國 立 交 通 大 學

電子工程學系 電子研究所碩士班

碩 士 論 文

二位元 SONOS 快閃式記憶體之物理機制

與可靠性探討

Investigation of the Mechanism and Reliability in a Two-Bit SONOS Flash Memory

研 究 生 ：郭建鴻

指導教授 ：莊紹勳 博士

中華民國 九十七 年 七 月

二位元 SONOS 快閃式記憶體之物理機制與可靠性探討

Investigation of the Mechanism and Reliability in a Two-Bit SONOS Flash Memory

研 究 生：郭建鴻　　Student : Jian-Hung Kuo

指導教授：莊紹勳 博士　　Advisor : Dr. Steve S. Chung

國立交通大學

電子工程學系 電子研究所碩士班

碩士論文

A Thesis

Submitted to Department of Electronics Engineering & Institute of Electronics

College of Electrical and Computer Engineering

National Chiao Tung University

In Partial Fulfillment of the Requirements

For the Degree of Master

In

Electronics Engineering

July 2008

Hsinchu, Taiwan, Republic of China.

中華民國 九十七 年 七 月

二位元 SONOS 快閃式記憶體之物理機制與可靠性探討

研究生：郭建鴻　　指導教授：莊紹勳　博士

國立交通大學　電子工程學系 電子研究所

摘要

就一個具有較佳資料保存特性的先進快閃式記憶體元件設計來說，氮化矽記憶體（SONOS memory）將是未來非揮發性記憶體元件的主流，它相較於快閃記憶體（flash），有著較簡單的結構、簡單的製程，而且可以比傳統浮動閘結構快閃記憶體有更佳的微縮能力（scalability）。現今的快閃式記憶體，因可攜式資訊系統的蓬勃發展，對低電壓操作、低耗電、以及快速等的要求愈趨殷切。傳統上多利用通道熱電子（CHEI: Channel Hot Electron Injection）的寫入機制來達到其中的要求，但由於離子碰撞產生的電子與電洞的交互作用可能導致穿隧氧化層的可靠性問題出現。在本論文中，將針對 SONOS 元件探討一種新的電子的注入方式，深入探討其物理機制以及可靠性議題。

對於 SONOS 記憶體元件的微縮來說，其一大優點就是每細胞二位元（2-bit-per-cell） 的操作。2-bit-per-cell 獨特的儲存結構是利用區域性的電荷注入，與不具傳導特性的電荷儲存材料。首先，我們發展出一個低電壓操作

的順偏壓電子注入方式— FBEI (Forward Bias induced Electron Injection)，相較於發表過的操作方式，該方法為目前最低的電壓，且有足夠大的電壓視窗 (operation window)。實驗結果比較，FBEI 具有與 CHEI 局部儲存電荷的相似行為。經由製作電荷密度儲存的輪廓中發現，FBEI 的儲存位置較 CHEI 更為靠近汲極端 (drain)。此外由於具有較佳的電荷保存 (data retention) 特性，使 FBEI 亦成為 2-bit 操作的新選擇。

Investigation of the Mechanism and Reliability in a Two-Bit SONOS Flash Memory

Student: Jian Hung Kuo **Advisor: Dr. Steve S. Chung**

Department of Electronics Engineering

& Institute of Electronics

National Chiao Tung University

Abstract

For the design of advanced flash memories with better data retention characteristics, SONOS (Silicon Oxide Nitride Oxide Silicon) will become the main stream of nonvolatile memory products because of its simplicity in structure and scalable by comparing with conventional floating gate cells. The flash memory today, due to the vigorous development of the portable information system, the requirements for low voltage operation, low power consumption, and high speed are becoming increasingly important. By using the conventional programming scheme of channel hot electron injection, the interaction of the generated electron and hole pairs could cause the reliability issue for the tunnel oxide. This thesis will be focused on a novel programming method for SONOS applications, in which its physical mechanism and reliability issues will be demonstrated.

For the scaling of SONOS memory, two-bit-per-cell operation has been one of the merits for SONOS devices. The unique feature of two-bit-per-cell storage is owing to the localized charge injection and the non-conducting property of charge storage material.

First, we developed a low voltage operation scheme, FBEI (Forward Bias induced Electron Injection). Comparing to those reported schemes, this FBEI scheme has features of low voltage and sufficient large operation window. We found that the FBEI and CHEI have a similar characteristic to store charge locally verified from our experiment. Moreover, the stored charge for FBEI is closer to the drain than CHEI from the profiling of the stored charge density distribution. In addition, a better data retention property also made FBEI to become a new candidate for 2-bit operation. The characteristics of endurance and data retention test have also been compared.

Acknowledgments

首先向指導教授莊紹勳教授表達無限的感激。老師總是在平常的話語、態度中教誨，以言教及身教指導學生做研究必需要有的嚴謹以及做事要抱持的積極態度。除此之外，他對學生的默默關懷，我必須在此表達感謝之意。

碩士兩年雖過得快，但是非常的充實。在此我要感謝已畢的學長們，包括元亨、耀賢、亞峻、大正，給予我在實驗上的建議及引航，特別是元亨學長，不僅傳授與教導我實驗上的各種經驗，生活上更是一大良友。感謝在兩年碩士生涯中，與我一起成長的同學，包括家銘、易叡、文彥與友良，我們一起經歷困苦的研究過程，相信數年後再聚首，一定倍感溫馨。再則，感謝汪老師研究群的學長們，包含小馬，阿斗肯，達達，志昌以及阿雄學長，即使我們身在不同的研究群，但我們的感情就像是一家人一樣，希望以後在業界相遇時還能保有我們現在的真誠。此外也感謝振鵬、米華、健宏、安舜與專題生禎晏，有你們使得我能夠在枯燥的研究外，在生活上找到歡樂與愉悅，輕遞一份真摯的祝福，希望明年也能看見你們順順利利的畢業。亦要感謝的是所有關心我的朋友們，我希望能和你們一起分享這個榮耀。

最後我要感謝我的家人給我的鼓勵與扶持，以及最要感謝的是我女友欣霖，在兩年中，每當我遭遇挫折、心灰失意時，總能諒解並體貼地給我關懷、支持與鼓勵，讓我碩士生涯充滿溫馨與甜蜜。沒有妳在背後當那個默默英雄，我無法有今日的成果。

謹將這份榮耀獻給培育我多年的父母親。

Contents

Figure Captions

BBHH erasing.

Fig. 5.16 Retention behaviors after 10k P/E cycles of Bit-1 and Bit-2 in four different states, respectively. The programming mechanism for **(a)** CHEI and **(b)** FBEI.

Table Captions

Chapter 1

Introduction

1.1 The Motivation of This Work

Next generation high density Electrically Erasable and Programmable Read Only Memories (EEPROMs) require an endurance in excess of 10^5 program/erase cycles with 10 years data retention and low programming voltages. Two basic types of EEPROMs exist, namely, the floating gate device and the SONOS (Silicon Oxide Nitride Oxide Silicon), as revealed in White's paper in 2000 [1]. The floating gate device stores charge in the poly-silicon as free carriers with a continuous spatial distribution in the conduction band and the SONOS stores charge in spatially isolated deep level traps

Although the relatively thick tunnel-oxide in the floating-gate type memories provides good 10-year data retention, the floating gate memory has limitations with respect to scaling the cell size and program/erase voltages. However, the demand for low power, low voltage electronics has accelerated the pace for nonvolatile semiconductor memory circuit designers to consider SONOS for low voltage, high density EEPROMs. The motivation for the interest in SONOS lies in low programming voltages, endurance to extended erase/write cycling, immunity to erratic charge loss, capability of 2-bit/transistor [2]-[8] storage, and compatibility with high density scaled CMOS technology.

In the past, many programming and erasing schemes have been studied for the

SONOS, such as CHEI (Channel Hot Electron Injection), FN tunneling (Fowler-Nordheim tunneling) for programming and BBHH (Band-to-Band tunneling induced Hot Hole injection) [8]-[9], FN tunneling for erasing in N-channel flash memory cells. Either of them needs a higher voltage or lower speed of operation.

In this thesis, FBEI (Forward Bias induced Electron Injection) will be employed for N-channel SONOS flash memory cells. The performance and the reliability issues based on this scheme will also be discussed. A new charge pumping method [10]-[15] will be used to extract the charge profile in the nitride [16]-[19]. Meanwhile, the charge profile during programming will also be detected. By using a three level charge pumping technique, the electron programming mechanism will be discussed. Finally, the reliability test for 2bit/cell operation will be compared between CHEI programming and FBEI.

1.2 Organization of the Thesis

The organization of this thesis consists of six chapters. After a brief introduction in Chapter 1, we will introduce the experimental devices and experimental setup in Chapter 2, including the operating scheme used to program the cells and the measuring method for charge pumping. In chapter 3, we will introduce the mechanisms for FBEI first and we will present the performance and the reliability by using FBEI and BBHH later. In Chapter 4, we will explain the phenomenon of charge pumping measurement, and we can make a V_T profile versus the gate length based on these concepts. Moreover, we will use this tool to calculte the trapping charge characteristic during different programming time. In Chapter 5, a three-level charge pumping is used to do the test, and FBEI programming mechanism is proposed; for two-bit applications, the SONOS endurance and retention characteristics will also be

discussed. In the last Chapter, the summary and conclusion will be given.

Chapter 2
Device Preparations and Equipment Setup

2.1 Introduction

This chapter is divided into four sections. First of all, the SONOS cells and the split conditions used in this study will be described. Second, the instrument setup and the experimental techniques to accurately control these instruments are illustrated. Third, we will discuss the program schemes of these cells. Finally, three charge pumping measurement technique setup used in this study will be demonstrated.

2.2 Device Fabrication

Table 2.1 (a) shows the dielectric thickness of the SONOS memory used in this study, and with several different channel widths and lengths, as shown in Table 2.1 (b). Moreover, we often used #15 and #25, in which the tunnel oxide is first grown by thermal oxidation with different thicknesses of 25 Å and 50 Å. Next, a layer of 60 Å LPCVD nitride film is grown. Finally, the LPCVD block oxide is grown with thickness of 70 Å.

2.3 Equipment Setup

The experimental setup for the I-V and transient characteristics measurement of SONOS is illustrated in Fig. 2.2. Based on the PC controlled instrument environment via HP-IB (GP-IB, IEEE-488 Standard) interface, the complicated and long-term characterization procedures during analyzing the intrinsic and degradation behaviors

(a)

SONOS 2nd Split

	#15	#17	#20	#23	#25
ONO = 25 / 40 / 50					
ONO = 25 / 60 / 50					
ONO = 25 / 40 / 70					
ONO = 25 / 60 / 70	V				
ONO = 50 / 40 / 50		V			
ONO = 50 / 60 / 50			V		
ONO = 50 / 40 / 70				V	
ONO = 50 / 60 / 70					V

(b)

W/L
0.7/0.5
0.7/0.3
0.7/0.22
0.2/0.5
0.2/0.22
0.2/0.2
0.2/0.16

Table 2.1 (a) The split conditions of samples used in this work, in which devices have several different ONO dielectric thickness

(b) Condition of samples with different channel widths and lengths

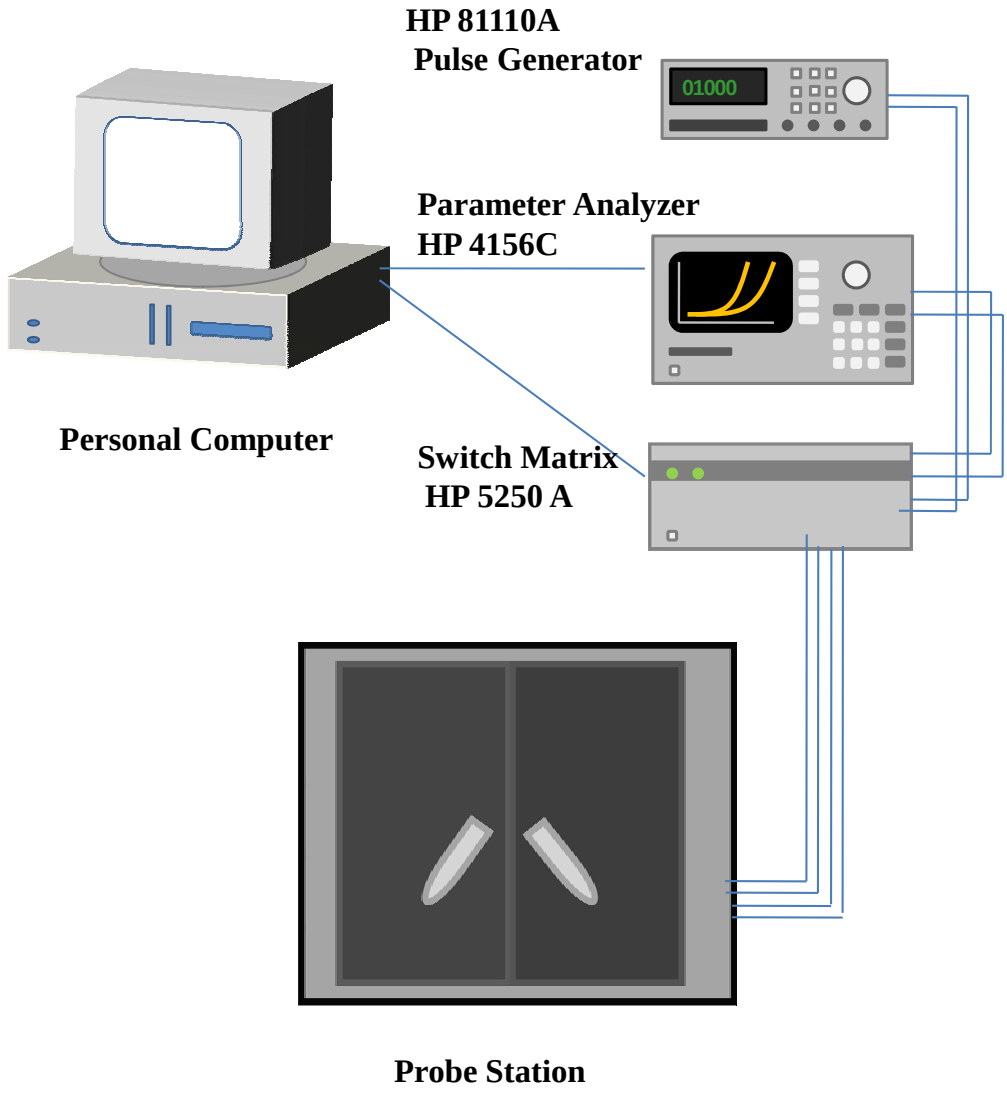


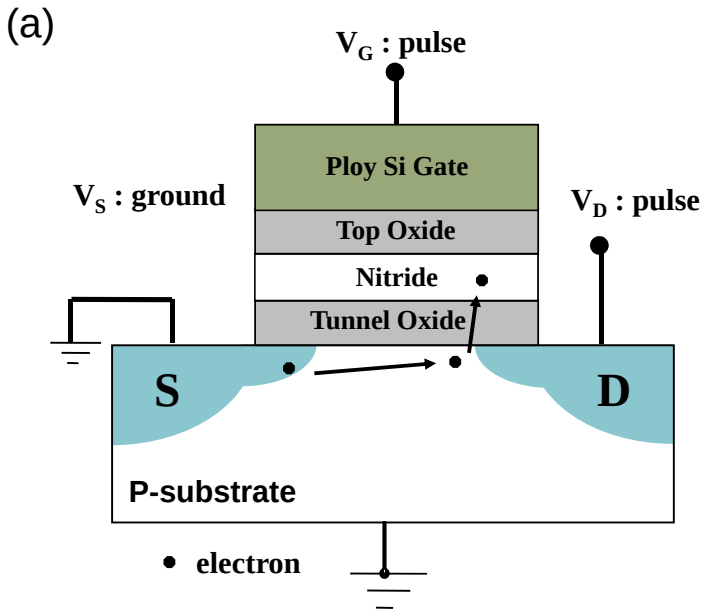
Fig. 2.2 The experimental setup of the current-voltage and the transient characteristics measurement in FLASH cells. An automatic controlled characterization system is setup based on the PC controlled instrument environment.

in SONOS cells can be easily achieved. As shown in Fig. 2.2, the characterization apparatus with semiconductor parameter analyzer (HP 4156C), dual channels pulse generator (HP 8110A), low leakage switch mainframe (HP E5250A), and a probe station provides an adequate capability for measuring the device I-V characteristics and executing the SONOS cell program/erase operation.

Source-monitor units (SMU) and provided the high current resolution to 10^{-15}A range facilitates the gate current measurement, sub-threshold characteristics extraction, and the saturation drain current measurement. The HP E5250A equipped with a 10-input (6 SMU ports and 4 AUX ports) × 12-output switching matrix, switches the signals from the HP 4156C and the HP 8110A to device under test (DUT) in probe station, automatically. In addition, the HT-Basic are used as the program languages to achieve the personal computer (PC) control of these measurement instruments.

2.4 Programming Schemes

The first scheme is a traditional program scheme, Channel Hot Electron Injection (CHEI), used for electron injection in SONOS cell, where the gate and drain are connected to pulse generator while source and substrate are grounded as shown in Fig. 2.3 (a). The pulse timing diagram for both gate and drain are shown in Fig.2.3 (b).The second scheme is the main idea which we propose a new program scheme to inject electron for SONOS cell applications, with the gate and drain are connected to pulse generator while substrate is grounded and source is opened in Fig 2.4 (a). The pulse timing diagram for gate and drain are shown in Fig 2.4 (b). The operation mechanism and relevant measurement will be discussed after this chapter.



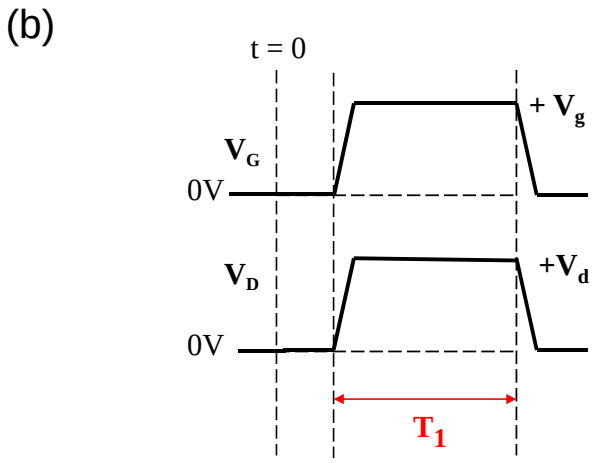


Fig. 2.3 (a) The operation scheme for CHEI
(b) The timing diagram of CHEI

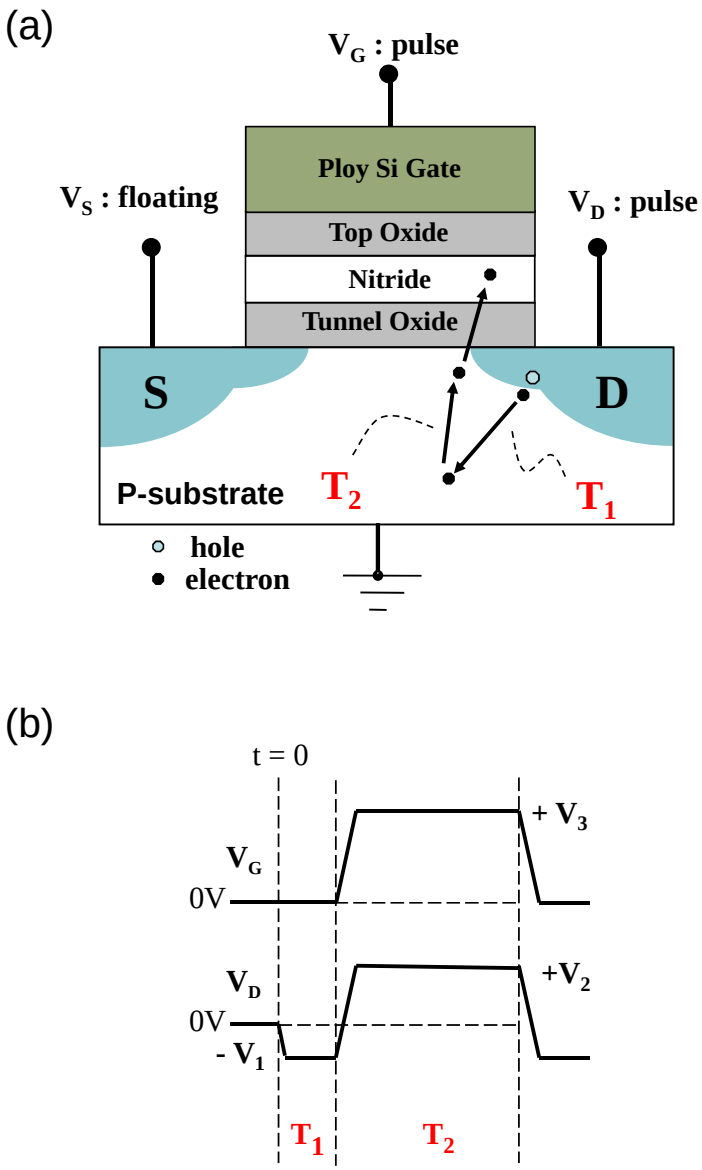
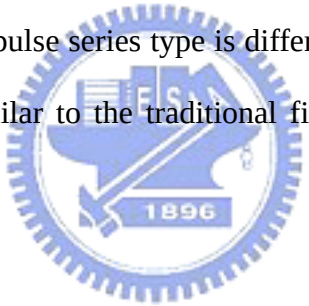


Fig. 2.4 (a) The operation scheme for FBEI

(b) The timing diagram of FBEI

2.5 Charge Pumping Measurement Technique Setup

The first scheme for charge pumping measurement technique setup is shown in Fig. 2.5, which we call fixed base charge pumping, but with some differences from the traditional one. The pulse generator is connected to the gate, and with the substrate and drain connecting to the HP4156C, while source is kept floating. Fig. 2.5 (b) shows the pulse series type sending out from the pulse generator by this setup. By using this setup, we can measure the charge pumping current from drain. If one wants to measure the charge pumping from drain (/source), we should connect the HP4156C with them and open the source (/drain) to get the drain side charge pumping current. The second scheme for charge pumping measurement technique setup is shown in Fig. 2.5 once again. However, the pulse series type is different this time. As shown in Fig. 2.5 (c), we used the way similar to the traditional fixed top charge pumping pulse series to do our measurement.



