electronics

MDPI

Review

Technique for Profiling the Cycling-Induced Oxide Trapped Charge in NAND Flash Memories

Yung-Yueh Chiu * and Riichiro Shirota

Department of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan; riichiro.shirota@gmail.com

* Correspondence: yungyueh.chiu@gmail.com

Abstract: NAND Flash memories have gained tremendous attention owing to the increasing demand for storage capacity. This implies that NAND cells need to scale continuously to maintain the pace of technological evolution. Even though NAND Flash memory technology has evolved from a traditional 2D concept toward a 3D structure, the traditional reliability problems related to the tunnel oxide continue to persist. In this paper, we review several recent techniques for separating the effects of the oxide charge and tunneling current flow on the endurance characteristics, particularly the transconductance reduction ($\Delta G_{m,max}$) statistics. A detailed analysis allows us to obtain a model based on physical measurements that captures the main features of various endurance testing procedures. The investigated phenomena and results could be useful for the development of both conventional and emerging NAND Flash memories.

Keywords: NAND Flash memory; endurance; reliability; oxide trapped charge

Citation: Chiu, Y.-Y.; Shirota, R. Technique for Profiling the Cycling-Induced Oxide Trapped Charge in NAND Flash Memories. *Electronics* **2021**, *10*, 2492. https://doi.org/10.3390/electronics10202492

Academic Editor: Changhwan Shin

Received: 2 September 2021
Accepted: 7 October 2021
Published: 13 October 2021

1. Introduction

The emergence of NAND Flash memories has revolutionized the data storage industry over the last few decades. NAND Flash devices are used in a wide range of applications in everyday consumer electronics such as laptops, tablets, and smart wearable devices. The first NAND-structured cell was invented in 1987 by Masuoka et al. [1] at Toshiba Corp. Since then, several improvements have been proposed to lower the power consumption of these cells and to enable the contents of the entire chip to be erased at once [2–7]. More recently, its application range has been expanded such that it has become the main storage element, in that solid-state drives (SSDs) are gradually replacing hard disk drives (HDDs) [8,9]. Furthermore, it is increasingly adopted for enterprise-class storage systems. As a result, the size of NAND cells has aggressively shrunk to continuously promote this evolution. However, the ever-shrinking dimensions of the NAND cell create additional challenges in terms of the endurance and retention characteristics, such as random telegraph noise (RTN) fluctuations of the threshold voltage (V_T) [10–12], charge trapping/detrapping mechanisms [13–15], electron injection statistics [16,17], and V_T distribution widening due to parasitic coupling effects [18,19].

Three-dimensional (3D) NAND Flash memories can be considered as a breakthrough to continue to deliver increasing bit density and reduce the bit cost [20]. 3D NAND Flash technology can utilize either floating gate (FG)- or charge trapping (CT)-type cells. Most of the 3D NAND reported to date are CT-type, owing to the simpler fabrication process [21]. The 3D NAND array architecture can be categorized into the following two classes depending on the direction of channel, as schematically shown in Figure 1: vertical gate 3D NAND architecture, which was proposed by Samsung Electronics in 2009 [22]; and vertical channel 3D NAND architecture. There are two main cell structure types that use vertical channels, namely bit cost scalable (BiCS), which was proposed by Toshiba Corp. in 2007 [23,24], and terabit cell array transistor (TCAT), which was developed by Samsung Electronics in 2009 [25]. TCAT subsequently evolved into V-NAND architecture, which has

32-stacked word line (WL) layers [26–28]. The industry has moved beyond 12x-stacked WL layers and achieved a 17x-stacked V-NAND [29,30]. As the memory industry transitions from planar to 3D scaling, traditional device reliability issues must still be considered. The Fowler Nordheim (FN) tunneling mechanism is commonly used in both planar and 3D NAND cells during programming and erasing (P/E) operations [31]. This mechanism leads to the formation of trap states in the tunneling oxide, and thus degrades the oxide reliability. Therefore, overcoming the reliability problems related to the oxide trap is critically important for the development of future advanced NAND Flash memories.

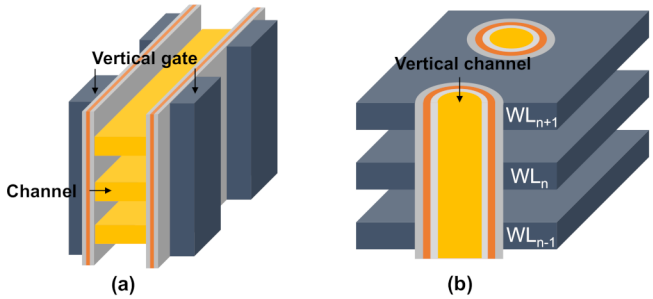

Figure 1. Schematic diagrams of 3D NAND architecture: (**a**) vertical gate and (**b**) vertical channel.

2. Shift in the Midgap Voltage

Generally, the midgap voltage (ΔV_{MG}) during P/E operations is described by a set of two components [32]: the first is the electrostatic shift (ES) that is caused by the creation of oxide trapped charges (Q_T), and the other is the tunneling shift (TS) that is related to the change in the number of floating-gate charges (Q_{FG}). Notably, these two components mutually influence each other. The former deforms the tunneling barrier for P/E operations and thus reduces the number of storage electrons. ΔV_{MG} can be expressed as the sum of these two components.

$$\Delta V_{MG} = \frac{Q_T}{C_i} + \frac{\Delta Q_{FG}}{C_{IPD}} \quad (1)$$

where C_i and C_{IPD} are the tunneling oxide and the interpoly dielectric capacitance, respectively.

Several approaches have been proposed to separate the ES and TS values from ΔV_{MG}. The first category of methods is based on indirect measurements. For example, the ΔV_{MG} in the programming and erasing states combined with tunneling-based modeling is commonly monitored to extract the Q_T distribution from Q_{FG} in NAND Flash memories. Q_T has been presented by a sheet charge located at fixed distance from the channel in the majority of the literature [32–34]. Under this assumption, the tunneling current is calculated straightforwardly along the direction perpendicular to the channel by using the Wentzel–Kramers–Brillouin (WKB) approximation, as schematically shown in Figure 2a. However, as the cell sizes are aggressively shrunk to the nanoscale regime, they are adversely affected by the discrete nature of Q_T. Thus, we must consider all possible tunneling paths across the defective oxide [35,36], as schematically shown in Figure 2b, which increases the computational time and complexity of the method.

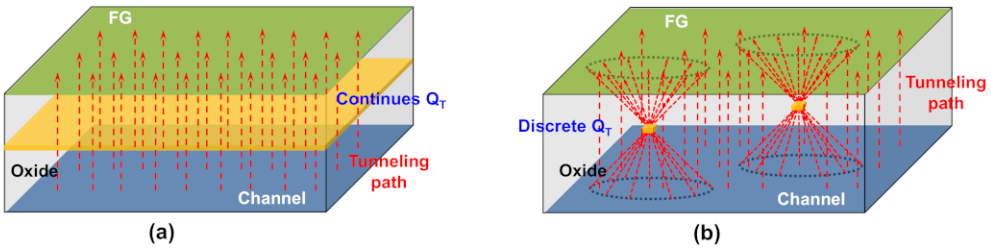

Figure 2. Schematic diagrams of all the possible tunneling paths by the (**a**) continuous and (**b**) discrete Q_T during P/E cycles.

The second category of methods is based on the direct extraction of Q_T and Q_{FG} using a special test device [37,38]. The cross-sectional view and equivalent circuit of the test structure are shown in Figure 3. The device is composed of two memory cells: one with a thick tunneling oxide, referred to as a high-voltage (HV) cell, and the other with a thin tunneling oxide, referred to as a low-voltage (LV) cell. Notably, these two cells have a common FG/common control gate (CG) configuration. During P/E operations, FN tunneling occurs only through the oxide of the LV cell, thus degrading the oxide of this cell. The ES resulting from Q_T is expressed as [38]:

$$\mathrm{ES} = \gamma[\Delta V_T(LV) - \Delta V_T(HV)] = -\frac{1}{\varepsilon_{ox}} \int_0^{T_{OX}} \rho \cdot x dx \quad (2)$$

where γ is the coupling ratio between the FG and CG, ρ is the density of Q_T, and $\Delta V_T(LV)$ and $\Delta V_T(HV)$ are the V_T shifts of the LV and HV cells after P/E cycles, respectively. Unfortunately, the size of the test device (L = 4 μm) is relatively large compared to that of conventional NAND Flash memories, yet it is necessary to continuously evaluate these miniaturized and new device structures. Moreover, this approach can only provide average information for a relatively large sample region rather than statistical information.

