Ph.D. DISSERTATION

Gated Twin-Bit (GTB) SONOS NAND Flash Memory for High Density Nonvolatile Memory

고집적 비휘발성 메모리 구현을 위한 차단 게이트를 갖는 2 비트 낸드 플래시 메모리

BY

WONBO SHIM

February 2013

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

Gated Twin-Bit (GTB) SONOS NAND Flash Memory for High Density Nonvolatile Memory

고집적 비휘발성 메모리 구현을 위한 차단 게이트를 갖는 2 비트 낸드 플래시 메모리

지도교수 박 병 국

이 논문을 공학박사 학위논문으로 제출함

2013 년 2 월

서울대학교 대학원

전기컴퓨터공학부

심 원 보

심원보의 공학박사 학위논문을 인준함

2013 년 1 월

위 원 장 : 신 형 철 (인)

부위원장 : 박 병 국 (인)

위 원 : 이 종 호 (인)

위 원 : 조 훈 영 (인)

위 원 : 심 재 성 (인)

ABSTRACT

In this dissertation, the gated twin-bit (GTB) SONOS NAND flash memory is studied. To meet the demand of high density nonvolatile memory, a GTB SONOS array having a cut-off gate and two storage nodes at a single wordline is designed. By using the forward-reverse read scheme, the memory density can be increased to 2 bit per $4F^2$ area which is twice the conventional NAND array.

Recent trend of nonvolatile memories are introduced, and the solutions of scaling down of NAND flash memory at this time is mentioned in Chapter 1 and 2. In Chapter 3, the GTB NAND device is introduced and the operation scheme of GTB NAND array is investigated with the TCAD simulation. In the case of program operation, the cut-off gate plays a role of preventing the channel potential transfer and enables separate programming. In the case of read operation, forward-reverse read scheme is used to read the bits with a common wordline. Fabrication processes are introduced, and several important issues in the GTB array are verified. The process flow and images of fabricated GTB SONOS devices are presented, and electrical characteristics are measured in Chapter 4.

In addition, a gated multi-bit (GMB) NAND array which can compete with recent 3-dimensional NAND structures is introduced in Chapter 5. This structure has single

crystalline silicon channel and 2N bit per $4F^2$ memory density. Fabrication processes are stated, and operations are verified with simulation.

Keywords: cut-off gate, gated multi-bit (GMB), gated twin-bit (GTB), 3D NAND flash memory, twin-bit operation

Student number: 2007-21010

CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Flash Memory Devices in Nanoelectronics Era

From the last part of the 20th century, many electronic devices such as personal computer, notebooks had been used popularly. In recent years, the consumption of smart phone, smart TV, tablet PC, digital camera, and mobile electronic devices has increased explosively. To meet these demands, semiconductor based memories are commonly used and researchers are trying to scale down the size of the device

Semiconductor based computer memories, especially in MOS based memory can be categorized as volatile memory and nonvolatile memory as shown in Fig. 1.1. Volatile memory requires power to maintain the stored information. In other words when the power is off, the stored memory is lost. A volatile memory will typically have a worst-case retention time of less than a second. On the contrary, a nonvolatile memory is usually specified as meeting a worst-case unpowered retention time of 10 years. The

two main forms of modern volatile memory are static RAM (SRAM) and dynamic RAM (DRAM). In SRAM, a bit of data is stored using the state of a flip-flop. SRAM, which is often used as cache memory for the CPU, is most expensive to produce because of its large size, but is generally faster and requires less power than DRAM. DRAM stores a bit of data using a transistor and a capacitor; the capacitor holds a low or high charge (0 or 1) and the transistor acts as a switch that lets the state of capacitor to be changed. Because DRAM is less expensive to produce than SRAM, it is the predominant form of computer memory used in modern computers.

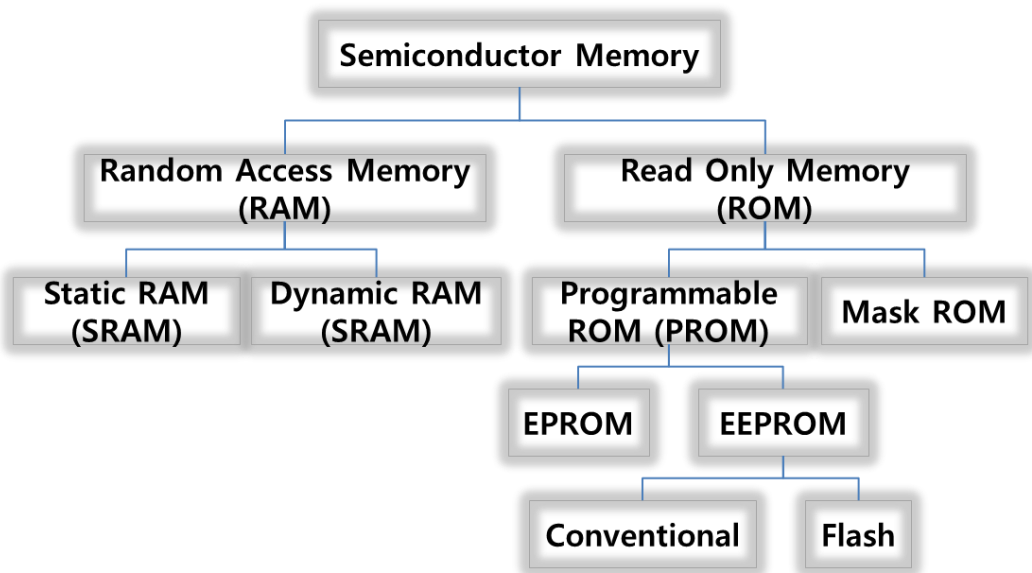


Fig. 1.1 Classification of semiconductor MOS memories.

Table 1.1 Nonvolatile Memory Functional Capability Classifications

Acronym	Definition	Description
ROM	Read-only memory	Memory contents are defined during manufacture and not modifiable.
EPROM	Erasable programmable ROM	Memory is erased by exposure to UV light and programmed electrically.
EEPROM	Electrically erasable programmable ROM	Memory can be both erased and programmed electrically. The use of "EE" implies block erasure rather than byte erasable.
E^2PROM	Electrically erasable programmable ROM	Memory can be both erased and programmed electrically as for EEPROM, but the use of "E^2" implies byte alterability rather than block erasable.

Integrated circuit nonvolatile memories are classified with the degree of functional flexibility available for modification of stored contents as indicated in Table 1.1 [1]. Flash memory is an EEPROM where the erase procedure of entire chip or subarray within the chip is performed at one time [2]. NAND flash is oriented toward data-block storage applications, and NOR flash is suited for code and word addressable data storage, are two typical flash memory [3].

Many common memory devices are charge based where charge can be injected into or removed from a critical region of a device (writing), and the presence or absence of the charge can be sensed (reading). As stated above, DRAM stores the charge on a capacitor and SRAM stores in a flip-flop, and flash memory stores the charge in a floating gate.

Flash memory is invented by Dr. Fujio Masuoka of Toshiba Corp. in 1984. Based on Masuoka's invention, Intel Corp. commercialized common ground NOR flash memory in 1988 as a non-volatile storage medium to store program codes including a PC's BIOs and firmware for various consumer products. NOR flash was also the basis for the first flash memory cards and non-volatile solid state drives in the 1990s. Toshiba Corp. introduced NAND flash memory in 1988 which promised lower cost per bit than NOR flash and faster program and erase throughput. Unlike NOR flash which is organized on a byte or word basis, NAND flash is organized into pages and erased on a block basis. A block consists of 64 or more pages. The organization of the NAND flash is conducive to lower cost per bit but not suitable for random access.

In recent years, flash memory is the mostly used nonvolatile memory because of its great scalability. Flash memory revenue is predicted to hit $32.8 billion in 2012, which is 11 percent increased from 2011 as shown in Fig. 1.2. Among them, NAND flash revenue comprises $29.5 billion whereas NOR flash comprises $3.3 billion.

Due to its relatively simple structure and high demand for higher capacity, NAND flash memory is the most aggressively scaled technology among electronic devices. The heavy competition among the top few manufacturers only adds to the aggressiveness in shrinking the design rule or process technology node [5]. While the expected shrink timeline is a factor of two every three years per original version of Moore's law, this has recently been accelerated in the case of NAND flash to a factor of two every two years.

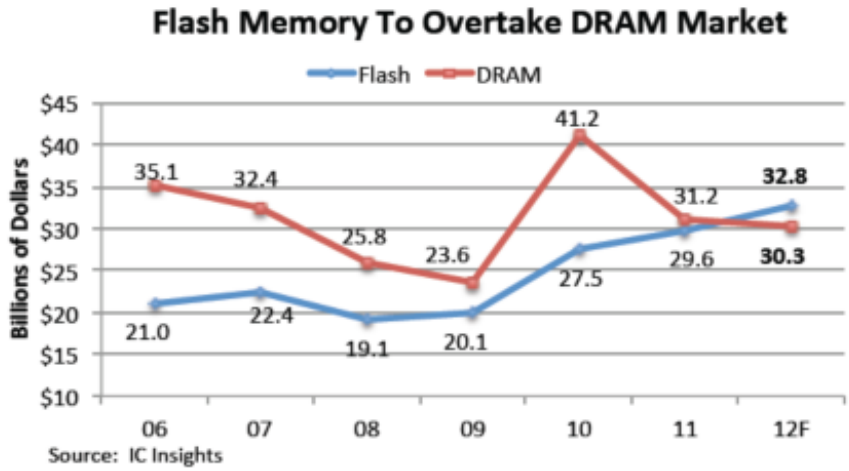


Fig. 1.2 Flash memory and DRAM market revenue.

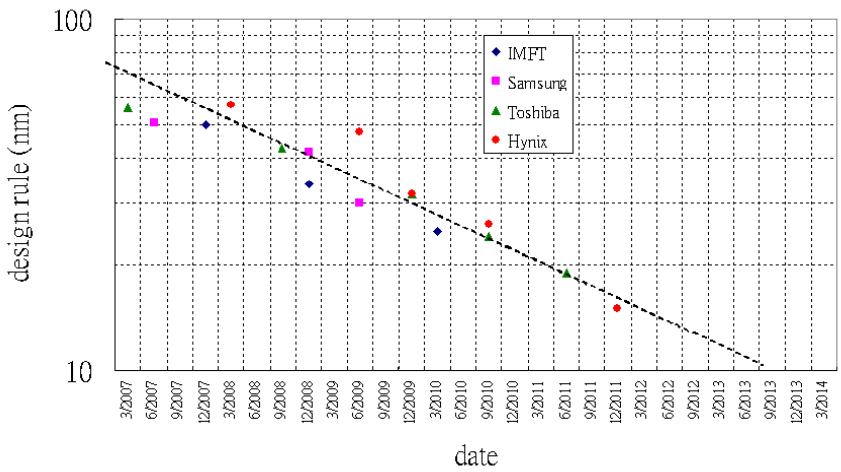


Fig. 1.3 Aggressive shrinking of design rule of NAND flash memory technology.

As shown in Fig. 1.4, NAND flash is used for data storage of mobile devices, media card, and digital consumer electronics. Furthermore, NAND flash is drawing attention as a replacement of hard disk drives recently. Flash memory does not have the

mechanical limitations and latencies of hard drives, so a solid-state drive (SSD) is attractive when considering speed, noise, power consumption, and reliability. In addition, SSDs have no moving parts and are resistant to shock and vibration. Flash drives are gaining traction as mobile device secondary storage devices; they are also used as substitutes for hard drives in high-performance desktop computers and some servers with RAID and SAN architectures.

Although there remain some aspects of flash-based SSDs that make them unattractive (for example, cost per gigabyte of flash memory remains significantly higher than that of hard disks) [6], the production cost decreases very rapidly ($0.48 USD per gigabyte in Oct. 2012). So many types of recent tablet PCs use SSD rather than HDD. Growth of SSD market can lead the explosive increase of NAND flash consumption.

Fig. 1.4 Applications of NAND flash memory.

1.2 Introduction to NAND Flash Memory

1.2.1 Introduction to Flash Memory

The history of nonvolatile memories began in the 1970s, with the introduction of the first EPROM memory. Since then, nonvolatile memories have always been considered one of the most important families of semiconductors and, up to the 1990s, their interest was tied up more to their role as a product of development for new technologies than to their economic value. Since the early 1990s, with the introduction of nonvolatile flash memories into portable products like mobile phones, palmtop, camcorders, digital cameras and so on, the market of these memories has experienced a drastic increase.

The most popular flash memory cell is based on the floating gate (FG) technology. As shown in Fig 1.5, a MOS transistor is built with two overlapping gates rather than a single one: the first one is completely surrounded by oxide, while the second one is contacted to form the gate terminal. The isolated gate constitutes an excellent "trap" for electrons, which guarantees charge retention for years. The operations performed to inject and remove electrons from the isolated gate are called program and erase, respectively. These operations modify the threshold voltage (V_{TH}) of the memory cell, which is a special type of MOS transistor. Applying a fixed voltage to cell's terminals, it is then possible to discriminate two storage levels: when the gate voltage is higher than the cell's V_{TH}, the cell is on ("1"), otherwise it is off ("0").

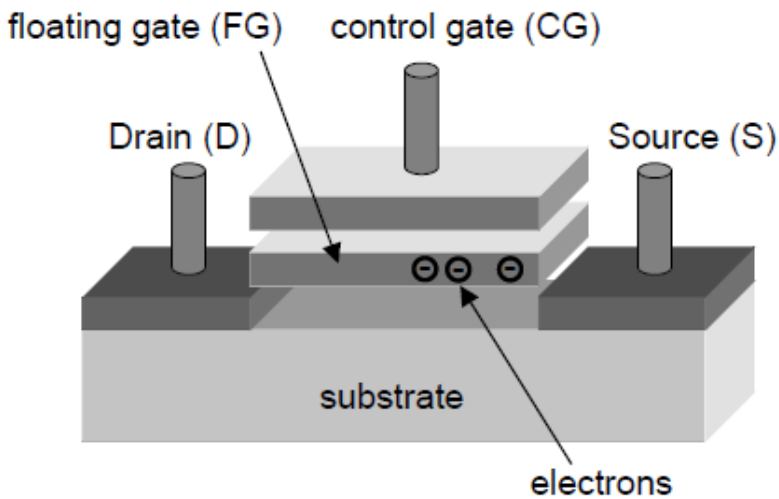


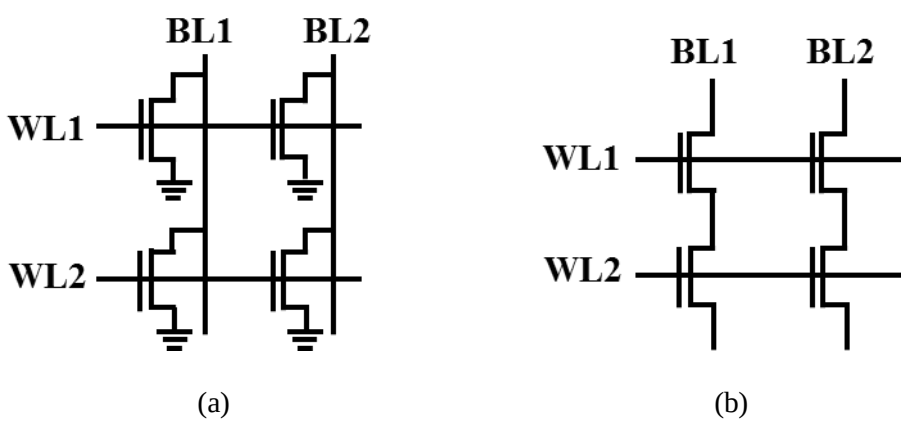
Fig. 1.5 Floating gate memory cell structure

1.2.2 NOR and NAND Array Architecture

The two main flash array architectures are shown in Fig. 1.6. NOR flash memory employs a parallel array architecture where each cell may be accessed via a contact. The direct cell access is the reason for the superior random performance of NOR flash. However, a metal contact is required for every cell in an array, that causes relatively large cell area about $10F^2$ where F is minimum feature size. Also, NOR cell array uses the channel hot electron injection (CHEI) mechanism for programming. So, it is impossible to program a large amount of cells at the same time because of the large power consumption from the low injection efficiency ($\sim 1\times10^{-6}$) of CHEI mechanism. Because random access is quite useful in memory that is used for the code storage and most of the

code storage applications do not need mass storage, NOR flash architecture is widely used for this type of memory.

In contrast, the memory cells in NAND flash are organized serially. 64 memory cells are connected in series between the two select transistors. Random performance is slow due to the fact that there are no contacts directly accessing the memory cells. However, because there are only two contacts every 64 memory cells, the effective cell size is much smaller than for NOR flash resulting in a smaller chip size and lower cost per bit. The cell size of NAND flash is generally in the $4F^2$ range. NAND exhibits fast page writes due to the ability to write 4-8kB simultaneously resulting in very high sequential write throughput. The serial architecture and small cell size make NAND flash optimized for low cost mass storage.



(a) (b)

Fig. 1.6 (a) NOR and (b) NAND architecture with circuit symbols.

Chapter 2

Scaling of NAND Flash Memory

Flash memory has a lot of advantages compared with other types of nonvolatile memory devices. Above all, the possibility of low cost production with higher density is attractive for both production and market. During the past decade, the integration density has increased twice per year outstripping Moore's law. However, continuous device scaling has encountered a difficulty because of several obstacles.

In this chapter, several obstacles of continuous scaling will be mentioned, and several solutions to scaling down the NAND flash memory below the 20 nm length will be introduced.

2.1 Obstacles in NAND Flash Scaling

Scaling down of the NAND flash memory is the most essential issue for the low cost production because much more memory transistors can be produced in the same area of silicon wafer. If the feature size of the NAND flash is decreased by 30 percent, the

memory density can be doubled.

However, several obstacles exist to scale down below the 20 nm channel length as shown in Fig. 2.1. The main problem is the limit of downscaling of line width in today's ArF photo lithography process. To overcome the limitation ArF photolithography, phase shift mask (PSM) [7], optical proximity correction (OPC) [8], immersion lithography [9], double patterning technique (DPT) [10] are used.

Although highly scaled NAND array is realized, electrical disturbances come to the fore as shown in Fig. 2.1(b) [11]. As the distance between the neighboring floating gates is getting closer, floating gate(FG) to floating gate interference makes severe disturbance during the program and read operation; V_T can be changed as we do not intend. To reduce the FG to FG interference, the height of the FG should be reduced. However, low height of FG reduces the CG-FG coupling ratio so that program and erase speed slows down. And scaled tunneling oxide thickness can degrade the retention characteristics.

Fig. 2.1(c) shows the number of electrons for a 100 mV V_T shift as the feature size scaled down [12]. The volume of the FG decreases with feature size, so that the small number of electrons can induce V_T shift. It means the fluctuation of the number of injected charge or bad retention characteristic of tunneling oxide should be taken seriously in small feature size structure.

To prevent the short channel effect, substrate doping concentration must be increased as the technology node decreases as shown in Fig. 2.1(d) [11]. High substrate doping concentration lowers the self-boosted potential of inhibit cell, that induces the

narrow V_{pass} window.

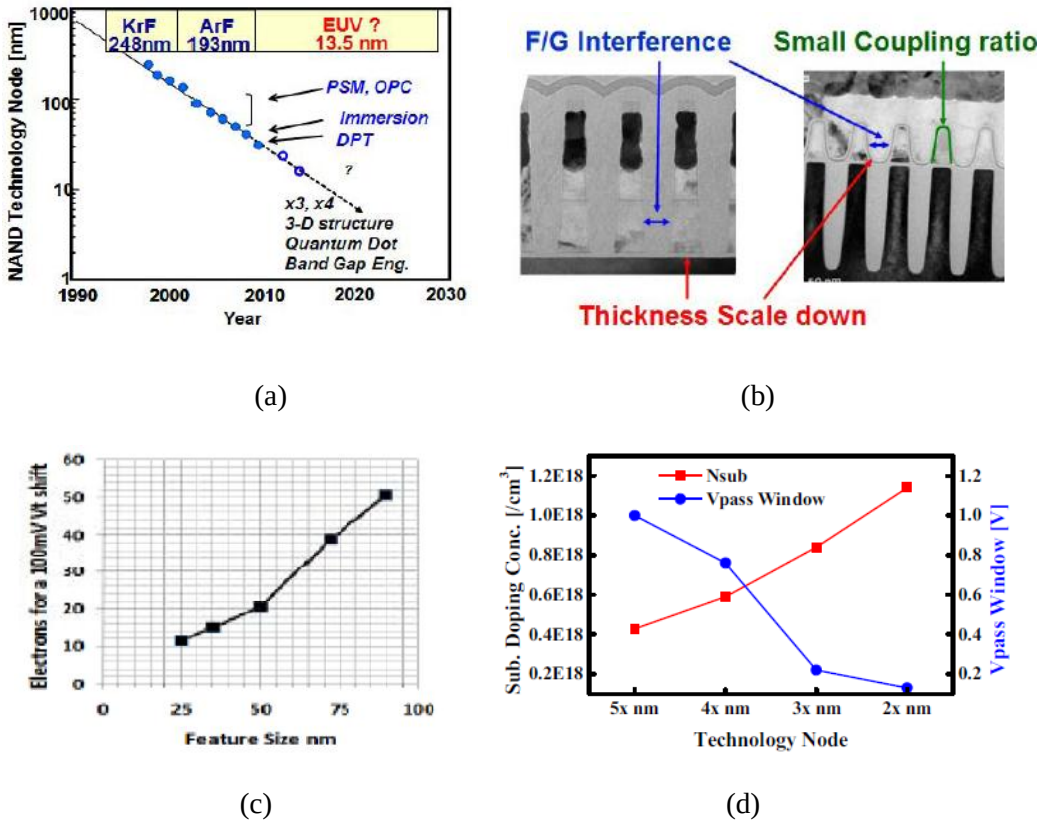


Fig. 2.1 Several obstacles in scaling down the NAND flash memory below the 20 nm channel length. (a) Limits in photolithography. (b) Electrical disadvantages due to small morphology. (c) The number of electrons in floating gate for 100 mV threshold voltage shift with the several feature size. (d) Substrate doping concentration and V_{pass} window with the technology nodes.

Such obstacles interrupt the continuous scaling down of the floating gate type

NAND flash memory. These problems are not easily solved, so other alternatives are proposed to develop high density NAND flash memory which are introduced as follows.

2.2 Multi-bit Cell Storage

Fig. 2.2 illustrates the concept of multi-level cell storage in flash memories. Conventional single-level (SLC) cell storage distinguishes between a '1' and '0' by having no charge or charge present on the floating gate of the flash memory cell. By increasing the number of charge or V_T levels, more than 1-bit per cell may be stored. Two bits per cell (MLC) storage is enabled by increasing the number of V_T levels to four representing 11, 10, 01 and 00 [13]. Similarly, by increasing the number of voltage threshold levels to 8 and 16, 3-bit per cell [14] and 4-bit per cell [15] storage is enabled. The benefit of multi-level cell storage is that storage capacity may be increased without a corresponding increase in process complexity. The same fab equipment used to manufacture silicon wafers for SLC products may be used to manufacture MLC, 3-bit per cell and 4-bit per cell devices. However, multi-level cell storage requires accurate placement of the V_t charge levels so that the charge distributions do not overlap as well as accurate sensing of the different charge levels. As the number of V_t levels increases the time it takes for accurate programming and sensing increases. Additional circuitry and programming algorithms are necessary to ameliorate the degradation of the performance and endurance of such devices. However, in effect, transition from SLC to MLC to 3-bit to 4-bit per cell

technology is equivalent to a partial shrink of the device from one process technology generation to the next without additional capital investment.