The third scheme is shown in Fig. 2.6, where the substrate, source and drain are identically connected to the ground. Again, the gate terminal is connected to the pulse generator in order to supply the new kind three-level pulse we needed, as shown in Fig. 2.6 (b). Additionally, to precisely control the pulse type of HP 8110A during charge pumping measurement, a special measurement technique, added mode measurement, is used. Fig. 2.7 shows the pulse series of the triggered pattern mode outputted by Output 1 and Output 2. By taking these time diagram as an example, with the use of added mode of HP8110A to combine the pulse of output-1 and output-2, we can get the three-level pulse.

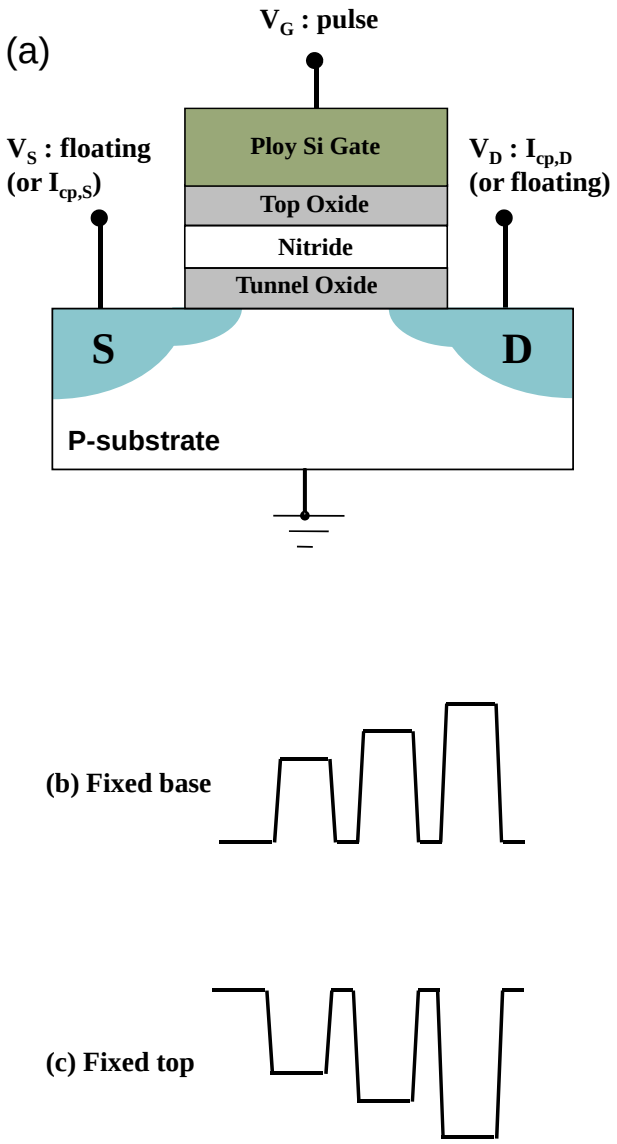
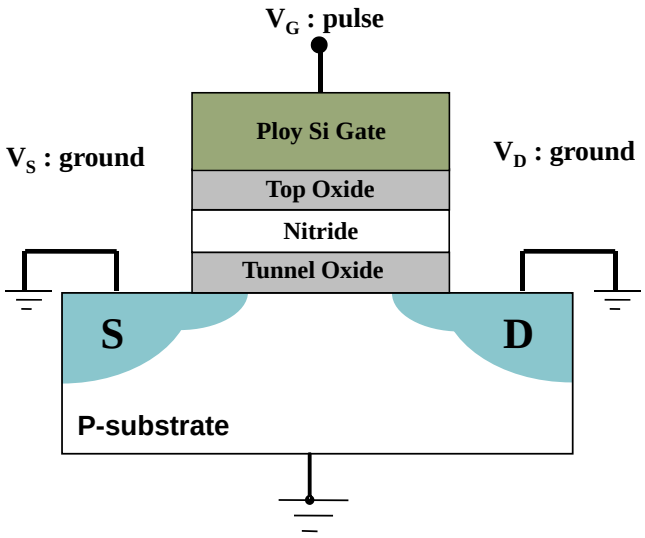


Fig. 2.5 (a) The operation scheme for charge pumping technique

(b) Fixed base pulse series sketch approach

(c) Fixed top pulse series sketch approach

(a)



(b)

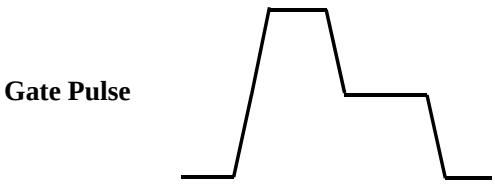


Fig. 2.6 (a) The operation scheme for three level charge pumping technique
(b) Three level pulse series sketch approach

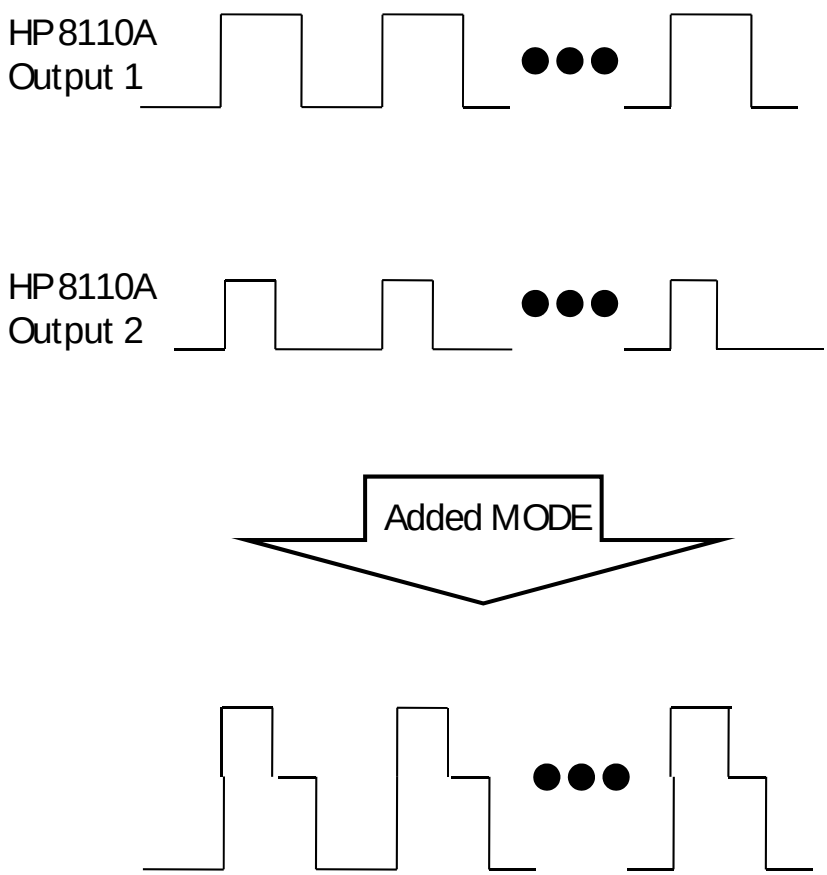


Fig. 2.7 Using the added mode of 8110A to combine the pulse series, and get three level pulse.

Chapter 3
Basic Results on the Cell Programming

3.1 Introduction

In this chapter, first, we will introduce the mechanism about FBEI (Forward Bias induced Electron Injection) for N-channel device. Then, we will show the optimized bias conditions of operation, including the gate and drain voltages for FBEI and how we enhance the injection capability and lowering the operation voltage.

3.2 The Operating Mechanisms of FBEI

FBEI was modified from PASHEI (Pulse Agitated Substrate Hot Electron Injection) [20]-[21]. The basic concept was coming from the SHEI (Substrate Hot Electron Injection) [22]-[23] which was discovered at least two decades ago. Initially, PASHEI has been used to program flash memory devices as well as to erase the DINOR [21],[23],[24] floating-gate flash memory devices. Normally, to implement the SHEI technique, an additional injector for emitting electrons is needed, which increases the complexity or the process and the circuit design. PASHEI was firstly used in the floating-gate devices, and it shows a great endurance characteristic and lower voltage operation. In our current work, we modify the timing diagram of V_G at time T_1 (see Fig. 2.4 (a) and (b) in Chapter 2), and this is the first time used in the SONOS devices. In the present scheme, we can inject the electrons into the gate dielectric nitride without the time T_1 of V_G and it can be found that the scheme exhibits some similar phenomena as CHEI (will be discussed in the next Chapter).

Fig. 2.4 (a) and Fig. 2.4 (b) in Chapter 2 show the schematic and timing diagram of FBEI, during the electron emitting phase T_1, in which the Sub/D junction is forward biased and electrons are injected into the substrate. Subsequently, the Sub/D is reverse biased to create a deep depletion region at collecting phase T_2, which will cause the previously injected electrons in the substrate (those that have not been recombined) to be accelerated across the depletion region and injected into the gate dielectric. The electron injection mechanism during the collecting phase is the same as the conventional SHEI, except that we use the drain junction as both the forward biased emitter and the reverse biased injector in FBEI.

3.3 Measurement Results and Basic Characteristics of FBEI

Figure 3.1 shows the characteristics of FBEI as a function of emitting bias on the drain for a given V_G= 6V. We can observe that a forward bias of more than 1V is needed during the emitting phase (T_1) before substantial injection takes place. Fig. 3.2 shows that a reverse bias of over 3.5V is required during the collecting phase (T_2) for efficient injection, due to the fact that hot electrons must surmount the barrier height of the Si/SiO_2 interface (~3.1eV). The threshold voltage decreases while the collecting voltage reaches 5V, since the lateral electric field is lager than the vertical electric field. By using these two measurement result, we can choose an optimum operation voltage.

The relationship between pulse rising/falling time and V_t is shown in Fig. 3.3. The strong pulse rising/falling time dependence can be seen herein, where the V_t drops about 2V for one order change of the rising/falling time. In Fig. 3.4, V_t after programming is plotted as a function of the emitting time, while the total pulse period

is kept at 10μs. For emitting times longer than 2μs, the programming rate seems to be independent of the emitting time. The pulse count dependence is then investigated with the results given in Fig. 3.5, where the total operating time is set to 1ms and changing the pulse count, the more pulse count the higher V_t is then achieved. One can suggest that the injecting rate depends primarily on the pulse count, regardless of the pulse period, at least within the range that we studied. These results imply that, within some limits which remain to be determined, the injection rate can be increased by raising the frequency of the pulse.

In Fig. 3.6, the transient behavior of FBEI can be investigated. As expected, the higher gate voltage shows a faster injecting speed comparing to collecting drain voltage. These results suggest that, if fast injecting speed is important, one may use a relatively high gate voltage to accomplish that. On the other hand, if low-voltage operation is important, one may use a lower gate voltage, as long as the moderate injecting speed is acceptable.

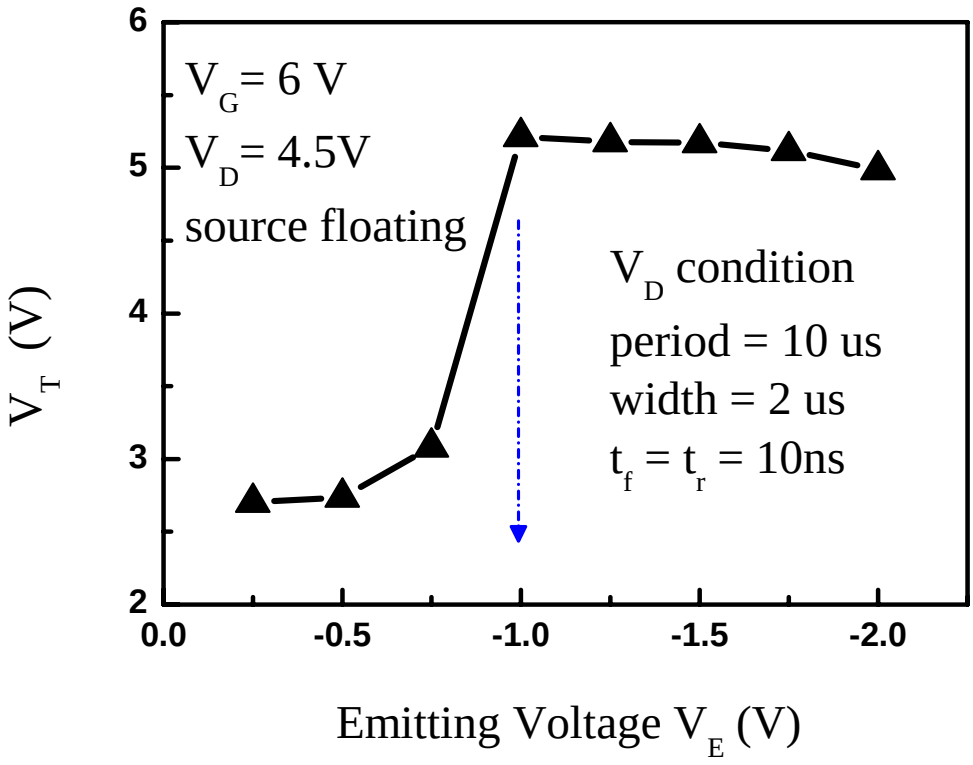


Fig. 3.1 The characteristics of FBEI as a function of emitting bias on the drain for a given V_G = 6V.

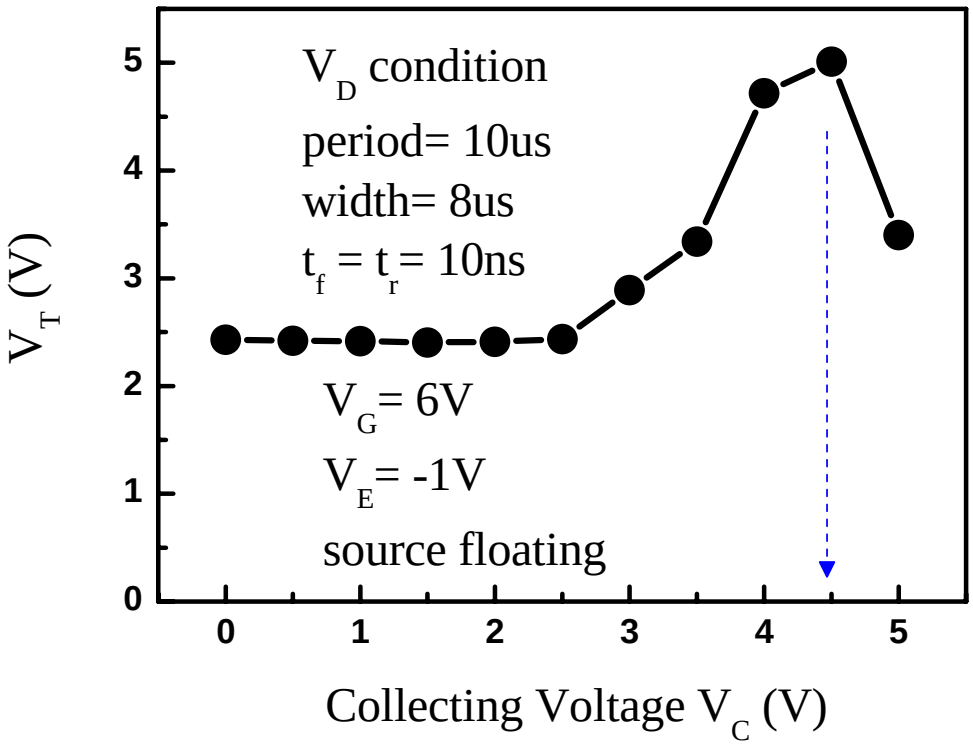


Fig. 3.2 The characteristics of FBEI as a function of collecting bias on the drain for a given V_G = 6V and V_e = -1V.

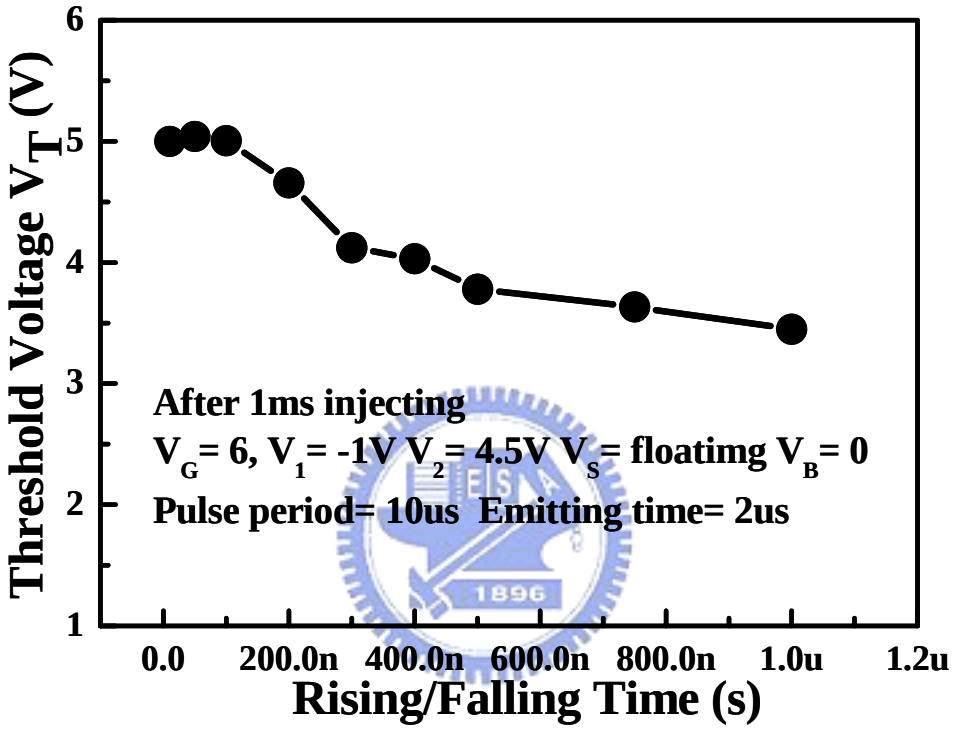


Fig. 3.3 Pulse rising / falling time dependence of programmed threshold voltage for FBEI technique

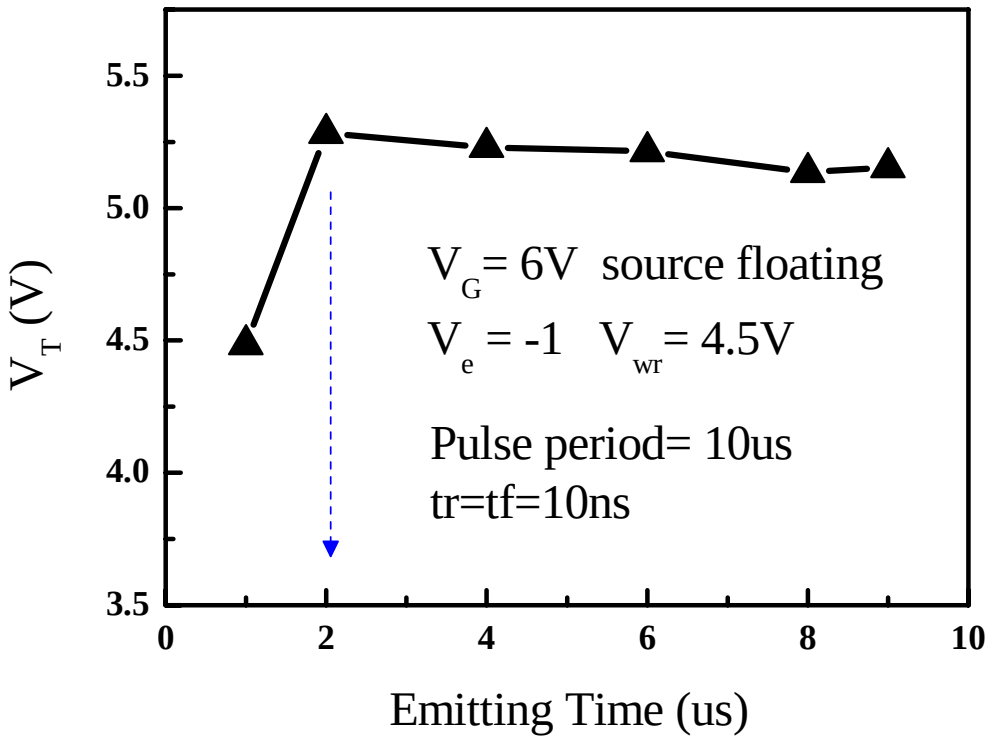


Fig. 3.4 Emitting time dependence of the threshold voltage for FBEI.

