行政院國家科學委員會專題研究計畫　成果報告

雙位元儲存氮化矽快閃式記憶元件技術及可靠性(3/3)

計畫類別：個別型計畫
計畫編號：NSC94-2215-E-009-004-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：國立交通大學電子工程學系及電子研究所

計畫主持人：汪大暉

報告類型：完整報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 95 年 10 月 14 日

雙位元儲存氮化矽快閃式記憶元件技術及可靠性

摘要

本篇計畫主要著重在研究以氮化矽(SiN)為電荷儲存之快閃式記憶元件可靠性議題(reliability issue)研究。一般而言，為了增加電荷保存能力(retentivity)，此元件通常選用較厚的底部氧化層(bottom oxide)。元件在未加壓之前，其具有相當優異的可靠度。然而，當經過多次寫入/抹除(program/erase)後，會使得氧化層造成傷害，進而對元件可靠性造成極大的影響。

第一章簡介此元件基本的結構以及寫入抹除的方式。對於二位元操作時反向讀取(reverse reading)的原理，也詳列於其中。第二章探討元件耐久性(endurance)機制。吾人發現，無論在寫入或抹除狀態之臨界電壓準位，皆會隨著寫入/抹除次數的增加而皆有向上揚升的現象。

第三章詳述一改良後之電荷幫浦(charge pumping)量測方法。藉由此法，在汲極/源極接面上之寫入電荷水平分佈，可獨立被萃取出來。吾人研究發現，在同一元件中，第二寫入的位元有著比第一寫入位元較寬的電荷分佈。此原因為在寫入第二位元時，第一寫入位元產生的電場會加速通道電子使其提早注入氮化矽層中。另外，實驗結果顯示，寫入電荷分佈會隨著寫入抹除次數增加而延伸到通道中央。

第四章中，吾人對抹除狀態臨界電壓不穩定(threshold voltage instability)、讀取擾動(read-disturb)，以及寫入狀態資料流失(charge loss)有著深入的探討。首先，對於一經加壓後的元件，抹除狀態之臨界電壓會隨著儲存時間而上升。此漂移現象與溫度有著微弱的關係並且隨寫入抹除次數呈現奇特的轉彎現象(turn-around)，這與底部氧化層中帶正電性缺陷的生成有著密切關係。吾人實驗結果發現，此臨界電壓漂移與時間呈現對數（logarithm）的相依性，並可用穿隧波前(tunneling front)模型來做描述。此外，若是讀取偏壓太大，正電性電荷幫助穿隧(PCAT)效應將會主導臨界電壓漂移，與時間將會轉變成指數(power-law)關係。藉由研究垂直電場以及溫度對氮化矽層電荷遺失之影響，吾人提出一解析之物理模型：Frenkel-Poole 蒸散進而透過氧化層缺陷穿隧。吾人可利用此模型提出一開極偏壓加速測試元件資料保存時間之方法。

第五章探討底部氧化層厚度以及加壓效應對於氮化矽記憶元件資料流失的影響。根據多電子捕捉（multiple electron trapping）模型，吾人利用一數值分析方法，分別對底部氧化層厚度為 1.8nm 到 5.0nm 的元件模擬其電荷保存特性。在吾人的模型中，假設氮化矽中缺陷為連續性的分佈。傳導帶(conduction band)與氮化矽中缺陷狀態(trap

state)的暫態行為，可用一連串 Frenkel-Poole 激發以及電子再度被捕陷(re-capture)來描述。電荷流失可分為兩種途徑：一是傳導帶電子透過正電性氧化層缺陷而流失；二是電子直接穿隧過底部氧化層而散失。透過大面積元件量測，吾人發現在較厚的底部氧化層元件中，經由加壓後引致的電荷逸失現象呈現兩階段(two stages)發展。第一階段電荷漏電流被氧化層缺陷幫助穿隧給限制住。第二階段中，因 Frenkel-Poole 散失使其而遵循 1/t 的時間關係。從第一階段到第二階段之間的過渡時間與氧化層缺陷幫助穿隧時間有關，但是將會被延長一個特定比例。基於以上的了解，吾人在第六章中，利用此 1/t 暫態電流，萃取出氮化矽材料中缺陷的密度。

第七章中，吾人針對一局部儲存(localized storage)、多準位(multi-level)氮化矽快閃式記憶元件中，因寫入/抹除加壓後產生隨機電報雜訊(random telegraph noise, RTN)導致之讀取電流擾動作一深入之探討。吾人發現，局部儲存方式明顯地增加 RTN 的擾動。而 RTN 的振幅隨著不同的寫入準位而改變。用一機率模型，可定義此 RTN 造成的讀取電流擾動分佈。例外，利用較好的底部氧化層製程方法，可有效降低此讀取電流雜訊。

最後於第八章，吾人將對本計畫做個總結。

關鍵詞：氮化矽快閃式記憶元件、寫入/抹除、耐久性、電荷幫浦、寫入電荷水平分佈、抹除狀態臨界電壓不穩定、讀取擾動、寫入狀態資料流失、穿隧波前、正電性電荷幫助穿隧、Frenkel-Poole 蒸散、多電子捕捉模型、隨機電報雜訊

Dual-bit Storage Nitride Trap Flash Memory Device Technology and Reliability

Abstract

This project will focus on the reliability issues of SONOS-type trapping storage flash memories. For today's SONOS cells, a thicker bottom oxide is employed to improve the retentivity. These cells exhibit excellent data retention behavior before stress. After P/E cycling, the bottom oxide is damaged, thereby degrading the reliability.

In Chapter 1, the device structure and program/erase methods of the cell are described. A reverse read scheme for two-bit operation is illustrated. With respect to the cell endurance, the threshold voltage in program-state or in erase-state may shift upward as P/E cycle number increases. The mechanism will be investigated in Chapter 2.

To expound the second-bit effect, a modified charge pumping technique to characterize programmed charge lateral distribution is proposed in Chapter 3. The stored charge distribution of each bit over the source/drain junctions can be profiled separately. Our result shows that the secondly programmed bit has a broader stored charge distribution than the first programmed bit. The reason is that a large channel field exists under the first programmed bit during the second bit programming. Such a large field accelerates channel electrons and causes earlier electron injection into the

nitride. In addition, we find that programmed charges spread further into the channel as program/erase cycle number increases.

Reliability issues including erase-Vt state threshold voltage instability, read-disturb, and high-Vt state charge loss will be addressed in Chapter 4. First, an erase-state threshold drift with storage time is observed in a P/E cycled cell. This drift has insignificant temperature dependence and exhibits an anomalous turn-around with P/E cycle number. This peculiar phenomenon is strongly related to the creation of positive charged defects in the bottom oxide. The temporal evolution of the threshold voltage drift has log(t) dependence on storage time and can be well described by the tunneling front model. Furthermore, at a sufficiently large read bias, positive charge assisted channel electron tunneling dominates the threshold voltage shift, causing a power-law time relation. By measuring the dependence of electric field and temperature, an analytical model based on Frenkel-Poole emission followed by oxide trap assisted tunneling successfully identifies the mechanism for charge loss. With use this model, a V_g acceleration method for retention lifetime test is also proposed.

Bottom oxide thickness and program/erase stress effects on charge retention in SONOS flash memory cells with FN programming are investigated. Utilizing a numerical analysis based on a multiple electron trapping model, the electron retention behavior in a SONOS cell with bottom oxide thickness from 1.8nm to 5.0nm is simulated. In our model, the nitride traps have a continuous energy distribution. A series of Frenkel-Poole excitation of trapped electrons to the conduction band and electron re-capture into nitride traps feature the transitions between the conduction band and trap states. Conduction band electron tunneling via positively charged

oxide traps created by high-voltage stress and trapped electron direct tunneling through the bottom oxide is included to describe various charge leakage paths. We measure the nitride charge leakage current directly in a large area device for comparison. Our study reveals that the charge retention loss in a high-voltage stressed cell with a thicker bottom oxide (5nm) exhibits two stages. The charge leakage current is limited by oxide trap assisted tunneling in the first stage and then follows a 1/t time dependence due to the Frenkel-Poole emission in the second stage. The transition time from the first stage to the second stage is related to oxide trap assisted tunneling time, but is prolonged by a factor. According to the above understanding, the silicon nitride trap density can be extracted from the 1/t transient current in Chapter 6.

In Chapter 7, program/erase stress induced read current fluctuation arising from random telegraph noise (RTN) in a localized, multi-level SONOS cells is explored. Our study shows that localized charge storage significantly enhances RTN. The amplitude of RTN varies in different program levels of a multi-level cell. The broadening of read current distribution caused by RTN is characterized and modeled. Better bottom oxide process can reduce read current noise.

Conclusions are finally made in Chapter 8.

Keywords: SONOS-type trapping storage flash memories, program/erase, endurance, charge pumping, programmed charge lateral distribution, erase state threshold voltage instability, read disturb, high-Vt state charge loss, read-disturb, tunneling front model, positive charge assisted tunneling (PCAT), Frenkel-Poole

emission, multiple trapping model, random telegraph noise (RTN).

Contents

Chinese Abstract I

English Abstract III

Contents VII

Chapter 1 Introduction **1**

1.1 Backgrounds 1

1.2 Uniform charge storage versus localized charge storage 2

1.3 Organization of the Project 5

Chapter 2 Program/Erase Cycling Induced Endurance Degradation in Nbit Flash Memory Cell **12**

2.1 Introduction 12

2.2 Endurance Failure in Nbit Cells 12

2.3 Evidence of Negative Charge Creation 13

2.4 Improvement of Cycling Endurance 15

Chapter 3 Characterization of Programmed Charge Lateral Distribution in Nbit Flash Memory Cell by Using Charge Pumping Technique **23**

3.1 Introduction 23

3.2 Device Structure and Measurement Setup 24

3.3 Measurement Result and Discussion 25

3.3.1 Single-bit Storage 25

3.3.2 Two-bit Storage 26
3.3.3 P/E Cycling Stress Effect 28
3.4 Summary 29
Chapter 4 Reliability Mechanisms of Data Retention and Read-Disturb in Nbit Flash Memory Cells 44
4.1 Introduction 44
4.2 Room-Temperature Threshold Voltage Drift 44
4.3 Read-Disturb Effects in Erase State 46
4.3.1 Commonality between Vt Drift and Read-Disturb 47
4.3.2 Gate and Drain Bias Dependences on Read-Disturb Behavior 47
4.4 Program state Charge Loss 48
4.5 Summary 49
Chapter 5 Numerical Simulation for Program-State Charge Retention in SONOS Flash Memory Cell 61
5.1 Introduction 61
5.2 Retention Loss Simulation Model 62
5.2.1 Nitride Charge Dynamics and Loss Mechanisms 62
5.2.2 Assumptions in Simulation 66
5.3 Numerical Simulation Method 67
5.3.1 Explicit and Implicit Methods 67
5.3.2 Non-linear and Stiff System 68
5.3.3 Jacobian Matrix 69
5.3.4 Simulation Flowchart 71

5.4 Results and Discussion 72
5.4.1 Bottom Oxide Blocking Effect 72
5.4.2 Post-Stress Two-Stage Retention Loss 73
5.4.3 Leakage Component Separation 75
5.5 Summary 76
Chapter 6 Extraction of Nitride Trap Density from Stress Induced Gate Leakage Current in SONOS Flash Memory 94
6.1 Introduction 94
6.2 Nitride Trap Characterization Technique 95
6.3 Summary 99
Chapter 7 Read Current Instability Arising from Random Telegraph Noise in Localized Storage, Multi-Level SONOS Flash Memory 105
7.1 Introduction 105
7.2 Experimental Setup 106
7.3 Two-region model 106
7.4 Results and Discussions 108
7.4.1 Localized charge storage 108
7.4.2 RTN in a multi-level cell 108
7.4.3 P/E stress and read current distribution modeling 109
7.5 Summary 110

Chapter 8 Conclusions **121**

References **124**

Appendix **137**

Chapter 1

Introduction

1.1 Backgrounds

Flash memory has become a kind of nonvolatile memory widely used in personal computers (PCs) and electronic products, as it is capable of storing, reading and erasing data for many times, and the data stored therein will be retained even after the supplied power is cut off. Data in flash memory devices are usually stored in a multi-layer gate structure in a field effect transistor. With respect to charge storage devices, two state-of-the-art techniques are categorized. (a) Floating Gate Devices: Charge is stored in a thin conducting or semiconductor layer or conducting particles sandwiched between insulators [1.1]-[1.4]. Kahng and Sze proposed the first floating gate device in 1967 [1.5]. Storage of the charge in the floating gate allows the threshold voltage (Vt) to be electrically altered between a low and a high value to represent logic 1 and 0, respectively. (b) Charge Trapping Devices: Charge is stored in the traps at the interface of a multi-layer gate structure and/or in the bulk of insulator. The major difference in charge loss mechanisms between charge trapping and FG flash is that the charge storage media is non-conductive in charge trapping devices and is conducting in FG flash. In a charge trapping cell, only trapped charges near an oxide trap can leak directly, as demonstrated in Fig. 1.1(a). Trapped charges elsewhere cannot escape via oxide trap assisted tunneling directly since trapped charge cannot move. Those trapped electrons need to be excited to its conduction band and then move to the oxide trap and escape. In FG cells, since the storage media is conductive, stored electrons can move freely to the weak spot of the oxide and

escape by trap-assisted tunneling (Fig. 1.1(b)). This causes a limitation in the bottom oxide thickness and thus the channel length and the operating voltage cannot be scaled down further in FG flash. According to the device simulation, the channel length of NOR type floating gate flash memory cannot be scaled below 0.13μm, as shown in Fig. 1.2 [1.6]. In order to solve the dilemma, charge trapping device becomes an attractive solution. Ref. [1.7]-[1.10] indicated various candidates for trapping material. Among them, nitride-based material has been considered to be the most promising candidate for successfully integrated into standard logic process.

Fig. 1.3 illustrates the development of the nitride-based nonvolatile memory device. The early MNOS (Metal Nitride Oxide Silicon) [1.11][1.12] devices of the 1960s and mid-1970s were used an aluminum gate technology, relatively thick nitride (45nm), and thin tunnel oxides (2nm). In the latter 1970s and early 1980s, the technology moved into ploy-silicon gates SNOS (Silicon Nitride Oxide Semiconductor) [1.13][1.14] cells with reduced nitride thickness (25nm). Later, silicon-gate SONOS (Silicon Oxide Nitride Oxide Semiconductor) [1.15]-[1.17] devices was proposed in the mid-1980s and 1990. SONOS/MONOS provides the most promising structures with an additional top oxide. Holes are blocked from gate injection to the nitride as a result of its high barrier (3.9eV). Therefore, memory window is widened and nitride thickness can be scaled to around 5nm to 15nm. With the thinner effective thickness, SONOS/MONOS can be operated at lower voltage.

1.2 Uniform charge storage versus localized charge storage

Recently, considerable research efforts have been made to study nitride trap storage flash memories for its lower voltage operation, smaller cell size, simpler

fabrication process, and no drain-induced turn-on [1.18]-[1.20]. These cells have generally categorized into two types; the first one utilizes uniform charge storage in a nitride layer, such as conventional SONOS/MONOS and the second one has localized charge storage at the source/drain junctions, such as NROM [1.19] or Nbit [1.21] technology. The operation principles between the two types are compared in Fig. 1.4. Among NROM or Nbit structure, attention is particularly paid to its two-bit storage capability. The two bits operation can be achieved by placing programmed charges in the nitride layer locally above the source or the drain junction by channel hot electron (CHE) program and band-to-band hot hole (BTB HH) erase. To allow for two-bit operation, the applied bit-line voltage in reverse-read must be sufficiently large (>1.5V) to be able to "read-through" the trapped charge in the neighboring bit. Table. 1.1 shows the operation biases for program, erase, and read, respectively. The IV characteristics of program state and erase state are shown in Fig. 1.5. A narrow charge trapping region, typically tens of nano-meter, deteriorates the sub-threshold swing in program state. Because of a thicker bottom oxide and the non-uniformity of charge stored in the nitride layer, the reliability issues in Nbit cell are quit different from that in conventional SONOS cell. Several concerns are believed to possibly hinder Nbit's development. (a) Second bit effect is the inevitable outcome in two-bit operation [1.22]. It is related to the programming sequence and the P/E stress condition. At present, no effective way can mitigate this intrinsic effect. As a result, second bit effect may result in a serious scaling problem in this technology. (b) Although excellent intrinsic integrity is possessed, data loss is still a reliability concern after stress. In low-Vt state, Vt up-drift was reported by the bottom oxide hole tunneling [1.21]. Much effort has been devoted to alleviating the issue. For

example, a “soft hot electron injection” combined with a circuit judgment [1.23] is demonstrated to neutralize the positive hole charge. Reduction of the bottom thickness is another way to reduce the drift [1.24]. In high-Vt state, Vt loss was identified by Frenkel-Poole emission subsequently followed by oxide trap assisted tunneling [1.21]. The two-step charge loss closely relates to the bottom oxide robustness and/or the nitride trap energy deepness. Nitrided-oxide [1.25] or deuterium treatment [1.26] can better the bottom oxide quality. The ratio of SiH_4 to N_2 during nitride film growth has a significant influence on trap energy distribution [1.27]. Several groups have investigated the impact of trap energy on retention characteristics and attempted to extract the energy distribution [1.28], but some deficiencies and misunderstandings exist in their methodology. All previous literatures are work on the SONOS cell with an ultra-thin bottom oxide (1.5nm~2.5nm). They hypothesized the time through the bottom oxide can be ignored and the trapped charges are almost thermally activated in high temperatures (>85C). However, in such thicknesses, direct tunneling (DT) rather than thermal emission governs the charge loss mechanism even in these high temperatures. On the other side, for today’s SONOS cells, a thicker bottom oxide is usually employed, causing above hypothesis impractical. With a larger oxide thickness, more factors should be taken into account in a charge loss model. For example, a pronounced stress-induced oxide trap assisted leakage current is exhibited. In addition, charges re-trapping in silicon nitride are also significant [1.29]. This effect is complicated and necessitates a numerical technique for accurate modeling. (c) For Nbit cell, the read current fluctuation will be worsened due to the threshold voltage non-uniformity [1.30]. The design margin will be degraded and cause a failure, especially in

multi-level operation.

1.3 Organization of the Project

The scope of this project mainly focuses on reliability concerns of nitride-based storage memory, which are schematically illustrated in Fig. 1.6. Following the introduction, the mechanism of endurance degradation after P/E cycling stress in Nbit cells is described in Chapter 2. Proposed in Chapter 3 is a modified charge pumping technique for direct characterization of programmed charge lateral distribution in silicon nitride. P/E cycling stress induced programmed charge broadening effect is also investigated in this measurement. Various reliability issues including erase-Vt state threshold voltage instability, read-disturb, and high-Vt state charge loss are identified and the responsible models are made in Chapter 4. To further explore the charge loss mechanisms in high-Vt state, a numerical algorithm is established in Chapter 5. With the use of the numerical analysis, the influence of bottom oxide thickness and program/erase stress on charge retention is retrieved. An anomalous two-stage feature during retention measurement and the responsible origins are also discussed here. From the study in Chapter 5, a simple technique based on Frenkel-Poole emission to extract a silicon nitride trap density from stress induced nitride charge leakage current is developed in Chapter 6. Lateral threshold non-uniformity induced read current noise is verified in Chapter 7. A simple but reasonable model for read current distribution is proposed. Finally, conclusions are drawn in Chapter 8.

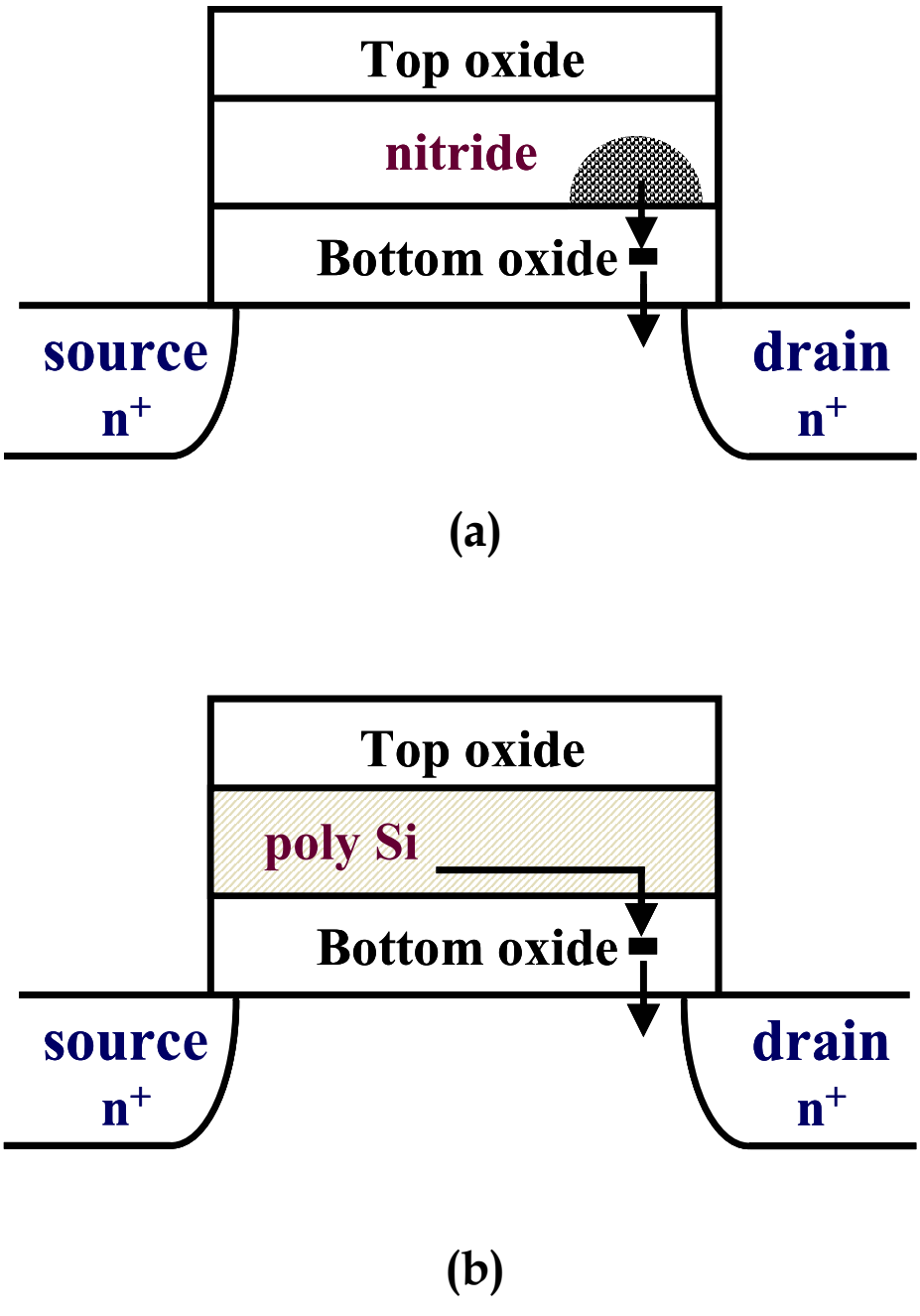


Fig. 1.1 Charge loss via the bottom oxide trap in a SONOS-type memory (a) and in a FG-type memory (b).

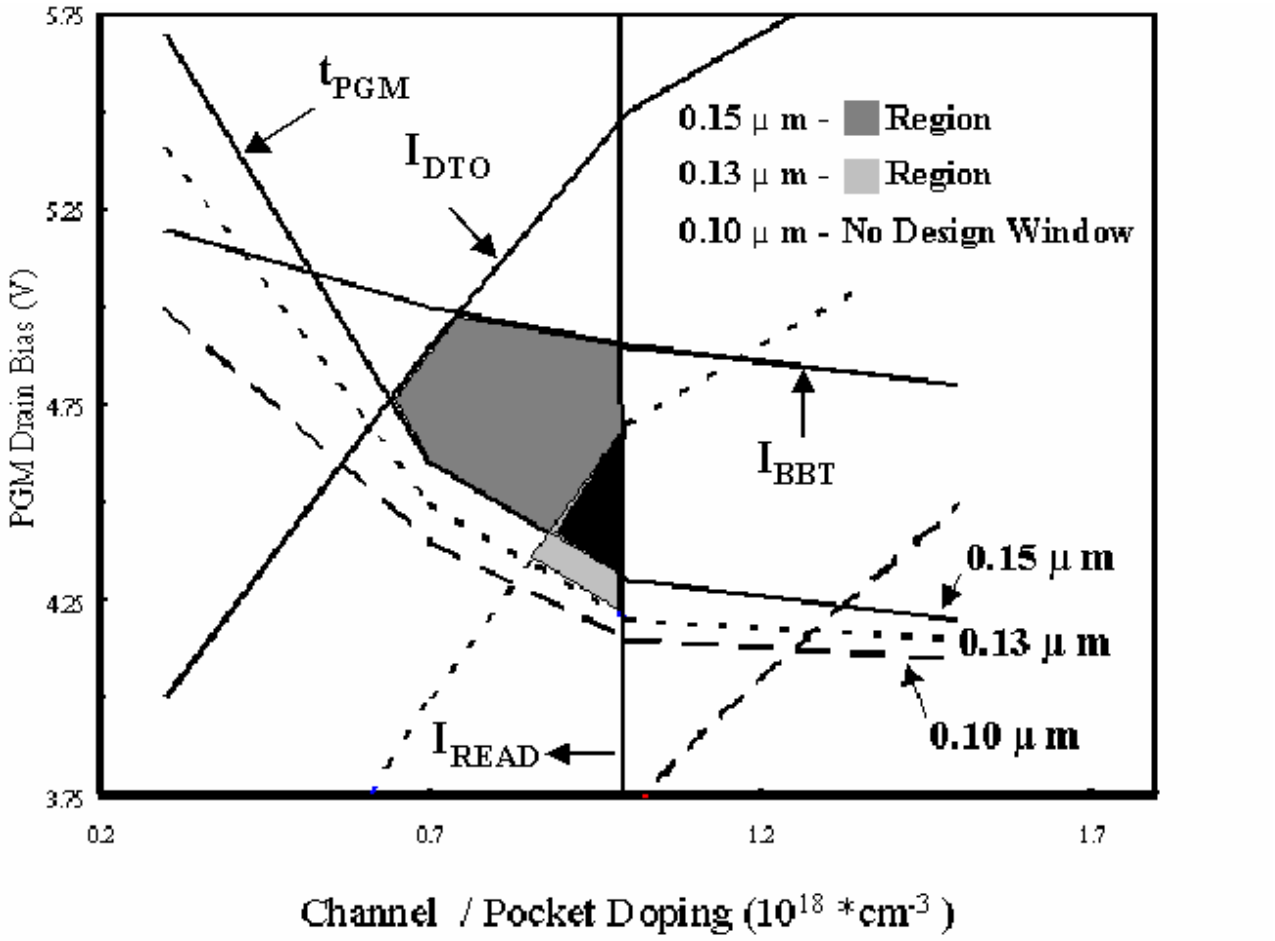


Fig. 1.2 Scaling limitation of NOR type floating gate flash. I_{DTO} is drain induced turn-on leakage current, I_{BBT} is band-to-band tunneling current, t_{PGM} refers to the hot electron programming time, and I_{READ} is the current needed for read. The shaded region refers to the design window. For the above design constraints, the design window will no longer exist when device channel length is below 0.13μm.

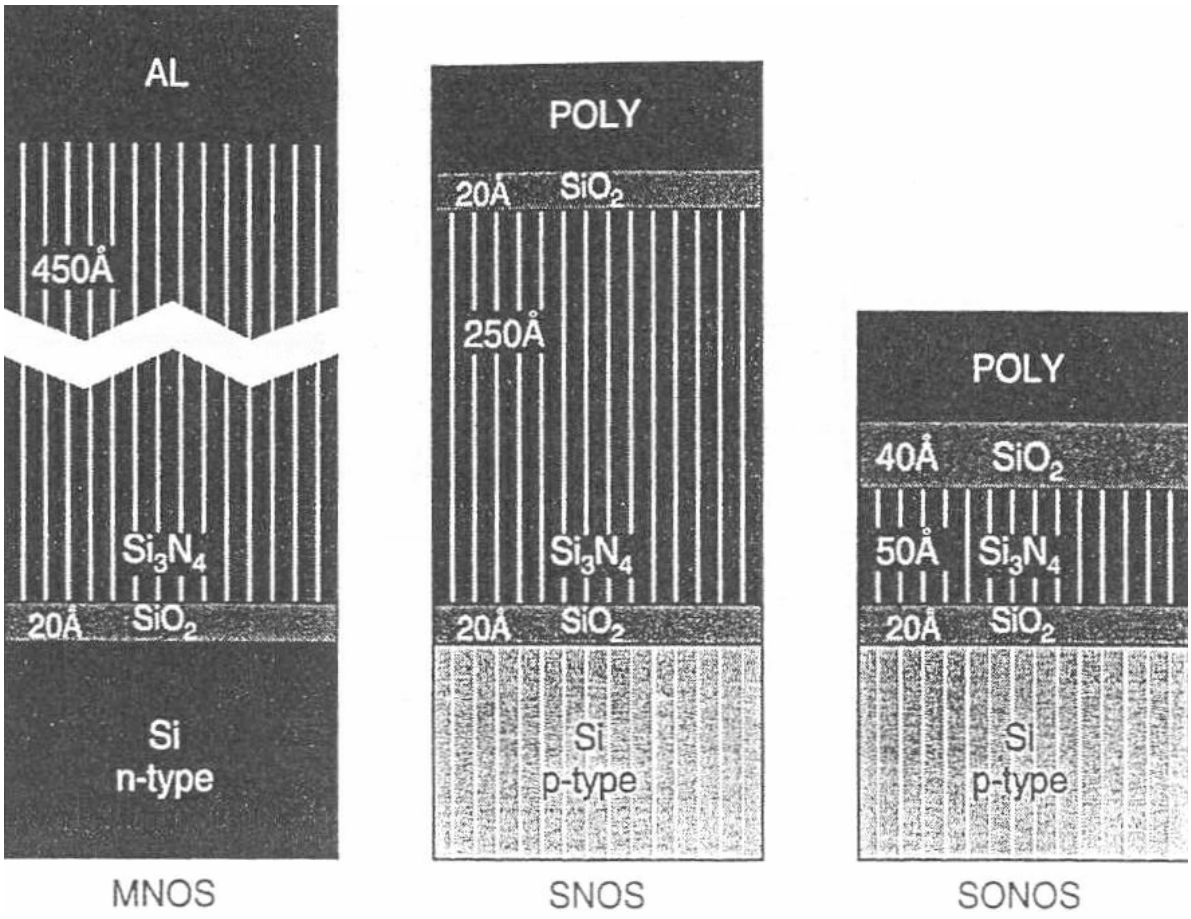
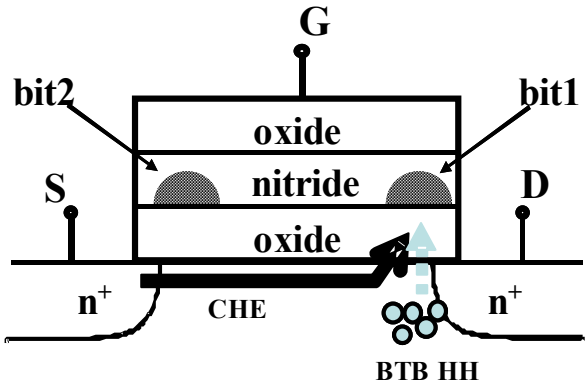
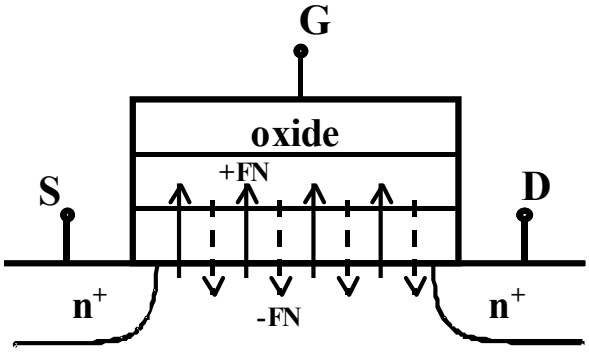


Fig. 1.3 Evolution of nitride-based NVSM.



(a)



(b)

	Conventional SONOS (15Å~35Å)	NROM or Nbit (>40Å)
	1 bit/cell	2 bits/cell
Program	FN	CHE
Erase	FN	BTB HH

(c)

Fig. 1.4 Program & Erase operations in Nbit cells (a) and in SONOS cells (b). The operation modes are summarized in (c).

Table. 1.1 Bias conditions for Nbit two-bit operation.

		Program	Erase	Read
Bit 1	Vg	11V	-3V	2.5V
	Vd	5V	8V	0V
	Vs	0V	0V	>1.5V
Bit 2	Vg	11V	-3V	2.5V
	Vd	0V	0V	>1.5V
	Vs	5V	8V	0V

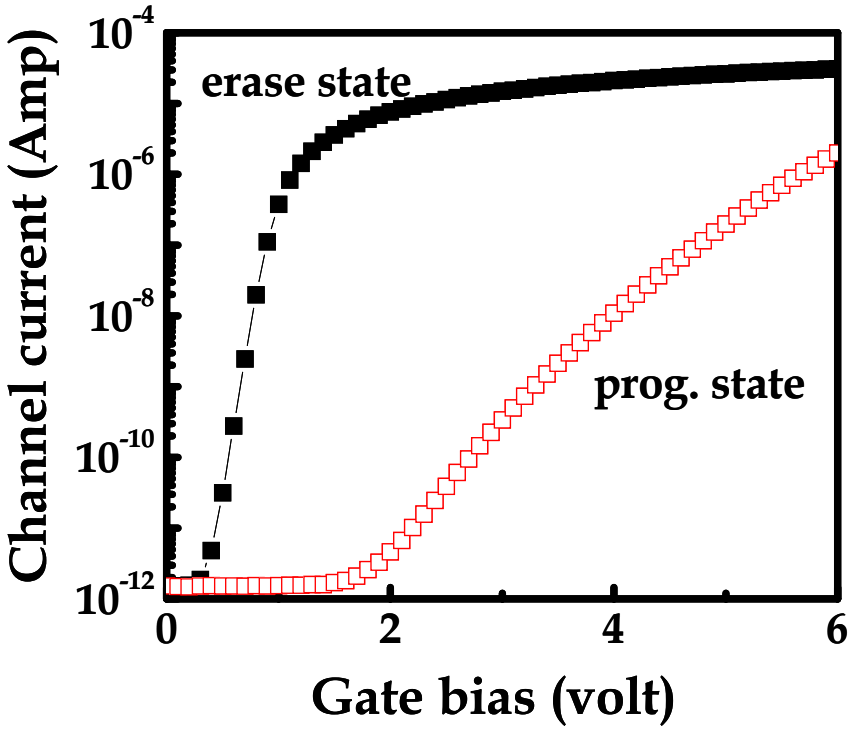


Fig. 1.5 Channel current versus gate bias in erase state and in program state.

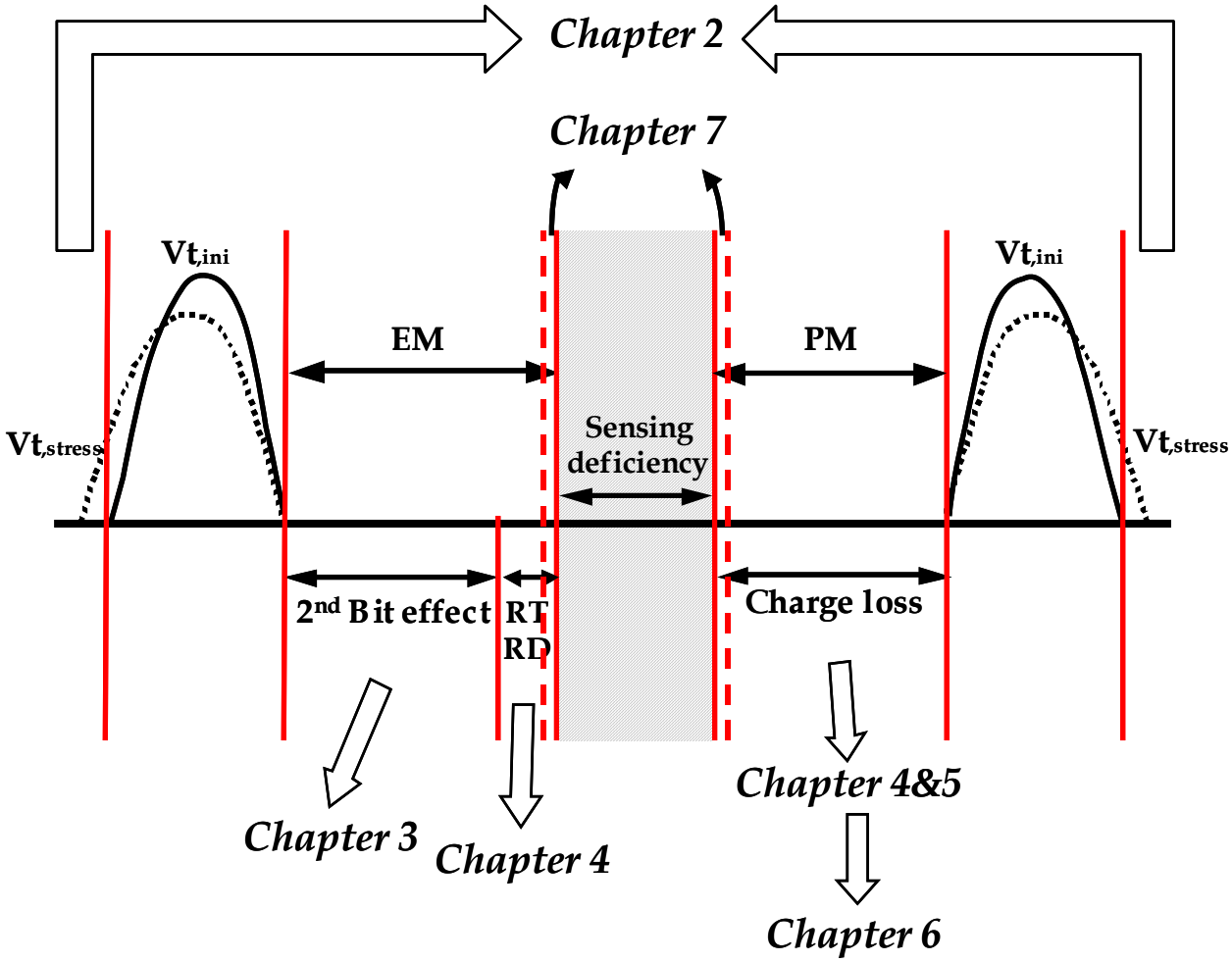


Fig. 1.6 The design margin for Nbit operation.

Chapter 2

Program/Erase Cycling Induced Endurance Degradation in Nbit Flash Memory Cell

2.1 Introduction

Endurance is an important issue in nonvolatile memory. It is defined by the number of data changes that can be performed in every cell of a given memory chip before one of the cells fails to meet the data sheet specifications. It describes the reliability of a device in terms of the number of program/erase (P/E) operations that can be performed on it without failure. Today, most commercially available nonvolatile memory products are guaranteed to withstand, at least 10,000 P/E cycles [2.1]. In a floating gate memory, electron trapping after numerous P/E cycles in tunnel oxide builds up a permanent negative charge, thereby reducing the electric field and injected tunneling current for the same applied terminal voltage. For a constant program voltage and program time, this reduces the program-state threshold voltage whereas increases the erase-state threshold voltage, resulting in a threshold voltage window closure problem [2.2]. In Nbit devices, the window closure is not observed (Fig. 2.1(a)). Instead, there appears an upward shift of both program-state and erase-state threshold voltages after P/E stress. We will discuss the cause for this endurance degradation in Nbit cells.

2.2 Endurance Failure in Nbit Cells

In conventional EEPROM memories, the degradation of the threshold voltage

window with P/E cycles is due to trap generation in the tunnel oxide. Threshold voltage window opening in the initial tens of cycles is caused by positive charge trapping, whereas window closing after 10^4-10^5 cycles is caused by electron trapping in the oxide (Fig. 2.1(b)) [2.3]. The Nbit cells apparently exhibit a different feature. Both program-state and erase-state threshold voltages move upward in parallel (Fig. 2.1(a)). This characteristic reveals that programming and erasing speeds are not degraded after cycling stress and implies that the failure of endurance may result from the aging of the tunnel oxide, including interface state creation [2.4] or oxide charge accumulation caused by cycling stress [2.5]. Interface state creation can be easily excluded since the sub-threshold swing in a 100k-P/E-cycled device is similar to that of the fresh one, no matter in program state or in erase state, as shown in Fig. 2.2. Based on the above observation, it seems that oxide trap/oxide charge creation should be responsible for the endurance failure.

2.3 Evidence of Negative Charge Creation

The sub-threshold characteristics at program state and erase state in a fresh device and in a cycled device are compared in Fig. 2.2. Although no significant swing degradation is observed, we do observe a parallel shift of the I-V curve before and after P/E stress. The parallel shift in Fig. 2.2 is usually attributed to negative trapped charge creation in the bottom oxide. Thus, the ONO dielectric layers actually consist of two kinds of charge in program state; negative trapped charge in the bottom oxide (Q_N) and negative charge in the nitride (Q_{ox}), as illustrated in Fig. 2.3. In erase state, the nitride charge (Q_N) can be neutralized by band-to-band hot hole injection through either electron-hole recombination or compensation while Q_{ox} is still present

in the bottom oxide. Fig. 2.3 is a diagram illustrating the charge distribution in the ONO in program state (Fig. 2.3(b)) and in erase state (Fig. 2.3(c)). Q_{ox} is negligible in a fresh device and increases with P/E stress. The build-up of the bottom oxide charge Q_{ox}, which cannot be removed by erase, explains the upward shift of the threshold window with P/E cycles.

The relative positions of Q_{ox} and Q_N depicted in Fig. 2.3 are constructed by the following measurements. A fresh device and a 100k cycling stressed device are used for this study. GIDL current and threshold voltage are measured. GIDL current is used as a monitor for the charge in the ONO layers above the n+ drain region while threshold voltage can be used to measure the ONO charge in the channel region (please refer to Fig. 2.4(a)) [2.6]. First, we adjust the programming bias condition for each device to make sure that program-state GIDL current is the same in each device. The measured program-state threshold voltage is 4.13V in the fresh device and 4.62V in the 100k device. The same GIDL current means the two devices have the same amount of charge above the n^+ drain. The higher threshold in the 100k cell is believed due to the additional stress created oxide charge Q_{ox}. Then, the erase characteristics of the two devices are measured. Fig. 2.4(b) shows the threshold voltage shift with erase time in the two devices. It is interesting to note that the two curves in Fig. 2.4(b) are almost identical. Two questions are brought about. First, why the two devices have the same erasing speed? This can be understood because the two devices have the same GIDL current. The readers should be reminded that GIDL reflects the strength of hot hole erasing. Second, why the erasing in the 100k cell stops around 1millisecond and the threshold voltage shift is the same as in the fresh cell although the program-state threshold in the 100k device is about 0.5V higher? This implies that

part of the ONO charge in the 100k device, amounting to 0.5V threshold voltage shift, cannot be erased. Why this charge cannot be erased? What kind of charge is it? We believe the answer is oxide trapped charge Q_{ox}. From the above arguments, the picture of the oxide charge and the nitride charge distributions in the 100k device is constructed as in Fig. 2.4(b).

In the second experiment, both the devices are programmed to have the same threshold voltage. The measured GIDL current in the 100k device is found to be smaller than that in the fresh device. Fig. 2.5(a) shows the measured result and the corresponding charge distributions. Again, the threshold voltage shift versus erase time in the devices is measured (Fig. 2.5(b)). Now, the fresh device has a larger threshold shift. Our explanation is as follows. Although the total ONO charge in the channel region in the two devices are equal (i.e. the same program-state threshold), the ONO charge in the fresh cell is completely the erasable nitride charge and thus the threshold voltage shift in the fresh device during erase is larger.