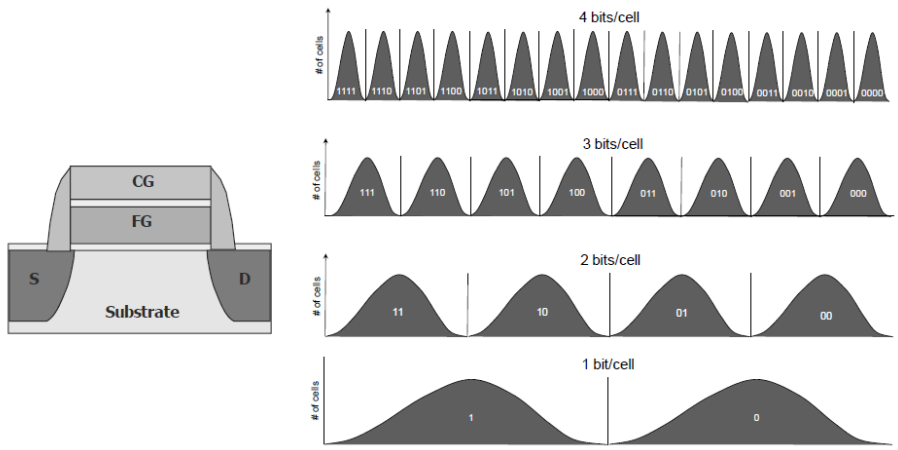


Fig. 2.2 Multi-level storage in floating gate type NAND flash memory

2.3 Charge Trap Flash Memory

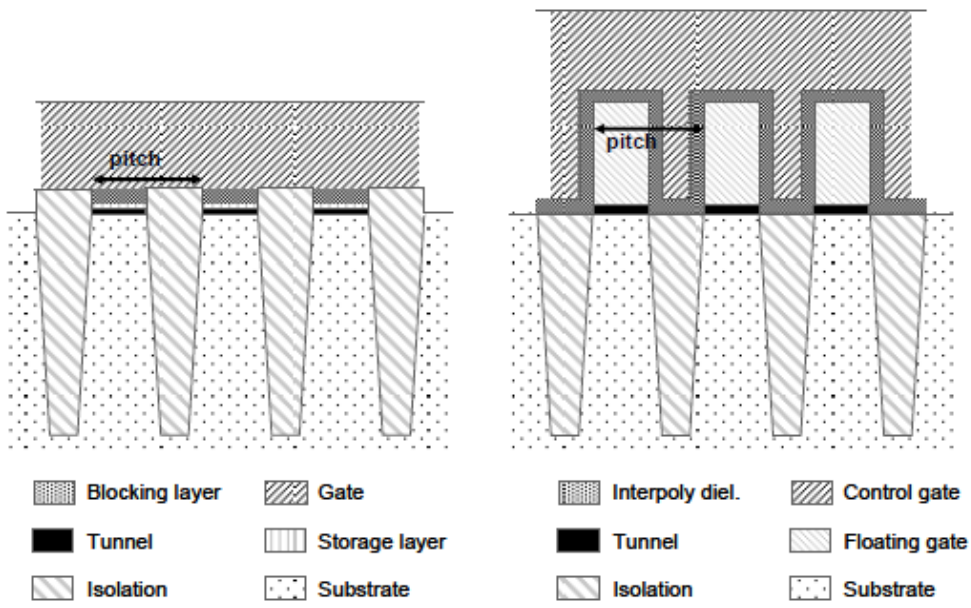
Even if FG NAND is the dominant technology and there is no advice of reduction in scaling pace, several physical roadblocks seem to limit future scalability as mentioned above. Charge trap (CT) memories may overcome some of these limitations and represent the best candidate to substitute FG devices for future nodes [16]. Differently from floating gate cells that have a semiconductor as storage element, in CT case electrons are trapped inside a dielectric layer. The different storage material changes the cell architecture drastically, impacting also on physical mechanisms for write operations (both

program and erase) and reliability.

Most important improvement of charge trap NAND flash memory compared with the FG type NAND flash is as follows: huge electrostatic interference reduction, higher immunity to active oxide defects, easier process integration. In FG NAND a lot of activity has been done on process [17,18] and algorithms to limit fringing issues, but a clear worsening of this phenomenon is observed scaling the technology node. Even if CT NAND are not completely unaffected by this problem [19], a drastic fringing improvement could be measured comparing CT versus FG at the same technology node.

The second big advantage of CT compared with FG NAND cells is the higher resistance to active dielectric defectivity both native or SILC. Stress induced leakage current (SILC) is the defectivity caused in active oxides by electrical stress. In case of a defect in tunnel or interpoly oxide, almost all the charge inside a FG cell is suddenly lost, whereas in CT NAND only the charge stored near the defect is detrapped. Superior robustness to defectivity issues allows CT NAND to greatly shrink tunnel oxide thickness without any major retention problem.

From process integration point of view, CT flow provides advantages compared with FG architecture resulting in higher cell scalability. As shown in Fig. 2.3, in FG technologies bitline pitch is equal to the sum of floating gate width, two interpoly layers and control gate coupling/shield layer width, and the scaling of all these parameters have a direct impact on cell performance and reliability. On the contrary, charge trap memories are characterized by a flat architecture that allows an easier shrink path.



(a) (b)

Fig. 2.3 Bitline crosssection of the (a) charge trap and (b) floating gate type NAND flash memory.

CT process has some beneficial effects also in wordline pitch reduction since floating gate removal simplifies wordline lithography and etch. A further opportunity to scale down wordline pitch could be obtained by the introduction of junctionless approach that has been demonstrated in CT architecture [20,21]. Due to floating gate removal, gate control of junction area is higher for CT cells. In a scaled technology node with an optimized channel doping profile engineering, the limited space between two adjacent cells could be inverted by the pass bias even without any junction and the removal of junction implant increases cell gate effective length allowing a further scaling for this

specific CT architecture.

2.4 3-Dimensional NAND Flash Memory

Even if planar charge trap memories seem a promising option to overcome scaling issues of FG products, more effective array organizations are continuously under development. 3D arrays could represent the most cost efficient solution for mass storage NAND products.

Fig. 2.4 shows the history of the 3-dimensional NAND flash memories [22-33]. In 2006, stacked NAND structure is proposed which has two NAND arrays stacked vertically, as shown in Fig. 2.5 [22]. This structure can increase the memory density as the number of stacked layers increases, but the process cost is high since many critical mask steps must be replicated for each vertical level.

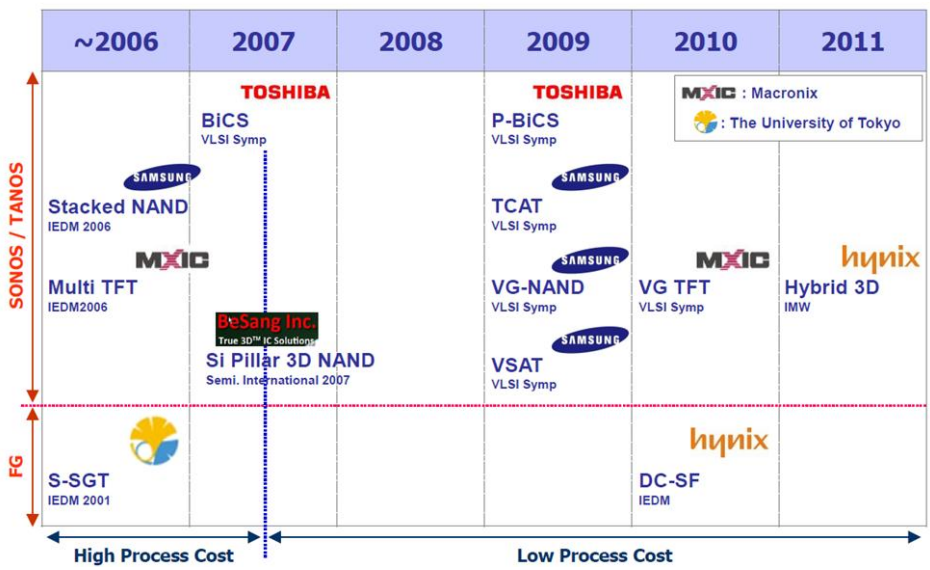


Fig. 2.4 The history of the 3D NAND flash memories.

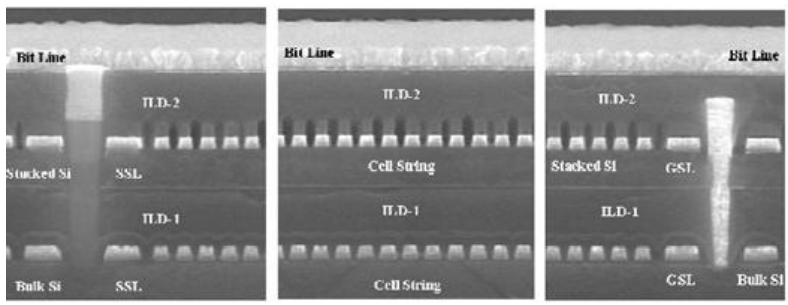


Fig. 2.5 Stacked NAND by Samsung [22].

In 2007, Toshiba Corp. proposed the bit-cost scalable (BiCS) structure as shown in Fig. 2.6 [23]. This structure has stacked gate structure and current flows in vertical direction. The fundamental feature common to these arrays is the reduced number of masks since critical masks are exposed only once to etch all vertical levels at the same time. Most research activities in 3D array development are now focused on vertical channel arrays because products realized with these technologies have a very good cost versus cell density ratio.

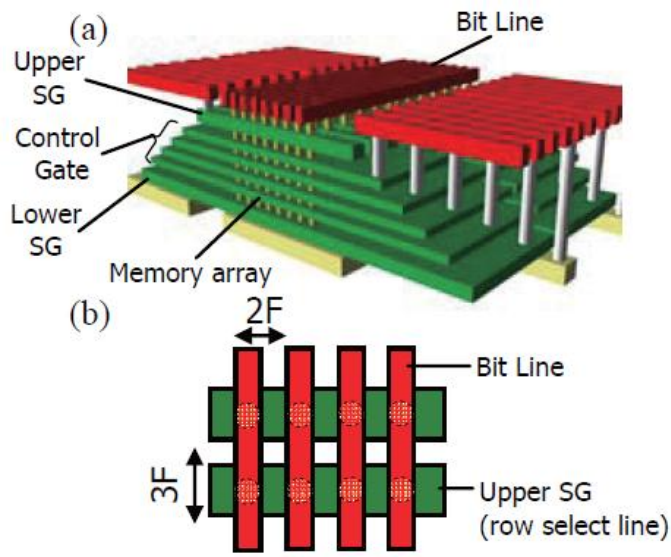
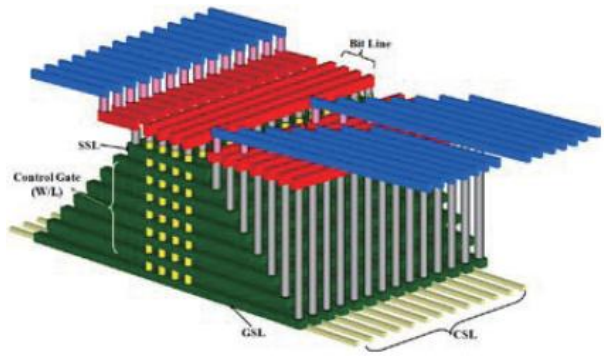


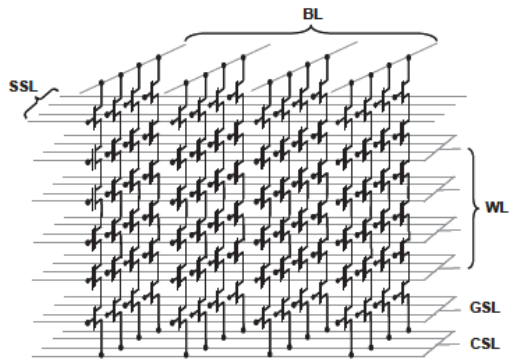
Fig. 2.6 Bit Cost Scalable (BiCS) NAND by Toshiba. (a) Bird's eye view and (b) top view.

Samsung Corp. presented the terabit cell array transistor (TCAT) in 2009 [24]. This structure has vertical directional channel that is similar to BiCS flash, but 'gate replacement' process is used to enable metal gate structure. And the channel poly plug is connected to the silicon substrate so that bulk erase is possible as conventional NAND array whereas BiCS structure use hole generation erase by GIDL current.

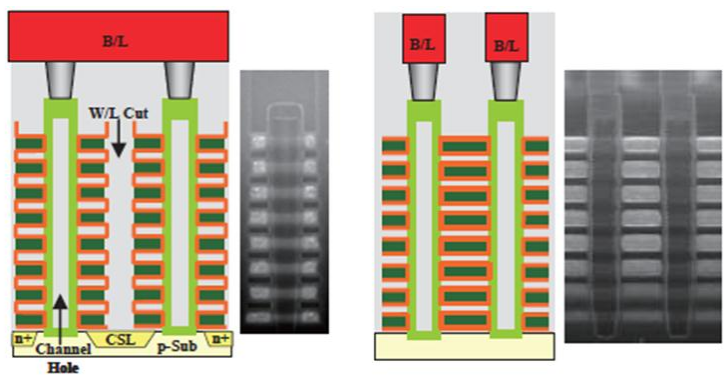
Above these, many 3D NAND structures are proposed and developed for the high density and low cost production [25-33].



(a) Bird's eye view



(b) Equivalent circuit



(c) Cross-sectional view and SEM image along the X and Y direction

Fig. 2.7 Terabit cell array transistor (TCAT) by Samsung

Chapter 3

Gated Twin-Bit (GTB) Array

In this chapter, we have proposed a gated twin-bit (GTB) array for high density NAND flash memory, and investigated its memory characteristics through the numerical simulation. This structure has a cut-off gate and two memory nodes at a single wordline, thereby 2 bits per $4F^2$ density (twice as high as the conventional NAND array) can be achieved.

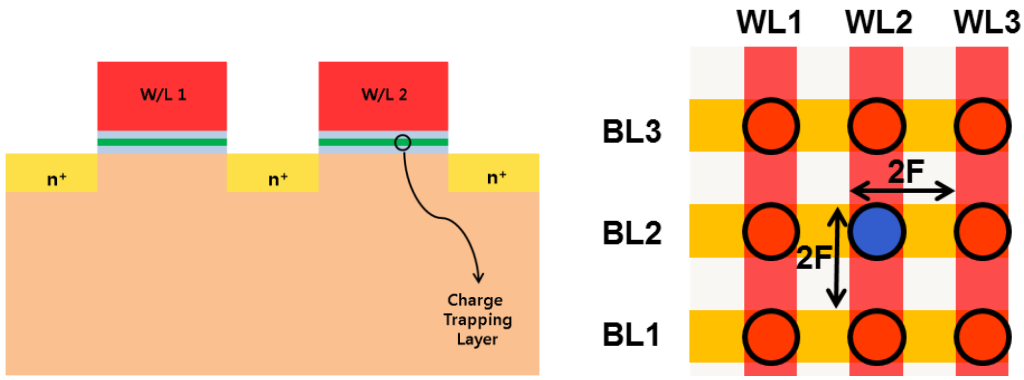
3.1 Introduction

Fig. 3.1 shows the motivation of the gated twin-bit NAND array. As shown in Fig. 3.1(a), conventional NAND array has the density of 1 bit per $4F^2$ area. To increase the density of the NAND array, folded NAND array was proposed as shown in Fig. 3.1(b), which appears to be made by applying lateral forces on the both ends of the bitline direction to form wrinkles [34]. Deep silicon trench is formed and polysilicon gate is formed as a sidewall of the trench. Since two wordlines can be formed in one trench, the number of

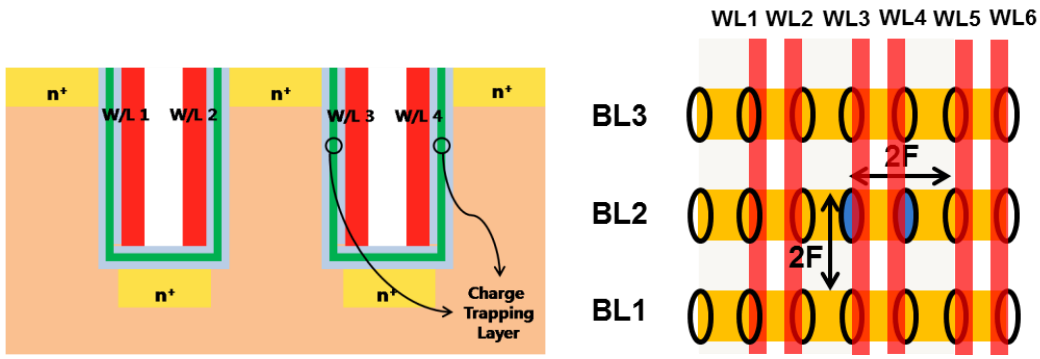
wordlines through the bitline direction is twice more than conventional NAND array. In the point of view of memory density, 2 bit can be integrated in $4F^2$ area. Thus, the folded NAND array is twice as dense as the conventional NAND array.

However, the folded NAND array requires wide trench width because two wordlines, two ONO layers, and isolating space is needed in one trench. Accordingly, there is a limit in minimizing the feature size. Wordline combined folded NAND array which has no isolation space is shown in Fig. 3.1(c). In this structure, feature size can be scaled than the folded NAND array but left and right side of the wordline cannot be controlled separately; memory density is 1 bit per $4F^2$ area.

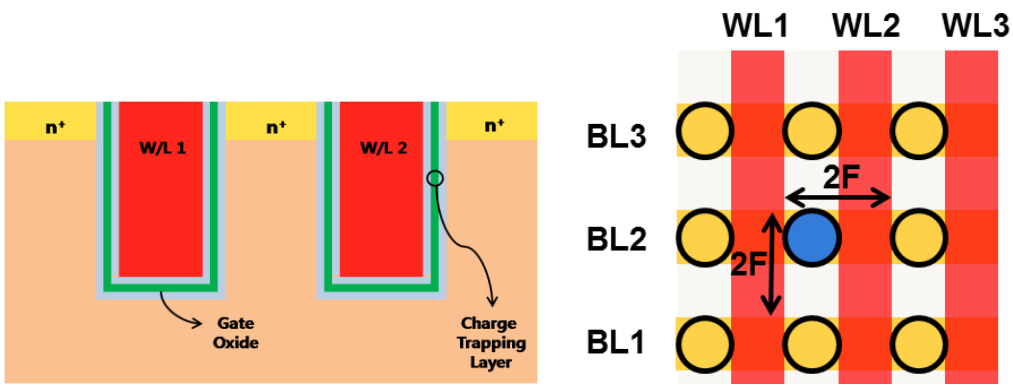
Fig. 3.1(d) shows the gated twin-bit (GTB) array which has a cut-off gate below the control gate (wordline). Cut-off gate enables two-bit operation; in other words, left and right side of the control gate can be operated separately. Since the cut-off gate exists below the control gate, it does not affect the memory density. So its memory density is same as that of folded NAND array (2 bit per $4F^2$ area), but feature size can be more scaled [35].



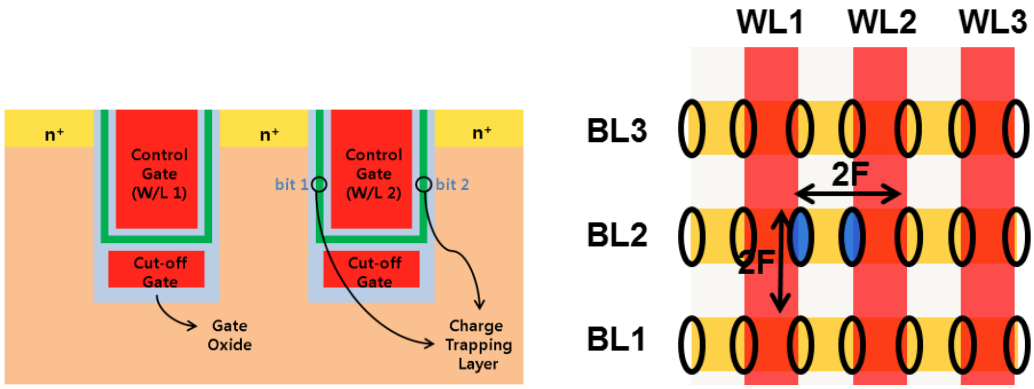
(a) Conventional NAND array



(b) Folded NAND array



(c) Wordline combined folded NAND array



(d) Gated twin-bit NAND array

Fig. 3.1 Motivation of gated twin-bit NAND array; Crosssectional view through bitline direction and top view of each array.

Fig. 3.2 shows the array schematic of conventional NAND array and GTB NAND array with the circuit symbols. In GTB array, one wordline and one cut-off gate can be represented as wordline - cut-off gate - wordline is connected in series. Since, in this series, three transistors need only 2F length of bitline direction, 128 storage nodes can be integrated in 128F length whereas 64 storage nodes exist in same length in conventional NAND array.

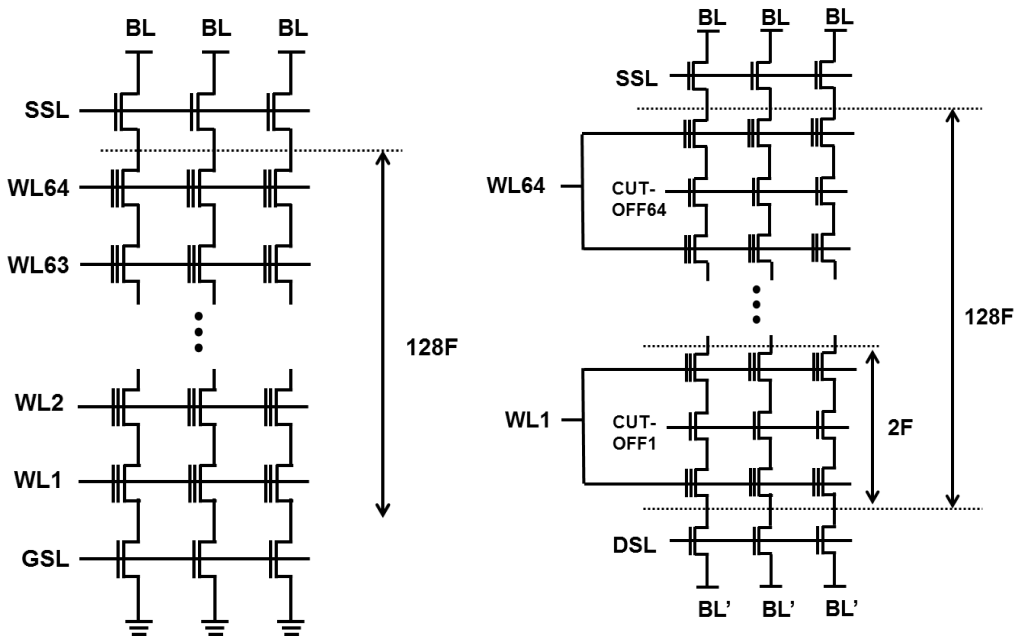


Fig. 3.2 Array schematic of (a) conventional NAND array and (b) GTB NAND array with the circuit symbols.

3.2 Operation Scheme of the GTB Array

As stated in Section 3.1, GTB NAND array has two storage nodes which share the same wordline. For the separate operation of two storage node, different schemes are required in program and read.

3.2.1 Program operation

In the case of program operation, cut-off gate plays an important role of separate

programming of two storage node which share the same control gate. Table 3.1 shows the programming bias condition of GTB array where only the left side storage node of WL2 is programmed.

Similar to the conventional NAND array, Fowler-Nordheim tunneling mechanism is used to trap the charge in the nitride layer and self-boosting scheme is used for inhibit cells. High program voltage is applied to the selected (want to program) wordline, and bitline voltage is set to 0 V to make the channel potential 0 V. However, if both the left side channel and right side channel of the wordline is set to 0 V, FN tunneling occurs at both side so that separate programming is impossible.

