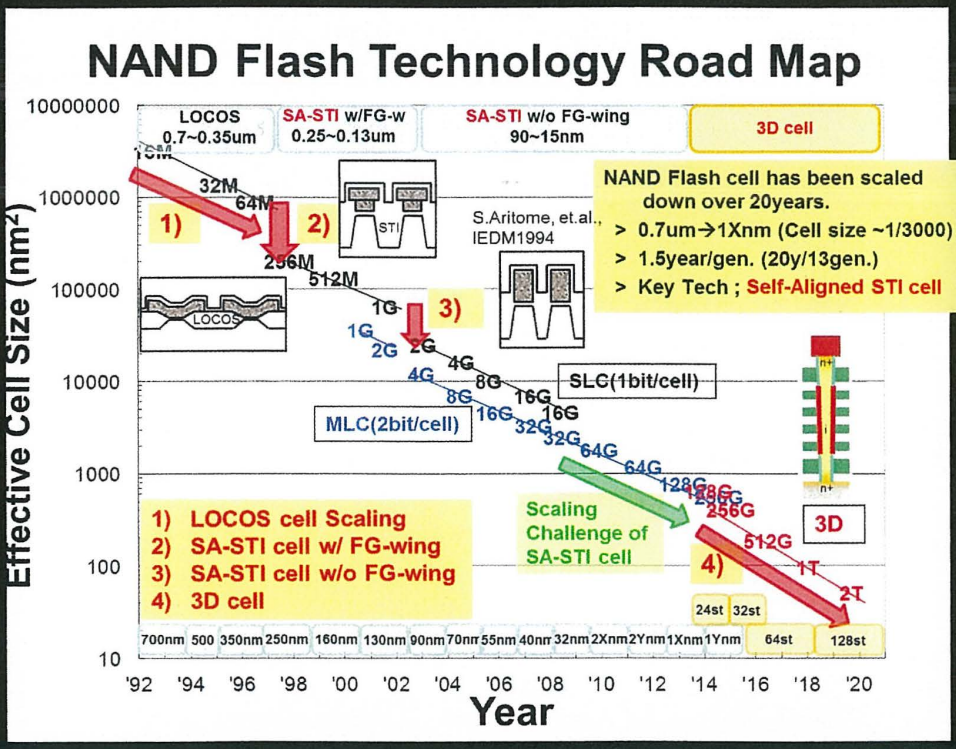
IEEE Press Series on Microelectronic Systems

R. Jacob Baker, Series Editor

NAND Flash Memory Technologies

Seiichi Aritome



IEEE PRESS

WILEY

NAND FLASH MEMORY TECHNOLOGIES

IEEE Press
445 Hoes Lane
Piscataway, NJ 08854

Technical Reviewers

Joe E Brewer, *PE, Electronic Engineering Consultant*
Chandra Mouli, *Director, R&D Device Technology, Micron Technology Inc, Boise ID, USA*
Gabriel Molas, *CEA LETI Minatec, Grenoble, France*

NAND FLASH MEMORY TECHNOLOGIES

SEIICHI ARITOME

IEEE Press Series on Microelectronic Systems

IEEE PRESS

WILEY

Library of Congress Cataloging-in-Publication Data is available.

ISBN: 978-1-119-13260-8

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

FOREWORD

I had an idea regarding NAND flash memory when I was in Washington, D.C. in 1986.

My stay in Washington, D.C. at that time was long. All of Japan's DRAM manufacturers were sued by the United States' International Trade Commission (ITC) with the demand that they be barred from exporting to the United States because they were infringing upon Texas Instruments' DRAM patents. I was dispatched to Washington, D.C. as the engineer handling the ITC lawsuit. Court session at the ITC was tough to hold continuously from 6 am to midnight. However, the court wasn't held every day. In those days, I had plenty of free time. I thought deeply about future semiconductor memory, so that NOR flash memory that I invented a few years earlier was too weak to drive out magnetic disks, and we needed to further reduce the bit cost and further shrink the per-bit space occupied. The answer was "NAND flash memory." I immediately wrote up the patent in 1986, and I filed the application on April 24, 1987 in Japan. The US patent for NAND Flash memory was registered as USP 5,245,566.

After returning to Japan, I started the development of NAND flash memory at the VLSI Research Center, Toshiba Corporation, in 1987. I involved several engineers to make the development team for NAND flash. The basic data of reading and writing were obtained in a short time. We immediately submitted a paper to the 1987 IEDM. For further acceleration of NAND flash development, I assigned several members—including Dr. Aritome, who is the author of this book—to a flash team from a DRAM team.

After that, I aggressively drove the development. I proposed the design of a prototype device of 4-Mbit NAND flash memory. However, unfortunately, there was not enough budget at the research center to proceed with this project. It was at

a crisis point to suspend the development of NAND flash. I had to somehow find someone to fund the NAND project so that it could continue. I first visited the division developing computers in Toshiba, but their response was that they wouldn't fund a dream-like project for replacing magnetic memory on a semiconductor. Meanwhile, I explained to Tajiri, the Director of the Consumer Electronics Laboratory, that if we managed to produce 4-Mbit NAND flash memory, cameras would no longer need film. I was indeed explaining that the digital cameras we know today would become possible. As a result, the Consumer Electronics Laboratory bore the development costs. We successfully developed 4-Mbit NAND flash memory in 1988 and announced 4-Mbit NAND flash memory at the ISSCC in February 1989. Thereafter, Consumer Electronics Laboratory Director Tajiri used 4-Mbit NAND flash memory to launch the world's first digital camera to replace conventional film with NAND flash memory. At the time, the price of the world's first flash-memory-based camera was high, more than two million yen ($20K), and thus it didn't sell well.

In 1992, a production of NAND flash was started. The first device was 0.7-μm rule 16-Mbit memory. The production volume was really small; however, it was an important milestone. For large volume production, we had to wait 4 or 5 years to create the market for flash memory cards mainly for digital cameras. After producing memory cards, the market for the NAND flash was amazingly huge. It was the disruptive innovations. Music players with cassette tape were replaced by flash-memory-based MP3 portable music players. USB memory arrived, and thus the floppy disk disappeared. Smartphone and tablet PC were designed based on the existence of NAND flash. Nowadays, the NAND flash memory became the standard nonvolatile memories that are used everywhere by everybody. However, the dream to replace magnetic memory (HDD, etc.) is along the way. I am expecting that the SSD will replace the HDD in the future.

I left Toshiba Corporation and was transferred to Tohoku University as a professor in 1994. I proposed the SGT (surrounding gate transistor) NAND flash memory, and I also started fundamental development of SGT for three-dimensional (3D) NAND flash. *Forbes* magazine published a structural drawing of the SGT NAND flash memory with my photograph in the cover of the June 24, 2002 edition. The cell structure of SGT NAND flash memory is currently used in 3D NAND flash memories that are in mass production. All NAND suppliers are intensively developing the next advanced 3D NAND flash based on SGT structure.

As the market for NAND flash memory expanded, engineers who are engaged in the development of NAND flash and its related products rapidly increased. This book is good for understanding the history, basic structure and process, scaling issues, 3D NAND flash, and so on. Dr. Aritome is one of the original members of the NAND flash development team and has over 27 years of experience as an engineer of development and production of NAND flash memory. I hope this book will contribute to the coming NAND flash technologies and products, including SSD.

Finally, I am grateful to the original team members of NAND flash development. NAND flash memory could not be realized without their contributions. I am very

happy and lucky that I could collaborate with them so that we could devote ourselves to developing NAND flash memories.

FUJIO MASUOKA

CTO of Semicon Consulting Ltd.
Professor Emeritus at Tohoku University

BIOGRAPHICAL NOTE

Dr. Fujio Masuoka is Chief Technology Officer of Semicon Consulting Ltd., which forms part of the "New Scope Group," a largely respected international group of companies actively pursuing and supporting advanced Research & Development of Breakthrough Technologies and translating these into commercial reality. He is also Professor Emeritus of the Research Institute of Electrical Communication at Tohoku University in Japan. He is the inventor of flash memory. He has spent most of his career working on the research and development of numerous kinds of semiconductor memory including flash memory, programmable read-only memory, and random access memory. He also possesses considerable knowledge in image sensing devices (such as charge-coupled devices) and high-speed semiconductor logic. He filled the original patents for both NOR and NAND flash memories, published the first paper on flash memory at the 1984 IEDM, and published the first paper on NAND flash memory at the 1987 IEDM.

CAREER PROFILE

1966 Graduated from Faculty of Engineering, Tohoku University
1971 Completed the doctoral course, Tohoku University
1971 Joined Toshiba Corporation
1994 Appointed Professor at Tohoku University
2005 Accepted as the Chief Technology Officer of Unisantis Electronics (Japan) Ltd.
2007 Appointed Honorary Professor at Tohoku University

AWARDS AND RECOGNITION

1977 Awarded the Watanabe Prize during the year of its inception
1980 Awarded the invention award, National Invention Awards
1985 Awarded the encouraging award for invention, Kanto district
1986 Awarded the encouraging award for invention, Kanto district
1988 Awarded the encouraging award, twice in the year, for invention, Kanto district
1991 Awarded the encouraging award for invention, Kanto district
1995 IEEE Fellow

1997 Awarded the IEEE Morris N. Liebmann Memorial Award
2000 Awarded the Ichimura-Sangyo Prize (major award)
2002 Awarded the 2002 SSDM Award
2005 Awarded Innovation Award by the *Economist*
2007 Awarded Medal with Purple Ribbon from Emperor Akihito of Japan
2009 Flash memory recognized in the IEEE as one of "25 Microchips That Shook The World"
2010 Computer History Museum
2011 Consumer Electronics Hall of Fame
2012 The winner of the Progress Medal, the highest honor of the Photographic Society of America (PSA)
2013 Awarded the Flash Memory Summit Lifetime Achievement Award
2013 Bunkakorosha (Person of Cultural Merits of Japan)

PREFACE

NAND flash memory became a standard semiconductor nonvolatile memory. Everyone in the world has widely used NAND flash memory in many applications, such as digital cameras, USB drives, MP3 music players, smartphones, and tablet PCs. The cloud data server starts to use SSD (Solid State Drive), which is based on NAND flash memory. Recently, three-dimensional (3D) NAND flash memory was developed and started mass production for reducing bit cost. By using 3D NAND flash memory, an advanced SSD has been intensively developed for high performance and low power consumption to avoid damaging the ecological environment.

As the production volume of NAND flash memory has increased, the number of engineers who are engaged in development and production of NAND flash memory has also increased. And a lot of people who are working for storage device are joining the industry of NAND flash memory. This book on NAND flash memory technologies was written to provide detailed views of NAND flash technologies for such individuals who are not only engineers of developing NAND flash memory, but are also NAND flash users, product engineers, application engineers, marketings, managers, technical managers, engineers for developing and producing SSD, engineers of other NAND flash-related storage devices such as data servers, and so on.

This book is also suitable for new engineers and graduate students to quickly study and to be familiar with NAND flash memory technologies. I expect this book to encourage newcomers to contribute to future NAND flash memory technologies and products.

The contents of this book include the starting history, memory cell technologies, basic structure and physics, principles of operations, history and trend of memory cell scaling, advanced operations for multilevel cells (2, 3, 4 bits/cell), scaling

challenges, reliability, 3D NAND flash memory cell, scaling challenge of 3D NAND flash memory cell, and future prospects of NAND flash memory.

After describing the background, the starting history of NAND flash is introduced in Chapter 1. The basic device structures and operations are described in Chapter 2.

Chapter 3 discusses the memory cell technologies focused on scaling. To scale down memory cell size, memory cell structure has been evolved from LOCOS isolation cells to self-aligned STI cells, along with reducing the feature size (design rule).

Chapter 4 introduces the advanced operations for multilevel cells. Tight V_t distribution width is very important for multilevel cells because of enough read window margin. Advanced operations have been mainly developed for this point.

By scaling down memory cell size below 20 nm, several physical limitation phenomena are exaggerated. Chapter 5 discusses the details of physical limitations for scaling. The floating-gate capacitive coupling interference has the worst impact on scaling, even when advanced operations are used, as shown in Chapter 4. And other physical limitation factors are also discussed, including an electron injection spread, RTN, structure limitations, high field problems, and so on.

Chapter 6 describes the reliability of NAND flash memory. A program/erase cycling degrades tunnel oxide quality by generating electron/hole traps and stress-induced leakage current (SILC). Thus all reliability aspects of the cycling endurance, data retention, read disturb, program disturb, and erratic over-programming are degraded by increasing the amount of cycling. In Chapter 6, the mechanism and impact on device reliabilities are discussed.

Chapter 7 shows three-dimensional (3D) NAND flash memory cells. Many types of 3D cells have been proposed. These 3D cells are introduced, and pros and cons in structure, process, operations, scalability, performance, and so on, are discussed.

The mass production of 3D NAND flash memory was started in 2013. Full-scale production will begin in 2016. However, for future 3D NAND flash memory, many problems still remain and have to be solved. In Chapter 8, challenges of 3D NAND flash memory are discussed. Increasing the number of stacked cells is essential for reducing the effective cell size in a 3D cell. High aspect ratio process and small cell current issues will be of utmost importance, as discussed in this chapter. I tried to show some possible solutions for these problems. Other challenges, such as new program disturbance issues, data retention, power consumption, and so on, are discussed.

In Chapter 9, I summarize and describe the prospect of technologies and market for the future of NAND flash memory.

I am convinced that this book is a significant contribution to the industry of NAND flash memory and related products. I sincerely hope you find this book useful in your future work.

SEIICHI ARITOME

Kawasaki, Japan

ACKNOWLEDGMENTS

The author would like to express special thanks to Professor Fujio Masuoka. The author is proud of their collaboration on NAND flash memory, since Professor Masuoka assigned the author to the development of NAND flash memory in VLSI Research Center, Toshiba Corporation, in the early stage of development in 1988.

The author would like to thank Mr. Kiyoshi Kobayashi, Mr. Shinichi Tanaka, Mr. Masaki Momodomi, Professor Riichiro Shirota, Professor Shigeyoshi Watanabe, Dr. Koji, Sakui, Professor Fumio Horiguchi, Mr. Kazunori Ohuchi, Dr. Junichi Matsunaga, Dr. Akimichi Hojo, and Dr. Hisakazu Iizuka for their continuous encouragement ever since I joined the Research & Development Center at the Toshiba Corporation.

The author is grateful to colleagues at the Toshiba Corporation for their contributions. Without their help, this work could not have been successful. I especially thank Mr. Ryouhei Kirisawa, Dr. Kazuhiro Shimizu, Mr. Yuji Takeuchi, Mr. Hiroshi Watanabe, Dr. Gertjan Hemink, Mr. Shinji Sato, Dr. Tooru Maruyama, Mr. Kazuo Hatakeyama, Professor Hiroshi Watanabe, Professor Ken Takeuchi, and Mr. Tomoharu Tanaka. I appreciate Mr. Ryozo Nakayama, Mr. Akira Goda, Mr. K. Narita, Mr. E. Kamiya, Mr. T. Yaegashi, Ms. K. Amemiya, Mr. Toshiharu Watanabe, Dr. Fumitaka Arai, Dr. Tetsuya Yamaguchi, Ms. Hideko Oodaira, Dr. Tetsuo Endoh, Mr. Susumu Shuto, Mr. Hirohisa Iizuka, Mr. Hiroshi Nakamura, Dr. Toru Tanzawa, Dr. Yasuo Itoh, Mr. Yoshihisa Iwata, Mr. Kenichi Imamiya, Mr. Kazunori Kanebako, Mr. Kazuhisa Kanazawa, Mr. Hiroto Nakai, Mr. Takehiro Hasegawa, Dr. Katsuhiko Hieda, Dr. Akihiro Nitayama, Mr. Koichi Fukuda, and Mr. Seiichi Mori for their fruitful discussion.

The author would like to thank Mr. Eli Harari, Mr. Sanjay Mehrotra, Dr. George Samachisa, Dr. Jian Chen, Mr. Tuan D. Pham, Mr. Ken Oowada, Dr. Hao Fang, and

Dr. Khandker Quader for their continuous encouragement and fruitful discussions since SanDisk-Toshiba joint development started in 1999.

The author would like to thank Dr. Kirk Prall, the late Mr. Andrei Mihnea, Mr. Frankie Roohparvar, Dr. Luan Tran, and Mr. Mark Durcan for their continuous encouragement since I joined Micron Technology, Boise, Idaho, USA, in 2003.

The author would like to thank Mr. Krishna Parat, Dr. Pranav Kalavade, Dr. Mark Bauer, Dr. Nile Mielke, and Dr. Stefan K. Lai for their continuous encouragement and fruitful discussions since Intel-Micron joint development started.

The author would like to thank Dr. T.-J. Brian Shieh, Dr. Alex Wang, Dr. Travis C.-C.Cho, Ms. Saysamone Pittikoun, Mr. Yoshikazu Miyawaki, Mr. Hideki Arakawa, and Mr. Stephen C. K. Chen for their continuous encouragement and support since I joined Powerchip Semiconductor Corp, Hsinchu, Taiwan.

The author would like to thank Dr. Sungwook Park, Dr. Sungjoo Hong, Dr. Seok Hee Lee, Dr. Seokkiu Lee, Dr. Seaung Suk Lee, Dr. Sungkye Park, Mr. Gyuseog Cho, Mr. Jongmoo Choi, Mr. Yoohyun Noh, Mr. Hyunseung Yoo, Dr. EunSeok Choi, Mr. HanSoo Joo, Mr. Youngsoo Ahn, Mr. Byeongil Han, Mr. Sungjae Chung, Mr. Keonsoo Shim, Mr. Keunwoo Lee, Mr. Sanghyon kwak, Mr. Sungchul Shin, Mr. Iksoo Choi, Mr. Sanghyuk Nam, Mr. Dongsun Sheen, Mr. Seungho Pyi, Mr. Jinwoong Kim, Mr. KiHong Lee, Mr. DaeGyu Shin, Mr. BeomYong Kim, Mr. MinSoo Kim, Mr. JinHo Bin, Mr. JiHye Han, Mr. SungJun Kim, Mr. BoMi Lee, Mr. YoungKyun Jung, Mr. SungYoon Cho, Mr. ChangHee Shin, Mr. HyunSeung Yoo, Mr. SangMoo Choi, Mr. Kwon Hong, Mr. SungKi Park, Ms. Soonok Seo, and Mr. Hyungseok Kim for their warm consideration ever since I joined the Research & Development division at the SK Hynix Inc.

The author would like to thank Mr. Angelo Visconti, Ms. Silvia Beltrami, Ms. Gabriella Ghidini, Dr. Emilio Camerlenghi, Mr. Roberto Bez, Mr. Giuseppe Crisenza, and Mr. Paolo Cappelletti for their continuous encouragement and fruitful discussion while we performed Numonyx–Hynix joint development.

The author is profoundly grateful to Professor Masataka Hirose, Professor Mizuho Morita, Professor Seiichi Miyazaki, and Professor Yukio Osaka for their warm consideration and encouragement since I joined the laboratory of Professor Masataka Hirose in Hiroshima University on 1982.

The author would like to thank Professor Takamaro Kikkawa, Professor Shin Yokoyama, and Professor Seiichiro Higashi, of Hiroshima University for their invaluable guidance and continuous encouragement.

Finally the author would like to express his heartfelt thanks to his wife, Miho Aritome, and his son, Santa Aritome, who have continuously supported him with their love.

SEIICHI ARITOME

Kawasaki, Japan

ABOUT THE AUTHOR

Seiichi Aritome received his B.E., M.E., and Ph.D degrees in electronic engineering from Hiroshima University, Japan, in 1983, 1985, and 2013, respectively.

He joined the Toshiba Research and Development Center, Kawasaki, Japan, in 1985. Since then, he has been engaged in the development of high-density DRAM. In 1988, he joined the EEPROM development group at the same research center. At that time, the EEPROM development group started to develop NAND-type flash memory for the first time in the world. His major work is the NAND flash memory device technology, process integration, characterization, and reliability.

He has contributed to NAND flash memory technologies over 25 years (1988 to the present) in several companies and nations. He has developed over 12 generations of NAND flash memories. Many technologies which he developed had become a standard of NAND flash memories.

He had contributed to the crucial decision of the proper uniform program/erase scheme for NAND flash memory by analyzing phenomena of program/erase cycling degradation. He clarified that the uniform program/erase scheme was appropriate reliability in comparison with other schemes. As a result, the uniform program/erase scheme was decided upon for NAND flash operation. The uniform program/erase scheme had another important advantage, namely, fast programming speed (~100 Mbyte/s) due to low power consumption during the program in the uniform program/erase scheme. Because of high reliability and fast programming, the uniform program/erase scheme became the de facto standard technology. All NAND suppliers (Toshiba, Samsung, Micron/Intel, SK Hynix, etc.) have used the uniform program/erase scheme for all NAND flash products over the past 20 years.

He proposed and developed the self-aligned shallow trench isolation cell (SA-STI cell) for the first time. The cell size of NAND Flash memory could be drastically shrunk to 66% (from $6F^2$ to $4F^2$; F stands for feature size), in comparison with that

of a conventional LOCOS cell. This technology could realize a 256-Mbit NAND flash for the first time in the world. Also, the SA-STI cell has an excellent reliability because of no STI corner in tunnel oxide. Therefore, the SA-STI cell has been widely used in the NAND flash product over the past 17 years. All NAND flash memory suppliers are using this cell technology.

Many of the NAND flash technologies he developed became de facto standard because of low cost, high reliability, and fast programming speed. Therefore, the NAND flash enabled us to launch the new market of a smartphone, tablet PC, and SSD (solid-state drive) and have had a large volume production, estimated to be $35 billion in 2015.

In 1998, he moved to the flash device engineering group of memory division at the Toshiba semiconductor company. While working for Toshiba (1988–2003), he had engaged many generations of NAND flash (0.7 μm, 0.4 μm, 0.2 μm, 0.16 μm, 0.12 μm, 90 nm, 70 nm). Also, he was the first technical coordinator of the Toshiba–SanDisk joint development of NAND flash memory in 1999.

He started to work at Micron Technology Inc., Boise, Idaho, USA, in December 2003. He engaged the NAND flash process and device development over several generations of 90 nm, 72 nm, 50 nm, and 34 nm. He was transferred to Powerchip Semiconductor Corporation, Hsinchu, Taiwan, on April 2007. He worked for the development of NAND flash of 70-nm and 42-nm generations as program manager. And then he moved to SK Hynix Inc., Icheon, Korea, on April 2009. He worked for NAND flash development of 26-nm, 20-nm, and mid-1X-nm generations and three-dimensional cell.

He holds 251 US patents and 76 Japanese patents on NAND flash memories, such as cell structure, process, operation scheme, and high reliability device technologies. He has authored or co-authored over 50 papers. Most of them are for NAND flash memory technologies.

He is an IEEE Fellow and a member of the IEEE Electron Device Society.

1

INTRODUCTION

1.1 BACKGROUND

Recent progress in computers and mobile equipment requires further efforts in developing higher-density nonvolatile semiconductor memories. A breakthrough in the field of nonvolatile memories was the invention of the flash memory [1], which is a new type of EEPROM (electrically erasable and programmable read-only memory), as shown in Fig. 1.1a. The first paper discussing the flash memory was presented in 1984 IEDM (International Electron Device Meeting). The flash memory has many advantages in comparison with other nonvolatile memories. Therefore, the flash memory explosively accelerated the development of higher-density EEPROMs.

In 1987, a NAND structured cell was proposed by Masuoka et al. [2]. This structure can reduce the memory cell size without scaling of device dimension. The NAND structure cell arranges a number of bits in series, as shown in Fig. 1.1b [2]. The conventional EPROM cell has one contact area per two bits. However, for a NAND structure cell, only one contact hole is required per two NAND structure cells (NAND string). As a result, the NAND cell can realize a smaller cell area per bit than the conventional EPROM.

Applications of flash memory became quite wide due to nonvolatility, fast access, and robustness. Flash memory application can be classified into two major markets (Fig. 1.1). One is for code storage applications, such as PC BIOS, cellular phones, and DVDs. The NOR-type cell is best suitable for this market due to its fast random access speed. The other is for file storage applications, such as the digital still camera

Nand Flash Memory Technologies, First Edition. Seiichi Aritome.

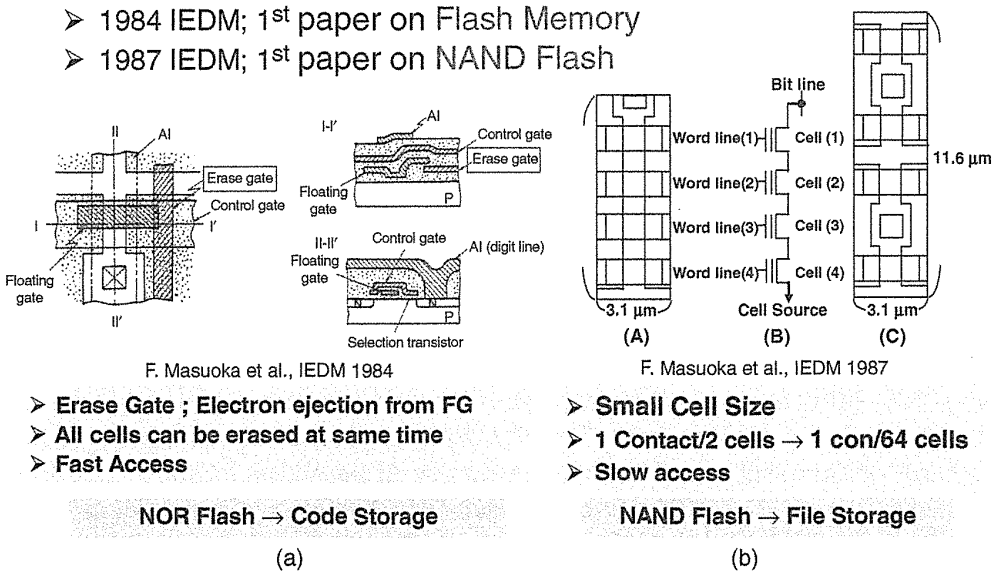


FIGURE 1.1 Invention of flash memory and NAND flash memory. (a) Flash memory. All cells in the memory chip can be erased at the same time by applying erase voltage to the erase gate [1]. (b) NAND flash memory [2]. Memory cells are connected in series to share contact area. Comparison between (A) NAND cell and (C) conventional EPROM (NOR flash cell). (B) shows the equivalent circuit of the NAND structure cell having 4 cells.

(DSC), silicon audio, the smartphone, and the tablet PC. The NAND-type cell is suitable for file storage market.

Figure 1.2 shows the memory hierarchy of computer system before mass production of NAND Flash. SRAM and DRAM had been used as cash memory and main memory, respectively. And magnetic memories, such as HDD, had been used as a nonvolatile mass-storage device. NAND flash memory had been targeted to replace magnetic memory [54]. Actually, from the production start of NAND flash memory in 1992, the NAND flash memory has been widely applied to new emerging applications and has replaced magnetic memory, as shown in Fig. 1.3. At first, a photo film had been completely replaced by the memory cards of NAND flash memory. Next, the floppy disk was replaced by USB drive memory. The mobile music equipment with cassette tape was replaced by the MP3 player using flash memory storage. Also, NAND flash memory had created new market of smartphones and tablet PCs. And now, the application is extending to the SSD (solid-state drive) market, not only for the consumer but also for the enterprise server. Therefore, over 20 years, NAND flash memory has created new large-volume markets and industries of consumer, computer, mass storage, and enterprise server. NAND flash production volume was tremendously increased. The overall NAND market is expected to reach $40 billion in 2016 [55]. NAND flash has become an explosive innovation and has greatly contributed to the improvement of our lives with the advent of convenient mobile equipment such as smartphones and tablet PCs.

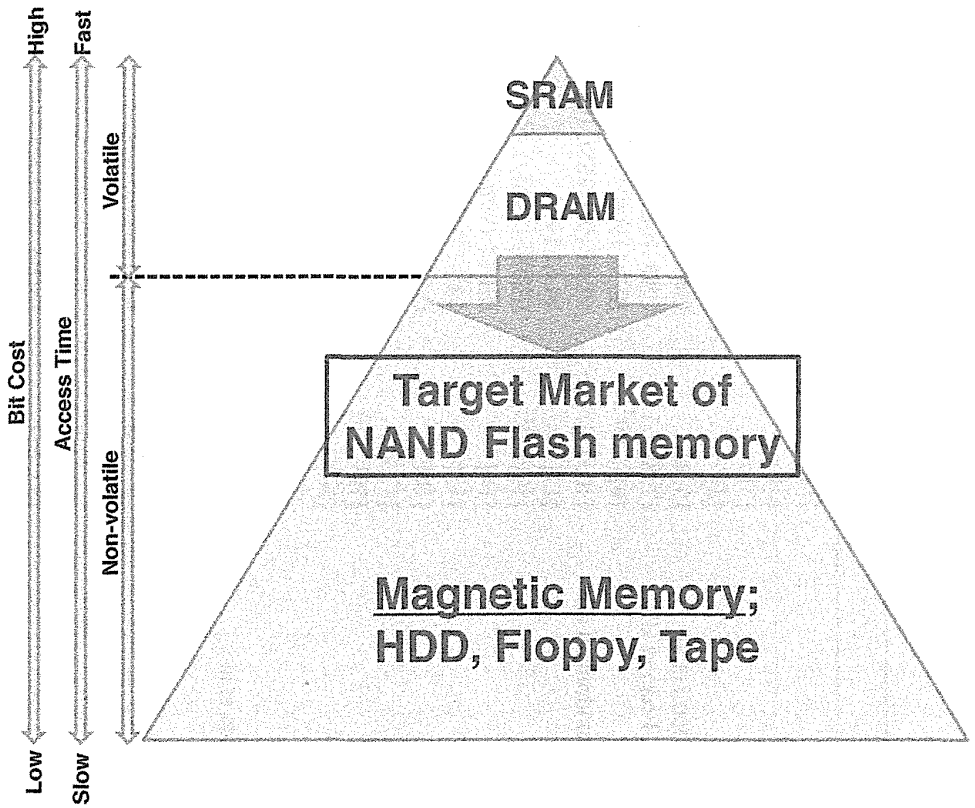


FIGURE 1.2 Target market of NAND flash memory.

Table 1.1 shows the history of NAND flash memory development, based on technical papers from 1987 to 1997. During the 10 years from the first NAND flash paper in 1987, all of the fundamental and important NAND flash technologies were established, such as page programming [7, 8], block erase, the uniform program and uniform well erase scheme [9, 12, 13], bit-by-bit verify [15, 21], the ISPP (incremental step pulse program) [25, 26, 29], the self-aligned STI cell [22, 51, 56], the shield

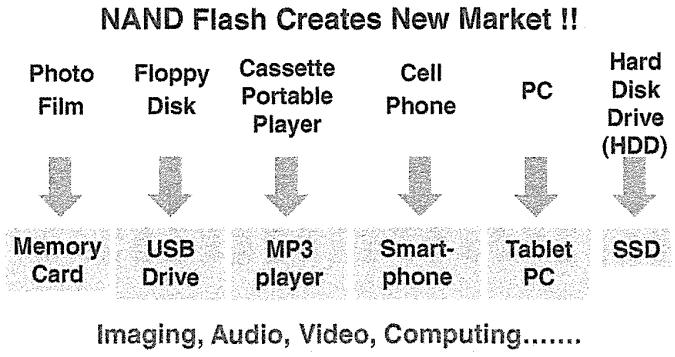


FIGURE 1.3 NAND flash memory creates a new market.

TABLE 1.1 History of the NAND Flash Memory (~1997)

Year		Authors	References	Conference/Journal
1984	Flash Memory, first paper	F. Masuoka et al.	[1]	IEDM 1984
1987	NAND-type flash memory, 1st paper	F. Masuoka et al.	[2]	IEDM 1987
1988	NAND-type flash memory	R. Shirota et al.	[3]	VLSI 1988
	Drain-FN program	M. Momodomi et al.	[4]	IEDM 1988
1989	4-Mb NAND-type flash memory	Y. Itoh et al./M. Momodomi et al.	[5, 6]	ISSCC1989/ JSSC
	Page program	M. Momodomi et al./Y. Iwata et al.	[7, 8]	CICC1989/ JSSC
1990	Well erase reliability	S. Aritome, et al	[9]	IRPS 1990
	4-Mbit tight V_t distribution	T. Tanaka et al./M. Momodomi et al.	[10, 11]	VLSI 1990/JSSC1991
	Well erase	R. Kirisawa et al.	[12]	VLSI 1990
	Bipolarity program/erase	S. Aritome et al.	[13]	IEDM 1990
	Double patterning	R. Shirota et al.	[14]	IEDM 1990
1992	Bit-by-bit verify	T. Tanaka et al.	[15]	VLSI 1992
1993	Reliability of flash	S. Aritome et al.	[16]	Proceedings of IEEE
	0.4-μm 64-Mb-cell technology	S. Aritome et al.	[17,18]	SSDM 93/ JJAP
1994	SILC	H. Watanabe et al.	[19]	VLSI 1994
	Cycling and data retention reliability	S. Aritome et al.	[20]	IEICE
	Intelligent program, shield BL scheme	T. Tanaka et al.	[21]	JSSC
	Self-aligned STI cell (SA-STI cell)	S. Aritome et al.	[22]	IEDM 1994
1995	32-Mb NAND	K Imamiya et al./Y. Iwata et al.	[23, 24]	ISSCC1995/ JSSC
	32-Mb NAND with increment step program pulse (ISPP), self-boost	K. D.Suh et al.	[25, 26]	ISSCC1995/ JSSC
	Read disturb, SILC	S. Satoh, et al	[27, 28]	ICMTS1995/ED
	Increment step program pulse (ISPP)	G. J. Hemink et al.	[29]	VLSI 1995
	Double V_t select gate	K Takeuchi et al.	[30, 31]	VLSI 1995/ JSSC
	SWATT cell	S. Aritome et al.	[32, 33]	IEDM 1995/ ED

1996	128-Mb MLC	T. S. Jung et al.	[34, 35]	ISSCC 1996/ JSSC
	64 Mb	J. K. Kim et al.	[36, 37]	VLSI1996/ JSSC
	SILC	G. J. Hemink et al.	[38]	VLSI 1996
	On-chip ECC	T. Tanzawa et al.	[39, 40]	VLSI1996/ JSSC
	Booster plate	J. D. Choi et al.	[41]	VLSI 1996
	High- speed NAND	D. J. Kim et al.	[42]	VLSI 1996
	SILC in STI	H. Watanabe et al.	[43]	IEDM 1996
	Shared bit line	W. C. Shin et al.	[44]	IEDM 1996
1997	Nonvolatile virtual DRAM using NAND	T. S. Jung et al.	[45, 46]	ISSCC 1997/ JSSC
	Three-level cell (1.5 bits/cell)	T. Tanaka et al,	[47]	VLSI 1997
	Multi-page cell	K Takeuchi et al.	[48, 49]	VLSI 1997/ JSSC
	Parallel program	H. S. Kim et al.	[50]	VLSI 1997
	0.25 μm SA-STI cell	K. Shimizu et al.	[51]	IEDM 1997
	Program disturb	S. Satoh et al.	[52]	IEDM 1997
	Triple poly-booster gate	J. D. Choi et al.	[53]	IEDM 1997

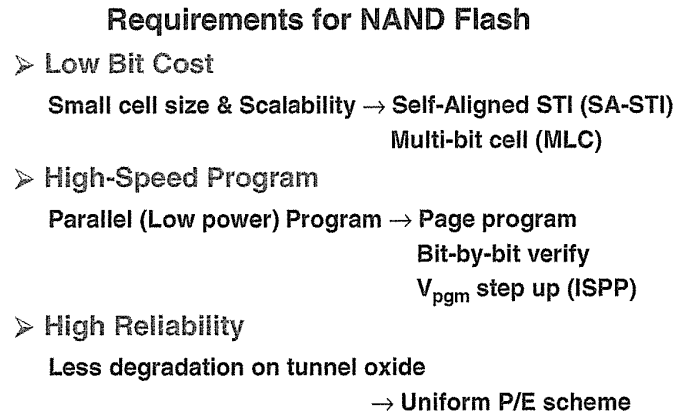


FIGURE 1.4 Requirements for NAND flash memory of the file storage market.

bit-line scheme [21], and so on. These technologies could satisfy the requirements of file storage memory.

Requirements for file storage memory are low bit cost, high-speed programming, and high reliability, as shown in Fig 1.4 [56].

The most important requirement for file storage applications is the low bit cost. The cost of a memory device is mainly determined by the die size of the memory chip and by the fabrication process cost, which is mainly dependent on depreciation of investment on factory. Then it is very important to combine small die size with a simple and low-cost fabrication process. In order to reduce the die size, reduction of unit memory cell size is as important as scaling feature size. Ideal memory cell size is $4 * F^2$ (F stands for feature size), because both *X* and *Y* directions are determined by line (F) and space (F). However, in early 1990s, it was difficult to realize $4 * F^2$ cell size of NAND flash memory due to wide ($>2 * F$) isolation width of LOCOS (local oxidation of Si). The self-aligned shallow trench isolation cell (SA-STI cell) was proposed and implemented to the NAND flash memory product. An isolation width could be scaled down from 2–3F in the LOCOS cell to F in the SA-STI cell. Therefore, the cell size could be drastically scaled down.

The SA-STI cell has been used in mass production for a long time, from 1998 to the present, because of a lot of advantages, such as small cell size, high reliability, and excellent scalability. However, below the 20-nm feature size, it is becoming very difficult to manage physical limitations, such as the floating gate capacitive coupling effect, RTN (random telegraph noise), the high-field problem, and so on. The recent feature size for production could reach to 15–16 nm [57]. It is not still clear whether memory cell size can be scaled down further or not.

Another way to reduce the effective cell size is the "multilevel cell." The logical bits are stored in one physical memory cell; for example, 2 logical bits are stored in one physical memory cell (MLC; 2 bits/cell). And 3 bits/cell and 4 bits/cell are called TLC and QLC. The mass-production start of the MLC was in 2000 by using 0.16-μm technology. The process technology to fabricate a multilevel cell device is basically as same as the process of single bit cell (SLC); however, the operations for

a multilevel cell are much different from SLC operation. It is very important to make a tight V_t distribution width in the multilevel cell, in order to have high performance and reliability.

The next requirement for file storage application is high-speed programming, as shown in Fig. 1.4. In NAND flash memory, the uniform program/erase (P/E) scheme has been used as a de facto standard over 20 years. Unlike a NOR flash, no huge hot-electron injection current is required for programming, but a uniform P/E scheme has produced very low power consumption for programs even when the number of memory cells to be programmed is increased. Therefore the NAND flash memory can be easily programmed in large pages (512-byte to 32-Kbyte cells) so that the programming speed per byte can be quite fast ($\sim$ 100 Mbytes/s). In addition, several advanced program operations, such as bit-by-bit verify, V_{pgm} step up (ISPP: incremental step pulse program), ABL (all bit line) architecture, and so on, had been developed for high-speed programming.

The other important requirement for file storage applications is "high reliability," as shown in Fig. 1.4. A high voltage (>20 V) is applied to a control gate to produce a Fowler–Nordheim (FN) tunneling current on the tunnel oxide during programming. The electric field in tunnel oxide reaches values greater than 10 MV/cm, which is normally caused by oxide breakdown in other semiconductor devices. This means that flash memory uses a breakdown-like operation in normal program and erase. Due to applying a high field, tunnel oxide has been degraded by an electron/hole trap, interface state generation, and stress-induced leakage current (SILC). Major reliability degradation aspects of flash memory are related to this tunnel oxide degradation by programming and erase cycling. Even if a tunnel oxide is degraded, stored data have to be sustained in memory cells for long time, as nonvolatile memory. Data retention time after programming and erase cycling is a key of NAND flash reliability.

In addition, read disturb and program disturb are also an important reliability phenomena in NAND flash [13,16]. During read and program operation, pass voltages are applied to unselected word lines (WLs) in the NAND string. Several kinds of disturb stress are applied to an unselected cell in a cell array. Read disturb and program disturb are caused in these unselected cells in a string in a cell array.

Reliability specifications for NAND flash memory are dependent on applications such as digital still cameras, MP3 players, SSDs (solid-state disks) for PCs, SSDs for data servers, and so on. Target specifications of a NAND flash are generally as follows. In order to guarantee the specifications of NAND, every effort has been made regarding devices, processes, operations, circuits, memory systems, and so on.

Program and erase cycles (P/E cycles): 1-K to 100-K cycles

Data retention: 1–10 years

Read cycles: 1E5 – 1E7 times

Number of page program time (NOP): 1 time for MLC,TLC,QLC, 2–8 times for SLC

In 2007, three-dimensional (3D) NAND flash device technology of BiCS (bit cost scalable) was proposed [58] in order to scale down the NAND flash memory cell

further. BiCS technology has a new low-cost process concept. The vertical poly-Si channel is fabricated by through-holes in stacked multilayer word lines. After the BiCS proposal, several three-dimensional (3D) NAND flash cells have been proposed [59–63]. Due to the vertical stacked cell structure, the 3D cell has an advantage of reducing effective cell size without scaling the feature size of F. In 2013, the mass-production start of 3D NAND flash was announced. The device was a 128-Gbit MLC 3D V-NAND flash with a 24-cell stacked charge trap cell [64]. To proceed to a lower bit cost of the 3D NAND cell, a number of stacked cells are needed to increase intensively. Many technical issues, such as a high-aspect etching, data retention of a charge trap cell, a new program disturb mode, cell current fluctuation, and so on, have to be solved or managed. After overcoming these critical issues, it is expected that a 1-terabit or 2-terabit NAND flash memory device will be available around 2020.

1.2 OVERVIEW

The NAND flash memory device technologies are reviewed in this book. The chapters focus on the scaling of the NAND flash memory cell, the high-performance operation of NAND flash, the improvement of NAND flash reliability, and three-dimensional (3D) NAND flash technologies, because they are very important for present and future NAND flash memory.

After describing a background of NAND flash technology in Chapter 1, Chapter 2 presents a basic structure and operations of NAND flash memory. The structures of single-cell and NAND-cell array are described. Cell operations of read, program, and erase are introduced. And then multilevel NAND cell technology is discussed to realize low-cost NAND flash memory.

The scaling history and scenario of planar (two-dimensional) NAND flash memory cells are reviewed in Chapter 3. The layout of the NAND flash memory cell is simple: Parallel word lines (WL) are perpendicular to parallel bit lines (BL). WL pitch is normally $2 * F$, (F: feature size), which is limited by lithography technology. However, BL pitch was normally $3 * F$ or more in the case of LOCOS isolation. This is because the isolation width needed to be $2 * F$ or more to prevent a relatively high (~8 V) punch-through between NAND cell channels (strings) during programming. Thus, it was crucial to scale down isolation width, in order to scale down memory cell size to satisfy the requirement of low bit cost.

First, LOCOS isolation cell technologies are presented (Section 3.2). The LOCOS isolation width can be minimized with improving device performance by the field-through implantation technique (FTI) after LOCOS formation. Next, the self-aligned STI cell with the FG (floating gate) wing is discussed (Section 3.3). The FG wing is applied to reduce the aspect ratio of cell structure. And then, the self-aligned STI cell without FG wing is discussed (Section 3.4). This cell has been used from the 90-nm generation cell to the present cell (1Y-nm cell), as a defacto standard. And the planar FG cell is introduced as an alternate cell structure (Section 3.5). Also, the sidewall transfer transistor cell (SWATT cell) is described (Section 3.6). Due to sidewall transfer transistor, the V_t read window margin can be greatly improved. Then,

fast programming speed can be expected. And then, recent advanced NAND flash memory cell technologies of the dummy word-line scheme and the *p*-type floating gate are discussed in Section 3.7.

Another important technology for low bit cost is the multilevel cell (MLC), which is a stored multilogical bit in a single memory cell. To implement MLC, smart operation schemes are crucial to produce reasonable performance and reliability. In Chapter 4, the advanced operations for a multilevel NAND flash are discussed. It is very important to make tight V_t distribution width during programming for better performance and reliability. Most of the operation schemes focus on this point.

For the scaling memory cell, it is becoming very difficult to control the V_t distribution width due to the occurrence of several physical limitations, including the floating-gate capacitive coupling effect, electron injection spread, RTN, and the high-field problem. These physical limitations make the V_t distribution width wider, and then the, cell V_t setting margin (read window margin) is degraded. The recent feature size for production could reach values below 20 nm. The read window margin is seriously degraded. Then it is important to clarify how much scaling limitation factors have an impact on the V_t margin beyond the 20-nm feature size. Thus, the scaling challenges of the self-aligned STI cell are discussed beyond 20 nm in Chapter 5.

The reliability of two-dimensional NAND flash memory cell is discussed in Chapter 6. Reliability of flash memory is attributed to data retention or read disturb after program/erase cycling endurance. Program and erase operation schemes have a serious impact on reliability of a flash memory cell. Then, many program/erase schemes were proposed to satisfy the requirement of reliability and performance. It is very important to clarify the cell degradation mechanism and the best scheme of program/erase in order to achieve the requirement of reliability and performance. Chapter 6 describes the reliability aspect of NAND flash cell. The uniform program/erase scheme has several advantages in NAND flash reliability by comparing program/erase endurance, data retention, and read disturb characteristics in several program and erase schemes.

The three-dimentional (3D) NAND flash cells are presented in Chapter 7. After describing motivation and history of 3D NAND flash, many types of three-dimensional cells are introduced. Advantages and performances are compared in several 3D cells, including BiCS cell, TCAT/V-NAND, SMArT cell, VG-NAND, and DC-SF cell.

After that, the challenges of 3D NAND cells are discussed in Chapter 8. To realize low-cost NAND flash memories, serious issues have to be solved or managed by improving process, structure, device, performance, and reliability

The future trend of NAND flash memory is discussed in Chapter 9. The perspectives on future NAND flash technologies are also discussed.

Corresponding to the above discussions, the following topics are described in this book:

1. Principle of NAND flash memory.
2. Scaling scenario of 2D NAND flash memory cell.

3. Practical framework of scaling down of 2D NAND flash memory cell.
4. The LOCOS isolation technology to scale down NAND flash memory cell.
5. The self-aligned STI technology.
6. Low-cost NAND flash process flow.
7. The planar FG cell.
8. The SWATT cell for MLC (multilevel cell).
9. Advanced operations for MLC.
10. Basic and advanced program operations for tight programmed V_t distribution width in MLC.
11. Page program sequence for MLC (2 bits/cell).
12. Page program sequence for TLC (3 bits/cell)
13. The scaling challenges of a 2D NAND flash cell.
14. The solutions to overcome the scaling limitation of a 2D NAND flash cell.
15. The factors analysis of physical scaling limitation of a 2D NAND flash cell.
16. Detail mechanism of the floating gate capacitive coupling interference, Electron injection spread, RTN, and so on., as scaling limiter.
17. Investigation on the reliability of NAND flash in several program and erase schemes, in order to clarify the dependence of program and erase scheme.
18. Investigation of the program disturb and read disturb phenomena to optimize operation of NAND flash cell.
19. Introduction of several three-dimensional (3D) NAND flash memory cells.
20. Scaling challenges of 3D NAND flash memory.
21. Detail mechanism of program disturb, cell current fluctuation, and so on., in 3D NAND flash cell.
22. Future trend of NAND flash memory technologies.

REFERENCES

[1] Masuoka, F.; Asano, M.; Iwahashi, H.; Komuro, T.; Tanaka, S. A new flash E^2PROM cell using triple polysilicon technology, *Electron Devices Meeting, 1984 International* vol. 30, pp. 464–467, 1984.

[2] Masuoka, F.; Momodomi, M.; Iwata, Y.; Shirota, R. New ultra high density EPROM and flash EEPROM with NAND structure cell, *Electron Devices Meeting, 1987 International* vol. 33, pp. 552–555, 1987.

[3] Shirota, R.; Itoh, Y.; Nakayama, R.; Momodomi, M.; Inoue, S.; Kirisawa, R.; Iwata Y.; Chiba, M.; Masuoka, F. New NAND cell for ultra high density 5v-only EEPROMs *Digest of Technical Papers—Symposium on VLSI Technology*, 1988, pp. 33–34.

[4] Momodomi, M.; Kirisawa, R.; Nakayama, R.; Aritome, S.; Endoh, T.; Itoh, Y.; Iwata, Y. Oodaira, H.; Tanaka, T.; Chiba, M.; Shirota, R.; Masuoka, F. New device technologies

for 5 V-only 4 Mb EEPROM with NAND structure cell, *Electron Devices Meeting, 1988. IEDM'88. Technical Digest International*, pp. 412–415, 1988.

[5] Itoh, Y.; Momodomi, M.; Shirota, R.; Iwata, Y.; Nakayama, R.; Kirisawa, R.; Tanaka, T.; Toita, K.; Inoue, S.; Masuoka, F. An experimental 4 Mb CMOS EEPROM with a NAND structured cell, *Solid-State Circuits Conference, 1989. Digest of Technical Papers, 36th ISSCC, 1989 IEEE International*, pp. 134–135, 15–17 Feb. 1989.

[6] Momodomi, M.; Itoh, Y.; Shirota, R.; Iwata, Y.; Nakayama, R.; Kirisawa, R.; Tanaka, T.; Aritome, S.; Endoh, T.; Ohuchi, K.; Masuoka, F. An experimental 4-Mbit CMOS EEPROM with a NAND-structured cell, *Solid-State Circuits, IEEE Journal of*, vol. 24, no. 5, pp. 1238–1243, Oct. 1989.

[7] Momodomi, M.; Iwata, Y.; Tanaka, T.; Itoh, Y.; Shirota, R.; Masuoka, F. A high density NAND EEPROM with block-page programming for microcomputer applications, *Custom Integrated Circuits Conference, 1989, Proceedings of the IEEE 1989*, pp. 10.1/1–10.1/4, 15–18 May 1989.

[8] Iwata, Y.; Momodomi, M.; Tanaka, T.; Oodaira, H.; Itoh, Y.; Nakayama, R.; Kirisawa, R.; Aritome, S.; Endoh, T.; Shirota, R.; Ohuchi, K.; Masuoka, F. A high-density NAND EEPROM with block-page programming for microcomputer applications, *Solid-State Circuits, IEEE Journal of*, vol. 25, no. 2, pp. 417–424, Apr. 1990.

[9] Aritome, S.; Kirisawa, R.; Endoh, T.; Nakayama, R.; Shirota, R.; Sakui, K.; Ohuchi, K.; Masuoka, F. Extended data retention characteristics after more than 10^4 write and erase cycles in EEPROMs, *International Reliability Physics Symposium, 1990. 28th Annual Proceedings*, pages 259–264, 1990.

[10] Tanaka, T.; Momodomi, M.; Iwata, Y.; Tanaka, Y.; Oodaira, H.; Itoh, Y.; Shirota, R.; Ohuchi, K.; Masuoka, F. A 4-Mbit NAND-EEPROM with tight programmed V_t distribution, *VLSI Circuits, 1990. Digest of Technical Papers, 1990 Symposium on*, pp. 105–106, 7–9 June 1990.

[11] Momodomi, M.; Tanaka, T.; Iwata, Y.; Tanaka, Y.; Oodaira, H.; Itoh, Y.; Shirota, R.; Ohuchi, K.; Masuoka, F. A 4 Mb NAND EEPROM with tight programmed V_t distribution, *Solid-State Circuits, IEEE Journal of*, vol. 26, no. 4, pp. 492–496, Apr. 1991.

[12] Kirisawa, R.; Aritome, S.; Nakayama, R.; Endoh, T.; Shirota, R.; Masuoka, F. A NAND structured cell with a new programming technology for highly reliable 5 V-only flash EEPROM, *1990 Symposium on VLSI Technology, 1990. Digest of Technical Papers*, pages 129–130, 1990.

[13] Aritome, S.; Shirota, R.; Kirisawa, R.; Endoh, T.; Nakayama, R.; Sakui, K.; Masuoka, F. A reliable bi-polarity write/erase technology in flash EEPROMs, *International Electron Devices Meeting, 1990. IEDM'90. Technical Digest*, pages 111–114, 1990.

[14] Shirota, R.; Nakayama, R.; Kirisawa, R.; Momodomi, M.; Sakui, K.; Itoh, Y.; Aritome, S.; Endoh, T.; Hatori, F.; Masuoka, F. A 2.3 μm^2 memory cell structure for 16 Mb NAND EEPROMs, *Electron Devices Meeting, 1990. IEDM'90. Technical Digest, International*, pp. 103–106, 9–12 Dec. 1990.

[15] Tanaka, T.; Tanaka, Y.; Nakamura, H.; Oodaira, H.; Aritome, S.; Shirota, R.; Masuoka, F. A quick intelligent program architecture for 3 V-only NAND-EEPROMs, *VLSI Circuits, 1992. Digest of Technical Papers, 1992 Symposium on*, pp. 20–21, 4–6 June 1992.

[16] Aritome, S.; Shirota, R.; Hemink, G.; Endoh, T.; Masuoka, F.; Reliability issues of flash memory cells, *Proceedings of the IEEE*, vol. 81, no. 5, pages 776–788, 1993.

[17] Aritome, S.; Hatakeyama, I.; Endoh, T.; Yamaguchi, T.; Shuto, S.; Iizuka, H.; Maruyama, T.; Watanabe, H.; Hemink, G. H.; Tanaka, T.; M. Momodomi, K. Sakui, and R. Shirota, A 1.13 μm^2 memory cell technology for reliable 3.3 V 64 Mb EEPROMs, *1993 International Conference on Solid State Device and Material (SSDM93)*, pp. 446–448, 1993.

[18] Aritome, S.; Hatakeyama, I.; Endoh, T.; Yamaguchi, T.; Susumu, S.; Iizuka, H.; Maruyama, T.; Watanabe, H.; Hemink, G.; Koji, S.; Tanaka, T.; Momodomi, M.; and Shirota, R. An advanced NAND-structure cell technology for reliable 3.3V 64 Mb electrically erasable and programmable read only memories (EEPROMs), *Jpn. J. Appl. Phys.* vol. 33, pp. 524–528, Jan. 1994.

[19] Watanabe, H.; Aritome, S.; Hemink, G. J.; Maruyama, T.; Shirota, R. Scaling of tunnel oxide thickness for flash EEPROMs realizing stress-induced leakage current reduction, *VLSI Technology, 1994. Digest of Technical Papers. 1994 Symposium on*, pp. 47–48, 7–9 June 1994.

[20] Aritome, S.; Shirota R.; Sakui, K.; Masuoka, F. Data retention characteristics of flash memory cells after write and erase cycling, *IEICE Trans. Electron.*, vol. E77-C, no. 8, pp. 1287–1295, Aug. 1994.

[21] Tanaka, T.; Tanaka, Y.; Nakamura, H.; Sakui, K.; Oodaira, H.; Shirota, R.; Ohuchi, K.; Masuoka, F.; Hara, H. A quick intelligent page-programming architecture and a shielded bitline sensing method for 3 V-only NAND flash memory, *Solid-State Circuits, IEEE Journal of*, vol. 29, no. 11, pp. 1366–1373, Nov. 1994.

[22] Aritome, S.; Satoh, S.; Maruyama, T.; Watanabe, H.; Shuto, S.; Hemink, G. J.; Shirota, R.; Watanabe, S.; Masuoka, F. A 0.67 μm^2 self-aligned shallow trench isolation cell (SA-STI cell) for 3 V–only 256 Mbit NAND EEPROMs, *Electron Devices Meeting, 1994. IEDM'94. Technical Digest., International*, pp. 61–64, 11–14 Dec. 1994.

[23] Imamiya, K.; Iwata, Y.; Sugiura, Y.; Nakamura, H.; Oodaira, H.; Momodomi, M.; Ito, Y.; Watanabe, T.; Araki, H.; Narita, K.; Masuda, K.; Miyamoto, J. A 35 ns-cycle-time 3.3 V-only 32 Mb NAND flash EEPROM, *Solid-State Circuits Conference, 1995. Digest of Technical Papers. 42nd ISSCC, 1995 IEEE International*, pp. 130–131, 351, 15–17 Feb. 1995.

[24] Iwata, Y.; Imamiya, K.; Sugiura, Y.; Nakamura, H.; Oodaira, H.; Momodomi, M.; Itoh, Y.; Watanabe, T.; Araki, H.; Narita, K.; Masuda, K.; Miyamoto, J.-I. A 35 ns cycle time 3.3 V only 32 Mb NAND flash EEPROM, *Solid-State Circuits, IEEE Journal of*, vol. 30, no. 11, pp. 1157–1164, Nov. 1995.

[25] Suh, K.-D.; Suh, B.-H.; Um, Y.-H.; Kim, J.-Ki; Choi, Y.-J.; Koh, Y.-N.; Lee, S.-S.; Kwon, S.-C.; Choi, B.-S.; Yum, J.-S; Choi, J.-H.; Kim, J.-R.; Lim, H.-K. A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme, *Solid-State Circuits Conference, 1995. Digest of Technical Papers. 42nd ISSCC, 1995 IEEE International*, pp.128–129, 350, 15–17 Feb. 1995.

[26] Suh, K.-D.; Suh, B.-H.; Lim, Y.-H.; Kim, J.-K.; Choi, Y.-J.; Koh, Y.-N.; Lee, S.-S.; Kwon, S.-C.; Choi, B.-S.; Yum, J.-S.; Choi, J.-H.; Kim, J.-R.; Lim, H.-K. A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme, *Solid-State Circuits, IEEE Journal of*, vol. 30, no. 11, pp. 1149–1156, Nov. 1995.

[27] Satoh, S.; Hemink, G. J.; Hatakeyama, F.; Aritome, S. Stress induced leakage current of tunnel oxide derived from flash memory read-disturb characteristics, *Microelectronic Test Structures, 1995. ICMTS 1995. Proceedings of the 1995 International Conference on*, pp. 97–101, 22–25 Mar. 1995.

[28] Satoh, S.; Hemink, G.; Hatakeyama, K.; Aritome, S.; Stress-induced leakage current of tunnel oxide derived from flash memory read-disturb characteristics, *IEEE Transactions on Electron Devices*, vol. 45, no. 2, pp. 482–486 1998.

[29] Hemink, G. J.; Tanaka, T.; Endoh, T.; Aritome, S.; Shirota, R. Fast and accurate programming method for multi-level NAND EEPROMs, *VLSI Technology, 1995. Digest of Technical Papers. 1995 Symposium on*, pp. 129–130, 6–8 June 1995.

[30] Takeuchi, K.; Tanaka, T.; Nakamura, H. A double-level-V_{th} select gate array architecture for multi-level NAND flash memories, *VLSI Circuits, 1995. Digest of Technical Papers., 1995 Symposium on*, pp. 69–70, 8–10 June 1995.

[31] Takeuchi, K.; Tanaka, T.; Nakamura, H. A double-level-V_{th} select gate array architecture for multilevel NAND flash memories, *Solid-State Circuits, IEEE Journal of*, vol. 31, no. 4, pp. 602–609, Apr. 1996.

[32] Aritome, S.; Takeuchi, Y.; Sato, S.; Watanabe, H.; Shimizu, K.; Hemink, G.; Shirota, R. A novel side-wall transfer-transistor cell (SWATT cell) for multi-level NAND EEPROMs, *Electron Devices Meeting, 1995. International*, pp. 275–278, 10–13 Dec. 1995.

[33] Aritome, S.; Takeuchi, Y.; Sato, S.; Watanabe, I.; Shimizu, K.; Hemink, G.; Shirota, R. A side-wall transfer-transistor cell (SWATT cell) for highly reliable multi-level NAND EEPROMs, *Electron Devices, IEEE Transactions on*, vol. 44, no. 1, pp. 145–152, Jan. 1997.

[34] Jung, T.-S.; Choi, Y.-J.; Suh, K.-D.; Suh, B.-H.; Kim, J.-K.; Lim, Y.-H.; Koh, Y.-N.; Park, J.-W.; Lee, K.-J.; Park, J.-H.; Park, K.-T.; Kim, J.-R.; Lee, J.-H.; Lim, H.-K. A 3.3 V 128 Mb multi-level NAND flash memory for mass storage applications, in Solid-State Circuits Conference, 1996. Digest of Technical Papers. 42nd ISSCC., 1996 IEEE International, pp. 32–33, 10-10 Feb. 1996.

[35] Jung, T.-S.; Choi, Y.-J.; Suh, K.-D.; Suh, B.-H.; Kim, J.-K.; Lim, Y.-H.; Koh, Y.-N.; Park, J.-W.; Lee, K.-J.; Park, J.-H.; Park, K.-T.; Kim, J.-R.; Yi, J.-H.; Lim, H.-K. A 117-mm^2 3.3-V only 128-Mb multilevel NAND flash memory for mass storage applications, *Solid-State Circuits, IEEE Journal of*, vol. 31, no. 11, pp. 1575–1583, Nov. 1996.

[36] Kim, J.-K.; Sakui, K.; Lee, S.-S.; Itoh, J.; Kwon, S.-C.; Kanazawa, K.; Lee, J.-J.; Nakamura, H.; Kim, K.-Y.; Himeno, T.; Jang-Rae, Kim; Kanda, K.; Tae-Sung, Jung; Oshima, Y.; Kang-Deog, Suh; Hashimoto, K.; Sung-Tae Ahn; Miyamoto, J. A 120 mm^2 64 Mb NAND flash memory achieving 180 ns/byte effective program speed, *VLSI Circuits, 1996. Digest of Technical Papers 1996 Symposium on*, pp. 168–169, 13–15 June 1996.

[37] Kim, J.-K.; Sakui, K.; Lee, S.-S.; Itoh, Y.; Kwon, S.-C.; Kanazawa, K.; Lee, K.-J.; Nakamura, H.; Kim, K.-Y.; Himeno, T.; Kim, J.-R.; Kanda, K.; Jung, T.-S.; Oshima, Y.; Suh, K.-D.; Hashimoto, K.; Ahn, S.-T.; Miyamoto, J. A 120-mm^2 64-Mb NAND flash memory achieving 180 ns/Byte effective program speed, *Solid-State Circuits, IEEE Journal of*, vol. 32, no. 5, pp. 670–680, May. 1997.

[38] Hemink, G. J.; Shimizu, K.; Aritome, S.; Shirota, R. Trapped hole enhanced stress induced leakage currents in NAND EEPROM tunnel oxides, *IEEE International Reliability Physics Symposium, 1996. 34th Annual Proceedings*, pp. 117–121, 1996.

[39] Tanzawa, T.; Tanaka, T.; Takeuchi, K.; Shirota, R.; Aritome, S.; Watanabe, H.; Hemink, G.; Shimizu, K.; Sato, S.; Takeuchi, Y.; Ohuchi, K. A compact on-chip ECC for low cost flash memories, *VLSI Circuits, 1996. Digest of Technical Papers, 1996 Symposium on*, pp. 74–75, 13–15 June 1996.

[40] Tanzawa, T.; Tanaka, T.; Takeuchi, K.; Shirota, R.; Aritome, S.; Watanabe, H.; Hemink, G.; Shimizu, K.; Sato, S.; Takeuchi, Y.; Ohuchi, K. A compact on-chip ECC for low cost flash memories, *Solid-State Circuits, IEEE Journal of*, vol. 32, no. 5, pp. 662–669, May 1997.

[41] Choi, J. D.; Kim, D. J.; Tang, D. S.; Kim, J.; Kim, H. S.; Shin, W. C.; Ahn, S. T.; Kwon, O.H. A novel booster plate technology in high density NAND flash memories for voltage scaling-down and zero program disturbance, *VLSI Technology, 1996. Digest of Technical Papers. 1996 Symposium on*, pp. 238–239, 11–13 June 1996.

[42] Kim, D.J.; Choi, J. D.; Kim, J.; Oh, H. K.; Ahn, S. T.; Kwon, O. H. Process integration for the high speed NAND flash memory cell, *VLSI Technology, 1996. Digest of Technical Papers. 1996 Symposium on*, pp. 236–237, 11–13 June 1996.

[43] Watanabe, H.; Shimizu, K.; Takeuchi, Y.; Aritome, S. Corner-rounded shallow trench isolation technology to reduce the stress-induced tunnel oxide leakage current for highly reliable flash memories, *Electron Devices Meeting, 1996. IEDM'96., International*, pp. 833–836, 8–11 Dec. 1996.

[44] Shin, W. C.; Choi, J. D.; Kim, D. J.; Kim, H. S.; Mang, K. M.; Chung, C. H.; Ahn, S. T.; and Kwon, O. H.. A new shared bit line NAND Cell technology for the 256 Mb flash memory with 12V programming, *Electron Devices Meeting, 1996. IEDM'96, International*, Dec. 1996.

[45] Jung, T.-S.; Choi, D.-C.; Cho, S.-H.; Kim, M.-J.; Lee, S.-K.; Choi, B.-S.; Yum, J.-S.; Kim, S.-H.; Lee, D.-G.; Son, J.-C.; Yong, M.-S.; Oh, H.-K.; Jun, S.-B.; Lee, W.-M.; Haq, E.; Suh, K.-D.; Ali, S.; Lim, H.-K. A 3.3 V 16 Mb nonvolatile virtual DRAM using a NAND flash memory technology, *Solid-State Circuits Conference, 1997. Digest of Technical Papers. 43rd ISSCC, 1997 IEEE International*, pp. 398–399, 493, 6–8 Feb. 1997.

[46] Jung, T.–S.; Choi, D.-C.; Cho, S.-H.; Kim, M.-J.; Lee, S.-K.; Choi, B.-S.; Yum, Jin-Sun; Kim, S.-H.; Lee, D.-G.; Son, J.-C.; Yong, M.-S.; Oh, H.-K.; Jun, S.-B.; Lee, W.-M.; Haq, E.; Suh, K.-D.; Ali, S. B.; Lim, H.-K.; A 3.3-V single power supply 16-Mb nonvolatile virtual DRAM using a NAND flash memory technology, *Solid-State Circuits, IEEE Journal of*, vol. 32, no. 11, pp. 1748–1757, Nov. 1997.

[47] Tanaka, T.; Tanzawa, T.; Takeuchi, K.; , A 3.4-Mbyte/sec programming 3-level NAND flash memory saving 40% die size per bit, *VLSI Circuits, 1997. Digest of Technical Papers., 1997 Symposium on*, pp. 65–66, 12–14 June 1997.

[48] Takeuchi, K.; Tanaka, T.; Tanzawa, T. A Multi-page Cell Architecture for high-speed programming multi-level NAND flash memories, *VLSI Circuits, 1997. Digest of Technical Papers, 1997 Symposium on*, pp. 67–68, 12–14 June 1997.

[49] Takeuchi, K.; Tanaka, T.; Tanzawa, T. A multipage cell architecture for high-speed programming multilevel NAND flash memories, *Solid-State Circuits, IEEE Journal of*, vol. 33, no. 8, pp. 1228–1238, Aug. 1998.

[50] Kim, H. S.; Choi, J. D.; Kim, J.; Shin, W. C.; Kim, D. J.; Mang, K. M.; Ahn, S. T. Fast parallel programming of multi-level NAND flash memory cells using the booster-line technology, *VLSI Technology, 1997. Digest of Technical Papers, 1997 Symposium on*, pp. 65–66, 10–12 June 1997.

[51] Shimizu, K.; Narita, K.; Watanabe, H.; Kamiya, E.; Takeuchi, Y.; Yaegashi, T.; Aritome, S.; Watanabe, T. A novel high-density $5F^2$ NAND STI cell technology suitable for 256 Mbit and 1 Gbit flash memories, *Electron Devices Meeting, 1997. IEDM'97. Technical Digest, International*, pp. 271–274, 7–10 Dec. 1997.

[52] Satoh, S.; Hagiwara, H.; Tanzawa, T.; Takeuchi, K.; Shirota, R. A novel isolation-scaling technology for NAND EEPROMs with the minimized program disturbance, *Electron Devices Meeting, 1997. IEDM'97. Technical Digest, International*, pp. 291–294, 7–10 Dec. 1997.

[53] Choi, J. D.; Lee, D. G.; Kim, D. J.; Cho, S. S.; Kim, H. S.; Shin, C. H.; Ahn, S. T. A triple polysilicon stacked flash memory cell with wordline self-boosting programming, *Electron Devices Meeting, 1997. IEDM'97. Technical Digest, International*, pp. 283–286, 7–10 Dec. 1997.

[54] F. Masuoka, flash memory makes a big leap, *Kogyo Chosakai*, vol. 1, pp. 1–172, 1992. (in Japanese).

[55] Aritome, S., NAND flash innovations, *Solid-State Circuits Magazine, IEEE*, vol. 5, no. 4, pp. 21,29, Fall 2013.

[56] Aritome, S. Advanced flash memory technology and trends for file storage application *Electron Devices Meeting, 2000. IEDM Technical Digest International*, pp. 763–766, 2000.

[57] Hwang, J.; Seo, J.; Lee, Y.; Park, S.; Leem, J.; Kim, J.; Hong, T.; Jeong, S.; Lee, K.; Heo, H.; Lee, H.; Jang, P.; Park, K.; Lee, M., Baik, S.; Kim, J.; Kkang, H.; Jang, M.; Lee, J.; Cho, G.; Lee, J.; Lee, B.; Jang, H.; Park, S.; Kim, J.; Lee, S.; Aritome, S.; Hong, S.; and Park, S. A middle-1X nm NAND flash memory cell (M1X-NAND) with highly manufacturable integration technologies, *Electron Devices Meeting (IEDM), 2011 IEEE International*, pp. 199–202, Dec. 2011.

[58] Tanaka, H.; Kido, M.; Yahashi, K.; Oomura, M.; Katsumata, R.; Kito, M.; Fukuzumi, Y.; Sato, M.; Nagata, Y.; Matsuoka, Y.; Iwata, Y.; Aochi, H.; Nitayama, A. Bit cost scalable technology with punch and plug process for ultra high density flash memory, *VLSI Symposium Technical. Digest., 2007*, pp. 14–15.

[59] Katsumata, R.; Kito, M.; Fukuzumi, Y.; Kido, M.; Tanaka, H.; Komori, Y.; Ishiduki, M.; Matsunami, J.; Fujiwara, T.; Nagata, Y.; Zhang, L.; Iwata, Y.; Kirisawa, R.; Aochi, H.; Nitayama, A. Pipe-shaped BiCS flash memory with 16 stacked layers and multi-level-cell operation for ultra high density storage devices, *VLSI Symposium Technology. Digest.*, pp. 136–137, 2009.

[60] Kim, J.; Hong, A. J.; Ogawa, M.; Ma, S.; Song, E. B.; Lin, Y.-S.; Han, J.; Chung, U.-I.; Wang, K. L. Novel 3-D structure for ultra high density flash memory with VRAT (vertical-recess-array-transistor) and PIPE (planarized integration on the same plane), *VLSI Symposium Technology. Digest., 2008*, pp. 122–123.

[61] Kim, W. J.; Choi, S.; Sung, J.; Lee, T.; Park, C.; Ko, H.; Jung, J., Yoo, I.; Park, Y. Multi-layered vertical gate NAND flash overcoming stacking limit for terabit density storage, *VLSI Symposium Technology. Digest., 2009*, pp. 188–189.

[62] Jang, J.; Kim, H.-S.; Cho, W.; Cho, H.; Kim, J.; Shim, S.I.; Jang, Y.; Jeong, J.-H.; Son, B.-K.; Kim, D. W.; Kim, K.; Shim, J.-J.; Lim, J. S; Kim, K.-H.; Yi, S. Y.; Lim, J.-Y.; Chung, D.; Moon, H.-C.; Hwang, S.; Lee, J.-W.; Son, Y.-H.; U-In Chung and Lee, W.-S. Vertical cell array using TCAT (terabit cell array transistor) technology for ultra high density NAND flash memory", *VLSI Symposium Technology. Digest, 2009*, pp. 192–193.

[63] Lue, H.-T.; Hsu, T.-H.; Hsiao, Y.-H.; Hong, S. P.; Wu, M. T.; Hsu, F. H.; Lien, N. Z.; Wang, S.-Y.; Hsieh, J.-Y.; Yang, L.-W.; Yang, T.; Chen, K.-C.; Hsieh, K.-Y.; Lu, C.-Y.; A highly scalable 8-layer 3D vertical-gate (VG) TFT NAND flash using junction-free buried channel BE-SONOS device, *VLSI Technology (VLSIT), 2010 Symposium on*, pp. 131–132, 15–17 June 2010.

[64] Park, K.-T.; Nam, S.; Kim, D.; Kwak, P.; Lee, D.; Choi, Y.-H.; Choi, M.-H.; Kwak, D.-H.; Kim, D.-H.; Kim M.-S.; Park, H.-W.; Shim, S.-W.; Kang, K.-M.; Park, S.-W.; Lee, K.; Yoon, H.-J.; Ko, K.; Shim, D.-K.; Ahn, Y.-L.; Ryu, J.; Kim, D.; Yun, K.; Kwon, J.; Shin, S.; Byeon, D.-S.; Choi, K.; Han, J.-M.; Kyung, K.-H.; Choi, J.-H.; Kim, K. Three-dimensional 128 Gb MLC vertical NAND flash memory with 24-WL stacked layers and 50 MB/s high-speed programming, *Solid-State Circuits, IEEE Journal of*, vol. 50, no. 1, pp. 204, 213, Jan. 2015.

2

PRINCIPLE OF NAND FLASH MEMORY

2.1 NAND FLASH DEVICE AND ARCHITECTURE

2.1.1 NAND Flash Memory Cell Architecture

Figure 2.1 shows the single-cell architecture of flash memory. The single cell has an Nch MOS transistor with poly-silicon floating gate (FG). FG is electrically isolated by tunnel oxide and interpoly dielectric (IPD). Charge is stored in FG, and potential of FG is controlled by control gate (CG) voltage with capacitive coupling between CG and FG. Figure 2.2 shows the NAND string structure [1]. The NAND string consists of 32 cells (typical) and two select transistors (SGD, SGS). Thirty-two cells are connected in series. SGD connects at a drain to isolate it from a bit line (BL), and SGS connects at a source to isolate it from a source line (SL). The number of cells in one NAND string has been increased with steps of 8 cells → 16 cells → 32 cells (0.12-μm generation~) → 64 cells (~43-nm generation).

Figure 2.3 shows (a) a top view and (b) a cross-sectional view of NAND flash memory cells. One NAND string consists of 32 series-connected stacked gate memory transistors and two select gate transistors. The entire memory array is formed by straight active area lines (horizontal lines) and straight gate lines (vertical lines). This simple cell structure enables easy fabrication of memory cells, which have small feature size. The memory cell is located at an intersection of the two cross lines. Thirty-two memory transistors are arranged between two select transistors, so that both the bit-line and source-line contacts are arranged every 34 gate lines (vertical lines).

Nand Flash Memory Technologies, First Edition. Seiichi Aritome.

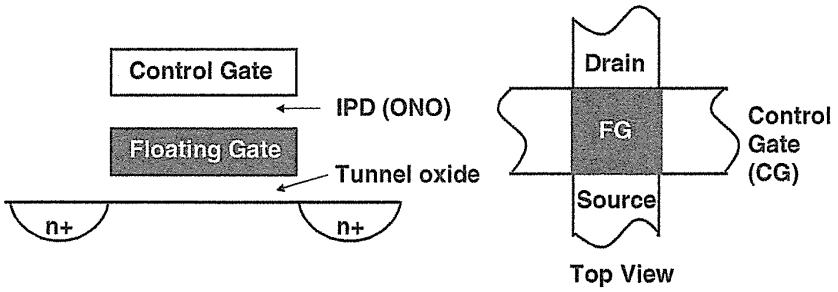


FIGURE 2.1 Single-cell architecture of floating-gate flash memory. Nch MOS transistor with poly-silicon floating gate (FG). The FG is electrically isolated by tunnel oxide and inter-poly dielectric (IPD). Charge is stored in FG.

Figure 2.4 shows the array architecture of a NAND flash memory cell. Page size is typically 2 K to 16 KByte (2- to 16-KByte cells + ECC code). Page size is increasing for enhancing the performance of read and program operations. The original page size was 512-Byte cells + ECC code [2–5]. And the page size was increased with steps of 512 B → 2 KB (0.12-μm generation~) → 4 KB (~56-nm generation) → 8 KB (~43-nm generation) → 16 KB (~2X- to 1X-nm generation, with an all bit-line scheme). The page size will be more increased for future devices because much higher performance will be required. The block size is basically page × 2 × 32 cells = 128-K to 256-KByte cells (physical) in this case. The block size was also increased by increasing the number of cells in the NAND string and increasing the page size. The read and program operations are performed per page. And the erase operation is performed per block.

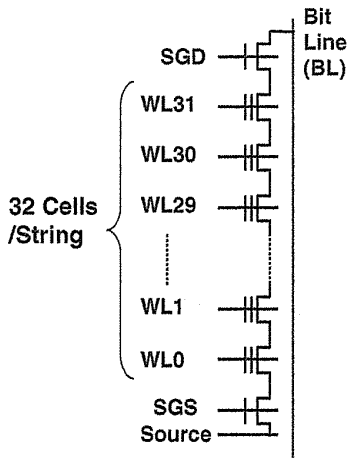


FIGURE 2.2 NAND cell string architecture. NAND cell string consists of 32 cells (for example) and two select transistors (SGD, SGS). Thirty-two cells are connected in series. SGD and SGS connect at drain and source to isolate from bit line and source line, respectively.

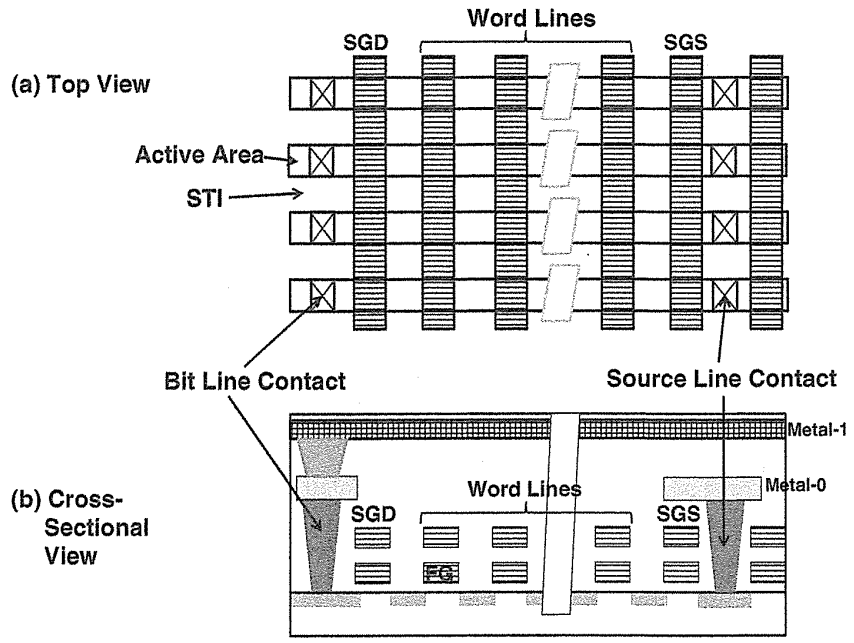


FIGURE 2.3 (a) Top schematic view and (b) cross-sectional view of NAND flash memory cells.

2.1.2 Peripheral Device

Figure 2.5 shows the cross-sectional view of a typical NAND flash memory device [6]. A memory cell array is fabricated on a double well consisting of a *p*-well and an *n*-well. Low-voltage transistors of N-ch (N-channel) and P-ch (P-channel) are on the *p*-well and the *n*-well, respectively. A high-voltage transistor of N-ch is fabricated

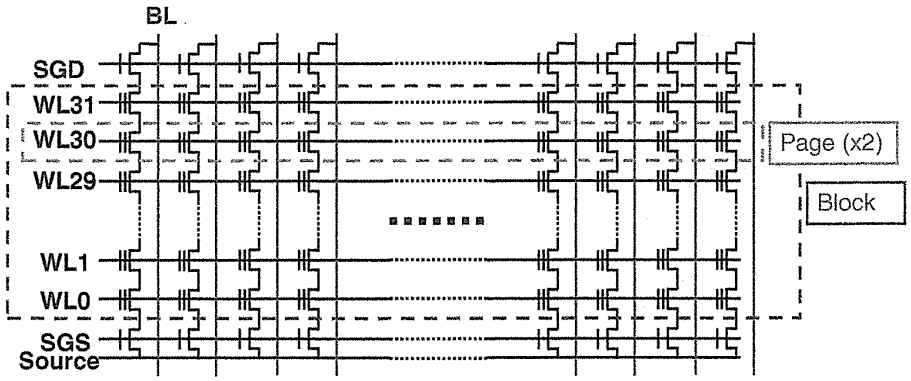


FIGURE 2.4 Array architecture of NAND flash memory cells. Page size is typically 2-K to 16-KByte (2- to 16-KByte cells + ECC code). Block size is typically page × 2 × 32 cells = 128-K to 256-KByte cell (physical). Read and program are performed in per page. Erase is performed in per block.

