Ph.D. DISSERTATION

Methods for Threshold Voltage Setting of String Select Transistors in Channel Stacked NAND Flash Memory

채널 적층 낸드 플래시 메모리에서 스트링 셀렉트 트랜지스터의 문턱 전압 설정 방법

BY

DO-BIN KIM

August 2017

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

Methods for Threshold Voltage Setting of String Select Transistors in Channel Stacked NAND Flash Memory

채널 적층 낸드 플래시 메모리에서 스트링 셀렉트 트랜지스터의 문턱 전압 설정 방법

지도교수 박 병 국

이 논문을 공학박사 학위논문으로 제출함

2017 년 8 월

서울대학교 대학원

전기정보공학부

김 도 빈

김도빈의 공학박사 학위논문을 인준함

2017 년 8 월

위 원 장 : 박 영 준 (인)

부위원장 : 박 병 국 (인)

위 원 : 이 종 호 (인)

위 원 : 조 성 재 (인)

위 원 : 정 민 규 (인)

ABSTRACT

Since recent mobile electronic devices such as tablets, laptops, smartphones, or solid-state drives (SSDs) have started to adopt the NAND flash memory as their main data storage device, the demand for low-cost and high-density NAND flash memories has experienced a rapid increase. However, some problems such as the limitations of photolithography technology, cell-to-cell interference, and reduction of the number of electrons stored in floating gates have hindered the downscaling of floating-gate NAND flash memories. To overcome the NAND scaling issues, several types of three-dimensional (3D) stacked charge-trap NAND flash memories, which have been developed based on the bit-cost scalable (BiCS) technology introduced by Toshiba, have been widely investigated, owing to their scalability, ease of fabrication, and coupling-free characteristics.

3D-stacked NAND flash memory architectures can be divided into two categories. The first is the gate-stacked NAND flash memory, in which current flows through a vertical channel while the gates are shared horizontally by all the

strings. The second category consists of channel-stacked NAND flash memories, in which the current flows through the horizontally stacked channel and the gates are shared vertically by all the strings. In 3D-stacked NAND flash memory architectures. The channel-stacked type presents several outstanding advantages in terms of minimal unit cell size, bit line (BL) pitch scaling, use of a single-crystalline Si channel by Si/SiGe epitaxial growth process, and degradation characteristics of read currents caused by the increase in the number of stacked layers. However, compared with the gate-stacked type, the channel-stacked type presents critical issues that hinder its use in commercial applications, such as complex array architectures and decoding of the stacked layers. To overcome these problems, our group has recently reported channel-stacked arrays with layer selection by multilevel (LSM) operation. However, the array architecture and operation scheme setting the string select transistors (SSTs) with multilevel states should be simplified further to enable commercialization. In this dissertation, a simplified channel-stacked array with LSM operation is proposed. In addition, new SST threshold voltage (V_{th}) setting methods to set all the SSTs on each layer to the

targeted V_{th}s values are introduced and verified by using technology computer-aided design (TCAD) simulations and measurements in fabricated pseudo-SLSM. Furthermore, various disturbance phenomena that could occur during basic memory operations such as erase, program, and read are analyzed, and schemes for mitigating these disturbances are proposed and verified.

Keywords: 3-D stacked NAND flash memory, NAND flash memory architecture, gate stacked NAND flash memory, channel stacked NAND flash memory, layer selection by multi-level operation

Student number: 2013-30221

CONTENTS

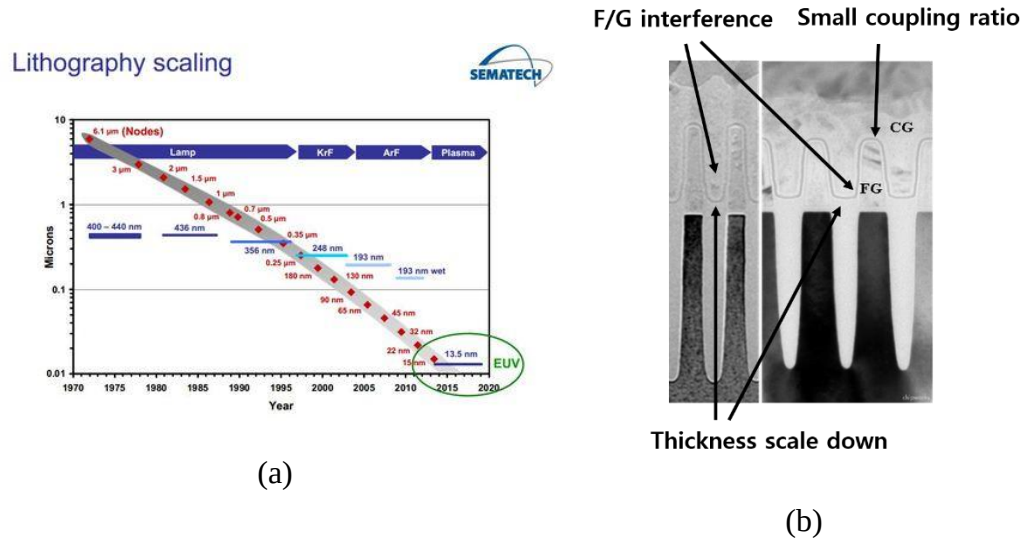
Chapter 1

Three-Dimensional Stacked NAND Flash Memory

1.1 Introduction to Three-Dimensional Stacked NAND Flash Memory

Since recent mobile electronic devices such as tablets, laptops, smartphones, and solid-state drives have started to adopt NAND flash memory as their main data storage device, the demand for low-cost, high-density NAND flash memory has been rapidly increasing. As the scaling down of NAND flash memory is accelerated, short channel effects and reliability problems of cell-to-cell interference and reduction of electrons stored in floating gates become more severe [1–5] and further scaling down faces process limitations such as imposed by current photolthography technology [6–8]. Fig. 1.1(a) shows the gap between required photolithography and NAND flash memory technology nodes. With simple line and space patterns, NAND flash has played a leading role in scaling design rules. However, there are

some concerns on whether it is possible to maintain the current scaling trend because of available lithography tools. Various lithography technologies and compensation technologies such as optical proximity correction (OPC), phase shift masking (PSM), and double patterning technology (DPT) are available for bridging the gap between actual lithography technologies and required design rules [6–10]. However, the gap will be widened beyond the 20 nm technology node until we find new breakthroughs. Until that time, it will be difficult to identify suitable solutions for compensating for the gap shrinking that will occur later.



(a) (b)

Fig. 1.1. (a) The gap between required photolithography and NAND flash memory technology nodes. (b) Major obstacles in scaling down NAND flash memory [10].

Moreover, up to now there have been many discussions about scaling barriers in NAND flash such as floating gate interference, edge word line (WL) disturbance, charge loss tolerence, small cell current, and coupling ratio [10, 11] as shown in Fig. 1.2(b). Fig. 1.2(b) shows the major obstacles in scaling down NAND flash memory. The volume of the floating gate (FG) decreases according to feature size, causing a shift in V_{th} induced by the small number of electrons. This means that the

fluctuation in the number of injected charges and bad retention characteristics caused by the tunneling oxide quality should be taken seriously in small feature size structures [9, 10]. Further scaling down also leads to short channel effect in NAND flash. As the gate length of the transistor shrinks, the doping concentration of the substrate must be increased to maintain the effective channel length and prevent source/drain (S/D) punch through. In NAND flash, however, high doping concentration can deteriorate the boosted potential of the channel at unselected bit lines (BLs) during the program-inhibit condition, which can narrow the V_{pass} bias operating window [10].

Despite these problems, maintaining a trend of increasing bit density and reducing bit cost of flash memory is in high demand. To achieve that, scaling of the device dimension has progressed aggressively and multilevel operation of cells has been developed. However, because the process cost per bit will increase more rapidly than the shrink rate, bit cost will not scale down in the near future, as shown in Fig. 1.2 [14]. Fig. 1.2 shows the relative per-bit manufacturing cost as a function of NAND flash memory capacity. As shown in Fig. 1.2, the pursuit of smaller

geometry has made it possible to reduce NAND flash memory cost while increasing capacity, but it may be possible that the pace of dimension shrinkage will slow or even stop completely from about the 2X nm generation.

To overcome these problems, three-dimensional (3D) stacked NAND flash memory has been proposed and reported [14–18]. A number of papers on 3D stacked NAND flash memory technology were presented. Fig. 1.3 shows the history of 3D stacked NAND flash memory [19]. In 2006, the vertically stacked NAND flash memory technologies faced a problem in that cost rose almost proportionally with the number of layers, because microfabrication was required one layer at a time.

In 2007, Toshiba proposed a bit-cost scalable (BiCS) 3D cell technology that held costs down even as the number of layers increased [15]. The technology alternately deposits interlayer dielectric films and polysilicon (poly-Si) electrodes for making many cells, then punches holes through the entire stack from top to bottom and fills them with oxide films, Si electrodes, and so on. As a result, manufacturing cost does not rise very much even if the number of layers increases.

In 2009, Toshiba modified this BiCS technology and used it to prototype a 32-Gbit NAND flash memory, about the same capacity as an actual product [20, 21]. Engineers combined a 16-layer cell with 2-bit-per-cell multivalue operation to shrink the cell area per bit to only 0.00082 μm^2 even with 60 nm manufacturing technology. The same cell area (2 bits/cell) would require 20 nm manufacturing technology to achieve without 3D stacked NAND flash memory technology. Samsung Electronics also proposed the terabit cell array transistor (TCAT) at VLSI 2009 [16]. The company listed all the shortcoming of BiCS technology, emphasizing that its proposed solution did not suffer from problems such as floating body issues and poly-Si gate issues [20].

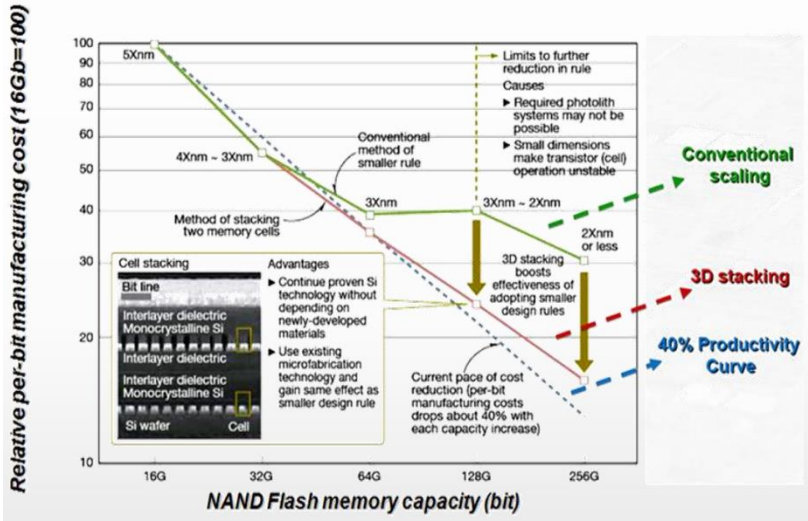


Fig. 1.2 Relative per-bit manufacturing cost as a function of NAND flash memory capacity [14].

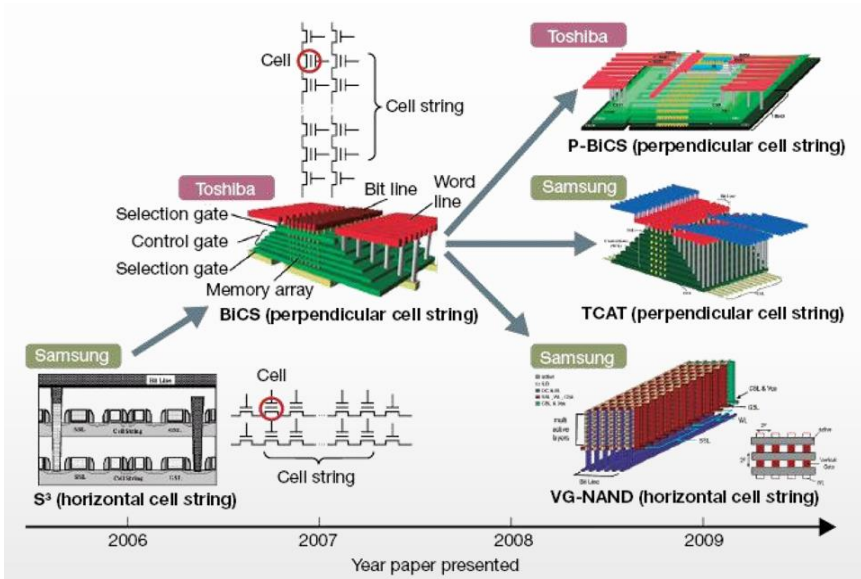


Fig. 1.3. History of 3D stacked NAND flash memory technology [19].

1.2 Gate Stack Type NAND Flash Memory

Several 3D stacked NAND flash memory architectures have been proposed so far [14–26]. The 3D stacked NAND flash memory architectures can be divided into two categories: gate stacked, in which current flows along a vertically formed channel and gates are shared horizontally by all strings, and channel stacked, in which current flows through the horizontally stacked channel and gates are shared vertically by all strings. Charge trapping SONOS (and MANOS) devices that use charge trapping for data storage are the most preferred as 3D stacked NAND flash memory architectures because of their easier process integration and smaller cell size. However, floating gate type 3D architecture has also been proposed for better cell memory windows [22].

Fig. 1.4 shows a schematic diagram and representative architectures of gate stack type NAND flash memory [16, 21]. As mentioned above, channel current flows through a vertically formed nanowire poly-Si channel and the cell structure is a gate-all-around structure. For the gate stack type, the bit density can be increased

by adding more gate plates, while the number of critical lithography steps remains constant because a whole stack of control gates and dielectric films is punched with only one lithography step.

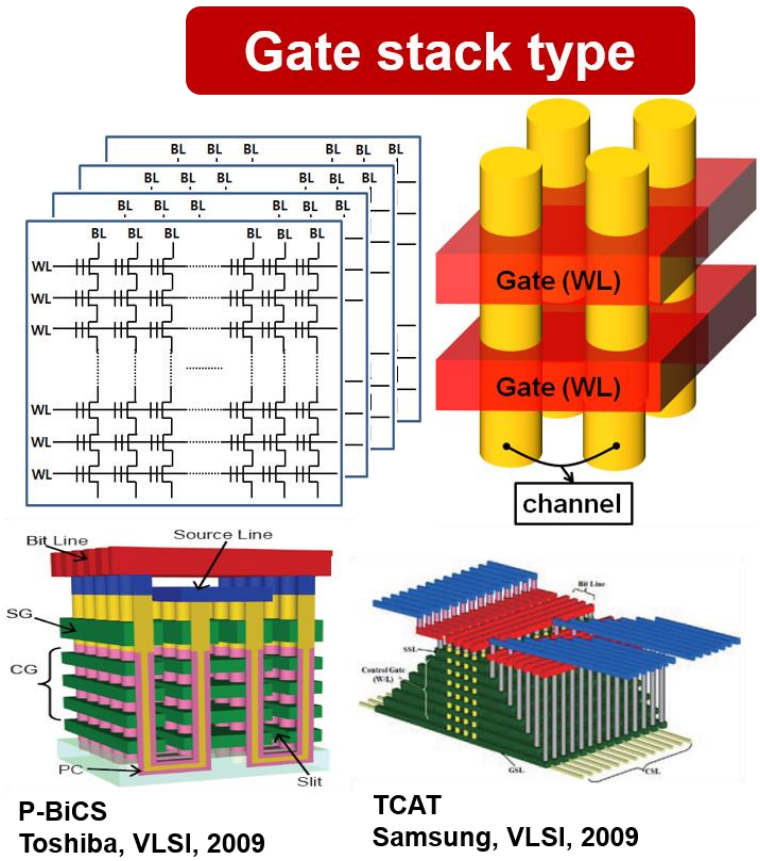


Fig. 1.4. Schematic diagram and representative architectures of gate stack type NAND flash memory [16, 21].

After the first pioneering work on BiCS flash memory in 2007, several gate stack type 3D NAND flash memory architectures were proposed. As shown in Fig 1.4, there are two representative structures of gate stack type NAND flash memory: pipe-shaped BiCS (P-BiCS) from Toshiba [21] and the TCAT from Samsung [16].

Fig. 1.5 shows the BiCS flash memory structure [15]. A string of cells is on the plugs located vertically in the holes punched through the whole stack of gate plates. Each plate acts as a control gate except for the lowest plate, which takes the role of the lower select gate (LSG). BiCS memory cell size is $6F^2/n$, where n is the number of control gate plates. A single cell is accessed by the control gate on the string, which is selected by a BL and a select gate (SG). The bottom of the memory plug is connected to source diffusion formed on the silicon substrate. For the erase operation of BiCS flash memory, the hole current generated by gate-induced drain leakage (GIDL) near the LSG is used. The edges of the control gate are processed into a stair-like structure for via holes, which are connected to driver transistors. The control gates are shared by several rows of strings, so that the area for the driver transistors does not increase even if more control gates are stacked.

In 2009, Toshiba introduced P-BiCS memory, which was developed from BiCS [21]. Fig. 1.6 shows a schematic diagram of P-BiCS NAND flash memory. The BiCS flash memory has three critical issues to be resolved. First, an SiN-based tunnel film is implemented to minimize the damage by diluted-HF treatment for the "gate-first" process; however, the program/erase (P/E) window which is vulnerable to read disturb and the data retention is insufficient to operate it as a multilevel cell (MLC). Second, the high-resistance source line is hardly clamped on its initial voltage owing to the large total current from a huge number of NAND strings during a read operation. Third, the LSG transistor is placed upon the heavily doped source diffusion and suffers from many thermal process steps during memory-film formation; therefore, the diffusion profile is not controllable. P-BiCS flash memory was proposed to overcome these issues.

As shown in Fig. 1.6, in P-BiCS memory, two adjacent NAND strings are connected at the bottoms by a 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 BL, and the other is bound by the source line (SL). The SL consists of the meshed

wiring of the third-level metal and is accessed by the first- and second-level metal as in conventional planar technology; therefore, the resistance of the SL is sufficiently low. Both the SG transistors are placed below the SL and the BL. The control gates (CGs) are isolated by the slit and face each other in a couple-of-combs pattern. The memory stack consists of a blocking layer, a charge trap layer, and an oxide-based layer as a tunnel dielectric. The oxide-based tunnel layer is implemented because the sequential processing from the deposition of the tunnel layer to the body silicon is done during the fabrication process for the P-BiCS NAND strings.

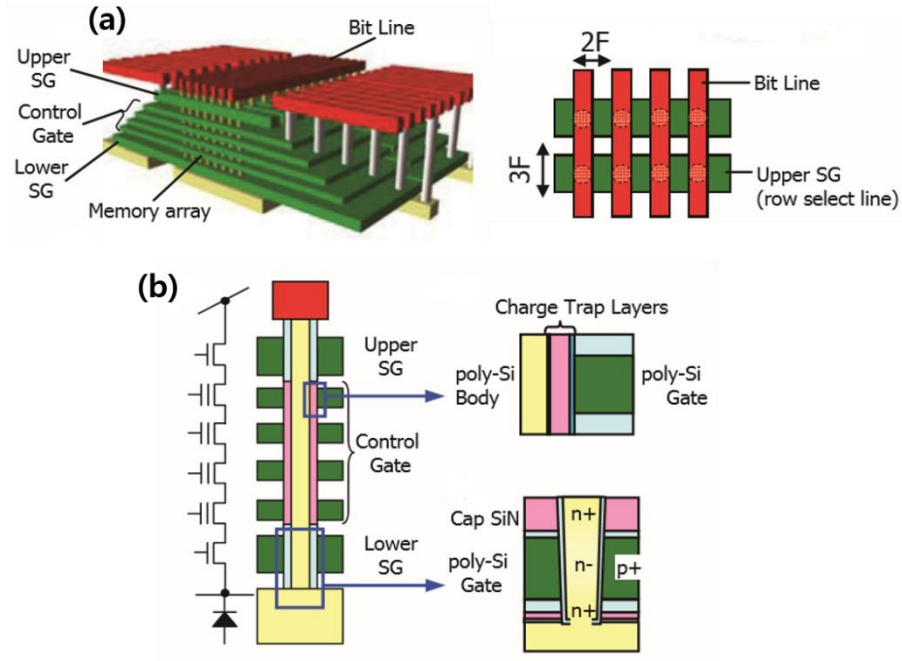


Fig. 1.5. BiCS NAND flash memory by Toshiba. (a) Bird's eye view and (b) schematic diagram of one string [15].

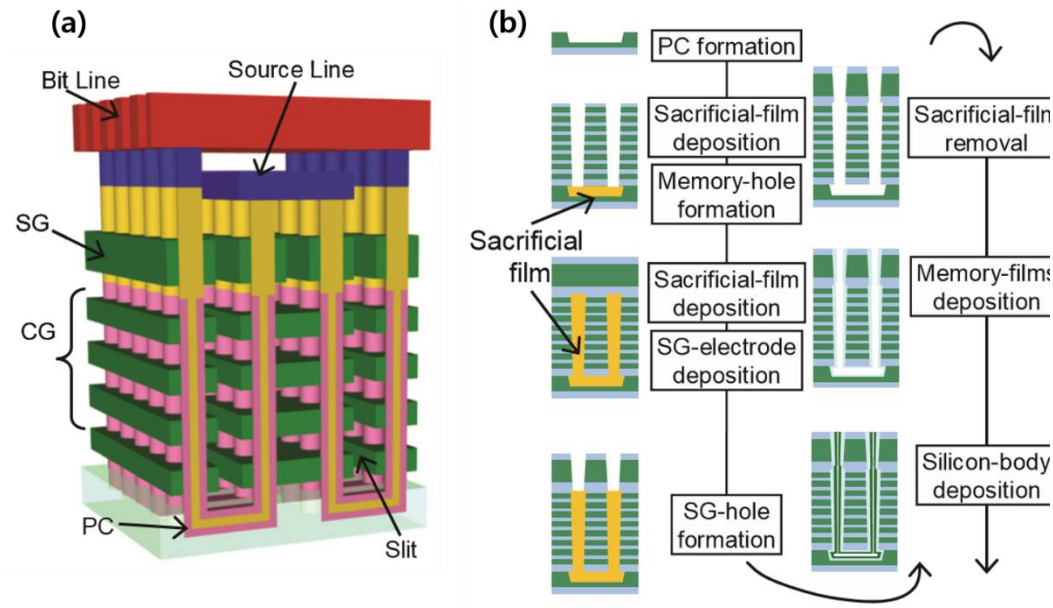
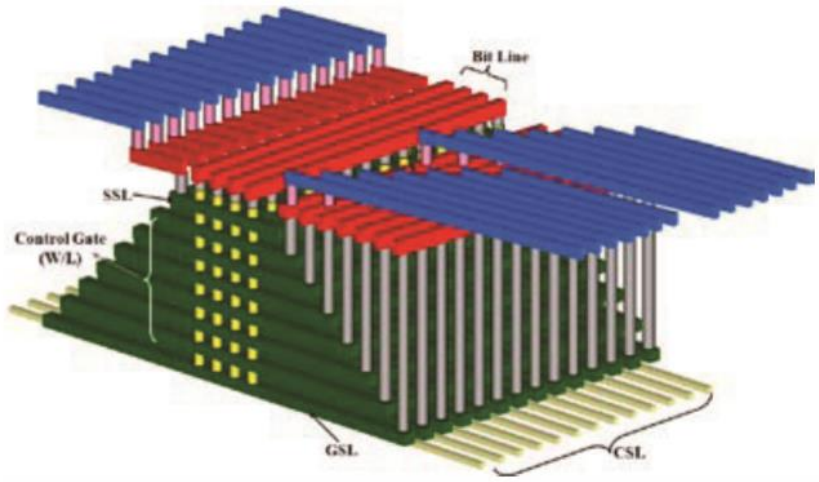


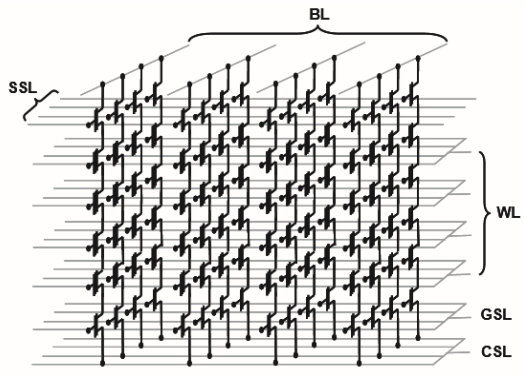
Fig. 1.6. P-BiCS NAND flash memory by Toshiba. (a) Bird's eye view and (b) process flow [21].

