Ph.D. DISSERTATION

3-Dimensional NAND flash memory having Tied Bit-line and Ground Select Transistor (TiGer)

비트라인과 그라운드 선택 트랜지스터가 연결된 구조를 가지는 3차원 낸드 플래시 메모리

BY

SEHWAN PARK

August 2014

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

3-Dimensional NAND flash memory having Tied Bit-line and Ground Select Transistor (TiGer)

비트라인과 그라운드 선택 트랜지스터가 연결된 구조를 가지는 3차원 낸드 플래시 메모리

지도교수 박 병 국

이 논문을 공학박사 학위논문으로 제출함

2014 년 7 월

서울대학교 대학원

전기컴퓨터공학부

박 세 환

박세환의 공학박사 학위논문을 인준함

2014 년 7 월

위 원 장 : 박 영 준 (인)

부위원장 : 박 병 국 (인)

위 원 : 이 종 호 (인)

위 원 : 송 기 환 (인)

위 원 : 조 성 재 (인)

ABSTRACT

The demand for flash memory in mobile devices such as smart phone and pads, and digital cameras, portable music players is constantly increasing. Moreover, many hard disk drives in desktop computer are replaced by the solid state drives. This trend accelerates the scaling down of NAND flash memory. As NAND flash memory is scaled down, short channel effect and reliability problems become more severe and further scaling down is faced with process limitations.

To overcome these problems, 3-D stacked NAND flash memory has been introduced for ultra-high-density storage device. Among the proposed structures, channel stacked array has most outstanding electrical characteristic, but has several drawbacks related with decoding difficulty, and additional process steps.

In this dissertation, Channel STacked ARray (CSTAR) having Tied bit-line and Ground Select Transistor (TiGer) structure is proposed and investigated to solve problems of conventional channel stacked array. In case of CSTAR having TiGer, the stacked layers can be distinguished by addressing of common source line (CSL) and bit-line (BL) only. Since ground select transistors (GSTs) are tied with BLs, there are no ground select line (GSL) and GSL decoder. This is main feature of TiGer structure. To execute memory operations successfully, novel program operation scheme for TiGer structure is introduced because the GST of TiGer structure can be handled independently unlike conventional

NAND flash memory. To verify the proposed operation scheme, TCAD simulation is performed.

To fabricate the stacked array, process flow and several unit processes are investigated, and finally 4-stacked single crystalline Si nanowire with GAA structure is fabricated. Si/SiGe epitaxial growth, isotropic SiGe selective etching, stacked nanowire formation, and damascene gate formation are demonstrated, and memory operation for TiGer structure is verified by fabricated device.

Keywords: 3-D stacked NAND flash memory, NAND flash memory architecture, channel stacked NAND flash array, nanowire charge trap flash

Student number: 2009-30189

CONTENTS

Chapter 1

Introduction

1.1 Flash Memory Technology

Demand for flash memory device has grown rapidly in recent years due to the expansion of digital computing beyond desktop computer to mobile devices. With the improvement of microprocessor, recent mobile devices such as smart phones and pads, digital cameras, portable music players, can handle a large amount of information. It means that these products require a large amount of non-volatile memory device to store the big data. Moreover, many hard disk drives (HDDs) in desktop computer are replaced by the solid state drives (SSDs). This trend accelerates the development of flash memory technologies.

Semiconductor memory devices are classified into two categories: volatile and nonvolatile, both based on the metal-oxide-semiconductor (MOS) technology, as shown in Fig. 1.1. The first category is the volatile memories, such as dynamic random-access memory (DRAM) and static random-access memory (SRAM). SRAM, which is most

expensive memory because of large size, is often used as cache memory for the CPU due to its high operation speed. DRAM, which stores a bit of data using a transistor and a capacitor, is used as main memory of computer since it has high density. These memories cannot maintain their data when power is turned off.

The second category is the nonvolatile memories (NVMs). The nonvolatile semiconductor memory includes programmable read only memory (PROM) and mask read only memory (Mask ROM). And, PROM includes electrically programmable read-only memory (EPROM), electrically erasable programmable read-only memory (EEPROM), and flash memory. Nonvolatile memories have the ability to retain stored data for a long time (about 10 years) without any power supply. EPROM can be electrically programmed, but is not electrically erasable. It has to be exposed to ultraviolet radiation for about twenty minutes in order to erase all bits in the memory array. EEPROM can be electrically erasable and programmable in-system, but has higher cost and lower density due to larger area required than EPROM. Flash memory is an EEPROM where the erase procedure of entire chip or sub-array within the chip is performed at one time [1, 2]. Flash memory has high density and high scalability. Due to the erase operation being much faster than the prior EPROM or EEPROM devices, these devices came to be called flash memory.

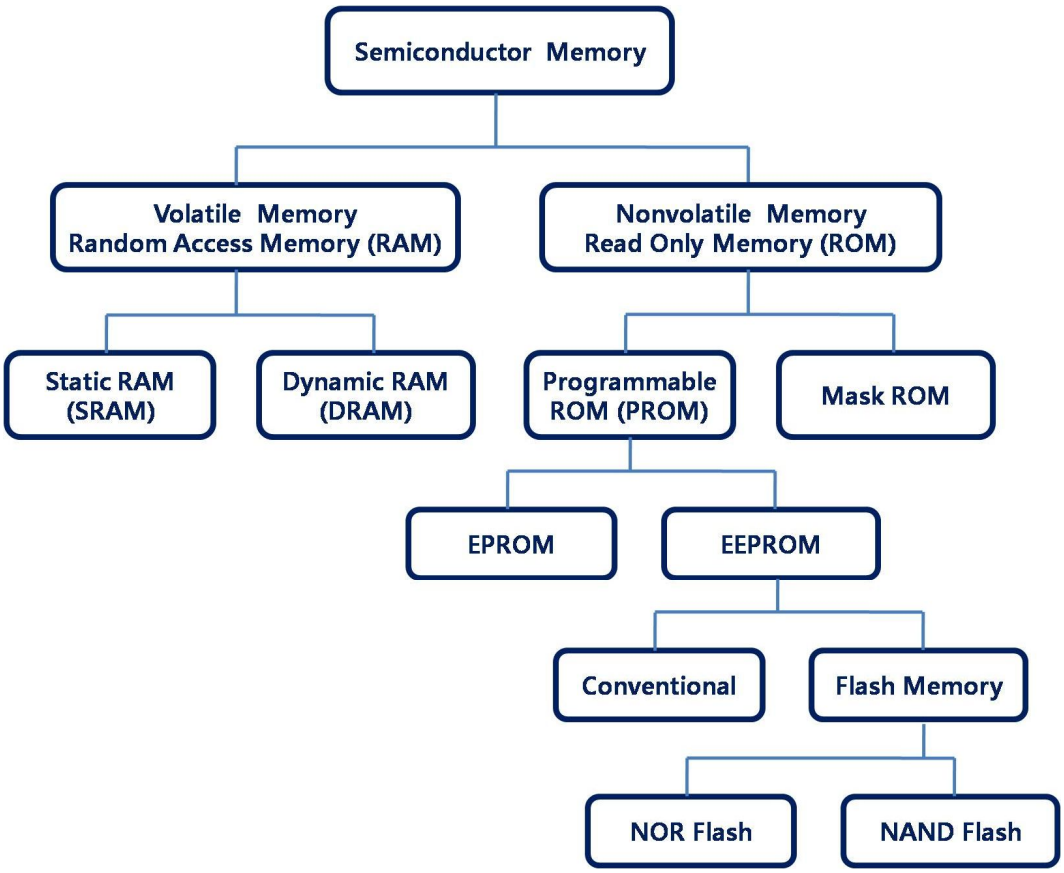


Fig. 1.1. Classification of semiconductor MOS memories.

Toshiba Corp. was a innovator of both NOR-type and NAND-type flash technology in the 1980's. 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. In terms of applications, initially Flash products were mainly used as an "EPROM replacement," offering the possibility to be erased on system, avoiding the cumbersome UV erase operation [1]. Toshiba Corp. introduced NAND flash memory in 1988 which promised lower cost per

bit than NOR flash and faster program and erase throughput.

In recent years, NAND flash memory is the mostly used nonvolatile memory because of its great scalability. Global NAND flash revenue in 2012 is projected to reach $22.9 billion, up from $21.2 billion in 2011. And with NAND flash consumption increasing in the three principal markets for smart phones, tablets and SSDs in ultrabooks, NAND revenue will climb continually during the next few years, hitting approximately $30.9 billion by 2016, as shown in Fig. 1.2 [3].

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 [2]. 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.

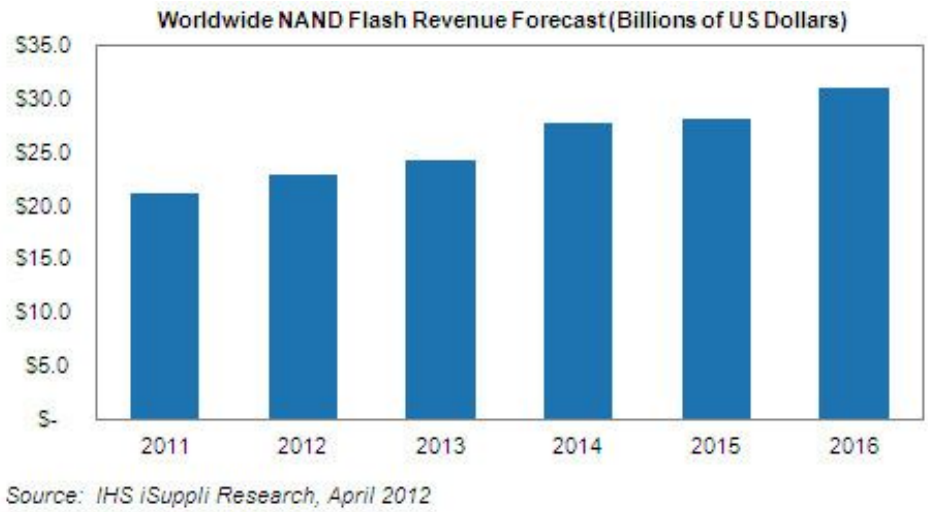


Fig. 1.2. Worldwide NAND flash revenue forecast [3].

As shown in Fig. 1.3 and Fig 1.4 [5], recent mobile electronic devices such as tablets, laptops, or smart phones have started to adopt NAND flash memory as their main data storage device, the demand for low-cost and high density NAND flash memory has been rapidly increasing. 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 durability. An SSD has no moving parts, so it is more likely to keep data safe. Flash drives are gaining traction as mobile storage devices; they are also used as substitutes for hard drives in high-performance desktop computers, laptops and some servers [4].

Fig. 1.3. Applications of NAND flash memory.

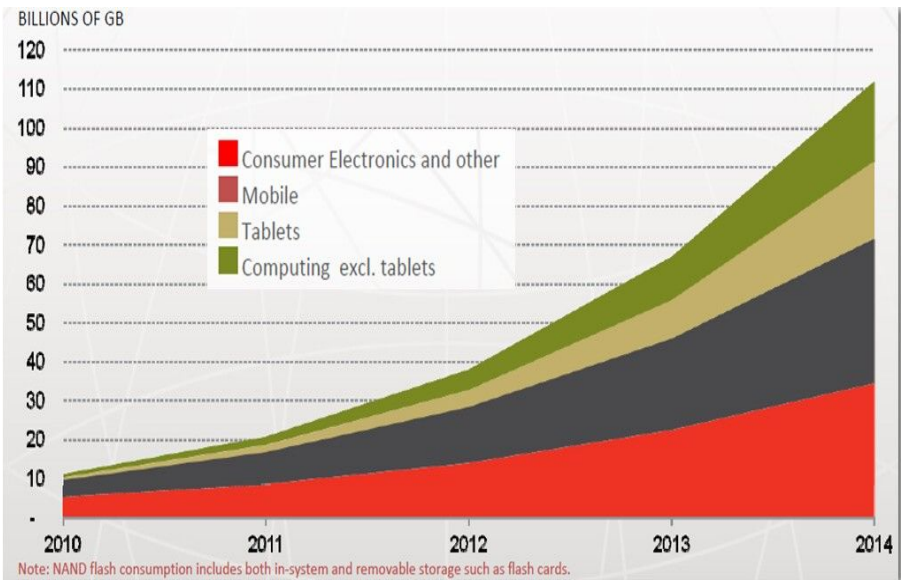


Fig. 1.4. NAND flash memory consumption [5].

1.2 Flash Memory Unit Cell and Array Structure

The most popular NAND memory cell is based on the floating gate (FG) technology. FG memory is used mainly in flash memory industry. Fig. 1.5 shows FG memory cell structure. The floating gate structure usually consists of two-gates; control gate and floating gate. The floating gate is inserted between the control gate and channel of conventional MOS transistor. The floating gate is completely surrounded by oxide, and the isolated gate constitutes an charge trap site. Once the charges are injected into the floating gate, they are retained for a long period of time. To occur the charge injection into the FG or removal from the FG, the electric field is used. When the high electric field is applied on floating gate by coupling, electrons are injected into or escaped from the FG. After the electric field is removed, electrons are retained intact. If the electron is injected, the threshold voltage of transistor increases, and V_{th} decreases as the electron is removed. This operations are called program and erase operation. Whether the electrons are injected or removed, the threshold voltages are switched between two distinct values, the logical "0" (written or programmed state) and logical "1" (erased state), as depicted in Fig. 1.6. The threshold voltage shift caused by the stored charge depends on the amount of charge per unit area, the distance between the control gate and the floating gate, and the permittivity of the blocking oxide layer. The quality and thickness of the oxide layers affect the write and erase characteristics of a memory cell. Usually Fowler-Nordheim

tunneling [7, 8] and hot-carrier injection mechanisms [9, 10] are used to modify the amount of charge stored in the FG.

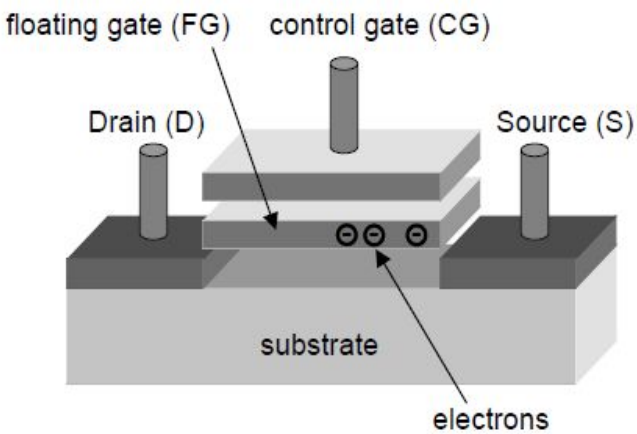


Fig. 1.5 The floating gate memory cell structure

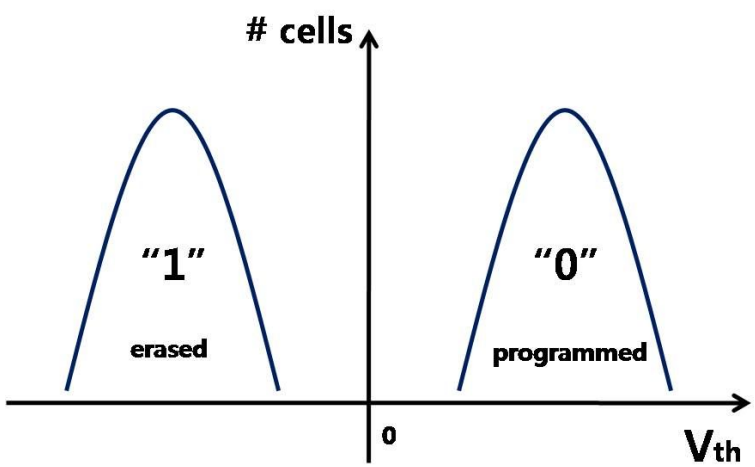
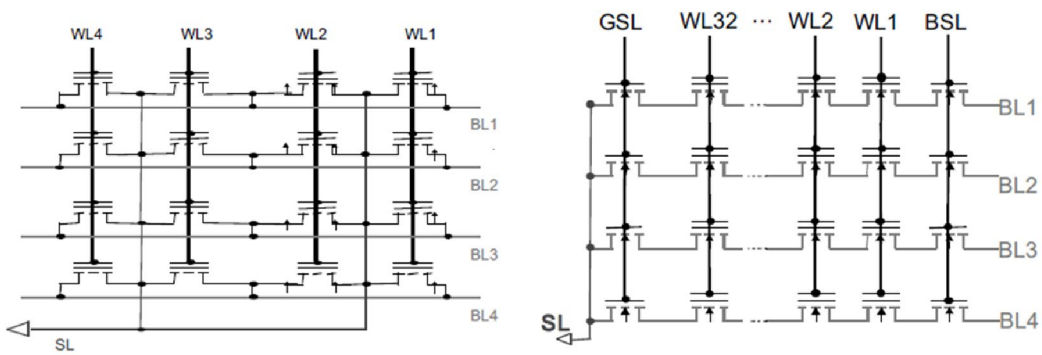


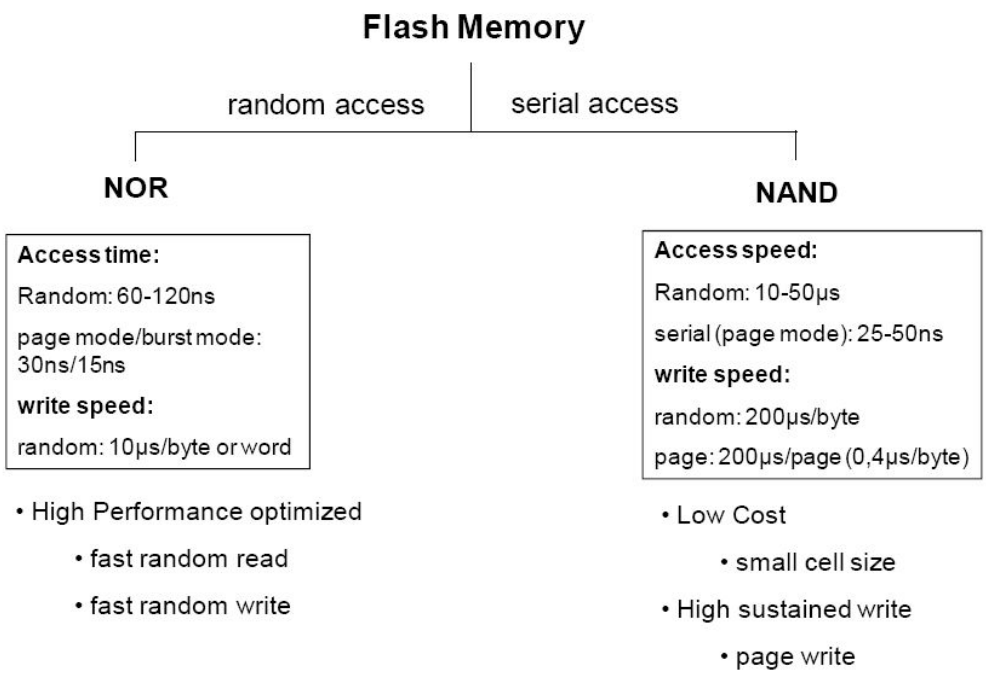
Fig. 1.6 The threshold voltage distribution of a flash memory device.

Fig. 1.7 (a) and (b) show NOR and NAND flash memory array architecture, respectively. In the case of NOR flash memory, the individual memory cells are connected in parallel. Each cell may be accessed via a contact. The direct cell access is the reason for the superior random performance of NOR flash. However, because the source and drain of each cell are all contacted by metal line the NOR cell array requires relatively large cell area. Moreover, because the programming mechanism of NOR flash memory, channel hot electron injection (CHEI), has low injection efficiency, it is impossible to program a large amount of cells at the same time [6, 9, 10]. Thus, NOR flash memory is ideal for performance code storage and excution, which require high speed random read and neglects a cell size.

In contrast, the NAND flash memory is organized serially. As shown in Fig. 1.7 (b), several memory cells in string sandwitched between two select transistors and two contact. Because there are no contacts directrly accessing the memory cells, random performance is slower than NOR flash memory. However, because there are only two contacts in one string, the effective cell size is much smaller than NOR flash memory. The cell size of NAND flash memory is generally in the $4F^2$ range where F is the minimum feature size. Due to the parallel architecture of NOR flash, its cell size is relatively large at $10F^2$. The serial architecture and samll cell size make NAND flash memory optimized for low-cost mass storage.



(a) (b)



(c)

Fig. 1.7. (a) NOR and (b) NAND architecture with circuit symbols

(c) NAND versus NOR [6].

In the NAND string, the cells are connected in series, in group of 64 (or 32). Two select transistors are placed at the edge of the strings for connections to the common source-line (through ground select transistor) and the bit-line (through string select transistor). Control gates are connected through word-lines (WLs).

The NAND flash memory array is organized in pages and blocks. The page is the unit for the simultaneous program and read operation, and the block is the smallest erase unit. Pages are made up by cells belonging to the same word-line. And all the NAND strings sharing the same group of word-lines are erased together, thus forming a flash block. Fig. 1.8 shows NAND Flash memory string and array schematic.

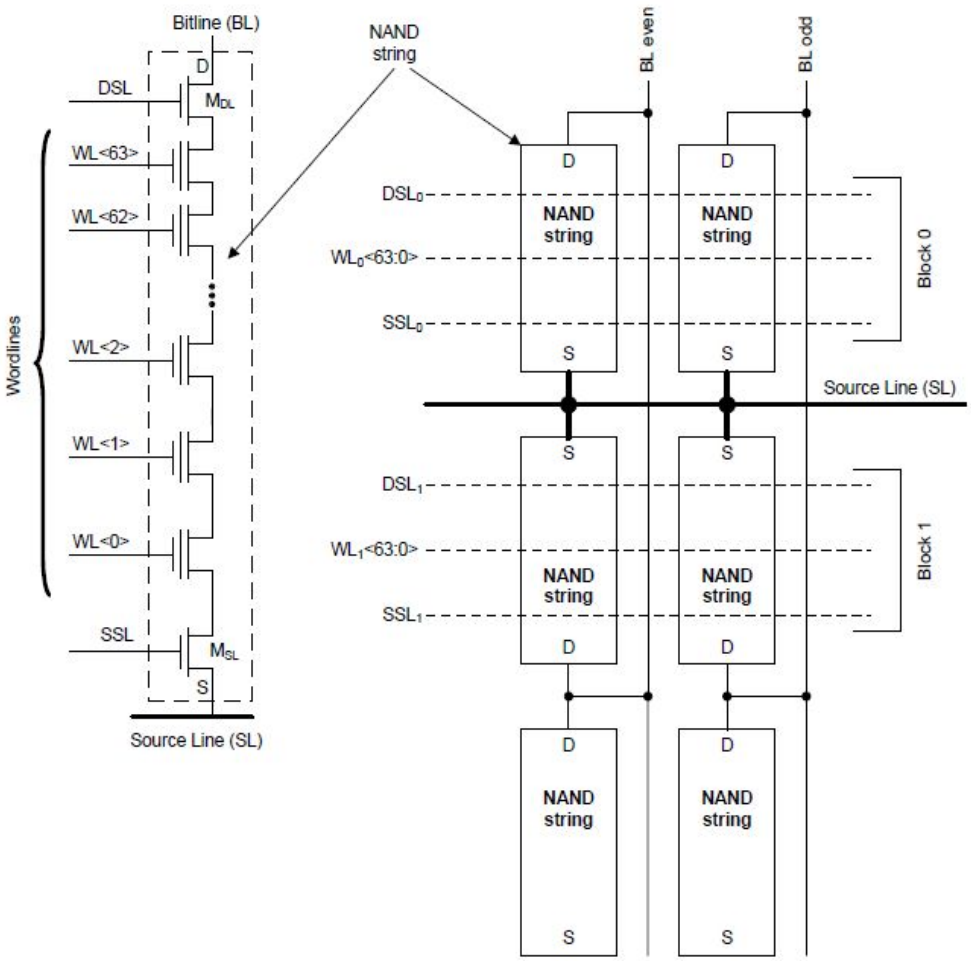


Fig. 1.8. The NAND Flash memory string and array schematic.

1.3 Basic Operation of NAND flash memory

1.3.1 Program and Erase Operation

The program and erase operation in the flash memory is performed by charge transport through gate dielectrics. The Fowler-Nordheim (FN) tunneling is the most widely used transport mechanism in NAND flash memory. When high electric field is applied between the channel and the control gate, the band of oxide is strongly bent. This effectively reduces the tunneling distance through the tunnel oxide barrier, thus the electrons transfer across the thin dielectric to the floating gate as depicted in Fig. 1.9

To induce FN tunneling on selected cells of NAND flash memory architecture, program scheme is investigated. The program operation in the NAND flash memory is performed by a page unit. When the programming voltage (Vpgm) is applied to the word-line of selected page, the channel potential of selected cell must be grounded to occur FN tunneling while that of unselected cell must retain high potential. The so-called self-boosting mechanism prevents an undesired program on the cells sharing the same gate with the addressed cell. When the bit-lines of unselected cells are driven to VDD, string-select transistors are turned-off and the corresponding strings are floating. When Vpass is applied to the unselected word-lines, parastic capacitors boost the potential of the channel. This reduces the voltage drop across the tunnel oxide of unselected cell and, hence, FN

tunneling on unselected cells are inhibited. Fig. 1.10 shows the NAND string biasing for program and inhibition during program operation.