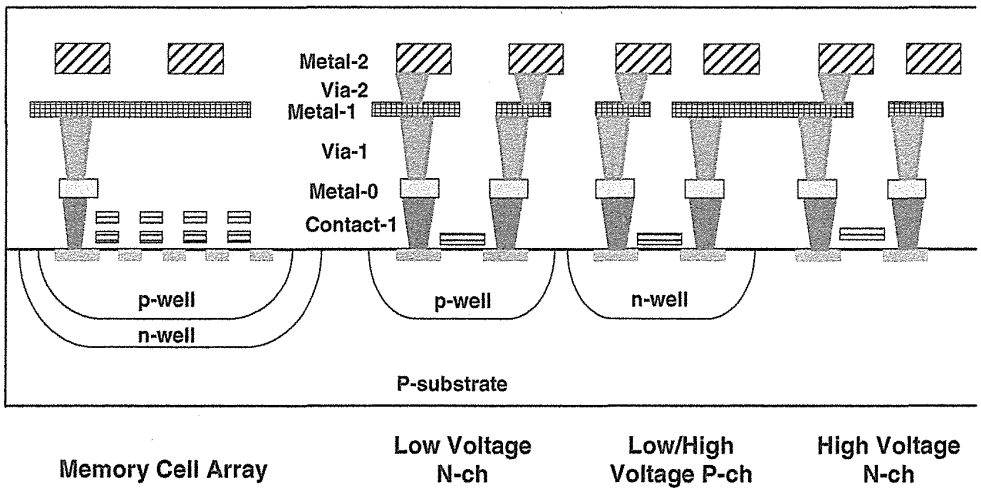


FIGURE 2.5 The cross-sectional view of NAND flash memory.

on a P-substrate. A high-voltage transistor of P-ch is fabricated on an *n*-well, whicl is basically the same *n*-well as for a low-voltage P-ch transistor. A NAND flasl memory device has normally three metal layers of Metal-0, Metal-1, and Metal-2. Cı metal layers with a damascene process are used for Metal-1 and Metal-2 in severa companies' products.

The peripheral transistors for a NAND flash memory device are listed in Fig. 2.6 The lineup of peripheral transistor types is dependent on circuit requirement. Fig ure 2.6 shows a typical case of a transistors list. A thick gate oxide of 25–40 nm is usec for a high-voltage transistor. A low-voltage transistor has a thin gate oxide (~9 nm which is the same as tunnel oxide in a memory cell. In order to realize low proces cost, it is important that peripheral Tr is fabricated without increasing process steps

NAND Flash Peripheral transistors

Low-voltage (LV) transistor N-ch E-type P-ch E-type
High-voltage (HV) transistor N-ch E-type N-ch I-type N-ch D-type P-ch E(+)-type

Where E-type = Enhancement type
I-type = Intrinsic type (V_t ~ 0 V)
D-type = Depletion type
E(+)-type = deep Enhancement type (V_t ~ –3 V)

FIGURE 2.6 List of peripheral transistors in NAND flash memory. V_t; threshold voltage.

The scaling of a high-voltage transistor (HV Tr) is one of the important challenges for NAND flash memories. The size scaling of a high-voltage transistor (HV Tr) is required for placing HV Tr and isolation on a certain determined cell pitch. There are two critical locations of HV Tr scaling. One is the row decoder/word-line driver, and HV Tr + isolation width have to be in one string pitch basically. The other location is the bit-line selector area. The size of an HV transistor has to be as small as possible to make a small area of bit line selector.

2.2 CELL OPERATION

2.2.1 Read Operation

Figure 2.7 shows basic read operation of single cell. Data (0 or 1) is judged by cell current (I_{cell}) flow "ON" or not flow "OFF" during V_{gate} = 0 V applied (for SLC (1 bit/cell) case). The erased state has a positive charge in an FG. The programmed state has a negative charge in FG.

Read operations are performed in page units, as shown in Fig. 2.8. A page corresponds to a row of cells and is accessed by a single word line. V_{passR} (~6 V) is applied to an unselected word line to be as a pass transistor. Random read speed (*tR*) is normally 25 μs for SLC and 60 μs for MLC (2 bits/cell). However, *tR* is becoming longer as a cell scaling because of decreasing cell current, I_{cell}.

A read disturb problem is mainly caused in unselected cells where V_{passR} is applied to a control gate (CG) and channel is 0 V. It is the weak electron injection mode, and then cell V_t (threshold voltage) is gradually increased. Detail read disturb characteristics are discussed in Chapter 6.

2.2.2 Program and Erase Operation

Several program and erase schemes were considered to use a NAND flash memory product in early stage of development, as shown in Fig. 2.9. There are only two

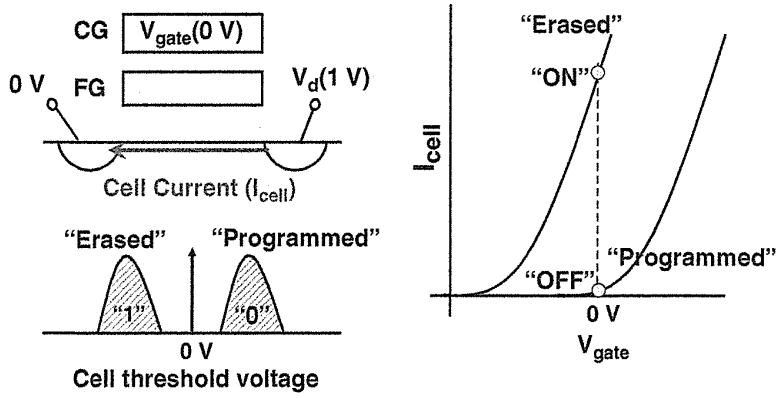


FIGURE 2.7 Principle of read of single cell. Erased: Positive charge in FG. I_{cell} "ON". Programmed: Negative charge in FG. I_{cell} "OFF".

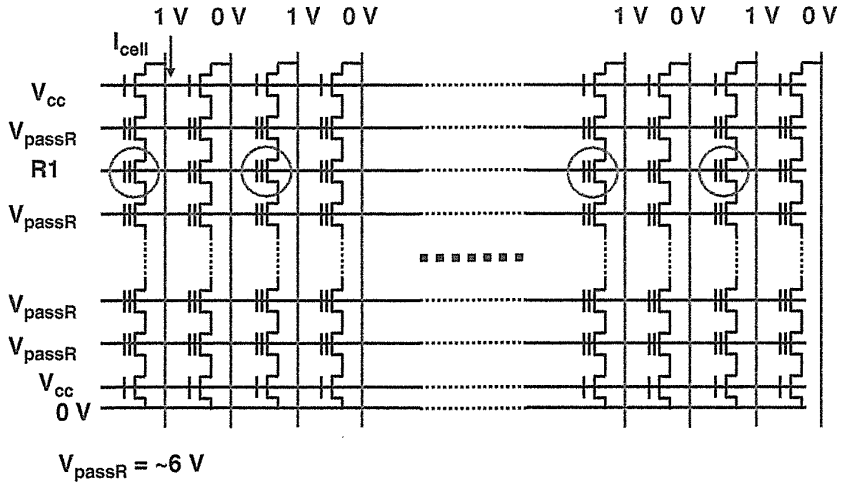


FIGURE 2.8 Read operation of NAND cell array. Page read: 2–16 KByte at the same time. Typical random read speed, *tR*: 25 μs for SLC, 60 μs for MLC.

ways of giving an electron injection to a floating gate (FG) of a channel hot electron (CHE) injection and a Fowler–Nordheim (FN) channel injection. And also, there are only two ways of electron ejection from an FG of a drain/source FN ejection and a channel FN ejection. Four schemes of programming and erase operations

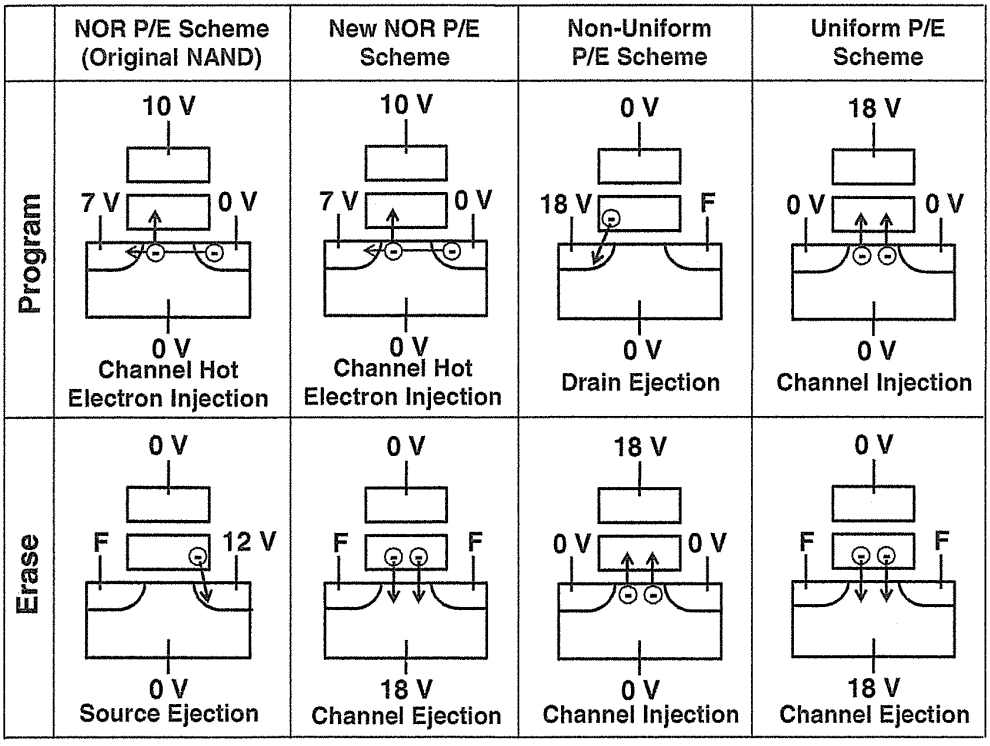


FIGURE 2.9 Program and erase schemes of NOR and NAND flash memory cells.

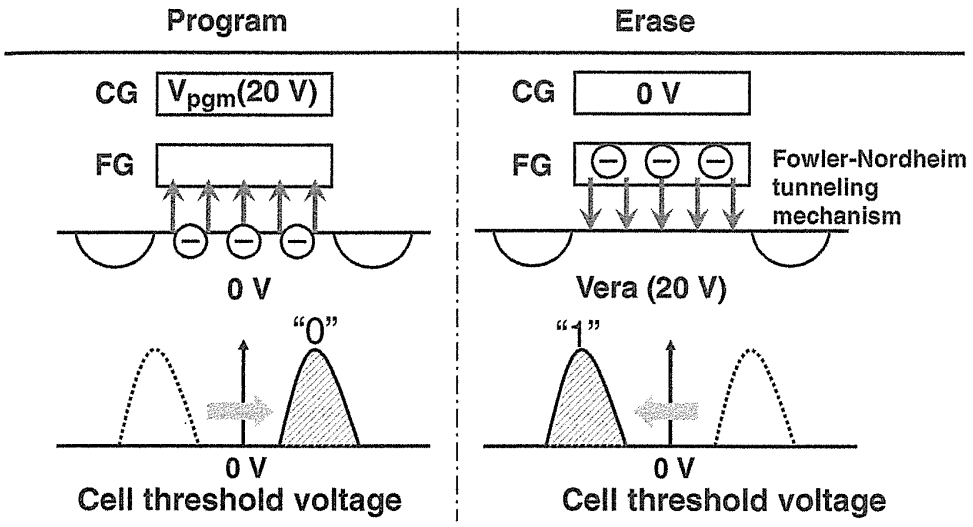


FIGURE 2.10 Principle of single-cell program and erase. Program: Injected electron to the FG through tunnel oxide. Erase: Ejected electron from the FG.

are possible by combination of these ways of injection and ejection, as shown in Fig. 2.9. The first one is the NOR-type flash program/erase scheme. CHE injection is used for programming, and source FN ejection is used for erase. The second one is the new NOR-type flash program/erase scheme [7]. CHE injection is used for programming, and channel FN ejection is used for erase. The third one is the old program/erase scheme for NAND flash [8, 9]. Drain FN ejection is used for programming, and channel FN injection is used for erasing. The fourth one is the current NAND-type flash program/erase scheme [7, 10, 11]. Channel FN injection is used for programming, and channel FN ejection is used for erasing.

Power consumption is worse in both CHE injection and drain/source FN ejection schemes due to large current flow. The memory cell scalability is also worse in CHE injection and drain/source FN ejection schemes due to applying large voltage to a drain/source. And reliability is worse in drain/source FN ejection schemes due to degradation by a generated hot hole, as described in Chapter 6. Therefore, the program/erase schemes for a NAND flash memory product use a uniform program/erase (P/E) scheme, as shown in Fig. 2.9 [7, 10, 11]. A detailed program and an erase operation are described in the following.

The program operation of a NAND flash cell is performed by applying high-voltage V_{pgm} to control gate (CG) while keeping substrate/source/drain 0 V, as shown in Fig. 2.10. Electrons are injected to the floating gate (FG) by a Fowler–Nordheim (FN) tunneling mechanism through the tunnel oxide. V_t of cell has a positive shift. An erase operation is performed by applying high-voltage V_{era} to substrate (p-well), while keeping CG at 0 V. Electrons in an FG are ejected to a substrate through a tunnel oxide. V_t has a negative shift.

Figure 2.11 shows a typical program characteristic. Higher program voltage (V_{pgm}) or longer program pulse width can cause faster programming.

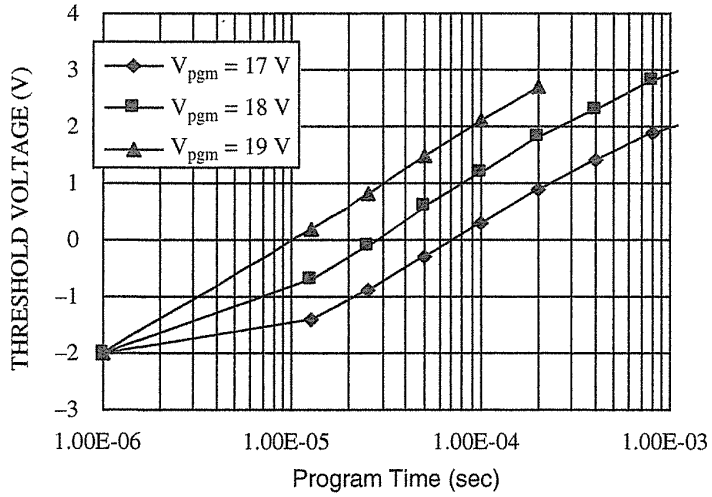


FIGURE 2.11 Typical program characteristics. High program voltage can be faster operation.

Figure 2.12 shows the cell array program scheme. At first, the programming starts at the source side cells. V_{pgm} is applied to the control gate (CG0; WL0) of the selected cells while V_{pass} is applied to the control gates of the unselected cells. These unselected cells act as pass transistors and make the boosting voltage in a channel to prevent boosting mode program disturb, as described in Section 2.2.4. A 0-V charge is applied to the bit line, and then electrons are injected from the bit line (channel) to a floating gate by the electric field between the bit line and the floating gate of the selected cell. The threshold voltage of the selected cell is pushed up into the

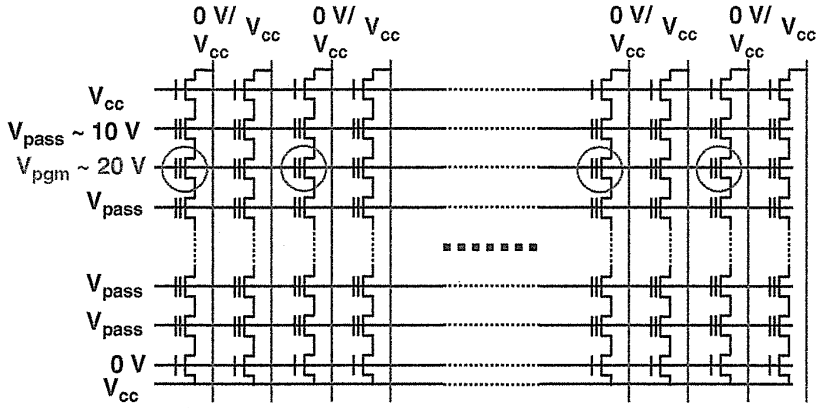


FIGURE 2.12 Program operation of NAND flash cell array. Page program: 2- to 16-KByte program at the same time. All cells in page should be programmed in one program sequence (prohibit partial page program). Program speed, t_{prog}: 200 μs for SLC, 800–1600 μs for MLC. Order of page programming is from source side page to drain side page (sequential page program in block). Random order of page programming is normally prohibited.

enhancement mode of approximately 2 V. In the case of a program inhibit mode, the bit line is raised to V_{cc}. The voltage of $V_{cc} - V_t$ (V_t of select Tr) is transferred to the channel of cell strings before applying V_{pass} to unselected CGs. Then V_{pass} are applied to the control gate of the unselected CGs to make the boosting voltage in the channel of unselected string. No electrons are injected from the channel (V_{boost}) to the floating gate, because the electric field between the bit line and the floating gate is insufficient to initiate tunneling. The threshold voltage of the selected cell remains at erase state of −3 V. After programming the cells connected with CG0 (WL0), the programming of the cells connected with CG1 (WL1) starts. V_{pgm} is applied to the control gate of the selected cell (CG1), and V_{pass} is applied to the control gates of the unselected cells (CG0, CG2–G31). Either 0 V or V_{cc} is applied to a bit line as corresponding with data. Subsequently, the programming continues from the source side cell to the bit-line side cell successively. Typical programming time per page, including the data load sequence, is 200 μs for SLC, 800–1600 μs for MLC.

Incremental step pulse programming (ISPP) achieves fast program performance under process and environmental variations while keeping a tight programmed cell V_t distribution [12–14]. Figure 2.13 shows the V_{pgm} waveform of an ISPP scheme. V_{pgm} has stepped up by each pulse. An ISPP scheme suppresses process variation issues effectively by allowing fast programmed cells to be programmed with a lower program voltage and slow program cells to be programmed with a higher program voltage. After an initial 15.5-V program pulse, each subsequent pulse (if required) is incremented in 0.5-V (ΔV_{pgm}) steps up to 20 V, for example. Since sufficiently programmed cells are automatically switched to the program inhibit state in the verification step, easily programmed cells are not affected by the higher program voltages. A 1-V program pulse increment is approximately as effective as five pulses without the increment. Thus, ISPP scheme has the effect of increasing pulse width

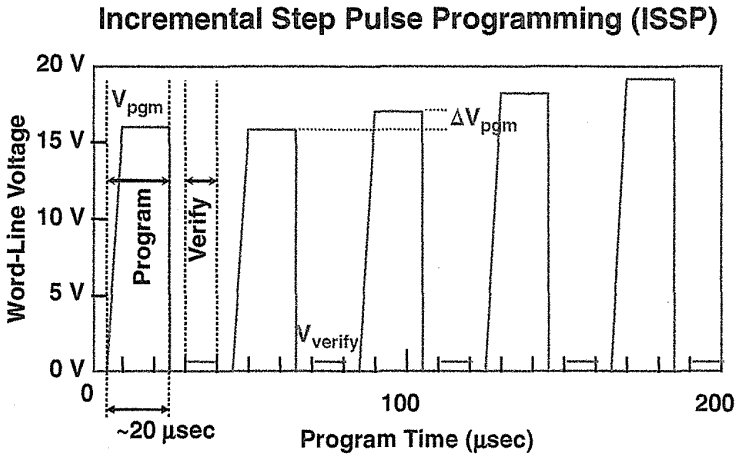


FIGURE 2.13 Incremental step pulse programming (ISSP) waveform. ISPP is used for page program to make a tight cell V_t distribution. Degradation of tunnel oxide can be mitigated due to relaxed maximum electric field in tunnel oxide during program.

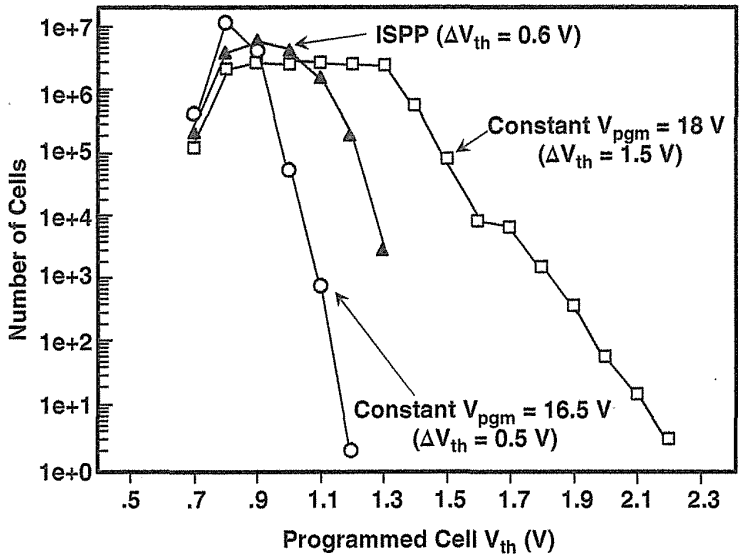


FIGURE 2.14 Comparison of programmed cell threshold voltage distribution in devices with ISPP and without ISPP (constant voltage programming with 16.5 and 18 V).

[15] without actually increasing the program time by dynamically optimizing program voltage to cell characteristics. Through program pulse/verify cycles, programmed cell V_t is maintained to within 0.6 V by using a 0.5-V ISPP step as shown in Fig. 2.14 [12]. With an ISPP scheme, a page is typically programmed within 2–6 program pulses/verify cycles for SLC (~300 μs). The constant 16.5-V program voltage device of Fig. 2.14 has the tightest V_t distribution (~0.5 V), however, program verify cycles are larger, consisting of 11–37 cycles. This means 2–16 times slower programming speed. Then an ISPP scheme provides an optimum combination of both a tight V_t distribution and a fast program time.

By effectively adjusting to process and environment variations, an ISPP scheme maintains a consistent program performance which helps to improve the yield of the device. Marginal cells that are previously out-of-spec when conditions are varied are brought within-spec with ISPP. An ISPP scheme is able to compensate for cell-by-cell variations that can exist within a die.

The program verify operation is performed just after program pulse (V_{pgm}) to check whether each cell V_t reach to verify level of V_{verify} or not, as shown in Fig. 2.13. The operation condition is almost the same as regular read operation, as shown in Fig. 2.8. The different point from regular read operation is that a control gate voltage is V_{verify}, replaced by R1 in Fig. 2.8.

The erase operation is performed in block units, as shown in Fig. 2.15. The wordlines of selected blocks are grounded and the wordlines of unselected blocks are floating. A high erase pulse (~ 20 V) is then applied to the *p*-well and *n*-well in the memory cell area (Fig. 2.5). In the selected blocks, the erase voltage creates a large (~20 V) potential difference between the *p*-well and the control gates. This causes FN tunneling of electrons from the floating gate into the *p*-well, resulting in a typical cell

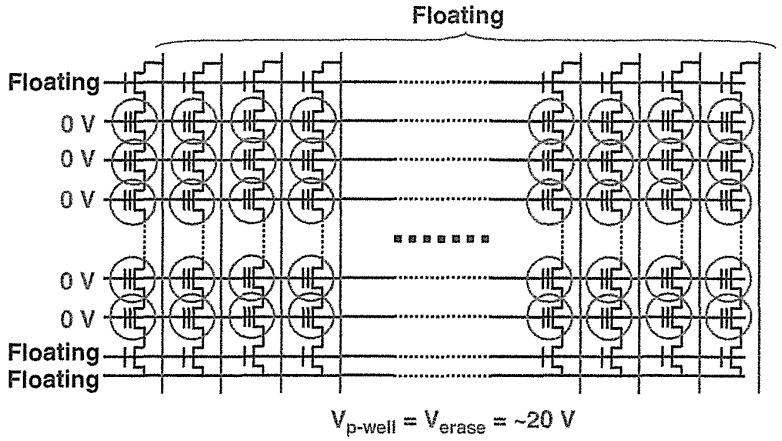


FIGURE 2.15 Erase operation of NAND flash cell array. Block erase: 128- to 256-KByte (2-KByte × 2 × 32) cells are erased at the same time. Erase speed, t_{erase}: 2–5 ms with erase verify.

threshold voltage that shifts negative. Since over-erasure is not a concern in NAND flash, cells are normally erased to −3 V. Also, the low erased cell threshold voltage provides an additional margin against upward threshold voltage shifts that arise from program/erase cycling, program disturb, read disturb, and V_t shift by floating gate capacitive coupling interference (see Chapter 5).

Figure 2.16 shows the typical erase characteristics. Cells can be erased to −3 V by applying a 17- to 18-V, 1-ms erase pulse width.

The erase verify operation is performed just after an erase pulse (V_{erase}) to check whether all cell V_t values in string become less than a certain V_t (for example, 0 V)

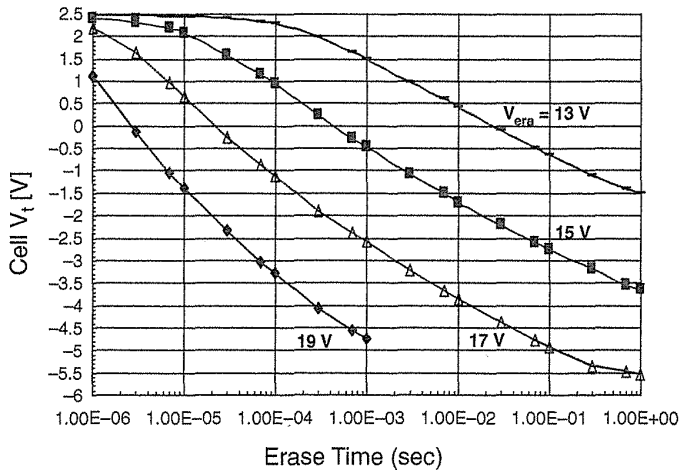


FIGURE 2.16 Erase characteristics. Higher erase voltage can be faster operation. Erase speed, t_{era}: 2–5 ms.

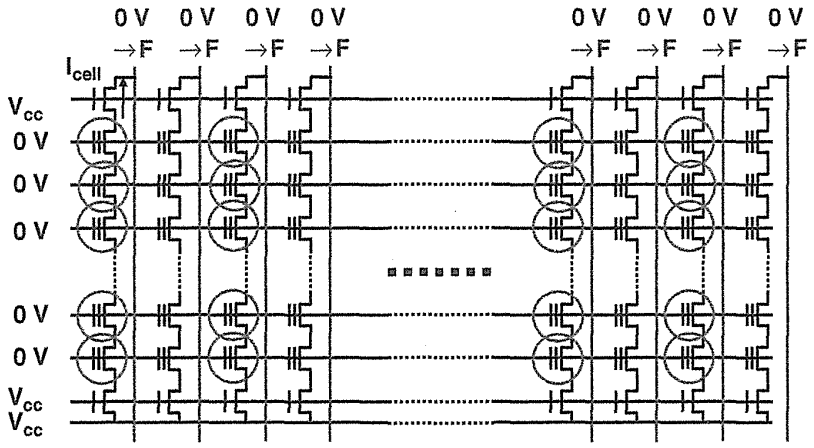


FIGURE 2.17 Erase verify operation.

or not. Figure 2.17 shows the operation condition of erase verify. A 0-V charge is applied to all control gates in a block, while V_{cc} and 0 V are applied to a source line and a bit line, respectively. The bit line is initially set to 0 V and then set to floating (F). During erase verify read, bit-line voltage is increased by cell current flow through a string, and then it judges whether all cells in the string are erased or not. If some cells are not be erased, an additional erase operation is performed in the same manner as that of the program/verify operation.

2.2.3 Program and Erase Dynamics

The device model is described for program and erase dynamics [16].

A. Calculation of Tunnel Current. The tunneling current density through the tunnel oxide is approximated by the well-known Fowler–Nordheim equation [17, 18].

$$J_{tun} = \alpha E_{tun}^2 (\exp(-\beta / E_{tun})) \tag{2.1}$$

where E_{tun} is the electric field in the tunnel oxide, and α and β are constants. The tunnel oxide field E_{tun} is given by

$$E_{tun} = |V_{tun}| / X_{tun} \tag{2.2}$$

where V_{tun} is the voltage drop across the tunnel oxide and X_{tun} is the thickness. V_{tun} can be calculated from a capacitive equivalent circuit of the cell.

B. Calculation of V_{tun}. In order to gain insight into the basic device operation, a simplified equivalent circuit is used, shown in Fig. 2.18. In Fig. 2.18, C_{pp} is the interpoly capacitance, C_{tun} is the thin oxide capacitance between the floating gate

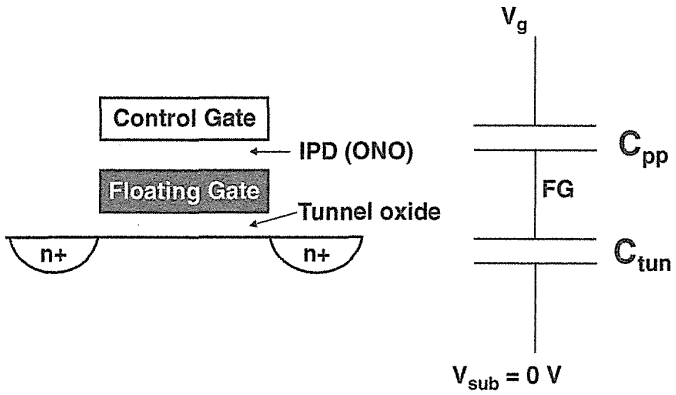


FIGURE 2.18 A simplified capacitive equivalent circuit of the NAND flash memory cell.

and the substrate. Q_{fg} is the stored charge on the floating gate. V_{tun} can be expressed for an electrically neutral floating gate in terms of simple coupling ratios:

$$|V_{tun}|_{write} = V_g * K_w \tag{2.3}$$

where

$$K_w = C_{pp}/(C_{pp} + C_{tun}) \tag{2.4}$$

and

$$|V_{tun}|_{erase} = V_{well} * K_e \tag{2.5}$$

where

$$K_e = 1 - C_{tun}/(C_{pp} + C_{tun}) \tag{2.6}$$

The coupling ratios K_w and K_e denote the fraction of the applied voltage that appears across the tunnel oxide. Note that (2.3) and (2.5) are applicable only when $Q_{fg} = 0$. During program operation, buildup of negative stored charge of the floating gate will reduce the tunnel-oxide voltage according to

$$|V_{tun}|_{write} = V_g * K_w + Q_{fg}/(C_{pp} + C_{tun}) \tag{2.3'}$$

In the ERASE operation, the initial negative stored charge on the floating gate will increase the tunnel-oxide voltage according to

$$|V_{tun}|_{erase} = V_{well} * K_e - Q_{fg}/(C_{pp} + C_{tun}) \tag{2.5'}$$

at the end of the erase operation when positive charge is built up on the floating gate, and the last term in (2.5′) will reduce the tunnel-oxide voltage.

C. Calculation of Threshold Voltages. The initial threshold voltage of the cell, corresponding to $Q_{fg} = 0$, is denoted by V_{ti}. Stored charge shifts the threshold according to the relation

$$\Delta Vt = -Q_{fg}/C_{pp} \tag{2.7}$$

Using (2.3′) and (2.5′) for Q_{fg} at the end of the program/erase pulse, the cell's threshold voltages are

$$V_{tw} = V_{ti} - Q_{fg}/C_{pp} = V_{ti} + V_g * (1 - V'_{tun}/(K_w * V_g)) \tag{2.8}$$

$$V_{te} = V_{ti} - Q_{fg}/C_{pp} = V_{ti} - V_{well} * (K_e/K_w - V'_{tun}/(K_w * V_{well})) \tag{2.9}$$

Here V_{tw} is the threshold of a programmed cell, and V_{te} is the threshold of an erased cell. V_g and V_{well} are the program/erase pulse amplitudes, respectively, and V'_{tun} is the tunnel-oxide voltage at the end of the pulse. Assuming that the program/erase pulse is sufficiently long, the thin-oxide field will be reduced to below about 1×10^7 V/cm, when tunneling practically "stops." An approximation of V'_{tun} can be calculated from (2.2), and it can be substituted in (2.8) and (2.9) to give the approximate programming window of the cell and its dependence on cell parameters and programming voltage. Typical results are shown in Fig. 2.19 [16].

In order to maximize the cell window at a given tunnel-oxide thickness and program/erase voltage, the coupling ratios should approach unity. Both coupling

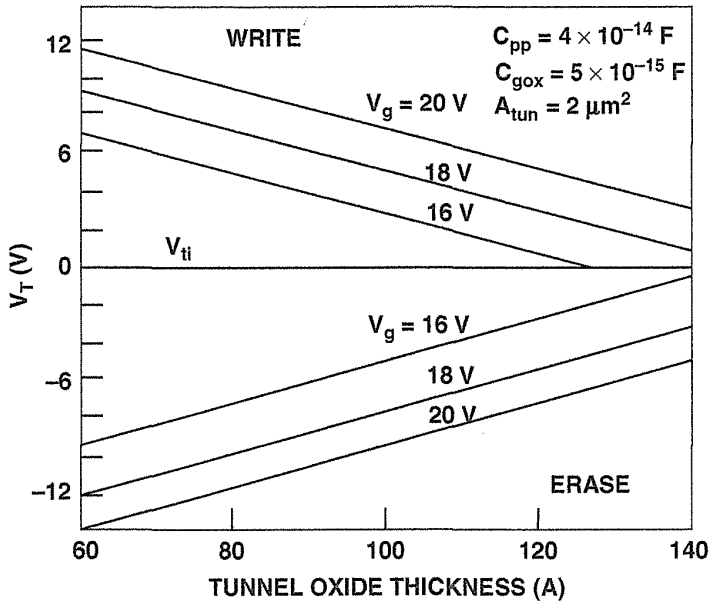


FIGURE 2.19 Program/erase threshold window versus tunnel-oxide thickness, calculated with the approximation of (2.8) and (2.9), assuming that $V_{tun}' = 1 \times 10^7 * X_{tun}$ at the end of the operation.

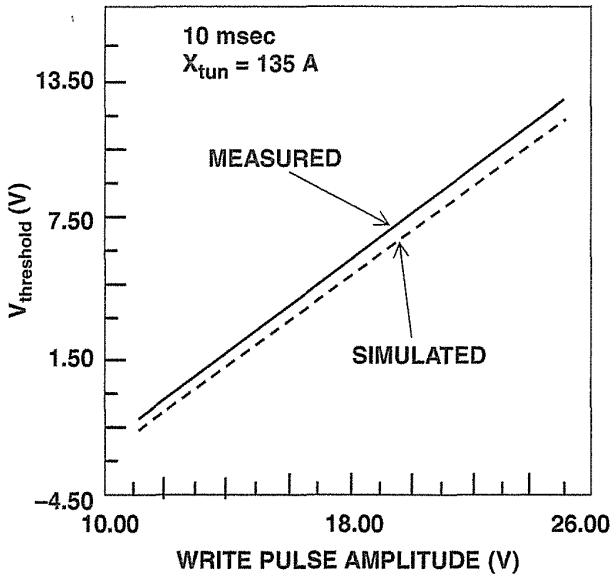


FIGURE 2.20 Measured and simulated threshold voltage as a function of program pulse amplitude, for a fixed program time of 10 ms.

ratios can be increased by reducing C_{tun} and increasing C_{pp}. At a given tunnel-oxide thickness, this is usually achieved by minimizing the thin-oxide area and adding extra poly–poly overlap area on the sides of the cell transistor. Typical coupling ratios are about 0.6.

Figure 2.20 shows the calculated and measured results for the program operation [16]. The threshold voltages as a function of program pulse amplitude are shown. The simulation results fit the measured data closely. And we can see write (program) pulse amplitude and V_t has linear relationship $\Delta V_t = \Delta V_{pgm}$ in same pulse width. This is important in the case of considering ISPP and V_t window setting.

2.2.4 Program Boosting Operation

Program disturb is a phenomenon in which the threshold voltage (V_t) of unselected cell is increased during program operation. The basic program disturb modes are shown in Fig. 2.21. There are two modes of program disturb. One is "V_{pass} mode" in selected NAND string, where bit-line (BL) and cell channel are 0 V. V_{pass} (~10 V) is applied to control gate (wordline), while a 0-V charge is applied to source and drain (channel). This condition is a weak electron injection mode to FG, and then V_t is increased, especially in the case of higher V_{pass}. The other is "boosting mode" in unselected NAND string. The program voltage V_{pgm} is applied to control gate, while channel is the boosting voltage. The boosting voltage (~8 V) is mainly generated by V_{pass} voltage of unselected wordlines in string. The "boosting mode" is also in weak electron injection mode. V_t of a memory cell is increased, especially in the case of high V_{pgm} and lower V_{pass} (lower boosing voltage). As indicated in Fig. 2.21, the V_t

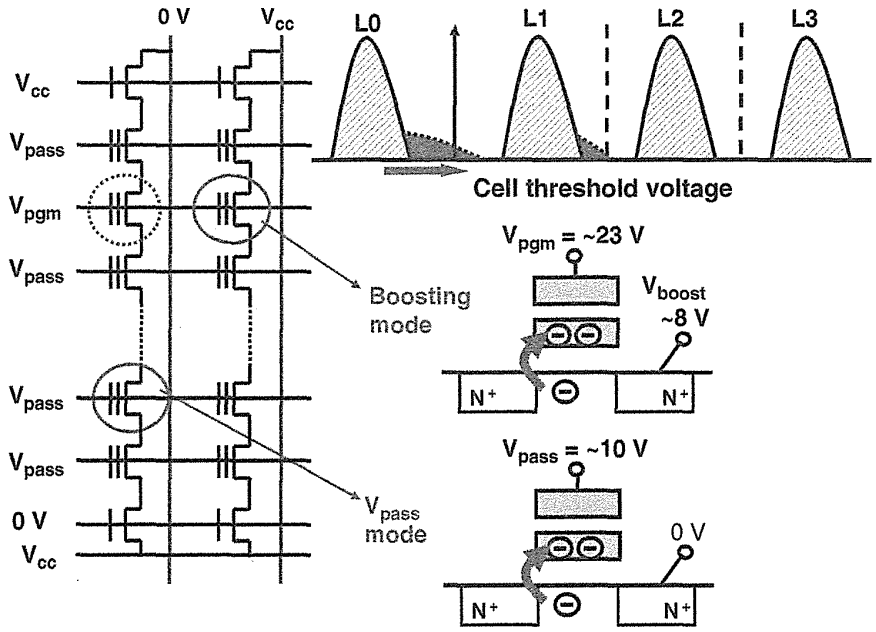


FIGURE 2.21 Program disturb of NAND flash memory cells. Weak electron injection mode is caused in an unselected cell. There are two modes of boosting mode and V_{pass} mode. The "self-boosting" scheme is used for unselected string. The channel voltage in unselected string has boosted up to V_{boost} ~8 V.

distribution tail is mainly increased in L0 or L1 states in MLC V_t setting due to higher tunnel-oxide field.

Figure 2.22 [19] shows program disturb data dependent on V_{pass}. In the case of higher V_{pass} (14–18 V), the V_t shift of a V_{pass} mode is increased. On the other hand, in the case of lower V_{pass} (~10 V), the V_t shift of boosting mode is increased. The middle of V_{pass} voltage (10–14 V) has to be used due to small V_t shift (small program disturb). Normally the range of available V_{pass} voltage is called "V_{pass} window", indeed, the V_{pass} region where V_{th} shifts by disturbs does not cause incorrect operation.

There are several program boosting schemes for multilevel NAND flash cells. Figure 2.23 shows three basic boosting schemes. The first one is a conventional self-boosting (SB) scheme. The SB scheme is mainly used for an SLC (1 bit/cell) device. The second one is a local self-boosting scheme [13]. The boosting voltage (V_{boost}) can be increased because an adjacent WL cell is cut off due to applied 0 V. Then a program disturb of boosting mode can be improved due to reducing voltage difference between CG (V_{pgm}) and channel (V_{boost}). The third one is the erase-area self-boosting scheme (EASB). EASB was widely used for MLC (multibit cell) to obtain a higher boosting voltage. The reason why higher boosting voltage can be obtained is separated erased cells and programmed cells in string. Due to page program order, source-side cells from selected WL have already been programmed. Boosting efficiency is lower in source side cells. However, drain-side cells from selected WL are still in an erased state. Boosting efficiency of drain side cells is higher than source-side due to lower cell V_t. Then it is very effective to obtain high V_{boost} by cutting off

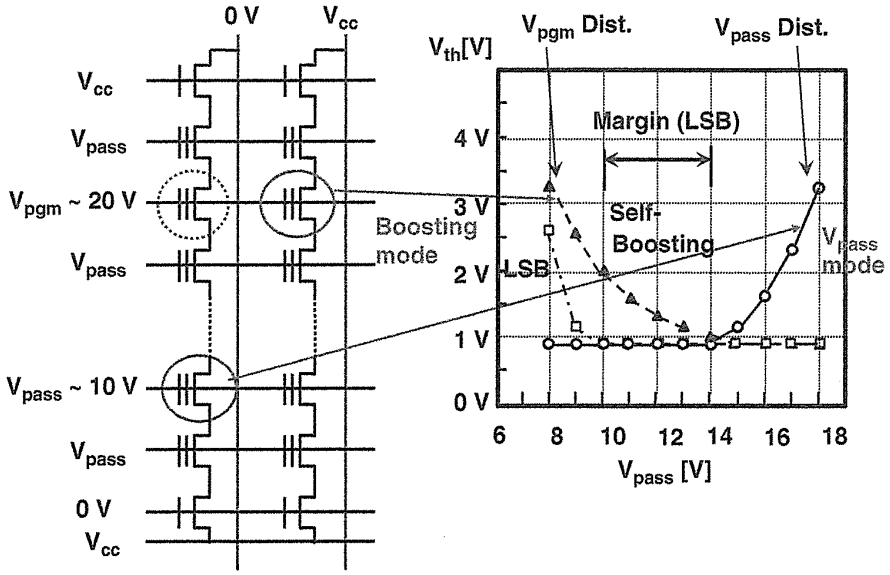


FIGURE 2.22 V_t shift of program disturb in the self-booting scheme. V_{pass} mode dominates in higher V_{pass}. Boosting mode dominates in lower V_{pass} due to low V_{boost}. V_{pass} window is between V_{pass} mode and boosting mode.

boosting node between drain-side and source-side cells. As NAND flash cell scaling, the boosting voltages produce a higher electric field in the source/drain area. A higher electric field generates hot electron and hot hole, and they are unexpectedly injected to FG. In order to relax high electric field, the self-boosting scheme is becoming a more advanced and complicated scheme for each generation of NAND flash memory.

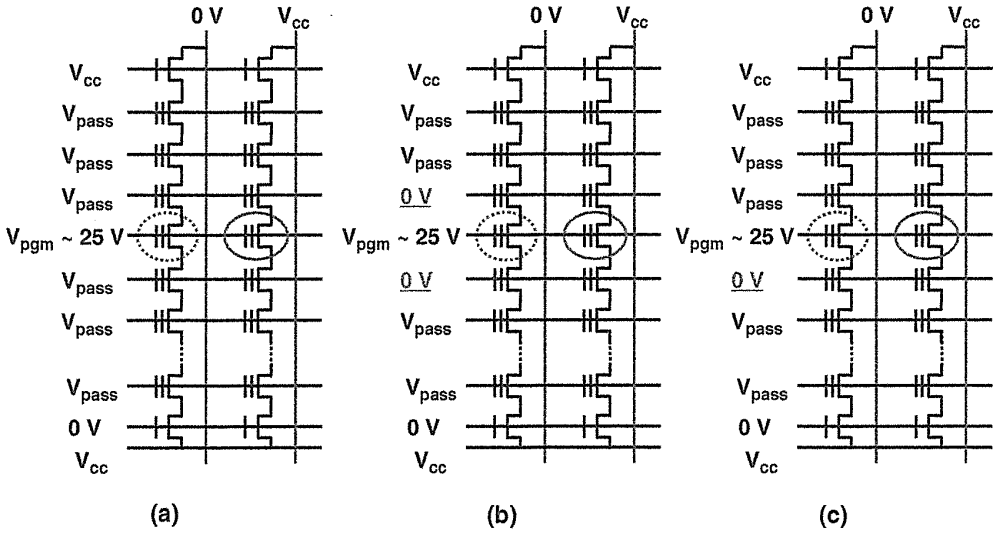


FIGURE 2.23 Self-boosting schemes. (a) Self-boosting scheme (SB). (b) Local self-boosting scheme (LSB). (c) Erase area self-boosting scheme (EASB).

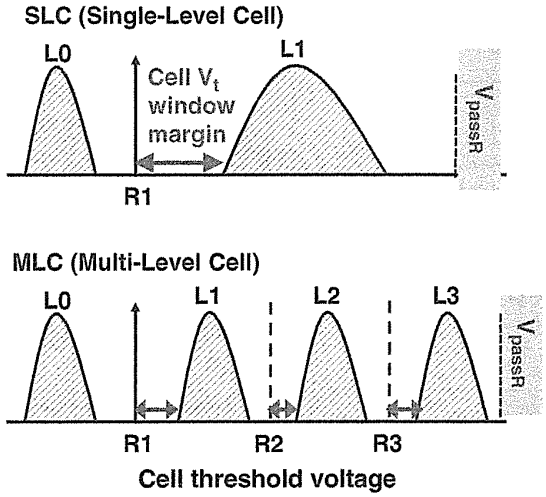


FIGURE 2.24 Cell V_t distribution of an SLC and an MLC.

2.3 MULTILEVEL CELL (MLC)

2.3.1 Cell V_t Setting

Figure 2.24 shows an image of cell V_t distribution setting of an SLC (single-level cell, 1 bit/cell) and MLC (Multilevel Cell, 2 bits/cell). An SLC has a wider cell V_t window margin, and thus SLC has a better program and read performance and also has a better reliability than MLC. As described in Section 2.2.1, in a read operation the unselected cells have to be pass a transistor during reading. Therefore, all cell V_t distributions have to be lower than V_{passR}, as shown in Fig. 2.24. This is one of limitation of cell V_t setting.

Figure 2.25 shows a read V_t window of MLC cells. V_t window is defined by the right side edge of erase V_t distribution and by the left side edge of L3 V_t distribution,

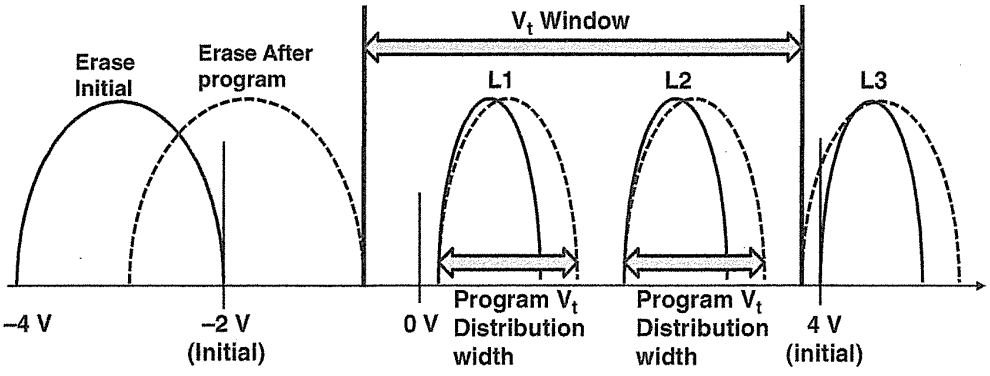


FIGURE 2.25 Read V_t window of an MLC NAND cell.

after programmed all pages in string (or block). Two programmed V_t distributions of L1 and L2 should be on the inside of a V_t window. And a read window margin (RWM) is defined as shown in Fig. 2.25 [20]. By the scaling down of memory cell size, the RWM has degraded due to an increasing impact of many inevitable physical phenomena, as discussed in Section 5.2. Major reliability issues might occur due to a narrow cell V_t window margin.

REFERENCES

[1] Masuoka, F.; Momodomi, M.; Iwata, Y.; Shirota, R. New ultra high density EPROM and flash EEPROM with NAND structure cell, *Electron Devices Meeting, 1987 International*, vol. 33, pp. 552–555, 1987.

[2] Itoh, Y.; Momodomi, M.; Shirota, R.; Iwata, Y.; Nakayama, R.; Kirisawa, R.; Tanaka, T.; Toita, K.; Inoue, S.; Masuoka, F. An experimental 4 Mb CMOS EEPROM with a NAND structured cell, *Solid-State Circuits Conference, 1989. Digest of Technical Papers. 36th ISSCC, 1989 IEEE International*, pp. 134–135, 15–17 Feb. 1989.

[3] Momodomi, M.; Itoh, Y.; Shirota, R.; Iwata, Y.; Nakayama, R.; Kirisawa, R.; Tanaka, T.; Aritome, S.; Endoh, T.; Ohuchi, K.; Masuoka, F. An experimental 4-Mbit CMOS EEPROM with a NAND-structured cell, *Solid-State Circuits, IEEE Journal of*, vol. 24, no. 5, pp. 1238–1243, Oct. 1989.

[4] Momodomi, M.; Iwata, Y.; Tanaka, T.; Itoh, Y.; Shirota, R.; Masuoka, F. A high density NAND EEPROM with block-page programming for microcomputer applications, *Custom Integrated Circuits Conference, 1989, Proceedings of the IEEE 1989*, pp. 10.1/1–10.1/4, 15–18 May 1989.

[5] Iwata, Y.; Momodomi, M.; Tanaka, T.; Oodaira, H.; Itoh, Y.; Nakayama, R.; Kirisawa, R.; Aritome, S.; Endoh, T.; Shirota, R.; Ohuchi, K.; Masuoka, F. A high-density NAND EEPROM with block-page programming for microcomputer applications, *Solid-State Circuits, IEEE Journal of*, vol. 25, no. 2, pp. 417–424, Apr. 1990.

[6] Takeuchi, Y.; Shimizu, K.; Narita, K.; Kamiya, E.; Yaegashi, T.; Amemiya, K.; Aritome, S. A self-aligned STI process integration for low cost and highly reliable 1 Gbit flash memories, *VLSI Technology, 1998. Digest of Technical Papers. 1998 Symposium on*, pp. 102–103, 9–11 June 1998.

[7] Aritome, S.; Shirota, R.; Kirisawa, R.; Endoh, T.; Nakayama, R.; Sakui, K.; Masuoka, F. A reliable bi-polarity write/erase technology in flash EEPROMs, *International electron devices meeting, 1990. IEDM'90. Technical Digest*, pp. 111–114, 1990.

[8] Shirota, R., Itoh, Y., Nakayama, R., Momodomi, M., Inoue, S., Kirisawa, R., Iwata, Y., Chiba, M., Masuoka, F. New NAND cell for ultra high density 5V-only EEPROMs, *Digest of Technical Papers—Symposium on VLSI Technology*, pp. 33–34, 1988.

[9] Momodomi, M.; Kirisawa, R.; Nakayama, R.; Aritome, S.; Endoh, T.; Itoh, Y.; Iwata, Y.; Oodaira, H.; Tanaka, T.; Chiba, M.; Shirota, R.; Masuoka, F. New device technologies for 5 V-only 4 Mb EEPROM with NAND structure cell, *Electron Devices Meeting, 1988. IEDM'88. Technical Digest, International*, pp. 412–415, 1988.

[10] Aritome, S.; Kirisawa, R.; Endoh, T.; Nakayama, R.; Shirota, R.; Sakui, K.; Ohuchi, K.; Masuoka, F. Extended data retention characteristics after more than 10^4 write and erase

cycles in EEPROMs, *International Reliability Physics Symposium, 1990. 28th Annual Proceedings.*, pp. 259–264, 1990.

[11] Kirisawa, R.; Aritome, S.; Nakayama, R.; Endoh, T.; Shirota, R.; Masuoka, F. A NAND structured cell with a new programming technology for highly reliable 5 V-only flash EEPROM, *1990 Symposium on VLSI Technology*. Digest of Technical Papers, pp. 129–130, 1990.

[12] Suh, K.-D.; Suh, B.-H.; Um, Y.-H.; Kim, J.-K.; Choi, Y.-J.; Koh, Y.-N.; Lee, S.-S.; Kwon, S.-C.; Choi, B.-S.; Yum, J.-S.; Choi, J.-H.; Kim, J.-R.; Lim, H.-K. A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme, *Solid-State Circuits Conference, 1995. Digest of Technical Papers. 42nd ISSCC, 1995 IEEE International*, pp. 128–129, 350, 15–17 Feb. 1995.

[13] Suh, K.-D.; Suh, B.-H.; Lim, Y.-H.; Kim, J.-K.; Choi, Y.-J.; Koh, Y.-N.; Lee, S.-S.; Kwon, S.-C.; Choi, B.-S.; Yum, J.-S.; Choi, J.-H.; Kim, J.-R; Lim, H.-K. A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme, *Solid-State Circuits, IEEE Journal of*, vol. 30, no. 11, pp. 1149–1156, Nov. 1995.

[14] Hemink, G.J.; Tanaka, T.; Endoh, T.; Aritome, S.; Shirota, R. Fast and accurate programming method for multi-level NAND EEPROMs, *VLSI Technology, 1995. Digest of Technical Papers. 1995 Symposium on*, pp. 129–130, 6–8 June 1995.

[15] Tanaka, T.; Tanaka, Y.; Nakamura, H.; Sakui, K.; Oodaira, H.; Shirota, R.; Ohuchi, K.; Masuoka, F.; Hara, H. A quick intelligent page-programming architecture and a shielded bitline sensing method for 3 V-only NAND flash memory, *Solid-State Circuits, IEEE Journal of*, vol. 29, no. 11, pp. 1366–1373, Nov. 1994.

[16] Kolodny, A.; Nieh, S.T.K.; Eitan, B.; Shappir, J. Analysis and modeling of floating-gate EEPROM cells, *Electron Devices, IEEE Transactions on*, vol. 33, no. 6, pp. 835–844, June 1986.

[17] Lenzlinger, M.; Snow, E. H. Fowler–Nordheim tunneling into thermally grown SiO, *Journal of Applied Physics*, vol. 40. p. 278, 1969.

[18] Weinberg, Z. A. On tunneling in metal-oxide–silicon structures, *J. Appl. Phys.*, vol. 53, p. 5052, 1982.

[19] Jung, T.-S.; Choi, D.-C.; Cho, S.-H.; Kim, M.-J.; Lee, S.-K.; Choi, B.-S.; Yum, J.-S.; Kim, S.-H.; Lee, D.-G.; Son, J.-C.; Yong, M.-S.; Oh, H.-K.; Jun, S.-B.; Lee, W.-M.; Haq, E.; Suh, K.-D.; Ali, S.B.; Lim, H.-K. A 3.3-V single power supply 16-Mb nonvolatile virtual DRAM using a NAND flash memory technology, *Solid-State Circuits, IEEE Journal of*, vol. 32, no. 11, pp. 1748–1757, Nov. 1997.

[20] Aritome, S.; Kikkawa, T. Scaling challenge of self-aligned STI cell (SA-STI cell) for NAND flash memories, *Solid-State Electronics* vol. 82, pp. 54–62, 2013.

3

NAND FLASH MEMORY DEVICES

3.1 INTRODUCTION

The most important requirement for NAND flash memory [1] is a low bit cost. In order to realize a low bit cost, the scaling down of memory cell size is essential. In this chapter, the NAND flash memory cell and its scaling technologies are discussed.

A NAND flash technology road map and structure scaling of a two-dimensional NAND flash memory cell are shown in Fig. 3.1 and Fig. 3.2, respectively.

Production of NAND flash memory was started in 0.7-μm technology in 1992. The line/space pitch of word line (WL) was ideal 2∗F (F: feature size); however, the line/space pitch of bit line (BL) was 3–4∗F because of limitation of the LOCOS (LOCal Oxidation of Silicon) isolation. Requirements for isolation in NAND flash memory cell are more severe than other devices due to high-voltage operation during programming. Therefore, it was difficult to scale down of LOCOS isolation width beyond 1.5-μm width due to boron diffusion from isolation bottom by LOCOS oxidation process. Then a new FTI process (field through implantation process) was developed [2, 3], as described in Section 3.2. Due to the FTI process, a LOCOS isolation width could be scaled down to 0.8 μm (2∗F of 0.4-μm rule) and the technology node could be scaled down to 0.35-μm technology, as indicated by "1) LOCOS cell scaling" in Fig. 3.1.

Next, the self-aligned shallow trench isolation cell (SA-STI cell) with floating gate (FG) wing had been developed [4, 5], as described in Section 3.3. Due to STI, isolation width could be drastically scaled down to 50% (0.8 μm to 0.4 μm) and then cell size could be scaled down to 67% (bit-line pitch: 1.2 μm to 0.8 μm) in the same design rule,

Nand Flash Memory Technologies, First Edition. Seiichi Aritome.

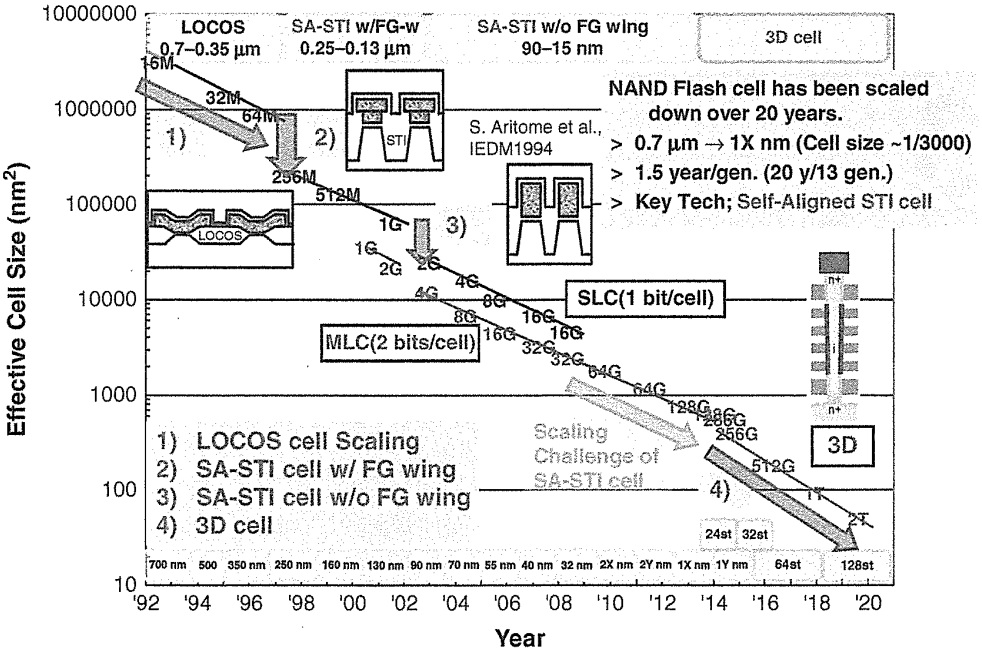


FIGURE 3.1 NAND flash memory technology road map.

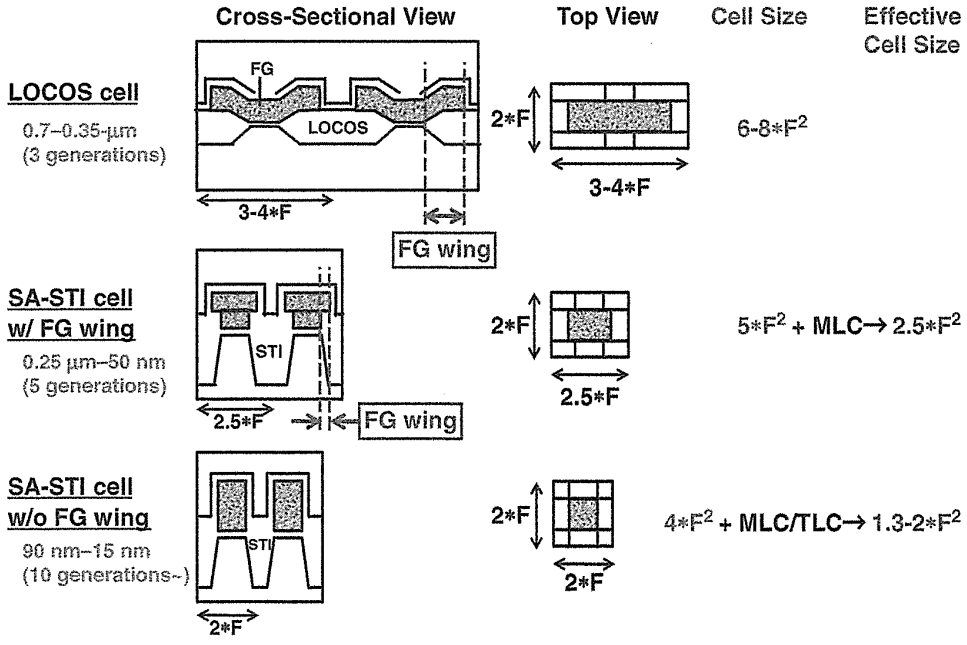


FIGURE 3.2 Structure scaling of two-dimensional NAND flash memory cell.

as indicated by "2) SA-STI cell w/ FG wing" in Fig. 3.1. Also, the isolation capability was much improved because of deep isolation of STI. Furthermore, reliability of tunnel oxide was improved due to no STI edge corner in tunnel oxide.

After that, SA-STI cell without FG wing had been developed [6], as described in Section 3.4. By using this cell structure, a high coupling ratio could be obtained due to large capacitance of IPD using an FG sidewall, as indicated by "3) SA-STI cell w/o FG wing" in Fig. 3.1. The SA-STI cell has a very simple structure, and layout allows for the formation of a very small cell size with bit-line and word-line pitch of 2∗F. Then the cell size becomes ideal $4*F^2$, as shown in Fig. 3.2. The SA-STI technology has also demonstrated an excellent reliability, because the FG does not overlap the STI corner because FG over the tunnel oxide is patterned as the shape of the active area. Therefore, the SA-STI cell realizes very low bit cost and high reliability [4–9]. The SA-STI cell has been extensively used for more than 15 years (since 1998), over 10 generations (0.25 μm to 1Y nm) of NAND flash memory product.

In Section 3.5, the planar FG cell [10–12] is presented. The planar FG cell has a very thin floating gate of around 10-nm thickness and the high-k block dielectric as IPD. The aspect ratio of stacked gate and control gate (CG) fill are much improved. The control gate (CG) fill issue, which is one of the serious problems in the SA-STI cell as described in Section 5.6, can be eliminated by planar FG structure. Also, the planar FG cell has a very small FG capacitive coupling interference due to thin FG. The planar FG cell also started to be used from a 2X-nm technology node in one supplier of NAND flash memory.

In Section 3.6, the side wall transfer-transistor (SWATT) cell [13, 14] is discussed as alternate memory cell technology for a multilevel NAND flash memory cell. By using a SWATT cell, a wide threshold voltage (V_{th}) distribution width could be allowed. The key technology that allows this wide V_{th} distribution width is the transfer transistor which is located at the side wall of the shallow trench isolation (STI) region and is connected in parallel with the floating-gate transistor. During read, the transfer transistors of the unselected cells (connected in series with the selected cell) work as pass transistors. So, even if the V_{th} of the unselected floating-gate transistor is higher than the control gate voltage, the unselected cell can be in the ON state. As a result, the V_{th} distribution of the floating-gate transistor can be wider and the programming speed can be faster because the number of program/verify cycles can be reduced.

Other advanced NAND flash device technologies are presented in Section 3.7.

First, a dummy word-line scheme in NAND flash memory [15–17] is discussed. Dummy word line (dummy cell) is located between edge word lines (edge memory cells) of NAND string and select transistors (GSL or SSL). The program disturb of a GIDL-generated hot electron injection mechanism can be suppressed. In addition, capacitive coupling noise between select gate and edge word line can be reduced. Then the program disturbance failure, read failure, and erase distribution width can be greatly improved by reducing coupling noise. The dummy word-line scheme was started to be used from 40-nm technology node due to stable operations in edge cells.

Second, the p-type doping floating gate (FG) [18–21] is introduced. The p-type FG has an advantage of better cycling endurance and data retention characteristics than an n-type floating gate. The p-type FG started to be used from a 2X-nm technology node due to better reliabilities.

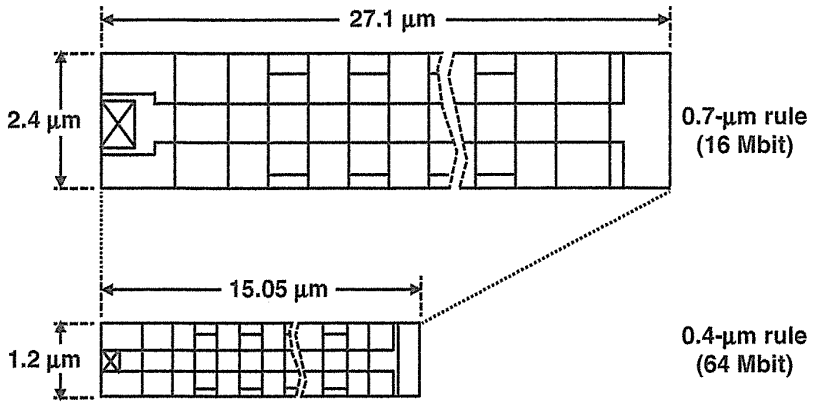


FIGURE 3.3 Top view of a 0.4-μm-rule NAND-structure cell in comparison with that of a 0.7-μm rule cell. This NAND-structured cell has 16 memory transistors arranged in series between two select transistors. Word-line and bit-line pitches are 0.8-μm (L/S = 0.4 μm/0.4 μm) and 1.2 μm, respectively. The cell size is 1.13 μm^2, including the select transistors and the drain contact area. Copyright 1994, The Japan Society of Applied Physics.

3.2 LOCOS CELL

3.2.1 Conventional LOCOS Cell

The mass production of the NAND flash memory [1] started in 1992. A 16-Mega-bit device was first produced by using 0.7-μm technology with a conventional LOCOS isolation process. In order to scale down memory cell size beyond 0.7-μm technology, the scaling down of LOCOS isolation width was a key issue. In a conventional LOCOS process in 0.7-μm generation, an isolation width was 1.7-μm and a bit-line pitch was 2.4 μm, as shown in Fig. 3.3 [2, 3]. It was hard to scale down isolation width beyond 1.5 μm, because of required high-voltage junction breakdown (>8 V) and high inverting voltage (V_t) of parasitic field transistor. Isolation stopper doping of boron is implanted before oxidation of LOCOS isolation. A dopant of boron is easy to diffuse during LOCOS oxidation process. Due to diffusion of boron, it became difficult to satisfy requirements of both a high-voltage junction breakdown (>8 V) and high inverting voltage (V_t) of a parasitic field transistor.

3.2.2 Advanced LOCOS Cell

A small NAND-structure cell (1.13 μm^2 per bit) had been developed in 0.4-μm technology [2, 3]. The chip size of a 64-Mb NAND flash memory using this cell was estimated to be 120 mm^2, which was 60% of a 64-Mb DRAM die size. In order to realize the small cell size, 0.8-μm width field isolation was developed with the field-through implantation (FTI) technique. A negative bias of -0.5 V to the *p*-well of the memory cell is applied during programming. Read disturb could be also ensured for more than 10 years even after 1 million write/erase cycles.

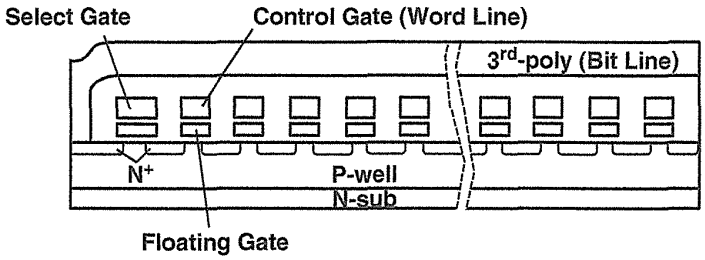


FIGURE 3.4 Cross-sectional view of a 0.4-μm-rule NAND-structure cell. Copyright 1994, The Japan Society of Applied Physics.

A. Scaling of NAND Cell Figure 3.3 compares the top view of the advanced 0.4-μm NAND cell with that of the 0.7-μm conventional one. Figure 3.4 shows the cross-sectional view of the 0.4-μm NAND cell [2, 3]. This NAND-structure cell has 16 memory transistors arranged between two select transistors in series. The word-line pitch is 0.8 μm (line/space = 0.4 μm/0.4 μm). The bit-line pitch could be reduced to 1.2 μm by using 0.8-μm field isolation technology. Figure 3.5 shows the cross-sectional SEM photograph of the 0.4-μm NAND-structure cell after the self-aligned stacked gate etching process. The floating gates are made of first-layer polysilicon (phosphorus-doped). The control gates are made of second-layer polysilicon (phosphorus-doped polysilicon/tungsten-silicide). The process technology is summarized in Table 3.1.

As the design rule of the word line is scaled, it becomes apparent that the NAND cell has an advantage of scaling down the gate length of the memory cell. This is because the NAND cell has punch-through free operation since there is no voltage difference between the drain and source during programming and erasing. And also, the NAND cell has a very small gate–drain overlap region because the impurity concentration of the diffusion layer can be less than 10^{18} cm^{-3} due to no hot electron injection programming and no source erase. As a result, the gate length of the NAND cell can be smaller as feature size is scaled down.

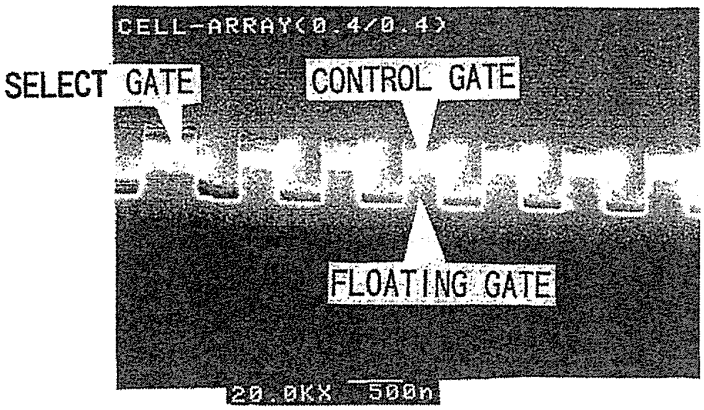


FIGURE 3.5 Cross-sectional SEM photograph of a NAND-structure cell after the self-aligned stacked etching process. Copyright 1994, The Japan Society of Applied Physics.

TABLE 3.1 Process Technology

Process	Double-well CMOS Triple-poly-silicon technology 1 layer-metal	
Cell	Cell size	1.13 μm^2
	W/L	0.4/0.4 μm
	Tunnel oxide	80 A
	Interpoly dielectric	ONO 200 A (effective)
Peripheral	L-poly N-ch	0.8 μm
	P-ch	0.8 μm

Source: Copyright 1994, The Japan Society of Applied Physics.

With respect to scaling in the bit-line direction (bit-line pitch), isolation technology is very important, as discussed in the Section 3.2.3.

B. Operation of NAND Cell The operating conditions are shown in Table 3.2. During writing, 18 V is applied to the selected control gate while the bit lines are grounded; electrons tunnel from the substrate to the floating gate, resulting in a positive threshold voltage shift. If a voltage of 7 V is applied to the bit line (not self-boosting operation), tunneling is inhibited, and the threshold voltage remains the same. The negative bias to the *p*-well is effective in preventing the parasitic field transistor from turning on. During erasing, 20 V is applied to both the *p*-well and the (*N*-type) substrate while keeping the bit lines floating and all the selected control gates grounded. Electrons tunnel from the floating gate to the substrate, and the threshold voltage of the memory cells becomes negative.

The reading method is also shown in Table 3.2. Zero volt is applied to the gate of the selected memory cell, while 3.3 V (V_{cc}) is applied to the gates of the other cells. Therefore, all of the other memory transistors, except for the selected transistor, serve as transfer gates. A cell current flows if a selected memory transistor is in the depletion mode. On the other hand, cell current does not flow if the memory cell is programmed to be in the enhancement mode.

TABLE 3.2 Operation Conditions

	Write 1	Erase	Read
Bit line	0/7 V	Open	1.5 V
Select gate	7 V	20 V	3.3 V
Control gate 1	7 V	0 V	3.3 V
Control gate 2	18 V(selected)	0 V	0 V
Control gate 16	7 V	0 V	3.3 V
Select gate	0 V	20 V	3.3 V
Source	0 V	Open	0 V
p-well	−0.5 V	20 V	0 V
n-sub	3.3 V	20 V	3.3 V

Source: Copyright 1994, The Japan Society of Applied Physics.

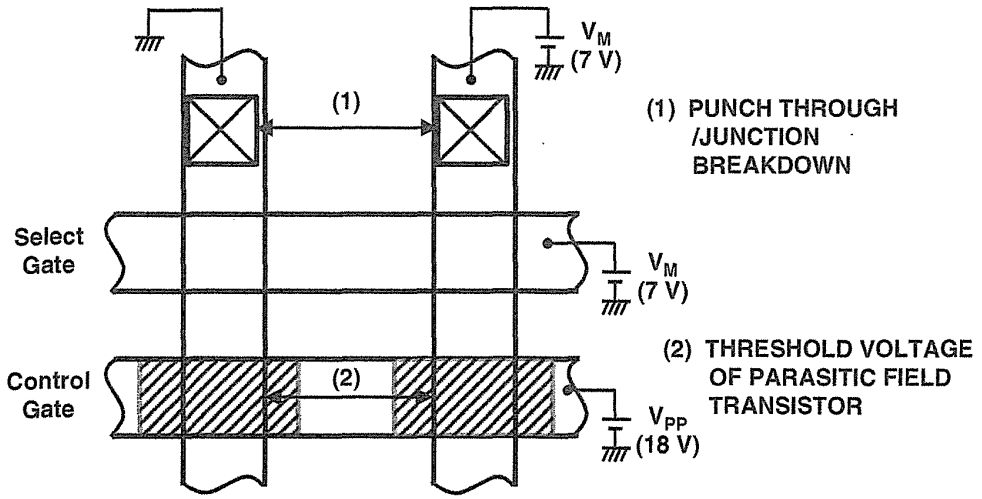


FIGURE 3.6 Isolation between two neighboring bit lines: (1) Punch-through or junction breakdown in the bit-line contact area. (2) The threshold voltage of the parasitic field transistor. Copyright 1994, The Japan Society of Applied Physics.

3.2.3 Isolation Technology

For the NAND cell, high-voltage field isolation technology is important to reduce the bit-line pitch. The isolation between the bit lines must satisfy two requirements, as shown in Fig. 3.6 [2, 3]. One is the punch-through or the junction breakdown voltage of the bit-line junction area. During programming, 7 V is applied to the bit line to prevent electron injection in cells that should remain in the erased state. Zero volt is applied to a bit line which is connected to a cell that should be programmed. The punch-through voltage between neighboring bit-line junctions must be higher than 7 V, as must the bit-line junction breakdown voltage. Another requirement is a high threshold voltage of the parasitic field transistor. During programming, the selected control gate is biased with a high voltage of 18 V, which may easily turn on the field transistor between neighboring bits (Table 3.2).

In order to avoid bit-line junction breakdown/punch-through and to prevent the field transistor from turning on, the field-through implantation process (FTI process) and *p*-well negative bias method have been developed [2, 3], as shown in Fig. 3.7. In the FTI process, the boron ions (160 keV, 1.13 cm^{-2}) are implanted to form a field stopper after LOCOS fabrication. The field-oxide thickness at field-through implantation is 420 nm. The punch-through voltage and the threshold voltage of the parasitic transistor increase without decreasing the junction breakdown voltage, because the lateral diffusion of the boron stopper impurity is decreased in comparison with the conventional field stopper implantation before LOCOS fabrication. A negative *p*-well bias prevents punch-through and increases the threshold voltage of the parasitic field transistor.

Figure 3.8 shows the breakdown voltage between two neighboring bit lines as a function of the bit-line contact distance. The breakdown voltage is ensured to be

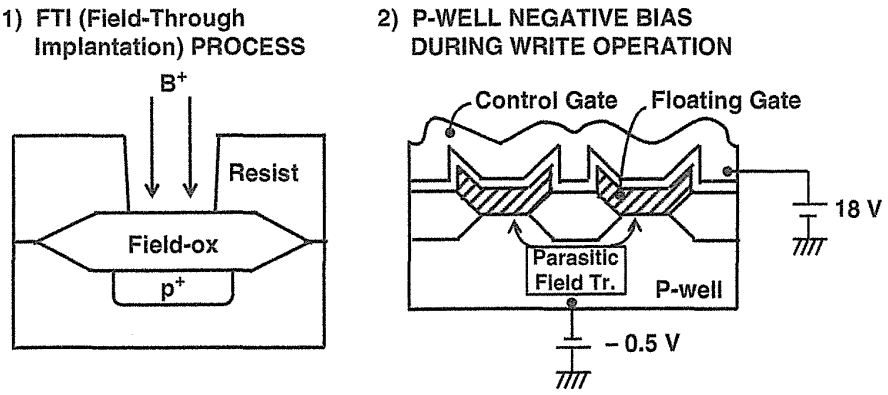


FIGURE 3.7 The 0.8-μm-wide high-voltage field isolation technology. Copyright 1994, The Japan Society of Applied Physics.

higher than 7 V, in the case of a negative-biased *p*-well. Figure 3.9 shows the threshold voltage of the parasitic field transistor. By using the FTI process, the threshold voltage of the 0.8-μm field transistor becomes more than 28 V. Moreover, a negative bias of −0.5 V is applied to the *p*-well to increase the threshold voltage of the parasitic field transistor (V_{tf}). In comparison with a zero-biased *p*-well, V_{tf} is increased to more than 30 V. Therefore, the field width margin is increased from 0.05 μm to 0.15 μm. As a result, a very small bit-line pitch of 1.2 μm could be realized.

The FTI process also reduces the body effect of the cell transistors, because the boron impurity concentration under the channel region of the cell transistors and the

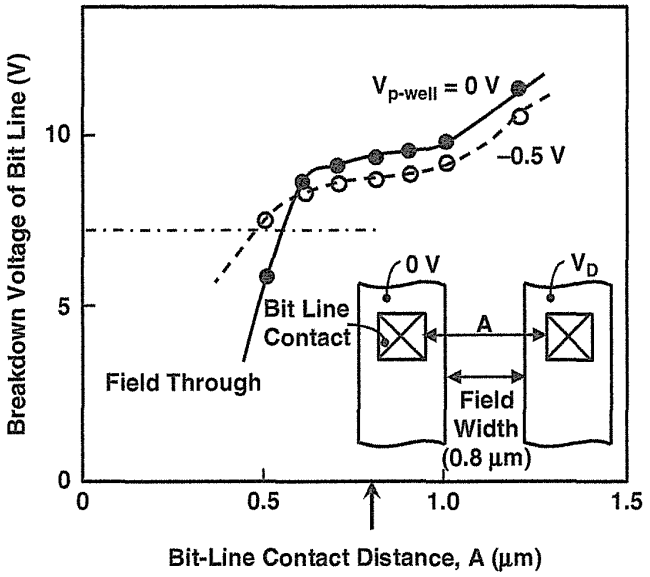


FIGURE 3.8 The breakdown voltage of the bit-line junction. The punch-through voltage and junction breakdown voltage are higher than 7 V at a 0.8-μm bit-line contact distance. Copyright 1994, The Japan Society of Applied Physics.

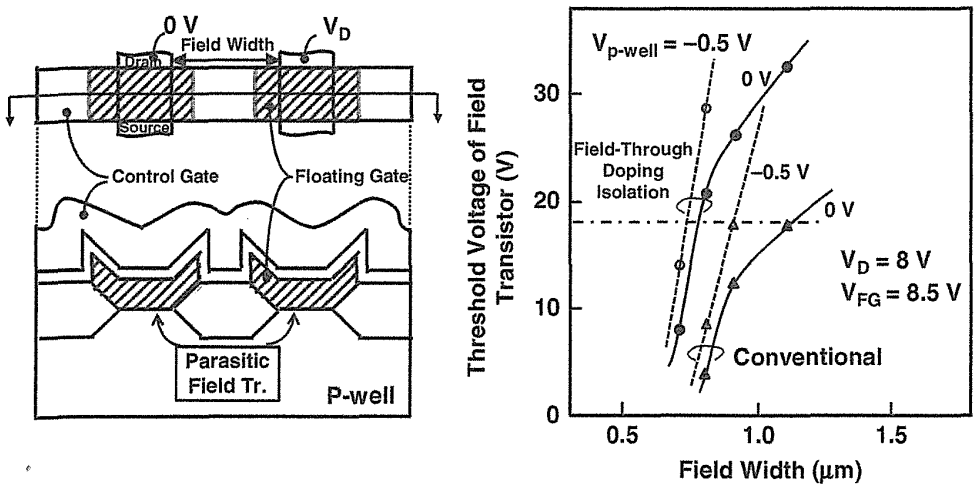


FIGURE 3.9 The threshold voltage of the parasitic field transistor as a function of the field width. By using a field-through implantation process (FTI process), the threshold voltage of a 0.8-μm field transistor becomes higher than 28 V when −0.5 V is applied to the *p*-well. Copyright 1994, The Japan Society of Applied Physics.

select transistor become lower. The decrease of the body effect results in an increased cell current. Figure 3.10 shows the cell current of a NAND cell as a function of the cell gate length. The cell currents were measured at the nearest cell from the bit-line contact side, which has the smallest cell current among all cells because that cell should be strongly influenced by the body effect due to the series resistance of the other

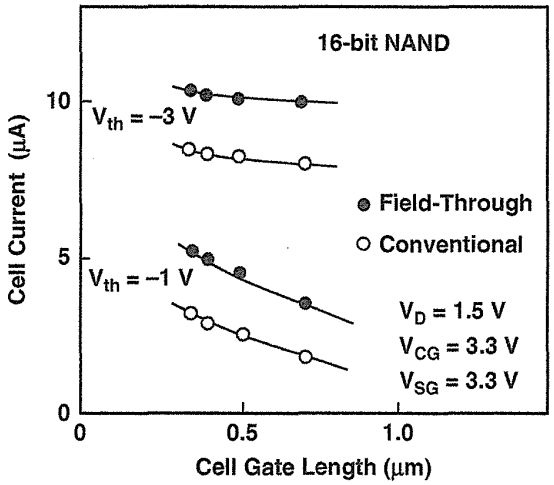


FIGURE 3.10 The cell current during a read operation using both the field-through implantation (FTI) process and the conventional process. The cell current of the FTI process is larger than that of the conventional process due to the suppression of boron diffusion to the channel region. Copyright 1994, The Japan Society of Applied Physics.

