Chapter 1
Introduction

1.1 The Motivation of This Work

Since 1990, nonvolatile semiconductor memory (NVSM) has been the technology driver of the semiconductor industry [1]. At the present time a floating gate FLASH EEPROM dominates the NVSM market. The floating gate type of FLASH EEPROM is impossible to be scaled down to beyond 0.18 um due to the difficulty in scaling the tunnel oxide [2]. On the contrary, the other multi-dielectric nonvolatile memory potentially can be scaled to below this size.

The development of the multi-dielectric nonvolatile memory device is shown in Fig. 1.1. The MNOS (metal – nitride – oxide – silicon) device was invented in 1967 [3] with a relative thick nitride layer and high program/erase voltage. However, MNOS device has its scaling limit with nitride thickness of 12.3nm [4]. For nitride thickness less than 12.3nm, the gate charge injection becomes significant, and proper initial charges for 10 years retention can not be stored in the nitride. Because of the scaling limit in MNOS, Suzuki et al [5] proposed a MONOS (metal-oxide-nitride-oxide-silicon) cell with a layer of oxide above the nitride. Later, silicon-gate SONOS (silicon-oxide-nitride-oxide-silicon) device was proposed by Chen and Yatsuda[6]. MONOS/SONOS provides one of the most promising structures with an added blocking oxide, holes are blocked from gate injection to the nitride as a result of its high barrier (3.9eV). So, memory window is widened and nitride thickness can be scaled. With thinner effective thickness, MONOS/SONOS can be operated at lower voltage.

The differences between these two types (floating gate type and SONOS type) of devices are

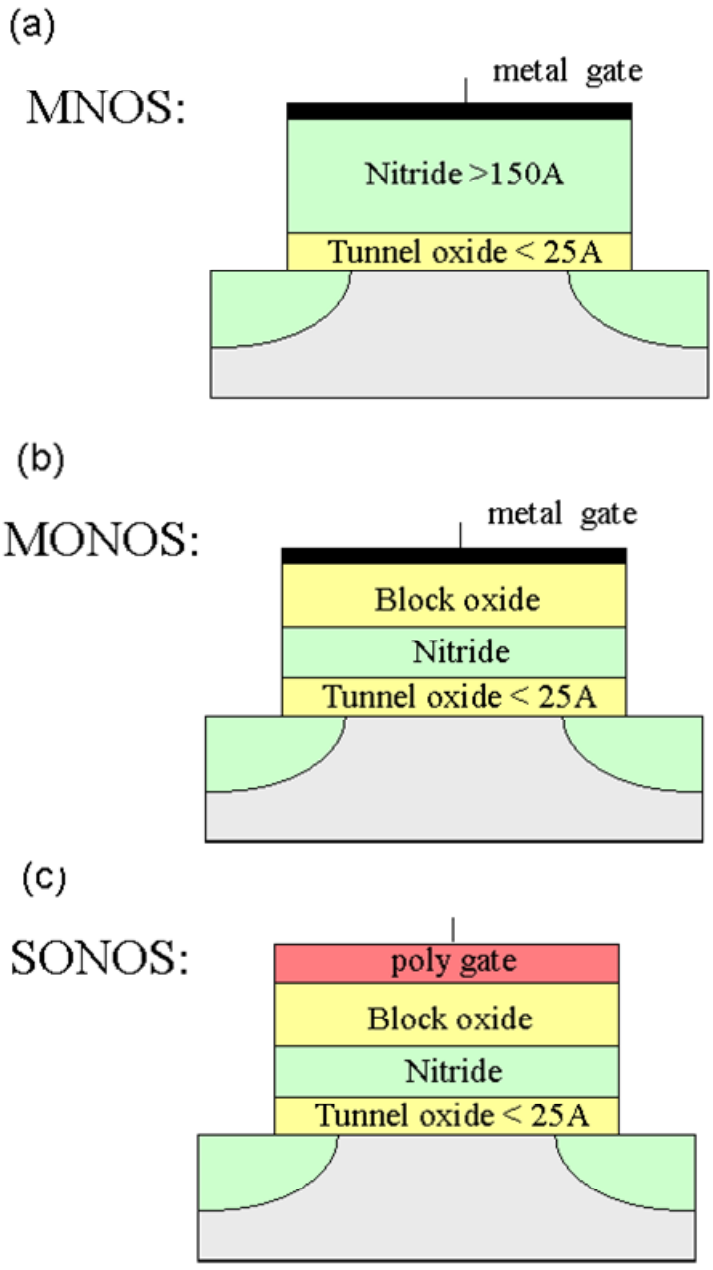


Fig. 1.1 (a) The MNOS device with a relative thick nitride(>150A) and high program/erase bias,
(b) MONOS device with metal gate and added blocking oxide, and
(c) SONOS device with poly-silicon gate and blocking oxide.

the gate storage dielectric and the thickness of the tunnel oxide. Comparing to floating gate flash memory, SONOS has a dielectric nitride on top of the tunnel oxide instead of a conductive poly silicon floating gate and with a blocking oxide between the top gate and nitride. Therefore, SONOS technology offers advantages of retention reliability because of pinhole immunity, low-voltage low-power operation and a simple fabrication process compatible with standard CMOS technology.

For a thin gate oxide SONOS cell in the direct tunneling regime, two leakage current components[7], thermionic and direct tunneling, in relating to the data loss, are the dominant mechanism. So far, most of the papers deal with data loss through the tunnel oxide. However, after cycling, another third component- the trap-to-trap tunneling current will be more serious and can be separated. In this thesis, data retention for various top (blocking) and bottom oxide (tunnel oxide) SONOS cells has been investigated. The leakage current separation has been presented to distinguish the three leakage components via thermionic, direct tunneling and trap-to-trap tunneling. The results can be useful toward an understanding of the scaling of SONOS cell with focus on its reliabilities.

On the other hand, several basic characteristics of pre-cycle and post-cycle have been compared. The consequences show that the short term and the long term endurance properties were governed by different mechanisms.

1.2 Organization of The Thesis

The organization of this thesis is separated into five chapters. After a brief introduction in chapter 1, we will introduce the devices and experimental setup in chapter 2, which presents the operating schemes used to program/erase the cells. In chapter 3, we will first discuss the basic

theory of SONOS cell, including trapping mechanism, tunneling, and emission models. Next, we will show the experimental results of the device retention characteristic with various blocking oxide thicknesses and after P/E cycling to separate the different charge loss mechanism. The endurance measurement and cycling effect will be studied in chapter 4. And finally, in chapter 5, the summary and conclusion will be given.

Chapter 2
Device Fabrication and Equipment Setup

2.1 Introduction

This chapter is separated into three parts. We will first describe the SONOS type flash memory devices and the split conditions we used. Second, the equipment setup and the experimental techniques to accurately control these instruments are explained. At last, we will discuss the program and erase schemes of these cells.

2.2 Device Preparation

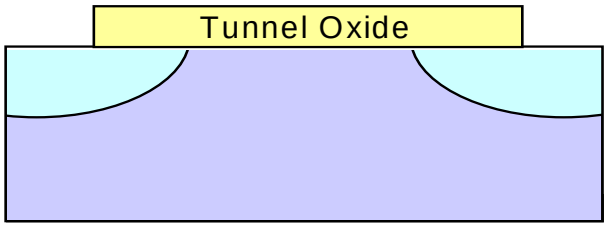
The cells used in this study are shown in Table 2.1, with different ONO thickness and W/L= 1.4/0.6 (um) and n-poly gate. Fig. 2.1 shows the major process flow for the fabrication of the ONO layer of the SONOS memory. As shown in Figs. 2.1(a), 1(b), and 1(c), the tunnel oxide is first grown by thermal oxidation with N_2 diluted (1%) at 850^{o}C, and has two different thickness of 20 Å and 25 Å. Next, a layer of LPCVD nitride film is grown with SiH_2Cl_2/NH_3=1:10 at 790^{o}C. Controlling the ratio of SiH_2Cl_2/NH_3 can determine the density of traps in nitride film. Calculated nitride thickness is 54 Å. Finally, the LPCVD block oxide is grown with SiH_2Cl_2/NH_3= 1:10~ 11:20 at 725~780^{o}C and has three different thickness, 34 Å, 46 Å, and 65 Å. Blocking oxide is used to block charge injection from the gate. We can use these cells to investigate the effect of blocking oxide thickness in the study of charge loss behavior. For a capacitance measurement, we use a larger size cell W/L= 240/240(um) to calculate the effective thickness of SONOS cells.

2.3 Equipment Setup

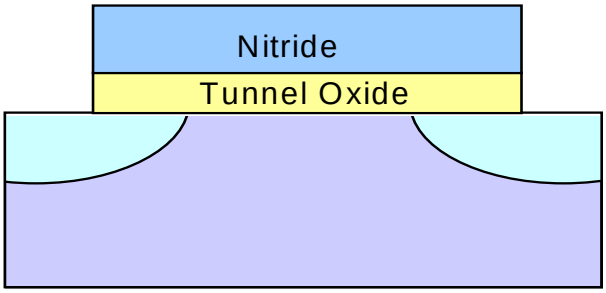
Split table

Wafer no. / ONO	1	2	3	4	5	6
$T_{tunnel\ oxide}$	25A			20A		
$T_{nitride}$	54A					
$T_{block\ oxide}$	34A	46A	65A	34A	46A	65A
W/L	1.4um / 0.6um					

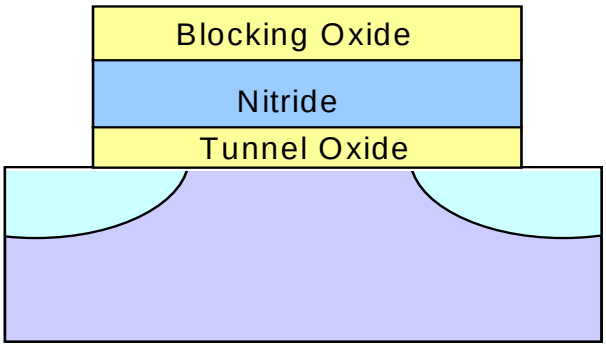
Table 2.1 The split conditions of samples used in this work, in which devices have three different blocking oxide thicknesses and two different tunnel oxide thicknesses.



Thermal oxidation with N_2 diluted (1%)at 850°C



LPCVD nitride with SiH_2Cl_2/NH_3 = 1:10 at 790°C



LPCVD oxide with SiH_2Cl_2/NH_3=1:10~11:20 at 725~780°C

Fig. 2.1 The process steps of SONOS memory to grow thermal tunnel oxide, LPCVD nitride, and LPCVD blocking oxide.

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 for analyzing the intrinsic and degradation behaviors 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 performing the SONOS cell program/erase operation.

Source-monitor units (SMU) with high current resolution to 10^{-15}A range facilitate the gate current measurement, sub-threshold characteristics extraction, and the saturation drain current measurement. The dual channels HP 8110A with high timing resolution provides two different pulse levels simultaneously for transient and P/E cycling endurance characterization. The HP E5250A configured 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 HP Visual Engineering Environment (VEE) and HT-Basic are used as the program languages to achieve the personal computer (PC) control of these measurement instruments.

In order to precisely control the pulse timing of HP 8110A, a special measurement technique, triggered pattern mode measurement, is used. Figs. 2.3(a) and (b) show the timing diagrams of the triggered pattern mode method during program and erase operations, respectively. By taking the programming time diagram as an example, the triggered pattern mode method can be explained as follows. In Fig. 2.3(a), the voltage source of Vs1 from HP 4156C (illustrated in Fig. 2.2) generates a signal, which is ramped to a state higher than the threshold trigger voltage of HP 8110A. The patterned pulses (defined as 01000 in Fig. 2.3(a) with the minimum period of 10 ns)

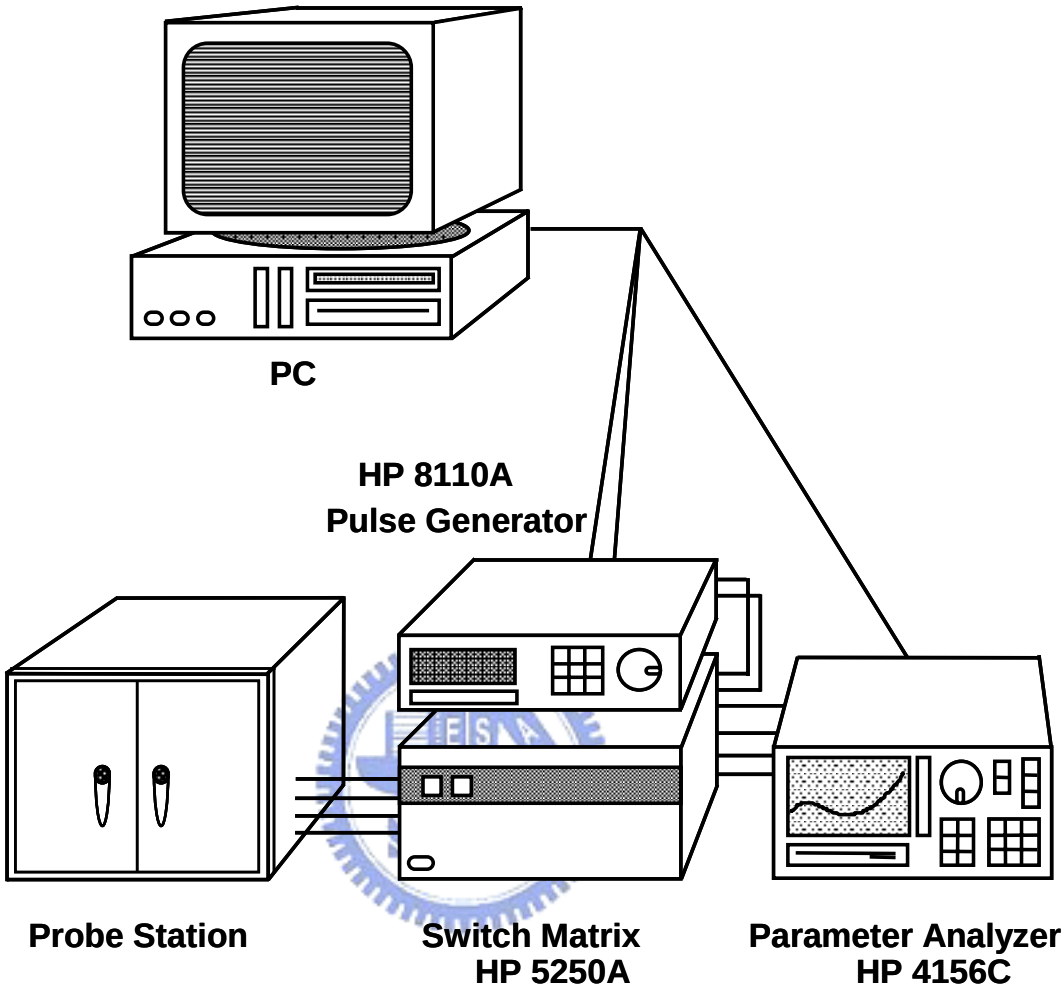


Fig. 2.2 The experimental setup of the current-voltage and the transient characteristics measurement in flash cells. An automatic controlled characterization system is setup based on the PC controlled instrument environment.

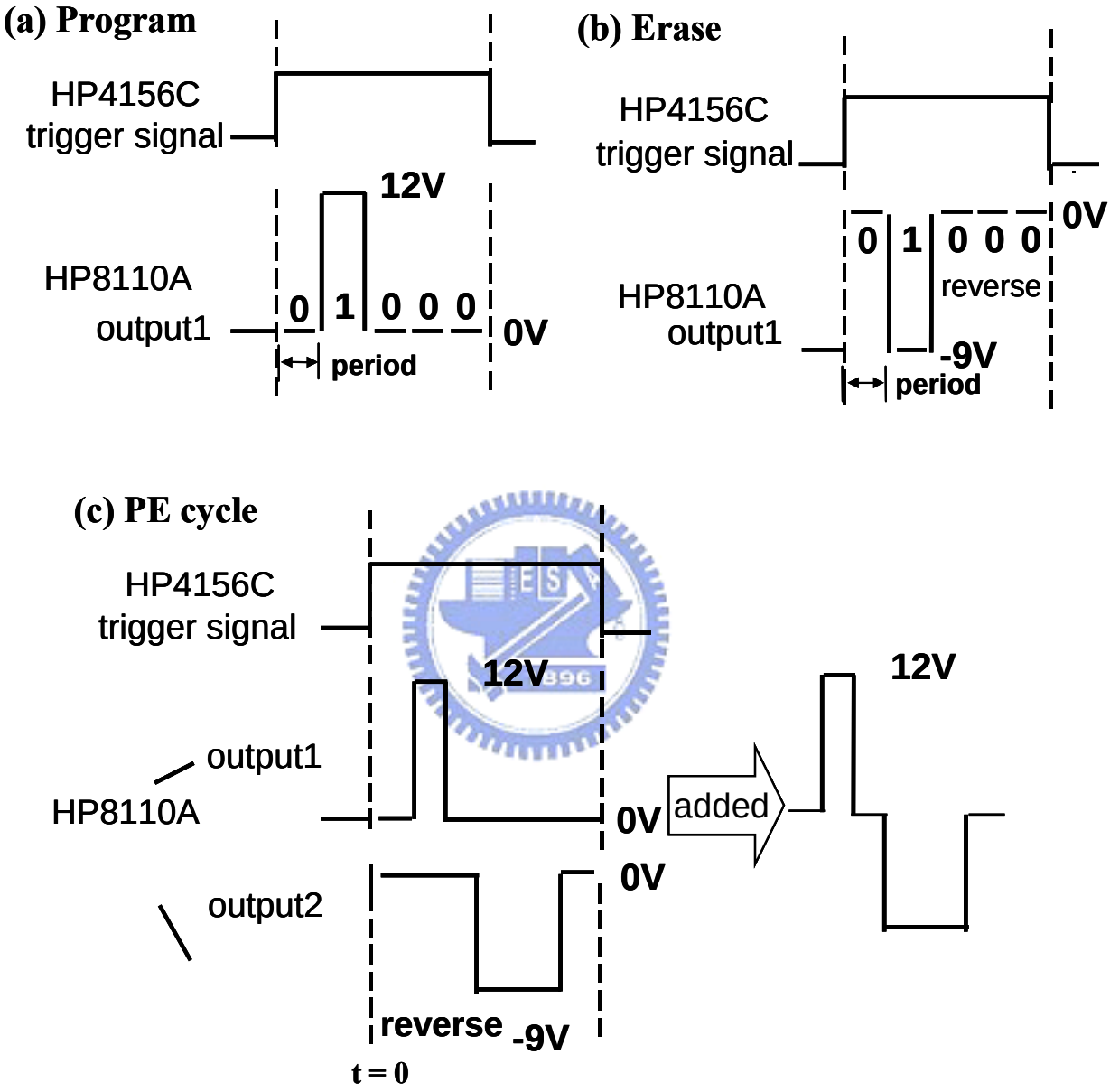


Fig. 2.3 The timing diagrams of the triggered pattern mode method during (a) program, (b) erase, and (c) PE cycling operations, respectively.

from HP 8110A are then triggered and the programming of a SONOS cell is performed. In our experience with the HP 8110A, we must set the period of pulse larger than 50 ~ 100ns to obtain a good square pulse and the leading or trailing edge larger than 25 to 50 ns to prevent the over-shoot and under-shoot of output pulse.