As previously mentioned, Samsung Electronics introduced the TCAT in 2009 [16]. There are two major concerns for BiCS flash memory. First, it is almost impossible to implement a metal gate structure for BiCS flash memory because it is very difficult to etch a metal/oxide multilayer simultaneously. BiCS flash lacks the various advantages of MONOS cell structure, such as faster erase speed, wider V_{th} margin, and better retention characteristics [16]. Another concern is GIDL erase

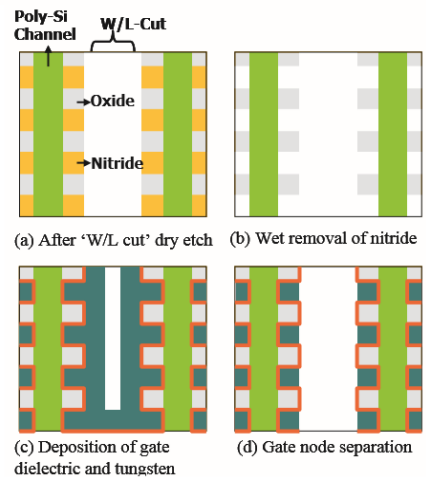
of BiCS flash memory. An extensive circuit change may be necessary to apply a negative bias on a word line during an erase operation. An area penalty and a limited erase voltage are expected. TCAT overcomes these problems. Fig. 1.7 shows a schematic birds-eye view and the equivalent circuit of the TCAT flash memory. The process sequence is shown in Fig. 1.7. There are a number of distinctive structural differences between BiCS flash memory and TCAT flash memory, including an oxide/nitride multilayer stack, a line type “W/L cut” etched through the whole stack between each row array of the channel poly plug, a line-type common source line (CSL) formed by an implant through the W/L cut, and replaced metal gate lines for each row of the poly plug. The most unique process is “gate replacement” to achieve the MONOS structure. Fig. 1.7(c) shows a detailed view. After the W/L cut dry etch and wet removal of the sacrificial nitride layer, gate dielectric layers and gate metal are deposited in the conventional order. This is not “gate-first” process as in BiCS flash memory fabrication. The cell string has NAND cell transistors with an string select line (SSL) transistor at the top and the a ground select line (GSL) transistor at the bottom.



(a) Bird's eye view



(b) Equivalent circuit



(c) Gate replacement process flow

Fig. 1.7. TCAT flash memory by Samsung Electronics [16].

1.3 Channel Stack Type NAND Flash Memory

Fig. 1.8 shows a schematic diagram and representative architectures of channel stack type NAND flash memory [23–26]. As mentioned above, channel current flows horizontally through a horizontal nanowire silicon channel and the cell structure is a double gate or a gate-all-around structure. For the channel stack type, the bit density can be increased by adding more stacked channels, while the number of critical lithography steps remains constant because the whole stack of the channel is dry etched with only one lithography step.

While research papers and proposals about gate stack type NAND flash memory such as BiCS, P-BiCS, and TCAT were published, several channel stack type 3D NAND flash memory architectures were also proposed. As shown in Fig. 1.8, there are two representative structures of channel stack type NAND flash memory: vertical gate (VG) NAND architecture [23] and channel stacked array architecture (CSTAR) [26]. Channel stacked NAND flash memory has good pitch scalability and thus is very attractive. However, it is more difficult to decode the BL in a

channel stack architecture, thus decoding innovations are required for a compact array architecture design. More rigorous analysis of gate stack and channel stack types will be discussed in the next section.

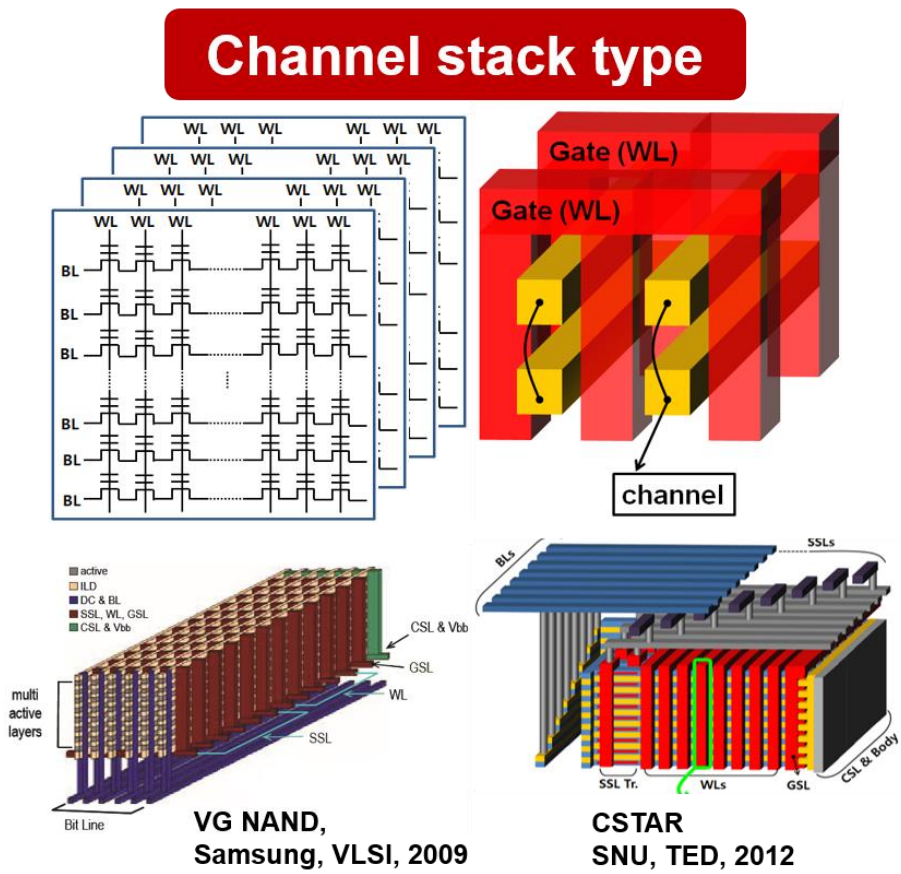


Fig. 1.8. Schematic diagram and representative architectures of channel stack type NAND flash memory [23, 26].

Samsung Electronics introduced vertical gate NAND flash memory (VG-NAND) in 2009 [23]. Fig. 1.9 shows a bird's eye view of the VG-NAND flash memory and its equivalent circuit. VG-NAND flash memory includes WLs, BLs, CSLs, horizontal active strings with a pattern, VGs (for SSLs, WLs, and GSLs), charge trap layers between the active and the vertical gate, vertical DC plugs, and a source and active body (V_{bb}), as shown in Fig. 1.9. WLs and BLs are formed at the beginning of fabrication before the cell array is constructed, making the interconnect between WLs and BLs and the decoder easier. The source and active body (V_{bb}) are electrically tied to the CSL for enabling the body erase operation. A positive bias is applied to the CSL during erase operation. The array schematic of each layer is identical to that of the planar NAND flash memory except for the SSL.

VG-NAND requires 6 SSLs for 8 active layers and 8 SSLs for 16 active layers. The reason for the multiple SSLs is to enable selection of data from a chosen layer out of multilayers since VG-NAND uses a common BL and a common WL between multi-active layers. The cell size of the VG-NAND is $4F^2$ per layer, as shown in Fig. 1.10, and the active dimension can be reduced without any scaling issue,

making the cell size <$4F^2$ per layer. Fig. 1.10 explains the VG-NAND integration sequence; the integration scheme is based on simple patterning and plugging. A BL having n^+ poly-Si is fabricated first and then an 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 alternating interlayer dielectrics are inserted between active layers. Then patterning is performed on the multiple active layers and charge trap layers (ONO) are deposited over the patterned active layers. Consecutively, a VG is formed and connected to the WL.

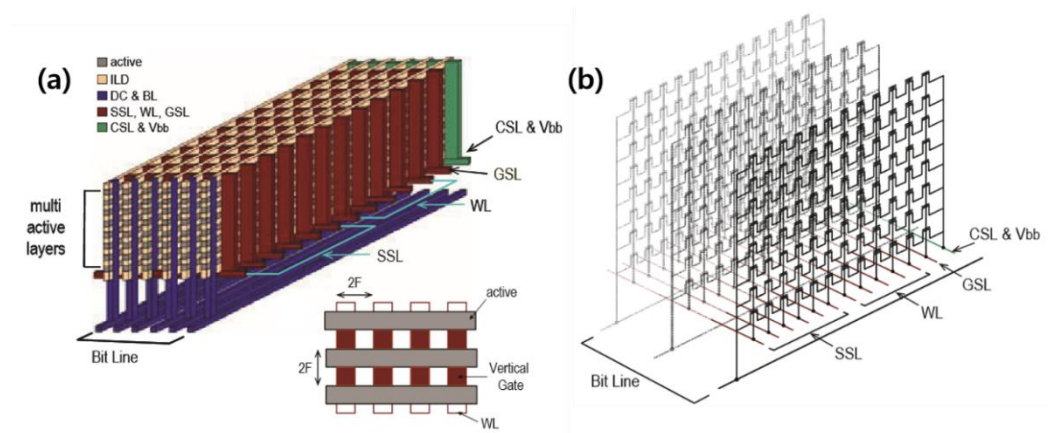


Fig. 1.9. Multiple layered VG-NAND. (a) Bird's eye view and (b) equivalent circuit.

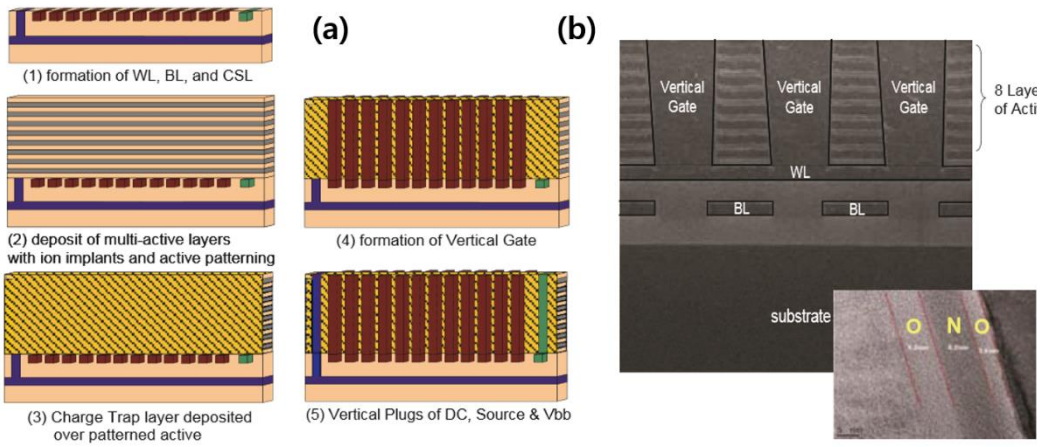
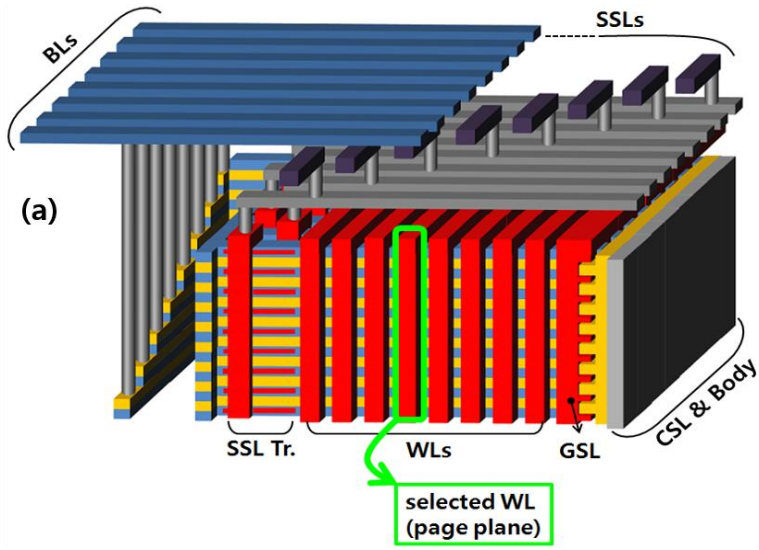


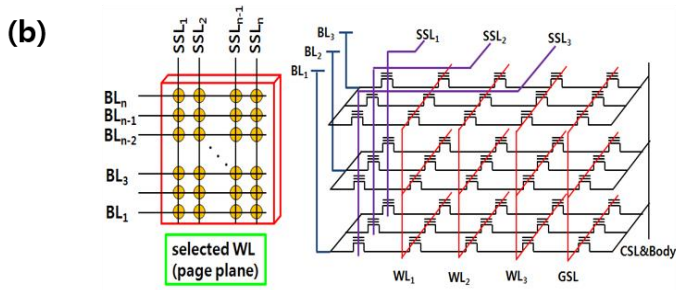
Fig. 1.10. (a) Process flow of VG-NAND flash memory. (b) Scanning electron microscopy image of VG-NAND flash memory along the WL direction.

CSTAR architecture was introduced in 2012 [26]. Fig. 1.11 shows the proposed CSTAR structure for the 3D NAND flash architecture. This unit structure is defined as “building.” The specific descriptions of each part of the building are as follows. The 3D stacked NAND flash memory, unlike conventional two-dimensional (2D) planar NAND flash memory, needs one more address that is determined by the SSLs. In conventional 2D NAND, the WL selects a “page (line),” whereas the WL of the 3D 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 of 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 that it uses a single-crystalline silicon channel and a gate-all-around structure. As mentioned earlier, VG-NAND has a poly-Si channel and a double gate structure. The CSTAR has the merit of more cell current drivability and short channel effect immunity.





	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

Fig. 1.11. Channel stacked NAND flash memory. (a) Bird's eye view of the unit building structure and (b) equivalent circuit and operation scheme [26].

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

Table 1.1 shows the various 3D stacked NAND flash memory architectures [16, 21, 23, 26]. In this section, we discuss the pros and cons beween gate stack type NAND flash memory and channel stack type NAND flash memory.

Gate stack type memory has gate-all-around structure and a macaroni poly-Si channel. This poly-Si channel shows poorer current drivability owing to its lower mobility than that of a single-crystalline silicon channel. For gate stacked NAND flash memory, the bit density can be increased by adding more gate stack layers. However, the greater the number of gate stack layers, the smaller the cell current is. Because of the lower 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. The number of cells in a string is constant even if the number of channel stack layers is increased. Consequently, the cell current is not degraded by increasing of

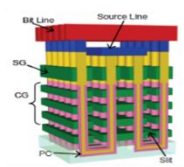
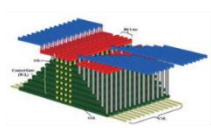
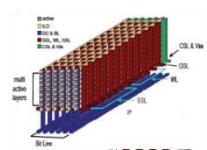
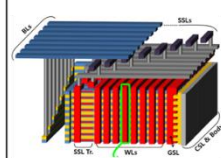
stack layers in the channel stacked flash memory. However, it is more difficult to decode stack layers in channel stacked NAND flash memory architectures. It can be easy to decode stack layers because of the common gate plate structure in gate stack type NAND flash memory. Therefore, decoding innovations are required for a compact array architecture design for channel stacked NAND flash.

Most published 3D nano-wire gate stack type NAND flash memory architectures have relatively large cell size ($6F^2$ with F > 60 nm). To offer a cost-competitive product compared with 20 nm floating gate planar type NAND flash memory, it must be stacked with more than 32 gate plate layers. If the pitch is not scalable below 50 nm, to provide a terabyte of memory requires a stacking of >128 layers. Although there is no direct physical limitation for fabricating 128 stacked layers, it would be very challengeing given actual etching capability, process throughput, cost, and yield loss.

Nonetheless, channel stack type NAND flash memory has better pitch scalability than gate stack type NAND flash memory. Fig. 1.12 shows the layout top view of gate stack type and channel stack type NAND flash memory. For gate stack type,

each drilled hole must be located inside one SSL line for decoding. The drilled hole size is limited by the minimal ONO thickness and the channel diameter. There are also stricter lithography constraints such as the overlay between the drill hole and the SSL and the minimal SSL/WL cut space. However, channel stack type is not similarly constrained in cell size. For channel stack type, the limiting factor is the finite ONO thickness that restricts the BL pitch.

Table 1.1 Various 3D stacked NAND flash memory architecture [16, 21, 23, 26]

	GATE STACK TYPE		CHANNEL STACK TYPE	
	P-BICS	TCAT	VG NAND	CSTAR
STRUCTURE				
CELL STRUCTURE	Macaroni & GAA	Macaroni & GAA	Double Gate	Gate-All-Around
CAHNNEL	Poly-Si channel	Poly-Si channel	Poly-Si channel	Single crystalline Si channel
KEY ISSUE	Low read current Reliability Pitch scaling	Low read current Reliability Pitch scaling	Number of lithography Implantation	Layer decoding Wiring complexity

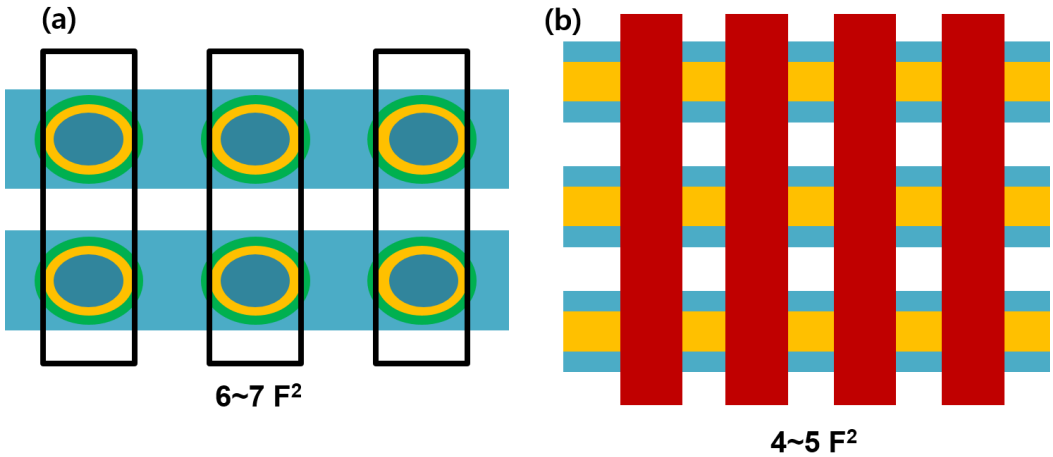


Fig. 1.12. Layout top view of (a) gate stack type and (b) channel stack type NAND flash memory.

Chapter 2

Channel Stacked NAND Flash Memory with Layer Selection by Multilevel Operation

In this chapter, the proposed channel stacked NAND flash memory with layer selection by multilevel operation (LSM) is discussed [27]. The 3D stack NAND flash architecture design with LSM and its advantages were studied via a multidirectional investigation, and a numerical simulation was performed to verify the operation scheme. The proposed stack NAND flash memory architecture with LSM can be an innovative solution to overcome various problems of conventional channel stacked NAND flash memory.

2.1 LSM and Channel Stacked NAND Flash Architecture Design

Fig. 2.1 shows a bird's eye views of gate stack type and channel stack type NAND flash memory [21, 26]. For the gate stack type, cell selection is relatively

easy. Because the gate plate is common, not only SSL patterning but also BL patterning is possible by lithography and dry etching. However, it is more difficult to decode the stacked layers and the cell in the channel stacked NAND flash memory. As the BL is common in the selected layers, the design of SSLs must be of island-type to separate leads to the various problems of wiring complexity, limitation of BL pitch scaling, low BL throughput, and restriction on building expandability [26]. Therefore, decoding innovations are required for a compact array architecture design for channel stacked NAND flash memory.

In this chapter, a novel layer selection method and stacked NAND flash architecture are proposed to overcome these problems. Fig. 2.2 shows a bird's eye view of the proposed stacked NAND flash architecture. The proposed architecture is a channel stacked NAND flash array, but there is no island-type SSL. Channels are formed in the horizontal direction, and SSLs are formed parallel to the WLs as in a planar type NAND flash array. The stacked layers can be distinguished by an electrical initialization of specific SSLs of the stacked layers. To access a selected layer electrically, the SSLs have O/N/O dielectric layers for initial P/E operations.

String select transistors (SSTs) in the vertical direction have different threshold voltages by programming the charge trapping layer to enable LSM.

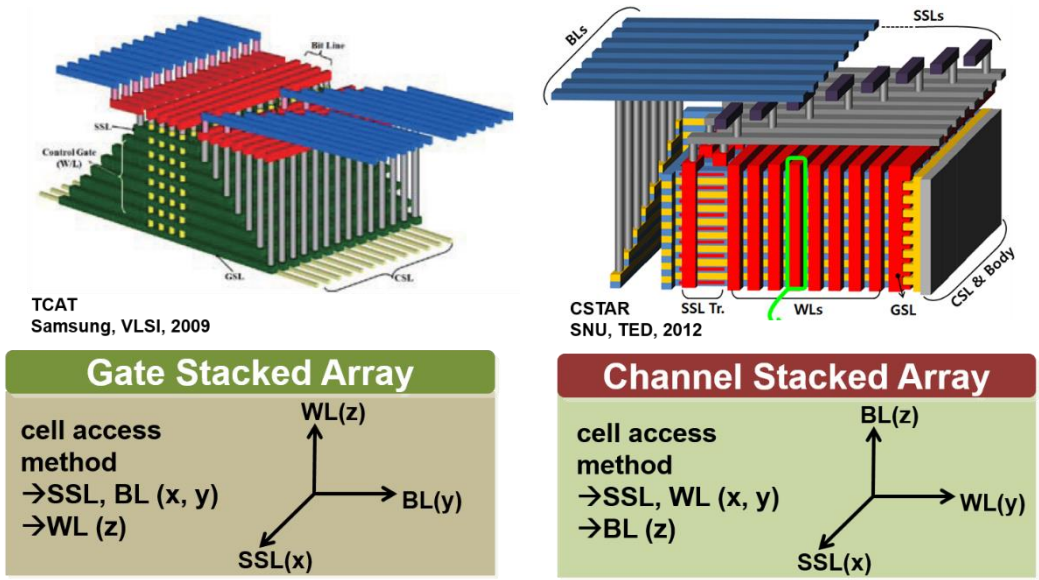


Fig. 2.1 Bird's eye view of the gate stacked array and channel stacked array [13, 20].

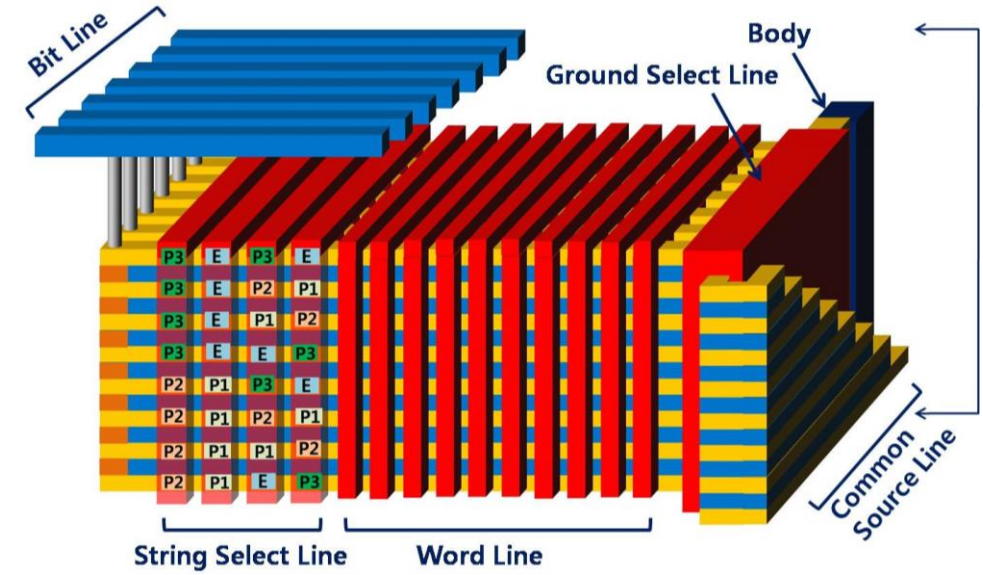


Fig. 2.2. Bird's eye view of the channel stacked NAND flash memory with LSM.