As indicated in Table 3.1 and Fig 3.3, different voltage is applied to bitline of each side. In the case of the left side of WL2 is programmed, 0 V is applied to the left side bitline and V_{DD} is applied to the right side bitline. And low voltage is applied to the cut-off gate so that the left side and right side channel potential is not transferred. In this condition, 0 V of left side bitline is transferred to the left side channel of WL2, while right side channel potential is boosted because V_{DD} is applied to the right side bitline and right side select gate. If the channel is cut under the cut-off gate, FN tunneling occurs only in left side and electrons are trapped at the left-side nitride layer. Since the channel potential of right side is boosted, FN tunneling cannot occur and the right side cell is inhibited.

Table 3.1 Programming bias condition of GTB array (only left side of WL2 is programmed).

CONTROL LINE	PROGRAM (WL2 → 01 state)
WL1(unselected)	HIGH
WL2(selected)	$V_{PROGRAM}$
WL3(unselected)	HIGH
Cufoff-G1	HIGH
Cufoff-G2	LOW
Cufoff-G3	HIGH
SSL	V_{DD}
DSL	V_{DD}
BL1(selected)	GND
BL1‘ (selected)	V_{DD}
BL2(unselected)	V_{DD}
BL2‘(unselected)	V_{DD}
Substrate	GND

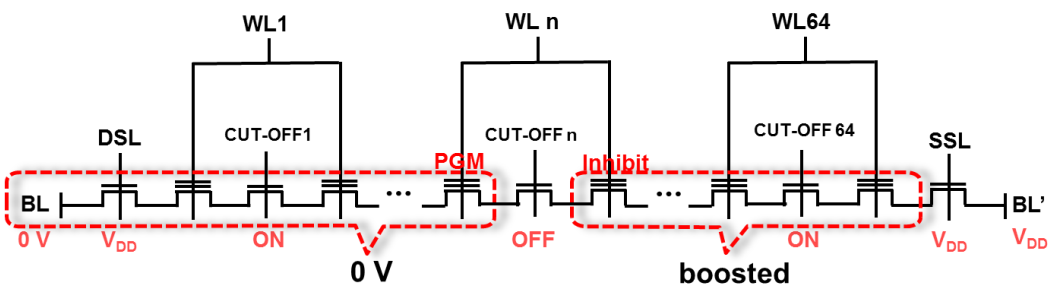
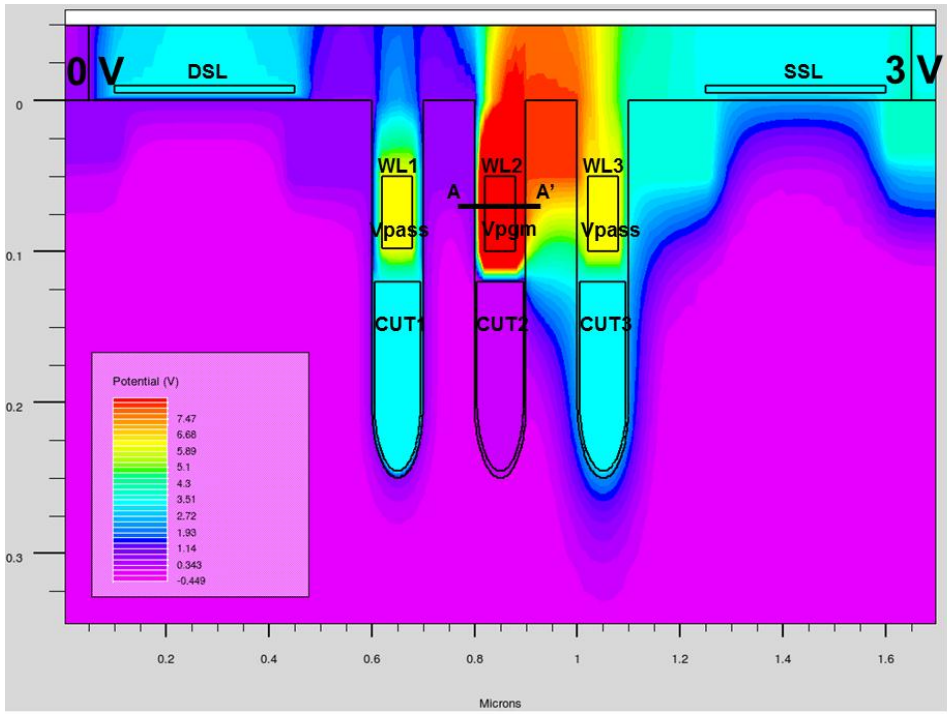
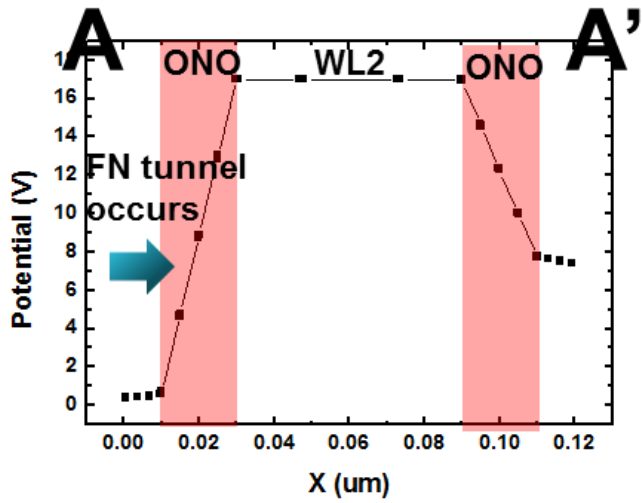


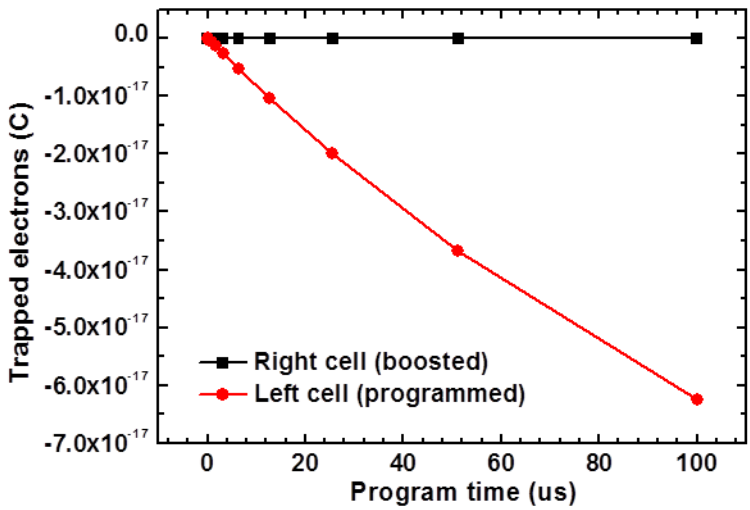
Fig. 3.3 Channel potential of programming case.


0 V
DSL
SSL
3 V
WL1
WL2
WL3
A
A'
Vpass
Vpgm
Vpass
CUT1
CUT2
CUT3
Potential (V)
7.47
6.68
5.89
5.1
4.3
3.51
2.72
1.93
1.14
0.343
-0.449
0
0.1
0.2
0.3
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
Microns

(a)


A
A'
ONO
WL2
ONO
FN tunnel
occurs
Potential (V)
18
16
14
12
10
8
6
4
2
0
0.00
0.02
0.04
0.06
0.08
0.10
0.12
X (um)

(b)



(c)

Fig. 3.4 (a) Potential distribution and (b) electric field along the A-A' cutline in the programming case. (c) Injected charge in each side of storage node.

Program operation scheme is validated by device simulation. Simulated GTB array has 60-nm long storage node and control gates, and 301-nm long cut-off gate (sum of two vertical direction 80 nm and half-circle of 141 nm), 50-nm deep n-type S/D regions. The thickness of tunneling oxide, charge trap nitride layer, blocking oxide is 5/5/10 nm, respectively. Program voltage is 15 V, and 2 V is applied to the cut-off gate to connect channel, and 0 V not to connect. 3 V is applied to the select gates and inhibit bitline (right side).

Fig. 3.4(a) shows the potential distribution of the programming case. Left-side

channel potential is 0 V, while right side potential is boosted enough to prevent programming as shown in Fig. 3.4(b). As can be seen in Fig. 3.4(c), large amount of electrons are trapped in the left side storage node than the right-side storage node.

In terms of view of leakage current of boosted channel, GTB array has great advantage compared with the conventional planar NAND array. Because the effective channel length of the cut-off gate is relatively long, the leakage current is thoroughly suppressed so that the boosted channel potential is well maintained.

3.2.2 Read Operation (Forward-reverse Read Scheme)

To read the left side and right side storage node of the same wordline, forward-reverse read scheme must be used [36]. This scheme was used to read the NROM device [37-43], which is one of NOR flash memories. The trapped electrons affect the threshold voltage (V_T) in the bidirectional reads. If electrons are trapped at the source side cell, it causes a larger V_T shift than the drain side cell.

Fig. 3.5 shows the read direction and reading storage node of GTB array. Since the source side trapped charge affects the current more than the drain side, 0 V is applied to the bitline where the want-to-read cell exists, and V_{DD} is applied to the opposite side bitline. As shown in Fig. 3.6, if we want to read the right-side cell, V_{DD} is applied to the left-side bitline and 0 V is applied to the right-side bitline. If we want to read the left-side cell, the polarity of the applied voltage should be swapped.

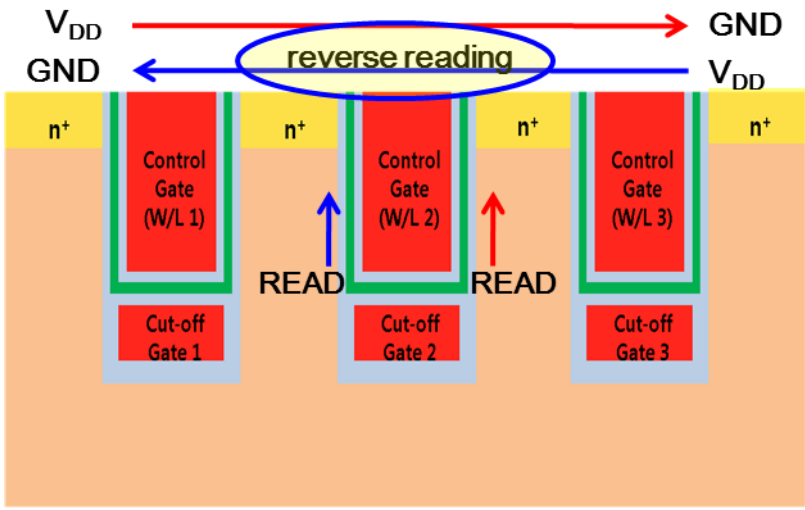


Fig. 3.5 Two bit read direction of reading storage node.

Table 3.2 Read bias condition of GTB array (WL2).

CONTROL LINE	READ(WL2)	
WL1(unselected)	HIGH	
WL2(selected)	0 V	
WL3(unselected)	HIGH	
Cufoff-G1	HIGH	
Cufoff-G2	HIGH	
Cufoff-G3	HIGH	
SSL	HIGH	
L side of BL1(selected)	GND	V_{DD}
R side of BL1(selected)	V_{DD}	GND
L side BL2(unselected)	GND	
R side BL2(unselected)	GND	
Substrate	GND	

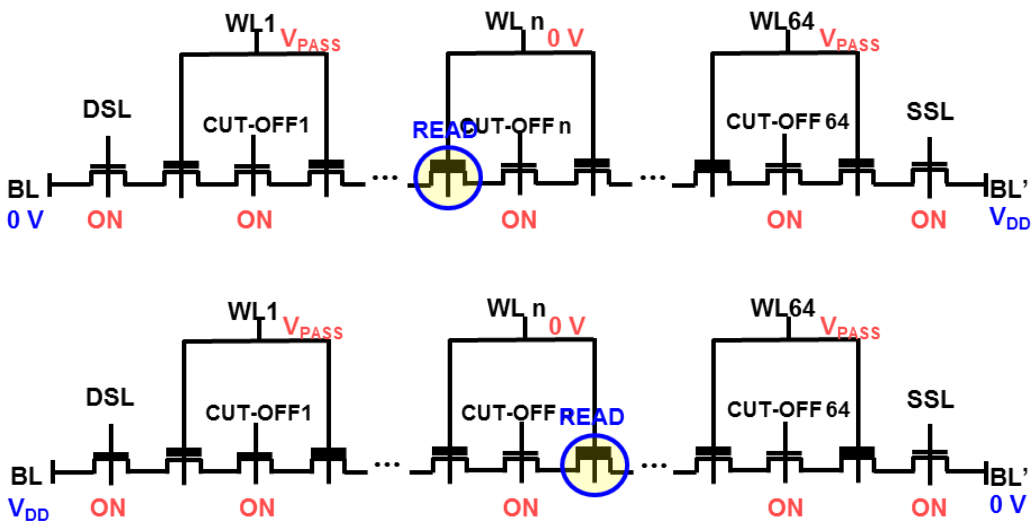


Fig. 3.6 Applied voltages to the contacts in the forward-reverse read operation.

To verify the forward-reverse read scheme, device simulation is carried out with one wordline and one cut-off gate. Fig. 3.7 shows the bias condition during the read operation for 4 possible states. Since one wordline has two storage nodes, 4 states can exist (P/P, P/E, E/P, E/E). I-V curve is achieved with the reverse read bias condition; current flows from the right side to the left side. The drain voltage is 2 V, and the length of the storage nodes is 40 nm.

10 and 11 states have low threshold voltage, whereas 00 and 01 state have high threshold voltage as shown in Fig. 3.8. That means V_T is mainly determined by the state of the left side node, and the right side node scarcely affects V_T.

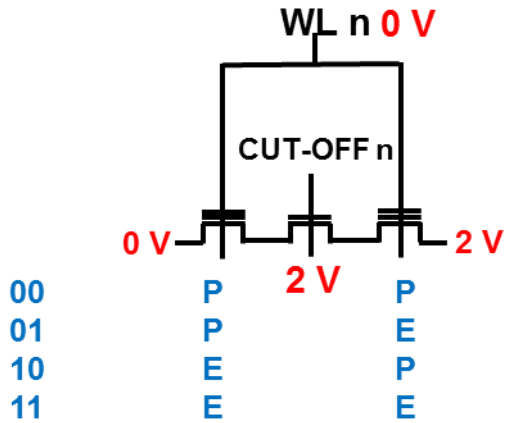


Fig. 3.7 Reverse read bias condition during the read operation for 4 possible states.

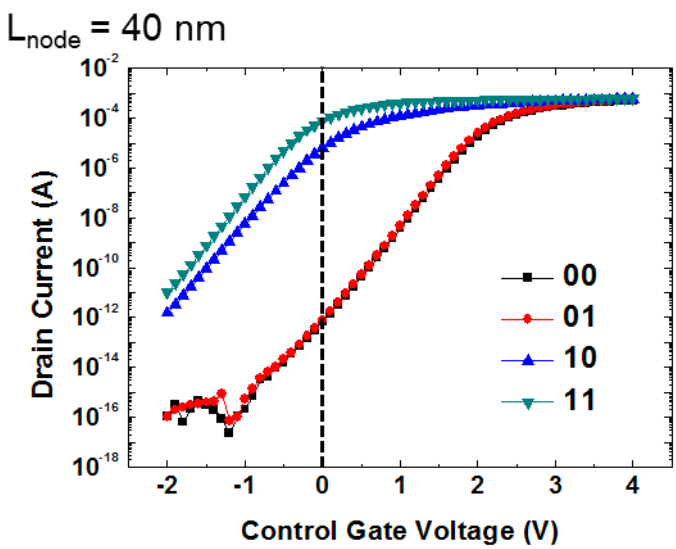


Fig. 3.8 I-V characteristics for 4 possible states in reverse read condition.

To investigate the reason of the phenomenon, the energy band diagram along the channel is analyzed. As shown in Fig. 3.9(a) and (b), electrons which move from the left

side to the right side meet high energy barrier at node L if the left side storage node is in programmed state. Therefore, if the left side node is programmed, V_T is high irrespective of the state of the right side node; whether the energy barrier is low or high at node R.

If the left side node is erased state, energy barrier does not exist at the node L. In the case of 10 state, energy barrier is not seen at the node R even though the right side node is programmed. It can be explained by barrier lowering caused by the large potential difference between both ends of the node R. Consequently, V_T is low even though the right side cell is programmed.

In the case of 11 state, all the nodes are erased so that there is no energy barrier through whole channel and V_T is low.

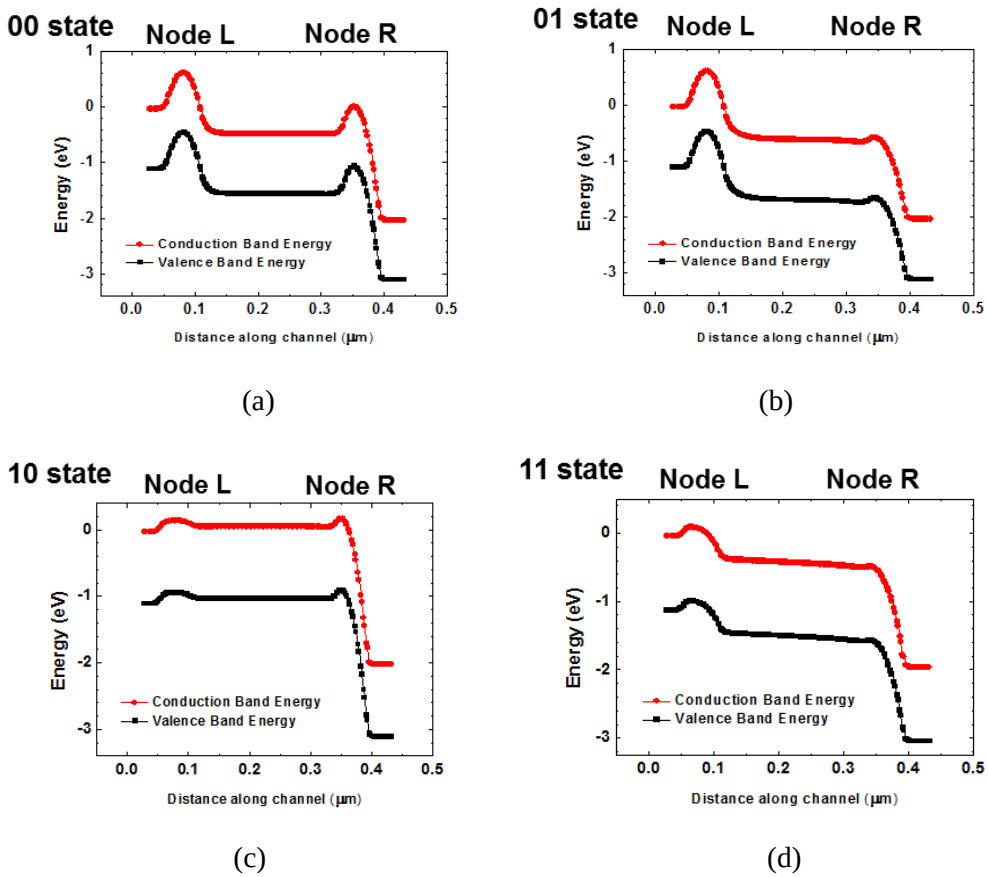
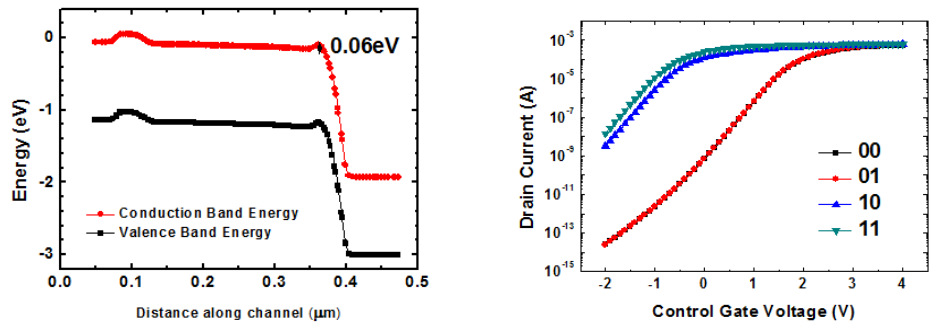


Fig. 3.9 Energy band diagram through the channel direction. (a) 00 state (b) 01 state (c) 10 state (d) 11 state.

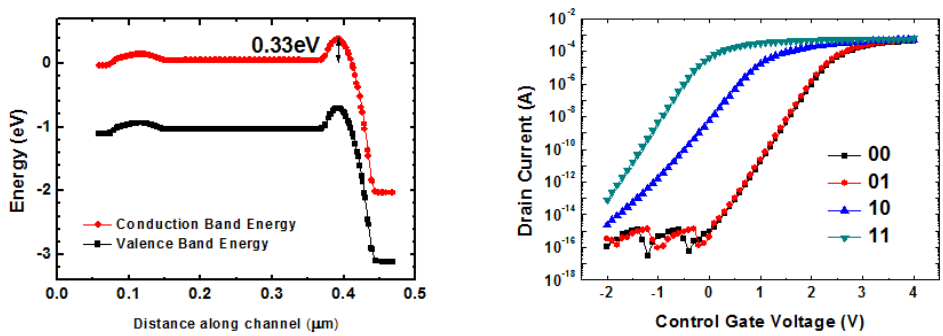
3.2.3 Barrier Lowering in Read Operation

To investigate further into barrier lowering during the read operation, several simulations are conducted with varying the parameters which are related to the barrier lowering. To investigate the effect of the node R, simulations are carried out in 10 state; only node R is programmed state which must be ignored.

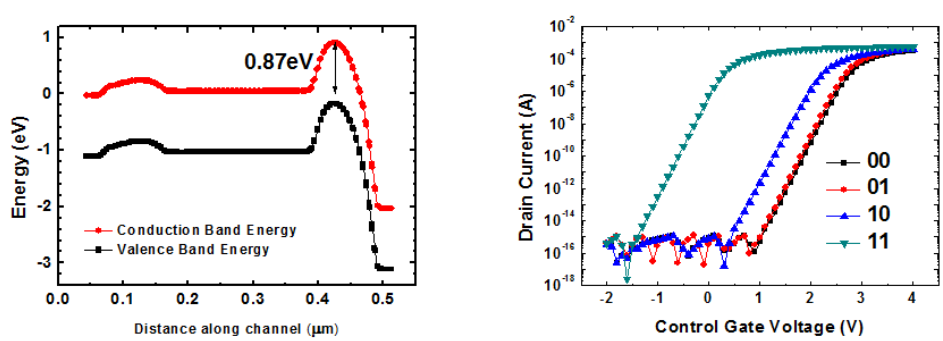
Although node R is programmed, energy barrier is 0.06 eV when the channel length is short enough as shown in Fig. 3.10(a). So the threshold voltage of 10 and 11 states are not significantly different. However, the longer the length of the storage node is, the higher the energy barrier becomes. And then the V_T difference of 10 and 11 state becomes larger which must be nearly the same for accurate two-bit read scheme. If the length of the storage node is too long, the V_T of 10 state approaches to that of 00 and 01 state as shown in Fig. 3.10(c).