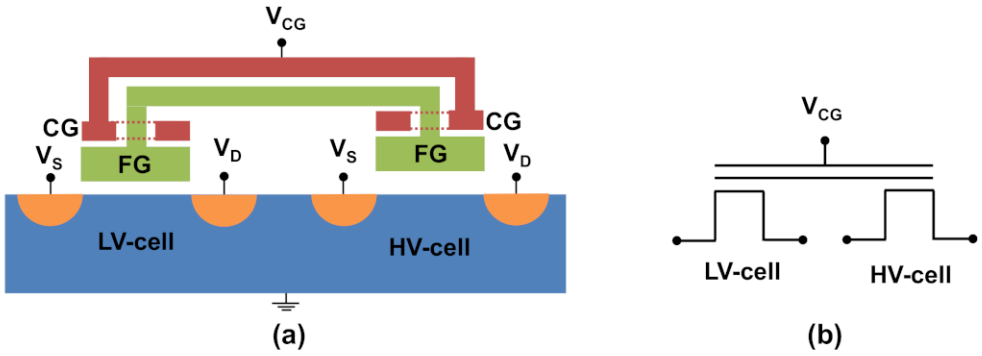

Figure 3. Schematic cross-section view (**a**) and equivalent circuit (**b**) of the test device. Adapted from [37,38].

3. $\Delta G_{m,max}$ Statistics

To overcome the limitations of the above-mentioned approaches, we proposed a statistical transconductance reduction ($\Delta G_{m,max}$) method [39], which enables the extraction of Q_T from Q_{FG} in both 2D and 3D NAND memories.

3.1. Experimental Setup

Experiments are carried out in 2D FG-type NAND Flash memory chip. In the NAND array, a string is composed of 32-unit cells, a source-select transistor, and a drain-select transistor, as schematically shown in Figure 4a. The control gates, source-select transistors, and drain-select transistors are connected across different strings to constitute the wordline (WL), source select line (SSL), and drain select (DSL), respectively. The strings are connected to a common sourceline (SL) and bitlines (BLs). The channel length (L) and width (W) are both 42 mm, and the tunneling oxide thickness (T_{ox}) is 8 nm. The measurement scheme was as follows: the program operation is performed by adopting the incremental step pulse programming (ISPP) technique [40] with a starting CG voltage ($V_{CG,0}$) in increments of 0.2 V with a duration of 10 μs, as schematically shown in Figure 4b, driving the selected cells to the desired V_T level. The erase operation is performed on blocks by adopting the incremental step pulse erasing (ISPE; similar to the ISPP) technique. Because it is not possible to apply high negative voltages in NAND chips, a high positive voltage is applied to the p-well. As a result, all cells in the block were erased simultaneously. During the read operation, the CG gate voltage was swept from 0 V to 5 V to harvest the maximum transconductance reduction ($\Delta G_{m,max}$). Figure 5a shows the $I_D - V_{CG}$ characteristics of the

200 randomly selected cells on WL15 in NAND strings before cycling and after 1 k, 3 k, and 30 k P/E cycles, respectively. Then, the corresponding $\Delta G_{m,max}$ distribution can be obtained, as shown in Figure 5b. Notably, the endurance test and $\Delta G_{m,max}$ monitoring were performed at room temperature. The mean value of $\Delta G_{m,max}$ ($\overline{\Delta G}_{m,max}$) is clearly observed to increase, and the distribution to become wider, as the number of cycles increases. This suggests that the $\Delta G_{m,max}$ distribution will be a good parameter for evaluating the oxide degradation.

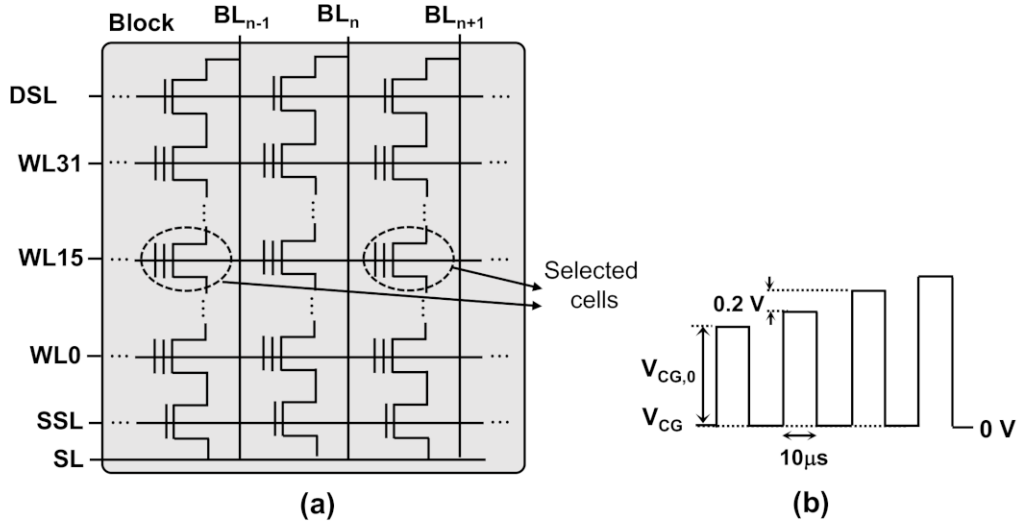

Figure 4. Schematic view of (**a**) NAND Flash array and (**b**) ISPP operation. Adapted from [39].

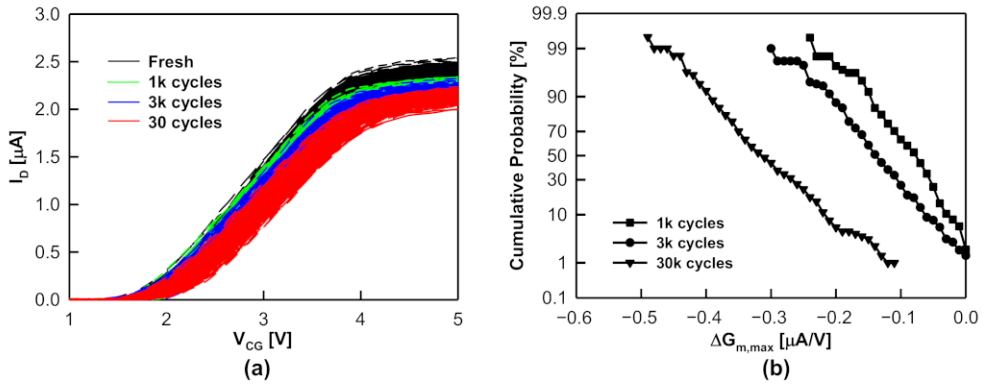

Figure 5. (**a**) $I_D - V_{CG}$ Characteristics and (**b**) cumulative $\Delta G_{m,max}$ statistics of the read cells on WL15 as a function of the number of P/E cycles. Adapted from [39].

3.2. Simulation Methodology

Monte Carlo simulations have been used in an attempt to extract information about Q_T from the measured $\Delta G_{m,max}$ distribution. A NAND string can be modeled to have a selected cell with equivalent source and drain resistances (R_S and R_D), as shown in Figure 6a. The equivalent R_S and R_D can be extracted from the monitoring of the transconductance of the read cells for different positions along the NAND string [41]. The equivalent R_S and R_D are 130 kΩ and 138.2 kΩ, respectively. The TCAD simulations used a 3D drift-diffusion equation and coupled with the Shockley–Read–Hall model for generation/recombination and mobility models (including the electric field dependence, doping-dependent modification, and surface mobility degradation). To determine the $\overline{\Delta G}_{m,max}$ statistics accurately, the simulated $I_D - V_{CG}$ characteristic of the fresh cell is calibrated with experimental data at a probability level $p = 50\%$, as shown in Figure 6b. The simulation was in good

agreement with the experimental results. After calibrating the equivalent resistances, the Monte-Carlo-based method was adopted to evaluate the concentration of Q_T (Q_T^C) after P/E cycles, as schematically shown in Figure 7. The step-by-step procedure is as follows: First, discrete Q_T is randomly generated following a uniform distribution in a cuboid volume 420 nm $\times$ 840 nm $\times$ 8 nm in size (i.e., 20 L $\times$ 10 W $\times$ T_{ox}), with an equivalent Q_T^C. Notably, the discrete Q_T is treated as a negative point charge corresponding to one electron because the electron mobility is degraded by the Coulomb repulsion.

