M.S. THESIS

Investigation of Three Dimensional NAND Flash Memory Based on Gate STacked ARray (GSTAR)

Gate 적층 구조를 가지는 3차원 NAND Flash Memory의 제안 및 검증

BY

Do-Bin Kim

February 2013

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

Investigation of Three Dimensional NAND Flash Memory Based on Gate STacked ARray (GSTAR)

Gate 적층 구조를 가지는 3차원 NAND Flash Memory의 제안 및 검증

지도교수 박 병 국

이 논문을 공학석사 학위논문으로 제출함

2013 년 2 월

서울대학교 대학원

전기컴퓨터공학부

김 도 빈

김도빈의 공학석사 학위논문을 인준함

2013 년 2 월

위 원 장 : 신 형 철 (인)

부위원장 : 박 병 국 (인)

위 원 : 이 종 호 (인)

ABSTRACT

We've developed a Gate STacked ARray (GSTAR) structure as a three-dimensional memory, which extremely reduces the chip costs by vertically stacking memory arrays with easier process steps. The proposed device is a gate stack type structure having U-shaped channel to achieve high memory density without shrinking cell channel length. The GSTAR, by using an ultra-thin body structure, can reduce the off-current level, and planar cell type of GSTAR assures insensitivity to process variables such as etch-slope. The performance of designed structure is described and the optimization of device parameter is performed by using TCAD simulation. To increase device performance and ease of fabrication, the modified fabrication method to reduce the spacing between gates is also introduced.

Keywords:3D NAND flash memory, word-line stacking, vertical channel, ultra-thin body.

Student number: 2011-20800

CONTENTS

1. Introduction

To satisfy the explosive demands of mobile storage having high capacity, the density of NAND flash memories has been increased dramatically. Therefore, the technical issues in scaling down of the NAND flash memory are still important. However, there are some physical limitations, such as Short Channel Effect (SCE), reduction of storage charge numbers, cell-to-cell interference, and coupling ratio.

To overcome these issues of conventional NAND flash memory type, various three dimensional stacked NAND flash architectures have been proposed, after the Bit-Cost Scalable (BiCS) technology was introduced [1]-[3]. There are two types according to stacking type in three-dimensional stackable memory devices; one is gate stack type, and the other is channel stack type [4],[5]. All proposed structures have pros and cons.

To expand flash memory array to a three dimensional structure, in this

paper, we propose the Gate-STacked-ARray (GSTAR), having vertical ultra-thin channel and stacked planar type word-lines, while most word-line stacked type arrays have Gate-All-Around (GAA) structure. The GSTAR has many advantages comparing with other GAA structures, such as easy fabrication, small unit cell size, insensitivity to process variables and high density. We compare these characteristics of GSTAR to the GAA structure. And we investigate characteristics of GSTAr and optimize the device parameters for good performance by suing the TCAD simulation.

2. Investigation of GSTAR

2. 1 Introduction

As cell size of NAND flash memory rapidly shrinks, photolithography and etch processes are became difficult to make device. Also reduction of stored charges, short channel effect and cell interference problems are come out. Therefore, 3D stack NAND flash memory structure has been proposed for solution to these problems. However, suggested structures have some problems that are body erase possibility, metal gate use possibility and difficulty of stacking extension. Moreover, in case of Gate-All-Around structure, there are difficulty of process steps and cell size variation problems. Therefore, it is necessity to suggest new 3D stack NAND flash memory for solution to these problems of proposed structure.

2. 2 Suggestion of 3D stacking NAND flash memory array structure

In this thesis, we propose new 3D NAND flash memory, which is named Gate STacked Array (GSTAR), and analyze electrical characteristics of proposed NAND flash memory structure. To solve above mentioned issues, newly suggested GSTAR have planar cell type. Then we can expect reduction of cell size variation by etch slope, possibility of device fabrication which use metal gate and ease of stacking expansion. We will verify operation of GSTAR, optimize device parameters for improved performance by TCAD simulation.

To achieve independent operation of selected string, bit-lines and SSLs cross each other perpendicularly. We adopt virtual S/D generated by fringing field at memory cell and select transistors (SSLs and GSL), and the Common Source lines (CSLs) exist in Si substrate. Thus, the bulk erase operation is available in the GSTAR.

2. 3 Operation scheme for GSTAR

Fig. 2.1 shows basic structure of Silicon-Oxide-Nitride-Oxide-Silicon (SONOS) flash memory. In SONOS structure, high voltage is supplied to word line to make strong electrical field from word line to substrate, then electron can penetrate through tunnel oxide to silicon nitride layer by FN tunneling and trapped in silicon nitride layer. The threshold voltage of cell is increased by existence of trapped electrons in silicon nitride layer. And erase operation is carried out by supplying high voltage to substrate. By using these phenomenons, program state and erase state are distinguished.

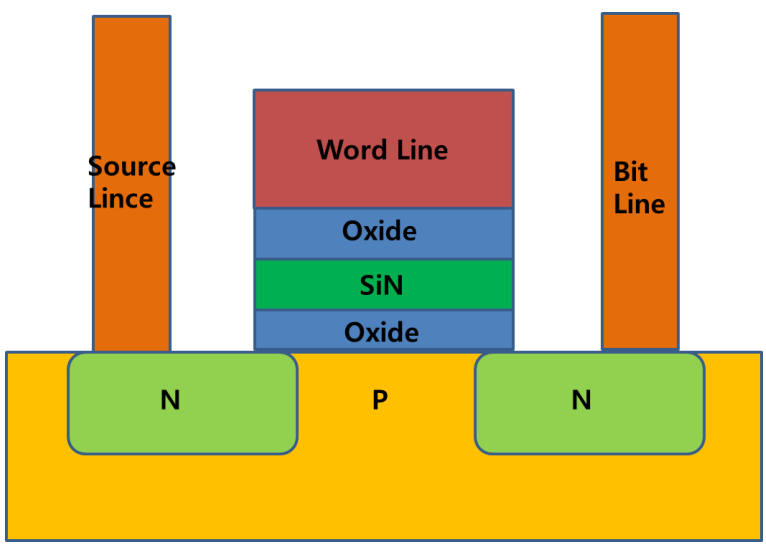


Fig. 2.1 Basic structure of SONOS flash memory

The conventional NAND flash memory array have organized intersecting arrangement of word line and bit line as shown in Fig. 2.2. To achieve each operation in selected cell, different voltages are supplied to each word line, bit line, GSL and SSL. By using a bias scheme which is constituted various voltages, we can operate a device properly.

