The Effect of Thermal Annealing on Stress-Induced Leakage Current in Gate Oxides

25th June, 2021

Alexander Keizer

***Abstract*—Stress-induced leakage current caused by electrical stress and the recovery due to annealing has been investigated in MOS capacitors present in flash memories. The thermal annealing process in air shows exponential recovery behaviour with respect to time. Additionally, the process exhibits Arrhenius behaviour, such that the rate of reaction of the annealing process is proportional to the annealing temperature. An average activation energy of 0.065 eV was calculated. However, significant recovery due to annealing is not expected at typical operating temperatures of flash memories.**

1 Introduction

When gate oxides in flash memories are electrically stressed, the oxide layer is damaged and the defects created will result in a phenomenon known as stress-induced leakage current (SILC). This leakage current is an issue because the charge trapped in the gate oxide needs to be contained in order for the flash memory to operate normally. The defects, also referred to as traps within literature, can be detrapped by thermal annealing. From previous research, it is known how the traps within the oxide form and what the most likely conduction mechanism is for SILC. Furthermore, the damage created by electrical stressing can be fully reversed using thermal annealing at temperatures of 250 °C [1] [2]. However, it is still not known how the SILC current decreases with respect to anneal time during the anneal phase. In addition, it is unclear whether a functional relation between the rate of recovery and anneal temperature exists. If such a relation would exist, it would be possible to create a reliability model that can predict the lifetime of a MOS device.

Before SILC was discovered, the earliest reported failure mode of the gate oxide present in flash memory was considered to be oxide breakdown [3]. In oxide breakdown, also referred to as time-dependent dielectric breakdown (TDDB), a conducting path is created as a result of enough defects being formed in the oxide layer due to stress. The electron tunneling current occurs when the gate oxide is operated at or above the specified maximum operating voltage. The difference between SILC and oxide breakdown is that the current increase as a result from oxide breakdown is several orders of magnitude greater compared to the increase seen with SILC. Furthermore, it is currently not known whether the damage from oxide breakdown can be repaired, or whether it is an irreversible process.

The phenomenon of interest (SILC) occurs before oxide breakdown, which means that as the oxide is electrically stressed, defects are created in the oxide which allow for a leakage current to occur. It is known that SILC is not distributed evenly throughout the I-V curve of the MOS capacitor that has been stressed. At lower voltages, relatively more SILC can be observed than at higher voltages [4] [5] [6] [7]. This behaviour can be observed in Fig. 2.

The mechanism through which SILC occurs is most commonly thought to be the trap assisted tunneling (TAT) mechanism. The TAT mechanism proposes that charge carriers tunnel from trap-to-trap in the oxide barrier in a field assisted way [8] [9]. The motivation behind using thermal annealing is to lower the density of the traps in the oxide such that the leakage current decreases.

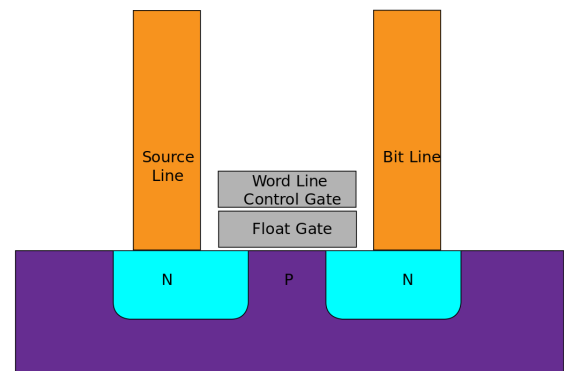

Fig. 1. Schematic cross-section of a flash memory cell [10]. The charge is stored in the floating gate structure [11]

Fig. 1 shows the schematic cross-section of a flash memory cell. In order to store data, a charge is stored in the floating gate structure by applying a voltage difference between the control gate and the substrate [11]. Through successive program/erase cycles, charges flow through the tunnel oxide between the floating gate structure and the substrate. If enough program/erase cycles have been completed, defects start to form in the tunnel oxide. It is through these defects that electrons can pass through the oxide at lower voltages through the TAT mechanism [8] [9].