In addition, the triggered pattern operation mode also provides a precisely and shortly negative single pulse from HP 8110A to program or erase a SONOS, in order to provide a negative single pulse from HP 8110A; a patterned pulse of 01000 with the period of 10 ns and reverse mode is applied. Owing to a reverse mode is applied, the “1” state is the low voltage of the pulse as -9V, and the “0” state is the high voltage as 0V during program. When the cell is during erase, there is also a negative bias at the control gate. The applying pulse input and the Vs1 signal are the same as that in Fig. 2.3(a). In P/E cycling, we use add-mode of HP8110A to combine the pulse of output-1 and output-2. Next, HP8110A is set to triggered burst mode to send numbers of pulses we want. Thus, we can send a numbers of positive and negative pulses when triggered by 4156C in Fig. 2.3(c). This can provide us a speed-up of the cycling method, which can reach 10000 P/E cycles in 8hours.

2.4 Program/Erase Schemes

The SONOS cell is programmed and erased by having gate connected to a pulse-generator while substrate, drain, and source are connected to ground in Figs. 2.4(a) and (b). The injection mechanism depends on electrical field of tunnel oxide as shown in Figs. 2.5(a), (b), and (c). When electrical field of tunnel oxide is very strong ($E_{ox} > \phi_1 / t_{ox}$), charge tunnels through triangular region of the oxide, i.e., by FN tunneling shown in Fig. 2.6. At lower field, ($(\phi_1 - \phi_2)/t_{ox} < E_{ox} < \phi_1 / t_{ox}$), charge directly tunnels through oxide. When $E_{ox} < (\phi_1 - \phi_2)/t_{ox}$, charge tunnels through triangular region of oxide and nitride by a modified FN tunneling[8].

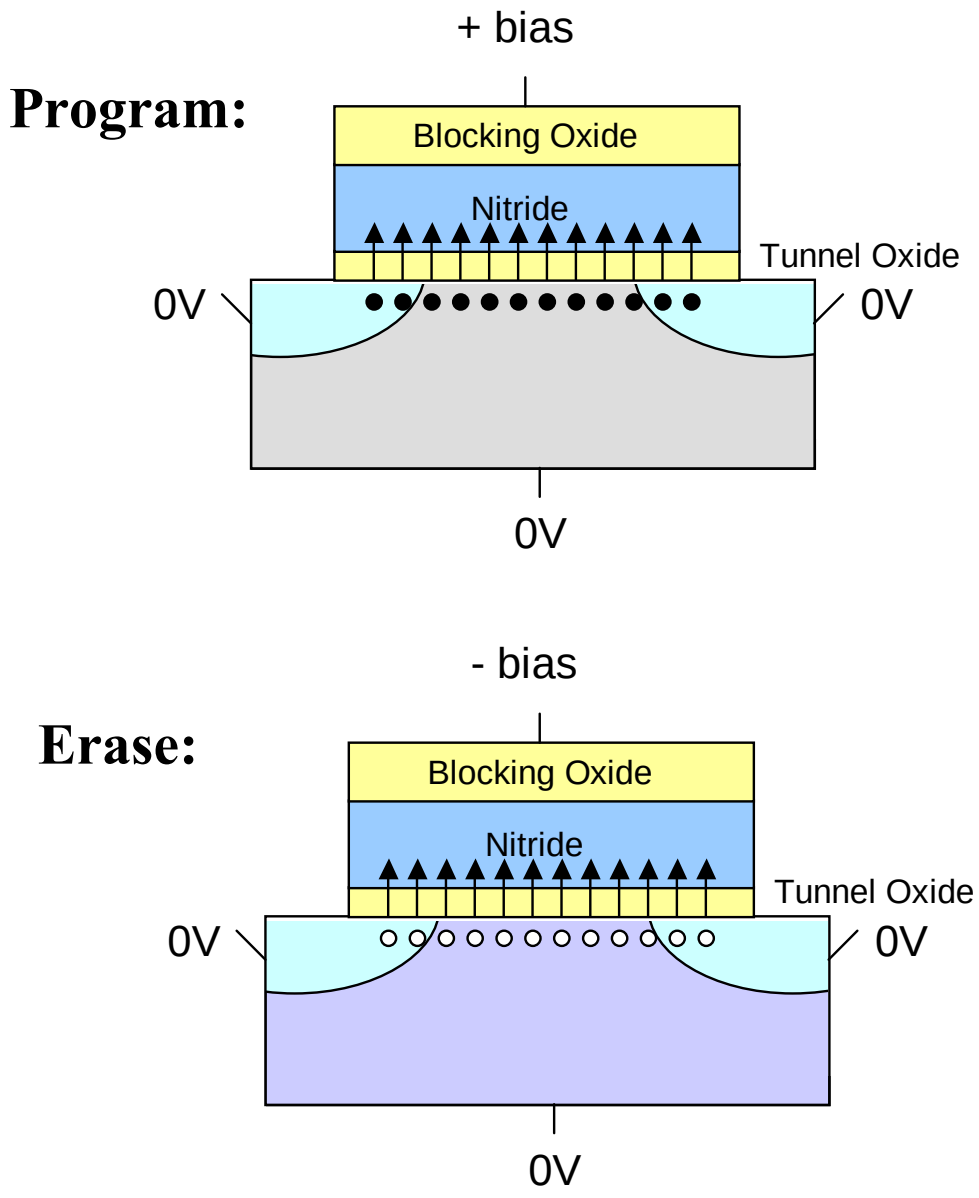


Fig. 2.4 The scheme we used to program/erase SONOS, with the drain, source, and substrate connected to ground.

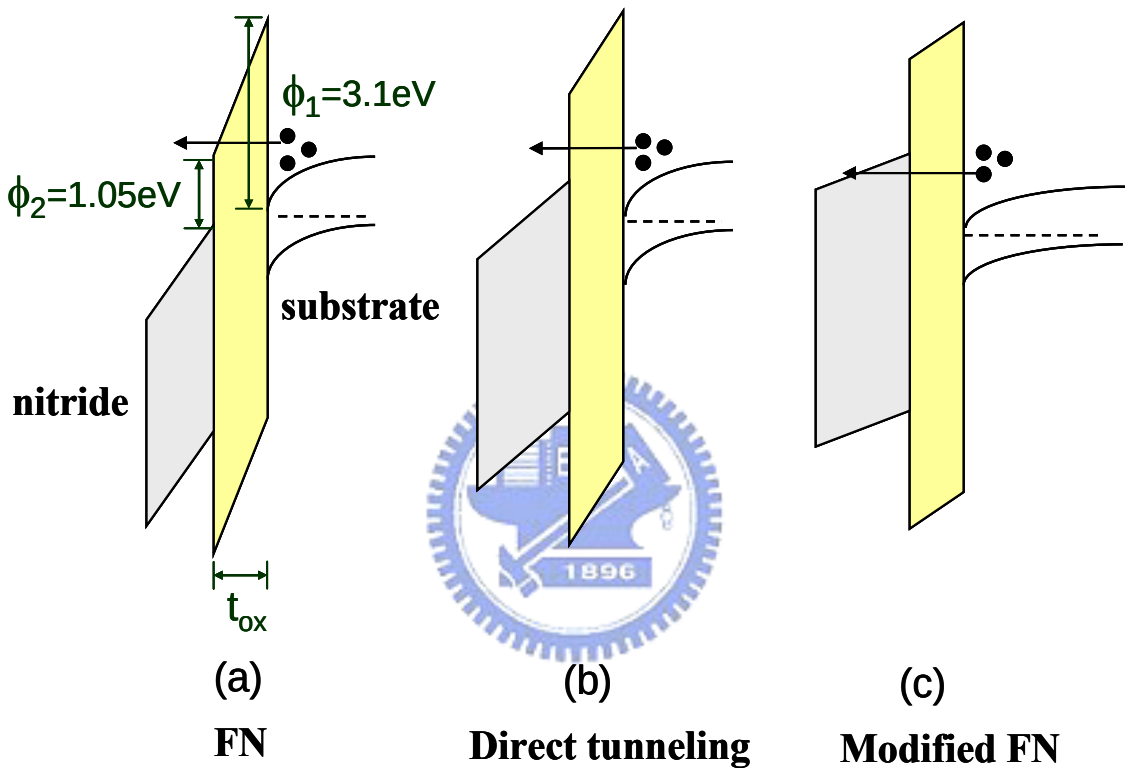


Fig. 2.5 The schematic diagram of FN tunneling, direct tunneling, and modified FN tunneling.

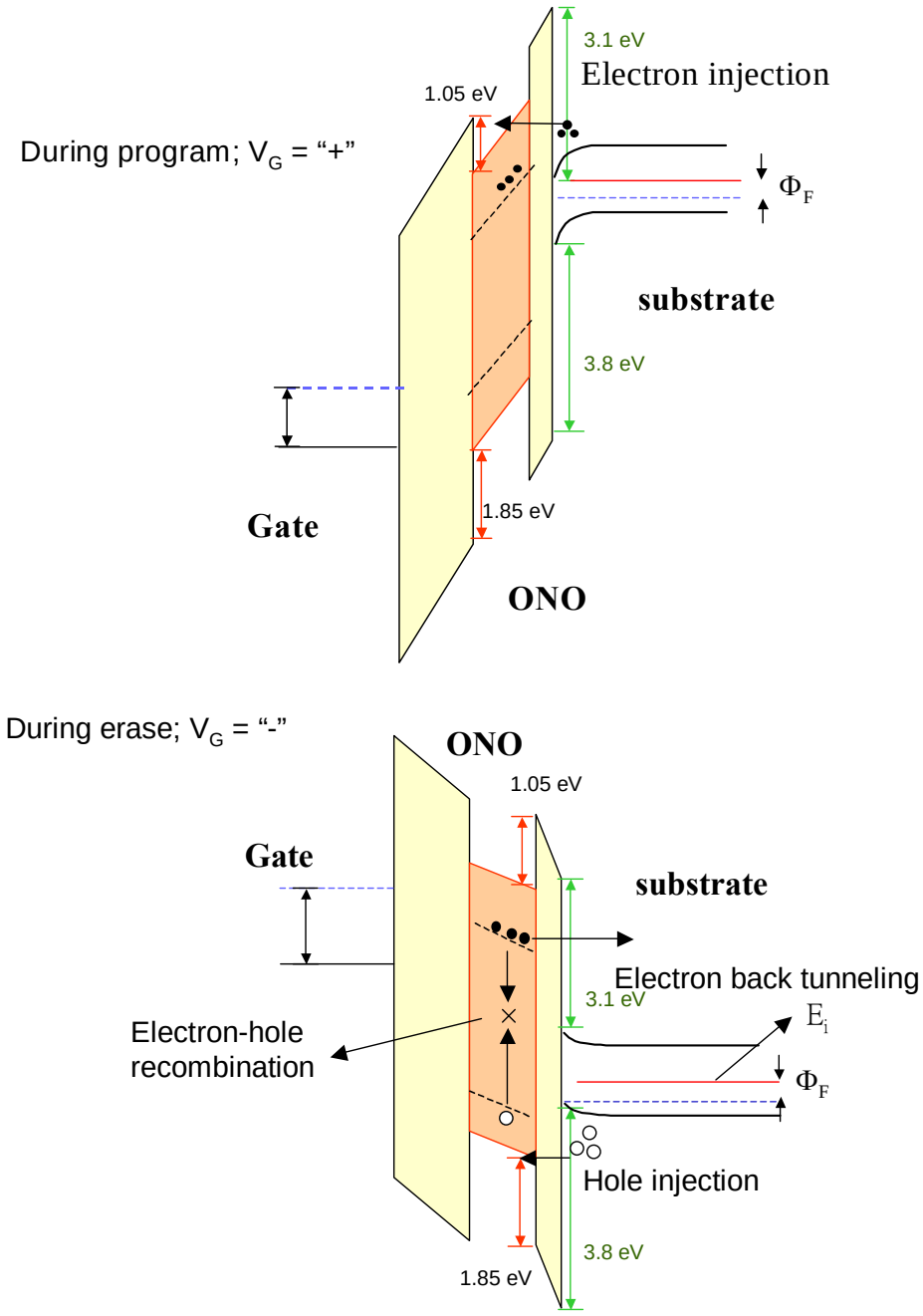


Fig. 2.6 The schematic diagram of FN program and FN erase for SONOS memory.

Generally, erasing the cell (negative bias) consumes much more time due to higher barrier height for holes. During programming, electrons tunnel from the band to the nitride, holes are blocked by blocking oxide due to a high barrier height between the poly-Si gate and the oxide. During erase, some electrons close to SiO_2/SiN interface back tunnels from the trap to the band initially [9], and then hole tunnels from the band to the nitride. Hole injection decreases in an exponential rate with time and electrical field of blocking oxide becomes stronger and stronger. Enhanced field causes electrons to tunnel from the gate to the nitride and finally reaches steady state.

Chapter 3
The Principle of Trapping and Retention Model

3.1 Introduction

Retention is a measure of the time that a memory cell retains its information whether it is powered or standby. The nonvolatile function of SONOS type memory is achieved by the nitride layer with rich independent traps. The charge should ideally be stored for more than 10 years under normal chip operating conditions.

In this chapter, we will first discuss the charge loss mechanism, including direct tunneling; trap assisted tunneling, and thermionic emission as well as the derivation of equations relating to these mechanisms. Finally, we will explain how to separate the three leakage components.

3.2 Trapping Mechanism of SONOS

The nonvolatile function of SONOS device is based on the storage of charges (includes electron and hole) in the discrete traps (that is, dangling Si bonds) in the silicon nitride layer of the multi-layer of oxide-nitride-oxide (ONO) dielectric stacked gate structure. In 1981, Robertson[10] found that the $\equiv$Si-H unit in nitride contributes to the silicon dangling bonds which have positive, neutral, and negative charge state with 0, 1 and 2 electrons, shown in Eqs. (3.1), (3.2), and (3.3,) as the following [11]:

$$\equiv \text{Si-H} \rightarrow \equiv \text{Si}: + \text{H}^+, \tag{3.1}$$

$$\equiv \text{Si}: + \text{h}^+ \rightarrow \equiv \text{Si}\bullet, \quad \equiv \text{Si}\bullet + \text{h}^+ \rightarrow \equiv \text{Si}, \tag{3.2}$$

$$\equiv \text{Si} + \text{e}^- \rightarrow \equiv \text{Si}\bullet, \quad \equiv \text{Si}\bullet + \text{e}^- \rightarrow \equiv \text{Si}:. \tag{3.3}$$

3.3 The Retention Model

Figure 3.1 shows the retention model of a SONOS cell. The potential of the nitride is raised when electrons are stored in the nitride. Thus, electrons attempt to escape from the nitride. Electron may loss via many approaches, such as direct tunneling of electrons, thermionic emission of electrons, trap-to-trap tunneling of electrons, and band-to-trap tunneling of hole. In a fresh cell with zero bias, we assume that charge loss is dominated by the direct tunneling (components (2) and (3)) and thermionic emission (component (1)), respectively, each represents the short term and long term charge loss respectively. After several P/E cycles, the charge loss via the trap-to-trap tunneling (components (4) and (5)) will become more and more important.