As mentioned above, erase of NAND flash memory is performed by a block unit. In order to extract the electron from the floating gate through tunnel oxide, the high volate is applied to channel and word-lines are grounded. The erase of unselected blocks is prevented leaving their word-lineds floating.

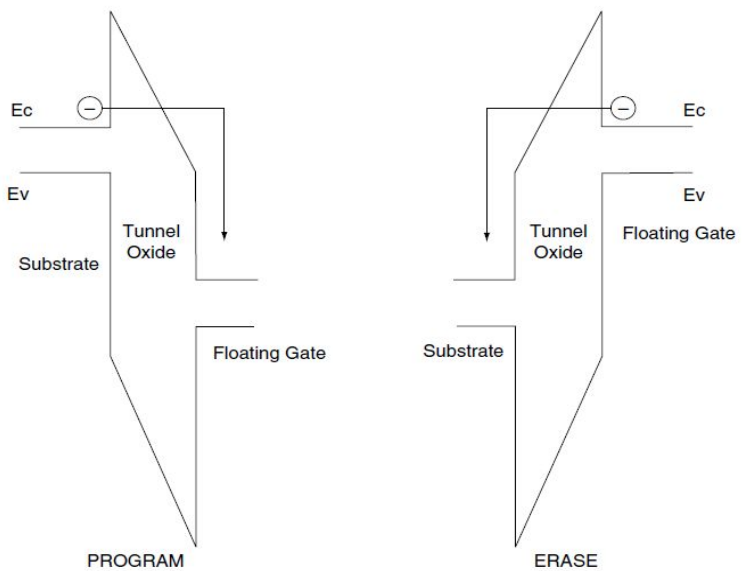


Fig. 1.9. The band diagram of tunneling effec

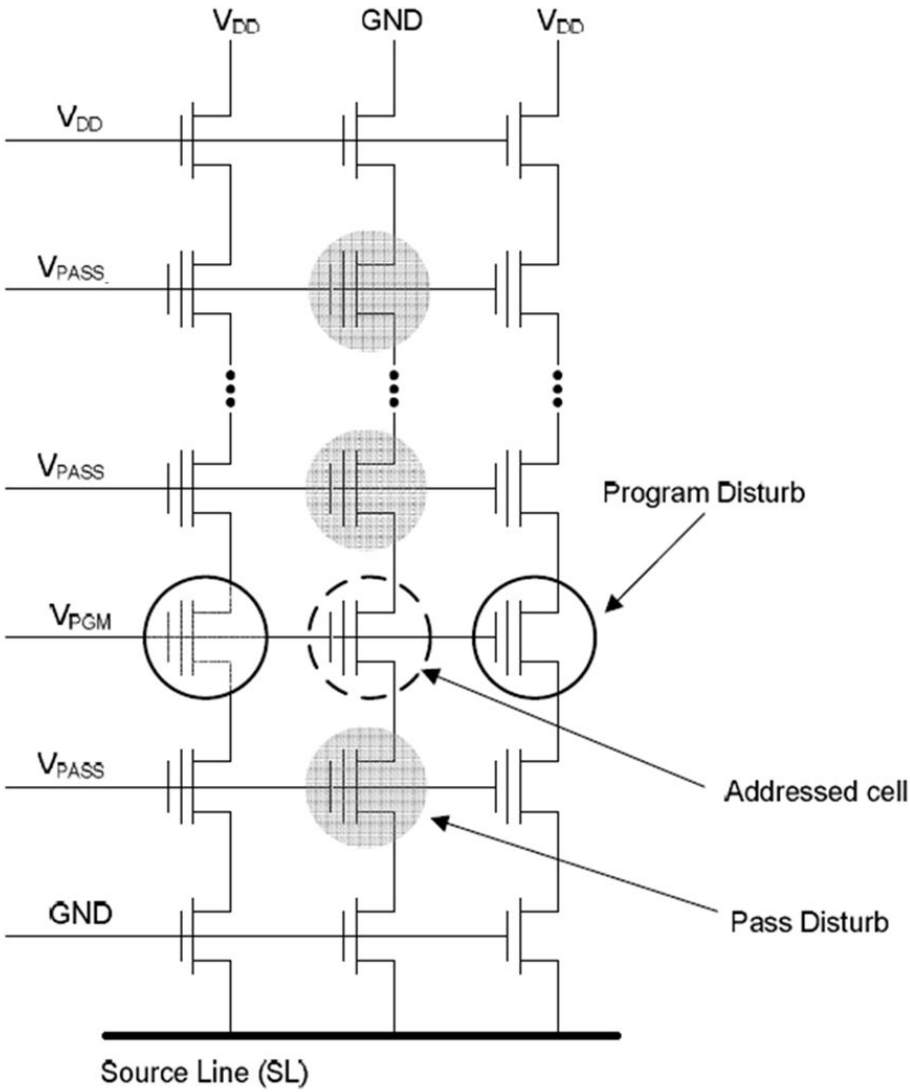


Fig. 1.10. NAND string biasing during program operation

1.3.2 Read Operation

The basic read scheme of NAND flash string is shown in Fig. 1.11. Although NAND flash string consists of several cells connected serially, cells act like usual MOS transistors when gates of usselected cells are biased at high voltage (V_{pass}). The unselected cell turned on by sufficiently high V_{pass} acts as resistor and, thus, bit-line current is a function of selected cell's threshold voltage. If the selected cell has a V_{th} belonging to erased distribution, it contains a logic "1", otherwise, if it belongs to the written distribution, it contains a logic "0".

Historically, the first sensing scheme used the parasitic capacitor of the bit-line. In this technique, the cell state is detected by sensing a voltage of bit-line capasitor. In Fig. 1.12 the basic sensing scheme exploiting bit-line capacitor and the related timing diagram is shown. At the beginning, CBL is charged up since ground select transistor (GST) is turned off. After C_{BL} is charged up to V_{PRE}, C_{BL} is left floating (T0). When the GST is turned on (T1), the string is enabled to sink current (I_{CELL}) from the bit-line capacitor. The cell gate is biased at V_{READ}. As shown in Fig. 1.12, if the cell is erased, the sink current discharges C_{BL}, and the voltage of C_{BL} node is lower than V_{SEN}. In this operation, the coupling of adjacent bit-lines have an effect on the CBL sensing .To solve this problem, the interleaving architecture is introduced. In this architecture, while the even (or odd) bit-lines are read, the odd (or even) bit-lines are grounded, acting as an electrical shield.

Another solution to the bit-line coupling issue is the All Bit-Line (ABL) architecture, as shown in Fig. 1.13. In the ABL architecture, the voltage of SO node is used for sensing instead of BL node. And, the bit-line voltage is maintained during the evaluation step so that the constant BL current flows. Whether BL current flows or not is detected at SO node connected with CSO and sense amplifier. Thus, ABL architecture is called by the "current sensing" as opposed to "voltage sensing" used by interleaving architecture. As all bit lines of selected page are read (or programed) simultaneously, ABL architecture can realize maximum parallelism. The ABL architecture has following advantages over interleaving architecture: evaluation time reduction, lower bit line voltage, simultaneous data process of even/odd cells, and reduction of floating gate interference. Howerver, the ABL architecture needs one sense amplifier for each bit-line while on sense amplifier is connected per two bit-lines.

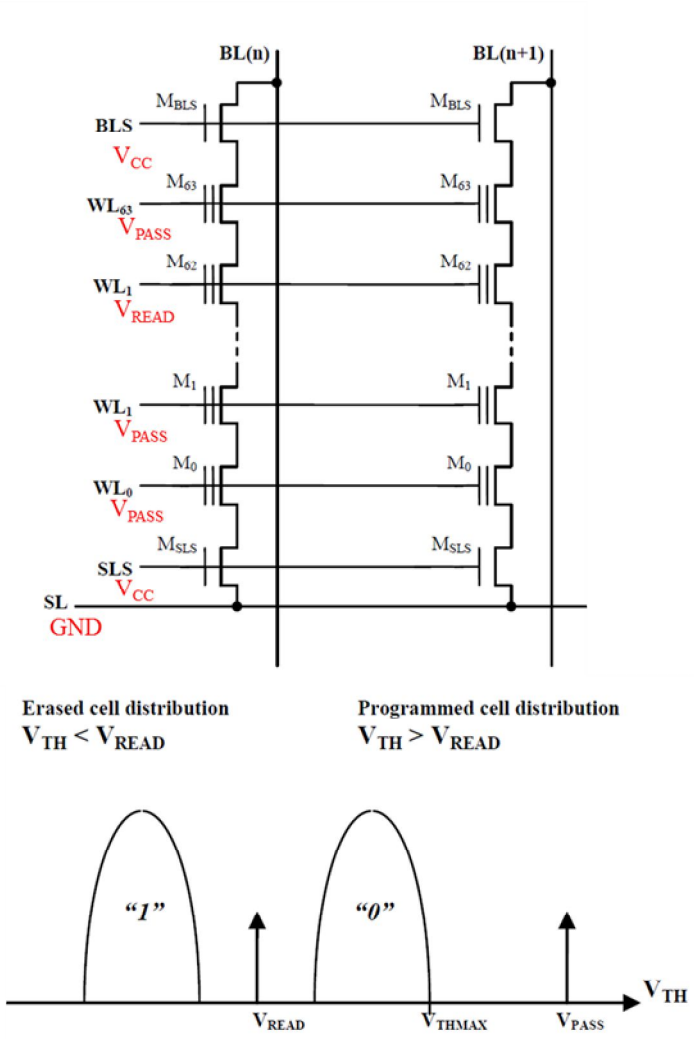


Fig. 1.11. The basic read scheme of NAND flash string

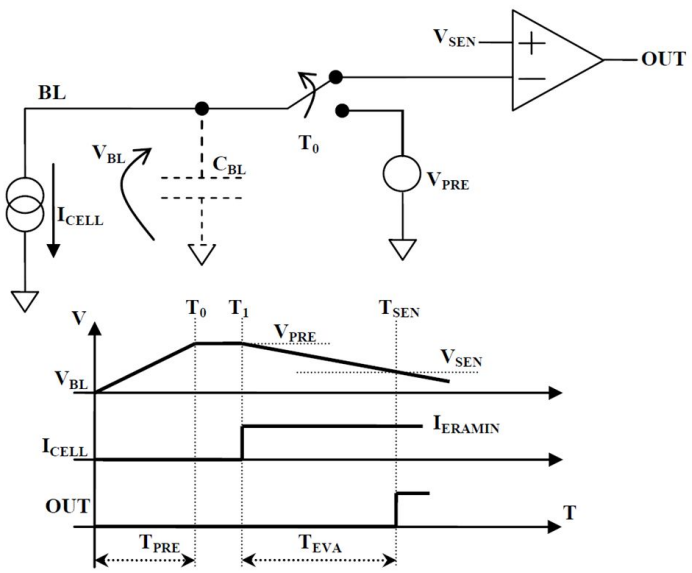


Fig. 1.12. The basic sensing scheme exploiting bit-line capacitor and the related timing diagram

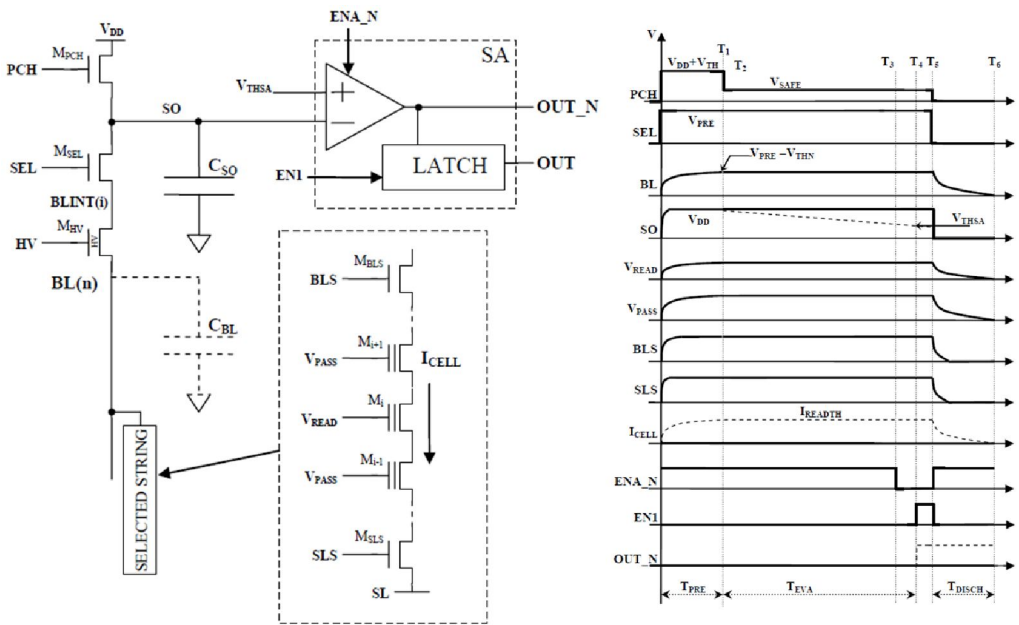


Fig. 1.13. The basic elements of ABL sensing scheme and the related timing diagram

1.4 Charge Trap Flash Memory

In recent years, the charge trap (CT) flash memory has been studied as new technology that could represent a possible alternative to conventional floating gate (FN) NAND flash memory [11-13]. In fact even if FG flash is still the dominant technoloty, several physical barriers seem to limit future scalability such as electrostatic interference among adjacent cells [14, 15]. The charge trap flash memory is a sandwich of tunneling dielectric, charge trapping dielectric, and blocking dielectric. Differently from FG flash which have a semiconductor as storage material, in the case of CT flash, electrons and holes are trapped inside a dielectric material. Due to differnet storage material, CT flash memory is relieved from the disturbance caused by coupling between storage nodes and the stored charge loss caused by stress induced leakage current (SILC) [14-19].

Moreover, the advantage of CT flash from process integration point is higher cell scalability than FG flash memory. From process integration point of view, CT flow provides advantages compared with FG architecture resulting in higher cell scalability. As shown in Fig. 1.14, 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.

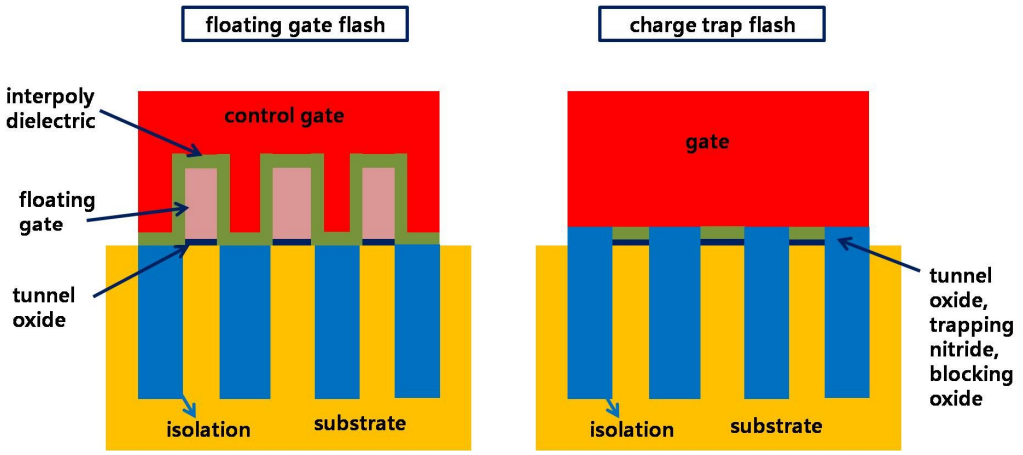


Fig. 1.14. Bit line cross-sectional view of the (a) floating gate and (b) charge trap NAND flash memory

However, more investigation is necessary to improve material properties of charge trap flash memory. The most difficulty issue of CT NAND flash is how to provide both fast FN erase speed and good data retension. Moreover, the threshold voltage saturation in program and erase operation causes typically smaller memory window than conventional FG one. Last critical issue of CT flash is cycling degradation due to oxides worsening caused by the higer current that flows through a CT stack.

Chapter 2

3D Stacked NAND Flash Memory

2.1 Introduction to 3D stacked NAND Flash Memory

In recent years, the demand for low-cost and high density NAND flash memory has been rapidly increasing. As the scaling down of NAND flash memory is accelerated, short channel effect and reliability problems become more severe [20, 21] and more complex and high cost fabrication technology are required for the scaling down [22-24]. Moreover, further scaling down is faced with process physical limitation. Fig. 2.1 (a) shows the gap between required photo lithography and NAND flash memory technology nodes. Up to now reducing the line and space pattern provides the scaling down of cell. To overcome the limitation of photo lithography tools, compensation technologies such as optical proximity correction (OPC), immersion lithography, and double patterning technology (DPT) are introduced [22-26].

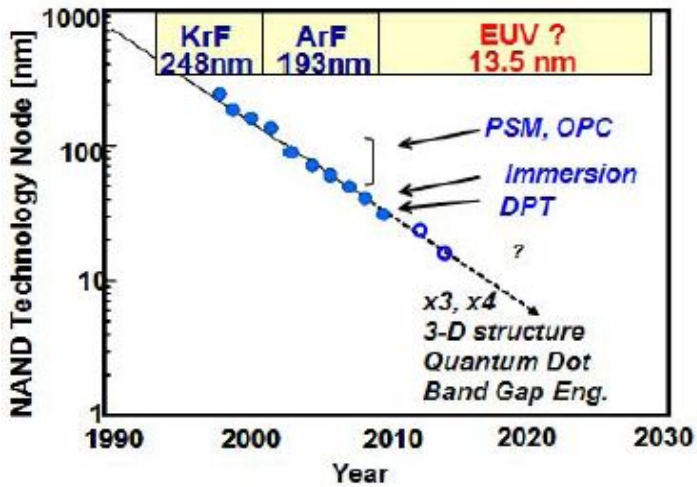


Fig. 2.1. The gap between required photo lithography and NAND flash memory technology nodes [26]

Moreover, the physical limitation blocks further scaling down. The volume of the storage node decreases with shrinking cell size, so that the small number of electrons can induce threshold voltage shift. And it means that the fluctuation of the number of injected charge or charge loss due to bad retention characteristics should have a bad effect on V_{th} shift of cell [27, 28].

Despite of these problems, it is strongly demanded to keep a trend of increasing bit density. However, cost per bit will not scale down near future, as shown in Fig. 2.2, because process cost to shrink of the pattern has steadily increased [29].

To overcome these problems, 3-dimensional (3D) stacked NAND flash memory has been introduced [29-33]. Fig. 2.3 shows the host of new 3D stacked NAND flash

memory [34]. In 2006, the first 3D stacked NAND flash memory is introduced by Samsung Electronics of Korea. In this technology, because each layer requires individual fabrication processes, the cost rises almost proportionally with the number of layers.

In 2007, Toshiba of Japan resolved this problem by developing the bit-cost scalable (BiCS) 3D cell technology that held costs down even as the number of layers increased [30]. The key technology of BiCS is poly-Si vertical channel through the hole on stacked gates. First, many poly-Si gate/SiO_2 dielectric stackes are deposited on substrate. Then, through-holes are punched for channel on gate/dielectric stacks by one photo lithography. The through-holes are filled with gate dielectric and poly-Si channel. As a result, manufacturing cost does not rise very much even if the number of layers increases. In 2009, Toshiba reports the P-BiCS architecture which is modified structure of BiCS to enhance the reliability [35].

In 2009, Tera-bit array transistor (TCAT) based on BiCS technology is presented by Samsung Electronics [31]. Compared with BiCS and P-BiCS, The major advantage of TCAT architecture is that gate replacement process for metal gate and bulk-erase operation are applied.

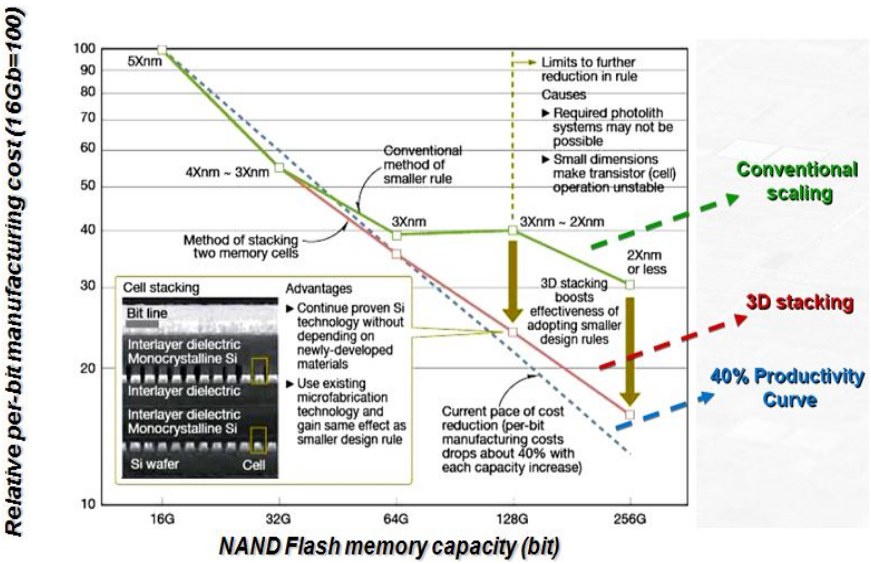


Fig. 2.2 Relative per-bit manufacturing cost as a function of NAND flash memory capacity [29]

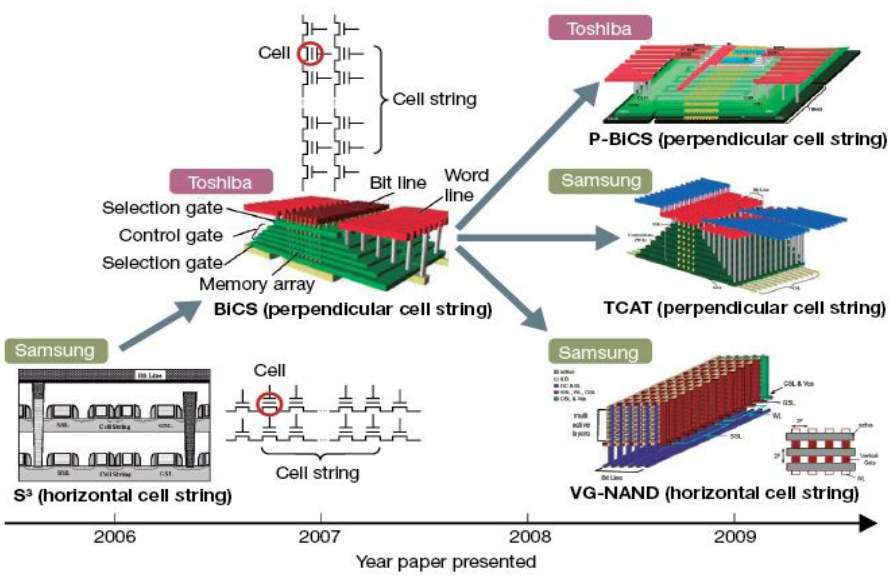


Fig. 2.3. Host of new 3D stacked NAND flash memory technology [34]

2.2 Gate Stack Type NAND Flash Memory

Several 3D stacked NAND flash memory architectures have been proposed so far [29-35]. The 3D stacked NAND flash memory architectures can be devided into two categories. One is gate stack type NAND flash memory, and the other is channel stack type NAND flash memory.

Fig. 2.4 shows the schematic diagram and representative architectures of gate stack type NAND flash memory [31, 35]. The through-holes are punched for channel on gate/dielectric stacks by one photo lithography step regardless of number of stacks. Because the channel is vertically formed through hole, channel material is poly-Si and macaroni structure is applied for uniform electrical characteristics of poly-Si channel.

After the proposal of BiCS technology in 2007, several gate stack type 3D NAND flash memory architectures has benn introduced. There are two representative structures of gate stack type NAND flash memory. One is P-BiCS of Toshiba [35], and the other is TCAT of Samsung Electronics [31] (Fig. 2.4).

As mentioned above, the main feature of BiCS technology is the vertical channel in the hole punched through whole stack of gate plates. Each plate acts as control gate except the top plate and lowest plate which play a role of the select gate. And, a single cell is accessed by the control gate on the string which is selected by a bit line and a select lines. In the case of BiCS structure, common source is formed by diffusion on Si substrate

and the bottom of poly-Si channel is connected common source. Because the channel is foremd after the gate dielectrics are deposited, gate dielectrics located on bottom of channel plug must be etched in order to connect channel with common source. Thus, gate dielectric stack of BiCS structure is N/O, not O/N/O. However the P/E window, the immunity to read disturb and the data retention are insufficient to operate it as a Multi-Level-Cell (MLC). Moreover, high-resistance source line and poor characteristics of lower select transistor due to source doping diffusion during many thermal processes (Fig. 2.5).