Thermal annealing of the device is known to be able to repair the damage done to the gate oxide by electrical stress. The first goal of the research in this paper is to investigate how the gate current decreases as a function of time during the thermal anneal phase. The second goal of the research is to investigate whether the characteristic time of the recovery process is temperature dependent and whether the process follows Arrhenius behaviour.
If the gate current is decreasing exponentially as a function of time, the next point of interest is what experimental conditions influence the time constant during the exponential decay. The reason why the decrease in SILC is thought to be exponential is because detrapping is a physical process and the reaction rate of a physical process decreases exponentially with respect to time. Furthermore it is assumed that the reaction rate of the detrapping process can be modelled using the Arrhenius equation, given in equation 1 [12].

$$k = Ae^{\frac{-E_{\mathrm{a}}}{k_{\mathrm{B}}T}} \tag{1}$$

Where k is the rate constant, E_{a} is the activation energy in eV, T is the temperature in Kelvin, k_{B} is the boltzmann constant (expressed in $eV \cdot K^{-1}$) and A is a pre-exponential factor.

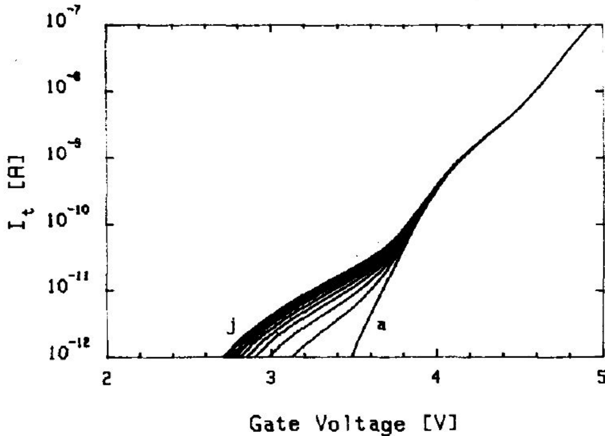

Fig. 2. I-V curves for a device before and after electrical stress induced by successive voltage ramps, with curve (a) showing the fresh device and curve (j) showing the I-V characteristic after all voltage ramps have been performed [4].

Fig. 2 shows the I-V characteristic of a stressed device compared to a fresh device, measured by P. Olivo et al. [4]. As the device is electrically stressed, defects are created which contribute to an increase in the gate current. Fig. 2 illustrates this, as the I-V curves shift to the left. At relatively lower voltages, ($2.8V - 3.2$ V) more SILC can be measured than at higher gate voltages (> 3.5V). The fact that relatively larger SILC can be measured at lower voltages is an issue for reliability in flash memories. This is because the charge contained in a flash memory cell needs to be contained sufficiently such that no data is lost from a programmed cell. However, if the device is stressed sufficiently and the induced leakage current is large enough, then a programmed cell can erroneously be measured as an erased cell at the same voltage conditions as before, resulting in data corruption.

2 Materials and Methods

The devices studied were MOS capacitors with a polysilicon area of 0.16 mm^2, and an oxide thickness of 6 nm [13]. For the measurements, a PM300 with a Keithley 4200-SCS were used. Measurements were performed in quiet mode, which means a longer integration time is used. This was done to ensure the measurements were as accurate as possible. Additionally, the sweep delay between voltage measurements was 5 seconds. The motivation for this is that if the sweep delay is shorter, then the measurements become noisy, probably because the high capacitive load leads to a long transient in the system. A full sweep from 0.0-5.5 V takes approximately 10 minutes in quiet mode. In order to reduce measurement times and prevent further stressing of the device during the measurements, a sweep was performed between 3.5-5.5 V, which reduced the measurement time to 3 minutes.

In order to measure the effect of thermal annealing, first an I-V measurement was performed on a fresh device. Then the device was stressed by applying a voltage of 6.5 V between the gate and drain of the device for 330 seconds. The stress voltage of 6.5 V was chosen based on preliminary tests. If the stress voltage is too low, then no significant SILC could be measured. If the stress voltage is too high, then the oxide layer can suffer from TDDB. An I-V measurement would be done after the device was stressed in order to gauge how much SILC was generated by the stress. Afterwards, the device was annealed at a set temperature using a hotplate. An external thermometer was used to ensure that the surface of the hotplate was at the desired anneal temperature. The device was annealed for a total of 4 hours, with I-V measurements happening regularly during the 4 hour anneal period at 25 °C. Because the expected gate current versus time behaviour was exponential, more I-V measurements were done at the start of the anneal process.