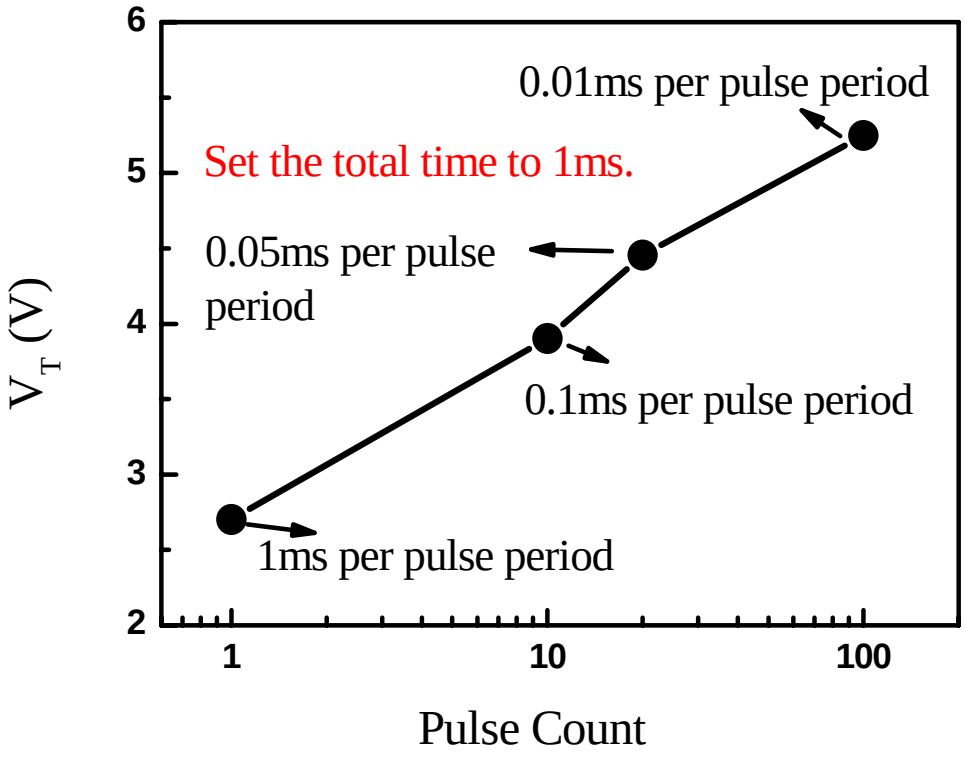


Fig. 3.5 Pulse count dependence of the threshold voltage for FBEI.

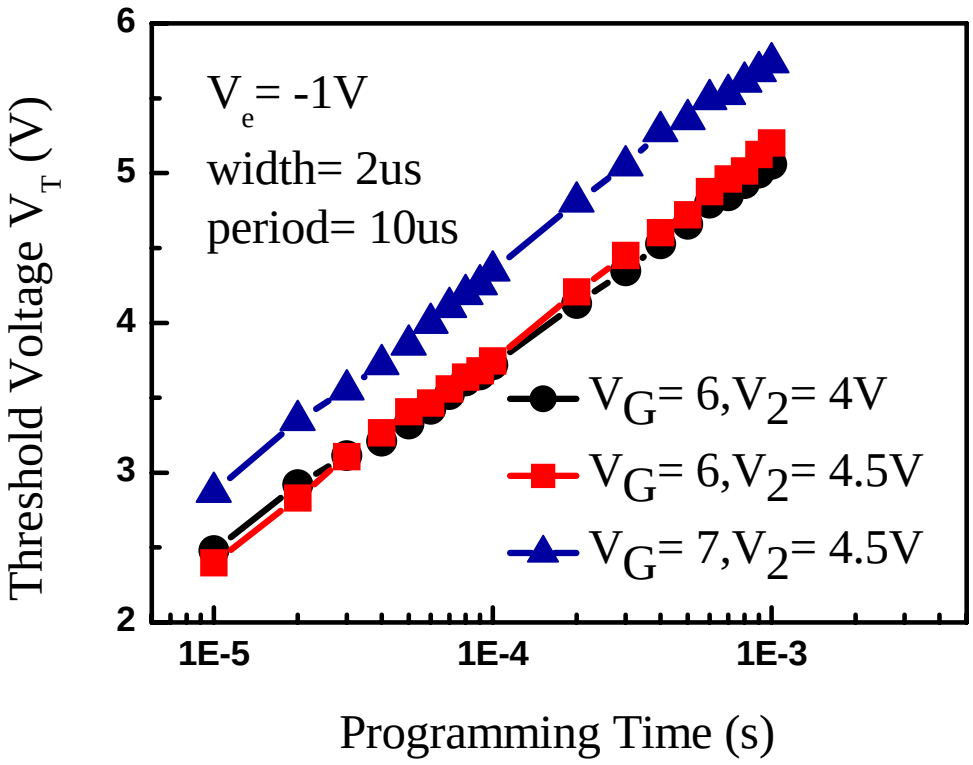


Fig. 3.6 The transient characteristics for three different operation conditions.

Chapter 4
Characteristics of Localized Trapping Charge

4.1 Introduction

SONOS memory has received much more attention as one of the alternatives to conventional poly-Si floating-gate nonvolatile devices. Difference from the conventional floating-gate flash memory, the storage sites of the SONOS memory exhibit dielectric property, which means the trapped charge is isolated from the neighboring sites. This SONOS concept has been evolved into a two bit-per-cell, trapping storage, nonvolatile memory (NROM) technology. In charge trapping memory, channel hot electron injection (CHEI) or channel initiated secondary electron (CHISEL) injection is usually used for programming, by which means electrons are injected and stored in a localized region of the silicon nitride layer along the channel. Meanwhile, we found a similar localized charge trapping phenomenon between CHEI and FBEI schemes.

To profile the actual stored charge distribution, several experimental methods such as the current-voltage (I-V) measurement, the GIDL method, and the charge-pumping method have been studied. In this chapter, a new different way of charge pumping method is developed to detect the lateral distribution of nitride trapped charge density and the V_t profile.

4.2 Charge profile in view of FBEI and CHEI Schemes

4.2.1 Principle of Charge Profile by Charge Pumping Method

The charge pumping method has been widely used for hot-carrier-related reliability characterization in MOSFETs. During a typical charge pumping measurement, a pulse string is applied to the gate terminal of a MOSFET while the substrate current (commonly called the “charge pumping current”) is monitored. Since this current is a result of the recombination of majority carriers (coming from the substrate when the gate is biased between flat-band and accumulation) with the trapped minority carriers at the interface (coming from the source/drain when the gate is biased to inversion), to first order the charge pumping current (I_{cp}) is nonzero only if the high level (V_h) and the base level (V_b) of the gate pulses cover both the threshold voltage (V_t) and the flat-band voltage (V_{fb}) [14].

Unlike the conventional charge pumping (CP) method, the other two basic ways of charge pumping test to attain the profile scheme are demonstrated. First one is the fixed base CP (fixed base level and varying the top level) method with one side of drain (or source) floating and the other one is the fixed top CP (fixed top level and varying the base level) method with also one side of drain (or source) floating, which are defined as FV_b and FV_t, respectively.

In FV_b CP method, the setup is shown in Fig. 2.5 (a) in chapter 2, the gate is applied with a pulse string, as shown in Fig. 2.5 (b), and the I_{cp} can be measured from drain or source side with source or drain floating respectively. When measuring the charge pumping current $I_{cp,d}$ from the drain side, the minority carrier only contributed from the drain side and vice versa with $I_{cp,s}$. Therefore, we can obtain more precious information about the drain and source side from $I_{cp,d}$ and $I_{cp,s}$. By combining these two currents, we can profile the asymmetrical V_t along the channel for both virgin and programmed cells.

Figure 4.1 (a) illustrates the V_t profile of a programmed nitride storage memory cell, which contains a narrow V_t peak near the drain side. Four regions are marked in this figure, and they are consistent with the I_{cp} curve tested from FV_b CP method in Fig. 4.1 (b). Fig. 4.1 (b) corresponds to the drain or source junction area in Fig. 4.1 (a). After programming, localized trapped charges enhance the threshold voltage near drain side, which forms the asymmetrical V_t profile in Fig. 4.1 (a). Therefore, the $I_{cp,d}$ and $I_{cp,s}$ curves shift toward the right, which corresponds to the regions B and C in Fig. 4.1 (b). The difference between curves B and C indicates the location and profile of the injected charges. As Fig. 4.1 (b) shows, the injection is closer to the drain side. It needs to be pointed out that I_{cp} keeps shifting rightward in region D, indicating a V_t peak here. Moreover, $I_{cp,d}$ and $I_{cp,s}$ overlap in this region, which means the minority carrier coming from drain or source is passing through the peak region under the channel. For this reason, the equivalent interface traps are sensed and contribute the same $I_{cp,d}$ and $I_{cp,s}$. Thus, in FV_b CP method, data obtained in region D cannot be used to extract the exact profile of V_t in large current region. We can, however, extract the accurate location of the peak using this method.

On the other hand, the equipment setup is similar to FV_b CP method expect that the top level of gate pulse string is a constant upon the threshold voltage. In contrast, the I_{cp} curve shift caused by the V_t peak, takes place in the low current region and has higher precision in FV_t CP method. Fig. 4.2 (a) illustrates I_{cp} test with FV_t CP method in logarithmic scale. Correspondently, region B, C and D in Fig. 4.2 (a) and Fig. 4.1 (b) also can be seen herein. I_{cp} in region D can be used to extract the accurate profile of narrow V_t peak due to its low testing current. However, V_h is set larger than the highest V_t along the whole channel, and I_{cp} current tested from drain and source are

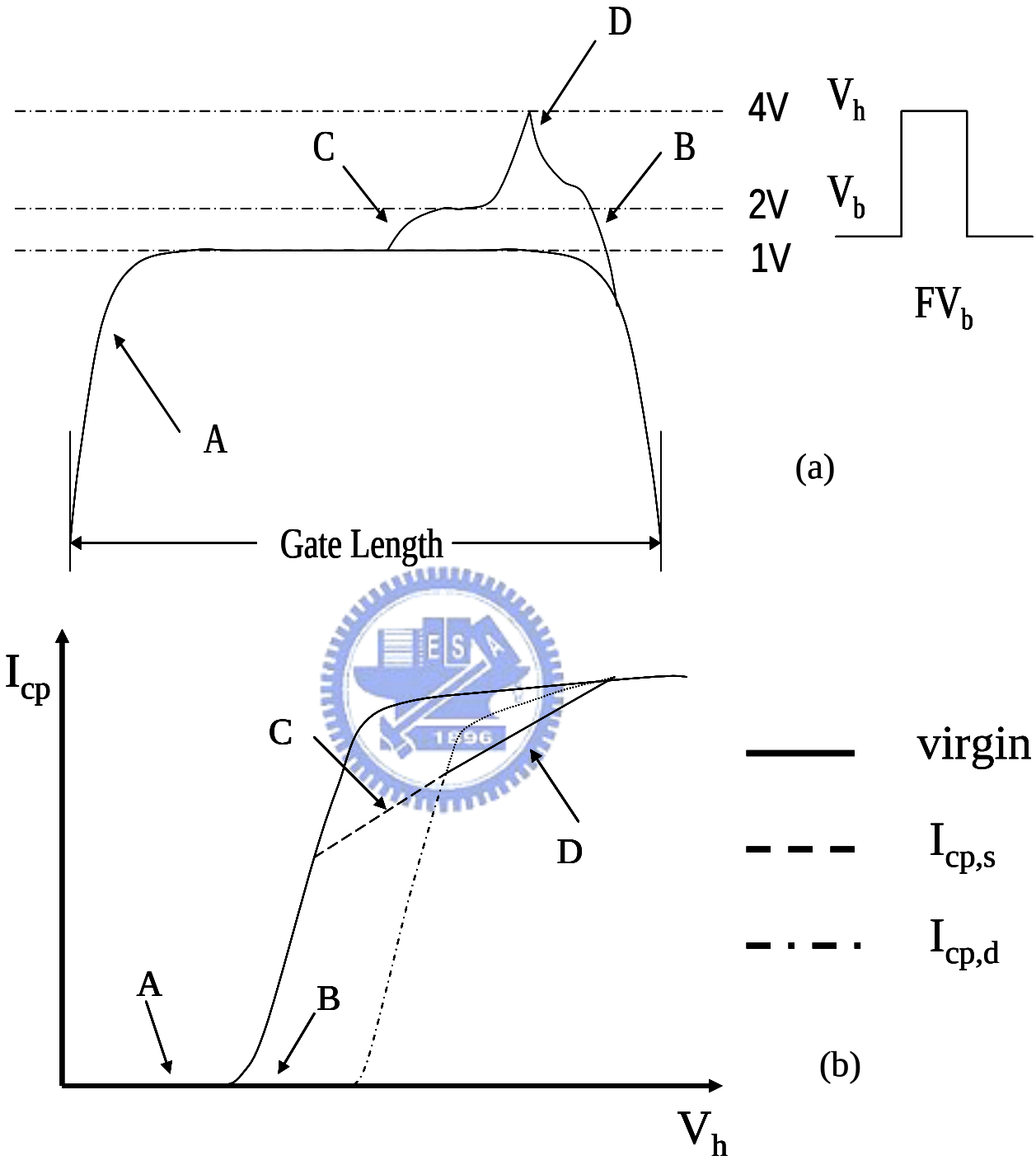


Fig. 4.1 (a) Diagram of V_T profile in a programmed nitride storage memory cell
(b) Illustration of I_{cp} curves versus V_h before and after programming

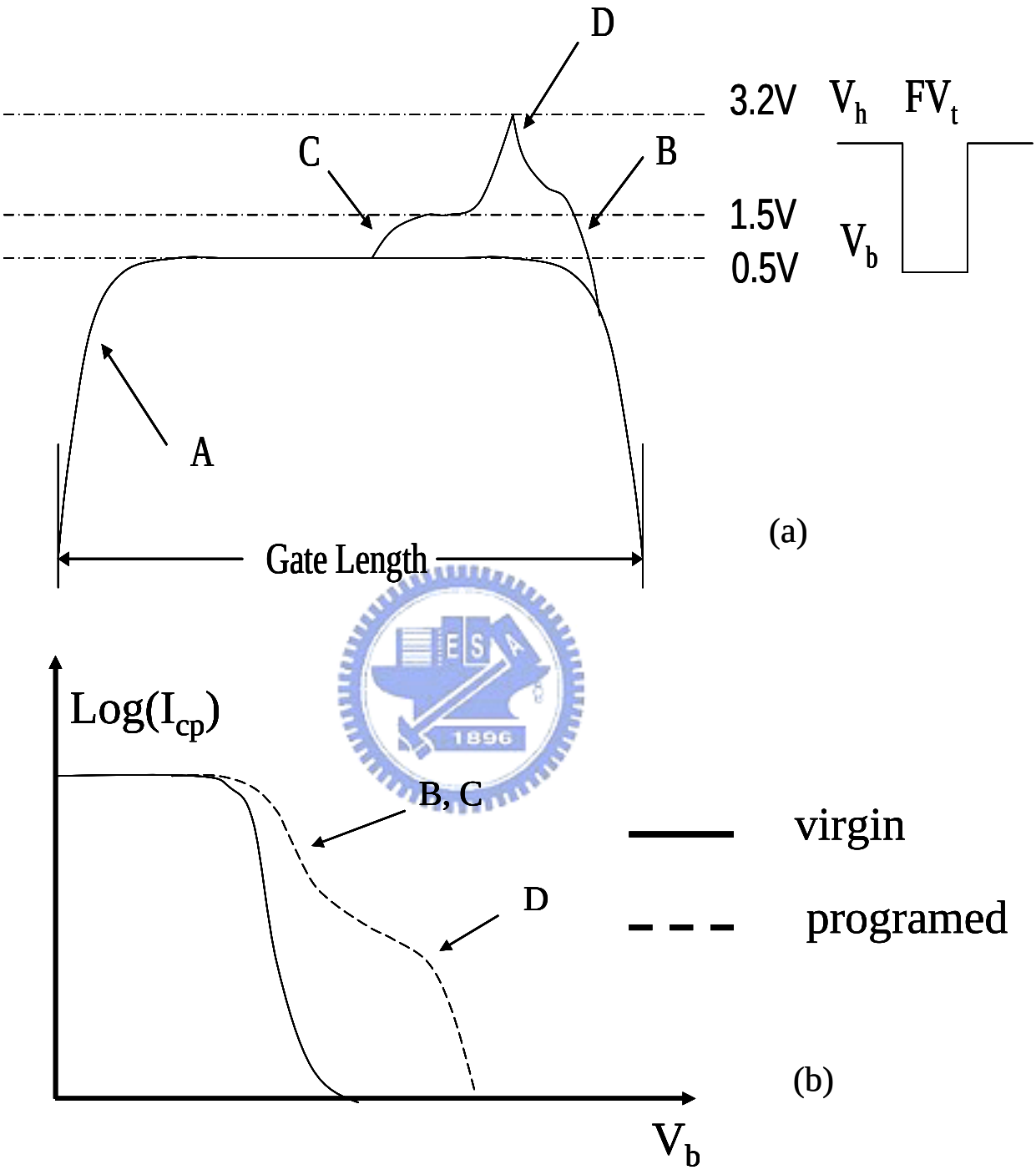


Fig. 4.2 (a) Diagram of V_T profile in a programmed nitride storage memory cell
(b) Illustration of I_{cp} curves versus V_b before and after programming

identical. FV_t CP method can only extract the width and value of narrow V_t peak but cannot be used to identify the location.

4.2.2 Derivation of Local V_t Distribution

Charge pumping technique was originally developed to detect the interface state density of MOSFET devices. Today, it is possible to profile the lateral distribution of interface state density $N_{it}(x)$ and oxide trapped charge density $Q_{ot}(x)$ in the MOSFET, using various charge pumping techniques [10]-[15][16]-[19]. However, we observed that during much shorter hot carrier programming (microseconds of milliseconds order) relative to hot carrier stressing (a few hundred seconds), additional interface states are hardly generated and the injected charges are only trapped within nitride layer in the SONOS memory. Meanwhile, threshold voltage distribution $V_t(x)$ of a device without electrical stress can be obtained easily because N_{it} for the fresh devices is known to be uniform. Therefore, we can easily find out the $V_t(x)$ of the SONOS memory hot carrier programmed shortly enough to suppress additional interface states. Since the local $V_t(x)$ reflects the trapped charge at the point of interest, we can know the nitride trapped charge density distribution $N_{ntr}(x)$.