To overcome this problem, P-BiCS architecture is introduced. As shown in Fig. 2.6, in P-BiCS, two adjacent NAND strings are connected at the bottoms by so-called pipe-connection (PC) which is gated by the bottom electrode. One of the terminals for the U-shaped pipe is connected to the bit line (BL), and the other is bound by the source line (SL). The SL consists of the meshed wiring of the third level metal and accessed by the first and the second level metal like a conventional planar technology, therefore the resistance of the SL is sufficiently low. Both of the SG transistors are placed below the SL and the BL. The memory stack consists of a blocking layer, a charge trap layer and the oxide-based layer as a tunnel dielectric. For the erase operation of BiCS flash memory, hole current which is generated by gate induced drain leakage (GIDL) near the lower select gate is used.

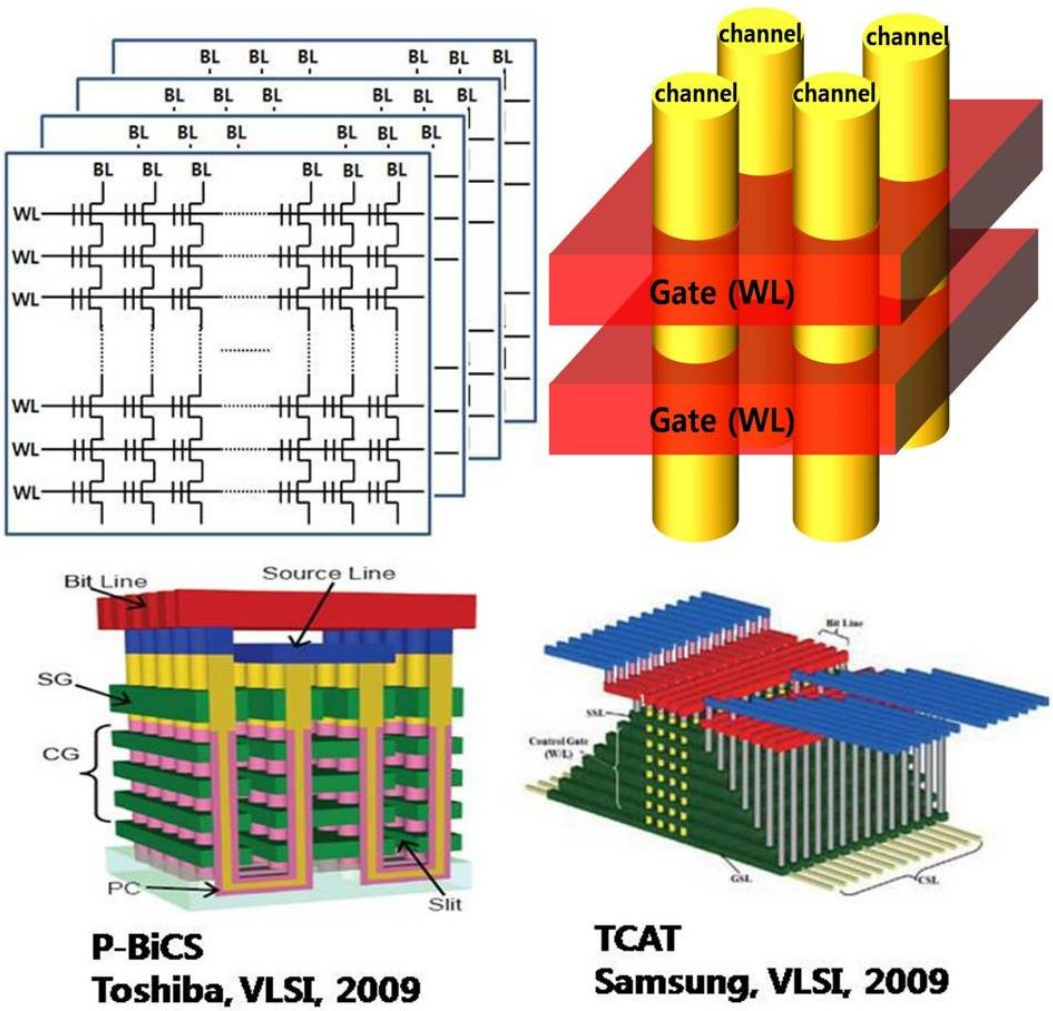
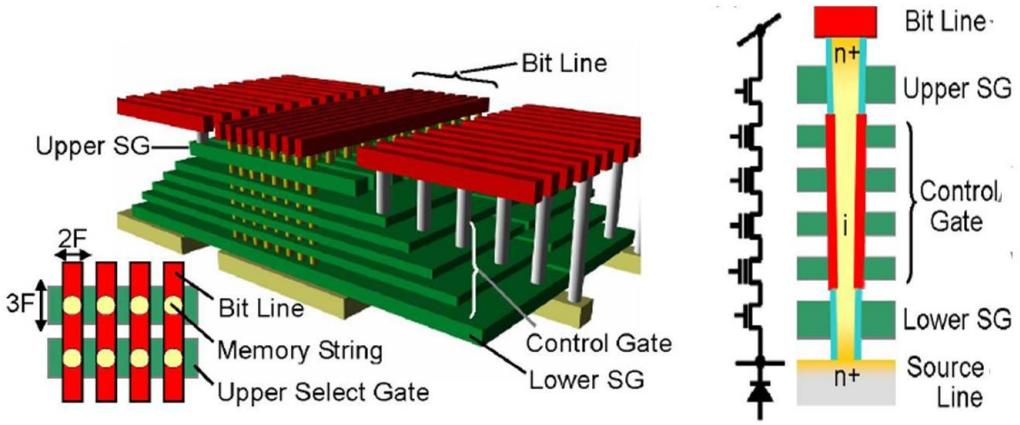
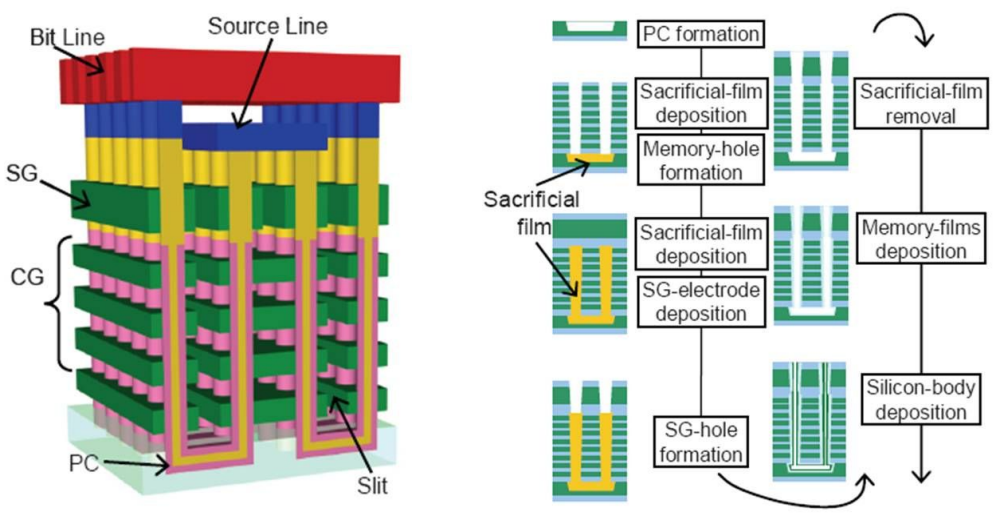


Fig. 2.4. Schematic diagram and representative architectures of gate stack type NAND flash memory [31, 35].



(a) (b)

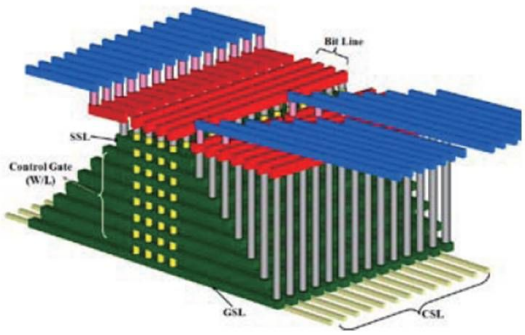
Fig. 2.5. Bit Cost Scalable (BiCS) NAND flash memory by Toshiba. (a) Bird's eye view and (b) schematic diagram of one string [30].



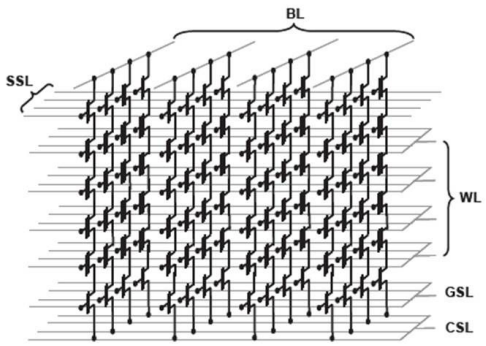
(a) (b)

Fig. 2.6. Pipe-shaped Bit Cost Scalable (P-BiCS) NAND flash memory by Toshiba. (a) Bird's eye view and (b) process flow [35].

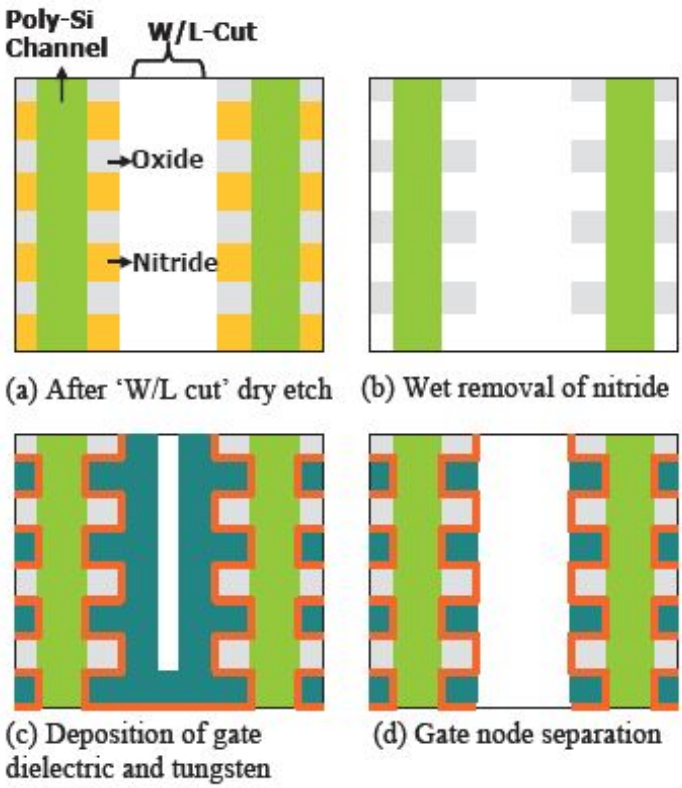
Samsung Electronics presented the terabit cell array transistor (TCAT) based on BiCS technology in 2009 (Fig.2.7). There are two major concerns on the BiCS technology. First of all, it is almost impossible to implement metal gate structure for BiCS architecture because it is very difficult to etch metal/oxide multi-layer simultaneously. To apply the metal gate on gate stack type, gate replacement process is performed in TCAT architecture. Distinctive structural differences with BiCS flash are oxide/nitride multilayer stack, line-type 'W/L cut' etched through the whole stack between the each row array of channel poly plug, line-type CSL formed by an implant through the 'W/L cut', replaced metal gate lines for each row of poly plug. The metal gate provides various advantages such as faster erase speed, wider V_{th} margin, and better retention characteristics. Another concern is that it is impossible to operate bulk-erase for BiCS architecutre. The channel poly plug in TCAT architecture is connected to Si substrate, not n+ common source diffusion layer as in the BiCS architecture. Therefore, conventional bulk erase operation can be achieved, as shown in Fig. 2.7(d) [31].



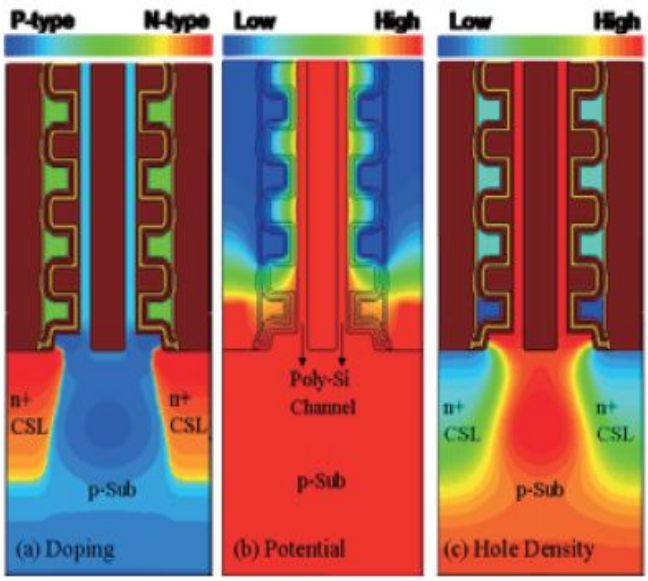
(a) Bird's eye view



(b) Equivalent circuit



(c) gate replacement process flow



(d) Simulated profiles during bulk-erase operation

Fig. 2.7. Terabit cell array transistor (TCAT) by Samsung Electronics. [31]

2.3 Channel Stack Type NAND Flash Memory

Fig. 2.8 shows the schematic diagram and representative architectures of channel stack type NAND flash memory [37, 38]. Channels are stacked horizontally, and bit density can be increased by adding more stacked channel, while the number of the critical lithography steps remains constant because whole stack of channel is dry etched with only one lithography step.

There are two representative structures of channel stack type NAND flash memory (Fig. 2.8). One is VG-NAND architecture [37], and the other is channel stacked array architecture (CSTAR) [38]. Channel stacked NAND flash memory has good pitch scalability thus is very attractive. On the other hand, it is more difficult to decode the bit line in a channel stack architecture, thus decoding innovations are required for a compact array architecture design.

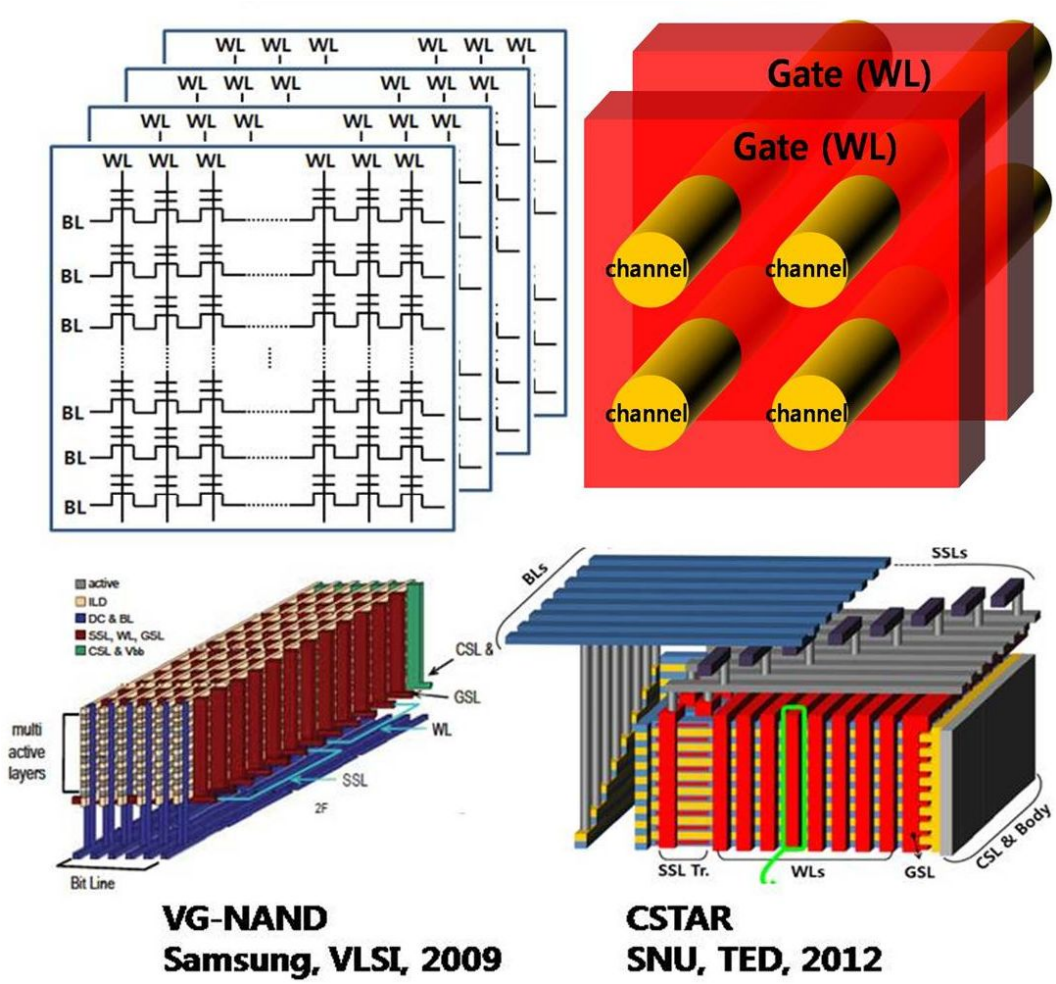
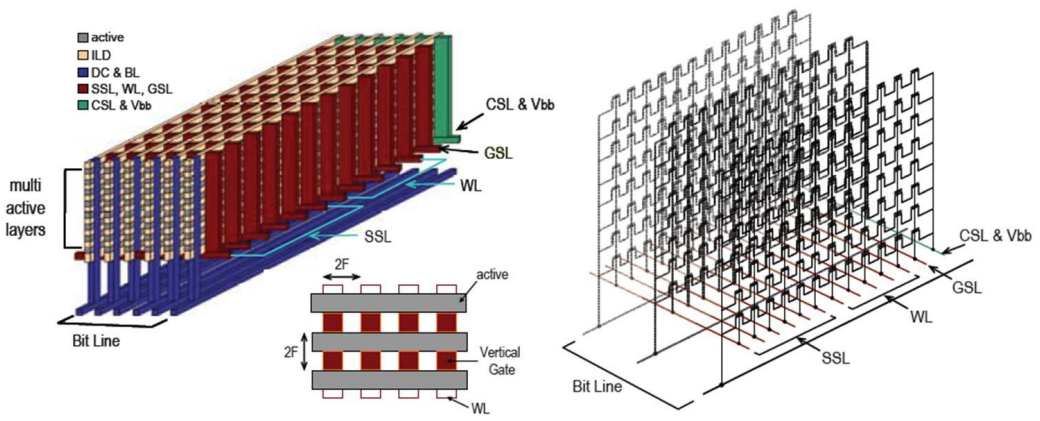


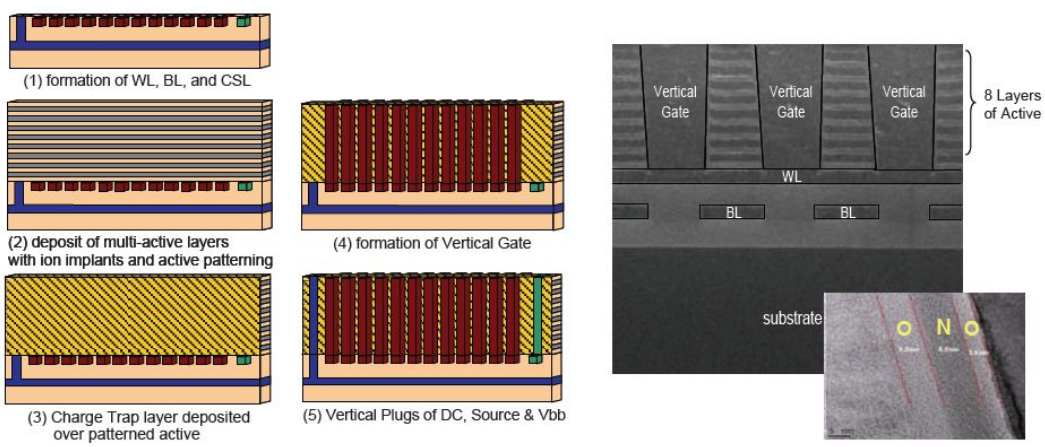
Fig. 2.8. Schematic diagram and representative architectures of channel stack type NAND flash memory [37, 38].

Samsung Electronics presented the vertical gate NAND flash memory (VG-NAND) in 2009 [37]. Fig. 2.9 shows the bird's eye view of the VG-NAND flash and its equivalent circuit. VG-NAND flash includes WL, BL, CSL, horizontal active string with pattern, VG (for SSL, WL, GSL), charge trap layers between active and vertical gate, vertical plugs of DC, source and active body (V_{bb}) as shown in Fig. 2.9. WL and BL are formed at the beginning of fabrication before cell array making interconnect between WL, BL and decoder easier. Source and active body (V_{bb}) are electrically tied to CSL for enabling body erase operation. A positive bias is applied to CSL during erase. Array schematic of each layer is identical to the planar NAND flash except for SSL. VG-NAND requires 6 SSLs for 8 active layers and 8 SSLs for 16 active layers. A reason for the multi-SSLs is to select data from a chosen layer out of multi-layers since our VG-NAND uses common BL and common WL between multi-active layers. Cell size of VG-NAND is $4F^2$ per layer shown in Fig. 2.10 and the active dimension can keep getting down without scaling issue, making the cell size below $4F^2$ per layer. Fig. 2.10 explains integration sequence of VG-NAND and the integration scheme is based on simple patterning and plugging. BL with n^+ poly-Si is fabricated first and then n^+ poly-Si WL is formed on top of it. Multiple active-layers with p-type poly-Si are formed with n-type ion implants for SSL layer selection and alternated inter-layer dielectrics are inserted between active layers. Then patterning is carried on the multiple active layers and charge trap layers (ONO) are deposited over the patterned active layers. Consecutively, VG is formed and connected to WL.



(a) (b)

Fig. 2.9.Multiple layered vertical gate NAND flash memory (VG-NAND). (a) Bird's eye view and (b) equivalent circuit.

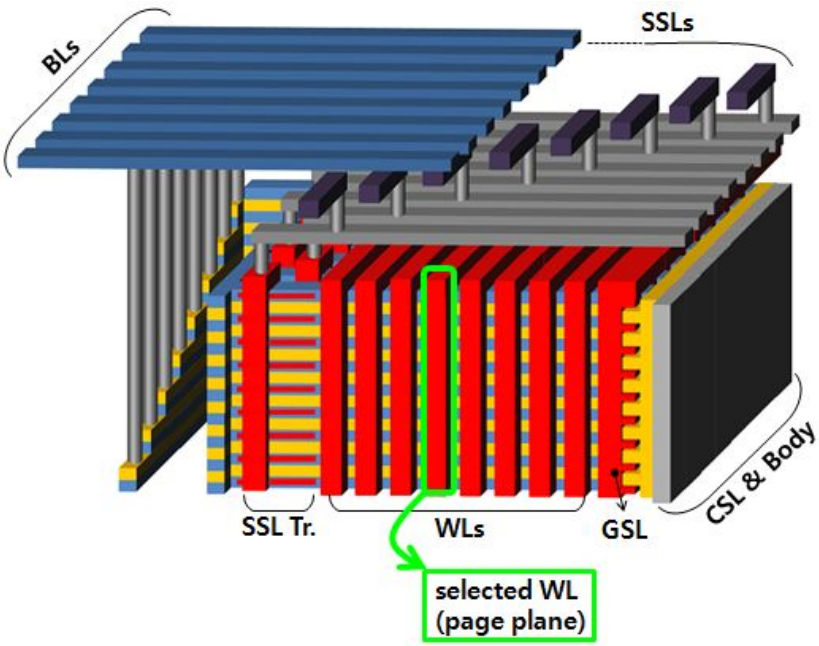


(a) (b)

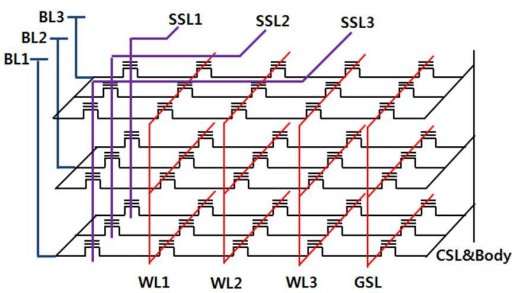
Fig. 2.10. (a) Process flow of VG NAND flash memory. (b) SEM image of the VG NAND flash memory along the WL direction.

Channel stacked array (CSTAR) architecture was presented in 2012 [38]. Fig. 2.11 shows the proposed structure of CSTAR for 3-D NAND flash architecture. This unit structure is defined as "building." The specific descriptions about each part of the building are as follows. The 3-D stacked NAND flash memory, unlike conventional 2-D planar NAND flash memory, needs one more address that is determined by SSLs. In conventional 2-D NAND, the WL selects a "page (line)," whereas the WL of 3-D stacked NAND selects a "page plane." Then, the column and row in the selected page plane are determined by the BL and SSL, respectively. The BLs are formed on the top floor of a building. In addition, BLs are perpendicular to other lines (WLs and SSLs). In a full array, all BLs of each block are connected with those of other blocks to the page buffer. Consequently, there must not be any overlapping with other metal lines or contact holes in the BL region. The WLs and SSLs are parallel with different levels, and they should be connected to a decoder.

The key features of CSTAR are using single-crystalline silicon channel and gate-all-around structure. As mentioned earlier, VG-NAND has poly-silicon channel and double gate structure. The CSTAR has the merit of more cell current drivablity and short channel effect immunity.