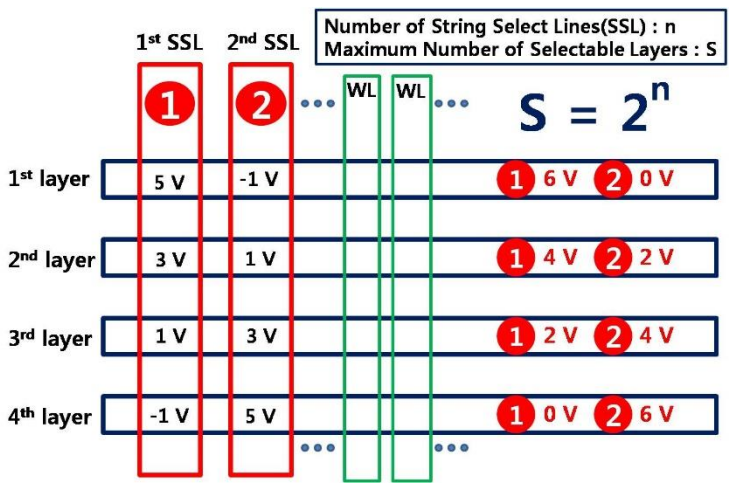


Fig. 2.3. LSM scheme. Example of the arrangement of multilevel SST V_{th} and bias scheme with four stacked layers and two string select lines.

For layer selection by LSM, the SSLs in a layer have different multilevel states and also different SSL bias levels. Fig. 2.3 shows an example of the arrangement of multilevel SSL V_{th} and bias schemes with four stacked layers and two SSLs. As shown in Fig. 2.3, the LSM method makes various layer selection cases through the combinations of SST V_{th} states and bias levels.

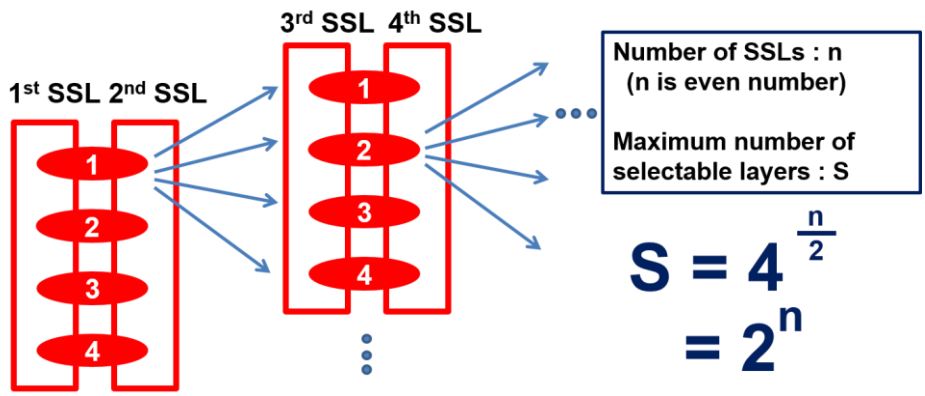


Fig. 2.4. Maximum number of selectable layers by LSM.

In the initially proposed method and array, the number of SSLs should be an even number. This is due to the pair SSL operation of the threshold voltage and the SSL bias. A pair of SSLs is able to select four stacked layers. This results in a maximum

number of selectable stacked layers of 2^n, where n is the number of SSLs. Because layer selection can be performed by the bias of the SSL, the placement of BLs and WLs is similar to that in the conventional planar structure. This resolves various issues of channel stacked NAND flash memory arrays such as BL pitch scaling, wiring complexity, and read throughput limitation.

However, owing to imperfect strategies for arranging the V_{th} states in the proposed LSM, more SSLs are required than necessary. Therefore, the layer selection by multilevel permutation (LSMP) method was proposed for maximizing the number of possible V_{th} arrangements with layer selection by permutation [22]. The relevant aspects that explain how to adopt the permutation method to LSM and the results for the maximum number of selectable layers by using LSMP are presented in Tables 2.1–2.3 [27–29].

Table 2.1. Multisets used in the LSMP method. The condition is that the sum s of all elements is equal to $[n/2 \times (k-1)]$, where n is the number of SSLs and k is the number of states [28].

number of SSLs	2 states	3 states	4 states		5 states		
2	{1,0}	{1,1} {2,0}	{2,1}	{3,0}	{2,2}	{3,1}	{4,0}
3	{1,0,0}	{1,1,1} {2,1,0}	{2,1,1} {2,2,0}	{3,1,0}	{2,2,2} {3,3,0}	{3,2,1} {4,2,0}	{4,1,1}
4	{1,1,0,0}	{1,1,1,1} {2,2,0,0} {2,1,1,0}	{2,2,2,0} {2,2,1,1} {3,3,0,0}	{3,2,1,0} {3,1,1,1}	{2,2,2,2} {3,3,2,0} {3,3,1,1}	{3,2,2,1} {4,4,0,0} {4,3,1,0}	{4,2,2,0} {4,2,1,1}
5	{1,1,0,0,0}	{1,1,1,1,1} {2,2,1,0,0} {2,1,1,1,0}	{2,2,2,1,0} {2,2,1,1,1} {3,3,1,0,0}	{3,2,2,0,0} {3,2,1,1,0} {3,1,1,1,1}	{2,2,2,2,2} {3,3,3,1,0} {3,3,2,2,0} {3,3,2,1,1}	{3,2,2,2,1} {4,4,2,0,0} {4,4,1,1,0} {4,3,3,0,0}	{4,3,2,1,0} {4,3,1,1,1} {4,2,2,2,0} {4,2,2,1,1}
6	{1,1,1,0,0,0}	{1,1,1,1,1,1} {2,2,2,0,0,0} {2,2,1,1,0,0} {2,1,1,1,1,0}	{2,2,2,2,1,0} {2,2,2,1,1,1} {3,3,3,0,0,0} {3,3,2,1,0,0}	{3,3,1,1,1,0} {3,2,2,2,0,0} {3,2,2,1,1,0} {3,2,1,1,1,1}	{2,2,2,2,2,2} {3,3,3,3,0,0} {3,3,3,2,1,0} {3,3,3,1,1,1} {3,3,2,2,2,0} {3,3,2,2,1,1}	{3,2,2,2,2,1} {4,4,4,0,0,0} {4,4,3,1,0,0} {4,4,2,2,0,0} {4,4,2,1,1,0} {4,4,1,1,1,1}	{4,3,3,2,0,0} {4,3,3,1,1,0} {4,3,2,2,1,0} {4,3,2,1,1,1} {4,2,2,2,2,0} {4,2,2,2,1,1}

Table 2.2. LSMP operation scheme example with three SSLs and three states ($V_m < V_{bm} < V_{m+1}$, for $m = 0, 1, 2$) [28].

	V_{th} of SST			SSL bias for selection		
Layer #	1st SSL	2nd SSL	3rd SSL	1st SSL	2nd SSL	3rd SSL
1	V_2	V_1	V_0	V_{b2}	V_{b1}	V_{b0}
2	V_2	V_0	V_1	V_{b2}	V_{b0}	V_{b1}
3	V_1	V_2	V_0	V_{b1}	V_{b2}	V_{b0}
4	V_0	V_2	V_1	V_{b0}	V_{b2}	V_{b1}
5	V_1	V_0	V_2	V_{b1}	V_{b0}	V_{b2}
6	V_0	V_1	V_2	V_{b0}	V_{b1}	V_{b2}
7	V_1	V_1	V_1	V_{b1}	V_{b1}	V_{b1}

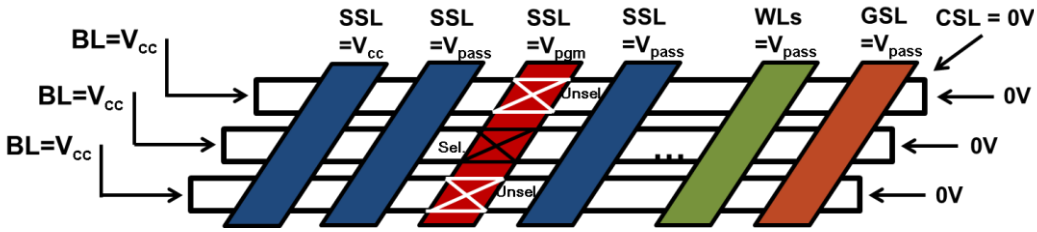
Table 2.3. Maximum number of selectable layers by layer selection by erase (LASER), LSM, and LSMP [27–29].

number of SSLs	maximum number of selectable layers							
	LASER	LSM			LSMP [This work]			
	2 states	3 states	4 states	5 states	2 states	3 states	4 states	5 states
2	2	3	4	5	2	3	4	5
3	3	-	-	-	3	7	12	19
4	6	9	16	25	6	19	44	85
5	10	-	-	-	10	51	155	381
6	20	27	64	125	20	141	580	1751

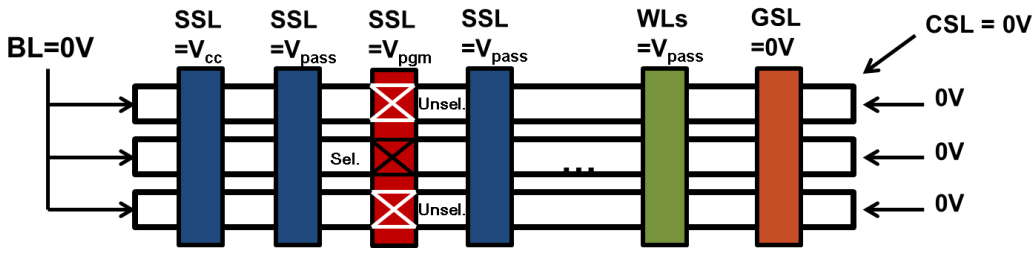
2.2 Operation Scheme of Channel Stacked NAND Flash Memory with LSM

2.2.1 Stacked SST Initialization to Enable LSM

To perform LSM, stacked SSTs should be initialized to have specific threshold voltages. However, owing to the characteristics of the proposed structure, bit-by-bit inhibit operations in the LSM structure are impossible. Unlike the proposed channel stacked NAND flash memory structure, the LSM has an SSL, which is not separated but made like a WL. Therefore, it is not possible to inhibit unselected SSTs that are located on the same layer or share a BL with the selected SST while the selected SST is under program operation. This problem cannot be solved by the previously proposed source side program scheme or the BL side program scheme [30–36]. The above description can be confirmed in detail in Fig. 2.5. Therefore, a new scheme or structure modified for a V_{th} setting of SSTs is needed to implement LSM for channel stacked NAND flash memory.



(a)



(b)

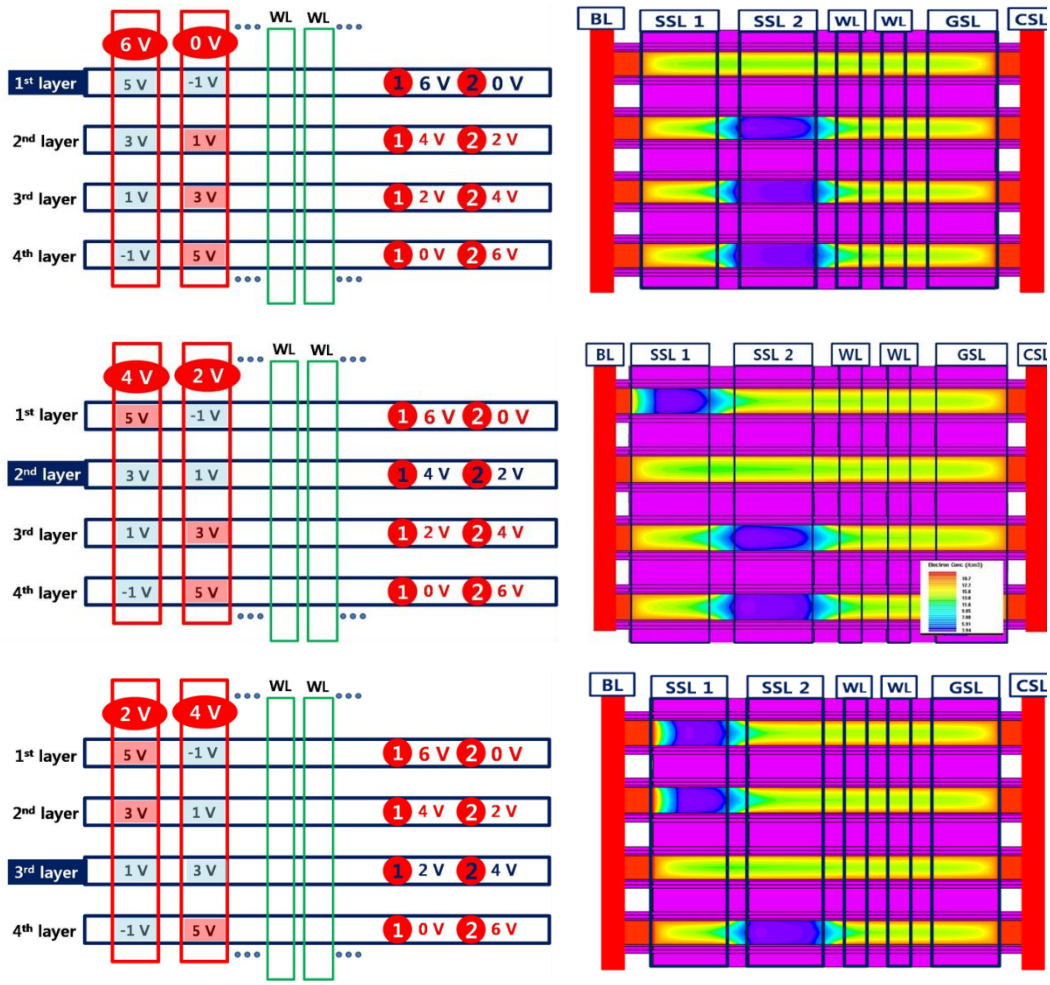
Fig. 2.5. SST initialization cannot be achieved by using conventional self-boosting schemes. (a) Bit-by-bit inhibition failure of unselected SSTs when source side program scheme is used for initialization. (b) BL side program scheme.

2.2.2 Read Operation with LSM

Once the SSTs are initialized to have a required threshold voltage level, LSM is possible. Fig. 2.6 shows the layer selection operation scheme and electron concentration of the stacked channels under the LSM bias condition. As shown in Fig. 2.6, by differently programming the charge storage layer of each of the SSTs, the first vertically stacked string selection transistors can be configured to form the first SSL threshold voltages that sequentially decrease toward the lower layers, while the second vertically stacked SSTs formed by the second string selection line have threshold voltages that are sequentially increased toward the lower layers. By combining the four multilevel states of the SST and LSM bias schemes, two string select lines can distinguish four stacked layers.

In other words, as shown in Fig. 2.6, when 6 and 0 V are applied to the first string select line and the second string select line, respectively, all string select transistors of the first layer having 5 and −1 V threshold voltage are turned on and the first layer is selected, namely, electrically connected to a BL, while the other unselected

layers cannot be electrically connected to BLs, because the second vertically stacked string select transistors having 1, 3, or 5 V threshold voltage are turned off.



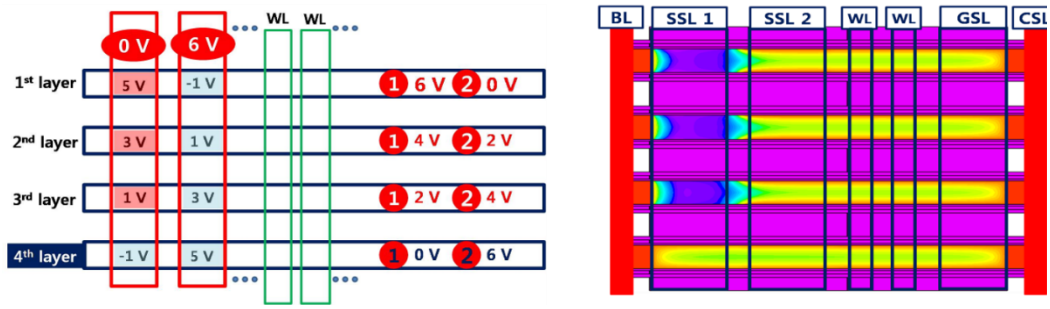


Fig. 2.6. Layer selection operation scheme and electron concentration of the stacked channels under the LSM bias condition.

By the same method, when 4 and 2 V are applied to the first string select line and the second string select line, respectively, all string select transistors of the second layer are turned on and the second layer is connected to a BL, while the other unselected layers cannot be connected to BLs because at least one SST formed on each of the other layers is turned off. The selection schemes for the third and fourth layers and simulation results are also depicted in Fig. 2.6. Fig. 2.7 shows the transfer characteristics of the cells in selected and unselected layers. For the selected layer, the cell is turned on, while the cell of the unselected layer is turned off under the LSM bias condition, as shown in Fig. 2.7.

As mentioned above, when there are two SSLs, four layers can be selected because the number of cases (namely, the number of combinations of threshold voltage arrangements) is four. If additional SSLs are used in the structure, a larger number of layer selection can be realized using LSMP than LSM, as shown in Table 2.3.

2.2.3 Program / Erase Operation with LSM

For the channel stacked NAND flash memory array with LSM, program and erase operation schemes are similar to those of conventional planar NAND flash memory except for layer selection. Fig. 2.7 shows the equivalent circuit diagram for program operation in the channel stacked NAND flash memory with LSM. Layer selection is performed by the LSM method. In the selected layer, BLs and WLs perform program operations similar to those of conventional NAND flash memory.

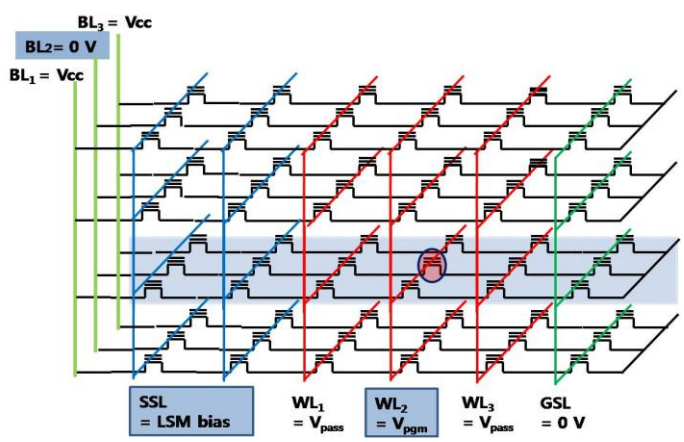


Fig. 2.7. Equivalent circuit diagram for program operation in LSM.

Figs. 2.8(a) and (b) show two different cases for program inhibition in the channel stacked NAND flash memory array. One is unselected BL program inhibition in the selected layer, and the other is unselected layer program inhibition in the selected BL.

In the case of unselected BL program inhibition in the selected layer, when V_{cc} is applied to the BL and SSLs of the layer is under the selection bias, self-boosting program inhibition occurs. Otherwise, in the case of unselected layer program inhibition in the selected BL, the BL is applied with 0 V. However, SSTs of the unselected layer are turned off, and then self-boosting program inhibition also occurs. For maintaining the self-boosting in the inhibit channel, ground select transistors must be turned off as in conventional NAND flash memory operation. In the case of a selected string, as 0 V is applied to the BL and SSTs are turned on, a 0 V bias of the BL is applied to the channel of the selected string. Therefore, it is possible to program the selected cell in the selected string, while the unselected cells are under the self-boosting program inhibition, as shown in Fig. 2.8.

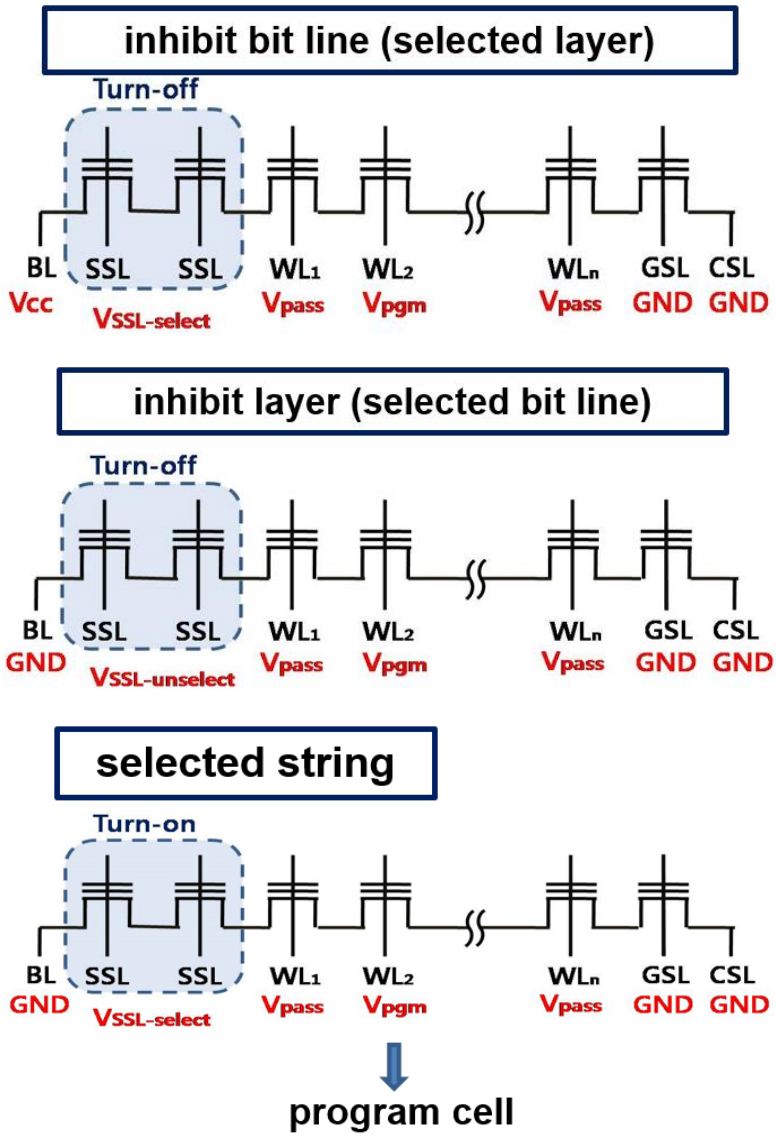


Fig. 2.8. (a) Unselected BL program inhibition in the selected layer, (b) unselected layer program inhibition in the selected BL, and (c) program operation in the LSM array.

The erase operation in the channel stacked NAND flash memory array with LSM is the same as that in the conventional NAND flash memory array, because the body and GSL are designed to connect the channel to the body. All the word lines are set to 0 V, and a high erase voltage is applied to the channel through the body contact to induce Fowler–Nordheim tunneling so that the erase operation is possible. All the CSL, SSL, and BL voltages are floated. The program and erase operations in the LSM array were successfully verified by the fabricated CSTAR with LSM, as shown in Fig. 2.9. The program pulse voltage is 13 V, and program pulse duration time is 80 μs. The erase pulse voltage is −12 V, and the erase pulse duration time is 3 ms. A voltage of 6 V is applied to the pass WL to perform the program operation. As shown in Fig. 2.9, a >3.5 V threshold voltage window can be obtained, though relatively lower program and erase pulse voltages are applied than in the conventional planar NAND flash memory device. This is due to the field concentration effect of the gate-all-around structure [37]. When a coaxial capacitor is biased, the electric field of the inner surface is larger than that of the outer surface by Gauss's law, and this phenomenon is referred to as the electric field concentration

(enhancement) effect.

