首都大学東京　博士（工学）学位論文（課程博士）

論 文 名　　半導体メモリの記憶電荷漏えい機構に関する研究

著　者　　蒲原　史朗

審査担当者

主　査

委　員

委　員

委　員

上記の論文を合格と判定する

平成　　年　　月　　日

首都大学東京大学院理工学研究科教授会

研究科長

DISSERTATION FOR A DEGREE OF
DOCTOR OF PHILOSOPHY IN ENGINEERING
TOKYO METROPOLITAN UNIVERSITY

TITLE :
Mechanism of Data Retention on Semiconductor Memories

AUTHOR : Shiro Kamohara

EXAMINED BY

Examiner in chief

Examiner

Examiner

Examiner

QUALIFIED BY THE GRADUATE SCHOOL
OF SCIENCE AND ENGINEERING
TOKYO METROPOLITAN UNIVERSITY

Dean

Date

Mechanism of Data Retention on Semiconductor Memories

Shiro Kamohara, 2008

Abstract

Semiconductor memories can be divided into two main categories: random-access memories (RAMs) and read-only memories (ROMs). Once the power supply is switched off, the former loses stored information and the latter keeps stored information and for that reason RAMs are categorized to volatile and non-volatile memories, respectively. DRAMs have been used as main memories for personal computers. SRAMs have been used as the cash memories for a variety of applications. DRAMs represent the main portion of the memory market, around 25000M$/year. Market of SRAMs is about one tenth for those of DRAMs. Nonvolatile memory are categorized to mask read-only memories (MROMs), electrically programmable read-only memories but erasable via ultraviolet (UV-EPROMs), electrically erasable and programmable read-only memories (EEPROMs) and Flash EEPROMs (flash memories). The market of flash memories are around 15000M$/year. The market share of DRAMs and flash memories is about 90% among all kinds of memory devices.

Both DRAMs and flash memories store electronic charges as the information. Therefore, one of the most critical reliability issues is the data retention characteristics attributed to the electron leakage. Mechanisms of the electron leakage for both DRAMs and flash memories, however, have not been clarified to date. For manufacturing of reliable semiconductor memories of high quality, we should measure, analyze, improve and control over the data retention characteristics. Under the current circumstances, we must repeat to measure, analyze and improve the data retention for every product generation by using the conventional ways. For ultimate control over the data retention characteristics, we must clear up origins of the data retention characteristics and find out the methods of the ultimate control over the data retention characteristics.

The purpose of present research is to understand mechanisms and origins of the data retention characteristics and to propose methodologies of ultimate control over the data retention. To understand the origins of the data retention, simple analytical models have been proposed based on mechanisms of the data retention characteristics. By comparisons between models and experimental results, physical parameters associated with the origins of the data retention characteristics are extracted. Extracted physical parameters are assessed and the origins of the data leakage are reckoned.

From Chapter 2 to Chapter 4, the data retention characteristics of flash memories are discussed. In Chapter 2, the structure, operation and reliability of flash memories are briefly described. As the nonvolatile memory, the data retention characteristics are one of the most important reliability issues for flash memories. The data retention

characteristics of flash memories are determined by two mechanisms, i.e., detrapping and stress induced leakage currents (SILCs). During the program/erase cycles, electrons are trapped inside the tunnel oxide, and the multiple-trap-paths between the floating gate and the silicon substrate are generated. The threshold voltage shift via the emission of the trapped electron is called as the data retention characteristics via the detrapping. In Chapter 3, the mechanism of the data retention characteristics via detrapping is discussed. The detrapping can be described by an analogous model equation with that of the chemisorptions process. The electron flow out via the multiple-trap-path is called as the data retention characteristics via SILCs. In Chapter 4, the mechanism of the data retention characteristics via SILCs is discussed. The data retention characteristics via SILCs can be described by the model equation of B-mode SILCs. In Chpater 3 and Chapter 4, the origins of electron traps and multiple-trap-paths are specified.

From Chapter 5 to Chapter 6, the data retention characteristics of DRAMs are discussed. In Chapter 5, the structure, operation and reliability of DRAMs are briefly described. The electrons stored to the capacitors flow out via some leakage mechanisms. By refresh operations, the voltages of all storage capacitances connected to the specific word line are read out, amplified, and reprogrammed, i.e., refreshed. Therefore, the voltages of all storage capacitances are refreshed even if the voltages are modulated by the leakage currents. Performances of the refresh operations govern the speed and the power consumption of DRAMs. The performance of the refresh operation is determined by *tail bits* which show the anomalous leakage characteristics among all bits in a DRAM chips. To attain the specification of the refresh operation, the retention characteristics of *tail bits* should be improved. In Chapter 6, the mechanisms of the anomalous leakage currents of *tail bits* are discussed. To derive model equations, we assume that some bits containing one specific trap become *tail bits* among all bits in a DRAM chip. The variation of the leakage current of *tail bits* is attributed to the fluctuation of the trap level.

In Chapter 7, origins of the data retention characteristics for both flash memories and DRAMs are discussed. For flash memories, oxygen vacancies play important roles to the data retention characteristics. The oxygen vacancies of the dimer configurations and the four-folded configurations are thought to be origins of the detrapping and SILCs, respectively. The metal contaminations are thought to be the origin of the anomalous leakage of *tail bits*. In Chapter 8, prospects for conventional and emerging memories are discussed. The data retention characteristics will become the roadblock for the scale down of conventional memories. Therefore, new memories of the retention-problem-free have been proposed, i.e., Ferroelectric RAMs (FeRAMs),

Magnetoresistive RAMs (MRAMs), and Phase change RAMs (PRAMs). Prospects of these emerging memories are briefly discussed in this chapter.

Finally, in Chapter 9, conclusions of this thesis are described. The retention mechanisms of flash memories via detrapping and via SILCs are attributed to the oxide traps of 0.37 [eV] and 3.6 [eV], respectively. The oxygen vacancies of the dimer configurations and the four-folded configurations are thought to be origins of the oxide trap of 0.37 [eV] and 3.6 [eV], respectively. Oxygen vacancies basically exist in the oxide and are hardly removed. The retention mechanism of DRAMs is attributed to the silicon trap of 0.68 [eV]. Origin of this trap is thought to be the metal contamination of 0.01 [ppb]. This level of contamination is beyond standard purity for semiconductor wafer. As a result, we can not remove these fundamental origins of charge loss for both flash memories and DRAMs. Therefore, we must continuously rely on the conventional methodology to modify the data retention for every successive generation, i.e., the reduction of the junction field and the increase of the repairable bits for DRAMs, and the constant thickness of oxide thickness, the increases of the repairable bits and even relax of the specification of program/erase number for flash memories. Shrink of devices make it difficult to satisfy the specifications of memories by only relying on the conventional methodologies. Therefore, development of new semiconductor memories of the retention-problem-free, i.e., FeRAMs, MRAMs and PRAMs, are crucial in the near future.

The detail summaries of each chapter are as follows:

The brief review of the semiconductor memories, the purpose of the present research and the outline of the thesis are described in Chapter 1. The purpose of the present research is to understand the mechanisms and origins of the data retention characteristics and to propose the methodologies of the ultimate control over the data retention for flash memories and DRAMs. To understand the origins of the data retention characteristics, simple analytical models have been proposed based on the data retention mechanisms. By comparisons between models and experimental results, physical parameters associated with origins of the data retention characteristics are extracted. Extracted physical parameters are assessed by using the previous works and origins of the data retention are reckoned.

Chapter 2 briefly summarizes the structures, operations, and reliability issues of flash memories. After quick review of the history for flash memories, the cell structures

and their operations, i.e., program, erase and read, are explained, including NOR-type, NAND-type and AND-type flash memories. For flash memories, the most critical reliability issue is the data retention. The data retention characteristics are determined by two mechanisms, i.e., detrapping and stress induced leakage currents (SILCs). In this thesis, origins of detrapping and SILCs have been studied. To understand the previous works and the present work, the direct tunneling, the Fowler-Nordheim (FN) tunneling, and the Poole-Frenkel (PF) emission are described. To understand the carrier emission from the multiple traps in the insulator, the tunnel front model and thermal (Poole-Frenkel) emission front model are also described.

The data retention characteristics of flash memories via detrapping are described in Chapter 3. Electron detrapping is one of the main causes of data retention in the state-of-the-art flash EEPROM. The log (*t*) dependence of ΔVth is a unique aspect of data retention characteristics via electron detrapping. To explain log (*t*) dependence, we have assumed that after electron detrapping, the positive-ionized trap reduces the probability of the electrons in the influence area being emitted from their site. Based on this assumption, we have developed a model for detrapping that is consistent with the experimental results.

Chapter 4 discusses the data retention characteristics of flash memories via stress-induced leakage currents (SILCs). A model of the stress-induced leakage currents (SILCs) based on the inelastic trap-assisted tunneling (ITAT) is developed by introducing a trap with a deep energy level of 3.6eV from the bottom of the conduction band. This model can explain both of two field dependencies, i.e., a field dependence of the direct tunneling (DT) for A-mode SILC and that of the Fowler-Nordheim (FN) tunneling for B-mode SILC by analytical equations of a common form. For simple analytical equations, we introduce the most favorable trap position (MFTP), which gives the largest contribution to the leakage current. The trap area density for A-mode SILC of around 1×10^{10} cm^{-2} and the are density of the leakage paths for B-mode SILC of 5×10^{2} cm^{-2} were obtained by comparisons with the experimental results and the present model.

Chapter 5 briefly summarizes structures, operations, and reliability issues of DRAMs. After brief review of DRAMs, the high density memory cell structures, i.e., a stacked capacitor cell (STC cell) and a trench capacitor cell (trench cell), are described. Then, the operations of DRAMs are explained, including read, program and refresh

operations. The refresh operation is unique for DRAMs. To reduce the power consumption and enhance the operation speeds, the refresh time should be as long as possible. The refresh time is determined the data retention characteristics of *tail* bits. The origin of the data leakage is p-n junction leakage currents. To understand the previous works and the present work, the generalized Shockley-Read-Hall recombination currents and gate induced drain leakage (GIDL) currents are described.

The data retention mechanism of DRAMs is described in Chapter 6. A new model for the leakage current of a single *tail bit* of DRAMs is developed. This model can explain the leakage current of each *tail bit* quantitatively. To derive model equations, we assume that some bits containing one specific trap become *tail bits* among all bits in a DRAM chip. The variation of the leakage current of *tail bits* is attributed to the fluctuation of the trap level. By introducing the trap level fluctuation model, we have successfully reproduced the distribution of the retention time for *tail bits*. We also have obtained a good agreement between model and experimental results of *tail* distributions as functions of process splits and the temperature by using the present model. As an example applied by the present model, we estimated the required number of the repairable bits for 1G DRAM.

Chapter 7 discusses the origins of leakage currents for flash memories and DRAMs. The charge loss of flash memories via detrapping and SILCs are caused by the oxide trap of 0.37 [eV] and 3.6 [eV], respectively. Trap levels of 0.37 [eV] and 3.6 [eV] are thought to originate in oxygen vacancies of dimer and fourfold configurations, respectively. For *tail bits* of DRAMs, the origin of anomalous leakage of *tail bits* is the silicon trap of 0.68 [eV]. By comparison with the previous data of silicon traps by metal contaminations, the elements which show close values of 0.677 [eV] are “Fe” and “Cu”. From the failure rate of DRAM cells, the contamination level of 0.01 [ppb] is obtained, which is below the purity level of silicon substrates.

The prospects for conventional and emerging memories are remarked in Chapter 8. Conventional memories, including NOR and NAND type flash memories, and DRAMs, would be believed that the data retention would be more serious problem for the future applications. Therefore, the new memories of the retention-problem-free are proposed including Ferroelectric RAMs (FeRAMs), Magnetoresistive RAMs (MRAMs) and Phase-Change RAMs (PRAMs). FeRAMs, MRAMs and PRAMs store the information via the polarization of the ferroelectric film, the magnetization of the

magnetic tunnel junction (MTJ), the phases, i.e., crystalline or amorphous, of the Chalcogenide glass, respectively. These emerging memories possess nearly ideal properties, i.e., superior data retention, fast random access, virtually unlimited usage. Excellent functional properties of the new memories offer possibilities to displace the existing memories or create the new types of applications.

Finally, in Chapter 9, conclusions of the thesis are described. The data retention of flash memories can be attributed to the detrapping from the trap with an energy level of 0.37 [eV] as well as the stress induced leakage currents (SILCs) via 3.6 [eV] trap in the oxide layer. The origins of these traps can be attributed to the oxygen vacancies. Oxygen vacancies basically exist in the oxide and are hardly removed. On the other hand, the retention of DRAMs originates 0.68 [eV]-trap in the Si p-n junction. This trap can be ascribed to the contaminated Fe atoms on the concentration order of 0.01 [ppb]. This level of contamination is below the purity of the silicon substrate. As a result, we can not remove these fundamental origins of charge loss for both flash memories and DRAMs. Therefore, development of new semiconductor memories including FeRAMs, MRAMs and PRAMs, are crucial in near future.

Contents

1

Introduction

Summary

In this chapter, the purpose of the present research and the outline of the thesis are described, after brief review of the semiconductor memories. *The purpose of the present research is to understand the mechanisms and origins of the data retentions and to propose the methodologies of ultimate control over the data retention.* To understand the origins of the data retention, simple analytical models have been proposed based on models of data retention mechanisms. By comparisons between models and experiments, physical parameters associated with the origins of the charge loss are extracted. I will identify the origins of the charge loss by assessing the extracted physical parameters. The prospects of DRAM and flash memories from the aspect of the data retention are discussed based on the present studies.

1.1 Semiconductor Memories

Semiconductor devices have a variety of products, including micro-control units (MCUs), micro-processing units (MCPs), system LSIs, analog devices, discrete devices, memories and so on [1-5]. Semiconductor memories can be divided into two main categories: random-access memories (RAMs) and read-only memories (ROMs). Once the power supply is switched off, the former loses stored information and the latter keeps stored information and for that reason RAMs are categorized to volatile and non-volatile memories, respectively.

RAMs are sorted into dynamic random-access memories (DRAMs) and static random-access memories (SRAMs), as shown in Fig. 1 [1]. DRAMs have been used as the main memories for personal computers. SRAMs have been used as the cash memories for a variety of applications. DRAMs of 1K bit memories were commercialized on 1971 by Intel Corp. and were used as the main memory of the computer. DRAMs use the capacitor as the storage node of the data. Because the data stored in the capacitor disappears for a short time, the re-write of the data is required within a certain period, which is called as the refresh action. The memory is always active during the refresh action, so the terminology of “dynamic” RAM is determined. DRAMs of Intel Corp. consists of one capacitor and three transistors. Meanwhile DRAMs of 4K bit was commercialized by TI. DRAMs of TI Corp. consists of one capacitor and one transistor, which becomes the standard structure of DRAMs memory cell. From this time, the memory size of DRAMs has been four times lager on every DRAMs generation. Here one generation of DRAMs is around three years. DRAMs represent the main portion of the memory market, around 25000M$/year, as shown in Fig. 1-2. The market of SRAMs is about one tenth for those of DRAMs.

Nonvolatile memories are categorized to mask read-only memories (MROMs), electrically programmable read-only memories but erasable via ultraviolet (UV-EPROMs), electrically erasable and programmable read-only memories (EEPROMs) and Flash EEPROMs (flash memories), as shown in Fig. 1 [2]. UV-EPROMs and EEPROMs are manufactured for specific applications only, since the use larger areas and, therefore, are more expensive. At the end of the 1980’s, flash memories were supposed to replace EPROM’s rapidly in ever applications. This was not the case, mainly due to reliability problems in the early stage. On the other hand, realization of a new generation of flash memories that can be erased by blocks of different sizes encourages new applications such as data storages, silicon audios and so on. The market share of nonvolatile memories has been continuously growing in the past few years, and a further growth in the near future is foreseen, especially for flash

memories. The market of flash memories is estimated to be around 15000M$/year, as shown in Fig. 1-2. The market share of DRAMs and flash memories is about 90% among all kinds of memory devices, as shown in Fig. 1-2.

Recently, new types of non-volatile memories are proposed to overcome the road stops of the conventional semiconductor memories, i.e., ferroelectric random access memories (FeRAMs), magnetoresistive random access memories (MRAMs) and phase change random access memories (PRAMs). These emerging types of non-volatile memories are discussed in the last section.

Table 1-1 lists equivalent circuits, storage nodes, storage elements, and data retentions of the conventional semiconductor memories. For writing each bit of DRAMs, both the word line (WL) and the bit line (BL) are set to “high”, and the BL information is stored to the capacitor via charges. For writing each bit of SRAMs, WL, BL and $\overline{BL}$ are set to “high”, “high” and “low”, respectively, and the information of BLs is stored to the flip-flop via currents. For writing each bit of flash memories, WL and BL are set to “high” and “low”, respectively. The information of BL is stored to the floating gate via charges. DRAMs and flash memories store charges in capacitors and floating gates, respectively. Actually, the stored charges must disappear in the course of time via some leakage paths or mechanisms. This phenomenon is called as the data retention. The data retention is one of the most important reliability issues of both DRAMs and flash memories as the memory devices.

Some important properties for the conventional semiconductor memories are summarized in Table 1-2, including endurances, write speeds, read speeds, cell sizes, densities, power supply voltages, program voltages, applications and refresh operations. Small cell sizes, high densities, low program voltages are advantages of DRAMs, while the requirement of refresh operations is disadvantage. SRAMs are superior to high read/write speeds and low program voltages. However, big cell sizes and low densities are disadvantage of SRAMs. Flash memories have advantages for small cell sizes and the high densities and, however, are inferior to low endurances and high program voltages. An individual application of the conventional semiconductor memories is determined by taking these advantages and disadvantages into account. DRAMs, SRAMs, flash memories are mainly used as main memories, cash memories, data storages, respectively, for PC, portable cellular, and so on.

Both DRAMs and flash memories store electronic charges as the data. Therefore, one of the most critical reliability issues is the data retention characteristics attributed to electron leakage. The data retentions of DRAMs and flash memories are crucial for applications of huge-data-storage. Recently, applications of these memories are growing

as the embedded memories for the micro control units (MCUs) and the system on chips (SoCs). Almost all semiconductor companies have strong concerns about the data retention of DRAMs and flash memories.

Mechanisms of the data retention of DRAMs and flash memories, however, have not been clarified to date. For manufacturing of reliable semiconductor memories of high quality, we should measure, analyze, improve and control over the data retention characteristics. Under the current circumstances, we must repeat to improve the data retention characteristics for every product generation by using the conventional ways. The conventional methodologies to improve the retention characteristics are empirically obtained. By shrinking of memories, the conventional methodologies have been less efficient. Therefore, new methodologies of the ultimate control over the data retention characteristics are required. For this purpose, the origins of the data retention characteristics should be clarified. However, the data retention mechanisms of DRAMs and flash memories are still under discussions. In the present thesis, I will disclose the origins of the data retention characteristics and discuss the prospective of the conventional memories, i.e, DRAMs and flash memories.

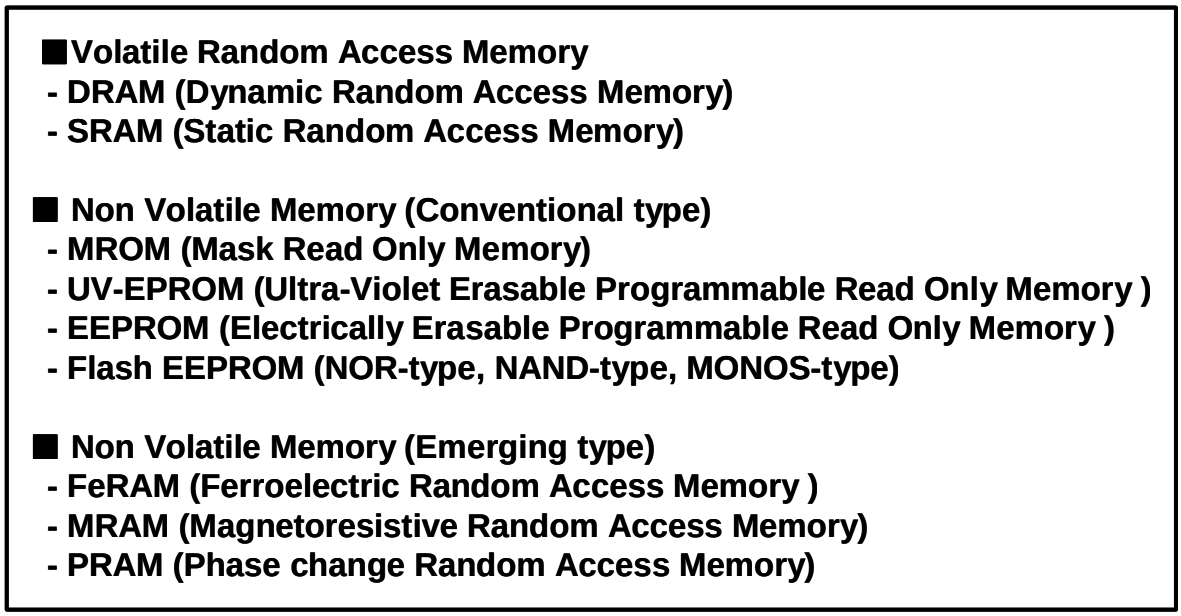


Fig. 1-1 Semiconductor Memories

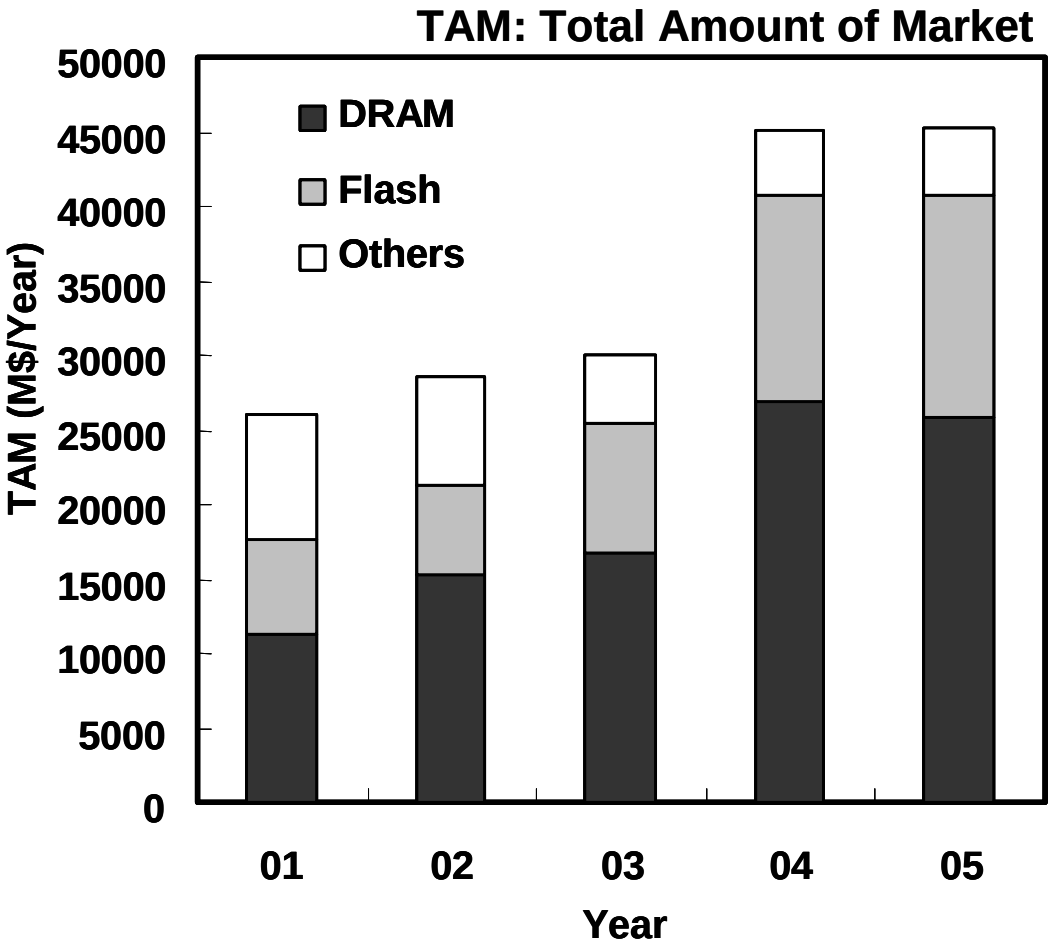


Fig. 1-2 Market of Semiconductor Memories

Table 1-1 Brief summary of conventional semiconductor memories.

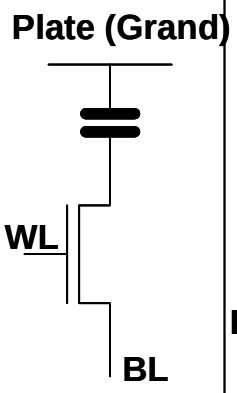
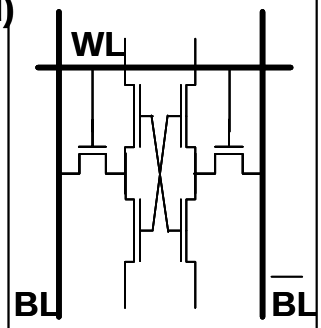
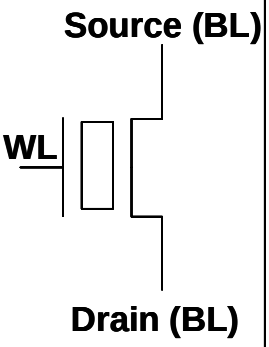
	DRAM	SRAM	Flash
Equivalent Circuit			
Storage Node	Capacitor	Flip-Flop	Floating Gate
Storage Element	Charge	Current	Charge
Data Retention	Big issue	Small issue	Big issue

Table 1-2 Properties of conventional semiconductor memories.

	Volatile RAM		Non Volatile RAM	
	DRAM	SRAM	Flash(NOR)	Flash(NAND)
Endurance	$>10^{15}$	$>10^{15}$	10^5-10^6	10^5-10^6
Write	15n-50n	1n-100n	10μs/B	10MB/s
Read	15n-50n	1n-100n	20n-100n	>10MB/s
Cell Size	$10F^2$	$40F^2$	$12F^2$	$6F^2$
Density	512M/1G	36M/72M	256M/512M	4G/8G
Power supply voltage	1.8V-2.5V	1.2V-1.8V	1.8V-2.7V	1.8V-2.7V
Program voltage	1.8V-2.5V	1.2V-1.8V	10V	18V
Aplication	PC/Cellular	PC/General	Embedded (General)	Memory card
Refresh	Neccesary	Non	Non	Non

1.2 Purpose of research

The purpose of the present research is to understand the mechanisms and the origins of the data retention characteristics and to propose the methodologies of the ultimate control over the data retention of semiconductor memories. The details are summarized as follows:

(1) The data retention characteristics of flash memories are determined by two mechanisms, i.e., the detrapping and stress induced leakage currents (SILCs). The data retention of DRAMs is dominated by the junction leakage of *tail bits*. At the first step, I will specify mechanisms for the detrapping, SILCs, and the junction leakage of *tail bits*.

(2) To verify the present mechanisms, I will derive simple analytical models for the detrapping, SILCs, and the junction leakage of *tail bits*. The present mechanisms will be verified by comparisons between simple analytical models and experimental results.

(3) The origins of the detrapping, SILCs, and the junction leakage of *tail bits* should be specified to ultimate control over the data retention. For this purpose, physical parameters associated with origins of the detrapping, SILCs, and the junction leakage of *tail bits* will be extracted by comparisons between simple analytical models and experimental results. I will identify the origins of the detrapping, SILCs, and the junction leakage of *tail bits* by assessing the extracted physical parameters by referring the previous studies.

(4) Finally, I will discuss prospects of flash memories and DRAMs from the aspects of the data retention based on the present studies. Especially I will discuss the possibility to ultimate control over the origins of the detrapping, SILCs, and the junction leakage of *tail bits*, which will suggest the importance of studies for emerging memories.

1.3 Outline of thesis

This thesis consists of nine chapters. From Chapter 2 to Chapter 4, the data retention of flash memories is discussed. In Chapter 2, the structure, operation and reliability of flash memories are briefly described. As the nonvolatile memory, the data retention is one of the most important reliability issues for flash memories. The data retention characteristics of flash memories are determined of two mechanisms, i.e., detrapping and stress induced leakage currents (SILCs). During the program/erase cycles, electrons are trapped inside the tunnel oxide, and multiple-trap-paths between the floating gate and the silicon substrate are generated. The threshold voltage shift via the emission of the trapped electron is called as the data retention characteristics via the detrapping. In Chapter 3, the mechanism of the data retention characteristics via detrapping is discussed. The electron flow out through the multiple-trap-path is called as the data retention via SILCs. In Chapter 4, the mechanism of the data retention characteristics via SILCs is discussed. In Chpater 3 and Chapter 4, the origin of the electron trap and the multiple-trap-path is specified.

From Chapter 5 to Chapter 6, the data retention characteristics of DRAMs are discussed. In Chapter 5, the structure, operation and reliability of DRAMs are briefly described. Electrons stored to the capacitors flow out via some leakage mechanisms. The refresh operation is required to retain the information. Performances of the refresh operations govern the speed and the power consumption of DRAMs. The performance of the refresh operation is determined by *tail bits* which show the anomalous leakage characteristics among all bits in a DRAM chip. To attain the speculation of the refresh operation, the retention characteristics of *tail bits* should be improved. In Chapter 6, the mechanisms of the data retention are discussed. To derive model equations, I assume that some bits containing one specific trap become *tail bits* among all bits in a DRAM chip. The variation of the leakage current of *tail bits* is attributed to the fluctuation of the trap level.

In Chapter 7, origins of data retentions characteristics for both flash memories and DRAMs are discussed. For flash memories, oxygen vacancies play important roles to the data retention characteristics. The oxygen vacancies of the dimer configurations and the four-folded configurations are the origins of the data retention characteristics via detrapping and SILCs, respectively. In Chapter 8, prospects for conventional and emerging memories are discussed. The data retention characteristics will become the roadblock for scale down of the conventional memories. Therefore, the new memories of retention-problem-free have been proposed, i.e., Ferroelectric RAMs (FeRAMs), Magnetoresistive RAMs (MRAMs), and Phase change RAMs (PRAMs). These

memories are briefly explained in this chapter. Finally, in Chapter 9, conclusions of this thesis are described.

The detail summaries of each chapter are as follows:

The brief review of the semiconductor memories, the purpose of the present research and the outline of the thesis are described in Chapter 1. *The purpose of the present research is to understand the mechanisms and origins of the data retentions and to propose the methodologies of ultimate control over the data retention.* To understand the origins of the data retention, simple analytical models have been proposed based on models of data retention mechanisms. By comparisons between models and experiments, physical parameters associated with the origins of the charge loss are extracted. I will identify the origins of the charge loss by assessing the extracted physical parameters. The prospects of DRAM and flash memories from the aspect of the data retention are discussed based on the present studies.

Chapter 2 briefly summarizes the structures, operations, and reliability issues of flash memories. After quick review of the history for flash memories, the cell structures and their operations, i.e., program, erase and read, are explained, including NOR-type, NAND-type and AND-type flash memories. For flash memories, the most critical reliability issue is the data retention. The data retention characteristics are determined by two mechanisms, i.e., detrapping and stress induced leakage currents (SILCs). In this thesis, origins of detrapping and SILCs have been studied. To understand the previous works and the present work, the direct tunneling, the Fowler-Nordheim (FN) tunneling, and the Poole-Frenkel (PF) emission are described. To understand the carrier emission from the multiple traps in the insulator, the tunnel front model and thermal (Poole-Frenkel) emission front model are also described.

The data retention characteristics of flash memories via detrapping are described in Chapter 3. Electron detrapping is one of the main causes of data retention in the state-of-the-art flash EEPROM. The log (*t*) dependence of ΔVth is a unique aspect of data retention characteristics via electron detrapping. To explain log (*t*) dependence, I have assumed that after electron detrapping, the positive-ionized trap reduces the probability of the electrons in the influence area being emitted from their site. Based on this assumption, I have developed a model for detrapping that is consistent with the experimental results.

Chapter 4 discusses the data retention characteristics of flash memories via stress-induced leakage currents (SILCs). A model of the stress-induced leakage currents (SILCs) based on the inelastic trap-assisted tunneling (ITAT) is developed by introducing a trap with a deep energy level of 3.6eV from the bottom of the conduction band. This model can explain both of two field dependencies, i.e., a field dependence of the direct tunneling (DT) for A-mode SILC and that of the Fowler-Nordheim (FN) tunneling for B-mode SILC by analytical equations of a common form. For simple analytical equations, I introduce the most favorable trap position (MFTP), which gives the largest contribution to the leakage current. The trap area density for A-mode SILC of around 1×10^{10} cm^{-2} and the are density of the leakage paths for B-mode SILC of 5×10^{2} cm^{-2} were obtained by comparisons with the experimental results and the present model.

Chapter 5 briefly summarizes structures, operations, and reliability issues of DRAMs. After brief review of DRAMs, the high density memory cell structures, i.e., a stacked capacitor cell (STC cell) and a trench capacitor cell (trench cell), are described. Then, the operations of DRAMs are explained, including read, program and refresh operations. The refresh operation is unique for DRAMs. To reduce the power consumption and enhance the operation speeds, the refresh time should be as long as possible. The refresh time is determined the data retention characteristics of *tail* bits. The origin of the data leakage is p-n junction leakage currents. To understand the previous works and the present work, the generalized Shockley-Read-Hall recombination currents and gate induced drain leakage (GIDL) currents are described.

The data retention mechanism of DRAMs is described in Chapter 6. A new model for the leakage current of a single *tail bit* of DRAMs is developed. This model can explain the leakage current of each *tail bit* quantitatively. To derive model equations, I assume that some bits containing one specific trap become *tail bits* among all bits in a DRAM chip. The variation of the leakage current of *tail bits* is attributed to the fluctuation of the trap level. By introducing the trap level fluctuation model, I have successfully reproduced the distribution of the retention time for *tail bits*. I also have obtained a good agreement between model and experimental results of *tail* distributions as functions of process splits and the temperature by using the present model. As an example applied by the present model, I estimated the required number of the repairable bits for 1G DRAM.

Chapter 7 discusses the origins of leakage currents for flash memories and DRAM. The charge loss of flash memories via detrapping and SILCs are caused by the oxide trap of 0.37 [eV] and 3.6 [eV], respectively. Trap levels of 0.37 [eV] and 3.6 [eV] are thought to originate in oxygen vacancies of dimer and fourfold configurations, respectively. For *tail bits* of DRAM, the origin of anomalous leakage of *tail bits* is the silicon trap of 0.68 [eV]. By comparison with the previous data of silicon traps by metal contaminations, the elements which show close values of 0.677 [eV] are “Fe” and “Cu”. From the failure rate of DRAM cells, the contamination level of 0.01 [ppb] is obtained, which is below the purity level of silicon substrates.

The prospects for conventional and emerging memories are remarked in Chapter 8. Conventional memories, including NOR and NAND type flash memories, and DRAMs, would be believed that the data retention would be more serious problem for the future applications. Therefore, the new memories of the retention-problem-free are proposed including Ferroelectric RAMs (FeRAMs), Magnetoresistive RAMs (MRAMs) and Phase-Change RAMs (PRAMs). FeRAMs, MRAMs and PRAMs store the information via the polarization of the ferroelectric film, the magnetization of the magnetic tunnel junction (MTJ), the phases, i.e., crystalline or amorphous, of the Chalcogenide glass, respectively. These emerging memories possess nearly ideal properties, i.e., superior data retention, fast random access, virtually unlimited usage. Excellent functional properties of the new memories offer possibilities to displace the existing memories or create the new types of applications.

Finally, in Chapter 9, conclusions of the thesis are described. The data retention of flash memories can be attributed to the detrapping from the trap with an energy level of 0.37 [eV] as well as the stress induced leakage currents (SILCs) via 3.6 [eV] trap in the oxide layer. The origins of these traps can be attributed to the oxygen vacancies. Oxygen vacancies basically exist in the oxide and are hardly removed. On the other hand, the retention of DRAMs originates 0.68 [eV]-trap in the Si p-n junction. This trap can be ascribed to the contaminated Fe atoms on the concentration order of 0.01 [ppb]. This level of contamination is below the purity of the silicon substrate. As a result, we can not remove these fundamental origins of charge loss for both flash memories and DRAMs. Therefore, development of new semiconductor memories including FeRAMs, MRAMs and PRAMs, are crucial in near future.

1.3 References

[1] K. Itoh: *VLSI Memory Chip Design* (Springer, New York, 2001) 1st ed.
[2] W. D. Brown and J. E. Brewer: *Nonvolatile Semiconductor Memory Technology* (IEEE, Inc., New York, 1998) 1st ed.
[3] Y. Taur and T. H. Ning: *Fundamentals of Modern VLSI Devices* (Cambridge University Press, Cambridge, 1998) 1st ed.
[4] S. M. Sze: *Physics of Semiconductor Devices* (Wiley, New York, 1981) 2nd ed.
[5] A. S. Grove: *Physics and technology of semiconductor devices* (John Wiely & Sons, New York, 1967) 1st ed.
[6] S. Wolf: Silicon processing for the VLSI era vol.3: the submicron MOSFET (Lattice Press, California, 1995) 1st ed.

2

Structure, operation and reliability of flash memories

Summary

In this chapter, structures, operations, and reliability issues of flash memories are briefly reviewed. After quick review of the history for flash memories, the cell structures and their operations, i.e., program, erase and read, are explained, including NOR-type, NAND-type and AND-type flash memories. For flash memories, the most critical reliability issue is the data retention. The data retention characteristics are determined by two mechanisms, i.e., detrapping and stress induced leakage currents (SILCs). In this thesis, origins of detrapping and SILCs have been studied. To understand the previous works and the present work, the direct tunneling, the Fowler-Nordheim (FN) tunneling, and the Poole-Frenkel (PF) emission are described. To understand the carrier emission from the multiple traps in the insulator, the tunnel front model and thermal (Poole-Frenkel) emission front model are also described.

2.1 Introduction

The first modern flash memories are proposed at 1984 International Electron Devices Meeting by Masuoka et al. [1-3]. At present, the name flash memories is used for all EEPROM in which all or a large number of cells, called a block or a page, are erased at the same time.

Flash memories array reported is based on an EPROM array. It has similar high density, low cost and high reliability advantages of EPROMs. The challenge in flash memories is to learn how to remove electrons from the floating gate by an electrical process instead of UV light illuminations used for EPROMs. The advantage of flash memories is that the erase time is less than 1 second, while the erase time for UV-EPROMs is about 10 minutes. Moreover, UV-EPROMs require expensive packages because of UV transparent quartz windows, and it must be taken out of the system for erasure. Flash memories, however, can be packaged in small, inexpensive plastic packages and reprogrammed in-system. The limitation of flash memories compared to traditional EEPROMs is that many bytes are erased simultaneously, instead of a single byte at a time.

In 1985, Mukherjee et al. proposed a source-erase type of flash memory cell called the stacked gate cell [4], as shown in Fig. 2-1. The structure of this cell is the same as that of the stacked gate UV-EPROMs, with two modifications: (1) the source junction is graded to support the high voltage during erase, and (2) the gate oxide is considerably thinner than UV-EPROMs to allow Fowler-Nordheim carrier tunneling during erase. Since 1985, the stacked gate cell has become the volume shipment leader of flash memories. Several variations of the stacked gate cell have been reported [5-9].

In 1987, Masuoka et al. proposed a NAND structured cell [10] to reduce the cell size of conventional NOR structures [1-4]. This structure reduces the cell size without scaling the device dimensions. The conventional NOR structures has one-half contact per bit. However, for a NAND structured cell, only one-half contact hole is required per one NAND structure of 8 bits. As a result, the NAND cell can realize a smaller cell area per bit than the current EPROM [11-12].

After that, several other cell structure and array architecture have been disclosed, i.e., AND [13], DINOR [14], and HICR [15]. In a last few years, cell structures converge to NOR and NAND types. NOR and NAND type have been applied to program storages of micro control units (MCUs) and high volume data storages, respectively.

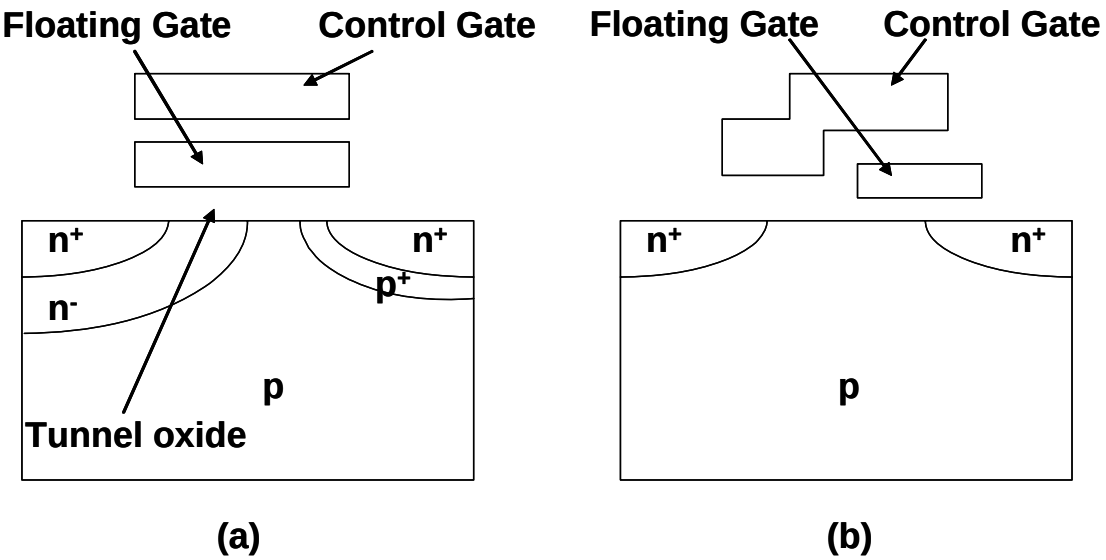


Fig. 2-1 Top view and cross-section view along the word-line and bit-line of NOR cells.

2.2 Cell structures of flash memories and their operations

2.2.1 NOR structure and its operation

Figure 2-2 shows a top view and cross-section views along the word-line and bit-line directions [16]. The drain region consists of an n^+ S/D diffusion, which is aligned with the edge of the control gate (poly 2). The source region consists of an n^+ S/D diffusion, which overlaps the floating gate (poly 1). A cell implant beneath the floating gate is used to control the intrinsic cell threshold (V_t) and the punch-through voltage. The control gate is separated from the channel by a 40 [nm] oxide. The floating gate is separated from the channel and the source diffusion by a thermally grown 15nm gate oxide. The floating gate is separated from the control gate by a 40nm oxide on the side-wall and a 200 [nm] oxide vertically between the control and the floating gates. A silicide or policide can be formed on the control gate to reduce the poly word-line resistance.

Figure 2-3 shows the cell array schematic by the equivalent circuit and the condition for the memory cell terminal during erase, program and read operations. A common source is used for each page; that is, each pair of bits sharing a common source along a row pair (even plus odd row). Programming is either byte by byte individually or for all bytes within the same page simultaneously.

The cell is erased by the F-N tunneling of electrons from floating gate through inter-poly oxide. During erase, the source and drain are grounded and the word-line is raised to 14 [V]. The conduction for the erase is given in Fig. 2-4. The low coupling ratio between the control gate and the floating gate provides a significant voltage drop across the inter-poly oxide. A high field is generated primarily in the area of the tunneling injector. Charge transfer is very rapid and is eventually limited by the accumulation of the positive charge on the floating gate. This positive charge raises the floating gate voltage such that there is the insufficient voltage drop of the poly-to-poly dielectric to sustain the FN tunneling. The removal of the charge can leave a net positive charge on the floating gate. The positive charge on the floating gate reduces the memory cell's threshold voltage to about the control gate V_t.

The cell is programmed using high-efficiency source-side channel hot-electron injection. The conditions for programming are given in Fig. 2-4. During the programming, a voltage of approximately V_t volts is placed on the control gate via the word-line. This is sufficient to turn on the channel under the select portion of the control gate. The drain is at approximately V_{ss} if the cell is to be programmed. If the drain is at V_{cc}, programming is inhibited. The drain voltage is transferred across the select channel because of the voltage on the control gate. The source is at approximately 12 [V]. The

source to drain voltage differential generates channel hot electrons. The source voltage is capacitively coupled to the floating gate. The electric field between the floating gate and the channel sweeps the channel hot electrons that cross the Si-SiO_2 barrier height of approximately 3.2 [V] to the floating gate very efficiently. The programming effect is eventually self-limiting as negative charge accumulates on the floating gate.

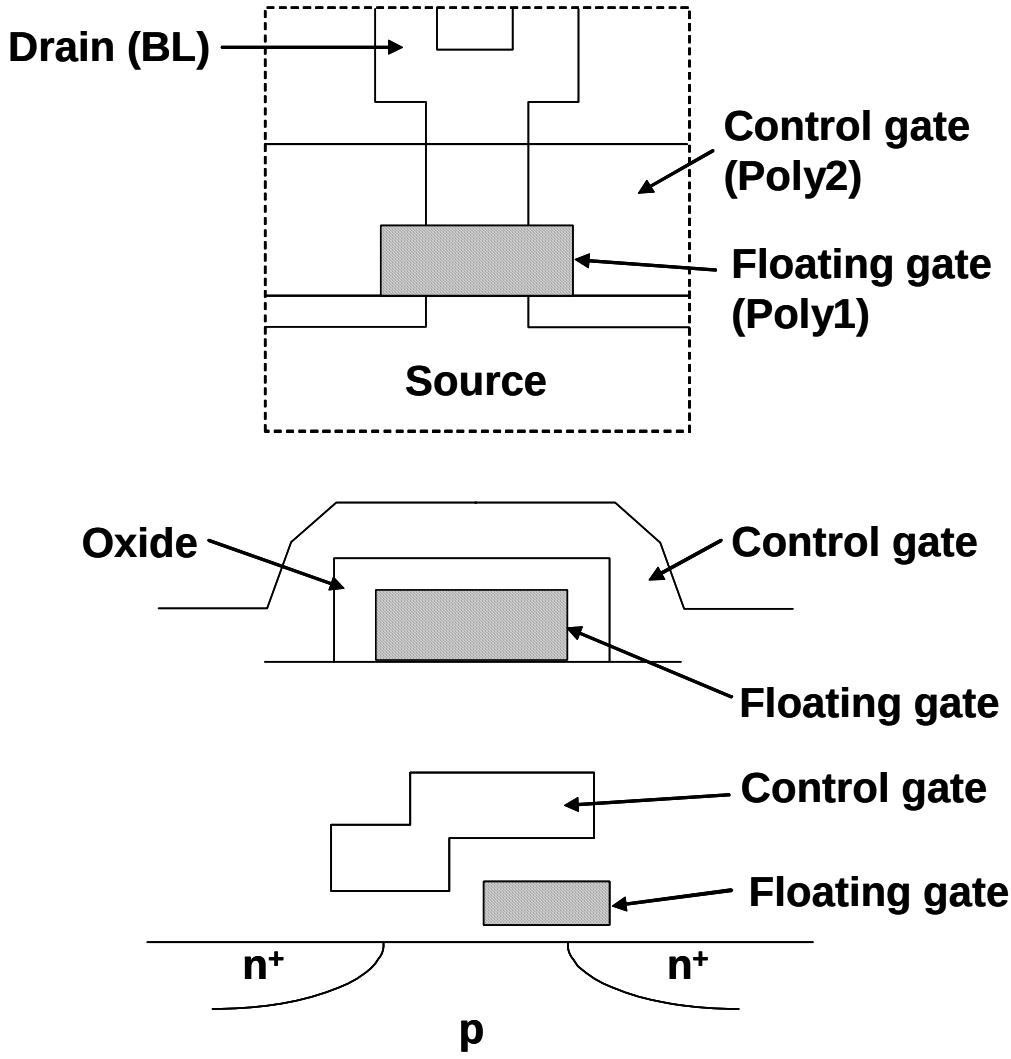


Fig. 2-2 Top view and cross-section view along the word-line and bit-line of NOR cells.

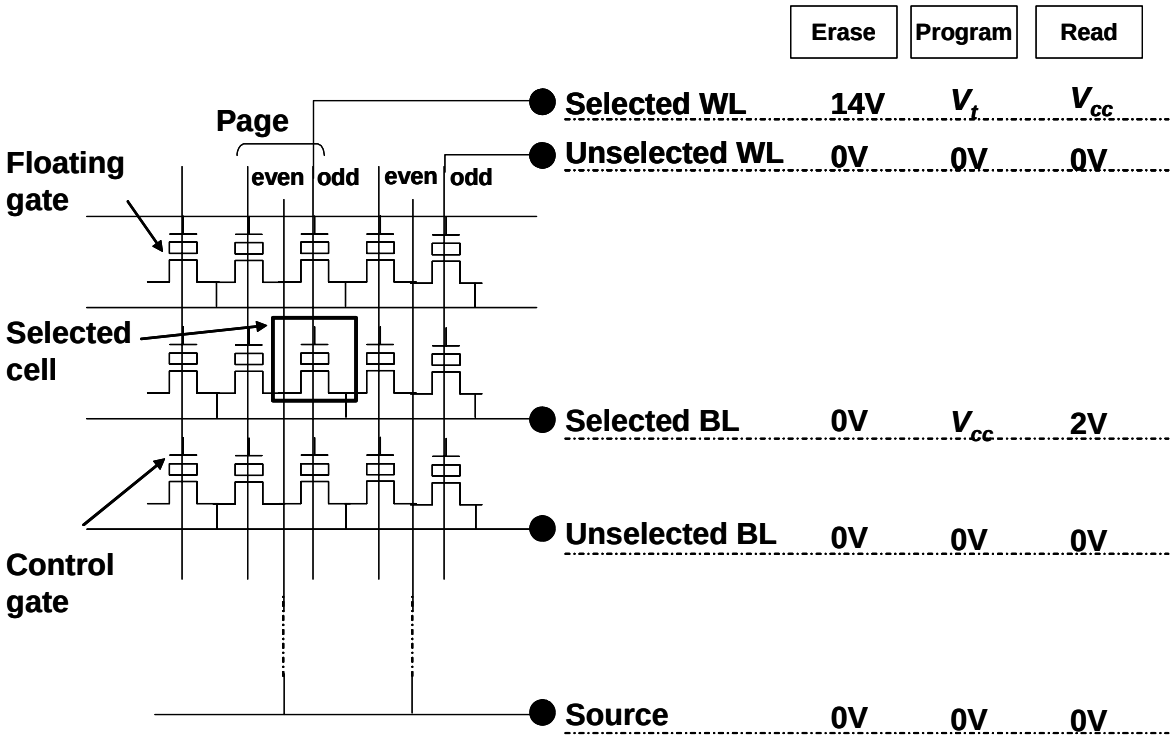


Fig. 2-3 Equivalent circuit and operations of NOR flash memory.

