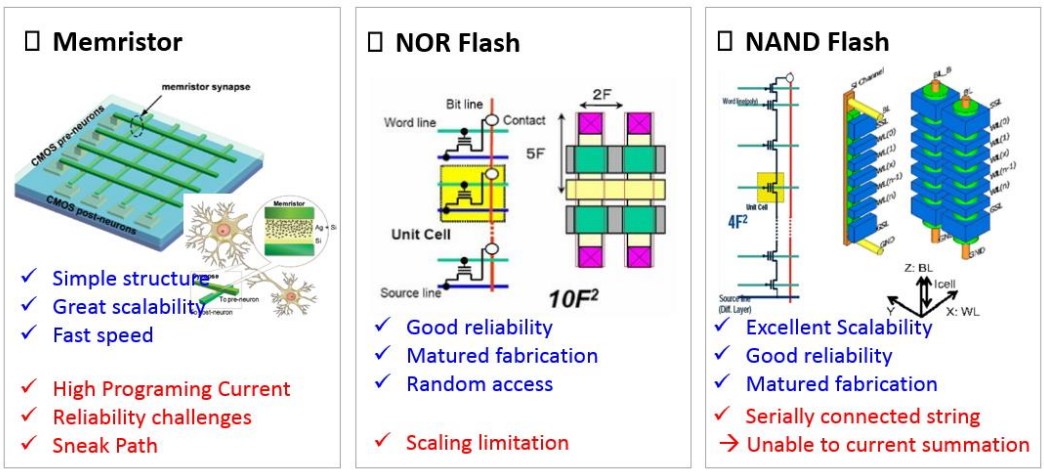


Figure 1.3. Comparison of pros and cons for various types of synaptic devices.

There are no devices that satisfy the requirements (high scalability, low power, high speed, high reliability, and non-volatility) of the Synaptic device[7][8][9][10].

1.2 Purpose of research

A neuromorphic computing is accomplished by the modification of connection strength of neurons and synapses [11]. Synapses have a memory function similar to that used in conventional computing. For a successful computation, a synaptic device should have following characteristics: high scalability, low power, high speed, high reliability, and non-volatility. Unfortunately, there are no standardized devices that satisfy all these conditions. In this dissertation, I propose a new synaptic device that can effectively perform neuromorphic computing in a 3-D stack structure while preserving the high-density features of the 3-D stack structure. In the architecture, memory behavior is verified by device simulations and characteristics of fabricated devices.

1.3 Thesis outline

This dissertation is composed as follows. Chapter 1 introduces the research background of the neural networks and the technology underlying the proposed device. The purpose of the research and the organization of thesis are also presented. In Charter 2, the new device structure and the operation methods are explained based on the simulation results. Chapter 3 deals with the device fabrication to implement the RDC flash memory device. In Chapter 4, the device characteristics including DC *I-V* are explained with failure analysis. Also, the new concept of process integration is explained.

Chapter 2

New synaptic device : Rounded Dual Channel (RDC) flash memory architecture

2.1 Architecture

In order to satisfy the requirements for the aforementioned synaptic device, A new 3-D synaptic device with stackable AND-type Rounded Dual Channel (RDC) flash memory architecture is proposed for neuromorphic computing. Figure 2.1 shows a 3-D schematic of stacked AND-type cells for the proposed synaptic devices. The Oxide/Nitride/Oxide (O/N/O) gate insulator stack is formed around the vertically standing plug-shaped gate electrode. A layer of polysilicon, the channel material of each cell stacked vertically, is formed around the gate insulator stack. In this architecture, bit lines (BL) and source lines (SL) are configured horizontally on both sides of the polysilicon in a particular layer. The charge stored in the nitride film of the O/N/O stack controls the current between the bit line and the source line.

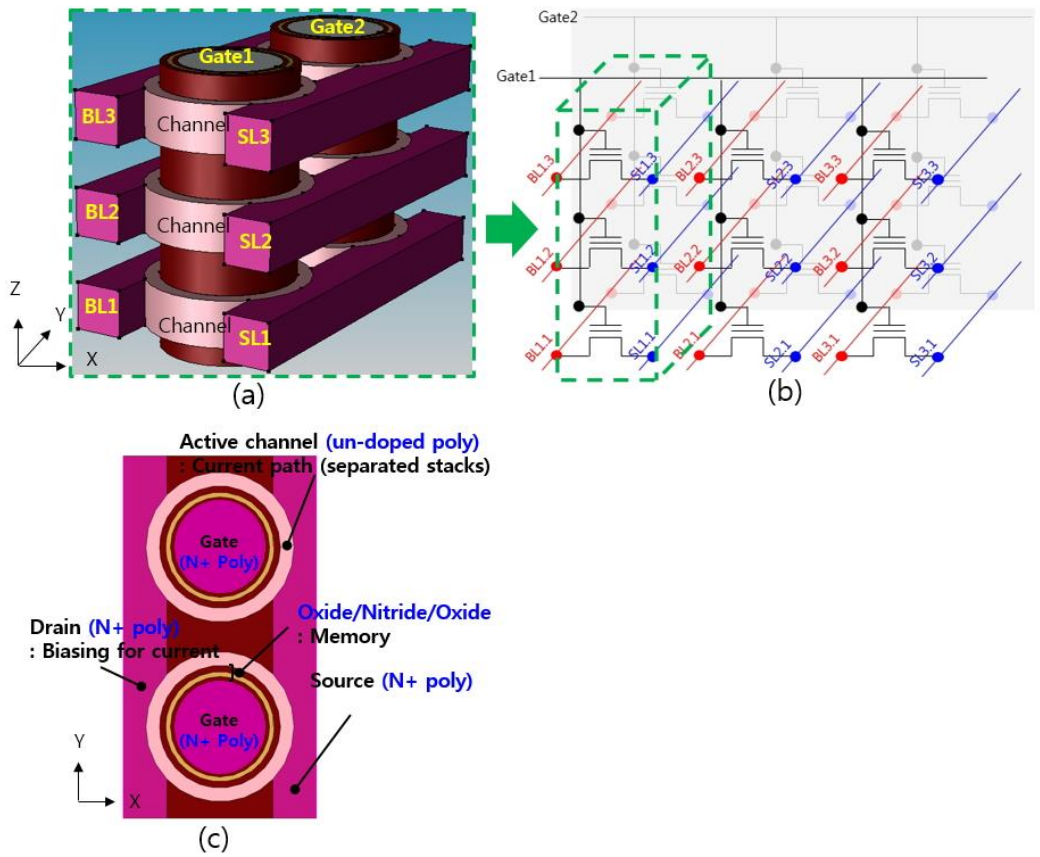


Figure 2.1. (a) Schematic 3D Rounded Dual Channel (RDC) flash device structure with three layers. (b) Equivalent circuit. (c) Top view schematics.

The proposed device has the following advantages; 1) High density can be realized by stacking. Figure 2.2 shows the effective area of the proposed device as a parameter of the number of stack layers based on the area of the memristor device without the selector. When stacked up to 15 floors or higher, the proposed device has a higher density than a memristor device having an area of 4F2 (F= 20 nm).

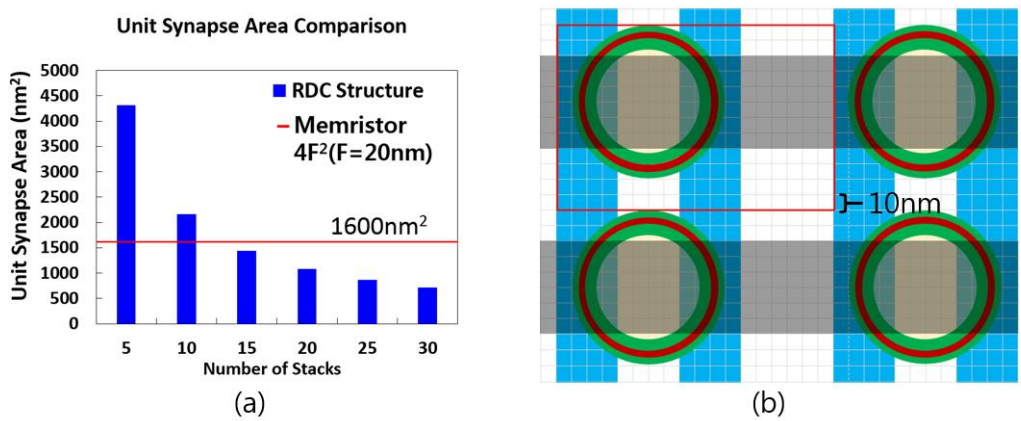


Figure 2.2. (a) Effective area of a 3D RDC flash device as a synapse. The reference memristor occupies 1600 nm2. No selector is used in a memristor. (b) An array of 3D RDC flash devices (rectangular box = one cell).

2) The proposed architecture allows for fast reads in sum-of-product operations for neuromorphic computing. For conventional memory operation, V_{read} is applied to the selected word line, and the unselected word line is turned off. Applying V_{read} to all word lines enables a sum-of-product operation for neuromorphic computing. All synaptic devices are connected in parallel, enabling read operation to be performed at once (Figure 2.3).

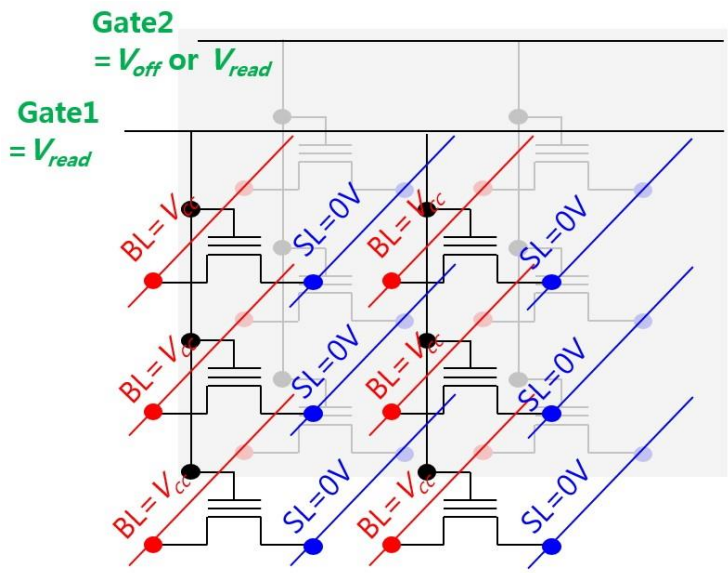


Figure 2.3. Schematic circuit diagram of 3D stack RDC flash devices. There are two methods of RDC flash devices: conventional memory reading and the neuromorphic operation.

3) Unlike those of NOR flash structure, source lines and bit lines of the proposed architecture are configured in parallel to allow FN program operation. Because of this arrangement, the structure is classified as AND-type instead of NOR-type [12] (Figure 2.4). This structure reduces power consumption as the program current is reduced relative to that in hot carrier program operation of NOR flash.

4) Since the proposed architecture is similar to that of V-NAND, most of the V-NAND processes currently in production could be utilized. (Chapter 3 is described in detail.)

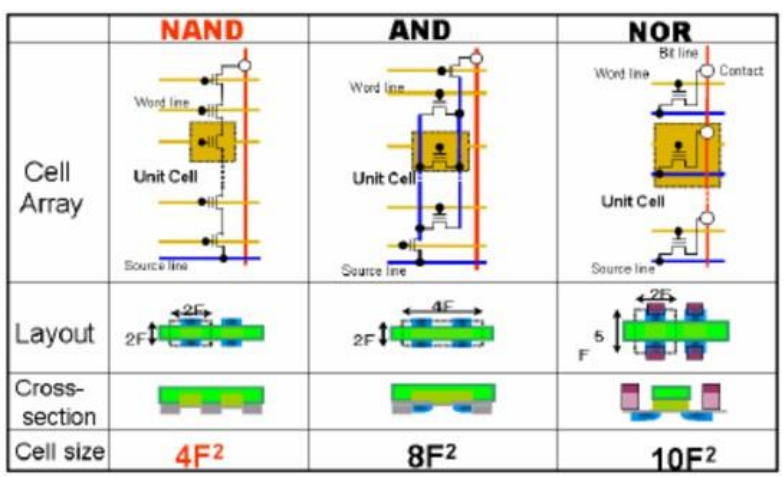


Figure 2.4. Various types of flash memory structure [13]. The AND structure differs from the NOR structure in that the source lines and the bit lines are parallel.

5) Due to the parallel connection of the plug-shaped gate electrodes, RC delay can be reduced by decreasing the length of the word line.

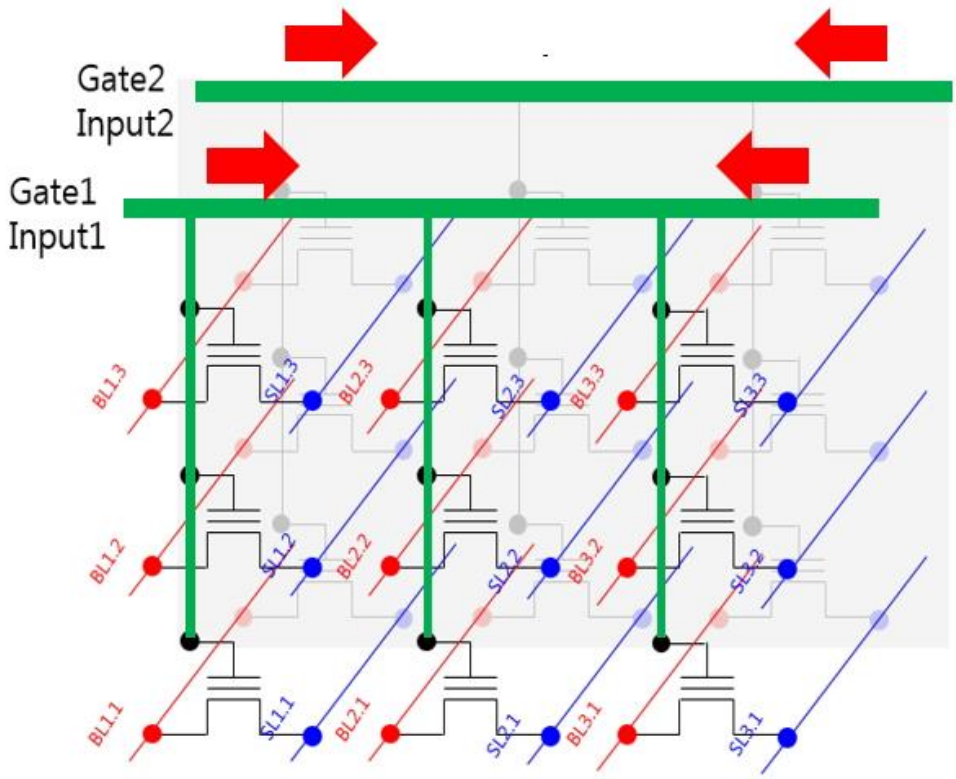


Figure 2.5. Schematic circuit diagram of 3D stack RDC flash devices. Due to the parallel connection of the plug-shaped gate electrodes, RC delay can be reduced by decreasing the length of the word line.

The proposed device can be of two types, considering the direction of charge

flow in the program (or erase): Inside Tunnel Oxide (ITOX) and Outside Tunnel Oxide (OTOX) types (Figure 2.6).

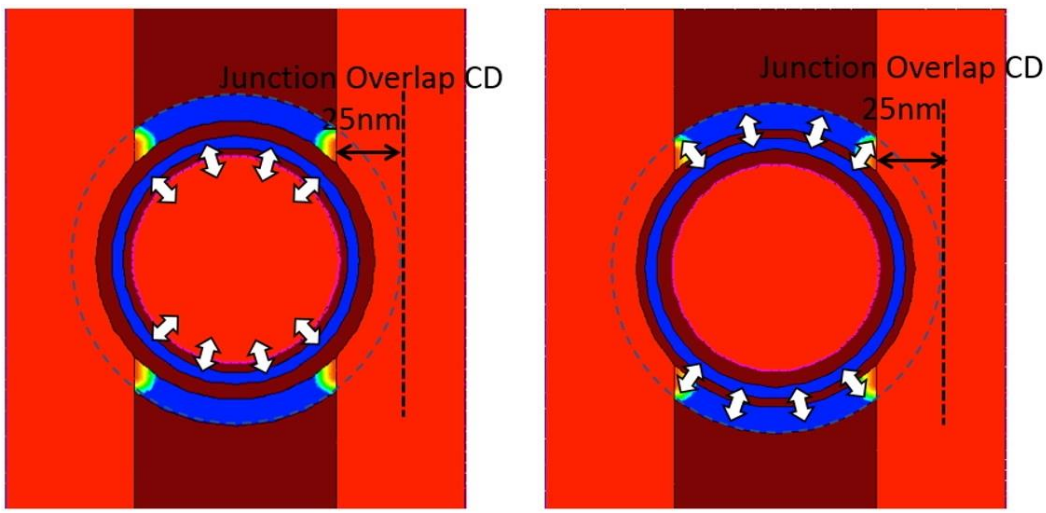


Figure 2.6. Two types of structures. (a) ITOX. (b) OTOX. Arrows indicate movement of electrons or holes.

