論文 / 著書情報
Article / Book Information

題目(和文)	高信頼性・大容量SRAMの微細化に向けた先端プロセス技術
Title(English)	Advanced Process Technologies for High-Reliability and High-Capacity SRAM Scaling
著者(和文)	石丸一成
Author(English)	Kazunari Ishimaru
出典(和文)	学位:博士（工学）, 学位授与機関:東京工業大学, 報告番号:甲第12273号, 授与年月日:2022年6月30日, 学位の種別:課程博士, 審査員:若林 整,筒井 一生,宮本 恭幸,菅原 聡,角嶋 邦之
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12273号, Conferred date:2022/6/30, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Advanced Process Technologies for High-Reliability and High-Capacity SRAM Scaling

by

Kazunari Ishimaru

Submitted to the
Electrical and Electronic Engineering Graduate Major,
Department of Electrical and Electronic Engineering, School of Engineering,
Tokyo Institute of Technology on May 27th, 2022
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Engineering

Author: _____

Electrical and Electronic Engineering Graduate Major,
Department of Electrical and Electronic Engineering

Certified by: _____

Hitoshi Wakabayashi
Professor of Electrical and Electronic Engineering
Thesis Supervisor

Certified by: _____

Kazuo Tsutsui
Professor of Electrical and Electronic Engineering
Thesis Supervisor

Certified by: _____

Yasuyuki Miyamoto
Professor of Electrical and Electronic Engineering
Thesis Supervisor

Certified by: _____

Satoshi Sugahara
Associate Professor of Electrical and Electronic Engineering
Thesis Supervisor

Certified by: _____

Kuniyuki Kakushima
Associate Professor of Electrical and Electronic Engineering
Thesis Supervisor

Advanced Process Technologies for High-Reliability and High-Capacity SRAM Scaling

by

Kazunari Ishimaru

Submitted to the
Electrical and Electronic Engineering Graduate Major,
Department of Electrical and Electronic Engineering, School of Engineering,
Tokyo Institute of Technology on May 27th, 2022
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Engineering

Abstract

This year marks 75 years since the birth of the transistor in 1947. During this time, semiconductor technology has progressed dramatically, continuing to miniaturize, increasing integration and improving performance with Moore's Law. The market for semiconductor devices has expanded from the original large computers for industrial use to include small office computers, personal computers, and mobiles in the pockets while changing the shapes. The metaverse has been attracting attention as a new market in recent years. While applications have shifted from industry to businesses, homes, and individuals in the past, the metaverse is once again targeting society.

In conjunction with this trend, the amount of information generated annually is also exploding. While there are continuous attempts to increase computational power to process this large amount of data, the increase in power consumption is a significant obstacle. In recent years, with calls for green and carbon neutrality, there has been a strong desire to reduce power consumption and improve energy efficiency in computing as well.

Memory is an essential component in computing and is responsible for storing information. Current computing systems use the von Neumann

architecture, and frequent data transfer between CPUs and memories consumes a lot of energy and degrades performance, known as the "memory wall" problem. Much research on parallel computing by using GPUs and the development of accelerators for efficient computing. If information processing can be performed inside the memory, the load on the CPU can be significantly reduced, and power-efficient computing systems can be realized. There is a strong requirement for a non von Neumann computing architecture to overcome power issues. A deep neural network (DNN) has been paid much attention to image recognition and classifications. Computing in memory (CiM) architecture with the DNN can significantly improve energy efficiency, and there is much research on this topic, including accelerator development. In order to accommodate exploding data, the CiM systems with high-capacity memory are strongly desired.

A NAND flash memory is widely used as a low-cost and high-capacity memory element for storage. Suppose it becomes possible to provide energy-efficient computer systems using a 3D NAND flash memory, which is more than two orders of magnitude larger capacity than the DRAM and cheaper than the DRAM. In order to realize this 3D NAND CiM, the high-capacity SRAM needs to be combined to improve performance and efficiency. This thesis aims to provide technology to embed high-capacity SRAM into 3D NAND without losing the low-cost and high-capacity advantages. From the manufacturing process point of view, package-level integration is the easiest approach. However, its size and cost are the concern. Thus, from a manufacturing cost view points, both 3D NAND and high-capacity SRAM on the same wafer are preferable. However, the manufacturing process and thermal budget between 3D NAND and high-capacity SRAM are different. Therefore, key process modules are extracted from front-end-of-line, middle-of-line, and back-end-of-line modules and studied for modification and/or optimization. The first base process technology node of 90 nm, which is similar to the current peripheral circuits of 3D NAND, was selected. Then, a scalable SRAM cell layout was studied. Mechanical stress affects SRAM cell reliability, and optimum cell layout with MOSFET structure was proposed. This SRAM cell design guideline was validated by the hardware and showed good scalability at least down

to 45 nm node. A gate insulator needed to be optimized to satisfy the standby leakage current specifications of SRAM, and a high-κ/metal gate (HK/MG) should be introduced. To accommodate the process temperature of 3D NAND, a gate-first HK/MG was the only option, and the Hf base gate insulator with poly-Si gate electrode was evaluated. Both optimum Hf and nitrogen concentrations were proposed down to a 32 nm node. A local interconnect (L.I.) technology is beneficial for reducing the SRAM cell size and relaxing bit-line pitches for speed improvement since the conventional copper interconnect cannot be used because of the thermal budget. The impact on device characteristics was studied, and an optimum shallow trench isolation height with MOSFET sidewall structure was proposed. A low-κ inter layer dielectric film was evaluated for future scaling, and material requirements were studied with barrier metal selection. A redundancy fuse process by copper blowing scheme was evaluated, and issues were identified. A high-capacity SRAM can be embedded into 3D NAND with down to 45 nm CMOS technology. Beyond 32 nm node SRAM integration with 3D NAND, a gate-last HK/MG should be introduced from both leakage current and CMOS scaling points of view. A die-to-die bonding between 3D NAND and CMOS peripheral circuits with high-capacity SRAM will be a possible solution. Preliminary results of die warpage impact on the device were presented. An SRAM keep-out area (KOA) is needed to be introduced with further studies. Using above-mentioned design guidelines, one package solid-state drive can be realized as a stepping stone for future one package servers. As a next step, one package server with cryogenic temperature operation will enable further energy reduction and combination with quantum computing systems. Manufacturing cost reduction of future LSIs is one of the issues to satisfy the carbon-neutral requirements. The 3D NAND introduced a breakthrough in the manufacturing process, which realized an efficient manufacturing process with low cost for storage memory. Surprisingly, future CMOS device structure, such as nanosheet CMOS, resembles that of 3D NAND cells. This means that the 3D NAND base manufacturing process can be applied to future CMOS devices, bringing lower cost and higher capacity simultaneously. Thus, the CiM systems realized by 3D NAND with high-capacity SRAM are indispensable for a future sustainable society.

Thesis Supervisor: Hitoshi Wakabayashi
Title: Professor of Department of Electrical and Electronic Engineering

Thesis Supervisor: Kazuo Tsutsui
Title: Professor of Department of Electrical and Electronic Engineering

Thesis Supervisor: Yasuyuki Miyamoto
Title: Professor of Department of Electrical and Electronic Engineering

Thesis Supervisor: Satoshi Sugahara
Title: Associate Professor of Department of Electrical and Electronic Engineering

Thesis Supervisor: Kuniyuki Kakushima
Title: Associate Professor of Department of Electrical and Electronic Engineering

Acknowledgements

The support and encouragement of many individuals contributed to the successful completion of this work. I would like to express sincere gratitude to Professor Hitoshi Wakabayashi for providing technical guidance. I was very fortunate to have had the opportunity to learn from him. I would like to thank Professor Kazuo Tsutsui, Professor Yasuyuki Miyamoto, Associate Professor Satoshi Sugahara and Associate Professor Kuniyuki Kakushima for reading this thesis.

I would like to express deep gratitude to Dr. Nobuo Hayasaka, CEO & President of KIOXIA Corporation, Mr. Masaki Momodomi, CTO of KIOXIA Corporation and Mr. Susumu Yoshikawa, Director of the Institute of Memory Technology Research & Development, KIOXIA Corporation for their managements, helpful advices, and encouragements of this study.

I would like to thank Dr. Yasuo Naruke, former President & CEO of KIOXIA Holding, who helped and encouraged my work since I joined Toshiba Corporation. I also would like to thank my former managers of Toshiba Corporation, Mr. Kenji Maeguchi, Dr. Masakazu Kakumu, Mr. Masaaki Kinugawa, Dr. Fumitomo Matsuoka, Mr. Hidemi Ishiuchi, Mr. Yoshiaki Toyoshima for continuous encouragement and support.

I would like to express my sincere gratitude to Professor Chenming Hu, Professor Tue-Jae King Liu, Dr. John F. Chen, Dr. James Chen, for their guidance during my research period at the University of California, Berkeley, as a visiting industrial fellow.

I also would like to thank to Mr. Kunihiro Kasai, Mr. Yasuhiro Fukaura, Mr. Yasunori Okayama, Ms. Yukari Unno, Mr. Toshiyuki Kondo, Ms. Mariko Takayanagi, Dr. Satoshi Inaba, Mr. Kazuya Ohuchi, Mr. Nobuaki Yasutake, Mr. Makoto Fujiwara, Mr. Akira Hokazono, Ms. Kanna Adachi, Dr. Takeshi Watanabe, Mr. Hirohisa Kawasaki, Dr. Naoyuki Shigyo, Dr. Nobutoshi Aoki, Dr. Hiroyoshi Tanimoto, Mr. Toshiyuki Enda, Dr. Yoshitaka Tsunashima, Dr. Kazuhiro Eguchi, Dr. Hideshi Miyajima, Dr. Atsushi Yagishita, Dr. Akio Kaneko, Dr. Hideki Shibata, Mr. Noriaki Matsunaga, Dr. Takamasa Usui, Mr. Naofumi Nakamura of Toshiba Corporation (some are now with KIOXIA

Corporation) for their significant discussions and practical advices.

The author has been strongly supported by my wife, Yoko, and daughters, Ai and Moe and would like to thank them gratefully.

May 27th, 2022

Kazunari Ishimaru

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Data explosion and energy crisis

In today's advanced information society, information is exploding, and the annual amount of data generated last year reached 79 ZB (zettabyte) [1]. The amount of data generated will continue to increase and is expected to reach 181 ZB by 2025, as shown in Figure 1.1. In 2019, the annual amount of data generated in 2025 was estimated to be 175 ZB, which means an increase of 6 ZB over the past two years. If this trend continues, the amount of data generated in the future will far exceed the current estimation. This data includes text data generated by people, such as e-mails and chats, photos and videos taken using smartphones and data generated by industrial equipment, and new data generated by secondary processing of acquired data. Today, far more data is generated by devices than by people, and this percentage is expected to increase in the future [2].

As the amount of data generated increases, the need to improve the computing performance to process the enormous amount of data increases. As proof of this, datacenters are being built worldwide [3]. As the number of datacenters increases, so does the amount of power they consume. The energy consumed by datacenters worldwide was 190 TWh (terawatt-hour) in 2017, which was about half of the previously predicted amount as shown in Figure 1.2 [4–6].

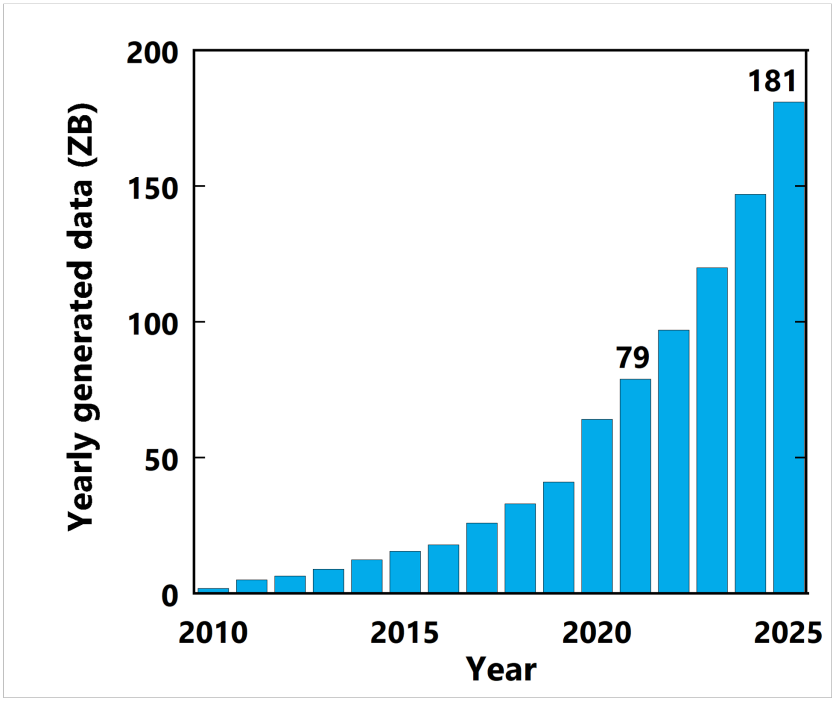


Figure 1.1: Trend of yearly data generation [1].

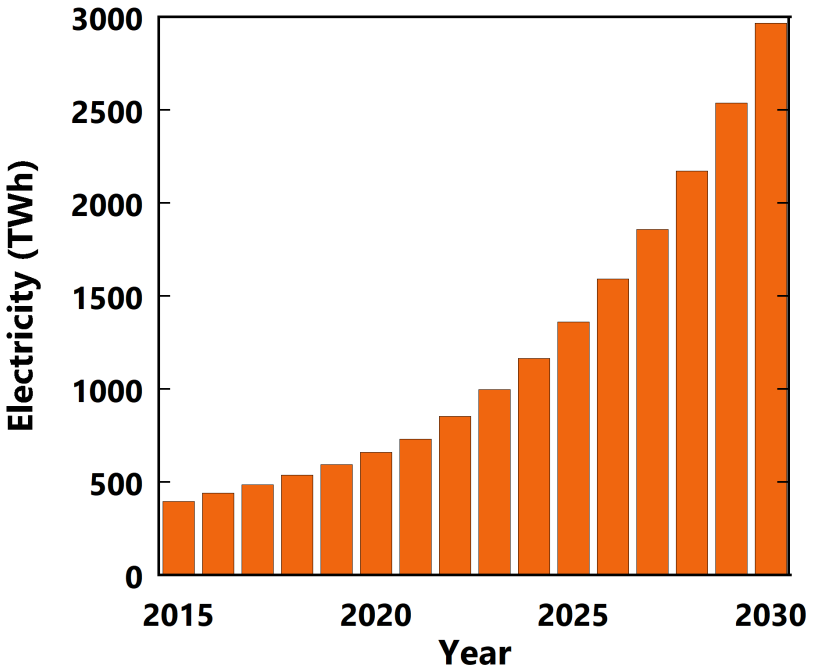


Figure 1.2: Trend of world-wide datacenter electricity usage [7].

It is expected to increase significantly to 3,000 TWh by 2030 due to the explosive growth of generated data [7]. Although current consumption is less than predicted, this amount of electricity dramatically exceeds the total amount of electricity generated in Japan. From preventing global warming point of view, it is essential to use renewable energy and reduce electricity consumption. Servers consume about 74% of the total power consumption in datacenters, and about 73% of the power consumption in servers is consumed by the CPU (Central Processing Unit) [8], as shown in Figure 1.3. Since these CPUs consume a large amount of power and generate heat, large-scale cooling systems are equipped in datacenters, accounting for nearly 13% of the total power consumption of datacenters.

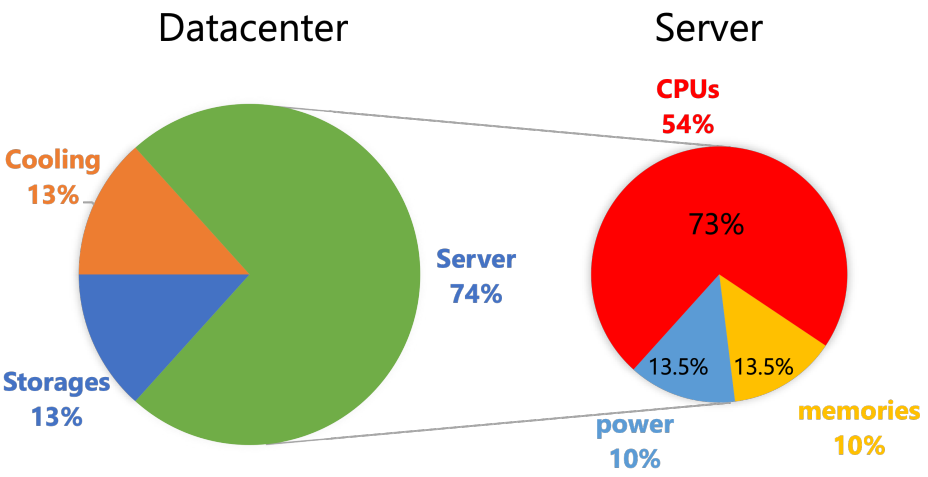


Figure 1.3: Breakdown of electricity usage for both datacenter and server [8].

A breakdown of the power consumed by the CPU is shown in Figure 1.4. Only 33% of the power is used for computation itself, while two-thirds is spent on accessing the cache SRAM (Static Random Access Memory) and main memory, DRAM (Dynamic Random Access Memory), which means "moving data for computation." In particular, machine learning, which has recently been the focus of much research, handles enormous amounts of data, and the frequent transfer of data from main memory has become a bottleneck in

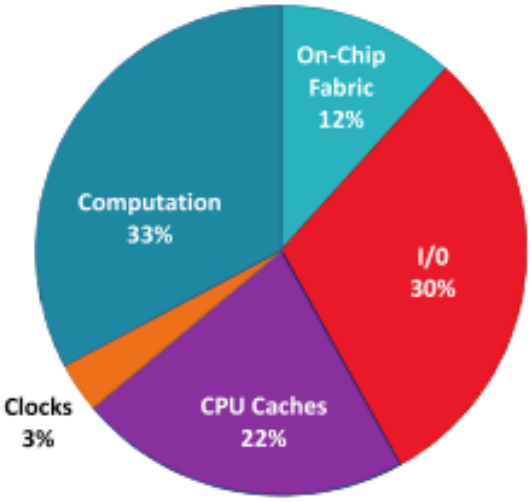


Figure 1.4: CPU power breakdown [8].

computing capability and power consumption, known as a "memory wall" [9–11].

In recent years, to improve this problem, stacked DRAM modules using TSV (Through Silicon Via) technology, so-called HBM (High Bandwidth Memory) [12], are directly attached to the processor. However, since DRAMs have a data retention time of less than one second, they require frequent refresh operations to overwrite the data, and as the capacity of DRAMs increases, the delay caused by these refresh operations becomes non-negligible, as shown in Figure 1.5 [13]. There have been attempts to increase the size of the SRAM on the CPU in order to reduce the data movement [14]. In this example, as shown in Figure 1.6, eight processor cores share a total of 96 MB of SRAM by using 7 nm node technology [15, 16], and it has been reported that this stacked SRAM is suitable for energy-efficient accelerators [17]. However, the TPD (Thermal Design Power) of this processor is still quite large at 105 W. Parallel processing is being done with more processor cores using GPU (Graphics Processing Unit), but it does not replace the traditional von-Neumann architecture, and dramatic power reductions are difficult to achieve.

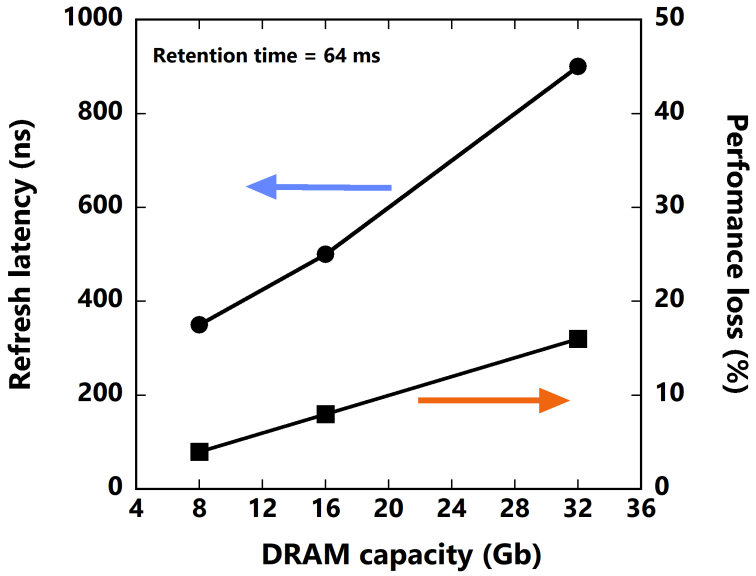


Figure 1.5: Trade-off between DRAM capacity and refresh latency (left axis) and percentage of performance loss (right axis) [13].

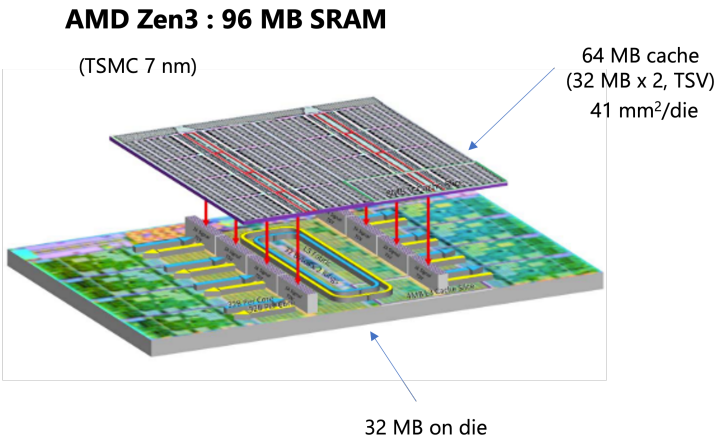


Figure 1.6: High-capacity SRAM stacked on the processor [16].

1.2 Advantage of embedding high-capacity SRAM into 3D NAND

In recent years, new computational methods utilizing artificial intelligence (AI) that use non von Neumann architecture are attracting much attention. This approach is being considered to efficiently process data at the edge side, rather than doing it on the conventional servers in the cloud, as shown in Figure 1.7. An architecture of computing in memory (CiM) or in-memory computing has also been attracting attention that can handle huge data efficiently. Since SRAMs have no endurance limit, studies are

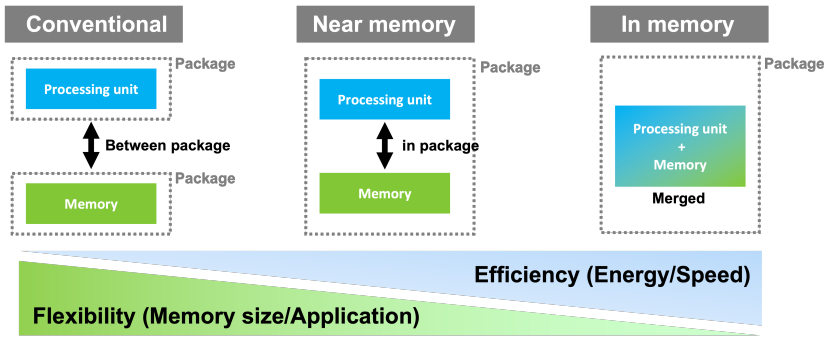


Figure 1.7: Trade-off between memory size flexibility and energy efficiency of computing. In-memory computing is suitable for energy efficient computing [40].© 2019 IEEE

being conducted using not only conventional 6T (six transistors) SRAMs but also 8T (eight transistors) and 9T (nine transistors) SRAMs in terms of energy efficiency [18–20].

On the other hand, various CiMs using non-volatile memory that can store large amounts of data are also being considered [21, 22]. There are also many studies on utilizing NAND flash memory for in-memory computing [23–30]. NAND flash memory is already a proven product and has the potential to realize high-capacity, low-cost CiM. One of the studies showed that a combination of 6.5 Gb (gigabit) of SLC-NAND (single-bit per cell NAND) and 3.2 MB (megabyte) of SRAM with 16 core system is effective [26]. Their study assumes that a CMOS (Complementary Metal Oxide Semiconductor) chip and 3D NAND (three-dimensional cell NAND

flash memory) are fabricated separately, and both SRAM and 3D NAND dies are bonded together, which causes process complexity and manufacturing cost increase.

In the early days of 3D NAND products, it was fabricated by first forming CMOS transistors and then NAND memory cells. As shown in Figure 1.8, after the CMOS transistors were formed, the NAND memory cell arrays were formed beside it. This method is called CMOS Next to cell Array (CNA) process. The source lines of the NAND cell arrays are formed below the memory cells by a process that etches a trench in the Si (silicon) substrate and fills it with electrodes [31, 32]. This meant that the area of the CMOS circuit was small, and it was impossible to embed high-capacity SRAMs.

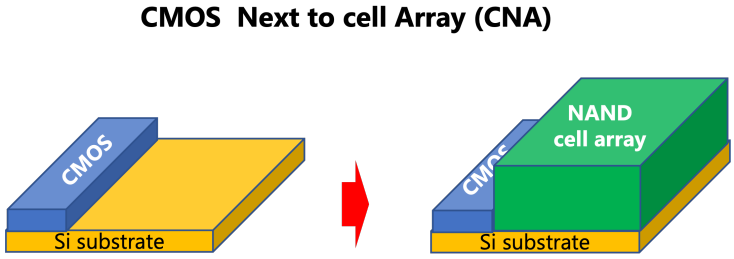


Figure 1.8: 3D NAND fabricated by CNA (CMOS Next to cell Array) process.

3D NAND has increased its capacity by increasing the number of memory layers in each generation. However, as memory capacity increases, the peripheral circuits (decoders and sense amplifiers) used to access the memory cells also increase in size. The conventional CNA process could no longer accommodate the peripheral circuits increase, so a CMOS circuit was proposed to be placed under the NAND cell arrays. This is called the CUA (CMOS Under cell Array) process, as shown in Figure 1.9. This process accommodates more CMOS circuits in the die [33, 34].

As 3D NAND capacity continues to increase, it will again become more difficult to accommodate CMOS circuits. In the future, CMOS circuits may be stacked like NAND cells. One such proposal is to create CMOS circuits and NAND cell arrays separately and finally stack them together to form one chip, as shown in Figure 1.10 [35, 36]. In this process, the CMOS circuit is

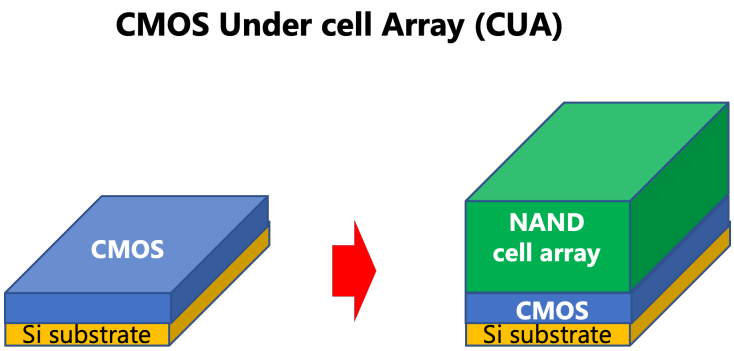


Figure 1.9: 3D NAND fabricated by CUA (CMOS Under cell Array) process.

not affected by the thermal process of NAND memory cell formation, thus increasing the choice of technology to be applied. However, this process only allows stacking two chips. Also, the manufacturing cost increases by fabricating separate chips. Since the bonding is done on the wafer rather than chip by chip, there are many issues that need to be addressed in order to achieve mass production, such as yield degradation. For these reasons, the CUA structure is suitable as the process for NAND with high-capacity SRAM at this moment.

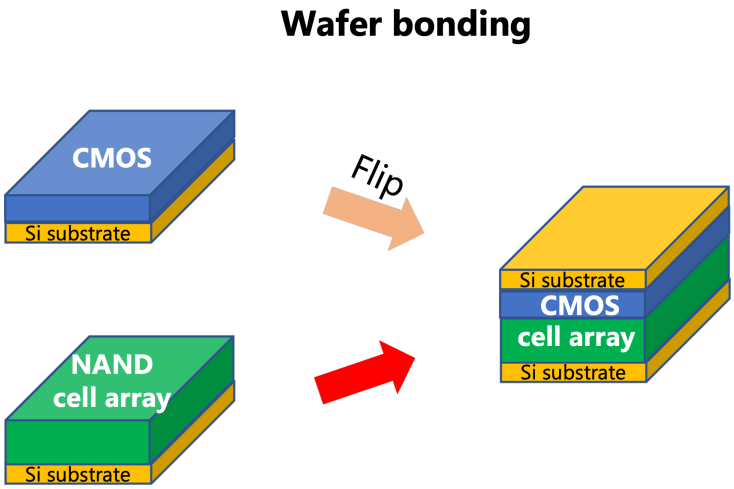


Figure 1.10: 3D NAND fabricated by wafer bonding process.

Another advantage of embedding high-capacity SRAM into 3D NAND is the replacement of DRAM used in the SSD (Solid-State Drive). Figure 1.11

shows the SSD image and components inside SSD used for a computing system. Typically, the SSD housing contains NAND flash memory according to its installed capacity, a controller that controls reading and writing, and DRAM as a buffer memory for exchanging data with the outside world at high speed. In a single NAND flash memory package, eight to sixteen NAND memory dies are stacked. Normally, the capacity of DRAM used as a cache is about one-thousandth of the SSD capacity. There was a study for embedding DRAM in NAND flash memory [37]. However, it was a planar capacitor DRAM and could not replace off-chip DRAM. If high-capacity SRAM could be embedded in 3D NAND, DRAM would no longer be needed, and the result would be a smaller, faster, and less expensive SSD.

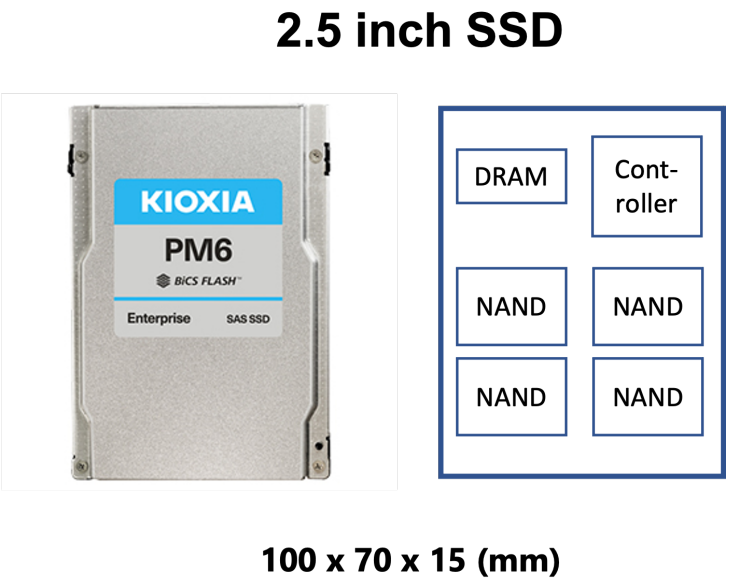


Figure 1.11: Photo of enterprise SSD drive [38] © KIOXIA and image of components inside SSD.

One approach is stacking NAND, SRAM, and controller by using TSV technology [39]. The other is embedding high-capacity SRAM in 3D NAND, as shown in Figure 1.12. If it is possible to create a process for integrating high-capacity SRAM that is compatible with the 3D NAND manufacturing process, it will be possible to significantly reduce the size of SSD, as shown in Figure 1.13.

In the future, together with the technology for stacking controller die with high-capacity SRAM embedded in 3D NAND, it will be possible to integrate server functions into a single package [40]. This will allow for smaller, more power-efficient servers and improved performance through massively parallel processing.

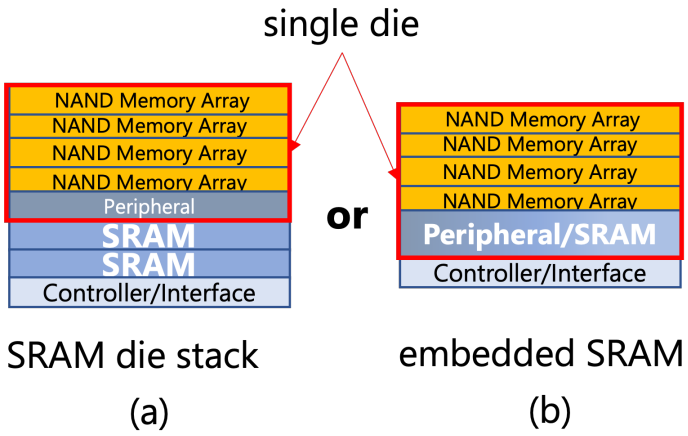


Figure 1.12: Image of two approaches to realize one package SSD. One is stacking 3D NAND, high-capacity SRAM and controller (a). The other is stacking SRAM embedded NAND and controller (b).

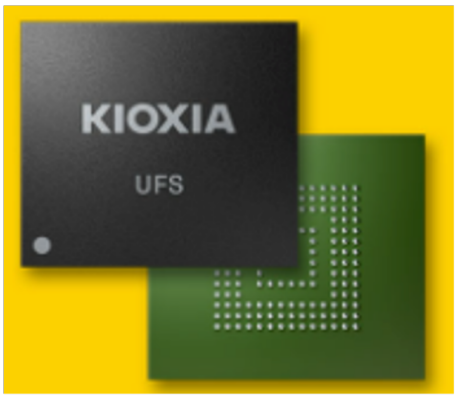


Figure 1.13: Image of one package SSD drive [38] © KIOXIA.

1.3 Process comparison between SRAM and 3D NAND

There are some differences in the fabrication process between embedded SRAM with the logic process, stand-alone SRAM, and 3D NAND. In comparing the logic process used for embedded SRAM, stand-alone SRAM process, and NAND processes, it is important to know which generation to compare. The required technology generation depends on the required CMOS transistor performance. Embedded SRAM is manufactured using logic processes and design rules. In state-of-the-art logic, CMOS transistors are formed using a combination of FinFET (Fin Filed-Effect Transistor) and high-κ/metal gate (HK/MG) [41]. On the other hand, stand-alone SRAMs are designed for high capacity while maintaining the high speed, and some special processes and design rules are used to keep manufacturing costs as low as possible.

NAND flash memory is mainly designed to reduce the cost per bit and is inferior in speed compared to DRAM. That is why the bit price of NAND flash is about 1/40th of that of DRAM (Table 1.1 [42]).

Table 1.1: Price per giga-bit of both DRAM and NAND [42].

Memory	Capacity	*Price ($)	Gb price($)
DRAM	16Gb DDR4	$7.5	$0.47
3D NAND	256Gb TLC	$3.07	$0.012

1/40

*Mar. 25th, 2022

The process is built with priority on the memory cells, and the CMOS transistors are formed in the process that forms the memory cells. Therefore, materials and transistor structures similar to those used in advanced logic are not applicable. Until now, it has been difficult to apply advanced logic processes to DRAM and NAND flash memory because the memory cells are

formed after CMOS transistor fabrication, and the process temperature for forming memory cells is more than 600°C, which is much higher than the conventional logic process. However, interface speeds have been increasing in recent years due to the demand for high-speed data processing, which in turn requires higher performance CMOS transistors. Therefore, DRAM is shifting to adopt HK/MG to improve CMOS performance. The interface speed of NAND flash memory follows DRAM as shown in Figure 1.14, and therefore the performance and process requirements for CMOS transistors are also following DRAM.

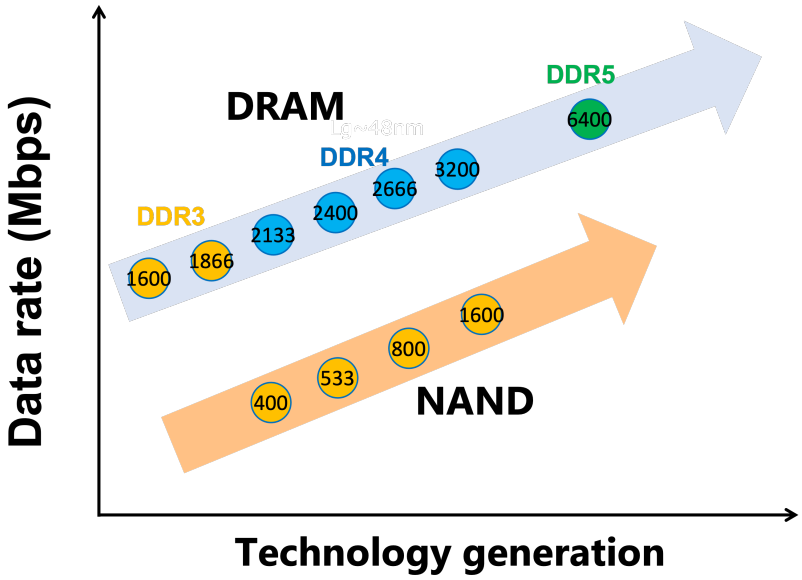


Figure 1.14: Interface speed trend of both DRAM and NAND.

The current CMOS generation used for 3D NAND is about 130 to 90 nm node technology. Therefore, as a starting point, it is beneficial to compare SRAM and 3D NAND processes based on the 90 nm generation CMOS technology and aims to extract the issues involved in the final integration of high-capacity SRAM into NAND flash memory.

1.3.1 Embedded SRAM process

Figure 1.15 shows a typical process flow of 90 nm node logic that is used for embedded SRAM fabrication [43, 44]. The process is divided into three parts. FEOL (Front End Of Line) process covers from STI (Shallow Trench Isolation) to before contact formation. MOL (Middle Of Line) process covers after silicide process to before first metal formation. BOEL (Back End Of Line) process covers the first interconnect layer and beyond. After gate electrode patterning, source and drain were formed, followed by a silicidation process. Cobalt silicide ($CoSi_2$) [45–49] was formed in both the active area and the gate poly-Si. Copper (Cu) is used for all the interconnect layers. Low-κ intermetal dielectric film was applied to reduce intermetal capacitance.

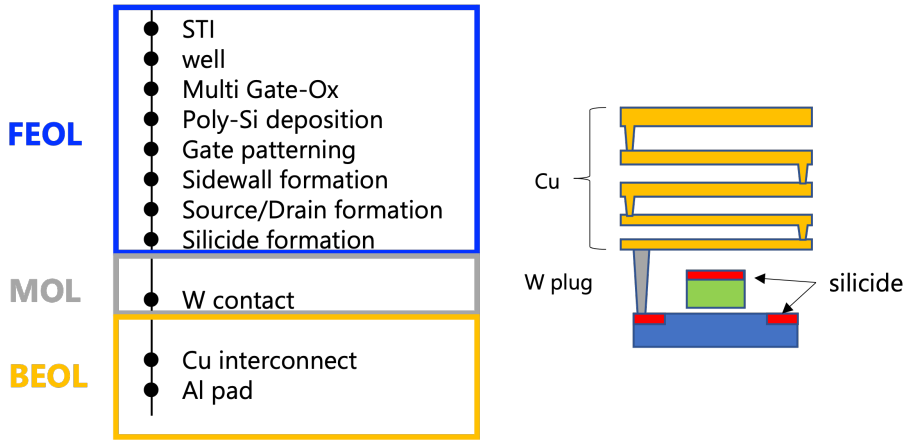


Figure 1.15: Process flow and schematic image for embedded SRAM.

1.3.2 Stand-alone SRAM process

The process flow of stand-alone SRAM is shown in Figure 1.16. The FEOL process is the same as that of embedded SRAM fabrication. The differences in both MOL and BEOL. The process difference is that the tungsten (W) used for contact formation is also used for short-distance wiring. As mentioned earlier, stand-alone SRAM requires not only performance but also cost reduction. Therefore, wiring is formed using tungsten used in contact embedding for short-distance wiring, the so-called "Local Interconnect (L.I.)" [50–52], and which does not affect performance. The details of the process are described in the MOL chapter, but by using L.I., it is possible to reduce the number of interconnect layers that form the SRAM cell by one layer. BEOL also differs slightly from the logic process. In high-capacity stand-alone SRAM, where low cost is required, aluminum (Al) interconnect is used for the bottom two layers. The upper two layers use copper interconnections for higher speed and power supply enhancement.

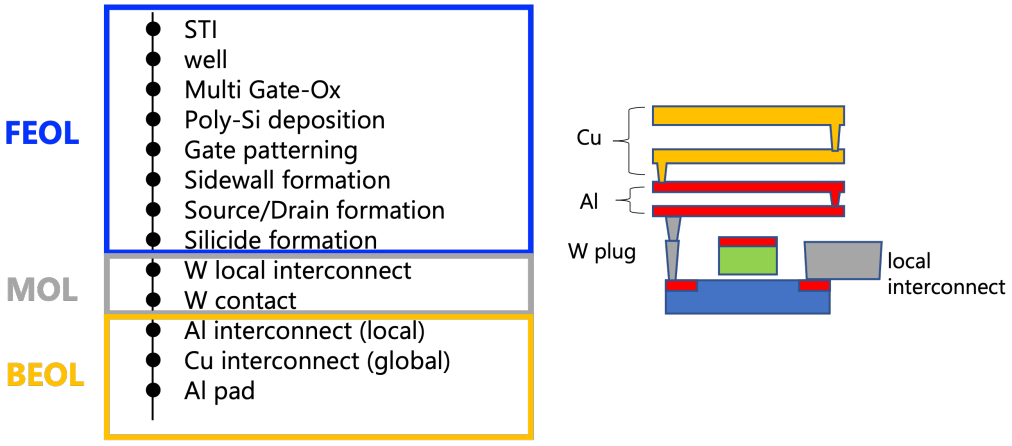


Figure 1.16: Process flow and schematic image for stand-alone SRAM.

1.3.3 3D NAND process

The process of 3D NAND is very different from that of the SRAM process, each in FEOL, MOL, and BEOL, as shown in Figure 1.17. First, FEOL requires a high voltage of nearly 30 V for programming NAND. Hence, an additional gate-oxide film for high-voltage transistors that is more than twice as thick as the gate oxide film for I/O (Input/Output) devices typically used in logic is necessary. Also, the gate electrode is not made of silicide but rather a polymetal gate made of polycrystalline silicon with tungsten layers. Furthermore, no silicide is formed in the source and drain diffusion layer regions at this step. The minimum gate length is also longer than that of conventional logic or SRAM because impurities in the source and drain diffuse during the high-temperature thermal process. After opening the contact hole, additional ion implantation to reduce contact resistance is carried out for the MOL process, followed by an activation anneal process. After that, cobalt is deposited in the contact hole to make $CoSi_2$ followed by a tungsten filling process [53]. For the BEOL process, since NAND cell formation requires a high-temperature thermal process of more than 600°C, multiple layers of tungsten, which has high thermal resistance, are used as interconnect instead of aluminum to connect CMOS devices. After NAND cell formation, copper interconnect for bit lines is formed. Due to the high-temperature thermal process after contact formation, the latest logic process, such as the gate-last HK/MG process, cannot be applied to CMOS. Low-κ interlayer dielectric (ILD) film also cannot be applied to interconnect layers before the NAND cell formation.

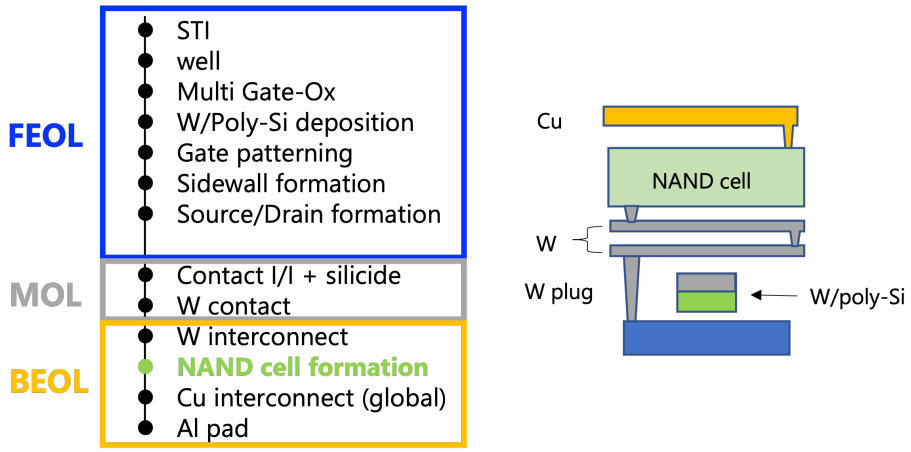


Figure 1.17: Process flow and schematic image for 3D NAND.