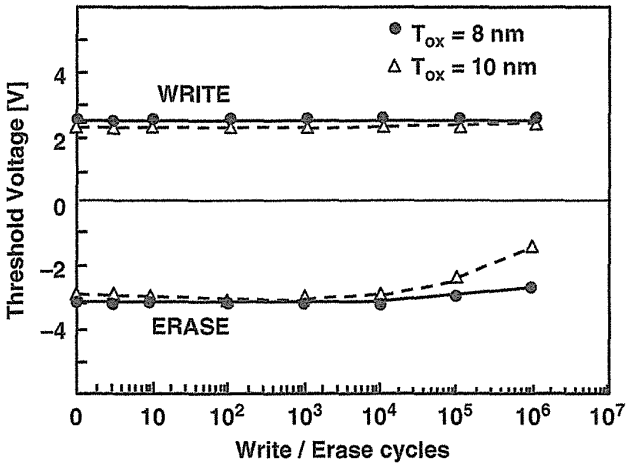


FIGURE 3.11 Program (write) and erase cycling endurance characteristics. The threshold voltage is defined as the control gate voltage which flows a drain current of 1 μA with a drain-source voltage of 1.5 V. For an 8-nm tunnel oxide, we have the following: Program (write): $V_{cg} = 18$ V, 0.1 ms. Erase: $V_{p\text{-well}} = 20$ V, 1 ms. For a 10-nm tunnel oxide, we have the following: Program (write): $V_{cg} = 20.4$ V, 0.1 ms. Erase: $V_{p\text{-well}} = 22.7$ V, 1 ms. Window narrowing is almost not observed up to 1 million program (write)/erase cycles. Copyright 1994, The Japan Society of Applied Physics.

cells. The threshold voltage of the selected cell is −1 V and −3 V, the threshold voltage of the unselected cells, which, connected in series, is from 0.5 to 1.5 V. In the case of the FTI process, the cell current is larger than the current in the conventional case.

3.2.4 Reliability

Figure 3.11 shows the program (write) and erase cycling endurance characteristics of a NAND cell with 8- and 10-nm-thick tunnel oxide [2, 3] using the bipolarity uniform program/erase scheme [22–25]. This scheme guarantees a wide cell threshold window of as large as 4 V, even after 1 million write/erase cycles. In the 8-nm tunnel-oxide cell, window narrowing can be hardly seen due to the small number of electron traps in the 8-nm tunnel oxide.

Read disturb occurs as a weak programming mode. When a certain positive voltage is applied to the control gate during read operation, a small Fowler–Nordheim tunneling current flows from the substrate to the floating gate. Unfortunately, the tunnel-oxide leakage currents, which are induced by the program and erase cycling stress, degrade the read disturb of the memory cell, as shown in Fig. 3.12. In order to suppress the read disturb, the applied gate voltage must be lowered. A reduction of the gate voltage from 5 V to 3.3 V allows the downscaling of the tunnel oxide from 10 nm to 8 nm (Fig. 3.13). Sensing at 3.3 V can be performed by a bit-by-bit verified programming scheme [26], which results in a written cell threshold voltage between 0.5 V and 1.8 V.

Even for an 8-nm tunnel-oxide thickness, read disturb suppression can be ensured for more than 10 years even after 10^6 W/E cycles, as presented in Fig. 3.13. By scaling

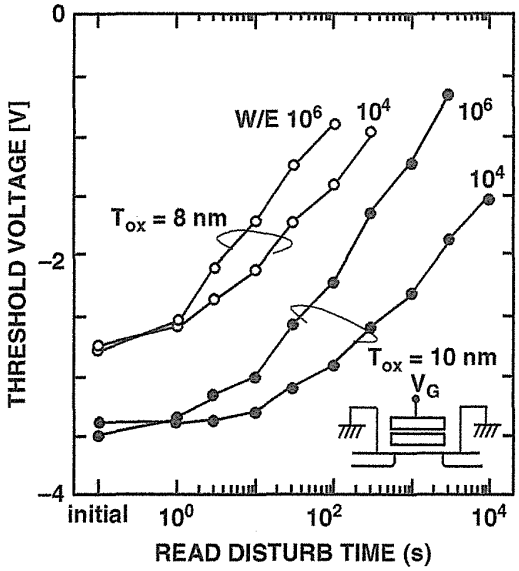


FIGURE 3.12 The read disturb characteristics of a NAND flash cell with 8- and 10-nm tunnel-oxide thickness. The voltage of the control gate (V_G) is 9 V as an accelerated condition. The threshold voltage is measured under the same condition as in Fig. 3.11. Copyright 1994, The Japan Society of Applied Physics.

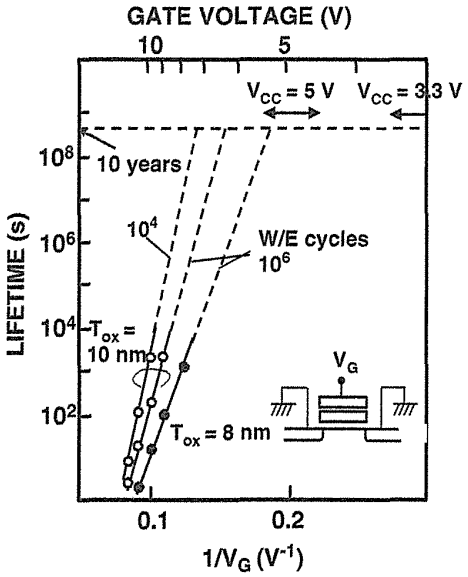


FIGURE 3.13 The read disturb lifetime is defined as the time at which V_{th} reaches −1.0 V during the applied gate voltage stress. Even if 8-nm tunnel oxide thickness is used, the read disturb time is far more than 10 years when a +3.3/−0.3 V supplied voltage (V_{cc}) is used. Copyright 1994, The Japan Society of Applied Physics.

down the tunnel-oxide thickness from 10 nm to 8 nm, the program voltage can be reduced from 21 V to 18 V, which allows the design of more compact peripheral circuits such as row decoders and sense amplifiers.

A 1.13-μm^2 memory cell for a 64-Mbit NAND flash memory had been successfully developed using 0.4-μm technology. High-voltage field isolation technology realizes a very small bit-line pitch of 1.2-μm. The tunnel-oxide thickness can be scaled down from 10 nm to 8 nm, and 3.3-V operation is possible using a bit-by-bit verify programming method. This technology was suitable for realizing a low-cost and highly reliable memory chip.

3.3 SELF-ALIGNED STI CELL (SA-STI CELL) WITH FG WING

A high-density $5*F^2$ (F: feature size) self-aligned shallow trench isolation cell (SA-STI cell) technology is described to realize low-cost and high-reliability NAND flash memories [4, 5]. The extremely small cell size of 0.31 μm^2 has been obtained for the 0.25 μm design rule. To minimize the cell size, a floating gate is isolated with a shallow trench isolation (STI) and a slit formation by a novel SiN spacer process, which has made it possible to realize a 0.55 μm-pitch isolation at a 0.25 μm design rule. Another structural feature to the cell and its small size is the borderless bit-line and source-line contacts which are self-aligned with the select gate. The proposed NAND cell with the gate length of 0.2 μm and the isolation space of 0.25 μm shows a normal operation as a transistor without any punch-through. A tight distribution of the threshold voltages (2.0 V) in 2-Mbit memory cell array is achieved due to a good uniformity of the channel width in the SA-STI cells. Also, the peripheral low-voltage CMOS transistors and high-voltage transistors can be fabricated at the same time by using the self-aligned STI process. The advantages are as follows: (1) The number of process steps is reduced to 60% in comparison with a conventional process, and (2) high reliability of the gate oxide is realized even at high-voltage transistors because a gate electrode does not overlap the trench corner. Therefore this SA-STI process integration combines a small cell size (a low cost) with a high reliability for a manufacturable 256-Mbit and l-Gbit flash memory.

3.3.1 Structure of SA-STI Cell

This section describes a novel high density $5*F^2$ (F: feature size) NAND STI cell technology [4] and peripheral transistors devices [5] which have been developed for a low bit-cost flash memories. The three key technologies to minimize the cell size have been introduced, as shown in Fig. 3.14. The self-aligned shallow trench isolation cell (SA-STI cell) has a high coupling ratio with a thick floating gate [6]. However, its high aspect ratio of the gate space has made it difficult to control the planarization process of the trench isolation by chemical mechanical polishing (CMP). To overcome this problem, a stacked floating gate structure is applied. A first thin poly-Si gate is self-aligned with the active area of the cells to control the

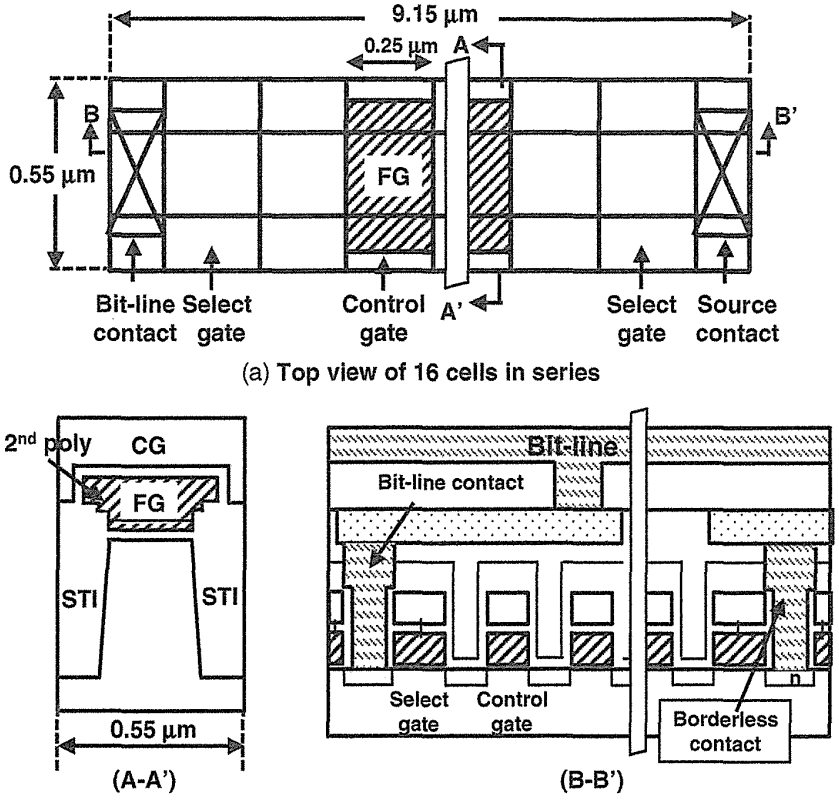


FIGURE 3.14 Schematic view of the SA-STI cell with floating-gate (FG) wing. (a) Top view of 16 cells in series. (b) Cross-sectional views of A–A′ and B–B′ in (a). Three key technologies to achieve $5*F^2$ (F: feature size) cell size has been introduced. (1) A stacked floating-gate structure has been applied in order to reduce the gate space aspect. A first poly-Si gate is self-aligned with the active area. A second poly-Si gate is formed over the exposed first gate to achieve a high coupling ratio (>0.6) with floating-gate wing. (2) The second poly-Si gate has been patterned with spacing of 0.15 μm by a novel SiN spacer process. This process has made it possible to realize 0.55-μm-pitch isolation. (3) The borderless bit-line and source-line contacts which are self-aligned with the select gate can eliminate a space between the contacts and the gate.

channel width precisely. A global planarization by CMP process is very controllable due to the reduction of the gate space aspect. A second poly-Si gate is formed on the first poly-Si gate to achieve a high coupling ratio (>0.6) of the cells. The second poly-Si gate is patterned with spacing of 0.15 μm by a novel SiN spacer process. This process has made it possible to realize 0.55-μm-pitch isolation. Another feature of integration to the cell and its small size is the borderless bit-line and source-line contacts which are self-aligned with the select gate. By the above technologies, an extremely small cell size of 0.31 μm^2 has been obtained for the 0.25 μm design rule.

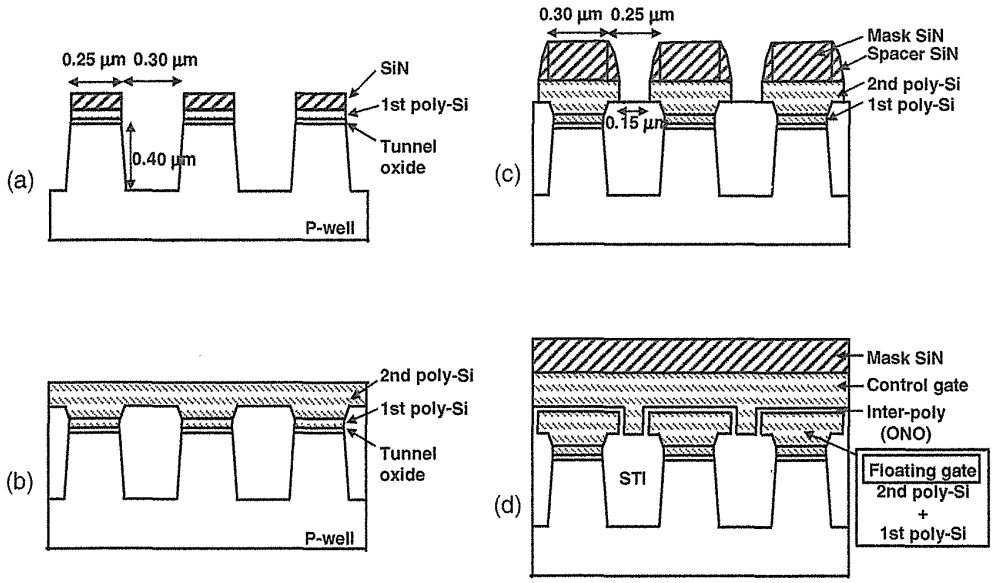


FIGURE 3.15 Process flow of the SA-STI cell with FG wing. (a) Trench etching. (b) LPCVD SiO_2 fill-in and planarization by CMP, second poly-Si gate deposited. (c) Floating-gate formation by SiN spacer process. (d) ONO and the control-gate formation.

3.3.2 Fabrication Process Flow

The process flow of the SA-STI cell with FG wing is described in Fig. 3.15. [4]. The active area is isolated by the STI formation using a self-aligned mask of a first thin poly-Si gate (a). After CVD SiO_2 deposition and planarization by CMP, the second poly-Si gate is deposited on the exposed first poly-Si layer, resulting in the stacked floating gate structure (b). The second poly-Si layer is striped with spacing of 0.15 μm, which is less than a design rule by a novel SiN spacer process as follows. A SiN mask is patterned at spacing of 0.25 μm, and a 50-nm-thick spacer SiN is then deposited. By etching the SiN mask back, a stripe mask pattern with 0.15-μm space is obtained (c). After removal of the SiN mask, an inter-poly dielectric (ONO) and the control gate are successively deposited (d). The control gate and the floating gate are continuously patterned, followed by deposition of a barrier SiN layer and an interlayer. The SiN layer covering the control gate prevents a short circuit between the gate and the borderless contacts. Finally, a doped poly-Si is filled within the bit-line contact and source-line contact and is etched back, followed by the metallization. Figures 3.16 and Fig. 3.17 show the cross-sectional SEM micrographs of the $5*F^2$ memory cell array with a cell size of 0.31 μm^2 using a 0.25-μm design rule. The key process parameters are listed in Table 3.3.

Figure 3.18 shows the schematic view of the SA-STI cell and peripheral transistors. The comparison between a novel process and a conventional process is schematically shown in Fig. 3.19. The number of fabrication steps of the process is reduced to about 60%. The memory cells and the peripheral transistors can be formed simultaneously without any additional process steps.

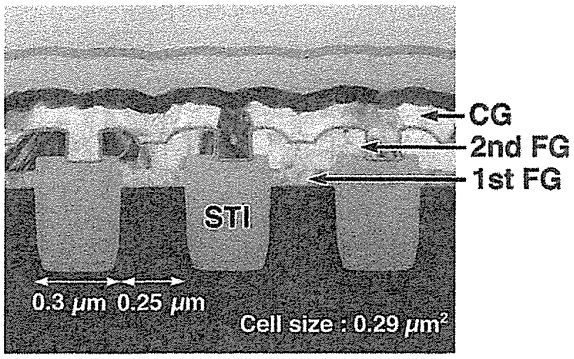


FIGURE 3.16 Cross-sectional SEM micrographs of the cell, parallel to the control gate.

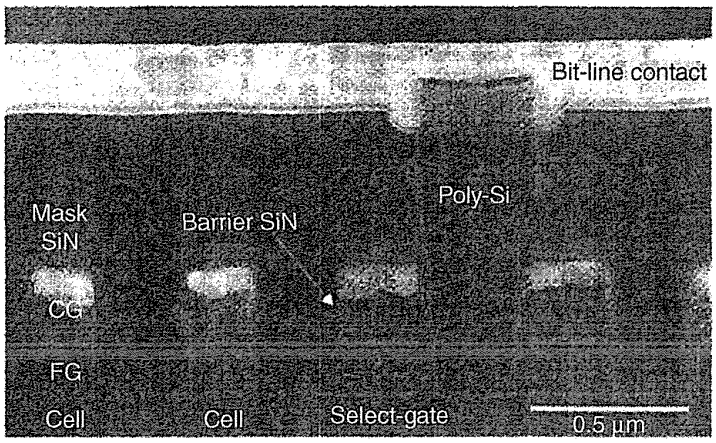


FIGURE 3.17 Cross-sectional SEM micrographs of the cell, parallel to the bit line.

TABLE 3.3 The Main Device Parameters

Technology	0.25-μm Double-well CMOS Self-aligned shallow trench isolation	
Cell	Tunnel oxide	9.0 nm
	Gate length	0.25 μm
	Channel width	0.25 μm
	Cell size	0.31 μm^2
Peripheral	High-Voltage gate oxide	40.0 nm
	Low-Voltage gate oxide	9.0 nm
	NMOS effective gate length	0.28 μm
	PMOS effective gate length	0.38 μm

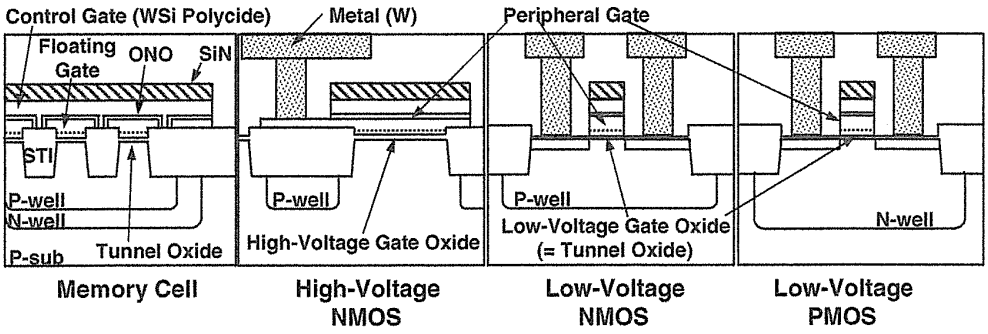


FIGURE 3.18 The schematic cross-sectional view of the SA-STI NAND flash memory.

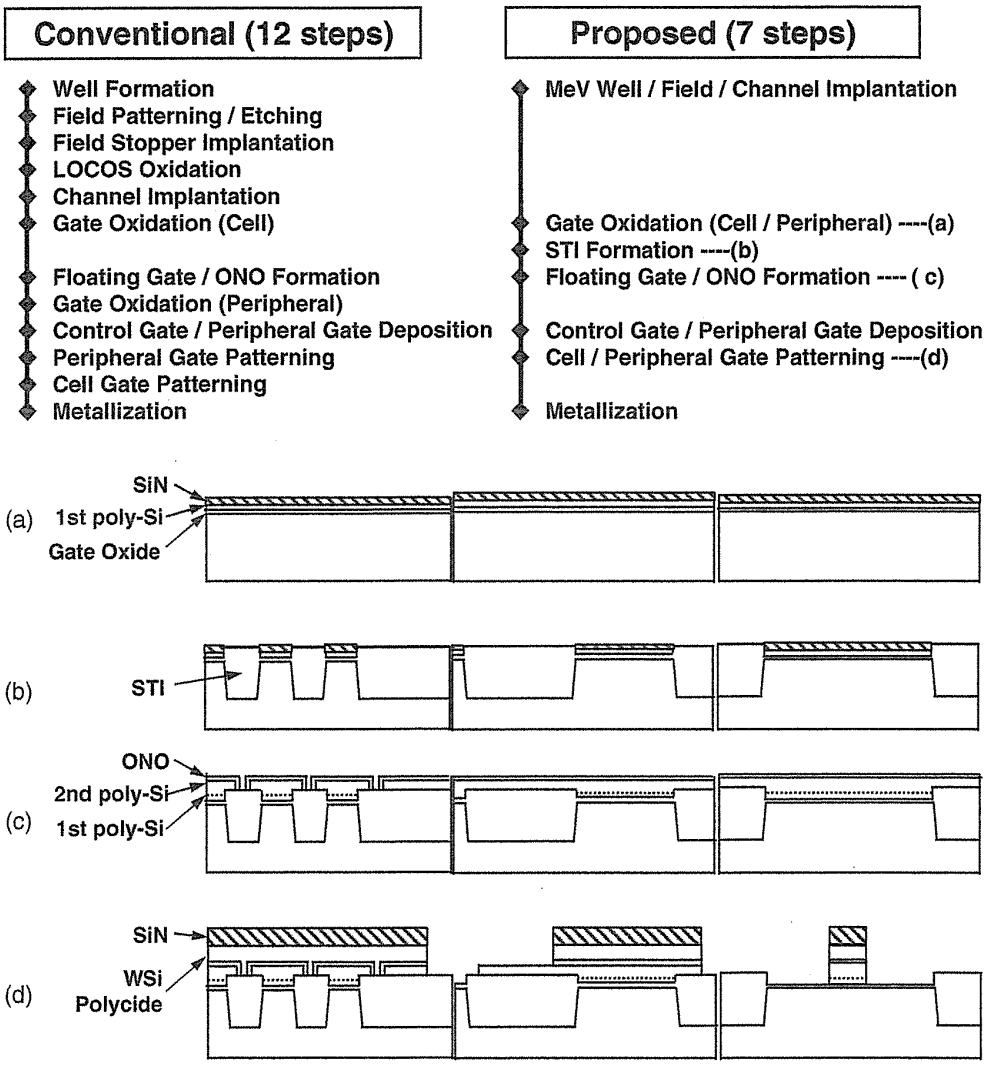


FIGURE 3.19 The process sequence of the SA-STI NAND flash memory.

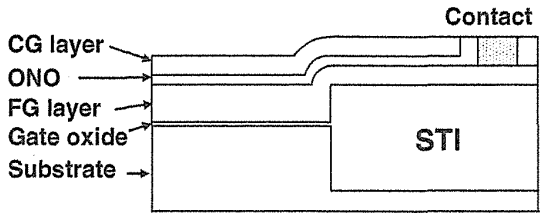


FIGURE 3.20 Cross-sectional view of peripheral transistor. The gate electrode (FG layer) is self-aligned with STI.

First, a retrograde well profile is formed by a high-energy ion implantation. Each implantation is carried out for a well formation, a field punch-through stopper, and a channel threshold adjustment. Next, 40-nm-thick gate oxides for the high-voltage transistors and 9-nm-thick gate oxides for the cells and the low-voltage transistors are formed, and then the first poly-Si layer for the floating gate and a SiN layer are successively deposited, as shown in Fig. 3.19a. The shallow trench, which is self-aligned with the first poly-Si layer, is etched, followed by SiO_2 filled-in planarization by CMP, as shown in Fig. 3.19b. After the SiN mask removal and second poly-Si layer deposition, the second poly-Si layers are patterned with a 0.15-μm space by a SiN spacer process. The stacked poly-Si structure acts as the floating gate of the cells and the gate electrode of the peripheral transistors. Then, ONO of an inter-poly dielectric is deposited, as shown in Fig. 3.19c.

Next, a WSi polycide layer for the control gate of the cell is deposited, and then the polycide layer, ONO, and the stacked poly-Si layer are continuously patterned, as shown in Fig. 3.19d. The peripheral gate electrode is also patterned with the cell. The polycide layer in the peripheral transistors is partially removed for a gate contact formation.

Finally, interconnections and peripheral contacts are formed by a dual damascene process. Figure 3.20. shows the cross-sectional view of the peripheral transistor [8]. The main device parameters of peripheral devices are also summarized in Table 3.3.

3.3.3 Characteristics of SA-STI with FG Wing Cell

Figure 3.21 shows a bit-line junction breakdown characteristics as a function of the isolation width. There occurs no field punch-through between the bit-line contacts at the STI width of up to 0.25 μm with an implantation of boron for the field stopper. Moreover, the 0.4-μm-thick STI field oxide results in a high threshold voltage (>30 V) of the parasitic field transistor between the neighboring bits. Figure 3.22. shows that the threshold voltage of the cell transistors shows a weak dependence on the channel width because no boron atoms implanted for the field stopper diffuses into the channel region from the trench bottom. From these results, the STI cell is very suitable for scaling of the isolation pitch.

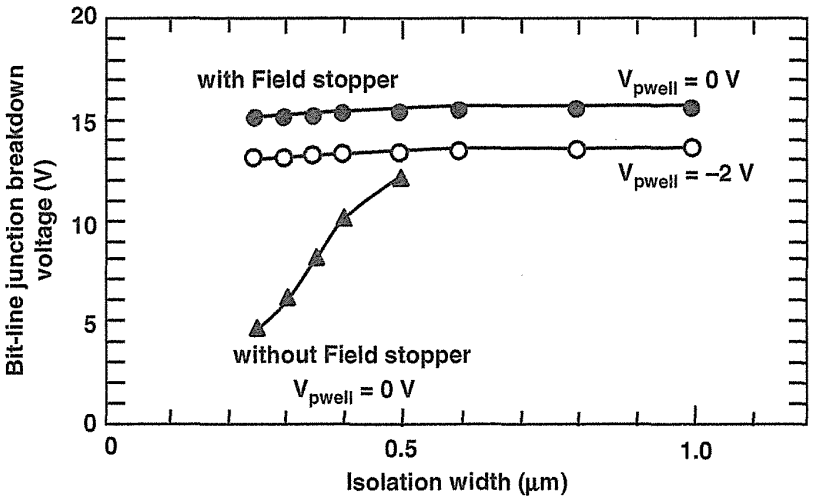


FIGURE 3.21 A breakdown voltage of the bit-line junctions, which is isolated by STI. There is no punch-through at less than 15 V, which is the junction breakdown voltage using boron field stopper implantation.

Figure 3.23 shows I_d–V_g characteristics of the SA-STI cell transistors. No anomalous hump is observed in the subthreshold characteristics of the wide channel transistors with W = 10 μm since the floating gate never overlaps the STI corners. In the case of the NAND cell transistors, the maximum drain voltage of around 1 V (less than V_{cc}) is applied only in the read operation. Figure 3.24 shows the I_d–V_g characteristics of the cell transistors with various gate lengths at $V_d = V_{cc}$ (2.5 V).

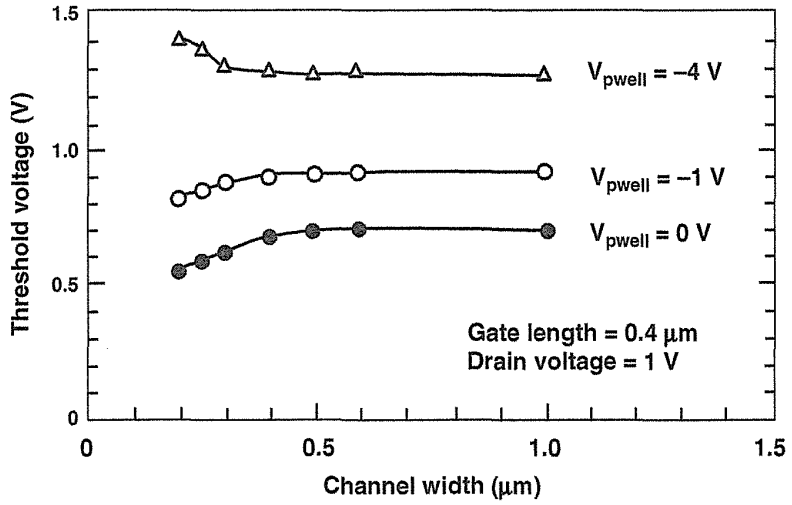


FIGURE 3.22 A threshold voltage of the SA-STI cell as a function of the channel width. The SA-STI cells show a weak dependence of the threshold voltage on the channel width. Therefore, the STI cell is suitable for scaling of the isolation pitch.

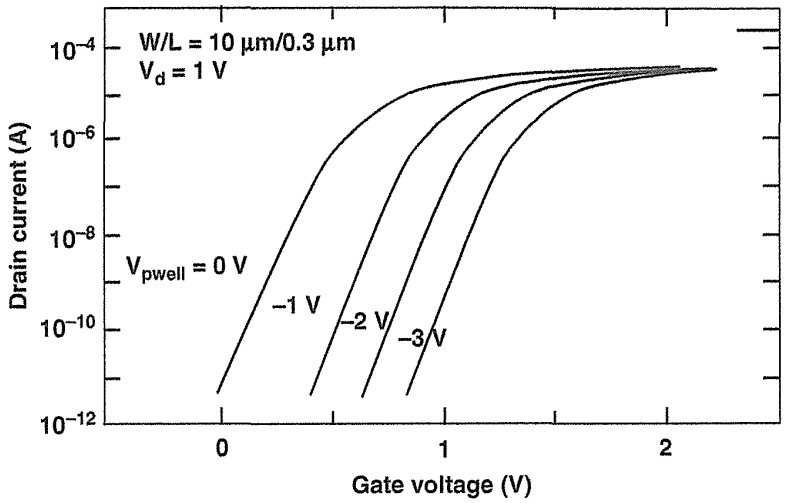


FIGURE 3.23 I_d–V_g characteristics of the wide channel width in cell transistors with various voltages of p-well. No anomalous hump is seen in the subthreshold characteristics since the floating gate does not overlap the STI comers.

There is a sufficient margin at the gate length of 0.2 μm for device operation. These results enable a 0.2-μm-rule SA-STI cell with the cell area of 0.31 μm^2. In the SA-STI cells, Fowler–Nordheim tunneling can achieve a fast programming (20 μs) by applying 17 V to the control gate and achieve a fast erasing (2 ms) by applying 18 V to a p-well, as shown in Fig. 3.25. Figure 3.26 shows the TDDB characteristics of the tunnel oxides. Since Q_{BD} in the stripe capacitors with trench edges is almost the

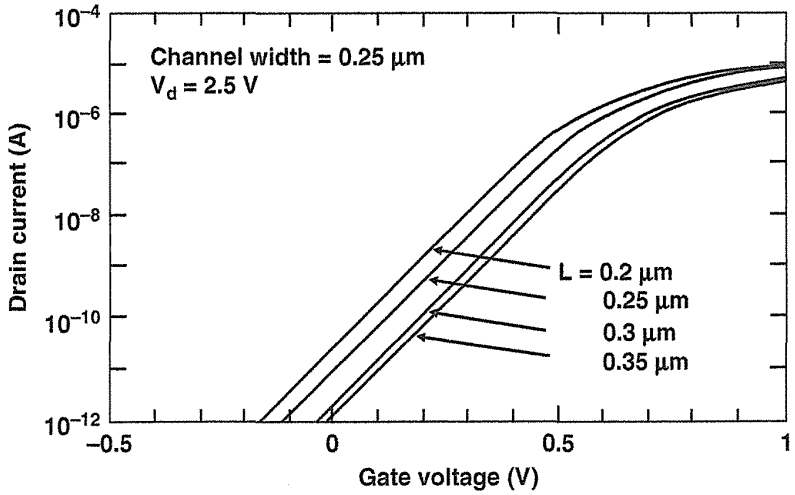


FIGURE 3.24 I_d–V_g characteristics of the short channel cell transistors at V_d = 2.5 V, which is applied in the read operation. There is a sufficient margin at the gate length of 0.2 μm for device operation.

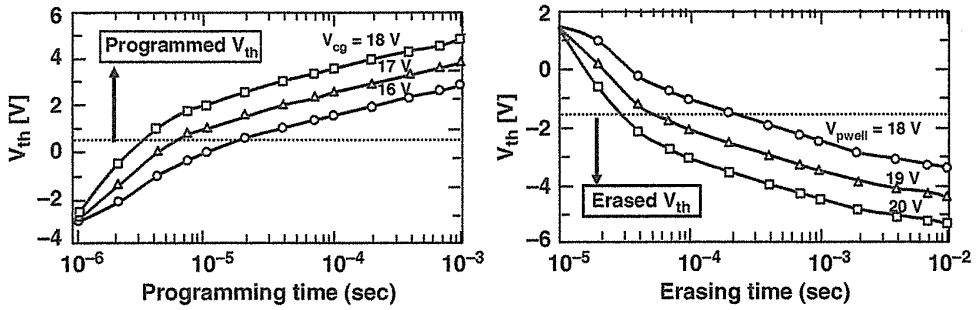


FIGURE 3.25 Programming and erasing characteristics of the SA-STI cells. Fast programming (20 μs) and erasing (2 ms) can be accomplished by Fowler–Nordheim tunneling, applying 17 V to the control gate during programming and 18 V to *p*-well during erasing, respectively.

same as that in the flat capacitors without trench edges, the process damages into the tunnel oxides during the SA-STI fabrication steps are negligibly small. Therefore, the endurance characteristics of the SA-STI cells are excellent, as shown in Fig. 3.27. The threshold voltage window narrowing has not been observed up to 1 million cycles.

A threshold voltage distribution of programmed and erased cells is evaluated by measuring a 2-Mbit cell array, as shown in Fig. 3.28. Both the programming and the erasing are performed by Fowler–Nordheim tunneling of electrons. A tight distribution of about 2.0 V is realized though the programming and the erasing are carried out by one pulse without verification, because of a good uniformity of the channel width in the memory cell by using the self-aligned STI structure.

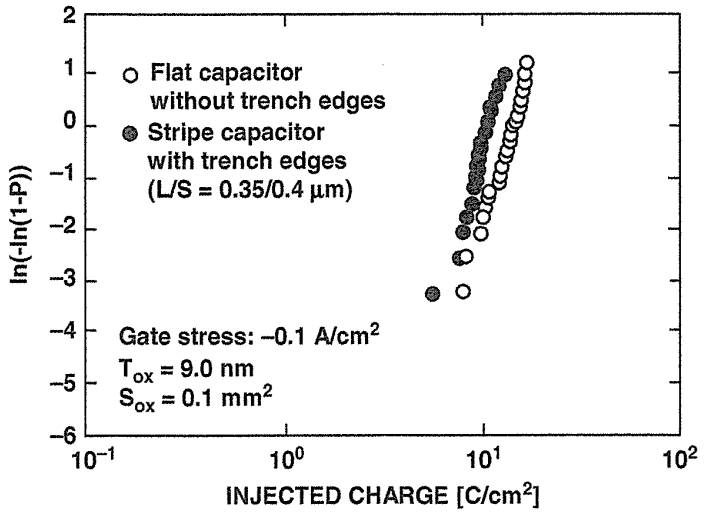


FIGURE 3.26 TDDB (Time Dependent Dielectric Breakdown) characteristics of the tunnel oxide in the STI stripe capacitor and a flat capacitor. The process damages into the tunnel oxides during the SA-STI fabrication steps are negligibly small.

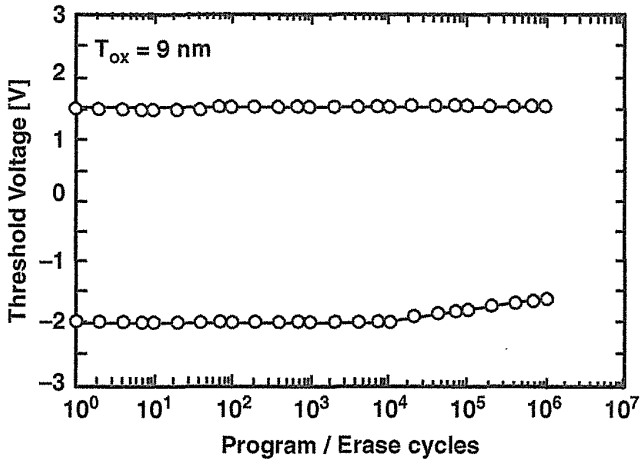


FIGURE 3.27 The program and erase cycling endurance characteristics of the SA-STI cell. The threshold voltage window narrowing has not been observed up to 1 million cycles.

3.3.4 Characteristics of Peripheral Devices

Figure 3.29 shows subthreshold characteristics of low voltage peripheral transistor as a function of the well voltage. No hump is observed in the subthreshold region as a result of avoiding to overlap the gate electrodes with the STI corners.

Figure 3.30 shows a junction breakdown voltage and a threshold voltage of a parasitic field transistor. The isolating ability of the STI is greatly higher than that of LOCOS. Furthermore, the breakdown voltage is higher than a demand (>22.5 V)

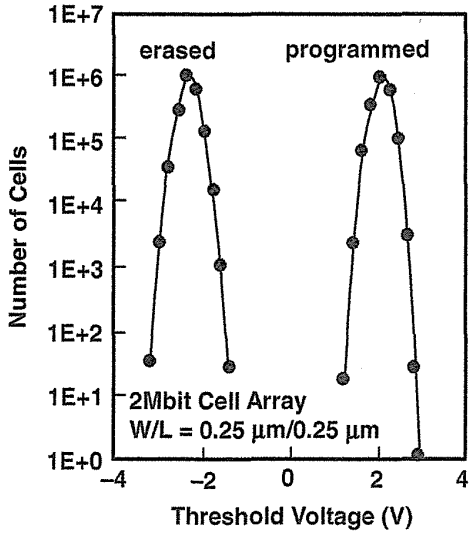


FIGURE 3.28 A cell threshold voltage distribution in one program and one erase pulse (no verify). Programming and erasing are carried out by 17 V, 10 µs and 18 V, 1 ms, respectively.

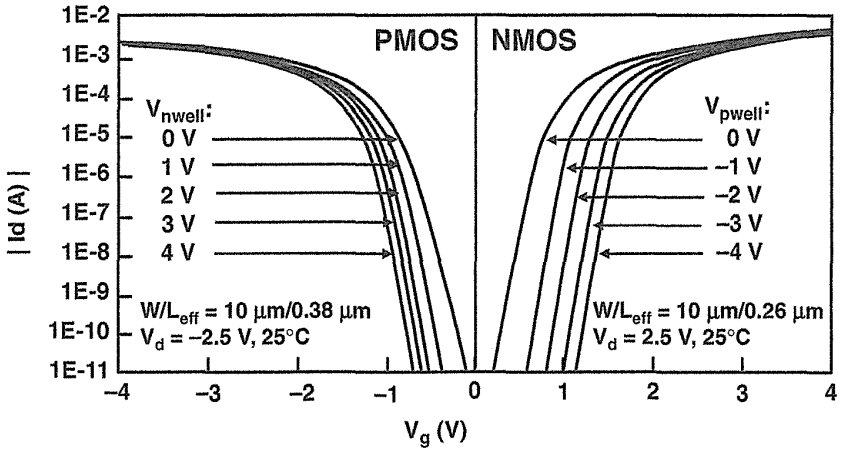


FIGURE 3.29 A subthreshold characteristics of the low-voltage peripheral transistors (NMOS and PMOS) as a function of well voltage.

with a sufficient margin. Therefore, the self-aligned STI process is suitable to the peripheral transistor as well as to the memory cell.

Figure 3.31 shows the TDDB characteristics of the gate oxides in the high-voltage transistors. The evaluated lifetime of the gate oxide is sufficiently long. The result implies that the process damages into the gate oxides are negligibly small.

A 0.31-μm^2 SA-STI cell with FG wing and peripheral integration process have been successfully developed using 0.25-μm design rules. This technology makes i

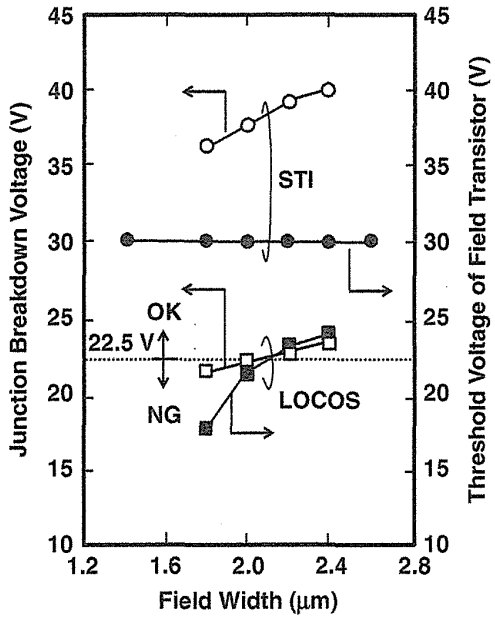


FIGURE 3.30 A punch-through voltage and threshold voltage of a parasitic field transisto for high-voltage isolation.

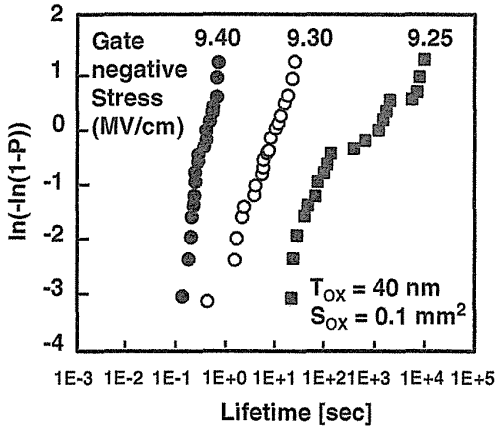


FIGURE 3.31 TDDB characteristics of the high-voltage gate oxide as a function of applied-gate electric field at negative gate voltage.

possible to realize reliable memory cell and peripheral devices with a simple process. Therefore, the SA-STI cell with FG wing is suitable for a low-cost flash memories of 256 Mbit and 1 Gbit for mass storage applications [7, 8]. The SA-STI cell with FG wing had been successfully adopted to a NAND flash memory product of 0.25-μm rule [7, 8], 0.15 to 0.16-μm rule [27], 0.12- to 0.13-μm rule [28, 29], and 90-nm rule [30], to achieve low-cost and reliable flash memory, as shown in Figs. 3.1 and 3.2.

3.4 SELF-ALIGNED STI CELL (SA-STI CELL) WITHOUT FG WING

An ultra-high-density NAND flash memory cell, using a self-aligned shallow trench isolation (SA-ST1) technology, had been developed for a high-performance and low-bit cost flash memory [6]. The SA-STI technology results in an extremely small cell size of ideal $4*F^2$ (F:feature size). The key technologies to realize a small cell size are (1) 0.4-μm (F) width shallow trench isolation (STI) to isolate neighboring bits and (2) a floating gate that is self-aligned with the STI, eliminating the floating-gate wings. Even though the floating-gate wings are eliminated, a high coupling ratio of 0.65 can be obtained by using the side walls of the floating gate to increase the coupling ratio. Using this self-aligned structure, a reliable tunnel oxide can be obtained because the floating gate does not overlap the trench corners, so enhanced tunneling at the trench corner is avoided. Therefore, the SA-STI cell combines a low bit cost with a high performance and a high reliability.

3.4.1 SA-STI Cell Structure

A self-aligned shallow trench isolation (SA-ST1) cell without FG wing is described for a high-performance and low-bit-cost NAND flash memory cell [6]. A small cell size of 0.67 μm^2, including the select transistor and drain contact area, was obtained

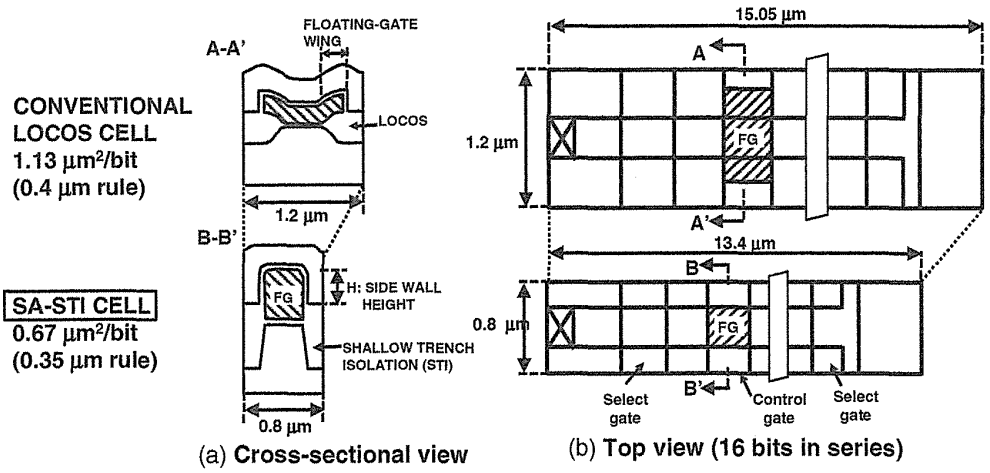


FIGURE 3.32 (a) The cross-sectional view and (b) top view of the self-aligned shallow trench isolation cell (SA-STI cell) without FG wing in comparison with that of the conventional LOCOS cell.

under a 0.35-μm design rule, in comparison with a 1.13-μm^2 LOCOS cell [3]. The key technology in obtaining small cell size is the bit-line isolation technology, which uses the shallow trench isolation (ST1) process. This technology also realizes a high reliability and a high performance.

Figure 3.32 compares the cross-sectional and top view of the SA-STI cell with that of the conventional LOCOS cell. This NAND structure cell has 16 memory transistors arranged between two select transistors in series. The word-line pitch is 0.7 μm. The bit-line pitch can be reduced to 0.8 μm by using 0.4-μm STI technology. As a result, the cell size of the SA-STI cell is about 60% of that of the conventional LOCOS cell [3].

In general, as the isolation width between the memory cells is reduced, the coupling ratio is reduced due to the decreased floating-gate wing area. However, in the SA-STI cell without FG wing, even if very tight 0.4-μm-width isolation is used, a high coupling ratio of 0.65 can be obtained because the 0.3-μm high side wall (H) of the floating gate is used to increase the coupling ratio, as shown in Fig. 3.33. Table 3.4 shows major cell parameters.

3.4.2 Fabrication Process

The fabrication of the SA-STI cell is simple and uses only conventional techniques, as shown in Fig. 3.34. First, a stacked layer of the gate oxide, the floating-gate poly-silicon, and the cap oxide is formed. Next, the trench isolation region is defined by patterning these three layers, followed by the trench etching, as shown in Fig. 3.34a and filling with LP-CVD SiO_2, as shown in Fig. 3.34b. Subsequently, the LP-CVD SiO_2 is etched back until the sidewall of the floating gate poly-silicon is exposed. After

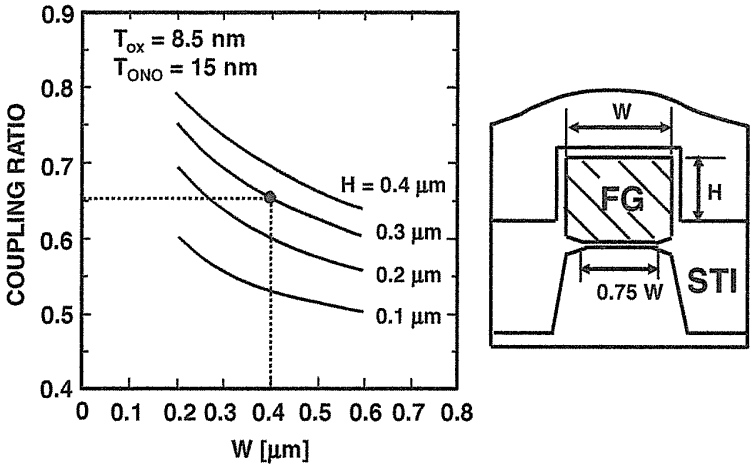


FIGURE 3.33 The coupling ratio of the SA-STI cell as a function of the gate width (*W*). A high coupling ratio of 0.65 can be obtained because the 0.3-μm-high sidewall (H) of the floating gate is used.

that, the inter-poly dielectric (ONO) (Fig. 3.34c) and the control-gate poly-silicon are formed (Fig. 3.34d), followed by the stacked-gate patterning. In this process, the floating-gate patterning and STI patterning are carried out by the same mask, so the number of fabrication steps for the SA-STI process can be decreased with about 10% in comparison with that for a conventional LOCOS process. Figure 3.35 shows cross-sectional SEM photograph of an SA-STI cell.

3.4.3 Shallow Trench Isolation (STI)

In the case of LOCOS isolation, the punch-through of the bit-line junctions occurs at a 0.5-μm isolation width, as shown in Fig. 3.36. However, in the case of STI, the punch-through and junction breakdown voltage are higher than 15 V even at a

TABLE 3.4 Memory Cell Parameters of the SA-STI Cell without FG Wing in 0.4-μm Technology

Cell size	0.67 μm^2 (including select Tr, etc.)
Gate length	0.35 μm
Gate width	0.4 μm
Trench isolation width	0.4 μm
Tunnel oxide	8.5 nm
Interpoly dielectric	ONO 15 nm (effective)
Programming time	0.195 μs/byte
Erase time	2 ms/sector
	2 ms/chip

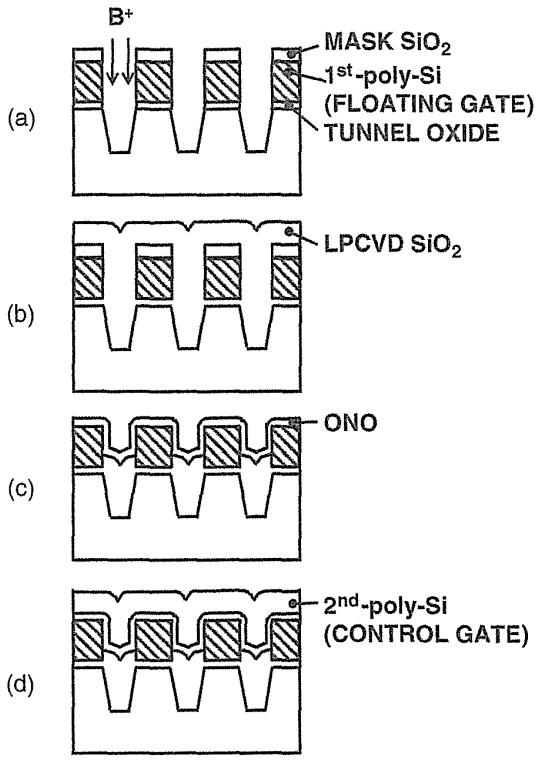


FIGURE 3.34 The process sequence of the SA-STI process without FG wing. (a) Trenc etching, B+ implantation. (b) LP-CVD SiO_2 fill-in. (c) Oxide etch-back and ONO formatior (d) Control-gate formation. The floating-gate and STI patterning are carried out by the sam mask, so the number of fabrication steps for the SA-STI process can be decreased with abou 10% in comparison with that for a conventional LOCOS process.

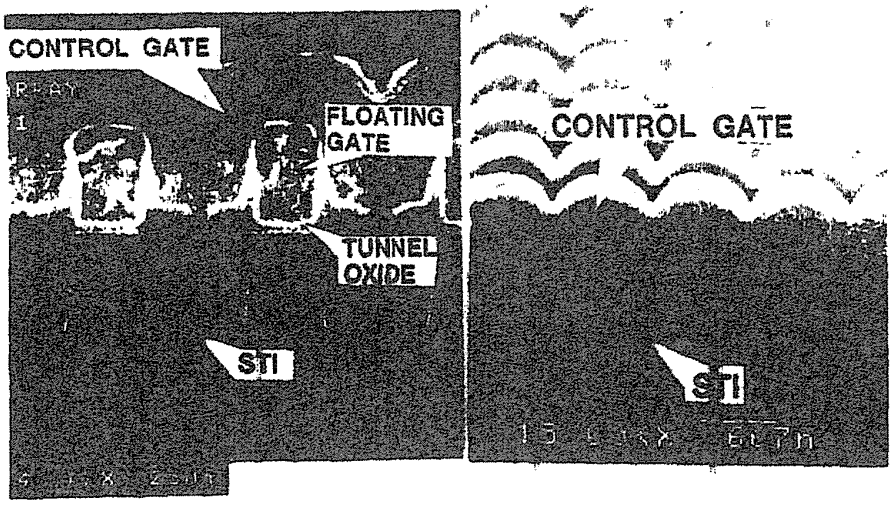


FIGURE 3.35 Cross-sectional SEM photograph of the SA-STI cell without FG wing.

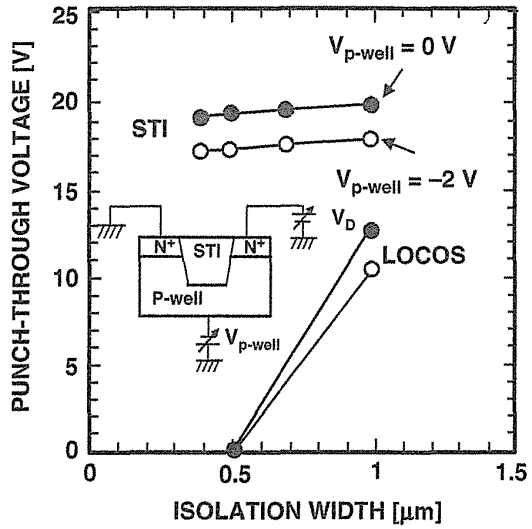


FIGURE 3.36 The punch-through voltage of the bit-line junction, which is isolated by shallow trench isolation (STI), in comparison with LOCOS isolation. The punch-through is higher than 15 V, which is high enough to realize 0.4-μm trench isolation.

0.4-μm isolation width, as shown in Fig. 3.36. Furthermore, the 0.7-μm-thick STI field oxide results in a high threshold voltage (>30 V) of the parasitic field transistor between the neighboring bits. As a result, a very tight 2∗F (0.8 μm) bit-line pitch can be realized by using STI.

Figure 3.37 illustrates the leakage of $n+-p$ junction for STI and LOCOS. The leakage current of STI is comparable to that of LOCOS.

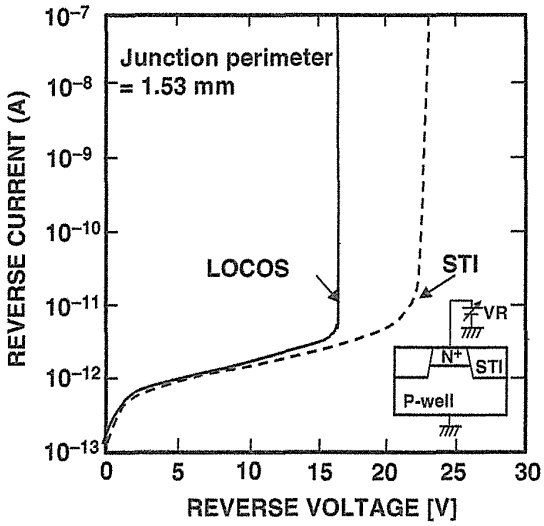


FIGURE 3.37 Junction leakage current of the SA-STI and LOCOS processes. The junction leakage current of the SA-STI process is comparable to that of LOCOS process.

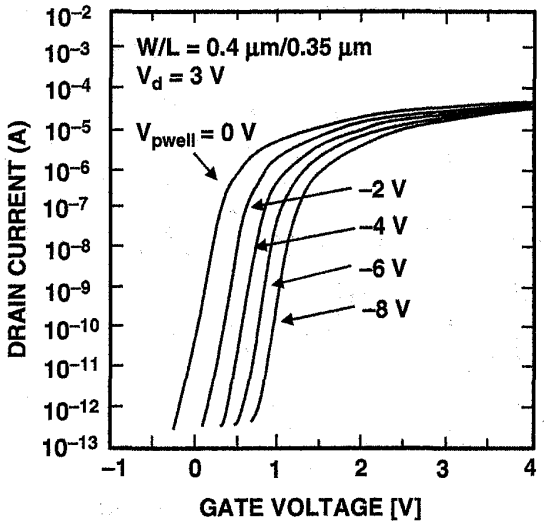


FIGURE 3.38 The subthreshold characteristics of the SA-STI cell with various substrate (*p*-well) bias conditions. Anomalous subthreshold characteristics (hump) cannot be seen because the floating-gate dose not overlap the trench comer.

3.4.4 SA-STI Cell Characteristics

Figure 3.38 shows the subthreshold characteristics of the SA-STI cell with a 0.4-μm channel width. Anomalous subthreshold characteristics (hump) cannot be seen as a result of the SA-STI structure.

Figure 3.39 shows the program and erase characteristics of an SA-STI cell. The fast programming (100 μs/512 byte) and fast erasing (2 ms) can be accomplished by Fowler–Nordheim tunneling, applying a positive voltage of 17 V to the control gate during programming and 17 V to the *p*-well during erasing, respectively, as shown in Fig. 3.39.

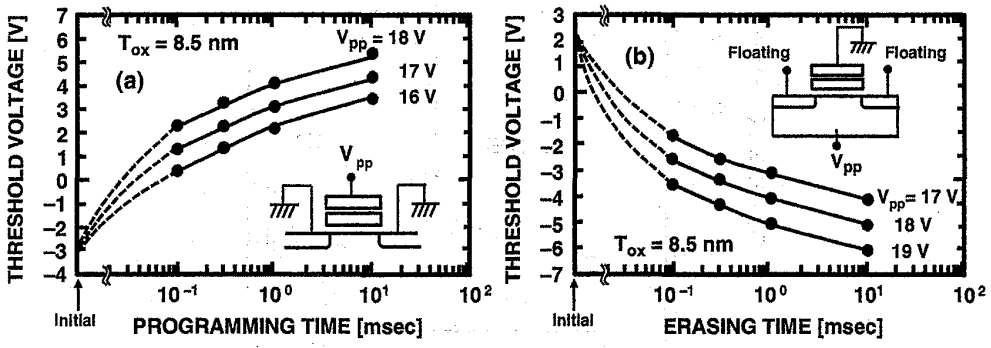


FIGURE 3.39 (a) Program and Erase characteristics of the SA-STI cell. The fast programming (100 μs) and erasing (2 ms) can be accomplished by Fowler–Nordheim tunneling over the channel area, applying a positive voltage of 17 V to the control gate during programming and 17 V to the *p*-well during erasing, respectively.

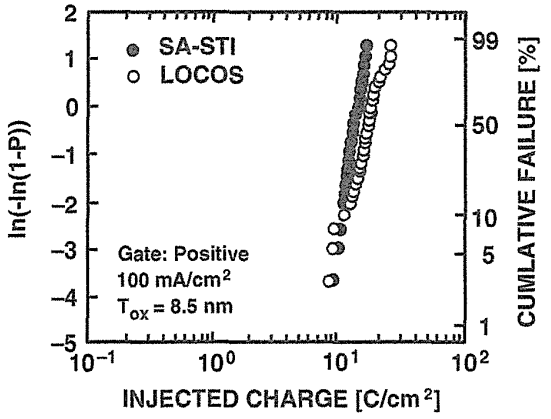


FIGURE 3.40 The TDDB characteristics of the 8.5-nm-thick tunnel oxide. The TDDB characteristics in the SA-STI process without FG wing are almost the same as that in the LOCOS process.

The TDDB characteristics of the tunnel oxide in the SA-STI process are almost the same as that in the LOCOS process, as shown in Fig. 3.40., because the floating gate does not overlap the trench edges. Therefore, the endurance characteristics of the SA-STI cell are comparable with that of the conventional LOCOS cell, as shown in Fig. 3.41. The SA-STI cell guarantees a wide cell threshold window as large as 3 V, even after 1 million write/erase cycles. Furthermore, read disturb characteristics can be ensured for more than 10 years even after 1 million write/erase cycles, as shown in Fig. 3.42.

The SA-STI cell without FG wing has a very simple cell structure and has a very small cell size of ideal $4*F^2$ with bit-line and word-line pitch of $2*F$ (F; feature size),

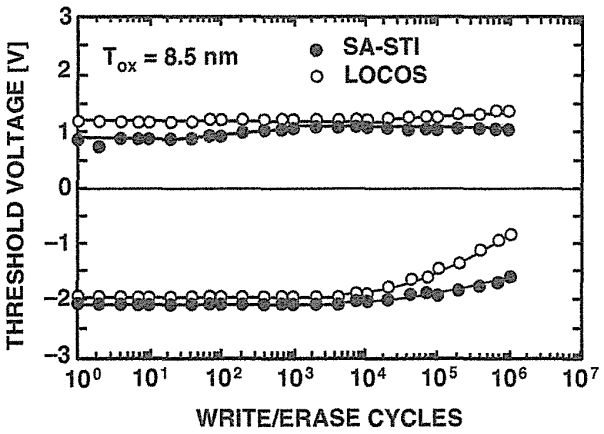


FIGURE 3.41 The program(write)/erase endurance characteristics of the SA-STI cell and the LOCOS cell. In the SA-STI cell, window narrowing has not been observed up to 1 million program (write)/erase cycles.

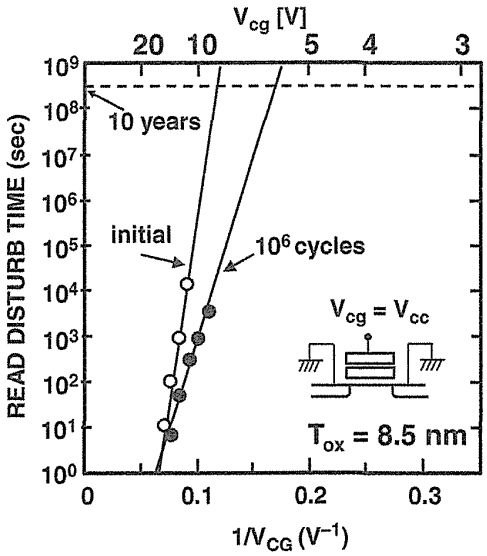


FIGURE 3.42 The read disturb characteristics of the SA-STI cell without FG wing. The read disturb time is more than 10 years when a V_{cc} of 3.0 V is used, even after 1 million program/erase cycles .

as shown in Fig. 3.2. The SA-STI technology has also demonstrated an excellent reliability and performance. Therefore, as shown in Fig. 3.1, the SA-STI cell without FG wing has been extensively used for more than 12 years since around (2002), over eight generations (90 nm to 1X nm) of NAND flash memory product, such as a 90 nm cell [31], 70- nm cell, a 50-nm cell, a 43-nm cell [32], a 30-nm cell [33], a 27-nm cell [34], a 20-nm cell, and a mid-1X-nm cell [21, 35].

3.5 PLANAR FG CELL

3.5.1 Structure Advantages

The conventional self-aligned STI cell (SA-STI cell) has a structure problem of control gate formation between floating gates, as shown in Fig. 3.43 (as described in Section 5.6). There is not enough space between floating gates (FGs) to fabricate a control gate in a scaled cell [36]. To solve this problem, two solutions were proposed, as shown in Fig. 3.43. One is the slimming FG width to obtain an enough space for a control gate [21, 35]. The other is the planar FG cell [10–12], which has a thin (~10 nm) floating gate with a high-k inter-poly dielectric (IPD). Thanks to a high-k IPD, the capacitance between CG and FG becomes large enough to operate a memory cell. Then FG thickness can be very thin.

Figure 3.44 shows cross sections of (a) a conventional SA-STI cell, (b) a planar FG cell [10], and (c) the stacked structure of a planar FG cell [11]. Thickness of

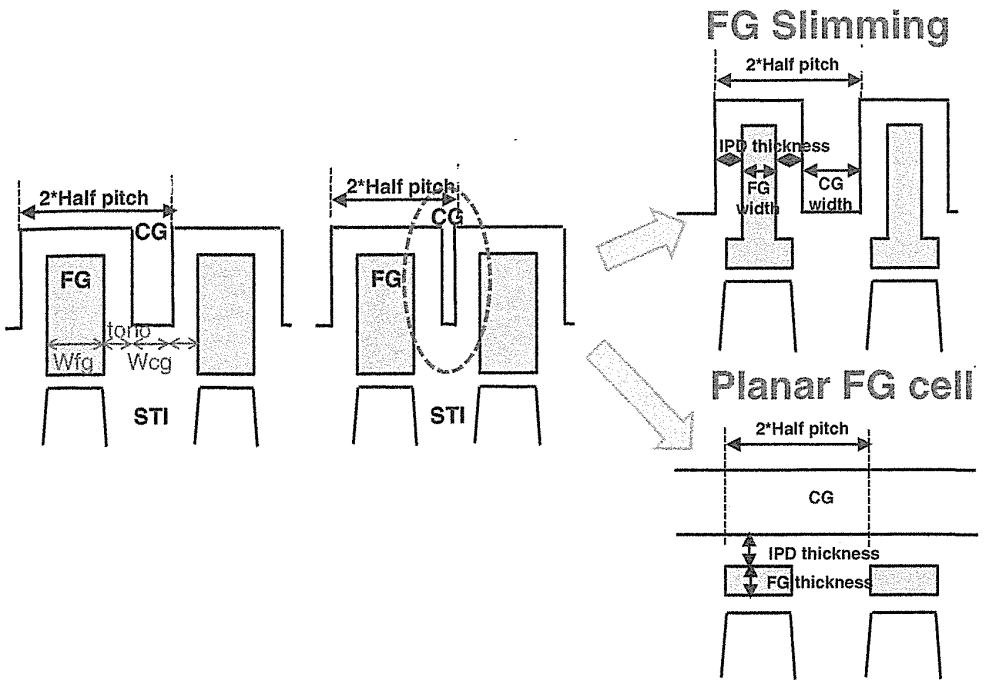


FIGURE 3.43 Structure problem of the self-aligned STI cell and two solutions of floating-gate (FG) slimming and the planar FG cell.

floating gate is very thin, around 10 nm. And the high-k block dielectric (BD) is stacked on thin FG as IPD. The aspect ratio of stacked gate and control gate (CG) fill are compared between conventional SA-STI cell (wrap cell) and a planar FG cell, as shown in Fig. 3.45 [12]. In the SA-STI cell, the aspect ratio becomes more than 10 for both the stacked gate and CG fill in a sub-20-nm cell. The planar FG cell can much mitigate this limitation.

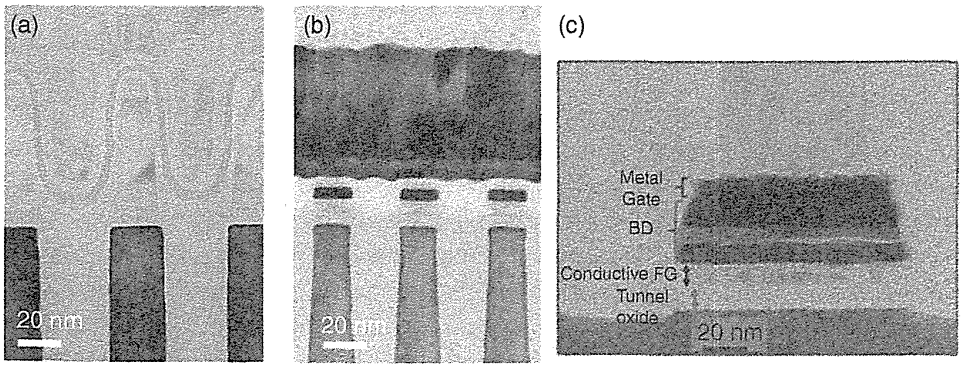


FIGURE 3.44 Cross sections of (a) a conventional SA-STI cell without FG cell, (b) a 20-nm planar FG cell, and (c) a stacked structure of planar FG cell.

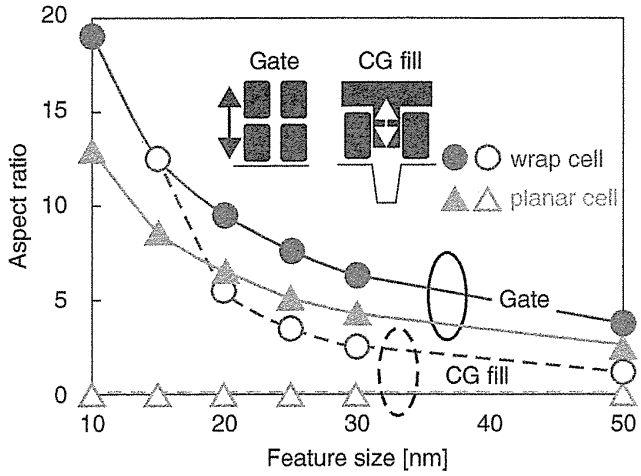


FIGURE 3.45 Aspect ratio (A.R.) of a conventional SA-STI cell (wrap cell) and a planar cell. A.R. increases with scaling. A.R. is >10 for both the word-line and the bit-line directions in a sub-20-nm SA-STI cell (solid: Gate, open: CG fill).

3.5.2 Electrical Characteristics

The planar FG cell can drastically reduce the floating-gate capacitive coupling interference (cell-to-cell interference), as shown in Fig. 3.46. Due to small floating-gate capacitive coupling interference, the read window margin (RWM) (see Section 5.2) can be much improved. Also, the erase V_t setting can be shallower V_t (higher V_t). Then program/erase cycling endurance and data retention can be expected to improve

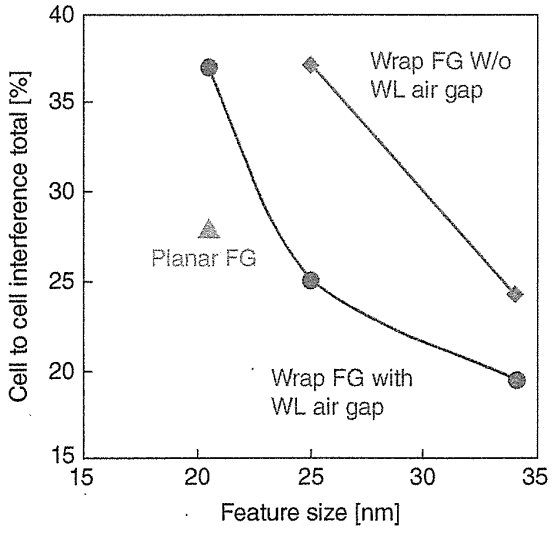


FIGURE 3.46 Cell-to-cell interference (floating-gate capacitive coupling interference) scaling. An ~30% total interference reduction is achieved with the planar FG cell.

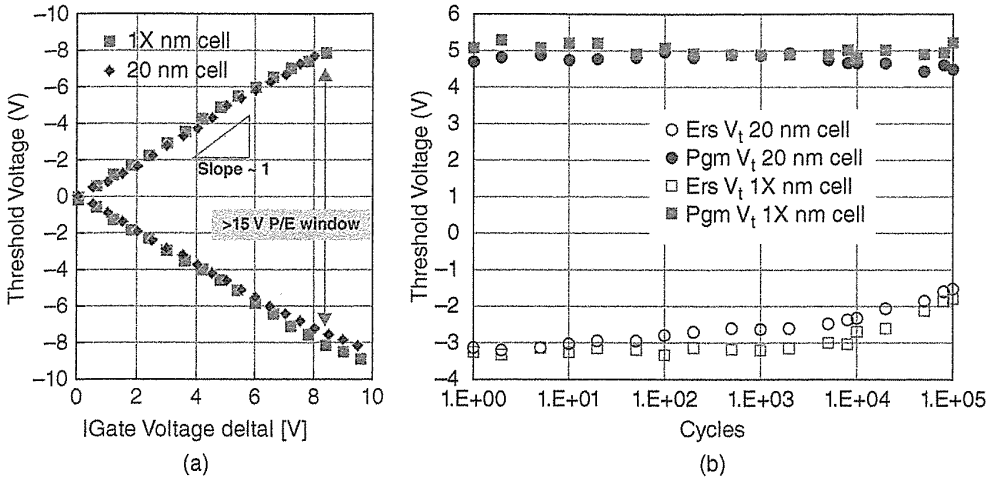


FIGURE 3.47 (a) Program/erase characteristics and (b) program/erase cycling endurance characteristics of a 20-nm and a 1X-nm planar FG cell.

because oxide stress is reduced due to smaller amount of charge through oxide during program/erase operations.

Figure 3.47a shows program/erase characteristics. Excellent program/erase window and program slope (~1) are demonstrated in the planar FG cell of a 20-nm cell and 1X-nm cell. Both of these characteristics are important for enabling a highly reliable MLC NAND flash memory. Figure 3.47b shows the program/erase cycling endurance characteristics. Excellent cycling endurance characteristics are also demonstrated.

The planar FG cell has a potential to extend NAND cell scaling very effectively by removing the structure problem and by small floating-gate capacitive coupling interference.

3.6 SIDEWALL TRANSFER TRANSISTOR CELL (SWATT CELL)

A multilevel NAND flash memory cell, using a sidewall transfer-transistor (SWATT) structure, had been developed for a high-performance and low-bit-cost flash memory [13, 14]. With the SWATT cell, a relatively wide threshold voltage (V_{th}) distribution width of about 1.1 V can be obtained for MLC (2 bits/cell) in contrast to a narrow 0.6-V distribution width that is required for a conventional cell. The key technology that allows this wide V_{th} distribution width is the transfer transistor, which is located at the side wall of the shallow trench isolation (STI) region and is connected in parallel with the floating-gate transistor. During read, the transfer transistors of the unselected cells (connected in series with the selected cell) work as pass transistors. So, even if the V_{th} of the unselected floating-gate transistor is higher than the control-gate voltage, the unselected cell will be in the ON state. As a result, the V_{th} distribution of

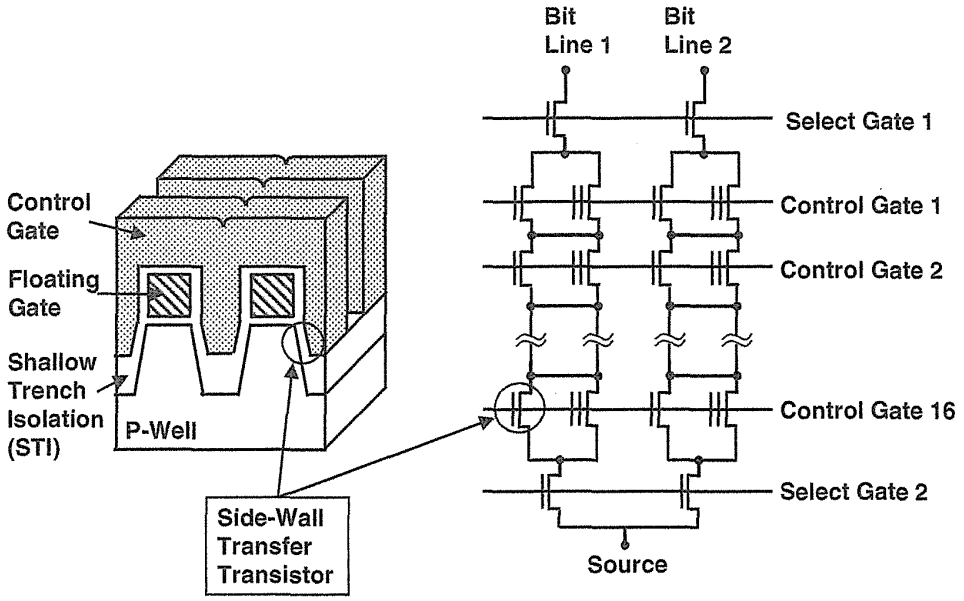


FIGURE 3.48 The schematic view and equivalent circuit of the sidewall transfer transistor cell (SWATT cell). A transfer transistor is located at the sidewall of the shallow trench isolation (STI) region and is connected in parallel with the floating-gate transistor.

the floating-gate transistor can be wider and the programming can be faster because the number of program/verify cycles can be reduced.

3.6.1 Concept of the SWATT Cell

The concept of a sidewall transfer-transistor cell (SWATT cell) for multilevel NAND flash memory [13, 14] is described. The schematic view and equivalent circuit of the SWATT cell are shown in Fig. 3.48. One cell consists of both a floating-gate transistor and a transfer transistor, which is located at the sidewall of the shallow trench isolation (STI) region. These two transistors are connected in parallel. Sixteen cells are connected in series between two select transistors to form a NAND cell string. The read conditions of a conventional NAND cell and a SWATT cell for the two-level scheme (SLC) are shown in Fig. 3.49. In a conventional cell, zero volt is applied to the gate of the selected memory cell, while 5.0 V is applied to the gates of the unselected cells in the NAND string. All the memory cells, except for the selected cell, serve as transfer gates. Therefore, for the conventional NAND cell, the threshold voltage of the in-series connected cells must be lower than the unselected control-gate (CG) voltage of 4.5–5.5 V. Thus, the V_t distribution of the cells in the programmed state must be narrow with a width of less than 3.0 V for two-level operation, as shown in Fig. 3.49a.

On the other hand, the sidewall transfer transistor in the SWATT cell works as a pass transistor instead of a floating-gate transistor, as shown in Fig. 3.49b. So the

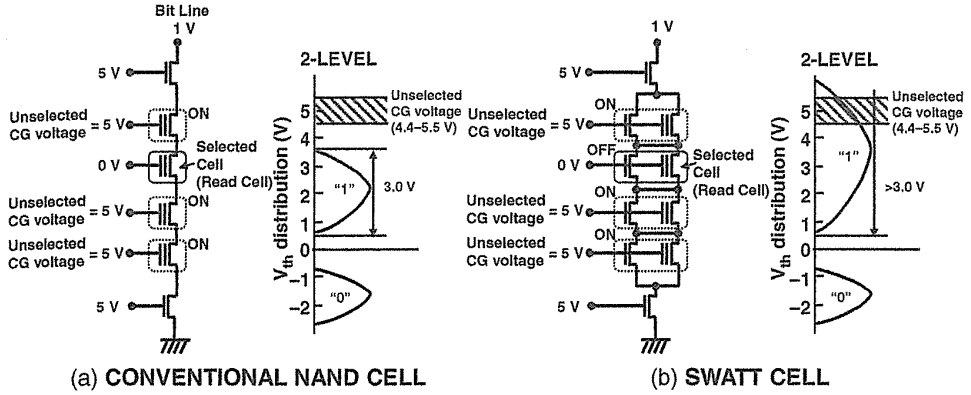


FIGURE 3.49 Read condition of (a) a conventional NAND cell and (b) the SWATT cell for a two-level scheme. In the conventional NAND cell, the V_{th} distribution of the cells in programmed state must be narrow with a width of 3.0 V or less, because the unselected cells must work as pass transistors for a control-gate voltage of 5 V. However, in the SWATT cell, the sidewall transfer transistor works as a pass transistor. Therefore, the V_{th} distribution of the floating-gate transistor in the programmed state is allowed to be very wide with a width of >3.0 V for two-level operation.

threshold voltage of the floatin-gate transistor does not have to be lower than the unselected CG voltage of 4.5–5.5 V. Therefore, the V_t distribution of the floating-gate transistor in the programmed state is allowed to be very wide with a width >3 V for two-level operation. As a result, the V_t distribution can be wider in comparison with the conventional NAND cell.

The threshold voltage distributions of the two-level (SLC) and four-level scheme (MLC) are compared in Fig. 3.50. In a conventional NAND cell, the V_t distribution of the cells in the programmed states ("1," "2," and "3") must be very narrow (0.6 V), because the in-series connected cells must work as pass transistors. However, in the SWATT cell, the V_t distribution of the floating-gate transistor in the programmed "1" and "2" states is allowed to be very wide (1.1 V). The V_t distribution in the programmed "3" state is allowed to be even wider than 1.1 V.

This wide threshold voltage distribution results in a high programming speed because of reducing the number of program/verify cycles and good data retention characteristics.

3.6.2 Fabrication Process

The developed SWATT cell has 16 memory transistors connected in series between two select transistors. The word-line pitch is 0.7 μm. A very narrow bit-line pitch of 0.8 μm can be realized by using 0.4-μm-width shallow trench isolation (STI) technology. As a result, a small cell size of $5.5*F^2$ (0.67 μm^2), including the select transistor and drain contact area, can be obtained under a 0.35-μm design rule. The thickness of sidewall dielectric (ONO) is 40 nm effective.

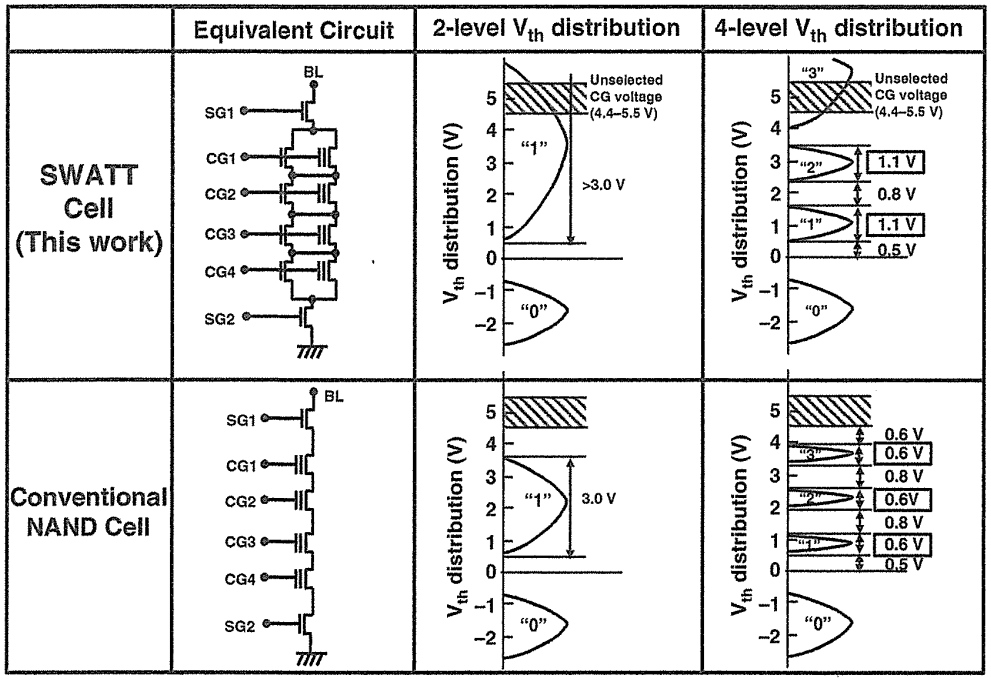


FIGURE 3.50 The cell threshold voltage distribution of the SWATT cell and the conventional NAND cell for two-level (SLC) and four-level (MLC) operation. In the SWATT cell, a wide threshold distribution of 1.1 V is allowed for four-level operation, in comparison with a 0.6-V distribution that is required for the conventional NAND cell. The range of the unselected CG (control gate) voltage is limited because of read disturb.