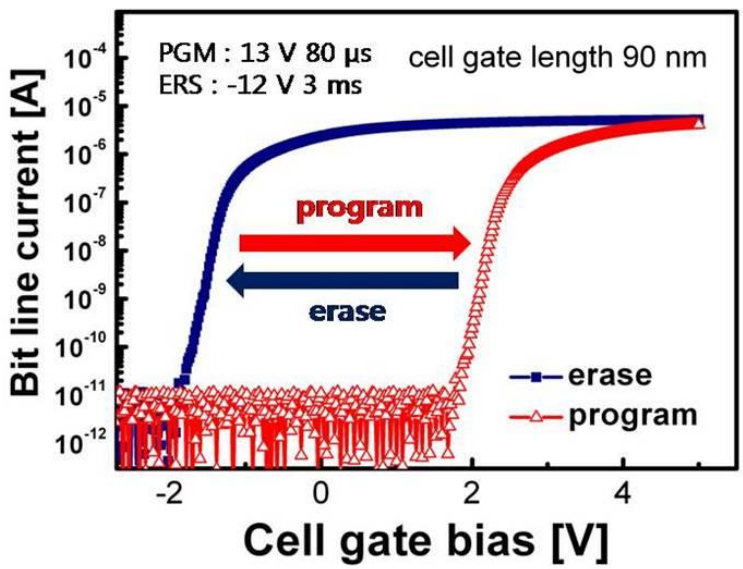


Fig. 2.9. Program and erase characteristics of fabricated NAND flash array.

2.3 Comparison with Conventional Channel Stacked NAND Flash Memory Architecture

The proposed CSTAR with LSM holds many advantages over the conventional CSTAR array. Fig. 2.10 shows a bird's eye view and a top view (layout) of the conventional CSTAR array and the CSTAR with LSM. In the case of the conventional CSTAR array, SSLs must be of island type for channel separation in a selected array, because the BL is common in a layer. So, minimum BL pitch is constrained by the lithography and dry etch margin to form the island-type SSLs. Moreover, according to the operation scheme, the SSL bias induces interference between the neighboring channel and the SSL. Additionally, the conventional CSTAR has complex wiring. Too many contacts are needed for the island-type SSL and the number of metal layers is increased to form different levels of WLs and SSLs.

However, in the case of CSTAR with LSM, the BL pitch can be scaled down compared with conventional CSTAR, because LSM enables line type SSLs. Scaling

down the BL pitch offers an area advantage of increased cell density. The CSTAR array is also free of interference between neighboring channels and SSLs owing to the shielding of the gate-all-around structure. In addition, it is not necessary to use different levels to form the WLs and SSLs, so wiring complexity can be reduced.

As mentioned above, it is more difficult to decode a stacked layer and cell in a conventional channel stacked NAND flash memory. This decoding difficulty leads to various problems such as wiring complexity, limitation of BL pitch scaling, low BL throughput, and restriction on building expandability [20]. Therefore, decoding innovations are required for a compact array architecture design for channel stacked NAND flash memory. The proposed CSTAR with LSM can be a promising solution to overcome these problems. The proposed architecture has no island-type SSL in spite of the channel stacked NAND flash array. Channels are formed in the horizontal direction, and SSLs are formed in the other horizontal direction to be parallel to WLs like a planar type NAND flash array. Layer selection can be performed by the electrical initialization and the SSL bias. By virtue of the above-mentioned features, the channel stacked NAND flash memory with LSM solves the

problems of conventional CSTAR.

However, owing to the characteristics of the proposed structure, bit-by-bit inhibit operation in the LSM structure is impossible. Therefore, a new scheme or structure modified for V_{th} setting of SSTs is needed to implement LSM to channel stacked NAND flash memory for exploiting various advantages of LSM and continuous NAND flash memory scaling.

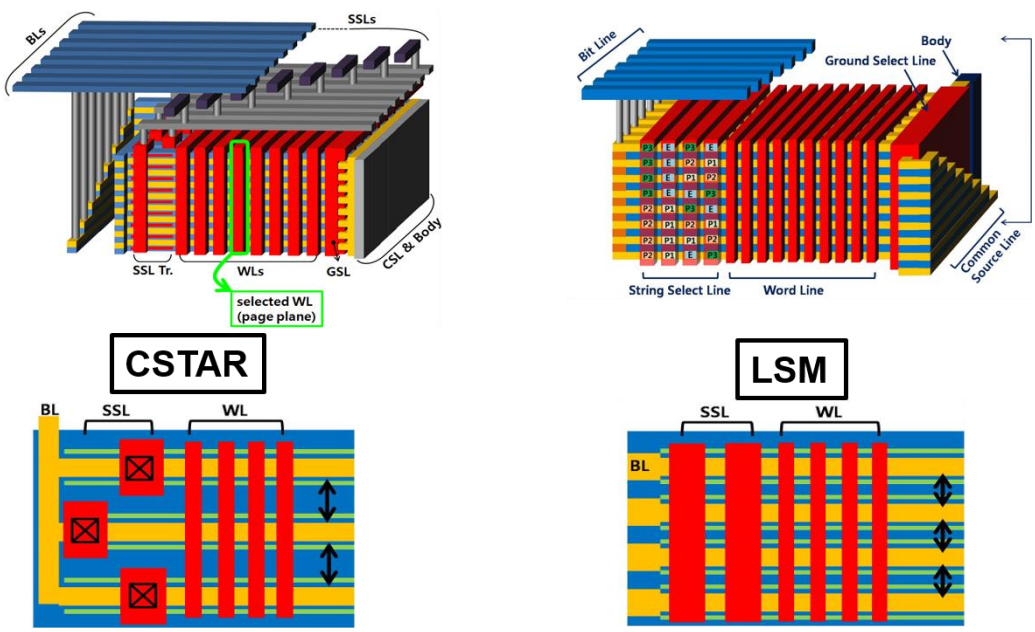


Fig. 2.10. Bird's eye view and top view (layout) of conventional CSTAR array and CSTAR with LSM.

Chapter 3

Methods for Setting String Select Transistors for Layer Selection in Channel Stacked NAND Flash Memory

In this chapter, we propose a simplified channel stacked array with LSM (SLSM) and new SST V_{th} setting methods that make all the SSTs have the targeted V_{th}s. The first method entails all the SSTs on each layer being set to the targeted V_{th}s simultaneously by one erase operation. The second method entails all the SSTs and dummy SSTs (DSSTs) on each layer being set to their targeted V_{th} values by using the incremental step pulse program (ISPP). Furthermore, we verify that basic memory operations are performed independently without interference between stacked layers in SLSM after setting the V_{th} values of SSTs and DSSTs with the new methods.

3.1 Method for Setting SST V_{th} Using One Erase Operation

Fig. 3.1 shows the proposed array architecture with LSM. The structure has stacked channels made of single-crystalline silicon and similar placement of BLs and WLs to that of conventional planar NAND flash memory. However, each vertically aligned string in all stacked layers shares one BL and each layer has its own CSL. The detailed fabrication process can be seen in a previous report [27].

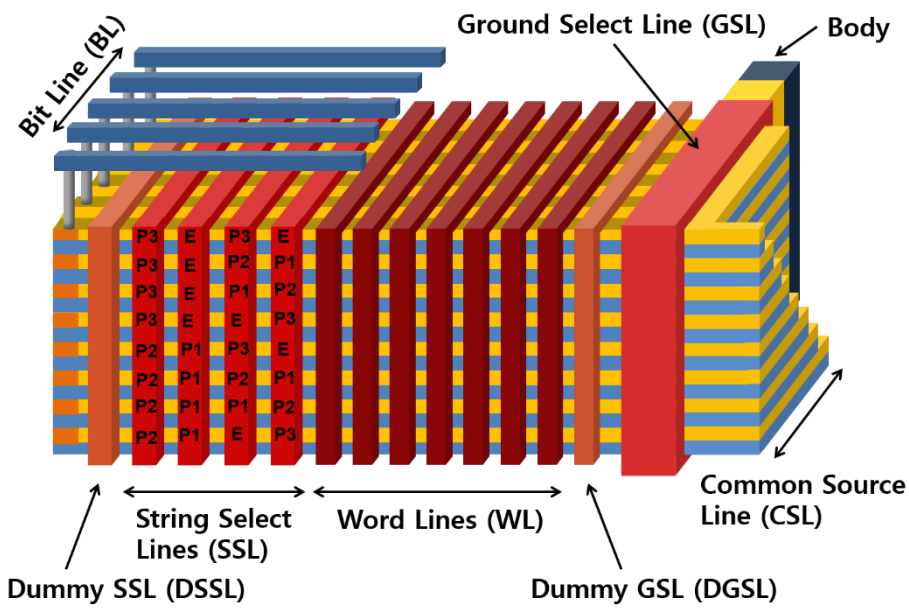


Fig. 3.1. Bird's eye view of the proposed channel stacked NAND array with

LSM.

To verify the new SST V_{th} setting methods and memory operations, measurements of the fabricated structure were made. The measured gate stack type 3D flash memory was fabricated at SK Hynix and has a thin tube-type poly-Si body, gate dielectric stacks including an SiN layer, and a virtual S/D. The cell string is similar to that of the p-BiCS structure [21, 38]. The gate length (L_g), space between vertical WLs, and tube diameter are all <100 nm.

Even though it is a gate stack type, the structure is used for verification since two strings of p-BiCS structure are equivalent to the simplified LSM structure with two stacked channels, if the p-BiCS structure is rotated by 90° as shown in Fig. 3.2. Although the use of a poly-Si channel may result in some disagreement in electrical characteristics such as current and subthreshold swing, this does not have much impact on our findings as the focus of this study is not on the exact values of the currents but more on verification of the new SST V_{th} setting methods and memory operations in the simplified LSM structure. If memory operations can be performed

in a poly-Si channel, more stable operations can be obtained in a single-crystalline silicon channel.

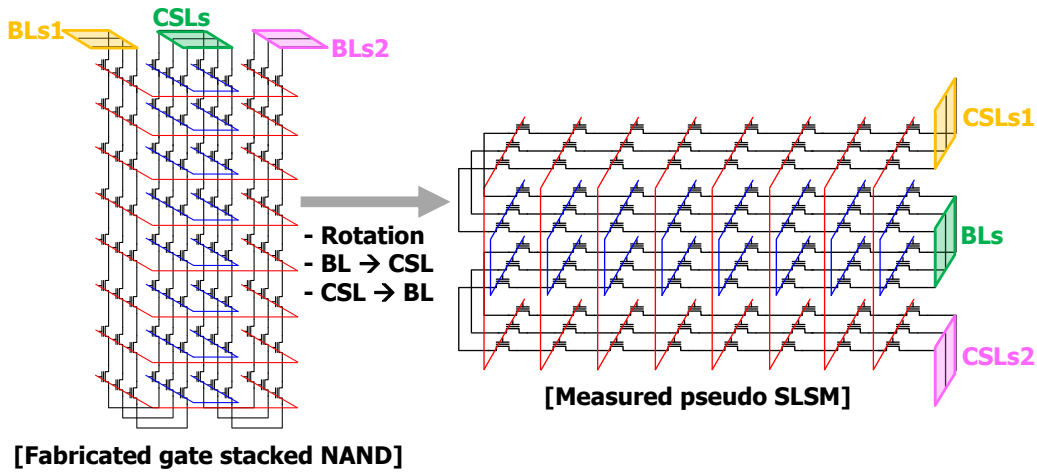


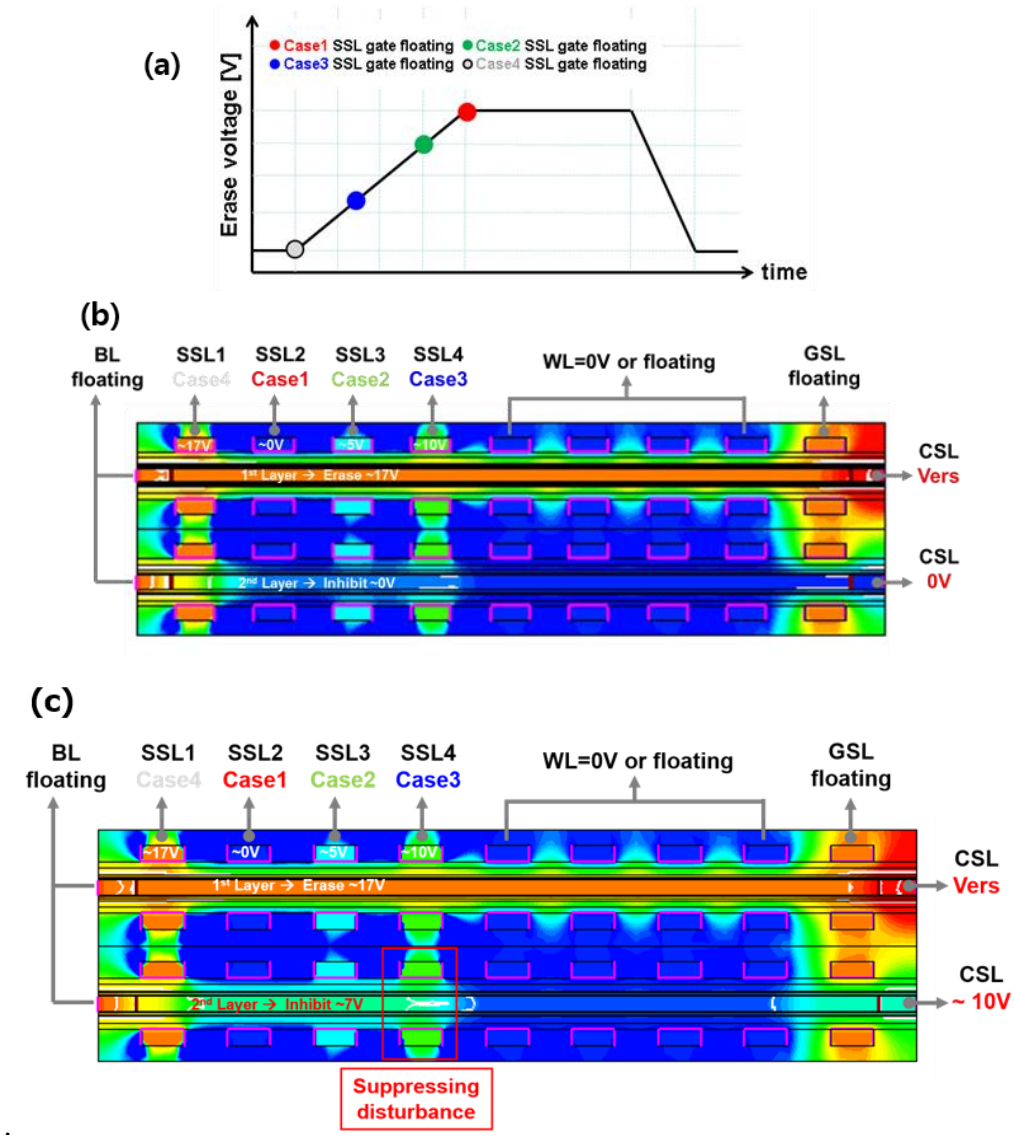
Fig. 3.2. Equivalent circuits of fabricated gate-stacked NAND flash memory and measured pseudo-SLSM.

In the SLSM structure, one or more layers can be erased independently by using GIDL at each CSL. By applying an erase voltage (V_{ers}) to the CSL of the selected layer with other CSLs grounded, only the selected layer can be erased with other layers inhibited. As with other 3D NAND flash devices, all the layers can also be erased simultaneously. Here, it should be noted that the SSTs of the selected layer

can also be erased. After the SSTs in all stacked layers are blind-programmed, the SSTs of the selected layer can be set to targeted V_{th}s simultaneously during the erase operation by adjusting the time for making each SST gate floated state, as shown in the erase pulse of Fig. 3.3(a). The simulated result of Fig. 3.3(b) shows that the SSTs have different gate-to-channel potential (V_{GC}) values according to the gate-floated time since each SST gate potential is coupled to the channel potential rising from the moment each SST gate is floated. The measurements in the simplified LSM structure demonstrate that the SSTs successfully achieve multilevel states by these V_{GC} differences, as can be seen in Fig. 3.3(d).

However, considering the inhibit condition of unselected layers during the erase of the selected layer, one can easily notice that serious erase disturbance can occur in the unselected layers that have already experienced the SST V_{th} setting. Fig. 3.3(c) illustrates that the SSTs of the unselected layers can be programmed by the large V_{GC} because the channels of the unselected layers are grounded by the grounded CSLs when the common gate potential of vertically shared SSTs is coupled to the channel potential of the selected layer. As expected, Fig. 3.4(a) confirms that the

SSTs of the unselected layers are soft-programmed by the large V_{GC} in the simplified LSM structure.



.

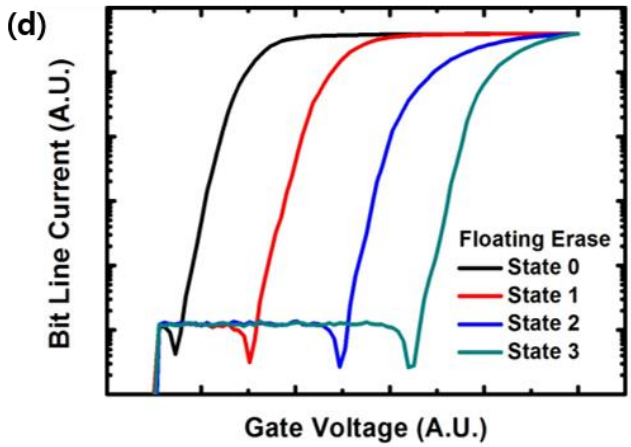


Fig. 3.3. (a) Erase pulse that shows the floated time of each SST gate. (b) Simulation result of the new SST V_{th} setting method by using one erase operation. The figure shows the cross-sectional images of the proposed SSTs, with color coding indicating electrostatic potential contours. The result shows that each SST has a different gate potential according to the gate-floated time and the SST gate potential induces a soft program in the unselected layer. (c) Simulation result when $V_{inhibit}$ is applied to a CSL of an unselected layer. (d) Transfer characteristics of the SSTs that are set to the targeted V_{th}s by the new V_{th} setting method.

To prevent the erase disturbance, $V_{inhibit}$ is applied to the CSLs of the unselected layers. Fig. 3.3(c) indicates that the $V_{inhibit}$ causes GIDL at the CSLs of the

unselected layers and increases the channel potential. However, if the channel potential increases too much, it induces an undesirable soft erase in the SSTs with a lower gate potential than the channel potential. Thus, V_{inhibit} needs to be optimized to minimize the soft erase and program at the same time. Based on the optimization results, Fig 3.4(a) demonstrates that V_{inhibit} = 10 V maximizes the improvement of the erase disturbance with negligible soft erase and program.

After setting the values of V_{th} for all the SSTs by using the new method with one erase operation, memory operations are performed. During the erase of cells, the same bias condition as in the new method is applied except that all the SSTs are floated from the start of erase pulse rising. It is verified that the erase disturbance of SSTs is negligible, similar to the results of Fig. 3.4(a). The cells in the selected layer can be programmed or read by applying higher SST gate voltages than their own V_{th}s. As shown in the equivalent circuit of Fig. 3.5, all the SSTs in the selected layer are turned on by the higher gate voltages, whereas the SST of the unselected layers is turned off. Therefore, the cells in the selected layer can be programmed or read by transferring the BL voltage to the channel without interference from the

unselected layer. Fig. 3.4(b) demonstrates that only the selected string in the selected layer is programmed and that other strings in the selected and unselected layers are all inhibited. Also, each layer is independently accessible for reading.

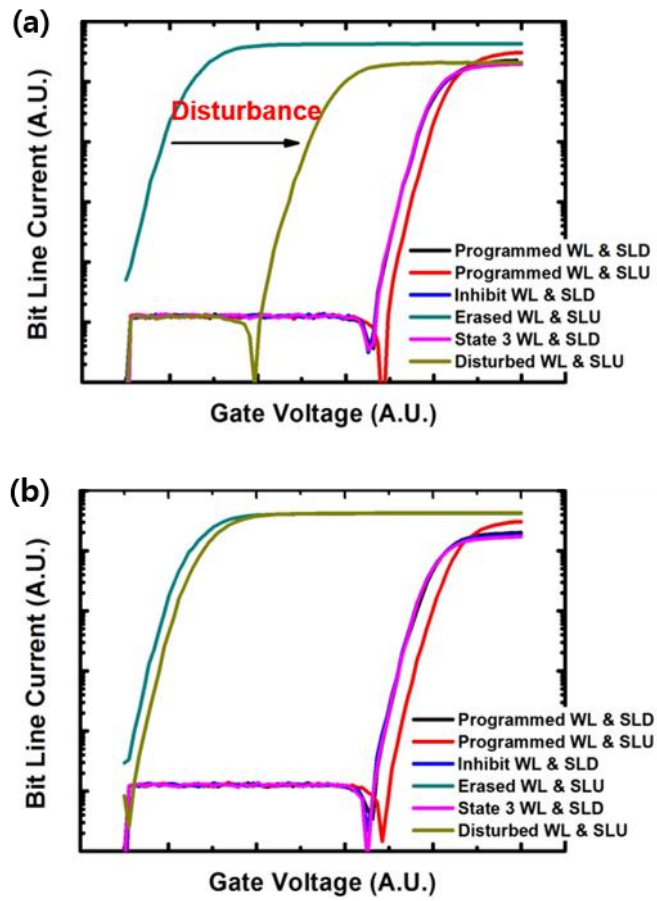


Fig. 3.4. Transfer characteristics of the programmed cell in the selected layer and the inhibited cells in selected and unselected layers. (a) Erase disturbance such as

a soft program and erase in the unselected layers. (b) By applying V_{inhibit} to the CSLs of unselected layers, the erase disturbance is significantly reduced.

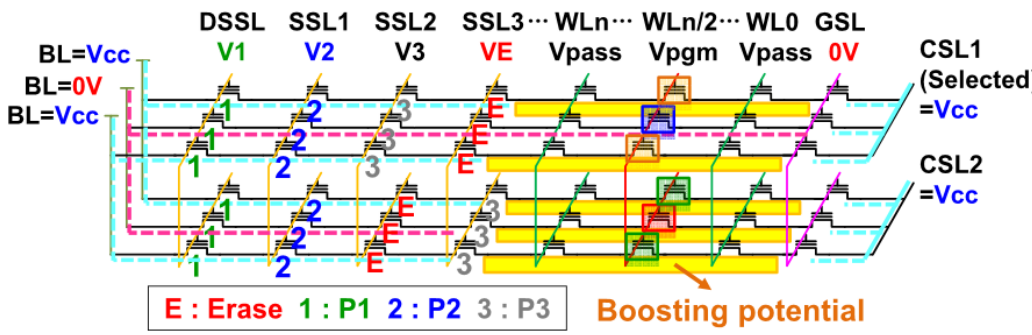


Fig. 3.5. Equivalent circuit that explains how to execute the program of the selected layer independently.

3.2 Method for Setting SST V_{th} Using Dummy SSTs

Fig. 3.6 shows the flowchart of the SST V_{th} setting method by using the ISPP in the SLSM. After erasing the selected layer, targeted SSTs (SST1s) are blind programmed and verified. Then, 0 V (V_{cc}) is applied to the BLs of programmed SST1s (not-programmed SST1s), as shown in Fig. 3.7(b). Here, it should be noted that only the selected layer can be verified by applying the same voltage as the BL voltage (0.7 V) to the CSLs of the unselected layers, as shown in Figs. 3.7(a) and (c).