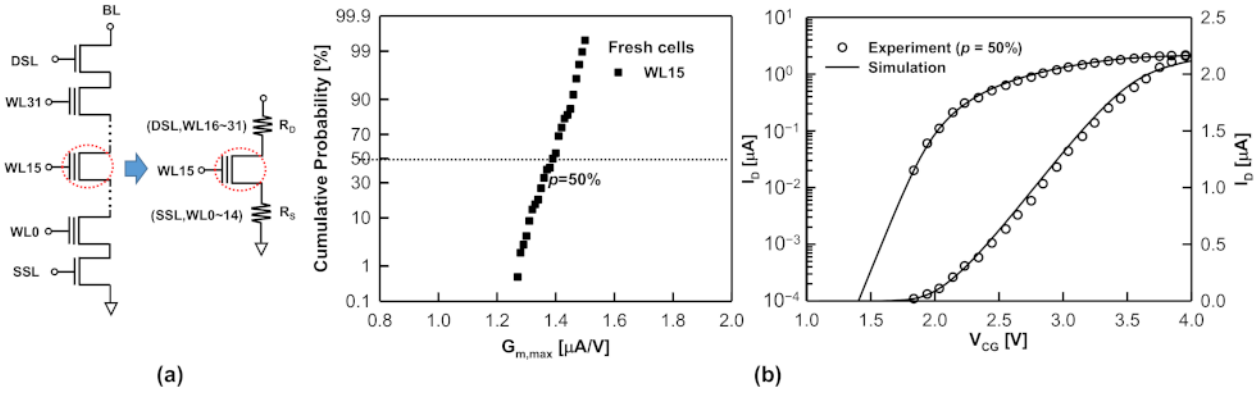

Figure 6. (**a**) Schematic circuit diagram of a NAND string and an equivalent model when cells on WL15 are read. (**b**) Comparison between measured and simulated $I_D - V_{CG}$ curve of cells on WL15 at $p = 50\%$, plotted on the linear and logarithmic scales. Adapted from [39].

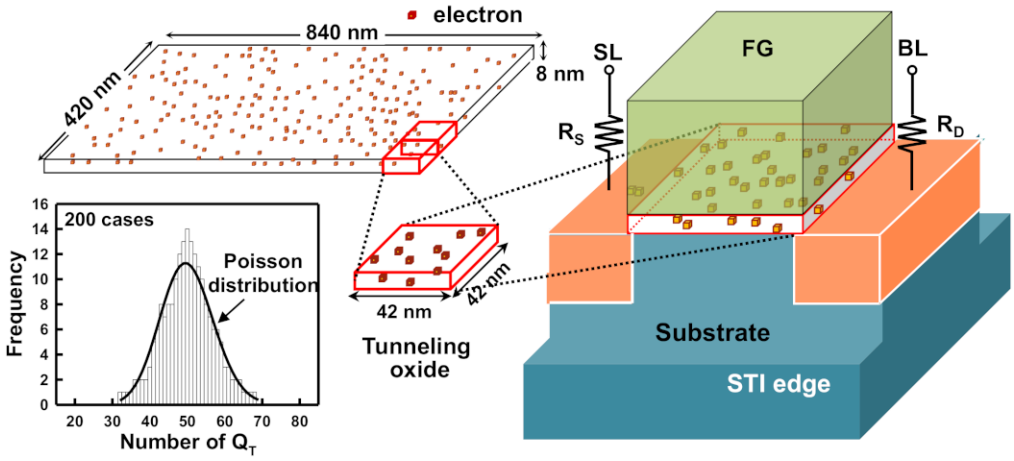

Figure 7. Schematic diagram of the random discrete Q_T generating algorithm. Adapted from [39].

Second, the cuboid is partitioned into 200 sub-cuboids and then mapped into the tunneling oxide region. Thus, the numbers of discrete Q_T in these 200 cases approximately follow a Poisson distribution, as shown in Figure 7. Finally, a comparison of the simulated and measured $\Delta G_{m,max}$ statistics allowed us to evaluate the Q_T^C during P/E cycles.

Moreover, even though the simulation does not directly account for interface trap (D_{it}) generation, the effect thereof is reflected in the model. The measured $I_D - V_{CG}$ characteristics indicated that the transconductance reached a maximum when V_{CG} slightly exceeded V_T; therefore, the occupied D_{it} can be considered as a fixed Q_T located at the silicon/oxide interface because the bending of the surface potential remains almost unchanged [42] (see Figure 8).

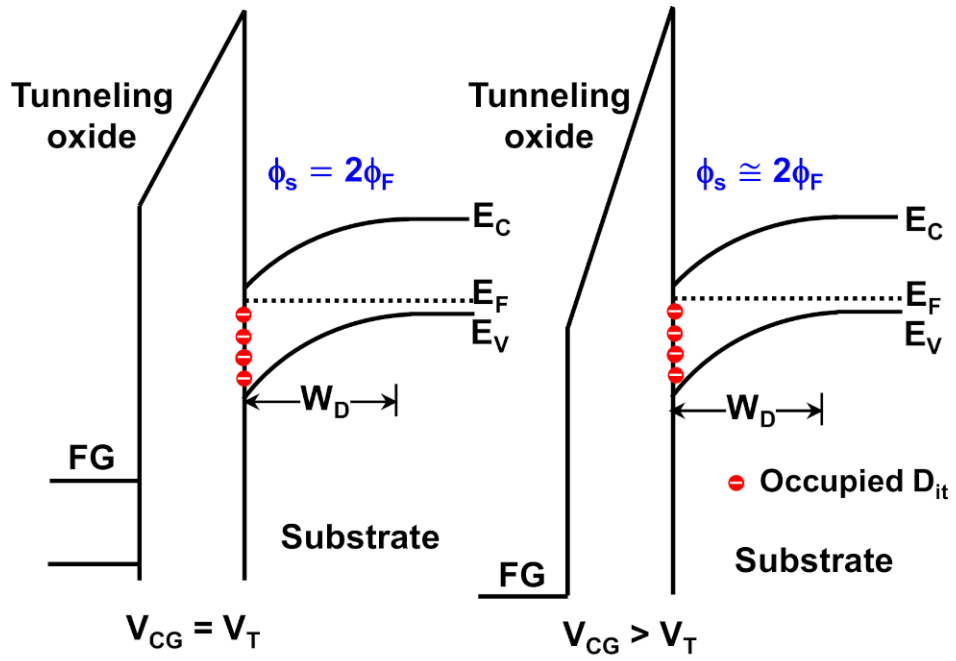

Figure 8. Band diagram and trap occupation of interface trap states at different biases. Reprinted from [39].

4. Endurance Characteristics

4.1. *QT* Extraction

Figure 9 indicates that simulations can reproduce the experimental results satisfactorily, where the extracted equivalent Q_T^C for 1 k, 10 k, and 30 k P/E cycles are 2.6×10^{18} cm^{-3}, 5×10^{18} cm^{-3}, and 1.9×10^{19} cm^{-3}, respectively. Furthermore, the proposed approach can be extended to include the array effect on $\Delta G_{m,max}$ statistics. Figure 10a shows the $\Delta G_{m,max}$ distribution as a function of the position of the WLs in a NAND string after 10 k P/E cycles. The simulations correspond well with the measurements when the following appropriate parameters are adopted: WL_1, $R_S = 16.3$ kΩ and $R_D = 251.9$ kΩ; WL_{15}, $R_S = 130$ kΩ and $R_D = 138.1$ kΩ; WL_{30}, $R_S = 251.7$ kΩ and $R_D = 16.3$ kΩ. Clearly, R_S increases as the position of the WLs changes from WL_0 to WL_{31} owing to the increase in the number of pass cells. Figure 10b shows the $\Delta G_{m,max}$ distribution as a function of V_{pass} after 10 k cycles. Again, a good agreement between the simulation and experimental results is found. The equivalent resistances of the pass cells are 8.2 kΩ/cell, 6.5 kΩ/cell, and 5.1 kΩ/cell for V_{pass} of 4 V, 5 V, and 6 V, respectively. It is clear that the pass-cell bias with a higher V_{pass} has a smaller equivalent resistance but a larger $\Delta G_{m,max}$. As a final verification, the $\Delta G_{m,max}$ statistics of two different V_T levels of the read cells were compared under the same cycling conditions. As shown in Figure 10c, the $\Delta G_{m,max}$ distributions almost overlap, indicating that Q_{FG} causes a simple parallel shift of the $I_D - V_{CG}$ curve.

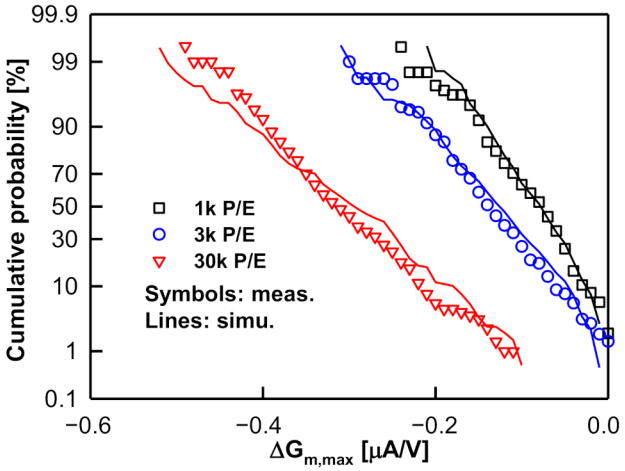

Figure 9. Simulated $\Delta G_{m,max}$ statistics for different number of P/E cycles. The simulations and experimental measurements are in good agreement. Reprinted from [39].