3.3.1 The Direct Tunneling Model

Let us start with the calculation of the direct tunneling component. According to **WKB equation**, the probability **P** of an electron tunneling through a one-dimensional barrier of height Φ_0 is:

$$\mathrm{P} = e^{-2\int(\sqrt{2m(q\Phi_0 - qxE_{ox})}/\hbar)\,dx}, \tag{3.4}$$

from which we can derive the following expression from Eq. (3.4). At low electrical field, $(q\phi_0 - qxE_{ox}) < 0$, electrons tunnel through dielectric by direct tunneling, and the tunneling probability P can be expressed as [12]:

$$\begin{aligned} \mathrm{P}_{DT} &= \mathrm{e}^{-\beta\left[(q\Phi_0)^{\frac{3}{2}} - (q\Phi_0 - qE_{ox}t_{ox})^{\frac{3}{2}}\right]/E_{ox}} \\ &\approx \mathrm{e}^{2\cdot\sqrt{2m(q\Phi_0 - qxE_{ox})}\cdot t_{ox}/\hbar}. \end{aligned} \tag{3.5}$$

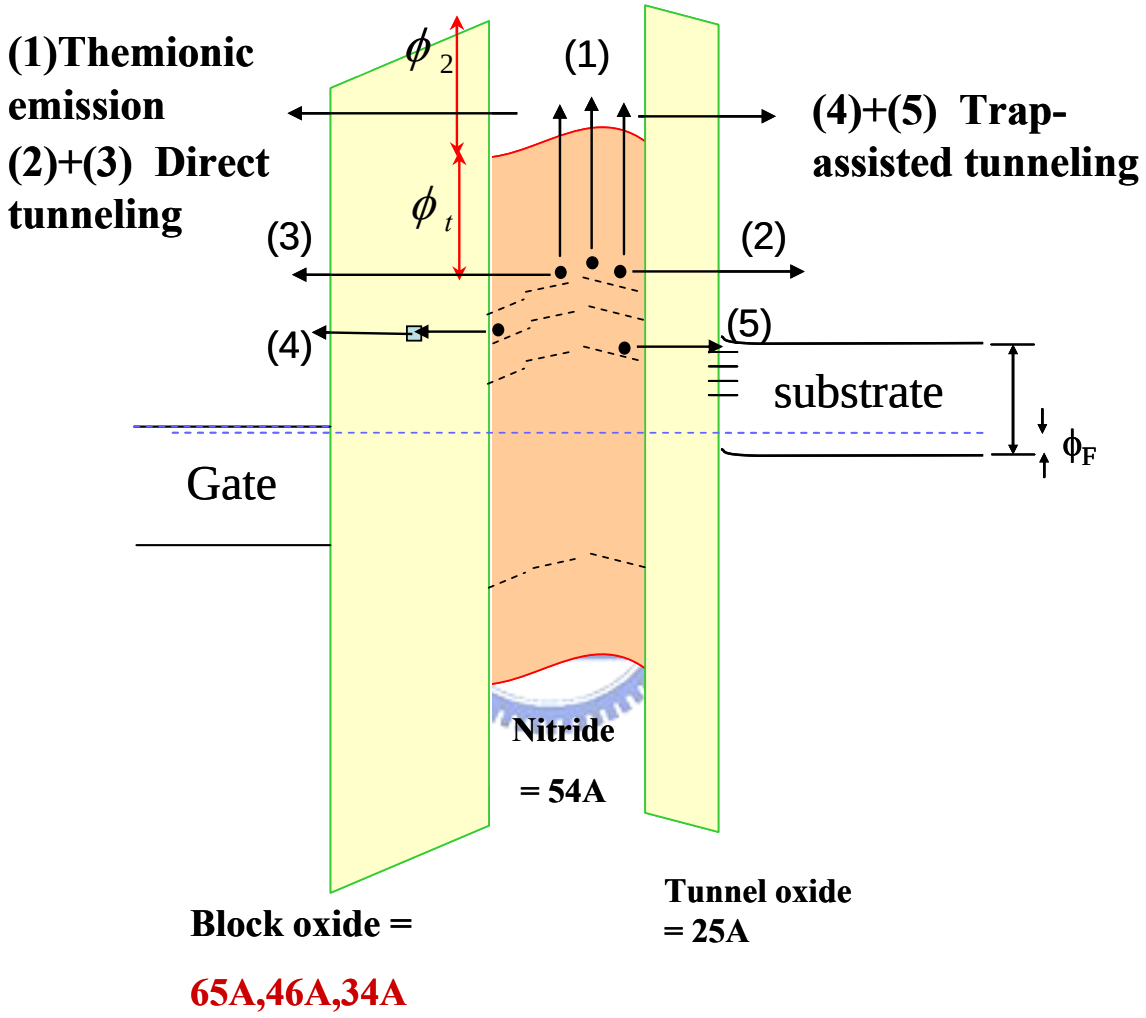


Fig. 3.1 The charge loss path and the two dominant leakage components in a SONOS cell, Component (1): thermionic emission , Components, (2) and (3) : direct tunneling, Components, (4) and (5) : trap-assisted tunneling.

At higher electrical field, $(q\phi_0 - qxE_{ox}) > 0$, electrons tunnel through the dielectric by FN tunneling, and the tunneling probability P can be expressed as:

$$\mathrm{P}_{FN} = \mathrm{e}^{-\beta(q\Phi_0)^{\frac{3}{2}}/E_{ox}}, \tag{3.6}$$

where $\beta = \frac{4\sqrt{2m}}{3\hbar q}$.

For a retention model, there is no bias added to the device, the electric field is small. So, charge loss is contributed by the direct tunneling. For electrons trapped in the nitride with trap depth ϕ_t, the time constant (τ) of tunneling through the nitride and oxide to the silicon substrate can be expressed as an inverse product of $\mathbf{P}_{\mathrm{ox}}$ and $\mathbf{P}_{\mathrm{N}}$ [13] as the following:

$$\frac{1}{\tau} = \frac{1}{\tau_0} \cdot P_{ox} \cdot P_N = \frac{1}{\tau_0} \cdot \mathrm{e}^{-2\alpha_{\mathrm{ox}} \cdot t_{ox}} \cdot \mathrm{e}^{-2\alpha_{\mathrm{N}} \cdot x} = \frac{1}{\tau'} \cdot \mathrm{e}^{-2\alpha_{\mathrm{N}} \cdot x}, \tag{3.7}$$

in which $\mathbf{P}_{\mathrm{ox}}$ and $\mathbf{P}_{\mathrm{N}}$ are the probability of tunneling through the oxide and nitride respectively and can be expressed by:

$$P_{ox} \approx \mathrm{e}^{-2\sqrt{2mq(\phi_2+\phi_{\mathrm{t}})} \cdot t_{ox}} = \mathrm{e}^{-2\alpha_{\mathrm{ox}} \cdot t_{ox}} \tag{3.8}$$

and

$$P_N \approx \mathrm{e}^{-2\sqrt{2mq\phi_{\mathrm{t}}} \cdot x} = \mathrm{e}^{-2\alpha_{\mathrm{N}} \cdot x}, \tag{3.9}$$

also, $\tau' = \tau_0 \cdot \mathrm{e}^{2\alpha_{\mathrm{ox}} \cdot t_{\mathrm{ox}}}$. With time constant τ, we can derive the trapped electron density n_t as given by:

$$n_t(\phi_t, t) = n_t(\phi_t, 0) \cdot e^{-\frac{t}{\tau}}. \tag{3.10}$$

Furthermore, ΔV_{th} can be expressed as an integral of n_t given by:

$$\Delta \mathrm{V}_{\mathrm{th}}(\mathrm{t}) = \int_0^{X_N} -qn_t(x)e^{-\frac{t}{\tau(x)}} \cdot \left(\frac{x_N - x}{\varepsilon_N} - \frac{x_{OB}}{\varepsilon_{ox}}\right)dx. \tag{3.11}$$

Differentiating Eq. (3.11), we have Eq. (3.12a) and Eq. (3.12b), which are proportional to $(1 - e^{-\frac{t}{\tau'}})$ [14], i.e.,

$$\frac{d\Delta V_{th}(t)}{d\log(t)} = -2.3 \cdot \int_0^{X_N} qn_t(x) \cdot \frac{t}{\tau} \cdot e^{-\frac{t}{\tau(x)}} \cdot \left(\frac{x_N - x}{\varepsilon_N} - \frac{x_{OB}}{\varepsilon_{ox}}\right) dx \tag{3.12a}$$

$$\approx -2.3 \cdot q \cdot n_t(x') \cdot \left(\frac{x_N - x'}{\varepsilon_N} - \frac{x_{OB}}{\varepsilon_{ox}}\right) \cdot \left(\frac{1 - e^{-\frac{t}{\tau'}}}{2\alpha_N}\right). \tag{3.12b}$$

We use samples no.1 and no. 3 with different blocking oxide to observe the charge loss behavior. First, we assume that charge is located at a region closed to the tunnel-oxide/nitride interface (proved later). From Fig. 3.2, we found that for the sample no.1 (34 Å block oxide), charge loss is proportional to $a + b(1 - e^{-\frac{t}{\tau_1}}) + (1 - e^{-\frac{t}{\tau_2}})$ within 1000 seconds, while the loss of sample no. 3 (65 Å block oxide) is proportional to $a + b(1 - e^{-\frac{t}{\tau}})$. Since the blocking oxide is deposited by LPCVD, the oxide quality is not as good as the tunnel oxide [15]. We conclude that for the thinner blocking oxide, charge may tunnel through both blocking oxide and tunnel oxide and thus have two time constants.

3.3.2 The Trap Assisted Tunneling Model

After several P/E cycles, the oxide layers will be damaged during programming and erasing, and generate interface states and oxide charges (traps) which degrade the retention ability. Through these damages, there are several leakage mechanisms. Fig. 3.3(a) shows the trap-to-trap tunneling current. The time constant of the electron decay due to trap-to-trap tunneling has been derived by Roy [16], and is given by

$$\tau_{T-T}(x, \phi_T) = \tau_{T-T0} \exp\left(\alpha_{ox}^e X_{ot}\right) \exp\left(\alpha_n^e x\right), \tag{3.13}$$

where X_{ot} is the oxide thickness, x is the tunnel distance inside nitride and

$$\tau_{T-T0}(\phi_T) = \frac{8\pi m_{ox,e}^* \left(X_{ot} + \alpha_{ox}^{e-1}\right)}{\alpha_{si}^e h^3 \overline{D_{it}}} \tag{3.14}$$

and

$$\alpha_{Si}^e(\phi_T) = \frac{4\pi}{h} \sqrt{2m_{Si,e}^* q\left(\phi_T - \phi_1^e + \phi_2^e\right)}, \tag{3.15}$$

$\overline{D_{it}}$ is the Si-SiO_2 interface tarp density, $m_{Si,e}^*$ is the electron effective mass in the silicon.

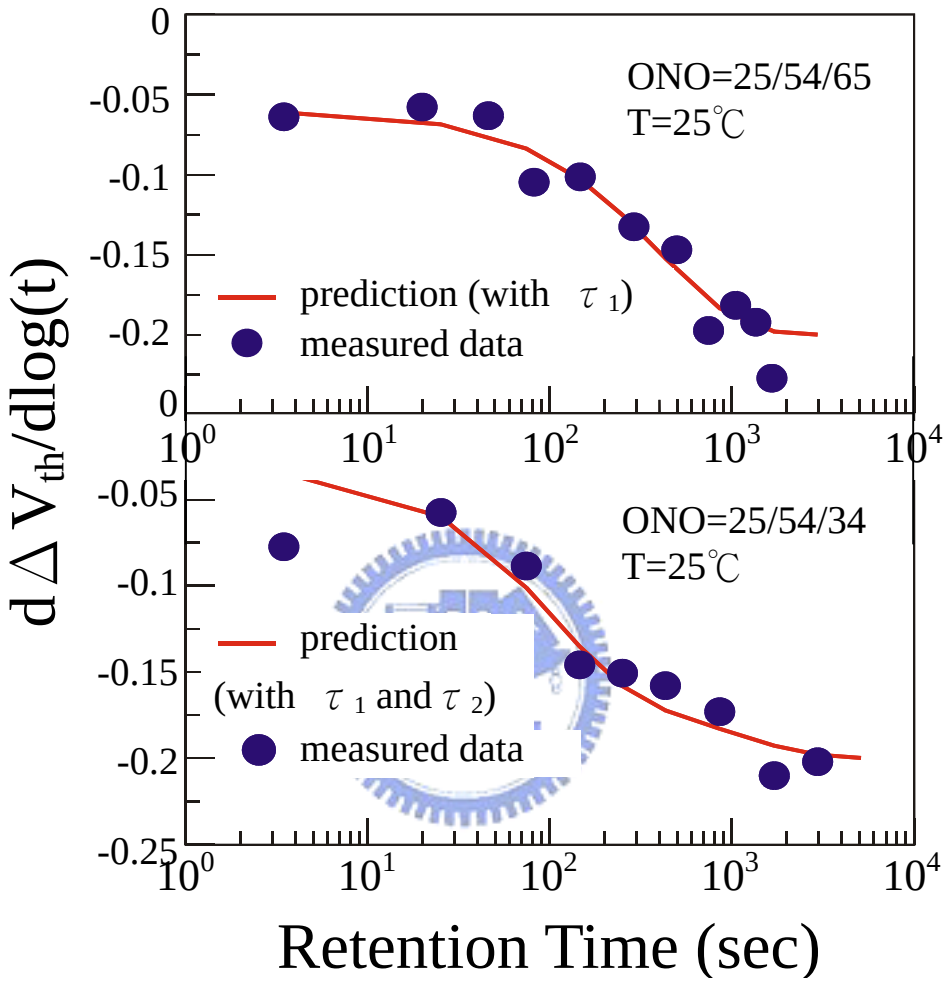


Fig. 3.2 The direct tunneling through ONO exhibits different time constants. Top figure has one time constant, while bottom figure, for a thin blocking oxide, shows two time constants.

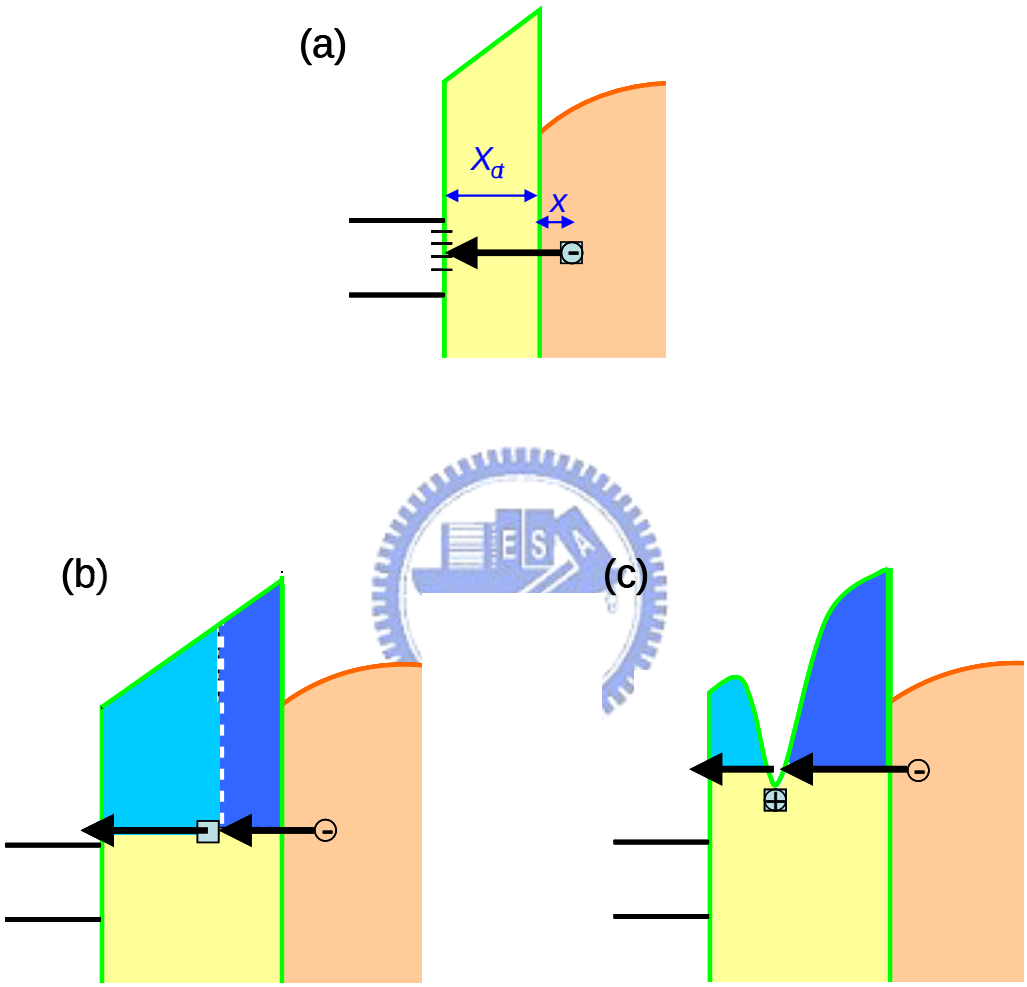


Fig. 3.3 The cycling-induced leakage components in a SONOS cell,
(a) Trap-to-trap tunneling
(b) Trap-assisted tunneling
(c) Positive charge-assisted tunneling

Fig. 3.3(b) is the sequential electron tunneling via neutral electron traps [17-21]. While the trap density generated within the oxide increases with P/E cycle numbers, there is a higher probability for electron tunneling into trap sites near the cathode. For thin oxide, there is also a high probability of direct tunneling out of trap sites into the anode and steady current flows when there is equilibrium between trap filling and emptying processes. As the oxide thickness increases, because a large part of the traps keep on filled for a long time, the initial charge loss during retention is higher than the steady state value.

In Fig. 3.3(c), the electrons tunnel out of nitride serially via trapped positive charges [22]. The trapped positive charges is caused by the hole generation due to electron impact ionization during programming and erasing. These generated positive charges in the oxide will diminish the tunneling barrier and enhance the tunneling probability.

3.3.3 The Thermionic Emission Model

With increasing temperatures, thermionic emission effect is more important. At high temperature, electrons emitted from the trap with energy level ϕ_t drift to the oxide/nitride interface, and then tunnel through the oxide. Because the barrier height between the nitride and the oxide is 1.05eV for electrons, emission rate is determined by the trap energy level ϕ_t in nitride and determined by the temperature. The time constants can be derived as the following:

$$r_1 = \sigma_n \cdot v_{th} \cdot n \cdot N_t \cdot (1 - f_T) \tag{3.16}$$

and

$$r_2 = e_n \cdot N_t \cdot f_T, \tag{3.17}$$

where r_1 is the capture rate of electron and r_2 is the thermal emission rate of electron from trap. At thermal equilibrium, $r_1 = r_2$, and from Eq. (3.16) and Eq. (3.17), the emission rate and time constant can be derived as:

$$e_n = \frac{1}{\tau} = v_{th} \cdot \sigma_n \cdot n \cdot \frac{1-f_T}{f_T} = v_{th} \cdot \sigma_n \cdot N_C \cdot \exp(\frac{E_T - E_C}{KT}) \propto e^{-\frac{\phi t}{KT}} \quad (3.18)$$

and

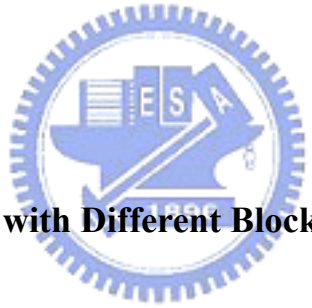
$$\tau \propto e^{\frac{\phi t}{KT}}. \quad (3.19)$$

respectively, which are strongly dependent on the trap energy level and temperature. For the retention model, we assume that the emission rate is the same as in equilibrium condition. Time constant can be expressed as a product of a constant τ_o and exponential item, i.e.,

$$\tau = \tau_0 \mathrm{e}^{q\Phi t/KT}, \quad (3.20)$$

and the trap density is exponentially proportional to $-\frac{t}{\tau}$, given by

$$n_t(\phi_t, t) = n_t(\phi, 0) \cdot e^{-\frac{t}{\tau}} = n_t(\phi, 0) \cdot \exp(-\frac{1}{\tau_0} \mathrm{e}^{-q\Phi t/KT} \cdot t). \quad (3.21)$$



3.4 The Results

3.4.1 Data Retention for Cell with Different Blocking Oxides

In the past, most reports of MONOS/SONOS retention studies assume that the leakage is through the tunnel oxide due to a thicker blocking oxide (40~50 Å) comparing to the tunnel oxide (20~25 Å). Here, we use samples 1, 2, and 3 with tunnel oxide 25 Å, nitride 54 Å, blocking oxide 34 Å, 46 Å, and 65 Å as shown in Table 2.1 to study the charge loss mechanism. First, we bake these samples in the oven for a period of time and then measure the threshold voltage. And, the procedure is repeated again. Finally, we have a ΔV_{th} *versus time* plot of three cells with different blocking oxide at a specified temperature. Next, we convert it into ΔQ *versus time* plot according to

$$\Delta Q = \Delta V_{th} \cdot \frac{A \cdot \varepsilon_{ox}}{t_{eff} - t_{tunnel-oxide} - x_c \cdot \frac{\varepsilon_{ox}}{\varepsilon_N}}, \quad (3.22)$$

and use the previous result [23]that charge is located at the tunnel-oxide/nitride interface.