To screen the programmed SST_1s, WL_ns (WL0s) are programmed by using the BL voltages, which are determined by the verify operation. The bias condition for supplying a program or inhibit voltage to the channels of the WL0s is shown in Fig. 3.8(a). With the DSSTs turned ON by V_{cc} and the ground select transistors (GSTs) turned OFF by 0 V, the BL voltage for the WL0s to be programmed is set to 0 V, while the BL voltage for the WL0 to be inhibited is set to V_{cc}. The BL voltage of 0 V ties the channel of the associated string to ground. When a program voltage (V_{pgm})

is applied to the gates of the WL0s, the WL0s with the grounded channel can be programmed by the large potential difference between the gate and the channel. In the strings with inhibited WL0s, the SSTs shut off and the channels become a floating node as the channel potentials are coupled to $V_{cc} - V_{th}$ (of the SST) by the BL voltage of V_{cc}. When the gate voltages (V_Gs) of WLs rise for the program of the WL0s, the channel potential is boosted automatically by V_G-induced capacitive coupling and the WL0s with the floated channel are inhibited by the boosted channel potential [30]. Fig. 3.8(c) shows that the WL0s in the strings with the programmed SST1 are programmed selectively and that V_{WL0} can be determined to distinguish the programmed WL0s from the inhibited WL0s.

Finally, SST1s are programmed under the bias condition of Fig. 3.8(b). In contrast to the program for the WL0s, 0 V is applied to the CSL of the selected layer and V_{CC} is applied to the CSLs of the unselected layers with the GSTs turned ON by V_{CC} and the SSTs turned OFF by 0 V. By applying V_{WL0} to the gates of the WL0s, only the not-programmed SST1s can be programmed in the selected layer, since the WL0s in the strings with the not-programmed SST1s are turned ON by V_{WL0} and

the CSL voltage of 0 V can be transferred to the associated channels, whereas the WL0s in the strings with the programmed SST1s are turned OFF and the potentials of the associated channels are boosted up. Furthermore, in the unselected layers, all the strings are inhibited by applying V_{CC} to the CSLs regardless of the states of the WL0s. This is because the programmed WL0s are turned OFF by V_{WL0}; otherwise, the channels are floated as the channel potentials are coupled to $V_{cc} - V_{th}$ (of the GST) by the CSL voltage of V_{CC}. Consequently, Fig. 3.8(d) shows that all the SST1s are inhibited except the not-programmed SST1s of the selected layer by using V_{WL0} and by applying V_{CC} to the CSLs of the unselected layers.

After the first program of the not-programmed SST1s is finished in the selected layer, the sequence from verify to SST1 program is repeated with incremental V_{pgm} until all the SST1s are verified to be programmed based on the verify level. By using this ISPP method after the erasure of each layer, as shown in Fig. 3.6 (flowchart), all the SSTs in all layers can have their targeted V_{th} values and an extremely narrow V_{th} distribution by adjusting ΔV_{pgm} of the ISPP [30]. Despite the successful V_{th} setting of all the SSTs by using the ISPP, LSM with DSSTs cannot

have the maximum number of selectable layers, because DSSTs cannot contribute to layer selection. [DSSTs have the function only of separating strings to be inhibited from strings to be programmed during the WL0s program, as shown in Fig. 3.8(a).] Fig. 3.9(a) (table) shows that the maximum number of selectable layers is $S = [n/2 \times (k - 1)]$ under a certain number of SSLs (n), including a dummy SSL and their number of states (k) [28]. In the equation, the square bracket means Gauss's notation or the floor function.

To obtain the DSSTs that can contribute to layer selection with multilevel states, after the DSSTs are blind programmed, the DSSTs of the selected layer are set to targeted V_{th} during the erase operation of Fig. 3.3(a) by controlling the erase voltage (V_{ers}), as shown in the erase pulses of Fig. 3.9(b). Fig. 3.9(c) shows that the SSTs and DSSTs can successfully achieve their targeted V_{th} values by using the ISPP and the erase operation with various V_{ers} values, respectively. In contrast to the SST V_{th} setting, the value of the DSST V_{th} should be set by the erase operation rather than the ISPP, including the repetition of program and verify sequence with incremental V_{pgm}, since BLs are directly connected to DSSTs and inhibit operations cannot be

performed as the consequence, as shown in Fig. 3.10. Thus, the DSST V_{th} distribution might be wider than that of the SSTs formed by using the ISPP though all the intrinsic electrical characteristics of DSSTs are similar to those of SSTs and cells, because DSSTs have the same structure (gate stacks, gate length, space between gates, and so on) as SSTs and cells.

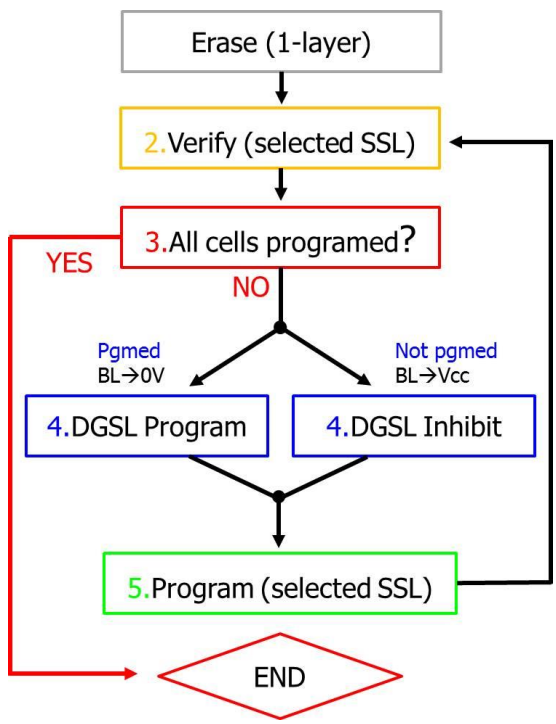


Fig. 3.6. Flowchart of SST V_{th} setting method by using the ISPP.

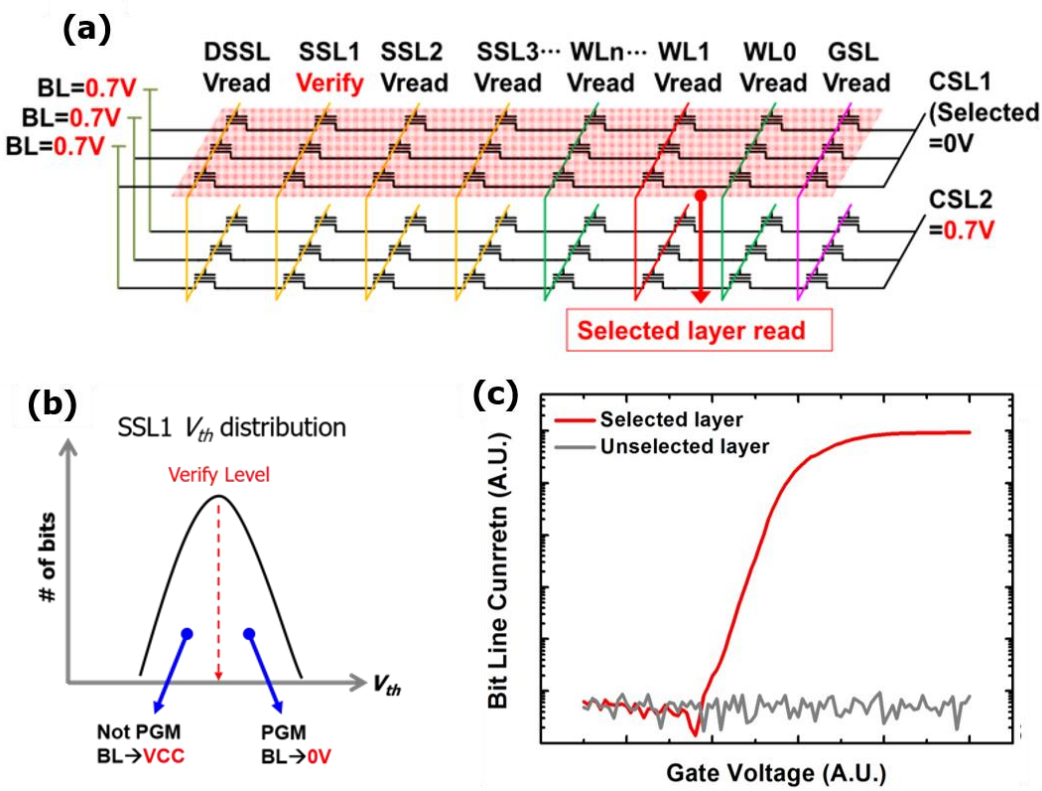


Fig. 3.7. (a) Equivalent SLSM circuit with two layers, explaining the method by which only the selected layer can be verified. (b) SST1 V_{th} distribution. Based on the verify level, SST1s are divided into two groups. (c) Simulated transfer characteristics, which prove that only the selected layer is accessible.

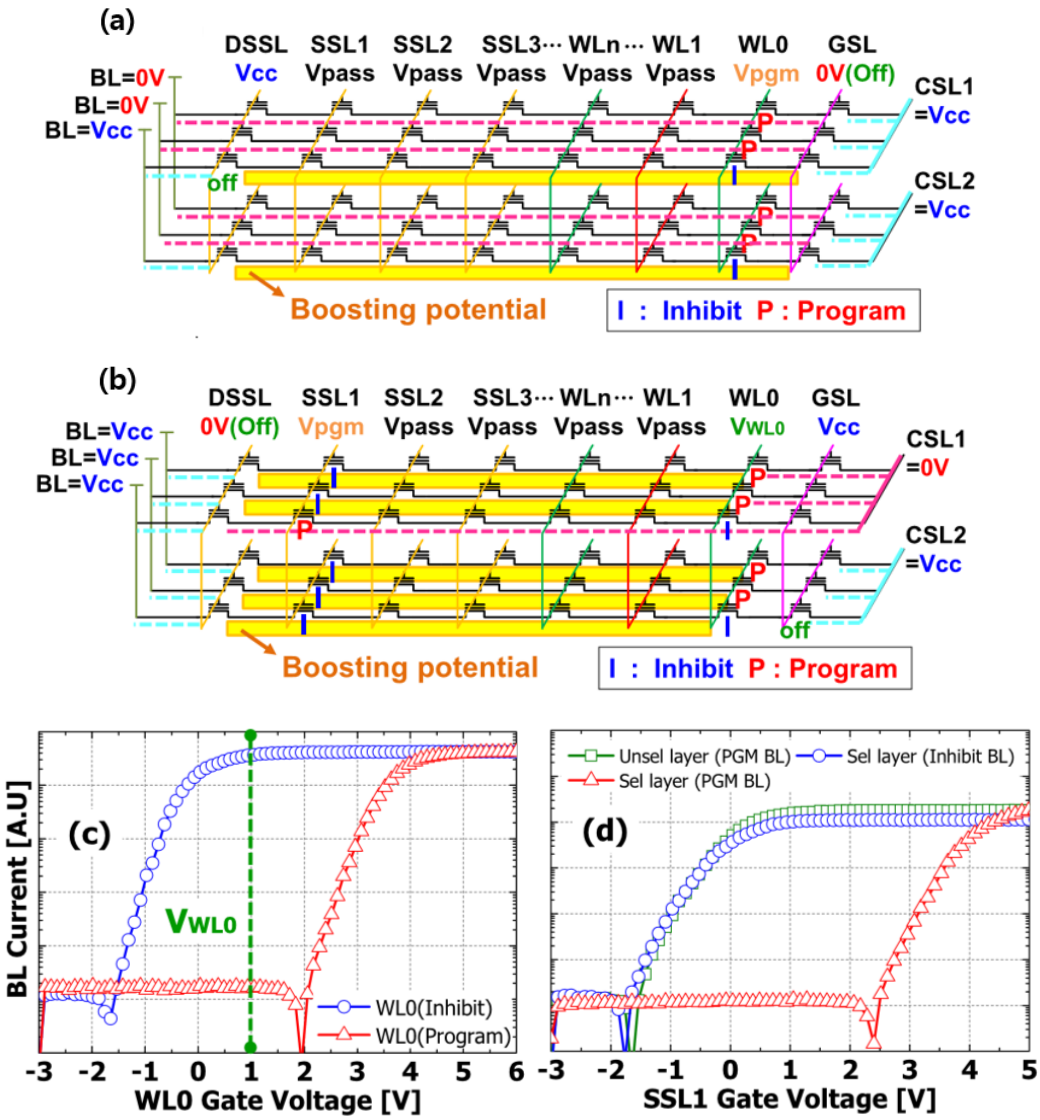


Fig. 3.8. (a) WL0 cells program for screening programmed SSTs. Only the WL0 cells in the strings with programmed SSTs are programmed. (b) Program of targeted not-programmed SST1s by applying V_{WL0} to the gates of WL0s. (c)

Measured transfer characteristics of WL0 cells after programming the WL0 cells. V_{WL0} is determined to distinguish programmed WL0 cells from inhibited WL0 cells. (d) Measure transfer characteristics of SST1s after programming not-programmed SST1s of the selected layer.

(a)	Maximum number of selectable layers			
# of SSLs	2 states	3 states	4 states	5 states
2	2	3	4	5
3	3	7	12	19
4	6	19	44	85
5	10	51	155	381
6	20	141	580	1751

Number of selectable layers = [n/2 x (k-1)]
n : Number of SSL k : Number of state

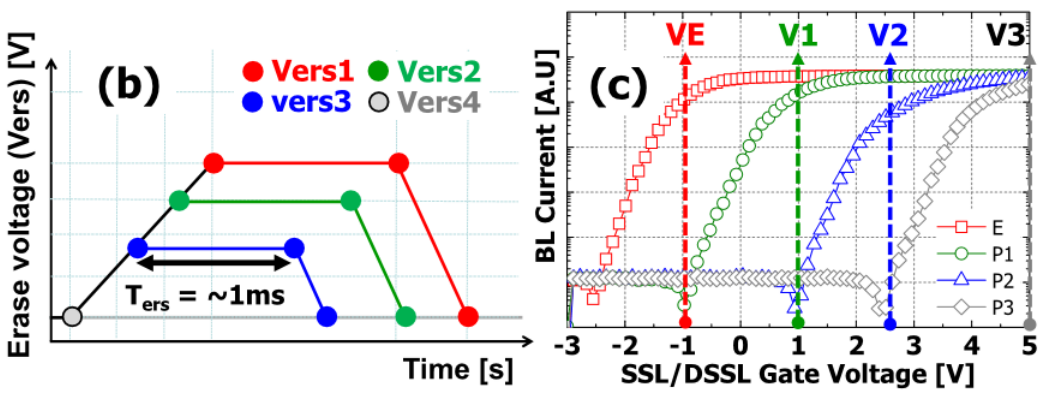


Fig. 3.9. (a) Table showing the maximum number of selectable layers according to the number of SSLs and the number of states in an SST. (b) Erase pulses with

various erase voltages that are applied to the CSL of the selected layer. (c) Measured transfer characteristics of SSTs and DSSTs, which are set to the targeted V_{th} values by the ISPP/one erase with various erase voltages, respectively.

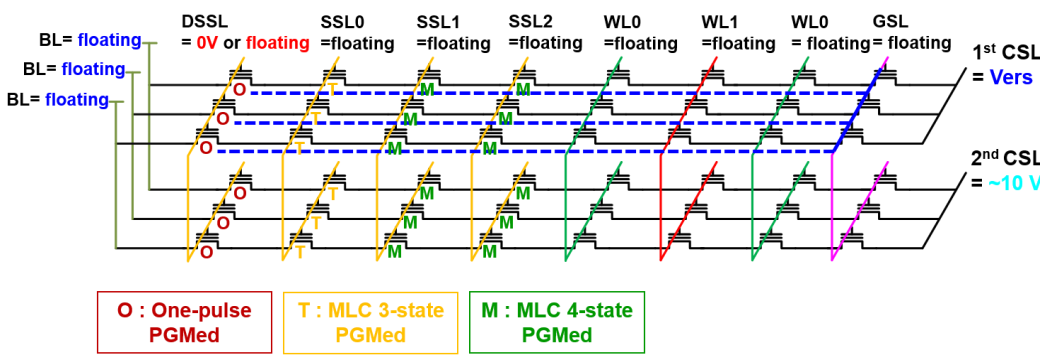


Fig. 3.10. Proposed V_{th} setting method according to each SSL position for stable operation.

Chapter 4

Reliability Issues During LSM in Channel Stacked NAND Flash Memory

In this chapter, we analyze mechanisms of program and read disturbance that can occur during memory operation in channel stacked NAND flash memory with LSM through 2D technology computer-aided design (TCAD) simulations and measurements in fabricated pseudo-SLSM, and we propose new program and read methods to suppress the program and read disturbance. Moreover, the effects of the proposed methods on reliability are investigated rigorously.

4.1 Program Disturbance in SLSM

In conventional NAND arrays, program inhibition is performed by applying V_{cc} to both unselected BLs and gates of SSTs [30]. With this self-boosting scheme, the channel potential becomes sufficiently high to inhibit carriers from tunneling

through the oxide layer. However, in SLSM, there are additional unselected cells that should be inhibited in different ways. Fig. 3.5 shows the inhibition cases under the program operation of SLSM according to BL and SSL gate biases. The unselected cells ($cell_{unsel}$) of BL(V_{cc})_SSLs (ON) are placed in the same plane with the selected cell ($cell_{sel}$), but, in different strings, all the SSTs and DSSTs are turned ON by the higher gate voltages (V_G) than their own V_{th}s, and V_{cc} is applied to their BLs. Thus, the channel potential is boosted similarly to what is done with conventional NAND arrays. The $cell_{unsel}$s of BL(V_{cc})_SSLs(OFF) are located in the unselected layers and connected to different BLs from that of the $cell_{sel}$; one or more SSTs or DSSTs are turned OFF, which causes boosting of the channel potential regardless of BL voltages. The $cell_{unsel}$s of BL(0 V)_SSLs(OFF) are positioned in the unselected layers and connected to the same BL as that of the $cell_{sel}$; they share a BL bias of 0 V. However, only $cell_{sel}$ can be programmed by transferring 0 V to the channel through all the SSTs and DSSTs, whereas the $cell_{unsel}$s are inhibited, because one or more SSTs or DSSTs are turned OFF and 0 V is blocked as the consequence. After setting V_{th}s of all the SSTs and DSSTs by using the SST and

DSST *V*th setting method as shown in Fig. 4.1(a), program operations are performed. Fig. 4.1(b) shows that the unselected cells have different program disturbances according to the inhibition cases. Interestingly, BL(0 V)_SSLs(OFF) $cell_{unsel}$s have the worst disturbance characteristic, while BL(V_{cc})_SSLs(ON) $cell_{unsel}$s have the best disturbance immunity [39].

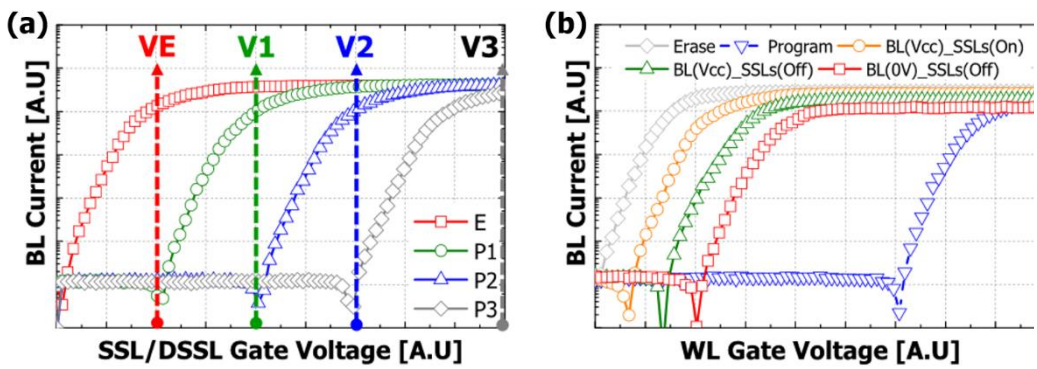
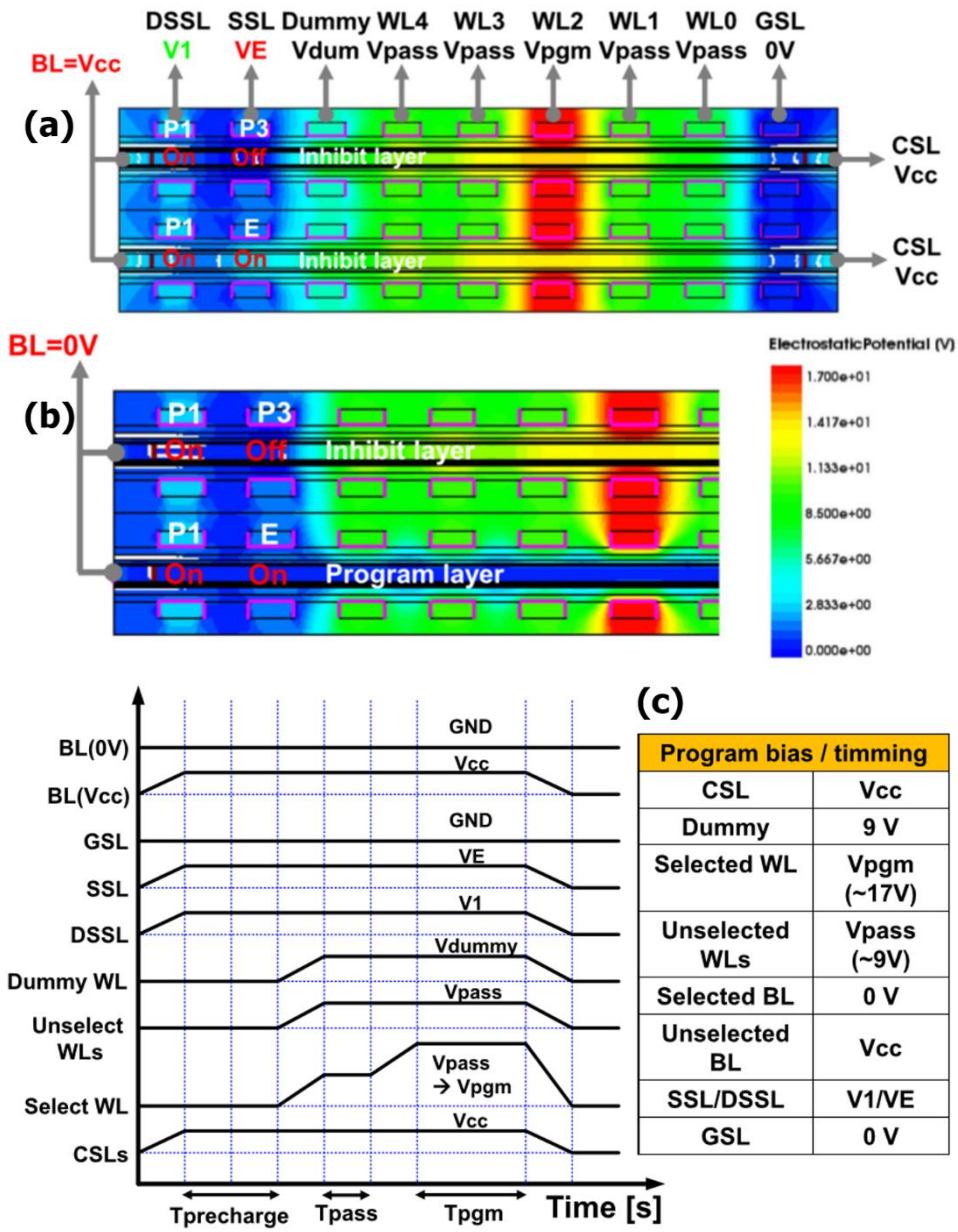


Fig. 4.1. (a) Measured transfer characteristics of SSTs and DSSTs, which are set by the SST and DSST method. (b) Measured transfer characteristics of the programmed cell in a selected layer and the inhibited cells in selected and unselected layers.