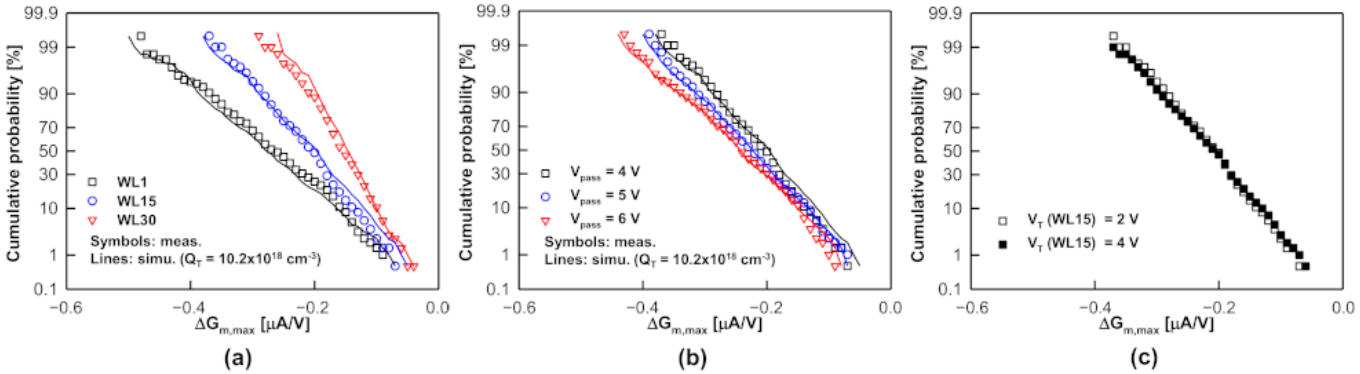

Figure 10. $\Delta G_{m,max}$ distributions for (**a**) different selected WLs in the string. All cells in the string with $V_{pass} = 6\ V$, (**b**) WL15 selected with different values of V_{pass}, and (**c**) WL15 selected with different V_T levels. Except for the read cells, the others are in the erased state. Reprinted from [39].

4.1. Endurance Degradation Model

We start this section with a description of an endurance degradation model that captures the features of the measurement. The evolution of Q_T^C can be conveniently described by the following modified power-law equation [39]:

$$Q_T^C = \frac{Q_0}{1 + (k \cdot N)^{-\alpha}} \tag{3}$$

$$k = k_0 \cdot \exp(-E_{A,G}/k_B T) \tag{4}$$

where Q_0 is the saturated value of Q_T^C, k is the reaction constant, N is the number of P/E cycles, α is the exponential coefficient, $E_{A,G}$ is the activation energy of Q_T creation, and $k_B T$ is the thermal energy. Figure 11 compares the results obtained with the endurance model and with the experimental results. The adopted parameters are as follows: $Q_0 = 1.5 \times 10^{20}$ cm^{-3}, $k = 1.0 \times 10^{-6}$, and $\alpha = 0.58$. When N is small ($\leq$ 30 k cycles), the exponential term in Equation (3) is much greater than 1, and Equation (3) can simply be expressed as a power law. On the other hand, when N is large ($>$ 30 k cycles), Q_T^C gradually approaches the saturated value Q_0. Overall, the proposed model is able to successfully describe the endurance characteristics over a wide range of N (up to 100 k cycles). Notably, Q_0 is supposed to be

related to the process condition, which determines the amount of weak Si-O or Si-H bonds that can be broken [43,44].

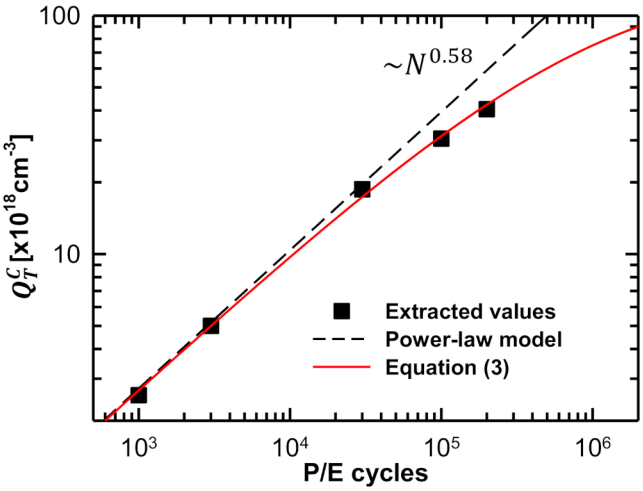

Figure 11. Comparison between extracted values (symbols) and model calculations (lines) according to Equation (3). Reprinted from [39].

To evaluate $E_{A,G}$ in Equation (4), we performed the experiments at various cycling temperatures. Figure 12 shows that $\Delta G_{m,max}$ increases as the cycling temperature (T_{cyc}) increases, suggesting that a higher T_{cyc} causes more oxide damage. The temperature-accelerated Q_T evaluations can be derived by using Equations (3) and (4) as follows [39]:

$$Q_T^C(T_{cyc}) = Q_T^C(T_R) \cdot exp\left[\alpha \cdot E_{A,G}\left(\frac{1}{k_B T_R} - \frac{1}{k_B T_H}\right)\right] \quad (5)$$

where T_R is the cycling performed at room temperature. A good linear relationship between the logarithm of Q_T^C and the reciprocal temperature was observed, and $E_{A,G}$ was evaluated, as shown in Figure 13. Ultimately, the theoretical result fitted the experimental results well, and the yield $E_{A,G}$ was approximately 100 mV, which agreed with that obtained by monitoring the stress-induced gate leakage current [13,45,46].

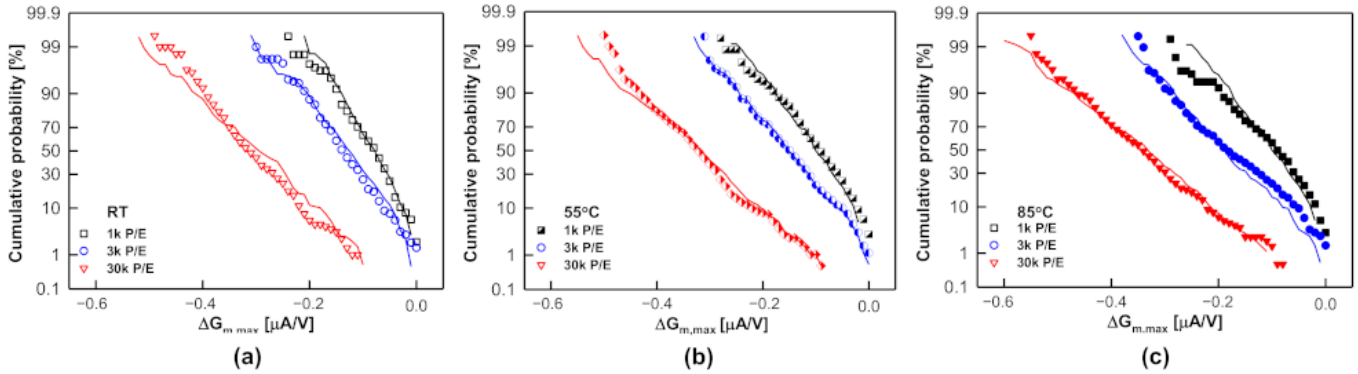

Figure 12. $\Delta G_{m,max}$ distributions for cells on WL15. Selected measured (symbols) and simulated (lines) at T_{cyc} of (**a**) 25 °C, (**b**) 55 °C, and (**c**) 85 °C are shown. Reprinted from [39].

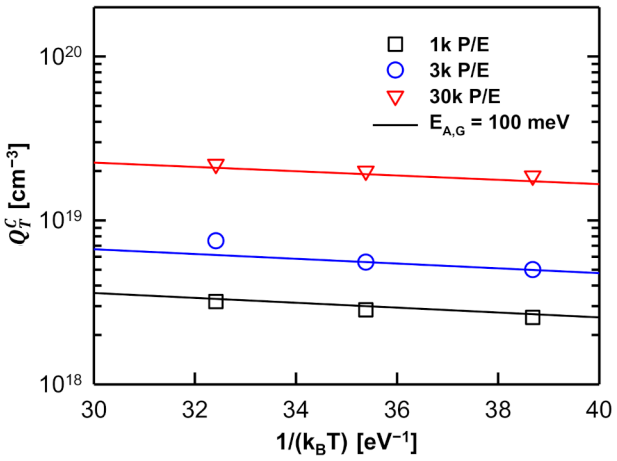

Figure 13. Q_T^C obtained by fitting experimental data using Equation (5) for different values of T_{cyc}. Reprinted from [29].