(a)



	PROGRAM	READ	ERASE
selected WL	Vpgm	0 V	0 V
unselected WL	Vpass	Vread	float
selected SSL	Vcc	Vread	float
unselected SSL	0 V	0 V	float
selected BL	0 V	Vcc	float
unselected BL	Vcc	0 V	float
GSL	0 V	Vcc	float
CSL&Body	0 V	0 V	Vers

(b)

Fig. 2.11. Channel stacked NAND flash memory (CSTAR). (a) Bird's eye view of the unit building structure and (b) equivalent circuit and operation scheme [38].

2.4 Comparison between Gate Stack Type NAND Flash and Channel Stack Type NAND Flash

Table 2.1 shows the various 3D stacked NAND flash memory architectures [31, 35, 37, 38]. In this chapter we discussed the advantages and disadvantages between gate stack type and channel stack type NAND flash memory.

Gate stack type has poly Si channel. To improve the poor current drivability and device uniformity than single crystalliane Si channel, gate-all-around structure and macaroni channel are applied. In the case of the gate stack NAND flash memory, as more gates are stacked, bit density can be increased but the cell current decreases. Due to low mobility of the poly-Si channel, the number of gate stack layers can be limited by the insufficient cell current issue and low read speed.

The channel stack type NAND flash memory architectures are free from this issue. In the case of the channel stack type, the cell current is not degraded by increasing of stacked layer because stacked layers are channels. Moreover, single crystalline silicon channel can be applied in the channel stack type such as CSTAR and channel stack type NAND flash memory has better pitch scalability than gate stack type (Fig. 2.12). The minimal unit cell size for the gate stack type (P-BiCS and TCAT) should be around $6F^2$ because of the "word line (WL)-cut" process. In addition, since polychannel and core filler are plugged in the punched deep hole for the macaroni channel structure

in P-BiCS and TCAT, the possible minimal feature size is ~50 nm. On the other hand, the unit cell size of channel stack type is smaller than that of gate stack type because top layout resembles the conventional planar NAND Flash memory. However, it is more difficult to decode stacked layers in channel stack type since it is hard to make BLs perpendicular to SSLs by fabrication. Therefore, decoding innovations are required for a compact array architecture design. In case of channel stack, the limiting factor is the finite ONO thickness that restricts the BL pitch.

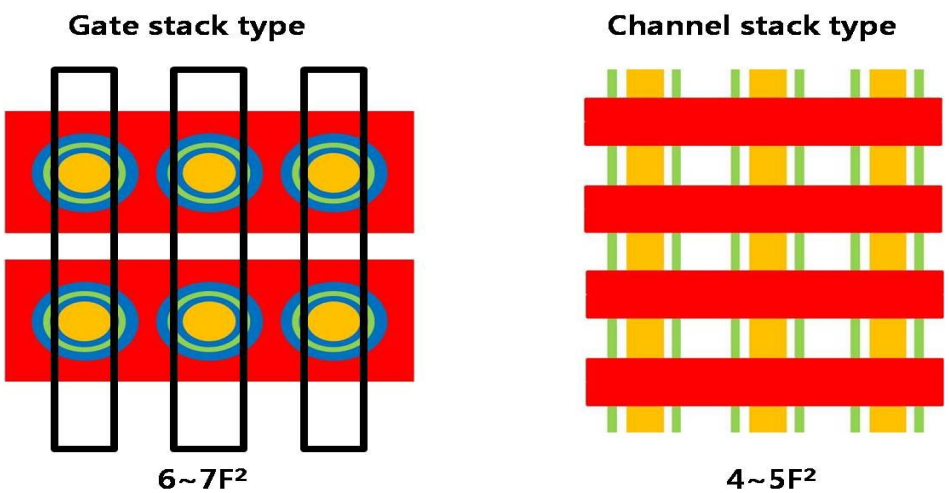
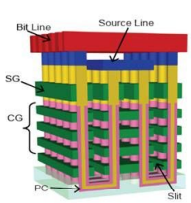
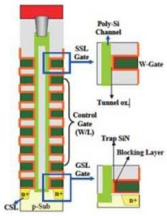
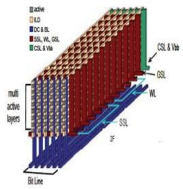
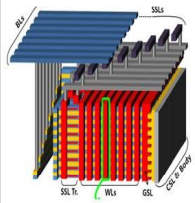


Fig. 2.12. Layout top view of (a) gate stack type and (b) channel stack type NAND flash memory.

Table 2.1. Various 3D stacked NAND flash memory architecture [31, 35, 37, 38].

	Gate stack type		Channel stack type	
	P-BiCS	TCAT	VG NAND	CSTAR
Structure				
Cell structure	Macaroni & GAA	Macaroni & GAA	Double gate	Gate-All-Around
Channel	Poly Si channel	Poly Si channel	Poly Si channel	Single crystalline Si channel
Key issue	Low read current & Reliability	Low read current & Reliability	Number of lithography, implantation	Layer decoding Layer separation

Chapter 3

Channel STacked ARray (CSTAR) having Tied bit-line and Ground Select Transistor (TiGer)

In this chapter, channel stacked NAND flash memory having Tied bit-line and Ground Select line Transistor structure (TiGer) is proposed. The characteristics of 3-D Channel STacked ARray (CSTAR) having TiGer are studied by multidirectional investigation, and the TCAD simulation is performed to verify the operation scheme.

3.1 3-D NAND Flash Memory Architecture Design of Channel STacked ARray having Tied Bit-line and Ground Select Transistor

Fig. 3.1 shows the bird's eye views of gate stack type and channel stack type NAND flash memory. In the case of gate stack type, common word-line (WL) gate plates are stacked up above ground select-line (GSL) gate plate and source select-line (SSL) gate plate is stacked at the top. SSL gate plate can be patterned easily by photo

lithography. And bit-lines (BLs) are patterned perpendicularly to SSLs. Thus, the string on cross-point of selected BL and SSL biasing is selected one. In the case of channel stack-type, it is more difficult to make BLs perpendicular to SSLs. To overcome this problem, a novel layer selection method and stacked NAND flash architecture is proposed. In the case of CSTAR, the string in same floor is connected same bit line and island-type SSLs selects the channel. And multi-layered metal line for BLs and SSLs are introduced to reduce the array size. The CSTAR with layer selection by multi-level operation (LSM) adopts several SSL having different threshold voltages to enable the layer selection.

In this dissertation, the CSTAR having Tied bit-line and Ground select transistor (TiGer) structure is proposed. Fig. 3.2 shows the bird eye's view of the proposed structure. Channels are stacked in a horizontal direction and each BL is connected to stacked channels in a line. The number of common source-lines (CSLs) is the same as stacked layers, and each CSL is connected to channels on the same floor. Word-lines are formed vertically and a SSL is formed in parallel to WLs, and island-type ground-select transistors (GSTs), which are separated in the lateral direction, are applied. To select a channel, GST and CSLs are perpendicularly crossing each other. The main feature of TiGer structure is that each island-type GST is connected to BL.

In spite of the introduction of three-dimensional NAND flash structure, it is difficult to make 3-D metal line because peripheral circuits are still two-dimensional. Therefore, compact design of metal lines of 3-D NAND flash memory is required. If the

GST is not connected to BL, the island-type GSTs demand individual GSLs which are parallel to BL. In order to reduce the array size, BL and GSL are combined. As GST is tied to BL, TiGer structure cannot accomplish same operation as the conventional NAND flash memory. Therefore, novel operation scheme is applied to proposed structure.

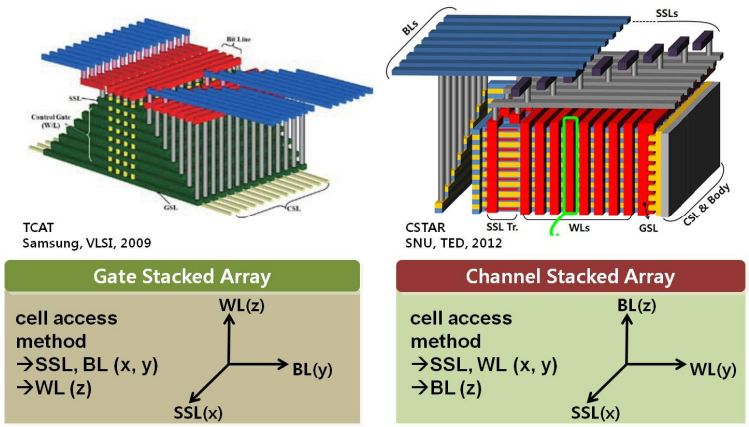


Fig. 3.1. Bird's eye view of the gate stacked array and channel stacked array

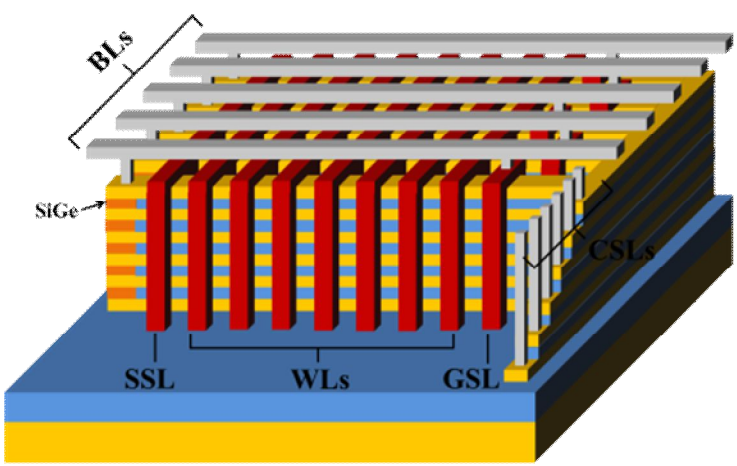


Fig. 3.2. Bird's eye view of CSTAR NAND flash memory having TiGer structure

3.2 Operation Scheme of Channel Stacked ARray having TiGer structure

Fig. 3.3 shows the equivalent circuit of CSTAR having TiGer structure. BLs separated in a horizontal direction are connected to stacked channels, and channels on the same floor share the CSL. One SSL and multiple WLs are connected in a vertical direction, and GSTs tied with BL are separated in a horizontal direction. Because GST cannot be controlled independently, CSTAR having TiGer structure requires novel operation scheme. In this chapter, operation schemes for TiGer structure are introduced and they are verified using TCAD simulation.

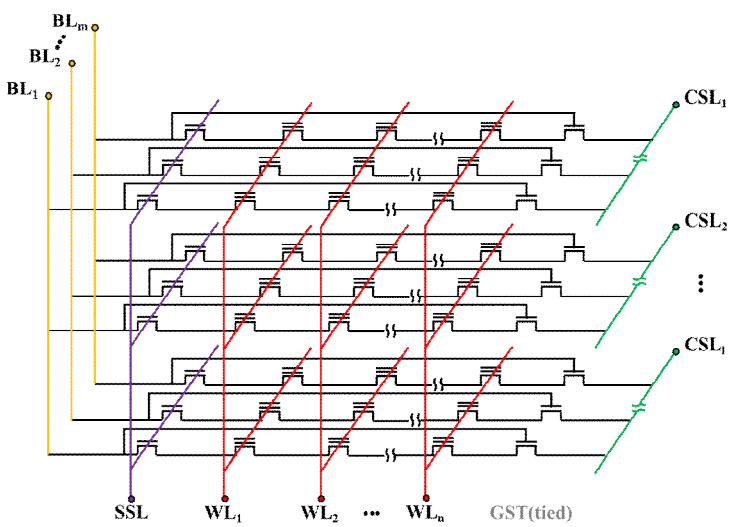


Fig. 3.3. Equivalent circuit of CSTAR NAND flash memory having TiGer structure

3.2.1 Erase operation of CSTAR having TiGer structure

In conventional NAND flash memory, erase operation is performed using Fowler-Nordheim (FN) tunneling. All word-lines are set to 0 V, and a high bias is applied to the channel through the body contact to induce FN tunneling. However, it is hard to make body contact at GAA nanowire channel. Thus, in the case of BiCS flash, for the erase operation, hole current generated by GIDL near the select gate is used to raise the body potential. And this scheme can be applied to TiGer structure for erase operation. Fig. 3.4 shows the timing diagram of erase scheme using GIDL current and the simulated channel potential during the erase operation.

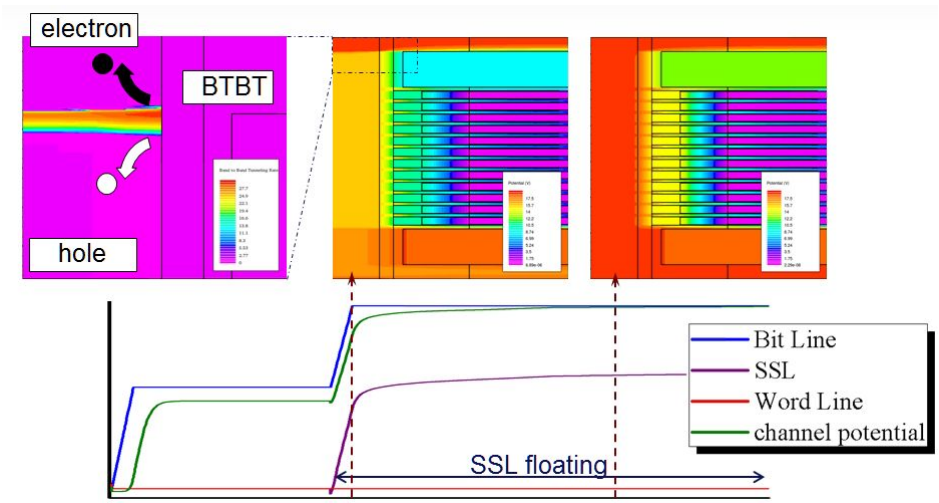
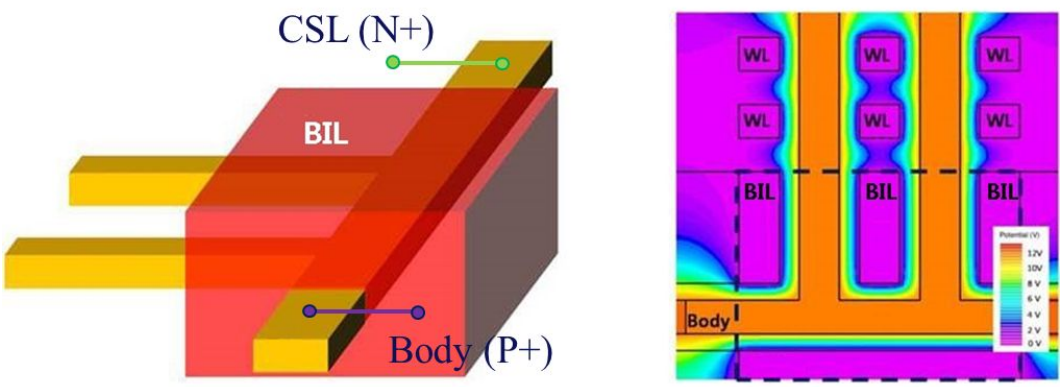


Fig. 3.4. The timing diagram and channel potential of erase scheme using GIDL current

Meanwhile, applying the additional select-line for body contact, bulk erase scheme in TiGer structure is available [38]. Fig. 3.5(a) shows the design of CSL and additional select line, which is named body-isolation line (BIL), for body contact. During program and read operation, BIL is turned on consistently to form inversion region like virtual source and body contact is floated. In the erase operation, BIL is floated and raised body potential transfers to channel (Fig. 3. 5(b)).

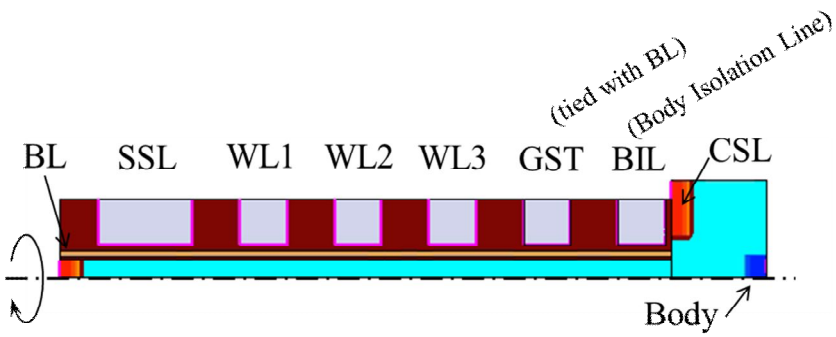


(a) (b)

Fig. 3.5. (a) BIL design for body contact (b) Simulated channel potential during bulk erase operation

Fig. 3.6 shows the simulated structure and device parameters. The 2D simulation is performed using cylindrical coordinates. In this structure, GST tied with BL and BIL for body contact are applied. Fig. 3.7 shows the channel potential during bulk erase operation and the trapped charge density on nitride layer after erase operation. The high

potential is applied at body and WLs are grounded while BL and GST, and BIL are left floating. Since the body potential transfers to channel, FN tunneling between the channel and the WLs is occurred and cells are erased successfully. The simulated transfer curve after erase operation is depicted on Fig. 3.8.



Device parameters	
Channel radius (r)	20 nm
Bottom Ox/ Nitride/ Top OX (O/N/O)	3/ 6/ 6 nm
WL gate length/space (L_g /L_s)	50/ 50 nm
SSL gate length/space ($L_{SSL,g}$ /$L_{SSL,s}$)	100/ 50 nm
GST gate length/space ($L_{GST,g}$ /$L_{GST,s}$)	50/ 50 nm
Body doping concentration (N_{body})	B, 5 x 10^{17}/cm^3
Inversion-type S/D structure	
Cylindrical coordinates	
Synopsys sentaurus TCAD simulation	

Fig. 3.6. The simulated structure and device parameters

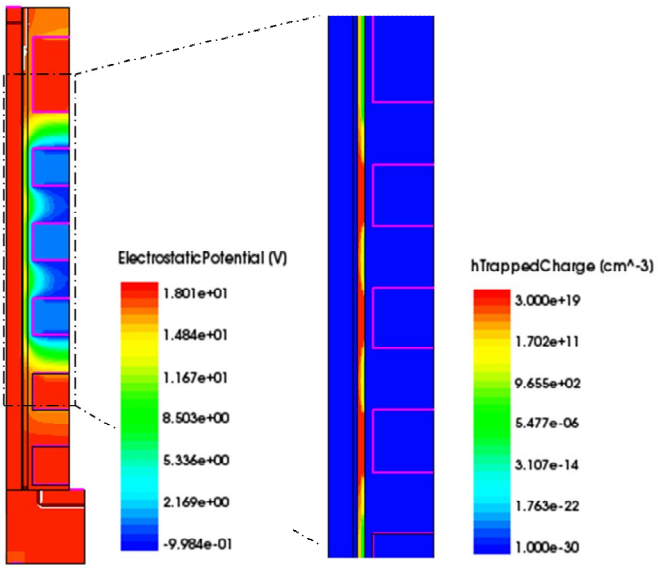


Fig. 3.7. Channel potential during the body erase operation and traped hole density on nitrde after erase operation

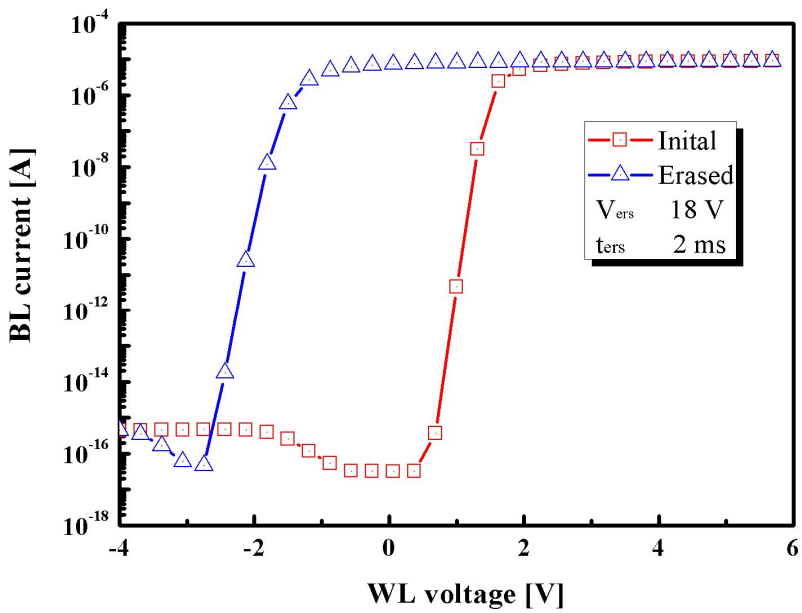


Fig. 3.8. The simulated transfer curves of initial and erased cell

3.2.2 Program operation of CSTAR having TiGer structure

The program operation in NAND flash memory is performed using FN tunneling induced by potential difference between channel and word-line. During the program operation, string containing selected cell maintains ground potential and word-line is set to high voltage. In order to avoid programming unselected cells, other strings are boosted to high potential by self-boosting scheme. The program operation consists of pre-charge stage to enhance the efficiency of self-boosting and program stage. In the case of conventional NAND flash memory, pre-charging is performed through BL. And SST becomes turn-off state as BL potential increases while GST is turned off consistently. (Fig. 3.9(a)) The floated channel can be boosted to high potential by coupling of WL potential. In the case of TiGer structure, if the same operation as conventional one is applied, GST became turn-on as BL potential increases and unselected string is not floated. Thus, pre-charging of TiGer structure is carried out through CSL and BL tied with GST is biased at Vcc. GST becomes turn-off as CSL potential increases and SSL is turned off constantly.

As mentioned in the previous chapter, for bulk erase operation BIL is applied to TiGer structre. In the program operation, BILs play a important role in preventing pre-charging of unselected strings. In the program operation of TiGer structure, if BILs don't exist, lots of blocks are pre-charged at the same time since CSL is connected to all strings

on the same floor. To reduce power consumption and RC delay, BILs of unselected strings must be tuned off. This pre-charge scheme of TiGer structure is depicted on Fig. 3.9(b).

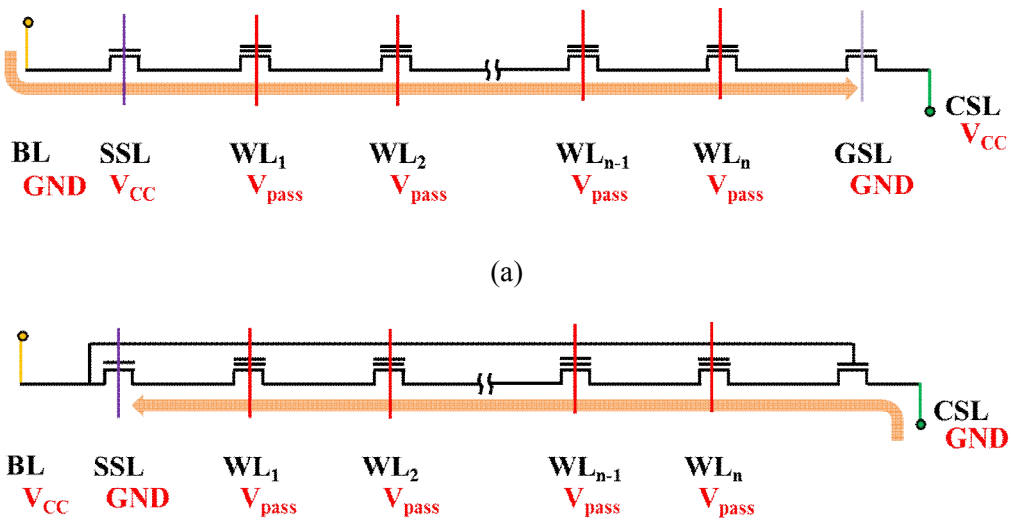


Fig. 3.9. The pre-charge phase of (a) conventional NAND flash memory and (b) CSTAR having TiGer structure

In planar NAND flash memory array, the only condition to decide selected string is whether BL is biased at V_{cc} or ground. However, in the case of 3-D NAND flash memory, several inhibition conditions exist. In Fig. 3.10, selected cell and inhibition cells are marked on equivalent circuit; inhibited cell 1: the cell sharing BL with selected one, inhibited cell 2: the cell sharing CSL with selected one, inhibited cell 3: the cell sharing

none with selected one.

In the case of inhibited cell 1, as mentioned above, GST tied with BL and CSL is biased at V_{cc} while SSL is turned off. In the case of inhibited cell 2, gate bias of GST must be 0 V since the CSL is shared with selected cell and biased at ground. Although BL and CSL of inhibited cell 2 is biased ground, the channel is boosted since SST and GST are turned off. And channel of the inhibited cell 3 is also boosted because SST and GST are turned off. The timing diagram of program operation and simulated channel potentials of selected/unselected string are depicted on Fig. 3.11 and Fig. 3.12.

Fig. 3.13 shows the simulated transfer curves of selected cell and inhibited cells after programming. The program disturbance is observed to the inhibit cells and the disturbance of inhibit cell 2 is larger than that of other cells. In the case of inhibit cell 2, during the self-boosting the channel potential is highly boosted while BL and CSL are grounded. Thus, the leakage current through select transistors, which decrease the efficiency of self-boosting, is larger than that of others. To reduce this phenomenon, the threshold voltage of GST is large enough to prevent leakage current and the distribution of V_{th} of GST must be well controlled.

