A New Technique for Hot Carrier Reliability Evaluations of Flash Memory Cell After Long-Term Program/Erase Cycles

Steve S. Chung, Senior Member, IEEE, Cherng-Ming Yih, Shui-Ming Cheng, and Mong-Song Liang

Abstract—In this paper, we provide a methodology to evaluate the hot-carrier-induced reliability of flash memory cells after long-term program/erase cycles. First, the gated-diode measurement technique has been employed for determining the lateral distributions of interface state \( N_{it} \) and oxide trap charges \( Q_{ox} \) under both channel-hot-electron (CHE) programming bias and source-side erase-bias stress conditions. A gate current model was then developed by including both the effects of \( N_{it} \) and \( Q_{ox} \). Degradation of flash memory cell after P/E cycles due to the above oxide damage was studied by monitoring the gate current. For the cells during programming, the oxide damage near the drain will result in a programming time delay, and we found that the interface state generation is the dominant mechanism. Furthermore, for the cells after long-term use under source-side FN erase, the oxide trap charge will dominate the cell performance such as read-disturb. In order to reduce the read-disturb, source bias should be kept as low as possible since the larger the applied source erasing bias, the worse the device reliability becomes.

Index Terms—Flash memory, hot carrier reliability.

I. INTRODUCTION

HOT-CARRIER reliability has been recognized as a major issue for the design of flash memories [1], [2]. In a certain type of flash memory cell, programming is performed by the use of channel-hot-electron (CHE) injection near the drain junction, while erase is achieved by Fowler–Nordheim (FN) tunneling through the gate oxide region above the source diffusion. Both will generate the so-called oxide damage, which includes the interface state \( N_{it} \) and the oxide trap charge \( Q_{ox} \). These oxide damages will cause serious reliability problems such as programming time delay [3], [4], operation window closure, and gate/read disturb [5]–[7], etc.

In terms of the oxide damage characterization, many efforts have been devoted to determining the spatial distributions of these damages in MOSFET’s [8]–[10]. But, most of them are focused on the profiling of the localized \( N_{it} \) near the drain region only [8]. Moreover, quite a few studies [9], [10] are able to simultaneously extract \( N_{it} \) and \( Q_{ox} \). However, they are not suitable for scaled thin-oxide devices or are not easy to implement. Recently, an improved gated-diode current measurement method was successfully developed which is able to determine both \( N_{it} \) and \( Q_{ox} \) in n-MOSFET’s under various stress bias conditions [11]. As a consequence, the study of these two oxide damages on the flash memory reliability becomes feasible.

In this paper, hot-carrier reliability of flash memories after programming/erase cycles will be investigated. The improved gated-diode current measurement method for characterizing \( N_{it} \) and \( Q_{ox} \) will be used. The effects of these damages on flash cell performance and reliabilities will then be identified.

In Section II, the memory cell and devices used in this study will be described. In Section III, the profiling technique and results will be described. The gate current model will be presented in Section IV. The influence of the oxide damage on device reliability, such as programming time delay and read disturb, will be discussed in Section V. Finally, the conclusion will be given in Section VI.

II. DEVICE PREPARATION

A conventional stacked-gate n-channel flash memory cell using 0.35-\( \mu \)m n-poly silicon gate technology was used in this study with channel length of 0.5 \( \mu \)m and channel width of 0.7 \( \mu \)m. The thickness of tunnel oxide and the effective interpoly oxide are 7 and 20 nm, respectively. The device threshold voltage is adjusted through the ion implantation with doping concentration of 2E17 atoms/cm\(^2\). Moreover, the MDD source and drain structure is used by ion implantation of phosphorus with dosage of 2.5E15/cm\(^2\) and energy of 35 KeV, followed by spacer formation and As implant of dosage 5E15/cm\(^2\) and energy 60 KeV. The calculated gate coupling ratio is 0.6 for the flash cell. In addition, the dummy cell (with the electrically connected control gate and floating gate) for oxide damage characterization was also used. In order to maintain good accuracy for low gate current measurement as well as \( N_{it}, Q_{ox} \) characterization, a dummy cell with large width (width = 20 \( \mu \)m) and the same channel length \( (L = 0.5 \, \mu \text{m}) \) as that of flash memory cell is used. As for the oxide damage measurements, two basic hot carrier stress conditions were performed on dummy cells. One is the hot-electron stress condition at the maximum gate current \( (I_{G,\text{max}}) \) bias (@ \( V_{DS} = 5 \, \text{V}, V_{GS} = 5.5 \, \text{V} \) for 1000 s). The other one is the hot-hole stress or the source-side erasing bias stress condition (@ \( V_{S} = 5 \, \text{V}, V_{G} = -4 \, \text{V} \) for 5000 s). To study the...
flash memory cell reliability after cycling, programming of the flash cell (width = 0.7 μm) by CHE injection was performed at $V_{DS} = 5$ V and $V_{GS} = 10$ V, while erase was accomplished by source FN erase at $V_{S} = 5$ V and $V_{G} = -7$ V.