(a) L=30 nm



(b) L=50 nm

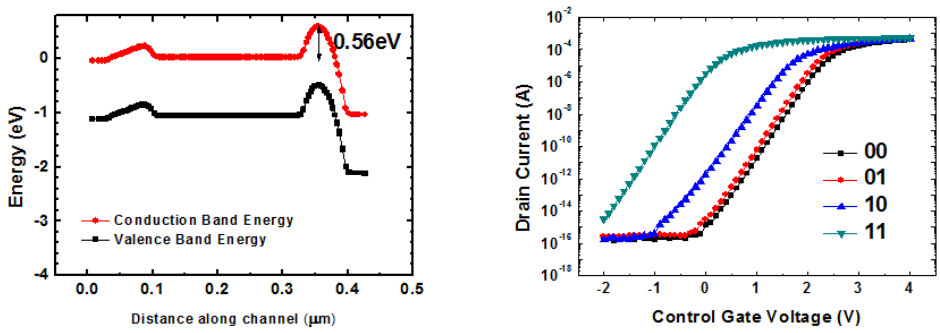


(c) L=80 nm

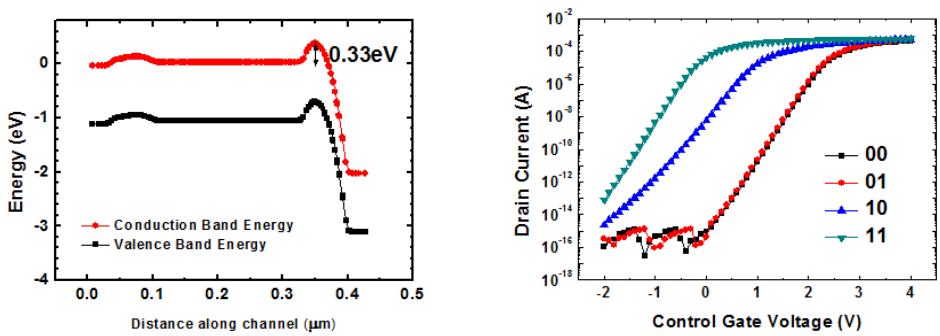
Fig. 3.10 Energy band diagram and I-V characteristic curve of 10 state with the different length of the storage node (a) 30 nm (b) 50 nm (c) 80 nm.

Next, simulations are carried out with various bitline voltages. When the bitline voltage is 1 V, the energy barrier is so high (0.56 eV) that it impedes the electron movement from the left to the right as shown in Fig. 3.11(a); V_T of 10 state is so close to that of 00 and 01 state, namely an error occurs in read operation. However, if the bitline voltage increases to 3 V as shown in Fig. 3.11(c), the energy barrier is only 0.26 eV so that V_T of 10 state is nearly the same as that of 11 state.

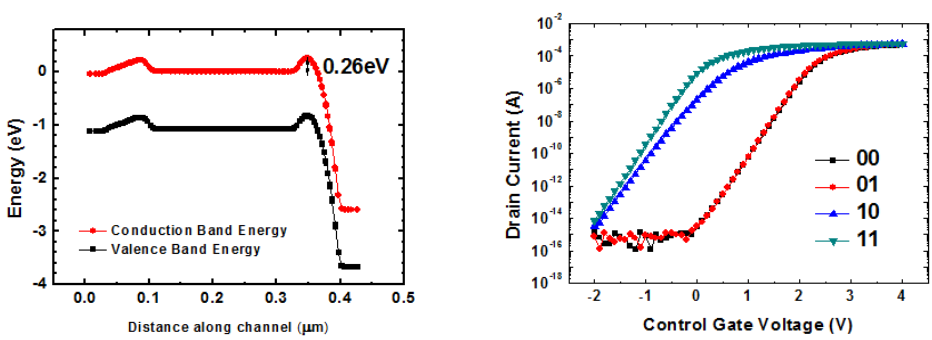
This dependence in bitline voltage is also related to the barrier lowering. As shown in Fig.3.11, the source voltage of 0 V is well transferred to the left side of the node R. So the voltage difference between the both ends of the node R is V_{BL}-$V_S \approx V_{BL}$. Therefore, as the bitline voltage increases, the barrier at the node R is lowered; trapped electrons at the node R cannot affect the V_T.



(a) V_{BL}=1 V



(b) V_{BL}=2 V



(c) V_{BL}=3 V

Fig. 3.11 Energy band diagram and I-V characteristic curve of 10 state with the different bitline voltage (a) 1 V (b) 2 V (c) 3 V.

3.2.4 Erase Operation

Erase operation in the GTB array is the same as that of the conventional NAND array. Table 3.3 shows the erase bias condition of the GTB array. All the wordlines are set to 0 V, and a high erase voltage is applied to the Si substrate to induce Fowler-Nordheim tunneling so that block erase is performed. All the cut-off gates and bitline voltage is floated.

Table 3.3 Erase bias condition of the GTB array.

CONTROL LINE	ERASE
WL1	0 V
WL2	0 V
WL3	0 V
Cufoff-G1	Floating
Cufoff-G2	Floating
Cufoff-G3	Floating
SSL	Floating
DSL	Floating
BL1	Floating
BL1‘	Floating
Substrate	V_{ERASE}

3.3 Fabrication Process of the GTB Device

Fig. 3.12 depicts the critical fabrication process of the GTB device. Key processes are summarized as follows.

Fig. 3.12(a): Active silicon fin patterning. To form the thin silicon fin, e-beam lithography and photo lithography mix-and-match technique is used.

Fig. 3.12(b): Shallow trench isolation (STI) process. Different from conventional NAND STI process, oxide is remaining to play a role of hard mask at the following silicon trench etch step.

Fig. 3.12(c): Oxide trench etching. The space for the gate polysilicon connection is formed. For the following deposition and etch-back process, the trench width must be narrow enough.

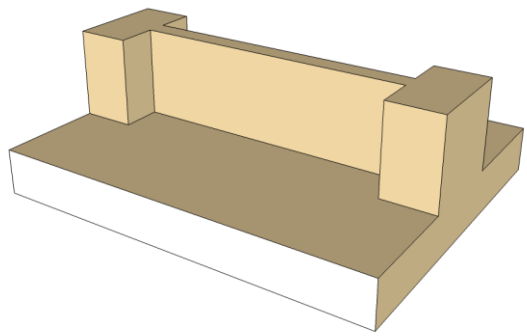
Fig. 3.12(d): Silicon trench formation.

Fig. 3.12(e): Gate oxidation and cut-off gate formation. To form a buried cut-off gate structure, n^+ doped polysilicon is deposited by low pressure chemical vapor deposition (LPCVD). If the width of the trench is narrower than the thickness of deposited polysilicon, the silicon trench is completely filled with polysilicon and the surface becomes planarized so that planarization process is not necessary. After that, etch back process using dry etch method is carried out.

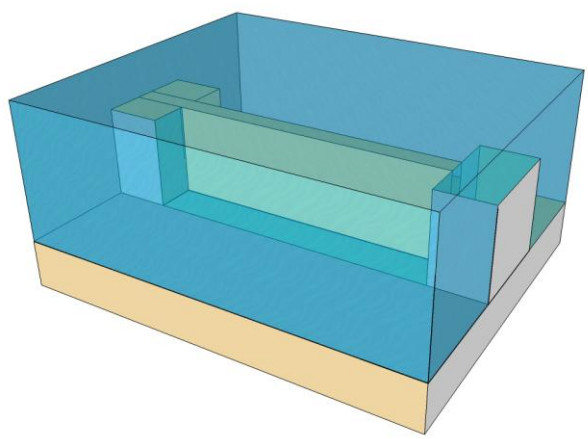
Fig. 3.12(f): Charge trapping layer deposition. First, the cut-off gate oxide is removed by isotropic etching process. Next, tunneling oxide/nitride/blocking oxide layers are deposited one after another.

Fig. 3.12(g): Control gate formation. Same as the cut-off gate formation, n^+ doped polysilicon is deposited by LPCVD and etched back by dry etch process.

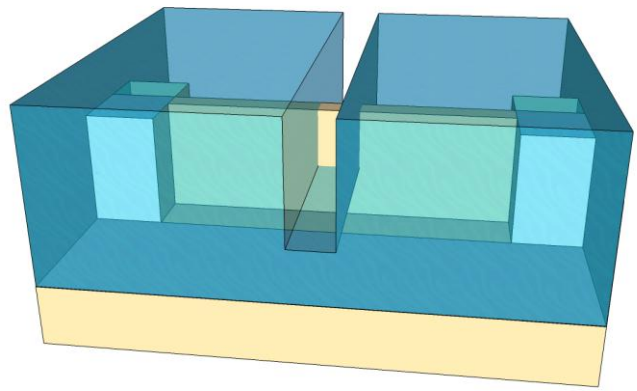
Fig. 3.12(h): Hard mask removal and S/D implantation.



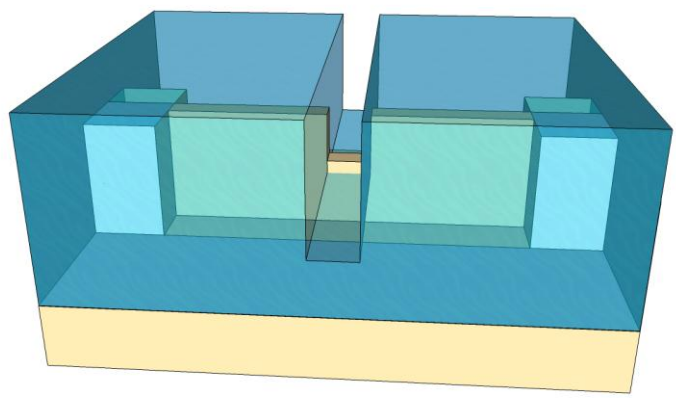
(a) Active silicon fin patterning.



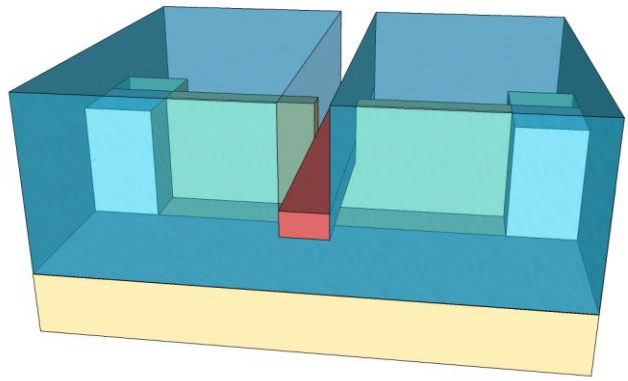
(b) Shallow trench isolation (STI) process.



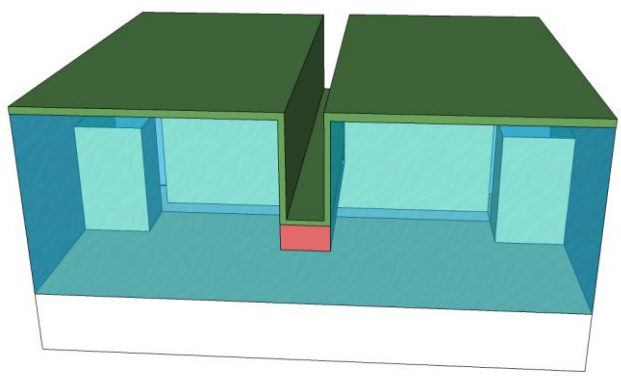
(c) Oxide trench etching.



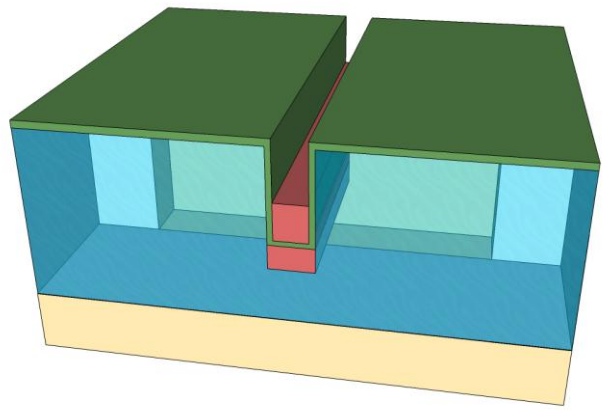
(d) Silicon trench formation.



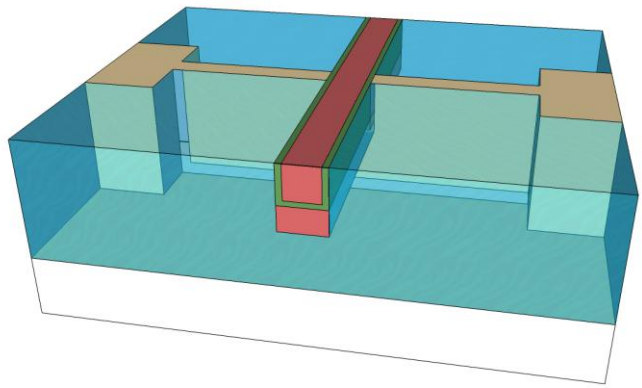
(e) Gate oxidation and cut-off gate formation.



(f) Charge trapping layer deposition.



(g) Control gate formation.



(h) Hard mask removal and S/D implantation.

Fig. 3.12 Key fabrication processes of GTB device.

3.4 Issues in GTB Array

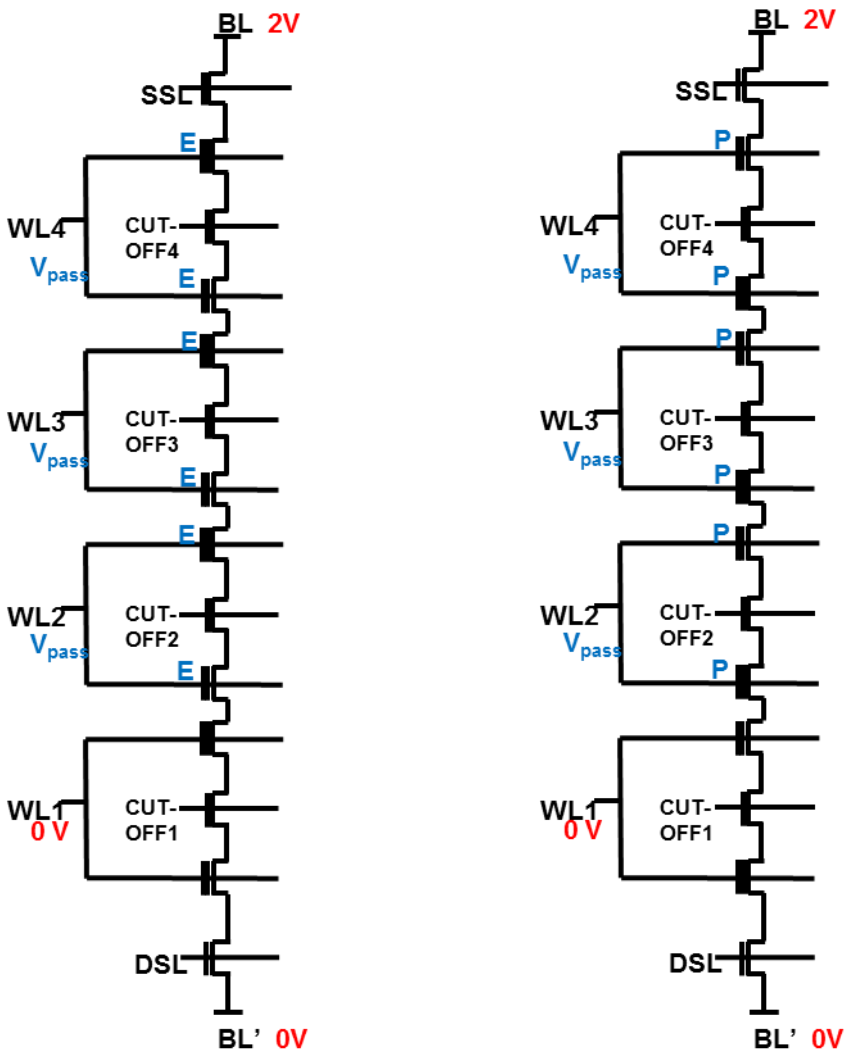
3.4.1 Voltage Drop in GTB Array

As indicated in Section 3.2.3, the barrier lowering at the drain side is very important for the accurate forward-reverse read scheme. Unintended decrease of the voltage can induce a read error. Since only a single device is simulated in Section 3.2.3, GTB array is simulated in this section to investigate the possibility of voltage drop when GTB devices are connected in series to comprise a NAND array.

In a NAND array, 64 storage cells, 2 dummy cells, and select gates are connected in series. If the channel resistance is not small, voltage drop occurs at the pass transistors (unselected cells). Unwanted voltage drop at the unselected cells can decrease the voltage that should be applied to the selected cell, then bring about read error.

This bitline voltage drop becomes the more serious as the selected cell is far from the bitline contacts. To transfer the bitline voltage to the selected cell, it must pass through all the unselected cells. The overdrive voltage (V_{GD}) is smaller in the unselected cells between the bitline and selected cell than that of the cells between the selected cell and ground (V_{GS}); that is, bitline voltage drop becomes larger as the number of unselected cells between the selected cell and bitline increases.

Fig. 3.13 shows the best case and the worst case of voltage drop in GTB array with the state of the unselected cells. As shown in Fig. 3.13(a), the voltage drop is minimized when all the unselected cells are in erased state. However, the voltage drop goes maximum value if all the unselected cells are in programmed state.



(a) Best case (b) Worst case

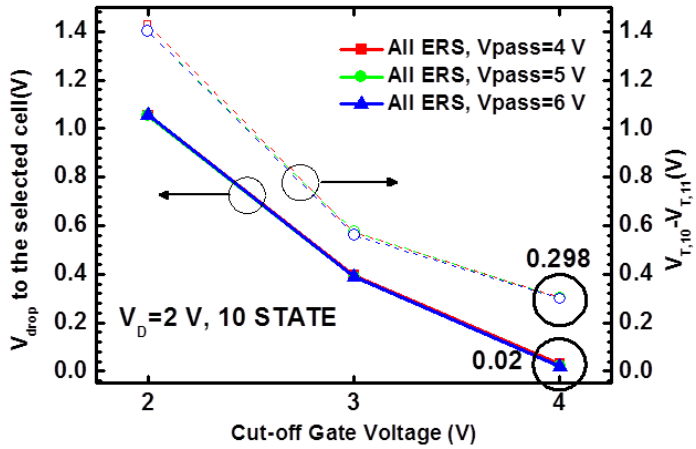
Fig. 3.13 Best and worst case with the state of unselected cells in the point of the view of voltage drop.

If the pass voltage of the unselected cell's control gate voltage is not high enough, unintended voltage drop becomes so serious. Or, low cut-off gate voltage of the

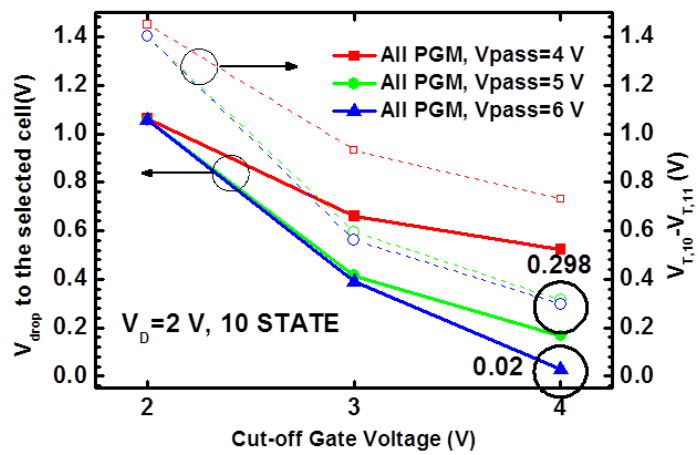
unselected cell can also induce severe voltage drop. The TCAD simulation is carried out to verify how much the bitline voltage drops while transferred to the selected cell with the cases indicated in Fig. 3.13.

Fig. 3.14 shows the results with various applied voltage to the unselected cut-off gate and pass voltage to the unselected wordlines. If all the unselected cells are in erased state, the amount of the voltage drop between the bitline and selected cell (WL1) does not vary with the pass voltage applied to the unselected wordlines as shown in Fig. 3.14(a). Voltage drop to the selected cell is 0.02 V, caused 0.298 V of threshold voltage difference between 10 and 11 state. However, if all the unselected wordlines are in programmed state, low pass voltage brings about high voltage drop because the low electron concentration at the channel makes high channel resistance. If the pass voltage is 4 V, the voltage drop is 0.522 V and it causes 0.731 V of V_T difference, even if cut-off gate voltage is high enough (4 V).

And, low cut-off gate voltage makes it difficult to transfer the bitline voltage to the selected cell. If the cut-off gate is 2 V, the voltage drop to the selected cell is approximately 1 V which is almost half of the bitline voltage that causes 1.4 V of V_T difference. As a result, the cut-off gate voltage and pass voltage applied to the unselected cells must be high to reduce the voltage drop and V_T difference of 11 and 10 state.



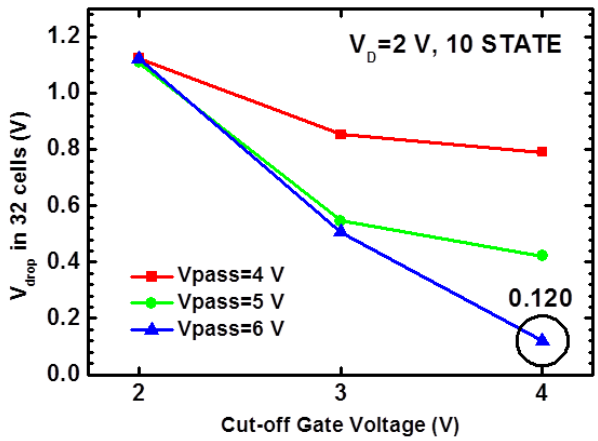
(a)



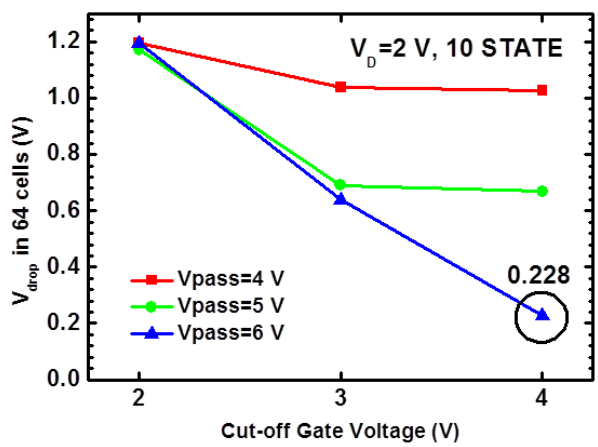
(b)

Fig. 3.14 Voltage drop to the selected cell with various applied voltage to the unselected cut-off gate and pass voltage to the unselected wordlines. (a) Best case: all the unselected cells are in erased state and (b) worst case: all the unselected cells are in programmed state. The channel length of each cell is 40 nm.

With these simulated results, the voltage drop in 64 cell array and 32 cell array is estimated. The whole channel resistance increases linearly as the number of the cells increase. As shown in Fig. 3.15, the voltage drop increases as the length of the array becomes longer. The voltage drop to the selected cell is 0.120 V in 32 cell array, and 0.228 V in 64 cell array.



(a)



(b)

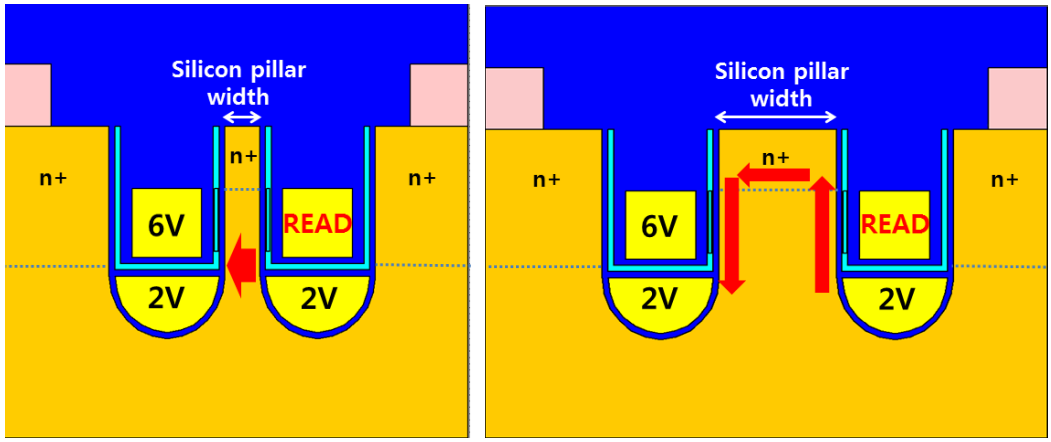
Fig. 3.15 Bitline voltage drop to the selected cell in array with (a) 32 cells (b) 64 cells.