Finally, we would like to compare the P/E stress effects in the conventional gate MOSFET and in the Nbit cell. Fig. 2.6(a) shows the ΔVt in the MOSFET and in the Nbit cell. Since oxide charge creation is the only cause for the shift in the nMOSFET, the similarity between the NMOSFET and the Nbit in Fig. 2.6(a) indicates that solely oxide charge creation can explain the observed endurance degradation in the Nbit cell

2.4 Improvement of Cycling Endurance

From the above study, we conclude that the shift of the threshold voltage window in the Nbit cell is attributed to bottom oxide charge creation. The endurance

of the operation window can be improved by increasing bottom oxide robustness. It has been well reported in literature that thinner oxide has less charge creation [2.7]. The reduction of bottom oxide thickness should be an effective approach to improving the endurance. Another approach is to strengthen the erase bias condition. A stronger erase will result in accumulation of excess holes in the nitride to compensate for the negative oxide charge or even increases the possibility of recombination of injected hot holes and oxide trapped electrons.

Fig. 2.6(b) shows the threshold voltage window by using a stronger erase bias. Endurance up to 10^6 cycles can be achieved.

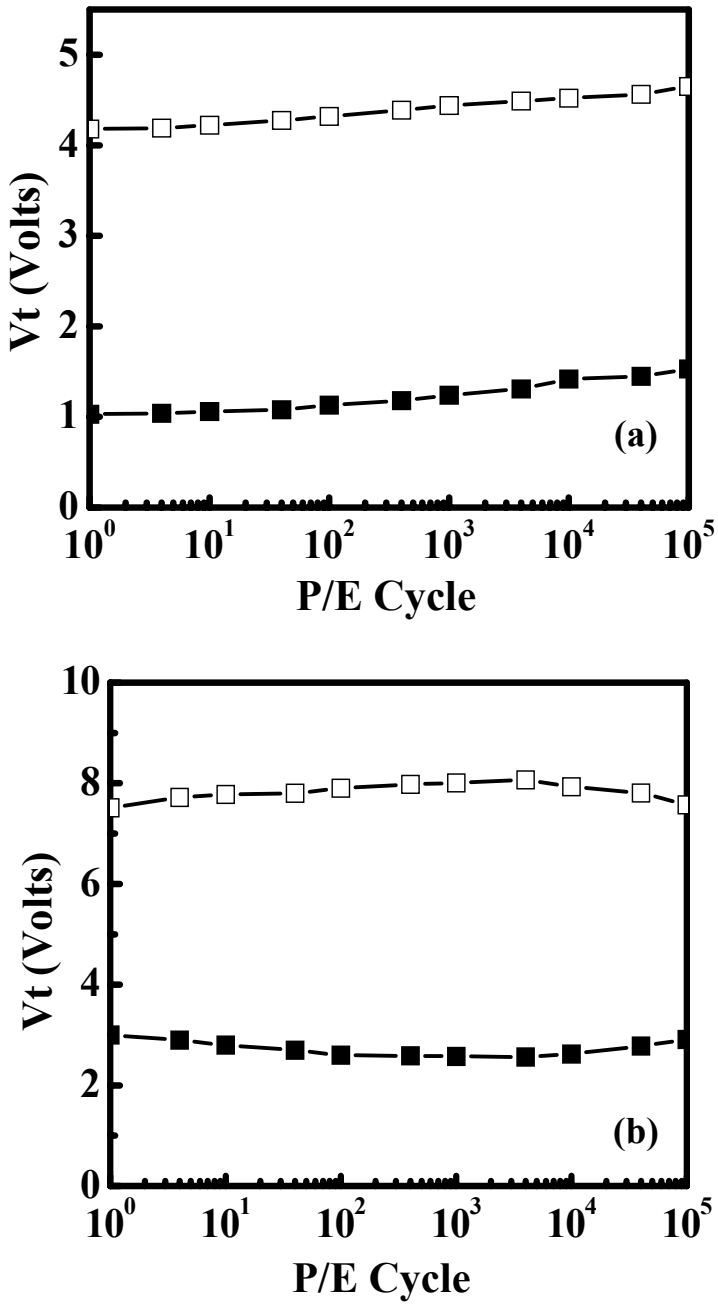


Fig. 2.1 (a) Endurance characteristics of a Nbit cell operation window. Program bias is V_g/V_d=11V/4.7V, 10μs and erase bias is V_g/V_d=-3V/, 3ms. The thickness of each ONO layer is 7nm, 6nm and 6nm. (b) Endurance characteristics of a floating-gate flash cell showing the window opening and closing.

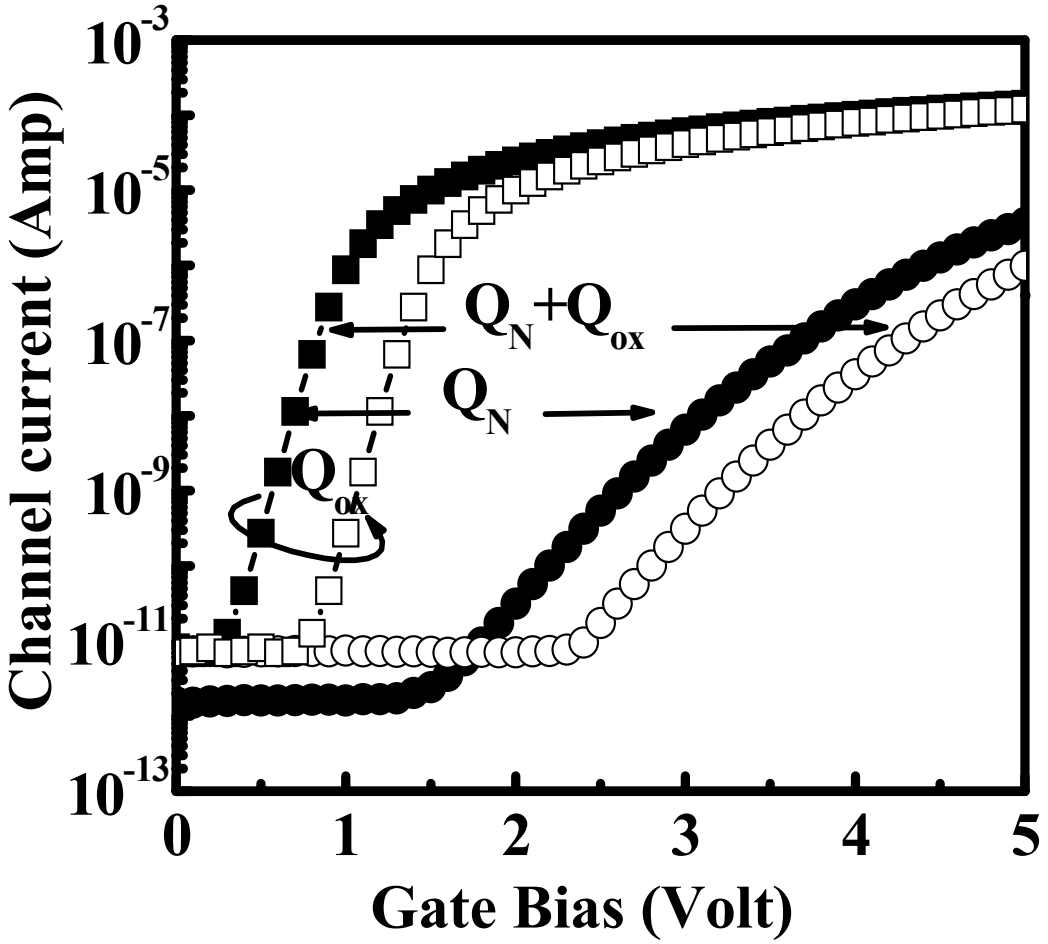


Fig. 2.2 Nbit cell subthreshold charcateristics in erase state and in program state before stress and after 100k cycling stress. Q_N and Q_{ox} represent the charge created by P/E stress and the injected charge by programming, respectively.

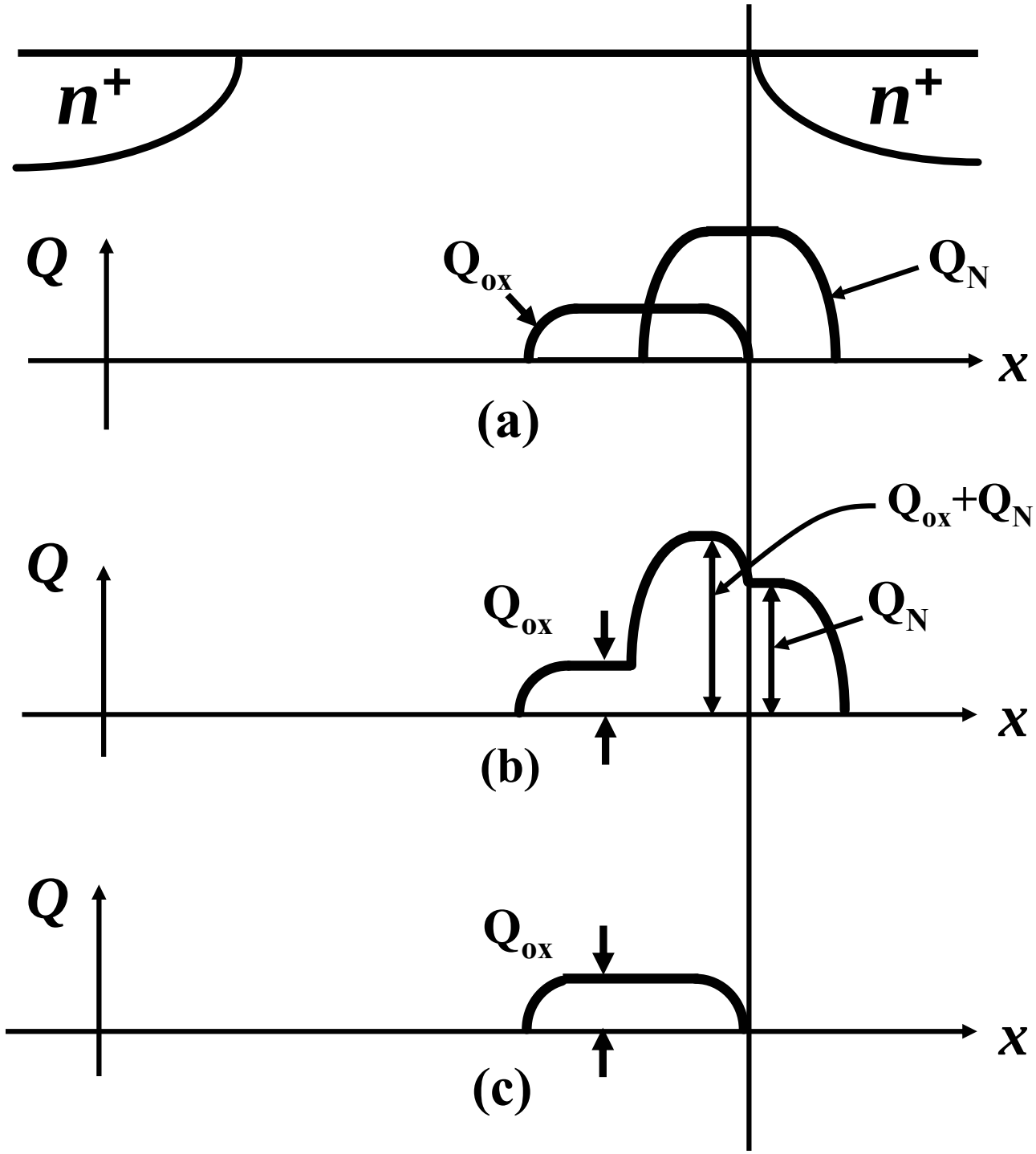
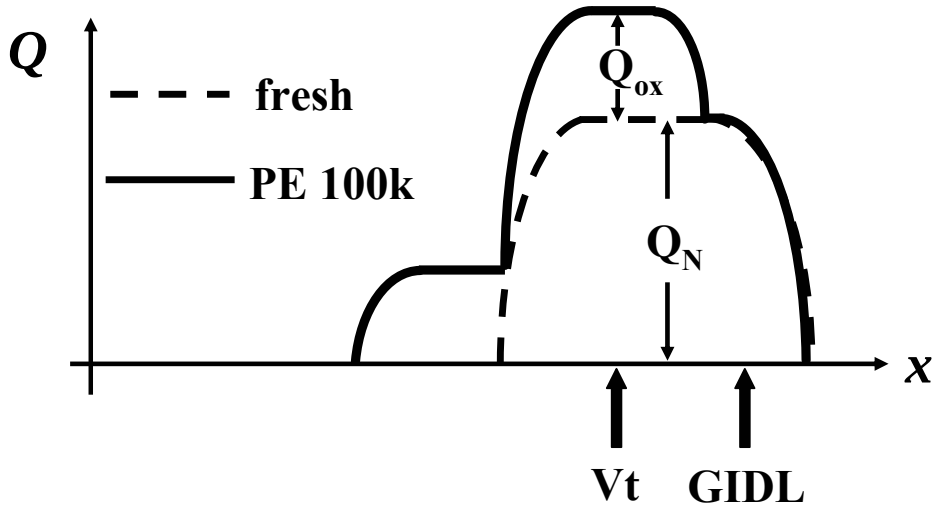
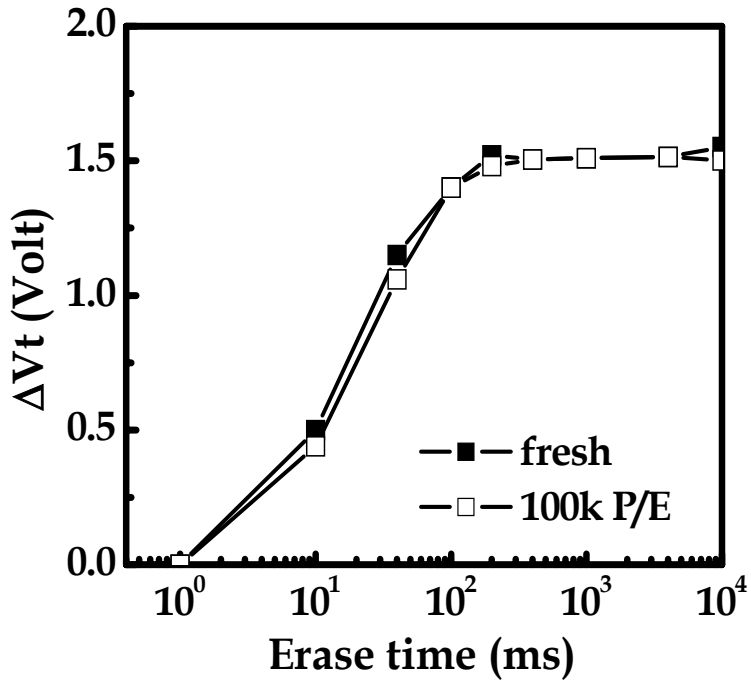


Fig. 2.3 (a) The relative position of Qox and Q_N in a cycled cell. The x-axis is the channel direction. There is an overlap between Qox and Q_N. (b) The charge distribution of Qox and Q_N in program state after cycling stress. (c) Qox in erase state in a stressed cell.



(a)



(b)

Fig. 2.4 (a) Illustration of the charge distribution in the fresh and the 100k P/E devices. The two devices are programmed to have the same GIDL current. (b) Threshold voltage shift with erase time in a fresh cell and in a 100k cycled cell.

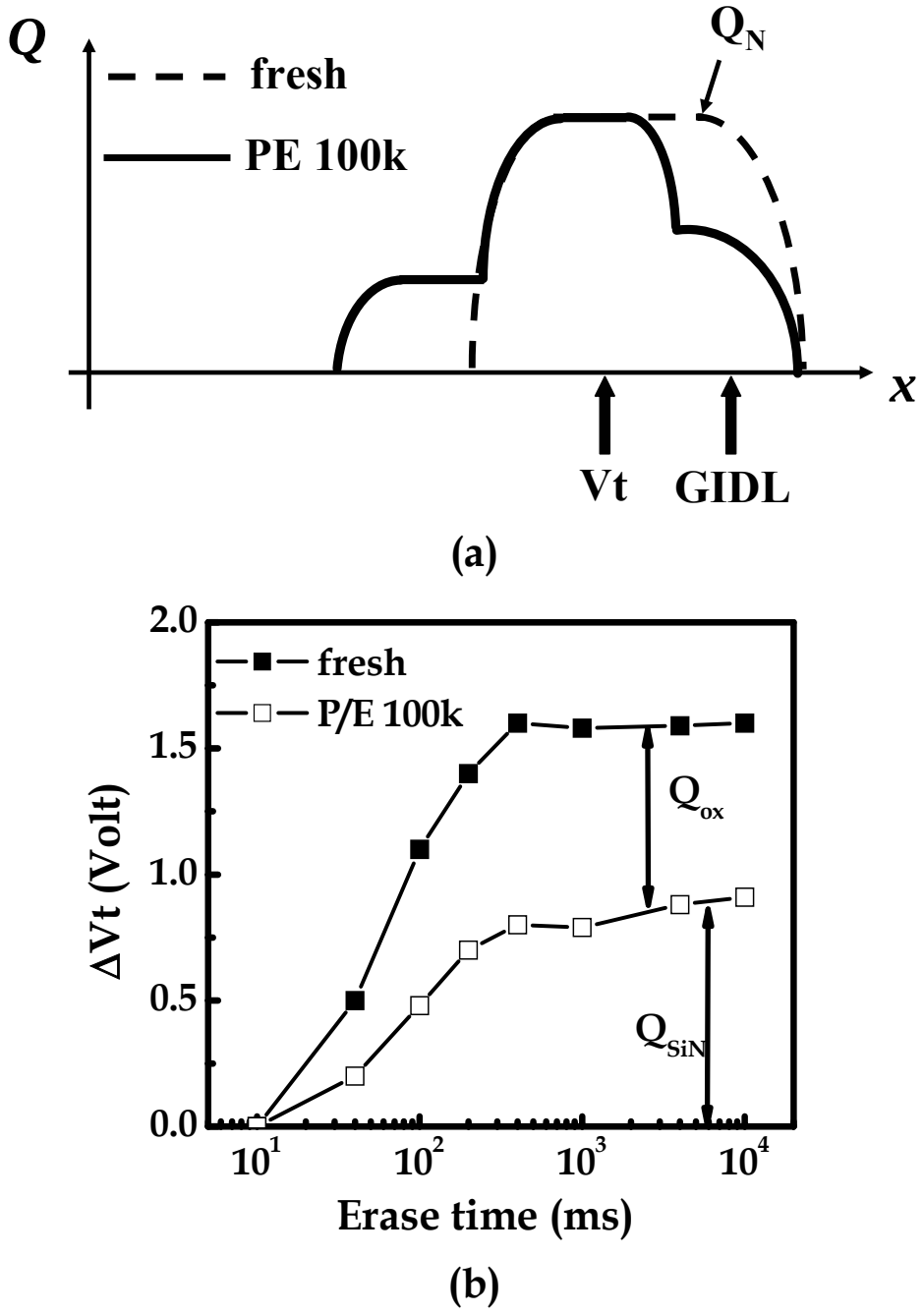


Fig. 2.5 (a) Illustration of the charge distribution in the fresh and 100k P/E devices. The two devices are programmed to have the same threshold voltage. (b) The erase characteristics in the fresh and in the 100k P/E devices. The two devices have the same program - state threshold voltage.

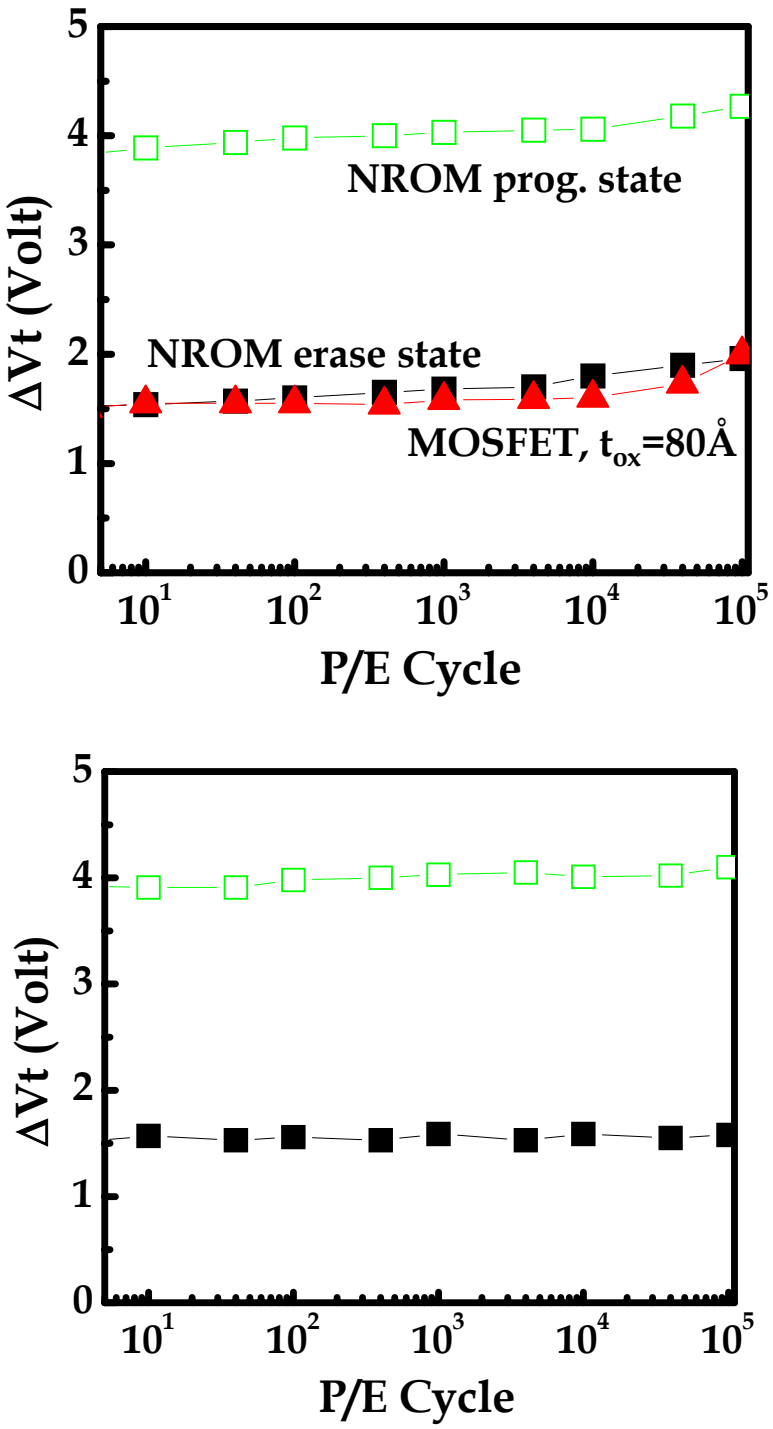


Fig. 2.6 (a) Threshold voltage variation versus P/E cycle number in a conventional gate MOSFET and in a Nbit. (b) Threshold voltage versus P/E cycle number in a Nbit cell with a stronger erase bias.

Chapter 3

Characterization of Programmed Charge Lateral Distribution in Nbit Flash Memory Cell by Using Charge Pumping Technique

3.1 Introduction

In two-bit operation, the control of programmed charge lateral distribution of each bit is a major concern for the scalability of the Nbit cell. The stored electrons at the first bit will affect the threshold voltage of the second bit in reverse read and vice versa. This phenomenon is referred to as the second bit effect [3.1] and is closely related to programmed charge lateral spread. Furthermore, the lateral spread of stored charges in nitride will result in the degradation of erase capability or erase speed due to a spatial mismatch between stored electrons and injected holes during erase [4.2]. For these reasons, comprehensive understanding of programmed charge spatial distribution is of vital importance in the optimization of the cell structure and operation bias.

Attempts have been made in the past to characterize the trapped charge distribution in a Nbit cell [3.3][3.4]. Larcher et al. used an inverse modeling approach to extract programmed charge distribution from measured I-V characteristics [3.3]. This method, however, suffers from some drawbacks, such as lack of precise information on device doping profile and extensive numerical calculation to reach a consistent solution between stored charge distribution and measurement result.

In this chapter, we will use a modified charge pumping technique [3.5] to probe

the lateral distribution of programmed charges at the source and drain junctions separately without using computer simulation. The devices and measurement setup throughout this study will be described. The charge pumping current for single bit storage and two-bit storage will be demonstrated. Program/erase cycling stress effect on stored charge distribution will be also examined.

3.2 Device Structure and Measurement Setup

The nitride flash cells used in this work have a gate length of 0.5μm and a gate width of 1.0μm. The thickness of each ONO layer is 9nm (top oxide), 6nm (nitride) and 6nm (bottom oxide). The cell intrinsic threshold voltage (Vt) is about 1.6V where Vt is defined as the gate voltage when the drain current is 1μA at a reverse read voltage of 1.6V. Channel hot electron program and band-to-band hot hole erase accompanied by a reverse read scheme are adopted to achieve two-bit per cell operation.

In charge pumping (CP) measurement, the voltage waveforms at gate and drain terminals supplied by a two-channel pulse generator are illustrated in Fig. 3.1. The employment of a dual-channel pulse generator can circumvent the misalignment of gate and drain signal. The gate pulse has a fixed high level (V_{gh}) and a variable low level (V_{gl}). The V_{gh} is sufficiently high (V_{gh}=6V here) to ensure that the entire channel at program state is inverted.

The charge pumping current, named hereafter Icp, is measured at the substrate. To probe the lateral extent of programmed charge in the drain side (or the source side), V_d (or V_s) is adjusted to modulate the drain (or source) depletion width while V_s (or V_d) is left floating. In this way, there is no channel current and thus no impact

ionization induced substrate current in CP measurement. Besides, the V_d is 180^o phase-shifted with respect to V_g so that the drain signal is applied only during the interface-trap (N_{it}) electron emptying period. Because the length of the channel hot electron injection region (i.e., programmed region) is only about a few nanometers [3.6], Icp contributed by interface traps in the programmed region is very low and close to the measurement limit of the current setup. Therefore, a higher frequency of 2.5MHz with 50% duty cycle and rise/fall times of 15ns each is selected in CP measurement. To make sure the measured Icp is still reliable at this frequency, the frequency dependence of normalized charge pumping current at program state is shown in Fig. 3.2. The constant Icp/f confirms the validity of the CP measurement.

3.3 Measurement Result and Discussion

3.3.1 Single-bit Storage

Fig. 3.3 shows the Icp versus V_{gl} curves in a virgin cell, after programming only, and after one P/E cycle, respectively. Only the first bit (drain side) is P/E cycled and the V_d in CP measurement is 0V. The threshold voltage window (ΔV_t) is 2V. The Icp in a virgin cell and in erase state are almost identical, whereas a noticeable Icp bump at program state is noticed. This Icp bump is attributed to a local increase of channel threshold voltage due to negative nitride charge trapping. To verify the nitride charge storage effect, the Icp characteristics for two different threshold voltage windows, ΔV_t=2V and 2.7V, are compared in Fig. 3.4. As expected, the larger threshold voltage window exhibits a larger Icp bump due to more stored electrons. The dependence of the program-state Icp bump on V_d in charge pumping measurement is shown in Fig. 3.5. The Icp bump is suppressed as V_d increases. The

reason is that the drain depletion region increases with Vd. At a sufficiently large Vd (in this figure 1.8V), interface traps underneath the program charges are completely "masked" by the drain depletion region. Thus, these interface traps can no longer go through inversion-accumulation cycles in CP measurement and do not contribute to Icp. As a result, the program-state Icp bump is totally suppressed. Our result here implies that a Vd about 1.8V is necessary in reverse read to avoid completely the second bit effect. In contrast, when Vs is applied in CP measurement, the Icp bump is not affected at all (Fig. 3.6). This indicates that programmed charge is highly localized near the drain edge.

3.3.2 Two-bit Storage

The Nbit cell stores two bits at different locations (drain-side and source-side). Each bit within a cell serves as binary unit of data that is totally mapped to four states in a memory cell. The Icp of the four states corresponding to "11", "10", "01" and "00" are shown in Fig. 3.7. "00" denotes both bits in program state. "10" (or "01") denotes the drain-side bit in erase state (or in program state) and the source-side bit in program state (or in erase state). To explore the influence of the programming sequence on trapped charge spatial profile, the charge pumping measurement is performed after each bit is programmed. Fig. 3.8 shows the Icp of the first programmed bit and the second programmed bit, respectively. Here, the Icp of the second programmed bit is obtained in two ways. One is to measure Icp after the first bit is erased. The second approach is to subtract the fist bit Icp from the "00" state Icp. Fig. 3.9 compares the second bit Icp from the above two approaches and the result is almost the same. It should be emphasized that a crossover of the first bit and the

second bit Icp in Fig. 3.8 is observed. This suggests that the second programmed bit has a wider charge distribution but a smaller peak density. The reason will be discussed later. To profile the nitride stored charge lateral distribution, a technique similar to [3.6] is performed. In profiling, we make the following assumptions. First, we assume that a fresh cell has uniform interface trap density (Nit) along the channel [3.6]. The charge pumping current thus should have a linear dependence on channel position x;

$$x = \frac{I_{cp}(V_{gl})}{I_{cp,max}} L \tag{3-1}$$

where x=0 is defined at the edge of the source or drain junction. L is the channel length and Icp,max denotes the saturated charge pumping current. The second assumption is that Nit generation after one program/erase cycle is negligible. Based on these assumptions, the nitride charge distribution is deduced as follows,

$$Q_N(x) = \frac{C_{ONO}}{q}(V_{gl} - V_{ti}) \tag{3-2}$$

where $Q_N(x)$ is the nitride charge density at the position x, and Vti stands for the threshold voltage of a fresh device. Fig. 3.10 depicts the extracted stored charge distribution of the first programmed bit and the second programmed bit versus a distance from the source/drain junction. The stored charges extend into the channel about tens of nanometers. This result is in agreement with the simulation in [3.3]. In addition, the CP measurement result shows that the second programmed bit exhibits

a broader distribution but a smaller peak density. To explain this result, a two-dimensional device simulation is performed. A rectangular charge distribution with a width of 30nm and a charge density of $1.6x10^{19}$ cm^{-3} is used. Fig. 3.11 shows the lateral channel electric field distribution in the second bit programming. A large channel field exists not only in the programmed region but also in the first bit region (drain side). Such a large field in the first bit region will accelerate channel electrons from the drain and cause earlier hot electron injection into the nitride, thus resulting in a broader second bit distribution. Finally, we would like to remark that Eq. (3-2) and consequently Fig. 3.10 are derived from a simplified one-dimensional model. For a narrow charge distribution by hot electron programming, Eq. (3-2) only serves as a first-order approximation. Accurate profiling of programmed charge distribution requires a more complicated 2D model.

The programmed charge lateral extent can be also probed by varying V_d (or V_s) in CP measurement. Fig. 3.12 shows the difference in Icp between program state and erase state versus V_d (or V_s). ΔIcp is measured at V_{gl}=1.6V. The 2nd programmed bit needs a larger junction bias to "mask" the programmed charge. The same conclusion that the second bit has a broader charge distribution is obtained.

3.3.3 P/E Cycling Stress Effect

The P/E cycling stress effect on programmed charge distribution is examined in Fig. 3.13 by using the variable V_d method. The Vt window keeps the same during cycling. Again, the ΔIcp is measured at V_{gl}=1.6V and is normalized to its value at V_d=0V to compensate for interface trap creation effect for different cycling stress. It has been reported that hot-carrier stress created oxide traps spread toward the

channel with stress time [3.7][3.8]. Such traps, especially positively charged traps, can effectively lower the Si/SiO2 injection barrier [3.9][3.10] and enhance the electron injection probability. Therefore, as cycle number (stress time) increases, the hot electron injection region expands toward the channel due to the spread of the bottom oxide damaged region. A larger Vd in CP measurement is necessary to screen programmed charge at a larger cycle number. The consequence of the broadening of programmed charge distribution is the degradation of the second bit effect. Fig. 3.14 shows the threshold voltage versus Vd in reverse read for different cycle numbers. The second bit effect is apparently worsened with increasing cycle number.

3.4 Summary

We have characterized the programmed charge lateral distribution in a two-bit storage nitride flash cell by channel hot electron program. The charge pumping measurement reveals that the charge distribution of each bit extends into the channel for tens of nanometers. This suggests the possibility of further scaling down the nitride flash cell with respect to the concern of the overlap of two bit charges. Our study also shows that the charge distribution of the secondly programmed bit is influenced by the stored charge of the first bit. A broader second bit charge distribution is obtained.

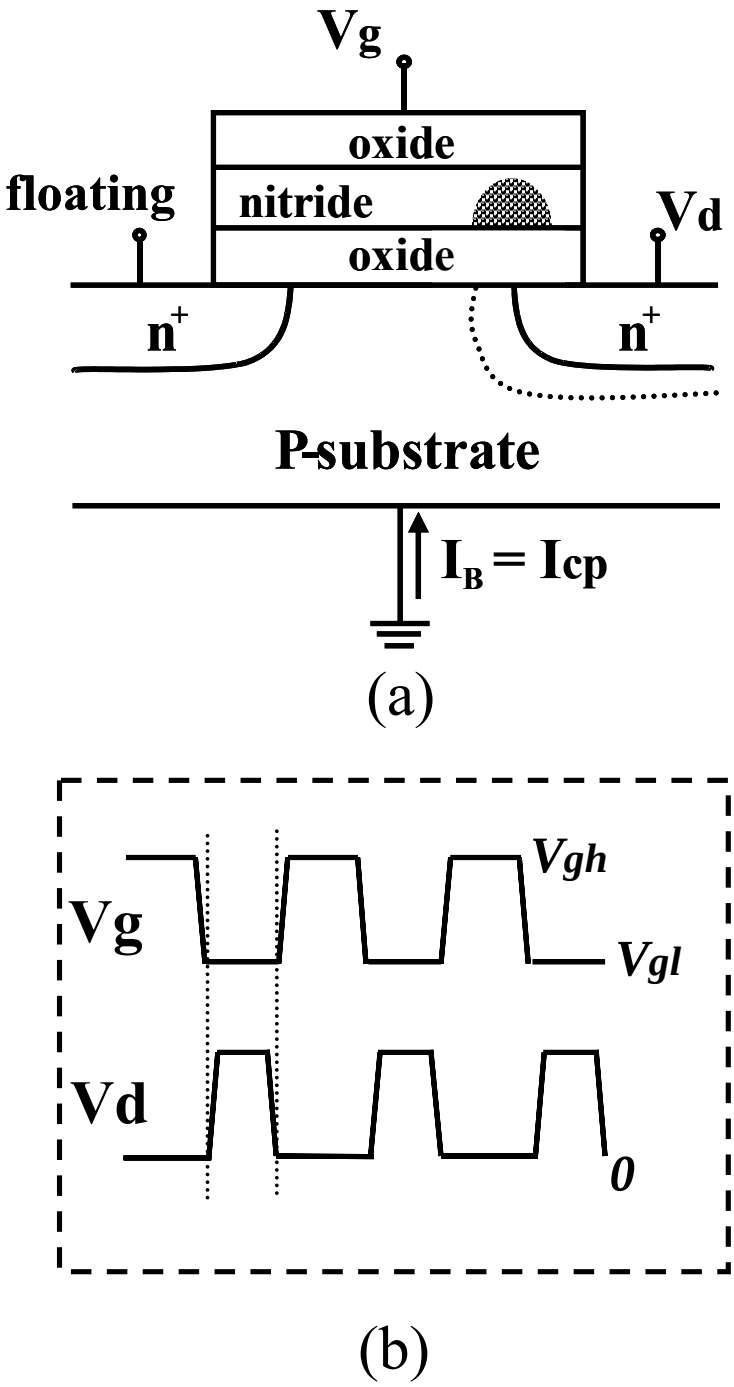


Fig. 3.1 (a) Schematic diagram of a two-bit storage nitride flash cell and (b) charge pumping measurement waveform. The dashed line in the substrate represents the depletion region caused by Vd. The thickness of the ONO gate stack is 9nm (top oxide), 6nm and 6nm, respectively.

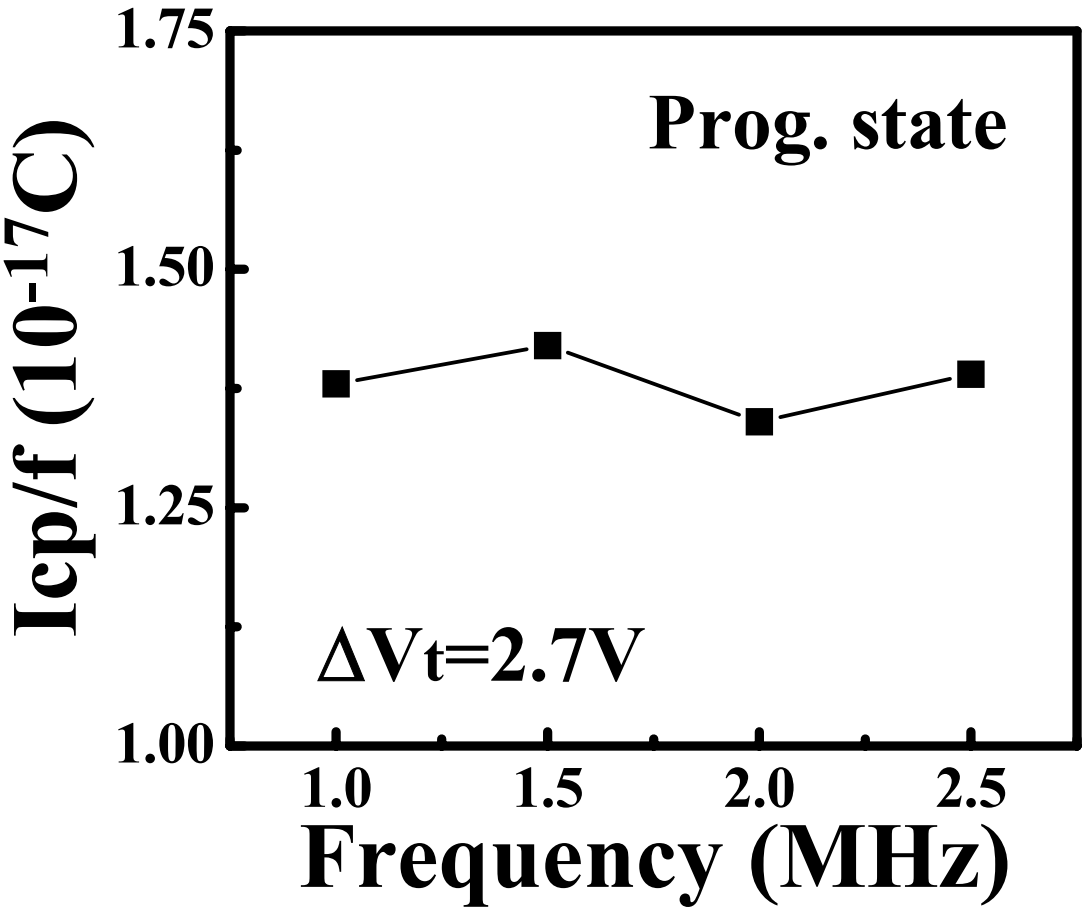


Fig. 3.2 The dependence of normalized charge pumping current (Icp/f) on measurement frequency. In charge pumping measurement, the Vt window (ΔVt) is 2.7V and Icp is measured at V_{gl}=2.5V.

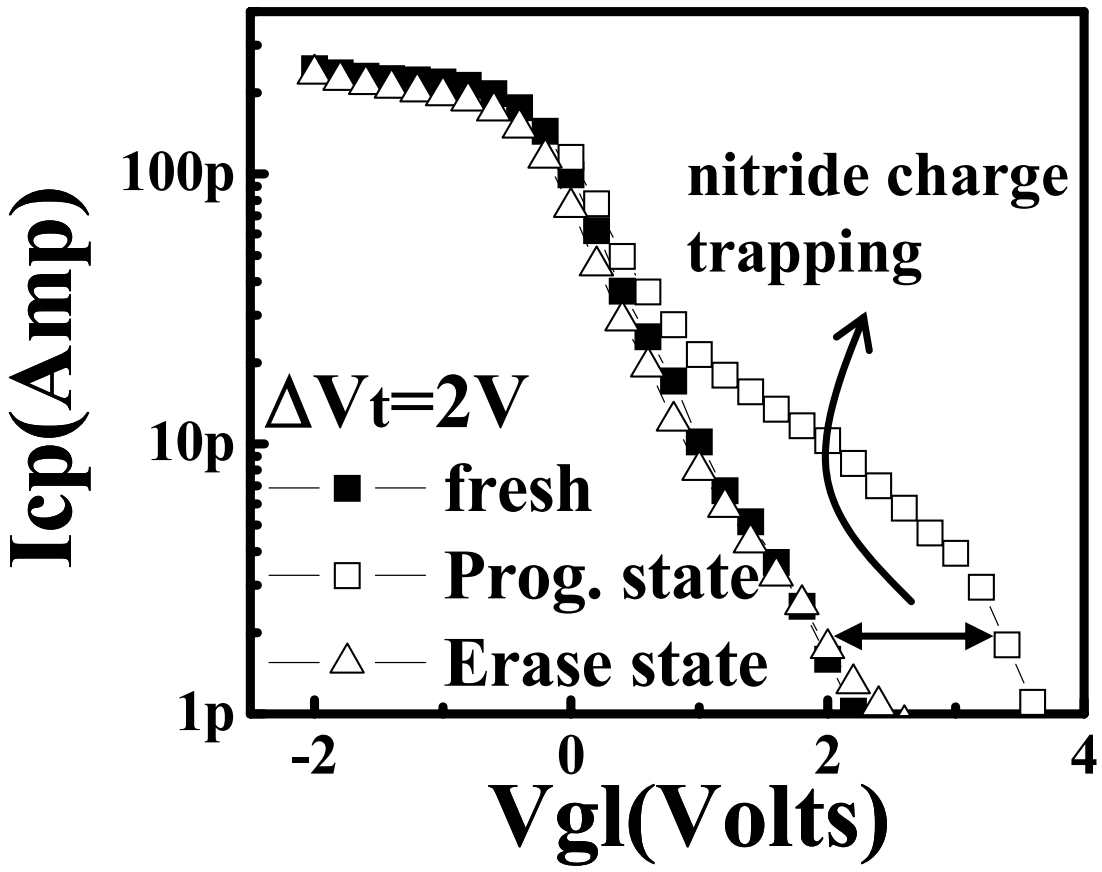


Fig. 3.3 Icp versus V_{gl} in a fresh cell, in program-state and in erase state, respectively. The Vt window (ΔVt) is 2V. V_d in CP measurement is 0V.

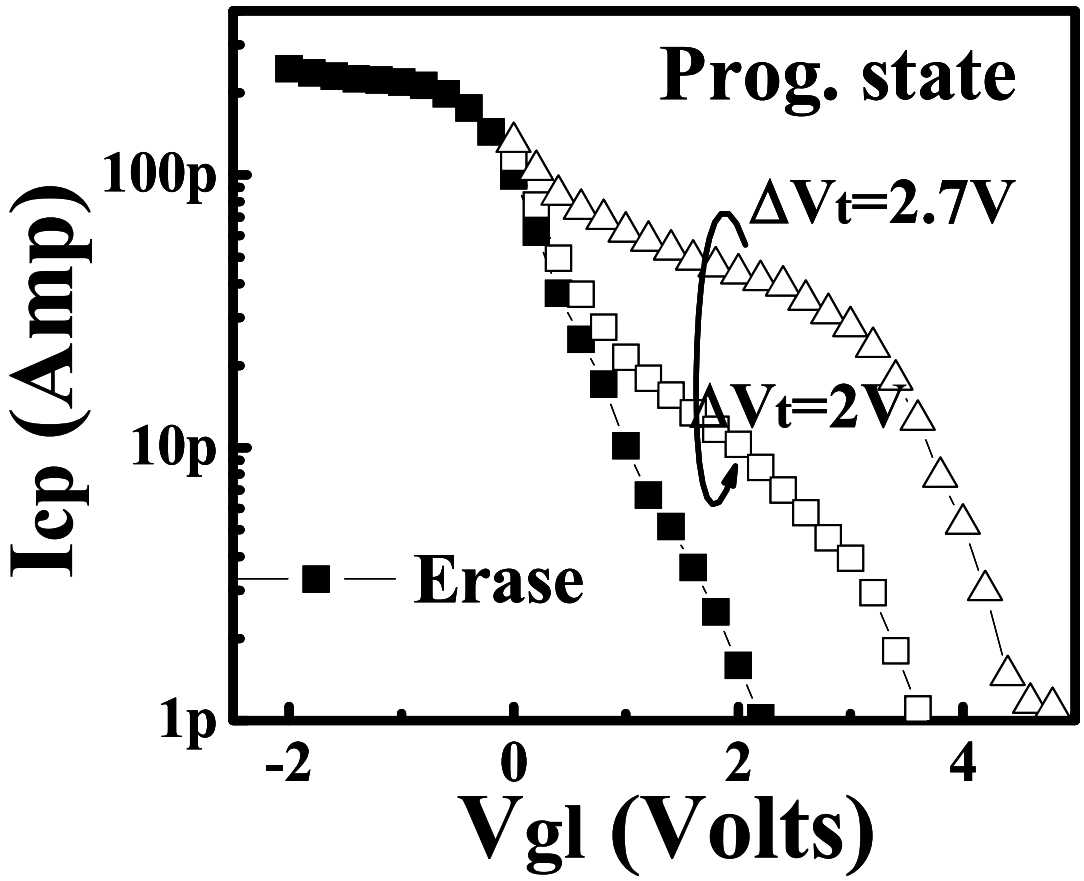


Fig. 3.4 Icp versus V_{gl} for different Vt window. The program-state Icp bump increases with Vt window due to more injected charges.