III. OXIDE DAMAGE PROFILING TECHNIQUE

Fig. 1(a) and (b) shows the schematic diagram of the programming and erase operations of a basic memory cell and the associated oxide damage during programming/erase. To study the generated oxide damage effects on both source and drain sides of a flash cell after P/E cycling, two different hot carrier stress conditions were performed. One is the hot-electron stress condition at the maximum gate current ($I_{G,\text{max}}$) during programming. The other one is the hot-hole stress (off-state stress) condition during erase. Both will generate the interface state and oxide trapped charge under these bias conditions.

To study the P/E cycling-induced oxide damage effects on flash memory performance and reliability, lateral distributions of these oxide damages will first be characterized by an improved gated-diode current measurement technique that we developed recently in [11]. Its basic principle and the characterization results will be described as follows.

Fig. 2 illustrates the new method using CHE stress at $I_{G,\text{max}}$ condition as an example. With a fixed small forward bias applied at the drain and by sweeping the gate voltages, the gated-diode currents ($I_{GD}$) are measured as a function of gate voltages for fresh and stressed devices. Here, the drain is biased at $-0.2$ V. The notations $\phi_e$ and $\phi_h$ represent the quasi-Fermi levels for electrons and holes, respectively, which coincide with the intrinsic level $E_i$. The region $\Delta \varepsilon$ at the surface between $\phi_e$ and $\phi_h$ shows where electron and hole recombination occurs. Moreover, we used the measured gated-diode current ($I_{GD}$) and gate-induced-drain-leakage current ($I_{GIDL}$) as in Fig. 3 and developed a method to separate $Q_{ox}$ from $N_{It}$ by the following steps.

1) Measure $I_{GD}$ of a fresh MOSFET (curve 1).
2) After hot-electron stress, we have $I_{GD}$ (curve 2), which includes both $N_{It}$ and $Q_{ox}$ effects.
3) Use a neutralization or detrapping step (@ $V_D = 3$ V, $V_G = -4.5$ V) to eliminate the effects of hot-electron-induced $Q_{ox}$ (curve 3).

Here, the neutralization or detrapping technique is by applying hot hole injection from the drain to the gate oxide, until $Q_{ox}$ was neutralized and eliminated totally. The $I_{GD}$ current will shift to the left (curve 3) with a decreasing of negative oxide trapped charge. If $Q_{ox}$ is totally eliminated, the $I_{GD}$ current difference between fresh (curve 1) and detrapped (curve 3) one is completely contributed from $N_{It}$. The difference between
the stressed (curve 2) and detrapped (curve 3) $I_{GD}$ currents is used to determine the values of $Q_{ox}$. In the measurement as above, GIDL currents in Fig. 3(b) are used as a monitor of the detrapping procedure to make sure that $Q_{ox}$ is eliminated in step 3). Based on the equations developed in [11], the spatial distribution of $N_{it}$ and $Q_{ox}$ can be determined. Profiling of $N_{it}$ and $Q_{ox}$ can be obtained as shown in Fig. 4. We see that both $N_{it}$ and $Q_{ox}$ are generated for an $I_{G_{\text{max}}}$ stressed devices, where both are localized inside the gate-drain overlap region.

The oxide damage which is mainly caused by hot-hole injection (source-side erase) has also been characterized. The $I_{GD} - V_G$ and $I_{GIDL} - V_D$ curves for fresh, after hot-hole stress, and after detrapping/neutralization are shown in Fig. 5. Here, we use the detrapping step ($@ V_G = 3$ V, $V_G = 3$ V, for 400 s) to detrap the positive $Q_{ox}$. The extracted distributions of $N_{it}(x)$ and $Q_{ox}(x)$ are shown in Fig. 6, which implies that the hot-hole-induced oxide trap charges are very significant and largely distributed in the channel close to the source-side during the source-erase operation.

IV. NUMERICAL MODEL OF GATE CURRENT DEGRADATION

The CHE-injection-generated oxide damage, including $N_{it}$ and $Q_{ox}$, has been considered as the dominant factor for the programming speed delay after P/E cycling. So far, no one has made clear what leads to the programming characteristics degradation. In this section, we will first investigate the correlation between CHE injection current (gate current, $I_G$) and its generated oxide damage. Furthermore, a numerical model of the gate current considering the generated oxide damage will be developed and the programming characteristics of flash cell before and after P/E cycles can then be simulated.