2.2 Operation methods

The programming methods for the RDC flash device is shown in Figure 2.7. An example of the figure is the program methods for the OTOX structure. The structure of the NAND flash performs a program on a page-by-page basis. Similarly, the RDC flash device also performs programming in one word line unit, so that one input node becomes one program unit. Figure 2.7 (a) is an example of the bias condition of the program unit. 0V is applied to the bit lines and source lines of the cells to be programmed, and V_{PASS} is applied to the remaining bit lines and source lines. At this time, V_{PGM} is applied to the gate. In this case, the programed cell stores electrons in the trap nitride due to the potential difference between V_{PGM} and 0V, and the inhibited cells are not programmed because the potential difference between V_{PGM} and V_{PASS} is not sufficient to trap electrons (Figure 2.7 (inhibit cell 1)). Also, 0V is applied to other word lines, 0V to the bit lines and the source lines, and V_{PASS} to the other bit lines and source lines. These are not sufficient potential difference and those are not programmed (Figure 2.7 (inhibit cell 2)).

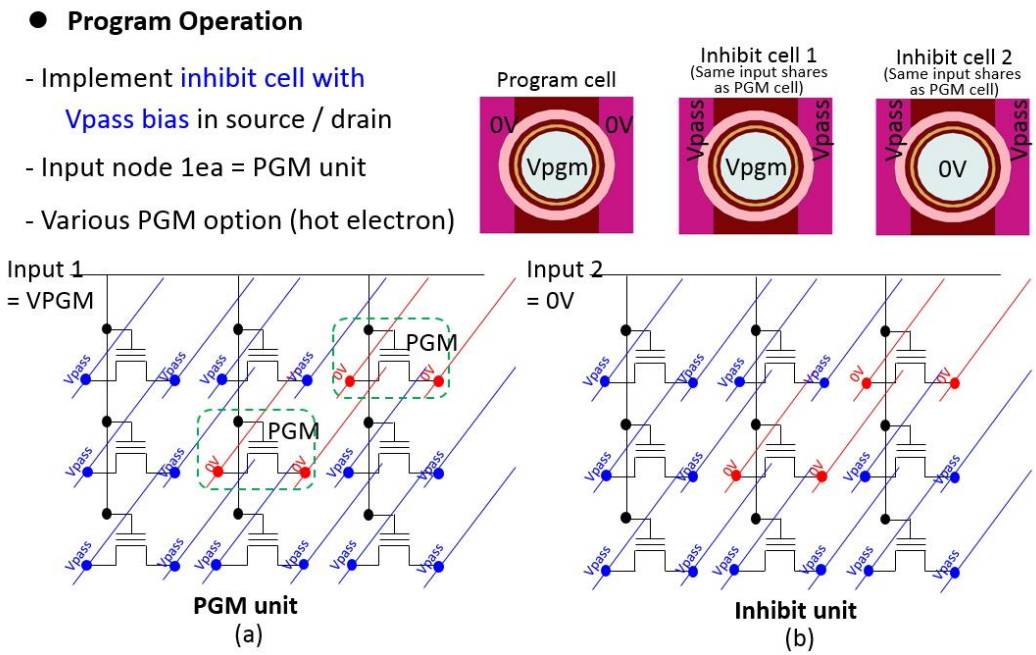


Figure 2.7. Program methods of OTOX structure. FN programming method was performed. (a) programed unit. (b) inhibited unit example. V_{PASS} is 1/2 V_{PGM}.

The Erase methods can be achieved by exchanging the nodes of the program methods with each other. That is, 0V is applied to the gate in the erase unit, and V_{ERS} is applied to the bit lines and the source lines. V_{PASS} is applied to inhibited cells like program methods. When the V_{ERS} bias is applied to the gate in the inhibit unit as in the programming methods, only the selected cell is erased. For the ITOX type, the programming methods are to exchange the terminals to which V_{PGM} and 0V are applied in Figure 2.7 respectively. In the case of erasing, the terminals which are VERS and 0V in Figure 2.8 are exchanged with each other.

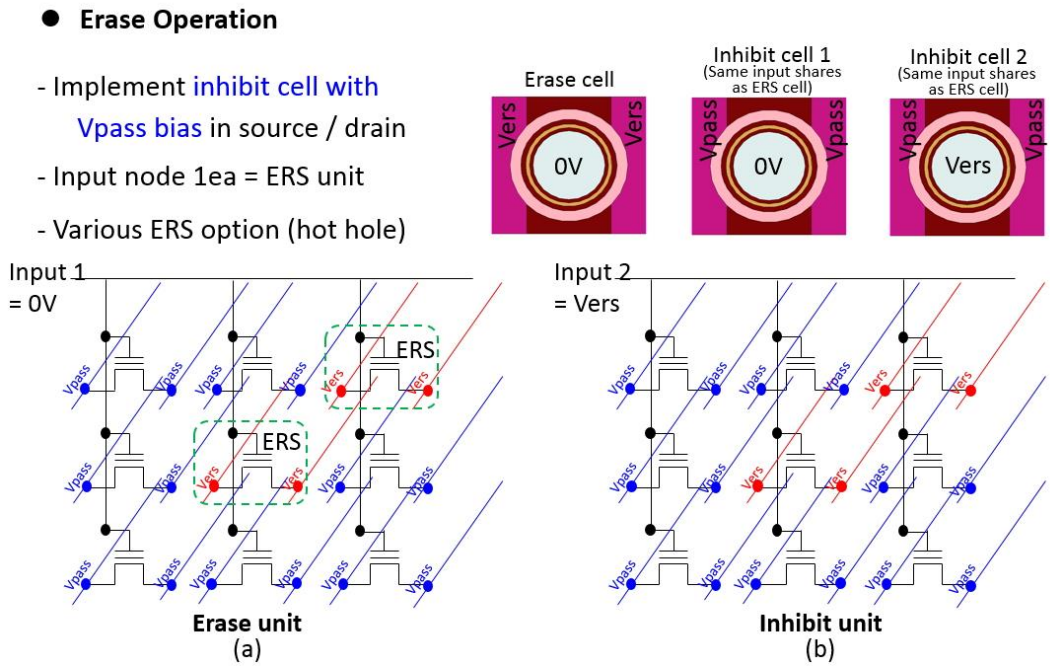


Figure 2.8. Erase methods of OTOX structure. FN erasing method was performed.

(a) erased unit. (b) inhibited unit example. V_{PASS} is 1/2 V_{ERS}.

2.3 Device simulation

To evaluate the program and erase characteristics of the synaptic device, TCAD simulation was performed. Figure 2.9 shows the voltage drop across the O/N/O insulator stack when program bias is applied to both ITOX and OTOX types. In the ITOX type, since the tunnel oxide is structurally located in the inner circle of the blocking oxide, the slope of the voltage profile across the tunnel oxide becomes steeper due to the electric field concentration, which works in favor of programming [14]. On the other hand, the threshold voltage is determined by the number of electrons trapped near the channel between the bit line and the source line. As a result, holes generated by GIDL are induced in the poly-Si channel. The failure of the channel potential to reach the applied program voltage indicates that the program time is not sufficient to transfer the potential to the channel with GIDL. In both types, the program efficiency with the diameter of the gate electrode (d_{gate}) changes as shown in Figure 2.10. Since the field effect becomes larger as the curvature increases, the smaller the d_{gate}, the better the program efficiency of the ITOX type.

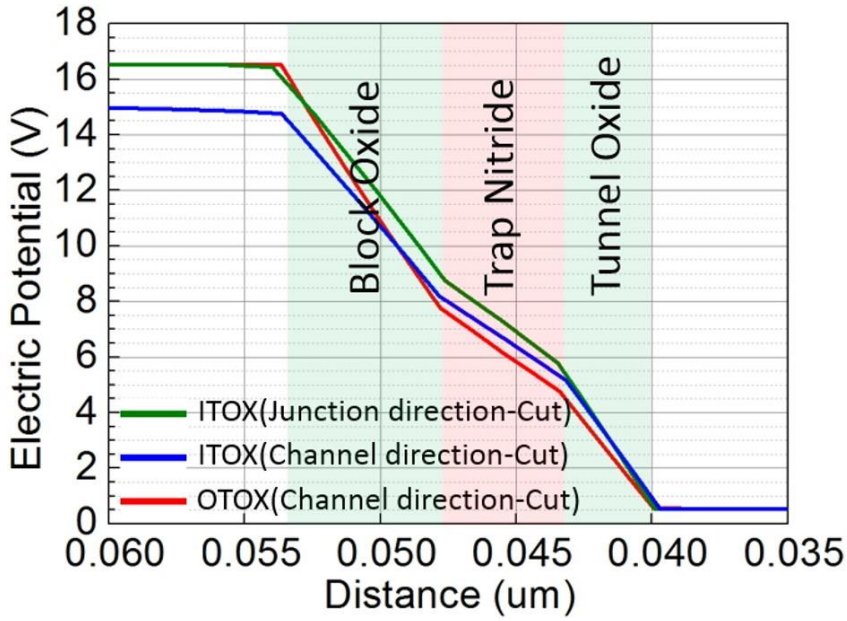


Figure 2.9. Potential profile in O / N / O insulator stack when programing (d_{gate}= 80 nm).

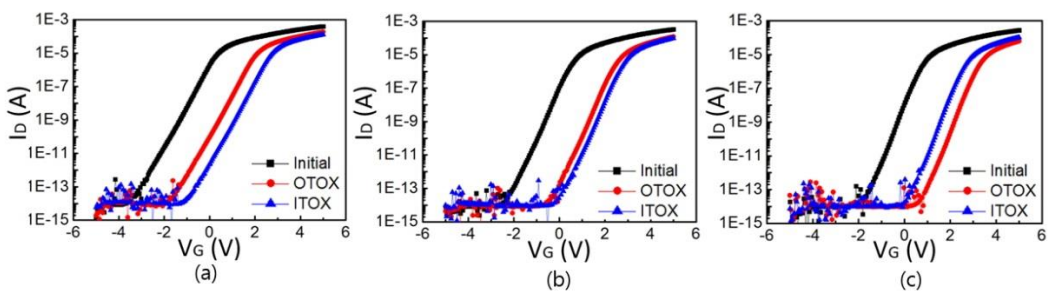


Figure 2.10. Simulated ID-VG characteristics of before and after program(+16V) with varying gate electrode diameter (d_{gate}): (a) d_{gate} = 60 nm, (b) d_{gate} = 80 nm, (c) d_{gate} = 100 nm.

The erase characteristics are shown in Figure 2.11. Even though the characteristics seem similar to program characteristics, the threshold voltage does not shift below the negative value, as no holes are supplied from the n+ doped poly gate in the ITOX type. The key characteristics of two types are listed in Table 2.1.

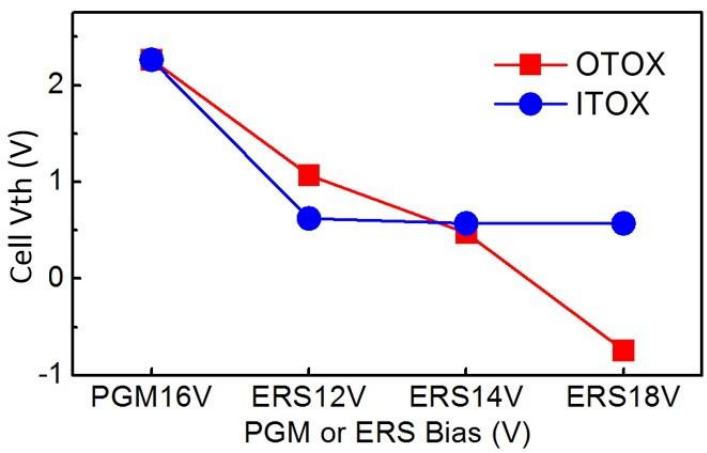


Figure 2.11. Cell Vth shift with increasing the erase bias

Table 2.1. Key characteristics of ITOX and OTOX structures.

Operation	Program		Erase	
Structure	Inside TOX	Outside TOX	Inside TOX	Outside TOX
Biasing Method	Gate : 0V BL/SL : Vpgm	Gate : Vpgm BL/SL : 0V	Gate : Vers BL/SL : 0V	Gate : 0V BL/SL : Vers
Electric field	Advantage	Disadvantage	Advantage	Disadvantage
Bias transfer to channel	Holes induced by GIDL	Electrons induced by Ch. inversion	Electrons induced by Ch. inversion	Holes induced by GIDL
Vth shift source	Electrons from N+ poly	Electrons induced by Ch. inversion	Electron de-trapping	Electron de-trapping + hole trapping
Remarks	**ITOX structure is advantageous under gate diameter 80nm**		**OTOX structure is suitable when negative cell Vth is required**	

The result of the device simulation in Figure 2.12 confirms that the program and erase of the selected device in an array is successfully performed, while the neighboring device is not affected during programming or erasing. Inhibition of the unselected cell is achieved by applying 8V to the BL and SL of this cell. While the V_{ISPP} is applied to the plug-shaped gate electrode for the program, 0V is applied to BL and SL of the selected device, and 8V is applied to the BL and SL of the neighboring devices for inhibition.

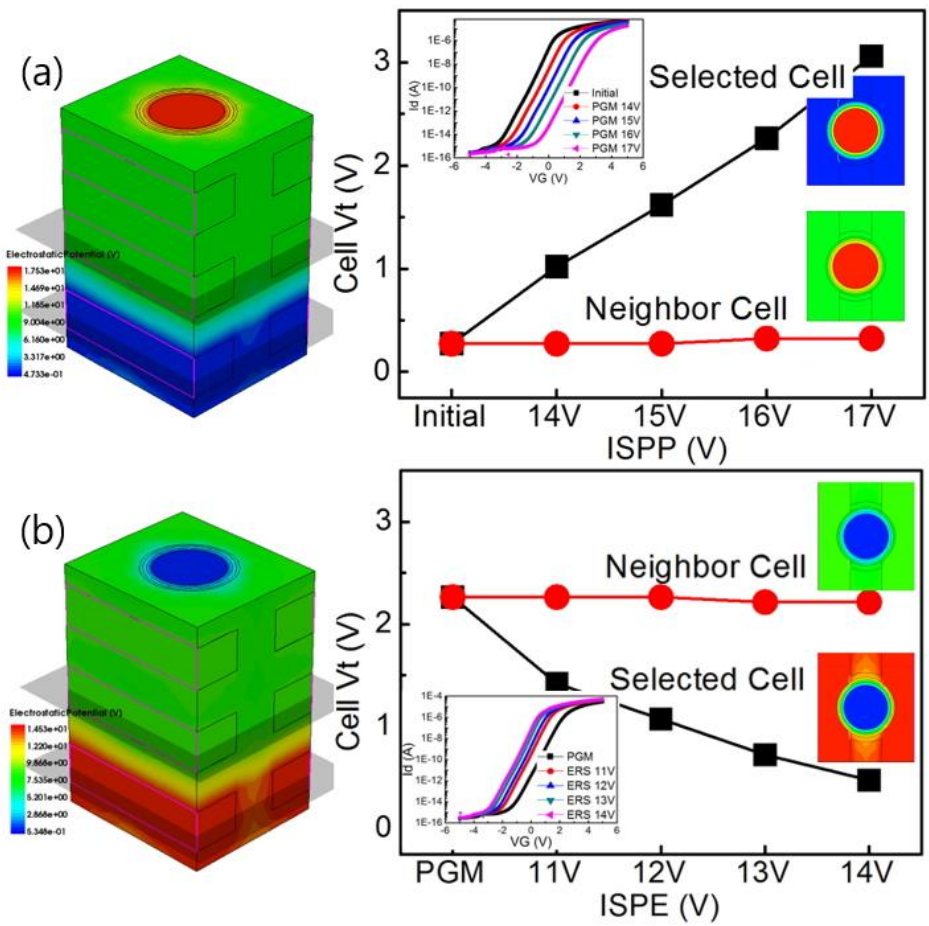


Figure 2.12. Simulated Vth shift behavior of selected and neighbor cells in an array when ISPP and ISPE pulses are applied for (a) program and (b) erase, respectively. In the structure of OTOX type, the cell with a gate diameter of 80 nm was simulated.