To extract the profile, we should base on some assumption. First, we assume that the interface trap density (including all the interface traps in the energy range of Si band-gap contributing all to I_{cp}), N_{it}, is spatially uniform along the channel. Second, the charge in the nitride is also spatially uniform for the virgin cell, thus the different value of local $V_t(x)$ is just due to the doping concentration in the substrate near the n+-p S/D diffusion region. Therefore,

$$I_{cp}(V_h) = qfN_{it}Wx \bullet h(V_h - V_t(x)) \qquad (1)$$

where q is the elementary charge, f the pulse frequency, W the effective channel width, L the gate length, and h can be approximated as a step function which can be expressed as

$$h(y) = \begin{cases} 1, & if \ y > 0 \\ 0, & if \ y < 0. \end{cases}$$

This means that the $N_{it}(x)$ will give a charge pumping current at a certain V_h equals to the local $V_t(x)$. Thus

$$V_h = V_t(x) \tag{2}$$

$$I_{cp}(V_h) = qfN_{it}Wx. \tag{3}$$

If the V_h is higher than the peak value of $V_t(x)$ in the whole channel, all the interface along the channel will "pump", so we can receive a maximum constant I_{cp},

$$I_{cp,\max} = qfN_{it}WL. \tag{4}$$

Comparing (1) and (4), we can derive

$$x = \frac{LI_{cp}(V_h)}{I_{cp,\max}}. \tag{5}$$

In other words, each incremental ΔV_h covers an additional incremental portion of the channel, which gives an incremental charge pumping current ΔI_{cp} corresponding to the interface traps in that portion of the channel, and the voltage V_h at which this increment occurs is the local V_t of that portion of the channel. Thus the derivative of (1) with respect to V_h

$$\frac{dI_{cp}(V_h)}{dV_h} = qfwN_{it}x \bullet \delta(V_h - V_t(x)) \tag{6}$$

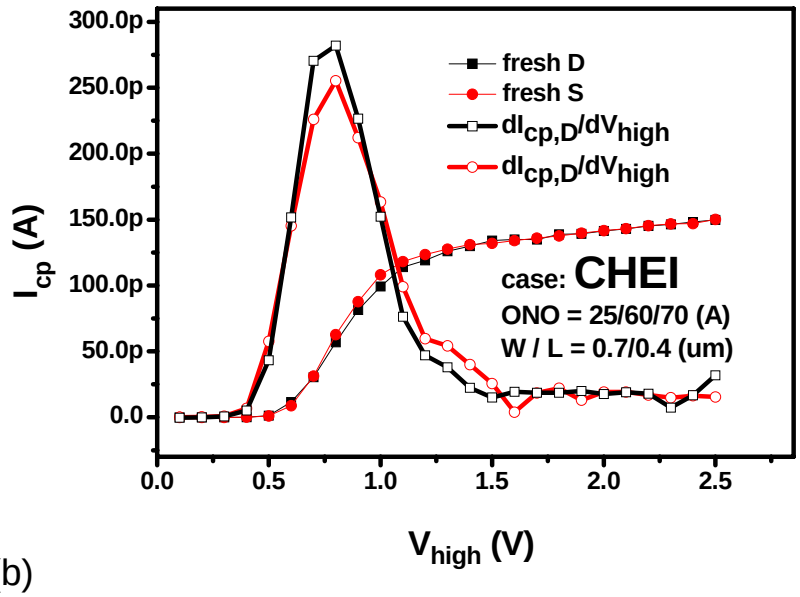
where δ, the Dirac delta function, reflects the distribution of interface traps over the threshold voltage. While in theory a uniform channel should give a sharp peak of the $dI_{cp}(V_h)/dV_h$ versus V_h plot, in general this curve exhibits a finite width due to the laterally non-uniform distribution of $V_t(x)$ (i.e. caused by the doping concentration). By using Eq. (6) we can evaluate the $I_{cp,max}$ at the maximum value of the derivative for the virgin cell.

For the programmed cells, Eq. (6) is not proper to evaluate the $I_{cp,max}$ because of the nitride charge does not still uniform distribute, and the V_t non-uniform distribution is caused by both the stored charge and substrate doping concentration. Instead, we can investigate the $I_{cp,max}$ from the fixed top charge pumping measurement, because of its maximum value is more stable than the fixed base charge pumping measurement for us to select in our devices.

4.2.3 V_t Profile and Trapping Charge Analysis

Fig. 4.3 (a) and Fig. 4.3 (b) show the FV_b CP method measured from CHEI and FBEI respectively. The charge pumping current measured from $I_{cp,d}$ and $I_{cp,s}$ are identical for the virgin cell. The right shift of $I_{cp,s}$ in low current region after programming indicating a slight injection near source side. Both $I_{cp,d}$ and $I_{cp,s}$ are overlapping at 2V indicating the location when the minority carriers, coming from the drain or source, passing through the V_t peak. Most important of all, Fig. 4.3 (a) and (b) show $dI_{cp}(V_h)/dV_h$ versus V_h curves, and we can obtain a $dI_{cp}(V_h)/dV_h$ peak for virgin cell. It means that the maximum local V_t distribution is 0.6V and 1.1V for the case of CHEI and FBEI respectively for the virgin cell. We would like to use this technique, so we can easily define the $I_{cp,max}$ for the virgin cell and the V_t distribution will be given by using Eq. (3) and Eq. (5).

(a)



(b)

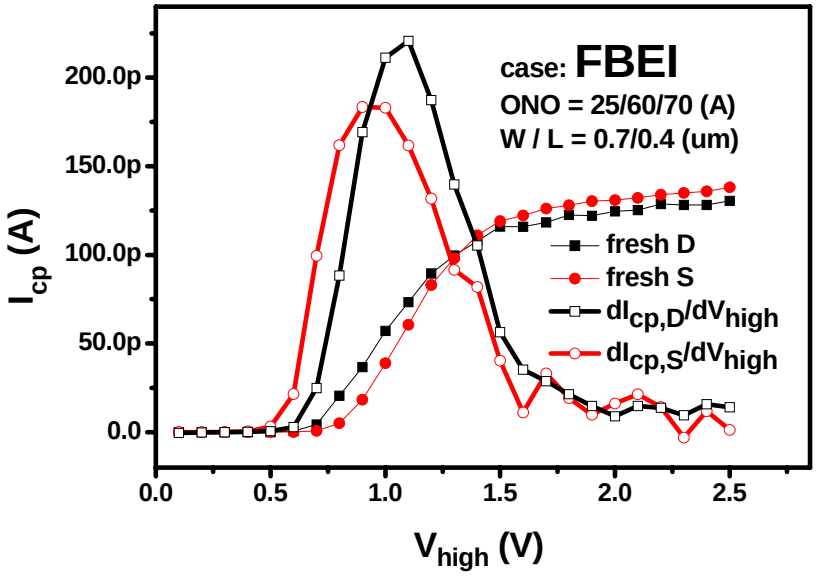


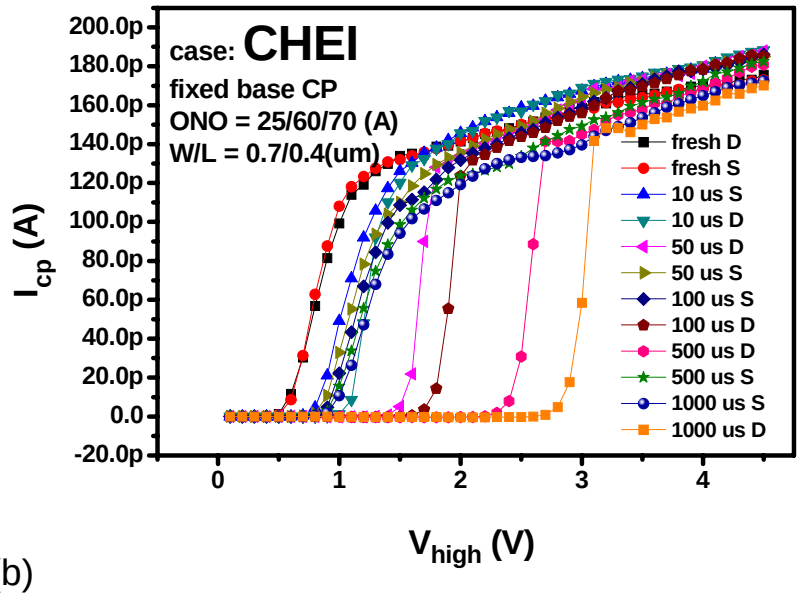
Fig. 4.3 From the maximum value of dI_{cp} / dV_h we can define the Icp,max for the virgin cells case **(a)** for CHEI writing and **(b)** for FBEI writing.

In Figs. 4.4 (a) and (b), we show the fixed base charge pumping method for CHEI and FBEI respectively, with different programming time to do the test. We can find out that the drain side I_{cp} using CHEI programming shifts larger than that of FBEI. Therefore, for CHEI the charge stored in the nitride close to the drain should be more saturated than that of FBEI. And in Figs. 4.5 (a) and (b), we show the fixed top charge pumping method for CHEI and FBEI respectively. This time we can observe that FBEI has more data under 10p than CHEI, which means that the "bump" of FBEI is easier to sense than that of CHEI.

We used $I_{cp,max}$ for the virgin cell in Fig. 4.3, but in the programmed cell, we can use the data of fixed top charge pumping measurement as a reference since the fixed base charge pumping current increases gradually in the high current region, and this is difficult for us to choose the $I_{cp,max}$. It is better to choose two different $I_{cp,max}$ to virgin and programmed cells because of the assumption of uniform distribution of the interface traps and inherent nitride charge for the virgin cell. Again, we can get the V_t profile by using Eq. (3) and Eq. (5) as shown in Fig. 4.6. The location of the passing through point from the source for the minority carrier was calculated to be 0.298 μm for the case of FBEI and 0.283 μm for the case of CHEI.

We can find the precise V_t peak value and width from fixed top charge pumping method due to its sensitive low current region in logarithmic scale. Since the passing through point for the minority carrier is taking place at 2V, so we just need the data beyond 2V. Figs. 4.7 (a) and (b) show the V_t peak for FBEI and CHEI. Finally, we insert the V_t peak in Fig. 4.6 (a) and 4.6 (b), we can obtain the whole V_t distribution for the virgin and programmed cells as shown in Fig. 4.8. The peak of FBEI is closer to the drain than that of CHEI.

(a)



(b)

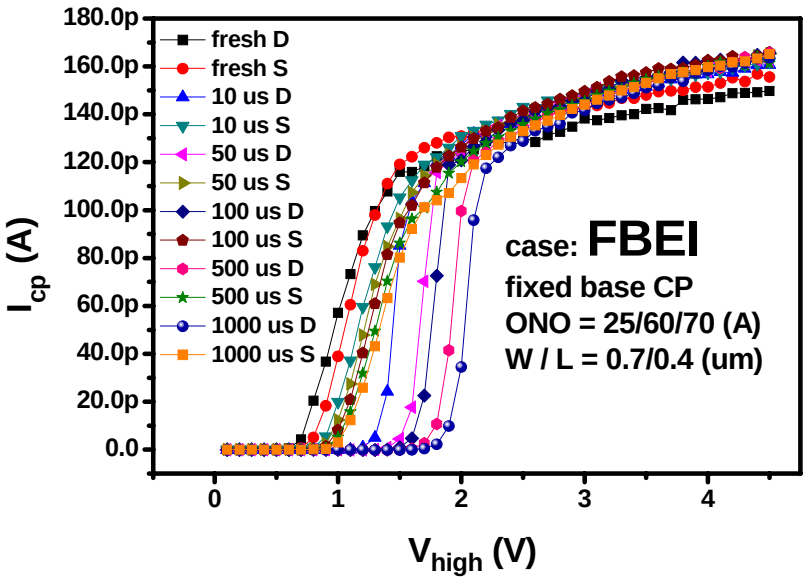
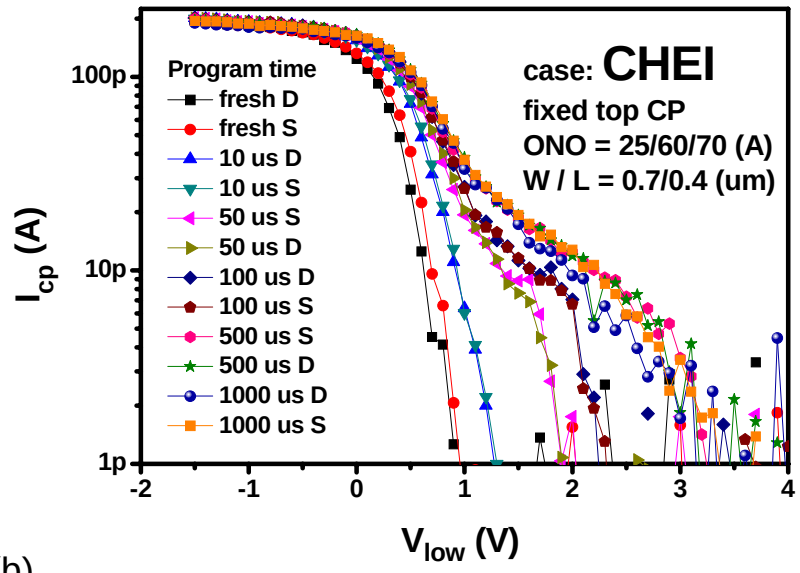


Fig. 4.4 Fixed base charge pumping measurement tested from drain and source, with different program time, **(a)** for CHEI and **(b)** for FBEI respectively.

(a)



(b)

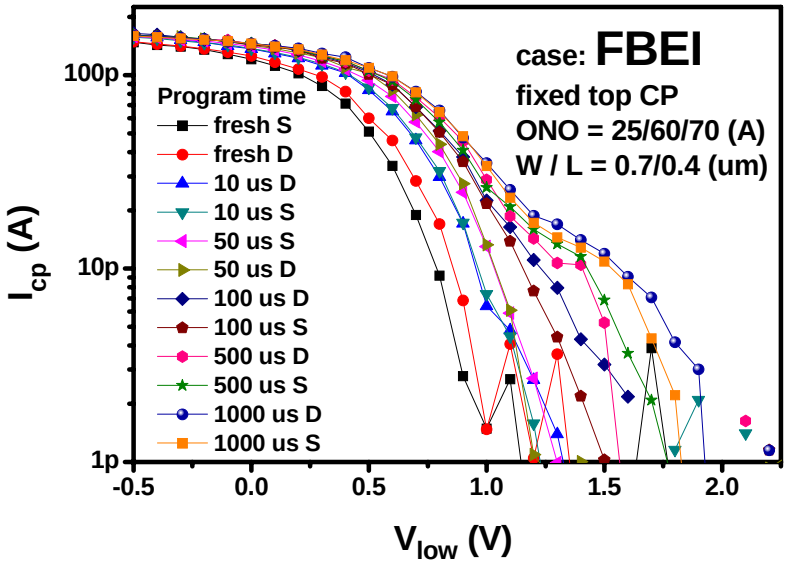
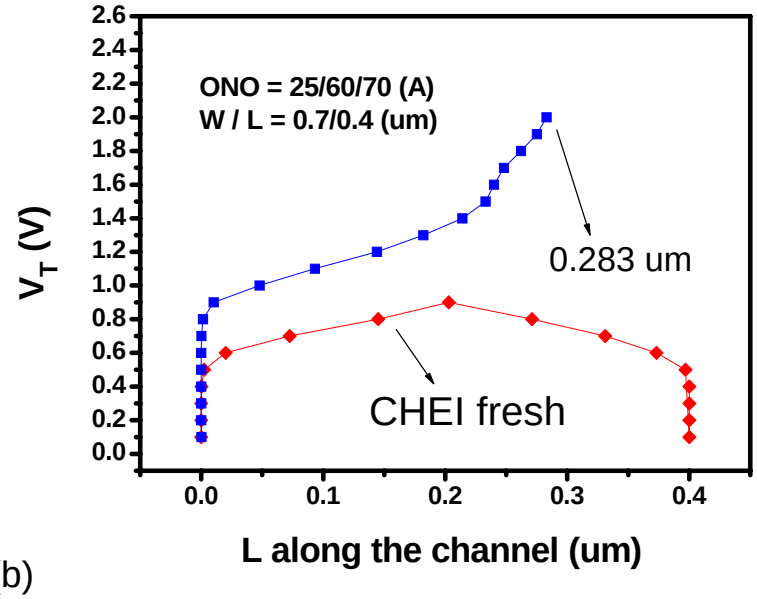


Fig. 4.5 Fixed top charge pumping measurement tested from drain and source, with different program time, **(a)** for CHEI and **(b)** for FBEI respectively.

(a)



(b)

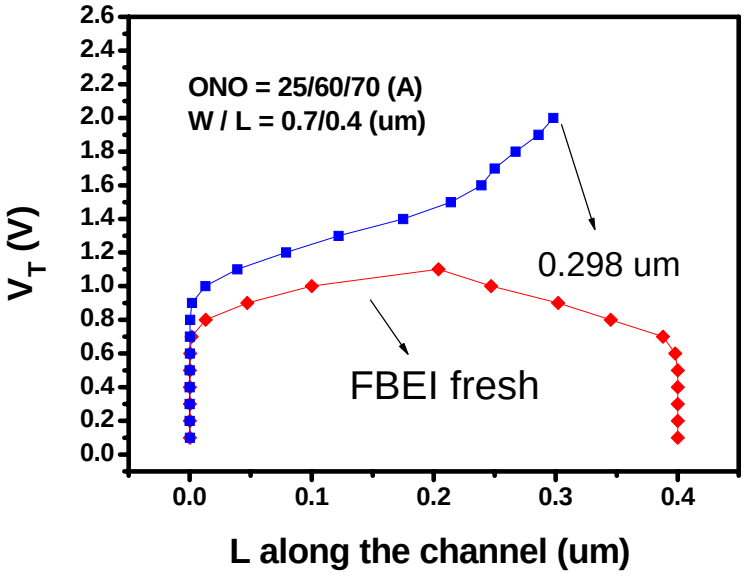
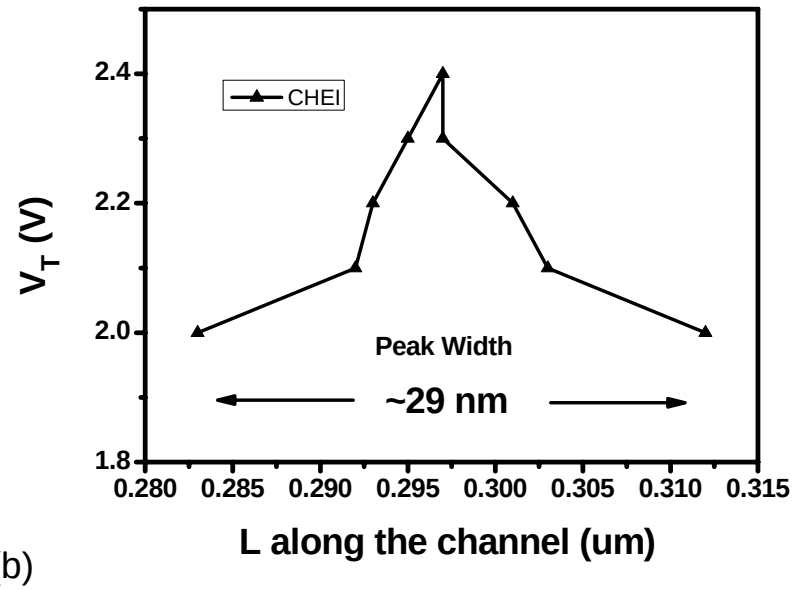


Fig. 4.6 Simplify the V_T profile by observing the passing through point tested from the source side, **(a)** for CHEI and **(b)** for FBEI.

(a)



(b)

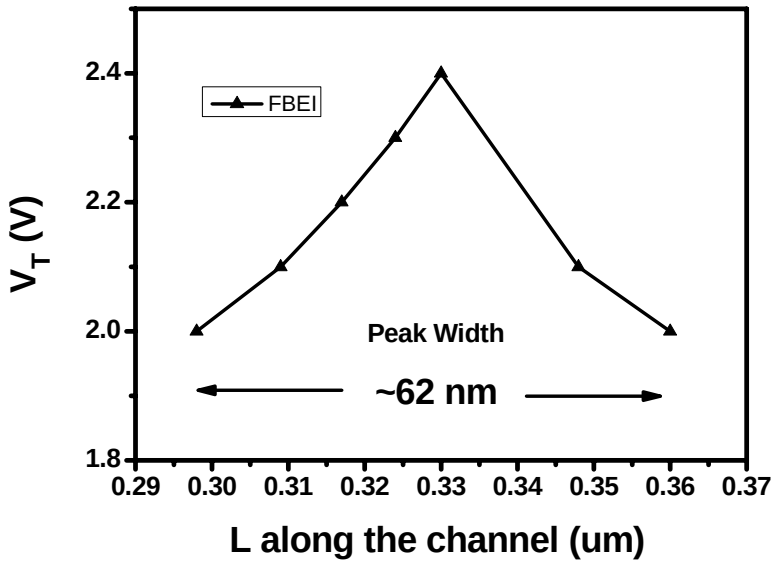


Fig. 4.7 (a) and **(b)** are the V_T peaks for CHEI and FBEI respectively.