Figures 3.4 is the $\triangle V_{th}$ *versus time* plot for three cells no.1, 2, and 3 at room temperatures. The charge loss($\triangle Q$) increases very fast initially and then reaches a steady state. These three cells are fresh cells that are programmed and erased for only several cycles and then programmed under the same field for 10ms. Therefore, it is reasonable to assume that equal injected charges are maintained in all three cells. We can observe that with decreasing blocking oxide thickness, $\triangle Q$ increases[24]. We think it may because of an upward leakage through the blocking oxide, which can be seen from the curve of Fig. 3.2, according to Eq. (3.12).

3.4.2 The Separation of Charge Loss Components

In order to separate the charge loss components, we measure the threshold voltage shift of O/N/O=25/54/65 sample after 24 hours baking at various temperatures and draw the figure of *threshold voltage shift versus bake temperature* shown in Fig. 3.5.

According to Eq. (3.7), we see that the time constant τ for direct tunneling is not sensitive to the temperature, though parameter τ_0 may depend weakly on the temperature. When temperature is sufficiently low, the brown curve maintains at a constant value of 0.84V. Therefore, we can divide $\triangle V_{th}$ into two components, with direct tunneling component of 0.84V and thermionic emission component of ($\triangle V_{th}$-0.84 V), as shown in Fig. 3.6.

To differentiate how direct tunneling and thermionic emission components contribute to the charge loss in SONOS memory cell, we plot the $\triangle Q$ versus *block-oxide thickness* curves for retention time from 10s to 24hour, as shown in Fig. 3.7. Notice that the $\triangle V_{th}$ in thin block oxide cell increases faster than in thick block oxide cell initially but with the same speed eventually.

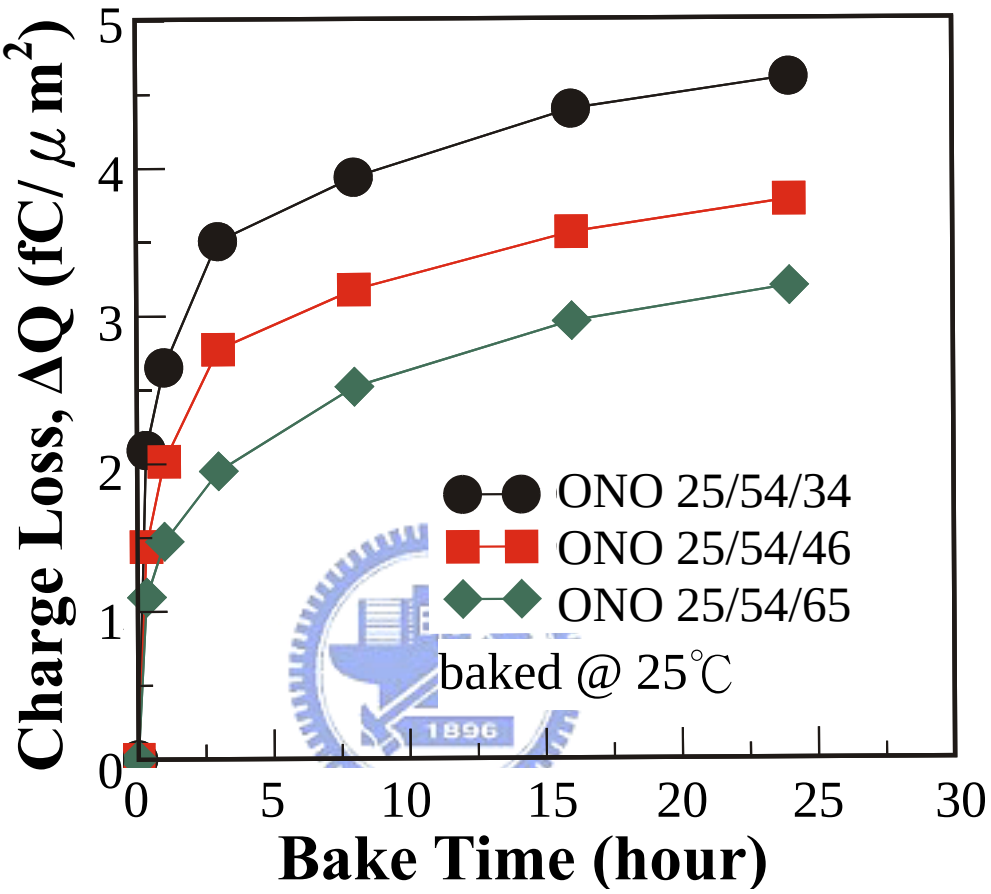


Fig. 3.4 The calculated charge loss versus bake time for different blocking oxide thickness. Thin blocking oxide sample exhibits larger charge loss.

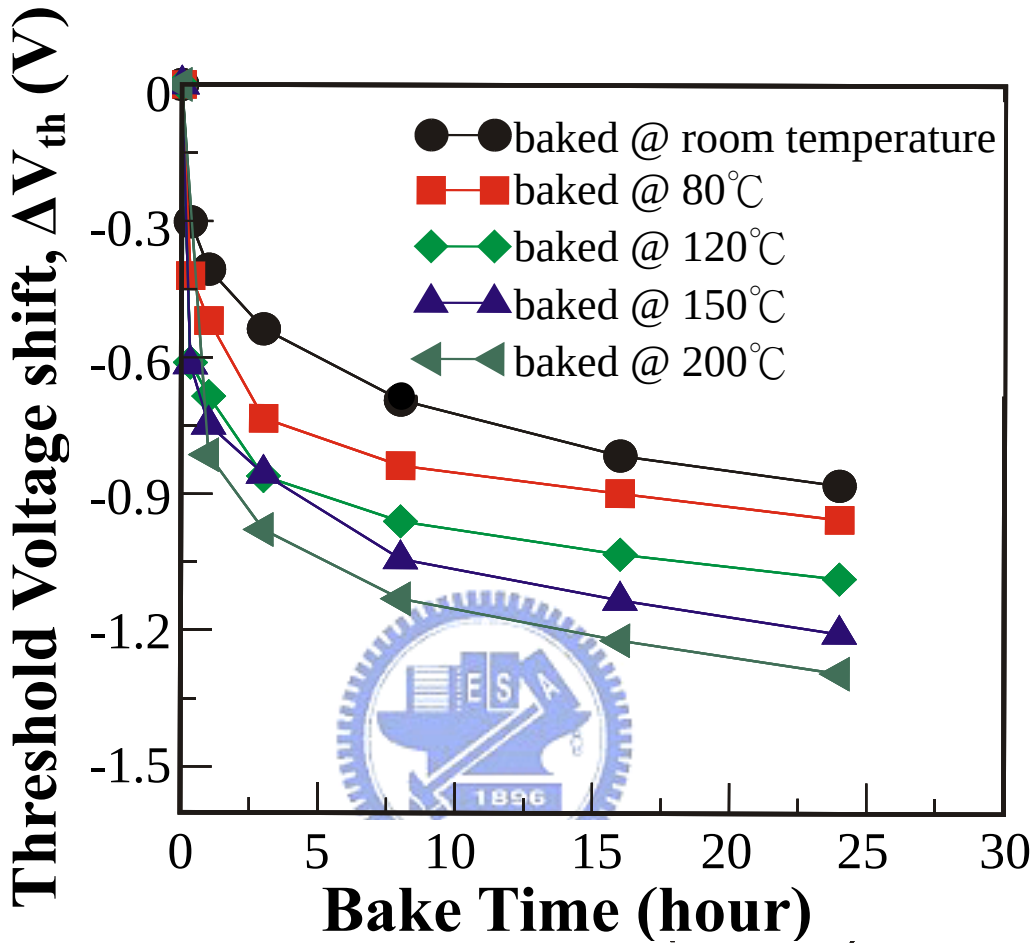


Fig. 3.5 The data retention of O/N/O = 25/54/65 sample at various temperatures.

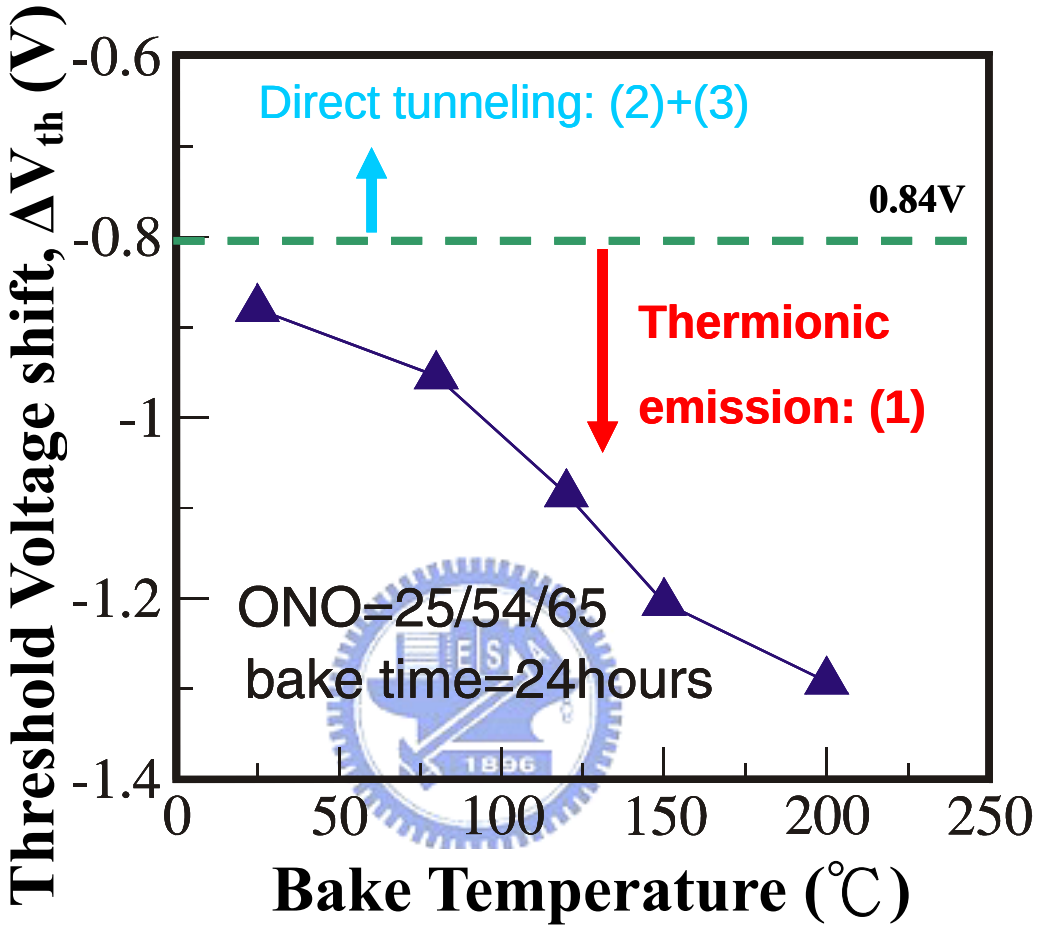


Fig. 3.6 The ΔV_{th} vs temperature plot, calculated from Fig. 4.7, showing that ΔVth has maximum slope between 80°C and 150°C, for sample 3- O/N/O = 25/54/65.

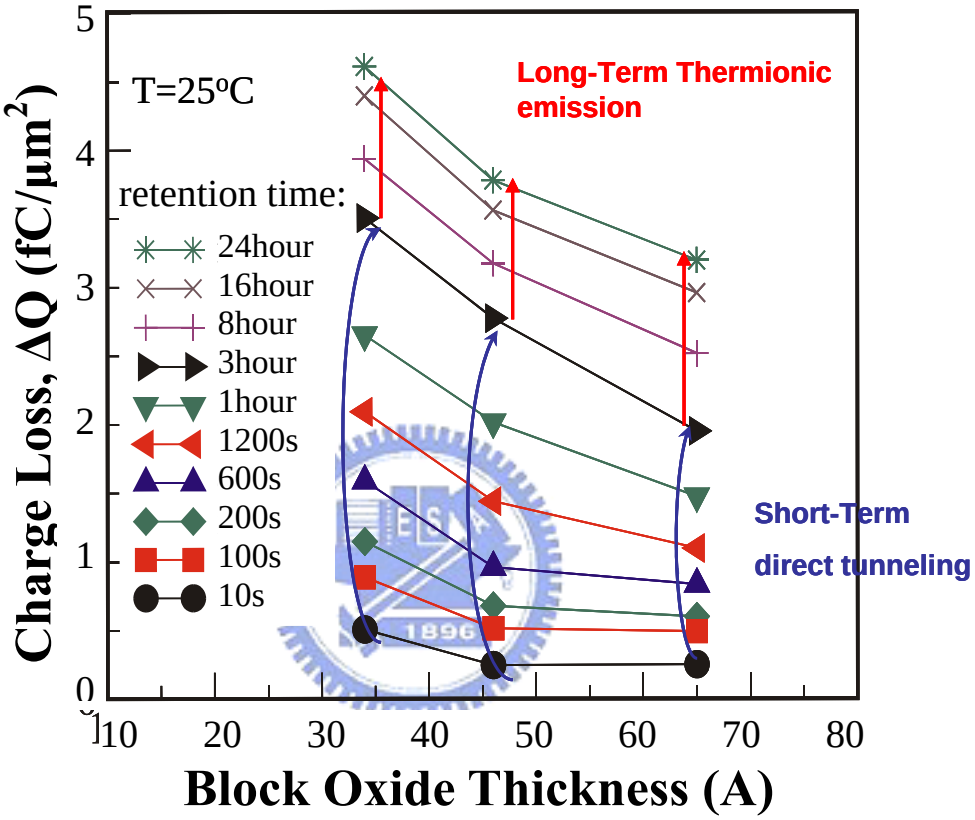


Fig. 3.7 Two dominant leakage mechanisms in a SONOS cell. Note that short term leakage is dominated by direct tunneling, while long term leakage is dominated by thermionic emission.

With the same tunnel oxide thickness, it can be concluded that more charges leak through the blocking oxide in thin block oxide cell initially. Fig. 3.8 shows the $\triangle Q$ versus *block-oxide thickness* curve for 24hours retention at various temperatures and shows no difference between each cell as temperature increases. Based on Fig. 3.7 and Fig. 3.8, we may conclude that after retention for 10000s, the leakage path is neither through blocking oxide nor through tunnel oxide for their shorter time constant. We think that it is because of thermionic emission leakage. The charge leaks via long-term thermionic emission and is not sensitive to oxide thickness. This result is the same as previously reported results that trap to band tunneling dominates the short-term loss and thermionic emission dominates the long-term loss.

3.4.3 The Trap-Assisted Tunneling Charge Loss Component

In this section, the cycling effect on data retention will be discussed. The measured I_{GIDL}-V_D curve for fresh and after 10^3 P/E cycles is shown in Fig. 3.9. Because the I_{GIDL} current increases after 10^3 P/E cycles due to the generation of Q_{ox}, cells after cycling have poorer data retention ability as shown in Fig. 3.10. It can be explained that more generated oxide damage will raise the probability of trap-assisted tunneling. We can further separate the trap-assisted tunneling after P/E cycles from this figure.

In Fig. 3.11, we know that the excess oxide damage will be generated after P/E cycles. It shows that thicker block oxide thickness can contain more oxide charge and cause poorer retention ability. The higher the blocking oxide thickness, the poorer its data decay rate (Fig. 11) is. Although the thicker blocking oxide thickness provides a better data retention in fresh cells, while after 10^3 P/E cycles, device degrades more seriously. We can also see the effect of trap-assisted tunneling. So, it is a trade-off for us to consider the data retention in both tunnel oxide and blocking oxide thicknesses for achieving a better performance of SONOS cells.

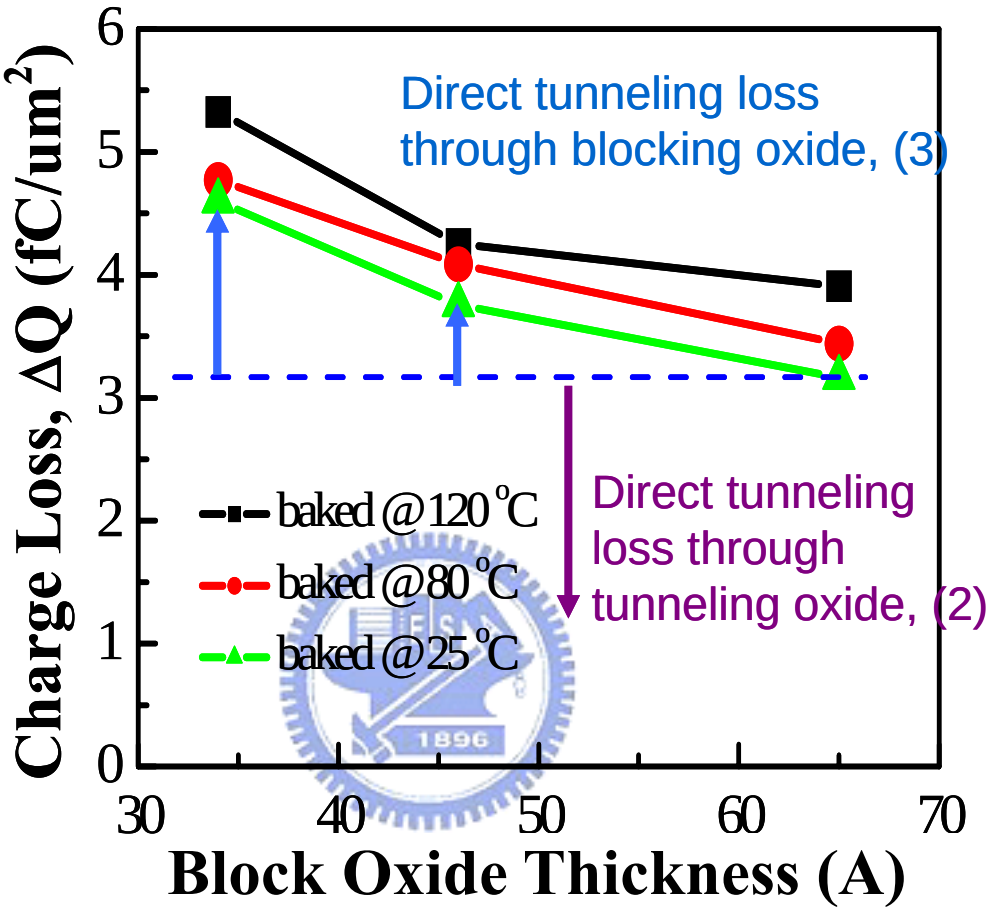


Fig. 3.8 Further calculations can be done by differentiating the charge loss through tunnel oxide (2) and blocking oxide (3).