The fabrication process of the SWATT cell is similar to that of the SA-STI cell. The process sequence of the SWATT cell is shown in Fig. 3.51. First, a stacked layer of the gate oxide, the floating-gate poly-silicon, and the cap oxide is formed. Next, the trench isolation region is defined by patterning these three layers, followed by the trench etching, trench bottom boron implantation, and filling with LP-CVD SiO_2, as shown in Fig. 3.51a. Subsequently, the LP-CVD SiO_2 is etched back until the sidewall of the STI is exposed (Fig. 3.51b). Boron (B+) ion implantation (60 KeV, 2E12/cm^2) is carried out for V_{th} adjustment of the sidewall transfer transistor. After that, the interpoly dielectric (ONO) and transfer-transistor gate oxide are formed at the same time, as shown in Fig. 3.51c. Then the control-gate poly-silicon is deposited, followed by the stacked gate patterning (Fig. 3.51d). In this process, the thermal oxide of the STI sidewall is about two times thicker than that on the poly-silicon due to oxidation enhancement at the STI sidewalls. As a result, breakdown of the control gate does not occur even if a high voltage of about 20 V is applied to the control gate during the program operation.

A cross-sectional TEM photograph is shown in Fig. 3.52. Both the trench isolation and channel width (gate width) of the floating gate transistor are 0.4 μm. The vertical channel width of the sidewall transfer transistor is about 0.2 μm.

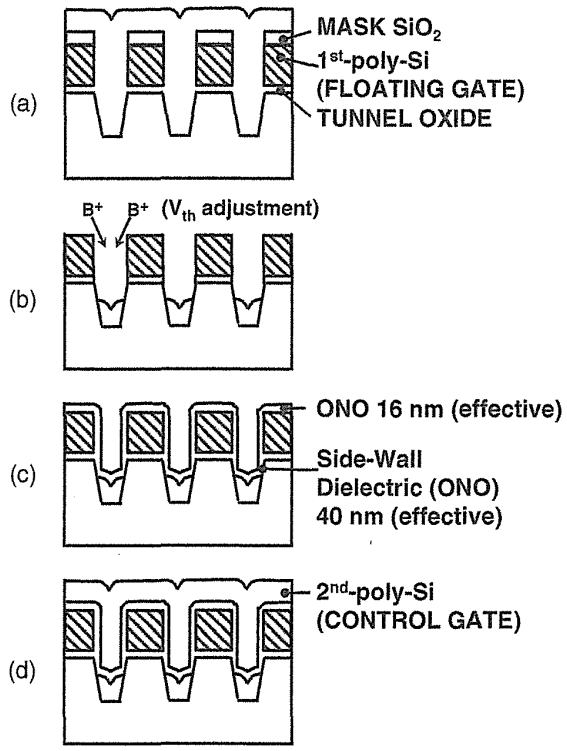


FIGURE 3.51 The process sequence of the SWATT process. (a) Trench etching, LP-CVD SiO_2 fill-in. (b) Oxide etch-back and B+ implantation of the V_{th} adjustment of the side wall transfer transistor. (c) ONO formation. (d) Control gate formation. The thermal oxide of the STI sidewall is about two times thicker than that on the poly-silicon due to oxidation enhancement at the STI sidewalls.

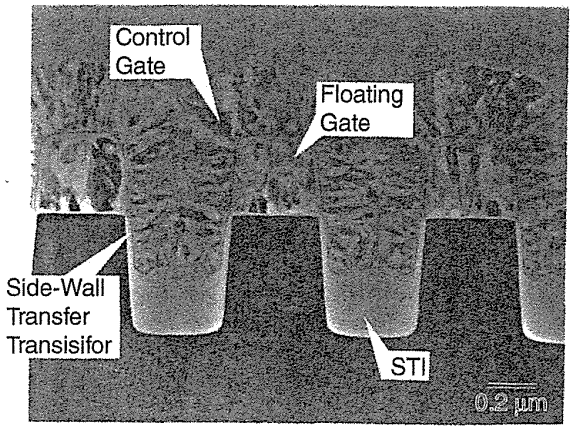


FIGURE 3.52 Cross-sectional TEM photograph of the SWATT cell along the word-line (Control Gate) direction.

An accurate control of the threshold voltage of the sidewall transfer transistor is important for the SWATT cell. The range of the threshold voltage is determined as follows. The sidewall transfer transistor must be in the ON state when the unselected CG voltage (4.5–5.5 V) is applied to the control gate. So, the upper limit of the V_{th} of the sidewall transistor is 4.5 V. On the other hand, the sidewall transfer transistor must be in the OFF state when the read voltage (about 3.9 V) between the "2" and "3" state for four-level operation is applied to the control gate. Therefore, the threshold voltage of the side-wall transfer transistor must be in the range from 3.9 V to 4.5 V for four-level operation (from 0 V to 4.5 V for two-level). The important statistical parameters of V_{th} of a sidewall transfer transistor are boron concentration in the channel region and the sidewall gate-oxide thickness. Boron concentration is well controlled by boron implantation, as shown in Fig. 3.51b. Also, the oxide thickness of the STI side wall is controlled within 10% variation. Therefore, the narrow range of the threshold voltage of the sidewall transfer transistor can be adjusted.

3.6.3 Electrical Characteristics

A. Isolation For the NAND flash cell, the high-voltage isolation technology is important to reduce the bitline pitch. The isolation between the bit lines must satisfy two demands. One is a high punch-through or junction breakdown voltage of the bit-line junction area (>10 V). The other is a high threshold voltage of the parasitic field transistor (>25 V) of the control gate (CG) in the memory cell.

The breakdown voltage of the bit-line junction occurs at about 19 V while no punch-through is observed. The breakdown voltage is higher than the required 10 V, which is high enough to apply NAND flash cell.

Figure 3.53 shows the threshold voltage of the parasitic field transistor in the SWATT cell. The 0.3-μm-thick STI field oxide results in a high threshold voltage (>30 V) of the parasitic field transistor between the neighboring bits.

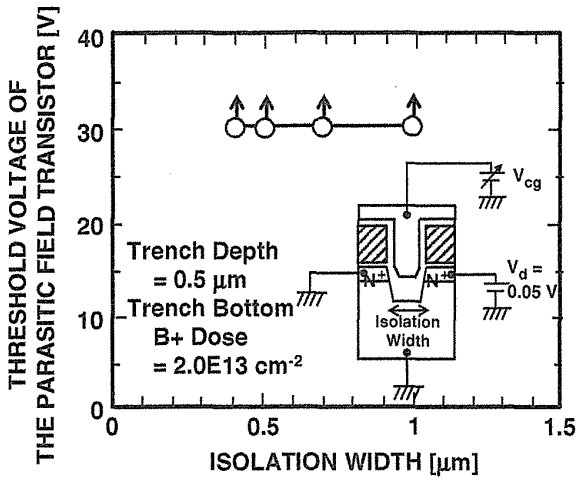


FIGURE 3.53 The threshold voltage of the parasitic field transistor in the SWATT cell, which is isolated by shallow trench isolation (STI). The threshold voltage of the field transistor is higher than 30 V, which is high enough to realize 0.4-μm-width trench isolation.

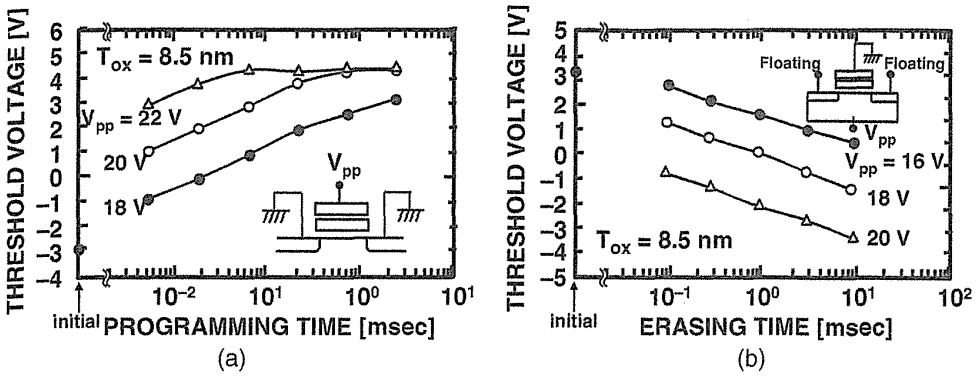


FIGURE 3.54 (a) The program and (b) erase characteristics of the SWATT cell. A short programming time of 200 μs and short erase time of 2 ms can be accomplished by Fowler–Nordheim tunneling over the channel area, applying a positive voltage of 21 V to the control gate during programming and 19 V to the *p*-well during erasing, respectively.

B. Cell Characteristics The program and erase characteristics of the SWATT cell are shown in Fig. 3.54a and Fig. 3.54b, respectively. The threshold voltage of programming saturates at about 4.2 V. This is explained as followed. In this memory cell (observed V_{th} = 4.2 V), a floating-gate transistor is programmed to high threshold voltage (V_{th} > 4.2 V), so a floating-gate transistor is in the OFF state for a measurement condition. On the other hand, the sidewall transfer transistor is in the ON state for V_{cg} > 4.2 V, because the V_{th} of the sidewall transfer transistor is about 4.2 V. Therefore, the V_{th} of sidewall transfer transistor is observed. Then, V_{th} saturates at about 4.2 V even after long programming time (>0.1 ms at 22 V). It can be seen that a fast programming (200 μs/512 byte) and erase operation (2 ms) can be obtained.

Figure 3.55 shows the subthreshold I_d–V_g characteristics of the SWATT cell at the erased "0" and programmed "1", "2", "3" states. In the programmed "3" state, the V_{th} of the floating-gate transistor is higher than 4.5 V, so only the I_d of the sidewall transfer transistor can be observed.

Figure 3.56 shows the coupling ratio of the SWATT cell as a function of the gate width (W). In general, as the isolation width between the memory cells is reduced, the coupling ratio is reduced due to the decreased floating-gate wing area. However, in the SWATT cell, even if very tight 0.4-μm-width isolation is used, a high coupling ratio of 0.65 can be obtained because the 0.3-μm-high sidewall (H) of the floating gate is used to increase the coupling ratio. Moreover, the coupling ratio increases as the gate width W is scaled down. This means that the programming voltage and erasing voltage (V_{pp}) can be reduced as the memory cell is scaled down, which allows the design of more compact peripheral circuits such as row decoders and sense amplifiers. Furthermore, the variation of the coupling ratio of the SWATT cell can be very small because the sidewall (H) of the floating gate is determined by the thickness of the floating-gate poly-silicon. Therefore, a very tight V_t distribution of the SWATT cell is expected.

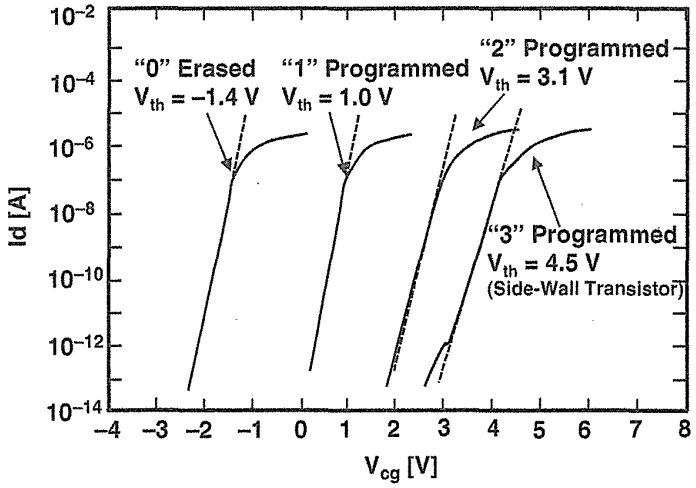


FIGURE 3.55 The subthreshold characteristics of the SWATT cell for the erased "0" and programmed "1," "2," "3" states. In the programmed "3" state, the side-wall transfer transistor is in the ON state.

C. Reliability Figure 3.57 shows the program/erase cycling endurance characteristics of a SWATT cell using the uniform program/erase scheme [22–25]. This scheme guarantees a wide cell threshold window of as large as 3 V, even after one million write/erase cycles. These endurance characteristics of the SWATT cell are comparable to that of the conventional NAND cell [4–6].

Read disturb occurs as a weak programming mode. The tunnel-oxide leakage currents, which are induced by the program and erase cycling stress, degrade the read disturb of the memory cell, as shown in Fig. 3.58. However, even after one million program/erase cycles, the read disturb time is more than 10 years when a V_{cg} of 5.0 V is used.

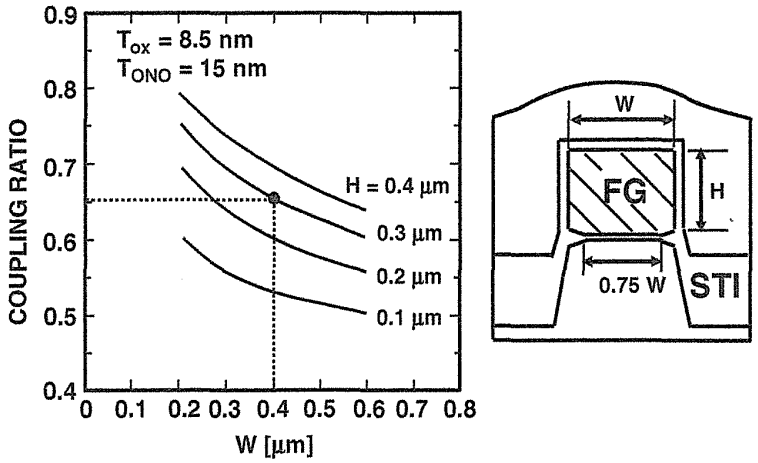


FIGURE 3.56 The coupling ratio of the SWATT cell as a function of the channel width (W). A high coupling ratio of 0.65 can be obtained because the 0.3-μm-high side-wall (H) of the floating gate is used to increase the coupling ratio.

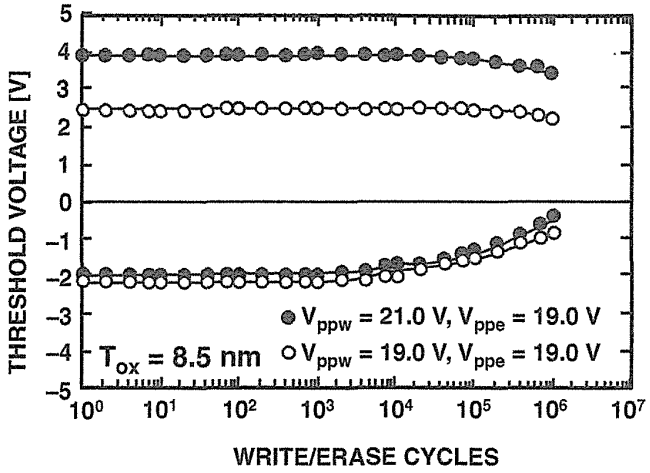


FIGURE 3.57 The program (write)/erase cycling endurance characteristics of the SWATT cell. Window narrowing has not been observed up to one million program/erase cycles.

3.7 ADVANCED NAND FLASH DEVICE TECHNOLOGIES

3.7.1 Dummy Word Line

A dummy word-line (dummy cell) scheme in NAND flash memory was proposed to eliminate abnormal program disturb of edge memory cell [15–17]. Dummy word line (dummy cell) is located between edge word lines (edge memory cells) of NAND

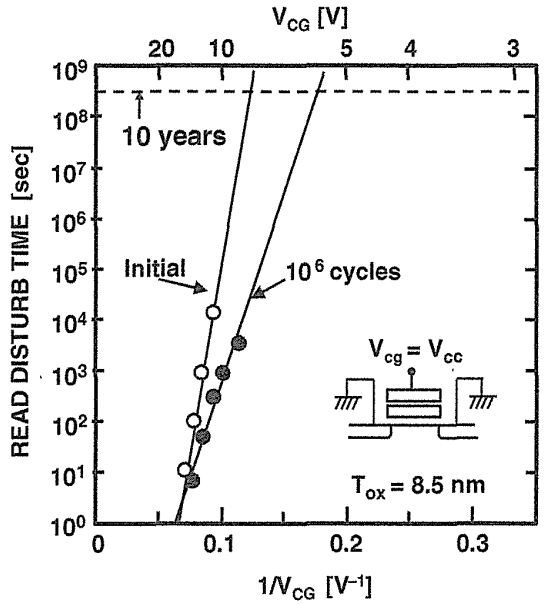


FIGURE 3.58 The read disturb characteristics of the SWATT cell. The read disturb time is more than 10 years when a V_{cc} of 5.0 V is used, even after 1 million program/erase cycles.

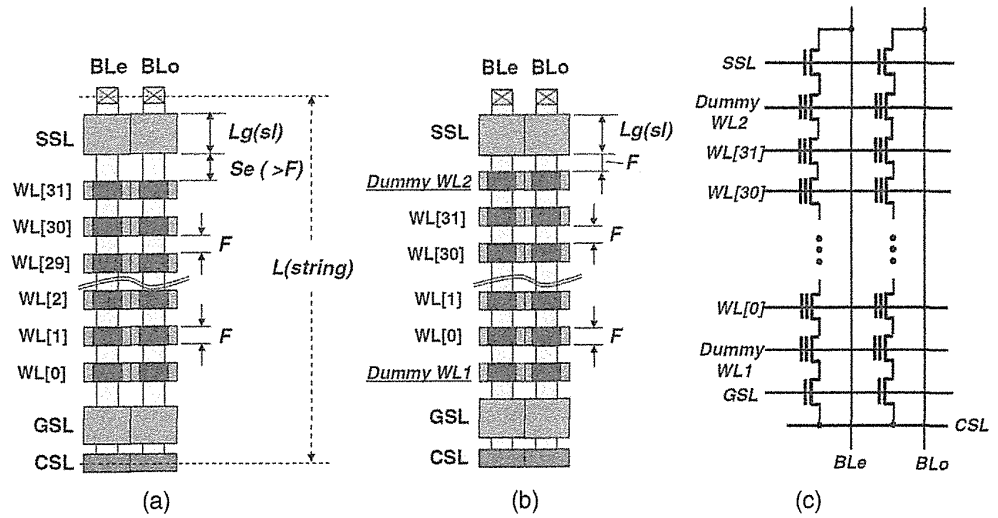


FIGURE 3.59 (a) Layout view of conventional NAND flash memory. (b) Layout view of dummy word lines in NAND flash memory. (c) Schematic diagram of dummy word lines in NAND flash memory. Copyright 2007, The Japan Society of Applied Physics.

string and select gate transistors (GSL or SSL). The program disturb of a GIDL generated hot electron injection mechanism [37] can be suppressed by increasing the distance between an edge cell and a select transistor. Also, the program boosting potential drop from edge cell to select transistor can be well controlled by using the proper dummy word-line voltage and proper dummy cell V_t. Therefore, abnormal program disturbance of an edge memory cell can be greatly suppressed. In addition, capacitive coupling noise between select transistor and edge memory cell can be reduced to less than 50%. The program disturbance failure, read failure, and erase distribution width can be reduced by reducing coupling noise. The dummy word-line scheme was started to be used from a 40-nm technology node due to stable operations in edge cells [17].

By scaling a NAND flash memory cell, area overhead of two select transistors in a NAND string is increasing because a select transistor cannot be scaled down as memory cell scale down due to the required punch-through immunity for program boosting voltage. This is one of the scaling problems for a NAND flash memory cell. Also, area of space (S_e) between select transistors (GSL, SSL) and edge word lines (WLs: WL[0] and WL[31]) is another area overhead problem, as shown in Fig. 3.59(a). Reducing the space S_e is hard to be scaled down because of the following two reasons. One is that the capacitive coupling noise between the select transistor and edge WLs is increased by reducing S_e. A boosting channel potential of program inhibit is decreased by a leakage current through a select transistor which is slightly turned on when V_{pass} and V_{pgm} voltages are applied to the edge WL due to large coupling between select transistor and edge WL. It causes program inhibit failure. Also, during read for an edge cell, the voltage of edge word line has a bump due to coupling with a select gate, which is ramped up after ramping up of voltage of edge

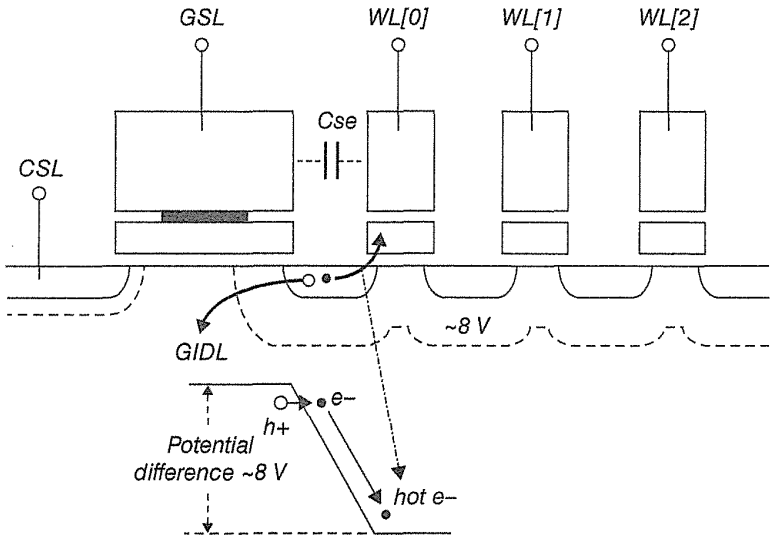


FIGURE 3.60 Enhanced program disturbance of edge memory cell by hot carrier induced by GIDL. Copyright 2007, The Japan Society of Applied Physics.

word lines [39]. It causes read failure. The other reason is that hot-carrier disturbance due to large electric field in junction between select transistor and edge WL [37], as shown in Fig. 3.60 (see Section 6.5.2). The hot carriers are mainly generated by a GIDL (gate-induced drain leakage) mechanism, and hot electrons are enhanced by an electric field between a select transistor and an edge cell. Some hot electrons are injected to the floating gate of an edge memory cell. It was reported that at least a larger than 110 nm S_e is required to avoid severe hot carrier program disturbance, as shown in Fig. 3.61 [37].

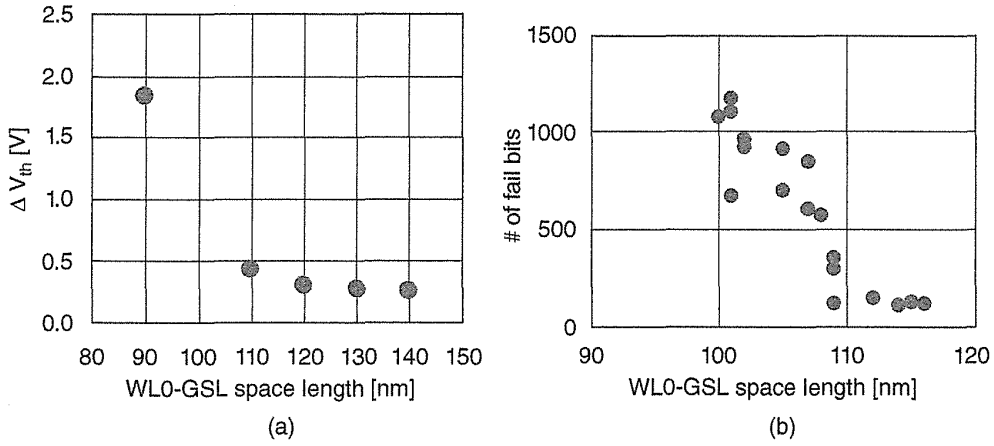


FIGURE 3.61 (a) Simulation result of the number of electrons injected to the WL0 cell for cell arrays having various WL0-GSL spaces. (b) Number of fail bits measured with 1 Mb block array at V_{pass} = 10 V. Copyright 2007, The Japan Society of Applied Physics.

TABLE 3.5 Read and Erase Condition for the Dummy Word-Line Scheme in NAND Flash Memory Cells

	Read	Erase
BL	V_{pc}	F
SSL	V_{cc}	F
Dummy WL2	V_{read}	0 V
Unselected WL	V_{read}	0 V
Selected WL	V_r	0 V
Dummy WL1	V_{read}	0 V
GSL	V_{cc}	F
CSL	0 V	F

V_{pc}, BL precharge voltage; V_r, read voltage for selected WL; V_{read}, read voltage for unselected WL; F, Floating).

Source: Copyright 2007, The Japan Society of Applied Physics.

Furthermore, an edge memory cell has a different condition compared with middle cells. In an edge cell, one side of a source/drain is connected to the select transistor while the other side is connected to a neighbor cell. However, in the middle cell, both sides are connected to neighbor cells. The potential of a floating gate is different between an edge cell and middle cells during operations. It causes abnormal electrical characteristics such as erase and program characteristics, as compared to middle cells. This is because the coupling ratio of the floating gate and the voltage condition applied around neighbor gates are different between an edge cell and a middle cell for each operation. It eventually results in wide V_{th} distributions of erase and program state.

To solve these scaling issues of an edge memory cell, a dummy word-line scheme and the new operation conditions were proposed [15–17]. Figures 3.59b and 3.59c show a structures of a dummy cell scheme. A dummy cell which is identical to normal memory cell is additionally placed between each select transistors (GSL, SSL) and the edge memory cell (WL[0], WL[31]). The space between the select transistor and the dummy cell is basically formed by F (feature size). By adjusting V_{th} of the dummy cell combined with an optimized dummy word-line bias condition, a nearly equal environment of the middle cell can be provided to the edge memory cell so that the unexpected edge memory cell effects can be eliminated. During read and erase operation, the dummy cell acts as a normal memory cell. The operation conditions of dummy word lines are shown in Table 3.5.

Figure 3.62 shows (a,b) a simulated band-to-band electron/hole generation contour and (c) a simulated lateral electric field in the case of both a conventional scheme (without dummy cell) and a dummy cell scheme during program inhibit condition [15]. A high lateral electric field generates large number of band-to-band carriers in case of a conventional scheme, however, in a dummy cell scheme, an electric field can be suppressed in between the dummy cell and the edge memory cell. It can be explained by that an optimized biased voltage and adjusted V_{th} of the dummy cell mitigate an electrostatic potential difference between a dummy cell and an edge memory cell, so that it results in a decreasing injection of hot electron carriers to

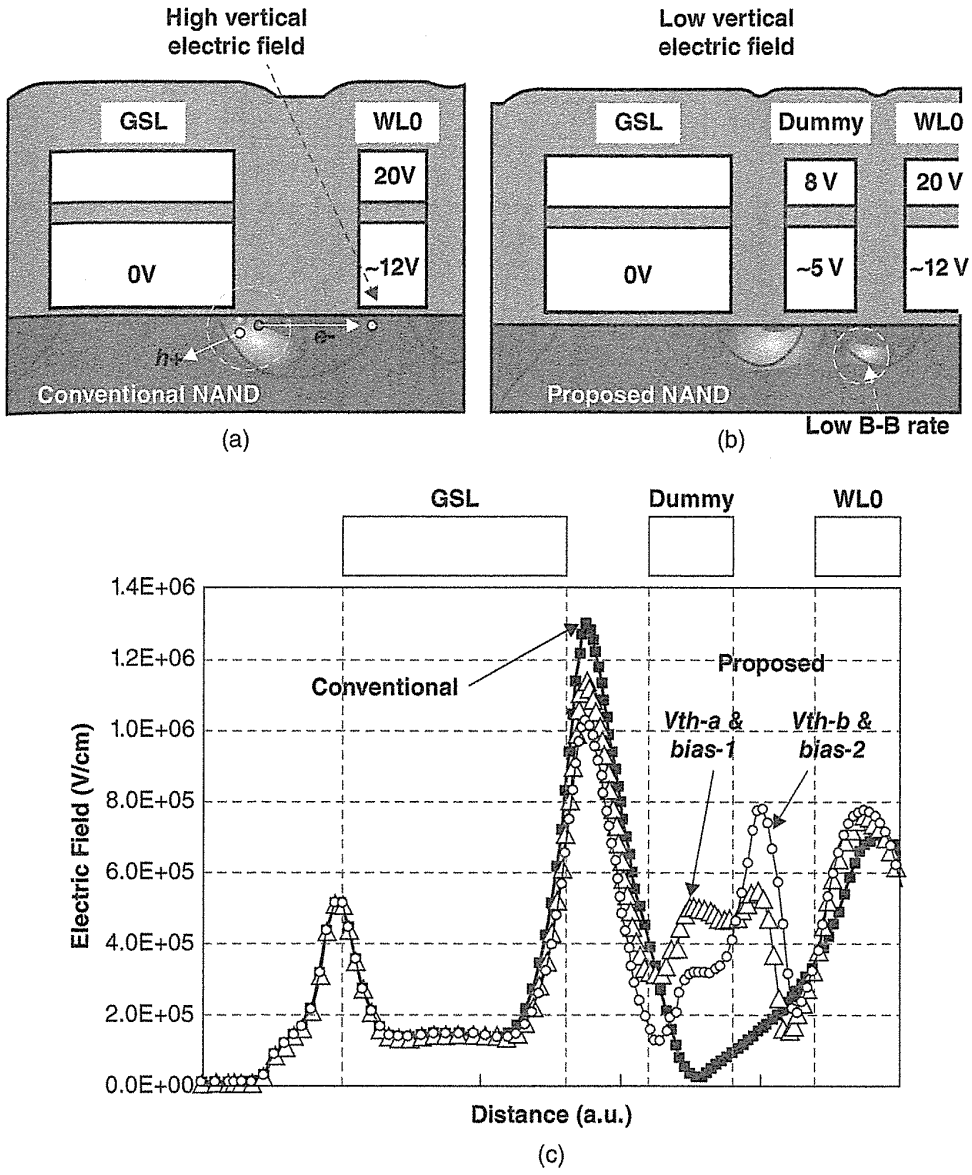


FIGURE 3.62 Simulated band-to-band electron/hole generation rate and lateral electric field during program: (a) Band-to-band electron/hole generation contour in conventional NAND (without dummy word line). (b) Band-to-band electron/hole generation contour with dummy word line scheme. (c) Lateral electric field across structure. Copyright 2007, The Japan Society of Applied Physics.

floating gate of edge cells. Depending on the dummy word-line bias voltage and adjusted V_{th} of the dummy cell, the generated lateral electric field can be reduced further, as shown in Fig. 3.62c.

The dummy cell scheme is also able to shield a memory cell from a high-voltage boosted select gate of GSL/SSL during an erase operation. Figure 3.63 shows a

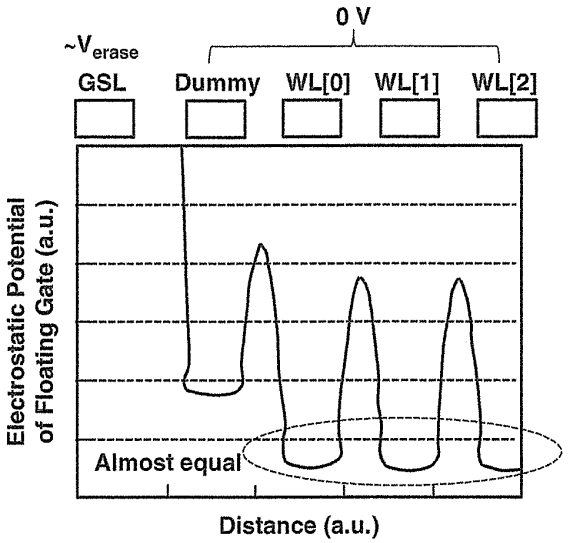


FIGURE 3.63 Simulated electrostatic potential of floating gate of memory cells during erase operation. Copyright 2007, The Japan Society of Applied Physics.

simulated electrostatic potential of a floating gate during an erase operation. The generated voltage of floated GSL during erasing is boosted up to almost the same of a high erase voltage so that the potential of a floating gate of dummy cell becomes slightly higher than that of middle cells due to capacitive coupling between GSL and the dummy cell. Thanks to the shielding effect of a dummy cell, the potentials of a floating gate of edge cells are almost equal to middle cells. Then it leads to improve the erase V_{th} distribution width. Figure 3.64 shows the measured V_{th} distributions of the erase state for each WL in the conventional and dummy word-line scheme. The erased V_{th} distribution of edge WLs in the conventional NAND is as high as 0.5–1.2 V compared to middle WLs. The erase V_{th} distribution is about 1.65 V wide. By using the dummy WL scheme, the difference of erased V_{th} distribution between edge WLs and middle WLs becomes negligible, as shown in Fig. 3.64(b). Thus, the erase V_{th} distribution of dummy word-line scheme is about 1.1 V, which is about 31% narrow width. This leads to better V_{th} distribution of programmed state cells compared to the conventional memory cells.

3.7.2 The P-Type Floating Gate

The *n*-type phosphorus-doped poly-Si floating gate is a legacy process from an initial production of NAND flash memory in 1992. As *n*-type poly-Si has several advantages of better dopant controllability, a better scalability of a surface channel nMOS cell, and a low sheet resistance for a select gate. Especially, in the NAND cell, it was important for *n*-type poly-Si layer to have lower sheet resistance due to a short RC (resistance and capacitance) delay of a select gate in a LOCOS cell and an SA-STI

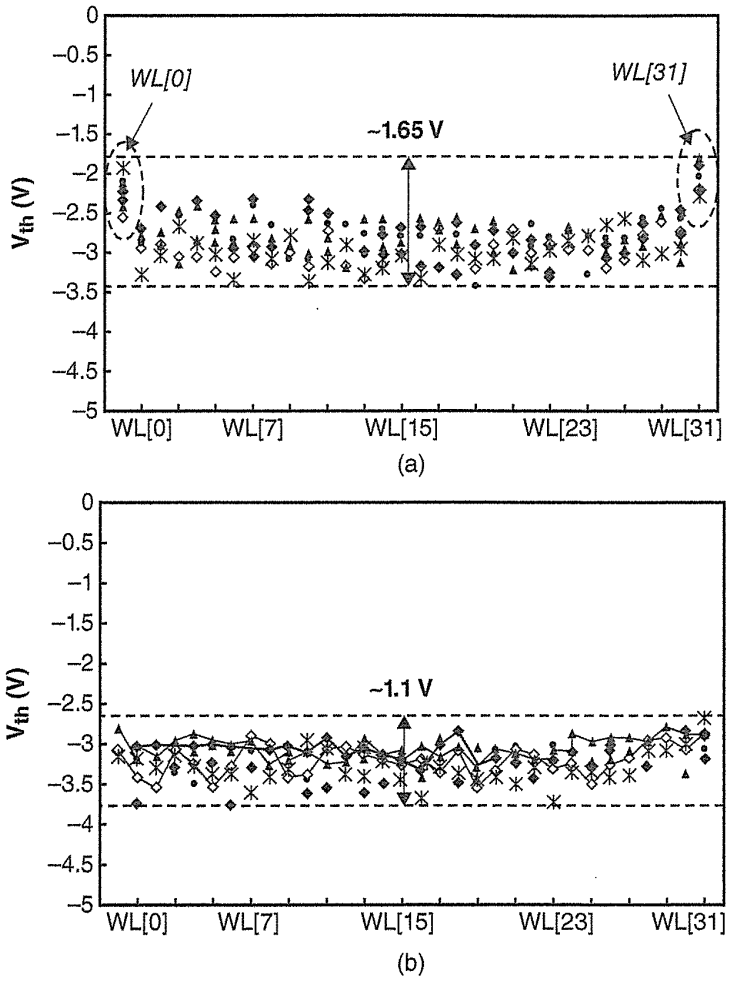


FIGURE 3.64 Measured erased V_{th} of memory cells in (a) conventional NAND flash memory cell without dummy word lines and (b) NAND flash memory cell with dummy word lines. Copyright 2007, The Japan Society of Applied Physics.

cell with FG wing (Section 3.2 and 3.3). Due to higher sheet resistance, *p*-type poly-Si could not be used for a NAND flash memory cell. However, in the SA-STI cell without FG wing, a high sheet resistance of a floating gate is not a problem because a floating gate and a control gate are directly connected as forming select gate transistor and peripheral transistors.

It had been reported that the *p*-type floating gate had an advantage to improve the data retention of flash memory cells [18]. However, the depletion effect of a *p*-type floating gate is not negligible because it is hard to maintain the required doping concentrations after subsequent heat budget processes because of faster inherent boron segregation compared with *n*-type phosphorus-doped poly-Si. Thus, the doping concentration of boron in the *p*-type floating gate is normally several times lower than that in an *n*-type floating gate. If the doping concentration is insufficient,

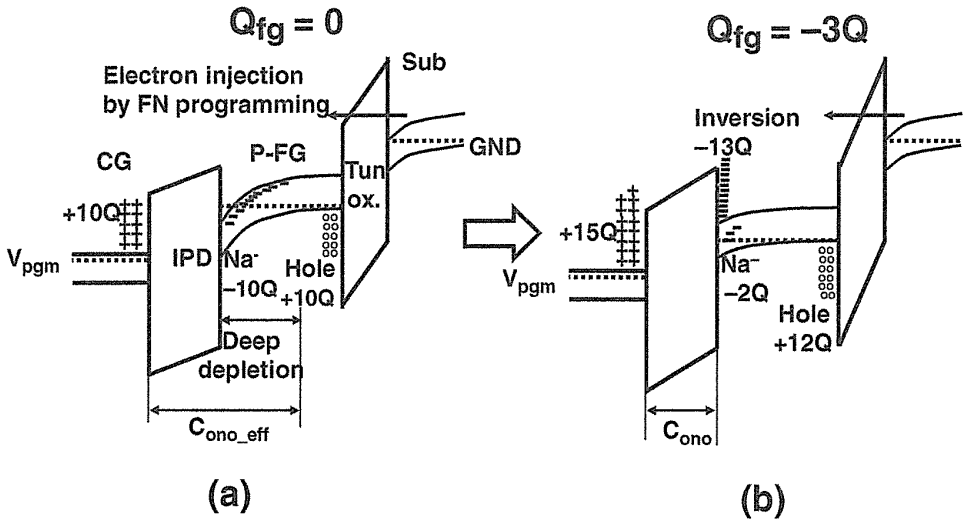


FIGURE 3.65 Schematic band diagram of a transient deep-depletion phenomenon in a *p*-type floating gate cell in (a) $Q_{fg} = 0$ and (b) $Q_{fg} = -3Q$. The charge values are not calibrated, just for conceptual illustrations.

the program speed is degraded due to a transient deep depletion phenomenon in a floating gate.

A transient deep depletion behavior has an impact on program and erase operations, which are based on a model of a nonequilibrium deep depletion phenomenon [38]. Figure 3.65 shows a conceptual model of the transient deep-depletion phenomenon [19, 20]. In the *p*-type floating gate, the amount of electrons is very low in the equilibrium state; thus when V_{PGM} is applied to a control gate (CG), negative charges in a floating gate are not available at the IPD/floating gate interface. Then, deep-depletion occurs and extends more deeply into the floating gate, as shown in Fig. 3.65a. In the nonequilibrium condition, the conduction band energy at the IPD/floating gate interface is much lower than that of the Fermi level of the floating gate, thus a large voltage drop occurs, resulting in the loss in the coupling ratio. There are several ways to break the deep-depletion conditions, such as electron injection through the tunnel oxide, electron generation by impact ionization by the injected electrons, thermal electron generation by a SRH, and BTBT electron generation by a strong electric field. All these mechanisms have a contribution toward breaking the deep-depletion.

In the case of a high enough *p*-type dopant concentration (N_a > 1E20), program and erase operations are successfully performed with a breaking deep-depletion condition. Figure 3.66 shows experimental program and erase characteristics of an *n*-FG/*n*-CG cell and a *p*-FG/*n*-CG cell in a 42-nm generation cell [19]. The *p*-FG/n-CG cell appears to have a slower erase speed with ~1.5 V than the *n*-FG/*n*-CG cell as mentioned above, while its program speed appears to be faster with ~1 V.

The program/erase cycling endurance of the *p*-type floating gate is better, compared with that of the *n*-type floating gate, as shown in Fig. 3.67 [19] and Fig. 3.68 [20]. The midgap voltage shift by N_{OT} (oxide trapping charge) of the *p*-type floating

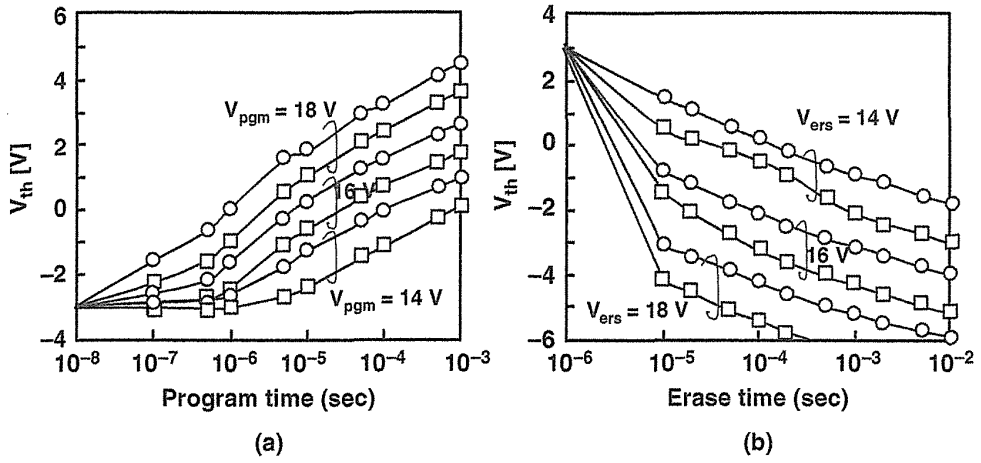


FIGURE 3.66 Experimental program and erase characteristics of *n*-FG and *p*-FG cells (circle is *p*-FG/n-CG and rectangle is *n*-FG/*n*-CG).

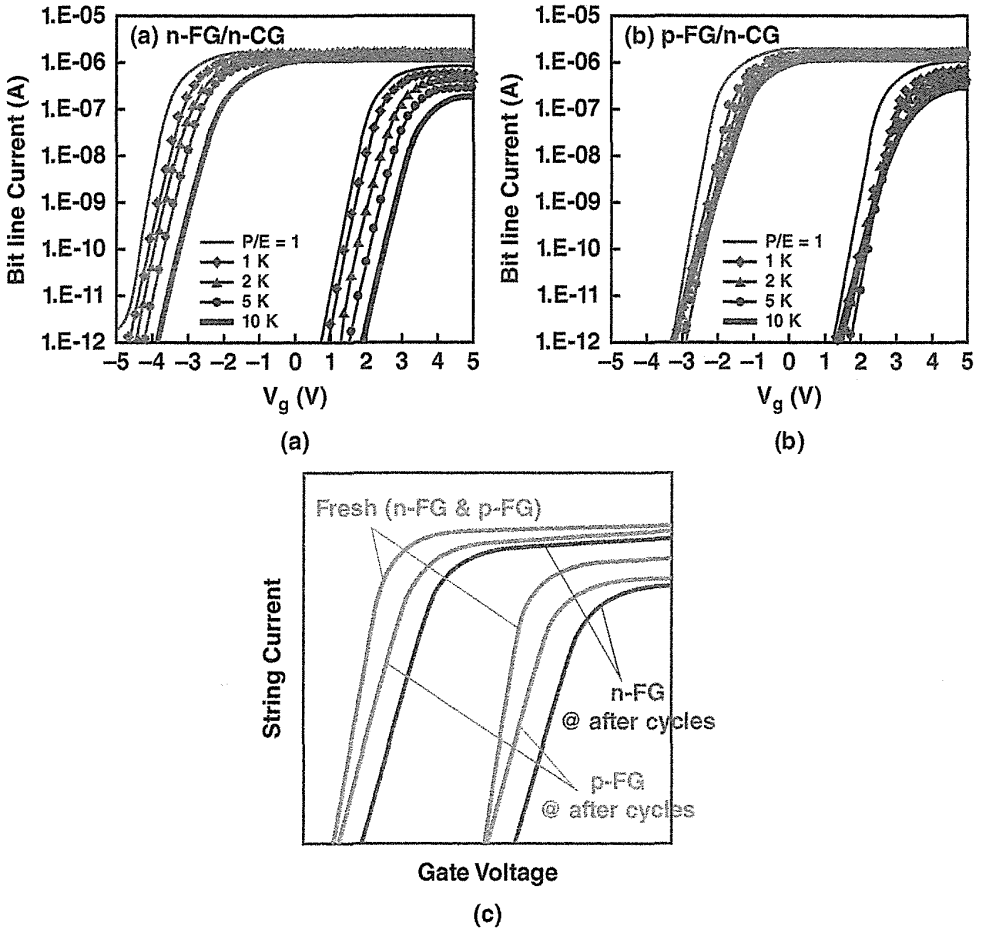


FIGURE 3.67 Experimental I_d–V_g characteristics of (a) *n*-FG and (b) *p*-FG cells ((a) *n*-FG/*n*-CG cell and (b) *p*-FG/*n*-CG cell) before and after P/E cycling. (c) Schematic I_d–V_g curve of cell transistor for *n*-type and *p*-type floating gate before/after P/E cycling.

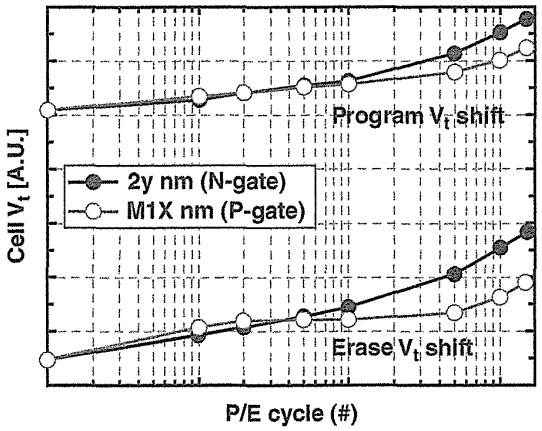


FIGURE 3.68 P/E cycling endurance characteristics. The $p+$ poly gate of the middle 1X nm cell shows an improved cycling endurance compared with 2y nm with p-type gate.

gate is very low, thus only N_{IT} (interface trapping charge) degradation is caused, as shown in Fig. 3.67. This can be explained by the injected electron/hole current ratio during erase operation, which mainly caused the degradation, as shown in Fig. 3.69 [20]. The erase voltage increases in a p-type floating gate because of low electron density at the floating gate, thus the hole current relatively increases to

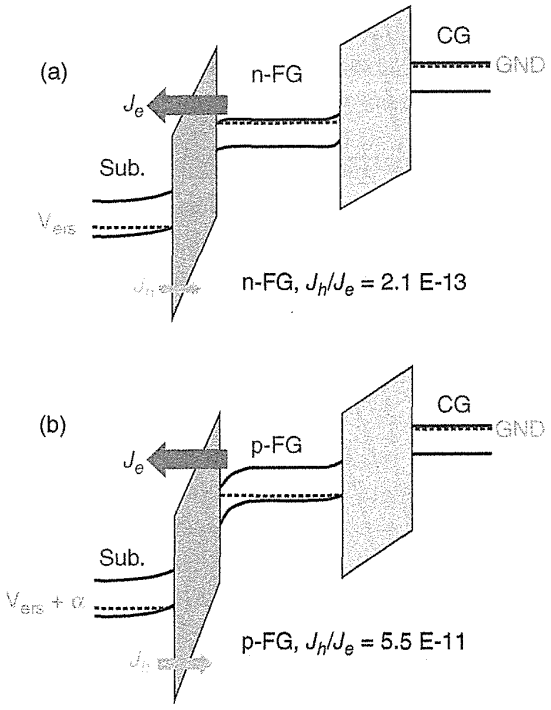


FIGURE 3.69 Schematic illustrations of endurance improvement in the p-type floating gate cell. (a) n-type FG and (b) p-type FG. The ratio of hole current in the p-type floating gate is increased with that in the p-type floating gate.

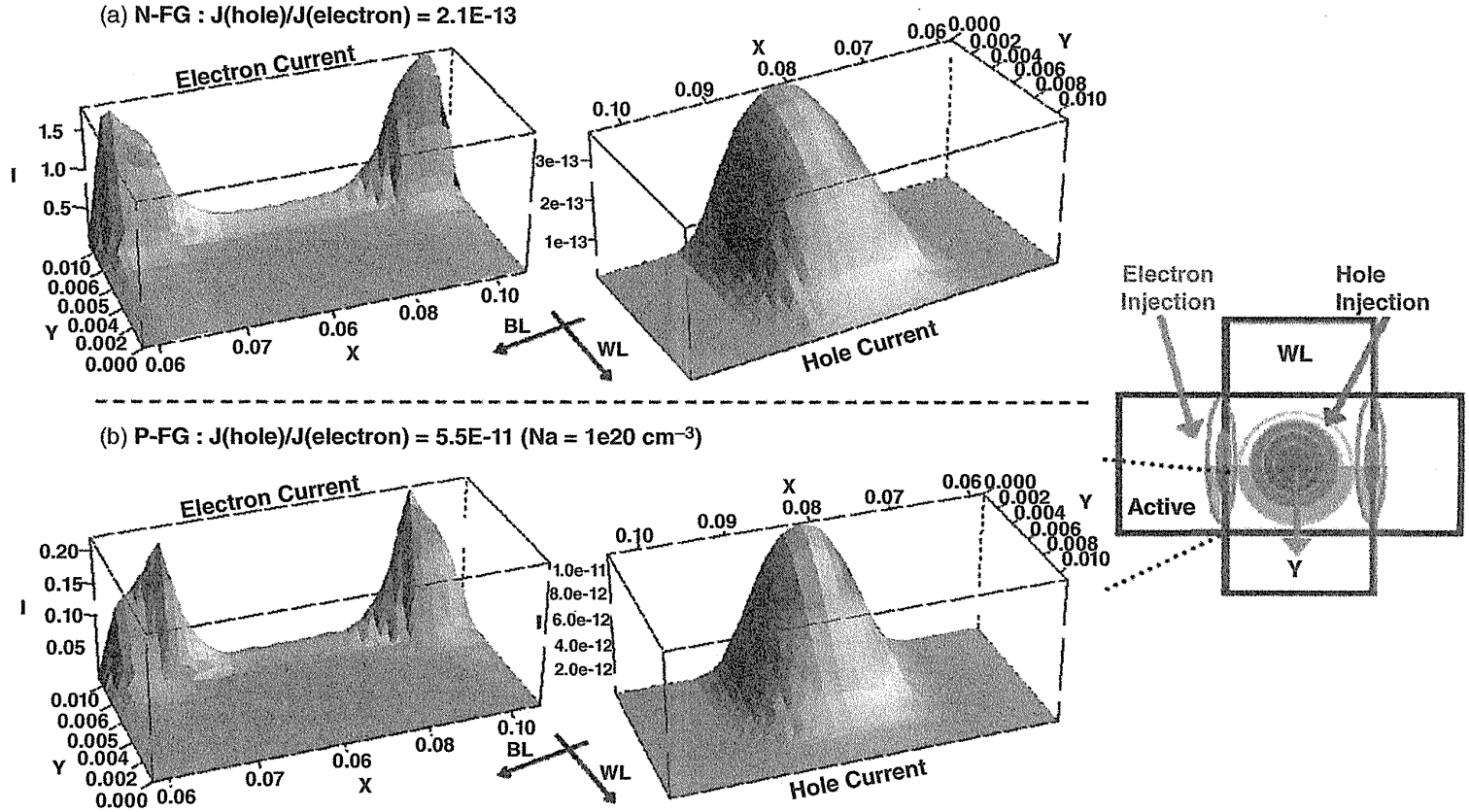


FIGURE 3.70 Distribution of hole/electron erasing tunneling currents with (a) *n*-FG cells and (b) *p*-FG cells with Na = 1e20cm^{-3}.

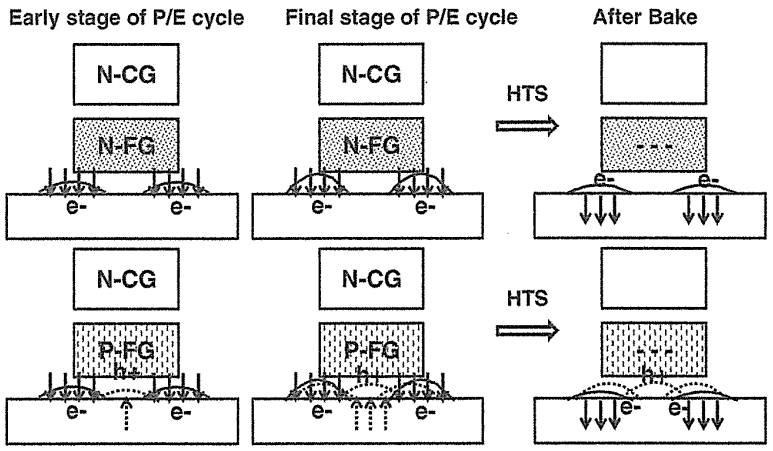


FIGURE 3.71 The model of endurance and data retention (HTS) of a *p*-FG cell.

have the required electron current for erasing, as shown in Fig. 3.69b. The ratio of hole current to the total erasing current increases, which subsequently increases the amount of hole trapping in the tunnel oxide. Therefore, the electron trapping in the tunnel oxide is mostly compensated by the hole trapping, and it results in negligible N_{OT} shift.

The two-dimensional distribution of an electron/hole current during erase operation is simulated, as shown in Fig. 3.70 [19]. The ratio of hole current injected from Si substrate to electron current emitted from FG is found to be 260 times higher in a *p*-type FG cell than in an *n*-type FG cell. With higher *p*-type doping concentration, the balance of both carriers contributing to erase operation becomes stronger for hole tunneling contribution.

The data retention characteristic of a *p*-type floating gate is similar to that of an *n*-type floating gate [19]. It is explained that the electron traps in both *p*-type/*n*-type FG cells are de-trapped from a tunnel oxide in the same manner, but the holes still remained in the hole trap sites without being de-trapped even after high temperature baking, resulting in the same charge loss in both cells, as shown in Fig. 3.71.

As described above, a *p*-type floating gate has better cycling endurance than does an *n*-type floating gate. However, doping type of the control gate has not been well discussed yet. In a realistic process in the case of a *p*-type floating gate, the doping type of the control gate should be *p*-type, because the floating gate has directly connected to the control gate in the select gate in the SA-STI cell. We should avoid to mixing and canceling out dopants of *p*-type and *n*-type. It had been reported that the *p*-type poly-Si is applied to the control gate [21] in a mid-1X-nm generation cell. The paper [21] pointed out a new problem of read bias sensitivity, caused by a severe control gate depletion. The *p*-type control gate, which is located on STI (between floating gates), is fully depleted during read due to low doping. The read bias sensitivity can be solved by increasing doping concentration in both floating gate and control gate [21].

REFERENCES

[1] Masuoka, F.; Momodomi, M.; Iwata, Y.; Shirota, R. New ultra high density EPROM and flash EEPROM with NAND structure cell, *Electron Devices Meeting, 1987 International*, vol. 33, pp. 552– 555, 1987.

[2] Aritome, S.; Hatakeyama, I.; Endoh, T.; Yamaguchi, T.; Shuto, S.; Iizuka, H.; Maruyama, T.; Watanabe, H.; Hemink, G.H.; Tanaka, T.; Momodomi, M.; Sakui, K.; and Shirota, R. A 1.13 um^2 memory cell technology for reliable 3.3 V 64 Mb EEPROMs, *1993 International Conference on Solid State Device and Material (SSDM93)*, pp. 446–448, 1993.

[3] Aritome S.; Hatakeyama I.; Endoh T.; Yamaguchi T.; Shuto S.; Iizuka H.; Maruyama T.; Watanabe H.; Hemink G.; Sakui K.; Tanaka T.; Momodomi, M.; and Shirota R. An advanced NAND-structure cell technology for reliable 3.3 V 64 Mb electrically erasable and programmable read only memories (EEPROMs), *Japanese Journal of Applied Physics*, vol. 33, part 1, no. 1B, pp. 524–528, Jan. 1994.

[4] Shimizu, K.; Narita, K.; Watanabe, H.; Kamiya, E.; Takeuchi, Y.; Yaegashi, T.; Aritome, S.; Watanabe, T. A novel high-density 5F^2 NAND STI cell technology suitable for 256 Mbit and 1 Gbit flash memories, *Electron Devices Meeting, 1997. IEDM '97. Technical Digest, International*, pp. 271–274, 7–10 Dec. 1997.

[5] Takeuchi, Y.; Shimizu, K.; Narita, K.; Kamiya, E.; Yaegashi, T.; Amemiya, K.; Aritome, S. A self-aligned STI process integration for low cost and highly reliable 1 Gbit flash memories, *VLSI Technology, 1998. Digest of Technical Papers. 1998 Symposium on*, pp. 102–103, 9–11 June 1998.

[6] Aritome, S.; Satoh, S.; Maruyama, T.; Watanabe, H.; Shuto, S.; Hemink, G. J.; Shirota, R.; Watanabe, S.; Masuoka, F. A 0.67 μm^2 self-aligned shallow trench isolation cell (SA-STI cell) for 3 V-only 256 Mbit NAND EEPROMs, *Electron Devices Meeting, 1994. IEDM '94. Technical Digest, International*, pp. 61–64, 11–14 Dec. 1994.

[7] Imamiya, K.; Sugiura, Y.; Nakamura, H.; Himeno, T.; Takeuchi, K.; Ikehashi, T.; Kanda, K.; Hosono, K.; Shirota, R.; Aritome, S.; Shimizu, K.; Hatakeyama, K.; Sakui, K. A 130 mm^2 256 Mb NAND flash with shallow trench isolation technology, *Solid-State Circuits Conference, 1999. Digest of Technical Papers. ISSCC. 1999 IEEE International*, pp. 112–113, 1999.

[8] Imamiya, K.; Sugiura, Y.; Nakamura, H.; Himeno, T.; Takeuchi, K.; Ikehashi, T.; Kanda, K.; Hosono, K.; Shirota, R.; Aritome, S.; Shimizu, K.; Hatakeyama, K.; Sakui, K. A 130-mm^2, 256-Mbit NAND flash with shallow trench isolation technology, *Solid-State Circuits, IEEE Journal of*, vol. 34, no. 11, pp. 1536–1543, Nov. 1999.

[9] Aritome, S. Advanced flash memory technology and trends for file storage application, *Electron Devices Meeting, 2000. IEDM Technical Digest. International*, pp. 763–766, 2000.

[10] Goda, A.; Parat, K. Scaling directions for 2D and 3D NAND cells, *Electron Devices Meeting (IEDM), 2012 IEEE International*, pp. 2.1.1, 2.1.4, 10–13 Dec. 2012.

[11] Ramaswamy, N.; Graettinger, T.; Puzzilli, G.; Liu H.; Prall, K.; Gowda, S.; Furnemont, A.; Changhan K.; Parat, K. Engineering a planar NAND cell scalable to 20nm and beyond, *Memory Workshop (IMW), 2013 5th IEEE International*, pp. 5,8, 26–29 May 2013.

[12] Goda, A. Recent progress and future directions in NAND flash scaling, *Non-Volatile Memory Technology Symposium (NVMTS), 2013 13th*, pp. 1,4, 12–14 Aug. 2013.

[13] Aritome, S.; Takeuchi, Y.; Sato, S.; Watanabe, H.; Shimizu, K.; Hemink, G. J.; Shirota, R. A novel side-wall transfer-transistor cell (SWATT cell) for multi-level NAND EEPROM's, in *IEEE IEDM Technical Digest*, pp. 275–278, 1995.

[14] Aritome, S.; Takeuchi, Y.; Sato, S.; Watanabe, I.; Shimizu, K.; Hemink, G.; Shirota, R. A side-wall transfer-transistor cell (SWATT cell) for highly reliable multi-level NAND EEPROMs, *Electron Devices, IEEE Transactions on*, vol. 44, no. 1, pp.145–152, Jan 1997.

[15] Park, K.-T.; Lee, S.C.; Sel, J.-S.; Choi, J.; Kim, K. Scalable wordline shielding scheme using dummy cell beyond 40 nm NAND flash memory for eliminating abnormal disturb of edge memory cell, *SSDM*, pp. 298–299, 2006.

[16] Park, K.-T.; Lee, S.C.; Sel, J.-S.; Choi, J.; Kim, K. Scalable wordline shielding scheme using dummy cell beyond 40 nm NAND flash memory for eliminating abnormal disturb of edge memory cell, *Japanese Journal of Applied Physics*, vol. 46, no. 4B, pp. 2188–2192, 2007.

[17] Kanda, K.; Koyanagi, M.; Yamamura, T.; Hosono, K.; Yoshihara, M.; Miwa, T.; Kato, Y.; Mak, A.; Chan, S.L.; Tsai, F.; Cernea, R.; Le, B.; Makino, E.; Taira, T.; Otake, H.; Kajimura, N.; Fujimura, S.; Takeuchi, Y.; Itoh, M.; Shirakawa, M.; Nakamura, D.; Suzuki, Y.; Okukawa, Y.; Kojima, M.; Yoneya, K.; Arizono, T.; Hisada, T.; Miyamoto, S.; Noguchi, M.; Yaegashi, T.; Higashitani, M.; Ito, F.; Kamei, T.; Hemink, G.; Maruyama, T.; Ino, K.; Ohshima, S. A 120 mm^2 16 Gb 4-MLC NAND flash memory with 43 nm CMOS technology, *Solid-State Circuits Conference, 2008. ISSCC 2008. Digest of Technical Papers. IEEE International*, pp. 430–625, 3–7 Feb. 2008.

[18] Shen, C.; Pu, J.; Li, M.-F.; Cho, J. Byung, P-Type Floating Gate for Retention and P/E Window Improvement of flash memory devices, *Electron Devices, IEEE Transactions on*, vol. 54, no. 8, pp. 1910, 1917, Aug. 2007.

[19] Lee, C.H.; Fayrushin, A.; Hur, S.; Park, Y.; Choi, J.; Choi, J.; Chung, C. Physical modeling and analysis on improved endurance behavior of *p*-type floating gate NAND flash memory, *Memory Workshop (IMW), 2012 4th IEEE International*, pp. 1,4, 20–23 May 2012.

[20] Park, Y.; Lee, J. Device considerations of planar NAND flash memory for extending towards sub-20 nm regime, *Memory Workshop (IMW), 2013 5th IEEE International*, pp. 1,4, 26–29 May 2013.

[21] Seo, J.; Han, K.; Youn, T.; Heo H.-E.; Jang, S.; Kim, J.; Yoo, H.; Hwang, J.; Yang, C.; Lee, H.; Kim, B.; Choi, E.; Noh, K.; Lee, B.; Lee, B.; Chang, H.; Park, S.; Ahn, K.; Lee, S.; Kim, J.; Lee, S. Highly reliable M1X MLC NAND flash memory cell with novel active air-gap and *p*+ poly process integration technologies, *Electron Devices Meeting (IEDM), 2013 IEEE International*, pp. 3.6.1,3.6.4, 9–11 Dec. 2013.

[22] Aritome, S.; Kirisawa, R.; Endoh, T.; Nakayama, R.; Shirota, R.; Sakui, K.; Ohuchi, K.; Masuoka, F. Extended data retention characteristics after more than 10^4 write and erase cycles in EEPROMs, *International Reliability Physics Symposium, 1990. 28th Annual Proceedings*, 1990, pp. 259–264, 1990.

[23] Kirisawa, R.; Aritome, S.; Nakayama, R.; Endoh, T.; Shirota, R.; Masuoka, F.; A NAND structured cell with a new programming technology for highly reliable 5 V-only flash EEPROM, *1990 Symposium on VLSI Technology, 1990. Digest of Technical Papers*, 1990, pp. 129–130, 1990.

[24] Aritome, S.; Shirota, R.; Kirisawa, R.; Endoh, T.; Nakayama, R.; Sakui, K.; Masuoka, F.; A reliable bi-polarity write/erase technology in flash EEPROMs, *International Electron Devices Meeting, 1990. IEDM '90. Technical Digest, 1990*, pp. 111–114, 1990.

[25] Aritome, S.; Shirota, R.; Hemink, G.; Endoh, T.; Masuoka, F.; Reliability issues of flash memory cells, *Proceedings of the IEEE*, vol. 81, no. 5, pp. 776–788, 1993.

[26] Tanaka, T.; Tanaka, Y.; Nakamura, H.; Oodaira, H.; Aritome, S.; Shirota, R.; Masuoka, F. A quick intelligent program architecture for 3 V-only NAND-EEPROMs, *VLSI Circuits, 1992. Digest of Technical Papers, 1992 Symposium on*, pp. 20–21, 4–6 June 1992.

[27] Choi, J.-D.; Lee, J.-H.; Lee, W.-H.; Shin, K.-S.; Yim, Y.-S.; Lee, J.-D.; Shin, Y.-C.; Chang, S.-N.; Park, K.-C.; Park, J.-W.; Hwang, C.-G. A 0.15 μm NAND flash technology with 0.11 μm^2 cell size for 1 Gbit flash memory," *Electron Devices Meeting, 2000. IEDM '00. Technical Digest. International*, pp. 767,770, 10–13 Dec. 2000.

[28] Arai, F.; Arai, N.; Satoh, S.; Yaegashi, T.; Kamiya, E.; Matsunaga, Y.; Takeuchi, Y.; Kamata, H.; Shimizu, A.; Ohtami, N.; Kai, N.; Takahashi, S.; Moriyama, W.; Kugimiya, K.; Miyazaki, S.; Hirose, T.; Meguro, H.; Hatakeyama, K.; Shimizu, K.; Shirota, R. High-density ($4.4F^2$) NAND flash technology using super-shallow channel profile (SSCP) engineering, *Electron Devices Meeting, 2000. IEDM '00. Technical Digest International*, pp. 775,778, 10–13 Dec. 2000.

[29] Choi, J.-D.; Cho, S.-S.; Yim, Y.-S.; Lee, J.-D.; Kim, H.-S.; Joo, K.-J.; Hur, S.-H.; Im, H.-S.; Kim, J.; Lee, J.-W.; Seo, K.-I.; Kang, M.-S.; Kim, K.-H.; Nam, J.-L.; Park, K.-C.; Lee, M.-Y. Highly manufacturable 1 Gb NAND flash using 0.12 μm process technology, *Electron Devices Meeting, 2001. IEDM '01. Technical Digest International*, pp. 2.1.1,2.1.4, 2–5 Dec. 2001.

[30] Kim, D.-C.; Shin, W.-C.; Lee, J.-D.; Shin, J.-H.; Lee, J.-H.; Hur, S.-H.; Baik, I.-G.; Shin, Y.-C.; Lee, C.-H.; Yoon, J.-S.; Lee, H.-G.; Jo, K.-S.; Choi, S.-W.; You, B.-K.; Choi, J.-H.; Park, D.; Kim, K. A 2 Gb NAND flash memory with 0.044 μm^2 cell size using 90 nm flash technology, *Electron Devices Meeting, 2002. IEDM '02. International*, pp. 919,922, 8–11 Dec. 2002.

[31] Ichige, M.; Takeuchi, Y.; Sugimae, K.; Sato, A.; Matsui, M.; Kamigaichi, T.; Kutsukake, H.; Ishibashi, Y.; Saito, M.; Mori, S.; Meguro, H.; Miyazaki, S.; Miwa, T.; Takahashi, S.; Iguchi, T.; Kawai, N.; Tamon, S.; Arai, N.; Kamata, H.; Minami, T.; Iizuka, H.; Higashitani, M.; Pham, T.; Hemink, G.; Momodomi, M.; Shirota, R. A novel self-aligned shallow trench isolation cell for 90 nm 4 Gbit NAND flash EEPROMs, *VLSI Technology, 2003. Digest of Technical Papers. 2003 Symposium on*, pp. 89,90, 10–12 June 2003.

[32] Noguchi, M.; Yaegashi, T.; Koyama, H.; Morikado, M.; Ishibashi, Y.; Ishibashi, S.; Ino, K.; Sawamura, K.; Aoi, T.; Maruyama, T.; Kajita, A.; Ito, E.; Kishida, M.; Kanda, K.; Hosono, K.; Miyamoto, S.; Ito, F.; Hemink, G.; Higashitani, M.; Mak, A.; Chan, J.; Koyanagi, M.; Ohshima, S.; Shibata, H.; Tsunoda, H.; Tanaka, S. A high-performance multi-level NAND flash memory with 43 nm-node floating-gate technology, *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, pp. 445, 448, 10–12 Dec. 2007.

[33] Kamigaichi, T.; Arai, F.; Nitsuta, H.; Endo, M.; Nishihara, K.; Murata, T.; Takekida, H.; Izumi, T.; Uchida, K.; Maruyama, T.; Kawabata, I.; Suyama, Y.; Sato, A.; Ueno, K.; Takeshita, H.; Joko, Y.; Watanabe, S.; Liu, Y.; Meguro, H.; Kajita, A.; Ozawa, Y.; Watanabe, T.; Sato, S.; Tomiie, H.; Kanamaru, Y.; Shoji, R.; Lai, C.H.; Nakamichi, M.; Oowada, K.; Ishigaki, T.; Hemink, G.; Dutta, D.; Dong, Y.; Chen, C.; Liang, G.; Higashitani, M.;

Lutze, J. Floating Gate super multi level NAND Flash Memory Technology for 30 nm and beyond, *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pp. 1,4, 15–17 Dec. 2008.

[34] Lee, C.-H.; Sung, S.-K.; Jang, D.; Lee, S.; Choi, S.; Kim, J.; Park, S.; Song, M.; Baek, H.-C.; Ahn, E.; Shin, J.; Shin, K.; Min, K.; Cho, S.-S.; Kang, C.-J.; Choi, J.; Kim, K.; Choi, J.-H.; Suh, K.-D.; Jung, T.-S. A highly manufacturable integration technology for 27 nm 2 and 3bit/cell NAND flash memory, *Electron Devices Meeting (IEDM), 2010 IEEE International*, pp. 5.1.1,5.1.4, 6–8 Dec. 2010.

[35] Hwang, J.; Seo, J.; Lee, Y.; Park, S.; Leem, J.; Kim, J.; Hong, T.; Jeong, S.; Lee, K.; Heo, H.; Lee, H.; Jang, P.; Park, K.; Lee, M.; Baik, S.; Kim, J.; Kkang, H.; Jang, M.; Lee, J.; Cho, G.; Lee, J.; Lee, B.; Jang, H.; Park, S.; Kim, J.; Lee, S.; Aritome, S.; Hong, S. and Park, S. A middle-1X nm NAND flash memory cell (M1X-NAND) with highly manufacturable integration technologies, *Electron Devices Meeting (IEDM), 2011 IEEE International*, pp. 199–202, Dec. 2011.

[36] Govoreanu, B.; Brunco, D. P.; Van Houdt, J. Scaling down the interpoly dielectric for next generation flash memory; Challenges and opportunities, *Solid-State Electronics*, vol. 49, pp. 1841–1848, Nov. 2005.

[37] Lee, J. D.; Lee, C. K.; Lee, M. W; Kim, H. S.; Park, K. C.; Lee, W. S. A new programming disturbance phenomenon in NAND flash memory by source/drain hot-electrons generated by GIDL current, *NVSMW*, pp. 31–33, 2006.

[38] Spessot, A.; Monzio Compagnoni, C.; Farina, F.; Calderoni, A.; Spinelli, A. S.; Fantini P. Effect of floating-gate polysilicon depletion on the erase efficiency of nand flash memories, *Electron Device Letters, IEEE*, vol. 31, no. 7, pp. 647, 649, July 2010.

[39] Takeuchi, K.; Kameda, Y.; Fujimura, S.; Otake, H.; Hosono, K.; Shiga, H.; Watanabe, Y.; Futatsuyama, T.; Shindo, Y.; Kojima, M.; Iwai, M.; Shirakawa, M.; Ichige, M.; Hatakeyama, K.; Tanaka, S.; Kamei, T.; Fu, J.Y.; Cernea, A.; Li, Y.; Higashitani, M.; Hemink, G.; Sato, S.; Oowada, K.; Lee S.-C.; Hayashida, N.; Wan, J.; Lutze, J.; Tsao, S.; Mofidi, M.; Sakurai, K.; Tokiwa, N.; Waki, H.; Nozawa, Y.; Kanazawa, K.; Ohshima, S. A 56 nm CMOS 99 mm^2 8Gb Multi-level NAND Flash Memory with 10MB/s Program Throughput, *Solid-State Circuits Conference, 2006. ISSCC 2006. Digest of Technical Papers. IEEE International*, pp. 507–516, 6–9 Feb. 2006.

4

ADVANCED OPERATION FOR MULTILEVEL CELL

4.1 INTRODUCTION

In order to reduce the cost per bit of flash memory, the multilevel memory cell technologies had been intensively developed [1–7], along with reduced memory cell size [8] (see Chapter 3). The multilevel cell technology was initially developed for MLC (2 bits/cell), but it was extended to TLC (3 bits/cell) and QLC (4 bits/cell). The chip size can be reduced to about 60% by using an MLC (2 bits/cell) scheme, compared with a single-level cell SLC (1 bit/cell) scheme. However, in a multilevel memory cell, a narrow threshold voltage (V_t) distribution width is necessary to have a enough margin between V_t distributions. Due to this narrow V_t distribution width, the programming time of the multilevel cell becomes longer than that of a conventional SLC. Also, reliability of the multilevel cell is worse than that of SLC due to less V_t margin (read V_t window margin). To avoid these problems, it is very important that the V_t distribution width be controlled to be as narrow as possible.

The memory cell structure and fabrication process of the multilevel cells are basically the same as that of SLC. Therefore, multilevel cell technology has been developed to focus on the operations of making narrow V_t distribution width. A lot of sophisticated techniques have been proposed and implemented to a NAND flash memory product [9]. Section 4.2 describes these techniques, such as the incremental step pulse program (ISPP), bit-by-bit verify operations, a two-step verify scheme, and a pseudo-pass scheme.

Nand Flash Memory Technologies, First Edition. Seiichi Aritome.

Even if narrow V_t distribution width is made during page programming, the V_t distribution is disturbed and is getting wider after programming neighbor cells due to the floating-gate capacitive coupling interference (cell-to-cell interference), and so on. Section 4.3 describes several page program sequences to reduce the effect of floating-gate capacitive coupling.

TLC (3 bits/cell) and QLC (4 bits/cell) technologies are described in Section 4.4 and Section 4.5, respectively. The three-level cell technology is introduced in Section 4.6 to compromise the performance and reliability of SLC and MLC.

Finally, in Section 4.7, the moving read algorithm is presented to compensate a V_t shift for minimizing a bit failure rate.

4.2 PROGRAM OPERATION FOR TIGHT V_t DISTRIBUTION WIDTH

4.2.1 Cell V_t Setting

Figure 4.1 shows the image of threshold voltage (V_t) setting for one program state. In order to avoid failure, V_t distribution width has to be tight enough, and a tail of distribution has to have enough margins from read voltage. However, by scaling memory cell size, V_t distribution width becomes much wider by several physical mechanisms, such as floating-gate capacitive coupling (FGC) interference, random telegraph signal noise (RTN), program electron injection spread (EIS), back pattern dependence, and so on, as shown in Fig. 4.1 (see Chapter 5 for details). The operation margins have been decreased as scaling memory cell, because each physical phenomenon becomes worse as scaling.

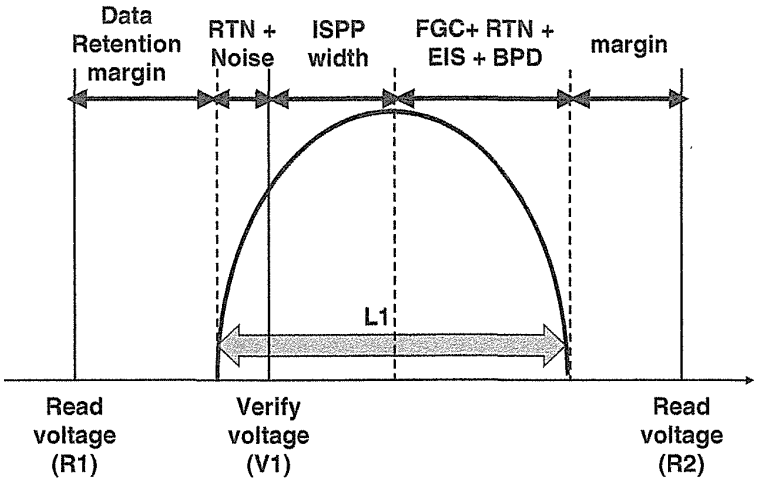


FIGURE 4.1 The image of threshold voltage (V_t) setting for one program state.

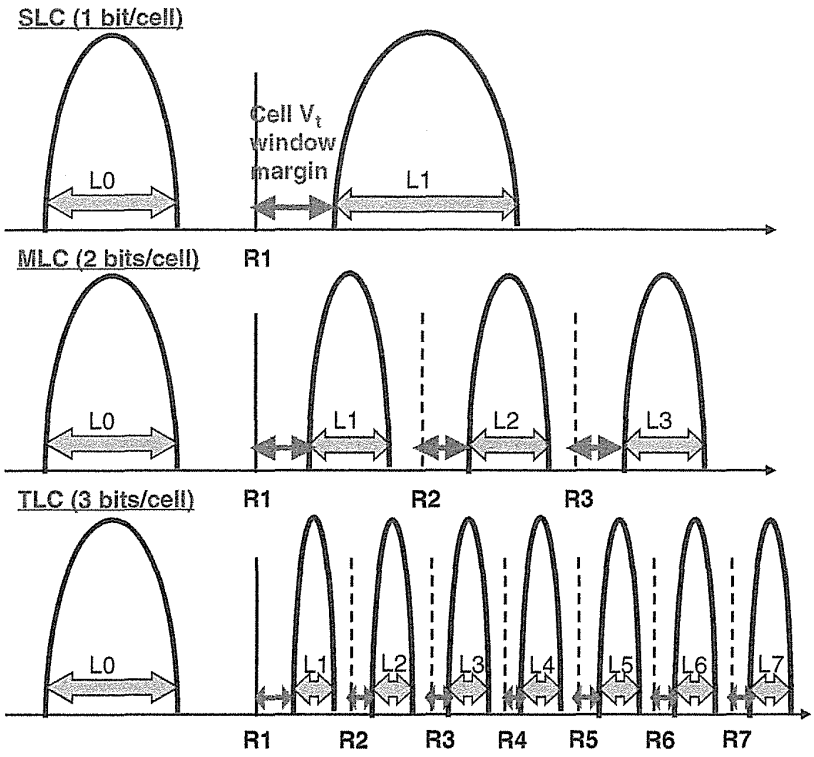


FIGURE 4.2 V_t distribution image of SLC (1 bit/cell), MLC (2 bits/cell), and TLC (3 bits/cell).

Figure 4.2 shows a V_t distribution image of SLC (1 bit/cell), MLC (2 bits/cell), and TLC (3 bits/cell). SLC has a wider cell V_t window margin, then SLC has a better reliability performance and also a better program and read performance than MLC and TLC. MLC and TLC have a very narrow margin to manage good enough reliability. To obtain a wider margin, it is important to make a tight V_t distribution width. Figure 4.3 shows one example of MLC threshold voltage (V_{th}) distributions of the four cell states [5]. Erase "11" cells are sufficiently "deep," and erase V_{th} distribution width is not needed to be controlled as tightly as the three program states. Each program state has a 0.4-V V_{th} distribution width and a 0.8-V margin separating them. The measured V_{th} distribution in a 0.4-μm cell is shown in Fig. 4.4. [5], which demonstrates that the V_{th} optimization results in a relatively tight 0.4 V V_{th} distribution width per state at normal operating condition.

4.2.2 Incremental Step Pulse Program (ISPP)

In order to make the tight programmed V_t distribution width, an incremental step pulse program (ISPP) scheme had been proposed [10, 11].

An incremental step pulse program (ISPP) scheme [10] (step-up programming scheme [11]) is shown in Fig. 2.13 (Chapter 2). The program pulses (V_{pgm}) are

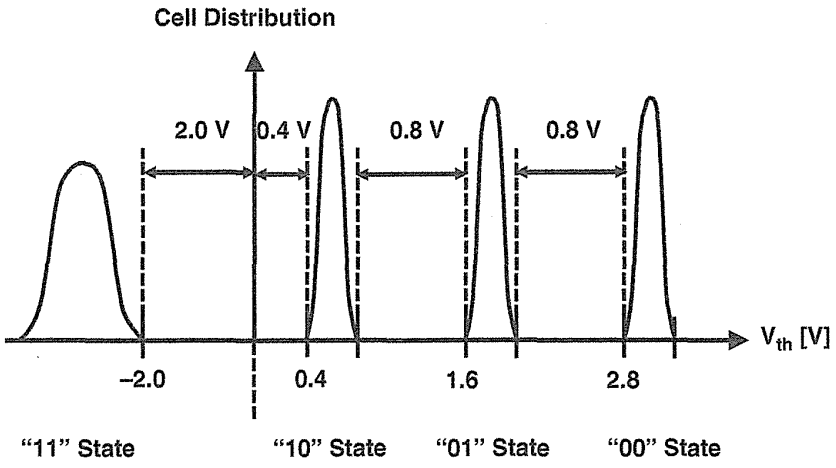


FIGURE 4.3 Target threshold voltage distribution of four states for MLC.

stepped up by ΔV_{pgm}. The ISPP scheme was compared with other schemes, as shown in Fig. 4.5 [11]. The conventional programming pulse (Fig. 4.5a) is the repeating pulses of the same program voltage V_{pp} (= V_{pgm}). There is a problem of increasing program time because many pulses are required to complete a page programming. On the other hand, in step-up program pulses (Fig. 4.5c), program speed can be drastically improved because the slow cells in page can be programmed by higher V_{pp}, and then page programming can be completed by the small number of program pulses.