1.3.4 Key process modules

As analyzed in previous sections, the processes can be divided into three areas: FEOL, MOL, and BEOL. The same process is basically used for both embedded SRAM and stand-alone SRAM, but in order to reduce the manufacturing cost, stand-alone SRAM uses tungsten for local (short distance) wiring to reduce the number of interconnect layers after contact formation. In some cases, Al interconnect is used instead of Cu interconnect for lower-layer interconnections when high-speed performance is not required. In contrast, 3D NAND has a different gate dielectric and gate electrode structure, and polymetal gate electrodes are used instead of silicide. In addition, tungsten, which has high thermal resistance, is basically used for CMOS device connections, and usage of Cu interconnect kept to a minimum. Table 1.2 summarizes differences of process modules among three devices, and their impact on embedding high-capacity SRAM in the 3D NAND process should be studied. Process modules discussed in this thesis are indicated in Figure 1.18.

Table 1.2: Process modules and differences among three devices.

Module	Process	Embedded SRAM	Stand alone SRAM	3D NAND
FEOL	Gate insulator	SiON/high-κ	SiON/high-κ	SiON
	Gate electrode	Silicide/Metal	Silicide/Metal	Polymetal
MOL	Local interconnect	not available	W	not available
BEOL	Interconnect	Cu/Low-k	Al Cu/Low-k	W Al Cu
MOL/BEOL	Temperature after contact	< 400℃	< 400℃	600℃ <

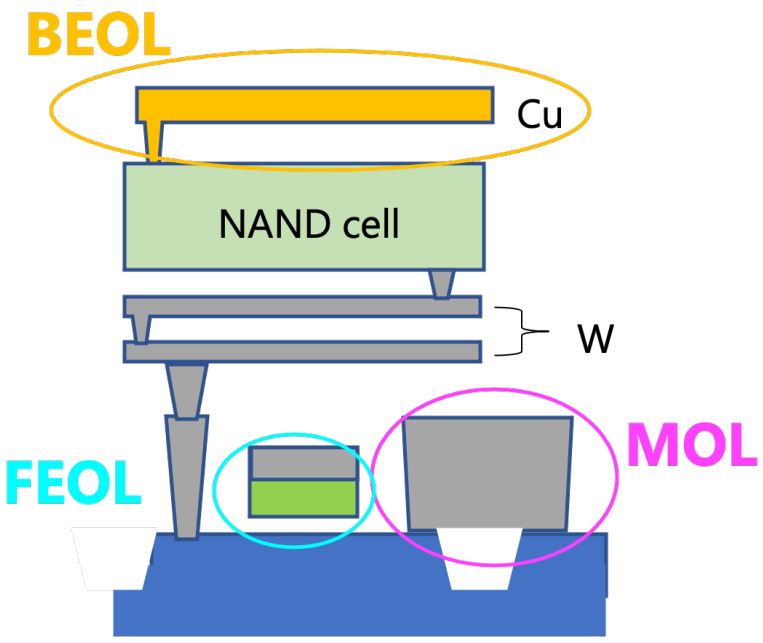


Figure 1.18: Schematic image of each process module.

1.4 Thesis focus and organization

As mentioned in the previous section, there are advantages to realizing a small and power-efficient computing system for the advanced information society. Computing-in-Memory is a promising technology to satisfy this requirement, and a 3D NAND-based system with high-capacity SRAM can play an important role. The critical manufacturing processes are identified in FEOL, MOL, and BEOL, respectively, to embed high-capacity SRAM into 3D NAND. This thesis discusses the process technology challenges and implementation methods required for integrating high-capacity SRAM with 3D NAND memory with scalability, as 90 nm node technology is a starting point. The scaling scenario until moving to 3D stacking is discussed.

Figure 1.19 shows the organization of this thesis.

After introductory discussions in Chapter 1, Chapter 2 discusses the cell layout optimization methodology to realize high-reliable high-capacity SRAM. The importance of mechanical stress control, which is caused by manufacturing processes and device structure, is discussed.

In Chapter 3, the reliability of the CMOS transistor used in the SRAM cell is analyzed. The minimum dimension MOS transistors show enhanced degradation. The mechanism and design guidelines are proposed to realize high-reliable high-capacity SRAM embedded in 3D NAND.

Chapter 4 describes FEOL and MOL process technologies that enable high-capacity SRAM in 3D NAND. Gate-first HK/MG process options for future scaling and impact of local interconnect with process guidelines are presented.

Chapter 5 proposes the BEOL processes to realize high-reliable high-capacity SRAM. Low-κ interlayer dielectric film options for future multi-layer copper interconnect, and redundancy fuse processes are discussed.

Chapter 6 discusses the remaining technical issues and proposes viable solutions for 3D integration which realize future single-package computing systems.

The conclusion of this thesis is shown in Chapter 7.

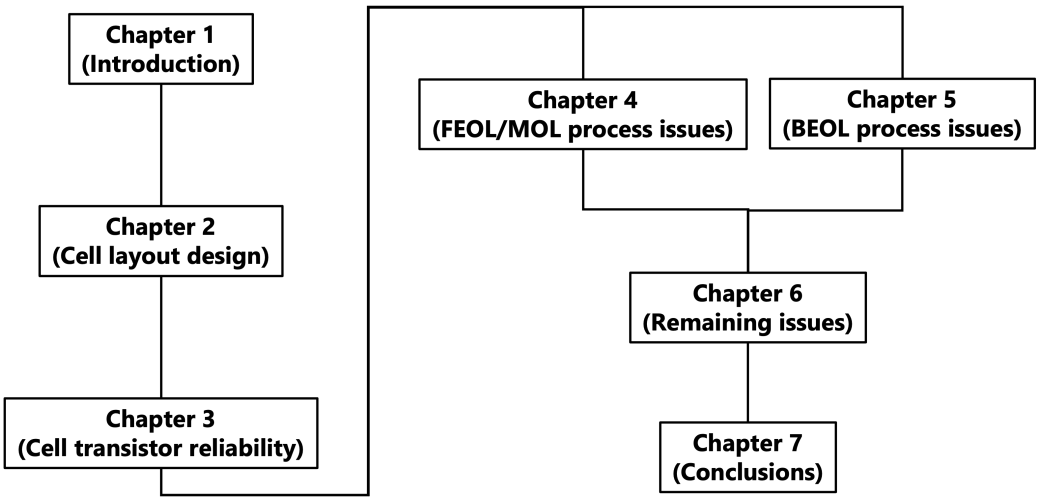


Figure 1.19: Thesis organization.

Chapter 2

Layout and transistor structure design for high-reliability SRAM

2.1 Background

This chapter aims to study a suitable layout for SRAM cells, considering that they are embedded in 3D NAND. As mentioned in the previous chapter, the CMOS process used for 3D NAND is very different from the process used for embedded or stand-alone SRAM. Various cell layouts for SRAM have been studied in the past. Figure 2.1 compares cell designs for high-speed SRAM at the 180 nm generation by M. Ishida *et al.* [54]. They showed that the Type-1b cell, in which the active area (AA) is arranged in a straight line and the gate electrode (GE) is also placed in parallel, is superior in terms of scalability and speed by shorter bit-line length compared to the Type-1a cell (referred to as conventional cell). However, no concrete studies have been conducted on Type-4 cells (referred to as thin cells).

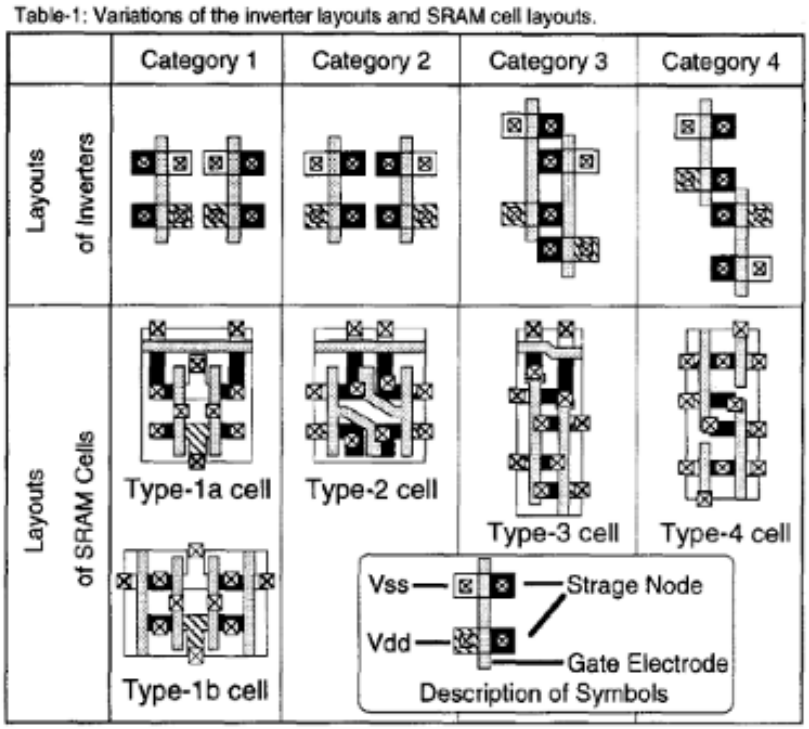


Figure 2.1: SRAM cell layouts study by M. Ishida *et al.* [54].© 1998 IEEE

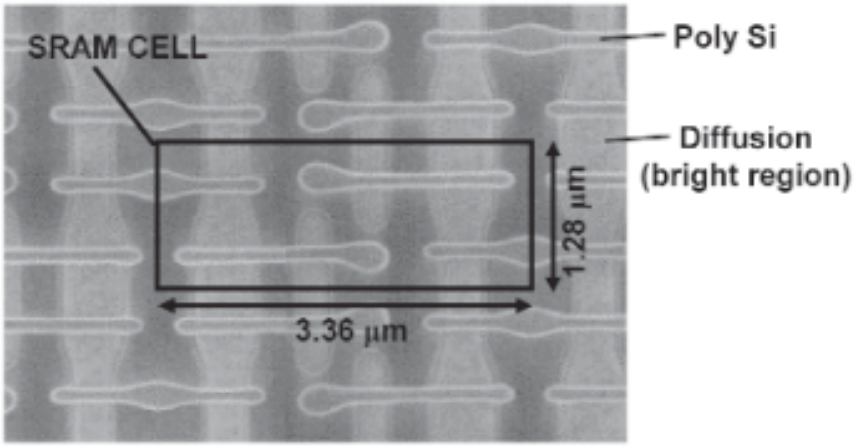


Figure 2.2: The first demonstration of SRAM cell used the Type-4 layout in Figure 2.1 by K. Osada *et al.* [55].© 2001 IEEE

In 2001, K. Osada *et al.* demonstrated the results of the first prototype of a 180 nm generation 32 kb high-speed cache SRAM using the thin cell, as shown in Figure2.2 [55]. While the bit-line (BL) length can be further shortened compared to Type-1b, the word-line (WL) length becomes longer, and the WL decoder circuit needs to be redesigned due to the lower cell height, resulting in the decoder circuit extending in the horizontal direction with chip size increase. However, conventional cells (Type-1a and Type-2) have still been used for high-capacity SRAM cells [56–58] and embedded applications at 90 nm generation, as shown in Figure 2.3 [59]. Therefore, it is important to select an SRAM cell layout with excellent scalability and verify that there are no characteristic problems with that SRAM cell. The following sections discuss the optimum SRAM cell layout to embed in the 3D NAND.

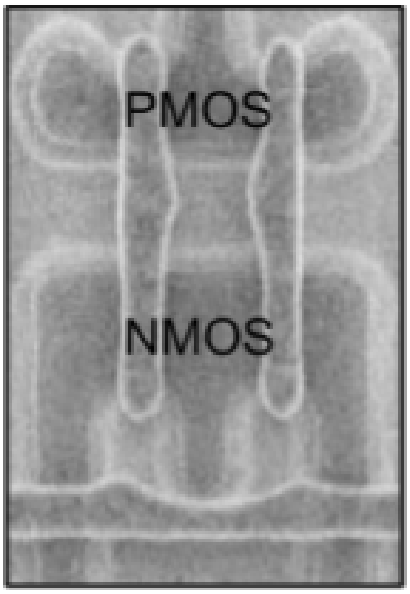


Figure 2.3: Example of Type-1a SRAM cell layout used for 90 nm node logic product [59]. The cell size is 1 μm^2.© 2002 IEEE

2.2 Selection of SRAM cell layout

2.2.1 Cell layout comparison

As mentioned in Chapter 1, the transistor technology used in 3D NAND is about 130 to 90 nm nodes, so it is important to consider scalability starting from the 90 nm node design rule. Since two types of SRAM cell layouts are proposed in 90 nm nodes, it is important to evaluate and decide on the scalable SRAM cell layout considering the extendibility. The advantages and disadvantages of both conventional and thin cells are summarized in Table 2.1. For a conventional cell (a), the aspect ratio of the unit cell is closed to one, which is easy to design large cell arrays with peripheral circuits. However, active areas have bending patterns that cause rounding shapes at corners resulting in dimension variation. The orthogonal layout of gate electrodes also needs attention to prevent pattern short by lithography process fluctuation. In contrast, thin cells are composed only of straight patterns. Therefore, the controllability of dimensions is considered superior to that of the conventional cell. The lower cell height shortens

Table 2.1: Advantages and disadvantages of both conventional and thin cells.

Cell type	Advantage	Disadvantage
Conventional (a)	Peripheral design cell spect ~ 1	Bending active area Orthogonal gate electorde
Thin (b)	Simple (straight) pattern Shorter bit line	Longer WL WL decorder congestion Chip size impact

the bit-line length, which is advantageous for high-speed operation but has the detrimental effect of increasing the word-line length. Furthermore, the height of the word-line decoder circuit must also be lowered as the cell height becomes lower, resulting in the cell array, including peripheral circuits becoming longer in the horizontal direction. This raises another concern that the lateral direction of the die may not fit into the package for high-capacity SRAM. Figure 2.4 shows SRAM cell layouts for evaluation, (a) conventional type layout used most of SRAM products down to 90 nm node and (b) thin type cell layout, which is composed of straight patterns. From the past technological trends [60], the cell size of 1 μm^2 is set as a target. Both cells satisfy the target cell size using the design rules summarized in Table 2.2. The gate electrodes (GEs) of conventional cell layouts are arranged orthogonally. As the cell size is scaled down, the space between WL and pulldown MOSFETs becomes narrow. The resist profile of WL is affected by the pulldown resist pattern, and the resist profile of WL becomes wider under the influence of the pulldown pattern, as shown in Figure 2.5 (b). If the distance is further reduced, in the worst-case scenario, the resist patterns of both WL and pulldowns are connected to each other, as shown in Figure 2.5 (c). From a manufacturability point of view, straight patterns for both AAs and GEs are preferable because it is easy to control dimensions with small size variations. Rectangular patterns tend to shrink in the longitudinal direction, which is usually suppressed by the optical proximity correction (OPC) [61] by adding patterns on the masks that are too small to be resolved on the wafer by the lithography process. Since those additional patterns do not affect each

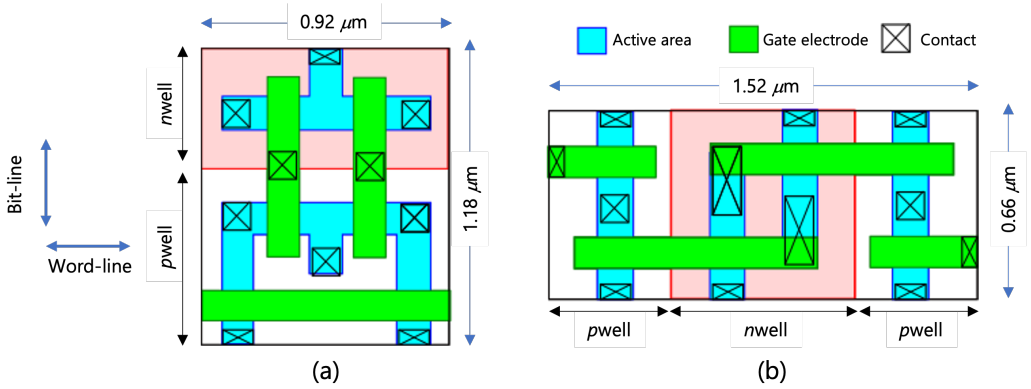


Figure 2.4: SRAM cell layouts with dimensions used for the scalability study. (a) Conventional cell layout used until 90 nm node and (b) low height (thin) cell that consists of all straight patterns.

other when placed in the longitudinal direction, pattern connection does not occur even if the spacing is narrow, as shown in Figure 2.6.

From the above measurement results, OPC models are calibrated, and WL gate length variability is simulated to investigate the process robustness under various lithography conditions. Figure 2.7 shows simulated results of WL gate length variation for both conventional and thin cell layouts. The figure shows that the thin cell shows tighter variation than the conventional cell layout. Therefore, a thin cell layout should be chosen as the SRAM cell layout for 90 nm nodes and beyond in consideration of scalability. Both AA and GE patterns are optimized by using an OPC combined with a manufacturability check [62] and a hotspot fixing system [63]. The area where AA, GE, and contacts are placed simultaneously, where contact connects both AA and GE, AA and GE is designed to have sufficient overlap to avoid open failure by misalignments. To confirm that this selection was correct, an SRAM test vehicle using both cell layouts was fabricated as a next step.

Table 2.2: Design rules used for 90 nm node SRAM cell.

Layers	Line (nm)	Space (nm)
Active area	120	140
Gate electrode	80	190
Contact	120	140
Contact - Gate	-	70
	Conventional	Low height
Cell size (μm^2)	1.0032	1.0856

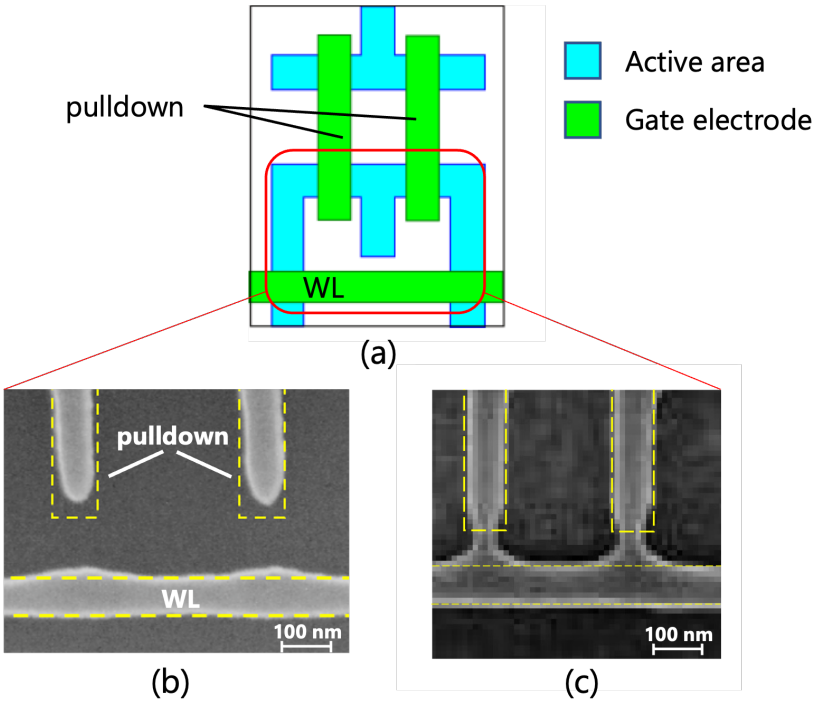


Figure 2.5: (a) Schematic layout of conventional cell, and (b) SEM photograph of the WL and pulldown gate resist pattern. (c) Resist pattern connection occurs due to the narrow distance between pulldown GEs and WL.

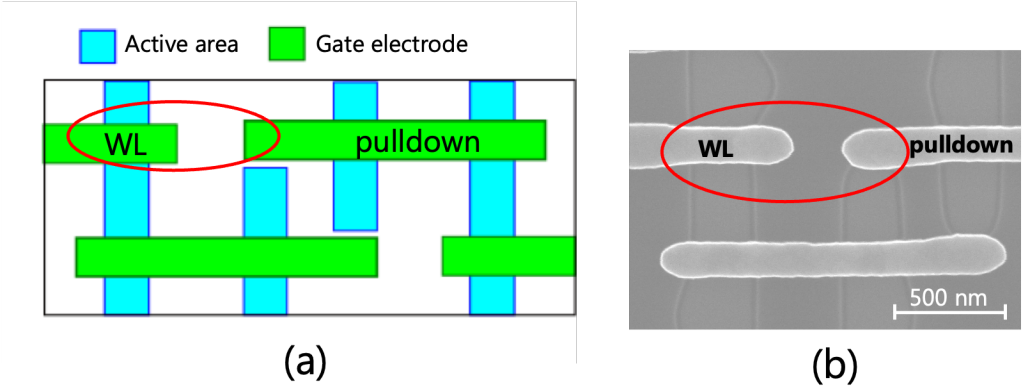


Figure 2.6: (a) Schematic image of cell layout, and (b) corresponding SEM photograph of the WL and pulldown gate patterns after gate patterning process in a thin cell. Good gate length uniformities without pattern connection were observed.

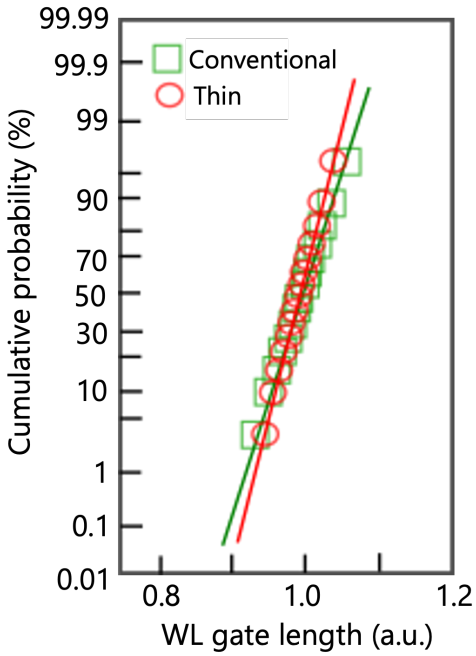


Figure 2.7: Simulated cumulative probability of WL gate length for both conventional and thin types SRAM cells.

2.2.2 Validation by the hardware

A 1 Mb SRAM test vehicle with conventional and thin cell layouts was fabricated and measured. In functional tests, only SRAM with thin cells showed single-bit failures. To investigate the root cause, failure analysis was carried out, and missing active areas in the *p*MOSFET region, which originated from crystal defects, were observed, as shown in Figure 2.8. One of the major causes of crystal defect generation is the damage caused by the high-dose ion implantation process [66, 67]. There are also several studies regarding SRAM bit failure caused by the mechanical stress with STI structures [64, 65]. However, those phenomena were observed in the *n*MOSFETs, and their models cannot explain the defects observed in the *p*MOSFET region. Furthermore, previous simulation studies for mechanical stress analysis used 2D (two-dimensional) models and could not accurately analyze the phenomenon of actual structures. Therefore, a 3D (three-dimensional) simulation of the mechanical stress by using the finite element simulation tool MARC [68] was carried out for to investigate the origin of this failure.

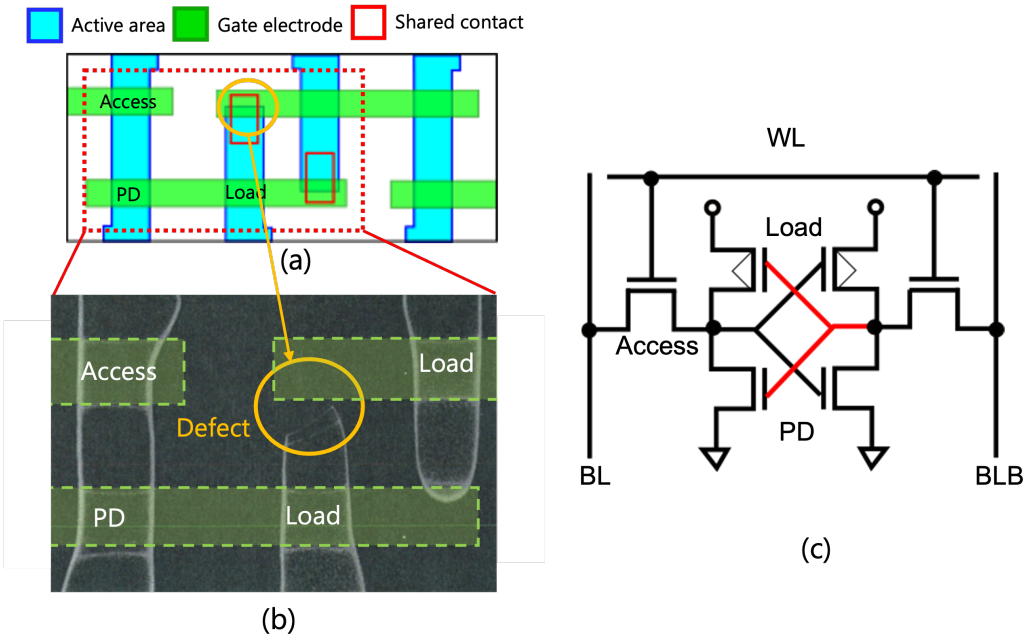


Figure 2.8: (a) SRAM cell layout and (b) corresponding SEM photograph after Wright etching where single-bit failure were observed. The active area edge of *p*MOSFET load transistor was missing where shared contact connected both GE and AA, indicated as red lines in SRAM cell circuit shown in (c).

2.3 Process flow and simulation setup

A process flow of fabricated SRAM is shown in Figure 2.9 with process steps that may generate crystal defects which are labeled through a to g. Parameters used for the simulation, such as Young's modulus and Poisson's ratio of Si, SiO_2, and SiN (silicon nitride), are summarized in Table 2.3.

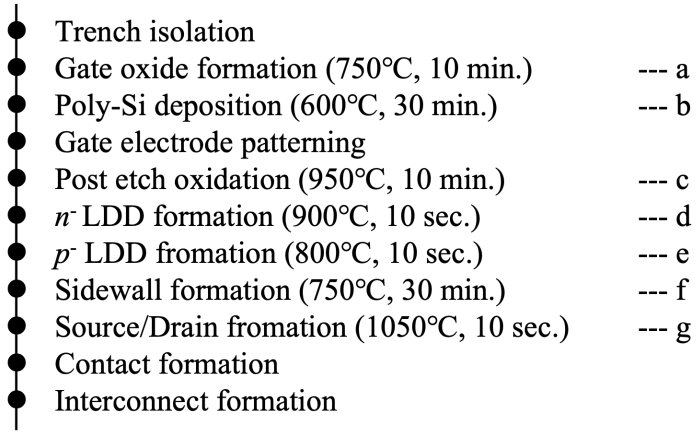


Figure 2.9: Process flow of fabricated SRAM. The process steps, a through g, are candidates that may generate crystal defects.

Table 2.3: Material parameters used for the simulation.

Parameters	Si	SiO_2	SiN
Young's modulus (GPa)	130	75	300
Poisson's ratio	0.2	0.25	0.28

Shear stresses were monitored at the top corner of the active area, AA, under the gate electrode, GE, as shown in Figure 2.10. The stress values are proportional to the process temperatures. Since the source/drain formation process showed the largest stress value, further stress analyses were carried out focusing on this process.

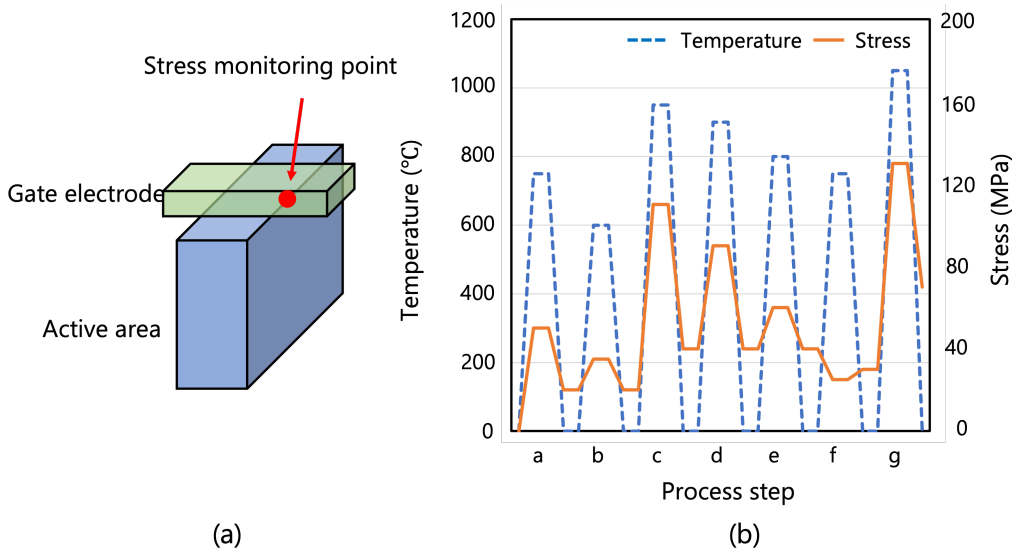


Figure 2.10: (a) Monitoring location of mechanical stress and (b) simulated stress values at each process steps, a - g, in Figure 2.9 with the temperature profile.

It is well known that structures used for simulation, mesh shape, and a number of meshes affect simulation results. Even using the OPC technique, fabricated devices have some deformations. Reflecting actual device shapes with complicated mesh structures cause simulation time to increase and does not converge in the worst case. Before moving to the detailed analysis, the effects of structure fidelity on the simulation results were evaluated. The top corner of the AA is rounded to prevent hump characteristics in MOSFETs [69]. First of all, von Mises stress values with and without 20 nm rounding shape at the top corner of STI were compared. Other dimensions are shown in Figure 2.11 (a).

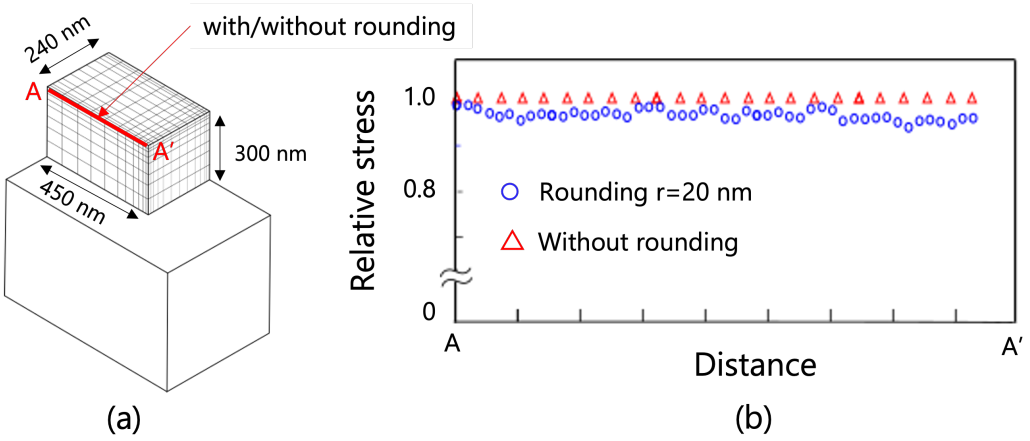


Figure 2.11: (a) Structure used for the simulation. Only AA portion was shown. (b) Simulated results of von Mises stress with and without top corner rounding. The stress values were normalized by the stress without top corner roundings.

Figure 2.11 (b) shows the simulated results where stress values were normalized by the stress without top corner roundings. The stress values with rounded top corners show slightly small stress values with some fluctuations due to mesh shapes. However, the stress differences with, and without rounded top corners were less than 5%. These differences are acceptable for comparing stress along the channel, and without rounding top corners show larger stress values.

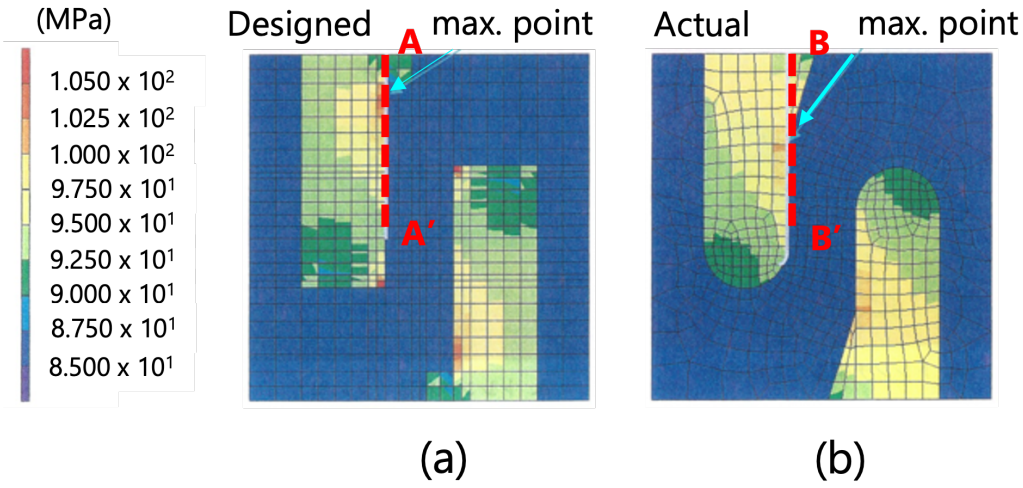


Figure 2.12: Contour plots of simulated von Mises stress of (a) designed layout and (b) actual device shape.

Figure 2.12 shows the contour plots of the simulation results of von Mises stress for each of (a) the designed layout and (b) manufactured shapes (shown in Figure 2.6 (b)). The number of mesh elements is kept as close as possible to obtain fair results, where 572 for the designed layout and 582 for the actual device. Both structures show similar stress distributions, and the maximum stress value is observed near the concave corner. The stress values along the active area indicated with the dotted lines A−A' and B−B' are shown in Figure 2.13. The differences between the designed structure and actual shape are less than 5% and acceptable to use the designed structure for further studies. Based on the above results, designed layouts without rounded top corners are used for the following simulation.

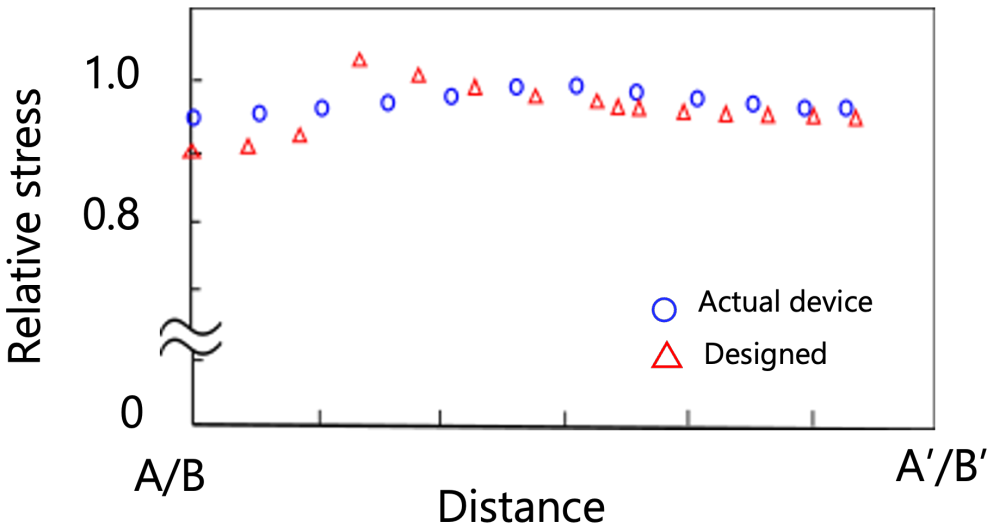


Figure 2.13: Relative stress values along the active area, indicated as dotted lines A−A' and B−B' in the Figure 2.12.

2.4 Simulation results

2.4.1 Mechanical stress and defect generation model

The results of the stress simulation in the cell structure where crystal defects were observed are shown in Figure 2.14. The colors represent the maximum resolved shear stress during the source/drain annealing process, where the maximum stress value was observed in the preliminary simulation. The displacements by the mechanical stress are multiplied 200 times for easy observation. The top corner regions of the AA show stress values of more than 300 MPa, which is large enough to generate crystal defects [65]. Moreover, the AA edge (labeled A') shows large displacements from the top surface to the bottom of the trench region. This region overlaps with the gate electrode, GE2, which indicates that mechanical stress by GE2 may be a root cause for this displacement. Figure 2.15 shows the displacement amounts of the AA under the gate electrode GE1 and GE2 along with the direction indicated as A−A' in Figure 2.14. The SRAM cell is rotated 90 degrees for easy understanding of displacement along the AA direction shown in Figure 2.15

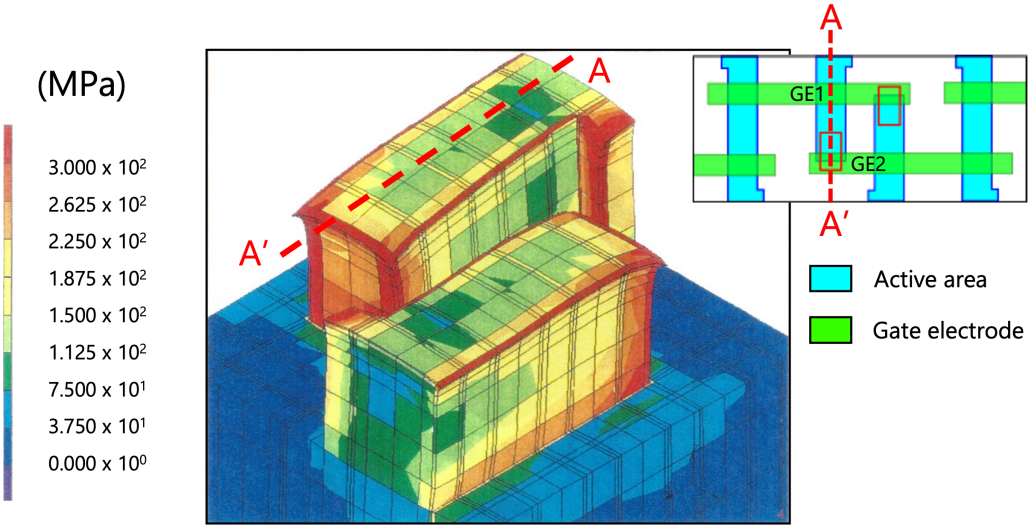


Figure 2.14: Simulated results of maximum resolved shear stress during the source/drain activation annealing process with displacement values. The displacement values are multiplied 200 × for easy observation.

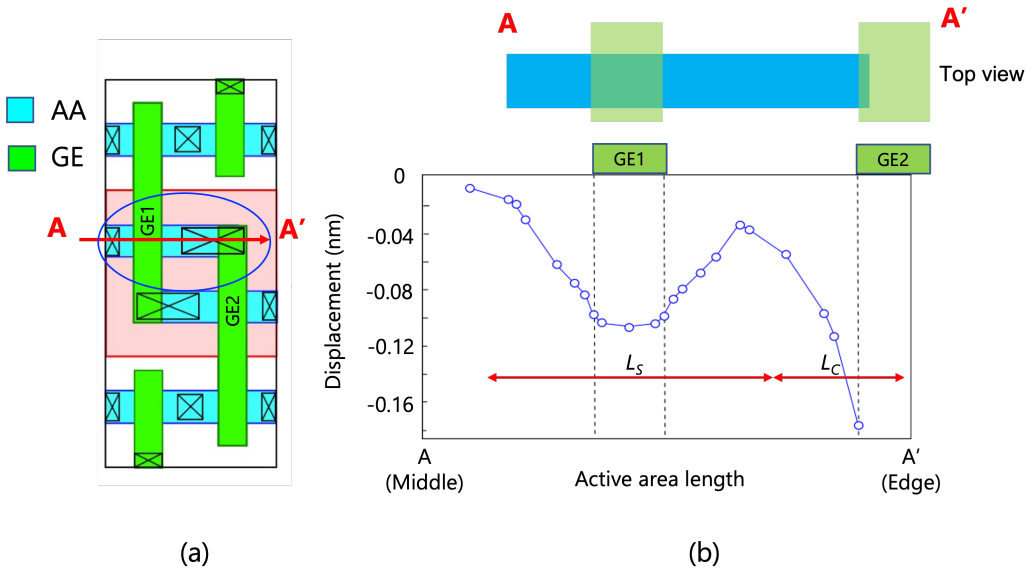


Figure 2.15: (a) Schematic layout of SRAM cell and (b) displacement of the AA surface along the dotted line A−A' in Figure 2.14.

(b). It clearly shows that GEs apply compressive stress to the AA because the stress values are negative. The gate electrode GE1 sits in the middle of the AA, while the gate electrode GE2 locates at the edge of the AA. Both stresses and displacements are in the same direction, and the displacements under GE2 are about 1.8 times larger than those under GE1. The red allows labeled as L_S, and L_C are the length affected by each GEs. The L_S is the length affected by the GE1, and L_C is the same by the GE2. The difference in substrate displacement between GE1 and GE2 is schematically shown in Figure 2.16. Because of the difference in thermal expansion coefficient (polycrystalline silicon (poly-Si) : 2.9×10^{-6}/K, single crystal silicon (Si) : 4.15×10^{-6}/K), wafer warps after the gate electrode material deposition, as shown in Figure 2.16 (a). Stress in the compressive direction is generated in the GE1, which pulls the AA towards the longitudinal direction, as shown in Figure 2.16 (b). Since the edge region of AA could easily be pulled, the displacement amount under the GE2 becomes larger than GE1, as shown in Figure 2.16 (c). To analyze the details of these displacements, models with beam structures are proposed. Figure 2.17 (a) and (b) show models that explain the deflection under GE1 and GE2. The simply supported beam

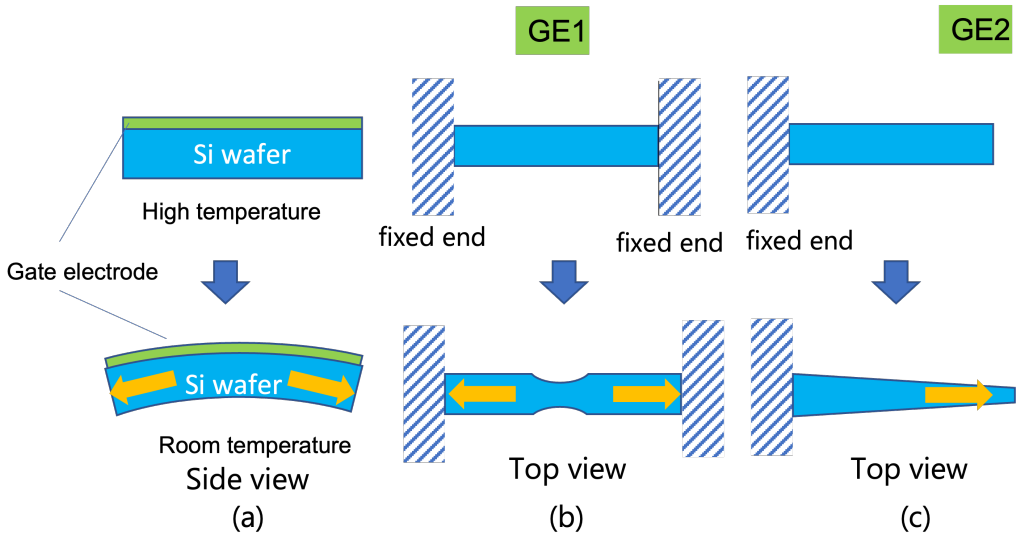


Figure 2.16: Schematic images of displacement difference between GE1 and GE2. (a) Wafer warpage after the gate electrode material deposition. (b) AA under GE1 is pulled by the fixed ends on both sides as it returns to room temperature. (c) AA under GE2 is pulled toward the open end when returning to room temperature.

model, shown as Figure 2.17 (a), is proposed for the deflection of the AA under GE1. The cantilever beam structure model, shown in Figure 2.17 (b), is proposed for the deflection of the AA edge under GE2. The displacement values of both d_C and d_S can be calculated with the following formulas [70]:

$$d_C = \frac{4FL_C^3}{EI} \tag{2.1}$$

$$d_S = \frac{FL_S^3}{4EI} = \frac{2FL_C^3}{EI} \tag{2.2}$$

$$(L_S = 2L_C)$$

F : Load
E : Young's modulus
I : Moment of inertia of area
L_C, L_S : Beam length
d_C, d_S : Displacement

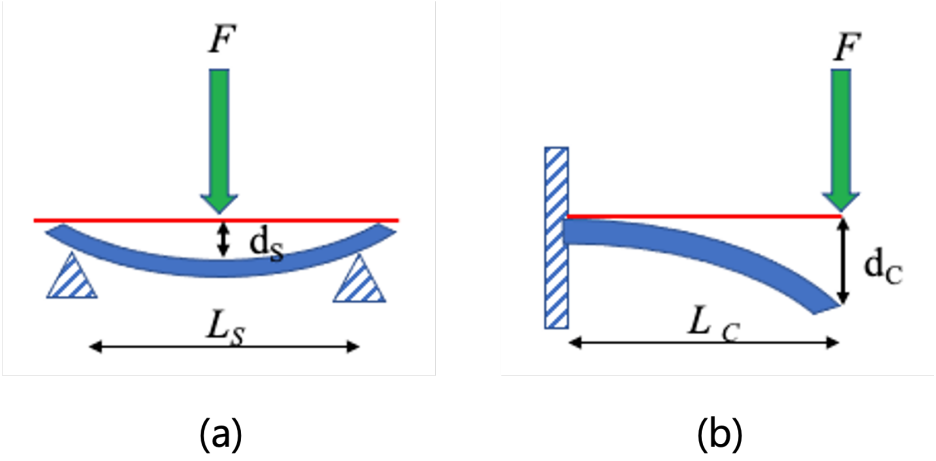


Figure 2.17: Models for explaining the mechanical stress and deflection under GE1 and GE2 shown in Figure 2.15. (a) A simply supported beam structure for under GE1 and (b) cantilever beam structure for under GE2.

where F is the load, E is Young's modulus, I is the moment of inertia of the area, L_S and L_C are the lengths of the simply supported beam and cantilever, respectively. Both d_S and d_C are the corresponding displacement values, respectively. The beam length L_S is about twice that of the L_C. When equation (2) is rewritten with the L_C, the d_C becomes twice the d_S. This is close to the simulation results that showed the displacement value under GE2 is about 1.8 times larger than that under GE1. The AAs of nMOSFETs in the SRAM cells are connected together and considered as both ends fixed beam structure. The amounts of deflection for both ends fixed beam structure, d_B, are given below.

$$d_B = \frac{FL_B^3}{16EI} = \frac{FL_C^3}{2EI} \tag{2.3}$$

$$(L_B = 2L_C)$$

Thus, the d_B is eight times smaller than the d_C. This is why defects were not observed in the nMOSFET regions but observed in the pMOSFET regions. Therefore it is concluded that the beam models can describe this phenomenon and is applicable to estimating the mechanical stress effects.