3.4.2 Leakage Current between the Cut-off Gates

Different from the conventional NAND array, current flows up and down through the bitline direction in GTB array. Therefore, there is a danger of punch through in the lateral direction when the silicon pillar thickness is too thin. As shown in Fig. 3.16(a), if the silicon pillar thickness is not thick enough, leakage current can flow directly from the cut-off gate area to the neighboring cut-off gate area or from control gate area to the neighboring control gate area.

Simulation is carried out for the two worst case conditions; one is the reading cell's left side is programmed and neighboring unselected cell's right side is erased (10 state), another is the reading cell's left side is erased and neighboring unselected cell's right side is also erased (11 state). The current ratio of these two conditions must be large enough for the accurate read operation. The reason why these are the worst case is that the erased neighboring cell can induce the most electrons in thin silicon pillar, bringing about punch-through leakage current. In this simulation, 6 V of pass voltage is applied to the unselected wordline and 2 V is applied to the cut-off gates.

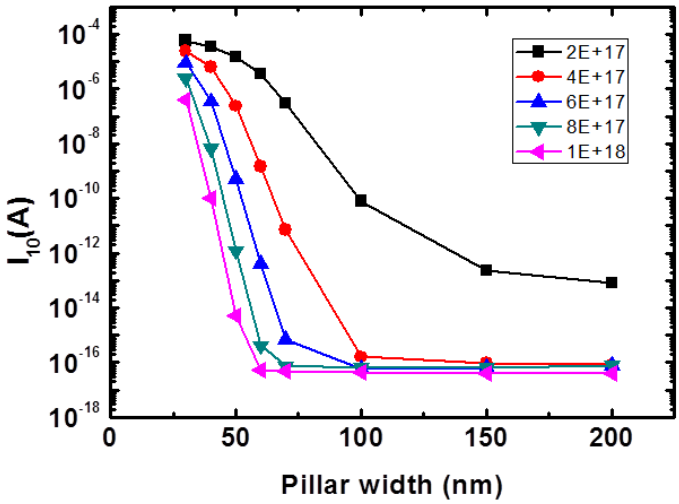
As shown in Fig. 3.17(a), the current of 10 state increases steeply when the silicon pillar thickness is thinner than 70 nm. Leakage current can be suppressed with the high channel doping concentration, but there is limitation when the silicon pillar thickness is near 30 nm range.



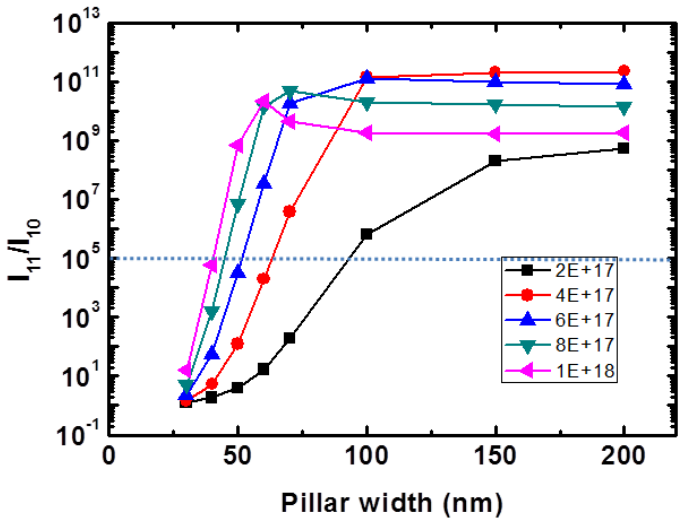
(a) 30 nm (b) 100 nm

Fig. 3.16 Current flow direction with the silicon pillar thickness of (a) 30 nm (b) 100 nm.

Current ratio of 11 state and 10 state with different silicon pillar width and channel doping concentration is shown in Fig. 3.17(b). The ratio means the read margin of P/E states. If the required standard for current ratio is established to 10^5, silicon pillar thickness must be thicker than 50 nm with the channel doping concentration of $6\times10^{17}/cm^3$. The P/E read margin can be increased as the channel doping concentration increases, but high channel doping concentration decreases the current in erased state and slows down the reading speed.



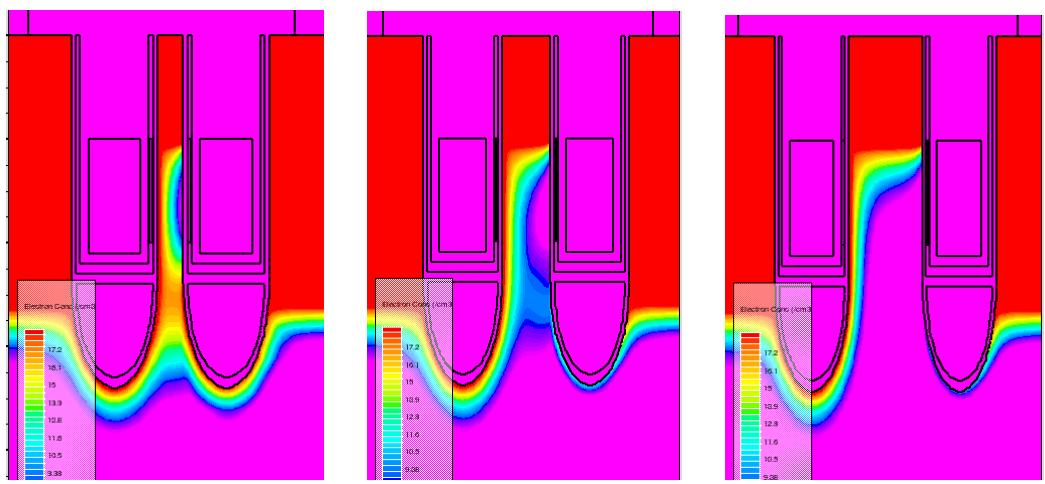
(a)



(b)

Fig. 3.17 (a) Flowing current of 10 state (b) ratio of 11 state current and 10 state current with different silicon pillar width and channel doping concentration.

Fig. 3.18 shows the electron concentration distribution during the read operation of 10 state with the different silicon pillar thickness. Because the reading cell is programmed state and read voltage is 0 V following the NAND array read scheme, electron concentration at the selected cell region is very low. The main leakage path (highest electron concentration that is punching through the silicon pillar) can be seen between the cut-off gates. To conclude, the main punch-through leakage current path goes from the cut-off gate area to the neighboring cut-off gate area.



(a) 30 nm (b) 60 nm (c) 100 nm

Fig. 3.18 Electron concentration distribution during the read operation of 10 state with the silicon pillar thickness of (a) 30 nm (b) 60 nm (c) 100 nm.

Meanwhile, when the fabrication process is considered, the silicon trench cannot be etched with 90° slope as shown in Fig. 3.16. As the slope becomes gradual, the distance

between the neighboring cut-off gates increases so that the punch-through leakage current reduces.

To reduce the punch-through current between the cut-off gates, downscaled cut-off gate structure is proposed. Fig. 3.19 shows the fabrication process of the structure. First, (a) silicon trench is formed by dry etch process same as normal GTB structure process. And (b) oxide sidewall is formed by deposition and etch process. The thickness of this sidewall equals to the decreased radius of the cut-off gate. And (c) silicon is etched for the space for cut-off gate, and (d) cut-off gate oxide is formed. After that, cut-off gate deposition and etch back process - oxide strip - ONO layer deposition - control gate deposition and etch back process is performed as in the case of the normal GTB structure.

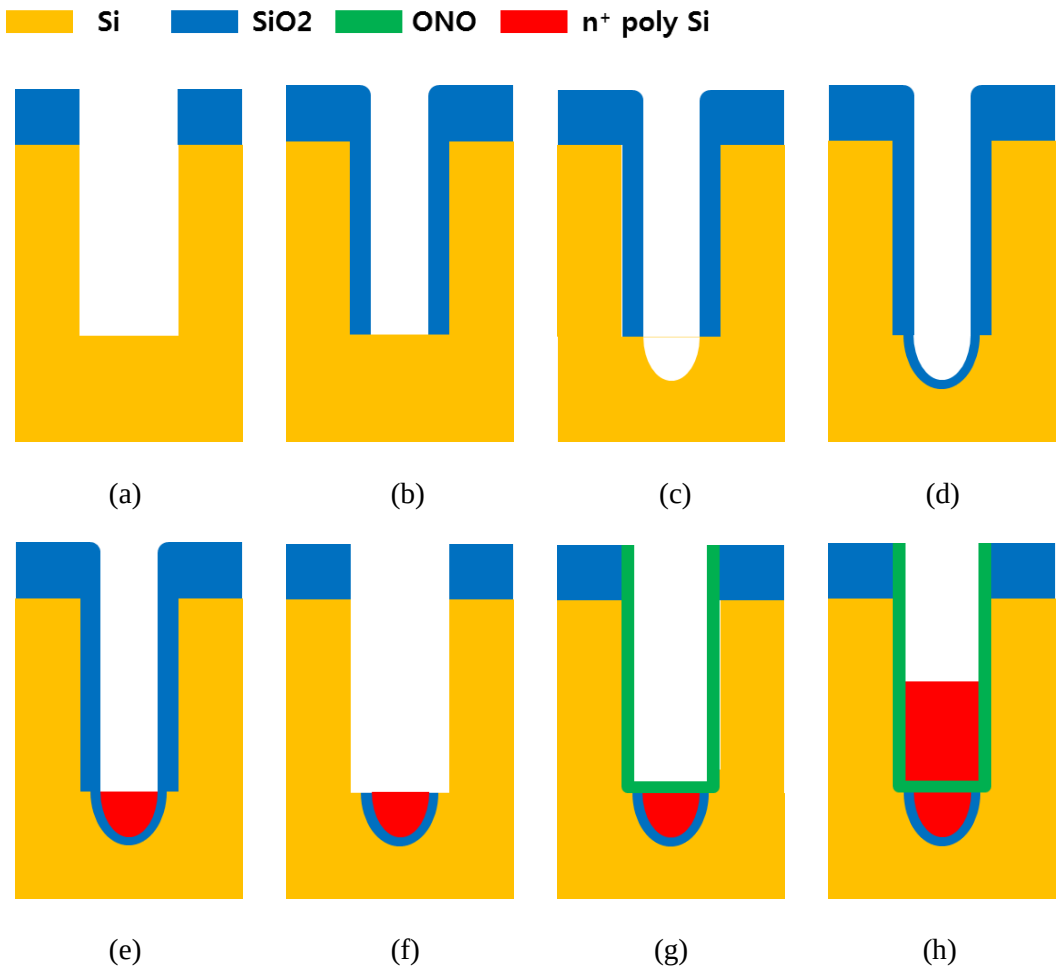
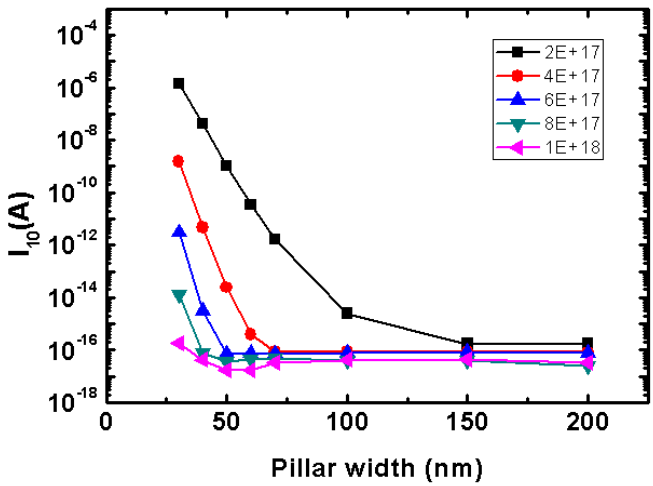


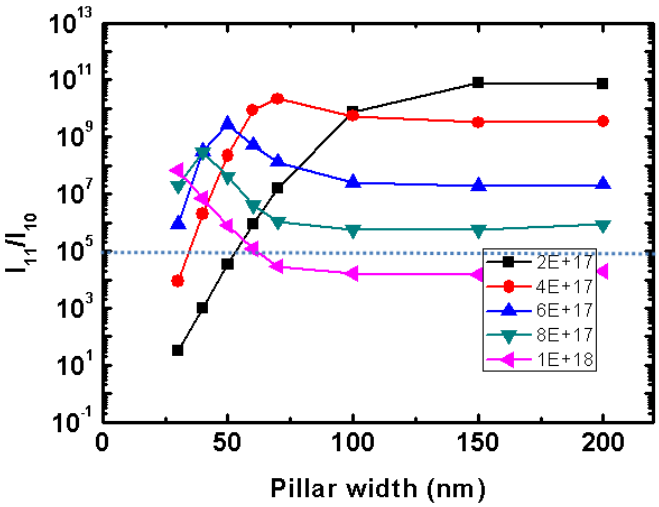
Fig. 3.19 Fabrication process of the downscaled cut-off gate GTB structure.

Leakage current of this structure is also investigated by the device simulation. The radius of the cut-off gates is scaled down by 15 nm, so the length between the cut-off gates is increased by 30 nm than normal GTB array structure. Fig. 3.20(a) shows the current flow of the 10 state of this structure. Compared to the leakage current of normal GTB array structure as shown in Fig. 3.17(a), punch-through leakage current can be

effectively suppressed because the distance between the neighboring cut-off gates is lengthened. Consequently, the current ratio of 11 state and 10 state is dramatically increased so that it is more than 10^6 with the 30 nm pillar width in channel doping concentration of $6\times10^{17}/cm^3$. Therefore, the required channel doping concentration to prevent punch-through leakage current is reduced so that the reading speed can greatly increase.



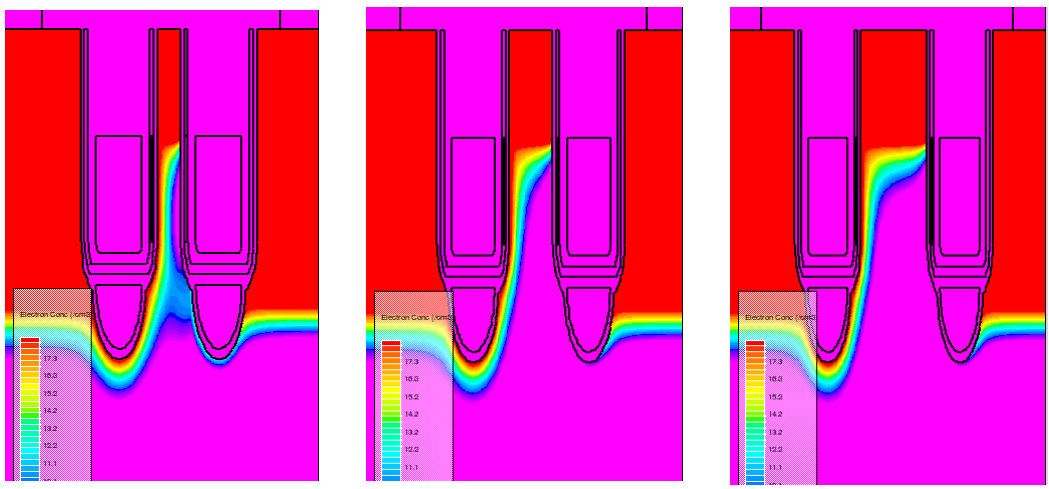
(a)



(b)

Fig. 3.20 (a) Flowing current of 10 state (b) ratio of 11 state current and 10 state current in downscaled cut-off gate structure with different silicon pillar width and channel doping concentration.

Fig. 3.21 shows the electron concentration distribution in the downscaled cut-off gate structure during the read operation of 10 state. Compared to the Fig. 3.16, electron concentration between the cut-off gates is decreased drastically, so that the punch-through leakage can be blocked well.



(a) 30 nm (b) 60 nm (c) 100 nm

Fig. 3.21 Electron concentration distribution in the downscaled cut-off gate structure during the read operation of 10 state with the silicon pillar thickness of (a) 30 nm (b) 60 nm (c) 100 nm.

Chapter 4

Device Fabrication of GTB SONOS

In this chapter, fabrication process and measured data of GTB SONOS NAND flash memory are presented. Cut-off gate and control gate are successfully formed by the etch-back process. And the fabricated device showed good twin-bit operation characteristics.

4.1 Fabrication Process

4.1.1 Active Fin Patterning and Hard Mask Remaining

The fabrication process follows the sequence which is explained in Section 3.3. The starting material is p-type bulk silicon wafer. After a standard cleaning and 100 Å of dry oxidation for sacrificial layer, boron ions (B^+) are implanted at energy of 80 keV and with a dose of 1×10^{13} cm^{-2} to adjust the threshold voltage. The sacrificial oxide is removed, and 100 Å of oxide, 800 Å of nitride are deposited by low pressure chemical vapor deposition (LPCVD). The nitride layer is essential for the uniform thickness of

hard mask after the chemical mechanical planarization (CMP) process. Next, 375 Å of oxide is deposited by plasma enhanced chemical vapor deposition (PECVD) which plays a role of protecting the lower nitride layer, and 700 Å of α-Si is deposited by LPCVD. Because the negative tone e-beam resist, which is called hydrogen silsesquioxane (HSQ), is etched during the oxide and nitride layers etching process, this α-Si dummy layer is first patterned by HSQ and it becomes a mask for oxide and nitride etching. The α-Si dummy layers are patterned by mix-and-match of e-beam and photo lithography, and α-Si, oxide, nitride, oxide, silicon layers are etched in order. As a result, 5500 Å of silicon fin is formed with the nitride hard mask.

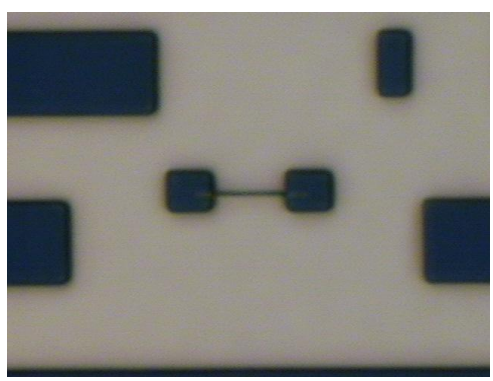


Fig. 4.1 Top view microscopic image of fabricated GTB device after active silicon fin formation. Neighboring dummy patterns help the uniform etching during the CMP process.

Then, the shallow trench isolation (STI) process is performed. 100 Å of oxide is

deposited to prevent damage to silicon during the following high density plasma chemical vapor deposition (HDPCVD) process. 8200 Å of oxide is deposited by HDPCVD for gap filling, and CMP is followed until the nitride layer is exposed. Different from the conventional STI process, nitride layer is not removed because it plays a role of hard mask of the silicon recess process as shown in Fig. 4.2.

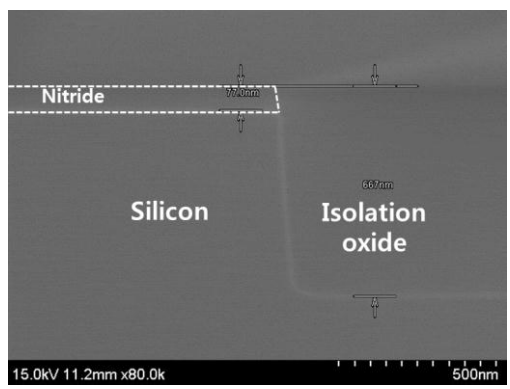


Fig. 4.2 Cross-sectional SEM image of the nitride hard mask and isolation oxide.

4.1.2 Recessed Channel Formation

As stated in Section 3.3, oxide trench etching precedes the silicon recess process. To increase the cut-off gate controllability, the oxide trench should be deeper than that of silicon. Deeper oxide trench enables the FiReFET [44] structure which can decrease the leakage current. To resist the oxide etching and silicon etching processes, additional hard mask is required; 700 Å of poly-Si is deposited.

Oxide and silicon trench is patterned as shown in Fig. 4.3. Green layer shows the

top view of the trench pattern. The cut-off gate poly-Si and control gate poly-Si will be contacted with metal at upper rectangle and lower rectangle area, respectively. This layer is patterned by e-beam lithography with ZEP which is one of the positive photoresist.

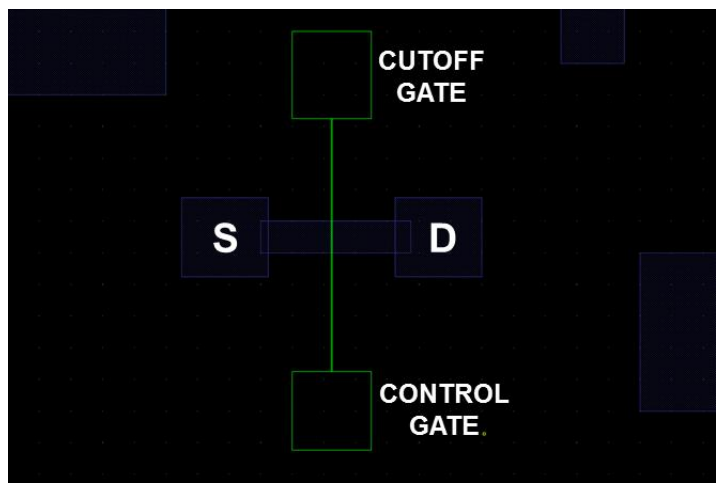
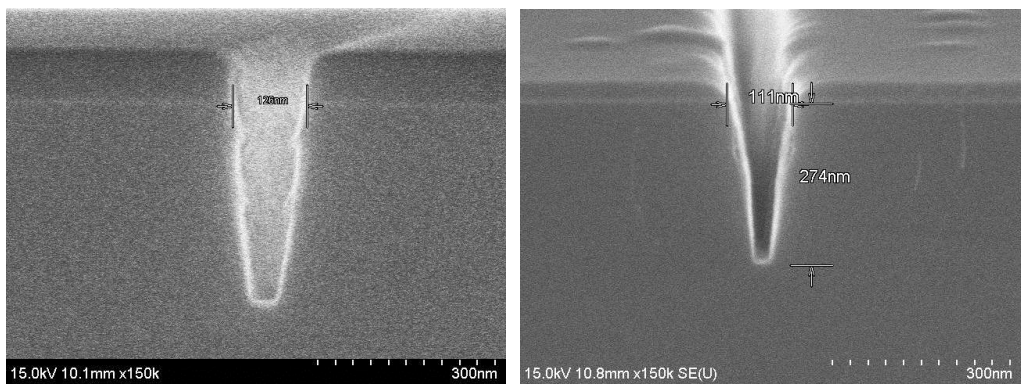


Fig. 4.3 Layout of GTB device (blue: active layer, green: trench layer).

The silicon recess process using the anisotropic etcher (ICP type) is followed by the oxide trench etching process. This process is a very critical step for the accurate formation of cut-off gate and control gate. The steep trench slope is favorable for the high density integration, but can induce bowing effect and distort the cut-off gate and control gate profile. Fig. 4.4 shows the cross-sectional SEM image of the silicon trench with oxide and nitride hard mask. Although nitride layer must be used as a hard mask because of the uniform thickness after the CMP process, the gradual slope of etched

nitride layer is transferred to the silicon trench as shown in Fig. 4.4(b).