To expand flash memory array to 3 dimensional structures, various three dimensional stacked NAND flash architectures have been proposed. There are two types according to stacking type in three-dimensional stackable memory devices; one is gate stack type, and the other is channel stack type. All proposed structures have pros and cons. In this thesis, we propose gate stack type and vertical channel structure. The proposed structure is shown in Figure 2. 3.

The proposed GSTAR is a three-dimensional NAND flash memory of the word-line stacked type. Poly-Si thin channels are vertically placed on mesa structures consisting of word-lines and Inter Word-line Dielectrics (IWDs). To distinguish between selected and unselected string, bit-lines and SSLs are perpendicularly crossed each other. The source/drain of each memory cell and

select transistors (SSL and GSL) is induced by the fringing field of adjacent gates, and the bulk exists on Si substrate. Thus, the bulk-erase operation is available at the GSTAR.

In case of GSTAR, bit lines are place in x direction, SSLs in y direction and word line in z direction. Figure 2.4 shows equivalent circuit for this arrangement of GSTAR. This construction can be considered that conventional NAND flash memory devices are placed in y direction additionally.

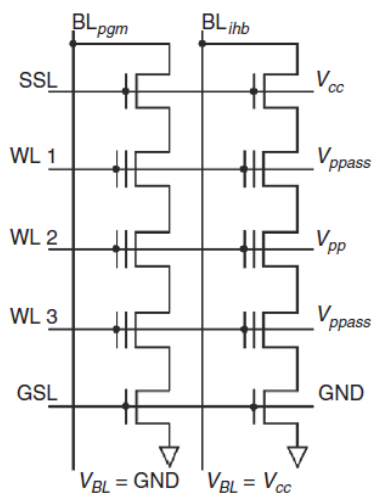


Fig. 2.2 NAND flash memory array structure and bias scheme condition

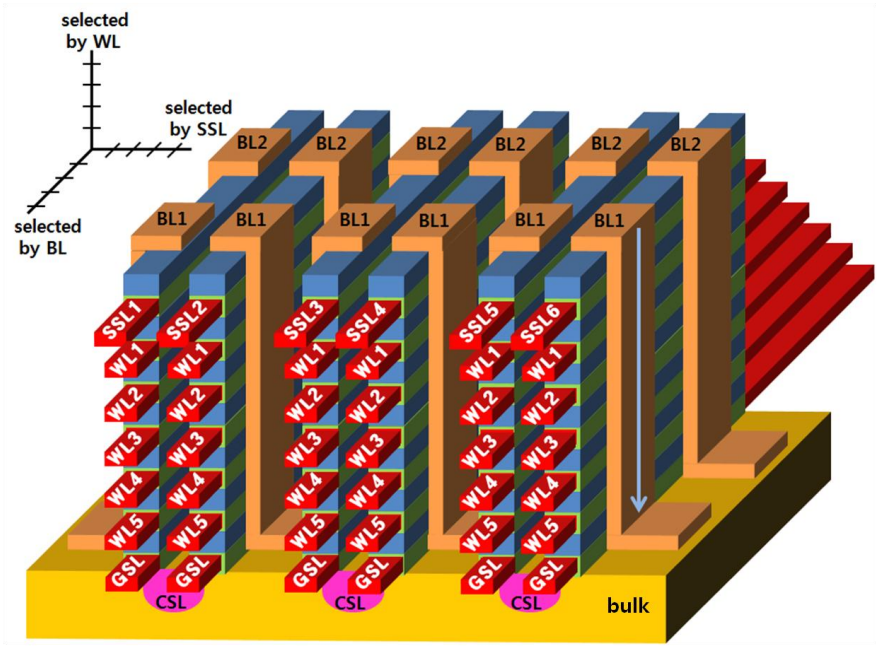


Fig. 2.3 Bird's eye view of GSTAR

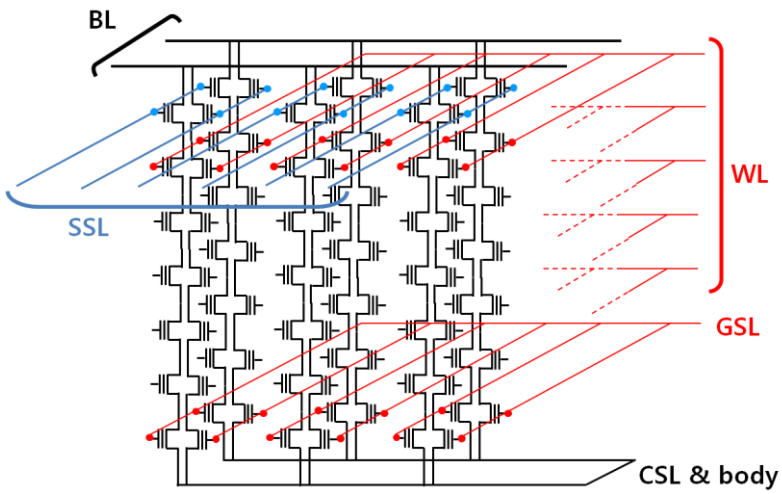


Fig. 2.4 Equivalent circuit of GSTAR

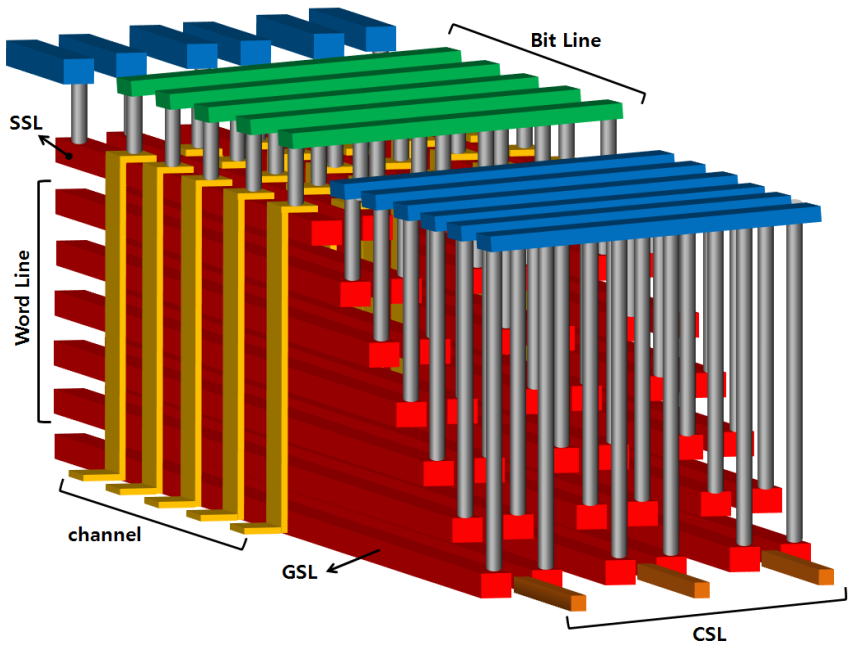


Fig. 2.5 3 Dimensional metal line of GSTAR

To minimize an extra area required for array integration, proper a 3-dimensional metal line construction which is shown in Fig. 2.5 is necessary [4]. In case of word line, we select stair-like structure at the array edge to supply same voltage to cells which are on same floor. The bit lines shares same metal line in x direction as shown in Figure 2. 3, and SSL use metal lines separately to operate each string. And CSLs are placed on substrate for body erase.