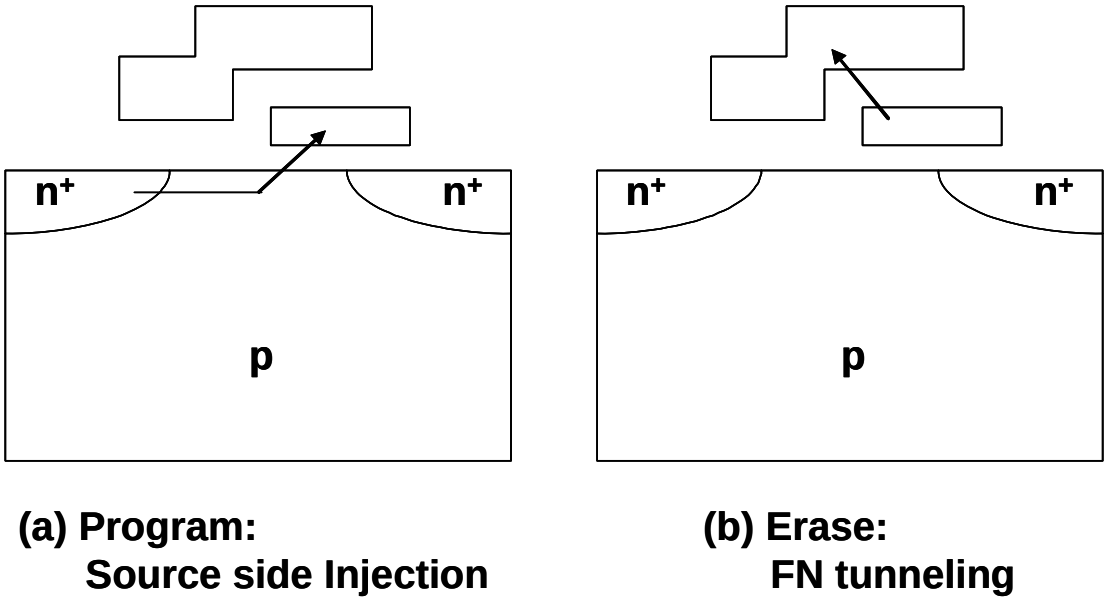


Fig. 2-4 Program and erase of NOR flash memory.

2.2.2 NAND structure and its operation

Figure 2-5 shows the layout and the equivalent circuit of NAND structured cell. As shown in this figure, NAND structured cell arrange eight or sixteen memory transistors in series which are sandwiched between two select gates, the select gate 1 and the select gate 2. The select gate 1 ensures selectivity, and the select gate 2 prevents the cell current from passing during a programming operation. The floating gate and the control gate are made of a first-level poly-silicon and a second-level poly-silicon, respectively. The dielectric between the floating gate and the control gate is an Oxide/Nitride/Oxide (ONO) stack.

Figure 2-6 shows the cross sectional view of NAND structured cell. The peripheral CMOS circuitry and NAND cell array are located in different p-wells which are electrically isolated from each other. The peripheral circuitry is fabricated in p-well (1). NAND cell arrays are located in p-well (2). In an erase operation, p-well (2) is raised to 20 [V] and p-well (1) is always grounded. Thus, p-well (2) can be biased to erase voltage, V_{pp}, during erase, while keeping p-well (1) and the n-well of the peripheral circuitry at “low” voltage, V_{ss}, and “high” voltage, V_{cc}, respectively.

Figure 2-7 shows the equivalent circuit and operating conditions during program, erase and read. The reading method is essentially the same as that of NAND-type MASK ROM. Erased cells, “0”, have a negative threshold voltage, and programmed cells, “1”, have a positive threshold. Zero volts is applied to the control gate of the selected memory cell, while 5 [V] are applied to the gates of the other cells. Therefore, all of the other memory transistors, except for the selected transistor, serve as transfer gates. As a result, in the case where a “0” is being written, the memory transistor is in the depletion mode, and current flows. On the other hand, current does not flow in the case where a “1” is being written because the memory transistor is in the enhancement mode. The state of the cell is detected by a sense amplifier that is connected to the bit-line.

In the erase operation, all control gates are grounded, 20 [V] are applied to the n-substrate and p-well (2), and the source and bit-lines are floating, as shown in Fig. 2-8. Erasing can be performed on the whole chip or selected blocks. Electrons are emitted from the floating gate to p-well (2) by FN tunneling, and the V_t of the erased cells becomes negative. During the erase operation, there is no voltage difference between the n^+-drain and p-well (2). Therefore, the gated-diode breakdown between the n^+-drain and p-well (2) does not occur. For 16MB NAND flash memory, the minimum erase block is 4KB. The typical erase time is 10ms (2.5μs/byte).

In the program operation, p-well (2) is grounded, 20 [V] is applied to the selected

control gate, and a 10 [V] is applied to the unselected control gates. Electrons are injected from p-well (2) to the floating gate by F-N tunneling, as shown in Fig. 2-8. The V_t of the selected cell becomes positive. The V_t of the unselected cell remains negative because the voltage across the tunnel oxide is inadequate to start the tunneling current. In this case, while a 10 [V] is applied between the n^+-drain and p-well (2) for unselected bit lines, gated-diode breakdown does not occur because the control gate is biased at 10 [V]. In both erase and program operations, the gated-diode break down is suppressed entirely. As a result, all high voltages are internally generated easily from a 5 [V] power supply through charge-pump circuits because there is no breakdown leakage current. Furthermore, oxide degradation [17] from hole trappings, due to gated-diode breakdown, is avoided. This result is improved program/erase endurance.

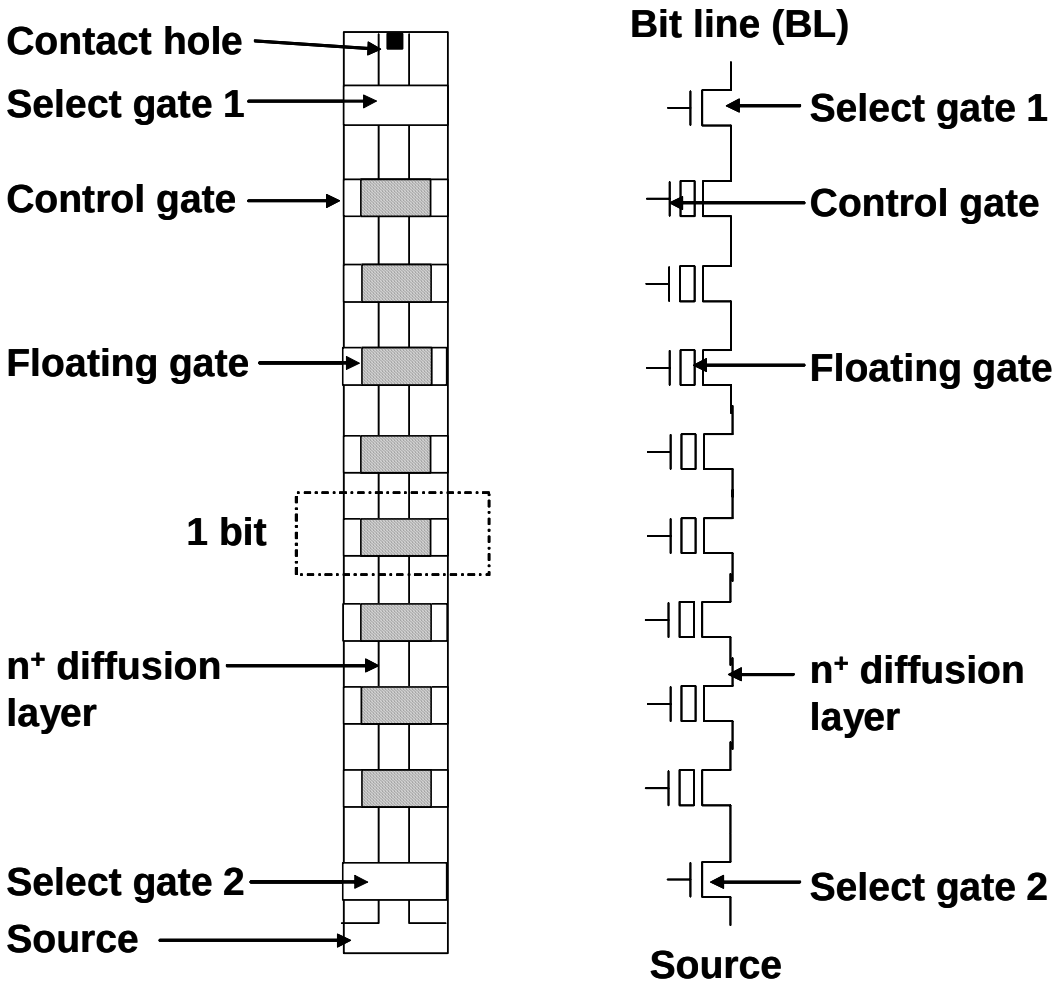


Fig. 2-5 Top view and equivalent circuit of NAND structured cell.

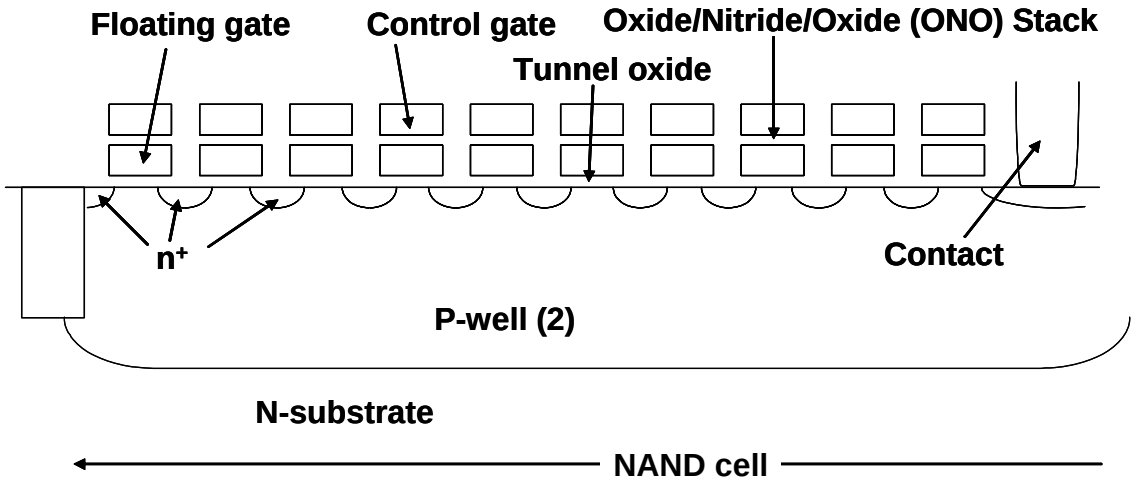


Fig. 2-6 Cross-section view of NAND cell.

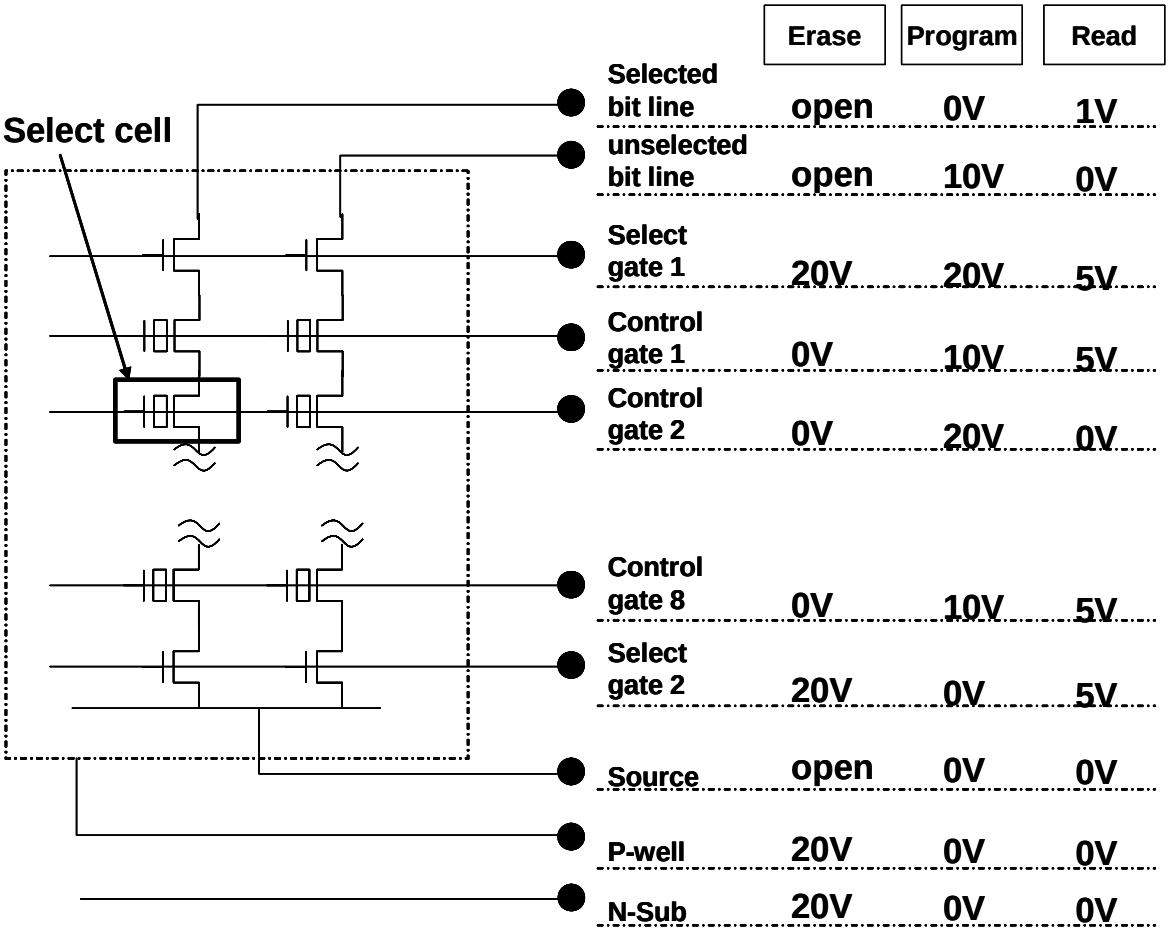


Fig. 2-7 Equivalent circuit and operation voltage of NAND flash memory

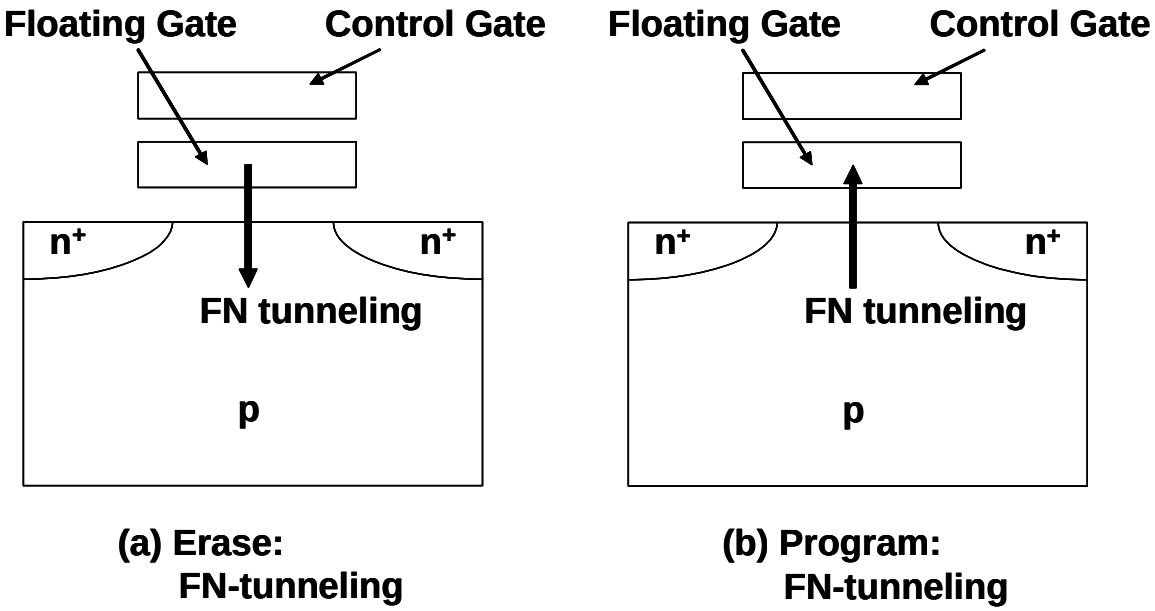


Fig. 2-8 Program and erase of NAND flash memory.

2.2.3 AND structure and its operation

Program and erase operations of the AND cell [18] are performed by F-N tunneling. The memory cell is programmed through the drain by tunnel injection of electrons and is erased by tunnel injection from the whole channel region, as shown in Fig. 2-9. The schematic diagram of this contact-less memory array and the program/ erase scheme are shown in Fig. 2-10. The memory cells are arranged in parallel between each pair of local data- and source-lines, and select transistor 1 and select transistor 2 are switches for connecting and disconnecting the target block to a global data-line and a common source-line, respectively. Both the local data-line and source-line are n^+-diffusions. Internal operating voltages applied to the selected word-line, -9 [V] for programming and +13 [V] for erasing, are generated from the single 3 [V] power supply voltage by on-chip voltage converters. The rewrite size is as small as 512 bytes (which corresponds to memory cells connected to each word-line), suitable for various silicon file applications. Each local source-line is disconnected from the common source-line when programming to allow program inhibit for the cells connected to the unselected word-lines. Adjacent bit-lines are separated by field oxide isolation.

The AND cell and the NAND cell is applied to the silicon data storage. The NAND cell is commercialized by Toshiba Corp. and Sumsung Electronics, while the AND cell is commercialized by Renesas Technology. In 2005, Renesas Technology retreated from the flash business. Therefore, the AND cell has disappeared from the market.

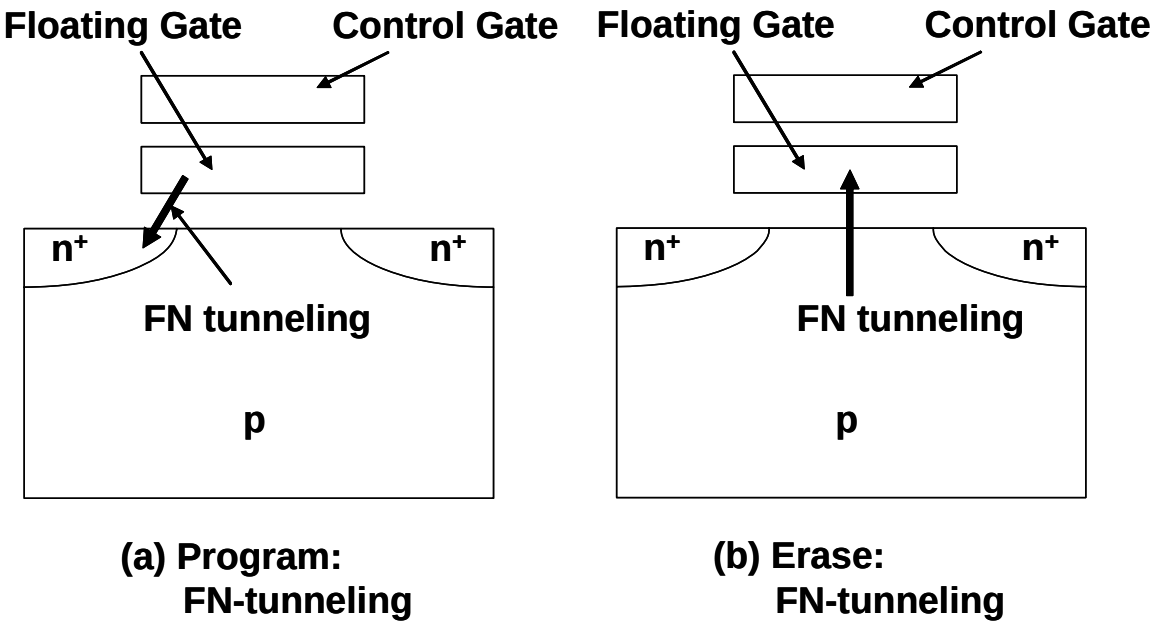


Fig. 2-9 Program and erase of AND flash memory.

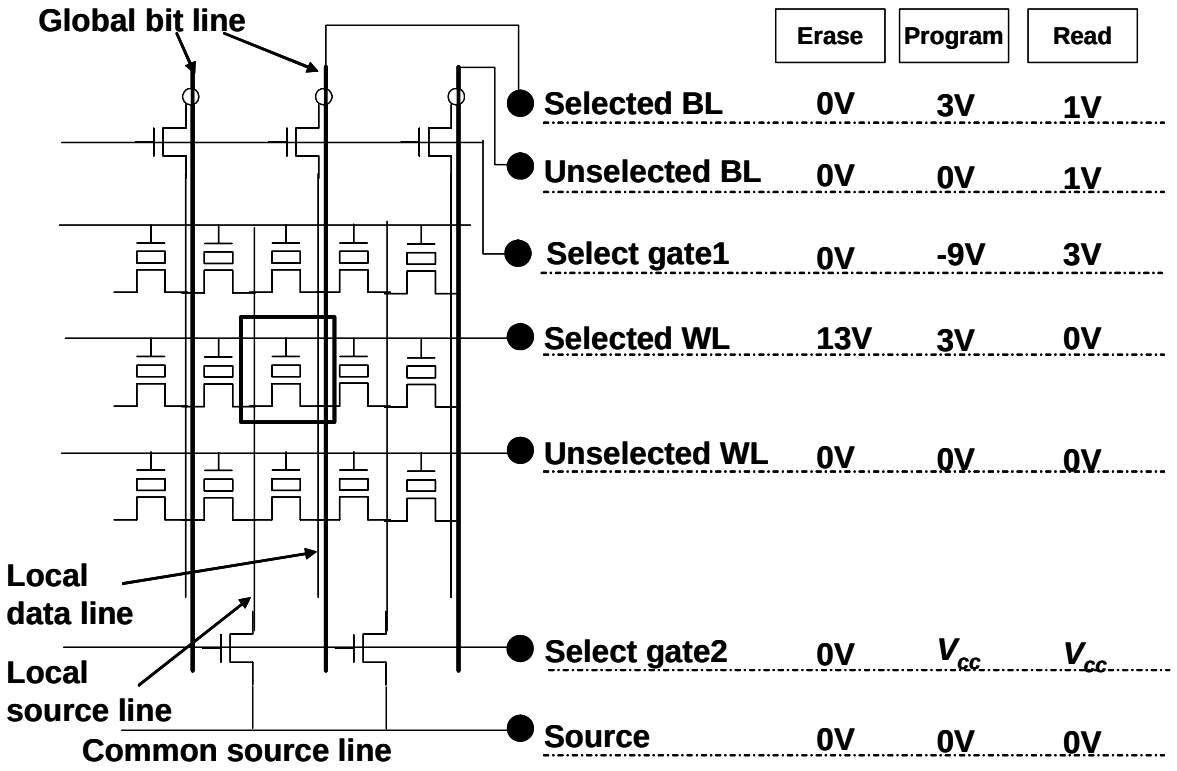


Fig. 2-10 Equivalent circuit and operations of AND flash memory.

2.5 Data retention

By programming or erase, threshold voltages of each memory cells are set to designed levels, i.e., “0” or “1” level. Threshold voltage, V_{th}, after programming or erasing should not shift beyond specifications. Specifications of V_{th}-shift are determined under some conditions, i.e., number of program/erase cycles, cycling temperatures, bake temperatures, bake times and gate voltages during bakes. Here, the shifts of threshold voltages under predetermined conditions are called as the data retention characteristics. For example, a modern flash memory cell stores on the order of a thousand electrons in its floating gate. The cell will, therefore, lose data within a few years if the floating gate loses only one electron a day, i.e., leakage current of about 10^{-24} [A]. No other kind of semiconductor devices has such a stringent leakage requirement. Simply put, the dielectric surrounding the floating gate must be nearly perfect.

The capability of the data retention is drastically deteriorated after program/erase cycles. For the data storage applications, the data retention is assessed after 10^4 program/erase cycles. Holding temperatures for assessments of the data retention are the room temperature and the high temperature of around 125˚C because the data retention mechanisms are different depending on the temperature. The reason why the data retention is drastically deteriorated is the degradation of the tunnel oxide by program/erase cycles. During program/erase cycles, many defects are generated inside the tunnel oxide via some mechanisms.

Many studies for the data retention of flash memories have been done [20-35] and it is general consensus that two mechanisms are dominant for the data retention characteristics. One is the detrapping [36, 37], and the other is the B-mode SILCs [38]. The detrapping and the B-mode SILCs are also called as the transient SILCs and the soft break down, respectively.

Figure 2-11 shows the two mechanisms of the data retention. After 10^4 program/erase cycles, many defects are generated inside the tunnel oxide. Almost all defects would trap electrons. Statistically, the multiple trap paths would be generated at the weak spot of the tunnel oxide. The threshold voltage, V_t, shifts via emissions of trapped electrons, i.e., the detrapping, and/or outflows of electrons inside the floating gate through the multiple trap path, i.e., B-mode SILC.

Figure 2-12 shows the cumulative plot for V_t of all cells as a function of holding times. Here, holding temperature is 125 [˚C]. This figure shows the typical V_t shift via the detrapping. The cumulative plots show the parallel shifts, which means V_t of almost all cells shift by high temperature holding. Therefore, all cells involve electron traps inside the tunnel oxides. Figure 2-13 also shows the cumulative plot for V_t of all cells as

a function of holding times. Here, holding temperature is the room temperature (R.T.). This figure shows the typical V_t shift via the B-mode SILC. The small portion of the flash cells shows the large V_t shift. Therefore, the very small portion of flash memory cells involves the multiple-trap-path.

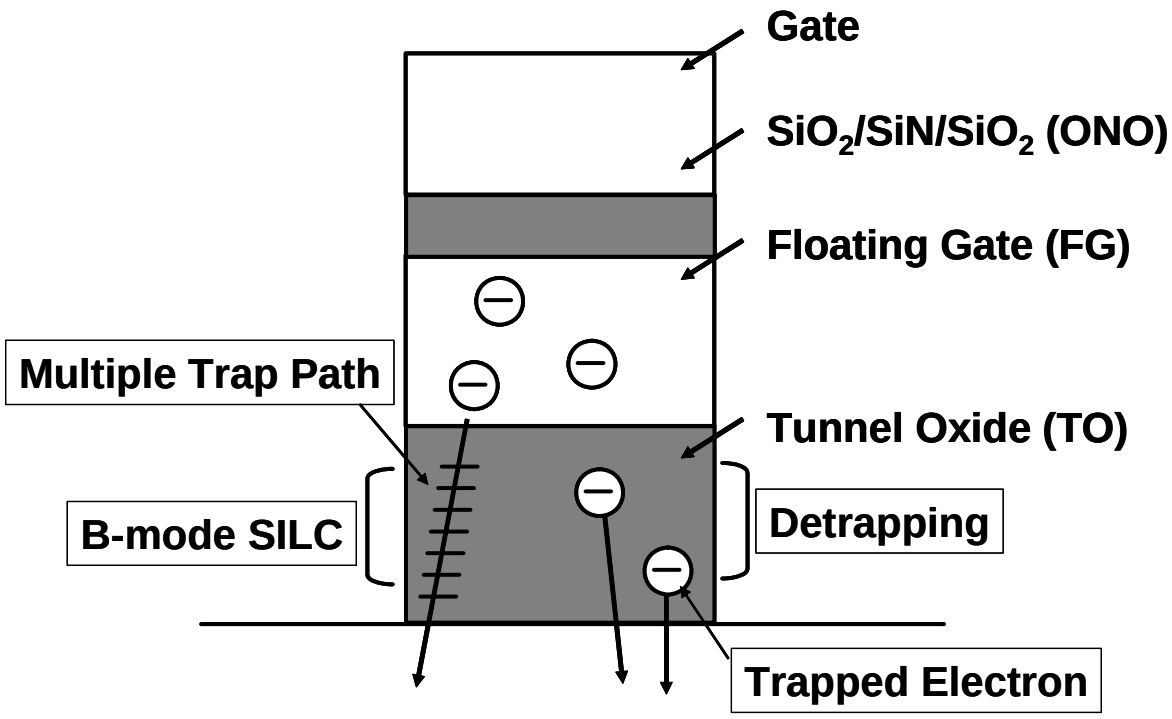


Fig. 2-11 Two mechanisms of the data retention.

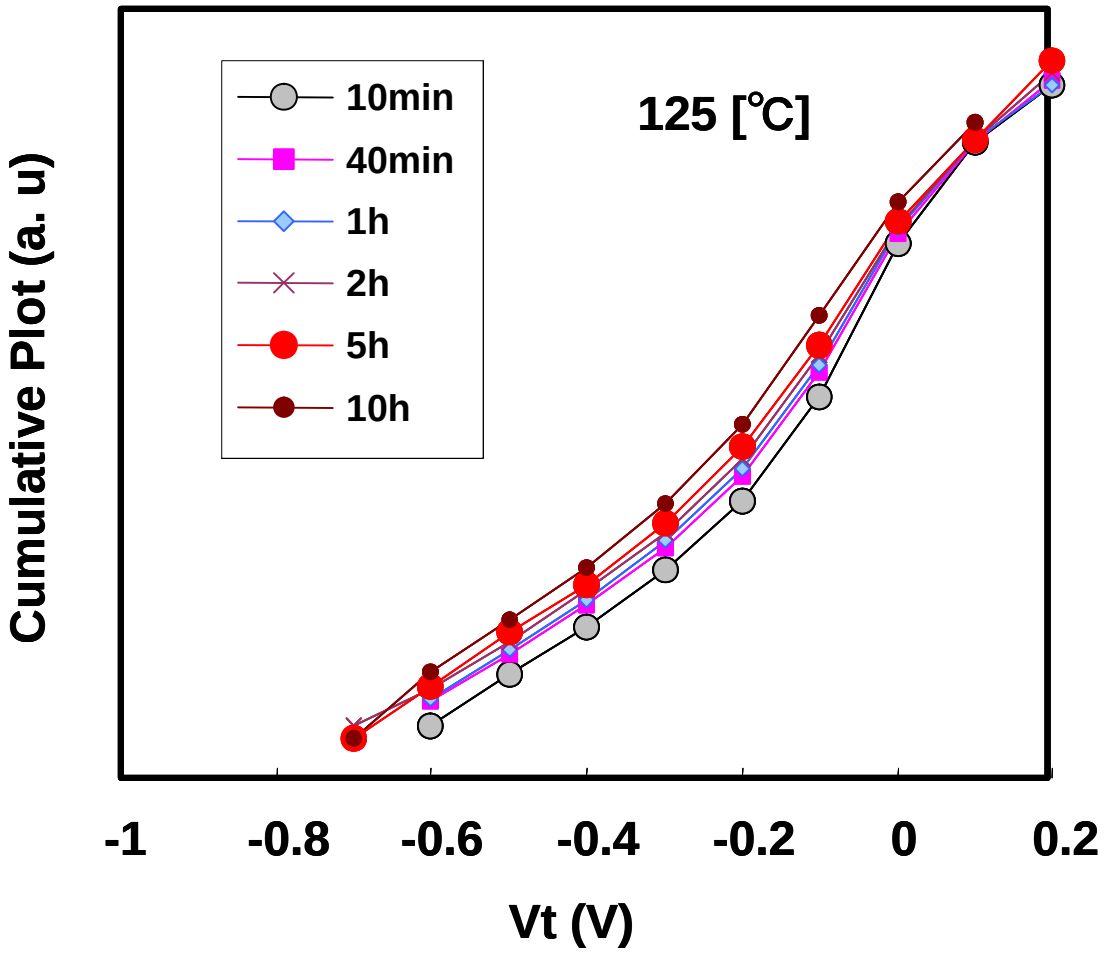


Fig. 2-12 Cumulative plot for V_t of all cells as a function of holding times. This figure shows the typical V_t shift via detrapping.

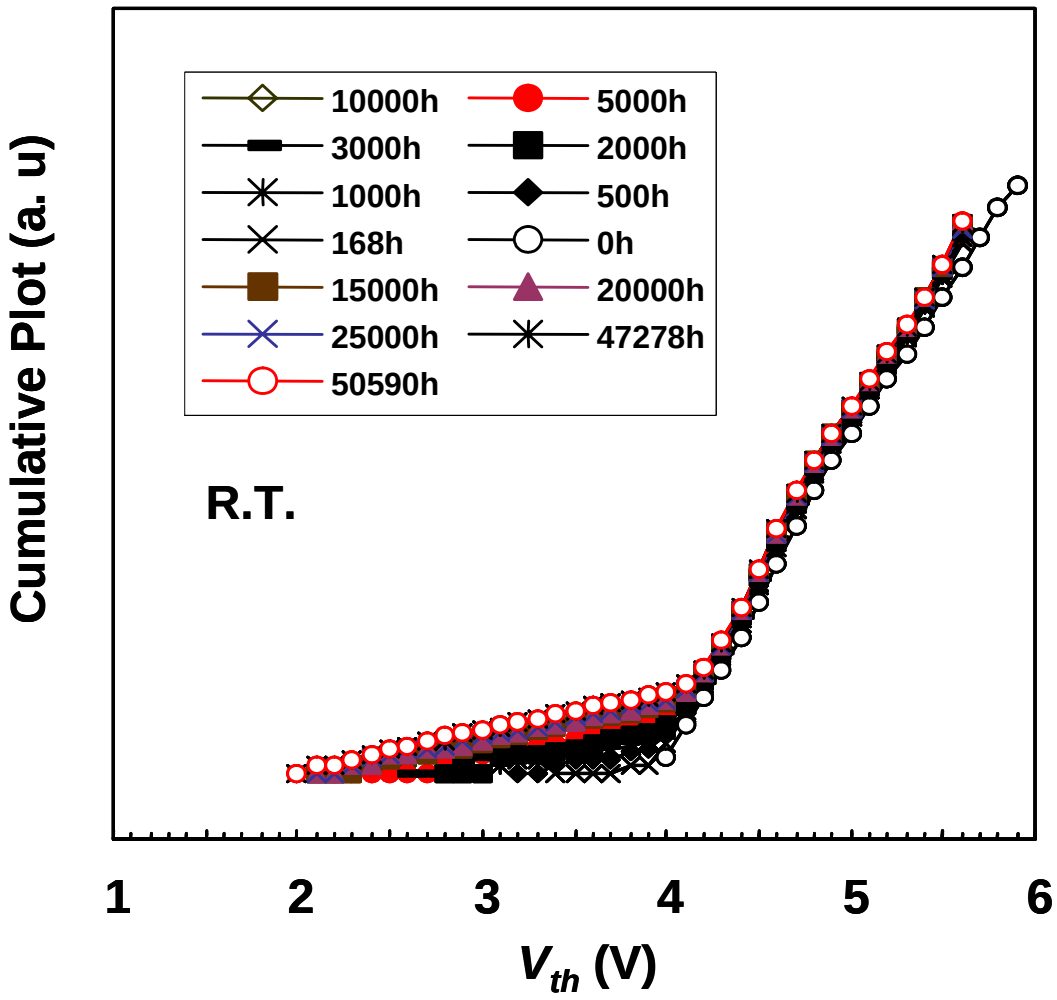


Fig. 2-13 Cumulative plot for V_t of all cells as a function of holding times. This figure shows the typical V_t shift via the B-mode SILCs.

2.4 Charge emission mechanisms

2.4.1 Direct tunneling, FN-tunneling and PF-emission

The direct tunneling (DT) tunneling has been studied extensively in MOS structures where it has been shown to be the dominant current mechanism, especially for thin oxides. The tunneling probability, P_t, can be given by the Wentzel-Kramers-Brillouin (WKB) method [39]:

$$P_t \approx \exp\left[-2\int_{x_1}^{x_2}|k(x)|dx\right], \tag{2-1}$$

where $|k(x)|$ is the absolute value of the wave vector of the carrier in the barrier, and x_1 and x_2 are the classical turning points.

For the direct tunneling as shown in Fig. 2-14 (a), $k(x)$ is described by

$$k(x)=\sqrt{\frac{2m^*}{\hbar^2}(E_0-qFx)}\,, \tag{2-2}$$

where E_0 is the band offset between conduction band of silicon and SiO_2, F the field applied to SiO_2, and m^* the effective mass of electrons. As shown in Fig. 2-14 (a), $x_1=0$ and $x_2=T_{ox}$. Substituting Eq. (2-2) to (2-1), the probability of the direct tunneling is obtained:

$$P_{DT} \approx \exp\left[-\frac{4}{3}\sqrt{\frac{2m^*}{\hbar^2}}\frac{1}{qF}\left[E_0^{\frac{3}{2}}-(E_0-qFT_{ox})^{\frac{3}{2}}\right]\right]. \tag{2-3}$$

The direct tunneling current can be obtained by introducing the suitable pre-factor. For example,

$$J_{DT}=\frac{qn}{\tau}P_{DT} \tag{2-4}$$

where τ is the relaxation time and n is the electron area density.

Fowler-Nordheim (FN) tunneling has been studied extensively in MOS structures where it has been shown to be the dominant current mechanism, especially for thick oxides. The basic idea is that quantum mechanical tunneling from the adjacent

conductor into the insulator limits the current through the structure. Once the carriers have tunneled into the insulator they are free to move within the valence or conduction band of the insulator. The calculation of the current is based on the WKB approximation. For the case of FN tunneling probability, x_1=0 and $x_2=E_0/qF$, as shown in Fig. 2-14 (b). Substituting (2-2) to (2-1), we can obtain

$$P_{FN} \approx \exp\left[-\frac{4}{3}\sqrt{\frac{2m^*}{\hbar^2}}\frac{E_0^{\frac{3}{2}}}{qF}\right]. \tag{2-5}$$

The tunneling current is obtained from the product of the carrier charge, velocity and density. The velocity with which on average the carriers approach to the barrier equals the Richardson velocity, v_R, while the carrier density equals the density of available electrons at the barrier, n. The tunneling probability multiplied with the Richardson velocity and the carrier density yields:

$$J_{FN} \approx qv_R n \exp\left[-\frac{4}{3}\sqrt{\frac{2m^*}{\hbar^2}}\frac{E_0^{\frac{3}{2}}}{qF}\right]. \tag{2-6}$$

The tunneling current, therefore, depends exponentially on the barrier height to the 3/2 power.

Finally, the Poole-Frenkel (PF) emission model is explained. The expression for Fowler-Nordheim tunneling implies that carriers are free to move through the insulator. Whereas this is indeed the case in thermally grown silicon dioxide it is frequently not so in deposited insulators, which contain a high density of structural defects. Silicon nitride (Si_3N_4) is an example of such material. The structural defects cause additional energy states close to the band edge, called traps. These traps restrict the current flow because of a capture and emission process, thereby becoming the dominant current mechanism, as shown in Fig. 2-14 (c). The current is a simple drift current described by

$$J_{PF} = qn\mu F\,, \tag{2-7}$$

where μ is the mobility. The carrier density depends exponentially on the depth of the

trap, which is corrected for the electric field

$$n = n_0 \exp\left[-\frac{q}{kT}\left(\phi_B - \sqrt{\frac{qF}{\pi\varepsilon}} \right) \right] \tag{2-8}$$

The total current then equals:

$$J_{PF} = (q\mu n_0) F \exp\left[-\frac{q}{kT}\left(\phi_B - \sqrt{\frac{qF}{\pi\varepsilon}} \right) \right] \tag{2-9}$$

The existence of a large density of shallow traps in CVD silicon nitride makes Poole-Frenkel emission a frequently observed and well-characterized mechanism.

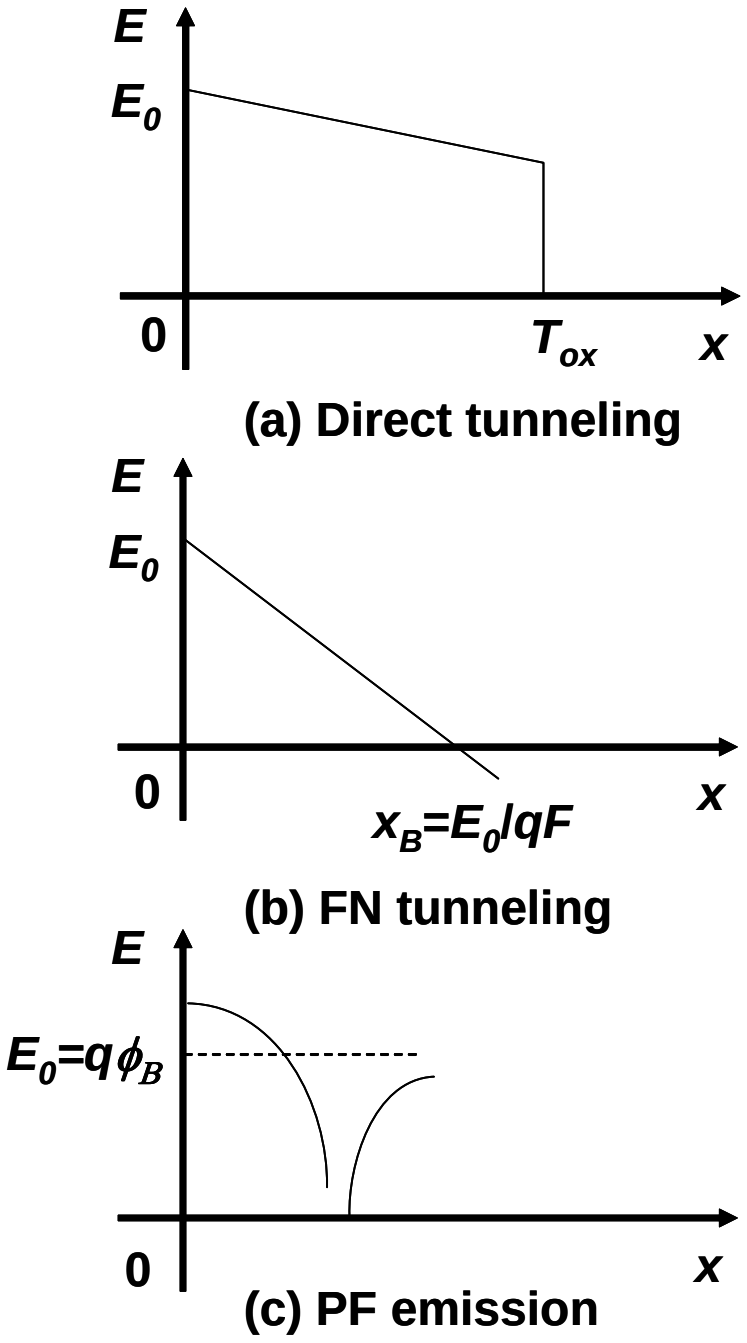


Fig. 2-14 Illustrations to explain direct tunneling, FN tunneling and PF emission.

2.4.2 Tunnel front model and thermal emission front model

Tunnel front model is described as follows. Suppose that trapped electrons spatially align inside oxide, as shown in Fig. 2-15 (a). Electrons emit at a stone's throw from the closest traps to the interface, x_0. The model equation for the time evolution of the total trapped charge inside oxide, $Q(t)$ are derived as follows.
The tunneling probability for the direct tunneling can be described as follows:

$$P_{DT} \approx \exp\left[-\frac{4}{3}\sqrt{\frac{2m^*}{\hbar^2}}\frac{E_0^{\frac{3}{2}}}{qF}\left[1-(1-\frac{qFx}{E_0})^{\frac{3}{2}}\right]\right]. \approx \exp\left[-2\sqrt{\frac{2m^*E_0}{\hbar^2}}x\right] = \exp(-\alpha x) \quad (2\text{-}10)$$

The rate equation for the probability of the trap occupation by electrons, f_t, is

$$\frac{\partial f_t}{\partial t} = -f_t\frac{1}{\tau\exp(-\alpha x)}, \quad (2\text{-}11)$$

where $\tau\exp(-\alpha x)$ is the escape time of trapped electrons at the location of x. Figure 2-15 (a) illustrates the schematic time evolution of f_t. By solving Eq. (2-11), we obtain

$$f_t = \exp(\frac{-t}{\tau\exp(-\alpha x)}). \quad (2\text{-}12)$$

Here, we define f_t =1/2 at x_b which means that f_t =1 at $x< x_b$ and f_t =0 at $x_b<x<x_0$. By using Eq. (2-12), this definition can be expressed,

$$\frac{1}{2} = \exp(\frac{-t}{\tau\exp(-\alpha x_b)}). \quad (2\text{-}13)$$

By solving Eq. (2-13), we can obtain

$$x_b = \frac{1}{\alpha}\ln\left(\frac{\tau\ln(2)}{t}\right). \quad (2\text{-}14)$$

Finally, the time evolution of the total trapped charge inside oxide, $Q(t)$, is described by

$$Q(t) = \rho_x x_b = \frac{\rho_x q}{\alpha} \ln\left(\frac{\tau \ln(2)}{t}\right), \tag{2-15}$$

where ρ_x is the area density of traps which is assumed to be uniformly distributed inside the oxide. Note that the time evolution of $Q(t)$ shows the logarithmic dependence on time.

Thermal emission front model is described as follows. Suppose that trapped electrons energetically align inside the forbidden band of the oxide, as shown in Fig. 2-15 (b). Trapped electrons emit at a stone's throw from the shallowest trap level, U_0. The model equation for the time evolution of the total trapped charge inside oxide, $Q(t)$ are derived as follows. Here, the thermal emission front model (TE front model) is also called as the Poole-Frenkel emission front model (PF front model).
The rate equation for the probability of the trap occupation by electrons, f_t, is

$$\frac{\partial f_t}{\partial t} = -f_t \frac{1}{\tau \exp(-\frac{U}{kT})}, \tag{2-15}$$

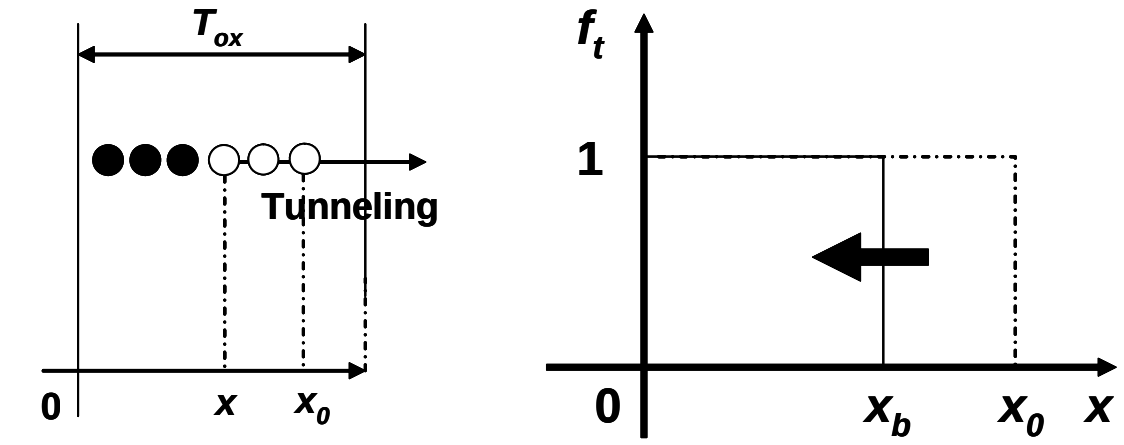
where $\tau \exp(-U/kT)$ is the escape time of trapped electrons at the trap level of U. Figure 2-15 (b) illustrates the schematic time evolution of f_t. Here, we define f_t =1/2 at U_b which means that f_t =1 at $U < U_b$ and f_t =0 at $U_b < U < U_0$. By using Eq. (2-12) and this definition, we can obtain

$$U_b = kT \ln\left(\frac{\tau \ln(2)}{t}\right). \tag{2-16}$$

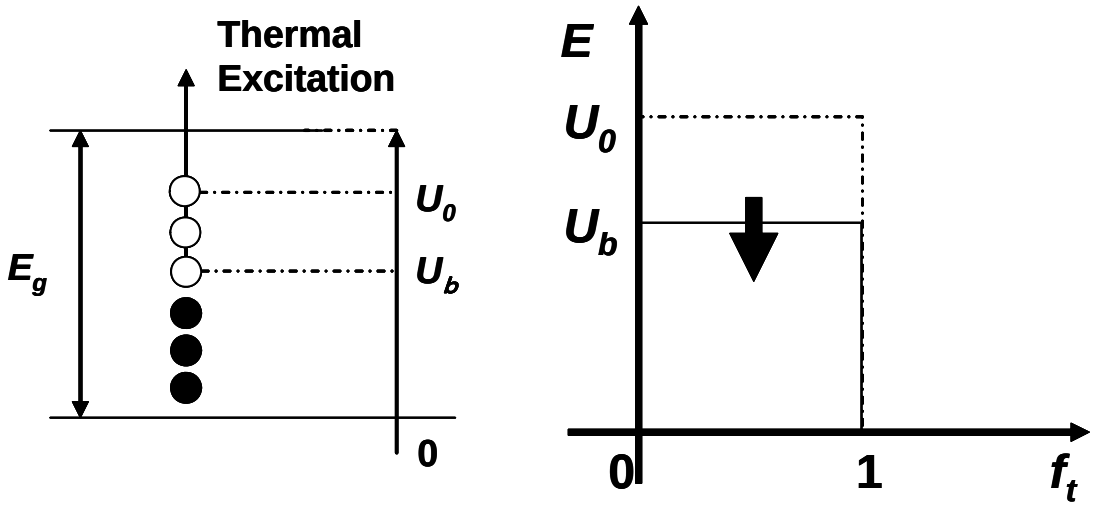
Finally, the time evolution of the total trapped charge inside oxide, $Q(t)$, is described by

$$Q(t) = \rho_E U_b = (q\rho_E) kT \ln\left(\frac{\tau \ln(2)}{t}\right), \tag{2-17}$$

where ρ_E is the area density of trap levels which is assumed to be uniformly distributed inside the forbidden band of the oxide. Note that the time evolution of $Q(t)$ shows the logarithmic dependence on time.