2.4.2 Cell layout design for 90 nm node

From the simulation results in the previous section, it is better to avoid the cantilever structure in AA under the GE2 region. The overlap lengths of AA and GE in cells vary due to misalignment in the lithography process and pattern shortening by defocusing. The optimization approach would either move AA and GE2 further apart or increase the overlap length. To realize a robust cell design, detailed simulations for layout optimization are carried out. There are several studies on the effects of mechanical stress on transistor characteristics and reliabilities. A time-dependent dielectric breakdown (TDDB) of the gate oxide degrades by the mechanical stress of more than 200 MPa [75,76]. Both hot carrier reliability and bias temperature instability also deteriorate by the compressive stress under the gate electrodes [77–79]. These degradations depend on the amount of stress, and they concluded that the stress value less than 200 MPa does not affect the reliabilities [80–83]. Another report showed that the mechanical stress in the AA increased by about 5% with a 10% reduction in dimensions [64]. Therefore, the maximum mechanical stress should be at least 15% less than 200 MPa for a possible extension to 65 nm nodes, which dimensions are 30% smaller than those of 90 nm nodes. Thus, the maximum allowable stress should be less than 170 MPa.

The AA edge of the initially designed cell had a 60 nm overlap with the GE2. However, the AA shrank in the longitudinal direction because of the pattern shortening, and the actual overlap length was about 50 nm. Since the gate length is 80 nm, the edge of the AA was located at about the middle of the GE2. According to the cantilever beam model, this is close to the conditions that cause the maximum deflection. The effect of the overlap length on the mechanical stress at the AA under the GE was investigated for further understanding of this phenomenon. Figure 2.18 (a) shows a cross-sectional schematic image of the overlap structure. The overlap length d varied from − 100 nm to + 120 nm. Since SRAM cells are arranged symmetrically, as shown in Figure 2.18 (c), the longer the overlap length of an AA, the shorter the distance between the AA and its counterpart. Because the minimum spacing between AAs must be maintained, if the overlap length

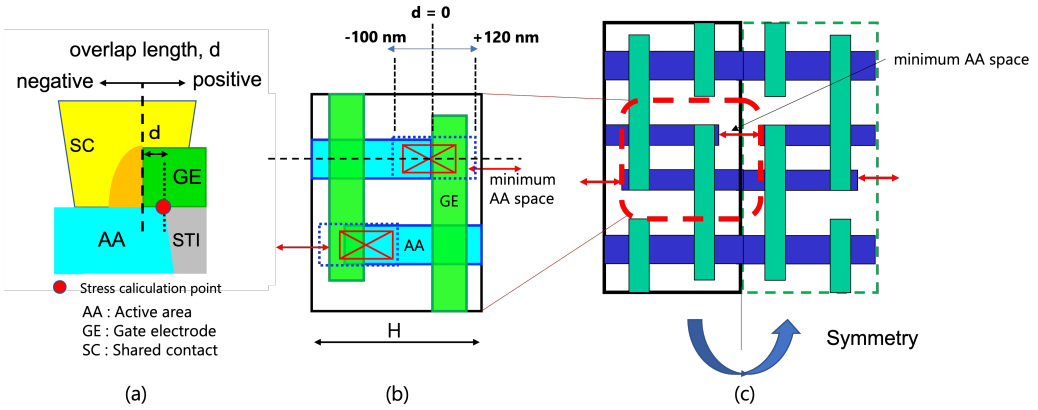


Figure 2.18: (a) Schematic cross-section image of AA−GE overlap. The stress monitoring point is indicated at red circle. (b) Top view of the SRAM cell. The overlap length d is varied from − 100 nm to + 120 nm. (c) Schematic image of SRAM cell arrangement.

is increased to the positive side, the cell height H, as shown in Figure 2.18 (b), must be increased, and it results in the cell size increases.

Figures 2.19 (a) through (d) are the simulated contour plots of the resolved shear stress for each overlap length, (a) d = − 60 nm, (b) d = 0 nm, (c) d = + 60 nm and (d) d = + 120 nm, respectively. For easy observation, only AAs are shown. The amounts of displacement towards the depth direction are also included in the figures.

The high-stress values were observed at the edges of the AAs where defects were observed, and those regions increased as the overlap lengths were increased from − 60 nm to + 60 nm. However, when the overlap length is increased to + 120 nm, high-stress regions at the AA edge are reduced. This is because the edge of the AA is apart from the GE, which shifted from a cantilever structure to a simply supported beam structure, as shown in Figure 2.17.

Figure 2.20 shows the AA−GE overlap length dependence on the maximum resolved shear stress at the stress monitoring point. With increasing the overlap length d from − 60 nm to + 60 nm, the amounts of stress also increase. However, when d is increased to + 120 nm, the amount of stress decreases. This is because the AA passes through under the GE, and the structure changes from a cantilever to a simple beam structure. Since an

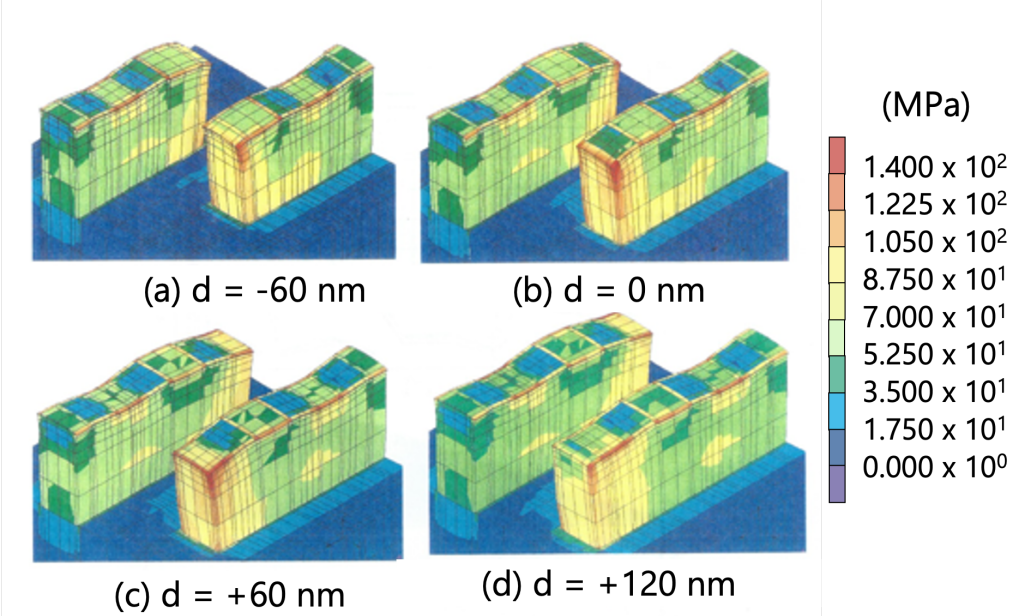


Figure 2.19: Simulated mechanical stress for each AA−GE overlap amount with AA deflection, where (a) d = − 60 nm, (b) d = 0 nm, (c) d = + 60 nm and (d) d = + 120 nm, respectively.

overlap length of more than + 80 nm increases the cell height, this design cannot be acceptable. If the overlap value is less than − 60 nm, the AA edges are off the shared contacts and do not overlap, resulting in contact failures. Therefore, the maximum allowable overlap lengths are between − 60 nm to 0 nm to ensure that the stress remains less than 170 MPa. These simulations were carried out without sidewall structure in the gate electrode. There is a report that the sidewall structure affects the channel stress [84]. Therefore, additional simulations are carried out with sidewall structures to investigate their effects.

Figure 2.21 shows the schematic images of sidewall structure which consists of (a) dual-layer structure with 20 nm SiN and 80 nm SiO_2 and (b) triple-layer structure with 20 nm SiO_2, 20 nm SiN, and 40 nm SiO_2. The sidewall materials affect the transistor reliabilities [85]. Therefore, the selection of an appropriate sidewall structure is important. Figure 2.22 the simulated results of shear stress at the AA surface under the gate electrode and sidewalls, along with the active area length. The triple-layer sidewall

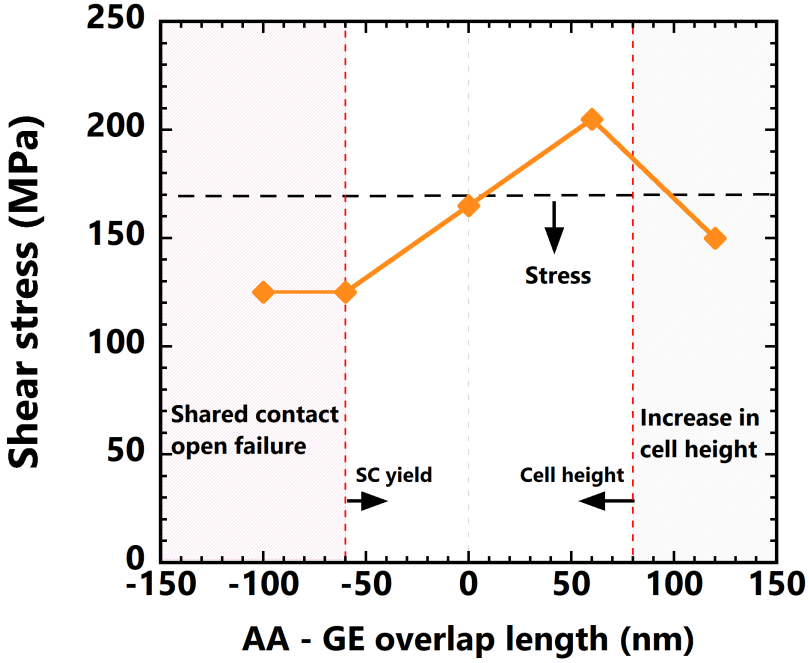


Figure 2.20: Simulated shear stresses and their AA − GE overlap length dependence shown in Figures 2.19 (a) - (d). The stress criteria of 170 MPa is indicated as a red dotted line.

structure shows the maximum stress under the GE region. In contrast, the dual-layer sidewall structure shows the maximum stress under the sidewall region. Moreover, the stresses under the GE region are 10% less for the dual-layer sidewall structure. For the dual-layer sidewall structure, it is considered that the compressive stress from the SiN film appears to have increased the stress under the sidewall. On the other hand, in the triple-layer sidewall structure, the SiO_2 film between GE and SiN reduces the stress from the SiN film, and the stress by the GE becomes dominant.

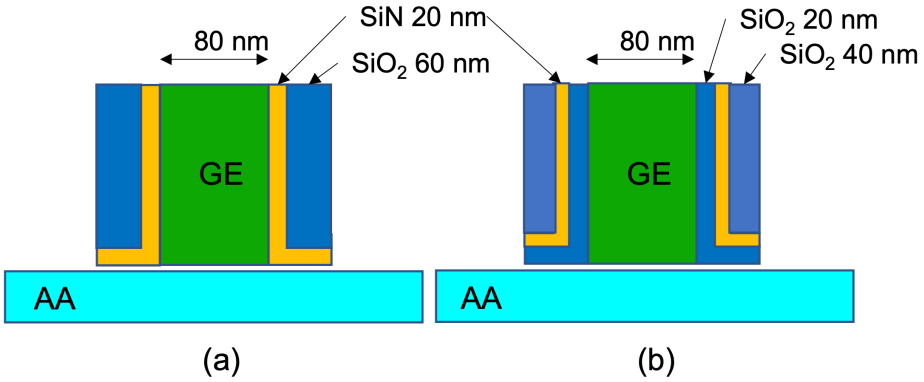


Figure 2.21: Sidewall configurations used for simulations: (a) dual-layer sidewall and (b) triple-layer sidewall structures.

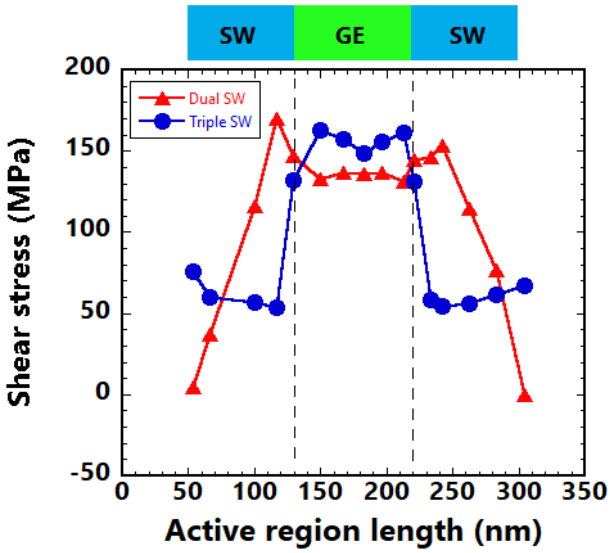


Figure 2.22: Simulated shear stress under GE along the active area for both sidewall structures. Dual-layer sidewall shows the maximum stress under the sidewall region while triple-layer sidewall shows the maximum stress uner GE region.

The criteria of the offset value are set as follows. If the overlap value is larger than − 60 nm, as shown in Figure 2.23 (a), there is no overlap between AA and shared contact (SC) that results in open failure. The overlap amount can be increased up to + 80 nm without any penalties, as shown in Figure 2.23 (b). However, if the overlap amount exceeds + 80 nm, it causes

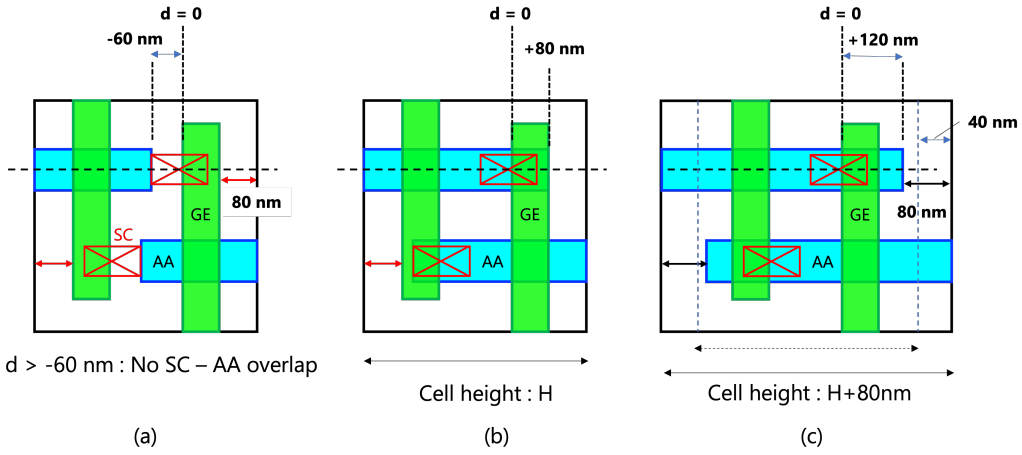


Figure 2.23: Schematic images of cell size impact by AA−GE overlap length. The overlap length d = − 60 nm (a), d = + 80 nm (b) and d = + 120 nm (c) are shown with cell height impact.

cell height increase (horizontal direction in this figure) because SRAM cells are placed in symmetry as shown in Figure 2.18 and are needed to keep minimum space (160 nm at 90 nm node) between AAs in the longitudinal direction. Therefore, in order to keep this minimum space, cell height needs to be increased and will be 80 nm larger if the overlap value is set as + 120 nm, as shown in Figure 2.23 (c).

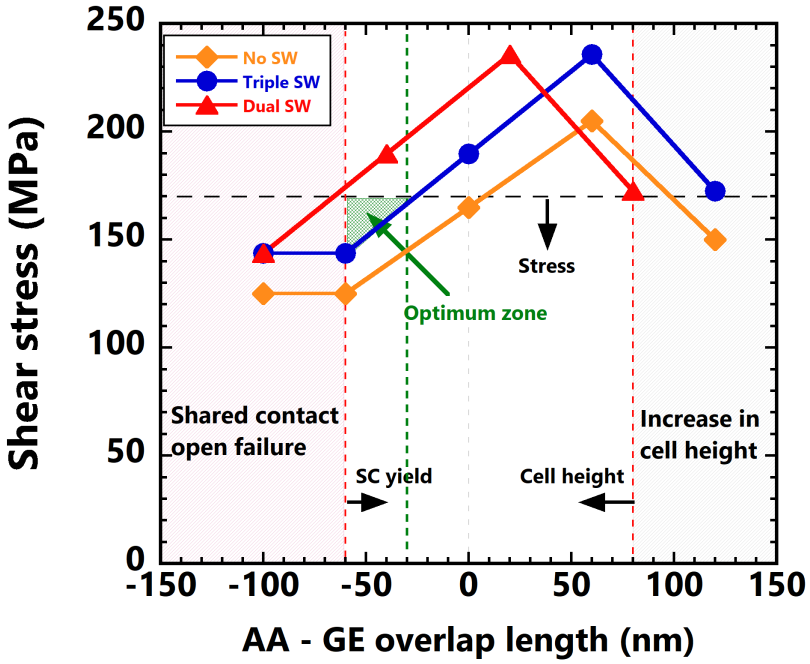


Figure 2.24: AA−GE overlap length dependence of simulated shear stresses under the GE with sidewall structure dependence. The optimum overlap length regions (− 60 nm to − 30 nm) are indicated as green triangle.

Figure 2.24 summarizes the design criteria of AA−GE overlap amount for both dual and triple sidewall structures. No sidewall structure is also shown as a reference. The overlap amount should be larger than − 60 nm from the SC yield criteria. In order to satisfy the mechanical stress criteria (< 170 MPa), the optimum overlap should be between − 60 nm to − 30 nm for a triple-layer sidewall structure to prevent cell height increase. There is no design area for a dual-layer sidewall structure. If cell height increase is permitted, more than + 80 nm overlap for dual-layer sidewall structure or more than + 120 nm overlap for triple-layer sidewall structure can be used. The optimum overlap length zone is indicated as a green triangle in the same figure. From a high-capacity SRAM realization point of view, a − 45 nm overlap length with a triple-layer sidewall is recommended for 90 nm node SRAM cells.

A 1 Mbit SRAM test vehicle was fabricated to confirm the robustness of the optimized cell layout. Figure 2.25 shows the cumulative distribution of

the shared contact resistance measure from the 6.5 k contact chain, which indicates a tight distribution and no open failures. The function bit counts with static noise margin (SNM) for both V_{DD} = 0.6 and 1.2 V are shown in Figure 2.26. The proposed design guideline was validated successfully. Since the number of SRAM test vehicles was limited during the R&D phase, no failures were confirmed and transferred to both memory and logic divisions. The development team in the fab ran many wafers for qualification, and they confirmed the robustness of this design guideline. One example of products, Figure 2.27 shows a 45 nm node embedded SRAM cell layout as one of the examples that followed this design guideline and were mass-produced.

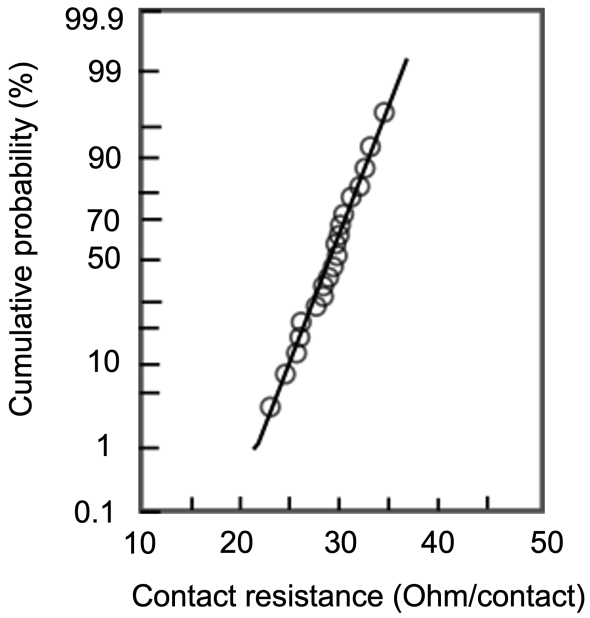


Figure 2.25: Cumulative probability of the shared contact resistance extracted from 6.5 k chain.

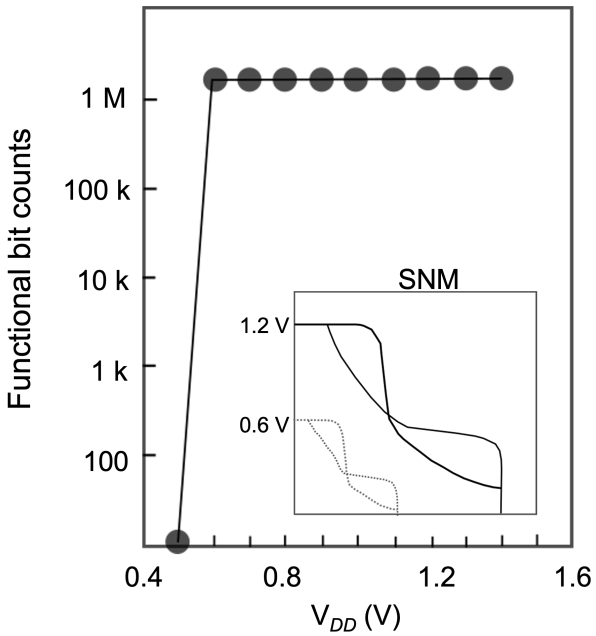


Figure 2.26: Functional bit counts of fabricated 1 Mb SRAM test vehicle with SNM of both $V_{DD} = 0.6$ and 1.2 V.

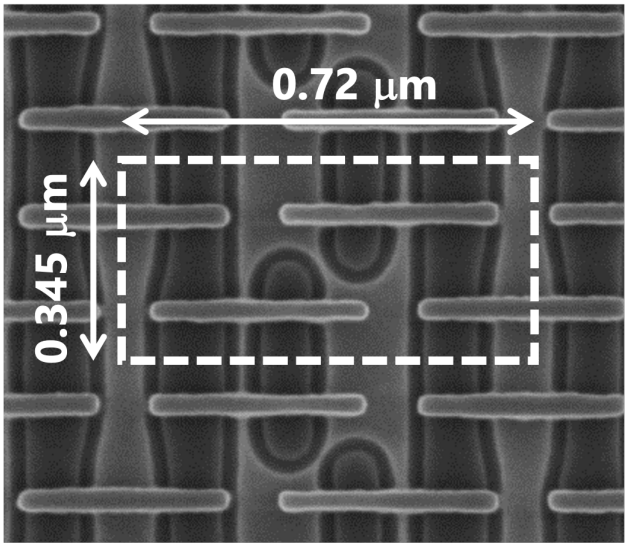


Figure 2.27: A 45 nm node embedded SRAM cell layout used for mass production [86].© 2006 IEEE

2.4.3 Universality of presented guideline

It is important to confirm the universality of the presented cell design guideline. M. L. Polignano *et al.* did a similar study using a test structure as shown in Figure 2.28 [65]. The width of AA was 180 nm. The simulated shear stress is shown in Figure 2.29. Fraction of the failed structure as a function of the calculated σ_{xx} and shear stress are shown in Figures 2.30 (a) and (b). Their shear stress criteria of 120 MPa are slightly lower than the 170 MPa obtained in this study. The difference is considered as follows.

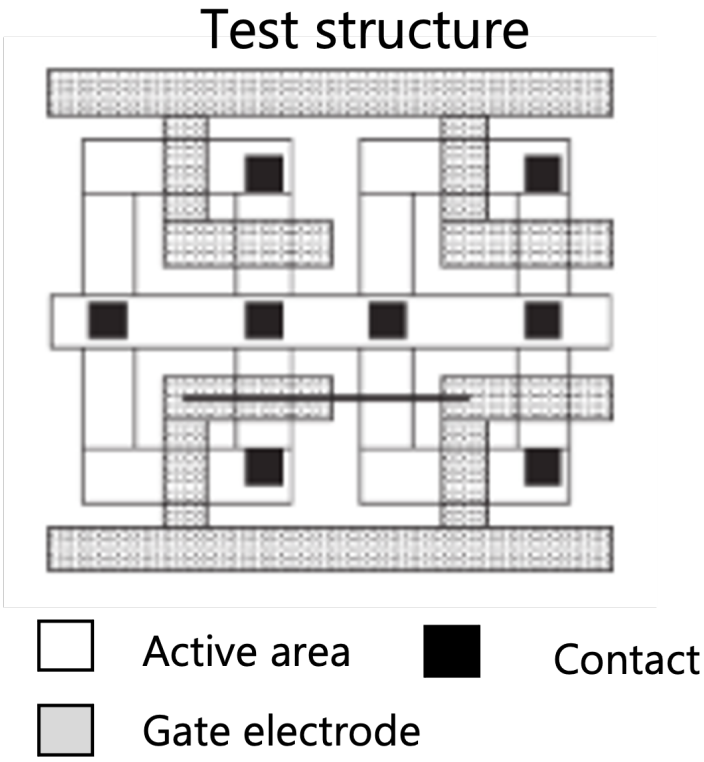


Figure 2.28: Layout of the structures electrically tested for defect-formation monitoring [65].© 2007 IEEE

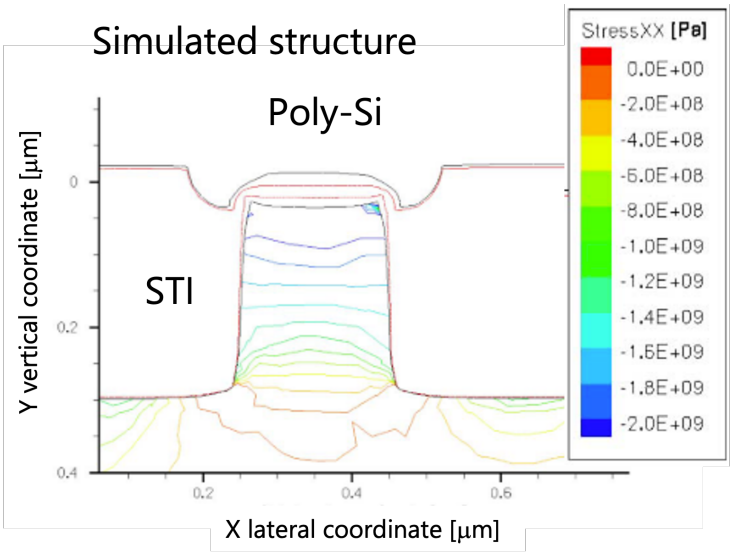


Figure 2.29: Cross-sectional the simulates structure and level curves of the σ_{xx} [65].© 2007 IEEE

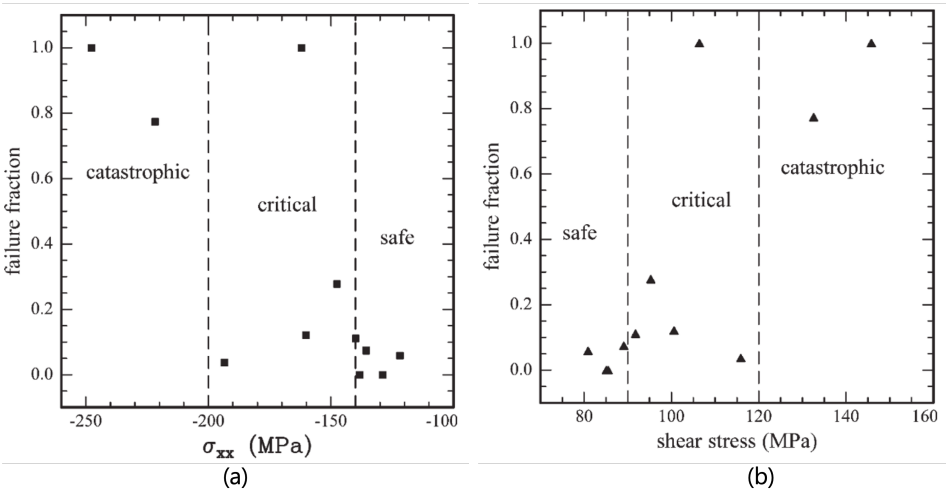


Figure 2.30: Calculated fraction of the failed structure for both (a) σ_{xx} and (b) shear stress [65].© 2007 IEEE

The past study [64] showed that a longer GE length causes a larger stress value resulting in defect generation than a shorter GE length. Their STI structure was not optimized, and the upper corner of the STI has a sharp corner. Moreover, the GE straddles the AA, making it more susceptible to

stress from the GE. From the above, it can be said that it is important to select appropriate parameters for device geometry and materials; if these device structures and materials are estimated correctly during simulation, stress control below 170 MPa as a boundary condition is common, and defect suppression is possible.

Past research suggested that the scaling of AA may increase the amount of mechanical stress. If the only width of AA is scaled, it causes about a 3 to 5% stress increase with 10% dimension scaling [64, 71–73]. However, both isolation width and AA width are scaled, which applies SRAM cell scaling. The amount of stress value is almost the same or decreases [74]. This model was calibrated and confirmed by the 65 nm node devices. Therefore the stress criteria less than 170 MPa can be feasible as far as planar MOSFET structure is used, such as down to 32 nm node.

The SRAM cell size can be calculated by using the design rules of key layers, as shown in Figure 2.31. The well isolation width is 4×GE, contact space is 1.5×CS, and GE-CS space is about 0.5×CS. Therefore, cell width is 7×CS + 8×GE. The cell height is also calculated as 2×(GE + CS) + 2×CS. The feature size F is defined as F = (GE + CS)/2. Therefore, cell width is about $16F^2$, and cell height is about $8F^2$, which makes cell size $128F^2$. It should be noted that the feature size F is the same as the technology node (F = 0.09 μm for 90 nm node).

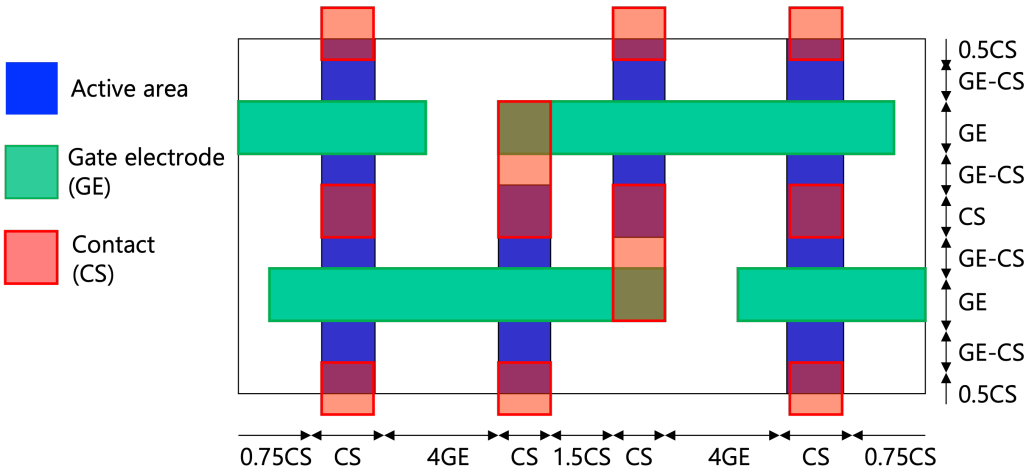


Figure 2.31: SRAM cell size definition by key design rules.

The ground rule of key layers with SRAM cell size and suggested AA-GE overlap length from 90 nm to 32 nm node are summarized in Table 2.4. Dual-layer sidewall structure requires cell height increase as described, which results in about a 6% increase in cell size.

Table 2.4: Design rules of key layers with target cell size and AA-GE overlap length from 90 nm to 32 nm technology node.

	Technology node			(T) Triple layer SW (D) Dual layer SW
Layer	**90 nm**	**65 nm**	**45 nm**	**32 nm**
Active area pitch (nm)	**260 (L/S:120/140)**	**200 (L/S:90/110)**	**140 (L/S:60/80)**	**100 (L/S:45/55)**
Gate electrode pitch (nm)	**240 (L/S:80/160)**	**180 (L/S:60/120)**	**140 (L/S:50/90)**	**100 (L/S:35/65)**
Contact pitch (nm)	**240 (L/S:100/140)**	**200 (L/S:80/120)**	**140 (L/S:60/80)**	**100 (L/S:40/60)**
1st Metal pitch (nm)	**240 (L/S:120/120)**	**180 (L/S:90/90)**	**140 (L/S:70/70)**	**100 (L/S:50/50)**
SRAM cell (μm^2)	**1.00 (T) 1.06 (D)**	**0.50 (T) 0.53 (D)**	**0.25 (T) 0.27 (D)**	**0.13 (T) 0.14 (D)**
AA-GE overlap (nm) & SW	**-60 to -30 nm (T) 80 to 120 nm (D)**	**-45 to -22 nm (T) 60 to 90 nm (D)**	**-30 to -15 nm (T) 50 to 70 nm (D)**	**-20 to -10 nm (T) 35 to 50 nm (D)**

2.4.4 DRC implementation

When SRAM cells are generated using current commercially available EDA (Electronic Design Automation) tools, the cells are automatically generated with AA and GE offset structure. However, it should be noted that this does not take into account the effect of mechanical stress. The standard DRC (design rule check) prohibits the placement of contacts in the area where AA and GE overlap in order to prevent AA and GE from shorting due to contacts, as shown in Figure 2.32.

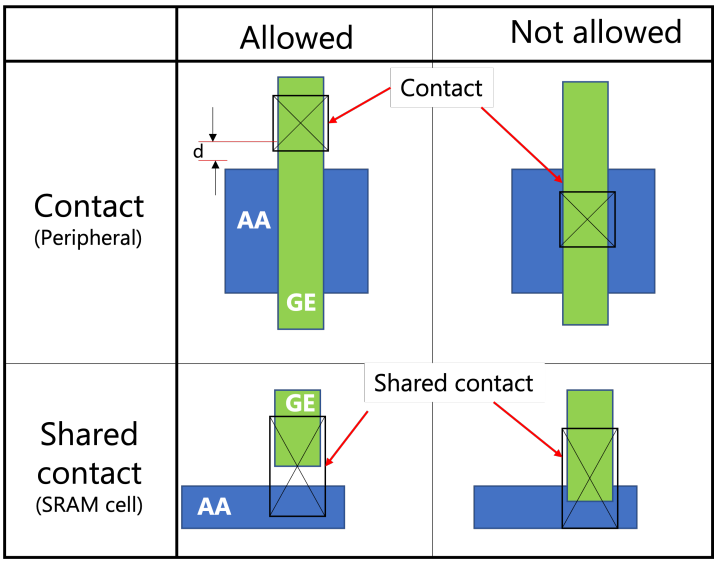


Figure 2.32: Layouts that are allowed and not allowed by conventional DRC.

Therefore, when shared contacts are formed, they are automatically formed with a layout separating AAs and GEs. In this case, the offset is set for each generation to take into account the misalignment and dimensional variation of AA and GE. If a DRC that allows shared contact is prepared specifically for SRAM cells, it is possible to form contacts in the region where AA and GE intersect, and this has been used in SRAMs that use FinFETs [87]. The DRC is implemented separately from the peripheral circuits in high-capacity SRAMs due to the large cell size. In many cases, a DRC is prepared exclusively for the SRAM cell. In the case of SRAM cell generation using standard EDA tools, it should be noted that the AA and GE offset value is not optimized, but only the offset structure is used for the above-mentioned reason. In order to realize highly reliable high-capacity SRAM, it is necessary to estimate the optimal amount of offset using mechanical stress simulation and verify whether it is acceptable to design with standard DRC.

2.5 Summary of this chapter

Optimization methodology of cell layout to embed high-reliable high-capacity SRAM into 3D NAND was presented. Two SRAM cell layouts, conventional and thin types, were compared from the viewpoint of scalability in the 90 nm node and beyond. It was found that mechanical stress from the STI and gate electrode causes crystal defects, and the defect generation mechanism was analyzed and modeled by using the finite element simulation. The active area (AA) edge was easily distorted by the stress from both STI and gate electrodes (GE). Compressive stress by the gate electrode was found to cause a large deflection in AA, and its effect was enhanced at the AA edge. This phenomenon can be described by models of cantilevers and simply supported beams. The sidewall structure modulates the mechanical stress and affects the optimum overlap length between AA and GE. The allowable maximum mechanical stress value for defect suppression was 170 MPa, which was shown to be reasonable in conjunction with results from other institutes. An overlap length between − 60 nm to − 30 nm with a triple-layer sidewall structure was recommended for 90 nm nodes. The validity of this cell design guideline was demonstrated by a 1 Mb SRAM using the 90 nm node rules, and the validity of this guideline was confirmed at the product level up to the 45 nm generation. The SRAM cell, considering the proposed design guideline, can be scaled down to the 32 nm nodes as far as planar structure is used. In addition, when SRAM cells are laid out using EDA tools, AA and GE are offset, as shown in this study when standard DRC is used, but this does not take into account the effect of mechanical stress. By fully utilizing the above-mentioned cell layout optimization methodology, high-reliable high-capacity SRAM can be embedded in the 3D NAND process.

Chapter 3

Reliability of SRAM cell transistors

3.1 Background

The introduction of the shallow trench isolation (STI) process that replaced the LOCOS (LOCal Oxidation of Silicon) isolation process enabled further device dimension scaling and became an inevitable component in current LSIs (Large Scale Integrations). In memory devices, such as SRAMs and DRAMs, CMOS transistors with minimum dimensions are used for memory cells in order to realize higher capacity. One of the issues caused by STI introduction is the hump characteristics in CMOSFETs (CMOS field-effect transistors). Many studies have been done about hump characteristics and suppression by processes [88, 89], and also for hot-carrier reliability analysis [91, 92]. Since suppressing hump characteristics leads to improvement in the reliability of CMOSFETs, efforts were focused on suppressing hump characteristics.

However, there were not many studies on CMOSFETs about the reliability in the narrow channel width regions, especially for pMOSFETs. Understanding the reliability in the narrow channel region is important to realize high-reliable SRAM. This chapter discusses the reliability of narrow channel width CMOSFETs in realizing high-reliable high-capacity SRAM.

3.2 Past study on nMOSFETs

A phenomenon has been found in scaled nMOSFETs with suppressed hump characteristics, in which the narrower the channel width, the lower the reliability [101]. Figure 3.1 shows variations of substrate current during the hot-carrier stress with several channel widths. Transistor was fabricated with 0.35 μm technology. The substrate current, which causes reliability degradation, increased rapidly, whose a channel width being less than 1 μm while the wide channel transistor (W_{eff} = 9.93 μm) slightly decreased. The substrate current change represents the impact ionization rate [94] change during the hot-carrier stress. The fact that the impact ionization rate increases with stress time mean that degradation due to the hot-carriers is accelerating. It is predicted that MOSFETs with narrower channel widths will have higher impact ionization rates from the initial state, and the channel width dependence of the initial impact ionization rate was measured, as shown in Figure 3.2. Contrary to expectations, the narrower the channel width MOSFETs showed a lower impact ionization rate. Thus, it was suggested that a different degradation model exists than the channel width dependence of the impact ionization rate.

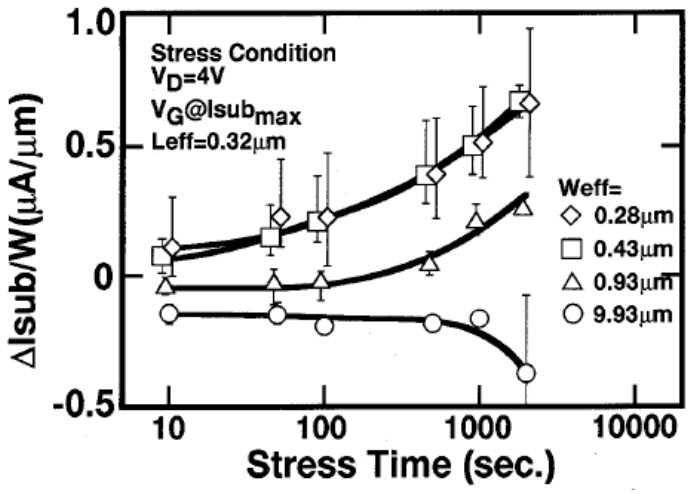


Figure 3.1: Stress time dependence on the substrate current change, delta I_{sub}, for various channel widths. W_{eff} is the effective channel width corrected from the electrical characteristics [101].© 1996 IEEE

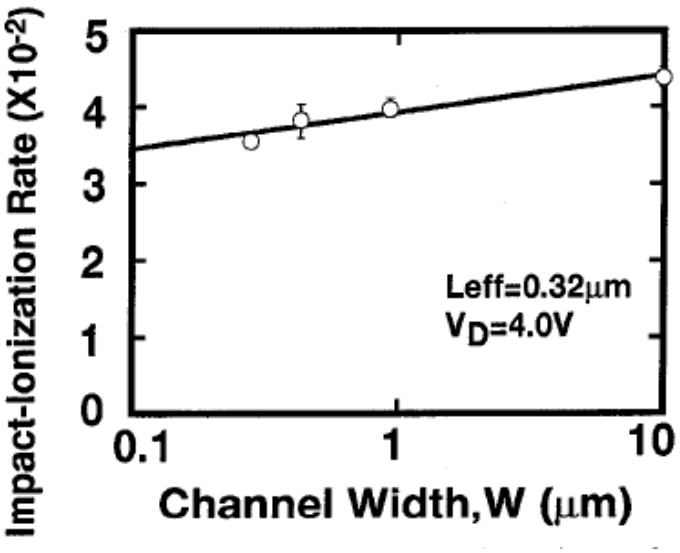


Figure 3.2: Channel width dependence of the impact ionization rate for measured nMOSFETs [101].© 1996 IEEE

Impact ionization depends on the lateral electric field between the source and drain and the vertical electric field between the gate and drain regions. Therefore, three-dimensional device simulations were performed to investigate this effect in detail.