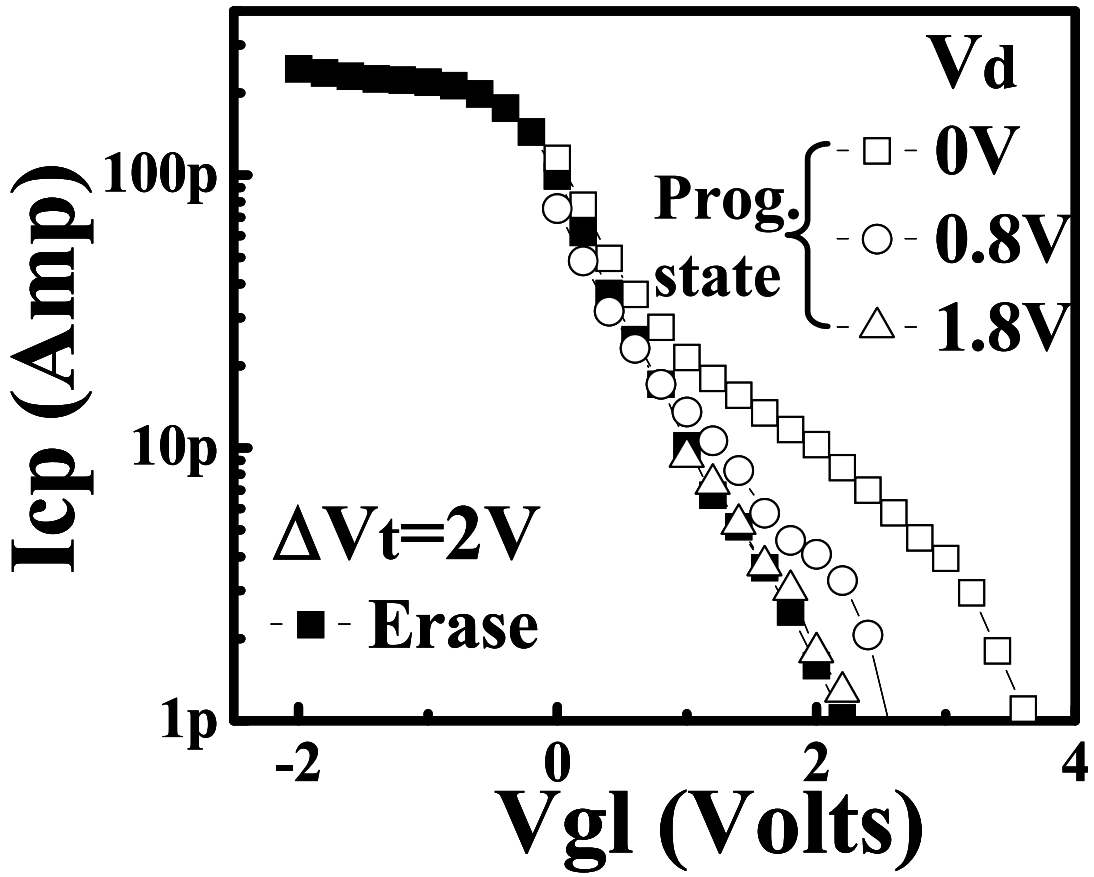


Fig. 3.5 Icp versus V_{gl} with different V_d in CP measurement. The program-state Icp bump decreases with V_d. The Vt window is 2V.

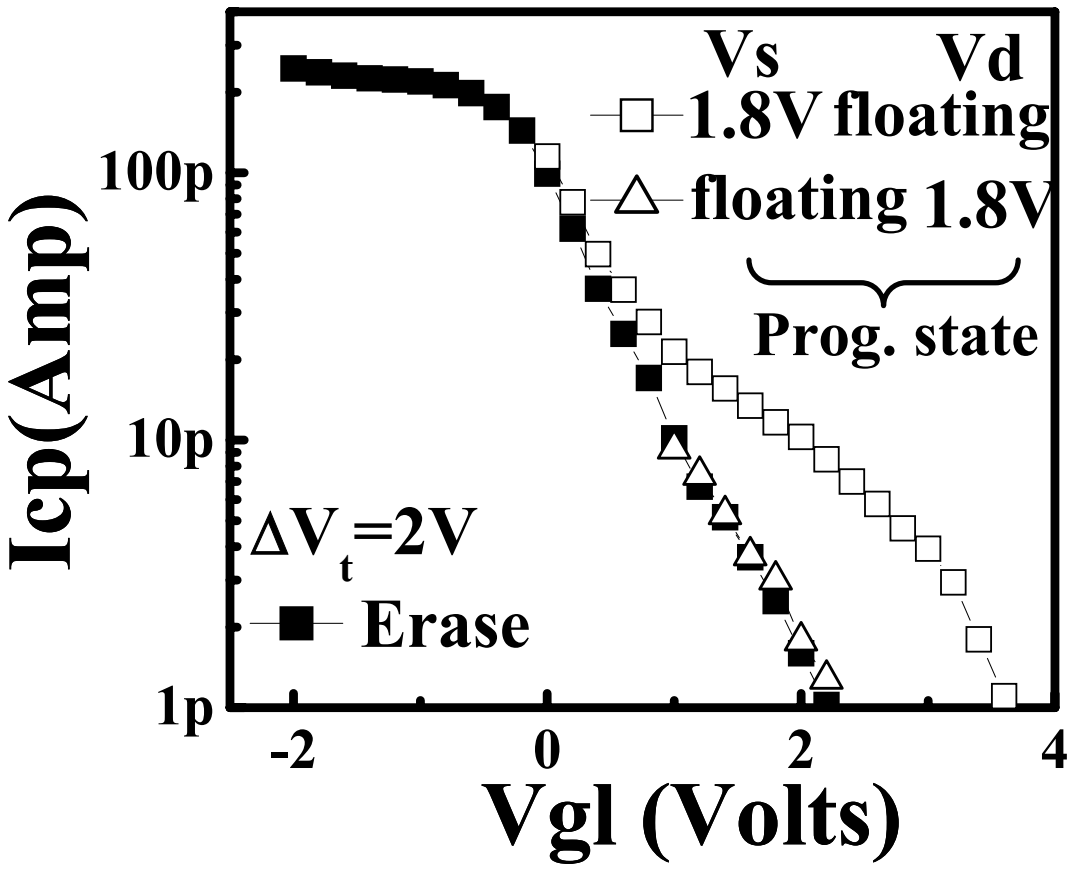


Fig. 3.6 The program state Icp measured at V_d=1.8V & V_s floating and at V_s=1.8V and V_d floating.

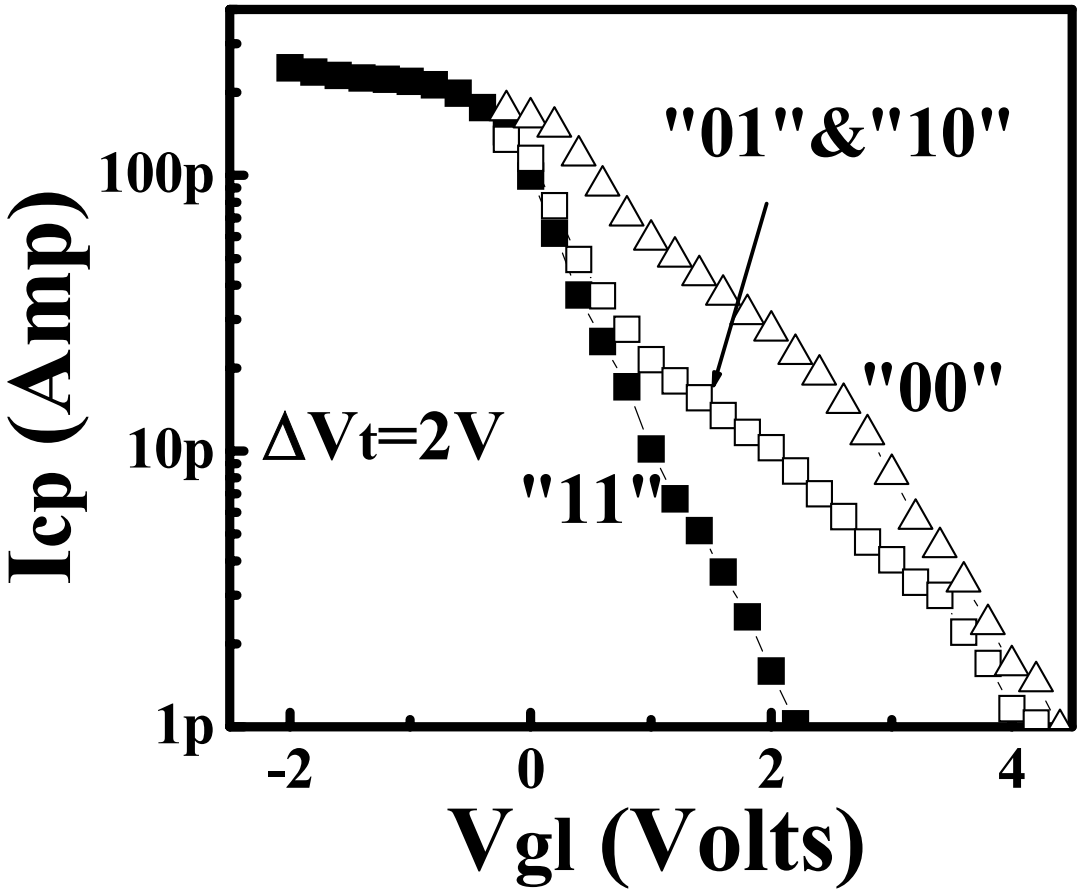


Fig. 3.7 The Icp versus V_{gl} of the four states of two-bit storage. “11” represents both bits in erase-state and “10” represents one bit in erase-state and one bit in program-state.

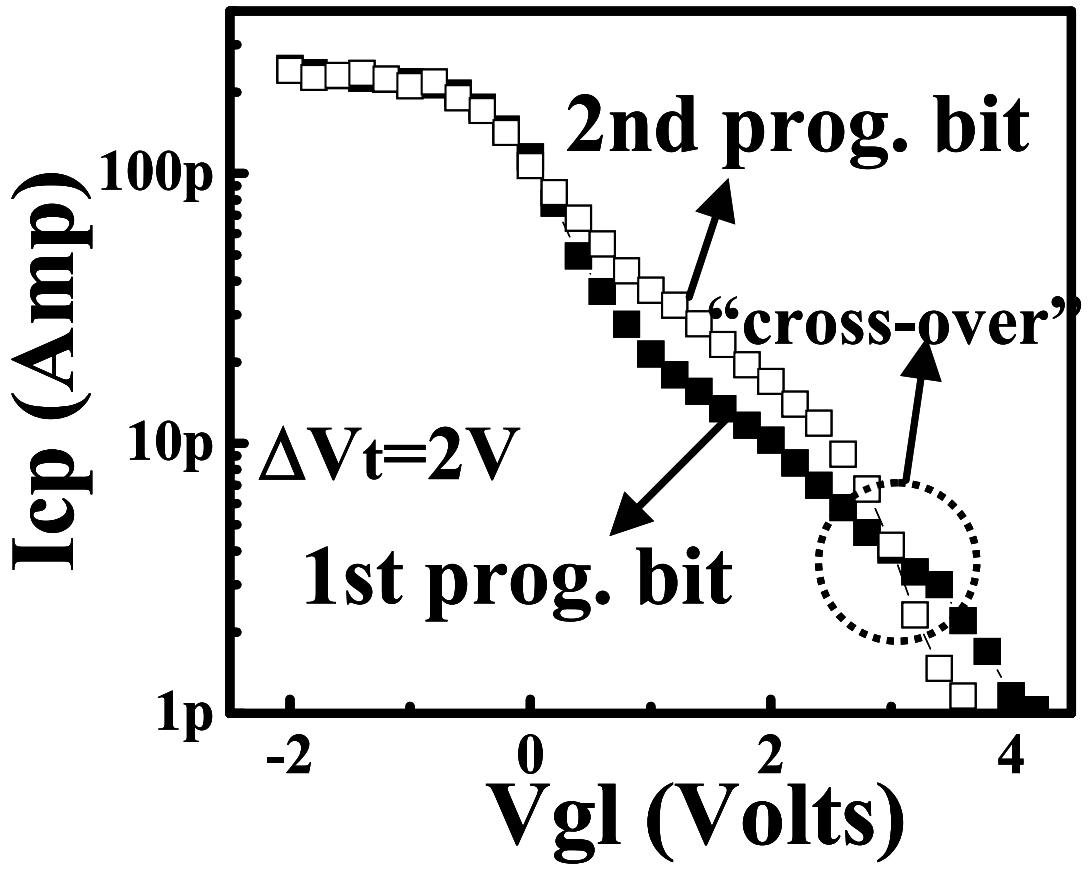


Fig. 3.8 Comparison of the Icp versus V_{gl} of the first programmed bit and the secondly programmed bit. The 2nd bit Icp is measured with the first bit erased.

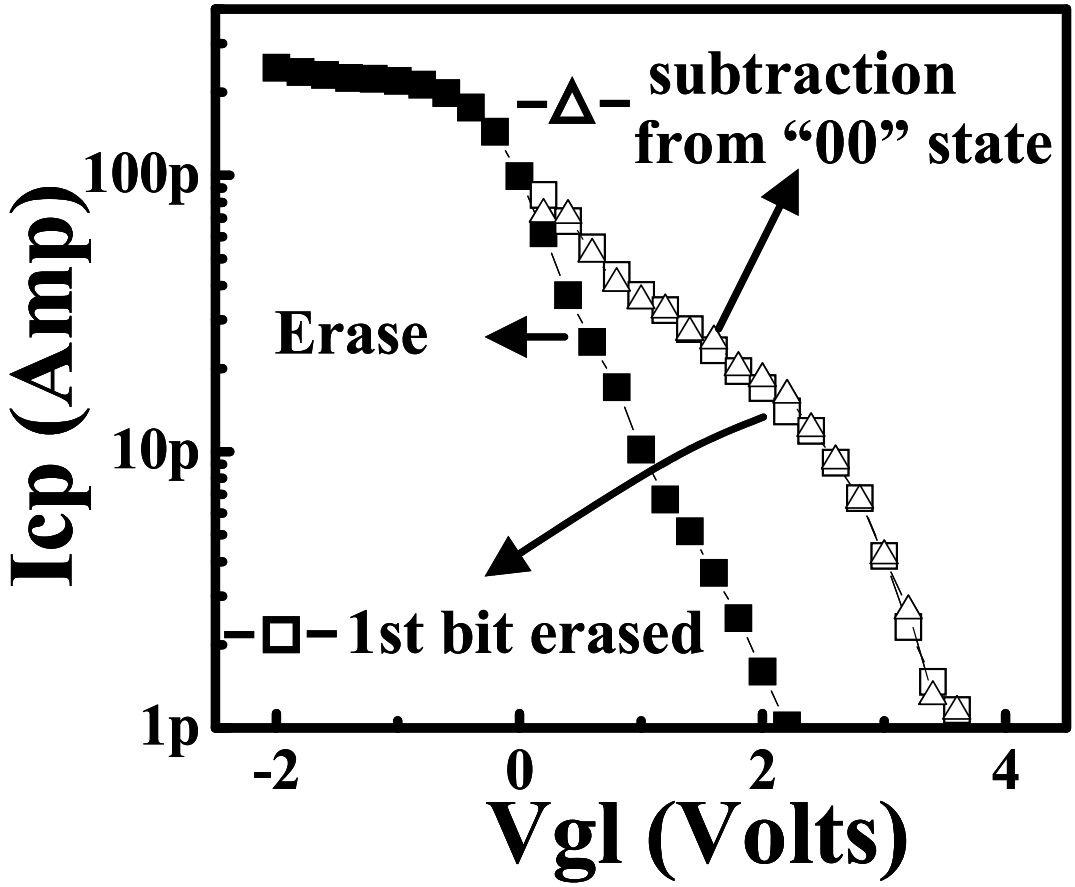


Fig. 3.9 Comparison of the second bit Icp-V_{gl} from two approaches. One is to measure Icp after the first bit is erased. The second approach is to subtract the fist bit Icp from the "00" state Icp.

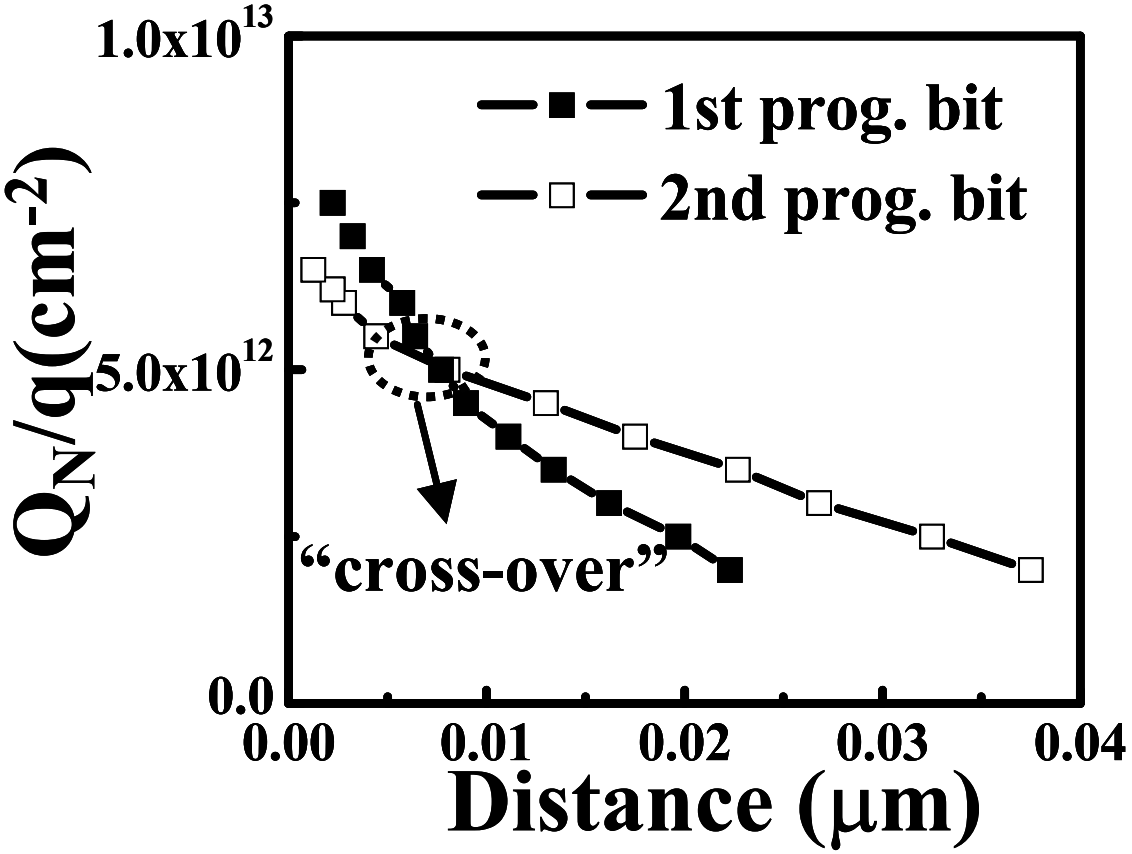


Fig. 3.10 Lateral profiling of the programmed charge distribution of the first programmed bit and the secondly programmed bit. A uniform interface trap distribution along the channel is assumed. Icp,max in Eq. (3-1) is 195pA.

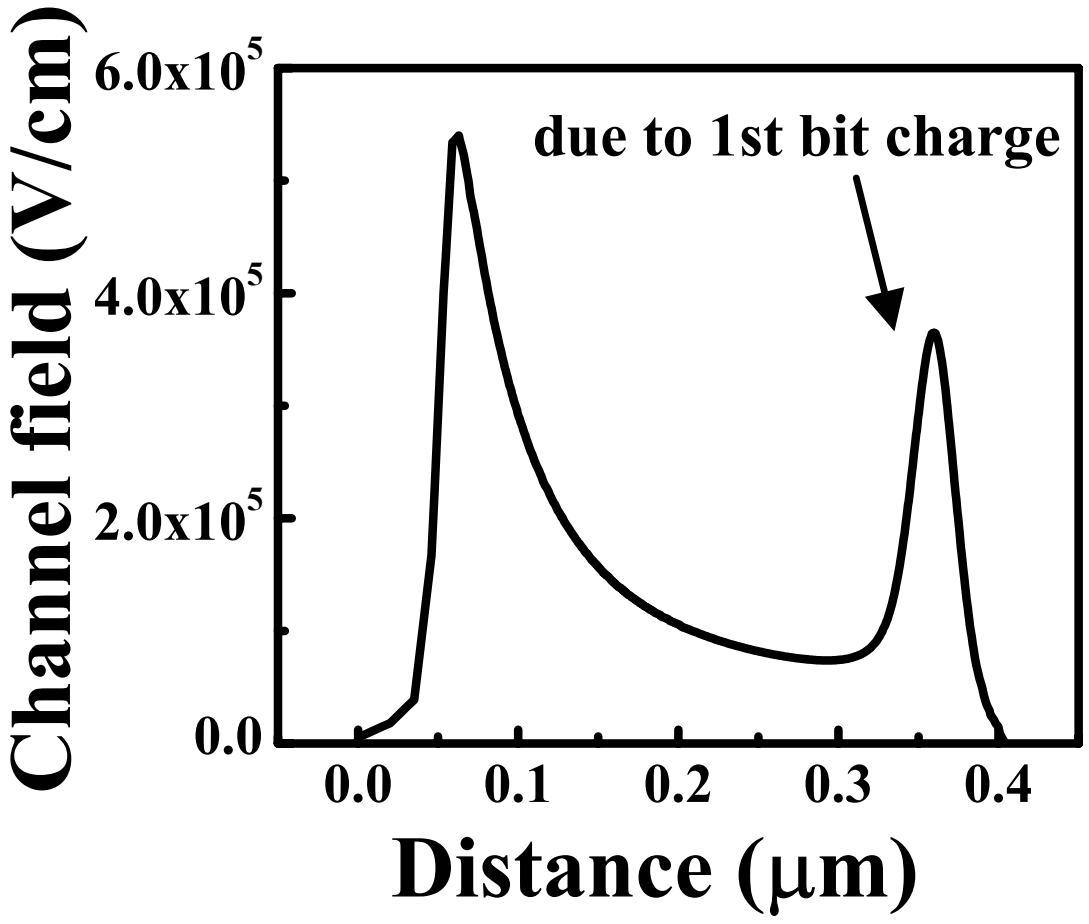


Fig. 3.11 Simulated channel field distribution in 2nd bit programming from 2D device simulation. x=0 is at the n^+ source edge and x=0.4 is at the n+ drain edge. Vs=6.5V and V_g=11 in 2nd bit programming.

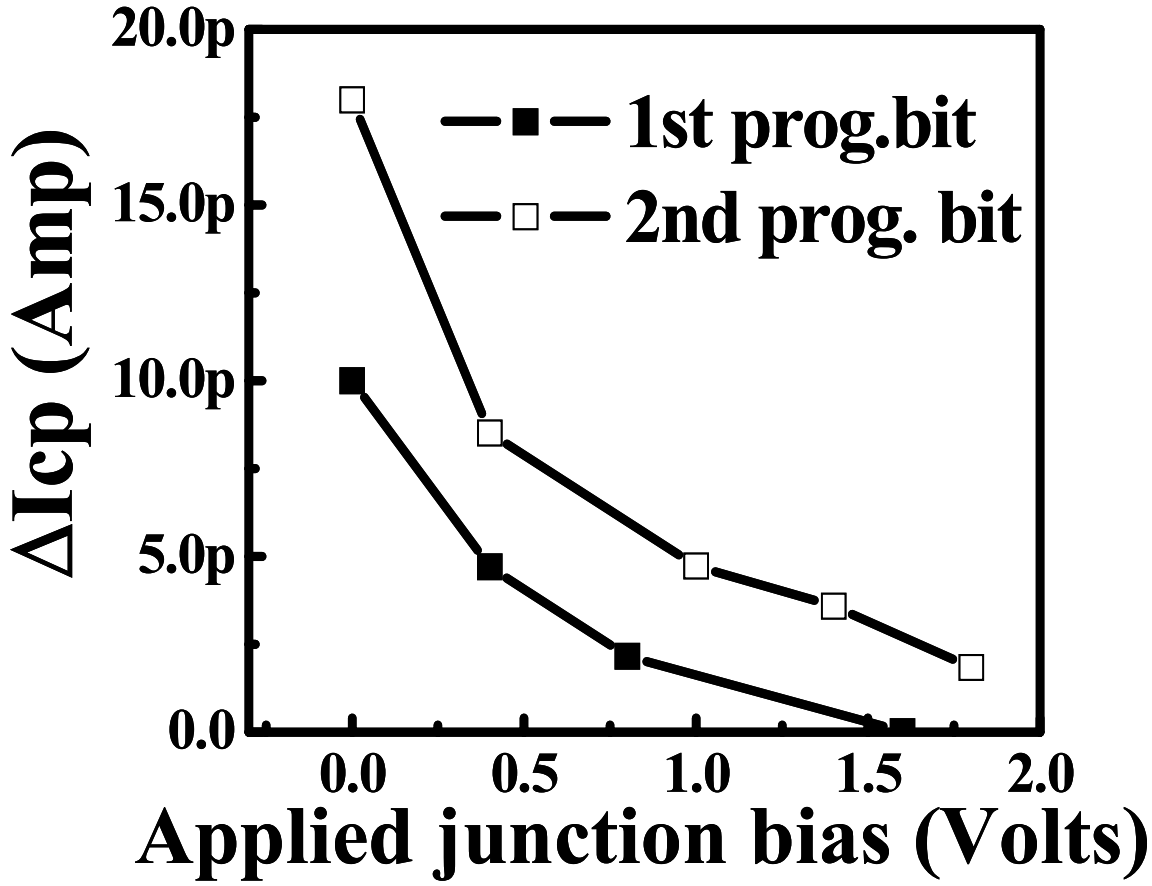


Fig. 3.12 The difference in Icp between program-state and erase-state as a function of drain bias for the 1st bit and source bias for the 2nd bit. The Δcp is obtained from Fig. 3.7 and Fig. 3.8 at V_{gl}=1.6V.

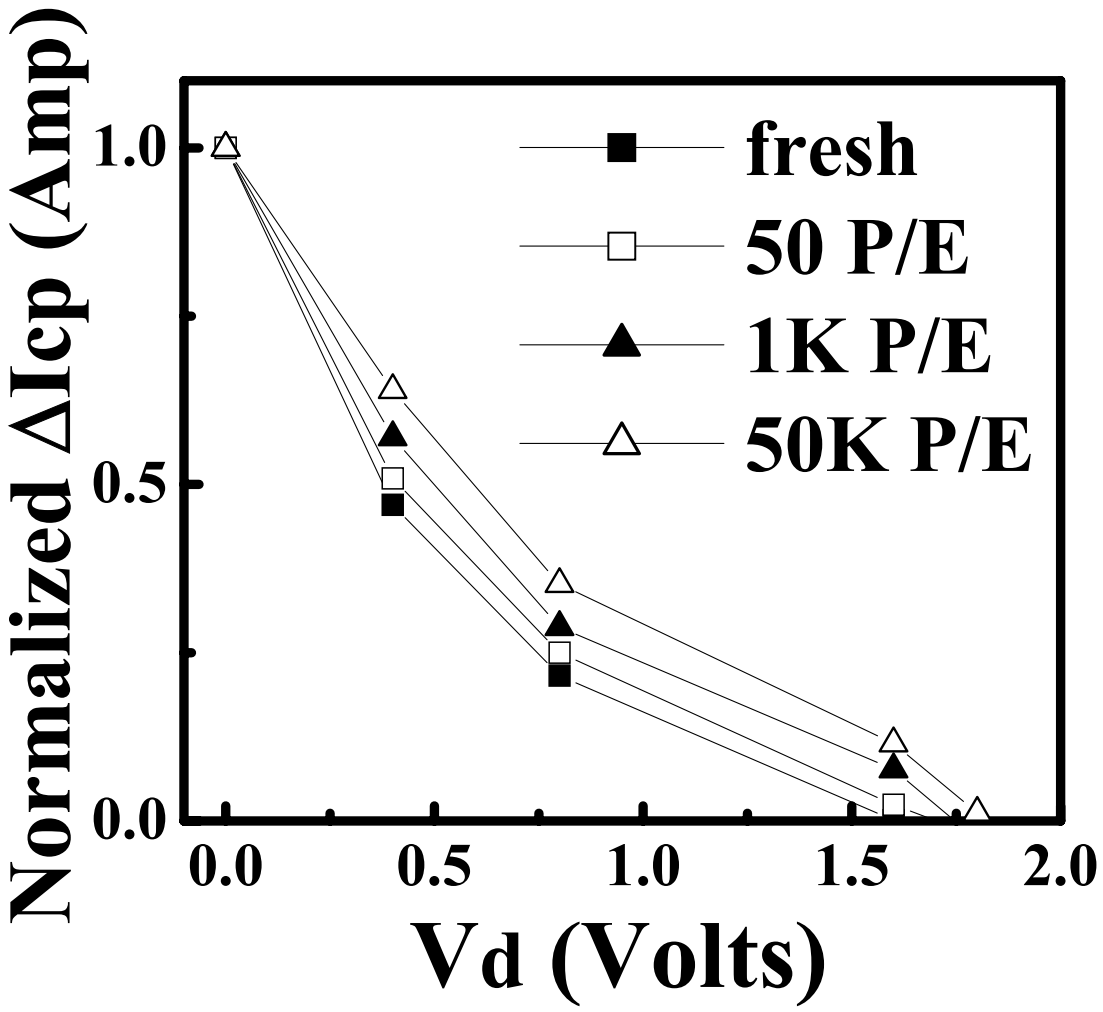


Fig. 3.13 The difference in Icp between program state and erase state as a function of V_d in CP measurement at various P/E cycle numbers. ΔIcp is measured at V_{gl}=1.6V and is normalized to its value at V_d=0V to take into account interface trap creation in cycling.

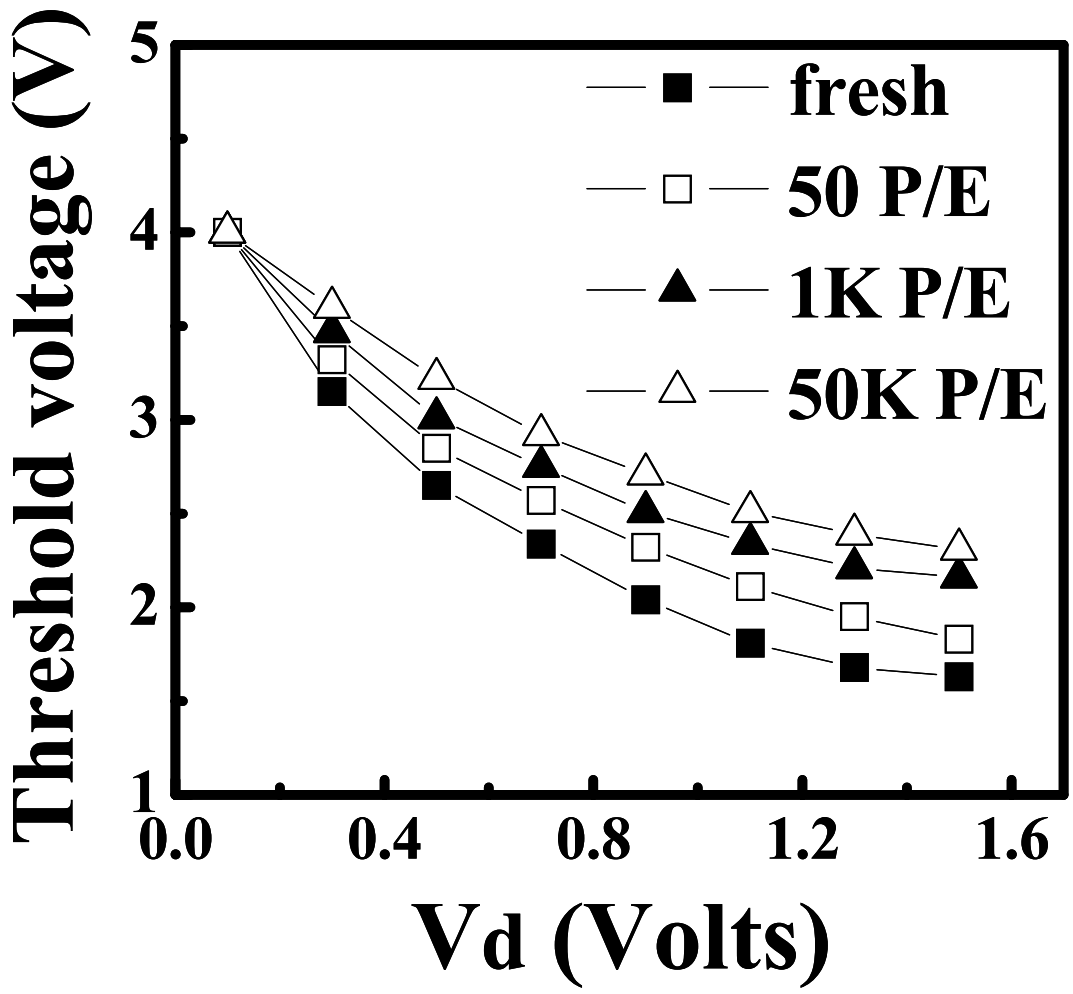


Fig. 3.14 Threshold voltage versus reverse read Vd for different cycle numbers.

Chapter 4

Reliability Mechanisms of Data Retention and Read-Disturb in Nbit Flash Memory Cells

4.1 Introduction

The major advantages of the Nbit cell, as compared with the conventional SONOS flash EEPROM, has a better retentivity due to a thick bottom oxide thickness [4.1]. The bottom oxide in the Nbit cell is normally thicker than 40Å. Such oxide is sufficiently thick to avoid charge direct tunneling [4.2]. The Nbit cell exhibits no window loss before P/E cycling. However, after P/E cycled stress, the retentivity behavior is degraded, showing a window closure with time (Fig. 4.1). As previously stated, the retention loss characteristics are determined by two factors. (a) oxide charge (Q_{ox}) de-trapping in erase state [4.3], strongly dependent on P/E stress and oxide quality [4.4], and (b) nitride charge (Q_N) loss in program state [4.5], closely related to temperature and electric field.

In this chapter, the reliability issues of the Nbit flash cell including low-Vt state threshold voltage instability (room temperature threshold voltage drift, RT drift), read-disturb, and high-Vt state charge loss will be reviewed. Responsible mechanisms and possible solution will be discussed.

4.2 Room-Temperature Threshold Voltage Drift

In a P/E stressed cell, the erase-state threshold is found to possess a positive drift with storage time (Fig. 4.1). This retention loss exhibits logarithmic

time-dependence but weak temperature dependence (Fig. 4.2). This is why the drift is referred to as RT drift [4.6]. Unlike a SONOS cell, the bottom oxide is sufficiently thick and thus this drift cannot be explained by nitride hole back tunneling [4.7]. Furthermore, we find that the Vt drift exhibits a peak around 10k P/E cycles in Fig. 4.3. To understand this peculiar cycle number dependence, the readers should be reminded that it is well published in literature that positive trapped charge creation is dominant in tunnel oxide in the initial period of P/E stress [4.8]. The appearance of the peak gives a clue that the Vt drift is related to positive charge creation in the bottom oxide. To explore the origin of the RT drift, Vt and GIDL techniques are used to monitor the charge, as shown in Fig.4.4 (a) and Fig.4.4 (b). From the change of Vt and GIDL, it can be deduced that after P/E stress the net ONO charge above the n^+ region is positive and the net ONO charge above the channel region is negative. More exactly speaking, the ONO charge in the channel region comprises positive oxide charge (Q_{ox}) and negative nitride charge (Q_N). In storage, trapped holes in the bottom oxide (Q_{ox}) can escape to the Si substrate with time. The total ONO charge in the channel region therefore becomes more negative and thus threshold voltage increases with time. To measure positive oxide charge de-trapping directly, the charge separation technique (Fig. 4.5(a)) is employed in large area devices with two different ONO processes (A and B). Process B is known to have a thinner bottom oxide, thereby causing a smaller Q_{ox}. The substrate current (I_b) before and after FN stress was measured in these two samples, as shown in Fig. 4.5(b). According to the hole tunneling front model [4.9], the post-stress substrate current resulting from positive oxide charge de-trapping follows a 1/t time-dependence,

$$I_b(t) = A\frac{Q_{ox}}{\alpha_h}t^{-1} \quad (4\text{-}1)$$

$$\alpha_h = 4\pi\sqrt{2m\phi_{ox}}\,/h \quad (4\text{-}2)$$

where Q_{ox} is the positive oxide charge density, ϕ_{ox} denotes the energy barrier of positive trapped charges and A is the area of the device. Note that process B exhibits a smaller post-stress substrate current because of less positive oxide charge creation. The corresponding Vt drift thus has a logarithmic time-dependence,

$$\Delta Vt(t) = 2.3\frac{1}{C_{ONO}}\frac{hQ_{ox}}{4\pi\sqrt{2m\phi_{ox}}}\log(t) \quad (4\text{-}3)$$

Fig. 4.6 shows the measured Vt drift in two 10k P/E cycled cells. By comparing the two ONO processes, a correlation between the Vt drift and I_b is obtained. Fig. 4.7 shows the RT drift versus bottom oxide thickness and demonstrates that an effective approach to reducing Q_{ox} is to use a thinner oxide. In addition, we can vary the erase time of the last-shot to study the dependence of the RT drift on Q_N. For a prolonged erase time, Q_N is less and the RT drift is smaller, as shown in Fig. 4.8.

4.3 Read-Disturb Effects in Erase State

Read-disturb effect is twofold in the NROM. The word-line voltage during read may enhance the RT drift in the neighboring bit. On the other side, the relatively large read bit-line voltage may cause channel hot electron injection and result in a significant threshold voltage shift of the neighboring bit. The hot electron injection caused Vt shift follows either power-law time-dependence or logarithmic

time-dependence. An analytical model based on positive oxide charge assisted channel hot electron injection is proposed to explain the observed power law time-dependence.

4.3.1 Commonality between Vt Drift and Read-Disturb

The RT drift and read-disturb have something in common. For example, the read-disturb caused Vt shift is also smaller when the bottom oxide thickness is reduced (Fig. 4.9). Secondly, we performed RT drift measurement and the read-disturb measurement in the same device (10k P/E cycles) sequentially. No matter the RT drift or the read-disturb is measured first, the subsequent read-disturb or RT drift is significantly reduced (Fig. 4.10). Fig. 4.10 gives strong evidence that the mechanisms of RT drift and read-disturb should share the same physical origin. From the study in the preceding section, we believe that read-disturb is also related to positive trapped charge in the bottom oxide.

4.3.2 Gate and Drain Bias Dependences on Read-Disturb Behavior

At read V_g, the channel is inverted and another current component flowing from the gate to the source and the drain (I_{sd}) arises due to positive oxide charge assisted electron tunneling (Fig. 4.11). When V_g is large, I_{sd} becomes dominant (Fig. 4.12). Our previous work has shown that Isd has a t^{-n} time-dependence with n~0.7 [4.10]. The gate-disturb induced $\Delta Vt \propto \int I_{sd}(t)dt$ should follow a power law time dependence with the power factor of 1-n~0.3. The gate-disturb effect and the RT Vt drift are compared in Fig. 4.13.

Moreover, since the bit-line voltage in reverse read must be sufficiently large to

overcome the stored charge of the second bit, hot electron injection during read should be considered. The hot electron read disturb is worsened in a P/E cycled cell because of positive oxide charge enhanced electron injection. Fig. 4.14 shows the Vt shift versus read bit-line voltage in a fresh device and in a 10k P/E cycled Nbit cell. The read disturb increases drastically as the read voltage is above 2V.

4.4 Program state Charge Loss

The most advantage of trapping storage cells, as compared to floating gate flash cells, is the better charge retentivity. Fig. 4.15 and Fig. 4.16 show that high-Vt state charge loss has temperature, cycle number and gate bias dependence, respectively. These observations suggest that the nitride charge escape is through thermionic-field emission and subsequently oxide trap assisted tunneling was proposed (Fig. 4.17). To characterize nitride charge escape current, a large area device was stressed at $-V_g$ (FN stress) and then was programmed to a high-Vt state by uniform FN injection. Trapped charge lateral migration in this case is excluded due to the uniform charge injection [4.11]. According to the FP emission model, the nitride trapped charge emission time is written below,

$$t = (A^* T^2) \exp((\phi_N - q(qE_N / \pi\varepsilon_N)^{1/2}) / kT) \qquad (4\text{-}4)$$

where A* is the Richardson constant, ϕ_N is the nitride trap energy, E_N is the electric field in nitride and other variables have their usual definition. The nitride charge emission current and the corresponding Vt shift can be derived as follows [4.1],

$$I_N = AQ_N \frac{d\phi_N}{dt} = \frac{AQ_N kT}{t} \quad (4\text{-}5)$$

$$\Delta V_t = \frac{Q_N}{C_{ONO}}[q(\frac{qE_N}{\pi\varepsilon_N})^{1/2} + kT\ln(\frac{t}{A^*T^2})] \quad (4\text{-}6)$$

where $Q_N(cm^{-2}eV^{-1})$ represents trapped charges in nitride. In the above derivation, we make the following assumptions. First, we assume that at measurement time t all nitride traps with time constants less than t are completely emptied and all other traps are unaffected. Secondly, we assume that nitride trapped charge escape is limited by the FP emission. Third, we assume that emitted nitride charges in the measurement interval from 1sec. to 100 sec. (Fig. 4.18) have a uniform distribution in trap energy, i.e., Q_N is a constant. This assumption is also reasonable since the trap energy span in the measurement period is only about kTln(100)~0.12eV. For a constant energy distribution of nitride-trapped charges, the nitride charge emission current obeys a 1/t relationship (Eq. (4-5) & Fig. 4.18) and the Vt loss is proportional to the square root of the electric field (Eq. (4-6) & Fig. 4.19).

4.5 Summary

In a low-Vt state, P/E stress created positive oxide charge plays a major role in various reliability issues. ONO process condition is critical to the improvement of the cell reliability. High-Vt state charge retention mechanism, FP excitation followed by oxide-trap assisted tunneling, is investigated. In next chapter, a detail numerical analysis for the program state charge loss is described.