(a) Tunnel Front Model



(b) Thermal Emission Front Model

Fig. 2-15 Illustration of TF and TEF model.

2.5 References

[1] F. Masuoka, M. Assano, H. Iwahashi, T. Komuro, and S. Tanaka: IEDM Tech Dig (1984) 464.
[2] F. Masuoka et al.: ISSCC Dig. Tech. Pap. (1985) 168.
[3] F. Masuoka et al.: IEEE J. Sol. St. Cir. **SC-22** (1987) 548.
[4] S. Mukherjee, T. Chang, R. Pan, M. Knecht, and D. Hu: IEDM Tech. Dig. (1985) 616.
[5] V. N. Kynett, A. Baker, M. Fandrich, G. Hosekstra, O. Jungroth, J. Kreifels and S. Wells: Proc. ISSCC Dig. Teck. Pap. (1994) 146.
[6] G. Samachisa, C. S. Su, Y. S. Kao, G. Smarandoiu, T. Wong, and C. Hu: Proc. ISSCC Dig. Teck. Pap. (1987) 76.
[7] G. Samachisa, C. S. Su, Y. S. Kao, G. Smarandoiu, C. Y. M. Wang, T. Wong, and C. Hu: IEEE J. Solid-State Circuits **SC-22** (1987) 676.
[8] S. Haddad, C. Chang, B. Swaninathan, and J. Lien: IEEE Elect. Dev. Lett. **EDL-10** (1989) 117.
[9] S. Kianian et al.: Symp. VLSI Tech. (1994) 71.
[10] F. Masuoka, M. Momodomi, Y. Iwata, and R. Shirola: IEDM Tech. Dig. (1987) 552.
[11] Y. Iwata et al.: Proc of Semiconductor Memory Workshop (1995) #4-2.
[12] K. D. Suh et al.: ISSCC Dig. Tech. Pap. (1995) 128.
[13] H. Kume et al.: IEDM Tech. Dig. (1992) 991.
[14] H. Onoda et al.: IEDM Tech. Dig. (1992) 599.
[15] Y. S. Hisamune et al.: IEEE J. of Solid State Circuits **SC-29** (1994) 461.
[16] S. Kianian et al.: Symp. Proc. of VLSI Tech. (1994) 71.
[17] V. N. Kynett, A. Baker, M. Fandrich, G. Hoekstra, O. Jungroth, J. Kreifels, and S. Wells: Proc. of ISSCC Dig. Tech.Pap. (1994) 132.
[18] H. Kume et al.: IEDM Tech. Dig. (1992) 991.
[19] Y. Miyawaki et al.: Symp. Proc. of VLSI Tech. (1990) 85.
[20] H. P. Belgal, N. Righos, I. Kalastirsky, J. J. Peterson, R. Shiner and N. Mielke: Proc. of IRPS (2002), 7.
[21] A. Brank, K. Wu, S. Pan and D. Chin: Proc of IRPS (1993) 127.
[22] A Modelli, F. Gilardoni, D. Ielmini and A. Spinelli: Proc of IRPS (2001) 61.
[23] J. D. Blauwe, J. V. Houdt, D. Wellekens, R. Degraeve, P. Roussel, L. Haspeslagh, L. Deferm, G. Groeseneken and H. E. Maes: Tech. Dig. of IEDM (1996) 343.
[24] A. Scarpa, G. Ghibaudo, G. Pananakakis, A. Paccagnella, and G. Ghidini: Electron

Letters **33** (1997) 1342.
[25] H. Kameyama, Y. Okuyama, S. Kamohara, K. Kubota, H. Kume, K. Okuyama, Y. Manabe, A. Nozoe, H. Uchida, M. Hidaka and K. Ogura: Proc. of IRPS (2000) 194.
[26] K. Naruke, S. Taguchi and M. Wada: Tech. Dig. of IEDM (1988) 424.
[27] D. Baglee and M. Smayling: Tech. Dig. of IEDM (1985) 624.
[28] E. Arai, T. Maruyama and R. Shirota: Proc. of IRPS (1998) 378.
[29] D. Wellekens, P. Hendrickx, F. Schuler and J. V. Houdt: Proc. of NVSM Workshop (2001) 120.
[30] Y. Manabe, K. Okuyama, K. Kubota, A. Nozoe, T. Karashima, K. Ujiie, H. Kannno, M. Nakashima, and N. Ajika: Proc of ICMTS (1998) 95.
[31] A. Scarpa, G. Tao, J. Dijkstra and F. Kuper: 2000 IRW Final Report (2000) 24.
[32] J. Yugami and T. Mine: Symposium on VLSI Tech. Dig. of Technical Paper (1997) 115.
[33] H. Watanabe, S. Aritome, G. Hemink, T. Maruyama and R. Shirota: Symposium on VLSI Tech. Dig. of Technical Paper (1994) 47.
[34] H. Watanabe, K. Shimizu, Y. Takeuchi, S. Aritome: Tech. Dig. of IEDM (1996) 833.
[35] T. Endo, K. Shimizu, H. Iizuka, S. Watanabe and F. Masuoka: Tech. Dig. of IEDM (1994) 49.
[36] R. Yamada, Y. Mori, Y. Okuyama, J. Yugami, T. Nishimoto and H. Kume: Proc. of IRPS 2000 (2000) 200.
[37] R. Yamada, Y. Mori, T. Sekiguchi, Y. Okuyama, J. Yugami and H. Kume: Tech. Dig. of VLSI Symp. 2001 (2001) 115.
[38] S. Kamohara, D. Park and C. Hu: Proc. of IRPS 98 (1998) 57.
[39] S. M. Sze: Physics of Semiconductor Devices (Wiley, New York, 1981) 2nd ed, p.520.

3

Data retention characteristics of flash memories via detrapping

Summary

In this chapter, the data retention of flash memories via detrapping are described. Electron detrapping is the main cause of data lekage in the state-of-the-art flash EEPROM. The log (*t*) dependence of *ΔVth* is a unique aspect of electron detrapping. To explain log (*t*) dependence, we have assumed that after electron detrapping, the positive-ionized trap reduces the probability of the electrons in the influence area being emitted from their site. Based on this assumption, we have developed a model for detrapping that is consistent with the experimental results.

3.1 Introduction

Data retention after ~10000 program/erase (P/E) cycles is one of the most important reliability issues in flash EEPROM. Data retention characteristics are determined by two mechanisms: stress induced leakage current (SILC) [1-5] and the release of electrons trapped inside the tunnel oxide (detrapping) [6, 7]. By using a tunnel oxide >8nm, we can avoid the SILC. Therefore electron detrapping is the remaining cause of data retention in the latest flash memory EEPROM.

Figure 3-1 shows the typical detrapping characteristics of 95˚C annealing after 10000 program/erase cycles obtained from the flash memory test chip. The important aspect is the log (*t*) dependence of *ΔVth*. This log (*t*) dependence means the detrapping current has 1/*t* dependence [6, 7].

To explain the $\log(t)$ dependence of *ΔVth* or 1/*t* dependence of the detrapping current, two models have been proposed to date: one is the tunnel front (TF) model [6, 7, 8], while the other is the Poole-Frenkel emission front (PF) mode [8]. Both these models are described by

$$\Delta Vth = -\alpha \ln\left(\frac{t}{\tau}\right), \qquad (1)$$

and successfully explain $\log(t)$ dependence. However, from a physical point of view, both models have problems. The TF model cannot explain the temperature dependence of the electron detrapping because the physical parameters, α and τ, should be independent of temperature. The experimental results clearly show the temperature dependence as illustrated in Fig. 3-2. The PF model assumes unrealistic traps that have a continuum trap-level. In this paper, we propose a new physical model that can explain $\log(t)$ and the temperature dependencies of the electron detrapping.

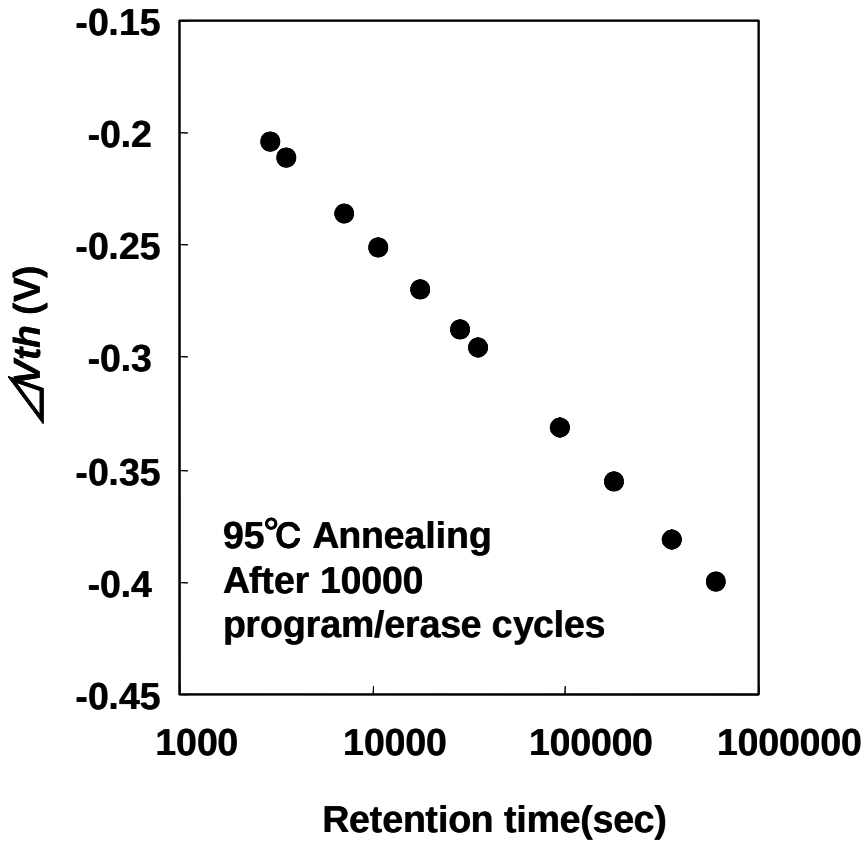


Fig. 3-1. Typical detrapping characteristics of 95˚C annealing after 10000 program/erase cycles obtained from the flash memory test chip.

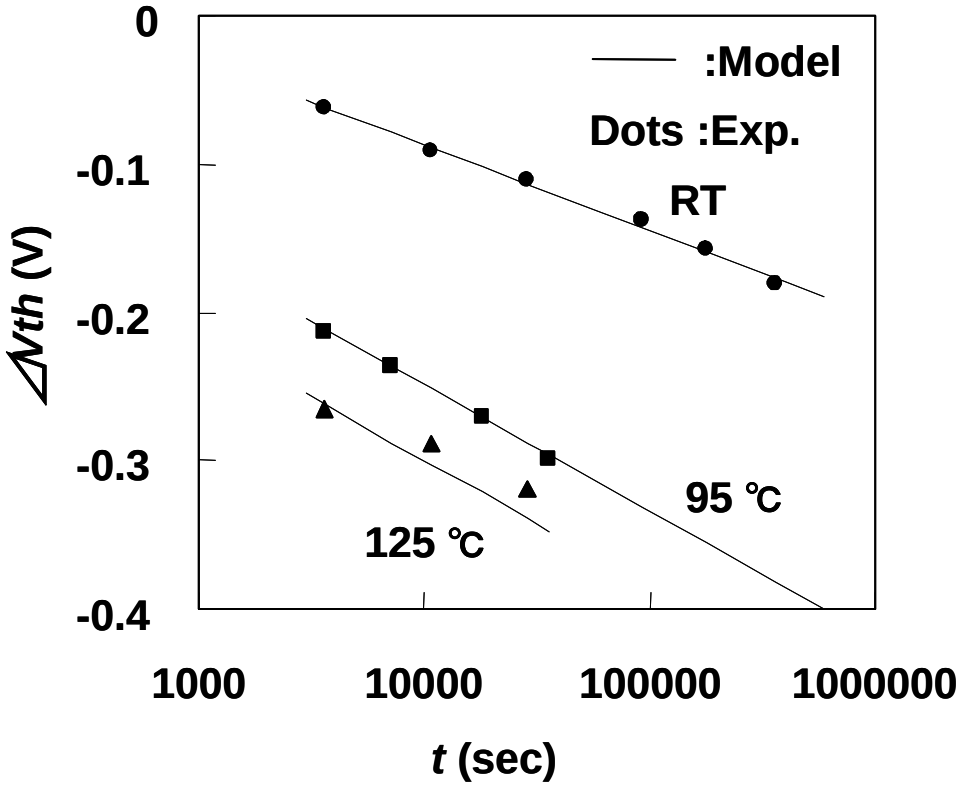


Fig. 3-2. Comparison between the model and the experiments to extract the model parameters.

3.2 Electron detrapping model

Figure 3-3 shows our model for electron detrapping. Fig. 3-3 (a) illustrates the electron trap sites inside the oxide, trapped electrons at the trap sites, the positive-ionized trap and the detrap electron. Suppose that an electron trapped at a certain site is thermally excited into the conduction band and is emitted from its trap site. After the electron detrapping, the trap would be positively charged and couples with the nearest trapped electron. The positive-ionized trap and the coupled electron are stabilized by the Columbic potential. Therefore, it is assumed that after electron detrapping, the positive-ionized trap reduces the probability of the electron in the influence area, *b*, being emitted from its site, as shown in Fig. 3-3 (b). Our model for electron detrapping is similar to the models introduced for the specific cases of chemisorption, oxidation and so on [9].

Derivation of the model equation is as follows. Electron detrapping via thermal excitation is described by

$$\frac{\partial Q}{\partial t} = \frac{qN}{\tau_{th}}, \tag{2}$$

where N is the area density of trap sites occupied by electrons outside the influence area, b, Q the area density of ionized traps, τ_{th} the emission time constant, t time and q the elementary charge. In accordance with the increase in ionized trap sites, the area density of trap sites occupied by electrons outside the influence area, N, decreases, as described by

$$-\frac{\partial N}{\partial t} = \frac{(bN)}{q}\frac{\partial Q}{\partial t}. \tag{3}$$

From Eq. (3), we can infer

$$N = C_1 \exp(-(b/q)Q). \tag{4}$$

Substituting Eq. (4) in Eq. (2), we obtain

$$Q(t) = \frac{q}{b}\ln\left(\frac{t}{\tau *} + const.\right)_{t >> \tau *} \approx \frac{q}{b}\ln\left(\frac{t}{\tau *}\right), \tag{5}$$

where

$$\tau^* = \tau^*_0 \exp\left(E_t / kT\right), \tag{6}$$

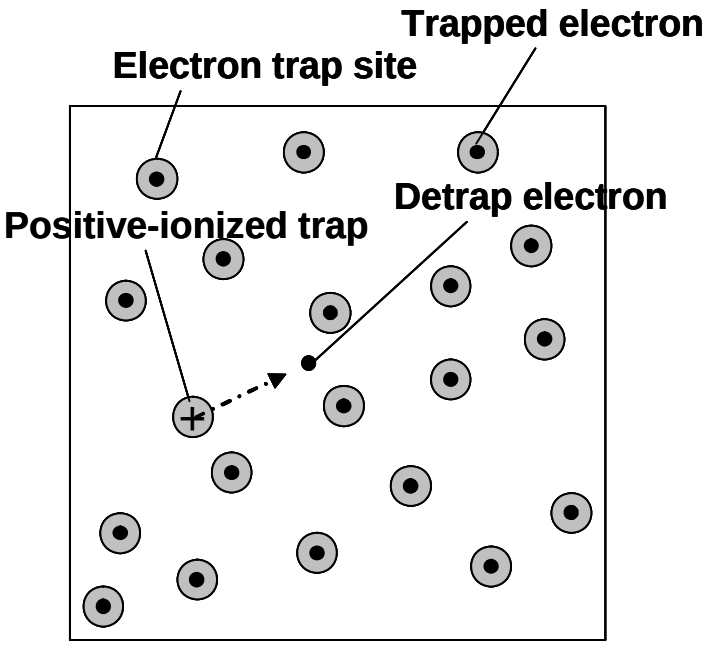
E_t is the trap level, T the temperature and $\tau *_0$ the renormalized time constant. Finally ΔVth can be described as follows:

$$\Delta Vth(t) = -\alpha * \ln\left(\frac{t}{\tau *}\right), \tag{7}$$

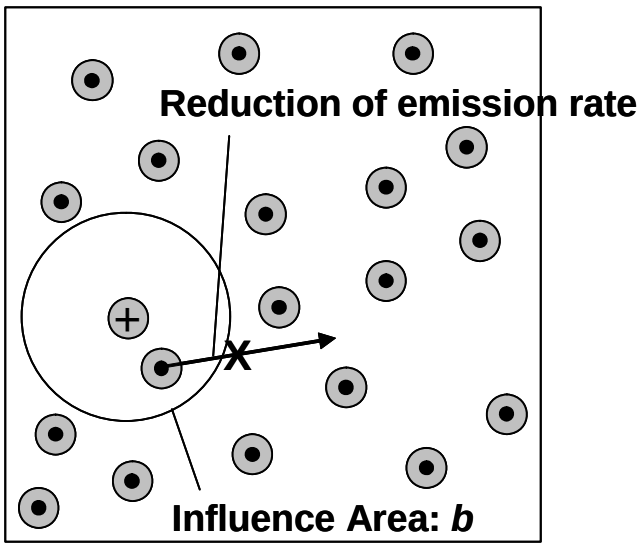
where

$$\alpha^* = q/\left(2bC_{CR}\left(\varepsilon_{ox} / t_{ox}\right)\right), \tag{8}$$

C_{CR} is the coupling ratio between the gate and the floating gate, and t_{ox} the oxide thickness of the tunnel oxide. To introduce Eq. (7), we assume that the traps are uniformly distributed inside the tunnel oxide.



(a)



(b)

Fig. 3-3. Proposed mechanism to explain the log(*t*) dependence of data retention via electron detrapping.

3.3 Extraction of model parameters and discussion

We can extract the trap level, E_t, and the influence area, b, by fitting equation (7) to the experimental results of the detrapping temperature dependence. Figure 3-2 shows the results of fitting the model to the experimental results. For our calculations, we use the following values: C_{CR}=0.5, t_{ox}=9 [nm]. Good agreement is achieved between the model and the experimental results. From the model fitting, we obtain the values of both α^* and τ^* for each temperature. Figure 3-4 illustrates the temperature dependence of α^* and τ,, and that the following parameter values are obtained: E_t=0.37 [eV], and b=1.31×10 $^{-11}$ [cm^2] (r=20.3 [nm]). Here, we discuss the validity of r=20nm. From Fig. 3-1, we obtain the time evolution of the ionized-trap density, $Q(t)/q$. As shown in Fig. 3-5, the data points seem to approach ~ 10 ^{12} [cm^{-2}], so the trap area density would be ~ 10 ^{12} [cm^{-2}] [10]. This means that the average distance of the traps is ~10nm. This value is very close to the r value obtained in Fig. 3-4.

Finally, we discuss a plausible mechanism for charging the trap. The electron trap site is negatively charged by forming the negative-U [11]. Upon detrapping, two electrons escape at the same time, and the empty site is positively charged. The positive empty site and several negative-U are stabilized by the Columbic potential.

Based on the model of the trap charging via the negative-U, the probability reduction of the electron emission by the Columbic interaction with the positive-charged trap was roughly estimated. An electron trapped at a certain site repulses the nearest trapped electron, thereby increasing the probability that this trapped electron will be emitted. Once the electron has been emitted from its site, the positive-ionized trap attracts the nearest trapped electron, thus reducing the probability that this trapped electron will be emitted. Using Columb's law, we can roughly estimate the difference in potential of the trapped electron before and after the nearest electron is emitted. The difference in potential for the average distance of the trapped electron, 10nm, is estimated to be ~0.076 [eV]. A trap level, E_t, of 0.37 [eV] is obtained under the condition of repulsive interaction. Therefore, the trap level under the condition of attractive interaction with the positive ionized trap would effectively be ~0.45 [eV]. As a result, the probability of the electron being emitting from its site is reduced to a value one order smaller at 125 [˚C]. In addition to the Columbic interaction, the deformation of the chemical structure by the electron detrapping would reduce the probability of the electrons emitting from its site, which is remaining issue on this paper.

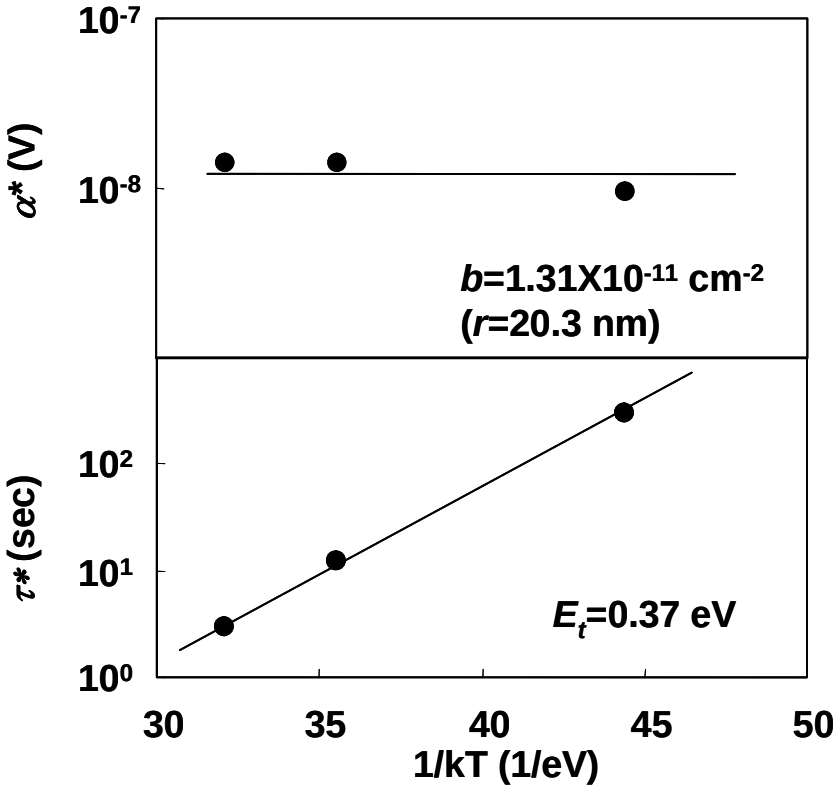


Fig. 3-4. Extraction of the influence area, *b*, and the trap level, E_t.

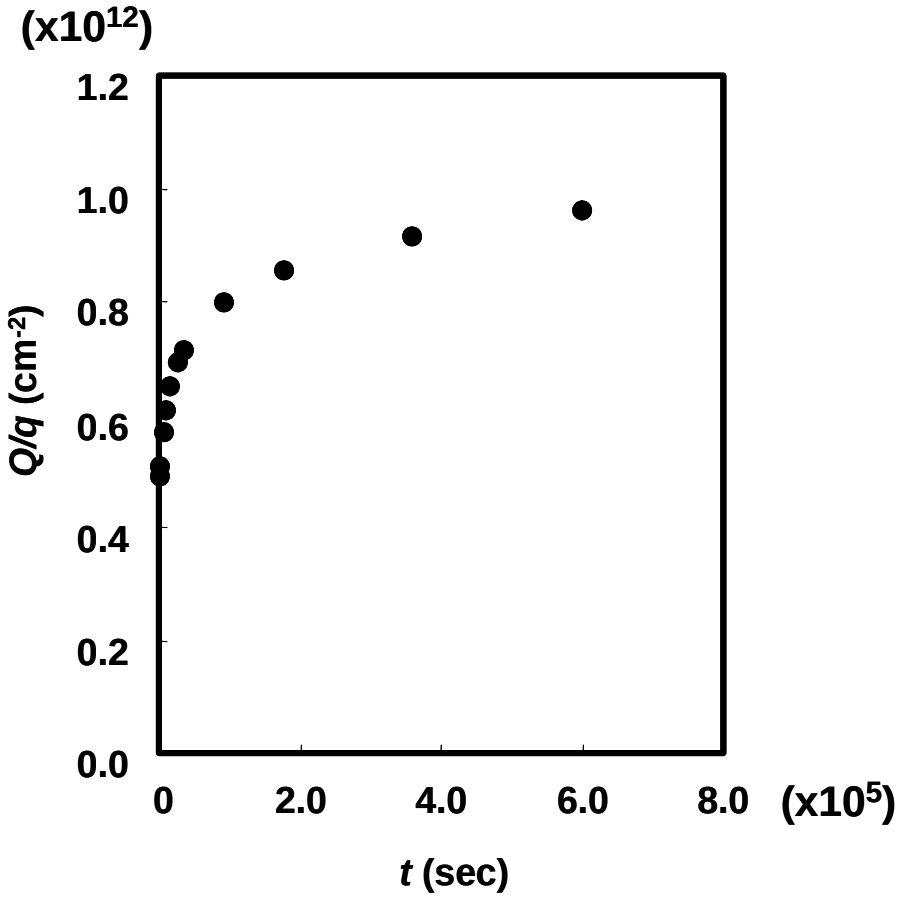


Fig. 3-5. Time dependence of the ionized-trap area density.

3.4 References

[1] D. A. Baglee and M. C. Smayling: IEDM Tech. Dig., (1985) 624.

[2] K. Naruke, S. Taguchi, and M. Wada: IEDM Tech. Dig., (1988) 424.

[3] D. Ielmini, A. S. Spinelli, M. A. Rigamonti, and A. L. Lacatia: IEEE Trans. Electron Devices **47** (2000) 1258.

[4] M. Kato et al.: IEDM Tech Dig. (1994) 45.

[5] S. Kamohara, D. Park and C. Hu: Proc. of IRPS 98 (1998) 57.

[6] R. Yamada, Y. Mori, Y. Okuyama, J. Yugami, T. Nishimoto and H. Kume: Proc. of IRPS 2000 (2000) 200.

[7] R. Yamada, Y. Mori, T. Sekiguchi, Y. Okuyama, J. Yugami and H. Kume: Tech. Dig. of VLSI Symp. 2001 (2001) 115.

[8] P. J. McWhorter, S. L. Miller and W. M. Miller: IEEE Trans. Nucl. Sci. **37** (1990) 1682.

[9] P. T. Landsberg: J. Chem., Phys. **23** (1955) 1079.

[10] A. Yokozawa, A. Oshiyama, Y. Miyamoto and S. Kumashiro: IEDM Tech Dig. (1997) 703.

[11] P. W. Anderson: Phys. Rev. Let. **34** (1975) 953.

4

Data retention characteristics of flash memories via SILC

Summary

In this chapter, the data retention of flash memories via he stress-induced leakage currents (SILCs) are described. A model of the stress-induced leakage currents (SILCs) based on the inelastic trap-assisted tunneling (ITAT) is developed by introducing a trap with a deep energy level of 3.6eV from the bottom of the conduction band. This model can explain both of two field dependencies, i.e. a field dependence of the direct tunneling (DT) for A-mode SILC and that of the Fowler-Nordheim (FN) tunneling for B-mode SILC by analytical equations of a common form. For simple analytical equations, we introduce the most favorable trap position (MFTP), which gives the largest contribution to the leakage current. The trap area density for A-mode SILC of around 1×10^{10} cm^{-2} and the are density of the leakage paths for B-mode SILC of 5×10^{2} cm^{-2} were obtained by comparisons with the experimental results and the present model.

4.1 Introduction

Reliability of nonvolatile memories is mainly affected by endurance of dielectrics for repeated write/erase cycles. Stress-induced leakage currents (SILCs) can lead to a progressive loss of the charge stored in the floating gate after write/erase cycles, thus limiting the oxide thickness reduction in EEPROM devices [1, 2]. It is well known that SILC phenomena exhibit two components, namely the transient SILC due to charge detrapping and the steady-state leakage current [3]. While the transient SILC affects the retention characteristics at oxide thicknesses higher than 13nm [4], the steady-state SILC is a real concern in the oxide thickness reduction in EEPROM devices. For this reason, a detailed physical description of the steady-state SILC is of utmost importance to design devices of a future generation. In this paper, we simply call the steady-state SILC as A-mode SILC.

According to the widely accepted interpretation, A-mode SILC is explained in terms of the trap-assisted tunneling [5], that is a two-step tunneling via a defect state in the oxide bulk. Despite this general agreement, the physical details of such a process have not been clearly understood. In a neutral-trap-assisted tunneling model [6], the oxide defects were assumed to behave as acceptor-like states. Recently, interpretations of A-mode SILC have been proposed in terms of a positive-trap-assisted tunneling process [7]-[10]. A detailed investigation of the capture and tunneling process involved in SILC has been still required.

A-mode SILC has been shown to lead to a soft breakdown (SBD) in oxides thinner than 5.5 nm [11]. The sudden loss of the insulating property of thin gate SiO_2 films used in a metal-oxide-semiconductor (MOS) structure is one of most important failure mechanisms. In addition to an abrupt breakdown (ABD), SBD [11]-[13] is the failure mechanisms of gate insulators for the state-of-art MOS technology. The main features of SBD are a huge leakage current in the direct tunneling voltage range, and a significant increase of the noise level. Because of the low-field leakage associated with SBD, terminology of B-mode SILC has also been used in the literature [14]. Some authors prefer to refer to SBD rather than to B-mode SILC because SBD and ABD are linked to a localized conduction mode. On the other hand, SILC is related to a mostly uniform increase of the current through the oxide [15]. In this paper we use terminology of A-mode and B-mode SILCs because both would be caused by the same defect been verified in this paper.

It is widely accepted that B-mode SILC consists in the formation of a low resistive path that connects the electrodes and that this conduction path is somehow related to the defects generated during write/erase cycles. It is not clear, however, what

kind of microscopic structure develops within the oxide at a spot of B-mode SILC and, hence, what theoretical approach could be suitable to deal with the conduction mechanism. In this regard, several models have been proposed and they are also mainly based on tunneling [13], [16], hopping [17], and percolation [18] mechanisms. A systematic experimental approach reveals striking features of current-voltage characteristics in B-mode SILC [19], [20]. However, the simple model to quantitatively describe the current-voltage characteristics has not been developed yet.

Recent studies suggest that A-mode SILC is caused by an inelastic trap-assisted tunneling (ITAT) [21-24]. The ITAT-based SILC model can qualitatively explain current-voltage characteristics of A-mode SILC [22], [23]. However, there are no quantitative ITAT-based models of A-mode SILC which can simply predict current-voltage characteristics. For B-mode SILC, models have never been discussed based on ITAT. There are also no quantitative models of B-mode SILC which can simply predict current-voltage characteristics.

While A-mode SILC shows a field dependence of the direct tunneling (DT), B-mode SILC shows that of the Fowler-Nordheim (FN) tunneling. The empirical models of SILCs are used to predict the retention time of the flash memory using phenomenological parameters such as the “reduced barrier height” [21] for the FN-field dependence or the “reduce oxide thickness” for the DT-field dependence. However, two kinds of field dependence result in such different behaviors that we should understand the mechanisms behind them.

In this work, we have developed analytical equations of a common form for both A-mode and B-mode SILCs by introducing the most favorable trap position (MFTP) which can quantitatively explain current-voltage characteristics. The equations are described with the area density of the trap for A-mode, that of the multiple-trap path for B-model and the same trap energy level of 3.6eV for both modes. The present model explains the difference of field dependences between A-mode and B-mode SILCs.

4.2 Experimental Details

For A-mode SILC, we refer to the experimental results reported by R. Moazzami and C. Hu [6]. MOS capacitors with 6.5nm oxide and n^+ polysilicon gate were processed on p-type silicon. The oxide was grown by the standard CMOS process. The capacitors were stressed with a constant voltage equivalent to 9.5MV/cm (Fowler-Nordheim stress). The electron fluences of the stress were varied between 0.1C/cm^2 and 10.0 C/cm^2.

For B-mode SILC, we used the experimental results of the samples fabricated at another line. MOS capacitors with 5.0nm oxide and n^+ polysilicon gate were processed on p-type silicon. The oxide was also grown by the standard CMOS process. The capacitors were stressed with a constant current equivalent to 0.1 A/cm^2. The time to the appearance of the soft-break down was ~100 sec, and the total fluence was ~10.0C/cm^2. The constant current stress was completed just after the soft-break down started. The maximum applied field during the I-V measurement after stress was limited to 7MV/cm to minimize the additional stress for both A-mode and B-mode.

Figure 4-1 shows A-mode and B-mode SILCs as a function of an applied electric field to the oxide, E_{ox}. While A-mode SILC shows a field dependence of the direct tunneling (DT), B-mode SILC shows that of the Fowler-Nordheim (FN) type. It is also distinct aspect that B-mode SILC is several orders of magnitude larger than A-mode SILC.

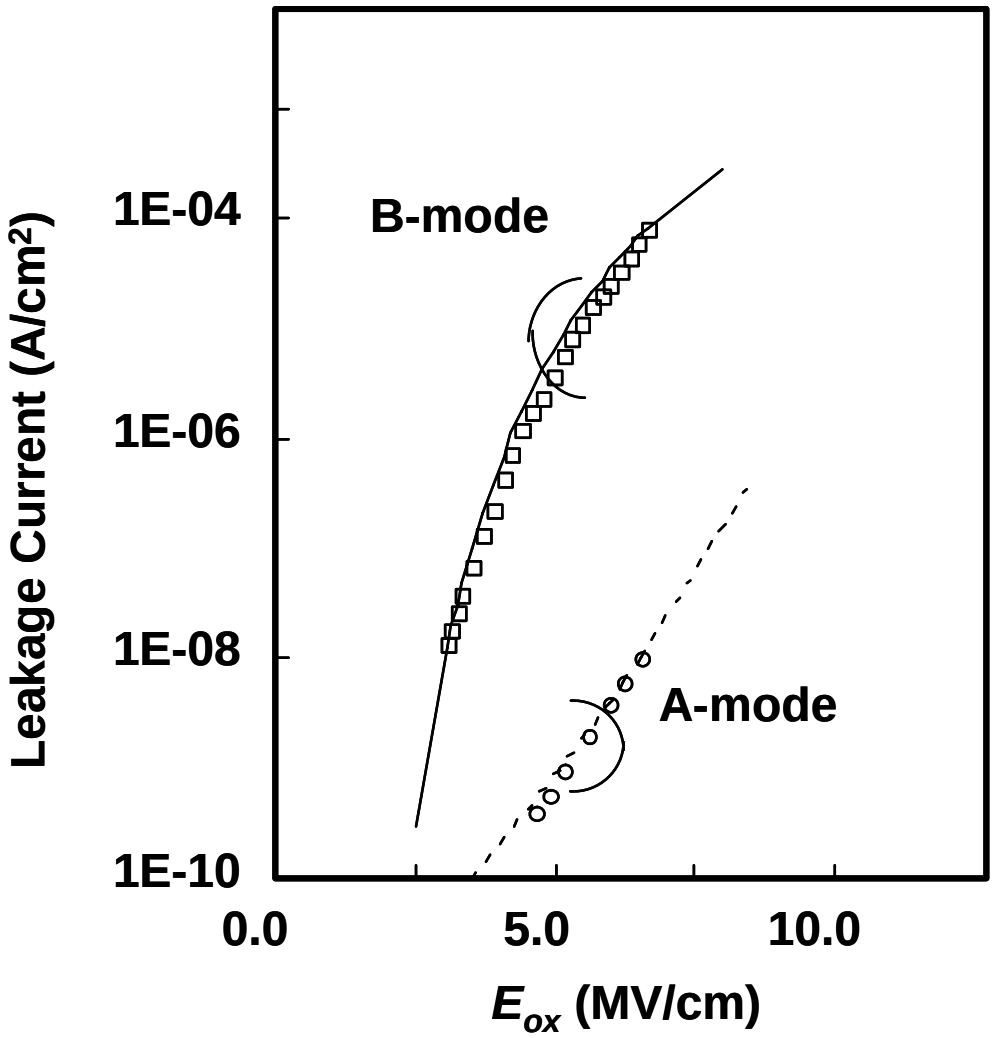


Fig. 4-1. Field dependence of A- and B-mode SILCs. A-mode SILC shows the DT-field dependence and B-mode SILC shows the FN-field dependence.

4.3 Deep Trap SILC Model

Figure 4-2 shows schematic models of electron paths for SILCs through the oxide layer. In Fig. 4-2, the upper and the lower figures show the disposition and the energy level, respectively, for the traps, the anode, and the cathode. Here, $E_{g,C}$ is the band gap offset at the cathode-oxide interface, $E_{g,A}$ the band gap offset at the anode oxide interface, E_t the trap level measured from the bottom of the conduction band, T_{ox} the oxide thickness, X_t the trap position in the oxide, and the other parameters are explained below in this section. As shown in Fig. 4-2 (a), electrons from the cathode are captured by the trap via the virtual tunnel state, and then emitted to the anode for A-mode [22], [25]. The capture and emission of the electron is subject to the ITAT mechanism and can be described by the phonon-assisted Shockley-Read-Hall (SRH) model between the virtual tunnel state and the trap state [22-24]. As shown in Fig. 4-2 (b), electrons hop to successively adjacent traps (multiple-trap path) before being emitted to the anode by tunneling for B-mode [25].

The allowable region of the trap state for A-mode is shown in Fig. 4-2 (a). If E_t is deeper than the allowable region, the electron tunneling is prohibited because the trap level is outside the conduction band of the anode. If E_t is shallower than the allowable region, the current-voltage characteristics becomes the different from that of A-mode SILC. Therefore, the trap state should be located within the allowable region. The allowable region depends on X_t, T_{ox}, and E_{ox}. For T_{ox}=6.5nm, the allowable region are 2.0 [eV]< E_t <5.1 [eV] for E_{ox}=6 [MV/cm] and 1.3 [eV]< E_t <4.5 [eV] for E_{ox}=6 [MV/cm], respectively. Here, the trap position in the oxide, X_t, is assumed at the middle of the oxide, which will be verified below in this section. In this section, we tentatively use the value of E_t =4.0 [eV] for the model calculations for both A-mode and B-mode. The selected value of E_t within the allowable region does not change the qualitative results obtained in this section. The extraction of E_t by using the experimental results is described in the next section for both A-mode and B-mode.

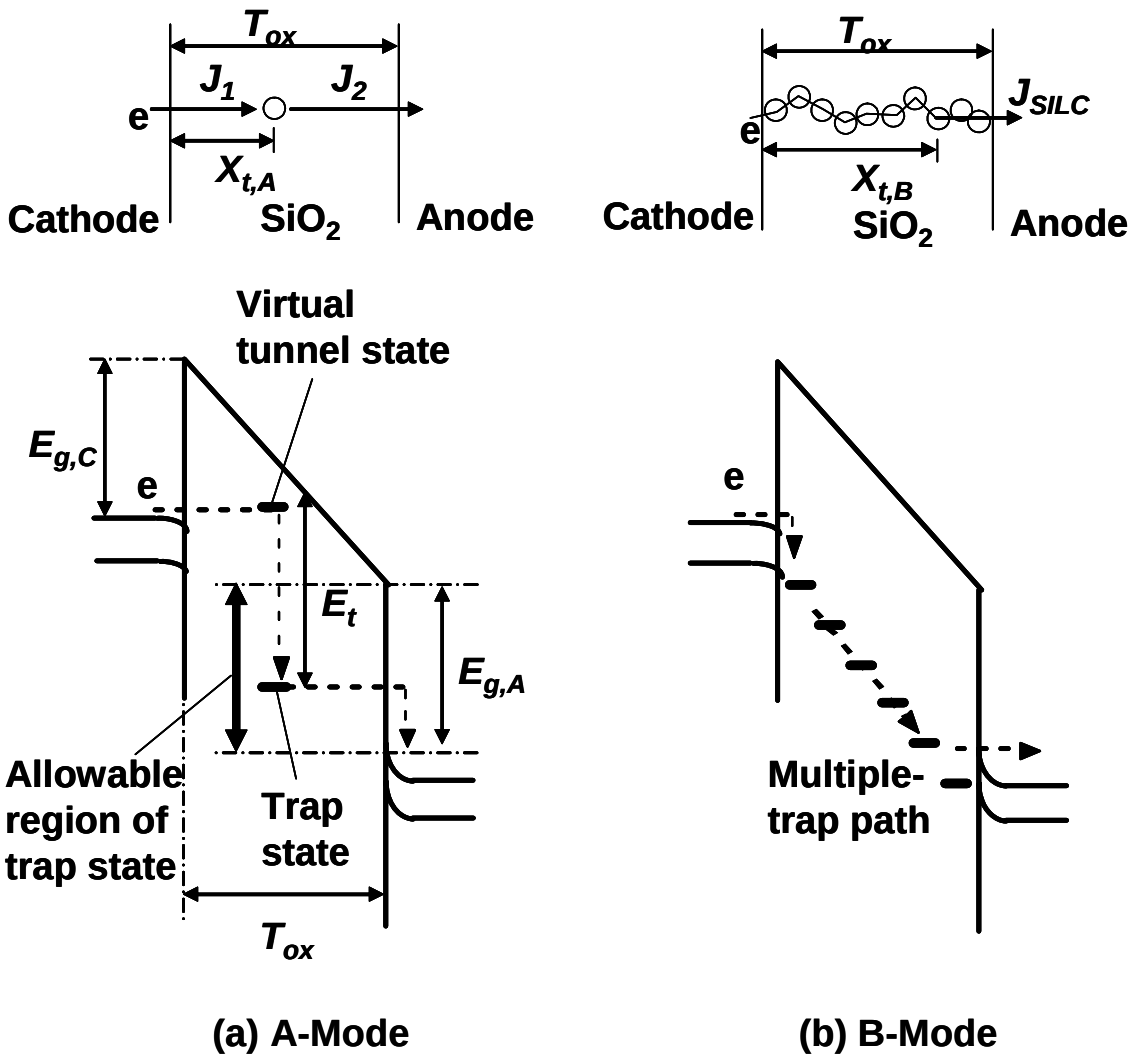


Fig. 4-2. Conduction mechanism of A- and B-mode SILC. (a) In A-mode, electrons at the cathode are captured by the trap, and are emitted to the anode by tunneling. (b) In B-mode, electrons hop to successively lower traps before being emitted to the anode by tunneling.

To simplify the theoretical equations for SILCs, we introduce "the most favorable trap position" (MFTP). For A-mode, we define the trap position which gives the largest contribution to the leakage current as the most favorable trap position. At the MFTP, the transition probability of electron from the cathode to the trap is equal to that from the trap to the anode. Based on the ITAT model, the rate equation of the tunneling electron is described with

$$q\frac{dn_t}{dt} = J_1 - J_2, \tag{1}$$

where $J_1 = (N_t - n_t)P_1$ and $J_2 = n_t P_2$. Here n_t is the area density of the electrons captured at trap, N_t is the area density of the traps, P_1 and P_2 are the electron current per one trap from the cathode to the trap and from the trap to the anode, respectively. At the steady state, i.e. J_1=J_2, the leakage current for A-mode SILC can be expressed as

$$J_{SILC} = N_t \frac{P_1 P_2}{P_1 + P_2}, \tag{2}$$

where

$$P_1 = \frac{q}{\tau}\exp\left\{-\frac{4}{3}\frac{(2m_{ox})^{0.5}}{\hbar}\frac{1}{qE_{ox}}\left(E_{g,C}{}^{3/2} - \left[E_{g,C} - qX_t E_{ox}\right]^{3/2}\right)\right\}, \tag{3}$$

$$P_2 = \frac{q}{\tau}\exp\left\{-\frac{4}{3}\frac{(2m_{ox})^{0.5}}{\hbar}\frac{1}{qE_{ox}}\left(E_t{}^{3/2} - \left[E_t - qE_{ox}(T_{ox} - X_t)\right]^{3/2}\right)\right\}. \tag{4}$$

Here, q is the electron charge, m_{ox} the electron effective mass in the oxide for the tunneling and τ the relaxation time. In this paper, we use the values for the above parameters as follows: $E_{g,C}$ =$E_{g,A}$ = 3.2 [eV], $(2m_{ox})^{0.5}/\hbar = 3.44\times10^7$ [V/cm eV$^{1.5}$], and $\tau = 1\times10^{-15}$ [s] [26]. Equations (3) and (4) are obtained by the WKB approximation for the electron tunneling through the trapezoidal potential.

Figure 4-3 shows the calculation results of the leakage currents normalized by N_t as a function of X_t and E_{ox}. The leakage currents are calculated with Eq. (2) in the case of T_{ox} = 6.5 [nm] and E_t = 4.0 [eV]. As shown in Fig. 4-3, the dominant contribution to

the leakage current is given by the traps at the symmetric position, which we have defined as the MFTP. The MFTP of A-mode SILC, $X_{t,A}$, is obtained by solving,

$$E_t^{3/2} - \left[E_t - qE_{ox}(T_{ox} - X_{t,A})\right]^{3/2} = E_{g,c}^{3/2} - \left[E_{g,c} - qE_{ox}(T_{ox} - X_{t,A})\right]^{3/2}. \quad (5)$$

Figure 4-4 shows the calculation results of $X_{t,A}$ as a function of E_{ox} with a parameter of three values of T_{ox}. For calculations, we use the values as follows: $E_{g,C} = E_{g,A} = 3.2$ [eV], $E_t = 4.0$ [eV], and T_{ox}=8.5, 6.5, 4.5 [nm]. The MFTP of A-mode, $X_{t,A}$, depends strongly on the oxide thickness, while it is almost independent of the oxide field, as shown in Fig. 4-4. In the inset of Fig. 4-4, $X_{t,A}$ is depicted as a function of the oxide thickness. Empirically, it is found that $X_{t,A}$=0.58T_{ox}–0.2[nm] at E_{ox} =5MV/cm.

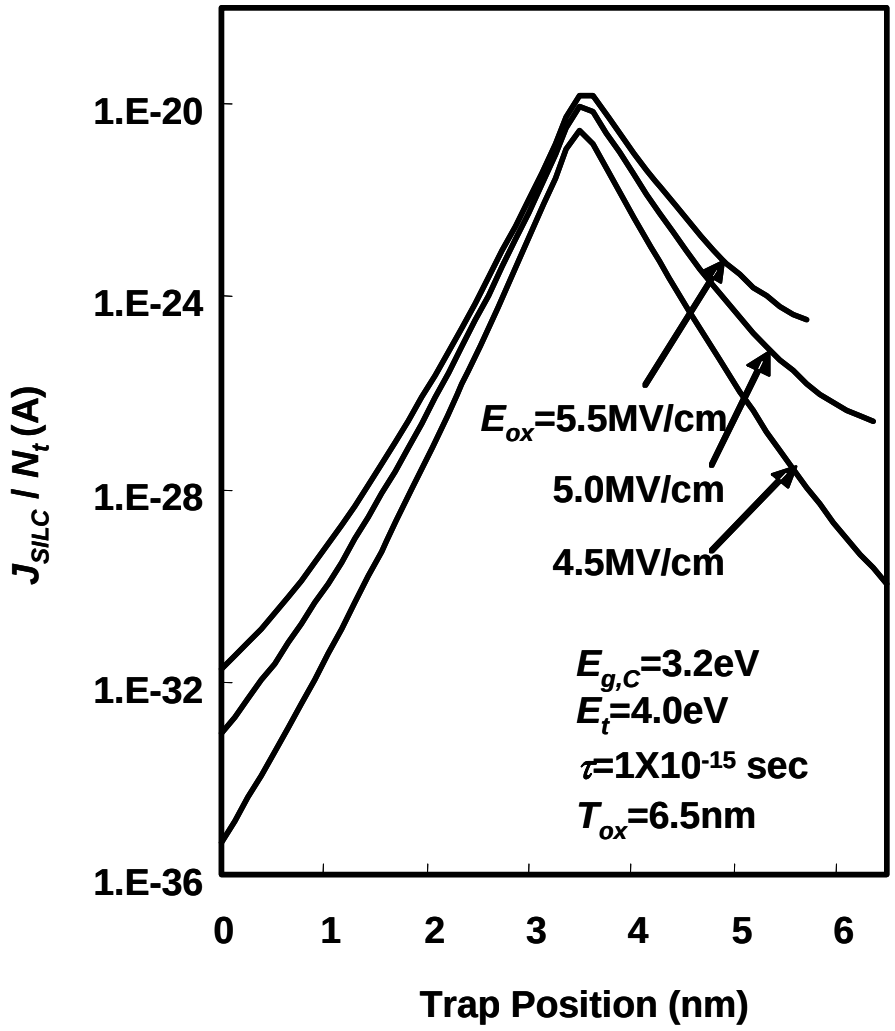


Fig. 4-3. The calculation results of the leakage currents normalized by N_t as a function of X_t and E_{ox}.