Because same voltage is supplied to line which are on same direction or

floor, thus we control bias scheme to each bit line, SSL and word line to operate the selected cell which is placed on intersected point of a face generated by supplied voltage independently. And this bias scheme for each memory operation is show in table. 1. The operation scheme is almost same with a conventional NAND flash memory, except the GSTAR requires a SSL decoder to access a selected channel among the strings that share bit-line.

In case of program operation, we supply program voltage to word line which contains selected cell, Vcc to SSL which is in same stack with selected cell and 0V to bit line which is on same stack with selected cell. Therefore, channel which wrap around selected stack will be separated to one which contain selected cell and another which is not contain. When selected cell is programmed, unselected cell which is in common word line become program inhibit state by self-boosting [6]. In case of read operation, adequate bias is applied to each word line, bit line, SSL to operate selected cell. In case of erase operation, all cells which is in common block can be erased by body erase.

	PROGRAM	READ	ERASE
selected WL	V_{pgm}	0 V	0 V
unselected WL	V_{pass}	V_{read}	float
selected SSL	V_{cc}	V_{read}	float
unselected SSL	0 V	0 V	float
selected BL	0 V	V_{cc}	float
unselected BL	V_{cc}	0V	float
GSL	0 V	V_{cc}	float
CSL	0 V	0V	float
bulk	0 V	0V	V_{ers}

Table. 1 Bias scheme for GSTAR

A formerly suggested 3D NAND flash memory structures have GAA cell type. On the other hand, GSTAR use planar type gate. As result of use of planar gate, we can fabricate suggested device relatively easy. Moreover, GSTAR has insensitivity to process variables such as etch slope. In case of GAA structure, size variation problem occurs by etch slope, electrical field to channel can be different although same voltage is applied to cells. Therefore, different result can be occurred in cells on which same operation perform. However, in case of GSTAR, although there are size variation by etch slope, electrical field to

adjacent cell is same. Therefore, each cell can perform same operation in GSTAR.

Table 2 shows that GSTAR compares with other 3D stack NAND flash memory on various characteristics. First, P-BiCS, TCAT and GSTAR are gate stack type, and 3D VG NAND is channel stack type. And 3D VG NAND has double gate cell type, GSTAR has planar cell type. Therefore GSTAR have some advantages with small unit cell size, size variation, staking extendability, use of metal gate and possibility of body erase. Especially, in case of unit cell size, GSTAR has about half cell area comparing to GAA structure. However, because GSTAR is gate stack type, as stack number is increased channel length lengthen. Therefore, read current will be reduced by longer channel. This problem should be solved in case of gate stack type 3D NAND flash memory.

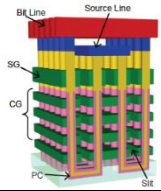
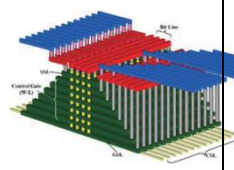
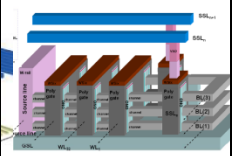
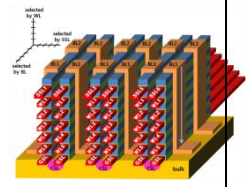
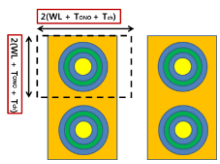
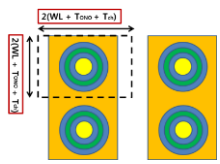
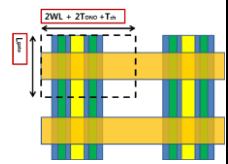
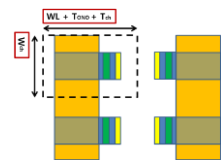
	P-BiCS	TCAT	3D VG NAND	GSTAR
Structure				
Cell type	GAA	GAA	Planar	Planar
Unit cell size				
Size variation by etch-slope	Bad	Bad	Bad	Good
Stacking extendability	Good	Good	Good	Good
Metal gate	X	O	O	O
Limitation of the number of stacks	Low read current	Low read current	No degradation of read current	Low read current

Table. 2 The comparison of GSTAR with other 3D NAND flash memory

2. 3 Electrical analysis for GSTAR by TCAD simulation

To analyze electrical characteristics of suggested 3D NAND flash memory we performed TCAD simulation by using SILVACO. The models which are used in simulation are band-to-band tunneling, dynasonos (tunneling mechanism, carrier transport and charge trap in silicon nitride layer), Shockley-Read-Hall model. Tungsten is used for gate material, length, height and width of gate are 50nm. And thickness of oxide, nitride and oxide layers is 3nm, 6nm and 6nm. And thickness of channel is 10nm. Because channel is very thin and vertical structure, uniform doping to channel is very difficult. Therefore, undoped poly silicon is used for channel material. Some other detail parameters are shown in Table. 3.

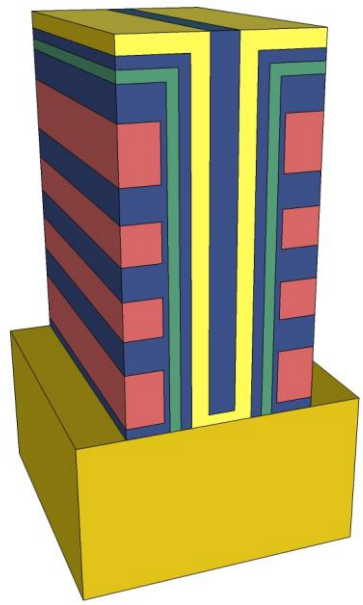


Fig. 2.6 GSTAR structure used in TCAD simulation.