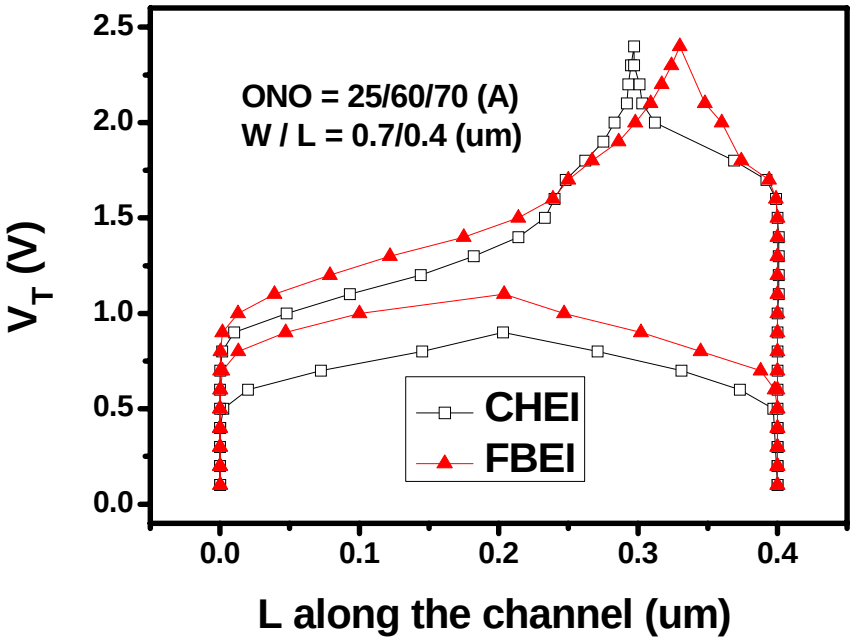


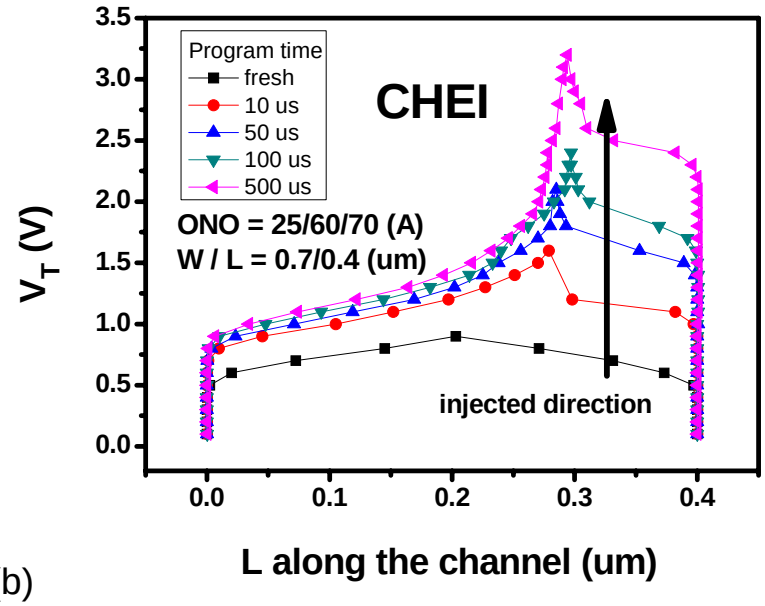
Fig. 4.8 The final V_T distribution for CHEI and FBEI.

4.2.4 Trapping Charge Behavior During Program

One of the applications by using the charge pumping method to obtain the charge profile is to observe the behavior of the trapping charge during programming. This application will be a powerful tool for us to clearly know more about the mechanism and the characteristics of the programming charges inside the nitride. Figs. 4.9 (a) and 4.9 (b) show the V_T profile after FBEI and CHEI programming, and different programming time are used to do the test.

As we can see, some differences between FBEI and CHEI programming are found and some are cross-checked. First, we can find out again at the same program time the peak width of the FBEI are much wider than that of the CHEI, which was compared in Fig.4.8. Second, the charge distribution near the drain side for the case of CHEI grows much faster than the case of FBEI. It will be explained later in Chapter 5 (two bits) why the V_T of 1ms CHEI programming is higher than that of FBEI and the location of charge storage will also be discussed. Finally, the most important dynamic analysis is to prove that the injecting direction of FBEI and CHEI are different. Let us throw our thoughts back to section 2.4 the programming scheme and as we know, the way CHEI programs is by accelerating the carrier until it is “hot” enough and then collide to the right direction to accomplish the injection. Unlike the CHEI, the way FBEI accelerate carrier is by using the forward bias at the program side and using positive gate bias to instantaneously suck the electrons up to accomplish the injection. The sketch map of CHEI and FBEI carrier injection assumption, as shown in Figs. 2.3 and 2.4 respectively, is now verified by direct measurement.

(a)



(b)

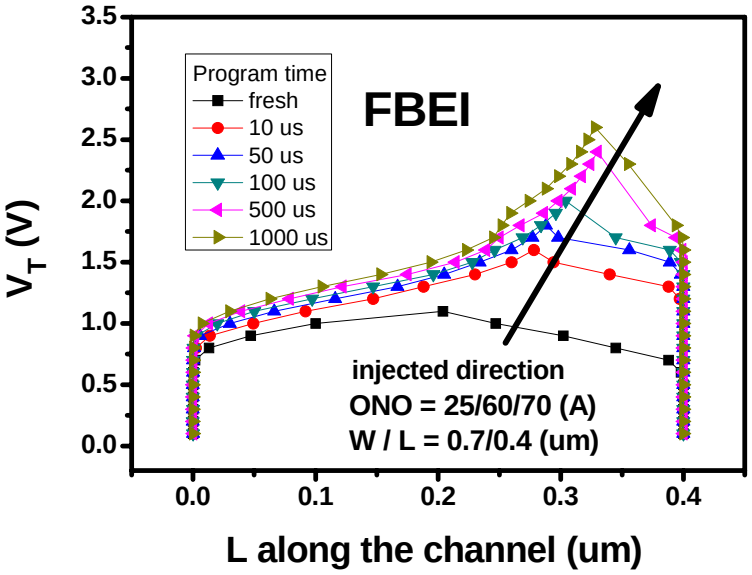


Fig. 4.9 The dynamic V_T distribution during different programming time for CHEI and FBEI respectively.

4.3 Temperature Effect on Read Current

4.3.1 Principle of Temperature Effect on Read Current

The temperature effect on the read current of a two-bit nitride storage Flash memory is investigated [25]. The characteristic of current-voltage (I-V) measurement has been a basic indicator of reading tool in all sorts of memories. With the addition of temperature, we gain more acquaintances about the storage charges. Besides, for commercial applications, cell operation at temperature up to 85^0C is required. The maintenance of sufficient read current window is necessary in a wide range of temperature.

Fig. 4.10 shows the I_D-V_G curves of the cell in erase state and program state at room temperature, and high temperature respectively. Here, if we read the data of the cell at V_G=3V, the read current window is about 8.77 uA at room temperature. But the read current window reduces to 7 uA when the cell operates at 85^0C. The degraded window may cause an error during data sensing at high temperature especially when the other reliability issues [26]-[27] are also taken into account. We can also see from the figure, it is found that the window narrowing mainly comes from the read current increment in the program state.

When a MOSFET is operated at high temperature, two mechanisms will affect the read current of device. One is the threshold voltage lowering, and the other is the mobility degradation [28]-[29]. The former mechanism may increase the currents in subthreshold and weakly on regions while the later mechanism may degrade the currents in fully on region. As a result, there is usually a crossover between the I-V

curves of device measured at room temperature and high temperature. In Fig. 4.10, the crossover behavior of the erase state cell happens at about V_G = 2.6V, and the current increment is insignificant in the weakly on region. But for the cell at program state, which has locally programmed charges at source side, the crossover behavior does not appear even when the applied V_G is as high as 3.5V. Because the impact of our 85^0C measurement on trapped charge profile is confirmed insignificant, it is inferred that the localized distribution of programmed charges enhance the subthreshold leakage current and results in a larger shift of ID-VG curve at high temperature. Thus, a delayed crossover point and a larger drain current increment are observed at the program state.

4.3.2 Experimental Results and Discussion

Fig. 4.11 shows not only the I-V curves representing fresh states, erase states and program states, but also the temperature effect on read current at different temperature, 25^0C and 85^0C. The shifting of the V_G, comparing to the fresh state, in region (1) and (2) are a manifestation of the magnitude of the stored charges for bit-1 and bit-2 respectively. At the same temperature, 25^0C, we can briefly say that CHEI programmed more charges than FBEI did not only at bit-1 but also at bit-2, which stands for CHEI will have narrower read window than FBEI as shown in region (3). Furthermore, comparing to 85^0C, CHEI degrades much more I_D than FBEI by the localized distribution of programmed charges which enhance the subthreshold leakage, as was discussed in the last section. In short, for commercial applications, it is easy to see that with the use of CHEI in high temperature, reliability will go worse after long-term operating stress such as P/E cycle, which probably bade much narrower window during the stress.

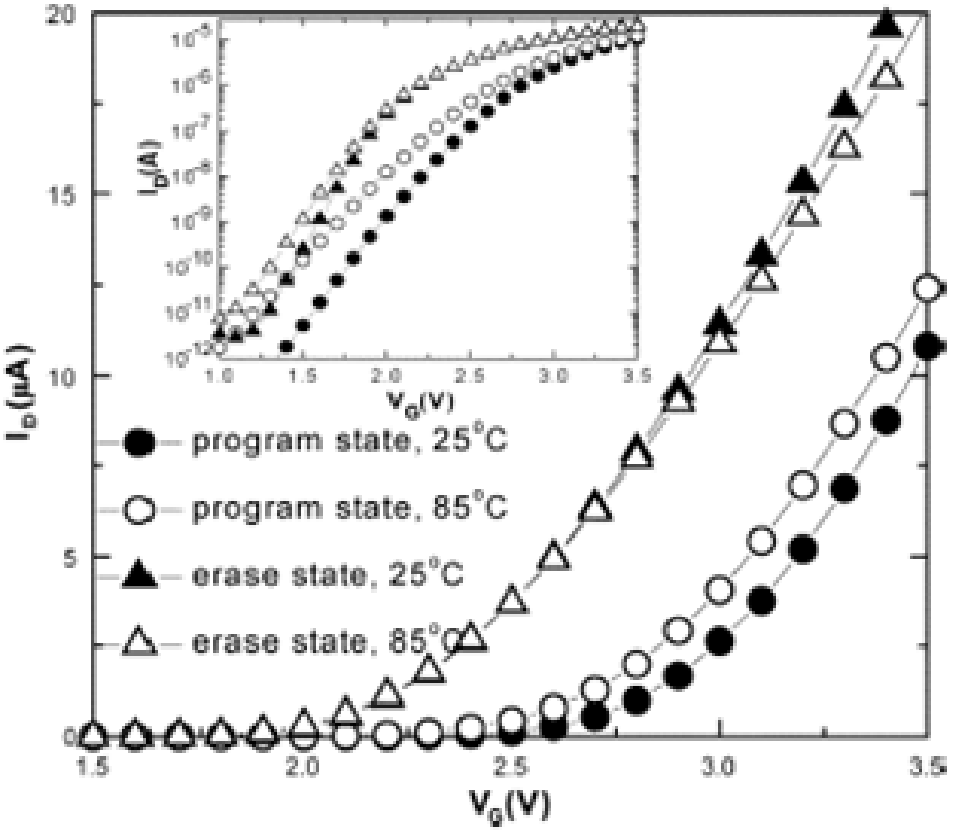


Fig. 4.10 Measured I_D-V_G curves for both linear and logarithmic scale, biased at V_D = 1.6V, of two cells in erase state and program state at room temperature and high temperature, respectively.

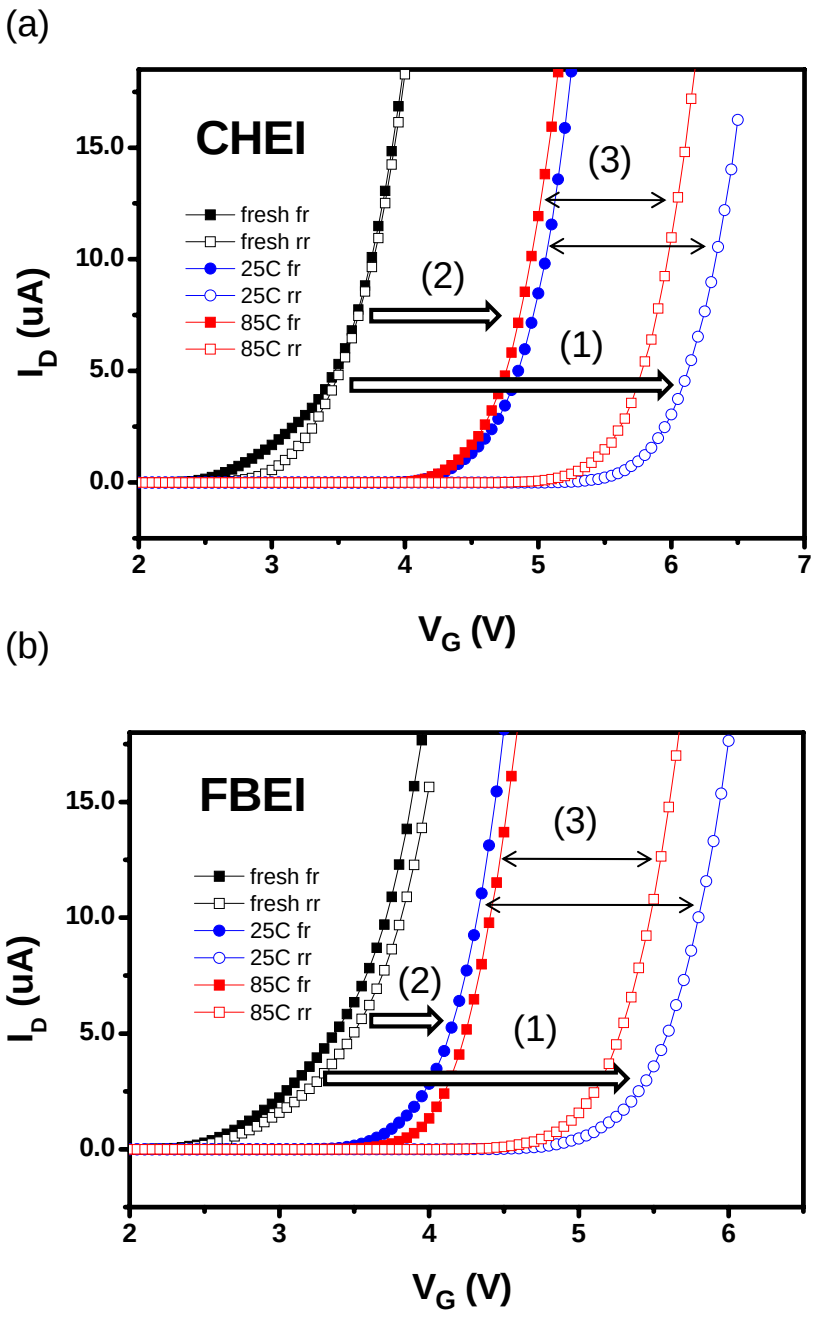


Fig. 4.11 Measured I_D-V_G curves for linear scale, biased at V_D = 2V, of two cells in fresh state, erase state and program state at room temperature and high temperature, respectively.

Chapter 5
Reliability Analysis of Endurance and Retention

5.1 Introduction

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 technologies. Channel hot electron (CHE) injection or channel initiated secondary hot electron (CHISEL) [30] injection is usually used for programming, since electron are injected and stored in a localized location in the silicon nitride layer along the channel. In most cases, band-to-band tunneling induced hot hole (BBHH) injection is usually used to erase the charge trapping memory, under which condition the drain is positively biased with the gate negatively biased. The injected holes may have different positions and ranges, which will cause the behavior of mismatch with the injected electrons. Furthermore, different location of charge distribution has different impacts on the cell's read characteristics, programming/erasing (P/E) performance and reliability.

In this chapter, we will also adopt BBHH to investigate the impacts after CHEI or BBEI, which also exhibit the localized characteristics presented in previous chapters. Finally, we will propose a well combination for programming and erasing schemes.

5.2 Three-Level Charge Pumping

5.2.1 Principle and Motivation of a Three-Level Charge Pumping

Charge pumping (CP) is a technique for studying traps at the Si-SiO_2 interface in MOS transistors. Compared to the traditional two-level techniques, an alternative approach, three-level CP, offers the possibility to obtain the interface trap parameters (emission times, interface trap density) as a function of the trap energy position in the silicon band-gap without any hypothesis concerning the values of trap capture cross sections [31]-[33]. A pulse with an additional level duration, t_h (hold time), is shown in Fig. 5.1 and is used to control the emission of electrons or holes.

In our near future work, three-level charge pumping method is used to separate the N_{it} and N_{ox}, which is hard to obtain using the traditional charge pumping.

In this section, the basic principle of the three-level charge pumping is described, some questionable assumptions are also discussed, and some phenomena different from two-level charge pumping are found. The experimental setup for the three-level charge pumping measurement is shown in Fig.2.6, where the three-level gate pulses are provided by a HP 8110A pulse generator, and a HP 4145C parameter analyzer is used to monitor the charge pumping current from the bulk. In this work, the source/drain and substrate electrodes are always grounded to prevent any undesirable stress caused by measurement.

The total three-level charge pumping current can be written as

$$I_{cp} = q \cdot W \cdot f \cdot \int_0^L N_{it}(x)dx \qquad \text{……..…(5a)}$$

where W is the channel width, f is the gate pulse frequency, N_{it} (x) is the local interface state density per area, and the integration from zero to L gives the lateral

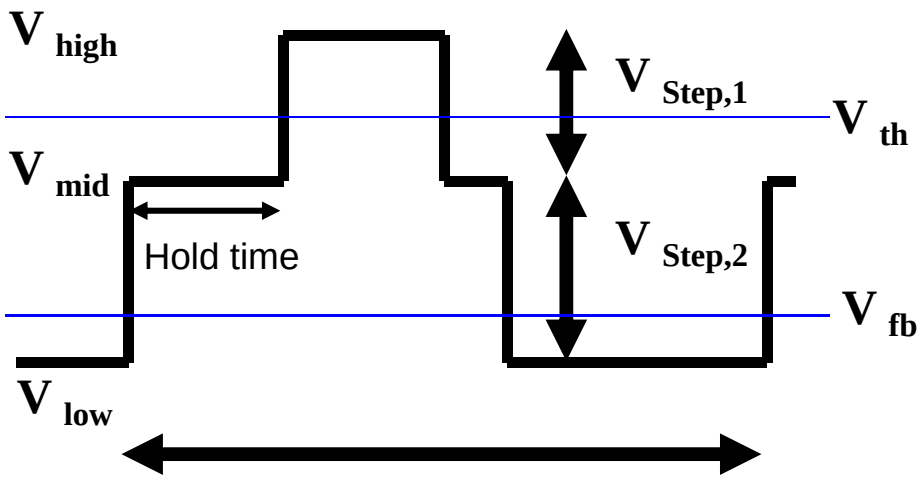


Fig. 5.1 Illustration of the parameters in a three level pulses.

length that contributes to the charge pumping current. Note that only the region where the surface potential fully swings from accumulation to inversion would contribute to the charge pumping current which is quite similar to the traditional one. Based on this fact, the separation of interface states and the oxide trapped charges on the Si-SiO_2 interface can be achieved.