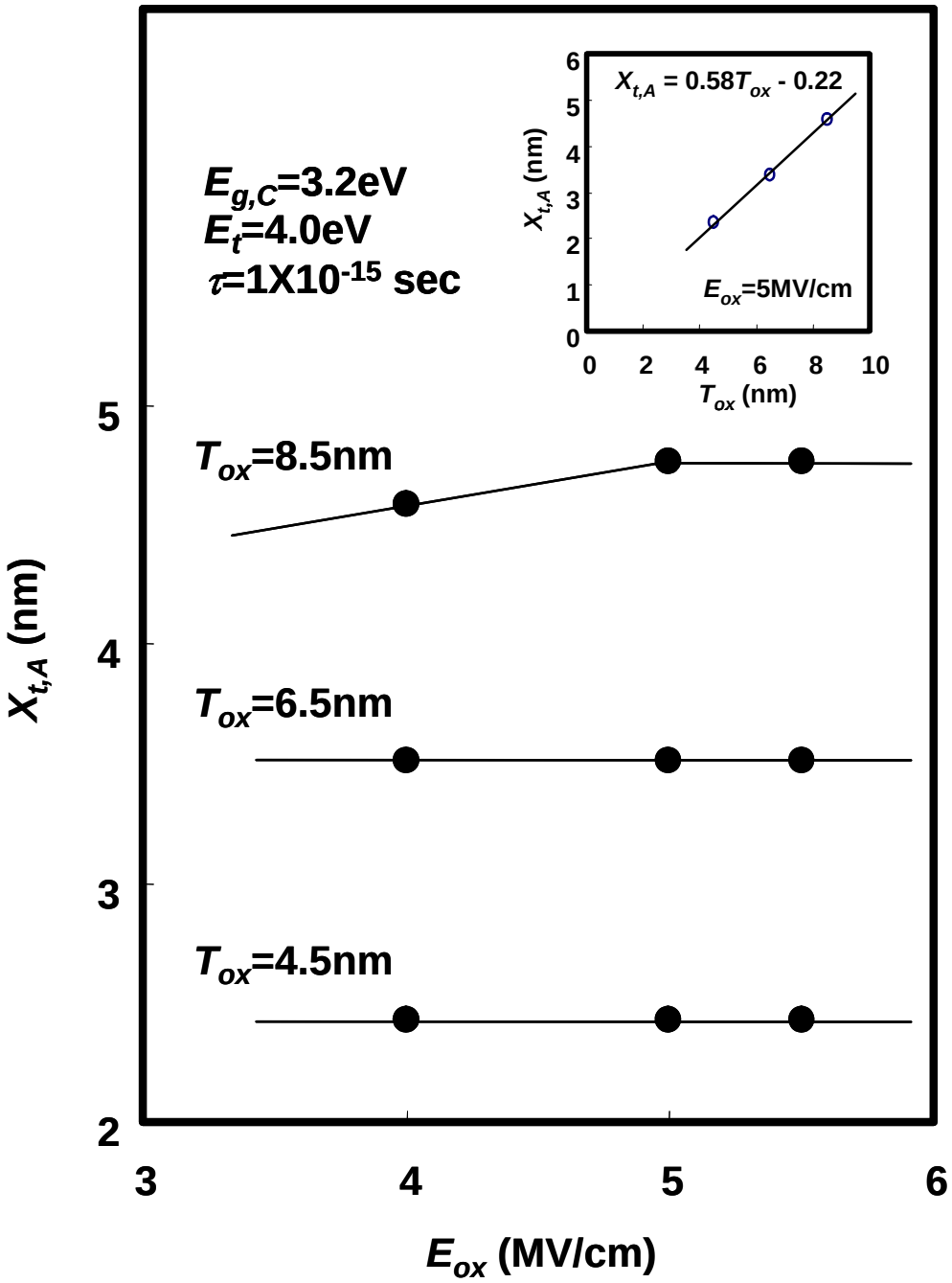


Fig. 4-4. The MFTP of A-mode SILC as a function of T_{ox} and E_{ox}.

For B-mode SILC, tunneling is considered to be the limiting process because the I-V curve is almost independent of temperature [25]. Figure 4-5 shows the electron conduction path in the energy diagram for the low and the high electric-field cases, respectively. In Fig. 4-5, the energy levels of the multiple-trap path, the anode and the cathode are depicted. Tunneling from the trap to the anode becomes the limiting process of the leakage current when the trap level, E_t, is deeper than $E_{g,A}$, as shown in Fig. 4-5. Based on the model shown in Fig. 4-5, electrons escape via tunneling from the trap whose level is within the conduction band of the anode. We define the MFTP for B-mode, $X_{t,B}$, as the trap position where the trap level is coincident with the conduction band edge of the anode, as shown in Fig. 4-5. The MFTP for B-mode SILC, $X_{t,B}$, for the low electric field is larger than that for the high electric field, as shown in Fig. 4-5. The MFTP for B-mode SILC, $X_{t,B}$, is simply obtained by

$$X_{t,B} = T_{ox} + (E_{g,A} - E_t)/qE_{ox}. \qquad (6)$$

Figure 4-6 shows the calculation results of $X_{t,B}$ as a function of E_{ox} with a parameter of T_{ox}. For calculations, we use the values as follows: $E_{g,C}$ =$E_{g,A}$ = 3.2 [eV], E_t = 4.0 [eV], and T_{ox}=8.5, 6.5, 4.5 [nm]. The $X_{t,B}$ shows a strong dependence on E_{ox}, while $X_{t,A}$ is almost independent of E_{ox} as shown in Fig. 4-4.

Finally, we can obtain a simple equation in a common manner for A-mode and B-mode SILCs by introducing the MFTP:

$$J_{SILC} = \alpha \frac{qN_t}{\tau} \exp\left\{ -\frac{4}{3} \frac{(2m_{ox})^{0.5}}{\hbar} \frac{1}{qE_{ox}} \left(E_t^{3/2} - \left[E_t - qE_{ox}(T_{ox} - MFTP) \right]^{3/2} \right) \right\}, \qquad (7)$$

where α = 0.5, MFTP= $X_{t,A}$ and N_t=$N_{t,A}$ for A-Mode, and α =1.0, MFTP= $X_{t,B}$ and N_t=$N_{t,B}$ for B-Mode. Here, $N_{t,A}$ is the area density of the trap for A-mode SILC, and $N_{t,B}$ is the area density of the multiple-trap path for B-mode SILC.

By using Eq. (7), both the DT-field dependence of A-mode and the FN-field dependence of B-mode can be explained. The MFTP, $X_{t,A}$, is almost independent of the oxide field as shown in Fig. 4-4. Therefore, Eq. (7) becomes a simple form for the direct tunneling of the trapezoidal potential. On the other hand, Fig. 4-6 shows the calculation

results of $X_{t,B}$ as a function of E_{ox} with a parameter of three values of T_{ox}. As shown in Fig. 4-6, the MFTP, $X_{t,B}$, is strongly dependent on the oxide field via Eq. (6). Substituting Eq. (6) to Eq. (7) leads to the FN-tunneling equation described by:

$$J_{SILC} = \frac{qN_{t,B}}{\tau} \exp\left\{ -\frac{4}{3} \frac{(2m_{ox})^{0.5}}{\hbar} \frac{1}{qE_{ox}} \left(E_t^{3/2} - E_{g,A}^{3/2} \right) \right\}. \quad (8)$$

This equation can be considered to be equivalent to that of the FN tunneling with the reduced barrier height of $E_t^{3/2}$-$E_{g,A}^{3/2}$.

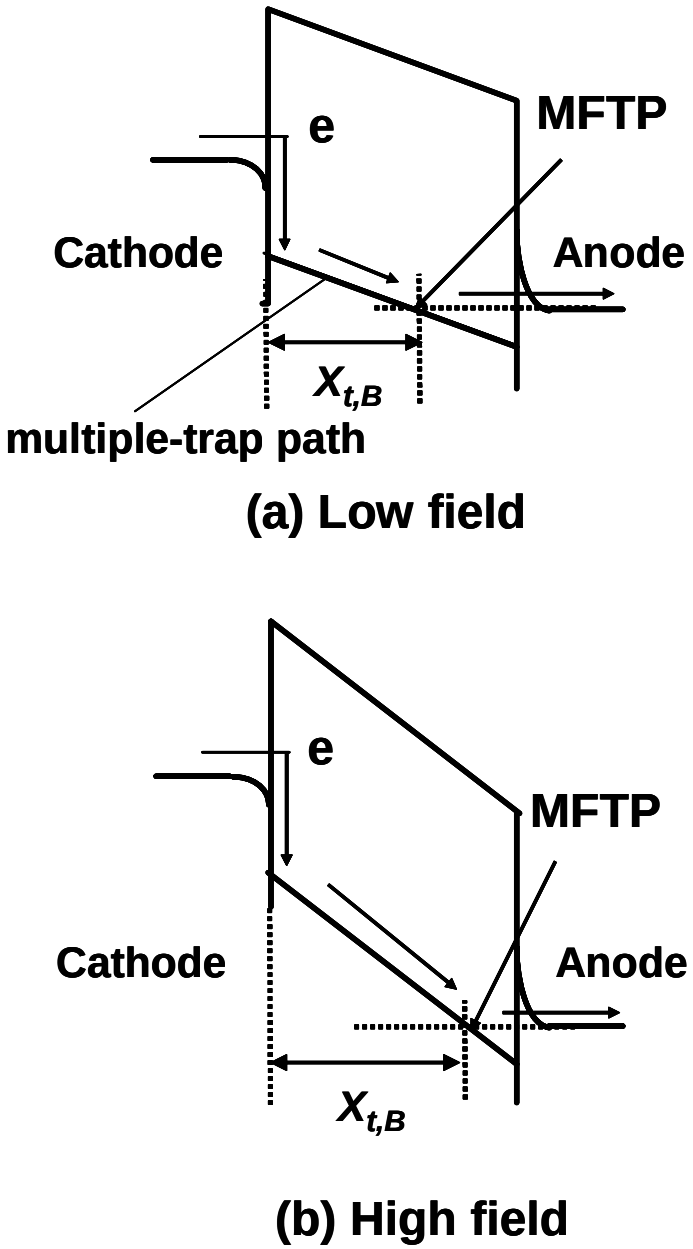


Fig. 4-5. The schematic illustrations of MFTP for B-mode SILC.

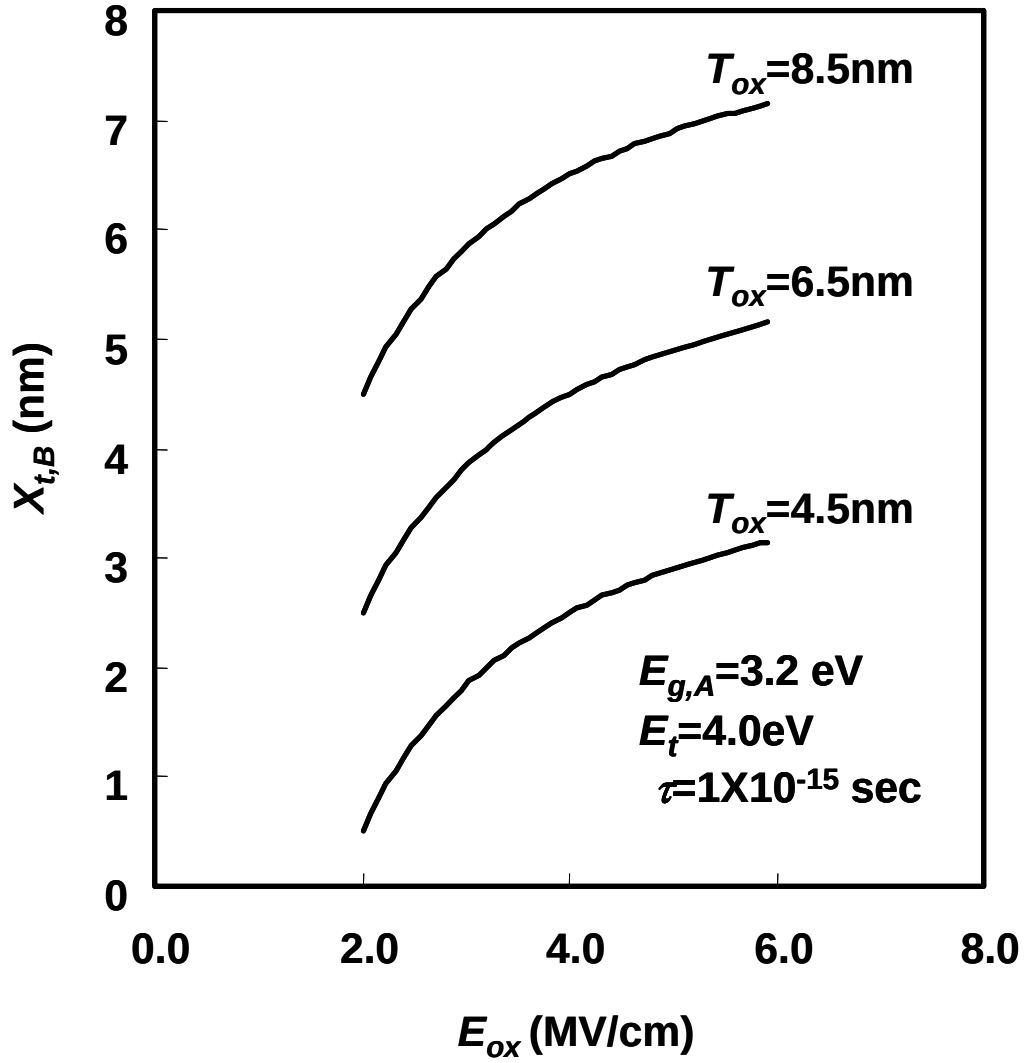


Fig. 4-6. The MFTP of B-mode SILC as a function of T_{ox} and E_{ox}.

4.4 Extraction of E_t and N_t from the experimental data

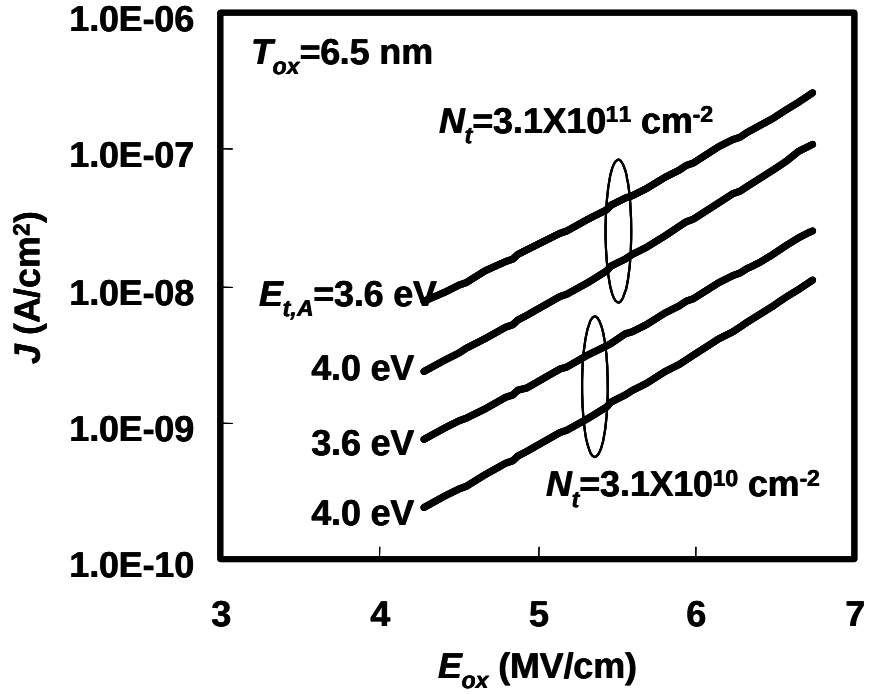
We can determine the model parameters, E_t and N_t, by comparing between the model calculations and the experimental results. Figure 4-7 shows the calculation results of SILCs as a function of E_{ox} with parameters of E_t and N_t. As shown in Fig. 4-7 (b), the slopes of log(J) for B-mode strongly depend on E_t and do not depend on $N_{t,B}$, while the slopes of log(J) for A-mode are independent of both E_t and $N_{t,A}$, as shown in Fig. 4-7 (a). Therefore, we can separately extract E_t and $N_{t,B}$ by fitting the model to the experimental results for B-mode SILC. The procedures of the parameter extraction are as follows. For the first step, we extract the values of E_t and $N_{t,B}$ by using the experimental results of B-mode. Then, we extract the value of $N_{t,A}$ for A-mode by using the determined value of E_t for B-mode. Here, we use the value as $E_{g,C}$= $E_{g,A}$=3.2 [eV] and τ =1×10^{-15} [sec] for the model calculation [26].

Figure 4-8 shows the leakage current of B-mode as a function of the reciprocal of E_{ox}. The comparison is made in Fig. 4-8 between the model and the experimental results for B-mode. As explained in Fig. 4-7 (b), we extracted the value of $E_{t,B}$ from the slope of log(J)-vs- E_{ox}^{-1} plot and then the value of $N_{t,B}$ by fitting the model to the experimental results. The resultant values of E_t and $N_{t,B}$ are 3.6 [eV] and 500 [cm^{-2}], respectively. $N_{t,B}$ for B-mode is the area density of the multiple-trap path, whose validity of the extracted value will be discussed in the later section.

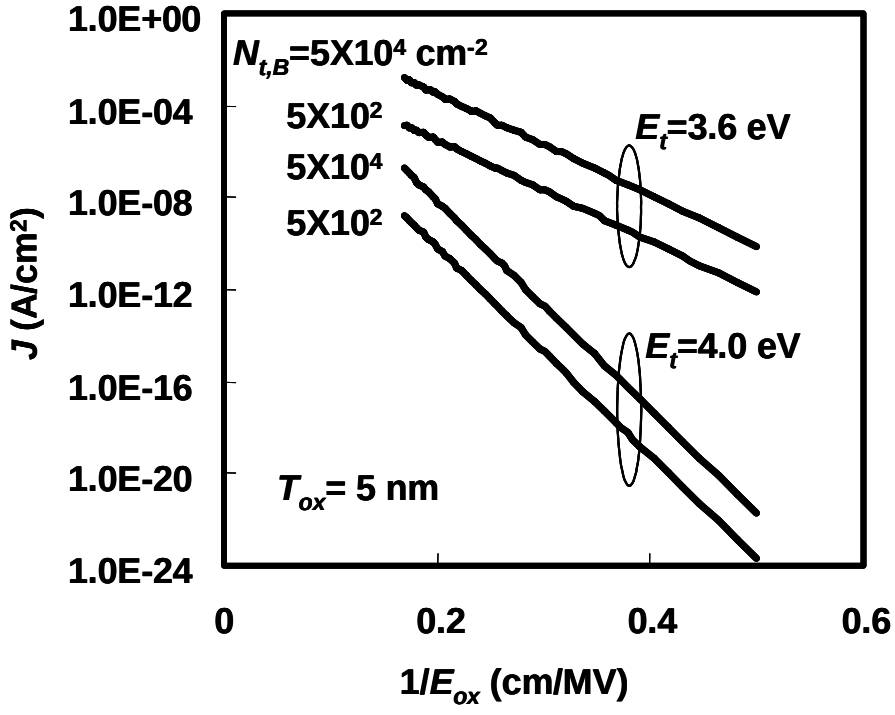
Figure 4-9 shows the leakage current of A-mode as a function of E_{ox}. The comparisons are made in Fig. 4-9 between the experimental results of A-mode and the model curves calculated with $E_{t,A}$=3.6 [eV] extracted from B-mode. The values of $N_{t,A}$ are extracted for each of the injected charge, Q_{inj}, as shown in Fig. 4-9. Extracted values of $N_{t,A}$ are as follows: $N_{t,A}$ =3.1×10^{10}, 1.9×10^{10}, 1.3×10^{10}, 0.75×10^{10} and 0.38×10^{10} [cm^{-2}] for Q_{inj}=10, 3, 1, 0.3, 0.1[C/cm^2], respectively. The dependence of the trap area density, $N_{t,A}$, on the injected charge, Q_{inj}, is shown in the inset of Fig. 4-9 and experimentally given by,

$$N_{t,A} = 1.4\times10^{10} + 1.1\times10^{10}\log(Q_{inj}). \qquad (9)$$

The origin of log(Q_{inj}) dependence of $N_{t,A}$ is the remaining issues.



(a) A-mode



(a) B-mode

Fig. 4-7 Calculation results by the present model for (a) A-mode SILC and (b) B-mode SILC as a function of trap levels and area densities.

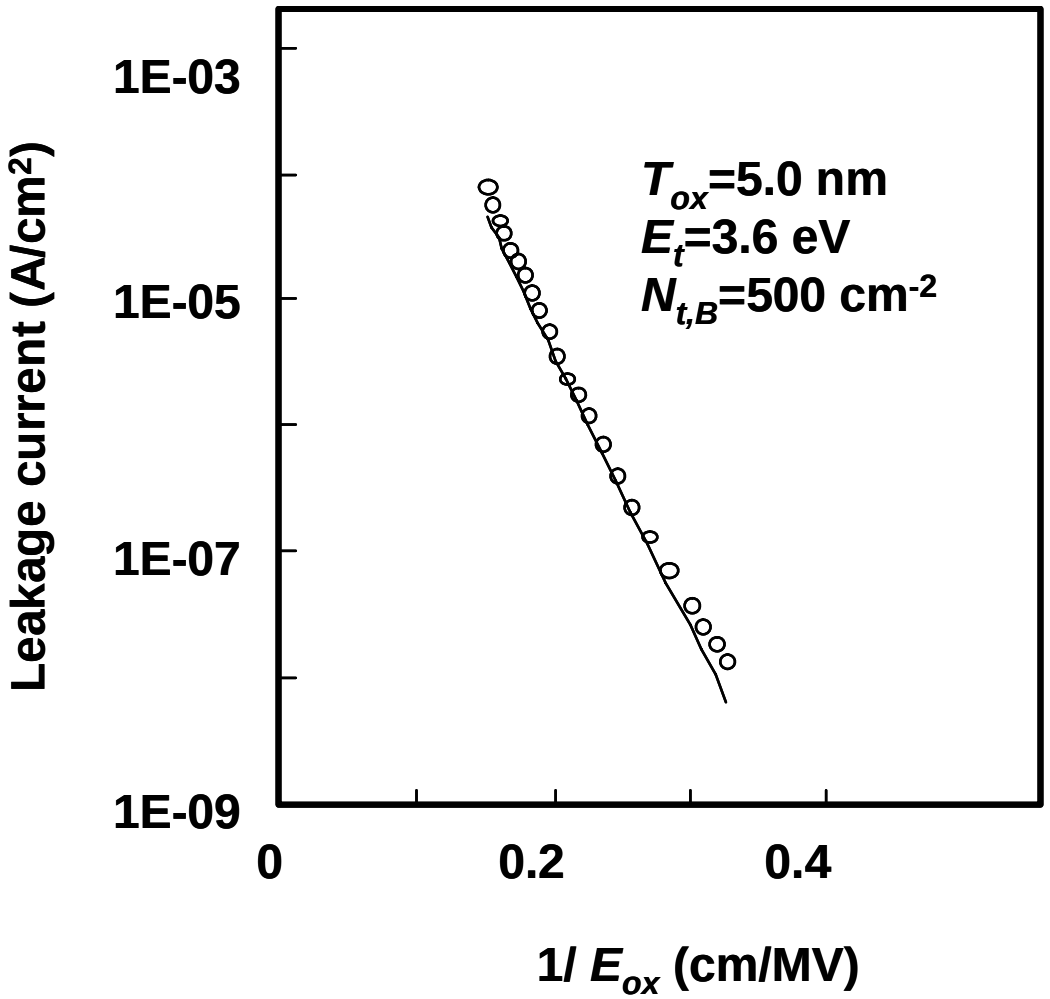


Fig. 4-8. The leakage current of B-mode as a function of the reciprocal of E_{ox}.

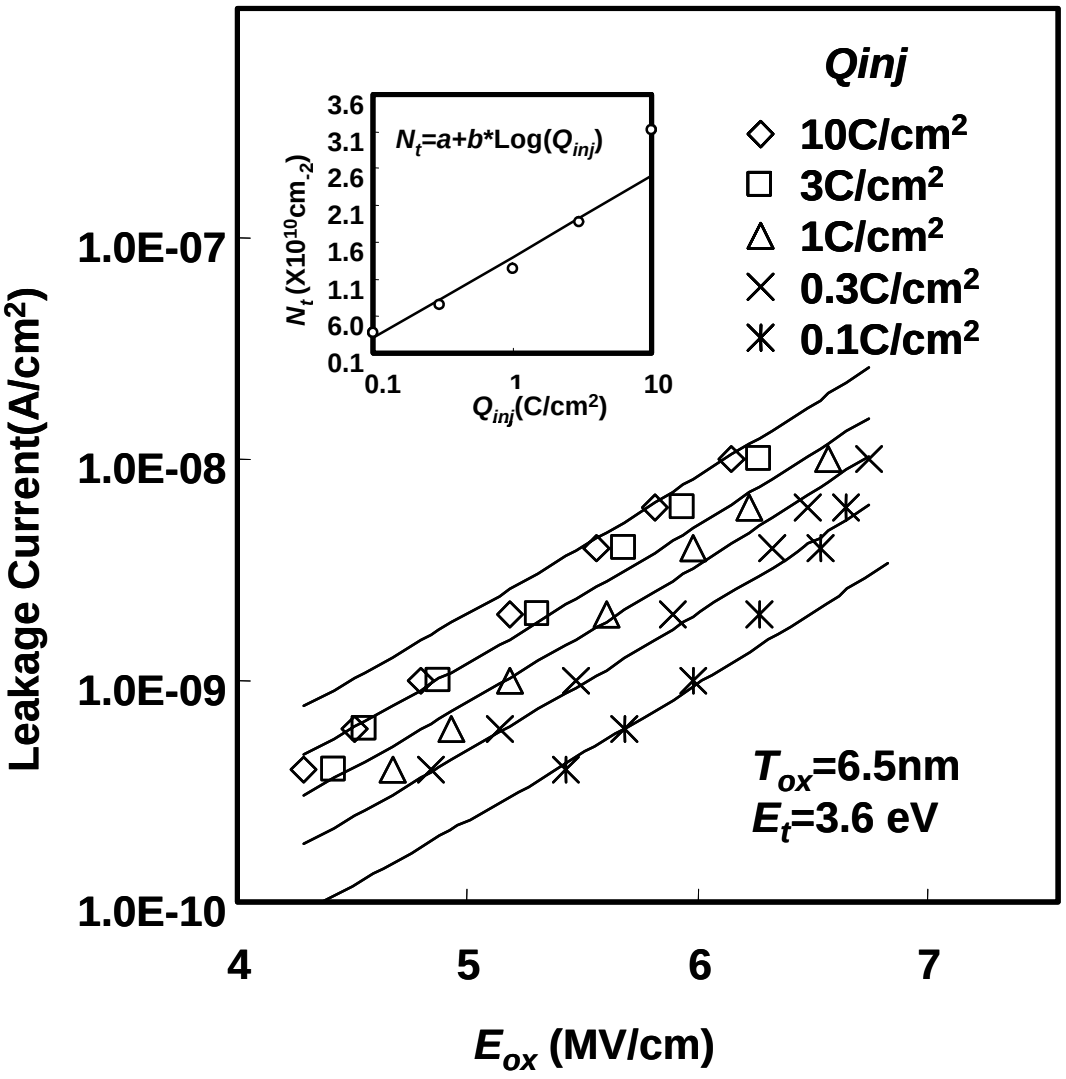


Fig. 4-9. The leakage current of A-mode as a function of E_{ox}.

4.5 Discussion

In the preceding section, the area density of the multiple-trap path for the B-mode SILC has been estimated to be 500 cm^{-2}, which is considered to be the density of the weak spots in the oxide. The density of the weak spots can be estimated using the charge-to-breakdown, Q_{BD}, distribution for the MOS capacitors of the small area. When the area of the MOS capacitor is small, the probability that each capacitor includes the weak spots will be very small. Suppose that we measure the distribution of Q_{BD} for N_c capacitors of the small area, S_c, N_{tail} samples show the extreme small Q_{BD}. Here, we call the samples which show the extreme small Q_{BD} as *tail samples*. By using these results, $N_{t,B}$ can be estimated as $N_{t,B} = N_{tail} / S_c N_c$. The Q_{BD} distributions of capacitors with area of 5.25×10^{-4} cm^2 have been shown, in Ref. [28]. Three samples out of 25 samples were *tail samples*. Therefore, from these experimental results, we obtains values as N_c =25, N_{tail} =3 and S_c $=5.25\times10^{-4}$ cm^2. By using $N_{t,B} = N_{tail} / S_c N_c$, we can estimate the area density of the weak spots of 228 cm^{-2}. This value is in the same order of magnitude as that with the area density of the weak spots obtained with our model, 500 cm^{-2}, which verifies our model.

As shown in the inset of Fig. 4-10, electrons from the cathode are captured by the trap via the virtual tunnel state, and then emitted to the anode for A-mode SILC. Therefore, electrons lose the energy corresponding to the difference between the virtual state and the trap state. Figure 4-10 shows the energy losses during the tunneling as a function of E_{ox} with a parameter of T_{ox}, which are calculated with the following equation:

$$\Delta E = (E_t + qE_{ox}X_{t,A}) - E_{g,c}\,. \tag{10}$$

As shown in Fig. 4-10, the energy loss for the case of T_{ox}=5.8 [nm] and E_{ox}=9.0 [MV/cm] is approximately estimated 3.1 [eV].

The energy loss of the electrons in A-mode SILC can be evaluated directly using the quantum yield of the impact ionization, γ [22]. The value of γ is determined by the ratio of the source (hole) current, I_s, to the gate (electron) current, I_g. The procedure of the measurement as follows. After the initial I_s and I_g are measured, the FN stress is applied to MOSFETs in order to cause A-mode SILC. Subsequently, the measurement of I_s and I_g are performed once more. While the incremental current in I_g before and after the stress, ΔI_g, corresponds to the hole currents generated by electrons in A-mode SILC, ΔI_s. Thus, the quantum yield of electrons in SILC is represented by the ratio of

the incremental currents, ΔI_s / ΔI_g. The electron energy as a function of γ is obtained by using the measurement of γ for FN tunneling [22]. With this method, the energy loss for the case of T_{ox}=5.8 [nm] and E_{ox}=9.0 [MV/cm] is approximately 1.5 [eV], which is almost half of the value calculated by using proposed model.

The reason for the above difference in the estimated energy loss would be explained as follows. Before and after applying the stress, the source current, I_s, under the effective field below 9MV/cm is difficult to measure which means the actual value of ΔI_s would become much smaller value than that of the reported ΔI_s. Therefore, γ would be estimated as much smaller values than the actual values, which corresponds to the much smaller estimation of energy loss.

Finally we discuss about the origin of the trap with $E_{t,A}$= $E_{t,B}$ =3.6 [eV]. Z. -Y. Lu et. al. have reported ab-initio total-energy calculations of O vacancies in several amorphous SiO_2 supercells [29]. The vast majority (~ 80%) of O vacancies are not bistable. Instead, the dimer configuration is the only stable configuration in both neutral and positively charged states after capture of a hole. The dimer energy level containing one or two electrons is shallow. Roughly 12% of O vacancies are bistable like those in quartz. In neutral states, the dimer configuration is stable. In positively charged states after capture of a hole, one of the adjoining Si atoms relaxes back past the plane defined by its three O neighbors and bonds with another network O atom. The latter becomes threefold coordinated whereas the puckered Si atom becomes fourfold coordinated. The localized energy level now become quite deep, nearly in the middle of the SiO2 energy gap. The trap with $E_{t,A}$= $E_{t,B}$ =3.6 [eV] would be attributed to this fourfold puckered configuration.

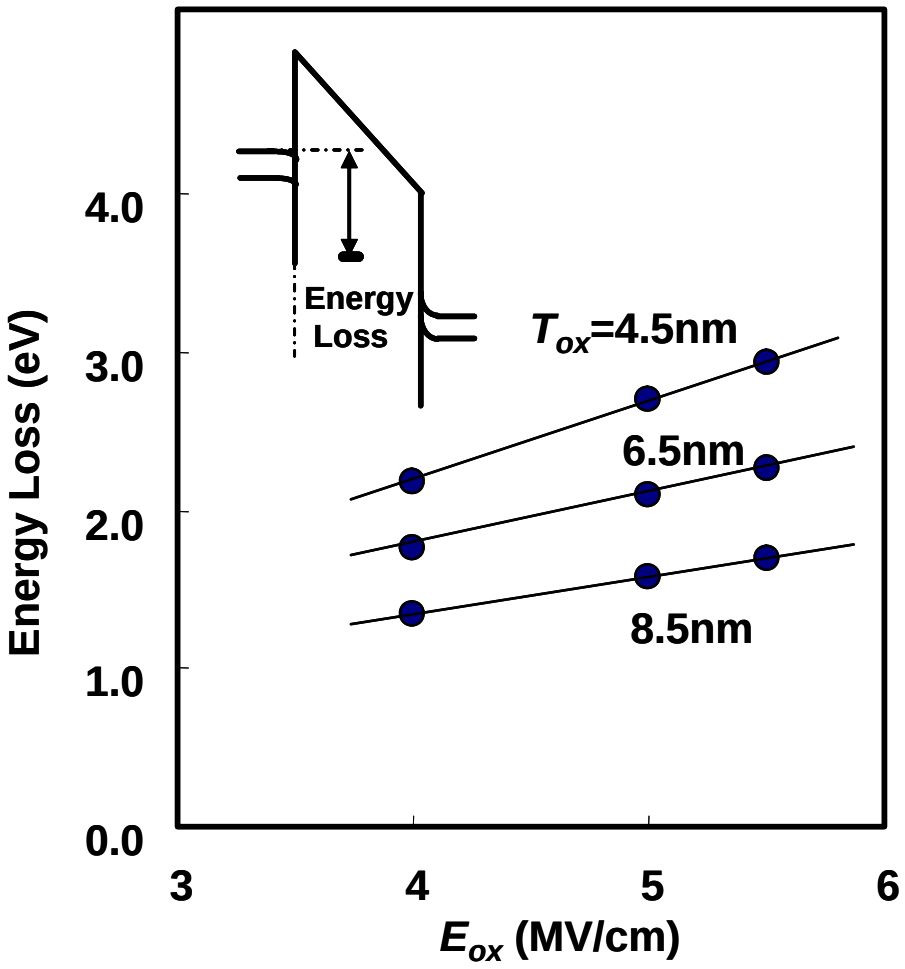


Fig. 4-10. The calculation results of the energy loss as a function of T_{ox} and E_{ox}.

4.6 References

[1] D. A. Baglee and M. C. Smayling: IEDM Tech. Dig. (1985) 624.

[2] K. Naruke, S. Taguchi, and M. Wada: IEDM Tech. Dig. (1988) 424.

[3] D. Ielmini, A. S. Spinelli, M. A. Rigamonti, and A. L. Lacatia: IEEE Trans. Electron Devices **47** (2000) 1258.

[4] M. Kato, N. Miyamoto, H. Kume, A. Satoh, T. Adachi, M. Ushiyama, K. Kimura: IEDM Tech Dig. (1994) 45.

[5] G. H. Paker and C. A. Mead: Appl. Phys. Lett. **14** (1969) 21.

[6] R. Moazzami and C. Hu: IEDM Tech. Dig. (1992) 139.

[7] N. Matsukawa, S. Yamada, K. Amemiya, and Hazama: IEEE Trans. Electron Devices **43** (1996)1924.

[8] G. J. Hemink, K. Shimizu, S. Aritome, and Shirota: Proc. IRPS (1996) 117.

[9] T. Wang, N. –K. Zous, J. –L. Lai and C. Huang: IEEE Trans. Electron Devices **19** (1998) 411.

[10] D. Ielmini, A. S. Spinelli, M. A. Rigamonti, and A. L. Lacatia: IEEE Trans. Electron Devices **47** (2000) 1266.

[11] M. Depas T. Nigam, and M. Heyns: IEEE Trans. Electron Devices **43** (1996) 1499.

[12] K. Okada: Jpn. J. Appl. Phys. **36** (1997) 1434.

[13] S. Lee, B. Chao, J. Kim, and S. Choi: IEDM Tech. Dig. (1994) 605.

[14] K. Okada, S. Kawasaki, and Y. Hirofuji: Ext. Abst. SSDM (1994) 565.

[15] B. Ricco, G. Gozzi, and M. Lanzoni: IEEE Trans. Electron Devices **45** (1998) 1554.

[16] T. Yoshida, S. Miyazaki, and M. Hirose: Ext. Abst. SSDM (1996) 539.

[17] K. Okada and K. Taniguchi: Appl. Phys. Lett. **70** (1997) 351.

[18] M. Houssa, T. Nigam, P. Mertens and M. Heyns: Appl. Phys. Lett. **73** (1998) 514.

[19] E. Miranda, R. Rodriguez, J. Sune, M. Nafria, and X. Aymerich: Proc. IRPS (1998) 42.

[20] E. Miranda, J. Sune, R. Rodriguez, M. Nafria, X. Aymerich, L. Fonseca and F. Campabadal: IEEE Trans. Electron Devices **47** (2000) 82.

[21] J. D. Blauwe , J. Van Houdt, D. Wellekens, R. Degraeve, Ph. Roussel, L. Haspeslagh, L. Deferm, G. Groeseneken and H. E. Maes: IEDM Tech. Dig. (1996) 343.

[22] S. Takagi, N. Yasuda and A. Toriumi: IEDM Tech. Dig. (1996) 323.

[23] E. Rosenbaum and L. F. Register: IEEE Trans. Electron Devices, **44** (1997) 317.

[24] K. Sakakibara, N. Ajika, M. Hatanaka and H. Miyoshi: Proc. IRPS (1996) 100.

[25] K. Okada: Proc. Int. Symp. on VLSI (1997) 143.

[26] I. Lundstorm and C. Svensson: J. Appl. Phys. **43** (1972) 5043.

[27] S. Yamada, K. Amemiya, T. Yamane, H. Hazawa and K. Hashimoto: Proc. IRPS (1996) 108.
[28] T. Nigam, R. Degraeve, G. Groeseneken, M. M. Heyns and H. E. Maes: Proc. IRPS (1998) 62.
[29] Z. -Y. Lu, C. J. Nicklaw, D. M. Fleetwood, R. D. Schrimpf and S. T. Pantelides: Phys. Rev. Lett. **89** (2002) 285505-1.

5

Structure, operation and reliability of DRAMs

Summary

In this chapter, structures, operations, and reliability issues of DRAMs are described. After brief review of DRAMs, the high density memory cell structures, i.e., a stacked capacitor cell (STC cell) and a trench capacitor cell (trench cell), are described. Then, the operations of DRAMs are explained, including read, program and refresh operations. The refresh operation is unique for DRAMs. To reduce the power consumption and enhance the operation speeds, the refresh time should be as long as possible. The refresh time is determined the data retention characteristics of *tail* bits. The origin of the data leakage is p-n junction leakage currents. To understand the previous works and the present work, the generalized Shockley-Read-Hall recombination currents and gate induced drain leakage (GIDL) currents are described.

5. 1 Introduction

The continuing research and development (R&D) effort directed toward VLSI memory technology has led to memory LSIs with lower cost, smaller size, higher speed, and more ease of use. The ever-larger memory capacities with smaller memory cells have contributed to these advances of memory LSIs [1]. The standard, i.e., commodity or stand-alone, DRAMs have increased memory-chip capacity to 1 to 4 Gb at the R&D level, and 64 to 256 Mb at the volume-production level through the use of high-density 1-T cells (one-transistor one-capacitor cells) until 2000 year. In addition, the throughput has been boosted, as exemplified by a 1-Gb chip of 1.6-Gbytes/s by new memory-subsystem architectures [2, 3]. The embedded DRAM for a system-on-chip has also shown great potential for high speed applications. This potential was verified by recent developments such as a 3.7-ns-access 8-Mb chip [4] that surpasses the performance of the embedded SRAM by its exceptional throughput of 128 Gbytes/s and smaller chip area.

The chip area has been increasing by 1.5 times for each successive chip generation up to 64 Mbits [1, 5]. However, experimental 256-Mb and 1-Gb DRAM chips have begun to depart from the trend. This departure indicates that device miniaturization is running up against difficulties. Increased chip area is also compensated by increased wafer area, i.e., 300-mm wafer production. None the less, the shrinking chip area is an urgent task for every DRAM manufacturer in order to reduce bit costs.

This chip-shrink trend can be clearly seen in a succession of drastic reductions in chip area of 64-Mb DRAMs. A 64-Mb DRAMs with an area less than 40 mm^2 is in volume production using 0.18-μm technology. Development of 1-Gb DRAMs also followed this strategy which means that the memory cell must be scalable with maintaining sufficient storage capacitance. Thus, in the next section, we review two DRAM memory cells that are presently used in production, focusing both on their scalability for future high-density DRAMs and extendibility to embedded DRAMs.

5.2 High density memory cell structures

There are two types of memory cell: a stacked capacitor cell (STC cell) and a trench capacitor cell (trench cell). Both cells have advantages and disadvantages as discussed below [1]. Figure 5-1 and 5-2 show the cross section view and the layout of STC cell, respectively. Regarding the STC cell, the biggest problem is the height difference between the memory array and the peripheral logic area, as shown in Fig. 5-1. Since the storage electrode must be high enough to provide a sufficient capacitor area, the height difference has been increasing as memory cell area decreases. The problem has been solved by several process technologies that are all used in commercial DRAMs. High-density plasma produces deep contact holes, with an aspect ratio more than 10, between the metal wiring located above the memory array and the substrate, as shown in Fig. 5-1 [16]. Sophisticated memory capacitor electrodes [l], a hemi-spherical-grain poly-silicon electrode [6] that doubles capacitor area by utilizing a rugged surface, and a Ta_2O_3 dielectric film help to reduce capacitor height [7]. Reducing capacitor process temperature to around 700˚C [8] suppresses the damage to MOSFETs underneath the capacitors.

Figure 5-3 and 5-4 show the cross section view and the layout of the trench cell, respectively. In the trench cell shown in Fig. 5-3, since the trench capacitor is buried in the Si substrate, the height difference is eliminated. Thus, a metal wiring with a finer pitch could be used. In addition, MOSFETs both in the memory array and the peripheral circuits do not experience the heat treatment (around 800 ˚C) necessary for the capacitor formation. Therefore, MOSFETs design can be easily optimized if performance is taken as a first priority. Thus, the trench cell is considered more compatible than the STC cell to logic LSI processes [9, 10]. However, the trench must be quite deep, so the trench process is the most serious problem facing every chip generation. The reason why the deep trench is necessary is attributed to an inherently small storage node and a low-permitivity capacitor film. As clearly shown in Fig. 5-4, the storage node of the trench cell is restricted in size, because the trench pattern cannot be overlapped on the active regions where the MOSFETs' channels are formed. On the other hand, the storage node pattern of the STC cell in Fig. 5-2 can be laid out as large as possible, taking only lithographic minimum space into account, since storage node are formed above MOSFET. In addition, several high-permitivity films useful for an STC cell are difficult to be applied to the trench cell. For example, a Ta_2O_3 film cannot be used because the film is vulnerable to the heat treatment necessary for fabricating the trench cell. Thus, a conventional SiO_2/Si_3N_4 film must be used for trench cell, making the trench deeper and deeper. A trench with a depth of 7 μm could be formed in a chip with a size of 0.25 μm

[11]. However, trench etching with the size about 0.1μm and the aspect ratio of more 40 or 50 aiming Giga-bit DRAMs are serious challenges [12].

To implement DRAMs into logic LSIs, several problems in both STC and trench cells must be overcome. These problems stem from DRAM-specific device structures and the processes for making them. The self-aligned contact (SAC), which is essential to reduce word line (gate electrode) pitch (Fig. 5-1), is a typical example. During the SAC process, when contact holes of data lines and storage node plugs are opened, a Si_3N_4 film (which works as an etching stopper of word lines) must surround the word lines. On the other hand, present high-performance logic LSIs employs a salicide-gate (self-alignedly formed silicide gate) electrode without the SAC structure. This is because a salicide gate produces a lower gate resistance (less than 5 Ω/sq.) than a policide gate (i.e., stack of silicide and polysilicon) used in DRAMs. Since all the gate structures on a chip must be the same, if high-density DRAMs are required, policide gate must be used in embedded DRAMs at the expense of logic performance. If logic performance has priority over DRAMs density, loosely laid-out memory cells must be used since the SAC process cannot be applied. Thus, a tungsten (W) /polysilicon stacked-gate structure has been intensively studied because it provides a favorably lower gate resistance of less than 5 Ω/sq. [13], while SAC process is applied to the gate.

Silicidation of the source/drain strongly affects the memory-cell area and the logic performance. It was experimentally demonstrated that this silicidation severely degrades data-retention characteristics probably because of point defects induced by silicide growth into substrate [14]. Thus, at present, memory area is protected during silicidation of peripheral logic devices by using a protection mask. Otherwise, the source/drain must be deep enough to avoid the effect of the silicidation on junction characteristics. This means that the memory cell must be large enough to accommodate such a deep source/drain. It has been estimated that a cell area around three times larger than a conventional memory cell of the standard DRAMs is suitable for an embedded DRAMs, because smaller memory cell requires the DRAMs specific processes and structures that are not so compatible to logic LSI [10].

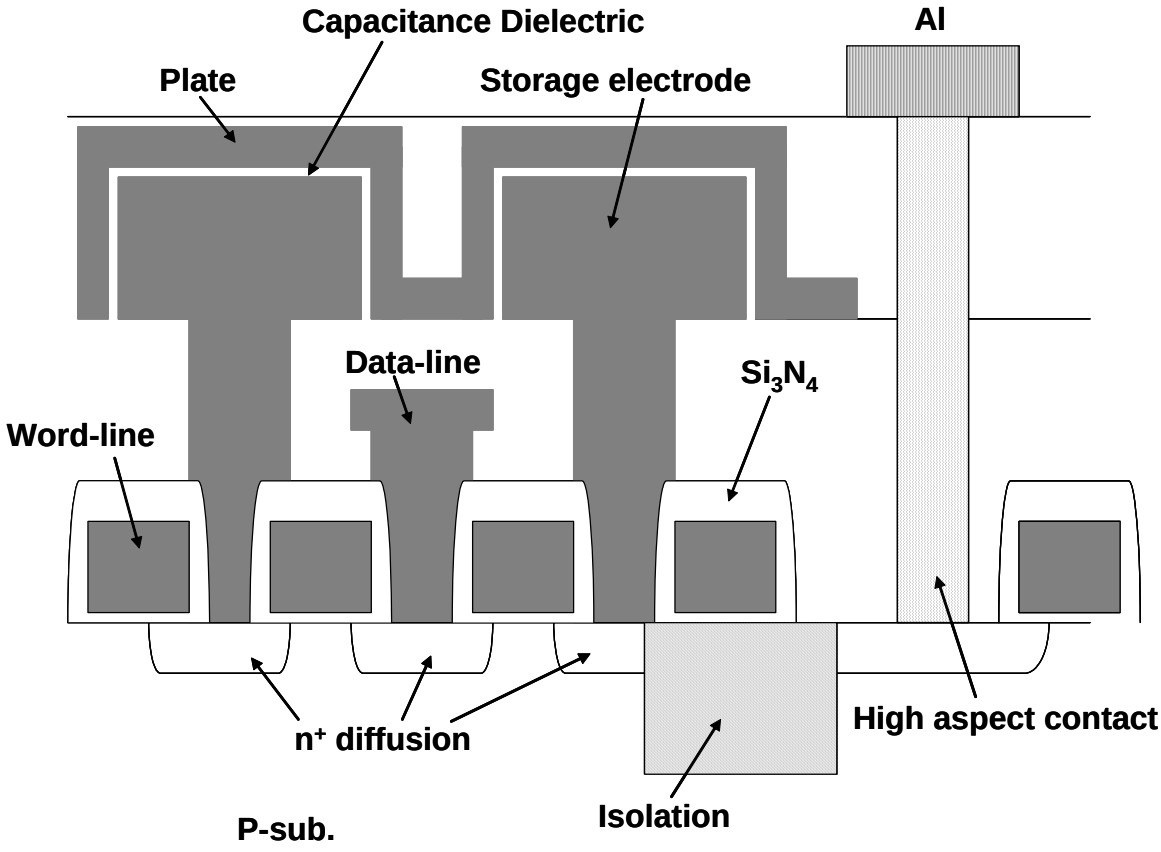


Fig. 5.1 Cross section view of STC cell.