Figure 3.3 shows the simulated contour plot of the electric field at the drain edge region for both (a) edge of the channel and (b) center of the channel, respectively. The initial current path is indicated as the gray arrow, and the current path after hot-carrier stress, which is shifted to a deeper region, is indicated as the black arrow. As shown in Figure 3.3 (a), in the channel region near the STI edge, the initial drain current flows in the region close to the Si substrate surface. When electrons are trapped in the gate dielectric film due to hot carrier stress, they weaken the electric field from the gate electrode, and the drain current gradually flows in the region away from the substrate surface. As a result, the drain current flows in a region where the transverse electric field is stronger than before the stress is applied, which is thought to result in an increase in the impact ionization rate.

On the other hand, as shown in Figure 3.3 (b), in the center of the channel, the drain current flows in the region where the transverse electric

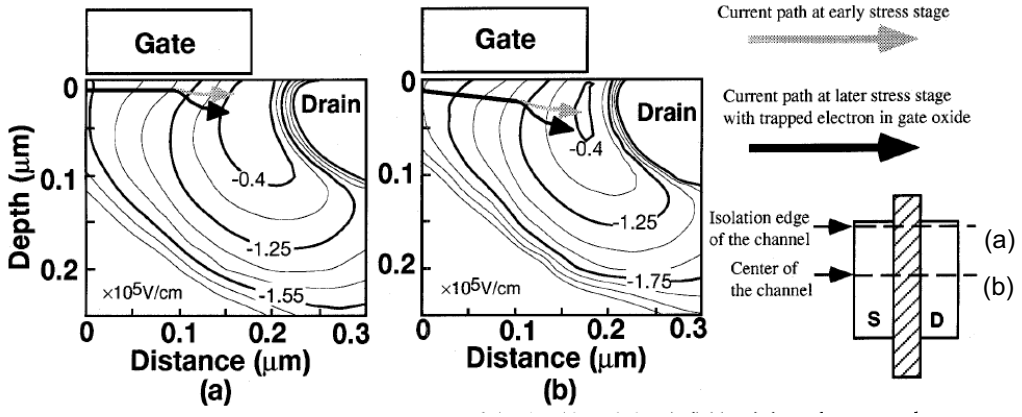


Figure 3.3: Simulated electric field with drain current path. (a) The channel near the STI edge and (b) center of the channel [101].© 1996 IEEE

field is maximum from the initial state, and as the electrons trapped in the gate dielectric increase with increasing stress time, the drain current path flows in a deeper region on the substrate side, that is, a region with a weaker transverse electric field. Therefore, the impact ionization rate decreases with stress time. Figure 3.4 shows the vertical electric field at the Si/SiO_2 interface, both the STI edge and center of the channels. The channel near the STI edge region shows more than twice the vertical electric field. Although the capture cross-section of the electron is small [93], this strong vertical electric field attracts more electrons to the gate insulator. The fringing field causes the difference of this vertical electric field from the gate electrode edge through the STI isolation region. Since the channel width becomes narrow, this edge region becomes dominant. Therefore, narrow channel MOSFETs show accelerated degradation by hot-carrier stress.

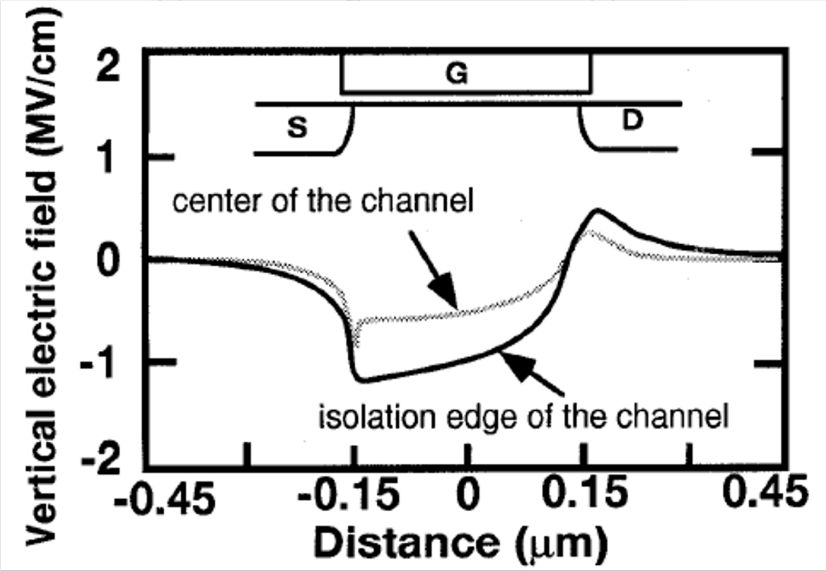


Figure 3.4: Simulated vertical electric field along the channel for both isolation edge of the channel and center of the channel [101].© 1996 IEEE

3.3 Reliability in narrow channel width pMOSFETs

There were many studies on the enhanced hot-carrier induced degradation for narrow width n and pMOSFETs with LOCOS isolation [90, 95–97]. Even after the introduction of STI, the hot-carrier reliability was still evaluated by using relatively larger channel width (W) MOSFETs such as W $\geq$ 1 μm. The importance of reliability measurement by using the minimum size of MOSFETs is described in the previous section. However, it was only for nMOSFETs, and still, there were not many studies on narrow channel width pMOSFETs. The minimum size MOSFETs are used not only for SRAM cells but also for data latches and decoders. Therefore, understanding the reliability of narrow channel width pMOSFETs is also important.

3.3.1 Characteristics of narrow channel width pMOSFETs

The MOSFETs used in this experiment were fabricated by a 0.35 μm dual-gate CMOS process [69], which is the same used for nMOSFETs in the previous chapter. The gate oxide thickness is 6 nm, and the WSi/poly-Si gate electrode has a SiN sidewall. Since the STI process was optimized [69], both wide and narrow channel width pMOSFETs show no hump characteristics even though substrate bias 1 V is applied, as shown in Figure 3.5.

The channel width dependence of hot-carrier reliability was measured by using MOSFETs with channel width W = 0.35, 1.0, and 20 μm, which gate length is 0.35 μm. The decrease of the drain current in the saturation region, I_{dsat}, under the bias $V_g = V_d = -3.3$ V condition and the amount of threshold voltage shift, ΔV_{th}, were monitored. The stress conditions for all devices were under the $V_d = -4.6$ V, and V_g varied from -0.3 to -1.5 V, which includes the maximum gate current (I_{gmax}) condition. The hot-carrier stress tests were interrupted periodically to measure the I_{dsat} degradation ($\Delta I_{dsat}/I_{dsat}$).

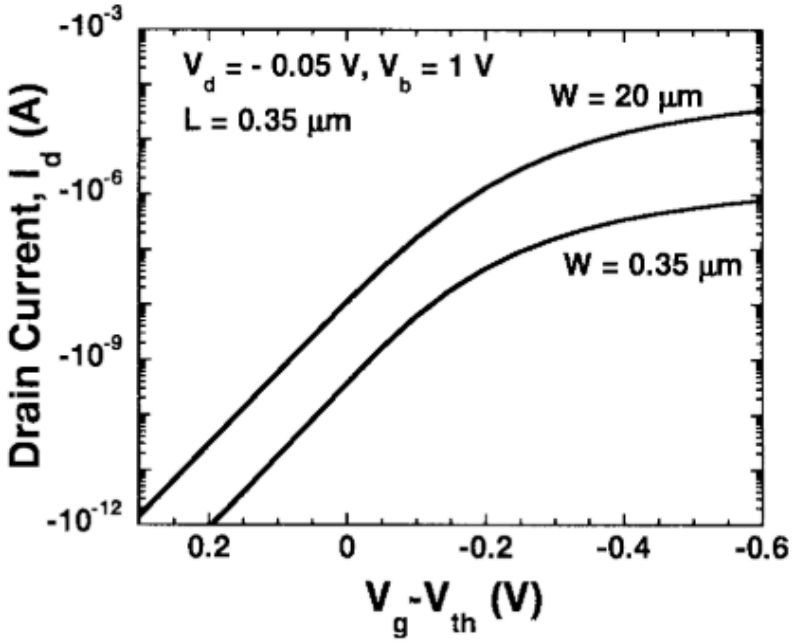


Figure 3.5: Initial I_d - V_g characteristics of pMOSFETs whose channel width W = 0.35 μm and 20 μm. Since STI process is optimized, no hump characteristics was observed.

3.3.2 Hot-carrier stress results

The channel width dependence of the drain current degradation, $\Delta I_{dsat}/I_{dsat}$, is shown in Figure 3.6 as a parameter of hot-carrier stress time under the stress condition of $V_g = I_{gmax}$ at $V_{ds} = -4.6$ V. It can be seen that the change in drain current $\Delta I_{dsat}/I_{dsat}$ increases as the channel width narrows. The amount of shit for W = 0.35 μm devices shows about three to five times larger than those of W = 20 μm devices. This tendency does not change during the stress time up to 50 minutes.

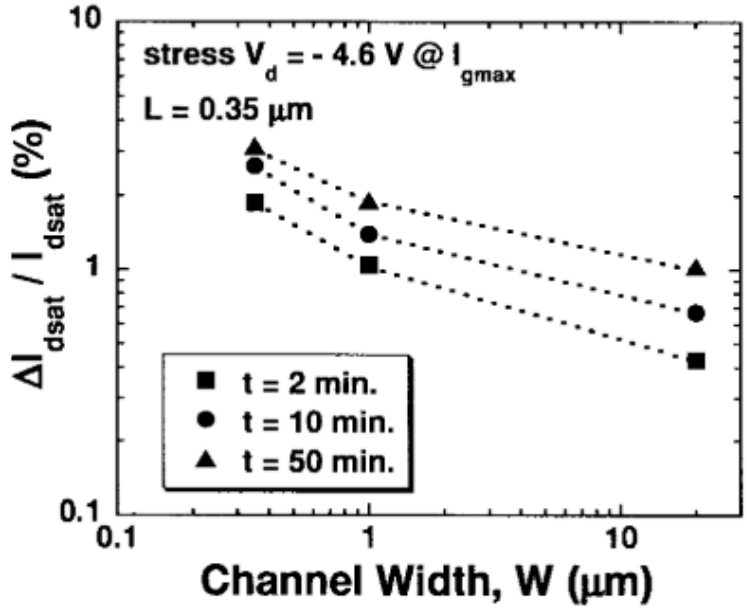


Figure 3.6: Channel width dependence of drain current change, $\Delta I_{dsat}/I_{dsat}$, monitored at stress time t = 2, 10 and 50 minutes.

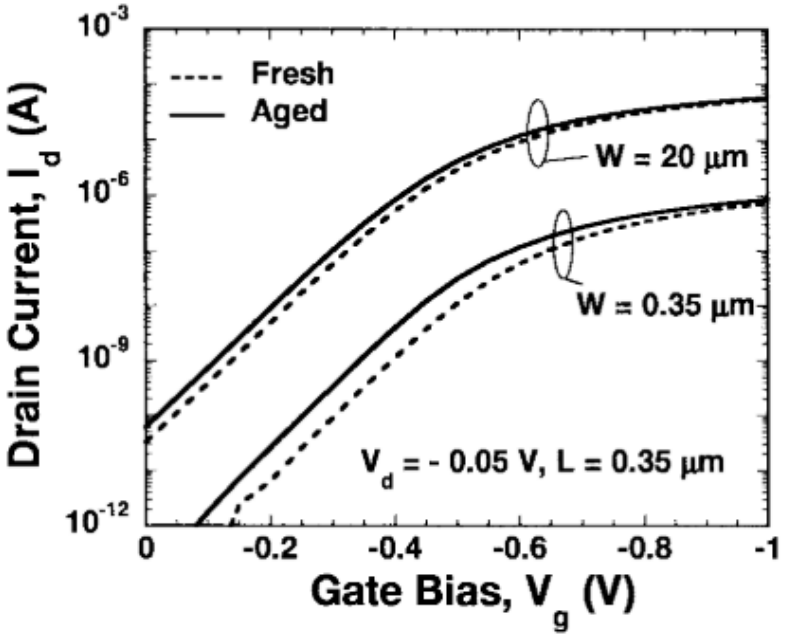


Figure 3.7: I_d - V_g characteristics of both W = 0.35 and 20 μm pMOSFETs after the hot-carrier stress (solid lines). Initial characteristics also shown as dotted lines.

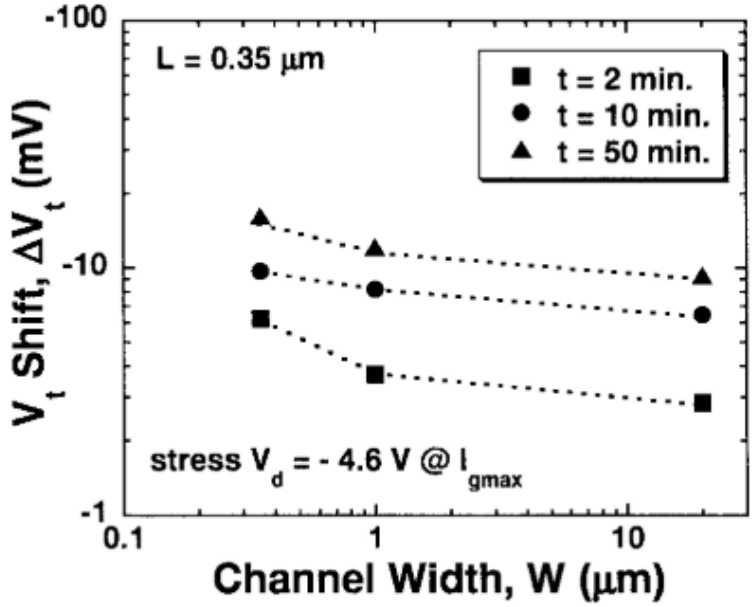


Figure 3.8: Channel width dependence of the threshold voltage degradation (ΔV_{th}) as a parameter of hot-carrier stress time. Narrow width devices show larger degradation as same as $\Delta I_{dsat}/I_{dsat}$.

Figure 3.7 shows I_d - V_g characteristics for both W = 0.35 and 20 μm pMOSFETs after fifty minutes of hot-carrier stress (Aged) indicated as solid lines. The I_d - V_g characteristics before the stress (Fresh) are also indicated in the same figure shown as dotted lines. The threshold voltage of both devices becomes lower that indicating the drain current degradation is caused by the trapped electrons in the gate oxide film [98]. Since the interface states are shielded by the negative charges [99, 100], the subthreshold slope does not change after the hot-carrier stress.

Figure 3.8 shows the channel width dependence of V_{th} shift, ΔV_{th}, as a parameter of stress time. The narrow width devices show larger ΔV_{th}. However, those amounts are less than 20 mV at the largest. These small V_{th} decreases cannot explain the drain current degradation shown in Figure 3.6. Figure 3.9 shows the stress time dependence of $\Delta I_{dsat}/I_{dsat}$ with various MOSFET dimensions. It should be noted that the smallest size MOSFET, such as W/L = 0.35 μm/0.35 μm, shows larger degradation

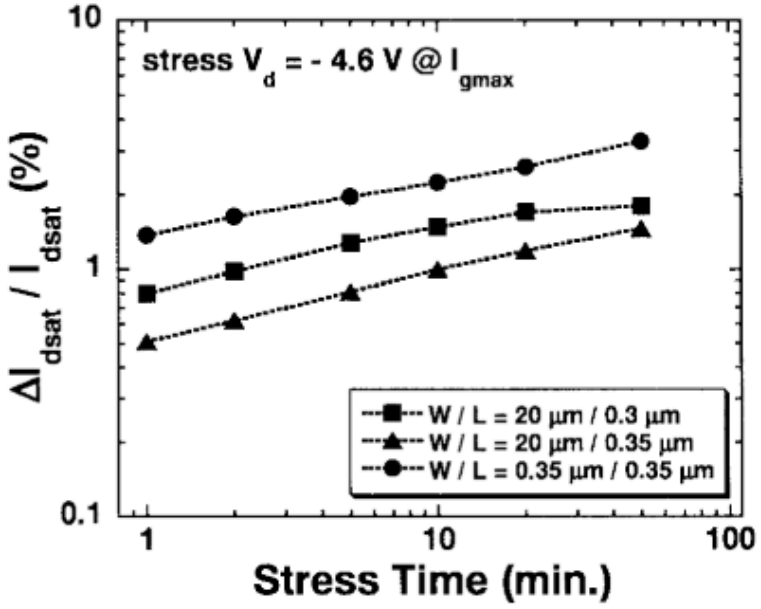


Figure 3.9: Stress time dependence of drain current degradation, $\Delta I_{dsat}/I_{dsat}$, for MOSFETs with various W/L dimensions.

compared to the MOSFET, which has a minimum gate length with wider channel width, such as W/L = 20 μm/0.3 μm. Under the same degradation rate in $\Delta I_{dsat}/I_{dsat}$, W/L = 0.35 μm/0.35 μm device shows more than ten times faster degradation compare to that of W/L = 20 μm/0.3 μm device. Generally, MOSFET with shorter gate length shows a shorter lifetime in the hot-carrier reliability test. However, this does not apply to the narrow channel width device.

To investigate this phenomenon in detail, the channel width dependence of drain current degradations was measured at different gate currents, such as I_{g0}/W = 12, 36, and 55 pA/μm, as shown in Figure 3.10. The I_{g0} represents the gate current at stress time t = 0 second and is normalized for the channel width. For each condition, the tendency for the amount of degradation to increase with narrower channel widths remains the same. Also, the amount of degradation increases as the gate current increases.

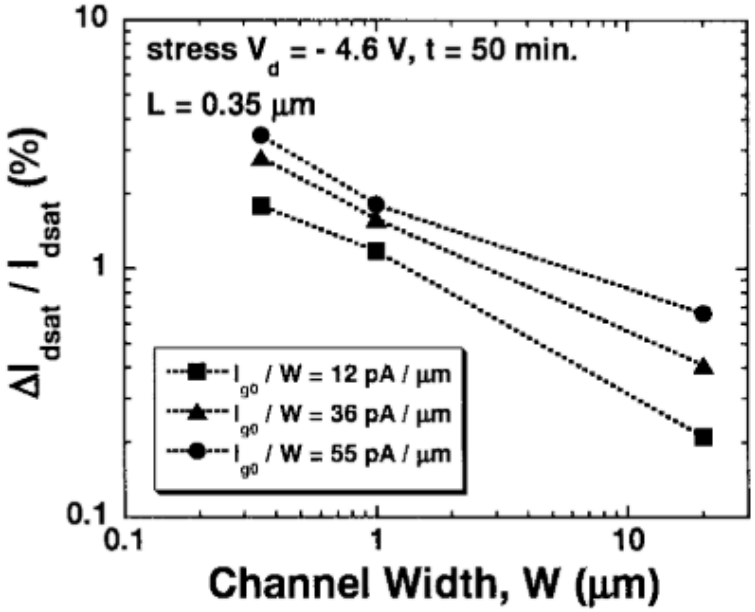


Figure 3.10: Channel width dependence of $\Delta I_{dsat}/I_{dsat}$ as a parameter of initial gate current normalized for W, I_{g0}/W. Narrow width devices show larger degradation for each stress condition.

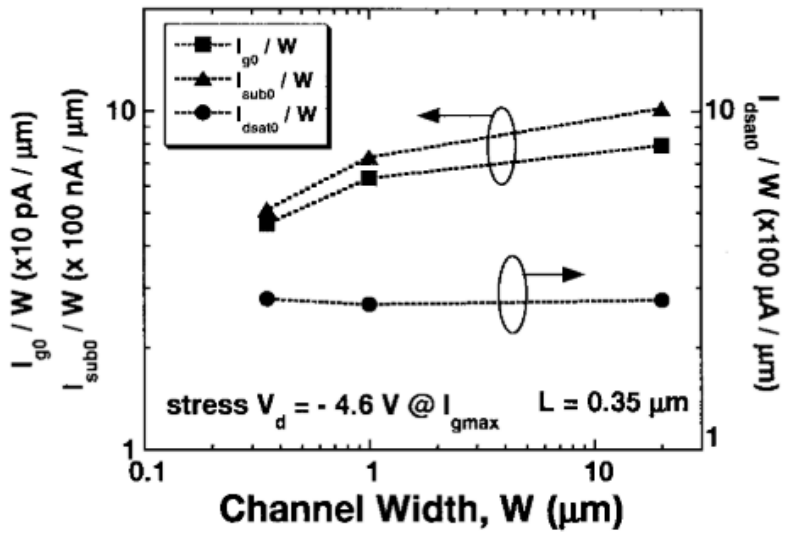


Figure 3.11: Channel width dependence of the normalized gate (I_{g0}), substrate (I_{sub0}), and drain (I_{dsat0}) current.

The above results suggest that the reason for the faster degradation of transistors with narrow channel widths is the large gate current. Therefore, the channel width dependence of both gate and substrate current is evaluated. Figure 3.11 shows the channel width dependence of both gate and substrate currents normalized by the channel width, I_{g0}/W and I_{sub0}/W. The normalized drain currents, I_{dsat0}/W, are also shown in the figure. Both I_{g0}/W and I_{dsat0}/W decrease with channel width decrease, while I_{dsat0} remain the same. Thus, the impact ionization of narrow channel width pMOSFET is lower than that of wide channel width pMOSFET, which is the same as nMOSFET case [101]. If the gate current in narrow width pMOSFET also increases during the stress, the same degradation mechanism in nMOSFET can also explain pMOSFET degradation. Figure 3.12 shows the stress time dependence of the gate current change, I_g/I_{g0} for each channel width. It is evident that the narrower channel width device shows a fast decrease in gate current compared to the wider channel width device, and the gate current decreases monotonically. These results are quite different from those observed in nMOSFETs. Despite the less injected charge Q_{inj}, $(= \int I_g dt)$, for a narrow channel width device, it shows larger degradation. If the electron trapping efficiency in a narrow channel width device is larger than that of a wide channel width device for some reason, this phenomenon can be explained. To confirm this hypothesis, Fowler-Nordheim (F-N) stress test is carried out as a next step.

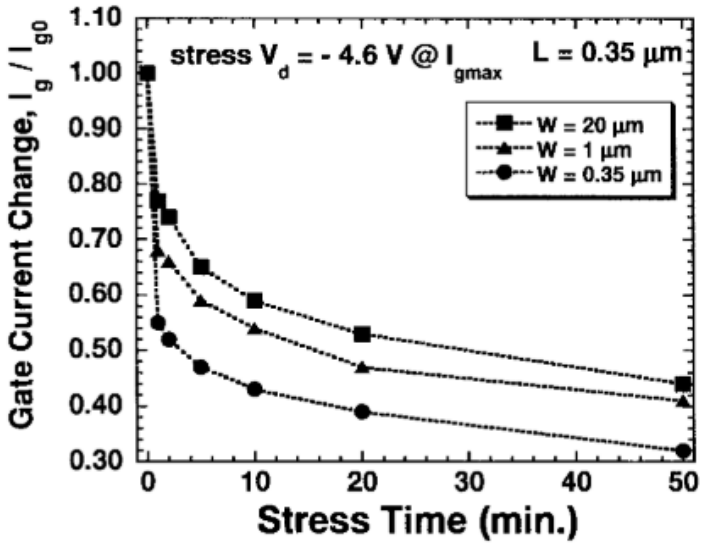


Figure 3.12: Stress time dependence of the gate current change for various channel width devices.

3.3.3 Fowler-Nordheim stress results

A constant gate current stress, known as Fowler-Nordheim stress, is applied to both W = 0.35 and 20 μm devices for the trap generation efficiency evaluation at the condition of $I_g/W = -50$ mA/cm^2. The source, drain, and substrate are connected to the ground level. Figure 3.13 shows the gate voltage shift, ΔV_g, during the stress time. The amount of ΔV_g increases with the stress time for both devices. It is significant that the W = 0.35 μm device shows a large ΔV_g compare to the W = 20 μm device. This result suggests that the W = 0.35 μm device has higher electron trapping efficiency in the gate oxide [102]. There is a report on the mechanical stress effect of the hot-carrier reliability [77]. However, they discussed the gate length dependence of the hot-carrier reliability and did not mention the channel width dependence. They concluded that the degradation caused by the mechanical stress becomes small along with device dimension scaling.

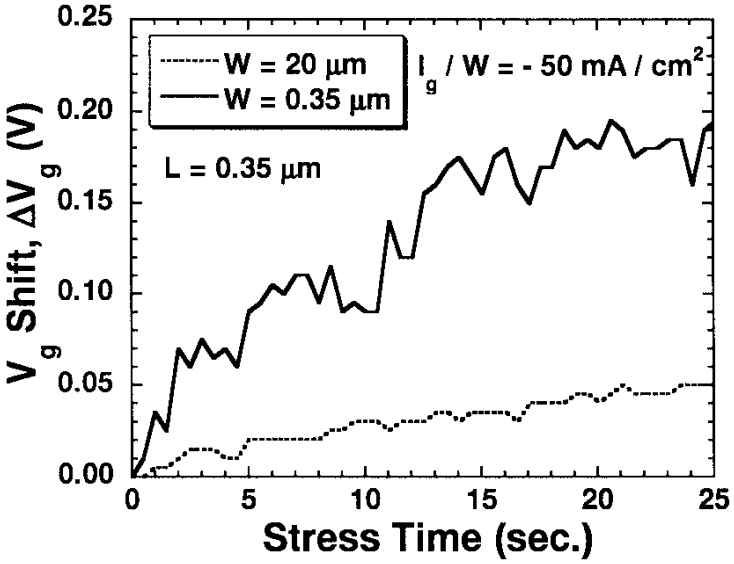


Figure 3.13: Fowler-Nordheim stress results for both wide and narrow width devices. The narrow device shows larger electron trapping efficiency, explained by the large V_g.

There is another study on the device characteristics degradation caused by mechanical stress. Miura *et al.* reported that the TDDB characteristic of the gate oxide degrades by the mechanical stress due to the enhanced electron

trapping efficiency [75]. This mechanical stress is attributed to the structure of the gate electrode, and its effectiveness decreases as the channel width narrows. Therefore, mechanical stress on the gate electrode alone cannot explain the accelerated degradation phenomenon in MOSFETs with narrow channel widths observed in this study. Another possible origin of mechanical stress is residual stress due to the filling material in STIs. It is known that in small dimension MOSFETs, mechanical stress due to the STI filling material can cause crystal defects [64]. Although the MOSFETs used in this study were carried out heat-treated at high temperatures to reduce residual stress, the stress value is not zero at room temperature. Therefore, the residual stress value due to the STI filling material is simulated.

3.3.4 Mechanical stress by STI

Figure 3.14 shows the channel width dependence of the simulated shear stress at the boundary of the gate oxide and the channel along the dotted line a - a' shown in the figure. A simulator used for this estimation was ATHENA by Silvaco. The depth of STI is 0.7 μm, and gate oxide thickness is 6 nm. The y-axis represents the percentage of the area where shear stress shows more than 150 MPa, which is considered a critical criterion, causing TDDB characteristics degradation [75]. The length of the channel is 1 μm. Therefore, when the channel width W is 1 μm, the y-axis of 40% represents more than 0.4 μm^2 of the channel area shows shear stress larger than 150 Mpa. The percentage of the area which shows more than 150 MPa increases as decreasing the channel width, and MOSFET with W = 0.35 μm shows more than 60% of the channel area is occupied by more than 150 MPa stress areas, while wide channel width MOSFET, such as W = 20 μm, shows almost negligible areas. These simulation results support measurement results and suggest that the mechanical stress by the STI filling material is responsible for the enhanced hot-carrier-induced degradation in narrow channel width pMOSFETs.

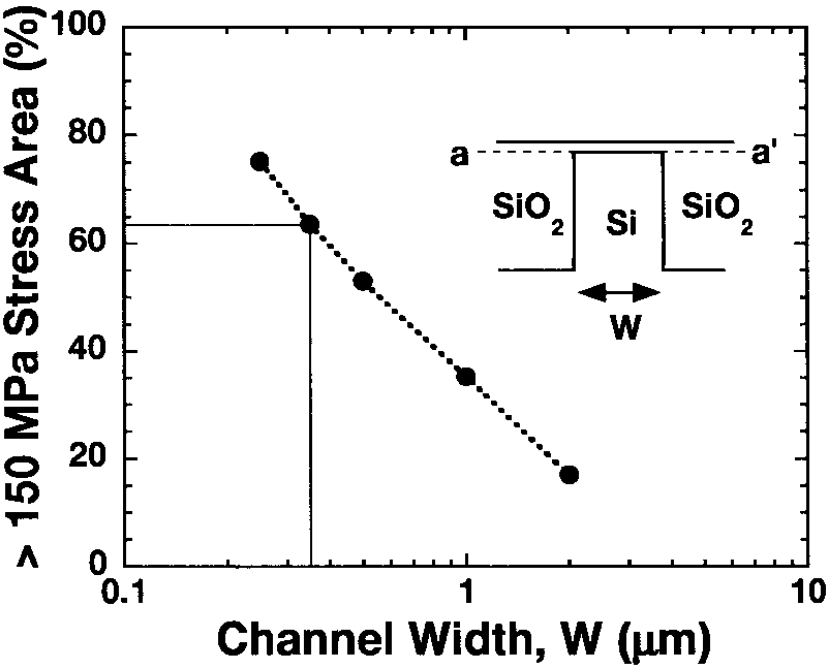


Figure 3.14: Simulated channel width dependence of shear stress at SiO_2/Si interface (shown as dotted line a - a' in the figure). Narrow channel width MOSFETs, such as below W = 0.5 μm, show more than 50% of channel area is occupied shear stress larger than 150 MPa by STI filling material.

Reducing the mechanical stress value at the narrow channel width region is essential for future scaling. In this study, the trench depth is 0.7 μm, which is deep, so the stress caused by STI is considered to be significant. Since it is possible to reduce the trench depth to 0.3 μm without degrading the pressure resistance, the influence will be reduced in the 90 nm generation and beyond. In parallel, a change to a low-stress embedding material should also be considered. O_3-TEOS SiO_2 was used in this study, but combinations of HDP (High Density Plasma) SiO_2, HDP SiO_2 with flowable oxide [103], and polysilazane [104] with wet oxidation are being considered for DRAM and NAND.

3.4 Summary of this chapter

The reliability of narrow channel width MOSFETs used for SRAM cells was studied. Narrow channel width MOSFETs showed larger degradation in hot-carrier reliability. Even though hump characteristics were suppressed, narrow channel width MOSFETs degrade faster than wide channel width MOSFETs. Therefore, reliability should be monitored by using SRAM cell size MOSFETs. Mechanical stress from both STI and gate electrodes may assist this accelerated degradation. To avoid this mechanical stress-enhanced degradation, the amount of stress value should be controlled to lower than 150 MPa. In the case of embedding high-reliable high-capacity SRAM into the 3D NAND, this phenomenon should be considered, and careful material selection and structure design are needed when device structures and/or materials are changed.

Chapter 4

FEOL and MOL process issues and optimization

4.1 Background

The design guidelines and reliability issues of SRAM cell transistors were discussed in the previous chapters to embed high-capacity SRAM into 3D NAND. It is important to adopt a technology compatible with the 3D NAND manufacturing process. The CMOS technology node to be applied depends on the required SRAM capacity and the area allowed in the 3D NAND. Increasing the capacity of embedded SRAM requires a reduction in cell size. As shown in Figure 2.31 in Chapter 2, gate length and contact pitch define the cell size. The gate length scaling should come along with the gate insulator thickness scaling. A thinner gate insulator causes an increase in gate leakage current, leading to an increase in power consumption in SRAM standby mode, especially for high-capacity SRAM. Memory cells are most important in high-capacity memory, and the manufacturing process is optimized for memory cells. Also, because the low cost is important, it is desirable to extend the conventional SiON-based dielectrics as much as possible.

Higher speeds have been required in recent years, even for high-capacity memory. As shown in Figure 1.14 in Chapter 1, the interface speed of NAND lags several generations behind that of DRAM. This comes from the current computing systems discussed in Chapter 1. Unlike logic, the

DDR5 (Double Data Rate 5) generation, which has recently started volume production, will gradually adopt CMOS with high-κ gate dielectrics for high-speed DRAMs [135]. In both DRAM and NAND, the memory cells are fabricated after the CMOS transistor formation. Compared to the standard logic process, the thermal process of forming the memory cells is much higher, such as more than 600°C, than the traditional logic process. Therefore, the high-κ dielectric fabrication process used in advanced logic, such as the gate-last high-κ/metal gate process, cannot be applied to DRAM and NAND in terms of allowable thermal budgets. A high temperature-resistant HK/MG process is strongly desired.

Another approach to reducing the SRAM cell size is contact pitch scaling. The local interconnect (L.I.) technology reduces the number of interconnect layers and relaxes the contact pitch without changing the SRAM cell size. The L.I. simultaneously connects active areas (AAs) and gate electrodes (GEs) in SRAM cell transistors. Therefore, the effect of the L.I. process on the CMOS transistors that consist of the SRAM cell must also be considered.

This chapter identifies new issues to be considered in the FEOL and MOL processes when embedding high-capacity SRAM with 3D NAND and discusses process optimization to realize high reliability.

4.2 Estimation of SRAM capacity and corresponding CMOS technology node

Estimation of the SRAM capacity and the necessary technology nodes are determined as follows. As mentioned in Chapter 1, the capacity of DRAM used in SSDs is 1/1000th of SSD capacity, and in the case of SRAM, only 1/5 to 1/10th of DRAM capacity is required. Assuming a 1 TB SSD, the required SRAM capacity is 100 to 200 MB. 1 TB SSD comprises 16 chips of 512 Gb 3D NAND stacked in a package. Since the maximum 200 MB of SRAM capacity is divided into 16 dies, each chip should have 12.5 MB (100 Mb) of SRAM. Figure 4.1 shows a typical 512 Gb 3D NAND chip [127, 128], consisting of 3bit/cell technology and 64 WL layers, with peripheral circuits formed by CNA technology. The chip size is approximately 130 mm^2. As can be seen in the figure, the cell occupancy is estimated to be approximately 75%, which means that the memory cell area is 100 mm^2 and the peripheral circuit area is 30 mm^2. Assuming that the peripheral circuit area can be accommodated under the memory cells using CUA technology, the remaining 70 mm^2 square millimeters is the area where SRAM can be placed. SRAM typically has a cell occupancy rate of about 50%, with some high-capacity SRAMs reaching 70%. Figure 4.2 shows the number of memory layers and chip size of the 3D NAND, NAND cell array area, and SRAM cell area when the cell occupancy is 50% and 70%. All 3D NAND is 3 bit/cell technology,

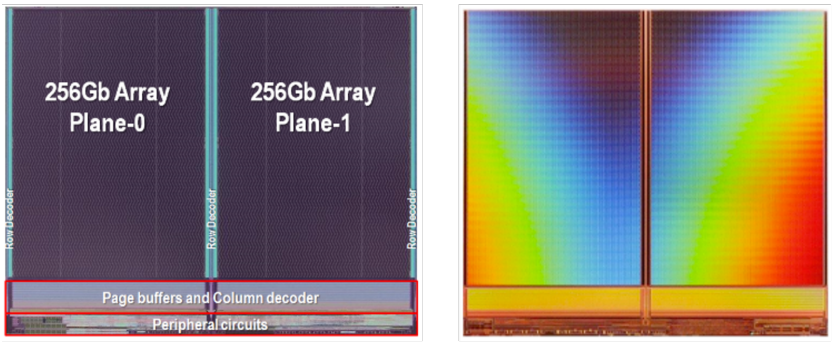


Figure 4.1: Die photos of 512 Gb 3D NAND presented at ISSCC 2017 (a) [127] and (b) [128].© 2017 IEEE

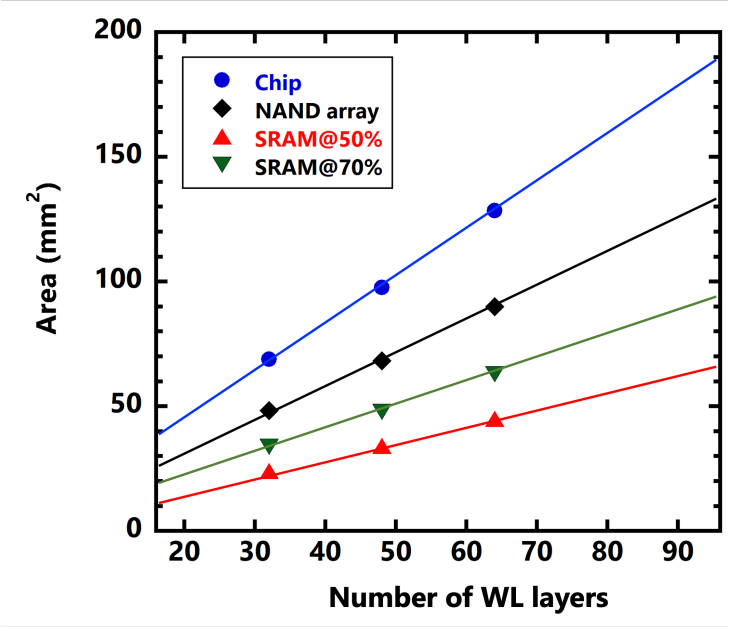


Figure 4.2: 3D NAND memory layer dependence of chip size, NAND array area, and SRAM cell areas.

and 32 layers correspond to 128 Gb, 48 layers to 256 Gb, and 64 layers to 512 Gb capacity. Figure 4.3 shows the SRAM cell occupancy required to achieve 100 Mb of SRAM capacity when using SRAM cells at the 90 to 32 nm node shown in Chapter 2. As can be seen, 90 nm node technology is not able to embed 100 Mb capacity. When using 65 nm node technology, it is necessary to increase the SRAM cell occupancy to 70%. In the case of 45 nm node technology and beyond, 100 Mb capacity is possible even if the SRAM cell occupancy is less than 50%.

For the CiM application, the SRAM capacity requirement is estimated as follows. As mentioned earlier, 3.2 MB of SRAM is required for 6.5 Gb SLC NAND to realize DNN accelerator [26], which was demonstrated by using 64 Gb SLC NAND. Current 512 Gb TLC NAND is equal to 170 Gb SLC NAND in cell capacity. By simply reducing the number of WL layers from 64 to 24, the 64 Gb SLC NAND can be obtained without changing the die size. The required SRAM capacity is about 32 MB, which corresponds to about 260 Mb. Therefore, at least 45 nm node CMOS technology with 70% cell occupancy is required. An alternative approach is a further reduction of

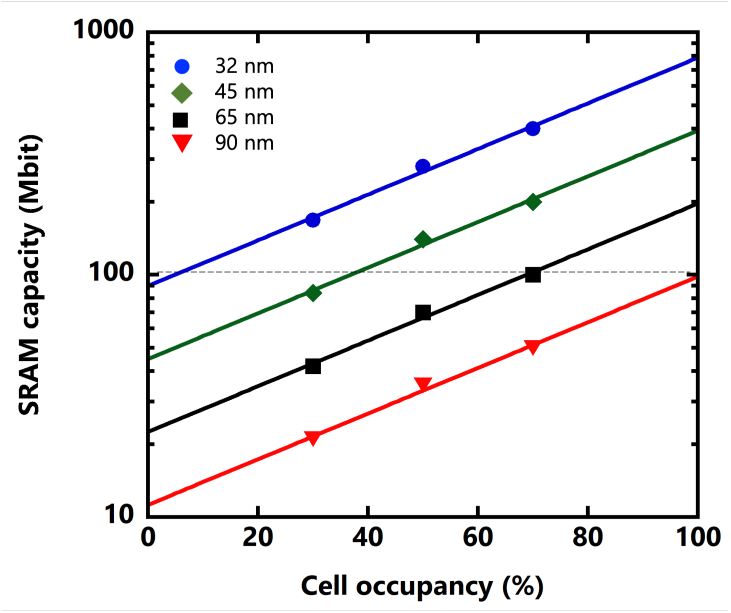


Figure 4.3: 3D NAND memory layer dependence of chip size, NAND array area, and SRAM cell areas.

WL layers. By reducing the number of WL layers down to 12 layers, which corresponds 32 Gb of capacity, the required SRAM capacity becomes half. This enables the adoption of 65 nm node CMOS technology for high-capacity SRAM fabrication and can be a viable option for low-cost CiM accelerators.

One of the obstacle components in MOSFET scaling is the gate insulator. A SiO_2 gate insulator has been used for a long time because of its excellent reliability with low interface state density, regardless of its simple fabrication process. Scaling of the gate insulator thickness improves MOSFET performance. However, it increases gate leakage current and also causes boron penetration from the p^+ poly-Si gate electrode into the substrate as a drawback. Nitridation of SiO_2 [105] was investigated to suppress the boron penetration, and SiON (silicon oxynitride) film was used as a gate insulator below 0.25 μm generation. Gate leakage current also affects standby leakage current in high-capacity SRAM. In the 90 nm generation, the extension of oxynitrides by NO (nitric oxide) gas was investigated [107, 108], and plasma nitridation of SiO_2 was introduced from the 65 nm node [134].

Figure 4.4 shows the equivalent oxide thickness (EOT) of the gate

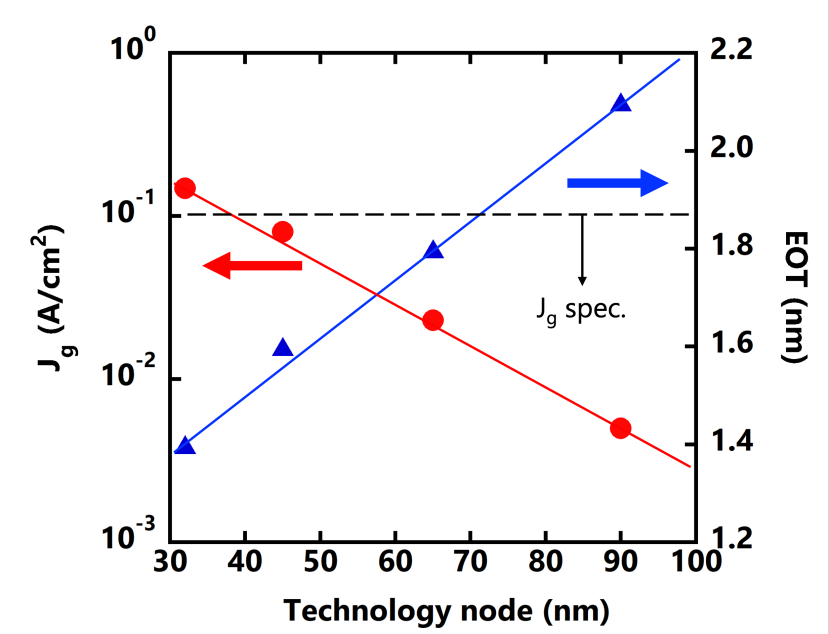


Figure 4.4: Target EOT and J_g for each technology node.

insulator for each technology node [43, 86, 129–131] and gate current density with 100 Mb SRAM leakage current spec. (J_g= 0.1 A/cm^2).

One of the issues of plasma nitrided SiON is the nitrogen penetration into the Si substrate, which degrades both CMOS performance and reliability [133], as shown in Figure 4.5.

A novel SiON process was proposed to overcome this issue [123, 124], as shown in Figure 4.6. The concept is to deposit Si_3N_4 at first, then oxidize that Si_3N_4 and re-nitride in the end. This process realizes low nitrogen concentration at the Si substrate interface.

However, those SiON gate insulators do not satisfy the gate leakage current requirement below 1.6 nm EOT, as shown in Figure 4.7, and the introduction of a high-κ gate insulator, such as HfSiON (nitrided hafnium silicate), is required. Target EOTs and recommended gate insulators are summarized in Table 4.1.