In the ISPP scheme, tight programmed V_t distribution width can be obtained by using narrower step ΔV_{pp} (= ΔV_{pgm}) without increasing the number of program pulses, as shown in Fig. 4.6 [11]. Furthermore, the ISPP scheme has another important advantage. During programming pulse, the electric field in tunnel oxide can be reduced by the lower starting V_{pp} in comparison with the conventional

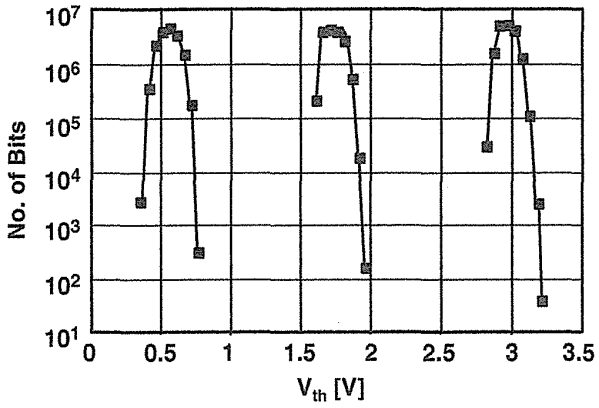


FIGURE 4.4 Measured V_{th} distribution of three programmed states for MLC.

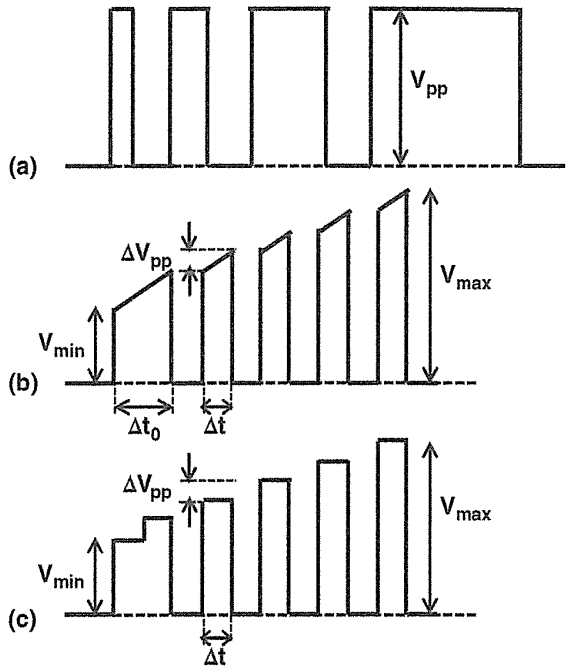


FIGURE 4.5 Program pulse waveforms: (a) conventional, (b) trapezoidal, and (c) staircase (incremental step pulse programming (ISPP)). A verify step is carried out after each pulse.

programming pulse. The degradation of tunnel oxide can be suppressed, and then the reliabilities of program/erase cycling, data retention, and read disturb can be greatly improved [12].

The ISPP scheme has been used in NAND flash memory products for a long time, more than 20 years, due to fast programming speed, tight V_t distribution, and excellent reliability.

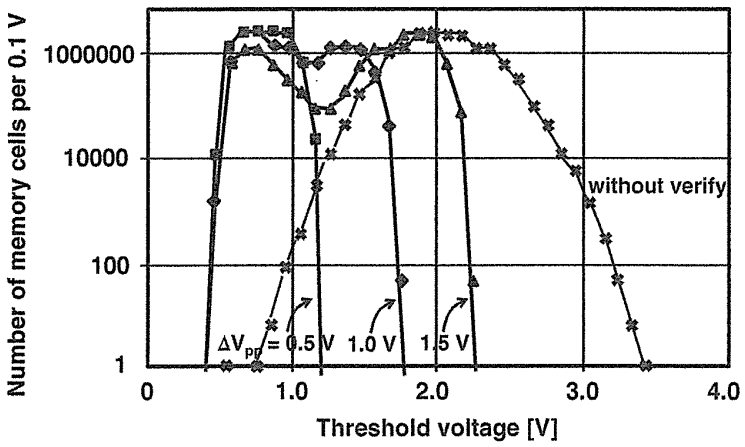


FIGURE 4.6 V_t distribution after programming in a 16-Mbit memory array, with/without verify, using staircase pulses (ISPP) with length of 20 μs and V_{pgm} step of 0.5, 1.0, 1.5 V.

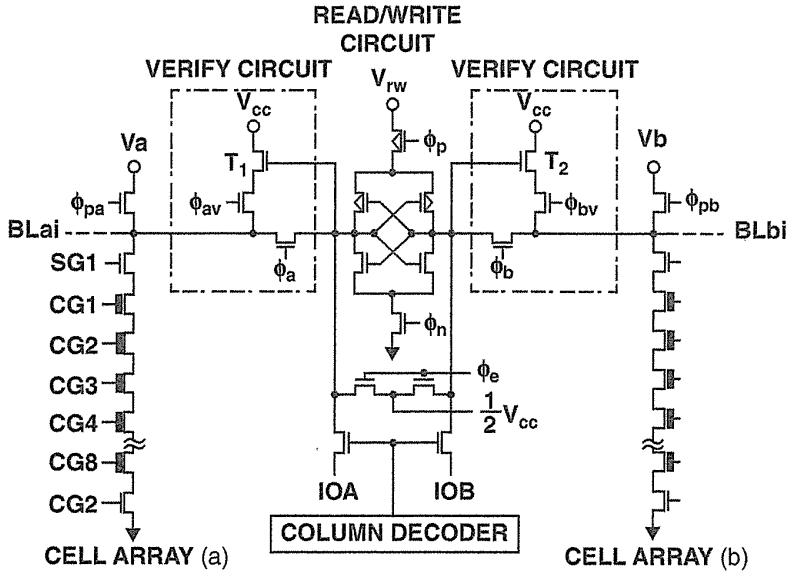


FIGURE 4.7 Intelligent verify circuit schematic for the bit-by-bit verify operation.

4.2.3 Bit-by-Bit Verify Operations

Another important and basic operation technique to make tight V_t distribution is the bit-by-bit verify operation.

An intelligent quick bit-by-bit verify circuit was proposed [13, 14] to realize fast page programming speed as well as tight programmed V_t distribution width. The new verify circuit is composed of adding only two transistors (T1, T2) to a conventional circuit, as shown in Fig. 4.7 [13, 14]. The program/verify operation could be much simplified in comparison with the conventional chip external verify operation. Detail operations are described in the following.

After the program operation, a verify operation is performed to detect the memory cells which require more time to reach the "1" programmed state. In the verify operation, the program data latched in the R/W (read/write) circuit is modified to the re-program data, according to the data modification rule shown in Fig. 4.8. As a result, a re-program operation is performed only on the memory cells which did not reach the "1"-programmed state.

In case of the program data "0" in the R/W circuit latch, the state of the transistor T1 in the verify circuit is "ON" (see Fig. 4.7). The bit line after "0"-programming is re-charged over $^1/_2 V_{cc}$ by the verify circuit. Therefore, the latched re-programmed data is "0" independent of the memory cell data in Fig. 4.8a,b.

In the case of the program data "1" in the R/W circuit latch, the state of the transistor T1 is "OFF." So the bit lines are not re-charged by the verify circuits even if the clock ϕ_{av} turns high. If the memory cell has been successfully programmed "1," the bit-line voltage after "1"-programming is over $^1/_2 V_{cc}$ ((d) in Fig. 4.8). On the other hand, if the memory cell does not reach the "1"-programmed state, the bit-line voltage decreases below $^1/_2 V_{cc}$ (in Fig. 4.8c). The latched re-program data is

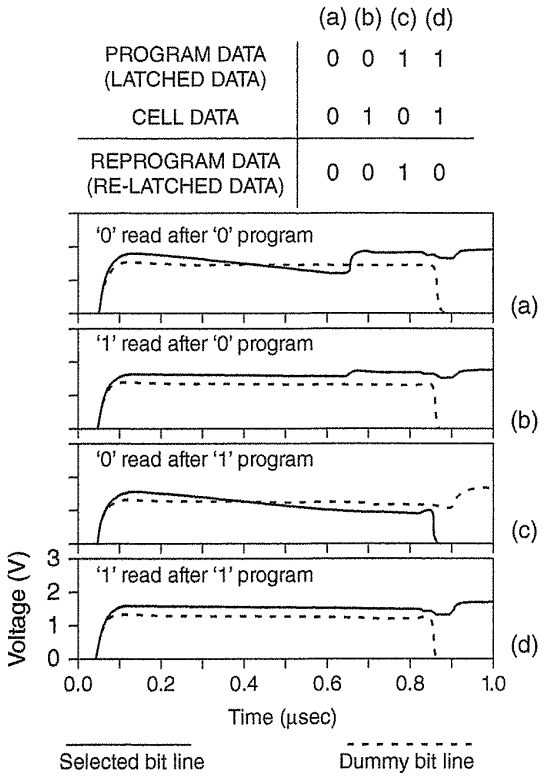


FIGURE 4.8 Data modification rule and simulated waveform for the bit-by-bit verify operation.

"0" for the memory cell which is in the "1"-programmed state (in Fig. 4.8d). The re-programmed data is "1" for the memory cell which did not reach the "1"-programmed state yet (in Fig. 4.8c).

By using the verify circuit, the program data is automatically and simultaneously modified to the re-program data according to Fig. 4.8.

The programmed V_t distribution could be tight with quick verify operation, and programming speed became fast due to chip internal verify operation, which replaced conventional chip external verify operation.

4.2.4 Two-Step Verify Scheme

To achieve a tight programmed V_{th} distribution width, it is important to control the cell V_{th} movement during ISPP program operation. The two-step verify scheme in program verify read operation is widely used [15] for a multilevel cell (MLC, TLC, QLC) to control V_t movement, as shown in Fig. 4.9a–c [16] and Fig. 4.10 [15]. In the two-step verify scheme, two times verify read operations are performed for each program level of P1–P7 in TLC (3 bits/cell), as shown in Fig. 4.9c. For example, for the program level of P1, two times verify read of first P1V and second P1V are performed (in Fig. 4.10, First step write verify voltage and Second step write verify

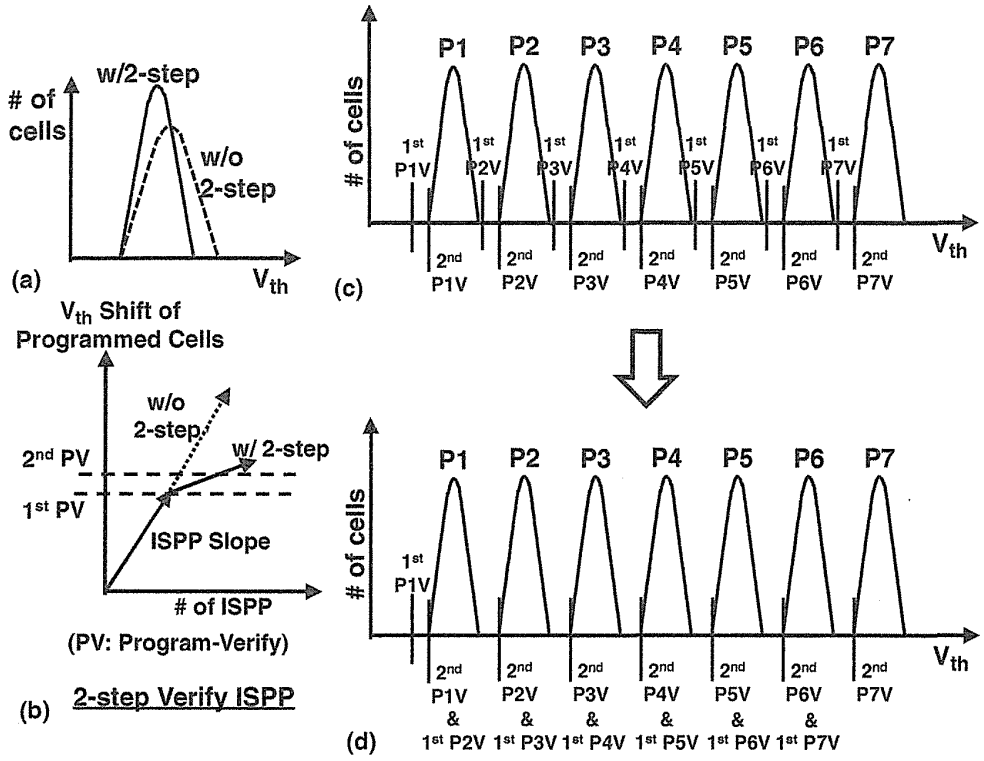


FIGURE 4.9 (a) V_t distribution image of a "with two-step verify" scheme and a "without two-step verify" scheme. (b) V_t shift of program cell in two-step verify scheme. (c) Two-step verify scheme. (d) Verify-skip two-step verify scheme.

voltage). The second P1V is the target verify voltage, and the first P1V is slightly lower than the target verify level. For cells that have V_{th} < first P1V, 0 V is applied to the bit line (BL) during the next program pulse to program (make V_{th} shift) normally. For cells that have V_{th} > second P1V, V_{cc} is applied to the bit line (BL) during the next program pulse to be the inhibit condition. For cells that have first P1V < V_{th} < second P1V, a predetermined low voltage of V_{fbl} (= 0.4 V, for example, in Fig. 4.10) is applied to the bit line (BL) during the next program pulse to make the smaller V_{th} shift than ISPP step voltage, as shown in Fig. 4.9b and Fig. 4.10. Due to the smaller V_t shift for the cells of just below target verify voltage (first P1V < V_{th} < second P1V), the programmed V_{th} distribution width of the two-step verify scheme can be tighter than that of the conventional verify scheme.

The two-step verify scheme, however, requires two times more verify operations for each target V_{th} state, causing an increase in program time. This is especially exaggerated for 3 bits/cell (TLC) NAND, where over 66% of total program time is spent in the verify operation. In order to reduce the extra verify overhead time, a verify-skip two-step tunneling ISPP scheme was proposed [16], as shown in Fig. 4.9d and Fig. 4.11. The verify-skip two-step tunneling ISPP scheme uses the second verify level of the previous target state as the first step verify for next target state. To obtain

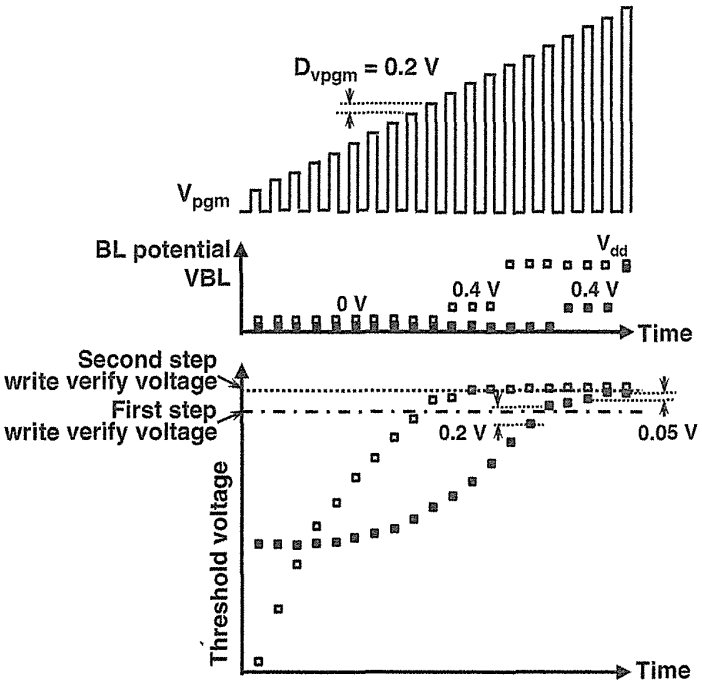


FIGURE 4.10 The two-step verify scheme of program waveform V_{pgm}, bit-line voltage V_{bl}, and V_{th} movement.

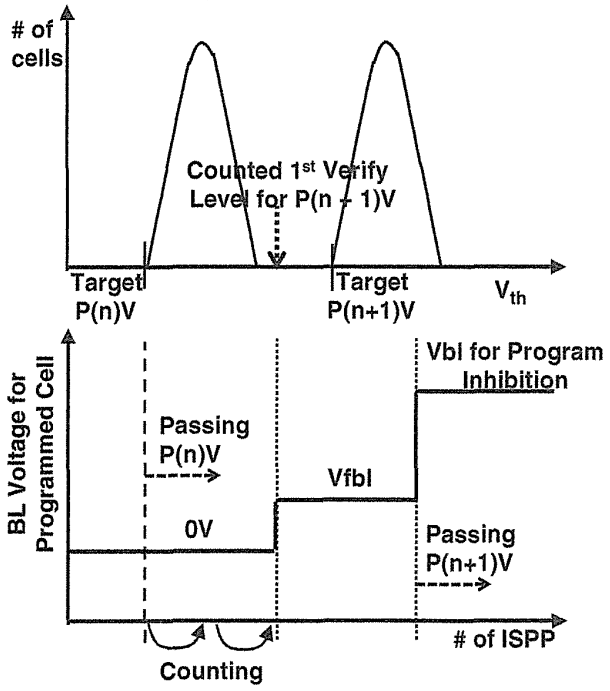


FIGURE 4.11 Forcing BL voltage by latch counting.

the effect of smaller V_{th} shift with V_{fbl}, the time of forcing the BL voltage should be delayed to after a few program pulses are applied. This is performed by a counting data latch in the page buffer, when passing the verify level of previous target state, as shown in Fig. 4.11. Thus, a tight V_{th} distribution width using a two-step tunneling rate is realized without an extra verify operation. Compared to the conventional two-step verify scheme, a verify-skip two-step tunneling ISPP scheme achieves 13% better program performance [16].

4.2.5 Pseudo-Pass Scheme in Page Program

The fast program speed essentially requires the more reduction of the time for one page programming. The duration of the page program is set sufficiently long to complete the program of all bits in a page. So, when any cells have unusually slow program characteristics in comparison with the majority of the cells, the page program speed becomes slower. As a solution to this problem, the pseudo-pass scheme (PPS) was proposed [17]. It allows the completion of a page programming operation even if a few bits are not programmed sufficiently. The error bits are corrected in a read operation by the ECC (error correction code). However, the conventional failure bit counting (FBC) operation is time-consuming, and so the PPS is not sufficiently effective. In order to realize the effective PPS, a high-speed FBC operation had been also proposed [17].

Figure 4.12a shows the flowchart of the conventional page program sequence. At first, memory cells in page are programmed according to loaded data. Then, they are verified consecutively. If all the cells, which should be programmed, are programmed, the program operation finishes and becomes a status pass. However, if they are not completed to program, the memory cells are programmed again. The judgment of "all cells programmed or not" is done by using the data that are stored in page buffer, as shown in Fig. 4.13. When the data in the buffer is "1," the cell that corresponds to the buffer is not programmed, however, when the data is "0," the cell is programmed

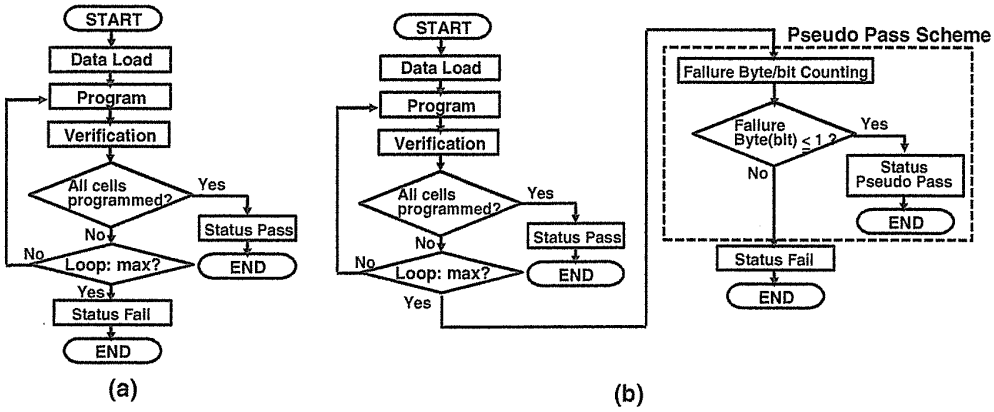


FIGURE 4.12 (a) Flowchart of a conventional page program sequence. (b) Flowchart of a page program sequence with the pseudo-pass scheme (PPS).

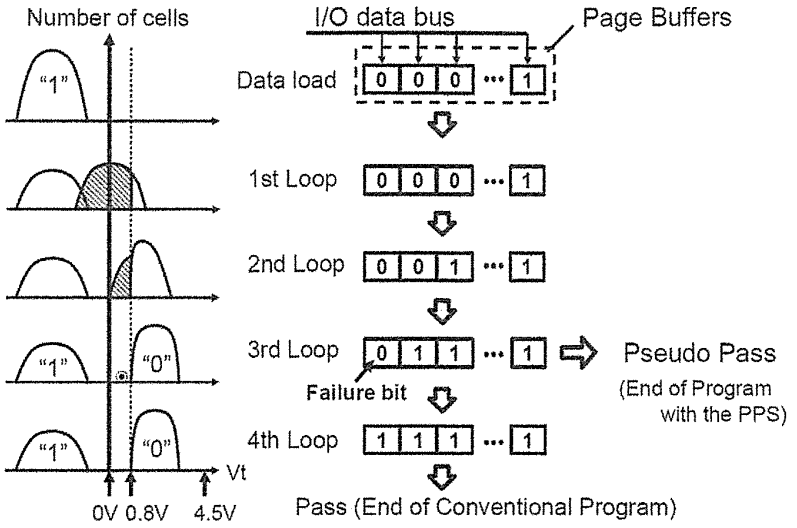


FIGURE 4.13 Change of V_t of memory cells and data in page buffers during a program sequence.

repeatedly. The data of the page buffers are revised in each verify operation. When the threshold voltage (V_t) of the programmed cell shifts up from the negative voltage of the erased status to more than the target value of 0.8 V, the data of the buffer is changed from "0" to "1" after the verify.

Figure 4.12b presents the flowchart of the pseudo-pass scheme (PPS). The PPS can be implemented just after the conventional program sequence. If the program operation doesn't complete after the predetermined iteration number of the program loops, the failure bit counting (FBC) circuit counts up the number of the page buffers whose data are "0." If the detected number of failure bits is less or equal to the allowed value, the status of "the pseudo pass" is output, and then the program sequence terminates. In Fig. 4.13, the predetermined iteration number of the program loops, which is enough programming for the majority of the cells, is assumed to be three. In this case, the program operation is finished with operation of the pseudo pass, as a result of the FBC after the third program loop without retrying an additional program loop, even though some insufficiently programmed cells are remained. Therefore, the iteration number of the program loops can be reduced by one or more in comparison with the conventional verify method, which doesn't permit any insufficient program bits.

Figure 4.14 compares the SLC program performance between the conventional program and the PPS program operation. The horizontal axis shows the worst program time. When the typical program time (tProg_typical) of the majority of cells is assumed to be 200 μs, there is a possibility that the worst program time of the conventional program becomes 250 μs or more because one or more program/verify sequence is required. However, the worst program time of the PPS operation using the new high-speed failure bit counter circuit [17] is limited to 200.8 μs, which is the sum of the typical program time of 200 μs and the counting-up time of 0.8 μs

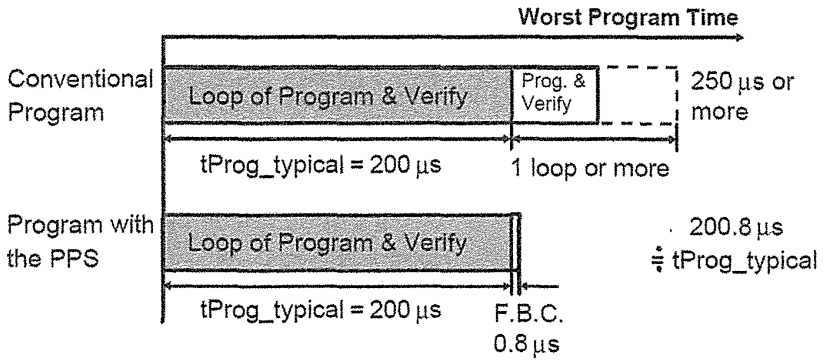


FIGURE 4.14 Program performance of a conventional page program and a page program with PPS.

in the FBC operation. In this case, the improvement of the worst program time is at least 20% in comparison with the conventional worst program time, by reducing the number of program loops for a few slowly programmed cells. The additional time of 0.8 μs for the PPS operation is negligibly small in comparison with the total program time.

The pseudo-pass scheme (PPS) has been implemented to the NAND flash product of SLC/MLC/TLC over 10 years due to fast page programming speed and less program disturb failure by avoiding excess program stress.

4.3 PAGE PROGRAM SEQUENCE

4.3.1 Original Page Program Scheme

In order to realize multi-bit cells (MLC) in scaled NAND flash memory cells, precise V_{th} distribution control is the key factor. The V_t distribution in a program state can be very tight by an ISPP and a bit-by-bit verify scheme. However, the distribution is eventually disturbed by well-known major parasitic effects, which are the background pattern dependency (BPD), source line noise (noise), and floating gate capacitive coupling interference (cell-to-cell interference), as shown in Fig. 4.15 [18, 19]. The background pattern dependency (BPD) can be minimized by various techniques such as fixed page program order and applied proper read voltage for unselected cells in a selected NAND string. And the source line noise can be also minimized by low resistance of mesh common source lines, as well as by low resistance of *p*-well structures of the memory array, such as retrograded doping profiled *p*-well. However, the cell-to-cell interference is mainly caused by floating gate capacitive coupling due to parasitic capacitances between cells, thus it is greatly affected by cell scaling (see Chapter 5). Figure 4.15b shows typical contributions of the three mentioned parasitic effects measured at a device at the 60-nm technology node [18–19]. Actually, the detailed portions of each effect can be different depending on the used NAND device structure and its operation condition, however, floating-gate

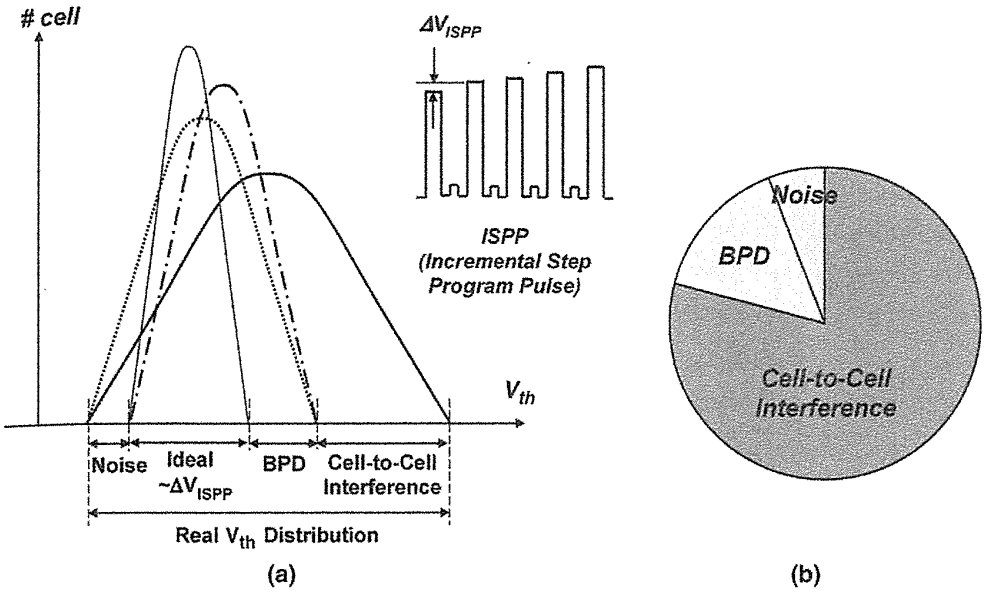


FIGURE 4.15 (a) Parasitic effects of a V_{th} distribution in NAND flash memory. (b) Contributed portion of each parasitic effect measured at a 60-nm technology node.

capacitive coupling interference is the most dominant effect, and will increase dramatically as scaling down NAND flash memory cells.

Figure 4.16a shows the memory cell array core architecture and page assignment of a conventional NAND flash memory device (original MLC product) [6,7][18,19]. Two BLs (bit lines) of even BL (BLe) and odd BL (BLo) are connected to sense amplifier (not shown in figure) through a switch. Either an even or odd BL cell is alternately selected and programmed sequentially in the order as described in Figure 4.16a. This BL scheme is called even/odd shield bit-line architecture [14, 21]. This scheme is effective to reduce BL noise shielding in read and program-verify

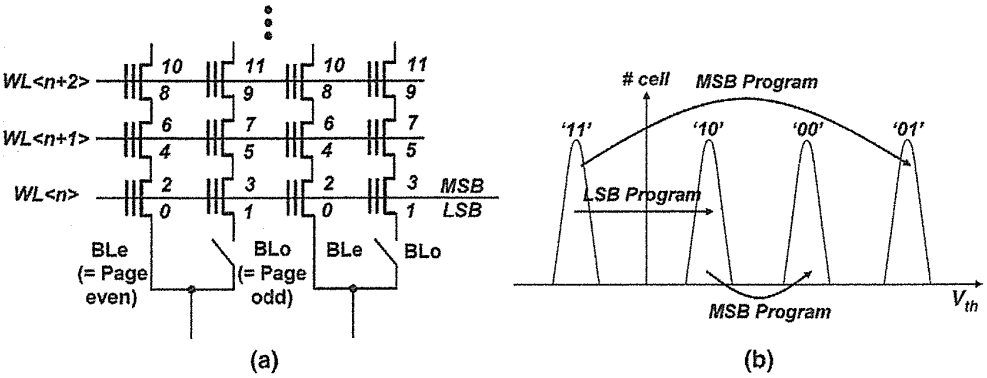


FIGURE 4.16 (a) Conventional core architecture and page assignment. (b) Conventional MLC program scheme. MSB; Most Significant Bit, and LSB; Least Significant Bit.

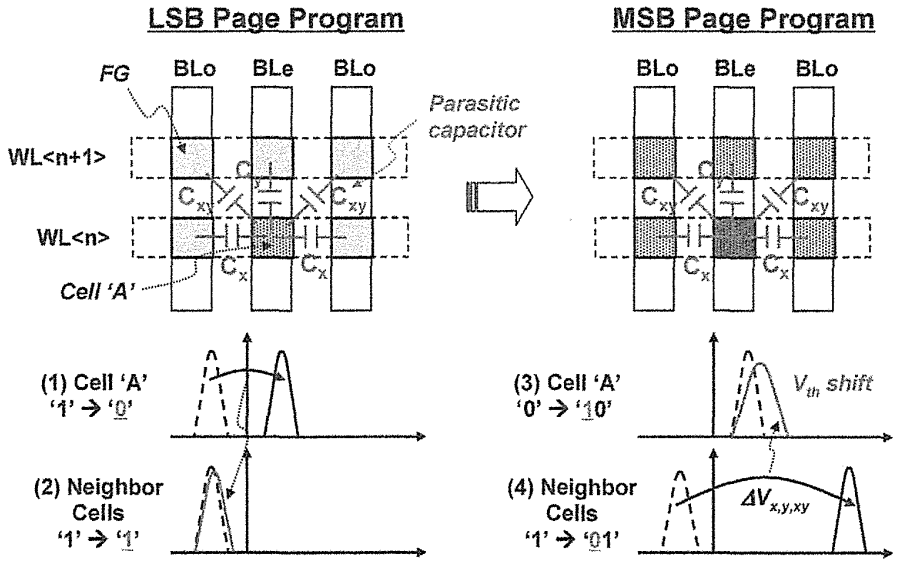


FIGURE 4.17 Worst-case cell-to-cell interference (floating-gate capacitive coupling interference) of conventional NAND architecture.

operations. A conventional MLC program scheme used in an original MLC NAND flash is shown in Fig. 4.16b. During the LSB program, V_{th} states of selected cells which have the erased V_{th} as the initial state move to the lowest programmed state '10'. Next, during the MSB program stage, two states, "00" and "01", are formed sequentially, depending on previous LSB data. After finishing the programming of four pages corresponding to a word line (WL<*n*>), the four pages corresponding to the next upper word line (WL<n + 1>) are programmed consecutively. It is noted that logical even and odd pages on the same word line are matched to physical even and odd BLs in the conventional architecture. The original MLC NAND architecture and page programming scheme shown in Fig. 4.16 was used in the first MLC NAND product of 0.16-μm 512-Mbit NAND flash memory in 2000.

Figure 4.17 shows the worst case of floating-gate capacitive coupling interference which occurs in original NAND architecture (see Fig. 4.16) [18, 19]. During LSB page programming, only selected the cell 'A' is programmed from '1' to '0', but all other surrounding neighbor cells are kept in the erase state ('1' → '1'). Subsequently at MSB page programming, if the data for the selected cell is '1', it is not programmed so that its state remains at '10'. Next, if the data for all neighbor cells are '0' and then all neighbor cells are programmed from the erased state '11' to the highest state '01', a large V_{th} shift is caused for the selected cell 'A' due to parasitic floating gate capacitive coupling interference, as shown in Fig. 4.17.

The widening of the distribution of the original NAND architecture caused by floating-gate capacitive coupling interference can be approximately expressed by equation "Original in Fig. 4.16" in Fig. 4.18. From the equation, it is found that not only reducing the parasitic capacitances, but also reducing the number of neighbor cells that are programmed after the programming of a selected cell and the amount

	Equation of floating-gate capacitive coupling
Original in Fig. 4.16	$\Delta V_x * (2C_x/C_{tot}) + \Delta V_y * (C_y/C_{tot}) + \Delta V_{xy} * (2C_{xy}/C_{tot})$
New Scheme (1) in Fig. 4.19	$(\Delta V_x/2) * (2C_x/C_{tot}) + (\Delta V_y/2) * (C_y/C_{tot}) + (\Delta V_{xy}/2) * (2C_{xy}/C_{tot})$
New Scheme (2) in Fig. 4.23	$(\Delta V_y/2) * (C_y/C_{tot}) + (\Delta V_{xy}/2) * (2C_{xy}/C_{tot})$

FIGURE 4.18 Approximated equations of floating-gate capacitive coupling interference in three page program schemes.

of shift at the MSB programming stage, is important to minimize the floating-gate capacitive coupling interference in a NAND flash memory cell.

4.3.2 New Page Program Scheme (1)

Figure 4.19 shows a new page program scheme (1) of new memory cell array core architecture and page assignment [22, 18, 19]. This scheme has been widely used in massproduction due to reducing V_{th} distribution width by decreasing an effect of the floating gate capacitive coupling interference. The floating-gate capacitive coupling interference by BL–BL direction (x-direction) can be reduced by performing the LSB program to the temporary state 'x0'. And the floating-gate capacitive coupling interference by WL–WL (y-direction) and diagonal neighbor cells can be reduced by performing MSB programming for a selected WL after LSB programming of its neighbor WL cell, as shown in Fig. 4.19a. The V_{th} shift by WL–WL and diagonal

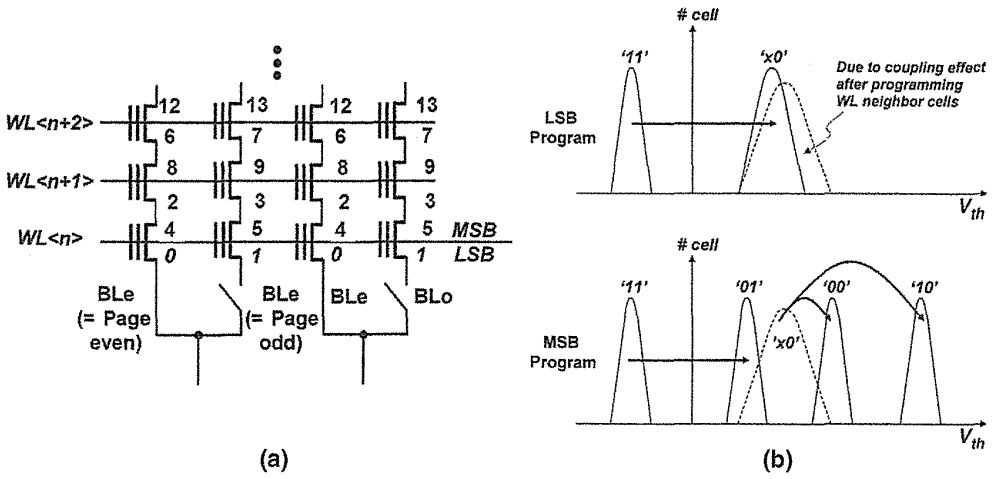


FIGURE 4.19 The new page program scheme (1). MLC program is performed after temporary LSB data storing. (a) New core architecture and page assignment. (b) New MLC program scheme.

interference can be almost reduced to half compared to the original program scheme of Fig. 4.16. With the new page program scheme (1), the achieved floating-gate capacitive coupling interference can be expressed as in "New Scheme (1) in Fig. 4.19" in Fig. 4.18. The new page program scheme (1) shown in Fig. 4.19 was first applied to a 70-nm 8-Gbit MLC NAND flash memory product in 2005 [23].

By using this new page program scheme (1), the worst case of the floating-gate capacitive coupling interference can be improved, compared with the original MLC program scheme. A new program scheme with temporary LSB data storing is used, as shown in Fig. 4.19b. At the LSB programming stage, the memory cell is programmed from "11" to "x0" as a temporary state just like SLC programming. After the WL neighbor cells are also LSB programmed, the V_{th} distribution is possibly widened as shown in Fig. 4.19b, uppergraph. Then, at the MSB programming stage, the 'x0' state is programmed to either '00' and '01' as the final state corresponding to the input data or either the '11' state is programmed to the final "01" state. All memory cells except '11' cells are programmed to their final states at the MSB programming stage from the temporary programmed state for LSB data. The V_{th} shift of neighbor cells is greatly reduced to around half in comparison with conventional page programming scheme shown in Fig. 4.16, so that the floating-gate capacitive coupling interference of neighbor cells can be greatly reduced. During MSB programming in this new page program scheme (1), a flag cell that is used for representing MSB programming and placed for each page is also programmed in order to distinguish LSB and MSB for read.

Other reports [24–26] also introduced the new scheme that reduced WL–WL interference by using programming to a temporary state. This programming scheme is that neighboring cells are roughly programmed before final programmed levels are programmed properly. Figure 4.20 shows the transient of a V_{th} distribution of cell "a" and neighboring cells, that is, cell "b". Figure 4.21 shows the programming order. At first, cell "a" is roughly programmed to lower levels than actual target level, as shown in Fig. 4.20 (1). The step voltage of incremental step pulse [10, 11] for this pre-programming is large, so the programming time of the operation is very short. Next, neighboring cells (cell "b") are programmed in the same way. The V_{th} distribution of cell "a" is widened because of the floating gate capacitive coupling effect, as shown in Fig. 4.20 (2). After this, cell "a" is programmed again with a smaller step voltage of incremental step pulse to proper levels, as shown in Fig. 4.20 (3). When next-neighboring cells (cell "c") and neighboring cells (cell "b") are programmed afterwards, the V_{th} distribution of cell "a" is widened by the floating-gate coupling effect, but the widening is very small, because the shift of neighboring cells are small, as shown in Fig. 4.20 (4), (5).

4.3.3 New Page Program Scheme (2)

The floating-gate capacitive coupling effect could be reduced by a new page program scheme (1), as shown in Fig. 4.19. In order to further reduce the floating-gate capacitive coupling interference between BLs (x-direction), the way to program adjacent cells (in both even page and odd page) at the same program pulse (sequence) is

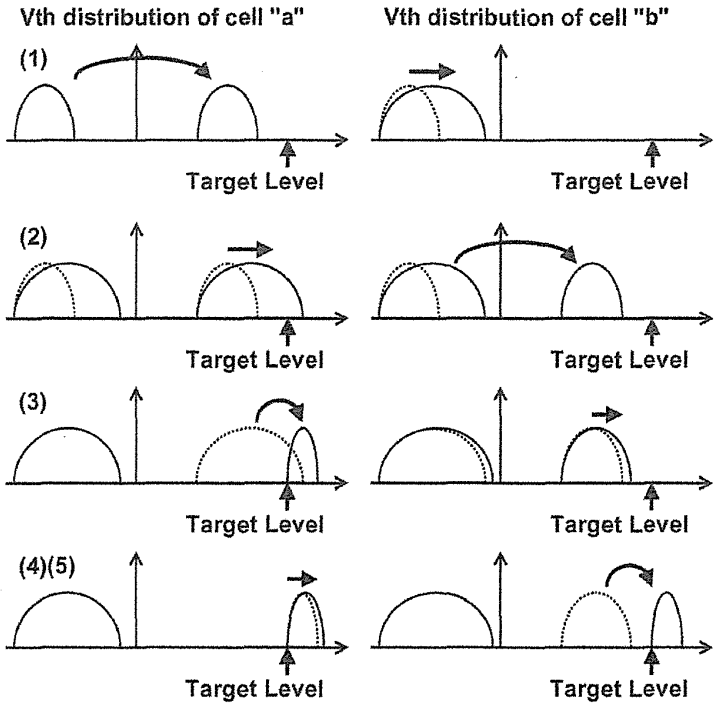


FIGURE 4.20 V_t distribution transition of cell "a" and cell "b."

effective [18, 19, 27, 28]. The concept of the new architecture is simply to reduce the number of neighbor cells as well as their amount of V_t shift at the MSB program stage. Figure 4.22 [18, 19] shows the concept of page assignment of the new page program scheme (2) to reduce the number of neighbor cells between BLs. In this new scheme (2), the logical even and odd pages on the same word line are each assigned to a physical group of memory BLs. By adopting this architecture, the same page address is assigned to adjacent memory cells on the same word line of a selected group of memory cells, which means that memory cells including adjacent cells in the BL direction can be programmed simultaneously. Accordingly, while logical odd

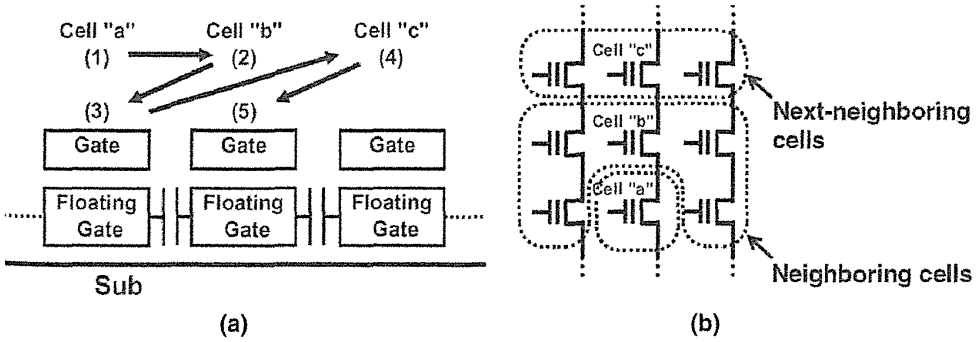


FIGURE 4.21 Programming order over word lines.

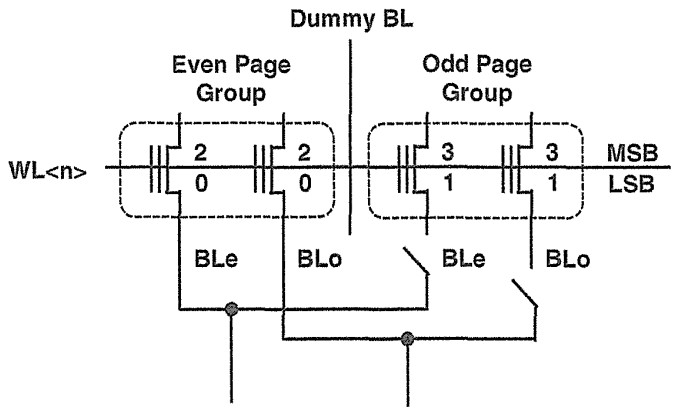


FIGURE 4.22 Concept of page assignment of the new page program scheme (2).

page data are programmed in the memory cells of the odd page group, memory cells in the even page group are not susceptible to the coupling effect at all. In order to remove the floating-gate capacitive coupling interference for edge memory cells of the page group, a dummy BL (dummy cell) can be simply used between the page groups, as shown in the figure.

Figure 4.23 shows (a) the simplified memory cell array core architecture and (b) page address ordering of the new page program scheme (2). Two BL selectors coupling to each even and odd page group are configured to transfer even and odd page data from the page buffer. The boundary between even and odd page groups is simply formed using dummy BL in order to remove the floating-gate capacitive coupling interference for edge memory cells of the page group. It should be noted that no additional area penalty arises in the proposed memory array. This is because the dummy BL which already exists for contacting the CSL (common source line) or wells in the conventional memory array can be used as the dummy BL for a page

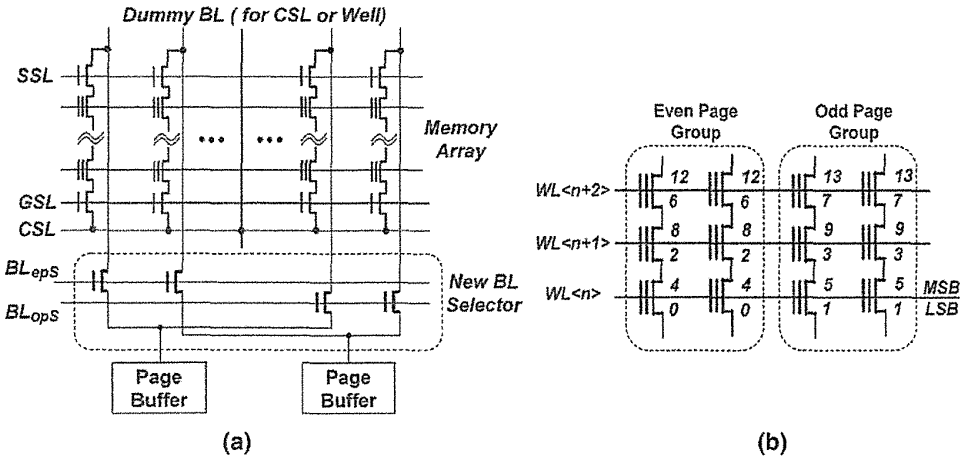


FIGURE 4.23 (a) Simplified core architecture and (b) page address ordering of the new page program scheme (2).

group boundary. With the new architecture, the floating-gate capacitive coupling interference can be much reduced as approximately expressed in equation "New Scheme (2) in Fig. 4.23" in Fig. 4.18.

4.3.4 All-Bit-Line (ABL) Architecture

The all-bit-line (ABL) architecture was firstly proposed in ISSCC 2008 [27,28]. In ABL architecture, all cells along a selected word line are programmed simultaneously, not separated to even or odd page groups. Then, the ABL architecture could reduce the floating-gate capacitive coupling interference due to reducing BL–BL interference, and also the ABL architecture could realize high-speed page programming with double page size.

Before ABL architecture was proposed, the conventional even/odd shield bit-line scheme was used for a NAND flash product. In the NAND cell array structure, the BL–BL coupling capacitance (not floating-gate coupling capacitance) is around 90% of the total bit-line capacitance [23]. For this reason, most NAND flash products utilize a conventional shielded bit-line scheme to do sensing [13, 21], where only half of the BLs are sensed and the other half of the BLs are at 0 V. Since only half of the cells on the same WL can be sensed at the same time, the data latches are designed to be shared by even/odd pairs, to save on die size. This architecture requires the separate programming and verification of even and odd BLs. The program speed was limited by small page size of even and odd BLs. Moreover, as memory cell size scales down, the program disturb issue was more aggravated due to a longer programming time on the same WL, and then reliability was degraded.

Figure 4.24 shows a schematic diagram of memory core circuits of ABL architecture [27,28]. The even and odd bit lines (namely, all bit lines) have their own sense

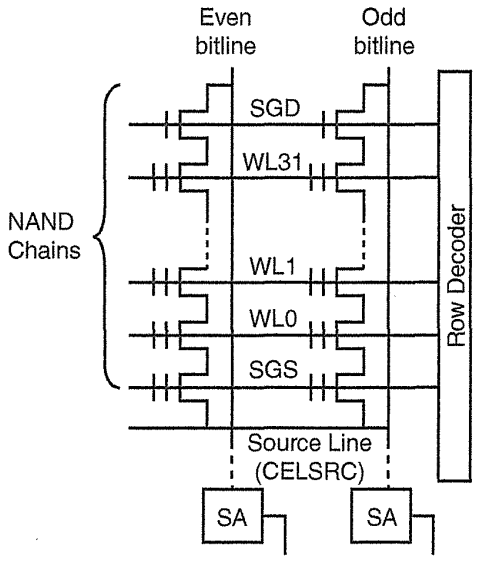


FIGURE 4.24 Simplified memory core architecture of all-bit-line (ABL) architecture.

amplifier (SA) attached. The sensing scheme was changed from voltage sensing in even/odd shield bit-line scheme to current sensing in the ABL scheme, to solve the large BL–BL coupling capacitance issues. The SA performs sensing operations in read, program verify, and erase verify operations. In ABL architecture, all bits on the same word line (WL) can be programmed and read at the same time. The total number of cells that can be programmed are double that of conventional even/odd bit-line architecture.

Therefore, the ABL architecture could lead to high performance of programming with double page size, accompanied by high reliability with shorter program disturb stress time. Also, in ABL architecture, the floating-gate capacitive coupling interference can be reduced, compared with a conventional even/odd shield bit-line scheme. Figure 4.25a,b shows the floating-gate capacitive coupling interference in a conventional even/odd shield bit-line scheme. The cells in an even bit line (even page) causes a V_t shift by programming neighbor cells in odd bit line (odd page) due to the floating-gate capacitive coupling interference, as shown in Fig. 4.25a,b. On the other hand, in an ABL scheme, all cells in even page and odd page are programmed at the same time. The V_t shift of the floating-gate capacitive coupling interference can be much reduced [27]. While the erase distribution can still encounter the full

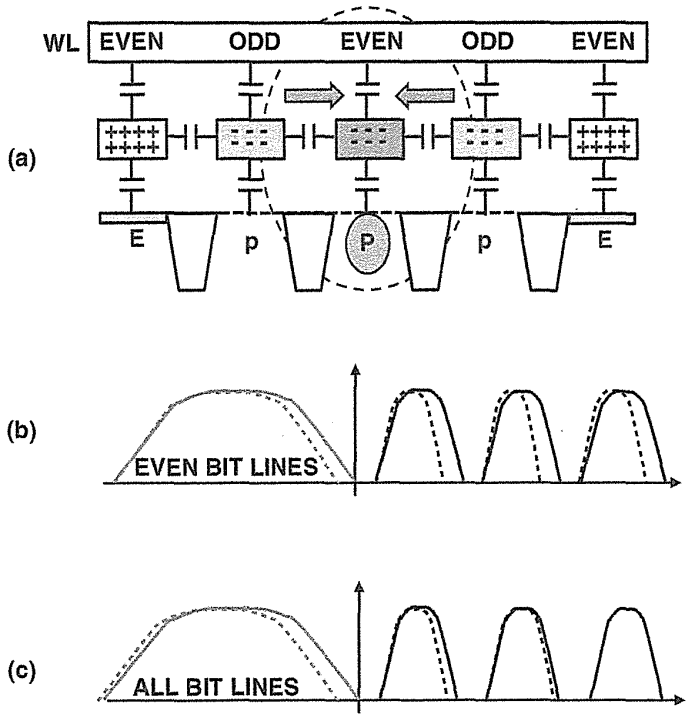


FIGURE 4.25 (a) Floating-gate capacitive coupling interference for an even page in a conventional even/odd programming scheme. (b) V_t distribution image of an even page in an even/odd programming scheme. (c) V_t distribution image in an ABL programming scheme. Floating-gate capacitive coupling interference in an all-bit-line programming scheme can be reduced, because both even page and odd page are programmed simultaneously.

the floating-gate capacitive coupling interference, the interference effect on the first programming state is reduced. The highest state has almost no V_t shift of interference, as shown in Fig. 4.25c [27].

4.4 TLC (3 BITS/CELL)

In order to decrease a bit cost of NAND flash memory, TLC (3 bits/cell) technologies had been developed [29–36, 16]. The first paper for massproduction of TLC was presented in 2008 ISSCC (International Solid-State Circuits Conference) by using 56-nm technology [30]. The key issue of TLC technologies is the page program sequence and method to achieve a very tight V_t distribution width for producing a margin between each V_t state.

The page address is assigned in a way that enables each page to be treated as an independent page for users. The same user commands can be used for all pages programmed in the array. The conventional page program sequence is shown in Fig. 4.26 [24, 25, 35, 29–31]. This sequence is implemented as applying the concept of a new program scheme (2) for MLC (Section 4.3.3) to TLC, in order to minimize an effect of the floating-gate capacitive coupling interference. The three pages on the same WL are called lower page (first), middle page (second), and upper page (third), respectively. The lower page (first) is programmed like a normal SLC program operation, where the erase cell "E" is programmed to state "A1". After the lower page programming, the middle page (second) program data can be brought in for programming. For the middle page programming, 2 bits (both lower and middle pages) are needed to program 3 program states. The lower page data can be read from a memory array into the data latch. The middle page is programmed similarly to that of 2-bit-per-cell programming that is from "E" to "B1" and from "A1" to "B2" or "B2". After the middle page programming, the upper page (Third) program data is brought in from outside. The upper page program requires the lower and middle page

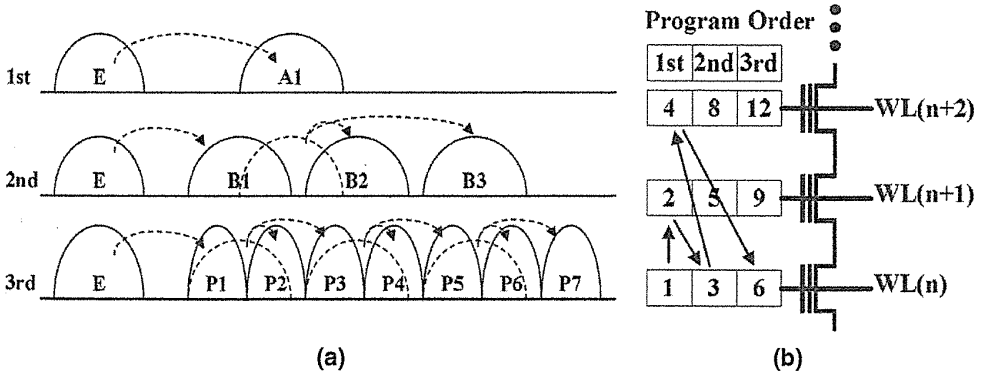


FIGURE 4.26 (a) The conventional page program sequence of TLC (3 bits/cell). Three pages of first (lower page), second (middle page), and third (upper page) are programmed on each word line. (b) Program order between word lines.

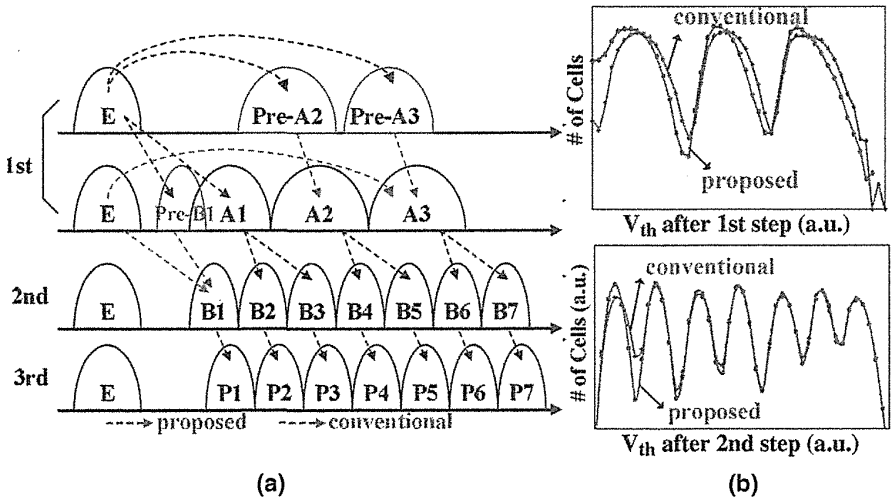


FIGURE 4.27 (a) The new page program sequence with pre-program scheme. (b) Measured V_t distribution after first (lower page) and second (middle page) program.

data to program seven program states. The lower and middle page information can be read from an array. The seven program states must fit into a V_t window similar to a MLC program V_t window, so the upper page program should have a very small V_{PGM} step size to achieve a well-controlled, narrow V_t distribution for all seven program states. Therefore, the upper page program is the slowest programming speed of the three pages. Figure 4.26b shows a page program sequence between world lines. This sequence is also implemented as applying the concept of new program scheme for MLC (Section 4.3.3) to TLC, in order to minimize an effect of the floating-gate capacitive coupling interference.

A new page program sequence with a pre-program scheme had been proposed for a 21-nm node cell [35], as shown in Fig. 4.27. In a new scheme, 5 states and 8 states are implemented in the first and second step program, respectively, so that it is minimized adjacent cell-to-cell interference (floating-gate capacitive coupling interference) at the third step program, as shown in Fig. 4.27a [35]. By using a pre-A2 and a pre-A3 program in the first step program, the V_t distribution width of A1/A2/A3 can be reduced by 15% reduction of adjacent BL-to-BL coupling interference compared to a sequential program, as shown in Fig. 4.27b, uppergraph. For the same reason, by applying a pre-B1 program in the first step program, adjacent WL-to-WL coupling interference can be reduced 10%. During the second step program, adjacent WL-to-WL coupling interference is minimized due to the effect of a pre-B1 program. Figure 4.27b (lowergraph) shows a measured V_{th} distribution of the second step program by using a pre-program program.

As memory cell geometry shrinks, floating-gate (FG) capacitive coupling interference is becoming worse. And smaller memory cells are also vulnerable to more cell-to-cell variations. These factors combine to negatively impact program performance. The FG capacitive coupling interference can be reduced by the air gap (AG)

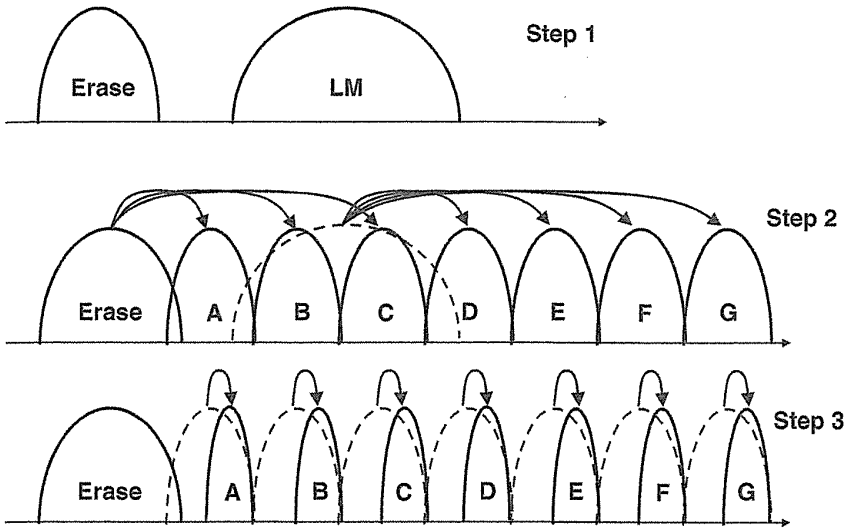


FIGURE 4.28 Three-step program algorithm for a 19-nm node cell.

process (see Section 5.3.4) between two adjacent WLs. The 19-nm AG technology has an FG-to-FG coupling ratio equivalent to that of 2X nm without AG. Also, for enhancing program performance in the 19-nm technology node, a new enhanced Three-Step Program (TSP) was applied to TLC 128-Gb NAND flash memory, as shown in Fig. 4.28 [36]. In a new enhanced TSP, cells are programmed from two states of Erase/LM to eight states of Erase/A-G, and then compaction program is performed to states of A–G (Step 3 in Fig. 4.28). Due to the skip of second page programming in the conventional program scheme in Fig. 4.26, program speed is enhanced. Therefore, a combination of a new enhanced Three-Step Program (TSP) and air gap allows fast program speed of 18 MB/s on TLC of the 19-nm technology node.

Figure 4.29 shows the measured V_t distributions of TLC in several generations of NAND flash memory cells, which are (a) 56-nm cell [29, 30], (b) 32-nm cell [31], (c) 20-nm-node cell (27-nm) [16], (d) 21-nm cell [35], and (e) 19-nm cell [36]. We can see that the read window margin is gradually degraded as scaling of memory cells, even if program operation is newly developed for each generation.

4.5 QLC (4 BITS/CELL)

QLC technology was presented in a 70-nm cell in the 2007 Symposium on VLSI Circuits [24] and 43-nm cell on 2009 ISSCC [37]. The papers were focused on the intelligent operation of achieving a tight V_t distribution width by reducing the floating-gate capacitive coupling interference. Figure 4.30 shows the V_{th} distribution transition of 16LC (16-level cell) case [24, 25]. At first, the cells are roughly programmed to lower levels (lower verify voltages) than the target levels (target verify voltages), as

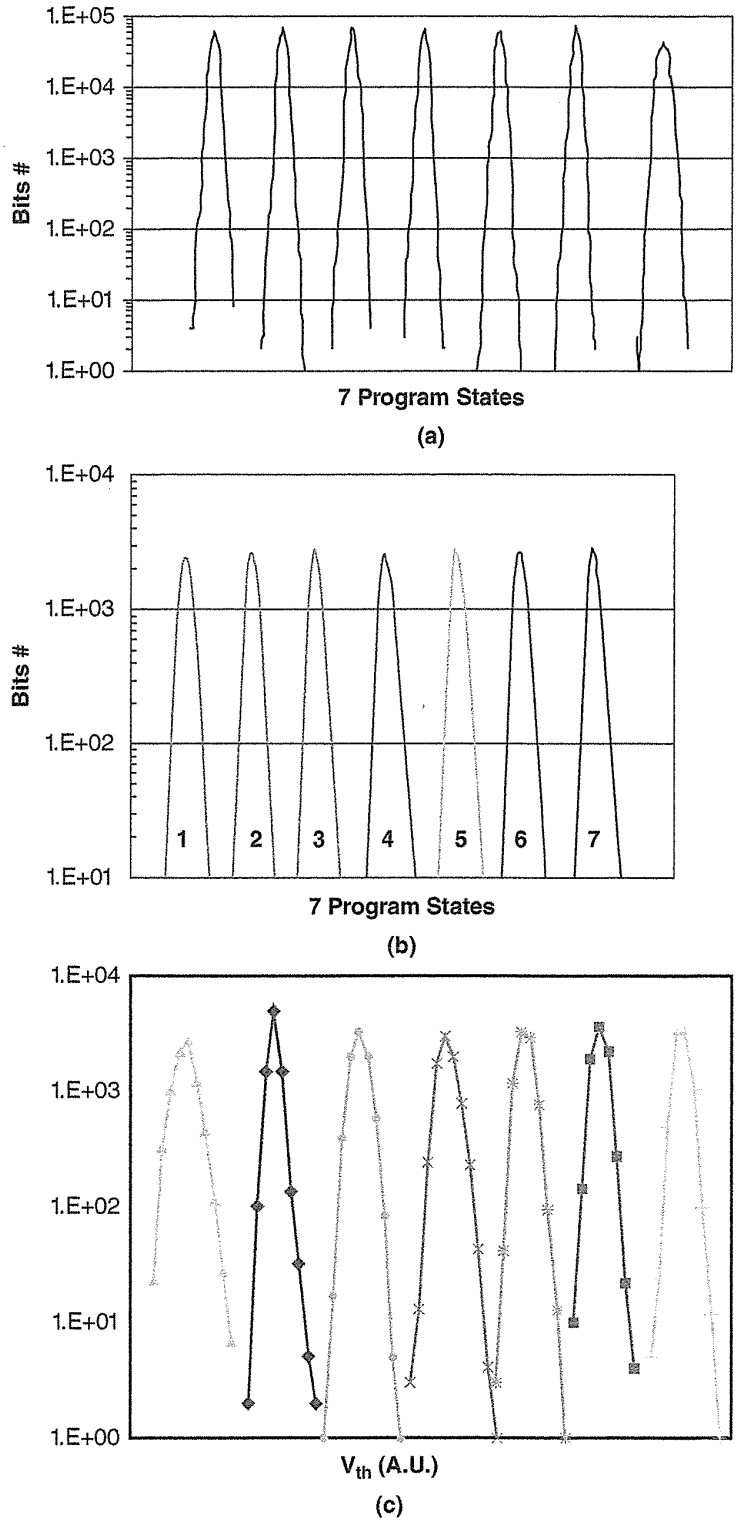
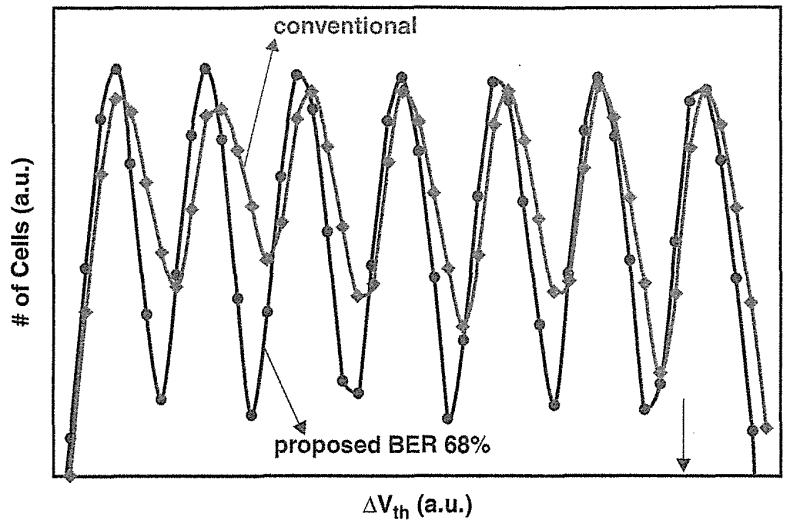
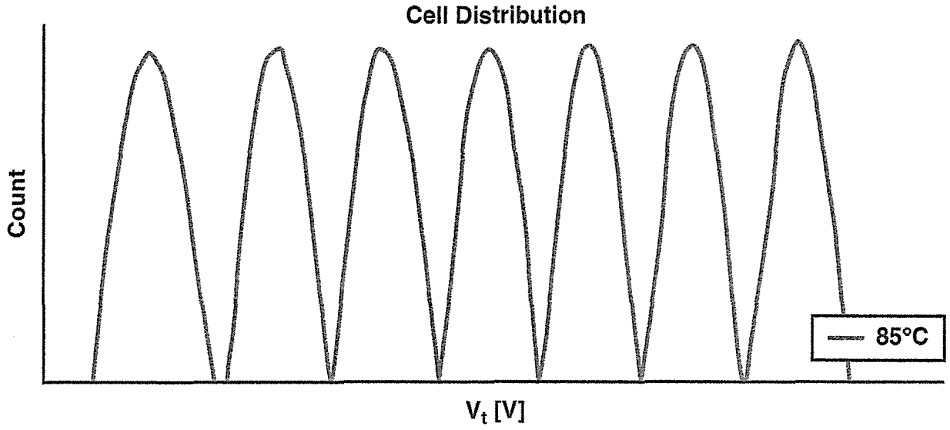


FIGURE 4.29 The V_t distribution of seven program states (TLC) in cell generations of (a) 56 nm, (b) 32 nm, (c) 20-nm node (27 nm), (d) 21 nm, and (e) 19 nm.



(d)



(e)

FIGURE 4.29 *(Continued)*

shown in Fig. 4.30 (1). Then, when neighboring cells are programmed, the distribution width is widened mainly due to the floating-gate capacitive coupling interference, as shown in Fig. 4.30 (2). After that, the cells are programmed again to 16 levels with target levels, as shown in Fig. 4.30 (3). And when neighboring cells are programmed again, the distribution of these cells is widened, but it is very small, because the shift of neighboring cell is small enough, as shown in Fig. 4.30 (4)(5). By using this method, very tight V_t distribution width for 16LC can be obtained.

QLC technology was developed in a 43-nm memory cell [37]. The page program sequence of 43-nm QLC is nearly the same as that of a 70-nm QLC cell [24, 25]. Figure 4.31 shows the V_T-distribution transition of cell "a" in the string and the programming order. Each cell goes through three steps of programming. First, cell "a" is programmed to three levels (Step 1), similar to a MLC device (see Fig. 4.31 (1)).

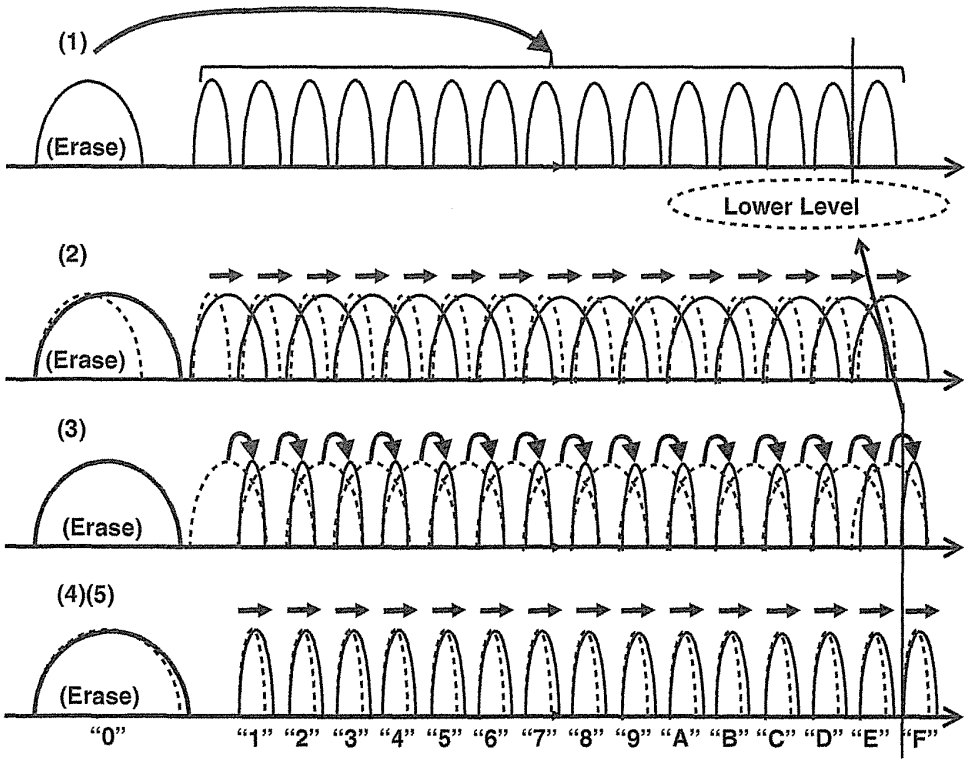


FIGURE 4.30 V_{th} distribution transition of 16LC (QLC, 4 bits/cell).

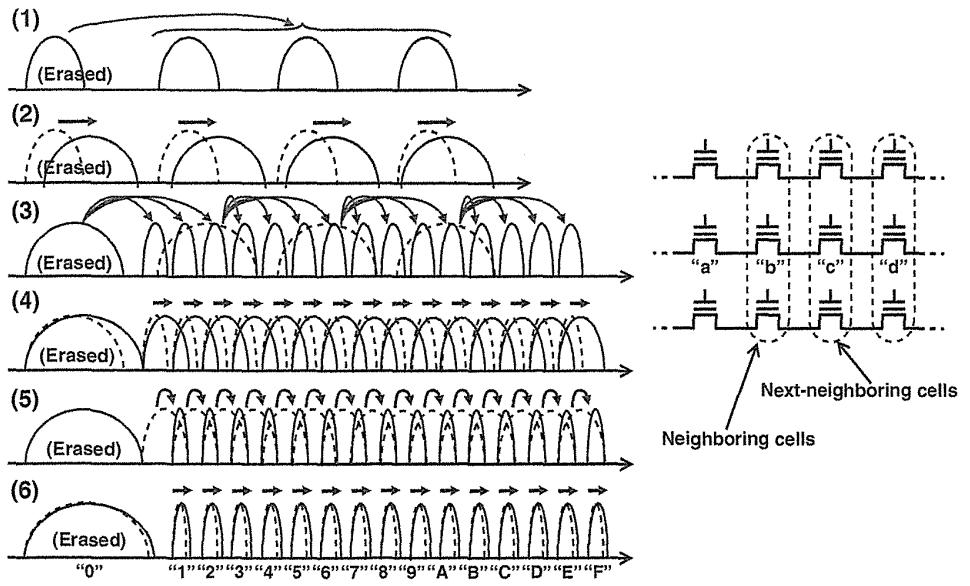


FIGURE 4.31 Three-step programming scheme of page programming order and V_t transition for QLC (4 bits/cell).

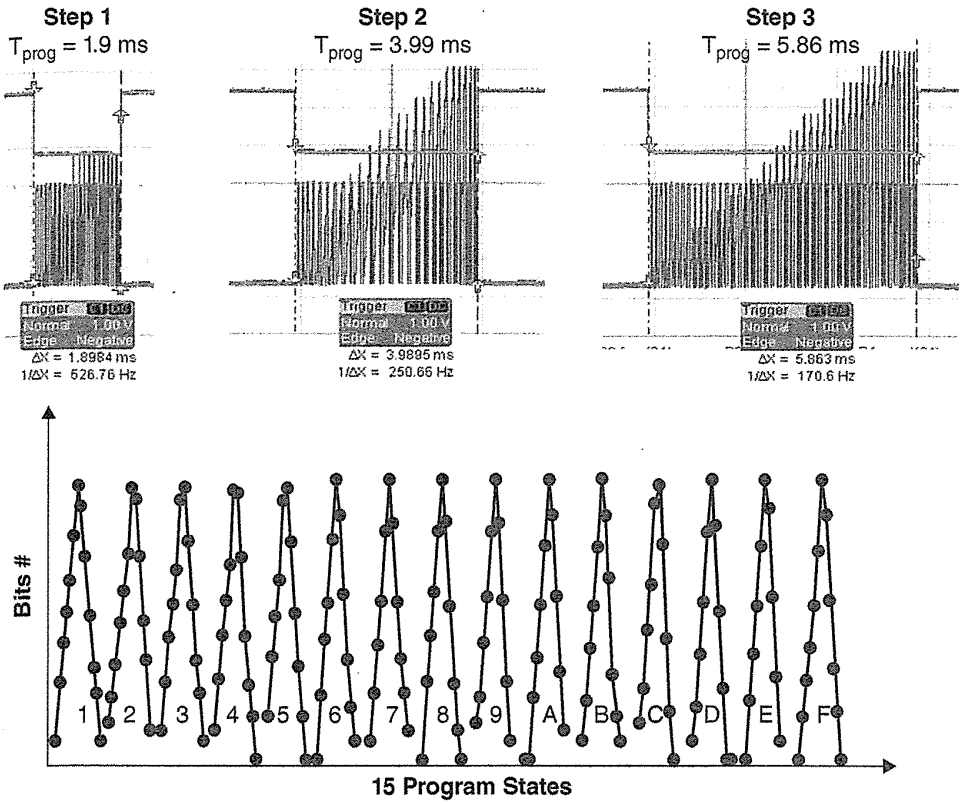
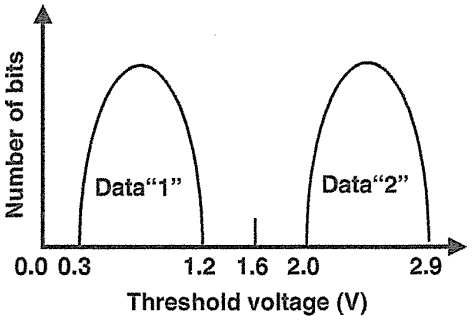


FIGURE 4.32 The measurement data of programming time for each page program step and V_t distribution for QLC.

Next, neighboring cells ("b") are programmed the same way. The V_T distribution of cell "a" is widened due to the FG coupling effect (see Fig. 4.31 (2)). Cell "a" is roughly programmed (Step 2) to 15 levels, lower than the targets (see Fig. 4.31 (3)). Next, neighboring cells ("c" and "b") are programmed to 3 and 15 rough levels, respectively, causing the V_T distribution of cell "a" to widen again (see Fig. 4.31 (4)). Finally, cell "a" is programmed (Step 3) to the 15 target levels (see Fig. 4.31 (5)). Neighboring cells ("b", "c", and "d") are then programmed again the same way (see Fig. 4.31 (6)). This page program sequence minimizes the FG coupling effect, even with large values due to technology scaling. Figure 4.32 shows the measured data of 15 program states, along with the programming time for each step. Total programming time of 11.75 ms translates to 5.6 MB/s, when two pages are programmed together in two cell arrays (two-page mode).

4.6 THREE-LEVEL (1.5 BITS/CELL) NAND FLASH

The 1.5-bit/cell technology was proposed to realize both a high performance and a low bit cost in the same product [38]. Targets of 1.5-bit/cell technology are (1) a high



3-bit data	Memory cells
000	0 0
001	0 1
010	0 2
011	1 0
100	1 1
101	1 2
110	2 0
111	2 1

(b)

FIGURE 4.33 (a) Threshold distribution for a 3-level memory cell (1.5 bits/cell). (b) Three-bit data and three-1evel data in memory cells.

program performance which is nearly the same as SLC (2 bits/cell) performance, (2) a better reliability rather than MLC, and (3) lower bit cost than SLC product.

MLC (2 bits/cell) NAND flash technology based on four-level V_t states is a most popular solution for demands. However, it is difficult for the four-level MLC to provide as good a reliability and performance as SLC due to MLC's narrow read window margin (RWM) and slow program speed. These problems would limit its application to be used in the market. To overcome the market limitation while achieving both cost and performance, three-level memory cell technology, which has at least 2 times wider read window margin than four-level MLC, is promising.

The three-level memory cell has data "0", "1", and "2" as shown in Fig. 4.33 [38] and Fig. 4.34 [39]. A "0"-state (erase state) corresponds to a threshold voltage of less than −1 V. A pair of memory cells stores 3-bit data as shown in Fig. 4.33b. Here, 528-byte page data including parity-check bits and several flag data are simultaneously transmitted from or to 2816 memory cells through 2816 compact intelligent three-level column latches.

Four intermediate code (i.e., 6-bit data), are loaded within a 25-ns cycle time. During the data load, a charge pump generates a high voltage for the first program pulse. The setup time of the high voltage is 5 μs. The first pulse duration is 20 μs and each of the subsequent pulses has 10 μs duration. A program recovery time and the program verify time are 1 μs and 16 μs, respectively. Total program time for 512-byte data is 704 × 25 ns + 20 μs + 1 μs + 16 μs + (5 μs + 10 μs + 1 μs +16 μs) × 3 = 150.6 μs. Then, the typical program throughput is 3.4 Mbyte/s and 68% of the two-level NAND flash, as shown in Fig. 4.35. In another paper [39], the page program speed is 45% of a 1-bit/cell (two-level) SLC NAND cell. Figure 4.35 shows the estimated program speed comparison between the three-level and conventional methods.

The die size is also estimated on the assumption that, in the case of the two-level flash memory, the memory cells and the column latches occupies 66% of the die size. A number of the memory cells and an area of the column latches are increased to 133.3% when a memory capacity is doubled. The die size of the three-level flash

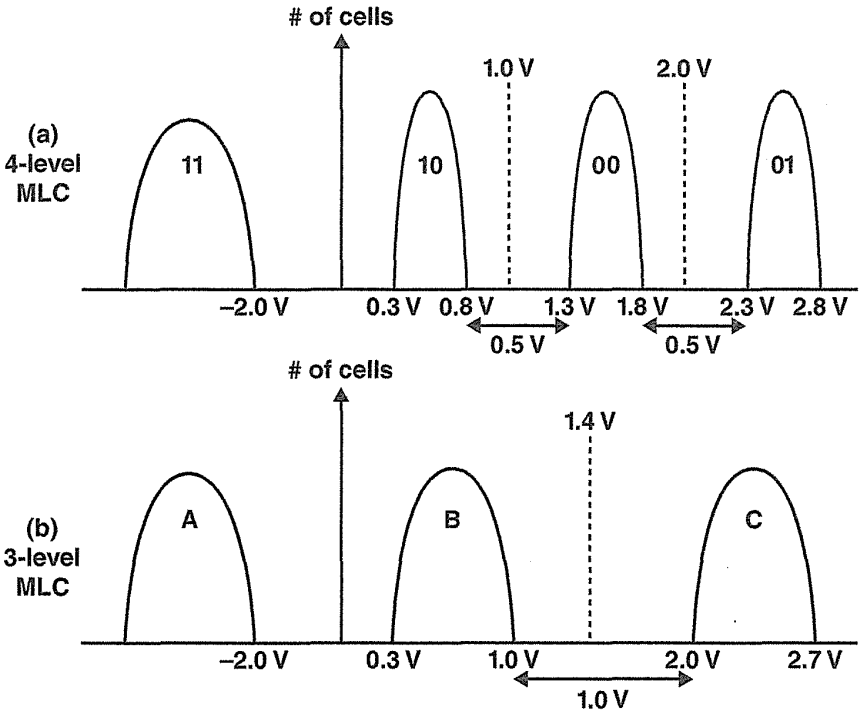


FIGURE 4.34 V_{th} distribution in NAND flash: (a) Four-level MLC. (b) Three-level cell (1.5 bits/cell).

memory chip is increased to 122%. Therefore, the die size per bit is reduced to 61%, as also shown in Fig. 4.35. Another paper shows that the estimated die size is 103 mm^2, which is reduced by 20% compared to SLC die, as shown in Fig. 4.35.