First of all, the motivation of the experiment is described as follows. As shown in Fig. 5.2, the pulse generator gives a series of pulses with the base-level fixed at V_{low} and varying high-level V_{high} starting from a value slightly larger than $V_{high,min}$ to a fixed high-level $V_{high,max}$, like the traditional charge pumping, to ensure that the charge pumping current is able to reach its saturated value. Different from the previous CP method, as shown in Fig. 5.1, a middle-level V_{mid} is added so that there is a gap $V_{step,1}$ between V_{high} and V_{mid} ($V_{step,1}$ between V_{low} and V_{mid}). There is also a hold time at the V_{mid}. After this step, the charge pumping current attributed by the region of the channel can be measured. The results of two different pulses are shown in Fig. 5.2. The V_{mid} of the two pulses are both sweeping from 0V to 4V, with a fixed V_{low} of -1V and a fixed $V_{step,1}$ of 2V. Besides, one of the pulses has 125 ns hold time with duty cycle 25%, i.e. 250 ns for V_{high}, while the other one has no hold time but with duty cycle 37%, i.e. 375 ns for V_{high}. Note that the two pulses have the same “charging time”, 375 ns, but the voltage is a little bit different. In addition, also shown in Fig. 5.2, the curve of three-level charge pumping behaves like a standard fixed base charge pumping curve, which consists of a rising edge and a saturation region with the saturation value.

The comparison between different hold time of 0ns, 125ns, and 250ns, with the same duty cycle is shown in Fig. 5.3. The most important phenomenon in Fig. 5.3 is

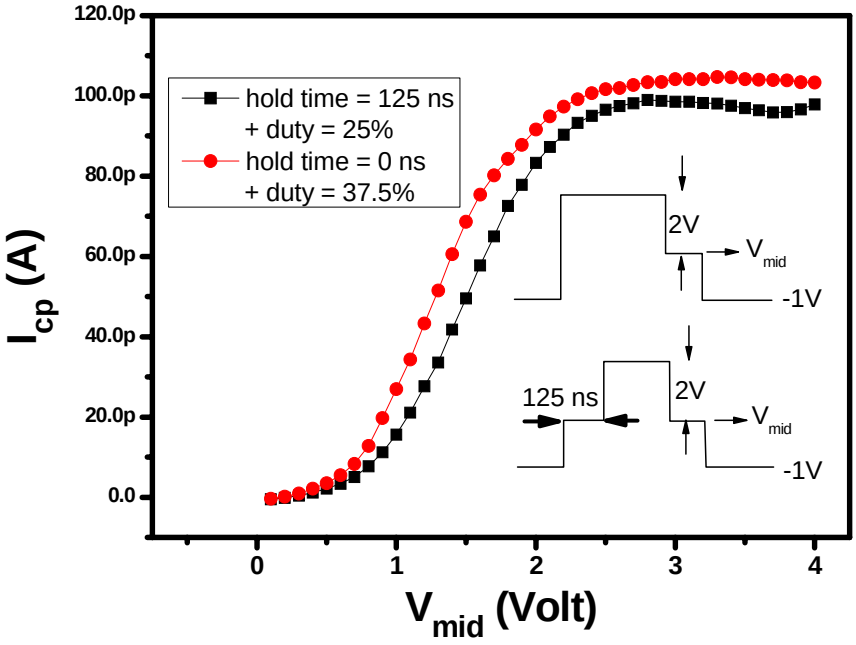


Fig. 5.2 I_{cp} as a function of V_{mid} but with conditions, 125 ns hold time with duty cycle 25% and no hold time but with duty cycle 37%.

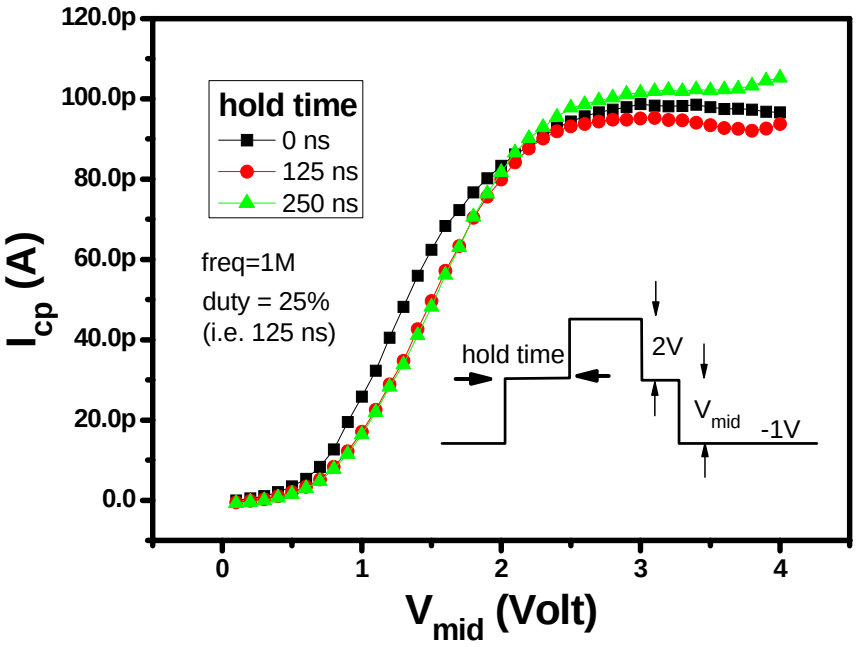


Fig. 5.3 I_{cp} as a function of V_{mid} but with different hold time of 0ns, 125ns, and 250ns.

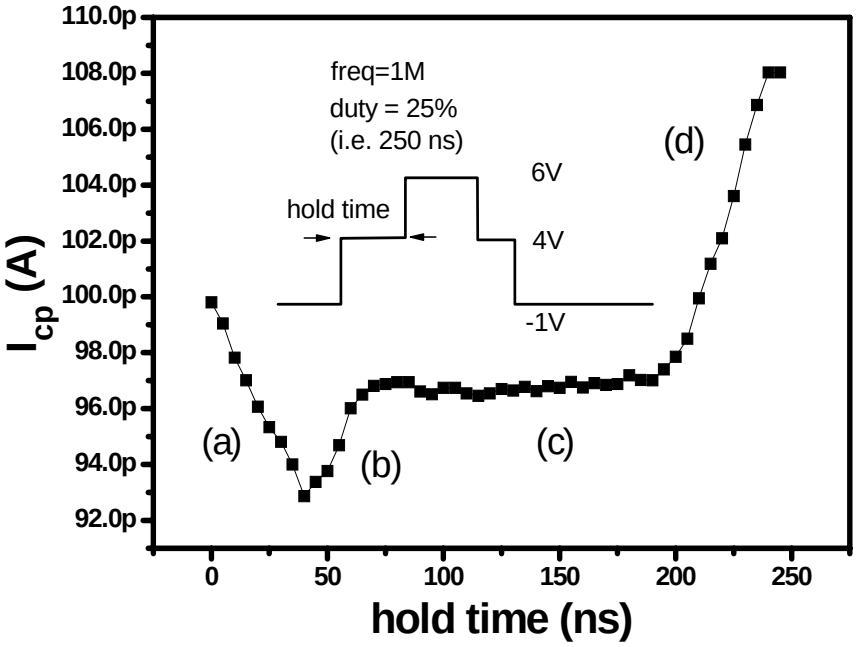


Fig. 5.4 I_{cp} as a function of hold time, as electrons and holes, several mechanisms happen in part **(a)**, **(b)**, **(c)** and **(d).**

the difference of the three curves after the crossover at 2.3V. It shows that with different holding time, we can get the slight differences of I_{cp}. To take a further study, by sweeping the hold time from 0 ns to 250 ns with the fixed voltage, 6V, 4V and 1V for V_{high}, V_{mid} and V_{low}, respectively, the result is shown in Fig. 5.4. This curve is the core of the main idea how we distinguish the interface states (N_{it}) from the oxide-trapped states (N_{ox}).

The current with different hold time measured by the three-level charge pumping is shown in Fig. 5.4. As we can see, there are four parts in the curve. In part (a), I_{cp} decreases rapidly as the hold time rises from 0ns to 40ns. Holes fill up the N_{it} and N_{ox} during V_{low}, and during V_{high} electrons recombine with the holes and be trapped in the N_{it}, N_{ox} to get the maximum I_{cp} at 0ns. However, as the hold time becomes longer, the emission of holes increase exponentially and some of the holes are recombined by the electrons during V_{mid} simultaneously, while most of them are recombined during V_{high} bias. I_{cp} will keep on decreasing until a turning point which comes to the part (b), and then rises up as the hold time goes on. This turning point is because the holes discharge is up to a maximum. Thus, with the recombining speed of the electrons at the V_{mid} bias still continue plus hole discharge no more, the I_{cp} at (b) zone will raise until the recombining speed of the electrons at V_{mid} bias goes to maximum, too. With no increase discharge and no increase recombination, the I_{cp} stays in the same level at part (c). After a long term of stable I_{cp} in part (c), it will rise again at part (d). This time I_{cp} rises is because of the filled electron discharge lesser as the hold time goes on. Two phenomena are noticeable. I_{cp} maximums are not the same when hold time is 0ns and 250ns; there is no turning point on the right side of the curve. Because the mobility of electron is much faster than that of the hole, so that the discharging behavior and the recombination will not be the same.

5.2.2 Experimental Results and Discussion

We used three-level charge pumping method to measure trap numbers and wish to understand more about the reliability of the SONOS bottom oxide during P/E cycle. Some of the experimental results will be discussed, and model of the programming mechanism will also be proposed.

In Figs. 5.5 (a) and (b), the device was programmed by CHEI and FBEI respectively and erased by BBHH (P/E cycle) several times. After erasing, it was measured by the three-level charge pumping as shown in the above. Obviously, there are several differences between the two illustrations, we will abstract some of the data and show them in the following figures. One of the obvious characteristic is that the amount of I_{cp} is proportional to the P/E cycles. As shown in Fig. 5.6(a), we took the I_{cp} data from region I in Figs. 5.5 and transformed into N_{it} by the equation 5a. The excess N_{it} is the traps produced by the P/E stress with a order of 10^{11}. Furthermore, Fig. 5.6(b) shows the device measured by the traditional charge pumping and is to compare with Fig. 5.6(a). As we can see in both Fig. 5.6 (a) and (b), there is the same tendency that the more P/E cycles, the more N_{it} generated. Most important of all, the excess traps caused by CHIE increase faster than those of FBEI. In short, CHEI produces much more traps than FBEI does during P/E stress which is not good for the reliability. Another overt characteristic that the I_{cp} slope of the hold time between 200 ns and 250 ns is quite different for CHEI and FBEI, as shown in Fig. 5.5 Region II. The slope is almost the same for FBEI, but not for CHEI whose slope becomes steeper every time after a P/E cycle. As shown in Fig. 5.7, we transform the I_{cp} of Region II into N_{it} again, but this time, this N_{it} stands for the interface-states which order is about 10^{9}. Comparing CHEI with FBEI in Fig. 5.7, we can find out that

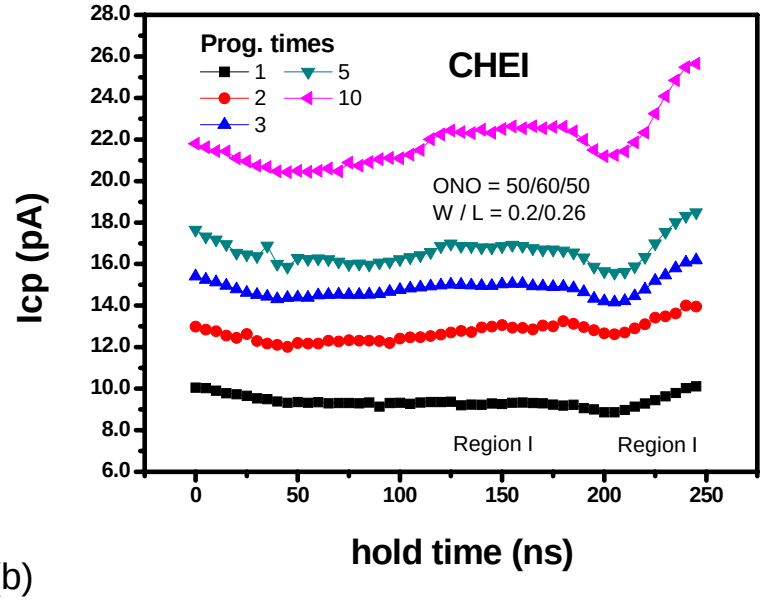
although FBEI creates more excess N_{it} in the beginning, the excess N_{it} caused by CHEI grows faster than that of FBEI again, which means CHEI creates more interface-states than FBEI in the long-term operation.

Fig. 5.8 shows the band diagram of the SONOS during program. Two paths of electron injection that will produce damages are considered. Path (3) is the way for electrons going directly through the bottom oxide which is usually called direct tunneling (DT) mechanism. Path (2) is the way that electrons first jumping into the oxide-trapped states, and then going through the bottom oxide, which is often called trap-assisted tunneling mechanism. The locations of the traps are what we are interested. Direct tunneling obviously creates the traps at the interface between Si and SiO_2 called N_{it}; trapped-assisted tunneling creates the traps more inside the bottom oxide called N_{ox}. Except for tunneling mechanisms mentioned above, another tunneling mechanisms called thermal emission, allows the electrons to jump over the oxide without damaging it.

From Figs. 5.6, Fig. 5.7, Chap 4, and the paths discussed above, we can now conclude that electrons prefer the path (2) CHEI programming mechanism. As we all know, CHEI programming mechanism is called “Lucky electron model”. Without enough energy, electron will not jump over the barrier. Even having enough energy, electrons do not tunnel without a correct direction, and that is why more damage is caused at the inside than the outside. Furthermore, with the more injection, the more damage will be created. On the other hand, the electrons programmed by FBEI might choose path (3) in the beginning, but they seem to have more energy than those of CHEI so that they may also choose the path (1), jumping across the oxide without creating a trap. Thus, even though P/E cycle times are the same, FBEI will cause less

damage either inside or outside the bottom oxide, which makes FBEI have better reliability than CHEI.

(a)



(b)

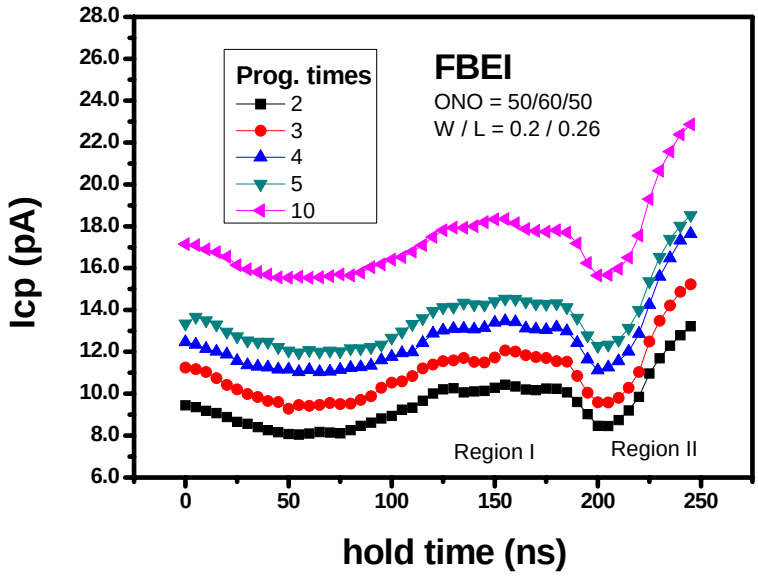
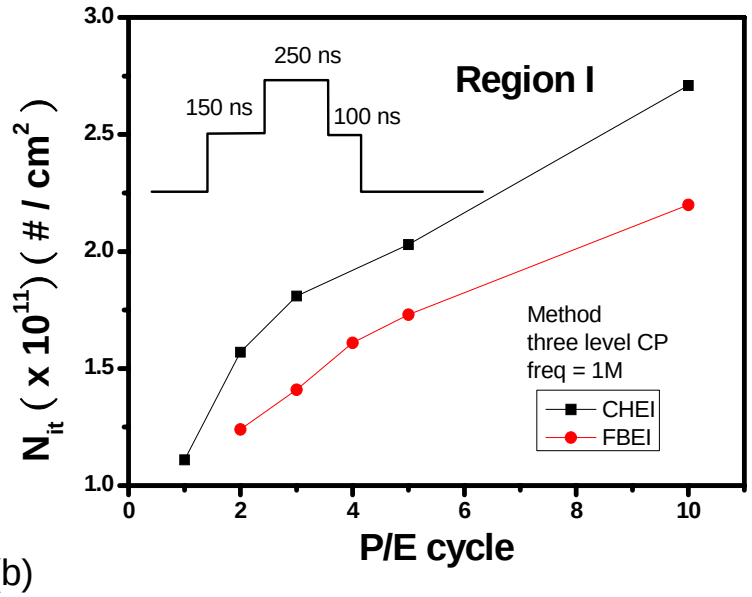


Fig. 5.5 After several cycles of P/E cycling, the measured I_{cp} vs hold time. Different signs of reliability issues are shown in Region I and Region II.

(a)



(b)

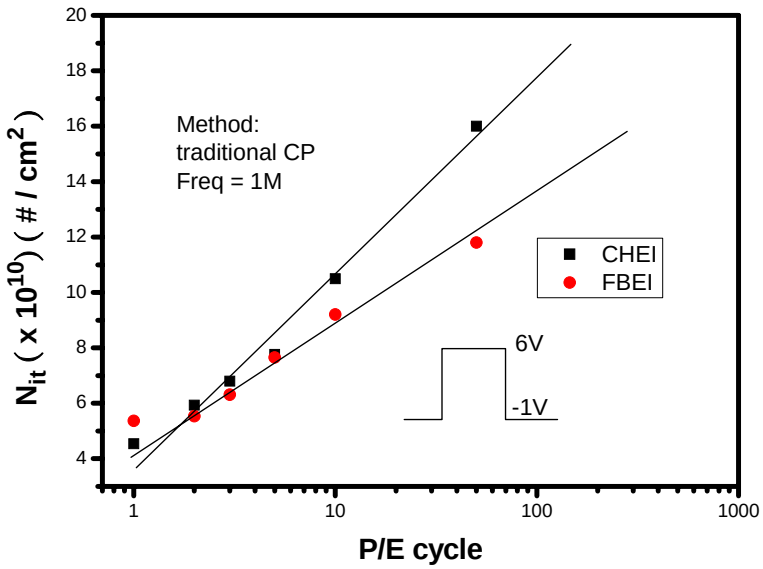


Fig. 5.6 N_{it} as a function of P/E cycle comparing CHEI with FBEI, for **(a)** using three level CP and for **(b)** using traditional CP.