Chapter 3

Device fabrication

3.1 Overall process flow

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

Most of the process flow of the proposed RDC synapse device is similar to TCAT process [1] of 3D NAND flash. Figure 3.1 shows a simplified process flow. First, the multiple oxide/nitride layers are deposited on Si substrate. Next, via holes are patterned and nitride layers are partially etched (a)~(b). Poly-Si for the channel is formed in the nitride etched grooves (c)~(d). Tunnel oxide, trap nitride, blocking oxide are deposited sequentially, and n+ poly-Si is deposited for the gate (e). Nitride layers are exposed after stack patterning for stack isolation (g)~(h). Then the nitride

layers are etched. n+ poly-Si is deposited, and this is formed only in the regions where nitride layers are etched such that BLs and SLs are isolated (i)~(k). After forming the interlayer insulating film (l), contact holes are formed, and metal wiring is performed. Most V-NAND flash processes currently in production can be utilized.

In order to fabricate the synaptic devices, the layout as shown in figure 3.2 was designed. For the fabrication of the 3-layer device, a total of 9 masks were used and a design rule was for consideration of the ISRC equipment situation (figure 3.3) The optimal align sequence was derived by considering the process margin between each layer. In particular, since phosphoric acid wet etch is used to form the source / drain, the most critical part of the design rule is the overlay between the GATE and the S/D mask, allowing it to be a primary alignment. In areas where no cells exist, CMP dummy is inserted as shown in Figure 3.4. The bar & space ratio of 1.1: 1 was used to prevent the dishing phenomenon that occurred during the CMP process. A gate hole pattern was also drawn on the CMP dummy to indirectly monitor the cell part during SEM analysis. Figure 3.5 shows total process flow, equipment, and detailed recipe information. A total of 90 steps were designed and five n^+ doped

poly-Si processes were performed in the NNFC.

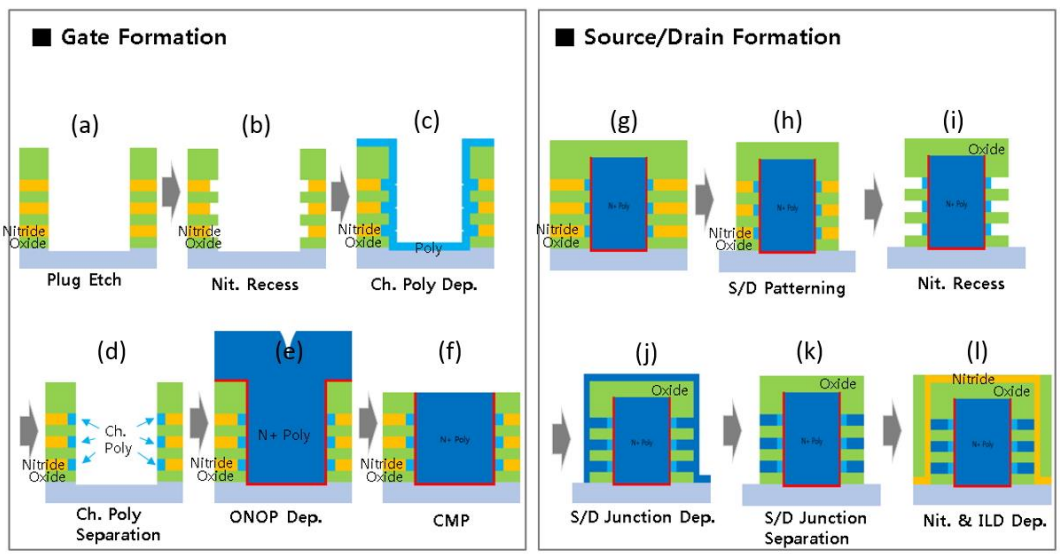
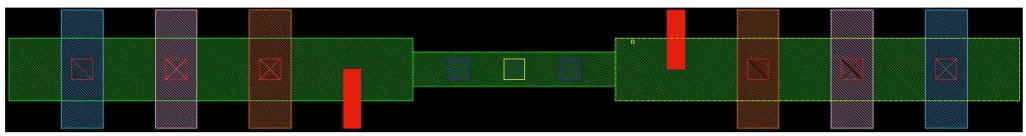
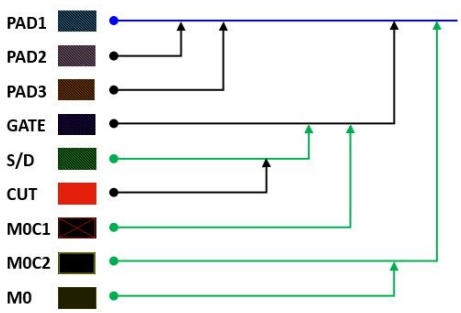


Figure 3.1. Key process steps for the fabrication of proposed architecture. (a) Via patterning and nitride groove formation (b). (c) Undoped poly-Si deposition. (c)~(d) Poly-Si in the grooves. (e) ONO and n+ poly deposition. (f) CMP process. (g) Cap oxide deposition (h) Stack patterning. (i) Nitride recess for BL and SL. (j) n+ poly deposition. (k) n+ poly separation. (l) interlayer insulating film deposition and CMP.

○ Base Module



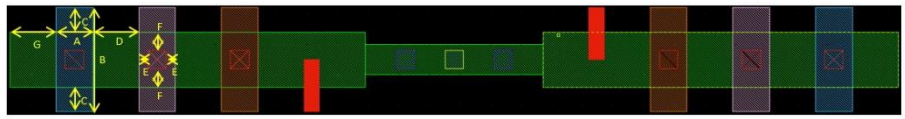
○ Align Sequence



Mask	Purpose	OL & CD Vari.	Sab
PAD1	Contact Pad	±100nm	PAD1 to S/D : 150nm (second)
PAD2	Contact Pad	±100nm	PAD2 to S/D : 187nm (third)
PAD3	Contact Pad	±100nm	PAD3 to S/D : 187nm (third)
GATE	Cell Plug Channel	±100nm	-
S/D	S/D Formation	±100nm	GATE to S/D : 112nm (first)
CUT	S/D Cut	±100nm	CUT to S/D : 112nm (first)
M0C1	Cell Plug Contact	±100nm	M0C1 to CPL : 112nm (first)
M0C2	S/D Contact	±100nm	M0C2 to PAD1 : 112nm (first) M0C2 to PAD2 : 158nm (second) M0C2 to PAD3 : 158nm (second)
M0	Metal Line	±100nm	M0 to M0C1 : 212nm (forth) M0 to M0C2 : 112nm (first)

Figure 3.2. Base module design and align sequence.

○ **Base Module**

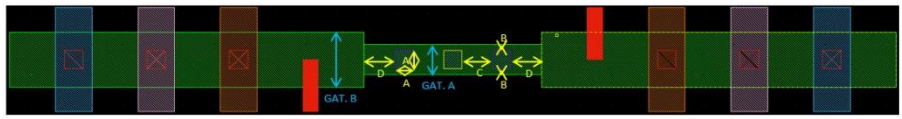


○ **PAD1~3**

Code	Mask	ITEM	Size	Remark
PAD. A	PAD1~3	PAD Width	1.2um	1.2 : M0C Enclosure (0.3x2) + M0C2 CD 0.6
PAD. B	PAD1~3	PAD Length	2.8um	2.8 : PAD Extension (0.5x2) + S/D CD 1.8
PAD. C	PAD1~3	PAD Extension to GAT	0.5um	>0.49 : PAD to S/D Sab 0.19 + PAD Patterning Margin 0.3
PAD. D	PAD1~3	PAD to PAD Space	1.5um	>0.35 : Conditions without ON stacks bending by Pad or >0.5 : Metal Line min Pitch 0.5 x 0.5
PAD. E	PAD1~3	M0C2 Enclosure of PAD (X)	0.3um	>0.16 : M0C2 to PAD Sab 0.16
PAD. F	PAD1~3	M0C2 Enclosure of PAD (Y)	0.6um	>0.16 : M0C2 to PAD Sab 0.16 or >0.4 : Cut Margin 0.4
PAD. G	PAD1	GAT Extension to PAD	1.5um	>0.65 : Conditions without ON stacks bending by Pad 0.35 + S/D Patterning Margin 0.3

(a)

○ **Base Module**

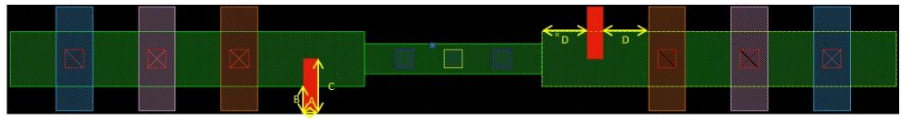


○ **CPL / GAT**

Code	Mask	ITEM	Size	Remark
GATE. A	GATE	Cell Plug Diameter	0.5um	Consideration of minimizing ON Recess at S/D (0.9)
GATE. B	GATE	CPL Enclosure of S/D	0.2um	>0.11 : GATE to S/D SAB 0.11
GATE. C	GATE	CPL to CPL Space	1.0um	>1.0 : Patterning Margin 1.0
GATE. D	GATE	Dummy GATE Space	1.0um	-
S/D. A	S/D	S/D Width in Cell	0.9um	GATE Enclosure (0.2x2) + GATE CD 0.5
S/D. B	S/D	S/D Width in PAD	1.8um	M0C2 Enclosure (0.6x2) + M0C2 CD 0.6

(b)

○ **Base Module**



○ **CUT / M0C1 / M0C2**

Code	Mask	ITEM	Size	Remark
CUT. A	CUT	CUT Width	0.5um	Pattering Margin (Larger Gap-fill burden: upper flexion)
CUT. B	CUT	CUT Extension to GAT	0.5um	>0.4 : CUT to S/D Sab 0.11 + PAD Patterning Margin 0.3
CUT. C	CUT	CUT Length	1.4um	>1.0 : CUT Extension 0.5 + CUT Margin 0.5
CUT. D	CUT	CUT Space	1.5um	>1.0 : After CUT Gap-fill, upper flexion free condition 1.0
M0C1	M0C1	M0C1 CD	0.5um	GATE CD 0.5um
M0C2	M0C2	M0C2 CD	0.6um	-

(c)

Figure 3.3. Design rule of PAD (a), GATE, S/D (b), CUT, M0C1, M0C2 (c).

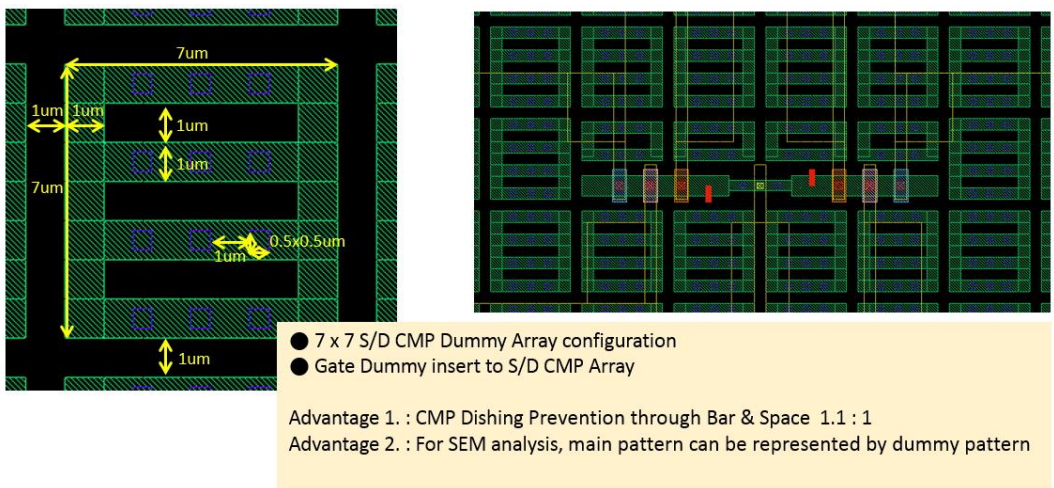


Figure 3.4. CMP dummy rule.

Module	No	PROCESS STEP	Equipment	Target	RCP
Pad	1	ON OXIDE1 DEP PRE CLN	Wet-Station		R-N1-N2-HF
	2	ON OXIDE1 DEP	FurnaceII	3000Å	Wet Oxidation #202 50min
	3	PAD POLY1 DEP	-	600Å	2E20 Doped Poly
	4	PAD1 MASK	Nikon Stepper	1.5um (Main Pattern)	300mS, -2.5
	5	PAD POLY1 ETCH	ICP Poly Etcher I	THK : Oxide Stop / CD : 1.5±0.2um	SMDL_JH 23"
	6	PAD POLY1 ETCH PR STRIP	PR Asher		O2 Ashing
	7	PAD POLY1 ETCH CLN	Wet-Station		R-R-N1-N2
	8	ON NIT1 DEP	LPCVD III	700Å	#333 26' 30"
	9	ON OXIDE2 DEP	-	800Å	LPTEOS
	10	PAD POLY2 DEP	-	600Å	2E20 Doped Poly
	11	PAD2 MASK	Nikon Stepper	1.5um (Main Pattern)	300mS, -2.5
	12	PAD POLY2 ETCH	ICP Poly Etcher I	THK : Oxide Stop / CD : 1.5±0.2um	SMDL_JH 23"
	13	PAD POLY2 ETCH PR STRIP	PR Asher		O2 Ashing
	14	PAD POLY2 ETCH CLN	Wet-Station		R-R-N1-N2
	15	ON NIT2 DEP	LPCVD III	700Å	#333 26' 30"
	16	ON OXIDE3 DEP	-	800Å	LPTEOS
	17	PAD POLY3 DEP	-	600Å	2E20 Doped Poly
	18	PAD3 MASK	Nikon Stepper	1.5um (Main Pattern)	300mS, -2.5
	19	PAD POLY3 ETCH	ICP Poly Etcher I	THK : Oxide Stop / CD : 1.5±0.2um	SMDL_JH 23"
	20	PAD POLY3 ETCH PR STRIP	PR Asher		O2 Ashing
	21	PAD POLY3 ETCH CLN	Wet-Station		R-R-N1-N2
	22	ON NIT2 DEP	LPCVD III	700Å	#333 26' 30"
	23	ON OXIDE4 DEP	-	3000Å	LPTEOS
Gate	24	GATE MASK	Nikon Stepper	600nm Hole : ~650nm (inner circle)	950ms / -2.5
	25	GATE ETCH	P-5000IV Etcher	THK : Si-Sub Stop	SiO2 270"
	26	GATE ETCH PR STRIP	PR Asher		O2 Ashing
	27	GATE ETCH CLN	Wet-Station		R-R-N1-N2
	28	GATE CH NIT RECESS	Wet-Station	moni. : 90nm (Pattern : form Oxide 70nm)	H3PO4 1400"
	29	GATE CH POLY DEP	LPCVD III	620Å	Un-Doped #200, 29'
	30	GATE INNER POLY ANNEAL	FurnaceII		P+ Anneal 600℃ 20H' N2
	31	GATE CH POLY SEP ETCH PRE CLN	Wet-Station	Native Oxide Remove	dHF 70"
	32	GATE CH POLY SEP ETCH	ICP Poly Etcher I	Side All Remove	SMDL_SF6 35" / 20sccm
	33	GATE CH POLY SEP CLN	Wet-Station		R-N1 (1:1:5 65℃)
	34	GATE CH TOX DEP (Outer Poly Scheme)	LPCVD I (MTO)	35Å	MTO_SMDL 3'
	35	GATE CH NIT DEP (Outer Poly Scheme)	LPCVD III	50Å	#391 3' 15" (195")
	36	GATE CH BOX DEP (Outer Poly Scheme)	LPCVD I (MTO)	110Å	MTO_SMDL 21'
	37	GATE INNER POLY DEP	-	6000Å	2E20 Doped Poly
	38	GATE INNER POLY CMP	CMP New	GATE upper 2000±500Å	70"
	39	GATE INNER POLY CMP CLN	Wet-Station		R-N1
	40	GATE INNER POLY ETCHBACK	ICP Poly Etcher I	GATE Stop CMP	70~100"
	41	GATE INNER POLY CMP CLN	Wet-Station		R-N1
	42	GATE CAPPING OXIDE DEP 1	LPCVD I (MTO)	430Å	MTO_SMDL 1h+20m
	43	GATE CAPPING OXIDE DEP 2	LPCVD I (MTO)	430Å	MTO_SMDL 1h+20m