4.2. Effect of the Time Delay between P/E Cycles

Reportedly, the time delay (t_{wait}) between P/E cycles is an important factor that affects the endurance characteristics [47–51]. Figure 14 shows that the $\Delta G_{m,max}$ statistics become larger as t_{wait} increases when the endurance test is performed at T_R; however, when T_{cyc} increases to 85 °C, the trend is completely the opposite due to recovery from oxide damage through thermal excitation. This suggests that, at high T_{cyc}, the endurance model should not only take into consideration the creation of damage but also the recovery from damage. The time-dependent damage recovery during t_{wait} can be described by a rate equation given by [52,53]

$$f = exp(-t_{wait}/\tau) \tag{6}$$

$$= \tau_0 \cdot \exp\left(E_{A,R}/k_B T_{cyc}\right) \tag{7}$$

where f is the occupation function, τ is the time constant, and $E_{A,R}$ is the activation energy for the recovery from the oxide damage. Therefore, Equation (3) can be rewritten as follows [37]:

$$Q_T^C = \frac{Q_0}{1 + (k \cdot N)^{-\alpha}} \cdot f \tag{8}$$

when t_{wait} is sufficiently short, for example, 0.1 s, Equation (8) tends to Equation (3) because the mechanism according to which recovery from oxide damage takes place plays a negligible role. Accordingly, we can extract the parameters (i.e., Q_0, α, k_0, and $E_{A,G}$) by using the approach described in Section 4.1. However, when t_{wait} becomes longer, the damage creation and recovery effects are mixed, which complicates the simultaneous extraction of $E_{A,G}$ and $E_{A,R}$. To simplify the situation, we assume that under the condition of T_{cyc} = 25 °C, Equation (8) approaches Equation (3) because the thermal excitation of Q_T is not noticeable. This allows us to evaluate the $E_{A,G}$ for longer t_{wait} values of 2 s and 4 s. Figure 15a shows that, by calibrating the $E_{A,G}$ values, Equation (3) can reproduce the characteristics of the extracted Q_T^C with different t_{wait} values. The relationship between the logarithm of t_{wait} and $E_{A,G}$ is linear, as shown in Figure 15b. Moreover, $E_{A,G}$ is in the approximate range of 60–100 meV, which agrees with the results obtained by monitoring the V_T transients after experiments at different T_{cyc} [13]. Once the value of $E_{A,G}$ is determined for different values of t_{wait}, the remaining parameters $E_{A,R}$ can be determined by using the change rate of the celebrated τ at different T_{cyc}. Figure 16 shows the experimental measurements fit the curve calculated with Equation (8).

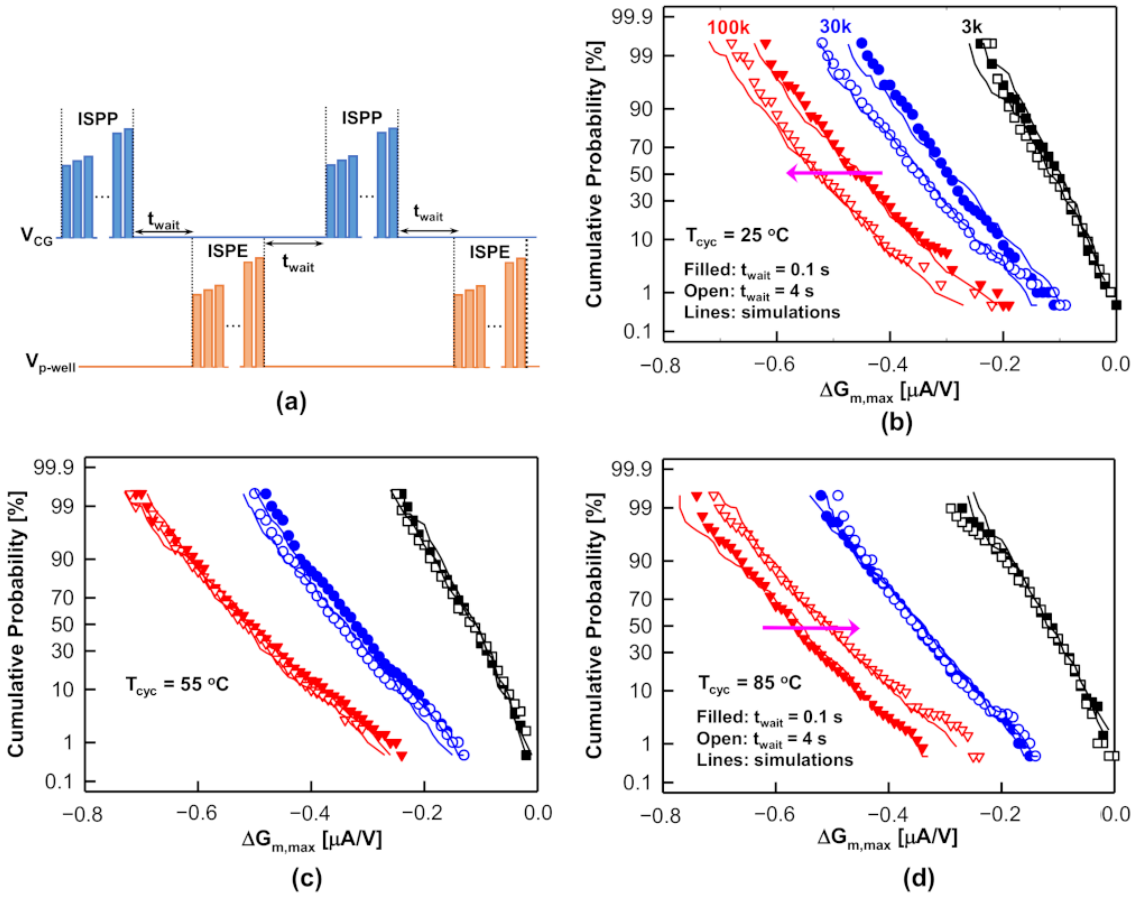

Figure 14. (**a**) Different values of t_{wait} introduced between P/E cycles. $\Delta G_{m,max}$ distributions for cells on WL15 selected measured (symbols) and simulated (lines) for different values of t_{wait} at T_{cyc} of (**b**) 25 °C, (**c**) 55 °C, and (**d**) 85 °C. Reprinted from [47].

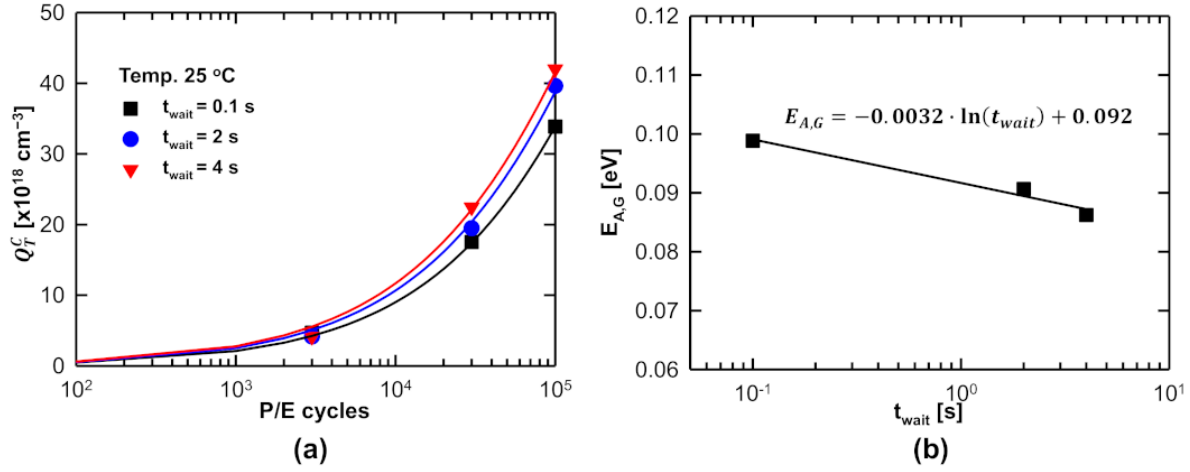

Figure 15. (**a**) Q_T^C obtained by fitting experimental data using Equation (3) for different values of t_{wait}. (**b**) Extracted $E_{A,G}$ for different values of t_{wait}. Reprinted from [47].

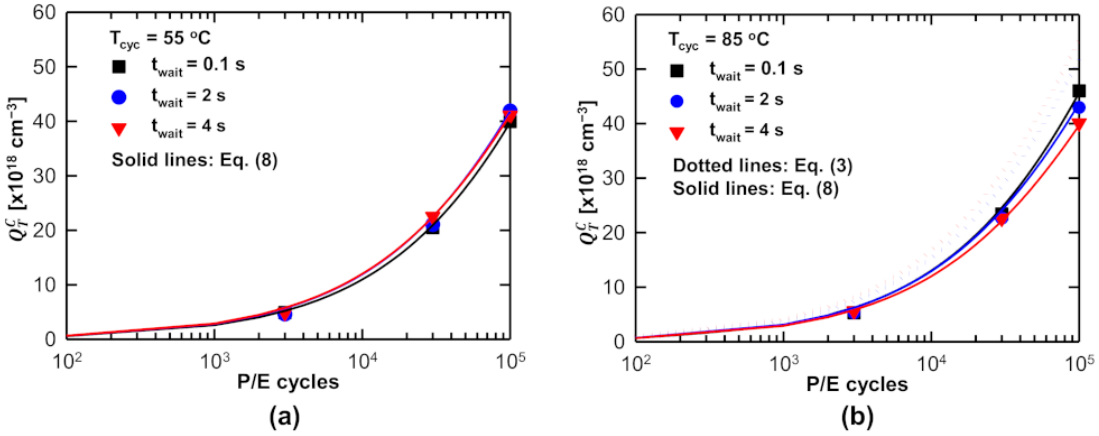

Figure 16. Q_T^C obtained by fitting experimental data using Equation (8) for different values of t_{wait} under T_{cyc} of (**a**) 55 °C and (**b**) 85 °C. Reprinted from [47].