Fig. 7 shows the measured $I_G$ after stress as a function of charge-pumping current variation (or interface states $N_{it}$). The stress condition is biased at the maximum substrate current ($I_{B_{\text{max}}} @ V_{GS} = 2$ V and $V_{DS} = 5$ V). For a conventional S/D device, it is well-known that the stress will generate $N_{it}$ only. According to Fig. 7 and the correlation of $I_{CP}$ and $N_{it}$, $\Delta I_{CP} = qA_{G} f \Delta N_{it}$, the correlation between gate current and $N_{it}$ is drawn as an exponential relationship. The factor $e^{-\alpha_{it} \Delta N_{it}}$ is proposed and regarded as Coulomb scattering due to trapped interface-state charges as described in [12]. In other words, the generated interface states filled with the electrons will serve as Coulomb scattering centers and then suppress the hot-electron injection capability. In addition, the generated $Q_{ox}$ trapped in the oxide after $I_{G_{\text{max}}}$ stress will also inhibit the hot-electron injection probability. Here, the influence of generated $Q_{ox}$ can be regarded as the changes of the effective potential barrier height of hot-electron injection at the Si/SiO$_2$ interface. Therefore, the gate current model

![Fig. 4. Extracted lateral distributions of $N_{it}$ and $Q_{ox}$ for devices with hot-electron stress at $I_{G_{\text{max}}}$.

![Fig. 5. Measured device (a) gated-diode and (b) GIDL currents for fresh, after source-side FN erase-bias stress, and after detrapping/neutralization.

![Fig. 6. Extracted lateral distributions of $N_{it}$ and $Q_{ox}$ for devices with source-side FN erase-bias stress in Fig. 5.

![Fig. 7. Correlation between measured gate current and charge-pumping current variation.](image-url)
Fig. 8. Measured and simulated gate current characteristics for devices at fresh, after hot electron stress, and after detrapped/neutralized conditions.

which considers the effects of generated \( N_{it} \) and \( Q_{ox} \) can be formulated as follows [13]:

\[
J_G = \int_{\Phi_b}^{\infty} q n_e(w_n) f(w_n) \exp(-\alpha_{nit} \Delta N_{it}(x)) \, dw_n 
\]

where \( q \), \( n_e \), and \( \Phi_b \) are the electronic charge, electron concentration, and potential barrier height for electrons to overcome the barrier, respectively. \( w_n \) is the electron energy calculated by energy balance equations. \( v_z(w_n) \) and \( f(w_n) \) represent the electron velocity in the direction normal to the Si/SiO\(_2\) interface and the electron energy distribution (EED) function [14], respectively. \( \alpha_{nit} \) is a fitting parameter for the gate current simulation by considering the \( N_{it} \) effect. The potential barrier height \( \Phi_b \), considering the \( Q_{ox} \) effect, can be suggested as follows:

\[
\Phi_b(x) = 3.2 - 2.7 \times 10^{-4} E^{3/2}_{ox} - 4 \times 10^{-5} E^{1/3}_{ox} + \frac{Q_{ox}(x)}{C_{ox}} 
\]

The quantity 3.2 V is the Si/SiO\(_2\) interface barrier height. The second and third terms in (2) represent the barrier lowering effect due to the image field and the finite probability of tunneling between the silicon and the silicon dioxide, respectively. The last term is the effect of \( Q_{ox} \) on the barrier height.

Fig. 8 shows the measured and simulated gate current characteristics for devices at fresh, after hot-electron stress, and after detrapping/neutralized conditions, respectively. The difference between the HE stressed (solid triangles) and the detrapped (solid rectangles) conditions is due to the \( N_{it} \) effect. On the other hand, the difference between the fresh (solid circles) and the detrapped (solid rectangles) curves is due to the \( Q_{ox} \) effect. Obviously, the \( I_G \) degradation due to \( N_{it} \) is much larger than that due to \( Q_{ox} \). As shown in Fig. 8, the degradation of \( I_G \) has been successfully simulated by using the model described in (1) and (2). The parameter \( \alpha_{nit} \) with value of \( 1.16 \times 10^{-12} \) cm\(^{-2}\) is used.