Module	No	PROCESS STEP	Equipment	Target	RCP
Source / Drain	44	S/D MASK	Nikon Stepper	1um	300ms / -2.5
	45	S/D ETCH 1	P-5000IV Etcher		SiO2 270"
	46	S/D ETCH 2	ICP Poly Etcher I	Pad Etch (△600Å+20%)	STI6IN 21"
	47	S/D ETCH 3	P-5000IV Etcher		SiO2 130"
	48	S/D ETCH PR STRIP	PR Asher		O2 Ashing
	49	S/D ETCH CLN	Wet-Station		R-R-N1-N2
	50	S/D JUNCTION RECESS	Wet-Station	300nm (Pattern : Oxide inner 250nm)	H3PO4 5000"
	51	S/D JUNCTION POLY DEP PRE CLN	Wet-Station	Native Oxide Remove	dHF 20"
	52	S/D JUNCTION POLY DEP 1	LPCVD III	100Å	Un-Doped #200, 5' 30"
	53	S/D JUNCTION POLY DEP 2	-	600Å	2E20 Doped Poly
	54	S/D JUNCTION ANNEAL 1	FurnaceII		P+ Anneal 600℃ 12H' N2
	55	GATE CH POLY SEP ETCH PRE CLN	Wet-Station	Native Oxide Remove	dHF 70"
	56	S/D JUNCTION POLY SEP ETCH	ICP Poly Etcher I	Side All Remove	SMDL_SF6 35" / 20sccm
	57	S/D JUNCTION POLY SEP ETCH CLN	Wet-Station		R-N1 (1:1:5 65℃)
	58	S/D JUNCTION ANNEAL 2	RTA		950℃ 10" RTA
	59	S/D SEALING NIT DEP	LPCVD III	600Å	#333 19'30"
	60	S/D SEALING OX DEP 1	HDPCVD II	7000Å	1h 20m
	61	S/D SEALING OX DEP 2	HDPCVD II	7000Å	1h 20m
	62	S/D SEALING OX CMP	CMP [SiO2-STI]	Nitride Stop CMP	
	63	S/D SEALING OX CMP CLN	Wet-Station		R-N1-N2
	64	S/D CUT MASK	Nikon Stepper		300mS, -2.5
	65	S/D CUT ETCH 1	P-5000IV Etcher	△4800Å	SiO2 270"
	66	S/D CUT ETCH 2	ICP Poly Etcher I	△1750Å	STI6IN 50"
	67	S/D CUT ETCH 3	P-5000IV Etcher	△1000Å	SiO2 25"
	68	S/D CUT ETCH 4	ICP Poly Etcher I	△1750Å	STI6IN 50"
	69	S/D CUT ETCH 5	P-5000IV Etcher	△1000Å	SiO2 25"
	70	S/D CUT ETCH 6	ICP Poly Etcher I	△1750Å	STI6IN 50"
	71	S/D CUT ETCH PR STRIP	PR Asher		O2 Ashing
	72	S/D CUT ETCH CLN	Wet-Station		R-R-N1-N2
	73	S/D CUT OXIDE DEP 1	LPCVD I (MTO)	50Å	MTO_SMDL 6' 30"
	74	S/D CUT OXIDE DEP 2	P-5000IV (TEOS)	1000±200Å	TEOS_DI 30"
Contact / Metal	75	M0C1 MASK	Nikon Stepper		900ms / -2.5
	76	M0C1 ETCH	P-5000IV Etcher	△1750Å	SiO2 83"
	77	M0C1 ETCH PR STRIP	PR Asher		O2 Ashing
	78	M0C1 ETCH CLN	Wet-Station		R-R-N1-N2
	79	M0C2 MASK	Nikon Stepper		900ms / -2.5
	80	M0C2 ETCH	P-5000IV Etcher	THK : 1F Pad Touch / CD : 600±150nm	SiO2 270"
	81	M0C2 ETCH PR STRIP	PR Asher		O2 Ashing
	82	M0C2 ETCH CLN	Wet-Station		R-R-N1-N2
	83	M0 BM TI DEP	Endura Sputter	300Å	
	84	M0 BM TIN DEP	Endura Sputter	300Å	
	85	M0 AL DEP	Endura Sputter	3500Å	
	86	M0 TIN DEP	Endura Sputter	300Å	
	87	M0 MASK	Nikon Stepper		300ms / -2.5
	88	M0 ETCH	ICP Etcher(Metal)	CD : 1.5±0.5um / △4400Å	92"
	89	M0 ETCH PR STRIP	PR Asher		
	90	M0 ETCH CLN	Wet-Station(WS-10)		

Figure 3.5. Total process flow, equipment, and detailed recipe information.

3.2 Cell process set up

There have been many trials and errors in proceeding with the aforementioned process integration. Therefore, I will introduce the issues and solutions that occurred during the process. In the fabricating cell stacks, 11 detailed steps were divided, and summarized the lessons in each process.

First, multiple oxide / nitride depositions are performed (figure 3.6). At this time, the oxide is deposited at 680 ° C LP (low pressure) -TEOS and the nitride is deposited at 780 ° C LP-CVD. The oxide was deposited at 640Å and the nitride at 750Å, and the top oxide was deposited at 1700Å to secure the gate CMP margin for the subsequent process.

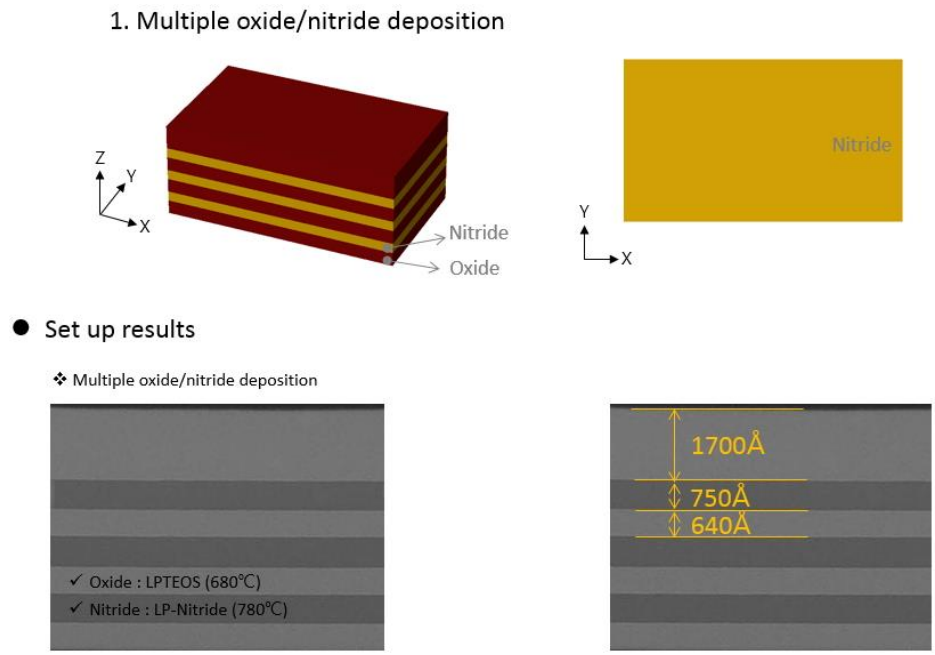


Figure 3.6. Cross-sectional image of multiple oxide/nitride deposition.

Second, via holes are patterned (figure 3.7). As a result of the energy-focus matrix split to secure the exposure conditions, the best conditions were obtained at energy 950ms, focus -2.5. And also, from the etching images, it was confirmed that etching loading occurred depending on the pattern size. The 700nm pattern has grown excessively to about 830nm for top CDs, while the 500nm pattern has obtained a drawing CD with about 500nm for top CDs. However, as a result of subsequent confirmation, the 500 nm hole had some patterning failures.

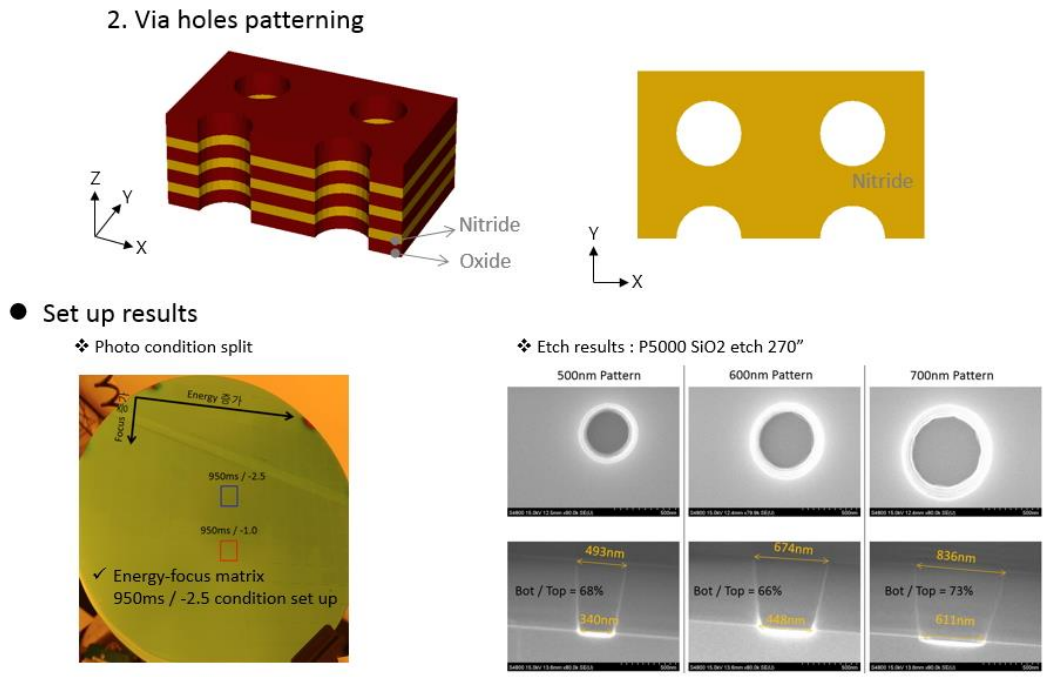


Figure 3.7. Cross-sectional image of via holes patterning.

Next, nitride layers are partially etched. This process step is to form a separated channel to be formed in a subsequent process. For this purpose, phosphoric acid (H_3PO_4) Wet etching was performed so that only the nitride was partially etching, using the selectivity of oxide and nitride. As a result of 720 seconds sample wet etching, it was found that the recess CD was about 30~40nm and the selectivity ratio between oxide and nitride was about 7: 1. To obtain sufficient channel separation margin, 1400 seconds wet etching was performed to main wafers.

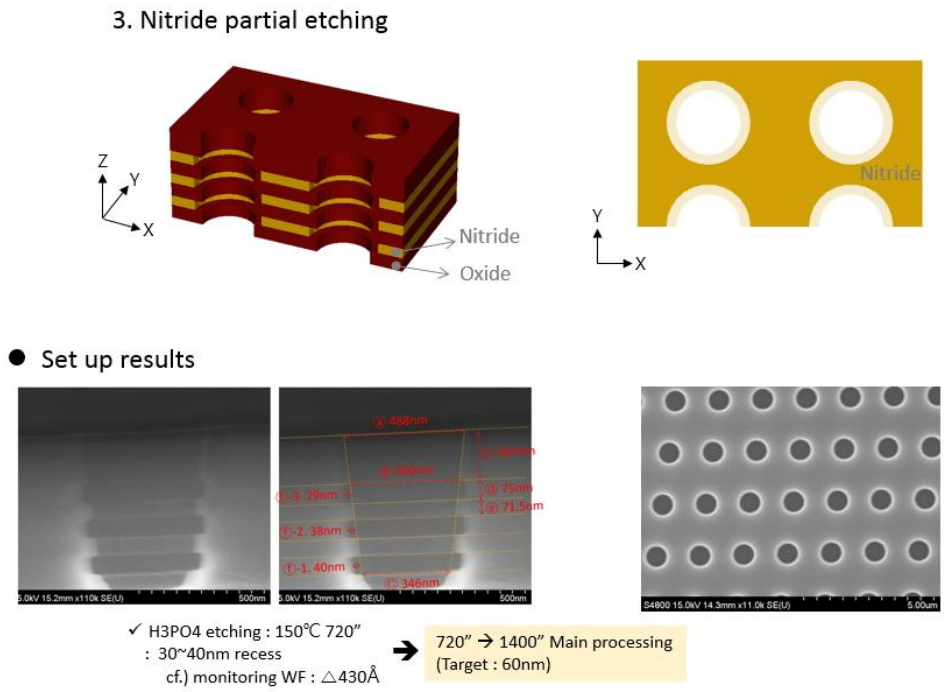


Figure 3.8. Cross-sectional image of partial nitride wet etching.

After partial nitride wet etching, poly-Si for the channel is formed in the nitride etched grooves (figure 3.9). If the channel is not properly separated, the channel between floors will be shared, resulting in short failure, which requires more attention. Firstly, 550° C un-doped channel poly-Si was deposited and annealed at 600° C for 12 hours for crystallization (figure 3.9 (a)).

There are various methods for interlayer separation, two methods were tested: Oxide etching through HF cleaning after oxidation, and isotropic etching using SF6 gas. As a result of the oxidation process, the oxidation rate of the lower layer was lower than that of the upper layer (Figure 3.9 (b)). This is because the stress is changed by the oxide film growth and the thickness of the oxide film is also changed. In other words, it can be explained that the oxide volume expands while it grows in the side and the bottom, and the stress concentrates on the corner. The results of isotropic etching using SF6 gas are shown in Figure 3.9 (c)~(e). Without proper cleaning prior to etching, etching will not be performed due to the native oxide present on the poly-Si. The results of an isotropic etching time change under the same conditions are shown in Figure 3.9 (d) and Figure 3.9 (e). Even though the

etching time was more than doubled, the etching amount of side region was not significantly increased, while bottom region was largely etched. It is shown that the side etching amount of SF6 gas is greatly reduced from the initial value. Based on the above experimental results, the isotropic etching was carried out for 30 seconds, and the final condition was to proceed with hot SC-1 cleaning in order to obtain additional process margin.

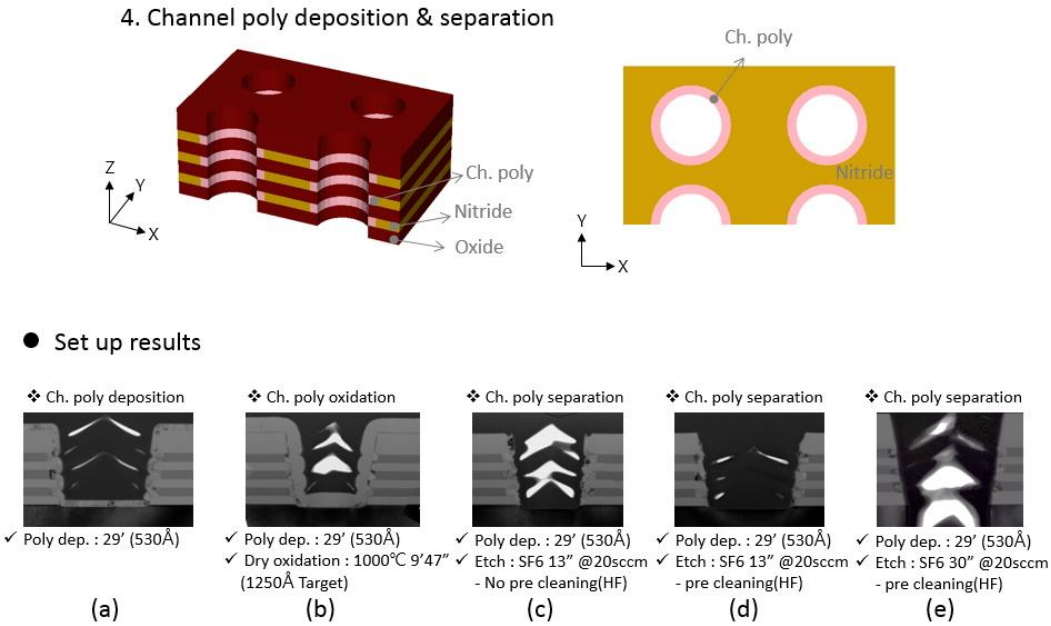


Figure 3.9. Cross-sectional image of channel poly deposition (a), oxidation (b), and separation (c)~(e).

After channel separation, Oxide / Nitride / Oxide (O / N / O) is deposited in order (Figure 3.10). O / N / O is tunnel oxide / trap nitride / blocking oxide, respectively, in the deposition order because the channel is outside the hole (For the OTOX scheme). Oxide and nitride were deposited by LPCVD at 780 ℃ and 750 ℃, respectively. Figure 3.9 (a) and (b) show the results of measuring the deposition thickness with time of oxide and nitride, respectively.