A high-κ gate insulator is a promising candidate to replace the SiON. Many materials were evaluated, and the hafnium-based insulator is becoming the most promising candidate as a high-κ material at present [109–111, 116–118, 120, 121].

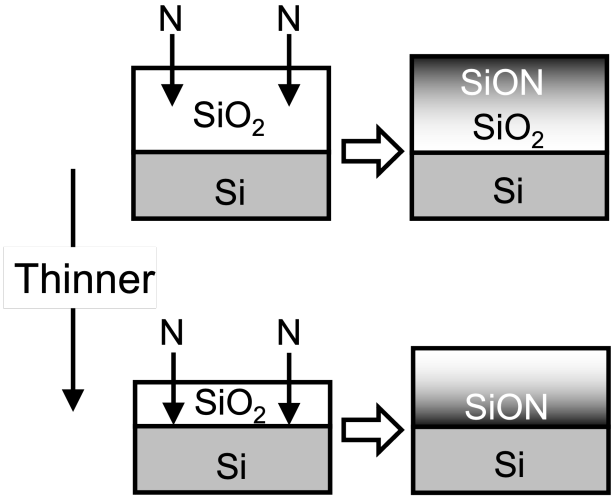


Figure 4.5: Issue of thickness scaling in plasma nitrided SiON.

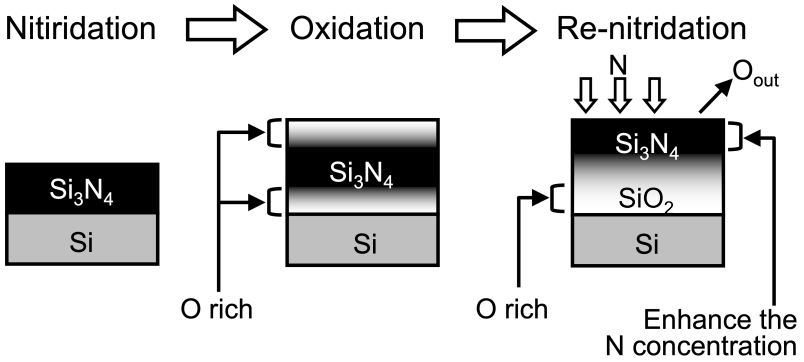


Figure 4.6: Schematic images of novel SiON process and nitrogen profile.

Next, process issues to implementing a high-κ gate insulator in the 3D NAND process will be discussed to embed high-capacity SRAM.

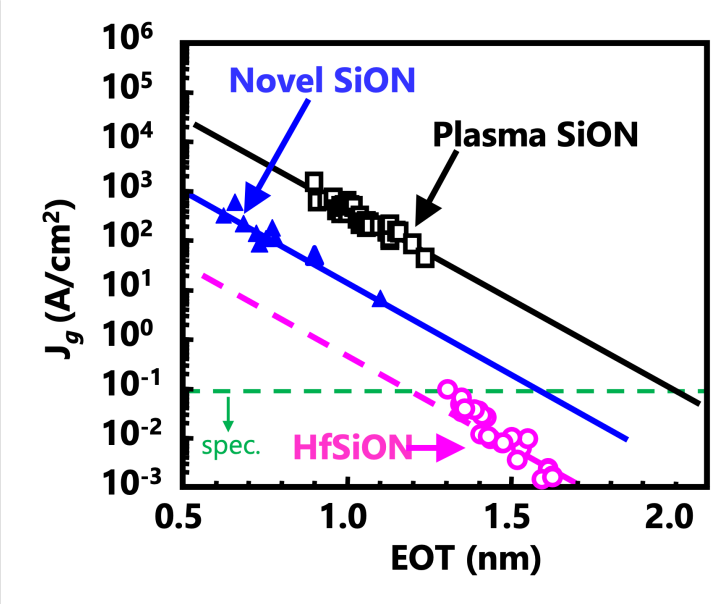


Figure 4.7: EOT dependence of J_g for plasma SiON, novel SiON, and HfSiON.

Table 4.1: Target EOT and recommended gate insulator for each technology node.

Node (nm)	EOT (nm)	Gate insulator
90	2.1	Plasma SiON
65	1.8	Novel SiON
45	1.6	Novel SiON/ HfSiON
32	1.4	HfSiON

4.3 Optimization of gate stack process

The dielectric constant can be varied by changing Hf (hafnium) contents. This means there exists a couple of combinations to realize given EOT [116–118, 120, 132]. One of the impacts of high Hf concentration on MOSFET is undesirable oxidation which causes severe reverse short channel effect [120]. Figure 4.8 (a) shows the schematic image of the undesirable oxidation as indicated bird's beak.

The bird's beak length, L_B, which was measured by a cross-sectional TEM photograph, depends on the Hf concentration and sidewall (SW) materials, as shown in Figure 4.8 (b). Two Hf concentrations, Hf/(Hf+Si) = 30% and 50%, are evaluated under the same EOT of 1.8 nm. In the Hf/(Hf+Si) = 30% with SIO_2 SW structure, the L_B, was 8.2 nm length. On the contrary, Hf/(Hf+Si) = 50% with SiO_2 SW shows 80 nm length, which means L_B extends the entire gate length. This is because the oxygen diffusion in the HfSiON film is much faster than poly-Si oxidation during the post gate etch oxidation process [111–114]. In order to prevent this undesirable oxidation, less post gate etch oxidation with SiN SW protection is effective [115]. Applying the

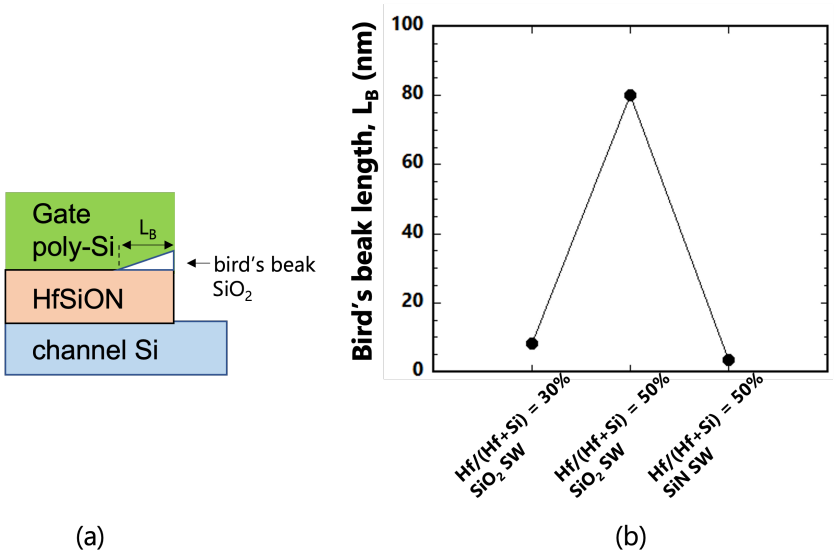


Figure 4.8: (a) Schematic image of undesired oxidation at the gate edge and (b) bird's beak length (L_B) dependence on Hf concentration with sidewall (SW) materials.

SiN SW process, the L_B drastically reduced to below 5 nm. This is the Hf/(Hf+Si) = 50% case, and if Hf concentration is reduced below 50%, it is expected to minimize L_B less than a few nm.

Keep in mind is that the SiN sidewall structure causes large stress under the sidewall region, as described in Chapter 2. Therefore the overlap amount of active area and gate electrode in SRAM cell layout should be re-optimized, such as more than 80 nm overlap with cell height increase, when HfSiON is introduced as a gate insulator. Even if the sidewall is covered by SiN films, Hf concentration needs to be defined carefully. Figure 4.9 shows the simulation result of accumulated charge at the gate edge region. Devices have the same EOT value but have different physical thicknesses by varying dielectric constant. Even in devices with the same EOT, a physically thick insulator shows a weak electric field at the gate edge region, resulting in less charge accumulation at the source region. This causes less gate controllability and higher parasitic resistance. Figure 4.10 shows the Hf concentration dependence on I_{on} on nMOSFET. The gate length is 65 nm, V_d = 1.2 V, and

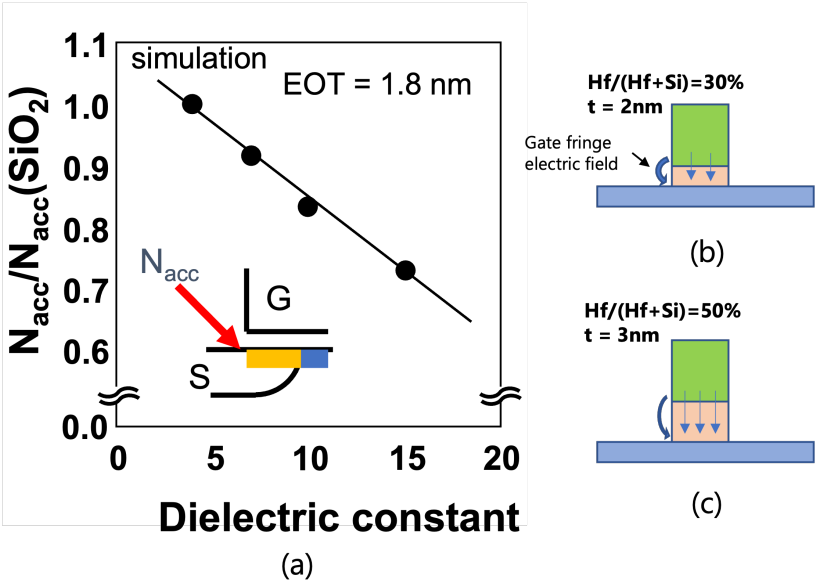


Figure 4.9: (a) Simulated Hf concentration dependence of accumulated charge at the gate edge region. Under the same EOT = 1.8 nm condition, Hf/(Hf+Si) = 30% shows more gate fringe electric field (b) compare to Hf/(Hf+Si) = 50% (c) because of the difference of physical thickness.

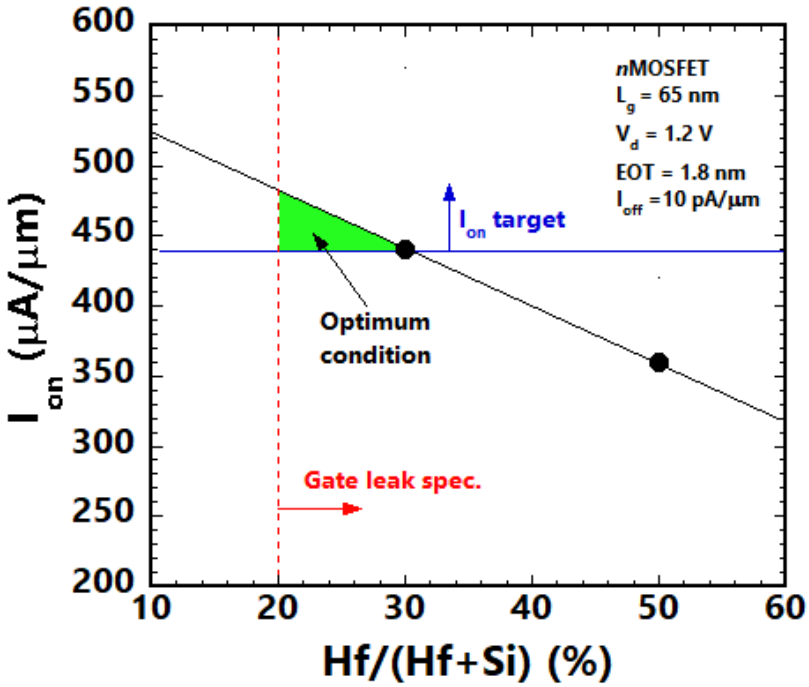


Figure 4.10: Hf/(Hf+Si) dependence on nMOSFET I_{on}.

I_{off} criteria is 10 pA/μm. As mentioned previously, higher Hf concentration shows lower I_{on}, and less than Hf/(Hf+Si) = 30% is required to achieve the I_{on} target. However, less than 20% does not satisfy the gate leakage current specification. Therefore, optimum Hf concentration should be between 20% and 30%. Figure 4.11 shows Hf/(Hf+Si) concentration dependence on hot carrier lifetime of nMOSFET whose gate length is 65 nm. In order to satisfy a 10 year lifetime, Hf/(Hf+Si) should be less than 25%. From the above studies, it is concluded that Hf concentration should be as low as possible at a given EOT to realize both high performance and better reliability.

In order to improve thermal stability and prevent boron penetration from the p^+ poly-Si gate electrode, nitrogen incorporation is required for HfSiO film. To prevent boron penetration, more than 20% nitrogen concentration is required [116]. The dielectric constant, κ, increases with increasing the nitrogen concentration. It is difficult to incorporate more than 50% nitrogen [132]. Considering the I_{on} of MOSFET, as shown in Figure 4.9, the physical thickness of HfSiON should be as thin as possible. Therefore, target

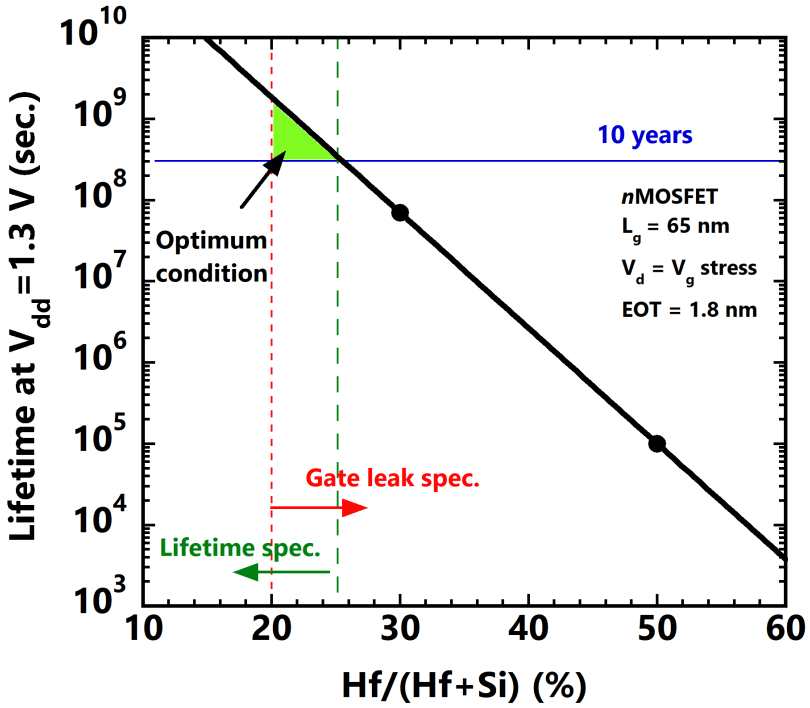


Figure 4.11: Hf concentration dependence on hot carrier lifetime of L = 65 nm nMOSFET.

Hf and nitrogen concentration should be set by taking into account process variability. Table 4.2 summarizes proposed Hf and nitrogen concentration for both 45 and 32 nm nodes. In addition to HfSiON, HfO and Al_2O_3 are also being considered as high-κ materials. However, it has been pointed out that these materials need to be improved in terms of thermal resistance [136], and based on the results of past studies [116, 120, 121], HfSiON is considered appropriate at this time. Various research institutes continue to study materials with high heat resistance, and the optimal material and process changes should be made while keeping a close eye on the results of these studies. As for SRAM to be embedded in 3D NAND, it is appropriate to start up the technology by applying SiON in the 90 nm generation technology, extend the life of SiON to the 65 nm generation, and then use HfSiON or alternative high thermal-resistant high-κ material at the time of embedding in the 45 nm generation SRAM.

Table 4.2: Proposed Hf and nitrogen concentration for 45 and 32 nm node.

Node (nm)	EOT (nm)	Hf/(Hf+Si) (%)	Nitrogen (%)	Dielectric constant (κ)	Physical thickness (nm)
45	1.6	22 (20 ~ 25)	30 (20 ~ 40)	8	3.3
32	1.4	22 (20 ~25)	30 (20 ~ 40)	8	2.9

4.4 Implementation of local interconnect

4.4.1 Advantage of local interconnect

Chapter 2 shows the difference between SRAM and NAND manufacturing processes. Since stand-alone SRAM uses tungsten local interconnect (L.I.) and 3D NAND uses tungsten interconnects, applying the tungsten L.I. process to 3D NAND seems feasible.

Firstly, the cell layout advantages of applying L.I. are explained. Figure 4.12 shows the cell layout of an embedded SRAM and the necessary interconnect layers. As seen in the figure, the first layer of interconnect is used as a writing node connection to form flip-flops in the cell and as a landing pad to lift to the upper layer interconnect. The bit lines and power lines are formed using the second interconnect layer, and the word-lines are formed using the third interconnect layer.

The cell layout of a stand-alone SRAM using L.I. is shown in Figure 4.13. Unlike contacts, local wiring can be routed over STI. Therefore, flip-flop node connections can be formed with only one layer of L.I. Furthermore, contacts brought up to the upper interconnect layer can be drawn on the active area (AA) and the STI. Therefore, there is more freedom in the routing of the

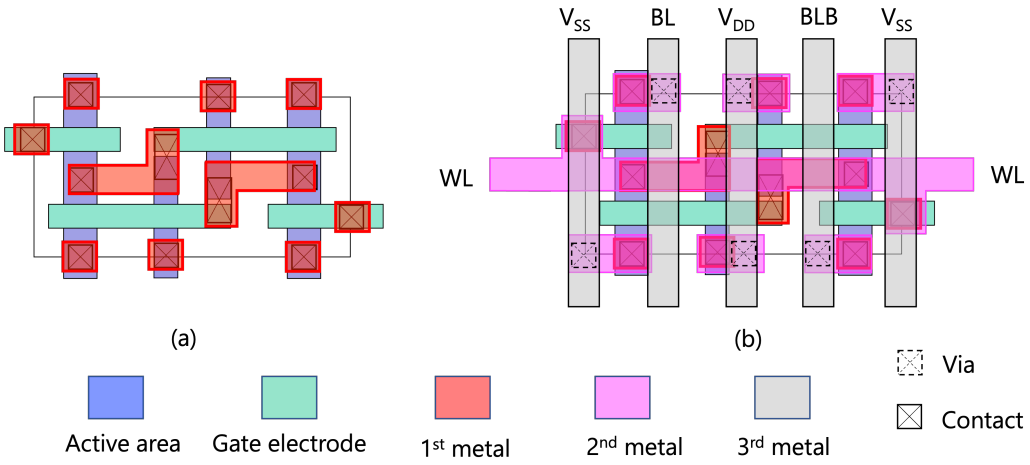


Figure 4.12: Cell layout of embedded SRAM and interconnect layer configuration. (a) Only AA, GE and L.I. layers are shown. (b) All layers are shown.

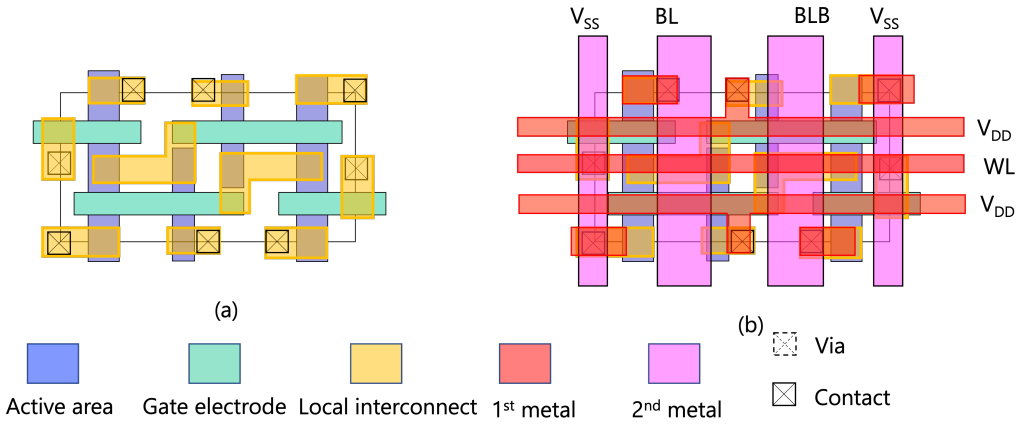


Figure 4.13: Cell layout of stand-alone SRAM and interconnect layer configuration. (a) Only AA, GE and L.I. layers are shown. (b) All layers are shown.

upper interconnect layer. The same interconnect layer configuration as in embedded SRAM can be used, but as shown in Figure 4.13 (b), it is possible to place the V_{DD} power lines parallel to the word lines in the first layer of interconnect. This relaxes the pitch of the bit lines formed by the second layer of interconnect and improves the RC (Resistance Capacitance products) delay of the BLs.

Figure 4.14 shows both the top and cross-sectional schematic view of SRAM cell with tungsten L.I. L shape L.I. connects both cell nMOSFET and pMOSFET and also connects node gate electrode. Junction leakage is a concern, since the L.I. is routed over the STI region. Moreover, the junction depth between n^+ and p^+ source/drain regions are different. Over etching of L.I., hole, or post etching treatment may cause junction leakage. Therefore, etching optimization is needed.

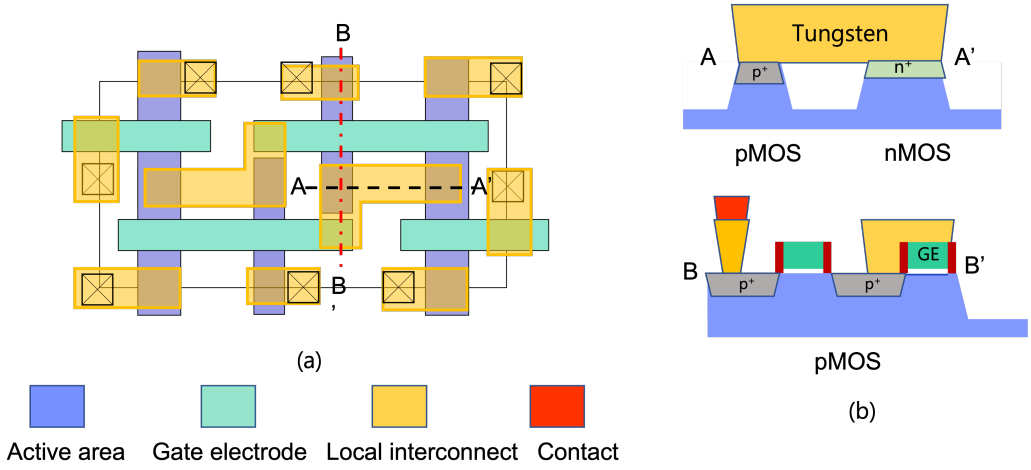


Figure 4.14: Schematic image of L.I. chain test structure. (a) Top view of cell. (b) Cross-sectional images of both directions along A−A' and B−B' shown in (a).

4.4.2 Process optimization

Figure 4.15 is a schematic image of the test structure of the L.I. chain. 50 k chain is used for both contacts to (a) n^+ and (b) p^+ regions. The size of the L.I. hole is 0.25 μm×1.1 μm.

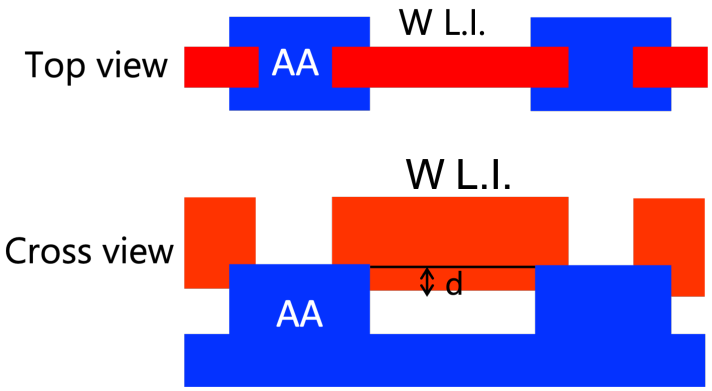


Figure 4.15: Schematic image of L.I. chain test structure of both top and cross views.

Figure 4.16 shows chain yield for both n^+ and p^+, with final STI height as a parameter. It is clearly shown that the bottom of L.I. should be above the active area surface. Cross-sectional SEM photos are also shown in

Figure 4.17. Since the contact process of NAND is different and contact ion implantation will be performed, those process impact on MOSFET needs to be considered

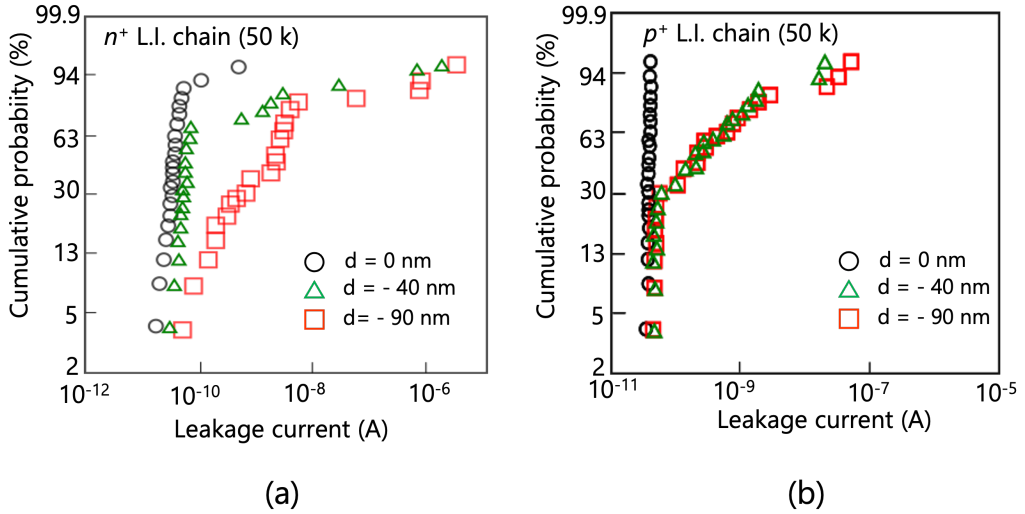


Figure 4.16: Cumulative probability of L.I. chain for both (a) n^+ and (b) p^+ contacts as a parameter of STI height.

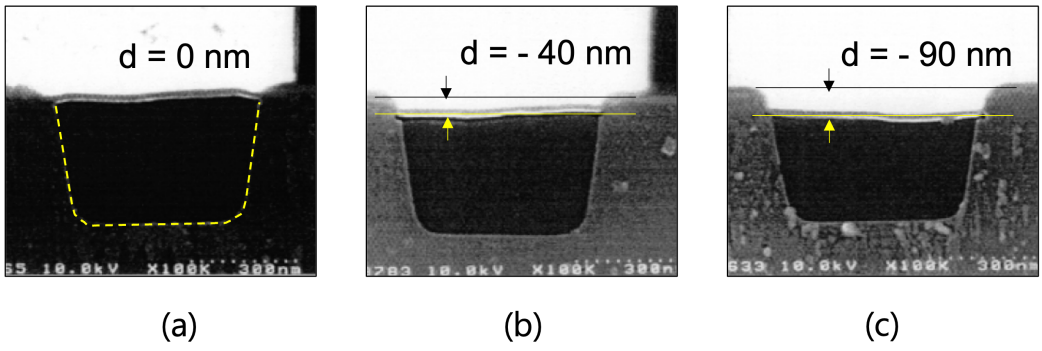


Figure 4.17: Cross sectional SEM images of fabricated L.I. with different STI height, d, (a) 0 nm, (b) − 40 nm, and (c) − 90 nm.

Figure 4.18 shows assumed process flow and schematic images of critical processes that may affect MOSFET characteristics. When the L.I. process is applied to 3D NAND, additional ion implantation processes with annealing are added after the L.I. contact etching required for contacts of peripheral circuits because both L.I.s and peripheral contacts are formed at the same time, as mentioned earlier. The cleaning process (a), activation annealing after the contact doping (b), and barrier metal layer formation for L.I. (c) may affect MOSFET characteristics and SRAM yields. In process step (a), if the sidewall of MOSFETs has a triple-layer sidewall structure, the SiN portion of the sidewall will be lifted off during the wet cleaning process that contains hydrofluoric acid. This causes contact failure due to the peeled-off SiN sidewalls. Even if the sidewalls are not lifted off, there is a concern about the barrier metal layer breaking because the oxide film portion is significantly recessed. The step-out of the barrier metal layer causes a decrease in reliability and junction leakage failure due to the diffusion of tungsten into the substrate. Another concern is the undesired oxidation when the poly-Si/HfSiON gate stack process is adopted to the SRAM cells. As described in the previous section, oxygen diffuses through the HfSiON much faster than the poly-Si oxidation. It diffuses more than one micrometer [111, 113]. Even if the optimized sidewall formation process prevents undesired oxidation, the contact doping activation anneal process may cause undesired oxidation to the entire gate length in the SRAM cells. Because the gate length in the SRAM cell, as shown in Figure 2.31, is less than one micrometer. From process robustness and MOSFET performance points of view, a dual-layer sidewall structure is recommended for the SRAM embedded in 3D NAND, even if the cell size increases.

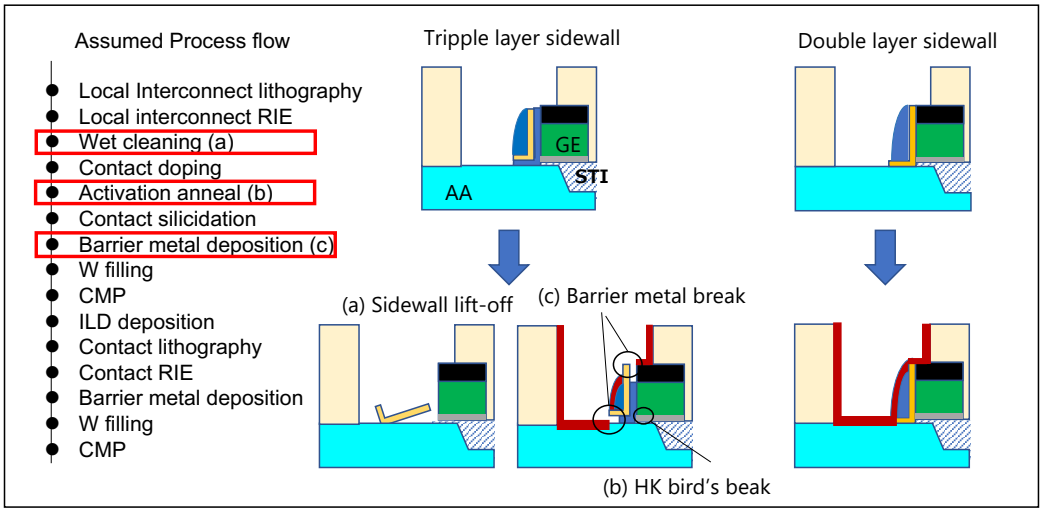


Figure 4.18: Assumed process flow and critical process (a) - (c) that may affect MOSFET characteristics.

4.5 Summary of this chapter

The available capacity of SRAM to be embedded in 3D NAND was estimated with the required CMOS technology node. When replacing DRAM used in SSDs, 100 Mb per 3D NAND die is required, which is not feasible with 90 nm node SRAM cell size. At least 65 nm node technology with 70% cell occupancy is required. 45 nm generation technology with 50% cell occupancy is also feasible. In addition, it was found that SiON dielectrics can be used up to 65 nm node, but from 45 nm node, high-κ, such as HfSiON, need to be used. When Hf is employed as a gate dielectric, it is clear that the Hf concentration should be adjusted in the range of 20 - 25% with nitrogen concentration between 20 and 40% to satisfy both transistor performance and reliability. The target concentration and physical thickness were proposed.

The tungsten L.I. shows the advantage of reducing the number of interconnect layers to form the SRAM cell. It also has the possibility to relax bit-line pitches with wider bit lines that improve SRAM access speed. It is important that when L.I. is applied, it is necessary to control the STI height before opening the contact holes to prevent the junction leakage current increase. From the process robustness point of view, MOSFET should have a dual-layer sidewall structure, and the overlap amount of the active area and gate electrode needs to be more than 80 nm. If these guidelines are followed, it will be possible to embed high-capacity SRAM with 3D NAND from 90 nm node technology to at least down to 45 nm node.

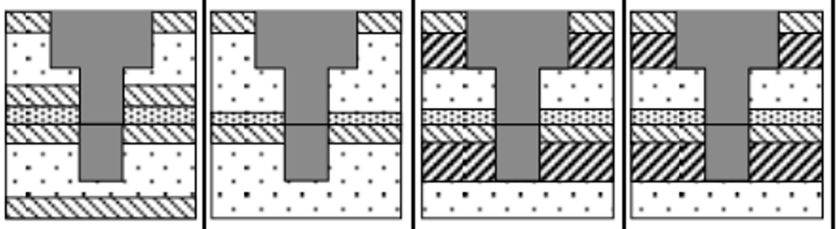
Chapter 5

BEOL process issue and optimization

5.1 Background

In high-capacity memory, the redundancy technique is used to improve memory yields. This is because the memory cell area is large and affected by tiny dust and defects, so replacing failed bits with good bits is necessary. In 3D NAND, information on redundant bits is stored in NAND memory cells. On the other hand, the fuse blow method has long been used in high-capacity SRAMs. In order to implement the redundancy module of high-capacity SRAMs in 3D NAND without affecting the 3D NAND process and circuitry as much as possible, it is important to examine the applicability of the conventional fuse blow process. Generally, the upper metal layer is cut by a laser-blown process in the case of aluminum interconnect because the melting temperature of aluminum is 660.3°C and easy to melt by a laser. However, the melting temperature of copper is 1085°C and raises concern about wheather current laser blow tools can be applied or not. The use of silicide agglomeration of poly-Si gate [137] and non-silicided poly-Si gate blown process by using electromigration was proposed [138] and used instead of laser fuse blow process. However, these processes do not apply to the high-capacity SRAM embedded in the 3D NAND because it uses a polymetal gate electrode.

Table 5.1: Example of BEOL structure roadmap for 130, 90, 65 and 45 nm node technologies [142].© 2004 IEEE

Node	130 nm	90 nm	65 nm	45 nm
Structure				
Cap	SiO_2	SiO_2	SiO_2	SiOC(2.9)
Trench	SiOF(3.4)	SiOC(2.9)	PAr(2.65)	PAr(2.2)
via			SiOC(2.5)	SiOC(2.2)
Bottom	SiO2			
Stopper	SiN(7.0)	SiCN(4.9)	SiC(3.5)	SiC(3.5)
keff	*4.0*	*3.3*	*3.0*	*2.7*

A via fuse process using electromigration of vias has also been proposed [139]. However, it cannot be applied when only one copper interconnect layer is used. Although writing redundancy information to the NAND cells and using that information to replace the good bits is possible, the redundancy method that does not affect the NAND circuit was strongly desired in terms of test cost and capital investment. It is also important to utilize the conventional redundancy scheme for SRAM, such as circuit design and fuse blow infrastructures. Therefore, an investigation of the laser blow process for copper fuse is needed.

As discussed in Chapter 1, high-capacity memories do not use the multilayer interconnect process used for advanced logic devices. Fewer interconnect layers and materials are used to keep manufacturing costs as low as possible. The L.I. described in the previous chapter is one such example. Therefore, the low-κ interlayer dielectric used in the conventional logic process is not applicable. However, it is important to consider the use of low-κ interlayer dielectrics to accommodate interface speed increase in the future along with the CMOS technology scaling, as shown in Table 5.1 [142].

This chapter discusses the challenges and solutions for applying the fuse blow process to the interconnect process in 3DNAND to embedding high-capacity SRAM.

5.2 Copper fuse experiments

5.2.1 Barrier metal evaluation

The fuse blow process by laser uses thermal energy. Therefore, using materials with low melting temperatures is preferable. The tantalum (Ta) is commonly used for copper interconnects as a barrier metal material. However, its melting temperature is 2985°C, which is much higher than copper (1085°C). The titanium (Ti) is also the candidate for barrier metal, and its melting temperature is 1668°C, which is less than the Ta. Moisture induced via failure (MIVF) is one of the issues for interconnects which use porous low-κ material [140]. The dry etching process causes damage to the low-κ film, and the damaged region absorbs moisture. This moisture oxidizes the barrier metal and induces failure known as SIV (stress-induced voiding) [141]. Since the oxidation resistance of Ta is less compared to that of Ti, it is beneficial to investigate Ti as a barrier metal material. Figure 5.1 shows electromigration test results for both Ta and Ti barrier layers. It should be noted that Ti shows superior reliability compared to Ta with no SIV degradation [141]. Moreover, Ti is less expensive than Ta, which is another advantage of the Ti barrier layer for high-capacity memories. Therefore, Ti barrier metal is chosen for fuse evaluation.

5.2.2 Fuse blow experiments

Figure 5.2 shows the schematic images of the evaluated test structure. The fuse metal width is varied from 0.6 to 1.0 μm with 0.2 μm step under the same fuse pitch of 3.5 μm. Fuse pitch is also varied from 2.5 to 4.5 μm with 0.5 μm step under the same fuse width of 1.0 μm. The thickness of cover layer SiO_2 is varied from 0.2 to 0.6 μm with 0.2 μm step, and the thickness of fuse metal is varied from 0.6 to 1.2 μm with 0.2 μm step, respectively. Figure 5.3 shows results after the fuse blow process in which the fuse thickness is 0.6 μm, and the cover SiO_2 thickness is 0.2 μm, which are considered the most blowable condition. Two different laser wavelengths, (a) $\lambda = 1321$ μm and (b) $\lambda = 1047$ μm, were evaluated. The significant leakage current variation observed in machine B with the pitch of 4.5 μm is caused

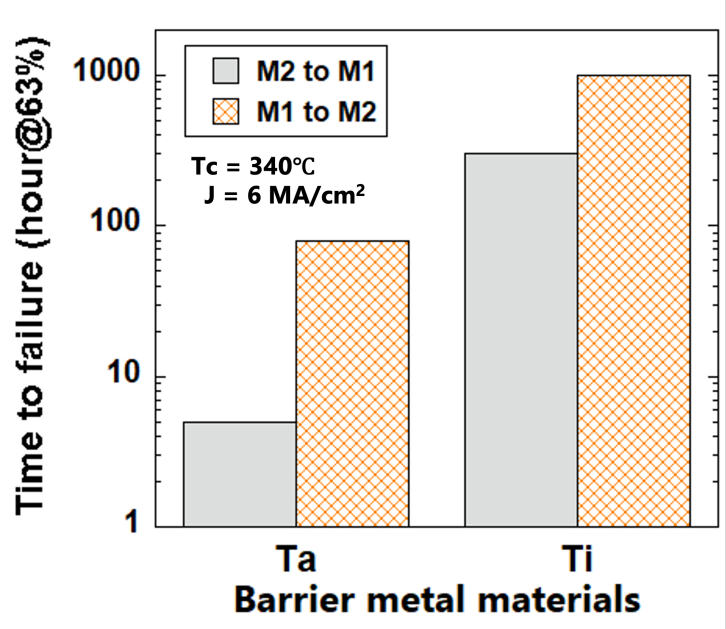


Figure 5.1: Barrier metal material dependence of electromigration. Ti shows better characteristics compare to the Ta.

by alignment inaccuracy during the laser blow process. Machine A with $\lambda = 1321$ μm shows a better yield compare to machine B with $\lambda = 1047$ μm. It is considered that this difference comes from the absorption efficiency difference. This is the preliminary result and laser blow conditions, such as power and pulse width, need to be optimized further. The thickness of the upper copper interconnect layer uses thicker copper, such as more than 1 μm. Therefore, more laser power will be needed and may cause more damage to the interconnect layers. This issue also needs to be addressed. Another issue with using the laser blow process for copper interconnect is the leakage current increase after the fuse blow process by copper corrosion. Figure 5.4 shows I-V characteristics of before and after fuse blow processes. In order to prevent copper corrosion, BTA (Benzotriazole) treatment was applied after the blowing process. However, small leakage currents were observed in the blown fuses and showed no BTA treatment dependence. It is considered that the BTA did not penetrate sufficiently into the blown area and that scattered copper particles corrode and form leakage current paths.

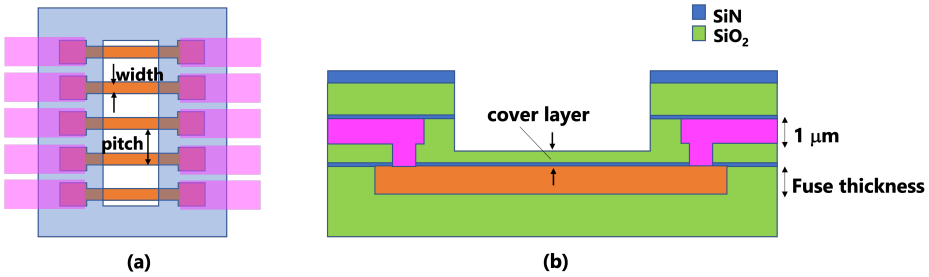


Figure 5.2: Schematic images of test structure of fuse blow evaluation, (a) top view and (b) side view.

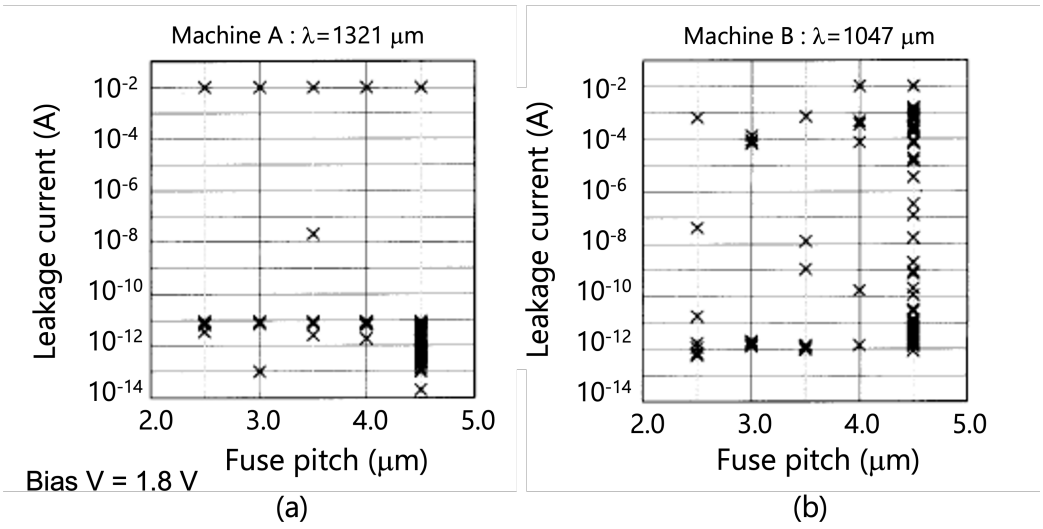


Figure 5.3: Pitch dependence of fuse blow yield. Wave lengths of the laser are (a) $\lambda = 1321$ μm and (b) $\lambda = 1047$ μm, respectively.

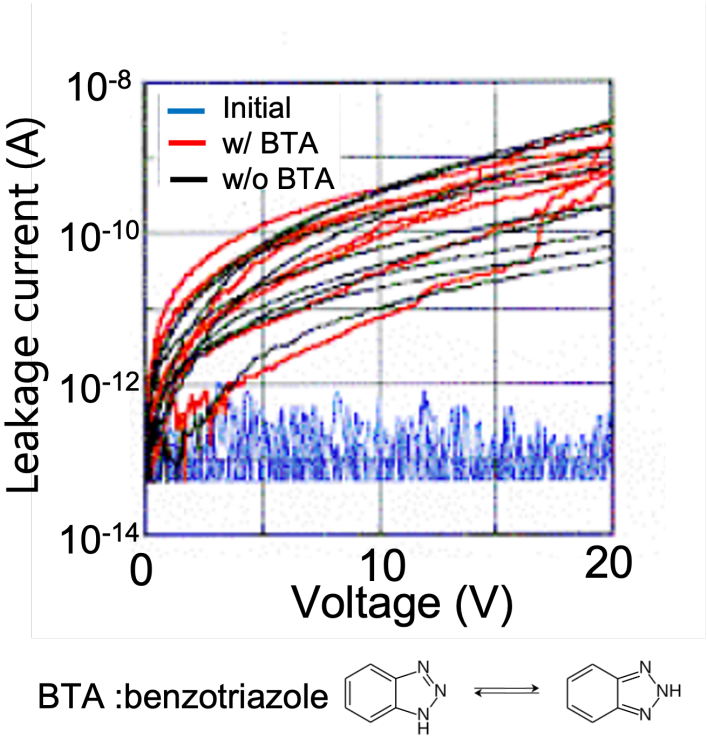


Figure 5.4: Leakage current characteristics of blown fuse with and without BTA treatment. Initial (before blow process) leakage characteristics is also shown.