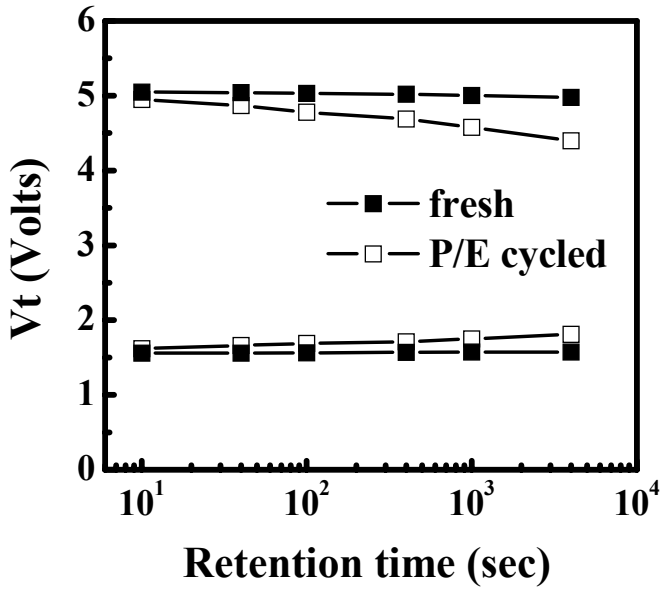


Fig. 4.1 Typical Vt retention characteristics in a fresh and a 100k P/E cycled cell.

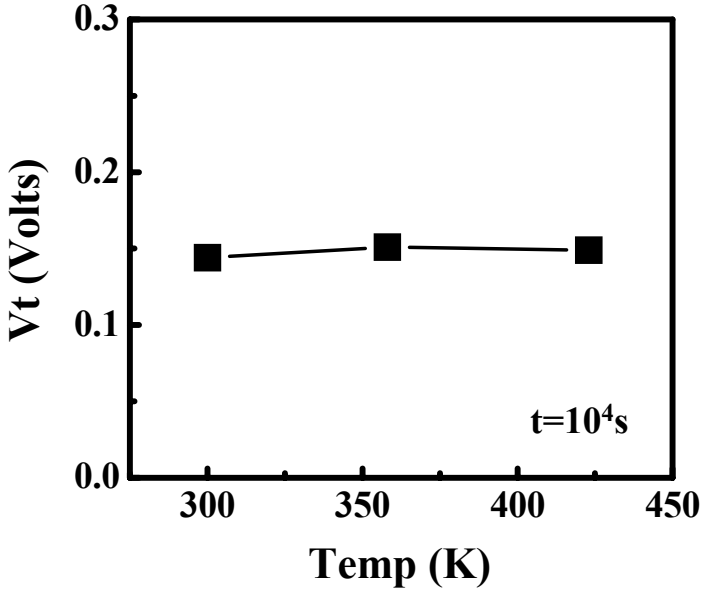


Fig. 4.2 Temperature-dependence of the erase-state Vt drift in a 10k P/E cycled cell at 10^4s. No significant change in Vt drift by varying temperature.

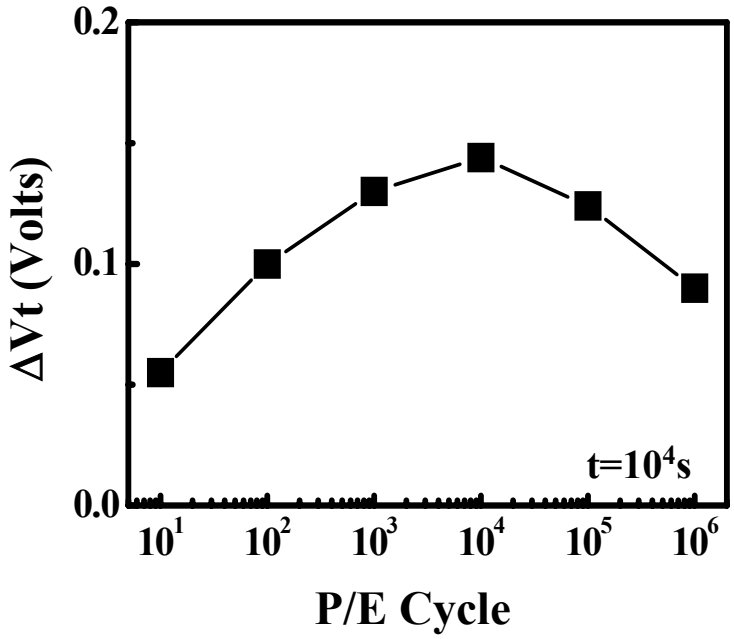
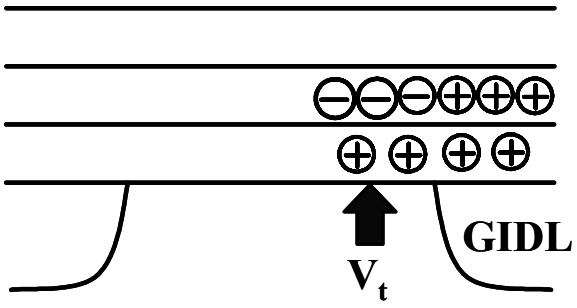


Fig. 4.3 Cycling number dependence of the erase-state Vt drift. The thickness of bottom oxide is 5nm. It shows that the peak is around 10k P/E cycles.

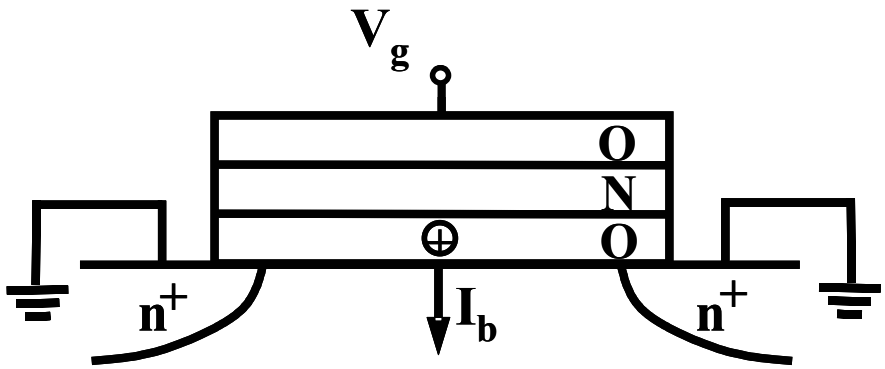


(a)

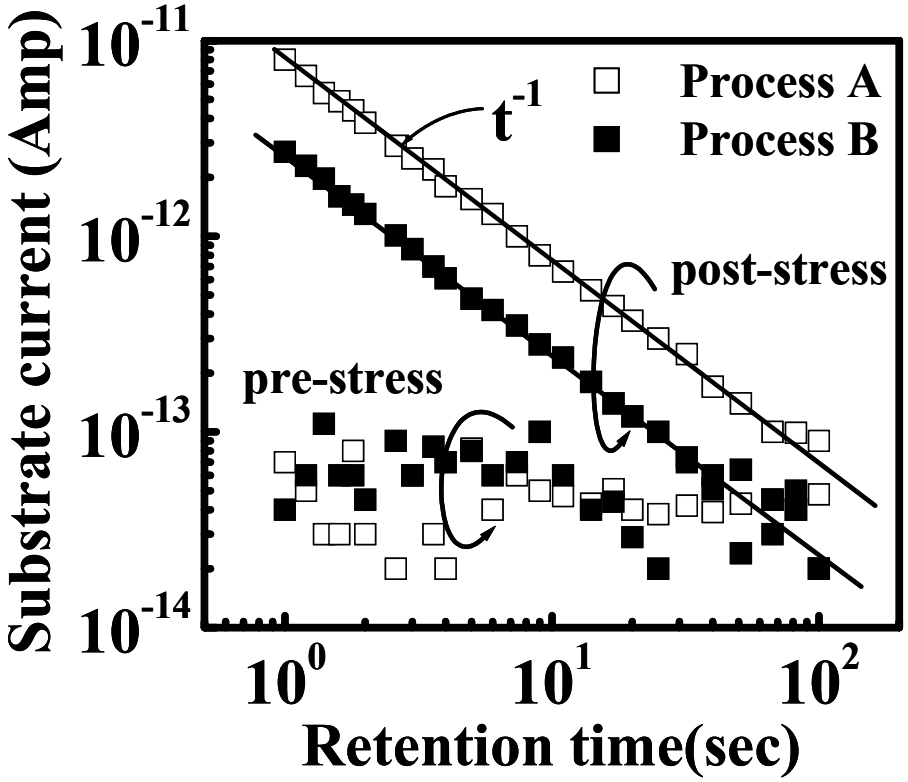
(b)

	Fresh	1k P/E	Net charge
V_t	1.6	1.74	negative
GIDL	28.9pA	21.4pA	positive

Fig. 4.4 (a) Illustration of charge distribution in a 10k P/E cell. (b) Measured Vt and GIDL in a fresh device and in a cycled device.



(a)



(b)

Fig. 4.5 (a) Measurement setup of positive oxide charge de-trapping induced substrate current (I_b). (b) Pre-stress and post-stress substrate currents in two large area devices (500μm×500μm). FN stress was at V_g=-18V for 3000s. Substrate current was measured at $V_g=V_{FB}$.

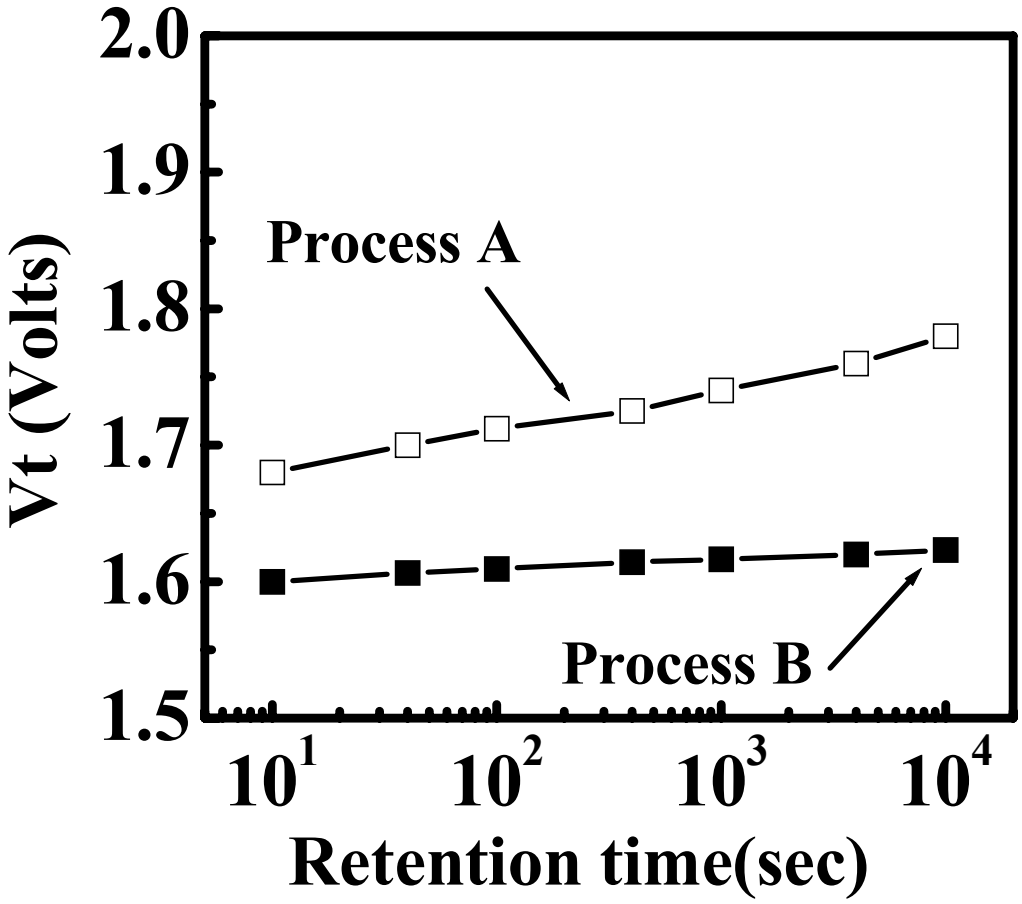


Fig. 4.6 Room temperature Vt drift in two 10k P/E cycled NROM cells fabricated with different ONO process. The cell size is L=0.5μm and W=0.35μm.

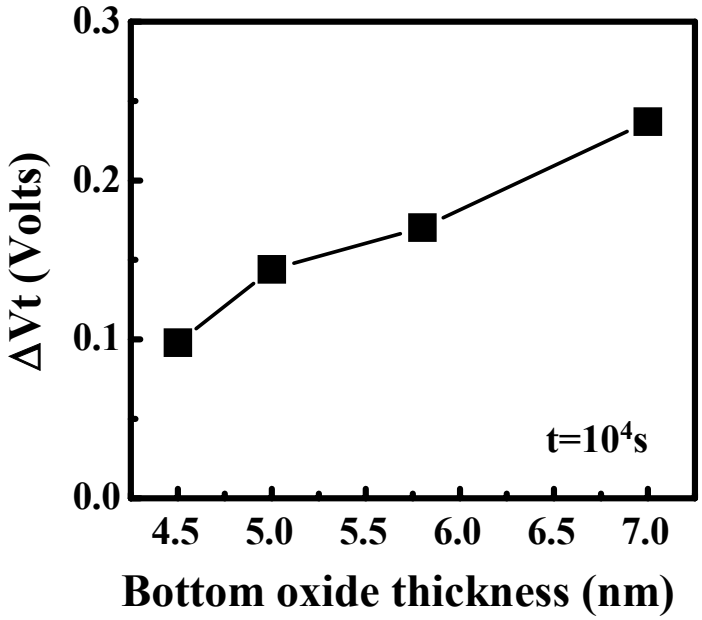


Fig. 4.7 RT drift versus bottom oxide thickness in a 10k P/E cycled device. Retention time is 10^4s.

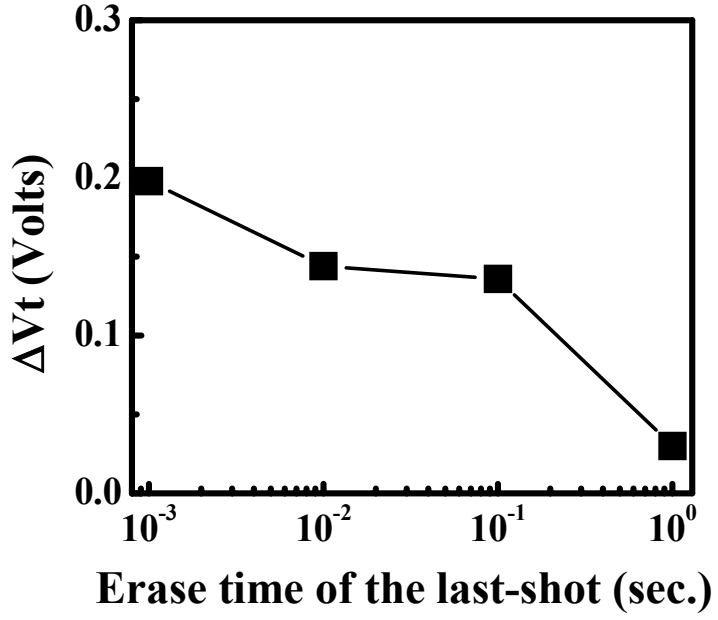


Fig. 4.8 Dependence of RT drift on erase time in a 10k P/E cycled cell.

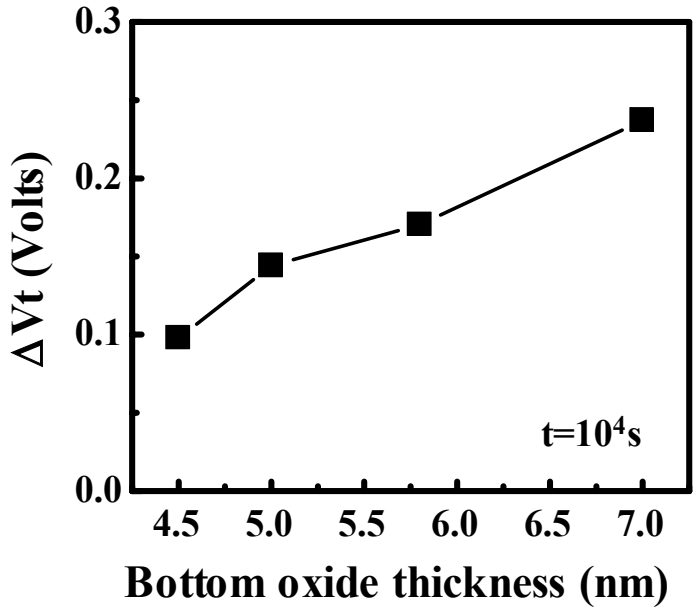


Fig. 4.9 Read-disturb caused Vt shift as a function of bottom oxide thickness in a 10k P/E cycled cell. The read bias condition is V_g=2.75V, V_d=1.6V and read-disturb time is 10^4 sec.

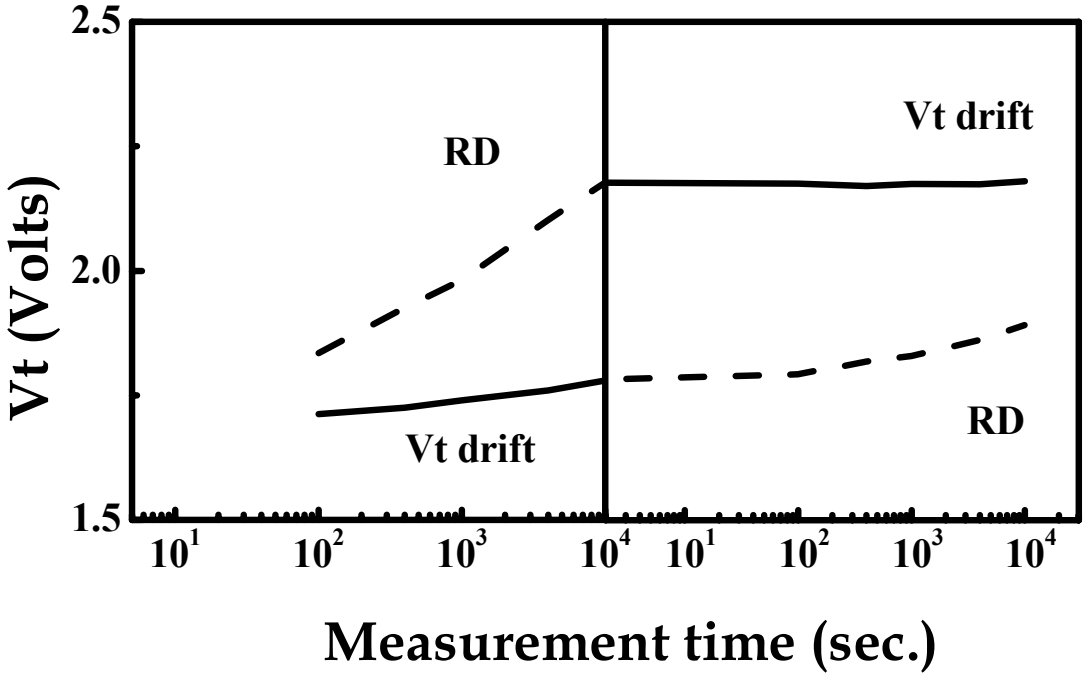


Fig. 4.10 Temporal evolutions of Vt drift and read-disturb caused Vt shift. These two measurements are performed sequentially.

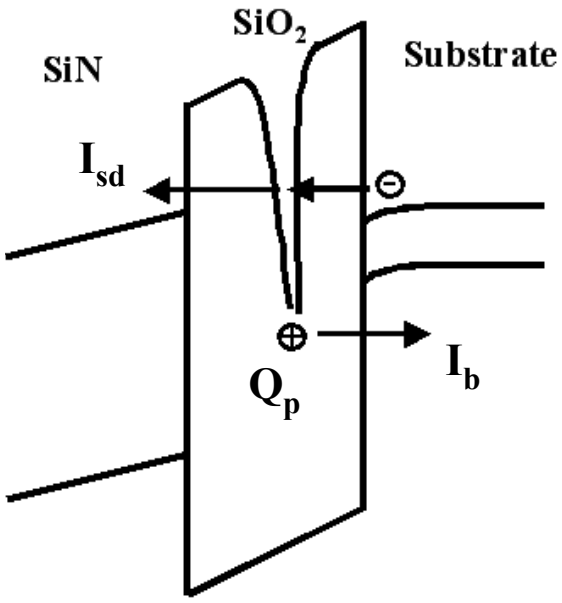


Fig. 4.11 Illustration of positive oxide charge assisted electron tunneling current (I_{sd}) at gate disturb. The time-dependence of Isd and corresponding Vt shift was derived in [4.1].

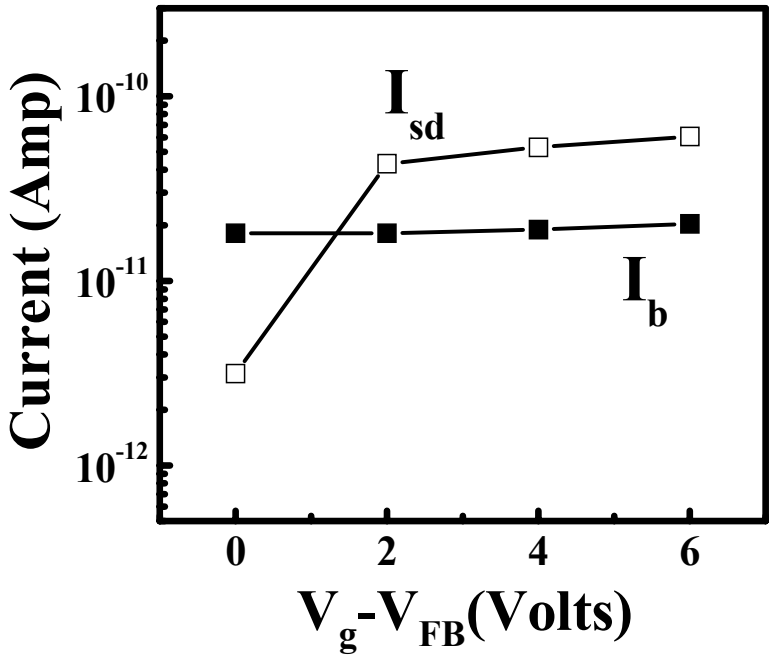


Fig. 4.12 Gate bias dependence of positive oxide de-trapping current (I_b) and electron tunneling current (I_{sd}). The current is measured at t=0.6sec after V_g is applied.

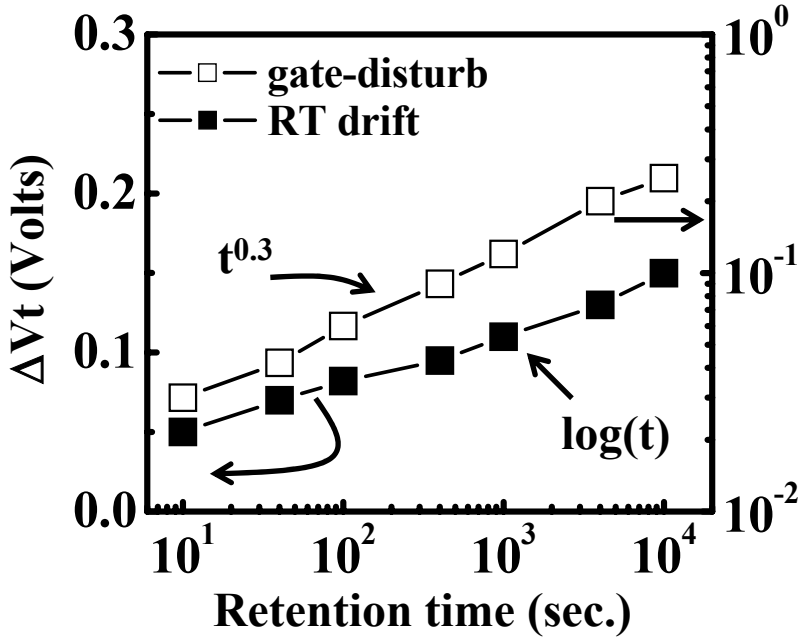


Fig. 4.13 Hot carrier read-disturb caused ΔVt in a fresh cell and in a 10k P/E cycled cell. The disturb time is 10^4s.

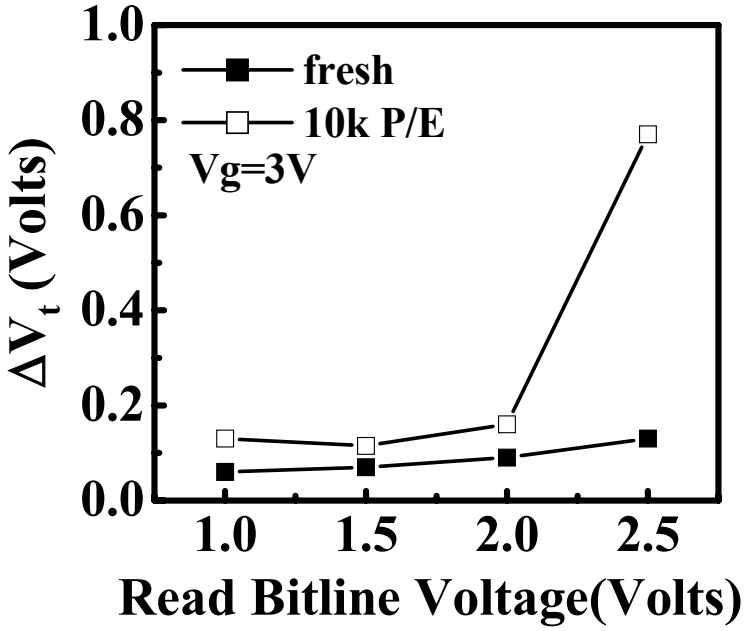


Fig. 4.14 Comparision of gate disturb caused ΔVt (V_g=3V, V_d=V_s=0V) and room temperature Vt drift (V_g=V_d=V_s=0V)

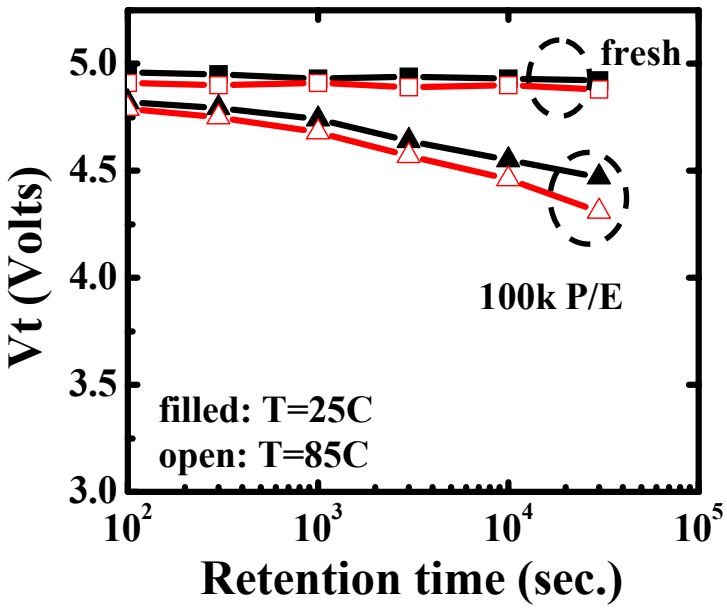


Fig. 4.15 Program-state charge loss in a fresh and in a 100k P/E cycled cells. T=25C and 85C.

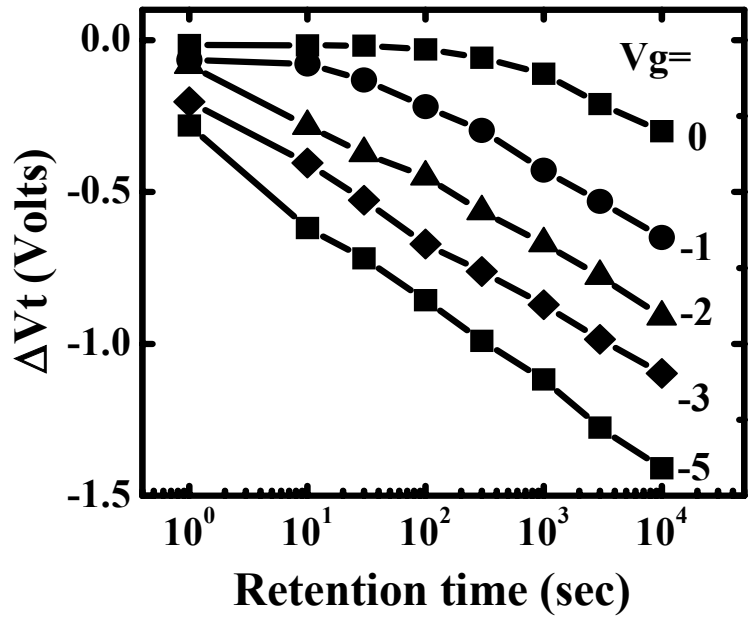


Fig. 4.16 Program-state charge loss at different applied gate bias. T=25C

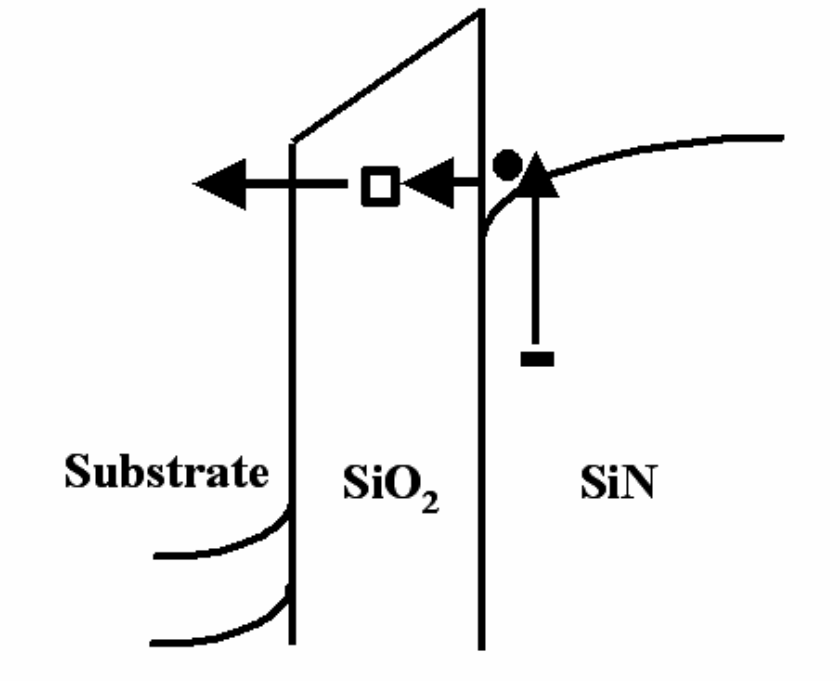


Fig. 4.17 Illustration of nitride trapped electron emission and oxide trap assisted tunneling.

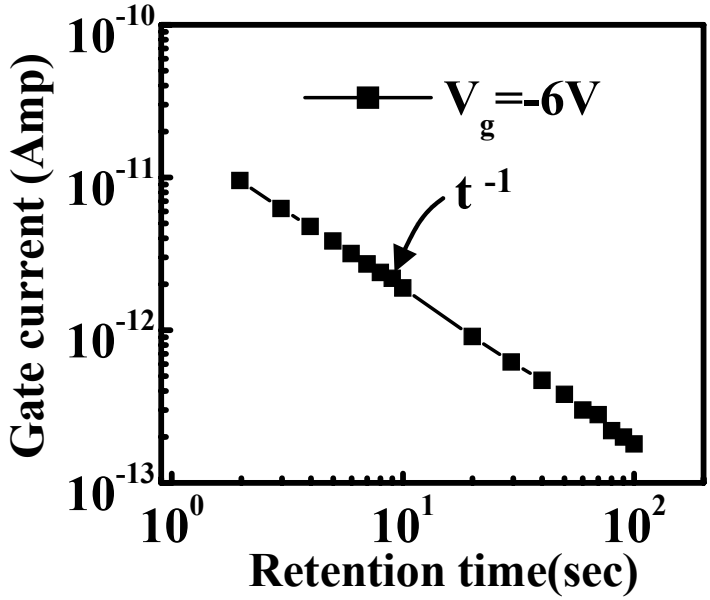


Fig. 4.18 Measured nitride charge de-trapping current (I_N) in a large area device at V_g=-6V. The device was subjected to FN stress at V_g=-20V for 2000s.

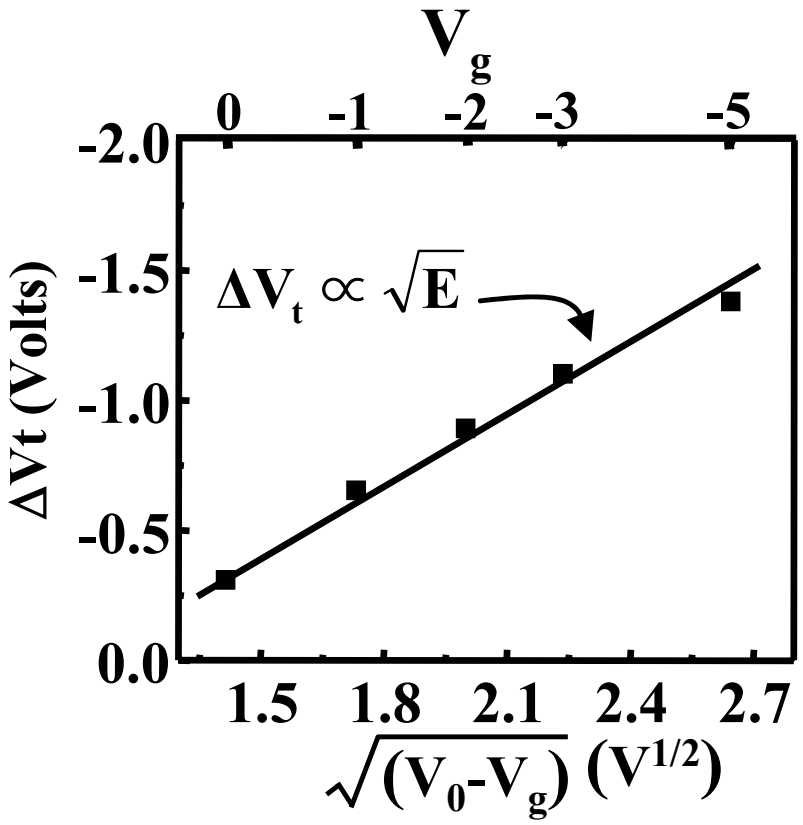


Fig. 4.19 Program-state Vt loss versus applied gate bias (top axis). The bottom axis corresponds to the square root of the nitride electric field. Vo is the flat-band voltage in high-Vt state.

Chapter 5

Numerical Simulation for Program-State Charge Retention in SONOS Flash Memory Cell

5.1 Introduction

Recently, considerable research efforts have been made to study the charge retention loss mechanisms in SONOS devices. Lundkvist et al. showed that trapped charge direct tunneling leads to charge loss at room temperature [5.1] and later they accounted for the increased decay rate of charge loss at elevated temperatures with a thermal-enhanced charge emission model [5.2][5.3]. Lehovec et al. delivered a simple analytical retention model through the Frenkel-Poole (FP) release of electrons from mono-energetic nitride traps [5.4]. Williams et al. regarded charge loss as direct tunneling of the charges out of the nitride together with nitride charge migration by a series of emission-capture events [5.5]. Recently, White et al. have combined trap-to-band tunneling and thermal excitation to reproduce measured device retention data [5.6][5.7]. Although a variety of models have been proposed, they have some deficiencies in common. For example, all the models are applicable only to ultra-thin bottom oxides (1.5nm~2.5nm). However, to improve data retention and to minimize gate disturb, a thicker bottom oxide is usually employed in today's SONOS cells [5.8][5.9]. With a larger oxide thickness, oxide traps created by program/erase (P/E) stress becomes important and should be taken into account in a charge loss model. Second, several groups fitted their experimental results with unphysical attempt-to-escape factors [5.4][5.5][5.10]. The values ranging from 10^6 s^{-1} to 5×10^8 s^{-1}

are extremely low and should be modified by Shockley-Read theory [5.11][5.12].

In this chapter, we develop a numerical approach to solve a set of rate equations governing the electron FP emission and re-trapping in the nitride and charge leakage through the bottom oxide [5.13]-[5.17]. Comprehensive charge loss mechanisms based on direct tunneling for trapped electrons and positive charge (oxide hole trap) assisted tunneling (PCAT) [5.13] for conduction band electrons in nitride are formulated. The effects of bottom oxide thickness and stress created oxide traps on charge retention are measured and simulated. The dominant leakage mechanisms for different bottom oxide thickness are identified.

5.2 Retention Loss Simulation Model

The samples used in this work are n-channel SONOS cells consisting of a top oxide of T_{TO}=9nm, a silicon nitride of T_N=6nm. The bottom oxide thickness (T_{BO}) ranges from 1.8nm to 5nm. The measurement data for the Tbo=1.8nm device are quoted from [5.18] and other oxide thickness results are measured in this work. FN tunneling is employed for programming and erase. This uniform FN injection excludes the possibility of stored charge lateral migration in the nitride.

5.2.1 Nitride Charge Dynamics and Loss Mechanisms

Charge transitions between the conduction band and trap states in the nitride and stored charge leakage paths are illustrated in the energy band diagram in Fig. 5.1. The change of conduction band and trapped electron densities in the nitride is described by the following two coupled rate equations [5.14]-[5.16]:

$$\frac{dn_t(y,\phi_N,t)}{dt} = R_c(y,\phi_N,t)\cdot n_c(t) - R_e(\phi_N)\cdot n_t(y,\phi_N,t) - R_t(y,\phi_N)\cdot n_t(y,\phi_N,t) \quad (5\text{-}1)$$

$$\frac{dn_c(t)}{dt} = \frac{1}{T_N}\iint [R_e(\phi_N)\cdot n_t(y,\phi_N,t) - R_c(y,\phi_N,t)\cdot n_c(t)]dyd\phi_N - (R_{PCAT}+R_{DT})\cdot n_c(t) \quad (5\text{-}2)$$

where $n_t(y,\phi_N,t)$ ($cm^{-3}eV^{-1}$) is the density of occupied nitride traps as a function of a distance to the SiN/bottom oxide interface (y), trap energy ϕ_N, and retention time t, respectively. The density of electrons in the nitride conduction band is denoted by $n_c(cm^{-3})$, which is assumed to be uniform in x because the nitride layer is sufficiently thin. R_e, R_c, R_t, R_{PCAT}, and R_{DT} are the rate coefficients for FP electron emission from nitride traps (path a in Fig. 5.1), free electrons capture into nitride traps (path b), trapped electron direct tunneling to the substrate (path c), conduction band electron escape through PCAT (path d) and via direct tunneling (path e), respectively. The FP excitation of electrons from nitride traps to the conduction band is [5.4]

$$R_e(\phi_N) = \nu_e \cdot \exp\left(\frac{\beta\sqrt{E_n} - \phi_N}{kT}\right) \quad \text{(unit: } s^{-1}\text{)} \quad (5\text{-}3)$$

where

$$\nu_e = Nc_n \cdot \upsilon_{th} \cdot \sigma_N . \quad \text{(unit: } s^{-1}\text{)} \quad (5\text{-}3a)$$

$$\tau_e = R_e^{-1} \quad \text{(unit: s)} \quad (5\text{-}3b)$$

The pre-factor ν_e is often referred to as the attempt-to-escape" frequency for emission and can be expressed as Eq. (5-3a) where Ncn is the effective density of states in the nitride conduction band, υ_{th} is the thermal velocity and σ_N is the nitride trap capture cross section. β is the FP constant and En is the average electric field in the nitride. $\tau_e(\phi)$ is the electron emission time from the trap energy of ϕ_N. The free electron capture rate coefficient is

$$R_c(y,\phi_N,t) = \upsilon_{th} \cdot \sigma_N \cdot (N_t - n_t) \qquad \text{(unit: } s^{-1}eV^{-1}\text{)} \qquad (5\text{-}4)$$

where Nt is the nitride trap density per unit trap energy. $(N_t - n_t)$ is the amount of available traps for free electron re-capture. The direct tunneling rate for nitride-trapped electron is

$$R_t(y,\phi_N) = \nu_t \cdot \exp(-\alpha_{ox} T_{BO}) \cdot \exp(-\alpha_N y) \qquad \text{(unit: } s^{-1}\text{)} \qquad (5\text{-}5)$$

$$\alpha_{ox} = \frac{2\sqrt{2m_{ox}^* q(\Phi_B + \phi_N)}}{\eta}; \alpha_n = \frac{2\sqrt{2m_N^* q\phi_N}}{\eta} \qquad (5.5a)$$

where m_{ox}^* and m_N^* are electron tunneling mass in the oxide and in the nitride. Other variables have their usually definitions.

As for conduction band electron leakage paths, positive oxide charge assisted tunneling [5.13] and direct tunneling through the bottom oxide are considered. Charge leakage via top oxide is not considered for a relatively large top oxide thickness. The built-in electric field in the ONO stack is neglected for simplification.

Electron tunneling from nitride traps to surface traps in Si band gap is also neglected. To simplify the evaluation of PCAT, a concept of the most favorable trapped charge position is employed [5.19][5.20]. With an oxide charge site at this favorable position, the tunneling probability from the conduction band to the trap site is the same as that from the trap site to the substrate conduction band. Such feature of the most efficient tunneling trap location leads to the following equation

$$R_{PCAT} = N_{ox} \cdot \upsilon_{th} \cdot \sigma_{ox} \cdot P_{PCAT} \qquad \text{(unit: } s^{-1}\text{)} \qquad (5\text{-}6)$$

$$\tau_{PCAT} = R_{PCAT}^{-1} \qquad \text{(unit: s)} \qquad (5\text{-}6a)$$

where N_{ox} is the stress-induced oxide hole trap (positively charged) density. P_{PCAT} is the tunneling probability from the nitride conduction band to the substrate. τ_{PCAT} is the electron tunneling time via PCAT. The way to calculate the tunneling probability P_{PCAT} is the same as in [5.13]. A one-dimensional Coulombic potential caused by a positive trapped charge is included in the electron tunneling barrier for the WKB approximation, i.e.,

$$\Phi_{coul}(y) = \frac{q}{4\pi\varepsilon_{ox}|y - y_h|} \qquad (5\text{-}7)$$

where ε_{ox} is permittivity of oxide and y_h denotes the location of a trapped positive charge from SiN/bottom oxide interface. Besides, the direct tunneling rate can be deduced from the supply function,

$$R_{DT} = \frac{\upsilon_{th}}{T_N} \cdot \exp(-\alpha_{ox} T_{BO}) \qquad \text{(unit: } s^{-1}\text{)} \qquad (5\text{-}8)$$

with α_{ox} (Eq. 5-5a) evaluated at the conduction band edge. With the above rate coefficients determined, the nitride charge leakage current J_N can be written as

$$J_N = q\iint R_t(y,\phi_N,t) \cdot n_t(y,\phi_N,t)dxd\phi_N + q \cdot R_{PCAT} \cdot n_c(t) \cdot T_N + q \cdot R_{DT} \cdot n_c(t) \cdot T_N \quad (5\text{-}9)$$

The first term on the right-hand side is the ensemble of out-tunneling electrons from nitride traps. The second and the third terms represent charge loss from the conduction band via PCAT and DT. It should be mentioned that in a P/E stressed device with a thicker oxide, direct tunneling is much smaller than PCAT and Eq. (5-9) reduces to

$$J_N = q \cdot R_{PCAT} \cdot n_c(t) \cdot T_N . \qquad (5\text{-}10)$$

5.2.2 Assumptions in Simulation

Several assumptions concerning the nitride traps in our model should be mentioned. First, we assume that the nitride traps have a continuous distribution in the bandgap and the trap density decreases exponentially with trap energy, i.e., $N_t(\phi)=N_{to}\cdot\exp(-\phi/\lambda)$, where N_{to} is set to be 3.0×10^{19} $cm^{-3}eV^{-1}$ and λ is the tail parameter (0.85eV). This exponential trap energy distribution is often used for

amorphous materials [5.14][5.21][5.22]. Second, the nitride traps below the Fermi-level are occupied in a fresh device in thermal equilibrium condition and thus are not available for the storage of injected charges. Third, nitride traps are considered to be isolated from each other and direct electron transitions among nitride traps are not taken into account.