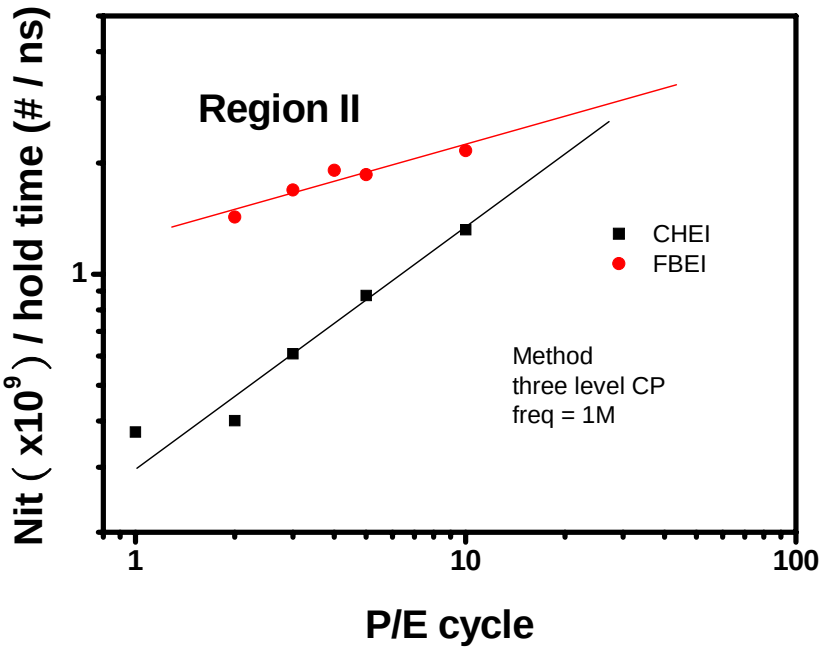


Fig. 5.7 N_{it} / hold time as a function of P/E cycles, comparing CHEI with FBEI, using three-level CP

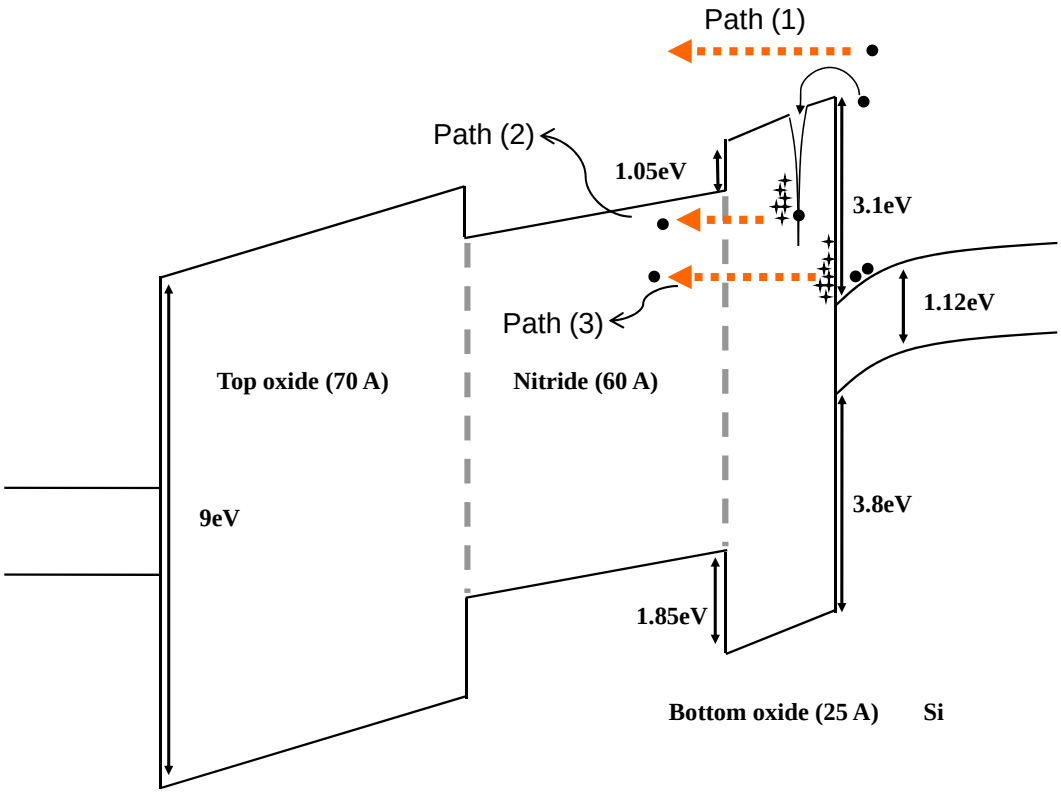


Fig. 5.8 Three different possible paths for the electrons programmed by FBEI and CHEI to jump across the barrier are shown in band diagram

5.3 Reliability Analysis of the Two Bits per cell Operation

5.3.1 The Scheme and Mechanism of BBHH

Band-to-band tunneling induced hot electron (BBHH) is employed in a P-channel cell which was firstly proposed by T. Ohnakado [34]-[35] for the application of the method to DINOR (Divided bit-line NOR) program operation. For N-channel cell, band-to-band tunneling will induce hot hole injection, and by using this method we can erase either N-channel SONOS or floating-gate flash memory cells. The operation scheme and band diagram are shown in Fig. 5.9 (a) and Fig. 5.9 (b), respectively, in which a positive drain voltage and a negative gate voltage are applied to the cell. The energy band diagram illustrates the band-to-band tunneling process in silicon in the gate/drain overlap region where a deep-depletion region is formed. Electron-hole pairs are generated by the tunneling of valence band electrons into the conduction band due to the high electric field induced band bending. The holes are accelerated by a lateral electric field toward the channel region and some of them obtain high energy. The injection of such hot holes into the nitride through the tunnel oxide is used for an erasing method to eliminate the localized trapped electron.

5.3.2 Proposed Program/Erase schemes for operation

With the understanding from the last section, we will propose a logic programming/erasing schemes, i.e., using FBEI for programming and BBHH for erasing operation, since FBEI has rather fast injection characteristic and low voltage operation so that it is suitable for programming.

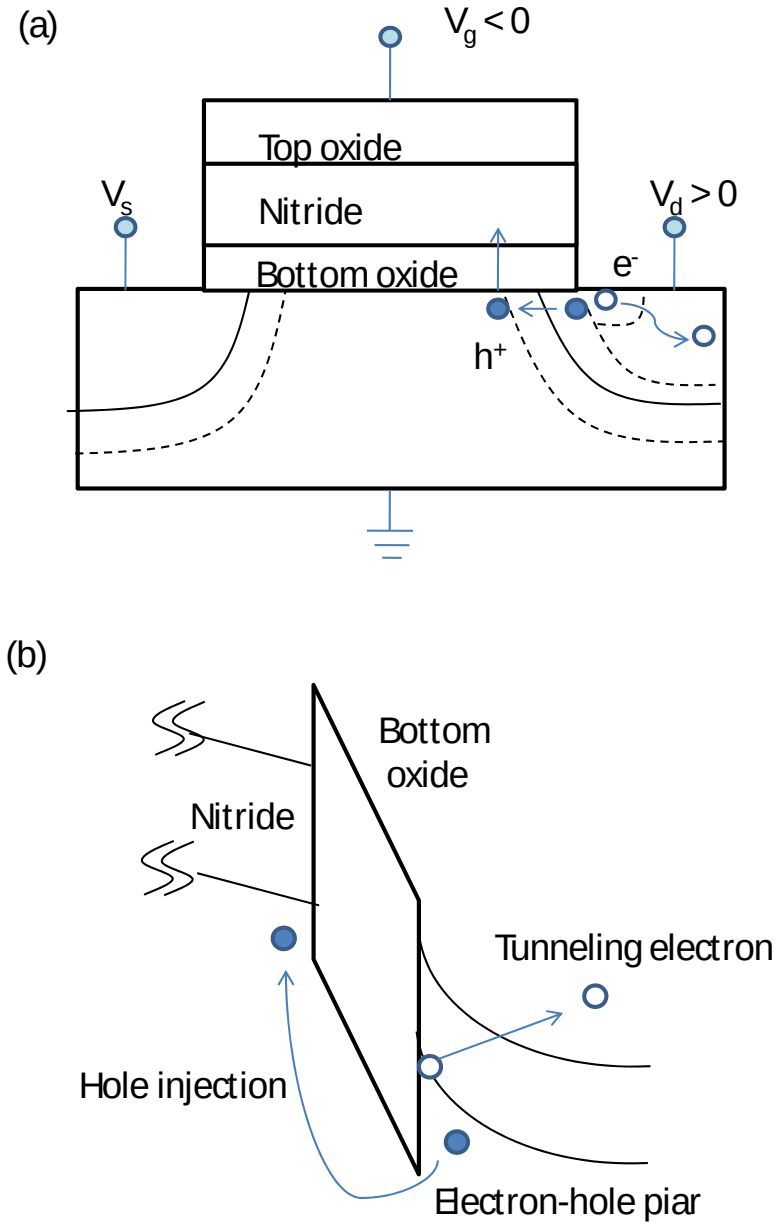


Fig. 5.9 (a) A deep depletion region is formed in the gate/drain overlap region.

(b) The energy band diagram illustrates the band-to-band tunneling process in the gate/drain overlap region. Electron-hole pairs are generated by the tunneling of valance band electrons into the conduction band and hole surmounts the oxide barrier into gate nitride. Electron is collected by the drain terminal.

5.3.3 Applications to Two Bits per cell Operation

First, we obverse the basic characteristics of P/E cycling combined with BBHH and FBEI in one bit per cell operation, as shown in Fig. 5.10. Not only the phenomenon of window closure is eliminated but also has a wider operation window about 3V. The data retention characteristic has been investigated in Fig. 5.11, in which about 2.8V window for 10 years after 10^4 cycling at room temperature cane be maintained.

Some of the two bits per cell issues such as cycling endurance and retention have been discussed [4][26]-[27]. However, we think that because some of the analysis from the above results of the experiment, FBEI could be another candidate for two bits storage operation due to its better characteristic. Moreover, we have proved that the localized charge behavior programmed by FBEI is much closer to the drain than CHEI in Chapter 4, by using the profiling techniques and I-V measurement with temperature effect. In this segment, we suggest this new programming method is an efficient instrument for two bits per cell applications, and in the following we will discuss the phenomena of two bit per cell operation for FBEI and CHEI.

First of all, a specific configuration with both source and drain tied together and by applying a higher voltage between S/D and gate will perform the injection of holes into the nitride layer, is shown in Fig. 5.12, with the so-called BBHH which can erase the cells to the low state. The energy band diagram had been discussed in section 5.3.2. Fig. 5.13 shows V_T as a function of V_{read} for bit-1 and bit-2. With the same $V_{read,}$ 1V, FBEI can reach 1V window at the state of programmed bit-1 and erased bit-2 compared to CHEI ones which can only reach 0.8V window. A smaller read drain

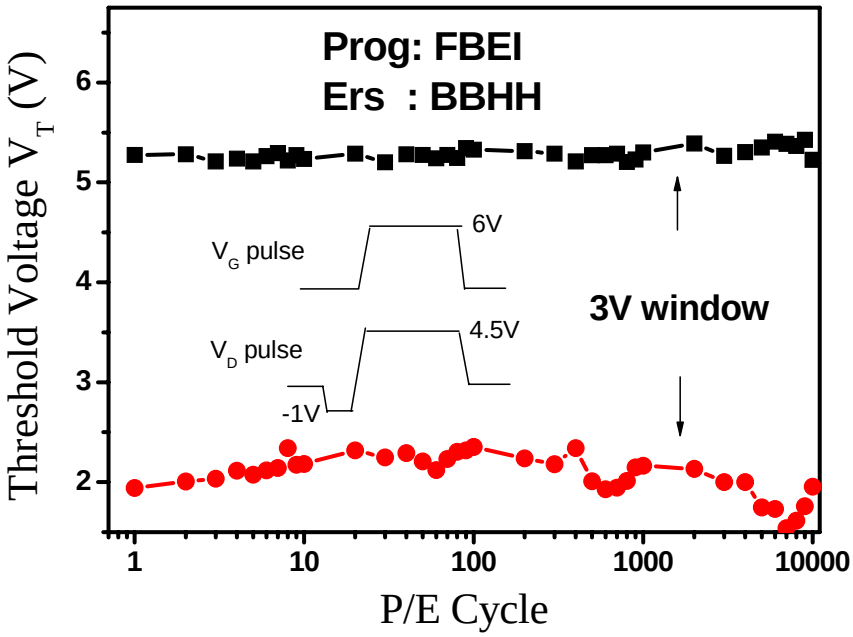


Fig. 5.10 P/E cycling characteristics after FBEI programming and BBHH erasing.

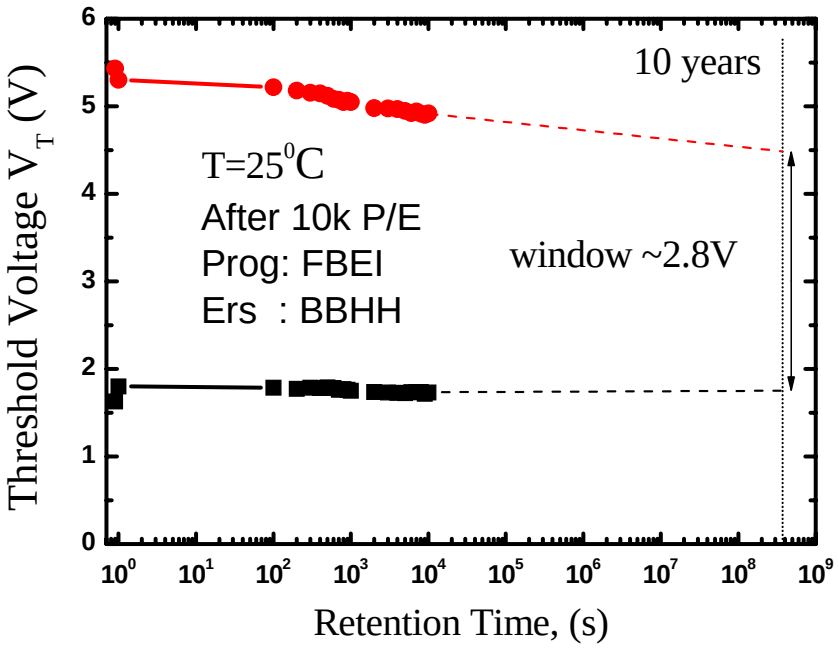


Fig. 5.11 Retention characteristics after 10000 FBEI and BBHH P/E cycles at room temperature.

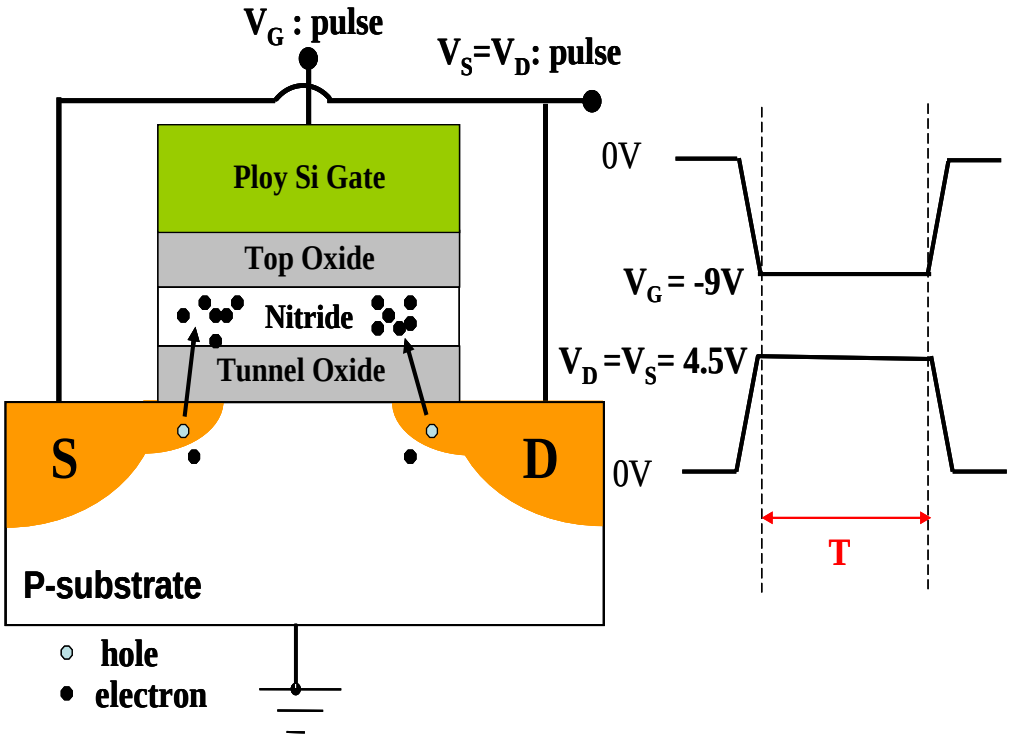


Fig. 5.12 Experimental set up and timing diagram for BBHI (band-to-band hot hole injection), with both source and drain being tied up together.

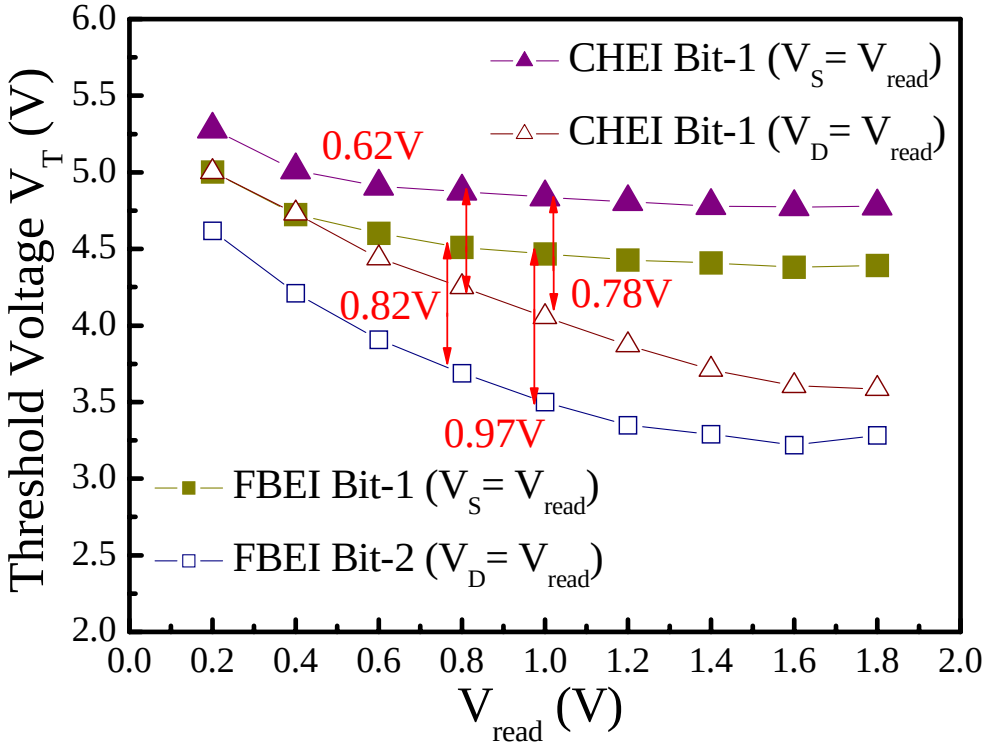


Fig. 5.13 V_T as a function of V_{read}, a larger window can be obtained by FBEI with smaller V_{read}.

voltage will lower the read disturb issues during long term reading. Consequently, it is an advantage, again, for FBEI comparing to CHEI to be a better choice for two bit per cell operation.

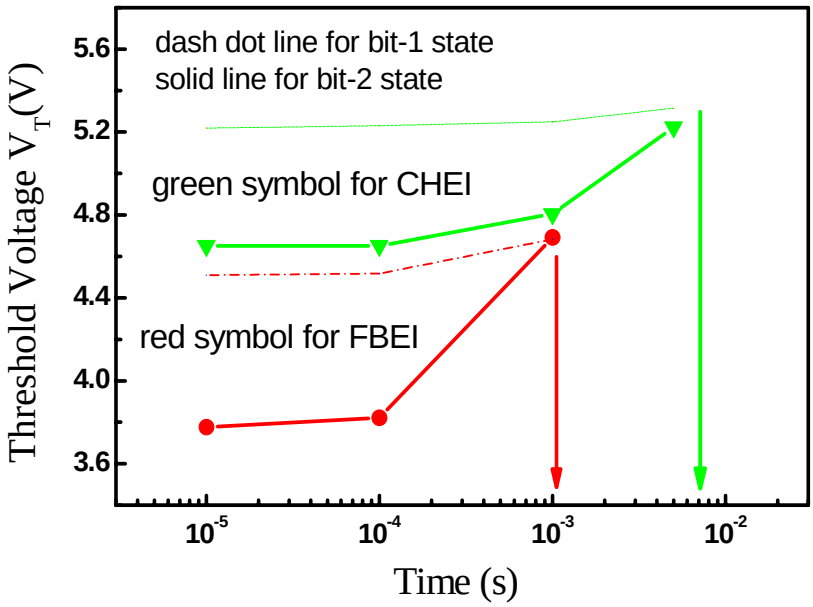
Fig. 5.14 (a) shows the bit-2 transient for FBEI and CHEI after bit-1 is programmed. Although CHEI has faster bit-1 programming speed, longer bit-2 programming time (~10m sec) is needed. On the other hand, using FBEI programming, it only needs ~1m sec reaching to the same high state as bit-1. In other words, shorter time is needed using FBEI to program bit-2, compared to CHEI programming, which again shows FBEI's advantage of the two bits applications. The physical mechanism why CHEI needs more programming time than FBEI was shown in Fig. 4.8, which the trapped charges at bit-1 will affect the charge inversion and acceleration in the channel near the drain side, such that will make the CHEI bit-2 programming time longer. Furthermore, in Fig. 5.14 (b), this mechanism proves again that 1ms FBEI programming time for bit-2 is still unchanged even if bit-1 is programmed by CHEI.

5.3.4 Endurance and Retention of 2-bit Operation

The cycling characteristics are shown in Fig. 5.15 (a) and Fig. 5.15 (b) with CHEI programming and FBEI programming, respectively, in which good endurance behavior is achieved. Both the cells, in Figs. 5.15 (a) and 5.15 (b), are programmed with 1 msec and the cycle sequence is program bit-1 => program bit-2 => erase bit-1 and bit-2 together, as shown in Fig. 5.12. The data retention after cycling is shown in Figs. 5.16 (a) and 5.16 (b), where an acceptable value of the four different states is achieved after 10 years. The retention behavior of cell is programmed by CHEI and

FBEI, respectively, which has shown better retention for FBEI. As we can see, FBEI still has 1V window after ten years in Fig. 5.16 (b) and is larger than the window of CHEI retention (~0.5V) in Fig. 5.16 (a). More importantly, the retention behavior can shows the reliability of the SONOS bottom oxide, which leads to with the using by FBEI scheme, less damage can be caused. In short, for the 2 bit per cell operation, the new scheme exhibits much better retention and endurance characteristics.