Gate length	50 nm
O /N / O	3 nm / 6 nm / 6 nm
ICD thickness	50 nm
GSL / SSL length	120 nm
IWD length	50nm
Channel doping concentration	Undoped poly silicon
Source / Drain doping concentration	$1x10^{20}$ cm^{-3}

Table. 3 Device parameters are used for TCAD simulation

When program voltage is applied to word lines for program of selected cell, there are concerns about program of unselected cells by sharing program voltage [7]. To solve this problem, self-boosting scheme is used to inhibit other cells from program by simple bias scheme [6]. Thus, self-boosting scheme is essential operation in NAND flash memory operation.

Because GSATR use undoped poly silicon channel, it has relatively low threshold voltage. Thus, GSTAR has weak point about leakage current by low threshold voltage. To reduce leakage current to bit line, we choose selector transistors with 120nm gate length and apply -1V to SSL [8]. When we performed simulation by using bias scheme which is shown in table 3, potential diagram on GSATR is shown in Fig. 2.8. When program operation is progressed during 1ms, we can confirm that channel potential of unselected cell which have common word line with selected cell is boosted effectively. However, there is still leakage current to bit line, it will be solved to operate device properly.

To make selected cell to program state, erase state, we use the bias scheme which is mentioned in table 1. We shift selected cell state from initial state to

programmed state by applying program voltage (16V) during 1ms to selected cell. And then, we shift selected cell state from programmed state to erased state by applying erase voltage (20V) during 10ms to body. Fig. 2.8 shows drain current-gate voltage transfer curve for each state. Threshold voltage is defined as selected gate voltage when drain current is 1e-7[A/um]. Therefore we can guarantee memory window about 6V for GSTAR. This memory window value is sufficient to distinguish each state, and 0~2V can be used for read voltage.

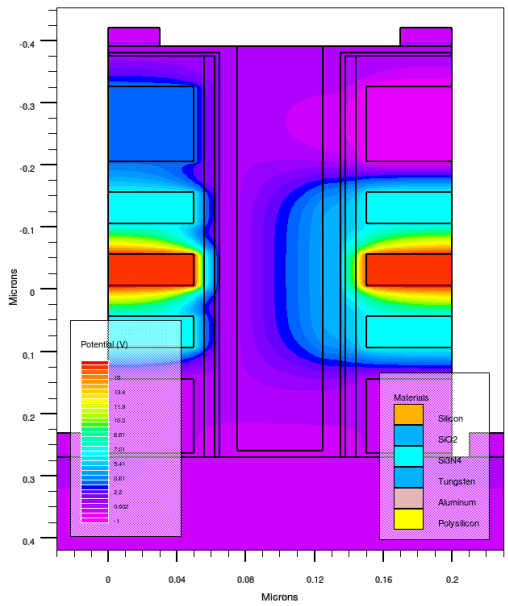


Fig. 2.7 Self-boosting operation of GSTAR

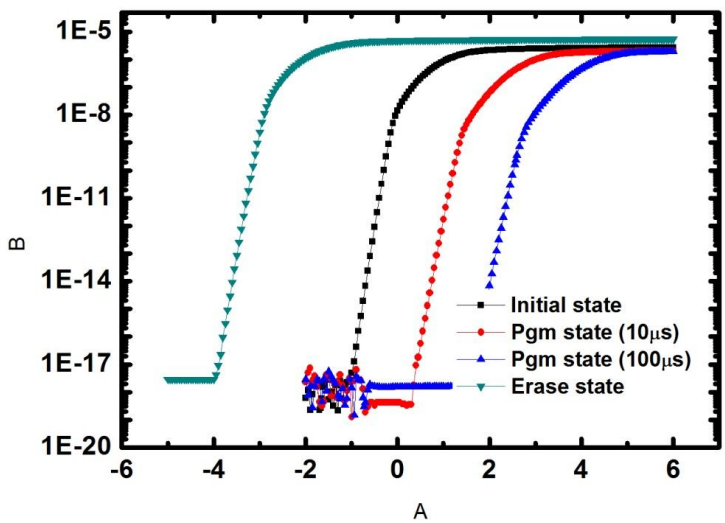
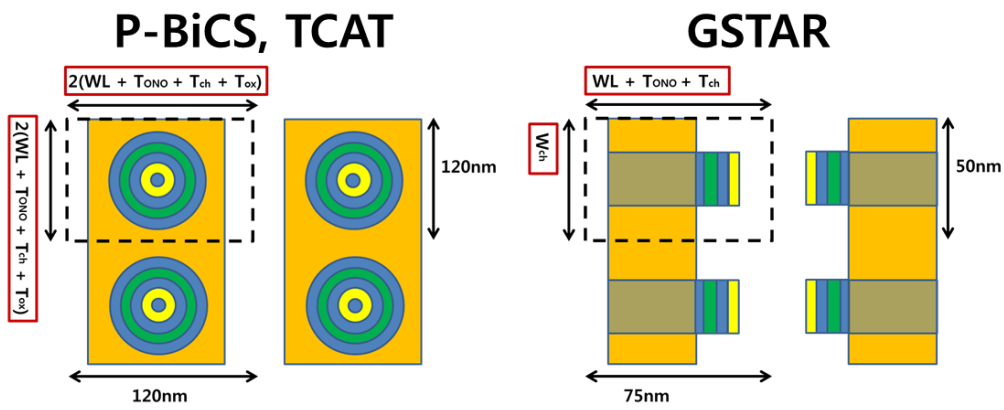
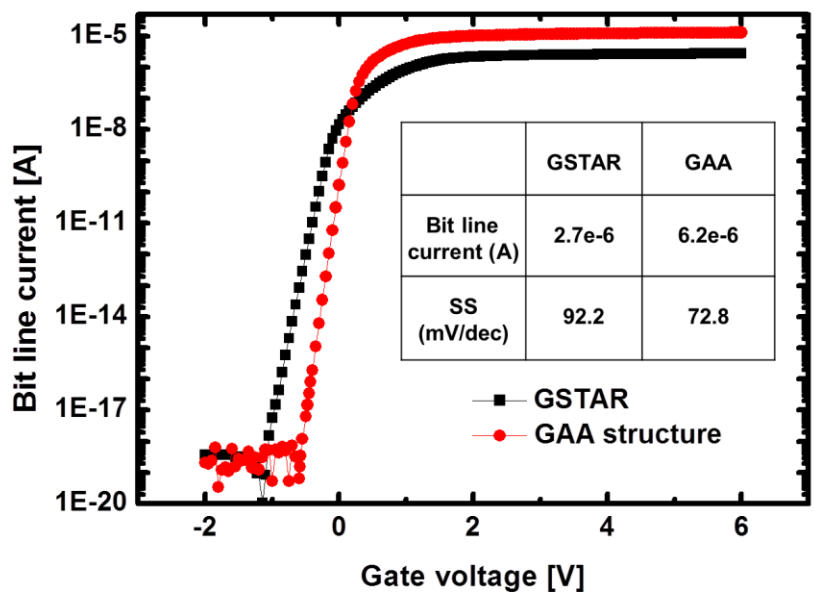