To achieve the proper profile for the two gates, some experiments have been carried out with the various etching condition changing the input parameters. Table 4.1 shows how the input parameters affect the trench profile.



(a) (b)

Fig. 4.4 Cross-sectional SEM image of the recessed silicon trench formation with (a) oxide hard mask (b) nitride hard mask.

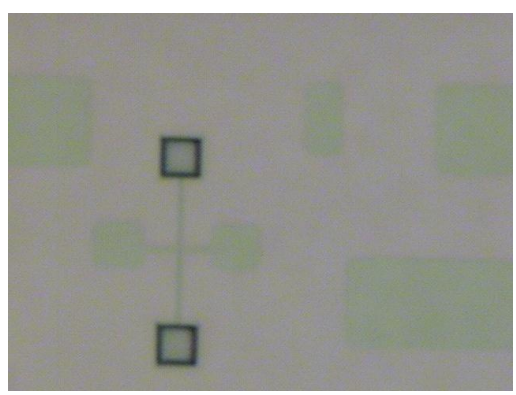


Fig. 4.5 Top view microscopic image of the fabricated device after the oxide and silicon trench etching process.

Table 4.1 Effects of some etching input parameters to the trench profile.

Input parameters	Slope	Etch rate	Selectivity to oxide
Platen Power ↑	↑	↑	↓
O_2 flow rate ↑	↓	↓	↑

With these results, new etching condition is suggested as shown in Table 4.2. To achieve the steeper trench slope, the platen power is increased to 150 W for ions having high directionality to the wafer. However, the high platen power degrades the selectivity to the oxide so that remaining hard mask can be removed during the silicon trench etching process. Therefore, O_2 flow rate is increased for the high selectivity.

Table 4.2 Newly suggested silicon trench etching condition.

	Coil power [W]	Platen power [W]	HBr flow rate [sccm]	O_2 flow rate [sccm]	Pressure [mTorr]
Conventional	900	80	40	2	3
New	900	150	40	3	3

However, if the width of the trench is too narrow, silicon etch rate suddenly decreases below the 70-nm width because of the microloading effect as shown in Fig. 4.6. If the etched silicon depth is shallower than the intended depth, the cut-off gate can disappear during the etch back process.

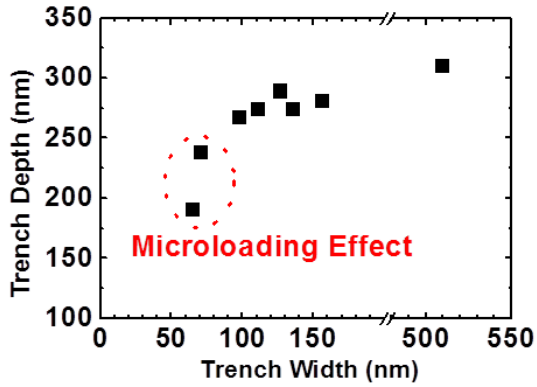


Fig. 4.6 Etched silicon depth with the different trench width.

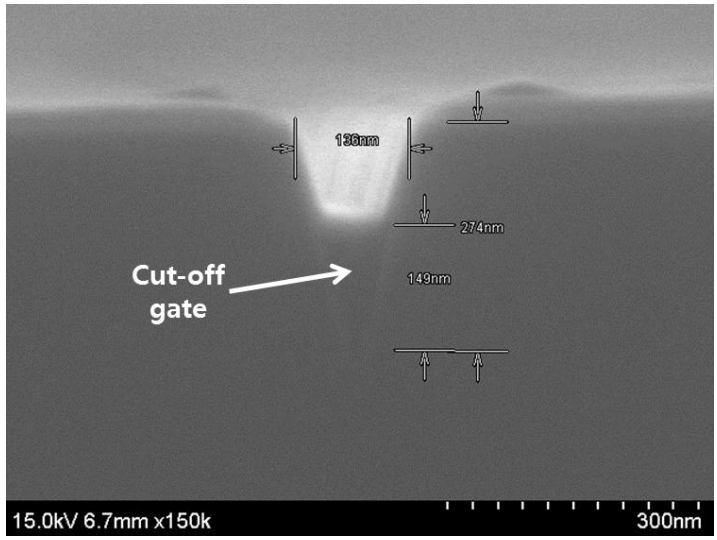
4.1.3 Gate Process: Deposition and Etch-back

After the silicon trench formation, 60 Å thermal dry oxidation is performed for 11 minutes at 850 °C for the cut-off gate oxide. Next, 3000 Å of n^+ doped poly-Si is deposited by LPCVD for the cut-off gate. If the deposited poly-Si is thick enough, the surface above the silicon trench becomes flat so that no planarization process is required after the doped poly-Si deposition. Doped poly-Si is etched back for the cut-off gate formation as shown in Fig. 4.7(a).

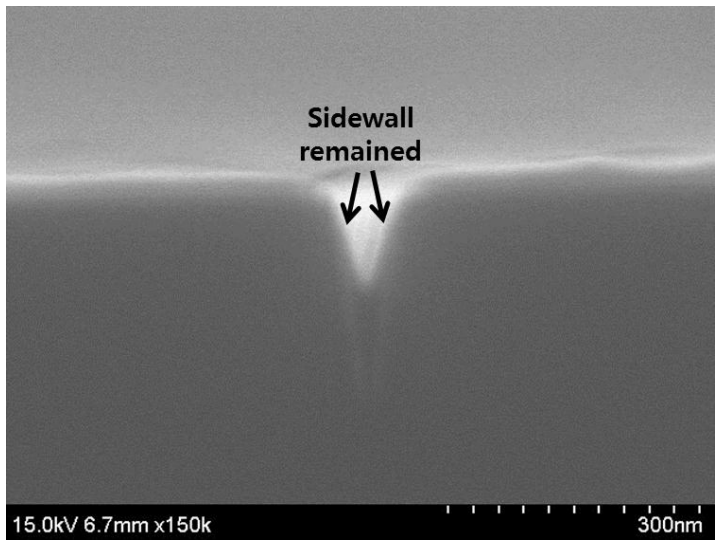
If the slope of the silicon trench is too steep, the poly-Si can remain at the sidewall of the silicon trench as shown in Fig. 4.7(b). This remaining silicon hinders the accurate channel length control of the cut-off gate and the control gate. It can be removed by isotropic etching process: in this experiment, SC-1 solution is used. In the SC-1 solution ($NH_4OH:H_2O_2$:D.I. water = 1:1:5) with 80 °C temperature, the etch rate of the phosphorus doped poly-Si is 1.2 nm/min. To clearly remove the remaining sidewall

poly-Si, the wafer is dipped for 10 minutes.

On the other hand, it should be noted that SC-1 solution etches the silicon dioxide with the etch rate of 0.4 nm/min for thermal oxide and much faster for CVD oxide. If the oxide trench is attacked by SC-1 solution for a long time, the width of the oxide trench becomes wider so that the oxide trench could not be filled completely by CVD process, and results in the disappearance of the WL poly-Si gates during the etch back process.



(a)



(b)

Fig. 4.7 Cross-sectional SEM image after the cut-off gate formation. (a) Well-defined structure and (b) poly-Si sidewall remained structure.

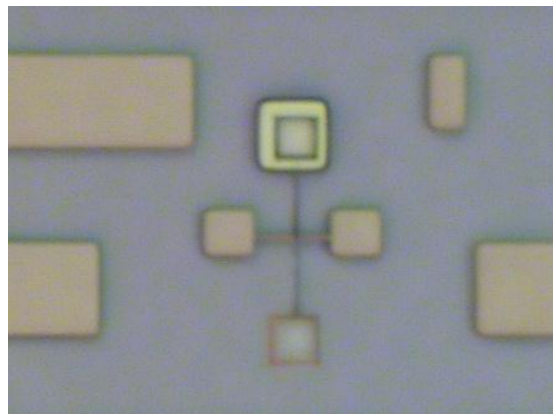


Fig. 4.8 Top view microscopic image of the fabricated device after the cut-off gate formation.

After the cut-off gate formation, the oxide removal process is carried out using diluted HF (HF:D.I. water = 1:100). Subsequently, 40 Å of tunneling oxide is formed by dry oxidation process and 80 Å of nitride for the charge trapping layer, 95 Å of blocking oxide layer is deposited by LPCVD. And then, 3000 Å of n^+ doped poly-Si is deposited for the control gate as shown in Fig. 4.9.

Similar to the cut-off gate formation, the control gate is formed by etch-back process using ICP dry etcher. The remaining sidewall of cut-off gate must be completely removed before the ONO and poly-Si deposition; otherwise the control gate cannot have the channel region as shown in Fig. 4.10(a).

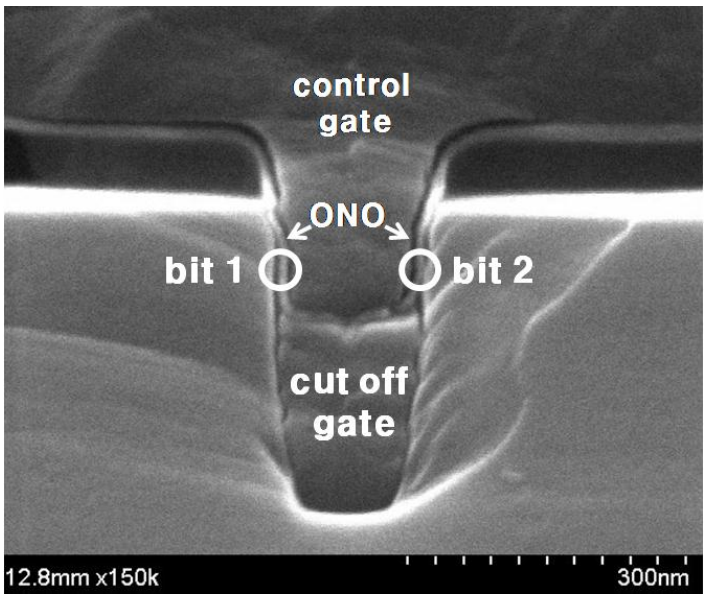
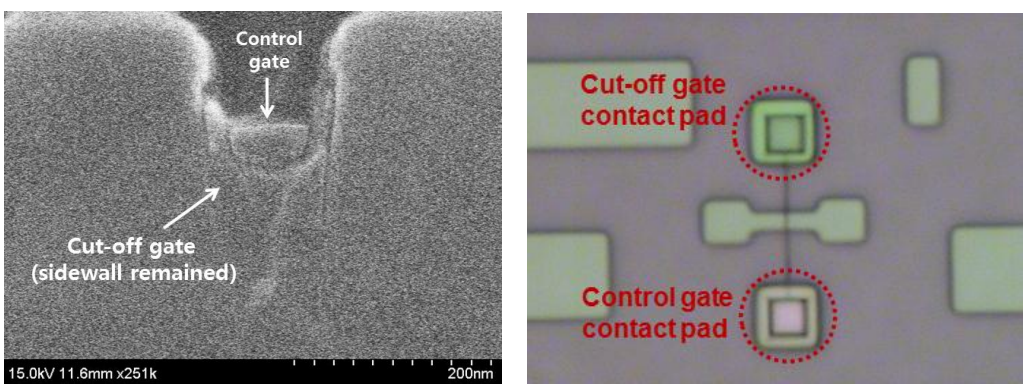


Fig. 4.9 SEM image of the GTB device after the control gate poly-Si deposition.



(a) (b)

Fig. 4.10 (a) SEM image of the GTB device after the control gate etch-back process (sidewall remained structure). (b) Top view microscopic image.

4.1.4 Source/Drain Implantation

Remaining hard mask on the silicon is removed by dry etch process. And source/drain is formed by As^+ ion implantation. As^+ ions are implanted with 10 keV energy and a dose of 1×10^{15} cm^{-2}. Next, rapid thermal annealing (RTA) process is carried out for 7 seconds at 1000 $^{\circ}C$ to activate the implanted dopants.

As verified by simulation in Chap. 3, channel length of the control gate is very critical factor for the exact twin-bit operation. The channel length is directly determined by the junction depth of the source and drain. To find the exact junction depth of the source and drain, the doping concentration is analyzed by secondary ion mass spectroscopy (SIMS) method. As shown in Fig. 4.11, the junction is formed at 90 nm below the silicon surface after the RTA process.

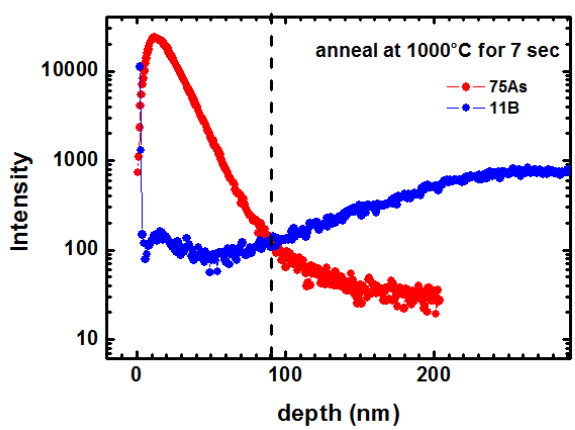


Fig. 4.11 SIMS profile after RTA process.

Next, the back-end-of-line (BEOL) processes are performed; inter-layer dielectric (ILD) deposition, contact formation and metallization. And finally, the alloy process in H_2 and N_2 ambient go after the BEOL process.

4.2 Electrical Characteristics

Electrical characteristics are discussed in this chapter. DC characteristics are analyzed using HP4156C parameter analyzer.

First, the cut-off gate characteristic is measured as shown in Fig. 4.12. The current flow is absolutely blocked when 0 V is applied to the cut-off gate regardless of the control gate voltage.

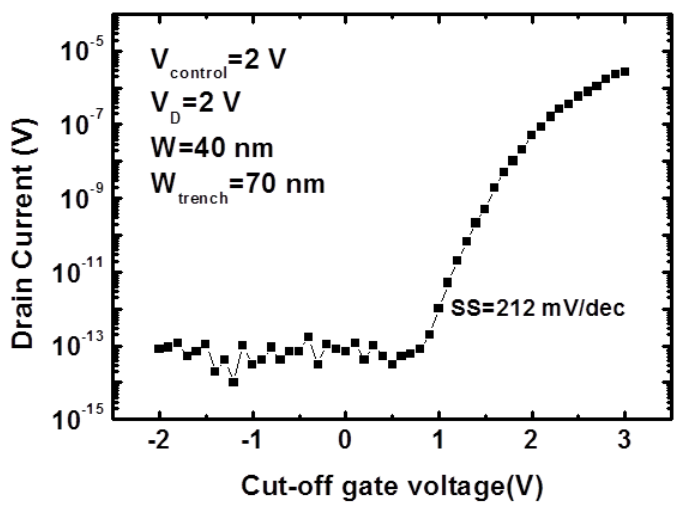


Fig. 4.12 Cut-off gate characteristic of the fabricated GTB device. 2 V is applied to the control gate.

Fig. 4.13 shows the control gate transfer characteristic when only the source side node is programmed. The program voltage is 14 V, and stress time is 200 μs. -4 V is applied to the cut-off gate to prevent the channel potential of each side is transferred. Because the fabricated device does not have the select gate which can induce self-boosting effect of inhibit cell, 6 V is applied to the drain so that Fowler-Nordheim tunneling does not occur at the drain side node. On the other hand, 0 V is applied to the source to induce FN tunneling.

During the read operation, forward-reverse reading scheme is used as mentioned in Chap. 3. To read the source side node, 2.5 V is applied to the drain and 0 V to the source. Because the source side node is programmed, threshold voltage is relatively high (black dot). On the contrary, if 0 V is applied to the drain and 2.5 V to the source, threshold voltage is relatively low because the injected charge at the source side cannot affect it.

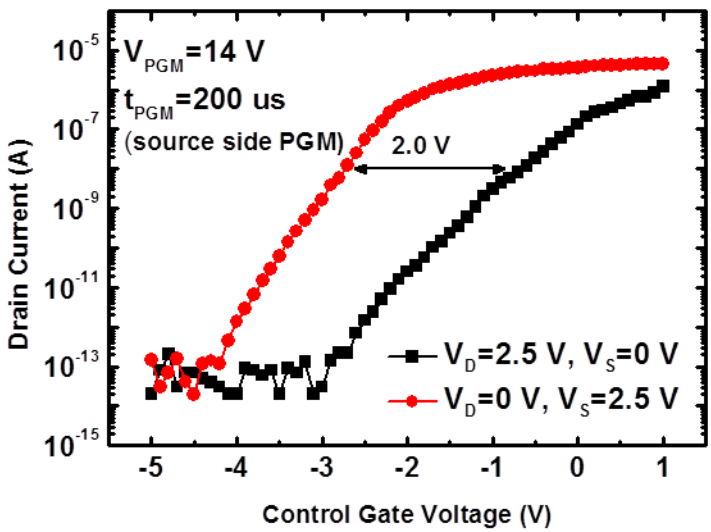
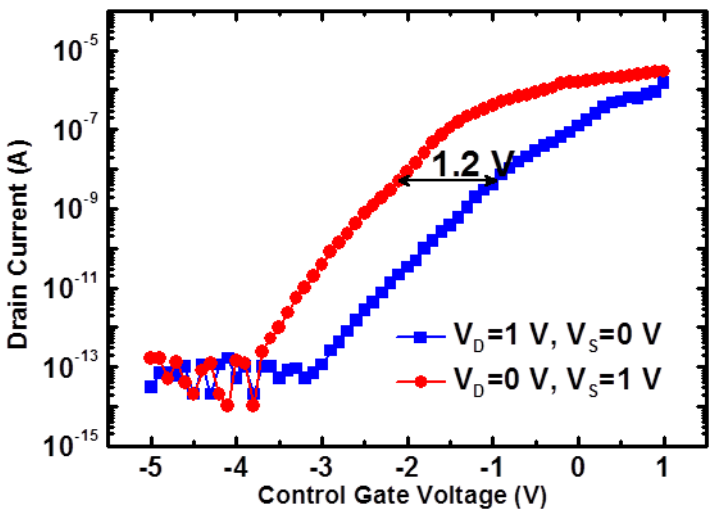
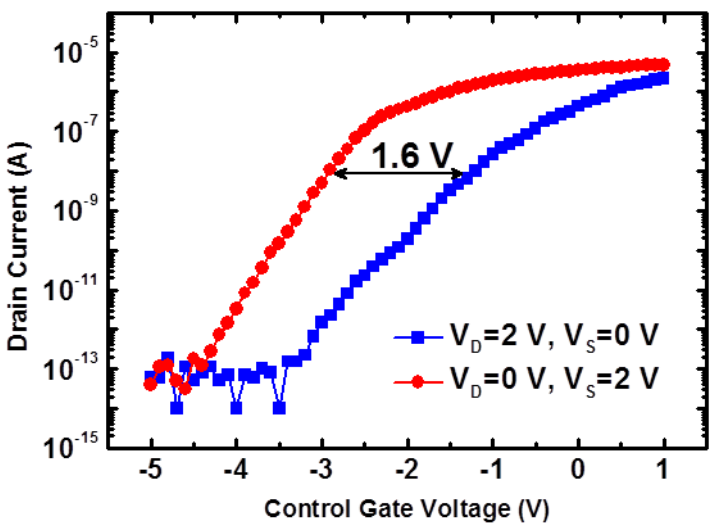


Fig. 4.13 I-V characteristic curve of the control gate when only the source side node is programmed.

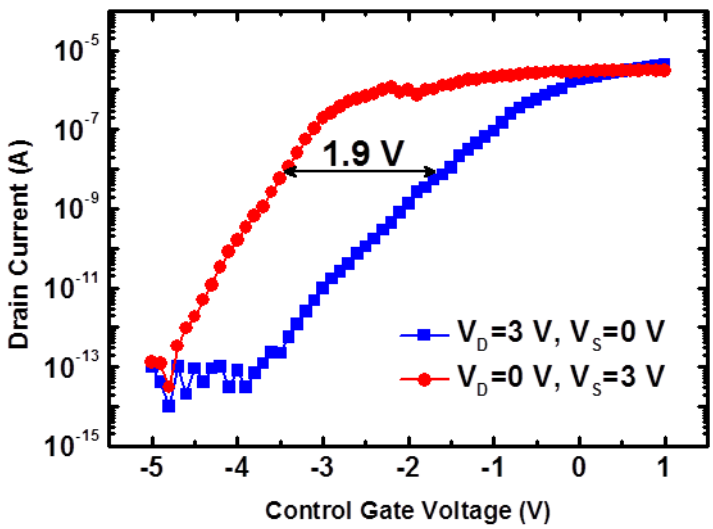
To verify the barrier lowering effect, forward-reverse read is performed with various source and drain voltage. As stated in Section 3.2.3, the higher the drain voltage, the more easily the barrier lowering occurs. It means the threshold voltage of 10 and 01 states are more clearly distinguished. In this fabricated GTB device, higher drain bias makes a larger threshold voltage difference as shown in Fig. 4.14.



(a) Forward read: V_D=1 V, V_S=0 V, reverse read: V_D=0 V, V_S=1 V



(b) Forward read: V_D=2 V, V_S=0 V, reverse read: V_D=0 V, V_S=2 V



(c) Forward read: V_D=3 V, V_S=0 V, reverse read: V_D=0 V, V_S=3 V

Fig. 4.14 Forward-reverse read operation with various source and drain voltage.

By measuring the V_T difference of the forward-reverse read case, the effective channel length of the fabricated device can be assumed. As stated in Section 3.2.3, the read error increases as the channel length of the storage node becomes longer; the V_T difference of forward-reverse read decreases. Compared to the simulated data in Table 4.3, V_T difference of the fabricated device is between that of 40 nm and 50 nm device. To conclude, the length of the storage node of the fabricated device is assumed to be 4x nm.

Table 4.3 V_T difference of forward-reverse read of fabricated device and simulated data (V_D=2 V).

L_{node} (nm)	$V_{T,F}$-$V_{T,R}$ (V)
30 (sim.)	2.20
40 (sim.)	2.01
Fabricated device	1.6
50 (sim.)	1.48
60 (sim.)	1.04
70 (sim.)	0.67
80 (sim.)	0.50

Fig. 4.15 shows the I-V characteristic curve of control gate with 4 states (00, 01, 10, 11). Bias condition for the 00 state is 14 V of program voltage during the 200 μs, 0 V to the source and the drain, and -4 V is applied to the cut-off gate. In the case of erase operation (11 state), -14 V is applied to the control gate for 1 ms.

For the read bias condition, the drain voltage is higher than the source voltage to read the state of source side storage node. The threshold voltage of 01 and 11 state is almost equal the each other and higher than that of 10 and 11 state; that is, only the state of the source side storage node affects the threshold voltage in this fabricated GTB device.

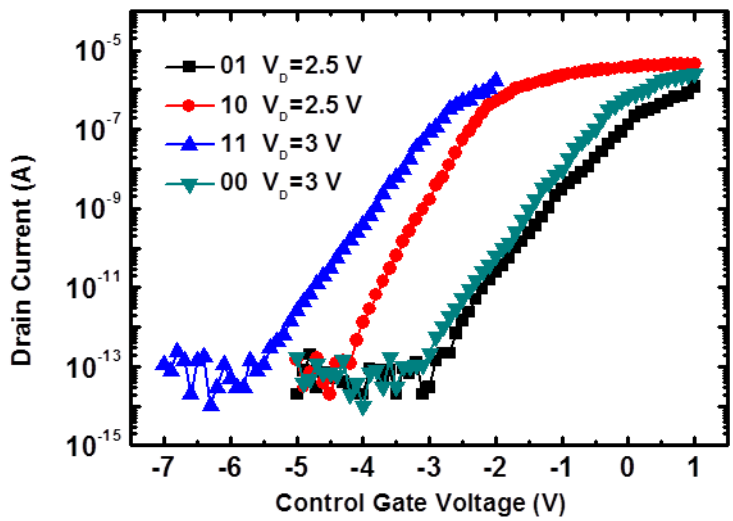


Fig. 4.15 I-V characteristic curve of control gate with 4 states.