During preliminary measurements, the device was heated to the anneal temperature using the thermochuck of the PM300. This way, it would be possible to measure the current at the same temperature as the anneal temperature. However, the annealing effect was only measurable at lower temperatures (25 °C), which is the motivation for the setup mentioned.

3 Results and Discussion

The I-V curve for a device annealed at 230 °C is shown in Fig. 3. The post anneal curve shown is after 4 hours of annealing. The I-V characteristic is very similar to the I-V curve shown in Fig. 2.

As mentioned previously, regular measurements were performed during the anneal period to observe how the leakage current decreases over time. The results of one of these measurements is shown Fig. 4, where a device was annealed at 230 °C. Since the x-axis in the plot is plotted in a logarithmic scale, it is clear that the current decreases exponentially with respect to time during the anneal process. It should be noted that in both Fig. 4 and Fig. 5 the current measured is the difference between the stressed device and the fresh device.

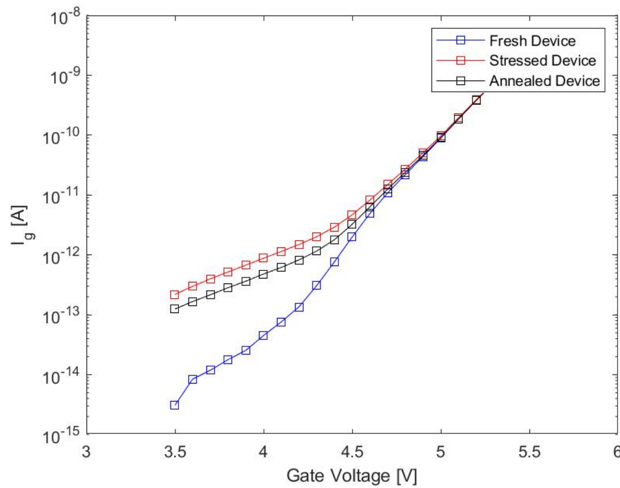

Fig. 3. I-V characteristic of the MOS capacitor before and after stress, with the post anneal curve as well. Annealing was performed at $T = 230$ °C over a period of 4 hours. The device was stressed at 6.5 V for 330 seconds at 25 °C. Measurements were performed at $T = 25$ °C

So if the current plotted is zero, the stressed device behaves exactly like the fresh device.

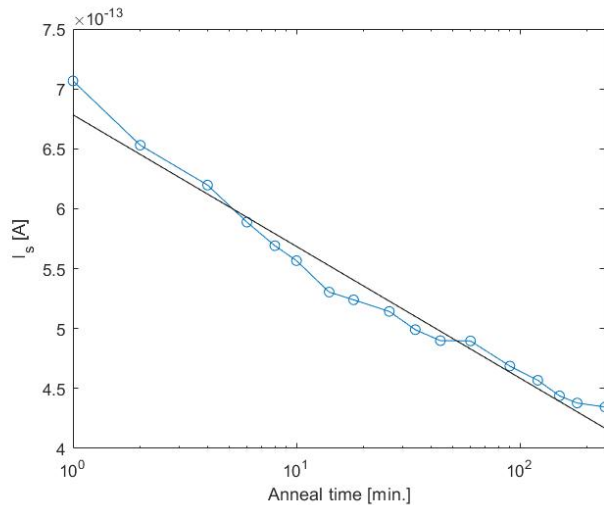

Fig. 4. SILC versus anneal time, $T_a = 230$ °C, with a total anneal time of 4 hours. SILC is measured at 4.0 V. The solid black line shows the line of best fit used to find the rate of reaction.

Similar behaviour can also be seen at lower annealing temperatures, as shown in Fig. 5. In Fig. 4 and 5 a line of best fit is drawn, the slope of which can be used to determine the rate of reaction of the annealing process. One difference observed between different temperatures is that the rate of reaction k is not the same for all temperatures. This is to be expected if it is assumed that the annealing process follows the Arrhenius equation as given in equation 1.