5.3 Numerical Simulation Method

5.3.1 Explicit and Implicit Methods

Several distinct numerical methods have been applied to solve the time-dependence rate equation. One of the most applicable methods is "explicit" forward Euler [5.23]. A concept of explicit is explained by a simple ordinary differential equation.

$$\frac{dy}{dt} = -\kappa y \tag{5-11}$$

Eq. (5-11) can be transferred to a first order Taylor expression.

$$\frac{y(t+\Delta t) - y(t)}{\Delta t} = -\kappa y(t)$$
$$\rightarrow y(t+\Delta t) = y(t)\cdot(1-\kappa\Delta t) \tag{5-11a}$$

In the explicit method, the slope at the time t is employed to determine the increment to the function. The advantage of explicit method is a direct computation without solving a complex iteration. However, this method always has a strict

limitation on time step size for stability. If the time step, Δt, is too large, the solution of $y(t+\Delta t)$ cannot converge toward the steady solution, exhibited in Fig. 5.2(a). Therefore, another "implicit" backward Euler method which permits the larger time step is required. It uses the unknown information at the time $t+\Delta t$ to calculate the solution as formula following.

$$\frac{y(t+\Delta t)-y(t)}{\Delta t} = -\kappa y(t+\Delta t)$$
$$\rightarrow y(t+\Delta t) = \frac{y(t)}{1+\kappa\Delta t} \qquad \text{(5-11b)}$$

In Fig 5.2(b), the implicit concept is demonstrated and there is no divergent solution compared with Fig 5.2(a). The only concern is that to choose an appropriate time step to reach the precise solution.

5.3.2 Non-linear and Stiff System

Here, some nomenclatures should be introduced. First, the differential equations in our case are non-linear, containing products of variables. To attain the solution, the Jacobian linearization for the non-linear system is essential.

In the amorphous material (silicon nitride), the eigen-values distribute much broadly. Such system is called a "stiff" system of equations. The emission time varying over many orders of magnitude is very short or long compared to the capture as well as tunneling time. Standard 'explicit' scheme is not feasible because only the time steps shorter than the critical time constant are used. To realize the simulation for long-term retention behavior, the time step must be adjustable. For

this reason, implicit Euler scheme is chosen as a good way to deal with the stiff problem.

5.3.3 Jacobian Matrix

Expecting the use of numerical program, the continuous distribution of nitride trap in space and in energy is replaced with discrete approximation. In our numerical calculation, the nitride traps are discretized into equi-distance segments in energy and in space, as shown in Fig. 5.3, transforming the differential equations into numerous sets of mutually related equations. After discretizating, implicit backward Euler method coupled with Newton iteration is selected to provide the required stability for large time step lengths, and to reach the steady state solution if the initial guess is reasonable. Prior to the introduction of the simulation procedure, the carrier density variables y(1) and y(n) are labeled, where

y(1)=n_c, the conduction band electron density

y(n)=the trapped electron density at each given site ($2 \le n \le M+1$)

M is the number of trap site and the corresponding (M+1) coupled equations for derivative function f(y) in vector form are listed below,

$$\frac{dy(n)}{dt} = f(y(n)) = \begin{bmatrix} f(y(1)) \\ f(f(2)) \\ M \\ f(y(M+1)) \end{bmatrix} \qquad (5\text{-}12c)$$

where

$$f(y(1)) = \frac{-\sum_{2}^{M+1} R_c(y(n)) \cdot y(1)}{T_n} dx + \frac{\sum_{2}^{M+1} R_e(n) \cdot y(n)}{T_n} dx - (R_{PCAT} + R_{DT}) \cdot y(1) \qquad (5\text{-}13a)$$

$$f(y(2)) = R_c(y(2)) \cdot y(1) - (R_e(2) + R_t(2)) \cdot y(2)$$

$$\text{N} \qquad (5\text{-}13b)$$

$$f(y(M)) = R_c(y(M+1)) \cdot y(1) - (R_e(M+1) + R_t(M+1)) \cdot y(M)$$

Now, the charge density y at time step number k+1 is denoted by y_{k+1}, Similarly, y_k is the value at time step k. The implicit first order equation is based on the Eq. (5-14)

$$y_{k+1} = y_k + f_{k+1} \cdot \Delta t \qquad (5.14)$$

where f_{k+1}, denoting $f(y_{k+1}(n))$, the derivative function vector at the end of the time step, is unknown, but can be found by iteration. By performing a Taylor's expansion of f_{k+1}, and substituting in Eq. (5-14), the iteration scheme is set up, giving for the (s+1) iterate on the (k+1) step,

$$(I - \Delta t \cdot J^s_{k+1}) \cdot \Delta y = \Delta t \cdot f^s_{k+1} - y^s_{k+1} + y_k \qquad (5.15)$$

where

$$\Delta y = y^{s+1}_{k+1} - y^s_{k+1} \qquad (5.15a)$$

I is the (M+1) (M+1) identity matrix and J is the (M+1) (M+1) Jacobian matrix of partial derivatives with elements

$$J_{ij})^{s}_{k+1} = \frac{\partial f_i}{\partial y_j})^{s}_{k+1} \qquad (5.15b)$$

The iteration proceeds by solving for Δy and updating Δy^{s}_{k+1} to Δy^{s+1}_{k+1} using Eq. (5-15a), (and hence updating f^{s}_{k+1} and J^{s}_{k+1}) repeatedly until the computed 'correction' Δy, is judged to be sufficiently small as the solution converges on the (k+1) time-step value, y_{k+1}.

It is useful to consider the structure of the Jacobian Matrix in detail, since an effortless simulation and acceptable computer memory is allowed. Fig. 5.4 illustrates the matrix structure. J_{11} is the summation of conduction electrons loss, J_{1n} is the top row of matrix, J_{n1} is electron trapping for each mesh and J_{nn} is the diagonal term; others are zero. These zero terms mean that the communication between each localized state does not occur.

5.3.4 Simulation Flowchart

Fig 5.5 shows our simulation flowchart for charge loss process. The parameters used in our model are given in Table 5.1. At the outset, the data of the device structure condition, including the discretized mesh size, ONO stack thickness, threshold voltage window, and gate voltage bias, are read. The total amount of stored electrons in the nitride is determined by measured Vt shift after programming.

The initial electron distribution in the nitride layer is obtained as the injected charges attain the dynamic balancing between the nitride conduction band and the trap states by setting all leakage paths equal to zero. Once the initial condition is determined, the conduction band electron and trapped electron densities at the subsequent instant can be evaluated step by step.

5.4 Results and Discussion

5.4.1 Bottom Oxide Blocking Effect

To investigate the bottom oxide thickness effect, the pre-cycling retention characteristics of four SONOS devices with different T_{BO} (1.8nm, 2.5nm, 3nm and 5nm) are compared in Fig. 5.6. All the devices are programmed to have an identical threshold voltage window of 1V. Good agreement between the measured (solid lines) and simulated data (symbols) confirms the validity of our simulation model. The program state Vt (for example, the 3nm oxide cell) remains almost unchanged for a certain period of time and then decreases with a log(t) dependence. The charge loss is retarded with increasing bottom oxide thickness. This feature can be well explained by a tunneling front model [5.24]. The corner time (τ_{COR}) in Fig. 5.6 corresponds to trapped electron direct tunneling time (process (c) in Fig. 1). The role of high-voltage stress created oxide traps in charge retention is examined in Fig. 5.7 for Tbo=5nm and 2.5nm. The two devices are subjected to a negative FN stress. As opposed to the thinner oxide (2.5nm) device, the 5nm oxide cell exhibits an apparent stress effect. The post-stress Vt loss in the 5nm oxide device exhibits an initial delay and then declines with a log(t) time-dependence. Our calculation in the next section will show that the dominant retention loss mechanism in the post-stress 5nm oxide cell is the

Frenkel-Poole emission followed by PCAT.

5.4.2 Post-Stress Two-Stage Retention Loss

To explain the above post-stress retention loss behavior in the 5nm oxide device, we measure the stress induced leakage current (i.e., Eq. (5-9)) in a large area device (500μm×500μm) directly. The programming window is 3V and the measurement gate voltage is 0V. The stress induced leakage current (gate current) and the evolution of the retention Vt are shown in Fig. 5.8(a) and 5.8(b). One of the samples in Fig. 5.8(a) is stressed lightly and the other is stressed more heavily. The symbols in Fig. 5.8 denote measured data and the solid lines represent simulation result. In the simulation, we use the positive trapped charge (hole trap) density as a fitting parameter. The hole trap density is $4.5\times10^{17}cm^{-3}$ for the lightly-stressed sample and $3\times10^{18}cm^{-3}$ for the heavily-stressed sample. The temperature dependence of Vt retention loss is shown in Fig. 5.9 for T=25C and 85C. The higher bake temperature yields a larger Vt loss, which is in agreement with the Frenkel-Poole model. Trap anneal effect should be considered as bake temperature further increases. It should be noted that the nitride charge leakage current exhibits two stages. In the first stage ($t<\tau_{COR}$), a DC-like characteristic, termed as current blocking effect, is observed. The current blocking effect persists for a longer time in the lightly-stressed sample. After τ_{COR}, a 1/t transient decay in the leakage current is observed, which accounts for the log(t) dependence of the Vt loss in Fig. 5.8(b). The evolution of nitride conduction band electron density (n_c) with retention time is plotted in Fig. 5.10. The pre-stress one is also shown as a reference. In the first stage, n_c remains almost the same as the pre-stress value because the oxide leakage current is limited by the amount of stress

created oxide traps. The conduction band electron density begins to have an apparent decrease in the second stage. The simulated nitride trap electron occupation factor (f_t) versus trap energy at different retention times are shown in Fig. 5.11. The FP emission front (ϕ_{fro}), defined as the trap energy with f_t=0.5, is displayed in Fig. 5.12 as a function of retention time. In the first stage, since the nitride charge leakage current is limited by oxide trap assisted tunneling, the FP emission front stays around 0.8eV. In the second stage, the oxide trap assisted tunneling is no longer a limiting mechanism and the leakage current is dictated by the Frenkel-Poole emission. As a consequence, the FP emission front moves downward in the nitride bandgap with a constant speed in a log(t) scale. For the heavily-stressed sample, the oxide leakage current is higher in the first stage and the corner time τ_{COR} is shorter. To further elaborate on the τ_{COR}, we use a simplified picture by assuming that the nitride charge transitions mainly occur between conduction band and the states in the vicinity of the FP emission front (in an energy range of ~kT) as illustrated in Fig. 5.13. $\tau_e(\phi_{fro})$, and $\tau_c(\phi_{fro})$ represent electron emission time and capture time between the conduction band and the emission front. It should be pointed out that in our measurement period τ_{PCAT} is much longer than $\tau_c(\phi_{fro})$ that the conduction band electron density is mainly determined by τ_e, τ_c, and the trapped charge density at the emission front $n_t(\phi_{fro})$. Therefore, the nitride charge leakage current in the second stage can be readily obtained in the following [5.22].

$$J \propto \frac{n_c}{\tau_{PCAT}} \propto \frac{n_t(\phi_{fro})}{\tau_{PCAT}} \cdot \frac{\tau_c(\phi_{fro})}{\tau_e(\phi_{fro}) + \tau_c(\phi_{fro})} = \frac{n_t(\phi_{fro})}{\tau_{eff}} \tag{5.18}$$

and

$$\tau_{eff} = \frac{\tau_e(\phi_{fro}) + \tau_c(\phi_{fro})}{\tau_c(\phi_{fro})} \cdot \tau_{PCAT} \qquad (5.18a)$$

where τ_{eff} is the effective time for nitride trapped charge at the emission front to escape from the ONO film. Fig. 5.14 shows our calculated τ_{eff} versus retention time. Three points are worth noting; (i) Since $n_t(\phi)$ has a pretty uniform distribution in our measurement period, J (Fig. 5.8(a)) and τ_{eff} (Fig. 5.14) have an inverse time-dependence from Eq. (5-18). (ii) The corner time τ_{COR} in Fig. 5.7&5.8 is equal to τ_{PCAT} multiplied by a factor $(\tau_e+\tau_c)/\tau_c$, which is determined by the initial FP emission front. Fig. 5.15 shows the calculated τ_{COR} and τ_{PCAT} versus Nox. For a larger program-state Vt, the initial ϕ_{fro} is smaller and so is the factor $(\tau_e+\tau_c)/\tau_c$. Thus, the τ_{COR} becomes shorter as the program Vt window increases. This feature is consistent with our measurement result. It also should be mentioned that the neglect of conduction band electron recapture (i.e., $\tau_c = \infty$) would lead to $(\tau_e+\tau_c)/\tau_c=1$ and an erroneous interpretation of $\tau_{COR}=\tau_{PCAT}$. (iii) At a sufficiently long retention time, Fig. 5.14 shows that τ_{eff} is actually equal to retention time, t. Thus, Eq. (5-18) becomes $J \propto n_t/t$ without regard to stress condition. This result is confirmed by our measurement in Fig. 5.8(a) that both samples (heavily-stressed and lightly-stressed) have the same leakage current in the second stage. A similar expression for the nitride leakage current in the second stage was published in [5.25].

5.4.3 Leakage Component Separation

In above discussion, the role of PCAT in charge loss has been substantiated for T_{bo}=5nm. On the other side, trapped charge direct tunneling is the dominant loss mechanism in ultra-thin oxides. It is interesting to investigate the transition between these two mechanisms with respect to oxide thickness. Since they are difficult to be separated experimentally, a numerical approach is performed. Positive oxide charges are placed at the most efficient PCAT position for different oxide thickness. Fig. 5.16 shows the percentage of the Vt loss from the two components versus oxide thickness at T=25C. For thinner oxides (<3nm), nitride charge loss is mainly via direct tunneling. The PCAT becomes dominant for oxide thickness exceeding 3.7nm.

5.5 Summary

A numerical method to simulate the retention characteristics in SONOS flash memories has been developed. The dominant Vt loss mechanism is identified for different oxide thickness. The Vt loss in a SONOS flash exhibits twos stages. For a thinner bottom oxide, the transition time between the two stages is equal to trapped charge direct tunneling time. For a thicker oxide, the transition time is related to hole trap assisted tunneling time but is prolonged by a factor. No matter of thinner or thicker oxides, the Vt retention loss in the second stage obeys a log(t) dependence. The dominant mechanism is direct tunneling for thinner oxides and the FP emission and oxide charge assisted tunneling can well explain the observed retention characteristics in thicker oxide cells.

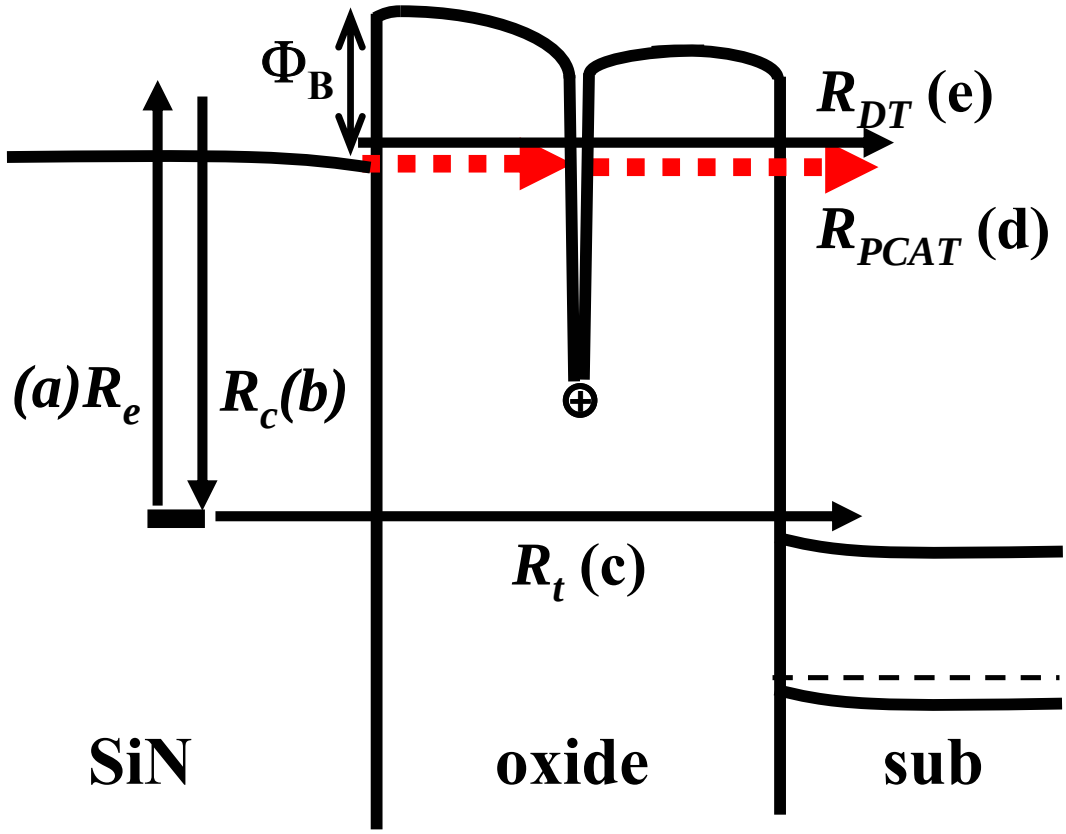
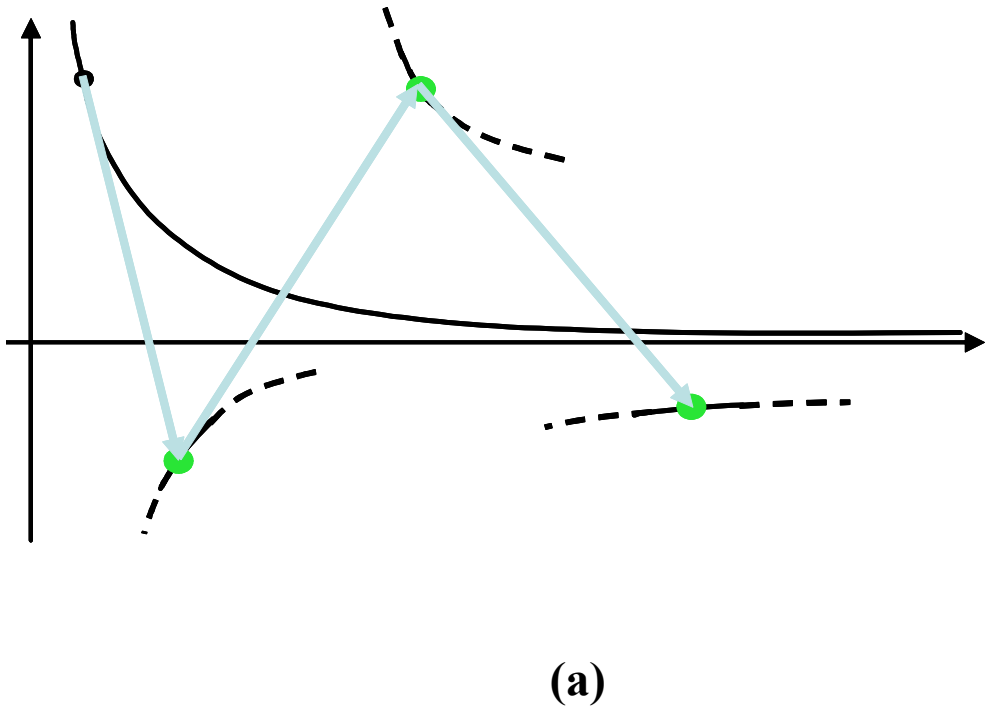
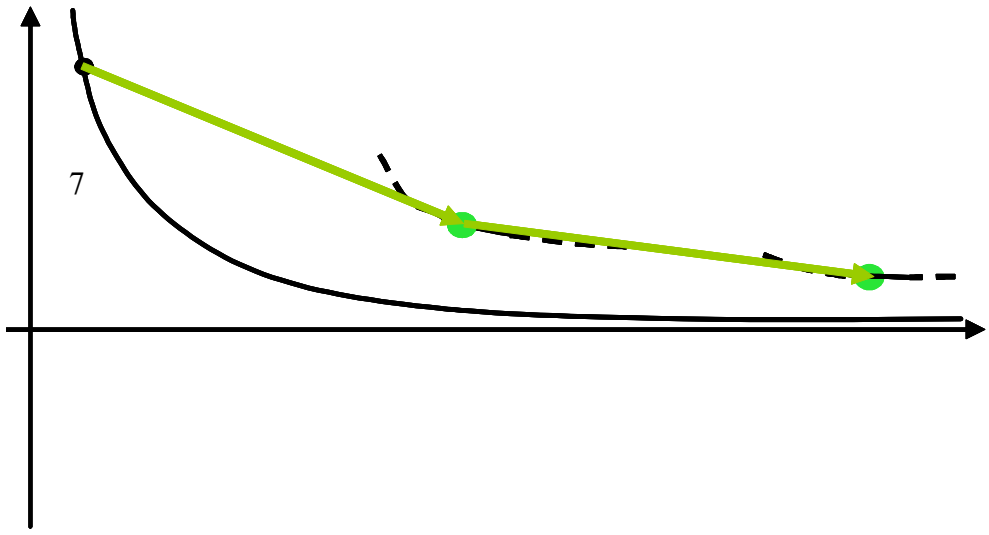


Fig. 5.1 Illustration of nitride electron leakage paths and electron transitions between the conduction band and trap states. (a) Frenkel-Poole excitation of trapped electrons to the SiN conduction band, (b) free electrons re-captured by nitride traps, (c) direct tunneling of nitride trapped electrons to Si substrate, (d) free electrons via PCAT to Si substrate, (e) direct tunneling of free electrons to Si substrate.



(a)



(b)

Fig. 5.2 (a) The "explicit" forward Euler. (b) The "implicit" backward Euler.

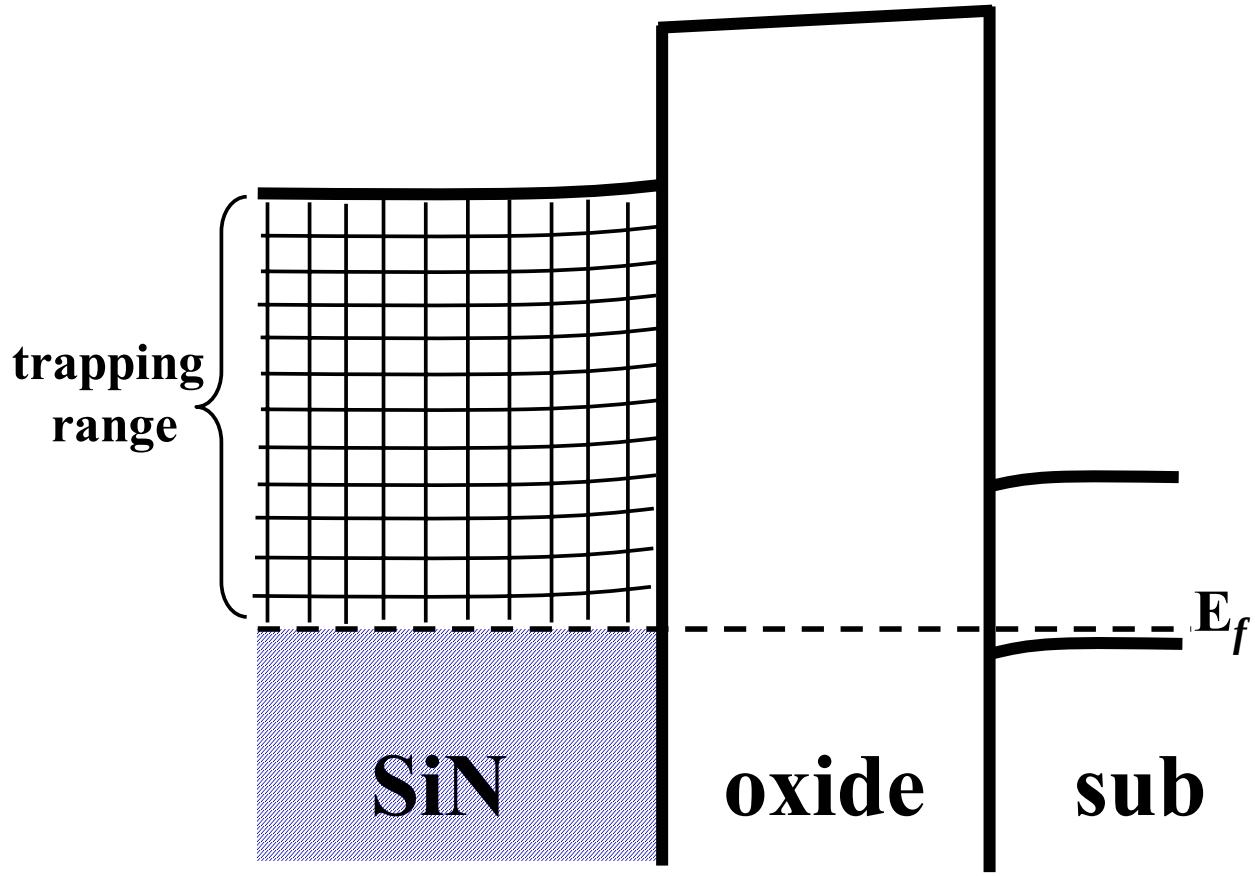


Fig. 5.3 Illustration of the discretization scheme in trap energy and space for numerical simulation.

$$\begin{bmatrix} -\frac{\sum_2^{Yn} R_d(N_t - y(n)_{k+1}^s)}{T_N} dx - R_{loss} \; (J_{11}) & \frac{R_d \cdot y(1)_{k+1}^s + R_e(2)}{T_N} \; (J_{12}) & K \;\; \cdots & \frac{R_d \cdot y(1)_{k+1}^s + R_e(Yn)}{T_N} \; (J_{1Y_n}) \\ R_d(N_t - y(2)_{k+1}^s) \; (J_{21}) & -R_d \cdot y(1)_{k+1}^s - R_e(2) - R_t(2) & 0 \;\; 0 & 0 \\ R_d(N_t - y(3)_{k+1}^s) \; (J_{31}) & 0 & O \;\; 0 & 0 \\ M & 0 & 0 \;\; O & 0 \\ R_d(N_t - y(Yn)_{k+1}^s) \; (J_{Y_n 1}) & 0 & 0 \;\; 0 & -R_d \cdot y(1)_{k+1}^s - R_e(Yn) - R_t(Yn) \; (J_{Y_n Y_n}) \end{bmatrix}$$

(a)

where

$$\text{a. } J_{11} = -\frac{\sum_2^{Yn} R_d(N_t - y(n)_{k+1}^s)}{T_N} dx - R_{loss} \qquad \rightarrow \text{sum of } n_c \text{ loss}$$

$$\text{b. } J_{1n} = \frac{R_d \cdot y(1)_{k+1}^s + R_e(n)}{T_N} dx \qquad \rightarrow \text{conduction transition}$$

$$\text{c. } J_{n1} = R_d(N_t - y(n)_{k+1}^s) \qquad \rightarrow \text{electron trapping}$$

$$\text{d. } J_{nn} = -R_d \cdot y(1)_{k+1}^s - R_e(n) - R_t(n) \qquad \rightarrow \text{diagonal}$$

(b)

Fig. 5.4 (a) The Jacobian matrix structure. (b) Each element of Jacobian matrix.

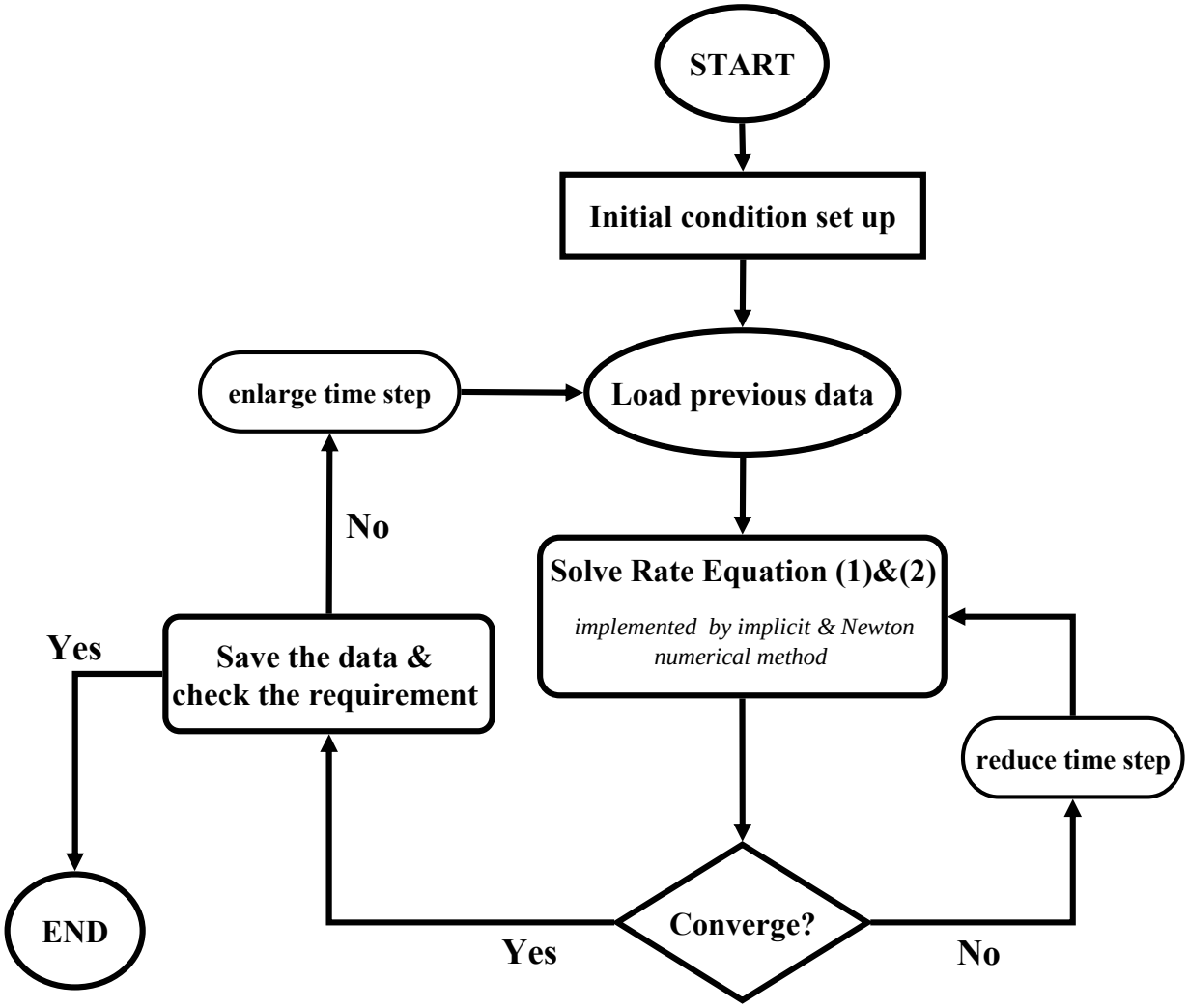


Fig.5.5. The flowchart for our simulation.

Table 5.1 Parameters used for the numerical simulation..

Parameters	Value
ε_{ox}	$3.9\varepsilon_o$
ε_N	$7\varepsilon_o$
m_{ox}	$0.42m_o$ [5.26]
m_N	$0.42m_o$ [5.26]
Φ_B	1.05 eV [5.6][5.7]
N_{t0}	3.0×10^{19} cm$^{-3}\cdot$eV^{-1}
σ_N	5×10^{-13} cm^2 [5.6][5.7]
σ_{ox}	1×10^{-14} cm^2
υ_{th}	10^7 cm/s
ν_e	5×10^{15} s^{-1}
ν_t	1.4×10^{14} s^{-1}

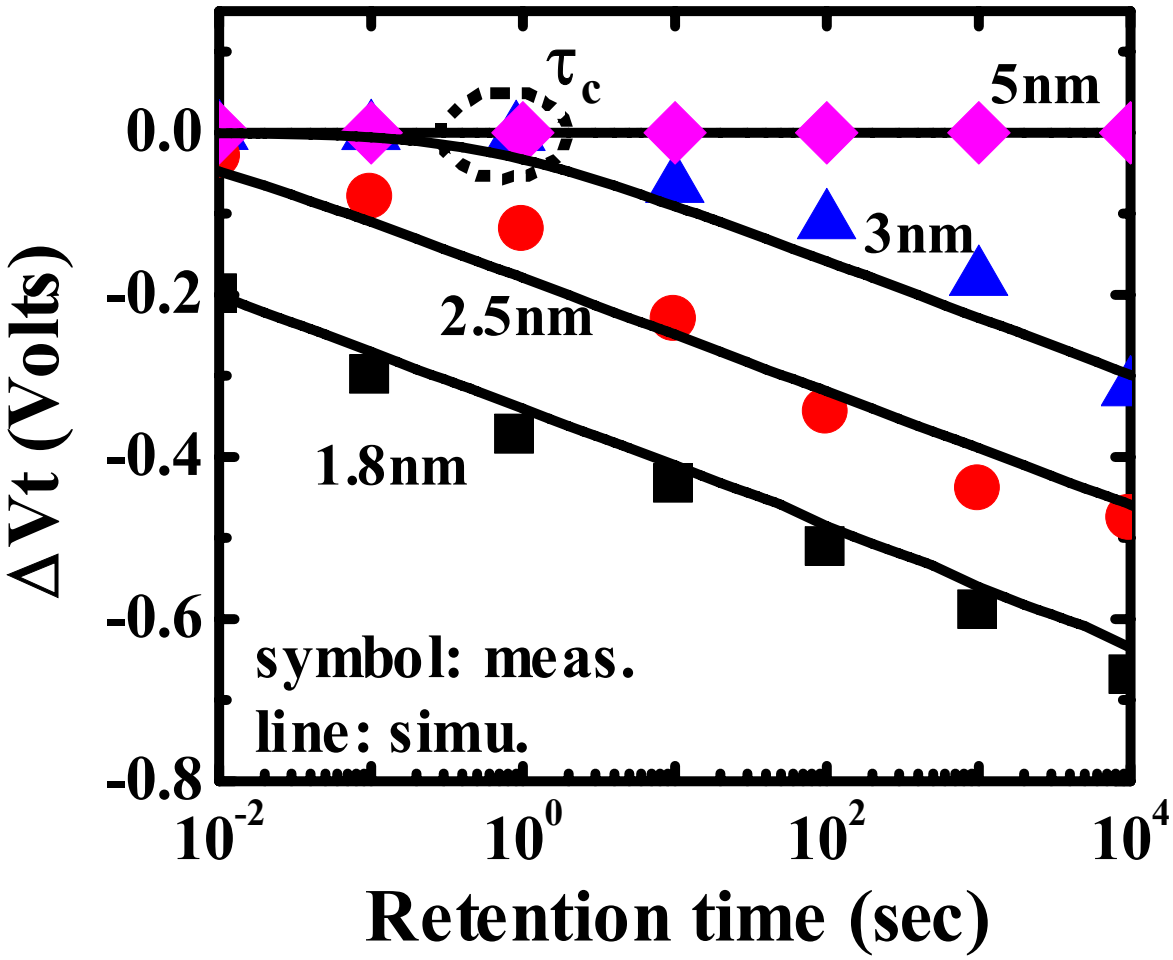


Fig. 5.6 Program-state Vt retention behavior for different bottom oxide thickness. All devices are programmed to an identical threshold voltage window of 1V. The symbols represent measured data and the lines are simulated results. The corner time (τ_{COR}) is indicated.

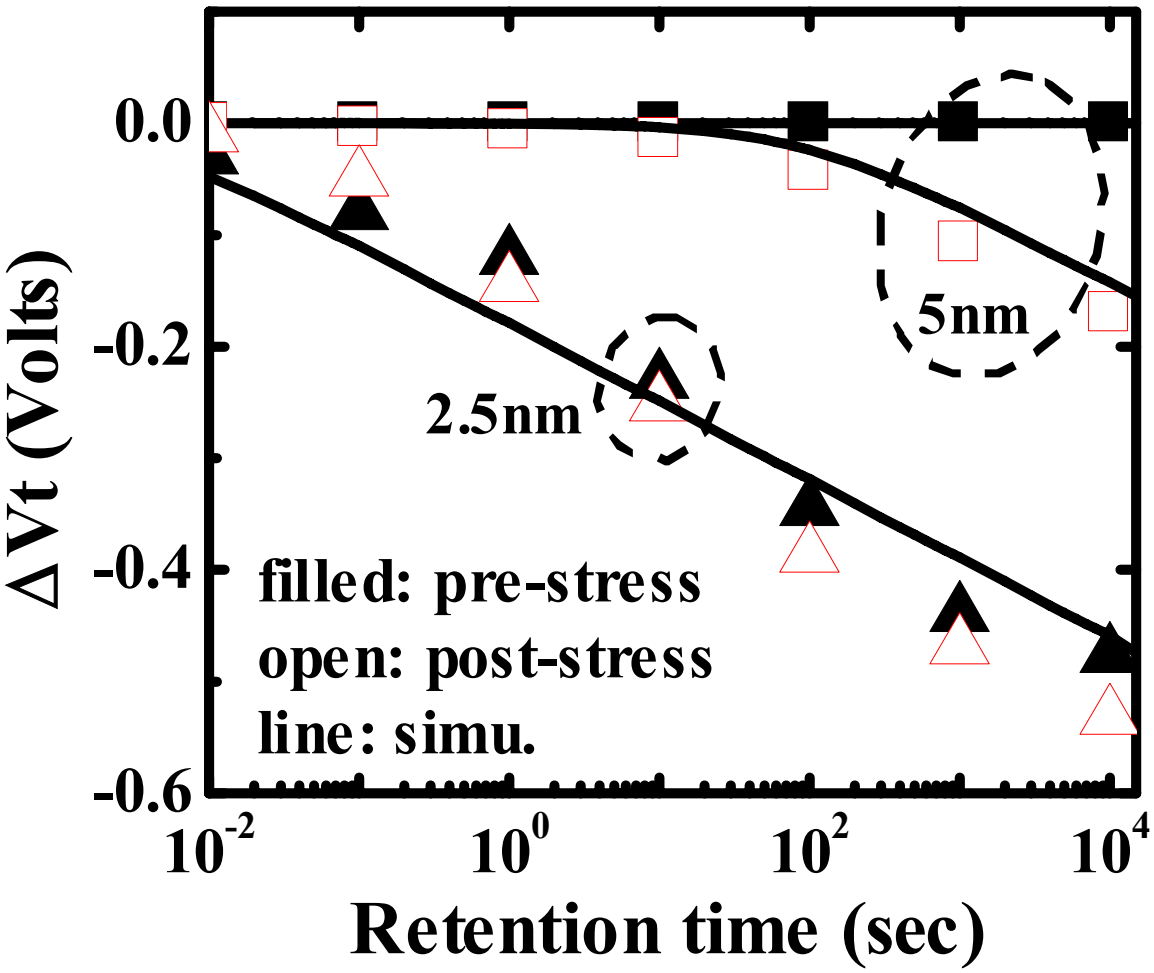


Fig. 5.7 Pre-stress and post-stress Vt loss for Tbo =2.5nm and 5nm. FN stress was performed at Eox= -11MV/cm for 100s.

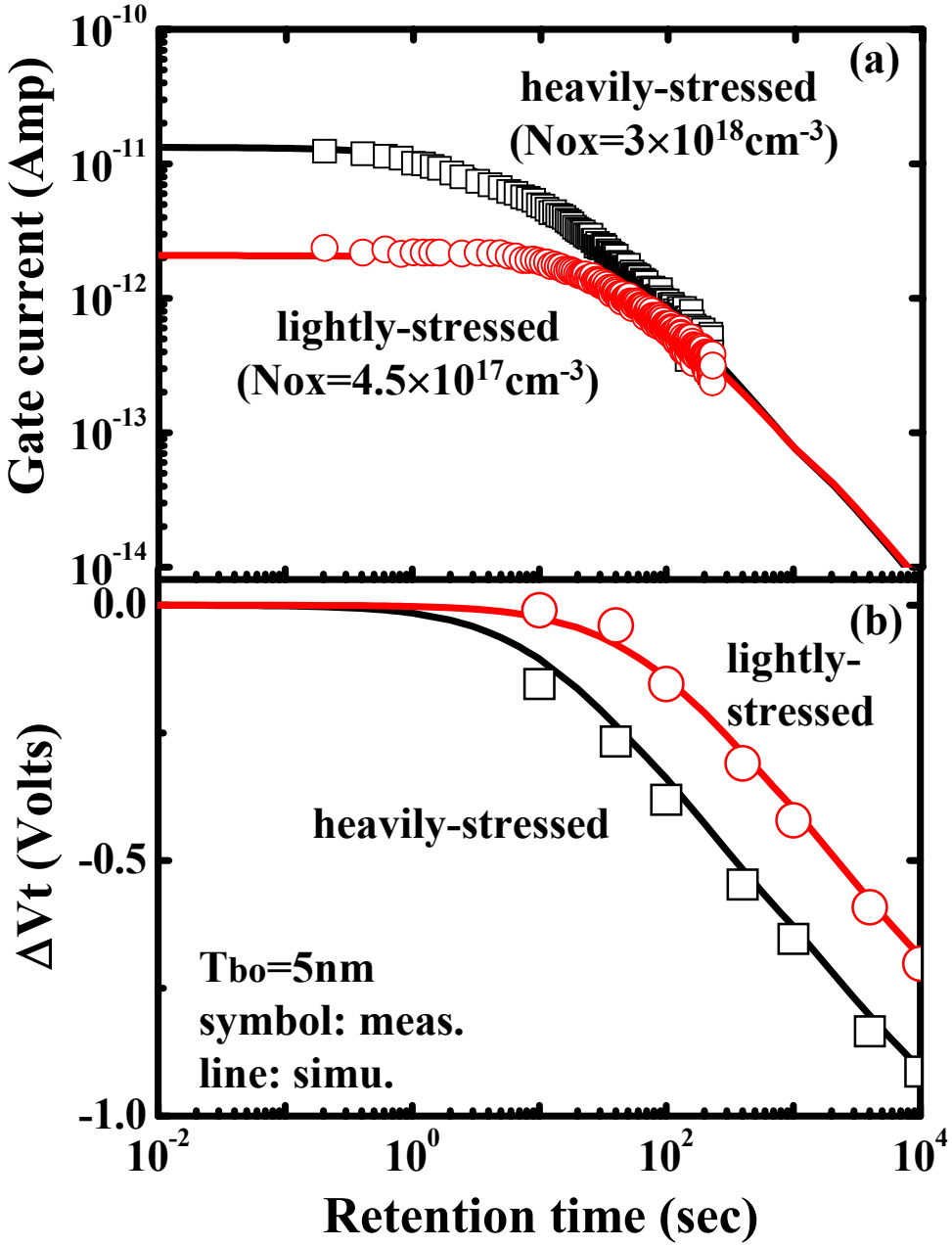


Fig. 5.8 (a) Measured stress induced leakage current (gate current) in a large area device (500μm×500μm) at V_g=0V. The lightly- and heavily-stressed conditions are at Eox=-11MV/cm for 1s and 1000s, respectively. Both devices are programmed to an identical threshold voltage window of 3V. (b) The corresponding Vt retention loss for the two stressed cells.