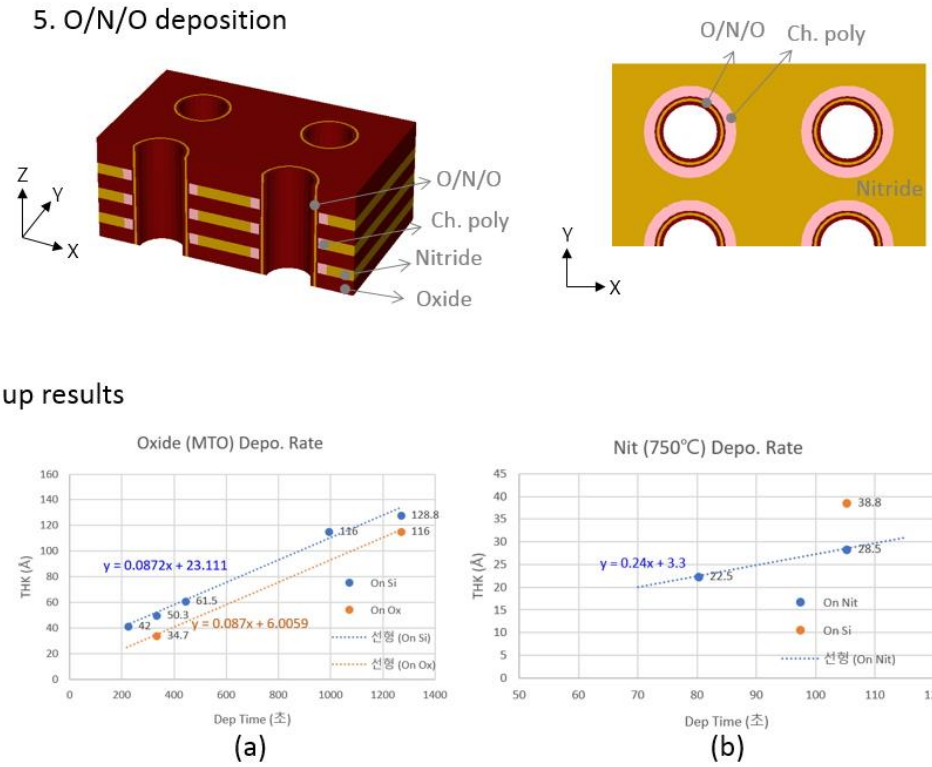


Figure 3.10. Cross-sectional image of O/N/O deposition. MTO deposition thickness (a) and nitride deposition thickness (b) over time.

Then, a gate poly-Si is deposited, and Chemical Mechanical Polishing (CMP) process is performed to remove the poly-Si and O / N / O deposited on the top region (Figure 3.11). The n^+ doped poly-Si was used as a gate poly, and 6000Å was deposited to fill the hole.

In order to minimize the thickness variation remaining after the CMP process in a wafer, a poly etch-back process was added after the CMP process (Figure 3.11 (a)). In this way, the variations in the wafer in the two processes can be offset. As the CMP process proceeds, the wafer color changes depending on the etching result, which is shown in Figure 3.11 (b). If the poly-Si remains on the oxide / nitride stacks, it looks red and if it is all removed, it is observed to be blue. During the process, adjusting the etch-back time while checking the wafer color can reduce process failures.

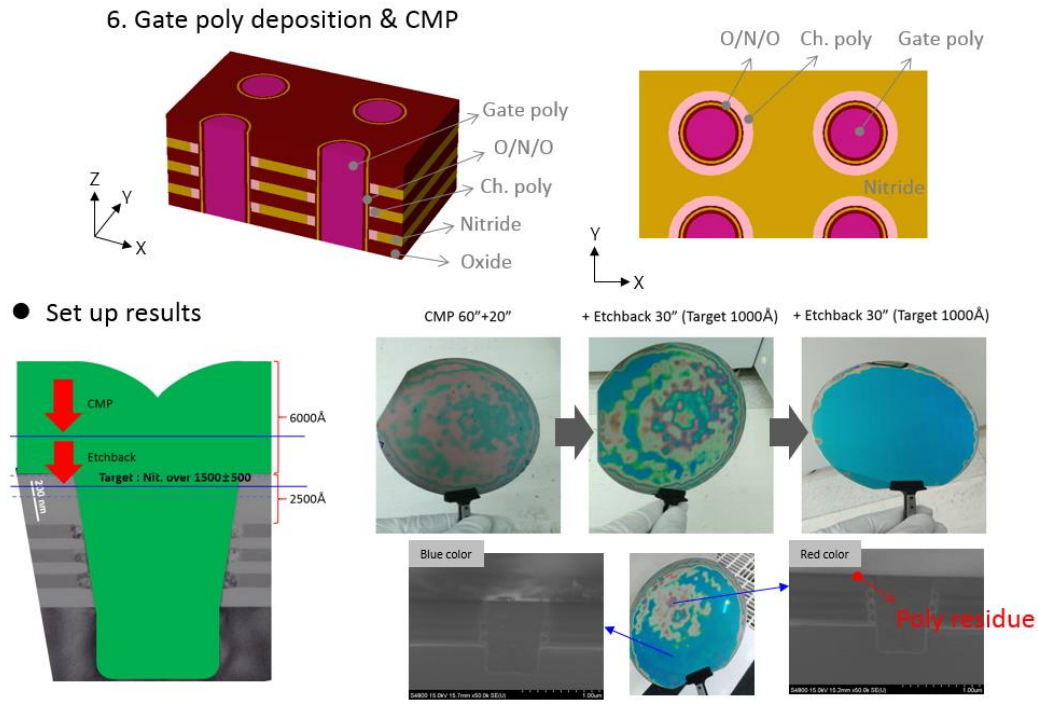


Figure 3.11. Cross-sectional image of CMP process. In order to minimize the thickness variation, a poly etch-back process was added (a). (b) shows the wafer color changes depending on the etching result.

The next step is to deposit capping oxide on top of the CMP process (Figure 3.12). The purpose of this step is to protect gate poly-Si and O / N / O films exposed on the top region during the subsequent nitride wet etching process. LPCVD oxide 800Å deposition was carried out.

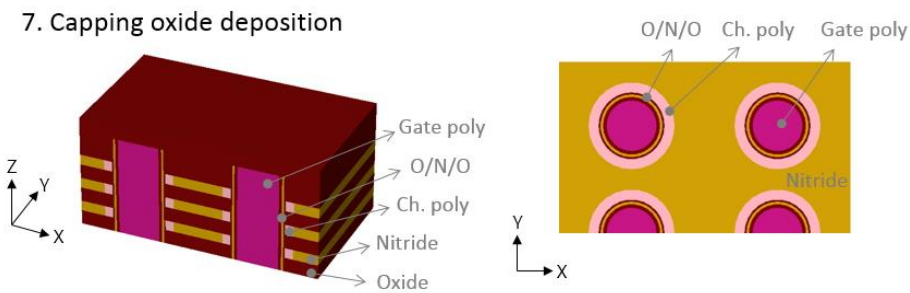


Figure 3.12. Cross-sectional image of capping oxide deposition.

As a process so far, the gate formation has been completed. Afterwards, the process of forming the bit line and source line that exist on both sides of the gate will be explained. Figure 3.13 shows bit line and source line patterning. It is a simple process that is conducted with photo-etching. However, if the overlay with the gate is out of range, it will adversely affect the device operation, which requires attention. In addition, due to the existence of n^+ doped poly-Si. in the pad region, which will be described in Section 3.3, the etching was carried out with 3 steps of oxide etching, poly etching and oxide etching. (Figure 3.13 (a).

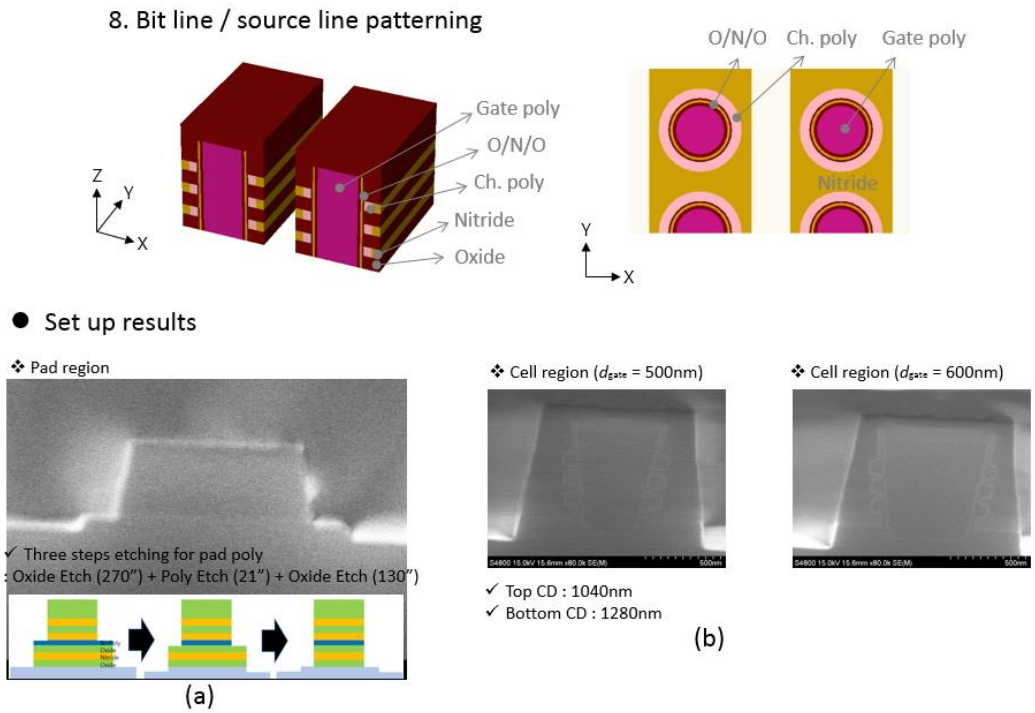


Figure 3.13. Cross-sectional image of bit line and source line patterning. SEM images of the pad region (a), and the cell region (b).

After patterning the source line and bit line, the oxides and nitrides on the side are exposed. Afterwards, phosphoric acid wet etching is carried out to create the space in which the bit line and source line are formed (Figure 3.14). At this time, the oxide and nitride selectivity for the phosphoric acid is important. because if the selectivity is not good, the separation between floors becomes difficult. The selectivity improvement experiment is shown in Figure 3.14 (a), (b). As result of phosphoric acid wet etching after conducted at 800°C 30 min N2 anneal (a) and 950°C 30 min N2 anneal (b), we confirmed that the selectivity between nitride and oxide increases from about 7:1 to 10:1.

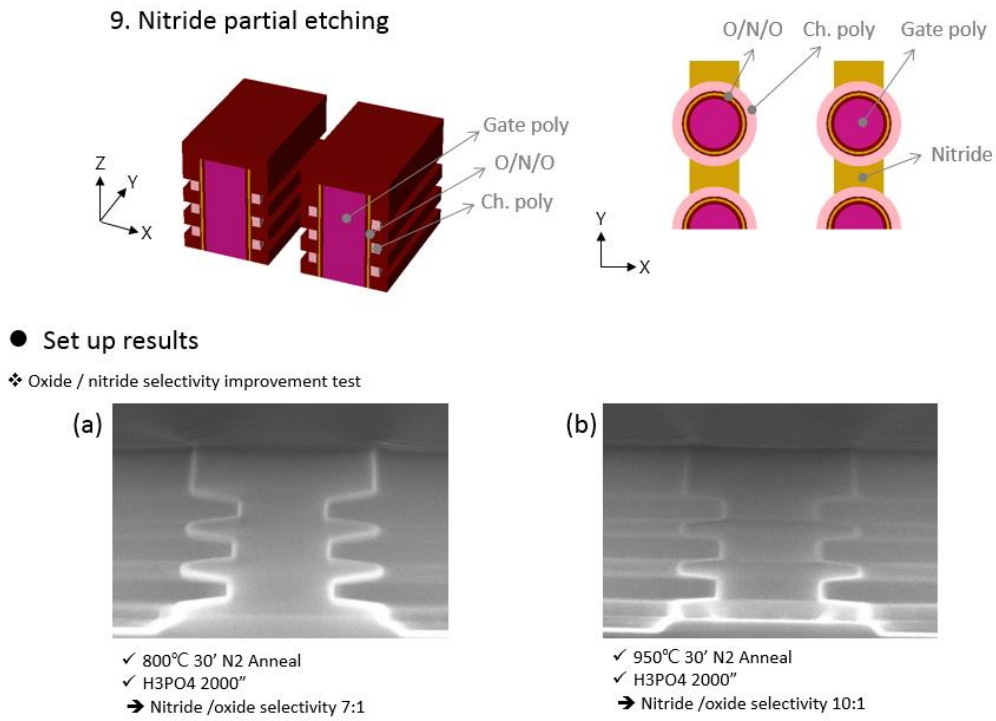


Figure 3.14. Cross-sectional image of nitride partial etching. 800°C 30 min N2 Anneal (a) and 950°C 30 min N2 Anneal (b) tests before the wet etching.

By deposing n^+ poly-Si in the etched region and separation, bit lines and source lines formation are completed. Gate formation is similar to the channel poly-Si separation process(Figure 3.9), but it is easier to process than the channel separation because it is no need to care about the etched surface. Figure 3.15 shows the test results of etching gas SF6. It has been observed that the side etching amount is highly dependent on the flow rate, compared to the time.

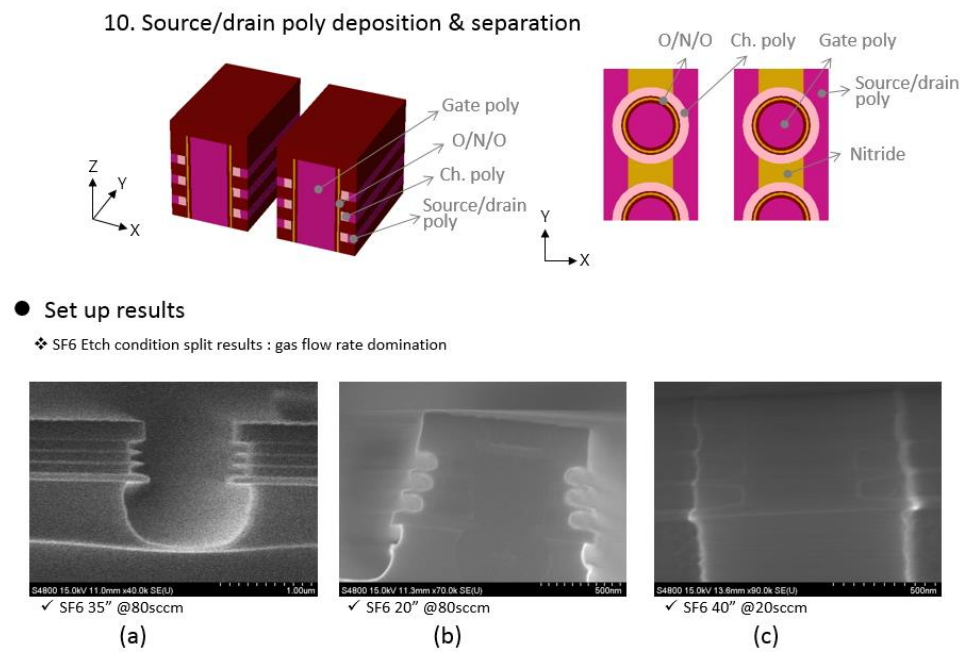


Figure 3.15. Cross-sectional image of junction separation. Etching condition tests: 35seconds at 80sccm (a), 20seconds at 80sccm (a) and 40seconds at 20sccm (c).

The next step is to fill the isolated region with an insulating material and perform CMP (Figure 3.16). High Density Plasma (HDP) CVD oxide was used as the insulation material and 500Å nitride was used as the CMP stopper. The entire stacks must be filled, so the amount of CMP is very large, about 15,000Å. Therefore, since the nitride exposed on the CD SEM can be distinguished from the oxide, the CD SEM was confirmed during the CMP process (Figure 3.16 (b)). The overall wafer uniformity was 15,000 ± 120Å and ± 10% control.