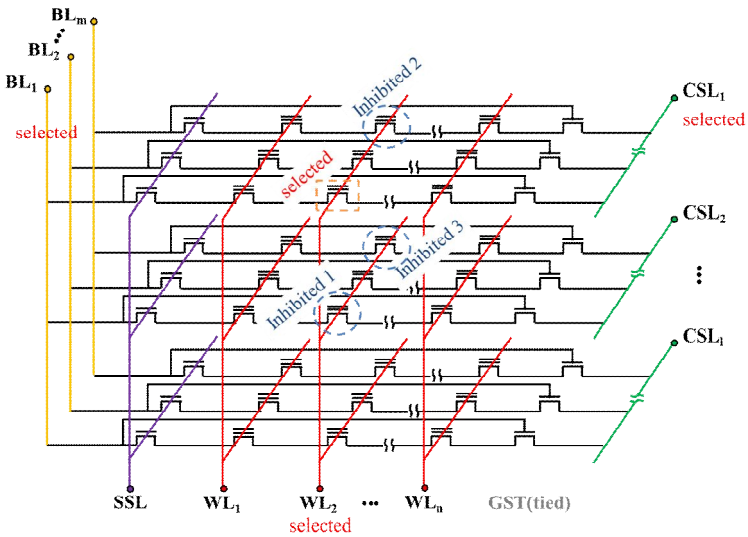


Fig. 3.10. The selected and inhibited cells are marked on equivanlent circuit of CSTAR having TiGer structure

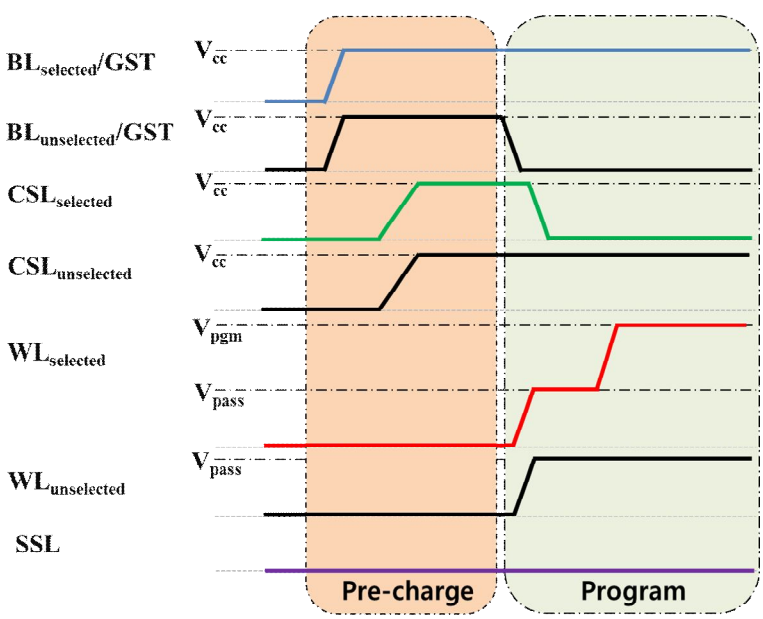


Fig. 3.11. The timing diagram of program operation for TiGer structure

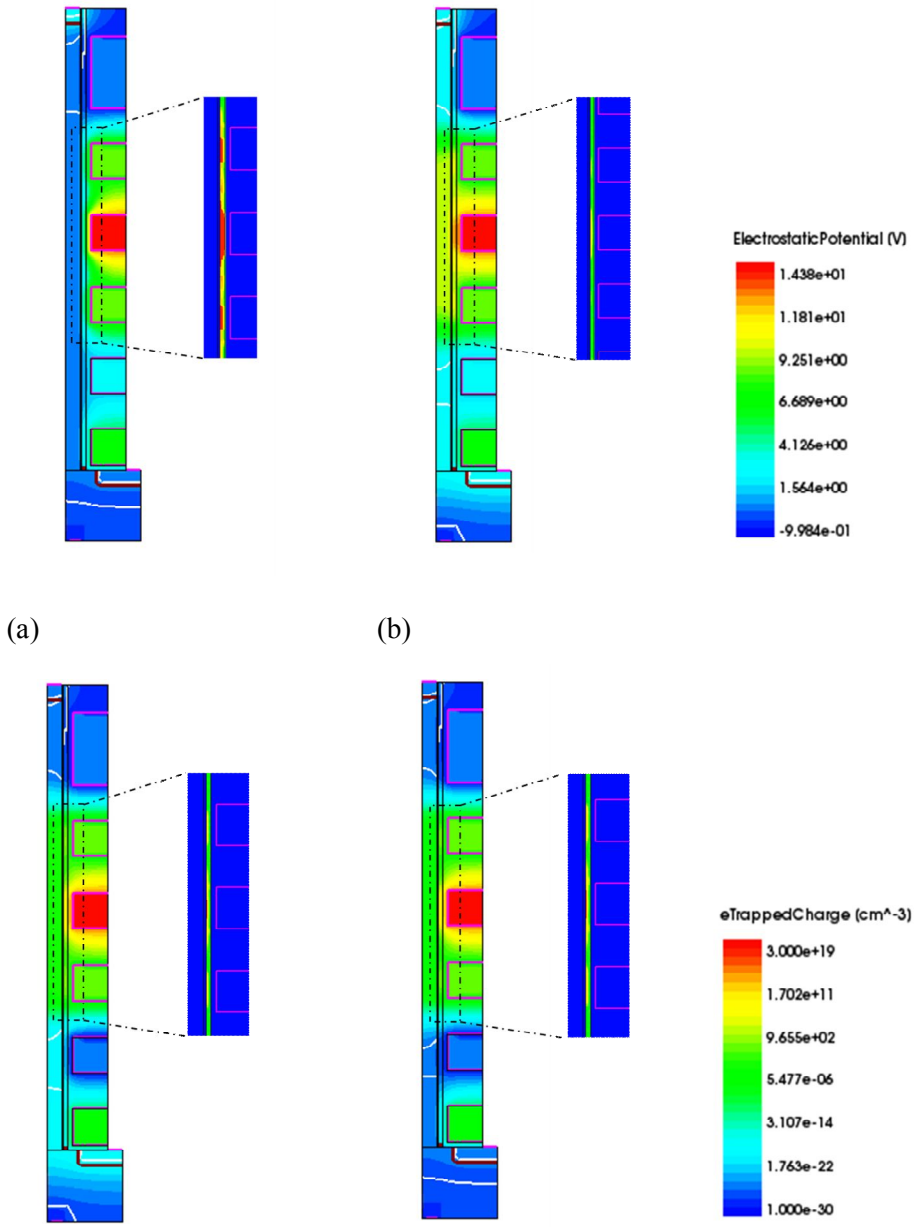


Fig. 3.12. The channel potential and trapped charge concentration of (a) selected cell, (b) inhibit cell 1, (c) inhibit cell 2, (d) inhibit cell 3

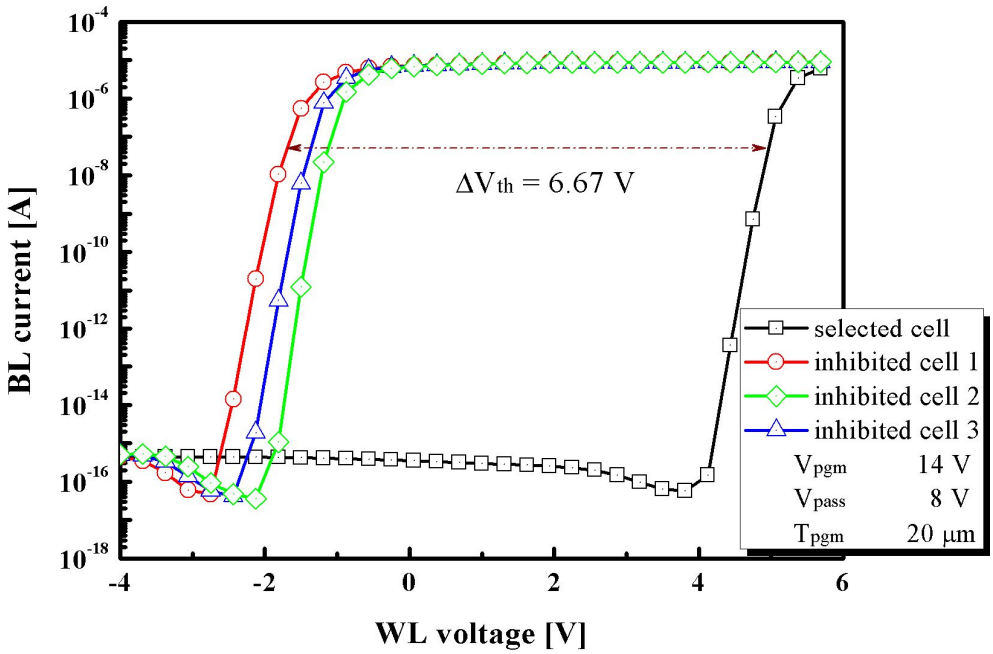


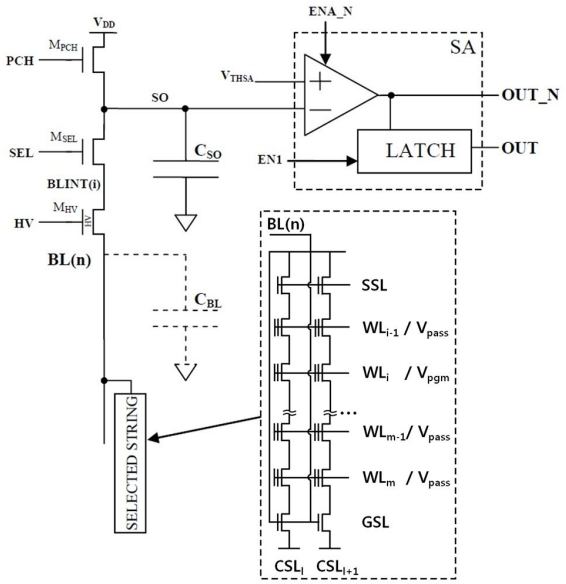
Fig. 3.13. The transfer curves of selected and inhibited cells after program operation

3.2.3 Read Operation of CSTAR having TiGer structure

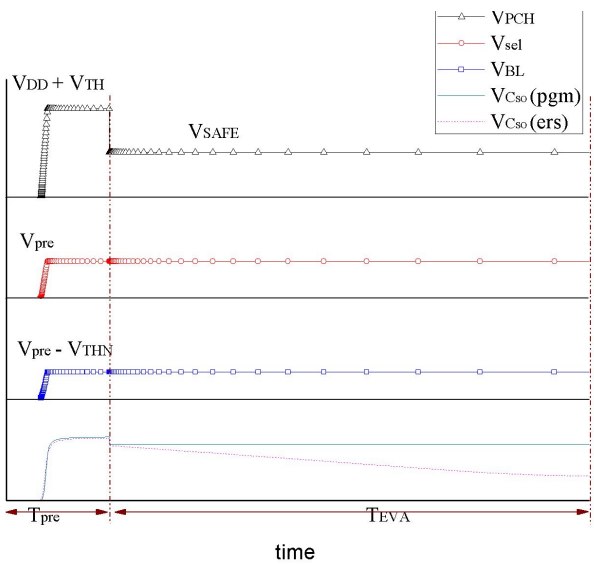
To select the string for read operation, CSL of selected layer is grounded and SSL is turned on for selected strings. BL of selected string is biased at read drain voltage (V_{BL}) and GST tied at BL is turned on at once. The selected WL is biased at V_{read} and other WLs are biased at V_{pass}.

As mentioned in the chapter 1, two sensing schemes for NAND flash memory are widely used. Of the two, the current sensing scheme, all bit-line (ABL) read scheme, is suitable for TiGer structure. In the case of voltage sensing scheme, the BL is left floating to sense the cell state using voltage drop caused by sink current. Thus, if the cell is erased, bit-line capacitor is discharged and the potential of BL decreases during the evaluation phase. That means GST is turned off and discharging of BL capacitor is interupted on the evalution phase.

Meanwhile, in the ABL read scheme, BL potential of read string is constantly maintained during read operation. Fig. 3.14 shows the main elements of the ABL sense amplifier and the timing diagram for TiGer structure. During the evaluation time, M_{PCH} is switches off and the cell current discharges the C_{SO} capacitor, when the selected cell is erased.



(a)



(b)

Fig. 3.14. (a) The main elements of the ABL sense amplifier and (b) the timing diagram of ABL read scheme for TiGer structure

In the conventional ABL read scheme, V_{BL} is decreased into 0.3~0.4 V, to reduce dynamic power consumption. However, bit-line voltage for TiGer structure cannot be scaled since V_{BL} must be higher than threshold voltage of GST. Although dynamic power consumption of TiGer structure is higher than conventional ABL scheme, it is much lower than one of interleaving read operation (Fig. 3.15).

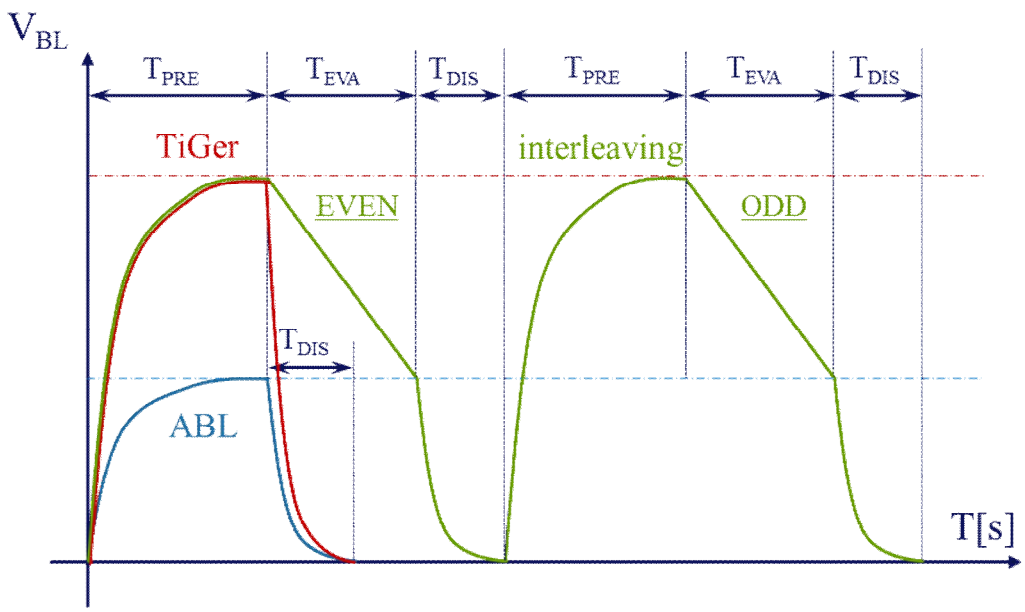


Fig. 3.15. The bit-line voltage versus time graph of interleaving and conventional ABL, and TiGer read scheme

3.2.4 Full Array Design Based on CSTAR having TiGer structure

Conventional NAND flash memory is organized in pages and blocks. A page is the smallest addressable unit for read and program operation, and a block is the smallest erasable unit. The page typically consists of cells sharing the same word line, and the block contains multiple pages. To access storage cells in NAND flash memory, the address of target cells have to be issued. The address is divided in row address and column address. The row address identifies the addressed page, and the column address is used to identify the bytes inside the page.

In the case of CSTAR having TiGer structure, the full array design is not quite different from the conventional planar NAND flash memory, while the addressed word-line influences strings in multiple stacked layer. The addressed word line of CSTAR seems to be made up of multiple pages. Thus, the unit structure as “building” is defined for three-dimensional NAND flash memory [38]. Fig. 16 shows the building for CSTAR having TiGer structure. The building consists of multiple stacked blocks. When the page operation is executed, the page plane is identified by the row address and CSL of addressed page is selected. Thus, for the page operation of CSTAR having TiGer, the CSL decoder and the CSL address are required. Fig. 17 shows the floor plan of the CSTAR having TiGer structure.

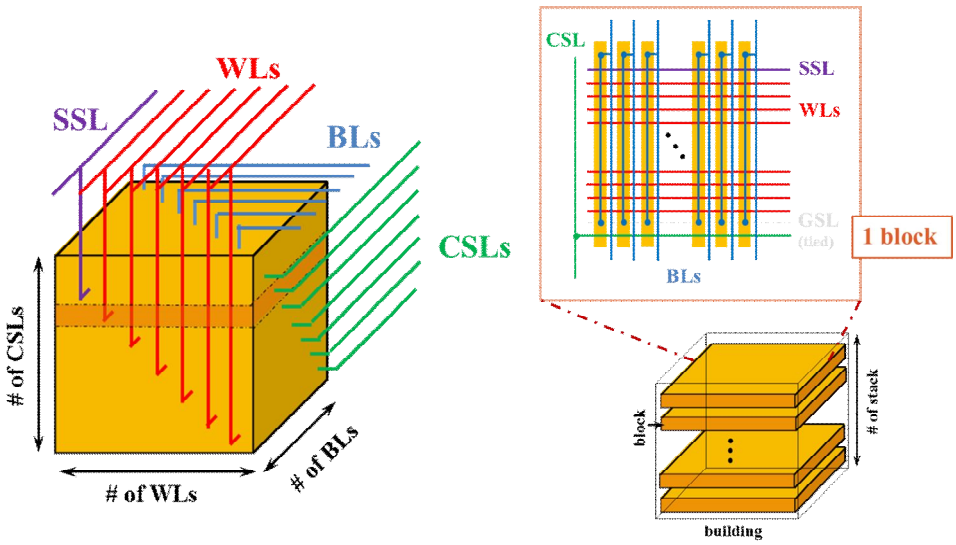


Fig. 3.16. The unit structure as “building” for CSTAR having TiGer structure

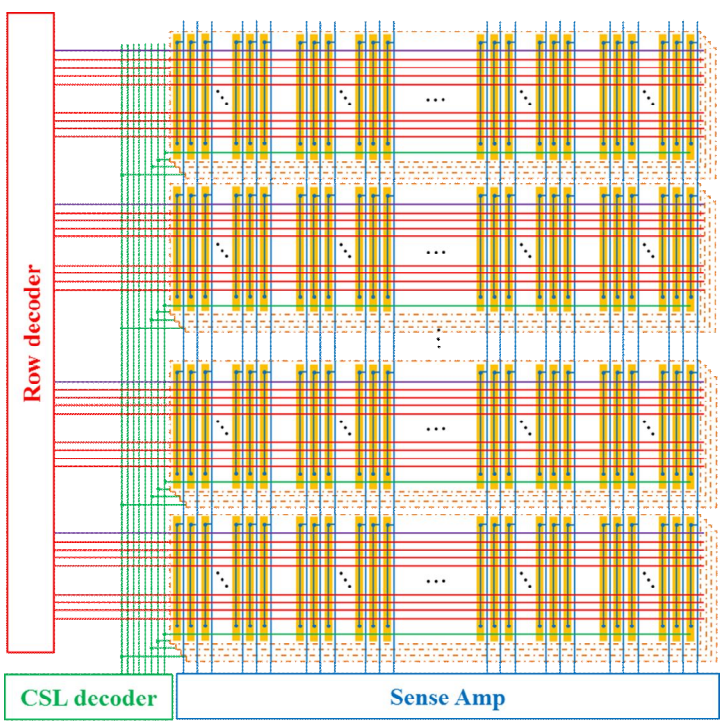


Fig. 3.17. The floor plan of the CSTAR having TiGer structure

Chapter 4

Fabrication and Measurement Results of CSTAR having TiGer structure

In this chapter, fabrication process and measured data of Channel STacked ARray having TiGer structure are presented. Si/SiGe epitaxial growth, SiGe selective etching, and stacked nanowire formation are successfully performed.

4.1 Fabrication Process of CSTAR having TiGer structure

To verify the memory operation, the channel stacked NAND flash memory array having TiGer is fabricated. There are several different methods for fabrication of the CSTAR.

In previous works, CSTAR was fabricated as following process flow.

Fig. 4.1(a)(b): At First, SiGe/Si/SiGe/Si… layers are stacked on the silicon substrate and active fin patterning for channel is performed. Si/SiGe layers are epitaxially grown to make the single crystalline silicon channel. To form the thin active fin and large

contact pad, electron-beam and photo lithography mix-and-match technique is performed. For e-beam lithography, hydrogen silsesquioxane (HSQ), a negative tone electron-beam resist, is used. HSQ is transformed to SiO_2 at elevated temperature, which makes it a hard mask for Si with suitable etching condition. Si etching is performed using inductive coupled plasma (ICP) dry etching in HBr plasma, and SiGe dry etching is also carried out by the same method. Thus, Si/SiGe/… channels are defined at a time.

Fig. 4.1(c): After the active fin patterning, SiGe selective isotropic etching is performed by the Chemical Dry Etching (CDE) and Ammonia-Peroxide Mixture (APM). During SiGe selective etching, silicon nanowire is rounded at once using low Si:SiGe selectivity of APM. 2μm long silicon nanowires can be suspended in midair.

Fig. 4.1(d): Tunneling oxide / charge trapping nitride / blocking oxide layers, and doped poly-Si for gate are deposited. For planarization, chemical mechanical polishing (CMP) is carried out.

Fig. 4.1(e)(f): E-beam lithography for patterning WLs, SSL, and GSL is performed. HSQ is also used as a hard mask for Si. After that, to remove the remained poly-Si between channels, isotropic chemical dry etching in CF_4 plasma is performed. After gate patterning, N+ regions for source/drain are formed by arsenic ion implantation.

The back-end-of-line (BEOL) process follows gate patterning and S/D formation; inter-layer dielectric (ILD) deposition, contact and metal line formation.

The above fabrication process is suitable for string having a couple of word lines because long floating silicon nanowire can be collapsed or adhered.

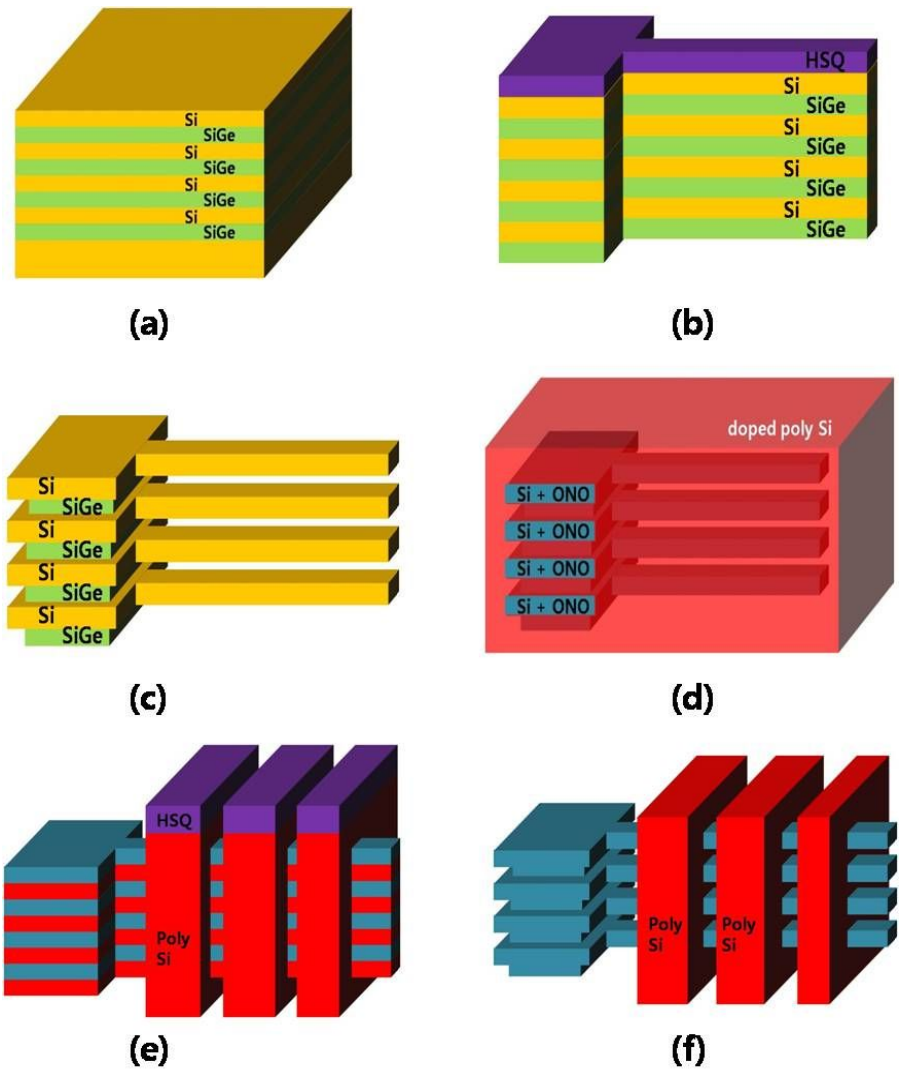


Fig. 4.1. The fabrication process flow of the CSTAR for a couple of word-lines

Fig. 4.2 shows the fabrication process using oxide buttress.

Fig. 4.2(a): SiGe/Si/SiGe/Si… layers are epitaxially grown on the silicon substrate and oxide/poly-Si/oxide layers are deposited as a hard mask to protect etching damage during buttress patterning. Next, using mix-and-match technique, thin channel and large contact pad is formed.