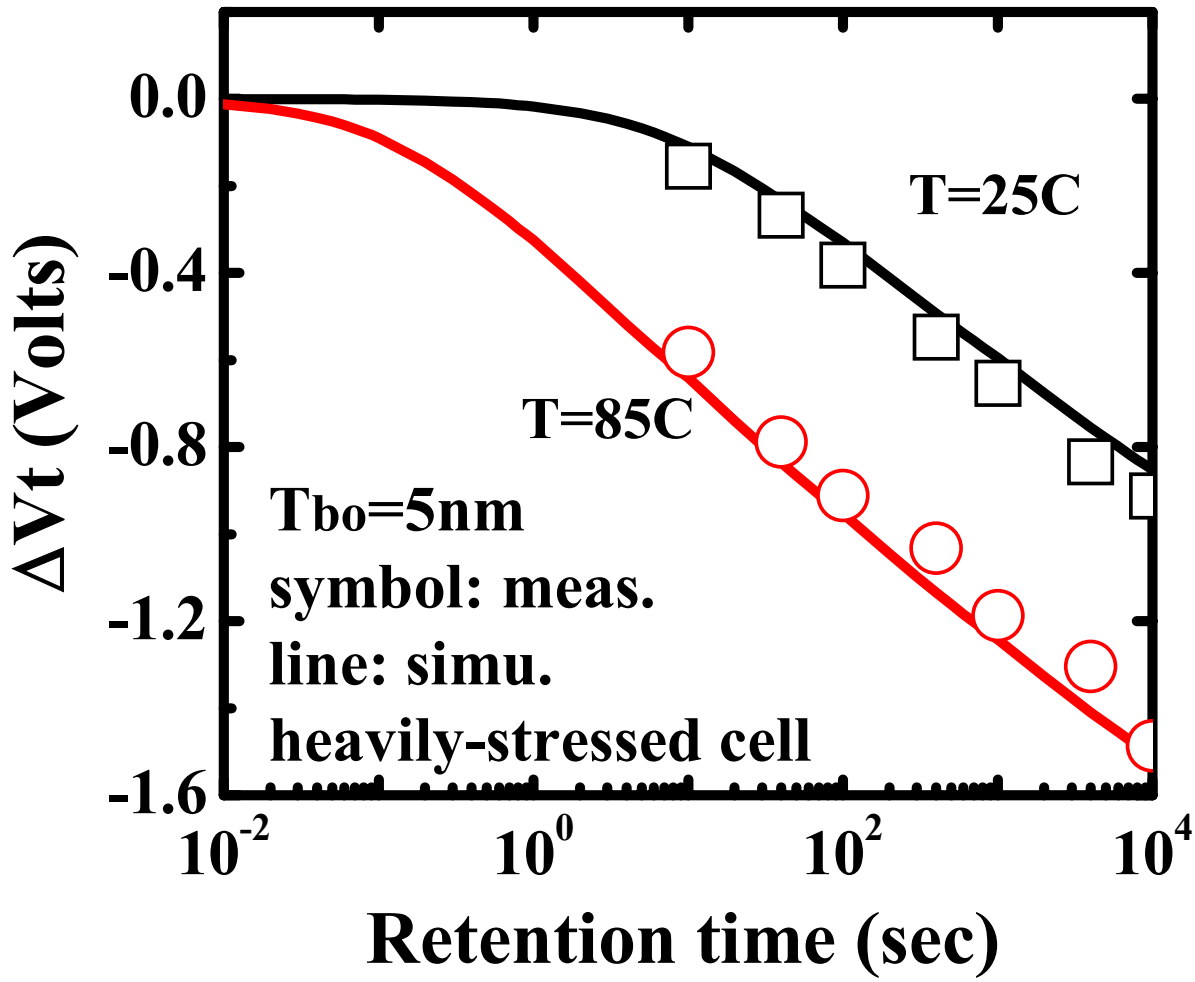


Fig. 5.9 Measured and simulated Vt loss versus retention time in the heavily stressed sample at T=25C and 85C.

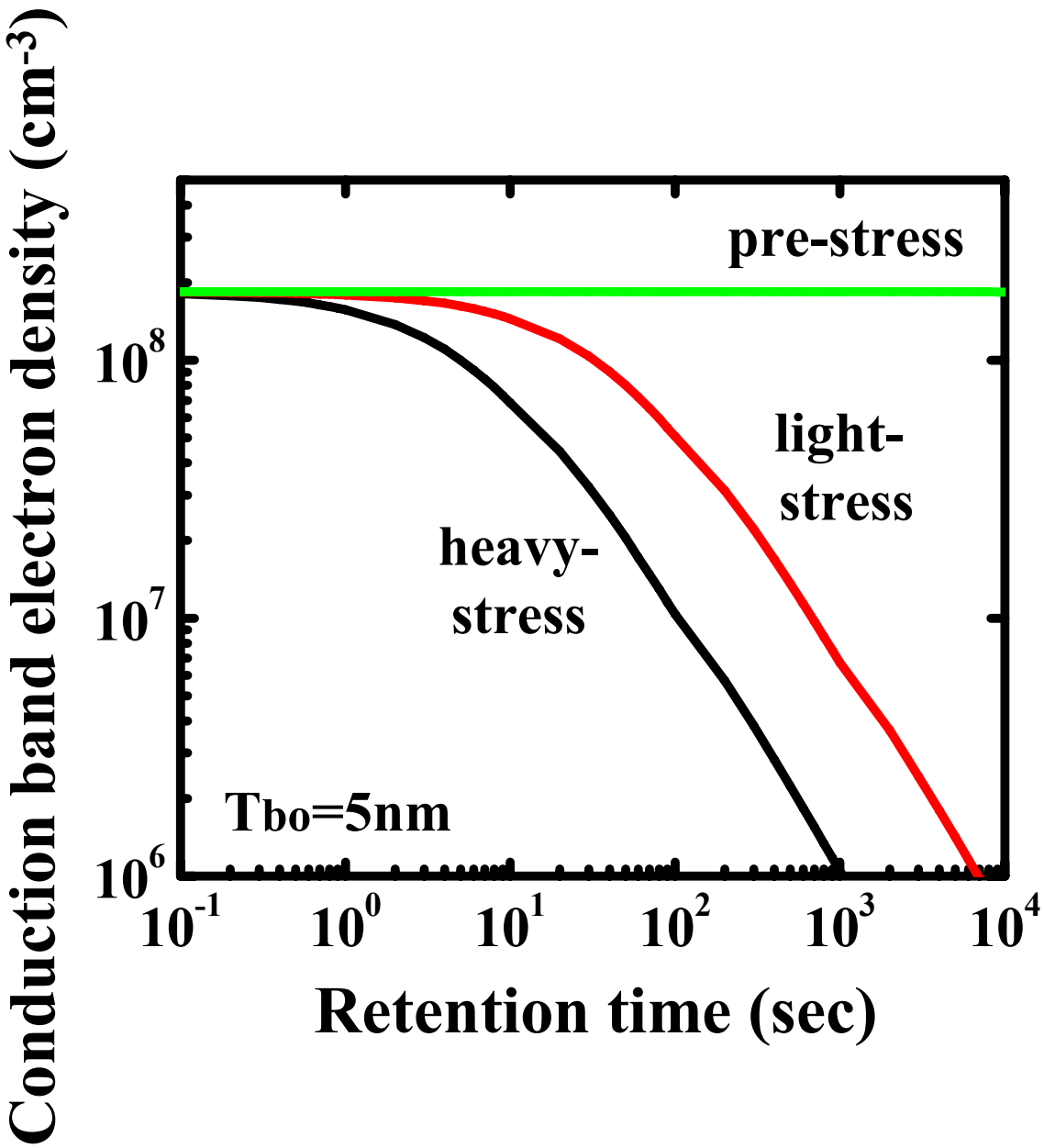


Fig. 5.10 Calculated conduction band electron density versus retention time before stress and after light-stress and heavy-stress, respectively.

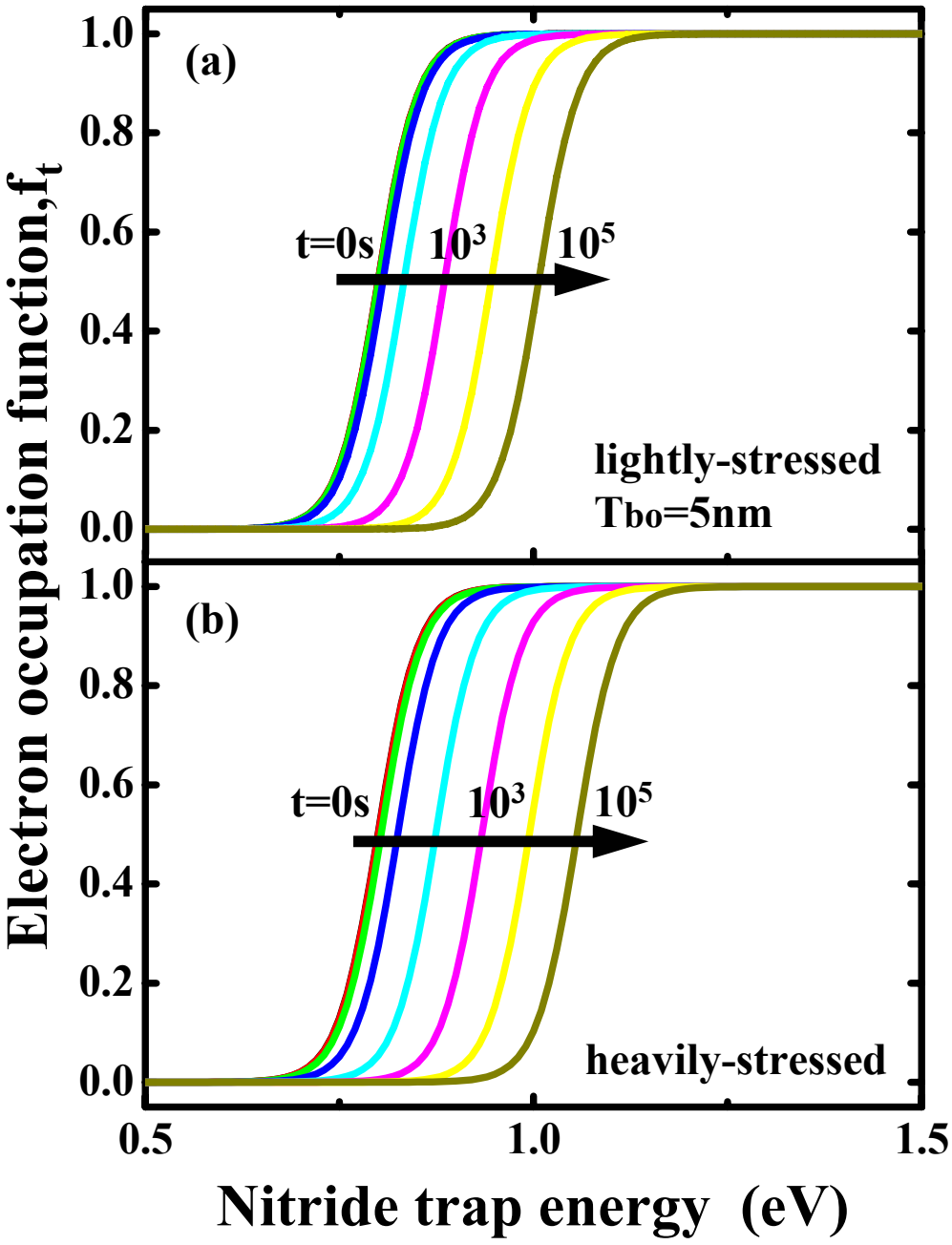


Fig. 5.11 Simulated electron occupation factor of nitride traps versus trap energy at different retention times. (a) lightly-stressed device and (b) heavily-stressed device.

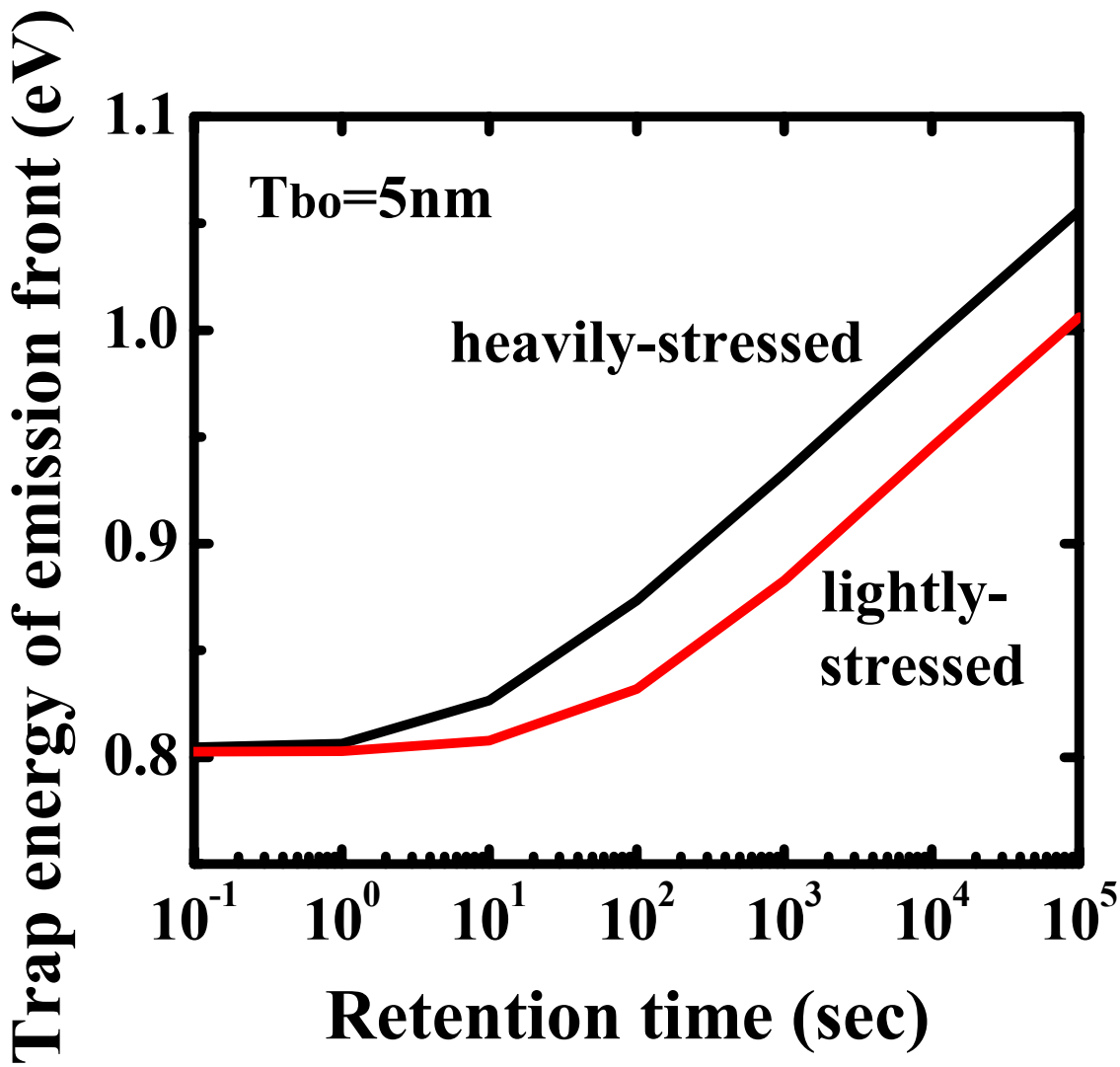


Fig. 5.12 The temporal evolution of the FP emission front, which is defined as the trap energy for the occupation factor in Fig. 5.11 equal to 0.5.

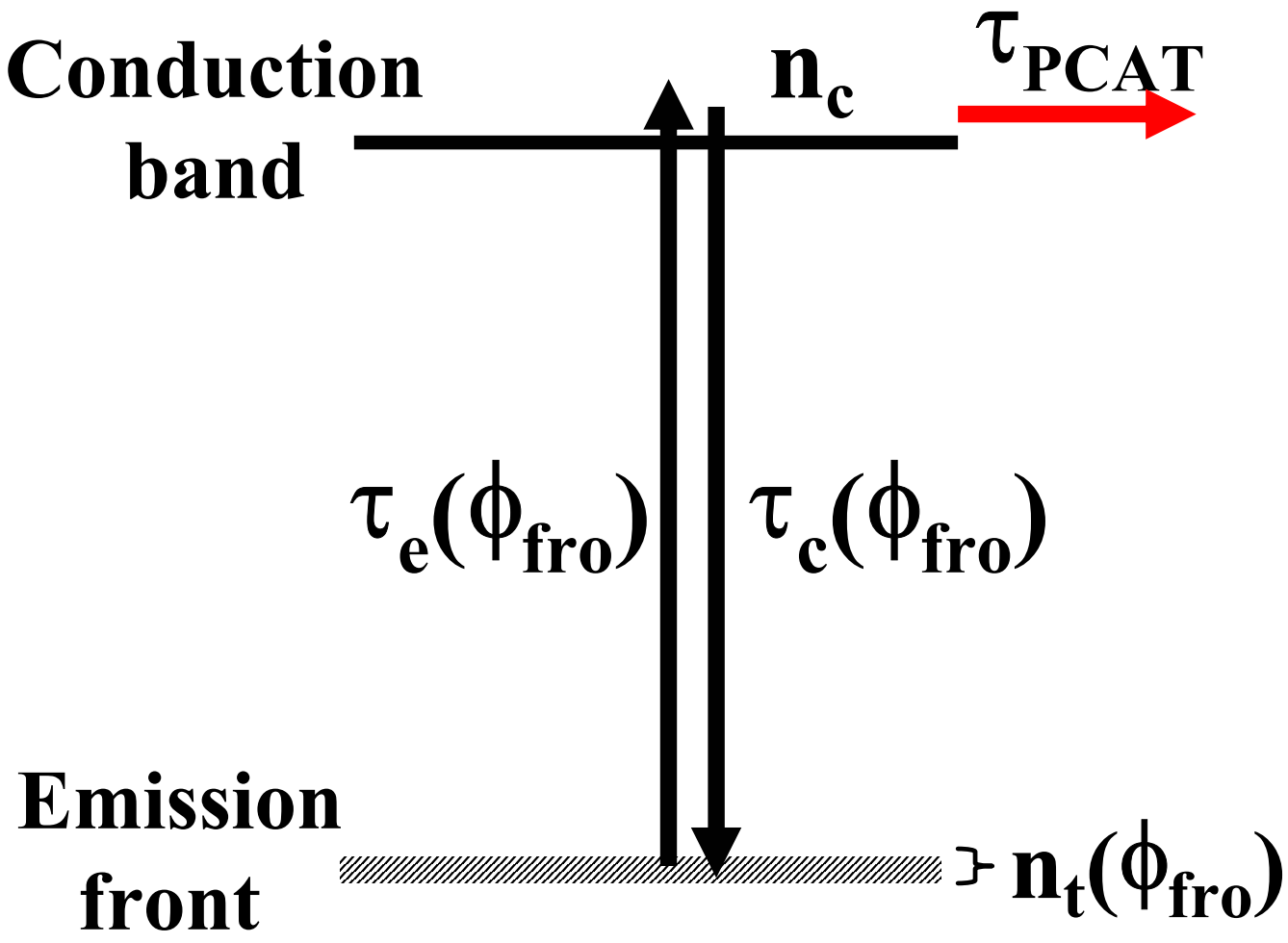


Fig. 5.13 Illustration of electron emission and capture between the conduction band and trap states around the FP emission front in an energy range of kT.

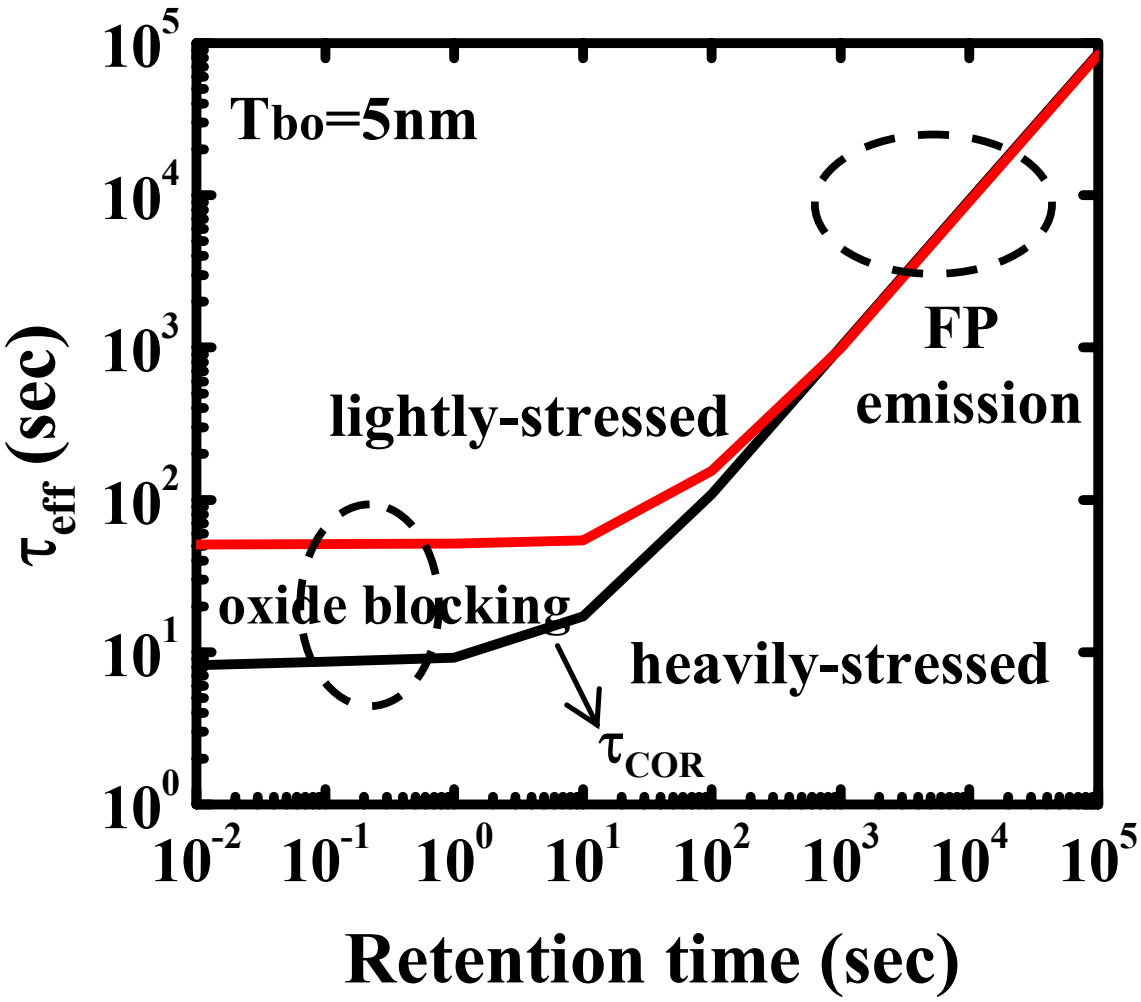


Fig. 5.14 Calculated effective time for nitride trapped charge at the FP emission front to escape from the ONO film.

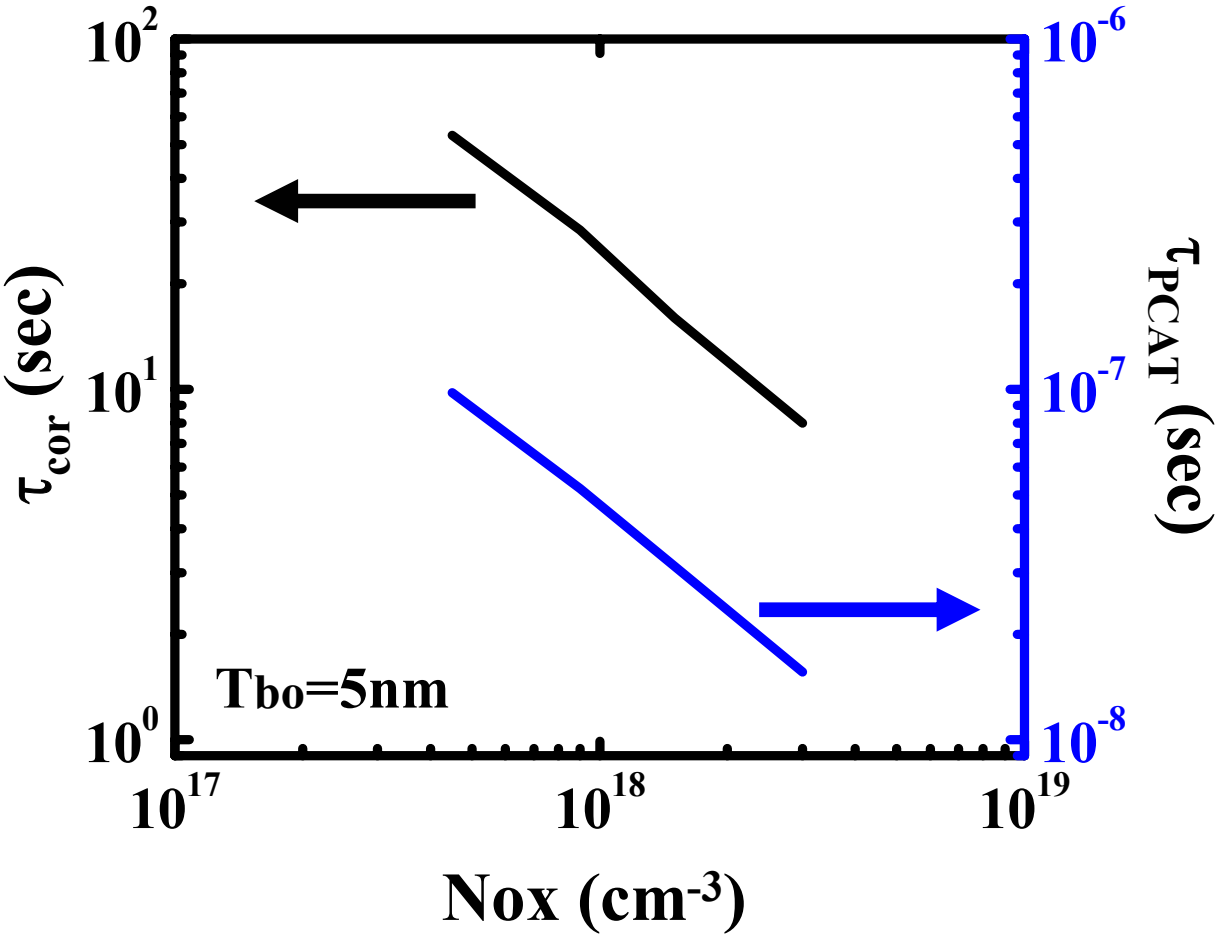


Fig. 5.15 The corner time (τ_{COR}) and positive oxide charge-assisted electron tunneling time (τ_{PCAT}) versus positively charged oxide trap density (Nox).

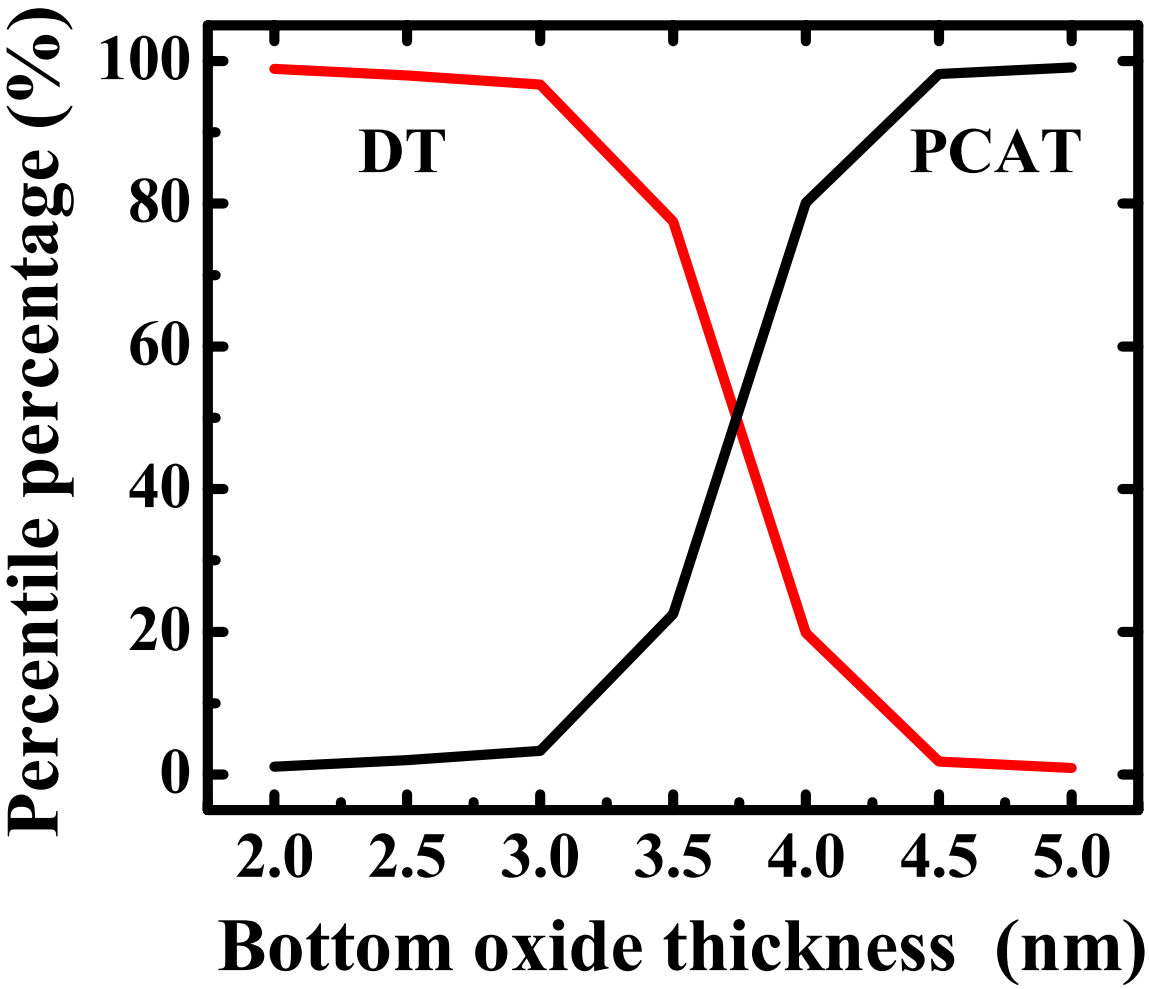


Fig. 5.16 The percentage of PCAT and DT caused retention Vt loss as a function of bottom oxide thickness. Positively charged oxide traps with a density of $3\times10^{18}cm^{-3}$ are placed at the most favorable position of PCAT. The dominant charge loss mechanism changes from DT to PCAT as oxide thickness increases.

Chapter 6

Extraction of Nitride Trap Density from Stress Induced Gate Leakage Current in SONOS Flash Memory

6.1 Introduction

For SONOS cells, programmed charges are stored in silicon nitride traps. The cell programming and retention characteristics are intimately related to nitride trap characteristics. Much research effort with respect to silicon nitride process optimization has been conducted to improve nitride trap properties. However, nitride trap characterization techniques are still very limited at present. Lundström et al. estimated a nitride trap density with a direct tunneling model [6.1]. White et al. employed a low-frequency (<1kHz) charge pumping (CP) technique to separate nitride traps from interfacial oxide traps by a difference in their time-constants [6.2]. Later, they extracted a nitride trap density from reverse modeling of threshold voltage (Vt) retention loss by combining trap-to-band tunneling and thermal excitation of trapped electrons in a nitride [6.3][6.4]. All the above methods, however, are restricted to ultra-thin bottom oxides (1.5nm~2.5nm). For today's SONOS cells, for example, NROM or Nbit technology, a thicker bottom oxide is usually employed to improve data retention. No appropriate nitride trap characterization methods are available for these cells.

Continued from the preceding understanding (Chapter 4&5), here, we develop an analytical model to extract a nitride trap density from the stress induced gate current. The effects of programming window (ΔVt) and stress condition on the

extraction result are evaluated. The nitride trap energy distribution is also profiled.

6.2 Nitride Trap Characterization Technique

The SONOS capacitors used in this work have a 9nm top oxide, a 6nm silicon nitride, and a 5nm bottom oxide. The capacitor area is 500μm×500μm. Uniform FN injection is employed for programming. The measured program-state gate leakage current versus retention time before and after a FN stress is shown in Fig. 6.1(a). The FN stress is performed at V_g=-20V for 1500s. The program Vt window is 3V. The corresponding Vt retention loss is shown in Fig. 6.1(b). A significant stress induced leakage current (SILC) is observed in Fig. 6.1(a). Unlike in a MOS structure, the SILC in a SONOS exhibits unique two-stage time dependence. In the first stage, the SILC exhibits a DC-like characteristic and in the second stage it follows a 1/t time dependence. Our numerical simulation in chapter 5 has shown that the first stage leakage current is limited by stress created oxide traps and the second stage is determined by the Frenkel-Poole (FP) emission of nitride trapped electrons. According to the FP emission model, the nitride trapped charge emission time is

$$\tau_e = (A^* T^2)^{-1} \exp\left(\frac{\phi_N - q\sqrt{qE_N/\pi\varepsilon_N}}{kT}\right)$$
$$= \tau' \exp(\phi_N/kT) \qquad (6\text{–}1)$$

where A* is the Richardson constant, ϕ_N is the nitride trap energy measured from the conduction band edge, E_N is the electric field, τ' is the proportionality constant and other variables have their usual definitions. In the FP emission limited condition

(second stage), the electron occupation factor (f_t) of a nitride trap state with an energy ϕ_N has a retention time dependence as follows [6.3][6.4]:

$$f_t(\phi_N) = \exp[-t/\tau_e(\phi_N)] = \exp[(-t/\tau')\exp(-\phi_N/kT)] \quad (6\text{–}2)$$

Because the above double exponential changes abruptly from 0 to 1 around $\phi_f = (kT)\ln(t/\tau')$, $f_t(\phi_N)$ can be approximated by a step-function at $\phi_N = \phi_{fro}$ and thus ϕ_{fro} is referred to as the FP emission front hereafter. This approximation translates into a "clear-cut" picture; at a certain time t, trap states above the emission front ϕ_{fro} are completely emptied while the states below ϕ_{fro} are occupied by electrons. The FP emission front moves downward in SiN band-gap with a speed of $d\phi_{fro}/dt = kT/t$, or 2.3kT per decade of time. The nitride charge leakage current in the second stage therefore can be expressed as

$$I_N = AqNt(\phi_{fro})\cdot d\phi_{fro}/dt = AqNt(\phi_{fro})\cdot kT/t$$

(6-3)

and the nitride trap density can be extracted,

$$Nt(\phi_{fro}) = I_N\cdot t/AqkT \quad (6\text{-}3a)$$

Since the emission front stays almost unchanged in the first stage ($\phi_{fro} = \phi_0$) and begins to move at the corner time (τ_{COR}) of the second stage with a constant speed in a log(t) scale, the relationship between ϕ_{fro} and retention time t is readily obtained,

$$\phi_{fro} = \phi_0 + kT\ln(t/\tau_{COR}) \quad (6\text{-}4)$$

where ϕ_0 is the emission front in the first stage and is determined by the total amount of programmed charges. From our numerical simulation, ϕ_0 is estimated to be 0.8eV for the present SONOS structure and programming window. From Eqs (6.3) and (6.4), the nitride trap density is characterized, as shown in Fig. 6.2. The solid line is directly from the SILC (Fig. 6.1(a)) and the dashed line is extrapolated from the Vt shift (Fig. 6.1(b)). Notably, Fig. 6.2 reveals that the trapped charges have a rather uniform distribution over an energy range of measurement (~0.2eV).

To further verify our nitride trap profiling technique, we change the FN stress condition and programming window. Fig. 6.3 shows the SILC for two different FN stresses. The two samples have the same programming window and thus the same ϕ_0. We should state again that the effective time for nitride trapped charges at the emission front to escape from the ONO film is [6.5]

$$\tau_{eff} = \frac{\tau_e(\phi_{fro}) + \tau_c(\phi_{fro})}{\tau_c(\phi_{fro})} \cdot \tau_{ox} \approx \frac{\tau_e(\phi_{fro})}{\tau_c(\phi_{fro})} \cdot \tau_{ox} \quad (6\text{-}5)$$

where $\tau_e(\phi_{fro})$ and $\tau_c(\phi_{fro})$ represent the electron emission time and capture time between the emission front and the nitride conduction band. τ_{ox} is the conduction band electron tunneling time via stress induced oxide traps. The lightly stressed sample (B) has a lower first-stage DC leakage because of less oxide trap creation and thus a longer τ_{ox}. It is worth pointing out that the SILC in the two samples (A&B)

converges in the second stage. The reason is that the ϕ_{fro} of the two samples at a certain time in the second stage is only shifted by $kTln(\tau_{COR}(B)/\tau_{COR}(A))$~0.06eV. As mentioned earlier, the nitride trap density Nt is almost a constant in the measurement range. The leakage current, which is proportional to Nt (Eq. (6-3)), is therefore nearly the same in the second stage no matter of stress conditions.

In addition, the SILC for different program windows is shown in Fig. 6.4. The program window is 3V for sample A and 2.5V for sample C. The sample A has a smaller ϕ_0 for more programmed charges. According to the Shockley-Read-Hall theory, $\tau_e/\tau c \propto \exp(\phi_{fro}/kT)$ and the difference in ϕo between samples A and C can be estimated from the ratio of the SILC in the first stage, i.e,

$$SILC(\text{first stage}) \propto 1/\tau_{eff} \propto \exp(-\phi_o/kT) \quad (6\text{-}6a)$$

and

$$\phi_0(A)-\phi_0(C) = -kTln(SILC(A)/SILC(C)) \quad (6\text{-}6b)$$

The ϕ_0 difference is still small. Consequently, samples A and C have almost the same leakage current in the second stage because $Nt(\phi)$ can be considered as a constant for such small ϕ_0 difference.

In summary, the extracted nitride trap density from the samples A, B, C is listed in Table 6.1. The extracted result is in reasonably good agreement without regard to stress condition and programming window.

6.3 Summary

A simple and reliable nitride trap profiling technique for a thicker bottom oxide SONOS is developed. We find that the SILC in a SONOS exhibits a distinguished two-stage feature. The nearly 1/t dependence in the second stage suggests that the nitride traps have a uniform energy distribution. The extracted nitride trap density is around 7.0×10^{12} $cm^{-2}eV^{-1}$.

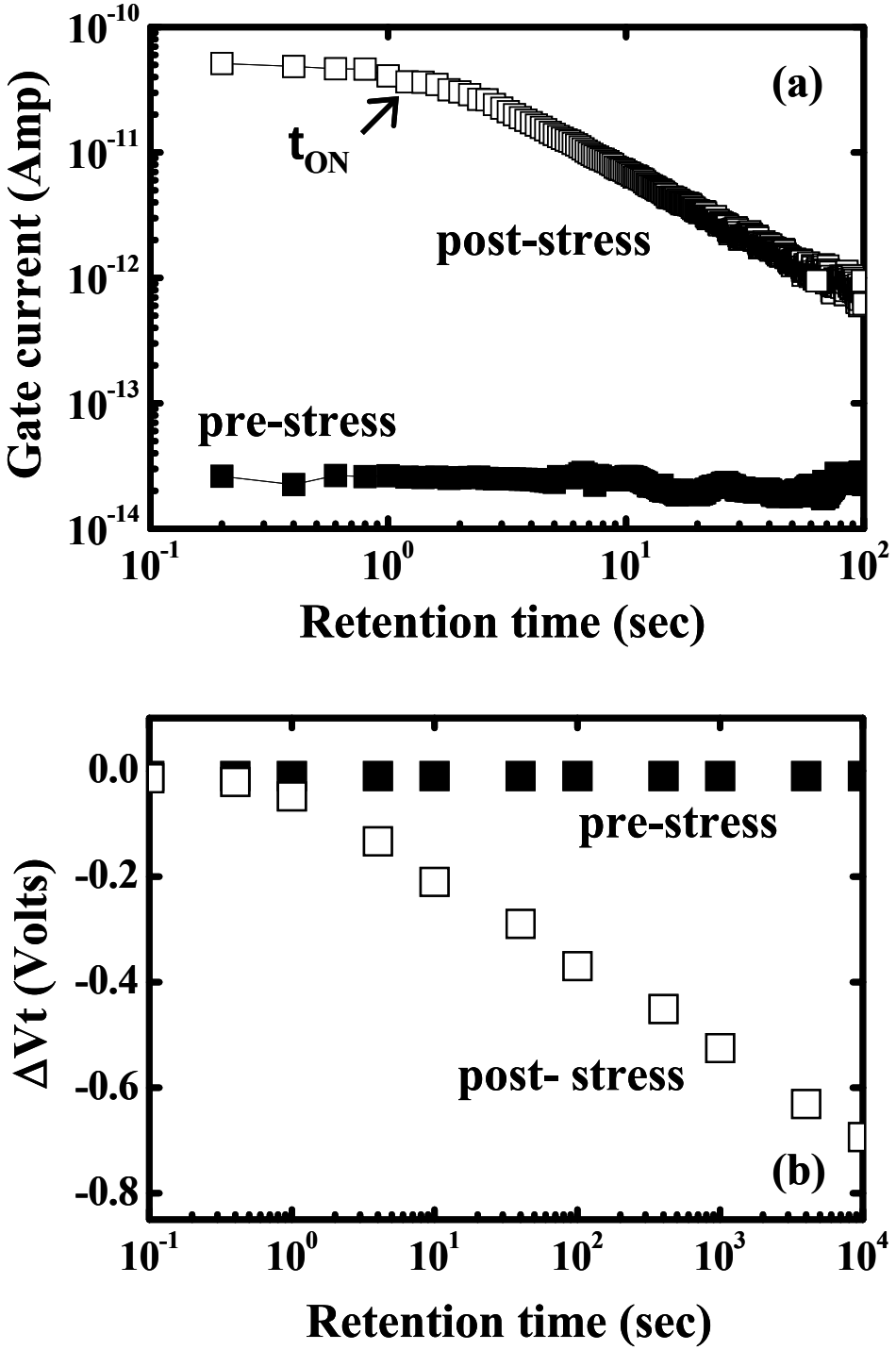


Fig. 6.1 (a) Measured gate leakage current in a large area SONOS (500μm×500μm) at Vg=0V. The stress condition is Vg=-20V for 1500s. Both devices are programmed to an identical threshold voltage window of 3V. (b) Corresponding Vt retention loss for the two samples.

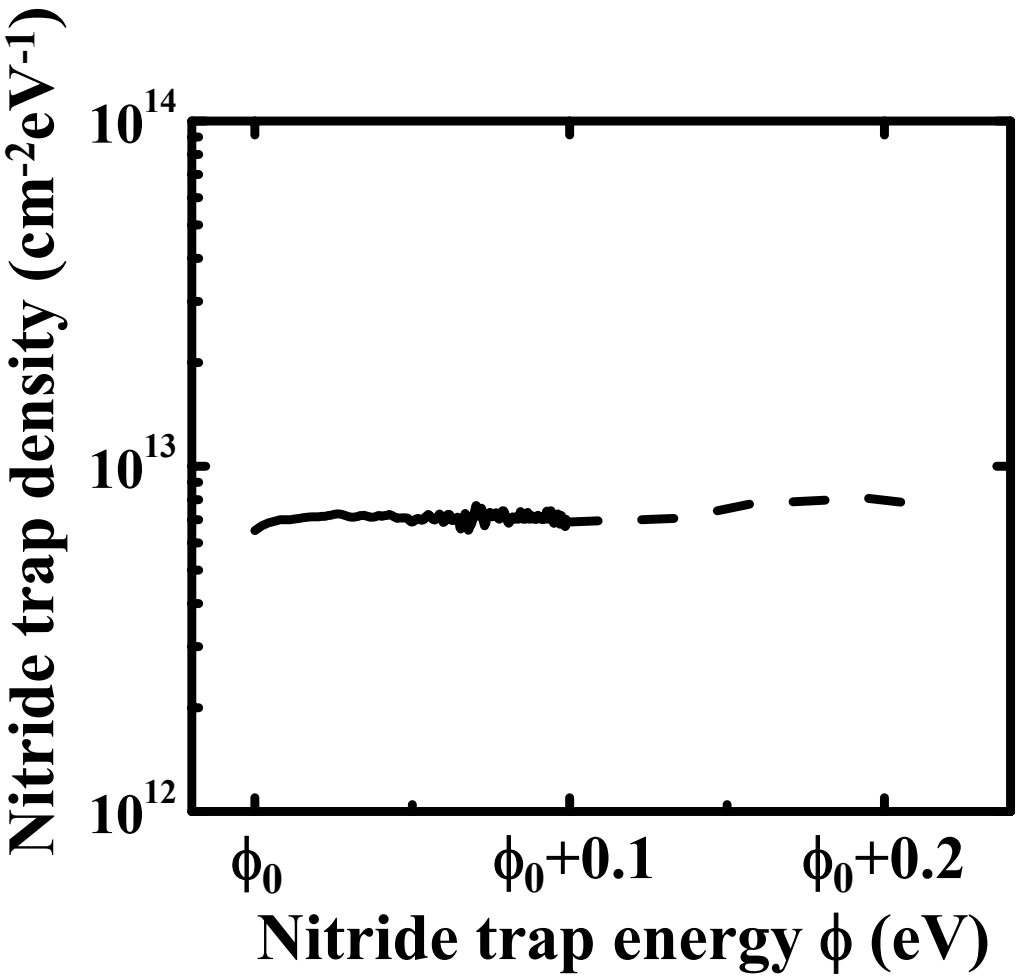


Fig. 6.2 Extracted nitride trap density distribution versus relative nitride trap energy. ϕ_0 is estimated to be 0.8eV.