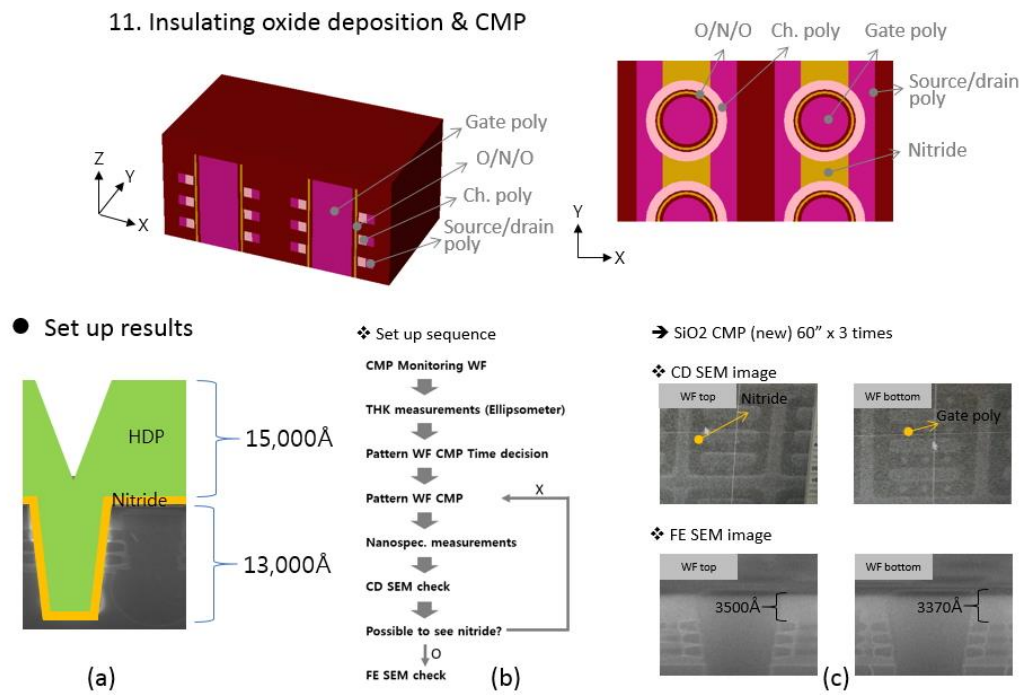


Figure 3.16 Cross-sectional image of insulating oxide deposition and CMP. (a) shows the image immediately after the deposition, (b) shows the CMP sequence, and (c) shows the CD SEM image.

Finally, a contact hole is formed, and metal wiring is done to complete the process. Contact hole etching is performed on the already formed gate with attention to mis-align.

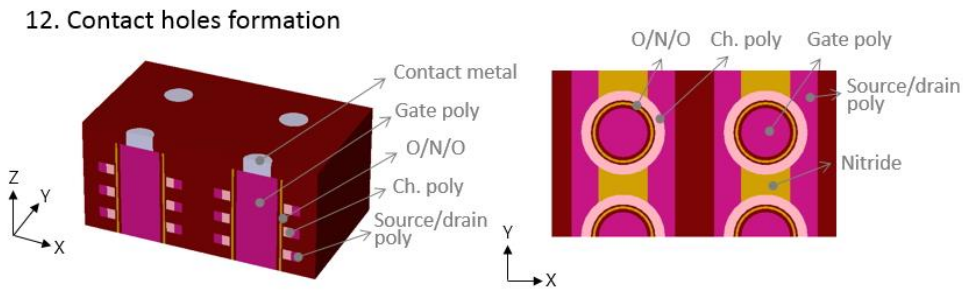


Figure 3.17 Cross-sectional image of contact holes formation.

The cross sectional images of proposed architecture are shown in figure 3.18. Cells with OTOX structures were fabricated in each layer of a three-layer stack.

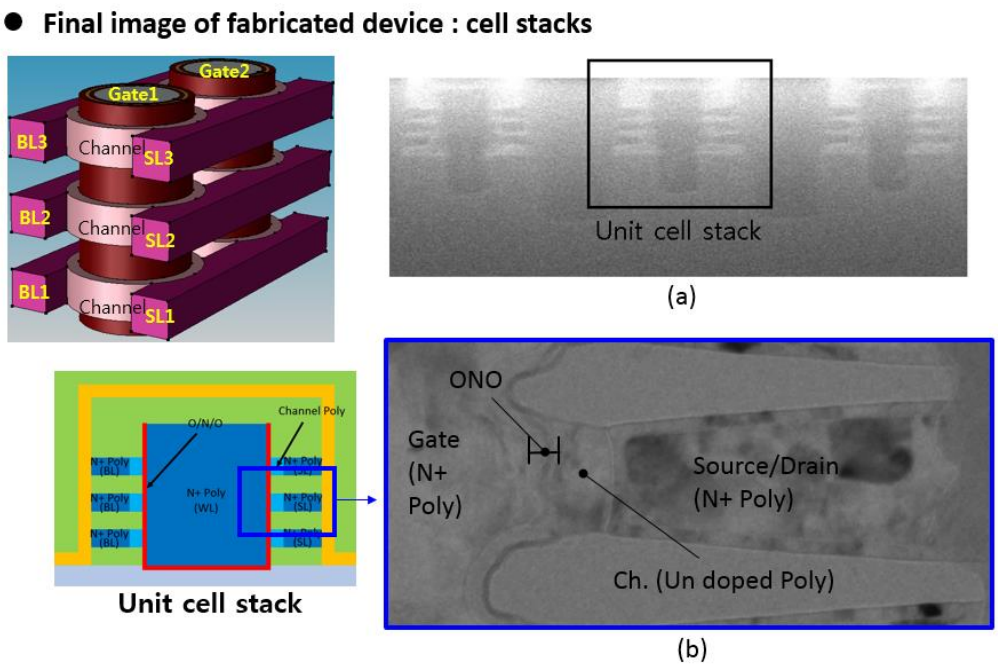


Figure 3.18. Cross sectional images of fabricated stack architecture. (a) SEM cross-section of three cell stacks as part of a cell array, (b) TEM cross-section image corresponding to the box region in the left figure.

3.3 Contact pad process set up

In a 3D NAND flash structure, the edge of the word lines is constructed a stair-shaped structure that forms a contact on it to give a signal (Figure 3.19) [15]. This process is called a resist slimming process, and by repeating RIE and resist slimming, it has the advantage of forming multiple layers with a single mask. However, this process leads to contact and mis-align failures if accurate slimming is not performed because the slimming amount is related to the staircase CD. Since the ISRC does not have equipment to perform accurate slimming, it is difficult to proceed with the sliming process on limited equipment. In addition, since only the bit lines and the source lines are replaced with poly-Si (about 200 nm), it is very difficult to form a contact thereon. Thus, a new process integration was designed to overcome these issues.

Figure 3.20 shows the contact pad process flow. Three masks were used to create three layers. The oxide / nitride stacking process is performed while patterning the n^+ poly-Si for each layer in the region where the contact is to be formed. In this case, n^+ poly-Si is naturally connected to bit lines and source lines

when forming bit lines and source lines.

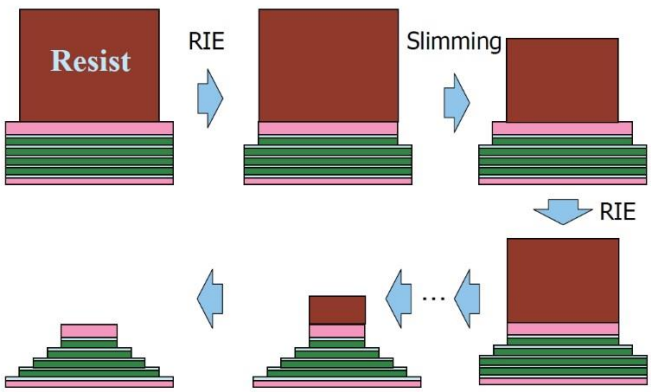


Figure 3.19. Fabrication Sequence of edge of control gates into stair-like structure [15].

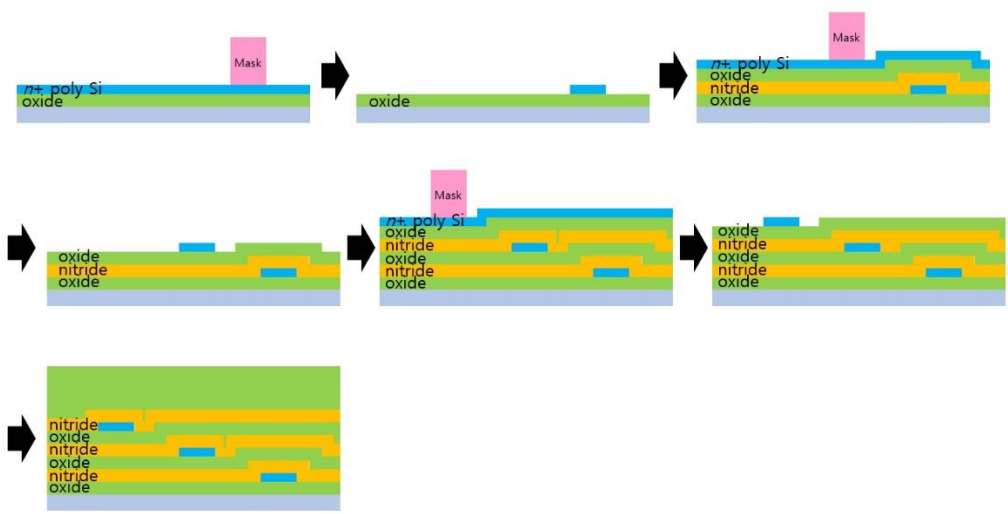


Figure 3.20. Fabrication Sequence of the new contact pad process.

Figure 3.21 shows the plane and vertical views when the final process is completed. In general, contact etching in a pickup region perform several layers at a time, so when the pad is punched, an interlayer short fail occurs. However, if a new pad process is carried out, even if the contact hole is punched at the top layer, there are only nitride and oxide stacks underneath, and the short fail is not occurred because the signal is transferred to bit line and source line at the edge. In practice, the profile of the contact pad after the process is as shown in Figure 3.22, with the upper pad showing excessive etching. The contact at the top layer were punctured, but no electrical failures occurred.

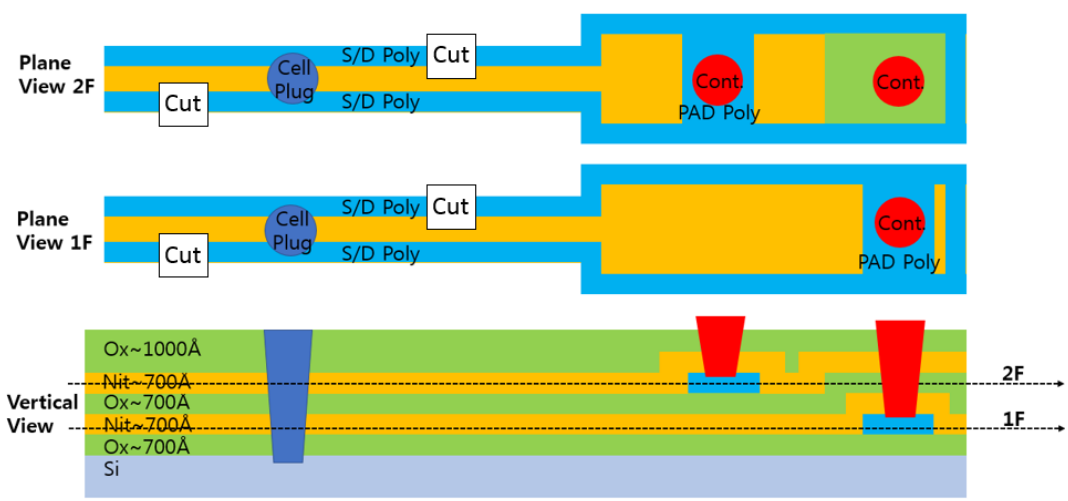


Figure 3.21. The plane and vertical views of contact pad region.

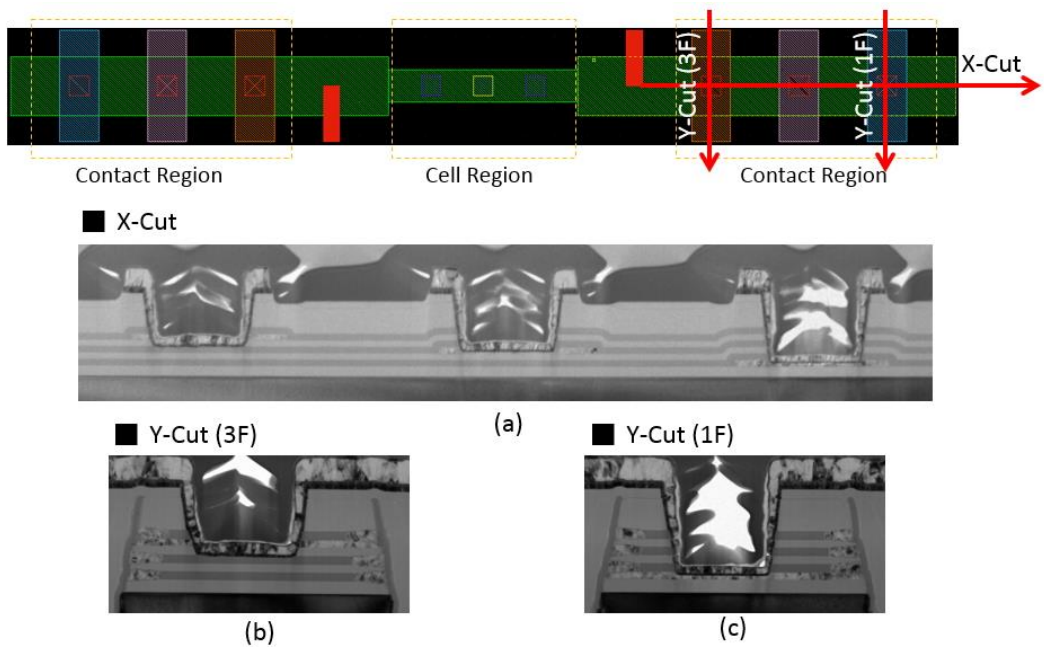


Figure 3.22. The cross sectional TEM images of contact region. X-cut image (a), 3F of Y-cut image (b), and 1F of Y-cut image (c). The contact at the top layer were punctured.

Chapter 4

DC Characteristics of RDC flash device

4.1 DC I-V characteristics

Figure 4.1 shows the initial I_D-V_G curves of fabricated devices. Devices in the first, second and third layers in a stack show similar I_D-V_G characteristics.

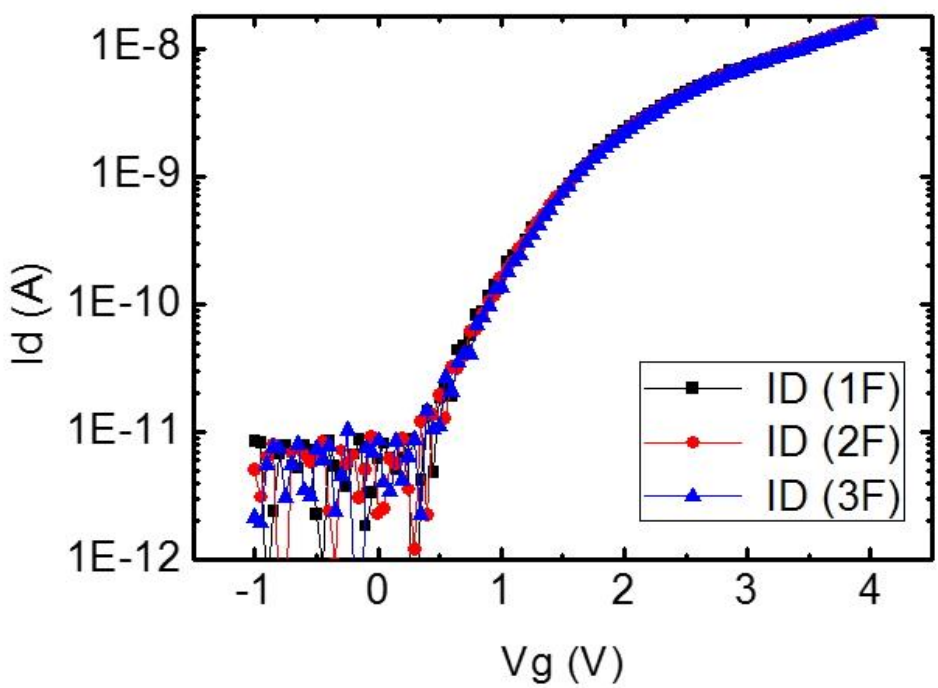


Figure 4.1. Measured I_D-V_G curves of cells in 1st, 2nd, and 3rd floors. It shows similar I_D-V_G characteristics.

In order to confirm the memory operation of the fabricated device, the I_D-V_G characteristics were confirmed while increasing the program bias in the program operation condition, as shown in Figure 4.2. Since the device was fabricated with OTOX structure, it was measured under the gate = V_{PGM} / the bit line and source line = 0V. The I_D-V_G curve is shifted depending on the program bias.