As shown above, a three-level cell has an advantage of 3–6 times improved program speed compared with MLC, along with 20–39% die size reduction compared with SLC. The three-level cell would have a possibility to use a certain application—for example, high-end enterprise server, and so on.

	Program Speed		Die Size/bit	
	Ref. 38	Ref. 39	Ref. 38	Ref. 39
1 bit/cell (SLC)	100%	100%	100%	100%
1.5 bit/cell	68%	45%	61%	80%
2 bit/cell (MLC)	11%	15%	50%	62%

FIGURE 4.35 Comparison of program speed and die size per bit.

4.7 MOVING READ ALGORITHM

As memory cell size is scaled down, the V_t shift of data retention becomes worse and worse, especially after a large amount of program/erase cycling. To compensate this data retention issues, the moving read algorithm was proposed [26]. Operation of moving read is to adjust a read voltage on the selected control gate according to cell V_t shift caused by data retention, and so on.

One example of the moving read algorithm for the program and read sequence is shown in Fig. 4.36 [26]. During page buffer setting in program operation, cells that will be programmed to PV3 are counted and stored in special extra cells of named FLG_PV3 in page. At the read operation, cells at PV3 are counted and compared

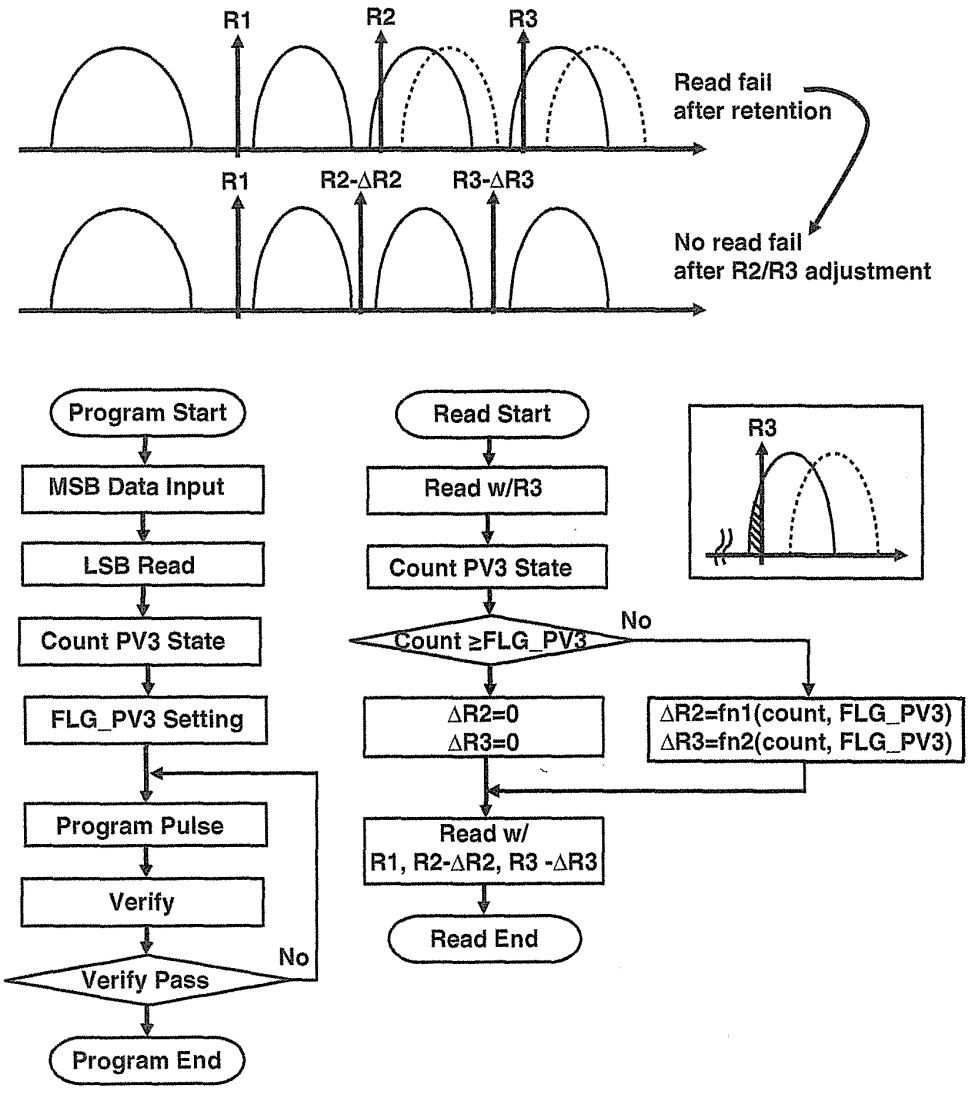


FIGURE 4.36 Moving read algorithm.

to ideal FLG_PV3. If count value meets FLG_PV3, the present read level can be applied. However, if count value does not meet FLG_PV3 (most cases are large V_t shift of data retention) to affect read results, the voltage of read level is calculated to be shifted down based on the difference between count value and FLG_PV3. The reason why PV3 cells are monitored is because data retention V_{th} shift is worst in the PV3 state. More than 30% error due to data retention V_{th} shift was improved.

Moving read operation in Fig. 4.36 is one example. There would be many alternate algorithms of the moving read to compensate cell V_t shifts which are caused not only by data retention but also by floating-gate capacitive coupling interference, program injection spread, RTN, program disturb, read disturb, and so on, as described in Chapters 5 and 6. The moving read operation has to be optimized because it is very effective to improve reliability of NAND flash memory product.

REFERENCES

[1] Bauer, M.; Alexis, R.; Atwood, G.; Baltar, B.; Fazio, A.; Frary, K.; Hensel, Ishac, M.; Javanifard, J.; Landgraf, M.; Leak, D.; Loe, K.; Mills, D.; Ruby, P.; Rozman, R.; Sweha, S.; Talreja, S., Wojciechowski, K. A multilevel-cell 32 Mb flash memory, *IEEE ISSCC*, pp. 132–133, 1995.

[2] Takeuchi, K.; Tanaka, T.; Nakamura, H. A double-level-V_{th} select gate array architecture for multi-level NAND flash memories, in *1995 Symposium on VLSI Circuits*, Technical Paper, pp. 69–70, 1995.

[3] Hemink, G. J.; Tanaka, T.; Endoh, T.; Aritome, S., Shirota, R. Fast and accurate programming method for multi-level NAND EEPROM's, in *1995 Symposium on VLSI Technology*, Technical Paper, pp. 129–130, 1995.

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

[5] Jung, T.-S.; Choi, Y.-J.; Suh, K.-D.; Suh, B.-H.; Kim, J.-K.; Lim, Y.-H.; Koh, Y.-N.; Park, J.-W.; Lee, K.-J.; Park, J.-H.; Park, K.-T.; Kim, J.-R.; Yi, J.-H.; Lim, H.-K. A 117-mm^2 3.3-V only 128-Mb multilevel NAND flash memory for mass storage applications, *Solid-State Circuits, IEEE Journal of,* vol. 31, no. 11, pp. 1575–1583, Nov. 1996.

[6] Takeuchi, K.; Tanaka, T.; Tanzawa, T. A Multi-page cell architecture for high-speed programming multi-level NAND flash Memories, *VLSI Circuits, 1997. Digest of Technical Papers, 1997 Symposium on*, pp. 67–68, 12–14 June 1997.

[7] Takeuchi, K.; Tanaka, T.; Tanzawa, T. A multipage cell architecture for high-speed programming multilevel NAND flash memories, *Solid-State Circuits, IEEE Journal of*, vol. 33, no. 8, pp. 1228–1238, Aug. 1998.

[8] Aritome, S.; Satoh, S.; Maruyama, T.; Watanabe, H.; Shuto, S.; Hemink, G. J.; Shirota, R.; Watanabe, S.; Masuoka, F. A 0.67 μm^2 self-aligned shallow trench isolation cell (SA-STI cell) for 3 V-only 256 Mbit NAND EEPROMs, *Electron Devices Meeting, 1994. IEDM '94. Technical Digest., International*, pp. 61–64, 11–14 Dec. 1994.

[9] Aritome, S. NAND Flash Innovations, *Solid-State Circuits Magazine, IEEE*, vol. 5, no. 4, pp. 21, 29, Fall 2013.

[10] Suh, K.-D.; Suh, B.-H.; Lim, Y.-H.; Kim, J.-K.; Choi, Y.-J.; Koh, Y.-N.; Lee, S.-S.; Kwon, S.-C.; Choi, B.-S.; Yum; J.-S. Choi, J.-H.; Kim, J.-R.; Lim, H.-K. A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme, *Solid-State Circuits, IEEE Journal of*, vol. 30, no. 11, pp. 1149–1156, Nov. 1995.

[11] Hemink, G. J.; Tanaka, T.; Endoh, T.; Aritome, S.; Shirota, R. Fast and accurate programming method for multi-level NAND EEPROMs, *VLSI Technology, 1995. Digest of Technical Papers. 1995 Symposium on*, pp. 129–130, 6–8 June 1995.

[12] Hemink, G. J.; Shimizu, K.; Aritome, S.; Shirota, R. Trapped hole enhanced stress induced leakage currents in NAND EEPROM tunnel oxides, *Reliability Physics Symposium, 1996. 34th Annual Proceedings., IEEE International*, pp. 117–121, April 30 1996–May 2 1996.

[13] Tanaka, T.; Tanaka, Y.; Nakamura, H.; Oodaira, H.; Aritome, S.; Shirota, R.; Masuoka, F. A quick intelligent program architecture for 3 V-only NAND-EEPROMs, *VLSI Circuits, 1992. Digest of Technical Papers, 1992 Symposium on*, pp. 20–21, 4–6 June 1992.

[14] Tanaka, T.; Tanaka, Y.; Nakamura, H.; Sakui, K.; Oodaira, H.; Shirota, R.; Ohuchi, K.; Masuoka, F.; Hara, H. A quick intelligent page-programming architecture and a shielded bitline sensing method for 3 V-only NAND flash memory, *Solid-State Circuits, IEEE Journal of*, vol. 29, no. 11, pp. 1366–1373, Nov. 1994.

[15] Tanaka, T.; Chen, J. US Patent, US6643188, 2003.

[16] Park, K.-T.; Kwon, O.; Yoon, S.; Choi, M.-H.; Kim, I.-M.; Kim, B.-G.; Kim, M.-S.; Choi, Y.-H.; Shin, S.-H.; Song, Y.; Park, J.-Y.; Lee, J.-e.; Eun, C.-G.; Lee, H.-C.; Kim, H.-J.; Lee, J.-H.; Kim, J.-Y.; Kweon, T.-M.; Yoon, H.-J.; Kim, T.; Shim, D.-K.; Sel, J.; Shin, J.-Y.; Kwak, P.; Han, J.-M.; Kim, K.-S.; Lee, S.; Lim, Y.-H.; Jung, T.-S. A 7 MB/s 64 Gb 3-bit/cell DDR NAND flash memory in 20 nm node technology, *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, pp. 212, 213, 20–24 Feb. 2011.

[17] Hosono, K.; Tanaka, T.; Imamiya, K.; Sakui, K. A high speed failure bit counter for the pseudo pass scheme (PPS) in program operation for Giga bit NAND flash, *Non-Volatile Semiconductor Memory Workshop, 2003. IEEE NVSMW 2003.* pp. 23–26, 16–20 Feb. 2003.

[18] Park, K.-T. A zeroing cell-to-cell interference page architecture with temporary LSB storing program scheme for sub-40 nm MLC NAND flash memories and beyond, *VLSI Circuits, 2007 IEEE Symposium on*, pp. 188–189, 14–16 June 2007.

[19] Park, K.-T.; Kang, M.; Kim, D.; Hwang, S.-W.; Choi, B. Y.; Lee, Y.-T.; Kim, C.; Kim, K. A zeroing cell-to-cell interference page architecture with temporary LSB storing and parallel MSB program scheme for MLC NAND flash memories, *Solid-State Circuits, IEEE Journal of*, vol. 43, no. 4, pp. 919–928, April 2008.

[20] Takeuchi, Y.; Shimizu, K.; Narita, K.; Kamiya, E.; Yaegashi, T.; Amemiya, K.; Aritome, S. A self-aligned STI process integration for low cost and highly reliable 1 Gbit flash memories, *VLSI Technology, 1998. Digest of Technical Papers. 1998 Symposium on*, pp. 102–103, 9–11 June 1998.

[21] Sakui, K.; Tanaka, T.; Nakamura, H.; Momodomi, M.; Endoh, T.; Shirota, R.; Watanabe, S.; Ohuchi, K.; Masuoka, F. A shielded bitline sensing technology for a high-density and

low-voltage NAND EEPROM design, in *International Workshop on Advanced LSI's*, pp. 226–232, July 1995.

22] Shibata, N.; Tanaka, T. US Patent 7,245,528. 7,370,009. 7,738,302.

23] Hara, T.; Fukuda, K.; Kanazawa, K.; Shibata, N.; Hosono, K.; Maejima, H.; Nakagawa, M.; Abe, T.; Kojima, M.; Fujiu, M.; Takeuchi, Y.; Amemiya, K.; Morooka, M.; Kamei, T.; Nasu, H.; Chi-Ming, Wang; Sakurai, K.; Tokiwa, N.; Waki, H.; Maruyama, T.; Yoshikawa, S.; Higashitani, M.; Pham, T. D.; Fong, Y.; Watanabe, T. A 146-mm^2 8-gb multi-level NAND flash memory with 70-nm CMOS technology," *Solid-State Circuits, IEEE Journal of*, vol. 41, no. 1, pp. 161, 169, Jan. 2006.

24] Shibata, N.; Maejima, H.; Isobe, K.; Iwasa, K.; Nakagawa, M.; Fujiu, M.; Shimizu, T.; Honma, M.; Hoshi, S.; Kawaai, T.; Kanebako, K.; Yoshikawa, S.; Tabata, H.; Inoue, A.; Takahashi, T.; Shano, T.; Komatsu, Y.; Nagaba, K.; Kosakai, M.; Motohashi, N.; Kanazawa, K.; Imamiya, K.; Nakai, H. A 70 nm 16 Gb 16-level-cell NAND Flash Memory, *VLSI Circuits, 2007 IEEE Symposium on*, pp. 190–191, 14–16 June 2007.

25] Shibata, N.; Maejima, H.; Isobe, K.; Iwasa, K.; Nakagawa, M.; Fujiu, M.; Shimizu, T.; Honma, M.; Hoshi, S.; Kawaai, T.; Kanebako, K.; Yoshikawa, S.; Tabata, H.; Inoue, A.; Takahashi, T.; Shano, T.; Komatsu, Y.; Nagaba, K.; Kosakai, M.; Motohashi, N.; Kanazawa, K.; Imamiya, K.; Nakai, H.; Lasser, M.; Murin, M.; Meir, A.; Eyal, A.; Shlick, M. A 70 nm 16 Gb 16-level-cell NAND flash memory, *Solid-State Circuits, IEEE Journal of*, vol. 43, no. 4, pp. 929–937, April 2008.

26] Lee, C.; Lee, S.-K.; Ahn, S.; Lee, J.; Park, W.; Cho, Y.; Jang, C.; Yang, C.; Chung, S.; Yun, I.-S.; Joo, B.; Jeong, B.; Kim, J.; Kwon, J.; Jin, H.; Noh, Y.; Ha, J.; Sung, M.; Choi, D.; Kim, S.; Choi, J.; Jeon, T.; Yang, J.-S.; Koh, Y.-H. A 32 Gb MLC NAND-flash memory with V_{th}-endurance-enhancing schemes in 32 nm CMOS, *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*, pp. 446–447, 7–11 Feb. 2010.

27] Cernea, R.-A.; Pham, L.; Moogat, F.; Chan, S.; Le, B.; Li, Y.; Tsao, S.; Tseng, T.-Y.; Nguyen, K.; Li, J.; Hu, J.; Yuh, J. H.; Hsu, C.; Zhang, F.; Kamei, T.; Nasu, H.; Kliza, P.; Htoo, K.; Lutze, J.; Dong, Y.; Higashitani, M.; Junnhui, Yang; Hung-Szu, Lin; Sakhamuri, V.; Li, A.; Pan, F.; Yadala, S.; Taigor, S.; Pradhan, K.; Lan, J.; Chan, J.; Abe, T.; Fukuda, Y.; Mukai, H.; Kawakami, K.; Liang, C.; Ip, T.; Chang, S.-F.; Lakshmipathi, J.; Huynh, S.; Pantelakis, D.; Mofidi, M.; Quader, K. A 34 MB/s MLC write throughput 16 Gb NAND with all bit line architecture on 56 nm technology *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 1, pp. 186–194, Jan. 2009.

28] Cernea, R.; Pham, L.; Moogat, F.; Chan, S.; Le, B.; Li, Y.; Tsao, S.; Tseng, T.-Y.; Nguyen, K.; Li, J.; Hu, J.; Park, J.; Hsu, C.; Zhang, F.; Kamei, T.; Nasu, H.; Kliza, P.; Htoo, K.; Lutze, J.; Dong, Y.; Higashitani, M.; Yang, J.; Lin, H.-S.; Sakhamuri, V.; Li, A.; Pan, F.; Yadala, S.; Taigor, S.; Pradhan, K.; Lan, J.; Chan, J.; Abe, T.; Fukuda, Y.; Mukai, H.; Kawakamr, K.; Liang, C.; Ip, T.; Chang, S.-F.; Lakshmipathi, J.; Huynh, S.; Pantelakis, D.; Mofidi, M.; Quader, K. A 34 MB/s-program-throughput 16 Gb MLC NAND with all-bitline architecture in 56 nm, *Solid-State Circuits Conference, 2008. ISSCC 2008. Digest of Technical Papers. IEEE International*, pp. 420–624, 3–7 Feb. 2008.

29] Li, Y.; Lee, S.; Fong, Y.; Pan, F.; Kuo, T.-C.; Park, J.; Samaddar, T.; Nguyen, H. T.; Mui, M. L.; Htoo, K.; Kamei, T.; Higashitani, M.; Yero, E.; Kwon, G.; Kliza, P.; Wan, J.; Kaneko, T.; Maejima, H.; Shiga, H.; Hamada, M.; Fujita, N.; Kanebako, K.; Tam, E.; Koh, A.; Lu, I.; Kuo, C. C.-H.; Pham, T.; Huynh, J.; Nguyen, Q.; Chibvongodze, H.; Watanabe, M.; Oowada, K.; Shah, G.; Byungki, Woo; Gao, R.; Chan, J.; Lan, J.;

Hong, P.; Peng, L.; Das, D.; Ghosh, D.; Kalluru, V.; Kulkarni, S.; Cernea, R.-A.; Huynh, S.; Pantelakis, D.; Wang, C.-M.; Quader, K. A 16 Gb 3-bit per cell (X3) NAND flash memory on 56 nm technology with 8 MB/s write rate, *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 1, pp. 195, 207, Jan. 2009.

[30] Li, Y.; Lee, S.; Fong, Y.; Pan, F.; Kuo, T.-C.; Park, J.; Samaddar, T.; Nguyen, H.; Mui, M.; Htoo, K.; Kamei, T.; Higashitani, M.; Yero, E.; Gyuwan, Kwon; Kliza, P.; Jun, Wan; Kaneko, T.; Maejima, H.; Shiga, H.; Hamada, M.; Fujita, N.; Kanebako, K.; Tarn, E.; Koh, A.; Lu, I.; Kuo, C.; Pham, T.; Huynh, J.; Nguyen, Q.; Chibvongodze, H.; Watanabe, M.; Oowada, K.; Shah, G.; Woo, B.; Gao, R.; Chan, J.; Lan, J.; Hong, P.; Peng, L.; Das, D.; Ghosh, D.; Kalluru, V.; Kulkarni, S.; Cernea, R.; Huynh, S.; Pantelakis, D.; Wang, C.-M.; Quader, K. A 16 Gb 3 b/cell NAND flash memory in 56 nm with 8MB/s write rate, *Solid-State Circuits Conference, 2008. ISSCC 2008. Digest of Technical Papers. IEEE International*, pp. 506–632, 3–7 Feb. 2008.

[31] Futatsuyama, T.; Fujita, N.; Tokiwa, N.; Shindo, Y.; Edahiro, T.; Kamei, T.; Nasu, H.; Iwai, M.; Kato, K.; Fukuda, Y.; Kanagawa, N.; Abiko, N.; Matsumoto, M.; Himeno, T.; Hashimoto, T.; Liu, Y.-C.; Chibvongodze, H.; Hori, T.; Sakai, M.; Ding, H.; Takeuchi, Y.; Shiga, H.; Kajimura, N.; Kajitani, Y.; Sakurai, K.; Yanagidaira, K.; Suzuki, T.; Namiki, Y.; Fujimura, T.; Mui, M.; Nguyen, H.; Lee, S.; Mak, A.; Lutze, J.; Maruyama, T.; Watanabe, T.; Hara, T.; Ohshima, S. A 113 mm^2 32 Gb 3b/cell NAND flash memory, *Solid-State Circuits Conference—Digest of Technical Papers, 2009. ISSCC 2009. IEEE International*, pp. 242–243, 8–12 Feb. 2009.

[32] Nobukata, H.; Takagi, S.; Hiraga, K.; Ohgishi, T.; Miyashita, M.; Kamimura, K.; Hiramatsu, S.; Sakai, K.; Ishida, T.; Arakawa, H.; Itoh, M.; Naiki, I.; Noda, M. A 144 Mb 8-level NAND flash memory with optimized pulse width programming, *VLSI Circuits, 1999. Digest of Technical Papers. 1999 Symposium on*, pp. 39–40, 1999.

[33] Nobukata, H.; Takagi, S.; Hiraga, K.; Ohgishi, T.; Miyashita, M.; Kamimura, K.; Hiramatsu, S.; Sakai, K.; Ishida, T.; Arakawa, H.; Itoh, M.; Naiki, I.; Noda, M. A 144-Mb, eight-level NAND flash memory with optimized pulsewidth programming, *Solid-State Circuits, IEEE Journal of*, vol. 35, no. 5, pp. 682–690, May 2000.

[34] Yang, J.; Park, M.; Jung, S.; Park, S.; Cho, S.; An, J.; Lee, J.; Cho, S.; Lee, H.; Cho, M. K.; Ahn, K. O.; Jin, K., Koh, Y. The operation scheme and process optimization in TLC (triple level cell) NAND flash characteristics, SSDM 2009.

[35] Shin, S.-H.; Shim, D.-K.; Jeong, J.-Y.; Kwon, O.-S.; Yoon, S.-Y.; Choi, M.-H.; Kim, T.-Y.; Park, H.-W.; Yoon, H.-J.; Song, Y.-S.; Choi, Y.-H.; Shim, S.-W.; Ahn, Y.-L.; Park, K.-T.; Han, J.-M.; Kyung, K.-H.; Jun, Y.-H. A new 3-bit programming algorithm using SLC-to-TLC migration for 8 MB/s high performance TLC NAND flash memory," *VLSI Circuits (VLSIC), 2012 Symposium on*, pp. 132, 133, 13–15 June 2012.

[36] Li, Y.; Lee, S.; Oowada, K.; Nguyen, H.; Nguyen, Q.; Mokhlesi, N.; Hsu, C.; Li, J.; Ramachandra, V.; Kamei, T.; Higashitani, M.; Pham, T.; Honma, M.; Watanabe, Y.; Ino, K.; Binh, Le; Woo, B.; Htoo, K.; Tseng, T.-Y.; Pham, L.; Tsai, F.; Kim, K.-h.; Chen, Y.-C.; She, M.; Yuh, J.; Chu, A.; Chen, C.; Puri, R.; Lin, H.-S.; Chen, Y.-F.; Mak, W.; Huynh, J.; Chan, J.; Watanabe, M.; Yang, D.; Shah, G.; Souriraj, P.; Tadepalli, D.; Tenugu, S.; Gao, R.; Popuri, V.; Azarbayjani, B.; Madpur, R.; Lan, J.; Yero, E.; Pan, F.; Hong, P.; Jang, Yong Kang; Moogat, F.; Fong, Y.; Cernea, R.; Huynh, S.; Trinh, C.; Mofidi, M.; Shrivastava, R.; Quader, K. 128 Gb 3b/cell NAND flash memory in 19 nm technology with 18 MB/s write rate and 400 Mb/s toggle mode, *Solid-State Circuits Conference*

Digest of Technical Papers (ISSCC), 2012 IEEE International, pp. 436, 437, 19–23 Feb. 2012.

[37] Trinh, C.; Shibata, N.; Nakano, T.; Ogawa, M.; Sato, J.; Takeyama, Y.; Isobe, K.; Le, B.; Moogat, F.; Mokhlesi, N.; Kozakai, K.; Hong, P.; Kamei, T.; Iwasa, K.; Nakai, J.; Shimizu, T.; Honma, M.; Sakai, S.; Kawaai, T.; Hoshi, S.; Yuh, J.; Hsu, C.; Tseng, T.; Li, J.; Hu, J.; Liu, M.; Khalid, S.; Chen, J.; Watanabe, M.; Lin, H.; Yang, J.; McKay, K.; Nguyen, K.; Pham, T.; Matsuda, Y.; Nakamura, K.; Kanebako, K.; Yoshikawa, S.; Igarashi, W.; Inoue, A.; Takahashi, T.; Komatsu, Y.; Suzuki, C.; Kanazawa, K.; Higashitani, M.; Lee, S.; Murai, T.; Lan, J.; Huynh, S.; Murin, M.; Shlick, M.; Lasser, M.; Cernea, R.; Mofidi, M.; Schuegraf, K.; Quader, K. A 5.6MB/s 64Gb 4b/Cell NAND Flash memory in 43 nm CMOS, *Solid-State Circuits Conference—Digest of Technical Papers, 2009. ISSCC 2009. IEEE International*, pp. 246–247, 247a, 8–12 Feb. 2009.

[38] Tanaka, T.; Tanzawa, T.; Takeuchi, K. A 3.4-Mbyte/sec programming 3-level NAND flash memory saving 40% die size per bit, *VLSI Circuits, 1997. Digest of Technical Papers., 1997 Symposium on*, pp. 65–66, 12–14 June 1997.

[39] Park, K.-T.; Choi, J.; Cho, S.; Choi, Y.; Kim, K. A high cost-performance and reliable 3-level MLC NAND flash memory using virtual page cell architecture, *Non-Volatile Semiconductor Memory Workshop, 2006. IEEE NVSMW 2006, 21st*, pp. 34–35, 12–16 Feb. 2006.

5

SCALING CHALLENGE OF NAND FLASH MEMORY CELLS

5.1 INTRODUCTION

Low-cost and highly reliable NAND flash memory technologies have been intensively developed [1–9] over 25 years, as described in Chapter 3. As a suitable memory cell structure for NAND flash, the self-aligned STI cell (SA-STI cell) had been developed [4–7] and implemented to NAND flash products [8]. This cell could reduce memory cell size to ideal $4*F^2$ [4], and had also demonstrated an excellent reliability, because the floating gate does not overlap the STI corner. Thus, the SA-STI cell structure and process have been used for more than 15 years and 10 generations of NAND flash product. The most advanced memory cell had presented as mid-1X-nm (15 to 16-nm) SA-STI memory cells [10], as shown in a cross-sectional TEM micrograph in Fig. 5.1. The effective cell size can be also reduced by multilevel cell technology, as described in Chapter 4. Therefore, the small physical cell size of $4*F^2$ combined with a multilevel cell can drastically reduce the bit cost of NAND flash memory.

However, by scaling memory cell size beyond the 20-nm generation, it is becoming very difficult to realize high-performance and highly reliable NAND flash memory, because many physical phenomena have a serious impact on the operation margin of NAND flash [11].

In Chapter 5, the scaling challenges of the NAND flash memory cell with a multilevel cell are discussed beyond 20-nm feature sizes. One important physical phenomenon is the floating-gate capacitive coupling interference [12] that causes a V_t shift by programming neighbor cells. An increase in V_t distribution width (Section 5.3)

Nand Flash Memory Technologies, First Edition. Seiichi Aritome.

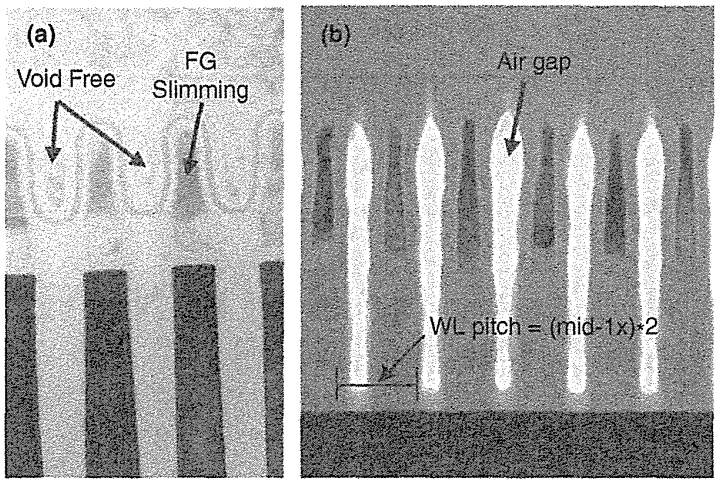


FIGURE 5.1 TEM photograph of mid-1X-nm SA-STI NAND flash cells.

will result in the degradation of read window margin (RWM). The other major physical phenomena to have an impact on RWM are electron injection spread (EIS) [13–15] (Chapter 5.4) and random telegraph noise (RTN) [16] (Section 5.5). Except for the RWM degradation, there are several other problems, such as CG formations between FGs [17] (Chapter 5.6), the WL high-field problem [11, 18] (Section 5.7), reducing the number of stored electrons [19] (Section 5.8), and so on.

The scaling capability of NAND flash memory has been discussed in several conferences and papers [20–33]. They pointed out major scaling limitations, such as floating-gate capacitive coupling interference [21–23,25–28,31,32], reduced number of electrons [21–23], lithograph/patterning [22–24], RTN and RDF (random dopant fluctuation) [23], structure limitation [25,28,30], air gap [34,35], V_t window margin [11,28,30], and so on.

In Chapter 5, several scaling problems and limitations have been widely discussed over 2X to 0X-nm generations [11]. As a result, there is a possibility that the NAND flash memory cell can be scaled down to 1Z-nm (10-nm) generation with an accurate control of FG/CG formation process and air-gap process to manage floating-gate capacitive coupling interference and the WL high-field problem.

5.2 READ WINDOW MARGIN (RWM)

The read window margin (RWM) of a self-aligned STI cell (SA-STI cell) is discussed for NAND flash memories over 2X to 0X-nm generations in Section 5.2 [11]. The RWM is investigated by extrapolating the physical phenomena of FG–FG coupling interference (floating-gate capacitive coupling interference), electron injection spread (EIS), back pattern dependence (BPD), and random telegraph noise (RTN). The RWM is degraded not only by increasing programmed V_t distribution width, but also by increasing the V_t of the erase state mainly due to the large FG–FG coupling

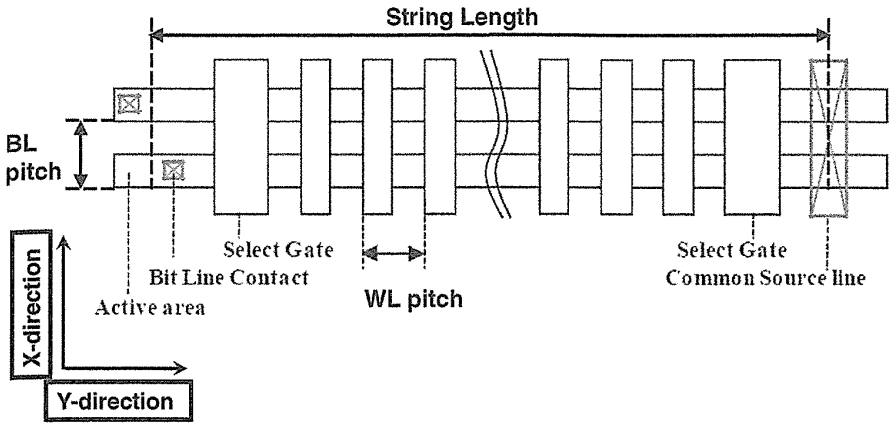


FIGURE 5.2 Top view of a NAND cell string. 64 cells are connected in series with two select gates. BL pitch and WL pitch are nearly equal to 2F (F: feature size). Then unit cell size is close to ideal $4*F^2$.

interference. However, RWM is still positive in the 1Z-nm (10-nm) generation with 60% reduction of FG–FG coupling interference by the air-gap process. Therefore, the SA-STI cell is expected to be able to scale down to the 1Z-nm (10-nm) generation, with the air gap of 60% reduced FG–FG coupling interference.

5.2.1 Assumption for Read Window Margin (RWM)

Figure 5.2 shows a top view of conventional NAND cell strings. In order to investigate the scaling of the NAND flash cell, cell dimensions beyond the 2X-nm (26-nm) generation are assumed, as shown in Table 5.1. Dimensions of 2X nm are given, 27 nm for the bit line (BL) half-pitch and 26 nm for the word-line (WL) half-pitch. Dimensions beyond 2X nm are assumed to scale down by a fixed scaling factor of ×0.85 for BL half-pitch and ×0.8 for the WL half-pitch. And also the channel width

TABLE 5.1 Assumption of Cell Dimensions and ONO (IPD) Thickness, in Generations of 2X–0X Nanometers[a]

Generation	2X	2Y	1X	1Y	1Z	0X	Scaling factor
BL half-pitch (nm)	*27*	23.0	19.5	16.6	14.1	12.0	×0.85 assumption
WL half-pitch (nm), Gate length *L*	26	20.8	16.6	13.3	10.6	8.5	×0.8 assumption
Channel W (nm)	20	18.0	16.2	14.6	13.1	11.8	×0.9 assumption
ONO thickness (nm)	*12*	11.4	10.8	10.3	9.8	9.3	×0.95 assumption

[a]Dimensions of 2X-nm generation are given, as 27 nm for BL half-pitch and 26 nm for WL half-pitch. Dimensions of ~2Y nm are assumed by scaling factors of ×0.85 for BL half-pitch (*x*-direction) and ×0.8 for the word line (WL) half-pitch (*y*-direction). And also, the scaling factors of channel width (*W*) and ONO thickness are assumed ×0.9 and ×0.95, respectively.

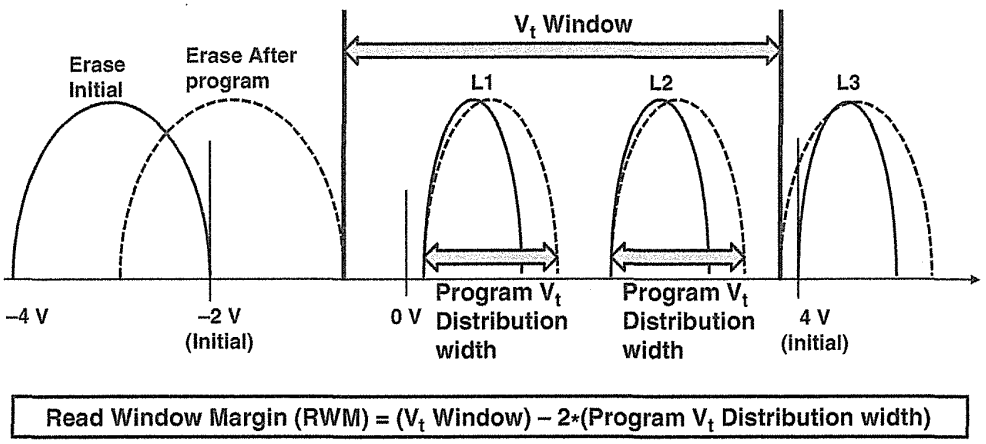


FIGURE 5.3 Read V_t window of an MLC NAND cell. V_t distributions of erase state and programmed L1, L2, L3 states are shifted up and become wider because of electron injection spread (EIS), FG–FG coupling interference, RTN, and back pattern dependence (BPD). Read window margin (RWM) is defined as RWM = (V_t window) − 2*(program V_t distribution width).

W and inter-poly dielectric (IPD) thickness are assumed to scale down by the factors of ×0.9 and ×0.95, respectively.

Figure 5.3 shows an image of a read V_t window in an MLC (2 bits/cell) NAND cell [11]. The "V_t window" is defined by a right-side edge of erase distribution and a left-side edge of L3 (highest programmed state) after completing all page program operations in block (strings). Two programmed V_t distributions of L1/L2 have to be inside of the V_t window to be a reliable read operation. Read window margin (RWM) is defined by RWM = (V_t window) − 2*(programmed V_t distribution width), so that RWM means the separation margin of V_t distributions of each states.

The RWMs have been seriously degraded by cell scaling down from 0.7 μm to 2X-nm generation, because several physical phenomena were getting worse. Therefore, for further scaling of a NAND cell, it is very important to analyze and foresee the RWM in a future scaled NAND cell. In order to investigate RWM, the scaling trend of physical phenomena of electron injection spread (EIS) [13–15], FG–FG coupling interference [12], RTN [16], and back pattern dependence (BPD) are assumed as follows. And other assumptions of the page program sequence, parameter setting, and so on, are also shown in the following.

Assumption of RWM Calculation

(a) V_t distribution width (@ $\pm 3\sigma$) is assumed to become wider by simple summation of values of electron injection spread (EIS), FG–FG coupling interference, RTN and back pattern dependence (BPD). Each value is given for 2X-nm generation and is extrapolated for 2Y-nm to 0X-nm generations with the following formulas.

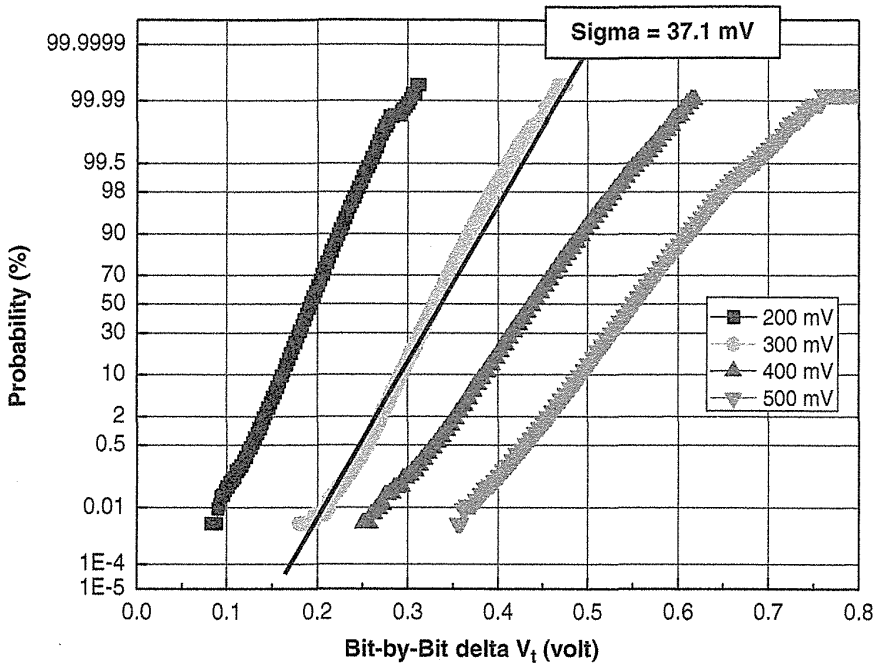


FIGURE 5.4 Electron injection spread (EIS) of a 2X-nm cell. In ISPP_step = 300 mV, the standard deviations (σ) is 37.1 mV. With the FG depletion effect, σ is assumed to be larger of 50%. Then the σ is assumed 55.6 mV for a 300-mV ISPP_Step.

(b) Program electron injection spread (EIS) [13–15] is caused during program operation due to statistical spread in a small number of injecting electrons during program pulse (see Section 5.4). The σ of EIS is linear with sqrt(q∗ISPP_Step/C_{IPD}) [13, 14], and three σ values are simply used for V_t distribution widening. Capacitance of inter-poly dielectric, C_{IPD}, is scaled down from 2X-nm generation by ×0.72 for each generation. The ISPP_step is a program voltage step of ISPP (increment step pulse program) [36, 37]. The measured standard deviations (σ) is 37.1 mV for ISPP_step = 300 mV, as shown in Fig. 5.4. A value of the sigma is assumed to be 50% larger due to FG depletion effects [38], and so on. Then the σ of 2X is assumed to be 55.6 mV for ISPP_Step = 300 mV, and 78.7 mV for ISPP_Step = 600 mV.

(c) FG–FG coupling interference (floating-gate capacitive coupling interference) [12] (see Section 5.3) is scaled from 2X-nm generation to each generation by ×(1/0.9) along WL, ×(1/0.8) along BL, and ×(1/0.85) diagonal. And spread effect (additional V_t shift) is assumed 10% of the FG–FG coupling value. FG–FG coupling values of 2X nm are assumed to be 94 mV/V for two sides of x-direction (between BL–BL), 85 mV/V for one side of y-direction (between WL–WL), and 25 mV/V for two sides of xy-direction (diagonal), based on measurement results. Scaling factors of FG–FG coupling (×(1/0.9) along WL, ×(1/0.8) along BL, and ×(1/0.85) diagonal) are the simple assumption by increasing FG–FG capacitance with decreasing FG–FG distance. This

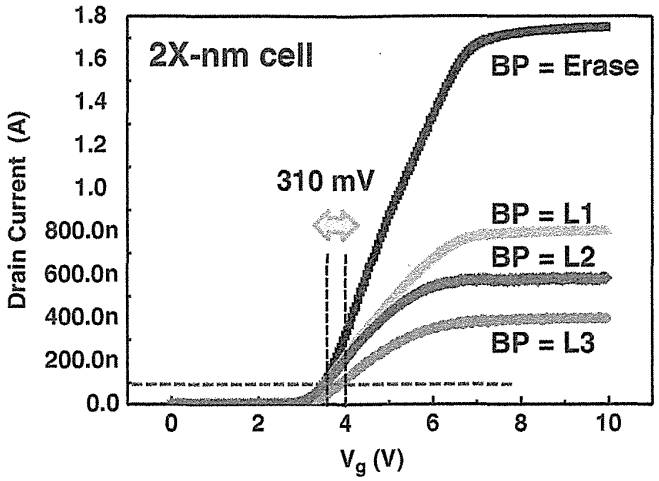


FIGURE 5.5 Back pattern dependence (BPD) of a 64-cell string in a 2X-nm cell. The V_t shift of BPD is assumed to be 310 mV as the worst case of BP = L3 (unselected cells are in L3).

simple assumption would be optimistic because it is not including the effect of decreasing total FG capacitance, which is mainly due to less scaling of IPD and tunnel-oxide thickness.

(d) RTN is linear with $1/(W*\text{sqrt}(L))$ [39] (see Section 5.5). The value of 2X nm is assumed to be ± 107 mV@ 3σ, based on measurement results.

(e) Back pattern dependence (BPD) is a V_t shift, which is caused by programming series-connected cells in the same string due to increasing series resistance in string. BPD is linear with L/W. The value of 2X nm is assumed to 310 mV, as shown in Fig. 5.5.

(f) The page program sequence uses the minimized FG–FG coupling program sequence [40,41], as shown in Fig. 5.6. It means that, before the MSB program, the surrounding pages (LSB of WLn − 1, WLn, WLn + 1, and MSB of WLn − 1) have already been programmed. FG–FG coupling interference can be minimized for the programmed V_t distributions (see Section 4.3).

(g) All-bit-line scheme (ABL) [42] (see Section 4.3). The V_t shift value of the x-direction FG–FG coupling interference is assumed that it is based on neighbor cell V_t shift of $3*\sigma*\text{SQRT}(2)*(1/2)$ [σ = (V_t distribution width ($\pm 3\sigma$) of one program pulse)/6 = 3V/6 = 0.5 V] [$*(1/2)$; factor of random data pattern], because neighbor cells are programmed with target cells at the same time.

(h) Random data pattern.

(i) Erase; initial V_t distribution; −3 V $\pm$1 V (V_t distribution width = 2 V). The right-side edge of erase initial is −2 V.

(j) L1 verify level = 0.5 V, L2 verify level = 2.25 V, L3 verify level = 4.0 V, LSB verify level = 0.8 V, except for the case of V_t setting dependence in Section 5.2.5.

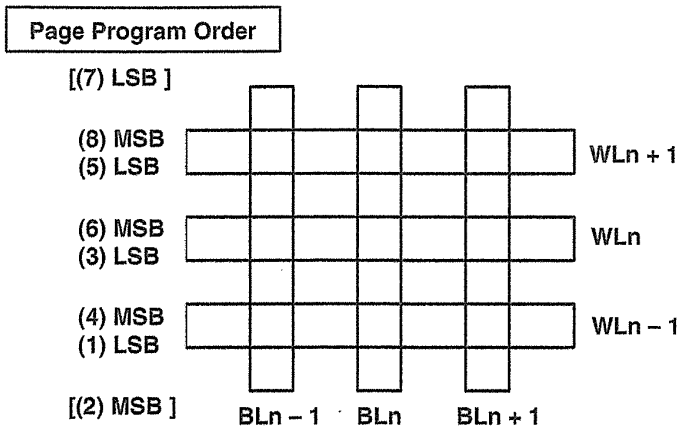


FIGURE 5.6 Page program sequence of the minimized FG–FG coupling interference. Before (6) MSB program of WLn, the surrounding pages [(1) LSB of WLn − 1, (3) LSB of WLn, (5) LSB of WLn + 1, and (4) MSB of WLn − 1] have already been programmed. Then the FG–FG coupling interference can be minimized for the programmed V_t distributions.

(k) ISPP step (increment step pulse program [36,37]) of LSB and MSB programming are 600 mV and 300 mV, respectively.

(l) The data retention V_t shift is not included in this RWM investigation, because it is expected to be managed by the multi-times read operations (moving read algorithm), as described in Section 4.7. Also, program disturb, read disturb, and other effects are not included in this RWM investigation.

5.2.2 Programmed V_t Distribution Width

In conventional program operation, a programmed V_t distribution width can be tight by using ISPP [36, 37] (see Section 4.2.2) and a bit-by-bit verify operation [43] (see Section 4.2.3). The initial programmed V_t distribution width is determined by ISPP_step + EIS. And then it becomes wider by RTN, FG–FG coupling interference, and BPD after all pages are programmed in a block (string).

The programmed V_t distribution width after all pages have been programmed in a block have been calculated based on the assumption of Section 5.2.1, as shown in Fig. 5.7. As memory cells are scaled down from 2X nm to 0X nm, the programmed V_t distribution width is increased from 1320 mV to 2183 mV. It is clear that major reasons to increase V_t distribution width are the FG–FG coupling and RTN.

In order to obtain appropriate V_t shift values of FG–FG coupling interference, the delta V_t of the neighbor attack cell (subject to target cell) have been derived, as shown in Fig. 5.8. V_t distributions of page programming steps are also described in Fig. 5.8. For FG–FG coupling for the programmed states, delta V_t of the attack cell is described as dVt_E_L1 or dV_t_LSB_L2 + dV_t_LSB_L3, as shown at (3) after the MSB program in Fig. 5.8. A dV_t_*E*_L1 means delta V_t shift from erase state (@2) before MSB program) to L1 state. Larger value of dV_t_*E*_L1 or dV_t_LSB_L2 + dV_t_LSB_L3 is used for calculation of FG–FG coupling V_t shift.

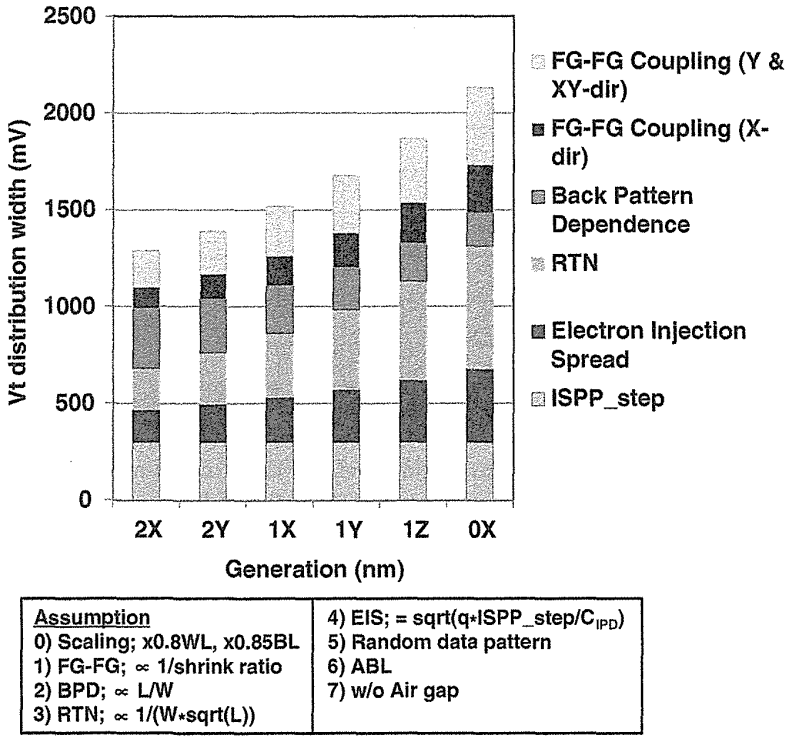


FIGURE 5.7 Calculated programmed V_t distribution width. The V_t distribution width is increased by cell dimension scaling. Major impact factors to increase V_t distributions width are the FG–FG coupling and RTN.

V_t shift value of x-direction FG–FG coupling interference is assumed that it is based on neighbor cell V_t shift of $3*\sigma*\text{SQRT}(2)*(1/2)$ [standard deviation; σ = (V_t distribution width ($\pm 3\sigma$) of one program pulse)/6 = 3 V/6 = 0.5 V] [$*(1/2)$; factor of random data pattern], as shown in Section 5.2.1g), because neighbor cells are programmed with target cells at the same program sequence in the all-bit-line scheme. V_t shift had been assumed as follows. Cells in programmed V_t distribution (3-V width) are programmed to shift up by ISPP program. A certain cell (cell A) stops programming by passing verify at threshold voltage of V_t_cell A, and neighbor cells (cell B) have not passed verify yet at threshold voltage of V_t_cell B. The neighbor cells (cell B) are programmed by following ISPP steps, then it causes FG–FG coupling on cell A with a V_t difference of (V_t_cell A − V_t_cell B). In this assumption, the distribution of V_t difference (V_t_cell A − V_t_cell B) is assumed to composition of V_t distribution (3-V width), then the σ of the V_t shift is $\text{SQRT}(\sigma^2 + \sigma^2) = \sigma*\text{SQRT}(2)$. Also, we assumed that the same FG–FG coupling V_t shift occurs between L1 and L2, by assuming to use preferable program operations, such as the ABL parallel program method [44], the BC state first program algorithm [45], and the P3-pattern pre-pulse scheme [46], to reduce FG–FG coupling for both L1 and L2.

Erase V_t distribution at (2) Before MSB program in Fig. 5.8 has already shifted up as $dV_t\_E_i\_E$ from erase initial V_t distribution, by FG–FG coupling with surrounding

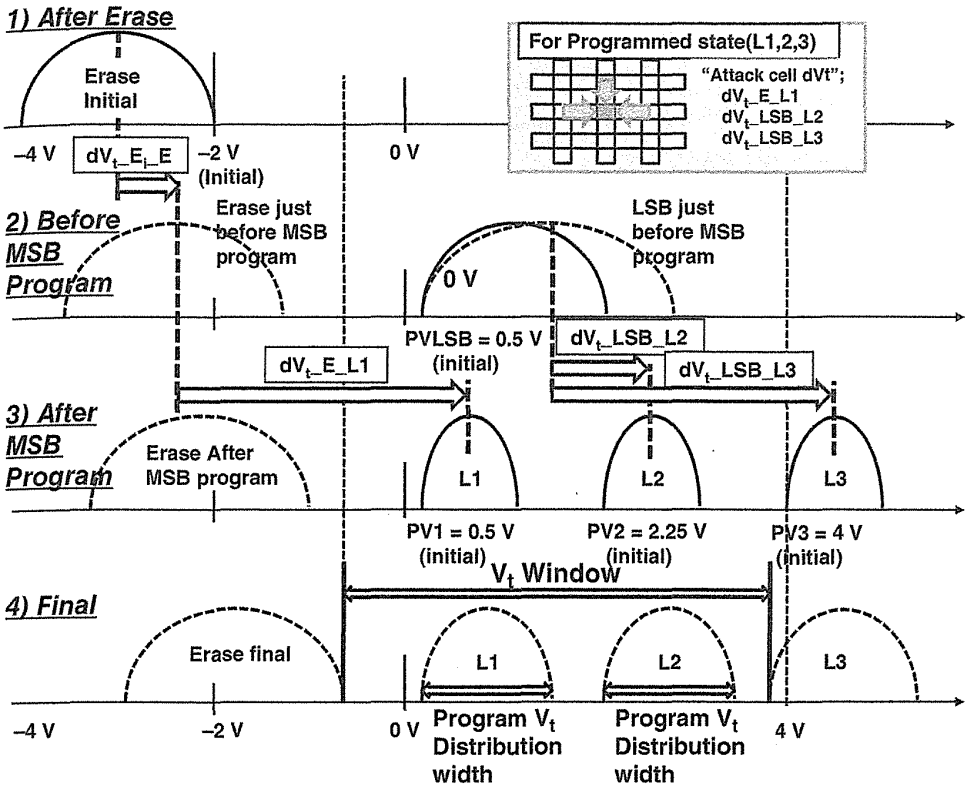


FIGURE 5.8 V_t distribution in page program steps. The attack cell delta V_t of $dV_t\_E\_L1$ or $dV_t\_LSB\_L2 + dV_t\_LSB\_L3$ are subjected to the programmed target cell of y-direction neighbor cells, resulting in wider V_t distribution width by the FG–FG coupling interference. Then distribution width of the programmed cell becomes wider from (3) After MSB program to (4) Final.

cells of LSB program (both sides of Y-direction/XY-direction/X-direction) and MSB program (one side of Y-direction/XY-direction), as shown in Fig. 5.6. By cell scaling, $dV_t\_E_i\_E$ becomes larger due to larger FG–FG coupling interference. Then $dV_t\_E\_L1$ becomes smaller in value as a result of cell scaling. Therefore, y-direction FG–FG coupling interference for programmed states is relatively smaller than expected, as shown in Fig. 5.7.

5.2.3 V_t Window

V_t window is defined from the right-side edge of erase V_t distribution to the left-side edge of L3 V_t distribution, as shown in Fig. 5.3. Figure 5.9 shows the calculation results of V_t window, the right-side edge of erase V_t distribution, and the left-side edge of L3 V_t distribution, in three cases of reducing FG–FG coupling of 0%, 30%, and 60% by air gap [34,35,47,48] or low-k dielectric. The reducing FG–FG coupling is assumed for both the x-direction (STI air-gap [35, 49]) and the y-direction (WL air-gap [34,47,48]).

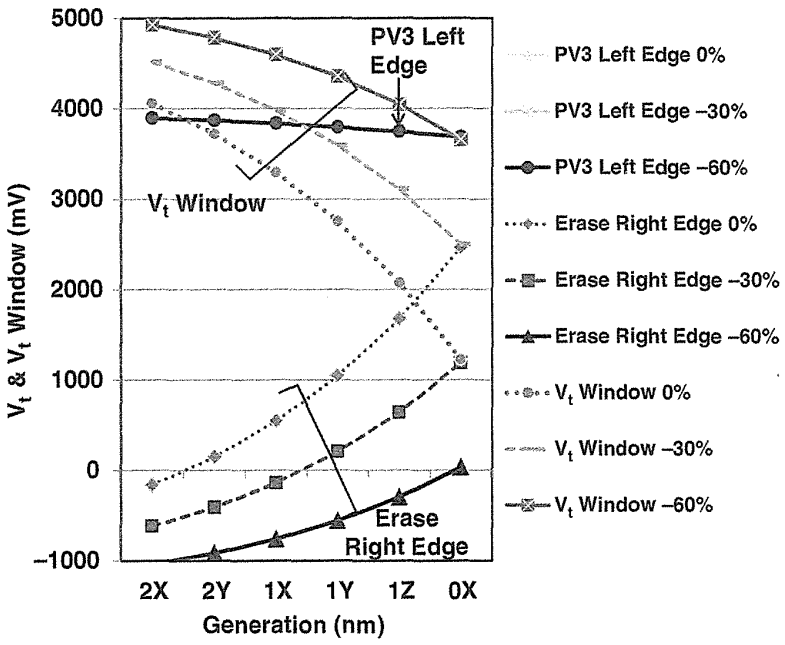


FIGURE 5.9 Calculated V_t window as a function of cell scaling down. V_t window is decreased mainly by increasing the erase right edge (right-side edge of erase). If FG–FG coupling interference can be reduced to −30% or −60%, the erase right edge can be much improved.

As shown in Fig. 5.9, the V_t window becomes seriously narrower as a result of cell scaling in the case of conventional 0% FG–FG coupling reduction (see "V_t window 0%"). This is because the right-side edge of the erase distribution is much increased as a result of scaling. However, in the case of −60% FG–FG coupling reduction, the right-side edge of erase distribution can be kept less than 0 V even in the 1Z-nm generation. Then, a V_t window can be kept more than 4000 mV in 1Z-nm generation.

In order to clarify the reason of increasing the right-side edge of erase, factors of increasing erase right-side edge are analyzed, as shown in Fig. 5.10. The erase right edge is increased mainly by FG–FG coupling, especially by *Y*- & *XY*-direction FG–FG coupling. For the erase state, the FG–FG coupling V_t shift is much larger than the FG–FG coupling V_t shift of the programmed states. Figure 5.11 shows the reasons of large FG–FG coupling for erase states. There are two reasons. One is the large delta-V_t of an attack cell, as shown in Fig. 5.11a. This is because a $dV_t\_E_i\_L1$, L2, L3 (attack cell delta V_t from an erase initial state to each programming state L1, L2, L3) is much larger, in comparison with attack cell V_t shift for a programmed state, such as $dV_t\_E\_L1$ or $dV_t\_LSB\_L2 + dV_t\_LSB\_L3$, as shown in Fig. 5.8. The other reason is that all of surrounding cells are subjected to cause FG–FG coupling V_t shift for the erase state (Fig. 5.11b). Conversely, for the programmed cell, only part of the surrounding cells (one side of *y*-direction [between WL–WL] and *x*-directions [between BL–BL]) have caused an FG–FG coupling V_t shift, as shown in Fig. 5.11b.

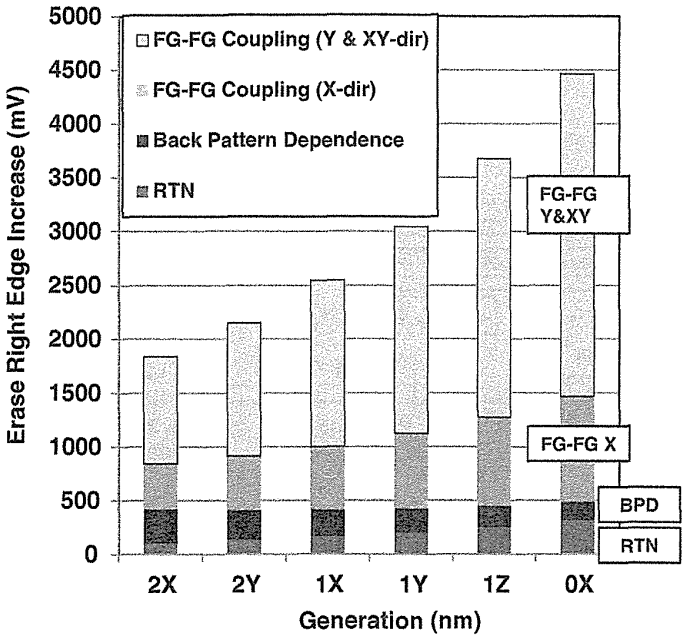


FIGURE 5.10 Increasing the right-side edge of the erase state distribution, as a function of scaling memory cells, in the case of w/o reduction of the FG–FG coupling. The erase right edge is increased mainly by the FG–FG coupling interference, especially by *Y* & *XY*–FG–FG coupling.

In order to obtain a wider V_t window for 1Y and 1Z generations, it is important to reduce FG–FG coupling, especially FG–FG coupling of the *Y*- & *XY*-directions. WL air gap (or low-*K*) [34,47,48] and STI air gap [35,49] have to be implemented as small FG–FG coupling (as small as possible) for future NAND cells.

Furthermore, the optimistic scaling factors of FG–FG coupling are used in this calculation, as described in Section 5.2.1c. Even if the optimistic values are used, the dominant factor of V_t window degradation is the FG–FG coupling. Therefore it is important to reduce FG–FG coupling for future scaled cells.

5.2.4 Read Window Margin (RWM)

Figure 5.12 shows the scaling trend of RWM, which is calculated by the programmed V_t distribution width in Fig. 5.7 and V_t window in Fig. 5.9. RWMs are degraded as a cell scaling. In the case of "no air gap," 1X nm has marginal RWM, and 1Y-nm generation has negative (−719 mV) RWM. In the case of "FG–FG coupling −30% air gap," 1Y nm becomes a marginal RWM, and 1Z-nm generation has a negative RWM. Also, in the case of "FG–FG coupling −60% air gap," 1Z-nm generation has still positive RWM. This means that 30% FG–FG coupling reduction is needed to implement a 1Y-nm generation cell, and 50–60% reduction is needed for a 1Z-nm cell.

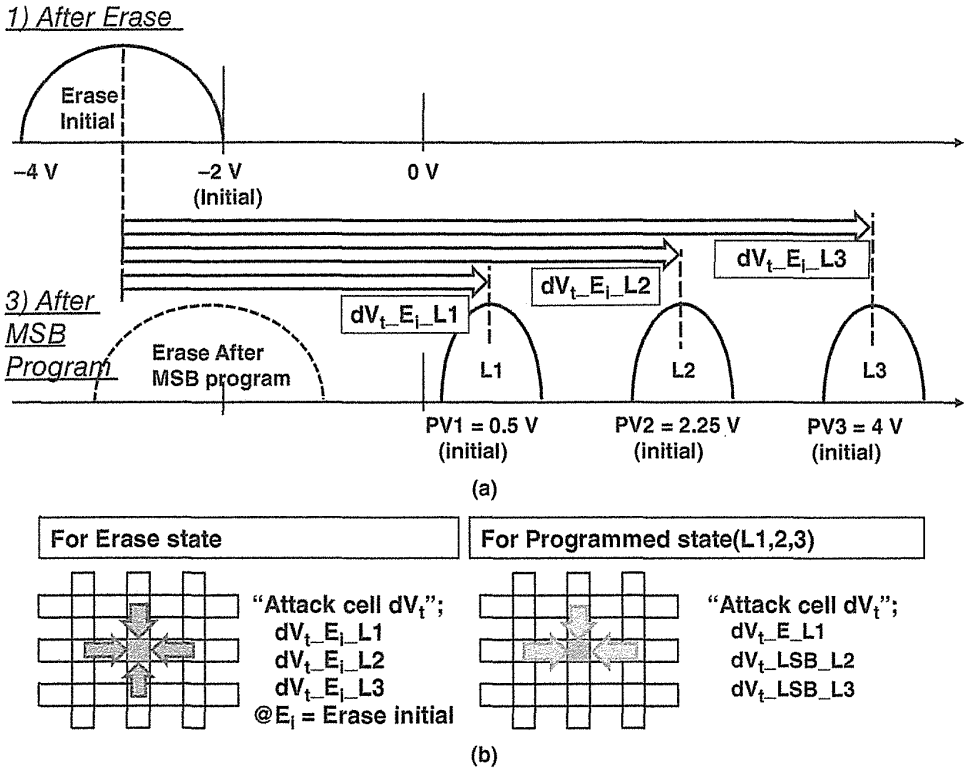


FIGURE 5.11 The FG–FG coupling for the erase state. (a) The attack cell (neighbor of target cell) V_t shift of $dV_t\_E_i\_L1$ or $dV_t\_E_i\_L2$ or $dV_t\_E_i\_L3$ are subjected to the erased target cell. The V_t shift of the erased state is larger than that of programmed state because an attack cell V_t shift for erased state is larger than that of programmed state (see Fig. 5.10), and also (b) all of the surrounding cells have been subjected to erased cells, compared to the fact that all of surrounding cells have not been subjected to a programmed state.

5.2.5 RWM V_t Setting Dependence

In order to find out other solutions for wider RWM, V_t setting dependence has been investigated. Figure 5.13 shows that RWM depends on V_t setting in the case of 1Z nm with −30% FG–FG coupling reduction. In a previous section, PV3 and the erase initial right edge are used for a fixed value of 4 V and −2 V, respectively. In this chapter, lower PV3 and lower erase initial right edge are assumed, as shown in Fig. 5.13.

RWM can be improved in the case of decreasing erase initial right edge, even if programmed V_t distribution widths are slightly increased. And RWM becomes positive in the case of erase initial right edge = −4 V. However, the decreasing erase V_t setting would degrade reliability because of subjecting higher erase voltage stress. Then, in order to obtain wider RWM in the 1Z-nm generation, the air-gap process of minimized FG–FG coupling should be combined with the decreasing erase V_t setting.

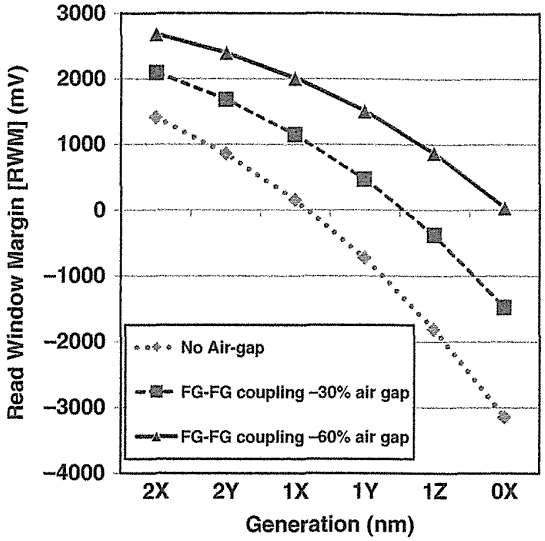
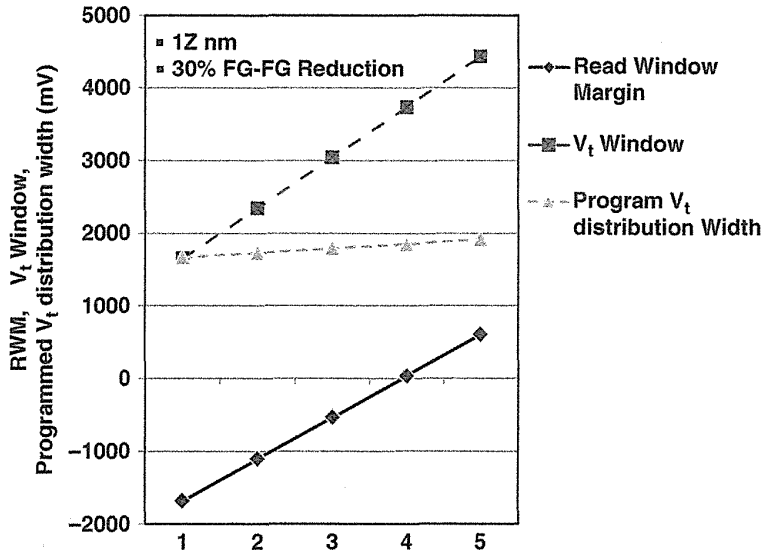


FIGURE 5.12 Calculated read window margin (RWM), in no air gap, 30% or 60% reduction of FG–FG coupling. RWM becomes less than 0 V beyond 1X-nm generation in the case of no air gap. However, by using an air gap with −60% FG–FG coupling reduction, RWM can be kept positive even if 1Z-nm generation is used.



PV3	2.0 V	3.0 V	4.0 V	←	←
PV2	0.75 V	1.5 V	2.25 V	2.0 V	1.75 V
PV1	−0.5 V	0 V	0.5 V	0 V	−0.5 V
PVLSB	−0.2 V	0.3 V	0.8 V	0.3 V	−0.2 V
Erase Initial Right Edge	−2.0 V	←	←	−3.0 V	−4.0 V

V_t Setting

FIGURE 5.13 RWM and V_t window in 1Z-nm generation in the case of −30% FG–FG coupling. RWM increases by decreasing erase V_t setting.

5.3 FLOATING-GATE CAPACITIVE COUPLING INTERFERENCE

Floating-gate capacitive coupling interference (FG–FG coupling) [12] is a major limitation issue to scale down floating-gate NAND flash memory cell, because the read window margin (RWM) is mainly degraded by the floating-gate capacitive coupling interference [11], as described in Section 5.2. As feature sizes (F) have scaled, the floating-gate to floating-gate space has become smaller to cause a V_t shift by V_t change of the eight adjacent cells. This scaling problem results in widening V_t distribution width.

5.3.1 Model of Floating-Gate Capacitive Coupling Interference

In the old concept of large cell size, the floating-gate voltage was determined by only the control-gate voltage with a coupling ratio of $CR = C_{IPD}/C_{total}$, where C_{IPD} is the control-gate to floating-gate capacitance and C_{total} is the total capacitance of the floating gate, as expressed in (5.1) (assuming floating-gate charge $Q_{FG} = 0$).

$$V_{FG} = \frac{C_{IPD}}{C_{TUN} + C_{IPD}} V_{CG} = \frac{C_{IPD}}{C_{total}} V_{CG} = CR * V_{CG} \tag{5.1}$$

where C_{TUN} is the capacitance of substrate to floating gate.

As the design rule of NAND flash memory is scaled down, parasitic capacitors (C_{FGX}, C_{FGY}, C_{FGXY}, C_{FGCG}, and C_{FGAA}) surrounding the floating gate, as shown in Fig. 5.14, have relatively become larger. They cannot be neglected. The floating-gate

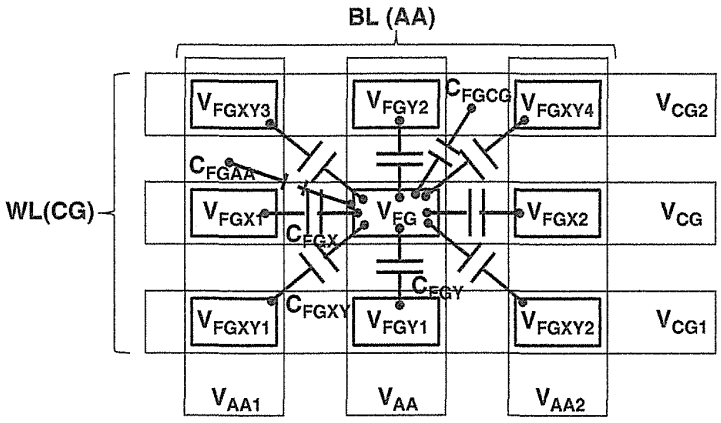


FIGURE 5.14 The model of floating-gate capacitive coupling interference based on parasitic capacitance coupling.

voltage is determined by not only the corresponding control-gate voltage but also the voltages of the surrounding floating gates, the control gates, and active area, as shown in (5.2) (assuming floating-gate charge $Q_{FG} = 0$),

$$V_{FG} = \frac{\begin{array}{c} C_{IPD}V_{CG} + C_{FGX}(V_{FGX1} + V_{FGX2}) + C_{FGY}(V_{FGY1} + V_{FGY2}) + \\ C_{FGXY}(V_{FGXY1} + V_{FGXY2} + V_{FGXY3} + V_{FGXY4}) + \\ C_{FGCG}(V_{CG1} + V_{CG2}) + C_{FGAA}(V_{AA1} + V_{AA2}) \end{array}}{C_{TUN} + C_{IPD} + 2C_{FGX} + 2C_{FGY} + 4C_{FGXY} + 2C_{FGCG} + 2C_{FGAA}} \quad (5.2)$$

where the variables are shown in Fig. 5.14. A phenomenon called "floating-gate capacitive coupling interference," occurs, in which a cell V_t change (ΔV_t) is caused by the threshold voltage shift of the adjacent cells by floating-gate voltage shift (ΔV_{fg}). In other words, the floating-gate voltage is coupled by the floating-gate voltage changes of the adjacent cells with parasitic capacitors in the same manner as the control-gate voltage, as shown in (5.1). For example, if the floating-gate voltage of upper y-direction (V_{FGY2}) is changed by ΔV_{FGY2}, the floating-gate voltage (V_t) change of the target cell causes ΔV_{FG}, as expressed in (5.3).

$$\begin{aligned} \Delta V_{FG} &= \frac{C_{FGY}}{C_{TUN} + C_{IPD} + 2C_{FGX} + 2C_{FGY} + 4C_{FGXY} + 2C_{FGCG} + 2C_{FGAA}} * \Delta V_{FGY2} \\ &= \frac{C_{FGY}}{C_{total}} * \Delta V_{FGY2} \end{aligned} \quad (5.3)$$

In the first report of the floating-gate capacitive coupling interferences [12], a three-dimensional (3-D) capacitance simulator was used to obtain the floating-gate capacitive coupling interference in 0.12-μm design rule cell (gate length = gate space = floating-gate height = channel width = 0.12 μm, tunnel-oxide thickness = 7.5 nm, IPD (ONO) thickness = 15.5 nm). If a neighbor cell is programmed from $V_t = -3$ V to $V_t = 2.2$ V, there are floating-gate interferences of 0.19 V in the y-direction, 0.04 V in the x-direction, and 0.01 V in the diagonal direction (xy-direction).

The floating-gate capacitive coupling interference has a linear characteristic with respect to the adjacent cell V_t change, as derived from (5.2) or (5.3). Figure 5.15 demonstrates the measurement results of the floating-gate capacitive coupling interference on a 0.12-μm design-rule cell [12]. The cell V_t shift by floating-gate interference is linearly proportional to the adjacent cell V_t change. The interference can be reduced significantly with a silicon oxide spacer as compared to a silicon nitride spacer, due to lower parasitic capacitance.

Figure 5.16 shows the simulation results of the V_{th} shift caused by floating-gate capacitive coupling interference with cell technology node scaling [50]. In 3-D TCAD simulations, a 63-nm memory cell has 8-nm tunnel-oxide thickness, 15-nm ONO thickness, and 85-nm floating-gate height. And the memory cell transistor has been scaled down from 63 to 20 nm. Along with cell size reduction, the field oxide recess is kept as +5 nm, and the doping concentration is adjusted to prevent

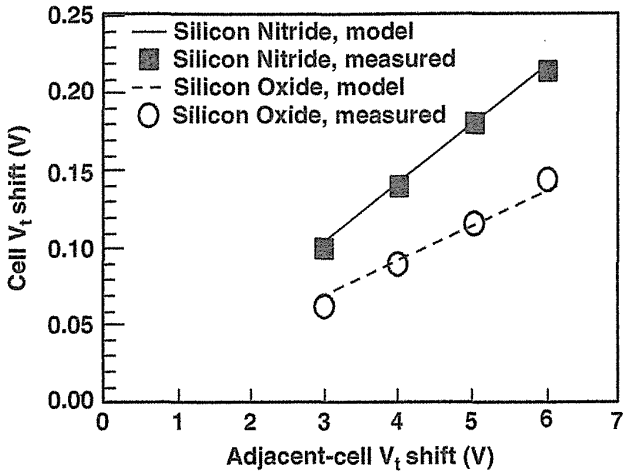


FIGURE 5.15 Floating-gate interference measurement results measured on a 0.12-μm design-rule cell array. The interference was measured on a WL8 cell as a function of WL9-cell V_t change for silicon nitride and silicon oxide spacer samples. Threshold voltage shift of the WL8 cell is monitored before and after WL9-cell programming from $V_t = -3$ V to 2.2 V. Each data consists of 15 points of a WL8 cell.

the cell transistor from punch-through. V_{th} shift by floating-gate capacitive coupling interference is drastically increased as technology node scaling.

Figure 5.17 shows the floating-gate capacitive coupling interference as a function of the technology node [23]. Floating-gate capacitive coupling interference as a percentage of the total V_t shift is remarkably increased by scaling the technology

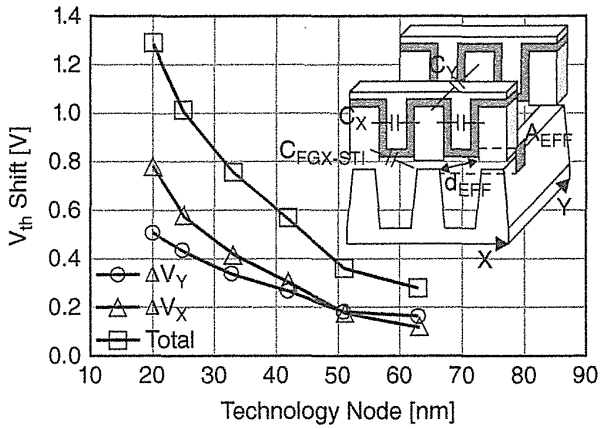


FIGURE 5.16 Simulation results of the V_{th} shift caused by cell-to-cell interference with cell size reduction. By changing the neighboring cell transistor V_{th} from −5 to 5 V, the V_{TH} shift of the reference cell is measured from 1.0 V of the initial V_{th}. Other word lines possess 6.5 V of the pass-gate voltage in a read operation. ΔV_X is the cell V_{th} shift induced by two adjacent cell transistors in the x-direction (word-line direction), while ΔV_Y is induced by a cell transistor in the y-direction (bit-line direction).

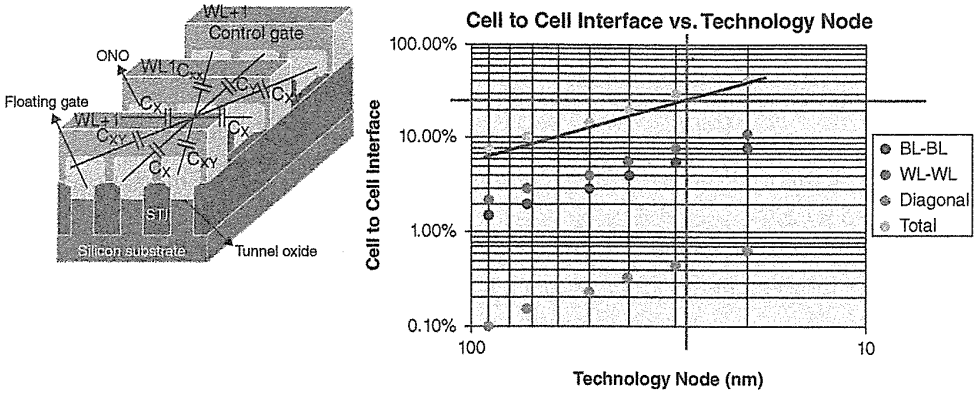


FIGURE 5.17 FG–FG coupling interference. Floating-gate interference as a percentage of the total V_t shift is shown as a function of the lithographic node. The BL–BL, WL–WL, and diagonal terms are per edge. FG–FG capacitive coupling interference exceeds 30% of FG capacitance beyond 30-nm generation. Need a solution beyond 30 nm for MLC 2 bits/cell, if assumed total interference <1 V in adjacent cell $dV_t = 4$ V → total <25%.

node. FG capacitive coupling interference exceeds 30% of FG capacitance beyond 30 nm generation. The interference approaches 50% of the total V_t shift of the cell at the 20-nm node. We need a solution to manage interference beyond 30 nm. It is estimated for MLC 2 bit/cell that the FG–FG capacitance has to be less than 20% of total FG capacitance, if assumed total interference is <1 V in an adjacent cell $\Delta V_t = 4$ V.

5.3.2 Direct Coupling with Channel

Based on the conventional theory of floating-gate capacitive coupling interference, the V_{th} shift in the y-direction (ΔV_Y) of a 63-nm technology node is more severe than that in the x-direction (ΔV_X), as shown in Fig. 5.16 [50], since the floating gates face each other directly in the y-direction, while it is shielded by a recessed control gate in the x-direction. However, it was observed that ΔV_X exceeds ΔV_Y at the node size of 50 nm, and it increases drastically as the technology node size reduces to 20 nm.

In the sub-50-nm technology nodes, the distance between the channel edge of a cell transistor and the floating gate of a neighboring cell is very close that the floating-gate voltage of the neighboring cell directly influences the channel edge, changing the electric field distribution on the channel edge. Then, V_{th} shift is caused by the direct field effect of the floating-gate potential of the neighboring cell. Since about 70% of the cell current flows on the channel edge, the V_{th} of the cell transistor is determined mostly on the condition of electric field crowding and the doping concentration of the channel edge [51]. Therefore, the memory cell suffers an intense V_{th} shift, particularly in the x-direction, where the floating gate faces the whole surface of the channel edge. This means that the observed floating-gate capacitive coupling interference is including the channel edge coupling with a neighbor cell.

The interference in the x-direction (ΔV_X) can be expressed as follows [50]:

$$\begin{aligned}\Delta V_X &= \Delta V_{X\text{—Indirect}} + \Delta V_{X\text{—Direct}} \\ &= 2*(C_{FGX}/C_{Tot})*\Delta V_{FGX} + \alpha*C_{FGX-STI}*\Delta V_{FGX}\end{aligned} \tag{5.4}$$

where ΔV_X is the total amount of the floating-gate capacitive coupling interference effect caused by ΔV_{FGX}, and ΔV_{FGX} is the V_{th} change of the adjacent cell transistor in the x-direction. ΔV_X is decomposed into two terms caused by indirect and direct field effects. The indirect field effect or parasitic capacitance-coupling effect produces a V_{th} shift ($\Delta V_{X\text{—Indirect}}$) with the ratio of C_{FGX}/C_{Tot}, where C_{FGX} is the FG–FG capacitance between two neighboring cell transistors in the x-direction and C_{Tot} is the total amount of capacitance of FG. This means that the indirect V_t shift of $\Delta V_{X\text{—Indirect}}$ is conventional floating-gate capacitive coupling interference. The direct field effect causes a V_{th} shift with the amount of $\alpha*C_{FGX-STI}*\Delta V_{FGX}$, where $C_{FGX-STI}$ is the capacitance between the floating gate of a neighboring cell transistor and the channel edge. α is constant, defining the influence of direct field effect, representing the doping profile and tunnel-oxide thickness on the channel edge. In the sub-100-nm technology nodes, $C_{FGX-STI}$ has been negligibly small due to a long distance between the floating gate of a neighboring cell and the channel edge. However, as the cell size reduces to below 50 nm, $C_{FGX-STI}$ increases in a large amount and builds up a large electric field on the channel edge. Therefore, combined with boron segregation on the channel edge, a large $C_{FGX-STI}$ causes an intense V_{th} shift on the channel edge, leading that $\Delta V_{X\text{—Direct}}$ exceeds $\Delta V_{X\text{—Indirect}}$ in the sub-50-nm technology nodes.