5.2.3 Study on low-κ dielectrics

It is important to consider using of low-κ interlayer dielectrics (ILD) to accommodate future interface speed increases. Reducing the dielectric constant of ILD is a trade-off between mechanical strength and dielectric property. In particular, it is difficult to reduce the dielectric constant of the dielectric film material between interconnect layers because it is subject to mechanical damage caused by the CMP (Chemical Mechanical Polishing) process. Therefore, a hybrid method is effective in which an insulating film with a relatively high dielectric constant and high mechanical strength is used in the interconnect layer, and a material with a low dielectric constant is used in the via layer to the extent possible, as shown in the Table 5.1 [142]. One is a homogeneous ILD structure, and the other is a hybrid ILD structure. Homogeneous SiOC (carbon-doped silicon oxide) ILD structure with the via-first process is widely used; however, etch profile control of both trench and via is difficult due to the same etching rate of low-κ film. On the contrary, a hybrid scheme has an excellent controllability of both trench and via profile because it uses different low-κ materials at interconnect and via portion. As shown in Figure 5.5 (a), SiOC (κ = 2.3) is used for via portion while SiOC (κ = 3.0)/PAr (κ = 2.3) is used between interconnects [140]. The etching rate of Polyarylene ether (PAr) is faster than that of SiOC (κ = 2.3) and enables a sharp profile for both trench and via. The problem with using PAr is its porosity and weakness during the CMP process. Therefore, SiOC (κ = 3.0) film is stacked on the PAr layer as a CMP protecting layer. Though the κ value of the SiOC capping layer is higher than that of PAr, effective κ value, κ_{eff} = 2.7 is obtained by this hybrid scheme. Figure 5.5 (b) shows the cumulative probability of via resistances which were measured by Kelvin structure. Three SiOC materials whose dielectric constant is the same κ = 2.3 but different material components were evaluated. It is important to note that, as shown in Figure 5.5 (b), even materials exhibiting the same κ value show different failure rates. The SiOC-C shows large via resistance distribution, and some samples showed more than 100 Ohm/via, and those samples were removed from this figure. It is known that low-κ dielectric films lose carbon in the film due to processing damage, resulting in

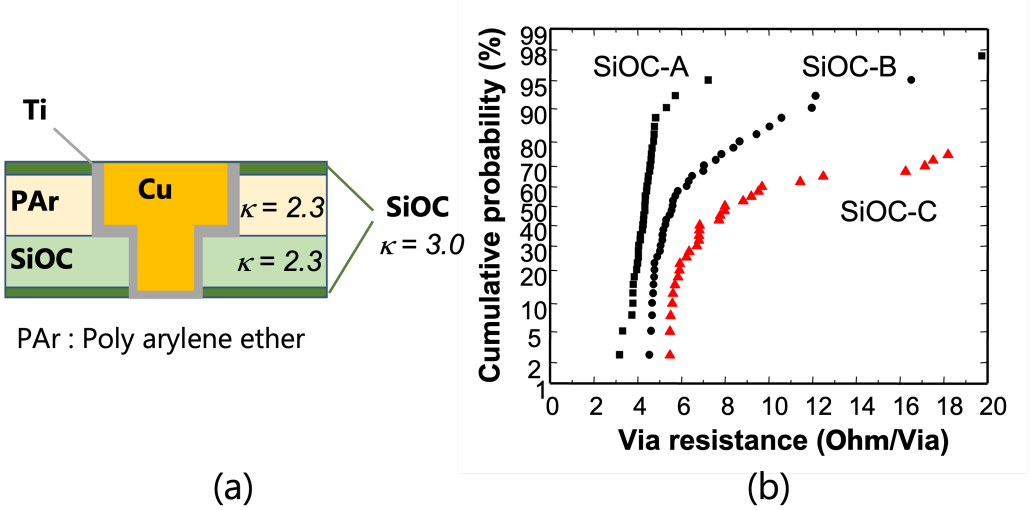


Figure 5.5: (a) Low-κ ILD configuration of the test structure. (b) Cumulative probability of via resistance of three SiOC materials, A, B and C that were measure by Kelvin structure. Samples more than 100 Ohm/via were removed in this figure.

a loss of strength. In this study, CMP damage was prevented by depositing SiO_2 film on the top of the low-κ film. Two types of processing damage can be considered: RIE (reactive ion etching) damage and ashing damage during resist stripping. It has been proposed to recover from this damage by applying treatment process [141, 142]. This damage treatment [143] is said to be effective in recovering from ashing damage, which is less damaging than RIE and is also used in this study. Therefore, it can be said that the material-dependent results in this study are dominated by the damage during the RIE processing, not by the ashing damage. By applying the above process optimizations, the BEOL process shown in Table 5.1 can be applied to high-capacity SRAM down to a 45 nm node. However, it is considered that the laser blow process will cause more damage in the low-κ materials, and a redundancy system other than laser blow of copper needs to be combined, such as anti-fuse [144], to realize the high-reliable high-capacity SRAM. With careful selection of low-κ material and new fuse systems such as anti-fuse, high-reliable high-capacity SRAM can be embedded into 3D NAND down to 45 nm node.

5.3 Summary of this chapter

The copper fuse process was evaluated for high-reliable high-capacity SRAM embedded in the 3D NAND. Using past architecture and infrastructures, a copper fuse blown by laser was desired. A Ti barrier metal was chosen instead of Ta, which has a low melting temperature, considered easy for laser blow. Two different laser wavelengths were evaluated, and the longer wavelength showed better blowing yield because of the difference in laser absorption efficiency. However, the blown fuses showed leakage current originating from the damage and copper corrosion.

Low-κ material was also evaluated to mitigate the future speed requirement. Even with the same dielectric constant, material composite affect the reliability. This degradation was caused mainly by the trench RIE process damages. By combining the damage recovery treatment process and robust low-κ material, BEOL process technology down to 45 nm node can apply to the high-capacity SRAM embedded in the 3D NAND. At that time, a new redundancy fuse process, such as anti-fuse, needed to be introduced.

Chapter 6

Remaining issues for embedding high-reliable high-capacity SRAM

6.1 Background

In the previous chapters, process technologies required to embed high-capacity SRAM in 3D NAND have been discussed. The capacity of SRAM needed to replace the DRAM used for SSD, which is necessary for the single-packaging of SSD realization, can be achieved by using 65 nm node technology and beyond. On the other hand, as discussed in Chapter 1, the realization of computing in memory, a new energy-efficient computation architecture that will become increasingly important in the future, will require even higher capacity SRAM. Figure 6.1 shows the maximum SRAM capacity in each technology node and the corresponding SLC-NAND capacity. The criteria are the same as discussed in Chapter 4. The die size of 3D NAND has assumed to be 100 mm^2, and the available SRAM cell area is 35 mm^2. Cell occupancy of SRAM, 70%, is used for this estimation.

Based on the discussion in Chapter 4, a 512 Gb TLC 3D NAND has an SLC equivalent capacity of approximately 170 Gb. A 3.2 MB of SRAM capacity per 6.5 Gb SLC 3D NAND would require approximately 84 MB of

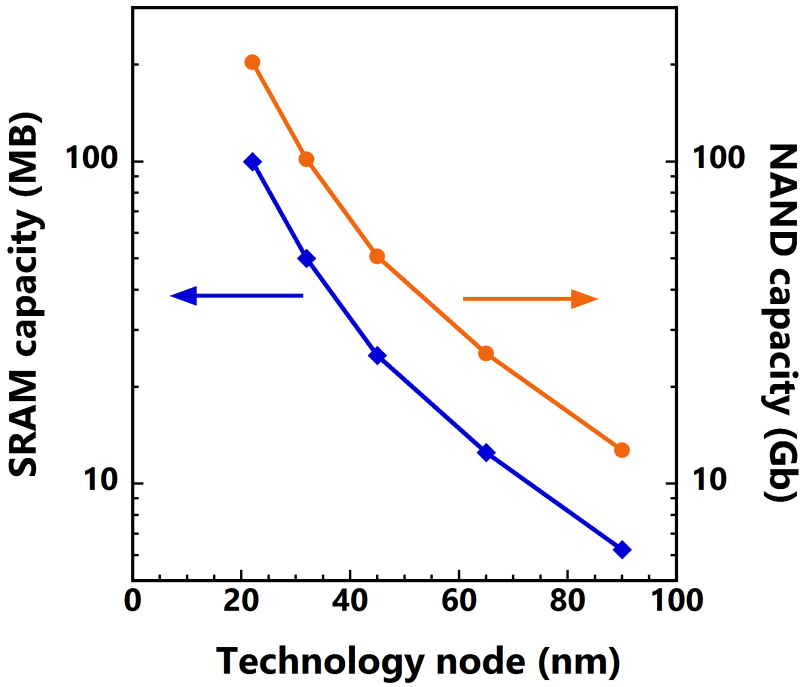


Figure 6.1: Available maximum SRAM capacity and corresponding 3D NAND capacity for each technology node based on the assumption used in Chapter 4.

SRAM. This is more than that used in the latest CPU [16]. Although further study is needed to determine the required SRAM capacity relative to NAND capacity, beyond 32 nm node technology is needed at least. Conversely, the capacity of SLC 3D NAND can be lowered according to the available SRAM capacity, and the number of WL layers can be reduced. For example, when using a 45 nm node SRAM, the required SLC NAND capacity is 51Gb, and since 128Gb SLC 3D NAND has 64 WL layers, it is possible to provide a low-cost CiM die by reducing the number of WLs by about 1/3, or 21 layers. Since beyond 22 nm node technology requires gate-last HK/MG process for MOSFETs, it is difficult to embed with 3D NAND, and other processes such as die stacking technology, either hybrid bondig [145–148] or TSV stacking [149], are needed. Although wafer-to-wafer bonding is preferable in process cost and throughput, selecting only good dies and bonding them together will effectively reduce product costs. Since multiple

dies are currently stacked and wire bonded to form a single package NAND, a thin-film die stacking technology has been established, and die-to-die or die-to-wafer bonding is a candidate for the diversion of this method. There are studies of issues on wafer-to-wafer bonding [145–148]. However, those focus on the characteristics and reliability of adhesion, such as TSV or copper pad. This chapter discusses an additional study requirement of process technology to enable future 3D integration to realize a future one-package server for the CiM systems

6.2 Technologies and issues for 3D integration

Compared to hard disk drives, SSDs have a smaller physical size, are lighter, and use less energy. Their lack of mechanically moving parts, which makes them more shock resistant, is suitable for mobile applications, a growing market. Another advantage is that it is easy to integrate with other LSIs and enables the addition of more functions. Multiple 3D NAND dies are contained in the current flash memory package. Wire bonding is used to connect each memory dies and stacking by shifting them so that wires do not interfere. The bonding pads and wires increase parasitic capacitance and inductance, affecting and degrading read and write speeds. Each die requires I/O circuits to drive signals to the bonding pad, which results in a die size increase. TSV technology enables one to stack dies vertically without shifting them. It does not need a bonding pad that can eliminate the I/O circuit for each die, as shown in Figure 6.2. Smaller die sizes reduce parasitic capacitance and inductance, improving speed and power consumption [149]. TSV can stack different dies, and it can integrate the memory controller and other system LSIs in the future, in addition to the 3D NAND.

Figure 6.3 shows one example of a future storage system. Current high-end SSDs consist of multiple 3D NAND packages, a memory controller, and DRAMs. These dies can be integrated into one package as a one package SSD. Furthermore, those one package SSD will be directly attached to accelerators or even combined with the accelerator, DRAM, SLC-NAND, or other storage-

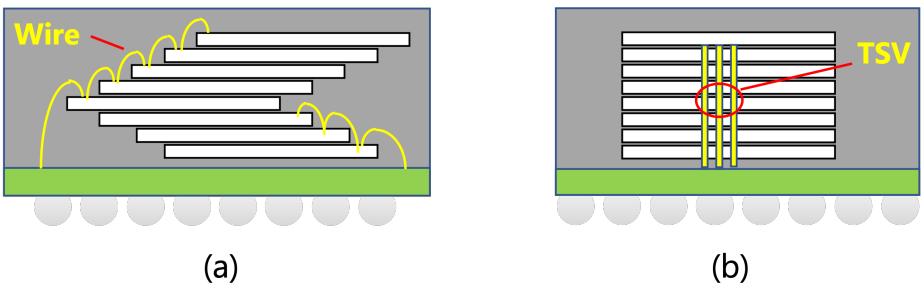


Figure 6.2: Schematic images of (a) conventional wire bonding and (b) TSV. TSV enables smaller package size.

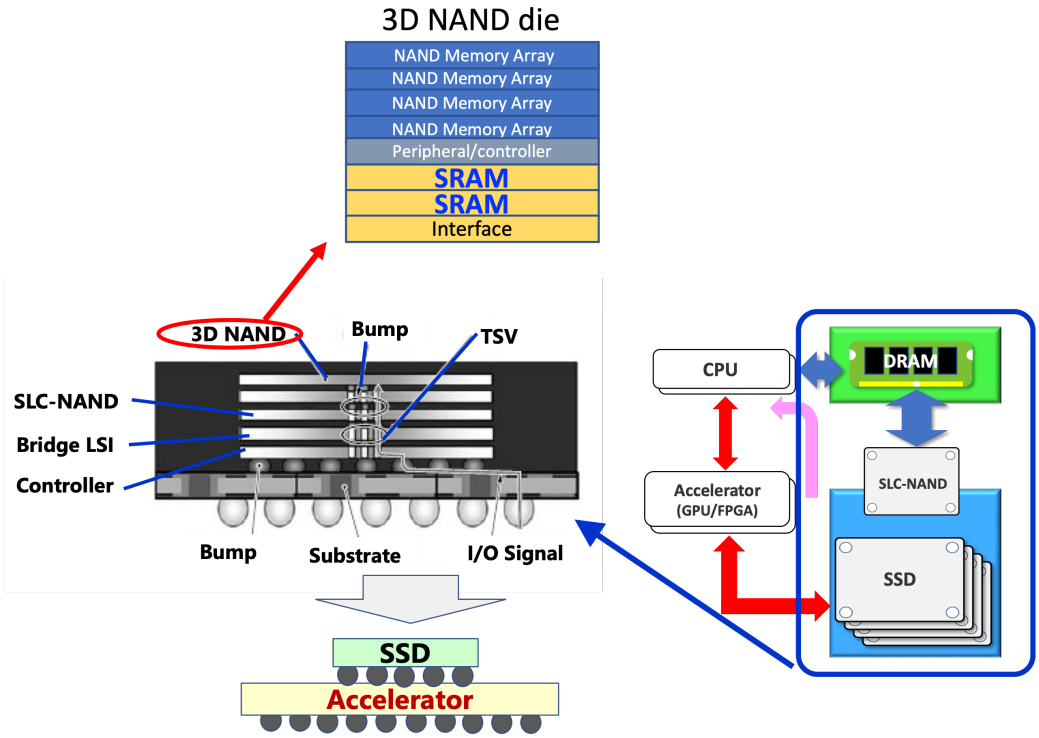


Figure 6.3: Concept of future one package storage by heterogeneous integration with high-capacity SRAM embedded 3D NAND.

class memory and realizes a tiny mobile edge server. There are many issues for realization, such as cost and thermal management. However, it is a viable future.

In order to realize such one package SSD, the impact on mechanical stress should also be considered. There are many studies on die-package interaction, such as shown in Figure 6.4 [150] as an example. Wafer warpage shown in Figure 6.5 is another issue for assembly process [151]. Not only the package or wafer but also the device itself has mechanical stress issues, as shown in Figure 6.6 [152].

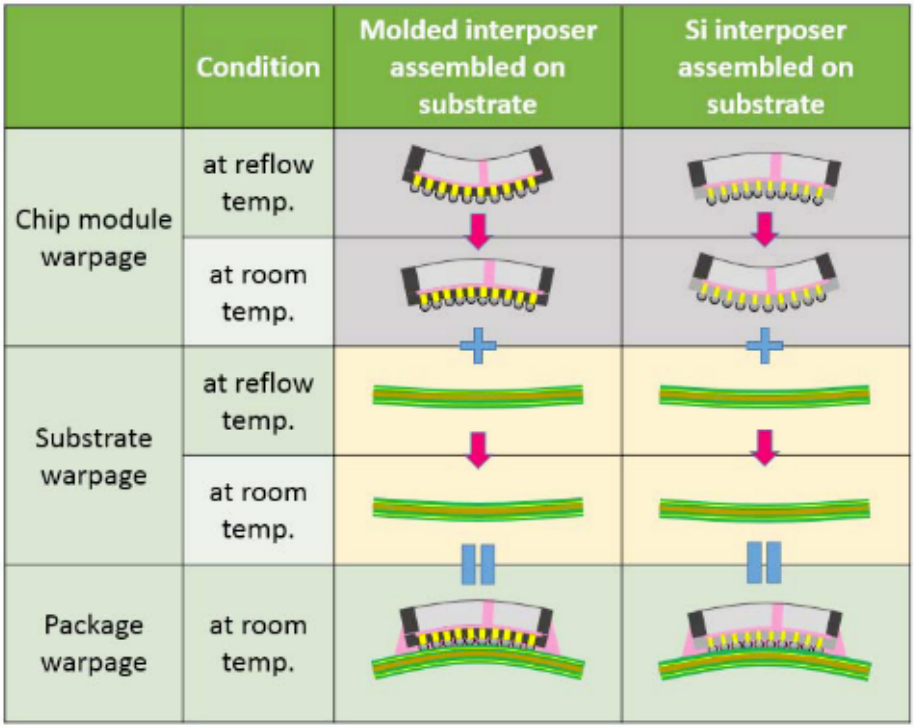


Figure 6.4: Example of chip module/substrate/package warpage from reflow to room temperature. [150].© 2019 IEEE

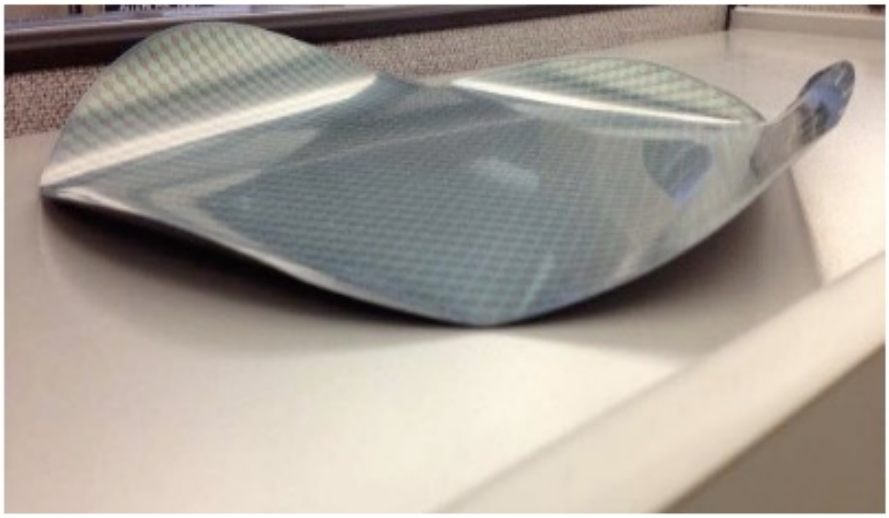


Figure 6.5: Example of wafer warpage after backside grinding [151].© 2013 IEEE

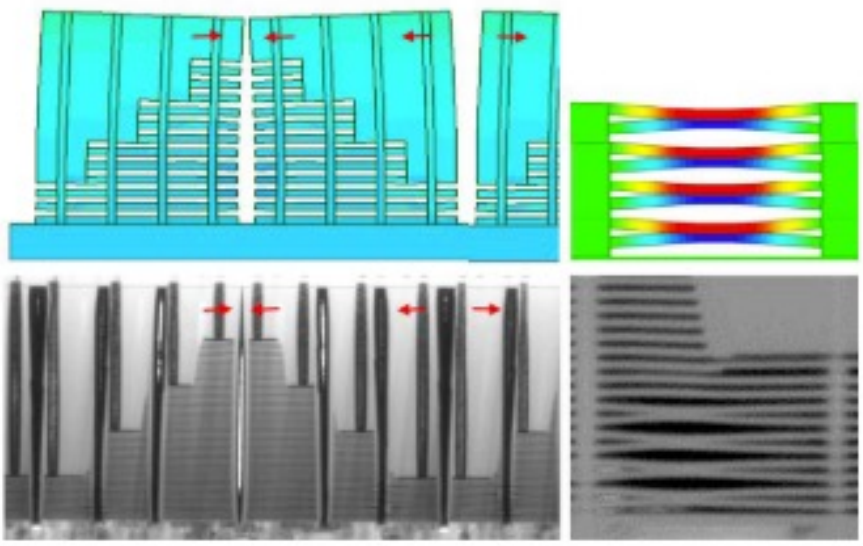


Figure 6.6: Process and device structure oriented deformation within device [152].© 2016 IEEE

However, there are few studies on the impact of die-to-die bonding stress on a memory device such as SRAM. Figure 6.7 (a) shows the schematic image used in the preliminary simulation study of warped die bonding. The upper die A has warpage and bonded on the flat die B. In this simulation, both dies are assumed as Si chiplets and have the same size of X = 10 mm and Y = 6 mm with the thickness T = 20 μm. The amount of the warpage, d, in die A is defined by the difference between the bottom (center of the die in this case) and top of the die (edge of the die in this case) as shown in Figure 6.7 (b). Figure 6.8 (a) shows the contour plot of simulated shear stress in die A bonded on die B in the case of warpage of die A is 0.1 mm. It shows a large amount of shear stress at die edge regions and decreases toward the die center. This is because die A has a large amount of warpage at the periphery and tries to recover the initial (warped) condition after bonding. Shear stress, along with the X direction at Y = 3 mm (shown as A-A' in Figure 6.8 (a)) and Y direction at X = 5 mm (shown as B-B' in Figure 6.8 (b)), are plotted in Figure 6.8 (b) for both warp amounts d = 0.1 and 0.5 mm, respectively. The stress criteria of 170 MPa proposed in Chapter 2 are indicated as a red line in the figure. In the case of initial warp amount

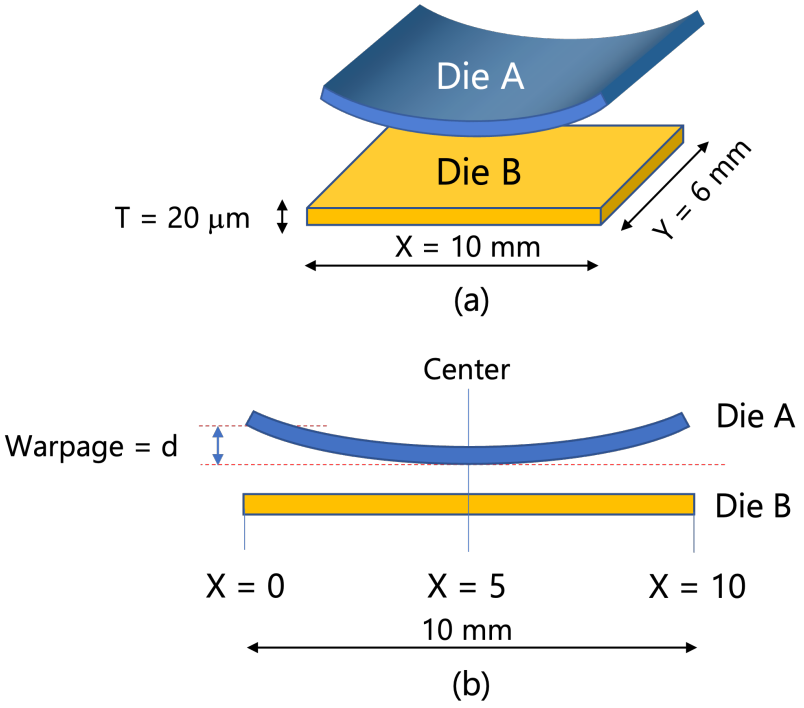


Figure 6.7: Schematic images of simulated structure. (a) Bird's eye view of both dies A and B with dimension. (b) Cross-sectional image of the die shape used in the simulation. The warp amount d is defined as shown in the figure. It is assumed the die A has a symmetrical warp structure in the X direction.

d = 0.1 mm, the amount of shear stress for both the X and Y direction does not show much difference (less than 0.1 mm). The stress value decreases monotonically as it moves away from the die edge and is below the stress criteria of 170 MPa at the region more than 0.5 mm away from the edge of the die. On the other hand, in the warpage amount d = 0.5 mm case, the stress value of the die edge region is increased and shows more than 100 MPa difference between X direction (along A-A') and Y direction (along B-B'). This is because the die warps only in the X direction. Similarly, in the d = 0.1 mm case, the stress decreases from the die edge toward the center and falls below 170 MPa in the region more than 1.2 mm away from the die edge for X direction while Y direction is around 0.5 mm that is similar to the warp amount d = 0.1 mm case. The study on the mechanical stress

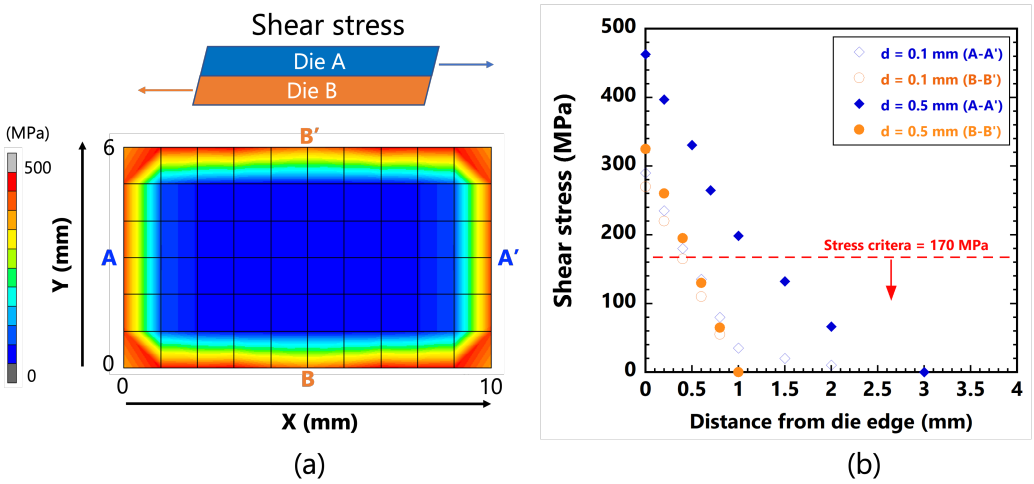


Figure 6.8: Simulation results of shear stress in die A beside die B. (a) Contour plot of shear stress in the case of die A warp amount d is 0.1 mm. (b) Shear stress in die A from the die edge for both X direction (along A-A') and Y direction (along B-B'). The die warpage amount d = 0.1 and 0.5 mm are plotted. Stress criteria, 170 MPa, is shown as a red dotted line.

impact on the device characteristics by nanoindentation technique also shows a similar tendency [83] with a stress decay length of about several hundred microns length. In reality, the situation is more complicated because of the internal residual stress by multilayer interconnect and deformation by the gate replacement process in 3D NAND, as shown in Figure 6.6. Those are the root cause of mechanical stress, and the balance of these stresses should be considered to embed high-capacity SRAM with high-capacity 3D NAND.

There are many studies on TSV, which is used as one of the dies connecting processes, regarding the impact on device characteristics [154, 155]. The mechanical stress near the TSV area affects MOSFET performance, and a keep-out-zone (KOZ) is proposed where devices should not be placed near the TSVs to prevent the degradation of device characteristics and the reliability [157]. However, this KOZ is a micrometer range (< 10 μm). On the contrary, the warpage impact propagates several hundred-micron ranges in this case, which is one or two orders of magnitude larger than KOZ. Therefore, when stacking dies using hybrid bonding, the manufacturing process should be constructed to consider the effects of this

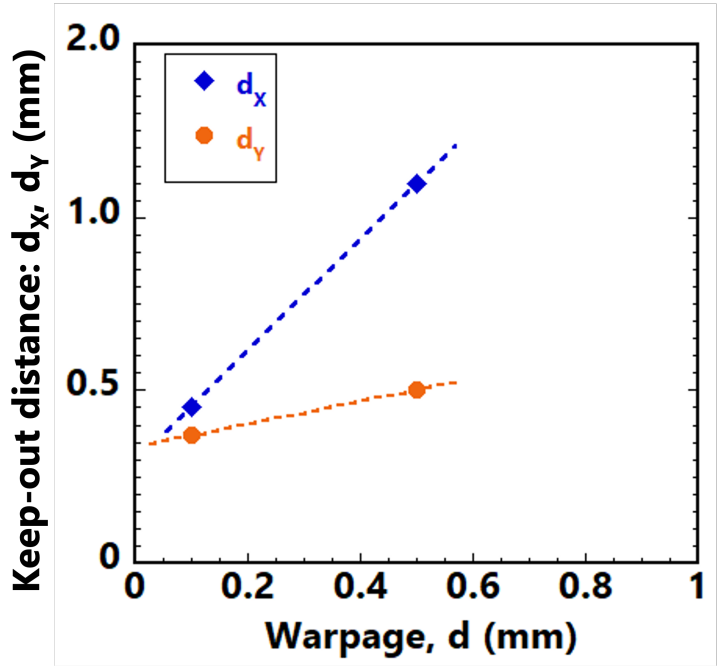


Figure 6.9: Keep-out distance, d_X and d_Y obtained from the simulation results shown in Figure 6.8. d_X shows strong dependence on the amount of warpage.

mechanical stress, and elements that are sensitive to mechanical stress should be arranged so that they are not affected by it. Figure 6.10 shows one example of a stress-sensitive device, such as SRAM, placement area by considering stacked die induced stress as shown in Figure 6.8. If the top die shows a concave shape, as die A, the SRAM placement area in die B should avoid the edge region where the stress value is maximum. From the simulation results, as shown in Figure 6.8, the keep-out-area (KOA) is defined by the allowable maximum stress value. A KOA is defined by d_X and d_Y, respectively, as shown in Figure 6.10. In the case of die warpage d = 0.5 mm, d_X = 1.2 mm and d_Y = 0.5 mm are keep-out distance. Figure 6.9 shows the warpage amount dependence of d_X and d_Y, obtained from simulation results shown in Figure 6.8. Since the die warp in the X direction, d_Y shows small dependence on warpage amount. The amounts of die warpage and shapes are affected by the grinding condition during the wafer thinning process [153, 156]. Therefore, the placement area of the

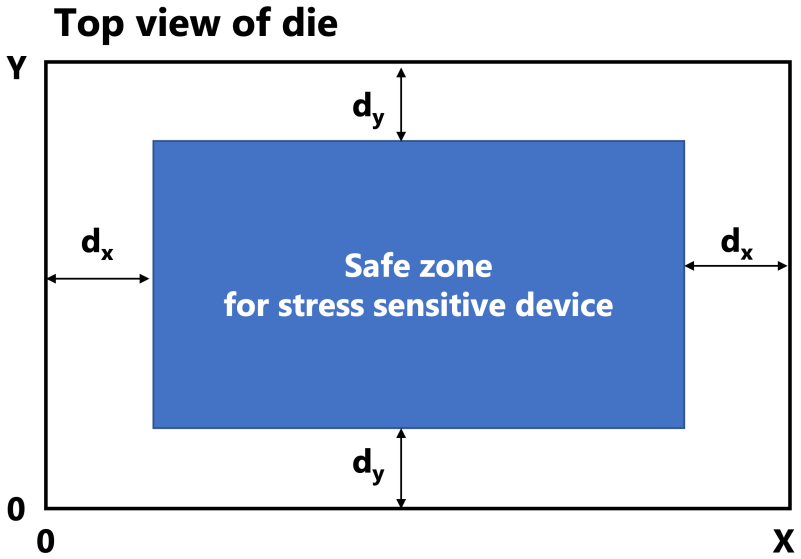


Figure 6.10: Example of SRAM placement area, indicate as safe zone, by considering mechanical stress caused by stacked die with warpage. The keep out distances, d_X and d_Y are also indicated in the figure.

mechanical stress-sensitive device, such as SRAM, needs to be modified by considering warpage conditions. To enhance the placement flexibility, the warpage amount should be controlled based on this guideline. In order to make high-reliable high-capacity embedded SRAM, this effect should be taken into account for 3D stacking cases. Generally, the die warps in a concave shape. If the die warps a convex shape, the characteristic becomes opposite (maximum stress at the die center) and significantly impacts SRAM placement. Therefore, control of the warpage direction is crucial and needs to study in advance when the manufacturing process and structure are modified. This topic is beyond the scope of this thesis. However, it is an extremely important topic, and I will continue to study it as a next step.

6.3 Summary of this chapter

The remaining issues for embedding high-reliable high-capacity SRAM into 3D NAND that enable one package server for future CiM were discussed. To embed higher-capacity SRAM with 3D NAND for CiM application requires beyond 32nm technology, and the gate-last HK/MG technology needs to be adopted. Therefore, the bonding process will need to be introduced to bond the 3D NAND and CMOS peripheral circuit. There are two possible bonding methods: wafer-to-wafer bonding and die-to-die bonding. The die-to-die bonding process is preferable from the yield point of view, as it allows the selection and stacking of good dies. Moreover, the current packaging scheme that stacks up to sixteen 3D NAND dies for wire bonding, as shown in Figure 6.2 (a), can be utilized for die handling. However, attention must be paid to the effect of warpage caused by the thinning of the die on the device characteristics of the dies to be stacked, especially mechanical stress. In addition to the internal stress of the 3D NAND itself, the impact of the stress caused by the stacked dies on yield and reliability must be taken into consideration. Since die warpage is strongly affected by the device configuration and fabrication process, it is necessary to determine the effect of warpage by stress simulation in advance and to present SRAM placement prohibited areas, KOA, like KOZ in TSV. These studies have not yet been widely conducted. Although they are not in the scope of this paper, they are essential and should be studied in the future.

Chapter 7

Conclusions

7.1 Thesis summary

Advanced process technologies for high-capacity and high-reliable SRAM are discussed. The object is to embedding high-capacity SRAM with 3DNAND because it can contribute to building inexpensive and energy-efficient computing systems. This thesis aims to identify the technologies required to embed high-capacity SRAM in 3D NAND, identify the technologies that need to be changed or newly prepared in the existing technologies, and provide a roadmap for realization.

The Chapter 1 discussed the importance of embedding high-capacity SRAM in 3D NAND to realize energy-efficient computing systems. The manufacturing processes of embedded SRAM, stand-alone SRAM, and 3D NAND were compared. Processes that require special consideration for embedding high-capacity SRAM in the FEOL, MOL, and BEOL modules were extracted.

In Chapter 2, the optimization method of SRAM cell layout was proposed. Two different cell layouts were evaluated, and the thin type cell that showed better scalability was selected. Fabricated SRAM showed crystal defects with leakage currents that led to bit fails. To determine the root cause, stress simulations were performed. The simulation results suggested that mechanical stresses from the STI filling materials and the gate electrode caused this failure. The overlap amounts between the active area and the

gate electrode were optimized with the sidewall structure of MOSFETs. Fabricated SRAM with optimized cell layout showed good functionality, and cell design guideline was transferred to the production group and validated by the products down to 45 nm node.

The Chapter 3 discussed the reliability of MOSFETs used in SRAM cells. It was found that MOSFETs with the smallest dimensions used in SRAM cells deteriorate faster than the MOSFETs with larger channel widths used in peripheral circuits. Both measured and simulated results suggested that mechanical stress may assist this accelerated degradation in small dimension MOSFETs. The importance of smallest size MOSFETs usage for reliability measurement was recommended to realize high-reliable high-capacity SRAM.

In Chapter 4, process technologies need to pay special attention in both FEOL and MOL were discussed. To embed a high-capacity SRAM into 3D NAND, gate insulator thickness scaling was necessary from both cell size scaling and gate leakage current reduction points of view. To introduce high-κ gate dielectrics, the gate-first process was the only option because of the total thermal budget, and HfSiON combined with polymetal gate electrode was recommended. The local interconnect reduced the number of interconnect layers in the SRAM cell. Design guideline of STI height with preferable gate sidewall structure was provided to realize high-reliable high-capacity SRAM that could be embedded in 3D NAND.

Chapter 5 discussed the BEOL technology. The redundancy fuse process for SRAM was evaluated. To apply the design architecture and infrastructure established for high-capacity SRAM products, laser blow of copper interconnect was investigated. The titanium barrier metal, whose melting temperature was lower than the tantalum, commonly used as a barrier metal for copper interconnect, was adopted. Although the titanium barrier metal showed better electromigration characteristics than the tantalum, the leakage currents caused by copper corrosion after the laser blow process were observed. Therefore, an alternative fuse process, such as anti-fuse, was proposed to realize high-reliable SRAM. The low-κ materials were also evaluated to accommodate future speed requirements. Even if the dielectric constants were the same, the resistance to etching varies depending on the constituent materials, indicating that material selection is essential.

In Chapter 6, the need to consider the 3D integration process for future one package server realization was discussed. The effects of die warpage for the die stacking process were simulated. When a warped die was bonded to a flat die, significant stress was applied to the surface of the flat die. The peripherals of the die showed large shear stress, and those regions varied depending on the amount of warpage. Based on the preliminary study, the requirement of a keep-out area for the stress-sensitive device, such as SRAM, was proposed.

Above mentioned SRAM process technologies with design guidelines enable embedding high-reliable high-capacity SRAM into 3D NAND with 90 nm to at least 45 nm node CMOS technologies.

7.2 Future directions

One of the issues, as mentioned in Chapter 1, is the server's power consumption. These days, liquid cooling of the entire server system is considered [158]. If technology can make it happen, it is possible to put a small edge server in a liquid nitrogen ambient. Liquid cooling has already been introduced to reduce the cooling energy, as shown in Figure 7.1.

The server blades were immersed in a large bathtub filled with cold-boiling liquid, such as Fluorinert ™, whose boiling temperature is less than 50°C. The vaporized liquid was condensed and returned to the bathtub by condenser on the top of the bathtub.

Further cooling, such as cryogenic temperature, is on the horizon. Figure 7.2 and 7.3 show preliminary results of the 77 K operation of 3D NAND. It shows tight V_{th} distribution and robust cycling characteristics compare to those at room-temperature operation. This enables a 6 bit per cell (Hexa Level Cell, HLC) operation. These are just preliminary results. However, cryogenic sever is a considerable future system. Cryogenic temperature is also an advantage for SRAM because of V_{min} improvement. This cryogenic server will help to improve the performance and reduce the physical size of the future quantum computing system, which uses very low temperature in the mK range. It can also expand the application to other

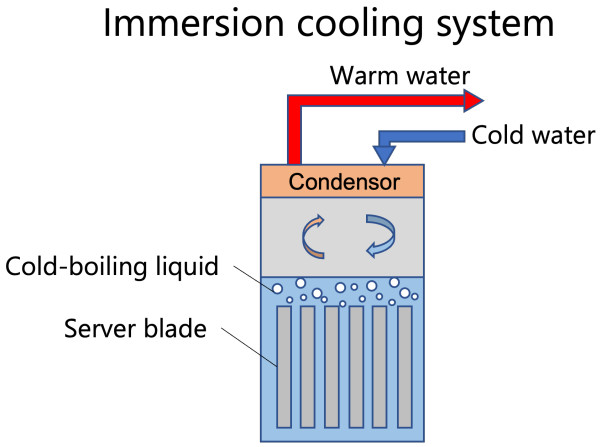


Figure 7.1: Schematic image of liquid cooling system that is commercially available for server system [3, 4].

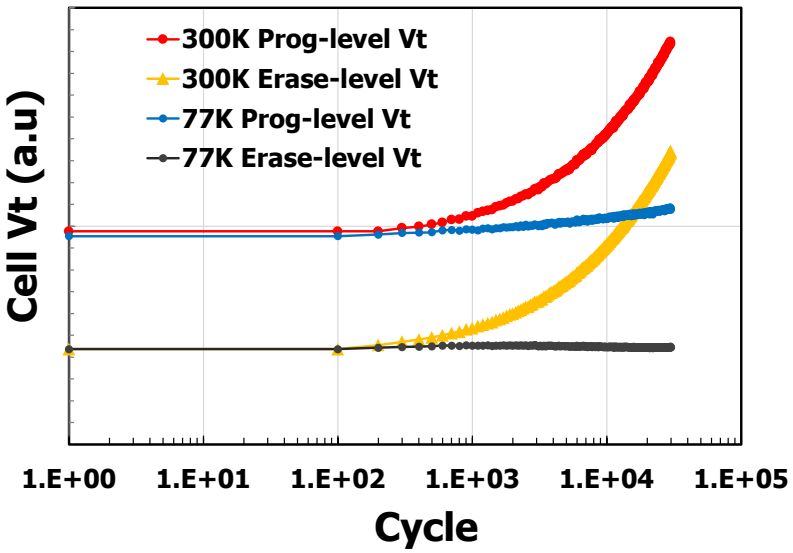


Figure 7.2: 77 K operation of 3D NAND shows better cycling characteristics compare to 300 K operation [161].© 2021 IEEE

areas such as aerospace, and people can build a datacenter on the Moon or Mars.

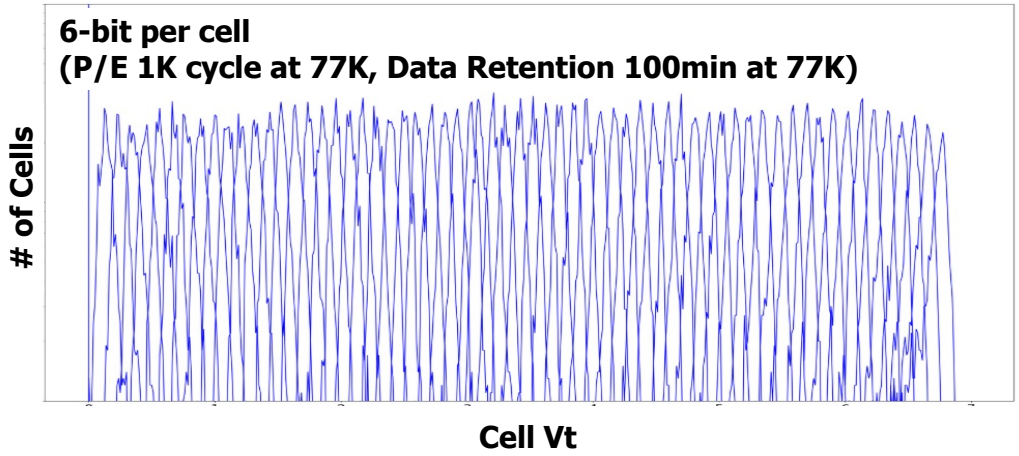


Figure 7.3: 3D NAND shows HLC capability under 77 K operation [161].© 2021 IEEE

Lastly, the future manufacturing process is mentioned. As mentioned earlier, to store the explosive growth of data, the storage capacity must be

increased to match the data growth. Over the last 20 years, NAND flash memory density per unit area (Gb/mm^2) has grown at a rate of about 40% per year, resulting in increased storage capacity. It will continue at least coming decade and will be more than 1 Tb/mm^2 in 2030. The transition from 2D to 3D structure enabled this continuous growth of NAND flash memory capacity, as shown in Figure 7.4.