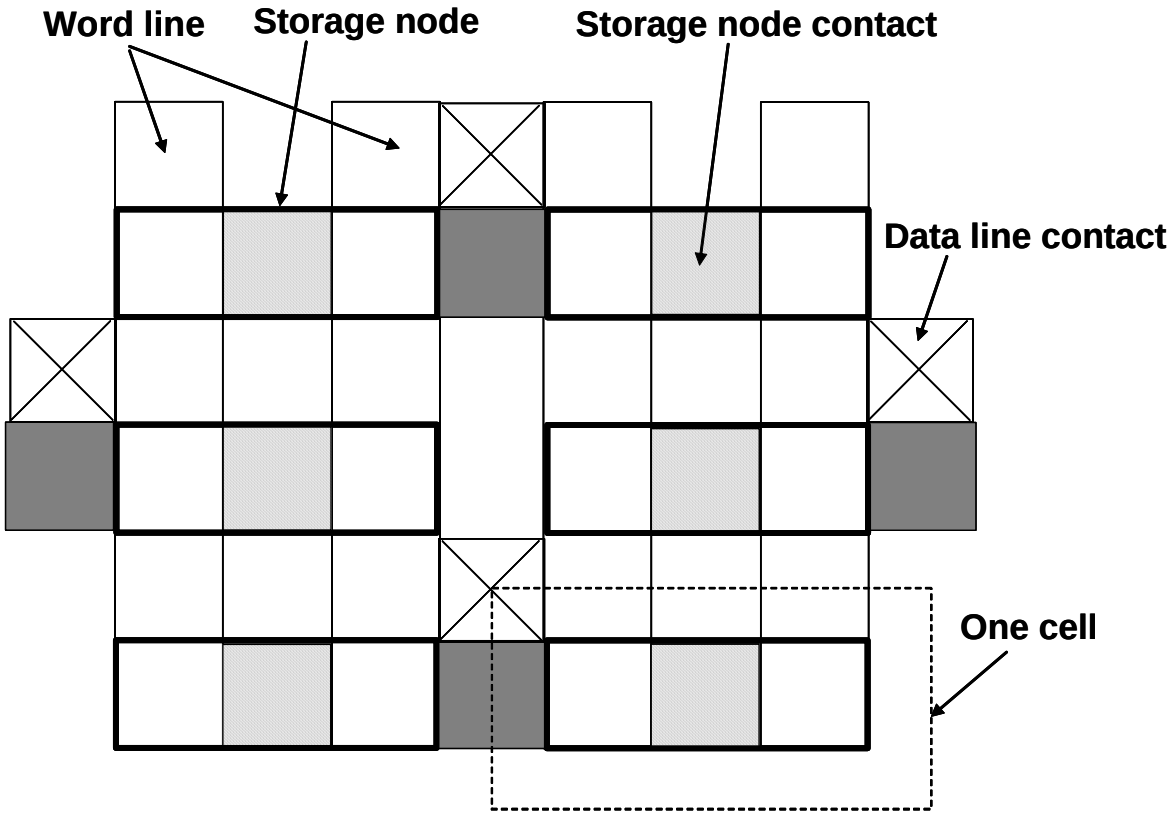


Fig. 5.2 Layour of STC cell.

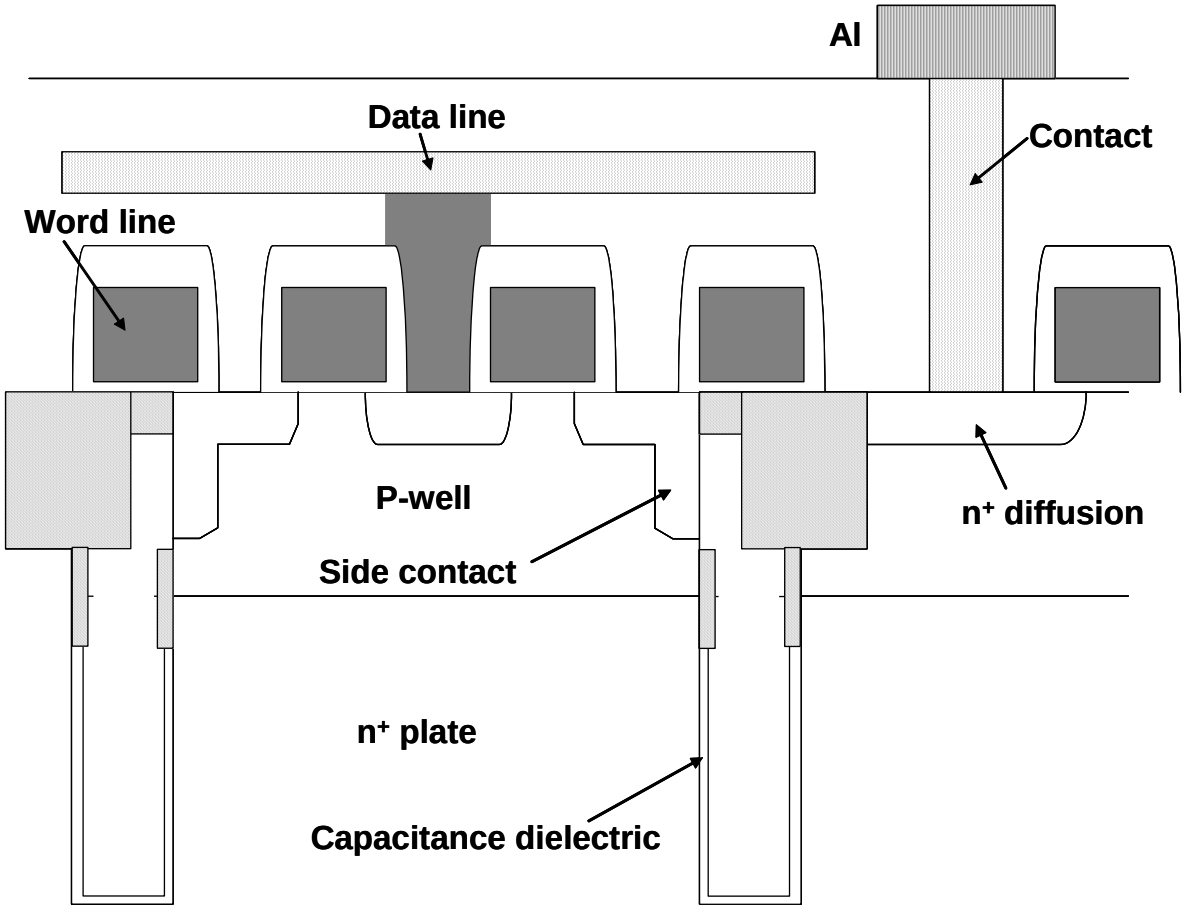


Fig. 5-3 Cross section view of a trench cell.

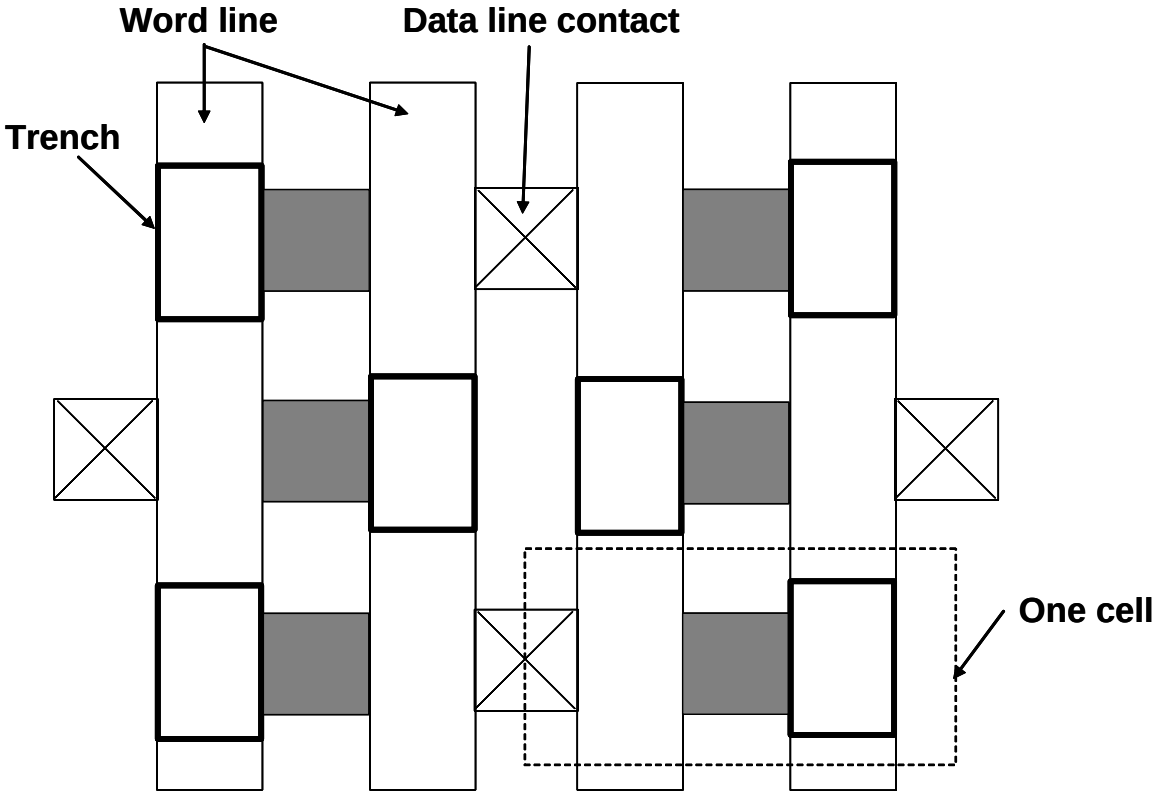


Fig. 5-4 Layout of a trench cell.

5.3 Operations of DRAM cells [14]

In this section, operations of memory cells, i.e., read, program and refresh, are briefly explained. Figure 5-5 shows the schematic illustration of DRAM operation circuits. Before operations of memory cells, by pre-charge operations, the voltage of the data-lines is set to predetermined value, V_p. The value of V_p is in general $V_D/2$. Here, the program voltages to the capacitors for "*high*" and "*low*" are V_D and 0, respectively. The reason to choose $V_p = V_D/2$ is to reduce the power consumption and to increase noise immunity. Pre-charges are performed when the pre-charge clock is "*high*". Just after the pre-charge clock become "*low*", data lines of the pre-charge voltage, V_p, are holding on the floating condition.

5.3.1 Read operations

For the case to read the specific memory cell, for example M_0, operations are as follows. After pre-charge operations, the word line, WL_0, of the memory cell, M_0, is selected. The voltages of $-v_s$, or v_s is superimposed to V_p corresponding to the voltages of the memory cell capacitor, i.e., *0* or V_D, respectively. The superimposed voltages are described by

$$v_s = \frac{V_D}{2}\frac{C_S}{C_S + C_D}, \tag{1}$$

where C_S and C_D are capacitances of the storage capacitor and the data line, respectively. In general, the value of C_S is much smaller than that of C_D which means the superimposed voltage, v_s, is much smaller than V_D . The small changes of V_p=V_D /2, i.e., $\pm v_s$, are detected and amplified by the sense-amplifier. By comparing between the reference voltage, $V_D/2$, and the data line voltage, the voltage of the data lines is amplified to V_D or 0 corresponding to the data line voltages of $V_D/2+ v_s$ or $V_D/2- v_s$, respectively. The amplified voltages, V_D or 0, are read out by the column select switch and the read operation is completed. The information voltages of all memory cells, i.e., V_D or 0, connected to the selected word line are corrupted, when the word line is selected. However, sense amplifiers connected to all data line detect the small change of voltage and recover the corrupted information by amplifying the voltages to V_D or 0. Therefore, the read operations consist of read-out, amplify and re-program for all memory cells connected to the selected word line. During these operations, the information of the selected data line is only read out.

5.3.2 Program operations

The program operation for the selected memory cell, M_0, is performed by selecting the corresponding word line, WL_0, and imposing the intended voltages, i.e., V_D or 0, to the data line, D_0. Because the selection of word line corrupts the data of the unselected memory cells, the read operation is performed before the program operation. That is, the recovered voltages of all memory cells by the read operation hold and the voltages of the selected data line are only programmed by the column select switch. Here, the column select switch is connected to the common I/O line and intended voltages are supplied via the common I/O line.

5.3.3 Refresh operations

Refresh operations are equal to read operations for all word lines, which is unique operations for only DRAMs. By refresh operations, the voltages of all storage capacitances connected to the specific word line are read out, amplified, and reprogrammed, i.e., refreshed. Therefore, the voltages of all storage capacitances are refreshed even if the voltages are modulated by the leakage currents. By selecting all word lines in turn, information of all memory cells on the chip is reproduced.

Figure 5-6 shows the schematic illustration of the refresh operation. Refresh cycle times to select each word line should be T_{ref}/n_{wl} to insure the information of all memory cell against the leakage currents, where T_{ref} is the refresh time and n_{wl} is the number of the word line. For 1Mb DRAM of T_{ref} =8 [ms] and n_{wl} =512, the refresh cycle time is 16μs. One refresh operation for the specific word line is performed on every 16μs. During the refresh operation, the other operations, i.e., read and program, is prohibited. The time required to refresh one word line is the same with the one cycle of the read or program operation, T_{RC}. Therefore, the performance loss by refresh operation, γ, is described by

$$\gamma = \frac{T_{RC}}{T_{ref} / n_{wl}}, \qquad (2)$$

which should be as small as possible. To compromise the performance degradation and the power reduction, the refresh time, T_{ref}, has doubled on every DRAM generation as shown in Fig. 5-7. To keep this trend, the increase of the storage capacitance and the reduction of the leakage current are crucial.

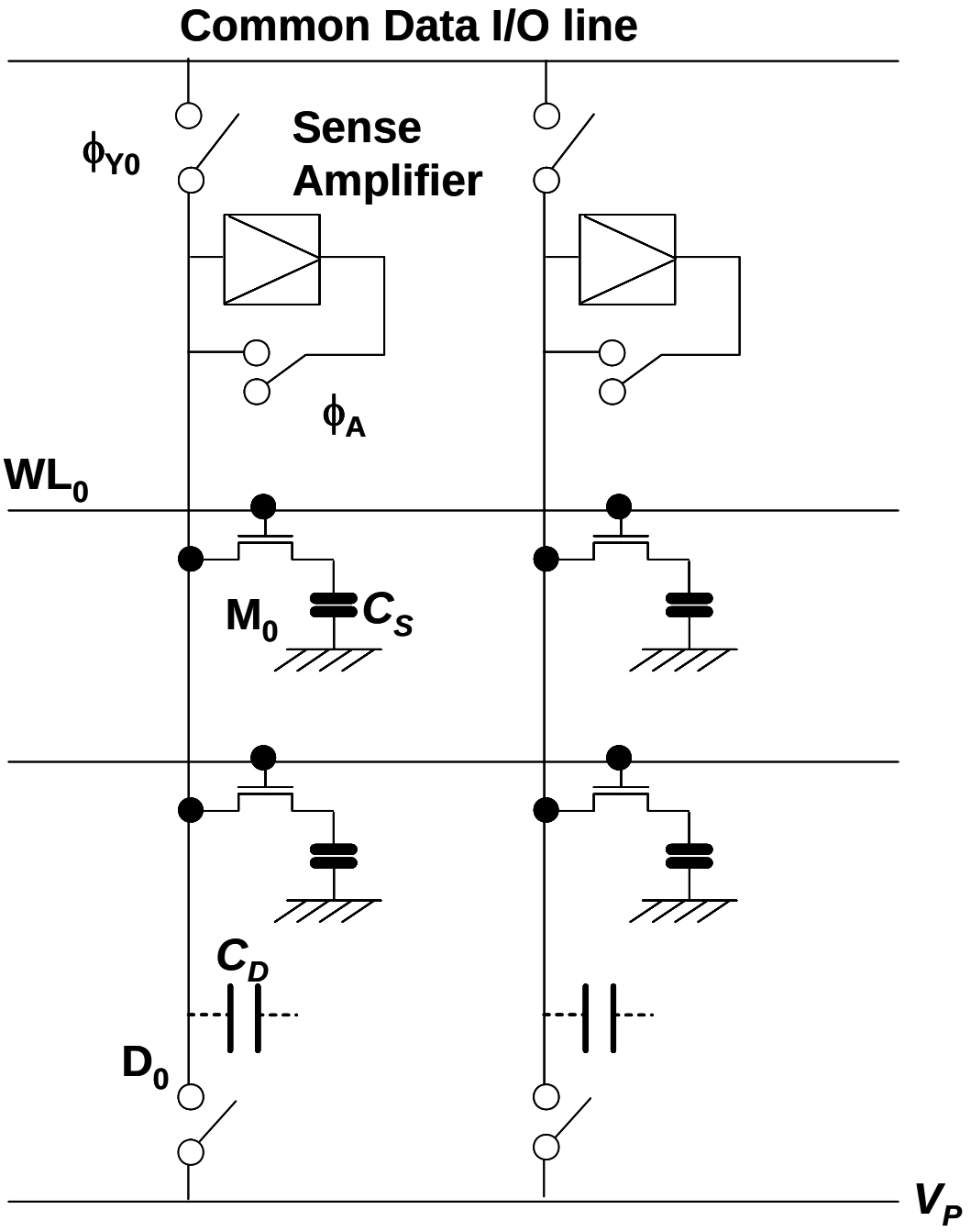


Fig. 5-5 Schematic illustration DRAM operation circuit

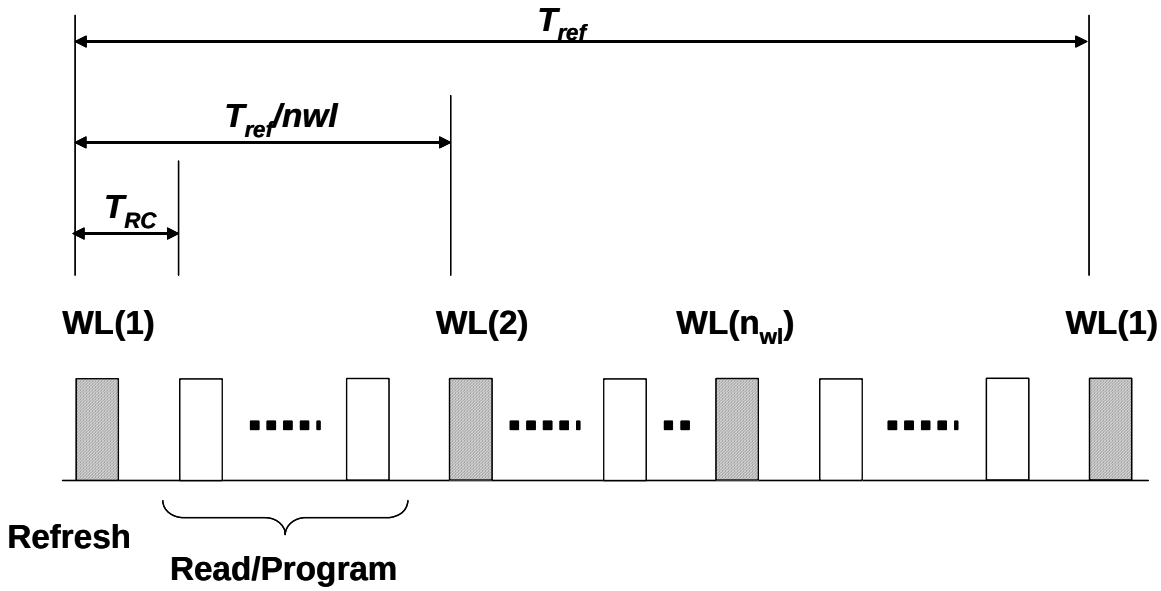


Fig. 5-6 Schematic illustration of refresh operations.

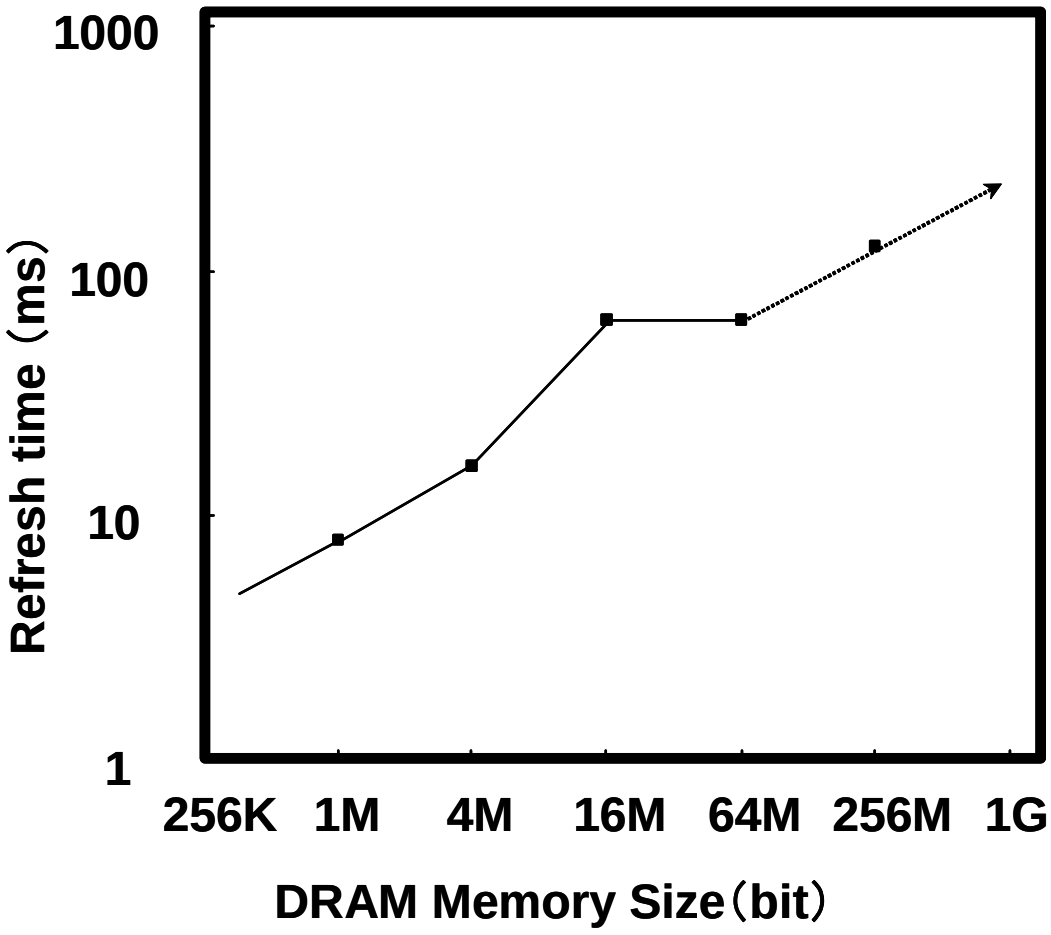


Fig. 5-7 Trend of refresh time.

5.4 Mechanisms of leakage current

5.4.1 Generalized Shockley-Read-Hall recombination current

Imperfections within the semiconductor can disrupt the perfect periodicity of the crystal lattice, and as a result can introduce energy levels into the forbidden gap much as donor and acceptor impurities do. These energy levels then act as "stepping-stones" in the transition of electrons and holes between the conduction and valence band. Because the probability of the transition depends on the size of the step, imperfections can make such transitions more probable and, therefore, can exert a drastic influence on the lifetime in the semiconductor.

The theory of the recombination-generation process taking place through the action of such intermediate energy-level has been worked out by Shockley and Read, and by Hall [15]. This theory, i.e., SRH process, has been remarkably successful in explaining a wide variety of phenomena in many semiconductors and semiconductor devices. However, the contribution of the tunneling has been neglected by the SRH process. In this section, the SRH process is generalized by implementing the contribution of the tunneling [16]. Figure 5-8 shows the various steps that occur in the recombination and generation process via the intermediate-level of traps. The dashed arrows and the solid arrows illustrate the SRH process and the SRH process with the tunneling, respectively. Process (a) is the capture of an electron from the conduction band by the trap. Process (b) is the reverse process, i.e., the emission of and electron from the trap to the conduction band. Process (c) is the capture of a hole from the valence band by the trap. Finally, process (d) is the emission of a hole from the trap to valence band.

Let us now consider the rates of these individual processes for the SRH process with the tunneling [15]. The rate of process (a) is described by

$$r_a = C_n n_t N_t (1 - f), \tag{3}$$

where C_n is the electron capture rate to the trap, n_t the density of the free electron including the contribution of the tunneling, N_t the density of traps and f the probability of occupation of a trap by an electron. The rate of the process (b) is described by

$$r_b = e_n N_t f, \tag{4}$$

where e_n is the electron emission rate from the trap to the conduction band including the contribution of the tunneling. The rate of process (c) is described by

$$r_c = C_p p_t N_t f \ , \tag{5}$$

where C_p is the hole capture rate to the trap, p_t is the density of the free hole including the contribution of the tunneling. The rate of process (d) is described by

$$r_d = e_p N_t (1 - f) , \tag{6}$$

where e_p is the electron emission rate from the trap to the conduction band including the contribution of the tunneling. Emission rates and densities of the free carriers are generalized by implementing the contribution of the tunneling [16, 17].

In a weak electric field, the carrier density at a certain location in a depletion layer is given by the conventional way [15]. However, in a strong electric field, the density of carriers at a certain location with in the depletion layer increased due to the finite probability of carriers tunneling into the gap, as shown Fig. 5-9 (a). Following the approach of Vincent et al. [17], the electron density including the contribution of the tunneling is described by

$$n_t(x) = n(x)\left[1 + \frac{1}{kT}\int_0^{\Delta En} \exp(\frac{E}{kT}) \frac{Ai^2(2m^*\gamma^{-2}\hbar^{-2}E)}{Ai^2(0)} dE\right], \tag{7}$$

where

$$\gamma = \left(2qFm^*\hbar^{-2}\right)^{1/3} , \tag{8}$$

F the electric field of the linear potential, m^* the effective mass, n the carrier density without the contribution of the tunneling, Ai the Airy function, $\Delta E_n = E_c - E_t$, E_c the conduction band edge, E_t the trap level, and the other parameters follow the conventional expressions. Based on the same approach as shown in Fig. 5-9 (b), the enhancement of the emission probability is given by

$$e_n = e_{n0}\left[1+\frac{1}{kT}\int_0^{\Delta En}\exp(\frac{E}{kT})\frac{Ai^2(2m^*\gamma^{-2}\hbar^{-2}E)}{Ai^2(0)}dE\right], \quad (9)$$

where e_{n0} is the emission rate without the contribution of the tunneling. For a linear potential, both the carrier concentration and the emission are enhanced by the same factor, i.e.,

$$\frac{e_n}{e_{n0}} = \frac{n_t}{n} \equiv \Gamma_n + 1, \quad (10)$$

$$\frac{e_p}{e_{p0}} = \frac{p_t}{p} \equiv \Gamma_p + 1, \quad (11)$$

where Γ_n and Γ_p is introduced as the field-effect functions which is described integration term of Eqs. (7) and (9). Using the asymptotic behavior of the Airy function, the expression for Γ_n is can be written as

$$\Gamma_n = \frac{\Delta E_n}{kT}\int_0^1 \exp\left[\frac{\Delta E_n}{kT}u - K_n u^{3/2}\right]du, \quad (12)$$

where

$$K_n = \frac{4}{3}\frac{\sqrt{2m^*\Delta E_n^3}}{q\hbar|F|}. \quad (13)$$

Analytical approximation for the integral in (12) is given as follows. For $K_n > 2/3\Delta E_n/kT$, i.e., for not too large values of the electric field (e.g., at room temperature and for ΔE_n=0.5 [eV], this criterion corresponds to $F < 9\times10^5$ V/cm) the maximum contribution to the integral in (12) comes from $u=u_m$, where $0<u_m<1$. In this case, the integral can be approximated by a second-order series expansion of the function of u in the exponent of Eq. (12) around its maximum at u_m. After setting the integration boundaries to $-\infty$ and ∞, integration yield

$$\Gamma_n = \Gamma_p = 2\sqrt{3\pi}\frac{|F|}{F_\Gamma}\exp\left[\left(\frac{|F|}{F_\Gamma}\right)^2\right], \tag{14}$$

where

$$F_\Gamma = \frac{\sqrt{24m^*(kT)^3}}{q\hbar}. \tag{15}$$

So, in the situation where the maximum contribution to the tunneling effect comes from energy levels above the minimum level at which electrons can tunnel, the integration interval is irrelevant. Obviously, the same reasoning holds for the tunneling of holes. If this situation holds for both electrons and holes, the field-effect functions are equal, i.e., $\Gamma_n=\Gamma_p=\Gamma$.

Let us now derive the generalized SRH recombination current. For the non-equilibrium steady-states, we can write by using four processes, i.e., process (a) to (d),

$$U = r_a - r_b = r_c - r_d, \tag{16}$$

where U is the generalized SRH recombination current. By substituting Eq. (3)-(6) into Eq. (16), we can solve for the occupancy factor, f, of the traps under a given non-equilibrium steady-state in terms of the electron and hole concentration as

$$f = \frac{C_n n_t + e_p}{C_n n_t + e_n + C_p p_t + e_p}. \tag{17}$$

By substituting Eq. (17) into Eq. (16), the generalized SRH recombination current, U, is obtained as follows:

$$U = \frac{n_t p_t - e_n e_p / C_n C_p}{\tau_p\left[n_t + \frac{e_n}{C_n}\right] + \tau_n\left[p_t + \frac{e_p}{C_p}\right]}, \tag{18}$$

where $\tau_n=C_n/N_t$ and $\tau_p=C_p/N_t$. By using the equilibrium condition, we can simplify Eq. (18). In equilibrium, $r_a=r_b$, $r_c=r_d$,

$$f = \frac{1}{1+\exp\left(\frac{E_t - E_F}{kT}\right)}, \tag{19}$$

$$n = n_i \exp\left(\frac{E_F - E_i}{kT}\right), \tag{20}$$

$$p = n_i \exp\left(\frac{E_i - E_F}{kT}\right), \tag{21}$$

where E_t is the trap level and E_F is the Fermi level. By using these equations in equilibrium, we can derive

$$e_{n0} = C_n n_i \exp\left(\frac{E_t - E_i}{kT}\right), \tag{22}$$

and

$$e_{p0} = C_p n_i \exp\left(\frac{E_i - E_t}{kT}\right). \tag{23}$$

Substituting Eq. (10), Eq. (11), Eq. (22) and Eq. (23) into Eq. (18), we can obtain the generalized form of SRH process including the contribution of the tunneling as follows:

$$U = \frac{pn - n_i^2}{\frac{\tau_p}{1+\Gamma_p}\left[n + n_i \exp\left[\frac{E_t - E_i}{kT}\right]\right] + \frac{\tau_n}{1+\Gamma_n}\left[n + n_i \exp\left[-\frac{E_t - E_i}{kT}\right]\right]}. \tag{24}$$

For weak electric field, Γ_n, $\Gamma_p<1$, Eq. (24) reduces to the conventional SRH recombination formula.

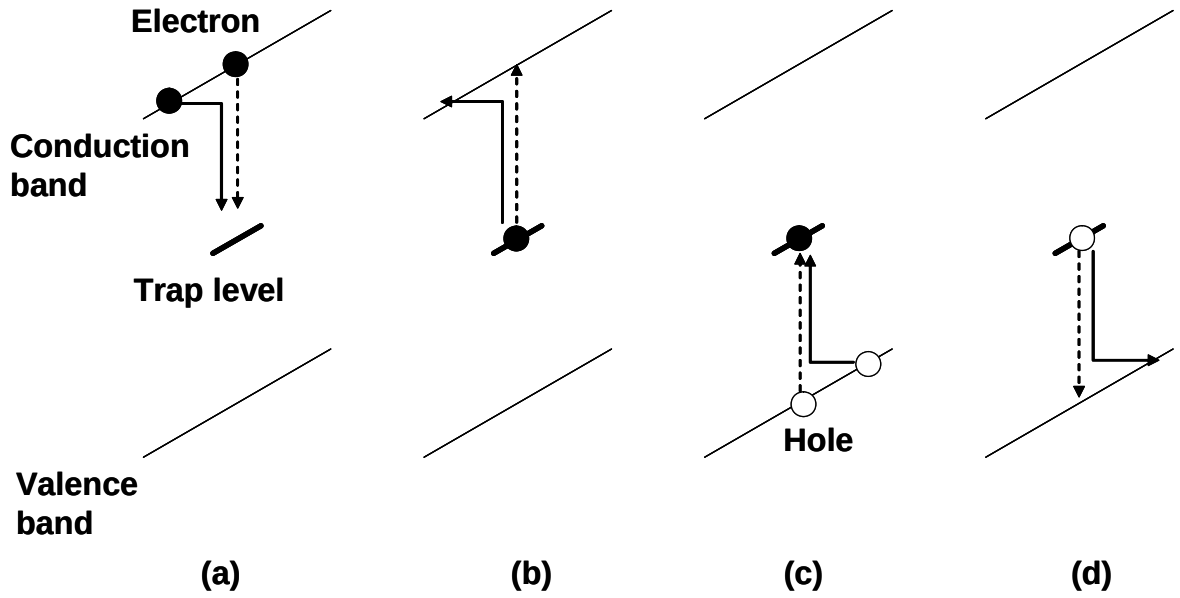


Figure 5-8 The various steps that occur in the recombination and generation process via intermediate-level of traps.

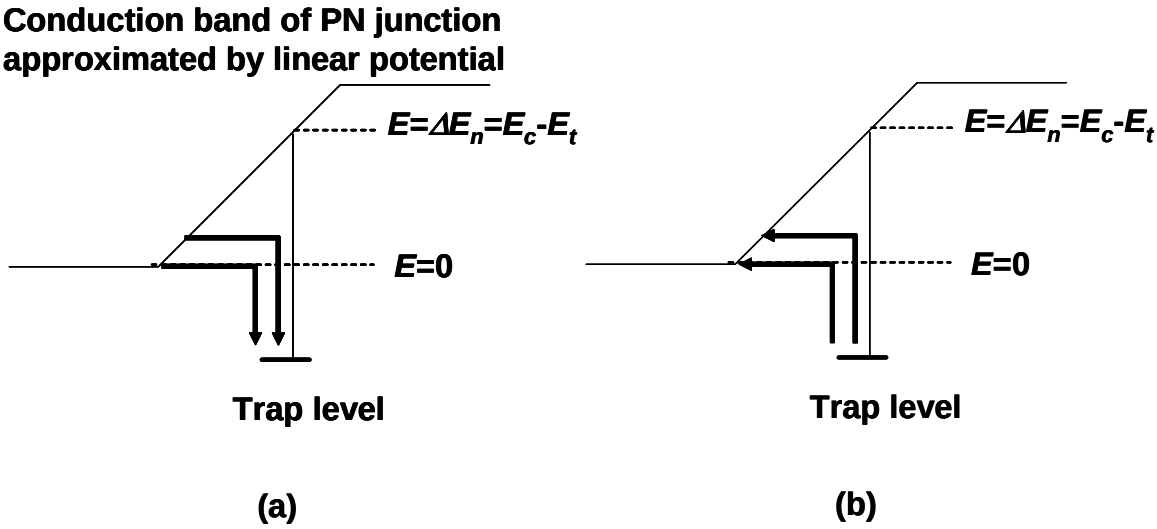


Fig. 5-9 Schematic illustrations for the contributions of the tunneling to (a) the density of the free carrier and (b) the emission rate.

5.4.2 Gate induced drain leakage current

Significant gate-induced drain leakage (GIDL) currents can be detected in thin gate oxide MOSFETs at drain voltages of much lower than the junction breakdown voltage. This current is found to be due to the band-to-band tunneling occurring in the deep-depletion layer in the gate-to-drain overlap region [18, 19], as shown in Fig. 5-10. The GIDL currents consist of two components, as shown in Fig. 5-11. One is the component of the x-direction, i.e., along the channel, the other is that of the y-direction, i.e., perpendicular to the channel [20, 21]. At the normal condition, the component of y-direction is dominant. However, the component of x-direction dose not become negligible after the channel hot electron stress [22-25], because the interface trap generated by the hot electron enhance the component of x-direction.

The band-to-band tunneling (BTBT) of y-direction is only possible in the presence of a high electric field and when the band bending is larger than the energy gap, E_g, as shown in Fig. 5-11 (a). The field in silicon at the Si-SiO_2 interface also depends on the doping concentration in the diffusion region and the difference between V_D and V_G, i.e., V_{DG}. A simple expression for the surface field at the dominant tunneling point can be expressed as

$$F_s = \frac{V_{DG} - E_g / q}{(\varepsilon_{si} / \varepsilon_{SiO2}) T_{ox}}. \tag{25}$$

where F_s is the vertical electrical field at silicon surface, T_{ox} the oxide thickness in the gate-drain overlap region, ε_{si} the silicon permittivity, ε_{SiO2} the oxide permittivity. By using the theory of the tunneling currents [26], the GIDL of y-direction is described by

$$J = AF_s \exp\left[-\frac{4}{3} \sqrt{\frac{2m^*}{\hbar^2}} \frac{E_g^{3/2}}{qF_s} \right]. \tag{26}$$

where A is the pre-exponential factor.

The band-to-band tunneling (BTBT) of x-direction is the two-step tunneling mechanism via interface traps, as shown in Fig. 5-11 (b). The tunneling rates for electrons and holes, T_e and T_h, are described by

$$T_e = f_t / \tau_e, \tag{27}$$

$$T_h = (1 - f_t)/\tau_h, \tag{28}$$

where f_t is the electron occupation factor of interface, τ_e and τ_h the time constant for the electron and the hole tunneling, respectively. According to the Fermi-Golden rule, time constants for electron and hole tunneling are derived as follows [27]:

$$\tau_e(E) = \tau_{0c} \exp\left[\frac{4}{3}\sqrt{\frac{2m_n^*}{\hbar^2}}\frac{(E_c - E_t)^{3/2}}{qF_l}\right], \tag{29}$$

$$\tau_h(E) = \tau_{0v} \exp\left[\frac{4}{3}\sqrt{\frac{2m_p^*}{\hbar^2}}\frac{(E_t - E_v)^{3/2}}{qF_l}\right]. \tag{30}$$

For the non-equilibrium steady-states, we can solve for the electron occupancy factor, f_t, of the traps via $T_e = T_h$ and obtain

$$f_t = \frac{\tau_e}{\tau_e + \tau_h}. \tag{31}$$

By substituting Eq. (31) to Eq. (27), we can obtain the GIDL of x-direction as follows:

$$J = \int_{bandgap} \frac{qN_t(E_t)}{\tau_e(E_t) + \tau_p(E_t)} dE_t. \tag{32}$$

Integrand of Eq. (32) becomes the sharply peaking function of E_t. The peak occurs at $E_t = E_{t,max}$ of $\tau_e = \tau_h$. By using Eq. (29) and Eq. (30), we obtain $E_{t,max} = (E_c + E_v)/2$ and finally derive the analytical equation for the GIDL of x-direciton,

$$J = \frac{qN_t}{2\tau_e((E_c + E_v)/2)} = \frac{qN_t}{2\tau_{0c}} \exp\left[-\frac{4}{3}\sqrt{\frac{2m_n^*}{\hbar^2}}\frac{((E_c - E_v)/2)^{3/2}}{qF_l}\right]. \tag{33}$$

The GIDL shows very weak temperature dependence [18, 19], because the GIDLs of both x- and y-direction are tunneling process.

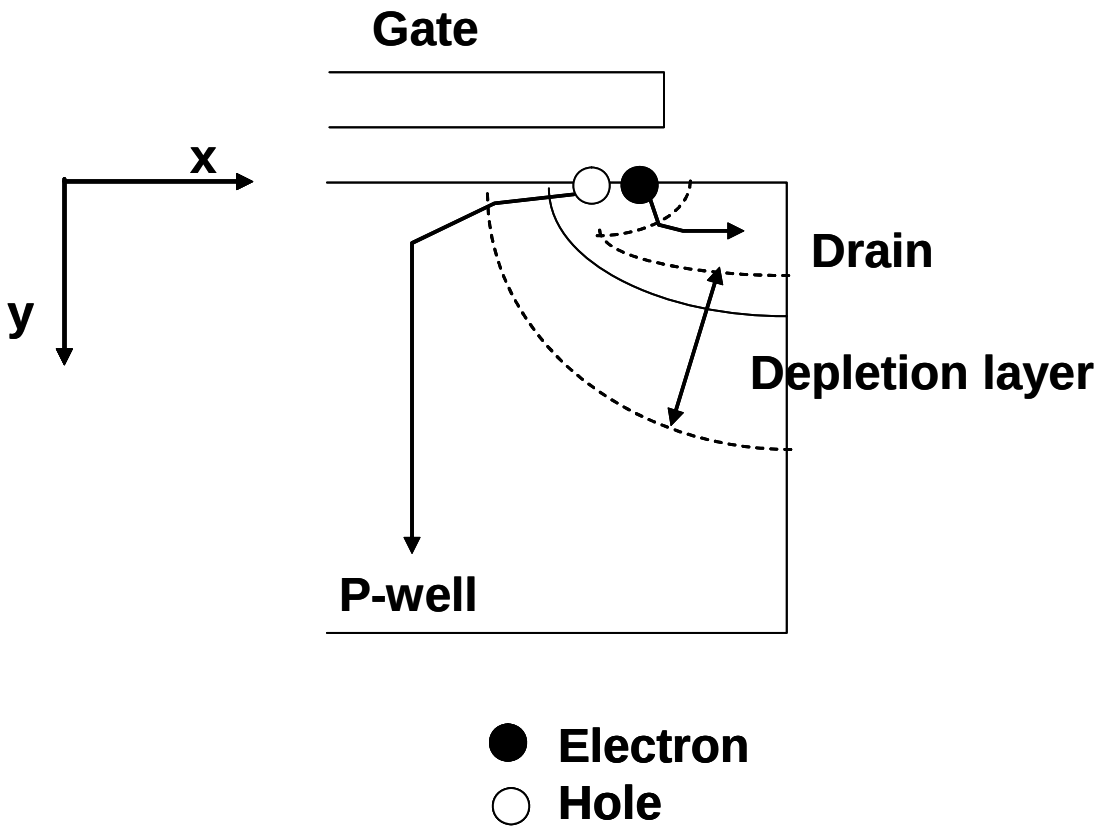


Fig. 5-10 Illustration for the electron-hole pair generation at a deep-depletion region, which is called as the gate induced drain leakage.

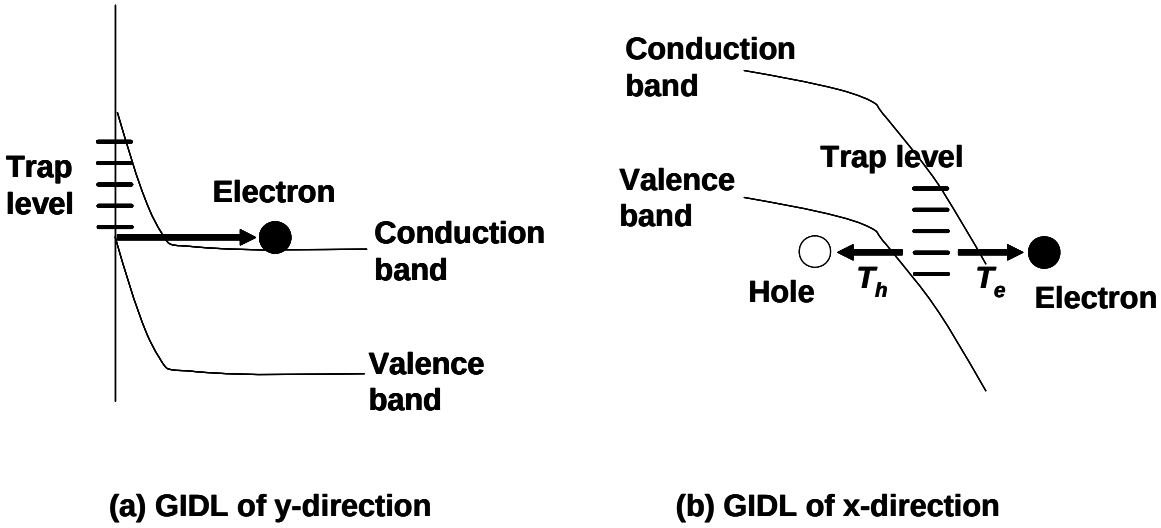


Fig. 5-11 Illustrations of the band bending and the leakage current for (a) x-direction and (b) y-direction.

5.5 References

[1] K. Itoh et al.: Proc. of ESSCIRC'99 (1999) 3.
[2] T. Kirihata et al.: ISSCC99 Dig. Tech. Papers (1999) 420.
[3] S. Takase and N. Kushiyama: ISSCC99 Dig. Tech. Papers (1999) 410.
[4] O. Takahashi et al.: ISSCC2000 Dig. Tech. Papers (2000) 396.
[5] K. Itoh et al.: IEEE J. Solid-State Circuits **32** (1997) 624.
[6] M. Sakao et al.: IEDM Tech. Dig. (1990) 655.
[7] I. Asano et al.: IEDM Tech. Dig. (1998) 755.
[8] I. Yamamoto et al.: Symp. VLSI Tech. Papers (1999) 157.
[9] H. Ishiuchi et al.: IEDM Tech. Dig. (1997) 33.
[10] S. S. Iyer et al.: IEEE Spectrum **35** (1999) 56.
[11] G. Bronner et al.: Symp. on VLSI Tech. Papers (1995) 15.
[12] T. Rupp et al.: IEDM Tech Dig. (1999) 33.
[12] K. Ohnishi et al.: IEDM Tech. Dig. (1998) 397.
[13] S. Crowder et al.: IEDM Tech Dig. (1997) 45.
[14] K. Itoh: *VLSI Memory Chip Design* (Springer, New York, 2001) 1st ed.
[15] A. S. Grove: *Physics and technology of semiconductor devices* (John Wiely & Sons, New York, 1967) 1st ed.
[16] G. A. M. Hurkx, D. B. M. Klaassen, M. P. G. Knuvers: IEEE Trans. Electron Devices 39 (1992) 331.
[17] G. Vincent, A. Chantre and D. Bois: J. Appl. Phys. **50** (1979) 5484.
[18] J. Chen, T. P. Chan, P. K. Ko and C. Hu: IEEE Electron Dev. Lett. **8** (1987) 515.
[19] T. P. Chan, J. Chen, P. K. Ko and C. Hu: IEDM Tech. Dig. (1987) 718.
[20] T. Wang, T. E. Chang and C. Hung: IEDM Tech. Dig. (1994) 161.
[21] T. E. Chang, C. Hung and T. Wang: IEED Trans. Electron Dev. **42** (1995) 738.
[22] H. Sasaki, M. Saitoh and K. Hashimoto: IEDM Tech. Dig. (1987) 726.
[23] C. Duvvury, D. J. Redwine and H. J. Stiegler: IEEE Electron Dev. **EDL-9** (1988) 574.
[24] G. Q. Lo, A. B. Joshi and D. –L. Kwong: IEEE Electron Dev. **EDL-12** (1991) 5.
[25] T. Hori: Proc. of Symp. on VLSI Tech. (1990) 69.
[26] S. M. Sze: Physics of Semiconductor Devices (Wiley, New York, 1981) 2nd ed, p.520.
[27] I. Lundsorm and C. Svensson: J. Appl. Phys. **43** (1972) 5045.

6

Data retention characteristics of DRAMs

Summary

In this chapter, the data retention mechanism of DRAMs is described. A new model for the leakage current of a single *tail bit* of dynamic random access memories (DRAMs) is developed. This model can explain the leakage current of each *tail bit* quantitatively. To derive model equations, we assume that some bits containing one specific trap become *tail bits* among all bits in a DRAM chip. The variation of the leakage current of *tail bits* is attributed to the fluctuation of the trap level. By introducing the trap level fluctuation model, we have successfully reproduced the distribution of the retention time for *tail bits*. We also have obtained a good agreement between model and experimental results of *tail* distributions as functions of process splits and the temperature by using the present model. As an example applied by the present model, we estimated the required number of the repairable bits for 1G DRAM.

6.1 Introduction

The control of the retention time for the stored charge is a key issue for realizing future dynamic random access memories (DRAMs) of high density, because the refresh time doubles with each successive generation. This requirement derives from needs to keep the refresh interval constant even if the number of memory bits increases. The retention time of each bit, T_{ret}, can be defined as the duration while the stored signal can be read out. To follow the trend of the refresh time, the retention time of each bit, T_{ret}, should be improved on every DRAM generation. The retention time is deteriorated by several leakage mechanisms. Among them, the junction leakage current (Fig. 6-1) is a critical subject for the retention time, because it is enhanced as an operation temperature increases. The maximum operation temperature of DRAMs is higher than 70˚C. Therefore, the control of the junction leakage current is one of the most important issues to improve T_{ret} [1], [2].

For DRAMs of every generation, the transistor has been shrunk to about 80-70%. To suppress the short channel effect for the shrunk transistor, the substrate doping concentration has increased which results in the increase of the electric field in the junction. Since the junction leakage increases with the junction field, it has been more difficult to improve T_{ret} with each successive generation.

Figure 6-2 shows the cumulative plot of T_{ret} which shows two distinct distributions known as *normal* and *tail* distribution. The bits belonging to *normal* and *tail* distributions are called as *normal* and *tail bits*, respectively in this paper. Many studies [1-17] have been concentrated on finding the mechanism of *tail* distributions. There have been also many simulation studies [1][7][8] to understand the reason for the existence of *tail bits*. In order to explain the abnormally large junction leakage currents of *tail bits*, the trap-assisted tunneling (TAT) model with local enhancement of electric field has been proposed [7-9]. It has been also reported that gate-induced drain leakage (GIDL) current due to band-to-band tunneling (BTBT) or band-to-defect tunneling (BTDT) is a plausible leakage source for the *tail* distribution [11-14]. Basically many researchers have shown that the dominant leakage component in DRAM cells is the generation current via traps derived by using generalized form of the Shockley-Read-Hall (SRH) model [20][21], and the wide variation in the retention time originates from the distribution in locations and/or energy levels of traps [7-10][17-19].

To follow the trend of T_{ret} toward the future generation of DRAMs, we should optimize the design parameters, including the storage capacitance, C_s, the substrate bias, V_{BB}, the number of the repair capability, process conditions and so fourth. Understanding the *tail* distribution of T_{ret} based on a physical model is crucial to

optimize these design parameters. The goal of this study is to provide a model to predict T_{ret} distributions of the current and the next generations for refresh-time-oriented DRAM design optimization. In this paper, we analyze, at a first step, the anomalous junction leakage of a single *tail bit* and then model the distribution of the junction leakage. Based on the analysis of the leakage for a single bit and the distribution of the junction leakage, we have derived a quantitative model of T_{ret} distribution for *tail bits*, which can available to determine the number of the repairable bits.