Clearly, if the damage-recovery mechanism is not taken into account, Equation (3) overestimates the experiment, and the discrepancy between them increases with t_{wait}. The optimized τ values were 150 s, 22 s, and 12 s for T_{cyc} of 25 °C, 55 °C, and 85 °C, respectively. Figure 17 shows that the relationship between the logarithm of τ and the reciprocal temperature is linear, and $E_{A,R} \simeq 0.4\ eV$ is obtained. Notably, $E_{A,R} \simeq 0.4\ eV$ is similar to the values reported in the literature for charge detrapping through thermal emission [54]. It is also easily verifiable that the assumption that f approaches one under the condition of T_{cyc} of 25 °C and t_{wait} of 0.1 s is satisfied.

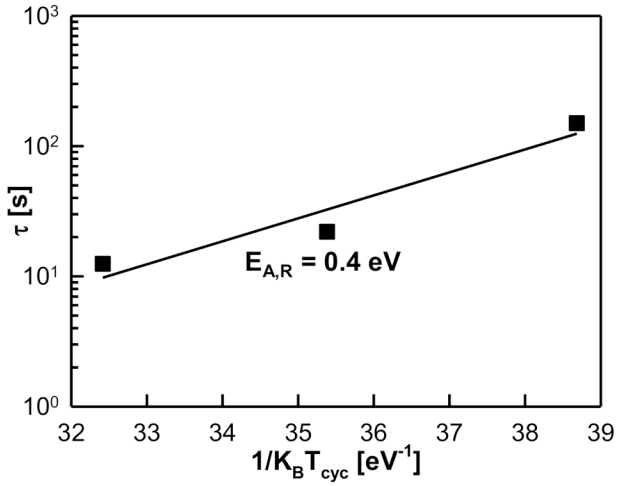

Figure 17. $E_{A,R}$ obtained using Equation (7) for different values of T_{cyc}. Reprinted from [47].

Although the above-mentioned model successfully describes the endurance characteristics, it still does not account for certain features according to more recent research [55]. A comparative analysis of the respective influence of t_{wait} from program to erase (P-to-E) and of that from erase to program (E-to-P) on the median $\Delta G_{m,max}$ ($\overline{\Delta G}_{m,max}$) after 30 k P/E cycles was reported [55] and is plotted in Figure 18, normalized to its initial value ($G_{m0,max}$). The E-to-P t_{wait} clearly had a more significant impact on the normalized $\overline{\Delta G}_{m,max}$ than P-to-E t_{wait}. Moreover, it is also shown that the normalized $\overline{\Delta G}_{m,max}$ increases as $V_{CG,0}$ increases. As a result, adopting ISPP with a lower $V_{CG,0}$ would be better for improving the oxide quality. The physical mechanism is graphically illustrated in Figure 19. Energetic electrons injected from the cathode result in anode hole injection during the P/E operations. During E-to-P t_{wait}, these holes drift near the silicon surface, where they recombine with

channel electrons [48,56]. This could create additional trap states, and subsequently, these traps are occupied by electrons.

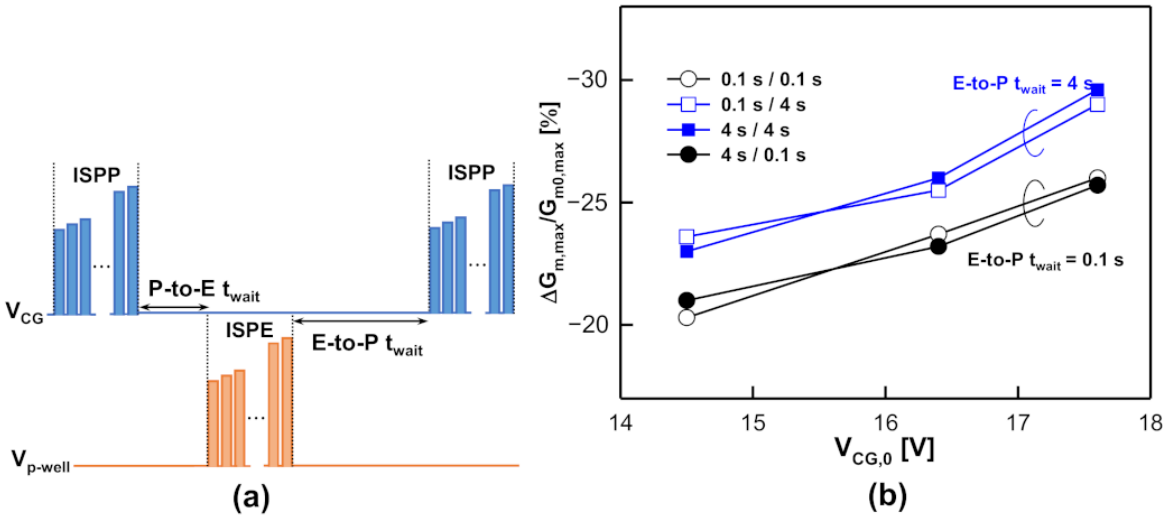

Figure 18. (**a**) Different P-to-E and E-to-P values of t_{wait} are introduced between P/E cycles. (**b**) Dependence of normalized $\Delta G_{m,max}$ on P-to-E and E-to-P values of t_{wait} after 30 k cycles. Reprinted from [55].

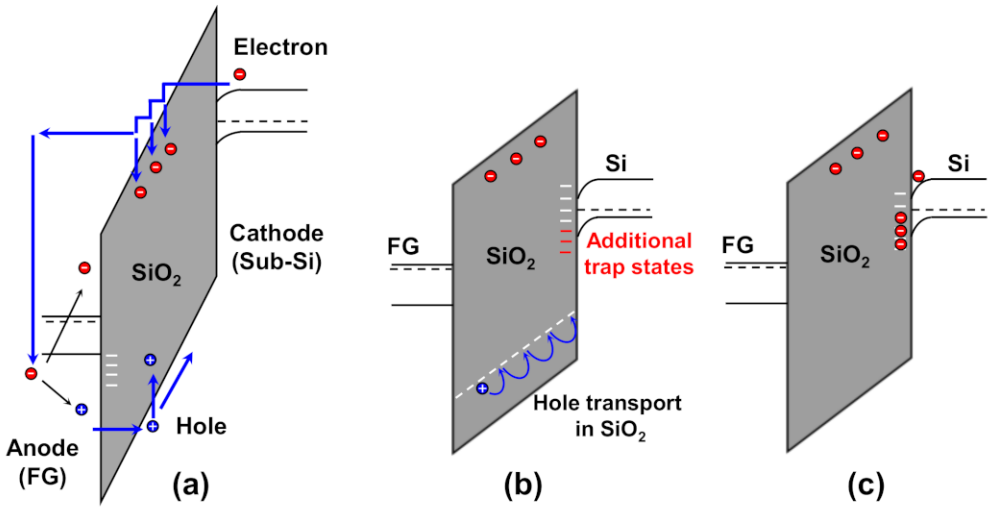

Figure 19. (**a**) Schematic diagram of anode hole injection during P/E cycles. During E-to-P t_{wait}, Q_T is generated via two possible physical mechanisms: (**b**) the holes migrate toward the silicon/oxide interface and recombine with inversion electrons, creating additional trap states. (**c**) Subsequently, electrons are trapped in these trap states. Reprinted from [55].

5. Conclusions

A method for characterizing the endurance characteristics of NAND Flash memories by monitoring the $\Delta G_{m,max}$ statistics is described. The discrete Q_T, gradually generated with P/E cycles, results in the reduction of $\Delta G_{m,max}$, and broadening of the distribution. Based on Monte Carlo simulations, an analytical model for the generation of Q_T, including the effects of T_{cyc} and t_{wait}, is then described. The model represents a powerful tool for the investigation and predictive analysis of next-generation NAND Flash technologies.

Author Contributions: Conceptualization and writing—original draft preparation, Y.-Y.C.; writing—review and editing, R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Winbond Electronics Corp. (Kanagawa, Japan).