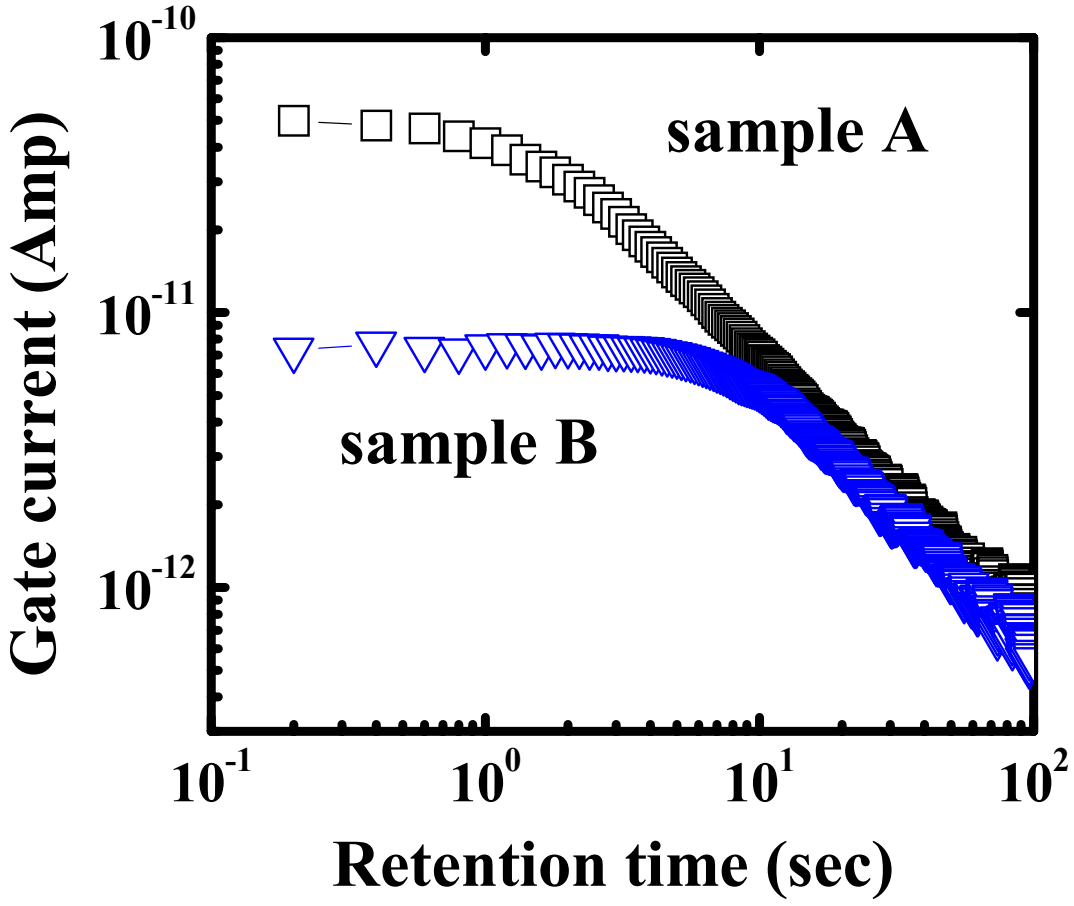


Fig. 6.3 The measured SILC in two differently stressed samples. Sample A is stressed at V_g=-20V for 1500s and sample B is for 5s. The programming window is 3V.

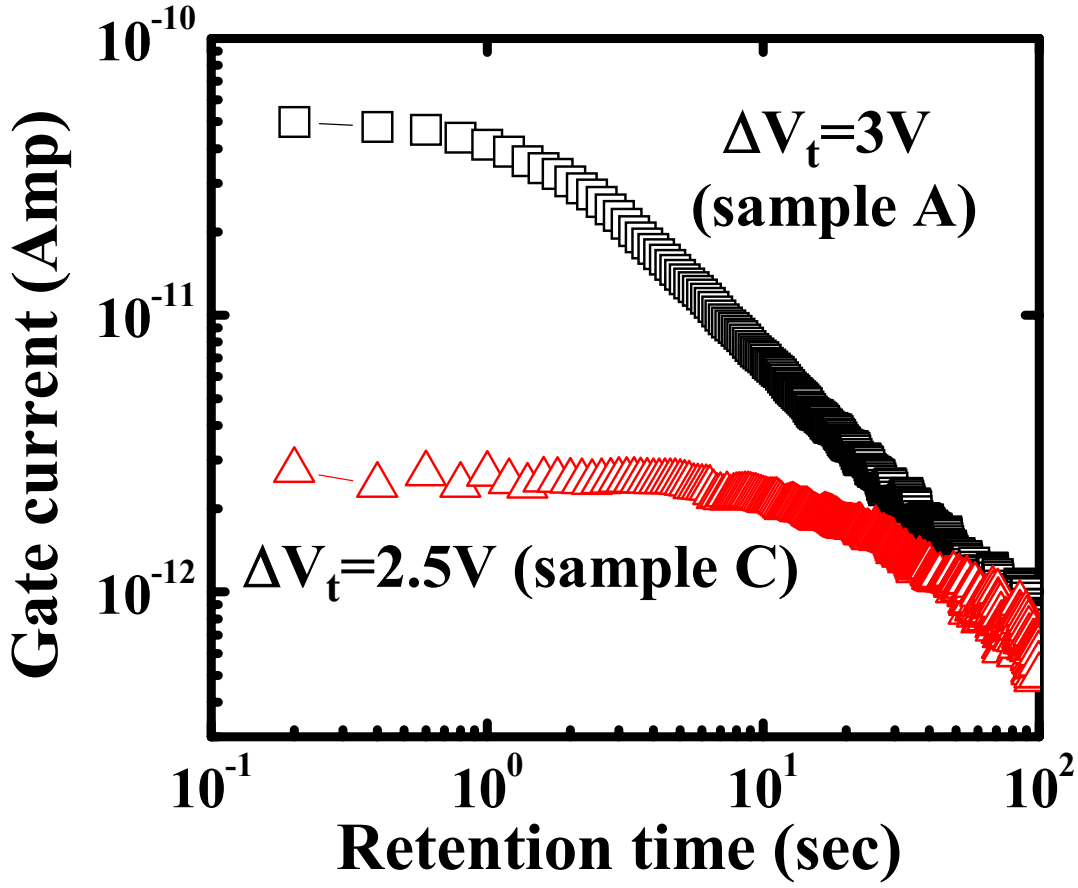


Fig. 6.4 The measured SILC for two different programming windows. Samples A and C are stressed under the same condition (V_g=-20V for 1500s) but have a program window of 3V and 2.5V respectively.

Table 6.1 The extracted nitride trap density from sample A, B, C respectively by using our technique.

Sample	Nt ($cm^{-2}eV^{-1}$)
A	7.1×10^{12}
B	6.3×10^{12}
C	6.2×10^{12}

Chapter 7

Read Current Instability Arising from Random Telegraph Noise in Localized Storage, Multi-Level SONOS Flash Memory

7.1 Introduction

Process technology has allowed significant density increases and lower cost-per-Mbyte since flash memory was first introduced in 1988. The primary technique for increasing the memory density is to downscale the cell size. Another approach to driving up memory density is to increase the number of possible states in a cell, called as Multi-Level Cell (MLC). For floating-gate cells, MLC is conceived by charging to different Vt [7.1]-[7.3]. For charge-trapping cells, 4-bit/cell SONOS flash memory can be realized [7.4][7.5] due to separated storage of charge packets at the two sides of a cell and four levels in each charge packet (Fig. 7.1). As bit size aggressively shrinks, the number of stored electrons in each program level greatly reduces. A single charge trapping/de-trapping will induce a large fluctuation in read current [7.6]. This fluctuation, referred to as random telegraph noise (RTN), will become a serious concern in advanced CMOS technologies and flash memories [7.7]-[7.9].

In this chapter, single charge phenomenon, such as RTN as well as nitride charge escape is observed by a fast-transient measurement system. Non-uniform threshold voltage distribution enhancing RTN is demonstrated and explained by a

simplified two-region model. A probability model to portray the read current distribution caused by RTN is proposed. Based on our understanding, the read current fluctuation can substantially alleviate by better bottom oxide process.

7.2 Experimental Setup

The SONOS flash cells used in this work have a gate width of 0.35μm and a gate length of 0.3μm. Channel hot electron program, band-to-band hot hole erase and reverse read are employed. The experimental setup for measuring read current variation is shown in Fig. 7.2(a). The sampling rate is 10KHz, which enables the observations of read current switching with time resolution up to 0.1ms. Fig. 7.2(b) depicts the waveforms of alternate measurement and stress to each terminal. The measurement time (t_{meas}) is chosen to 50ms such that it is long enough for integrating reliable signal as well as short enough to avoid introducing additional stress. Fig. 7.3 shows the read current instability in a SONOS cell after program/erase cycling. The read current instability is manifested by a stepwise increase with retention time, superimposed by RTN. The stepwise increase is attributed to stored charge escape from nitride traps to the substrate. Each current jump corresponds to a single nitride charge loss. The RTN in Fig. 7.3 results from the charging/discharging of oxide (interface) traps created by P/E stress. The fluctuations in read current imply we must allow more margins for reliable sensing in cell design. The impact of RTN in different programmed Vt levels of a multi-level cell will be studied below.

7.3 Two-region model

Since the charges are programmed locally, the channel can be roughly separated

into two electrically regions, as shown in Fig. 7.4. Similar to Ref. [7.10], we model the noise power as a two-region SONOS by converting the voltage fluctuations into current fluctuations,

$$\frac{S_I(f)}{G^2} = \frac{S_{I_1}}{G_1^{\,2}} + \frac{S_{I_2}}{G_2^{\,2}} \tag{7-1a}$$

$$\frac{\Delta I^2}{G^2} \propto \frac{\Delta I_1^2}{G_1^{\,2}} + \frac{\Delta I_2^2}{G_2^{\,2}} \tag{7-1b}$$

where subscripts 1 and 2 denote the two separated regions. G_i and ΔI_i is the channel conductance and the channel current fluctuation in each region; $G=G_1G_2/(G_1+G_2)$. Here, G_i and ΔI_i can be written as

$$G_i = \mu_i Q_i \frac{W}{L_i} = \mu_i \frac{qN_{chi}}{WL_i}\frac{W}{L_i} = \frac{q\mu_i N_{chi}}{L_i^{\,2}} \tag{7-2a}$$

$$\Delta I_i = \frac{q\mu_i V_{s_i}}{L_i^{\,2}} \tag{7-2b}$$

The Q_i and N_{chi} is the area charge density and the channel carrier number in different region, respectively. V_{si} is the voltage drop in each region. Putting Eq. (7-2a) and (7-2b) into (7-1b), we obtain

$$\Delta I^2 \propto G^2\left(\frac{V_{s_1}^{\,2}}{N_{ch1}^{\,2}} + \frac{V_{s_2}^{\,2}}{N_{ch2}^{\,2}}\right) \tag{7-3}$$

Because of the higher threshold voltage in region 2, V_{s2} is larger than V_{s1}, while N_{ch2} is much smaller than N_{ch1}. Eq. (7-3) can be rewritten as

$$\Delta I \propto \frac{q\mu_2 V_{s_2}}{L_2^{\,2}} \tag{7-4}$$

Thus, region 2 (high Vt region) dominates the noise fluctuation. The fluctuation is increased by a non-uniform Vt distribution due to a small L_2. We will verify the concept in next section.

7.4 Results and Discussions

7.4.1 Localized charge storage

To study the impact of operation mode and charge storage condition on RTN, two different programming techniques for SONOS (FN program and CHE program) are compared (Fig. 7.5). The former has uniform charge storage and the latter has localized charge storage. The two cells experience the same cycling procedure. Notably, the RTN amplitude in the localized storage cell (CHE programming) is much larger than that in uniform storage (FN programming). This result implies the impact of RTN in a localized storage SONOS is more serious than that in a conventional floating-gate flash. The two-region model can also response to this behavior.

7.4.2 RTN in a multi-level cell

The amplitude of RTN in different Vt levels of a multi-level cell is evaluated

(Fig. 7.6). The read gate voltage is adjusted that the read current for each program level is kept about the same (10μA). Fig. 7.6 reveals that the RTN amplitude increases with program-state Vt. The reason is that a higher programmed Vt state has stronger non-uniformity in channel Vt distribution and thus a larger amplitude in RTN according to the two-region model. The RTN effect is more prominent as device scaling advances. Fig. 7.7 shows the RTN amplitude versus gate length in program state "01".

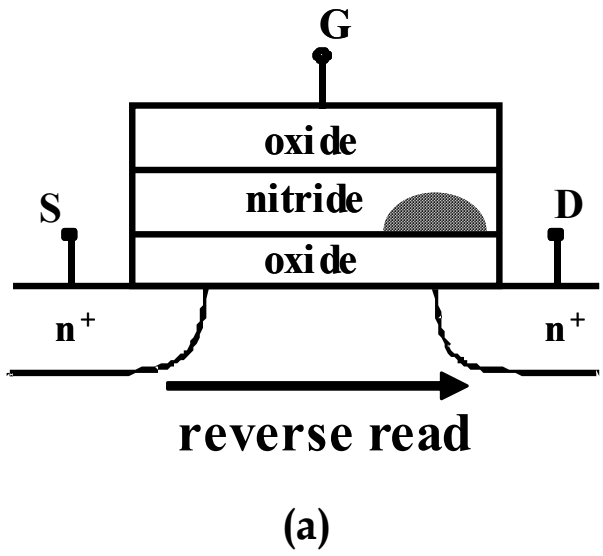
7.4.3 P/E stress and read current distribution modeling

The P/E cycling effect on RTN is shown in Fig. 8. RTN is rarely observed in an un-cycled cell (Fig. 7.8(a)). At 100 P/E cycles, typical RTN patterns exhibit two-level switching (Fig. 7.8(b)). At 100k P/E cycles, multi-level RTN is observed (for example, 5-level in Fig. 7.8(c)). This multi-level RTN, resulting from the overlap of several independent two-level RTN waveforms, may cause a large read current shift and thus a read failure. Fig. 7.9 shows RTN induced read current distribution at different P/E cycles. The number of readings is 10^5 for each cycle number. The read current distribution is broadened as cycle number increases due to more oxide trap creation. A probability model to account for RTN induced read current distribution is developed (Fig. 7.10). P_H and P_L are the probability of the current in high state and low state, respectively. These probabilities are obtained from two-level RTN measurement. The modeled result is shown in Fig. 7.9(c) (heavy line). Reasonable agreement between modeled and measured results is obtained. At 100k P/E cycles, the maximum current shift due to multi-level RTN exceeds 1.0μA for a sample size of 10^5 readings, which corresponds to a Vt shift of more than 0.1V in a 0.3μm SONOS.

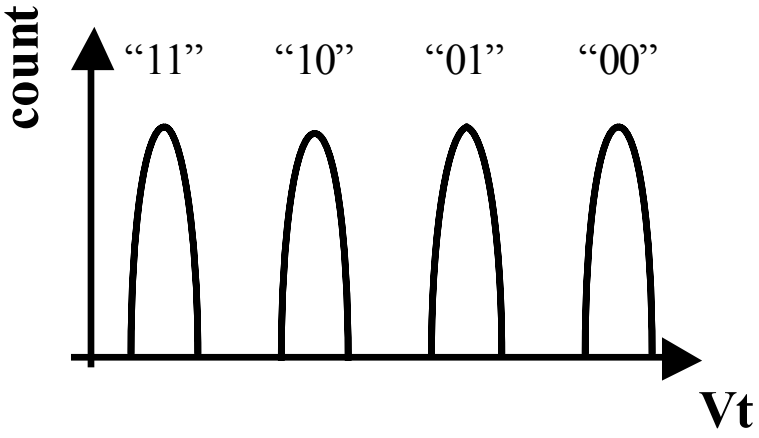
Two different oxide processes (oxide A and oxide B) are compared about RTN. Oxide B is known to have better oxide endurance from post-cycling charge pumping measurement because of NO treatment (Fig. 7.11(a)). The read current with oxide B apparently has a tighter distribution (Fig. 7.11(b)) because of less oxide (interface) traps.

7.5 Summary

The impact of RTN is more significant in a localized storage, multi-level SONOS flash memory cell. A probability model for the RTN induced read current distribution is proposed. The factors affecting RTN, such as charge storage condition, programmed Vt level, gate length and P/E cycling, have been characterized and discussed. Improvement of bottom oxide robustness can significantly reduce the RTN effect.



(a)



(b)

Fig. 7.1 (a) Schematic representation of separated storage of charge packets at the two sides of a SONOS cell with reverse read. (b) Each charge packet has four Vt levels in a multi-level cell.

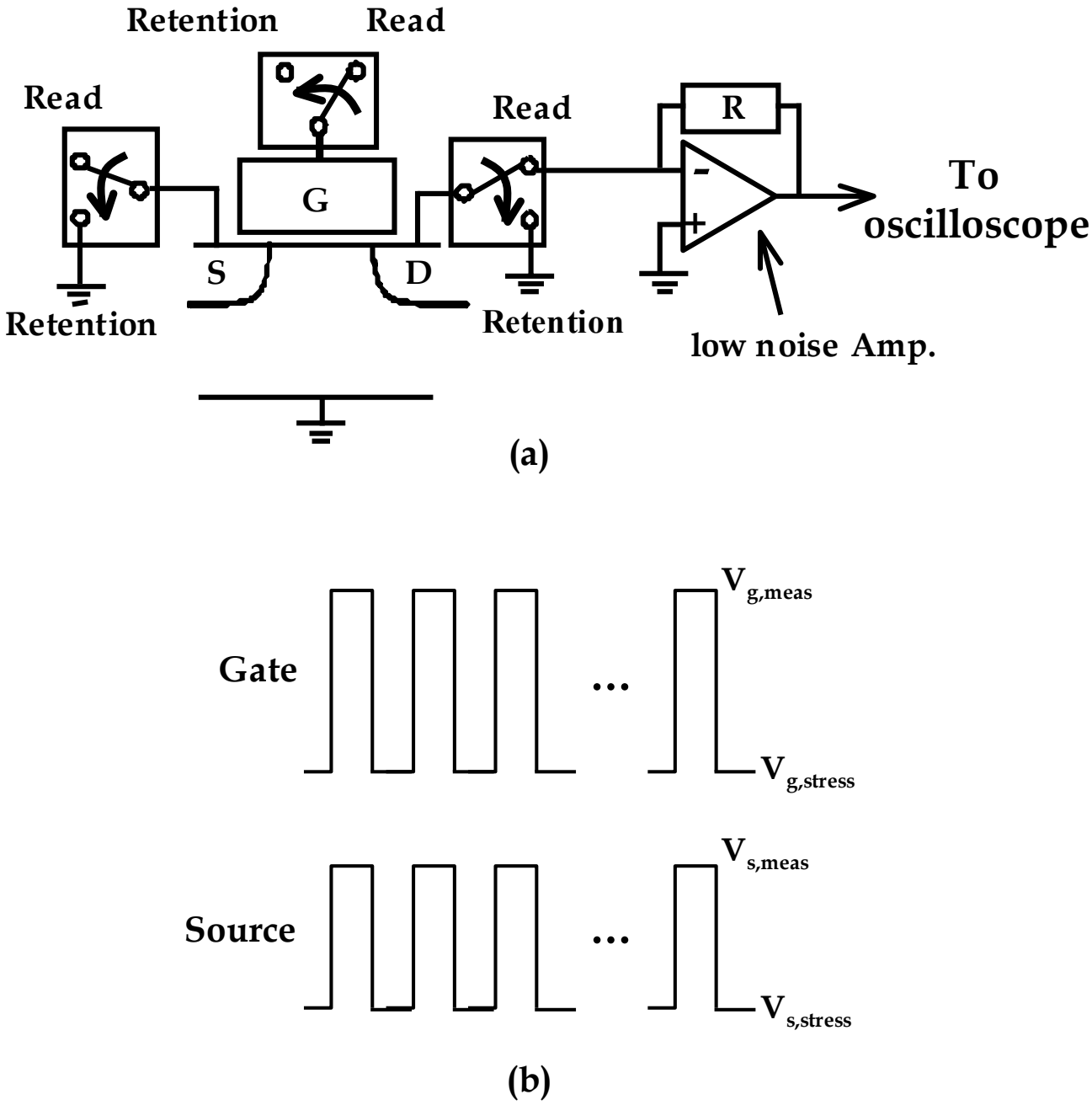


Fig. 7.2 (a) Experimental setup for read current measurement. The measurement consists of two alternating phases, retention phase and reverse read phase. In retention phase, a negative gate voltage (-3.5V) is applied to accelerate nitride charge loss. The sampling rate is 10kHz. (b) The waveforms for the measurement.

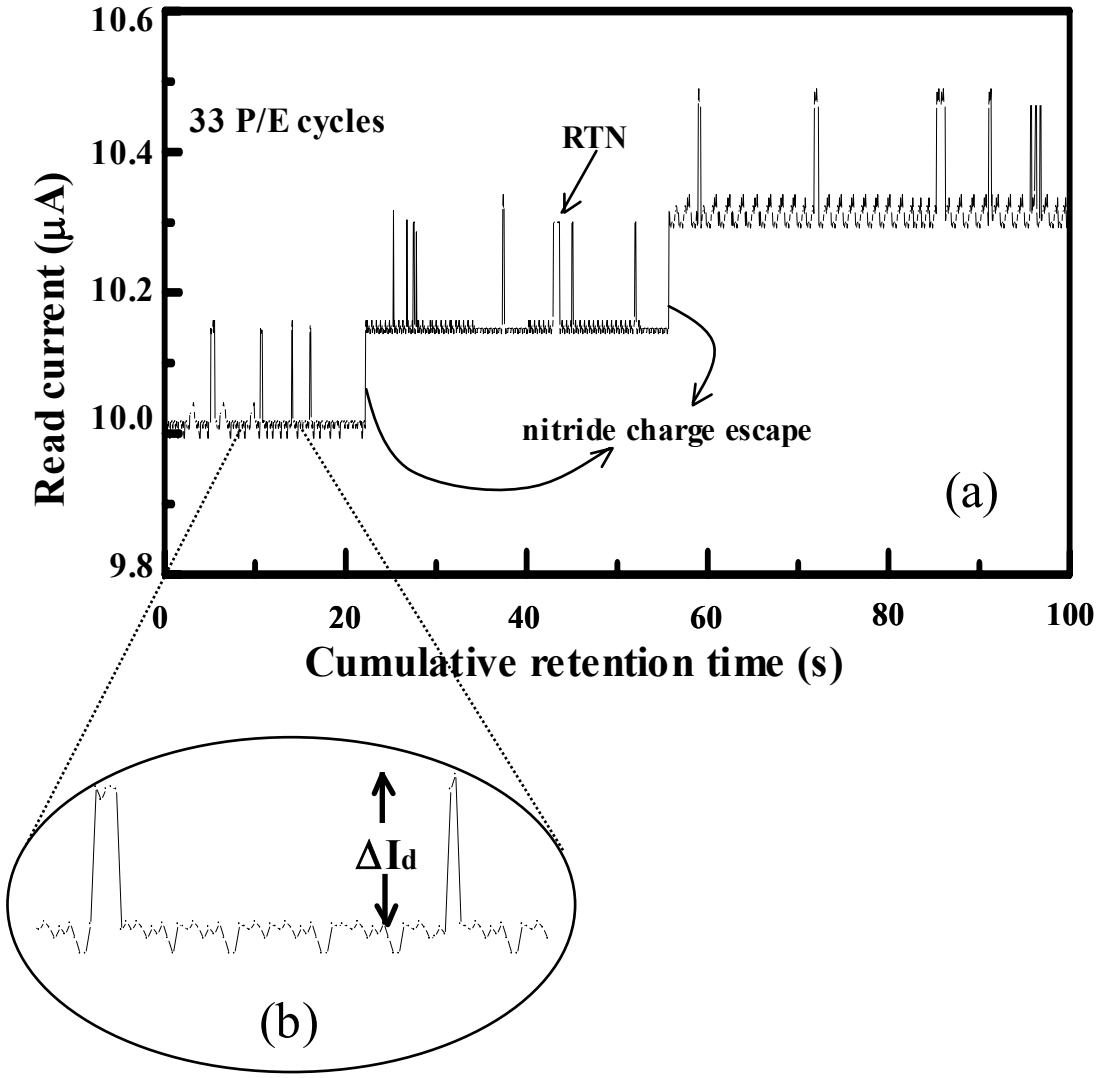
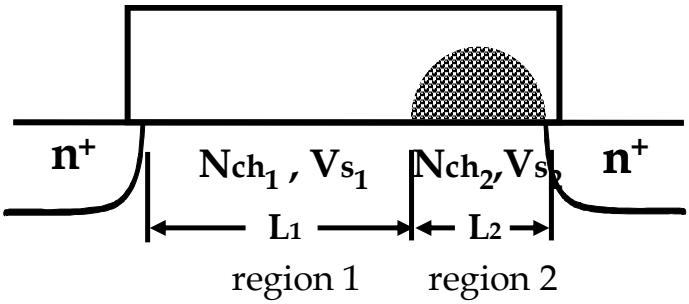


Fig. 7.3 (a) Read current variation with cumulative retention time for program state "01". The P/E cycle number is 33. Two-phase measurement is employed. The program ΔVt is about 2V. (b) Two-level random telegraph noise is highlighted.



$$\Delta I(RTN) \propto \begin{cases} \dfrac{1}{(L_1 + L_2)^2} & \text{for uniform storage} \\ \dfrac{1}{L_2^{\,2}} \end{cases}$$

Fig. 7.4 Simplified two-region model for RTN amplitude in a localized storage SONOS. N_{ch1}(N_{ch2}) and V_{s1}(V_{s2}) represent channel electron number and voltage drop in region 1 (region 2). L_1 and L_2 denote the length of each region.

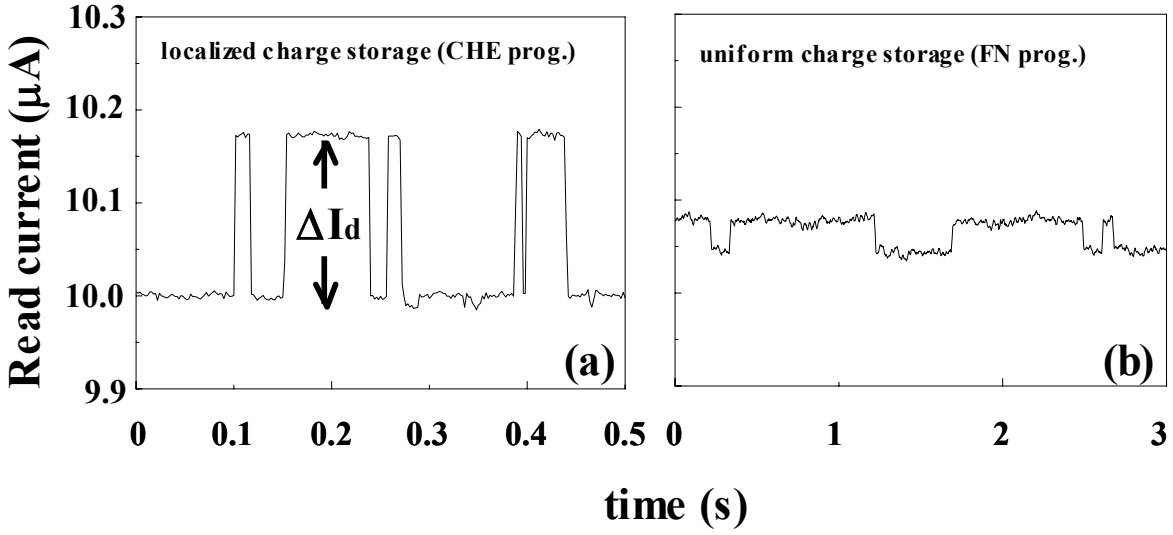
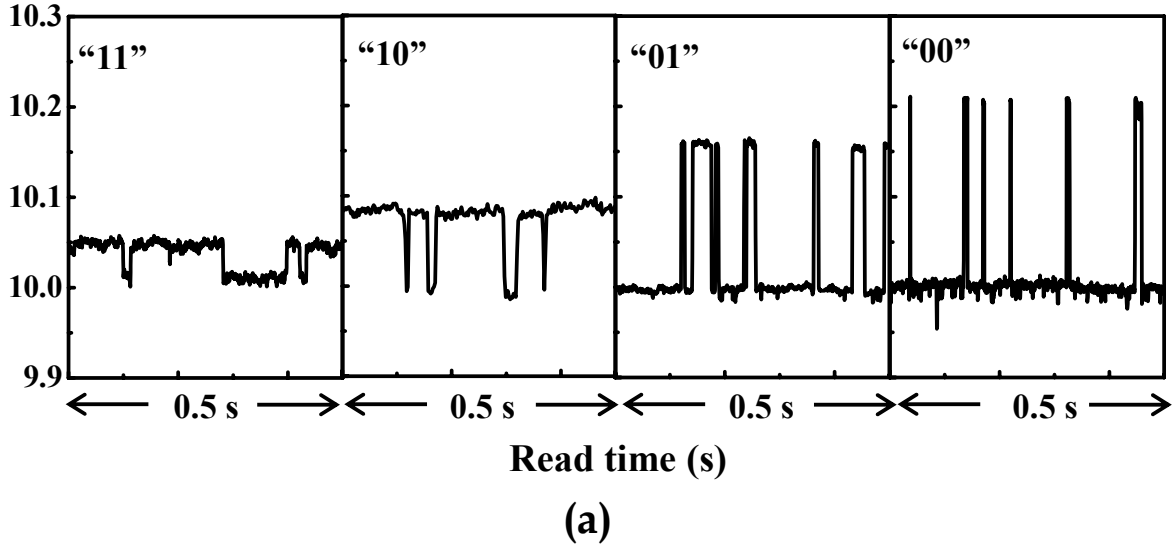


Fig. 7.5 Comparison of the amplitude of RTN in localized charge storage (a) and in uniform charge storage (b). The two cells experience the same P/E cycling procedure



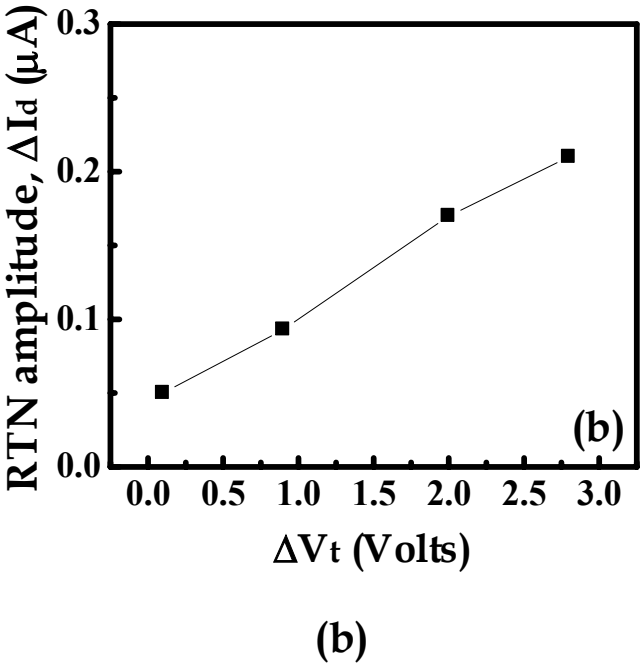


Fig. 7.6 (a) Comparison of RTN amplitude in four different Vt levels of a multi-level cell. ΔVt is 3V for "00", 2V for "01", 1V for "10" and 0.1V for "11". The P/E cycle number is 100. (b) RTN amplitude versus ΔVt.

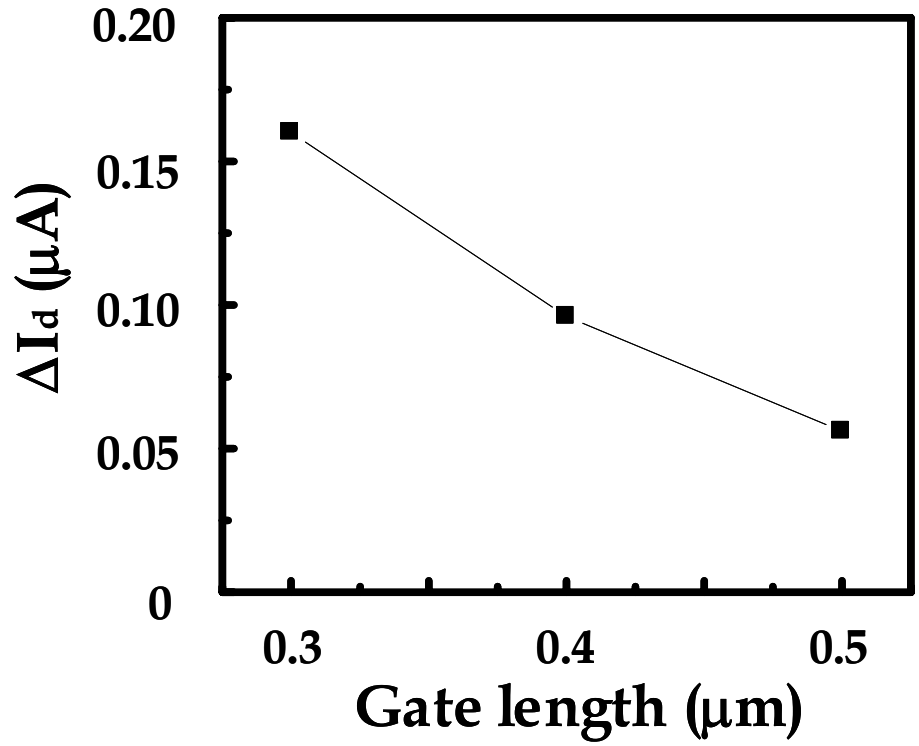


Fig. 7.7 Device scaling effect on the amplitude of RTN. The read current level is 10μA.

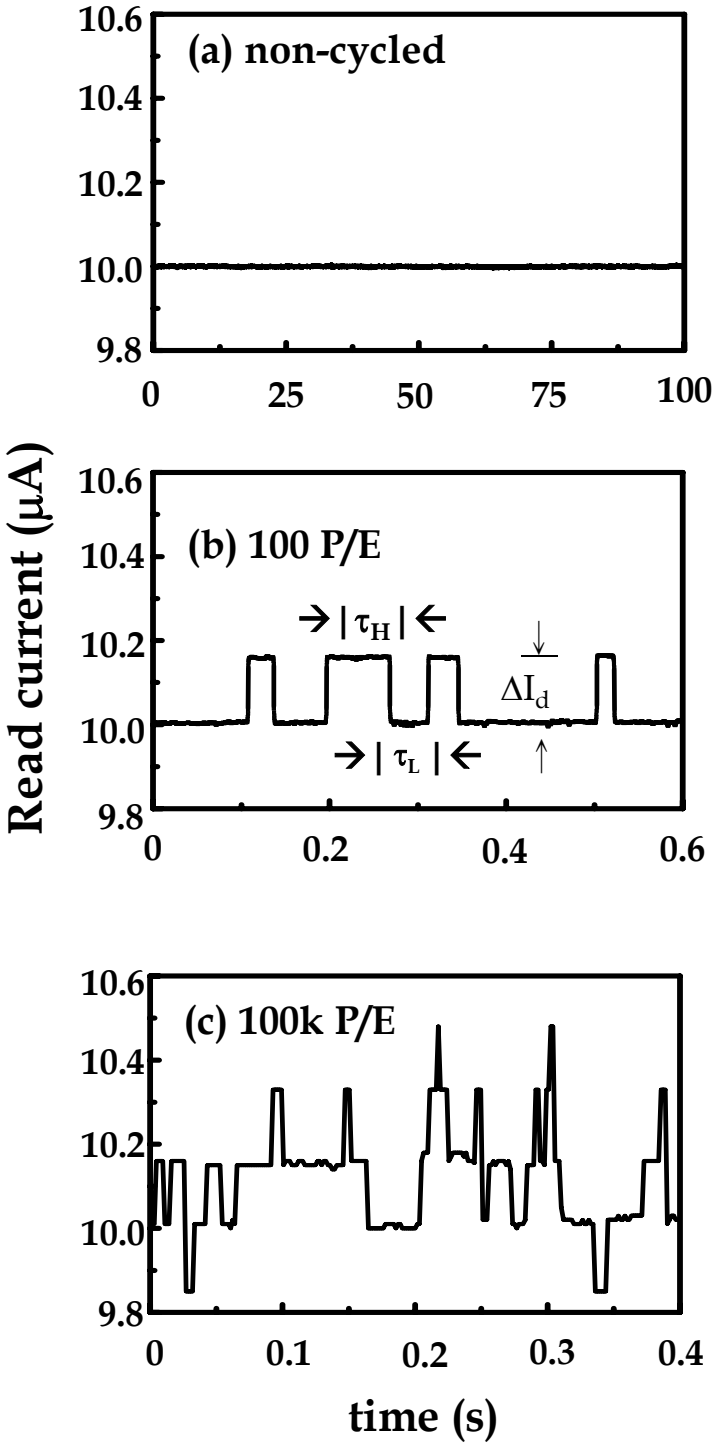


Fig. 7.8 Representative RTN patterns in "01" level in a localized storage, multi-level SONOS cell. (a) no RTN in a non-cycled cell, (b) two-level RTN after 100 P/E cycles (c) multi-level (5-level) RTN after 100k P/E cycles.

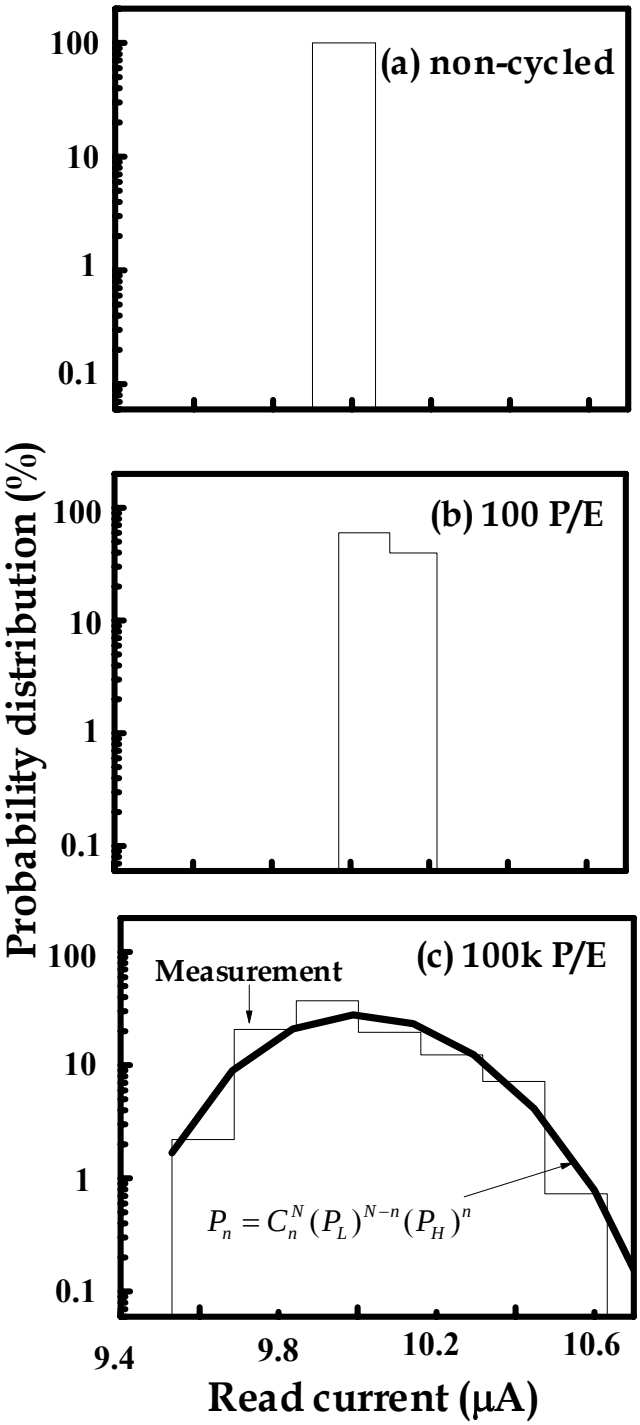


Fig. 7.9 RTN induced read current distributions in "01" level in a localized storage, multi-level SONOS cell. The number of readings for each P/E cycle number is 10^5. The heavy line in (c) is from the model in Fig. 7.10 with P_L=0.6, P_H=0.4 and N=8.

$$P_n = C_n^N (P_L)^{N-n} (P_H)^n$$

$$C_n^N = \frac{N!}{n!(N-n)!}\ ,\ P_L = \frac{<\tau_L>}{<\tau_L> + <\tau_H>},$$

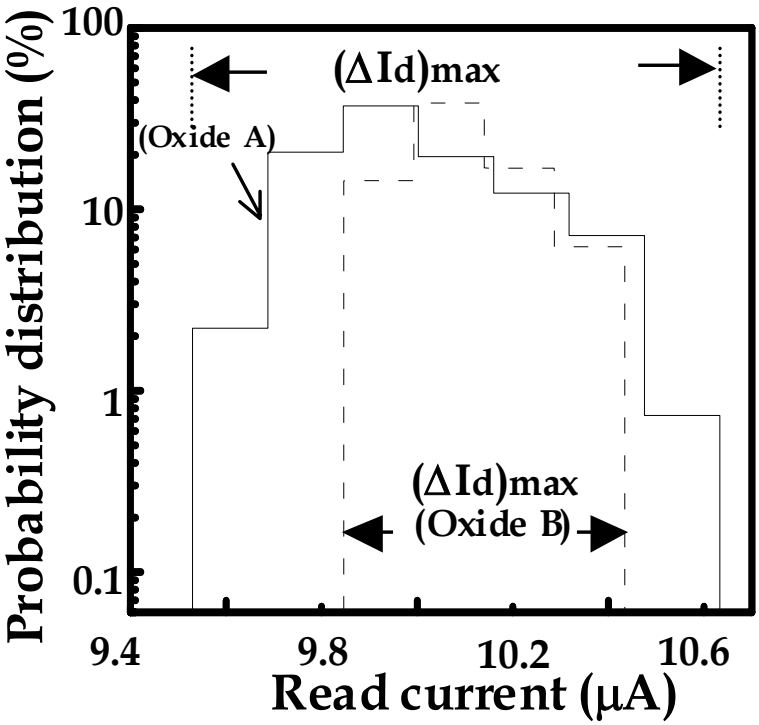
$$P_H = 1 - P_L$$

$$\Delta I_d = n \cdot \Delta I_d (two - level\ RTN)$$

Fig. 7.10 A probability model for RTN induced read current distribution. Pn is the probability of the overlap of n independent two-level RTN waveforms. τ_H and τ_L are the durations of read current in high-current state and in low-current state. P_H and P_L are the corresponding probabilities. The average values of τ_H and τ_L can be extracted from the RTN pattern in Fig. 7.8(b). C_n^N means the number of ways of selecting n independent two-level RTN waveforms from N traps without regard to order.

SONOS	Oxide A	Oxide B
Δ Icp	164pA	72pA
(Δ Id)max	1.11μA	0.58 μ A

(a)



(b)

Fig. 7.11 (a) Charge pumping current after 100k P/E cycles for oxide A and oxide B. (ΔI_d)max in the table denotes the width of RTN induced read current distribution. (b) Measured read current distributions in SONOS cells with oxides A and B.

Chapter 8

Conclusions

In short, this dissertation has involved the reliability issues in a trapping nitride storage Flash memory cell. The subjects that have been comprehensively discussed including the endurance behavior, charge gain in a low-Vt cell, charge loss in a high-Vt cell, programmed charge broadening after P/E stress, bottom oxide thickness effect on retention loss, and RTN caused read current fluctuation. Contributions of each subject in this work are summarized as follows.