Fig. 2.8 Id-Vg curve on initial, program, erase state

We verify that GSTAR has advantages comparing with other GAA structures in density and insensitivity process variations. Because the punch & plug technology and macaroni structure channel is used in GAA structure fabrication, O/N/O and gate thickness doubles in unit cell size. Therefore, in terms of the array density, the GSTAR has an advantage of scalability as shown in Figure 2.9 a). However, Figure 2.9 b) shows that the electrical characteristics of GAA structure are better than GSTAR. When structures which have the same thickness of O/N/O and channel are used in simulation, however, unit cell size of GAA structure is bigger than that of GSTAR.



a)

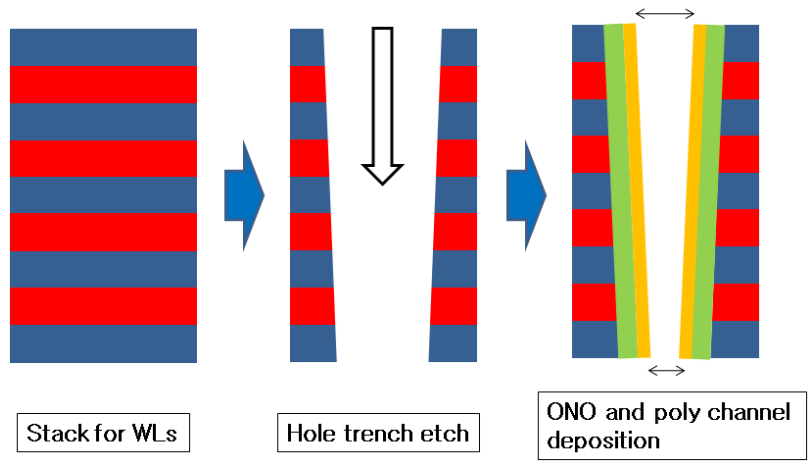


b)

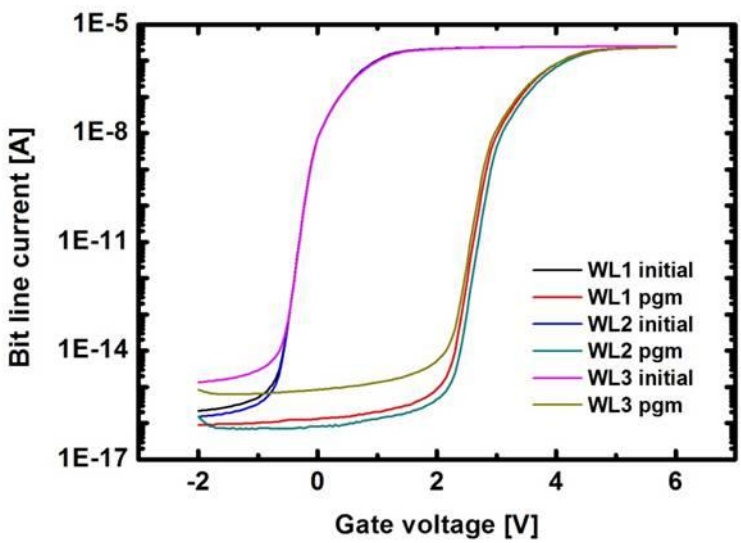
Fig. 2.9 Comparison between GSTAR and GAA structure (a) Unit cell size

(b) Id-Vg curve

Figure 2.10 a) shows etch slope problem by hole trench etch process. To investigate insensitivity to process variables of GSTAR, simulation is performed. When etch slope of about 89° is occurred by hole trench etch, there is difference about 30nm thickness between bottom and top word line in simulated each structure. In case of GAA structure, the different electric field is applied to each cell during program operation by field concentration effect, which causes the threshold voltage distribution problem. In contrast, we confirm little difference in memory operations of GSTAR. Because the channel of GSTAR is formed along the mesa structure instead of a hole, uniform cell characteristics can be achieved regardless of etch slope by almost same lateral electric field.



a)



b)

Fig. 2.10 a) Size variation by hole trench etch process b) Id-Vg curve of each word line by size variation

Fig. 2.11 shows change of threshold voltage according to program voltage to selected cell. As program voltage increases, threshold voltage increases by increment of trapped electrons in silicon nitride layer.

Program voltage and pass voltage are applied to each word line and bit line during program operation. Therefore, disturbance phenomenon which means weak program of unselected cell occurs by applied voltages. Program disturbance and pass disturbance are typical problem in NAND flash memory.

Undoped poly-Si is used as a material of channel of GSTAR by concern about uniform doping on channel in implantation process. The leakage current which flows to bit line due to low threshold voltage is predicted, so that we can verify self-boosting efficiency and disturb effects. Figure 2.12 a) and b) show program/pass disturb effect in NAND flash memory. The SSL cannot turn off the channel completely by low V_{TH} of channel, when self-boosting scheme is operated. Therefore, GSTAR cannot achieve sufficient channel potential. As shown in Fig. 2.12 a), there is program disturbance phenomenon when sufficient pass voltage is applied to adjacent WLs. In case of pass disturbance, there are

trivial variations by change of pass voltage. To solve this problem, selective boron implantation to channel on SSL can be used. Channel on SSL is doped with boron of $1x10^{13}$ cm^{-2} dose and 20keV. We achieve increase of V_{TH} in SSL about 0.5V as shown in Figure 2.13, while V_{TH} in word-lines is not changed much.