Fig. 4.2(b), (c): To form oxide buttress, oxide layer is deposited by high density plasma chemical vapor deposition (HDPCVD) and e-beam lithography is performed. After making the oxide buttress, isotropic selective etching using CDE and APM is carried out to remove SiGe between channels. Oxide buttress sustains the long channels not to be collapsed or adhered. After the etching using APM, the width of oxide buttress, the space between word lines, decreases.

Fig. 4.2(d): Tunneling oxide / charge trapping nitride / blocking oxide layers, and doped poly-Si for gate deposition and planarization is performed to make WLs, SSL and GSL.

Fig. 4.2(e): To remove conductive SiGe layer between Si wires, oxide buttress is etched. In this process, gates play a role as buttresses for channels. After oxide etch-back porcess is carried out, revealed SiGe layers are removed by CDE and APM etching.

Fig. 4.2(f): After SiGe layer is removed, N+ S/D implantation and the BEOL process is followed.

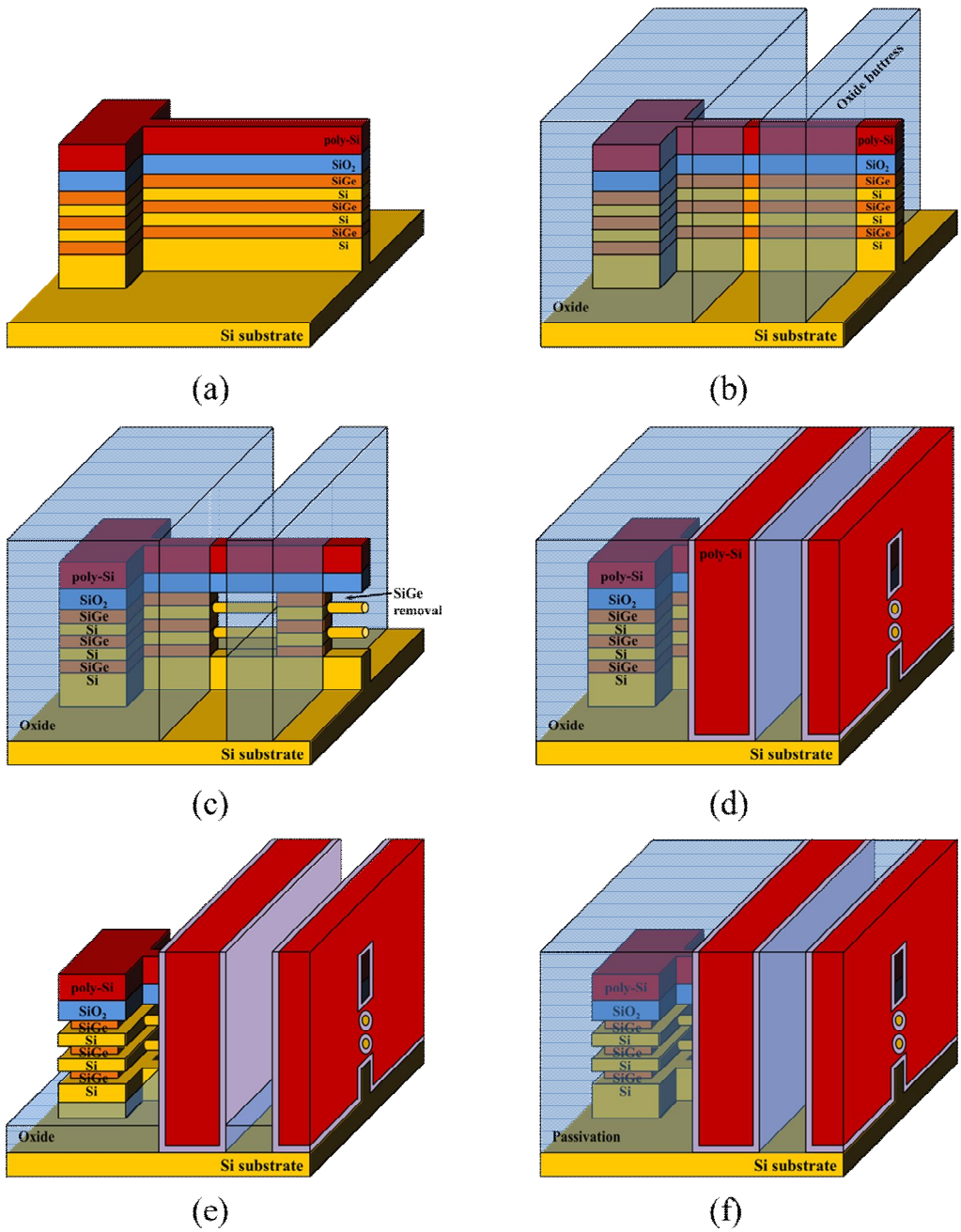


Fig. 4.2. The fabrication process flow of the CSTAR using oxide buttress

Another method to make the CSTAR is fabrication using SiGe buttress. The advantage of using SiGe buttress is that the unnecessary buttresses after gate formation and SiGe layers between Si channels are removed at once. The process flow using SiGe buttress is depicted in Fig. 4.3. The process is similar to Fig. 4.2 except that SiGe is used for buttress

Fig. 4.3(a): SiGe/Si/SiGe/Si… layers are epitaxially grown on the substrate and oxide hard mask is deposited. Using mix-and-match technique, thin channel and large contact pad are defined. Next, SiGe is deposited by ultrahigh vacuum chemical vapor deposition to make the SiGe buttress. CMP is followed for the planarization.

Fig. 4.3(b): After SiGe deposition, e-beam lithography and trench patterning for formation of SiGe buttresses. For e-beam lithography, a positive tone e-beam resist, polymethyl methacrylate (PMMA), is used. SiGe etching is performed using anisotropic dry etching in HBr plasma, and oxide hard mask on channel prevent channel from etching damage during buttress patterning. After patterning the SiGe buttress, SiGe selective isotropic etching is performed using CDE and APM (NH_4OH:H_2O_2 solution). Isotropic SiGe etching removes exposed SiGe between channels, as well as decreases SiGe buttress width. The width of SiGe buttress determines gap between gates, and the narrow WL gap enhances current drivability when the inversion-type source and drain structure by the fringing field is used.

Fig. 4.3(c): Tunneling oxide / charge trapping nitride / blocking oxide layers, and doped poly-Si for gate are deposited. For planarization, CMP is carried out.

Fig. 4.3(d): After the etch-back process to expose the SiGe layer, SiGe is removed by isotropic selective etching using CDE. CDE in fluorocarbon plasma shows very high selectivity and etch-rate. High selectivity of CDE, SiGe buttresses and remained SiGe between Si channels are removed at a time. Next, S/D implantation is performed.

The BEOL process is followed; ILD deposition, contact hole etching and metallization are performed.

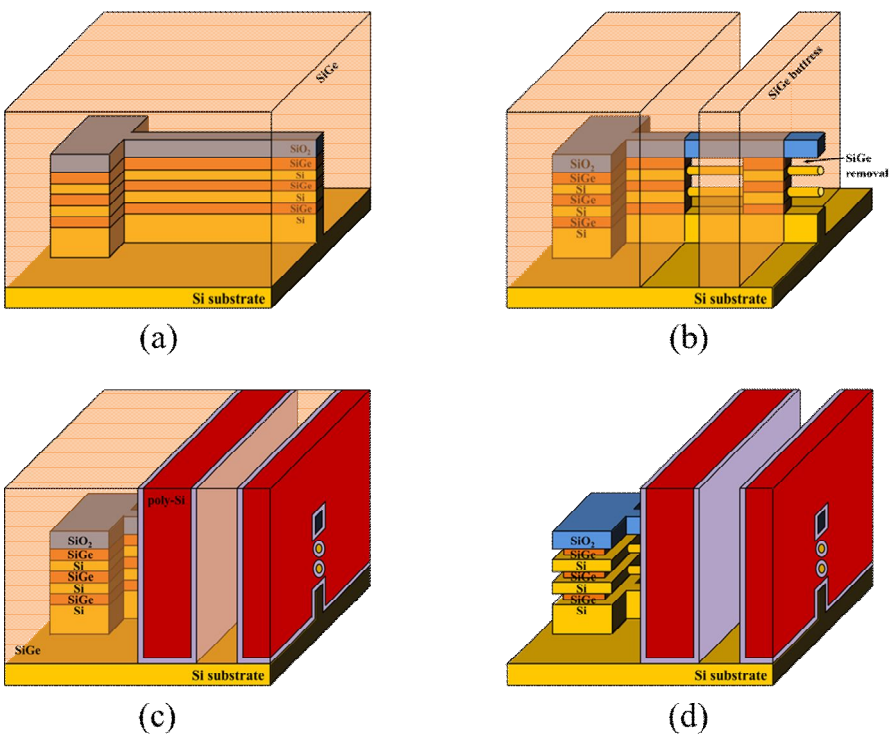


Fig. 4.3. The fabrication process flow of the CSTAR using SiGe buttress

4.2 Key Process Issues of CSTAR fabrication

4.2.1 SiGe/Si Epitaxial Growth for Single-crystalline Si Channel

In order to make single-crystalline silicon channels, Si and SiGe layers are epitaxially grown on Si substrate. Single-crystalline Si layers can be grown on SiGe layers below a critical thickness. Due to the lattice mismatch between Si and Ge of 4.2%, the epitaxial SiGe layer grown on Si is under strain and the critical thickness decreases as the Ge fraction of SiGe increases. If the epilayer thickness exceeds a critical value, the strain is relieved by the generation of misfit dislocations at the Si/SiGe interface. Thus, in the CSTAR fabrication, Ge fraction of SiGe layer is 28% and layer thickness of SiGe layer is 40nm. Si/SiGe layers are in-situ doped with boron to have the doping concentration of $5x10^{17}/cm^3$. Fig 4.4 shows the transmission electron microscope (TEM) image of the Si/SiGe epilayers on Si substrate by UHVCVD.

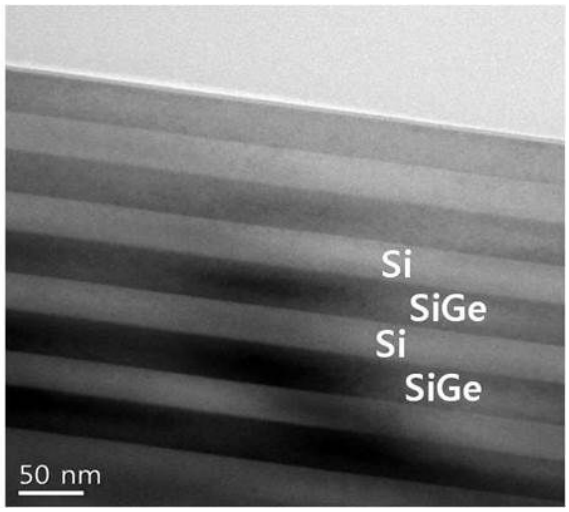


Fig. 4.4. Transmission electron microscope (TEM) image of the Si/SiGe stacked layer

4.2.2 SiGe Selective Isotropic Etching

SiGe is suitable for the sacrificial layer between Si layers. As well as single-crystalline Si layer can be epitaxially grown on SiGe layer, SiGe layers are removed in highly selective isotropic etching process because the presence of germanium offers the possibility to distinguish it from Si. There are several methods to selectively etch the SiGe layer. Among them, in the fabrication of CSTAR, APM wet etching and CDE in fluorocarbon plasma are performed.

4.2.2.1 Chemical Dry Etching in Fluorocarbon Plasma

Chemical Dry Etching (CDE) is performed using remote plasma etching tool. Because generation of reactive radicals is occurred in remote chamber, the etching is only chemical process, just as in the wet etching. CF_4 is used as a main etchant for SiGe selective etching. When the SiGe and Si are exposed simultaneously to fluorine species, SiGe is preferentially etched. Due to the difference between Si-Si and Si-Ge binding energies (2.31eV and 2.12eV, respectively), Si-Ge bond is weaker and more reactive to the F chemical action [39]. Table 4.1 shows the process condition of SiGe selective etching by CDE in fluorocarbon plasma. Adding a small amount of oxygen enhances the selectivity, and N_2 is added to reduce Si and SiGe etch-rates (ERs). Fig. 4.5 show lateral etching of SiGe. SiGe layer is grown on Si substrate, and SiO2/poly-Si hard masks are deposited on it. After CDE, SiGe is etched laterally with a good Si:SiGe, Si:SiO_2 selectivity. ER of SiGe is about 200 Å/s.

Table 4.1. Process condition of SiGe selective etching using CDE

Power (W)	Pressure (mTorr)	CF_4 (sccm)	O_2 (sccm)	N_2 (sccm)
700	350	80	12	12

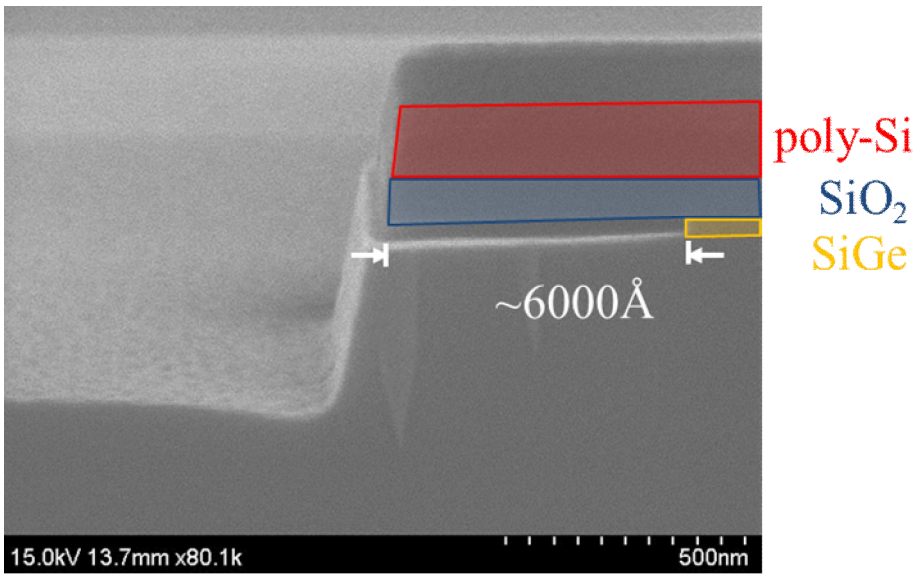


Fig. 4.4. lateral selective etching of SiGe using CDE

remarks – etching condition is described on Table 4.1, ething time is 20s.

4.2.2.2 Ammonia-Peroxide Mixture (APM) wet etching

Many method of SiGe wet etching uses the rapid oxidation of germanium. This characteristic results from the difference between Si-Si and Si-Ge binding energies. Most SiGe etching solutions consist of oxidizer and oxide etchant, and Ammonia-Peroxide Mixture (APM) also does that. APM is NH_4OH:H_2O_2 solution. H_2O_2 oxidizes Si and SiGe, and oxide is etched by NH_4OH. In this fabrication, wet etching is performed in a 1:8:64 ammonium hydroxide, hydrogen peroxide, and deionized water solution at 65℃. Fig. 4.6 shows the etch profile by APM wet etching. Because the Si:SiGe selectivity of APM wet etching is lower than CDE in fluorocarbon plasma, the corner of Si edge is rounded

during wet etching. Moreover, $NH_4OH:H_2O_2$ solution is used in surface cleaning. Thus, APM wet etching is used to remove SiGe layer in conjunction with CDE for nanowire rounding and surface cleaning. Fig. 4.7 depicts the etch rate of APM wet cleaning

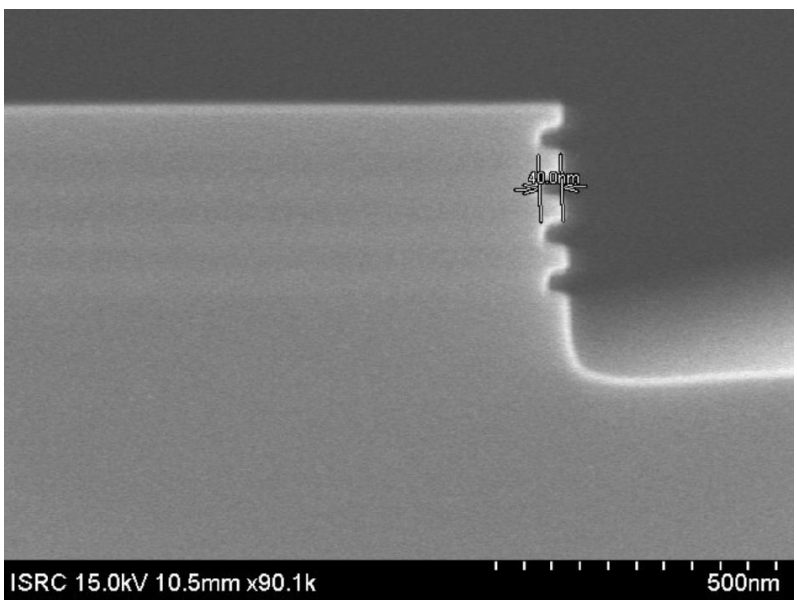


Fig. 4.6 The etch profile by APM wet etching

remarks – etching condition is 1:8:64 ammonium hydroxide, hydrogen peroxide, and deionized water solution at 65 ℃, and ething time is 30s.

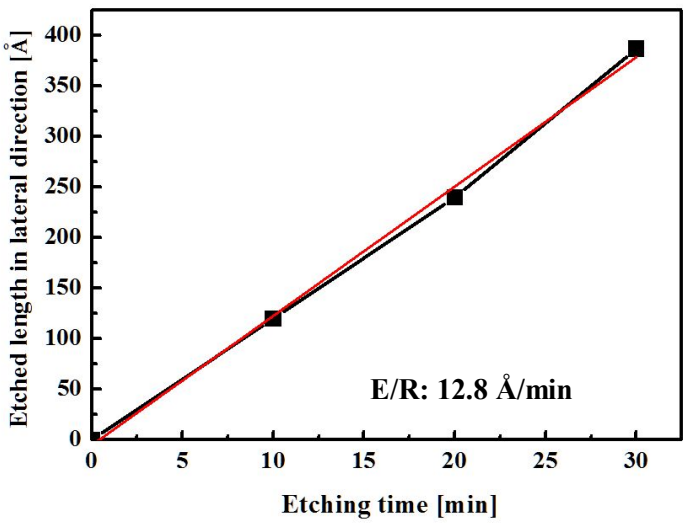


Fig. 4.7. The etch rate of APM wet etching

4.2.3 Stacked Nanowire Formation

To form the stacked nanowire, first, the active fin patterning on Si/SiGe/Si/SiGe… layers is carried out by e-beam lithography. Next, SiGe layers are removed by CDE and rounded by wet etching. After the formation of Si nanowire channels, tunneling oxide/trapping nitride/blocking oxide and doped poly-Si for gate are deposited. To form the stacked gate-all-around cell successfully, the thickness of SiGe layer must be larger than the twice of ONO thickness, and it cannot exceed critical thickness to avoid the generation of misfit dislocations. Thus, accurate thickness control of epilayer and ONO layers are required. Another key process to patterning the stacked channel is control of fin etch slope. The active fin etch slope affects nanowire diameters

of stacked layer and electrical characteristics of stacked string of NAND flash memory. When the patterning is performed using HSQ mask only, active fin is not steep. Both ends of HSQ mask by e-beam lithography are thin, and the slope of HSQ is transferred onto the fin profile. Fig. 4.8 shows the four stacked gate-all-around nanowires using HSQ mask only. The diameter of 1st layer cell is smaller than 2nd, 3rd, 4th layer cells. The difference of diameter of GAA structure affects variation of program/erase speed as well as distribution of initial threshold voltage, subthreshold swing, and on-cell current. To avoid this problem, additional hard mask is introduced during active fin patterning. This hard mask also acts as preventing the etch damage from the buttress patterning. As mentioned in the Fig. 4.3, before the patterning of active fin, oxide/poly-Si hard mask layers are deposited. When the patterning is performed on thin poly-Si layer, the HSQ mask slope is not transferred excessively. And oxide layer below poly-Si layer endures well during active fin etching. After the active fin patterning, hard mask layers protect the channel from etch damage of buttress patterning. When the buttress patterning is carried out, if the hard mask doesn't exist, 1st channel layer is exposed directly to etchant for buttress material etching. In the case of using SiGe buttress, oxide layer protects the 1st channel from the etch damage during patterning the SiGe. Thus, four stacked GAA nanowires are fabricated successfully and the diameter of them are almost the same, as depicted in Fig. 4.9.

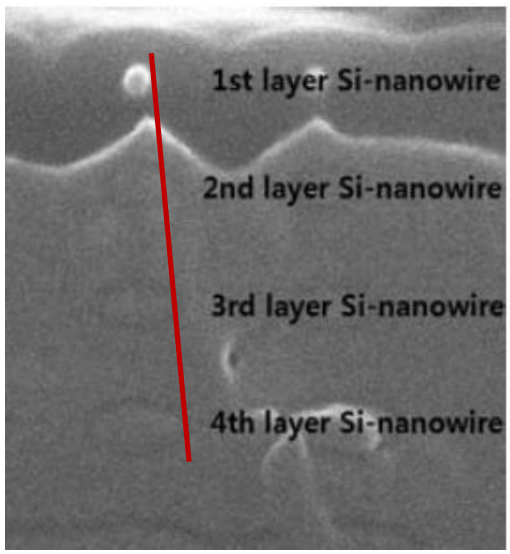


Fig. 4.8. The four stacked gate-all-around nanowires using HSQ mask only

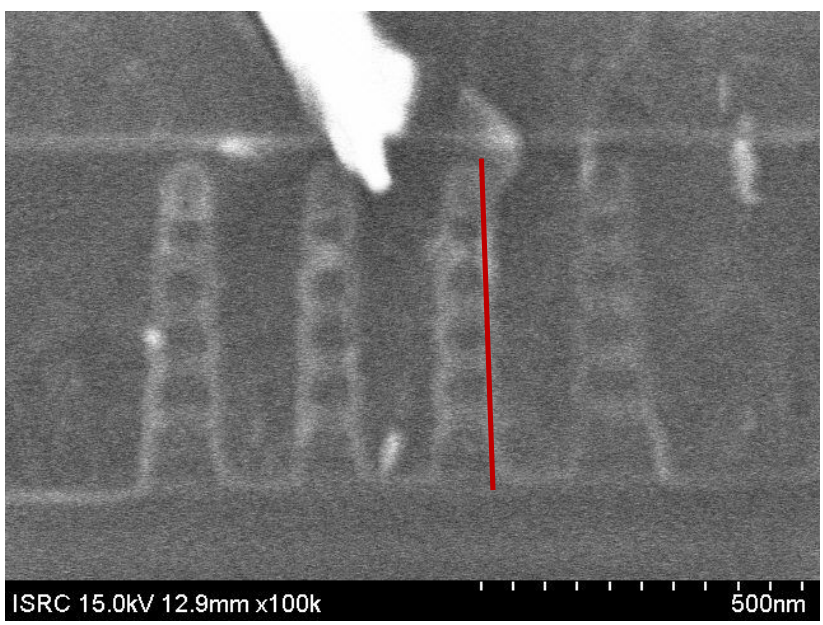


Fig. 4.9. The four stacked gate-all-around nanowires using oxide hard mask

4.2.4 Gate Patterning – Conventional Patterning and Damascene Gate Formation

In the CSTAR fabrication, two methods for forming the GAA cell are introduced. One is depicted in Fig. 4.1. The doped poly-Si is deposited on silicon nanowires and gate patterning using E-beam lithography by HSQ is performed. In this process, to make gates successfully, conductive gate material remaining between nanowires must be eliminated. Thus, after gate patterning, CDE is carried out for cutting the connection of gates. To apply inversion-type source/drain structure, the narrow WL gap is required to enhance the current drivability, but CDE process extents the WL gap. And the reducing the gap by patterning is limited due to proximity effect in E-beam lithography. Fig. 4.10 shows the linked pattern by proximity effect. To overcome this problem, extended WL patterning is introduced [40]. Fig. 4.11 shows the process flow of extended WL patterning.

Another method is damascene gate patterning (Fig. 4.2, 4.3). First, the buttress material is deposited and patterning is performed. After the removing exposed SiGe between Si layers, ONO layers and gate material are deposited. Using the damascene process, gate is separated automatically and WL gap can be narrower during the removing the exposed SiGe layer than conventional one. Moreover, in this process, High-K dielectric and Metal gate can be applied. Fig. 4.12 shows the gate trench having narrow pitch using damascene patterning and APM wet etching. the word-line gap is narrow

enough to induce the inversion source/drain by fringing field.