This effect was confirmed with 3D TCAD simulations [50]. The simulated cell had a 45-nm design rule with a gate pitch and an active pitch of 90 nm, as shown in Fig. 5.18. The potential distribution of the tunnel oxide and field oxide of a selected

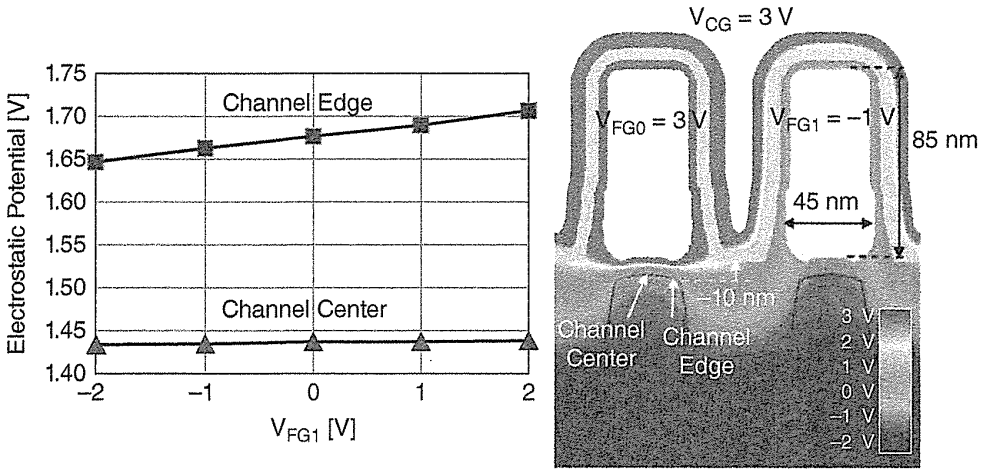


FIGURE 5.18 Simulation results depicting the potential distribution of the tunnel/field oxide of a selected cell transistor FG0 and its change with the floating-gate potential of a neighboring cell transistor FG1. The figure on the right side shows a representative electrostatic potential distribution in the case of $V_{FG1} = -1$ V.

cell transistor FG0 and its change with the floating-gate potential of a neighboring cell transistor FG1 are shown in Fig. 5.18 [50]. On the change of V_{FG1} from −2 to 2 V, it was observed that the potential on the channel edge increases from 1.65 to 1.71 V, while the potential on the channel center is kept constant on 1.43 V, showing the influence of the neighboring cell transistor potential on the channel edge. V_t is also simulated to be 0.62 V at $V_{FG1} = 2$ V, and it increases to 0.82 V in the case where $V_{FG1} = -2$ V. While the potential on the channel edge changed small from 1.65 to 1.71 V, the cell transistor V_{TH} is shifted largely from 0.62 to 0.82 V. This is because severe boron segregation occurs on the channel edge. Practically, the V_{th} shift will be twofold when considering the capacitance-coupling ratio as 0.5. This result indicates that the direct field effect of the adjacent cell transistor changes the cell V_{th} intrinsically, and it is larger than the effect of the FG–FG capacitive coupling.

Experimental data of cell-to-cell interference in a 45-nm cell are shown in Fig. 5.19 [50]. As the field oxide recess decreases, the direct field effect of the neighboring cell on the channel edge increases so that the V_{th} shift becomes larger. There are three lines in Fig. 5.19, namely, the V_{th} shift of conventional floating-gate capacitive coupling interference (ΔV_X—Indirect), V_{th} shift of direct field effect (ΔV_X—Direct), and V_{th} shift measured in experiments (ΔV_X). Both (ΔV_X—Indirect) and (ΔV_X—Direct) are calculated and classified by a 3D device simulator. While (ΔV_X—Indirect) varies a little with the field oxide recess and shows 0.28 V at −25 nm of field oxide recess, it is seen that (ΔV_X—Direct) increases drastically and reaches 0.67 V at −25 nm of field oxide recess. Moreover, the summation of two terms generates 0.95 V with an error of 0.08 V when compared with the experimental result, ΔV_X of 0.87 V. This experimental result demonstrates the strong influence of the direct field effect on floating-gate capacitive coupling interference in the x-direction. Therefore, in order

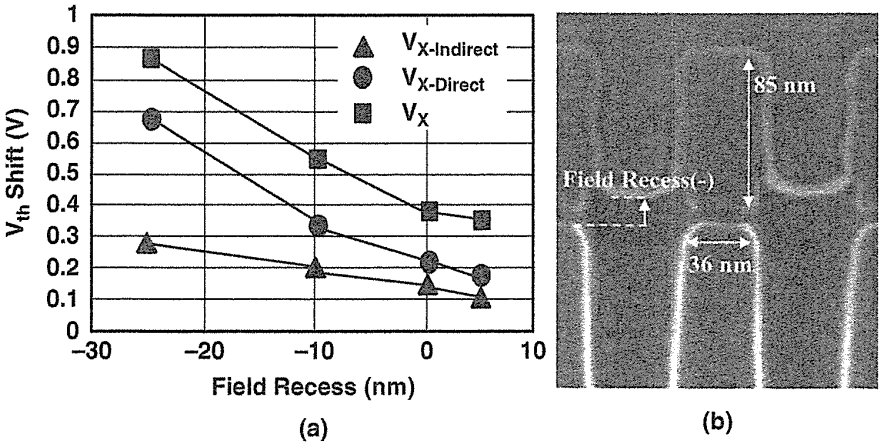


FIGURE 5.19 (a) V_{th} shift dependence on field recess. Four field recess conditions are prepared with field oxide recesses of −25, −10, 0, and +5 nm and the effect of cell-to-cell interference is measured by changing the neighboring cell transistor V_{th} from −5 to 5 V. (b) A representative cross-sectional SEM photograph of a NAND flash cell transistor with a 45-nm node size.

to reduce this effect in a NAND flash cell below a 50-nm cell, it is necessary to maximize the field oxide recess because keeping its balance to avoid the abnormal negative V_t shift effect in a large (deep) field recess [52], as described in Section 6.7.

5.3.3 Coupling with Source/Drain

A new cell-to-cell interference phenomenon of floating-gate induced barrier enhancement (FIBE) had been reported [53] in scaled cells below the 40-nm design rule. Unlike conventional capacitive coupling between floating gates, the threshold voltage (V_{th}) shift of the interfered cell becomes significantly large beyond some V_{th} of the interfering cell. This is due to modulation of the conduction band at the source and drain regions by capacitive coupling between source/drain and the floating gate of the interfering cell. The model was confirmed by experiment and simulation. In order to reduce the FIBE effect, the higher doping for S/D junction and the higher V_{read} scheme in neighbor WL could be effective.

Figure 5.20 shows the cell-to-cell interference between WLs (y-direction) [53]. The V_{th} of an interfered cell increases abnormally in region B (higher Interfering cell V_{th}), while the V_{th} of an interfered cell has a linear dependency on the lower V_{th} of the interfering cell at region A, which shows the conventional floating-gate capacitive coupling. This abnormal V_t increase at region B is observed only in higher V_t of an interfering cell for scaled dimension of NAND flash memory cell beyond 40 nm.

A new model of floating-gate induced barrier enhancement (FIBE) was proposed [53] for this phenomenon. The programming of an interfering cell causes the higher potential of the channel conduction band of an interfered cell at the drain-side region due to direct capacitive coupling between the source/drain region and the floating gate of an interfering cell, resulting in the increase of the cell V_{th}, as shown in Fig. 5.21a.

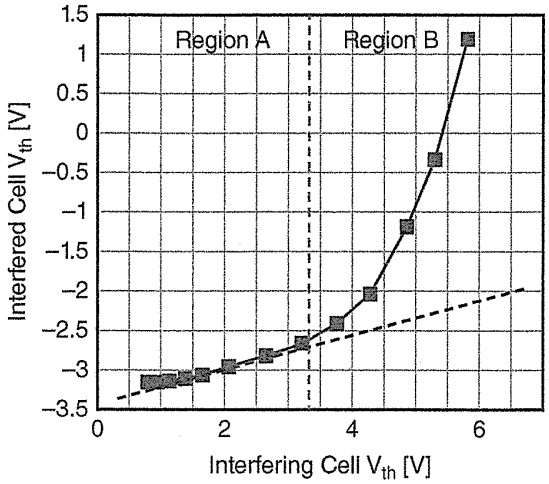


FIGURE 5.20 The V_{th} shift of interfered cell dependence on the interfering cell V_{th}. The initial V_{th} of an interfered cell is set to −3 V and interfering cell is 0.6 V. Linearity of region A can be explained by the conventional parasitic capacitors coupling.

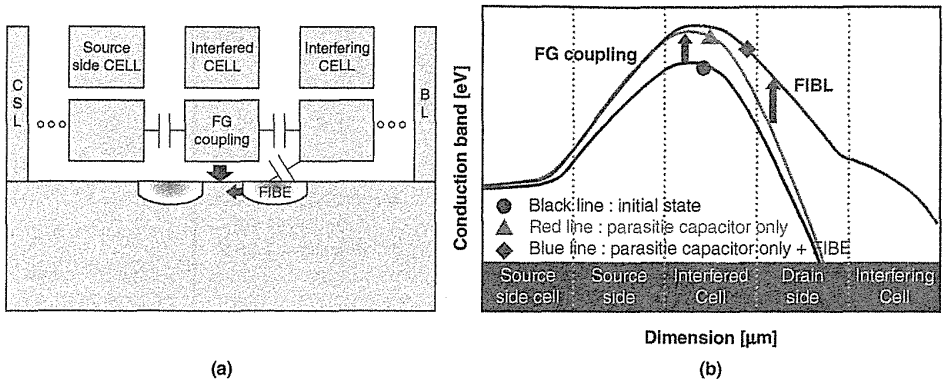


FIGURE 5.21 (a) Simplified image of parasitic coupling capacitances in the WL direction of a NAND flash cell. The source/drain junction is coupled with a floating gate of an interfering cell. (b) Conduction band profile beneath tunnel oxide. The word-line interference (*y*-direction interference) consists of two factors of the conventional floating-gate capacitive coupling and the FIBE (floating-gate induced barrier enhancement).

Figure 5.21b shows the simulated conduction band contour of the interfered cell in the case of a pre-programming interfering cell, with parasitic capacitance and with parasitic capacitance + FIBE effect. The FG potential change of the interfered cell with conventional floating-gate capacitive coupling appears to increase only the conduction band at the channel center as shown in Fig. 5.21b. However, the FG potential change of the interfering cell enhances the conduction band of the drain region and affect the channel conduction band of an interfered cell as shown in Fig. 5.21b.

The FIBE appears at a relatively high V_{th} region of the interfering cell where the drain region conduction band was enhanced sufficiently. Figure 5.22 shows the measured I_d–V_g curve influenced by the floating-gate coupling and FIBE in 27-nm node NAND flash cell [53]. If only the conventional floating-gate capacitance coupling interference is considered, we can see just mid-gap voltage shift in Line 2 from the original Line 1 without the slope change of the I_d–V_g curve in Fig. 5.22. However, in Line 3, the saturation region of I_d–V_g curve is distorted, and this distortion is extended even to the linear region of I_d–V_g curve when the interfering cell is programed highly to 5.5 V. This phenomenon makes V_t abnormally higher, as shown in Fig. 5.20.

5.3.4 Air Gap and Low-*k* Material

It had been introduced that the floating-gate capacitive coupling interference improved by using a low-*k* dielectric of gate spacer, such as low-*k* oxide and air gap [47,48,54, 55].

One example of a process flow to form an air gap between gates [47] is shown in Fig. 5.23A. After gate patterning, buffer oxide/nitride is deposited (Fig. 5.23A, part b), and then oxide is deposited to fill gate space (Fig. 5.23A, part c). After that,

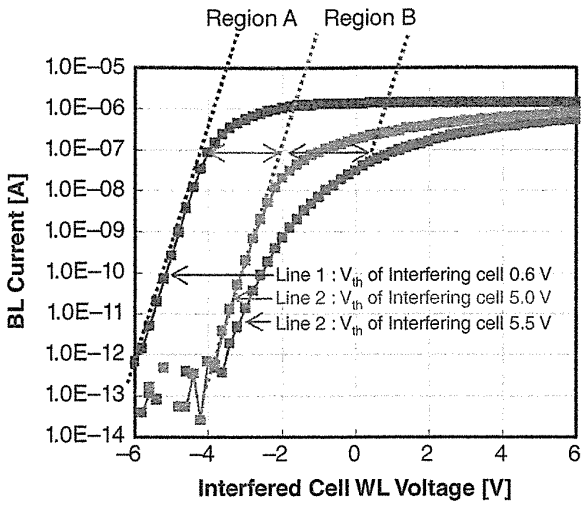


FIGURE 5.22 BL current versus WL voltage of an interfered cell. The V_{th} of an initial interfered cell is set to be −4 V, and the V_{th} of an interfering cell is set to 0.6 V. After the interfering cell is programmed up to 5.0 V, the BL current is parallel shifted up (Line 2). And after the interfering cell is programmed up to 5.5 V (V_{read}: 7 V), BL current is distorted (Line 3) by the FIBE effect.

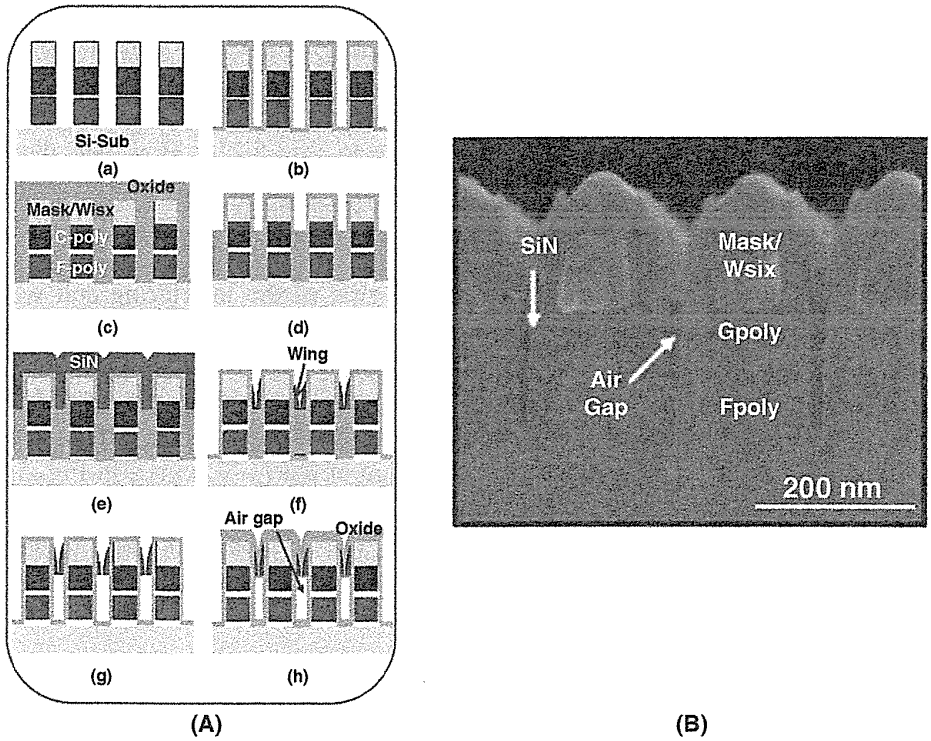


FIGURE 5.23 (A) The process flow of the air gap: (a) The gate patterning and the barrier silicon dioxide deposition (150 Å), (b) the barrier SiN deposition (200 Å), (c) thick-oxide deposition (1000 Å), (d) thick oxide is removed by dry etch, (e) SiN deposition (150 Å), (f) The wing is formed, (g) The thick oxide inside gate to gate space is removed by wet etch. (h) The air gap is formed. (B) The SEM image of air gap in a 90-nm cell (gate length = gate space = floating gate height = channel width = 90 nm, tunnel oxide thickness = 6.5 nm, ONO thickness = 16 nm).

the oxide over the gate poly is removed by dry etching (Fig. 5.23A, part d). Then, to form the gate wing inside gate-to-gate space, the SiN is deposited and etched (Fig. 5.23A, part f). After the oxide inside the spacer wings are removed by wet etching (Fig. 5.23A, part g), air gaps inside gate-to-gate space are formed by oxide deposition (Fig. 5.23A, part h). In Fig. 5.23B, the air gap can be clearly observed from SEM image of the fabricated device.

Figure 5.24a–c shows the cell V_{th} distribution of WL30 and WL31 in 90-nm cells with gate-space materials of SiN, oxide, and air gap, respectively [47]. The cell V_{th} on WL30/even bit line are shifted by programming the adjacent cells in the same word line (WL30)/odd bit line and WL31/even bit-line cell. As a cell is programmed from $V_{th} = -3$ V to $V_{th} = 1.5$ V, the threshold voltage changes for gate space material of SiN, oxide, and air gap are 0.16 V, 0.07 V, and 0.02 V, respectively. The V_{th} shift nearly corresponds to those of the dielectric constant (SiN:oxide:air = 8:4:1). The reduced V_{th} shift with air gap is due to its lower parasitic capacitance between floating gates. Figure 5.24d compares the cell V_{th} distribution of a 1Gbit cell by a single pulse program. The cell V_{th} distribution is improved with air gap due to the improved floating-gate capacitive coupling interference.

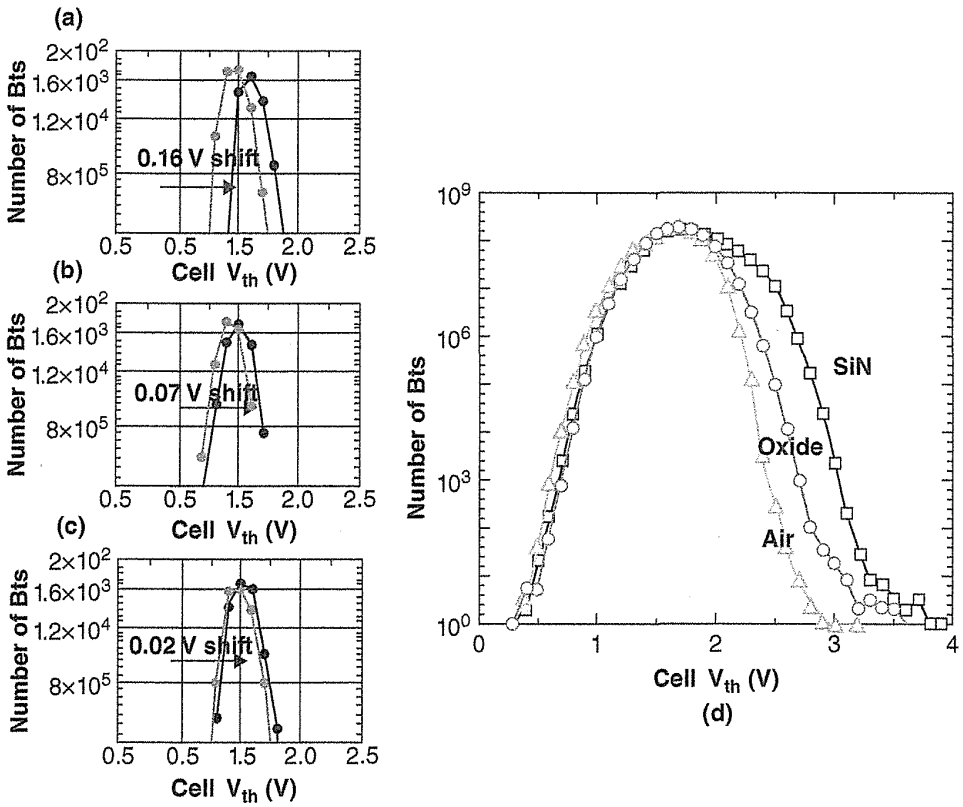


FIGURE 5.24 Threshold voltage shift by floating-gate interference was measured on WL30. (a) SiN spacer, (b) oxide, (c) air gap, and (d) the cell V_{th} distribution shifted by floating-gate interference.

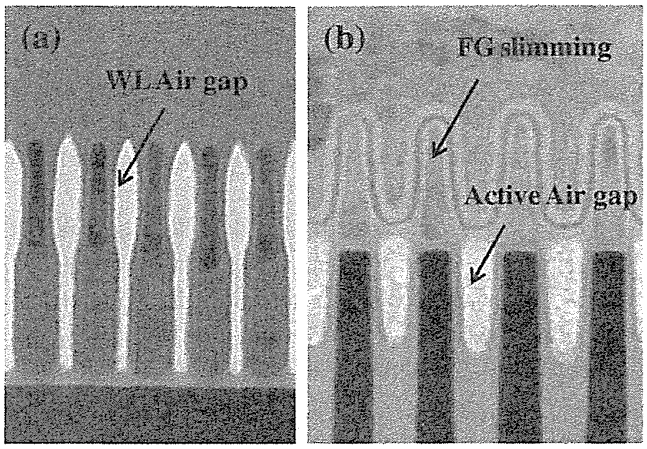


FIGURE 5.25 The cross-sectional TEM micrograph of (a) word-line air gap and (b) STI air gap in a mid-1X-nm SA-STI NAND flash cell. Air gap is key technology to improve a floating-gate capacitive coupling interference, a WL high-field problem, and a program disturb. WL air gap and STI air gap are applied to the product from 25-nm generation and 20-nm STI half-pitch, respectively.

Figure 5.25 shows the cross-sectional TEM micrograph of word-line (WL) air-gap and STI air-gap structure in the middle 1X-nm cell [49]. The WL and STI air gap were successfully fabricated. The WL air gap was started to be used from 25-nm generation product to reduce floating-gate capacitive coupling interference [34]. And WL air gap could also improve WL high-field problem [10, 11], as described in Section 5.7.

The STI air gap was started to be used from 20-nm bit-line half-pitch in the middle 1X-nm cell, as shown in Fig. 5.25 [49] and Fig. 5.26 [49]. STI air gap is very effective to improve not only floating-gate capacitive coupling interference but also

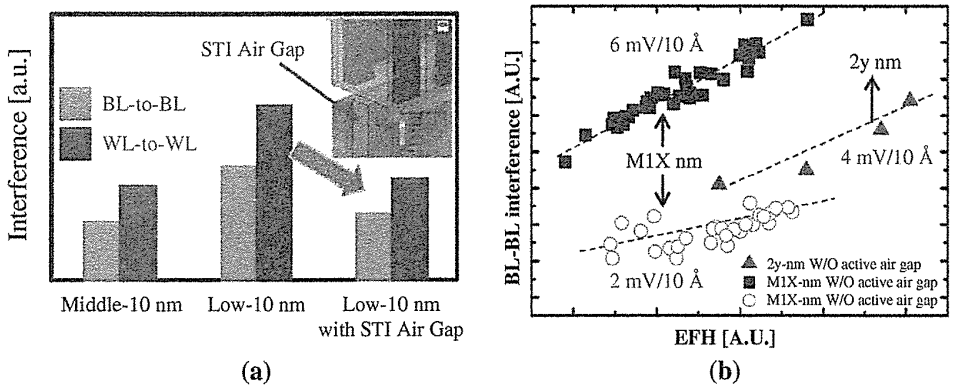


FIGURE 5.26 (a) A simulation result shows that the BL–BL interference value in a low-10-nm cell could be similar to that of a middle-10-nm cell (middle 1X-nm cell) by using STI air gap. (b) BL–BL interference of middle 1X-nm and 2y-nm design rule. EFH is "effective field height" of STI buried oxide (see FH in Fig. 6.72).

program disturb, which is related on channel–channel coupling [31,49,56] described in Section 6.5.3.

These air-gap technologies are the key to implement small cell size below 20-nm design rule, because of improvement of floating-gate capacitive coupling interference [35], WL high-field problem [10], and program disturb [49].

5.4 PROGRAM ELECTRON INJECTION SPREAD

5.4.1 Theory of Program Electron Injection Spread

By scaling memory cell size, the number of stored electrons in floating gate is reduced, as shown in Fig. 5.27 [19, 57]. In 1X-nm memory cell, number of stored electron reaches close to 100, which is corresponding to 3-V V_t shift. It means that only 10 electrons are injected to floating gate in one programming pulse which has 300-mV step-up between each pulse. A small number of 10 electrons should make a large statistical variation on number of injected electrons to floating gate, resulting in wider V_t distribution width of program states.

It had been reported that the programmed V_t distribution width became wider by statistical electron injection spread during program pulse [13–15]. The electron injection process is ruled by the Poisson statistics when small number of electrons is injected during program pulse. This can be explained by the reduction of the tunnel-oxide field that follows the electrons injection to floating gate and then reduced the electron injection rate. The results are explained by means of a Monte Carlo model, which is able to correctly describe the main physics behind the program operation.

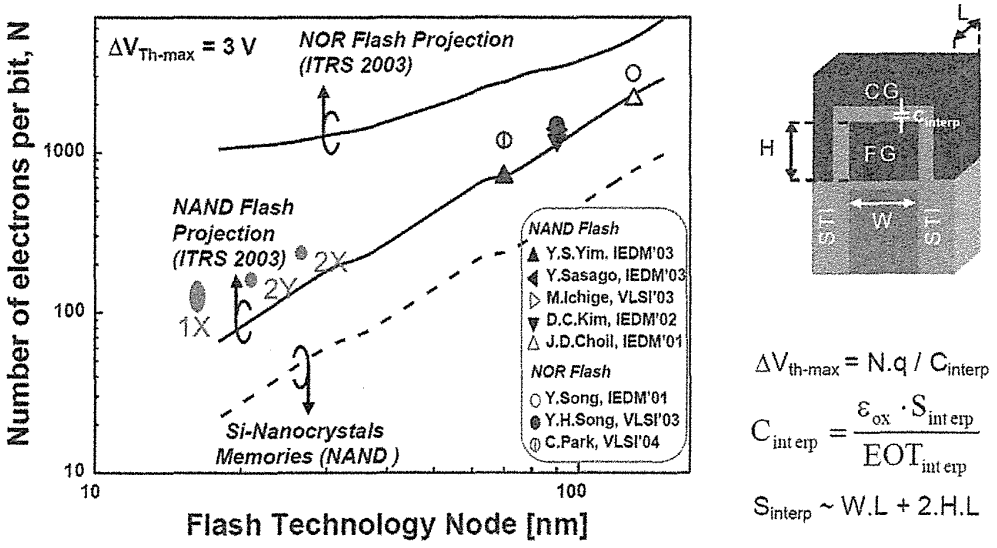


FIGURE 5.27 Number of stored electrons in FG, as a function of the flash memory technology node according to the ITRS 2003 edition. The number of electrons is decreased as scaling memory cell size.

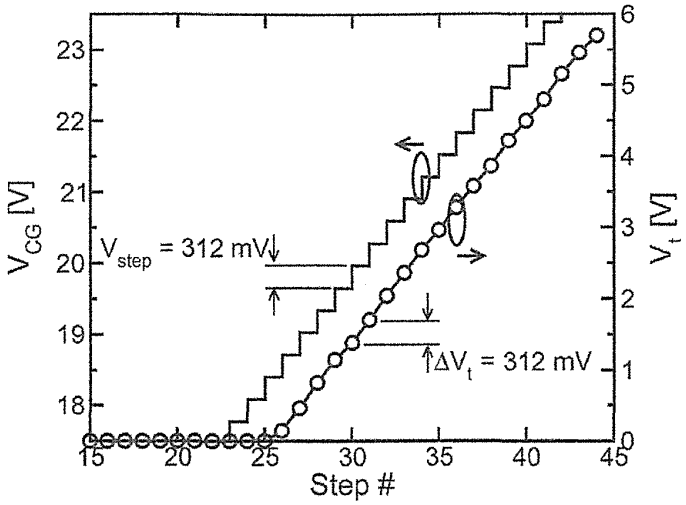


FIGURE 5.28 Example for a control-gate voltage waveform used to program a NAND cell and resulting V_t transient on a 60-nm device. Note that only positive V_t values can be sensed in a conventional NAND cell array.

The distribution is shown to broaden as a consequence of the injection statistical spread, leading some cells to displace from the verify level more than V_{step}. The injection statistical spread is considered to be larger as scaling the NAND memory cell due to reduced number of electrons in one program pulse.

An experiment of the ΔV_t spread had been studied by using the ramped programming (ISPP: incremental step pulse programming [36,37,58]), as shown in Fig. 5.28 [13,14]. ΔV_t was defined as the V_t shift obtained after n_s programming pulse steps. For example, Fig. 5.29a shows $\Delta V_t = V_{t,29+ns} - V_{t,29}$ transients, obtained using V_t at

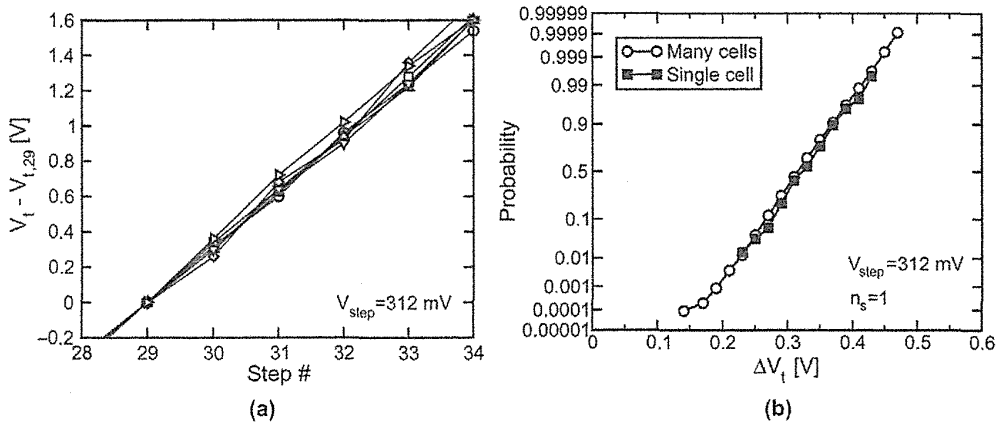


FIGURE 5.29 (a) Example of ΔV_t transients (assuming V_t at step 29 of the staircase as reference) measured on the same 60-nm technology NAND cell. (b) ΔV_t evaluated using many programming ramps on the same cell or a single programming ramp on a large number of cells, for $V_{step} = 312$ mV and $n_s = 1$.

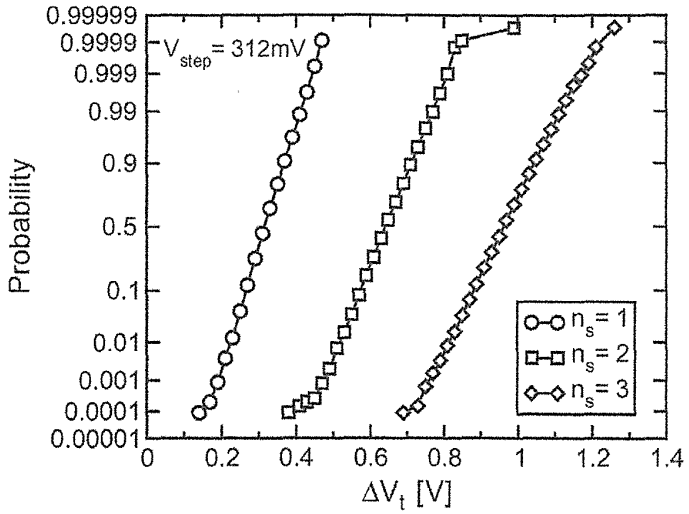


FIGURE 5.30 Experimental ΔV_t distributions for $V_{step} = 312$ mV and increasing n_s, showing the increase of $\sigma_{\Delta V_t}$ that follows the larger average ΔV_t value.

the 29th step of the control-gate ramp as a reference, on the same NAND memory cells in 60-nm technology. The statistical character of ΔV_t is clearly appeared. And the cumulative distribution of ΔV_t is shown in Fig. 5.29b for $n_s = 1$ (i.e., $V_{t,30} - V_{t,29}$), using nearly 100 V_t transients on the same cell. These results of single cell are compared with the ΔV_t distribution obtained from a V_t programming transient on a page (16 kb) of the NAND memory cell array. It can be seen that there is very good agreement between the two distributions, confirming that we are observing the same statistical distribution on single- and many-cell results. Also, it can be seen that the distributions clearly show a Gaussian behavior, with a spread nearly equal to the standard deviation $\sigma_{\Delta V_t} = 41$ mV.

Figure 5.30 [14] shows the ΔV_t distribution from many cell statistics in the case of $V_{step} = 312$ mV. The distribution spread of ΔV_t is clearly increased with increasing staircase pulse, along with increasing the average of ΔV_t. The injection spread is strictly related to the ΔV_t statistics, based on the following relation:

$$\sigma_{\Delta V_t} = \frac{q}{C_{pp}}\sqrt{\sigma_n^2}$$

where q is the electron charge, n is the number of injected electrons, and C_{pp} is inter-poly capacitance. By assuming that n is ruled by Poisson statistics, its variance σ_n^2 is equal to its average value n, and the previous equation becomes

$$\sigma_{\Delta V_t} = \frac{q}{C_{pp}}\sqrt{\bar{n}} = \sqrt{\frac{q}{C_{pp}}\overline{\Delta V_t}} \tag{5.5}$$

A square-root dependence of $\sigma_{\Delta V_t}$ on ΔV_t is expected from (5.5). However, the further consideration derives from the hypothesis of Poissonian injection. In fact,

when an electron is injected to the floating gate, the floating-gate potential energy rises. This reduces the tunnel-oxide field and the electron injection rate, thus causing a sub-Poissonian electron injection process.

In order to involve the effect of the tunnel-oxide field feedback, Monte Carlo simulations of the electron injection process had been performed. For each floating-gate potential, the average electron injection rate can be calculated as JA/q, where A is the cell area and J is the tunneling current density through the tunnel oxide. To obtain a reasonable accuracy, the tunneling current–floating-gate voltage (J–V_{FG}) characteristics were experimentally extracted from constant control-gate voltage programming transients [13]. The average injection rate was used to extract the time of the next electron injection event from the substrate. Simulations are performed with considering the control-gate steps (increasing the floating-gate potential and the average injection rate) and the electron injection events (decreasing the tunnel-oxide field and the average injection rate), causing a nonstationary Poisson process. The $\sigma_{\Delta V_t}$ is then extracted from the simulation of many V_t transients.

Figure 5.31 shows the experimental and calculated $\sigma_{\Delta V_t}$ as a function of ΔV_t, extracted in the stationary part of the programming transient for different V_{step}, n_s, and step durations. As the ΔV_t spread depends only on the electron injection process, $\sigma_{\Delta V_t}$ depends only on ΔV_t and does not depend on the duration of the steps of the control-gate staircase or the number of steps used to reach the ΔV_t value [13, 14]. For low ΔV_t, both experiments and Monte Carlo calculations well match the values predicted assuming a Poisson statistics for the electron injection process. This is because the feedback to the tunnel-oxide field is negligibly small when the number of injected electrons is small. However, when ΔV_t is increased, a saturating behavior clearly appears both from experiments and from Monte Carlo simulation.

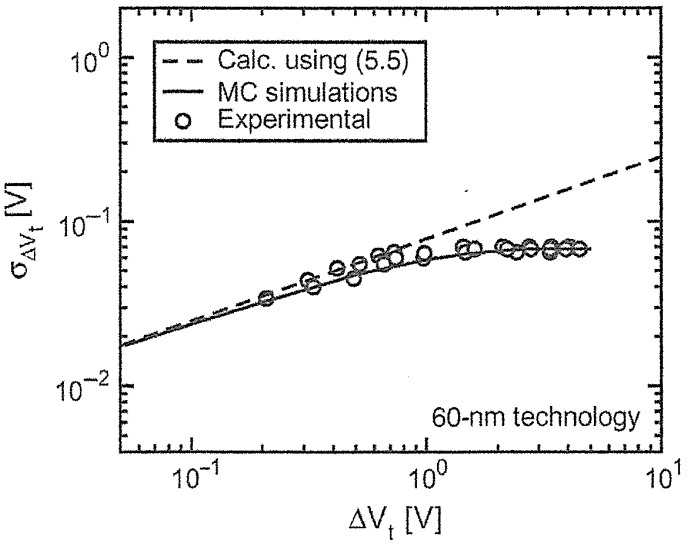


FIGURE 5.31 Experimental and calculated $\sigma_{\Delta V_t}$ as a function of the average ΔV_t in 60-nm NAND flash cell technology.

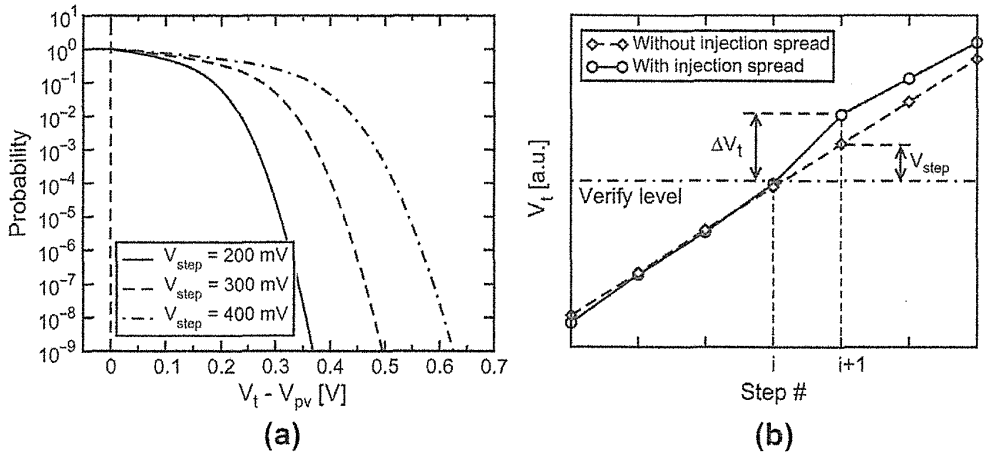


FIGURE 5.32 (a) Calculated 1-cumulative probability for $V_t - V_{pv}$ assuming constant-current NAND programming with different V_{step} values in the 60-nm NAND flash technology. (b) V_t evolution for increasing step numbers with and without considering the injection spread. In the presence of a verify level, the worst situation is obtained when, at step i, cell V_t is slightly lower than V_{pv}, thus requiring an additional program step. This shifts cell V_t to $V_{pv} + V_{step}$ when the injection spread is not considered, but to larger values when the injection spread is included.

This reveals the sub-Poissonian nature of the electron injection process, determined by the tunnel-oxide field feedback following each electron injection event. The starting point separation between the $\sigma_{\Delta V_t}$ curve from the Poissonian spread and its saturation are dependent on the shape of the J–V_{FG} characteristics, whose slope around the programming condition determines the field variation.

Figure 5.32a shows the effect of the injection spread on the V_t distribution by using bit-by-bit program verify operation [43] with program-verify voltage of V_{pv}. Results had been calculated in the case of three staircase steps, 200, 300, and 400 mV on a 60-nm NAND flash cell. If the injection spread is neglected, the ISPP programming algorithm should make all V_t's to fit between V_{pv} and $V_{pv} + V_{step}$ in principle [59] (see Section 2.2.3). Cells which have slightly lower than V_{pv}, are required to be subjected an additional program pulse to be higher than the verify level. This additional program pulse shifts V_t by V_{step}, projecting the cell to $V_{pv} + V_{step}$, as shown in Fig. 5.32b. However, when the injection spread is included, V_t values larger than $V_{pv} + V_{step}$ are caused. An example of this situation (shown in Fig. 5.32b) is due to the possibility to have single-step ΔV_t values larger than V_{step}, thus causing a cell to move further away from the verify level.

As a scaling of the NAND cell, inter-poly capacitance C_{pp} is decreased (namely number of stored electrons are decreased), then $\sigma_{\Delta V_t}$ is increased. This trend is clearly shown in Fig. 5.33a, where $\sigma_{\Delta V_t}$ is shown for different NAND cell technologies of 90-nm, 70-nm and 60-nm nodes. As the tunnel-oxide conduction characteristics are kept the same with technology scaling, the same saturating behavior is observed for all the curves. The spread corresponding to ΔV_t equal to 200, 300, and 400 mV is

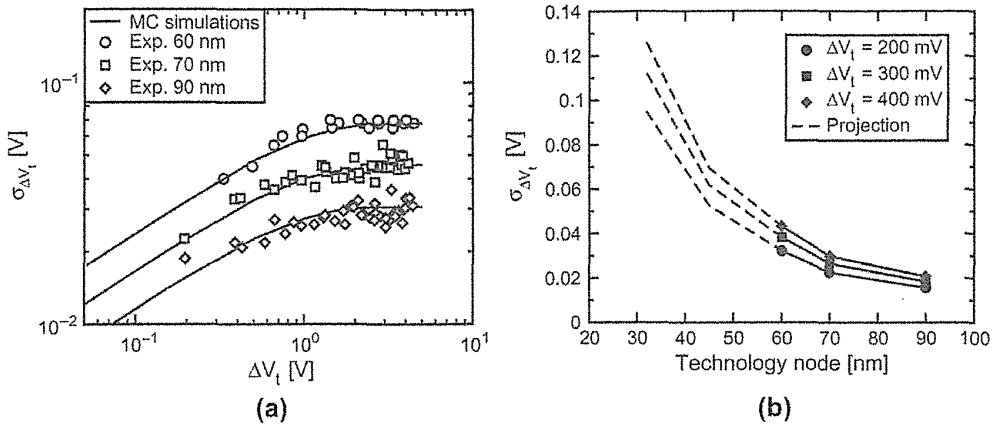


FIGURE 5.33 (a) Experimental and calculated $\sigma_{\Delta V_t}$ as a function of the average ΔV_t value for different technology nodes from 90 to 60 nm. (b) $\sigma_{\Delta V_T}$ for different ΔV_t values in the Poissonian region of the electron injection process as a function of the NAND technology nodes. Continuous curves and symbols refer to the available technologies; dashed curves are calculated projections assuming that C_{pp} scales with cell area. As number of stored electron in FG is decreased (C_{pp} decrease) as memory cell scaling, program V_t width is larger (worse) due to program variation.

shown in Fig. 5.33b for the available technologies, drawing also possible scaling projections. The $\sigma_{\Delta V_t}$ is drastically increased with scaling technology node. This result shows that electron injection spread would present a serious problem to make a tight V_t distribution width in multilevel cell for the scaled NAND flash memory. In order to keep the V_t distribution width as close as possible to the verify level without using very small V_{step} amplitudes, the scaling of C_{pp} should be carefully considered.

5.4.2 Effect of Lower Doping in FG

The electron injection spread is enhanced by lower doping concentration in floating gate (FG) [38]. Larger ΔV_t distribution has been observed in a cell which has a low floating-gate doping concentration in a 40-nm design rule cell, as shown in Fig. 5.34a, where ΔV_t means the V_t shift from j to $j + 1$ step-up programming pulses [$\Delta V_t \equiv V_t(j + 1) - V_t(j)$]. The ΔV_t distribution of the low FG doping shows a wider spread than that of the high doping, and tail bits are observed at the higher ΔV_t in the case of the low FG doping.

The reason why ΔV_t distribution in low doping is larger can be explained by the dynamics of forming an inversion layer in FG and electron–hole generation by FN (Fowler–Nordheim) tunneling electron injection, as follows. The details of the time dependence of the FG potential are considered in each programming step. At the beginning of the Nth programming pulse duration, the tunnel-oxide interface in the FG is deeply depleted due to the large electric field in the tunnel oxide. And a large band bending is caused at the tunnel-oxide interface in the FG, as schematically shown in Fig. 5.35a. The tunneling electrons become energetic at FG, and they

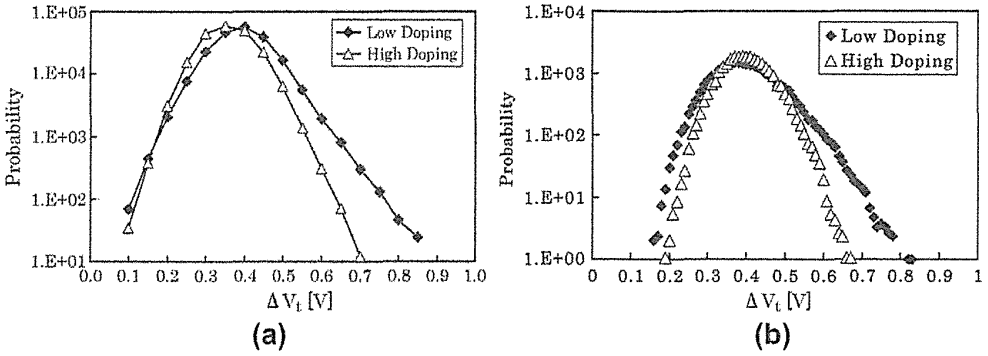


FIGURE 5.34 (a) Measurement results of bit-by-bit V_t transient (ΔV_t) distribution from i to $i + 1$ staircase programming pulses. ΔV_t distribution sampling points are accumulated by adding the data of $i = 12$–17. V_{step} is 400 mV. The NAND cell array with low FG phosphorus doping shows wider ΔV_t distribution than that with high doping. (b) Calculated ΔV_t distribution considering the effect of FN tunneling statistics in both cases with low and high FG phosphorus doping.

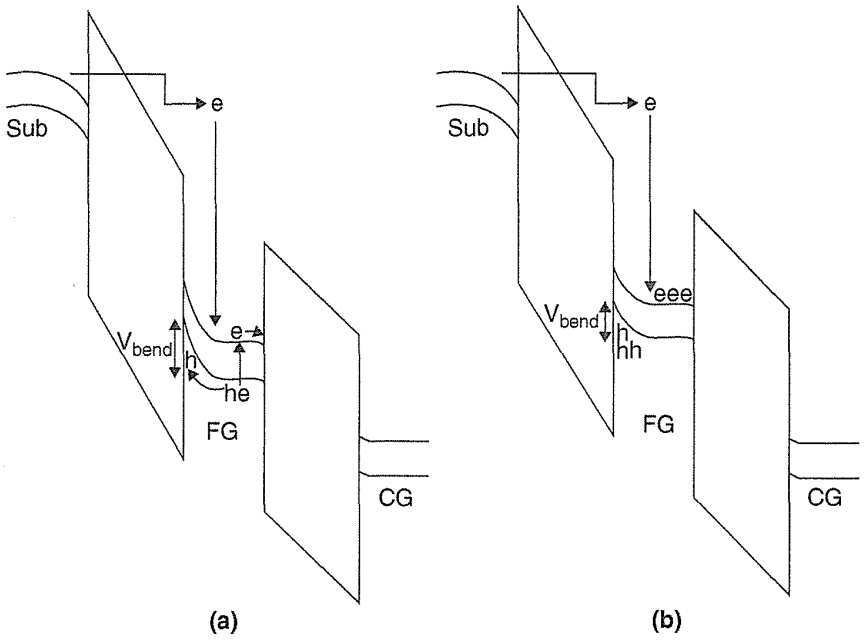


FIGURE 5.35 Schematic 1-D band diagram for the NAND cell during programming. (a) Band diagram at the initial stage of programming and the process of FN tunneling currents that generate e–h pairs. Generated holes move to the tunnel-oxide interface, and generated electrons move to the interface of ONO. (b) Band diagram after the passage of programming time. The inversion layer of holes has formed and degenerated the deep-depletion region (band bending height becomes smaller, as compared with that of initial case).

generate the electron–hole pairs (e–h pairs). These generated holes are gathered at the tunnel-oxide interface. Therefore, as the programming time (programming pulse width) has become longer, the gathered holes create the inversion layer and reduce the depletion width in the FG, which results in the reduction of the band bending voltage V_{bend}. Generally, in the case of longer programming time, the injected charges to the FG reduce the tunnel-oxide electric field. However, the V_{bend} reduction makes the tunnel-oxide field reduction slower. This V_{bend} reduction is schematically shown in Fig. 5.35b. Thus, the electric-field enhancement on the tunnel oxide increases the FN tunneling current at the latter period of the Nth programming pulse duration. Before applying the next $(N + 1)$th programming pulse, a verify read sequence is inserted in actual operation of NAND flash programming operation. Then, the generated holes during programming pulse diffuse into the entire FG area and will almost recombine with electrons during the relatively long verify read period. Therefore, the deep depletion in the FG repeatedly occurs at the beginning of the next $(N + 1)$th programming pulse duration. This electric-field enhancement effect through the tunnel oxide is exaggerated to appear in the case of lower phosphorus doping in the FG because of the larger V_{bend} at the beginning of each programming pulse. The wider ΔV_t distribution in the lower doping FG can be analyzed by this effect.

This phenomenon was simulated based on this model of combining the effect of the band bending reduction due to the holes stored in the FG and the FN tunneling statistics [13,59]. Monte Carlo simulation was carried out where each program pulse was divided into many small segments, and the calculation was carried out by each segment. The simulated ΔV_t distributions with high and low FG phosphorus doping are shown in Fig. 5.34b. It is clear that the ΔV_t distribution with the low FG doping shows a wider spread in comparison with the case of the high doping, due to the tunnel-oxide electric-field enhancement effect.

Figure 5.36 shows the phosphorus doping dependence of $\sigma(\Delta V_t)$, where V_{step} is fixed at 400 mV [38]. The ΔV_t distribution significantly spreads in lower phosphorus doping in the FG. This ΔV_t distribution widening mainly comes from the existence

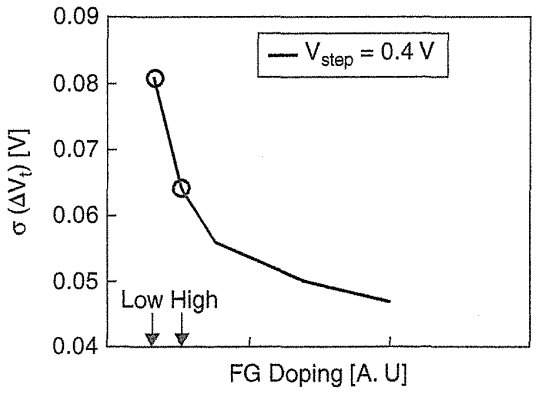


FIGURE 5.36 Calculated $\sigma(\Delta V_t)$ as a function of FG phosphorus doping. V_{step} is fixed to 400 mV.

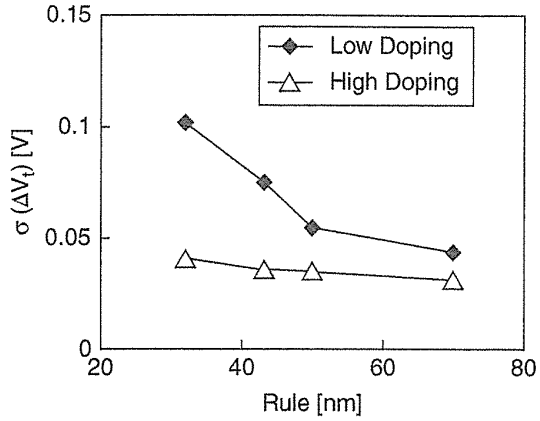


FIGURE 5.37 Calculated $\sigma(\Delta V_t)$ as a function of technology nodes. $\sigma(\Delta V_t)$ values with low and high FG phosphorus doping are compared, where V_{step} is fixed at 400 mV.

of the tail bits at the high ΔV_t side, as shown in Fig. 5.34b. Conversely, the amount of tail bits at the high ΔV_t side can be reduced as the FG doping increases. It means that, when the FG doping is higher, the σ plot of the ΔV_t distribution shows clearer linearity. This result provides a new guideline for the design of NAND cell process.

Generally, the $\sigma(\Delta V_t)$ is increased as scaling down of the NAND cell, because of the reduction of C_{pp}. Also, the tunnel-oxide electric-field enhancement effect at the lower FG impurity doping accelerates the increase in $\sigma(\Delta V_t)$ as the cell size scaling, as shown in Fig. 5.37, where the effect of FN tunneling statistics is considered. The increase in $\sigma(\Delta V_t)$ introduces a new reliability constraint to the design of read window margin (RWM) of the future NAND technologies. The upper limit of the impurity doping would come from the tunnel-oxide reliability degradation. Conversely, the lower limit would come from this ΔV_t distribution spread.

5.5 RANDOM TELEGRAPH SIGNAL NOISE (RTN)

5.5.1 RTN in Flash Memory Cells

Random telegraph noise (RTN) in a MOSFET is the drain current or threshold voltage fluctuation caused by electron capture and emission events at a charge trap site near the gate-oxide interface, as shown in Fig. 5.38 [69]. The amplitude of the threshold voltage fluctuation by each trap site ($\Delta V_{t_{trap}}$) in a flash memory cell is approximately [60–62]

$$\Delta V_{t_{trap}} = \frac{q}{L_{eff} W_{eff} \gamma C_{ox}} \tag{5.6}$$

where q is the elementary charge, L_{eff} and W_{eff} are the effective channel length and width respectively, γ is coupling ratio between the control and floating gates, and C_{ox}

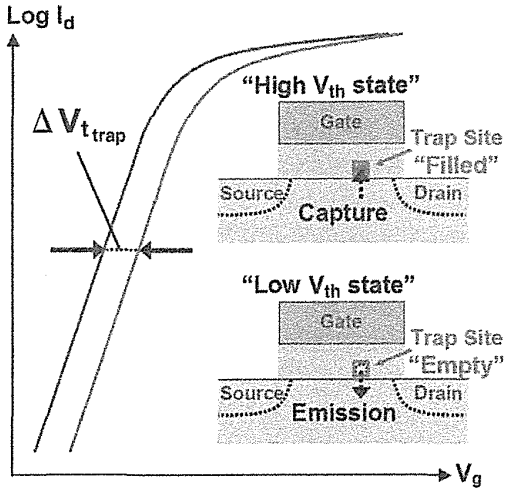


FIGURE 5.38 Threshold voltage fluctuation due to random telegraph noise (RTN) in a MOSFET is caused by electron capture and emission events at an oxide electron trap site.

is the gate capacitance. The amplitude of RTN is generally larger in a floating-gate flash memory cell than in a CMOS logic device, because of the very small C_{ox}, due to the relatively thick tunnel oxide (~10 nm thick). Also, in NAND flash memory, memory cell size has been intensively scaled down; thus dimensions of L and W, especially W, are much smaller than conventional CMOS logic device. Moreover, the amplitude of RTN can be larger than expected from Eq. (5.6) due to current-path percolation mechanism [63]. Therefore, RTN is a potential source of read failure in scaled NAND flash memory.

It had been reported for the first time that the threshold voltage (V_{th}) fluctuation due to random telegraph signal noise (RTN) had been observed in flash memory [16]. Figure 5.39 shows an example of RTN measured in 90-nm cell. Drain current shows a switching behavior as same as RTN in logic CMOS transistor.

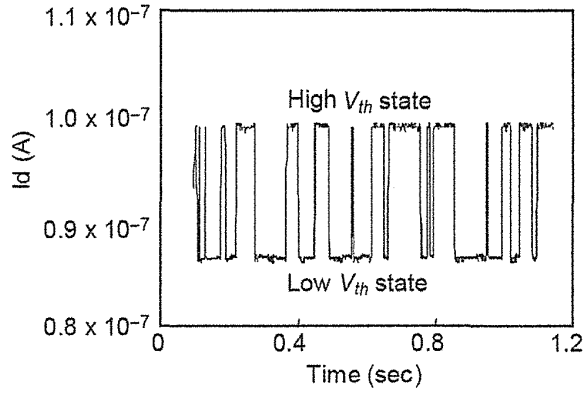


FIGURE 5.39 An example of time-series change in drain current in 90-nm-node flash memory.

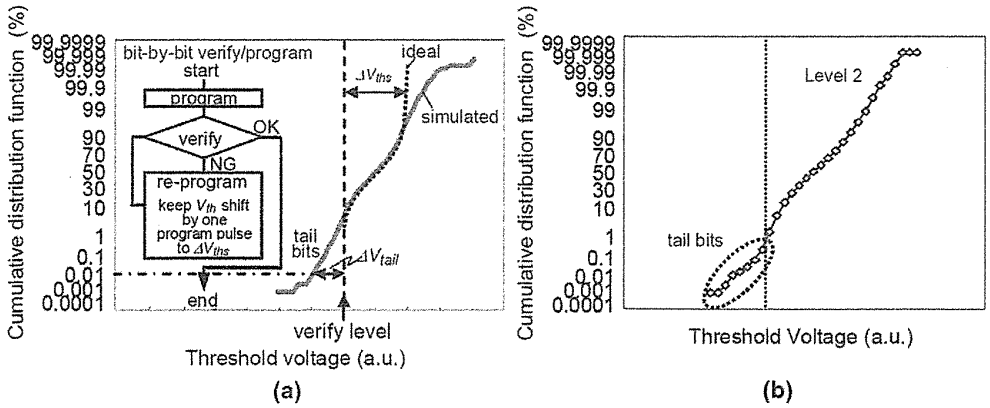


FIGURE 5.40 (a) Monte Carlo simulation results of programmed V_{th} distribution using bit-by-bit verify scheme. (b) Measured V_{th} distribution of Level 2 enlarged.

The influence of RTN for the multilevel flash memory was investigated by Monte Carlo simulation [16]. In multilevel flash memory, it is necessary to control the V_{th} precisely and make a tight V_{th} distribution. For such requirements, a bit-by-bit program/verify technique has been used [43] (Section 4.2.3). In this technique, the programming bias is applied only to the memory cells judged to "fail" in the verify operation. In addition, to keep V_{th} shifts by one programming pulse constant, the ISPP (incremental step pulse program) scheme (Section 4.2.2) is used. Figure 5.40a shows the simulated results of V_{th} distribution that is programmed by using bit-by-bit program/verify technique and ISPP scheme. In the simulated distribution with the RTN model, the tail bits appear in upper and lower of the V_{th} distribution in contrast with the ideal distribution without the RTN model. Figure 5.40b shows the measured V_{th} distribution. By comparison between Fig. 5.40a and 5.40b, we confirmed the existence of the tail bits generated by RTN in flash memory for the first time [16].

The properties of traps in the SiO_2 was investigated by means of a statistical analysis of random telegraph signal noise in flash memory arrays [65, 66]. A new physical model for the statistical superposition of the elementary Markov processes describing traps occupancy was developed. The comparison of modeling results with measured data is able to estimate the energy and space distribution of oxide defects, which are related to cell threshold voltage instability.

The random telegraph signal process [67] is schematically described in Fig. 5.41a [65]. An oxide trap has a distance x_t from the substrate/SiO_2 interface and energy E_t from the SiO_2 conduction band. An oxide trap can capture and emit single electrons from/to the substrate with average time constants τ_c and τ_e, giving rise to the typical behavior for the drain current shown in Fig. 5.41b and affecting V_t [67]. The properties of the trap responsible for the RTN can be experimentally extracted from the V_G dependences of τ_c and τ_e. Flash memory array was used to collect data, which could effectively evaluate a large number of devices. The RTN statistical distribution is then directly extracted from the V_t distribution, without any need to thoroughly characterize τ_c and τ_e for any single trap.

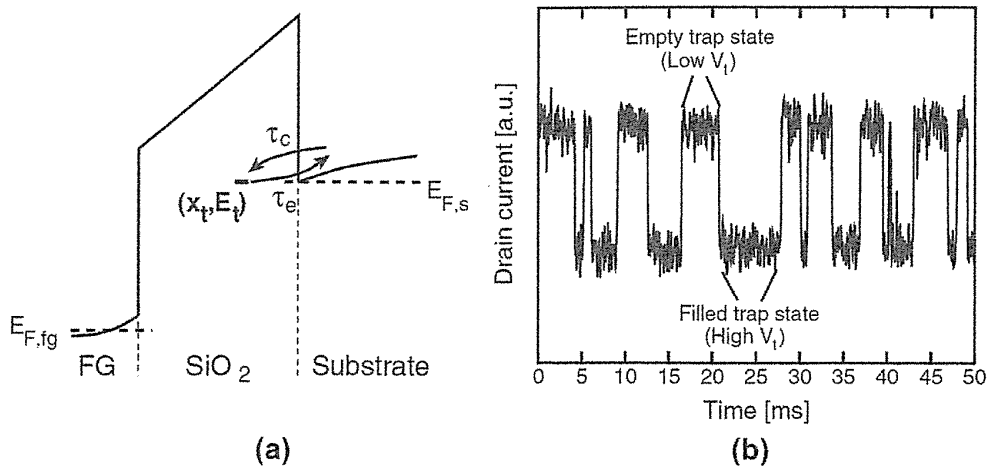


FIGURE 5.41 (a) Conduction band profile for a MOS structure under positive gate bias (read operation on a flash memory cell) and trap capture/emission processes. (b) Drain current in a flash memory cell as a function of time for fixed control-gate bias. Two-state RTN fluctuations can be clearly seen, associated to the empty and filled trap states.

A 512-Kbit NOR flash memory array in 65-nm technology was used to evaluate by sequentially reading the V_t for all the cells in the array up to 1000 times [66]. Figure 5.42a shows the measured V_t cumulative distribution at the first and the 100th read access on the array. No significant change in the V_t distribution between the first and the 100th read can be observed in the main distribution and in the tail. However, when ΔV_t between two read operations is evaluated for each cell, the V_t instability becomes clear, as shown in Fig. 5.42b, where the cumulative distribution (F) of ΔV_t between the first and the nth map. F is markedly different from the ideal step-like function centered in $\Delta V_t = 0$ that would be obtained for a stable V_t. And the cumulative distribution (F) also presents nearly exponential tails in its lower and upper parts. Moreover, these tails drift with time moving upward in the distribution.

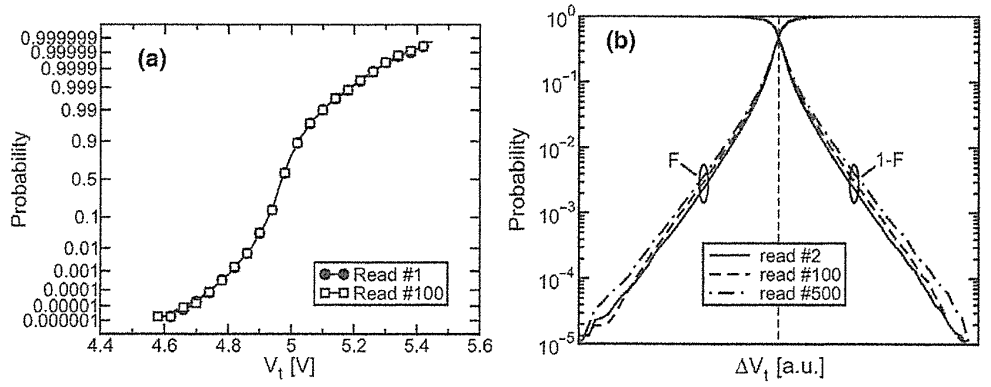


FIGURE 5.42 (a) V_t cumulative distribution at the first and the 100th read access to the array. (b) Experimental results for the cumulative probability F (and $1 - F$) of $\Delta V_t = V_t(n) - V_t(1)$, for read number $n = 2, 100, 500$. Cell V_t may shift randomly as $1/F$ noise.

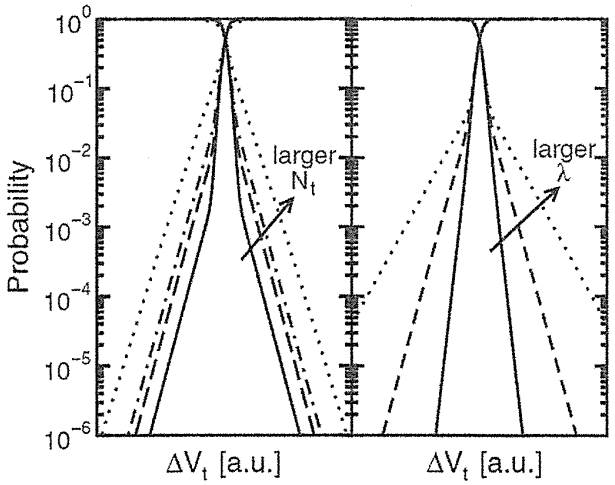


FIGURE 5.43 Cumulative probability distribution of ΔV_t calculated according to the model for different values of (left) N_t and (right) λ.

Figure 5.43 shows the cumulative distribution of ΔV_t, which is calculated by the model [66], as a function of the RTN-trap density N_t and the decay constant λ. Note that the model well reproduces the exponential behavior of the experimental ΔV_t distribution shown in Fig. 5.42b, with the tail amplitude determined by N_t and the tail slope related to λ. An increase in trap density N_t causes only an increase of the distribution tails, while the decay constant λ causes the "slope" of the exponential tails [68].

The model can be used for analysis of the traps location and energy in the oxide involved in the RTN V_t instability, as shown in Fig. 5.44. The increase in the elapsing time between the two read operations results in the activation of more traps, having

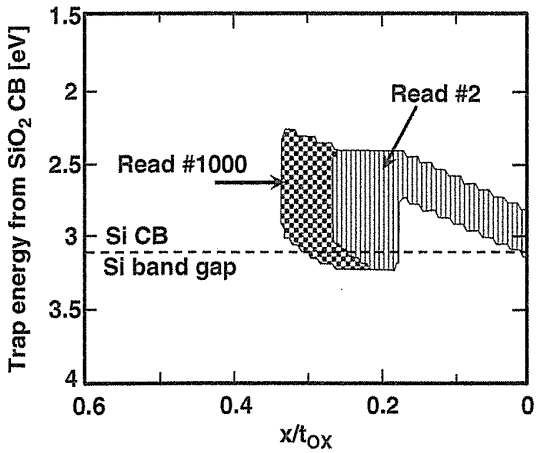


FIGURE 5.44 Tunnel-oxide region where traps that are active in the RTN V_t instability are located for read numbers 2 and 1000. An enlargement of the region inside the tunnel oxide appears for the increasing read number, corresponding to the activation of new traps in the RTN process. t_{ox} is the tunnel oxide thickness.

longer capture and emission time constants. This effect increases the randomness of cells V_t, thus increasing the width of the ΔV_t distribution. In Fig. 5.44, the RTN active traps are shown in the second and the 1000th read access, displaying an enlargement of the involved oxide region toward larger depths inside the tunnel oxide.

5.5.2 Scaling Trend of RTN

There are several reports that describe the scaling trend of RTN in flash memory cells [16, 39, 69, 70]. The dependences on gate length (L) and channel width (W) do not follow the same trend in these reports.

The RTN in flash memory becomes large as the device size is scaled down. Figure 5.45 shows the threshold voltage shifts estimated in each process node [16]. It was estimated that, if sense budget in multilevel flash memory is limited to about 1 V, total V_{th} shift exceeds the limitation of 1 V in a 45-nm process node.

Fukuda et al. [69] presented the statistical model of V_t fluctuation ($\Delta V_{t\text{cell}}$) in 20 to 90-nm design rule floating-gate NAND flash memory cells. It considers current-path percolation, which generates a large-amplitude-noise tail, caused by dopant-induced surface potential nonuniformity.

The scaling cell size reduces the average number of trap sites in a memory cell and increases the noise contribution of each trap site, as shown in Fig. 5.46. It is interesting to note that smaller cells have larger 3-σ $\Delta V_{t\text{cell}}$ but smaller mean $\Delta V_{t\text{cell}}$ than larger cells. This results in widening of $\Delta V_{t\text{cell}}$ distribution with cell size scaling, as shown in Fig. 5.47a. The 3-σ $\Delta V_{t\text{cell}}$ extracted from Fig. 5.47a increases by 1.8× from 90 nm to 20-nm technology nodes as shown in Fig. 5.47b. This is a much smaller increase than $\propto 1/LW$ suggests (>10×) and $\propto (LW)^{-1/2}$ suggests (>3x). In other words, the prospect of scaling is less pessimistic than $\propto 1/LW$ and $\propto (LW)^{-1/2}$. It would indicate $\Delta V_{t\text{cell}} \propto (LW)^{-0.24}$, which is a much slower scaling trend than the commonly accepted $1/L_{\text{eff}}W_{\text{eff}}$ trend, as shown Eq. (5.6).

Ghetti et al. [39,70] had also shown the scaling trend of NAND and NOR floating-gate cells. The scaling trend of RTN instabilities was investigated by using the Monte

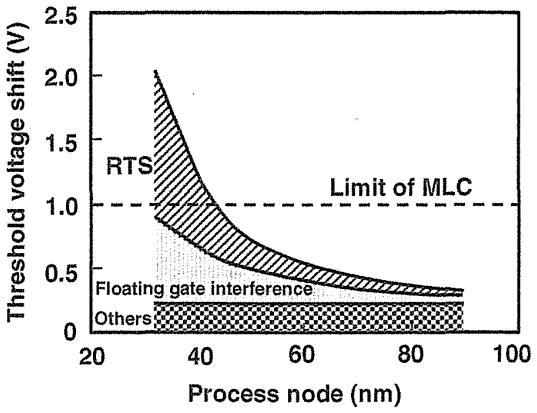


FIGURE 5.45 Estimation of threshold voltage shift as a function of process node.

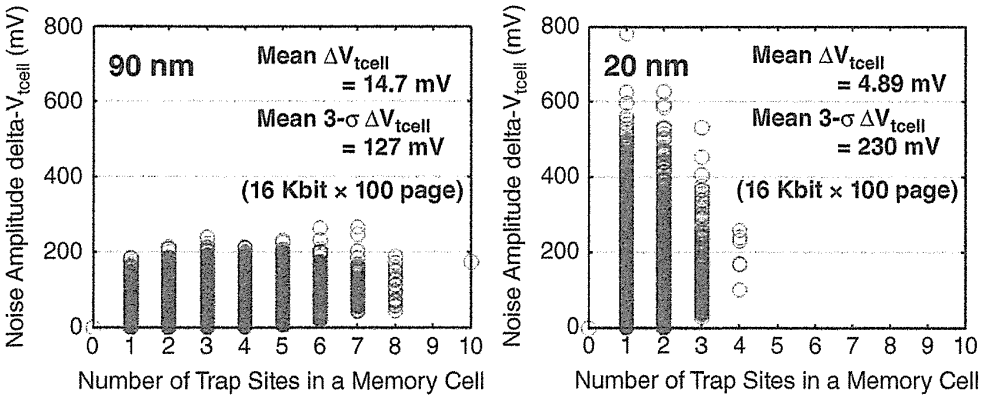


FIGURE 5.46 16-Mbit memory cell (1 Kbit × 100 page) Monte Carlo simulation results for 90 nm and 20-nm technology nodes. $N_D = 7E + 17/cm^3$ and $N_{trap} = 2E + 10/cm^2$ are assumed.

Carlo procedure with varying L, W, t_{ox}, and N_a, assuming discrete dopant atoms randomly placed according to a uniform distribution. The calculated slope of the RTN tails was divided by the control-gate to floating-gate capacitive coupling ratio α_G, to determine the real λ value of the flash cell.

Figure 5.48a shows the scaling trend for slope λ (see Fig. 5.43; unit is mV/dec) assuming $W = L$: a power-law $(W = L)^{-1.5}$ can well describe the dependence of λ on cell dimensions. This dependence is lower than the $(W = L)^{-2}$ (i.e., $1/WL$) expected from pure 1D electrostatics, but stronger than the $1/\sqrt{WL}$ dependence proposed in

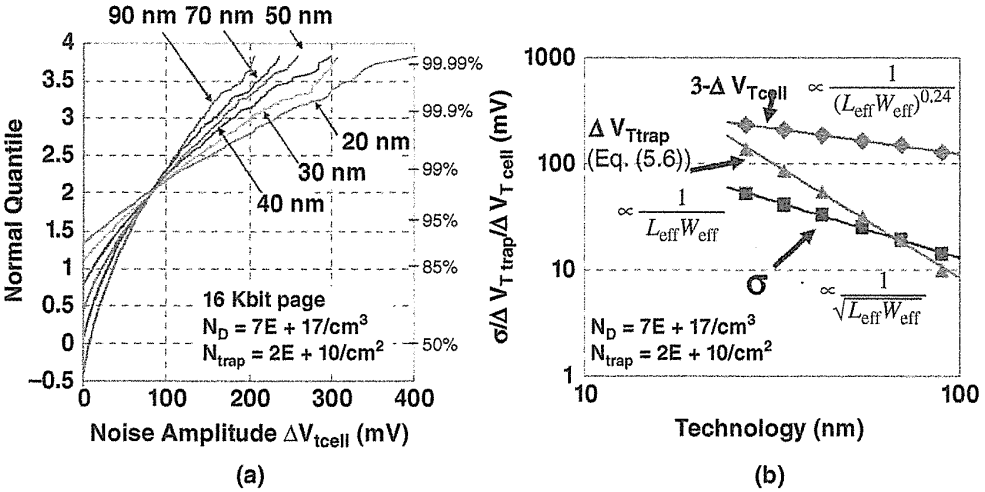


FIGURE 5.47 Random telegraph signal noise (RTN). dV_t is a linear with $(L_{eff} * W_{eff})^{-0.24}$. (a) Typical noise distributions of six technology nodes. Each trace represents the noise distribution of one 16-Kbit page. $N_D = 7E + 17/cm^3$ and $N_{trap} = 2E + 10/cm^2$ are assumed for all technology nodes. (b) Scaling trends of ΔV_{ttrap} of Eq. (5.6), σ, and 3σ of ΔV_{Tcell}. $N_D = 7E \pm 17/cm^3$ and $N_{trap} = 2E + 10/cm^2$ are assumed for all technology nodes.

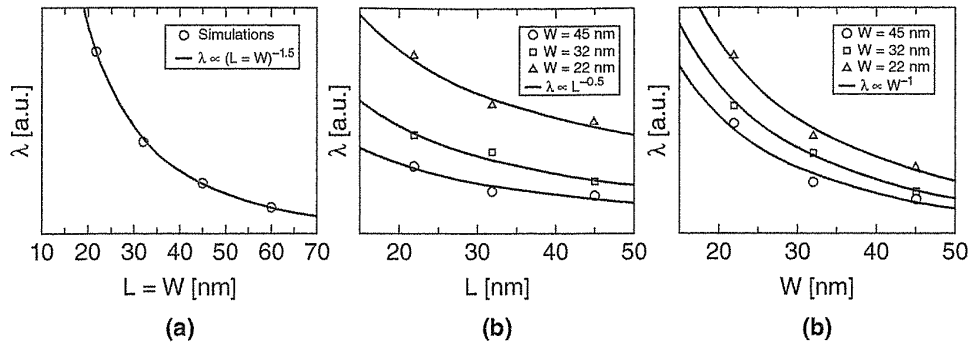


FIGURE 5.48 (a) Scaling trend for λ assuming $W = L$. (b) λ dependence on L with W as a parameter. (c) λ dependence on W with L as a parameter.

reference [69]. The separate dependences of λ on W and L were also investigated, as shown in Fig. 5.48b,c. Results indicate that the $(W = L)^{-1.5}$ power law can be decomposed in the form $W^{-1} \times L^{-0.5}$. This means that W has a stronger impact on λ due to the higher probability for a trap to effectively quench a percolation conduction path in narrower channels.

Figure 5.49a shows that the larger average substrate doping causes an increase of λ, due to the possibility to cause larger dis-uniformities of number of dopants in the channel inversion layer. This enhances the percolation effect and the current crowding at the cell channel edges which are responsible for the slope of the RTN exponentials, resulting in a square-root dependence of λ on Na.

Figure 5.49b shows that the scaling of the tunnel-oxide thickness t_{ox} reduces λ according to a slightly sublinear dependence $t_{ox}^{0.9}$. This is attributed to a less uniform current conduction over the active area as the gate electrode is placed further away from the channel, increasing the current crowding at the cell channel edges and giving

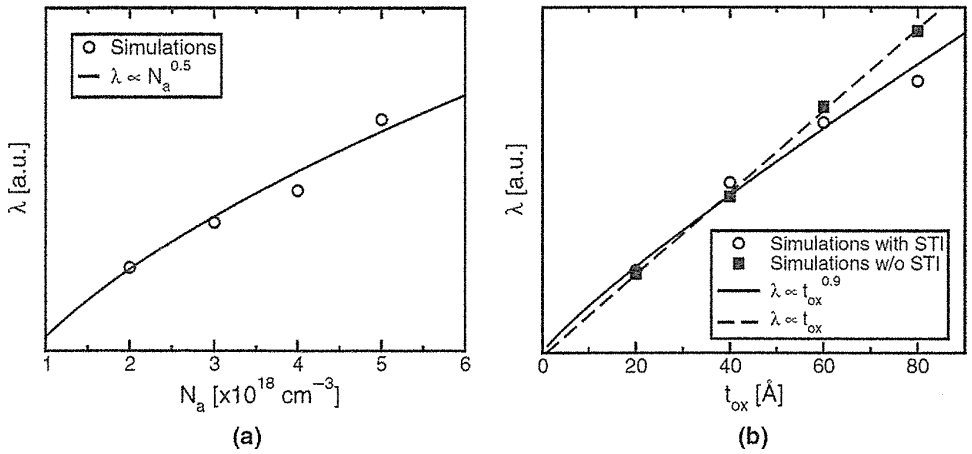


FIGURE 5.49 (a) λ dependence on Na for fixed cell geometry. (b) λ dependence on t_{ox} from 3D simulations including STI isolations (o) and for 2D-extruded structures neglecting cell active area edges (■).

a slightly weaker dependence with respect to the law of $\lambda \propto t_{ox}$. To confirm this result, Fig. 5.49b also shows simulation results obtained neglecting STI edges, that is, using 2D structures extruded in the W direction; in this case, a perfectly linear dependence of λ on t_{ox} is observed. The exponent 0.9 is therefore strictly related to the STI corner geometry, and a more general dependence ${t_{ox}}^{\alpha}$ with α slightly less than 1 can be adopted.

In summary of references [70 and 39], RTN scaling can be described by the following compact expression for λ that captures its dependence on all the main cell parameters:

$$\lambda = \frac{K}{\alpha_G} \frac{t_{ox}^{\alpha} \sqrt{N_a}}{W\sqrt{L}} \tag{5.7}$$

This equation represents a powerful result to investigate the scaling trend of the RTN instabilities and to derive scaling guidelines to optimize the design of future technologies with respect to RTN.

Figure 5.50 [39] shows a comparison between simulations for λ and the experimental data for NAND and NOR technologies in different feature sizes. It can be seen that a good agreement is reached between experimental results and TCAD simulation value of λ. In addition, solid lines show the dependence predicted by (5.7) for a constant value of K, determined by fitting all the simulated cases in Figs. 5.48 and 5.49; and using the real values for cell parameters of NAND and NOR devices. A good agreement of (5.7) with the experimental trend of λ is achieved on both technologies of NAND and NOR, demonstrating its validity to make RTN extrapolations on future technology nodes.

In order to investigate the RTN phenomena in future scaled NAND flash memory cells, there are many reports regarding RTN, such as random discrete doping [71],

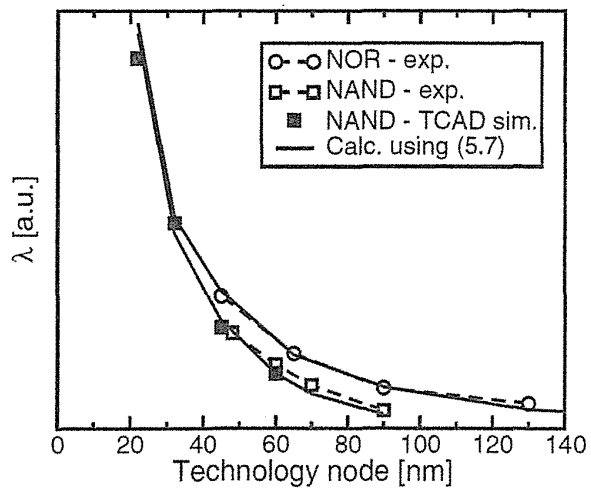


FIGURE 5.50 Experimental and simulation results for λ for different technology nodes. Results calculated by means of Eq. (5.7) are also shown and used for scaling projections.

non-equilibrium trap state [68], analytical investigation of the special and energetic position of trap [60], cycling impact on RTN [72], RTN impact on ISPP distribution [73], quick electron detrapping and random discrete dopants [74], tunnel-oxide nitridation effect [75], and inverse scaling phenomena due to the source/drain implantation condition effect in a 25-nm cell [76]. These reports are results of relatively larger device dimension (>25-nm cell); a practical data below 20-nm dimension cell will be presented in the future to clarify RTN mechanism in extremely scaled device.