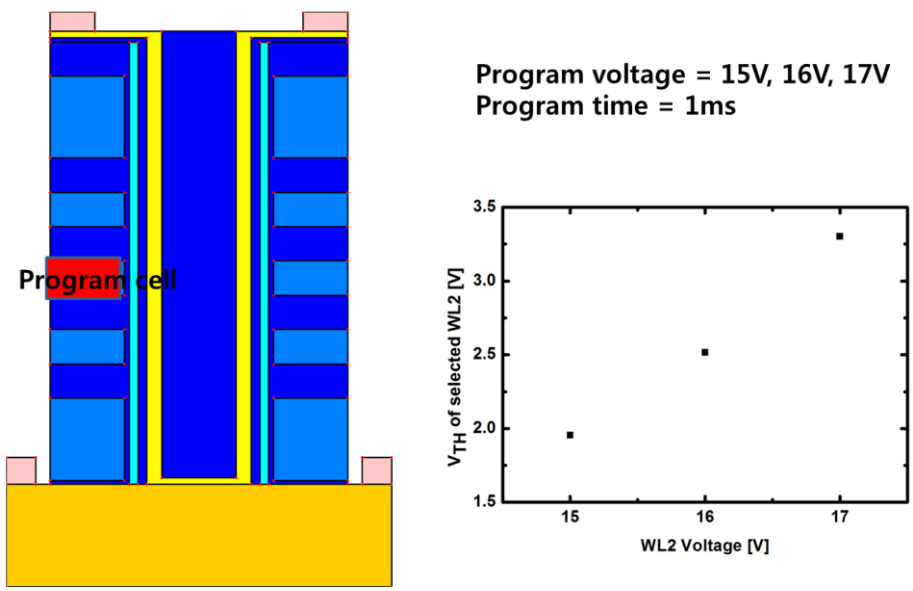
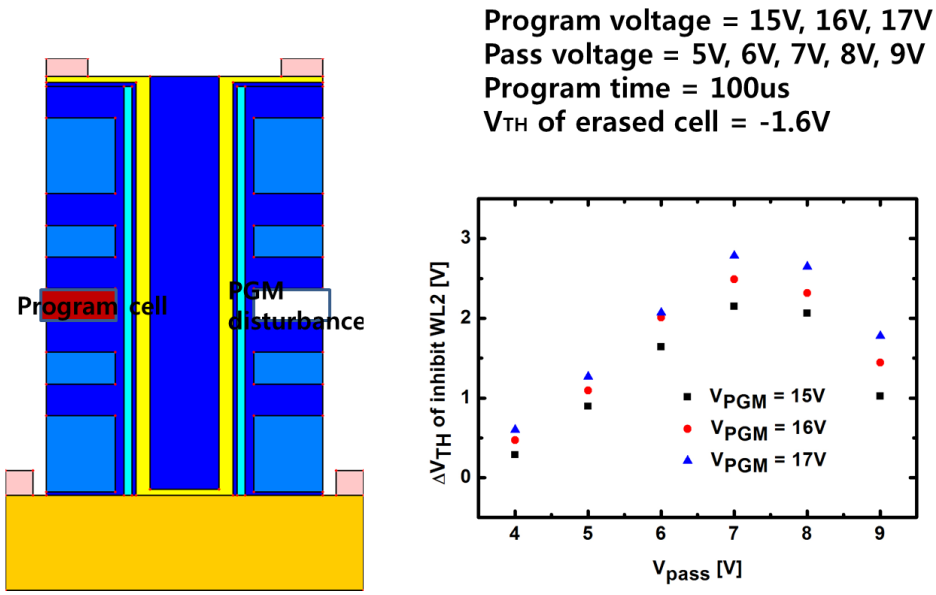
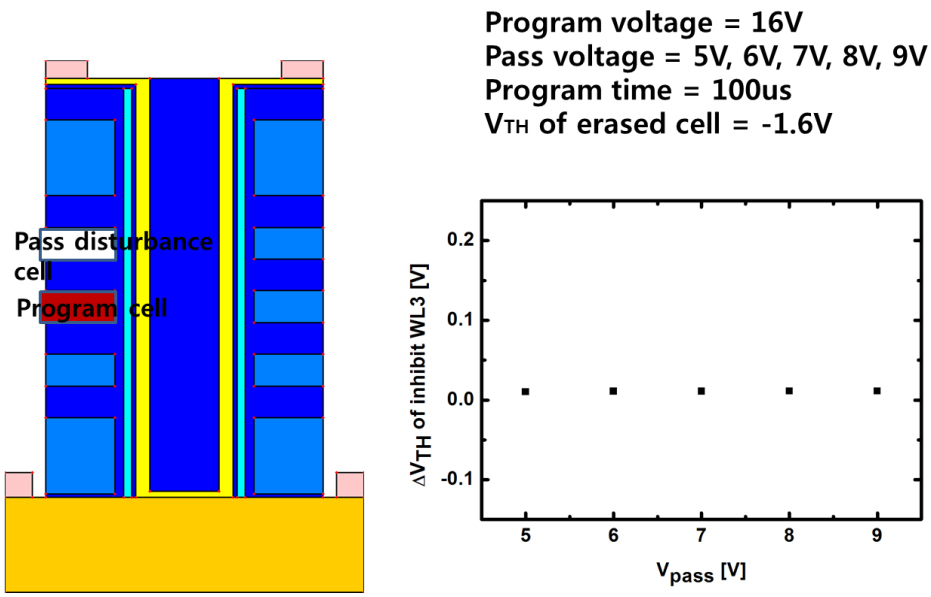


Fig. 2.11 Change of threshold voltage according to program voltage



a)



b)

Fig. 2.12 a) Program disturbance b) Pass disturbance

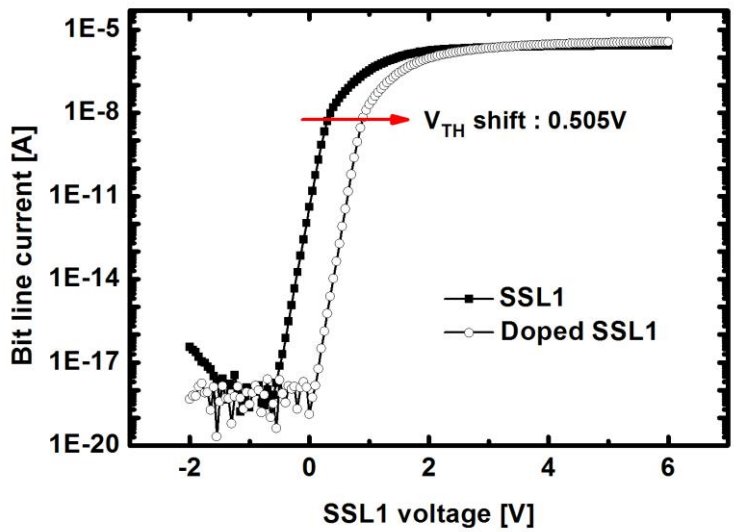
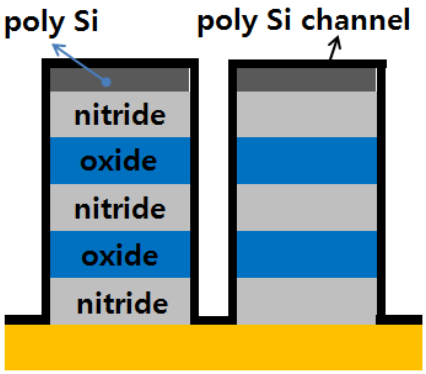


Fig. 2.13 V_{TH} of SSL comparison doped channel with undoped channel.