The number of memory layers in 3D flash memory has increased from 96 layers to 170 layers over the past three years [34]. If the number of layers continues to grow at this rate, it will reach around 1000 layers by 2030. There is no manufacturing process for 3D NAND with more than one thousand WL layers. However, it is not impossible to achieve this through continuous manufacturing process innovation. This low-cost manufacturing process for a vertical stack of many transistors can be used not only for 3D NAND but also for future logic processes. As can be seen in Figure 7.5, the CMOS transistor structure has shifted from planar to Fin-type transistors, and a shift to gate-all-around (GAA), nanosheet, and nanoribbon structures [163, 164] is being considered in the future. These transistors are also considered for stacking channel regions in the height direction, making the 3D NAND

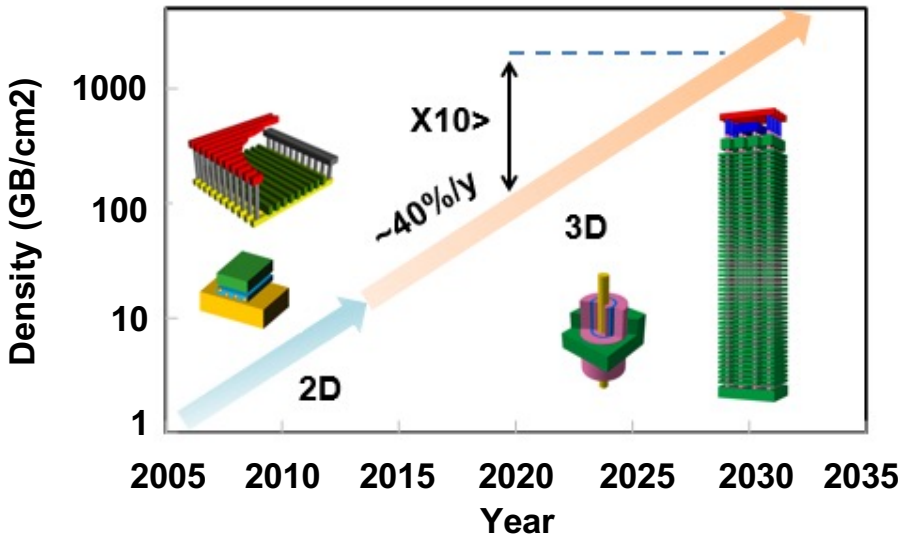


Figure 7.4: NAND flash memory density increases roughly 40%/year by shifting 2D to 3D structure.

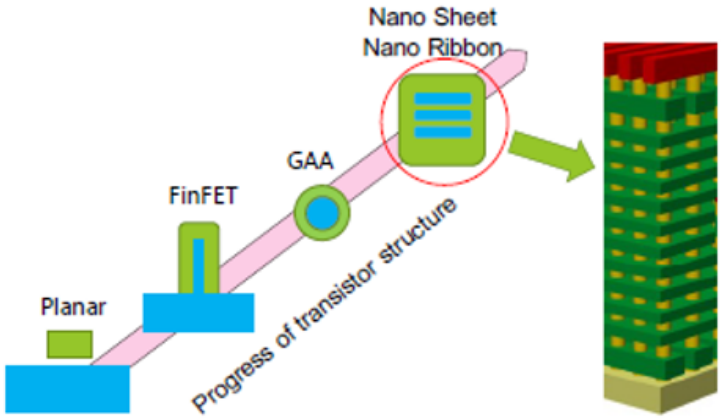


Figure 7.5: Progress of transistor structure. The best practice to reduce manufacturing costs of 3D NAND can be applied for future transistor manufacturing process.

formation method applicable. The channel regions of nanosheet devices are formed by epitaxial growth, which is expensive. The process of converting amorphous-Si or poly-Si channels to crystalline silicon by metal-assisted solid-phase crystallization (MILC) process is considered to apply to the 3D NAND manufacturing process [159]. This process can also apply to this nanosheet transistor formation. This process reduces manufacturing costs significantly because the MILC process costs much lower than the epitaxial growth process. The same MILC process can be applied to the transistor channels of both nanosheet transistors and 3D NAND. Figure 7.6 shows the images of the future 3D NAND embedding CMOS peripheral circuits and SRAM composed of nanosheet transistors. The nanosheet transistors are rotated 90 degrees and make vertical channel nanosheet transistors. Fine pitches of nanosheet channels can be manufactured using a double or quadruple patterning process used for 2D NAND manufacturing. After the channel crystallization by the MILC process, HK/MG process is applied to both nanosheet transistors and 3D NAND cells. The HK/MG process is first applied to the 3D NAND region using the current replacement gate process. Then HK for both n-type nanosheet and p-type nanosheet is deposited,

followed by meta gate deposition. These nanosheet transistor structures can be stacked vertically by utilizing a multi-tier process [167] used in current 3D NAND manufacturing. There are many challenges to be solved. However, it is possible enabling nanosheet-based high-reliable high-capacity SRAM to merge with 3D NAND, and it will realize a sustainable computing systems.

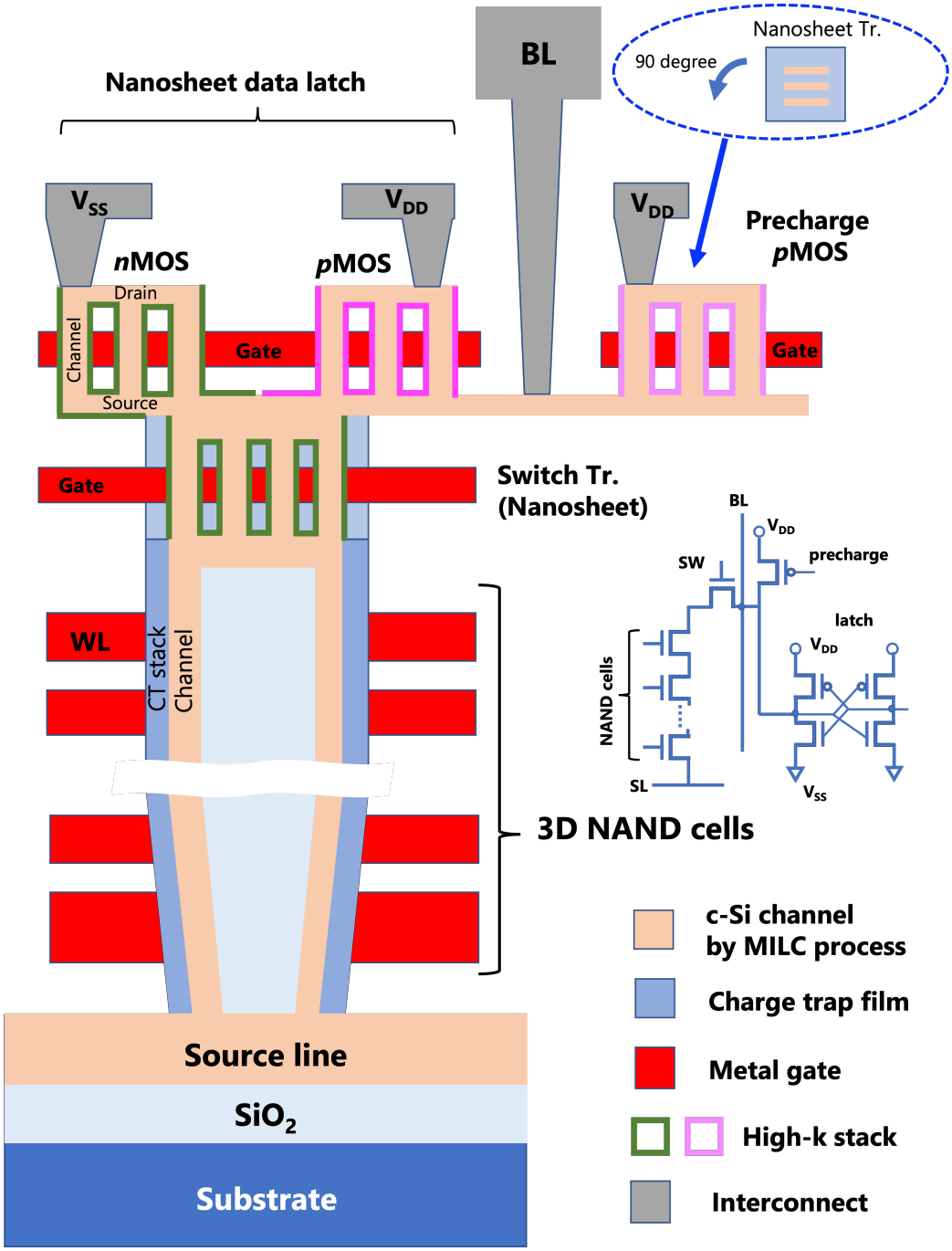


Figure 7.6: Future structure of the 3D NAND embedded with high-capacity SRAM by nanosheet transistors.

Bibliography

[1] Statista, "Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025," 2021. https://www.statista.com/statistics/871513/worldwide-data-created/

[2] D. Reinsel, J. Gantz and J. Rydning, "The Digitization of the World - From Edge to Core," *An IDC White Paper - #US44413318*, November 2018.

[3] L. A. Barroso, U. Hölzle and P. Ranganathan, "The Datacenter as a Computer: Designing Warehouse-Scale Machines, Third Edition," *SYNTHESIS LECTURES ON COMPUTER ARCHITECTURE Lecture #46*, Morgan & Claypool Publishers, October 2018.

[4] A. Shehabi, S. J. Smith, E. Masanet and J. Koomey, "Data center growth in the United States: Decoupling the demand for services from electricity use," *Environ. Res. Lett.*, vol. 13, p. 124030, 2018. doi:10.1088/1748-9326/aaec9c

[5] E. Masanet, A. Shehabi, N. Lei, S. Smith and J. Koomey, "Recalibrating global data center energy-use estimates," *Science*, vol. 367, no. 6481, pp. 984–986, 2020.

[6] M. Koot and F. Wijnhoven, "Usage impact on data center electricity needs: A system dynamic forecasting model," *Applied Energy*, vol. 291, p. 116798, 2021.

[7] A. Andrae and T. Edler, "On Global Electricity Usage of Communication Technology: Trends to 2030," *Challenges*, vol. 6, no. 1, pp. 117–157, 2015.

[8] K. Saegusa, "Impact of Progress of Information Society on Energy Consumption (Vol. 4): Feasibility Study of Technologies for Decreasing Energy Consumption of Data Centers," vol. 4, February, 2022.

[9] J. Backus, "Can Programming Be Liberated from the von Neumann Style? A Functional Style and Its Algebra of Programs," *Communicaions ACM*, vol. 21, no. 8, pp. 613–641, 1978.

[10] W. A. Wulf and S. A. McKee, "Hitting the memory wall," *ACM SIGARCH Computer Architecture News*, vol. 23, no. 1, pp. 20–24, 1995.

[11] M. Horowitz, "Computing's Energy Problem (and what we can do about it)," *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2014, pp. 10–14, doi: 10.1109/ISSCC.2014.6757323.

[12] H. Jun *et al.*,"HBM (High Bandwidth Memory) DRAM Technology and Architecture," *2017 IEEE International Memory Workshop (IMW)*, 2017, pp. 1–4, doi: 10.1109/IMW.2017.7939084.

[13] K. K. W. Chang *et al.*, "Improving DRAM performance by parallelizing refreshes with accesses," *2014 IEEE 20th International Symposium on High Performance Computer Architecture (HPCA)*, 2014, pp. 356–367, doi: 10.1109/HPCA.2014.6835946.

[14] L. T. Su, S. Naffziger and M. Papermaster, "Multi-chip technologies to unleash computing performance gains over the next decade," *2017 IEEE International Electron Devices Meeting (IEDM)*, 2017, pp. 1.1.1–1.1.8, doi: 10.1109/IEDM.2017.8268306.

[15] C. C. Hu, M. F. Chen, W. C. Chiou and D. C. Yu, "3D Multi-chip Integration with System on Integrated Chips (SoIC™)," *2019 Symposium on VLSI Technology*, 2019, pp. T20–T21, doi: 10.23919/VLSIT.2019.8776486.

[16] T. Burd *et al.*, "Zen3: The AMD 2^{nd} -Generation 7nm x86-64 Microprocessor Core," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, 2022, pp. 1–3, doi: 10.1109/ISSCC42614.2022.9731678.

[17] C. Y. Lo, P. T. Huang and W. Hwang, "Energy-Efficient Accelerator Design with 3D-SRAM and Hierarchical Interconnection Architecture for Compact Sparse CNNs," *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 2020, pp. 320–323, doi: 10.1109/AICAS48895.2020.9073944.

[18] H. C. Chen, J. F. Li, C. L. Hsu and C. T. Sun, "Configurable 8T SRAM for Enbling in-Memory Computing," *2019 2nd International Conference on Communication Engineering and Technology (ICCET)*, 2019, pp. 139–142, doi: 10.1109/ICCET.2019.8726871.

[19] A. K. Rajput and M. Pattanaik, "Energy Efficient 9T SRAM With R/W Margin Enhanced for beyond Von-Neumann Computation," *2020 24th International Symposium on VLSI Design and Test (VDAT)*, 2020, pp. 1–4, doi: 10.1109/VDAT50263.2020.9190473.

[20] C.-J. Jhang, C.-X. Xue, J.-M. Hung, F.-C. Chang and M.-F. Chang, "Challenges and Trends of SRAM-Based Computing-In-Memory for AI Edge Devices," *2021 IEEE 14th International Conference on ASIC (ASICON)*, 2021, pp. 1–4, doi: 10.1109/ASICON52560.2021.9620429.

[21] G. W. Burr *et al.*, "Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses), using phase-change memory as the synaptic weight element," *2014 IEEE International Electron Devices Meeting (IEDM)*, 2014, pp. 29.5.1–29.5.4, doi: 10.1109/IEDM.2014.7047135.

[22] H. Li *et al.*, "Hyperdimensional computing with 3D VRRAM in-memory kernels: Device-architecture co-design for energy-efficient, error-resilient language recognition," *2016 IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 16.1.1–16.1.4, doi: 10.1109/IEDM.2016.7838428.

[23] M. Abe, C. Matsui, K. Mizushina, S. Suzuki and K. Takeuchi, "Computational Approximate Storage with Neural Network-based Error Patrol of 3D-TLC NAND Flash Memory for Machine Learning

Applications," *2020 IEEE International Memory Workshop (IMW)*, 2020, pp. 1–4, doi: 10.1109/IMW48823.2020.9108136.

[24] S. K. Gonugondla, M. Kang, Y. Kim, M. Helm, S. Eilert and N. Shanbhag, "Energy-Efficient Deep In-memory Architecture for NAND Flash Memories," *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018, pp. 1–5, doi: 10.1109/ISCAS.2018.8351458.

[25] H. T. Lue *et al.*, "Optimal Design Methods to Transform 3D NAND Flash into a High-Density, High-Bandwidth and Low-Power Nonvolatile Computing in Memory (nvCIM) Accelerator for Deep-Learning Neural Networks (DNN)," *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019, pp. 38.1.1–38.1.4, doi: 10.1109/IEDM19573.2019.8993652.

[26] H. T. Lue, P. K. Hsu, K. C. Wang and C. Y. Lu, "Introduction of Non-Volatile Computing in Memory (nvCIM) by 3D NAND Flash for Inference Accelerator of Deep Neural Network (DNN) and the Read Disturb Reliability Evaluation: (Invited Paper)," *2020 IEEE International Reliability Physics Symposium (IRPS)*, 2020, pp. 1–6, doi: 10.1109/IRPS45951.2020.9128340.

[27] M. M. Hasan, M. Raquibuzzaman, I. Chatterjee and B. Ray, "Radiation Tolerance of 3-D NAND Flash Based Neuromorphic Computing System," *2020 IEEE International Reliability Physics Symposium (IRPS)*, 2020, pp. 1–4, doi: 10.1109/IRPS45951.2020.9128219.

[28] W. Shim and S. Yu, "Technological Design of 3D NAND-Based Compute-in-Memory Architecture for GB-Scale Deep Neural Network," in *IEEE Electron Device Letters*, vol. 42, no. 2, pp. 160–163, Feb. 2021, doi: 10.1109/LED.2020.3048101.

[29] L. Zhao *et al.*, "A Compute-in-Memory Architecture Compatible with 3D NAND Flash that Parallelly Activates Multi-Layers," *2021 58th ACM/IEEE Design Automation Conference (DAC)*, 2021, pp. 193–198, doi: 10.1109/DAC18074.2021.9586271.

[30] M. Kim, M. Liu, L. R. Everson and C. H. Kim, "An Embedded nand Flash-Based Compute-In-Memory Array Demonstrated in a Standard Logic Process," in *IEEE Journal of Solid-State Circuits*, vol. 57, no. 2, pp. 625–638, Feb. 2022, doi: 10.1109/JSSC.2021.3098671.

[31] H. Tanaka *et al.*, "Bit Cost Scalable technology with and plug process for ultra high density flash memory," *2007 IEEE Symposium on VLSI Technology*, 2007, pp. 14–15, doi: 10.1109/VLSIT.2007.4339708.

[32] 5Y. Fukuzumi *et al.*, "Optimal Integration and Characteristics of Vertical Array Devices for Ultra-High Density, Bit-Cost Scalable Flash Memory," *2007 IEEE International Electron Devices Meeting (IEDM)*, 2007, pp. 449–452, doi: 10.1109/IEDM.2007.4418970.

[33] K. Parat and A. Goda, "Scaling Trends in NAND Flash," *2018 IEEE International Electron Devices Meeting (IEDM)*, 2018, pp. 2.1.1–2.1.4, doi: 10.1109/IEDM.2018.8614694.

[34] T. Higuchi *et al.*, "A 1Tb 3b/Cell 3D-Flash Memory in a 170+ Word-Line-Layer Technology," *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, 2021, pp. 428–430, doi: 10.1109/ISSCC42613.2021.9366003.

[35] YMTC 3D NAND,
https://www.ymtc.com/en/technicalintroduction.html

[36] A. Goda, "Recent progress on 3D nand flash technologies," *Electronics*, vol. 10, no. 3516, pp. 1–16, 2021, doi:10.3390/electronics10243156.

[37] D. Takashima, M. Noguchi, N. Shibata, K. Kanda, H. Sukegawa and S. Fujii, "An embedded DRAM technology for high-performance NAND flash memories," in *IEEE Journal of Solid-State Circuits*, vol. 47, no. 2, pp. 536–546, Feb. 2012, doi: 10.1109/JSSC.2011.2170779.

[38] KIOXIA Website
https://business.kioxia.com/ja-jp/ssd.html

[39] K. Ishimaru, "Non-Volatile Memory Technology for Data Age," *2018 14th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, 2018, pp. 1–4, doi: 10.1109/ICSICT.2018.8564815.

[40] K. Ishimaru, "Future of Non-Volatile Memory -From Storage to Computing-," *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019, pp. 1.3.1–1.3.6, doi: 10.1109/IEDM19573.2019.8993609.

[41] G. Yeap *et al.*, "5nm CMOS Production Technology Platform featuring full-fledged EUV, and High Mobility Channel FinFETs with densest 0.021μm^2 SRAM cells for Mobile SoC and High Performance Computing Applications," *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019, pp. 36.7.1–36.7.4, doi: 10.1109/IEDM19573.2019.8993577.

[42] DRAMeXchange, https://www.dramexchange.com/

[43] K. Miyashita *et al.*, "A high performance 100 nm generation SOC technology (CMOS IV) for high density embedded memory and mixed signal LSIs," *2001 Symposium on VLSI Technology. Digest of Technical Papers*, 2001, pp. 11–12, doi: 10.1109/VLSIT.2001.934922.

[44] C. C. Wu *et al.*, "A 90-nm CMOS device technology with high-speed, general-purpose, and low-leakage transistors for system on chip applications," *2002 IEEE International Electron Devices Meeting (IEDM)*, 2002, pp. 65–68, doi: 10.1109/IEDM.2002.1175780.

[45] M. C. Poon, C. H. Ho, F. Deng, S. S. Lau and H. Wong, "Thermal stability of cobalt and nickel silicides," *Microelectronics Reliability*, vol. 38, no. 9, pp. 1495–1498, 1998, doi:10.1016/S0026-2714(98)00045-6.

[46] A. Alberti, F. La Via, M. G. Grimaldi and S. Ravesi, "Cobalt silicide thermal stability: From blanket thin film to submicrometer lines," *Solid-State Electronics*, vol. 43, no. 6, pp. 1039–1044, 1999, doi:10.1016/S0038-1101(99)00021-0.

[47] O. A. Kirillov, "Development of $CoSi_2$ Salicide Process," *19th Annual Microelectronics Engineering Conference*, pp. 45–50, May 2001.

[48] A. Lauwers *et al.*, "Silicides for the 100-nm node and beyond: Co-silicide, Co(Ni)-silicide and Ni-silicide," *Microelectronic Engineering*, vol. 64, no. 1-4, pp. 131–142, 2002.

[49] J. A. Kittl *et al.*, "Ni- and Co-based silicides for advanced CMOS applications," *Microelectronic Engineering*, vol. 70, no. 2-4, pp. 158–165, 2003.

[50] J. Mendonca *et al.*, "Interconnect material and CMP process change effects on local interconnect planarity," *Proceedings of the IEEE 1998 International Interconnect Technology Conference (IITC)*, 1998, pp. 196–198, doi: 10.1109/IITC.1998.704790.

[51] J. Chapple-Sokol, R. Phelps, T. Krywanczyk, C. Sanetra and D. Sturtevant, "Damascene tungsten process for local interconnects with improved reliability performance," *2004 IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop (ASMC)*, 2004, pp. 472–476, doi: 10.1109/ASMC.2004.1309617.

[52] V. Kamineni *et al.*, "Tungsten and cobalt metallization: A material study for MOL local interconnects," *2016 IEEE International Interconnect Technology Conference / Advanced Metallization Conference (IITC/AMC)*, 2016, pp. 105–107, doi: 10.1109/IITC-AMC.2016.7507698.

[53] J. H. Park *et al.*, "Effect of $CoSi_2$ Formation Process on CMOS Transistor Electrical Properties for Sub-100-nm Memory Applications," *ECS Journal Solid State Science Technology*, vol. 5, no. 5, pp. 264–271, 2016, doi: 10.1149/2.0181605jss.

[54] M. Ishida, T. Kawakami, A. Tsuji, N. Kawamoto, M. Motoyoshi and N. Ouchi, "Novel 6T-SRAM cell technology designed with rectangular patterns scalable beyond 0.18 μm generation and desirable for ultra high speed operation," *1998 IEEE International Electron Devices Meeting (IEDM)*, 1998, pp. 201–204, doi: 10.1109/IEDM.1998.746322.

[55] K. Osada *et al.*, "Universal-Vdd 0.65-2.0V 32 kB cache using voltage-adapted timing-generation scheme and a lithographical-symmetric cell," *2001 IEEE Int. Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2001, pp. 168–169, doi: 10.1109/ISSCC.2001.912589.

[56] S. Parihar *et al.*, "A high density 0.10 μm CMOS technology using low k dielectric and copper interconnect," *2001 IEEE International Electron Devices Meeting (IEDM)*, 2001, pp. 249–252, doi: 10.1109/IEDM.2001.979477.

[57] S. M. Jung *et al.*, "A novel 0.79 μm^2 SRAM cell by KrF lithography and high performance 90 nm CMOS technology for ultra high speed SRAM," *2002 IEEE International Electron Devices Meeting (IEDM)*, 2002, pp. 419–422, doi: 10.1109/IEDM.2002.1175868.

[58] Y. W. Kim *et al.*, "Robust Process Integration of 0.78 μm^2 embedded SRAM with NiSi gate and low-K Cu interconnect for 90nm SoC applications," *2003 Symposium on VLSI Technology. Digest of Technical Papers*, 2003, pp. 69–70, doi: 10.1109/VLSIT.2003.1221090.

[59] S. Thompson *et al.*, "A 90 nm logic technology featuring 50 nm strained silicon channel transistors, 7 layers of Cu interconnects, low k ILD, and 1 μm^2 SRAM cell," *2002 IEEE International Electron Devices Meeting (IEDM)*, 2002, pp. 61–64, doi: 10.1109/IEDM.2002.1175779.

[60] Y. Fukaura *et al.*, "A highly manufacturable high density embedded SRAM technology for 90 nm CMOS," *2002 IEEE International Electron Devices Meeting (IEDM)*, 2002, pp. 415–418, doi: 10.1109/IEDM.2002.1175867.

[61] S. Tanaka, S. Inoue, T. Kotani, K. Izuha and I. Mori, "Impact of OPC aggressiveness on mask maufacturability," *Proceedings of SPIE, Photomask Next-Generation Lithography Mask Technology X*, vol. 5130, August 2003, pp. 23–32, 2003, doi:10.1117/12.504253.

[62] T. Kotani, S. Tanaka, S. Nojima, K. Hashimoto, S. Inoue and I. Mori, "Yield-enhanced layout generation by new design for manufacturability

(DfM) flow," *Proceedings of SPIE, Design and Process Integration for Microelectronic Manufacturing II*, vol. 5379, May 2004, pp. 128–138, 2004, doi: 10.1117/12.536254.

[63] S. Kobayashi, "Automated hot-spot fixing system applied to the metal layers of 65-nm logic devices," in *Journal of Micro/Nanolithography, MEMS, MOEMS*, vol. 6, no. 3, p. 031010, Jul.-Sep. 2007.

[64] K. Ishimaru *et al.*, "Mechanical stress induced MOSFET punch-through and process optimization for deep submicron TEOS-O_3 filled STI device," *1997 Symposium on VLSI Technology*, 1997, pp. 123–124, doi: 10.1109/VLSIT.1997.623729.

[65] M. L. Polignano, G. P. Carnevale, I. Mica and C. Pastore, "Mechanical Stress and Defect Formation in Device-Processing: Validity of the Numerical Models for Mechanical Stress Calculation," in *IEEE Transactions on Electron Devices*, vol. 54, no. 5, pp. 1108–1114, May 2007, doi: 10.1109/TED.2007.892948.

[66] S. Ikeda, H. Ohta, H. Miura and Y. Hagiwara, "Mechanical stress control in a VLSI-fabrication process: A method for obtaining the relation between stress levels and stress-induced failures," in *IEEE Transactions on Semiconductor Manufacturing*, vol. 16, no. 4, pp. 696–703, Nov. 2003, doi: 10.1109/TSM.2003.818979.

[67] P. Ferreira, R.-A. Bianchi, F. Guyader, R. Pantel and E. Granger, "Elimination of stress induced silicon defects in very high density SRAM structures," *31st European Solid-State Device Research Conference*, 2001, pp. 427–430, doi: 10.1109/ESSDERC.2001.195292.

[68] "MSC MARC" https://www.mscsoftware.com/product/marc

[69] K. Ishimaru, H. Gojohbori, H. Koike, Y. Unno, M. Sai, F. Matsuoka and M. Kakumu, "Trench isolation technology with 1 μm depth n- and p-wells for a full-CMOS SRAM cell with a 0.4 μm n^+/p^+ spacing," *Proceedings of 1994 VLSI Technology Symposium*,1994, pp. 97–98, doi: 10.1109/VLSIT.1994.324437.

[70] J. M. Gere and B. J. Goodno, *Mechanics of Materials*, 8th ed. Stamford, CT: Cengage Lerning, 2013.

[71] R. A. Bianchi, G. Bouche and O. Roux-dit Buisson, "Accurate modeling of trench isolation induced mechanical stress effects on MOSFET electrical performance," *2002 IEEE International Electron Devices Meeting (IEDM)*, 2002, pp. 117–120, doi: 10.1109/IEDM.2002.1175792.

[72] Ke-Wei Su *et al.*, "A scaleable model for STI mechanical stress effect on layout dependence of MOS electrical characteristics," *Proceedings of the IEEE 2003 Custom Integrated Circuits Conference (CICC)*, 2003, pp. 245–248, doi: 10.1109/CICC.2003.1249396.

[73] M. Miyamoto, H. Ohta, Y. Kumagai, Y. Sonobe, K. Ishibashi and Y. Tainaka, "Impact of Reducing STI-induced stress on layout dependence of MOSFET characteristics," in *IEEE Transactions on Electron Devices*, vol. 51, no. 3, pp. 440–443, 2004, doi: 10.1109/TED.2003.822877.

[74] J. Y. Xue, Y. D. S. Deng, Z. C. Ye, L. Yang and Z. P. Yu, "Modeling of layout-dependent STI stress in 65nm technology," *2009 IEEE 8th International Conference on ASIC*, 2009, pp. 670–673, doi: 10.1109/ASICON.2009.5351335.

[75] H. Miura, S. Ikeda and N. Suzuki, "Effect of mechanical stress on reliability of gate-oxide film in MOS transistors," *1996 IEEE International Electron Devices Meeting (IEDM)*,1996, pp. 743–746, doi: 10.1109/IEDM.1996.554087.

[76] E. Morifuji, "Impact of mechanical stress on hot-carrier lifetime and time-dependent dielectric breakdown in downscaled complimentary metal-oxide-semiconductor," in *Japanese Journal of Applied Physics*, vol. 48, no. 2, pp. 021206-1–021206-5, 2009, doi: 10.1143/JJAP.48.021206.

[77] A. Hamada, T. Furusawa, N. Saito and E. Takeda, "A new aspect of mechanical stress effects in scaled MOS devices," in *IEEE*

Transactions on Electron Devices, vol. 38, no. 4, pp. 895–900, April 1991, doi: 10.1109/16.75220.

[78] J. R. Shih, J. J. Wang, K. Wu, Y. Peng and J. T. Yue, "The study of compressive and tensile stress on MOSFET's I-V, C-V characteristics and it's impacts on hot carrier injection and negative bias temperature instability," *2003 IEEE International Reliability Physics Symposium*, 2003, pp. 612–613, doi: 10.1109/RELPHY.2003.1197831.

[79] Z. Q. Teo, D. S. Ang, K. S. See and P. Z. Yang, "Effect of mechanical strain on the NBTI of short-channel p-MOSFETS: Role of impact ionization," *2009 IEEE International Reliability Physics Symposium*, 2009, pp. 1019–1022, doi: 10.1109/IRPS.2009.5173403.

[80] R. Degraeve, G. Groeseneken, I. De Wolf and H. E. Macs, "The effect of externally imposed mechanical stress on the hot-carrier-induced degradation of deep-sub micron nMOSFET's," in *IEEE Transactions on Electron Devices*, vol. 44, no. 6, pp. 943–950, June 1997, doi: 10.1109/16.585549.

[81] A. Shickova *et al.*, "Negligible effect of process-induced strain on intrinsic NBTI behavior," in *IEEE Electron Device Letters*, vol. 28, no. 3, pp. 242–244, March 2007, doi: 10.1109/LED.2007.891277.

[82] D. P. Ioannou and G. La Rosa, "Mechanical stress effects on p-channel MOSFET performance and NBTI reliability," *2014 IEEE International Reliability Physics Symposium (IRPS)*, 2014, pp. XT.19.1-XT.19.4, doi: 10.1109/IRPS.2014.6861196.

[83] A. Clausner *et al.*, "Analysis of 28 nm SRAM cell stability under mechanical load applied by nanoindentation," *2018 IEEE International Reliability Physics Symposium (IRPS)*, 2018, pp. 5B.1-1–5B.1-6, doi: 10.1109/IRPS.2018.8353607.

[84] Kwan-Yong Lim *et al.*, "Impact of gate sidewall spacer structures on DRAM cell transistors under Fowler-Nordheim and gate-induced drain leakage stress conditions," *2004 IEEE International*

Reliability Physics Symposium (IRPS), 2004, pp. 485-488, doi: 10.1109/RELPHY.2004.1315376.

[85] Y. Sambonsugi and T. Sugii, "Hot-carrier degradation mechanism and promising device design of nMOSFETs with nitride sidewall spacer," *1998 IEEE International Reliability Physics Symposium (IRPS)*, 1998, pp. 184-188, doi: 10.1109/RELPHY.1998.670531.

[86] H. Nii *et al.*, "A 45nm high performance bulk logic platform technology (CMOS6) using ultra high NA(1.07) immersion lithography with hybrid dual-damascene structure and porous low-k BEOL," *2006 International Electron Devices Meeting (IEDM)*, 2006, pp. 1–4, doi: 10.1109/IEDM.2006.346878.

[87] L. T. Clark, V. Vashishtha, L. Shifren, A. Gujja, S. Sinha, B. Cline, C. Ramamurthy and G. Yeric, "ASAP7: A 7-nm finFET predictive process design kit," *Microelectronics Journal*, vol. 53, pp. 105–115, July 2016, doi:10.1016/j.mejo.2016.04.006.

[88] N. Shigyo *et al.*, "Three-Dimensional Analysis of Subthreshold Swing and Transconductance for Fully Recessed Oxide (Trench) Isolated 1/4-μm-Width MOSFET's," in *IEEE Transactions on Electron Devices*, vol. 35, no. 7, pp. 945-951, July 1988, doi: 10.1109/16.3349.

[89] B. Davari *et al.*, "A variable-size shallow trench isolation (STL) technology with diffused sidewall doping for submicron CMOS," *2006 International Electron Devices Meeting (IEDM)*, 1988, pp. 92-95, doi: 10.1109/IEDM.1988.32759.

[90] E. Takeda, T. Hagiwara and A. Shimizu, "Role of Hot-Hole Injection in Hot-Carrier Effects and the Small Degraded Channel Region in MOSFET'S," in *IEEE Electron Device Letters*, vol. 4, no. 9, pp. 329–331, Sept. 1983, doi: 10.1109/EDL.1983.25751.

[91] W. Tonti, R. Bolam and W. Hansch, "Impact of shallow trench isolation on reliability of buried- and surface-channel sub-μm PFET," *995 IEEE International Reliability Physics Symposium (IRPS)*, 1995, pp. 24-29, doi: 10.1109/RELPHY.1995.513648.

[92] K. Hieda, F. Horiguchi, H. Watanabe, K. Sunouchi, I. Inoue and T. Hamamoto, "New Effects of Trench Isolated Transistor Using Side-Wall Gates." *1987 IEEE International Electron Devices Meeting (IEDM)*, 1987, pp. 736–739, doi: 10.1109/IEDM.1987.191536.

[93] S. Baba, A. Kita and J. Ueda, "Mechanism of Hot Carrier Induced Degradation in MOSFET's," *1986 IEEE International Electron Devices Meeting (IEDM)*, 1986, pp. 734–737, doi: 10.1109/IEDM.1986.191298.

[94] S.M. Sze and M-K. Lee, "Semiconductor Devices: Physics and Technology, 3rd Edition" *WILEY*, 2012.

[95] H. Mingam, J. C. Marchetaux and A. Boudou, "Hot-Carrier Stressing Damage in Wide and Narrow LDD NMOS Transistors," in *IEEE Electron Device Letters*, vol. 10, no. 3, pp. 132–134, Mar. 1989, doi: 10.1109/55.31692.

[96] Y. Nishioka, K. Ohyu, Y. Ohji and T. P. Ma, "Channel Length and Width Dependence of Hot-Carrier Hardness in Fluorinated MOSFET's," in *IEEE Electron Device Letters*, vol. 10, no. 12, pp. 540–542, Dec.1989, doi: 10.1109/55.43133.

[97] H. Hwang, J. Lee, P. Fazan, C. Dennison, "Hot-carrier reliability characteristics of narrow-width MOSFETs," in *Solid-State Electronics*, vol. 36, No. 4, pp.665–666,1993.

[98] D. J. DiMaria and J. W. Stasiak, "Trap creation in silicon dioxide produced by hot electrons," in *Journal of Applied Physics*, vol. 65, no. 6, pp. 2342–2356, 1989.

[99] F. Matsuoka, K. Kasai, H. Oyamatsu, M. Kinugawa and K. Maeguchi, "Drain structure optimization for highly reliable deep submicron nMOSFETs with 3.3V high performance operation on the scaling trend," *1990 IEEE International Electron Devices Meeting (IEDM)*, 1990, pp. 833–836, doi: 10.1109/IEDM.1990.237033.

[100] R. Woltjer, A. Hamada and E. Takeda, "Time dependence of p-MOSFET hot-carrier degradation measured and interpreted consistently over ten orders of magnitude," in *IEEE Transactions on Electron Devices*, vol. 40, no. 2, pp. 392–401, 1993, doi: 10.1109/16.182519.

[101] M. Nishigohri *et al.*, "Anomalous hot-carrier induced degradation in very narrow channel nMOSFETs with STI structure," *1996 IEEE International Electron Devices Meeting (IEDM)*, 1996, pp. 881–884, doi: 10.1109/IEDM.1996.554120.

[102] D. A. Buchanan and D. J. DiMaria, "Interface and bulk trap generation in metal-oxide-semiconductor capacitors," in *Journal of Applied Physics*, vol. 67, no. 12, pp. 7439–7452, 1990, doi: 10.1063/1.344534.

[103] S. W. Chung *et al.*, "Novel shallow trench isolation process using flowable oxide CVD for sub-100 nm DRAM," *2002 IEEE International Electron Devices Meeting (IEDM)*, 2002, pp. 233–236, doi: 10.1109/iedm.2002.1175820.

[104] L. Peng, H. Li, T. Sun, X. Gao and Q. Sun, "Study of shallow trench isolation gap fill for 19nm NAND flash," *2020 China Semiconductor Technology International Conference (CSTIC)*, 2020, pp. 27–29, doi: 10.1109/CSTIC49141.2020.9282478.

[105] H. S. Momose *et al.*, "Very lightly nitrided oxide gate MOSFETs for deep-sub-micron CMOS devices," *1991 IEEE International Electron Devices Meeting (IEDM)*, 1991, pp. 359–362, doi : 10.1109/IEDM.1991.235379.

[106] ITRS Roadmap http://www.itrs2.net/

[107] M. Fujiwara, M. Takayanagi and Y. Toyoshima, "New optimization guidelines for sub-0.1 μm CMOS technologies with 2 nm NO gate oxynitrides," *1999 Symposium on VLSI Technology*, 1999, pp. 121-122, doi: 10.1109/VLSIT.1999.799373.

[108] M. Fujiwara, M. Takayanagi, T. Shimizu and Y. Toyoshima, "Extending gate dielectric scaling limit by NO oxynitride: Design and process issues for sub-100 nm technology," *2000 IEEE International Electron Devices Meeting (IEDM)*, 2000, pp. 227–230, doi: 10.1109/iedm.2000.904298.

[109] C. Hobbs *etal.*, "80 nm poly-Si gate CMOS with HfO_2 gate dielectric," *2001 IEEE International Electron Devices Meeting (IEDM)*, 2001, pp. 651–654, doi: 10.1109/iedm.2001.979592.

[110] C. Hobbs *etal.*, "Sub-quarter micron Si-gate CMOS with ZrO_2 gate dielectric," *International Symposium on VLSI Technology System and Application*, pp. 204–207, 2001, doi: 10.1109/vtsa.2001.934520.

[111] G. D. Wilk, R. M. Wallace, and J. M. Anthony, "High-κ gate dielectrics: Current status and materials properties considerations," in *Journal of Applied Physics*, vol. 89, no. 10, pp. 5243–5275, 2001, doi: 10.1063/1.1361065.

[112] Y. Kim *etal.*, "Conventional n-channel MOSFET devices using single layer HfO2 and ZrO2 as high-k gate dielectrics with polysilicon gate electrode," *2001 IEEE International Electron Devices Meeting (IEDM)*, 2001, pp. 455–458, doi: 10.1109/iedm.2001.979538.

[113] T. Iwamoto *etal.*, "A Highly Manufacturable Low Power and High Speed HfSiO CMOS FET with Dual Poly-Si Gate Electrodes," *2003 IEEE International Electron Devices Meeting (IEDM)*, 2003, pp. 639–642, doi: 10.1109/iedm.2003.1269362.

[114] Y. Yasuda *etal.*, "A 65nm-node LSTP (Low standby power) poly-Si/a-Si/HfSiON transistor with high I_{on}-$I_{standby}$ ratio and reliability," *2004 Symposium on VLSI Technology*, 2004, pp. 40–41, doi: 10.1109/VLSIT.2004.1345381.VLSI

[115] Y. Kim *etal.*, "Conventional poly-Si gate MOS-transistors with a novel, ultra-thin Hf-oxide layer," *2003 Symposium on VLSI Technology*, 2003, pp. 167–168, doi: 10.1109/vlsit.2003.1221138.

[116] M. Koyama *et al.*, "Effects of nitrogen in HfSiON gate dielectric on the electrical and thermal characteristics," *2002 IEEE International Electron Devices Meeting (IEDM)*, 2002, pp. 849–852, doi: 10.1109/iedm.2002.1175970.

[117] M. Koyama *et al.*, "Degradation Mechanism of HfSiON Gate Insulator and Effect of Nitrogen Composition on the Statistical Distribution of the Breakdown," *2003 IEEE International Electron Devices Meeting (IEDM)*, 2003, pp. 931–934, doi: 10.1109/iedm.2003.1269431.

[118] M. Koyama *et al.*, "Careful examination on the asymmetric Vfb shift problem for poly-Si/HfSiON gate stack and its solution by the Hf concentration control in the dielectric near the poly-Si interface with small EOT expense," *2004 IEEE International Electron Devices Meeting (IEDM)*,2004, pp. 499–502, doi: 10.1109/iedm.2004.1419200.

[119] T. Watanabe *et al.*, "Design guideline of HfSiON gate dielectrics for 65 nm CMOS generation," *2003 Symposium on VLSI Technology*, 2003, pp. 19–20, doi: 10.1109/VLSIT.2003.1221065.

[120] T. Watanabe *et al.*, "Impact of Hf concentration on performance and reliability for HfSiON-CMOSFET," *2004 IEEE International Electron Devices Meeting (IEDM)*, 2004, pp. 507–510, doi: 10.1109/iedm.2004.1419202.

[121] M. A. Quevedo-Lopez *et al.*, "Thermal stability of hafnium-silicate and plasma-nitrided hafnium silicate films studied by Fourier transform infrared spectroscopy," *Applied Physics Letters*, vol. 87, no. 1, May 2005, doi: 10.1063/1.1977184.

[122] C. Hobbs *et al.*, "Fermi level pinning at the polySi/metal oxide interface," *2003 Symposium on VLSI Technology*, 2003, pp. 9–10, doi: 10.1109/VLSIT.2003.1221060.

[123] D. Matsushita *et al.*, "Novel fabrication process to realize ultra-thin (EOT = 0.7nm) and ultra-low leakage SiON gate dielectrics," *2004 Symposium on VLSI Technology*, 2004, pp. 172–173, doi: 10.1109/VLSIT.2004.1345462.

[124] D. Matsushita *et al.*, "Dramatic improvement of V_{fb} shift and ${G_m}^{max}$ with ultra-thin and ultra-low-leakage SiN-based SiON gate dielectrics," *2005 IEEE International Electron Devices Meeting (IEDM)*, 2005, pp. 828–831, doi: 10.1109/IEDM.2005.1609484.

[125] J. W. Im *et al.*, "A 128Gb 3b/cell V-NAND flash memory with 1Gb/s I/O rate," *2015 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2015, pp. 130–131, doi: 10.1109/ISSCC.2015.7062960.

[126] D. Kang *et al.*, "256Gb 3b/cell V-NAND flash memory with 48 stacked WL layers," *2016 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2016, pp. 130–131, doi: 10.1109/ISSCC.2016.7417941.

[127] C. Kim *et al.*, "A 512Gb 3b/cell 64-stacked WL 3D V-NAND flash memory," *2017 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2017, pp. 202–203, doi: 10.1109/ISSCC.2017.7870331.

[128] R. Yamashita *et al.*, "A 512Gb 3b/cell flash memory on 64-word-line-layer BiCS technology," *2017 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2017, pp. 196–197, doi: 10.1109/ISSCC.2017.7870328.

[129] K. Utsumi *et al.*, "A 65nm low power CMOS platform with 0.495 μm^2 SRAM for digital processing and mobile applications," *2005 Symposium on VLSI Technology*, 2005, pp. 216–217, doi: 10.1109/.2005.1469273.

[130] R. Watanabe *et al.*, "A low power 40nm CMOS technology featuring extremely high density of logic (2100kGate/mm^2) and SRAM (0.195 μm^2) for wide range of mobile applications with wireless system," *2008 IEEE International Electron Devices Meeting (IEDM)*, 2008, pp. 18–21, doi: 10.1109/IEDM.2008.4796773.