(a)



(b)

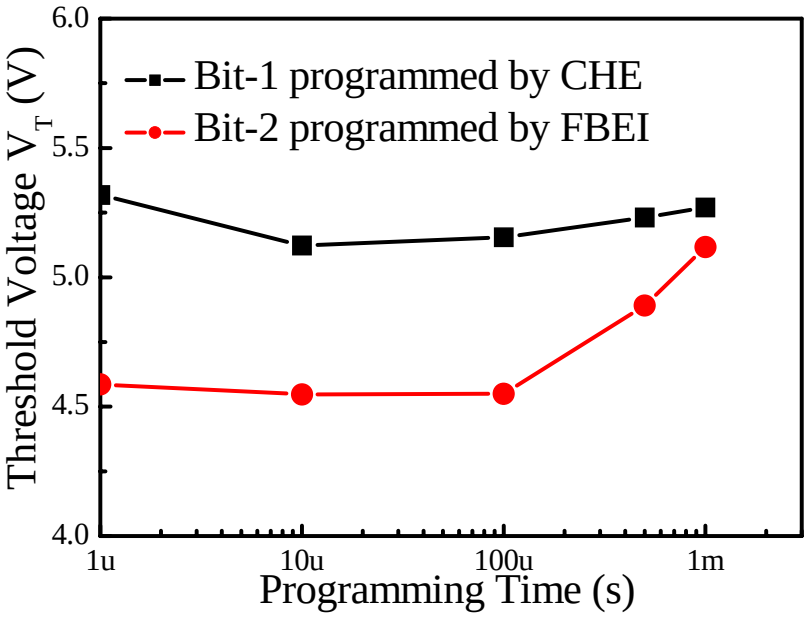
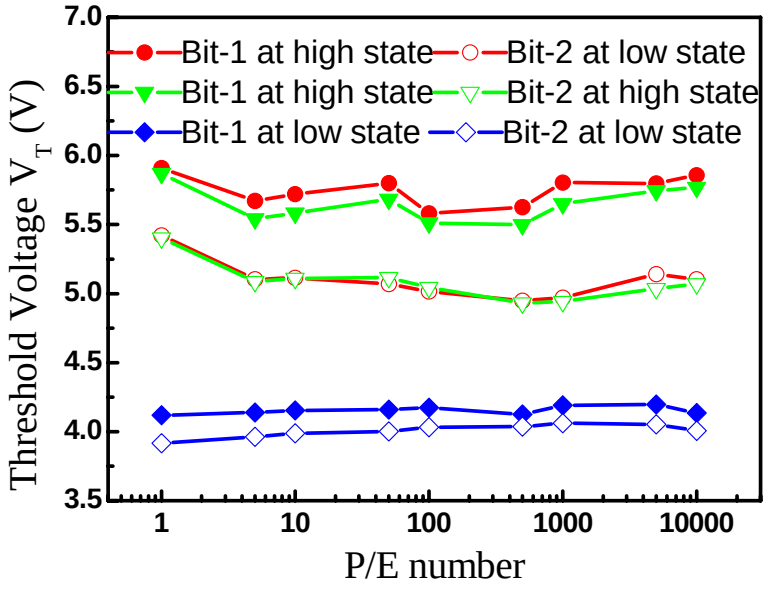


Fig. 5.14 (a) The bit-2 transient for FBEI and CHEI after bit-1 is programmed. Bit-2 programmed by FBEI can reach the same threshold voltage as bit-1 in shorter time compared with CHEI.

(b) The programming speed for bit-2 by FBEI while bit-1 has been programmed by CHEI. It is found that 1m sec FBEI programming time is still unchanged.

(a)



(b)

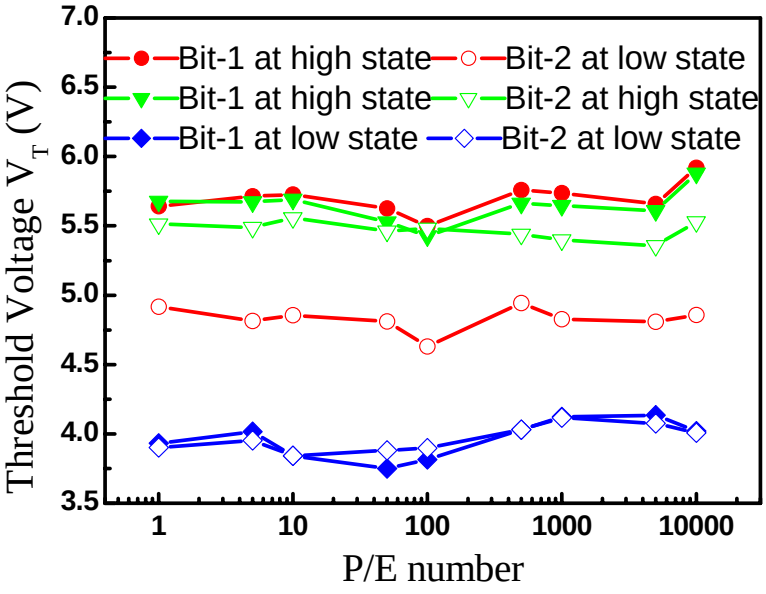


Fig. 5.15 Endurance characteristics of two-bit-per-cell application, **(a)** using CHEI programming and BBHH erasing and **(b)** using FBEI programming and BBHH erasing.

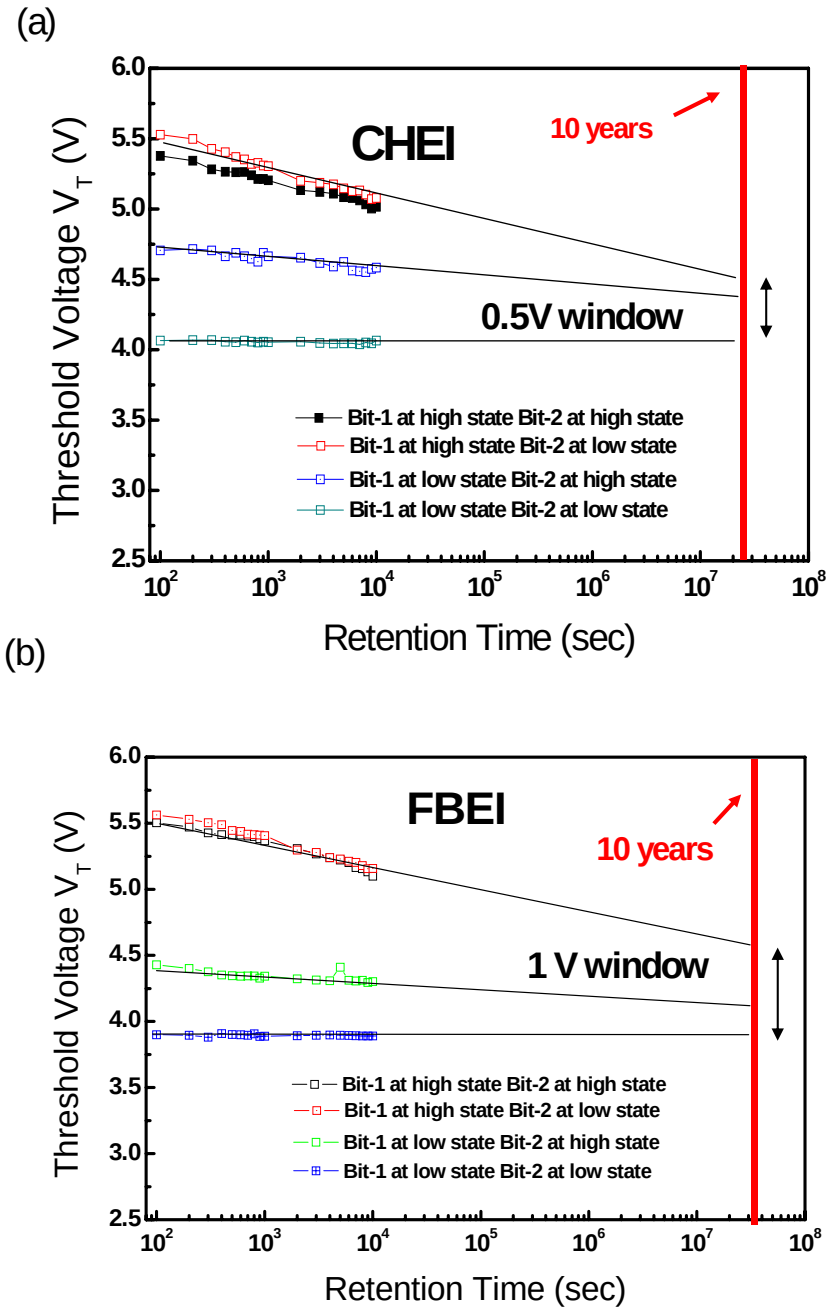


Fig. 5.16 Retention behaviors after 10k P/E cycles of Bit-1 and Bit-2 in four different states, respectively. The programming mechanism for **(a)** CHEI and **(b)** FBEI.

Chapter 6
Summary and Conclusion

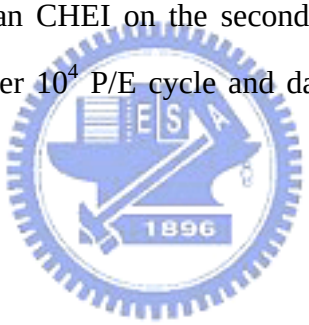
SONOS devices with trapped charges in the nitride layer have received much more attention, since SONOS non-volatile-memory devices are inherently free from drain to floating-gate coupling and are able to show a 2-bit/cell storage scheme that utilizes different physical locations to store the injected charges. In this thesis we proposed a new scheme for electron injection (FBEI). Comparing to PASHEI, FBEI has similar behaviors since they are pulse number dependence. Moreover, FBEI and CHEI are very promising since they make the device be able to operate as a 2-bit/cell, due to a large number of charges stored in the specific regime of a cell.

First, we used the I-V measurement with different temperature to observe the V_t shift, both forward read (FR) and reverse read (RR), after CHEI and FBEI writing. FBEI shows better window between the difference of FR and RR than CHEI, and the bigger window is still unchanged after 85^0C during the reading process. By using FV_b CP method and FV_t CP method, we can extract the V_T profile, in which it reveals that most of the charge distributions are close to the drain for FBEI and CHEI. However, the stored charge after FBEI writing shows more concentrated near the drain side than CHEI. Because of this characteristic, it proves and explains again why the window of FBEI is larger than that of CHEI. Moreover, from the method as above, we can obtain the injection direction for both FBEI and CHEI during programming.

Then, we utilized a three-level charge pumping to recognize the programming mechanism. After the measurement, we made an inference of the programming mechanism. With the illustration of the band diagram, we think that most of the charges injected by FBEI might jump over the tunnel oxide barrier. And, the charges

injected by CHEI will tunnel through the bottom oxide and bump into the nitride as lucky electron model which will cause more damage than FBEI. Therefore, FBEI has better reliability than CHEI does.

Furthermore, based on the experimental results, it was found that FBEI might become a new candidate for 2-bit operation. Thus, we also made the study on 2-bit operation for FBEI comparing with CHEI, with a specific configuration with both source and drain tied together to erase by BBHH (Band-to Band induced Hot Hole), which is utilized to neutralize the stored electrons. The transient characteristic of the second bit shows that we can obtain faster programming speed by FBEI than CHEI. Even though the first bit is programmed by CHEI, programming FBEI on the second bit can still make it faster than CHEI on the second bit. Finally, we can receive a better endurance for FBEI after 10^4 P/E cycle and data retention (~1V for program and erase state) for 10 years.



References

[1] M. H. White, D.A. Adams, and J. Bu, “On the go with SONOS,” *Circuits and Devices Magazine*, IEEE , Vol. 16, No. 4, pp. 22-31, July 2000.

[2] W. J. Tsai, C. C. Yeh, N. K. Zous, C. C. Liu, S. K. Cho, T. W. S. Pan, and C. Y. Lu, “Positive oxide charge-enhanced read disturb in a localized trapping storage flash memory cell,” *IEEE Trans. on Electron Devices*, Vol. 51, no. 3, Mar. 2004.

[3] B. Y. Choi, B. G. Park, J. D. Lee, H. Shin, Y. K. Lee, K. H. Bai, D. D. Kim, D. W. Kim, C. H. Lee, and D. Park, “Reliable 2-bit/cell NVM technology using twin SONOS memory transistor,” *Electronics Letters*. Vol. 41 no. 19, 15th Sep. 2005.

[4] M. K. Cho and D. M. Kim, “High performance SONOS memory cells free of drain turn-on and over-erase: Compatibility issue with current Flash technology,” *IEEE Electron Device Lett*., vol. 21, pp. 399-401, Oct. 2000.

[5] T. Y. Chan, J. Chen, P.K. Ko, and C. Hu, “The impact of gate-induced drain leakage current on MOSFET scaling,” in *IEDM Tech. Dig*., pp. 718-721, 1987.

[6] T. Y. Chan, K. K. Young, and C. Hu “A true single-transistor oxide-nitride-oxide EEPROM device,” *IEEE Electron Device Lett*., vol. EDL-8, pp. 93-95, 1987.

[7] S. H. Gu and T. Wang, “Characterization of programmed charge lateral distribution in a two-bit storage nitride flash memory cell by using a charge-pumping technique,” *IEEE Trans. on Electron Devices*, vol. 53, no. 1, Jan. 2006.

[8] C. C. Yeh, W. J. Tsai, M. I. Liu, T. C. Lu, S. K. Cho, C. J. Lin, T. Wang, S. Pan, and C. Y. Lu, “PHINES: a novel low power program/erase, small pitch, 2-bit per cell Flash memory,” in *IEDM Tech. Dig*., pp. 931-934, 2002.

[9] M. S. Liang and T. C. Lee, “A hot-hole erasable memory cell,” *IEEE Electron Device Lett*., Vol. 7, no. 8, pp. 465-467, 1986.

[10] C. 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 Electron Devices*, Vol. 45, no. 2,pp. 512- 520, 1998.

[11] W. Chen and T. P. Ma, “Channel hot-carrier induced oxide charge trapping in NMOSFET'S,” in *IEDM Tech. Dig.*, pp. 731-734, 1991.

[12] W. Chen, A. Balasinski, and T. P. Ma, “Lateral profiling of oxide charge and interface traps near MOSFET junctions,” *IEEE Trans. on Electron Devices*, Vol. 40, no. 1, pp. 187-196, 1993.

[13] C. Chen and T. P. Ma, “Direct lateral profiling of both interface traps and oxide charge in thin gate MOSFET devices,” in Symp. on *VLSI Tech Dig*., pp. 230-231, 1996.

[14] A. M. Martirosian and T. P. Ma, “Improved charge-pumping method for lateral profiling of interface traps and oxide charge in MOSFET devices,” in *IEDM Tech. Dig.* pp. 93-96. 1999.

[15] C. Y. Lu and K. S. Chang-Liao, “Minimized constrains for lateral profiling of hot-carrier-induced oxide charge and interface traps in MOSFETs,” *IEEE Electron Device Lett*., Vol. 25, no. 2, pp. 98-100, 2004.

[16] H. Pang, L. Pan, L. Sun, Y. Zeng, Z. Zhang, and J. Zhu, “A new method based on charge pumping technique to extract the lateral profiles of localized charge trapping in nitride,” *ESSDERC*, 2005.

[17] L. Sun, L. Pan, H. Pang, U. Zeng; Z. Zhang, John Chen, and J. Zhu, “Characteristics of band-to-band tunnel hot hole injection for erasing operation in charge-trapping memory,” *Jpn. J. Appl. Phys*., Vol. 45, No. 4B, 2006.

[18] M. Rosmeulen et al., “Characterization of the spatial charge distribution in local charge-trapping memory devices using the charge-pumping technique,” *Solid-State Electr.*, vol. 48, pp. 1525-1530, 2004.

[19] J. S. Sim, “Observation of the lateral redistribution of locally trapped charge in SONOS memory cells,” *SMDL Annual Teprot 2003*.

[20] L. Zhizheng and T. P. Ma, “A new programming technique for flash memory devices,” in *Symp. on VLSI Tech*., pp. 195-198, 1999.

[21] L. Zhizheng and T. P. Ma, “A low voltage erase technique for DINOR flash memory devices,” in *Symp. on VLSI Tech. Dig*., pp. 17-18, 1999.

[22] N. Tsuji, N. Ajika, K. Yuzuriha, Y. Kunori, M. Hatanaka, and H. Miyoshi, “New erase scheme for DINOR flash memory enhancing erase/write cycling endurance characteristics,” in *IEDM Tech. Dig*., pp. 53-56, 1994.

[23] S. Haddad, V. H. Chan, H. Fang, Y. Tang, M. Ramsbey, A. Wang, S. Yu, C. Chang, and J. Lien, “New erase scheme suitable for low power flash memory application,” in *Symp. on VLSI Tech. Dig*., pp. 52-53, 1996.

[24] T. Ohnakado, H. Onoda, O. Sakamoto, K. Hayashi, N. Nishioka, H. Takada, K. Sugahara, N. Ajika, and S. Satoh, “Device characteristics of 0.35 μm P-channel DINOR flash memory using band-to-band tunneling-induced hot electron (BBHE) programming,” *IEEE Trans. on Electron Devices*, Vol. 46, no. 9, pp. 1866-1871, 1999.

[25] M. Y. Liu, Y. W. Chang; N. K. Zous, I. Yang, and T. C. Lu et al., “Temperature effect on read current in a two-bit nitride based trapping storage flash EEPROM cell,” *IEEE Electron Device Letters*, Vol. 25, no. 7, July 2004.

[26] W. J. Tsai, N. K. Zous, C. J. Liu, C. C. Liu, C. H. Chen, T. H. Wang; S. Pan, C. Y. Lu, and S. H. Gu, “Data retention behavior of a SONOS type two-bit storage Flash memory cell,” *in IEDM Tech. Dig.*, pp.32.6.1-32.6.4, 2001.

[27] C. C. Yeh, W. J. Tsai, T. C. Lu, S. K. Cho, T. Wang, S. Pan, and C. Y. Lu, “A modified read scheme to improve read disturb and second bit effect in a scaled MXVAND Flash memory cell,” *in Proc. Non-Volatile Semiconductor Memory Workshop*, pp.44-45, 2003.

[28] F. S. Shoucair, ” Scaling, subthreshold, and leakage current matching characteristics in high-temperature (25°C-250°C) VLSI CMOS devices,” *IEEE Trans. COMP., Hybirds, Manufact. Technol.*, Vol. 12, pp. 780-788, Dec. 1989.

[29] F. S. Shoucair, “Design considerations in high temperature analog COMS integrated circuits,” *IEEE Trans. COMP., Hybirds, Manufact. Technol.*, Vol. CHMT-9, pp. 242-251, Sept. 1986.

[30] J. D. Bude *et* al., "Secondary electron Flash - a high performance, low power Flash technology for 0.35 pm and below," in *IEDM Tech. Dig.*, pp. 279-282. 1997.

[31] W. L. Tseng, “A new charge pumping method of measuring Si-SiO_2 interface states,” *J. Appl. Phys.*, Vol. 62, p.591, July 1987.

[32] N. S. Saks and M. G.. Ancona, “Determination of interface trap capture cross sections using three-level charge pumping,” *IEEE Electron Device Letters*, Vol. 11, no. 8, Aug. 1990.

[33] M. J. Kivi and S. Taylor, “Assessment of nMOSFET degradation using three level charge pumping,” *in International Conference on Microelectronics*, Vol. 1, Sept. 1995.

[34] T. Ohnakado, K. Mitsunaga, M. Nunoshita, H. Onoda, K. Sakakibara, N. Tsuji, N. Ajika, M. Hatanaka, and H. Miyoshi, ” Novel electron injection method using band-to-band tunneling induced hot electrons (BBHE) for flash memory with a P-channel cell,” in *IEDM Tech.Dig.*, pp. 279-282, 1995.

[35] T. Ohnakado, H. Onoda, O. Sakamoto, K. Hayashi, N. Nishioka, H. Takada, K. Sugahara, N. Ajika, and S. Satoh, “Device characteristics of 0.35 μm P-channel DINOR flash memory using band-to-band tunneling-induced hot electron (BBHE) programming,” *IEEE Trans. on Electron Devices*, Vol. 46, no 9, pp. 1866-1871, 1999.