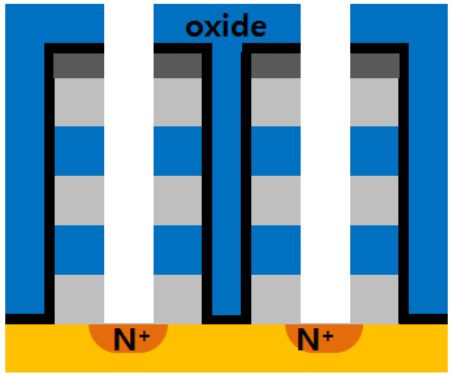
3. Device fabrication method

3. 1 Introduction

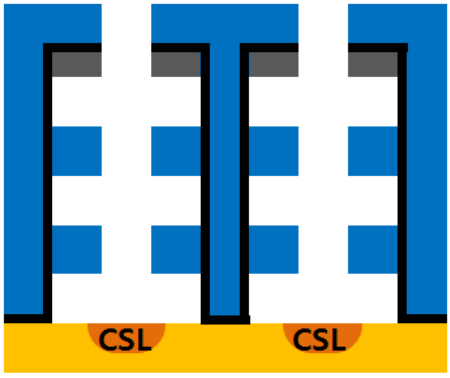
Figure 2.3 and 2.5 show schematic Birds-eye views and the equivalent circuit of the GSTAR flash memory. The key process sequence of the GSTAR fabrication is almost same to TCAT process as shown in figure 3.1. The nitride and oxides films are stacked on Si substrate to use dummy layers and IWDs. Then a trench etch process to form a vertical channel is performed, and a thin poly-Si layer is deposited. Nest, Inter Channel Dielectric (ICD) deposition, active patterning, and CSL implantation is achieved. After an isotropic dummy nitride removal is performed, gate dielectric layers and gate metal are deposited in the conventional order. By adapting a gate replacement method [2], the GSTAR can utilize metal gate structure.



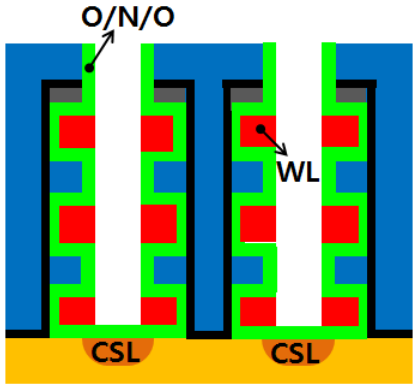
a) Stacking layers and formation of a channel



b) Trench etch and CSL implantation



c) Selective removal of dummy nitride layers

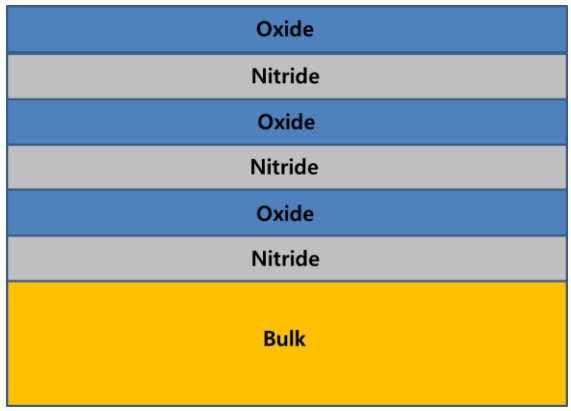


d) Deposition of gate dielectrics and WLs

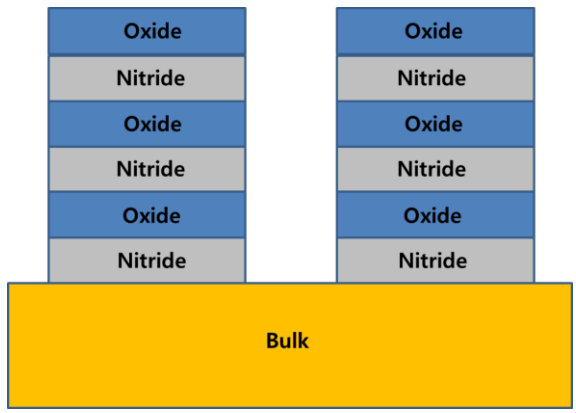
Fig. 3.1 Critical fabrication steps

3.2 Device process scheme

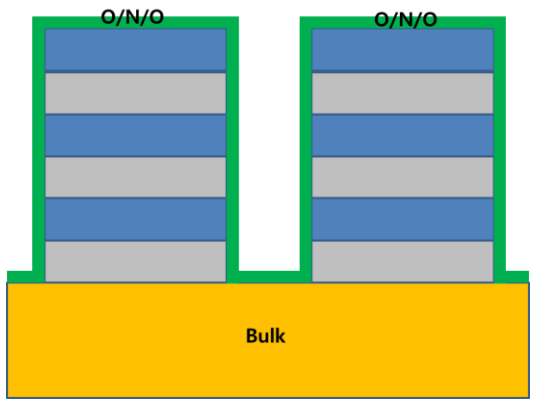
The spacing between gates in designed TCAT cannot be scaled down to sub~50nm since two gate dielectric layers (O/N/O) are inserted between word-lines. A wide band gap between word-lines leads to several problems. At first, it is hard to achieve stable virtual S/D characteristic and enough bit-line current level. A total height, also, is a burden since constant through-layer etch cannot be guaranteed as the aspect ratio increases. To solve this drawback, we proposed additionally modified process for GSTAR fabrication in figure 3.2. Using modified process method, O/N/O layers in the word-lines gap can be eliminated. The proposed fabrication process is as follows. After forming mesa structure consisting of nitride and oxide layers, O/N/O layers and poly-Si channel are deposited. Nest, ICD deposition, trench-etch, and CSL implantation is achieved. Then, dummy nitride layers are removed and word-lines are formed. As a consequence the modified process can reduce the total stack height and channel resistance.