Acknowledgments: The present work was performed with respect to the joint development agreement between Winbond Electronics Corp. and the National Yang Ming Chiao Tung University. The authors would like to thank T. Takeshita and M. Yano.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Masuoka, F.; Momodomi, M.; Iwata, Y.; Shirota, R. New ultra high density EPROM and Flash EEPROM with NAND structure cell. In Proceedings of the 1987 International Electron Devices Meeting, New York, NY, USA, 6–9 December 1987; pp. 552–555.
2. Shirota, R.; Itoh, Y.; Nakayama, R.; Momodomi, M.; Inoue, S.; Kirisawa, R.; Iwata, Y.; Chiba, M.; Masuoka, F. A new NAND cell for ultra high density 5 V-only EEPROMs. In Proceedings of the 1988 Symposium on VLSI Technology—Digest of Technical Papers, San Diego, CA, USA, 10–13 May 1988; pp. 33–34.
3. Momodomi, M.; Kirisawa, R.; Nakayama, R.; Aritome, S.; Endoh, T.; Itoh, Y.; Iwata, Y.; Oodaira, H.; Tanaka, T.; Chiba, M.; et al. New device technologies for 5 Vonly 4 Mb EEPROM with NAND structure cell. In Proceedings of the 1988 International Electron Devices Meeting, San Francisco, CA, USA, 11–14 December 1988; pp. 412–415.
4. Momodomi, M.; Itoh, Y.; Shirota, R.; Iwata, Y.; Nakayama, R.; Kirisawa, R.; Tanaka, T.; Aritome, S.; Endoh, T.; Ohuchi, K.; et al. An experimental 4-Mbit CMOS EEPROM with a NAND structure cell. *IEEE J. Solid-State Circuits* **1989**, *24*, 1238–1243. [CrossRef]
5. Iwata, Y.; Momodomi, M.; Tanaka, T.; Oodaira, H.; Itoh, Y.; Nakayama, R.; Kirisawa, R.; Aritome, S.; Endoh, T.; Shirota, R.; et al. A high-density NAND EEPROM with block-page programming for microcomputer applications. *IEEE J. Solid-State Circuits* **1990**, *25*, 417–424. [CrossRef]
6. Kirisawa, R.; Aritome, S.; Nakayama, R.; Endoh, T.; Shirota, R.; Masuoka, F. A NAND structured cell with a new programming technology for highly reliable SV-only Flash EEPROM. In Proceedings of the 1988 Symposium on VLSI Technology—Digest of Technical Papers, Honololu, HI, USA, 4–7 June 1990; pp. 129–130.
7. Aritome, S.; Shirota, R.; Kirisawa, R.; Endoh, T.; Nakayama, N.; Sakui, K.; Masuoka, F. A reliable bi-polarity write/erase technology in flash EEPROMs. In Proceedings of the 1988 International Electron Devices Meeting, San Francisco, CA, USA, 9–12 December 1990; pp. 111–114.
8. Micheloni, R.; Marelli, A.; Eshghi, K. *Inside Solid State Drives (SSDs)*; Springer: New York, NY, USA, 2013.
9. Spinelli, A.; Compagnoni, C.; Lacaita, A. Reliability of NAND Flash Memories: Planar Cells and Emerging Issues in 3D Devices. *Computers* **2017**, *6*, 16. [CrossRef]
10. Ghetti, A.; Monzio Compagnoni, C.; Spinelli, A.S.; Visconti, A. Comprehensive analysis of random telegraph noise instability and its scaling in deca-nanometer Flash memories. *IEEE Trans. Electron Devices* **2009**, *56*, 1746–1752. [CrossRef]
11. Kurata, H.; Otsuga, K.; Kotabe, A.; Kajiyama, S.; Osabe, T.; Sasago, Y.; Narumi, S.; Tokami, K.; Kamohara, S.; Tsuchiya, O. The impact of random telegraph signals on the scaling of multilevel Flash memories. In Proceedings of the 2006 Symposium on VLSI Technology (VLSI-Technology), Honolulu, HI, USA, 13–15 June 2006; pp. 112–113.
12. Tega, N.; Miki, H.; Osabe, T.; Kotabe, A.; Otsuga, K.; Kurata, H.; Kamohara, S.; Tokami, K.; Ikeda, Y.; Yamada, R. Anomalously large threshold voltage fluctuation by complex random telegraph signal in floating gate Flash memory. In Proceedings of the 2006 International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 11–13 December 2006; pp. 491–494.
13. Resnati, D.; Nicosia, G.; Paolucci, G.M.; Visconti, A.; Monzio Compagnoni, C. Cycling-induced charge trapping/detrapping in Flash memories—Part I: Experimental evidence. *IEEE Trans. Electron Devices* **2016**, *63*, 4753–4760. [CrossRef]
14. Mielke, N.; Belgal, H.; Kalastirsky, I.; Kalavade, P.; Kurtz, A.; Meng, Q.; Righos, N.; Wu, J. Flash EEPROM threshold instabilities due to charge trapping during program/erase cycling. *IEEE Trans. Device Mater. Reliab.* **2004**, *4*, 335–344. [CrossRef]
15. Paolucci, G.M.; Monzio Compagnoni, C.; Miccoli, C.; Spinelli, A.S.; Lacaita, A.L.; Visconti, A. Revisiting charge trapping/detrapping in Flash memories from a discrete and statistical standpoint—Part I: VT instabilities. *IEEE Trans. Electron Devices* **2014**, *61*, 2802–2810. [CrossRef]
16. Monzio Compagnoni, C.; Spinelli, A.S.; Gusmeroli, R.; Beltrami, S.; Ghetti, A.; Visconti, A. Ultimate accuracy for the NAND Flash program algorithm due to the electron injection statistics. *IEEE Trans. Electron Devices* **2008**, *55*, 2695–2702. [CrossRef]
17. Monzio Compagnoni, C.; Gusmeroli, R.; Spinelli, A.S.; Visconti, A. Analytical model for the electron-injection statistics during programming of nanoscale NAND Flash memories. *IEEE Trans. Electron Devices* **2008**, *55*, 3192–3199. [CrossRef]
18. Nishi, Y. (Ed.) *Advances in Non-Volatile Memory and Storage Technology*; Woodhead Publishing: Cambridge, UK, 2014.
19. Lee, J.-D.; Hur, S.-H.; Choi, J.-D. Effects of floating-gate interference on NAND flash memory cell operation. *IEEE Electron Device Lett.* **2002**, *23*, 264–266.
20. Goda, A.; Parat, K. Scaling directions for 2D and 3DNAND cells. In Proceedings of the 2012 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 10–13 December 2012.
21. Goda, A. 3-D NAND technology achievements and future scaling perspectives. *IEEE Trans. Electron Devices* **2020**, *67*, 1373–1381. [CrossRef]