Chapter 5

Gated Multi Bit (GMB) Array

In this chapter, gated multi-bit (GMB) array is introduced and operation schemes are verified. To compete with the today's 3-dimensional NAND arrays, we have proposed a solution of stacking wordlines of GTB in the vertical direction. By stacking the vertical gates, 2N bit per $4F^2$ size memory density can be achieved with single crystalline silicon channel where N is the number of the stacked layers.

5.1 Introduction

Today, various types of 3D NAND flash memories are proposed. 3D NAND flash can increase the memory density drastically as we stack the layers in vertical direction. Gate stack type 3D flash memories such as BiCS and TCAT have N bit per $6F^2$ area memory density as shown in Fig. 2.6 and Fig. 2.7.

Although GTB device can be integrated at a density of 2 bit per $4F^2$ area that is twice

as high as the conventional NAND array, it is inferior to the 3D NAND flash memory where the number of stacked layers is more than 4.

The minimization of feature size has a limitation because of two reasons; first, since the GTB array has memory nodes at both sides of a trench, minimum space for two ONO stacks and poly-Si control gate is required as shown in Fig. 3.1. Second, too thin silicon pillar thickness induces the punch-through leakage current between the neighboring cut-off gates as mentioned in Section 3.4.2.

To compete with BiCS and TCAT, gated multi-bit (GMB) structure is proposed in this chapter [45]. Fig. 5.1 shows the gated multi-bit array structure with 4 stacked gates. Deeper trenches are formed to secure enough height for multiple stacked gates. Performing dielectric deposition, poly-Si gap filling, and etch-back processes in sequence, a GMB device is constructed.

Like a GTB device, the storage nodes exist at the left and right side of each wordline. If N is 4 as shown in Fig. 5.1, 4 stacked wordlines in $4F^2$ area have 8 storage nodes. Therefore, 2N bit per $4F^2$ area memory density can be achieved with gated multi bit array.

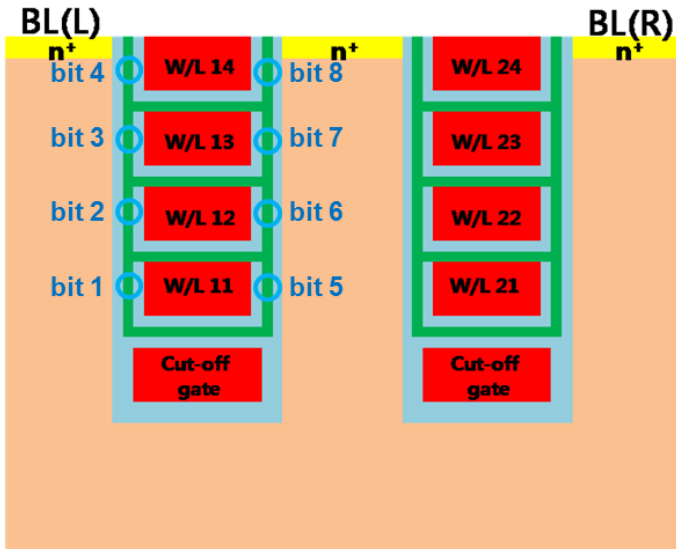


Fig. 5.1 Gated multi bit (GMB) NAND flash memory with 4 stacked gates.

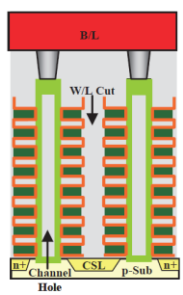
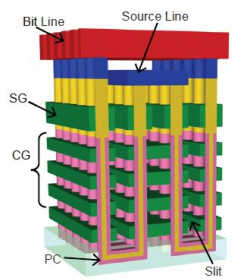
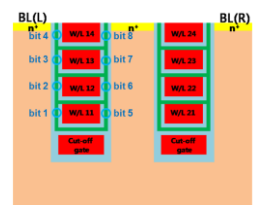
Table 5.1 compares the GMB array with TCAT and P-BiCS which are the most promising gate stack type 3D NAND flash memory recently. TCAT and P-BiCS has gate all around (GAA) channel structure. GAA channel has good gate control ability, but it has disadvantage that threshold voltage can change as the radius of the channel poly-Si varies. During the etch process, 90° slope is hardly achieved so that the poly-Si channel cannot have regular radius; V_T variation of each layer becomes serious problem.

In the point of the view of memory density, GMB array has three times higher density than BiCS and TCAT. As shown in top view of Fig. 2.6(b), existing gate stack type NAND arrays have a unit cell in $6F^2$ where 3F along the bitline direction and 2F along the select gate direction. Therefore the required area for 1 bit is $6F^2/N$ where N is the number of the stacked layers. Meanwhile, GMB array has N wordlines in $4F^2$ area and

each wordline has two storage nodes so that 2N bit can be integrated in $4F^2$ area; required area for 1 bit is $2F^2/N$.

BiCS and TCAT uses punch and plug process to form the silicon channel. So the silicon substrate cannot be used for the channel, and the low pressure chemical vapor deposition (LPCVD) process is used for the plug process so that only polycrystalline silicon can be used. Poly crystalline silicon channel has bad reliability characteristics compared to the single crystalline silicon channel; grain boundary can induce leakage current and electron mobility degradation, and subthreshold swing (SS) characteristic is poor.

Table 5.1 shows the comparison of the gate stack type 3D NAND flash memories.

	TCAT	P-BiCS	GMB
Structure			
Etch slope V_T variation	Bad	Bad	Good
Area per 1 bit	$6F^2/N$	$6F^2/N$	$2F^2/N$
Channel Si crystalline	Poly crystalline	Poly crystalline	Single crystalline

5.2 Operation Scheme of the GMB array

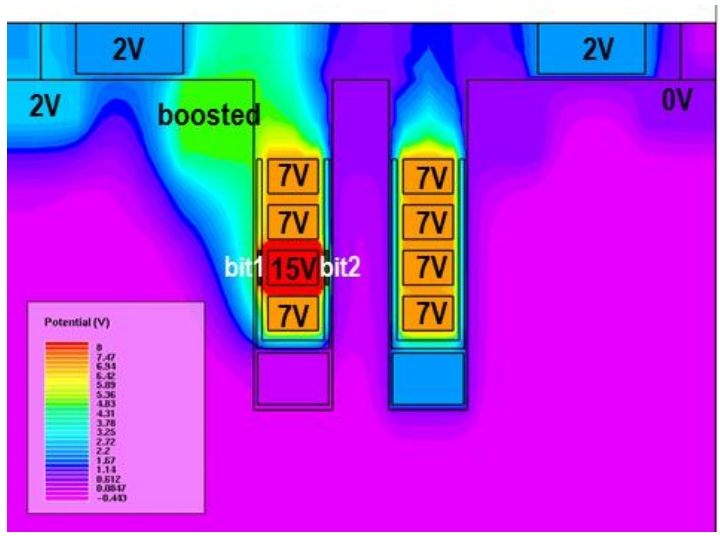
5.2.1 Program Operation

Operation scheme of the GMB array is similar to that of GTB array. Table 5.2 shows the programming bias condition of the GMB array. During the program operation, 0 V is applied to the cut-off gate for the separate programming to prevent the channel potential to be transferred. And, 0 V is applied to the bitline of want-to-program side, and V_{DD} is applied to the other side. Unselected layers' wordlines are biased at V_{PASS}, to transfer the bitline potential to the channel of the selected wordline.

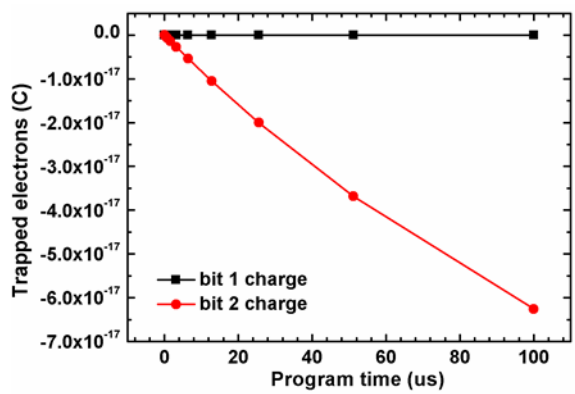
Table 5.2 Programming bias condition of the GMB array.

CONTROL LINE	PROGRAM (R side of WL12)
WL11(unselected layer)	V_{PASS}
WL12(selected)	$V_{PROGRAM}$
WL13(unselected layer)	V_{PASS}
WL14(unselected layer)	V_{PASS}
WL21(unselected)	V_{PASS}
WL22(unselected)	V_{PASS}
WL23(unselected)	V_{PASS}
WL24(unselected)	V_{PASS}
CUT-OFF GATE 1(selected)	LOW
CUT-OFF GATE 2(unselected)	HIGH
SSL	V_{DD}
DSL	V_{DD}
L side of selected BL	V_{DD}
R side of selected BL	GND
L side of unselected BL	V_{DD}
R side of unselected BL	V_{DD}
Substrate	GND

Fig. 5.2(a) shows the potential distribution of the programming operation. Bit 2 of the WL12 (red color) is selected for the programming cell. 0 V of right side bitline is properly transferred to the channel of bit 2, and the channel of bit 1 is boosted so that programming is inhibited as shown in Fig. 5.2(b). 7 V is applied to the unselected layers' wordlines to pass the channel potential and help the inhibit cell side channel to be boosted.



(a)



(b)

Fig. 5.2 (a) Potential distribution of the programming operation. The case of the right side storage node of the WL12 (red color) is programmed. (b) Injected charge of bit 1 and bit 2.

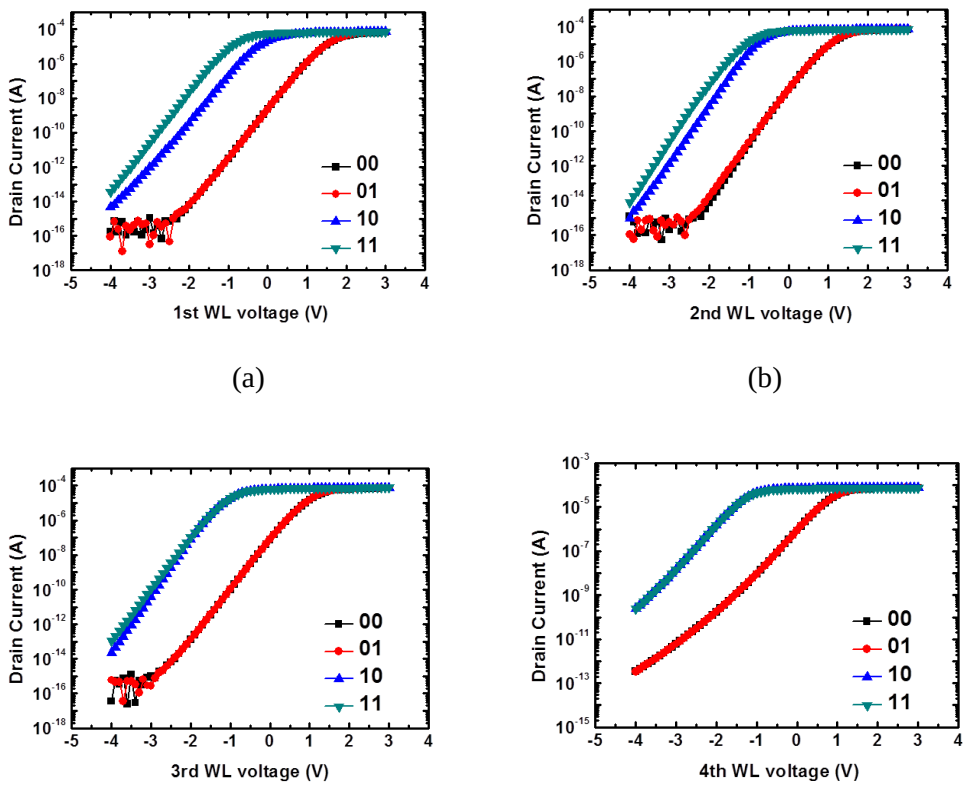
5.2.2 Read Operation

Table 5.3 shows the read bias condition of the GMB array. As in the case of GTB array, the forward-reverse reading scheme is used. V_{READ} is applied to the unselected wordlines to turn on all the gates.

Table 5.3 Read bias condition of the GMB array.

CONTROL LINE	READ (R side of WL12)
WL11(unselected layer)	V_{READ}
WL12(selected)	0
WL13(unselected layer)	V_{READ}
WL14(unselected layer)	V_{READ}
WL21(unselected)	V_{READ}
WL22(unselected)	V_{READ}
WL23(unselected)	V_{READ}
WL24(unselected)	V_{READ}
CUT-OFF GATE 1(selected)	HIGH
CUT-OFF GATE 2(unselected)	HIGH
SSL	V_{DD}
DSL	V_{DD}
L side of selected BL	V_{DD}
R side of selected BL	GND
L side of unselected BL	-
R side of unselected BL	-
Substrate	GND

Fig. 5.3 shows the I-V characteristic curve of the GMB array for each layer. In all the I-V characteristics curves, V_T is determined by the left side cell and the effect of right side cell is screened successfully.



(a) (b)

(c) (d)

Fig. 5.3 I-V characteristic curve of the GMB device for each layer. (a) 1st layer (b) 2nd layer (c) 3rd layer (d) 4th layer (V_{READ}=7 V, V_{BL}=3 V).

The pass voltage is applied to all unselected wordlines to induce high electron concentration as shown in Fig. 5.4, so that minimized channel resistance prevents the unintended voltage drop. Therefore, accurate twin-bit operation of selected wordline is possible.

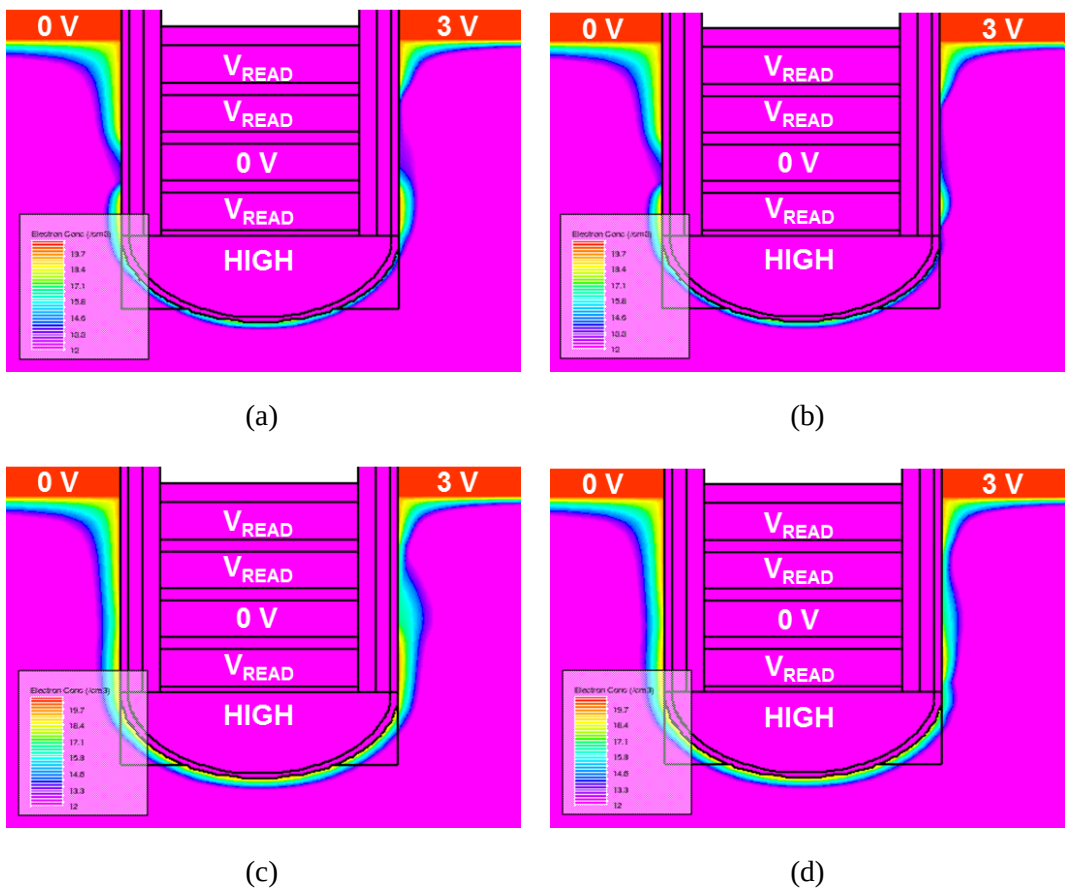


Fig. 5.4 Electron concentration distribution of GMB device when reading the 2nd layer's state. (a) 00 state (b) 01 state (c) 10 state (d) 11 state.

5.3 Fabrication Process of the GMB Device

There are two ways to fabricate the GMB device. The first is polysilicon etch-back and charge trap layer re-deposition method. The other is the epitaxial growth of silicon channel method.

Fig. 5.5 shows the fabrication process of polysilicon etch-back and charge trap layer re-deposition method after the STI process of the GMB device. Key processes are summarized as follows:

Fig. 5.5(a): Deep trench formation.

Fig. 5.5(b): Gate oxidation and n^+ doped poly-Si deposition by LPCVD for cut-off gate.

Fig. 5.5(c): Cut-off gate is formed by doped poly-Si etch-back process. If the deposited thickness of doped poly-Si at process (b) is thick enough, no planarization process is required.

Fig. 5.5(d): Thin gate oxide removal. Thin oxide on the silicon pillar could be damaged during the etch-back process, so it is clearly removed with isotropic etching such as HF solution.

Fig. 5.5(e): ONO layer deposition. Tunneling oxide is formed by oxidation, and charge trap nitride and blocking oxide layer is deposited by LPCVD.

Fig. 5.5(f): Doped poly-Si deposition for 1^{st} layer wordline.

Fig. 5.5(g): 1^{st} layer wordline is formed by doped poly-Si etch back process as same as process I.

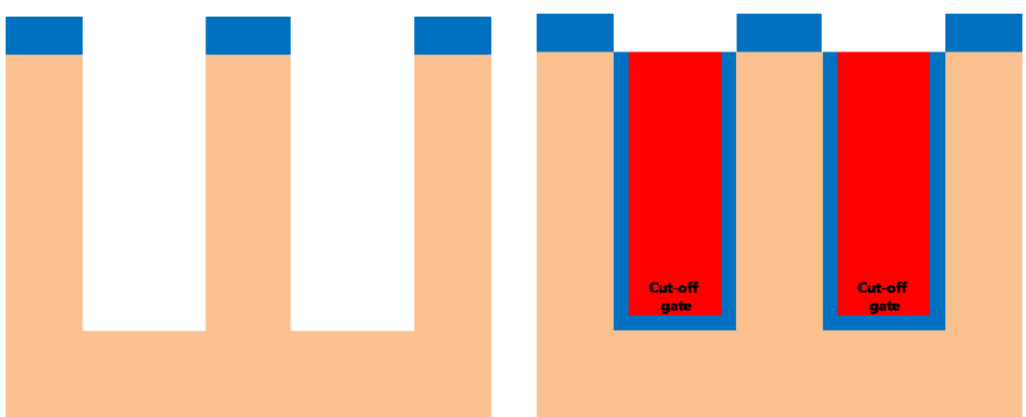
Fig. 5.5(h): Damaged blocking oxide and nitride layers are stripped with isotropic etching process.

Fig. 5.5(i): Nitride and oxide layers are re-deposited. These layers isolate the neighboring wordlines electrically, and protect the side surface during the etch-back process.

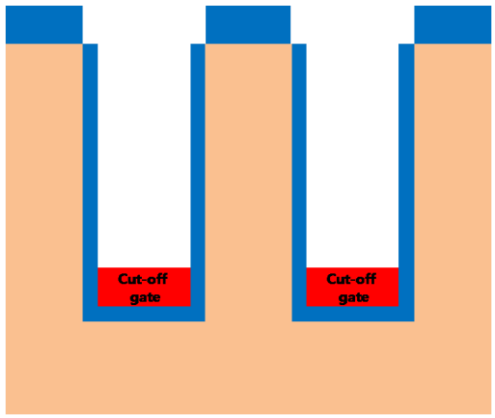
Fig. 5.5(j): Doped poly-Si deposition for 2nd layer wordline.

Fig. 5.5(k): (g)-(j) processes are repeated several times as the number of vertically stacked wordlines.

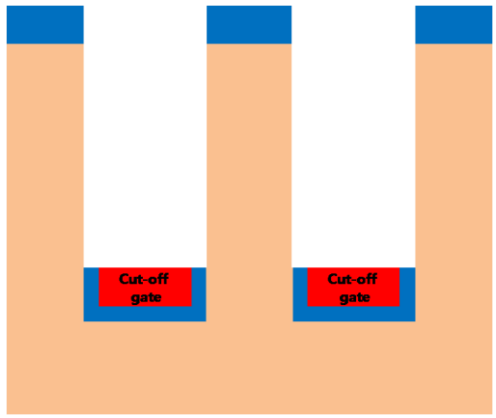
Fig. 5.5(l): Hard masks are removed and n-type dopants are implanted to form the source and drain junctions.