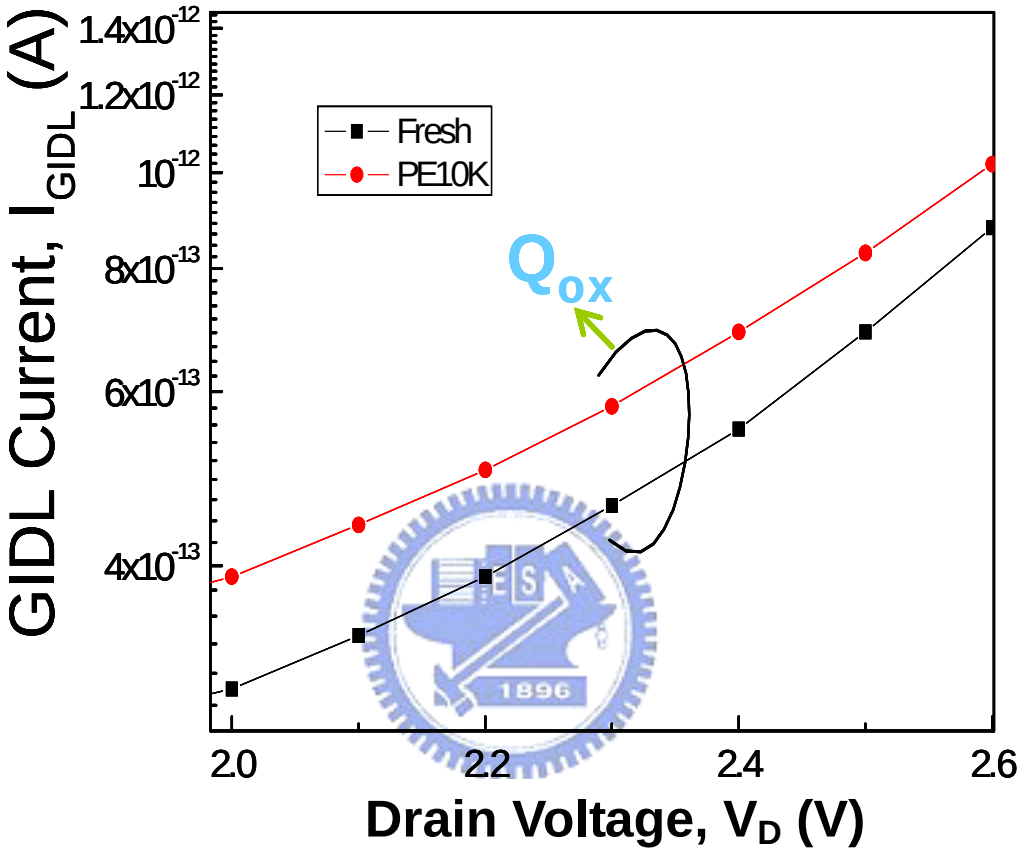


Fig. 3.9 The comparison of GIDL current for fresh and after 10k PE cycles. The increasing of GIDL current indicates that there are oxide traps be generated after cycling.

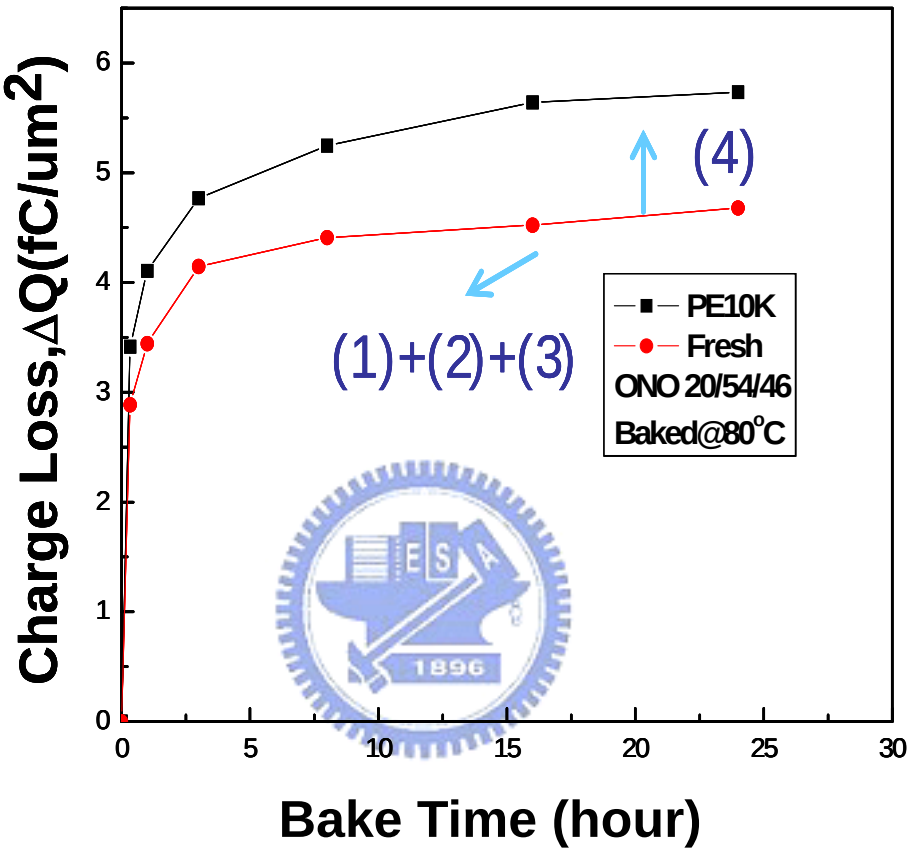


Fig. 3.10 The charge loss of fresh cell and after cycling. We can separate the trap assisted tunneling component from the direct tunneling and thermionic emission components.

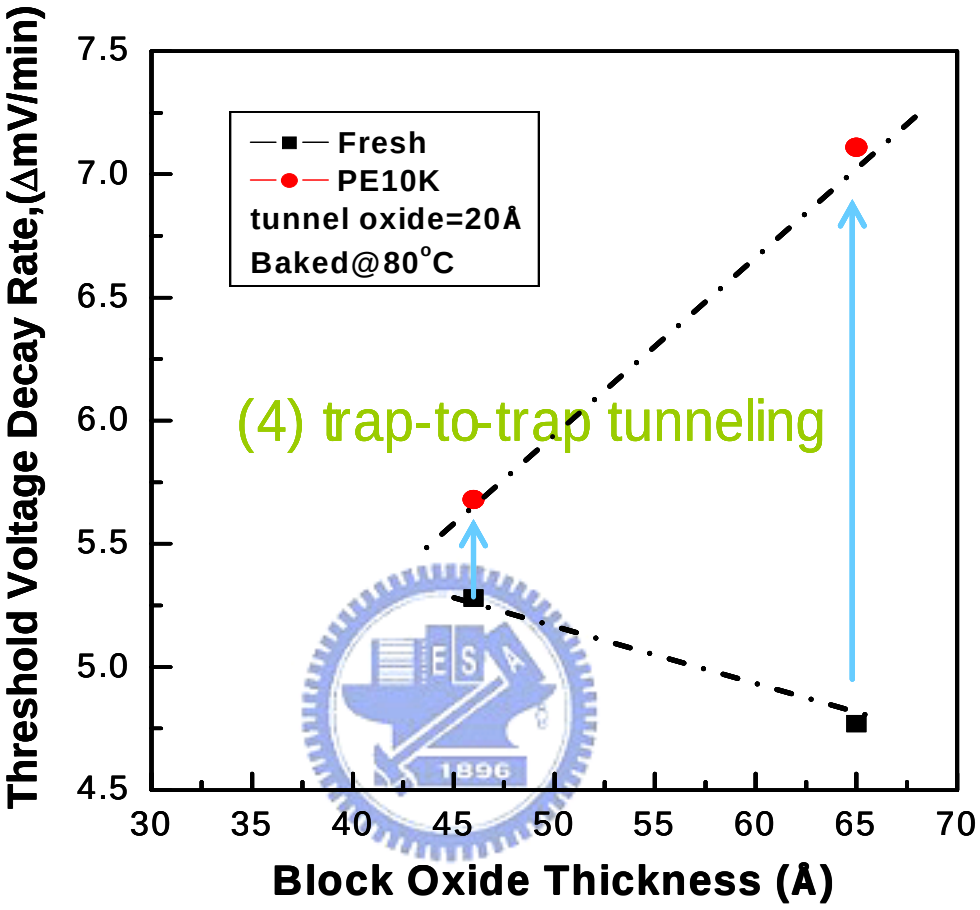


Fig. 3.11 The calculated decay rate versus block oxide thickness for sample no.5 and no.6 measured at fresh and 1000 PE cycles at 80ºC

Besides the lowing of retention ability, the more cycling effects upon the SONOS type memory device will be discussed in the next chapter. It will help us to understand the degradation mechanism during the whole cycling process and to suggest the improvement aspect for further scaling.

Chapter 4
Endurance of SONOS Cell

4.1 Introduction

Endurance is the most important failure rate because the actual endurance is a direct function of the application. Endurance of a SONOS device is a measure of the device's ability to meet a specified retention time as a function of accumulated P/E cycles.

The degradation of the ONO layer of the SONOS structure due to P/E cycles can be divided into three parts: (1) deterioration of the Si-SiO_2 interface of the tunnel oxide in the form of the creation of interface states (traps) with cycling, and (2) deterioration of the bulk Si_3N_4 layer trap density, which manifests itself in the form of increased charge centroid penetration into the nitride film as a function of P/E cycles, and (3) deterioration of the blocking oxide in the form of positive traps enhanced tunneling probability resulting the increase of leakage and the decrease of the programming efficiency.

In this chapter, we will first discuss the endurance, which shows better endurance in thinner tunnel oxide cell. Second, we will investigate the cycling stress effect on the nitride layer. The block oxide degradation of the block oxide will be analyzed. Finally, we define the layout window of the two oxide layers for the further scaling rule of the device according to the criteria of programming speed and charge loss of retention.

4.2 Data Endurance for Cells with Different Tunneling Oxides

4.2.1 The Qualitative Analysis

The samples used in figure 4.1 are SONOS device. The thickness of each ONO stack is 34 Å (blocking oxide), 54 Å (nitride) and 20 Å and 25 Å tunneling oxide respectively. According to Eq. (3.16), we select program/erase bias condition corresponding to the same electrical field of tunnel oxide for each cell. The gate bias is 10.6V and 10V for programming, -7.4V and -7V for erase. Note that two cells keep good operation window up to 10^3 P/E cycles in proportional to their effective thickness between the charge centroid and the gate. This is consistent with our assumption that injected charge is proportional to the initial tunnel oxide field. After 10^3 P/E cycles, the erased threshold voltage increases and reduces the operation window. The operation window decreases more seriously with increasing tunnel oxide thickness as shown in Fig. 4.1.

Because the tunnel oxide is bombarded directly by electrons during P/E cycling, the damage caused by carriers will be accumulated in the oxide after a large number of cycling. However, if the thickness of the tunnel oxide is scaled down, most carriers will directly tunnel through the oxide, without generating damage. The thicker of the tunnel oxide, the more damage is accumulated in the oxide. This result can be seen in charge pumping current measurement of these cells after P/E 10^4 cycles in Fig. 4.2. The technique of charge pumping is often used for detecting the amounts of interface traps. We can see that for a thinner oxide the charge pumping current is still small after 10^4 cycles, it indicates that the carriers hardly interacts with tunnel oxide, thereby creating little damage. On the other hand, the charge pumping current increases with tunnel oxide thickness, which indicates that there are more N_{it} generated on substrate/thicker tunnel oxide interface.

We can justify this phenomenon from the subthreshold swing, S, shown in Fig. 4.3. It is obvious that for the 25 Å tunnel oxide, the subthreshold swing becomes larger after cycling while for the 20 Å case, it is almost the same before and after the cycling.

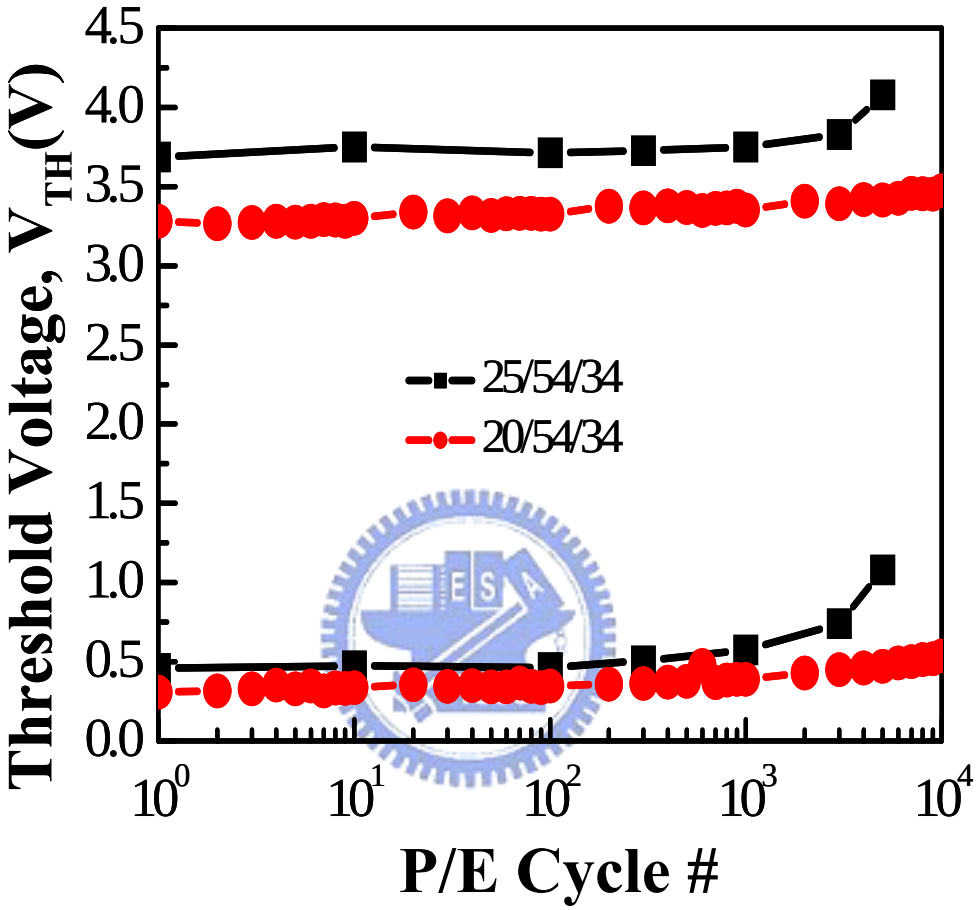


Fig. 4.1 The endurance of three cells, samples 1 and 4, with different tunneling oxide thickness, in which thin tunneling oxide cell shows better endurance characteristic.

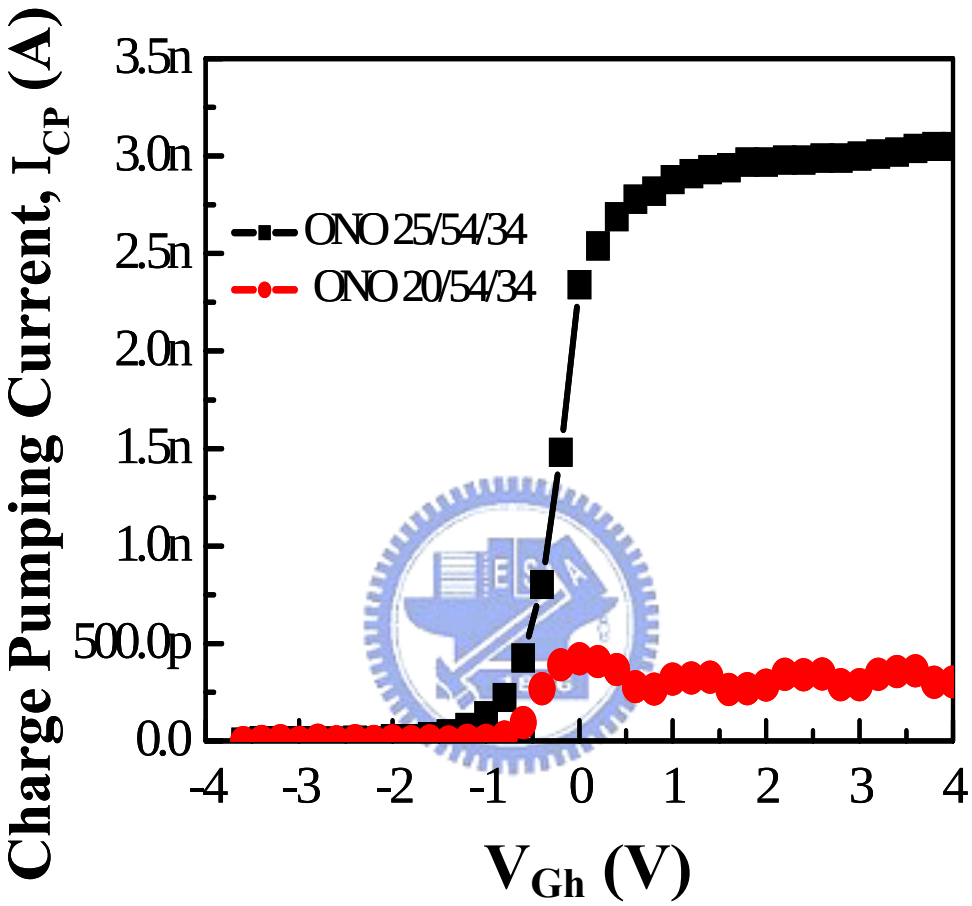


Fig. 4.2 The charge pumping measurement of two cells, samples 1 and 4, with different tunneling oxide thickness. It reavels that the thinner tunneling oxide cell shows much lower charge pumping current.