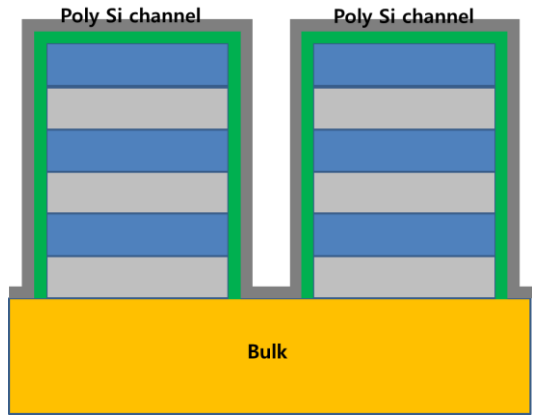
a) Stacking layer of dummy nitride and oxides films



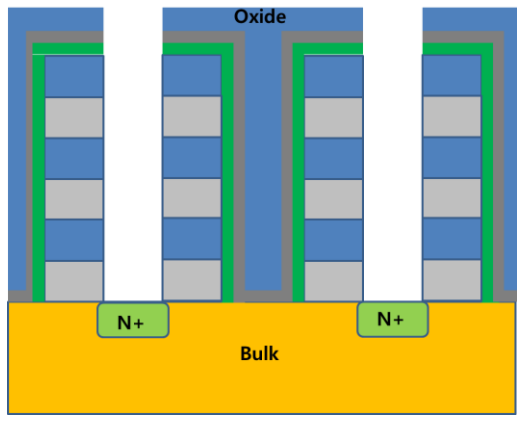
b) Trench etch



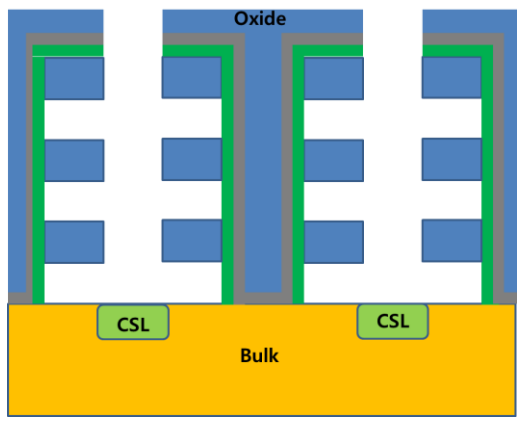
c) Formation of a gate dielectrics



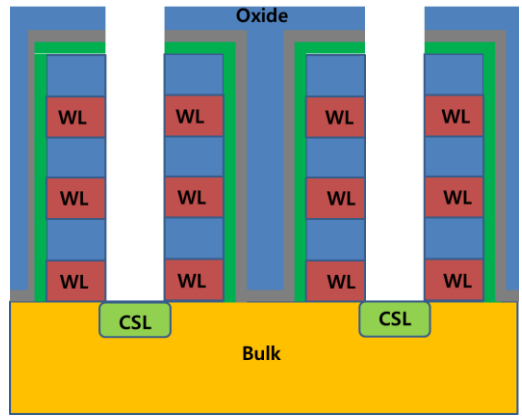
d) Formation of a channel



e) Trench etch and CSL implantation



f) Selective removal of dummy nitride layers



g) Formation of metal gates

Fig. 3.2 Fabrication steps of the modified GSTAR

6. Conclusion

In this paper, the three dimensional gate stack type NAND flash array is proposed for low cost and high density non-volatile memory. To increase the memory density without the channel length scaling, channels of the GSTAR are placed on stacked WLs vertically. By adapting a gate replacement method, the metal word-lines are utilized and staircase structure to contact each stacked word line-is used. Also the modified fabrication method to reduce the spacing between gates is also introduced.

Compared with previous 3D GAA structure, GSTAR provides advantages which are insensitivity to process variables and high density as well as ease of fabrication. Furthermore, almost same operations with a conventional NAND flash memory, including bulk-erase, can be available. And channel implantation is provided to improve self-boosting efficiency by increasing threshold voltage

of channel on SSL.

Therefore, the GSTAR will be a promising candidate that can solve drawbacks of previous stacked NAND flash memories.

Bibilography

[1] H. Tanaka et al., “Bit Cost Scalable Technology with punch and Plug Process for Ultra High Density Flash Memory,” in VLSI Tech. Dig., 2007, pp. 14-15.

[2] R. Katsumata et al., “Pipe-shaped BiCS Flash Memory with 16 stacked Layers and Multi-Level-Cell Operation for Ultra High Density Storage Devices,” in VLSI Tech. Dig., 2009, pp. 136-137

[3] A. Hitayama and H.Aochi., “Bit Cost Scalable (BiCS) Flash Technology for Future Ultra High Density Storage Devices,” in VLSI Technology Systems and Applications. Dig., 2010, pp. 130-131.

[4] J. Jang et al., “Vertical Cell Array using TCAT (Terabit Cell Array Transistor) Technology for Ultra High Density NAND Flash memory,” in VlSI Tech. Dig., 2009, pp. 192-193

[5] H.T. Lue et al., “A Highly Scalable 8-layer 3D Vertical-Gate (VG) TFT NAND Flash Using Junction Free Buried Channel BE-SONOS Device,” in

VLSI Tech. Dig., 2010, pp. 131-132

[6] K. -D Suh et al., “A 3.3 V 32 Mb NAND Flash memory with Incremental Step Pulse Programming Scheme,” Solid-State Circuits Conference, 1995. Digest of Technical Papers. 42nd ISSCC, 1995 IEEE International , pp. 128-129, 350, Feb 1995

[7] K. -D Suh et al., “A 3.3 V 32 Mb NAND Flash memory with Incremental Step Pulse Programming Scheme,” IEEE Journal of Solid-State Circuits, vol. 30. No. 11, pp. 1149-1156, Nov. 1995

[8] J. –D. lee et al., “A New Programming Disturbance Phenomenon in NAND Flash Memory By Source/Drain Hot-Electrons Generated By GIDL Current,” Non-Volatile Semiconductor Memory Workshop, 2006. IEEE NVSMW 2006. 21st, pp. 31-33, 2006.

초 록

3 차원 메모리로서 좀 더 쉬운 공정으로 공정 단가를 크게 절감할 수 있는 적층형 낸드 플래시 메모리, GSTAR 를 제안한다. 이 제안된 소자는 높은 집적도 달성을 위해 채널의 길이가 감소함에 따라 발생하는 문제점들을 방지하고자, 채널 길이가 줄어들지 않도록 U 자형 채널을 가지는 구조로 설계되었다. 그리고 GSTAR 는 박막형 몸체를 가지기 때문에 off 상태에서의 전류를 감소시킬 수 있으며, 기본 소자가 다른 3 차원 낸드 플래시 메모리와는 다르게 평면 형태로 되어 있어 공정 상에서 발생할 수 있는 변수들 (etch slope 등)에 대해 민감하게 반응하지 않게 된다. 제안된 소자의 성능 향상을 위해 TCAD simulation 을 이용하여 여러 변수들을 최적화 하였다. 그리고 소자의 성능과 공정상의 편이를 위하여 게이트들 같의 간격을 줄이는 개선된 공정 방법을 소개하고자 한다.

주요어 : 3차원 낸드 플래시 메모리, 워드 라인 적층, 수직 채널, 박막형 몸체

학 번: 2011-20800