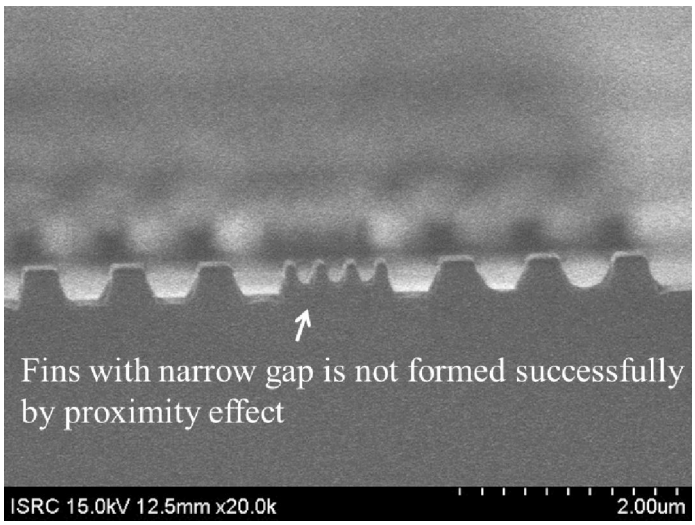


Fig. 4.10. The linked narrow pattern by proximity effect

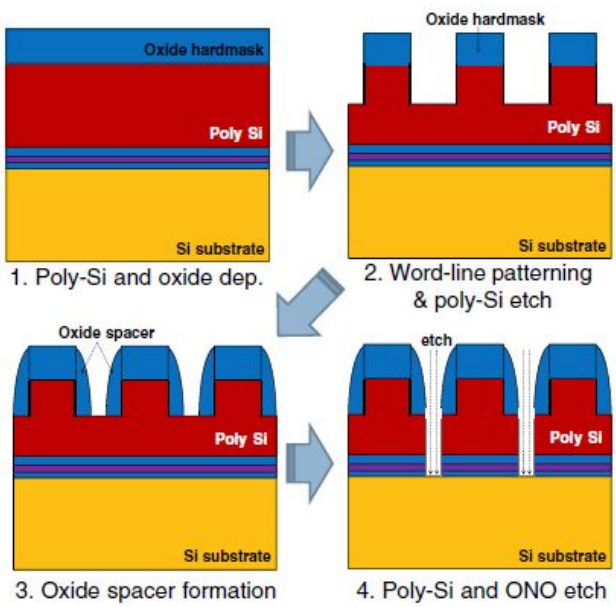
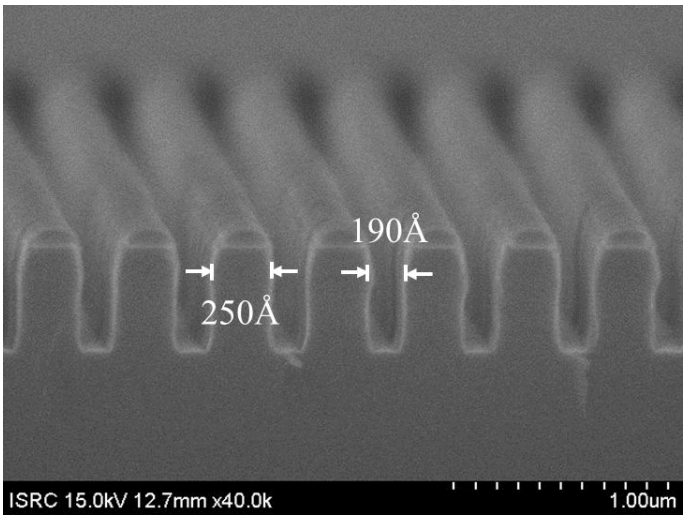
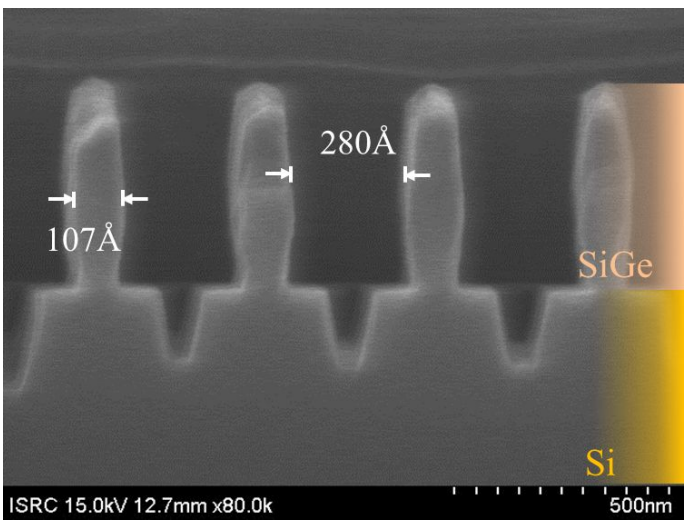


Fig. 4.11. The process flow for the extended word-line patterning



(a)



(b)

Fig. 4.12. The trench gate profile (a) after the patterning using E-beam litho. by PMMA (b) after the APM wet etching

remarks – the patterning is performed deposited poly-SiGe on Si substrate to examine the ER of SiGe and SiGe:Si selectivity.

4.3 Electrical Characteristics of CSTAR having TiGer Structure

Electrical characteritics of fabricated CSTAR having TiGer structure are discussed in this chapter. To secure a high yield of fabricated devices, a one-story 2x3 arrays are fabricated. And to prevent the back tunneling issue of blocking oxide, Al_2O_3 layer as blocking dielectric is used. As mentioned in the previous chapter, in the CSTAR structure, thickness of dielectric material is important factor for succesful device fabrication. Thus, HfO_2 is used to reduce the thickness of charge trapping layer. Fig.4.13 shows TEM image of cross section of the fabricated device. HfO_2 and Al_2O_3 is deposited using atomic layer deposition (ALD), and the thickness of SiO_2/ HfO_2/ Al_2O_3 layers are 35/45/120 Å. The diameter of Si nanowire is about 20 nm.

The fabricated array has 2 strings having 3 word-lines and SSL, and GST tied with BL. To verify the all operation of TiGer structure, array has two-types. The one is that 2 sources are connected to same CSL, and the other is that 2 drains are connected to same BL. Fig. 4.14 shows the equivalent circuits of fabricated devices.

Fig. 4.15 shows the basic characteristics of fabricated device. The initial threshold voltage of fabricated cells is about 1.1 V. Because BL is tied with GST, V_{BL} must be higer than V_{th} of GST to obtain sufficiently high on-cell current. Thus, V_{BL} of fabricated device is set at 2.5 V, which is higher than conventional bit-line bias. Due to high V_{BL}, GIDL current is observed. And, because single crystalline Si channel is used,

fabricated devides have low Subthreshold Swing (about 180 mV/dec) and high on-cell current (about 10μA).

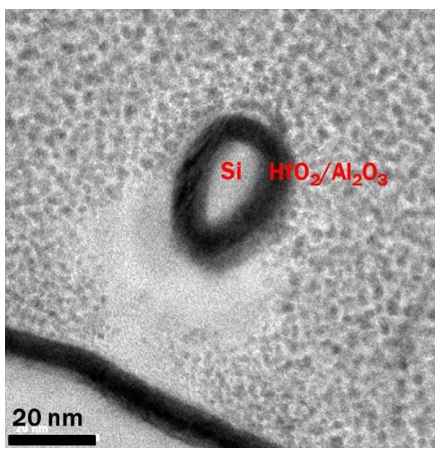


Fig. 4.13. TEM image of cross section of the fabricated device

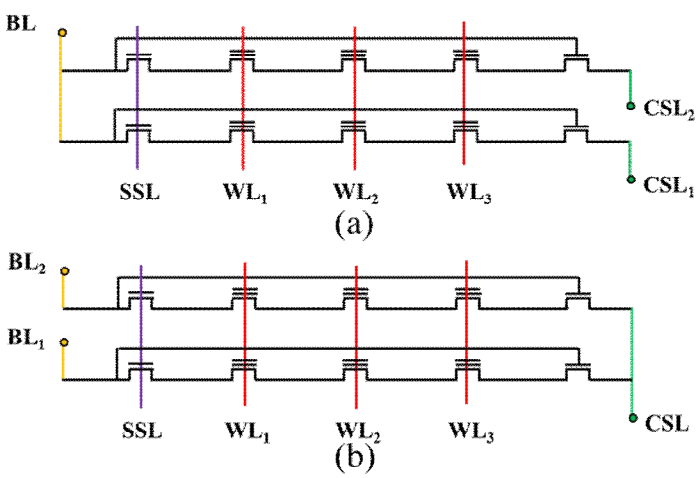


Fig. 4.14. The two-types of fabricated array; (a) 2 strings with common CSL, and (b) 2 strings with common BL

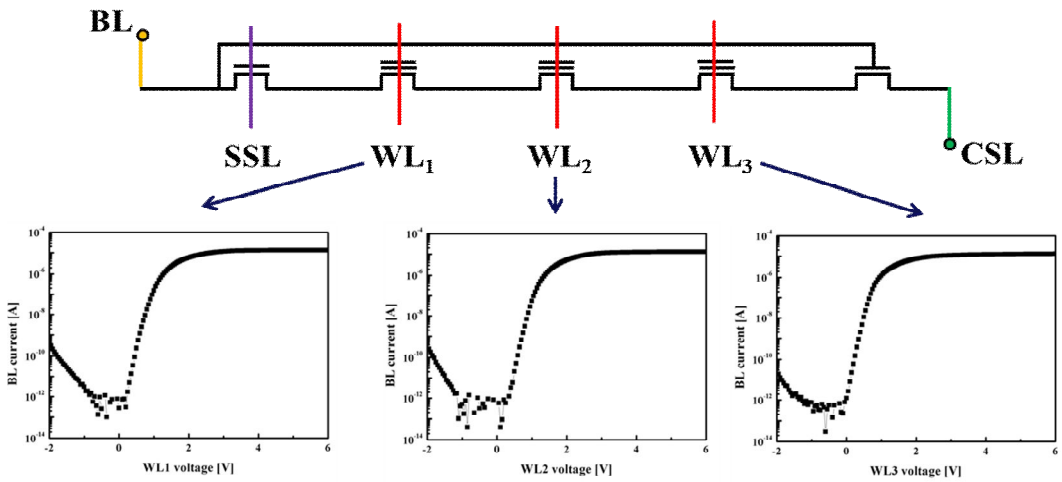


Fig. 4.15. The transfer curves of fabricated device

remarks – V_{BL} = 2.5 V, V_{SSL} = 6 V, V_{pass} = 6 V.

Fig. 4.16 shows the program/inhibition scheme and timing diagram for fabricated TiGer structure. Due to the limitation of measuring equipment, pre-charging stage is not performed sufficiently. The inhibit cell sharing the bit-line with selected cell is self-boosted since the CSL and GST tied with BL are biased at V_{cc} and SST is turned-off. In the case of inhibit cell sharing the CSL with selected one, channel is boosted because SST and GST are all turned-off, but self-boosting efficiency is lower than other inhibition cases due to the leakage through the SST and GST. Fig. 4.17 shows the transfer curve of selected and inhibited cells of TiGer structure. The program disturbance of inhibited cells is observed. To perform bulk erase operation, GST of fabricated device is designed for the body contact as shown in Fig. 3.5. Thus, bulk erase scheme is available.

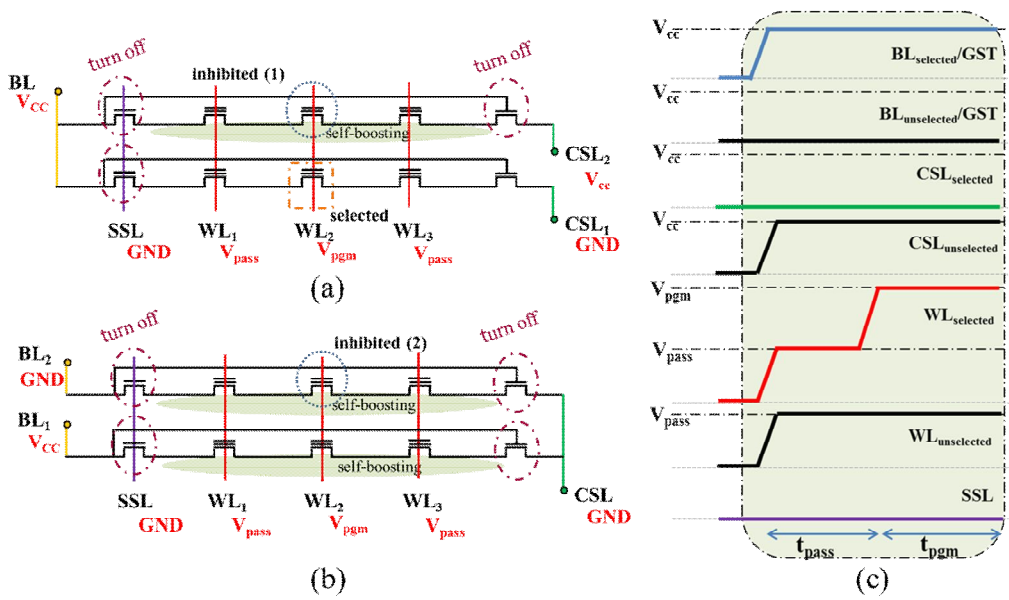
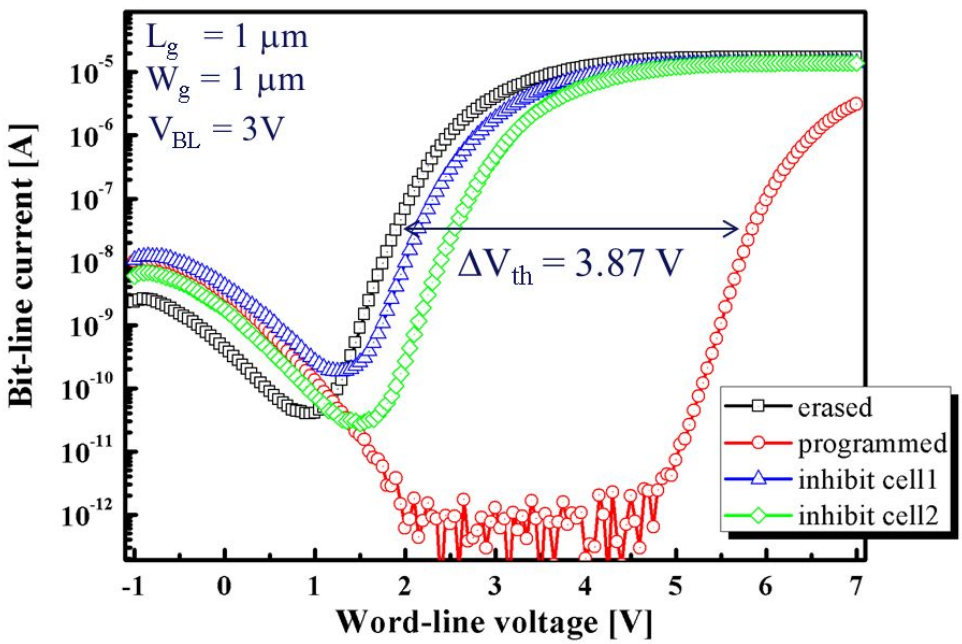


Fig. 4.16. (a), (b) The program and inhibition scheme and (c) timing diagram for fabricated TiGer structure

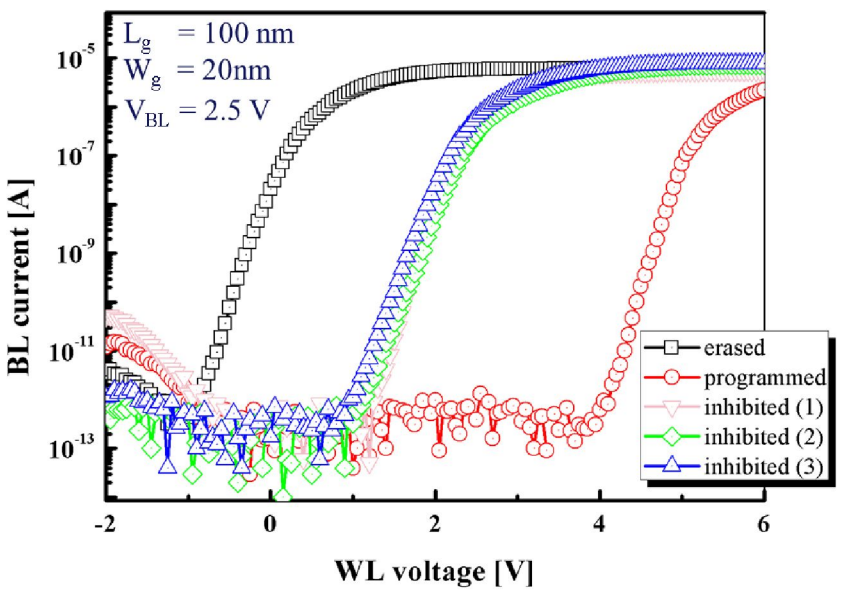


(a)

remarks – L_g, W_g = 1 μm(defined by photo lithogrpahy)

program conditions: V_{pass} = 6 V, V_{pgm} = 17 V, t_{pass} = 100μs, t_{pgm} =1ms,

erase condition: V_{ers} = 16 V, t_{ers} = 1s



(b)

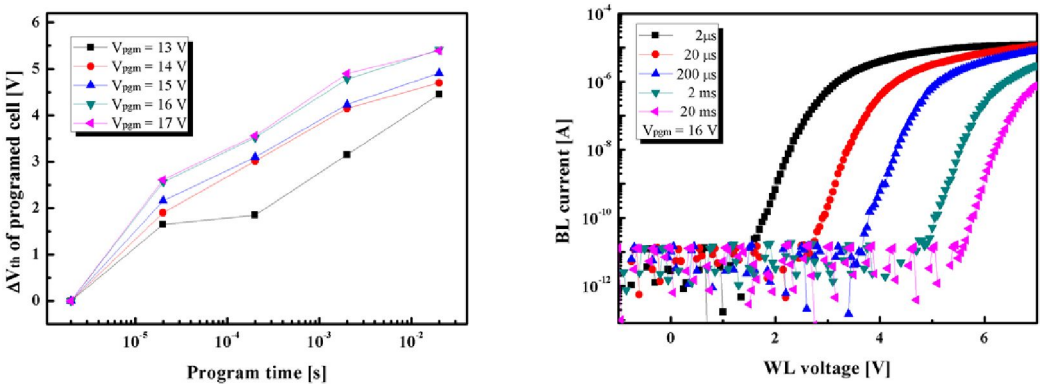
remarks – L,= 100 nm, W_g =20 nm(defined by E-beam lithogrpahy)

program conditions: V_{pass} = 7 V, V_{pgm} = 15 V, t_{pass} = 10μs, t_{pgm} =1ms,

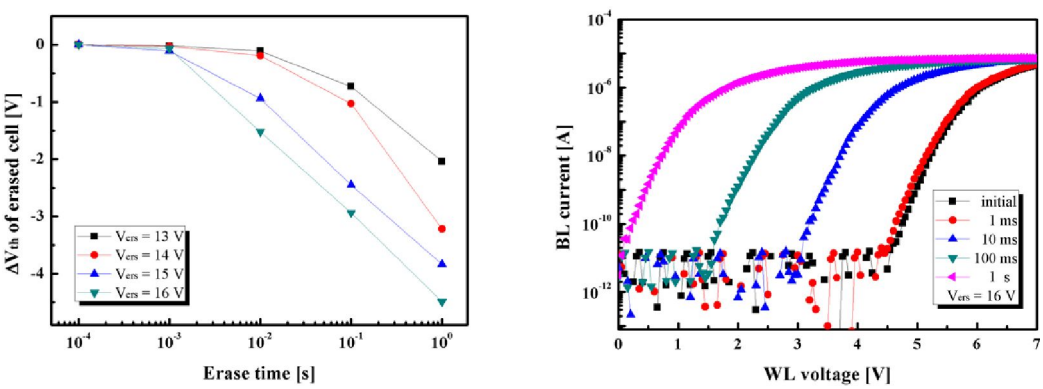
erase condition: V_{ers} = 16 V, t_{ers} = 2s

Fig. 4.17. The transfer curve of selected and inhibited cells of TiGer structure

The program and erase characteristics as a function of time and voltage are depicted in Fig. 4.18. The field concentration effect of GAA structure enables lower program and erase voltage thant the conventional planar devices.



(a)



(b)

Fig. 4.18. (a) The program and (b) erase characteristics of fabricated device

Fig. 4.19 demonstrates the retention characteristics of the fabricated device. The tested cell for retention is programed at V_{pgm} = 14 V and V_{th} window is 3.2 V. The cell is baked at 85 °C. The extracted memory winodow by extrapolating threshold voltage up to 10 years is about 1.5V.

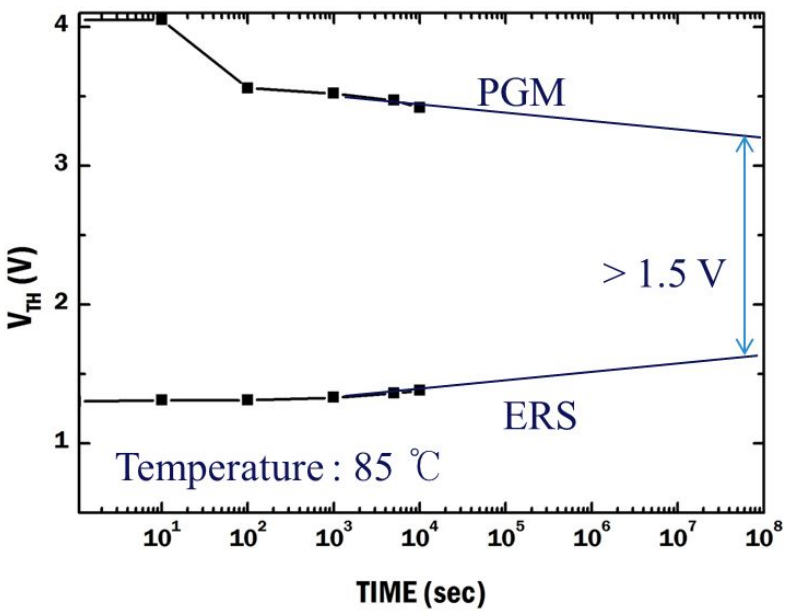


Fig. 4.19. The retention characteristics of fabricated device.

Chapter 5

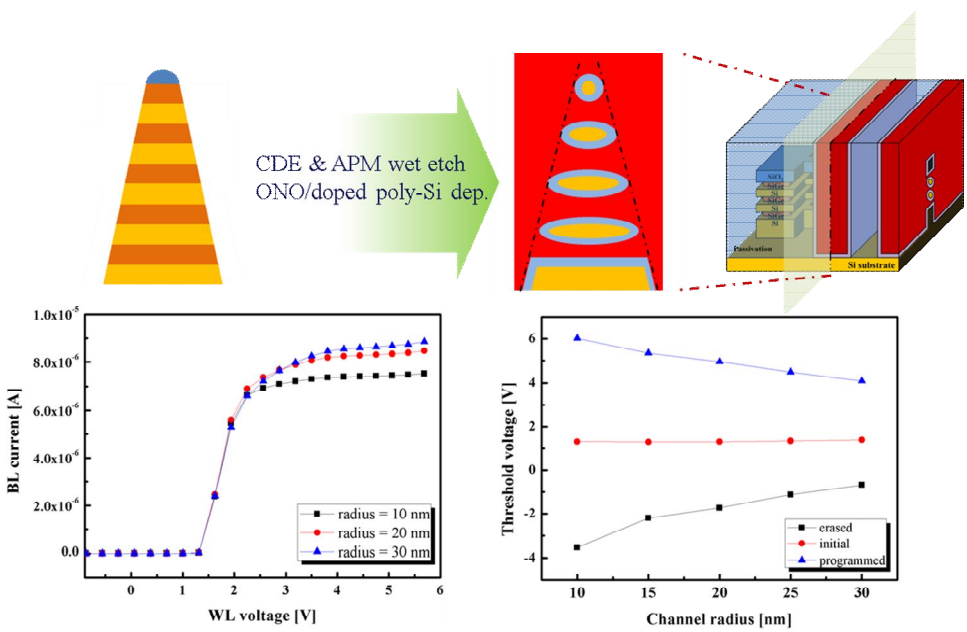
Optimization for TiGer Structure

In this chapter, significant conditions for reliable TiGer structure operation are demonstrated. First, effects of etching slope are demonstrated using simulation, and improvement method is introduced. Next, influence of threshold voltage of GST variation is studied. Finally, insertion of dummy WL is introduced.