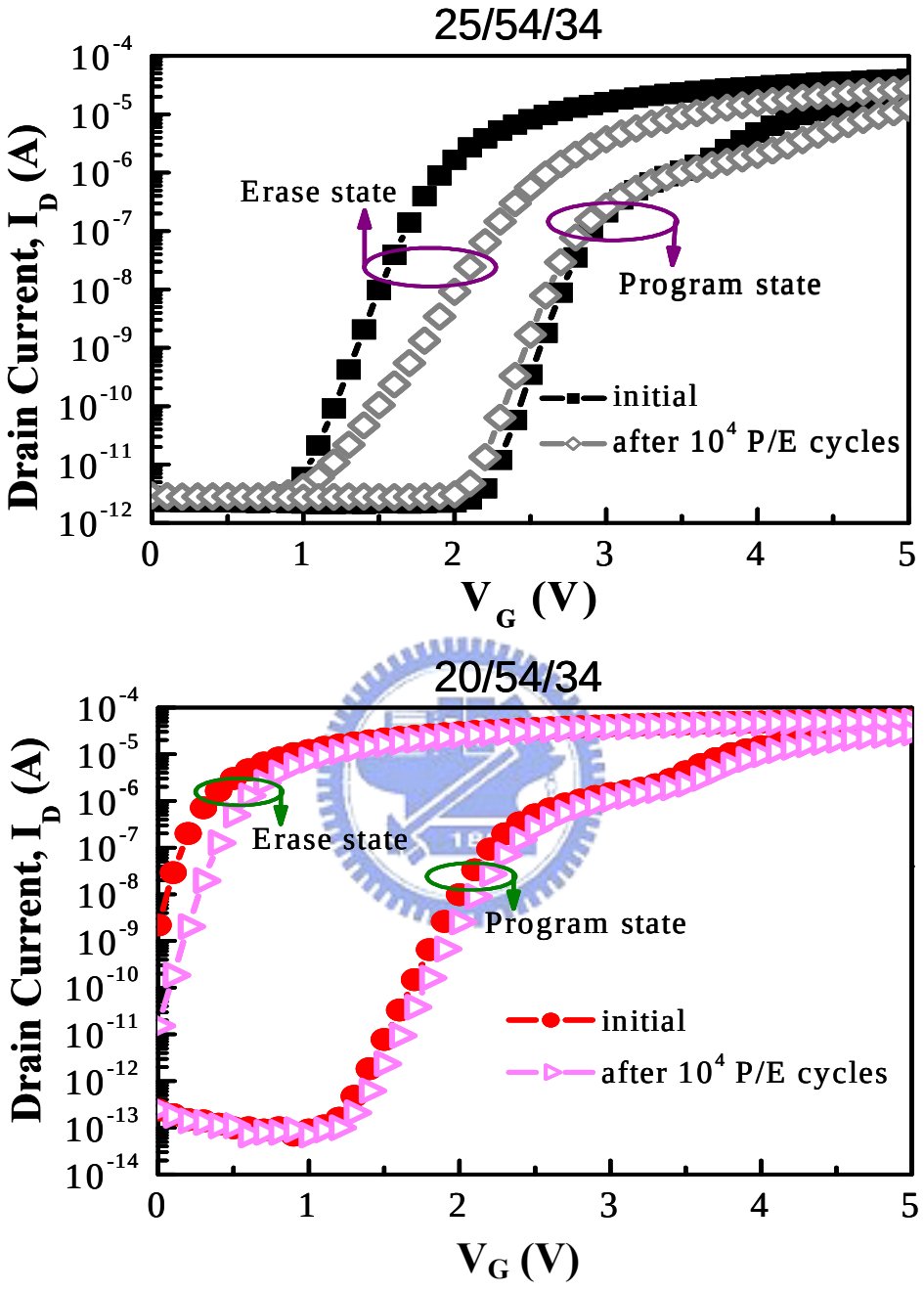


Fig. 4.3 The subthreshlod characteristic of two cells with different tunneling oxide thickness. There is no difference before and after cycling for thinner tunneling oxide.

4.2.2 The Quantitative Analysis

The subthreshold swing is related to the interface trap density with the following equations[25]:

$$S = S_0 \times \frac{1+\frac{C_D + C_{it}}{C_{ox}}}{1+\frac{C_D}{C_{ox}}}, \tag{4.1}$$

where S_0 and S are the subthreshold swing before and after cycling respectively, C_D and C_{ox} represent the depletion-layer and oxide layer capacitance, individually. Here, $C_{it}=qD_{it}$ and D_{it} is the interface trap density. Equation (4.1) can be rewritten as:

$$\Delta D_{it} = \frac{\Delta S}{S_i q}(C_{ox} + C_D). \tag{4.2}$$

We can have a rough estimation of the D_{it} generated by cycling. Here we suppose that substrate doping density N_A is 10^{17} cm^{-3} and get the Debye length L_D and the depletion-layer capacitance C_D at flat band.

$$L_D \equiv \sqrt{\frac{kT\varepsilon_S}{N_A q^2}} = 13.05nm \tag{4.3}$$

and

$$C_D(flat \quad bamd) = \frac{\varepsilon_S}{L_D} = 8.07mF. \tag{4.4}$$

The effective thickness of the ONO layer used in this work is 118 Å, and

$$C_{ox} = \frac{\varepsilon_S}{d_{effect}} = 2.93mF. \tag{4.5}$$

From the Fig. 4.4,

$$S_i = 0.090031 \ , \ S = 0.1009475 \tag{4.6}$$

and

$$\Delta S = S - S_i = 0.0109165. \tag{4.7}$$

The parameters in (4.2) are replaced with the values extracted from eq(4.4) ~ eq(4.7) and we can obtain the value of the interface trap density induced by cycling:

$$\Delta D_{it} = 8.33 \times 10^{11} cm^{-2}. \tag{4.8}$$

These interface traps will contribute to the charge pump current:

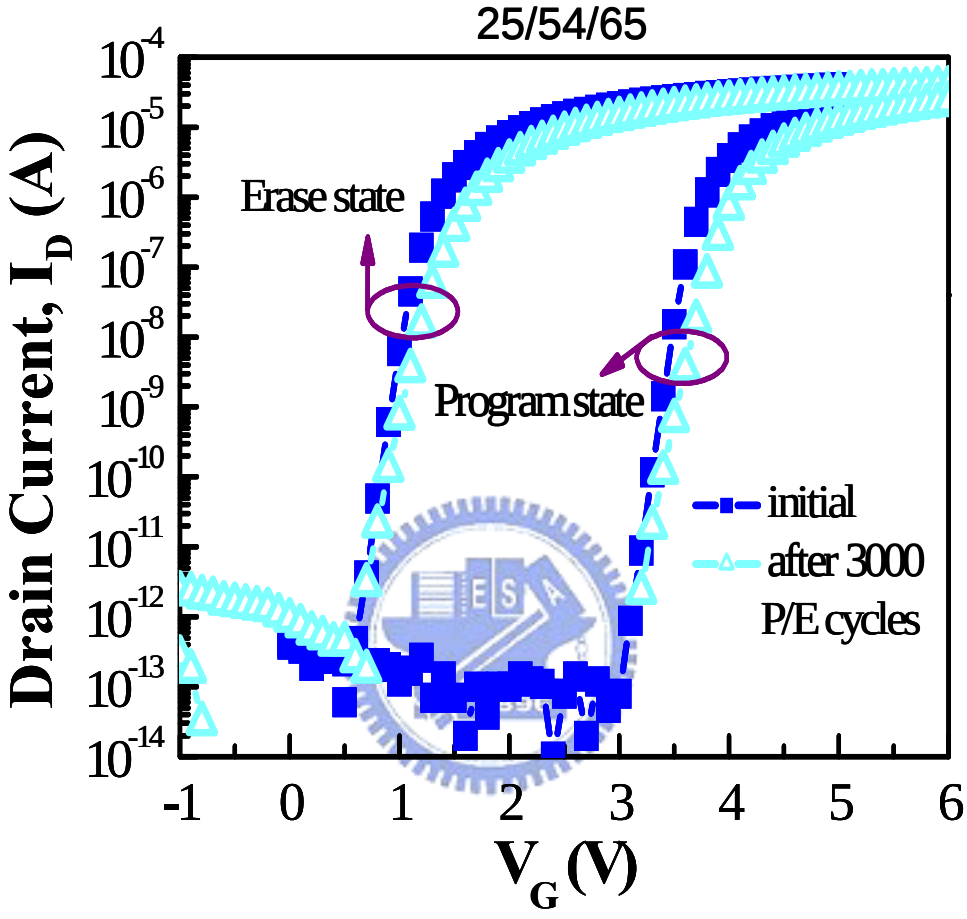


Fig. 4.4 The subthreshlod characteristic of sample no.3. The subthreshold swing changes after 3000 P/E cycles

$$\Delta I_{CP} = q\Delta D_{it} f A_G . \quad (4.9)$$

and the value of ΔI_{CP} equals 1.12nA for f= 1MHz. This is consistent with the actual charge pumping measurement shown in Fig. 4.5.

At last, we used samples no. 1, 2 and 3 of different block oxide thickness to compare the endurance characteristic. We select program/erase bias condition corresponding to the same electrical field of tunnel oxide for each cell. The gate bias is 10.6V, 14.4V and 12.1V for programming, -7.4V -8.5V and -10V for erase. The result is shown in Fig. 4.6. We find that there is no obvious distinction of the endurance between these cells. It shows that the degradation of a thicker tunnel oxide of SONOS memory is caused mainly by the degradation of its tunnel oxide.

4.3 The Influence of P/E Cycles on The Nitride Layer

4.3.1 The Qualitative Analysis

In the past, many authors assume that the major damage part is tunnel oxide either by FN or HC operation. In this section, we will discuss the influence of P/E cycles on the nitride layer.

Here we use sample no. 6 with 20 Å tunnel oxide, 54 Å nitride and 34 Å block oxide to measure the programming transient characteristic before and after P/E cycles, shown in Fig. 4.7. We find a peculiar phenomenon that several P/E cycles can enhance the program and erase speed. It is considerably different from the floating gate flash memory. The wear-out of floating gate memory performances induced by P/E cycling is due to gate oxide degradation. The reduction of the programming efficiency is attributed to oxide traps and interface state generation at drain side of the memory cell as show in Fig. 4.8, a degradation mechanism which is inherent to CHE programming. These interface states reduce electron mean free path, impacting on the hot carrier

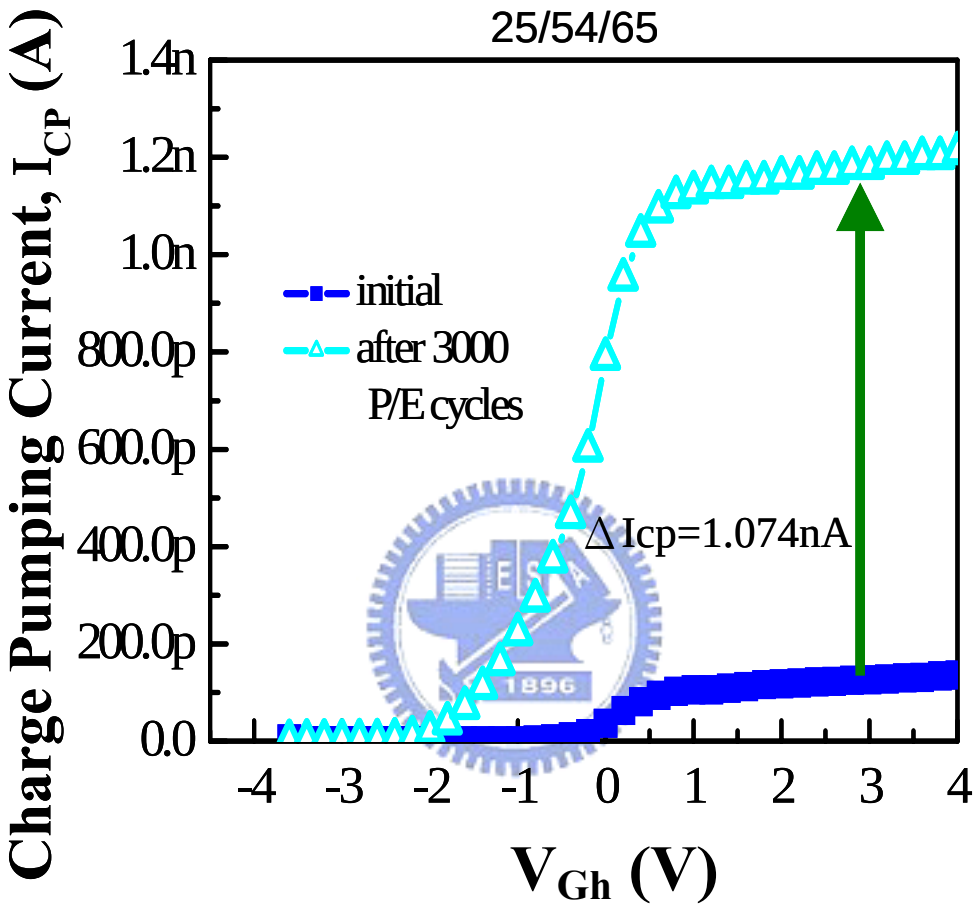


Fig. 4.5 The charge pumping measurement of sample no. 3. The charge pumping current shift is consistent with the value calculated from subthreshold swing in Fig. 4.4.

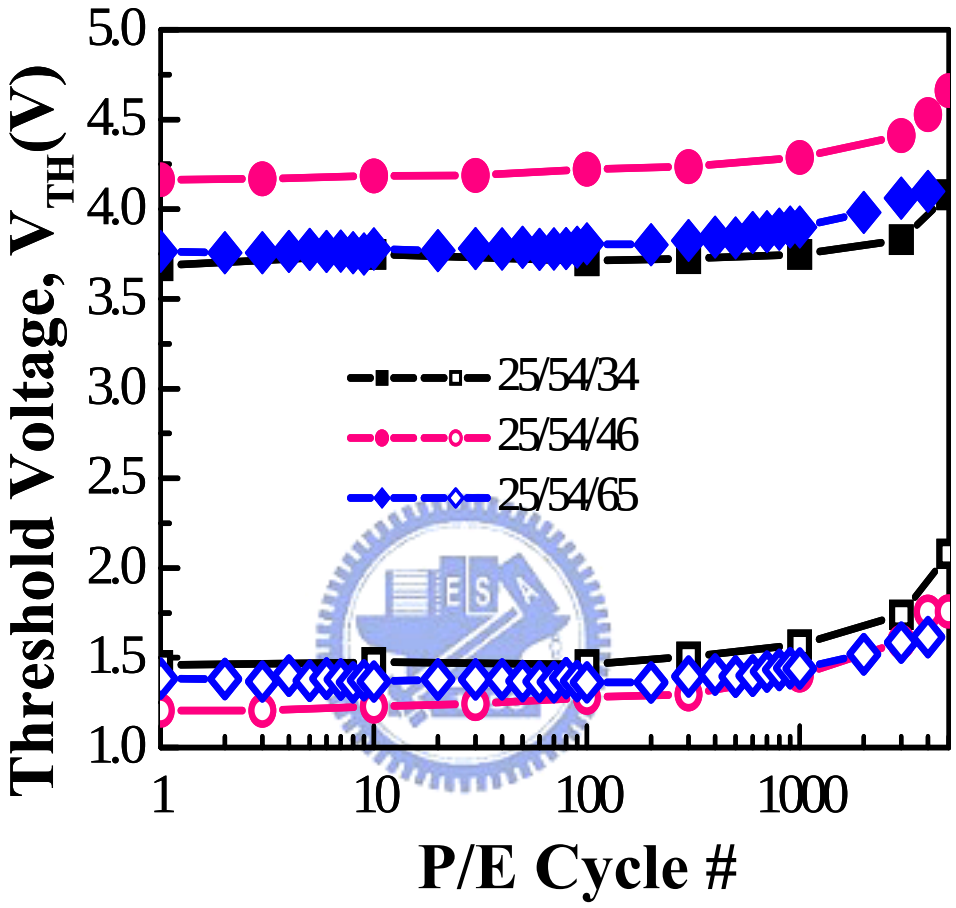


Fig. 4.6 The endurance of three cells with different blocking oxide thickness. There is no obvious distinction between these cells.

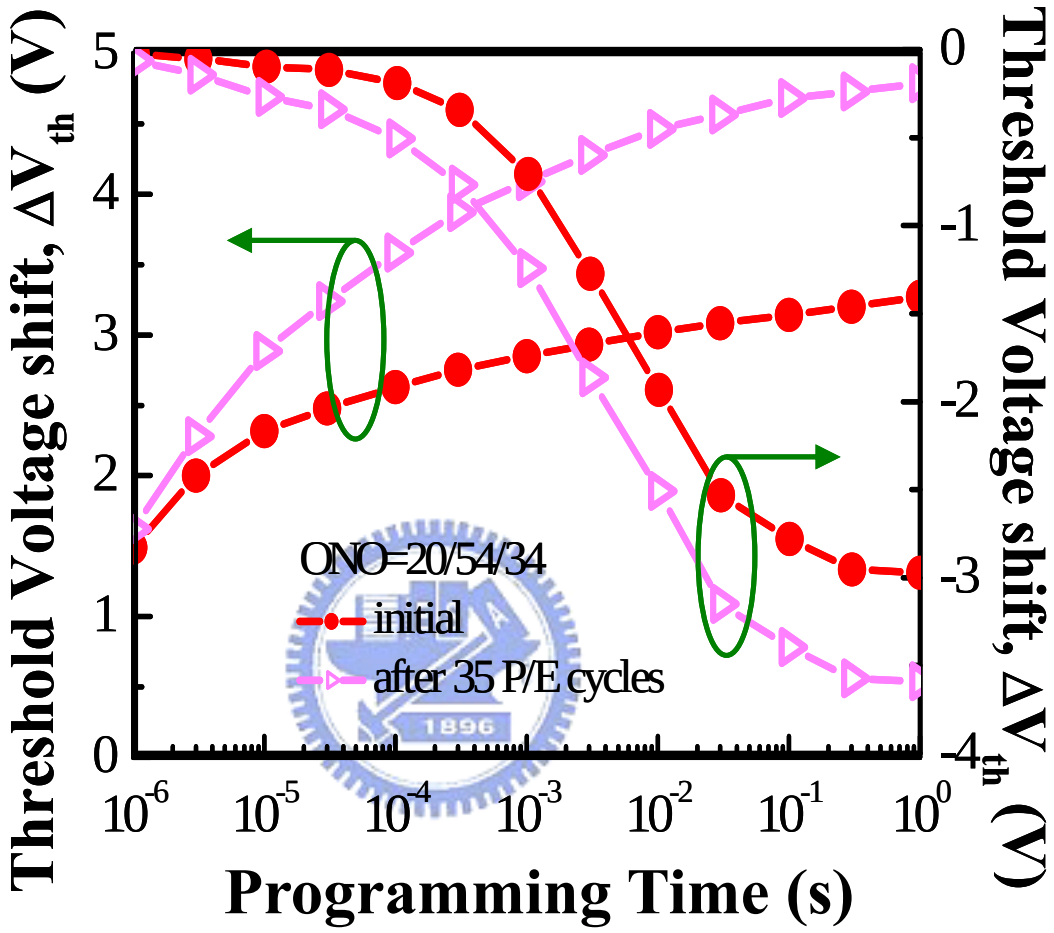


Fig. 4.7 The programming transient of sample 6 before and after cycling. The programming and erase efficiency is enhanced apparently.