The programming characteristic simulation is described in the flowchart of Fig. 10. The spatial distributions of \( N_{it} \) and/or \( Q_{ox} \) are incorporated into the drift-diffusion (DD) simulator, energy balance equation, and the gate current model equations. The whole programming operation, beginning at \( t_o \), is divided into time steps with variable duration \( \Delta t \). At a given step \( i \), the drift-diffusion model (DD simulator) and energy balance equation are solved with charges \( Q_{FG}(i-1) \) on the floating gate to calculate the carrier concentration \( n_e \), electric field \( E \), and electron energy \( w_n \), respectively. From the calculated \( n_e \), \( E \), and \( w_n \), \( I_G \) can be simulated. Here, we assumed that \( I_G \) is constant during \( \Delta t \) so that \( I_G \Delta t \) represents the charges \( \Delta Q_{FG} \) injected into the floating gate during \( \Delta t \). The values of \( Q_{FG}(i) \) and threshold voltage shift \( \Delta V_d \) are then calculated and the procedure is iterated until the last time step. In the flowchart, \( T_{ino} \) and \( e_{ho} \) represent the interpoly dielectric thickness and its dielectric constant, respectively.

V. RESULTS AND DISCUSSION

A. Transient Characteristic Simulation

The programmed threshold voltage \( V_{TH} \) degradation (or programming speed delay) of flash memories is one of the major reliability issues caused by CHE-injection programming operation. The degradation, obviously, is mainly due to the reduction of CHE injection probability into the floating gate. From the results shown in Fig. 8, we see that the oxide damage, especially \( N_{it} \), is mainly responsible for the \( I_G \) degradation. Hence, the floating charge \( Q_{FG} \) as well as the threshold voltage will vary as a result of the generated interface states. From the calculated programming characteristics of flash cells before and after P/E cycles, time delay can then be determined.

Fig. 9 shows the measured and simulated programming characteristics of flash memory among fresh, after \( 10^5 \) P/E cycles, and after oxide charge detrapping/neutralization. The device \( V_{TH} \) is defined as the gate voltage required to achieve drain current of 1 \( \mu A \) at \( V_{DS} = 0.1 \) V. In this figure, we see that the degradation of programming characteristics due to
CHUNG et al.: HOT CARRIER RELIABILITY EVALUATIONS OF FLASH MEMORY CELL

Fig. 10. Schematic flowchart of the procedure used for simulation of flash memory cell transient characteristics.

Nth is much larger than that due to $Q_{occ}$. It can be explained that the degradation of gate current is dominated by the $N_{th}$ shown in Fig. 8. Fig. 12 further shows the programming time as a function of P/E cycles, in which the programming time is defined as the time for $V_{TH}$ to change from its initial value $V_{i}$ to $V_{f}$. The individual contributions of simulated programming time by considering the individual effect of $N_{th}$ and $Q_{occ}$ are also shown. It is noted that the $N_{th}$ effect on the programming time delay is larger than the $Q_{occ}$ ones. In other words, $N_{th}$ is the dominant mechanism for the programming time delay of flash memory devices using CHE-injection programming scheme.

C. Erasing Characteristics

In a certain type of flash memory cell, the erase operation is intended to remove electrons from the floating gate and to bring the device back to its low threshold voltage state. FN tunneling is the most typical way used to erase the flash memory array. Depending on the bias configuration, the tunneling process could take place either through the gate oxide over the entire channel or through a small portion of the gate oxide above the source. For the former case (called the channel erase operation), it is accomplished by applying a negative voltage to the control gate with respect to the substrate so that the FN current flows through the gate oxide more or less uniformly over the entire channel. For the latter (called the source-side erase operation), a positive bias is applied to the source junction with respect to the substrate [shown in Fig. 1(b)], which confines the FN current to a small region near the transistor junction region. The advantage of the source-side erase scheme is that the main channel is not damaged during erase, and the $V_{TH}$ degradation can thus be minimized. However, it has been demonstrated that hole injection is difficult to avoid during source-side erase operation. The resulting hole injection will give rise to the creation of oxide charges/traps in the oxide and is a strong function of source erasing voltage. Therefore, for the consideration of erase performance and reliability, the applied source voltage during erase should be optimized more carefully [15], [16].

In order to evaluate the source bias effects on hot carrier reliability of flash memory cells, the hot-hole-injection-induced $Q_{occ}(x)$ for dummy cells are shown in Fig. 13 with three different source erase schemes. The source-to-gate voltages are kept constant (9 V) for constant oxide field and the source voltages are varied from 3–7 V. It can be seen that the oxide charges are almost distributed in the channel region with a wide range. The peak density of oxide charges increases and is shifted toward the channel for large source erasing bias. Therefore, the higher the applied source bias during erase, the worse the hot carrier reliability of flash memory devices becomes.