[131] S. Hasegawa *et al.*, "A cost-conscious 32nm CMOS platform technology with advanced single exposure lithography and gate-first metal gate/

high-K process," *2008 IEEE International Electron Devices Meeting (IEDM)*, 2008, pp. 21–23, doi: 10.1109/IEDM.2008.4796776.

[132] M. Koike *et al.*, "Effect of Hf-N Bond on Properties of Thermally Stable Amorphous HfSiON and Applicability of this Material to Sub-50nm Technology Node LSIs," *2003 IEEE International Electron Devices Meeting (IEDM)*, 2003, pp. 107–110, doi: 10.1109/iedm.2003.1269177.

[133] Y. Mitani, "Influence of nitrogen in ultra-thin SiON on negative bias temperature instability under AC stress," *2004 IEEE International Electron Devices Meeting (IEDM)*, 2004, pp. 117–120, doi: 10.1109/IEDM.2004.1419082.

[134] Y. Mitani, H. Satake and A. Toriumi, "Influence of nitrogen on negative bias temperature instability in ultrathin SiON," in *IEEE Transactions on Device and Materials Reliability*, vol. 8, no. 1, pp. 6–13, Mar. 2008, doi: 10.1109/TDMR.2008.917314.

[135] M. Sung *et al.*, "Gate-first high-k/metal gate DRAM technology for low power and high performance products," *2015 IEEE International Electron Devices Meeting (IEDM)*, 2015, pp. 680–683, doi: 10.1109/IEDM.2015.7409775.

[136] R. Ritzenthaler *et al.*, "Low-power DRAM-compatible Replacement Gate High-k/Metal Gate stacks," *2012 Proceedings of the European Solid-State Device Research Conference (ESSDERC)*, 2012, pp. 242–245, doi: 10.1109/ESSDERC.2012.6343378.

[137] M. Alavi *et al.*, "PROM element based on salicide agglomeration of poly fuses in a CMOS logic process," *1997 IEEE International Electron Devices Meeting (IEDM)*, pp. 855–858, 1997, doi: 10.1109/iedm.1997.650515.

[138] K. S. Wu *et al.*, "The improvement of electrical programmable fuse with salicide-block dielectric film in 40nm CMOS technology," *roceedings of the IEEE 2010 International Interconnect Technology Conference (IITC)*, 2010, pp. 5–7, doi: 10.1109/IITC.2010.5510461.

[139] H. Takaoka, T. Ueda, H. Tsuda, and A. Ono, "A novel via-fuse technology featuring highly stable blow operation with large on-off ratio for 32nm node and beyond," *2007 IEEE International Electron Devices Meeting (IEDM)*, 2007, pp. 43–46, doi: 10.1109/IEDM.2007.4418858.

[140] N. Matsunaga *et al.*, "BEOL process integration technology for 45 nm node porous low-κ/copper interconnects," *Proceedings of the IEEE 2005 International Interconnect Technology Conference (IITC)*, 2005, pp. 6-8, doi: 10.1109/IITC.2005.1499903.

[141] N. Nakamura *et al.*, "A plasma damage resistant ultra low-k hybrid dielectric structure for 45nm node copper dual-damascene interconnects," *Proceedings of the IEEE 2004 International Interconnect Technology Conference (IITC)*, 2004, pp. 228–230, doi: 10.1109/IITC.2004.1345756.

[142] H. Miyajima *et al.*, "Challenge of low-k materials for 130, 90, 65 nm node interconnect technology and beyond," *2004 IEEE International Electron Devices Meeting (IEDM)*, 2004, pp. 329–332, doi: 10.1109/IEDM.2004.1419147.

[143] A. Bhanap et al.,"Repairing Process-Induced Damage to Porous Low-k ILDs by Post Ash Treatment," *Proceedings of Advanced Metallization Conference (AMC)*, pp. 519–523, 2003.

[144] T. Sasaki, N. Otsuka, K. Hisimo, and S. Fujii, "Melt-segregate-quench programming of electrical fuse," *2005 IEEE International Reliability Physics Symposium (IRPS)*, 2005, pp. 347–351, doi: 10.1109/RELPHY.2005.1493110.

[145] C. A. Cheng, C. T. Ko and K. N. Chen, "Investigation of bonding temperature for SU-8 materials in wafer-level hybrid bonding technology for 3D IC," *The 4th IEEE International NanoElectronics Conference*, 2011, pp. 1–2, doi: 10.1109/INEC.2011.5991657.

[146] S. W. Kim *et al.*, "Ultra-Fine Pitch 3D Integration Using Face-To-Face Hybrid Wafer Bonding Combined with a Via-Middle Through-Silicon-

Via Process," *Electronic Components and Technology Conference (ECTC)*, 2016, pp. 1179–1185, doi: 10.1109/ECTC.2016.205.

[147] D. W. Fisher *et al.*, "Face to Face Hybrid Wafer Bonding for Fine Pitch Applications," *Electronic Components and Technology Conference (ECTC)*, 2020, pp. 595–600, doi: 10.1109/ECTC32862.2020.00099.

[148] W. L. Chiu, O. H. Lee, C. W. Chiang and H. H. Chang, "Low Temperature Wafer-To-Wafer Hybrid Bonding by Nanotwinned Copper," *Electronic Components and Technology Conference (ECTC)*, 2021, pp. 365–370, doi: 10.1109/ECTC32696.2021.00068.

[149] K. Matsudera and K. Kawasaki, "World's first 16-Die Stacked NAND Flash Memory Package Fabricated Using TSV Technology," *Toshiba Review*, vol. 71, no. 6, p. 20, 2016.

[150] C. Y. Huang *et al.*, "Process induced wafer warpage optimization for multi-chip integration on wafer level molded wafer," *Electronic Components and Technology Conference (ECTC)*, 2019, pp. 1287–1293, doi: 10.1109/ECTC.2019.00199.

[151] A. H. Abdelnaby *et al.*, "Numerical simulation of silicon wafer warpage due to thin film residual stresses," *IEEE Workshop on Microelectronics and Electron Devices, WMED*, pp. 9–12, 2013, doi: 10.1109/WMED.2013.6544506.

[152] S. H. Lee, "Technology scaling challenges and opportunities of memory devices," *2016 IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 1.1.1–1.1.8, doi: 10.1109/IEDM.2016.7838026.

[153] S. Gao, Z. Dong, R. Kang, B. Zhang, and D. Guo, "Warping of silicon wafers subjected to back-grinding process," *Precision Engineering*, vol. 40, October, pp. 87–93, 2015, doi: 10.1016/j.precisioneng.2014.10.009.

[154] W. Guo *et al.*, "Impact of through silicon via induced mechanical stress on fully depleted Bulk FinFET technology," *2012 IEEE International*

Electron Devices Meeting (IEDM), pp. 18.4.1–18.4.4, 2012, doi: 10.1109/IEDM.2012.6479066.

[155] Y. Zhu *et al.*, "On the origins of near-surface stresses in silicon around Cu-filled and CNT-filled through silicon vias," *Semiconductor Science and Technology*, vol. 31, no. 5, pp. 1–6, 2016, doi: 10.1088/0268-1242/31/5/055008.

[156] J. Sun, F. Qin, P. Chen, and T. An, "A predictive model of grinding force in silicon wafer self-rotating grinding," *International Journal of Machine Tools & Manufacture*, vol. 109, pp. 74–86, 2016, doi: 10.1016/j.ijmachtools.2016.07.009

[157] C. S. Premachandran *et al.*, "Impact of 3D Via Middle TSV Process on 20nm Wafer Level FEOL and BEOL Reliability," in *Electronic Components and Technology Conference (ECTC)*, 2016, pp. 1593–1598, doi: 10.1109/ECTC.2016.254

[158] Fujitsu Website
https://www.fujitsu.com/global/about/resources/news/press-releases/2018/0906-01.html

[159] H. Miyagawa *et al.*, "Metal-Assisted Solid-Phase Crystallization Process for Vertical Monocrystalline Si Channel in 3D Flash Memory," *2016 IEEE International Electron Devices Meeting (IEDM)*, 2019, pp. 28.3.1–28.3.4, doi: 10.1109/IEDM19573.2019.8993556.

[160] Y. Aiba *et al.*, "Bringing in cryogenics to storage: Characteristics and performance improvement of 3D flash memory," *2021 IEEE International Memory Workshop (IMW)*, 2021, pp. 1–4, doi: 10.1109/IMW51353.2021.9439594.

[161] Y. Aiba *et al.*, "Cryogenic Operation of 3D Flash Memory for New Applications and Bit Cost Scaling with 6-Bit per Cell (HLC) and beyond," *2021 5th IEEE Electron Devices Technology & Manufacturing Conference (EDTM)*, 2021, pp. 9–11, doi: 10.1109/EDTM50988.2021.9421051.

[162] S. Barraud *et al.*, "7-Levels-Stacked Nanosheet GAA Transistors for High Performance Computing," *2020 Symposium on VLSI Technology*, 2020, pp. 7–8, doi: 10.1109/VLSITechnology18217.2020.9265025.

[163] N. Loubet *etal.*, "Stacked Nanosheet Gate-All-Around Transistor to Enable Scaling Beyond FinFET T230 T231," *2017 Symposium on VLSI Technology*, 2017, pp. 14–15, doi: 10.23919/VLSIT.2017.7998183.

[164] J. Ryckaert *etal.*, "The Complementary FET (CFET) for CMOS scaling beyond N3," *2018 Symposium on VLSI Technology*, 2018, pp. 141–142, doi: 10.1109/VLSIT.2018.8510618.

[165] P. Weckx, M. Gupta, Y. Oniki, L. A. Ragnarsson, N. Horiguchi, A. Spessot, D. Verkest, J. Ryckaert, E. D. Litta, D. Yakimets, P. Matagne, P. Schuddinck, D. Jang, B. Chehab, and R. Baert, "Novel forksheet device architecture as ultimate logic scaling device towards 2nm," in *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019, pp. 49–52, doi: 10.1109/IEDM.2017.8268430.

[166] P. Y. Du, H. T. Lue, Y. H. Shih, K. Y. Hsieh, and C. Y. Lu, "Overview of 3D NAND Flash and progress of split-page 3D vertical gate (3DVG) NAND architecture," in *2014 10th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, 2014, pp. 3–6, doi: 10.1109/ICSICT.2014.7021429.

[167] S. Inaba, "3D Flash Memory for Data-Intensive Applications," *2018 IEEE International Memory Workshop (IMW)*, pp. 1–4, 2018, doi: 10.1109/IMW.2018.8388775.

List of papers and presentations

Papers in international journals as first author

1. K. Ishimaru, J.F. Chen and Chenming Hu, "Channel Width Dependence of Hot-Carrier Induced Degradation in Shallow Trench Isolated PMOSFET's," in *IEEE Transactions on Electron Devices*, vol. 46, pp. 1532–1536, July 1999, doi: 10.1109/16.772507.

2. K. Ishimaru, "45nm/32nm CMOS -challenges and perspective-," in *Journal of Solid-State Electronics*, vol. 52, no. 9, pp. 1266–1273, September 2008, doi: 10.1016/j.sse.2008.04.034.

3. K. Ishimaru, M. Tamura, O. Fujii, "Analysis and Optimization of Defect Generation Due to Mechanical Stress in High-Density SRAM," in *IEEE Journal of the Electron Devices Society*, pp. 1103–1109, November 2021, doi: 10.1109/JEDS.2021.3127953.

4. K. Ishimaru, M. Fujiwara, H. Miyagawa and Y. Aiba, "Flash Memory and its Manufacturing Technology for Sustainable World," in *IEEE Journal of the Electron Devices Society (Early Access)*, Dec. 2021, doi: 10.1109/JEDS.2021.3139212.

5. K. Ishimaru, T. Tatsumi, I. Kato, "Characteristics of a-Si:H Thin Films Fabricated at Various Substrate Bias Outside of Microwave Discharge Plasma," in *Transactions on Electronics, IEICE*, vol. E71, no. 4, pp. 299–300, 1988.

Papers in international journals as coauthor

1. J.F. Chen, K. Ishimaru, Chenming Hu, "Enhanced hot-carrier induced degradation in shallow trench isolated narrow channel PMOSFET's," in *IEEE Electron Device Letters*, vol. 19, pp. 332–334, Sept. 1998, doi: 10.1109/55.709632.

2. S. Inaba, K. Miyano, H. Nagano, A. Hokazono, K. Ohuchi, I. Mizushima, H. Oyamatsu, Y. Tsunashima, K. Ishimaru,

Y. Toyoshima and H. Ishiuchi, "SODEL FET: novel channel and source/drain profile engineering schemes by selective Si epitaxial growth technology," in *Transactions on Electron Devices*, vol. 51, issue 9, pp. 1401–1408, Sept. 2004, doi: 10.1109/TED.2004.833573.

3. S. Inaba, H. Nagano, K. Miyano, I. Mizushima, Y. Okayama, T. Nakauchi, K. Ishimaru and H. Ishiuchi, "Low power logic circuit and SRAM cell applications with silicon on depletion layer CMOS (SODEL CMOS) technology," in *IEEE Journal of Solid State Circuit*, pp. 1455–1462, June 2006, doi: 10.1109/JSSC.2006.874335.

4. A. Hokazono, S. Balasuburamanian, K. Ishimaru, H. Ishiuchi, Tue-Jae King Liu and Chenming Hu, "MOSFET design for forward body biasing scheme," in *IEEE Electron Device Letters*, vol. 27, pp. 387–389, May 2006, doi: 10.1109/LED.2006.873382.

5. A. Hokazono, S. Balasuburamanian, K. Ishimaru, H. Ishiuchi, Chenming Hu and Tue-Jae King Liu, "MOSFET hot-carrier reliability improvement by forward-body bias," in *IEEE Electron Device Letters*, vol. 27, pp. 605–608, July 2006, doi: 10.1109/LED.2006.877306.

6. L. Zhang, K. Ohuchi, K. Adachi, K. Ishimaru, M. Takayanagi and A. Nishiyama, "High-resolution characterization of ultrashallow junction by measuring in vacuum with scanning spreading resistance microscopy," in *Applied Physics Letters*, vol. 90, issue 19, pp. 192103-1–192103-3, 2007

7. A. Hokazono, S. Balasuburamanian, K. Ishimaru,H. Ishiuchi, Chenming Hu and Tue-Jae King Liu, "Forward Body Bias as a Bulk-Si CMOS Technology Scaling Strategy," in *IEEE Transactions on Electron Devices*, vol. 55, pp. 2657–2664, Oct. 2008, doi: 10.1109/TED.2008.2003029.

Oral presentations in international conferences as first author

1. K. Ishimaru, F. Matsuoka,T. Maeda, H. Satake, T. Fuse, M. Matsui, Y. Urakawa and H. Momose, "A reverse base current under high level injection and its influence on BiCMOS circuit," *1991 IEEE International Electron Devices Meeting (IEDM)*, pp. 865–868, 1991, doi: 10.1109/IEDM.1991.235288.

2. K. Ishimaru, H. Gojohbori, H. Koike, Y. Unno, M. Sai, F. Matsuoka and M. Kakumu, "Trench isolation technology with 1 μm depth n- and p-wells for a full-CMOS SRAM cell with a 0.4 μm n^+/p^+ spacing," in *1994 Symposium on VLSI Technology*, pp. 97–98, 1994, doi: 10.1109/VLSIT.1994.324437.

3. K. Ishimaru, M. Takahashi, M. Nishigohri, Y. Okayama, Y. Unno, F. Matsuoka and M. Kakumu, "Bipolar installed CMOS technology without any process step increase for high speed cache SRAM," in *1995 IEEE International Electron Devices Meeting (IEDM)*, pp. 673–676, 1995, doi: 10.1109/IEDM.1995.499309.

4. K. Ishimaru, F. Matsuoka, M. Takahashi, M. Nishigohri, Y. Okayama, Y. Unno, M. Yabuki, K. Umezawa, N. Tsuchiya, O. Fujii and M. Kinugawa, "Mechanical Stress Induced MOSFET Punch-through And Process Optimization For Deep Submicron TEOS-O_3 Filled STI Device," in *1997 Symposium on VLSI Technology*, pp. 123–124, 1997, doi: 10.1109/VLSIT.1997.623729.

5. K. Ishimaru, J.F. Chen and C. Hu, "Channel Width Dependence of Hot-Carrier Induced Degradation in Shallow Trench Isolated PMOSFET's," in *28th European Solid-State Device Research Conference*, pp. 240–243, 1998.

6. K. Ishimaru, K. Kasai, Y. Fukaura, Y. Okayama, T. Imamura, S. Irie, T. Hirano, K. Watanabe, M. Ueno, K. Hashimoto and F. Matsuoka, "Combination of TCAD and physical MOSFET model for LSI development time reduction," in *2000 IEEE/SEMI*

Advanced Semiconductor Manufacturing Conference and Workshop, 2000, pp. 103–107, doi: 10.1109/ASMC.2000.902567.

7. K. Ishimaru, M. Takayanagi, T. Watanabe, S. Inaba, M. Fujiwara and D. Matsushita, "Scaled CMOS with SiON and High-k," in *ECS Transactions*, volume 2, Issue 2, pp. 317–327, 2006, doi: 10.1149/1.2195669.

8. K. Ishimaru, "45nm/32nm CMOS -challenges and perspective-," in *Proceedings of European Solid State Device Research Conference*, pp. 32–35, 2007, doi: 10.1109/ESSDERC.2007.4430877.

9. K. Ishimaru, "Non-Volatile Memory Technology for Data Age," in *2018 14th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, 2018, pp. 1215–1218, doi: 10.1109/ICSICT.2018.8564815.

10. K. Ishimaru, "Future of Non-Volatile Memory -From Storage to Computing-," in *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019, pp. 1.3.1–1.3.6, doi: 10.1109/IEDM19573.2019.8993609.

11. K. Ishimaru, "Challenges of Flash Memory for Next Decade," in *2021 IEEE International Reliability Physics Symposium (IRPS)*, 2021, pp. 1–5, doi: 10.1109/IRPS46558.2021.9405182.

12. K. Ishimaru ""Memory" for Sustainable Society," in *2021 20th International Workshop on Junction Technology (IWJT)*, 2021, pp. 1–3, doi: 10.23919/IWJT52818.2021.9609367.

Oral presentations in international conferences as coauthor

1. T. Maeda, K. Ishimaru and H. Momose, "Lower submicron FCBiMOS (fully complementary BiMOS) process with RTP and MeV implanted 5 GHz vertical PNP transistor," in *Digest of Technical papers, Symposium on VLSI Technology*, pp. 79–80, 1990, doi: 10.1109/VLSIT.1990.111017.

2. T. Maeda, H. Gojohbori, K. Inoue, K. Ishimaru, A. Suzuki, H. Kato and M. Kakumu, "High performance BiCMOS technology design for sub-10ns 4 Mbit BiCMOS SRAM with 3.3 V operation," in *Digest of Technical papers, Symposium on VLSI Technology*, pp. 32–33, 1992, doi: 10.1109/VLSIT.1992.200632.

3. T. Yabe, F. Matsuoka, K. Sato, S. Hayakawa, M. Matsui, A. Aono, H. Yoshimura, K. Ishimaru, H. Gojohbori, S. Morita, Y. Unno, M. Kakumu and K. Ochii, "High-speedand low-standby-power circuit design of 1 to 5 V operating 1 Mb full CMOS SRAM," in *Digest of Technical papers, Symposium on VLSI Circuit*, pp. 107–108, 1993, doi: 10.1109/VLSIC.1993.920564.

4. H. Koike, Y. Unno, K. Ishimaru, F. Matsuoka and M. Kakumu, "Dual Polycide gate and dual buried contact technology achieving a 0.4μm nMOS/pMOS spacing for a 7.65μm^2 full-CMOS SRAM cell," in *Technical Digest of IEDM*, pp. 855–858, 1994, doi: 10.1109/IEDM.1994.383278.

5. A. Suzuki, T. Kobayashi, T. Hamano, H. Hatada, A. Kawasumi, F. Matsuoka, K. Ishimaru, M. Takahashi, M. Nishigohri, Y. Okayama, Y. Unno, M. Kakumu and J. Tsujimoto, "A 400 MHz 4.5 Mb synchronous BiCMOS SRAM with alternating bit-line loads," in *Digest of Technical papers, ISSCC*, pp. 146–147, 1996, doi: 10.1109/ISSCC.1996.488546.

6. M. Nishigohri, K. Ishimaru, M. Takahashi, Y. Unno, Y. Okayama, F. Matsuoka and M. Kinugawa, "Anomalous hot-carrier induced degradation in very narrow channel nMOSFETs with STI structure," in *Technical Digest of IEDM*, pp. 881–884, 1996, doi: 10.1109/iedm.1996.554120.

7. E. Morifuji, T. Ohguro, H. Kimijima, T. Yoshitomi, H.S. Momose, Y. Katsumata, K. Ishimaru, F. Matsuoka, M. Kinugawa and H. Iwai, "RF modeling for 0.1μm gate length MOSFETs," in *Proceedings of ESSDERC*, pp. 656–659, 1999.

8. T. Yoshitomi, Y. Ebuchi, H. Kimijima, T. Ohguro, E. Morifuji, H.S. Momose, K. Kasai, K. Ishimaru, F. Matsuoka, Y. Katsumata, M. Kinugawa and H. Iwai, "High performance MIM capacitor for RFBiCMOS/CMOS LSIs," in *Proceedings of BCTM*, pp. 133–136, 1999, doi: 10.1109/BIPOL.1999.803543.

9. K. Miyashita, T. Nakayama, A. Oishi, R. Hasumi, M. Owada, S. Aota, Y. Okayama, M. Matsumoto, H. Igarashi, T. Yoshida, K. Kasai, T. Yoshitomi, Y. Fukaura, H. Kawasaki, K. Ishimaru, K. Adachi, M. Fujiwara, K. Ohuchi, M. Takayanagi, H. Oyamatsu, F. Matsuoka, T. Noguchi and M. Kakumu, "A high performance 100 nm generation SOC technology (CMOS IV) for high density embedded memory and mixed signal LSIs," in *Digest of Technical papers, Symposium on VLSI Technology*, pp. 11–12, 2001, doi: 10.1109/vlsit.2001.934922.

10. Y. Fukaura, K. Kasai, Y. Okayama, H. Kawasai, K. Isobe, M. Kanda, K. Ishimaru and H. Ishiuchi, "A highly manufacturable high density embedded SRAM technology for 90 nm CMOS," in *Technical Digest of IEDM*, pp. 415–418, 2002, doi: 10.1109/iedm.2002.1175867.

11. M. Matsuo, K. Kasai, Y. Okayama, K. Ishimaru, N. Matsunaga, H. Yamaguchi, N. Ohtsuka, and N. Hayasaka, "Chip-on-chip technology with copper through-plug for 0.15 μm SRAM," in *Proceedings of IITC*, pp. 174–176, 2003, doi: 10.1109/IITC.2003.1219746.

12. T. Ohguro, Y. Okayama, K. Matsuzawa, K. Matsunaga, N. Aoki, K. Kojima, H.S. Momose and K. Ishimaru, "The impact of oxynitride process, deuterium annealing and STI stress to 1/f noise of 0.11 μm CMOS," in *Digest of Technical papers, Symposium on VLSI Technology*, pp. 37–38, 2003, doi: 10.1109/VLSIT.2003.1221074.

13. T. Watanabe, M. Takayanagi, R. Iijima, K. Ishimaru, H. Ishiuchi and Y. Tsunashima, "Design guideline of HfSiON gate dielectrics for 65 nm CMOS generation," in *Digest of Technical papers, Symposium on VLSI Technology*, pp. 19–20, 2003, doi: 10.1109/VLSIT.2003.1221065.

14. M. Iwai, A. Oishi, T. Sanuki, Y. Takegawa, T. Komoda, Y. Morimasa, K. Ishimaru, M. Takayanagi, K. Eguchi, D. Matsushita, K. Muraoka, K. Sunouchi and T. Noguchi, "45nm CMOS platform technology (CMOS6) with high density embedded memories," in *Digest of Technical papers, Symposium on VLSI Technology*, pp. 12–13, 2004, doi: 10.1109/VLSIT.2004.1345364.

15. N. Yasutake, K. Ohuchi, M. Fujiwara, K. Adachi, A. Hokazono, K. Kojima, N. Aoki, H. Shuto,T. Watanabe, T. Morooka, H. Mizuno, S. Magoshi, T. Shimizu, S. Mori, H. Oguma, T. Sasaki, M. Ohmura, K. Miyano, H. Yamada, H. Tomita, D. Matsushita, K. Muraoka, S. Inaba, M. Takayanagi, K. Ishimaru and H. Ishiuchi, "A hp22 nm node low operating power (LOP) technology with sub-10 nm gate length planar bulk CMOS devices," in *Digest of Technical papers, Symposium on VLSI Technology*, pp. 84–85, 2004, doi: 10.1109/VLSIT.2004.1345407.

16. T. Ohguro, N. Sato, M. Matsuo, K. Kojima, H.S. Momose, K. Ishimaru and H. Ishiuchi, "Ultra-thin chip with permalloy film for high performance MS/RF CMOS," in *Digest of Technical papers, Symposium on VLSI Technology*, pp. 220–221, 2004, doi: 10.1109/VLSIT.2004.1345490.

17. S. Inaba, H. Nagano, K. Miyano, I. Mizushima, Y. Okayama, T. Nakauchi, K. Ishimaru and H. Ishiuchi, "Low power logic circuit and SRAM cell applications with silicon on depletion layer CMOS (SODEL CMOS) technology," in *Proceedings of CICC*, pp. 225–228, 2004, doi: 10.1109/CICC.2004.1358783.

18. H. Suto, S. Inaba and K. Ishimaru, "Error evaluation of C-V characteristic measurements in ultra-thin gate dielectrics," in *Proceedings of ICMTS*, pp. 221–226, 2004, doi: 10.1109/ICMTS.2004.1309483.

19. M. Takayanagi, T. Watanabe, R. Iijima, K. Ishimaru and Y. Tsunashima, "Investigation of hot carrier effects in n-MISFETs with HfSiON gate dielectric," in *Proceedings of IRPS*, pp. 13–17, 2004, doi: 10.1109/RELPHY.2004.1315294.

20. T. Watanabe, M. Takayanagi, R. Iijima, K. Ishimaru, H. Ishiuchi and Y. Tsunashima, "Impact of Hf concentration on performance and reliability for HfSiON-CMOSFET," in *Technical Digest of IEDM*, pp. 507–510, 2004, doi: 10.1109/iedm.2004.1419202.

21. H. Kawasaki, K. Ohuchi, A. Oishi, O. Fujii, H. Tsujii, T. Ishida, K. Kasai, Y. Okayama, K. Kojima, K. Adachi, N. Aoki, T. Kanemura, D. Hagishima, M. Fujiwara, S. Inaba, K. Ishimaru, N. Nagashima and H. Ishiuchi, "Impact of parasitic resistance and silicon layer thickness scaling for strained-silicon MOSFETs on relaxed Si_{1-x}/Ge_x virtual substrate," in *Technical Digest of IEDM*, pp. 169–172, 2004, doi: 10.1109/IEDM.2004.1419098.

22. K. Okano, T. Izumida, H. Kawasaki, A. Kaneko, A. Yagishita, T. Kanemura, M. Kondo, S. Ito, N. Aoki, K. Miyano, T. Ono, K. Yahashi, K. Iwade, T. Kubota, T. Matsushita, I. Mizushima, S. Inaba, K. Ishimaru, K. Suguro, K. Eguchi, Y. Tsunashima and H. Ishiuchi, "Process integration technology and device characteristics of CMOS FinFET on bulk silicon substrate with sub-10 nm fin width and 20 nm gate length," in *Technical Digest of IEDM*, pp. 721–724, 2005, doi: 10.1109/IEDM.2005.1609454.

23. A. Kaneko, A. Yagishita, K. Yahashi, T. Kubota, M. Omura, K. Matsuo, I. Mizushima, K. Okano, H. Kawasaki, S. Inaba, T. Izumida, T. Kanemura, N. Aoki, K. Ishimaru, H. Ishiuchi, K. Suguro, K. Eguchi and Y. Tsunashima, "Sidewall transfer process and selective gate sidewall spacer formation technology for sub-15nm finfet with elevated source/drain extension," in *Technical Digest of IEDM*, pp. 844–847, 2005, doi: 10.1109/IEDM.2005.1609488.

24. M. Fujiwara, T. Morooka, N. Yasutake, K. Ohuchi, N. Aoki, H. Tanimoto, M. Kondo, K. Miyano, S. Inaba, K. Ishimaru and H. Ishiuchi, "Impact of BOX scaling on 30 nm gate length FD SOI MOSFET," in *Proceedings of SOI Conference*, pp. 180–182, 2005, doi: 10.1109/SOI.2005.1563581.

25. K. Adachi, K. Ohuchi, N. Aoki, H. Tsujii, T. Ito, H. Itokawa, K. Matsuo, K. Suguro, Y. Honguh, N. Tamaoki, K. Ishimaru and H. Ishiuchi, "Issues and optimization of millisecond anneal process for 45 nm node and beyond," in *Digest of Technical papers, Symposium on VLSI Technology*, pp. 142–143, 2005, doi: 10.1109/.2005.1469245.

26. H. Tsuji, K. Adachi, K. Ohuchi, N. Aoki, T. Ito, K. Matsuo, K. Suguro, K. Ishimaru and H. Ishiuchi, "Design guideline for halo condition on CMOSFETs utilizing FLA," in *Proceedings of IWJT*, pp. 111–114, 2005, doi: 10.1109/IWJT.2005.203897.

27. M. Takayanagi, T. Watanabe, R. Iijima, M. Koyama, M. Koike, T. Ino, Y. Kamimuta, K. Sekine, K. Eguchi, A. Nishiyama and K. Ishimaru, "HfSiON-CMOSFET technology for low standby power application," in *Technical Digest of IEDM*, pp. 882–885, 2005, doi: 10.1109/IEDM.2005.1609499.

28. H. Kawasaki, K. Okano, A. Kaneko, A. Yagishita, T. Izumida, T. Kanemura, K. Kasai, T. Ishida, T. Sasaki, Y. Takeyama, N. Aoki, N. Ohtsuka, K. Suguro, K. Eguchi, Y. Tsunashima, S. Inaba, K. Ishimaru and H. Ishiuchi, "Embedded Bulk FinFET SRAM Cell Technology with Planar FET Peripheral Circuit for hp32 nm Node and Beyond," in *Digest of Technical Papers, Symposium on VLSI Technology*, pp. 70–71, 2006, doi: 10.1109/VLSIT.2006.1705221.

29. K. Adachi, K. Ohuchi, N. Aoki, H. Tanimoto, H. Tsujii, P. Eyben, D. Vanhaeren, W. Vandervorst, K. Ishimaru and H. Ishiuchi, "Direct Observation of 2-D Dopant Profiles of MOSFETs Activated by Millisecond Anneal," in *Proceedings of IWJT*, pp. 104–107, 2006, doi: 10.1109/IWJT.2006.220871.

30. L. Zhang, K. Ohuchi, K. Adachi, K. Ishimaru, M. Takayanagi and N. Fukushima, "Reproducible and High-Resolution Analysis on ultra-Shallow-Junction CMOSFETs by Scanning Spreading Resistance Microscopy," in *Proceedings of IWJT*, pp. 108–111, 2006, doi: 10.1109/IWJT.2006.220872.

31. T. Ohguro, K. Kojima, R. Iijima, T. Watanabe, M. Takayanagi, H.S. Momose, K. Ishimaru and H. Ishiuchi, "Analysis of 1/f noise for CMOS with high-k gate dielectrics," in *Proceedings of ICSICT*, pp. 400–403, 2006, doi: 10.1109/ICSICT.2006.306262.

32. T. Kanemura, T. Izumida, N. Aoki, M. Kondo, S. Ito, T. Enda, K. Okano, H. Kawasaki, A. Yagishita, A. Kaneko, S. Inaba, M. Nakamura, K. Ishimaru, K. Suguro, K. Eguchi and H. Ishiuchi, "Improvement of DriveCurrent in Bulk-FinFET using Full 3D Process/Device Simulations," in *Proceedings of SISPAD*, pp. 131–134, 2006, doi: 10.1109/SISPAD.2006.282855.

33. S. Inaba, K. Okano, T. Izumida, A. Kaneko, H. Kawasaki, A. Yagishita, T. Kanemura, T. Ishida, N. Aoki, K. Ishimaru, K. Suguro, K. Eguchi, Y. Tsunashima, Y. Toyoshima and H. Ishiuchi, "FinFET: the prospective multi-gate device for future SoC application," in *Proceedings of ESSDERC*, pp. 49–52, 2006, doi: 10.1109/ESSDER.2006.307635.

34. A. Kaneko, A. Yagishita, K. Yahashi, T. Kubota, M. Omura, K. Matsuo, I. Mizushima, K. Okano, H. Kawasaki, T. Izumida, T. Kanemura, N. Aoki, A. Kinoshita, J. Koga, S. Inaba, K. Ishimaru, Y. Toyoshima, H. Ishiuchi, K. Suguro, K. Eguchi and Y. Tsunashima, "High-Performance FinFET with Dopant-Segregated Schottky Source/Drain," in *Technical Digest of IEDM*, pp. 893–896, 2006, doi: 10.1109/IEDM.2006.346926.

35. A. Hokazono, S. Kawanaka, K. Tsumura, Y. Hayashi, H. Tanimoto, T. Enda, N. Aoki, K. Ohuchi, S. Inaba, K. Okano, M. Fujiwara, T. Morooka, M. Goto, A. Kajita, T. Usui, K. Ishimaru and Y. Toyoshima, "Guideline for Low-temperature-operation Technique to Extend CMOS Scaling," in *Technical Digest of IEDM*, pp. 675–678, 2006, doi: 10.1109/IEDM.2006.346875.

36. N. Yasutake, T. Ishida, K. Ohuchi, N. Aoki, N. Kusunoki, S. Mori, I. Mizushima, T. Morooka, K. Yahashi, S. Kawanaka, K. Ishimaru and H. Ishiuchi, "A High Performance pMOSFET with Two-step Recessed

SiGe-S/D Structure for 32nm node and Beyond," in *Proceedings of ESSDERC*, pp. 77–80, 2006, doi: 10.1109/ESSDER.2006.307642.

37. T. Ikehashi, T. Ohguro, E. Ogawa, H. Yamazaki, K. Kojima, M. Matsuo, K. Ishimaru and H. Ishiuchi, "A Robust RF MEMS Variable Capacitor with Piezoelectric and Electrostatic Actuation," in *Proceedings of MTT-S*, pp. 39–42, 2006, doi: 10.1109/MWSYM.2006.249903.

38. S. Inaba, H. Kawasaki, K. Okano, T. Izumida, A. Yagishita, A. Kaneko, K. Ishimaru, N. Aoki and Y. Toyoshima, "Direct evaluation of DC characteristic valiability in FinFET SRAM Cell for 32 nm node and beyond," in *Technical Digest of IEDM*, pp. 487–490, 2007, doi: 10.1109/IEDM.2007.4418980.

39. H. Kawasaki, A. Kaneko, A. Yagishita, K. Okano, T. Izumida, T. Kanemura, K. Suguro, K. Eguchi, Y. Tsunashima, S. Inaba, N. Aok, K. Ishimaru and Y. Toyoshima, "FinFET Process and Integration Technlogy for High Performance LSI in 22 nm node and beyond," in *Proceedings of IWJT*, pp. 3–8, 2007, doi: 10.1109/IWJT.2007.4279933.

40. Y. Haizhou, C.Y. Sung, K.L. Saenger, M. Hamaguchi, R. Hasumi, K. Ohuchi, H. Ng, R. Zhang, K.J. Stein, T.A. Wallner, J. Li, J.A. Ott, X. Chen, Z.J. Luo, N. Rovedo, K. Fogel, G. Pfeiffer, R. Klemhenz, R. Bendernagel, D.K. Sadana, M. Takayanagi, K. Ishimaru, S.W. Crowder, D. Park, M. Khare and G. Shahidi, "Scalability of Direct Silicon Bonded (DSB) Technology for 32 nm Node and Beyond," in *Digest of Technical papers, Symposium on VLSI Technology*, pp. 222–223, 2007, doi: 10.1109/VLSIT.2007.4339701.

41. K. Ohuchi, C. Lavoie, C.E. Murray, C.P. D ' Emic, J.O. Chu, B. Yang, P. Besser, L.M. Gignac, J. Bruley, G.U. Singco, F. Pagette, A.W. Topol, M.J. Rooks, J.J. Bucchignano, V. Narayanan, M. Khare, M. Takayanagi, K. Ishimaru, D.G. Park, G. Shahidi and P.M. Solomon, "Extendibility of NiPt Silicide Contacts for CMOS Technology Demonstrated to the 22-nm Node," in *Technical Digest of IEDM*, pp. 1029–1032, 2007, doi: 10.1109/IEDM.2007.4418888.

42. H. Kawasaki, M. Khater, M. Guillorn, N. Fuller, J. Chang, S. Kanakasabapathy, L. Chang, R. Muralidhar, K. Babich, Q. Yang, J. Ott, D. Klaus, E. Kratschmer, E. Sikorski, R. Miller, R. Viswanathan, Y. Zhang, J. Silverman, Q. Ouyang, A. Yagishita, M. Takayanagi, W. Haensch and K. Ishimaru, "Demonstration of highly scaled FinFET SRAM cells with high-k/metal gate and investigation of characteristic variability for the 32 nm node and beyond," in *Technical Digest of IEDM*, pp. 237–240, 2008, doi: 10.1109/IEDM.2008.4796661.

43. K. Ohuchi, C. Lavoie, C.E. Murray, C.P. D ' Emic, I. Lauer, J.O. Chu, B. Yang, P. Besser, L.M. Gignac, J. Bruley, G.U. Singco, F. Pagette, A.W. Topol, M.J. Rooks, J.J. Bucchignano, V. Narayanan, M. Khare, M. Takayanagi, K. Ishimaru, D.G. Park, G. Shahidi, P.M. Solomon, "Extendibility of NiPt Silicide to the 22-nm node CMOS technology," in *Extended Abstract IWJT*, pp. 150–153, 2008, doi: 10.1109/IWJT.2008.4540037.

44. M. Hamaguchi, H. Yin, K.L. Saenger, C.Y. Sung, R. Hasumi, R. Iijima, K. Ohuchi, Y. Takasu, J.A. Ott, H. Kang, M. Biscardi, J. Li, A.G. Domenicucci, Z. Zhu, P. Ronsheim, R. Zhang, N. Rovedo, H. Utomo, K. Fogel, J.P. De Souza, D.K. Sadana, M. Takayanagi, D. Park, G. Shahidi and K. Ishimaru, "Higher hole mobility induced by twisted Direct Silicon Bonding (DSB)," in *Digest of Technical papers, Symposium on VLSI Technology*, pp. 178–179, 2008, doi: 10.1109/VLSIT.2008.4588609.

45. Y. Haizhou, M. Hamaguchi, K.L. Saenger, C.Y. Sung, R. Hasumi, K. Ohuchi, R. Zhang, J. Cai, J.A. Ott, X. Chen, Z.J. Luo, N. Rovedo, K. Fogel, G. Pfeiffer, R. Kleinhenz, D.K. Sadana, M. Takayanagi, K. Ishimaru, T.H. Ning, D.-G. Park, M. Khare and G. Shahidi, "Effect of End-of-Range Defects on Device Leakage in Direct Silicon Bonded (DSB) Technology," in *Proceedings of VLSI-TSA* , pp. 34-35, 2008, doi: 10.1109/VTSA.2008.4530786.

46. H. Kawasaki, V.S. Basker, T. Yamashita, C.-H. Lin, Y. Zhu,

J. Faltermeier, S. Scumitz, J. Cummings, S. Kanakasabapathy, H. Adhikari, H. Jagannathan, A. Kumar, K. Maitra, J. Wang, C.-C. Yeh, C. Wang, M. Khater, M. Guillorn, N. Fuller, J. Chang, L. Chang, R. Muralidhar, A. Yagishita, R. Miller, Q. Ouyang, Y. Zhang, V.K. Parchuri, H. Bu, B. Doris, M. Takayanagi, W. Haensch, D. McHerron, J. O ' Neill and K. Ishimaru, "Challenge and solutions of FinFET integration in an SRAM cell and logic circuit for 22 nm node and beyond," in *Technical Digest of IEDM*, pp. 289–292, 2009, doi: 10.1109/IEDM.2009.5424366.

47. A. Isobayashi, J.J. Kelly, T. Watanabe, M. Fujiwara, C. Koburger, J. Maniscalco, V.Tuan, S.K. Chiang, J. Ren, T. Spooner, M. Takayaagi, T. Usui and K. Ishimaru, "Robust and low cost copper contact application for low power device at 32 nm-Node and beyond," in *Proceedings of IITC*, pp. 5–7, 2009, doi: 10.1109/IITC.2009.5090325.

48. Q. Liu, A. Yagishita, N. Loubet, A. Khakifirooz, P. Kulkarni, T. Yamamoto, K. Cheng, M. Fujiwara, J. Cai, D. Dorman, S. Mehta, P. Khare, K. Yako, Y. Zhu, S. Mignot, S. Kanakasabapathy, S. Monfray, F. Boeuf, C. Koburger, H. Sunamura, S. Ponoth, A. Reznicek, B. Haran, A. Upham, R. Johnson, L.F. Edge, J. Kuss, T. Levin, N. Berliner, E. Leobandung, T. Skotnicki, M. Hane, H. Bu, K. Ishimaru, W. Kleemeier, M. Takayanagi, B. Doris and R. Sampson, "Ultra-thin-body and BOX (UTBB) fully depleted (FD) device integration for 22nm node and beyond," in *Digest of Technical papers, Symposium on VLSI Technology*, pp. 61–62, 2010, doi: 10.1109/VLSIT.2010.5556120.

49. K. Kakehi, H. Aikawa, T. Tadokoro, H. Eguchi, T. Hirayu, H. Yoshimura, T. Asami and K. Ishimaru, "An efficient manufacturing technique based on process compact model to reduce characteristics variation beyond process limit for 40 nm node mass production," in *Digest of Technical papers, Symposium on VLSI Technology*, pp. 90–91, 2011.

50. Q. Liu, F. Monsieur, A. Kumar, T. Yamamoto, A. Yagishita,

P. Kulkarni, S. Ponoth, N. Loubet, K. Cheng, A. Khakifirooz, B. Haran, M. Vinet, J. Cai, J. Kuss, B. Linder, L. Grenouillet, S. Mehta, P. Khare, N. Berliner, T. Levin, S. Kanakasabapathy, A. Upham, R. Sreenivasan, Y. Le Tiec, N. Posseme, J. Li, J. Demarest, S. Smalley, E. Leobandung, S. Monfray, F. Boeuf, T. Skotnicki, K. Ishimaru, M. Takayanagi, W. Kleemeier, H. Bu, S. Luning, T. Hook, M. Khare, G. Shahidi, B. Doris and R. Sampson, "Impact of back bias on ultra-thin body and BOX (UTBB) devices," in *Digest of Technical papers, Symposium on VLSI Technology*, pp. 160–161, 2011.

51. S. Shimizu, H. Aikawa, S. Okamoto, K. Kakehi, K. Ohsawa, H. Yoshimura, T. Asami and K. Ishimaru, "Comprehensive study of systematic and random variation in Gate-Induced Drain Leakage for LSTP applications," in *Digest of Technical papers, Symposium on VLSI Technology*, pp. 196–197, 2011.

52. T. Miyata, H. Tanaka, K. Kagimoto, M. Kamiyashiki, M. Kamimura, A. Hidaka, M. Goto, K. Adachi, A. Hokazono, T. Ohguro, K. Nagaoka, Y. Watanabe, S. Hirooka, Y. Ito, S. Kawanaka and K. Ishimaru, "150 GHz FMAX with high drain breakdown voltage immunity by multi gate oxide dual work-function (MGO-DWF)-MO SFET," in *Technical Digest of IEDM*, 2015, pp. 25.8.1–25.8.4, doi: 10.1109/IEDM.2015.7409769.