To clarify the physical origin of the different program disturbances, TCAD

simulations are performed. Figs. 4.2(a) and (b) show the simulated cross-sectional structures of SLSM, which are composed of six WLs, one SSL, and one dummy string select line (DSSL) for simplicity. Program operations are simulated under the bias and timing conditions of Fig. 4.2(a)–(c) after block erase. The biases of Fig. 4.2(a) are applied to the strings with BL(V_{cc})_SSLs(OFF) $cell_{unsel}$s or BL(V_{cc})_SSLs(ON) $cell_{unsel}$s. Identical biases are used on the strings with BL(0 V)_SSLs(OFF) $cell_{unsel}$s except for the BL bias, as shown in Fig. 4.2(b). Fig. 4.2(a) shows the boosted channel potentials of the $cell_{unsel}$s, which are extracted at the interface between the channel and the tunneling oxide of the programmed WL throughout program operation. BL(V_{cc})_SSLs(ON) $cell_{unsel}$s and BL(0 V)_SSLs(OFF) $cell_{unsel}$s have the highest and lowest boosted channel potentials, respectively, which is consistent with the measured program disturbance.



Program bias / timming	
CSL	Vcc
Dummy	9 V
Selected WL	Vpgm (~17V)
Unselected WLs	Vpass (~9V)
Selected BL	0 V
Unselected BL	Vcc
SSL/DSSL	V1/VE
GSL	0 V

Fig. 4.2. (a) Simulated SLSM cross-sectional structure that shows the bias condition of strings with BL(V_{cc})_SSLs(ON) or BL(V_{cc})_SSLs(OFF) unselected

cells. (b) Bias condition of strings with BL(0 V)_SSLs(ON) or BL(0 V)_SSLs(OFF) unselected cells. The same biases are applied as in (a) except for the BL bias; the color code indicates electrostatic potential contours. (c) Bias and timing conditions used during program operations.

In Fig. 4.3(a) (inset), it should be noted that BL(V_{cc})_SSLs(ON) $cell_{unsel}$s with the highest channel potential are initially precharged only. Furthermore, the strings connected to 0 V of the BL undergo a significant potential drop after applying program voltage (V_{pgm}). A new program scheme is proposed to precharge all the unselected strings (PAUSs). As shown in the timing diagram of Fig. 4.3(b), all BLs are biased to V_{cc}, and $V3$ is applied to SSLs and DSSLs during the precharge time ($T_{prechage}$). Then, they are restored to their own biases before applying the pass voltage (V_{pass}) to satisfy each condition.

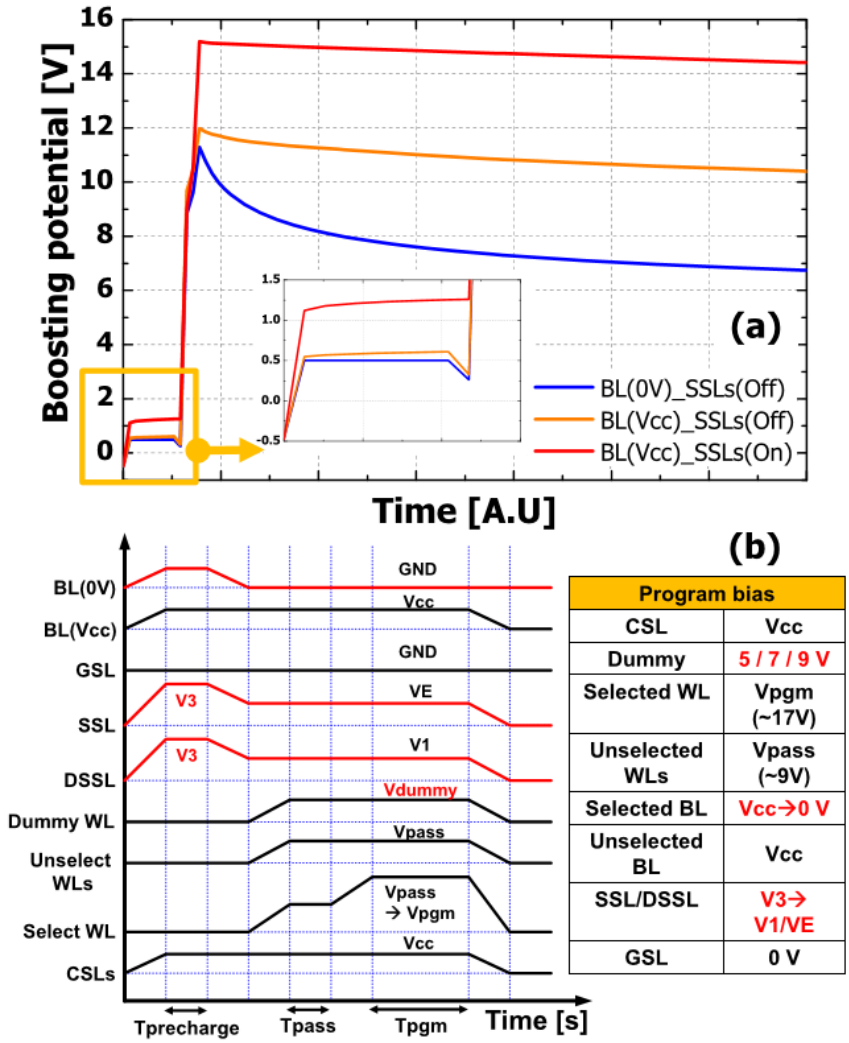


Fig. 4.3. (a) Transient channel potential change of unselected cells with different inhibition cases during program operation. (Inset) Channel potential variation by the precharged channel. (b) Program bias and timing condition designed to suppress program disturbance of SLSM.

Fig. 4.4(a) shows that the channel potential is initially boosted by the precharge, such that it is easier to increase it when V_{pass} and V_{pgm} are applied consecutively. However, the precharge-enhanced potential is saturated toward the channel potential of the default program as program time (T_{pgm}) gets increased in BL(0 V)_SSLs(OFF) $cell_{unsel}$s. This implies that the leakage current flowing from the channel to the BL limits the enhancement of the channel potential, because the increased channel potential is sustained in BL(V_{cc})_SSLs(OFF) $cell_{unsel}$s. Thus, various biases (V_{dummy}s) are applied to the gates of the dummy WL, since the leakage current should be decreased if the potential difference between the channel and the BL is reduced by the V_{dummy} change. Expectedly, Figs. 4.4(b) and (c) show that the leakage current gets reduced and the potential drop during T_{pgm} becomes alleviated with lower values of V_{dummy}. From these simulated results, it is clearly found that the precharging effect and the leakage current from the channel to the BL are the main factors for determining the boosted channel potential in SLSM. In the fabricated structures, the effects of the PAUS with various V_{dummy}s are confirmed.

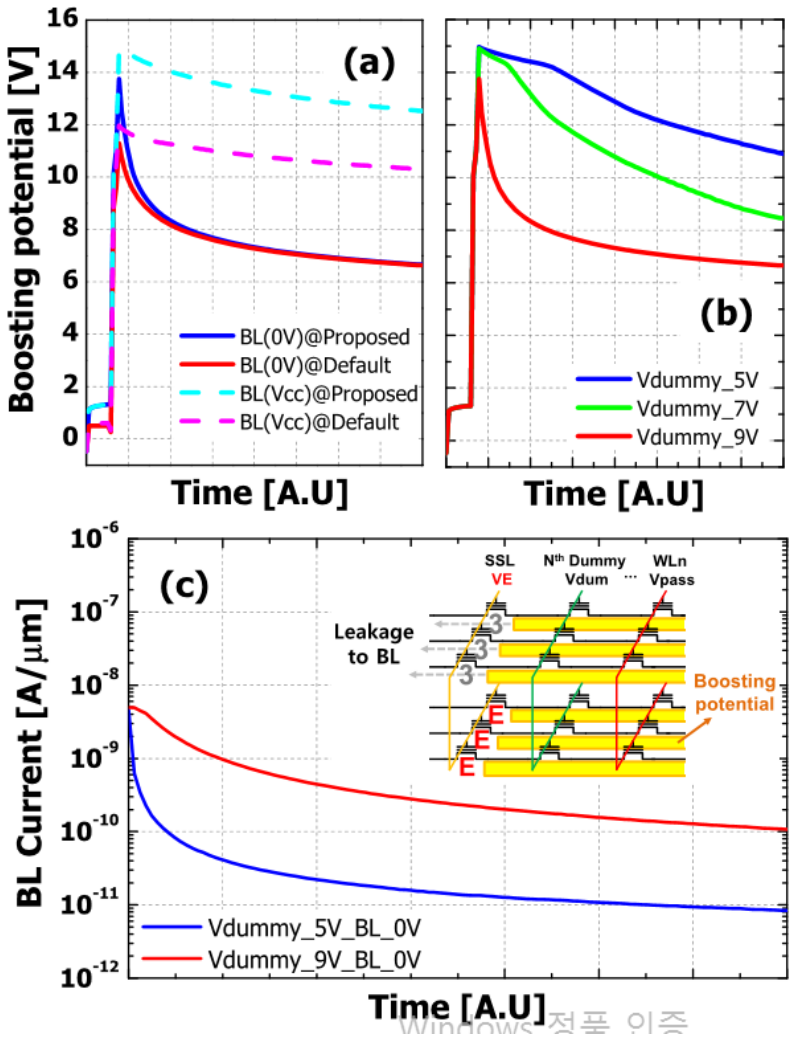


Fig. 4.4. (a) Boosted channel potential of BL(0 V)_SSLs(OFF) unselected cells. Proposed and default program methods are used, respectively. V_{dummy} = 9 V is applied in both program methods. (b) Boosted channel potentials of BL(0 V)_SSLs(OFF) unselected cells. The proposed program method (PAUS) is used

with various V_{dummy}s. (c) Leakage currents flowing from the channel to the BL. As a function of time, the currents are extracted under the condition of PAUS with various V_{dummy}s after V_{pgm} is applied.

First, the variation of the transfer curve is measured in BL(V_{cc})_SSLs(OFF) $cell_{unsel}$s. Fig. 4.5(a) shows that the program disturbance is improved by the PAUS owing to the precharged channel. In addition, a low V_{dummy} value causes additional improvement, which is the result of reduced leakage current flowing from the channel to the BL. Subsequently, the same measurements are repeated in BL(0 V)_SSLs OFF) $cell_{unsel}$s. In contrast to the BL(V_{cc})_SSLs OFF) $cell_{unsel}$s, the initial precharging has negligible effect on the program disturbance, as shown in Fig. 4.5(b). As mentioned in the simulation results, this can be explained by the leakage current limiting the boosting of the channel potential. Fig. 4.5(b) shows that the program disturbance can be improved only by the PAUS along with a low V_{dummy} value. Consequently, the program disturbance of all cases of the $cell_{unsel}$s is significantly improved by the initially precharged channel and the reduced leakage

current, as shown in Fig. 4.5(c).

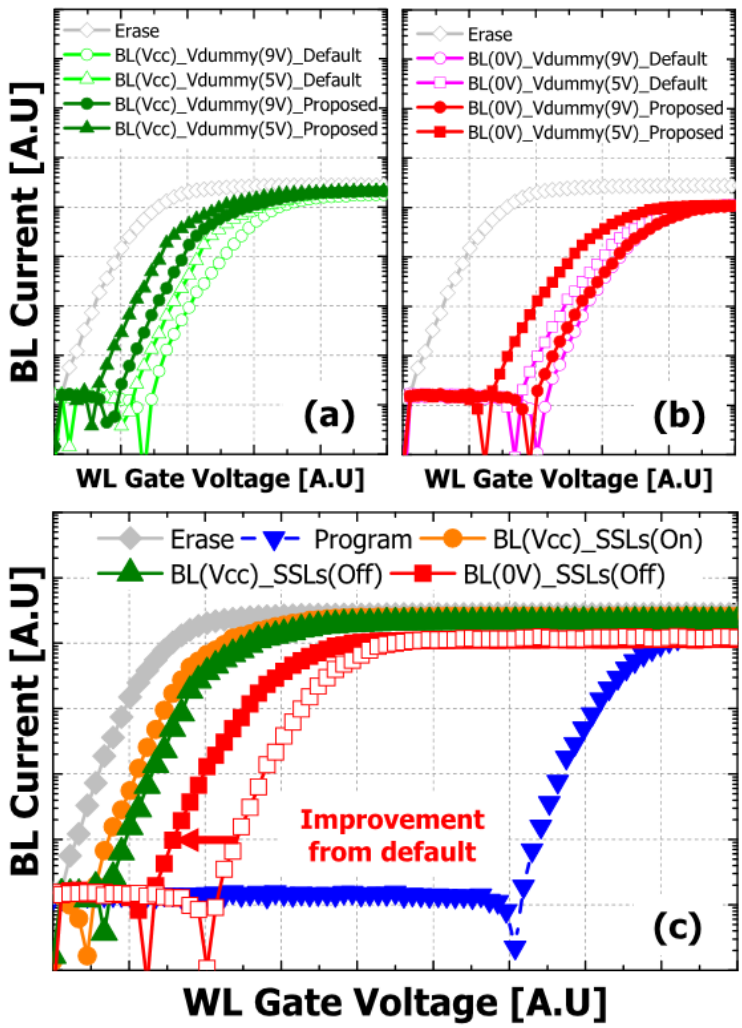
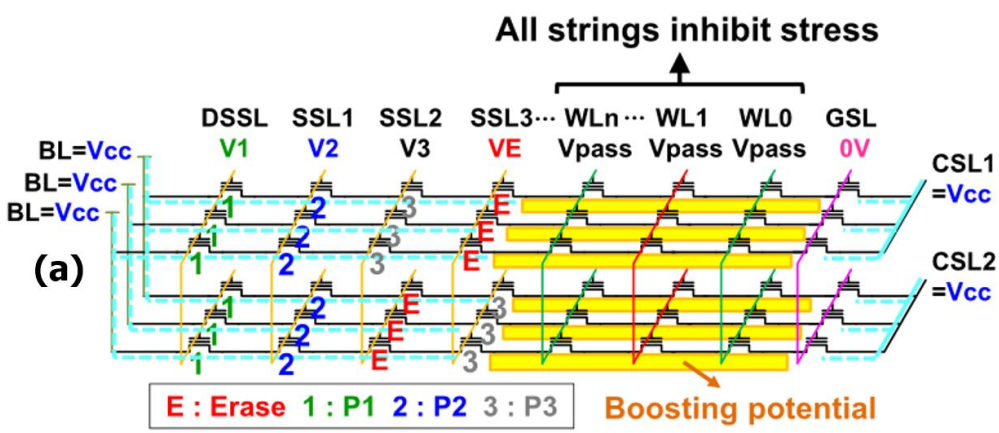


Fig. 4.5. (a) Transfer curve variation of BL(V_{cc})_SSLs(OFF) $cell_{unsel}$ by using the proposed program method (PAUS) with various V_{dummy}s. (b) Transfer curve

variation of BL(0 V)_SSLs(OFF) $cell_{unsel}$ by using the PAUS with various V_{dummy}s. (c) Measured transfer characteristics of a programmed cell in a selected layer and inhibited cells in selected and unselected layers. PAUS and V_{dummy} = 5 V are used.

All the of SST and DSST reliabilities are similar to those of cells except for the inhibit stress derived from unique LSMs. In Fig. 4.6(a), the inhibit stress is applied to all strings simultaneously by boosting the potential of all channels. Considering the condition of the inhibit stress, one can simply notice that electron–hole pairs (EHPs) can be generated by GIDL, and impact ionization (II) between an SSL and a dummy WL and the SSTs and DSSTs and a dummy WL can be degraded by the generated EHPs [40–43]. However, fortunately, a low V_{dummy} value is used to improve channel boosting, and this low V_{dummy} value can also decrease band-to-band tunneling (BTBT) and II by reducing the potential difference between the SSL gate and the channel, as shown in the simulated results of Figs. 4.6(e) and (f). In Figs. 4.6(c) and (d), the SST and DSST V_{th} and dummy WL changes are measured

as a degradation parameter by the inhibit stress of Fig. 4.6(a). The SSTs and DSSTs and a dummy WL are first set to their targeted V_{th}s. Then, the SST and DSST V_{th} and dummy WL changes are monitored by increasing the number of the inhibit stress. Here, all the V_{th}s are extracted by using the constant-current method. Figs. 4.6(c) and (d) show that the inhibit stress has a negligible effect on the SST and DSST V_{th} and dummy WL changes owing to GIDL and II suppression [Fig. 4.6(b)] by controlling the gate voltages of the dummy WL.



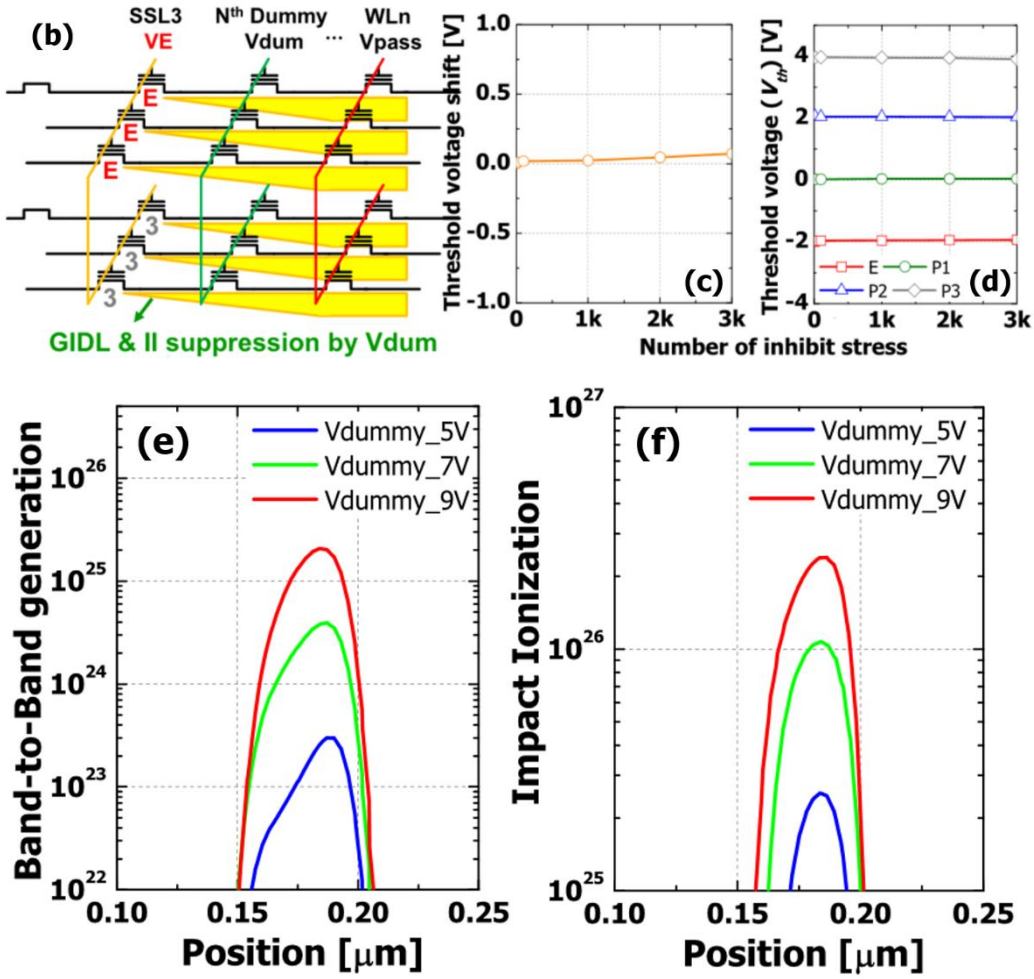


Fig. 4.6. (a) Equivalent LSM circuit that explains how to apply inhibit stress to all the strings. (b) GIDL and II suppression by controlling the gate voltages of dummy WLs. (c) Threshold voltage shift of a dummy WL as a function of inhibit stress. (d) Threshold voltage changes of SSTs and DSSTs with multilevel states as a function of inhibit stress. (e) BTBT generated between an SSL and a

dummy WL in inhibit strings. (f) II generated between an SSL and a dummy WL in inhibit strings. Various V_{dummy}s are applied to change the BTBT generation and the II.

4.2 Read Disturbance in SLSM

To verify the mechanisms of the read disturbance in the SLSM structure, measurements of the fabricated structure are made. The bias and timing conditions of Figs. 4.7(a) and 4.8 are used in the read operations of SLSM. As can be seen in Figs. 4.7(a) and (b), the V_{th}s of the SSTs and the selected cells are set (where, unless otherwise mentioned, the selected cells are in the P3 state) and all the unselected cells are in the erase state before the read operations. Then, the V_{th} changes of case1, case2, case3, and case4 cells ($\Delta V_{th,case1}$, $\Delta V_{th,case2}$, $\Delta V_{th,case3}$, and $\Delta V_{th,case4}$) are measured as a function of the number of read cycles at 90°C.

Fig. 4.9(a) shows that V_{th} changes as a result of the read disturbance by the conventional read operation where a constant selected WL voltage (V_{sel}) is applied throughout the read operation. All the V_{th}s are extracted in the drain current–gate voltage (I_D–V_G) curves by using the constant-current method at $I_D = 1 \times 10^{-8}$ A. It should be noted that each case has different read disturbance values (namely, $\Delta V_{th,case1} > \Delta V_{th,case2} > \Delta V_{th,case3} > \Delta V_{th,case4}$). Given that GSTs (SSTs) turn on and

0 V (0.5 V) is transferred from the CSL (BL) to the channel of the case2 cell (case3 cell), it can be expected that the V_{th}s of the case2 and case3 cells are changed by typical read disturbance where a small amount of charge may be injected into the storage node with repetitive F-N tunneling stress in all unselected cells, although a read bias (V_R) applied to unselected WLs is much smaller compared to a program voltage. Furthermore, the worst disturbance of the case1 cell is caused by GIDL-induced hot carrier injection (GHCI) in Fig. 4.7(c), which is the result of channel boosting in the unselected layers by a combination of turned-off SSTs and selected cells and the rising of the voltage (V_{unsel}) applied to unselected WLs [40, 44, 45]. Also, the best disturbance immunity of the case4 cell results from the reduced F-N tunneling stress by V_{unsel}-induced channel boosting [39].

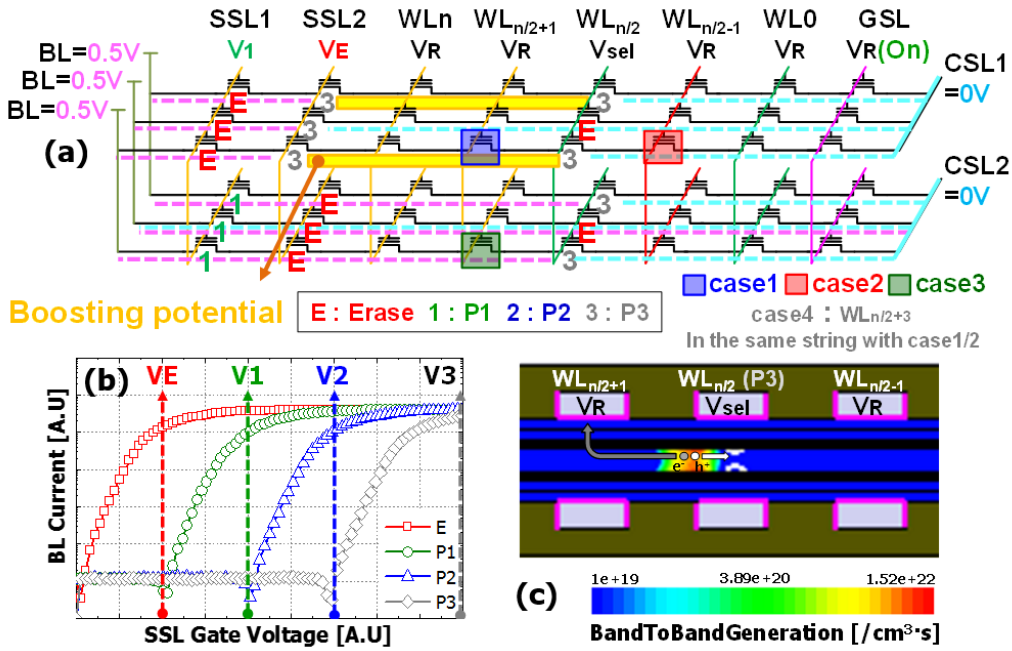


Fig. 4.7. (a) Equivalent circuit that explains how to read the selected layer only without interference from unselected layers and the occurrence of channel potential boosting in unselected layers. (b) Measured transfer characteristics of SSTs and cell_{sel}, which are set to their targeted V_{th}s. (c) Simulation result that indicates that GHCI can occur by the positively boosted channel potential.