First, unlike the window closure in conventional floating gate flash memories, the upward shift in program state or in erase state is observed. Negative oxide charge accumulation and interface state generation are the two possible causes for this interesting shift. The former is identified as the dominant component because the sub-threshold swing has no change before and after stress, implying a little contribution from interface state generation.

Regarding the unique two-bit operation, programmed charge lateral distribution has a large impact on second bit reading and can be probed by using the modified charge pumping technique. Our study shows that the charge distribution of the secondly programmed bit would extend more toward the channel because the field built by the first programmed bit accelerates the channel electrons and cause earlier electron injection into the nitride. In addition, electron injection region also spreads further into the channel with cycle number due to the extension of the stress-created oxide traps in the bottom oxide damaged region.

The analytical reliability models for a low-Vt and a high-Vt state are reported,

individually. Intrinsic charge retentivity is superior before P/E cycling. After P/E stress, a positive erase-state Vt drift with a logarithmic time-dependence is observed. This drift shows a turn-around feature with the cycle numbers, and is found to be insensitive to the bake temperature. Positive oxide charge tunnel de-trapping was proposed to explain this phenomenon. As the word-line or the bit-line bias increases, the positive charge will assist electron tunneling, causing a power law time dependence.

In high-Vt state, charge loss increases with cycle number and shows strong dependence on bake temperature and the applied gate bias. Frenkel-Poole emission followed by (positive) oxide trap assisted tunneling accounts for the retention loss. The square root of the nitride electric field on the Vt loss can confirm the model very well.

In order to further investigate the detailed nitride charge transient behavior during retention, a numerical approach describing the electron FP emission and re-trapping in the nitride and charge leakage through the bottom oxide is developed. The nitride charge leakage current after stress is found to have two-stage feature. The first stage is dominated by the stress-induced (positive) oxide trap assisted featuring a DC-like evolution and a strong stress dependence, while the second stage is dictated by nitride charge Frenkel-Poole emission exhibiting a 1/t time relation and almost no stress dependence. We can use the stress-independence 1/t transient current to extract the nitride trap density. In our case, the extracted nitride trap density is around 7.0×10^{12} $cm^{-2}eV^{-1}$.

Finally, the read current fluctuation due to random telegraph noise in a localized storage, multi-level SONOS flash memory cell is investigated. Threshold

voltage non-uniformity, different programmed Vt level, channel length, and P/E cycling affect the RTN have been characterized. We firstly propose a probability model for the RTN induced read current distribution and obtain an excellent match with measurement result. To mitigate the RTN effect, a better bottom oxide process is necessary.

References

Chapter 1

[1.1] R. Bez, E. Camerlenghi, A. Modelli, and A. Visconti, "Introduction to Flash memories," *Proc. IEEE,* vol. 91, pp. 489-502, 2003

[1.2] P. Pavan, R. Bez, P. Olivo, and E. Zanoni, "Flash memory cells – An overview," *Proc. IEEE,* vol. 85, pp. 1248-1271, 1997

[1.3] J. D. Lee, S. -H. Hur, and J. –D. Choi, "Effects of the floating-gate interferences on NAND Flash memory cell operation," *IEEE Elec. Dev. Lett.*, vol. 23, pp. 264-267, 2003

[1.4] H. Maes and J. F. Van Houdt, "Silicon scaling and its consequences for memory technology," *Proc. IEEE NVSM.*, pp. 7-9, 2003

[1.5] D. Kahng and S. M. Sze, "A Floating Gate and Its Application to Memory Devices," *Bell Syst. Tech. J.,* pp. 1283-1288, 1967

[1.6] Sam Pan, C. C. Yeh, Rich Liu, and C. Y. Lu, "Nonvolatile Memory Challenges toward Gigabit and Nano-scale Era and a Nan-scale Flash Cell: PHINES", *Solid State Devices and Materials*, pp. 152-153, 2002

[1.7] T. Sugizaki, M. Kobayashi, M. Ishidao, H. Minakata, M. Yamaguchi, Y. Tamura, Y. Sugiyama, T. Nakanishi, and H. Tanaka, "Novel Multi-bit SONOS Type Flash Memory Using a High-k Charge Trapping Layer," *VLSI Tech. Dig.,* pp. 27-28, 2003

[1.8] C. H. Lai, Albert Chin, K. C. Chiang, W. J. Yoo, C. F. Cheng, S. P. McAlister, C. C. Chi and P. Wu, "Novel SiO_2/AlN/HfAlO/IrO_2 Memory with Fast Erase, Large ΔVth and Good Retention," *VLSI Tech. Dig.,* pp. 210-211, 2005

[1.9] Jan Van Houdt, "HIGH-K MATERIALS FOR NONVOLATILE MEMORY APPLICATIONS," *Proc. Int. Reliab. Phys. Symp.*, pp. 234-239, 2005

[1.10] R. van Schaijk, M. van Duuren, N. Akil, A. Huerta, S. Beckx, F. Neuilly, Z. Rittersma, M. Slotboom, S. Van Elshocht and J. Wouters, "A Novel SONOS Memory with HfSiON/Si_3N_4/HfSiON Stack for Improved Retention," *Proc. IEEE NVSM.*, pp. 50-51, 2006

[1.11] H. C. Pao and O'Connell, "Memory Behavior of an MNS Capacitor," *Appl. Phys. Lett.,* vol. 12, pp. 260-262, 1968

[1.12] H. A. R. Wegener, A. J. Lincoln, H. C. Pao, M. R. O'Connel, and R. E. Oleksiak, "The Variable Threshold Transistor, A New Electrically-Alterable, Non-destructive Read-only Storage Device," *IEDM Tech. Dig.*, pp. 420-423, 1967

[1.13] Mehran Aminzadeh, Shinji Nozaki, R. V. Giridhar, "Conduction and charge trapping in polysilicon-silicon nitride-oxide-silicon structures under positive gate bias," *IEEE Trans. on Elec. Dev.*, vol. 35, pp. 459-467, 1988

[1.14] James A. Topich, "Long-term retention of SNOS nonvolatile memory devices," *IEEE Trans. on Elec. Dev.*, vol. 31, pp. 1908-1910, 1984

[1.15] Y. –N. Tan, Wai-Kin Chim, B. J. Cho, and Wee-Kiong Choi, "Over-erase phenomenon in SONOS-type flash memory and its minimization using a hafnium oxide charge storage layer," *IEEE Trans. on Elec. Dev.*, vol. 51, pp. 1143-1147, 1984

[1.16] Myung Kwan Cho, and Dae M. Kim, "High performance SONOS memory cells free of drain turn-on and over-erase: Compatibility issue with current flash technology," *IEEE Elec. Dev. Lett.*, vol. 21, pp. 399-401, 2000

[1.17] F. R. Libsch, A. Roy, and M. H. White, "Charge transport and storage of low programming voltage SONOS/MONOS memory devices," *Sol. State Elec.*, vol. *33*, pp. 105-109, 1990

[1.18] T. K. Chan, K.K. Young, and C. Hu, "A True Single-transistor Oxide-Nitride-Oxide EEPROM Device", *IEEE Elec. Dev. Lett.*, Vol.8, pp. 93-95, 1987

[1.19] B. Eitan, P. Pavan, I.Bloom, E. Aloni, A. Frommer, and D. Finzi, " NROM: A Novel Localized Trapping, 2-Bit Nonvolatile Memory Cell", *IEEE Electron Dev. Lett.*, Vol.21., No. 11, pp. 543-545, 2000

[1.20] W.J. Tsai, N.K. Zous, C.J. Liu, C.C. Liu, C.H. Chen, Tahui Wang, Sam Pan, and Chih-Yuan Lu, "Data Retention Behavior of a SONOS Type Two-Bit Storage Flash Memory Cell", *IEDM Tech. Dig.*, pp. 719-722, 2001

[1.21] Tahui Wang, W. J. Tsai, S. H. Gu, C. T. Chan, C. C. Yeh, N. K. Zous, T C. Lu, Sam Pan, and Chih-Yuan Lu, "Reliability Models of Data Retention and Read-Disturb in 2-bit Nitride Storage Flash Memory Cells (Invited Paper)," *IEDM Tech. Dig.*, pp. 169-172, 2003

[1.22] H. T. Lue, T. H. Hsu, M. T. Wu, K. Y. Hsueh, Rich Liu, and Chih-Yuan Lu, "Studies of the cReverse Read Method and Second-Bit Effect of 2-Bit/Cell Nitride-Trapping Device by Quasi-Two-Dimensional Model," *IEEE Trans. on Elec. Dev.*, vol. 53, pp. 119-125, 2006

[1.23] W. P. Lu and S. H. Gu, "A Novel Hole Annealing Method," patent pending, Taiwan, Republic of China, 2005

[1.24] Yakov Roizin, Evgeny Pikhay, and Micha Gutman, "Suppression of Erased State V_t Drift in Two-Bit Per Cell SONOS Memories," *IEEE Elec. Dev. Lett.*, vol.

26, pp. 35-37, 2005

[1.25] Takashi Hori and Takatoshi Yasui, and Susumu Akamatsu, "Hot-Carrier Effects in MOSFET's with Nitrided-Oixde Gate-Dielectrics Prepared by Rapid Thermal Processing," *IEEE Trans. on Elec. Dev.*, vol. 39, pp. 134-147, 1992

[1.26] Jiankang Bu and Marvin H. White, "Improvement in retention reliability of SONOS nonvolatile memory devices by two-step high temperature deuterium anneals," *Proc. Int. Reliab. Phys. Symp.*, pp. 52-56, 2001

[1.27] Kuo-Hong Wu, Hua-Ching Chien, Chih-Chiang Chan, Tung-Sheng Chen, and Chin-Hsing Kao, " SONOS device with tapered bandgap nitride layer,"*IEEE Trans. on Elec. Dev.*, vol. 52, pp. 987-992, 2005

[1.28] Yu Wang and M. H. White, "An Analytical retention Model for SONOS Nonvolatile Memory Devices in the Excess Electron State," *Semiconductor Dev. Research Sym.*, pp. 157-157, 2003

[1.29] S. H. Gu, Tahui Wang, Wen-Pin Lu, Yen-Hui Joseph Ku, and Chih-Yuan Lu, "Numerical Simulation of Bottom Oxide Thickness Effect on Charge Retention in SONOS Flash Memory Cells with Fowler-Nordheim Programming," *IEEE Trans. on Elec. Dev.*, vol. 19, pp. 826-665, 2006

[1.30] S.H. Gu, C.W. Li, Tahui Wang, W.P. Lu, K.C. Chen, Joseph Ku, and Chih-Yuan Lu, "Read Current Instability Arising from Random Telegraph Noise in Localized Storage, Multi-Level SONOS Flash Memory," *sumitted to IEDM Tech. Dig.*, 2006

Chapter 2

[2.1] P. Cappelletti, C. Golla, P. Olivo, and E. Zanoni, "Flash MEMORIES," in

KLUWER ACADEMIC, 2000

[2.2] B. Euzent, N. Boruta, J. Lee, C. Jeng, "Reliability aspects of a floating gate EEPROM", *Proc. Int. Reliability Phys. Symp.*, pp. 11-16, 1981

[2.3] N. Mielke, A. Fazio, and H. C Liou, "Reliability comparison of FLOTOX and textured polysilicon EEPROM's," *Proc. Int. Reliability Phys. Symp.*, pp. 85-90, 1987

[2.4] C. W. Tsai, S. H. Gu, L. P. Chiang, T. Wang, Y. C. Liu, L. S. Huang, M. C. Wang and L. C. Hsia, "Valence-Band Tunneling Enhanced Hot Carrier Degradation in Ultra-Thin Oxide nMOSFETs", *IEDM Tech. Dig.*, pp. 139-142, 2000

[2.5] Tahui Wang, Senior Member, IEEE, Lu-Ping Chiang, Nian-Kai Zous, Charng-Feng Hsu, Li-Yuan Huang, and Tien-Sheng Chao," A Comprehensive Study of Hot Carrier Stress-Induced Drain Leakage Current Degradation in Thin-Oxide n-MOSFET's," *IEEE Trans. on Elec. Dev.*, vol. 46, pp. 1877-1882, 1999

[2.6] Tahui Wang, L. P. Chiang, T. E. Chang, N. K. Zous, K. Y. Shen, and C. Huang, "Reliability comparison of FLOTOX and textured polysilicon EEPROM's," *Proc. Int. Reliability Phys. Symp.*, pp. 292-295, 1997

[2.7] E. F. Runnion, S. M., IV Gladstone, R. S. Scott, D. J. Dumin, L. Lie, and J. Mitros, " Limitations on oxide thicknesses in flash EEPROM applications", *Proc. Int. Reliability Phys. Symp.*, pp. 93-99, 1996

Chapter 3

[3.1] Y. W. Chang, T. C. Lu, Sam Pan, and Chih-Yuan Lu, "Modeling for the 2nd-Bit Effect of a Nitride-Based Trapping Storage Flash EEPROM Cell Under Two-Bit

Operation," *IEEE Elec. Dev. Lett.*, Vol. 25, pp. 95, 2004

[3.2] W. J. Tsai, N. K. Zous, M. H. Chou, Smile Huang, H. Y. Chen, Y. H. Yeh, M. Y. Liu, C. C. Yeh, T. Wang, Joseph Ku, and Chih-Yuan Lu, "Cause of Erase Speed Degradation during Two-Bit Per Cell Operation of A Trapping Nitride Storage Flash Memory Cell," *Proc. Int. Reliability Phys. Symp.*, pp. 522-526, 2004

[3.3] L. Larcher, G. Verzellesi, P. Pavan, E. Lusky, I. Bloom and B. Eitan, "Impact of Programming Charge Distribution on Threshold Voltage and Subthreshold Slope of NROM Memory Cells," *IEEE Trans. on Elec. Dev.*, pp.1939-1946, 2002

[3.4] S. H. Gu, M. T. Wang, C. T. Chan, N. K. Zous, C. C. Yeh, W. J. Tsai, T. C. Lu, Tahui Wang, Joseph Ku, and Chih-Yuan Lu, "Investigation of Programmed Charge Lateral Spread in a Two-bit Nitride Storage Flash Memory Cell by Using a Charge Pumping Technique," *Proc. Int. Reliability Phys. Symp.*, pp. 639-640, 2004

[3.5] S. H. Gu, Tahui Wang, W. P. Lu, Wenchi Ting, Y. Hui Joseph Ku, and Chih Yuan Lu, "Characterization of Programmed Charge Lateral Distribution in a Two-bit Storage Nitride Flash Memory Cell by Using a Charge Pumping Technique," *IEEE Trans. on Elec. Dev.*, vol. 53, pp. 103-108, 2006

[3.6] Chun Chen and T. P. Ma, "Direct Lateral Profiling of Hot-Carrier-Induced Oxide Charge and Interface Traps in Thin Gate MOSFET's," *IEEE Trans. on Elec. Dev.*, vol. 45, pp.512-520, 1998

[3.7] W. Chen and T. P. Ma, "Oxide charge buildup and spread-out during channel-hot-carrier injection in N-MOSFET's," *IEEE Elec. Dev. Lett.*, vol. 13, pp. 319-321, 1992

[3.8] M. G. ANCONA, N. S. SAKS, and D. Mccarthy, "Lateral Distribution of Hot-Carrier-Induced Interface Traps in MOSFET's," *IEEE Trans. on Elec. Dev.*, vol. 35, pp. 2221-2228, 1988

[3.9] N. K. Zous, Y. J. Chen, C. Y. Chin, W. J. Tsai, T. C. Lu, M. S. Chen, W. P. Lu, T. Wang, Sam Pan, and C. Y. Lu, "An Endurance Evaluation Method for Flash EEPROM," *IEEE Trans. on Elec. Dev.*, vol. 51, pp. 720-725, 2004

[3.10] J. Wu, L. F. Register, and E. Rosenbaum, "Trap-Assisted Tunneling Current Through Ultra-Thin Oxide," *Proc. Int. Reliability Phys. Symp*., pp. 389-395, 1999

Chapter 4

[4.1] Tahui Wang, W. J. Tsai, S. H. Gu, C. T. Chan, C. C. Yeh, N. K. Zous, T C. Lu, Sam Pan, and Chih-Yuan Lu, "Reliability Models of Data Retention and Read-Disturb in 2-bit Nitride Storage Flash Memory Cells (Invited Paper)," *IEDM Tech. Dig*., pp. 169-172, 2003

[4.2] B. Eitan, P. Pavan, I.Bloom, E. Aloni, A. Frommer, and D. Finzi, " NROM: A Novel Localized Trapping, 2-Bit Nonvolatile Memory Cell", *IEEE Elec. Dev. Lett*., vol.21., No. 11, pp. 543-545, 2000

[4.3] D. J. Dumin and J. Maddux, "Correlation of Stress-Induced Leakage Current in Thin Oxide with Trap Generation Inside the Oxide," *IEEE Trans. on Elec. Dev*., vol. 40, pp986, 1993

[4.4] K. Lehovec and A. Fedotowsky, "A Charge retention of MNOS devices limited by Frenkel-Poole detrapping", *J. Appl. Phys*., Vol. 32, pp. 335, 1978.

[4.5] B. Riccò, G Gozzi, and M. Lanzoni, " Modeling and Simulation of Stress-Induced Leakage Current in Ultrathin SiO_2 Films," *IEEE Trans. on Elec. Dev*., vol. 45, pp. 1554-1560, 1998.

[4.6] W. J. Tsai, N. K. Zous, C. J. Liu, C. C. Liu, C. H. Chen, Tahui Wang, Sam Pan and Chih-Yuan Lu, "Data Retention Behavior of a SONOS Type Two-Bit Storage Flash Memory Cell," *IEDM Tech. Dig*., pp. 719-722, 2001.

[4.7] A. Roy and M. H. White, " A New Approach to Study Electron and Hole Charge Separation at the Semiconductor-Insulator Interface," *IEEE Trans. on*

Elec. Dev., vol. 37, pp. 1504-1513, 1990

[4.8] W. J. Tsai, C. C. Yeh, N. K. Zous, C. C. Liu, S. K. Cho, Tahui Wang, Samuel C. Pan and Chin-Yuan Lu, " Positive Oxide Charge-Enhanced Read Disturb in a Localized Trapping Storage Flash Memory Cell," *IEEE Trans. on Elec. Dev.*, vol. 51, pp. 434-439, 2004

[4.9] Tahui Wang, N. K. Zous, J. L. Lai and C. Huang, "Hot hole Stress Induced Leakage Current (SILC) Transient in Tunnel Oxides," *IEEE Elect. Dev. Lett.*, vol. 19, pp. 411-413, 1998.

[4.10] Taihu Wang, N. K. Zous, and C. C. Yeh, "Role of positive trapped charge in stress-induced leakage current for Flash EEPROM devices," *IEEE Trans. on Elec. Dev.*, pp.1910-1916, 2002

[4.11] S. Jeon, J. H. Han, J. Lee, S. Choi, H. Hwang, and Chungwoo Kim, "Impact of metal work function on memory properties of charge-trap flash memory devices using fowler-nordheim P/E mode," *IEEE, Elect. Dev. Lett.*, vol. 27, pp. 486-488, 2006

Chapter 5

[5.1] L. Lundkvist, I. Lundstörm, and C. Svensson," Discharge of MNOS Structures," *Sol. State Elec.*, vol. 16, pp. 811-823, 1973.

[5.2] L. Lundkvist, C. Svensson, and B. Hansson," Discharge of MNOS Structures at Elevated Temperatures," *Sol. State Elec.*, vol. 19, pp. 221-227, 1976.

[5.3] P. J. McWhorter, S. L. Miller, and T. A. Dellin, "Modeling the memory retention characteristics of silicon-nitride-oxide-silicon nonvolatile transistors in a varying thermal environment," *J. Appl. Phys.*, vol. 68, pp. 1902-1909, 1990.

[5.4] K. Lehovec and A. Fedotowsky, "Charge Retention of MNOS Devices Limited

by Frenkel-Poole Detrapping", *Appl. Phys. Lett.*, vol. 32, pp. 335-338, 1978.

[5.5] Ross A. Williams and Moiz M. E. Beguwala, "The Effect of Electrical Conduction of Si_3N_4 on the Discharge of MNOS Memory Transistors," *IEEE Trans. on Elec. Dev.*, vol. 25, pp.1019-1023, 1978.

[5.6] Y. Yang, M. H. White, "Charge Retention of Scaled SONOS Nonvolatile Memory Devices at Elevated Temperatures," *Sol. State Elec.,* vol. 44, pp. 948-958, 2000.

[5.7] Yu Wang and M. H. White, "An Analytical retention Model for SONOS Nonvolatile Memory Devices in the Excess Electron State," *Sol. State Elec.*, vol. 49, p.97, 2005.

[5.8] Gowrishankar L. Chindalore, C. T. Swift, and David Burnett, "A New Combination-Erase Technique for Erasing Nitride Based (SONOS) Nonvolatile Memories," *IEEE, Elect. Dev. Lett.*, vol. 24, p.257-259, 2003

[5.9] M. Specht, U. Dorda, L.Dreeskornfeld, J.Kretz, F. Hofmann, M.Städele, R.J.Luyken, W.Rösner, H.Reisinger, E. Landgraf, T.Schulz, J.Hartwich, R. Kömmling, L.Risch, "20 nm tri-gate SONOS memory cells with multi-level operation," *VLSI Tech. Dig.*, 2004

[5.10] V. A. Gritsenko, "Nonstationary Electrons and Holes Transport by Depolarization of MNOS Structures: Experiment and Numerical Simulation," *Microelectronics*, vol. 16, pp.42-50. 1987

[5.11] W. Schockley and W. T. Read, Jr., "Statistics of the Recombinations of Holes and Electrons," *Phys. Rev.*, vol. 87, pp. 835-842, 1952.

[5.12] P. C. Arnett, "Transient conduction in insulators at high fields," *J. Appl. Phys.*, vol. 46, pp. 5236-5243, 1975

[5.13] Taihu Wang, N. K. Zous, and C. C. Yeh, "Role of positive trapped charge in stress-induced leakage current for Flash EEPROM devices," *IEEE Trans. on Elec. Dev.*, pp.1910-1916, 2002

[5.14] V. Vasudevan and J. Vasi, "A simulation of multiple trapping model for continuous time random walk transport," *J. Appl. Phys.*, vol. 74, pp. 3224-3230,

1993

[5.15] D. Debuf, "General theory of carrier lifetime in semiconductors with multiple localized states," *J. Appl. Phys.*, vol. 96, pp. 6454-6469, 2004

[5.16] K. A. Nasyrov, V. A. Gritsenko, M. K. Kim, H. S. Chae, S. D. Chae, W. I. Ryu, J. H. Sok, J.-W. Lee, and B. M. Kim, "Charge Transport Mechanism in Metal-Nitride-Oxide-Silicon Structures," *IEEE Elec. Dev. Lett.*, vol. 23, pp. 336-338, 2002

[5.17] B. De Salvo, G. Ghibaudo, G. Pananakakis, P. Masson, T. Baron, N. Buffet, A. Fernandes, and B. Guillaumot, "Experimental and theoretical investigation of nano-crystal and nitride-trap memory devices," *IEEE Trans. Elec. Dev.*, vol. 48, pp.1789-1799, 2001

[5.18] M. H. White, D. A. Adams, J. R. Murray, S. Wrazien, Yijie Zhao, Yu Wang, B. Khan, W. Miller, R. Mehrotra, "Characterization of scaled SONOS EEPROM memory devices for space and military systems," *Proc. IEEE NVSM.*, pp. 51-59, 2004

[5.19] S. Kamohara, D. Park, and C. Hu, "Deep-Trap SILC(Stress Induced Leakage Current) Model For Nominal and Weak Oxides," *Proc. Int. Reliability Phys. Symp.*, pp. 57-61, 1998

[5.20] T. K. Kang, M. J. Chen, C. H. Liu, Chang Y.J., and S. K. Fan, "Numerical confirmation of inelastic trap-assisted tunneling (ITAT) as SILC mechanism," *IEEE Trans. on Elec. Dev*, vol. 48, pp.2317-2322, 2001

[5.21] M. Silver and L. Cohen, "Monte Carlo simulation of anomalous transit-time dispersion of amorphous solids," *Phys. Rev. B.*, vol. 15, pp. 3276-3278, 1977.

[5.22] C. Main, S. Reynolds, and R. Brüggemann, "Decay from steady-state photocurrent in amorphous semiconductors," *Phys.stat. sol.(c)*, vol. 5, pp. 1194-1207, 2004

[5.23] Curtis F. Gerald, "Applied Numerical Analysis Second Efition," in *Addison-Wesley*, 1977

[5.24] T. R. Oldham, A. J. Lelis, and F. B. Mclean, "Spatial dependence of trapped

holes determined from tunneling analysis and measured annealing," *IEEE Trans. Nucl. Dev.*, vol. 40, pp. 1203-1207, 1986

[5.25] Tahui Wang, W. J. Tsai, S. H. Gu, C. T. Chan, C. C. Yeh, N. K. Zous, T. C. Lu, Sam Pan and C. Y. Lu, "Reliability Models of Data Retention and Read-Disturb in 2-bit Nitride Storage Flash Memory Cells," *IEDM Tech. Dig.*, pp. 169-172, 2003

[5.26] K. I. Lundström, and C. M. Svensson, "Properties of MNOS structures," *IEEE Trans. Elec. Dev.*, Vol. 19, pp.826-836, 1972.

Chapter 6

[6.1] K. I. Lundström, and C. M. Svensson, "A unified model for the flicker noise in metal-oxide-semiconductor field-effect transistors," *IEEE Trans. on Elec. Dev.*, vol. 19, pp. 826-665, 1972

[6.2] R. E. Paulsen, R. R. Siergiej, M. L. French, and M. H. White, "A unified model for the flicker noise in metal-oxide-semiconductor field-effect transistors," *IEEE Elec. Dev. Lett.*, vol. 13, pp. 627-338, 1992

[6.3] Y. Yang, M. H. White, "Charge Retention of Scaled SONOS Nonvolatile Memory Devices at Elevated Temperatures," *Sol. State Elec.*, vol. 44, pp. 948-958, 2000

[6.4] Yu Wang and M. H. White, "An Analytical retention Model for SONOS Nonvolatile Memory Devices in the Excess Electron State," *Sol. State Elec.*, vol. 49, p.97, 2005

[6.5] S. H. Gu, Tahui Wang, Wen-Pin Lu, Yen-Hui Joseph Ku, and Chih-Yuan Lu, "Extraction of Nitride Trap Density from Stress Induced Leakage Current in Silicon-Oxide-Nitride-Oxide-Silicon (SONOS) Flash Memory," *Appl. Phys. Lett.*,

vol. 12, pp. 260-262, 1968

Chapter 7

[7.1] K. C. Smith, "The prospects for multivalued logic: a technology and application view, " *IEEE Trans. Computers,* vol. C-30, pp. 619-634, 1981.

[7.2] M. Horiguchi, M. Aoki, Y. Nakagome, S. Ikenaga, and K. Shimohigashi, "An experimental large-capacity semiconductor file memory using 16-level/cell storage," *IEEE J. Solid-State Circuits,* vol. SC-23, pp. 27-33, 1988.

[7.3] E. Harari, "Flash EEPROM memory system having multi-level storage cells," U.S. Patent #5,043,940, 1991.

[7.4] B. Eitan, "4-bit per cell NROM reliability," *IEDM Tech. Dig.*, pp. 539-542, 2005.

[7.5] Y. Polansky, A. Lavan, R. Sahar, O. Dadashev, Y. Betser, G. Cohen, E. Maayan, B. Eitan, T. C. T. Chen, C. H. Chang, C. Y. Liao, C. K. Chen, W. C. Ho, Y. S., W. P. Lu, W. C. Ting, F. L. Ni, Y. H. Joseph Ku, Chih-Yuan Lu, "A 4 bits per cell NROM 1Gb Data Storage Memory," *ISSCC Dig. Tec*h., pp. 34-38, 2006

[7.6] K. S. Ralls, W. J. Skocpol, L. D. Jackel, R. E. Howard, L. A. Fetter, R. W. Epworth, and D. M. Tennant, "Discrete resistance switching in submicron silicon inversion layers: individual interface traps and low frequency (1/f?) noise," *Phys. Rev. Lett.*, vol. 52, pp. 228-, 1984

[7.7] M. J. Kirton and M. J. Uren, "Noise in solid state microstructure: a new perspective on individual defects, interface states, and low-frequency (1/f) noise," Adv. in Phys., vol. 38, pp. 367-468, 1989.

[7.8] H. Kurata, "20 nm tri-gate SONOS memory cells with multi-level operation," *VLSI Tech. Dig.*, 2006

[7.9] S. H. Gu, C. W. Li, Tahui Wang, W. P. Lu, K. C. Chen, Y. Hui Joseph Ku, and Chih Yuan Lu, "Read Current Instability Arising from Random Telegraph Noise in Localized Storage, Multi-Level SONOS Flash Memory," to be published in *IEEE International Electron Devices Meeting (IEDM)*, San Franciso, U.S.A., 2006

[7.10] M. H. Tsai and T. P. Ma, "The Impact of Device Scaling on the Current Fluctuations in MOSFET's," *IEEE Trans. on Elec.* Dev., vol. 41, pp. 2061-2068, 1994

Appendix

[1] Tahui Wang et al.,, "Reliability Models of Data Retention and Read-Disturb in 2-bit Nitride Storage Flash Memory Cells (Invited Paper)," *IEDM Tech. Dig.*, pp. 169-172, 2003

Reliability Models of Data Retention and Read-Disturb in 2-bit Nitride Storage Flash Memory Cells (Invited Paper)

Tahui Wang*†, W.J. Tsai†*, S.H. Gu*, C.T. Chan*, C.C. Yeh†*, N.K. Zous†, T.C. Lu†, Sam Pan†, and C.Y. Lu†

†Macronix International Co. Ltd, No. 16, Li-Hsin Road, Science-Based Industrial Park, Hsin-Chu, Taiwan
*Dept. of Electronics Engineering, National Chiao-Tung University, Hsin-Chu, Taiwan
email: tahuiwang@faculty.nctu.edu.tw, tahuiwang@mxic.com.tw

Abstract

The reliability issues of two-bit storage nitride flash memory cells including low-V_t state threshold voltage instability, read-disturb, and high-V_t state charge loss will be addressed. Responsible mechanisms and reliability models will be discussed. Our study shows that the cell reliability is strongly dependent on operation methods and process conditions.

Introduction

Nitride flash memory cells with 2-bit storage have received much interest due to their simpler fabrication process, smaller bit size and absence of drain induced turn-on. These cells have an ONO gate dielectric stack, as shown in Fig. 1(a). The signal charge of each bit is stored in nitride above the source and the drain junctions, respectively. Various program and erase methods have been proposed for 2-bit/cell operation (NROM [1] and PHINES [2], Fig. 1(b)). In this paper, the reliability issues of the nitride flash cells will be reviewed. Responsible mechanisms and possible solutions will be discussed. Typical bias conditions used in this work for a NROM cell are given in Table 1.

Low-V_t State Reliability Issues

(a) room temperature (RT) V_t drift

The NROM cell exhibits excellent data retention behavior before P/E cycling. After P/E stress, a positive erase-state V_t drift with a logarithmic time-dependence is observed (Fig. 2). This V_t drift *does not have temperature dependence*. Positive oxide charge tunnel detrapping was proposed [3] to explain this phenomenon. In order to measure positive oxide charge detrapping current directly (Fig. 3(a)), large area devices were fabricated with two different ONO processes (A and B). Process B is known to have better oxide endurance. The substrate current (I_b) before and after FN stress was measured in these two samples. According to the hole tunneling front model [4], the post-stress substrate current resulting from positive oxide charge detrapping follows a 1/t time-dependence,

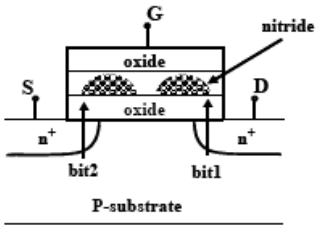
$$I_b(t)=A\frac{Q_{ox}}{\alpha_h}\,t^{-1} \quad (1)$$

$$\alpha_h=4\pi\sqrt{2m\phi_{ox}}/h \quad (2)$$

where Q_{ox} is the positive oxide charge density, ϕ_{ox} denotes the energy barrier of positive trapped charges and A is the area of the device. Fig. 3(b) shows the measured pre-stress and post-stress substrate currents versus time. Note that process B exhibits a smaller post-stress substrate current because of less positive oxide charge creation. The corresponding V_t drift thus has a logarithmic time-dependence,

$$\Delta V_t(t)=2.3\frac{1}{C_{ONO}}\frac{hQ_{ox}}{4\pi\sqrt{2m\phi_{ox}}}\log(t) \quad (3)$$

Fig. 4 shows the measured V_t drift in two 10k P/E cycled cells. By comparing the two ONO processes, a correlation between the V_t drift and I_b is obtained.



(a)

Cell Operation	NROM[1]	PHINES[2]
Program	CHE	BTBT Hot Hole
Erase	BTBT Hot Hole	FN

(b)

Fig. 1 (a) Schematic diagram of a two-bit storage nitride flash cell. Typical thickness for the ONO layers is 9nm (top oxide), 6nm (nitride) and 7nm, respectively. (b) Operation methods of two nitride flash cells, NROM [1] and PHINES [2].

Table 1 Typical bias conditions for program, erase and read in a NROM cell.

		Program (CHE)	Erase (BTBT HH)	Read (Reverse)
Bit 1	V_g	11V	-3V	3V
	V_d	5V	7V	0V
	V_s	0V	0V	>1.5V
Bit 2	V_g	11V	-3V	3V
	V_d	0V	0V	>1.5V
	V_s	5V	7V	0V

[2] S. H. Gu et al., "Investigation of Programmed Charge Lateral Spread in a Two-bit Nitride Storage Flash Memory Cell by Using a Charge Pumping Technique," *Proc. Int. Reliability Phys. Symp.*, pp. 639-640, 2004

Investigation of Programmed Charge Lateral Spread in a Two-bit Storage Nitride Flash Memory Cell by Using a Charge Pumping Technique

S.H. Gu, M.T. Wang*, C.T. Chan*, N.K. Zous†, C.C. Yeh†, W.J. Tsai†, T.C. Lu†, Tahui Wang*, Joseph Ku†, and Chih-Yuan Lu†*

†Macronix International Co., Ltd, No. 16, Li-Hsin Road, Science-Based Industrial Park, Hsin-Chu, Taiwan
*Dept. of Electronics Engineering, National Chiao-Tung University, Hsin-Chu, Taiwan
email:twang@cc.nctu.edu.tw

Abstract

The lateral distribution of programmed charge in a hot electron program/hot hole erase nitride storage flash cell is investigated by using a charge pumping technique. Our study shows that the secondly programmed bit has a wider trapped charge distribution than the first programmed bit. In addition, we find programmed charge spreads further into the channel with program/erase cycle number.

Introduction

Nitride storage flash cells have received much interest recently due to their smaller bit size, simpler fabrication process and absence of drain induced turn-on. By taking advantage of localized charge trapping in nitride above the source and the drain junctions (Fig. 1), two-bit storage of a nitride cell can be realized by hot electron program and band-to-band hot hole erase with a reverse read scheme [1]. In two-bit storage, the control of programmed charge lateral distribution is particularly important since stored electrons at the first bit will affect the threshold voltage of the second bit in reverse read and vice versa. This phenomenon is referred to as the second bit effect. Furthermore, the lateral spread of programmed charge will cause a mismatch between trapped electron distribution and injected hole distribution in erase, thus resulting in degradation of erase capability or erase speed. In this work, we will use a charge pumping technique to explore the programmed charge profile of each bit. The P/E cycling stress effect on the distribution of programmed charge is also investigated.

Charge Pumping Measurement

The samples have a gate length of 0.5μm and a gate width of 1.0μm. The voltage waveforms in charge pumping (CP) measurement are illustrated in Fig. 1. The gate pulse has a fixed high level (V_{gh}=6V) and a variable low level (V_{gl}). To profile the lateral extent of programmed charge in the drain side (or the source side), V_d (or V_s) is adjusted to modulate the drain (or source) depletion width while V_s (or V_d) is floating and the charge pumping current (I_{cp}) is measured at the substrate. The drain pulse is 180° phase-shifted with respect to V_g that the drain bias is applied only during the trapped electron emission cycle. The frequency in CP measurement is 2.5MHz.

Results and Discussion

(a) first bit I_{cp}

Fig. 2 shows I_{cp} versus V_{gl} in a virgin cell, after programming only and after one P/E cycle. Only the first bit (drain side) is P/E cycled. The threshold voltage window (ΔV_t) is 2V. Here, V_t is defined as the gate voltage when the drain current is 1μA at a reverse read voltage of 1.6V. The observed I_{cp} bump in program state is caused by negative nitride charge trapping. Fig. 3 shows that the I_{cp} bump increases with a V_t window due to more trapped electrons. In Fig. 4, the dependence of the program-state I_{cp} bump on V_d in CP measurement is shown. At a sufficiently large V_d, interface traps underneath programmed charge are "masked" by the drain depletion region. Thus, the program-state I_{cp} bump is completely suppressed. In contrast, when V_s is applied in CP measurement, the I_{cp} bump is not affected at all (Fig. 5). This indicates that programmed charge is highly localized near the drain edge.

(b) two-bit storage I_{cp}

The I_{cp} of four two-bit storage states, "11", "10", "01" and "00", is shown in Fig. 6. "00" denotes both bits in program state. Fig. 7 compares the I_{cp} of the first programmed bit and the secondly programmed bit. The 2nd programmed bit I_{cp} is measured with the first bit erased. Notably, a cross-over in Fig. 7 is observed. This cross-over suggests that the secondly programmed bit exhibits a wider charge distribution but a smaller peak density. By using a charge spatial profiling technique similar to [2], the nitride charge spatial distribution can be obtained as follows;

$$Q_N(x) = \frac{C_{ONO}}{q}(V_{gl} - V_{ti}), \quad x = \frac{I_{cp}(V_{gl})}{I_{cp,max}} L_{ch}$$

where $Q_N(x)$ is the nitride charge density, L_{ch} is the channel length and V_{ti} is the threshold voltage of a fresh device. x=0 is at the source or the drain edge. The extracted charge profile of the first programmed bit and the secondly programmed bit is shown in Fig. 8. The distribution of the secondly programmed bit (source side) is broader because a large channel field exists in the drain side during 2nd bit programming (Fig. 9). Such a large drain field results from stored electrons of the first programmed bit and will cause channel electrons to inject into the nitride earlier. It should be remarked that the above equation is derived from a 1D V_t model. For a narrow charge distribution by hot electron programming, it only serves as a first-order approximation. Accurate profiling of programmed charge distribution requires a 2D device simulation.

The programmed charge lateral extent can be also probed by varying V_d (or V_s) in CP measurement. The decrease of the program-state I_{cp} bump with V_d (or V_s) is shown in Fig. 10. The 2nd programmed bit needs a larger junction bias to "mask" the programmed charge. The same conclusion that the second bit has a broader charge distribution is reached.

(c) P/E cycling stress

The P/E cycling stress effect on programmed charge distribution is examined in Fig. 11. The V_t window keeps the same during cycling. In order to eliminate interface trap creation effect in cycling, the I_{cp} bump is normalized to its value at V_d=0V. As cycle number increases, the bottom oxide damaged region becomes broader and so does the hot electron injection region. Thus, a larger V_d in CP measurement is necessary to screen the programmed charge. In other words, the second bit effect is worsened after P/E cycling.

0-7803-8315-X/04/$20.00 ©2004 IEEE

IEEE 04CH37533 42nd Annual International Reliability Physics Symposium, Phoenix, 2004

[3] S. H. Gu, et al.,, "Characterization of Programmed Charge Lateral Distribution in a Two-bit Storage Nitride Flash Memory Cell by Using a Charge Pumping Technique," *IEEE Trans. on Elec. Dev.*, vol. 53, pp. 103-108, 2006

Characterization of Programmed Charge Lateral Distribution in a Two-Bit Storage Nitride Flash Memory Cell by Using a Charge-Pumping Technique

Shaw-Hung Gu, Tahui Wang, *Senior Member, IEEE*, Wen-Pin Lu, Wenchi Ting, Yen-Hui Joseph Ku, and Chih-Yuan Lu, *Fellow, IEEE*

Abstract—In this paper, we use a modified charge pumping technique to characterize the programmed charge lateral distribution in a hot electron program/hot hole erase, two-bit storage nitride Flash memory cell. The stored charge distribution of each bit over the source/drain junctions can be profiled separately. Our result shows that the second programmed bit has a broader stored charge distribution than the first programmed bit. The reason is that a large channel field exists under the first programmed bit during the second bit programming. Such a large field accelerates channel electrons and causes earlier electron injection into the nitride. In addition, we find that programmed charges spread further into the channel as program/erase cycle number increases.

Index Terms—Charge pumping (CP), cycling stress, programmed charge distribution, two-bit storage nitride Flash cell.

I. Introduction

NITRIDE-BASED trapping storage Flash memory has received much interest recently for its smaller bit size, simpler manufacturing process, and no drain-induced turnon [1], [2]. In a conventional SONOS cell, programmed charges are stored uniformly in a nitride layer. This SONOS concept has recently evolved into a localized trapping and two-bit storage cell, such as NROM [3] and Nbit [4] technologies. These special-type nitride Flash cells resemble a standard MOS transistor except that the gate oxide is replaced by an oxide-nitride-oxide gate dielectric stack. Two bits operation can be achieved by placing programmed charges in the nitride layer locally above the source or the drain junction by channel hot electron program and band-to-band hot hole erase. A reverse read scheme is employed to maximize the effect of the stored charge on the threshold voltage window [3]. Because of a symmetrical cell structure and a nonconductive storage element, the Nbit technology has been engineered to take advantage of a higher packing density memory array without compromising device endurance, performance and reliability.

In two-bit operation, the control of programmed charge lateral distribution of each bit is a major concern for the scalability of this memory cell. The stored electrons at the first bit will affect the threshold voltage of the second bit in reverse read and vice versa. This phenomenon is referred to as the second bit effect [5] and is closely related to programmed charge lateral spread. Furthermore, the lateral spread of stored charges in nitride will result in the degradation of erase capability or erase speed due to a spatial mismatch between stored electrons and injected holes during erase [6]. For these reasons, comprehensive understanding of programmed charge spatial distribution is

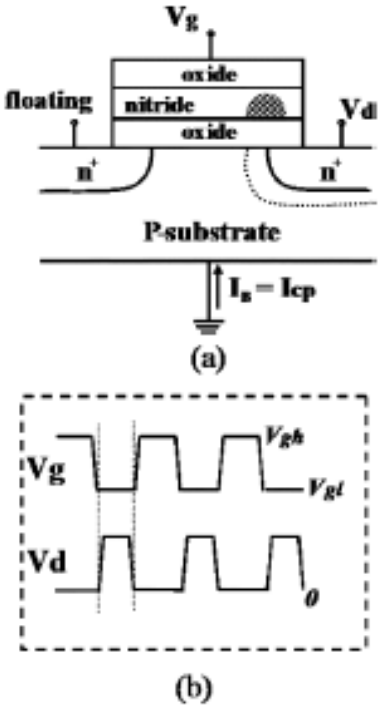


Fig. 1. (a) Schematic diagram of a two-bit storage nitride Flash cell and (b) CP measurement waveform. The dashed line in the substrate represents the depletion region caused by V_{gh}. The thickness of the ONO gate stack is 9 (top oxide), 6, and 6 nm, respectively

Manuscript received May 4, 2005; revised July 27, 2005. This work was supported by the National Science Center, Taiwan, R.O.C., under Contract NSC 92-2215-E-009-056. The review of this paper was arranged by Editor S. Kimura.

S.-H. Gu and T. Wang are with the Department of Electronics Engineering, National Chiao-Tung University, Hsinchu, Taiwan, R.O.C. (e-mail: twang@ec.nctu.edu.tw).

W.-P. Lu, W. Ting, Y.-H. J. Ku and C.-Y. Lu are with the Macronix International Company Ltd., Hsinchu 300, Taiwan, R.O.C.

Digital Object Identifier 10.1109/TED.2005.860632