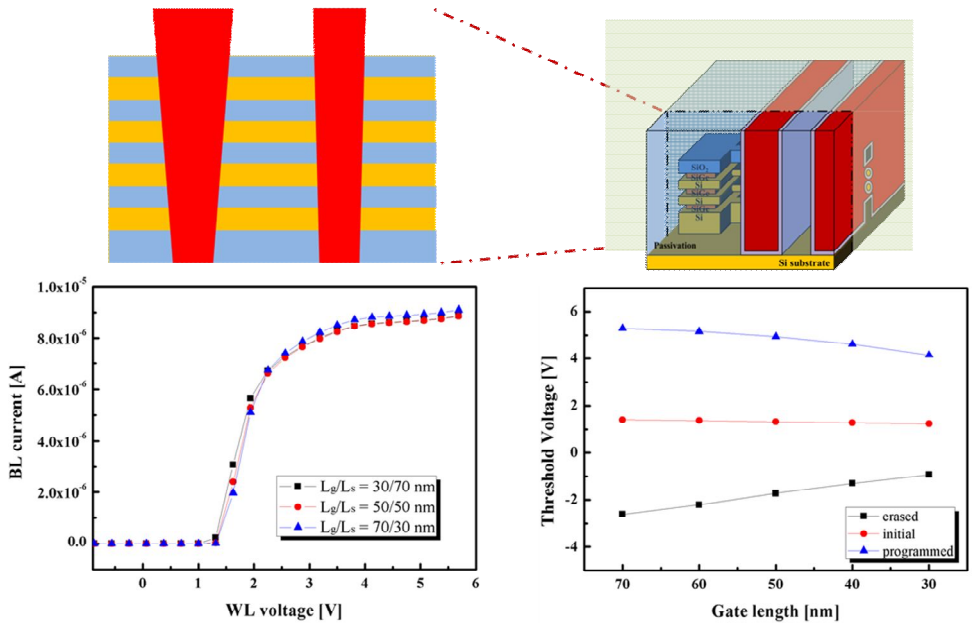
5.1 The Effects of Etching Slope

To make Si nanowires, the fin which consists of Si/SiGe/… layers are formed using dry etching, and SiGe layers are selectively removed. If the fin has a gentle slope, the channels on ther lower floor have the larger diameter than channels on the top floor. When the gate pattering is performed, etching slope causes variation on channel length and gate space. The variation on channel radius, channel length, and gate space has a bad effect on the distribution of on-cell current and program/erase speed, as shown in Fig. 5.1. To improve the etching slope of Si fin formed using E-beam lithography, oxide and thin α-Si hard mask are deposited on Si/SiGe/… substrate. Patterning on thin α-Si hard mask

prevents that HSQ e-beam resist slope is transffered excessively. Next, oxide hard mask pattering is performed using α-Si hard mask which having more steep etching slope than HSQ. Because oxide layer is outstanding hard mask for etching in HBr plasma, Si/SiGe/… active fin has a very steep slope (Fig. 5.2). Gate patterning is carried out using same method.



(a)



(b)

Fig. 5.1. The effects of etching slope of (a) active fin and (b) gate patterning on on-cell current and program/erase speed

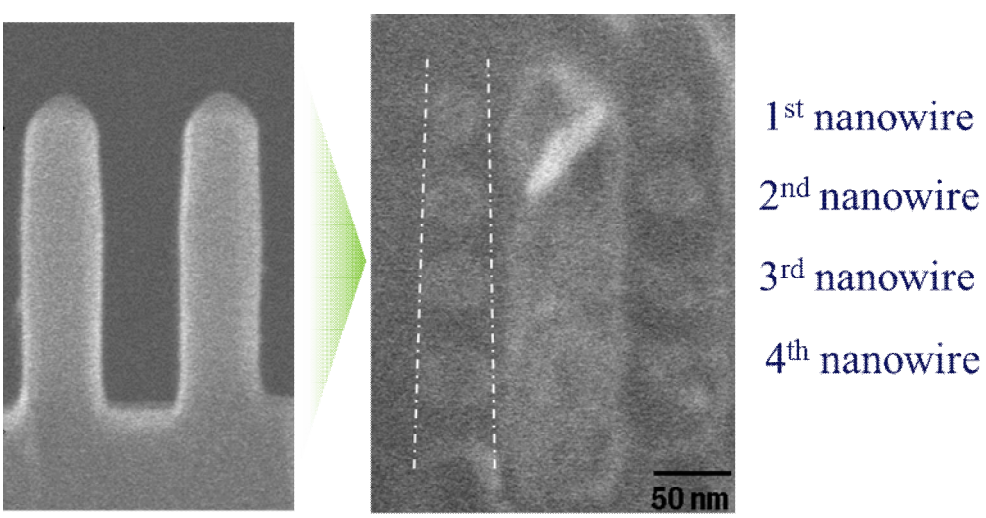
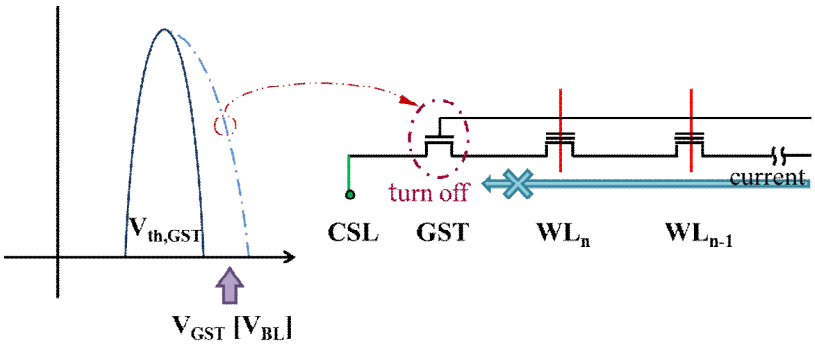


Fig. 5.2. quite uniform diameter of stacked nanowires due to a very steep etching slope

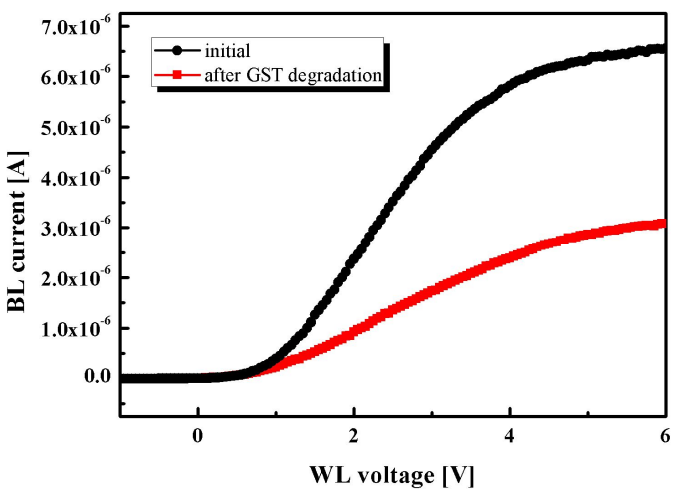
5.2 The Effects of Threshold Voltage of GST Variation

In the case of TiGer structure, the threshold voltage of GST is important to perform reliable memory operation. Fig. 5.3(a) shows the read operation failure by increased threshold voltage of GST. Because the gate voltage of GST is same as bit-line voltage, increased V_{th} of GST brings about on-cell current decreasing. In the case of fabricated device, because the charge trap lyaer, HfO_2, exists in the select transistor, threshold voltage of GST can increase after program/erase cycling. Fig. 5.3(b) shows the decrease of BL current due to increased V_{th} of GST. On the other hand, decreased V_{th} of GST causes program operation failure. During the program operation, GST must be turned-off. If the leakage current through GST increases, the self-boosted potential decreaes. Fig. 5.4(b) shows the self-boosting failure due to V_{th} decrease of GST.

As mentioned above, because threshold voltage distribution of GST must be controlled well, the separate gate patterning for select transistors is required. Separate select transistors patterning can be applied to CSTAR architecture, in contrast with GSTAR. First, only WL gate patterning is carried out and WL gates are deposited on ONO layers. Next, select transistors patterning is performed separately after forming WL gates and select gates are deposited on only gate oxide layer.

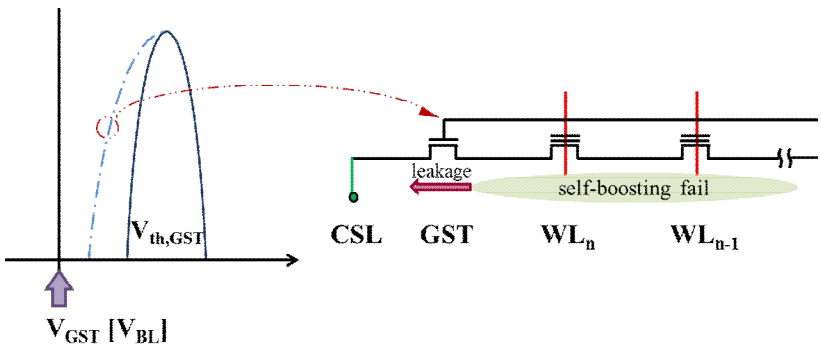


(a)

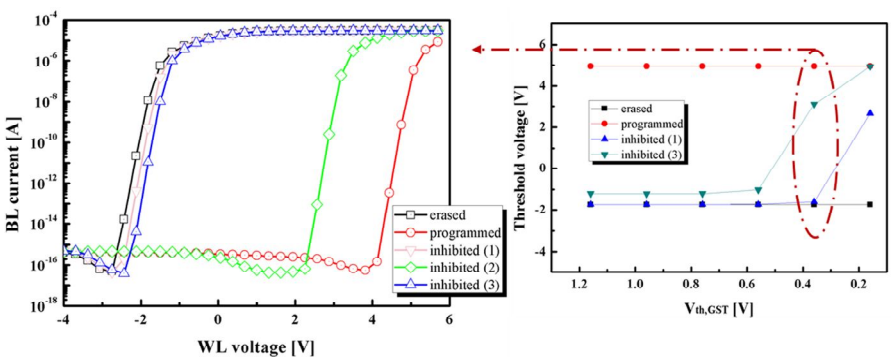


(b)

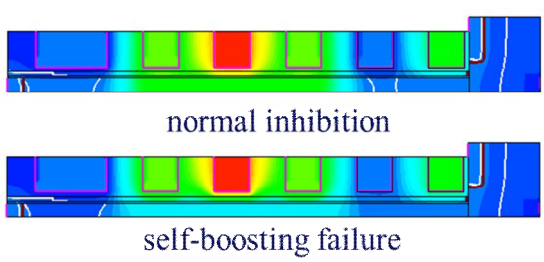
Fig. 5.3. (a) The read failure due to increased V_{th} of GST and (b) transfer curves of initial and after program/erase cycling. The BL current decreases since V_{th} of GST increases due to degradation.



(a)



(b)



(d)

Fig. 5.4. (a) The progrma failure due to decreased V_{th} of GST, (b) the program windows as a function of $V_{th,GST}$, and (c) the channel potential of normal inhibition case and self-boosting failure case.

5.3 Dummy Word-Lines and BIL

During the inhibition operation, hot-carriers induced by GIDL increase the program disturbance of edge memory cell. To prevent undesired program disturbance on edge cell, dummy WL is additionally placed between select transistor and edge cell (Fig. 5.5) [41]. In the case of TiGer structure, one of dummy WL can be utilized as BIL. Fig. 5.6 shows the bird's eye view and equivalent circuit of proposed structure.

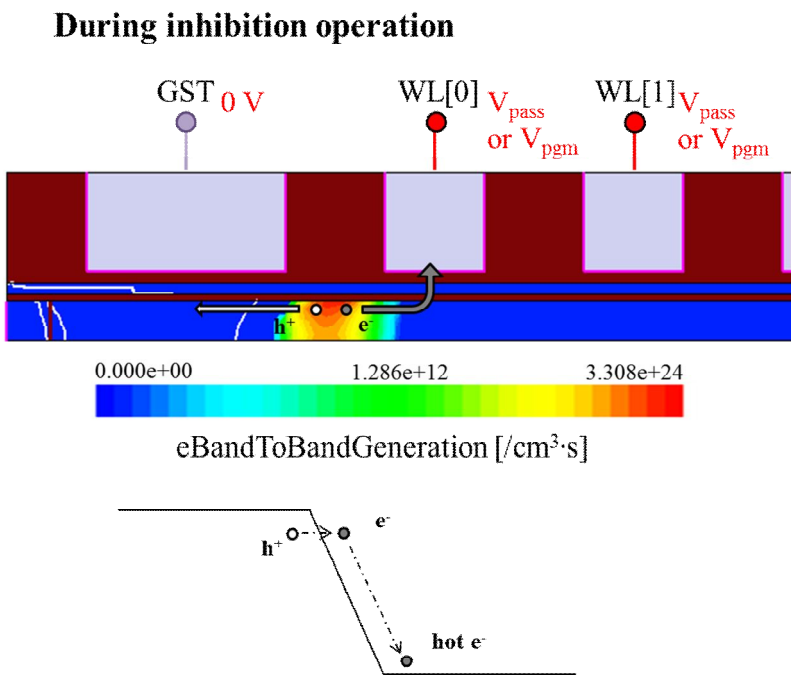
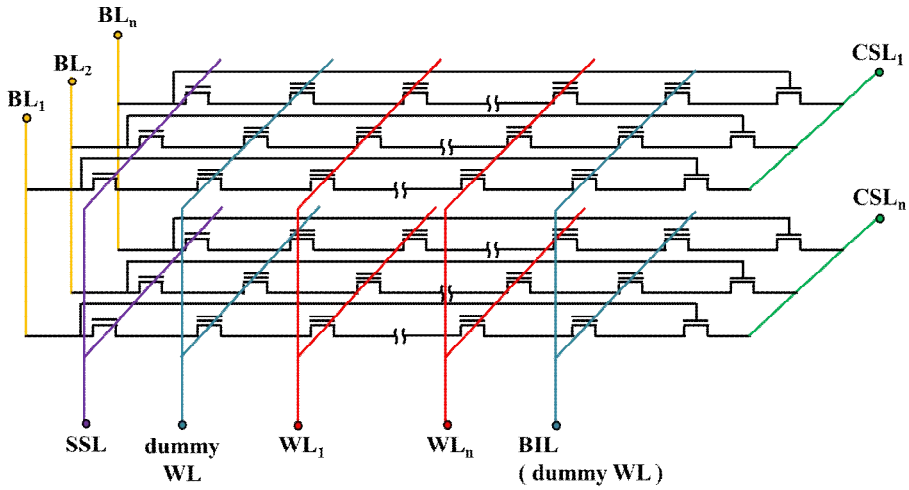
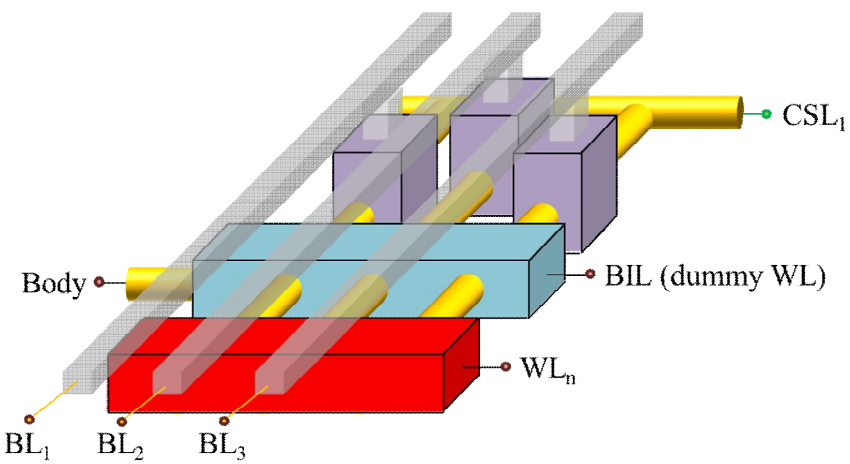


Fig. 5.5. The undesired program disturbance on edge cell



(a)



(b)

Fig. 5.6. (a) The bird's eye view and (b) equivalent circuit of proposed structure

Chapter 6

Conclusions

In this dissertation, Channel Stacked ARray (CSTAR) having Tied Bit-line and Ground Select Transistor (TiGer) is proposed, fabricated, and characterized. In spite of the introduction of three-dimensional NAND flash structure, it is difficult to make 3-D metal line because peripheral circuits are still two-dimensional. Therefore, a novel architecture of 3-D flash memory for compact metal-line is proposed. In comparison with gate stacked type 3-D flash memories, the optimized architecture of CSTAR is not established. The proposed CSTAR having TiGer structure can be the promising solution for compact design and exact memory operation.

In the CSTAR having TiGer structure, channels are stacked in a horizontal direction and each BL is connected to stacked channels in a line. Each CSL is connected to channels on the same floor, and the block is defined by CSL decoder. The main feature of TiGer structure is that each island-type GST separated in the lateral direction is connected to BL. In this way, the CSTAR having TiGer structure accomplishes the memory operation successfully without complicating SSL decoding.

The CSTAR having TiGer requires a novel programing scheme since the

architecture of that is quite different from conventional NAND flash memory. To verify the programing scheme for TiGer structure, TCAD simulation is carried out. By applying proposed programing scheme, the selected cell is only programed and inhibited cells are not programmed by self-boosting. The read and erase scheme of conventional NAND flash memory is applied for TiGer structure, and these are also verified using TCAD simulation.

Finally, 4-layer stacked single-crystalline silicon nanowire channel with GAA structure for CSTAR having TiGer structure is fabricated. Stacked nanowire formation, isotropic SiGe selective etching, damascene gate formation using SiGe buttress, and device fabrication process flow are demonstrated. Electrical characteristics of fabricated array show that proposed program and read, and erase scheme for CSTAR having TiGer structure are accomplished. Performing various measurements, the feasibility of designed CSTAR having TiGer is confirmed successfully.

Bibliography

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

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

[3] Michael Yang, “Aided By Ultrabook Sales, Global NAND Flash Revenue Expected to Grow 8 Percent This Year,” isuppli.com, April 19, 2012.

[4] Lyths, "SSD vs. HDD". elitepcbuilding.com, July 11, 2011.

[5] Gartner Inc., “NAND Flash supply and demand,” gartner.com, Oct. 10, 2012.

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

[7] M. Lenzlinger and E. H. Snow, “Fowler-Nordheim tunneling into thermally grown SiO,” *Journal of Applied Physics*, vol. 40, no. 1, pp. 278-283, 1969.

[8] J. Sune, P. Olivo, and B. Ricco, “Quantum-mechanical modeling of accumulation layers in MOS structure,” *IEEE Transactions on Electron Devices*, vol. 39, no. 7, pp.1732-1739, July 1992.

[9] S. Tam, P. K. Ko, and C. Hu, “Lucky-Electron Model of Channel Hot-Electron Injection in MOSFET’s,” *IEEE Transactions on Electron Devices*, vol. 31, no. 9, pp.1116-1125, Sept. 1984.

[10] Sumio Tanaka, and Mitsuaki Ishikawa, “One-Dimensional Writing Model of n-Channel Floating Gate Ionization-Injection MOS (FIMOS),” *IEEE Transactions on Electron Devices*, vol. 28, no. 10, pp.1190-1197, Oct. 1981.

[11] Hang-Ting Lue, Sheng-Chih Lai, Tzu-Hsuan Hsu, Yi-Hsuan Hsiao, Pei-Ying Du, Szu-Yu Wang, Kuang-Yeu Hsieh, Rich Liu, and Chih-Yuan Lu, “A critical review of charge-trapping NAND Flash devices,” in *IEEE ICSICT Tech. Dig.*, pp. 807-810, Oct. 2008.

[12] Y. Shin, J. Choi, C. Kang, C. Lee, K.T. Park, J.S. Lee, J. Sel, V. Kim, B. Choi, J. Sim, D. Kim, H.J. Cho and K. Kim, “A Novel NAND-type MONOS Memory using 63nm Process Technology for Multi-Gigabit Flash EEPROMs,” in *IEDM Tech. Dig.*, pp. 337-340, 2005.

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

[14] Jae-Duk Lee, Sung-Hoi Hur, and Jung-Dal Choi, “Effects of Floating-Gate Interference on NAND Flash Memory Cell Operation,” *IEEE Electron Device Letters*, Vol. 23, No. 5, pp. 264-266, May. 2002.

[15] E. Vianello, F. Driussi, A. Arreghini, P. Palestri, D. Esseni, L. Selmi, N. Akil, M. J. van Duuren, and D. S. Golubovi´c, “Experimental and Simulation Analysis of Program/Retention Transients in Silicon Nitride-Based NVM Cells,” *IEEE Transactions on Electron Devices*, vol. 56, no. 9, pp.1980-1990, Sep. 2009.

[16] J. E. Brewer, M. Gill, “Nonvolatile Memory Technologies with Emphasis on Flash”,

IEEE Press, pp. 2, 2008.

[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] Eun Suk Cho, Hyun Jung Kim, Byoung Taek Kim, Jai Hyuk Song, Du Heon Song, Jeong-Hyuk Choi, Kang-Deog Suh, and Chilhee Chung, "Optimal Cell Design for Enhancing Reliability Characteristics for sub 30 nm NAND Flash Memory," in *IEEE IRPS Tech. Dig.*, pp.611-614, May 2010.

[22] 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.

[23] 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.

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

[25] 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.

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

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

[28] K. Prall, “Scaling Non-Volatile Memory Below 30 nm,” *NVSMW Tech Dig*., pp. 5-10, Aug. 2007.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] Masahide Kimura, "3D cells make terabit NAND flash possible," techon.nikkeibp.co.jp, Sept 17, 2009.

[35] 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.

[36] Sungjin Whang, et al., "Novel 3-Dimensional Dual Control-Gate with Surrounding Floating-Gate (DC-SF) NAND Flash Cell for 1Tb File Storage Application," in *IEDM Tech. Dig.*, pp. 29.7.1~4, Dec. 6~8, 2010.

[37] W. Kim, S. Choi, J. Sung, T. Lee, C. Park, H. Ko, J. Jung, I. Yoo, and Y. Park, "Multi-layered vertical gate NAND flash overcoming stacking limit for terabit density storage," in *VLSI Symp. Tech. Dig.,* pp. 188–189, 2009.

[38] Y. Kim, Jang-Gn Yun, Se Hwan Park, Wandong Kim, Joo Yun Seo, Myounggon Kang, Kyung-Chang Ryoo, Jeong-Hoon Oh, Jong-Ho Lee, Hyungcheol Shin, and Byung-Gook Park, "Three-Dimensional NAND Flash Architecture Design Based on Single-Crystalline STacked ARray," *IEEE Transactions on Electron Devices*, vol. 59, no. 1, pp. 35-45, Jan. 2012.

[39] Stéphan Borel, Christian Arvet, Jeremy Bilde, Véronique Caubet, and Didier Louis, "Control of Selectivity between SiGe and Si in Isotropic Etching Processes," Japanese Journal of Applied Physics, Vol. 43, No. 6B, pp. 3964-3966, 2004.

[40] Jang-Gn Yun, Il Han Park, Wandong Kim, Jong Duk Lee, and Byung-Gook Park, "Extended Word-Line NAND Flash Memory," Japanese Journal of Applied Physics, Vol. 48, No. 8R, pp. 081203, 2009

[41] Ki-Tae Park, SeungChul Lee, Jong-Sun Sel, Jungdal Choi and Kinam Kim, "Scalable Wordline Shielding Scheme using Dummy Cell beyond 40 nm NAND Flash Memory for Eliminating Abnormal Disturb of Edge Memory Cell," Japanese Journal of Applied Physics, Vol. 46, No. 4B, pp. 2188-2192, 2007

초 록

최근 스마트폰과 스마트패드, 디지털 카메라, 포터블 음악 재생기 등에서 낸드 플래시 메모리의 요구가 끊임없이 증가하고 있다. 뿐만 아니라 데스크탑의 하드 디스크 드라이브가 낸드 플래시 메모리를 사용하는 솔리드 스테이트 드라이브로 대체되고 있다. 이러한 추세로 인하여 낸드 플래시 메모리의 축소화가 가속화하고 있다. 낸드 플래시 메모리가 축소화 되면서, 짧은 채널 효과와 신뢰성 문제가 더욱 심각해졌으며, 계속된 축소화로 인하여 공정 한계에 부딪히게 되었다.

이러한 문제들을 해결하기 위하여, 초고용량 저장 장치로서 3차원 적층형 낸드 플래시 메모리 구조들이 개발되고 소개되고 있다. 제안된 구조들 중에서 채널 적층형 어레이 (CSTAR)는 가장 뛰어난 전기적 특성을 보이나, 접근 방식의 복잡성이나 추가적인 공정 단계의 요구 등의 문제점이 보고 되고 있다.

본 논문에서는 기존의 채널 적층형 어레이가 가지는 문제점들을 해결하기 위해, 비트라인과 그라운드 선택 트랜지스터가 연결된 구조 (TiGer)를 가지는 채널 적층형 어레이를 제안하였다. TiGer 구조는 공통 소스 라인(CSL)과 비트라인(BL)의 어드레싱 만으로 적층된 층을 구분할 수 있다. 그라운드 선택 트랜지스터 (GST)가 비트라인과 연결되어 있으므로 그라운드 선택 라인 (GSL)과 그라운드 선택 라인 디코더가 필요하지 않으며, 이것이 TiGer 구조의 주요한 특징이다. TiGer 구조는 기존의 낸드 플래시 메모리와

다르게 GST를 개별적으로 컨트롤 할 수 없기 때문에, 메모리 동작을 성공적으로 수행하기 위해서 새로운 프로그램 동작 방법을 적용되어야 한다. 제안한 새로운 동작 방법을 검증하기 위하여 TCAD 시뮬레이션을 시행하였다.

적층 어레이를 만들기 위한 공정 플로우와 여러 단위 공정들을 개발하였고, 게이트-올-어라운드 (GAA) 형태의 단결정 실리콘 채널을 갖는 4층 적층 구조를 만들었다. 논문을 통해 Si/SiGe 에피택시얼 성장, SiGe의 등방성 선택적 식각, 적층된 나노 와이어의 형성, 그리고 다마신 공정을 보여주었고, 제작된 소자로 TiGer 구조의 메모리 동작을 검증하였다.

주요어: 3차원 적층형 낸드 플래시 메모리, 낸드 플래시 메모리 아키텍처, 채널 적층형 낸드 플래시 메모리, 나노와이어 전하 트랩형 낸드 플래시 메모리

학 번: 2009-30189