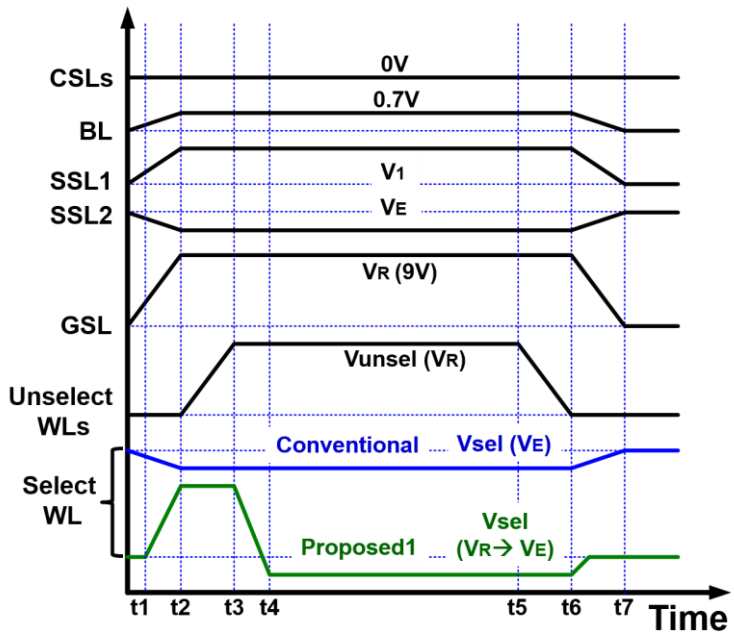


Fig. 4.8. Bias and timing conditions used during read operations (conventional and proposed1).

To suppress the GHCI-induced read disturbance of the case1 cell, V_{sel} is maintained at V_R during V_{unsel} rising and then V_{sel} is restored to its own bias, as shown in the proposed1 read operation of Fig. 4.8. Fig. 4.9(b) indicates that the read disturbance of the case1 cell is significantly decreased and thus the disturbance is similar to that of the case3 cell since GHCI cannot occur under the proposed1 operation. However, interestingly, the disturbance of the case2 cell is aggravated by

the proposed1 condition. This phenomenon can be explained by the conduction energy band diagrams of Fig. 4.10(b). Similar to the channel boosting by the rise in V_{unsel} with the turned-off SSTs and the selected cells (positive boosting), the channel potential can be boosted negatively by V_{unsel} falling (negative boosting), as depicted in Fig. 4.10(a). Fig. 4.10(b) shows that the BL-side conduction energy band (BCEB) is raised by the negative boosting and thus electrons can be injected into the CSL-side conduction energy band (CCEB). The injected electrons are accelerated by the CSL-side energy slope of Fig. 4.10(b) and can obtain enough energy to be injected into the storage node of the case2 cell. Consequently, the read disturbance of the case2 cell is degraded owing to the hot carrier injection (HCI) induced by the negative boosting (HCI-NB). Additionally, Fig. 4.9(b) indicates that the case2 cell disturbance characteristics get improved as the selected cell has a lower V_{th} because the probability of HCI-NB drops because of the reduced CSL-side energy barrier with the lower V_{th} of the selected cell. The step-by-step status of the phenomenon described above can be seen in Fig. 4.11 and can be summarized as follows:

V_{unsel} = 6 V: The pass voltage is applied to the unselected WLs during the verify

operation. At this time, all main cell transistors are turned on and their channel potential is derived from the CSL at 0 V and the BL at 0.7 V.

V_{unsel} = 3 V: When the verify operation is complete, V_{PASS} is applied to the unselected WL, which drops to 0 V (WL recovery). When the WL voltage drops from 6 to 3 V (the threshold voltages of $WL_{n/2}$ and SSL2, respectively), the $WL_{n/2}$ and SSL2 transistors are turned off and the WL_n to $WL_{n/2}$ channels change to their floating states.

V_{unsel} = 0 V: The WL_n to $WL_{n/2}$ channel is not connected to the BL. Consequently, the channel potential is changed along with the WL bias. Therefore, the channel potential falls to −3 V at V_{PASS} = 0 V because of WL capacitive coupling.

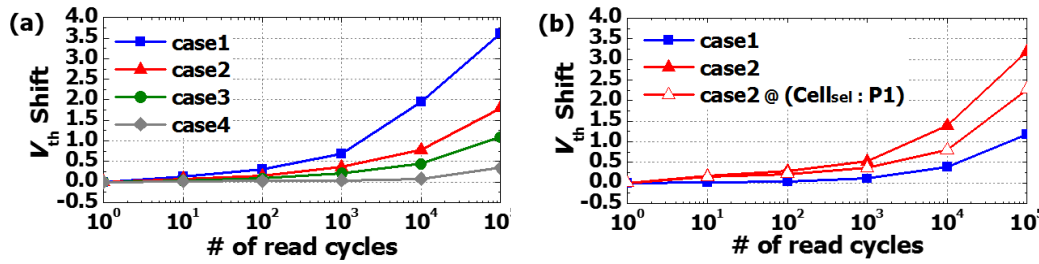
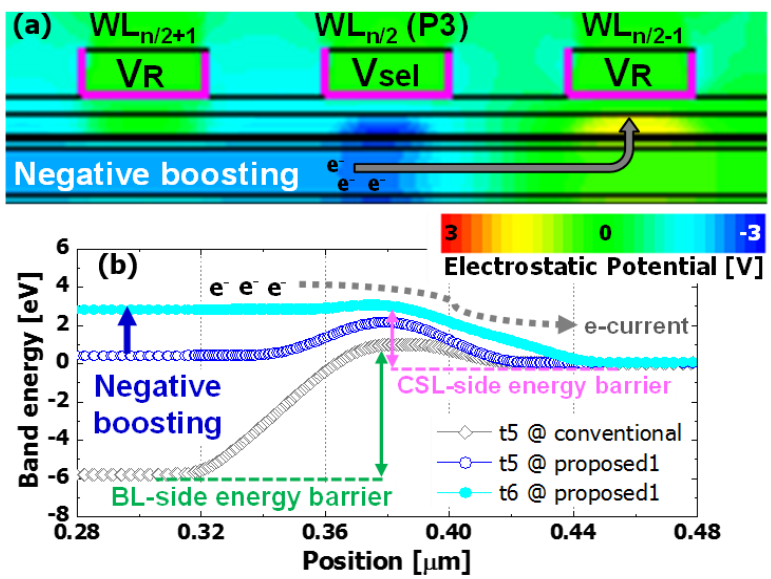


Fig. 4.9. V_{th} shifts as a function of the number of read cycles, which are measured as a result of the read disturbance in (a) conventional and (b) proposed1 read operations.



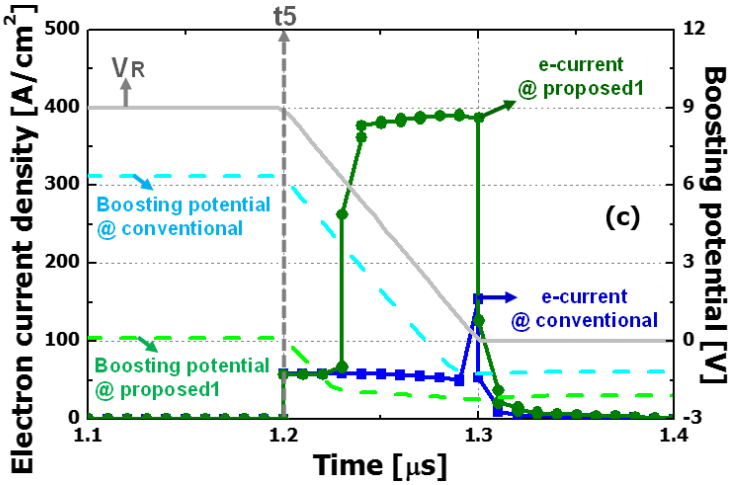


Fig. 4.10. (a) Simulation result and (b) conduction energy band diagrams that explain that HCI can occur as a result of injected electrons from CB_{BL} to CB_{CSL} caused by negative boosting. (c) Channel potentials and electron currents flowing from CB_{BL} to CB_{CSL} monitored throughout conventional and proposed1 read operations.

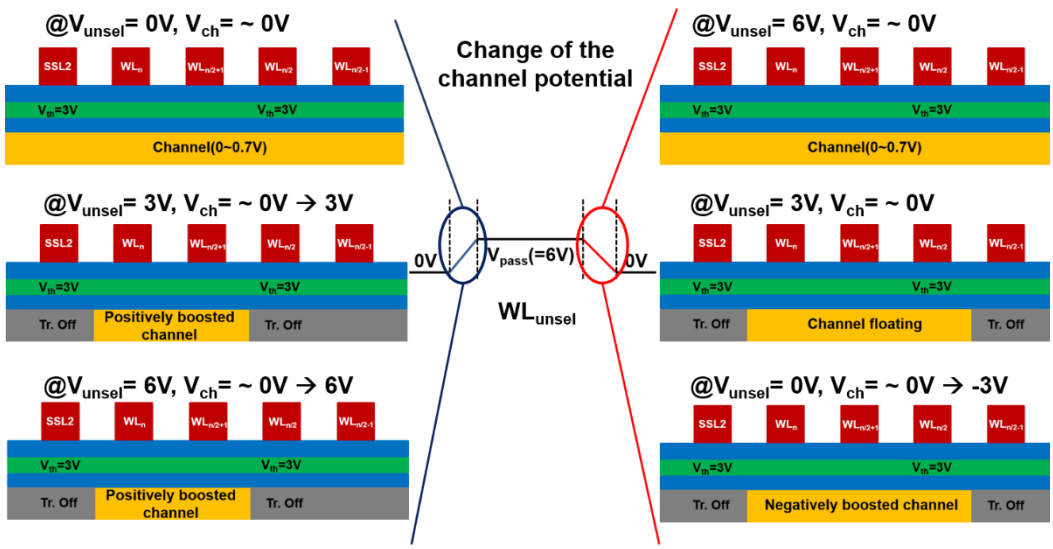


Fig. 4.11. Change of the channel potential during the verify operation.

To clarify the reason why the case2 cell disturbance is seriously generated in the proposed1 read operation, channel potentials and electron currents (e-currents) flowing from the BCEB to the CCEB are monitored throughout the conventional and proposed1 read operations. (The e-currents represent the injected electrons, and the channel potentials are extracted at the interface between the channel and the tunneling oxide.) Fig. 4.10(c) shows that a large e-current is detected in the proposed1 read operation not long after the start of the V_{unsel} falling transition, while the conventional read operation exhibits a small e-current only at the end of the

V_{unsel} falling transition. This implies that HCI-NB occurs much more seriously in the proposed1 read operation since HCI-NB is strongly related to the number of injected electrons.

Moreover, the e-current difference between the conventional and the proposed1 read operations can be explained by the channel potential before the V_{unsel} falling transition (t5). Fig. 4.10(c) shows that the conventional scheme has a high channel potential (~6 V) at t5 by positive boosting, whereas a channel potential of ~0 V is formed in the proposed1 read operation owing to suppression of positive boosting. Therefore, electron injection is significantly reduced in the conventional scheme despite the BCEB rising by negative boosting because the high channel potential at t5 causes a large BL-side energy barrier, as shown in Fig. 4.10(c).

A novel read operation (proposed2) is proposed to suppress both GHCI and HCI-NB. In the proposed2 read operation, V_{sel} is maintained at V_R during V_{unsel} rising and falling and then V_{sel} is restored to its own bias, as shown in Fig. 4.12(a). Fig. 4.12(b) shows that the read disturbances by GHCI and HCI-NB are successfully removed in case1 and case2 cells.

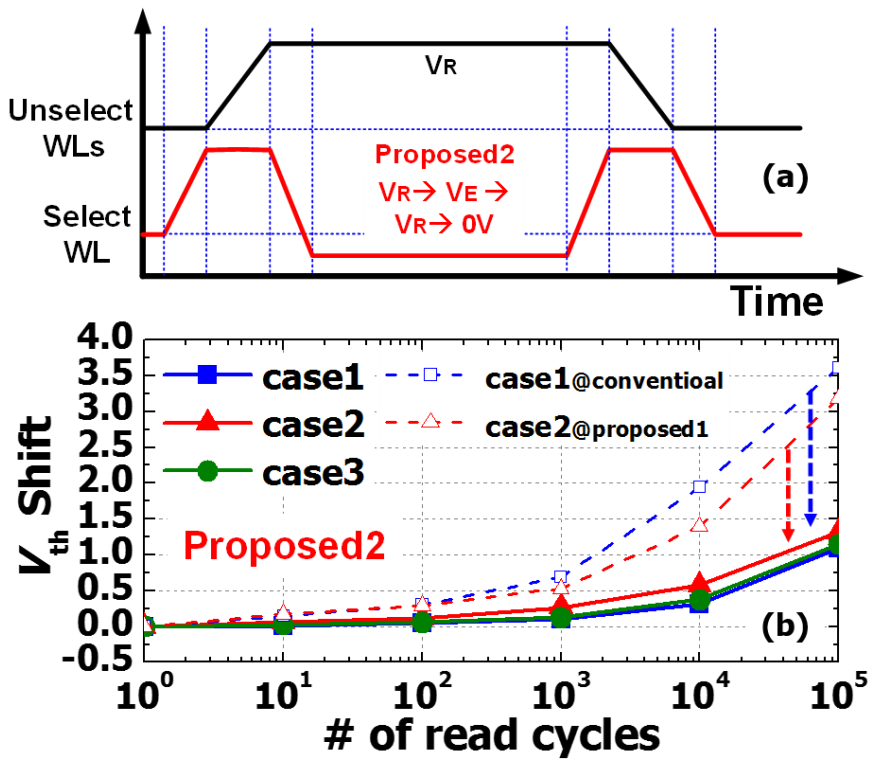


Fig. 4.12. (a) Bias and timing conditions used in the proposed2 read operation to suppress HCI_{GIDL} and HCI_{NB}. (b) V_{th} shift as a function of the number of read cycles measured in the proposed2 read operation.

Chapter 5

Application to General NAND Flash Memory

Channel stack type NAND flash memories have very good characteristics in terms of their small unit cell size, bit line pitch scaling, use of a single-crystalline Si channel, and alleviation of the current reduction problem caused by increasing in stacking number. However, compared to the gate stack type, the channel stack type has critical obstacles in the way to commercialization such as complex array architecture and decoding of the stacked layers. To overcome these problems, our group proposed LSM. LSM does not physically divide the SSLs for layer selection, but it uses the electrical properties of the SSTs to select layers. In other words, LSM is not only available in the channel stack but is an operation method applicable to all types of NAND flash memory. In addition, it has the advantage of lowering the degree of difficulty in the process since SSLs are not separated and used in the same form as WLs. Another important feature of LSM is that efficiency is greatly increased depending on the number of V_{th} states that SST can have [28]. Currently,

eight states are used in NAND flash memory. Therefore, in this experiment, we set the SST V_{th} to four states (MLC) for stable LSM. However, techniques for keeping more data in one cell and keeping it stable are continuously being proposed and developed [46]. Therefore, the possibilities and efficiency of LSM continue to increase with the development of NAND flash memory.

Given the characteristics of LSM and the current state-of-the-art development, it is quite possible to apply it to the gate stack type. Three-dimensional NAND flash is divided into two types, but the basic structure of both is the same. As shown in Fig. 5.1, BL, SSL, WL, and CSL are arranged so that they cross each other and be used to access selected cells. At this time, the structural difference between the conventional gate stack type and the SLSM is only the separation of the SSL and CSL. Therefore, if the CSL is separated in the gate stack type, forming an SSL like a WL, then LSM can be applied.

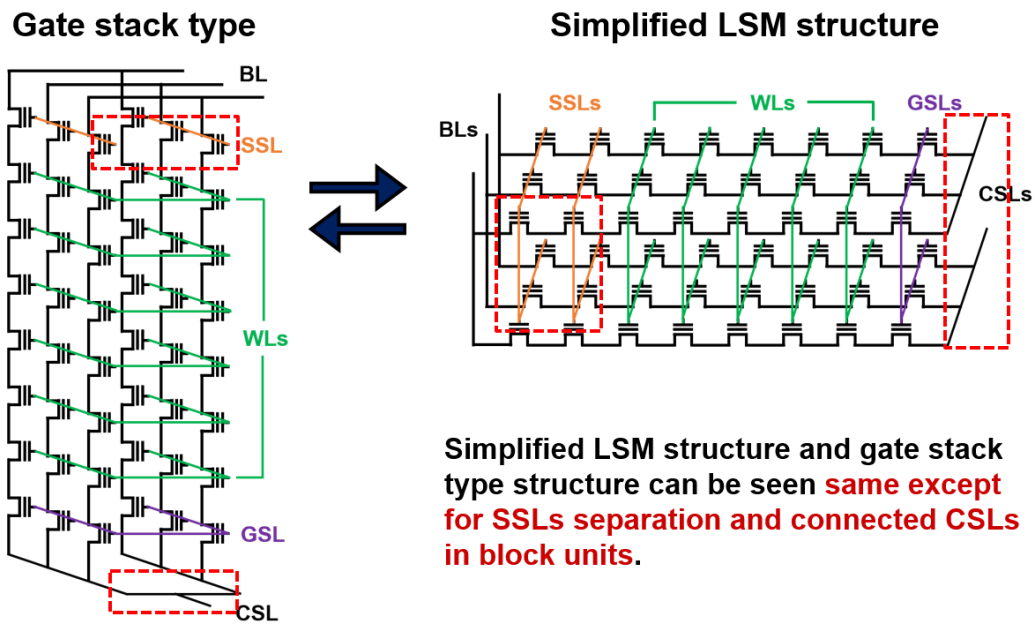


Fig. 5.1. Equivalent circuits according to stack types. SLSM structure and gate stack type structure can be seen to be the same except for SSL separation and connected CSLs in block units.

To verify the proposed LSM and mechanisms of the disturbances in the SLSM structure, measurements of the fabricated structure were made. The measured gate stack type 3D flash memory has gate dielectric stacks including an SiN layer, a virtual S/D, and a thin tube-type poly-Si body, although the proposed structure can have a single-crystalline silicon body. For the gate stack type, poly-Si is used as the

channel material for the transistor owing to issues of fabrication and structure. Thinner poly-Si channels are especially preferred for the gate stack type to improve controllability and suppress short channel and grain boundary effects. In particular, grain size and trap sites in grain boundaries have a great influence on electrical properties, and studies are actively proceeding [47–49]. To apply LSM to a general NAND flash memory, it is important to accurately understand the electrical characteristics of the poly-Si channel.

Fig. 5.2(a) shows that the grain size depends on poly-Si thickness, as reported in thick poly-Si films, if the conditions of deposition and annealing are the same [50]. To distinguish between grain size and thickness effects in thin poly-Si channels, four samples were prepared, as indicated in Table 5.1 [47]. Figs. 5.2(b) and (c) show poly-Si thickness dependencies on the ON current and effective mobility, respectively. These graphs indicate that the ON current and effective mobility of poly-Si decrease with decreasing grain size of the poly-Si channel (samples 1, 2, and 4). However, these values are independent of poly-Si thickness down to ~88 Å (samples 1 and 3) if the grain sizes are the same. These results indicate that charge

transport in the inversion layer is disturbed by the trap sites in grain boundary of the poly-Si channel.

The subthreshold swing property of thin poly-Si channels is shown in Fig. 5.2(d). Owing to the thin body that suppresses the short channel effect, the thinner poly-Si channel is expected to have a lower swing value. Although the swing decreases with decreasing poly-Si thickness regardless of grain size, the swing of a small grain size sample is larger than that of a large grain size sample. The reliability of the poly-Si channel is measured by applying a bipolar pulse stress to the gate. I_d–V_g characteristics of samples 1 and 4 are shown in the inset of Fig. 5.2(e). Swing degradation is the primary factor affecting poly-Si reliability under cycle stress. V_{th} shifts after a stress of 10,000 cycles are shown in Fig. 5.2(e). The poly-Si channel with a large grain size has a large resistance to cycle stress regardless of poly-Si thickness. These measurement results indicate that the electrical properties are dominantly affected by grain size and not by poly-Si thickness. Increasing grain size is the key for large ON current and better reliability of 3D NAND flash memory using thin poly-Si channels.

Table 5.1. Poly-Si channel samples. Various thickness poly-Si channel were made using slope etch of sample 2,3 and the grain size were controlled by changing the as-deposited a-Si thickness.

Sample Number	Poly-Si thickness (Å)	Grain size
1	185	Large
2	88 ~ 138	Medium (variable)
3	87 ~ 130	Large
4	77	Small

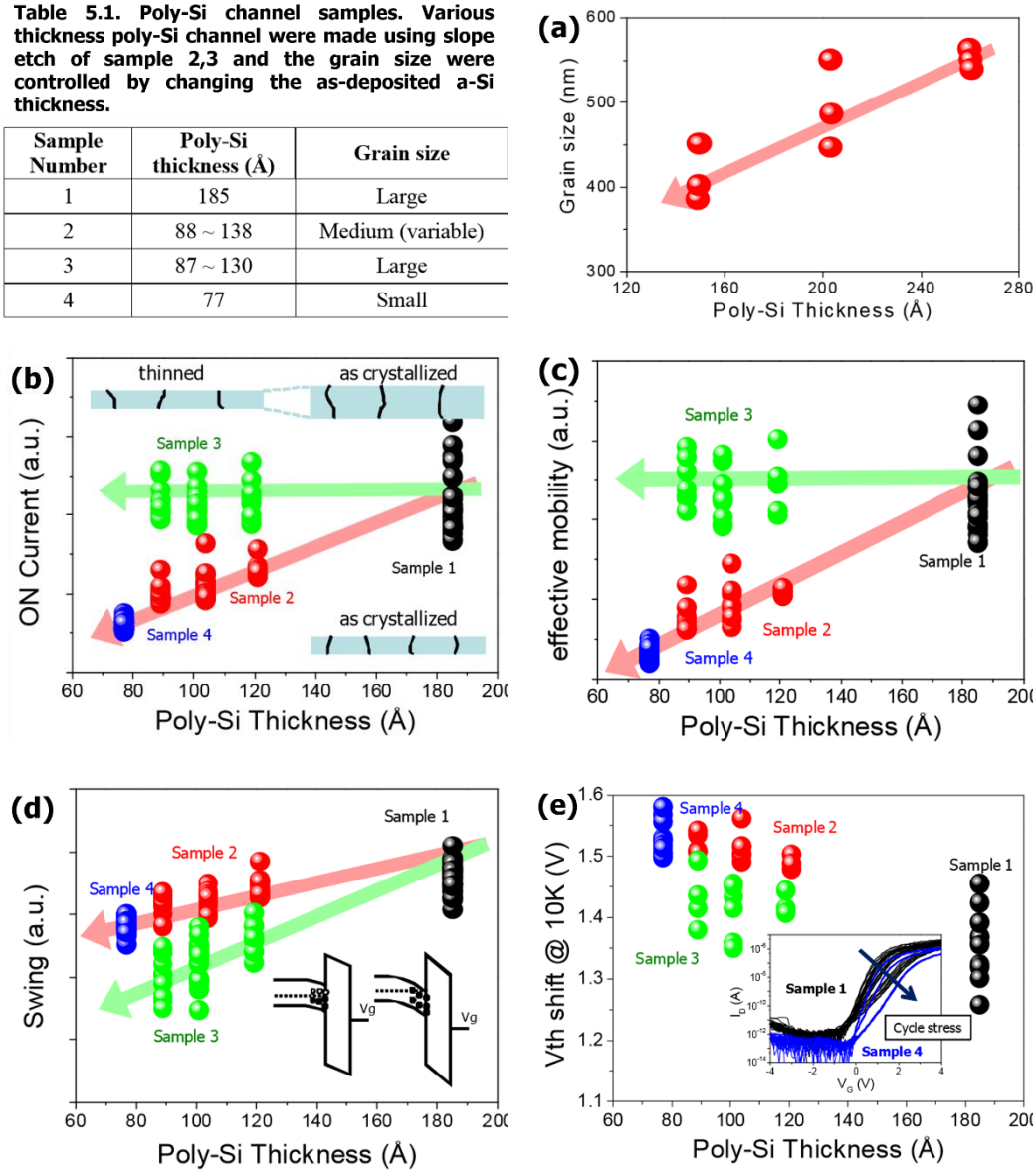


Fig. 5.2. (a) Poly-Si thickness dependence on grain size. Grain size is evaluated from EBSD and the poly-Si thickness is as-deposited thickness. (b) Poly-Si thickness dependence on current. Sample 3 shows large current. (c) Poly-Si

thickness dependence on effective mobility. (d) Subthreshold swing as a function of poly-Si thickness. (e) V_{th} shift after a stress of 10,000 cycles as a function of poly-Si thickness. (Inset) I_d–V_g characteristics of samples 1 and 4 at initial state and after stresses of 1000 and 10,000 cycles. [47]

In addition, two intrinsic fluctuations (random trap fluctuation and random telegraph noise) induced by the poly-Si channel in VNAND devices have been modeled and quantified for structure optimization and statistical prediction in gate stack type [48].