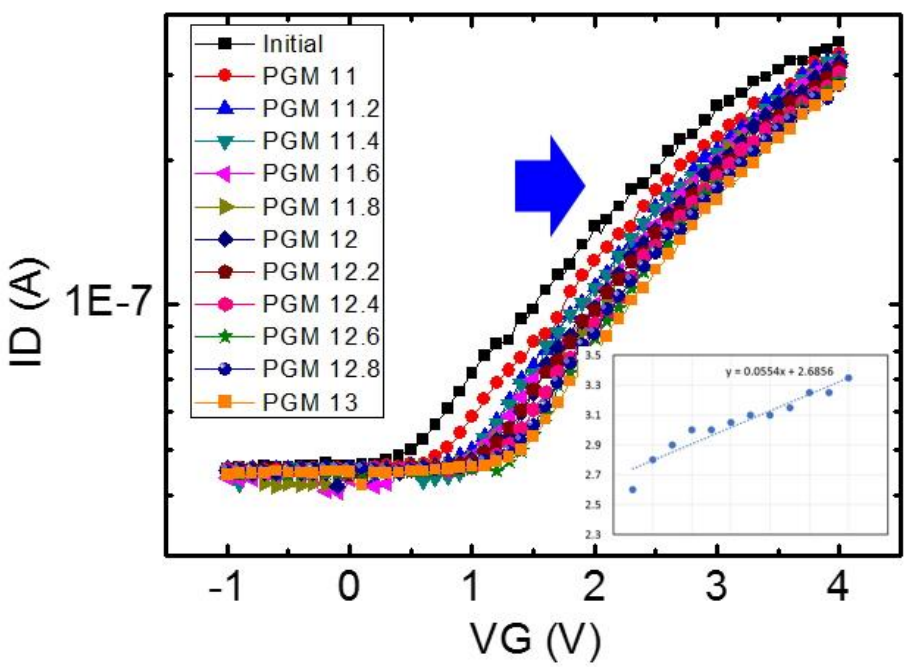


Figure 4.2. Measured ISPP programing characteristics. As the programming voltage increases from 11V to 13V in 0.2V step, the V_{th} of the device changes by about 1V.

Erase characteristics are shown in Figure 4.3. When an erase bias of 8V is applied four times, the cell device has an initial V_{th}. it was shown that Vth was shifted more in the initial pulse when Erasing than programming. As a result, the basic operation of memory was confirmed, and the program-erase window was secured about 1V.

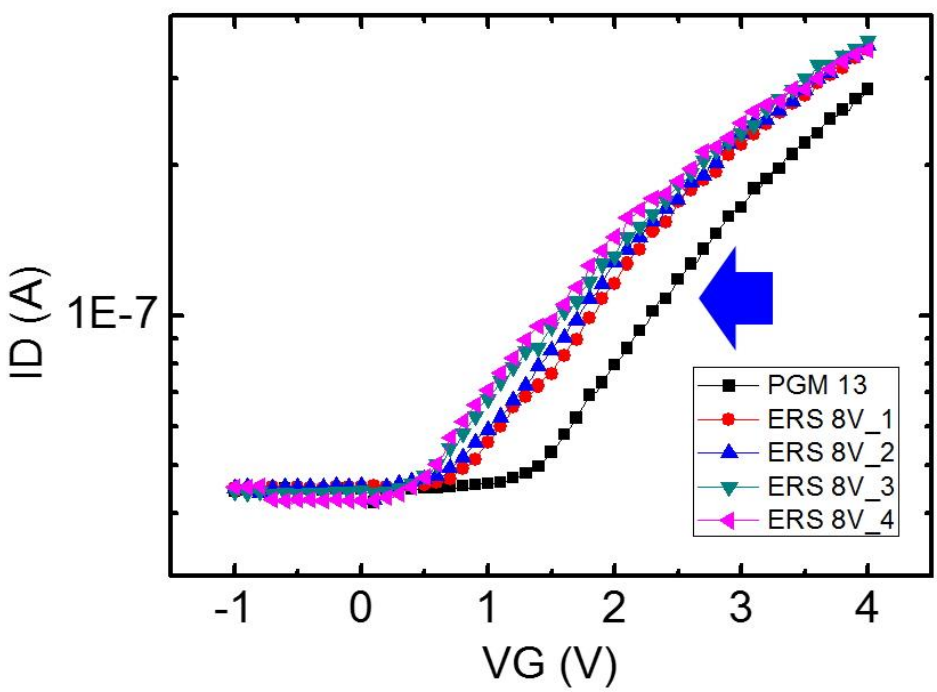


Figure 4.3. Measured erase characteristics. When an erase bias of 8V is applied four times, the cell device has an initial V_{th}.

4.2 Failure analysis

The test modules that can identify the stable Id-Vg curve in the fabricated wafer are about 20%. The others have various types of failures. In this section, we describe the cause of the electrical fail.

Figure 4.4 shows the results of plotting the currents of each measurement terminal under the condition of I_D-V_G measurement (drain = 1V, Source = 0V under Gate bias = sweep). In the normal case, most currents should show a tendency to increase the drain current as the gate bias increases (Figure 4.4 (a)). However, in some test modules, the gate currents flow into the source (Fig. 4.4 (b)) and flow into the source and drain (Fig. 4.4 (c)). In addition, in some cases it was shown that the applied bias was not transmitted at all (Fig. 4.4 (d)).

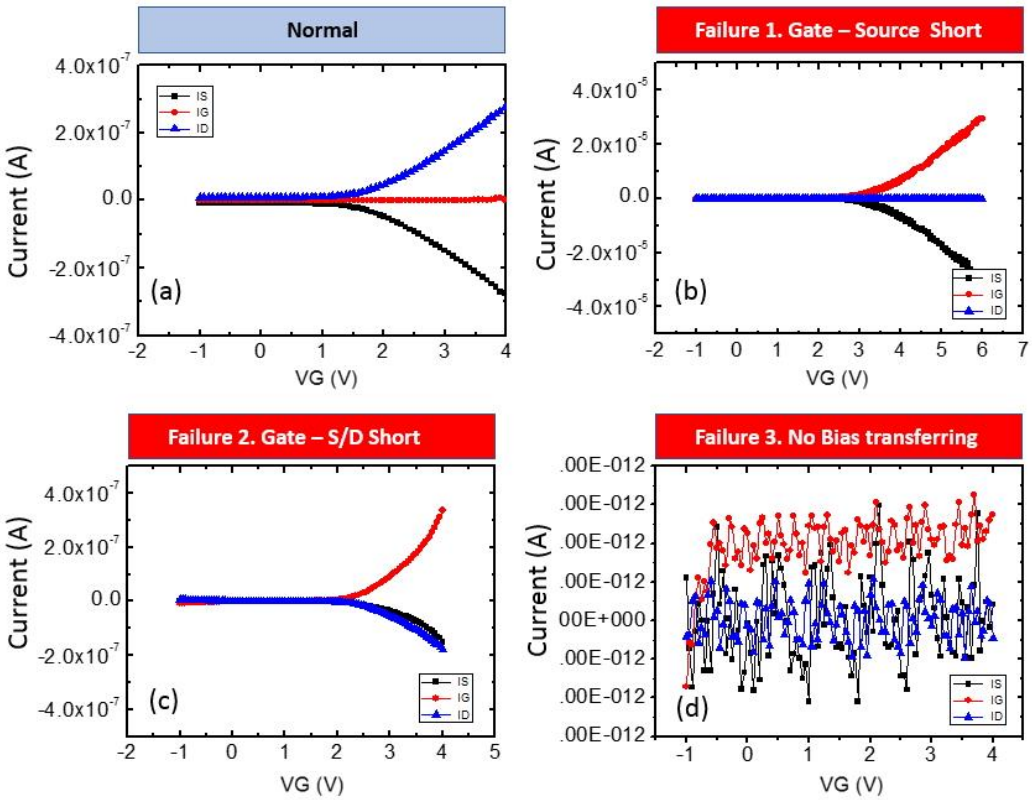


Figure 4.4. Failure mode. Plotting the currents of each measurement terminal under the condition of I_D-V_G measurement (drain = 1V, source = 0V under gate bias = sweep).

From the results of (a) and (b) in Figure 4.4, it can be seen that there are leakage currents between gate and junctions, which can be explained by the weakness in the bit lines and source lines formation process. In the process of forming the source

lines and the bit lines, phosphoric acid wet etching process is performed. During this process, over etching is performed so as to sufficiently meet with the channel poly. At this time, if the selectivity between oxide and nitride is insufficient, O/N/O is attacked by phosphoric acid wet etching (Figure 4.5). If the overlay between the gate and the junctions line goes to one side, this type of failure increases further.

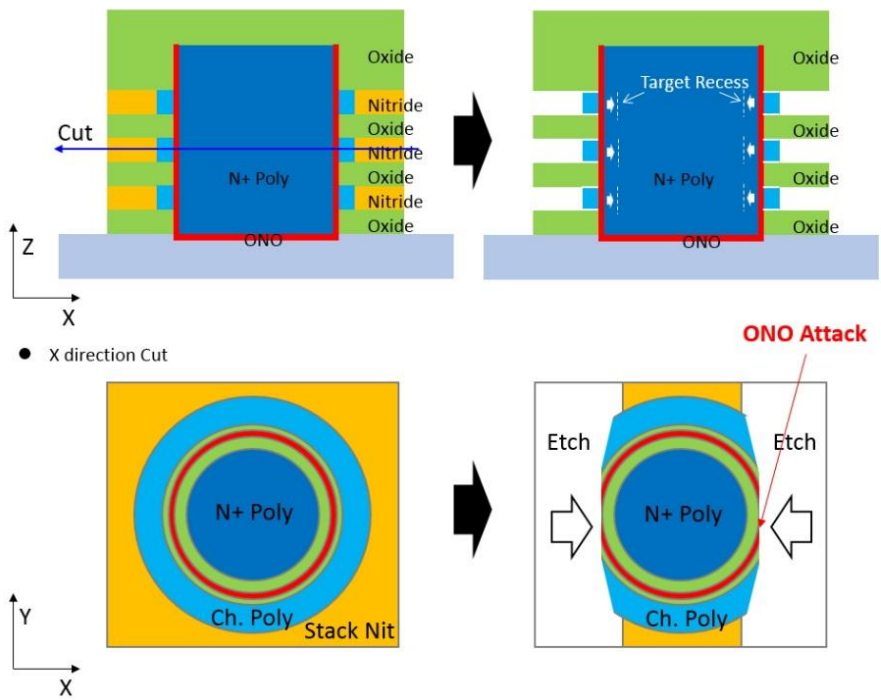


Figure 4.5. Expected cause process for short fail between gate and source. if the selectivity between oxide and nitride is insufficient, O/N/O is attacked by phosphoric acid wet etching.

These results were also confirmed in the cross sectional results. The TEM image shows that the O / N / O of the third layer cell was attacked (Figure 4.6). The reason for the weakness of the upper layer is that when the junction line patterning, etch slope occurs, resulting in longer exposure time to the O/N/O for subsequent phosphoric acid wet etching.

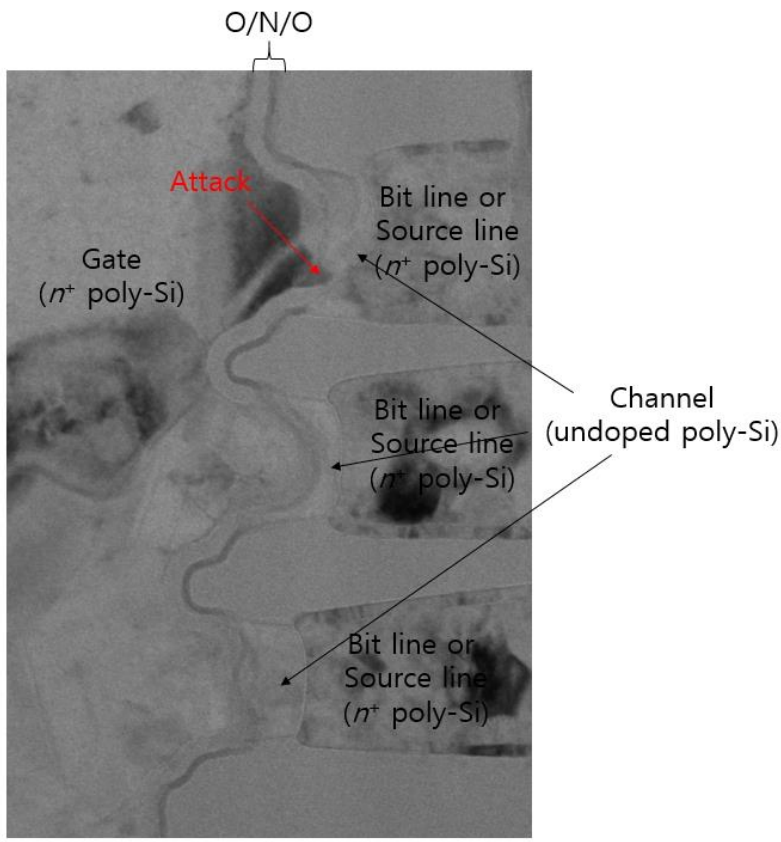


Figure 4.6. The cross-sectional TEM image of failure cell. the O / N / O of the third layer cell was attacked.

Failure 3 in Figure 4.4 (d) can be explained by the TEM results in Figure 4.7. During gate etching and bit line and source line etching, if the first layer does not properly wet etching the first layer due to the etching slope, the channel poly and the bit line and source line poly do not meet each other, and the corresponding fail occurs. These failures are more likely to occur if the overlay between the gate and the junction line is outside the allowable range.

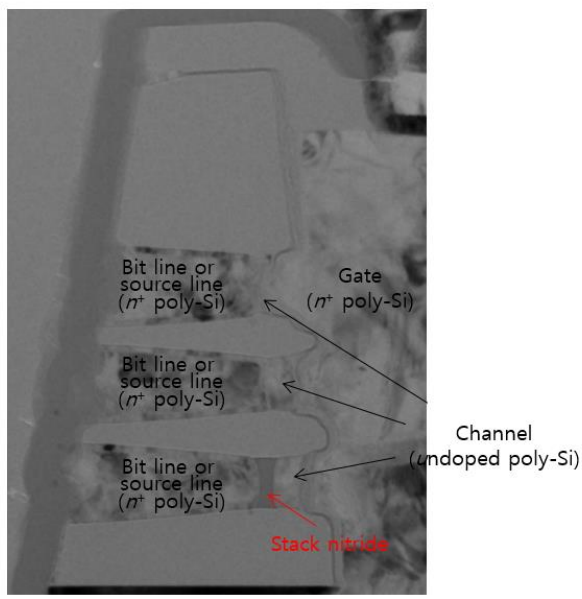


Figure 4.7. The cross-sectional TEM image of failure cell. the O / N / O of the third layer cell was attacked. The stack nitride on the first floor is not properly etched and remains.

Finally, the reasons for the deterioration of the electrical characteristic variations between measurement modules can be explained in Figure 4.8 plane TEM image. The surface of the channel is not smooth. This is due to the isotropic channel separation etching process using SF6 gas.

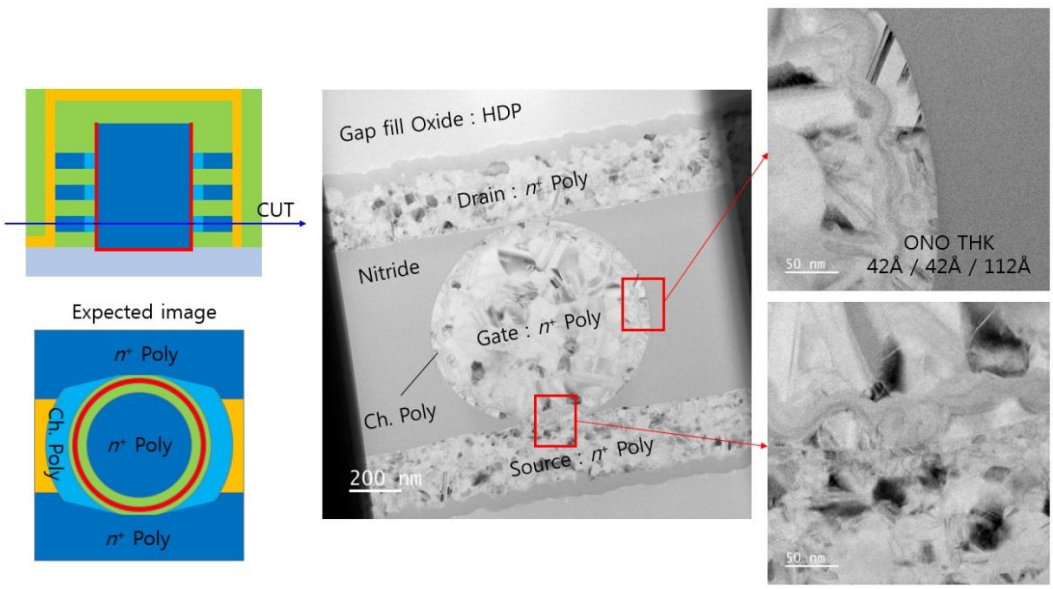


Figure 4.8. The plane TEM image of the cell. The surface of the channel is not smooth. This is due to the isotropic channel separation etching process using SF6 gas.