22. Kim, W.; Choi, S.; Sung, J.; Lee, T.; Park, C.; Ko, H.; Jung, J.; Yoo, I.; Park, Y. Multi-layered vertical gate NAND Flash overcoming stacking limit for terabit density storage. In Proceedings of the 2009 Symposium on VLSI Technology, Kyoto, Japan, 15–17 June 2009; pp. 188–189.
23. Tanaka, H.; Kido, M.; Yahashi., K.; Oomura, M.; Katsumata, R.; Kito, M.; Fukuzumi, Y.; Sato, M.; Nagata, Y.; Matsuoka, Y.; et al. Bit cost scalable technology with punch and plug process for ultra high density flash memory. In Proceedings of the 2007 Symposium on VLSI Technology, Kyoto, Japan, 12–14 June 2007; pp. 14–15.
24. Fukuzumi, Y.; Katsumata, R.; Kito, M.; Kido, M.; Sato, M.; Tanaka, H.; Nagata, Y.; Matsuoka, Y.; Iwata, Y.; Aochi, H.; et al. Optimal integration and characteristics of vertical array devices for ultra-high density, bit-cost scalable flash memory. In Proceedings of the 2007 International Electron Devices Meeting (IEDM), Washington, DC, USA, 10–12 December 2007; pp. 449–452.
25. Jang, J.; Kim, H.-S.; Cho, W.; Cho, H.; Kim, J.; Shim, S.I.; Jang, Y.; Jeong, J.-H.; Son, B.-K.; Kim, D.-W.; et al. Vertical cell array using TCAT (terabit cell array transistor) technology for ultra high density NAND flash memory. In Proceedings of the 2009 Symposium on VLSI Technology, Kyoto, Japan, 15–17 June 2009; pp. 192–193.
26. Elliott, J.; Jung, E.S. Ushering in the 3D Memory Era with V-NAND. In Proceedings of the Flash Memory Summit, Santa Clara, CA, USA, 13–15 August 2013.
27. Park, K.T.; Byeon, D.S.; Kim, D.H. A world's first product of three-dimensional vertical NAND Flash memory and beyond. In Proceedings of the 2014 14th Annual Non-Volatile Memory Technology Symposium (NVMTS), Jeju Island, Korea, 27–29 October 2014; pp. 1–5.
28. Park, K.T.; Nam, S.; Kim, D.; Kwak, P.; Lee, D.; Choi, Y.H.; Choi, M.H.; Kwak, D.H.; Kim, D.H.; Kim, M.S.; et al. Three-Dimensional 128 Gb MLC Vertical nand Flash Memory With 24-WL Stacked Layers and 50 MB/s High-Speed Programming. *IEEE J. Solid State Circuit* **2015**, *50*, 204–213. [CrossRef]
29. Kang, D.; Kim, M.; Jeon, S.-C.; Jung, W.; Park, J.; Choo, G.; Shim, D.-K.; Kavala, A.; Kim, S.-B.; Kang, K.-M.; et al. A 512Gb 3-bit/cell 3D 6th-generation V-NAND flash memory with 82MB/s write throughput and 1.2Gb/s interface. In Proceedings of the 2019 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 17–21 February 2019; pp. 216–218.
30. Cho, J.; Chris Kang, D.; Park, J.; Nam, S.-W.; Song, J.-H.; Jung, B.-K.; Lyu, J.; Lee, H.; Kim, W.-T.; Jeon, H.; et al. A 512Gb 3b/Cell 7th-generation 3D-NAND flash memory with 184MB/s write throughput and 2.0Gb/s interface. In Proceedings of the 2021 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 13–22 February 2021; pp. 426–428.
31. Monzio Compagnoni, C.; Goda, A.; Spinelli, A.S.; Feeley, P.; Lacaita, A.L.; Visconti, A. Reviewing the evolution of the NAND Flash technology. *Proc. IEEE* **2017**, *105*, 1609–1633. [CrossRef]
32. Fayrushin, A.; Lee, C.-H.; Park, Y.; Choi, J.-H.; Chung, C. Unified endurance degradation model of floating gate NAND flash memory. *IEEE Trans. Electron Devices* **2013**, *60*, 2031–2037. [CrossRef]
33. Xia, Z.; Kim, D.S.; Jeong, N.; Kim, Y.-G.; Kim, J.-H.; Lee, K.-H.; Park, Y.-K.; Chung, C.; Lee, H.; Han, J. Comprehensive modeling of NAND flash memory reliability: Endurance and data retention. In Proceedings of the IEEE International Reliability Physics Symposium (IRPS), Anaheim, CA, USA, 15–19 April 2012; pp. MY.5.1–MY.5.4.
34. Yang, B.-J.; Wu, Y.-T.; Chiu, Y.-Y.; Kuo, T.-M.; Chang, J.-H.; Wang, P.-Y.; Shirota, R. Evaluation of the role of deep trap state using analytical model in the program/erase cycling of NAND flash memory and its process dependence. *IEEE Trans. Electron Devices* **2018**, *65*, 499–506. [CrossRef]
35. Watanabe, H.; Yao, K.; Lin, J. Numerical study of very small floating islands. *IEEE Trans. Electron. Devices* **2014**, *61*, 1145–1152. [CrossRef]
36. Lin, P.-J.-J.; Lee, C.-A.-A.; Yao, C.-W.-K.; Lin, H.-J.-V.; Watanabe, H. Localized tunneling phenomena of nanometer scaled high-K gate-stack. *IEEE Trans. Electron Devices* **2017**, *64*, 3077–3083. [CrossRef]
37. Shirota, R.; Yang, B.-J.; Chiu, Y.-Y.; Chen, H.-T.; Ng, S.-F.; Wang, P.-Y.; Chang, J.-H.; Kurachi, I. New accurate method to analyze both floating gate charge and tunnel oxide trapped charge profile in NAND flash memory. In Proceedings of the IEEE International Memory Workshop (IMW), Taipei, Taiwan, 18–21 May 2014; pp. 55–58.
38. Shirota, R.; Yang, B.-J.; Chiu, Y.-Y.; Chen, H.-T.; Ng, S.-F.; Wang, P.-Y.; Chang, J.-H.; Kurachi, I. New method to analyze the shift of floating gate charge and generated tunnel oxide trapped charge profile in NAND flash memory by program/erase endurance. *IEEE Trans. Electron Devices* **2015**, *62*, 114–120. [CrossRef]
39. Chiu, Y.-Y.; Lin, I.-C.; Chang, K.-C.; Yang, B.-J.; Takeshita, T.; Yano, M.; Shirota, R. Transconductance distribution in program/erase cycling of NAND flash memory devices: A Statistical Investigation. *IEEE Trans. Electron Devices* **2019**, *66*, 1255–1261. [CrossRef]
40. Hemink, G.J.; Tanaka, T.; Endoh, T.; Aritome, S.; Shirota, R. Fast and accurate programming method for multi-level NAND EEPROMs. In Proceedings of the 1995 Symposium on VLSI Technology—Digest of Technical Papers, Kyoto, Japan, 6–8 June 1995; pp. 129–130.
41. Joe, S.-M.; Yi, J.-H.; Park, S.-K.; Kwon, H.-I.; Lee, J.-H. Position–dependent threshold–voltage variation by random telegraph noise in NAND flash memory strings. *IEEE Trans. Electron Devices* **2010**, *31*, 635–637.
42. van Langevelde, R.; Klaassen, F.M. An explicit surface-potentialbased MOSFET model for circuit simulation. *Solid-State Electron.* **2000**, *44*, 409–418. [CrossRef]
43. Walters, M.; Reisman, A. Radiation-induced neutral electron trap generation in electrically biased insulated gate field effect transistor gate insulators. *J. Electrochem. Soc.* **1991**, *138*, 2756–2762. [CrossRef]
44. Nicklaw, C.J.; Lu, Z.-Y.; Fleetwood, D.M.; Schrimpf, R.D.; Pantelides, S.T. The structure, properties, and dynamics of oxygen vacancies in amorphous SiO_2. *IEEE Trans. Nucl. Sci.* **2002**, *49*, 2667–2673. [CrossRef]

45. Satake, H.; Toriumi, A. Common origin for stress-induced leakage current and electron trap generation in SiO_2. *Appl. Phys. Lett.* **1995**, *67*, 3489–3490. [CrossRef]
46. DiMaria, D.J.; Stasiak, J.W. Trap creation in silicon dioxide produced by hot electrons. *J. Appl. Phys.* **1989**, *65*, 2342–2356. [CrossRef]
47. Chiu, Y.-Y.; Chang, K.-C.; Lin, H.-J.; Tsai, H.-T.-E.; Lin, P.-J.; Li, H.-C.; Takeshita, T.; Yano, M.; Shirota, R. Impact of program/erase Cycling Interval on the transconductance distribution of NAND flash memory devices. *IEEE Trans. Electron Devices* **2020**, *67*, 4897–4903. [CrossRef]
48. Shirota, R.; Yang, B.-J.; Chiu, Y.-Y.; Wu, Y.-T.; Wang, P.-Y.; Chang, J.-H.; Yano, M.; Aoki, M.; Takeshita, T.; Wang, P.-Y.; et al. Improvement of oxide reliability in NAND flash memories using tight endurance cycling with shorter idling period. In Proceedings of the IEEE International Reliability Physics Symposium (IRPS), Monterey, CA, USA, 19–23 April 2015; pp. MY.12.1–MY.12.5.
49. Miccoli, C.; Monzio Compagnoni, C.; Beltrami, S.; Spinelli, A.S.; Visconti, A. Threshold-voltage instability due to damage recovery in nanoscale NAND flash memories. *IEEE Trans. Electron Devices* **2011**, *58*, 2406–2414. [CrossRef]
50. Lee, M.C.; Wong, H.Y. Investigation on the impact of program/erase cycling frequency on data retention of nanoscale charge trap nonvolatile memory. *IEEE Electron Device Lett.* **2014**, *35*, 918–920.
51. Mielke, N.; Belgal, H.P.; Fazio, A.; Meng, Q.; Righos, N. Recovery effects in the distributed cycling of flash memories. In Proceedings of the IEEE International Reliability Physics Symposium (IRPS), San Jose, CA, USA, 26–30 March 2006; pp. 29–35.
52. Wang, Y.; White, M.H. An analytical retention model for SONOS nonvolatile memory devices in the excess electron state. *Solid-State Electron.* **2005**, *49*, 97–107. [CrossRef]
53. Chiu, Y.-Y.; Yang, L.F.-H.; Chang, R.-W.; Sun, W.-T.; Lo, C.-Y.; Hsu, C.-J.; Kuo, C.-W.; Shirota, R. Characterization of the charge trapping properties in pchannel silicon–oxide–nitride–oxide–silicon memory devices including SiO_2/Si_3N_4 interfacial transition layer. *Jpn. J. Appl. Phys.* **2015**, *54*, 104201.1–104201.6. [CrossRef]
54. Thompson, S.E.; Nishida, T. Tunneling and thermal emission of electrons from a distribution of shallow traps in SiO_2. *Appl. Phys. Lett.* **1991**, *58*, 1262–1264. [CrossRef]
55. Chiu, Y.-Y.; Tsai, H.-T.-E.; Chang, K.-C.; Kumari, R.; Li, H.-C.; Takeshita, T.; Yano, M.; Shirota, R. The origin of oxide degradation during time interval between program/erase cycles in NAND Flash memory devices. *Jpn. J. Appl. Phys.* **2021**, *60*, 074004.1–074004.5. [CrossRef]
56. Lai, S.K. Interface trap generation in silicon dioxide when electrons are captured by trapped holes. *J. Appl. Phys.* **1983**, *54*, 2540–2546. [CrossRef]