5.6 CELL STRUCTURE CHALLENGE

A structural challenge of the SA-STI cell is also investigated, based on assumption in Table 5.1 in Section 5.2.1. The critical structure of the SA-STI cell is "CG formation margin," which is fabrication margin of CG between FGs [17]. Figure 5.51 shows an estimation of CG fabrication margin in the FG slimming structure beyond 2X-nm generation. FG width and CG width are assumed to be equal in this estimation. From

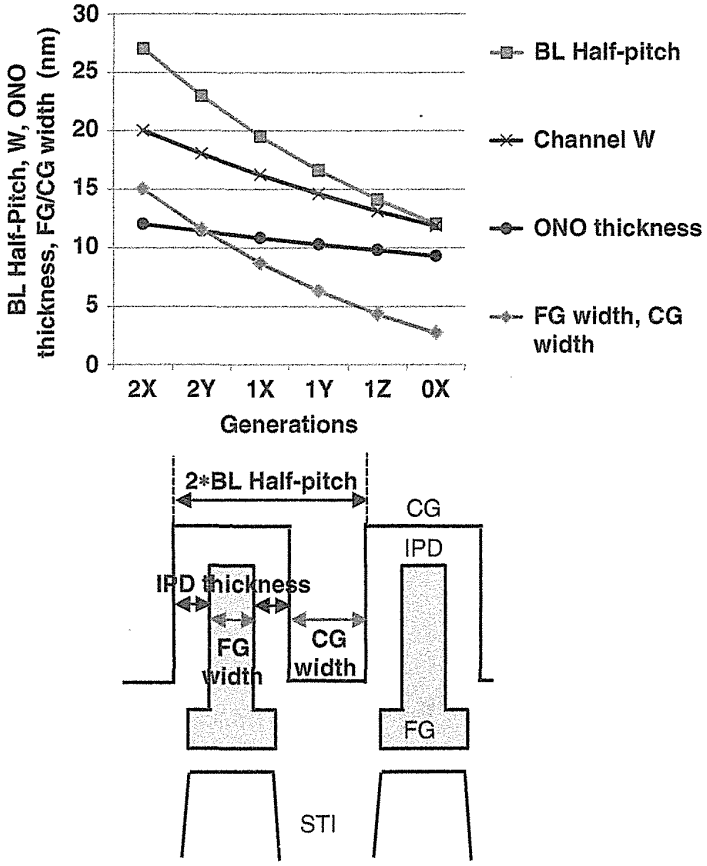


FIGURE 5.51 Estimated margin of CG fabrication between FGs. Very narrow CG and FG width of around 5 nm have to be controlled in 1Z-nm generation.

this estimation, as scaling down of the SA-STI cell, FG width and CG width are decreased to less than 10-nm width in 1X-nm cell. The FG and CG width have to be controlled around 5 nm in 1Y-nm and 1Z-nm generation, even ONO thickness is scaled down by the ratio of ×0.95 for each generation. The depletion effects in FG [38] and CG during programming and erasing have to be also suppressed. Metal or silicide material [77] will be applicable to FG and CG in future NAND cell.

In order to solve cell structure issues, so-called "Planar FG cell" has been proposed [28]. The planar FG cell has very thin FG thickness (~10 nm) with high-k inter-poly dielectric (IPD), as described in Section 3.5.

5.7 HIGH-FIELD LIMITATION

A program and erase voltage of NAND flash memory is high (~22 V) and cannot be drastically decreased because a high electric field (~10 MV/cm) in tunnel oxide is required for the Fowler–Nordheim tunneling mechanism during program and erase.

It had been reported that the new program interference phenomenon [18] occurred due to the high electric field between the program word line (WL) and the adjacent WL. This new program interference is that the V_{th}'s of the adjacent word lines are decreased during programming. This program interference became more severe as scaling memory cell, because this phenomenon is seriously aggravated as the gate space is decreased.

Figure 5.52 shows measurement conditions of new program interference phenomena. When WL(n) is programmed to the high V_t state, the high program voltage

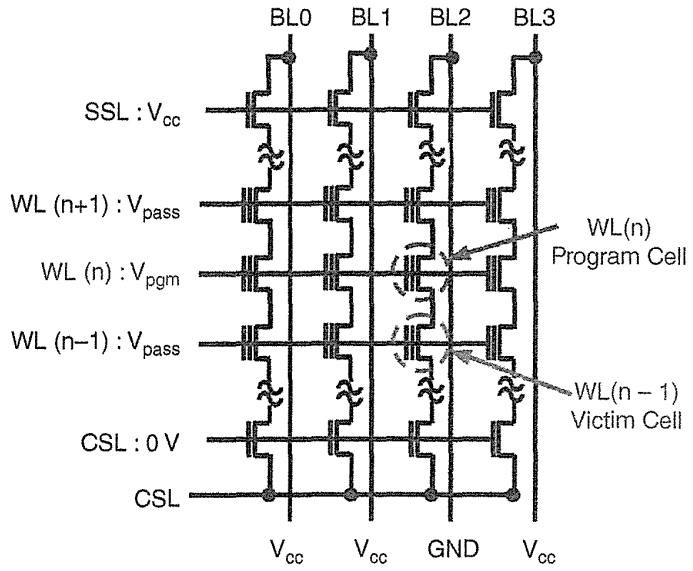


FIGURE 5.52 The schematic diagram for the test module with 4 bit lines and the basic program condition for the self-boosting scheme.

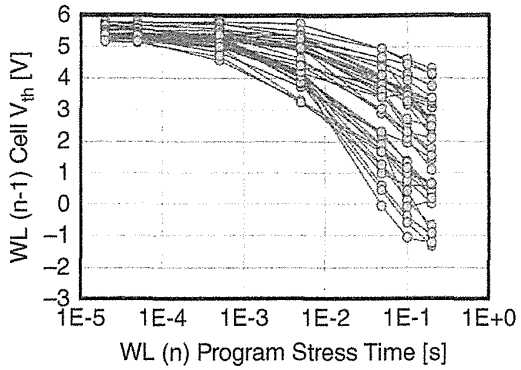


FIGURE 5.53 The V_{th} reduction of the victim cells in WL(n − 1), with the program stress times for 40 cells. The program stress voltage is 26 V and the pass voltage is 4.5 V.

(V_{pgm}) is applied to the control gate of WL(n) and the pass voltage (V_{pass}) is applied to other control gates for the program inhibit cells in the selected and unselected strings. Also normally, NAND flash memory is programmed from a lower word line which is closer to the CSL (common source line) to a higher word line which is closer to the BL (bit line). This new program interference phenomenon had been observed in the sub-30-nm memory cell under applying the relatively low pass voltage.

Figure 5.53 shows the victim cell V_t under acceleration condition of WL(n) program stress of V_{pgm} = 26 V and V_{pass} = 4.5 V. V_{th} reductions are clearly observed as increasing stress time with large variations in the measured 40 cells. These new program interference phenomena are observed in cell array as the under tail bits of the V_{th} distribution when all word lines are programmed to high state in the multilevel cell operation, as shown in Fig. 5.54. The gate design rule of the cell array is sub-30 nm. The under tail bits of Fig. 5.54 is generated at the $WL(n-1)$ when the $WL(n)$ is programmed. The final word line in the string without the upper word line does not have the under tail bits of the V_{th} distribution. As the pass voltage is increasing, the tail of V_{th} distribution is decreasing, as shown in Fig. 5.54.

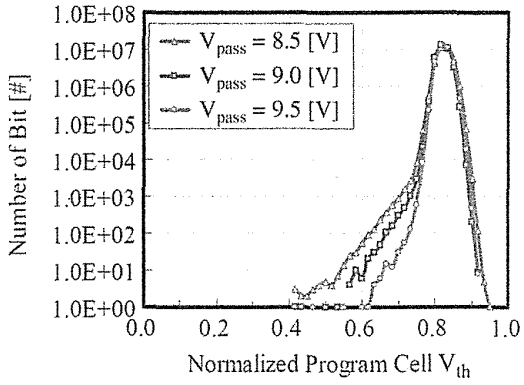


FIGURE 5.54 The distribution of program cell V_{th} after program of the cell array, where V_{pass} is 8.5 V, 9.0 V, and 9.5 V, respectively.

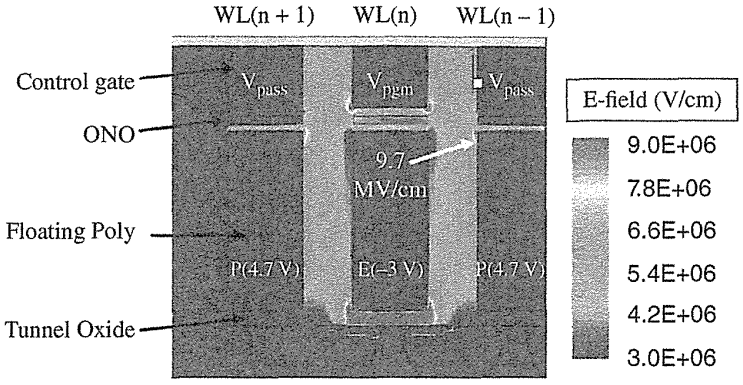


FIGURE 5.55 The simulation result of the electric field distribution under the program operation at $t = 0$, where the V_{pgm} is 24 V and the Vpass is 8.5 V.

Figure 5.55 shows a simulation result of the electric field in WLs for a 30-nm memory cell. The simulation was performed in 3D structure with practical dimension reflecting doping on Si channel, poly-Si floating gates, and control gates. The target V_{th}'s of the floating gates are adjusted from I_d–V_g curve by controlling charges of the floating gates. The maximum electric field of 9.7 MV/cm is observed in between the top edge of the floating gate and the bottom edge of the control gate, in condition of $V_{pgm} = 24$ V and $V_{pass} = 8.5$ V. It is confirmed that the edge field is large enough to generate FN tunneling current between a control gate and a neighbor floating gate.

V_{th} reductions had been evaluated in different gate design rules which are from sub-30 nm to sub-50 nm. Figure 5.56a shows the gate space dependence of this phenomenon. The space between the adjacent WL gates is very critical to this phenomenon. Although the V_{th} reductions are measured at different program voltage

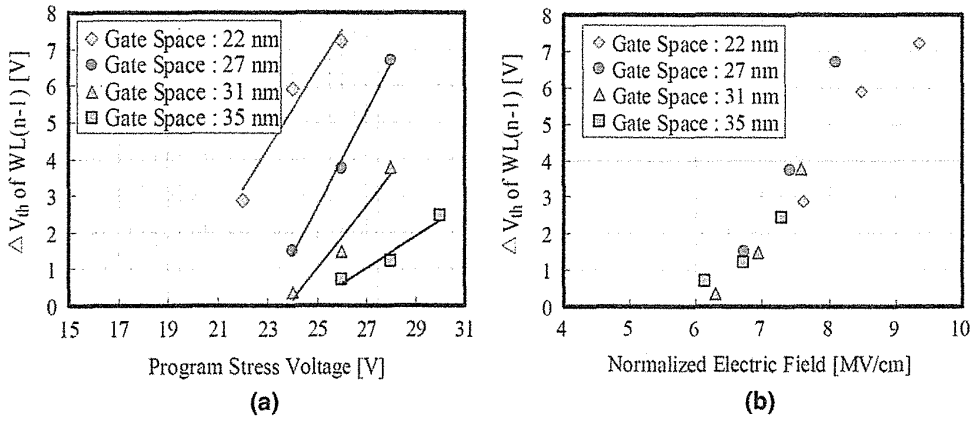


FIGURE 5.56 (a) The V_{th} reduction of WL(n – 1) for samples with different gate space after 0.2 s of program stress; the pass voltage is 4.5 V. (b) The V_{th} reduction of WL(n – 1) for samples with different gate space. The x-axis is the electric field between the control gate of WL(n) and the WL(n – 1).

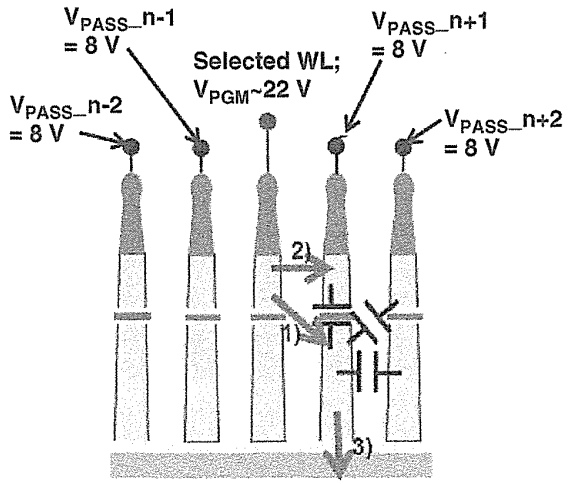


FIGURE 5.57 Word line (WL) high-field problem. (1) Charge loss from neighbor FG to selected WL (CG), (2) leakage or breakdown between selected WL and neighbor WL, and (3) program disturb in neighbor cell. Charges (electron) has injected from substrate to FG.

according to the gate design rule, all of the measured results are simply generalized with the electric field between the control gates. The normalized results with the electric field are shown in Fig. 5.56b, where the y-axis is the decrease of the victim cell V_{th} after program stress for 0.2 s. As shown in Fig. 5.55, the electric field of the WL edge is susceptible to the shape and profile. If the coupling ratio of the floating gate cell is similar, the electric field between the control gates is a simple and proper parameter for representing this phenomenon. In Fig. 5.56b, the V_{th} reduction in WL($n-1$) was observed above 6.0 MV/cm regardless of the gate design rules. This means that the V_{th} reduction of WL($n-1$) is the general phenomenon related to the electrical field between the floating gate and the adjacent control gate.

In NAND flash memory cell, the most serious high-field problem is caused in between selected-word line (WL) and neighbor-WL during programming. In 2X-nm cell, the selected WL is in V_{pgm} (~22 V) and the neighbor WL is in V_{pass} (7–10 V), as shown in Fig. 5.57. There are three problems: (1) charge (electron) loss; charges in FG of neighbor cell is discharged to selected WL [18], as shown in Figs. 5.52–5.56; (2) WL leakage or breakdown; high field between WLs (V_{PGM}–V_{PASS}_$n + 1/n - 1$ >10 V) may cause leakage or breakdown; (3) program disturb; charge (electron) has injected from substrate to FG. In order to mitigate these problems, it will be important for future generation to optimize V_{PASS}_$n + 1/n - 1$.

Figure 5.58 shows the estimated electric field as a function of V_{pgm}. Criteria of maximum electric field between WLs, which is determined by (1) charge loss between FG and selected WL [18], can be increased from 6 MV/cm [18] to 9.5 MV/cm [10] by using a WL air gap. Even if WL air gap is used, an available range of (V_{pgm}–V_{pass}) is reduced from 15 V of 2X-nm cell to 10 V of 1Z-nm cell (V_{pgm} = 23 V/V_{pass} = 8 V to V_{pgm} = 18 V/V_{pass} = 8 V). However, if 1Z-nm generation (word-line half-pitch; 10.6 nm, see Table 5.1) is used, V_{PGM} = 18 V is available to program in case of

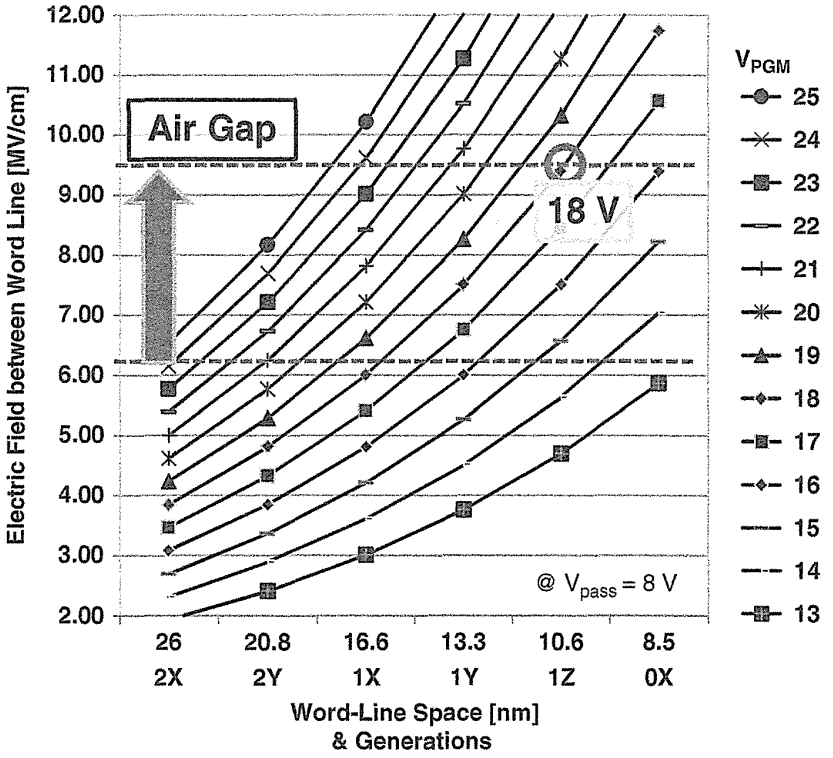


FIGURE 5.58 Estimated electric field between word lines during programming. In 1Z-nm generation (word-line space 10.6 nm), high 18 V can be applied due to using an air gap in the word-line space.

neighbor $V_{PASS}\_n + 1/n - 1 = 8$ V. It means that enough high voltage of $V_{PGM} = 22$ V/$V_{PASS}\_n + 1/n - 1 = 12$ V is available for programming with decreasing $V_{PASS}\_n + 2/n - 2$ to prevent (3) program disturb to manage high-field problems.

5.8 A FEW ELECTRON PHENOMENA

By scaling down memory cell size, the number of electrons stored on the floating gate is significantly decreased due to the decrease of inter-poly dielectric ONO capacitance. Figure 5.27 represents the number of electrons per bit (for ΔV_t of 3 V) as a function of the technology node for NAND and NOR flash memory cells [19, 57]. It is expected that around 100 electrons are stored in 1X-nm design rule cell. By scaling down memory cell size further, the number of stored electrons will be much less than 100 electrons. It will be sufficiently small enough to make few electron phenomena observable. Then the impact of these single electron phenomena on the performance of floating-gate (FG) memory cell had been studied [19, 57]. The charging and discharging of scaled FG memory cells should no longer be considered as a continuous phenomenon but as a sum of discrete stochastic events. This results

in an intrinsic dispersion of both the retention time and of the memory programming window.

The stochastic character of the charging process in a few-electron memory had been addressed in reference 78 in the case of a nanometer-size storage node. It had been also demonstrated that there was an uncertainty regarding the number of charged electrons in the FG after programming due to the Poisson nature of the electron charging. Moreover, it was shown that Coulomb blockade modified the charging kinetic.

For simulation in references 19 and 57, it was assumed that in a scaled memory device, the charging (discharging) of one electron to (from) the FG could be described by a Poisson process with an exponential law in time and a lifetime τ_d, depending on the charge stored in the FG and on the tunnel-oxide transparency (τ_c, the electron capture time constant, depending on the oxide thickness and the programming voltage). However, no Coulomb blockade was taken into account. Indeed, in the case of continuous FG memory cell, the charging energy was negligible due to the large dimensions of the storage node.

Figure 5.59a represents the calculated retention-time distribution for various numbers (N) of electrons per bit. By decreasing the number (N) of electrons per bit, the retention time T_R probability density is strongly evolved from a Gaussian-like distribution (when $N \sim 250$) to a pure exponential/ Poisson-like distribution (when $N \sim 5$). We can also see that the dispersion around the mean value increases as N is decreased.

Figure 5.59b shows the cumulative probabilities of the retention time T_R for different values N of electrons per bit. For large values of N, the cumulative-probability evolution is very tight distribution. On the other hand, we can see that as N decreases, the distribution tails have much shorter retention time, which means that the number of discharged memory cells will become critical. For example, in the case of a 1-Mb memory array with memory cells having a mean retention time of 10 years and

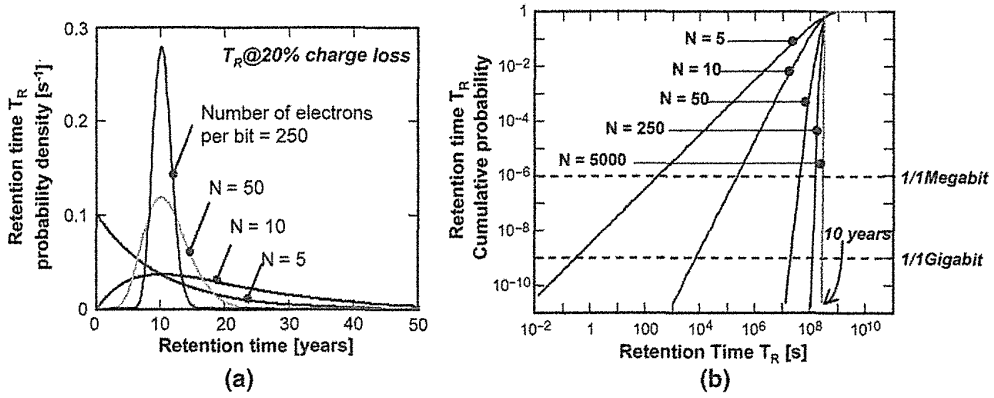


FIGURE 5.59 (a) Probability density of the retention time T_R for memories with reduced number of electrons per bit N. The mean T_R is fixed at 10 years. (b) Cumulative probability of retention time T_R (from (a)) of memories with reduced number of electrons per bit N.

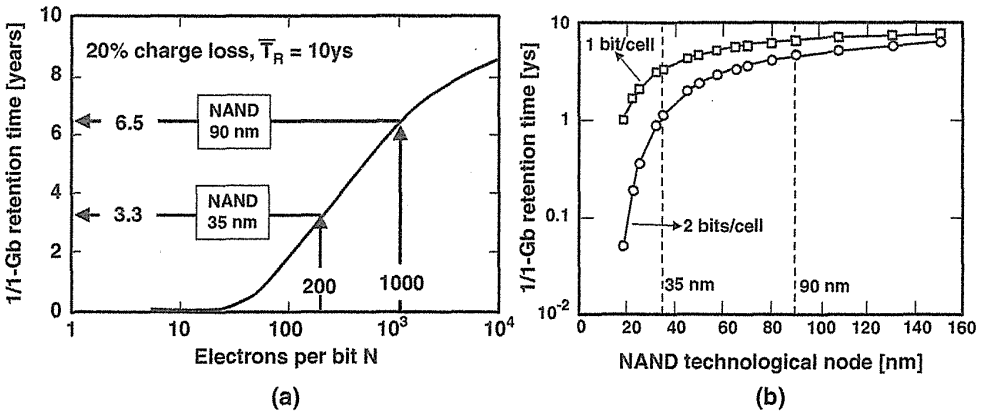


FIGURE 5.60 (a) Calculated failure time (i.e., retention time of 1 bit over 1-Gb array) due to single-electron discharging of the FG plotted as a function of the number of electrons per bit. The mean retention time is ten years for each cell array. (b) Calculated failure time (i.e., retention time of 1 bit over 1-Gb array) due to single-electron discharging of the FG plotted as a function of the technological node for 1-bit/cell and 2-bit/cell memory devices.

containing five electrons per bit, we can observe that one cell will be discharged after only a few hundreds of seconds.

These dispersion results were extrapolated to future memory generations with a smaller number of stored electrons in FG [19, 57].

Figure 5.60a shows the evolution of this failure time as a function of the number (N) of electrons per bit. The failure criterion is defined as the retention time of 1 bit over a 1-Gb array, and the retention criterion corresponds to 20% of the charge loss. We can see that the failure time reduction can become relevant when the number of electrons per bit is decreased and becomes really critical in few electron memories. If we consider the 90-nm NAND flash technology node, a threshold-voltage shift of 3 V corresponds to about 1000 electrons per bit. Thus, in this case, the retention time of one erratic bit over 1 Gb will be equal to 6.5 years. However, if the 35-nm NAND flash technology node, which corresponds to about 200 electrons per bit, is considered, the retention time of one erratic bit decreases drastically to 3.3 years, which could be very critical.

Moreover, in multilevel cells, the retention-time operation margins will be further reduced, with the number of electrons per bit being decreased by $2^{\text{bit/cell}}-1$. Figure 5.60b shows the retention time of 1 bit/1 Gb as a function of the NAND flash technology node for 1-bit/cell and 2-bit/cell memory technologies. This plot illustrates that in future technology nodes, the multilevel memories will decrease critically the failure time of high-density memory arrays. Thus, for the 35-nm memory node, the introduction of 2 bits/cell induces a reduction of the failure time from 3.3 years to one year.

As shown above, the decreasing stored electron has intrinsically degraded the data retention time of tail bits in scaled memory cells. These tail bits would not be a serious problem in an actual NAND flash usage with system solutions, such as ECC,

and so on, if the number of stored electrons is more than 50. However, in a future device—for example, a 3D SONOS cell with very small channel diameter and short channel—the number of stored electrons would be much reduced. A few electron phenomena have to be considered as being one of the scaling obstacles to keeping the appropriate reliability.

5.9 PATTERNING LIMITATION

A NAND flash memory cell can be easily scaled down as scaling a minimum device dimension, without any electrical, operational, and reliability limitations due to an excellent scalability of the SA-STI cell [4] as well as an excellent gate length (L) scalability of the uniform program/erase operation scheme [79–83]. Therefore, the memory cell size could decrease straightforward as feature size decreased from 0.7 μm to current 1Y-nm generation, as shown in Chapter 3.

The cell size of the NAND flash became ideal $4*F^2$ by the SA-STI cell. The feature size (F) is normally determined by the capability of the lithography tool. At present, the most advanced lithography tool is the ArF immersion (ArFi) stepper. Minimum feature size is 38–40 nm. Then the scaling of feature size (F) had been limited by 38–40 nm. In order to accelerate to scale-down NAND flash memory cell size further, the double patterning process [84,85] had started to be used from 3X-nm generation. The side-wall spacer was used as a patterning mask in the conventional double patterning process, as shown by the SPT process in Fig. 5.61 [10]. Thanks to double patterning, feature size could be scaled down from 38–40 nm to 19–20 nm. Furthermore, quadruple (×4) patterning has been used beyond 20 nm, as shown by the QSPT process in Fig. 5.61 [10]. Feature size (F) can be scaled down from 19–20 nm to 9.5–10 nm by using ArFi if serious physical limitations described in this chapter

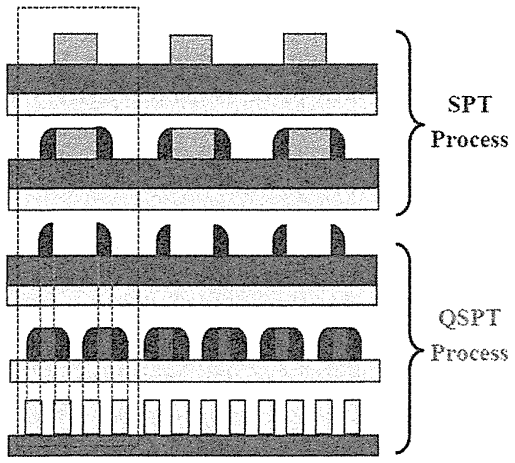


FIGURE 5.61 Schematic diagram of SPT (double patterning) and QSPT (quadruple patterning) key fabrication steps. Two times spacer patterning (QSPT) are used to make mid-1X patterning. SPT; Spacer Patterning Technology, QSPT; Quadruple Spacer Patterning Technology.

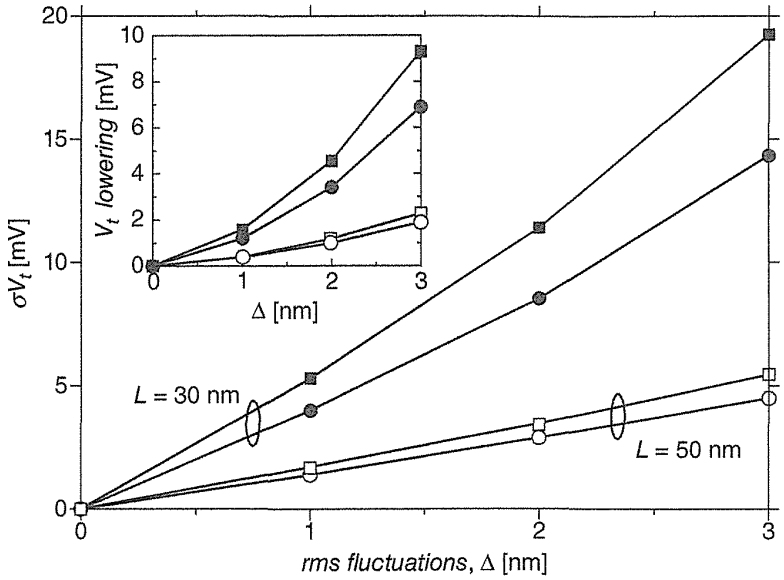


FIGURE 5.62 Standard deviation of threshold voltage σV_t for 30 × 50 and 50 × 50 nm MOSFETs as a function of RMS fluctuations Δ, at $V_D = 1.0$ V (squares) and $V_D = 0.1$ V (circles).

are not considered. For scaling beyond 9.5–10 nm, new tool or technology to make fine pattern is absolutely required. One candidate is the EUV (extreme ultraviolet) lithograph tool. However, EUV tool is not available at least in 2014.

Line edge roughness (LER) is one of the serious problems in fine patterning [86]. The LER has tended to stay relatively constant as a device scale. In the case of an aggressive scaling, the LER has become a larger fraction of the channel width or length. Figure 5.62 shows the variation in standard deviation of V_t as a function of RMS LER parameters [86]. As RMS LER increased, V_t variation is increased. This would cause a large variation of cell V_t in a <20-nm NAND flash cell.

5.10 VARIATION

One of the obstacles to scale down memory cell size is the increasing role played by variability effects [87], that strongly influence the threshold voltage (V_t) distribution of NAND flash memory cells [88, 89], affecting their performance and reliability.

To investigate the variation effect, the compact model for the NAND flash memory array was presented [90]. The model includes 3 NAND strings of 32 cells with the select transistors, and it accounts for floating-gate capacitive coupling interference effects. Only the central string is simulated, while the two neighbor NAND strings set the boundary conditions for the electrostatic couplings among the cells. The floating-gate devices are described via capacitors in series to MOS transistors, whose parameters were extracted as detailed in [90] for any technology node.

The compact SPICE model was used in a Monte Carlo framework to obtain the V_t distribution from the calculated string current in read conditions. In simulations, the device parameters were randomly changed to account for the different variability effects. In particular, both process-induced fluctuations in the cell geometry and more fundamental (intrinsic) ones were considered, for example, due to the discrete nature of the charge. The former of process-induced fluctuations in the cell geometry include W, L, tunnel, and inter-poly dielectric thickness fluctuations (indicated as WF, LF, TOXF, IPDF, respectively) as well as fluctuations in the control to floating-gate coupling coefficient. The latter of the fundamental (intrinsic) ones account for random dopant (RDF) and oxide trap fluctuations (OTF). Process-induced fluctuations are directly inserted in the compact model by changing the device parameters (W, L, etc.), according to Gaussian distributions whose spreads are extracted from process data. The implementation of the fundamental (intrinsic) contributions is carried out as follows: The RDF effect on V_t was accounted for by the analytical formula reported in reference 87, while the V_t variability due to OTF was implemented as $\sigma_{\mathrm{OTF}} = K_{\mathrm{ox}} Q^{\alpha}{}_{ox} / \sqrt{WL}$ with $\alpha \approx 0.5$ and K_{ox} and Q_{ox} fitted on cycled distribution data.

Figure 5.63 shows the simulation results including the spread of neutral cell V_t with the experimental V_t distribution measured on a page of a 41-nm NAND flash array [88, 89]. This result shows a good agreement between measurement and simulation with support of the correctness of the variability models. The slight underestimation of the spread is probably caused by the soft erase operation from programmed V_t distribution.

Figure 5.64 shows the simulated and experimental standard deviation of the V_t distribution σ_{V_t} as a function of the NAND flash technology node from 100-nm to 25-nm rule. The results show an increase in V_t variability as scaling memory cell due to the degradation of all the fluctuations factors impacting array functionality. Figure 5.65 shows the detailed relative weight of the major variability factors as a function of the technology node, represented by RDF (random dopant fluctuation), OTF (oxide trap fluctuation), and fluctuations in W and L. The results clarify that the variability is dominated by several factors, not dominated by single killer factor.

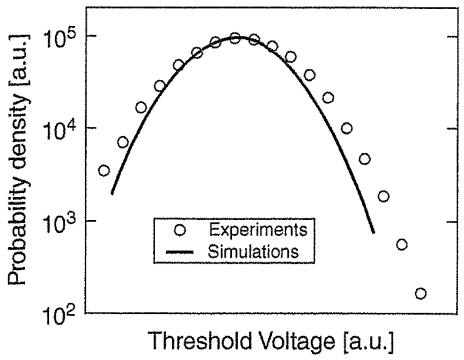


FIGURE 5.63 V_t distribution for a page of a 41-nm NAND flash array and the corresponding simulation results.

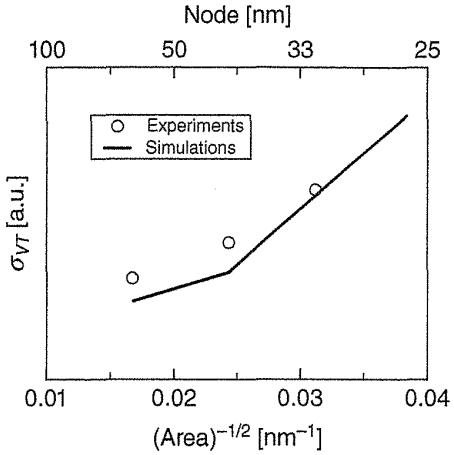


FIGURE 5.64 Comparison between modeling results and experimental data for σ_{V_t} as a function of technology node.

In a future device of less than 25-nm technology node, RDF is expected to play a more important role on V_t statistical dispersion. The consequent increase in the V_t spread should be carefully considered when the array functionality is implemented in program, erase, and read conditions [91].

The model, which was used to investigate the behavior of the neutral cell V_t, was expanded to simulate the program and erase operation. One of the most important parameters that play a fundamental role in the σ_{VT} during P/E operations is the control-gate to floating-gate coupling coefficient (α_G). This parameter is related to the structure adopted for the floating-gate definition, and its fluctuations depend on the spread in the geometrical parameters, which are schematically shown in Fig. 5.66a. The contributions of such factors to the spread in α_G are shown in Fig. 5.66b on a normalized scale. In this simulation results, fluctuations in t_{FG} and t_R play the major role for all technology nodes.

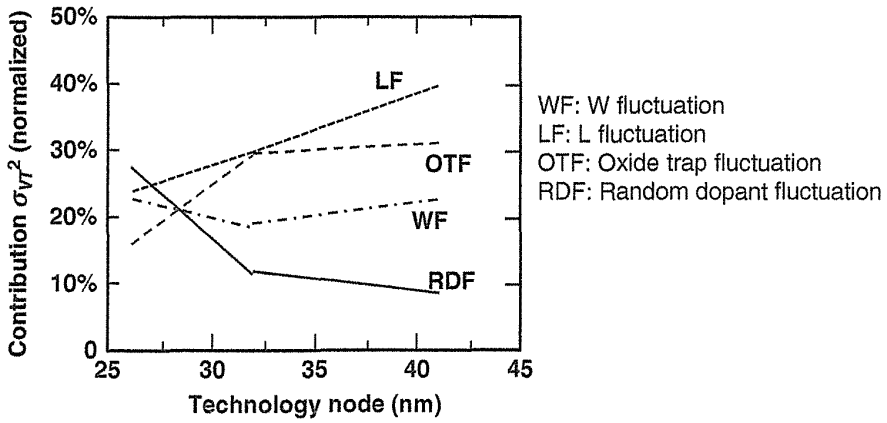


FIGURE 5.65 Most important contributions to σ_{VT}^2 of neutral cells (normalized) for different technology nodes.

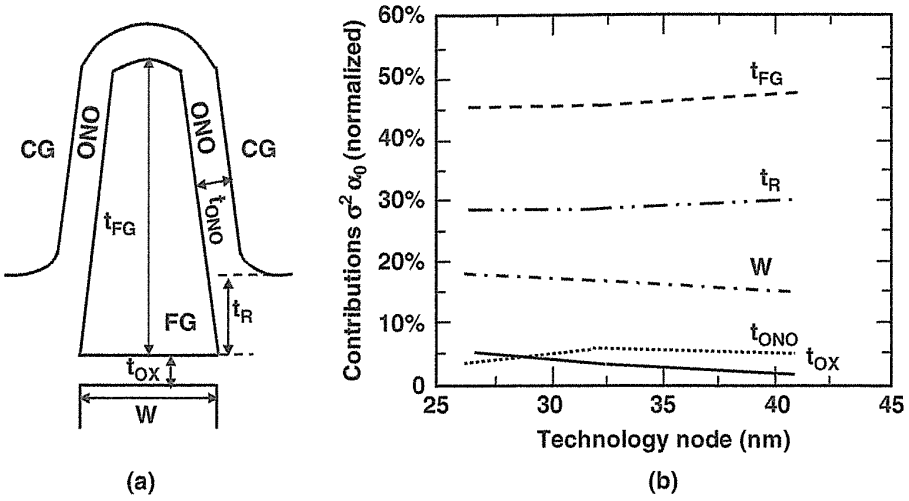


FIGURE 5.66 (a) Schematic view of the cross section of a memory cell along the W direction, showing the floating-gate geometry. (b) Individual contributions to the spread in the coupling coefficient α_G for the different technology nodes.

An RDF (random dopant fluctuation) is major contributor to neutral V_t variation in less than 25-nm memory cells, as shown in Fig. 5.65. As memory cells are scaled down, the number of dopant atoms per cell decreases, resulting in a larger standard deviation in the threshold voltage. Figure 5.67 shows the number of boron atoms per cell versus technology node and increasing variation in dose per cell at smaller nodes

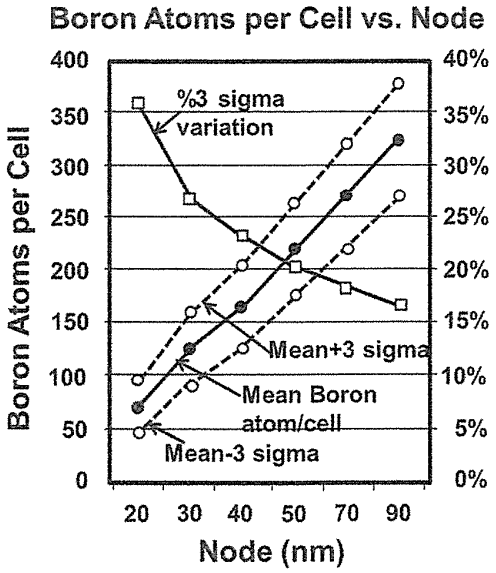


FIGURE 5.67 Number and 3σ of channel dopant atoms versus transistor size with constant V_t scaling.

[23,92]. The atomistic nature of substrate doping has been clearly shown to result in a fundamental V_t spread for MOS field effect transistors given by reference 87.

$$\sigma_{\mathrm{RDF}} = 3.19 \times 10^{-8} \frac{t_{\mathrm{ox}} N_A^{0.4}}{\sqrt{WL}}$$

5.11 SCALING IMPACT ON DATA RETENTION

It had been reported that data retention characteristics had the neighbor cells data pattern (back-pattern) dependence [93]. A programmed cell has a larger threshold-voltage (V_t) loss when its neighbor cells are in the erased state than when they are in the program state. The cells on the same bit line and word line have a similar impact on the acceleration of the V_t loss. This phenomenon is explained by an influence of charge in a neighbor cells, so that a stored charge in a neighbor cell has an impact on the electric field of tunnel oxide at a corner in the gate and active area of target cell.

The cell arrays in 60-nm technology were used to analyze for the electrical characterization of data retention on single cells. These arrays allow the arbitrary bias conditions of three adjacent WLs and BLs in the central part of the NAND cell matrix, as shown in Fig. 5.68a. The shaded circle in the figure shows the selected cell

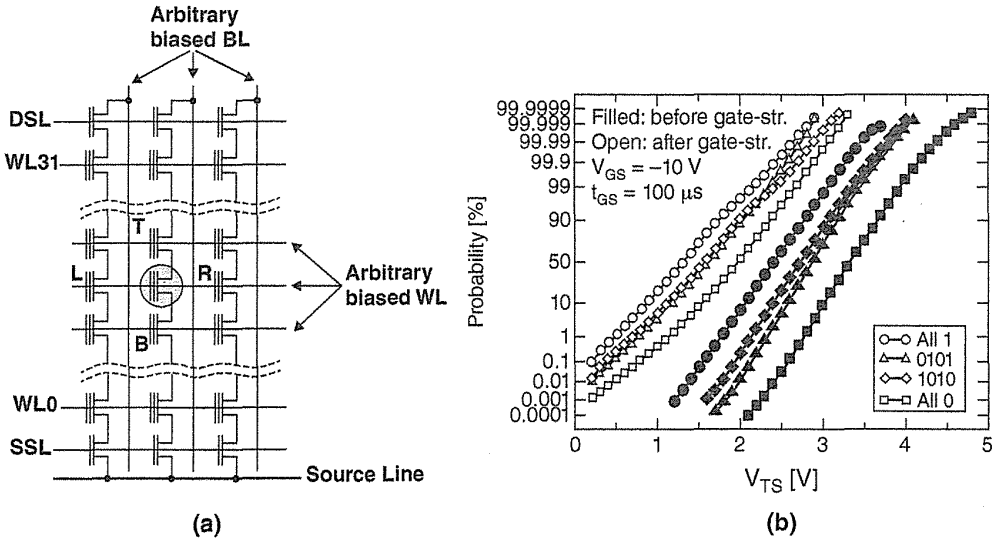


FIGURE 5.68 (a) Schematic for the cell connection in the NAND array, evidencing the nine cells whose WL and BL can be arbitrarily biased in the analytical cell array. The shaded circle highlights the selected cell whose V_{TS} was monitored during the gate-stress experiments for different V_{TA}'s of the adjacent cells at its (L) left, (T) top, (R) right, and (B) bottom. (b) Cumulative V_{TS} distributions measured on a 70-nm NAND test chip (filled symbols) before and (open symbols) after a 100-μs negative gate-stress experiment at $V_{GS} = -10$ V. Results for different background patterns are shown.

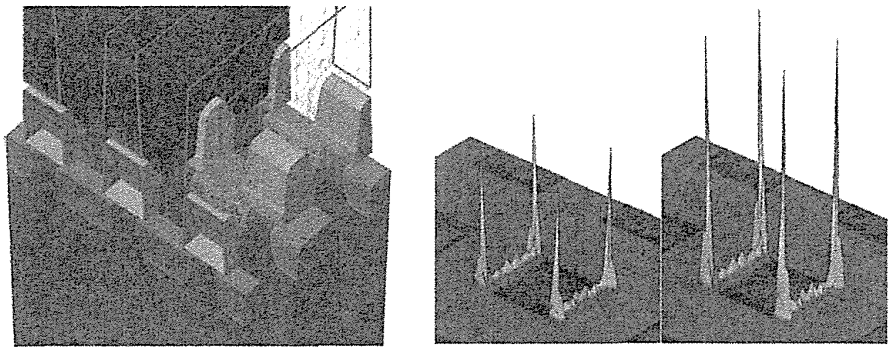


FIGURE 5.69 (a) Template NAND device structure for 3D TCAD (Technology Computer Aided Design) simulations. (b) Calculated tunneling current density at the edge of the selected cell active area along the WL direction, for the All 0 and the All 1 background patterns. A fixed negative V_{FG},*s* is used.

for the gate-stress experiments by applying V_{GS} to its WL with all the other WLs of the NAND string at V_{pass}. Prior to the gate stress, specific array V_t data patterns were set by selective programming of the cells at the left (**L**), top (**T**), right (**R**), and bottom (**B**) of the selected one. Cell V_t is set from low-V_t erased level nearly equal to −3.5 V (referred to as state 1) to a high-V_t level nearly equal to 3 V (referred to as state 0).

Figure 5.68b shows that the negative shift of the V_{TS} distribution is larger in the All 1 (erase) than in the All 0 (program) case, with intermediate results obtained for the 0101 and the 1010 patterns [left (**L**), top (**T**), right (**R**), and bottom (**B**)].

In order to investigate neighbor cell data pattern dependence, 3D TCAD simulation had been carried out, as shown in Fig. 5.69. The template NAND device structure for 3D TCAD simulations is shown in Fig. 5.69a. The calculated tunneling current density over the selected cell active area is presented in Fig. 5.69b for neighbor cells of the All 0 (left) and the All 1 (right) patterns in 60 to 70-nm design NAND flash memory cell. A larger current flow with four peaks is caused at the corners of the active area, and the source/drain junction overlaps with the cell floating gate. These peaks are higher for the All 1 case because an electrostatic profile at the edges of the active area is strong when the neighboring floating gates are charged positive. The current density profile along the BL and WL directions at the corners of the cell active area are shown in Fig. 5.70. It was confirmed that the larger tunneling current flows over the source/drain junctions in the negative gate-stress conditions.

The above results were presented to be analyzed on 60 to 70-nm NAND flash memory cells. By scaling a memory cell, this phenomenon of tunneling current confining at the FG corner and edge has been exaggerated. The current 1X-nm memory cell would have a strong impact on this phenomenon. Data retention characteristics would be much worse in future memory cell scaling.

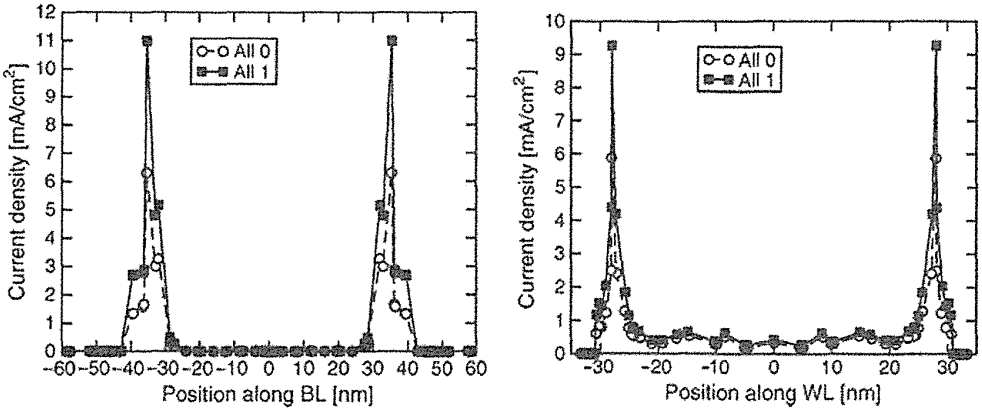


FIGURE 5.70 Calculated tunneling current density at the edge of the selected cell active area along the BL direction (left) and the WL direction (right), for the All 0 and the All 1 background patterns. A fixed negative V_{FG},*s* is used.

5.12 SUMMARY

The scaling limitations and challenges over 90 to 0X-nm generations was discussed for two-dimensional (2D) floating-gate NAND flash memories. The scaling challenges were categorized as (1) narrow read window margin (RWM) problem, (2) floating-gate capacitive coupling interference, (3) program electron injection spread, (4) random telegraph signal noise (RTN), (5) cell structural challenge, (6) high-field (5–10 MV/cm) limitation, (7) a few electron phenomena, (8) patterning limitation, (9) variation, and (10) scaling impact on data retention.

First, (1) the narrow RWM was discussed by extrapolating physical phenomena of FG–FG coupling interference, electron injection spread (EIS), back pattern dependence (BPD), and random telegraph noise (RTN). The RWM is degraded not only by increasing programmed V_t distribution width, but also by increasing V_t of erase state mainly due to large FG–FG coupling interference. However, RWM is still positive in 1Z-nm (10-nm) generation with 60% reduction of FG–FG coupling by the air-gap process.

Second, floating-gate capacitive coupling interference, which was a major contributor to degrade RWM, was discussed. Air gaps between word lines and in STI are the solutions to improve of floating-gate capacitive coupling interference.

Next, (3) program electron injection spread and (4) random telegraph signal noise (RTN) were described as contributors to RWM.

Then, (5) structural challenge was discussed. The control gate (CG) fabrication margin between floating gates (FGs) is becoming much more severe beyond 1X-nm generation. Very narrow 5-nm FG width/space has to be controlled. And for (6) the high-field problem, the high field between CGs (word lines: WLs) is critical during the program. By using WL air gap, the high-field problem can be mitigated, and 1Y/1Z-nm generations could be realized.

After that, several scaling problems of (7) a few electron phenomena, (8) patterning limitation, (9) variation, and (10) scaling impact on data retention were discussed. These problems are inevitable to scale down NAND flash memory cells.

To improve RWM margin and reliability margin, operational techniques and system solutions are effective to manage these margins. One example is the "randomization" [94,95]. The data pattern that is programming to memory cell is randomized by using code data. The "0" and "1" data become random data, and nearly 50% for both "0" and "1". Therefore, worst case of data pattern can be avoided. The randomization can improve RWM with mitigating floating-gate capacitive coupling (Section 5.2), back pattern dependence (Section 5.2), and scaling impact on data retention (Section 5.11). The other example is the moving read algorithm, as described in Section 4.7. The moving read operation can greatly improve the failure rate of the V_t shift of data retention as well as that of V_t distribution widening by floating-gate capacitive coupling interference, program electron injection spread, and so on.

REFERENCES

[1] Masuoka, F.; Momodomi, M.; Iwata, Y.; Shirota, R. New ultra high density EPROM and flash EEPROM with NAND structure cell, *Electron Devices Meeting, 1987 International*, vol. 33, pp. 552–555, 1987.

[2] Aritome, S. NAND Flash Innovations, *Solid-State Circuits Magazine, IEEE*, vol. 5, no. 4, pp. 21, 29, Fall 2013.

[3] Aritome, S.; Hatakeyama, I.; Endoh, T.; Yamaguchi, T.; Shuto, S.; Iizuka, H.; Maruyama, T.; Watanabe, H.; Hemink, G.; Sakui, K.; Tanaka, T.; Momodomi, M., and Shirota, R. An advanced NAND-structure cell technology for reliable 3.3 V 64 Mb electrically erasable and programmable read only memories (EEPROMs), *Japanese Journal of Applied Physics*, vol. 33, part 1, no. 1B, pp. 524–528, Jan. 1994.

[4] Aritome, S.; Satoh, S.; Maruyama, T.; Watanabe, H.; Shuto, S.; Hemink, G. J.; Shirota, R.; Watanabe, S.; Masuoka, F. A 0.67 μm^2 self-aligned shallow trench isolation cell (SA-STI cell) for 3 V-only 256 Mbit NAND EEPROMs, *Electron Devices Meeting, 1994. IEDM '94. Technical Digest., International*, pp. 61–64, 11–14 Dec. 1994.

[5] Shimizu, K.; Narita, K.; Watanabe, H.; Kamiya, E.; Takeuchi, Y.; Yaegashi, T.; Aritome, S.; Watanabe, T. A novel high-density $5F^2$ NAND STI cell technology suitable for 256 Mbit and 1 Gbit flash memories, *Electron Devices Meeting, 1997. IEDM '97. Technical Digest., International*, pp. 271–274, 7–10 Dec. 1997.

[6] Takeuchi, Y.; Shimizu, K.; Narita, K.; Kamiya, E.; Yaegashi, T.; Amemiya, K.; Aritome, S. A self-aligned STI process integration for low cost and highly reliable 1 Gbit flash memories, *VLSI Technology, 1998. Digest of Technical Papers. 1998 Symposium on*, pp. 102–103, 9–11 June 1998.

[7] Aritome, S. Advanced flash memory technology and trends for file storage application, *Electron Devices Meeting, 2000. IEDM Technical Digest. International*, pp. 763–766, 2000.

[8] Imamiya, K.; Sugiura, Y.; Nakamura, H.; Himeno, T.; Takeuchi, K.; Ikehashi, T.; Kanda, K.; Hosono, K.; Shirota, R.; Aritome, S.; Shimizu, K.; Hatakeyama, K.; Sakui, K. A

130-mm^2, 256-Mbit NAND flash with shallow trench isolation technology, *Solid-State Circuits, IEEE Journal of*, vol. 34, no. 11, pp. 1536–1543, Nov. 1999.

[9] Ichige, M.; Takeuchi, Y.; Sugimae, K.; Sato, A.; Matsui, M.; Kamigaichi, T.; Kutsukake, H.; Ishibashi, Y.; Saito, M.; Mori, S.; Meguro, H.; Miyazaki, S.; Miwa, T.; Takahashi, S.; Iguchi, T.; Kawai, N.; Tamon, S.; Arai, N.; Kamata, H.; Minami, T.; Iizuka, H.; Higashitani, M.; Pham, T.; Hemink, G.; Momodomi, M.; Shirota, R. A novel self-aligned shallow trench isolation cell for 90 nm 4 Gbit NAND flash EEPROMs, *VLSI Technology, 2003. Digest of Technical Papers. 2003 Symposium on*, pp. 89,90, 10–12 June 2003.

[10] Hwang, J.; Seo, J.; Lee, Y.; Park, S.; Leem, J.; Kim, J.; Hong, T.; Jeong, S.; Lee, K.; Heo, H.; Lee, H.; Jang, P.; Park, K.; Lee, M.; Baik, S.; Kim, J.; Kkang, H.; Jang, M.; Lee, J.; Cho, G.; Lee, J.; Lee, B.; Jang, H.; Park, S.; Kim, J.; Lee, S.; Aritome, S.; Hong, S., and Park, S. A middle-1X nm NAND flash memory cell (M1X-NAND) with highly manufacturable integration technologies, *Electron Devices Meeting (IEDM), 2011 IEEE International*, pp. 199–202, Dec. 2011.

[11] Aritome, S.; Kikkawa, T. Scaling challenge of self-aligned STI cell (SA-STI cell) for NAND flash memories, *Solid-State Electronics*, vol. 82, 54–62, 2013.

[12] Lee, J.-D.; Hur, S.-H.; Choi, J.-D. Effects of floating-gate interference on NAND flash memory cell operation, *Electron Device Letters, IEEE*, vol. 23, no. 5, pp. 264–266, May 2002.

[13] Compagnoni, C. M.; Spinelli, A. S.; Gusmeroli, R.; Lacaita, A. L.; Beltrami, S.; Ghetti, A.; Visconti, A. First evidence for injection statistics accuracy limitations in NAND flash constant-current Fowler–Nordheim programming, *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, pp. 165–168, 10–12 Dec. 2007.

[14] Compagnoni, C. M.; Spinelli, A. S.; Gusmeroli, R.; Beltrami, S.; Ghetti, A., and Visconti, A. Ultimate accuracy for the NAND Flash program algorithm due to the electron injection statistics, *IEEE Trans. Electron Devices*, vol. 55, no. 10, pp. 2695–2702, Oct. 2008.

[15] Compagnoni, C. M.; Gusmeroli, R.; Spinelli, A. S.; Visconti, A. Analytical model for the electron-injection statistics during programming of nanoscale NAND flash memories, *Electron Devices, IEEE Transactions on*, vol. 55, no. 11, pp. 3192–3199, 2008.

[16] Kurata, H.; Otsuga, K.; Kotabe, A.; Kajiyama, S.; Osabe, T.; Sasago, Y.; Narumi, S.; Tokami, K.; Kamohara, S.; Tsuchiya, O. The impact of random telegraph signals on the scaling of multilevel flash memories, *VLSI Circuits, 2006. Digest of Technical Papers. 2006 Symposium on*, pp. 112–113.

[17] Govoreanu, B.; Brunco, D. P.; Van Houdt, J. Scaling down the interpoly dielectric for next generation flash memory: Challenges and opportunities, *Solid-State Electronics*, vol. 49, no. 11, pp. 1841–1848, Nov. 2005.

[18] Kim, Y. S.; Lee, D. J.; Lee, C. K.; Choi, H. K.; Kim, S. S.; Song, J. H.; Song, D. H.; Choi, J.-H.; Suh, K.-D.; Chung, C. New scaling limitation of the floating gate cell in NAND flash memory, *Reliability Physics Symposium (IRPS), 2010 IEEE International*, pp. 599–603, 2–6 May 2010.

[19] Molas, G.; Deleruyelle, D.; De Salvo, B.; Ghibaudo, G.; GelyGely, M.; Perniola, L.; Lafond, D.; Deleonibus, S. Degradation of floating-gate memory reliability by few electron phenomena, *Electron Devices, IEEE Transactions on*, vol. 53, no. 10, pp. 2610–2619, Oct. 2006.

[20] Kinam, K.; Jeong, G. Memory technologies in the nano-era: Challenges and opportunities, *Solid-State Circuits Conference, 2005. Digest of Technical Papers. ISSCC. 2005 IEEE International*, vol. 1, pp. 576, 618, 10–10 Feb. 2005.

[21] Kim, K. Technology for sub-50nm DRAM and NAND flash manufacturing, *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, pp. 323, 326, 5–5 Dec. 2005.

[22] Kim, K.; Choi, J. Future outlook of NAND flash technology for 40 nm node and beyond, *Non-Volatile Semiconductor Memory Workshop, 2006. IEEE NVSMW 2006. 21st*, pp. 9, 11, 12–16 Feb. 2006.

[23] Prall, K. Scaling non-volatile memory below 30 nm, *Non-Volatile Semiconductor Memory Workshop, 2007 22nd IEEE*, pp. 5, 10, 26–30 Aug. 2007.

[24] Kim, K.; Jeong, G. Memory Technologies for sub-40nm Node, *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, pp. 27, 30, 10–12 Dec. 2007.

[25] Parat, K. Recent developments in NAND flash scaling, *VLSI Technology, Systems, and Applications, 2009. VLSI-TSA '09. International Symposium on*, pp. 101, 102, 27–29 April 2009.

[26] Kim, K. Technology challenges for deep-nano semiconductor, *Memory Workshop (IMW), 2010 IEEE International*, pp. 1, 2, 16–19 May 2010.

[27] Kim, K. From the future Si technology perspective: Challenges and opportunities, *Electron Devices Meeting (IEDM), 2010 IEEE International*, pp. 1.1.1, 1.1.9, 6–8 Dec. 2010.

[28] Goda, A.; Parat, K. Scaling directions for 2D and 3D NAND cells, *Electron Devices Meeting (IEDM), 2012 IEEE International*, pp. 2.1.1, 2.1.4, 10–13 Dec. 2012.

[29] Goda, A. Opportunities and challenges of 3D NAND scaling, *VLSI Technology, Systems, and Applications (VLSI-TSA), 2013 International Symposium on*, pp. 1, 2, 22–24 Apr. 2013.

[30] Goda, A. Recent progress and future directions in NAND Flash scaling, *Non-Volatile Memory Technology Symposium (NVMTS), 2013 13th*, pp. 1, 4, 12–14 Aug. 2013.

[31] Park, Y.; Lee, J. Device considerations of planar NAND flash memory for extending towards sub-20 nm regime, *Memory Workshop (IMW), 2013 5th IEEE International*, pp. 1, 4, 26–29 May 2013.

[32] Park, Y.; Lee, J.; Cho, S. S.; Jin, G.; Jung, E. S. Scaling and reliability of NAND flash devices, *Reliability Physics Symposium, 2014 IEEE International*, pp. 2E.1.1, 2E.1.4, 1–5 June 2014.

[33] Aritome, S. 3D flash memories, International Memory Workshop 2011 (IMW 2011), short course.

[34] Prall, K.; Parat, K. 25 nm 64 Gb MLC NAND technology and scaling challenges invited paper, *Electron Devices Meeting (IEDM), 2010 IEEE International*, pp. 5.2.1–5.2.4, 6–8 Dec. 2010.

[35] Seokkiu, L. Scaling challenges in NAND flash device toward 10 nm technology, *Memory Workshop (IMW), 2012 4th IEEE International*, pp. 1–4, 20–23 May 2012.

[36] Suh, K.-D.; Suh, B.-H.; Lim, Y.-H.; Kim, J.-K.; Choi, Y.-J.; Koh, Y.-N.; Lee, S.-S.; Kwon, S.-C.; Choi, B.-S.; Yum, J.-S.; Choi, J.-H.; Kim, J.-R.; Lim, H.-K. A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme, *Solid-State Circuits, IEEE Journal of*, vol. 30, no. 11, pp. 1149–1156, Nov. 1995.

[37] Hemink, G. J.; Tanaka, T.; Endoh, T.; Aritome, S.; Shirota, R. Fast and accurate programming method for multi-level NAND EEPROMs. *VLSI Technology, 1995. Digest of Technical Papers. 1995 Symposium on*, pp. 129–130, 6–8 June 1995.

[38] Shirota, R.; Sakamoto, Y.; Hsueh, H.-M.; Jaw, J.-M.; Chao, W.-C.; Chao, C.-M.; Yang, S.-F.; Arakawa, H. Analysis of the correlation between the programmed threshold-voltage distribution spread of NAND flash memory devices and floating-gate impurity concentration, *Electron Devices, IEEE Transactions on*, vol. 58, no. 11, pp. 3712–3719, Nov. 2011.

[39] Ghetti, A.; Compagnoni, C. M.; Spinelli, A. S.; Visconti, A. Comprehensive analysis of random telegraph noise instability and its scaling in deca–nanometer flash memories, *Electron Devices, IEEE Transactions on*, vol. 56, no. 8, pp. 1746–1752, Aug. 2009.

[40] Shibata, N.; Tanaka, T. US Patent 7,245,528. 7,370,009. 7,738,302.

[41] Park, K.-T.; Kang, M.; Kim, D.; Hwang, S.-W.; Choi, B. Y.; Lee, Y.-T.; Kim, C.; Kim, K. A zeroing cell-to-cell interference page architecture with temporary LSB storing and parallel MSB program scheme for MLC NAND flash memories, *Solid-State Circuits, IEEE Journal of*, vol. 43, no. 4, pp. 919–928, April 2008.

[42] Cernea, R.-A.; Pham, L.; Moogat, F.; Chan, S.; Le, B.; Li, Y.; Tsao, S.; Tseng, T.-Y.; Nguyen, K.; Li, J.; Hu, J.; Yuh, J. H.; Hsu, C.; Zang, F.; Kamei, T.; Nasu, H.; Kliza, P.; Htoo, K.; Lutze, J.; Dong, Y.; Higashitani, M.; Yang, J.; Lin, H.-S.; Sakhamuri, V.; Li, A.; Pan, F.; Yadala, S.; Taigor, S.; Pradhan, K.; Lan, J.; Chan, J.; Abe, T.; Fukuda, Y.; Mukai, H.; Kawakami, K.; Liang, C.; Ip, T.; Chang, S.-F.; Lakshmipathi, J.; Huynh, S.; Pantelakis, D.; Mofidi, M.; Quader, K. A 34 MB/s MLC write throughput 16 Gb NAND with all bit line architecture on 56 nm technology, *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 1, pp. 186–194, Jan. 2009.

[43] Tanaka, T.; Tanaka, Y.; Nakamura, H.; Oodaira, H.; Aritome, S.; Shirota, R.; Masuoka, F. A quick intelligent program architecture for 3 V-only NAND-EEPROMs, *VLSI Circuits, 1992. Digest of Technical Papers, 1992 Symposium on*, pp. 20–21, 4–6 June 1992.

[44] Kim, T.-Y.; Lee, S.-D.; Park, J.-S.; Cho, H.-Y.; You, B.-S.; Baek, K.-H.; Lee, J.-H.; Yang, C.-W.; Yun, M.; Kim, M.-S.; Kim, J.-W.; Jang, E.-S.; Chung, H.; Lim, S.-O.; Han, B.-S.; Koh, Y.-H. A 32 Gb MLC NAND flash memory with V_{th} margin-expanding schemes in 26 nm CMOS, *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, pp. 202–204, 20–24 Feb. 2011.

[45] Kanda, K.; Shibata, N.; Hisada, T.; Isobe, K.; Sato, M.; Shimizu, Y.; Shimizu, T.; Sugimoto, T.; Kobayashi, T.; Kanagawa, N.; Kajitani, Y.; Ogawa, T.; Iwasa, K.; Kojima, M.; Suzuki, T.; Suzuki, Y.; Sakai, S.; Fujimura, T.; Utsunomiya, Y.; Hashimoto, T.; Kobayashi, N.; Matsumoto, Y.; Inoue, S.; Suzuki, Y.; Honda, Y.; Kato, Y.; Zaitsu, S.; Chibvongodze, H.; Watanabe, M.; Ding, H.; Ookuma, N.; Yamashita, R. A 19 nm 112.8 mm^2 64 Gb multi-level flash memory with 400 Mbit/sec/pin 1.8 V toggle mode interface, *Solid-State Circuits, IEEE Journal of*, vol. 48, no. 1, pp. 159–167, Jan. 2013.

[46] Lee, D.; Chang, I. J.; Yoon, S.-Y.; Jang, J.; Jang, D.-S.; Hahn, W.-G.; Park, J.-Y.; Kim, D.-G.; Yoon, C.; Lim, B.-S.; Min, B.-J.; Yun, S.-W.; Lee, J.-S.; Park, I.-H.; Kim, K.-R.; Yun, J.-Y.; Kim, Y.; Cho, Y.-S.; Kang, K.-M.; Joo, S.-H.; Chun, J.-Y.; Im, J.-N.; Kwon, S.; Ham, S.; Ansoo, P.; Yu, J.-D.; Lee, N.-H.; Lee, T.-S.; Kim, M.; Kim, H.; Song, K.-W.; Jeon, B.-G.; Choi, K.; Han, J.-M.; Kyung, K. H.; Lim, Y.-H.; Jun, Y.-H. A 64 Gb 533 Mb/s DDR interface MLC NAND flash in sub-20 nm technology, *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, pp. 430–432, 19–23 Feb. 2012.

[47] Kang, D.; Jang, S.; Lee, K.; Kim, J.; Kwon, H.; Lee, W.; Park, B. G.; Lee, J. D.; Shin, H. Improving the cell characteristics using low-*k* gate spacer in 1 Gb NAND flash memory, *Electron Devices Meeting, 2006. IEDM '06. International*, pp. 1–4, 11–13 Dec. 2006.

[48] Kim, S.; Cho, W.; Kim, J.; Lee, B.; Park, S. Air-gap application and simulation results for low capacitance in 60 nm NAND flash memory, *Non-Volatile Semiconductor Memory Workshop, 2007 22nd IEEE*, pp. 54–55, 26–30 Aug. 2007.

[49] Seo, J.; Han, K.; Youn, T.; Heo, H.-E.; Jang, S.; Kim, J.; Yoo, H.; Hwang, J.; Yang, C.; Lee, H.; Kim, B.; Choi, E.; Noh, K.; Lee, B.; Lee, B.; Chang, H.; Park, S.; Ahn, K.; Lee, S.; Kim, J.; Lee, S. Highly reliable M1X MLC NAND flash memory cell with novel active air-gap and p+ poly process integration technologies, *Electron Devices Meeting (IEDM), 2013 IEEE International*, pp. 3.6.1, 3.6.4, 9–11 Dec. 2013.

[50] Park, M.; Kim, K.; Park, J.-H.; Choi, J.-H. Direct field effect of neighboring cell transistor on cell-to-cell interference of NAND flash cell arrays, *Electron Device Letters, IEEE*, vol. 30, no. 2, pp. 174–177, Feb. 2009.

[51] Park, M.; Suh, K.; Kim, K.; Hur, S.; Kim, K., and Lee, W.; The effect of trapped charge distributions on data retention characteristics of NAND flash memory cells, *IEEE Electron Device Letters*, vol. 28, no. 8, pp. 750–752, Aug. 2007.

[52] Aritome, S.; Seo, S.; Kim, H.-S.; Park, S.-K.; Lee, S.-K.; Hong, S. Novel negative V_t shift phenomenon of program–inhibit cell in 2*X*–3*X*-nm self-aligned STI NAND flash memory, *Electron Devices, IEEE Transactions on*, vol. 59, no. 11, pp. 2950, 2955, Nov. 2012.

[53] Cho, B.; Lee, C. H.; Seol, K.; Hur, S.; Choi, J.; Choi, J.; Chung, C. A new cell-to-cell interference induced by conduction band distortion near S/D region in scaled NAND flash memories, *Memory Workshop (IMW), 2011 3rd IEEE International*, pp. 1, 4, 22–25 May 2011.

[54] Park, M.; Choi, J.-D.; Hur, S.-H.; Park, J.-H.; Lee, J.-H.; Park, J.-T.; Sel, J.-S.; Kim, J.-W.; Song, S.-B.; Lee, J.-Y.; Lee, J.-H.; Son, S.-J.; Kim, Y.-S.; Chai, S.-J.; Kim, K.-T.; Kim, K. Effect of low-*k* dielectric material on 63 nm MLC (multi-level cell) NAND flash cell arrays, *VLSI Technology, 2005. (VLSI-TSA-Tech). 2005 IEEE VLSI-TSA International Symposium on*, pp. 37–38, 25–27 April 2005.

[55] Kang, D.; Shin, H.; Chang, S.; An, J.; Lee, K.; Kim, J.; Jeong, E.; Kwon, H.; Lee, E.; Seo, S.; Lee, W. The air spacer technology for improving the cell distribution in 1 giga bit NAND flash memory, *Non-Volatile Semiconductor Memory Workshop, 2006. IEEE NVSMW 2006. 21st*, pp. 36–37, 12–16 Feb. 2006.

[56] Lee, C.; Hwang, J.; Fayrushin, A.; Kim, H.; Son, B.; Park, Y.; Jin, G.; Jung, E. S. Channel coupling phenomenon as scaling barrier of NAND flash memory beyond 20 nm node, *Memory Workshop (IMW), 2013 5th IEEE International*, pp. 72, 75, 26–29 May 2013.

[57] Molas, G.; Deleruyelle, D.; De Salvo, B.; Ghibaudo, G.; Gely, M.; Jacob, S.; Lafond, D.; Deleonibus, S. Impact of few electron phenomena on floating-gate memory reliability, *Electron Devices Meeting, 2004. IEDM Technical Digest. IEEE International*, pp. 877–880, 13–15 Dec. 2004.

[58] Suh, K.-D.; Suh, B.-H.; Um, Y.-H.; Kim, J.-K.; Choi, Y.-J.; Koh, Y.-N.; Lee, S.-S.; Kwon, S.-C.; Choi, B.-S.; Yum, J.-S.; Choi, J.-H.; Kim, J.-R.; Lim, H.-K. A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme, *Solid-State Circuits Conference, 1995. Digest of Technical Papers. 42nd ISSCC, 1995 IEEE International*, pp. 128–129, 350, 15–17 Feb. 1995.

[59] Kolodny, A.; Nieh, S. T. K.; Eitan, B.; Shappir, J. Analysis and modeling of floating-gate EEPROM cells, *Electron Devices, IEEE Transactions on*, vol. 33, no. 6, pp. 835–844, June 1986.

[60] Compagnoni, C. M.; Gusmeroli, R.; Spinelli, A. S.; Visconti, A. RTN V_T instability from the stationary trap-filling condition: An analytical spectroscopic investigation, *Electron Devices, IEEE Transactions on*, vol. 55, no. 2, pp. 655–661, 2008.

[61] Kirton, M. J., et al., *Advances in Physics*, vol. 38, no. 4, pp. 367–468, 1989.

[62] Roux dit Buisson, O.; Ghibaudo, G., and Brini, J. Model for drain current RTS amplitude in small-area MOS transistors, *Solid-State Electronics*, vol. 35, no. 9, pp. 1273–1276, Sept. 1992.

[63] Tega, N.; Miki, H.; Osabe, T.; Kotabe, A.; Otsuga, K.; Kurata, H.; Kamohara, S.; Tokami, K.; Ikeda, Y.; Yamada, R. Anomalously large threshold voltage fluctuation by complex random telegraph signal in floating gate flash memory, *Electron Devices Meeting, 2006. IEDM '06. International*, pp. 491–494, 11–13 Dec. 2006.

[64] Tanaka, T.; Momodomi, M.; Iwata, Y.; Tanaka, Y.; Oodaira, H.; Itoh, Y.; Shirota, R.; Ohuchi, K.; Masuoka, F. A 4-Mbit NAND-EEPROM with tight programmed V_t distribution, *VLSI Circuits, 1990. Digest of Technical Papers, 1990 Symposium on*, pp. 105–106, 7–9 June 1990.

[65] Gusmeroli, R.; Compagnoni, C. M.; Riva, A.; Spinelli, A. S.; Lacaita, A. L.; Bonanomi, M.; Visconti, A.; Defects spectroscopy in SiO2 by statistical random telegraph noise analysis, *Electron Devices Meeting, 2006. IEDM '06. International*, pp. 483–486, 2006.

[66] Compagnoni, M. C.; Gusmeroli, R.; Spinelli, A. S.; Lacaita, A. L.; Bonanomi, M.; Visconti, A. Statistical model for random telegraph noise in flash memories, *Electron Devices, IEEE Transactions on*, vol. 55, no. 1, pp. 388–395, Jan. 2008.

[67] Ralls, K. S.; Skocpol, W. J.; Jackel, L. D.; Howard, R. E.; Fetter, L. A.; Epworth, R. W.; Tennant, D. M. Discrete resistance switching in submicrometer silicon inversion layers: Individual interface traps and low-frequency (1f?) noise, *Physical Review Letters*, vol. 52, no. 3, pp. 228–231, 1984.

[68] Compagnoni, C. M.; Gusmeroli, R.; Spinelli, A. S.; Lacaita, A. L.; Bonanomi, M.; Visconti, A. Statistical investigation of random telegraph noise ID instabilities in flash cells at different initial trap-filling conditions, *Reliability physics symposium, 2007. proceedings. 45th annual. IEEE international, 2007*, pp. 161–166.

[69] Fukuda, K.; Shimizu, Y.; Amemiya, K.; Kamoshida, M.; Hu, C. Random telegraph noise in Flash memories—Model and technology scaling, *IEDM Technology Digest*, pp. 169–172, 2007.

[70] Ghetti, A.; Compagnoni, C. M.; Biancardi, F.; Lacaita, A. L.; Beltrami, S.; Chiavarone, L.; Spinelli, A. S.; Visconti, A. Scaling trends for random telegraph noise in deca-nanometer flash memories, *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, 2008, pp. 835–838.

[71] Ghetti, A.; Bonanomi, M.; Compagnoni, C. M.; Spinelli, A. S.; Lacaita, A. L.; Visconti, A. Physical modeling of single-trap RTS statistical distribution in flash memories, *Reliability Physics Symposium, 2008. IRPS 2008. IEEE International*, 2008, pp. 610–615.

[72] Compagnoni, C. M.; Spinelli, A. S.; Beltrami, S.; Bonanomi, M.; Visconti, A. Cycling effect on the random telegraph noise instabilities of NOR and NAND flash arrays, *Electron Device Letters, IEEE*, vol. 29, no. 8, pp. 941–943, 2008.

[73] Compagnoni, C. M.; Ghidotti, M.; Lacaita, A. L.; Spinelli, A. S.; Visconti, A. Random telegraph noise effect on the programmed threshold-voltage distribution of flash memories, *Electron Device Letters, IEEE*, vol. 30, no. 9, pp. 984–986, 2009.

[74] Kim, T.; He, D.; Porter, R.; Rivers, D.; Kessenich, J.; Goda, A. Comparative study of quick electron detrapping and random telegraph signal and their dependences on random discrete dopant in sub-40-nm NAND flash memory, *Electron Device Letters, IEEE*, vol. 31, no. 2, pp. 153–155, Feb. 2010.

[75] Kim, T.; He, D.; Morinville, K.; Sarpatwari, K.; Millemon, B.; Goda, A.; Kessenich, J. Tunnel oxide nitridation effect on the evolution of V_t instabilities (RTS/QED) and defect characterization for sub-40-nm flash memory, *Electron Device Letters, IEEE*, vol. 32, no. 8, pp. 999, 1001, Aug. 2011.

[76] Kim, T.; Franklin, N.; Srinivasan, C.; Kalavade, P.; Goda, A. Extreme Short-channel effect on RTS and inverse scaling behavior: Source–drain implantation effect in 25-nm NAND flash memory, *Electron Device Letters, IEEE*, vol. 32, no. 9, pp. 1185, 1187, Sept. 2011.

[77] Raghunathan, S.; Krishnamohan, T.; Parat, K.; Saraswat, K. Investigation of ballistic current in scaled floating-gate NAND FLASH and a solution, *Electron Devices Meeting (IEDM), 2009 IEEE International*, pp. 1–4, 7–9 Dec. 2009.

[78] Yano, K.; Ishii, T.; Sano, T.; Mine, T.; Murai, F.; Hashimoto, T.; Kobayashi, T.; Kure, T.; Seki, K. Single-electron memory for giga-to-tera bit storage, *Proceedings of the IEEE*, vol. 87, no. 4, pp. 633–651, April 1999.

[79] Aritome, S.; Kirisawa, R.; Endoh, T.; Nakayama, R.; Shirota, R.; Sakui, K.; Ohuchi, K.; Masuoka, F. Extended data retention characteristics after more than 10^4 write and erase cycles in EEPROMs, *International Reliability Physics Symposium, 1990. 28th Annual Proceedings*, 1990, pp. 259–264.

[80] Kirisawa, R.; Aritome, S.; Nakayama, R.; Endoh, T.; Shirota, R.; Masuoka, F. A NAND structured cell with a new programming technology for highly reliable 5 V-only flash EEPROM, *1990 Symposium on VLSI Technology, 1990. Digest of Technical Papers, 1990*, pp. 129–130.

[81] Aritome, S.; Shirota, R.; Kirisawa, R.; Endoh, T.; Nakayama, R.; Sakui, K.; Masuoka, F. A reliable bi-polarity write/erase technology in flash EEPROMs, *International Electron Devices Meeting, 1990. IEDM '90. Technical Digest., 1990*, pp. 111–114.

[82] Aritome, S.; Shirota, R.; Sakui, K.; Masuoka, F. Data retention characteristics of flash memory cells after write and erase cycling, *IEICE Transactions on Electronics*, vol. E77-C, no. 8, pp. 1287–1295, Aug. 1994.

[83] Aritome, S.; Shirota, R.; Hemink, G.; Endoh, T.; Masuoka, F. Reliability issues of flash memory cells, *Proceedings of the IEEE*, vol. 81, no. 5, pp. 776–788, May 1993.

[84] Shirota, R.; Nakayama, R.; Kirisawa, R.; Momodomi, M.; Sakui, K.; Itoh, Y.; Aritome, S.; Endoh, T.; Hatori, F.; Masuoka, F. A 2.3 μm^2 memory cell structure for 16 Mb NAND EEPROMs, *Electron Devices Meeting, 1990. IEDM '90. Technical Digest, International*, pp. 103–106, 9–12 Dec. 1990.

[85] Lee, C.-H.; Sung, S.-K.; Jang, D.; Lee, S.; Choi, S.; Kim, J.; Park, S.; Song, M.; Baek, H.-C.; Ahn, E.; Shin, J.; Shin, K.; Min, K.; Cho, S.-S.; Kang, C.-J.; Choi, J.; Kim, K.; Choi, J.-H.; Suh, K.-D.; Jung, T.-S. A highly manufacturable integration technology for 27 nm 2 and 3 bit/cell NAND flash memory, *Electron Devices Meeting (IEDM), 2010 IEEE International*, pp. 5.1.1, 5.1.4, 6–8 Dec. 2010.

[86] Asenov, A.; Kaya, S.; Brown, A. R. Intrinsic parameter fluctuations in decananometer MOSFETs introduced by gate line edge roughness, *Electron Devices, IEEE Transactions on*, vol. 50, no. 5, pp. 1254–1260, May 2003.

[87] Asenov, A.; Brown, A. R.; Davies, J. H.; Kaya, S.; Slavcheva, G. Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs, *Electron Devices, IEEE Transactions on*, vol. 50, no. 9, pp. 1837–1852, Sept. 2003.

[88] Spessot, A.; Calderoni, A.; Fantini, P.; Spinelli, A. S.; Compagnoni, C. M.; Farina, F.; Lacaita, A. L.; Marmiroli, A. Variability effects on the VT distribution of nanoscale NAND flash memories, *Reliability Physics Symposium (IRPS), 2010 IEEE International*, pp. 970–974, 2–6 May 2010.

[89] Spessot, A. M.; Compagnoni, C. M.; Farina, F.; Calderoni, A.; Spinelli, A. S.; Fantini, P. Compact modeling of variability effects in nanoscale NAND flash memories, *Electron Devices, IEEE Transactions on*, vol. 58, no. 8, pp. 2302, 2309, Aug. 2011.

[90] Larcher, L.; Padovani, A.; Pavan, P.; Fantini, P.; Calderoni, A.; Mauri, A.; Benvenuti, A. Modeling NAND Flash memories for IC design, *IEEE Electron Device Letters*, vol. 29, pp. 1152–1154, Oct. 2008.

[91] Miccoli, C.; Compagnoni, C. M.; Amoroso, S. M.; Spessot, A.; Fantini, P.; Visconti, A., and Spinelli, A. S. Impact of neutral threshold voltage spread and electron-emission statistics on data retention of nanoscale NAND flash, *IEEE Electron Device Letters*, vol. 31, no. 11, pp. 1202–1204, Nov. 2010.

[92] Mouli, C.; Prall, K.; Roberts, C. Trend in memory technlogy—reliability perspectives, challenges and opportunities, *Proceedings of 14th IPFA* 2007, pp. 130–134.

[93] Compagnoni, C. M.; Ghetti, A.; Ghidotti, M.; Spinelli, A. S.; Visconti, A. Data retention and program/erase sensitivity to the array background pattern in deca-nanometer NAND flash memories, *IEEE Transactions on Electron Devices*, vol. 57, no. 1, pp. 321–327, 2010.

[94] Park, B.; Cho, S.; Park, M.; Park, S.; Lee, Y.; Cho, M. K.; Ahn, K.-O.; Bae, G.; Park, S. Challenges and limitations of NAND flash memory devices based on floating gates, *Circuits and Systems (ISCAS), 2012 IEEE International Symposium on*, pp. 420, 423, 20–23 May 2012.

[95] Kim, C.; Ryu, J.; Lee, T.; Kim, H.; Lim, J.; Jeong, J.; Seo, S.; Jeon, H.; Kim, B.; Lee, I. Y.; Lee, D. S.; Kwak, P. S.; Cho, S.; Yim, Y.; Cho, C.; Jeong, W.; Park, K.; Han, J.-M.; Song, D.; Kyung, K.; Lim, Y.-H.; Jun, Y.-H. A 21 nm high performance 64 Gb MLC NAND flash memory with 400 MB/s asynchronous toggle DDR interface, *Solid-State Circuits, IEEE Journal of*, vol. 47, no. 4, pp. 981, 989, April 2012.