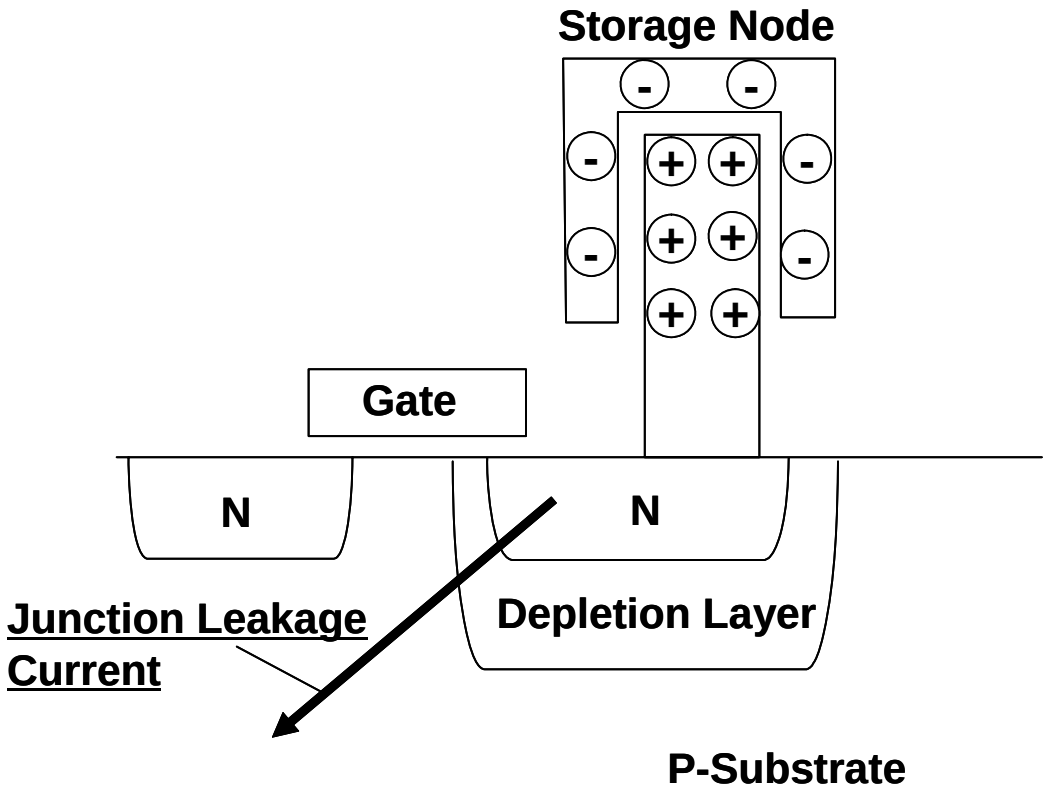


Fig. 6-1. Schematic illustration of DRAM structure and data retention.

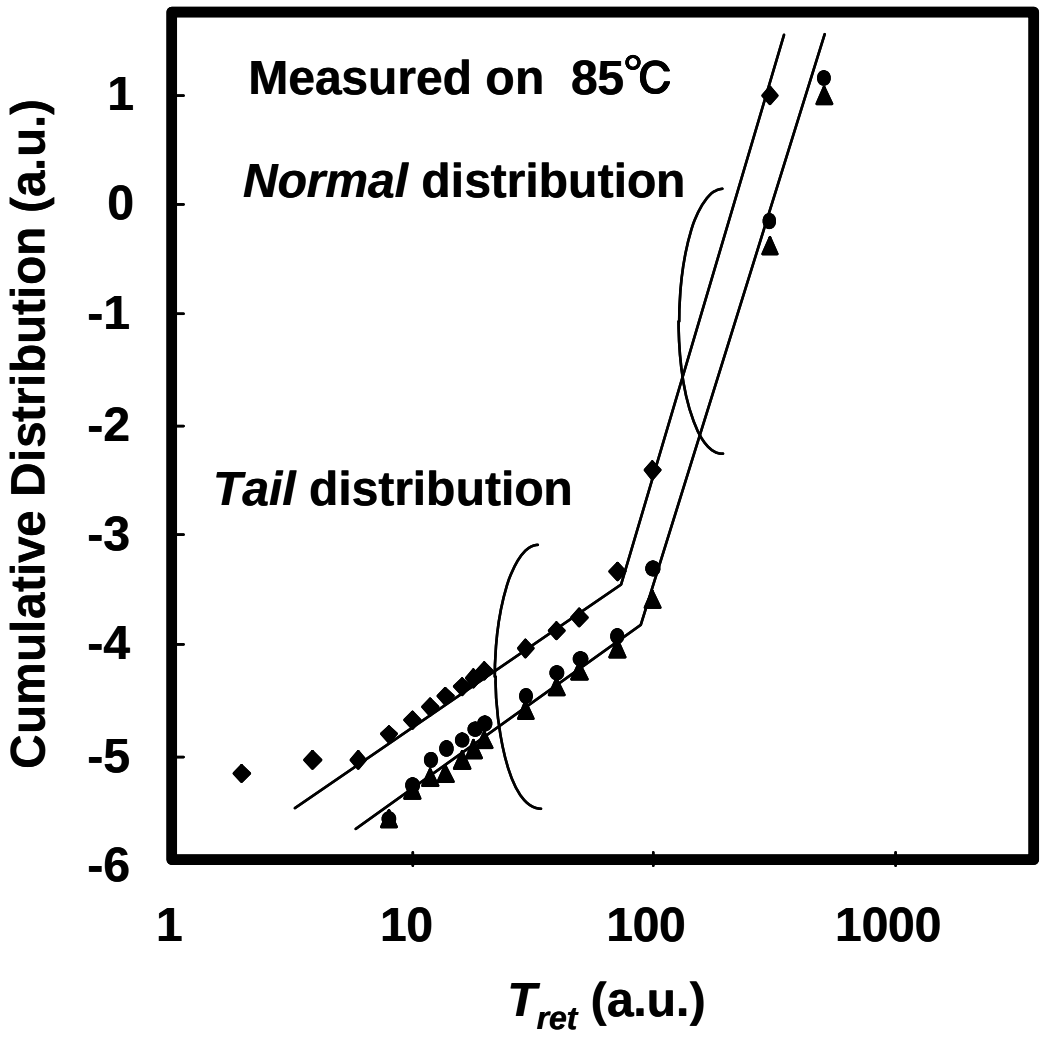


Fig. 6-2. Typical cumulative distribution of T_{ret}.

6.2 Leakage current of a single bit

At a first step, the leakage current of a single *tail bit* is obtained by

$$I_L = Q / T_{ret}, \quad (1)$$

where Q is the charge stored in the storage node. The variation of Q for each bit can be neglected, because capacitances of storage nodes are well controlled on mass productions. The values of T_{ret} can be obtained as functions of the temperature and the substrate bias. Figure 6-3 shows I_L of a *tail bit* as a function of the junction voltage, V_j, at 85˚C. Here, the junction voltage, V_j, is obtained by the substrate bias plus the built-in potential of the p-n junction, which is obtained with the device simulator. As shown in Fig. 6-3, the values of I_L increase with V_j due to the field enhancement mechanism [22][23] and I_L at V_j=4.3 [V] is 1×10^{-13} [A] at 85˚C [16].

Figure 6-4 shows the activation energy of the leakage current for a *tail bit* as a function of V_j. Activation energies are obtained from the values of I_L obtained from measurement results of T_{ret} at three temperatures, 115˚C, 85˚C, and 55˚C. Here, the results shown both in Fig. 6-3 and Fig. 6-4 are obtained from the identical single *tail bit*. The values of the activation energy decrease with increase of V_j via the field enhancement mechanism [22][23], as shown in Fig. 6-4. We assume that memory bits containing a specific trap center become *tail bits* among all memory bits in a DRAM chip. Therefore, the activation energy at V_j,=0 is 0.62 [eV], which means the *tail bit* contains the specific trap of E_t=0.62 [eV], as shown in Fig. 6-4.

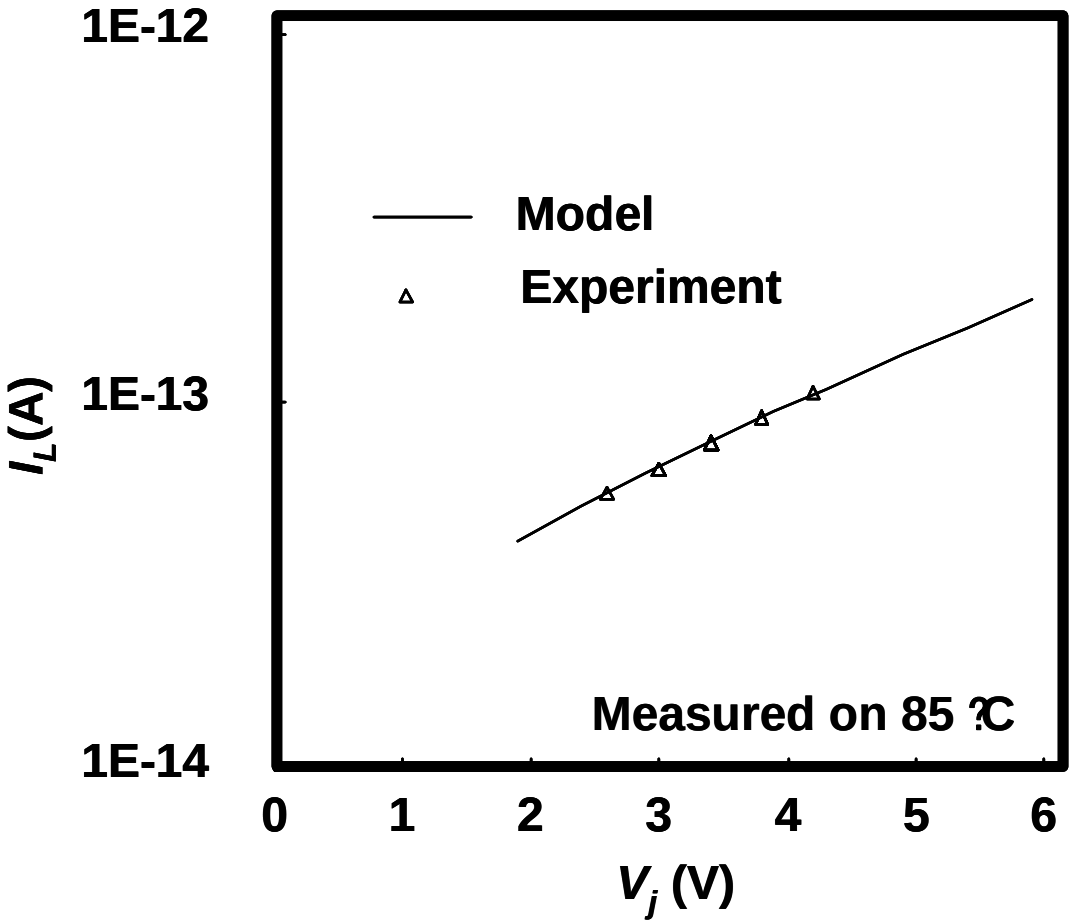


Fig. 6-3. Experimental and calculation results of I_L for a *tail bit* as a function of the junction voltage, V_j, at 85˚C.

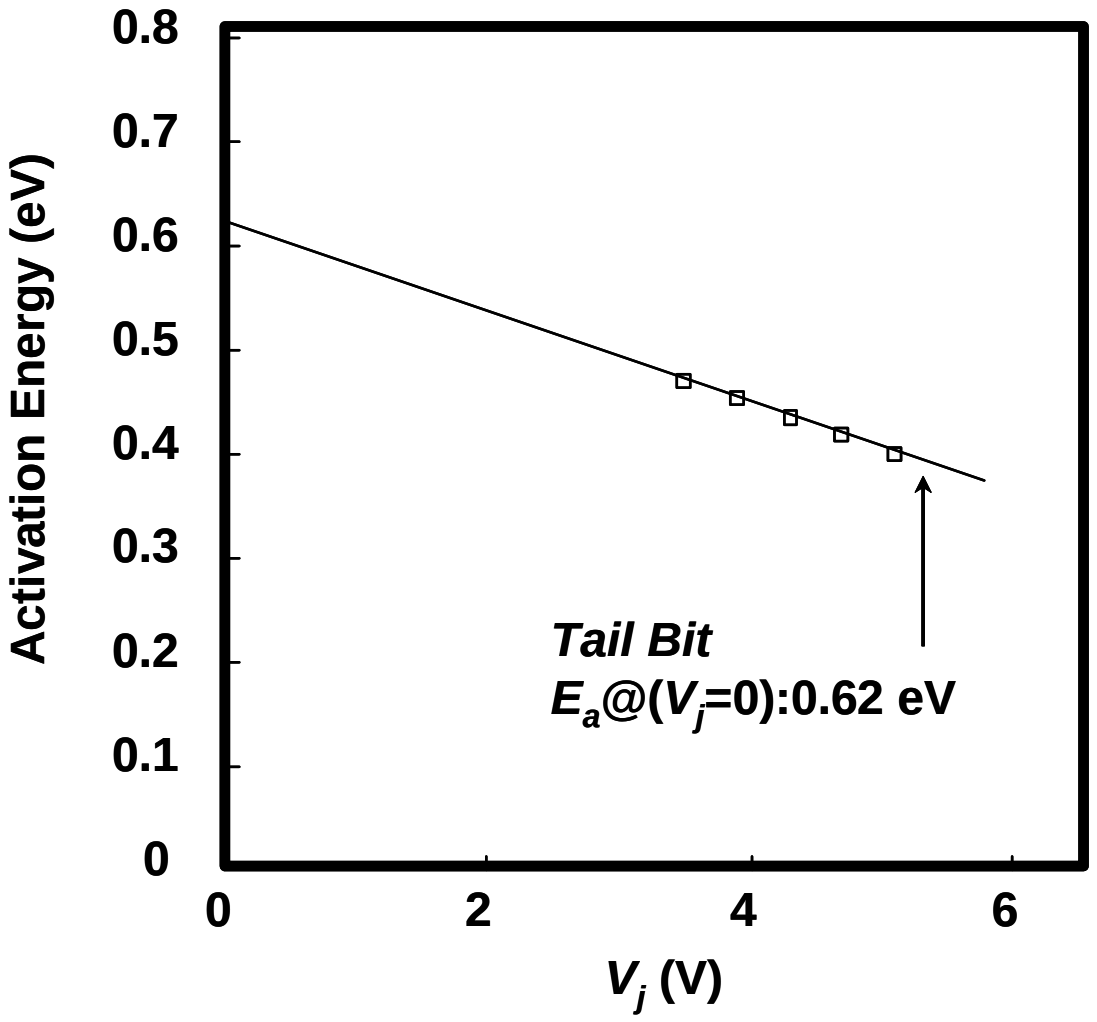


Fig. 6-4. Activation energy of the *tail bits*.

6.3 Modeling of the leakage current

In a generalized form of the Shockley-Read-Hall (SRH) model, the leakage current based on the thermal carrier generation enhanced by the applied electric field can be described as follows [16-23],

$$G = \frac{n_i^2 - np}{\frac{\tau_P}{1+\Gamma_P(F)}\left(n + n_i \exp\left(-\frac{\Delta E_t}{kT}\right)\right) + \frac{\tau_n}{1+\Gamma_n(F)}\left(p + n_i \exp\left(\frac{\Delta E_t}{kT}\right)\right)}, \tag{3}$$

where

$$\Gamma_n(F) = 2\sqrt{3\pi}(F / F_{\Gamma,n}) \exp\left[(F / F_{\Gamma,n})^2\right], \tag{4}$$

$$\Gamma_p(F) = 2\sqrt{3\pi}(F / F_{\Gamma,p}) \exp\left[(F / F_{\Gamma,p})^2\right], \tag{5}$$

$$F_{\Gamma,n} = \sqrt{24 m_n^{*} (kT)^3} / q\hbar, \tag{6}$$

$$F_{\Gamma,p} = \sqrt{24 m_p^{*} (kT)^3} / q\hbar, \tag{7}$$

$$\tau_n = 1/\left(v_{th,n}\sigma_n N_t\right), \tag{8}$$

$$\tau_p = 1/\left(v_{th,p}\sigma_p N_t\right), \tag{9}$$

$$\Delta E_t = E_i - E_t. \tag{10}$$

Here, F is the electric field, m^* the effective mass, T the temperature, v_{th} the thermal velocity, σ the capture cross section, N_t the trap density, E_i the intrinsic Fermi-level, E_t the trap level, and the other parameters are used as the conventional ways.

Because of $n_i >> n$, p in the depletion layer under a reverse bias, Eq. (3) becomes more

simple form:

$$G_L = \frac{G_1 G_2}{G_1 + G_2}, \tag{11}$$

where

$$G_1 \approx \left(1 + \Gamma_n(F)\right)\frac{n_i}{\tau_n}\exp\left(-\frac{\Delta E_t}{kT}\right), \tag{12}$$

$$G_2 \approx \left(1 + \Gamma_p(F)\right)\frac{n_i}{\tau_p}\exp\left(\frac{\Delta E_t}{kT}\right), \tag{13}$$

G_1 corresponds to the electron transition rate from the trap level to the conduction band, and G_2 to that from the valence band to the trap level, as depicted in Fig. 6-5.
As a next step, we derive the leakage current of a single *tail bit*. In order to derive model equations, we assume that memory bits containing a specific trap center in the depletion layer of the junction become *tail bits*. Since the number of *tail bits* is significantly small, it is provided that a particular *tail bit* contains only one trap center. Under this assumption, we have derived the model equation for leakage currents of a single *tail bit* by replacing N_t in Eq. (8) and Eq. (9) to "1", as follows:

$$I_L = q\frac{G_1 G_2}{G_1 + G_2}\bigg|_{N_t=1}. \tag{14}$$

Note that the dimension of I_L is "Ampere". For the calculation of the leakage current, the temperature dependent parameters are described as [24],

$$n_i = 1.45\times10^{10}\ [\mathrm{cm}^{-3}]\left(\frac{T}{300.15\,[\mathrm{K}]}\right)^{1.5}\exp\left(21.6\text{-}\frac{E_g}{2kT}\right), \tag{15}$$

$$E_g = 1.16\,[\mathrm{eV}] - 0.000702\,[\mathrm{eV}]\frac{T^2}{T + 1108\,[\mathrm{K}]}, \tag{16}$$

$$v_{th}=\sqrt{3kT/m^*}\ , \tag{17}$$

where $m^*=0.25m_0$ [23], and $m_0=9.1095\times10^{-31}$ [kg]. At 85˚C, we use the values: E_g=1.10 [eV], $n_i=8.16\times10^{11}$ [cm^3], $v_{th,n}= v_{th,p}=2.55\times10^7$ [cm/s]. The field enhancement factor, $\Gamma_n(F)$ is calculated by using Eq. (4) and Eq. (6), while we assume $\Gamma_p(F)\sim \Gamma_n(F)$ [23]. At the operation bias condition of DRAMs, $\Gamma_n(F)$= 15.4. The electric field, F, is the maximum electric field around the gate edge obtained by carrying out 3D-device simulator with CADDETH [25].

Figure 6-3 shows the comparison between the model and the experimental results of the leakage current, I_L, as the function of V_j. For model calculations, we use the values as T=85˚C, E_t=0.62 [eV], and F from the device simulation. A good agreement between the model calculation using Eq. (14) and the experimental results is obtained as shown in Fig. 6-3. In order to obtain the best fit of the data, we chose $\sigma_n=\sigma_p= 2.50\times10^{-14}$ cm^2, which has been kept constant for all model calculations.

]

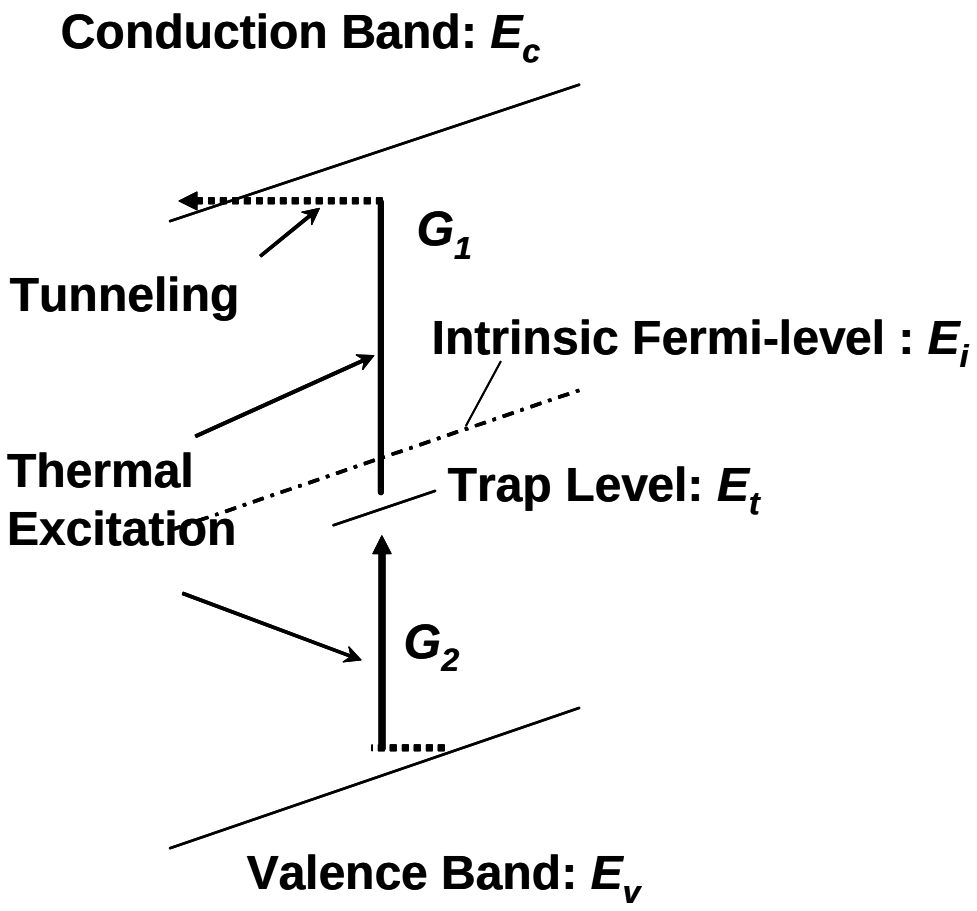


Fig. 6-5. Illustration of the leakage current at the depletion region.

6.4 Origin of the *tail*-distribution

Figure 6-6 shows a cumulative plot of the *tail* distribution. The result suggests that the logarithm of T_{ret} follows the Gaussian distribution. It is not irrational to speculate that the location of a trap center and hence its surroundings statistically distribute in the p-n junction. Consequently the effective energy level of the trap might be perturbed by the lattice distortion due to the stress and the other physical condition around the trap. Therefore, the *tail* distribution in Fig. 6-6 can be attributed to the fluctuation of the trap level. Based on the trap level fluctuation model, *tail bits* distribution is described by

$$\int f(T_{ret})dT_{ret} = \int (\eta N_{cell} / \sqrt{2\pi}\delta) \exp(-\frac{(E_t - \langle E_t \rangle)^2}{2\delta^2}) \frac{\partial E_t}{\partial T_{ret}} dT_{ret}, \qquad (18)$$

where $<E_t>$ is the average trap level of *tail bits*, η ratio of *tail bits* to total bits, δ the root mean square (RMS) for the trap levels of *tail bits*, and N_{cell} the total bits number of DRAM chip. By fitting Eq. (18) to the experimental results in Fig. 6-6, we can obtain the trap level distribution of *tail bits*. As shown in Fig. 6-6, a good agreement between the experimental result and the model calculation for *tail* distributions are obtained. Extracted values of $<E_t>$ and δ are 0.677 [eV] and 0.025 [eV], respectively. Here, the values of η would depend on fabrication lines, which we do not specify this value in this paper. Only values of the electric field and temperature were changed for all model calculations in the later section, while the other model parameters are kept constant.

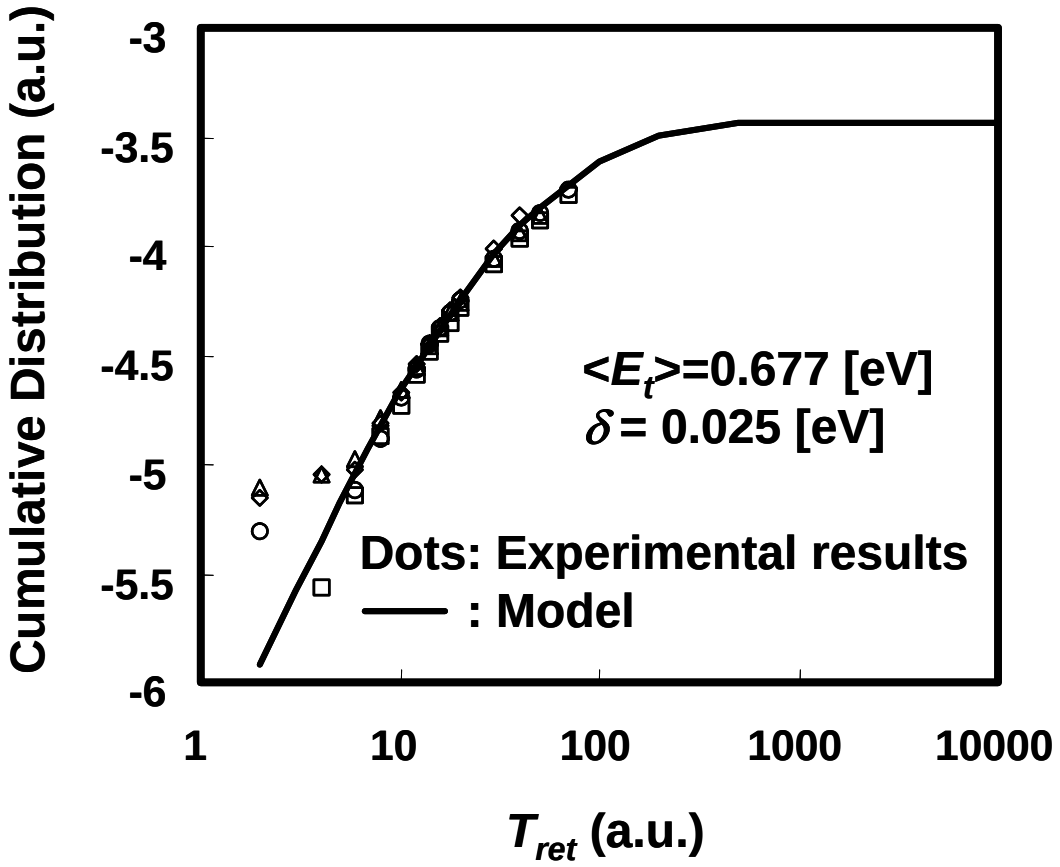


Fig. 6-6. Experimental and calculation results for the cumulative distribution of T_{ret}.

6.5 Verification of the model and its application

We compare the results of model calculations with those of experiments for several DRAM test chips. Table 1 lists the process splits of the test chips. Figure 6-7 shows the cumulative plots of *tail* distributions as the parameter of the process splits. For all model calculations, we used only a different value of the electric field by using a device simulator, while the other parameters are constant. As shown in Fig. 6-7, we obtain good agreements between the model and the experimental results. Figure 6-8 shows the cumulative plots of T_{ret} as the parameter of the operation temperature. Good agreements are also obtained for the temperature dependence of *tail bits* distributions as shown in Fig. 6-8.

We have applied the present model to make a scenario for improvement of T_{ret} for the state-of-the-art technology. The design parameters are implemented to the present model via the charge stored in the storage node, Q, as follows:

$$Q = C_S (V_{DL} / 2 - \Delta V_S (C_S + C_D) / C_S), \tag{19}$$

where C_S and C_D are the storage and the data-line capacitance, V_{DL} the storage voltage, ΔV_S the sensitivity of the sense-amplifier. The process condition can be implemented via the electric field by using a device simulator. For example, we calculated the required number of the repairable bits as a function of *Cs* for 1G DRAM. Figure 6-9 shows the refresh time as the function the number of the repairable bits which is calculated by using our model. For all model calculations, we used only a different value of the electric field by using a device simulator, while the other parameters are kept constant. As shown in Fig. 6-9, in the case of C_s =50fF, the required number of the repairable bits would be more than 300.

Finally we discuss the specific trap which is contained in *tail bits*. As shown in Fig. 6-6, we have extracted the average trap level of 0.677 [eV]. One of the most plausible mechanisms for the trap generation is the contamination of transition metals during processing. The elements which show close values of 0.677 [eV] are “Fe” and “Cu” [26] [27]. To get good retention characteristics, hence, the reduction of the contamination is crucial in addition to the junction field reduction.

Table 1 Process splits of the DRAM test chips.

Process	P-well	Cs
A	N_1	C_1
B	N_2	C_1
C	N_3	C_2
D	N_4	C_2

$N_1 > N_2 > N_3 > N_4$

$C_1 < C_2$

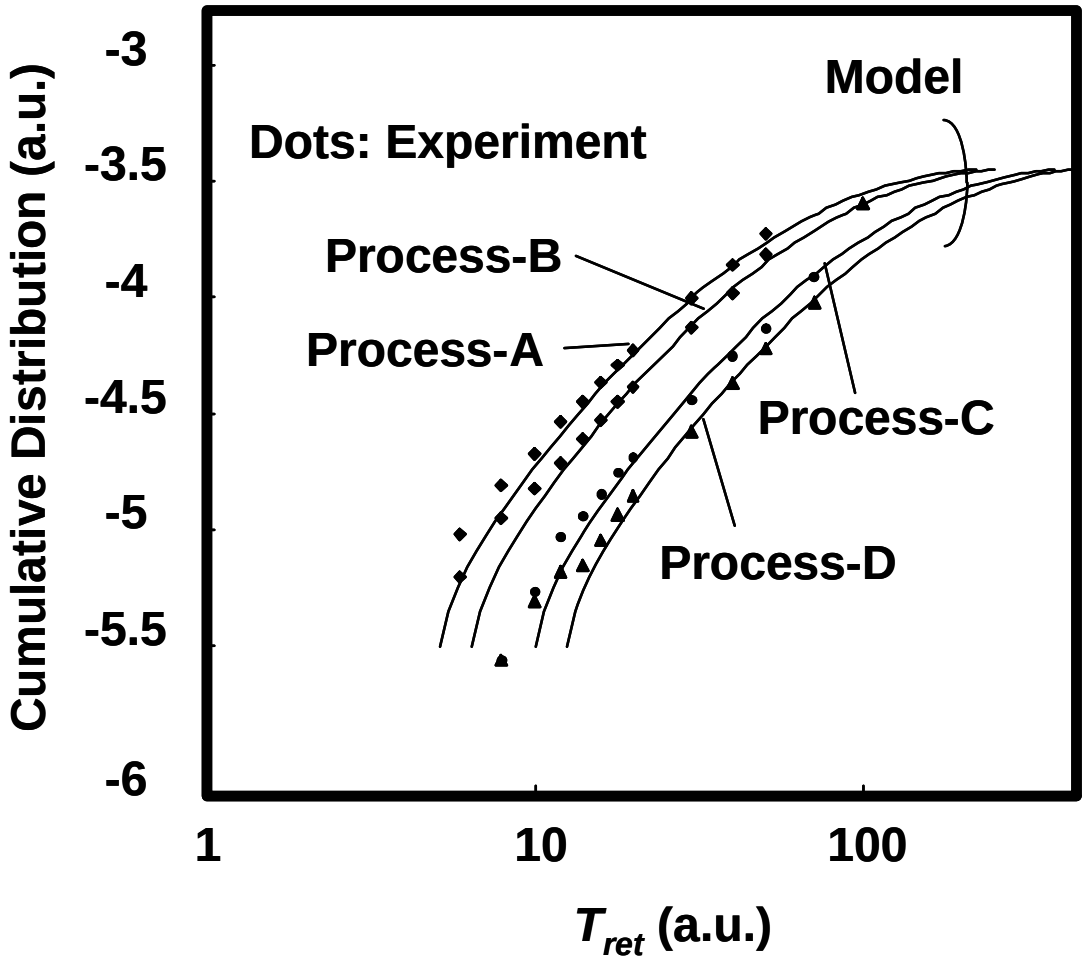


Fig. 6-7. Comparisons between the model and the experimental results for four process splits.

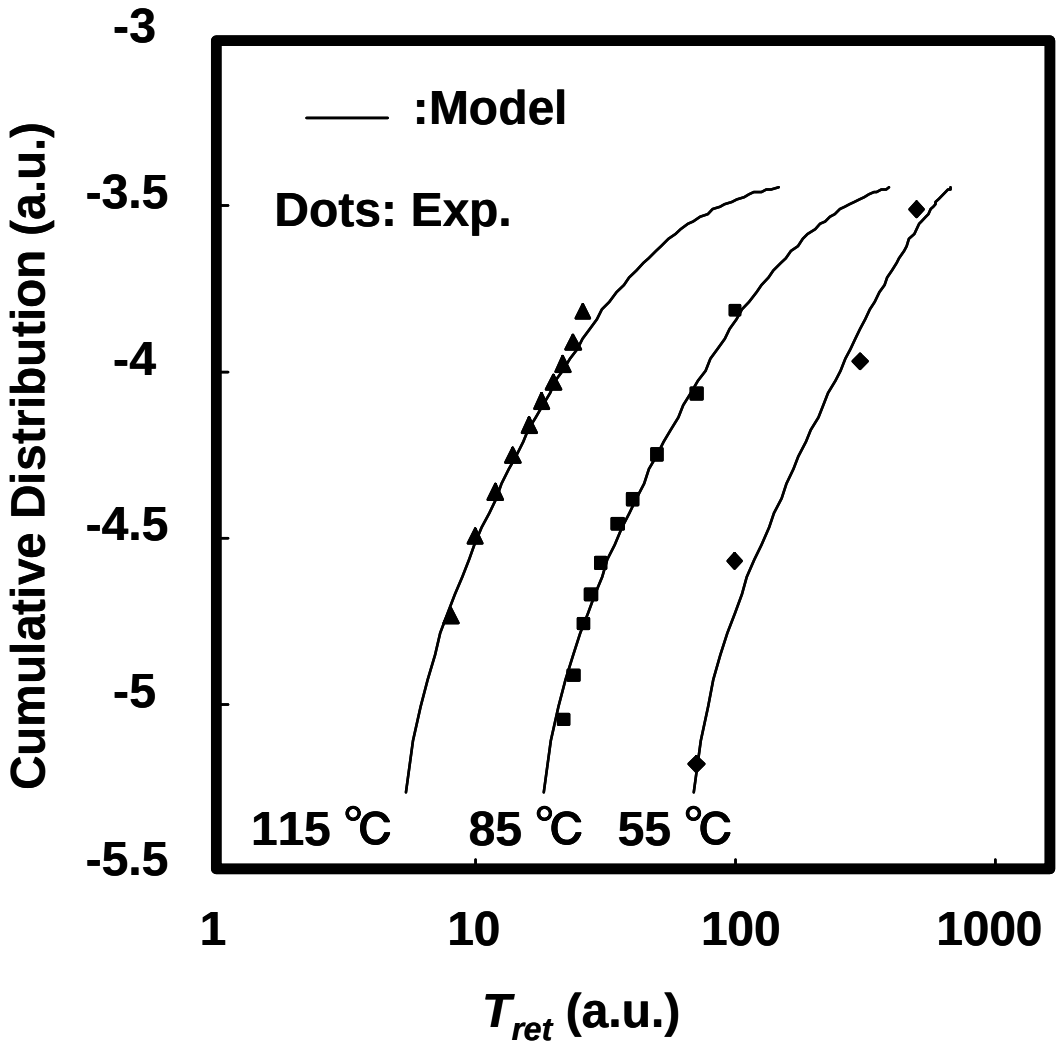


Fig. 6-8. Comparisons between the model and the experimental results for the temperature dependence.

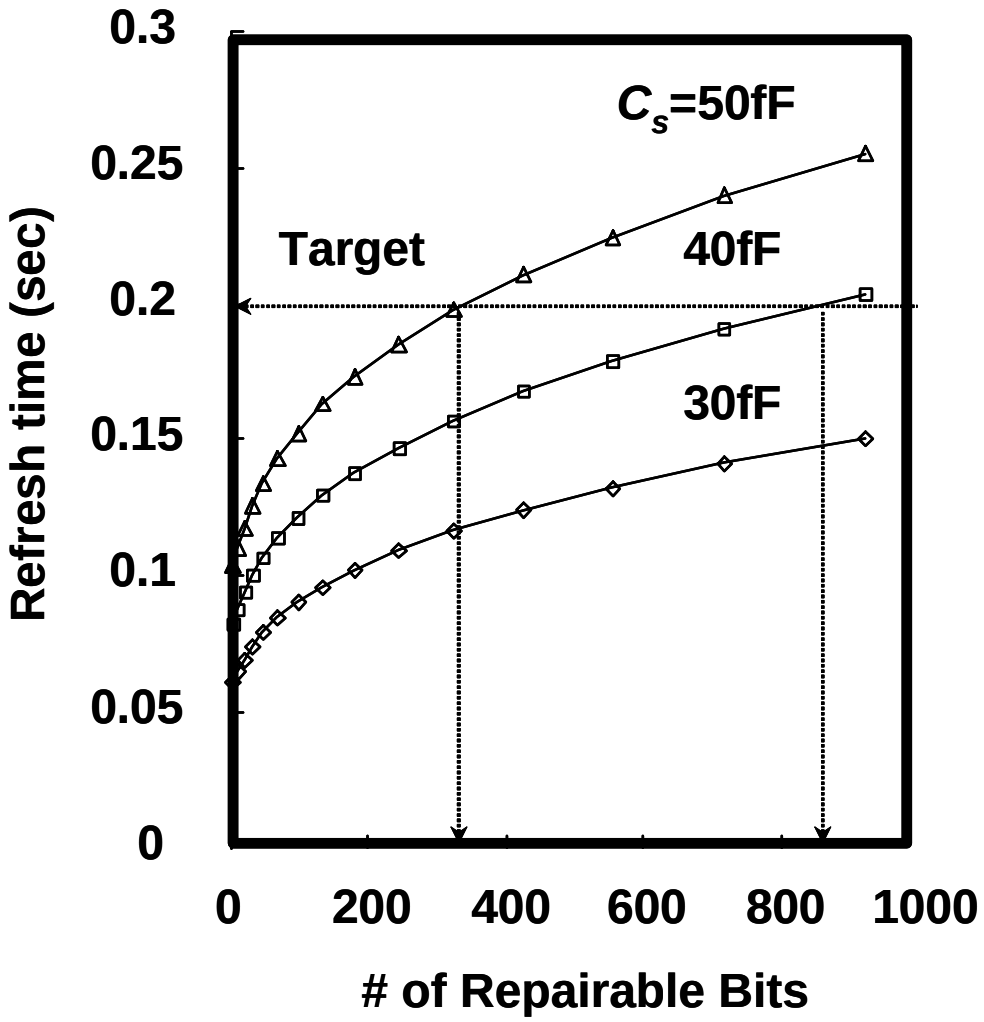


Fig. 6-9. Estimations of the required number of repairable bits for 1G-DRAM.

6.6 References

[1] T. Hamamoto, S. Sugiura, S. Sawada: IEDM Tech. Dig. (1995) 377.

[2] T. Hamamoto, S. Sugiura, S. Sawada: IEEE Trans. Electron Devices **45** (1998) 1300.

[3] Y. P. Kim, S. T. Kim, J. T. Moon, and S. U. Kim: IEEE Trans. Electron Devices Mater. Rel. **1** (2001) 104.

[5] Y. Mori, R. Yamada, S. Kamohara, M. Moniwa, K. Ohyu, and T. Yamanaka: Proc. of IRPS2001 (2001) 167.

[6] Y. Mori, S. Kamohara, M. Moniwa, K. Ohyu, T. Yamanaka, R. Yamada: IEEE Trans. Electron Devices **53** (2006) 398.

[7] K. Yamaguchi: IEEE Trans. Electron Devices **47** (2000) 774.

[8] A. Hiraiwa, M. Ogasawara, N. Natsuaki, Y. Itoh, H. Iwai: IEDM Tech. Dig. (1998) 157.

[9] S. Ueno, Y. Inoue, and M. Inuishi: IEDM Tech. Dig. (1999) 37.

[10] S. Kamohara, K. Kubota, M. Moniwa, K. Ohyu, and A. Ogishima: IEDM Tech. Dig. (1999)539.

[11] K. Saino, K. Okonogi, S. Horiba, M. Sakao, M. Komuro, Y. Takaishi, T. Satoh, K. Yoshida, K. Koyama: IEDM Tech. Dig. (1998) 149.

[12] M. Chang, J. Lin, S. N. Shil, T. Wu, B. Huang, J. Yang, and P. Lee: IEEE Trans. Electron Devices **50** (2003) 1036.

[13] H. W. Seo, G. Y. Jin, K. Yang, Y. Lee, J. Lee, D. Song, Y. Oh, J. Noh, S. Hong, D. kim, J.Kim, J. Kim, H. Kim, D. Won, and W. Lee: Proc of IRPS2002 (2002) 287.

[14] H. Suzuki, M. Kojima, and Y. Nara: Proc. of Int. Conf. Solid State Materials (1998) 32.

[15] S. Ueno, T. Yamashita, H. Oda, S. Komori, Y. Inoue, T. Nishimura: IEDM Tech. Dig. (1998) 153.

[16] J- H. Yi, S- K Park, Y- J Park, and H. S. Min: IEEE Trans. Electron Devices **52** (2005) 554.

[17] S. Jin, J- H Yi, J. H. Choi, D. G. Kang, Y. J. Park, and H. S. Min: IEEE Trans. Electron Devices **52** (2005) 2422.

[18] S. Jin, J- H Yi, Y. J. Park, and H. S. Min: Proc. of Int. Conf. Simulation Semiconductor Processes Devices (2004) 315.

[19] S. Jin, J- H Yi, J. H. Choi, D. G. Kang, Y. J. Park, and H. S. Min: IEDM Tech. Dig. (2004) 339.

[20] W. Schokeley and W. T. Read: Phys. Rev. **87** (1952) 835.

[21] R. N. Hall: Phys. Rev. **87** (1952) 387.

[22] G. A. M. Hurkx, H. C. de Graaff, W. J. Kloosterman, M. P. G. Knuvers: IEEE Trans. Electron Devices **39** (1992) 2090.
[23] G. A. M. Hurkx, D. B. M. Klaassen, M. P. G. Knuvers: IEEE Trans. Electron Devices **39** (1992) 331.
[24] X. Xi, K. M. Cao, H. Wan, M. Chan, and C. Hu: *BSIM4.2.1 MOSFET Model -User's Manual* (University of California, Berkeley, CA, 2001), p.12-8.
[25] T. Toyabe, H. Masuda, Y. Aoki, H. Shukuri and T. Hagiwara: IEEE Trans. Electron Devices **32** (1985) 2038.
[26] S. M. Sze and J. C. Irvin: Solid State Electron, **11** (1969) 599.
[27] S. M. Sze: *Physics of Semiconductor Devices* (Wiley, New York, 1981) 2nd ed, p.21.

7

Origins of leakage currents

Summary

In this chapter, we discuss the origins of leakage currents for flash memories and DRAMs. The charge loss of flash memories via detrapping and SILCs are caused by the oxide trap of 0.37 [eV] and 3.6 [eV], respectively. Trap levels of 0.37 [eV] and 3.6 [eV] are thought to originate in oxygen vacancies of dimer and fourfold configurations, respectively. For *tail bits* of DRAMs, the origin of anomalous leakage of *tail bits* is the silicon trap of 0.68 [eV]. By comparison with the previous data of silicon traps by metal contaminations, the elements which show close values of 0.677 [eV] are “Fe” and “Cu”. From the failure rate of DRAM cells, the contamination level of 0.01 [ppb] is obtained, which is below the purity level of silicon substrates.

7.1 Origins of leakage currents for flash memories

7.1.1 Oxygen vacancies in the oxide

Silicon dioxide has been the subject of extensive experimental investigations in both its crystalline and amorphous forms. Several of its properties are dominated by a single point defect, known as E_1' in crystalline quartz [1] and E_γ' in amorphous SiO_2 [2], with a characteristic electron-paramagnetic-resonance (EPR) signature.

Theoretical calculations [3–5] for small clusters, for crystalline quartz, and amorphous supercells led to the identification of the defect as the oxygen vacancy. This defect exhibits a unique bistability: In the neutral state, the two adjoining Si atoms rebond into a ‘‘dimer configuration’’, as shown in Fig. 7-1 (a). In the EPR-active positively charged state, one of the Si atoms relaxes back past the plane of its three O neighbors and bonds with another network O atom (‘‘puckered configuration’’), as shown in Fig. 7-1 (b).

In the past decade, the research on radiation or the high-field stress induced defects in the amorphous SiO_2 layer of MOS transistors has revealed a broad range of complex dynamical phenomena in both the bulk oxide film and near the Si-SiO_2 interface. The origin of many of these phenomena has been traced to O vacancies [6–10]. The underlying atomic scale processes, however, remain elusive. The phenomena are often associated with the dynamics of the E_γ' center and of a second EPR-active defect that has been labeled E_δ' and identified as an O vacancy in the dimer configuration [6, 7, 11].

Examples of these phenomena are as follows: (a) E_γ' centers are found to be more stable thermally than E_δ' centers; at room temperature, after irradiation or hole injection, the density of E_δ' centers has been observed to decrease, in some cases with a concomitant increase in the density of E_γ' centers [7]. (b) After irradiation or high field stress, the accumulated positive charge in the oxide can often be neutralized by a high-temperature anneal at large positive bias, but much of the original positive charge can be restored by the reverse-bias annealing [12–14]. The restored positive charge can cycle reversibly when the bias is cycled from positive to negative, but the total positive charge that gets restored during the cycling can gradually decrease with repeated cycling.

Z- Y. Lu et al. report ab initio total-energy calculations of O vacancies in several amorphous SiO_2 supercells [15]. The results and analysis reveal that O vacancies in amorphous SiO_2 exhibit a very rich structure because of the large variations in local bonding that are possible in the amorphous network as opposed to crystalline quartz. Three distinct types of O vacancies were found. Examinations of the variations in local bonding in the amorphous supercells allow an analysis of the statistical probability of

distinct modes of behavior. For the statistical study, the local topologies of a million-atom cell were analyzed by theoretical calculations.

The results are as follows.

(a) The vast majority (roughly 80%) are not bistable like O vacancies in quartz. Instead, the dimer configuration, as shown in Fig. 7-1 (a) is the only stable configuration in both the neutral and positively charged state. The dimer energy level containing one or two electrons is shallow.

(b) Roughly 12% are bistable in the same way as an O vacancy in quartz: In the neutral state, the dimer configuration is stable; in the positively charged state, one of the adjoining Si atoms relaxes back past the plane defined by its three O neighbors and bonds with another network O atom. The latter becomes threefold coordinated whereas the puckered Si atom becomes fourfold coordinated, as shown in Fig. 7-1 (b). The localized energy level is now quite deep, nearly in the middle of the SiO_2 energy gap. When this center is given an electron back, it stays in a metastable puckered configuration, with an energy barrier ranging from 0.2 to 1.2 [eV]. The electron, in fact, fills the dangling bond, making it negative, whereas the hole is still left on the puckered Si side, resulting in a dipole. The result that a defect with such small spatial extent can sustain such a dipole is somewhat surprising, but the idea was in fact invoked in previous studies [12–17].

(c) Roughly 8% are also bistable, but with the following key differences: In the positively charged state, the puckered Si atom now bonds both to a network O and a network Si atom, becoming fivefold coordinated; the localized energy level is again deep and the unpaired electron has essentially the same distribution as before (suggesting a nearly identical EPR signature). However, when this center is given back an electron, it collapses immediately to the dimer configuration without an energy barrier.

As a result, oxygen vacancies as electron traps have two stable states, i.e., the shallow state by the dimer configuration and the deep state by the fourfold puckered configuration. These results can account for many complex dynamical phenomena associated with electron capture and release after hole trapping in irradiated or stressed SiO_2. Existence of two stable states of oxygen vacancies should be attributed to existence of two leakage mechanisms of flash memories.

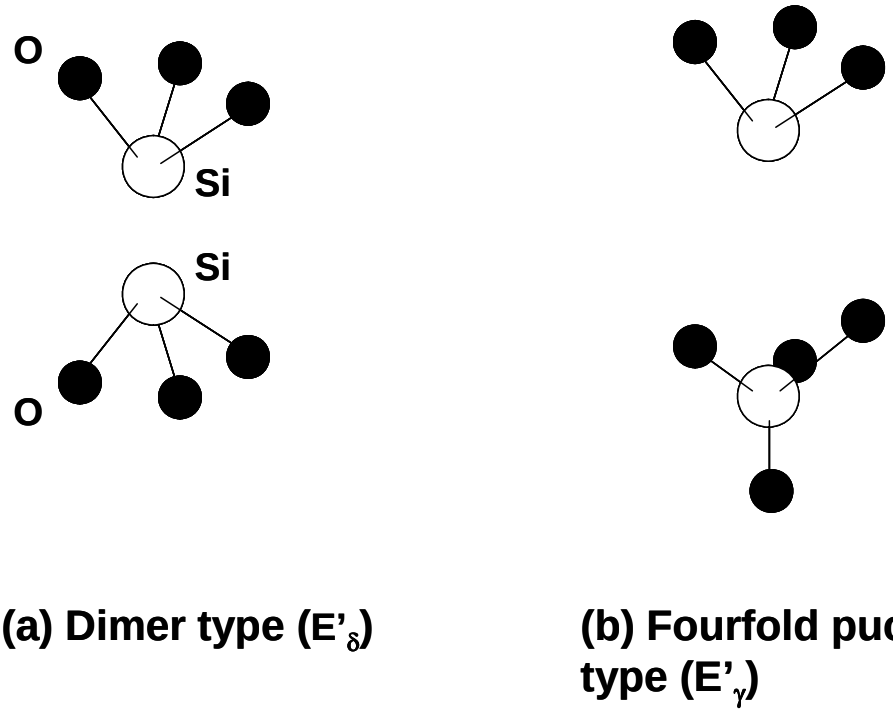


Fig. 7-1. Schematics of (a) dimmer, (b) fourfold puckered oxygen vacancies.