These reported results show the characteristics of the poly-Si channel used in the current gate stack type NAND flash memory. Although poly-Si is inferior in electrical properties compared to single-crystal Si, it already offers good ON current, reliability, and V_{th} distribution [47]. In addition, the V_{th} distribution can be predicted through analysis and modeling, and electrical properties can be further improved by application of software such as early prediction of error correction [48].

To verify the effect on of LSM on the electrical properties of a transistor, we

measured a device having a poly-Si channel that is different from the proposed SLSM structure. The present gate stack type NAND flash memory has a sufficiently narrow V_{th} distribution and a small subthreshold swing, so that stable LSM can be expected through the experimental results. It is also expected that use of single-crystal Si as a channel material, as the proposed structure, will have better performance than that indicated by the above results. Therefore, it is confirmed that the proposed LSM can be applied regardless of the structure of the 3D NAND flash and the material used for the channel and can operate stably.

Chapter 6

Conclusions

New SSTs V_{th} setting methods are proposed in SLSM. In the first method, all the SSTs on each layer are set to the targeted V_{th} values simultaneously using a single erase operation. In the second method, all the SSTs/DSSTs on each layer are set to their targeted V_{th} values by the incremental step pulse program (ISPP). By using the new methods, SSTs having the proper V_{th} distribution can be obtained, and the number of selectable layers can be increased with a fixed number of SSLs (if the number of layers is determined, the number of SSLs can be reduced). Furthermore, in the fabricated pseudo-LSM structure, basic memory operations such as erase, program, and read are performed and optimized after setting the V_{th} values of all the SSTs by using the new V_{th} setting methods. As a result, stable memory operations are successfully obtained without the undesirable disturbances derived from unique LSM operations and interference between stacked layers.

After setting the SSTs/DSSTs with multilevel states, the program disturbance is

analyzed in SLSM. Three different inhibition cases exist according to the SST gate and BL biases. Out of the three cases, the BL(V_{cc})_SSLs(OFF) and BL(0 V)_SSLs(OFF) cases do not occur in conventional NAND arrays; moreover, they have worse disturbance characteristics than the BL(V_{cc})_SSLs(ON) case. TCAD simulations and measurements were performed to investigate the disturbance mechanism of the additional cases. Initially, non-precharged channel and large leakage current degraded the disturbance. To precharge all the strings and reduce the leakage current, PAUS is proposed along with a low V_{dummy} value. As a result, the program disturbance significantly improved, compared with that of a default program. In addition, reliability was also enhanced by reducing the potential difference between the SSL gate and the channel.

Furthermore, the read disturbance derived from the unique read operation of SLSM was rigorously investigated. Based on the simulation results, it is found that case1 cell disturbance is degraded by the HCI caused by the gate-induced drain leakage, resulting from the positively boosted channel potential by the combination of turned-off SSTs and selected cell and V_{unsel} rising. Furthermore, the case2 cell is

seriously disturbed, as the channel potential is negatively boosted by V_{unsel} falling, and a different type of HCI is generated owing to the electrons injected by the negative boosting.

A novel read operation, in which V_{sel} is maintained at V_R during V_{unsel} rising/falling and then V_{sel} is restored to its own bias, is proposed to suppress these HCIs. Upon application of the proposed read method, the read disturbances by the HCIs are successfully removed in case1 and case2 cells.

Through the above discussion, we verified that the LSM operation in channel-stacked NAND flash memories works stably by adopting the proposed schemes.

Bibliography

[1] 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,

[2] 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.

[3] Y.W. Chang, G.W. Wu, P.C. Chen, C.H. Chen, I.C. Yang, C.Y. Chin, I.J. Huang, W.J. Tsai, T.C. Lu, W.P. Lu, K.C. Chen, and C.Y. Lu, "A new interference phenomenon in sub-60nm nitride-based Flash memory," *Proceedings of NVSMW,* pp. 81-82, 26-30 Aug. 2007.

[4] C.-H. Lee, J.Choi, Y. Park, C. Kang, , "B.-I. Choi, H. Kim, H. Oh, and W.-S. Lee, "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.

[5] E.S. Cho, H.J. Kim, B.T. Kim, J.H. Song, J.H. Song, D.H. Song, J.-H. Choi, K-D. Suh, and C. Chung, "Optimal Cell Design for Enhancing Reliability Characteristics for sub 30 nm NAND Flash Memory," in *IEEE IRPS Tech.* pp.611-614, May. 2010.

[6] J.N. Wileu, T.-Y. Fu, T. Tanaka, S. Takeuchi, S. Aoyama, J. Miyazaki, and Y. Watakabe, "Phase-shift mask pattern accuracy requirements and inspection technology," *Proc. SPIE 1464, Integrated Circuit Metrology, Inspection, and Process Control V,* 346, Jul. 1991.

[7] M. Garza, J.V. Jensen, N.K. Eib, and K.K. Chao., "Optical Proximity Correction Method and Apparatus," U. S. Patent No. 5,723,233, Mar. 3, 1998.

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

[9] C.-M. Lim, S.-M. Kim, Y.-S. Hwang, J-S. Choi, K.-D. Ban, S.-Y. Cho, J.-K. Jung, E.-K. Kang, H.-Y. Lim, H.-S. Kim, and S.-C. Moon, "Positive and negative tone double patterning lithography for 50 nm flash memory." *Proc. SPIE 6154, Optical Microlithography XIX, 615410,* Mar. 2006

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

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

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

[13] Kirk. Prall, "Scaling Non-Volatile Memory Below 30 nm," *NVSMW Tech Dig.,* pp.5-10, Aug. 2007.

[14] S.-M. Jung, J. Jang, W. Cho, H. Cho, J. Jeong, Y. Chang, J. Kim, Y. Rah, Y. Son, J. Park, M.-S. Song, K.-H. Kim, J.-S. Lim, and K. Kim, "Three dimensionally stacked NAND Flash memory technology using stacking single

crystal Si layers on ILD and TANOS structure for beyond 30 nm node," *IEDM tech. digest,* pp. 1-4, 11-13 Dec. 2006.

[15] H. Tanaka, M. Kido, K. Yahashi, M. Oomura, R. Katsumata, M. Kido, Y. Fukuzumi, M. Sato, Y. Nagata, Y. Matsuoka, Y. Iwata, H. Aochi, and A. Nitayama, "Bit Cost Scalable technology with punch and plug process for ultra high density Flash memory," *Symposium on VLSI technology,* pp. 14-15, 2007.

[16] J. Jang, H.-S. Kim, W. Cho, H. Cho, J. Kim, S.I. Shim, Younggoan, J.-H. Jeong, B.-K. Son, D.W. Kim, Kihyun, J.-J. Shim, J.S. Lim, K.-H. Kim, S.Y. Yi, J.-Y. Lim, D. Chung, H.-C. Moon, S. Hwang, J.-W. Lee, Y.-H. Son, U.-I. Chung, and W.-S. Lee, "Vertical cell array using TCAT (Terabit Cell Array Transistors) technology for ultra high density NAND Flash memory," *Symposium on VLSI technology,* pp. 192-193, 2009.

[17] E.-K. Lai, H.-T. Lue, Y.-H. Hsiao, J.-Y. Hsieh, C.-P. Lu, S.-Y. Wang, L.-W. Yang, T. Yang, K.-C. Chen, J. Gong, K.-Y. Hsieh, R. Liu, and C.-Y. Lu, "A Multi-Layer Stackable Thin-Film Transistor (TFT) NAND-Type Flash Memory," *IEEE International Electron Devices Meeting,* pp. 1-4, 11-13 Dec. 2006.

[18] T. Endoh, K. Kinoshita, T. Tanigami, Y. Wada, K. Sato, K. Yamada, T. Yokoyama, N. Takeuchi, K. Tanaka, N. Awaya, K. Sakiyama, and F. Masuoka, "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. 945-961, April 2003.

[19] Masahide Kimura, "3D cells make terabit NAND flash possible," techno.hikkeibp.co.jp, Sept 17, 2009.

[20] Y. Komori, M. Kido, M. Kito, R. Katsumata, Y. Fukuzumi, H. Tanaka, Y. Nagata, M. Ishiduki, H. Aochi, and A. Nitayama, "Disturbless flash memory due to high boost efficiency on BiCS structure and optimal memory film stack for ultra high density storage device," in *IEDM Tech. Dig.*, pp. 1-4, Dec. 15~17, 2008.

[21] R. Katsumata, M. Kito, Y. Fukuzumi, M. Kido, H. Tanaka, Y. Komori, M. Ishiduki, J. Matsunami, T. Fujiwara, Y. Nagata, L. Zhang, Y. Iwata, R. Kirisawa, H. Aochi, and A. Nitayama, "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.

[22] S. Whang, K. Lee, D. Shin, B. Kim, M. Kim, J. Bin, J. Han, S. Kim, B. Lee, Y. Jung, S. Cho, C. Shin, H. Yoo, S. Choi, K. Hong, S. Aritome, S. Park, and S. Hong, "Novel 3-Dimensional Dual Control-Gate with Surrounding Floating-Gate (DC-SF) NAND Flash Cell for 1 Tb File Storage Application," in *IEDM Tech. Dig.*, pp. 29.7~4, Dec. 6~8, 2010.

[23] 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.

[24] H.-T. Lue, T.-H. Hsu, Y.-H. H, S.P. Hong, M.T. Wu, F.H. Hsu, N.Z. Lien, S.-Y. Wang, J.-Y. Hsieh, L.-W. Yang, T. Yang, K.-C. Chen, K.-Y. Hsieh, and C.-Y. Lu, "A highly scalable 8-layer 3D vertical-gate (VG) TFT NAND Flash using junction-free buried channel BE-SONOS device," in *VLSI Symp. Tech. Dig.*, pp. 131-132, 2010.

[25] J.-G. Yun, G. Kim, J.-E. Lee, Y. Kim, W.B. Shim, J.-H. Lee, H. Shin, B.-G. Park, and J.D. Lee, “Single-Crystalline Si STacked ARray (STAR) NAND Flash Memory,” *IEEE Transactions on Electron Devices,* vol. 58, no. 4, pp. 1006-1014, Mar. 2011.

[26] Y. Kim, J.-G. Yun, S.H. Park, W. Kim, J.Y. Seo, M. Kang, K.-C. Ryoo, J.-H. Oh, J.-H. Lee, H. Shin, and B.-G. Park, “Three-Dimensional NAND Flash Architecture Design Based on Single-Crystalline STacked ARray,” *IEEE Transactions on Electron Devices,* vol. 59, no. 1, pp. 34-45, Jan. 2012.

[27] W. Kim, J.Y. Seo, Y. Kim, S.H. Park, S.H. Lee, M.H. Baek, J.-H. Lee, and B.-G. Park, “Channel-stacked NAND flash memory with layer selection by multi-level operation (LSM),” in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2013, pp. 3.8.1–3.8.4.

[28] S.-H. Lee, W. Kim, D.W. Kwon, J.Y. Seo, M.H. Baek, S. Lee, J. Kang, W. Jang, J.-H. Lee, B.-G. Park, “Layer selection by multi-level permutation in 3-D stacked NAND flash memory,” *IEEE Electron Device Lett*., vol. 37, no. 7, pp. 866–869, Jul. 2016

[29] J.-G. Yun, S.H. Park, and B.-G. Park, “LAyer Selection by ERase (LASER) With and Etch-Through-Spacer Technique in a Bit-Line Stacked 3-D NAND Flash Memory Array,” *IEEE Transactions on Electron Devices,* vol. 58, no. 7, pp. 1892-1897, Jul. 2011.

[30] K.-D. Suh B.-H. Suh, Y.-H. Lim, J.-K. Kim, Y.-J. Choi, Y.-N. Koh, S.S. Lee, S.-C. Kwon, B.-S. Choi, J.-S. Yum, J.-H. Choi, J.-R. Kim, H.-K. Lim, ”A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme,” in *ISSCC Tech. Dig.*, Feb. 1995, pp. 128-129.

[31] K. Takeuchi, S. Satoh, K. Imamiya, and K. Sakui, “A source-line programming scheme for low-voltage operation NAND Flash memories,” *IEEE J. Solid-State Circuits*, vol. 35, no. 5, pp. 672-681, May. 2000.

[32] T. Cho, Y.-T. Lee , E.-C. Kim, J.-W. Lee, S. Choi, S. Lee, D.-H. Kim, W.-G. Hna, Y.-H. Lim, J.-D. Lee, J.-D. Choi, and K.-D. Suh, “A dual-mode NAND flash memory: 1-Gb multilevel and high-performance 512-Mb single-level modes,” *IEEE J. Solid-State Circuits*, vol. 36, no. 11, pp. 1700-1706, Nov. 2001.

[33] J. Lee, S.-S. Lee, O.-S. Kwon, K.-H. Lee, K.-H. Lee, D.-S. Byeon, I.-Y. Kim, Y.-H. Lim, B.-S. Choi, J.-S. Lee, W.-C. Shin, J.-H. Choi, and K.-D. Suh, "A 90-nm CMOS 1.8-V 2-Gb NAND flash memory for mass storage applications," *IEEE J. Solid-State Circuits*, vol. 38, no. 11, pp. 1934-1942, Nov. 2003.

[34] T.-S. Jung, Y.-J. Choi, K.-D. Suh, B.-H. Suh, J.-K. Kim, Y.-H. Lim, Y.-N. Koh, J.-W. Park, K.-J. Lee, J.-H. Park, K.-T. Park, J.-R. Kim, J.-H. Lee, and H.-K. Lim, "A 3.3 V 128 Mb multi-level NAND flash memory for mass storage applications," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 1996, pp. 32-33.

[35] J.-K. Kim, H.-B. Pyeon, H. Oh, R. Schuetz, and P. Gillingham, "Low stress Program and Single Wordline Erase Schemes for NAND Flash Memory," in *Proc. IEEE NVSMW 22st Non-Volatile Semicond. Memory Workshop*, Feb. 2007, pp. 19-20

[36] K.-T. Park, Y. Song, M. Kang, S. Lee, Y.-H. Lim, K.-D. Suh, C. and Chung, "Dynamic Vpass Controlled Program Scheme and Optimized Erase Vth Control for High Program Inhibition in MLC NAND Flash Memories, " *IEEE J. Solid-State Circuits*, vol. 45, no. 10, pp. 2165-2172, Oct. 2010.

[37] W. Kim, J.H. Lee, J.-G. Yun, S.e Cho, D.H. Li, Y. Kim, D.-H. Kim, G.S. Lee, S.H. Park, W.B. Shim, J.-H. Lee, H. Shin, and B.-G. Park, "Arch NAND Flash Memory Array With Improved Virtual Source/Drain Performance," *IEEE Electron Device Letters,* vol. 31, no. 12, pp. 1374-1376, Dec. 2010.

[38] E.-S. Choi, and S.-K. Park, "Device considerations for high density and highly reliable 3D NAND flash cell in near future," in *IEDM Tech. Dig.*, 2012, pp.9.4.1–9.4.4.

[39] H. Yoo, E.-S. Choi, J. Oh, K. Park, S. Jung, S. Kim, K. Shim, H. Joo, K. Jeon, M. Seo, Y. Jang, S. Lee, J. Lee, S. Oh, G. Cho, S. Park, S. Lee, and S. Hong, "Modeling and optimization of the chip level program disturbance of 3D NAND Flash memory." in *Proc. IEEE Int. Memory Workshop,* 2013, pp. 147–150.

[40] J.-D. Lee, C.-K. Lee, M.-W. Lee, H.-S. Kim, K.-C. Park, and W.-S. Lee, "A new programming disturbance phenomenon in NAND flash memory by source/drain hot-electron generated by GIDL current," in *Proc. IEEE NVSMW 21st Non-Volatile Semicond. Memory Workshop,* Feb. 2006, pp. 31-33.

[41] S.J. Joo, H.J. Yang, K.H. Noh, H.G. Lee, W.S. Woo, J.Y. Lee, M.K. Lee, W.Y. Choi, K.P. Hwang, H.S. Kim, S.Y. Sim, S.K. Kim, H.H. Chang, G.H. Bae, "Abnormal Disturbance Mechanism of Sub-100nm NAND Flash Memory," *Jpn. J. Appl. Phys*., vol. 45, no. 8A, pp. 6210–6215, Aug. 2006.

[42] K.-T. Park, S. Lee, J.-S. Sel, J. Choi, and K. Kim, "Scalable wordline shielding scheme using dummy cell beyond 40 nm NAND flash memory for eliminating abnormal disturb of edge memory cell," *Jpn. J. Appl. Phys*., vol. 46, no. 4B, pp. 2188–2192, 2007.

[43] T. Melde, M. F. Beug, L. Bach, A. R. Tilke, R. Knoefler, U. B. Sarlette, V. Beyer, M. Czernohorsky, J. Paul, T. Mikolajick, "Select Device Disturb Phenomenon in TANOS NAND Flash Memories," *IEEE Electron Device Lett.*, Vol. 30, No. 5, pp. 568-570, May. 2009.

[44] H.-H. Wang P.-S. Shieh, C.-T. Huang, K. Tokami, R. Kuo, S.-H. Chen, H.-C. Wei, S. Pittikoun, S. Aritome, "A new read-disturb failure mechanism caused by boosting hot carrier injection effect in MLC NAND Flash memory," in *Proc. IEEE Int. Memory Workshop*, May 2009, pp. 40-41.

[45] B.-I. Choe, J.-K. Lee, B.-G. Park, and J.-H. Lee, “Suppression of read disturb fail caused by boosting hot carrier injection effect for 3-D stack NAND flash memories,” *IEEE Electron Device Lett*, vol. 35, no. 1, pp. 42-44, Jan. 2014.

[46] Shijun Liu, and Xuecheng Zou, “QLC NAND study and enhanced Gray coding methods for sixteen-level-based program algorithms,” *Microelectronics Journal*, vol. 66, pp.58-66, Aug. 2017.

[47] B. Kim, S.-H. Lim, D.W. Kim, T. Nakanishi, S. Yang, J.-Y. Ahn, H. Choi, K. Hwang, Y. Ko, and C.-J. Kang, “Investigation of ultra thin polycrystalline silicon channel for vertical NAND flash,” Proc. IEEE Int. Reliability Physics Symposium, Apr. 2011, pp. 2E.4.1-2E.4.4.

[48] E. Nowak, J.-H. Kim, H. Kwon, Y.-G. Kim, J.S. Sim, S.-H. Lim, D.S. Kim, K.-H. Lee, Y.-K. Park, J.-H. Choi, and C. Chung, “Intrinsic fluctuations in Vertical NAND flash memories,” in *VLSI Symp. Tech. Dig.*, pp. 21-22, 2012.

[49] W.-J. Tsai, W.L. Lin, C.C. Cheng, S.H. Ku, Y.L. Chou, L. Liu, S.W. Hwang, T.C. Lu, K.C. Chen, T. Wang, and C.-Y. Lu, “Polycrystalline-Silicon channel

trap induced transient read instability in 3D NAND flash cell string," in *IEDM Tech. Dig.*, pp. 11.3-4, Dec. 3-7, 2016.

[50] M. K. Hatalis and D. W. Greve., "Large grain polycrystalline silicon by low-temperature annealing of low-pressure chemical vapor deposited amorphous silicon films," *J. Appl. Phys.* vol. 63, pp. 2260-2266, 1988.

초 록

최근 태플릿, 랩탑, 스마트폰 또는 솔리드 스테이드 드라이브와 같은 전자 기기의 주 저장 장치로 낸드 플래시 메모리가 사용됨에 따라 저비용, 고용량의 낸드 플래시 메모리에 대한 수요가 매우 빠르게 증가하고 있다. 그러나 사진공정 기술의 한계, 셀 간 간섭현상, 그리고 플로팅 게이트에 저장되는 전자 수의 감소와 같은 문제들이 지속적인 플로팅 게이트 낸드 플래시 메모리의 축소화를 어렵게 한다. Toshiba 사가 bit-cost scalable (BiCS) technology 를 제안한 이후, 다양한 유형의 삼차원 적층 차지 트랩 낸드리 플래시 메모리가 축소화, 간단한 공정, 그리고 커플링 현상 부재의 이유로 앞에서 말한 문제점들을 해결할 수 있는 수단으로 고려되어 왔다.

삼차원 낸드 플래시 메모리 구조는 두 종류로 나뉠 수 있다. 첫 번째는 게이트 적층 낸드 플래시 메모리다. 여기서 전류는 수평하게 형성된 게이트들과 수직하게 만들어진 채널을 따라 흐른다. 나머지 하나는 채널 적층 낸드 플래시 메모리로, 여기서 전류는 수평하게 형성된 채널을 따라 흐른다. 채널 적층

형태는 작은 유닛 셀 크기, 빗라인 핏치 축소화, 실리콘/실리콘저머늄의 에피택셜 성장 공정에 따른 단결정 실리콘 채널을 사용할 수 있으며, 적층 수가 증가하여도 전류 감소 문제가 없다는 점에서 장점을 가진다. 하지만 게이트 적층 형태에 비해 채널 적층 형태는 구조가 복잡하며 셀 접근 방법이 복잡하다는 문제점을 가진다. 이러한 문제들을 해결하기 위해, 우리 연구 그룹에서는 다중 레벨 동작을 이용한 층선택 방법을 제안하였다. 하지만 기존에 제안된 구조와 동작 방법으로는 층선택 동작을 위한 스트링 셀렉트 트랜지스터들의 초기화 작업을 수행할 수 없으므로 이에 대한 개선이 필요한 상황이었다. 따라서 이 논문에서는 층선택 동작을 위한 채널 적층 형태의 낸드 플래시 메모리 구조를 제안하고 스트링 셀렉트 트랜지스터들의 문턱전압 값을 설정할 수 있는 동작 방법들을 새롭게 제안하였다. 그리고 이를 검증하기 위해 티캐드 시뮬레이션과 제작된 소자의 측정을 진행하였다. 또한 기본적인 메모리 동작 중에 발생할 수 있는 다양한 디스터번스 현상에 대해 분석하고 이를 완화시킬 수 있는 방법을 제안하고 검증하였다.

주요어: 3차원 적층 낸드 플래시 메모리, 낸드 플래시 메모리 아키텍쳐, 게이트 적층 낸드 플래시 메모리, 채널 적층 낸드 플래시 메모리, 다중 레벨 동작에 따른 층 선택

학 번: 2013-30221