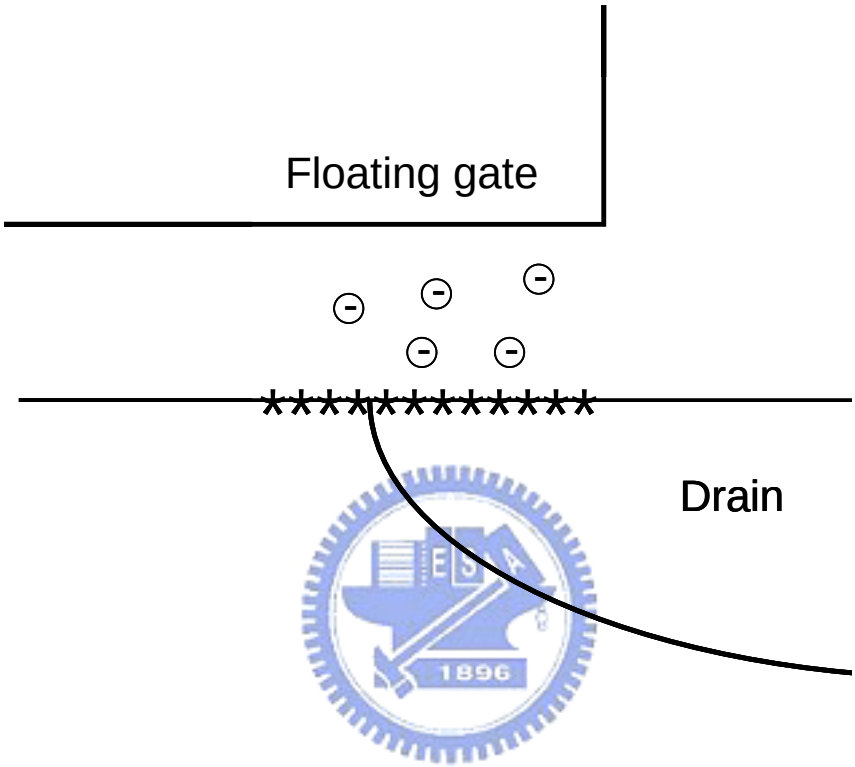


Fig. 4.8 Schematic representation of the location of oxide charge and interface states generated by CHE injection under various voltage condition.

generation mechanism, and electrons trapped in the tunnel oxide modify the electric field at the injection point, reduction programming curves, Fig. 4.9 [26], caused by P/E cycling.

What is the reason that causes the result in Fig. 4.7? We use another wafer to measure the charge pumping current variation due to P/E cycling. The values of charge pumping current measured with different frequency from 20 kHz to 10 MHz before and after 1500 P/E cycles were divided by the frequency to obtain the charge recombined per cycle (Q_{CP}). The result is shown in Fig. 4.10. For the fresh cell, only the interface traps charge and discharge with a charge pumping current I_{CP} directly proportional to frequency f; however, Q_{CP} remains the same irrespective of the measurement frequency, so the curve of I_{CP}/f versus f is almost a horizontal line.

If there are traps located spatially within nitride near the thin tunnel oxide, these traps may contribute to charge pumping current via direct tunneling when their tunneling time constant is shorter than the period of time for the device is held in inversion. This results in an additional current component and concomitant increase in Q_{CP} [27]. After 1500 cycles, we see at Fig. 4.10 that the Q_{CP} increase when the frequency is lower than 60 kHz. We believe that the SiN traps will increase after cycling. This contributes to the increment of charge pumping current.

4.3.2 The Quantitative Analysis

Now, we will calculate the trap level near the nitride/tunnel oxide interface. The time constant of the electron direct tunneling detrapping process has been derived by Lundkvist et al. [28] with a WKB analysis, and given by:

$$\tau_{DT}(x,\phi_T) = \tau_{DT0}\exp\left(\alpha_{ox}^{e} X_{OT}\right)\exp\left(\alpha_{n}^{e} x\right) = \tau_{DT}^{*}\exp\left(\alpha_{n}^{e} x\right), \tag{4.10}$$

where τ_{DT0} is a time constant for the traps studied, ϕ_T is the normalized trap energy, X_{OT} is the thickness of the tunnel oxide, x is the tunneling distance inside the nitride. The coefficients α_{ox}^{e}

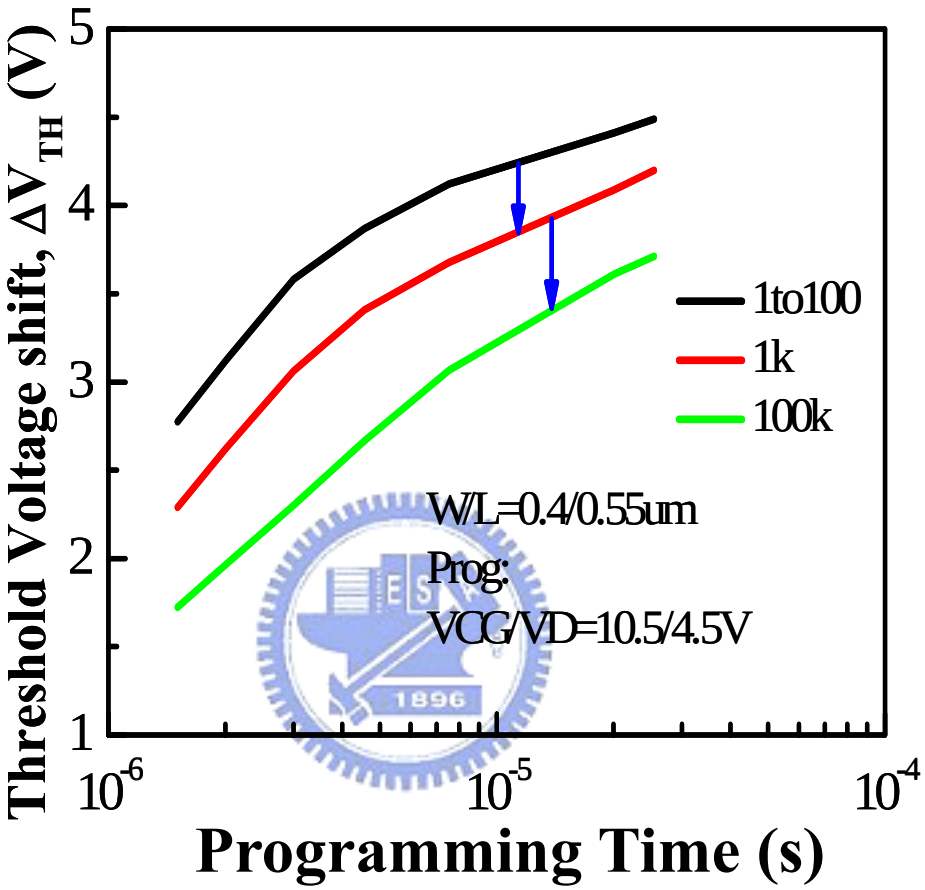


Fig. 4.9 Cycling effect for the floating gate memory reduces the programming curve.

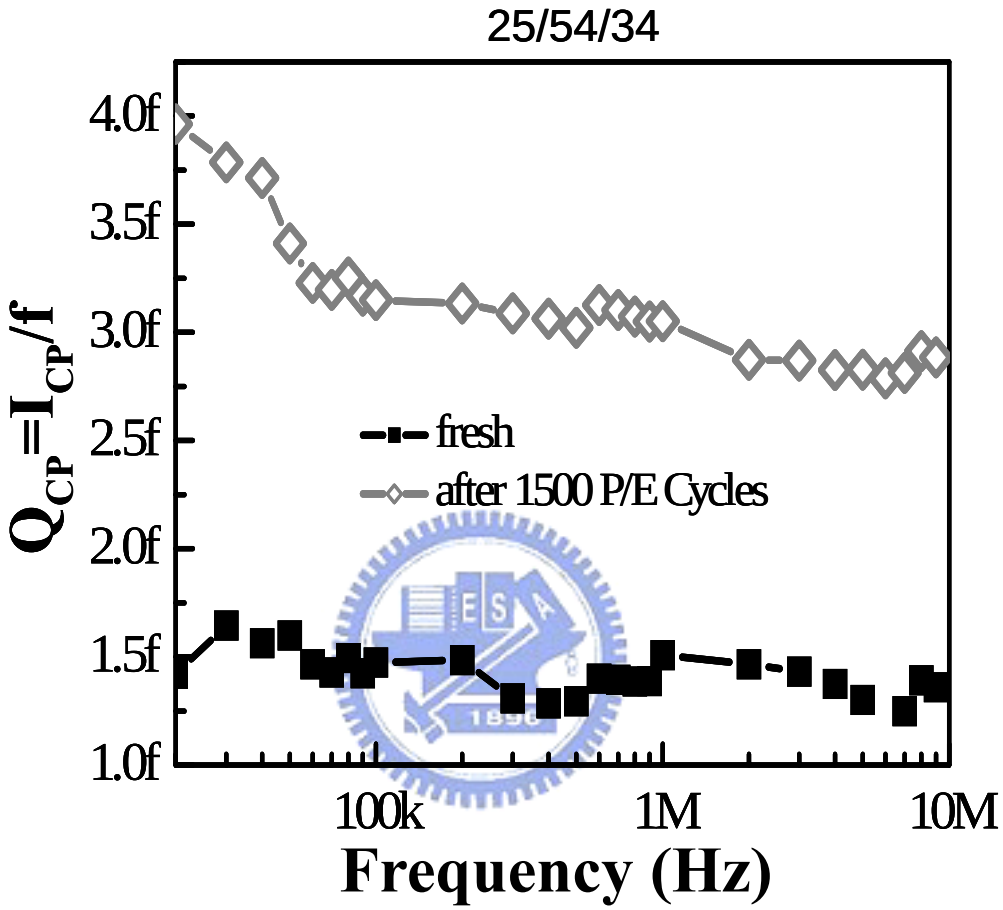


Fig. 4.10 The recombined charge per cycle (Q_{CP}) for the fresh and cycled conditions. The charge increases for lower frequencies confirming the cycling-induced nitride traps behavior.

and α_n^e weakly depend on the electric field across the gate dielectrics. Under a weak internal electric field, they can be approximated as:

$$\alpha_{ox}^{e}(\phi_T) = \frac{4\pi}{h}\sqrt{2m_{ox,e}^{*}q(\phi_2^e + \phi_T)} \tag{4.11}$$

and

$$\alpha_{n}^{e}(\phi_T) = \frac{4\pi}{h}\sqrt{2m_{n,e}^{*}q\phi_T}\,, \tag{4.12}$$

where h is the Plank's constant, $m_{ox,e}^{*}$ and $m_{n,e}^{*}$ are the electron effective mass in the oxide and nitride respectively, and ϕ_2^* is the normalized conduction band offset of the oxide and nitride.

We suppose the values of these parameters are q=1.6 × 10-19C, m_0=9.11 × 10-31kg, $m_{ox,e}^{*}$=5.01×10-31kg, $m_{n,e}^{*}$=3.83×10-31kg, ϕ_2^*=1.05V, h=6.626×10-34 J-s, τ_{DT0}=3.57×10-14s. The direct tunneling time constant equals the inverse of the charge pumping measurement frequency, τ_{DT}=1/f=1/60kHz=16.7 μ s. At the nitride/tunnel oxide interface X_{OT}= 0, then we can obtain the trap level ϕ_T=40mV. It means that when charge pumping frequency is lower than 60 kHz, the increased traps in the nitride begin to contribute to the charge pumping current.

4.4 Data Endurance for Cell with Different Blocking Oxides

Up to the present, we know that the endurance characteristic will be degraded with the tunnel oxide thickness. But the cells with thinner tunnel oxide still have some degradation. What is the cause of the degradation?

Three samples of no. 4, 5, and 6 are used to compare their endurance characteristics shown in Fig. 4.11. The gate bias is 11V, 11.8V and 13.8V for programming, -8V, -8.5V, and -9.6V for erase. After 10^3 P/E cycles, the erased threshold voltage increases and reduces the operation window. The operation window decreases more seriously with increasing blocking oxide thickness as shown in Fig. 4.11. This result can be seen in CP current measurement of these

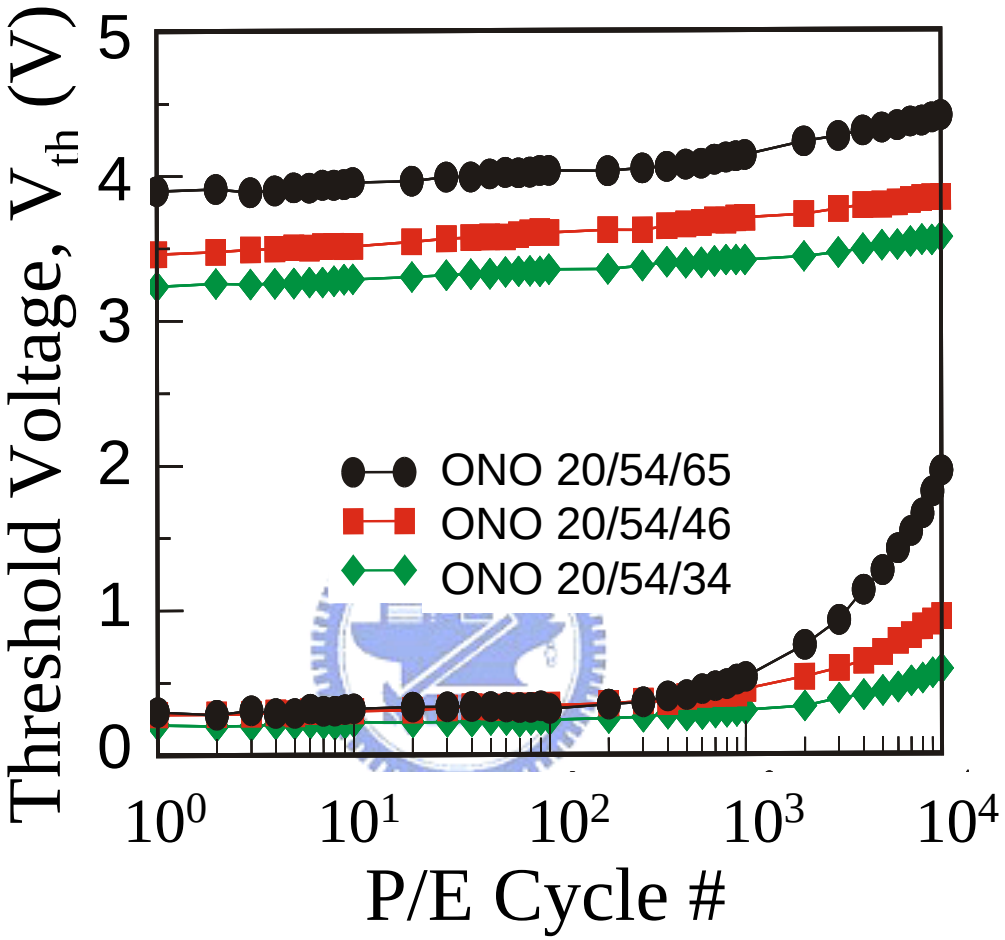


Fig. 4.11 Endurance of cells with various blocking oxide thickness. This shows good operation window for 10^4 PE-cycles and poorer endurance for thicker blocking oxide cell.

cells after P/E 10^5 in Fig. 4.12. Note that CP current and ΔV_{th} increase with increasing blocking oxide thickness, which indicates that either negative charge remains on the substrate or N_{it} exists on substrate/oxide interface.

How to explain this increasing damage in sample no.6, especially for equal applied tunnel oxide field? It may be explained that all three cells have the same initial electrical field during program and erase transients. After several cycles of program/erase, the erased threshold voltage rise. This implies that negative charges remain upon the substrate as shown in Fig. 4.13, and all three cells have the same structure except the blocking oxide thickness. Thus, we think that it is due to electrons trapped in the blocking oxide. Negative charges in block oxide enhance the electrical field of tunnel oxide during erase. On the contrary, oxide field is reduced during program. The more enhanced erase field causes much more damage to the device, and a lower programming field will give rise to a lower operation window. For the cell with thicker blocking oxide, there are more charges trapped in oxide, and hence, more damage and erased threshold voltage shift are observed.