7.1.2 Anode hole injection model [18, 19]

The maximum potential energy that can be gained by the tunneling electrons in the silicon dioxide is $q(V_{ox}\text{-}\phi_n)$, where ϕ_n is the barrier height offset between the silicon and silicon dioxide. The value of $q(V_{ox}\text{-}\phi_n)$ is too small to cause impact ionization in the silicon dioxide. Therefore, the oxide breakdown is not a result of the hot electron effect in the silicon dioxide itself.

However, the electrons will enter the anode electrode with qV_{ox} of energy if scattering in the silicon dioxide is negligible as shown in Fig. 7-2. Since this energy can be substantially greater than the band-gap of silicon, it is possible to generate hole-electron pairs in the anode electrode [20]. It is thus possible to generate hot holes. A fraction of hot holes can tunnel into the silicon dioxide as originally proposed by Weinberg et al. [21]. These holes will tend to increase the cathode electric field by trapping in the oxide which in turn could lead to the oxide breakdown.

In the n^+-gate (n-type anode) case, the valence-band electrons must be excited to the bottom of the conduction band. However, in the p^+-gate (p-type anode) case, the valence-band electrons only need to excite to the top of the valence band since there are available states there. Thus, for the p^+-case the barrier height for hole tunneling will be lower than in the n^+-case by 1.1 [eV] (the silicon band-gap), as shown in Fig. 7-2. If the fraction of tunneling electrons which give up all of their energy by creating hole-electron pairs is α_H, then the hole-tunneling current arising from hole generation in the anode will be given by $J_p = \alpha_H J_n \theta_H$, where J_n is the electron tunneling current from the cathode and θ_H is the hole tunneling probability.

The charge associated with the trapped holes at time t is thus $Q_p(t) = \eta J_p t$ where η is the hole trapping efficiency. Assuming that breakdown occurs when a critical number of holes have tunneled into the SiO_2 and have been trapped [22], then Q_{BD} will be given by

$$Q_{BD} = J_n t_{BD} = C\theta^{-1} = C\exp\left[\frac{B'T_{ox}}{V_{ox}}\left(\phi_P(V_{ox})\right)^{3/2}\right], \tag{1}$$

where

$$\phi_P(V_{ox}) = E_{g,SiO2} - q\phi_n - qV_{ox} - E_{g,Si} \quad \text{for } p^+\text{gate}, \tag{2}$$

$$\phi_P(V_{ox}) = E_{g,SiO2} - q\phi_n - qV_{ox} \quad \text{for } n^+\text{gate}, \tag{3}$$

,V_{ox} is the voltage supplied to the oxide, C a constant, T_{ox} the oxide thickness, $B' = B/\phi_n^{3/2}$ the Fowler-Nordheim slope parameter, $E_{g,SiO2}$ and $E_{g,Si}$ the band-gap energies of silicon dioxide and silicon, respectively, ϕ_n and $\phi_p(V_{ox})$ the barrier heights for electrons and holes, respectively. Here, Equation (1) is derived assuming Fowler-Nordheim tunneling and, equal electron and hole effective masses. The field dependence in the pre-exponential factor of the Fowler-Nordheim expression has been neglected for simplicity, i.e., $J_n = A \exp [-BT_{ox}/ V_{ox}]$ where A is a material constant. The dependence of t_{BD} on V_{ox} can be determined by dividing Eq. (1) by J_n.

Hole injection by anode hole injection play a important role to activate oxygen vacancies as the electron traps.

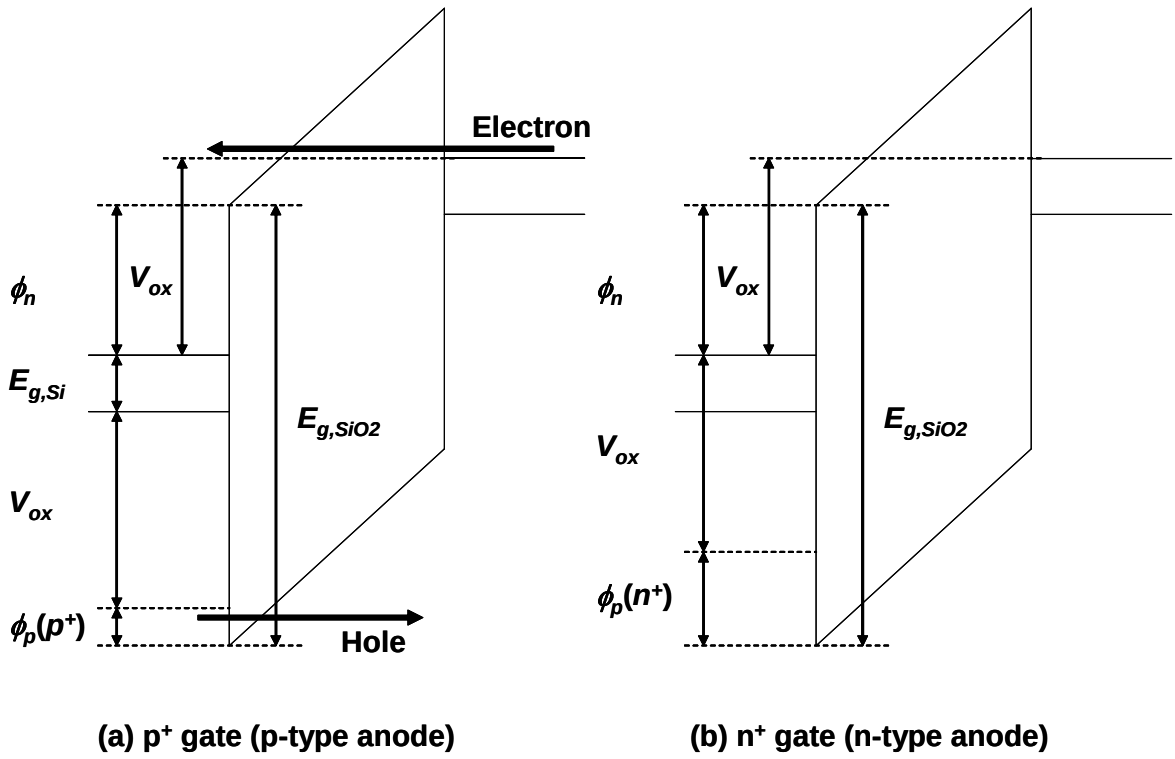


Fig. 7-2. Energy-band diagram depicting the mechanism of the anode hole injection.

7.1.3 Origins of leakage currents

While program operations of flash memories depend on the cell structures, i.e., the channel hot electron, the source side injection, and the Fowler-Nordheim (FN) tunneling, the erase operations is independent of the cell structure, i.e., FN-tunneling. The FN-tunneling for erase operation corresponds to the anode hole injection of p^+-gate, which is more efficient than that of n^+-gate. Therefore, during program/erase cycles, the holes are mainly injected during the erase operations.

Just after the production, i.e. without any stress, the tunnel oxide includes a specific amount of oxygen vacancies, which become the origin of electron traps. During the program/erase cycles, oxygen vacancies capture holes which are generated by anode hole injections. Oxygen vacancies which capture holes became three types of the electron traps, i.e., the dimer configuration, the fourfold coordinated configuration, and the fivefold coordinated configuration. Traps of the dimmer configuration are shallow, which becomes responsible traps for the detrapping. Traps of the fourfold coordinated configuration are deep, which become responsible traps for the stress induced leakage currents (SILC). Traps of the fivefold coordinated configuration become immediately dimmer configuration after captures of electrons.

Our analysis show that trap levels measured form the conduction band edge of the dimmer configuration and the fourfold coordinated configuration are 0.37 [eV] and 3.6 [eV], respectively, as shown in Fig.7-3. These values have not been verified by the *ab initio* calculations method, because the accuracy of the *ab initio* calculations is not sufficient to quantitative discussion of trap level. However, we believe oxygen vacancies of the dimmer and the fourfold coordinated configuration have the energy level of 0.37 [eV] and 3.6 [eV], respectively.

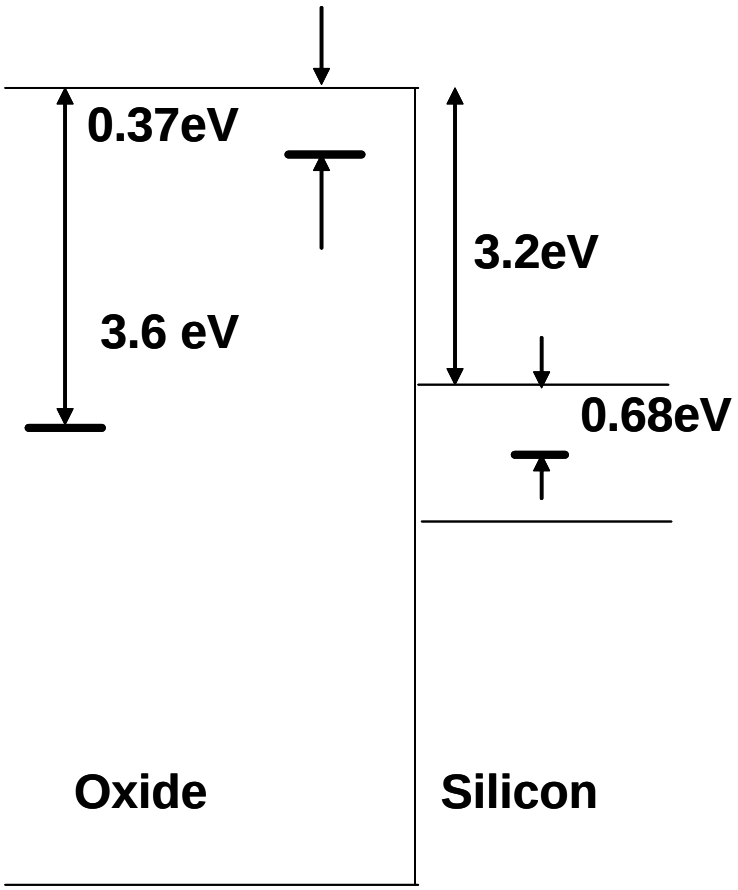


Fig. 7-3. Trap levels related with the data retention of semiconductor memories.

7.2 Origins of leakage currents for DRAMs

When a semiconductor is doped with donor or acceptor impurities, impurity energy levels are introduced [22]. A donor level is defined as being neutral if filled by an electron, and positive if empty. An acceptor level is neutral if empty, and negative if filled by an electron.

The simplest calculation of impurity energy levels is based on the hydrogen-atom model. The ionization energy for the hydrogen atom is

$$E_H = \frac{m_0 q^4}{32\pi^2 \varepsilon_0^2 \hbar^2} = 13.6\text{eV}, \tag{4}$$

where ε_0 is the free-space permittivity. The ionization energy for the donor E_d can be obtained by replacing m_0 to the conductivity effective mass [23] of electrons

$$m_{ce} = 3\left(\frac{1}{m_1^*} + \frac{1}{m_2^*} + \frac{1}{m_3^*}\right)^{-1}, \tag{5}$$

and by replacing ε_0 to the permittivity of the semiconductor ε_s in Eq. (4) as follows:

$$E_d = \left(\frac{\varepsilon_0}{\varepsilon_s}\right)^2 \left(\frac{m_{ce}}{m_0}\right) E_H . \tag{6}$$

The ionization energy for donors as calculated from Eq. (6) is 0.006 [eV] for Ge, 0.025 [eV] for Si, and 0.007[eV] for GaAs. The hydrogen-atom calculation for the ionization level for the acceptor is similar to that for the donors. We consider the unfilled valence band as a filled band plus an imaginary hole in the central force field of a negatively charged acceptor. The calculated acceptor ionization energy is 0.015 [eV] for Ge, 0.05 [eV] for Si, and about 0.05 [eV] for GaAs.

The simple hydrogen-atom model cannot account for the details of the ionization energy, particularly the deep levels in semiconductor [24-26]. However, the calculated values do predict the correct order of magnitude of the true ionization energies for shallow impurities. Figure 7-4 shows the measured ionization energies for various impurities in Si [22, 27]. Note that it is possible for a single atom to have many levels; for example, gold in Ge has three acceptor level and one donor level in the forbidden

energy gap [29].

As shown in Fig. 7-3, we have extracted the average trap level of 0.677 [eV] for the anomalous leakage currents of *tail bits*. One of the most plausible mechanisms for the trap generation is the contamination of transition metals during processing. The elements which show close values of 0.677 [eV] are “Fe” and “Cu” [26] [27]. To get good retention characteristics, hence, the reduction of the contamination, maybe “Fe” and “Cu”, is crucial in addition to the junction field reduction.

By using the failure rate of DRAM cells, the contamination density of about 1×10^{10} cm^{-3} can be approximately estimated. Purity of silicon substrate is 9.999999999%, or 0.1 [ppb], which means the contamination density of about 1×10^{10} cm^{-3}, or 0.01 [ppb], is below the purity level of silicon substrates. Therefore, the tail-distribution is unavoidable by reducing the contamination. Only possible solution to modify the retention characteristics is the reduction of the electric field and the increase of the repairable bits, which is the conventional methodology DRAM companies have used.

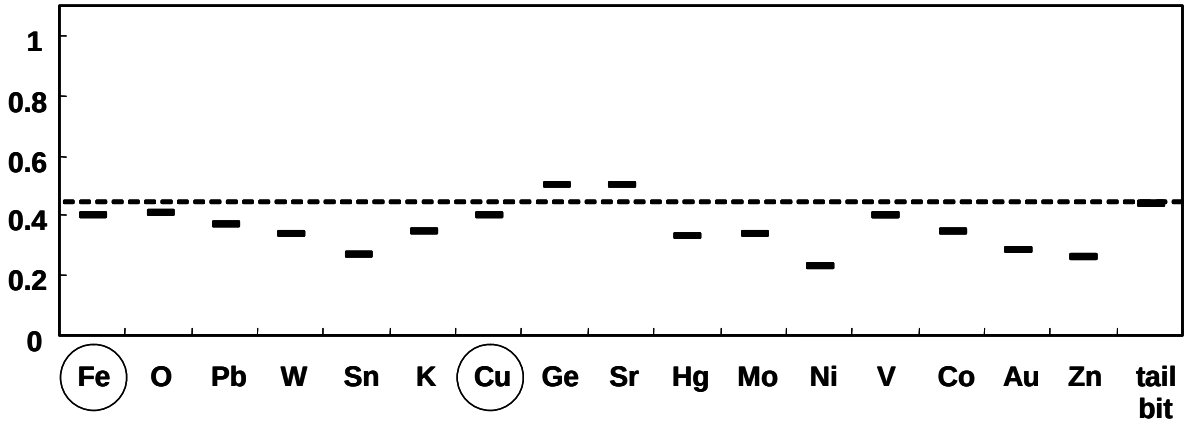


Fig. 7-4. Trap levels related metal contaminations in silicon substrate.

7.3 References

[1] R. A.Weeks: J. Appl. Phys. **27** (1956) 1376.
[2] D. L. Griscom: Phys. Rev. **B20** (1979) 1823.
[3] F. J. Feigl, W. B. Fowler, and K. L. Yip: Solid State Commun. **14** (1974) 225.
[4] J. K. Rudra and W. B. Fowler: Phys. Rev. **B35** (1987) 8223.
[5] M. Boero, A. Pasquarello, J. Sarnthein, and R. Car: Phys. Rev. Lett. **78** (1997) 887.
[6] W. L.Warren et al.: J. Electrochem. Soc. **139** (1992) 872.
[7] W. L. Warren et al.: IEEE Trans. Nucl. Sci. **41** (1994) 1817.
[8] E. P. O'Reilly and J. Robertson: Phys. Rev. **B27** (1983) 3780.
[9] P.M. Lenahan and P.V. Dressendorfer: J. Appl. Phys. **55** (1984) 3495.
[10] P.M. Lenahan and J.M. Conley, Jr.: IEEE Trans. Nucl. Sci. **45** (1998) 2413.
[11] J. R. Chavez et al.: IEEE Trans. Nucl. Sci. **44** (1997) 1799.
[12] J. R. Schwank et al.: IEEE Trans. Nucl. Sci. **31** (1984) 1434.
[13] A. J. Lelis, H. E. Boesch, T. R. Oldham, and F. B. McLean: IEEE Trans. Nucl. Sci. **35** (1988) 1186.
[14] D.M. Fleetwood et al.: Appl. Phys. Lett. **74** (1999) 2969.
[15] Z- Y. Lu, C. J. Nicklaw, D. M. Fleetwood, R. D. Schrimpf, and S. T. Pantelides: Phys. Rev. Lett. **89** (2002) 285505-1.
[16] A. Yokozawa, A. Oshiyama, Y. Miyamoto, and S. Kumashiro: IEDM Tech. Dig. (1997) 703.
[17] H. S. Kim, C. K. Williams and A. Reisman: J. Appl. Phys. **81** (1997) 1566.
[18] S. Holland, I. C. Chen and Chenming Hu: IEEE Elec. Dev. Let. **8** (1987) 572.
[19] K. F. Schuegraf and C. Hu: Proc of IRPS (1993) 7.
[19] C. Chang, C. Hu and R. W. Broderson: J. Appl. Phys. **57** (1985) 302.
[20] Z. A. Weinberg, W. C. Johnson and M. A. Lampert: J. Appl. Phys **47** (1976) 248.
[21] I. C. Chen, S. Holland and C. Hu: IEEE Electron Dev. Lett. **EDL-7** (1986) 164.
[22] S. M. Sze: *Physics of Semiconductor Devices* (Wiley, New York, 1981) 2nd ed, p.21.
[23] R. A. Smith: *Semiconductors* (Cambridge University Press, London, 1979) 2nd ed.
[24] A. G. Milnes: *Deep Impurities in Semiconductors* (Wiley, New York, 1973).
[25] J. Hermanson and J. C. Philips: Phys. Rev. **150** (1966) 652.
[26] J. Callaway and A. J. Hughers: Phys. Rev. **156** (1967) 860.
[27] S. M. Sze and J. C. Irvin: Solid State Electron **11** (1969) 599.
[28] W. M. Bullis: Solid State Electron **9** (1966) 143.

8

Prospects for conventional and emerging memories

Summary

In chapter 8, the prospects for conventional and emerging memories are remarked. Conventional memories, including NOR and NAND type flash memories, and DRAMs, would be believed that the data retention would be more serious problem for the future applications. Therefore, the new memories of the retention-problem-free are proposed including Ferroelectric RAMs (FeRAMs), Magnetoresistive RAMs (MRAMs) and Phase-Change RAMs (PRAMs). FeRAMs, MRAMs and PRAMs store the information via the polarization of the ferroelectric film, the magnetization of the magnetic tunnel junction (MTJ), the phases, i.e., crystalline or amorphous, of the Chalcogenide glass, respectively. These emerging memories possess nearly ideal properties, i.e., superior data retention, fast random access, virtually unlimited usage. Excellent functional properties of the new memories offer possibilities to displace the existing memories or create the new types of applications.

8.1 Introduction [1]

Conventional semiconductor memories such as DRAMs and flash memories have successfully evolved in the direction of high density, high performance and low cost. However, in the view of the growing technical complexity, fabrication cost and physical limit, there have been concerns about whether this successful progress can be maintained in the future [2, 3]. There have been extensive efforts to study the technical and physical limits of conventional memories and to find the ways to overcome the predicted technical barriers. The data retention is one of the most important roadblocks. Based on our studies, we discuss the prospects of the conventional memories.

Recently, many research groups and companies have tried different ways, developing new types of memories aiming less technical barriers and ideal memory characteristics such as non-volatility, high density, high speed and low power consumption, which none of the conventional memories can be satisfy at the same time [4, 5]. Among the many candidates of emerging new memories, Ferroelectric Random Access Memories (FeRAMs), Magnetoresistive Random Access Memories (MRAMs) and Phase change Random Access Memories (PRAMs) appears to be promising because it is expected that these can be commercialized because in the near future although it is longevity and technical barriers are not fully known.

8.2 Prospect of conventional memories

8.2.1 Prospect of flash memories in terms of data retention

The data retention characteristics of flash memories are determined by the detrapping and SILCs. In our study of data retention characteristics via detrapping, during program/erase cycles, oxygen vacancies of the dimer configuration capture holes injected via the anode-hole-injection and become active as electron traps. The anode-hole-injection is caused during FN tunneling of electrons. After oxygen vacancies become active, electrons are captured by these traps. The trap level is 0.37 [eV] measured from the conduction band edge. The density of oxygen vacancies are about 1×10^{12} [cm^{-2}]. Therefore, electrons of about 1×10^{12} [cm^{-2}] are captured inside the tunnel oxide. During the high temperature holding at 125 [˚C], captured electrons are released via the thermal excitation which causes the shift of the threshold voltage.

Fundamental solutions of the data leakage via detrapping are to decrease oxygen vacancies inside the tunnel-oxide and to stop to use FN tunneling for erase and/or program operations. To decrease oxygen vacancies, the post annealing by several atmospheres, i.e., NO, N_2O, H_2 and D_2, have been examined, which shows only the small effect within the author’s knowledge. The erase operation is only possible by

FN-tunneling. Therefore, the fundamental solutions are hard to find.

In our study of data retention characteristics via SILCs, electrons in the floating gate flow out through the multiple-trap path. The area density of oxygen vacancies are about 1×10^{12} [cm^{-2}], which would be order of 1×10^{18} [cm^{-3}]. However, the oxygen vacancies of the dimer configuration statistically get lined up from the cathode to the anode against long odds, which we call the multiple-trap-path or the weak spot. The area density of the multiple-trap-path is about 5.0×10^{2} cm^{-2}. During program/erase cycles, oxygen vacancies of the dimer configuration capture holes injected via the anode-hole-injection and become the fourfold configuration. The trap level of the fourfold configuration is 3.6 [eV]. The flash cells which contain the multiple-trap-path inside the tunnel oxide become *tail cells* by B-mode SILC. The *tail cells* show the anomalous leakage currents at the room temperature.

Empirically the tunnel oxides of thicker than 8nm can avoid B-mode SILC. By using oxide of thicker than 8nm, we can drastically reduce the probability to form the multiple-trap-path statistically. This means the tunnel oxide thickness must be constant even if the memory cells scale down, which deteriorates the short channel effect of MOSFETs and increases the amplitude of the random telegraph signal which reduces V_{th} windows.

8.2.2 Other prospects of flash memories

Fundamental limitation in NOR flash memory scaling comes from the non-scalable Si-SiO_2 energy barrier height [6, 7]. NOR flash memory requires high drain voltage for the efficient channel hot electron generation and the injection of hot electrons into floating gate over the Si-SiO_2 energy barrier. But high drain voltage needed for programming can cause unwanted drain disturbance in the other cells connected to the same bit-line. Drain disturbance problem includes cell transistor current leakage and hot hole injection into the floating gates in the unselected cells connected to the high voltage bit-line. As we scale down the cell transistor dimension, the transistor suffers severe short channel effect and need high channel doping concentration, which reduces the drain voltage allowed to prevent the drain disturbance. However drain voltage scaling with the cell transistor dimension cannot be continued since the drain voltage cannot be lower than the Si-SiO_2 barrier height of 3.1eV and this imposes the fundamental limit of NOR flash memory scaling. This limit is expected to be critical below 65nm technology node.

NAND flash has more scaling capability than DRAMs and NOR Flash because of its compact cell architecture and its different device physics [8]. But cell-to-cell

interference (the inter-floating-gate coupling) in NAND flash memory eventually limits the technology scaling. Word line space is affected by the states of adjacent cells due to capacitive coupling between floating gates, which results in over-program or under-erase failures [9]. The floating gate interference for the coupling ratio increase as word line space becomes narrow. We can see that floating gate interference coupling ratio steeply increase when the word line space is narrower than 60nm.

Among digital devices, floating-gate (FG) flash memory is considered to be the first device that is affected by the random telegraph signal (RTS) during device scale-down by following reason [10]. Since the tunnel oxide thickness cannot be decreased due to stress-induced leakage current, C_{ox} is constant during scaling down. Moreover, the gate length and width are as small as a feature size, which is almost the smallest among the MOSFETs fabricated by same technology node. From the above characteristics, the ΔV_{th} due to the RTS of the FG flash memory is the largest among digital devices fabricated by the same technology node.

Windows of threshold voltage, V_{th}, become smaller for every flash generation because of the capacitive coupling between floating-gates, the random telegraph signal and the variety of disturb mechanism. In near future, flash memories will not maintain the sufficient V_{th}-window. Under such circumstances, relax of the specification of program/erase number are only solutions for the data retention problem.

To scale down the oxide thickness, one solution is to use Si_2N_3 layer as the storage, i.e., SONOS structure, instead of the floating gate. The electrons stored in Si_2N_3 layer are captured by traps. Therefore, the electron can not flow out from Si_2N_3 layer even if the multiple-trap-path is generated. Furthermore, if we use SONOS type flash cell structure [11], we can drastically reduce the interference arising from the capacitive coupling between floating gates. The SONOS cell will be the solutions for the technology node below 65nm. However, SONOS cell have the other difficulties as the memory devices, which is out of scope in this thesis. So we do not think SONOS cell is fundamental solutions for the future semiconductor memories.

8.2.3 Prospect of DRAMs memories in terms of data retention

One of the most plausible mechanisms for the trap generation is the contamination of transition metals. The elements which show close values of 0.677 [eV] are “Fe” and “Cu”.

By using the failure rate of DRAM cells, the contamination density of about 1×10^{10} [cm^{-3}] can be approximately estimated. Purity of silicon substrate is 9.999999999%, or 0.1 [ppb], which means the contamination density of about 1×10^{10}

cm^{-3}, or 0.01 [ppb], is below the purity level of silicon substrates. Therefore, the *tail*-distribution is unavoidable by reducing the contamination. Only possible solution to modify the retention characteristics is the reduction of the electric field and the increase of the repairable bits, which is the conventional methodology DRAM companies have used.

As we scale down the DRAM cells dimension, cell array transistor will be the most critical. Data retention time of DRAMs has always been under pressure of ever-increasing trend and is critically affected by dimension of the cell array transistor [12]. Decreased transistor channel length requires increased channel doping concentrations to prevent channel punch-through. But the increased doping concentration is accompanied with the increased of electric field across the junctions boundary and the increase of junction leakage current which results in decrease of data retention time. We expect significant degradation of data retention time below 90nm node due to rapid increase in junction electric field.

8.2.4 Summary of prospect of conventional memories

The data retention of flash memories and DRAMs would be attributed to the existence of oxygen vacancies and the metal contamination of 0.01 [ppb], respectively. We can not remove these fundamental origins of the charge loss. We must continuously rely on the conventional methodology to modify the data retention for every successive generation. Shrink of devices make it difficult to satisfy the specification of memories by only relying on the conventional methodologies. Therefore, development of new semiconductor memories including FeRAMs, MRAMs and PRAMs, are crucial in near future.

8.3 Emerging new memories

Table 8-1 lists the brief summary of the emerging memories, i.e., Ferroelectric RAMs (FeRAMs), Magnetoresistive RAMs (MRAMs) and Phase-Change RAMs (PRAM). As shown in Table 8-1, FeRAMs, MRAMs and PRAMs store the information via the polarization of the ferroelectric film, the magnetization of the magnetic tunnel junction (MTJ), and the phases, i.e., crystalline or amorphous, of the Chalcogenide glass, respectively. These new types of memories possess nearly ideal properties, i.e., superior data retention, fast random access, virtually unlimited usage as shown in Table 8-2 [4], [13]-[15].

Excellent functional properties of the new memories offer possibilities to displace the existing memories or create the new types of applications. For example, new types of memories can revamp memory management in systems, i.e., from the complicated memory solution which uses many different types of memories at the same time, to simple memory solution where only a single type of new memory may be of necessity. However, the technical leverage of the new types of memories is at early stage of verification. To realize the great potential capacities of new memories, they seem to need further investigations and breakthroughs in technology.

Table 8.1 Brief summary of emerging semiconductor memories.

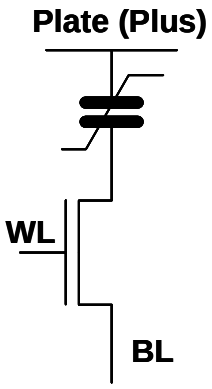
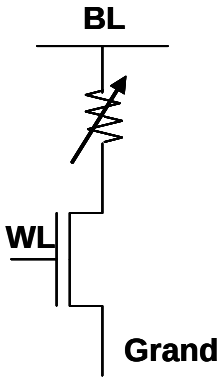
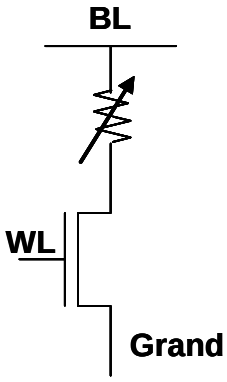
	FeRAM	MRAM	PRAM
Equivalent Circuit			
Storage Node	Ferroelectric Film	Magnetic Tunnel Junction (MTJ)	Chalcogenide Glass
Storage Element	Polarization	magnetization	Phase (crystalline/ Amorphous)
Data Retention	Superior	Superior	Superior

Table 8-2 Properties of emerging semiconductor memories.

	Non Volatile RAM (Emerging)		
	FeRAM	**MRAM**	**PRAM**
Endurance	**10^{12}-10^{16}**	**>10^{15}**	**>10^{12}**
Write	**40n-100n**	**20n-100n**	**10n-50n**
Read	**40n-100n**	**20n-100n**	**20n**
Cell Size	**$20F^2$**	**$20F^2$**	**$10F^2$**
Density	**64M/128M**	**64M/256M**	**64M**
Power supply voltage	**1.8V-2.5V**	**1.8V-3.3V**	**1.8V-3.3V**
Program voltage	**1.2V-2.5V**	**5V-10V**	**1.2V-2.5V**
Aplication	**Embedded (Secure)**	**Embedded (Car)**	**Embedded (Cellular)**
Refresh	**Non**	**Non**	**Non**

8.3.1 Ferroelectric RAMs (FeRAMs)

Figure 8-1 shows the cross section view and the equivalent circuit of FeRAMs. FeRAMs store the information by using the polarization of the ferroelectric film. The polarizations are detectable by the sense amplifier via voltages of bit lines versus the reference voltage.

FeRAMs are to utilize the positive or negative remnant polarization charge state of ferroelectric dielectrics as either data “1” or data “0”, as shown in Fig. 8-2. In the case of FeRAMs technology scaling, ferroelectric capacitor appears to be more critical limiting factor than the array transistor. Sensing signal of FeRAMs is proportional to capacitor area and remnant polarization charges of ferroelectric film. Therefore, when we scale down the FeRAM cell area for high density memory, we need technologies to at least maintain or increase the capacitor area and the remnant polarization charges of the ferroelectric films to ensure proper signal sensing at the same time.

Figure 8-3 shows schematics of FeRAM cell area scaling trend. With planar ferroelectric capacitor, we can scale down the cell area by improving the capacitor slope and by reducing the ferroelectric film thickness while maintaining the same signal sensing margin. But for further scaling down the cell area, we need to follow the similar way which DRAMs capacitor already experienced, i.e. we need three dimensional capacitor structures. For this purpose, securing ferroelectric film technology with nano-scaled thickness and excellent conformal deposition capability along the inside walls of a high aspect-ratio trench will be essential.

Application suitable for FeRAMs is the security data storage, because the data stored in FeRAMs is hard to be read out.

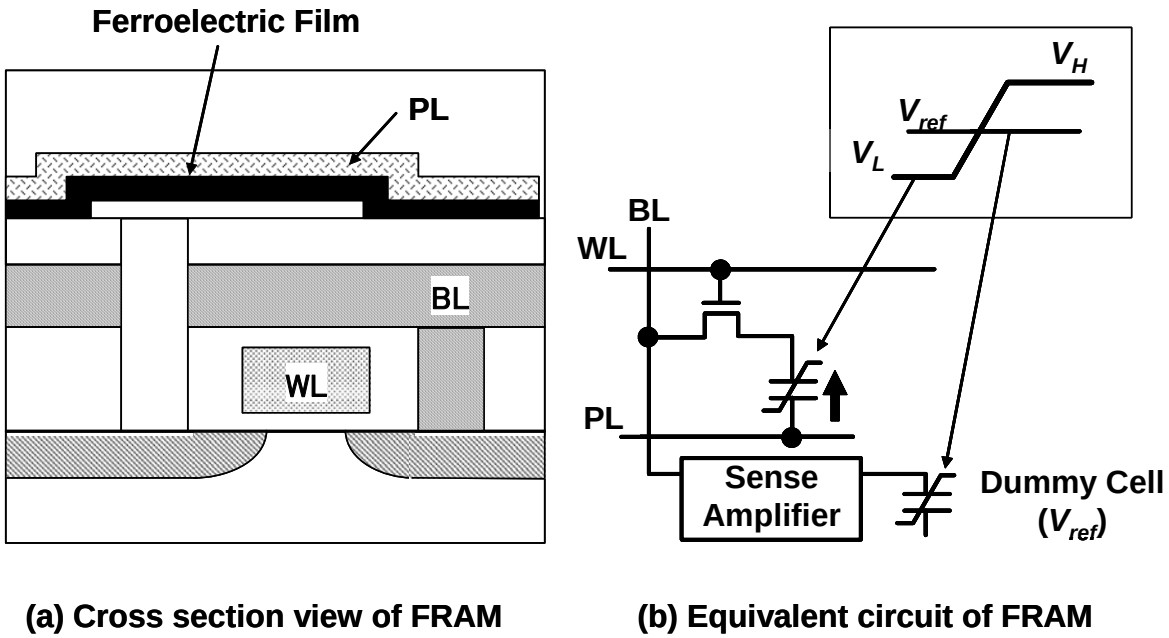


Fig. 8-1. Cross section view and the equivalent circuit of FeRAMs.

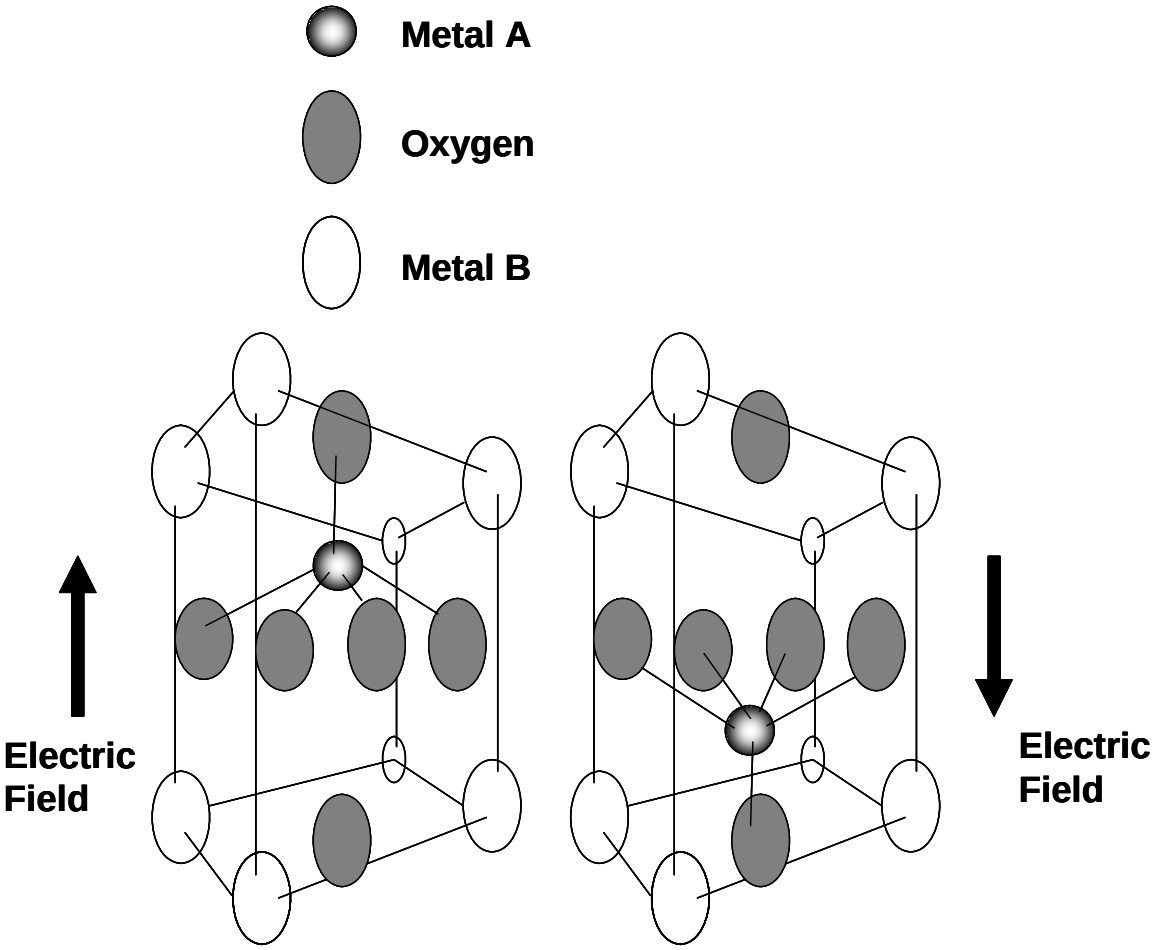


Fig. 8-2. Polarization of ferroelectric film, i.e., perovskite compound film.

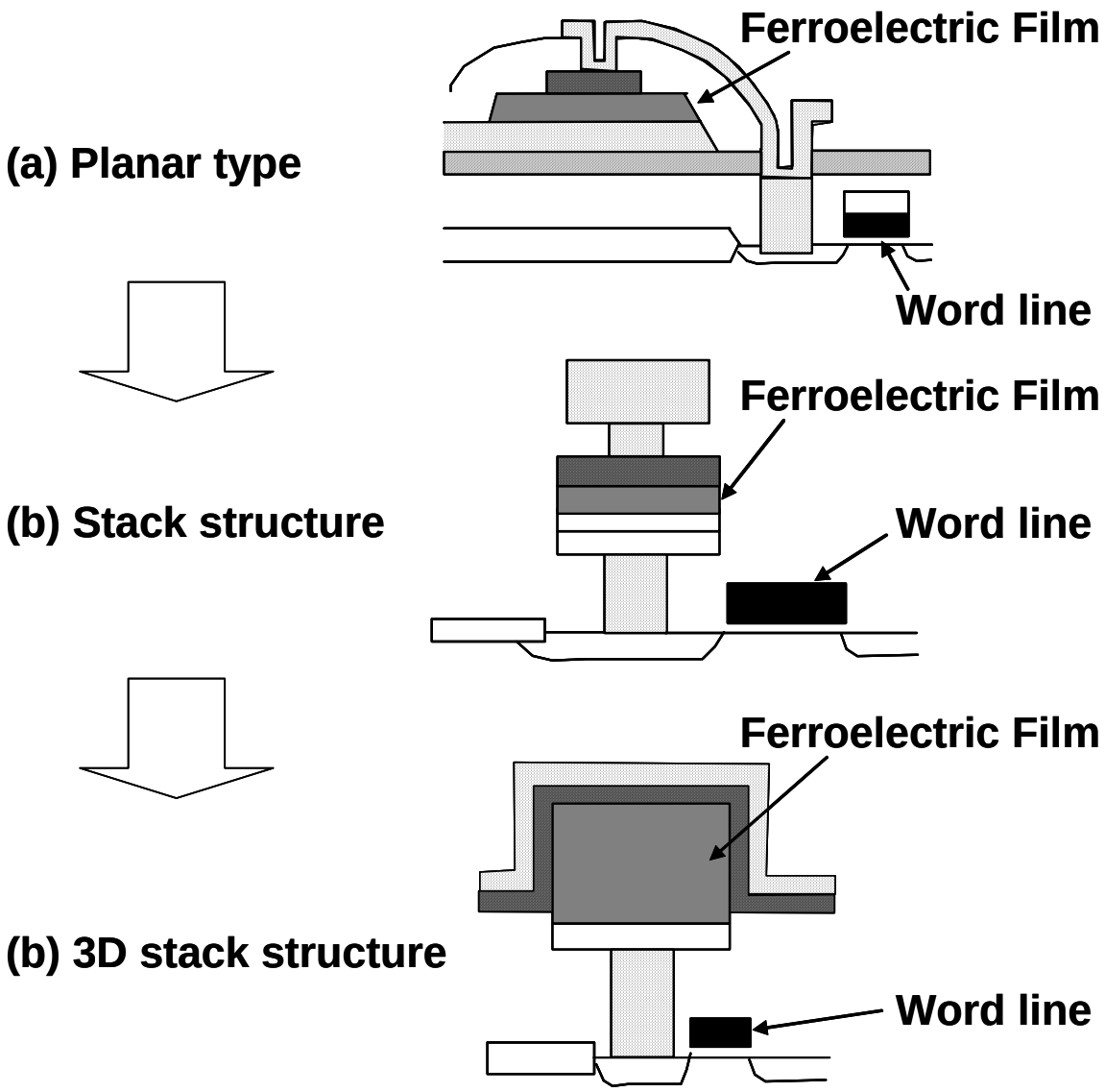


Fig. 8-3. Schematics of FRAM cell area scaling trend.

8.3.2 Magnetoresistive RAMs (MRAMs)

Figure 8-4 shows the cross section view and the equivalent circuit of MRAMs. MRAMs store the information by using the magnetization of the magnetic tunnel junction (MTJ). The magnetizations are detectable from voltages of the bit lines.

MRAMs are to utilize the resistance change of MTJ (Magnetic tunnel junction), which is determined by relative magnetization directions of two ferromagnetic electrodes, as shown Fig.8-5. The parallel magnetizations show the low resistance, while the anti-parallel magnetizations show the high resistance. Owing to the excellent performance such as non-volatility, fast read/write time, unlimited write endurance, MRAMs are considered to be a strong candidate for the future ideal memory [14]. However MRAMs are facing many technical barriers to be solved for the present. The urgent issues of obtaining uniform MTJ resistance and preventing writing disturbance to unselected cells seem to be solved considering the progress made by many research groups [16-19].

In view of making high density MRAMs, the most critical limitation comes from the large cell size and poor scalability of MTJ size. MRAM cell structure has additional cell area is needed for MTJ connection. As a result, MRAM has typically large cell are factor of $20F^2$, which makes it hard to achieve high density MRAMs. Recently, new MRAMs structure was suggested to solve this problem. With the given size of MTJ, we can reduce the cell area to $8F^2$ by adopting new MRAM cell structure with split digit-lines [1]. The fundamental scalability limit of MTJ comes from the fact that the smaller the MTJ size, the higher switching field is required [1].

Large switching field requires large writing current and causes the increase in power consumption and chip size and the reliability issues arising from the electro-migration phenomena in the metal lines. The writing current can be reduced by using magnetic flux concentrating structures in digit-line and bit-lines [18] and optimization of MTJ magnetic properties [19]. Future of MRAM technology scaling seems to depend on how far we can shrink the MRAM writing current.

Application suitable for MRAMs is the embedded MCU for the automobile, because MRAMs shows a good stability even on the high temperature of more than 150˚C.

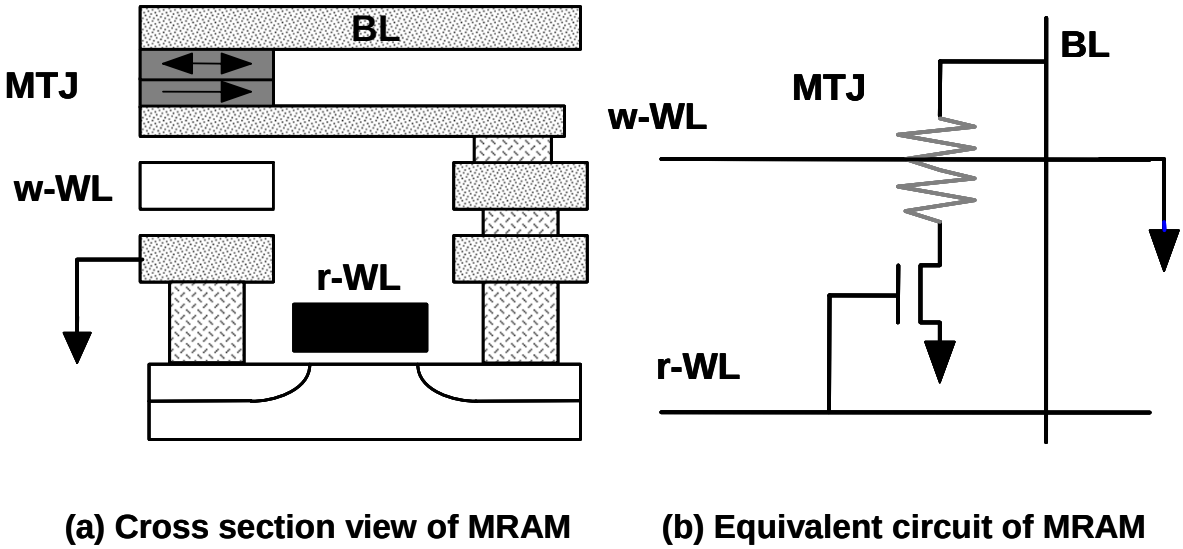


Fig. 8-4. Cross section view and equivalent circuit of MRAMs.

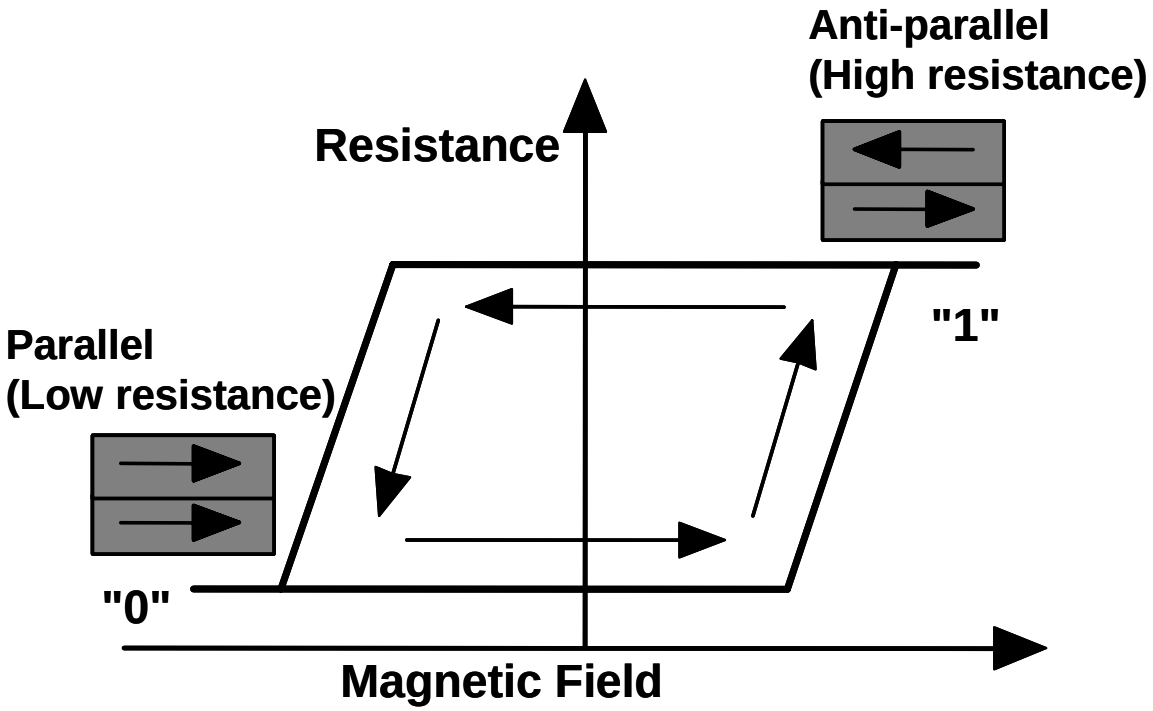


Fig. 8-5. Resistance change of MTJ (Magnetic tunnel junction), which is determined by relative magnetization directions of two ferromagnetic electrodes.

8.3.3 Phase change RAMs (PRAMs)

Figure 8-6 shows the cross section view and the equivalent circuit of PRAMs. PRAMs store the information by using the phases, i.e., crystalline or amorphous, of the Chalcogenide glass [20]. The phases are detectable from voltages of the bit lines.

By proper control of heating current, we can reversibly amorphize (reset) or crystallize (set) the GST. Resistance ratio between reset and set states is about $10\sim10^4$, which gives large signal sensing margin. There are no physical limits to PRAM scaling down to CMOS limit as far as we know. The most important technical barrier to the PRAM scaling comes from large current consumption during reset process of GST in order to change the crystalline state into the amorphous state. Large reset current requires a large size of cell transistor, which makes PRAM cell size larger than ideal size of $10F^2$ and makes it hard to achieve high density PRAMs.

Another barrier to the PRAM scaling is that set resistance shows tendency to rise as we decrease the reset current. Lager set resistance reduces the sensing signal and results in worse noise immunity and longer reading time. Approaches to reduce reset current may be categorized as the reduction of cell dimension, the modification of cell structure and the GST material. It is known that a reset current can be scaled down with GST size scaling and GST-electrode contact size scaling [21]. Thus PRAMs has advantages in cell size scaling over other memories like MRAMs where cell size scaling increases writing current.