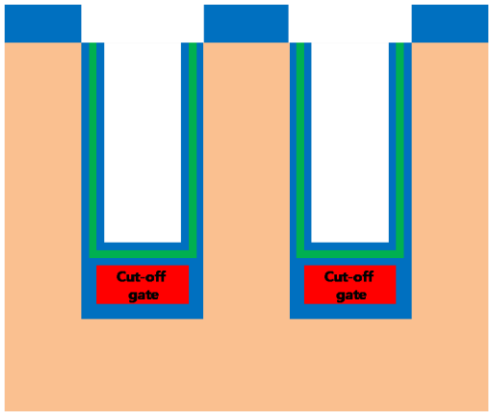
(a) (b)


Cut-off gate
Cut-off gate

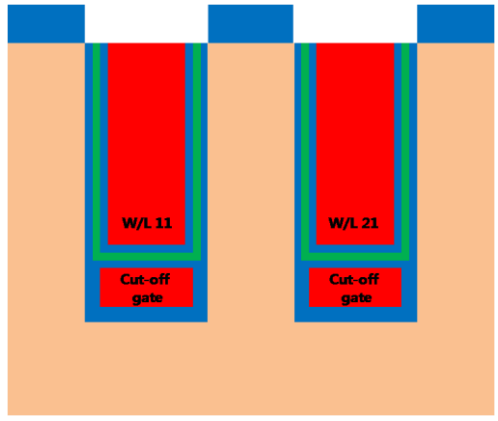
(c)


Cut-off gate
Cut-off gate

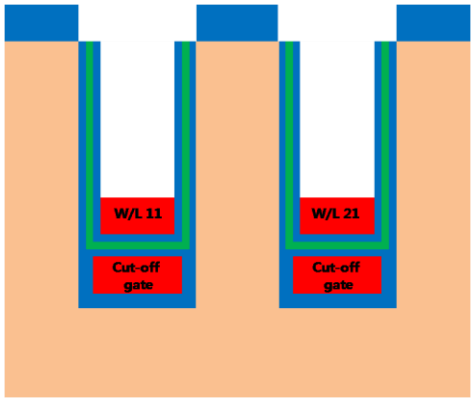
(d)


Cut-off gate
Cut-off gate

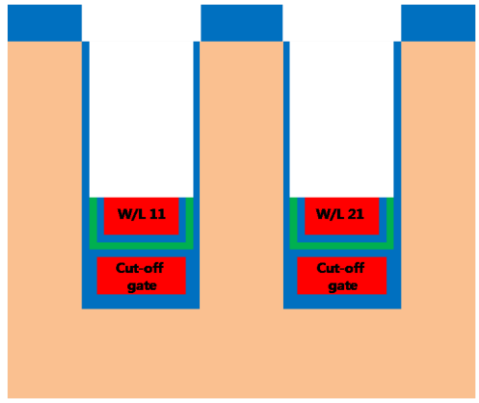
(e)


W/L 11
W/L 21
Cut-off gate
Cut-off gate

(f)


W/L 11
W/L 21
Cut-off gate
Cut-off gate

(g)


W/L 11
W/L 21
Cut-off gate
Cut-off gate

(h)

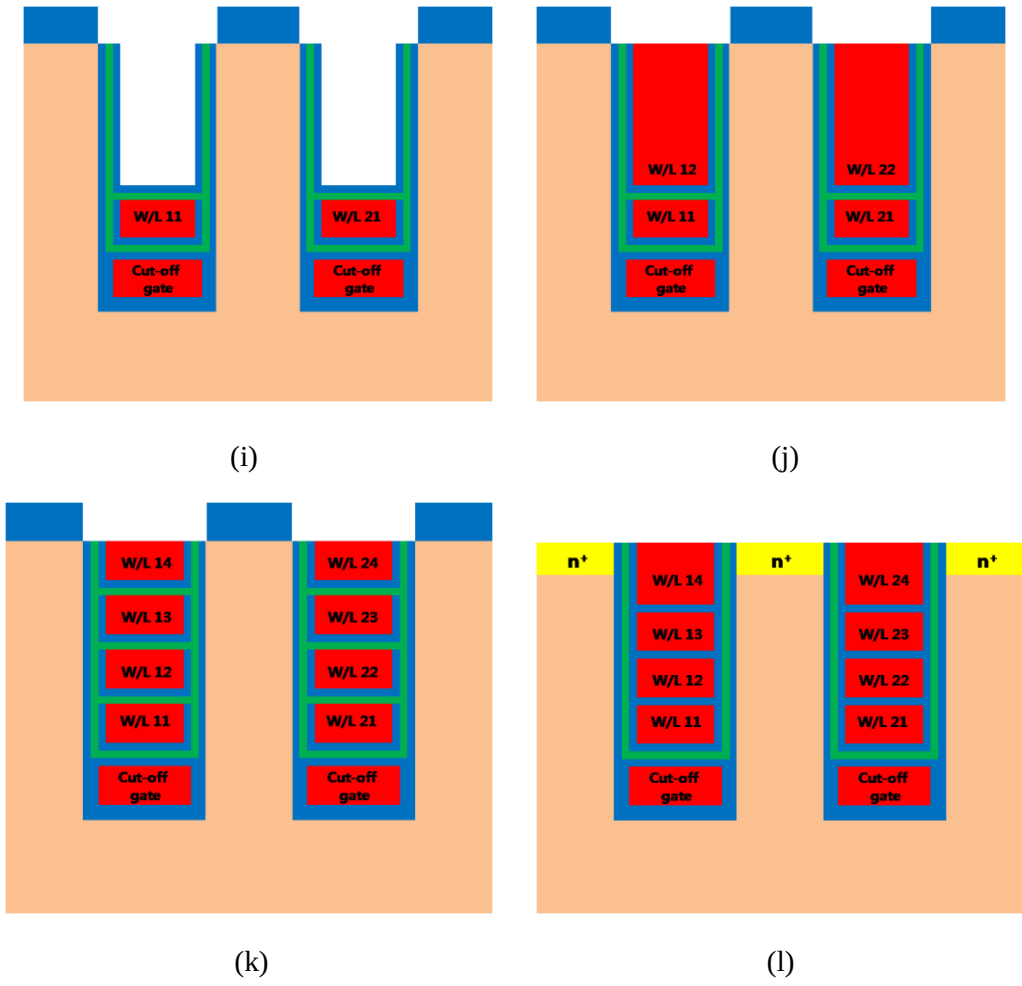


Fig. 5.5 Fabrication process flow of the GMB array with polysilicon etch-back and charge trap layer re-deposition method after a STI process.

This method can realize the single crystalline silicon channel that has better electrical performance than the polycrystalline silicon channel devices. However, this method have disadvantage in that too many LPCVD processes are required because of the re-deposition process. Therefore, epitaxial growth of the silicon channel method can be

used to simplify the fabrication process. Fig. 5.6 shows the fabrication process flow of the GMB array with epitaxial growth of the silicon channel method after a STI process. Key processes are summarized as follows:

Fig. 5.6(a): Cut-off gate oxidation.

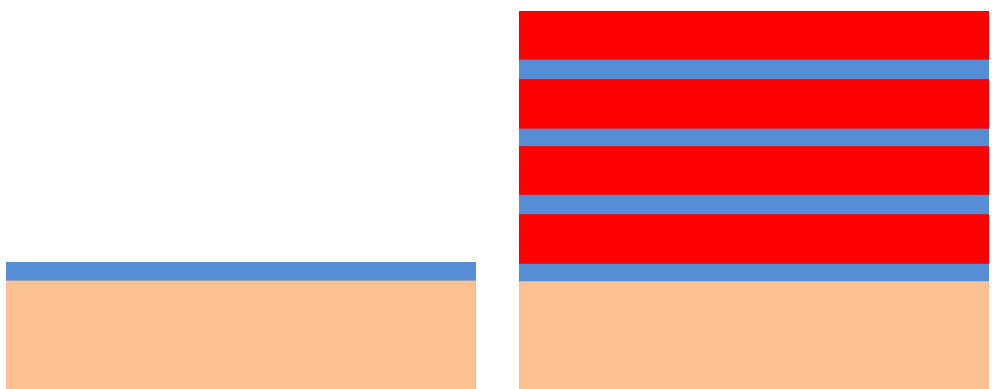
Fig. 5.6(b): n^+ doped polysilicon and oxide layers are deposited as the number of the wordlines.

Fig. 5.6(c): Wordlines are patterned.

Fig. 5.6(d): ONO layer deposition.

Fig. 5.6(e): Single crystalline silicon epitaxial growth.

Fig. 5.6(f): n-type dopants implantation.



(a) (b)

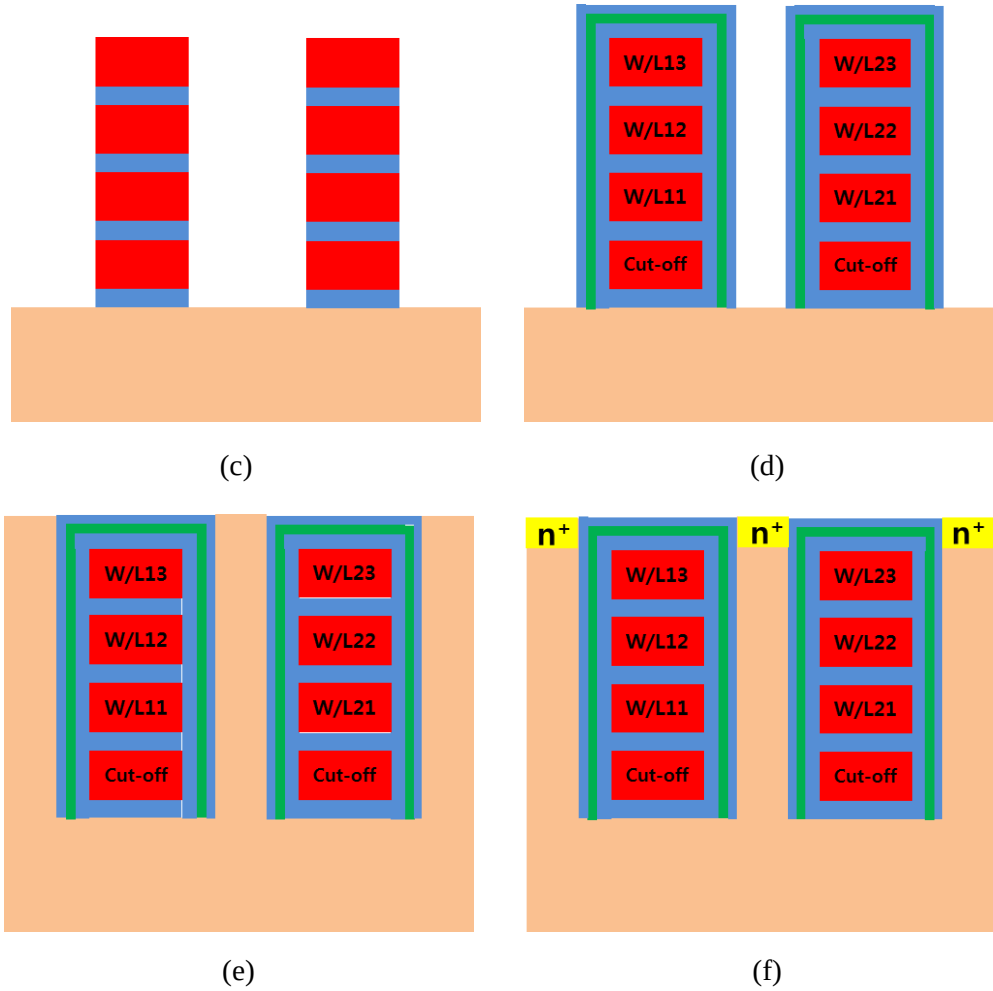


Fig. 5.6 Fabrication process flow of the GMB array with epitaxial growth of the silicon channel method after a STI process.

Chapter 6

Conclusions

In this dissertation, gated twin-bit (GTB) device with SONOS type NAND flash memory was proposed, fabricated, and characterized. To increase the memory density of NAND flash memory, GTB device has recessed channel structure and cut-off gate that enables the two-bit operation. With this 3-dimensional structure, the memory density of 2 bit per $4F^2$ can be realized that is twice that of conventional NAND flash memory.

The method of two-bit operation by cut-off gate was proposed, and its mechanism was verified by the TCAD simulation. Cut-off gate cuts the channel of the left and right side of the wordline so that separate programming is possible. During the read operation, forward-reverse read scheme was used; barrier lowering at the drain side cell suppress the effect of the trapped charge and enables the twin-bit operation.

The GTB device was fabricated with the gate deposition and etch-back method. The cut-off gate of the fabricated device showed good blocking characteristic of the current flow. And in the case that only the source side cell was programmed, the V_T of forward

and reverse reading was clearly differentiated. It was also verified by the increase of V_T difference with various drain voltage.

The gated multi-bit (GMB) array which has single crystalline silicon channel and 2N bit per $4F^2$ memory density was introduced to compete with the recently proposed 3-dimensional gate stacked (or channel stacked) NAND flash structures. Two fabrication methods for GMB array structures were demonstrated.

Bibliography

[1] IEEE Std. 1005-1998, IEEE Standard Definitions and Characterization of Floating Gate Semiconductor Arrays, Feb. 1999.

[2] R. Bez, E. camerlenghi, A. Modelli, and A. Visconi, “Introduction to Flash Memory,” *Proc IEEE*, Vol. 91, pp.489-502, Apr. 2003.

[3] J. E. Brewer, M.Gill, “Nonvolatile Memory Technologies with Emphasis on Flash”, *IEEE Press*, pp. 2, 2008.

[4] R. Micheloni, L. Crippa, A. Marelli, “Inside NAND Flash Memories”, *Springer*, 2010.

[5] Kawamatus, Tatsuya. "TECHNOLOGY FOR MANAGING NAND FLASH". Hagiwara sys-com co., LTD., August 1, 2011.

[6] Lyth0s, "SSD vs. HDD". elitepcbuilding.com, July 11, 2011

[7] J. Wileu, T.Y. Fu, T Tanaka, et al., “Phase-shift mask pattern accuracy requirements and inspection technology,” *Proc. SPIE 1464, Integrated Circuit Metrology, Inspection, and Process Control V, 346,* Jul., 1991.

[8] Mario Garza, Nicholas K. Eib, Keith K. Chao, et al., “Optical Proximity Correction Method and Apparatus,” U.S. Patent No. 5,723,233, Mar. 3, 1998.

[9] Soichi Owa, Hiroyuki Nagasaka, “Immersion lithography; its potential performance and issues,” *Proc. SPIE 5040, Optical Microlithography XVI, 724,* Jun. 2003.

[10] C. M. Lim, S. -M. Kim, Y. -S. Hwang, et al., “Positive and negative tone double patterning lithography for 50nm flash memory,” *Proc. SPIE 6154, Optical Microlithography XIX,* 615410, Mar. 2006.

[11] Yohwan Koh, "NAND Flash Scaling Beyond 20nm," *Memory Workshop, 2009. IMW '09. IEEE International* , pp.1-3, 10-14 May 2009.

[12] Bart van Schravendijk, Dong Niu, Keith Fox, et al., “Vertically Integrated Memory Processes,” *Semicon West,* 7 July 2011.

[13] Jong-Ho Park, Sung-Hoi Hur, Joon-Hee Lee, et al., "8 Gb MLC (multi-level cell) NAND flash memory using 63 nm process technology," *IEEE International Electron Devices Meeting,* pp. 873- 876, 13-15 Dec. 2004.

[14] Yan Li, Seungpil Lee, Yupin Fong, et al., "A 16 Gb 3-Bit Per Cell (X3) NAND Flash Memory on 56 nm Technology With 8 MB/s Write Rate," *IEEE Journal of Solid-State Circuits,* vol.44, no.1, pp.195-207, Jan. 2009

[15] Trinh, C., Shibata, N., Nakano, T., et al., "A 5.6MB/s 64Gb 4b/Cell NAND Flash memory in 43nm CMOS," *IEEE International Solid-State Circuits Conference - Digest of Technical Papers,* pp.246-247,247a, 8-12 Feb. 2009.

[16] K. Kim, “Future outlook of NAND Flash technology for 40nm node and beyond”, P*roceedings of NVSMW,* pp.09-11, 2006.

[17] M. Park et al., “Effect of low-K dielectric material on 63nm MLC (Multi-Level Cell)

NAND Flash cell arrays", *Proceedings of tech. papers of VLSI-TSA,* pp. 37-38, Apr. 2005.

[18] D. Kang et al., "The air spacer technology for improving the cell distribution in 1 Giga bit NAND Flash memory", *Proceedings of NVSMW*, pp. 36-37, 12-16 Feb. 2006.

[19] Y.W. Chang et al., "A new interference phenomenon in sub-60nm nitride-based Flash memory", *Proceedings of NVSMW*, pp. 81-82, 26-30 Aug. 2007.

[20] C.H. Lee et al., "Highly scalable NAND Flash memory with robust immunity to program disturbance using symmetric inversion-type source and drain structure", *Symposium on VLSI technology,* pp. 118-119, 17-19 June 2008.

[21] H.T. Lue et al., "A novel junction-free BE-SONOS NAND Flash", *Symposium on VLSI technology,* pp. 140-141, 17-19 June 2008.

[22] S.M. Jung et al., "Three dimensionally stacked NAND Flash memory technology using stacking single crystal Si layers on ILD and TANOS structure for beyond 30nm node", *IEDM tech. digest*, pp. 1-4, 11-13 Dec. 2006.

[23] H. Tanaka et al., "Bit Cost Scalable technology with punch and plug process for ultra high density Flash memory", *Symposium on VLSI technology,* pp.14-15, 2007.

[24] J. Jang et al., "Vertical cell array using TCAT (Terabit Cell Array Transistor) technology for ultra high density NAND Flash memory", *Symposium on VLSI technology,* pp.192-193, 2009.

[25] Erh-Kun Lai, Hang-Ting Lue, Yi-Hsuan Hsiao, et al., "A Multi-Layer Stackable Thin-Film Transistor (TFT) NAND-Type Flash Memory," *IEEE International Electron Devices Meeting,* pp.1-4, 11-13 Dec. 2006.

[26] Endoh, T., Kinoshita, K., Tanigami, T., et al., "Novel ultrahigh-density flash memory with a stacked-surrounding gate transistor (S-SGT) structured cell," *IEEE Transactions on Electron Devices*, vol.50, no.4, pp. 945- 951, April 2003.

[27] Katsumata, R., Kito, M., Fukuzumi, Y., et al., "Pipe-shaped BiCS flash memory with 16 stacked layers and multi-level-cell operation for ultra high density storage devices," *Symposium on VLSI Technology,* , pp.136-137, 16-18 June 2009.

[28] Wonjoo Kim, Sangmoo Choi, Junghun Sung, et al., "Multi-layered Vertical Gate NAND Flash overcoming stacking limit for terabit density storage," *Symposium on VLSI Technology*, pp.188-189, 16-18 June 2009.

[29] Jiyoung Kim, Hong, A.J., Sung Min Kim, et al., "Novel Vertical-Stacked-Array-Transistor (VSAT) for ultra-high-density and cost-effective NAND Flash memory devices and SSD (Solid State Drive)," *Symposium on VLSI Technology*, pp.186-187, 16-18 June 2009.

[30] Jiyoung Kim, Hong, A.J., Ogawa, M, et al., "Novel 3-D structure for ultra high density flash memory with VRAT (Vertical-Recess-Array-Transistor) and PIPE (Planarized Integration on the same PlanE)," *Symposium on VLSI Technology*, pp.122-123, 17-19 June 2008.

[31] Hang-Ting Lue, Tzu-Hsuan Hsu, Yi-Hsuan Hsiao, et al., "A highly scalable 8-layer 3D vertical-gate (VG) TFT NAND Flash using junction-free buried channel BE-SONOS device," *Symposium on VLSI Technology (VLSIT)*, pp.131-132, 15-17 June 2010.

[32] SungJin Whang, KiHong Lee, DaeGyu Shin, et al., "Novel 3-dimensional Dual

Control-gate with Surrounding Floating-gate (DC-SF) NAND flash cell for 1Tb file storage application," *IEEE International Electron Devices Meeting*, pp.29.7.1-29.7.4, 6-8 Dec. 2010.

[33] Eun-Seok Choi, Hyun-Seung Yoo, Han-Soo Joo, et al., "A Novel 3D Cell Array Architecture for Terra-Bit NAND Flash Memory," *IEEE International Memory Workshop (IMW), 2011 3rd*, pp.1-4, 22-25 May 2011.

[34] Seongjae Cho, Won Bo Shim, Yoon Kim, et al., "A Charge Trap Folded NAND Flash Memory Device With Band-Gap-Engineered Storage Node," *IEEE Transactions on Electron Devices*, Vol. 58, No. 2, pp. 288-295, Feb. 2011.

[35] Seongjae Cho, Il Han Park, Yoon Kim, et al., "A Gated Twin-Bit (GTB) Nonvolatile Memory Device and Its Fabrication Method," *IEEE Trans. Nanotechnol.*, Vol. 8, No. 5, pp. 595-602, Sep. 2009.

[36] Hang-Ting Lue, Tzu-Hsuan Hsu, Min-Ta Wu, et al., "Studies of the reverse read method and second-bit effect of 2-bit/cell nitride-trapping device by quasi-two-dimensional model," *IEEE Transactions on Electron Devices*, vol.53, no.1, pp. 119- 125, Jan. 2006.

[37] B. Eitan, P. Pavan, I. Bloom, et al., "NROM: A novel localized trapping, 2-bit nonvolatile memory cell," *IEEE Electron Device Lett.*, vol. 21, no. 11, pp. 543-545, Nov. 2000.

[38] Tsai, W.J., Zous, N.K., Liu, C.J., et al., "Data retention behavior of a SONOS type two-bit storage flash memory cell," *IEEE International Electron Devices Meeting*,

pp.32.6.1-32.6.4, 2001.

[39] Yen-Hao Shih, Hang-Ting Lue, Kuang-Yeu Hsieh, et al., "A novel 2-bit/cell nitride storage flash memory with greater than 1M P/E-cycle endurance," *IEEE International Electron Devices Meeting,* pp. 881- 884, 13-15 Dec. 2004.

[40] Hang-Ting Lue, Yen-Hao Shih, Kuang Yen Hsieh, et al., "Novel soft erase and re-fill methods for a P^+-poly gate nitride trapping non-volatile memory device with excellent endurance and retention properties," *43rd IEEE International Reliability Physics Symposium, Proceedings.* pp. 168- 174, April 17-21, 2005.

[41] E. Lusky, Y. Shacham-Diamand, I. Bloom, et al., "Characterization of channel hot electron injection by the subthreshold slope of NROM/sup TM/ device," *IEEE Electron Device Letters,* vol.22, no.11, pp.556-558, Nov. 2001.

[42] Larcher, L., Verzellesi, G., Pavan, P., et al., "Impact of programming charge distribution on threshold voltage and subthreshold slope of NROM memory cells," *IEEE Transactions on Electron Devices,* vol.49, no.11, pp. 1939- 1946, Nov 2002.

[43] Yao-Wen Chang, Tao-Cheng Lu, Sam Pan, et al., "Modeling for the 2nd-bit effect of a nitride-based trapping storage flash EEPROM cell under two-bit operation," *IEEE Electron Device Letters,* vol.25, no.2, pp. 95- 97, Feb. 2004.

[44] Jae Young Song, Jong Pil Kim, Sang Wan Kim, et al., “Fin and recess-channel metal oxide semiconductor field effect transistor for sub-50nm dynamic random access memory cell,” *Japanese Journal of Applied Physics: Regular Papers*, Vol. 49, No. 10, pp. 1042021-1042025, Oct. 2010.

[45] Won Bo Shim, Seongjae Cho, Jung Hoon Lee, et al., "Stacked Gated Twin-Bit (SGTB) SONOS Memory Device for High-Density Flash Memory," *IEEE Transactions on Nanotechnology*, Vol. 11, No. 2, pp. 307-313, Mar. 2012.

초 록

본 논문에서는 차단 게이트를 갖는 2 비트 낸드 플래시 메모리에 관한 연구를 진행하였다. 고집적화된 비휘발성 메모리에 대한 수요를 충족시키기 위해서, 차단 게이트를 통해 하나의 워드 라인에서 2 비트 동작이 가능한 gated twin-bit (GTB) 어레이가 설계되었다. 이 구조는 $4F^2$ 면적 당 2 비트를 구현할 수 있다는 점에서 기존의 낸드 어레이보다 메모리 집적도를 두 배 증가시킬 수 있는 구조이다.

1, 2장에서는 최근 비휘발성 메모리의 동향과, 현재 낸드 플래시 메모리의 고집적화를 위해 제안된 여러 해결책들에 대해 소개하였다. 3장에서는 본 논문에서 제안하는 GTB 소자에 대해 소개하고, 시뮬레이션을 통해 쓰기와 읽기 동작 방법을 검증하였다. 쓰기 동작 시에는 차단 게이트를 활용하여 워드 라인 양쪽의 채널 포텐셜이 전달되지 않도록 함으로써 양 쪽의 셀을 따로 동작시키며, 읽기 동작 시에는 드레인 쪽에서의 barrier lowering 효과를 이용한 forward-reverse 방법을 사용한다. 이 외에도, 공정 과정과 어레이를 구성하였을 때의 문제에 대해 검토하였다. 4장에서는 GTB 소자의 공정 진행 내용과 측정된 데이터들을 수록하였다. 제작된 소자의 차단 게이트가 제 역할을 함으로서 twin-bit 동작이 잘 되는 것을 확인하였고, 드레인 전압에 따른 barrier lowering 영향을 검증하였다.

또한, 5장에서는 최근 제안되고 있는 여러 가지 3차원 낸드 구조와 보다

세 배 높은 메모리 집적도를 갖는 gated multi-bit (GMB) 구조를 소개하였다. 이 구조는, $4F^2$ 면적 당 2N 비트를 저장할 수 있는 구조이며, 단결정 실리콘 채널을 이용한다는 장점이 있다. 공정 과정을 소개하였으며 시뮬레이션을 통해 각 층의 동작을 검증하였다.

주요어: 차단 게이트, gated twin-bit (GTB), gated multi-bit (GMB), 3차원 낸드 플래시 메모리, 2 비트 동작

학 번: 2007-21010