Fig. 4.14 is the programming characteristic of sample no. 4 for the different cycle number. We can see that the programming efficiency after 30 P/E cycles is enhanced; it is the evidence of the increased nitride traps. The following P/E cycling can't generate more traps within nitride, it means that the trap density in the nitride is saturated, and program speed can't be faster. After hundreds of P/E cycling, the threshold voltage shift, ΔV_{TH}, gradually drop down at the later period while it has no change nearly at the beginning. The traps generated by cycling will not disappear, so the programming efficiency almost has no change at beginning. We believe that the latter degradation is related to the degradation of the block oxide. When the number of the P/E cycle times increases, the block oxide will be damaged by the carriers with higher kinetic energy, these damages will lower the "block" ability of the block oxide. After the cycling, when the

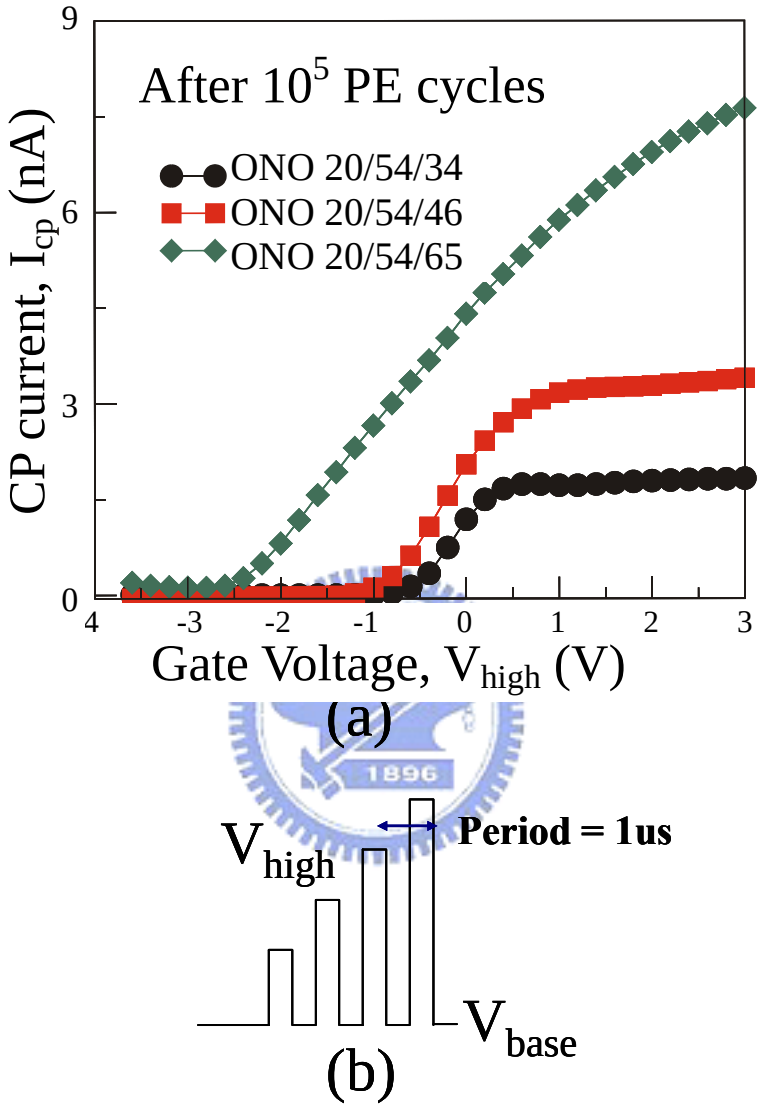


Fig. 4.12 (a) Measured charge pumping current after 10^5 PE cycles for samples 4, 5, and 6. (b) Gate-pulse representation in a CP measurement.

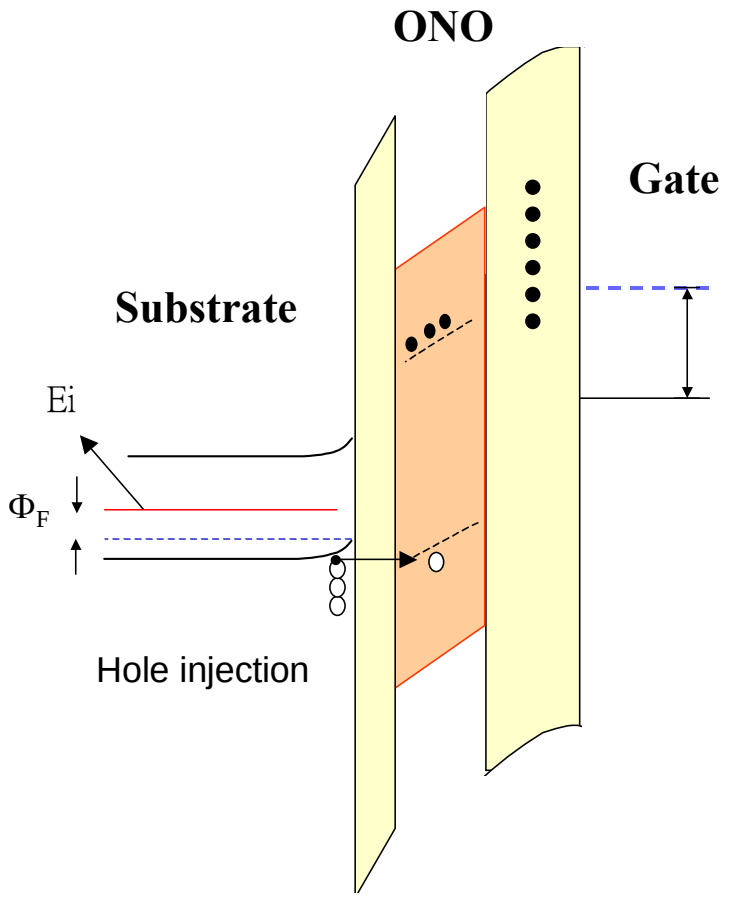


Fig. 4.13 The band diagram of SONOS cell during negative gate bias when electrons are trapped in the blocking oxide.

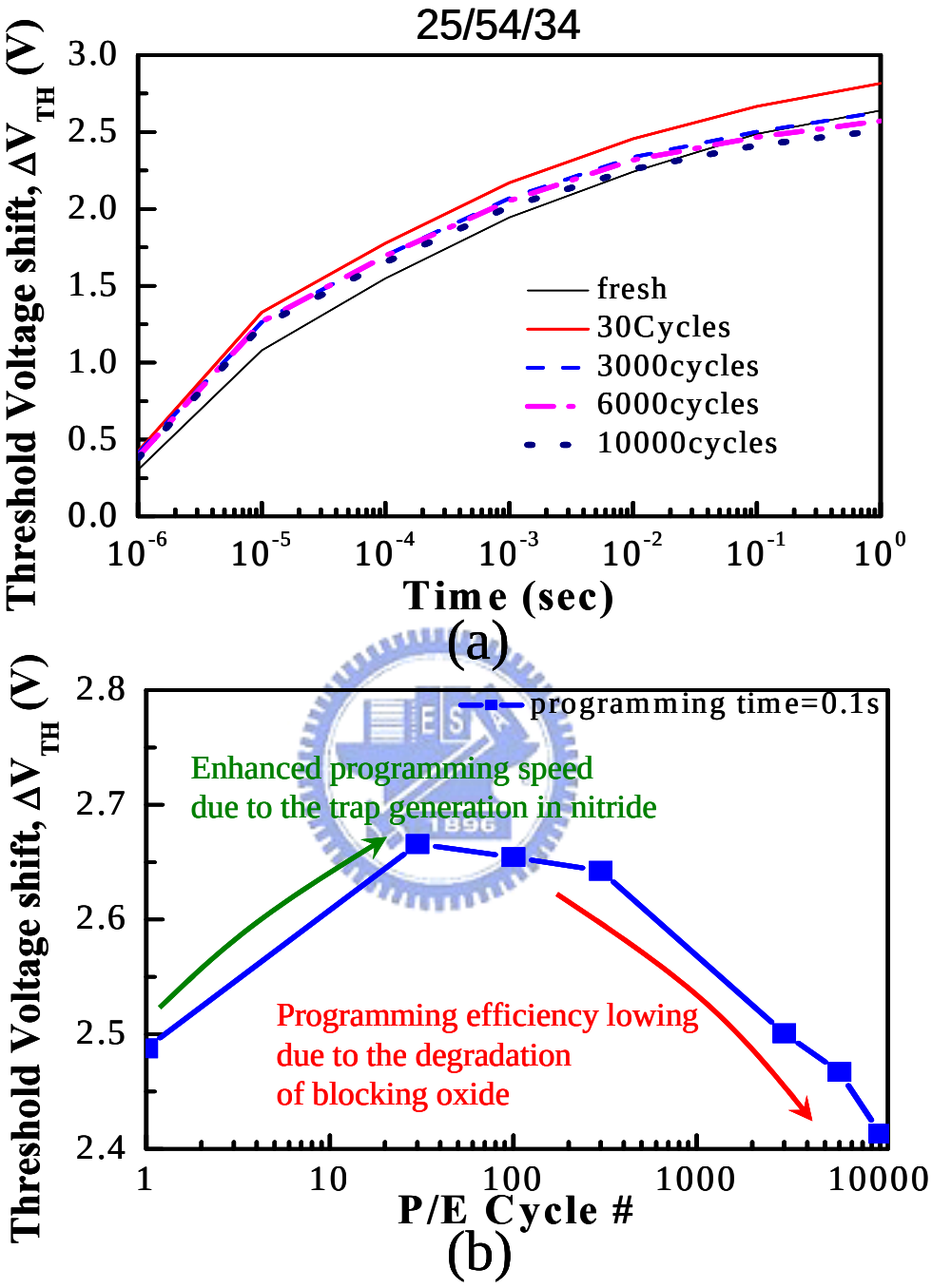


Fig. 4.14 (a) The programming characteristic of sample no. 4 for the different cycle numbers. (b) The enhancement and degradation of the programming efficiency is shown in this figure.

programming time rises, the electrons trapped in nitride have the chance to be swept to the gate through the block oxide along the field build up by the programming voltage. Thus, the amount of stored charge can not reach the saturation and the threshold voltage shift saturate at a lower value.

4.4 Summary

From the results as aforementioned, we propose a rough degradation mechanism about the ONO layer during the whole cycling process. Because of the direct tunneling programming, the thinner tunnel oxide is almost not damaged while the degeneration of the tunnel oxide is the main degradation cause for the thicker tunnel oxide. At early periods of P/E cycling, nitride layer will be damaged by the electron injected from substrate and generate many traps to enhance programming efficiency, and then attain to saturation since the limited volume of nitride bulk. When the cell is programmed and erased more and more times, the degradation of the block oxide gradually becomes serious and dominates the degraded characteristic.

So, in order to improve the endurance of the SONOS flash memory, we have analyzed the degradation mechanism and know that the quality of the block oxide is the important parameter for the continued scaling of tunnel oxide.

Chapter 5
Summary and Conclusion

Based on the experimental results, several issues have been addressed in this thesis. Two major results have been accomplished. First, we have successfully identified the charge loss path and the associated leakage components. These leakage components include the thermionic emission, the direct tunneling, and the trap assisted tunneling. The result helps us to understand how charge leaks in a SONOS cell under a given temperature and blocking oxide. Second, the degradation mechanism of the ONO layer during the period of P/E cycling has been identified.

At first, it was found that the cell with thinner block oxide has more charge loss in various baking temperature. We think that more charge loss is contributed by leakage through blocking oxide, especially when blocking oxide is thinner than 40A. Next, a novel leakage current separation technique has been developed to separate the trap assisted tunneling from the direct tunneling and the thermionic emission. Thermionic emission dominates the long-term leakage while short-term leakage is dominated by the direct tunneling. This will help us to understand the dominant leakage during the scaling of SONOS cells. We also distinguished the charge loss component via traps generated by PE cycling. The retention comparison of cycling effect using various blocking oxides shows that the retention degradation is more serious for thicker blocking oxide under the same cycling times because of its larger capacity for oxide damages generated by PE cycles.

In the second part of this thesis, we discussed the cycling effect on the ONO layer. The result shows that the reduction of the tunneling oxide quality is the main origin of the endurance degradation for cells with thicker tunneling oxide while the less damage can be generated inside thinner tunneling oxide during cycling. The other interesting phenomenon is that there are more

traps generated in the nitride layer after several PE cycles (the short term mechanism of endurance) and enhance the programming speed, and then reach saturated situation. The long term degradation of endurance is dominated by the quality of the blocking oxide. This mechanism is more important and more obvious for thinner tunneling oxide. Since for the same amount of injected charges, the charges are easier to damage the blocking oxide for the thinner tunneling oxide. And by CP measurement, we found that the cell with thicker block oxide suffers much more damage after P/E cycling. We believe it is attributed to a more charge trapping in the blocking oxide that enhances the field of tunnel oxide during erase. In fact, it is a trade-off in terms of the blocking oxide thickness of SONOS memory. Thick blocking oxide cell has larger operation window and less charge loss, but needs larger gate voltage during program and erase and has poorer endurance.

In the future, we would like to investigate SONOS retention and reliability behavior under different operation schemes. Different operation schemes cause different damage location in SONOS cell and may have other leakage path and mechanisms in addition to the direct tunneling and thermionic emission we have examined.

References

[1] S. M. Sze, “Evolution of nonvolatile semiconductor memory: from floating-gate to single-electron memory cell.” In Luryi S, Xu J, Zaslavsky A, editors. Future trends in micro-electronics. John Wiley & Sons; 1999.

[2] I. Fujiwara, H. Aozasa, A. Nakamura, Y. Hayashi, “0.13 μ m MONOS single transistor memory cell with separated source lines,” in *IEDM*, pp. 995-998, 1998.

[3] D. Kahng and S. M. Sze. “ A Floating Gate and its Application to Memory Devices”. *Bell System Technical Journal*, vol. 46, p. 1288, 1967.

[4] S. Minami and Y. Kamigaki, “New Scaling Guidelines for MNOS Nonvolatile Memory Devices,” *IEEE Trans. on ED.*, p. 2519, 1991.

[5] E. Suzuki et al., “A Low Voltage Alterable EEPROM with Metal Oxide Nitride Oxide Semiconductor (MONOS) Structures,” *IEEE Trans. on ED.*, vol. ED-30, pp. 122-128, 1982.

[6] A. Roy, “Characterization and Modeling of Charge Trapping and Retention in Novel Multi-dielectric Nonvolatile Semiconductor Memory Devices,” Lehigh University, 1989.

[7] Y. Yang and M. H. White, “Charge Retention of Scaled SONOS Nonvolatile Memory Devices at Elevated Temperatures,“ *Solid State Electronics*, vol. 44, pp. 948-958, 2000.

[8] H. Bachhofer, H. Reisinger, E. Bertagnolli, and H. von Philipsborn, “Transient Conduction in Multidielectric Silicon–Oxide–Nitride–Oxide Semiconductor Structures,” *J. Applied Physics*, vol 89, p. 2791, 2001.

[9] F. R. Libsch and M. H. White, “Charge Transport and Storage of Low Programming Voltage SONOS/MONOS Memory Devices,” *Solid-State Electronics*, pp. 105-126, 1990

[10] J. Robertson. “The Electronic Properties of Silicon Nitride”, Philosophical Magazine B 44, pp. 215-237, 1981.

[11] W. D. Brown and J. E. Brewer, “Nonvolatile Semiconductor Memory Technology,” *IEEE* press.

[12] L. Lundkvist, C. Svensson, and B. Hansson, *Solid-State Electronics*, vol. 10, pp. 221-227, 1976.

[13] Y. Yang and M. H. White, "Charge Retention of Scaled SONOS Nonvolatile Memory Devices at Elevated Temperatures," *Solid State Electronics*, Vol. 44, pp. 948-958, 2000.

[14] Y. Hu and M. H. White, "Charge Retention in Scaled SONOS Nonvolatile Semiconductor Memory Devices-Modeling and Characterization," *Solid State Electronics,* vol. 36, No. 10, pp. 1401-1416, 1993.

[15] Y. Yang and M. H. White, "Reliability Considerations in Scaled SONOS Nonvolatile Memory Devices," *Solid State Electronics,* vol. 43, pp. 2025-2032, 1999.

[16] A. Roy, "Characterization and modeling of charge trapping and retention in novel multi-dielectric nonvolatile semiconductor memory devices," Ph.D. Dissertation, Lehigh University, 1989.

[17] R. Rofan and C. Hu, "Stress-induced oxide leakage," *IEEE Electron Device Lett.*, vol. 12, pp. 632-634, 1991.

[18] R. Moazzami and C. Hu, "Stress-induced current in thin silicon dioxide films," in *IEDM Tech. Dig.*, pp. 139-142, 1992.

[19] D. J. Dumin and J. R. Madux, "Correlation of stress-induced leakage current in thin oxides with trap generation inside the oxides," *IEEE Trans. Electron Devices*, vol. 40, pp. 986-993, 1993.

[20] M. Kimura and H. Koyama, "Stress-induced low-level leakage mechanism in ultrathin silicon dioxide films caused by neutral oxide trap generation," in *Proc. IEEE Reliability Phys. Symp.*, pp.167-172, 1994.

[21] K. Kobayashi, A. Teramoto, and M. Hirayama, "Electron traps and excess current induced by hot-hole injection into thin SiO_2 films," *in Proc. IEEE Reliability Phys. Symp.*, pp.167-172, 1994.

[22] J. Maserjian and N. Zamani, "Observation of positively charged state generation near the

Si/SiO2 interface during Fowler-Nordheim tunneling," *J. Vac. Sci. Technol.*, vol. 20, pp. 743-764, 1982.

[23] S. S. Chung, P. Y. Chiang et al., "A novel leakage current separation technique in a direct tunneling regime gate oxide SONOS memory cell," in *IEDM Tech. Dig.*, pp. 26.6.1-26.6.4, 2003.

[24] S. Mori, Y. Y. Araki, M. Sato, H. Meguro, H. Tsunoda, E. Kamiya, K. Yoshikawa, N. Arai, and E. Sakagami, "Thickness Scaling Limitation Factors of ONO Interpoly Dielectric for Nonvolatile Memory Devices," *IEEE Trans. on ED*., Vol. 43, No. 1, pp. 47-53, Jan. 1996.

[25] S. M. Sze, "*Phys. of Semiconductor Devices",* John Wiley, 1981.

[26] P. Cappelletti et al., "Flash memory", KAP, 1999.

[27] R. E. Paulsen, R. R. Siergiej, M. L. French, M. H. White, "Observation of near-interface oxide traps with the charge-pumping technique," *IEEE Electron Device Lett.*, vol. 13, pp. 627-639, 1992.

[28] L. Lundkvist, C. Svensson, B. Hansson, "Discharge of MNOS structures at elevated temperatures," *Solid State Electronics,* pp. 221-227, 1976.

[29] C. A. Neugebauer and J. F. Burgess "Endurance and Memory Decay of MNOS Devices", *J. Applied Physics*, vol. 47, pp. 3182, 1976.