In source-side erased flash memories, hot-hole-injection-induced read disturb failure is another important reliability issue [6], [7]. Basically, read disturb occurs in an erased cell, in which its $V_{TH}$ increases unintentionally during read operation. Fig. 14 shows the read disturb failure and the characterized $Q_{occ}(x)$ for a dummy cell after different read disturb times. Hot-hole stress is performed at $V_{C} = -4$ V and $V_{S} = 5$ V for 5000 s and then read is achieved at $V_{C} = 3$ V and $V_{D} = 1$ V (i.e., $V_{CC} = 5$ V, $V_{D} = 1$ V for flash memory device). For a flash memory device after P/E cycles, hot-hole-
injection-induced oxide charges and oxide traps in the oxide during erase under a small positive gate voltage will result in 1) hole emission from the oxide to the substrate (Mechanism I in Fig. 15) and 2) electron trapping from the substrate to the generated oxide traps (Mechanism II) [7]. Both will give rise to the oxide charge fluctuation and then increase cell threshold voltage.

From Fig. 14, it can be seen clearly that the effective positive gate voltage decreases gradually with the increase of read disturb times. The decrease of $V_{TH}$ in the channel will result in an increase of cell $V_{TH}$. From Fig. 15, it can be seen clearly that the effective positive $Q_{ox}$ decreases gradually with the increase of read disturb times. The decrease of $Q_{ox}$ in the channel will result in an increase of device $V_{TH}$. Fig. 15 shows the read disturb characteristics of flash memory devices after $10^5$ P/E cycles. Source bias effect on read disturb failure is also shown for three different source erase schemes given in Fig. 13. We see that the device threshold voltage increases with the read times. In addition, the read disturb failure becomes worse for a larger source erasing bias.

VI. DISCUSSION

In the study of CHEI gate current degradation and the flash cell programming delay, two different results were reported. One is proposed by Peng et al. [4], who claimed that the oxide charge creation is the major cause of the flash cell degradation. However, it cannot explain the gate current degradation due to $N_{it}$ as given in Fig. 8. The other one is proposed by Yamada [3], who observed that $N_{it}$ located in the drain overlap region dominates the programming delay.

In this work, the individual effect of $N_{it}$ and $Q_{ox}$ on the flash cell programming characteristics has been clearly identified. Results tell us that Yamada’s paper is correct. Peng’s result is wrong. Also, we proposed a complete methodology to verify flash cell delay after cycling that is more practical and has never been reported before.

For the study of hot-hole-injection-induced reliability in flash memories, in the past, San et al. [16] pointed out indirectly that the oxide charge was the main cause to degrade flash performance by the use of GIDL current measurement in a dummy cell (MOSFET). However, the correlation between $Q_{ox}$ and the reliability (such as read disturb) cannot be obtained in [16]. In this study, we quantitatively determine $Q_{ox}(x)$ before and after read disturb. Also, the effect of $Q_{ox}(x)$ on read disturb characteristics can be simulated and analyzed quantitatively. Our results verify that the read disturb failure in source-erased flash memories is mainly due to the electron-trapping/hole-detrapping-induced $Q_{ox}$.

Moreover, we should point out that San in [16] suggested that source voltage of 2 V during erase has the minimum damage in the fixed $V_{th}$ condition. However, Huang in [17] observed that the hot-hole injection current is dependent on the source structure (for hole current of 100 pA, $V_{th} = 1.4$ V in $n^+$ dose of $4E15$ and 4 V in $2E15$). Therefore, the optimized erase source bias is strongly dependent on the source injection structure.

VII. CONCLUSION

In summary, the hot-carrier-induced reliabilities on flash memories after P/E cycles have been studied by an improved gated-diode current measurement technique. This technique is able to simultaneously determine the lateral distributions of both interface states and oxide trap charges near both the source and drain sides for flash memory cells after cycling. Two major memory cell characteristics such as programming time delay and read-disturb as a result of these oxide damage were then studied.

The interface state and oxide trapped charge distributions were first calculated for memory cells after $L_{G_{max}}$ and off-state stresses, respectively. A gate current model was then developed by including both the distributions of $N_{it}$ and $Q_{ox}$. 
which is able to study their individual effect on the cell characteristics after cycling. Degradation of flash memory cell after P/E cycles due to the oxide damage has been identified. It was found that the interface state will dominate the device degradation during programming, while the oxide trap charge will dominate the cell performance during source-side FN erase operation. Moreover, to reduce the effect of source erasing bias on the cell read-disturb characteristics, source bias should be kept as low as possible since the larger the applied source erasing bias, the more the oxide trapped charges will be generated in the channel near the source-side. This will cause larger threshold voltage shift and leads to poorer cell reliability after long term cycling.

REFERENCES