In order to reduce reset current with the contact dimension shrink, the various ideas and techniques have been proposed. For instance, an edge contact structure [22] can reduce the reset current as much as 50% and more. We can also reduce the reset current by modifications of GST material. The reset current of nitrogen doped GST is reduced to half compared to that of un-doped GST [23]. Reduction of the reset current while suppressing the set resistance below tolerance limit will be key success factors of future PRAM scaling.

Application suitable for PRAMs is the embedded MCU for the cellular phone, because PRAM shows the much higher program speed than NOR-type flash memory.

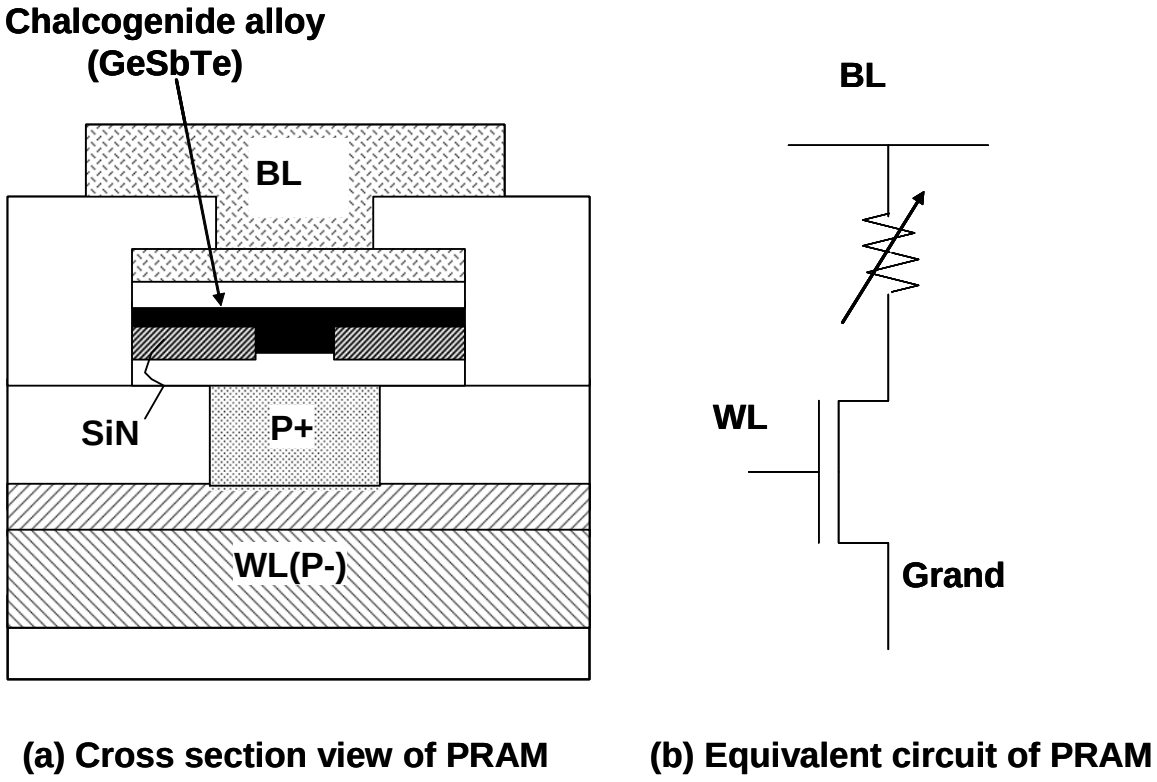


Fig. 8-6. Cross section view and equivalent circuit of PRAMs.

8.4 References

[1] K. Kim and G- H. Koh: Proc of International Conf. on Microelectronics (2004) 377.
[2] S Thomson et al.: Intel Tech. Journal (1998).
[3] ITRS Road map (2006).
[4] B. Prince: *Emerging memories: Technologies and trends* (Kluwer Academic Publications, Massachusetts, 2002).
[5] J. A. Hutchby et al.: IEEE Circuits and Device Magazine (2002) 28.
[6] P. Pavan et al.: Proc. IEEE **91** (2003) 1248.
[7] R. Bez et al.: Proc. IEEE **91** (2003) 489.
[8] P. Pavan et al.: Proc of IEEE **85** (1997) 1248.
[9] J. D. Lee et al.: Electron Device Lett. **23** (2002) 264.
[10] N. Tega, H. Miki, T. Osabe, A. Kotabe, K. Otsuga, H. Kurata, S. Kamohara, Y. Ikeda, R. Yamamda: IEDM Tech. Dig. (2006) 1.
[11] C. H. Lee et al.: IEDM Tech. Dig. (2003) 613.
[12] K. Kim et al.: IEEE Trans. Electron Dev. **45** (1998) 598.
[13] K. Kim and Y. J. Song: Micro-electronics Reliability **43** (2003) 385.
[14] S. Tehrani et al.: Proc. IEEE **91** (2003) 703.
[15] W. Reohr et al.: IEEE Circuits & Device Magazine (2002) 17.
[16] W. Motoyoshi et al.: VLSI Symp. Tech. Dig. (2002) 212.
[17] M. Durlam et al.: IEDM Tech. Dig. (2003) 995.
[18] M. Durlam et al.: J. Solid0state Circuit **38** (2003) 769.
[19] K. Tsuji: IEDM Tech. Dig. (2001) 36.4.1.
[20] M. Gill et al.: ISSCC Tech. Dig. (2002) 202.
[21] Y. N. Hwang et al.: IEDM Tech. Dig. (2004) 893.
[22] Y. H. Ha et al.: VLSI Symp. Tech. Dig. (2003) 175.
[23] H. Horii et al.: VLSI Symp. Tech. Dig. (2003) 177.

9

Conclusions

Summary

The data retention of flash memories can be attributed to the detrapping from the trap with an energy level of 0.37 [eV] as well as the stress induced leakage currents (SILCs) via 3.6 [eV] trap in the oxide layer. The origins of these traps can be attributed to the oxygen vacancies. Oxygen vacancies basically exist in the oxide and are hardly removed. On the other hand, the retention of DRAMs originates 0.68 [eV]-trap in the Si p-n junction. This trap can be ascribed to the contaminated Fe atoms on the concentration order of 0.01 [ppb]. This level of contamination is below the purity of the silicon substrate. As a result, we can not remove these fundamental origins of charge loss for both flash memories and DRAMs. Therefore, development of new semiconductor memories including FeRAMs, MRAMs and PRAMs, are crucial in near future.

9.1 Data retention of flash memories

I have shown that the data retention of flash memories is determined by the electron detrapping and the stress induce leakage currents. During program/erase operations, electrons are captured by the traps and multiple-trap-paths are generated. The trapped electrons are thermally exited and released from traps during the high temperature retention test. On the other hand, on the low temperature retention test, electrons inside floating gates are released via multiple-trap-paths.

Retention mechanism via detrapping

I proposed the model for the time evolution of the threshold voltage due to the detrapping as follows: During program/erase operations, electrons are captured by the traps. Suppose that an electron trapped at a certain site is thermally excited into the conduction band and is emitted from its trap site. After the electron detrapping, the trap would be positively charged and couples with the nearest trapped electron. The positively ionized trap and the coupled electron are stabilized by the Columbic interaction. It is assumed that after electron detrapping, the positively ionized trap reduces the probability of the electron in the influence area, *b*, being emitted from its site. Present model for electron detrapping is similar to the models introduced for the specific cases of chemisorption, oxidation and so on.

Analytical model equation of detrapping

I derived a simple analytical model for the time evolution of the threshold voltage due to the detrapping. Based on the model I proposed, the time evolution of the threshold-voltage, ΔV_{th} , can be described as follows:

$$\Delta Vth(t) = -\alpha^* \ln\left(\frac{t}{\tau^*}\right), \qquad (1)$$

where $\tau^* = \tau^*_0 \exp(E_t / kT)$, $\alpha^* = q/(2bC_{CR}(\varepsilon_{ox} / t_{ox}))$, b is the influence area, C_{CR} the coupling ratio between the gate and the floating gate, E_t the trap level, T the temperature, τ^*_0 the renormalized time constant, and t_{ox} the oxide thickness of the tunnel oxide. Equation (1) has successfully reproduced the experimental results, which means the validity of our model.

By comparisons between models and experiments, the following parameter

values are obtained: E_t=0.37 [eV], and b=1.31×10 $^{-11}$ [cm^2], i.e., radius of the influence area, r=20.3 [nm].

Origin of detrapping

From the value of trap level, i.e, 0.37 [eV], I have identified the origin of the detrapping is the active oxygen vacancies of the dimer configuration. During program/erase cycles, oxygen vacancies of the dimer configuration capture holes injected by the anode-hole-injection and become active as electron traps. The anode-hole-injection is caused during programs and/or erases, i.e, the Fowler-Nordheim (FN) tunneling of electrons. As a result, the tunnel oxide is negatively charged after program/erase cycles. The trap level of the active dimer configuration would be 0.37 [eV] measured from the conduction band edge. The density of oxygen vacancies are about 1×10^{12} [cm^{-2}]. Therefore, electrons of about 1×10^{12} [cm^{-2}] are captured inside the tunnel oxide. During the high temperature holding at 125 [˚C], captured electrons are released via the thermal excitation which causes the shift of the threshold voltage.

Retention mechanism via SILCs

I proposed the mode for the charge loss via SILCs as follows: During program/erase cycles, multiple-trap-paths are generated. Electrons inside the floating gate flow out to the silicon substrate via the multiple-trap-path. I named the stress induced leakage currents via the multiple-trap path as B-mode SILC. Tunneling from the trap to the anode becomes the limiting process of the leakage current when the trap level, E_t, is deeper than the anode conduction band, $E_{g,A}$. Under this assumption, the leakage currents via multiple-trap-paths can be described by the tunneling of trapezoidal potential and show the FN field-dependence.

Analytical model equation of B-mode SILCs

I derived a simple analytical model for the stress induced leakage currents via the multiple-trap path, i.e, B-mode SILC. Based on the present model, B-mode SILC can be analytically described by

$$J_{SILC} = \frac{qN_{t,B}}{\tau}\exp\left\{-\frac{4}{3}\frac{(2m_{ox})^{0.5}}{\hbar}\frac{1}{qE_{ox}}\left(E_t^{\,3/2} - E_{g,A}^{\,3/2}\right)\right\}, \qquad (2)$$

where $N_{t,B}$ is the area density of multiple-trap-path, E_t the trap level measured from the

conduction band of oxide and the other parameters are explained in chapter 4. I have confirmed the accuracy of our model by comparisons between the simple analytical model and the experimental results.

By comparisons between models and experiments, the following parameter values are obtained: E_t=3.6 [eV], and $N_{t,B}$ =5.0×10 ^{2} cm^{-2}.

Origin of B-mode SILCs

From the trap level of 3.7 [eV], I have identified the origin of the multiple-trap-path is the active oxygen vacancies of the fourfold configuration. The area density of oxygen vacancies are about 1×10^{12} [cm^{-2}], whose bulk density would be order of 1×10^{18} [cm^{-3}]. The oxygen vacancies of the dimer configuration statistically get lined up from the cathode to the anode against long odds, which is called as the multiple-trap-path or the weak spot. The area density of the multiple-trap-path is about 5.0×10^{2} [cm^{-2}]. During program/erase cycles, oxygen vacancies of the dimer configuration capture holes injected via the anode-hole-injection and become the fourfold configuration. The trap level of the fourfold configuration is 3.6 [eV]. The flash cells which contain the multiple-trap-path inside the tunnel oxide become *tail* cells. The *tail* cells show the anomalous leakage at the room temperature.

Prospects of flash memories

Based on the present studies, I discussed prospects of flash memories. Empirically, the tunnel oxides of thicker than 8nm can avoid B-mode SILC. That is why by using oxides of thicker than 8nm we can drastically reduce the probability to form the multiple-trap-path statistically. This means the tunnel oxide thickness must be keep constant even if the memory cells scale down, which deteriorates the short channel effect of MOSFETs and increases the random telegraph signal which reduces V_{th} windows.

Fundamental solutions of the detrapping are to decrease oxygen vacancies inside the tunnel-oxide or to stop to use FN tunneling for both erase and program operations. To decrease oxygen vacancies, the post annealing by several atmospheres, i.e., NO, N_2O, H_2 and D_2, has been examined, which shows only a small effect within the author's knowledge. The erase operation is only realized by FN-tunneling. Therefore, the fundamental solutions are hard to be found out. Windows of threshold voltage, V_{th}, become smaller for every flash generation because of the capacitive coupling between floating gates, the random telegraph signal and the variety of disturb mechanism. In the near future, flash memories will be not able to maintain the sufficient V_{th}-window.

Under such circumstances, relax of the specification of program/erase number is only solution for the data retention.

Only solution is to use Si_2N_3 layer as the storage, i.e., a MONOS structure, instead of the floating gate. For MONOS cells, electrons are stored in traps of the Si_2N_3 layer. Therefore, electrons can not flow out from the Si_2N_3 layer even if the multiple-trap-path is generated. We can scale down the oxide thickness against B-mode SILC, which means the reduction of the random telegraph signal. In addition, the capacitive coupling between floating gates is drastically reduced. Therefore, the sufficient V_{th} window will be easily obtained even if the detrapping exists. However, a MONOS structure cannot overcome the problems of the conventional flash memories.

9.2 Data retention of DRAMs

I have shown the refresh time of DRAMs is governed by *tail bits*. The *tail bits* show the anomalous leakage currents and the distribution for their leakage level. Until now the mechanism of *tail bits* was not clear.

Retention mechanism of tail bits

I proposed the model of the leakage characteristics of *tail bits* and their distribution. In order to explain the anomalous leakage currents of *tail bits*, I assume that memory bits containing a specific trap center in the depletion layer become *tail bits*. Since the number of *tail bits* is significantly small, it is provided that a particular *tail bit* contains only one trap center. To explain the origin of the *tail*-distribution, it is not irrational to speculate that the location of a trap center and hence its surroundings statistically distribute in the p-n junction. Consequently, the effective energy level of the trap might be perturbed by the lattice distortion due to the stress and the other physical condition around the trap.

Analytical model equation of tail bits

I derived simple analytical models for both the leakage characteristics of *tail bits* and their distribution. Anomalous leakage currents of *tail bits*, I_L, can be described by the extended SRH recombination currents. The retention time of the specific cell is described by $T_{ret}=Q/I_L$, where Q is the charge stored in the capacitor. Finally, the *tail*-distribution can be described by

$$\int f(T_{ret})dT_{ret} = \int (\eta N_{cell} / \sqrt{2\pi}\delta) \exp(-\frac{(E_t - \langle E_t \rangle)^2}{2\delta^2}) \frac{\partial E_t}{\partial T_{ret}} dT_{ret}, \qquad (3)$$

where $<E_t>$ is the average trap level of *tail bits*, η ratio of *tail bits* to total bits, δ the root mean square (RMS) for the trap levels of *tail bits*, and N_{cell} the total bits number of DRAM chip. I have confirmed the accuracy of our model by comparisons between the simple analytical model and the experimental results.

By comparisons between models and experiments, extracted values of $<E_t>$ and δ are 0.677 [eV] and 0.025 [eV], respectively.

Origin of tail bits

From the value of trap level, i.e, 0.677 [eV], I have identified the origin of the anomalous leakage of *tail bits* is the contamination of transition metal impurities. The elements which show close values of 0.677 [eV] are "Fe" and "Cu". For some reasons, we believe the "Fe" contamination is the most plausible.

Prospects of DRAMs

Based on the present studies, I discussed prospects of DRAMs. By using the failure rate of DRAM cells, the contamination density of about 1×10^{10} cm^{-3} can be approximately estimated. Purity of silicon substrate is 9.999999999%, or 0.1 [ppb], which means the contamination density of about 1×10^{10} cm^{-3}, or 0.01 [ppb], is below the purity level of silicon substrates. Therefore, the tail-distribution is unavoidable up to now by reducing the contamination. Only possible solution to modify the retention characteristics is the reduction of the electric field and the increase of the repairable bits, which is the conventional methodology DRAM companies have used.

As we scale down the DRAM cell dimension, cell array transistors will be the most critical. The refresh time of DRAMs has always been under pressure of ever-increasing trend and is critically affected by dimension of the cell array transistor. The decreased transistor channel length requires increased channel doping concentrations to prevent channel punch-through. But the increased doping concentration is accompanied with an increase of the electric field across the junction boundary and an increase of junction leakage current which results in a decrease of the data retention time. We expect significant degradation of the data retention time below 90nm node due to a rapid increase in the junction electric field.

9.3 Prospects of semiconductor memories

Finally, I discussed the prospects of the conventional semiconductor memories

and an importance of studies for the emerging new semiconductor memories. The data retention of flash memories and DRAMs would be attributed to the existence of oxygen vacancies and the metal contamination of 0.01 [ppb], respectively. We can not remove these fundamental origins of charge loss. Therefore, we must continuously rely on the conventional methodology to modify the data retention for every successive generation. Shrink of devices makes it difficult to satisfy the specification of memories just by relying on the conventional methodologies. Therefore, new semiconductor memories of the retention-problem-free, such as FeRAMs, MRAMs and PRAMs, are crucial in the near future.

Acknowledgements

My gratitude is especially extended to Professor Tsugunori Okumura of Department of Electrical and Electronic Engineering, Tokyo Metropolitan University for his continuous suggestions and encouragements as my examiner in chief.
I would like to acknowledge Professor Takasi Ito, and Professor Michihiko Suhara of Department of Electrical and Electronic Engineering, Tokyo Metropolitan University for their fruitful discussions and suggestions as my examiner.

My great appreciation is to Mr. Katsuhiko Kubota. As my supervisor, he led and organized my work over ten years. Through working with him, I gained a good attitude and approach for research and development. He also taught me how to face problems and solve them.

I would like to acknowledge to my previous supervisor, Dr. Akio Yajima, Mr. Hirohumi Jonishi, Dr. Hotta Masao, Mr. Goichi Yokomizo, Mr. Yuzuru Ohji, Mr. Kosuke Okuyama, Dr. Shigeto Maegawa, Dr. Masahide Inuishi, Dr. Katsuhiko Shimohigashi, Dr. Tadashi Nishimura, for their valuable suggestions.

I would like to acknowledge to previous and present colleagues, Ms. Megumi Kawakami, Mr. Masahiro Sugaya, Ms. Nobue Nakajima, Mr. Mitsuo Namba, Mr. Chikashi Suzuki, Mr. Yasuyuki Okura, Mr. Fumitoshi Ito, Mr. Keiichi Maekawa, Ms. Mio Fujioka, Mr. Tsuyoshi Sendoda, Ms. Yasuko Yoshida, Mr. Takashi Takeuchi, Dr. Hiroshi Yanagida, Mr. Takao Kumihashi, Mr. Takeshi Sakai, Mr. Takumi Hasegawa, Dr. Hiroshi Uozaki, Mr. Shinobu Okanishi, Mr. Shoji Yoshida, Mr. Kazuharu Yamabe, Dr. Akio Nishida, Mr. Takaaki Tsunomura, and Mr. Yasuhiko Kawashima for good times we shared to attack the challenging problems.

I would like to acknowledge to the advisors of universities, Professor Ryogo Kubo and Professor Hiroshi Takano of Keio University, Professor Toshiyuki Hattori, Professor Eiichi Arai, and Dr. Kiyokazu Sato of Tokyo Institute of Technology,

Professor Chenming Hu, Professor Ya-Chin King, Professor Hiroshi Fujioka, Dr. Kiyotaka Imai, Ms. Ingrid Ma, Dr. Donggan Parik, Dr. Kai Chen of University of California at Berkeley, Professor Souvik Mahapatra of India Institute of Technology, for their valuable discussion and friendship during my stay in each universities.

During my careers in the company, I am involved in a variety of topics. I would like to convey thanks to all co-workers,

Professor Yoshiaki Kamigaki for study of oxidation kinetics of silicon substrate,

Professor Toru Toyabe, Professor Shigeo Ihara, Mr. Shuichi Yamamoto, Mr. Yuzuru Fujita, Dr. Kozo Katayama, Dr. Satoshi Ito, Ms. Eriko Tange, Mr. Takayuki Miyashita for modeling, simulation, and robust design,

Dr. Mitsuru Hiraki, Mr. Kazuhiro Onozawa, Mr. Takeshi Akioka, Mr. Hisaaki Kunitomo, Mr. Masao Shinozaki, Dr. Peter M. Lee for compact modeling of MOSFETs,

Dr. Shuji Ikeda, Dr. Kenichi Osada, Mr. Akira Kotabe, Dr. Eiichi Murakami, Mr. Hideki Aono, Mr. Kenichi Kuroda for design of CMOS and SRAM device,

Ms. Yuki Mori, Dr. Masahiro Moniwa, Mr. Kiyonori Ohyu, Dr. Renichi Yamada, Dr. Ken Yamaguchi, Mr. Atsushi Ogishima, Dr. Atsushi Hiraiwa, for study of DRAM refresh model,

Dr. Isao Yoshida, Ms. Yuki Kusakari, Mr. Yutaka Hoshino, Mr. Toru Fyjioka, Mr. Masatoshi Morikawa, Mr. M. Kondo, Dr. N. Sugii for design of RF and Power device,

Dr. Takashi Kobayashi, Dr. Yoshitaka Sasago, Mr. T. Arigane, Dr. H. Kurata, Dr. Yutaka Okuyama, Mr. Makoto Ogasawara, Ms. Yukiko Manabe, Mr. Hideaki Kameyama, Mr. Hitoshi Kume, Dr. Naoki Tega, Dr. Hiroshi Miki, Mr. Taro Osabe, Dr. Hideaki Kurata, Mr. Yoshihiro Ikeda, Mr. Osamu Tsuchiya for development of flash meories,

Professor Hideo Miura, Mr. Yukihiro Kumagai, Dr. Hiroyuki Ohta, Norio Ishiduka for stress analysis for silicon devices,

Dr. Yuichi Hamamura, Dr. Kazunori Nemoto, Mr. Kazuyuki Tsunokuni, Mr. Tomohiko Ouchi, Ms. Hide Yamaguchi, Ms. Yumiko Kamei for yield modeling,

Mr. Satoshi Meguro, Dr. T. Ishimaru, Dr. N. Matsuzaki, Dr. T. Mine, Dr. K. Watanabe, Mr. Makoto Mizuno, Dr. Yoshiharu Kanegae for development of MONOS devices.

I would like to acknowledge to members of the robust transistor program of MIRAI-PJ, Dr. Toru Mogami, Professor Toshiro Hiramoto, Dr. Fumiko Yano, Dr. Kiyoshi Takeuchi, Mr. Arifin Tamsir Putra for fruitful discussions.

Finally I would like to extend my deepest thanks to my wife Mikako Kamohara, our children Kazuto and Mizuki for their strong supports, deep understandings and affectionate encouragements.

Education

1988	M.S.	Tokyo Institute of Technology (Nuclear engineering department)	Tokyo, Japan
1986	B.S.	Keio University (Physics department)	Tokyo, Japan

Professional History

2003-present	Renesas Technology Corp.
1995-1996	Visiting industrial fellow of University of California at Berkeley
1988-2003	Hitachi, Ltd.

Awards

2000	Best conference paper award (Asia South Pacific Design Automation Conference 2000)
2008	Best paper award (Journal of Solid Mechanics and Materials Engineering)

List of Publications

A. Journal papers

(1) Ryogo Kubo, Takahiko Endo, S. Kamohara, Mitsuaki Shimizu, Masataka Fujii and Hiroshi Takano, “Long-Time Spin Relaxation in Zero-Field,” Journal of the Physical Society of Japan, vol. 56, pp.1172-1177, 1987.

(2) S. Kamohara, Y. Kamigaki, “Activation Energy Enhancement During Initial Silicon-Oxide Growth in Dry Oxgen,” Journal of Applied Phys., vol. 69, pp.7871-7875, 1991.

(3) Ya-Chin King, Hiroshi Fujioka, S. Kamohara, Chenming Hu., "Small-signal electron charge centroid Model for quantization of Inversion Layer in a metal-on-insulator field-effect transistor, " Applied Physics Letters, vol.72, p. 3476, 1998.

(4) M. Miyama, S. Kamohara, M. Hirakai, K. Onozawa, H. Kunitomo, “Pre-Silicon Parameter Generation Methodology Using BSIM3 for Circuit Performance-Oriented Device Optimization,” IEEE Trans. of Semiconductor Manufacturing, vol. 14, pp.134-142, 2001.

(5) M. P. Pagey, R. D. Schrimpf, K. F. Galloway, C. J. Nicklaw, S. Ikeda, S. Kamohara, “A Hydrogen-transpot-based Interface-trap-generation Model for Hot-carrier Reliability Prediction,” IEEE EDL, vol. 22, pp.290-292, 2001.

(6) S. Ikeda, Y. Yoshida, S. Kamohara, K. Imato, K. Ishibashi, K. Takahashi, “Threshold Voltage-Related Soft Error Degradation in a TFT SRAM Cell,” IEEE Trans. Of ED, vol. 50, pp.391-396, 2003.

(7) A.Kotabe, K. Osada, N. Kitai, M. Fujioka, S. Kamohara, M, Moniwa, S. Morita, Y. Satoh, “A Low-Power Four-Transistor SRAM Cell With a Stacked Vertical Poly-Silicon PMOS and Dual-Word-Voltage Scheme,” IEEE Journal of Solid-State Circuit, vol. 40, pp.870-876, 2005.

(8) P. B Kumar, P. R.Nair, R. Sharma, S. Kamohara, S. Mahapatra, “ Lateral profiling of trapped charge in SONOS flash EEPROMs programmed using CHE injection, “IEEE Trans. of ED, vol. 53, pp.698-705, 2006.

(9) Y. Mori, S. Kamohara, M. Moniwa, K. Ohyu, T. Yamanaka, R. Yamada, “Direct observation of worst-bit leakage currents of DRAM, “ IEEE Trans. of ED, vol. 53, pp.398-400, 2006.

(10) M. Kondo, N. Sugii, Y. Hoshino, W. Hirasawa, Y. Kimura, M. Miyamoto, S. Kamohara , Y. Kondo, S. Kimura, I. Yoshida, “Thick-Strained-Si/Relaxed-SiGe Structure of High-Performance RF Portable Cellular Handsets,” IEEE Trans. of ED, vol. 53, pp.3136-3145, 2006.

(11) H. Kurata, K. Otsuga, A. Kotabe, S. Kajiyama, T. Osabe, Y. Sasago, S. Kamohara, P. Tuchiya, “Random Telegraph Signal in Flash Memory: Its Impact on Scaling of Memory Beyond the 90-nm Node,” IEEE J. of Solid-State Circuits, vol. 42, pp. 1362-1369, 2007.

(12) Y. Kumagai, H. Ohta, H. Miura, A Simizu, S. Kamohara, K. Maekawa, "Development of Evaluation Method for Estimation Stress-Induced Change in Drain Current in Deep-sub-micron MOSFETs", J of Solid Mechanics and Materials Eng., vol. 1, pp. 93-101, 2007.

(13) S. Kamohara, T. Okumura, “New Physical Model to Explain Logarithmic Time Dependence of Data Retention,” to be accepted by Appl. Surf. Sci., 2008.

(14) S. Kamohara, C. Hu, T. Okumura, “Deep Trap Stress Induce Leakage Current Model For Nominal and Weak Oxides,” to be submitted to JJAP.

(15) S. Kamohara, K. Kubota, M. Moniwa, K. Ohyu, A. Ogishima, T. Okumura, “Stastical PN Junction Leakage Model with Trap Level Fluctuation for Tref(Refresh Time)-Oriented DRAM Design,” to be submitted to JJAP.

B. International conference papers

(1) S. Kamohara, T. Kobayashi, “Theoretical Model to Explain Dopant Segregation to the Grain Boundaries and Grain Growth,” Proc. Of VLSI process/device Modeling Workshop, pp.52-53, 1990.

(2) S. Kamohara, Y. Kamigaki, “Discovery of Activation Energy Enhancement during Initial Silicon-Oxide Growth and Its Physical Modeling for Accurate Process Simulation”, SSDM90, pp.1087-1090 ,1990.

(3) K. Katayama, S. Kamohara, S. Itoh, “ An efficient quantum Monte Carlo simulation method-path integral approach for dissipative transport in quantum devices,” Proc. of IEDM90, pp. 775-778, 1990.

(4) S. Kamohara, T. Kobayashi, M. Sugaya, S. Yamamoto, “New Model for the Simulation of Polysilicon Impurity Diffusion Sources for a Wide Range of Process Condition,” Proc. of BCTM92, pp.126-129, 1992.

(5) S. Kamohara, M. Sugaya, H. Matsuo, “A Very Fast Three-Dimensional Impurity Profile Simulation Incorporating an Accumulated Diffusion Length and Its Application to the Design of Power MOSFETs,” Proc. of SISPAD95, pp.194-197, 1995.

(6) S. Kamohara, M. Kawakami, “A New Hydrodynamic Equation for Ion-Implantation Simulation,” Proc. of SISPAD95, pp.180-483, 1995.

(7) M. Kawakami, M. Sugaya, S. Kamohara, “ A new high-speed non-equilibrium point defect model for annealing simulation,” Proc. of SISPAD96, pp. 93-94, 1996.

(8) S. Kamohara, A. Shimizu, S. Yamamoto, K. Kubota, “ TED Model Including the Dissolution of Extended Defects,” Proc. of SISPAD96, pp.85-86, 1996.

(9) Y. -C. King, H. Fujioka, S. Kamohara, W. -C. Lee, C. Hu, “AC Charge Centroid Model For Quantization Of Inversion Layer In NMOS,” Proc. of VLS Symp. on VLSI Technology, pp. 245-249, 1997.

(10) S. Kamohara, K. Imai, I. Ma, Y. Cheng, C. Hu, “Studies of Coulomb Blockade,

Single-Trangister Dynamic Memory, and Oxide Trap Charge State Based on the Random Telegraph Noise Modeling," Proc. of ICVC97, pp.295-297, 1997.

(11) S. Kamohara, Y. C. King, K. Chen, D. Park, C. Hu, " MOSFET Carrier Mobility Model Based on the Density-of-state at the DC-Centroid in the Quantized Inversion Layer," Proc. of ICVC97, pp.171-173, 1997.

(12) Y. Okuyama, S. Kamohara, Y. Manabe, K. Okuyama, K. Kubota, " Monte Carlo simulation of stress-induced leakage current by hopping multi-traps in oxide, " Proc. of IEDM98, pp. 905-908, 1998.

(13) K. Yamaguchi, M. Ogasawara, S. Kamohara, T. Teshima, " Two- and three-dimensional simulations of anomalous current in small junctions integrated with an ultra-high density DRAM," Proc. of IWCE-6, pp. 182-185, 1998.

(14) S. Kamohara, D. Park, C. Hu, "Deep Trap SILC (Stress Induce Leakage Current) Model For Nominal and Weak Oxides," Proc. of IRPS98, pp.57-61, 1998.

(15) S. Kamohara, K. Kubota, M. Moniwa, K. Ohyu, A. Ogishima, "Stastical PN Junction Leakage Model with Trap Level Fluctuation for Tref(Refresh Time)-Oriented DRAM Design,", Proc. Of IEDM 99, pp. 539-542, 1999.

(16) S. Kamohara, Yutaka Okuyama, Yukiko Manabe, Kosuke Okuyama, Katsuhiko Kubota, Donggun Park, Chenming Hu, "Quantitative analysis of SILCs (stress-induced leakage currents) based on the inelastic trap-assisted tunneling model," Proc. SPIE Vol. 3881, pp. 206-214, 1999.

(17) M. Miyama, S. Kamohara, M. Hiraki, K. Onozawa, H. Kunitomo, " Pre-silicon parameter generation methodology using BSIM3 for device design, " pp. 359-363, Proc. of CICC99, pp. 359-363, 1999.

(18) S. Ikeda, Y. Yoshida, K. Shoji, K. Kuroda, K. Komori, N. Suzuki, K. Okuyama, S. Kamohara, N. Ishitsuka, E. Murakami, T. Yamanaka, " A highly manufacturable 0.18 μm generation logic technology," Proc. of IEDM99, pp. 675-678, 1999.

(19) Y. Hoshino, M. Morikawa, S. Kamohara, M. Kawakami, T. Fujioka, S. Ikeda, I.

Yoshida, S. Shimizu, “ High performance scaled down Si LDMOSFET with thin gate bird's beak for RF power amplifiers,” Proc. of IEDM99, pp. 205-208, 1999.

(20) M. Miyama, S. Kamohara, K. Nakura, M. Shinozaki, T. Akioka, K. Okuyama, “ Statistical BSIM3 model parameter extraction and fast/slow model parameter determination for high speed SRAM parametric yield estimation, “ Proc. of IWSM2000, pp. 42-45, 2000.

(21) H. Kameyama, Y. Okuyama, S. Kamohara, K. Kubota, H. Kume, K. Okuyama, “ A new data retention mechanism after endurance stress on flash memory,”, Proc of IRPS2000, pp. 194-199, 2000.

(22) M. Miyama, S. Kamohara, “ Circuit performance oriented device optimization using BSIM3 pre-silicon parameters, “ Proc. of ASP-DAC 2000, pp. 371-374, 2000.

(23) S. Kamohara, T. Ouchi, K. Okuyama, Y. Kawashima, Y. Ohji, K. Kubota, " A new extraction method of device parameters for mass production, "Proc. of IWSM 2001, pp. 43 - 46, 2001.

(24) S. Kamohara, M. Kawakami, E. Kosaka, F. Ito, K. Okuyama, Y. Ohji, "Robust Design of 0.18um ASIC MOSFETs Using Taguchi Method with TCAD,” Proc of IWSM 2001, pp.21-24, 2001.

(25) M. Miyama, S. Kamohara, K. Okuyama, Y. Oji, “ Parametric yield enhancement system via circuit level device optimization with statistical circuit simulation, “ Proc. of VLSI symp. on VLSI Circuits, pp. 163-166, 2001.

(26) Y. Mori, R. Yamada, S. Kamohara, M. Moniwa, K. Ohyu, T. Yamanaka, “ A new method for predicting distribution of DRAM retention time,” Proc. of IRPS, pp. 7-11, 2001.

(27) Y. Sasago, T. Arigane, H. Kurata, S. Saeki, Y. Goto, S. Kamohara, T. Kobayashi, “ 10-MB/s multi-level programming of Gb-scale flash memory enabled 90nm technology, “ IEDM02, pp.952-954, 2002.

(28) Y. Hamamura, K. Nemoto, T. Kumazawa, H. Iwata, K. Okuyama, S. Kamohara,”

Repair yield simulation with iterative critical area analysis for different technology," Proc of DFT2002, pp.305-313, 2002.

(29) P. R. Nair, P. B. Kumar, R. Sharma, S. Mahapatra, S. Kamohara, E. Murakami, " A comprehensive trapped charge profiling technique for SONOS flash EEPROMs," Proc. of IEDM04, pp. 403-406, 2004.

(30) T. Ishimaru, N. Matsuzaki, Y. Okuyama, T. Mine, K. Watanabe, Y. Yugami, Y. Kawashima, T. Sakai, Y. Kanamaru, Y. Ishii, M. Mizuno, S. Kamohara, K. Okuyama, K. Kuroda, K. Kubota, " Impact of SiON on embedded nonvolatile MNOS memory" Proc. of IEDM04, pp.885-888, 2004.

(31) K. Osada, N. Kitai, S. Kamohara, T. Kawahara, " Analysis of SRAM neutron-induced errors based on the consideration of collection and parasitic-bipolar failure modes," CICC04, pp.357-360, 2004

(32) A. Kotabe, K. Osada, N. Kitai, M. Fujioka, S. Kamohara, M. Moniwa, " A 0.13-/spl mu/m, 0.78-/spl mu/m/sup 2/ low-power four-transistor SRAM of vertically stacked poly-silicon MOS and a dual-word-voltage scheme," Proc. of VLSI symp. on VLSI Circuits, pp. 60-63, 2004.

(33) M. Kondo, N. Sugii, Y. Hoshino, W. Hiraiwa, Y. Kimura, M. Miyamoto, S. Kamohara, Y. Kondo, S. Kimura, I. Yoshida, " High performance RF power LDMOSFETs for cellular handsets formed on Si/relaxed-SiGe structure," Proc. of IEDM05, pp. 365-368, 2005.

(34) P. B. Kumar, R. Sharma, P. R. Nair, D. R. Nair, S. Kamohara, S. Mahapatra, " Mechanism of drain disturb in SONOS flash EEPROMs, Proc. of IRPS, pp.186-190, 2005.

(35) K. Sridhar, P. B. Kumar, S. Mahapatra, E. Murakami, S. Kamohara, " Controlling injected electron and hole profiles for better reliability of SONOS flash EEPROMs," Proc. of IPFA2005, pp. 190-194, 2005.

(36) N. Tega, H. Miki, T. Osabe, A. Kotabe, K. Otsuga, H. Kurata, S. Kamohara, Y. Ikeda, R. Yamamda, " Anomalously Large Threshold Voltage Fluctuation by Complex

Rand in Floating Gate Flash Memory,” Proc. of IEDM06, pp. 1-4, 2006.

(37) H. Kurata, K. Otsuga, A. Kotabe, S. Kajiyama, T. Osabe, Y. Sasago, S. Kamohara, O. Tsuchiya, “The Impact of Random Telegraph Signals on the Scaling of Multilevel Flash Memory,” Proc. of VLSI Symp. on VLSI Circuits, pp. 112-113, 2006.

(38) P. K. Bharath, E. Murakami, S. Kamohara, S. Mahapatra, “Endurance and Retention Characteristics of SONOS EEPROMs Operation Induced Hot Hole Erase,” Proc of IRPS, pp.699-700, 2006.

(39) H. Miki, T. Osabe, N. Tega, A. Kotabe, H. Kurata, K. Tokami, Y. Beceda, S. Kamohara, R. Yamada, ”Quantitative Analysis of Random Telegraph Signals as Fluctuations Voltages in Scaled Flash Memory Cells,” Proc. of IRPS, pp. 29-35, 2007.

(40) S. Kamohara, T. Okumura, “New Physical Model to Explain Logarithmic Time Dependence of Data Retention in Flash EEPROM,” Proc. of ISCSI-V, pp. 57-58, 2007.

(41) K. Takeuchi, T. Fukai, T. Tsunomura, A. T. Putra, A. Nishida, S. Kamohara, T. Hiramoto, "Understanding random threshold voltage fluctuation by comparing multiple Fabs and technologies," Tech. Dig. of IEDM 2007, pp. 467-470, 2007.

(42) T. Tsunomura, A. Nishida, F. Yano, A. T. Putra, K. Takeuchi, S. Inaba, S. Kamohara, K.Terada, T. Hiramoto, T. Mogami, "Analysis of 5σ V_{th} fluctuation in 65nm MOSFETs using Takeuchi plot," to be accepted by VLSI 2008.

C. Domestic conference and journal papers

(1) 鈴木寛光, 服部俊幸, 木下仁史 , 蒲原史朗,"IHQ 線形加速器の開発･研究(3)," 応用物理学会予稿集, vol. 1986, p.69, 1986.

(2) 蒲原史朗, 佐藤潔和, 服部俊幸, 新井栄一, "表面電荷法によるドリフトチューブ静電容量の計算," 応用物理学会予稿集, Vol.1986, p. 69, 1986.

(3) 菅井勲, 石井三郎, 服部俊幸, 蒲原史朗, 武藤英, 高橋陽介, 加藤一, 山崎邦明, "長寿命高強度ハイブリッド炭素ストリッパー膜の開発 I," 応用物理学会予稿集, Vol.1987, p. 43, 1987.

(4) 蒲原史朗, 佐藤潔和, 服部俊幸, 新井栄一, " 等価回路計算による IH-型線形加速器の研究," 応用物理学会予稿集, Vol.1987, p. 43, 1987.

(5) 蒲原史朗, 菅谷正弘, 松尾仁司, 井原茂男, "蓄積拡散長を用いた高速三次元不純物分布シミュレーション,"電子情報通信学会技術研究報告.ED, Vol.94, pp. 35-42, 1994.

(6) 河上恵, 菅谷正弘, 蒲原史朗, "Hydrodynamic 方程式を用いた高速イオン打ち込みシミュレーション方法の提案," 電子情報通信学会技術研究報告.VLD, Vol.95, pp. 39-46, 1995.

(7) 宮下孝幸, 河上恵, 菅谷正弘, 蒲原史朗, "プロセス･デバイス統合シミュレーションシステム : VFL の開発とデバイス設計への適用,"電子情報通信学会総合大会講演論文集, Vol.1996 年_基礎・境界, p. 447, 1996.

(8) 藤岡徹, 蒲原史朗, 森川正敏, 吉田功, " BSIM を用いた RF-LDMOS モデル," 電子情報通信学会総合大会講演論文集, Vol.2000 年.エレクトロニクス, p. 91, 2000.

(9) 蒲原史朗, 久保田勝彦, 茂庭昌弘, 大湯静憲, 荻島淳史, "DRAM リフレッシュエンジニアリングモデル,"電子情報通信学会技術研究報告.SDM, Vol.100, pp. 41-47, 2000.

(10) 小田部晃, 長田健一, 北井直樹, 藤岡美緒, 蒲原史朗, 茂庭昌弘, 森田貞幸,

斉藤良和, "従来型メモリセルの 1/3 の面積を実現する縦型 MOS を用いた 4 トランジスタ SRAM セル(VLSI 回路,デバイス技術(高速,低電圧,低電力)), 電子情報通信学会技術研究報告.SDM, Vol.104, pp. 7-11, 2004.

(11) 長田健一, 北井直樹, 蒲原史朗, 河原尊之, "電荷収集と寄生バイポーラ効果を考慮した SRAM の中性子ソフトエラー解析(新メモリ技術, メモリ応用技術, 一般)," 電子情報通信学会技術研究報告.ICD, Vol.105, pp. 31-36, 2005.

(12) 倉田英明, 大津賀一雄, 小田部晃, 梶山新也, 長部太郎, 笹子佳孝, 鳴海俊一, 冨上健司, 蒲原史朗, 土屋修, " 多値フラッシュメモリのスケーリングにおけるランダム・テレグラフ・シグナルの影響(VLSI 回路,デバイス技術(高速,低電圧,低消費電力))," 電子情報通信学会技術研究報告.SDM, Vol.106, pp. 161-166, 2006.

(13) 熊谷幸博, 太田裕之, 三浦英生, 清水昭博, 蒲原史朗, 前川径一, " ディープサブミクロン MOSFET の応力起因ドレイン電流変動評価手法の開発,"日本機械学會論文集, Vol.72, pp. 47-54, 2006.

(14) 井原茂男, 蒲原史朗, 鳥谷部達, 横溝剛一, ”電気系ＣＡＤへの応用,” 電子情報通信学会, Vol. 75, pp. 148-154, 1992.

D. Patents

(1) 蒲原史朗, 上西博文, "半導体の酸化制御方法, "特開平 3-274729, 1991.

(2) 蒲原史朗, 伊藤智, "多結晶シリコン膜の熱処理工程制御方法," 特開平 4-92442, 1992.

(3) 山本秀一, 蒲原史朗, "三次元素子構造の生成方法,"特開平 4-369760, 1992.

(4) 田中順子, 鳥谷部達, 松尾仁司, 井原茂男, 土井一平, 蒲原史朗," 半導体装置, 特開平 5-291244, 1993.

(5) 鳥谷部達, 田中順子, 松尾仁司, 蒲原史朗, 井原茂男, 土井一平, "半導体集積回路装置, 特開平 5-291392, 1993.

(6) 松尾仁司, 鳥谷部達, 田中順子, 蒲原史朗, 井原茂男, 土井一平, "半導体装置およびその製造方法, "特開平 5-326694, 1993.

(7) 蒲原史朗, 松尾仁司, 菅谷正弘, "プロセスシミュレータ," 特開平 6-97020, 1994.

(8) 高橋誠, 鳥谷部達, 蒲原史朗, 松尾仁司, 大倉康幸, 伊藤智、田中順子, 丸山貴志子, "絶縁ゲート型バイポーラトランジスタ,"特開平 6-268227, 1994.

(9) 奥山裕, 蒲原史朗, 松尾仁司, 井原茂男, "不揮発性半導体記憶装置,"特開平 7-106448, 1995.

(10) 蒲原史朗, 松尾仁司, 菅谷正弘, 河上恵, "高速三次元不純物分布計算方法," 特開平 7-147254, 1995.

(11) 船場誠司, 横溝剛一, 蒲原史朗, "回路設計支援方法, "特開平 8-161363, 1996.

(12) 菅谷正弘, 蒲原史朗, 新美敏男, 横溝剛一, 河上恵, "プロセスシミュレーション装置,"特開平 8-194730, 1996.

(13) Ｐ・リー, 横溝剛一, 新美敏男, 蒲原史朗, "半導体装置設計方法,"特開平

8-263540, 1996.

(14) 蒲原史朗, 田中聡, 新美敏男, 横溝剛一, "高周波増幅回路,"WO96/027905, 1998.

(15) 蒲原史朗, 坂本光造, 目黒怜, 藤田譲, 飯島哲郎, 矢ノ倉栄二, "縦型電界効果トランジスタの製造方法, "WO97/011497, 1998.

(16) 蒲原史朗, 小川哲生, 藤田譲, 中嶋伸恵, 小林正義, 飯島哲郎, "縦型電界効果トランジスタとその製造方法,"特開平 11-177086, 1999.

(17) 見山美可子, 蒲原史朗, "モデルパラメータ抽出方法及び装置, "特開 2000-322456, 2000.

(18) 森川正敏, 星野裕, 蒲原史朗, 吉田功, 池田修二, 河上恵, 三宅智之, "半導体装置, 特開 2001-094094, 2001.

(19) 見山美可子, 蒲原史朗, 那倉康一, "統計モデルパラメータ抽出方法および装置,"特開 2001-291778, 2001.

(20) 大竹成典, 横溝剛一, 蒲原史朗, "半導体集積回路および半導体集積回路の設計方法, "特開 2002-009244, 2002.

(21) 菅谷正弘, 蒲原史朗, "半導体装置の設計方法,"特開 2002-009260, 2002.

(22) 星野裕, 池田修二, 蒲原史朗, 森川正敏, 菅谷正弘 , "半導体装置及び半導体装置の製造方法," 特開 2002-076337, 2002.

(23) 河上恵, 蒲原史朗, 武田康裕, 小坂恵理子, 山中俊明, "シミュレータのキャリブレーション方法,"特開 2002-134373, 2002.

(24) 松浦達治, 久保征治, 堀田正生, 蒲原史朗, 中嶋伸恵, "半導体集積回路装置およびその製造方法, "特開 2002-151599, 2002.

(25) 朝山匡一郎, 三井裕, 荒川史子, 蒲原史朗, 大路譲, "半導体装置およびその測定方法、ならびに半導体装置の製造方法,"特開 2002-217258, 2002.

(26) 見山美可子, 蒲原史朗, 大路譲, 奥山幸祐, 河上恵, "半導体集積回路の製造方法," 特開 2002-353083, 2002.

(27) 見山美可子, 蒲原史朗, 大路譲, 奥山幸祐, 河上恵, "情報処理システム及び情報処理方法,"特開 2002-353084, 2002.

(28) 河上恵, 蒲原史朗, 仙洞田剛士, 見山美可子, "半導体装置の製造方法,"特開 2002-368056, 2002.

(29) 山中俊明, 西田彰男, 吉田安子, 池田修二, 黒田謙一、蒲原史朗, 木村紳一郎, 村上英一, 松岡秀行, 南正隆, "半導体集積回路装置の製造方法," 特開 2002-368126, 2002.

(30) Ｐ・リー, 横溝剛一, 新美敏男, 蒲原史朗, "半導体装置の信頼性設計支援システム,"特開 2003-224258, 2003.

(31) 坂井健志, 志波和佳, 蒲原史朗, 中嶋伸恵, 黒田謙一, "半導体装置の製造方法," 特開 2003-031680, 2003.

(32) 桧上竜也, 伊藤文俊, 蒲原史朗, "半導体集積回路装置およびその製造方法," 特開 2003-152101, 2003.

(33) 森川正敏, 星野裕, 蒲原史朗, 吉田功, 池田修二, 河上恵, 三宅智之, "半導体装置およびその製造方法, 特開 2004-096118, 2004.

(34) 森川正敏, 星野裕, 蒲原史朗, 吉田功, 池田修二, 河上恵, 三宅智之, "半導体装置およびその製造方法, 特開 2004-096119, 2004.

(35) 藤岡美緒, 蒲原史朗, 飯田哲也, 藤岡徹, 黒谷欣吾, "半導体装置,"特開 2004-063922, 2004.

(36) 小笠原誠, 蒲原史朗, "半導体集積回路装置の製造方法,"特開 2004-071593, 2004.

(37) 藤岡美緒, 蒲原史朗, "半導体装置およびその製造方法,"特開 2004-103637,

2004.

(38) 伊藤文俊, 蒲原史朗, 中嶋伸恵, "半導体装置およびその製造方法,"特開 2004-111749, 2004.

(39) 蒲原美可子, 蒲原史朗, "回路の設計方法および回路設計支援システム,"特開 2004-145410, 2004.
]
(40) 桧上竜也, 伊藤文俊, 蒲原史朗, "半導体集積回路装置,"特開 2005-333164, 2005.

(41) 中嶋伸恵, 蒲原史朗, "半導体装置およびその製造方法,"特開 2005-064178, 2005.

(42) 小田部晃, 長田健一, 茂庭昌弘, 蒲原史朗, "半導体記憶装置,"特開 2005-078741, 2005.

(43) 前川径一, 蒲原史朗, "半導体装置およびその製造方法,"特開 2005-116582, 2005.

(44) 前川径一, 南眞一, 渡部浩三, 蒲原史朗, 吉田省史, "半導体装置及びその製造方法, "特開 2006-165451, 2006.