4.3 New integration design

The causes of the electrical failures were investigated. The leakage problem between gate and junction line can be solved by improving the etching slope and increasing the selectivity between oxide and nitride for phosphoric acid. In addition, the problem that the bias is not transmitted should improve the overlay between gate and junction line, and the electrical variation issues require a new method for channel separation. This chapter introduces new process integration to solve this problem.

The new process integration is illustrated in Figures 4.9 to 11. The key point is that the gate last process is designed to overcome the leakage problem and the straight dry etch method is used to improve the channel profile. First, the multiple oxide/nitride layers are deposited on Si substrate. The junction lines are patterned, and nitride layers are partially etched (a)~(c). n^+ poly-Si is deposited and separated each layer (d). The nitride film is deposited for CMP stopping layer (e). Then, the insulating film is deposited, and CMP is performed (f). Next, via holes are patterned (g)~(i). and nitride layers are partially etched until encountering the junction line(j).

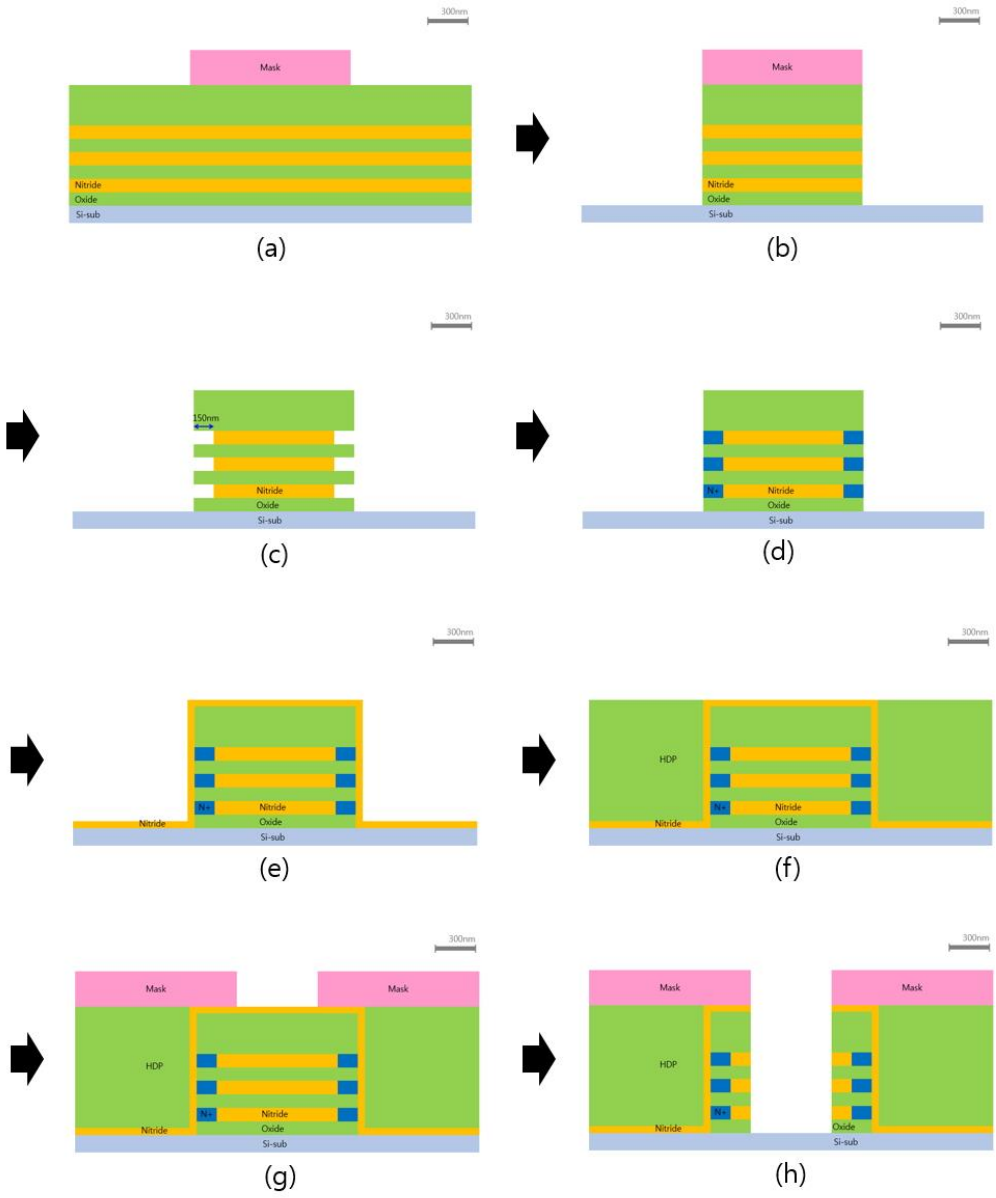


Figure 4.9. The junction line formation in new process integration. In order to improve the leakage problem between gate and junction line, the gate last process has been designed.

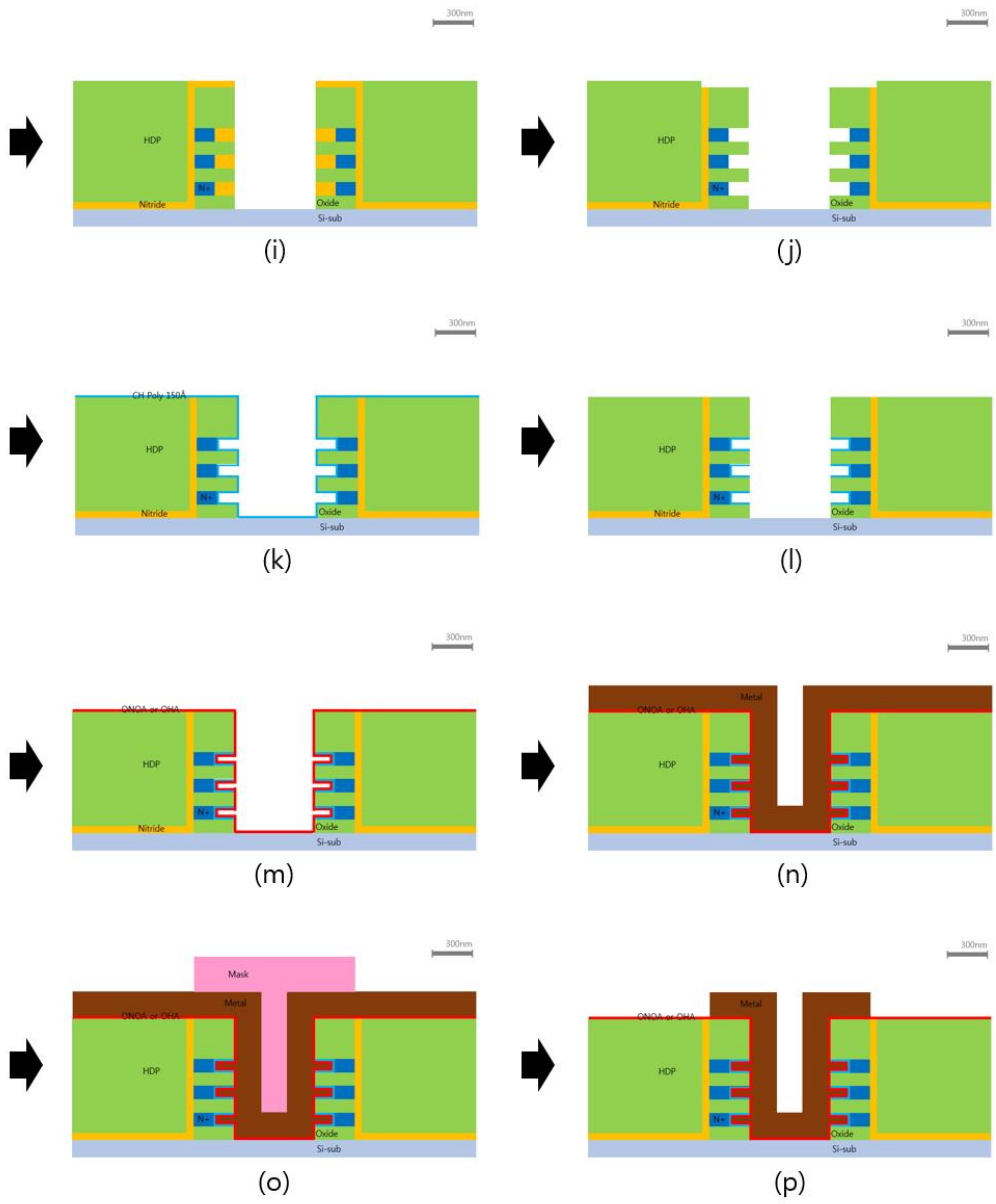


Figure 4.10. The gate formation in new process integration. For the smooth channel profile, Straight-line dry etching method was used in channel separation process.

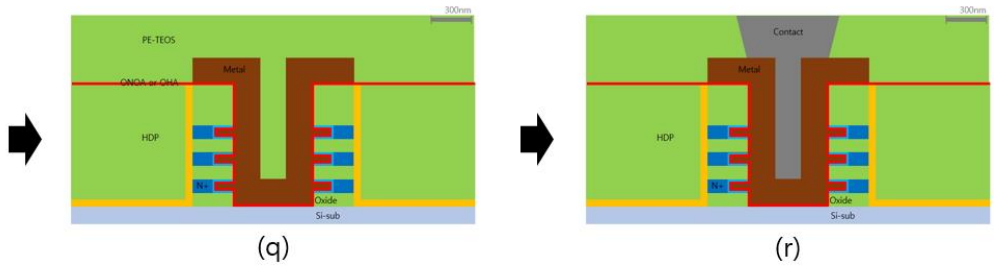


Figure 4.11. The contact formation in new process integration.

poly-Si for the channel is formed in the nitride etched grooves (k) and separated by straight-line dry etching method (l). The tunnel oxide, trap nitride, blocking oxide or high-*k* metal are deposited sequentially, and the metal is deposited for the gate (n)~(p). Finally, after forming the interlayer insulating film (q), contact holes are formed, and metal wiring is performed (r).

Chapter 5

Conclusion

I have proposed a highly integrated AND stack architecture that can be useful in neuromorphic computing and verified the operation of cells that can be used as synaptic devices in each layer of the stack. In addition, device operation for two gate insulator stacks (ITOX and OTOX types) was successfully verified through TCAD simulation. I have successfully verified the program and erase of selected devices in the array and the inhibition of unselected devices using device simulation. Stacks with three layers were fabricated and the operation of the cells in each layer was verified. Also, the cause of fail through TEM analysis was performed well, and the solution was suggested. The proposed 3-D stacked AND architecture and device structure proposed in this work will be a promising candidate for highly integrated synaptic device.

Bibliography

[1] G. E. Hinton, R. R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks”, *Science,* Vol. 313, Issue 5786, pp. 504-507, 28 Jul 2006.

[2] Bing Chen, Fuxi Cai, Jiantao Zhou, Wen Ma, Patrick Sheridan, and Wei D. Lu, “Efficient in-memory computing architecture based on crossbar arrays,” *Electron Devices Meeting (IEDM)*, 2015.

[3] Duygu Kuzum, Shimeng Yu, and H-S PhilipWong, “Synaptic electronics: Materials, devices and applications,” *Nanotechnology*, vol. 24, no. 38, p. 382001, 2013..

[4] Peyman Pouyan, Esteve Amat, and Antonio Rubio "Reliability Challenges in Design of Memristive Memories," *2014 5th European Workshop on CMOS Variability (VARI)*, 29 Sep.-1 Oct.

[5] Roberto Bez, Emilo Camerlenghi, Alberto Modelli, and Anglo Visconti, “Introduction to Flash Memory,” *Proceedings of the IEEE*, Vol. 91, No. 4, April 2003.

[6] Dmitri Strukov, Farnood Merrikh-Bayat, Mirko Prezioso, Xinjie Guo, Brian Hoskins, and Konstantin Likharev “Memory Technologies for Neural Networks,” *2015 IEEE International Memory Workshop (IMW)*, 2015.

[7] “Memristor overview”, https://www.slideshare.net/SCU_ECE_Staff/memristor-overview. Accessed April 16, 2019.

[8] Peyman Pouyan, Esteve Amat, and Antonio Rubio "Reliability Challenges in Design of Memristive Memories," *2014 5th European Workshop on CMOS*

Variability (VARI), 29 Sep.-1 Oct.

[9] “NAND Flash memory”, https://docplayer.net/20969248-Nand-flash-memory-samsung-electronics-co-ltd.html. Accessed April 16, 2019.

[10] “Overview of 3D NAND Flash and progress”, https://www.semanticscholar.org/paper/Overview-of-3D-NAND-Flash-and-progress-of-3D-gate. Accessed April 16, 2019.

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

[12] Hang-Ting Lue, Weichen Chen, Hung-Sheng Chang, Keh-Chung Wang, and Chih-Yuan Lu, “A Novel 3D AND-type NVM Architecture Capable of High density, Low power In Memory Sum of Product Computation for Artificial Intelligence Application,” *2018 Symposium on VLSI Technology*, 2018.

[13] “Structure and principle of flash memory”, https://imsosimin.com/5. Accessed April 16, 2019

[14] Akihiro Nitayama and Hideaki Aochi, “Bit Cost Scalable (BiCS) Technology for Future Ultra High Density Memories,” *International Symposium on VLSI Technology, Systems and Application (VLSI-TSA)*, 22-24 April 2013.

[15] H.Tanaka, M.Kido, K.Yahashi, M.Oomura, R.Katsumata, M.Kito, Y.Fukuzumi, M.Sato, Y.Nagata, Y.Matsuoka, Y.Iwata, H.Aochi and A.Nitayama “Bit Cost Scalable Technology with Punch and Plug Process for Ultra High Density Flash Memory”, *2007 Symposium on VLSI Technology*, 2007.

초 록

기존의 "Von Neumann 아키텍처"는 버스를 통해 메모리 장치에 연결된 컴퓨터에서 계산을 수행한다. 이와 같은 구조는 전력 및 속도의 단점을 보이는데, 오늘날 고용량 컴퓨팅의 가장 큰 한계라 할 수 있다. 반면, 뇌의 계산을 모방하는 뉴로모픽 컴퓨팅은 메모리와 로직을 병렬로 동작 시키므로 고용량의 컴퓨팅에 매우 바람직하다.

뉴로모픽 컴퓨팅은 뉴런과 시냅스의 연결 강도를 이용하여 수행된다. 시냅스는 기존 컴퓨팅에서 사용되는 메모리와 유사한 기능을 가지고 있는데, 성공적인 뉴로모픽 컴퓨팅을 위해 시냅스 소자는 고집적, 저전력, 빠른 속력, 높은 안정성 및 비 휘발성과 같은 특성을 가져야 한다. 하지만 아직까지 이러한 모든 조건을 충족하는 표준화된 소자는 없다. 다양한 멤리스터 기반의 장치가 시냅스 소자로 보고되었지만 여전히 소자 안정성과 산포 문제가 존재한다. NOR 플래시는 훨씬 성숙된 소자이나, 메모리 셀(~ 10F2)의 큰 사이즈 때문에 집적도가 떨어진다. 한편, 고집적 NAND 플래시 메모리는 셀이

연속으로 직렬 연결되기 때문에 뉴로모픽 연산에 필요한 전류 곱의 합을 구현하기가 어렵다.

이 논문에서는 효과적으로 뉴로모픽 컴퓨팅을 수행할 수 있는 새로운 3차원 스택 구조 시냅스 소자를 제안한다. 스택으로 구성된 AND 형 RDC (Rounded Dual Channel) 플래시 메모리 소자는 FN 프로그램/삭제 방법을 사용하여 저전력으로 작동하고, 병렬 읽기 동작으로 빠른 속도를 보이며, 스택을 통해 고집적 구현이 가능하다. 제안된 3차원 AND 구조 소자의 시냅스 가능성을 검증하기 위해, 본 논문에서는 제안된 소자의 주요 제조 단계가 설명되며, 3차원 시뮬레이션을 통해 적층 구조에서 소자의 동작을 확인한다. 또한, 3층 소자를 제작하여 다수층에 대한 적층 가능성과 실제 동작을 검증한다.