Knowing this, it is possible to calculate the activation energy in equation 1. If the natural logarithm is taken on both sides, an equation of the form $y = mx + b$ is found, where $x = 1/T$ and $y = ln(k)$. This is shown in equation 2. The slope of the straight line created from this plot will be equal to $\frac{-E_{\mathrm{a}}}{k_{\mathrm{B}}}$. This equation is plotted in Fig. 6 for a gate-drain voltage of $V_{\mathrm{gd}} = 4V$ using the rates of reaction measured at 5 different temperatures.

$$ln(k) = \frac{-E_a}{k_B}\frac{1}{T} + ln(A) \tag{2}$$

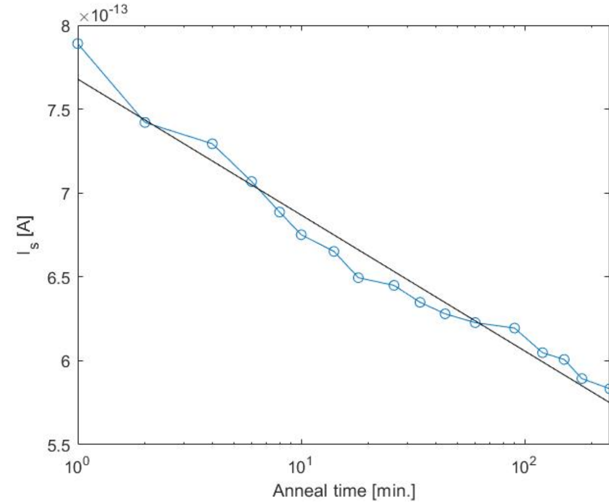

Fig. 5. SILC versus anneal time, $T_{\mathrm{a}} = 170$ °C, with a total anneal time of 4 hours. SILC is measured at 4.0 V. The solid black line shows the line of best fit used to find the rate of reaction.

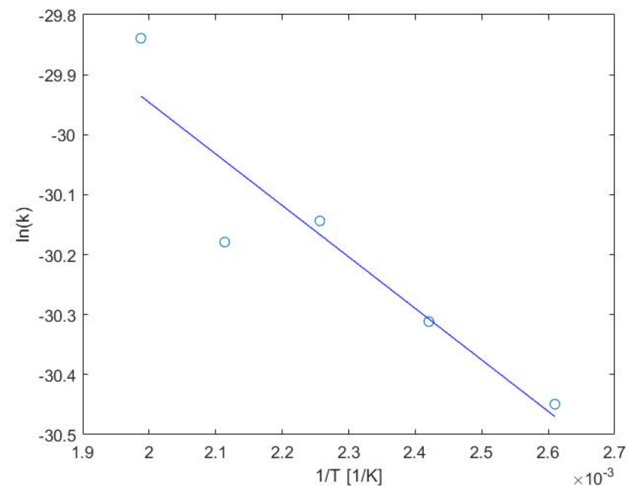

Fig. 6. Rate of reaction plotted against temperature, in the form $ln(k)$ versus $\frac{1}{T}$, for a gate-drain voltage of $V_{\mathrm{gd}} = 4$ V. The solid blue line shows the line of best fit used to find the activation energy E_{a}.

Fig. 6 shows a downward sloping best fit curve, with the measured data points plotted as well. The graph indicates that the rate of reaction decreases as the temperature decreases, which can be described by Arrhenius behaviour. As mentioned earlier, the slope of the best fit curve is equal to $-\frac{E_{\mathrm{a}}}{k_{\mathrm{B}}}$. Calculating the activation energy from Fig. 5 yields a value of 0.074 eV. For high measurement voltages (> 4.5 V), the activation energy drops dramatically and becomes negative. This means that as the temperature increases that the rate of reaction does not increase, or even decreases, if

the activation energy is negative.

Because the entire I-V curve was measured during the annealing process, it is also possible to plot the activation energy as a function of the gate voltage. This is done in Fig. 7. The dotted line plots the average activation energy in the range $3.5 - 4.5$ V, which has a value of 0.065 eV.

For the process in question this is a relatively low value, when compared to similar processes. One such a degradation process is oxide-trap charge generation due to negative bias temperature instability (NBTI) stress conditions present in pMOSFETs. In this process, similar to SILC, defects are generated in the oxide of a pMOSFET due to electrical stress in certain conditions. For this process however, an activation energy of around 0.2 eV has been found in the literature [14] [15]. This seems to suggest that the type of traps being annealed in the case of SILC are different in nature than the traps encountered in the case of NBTI degradation in pMOSFETs.

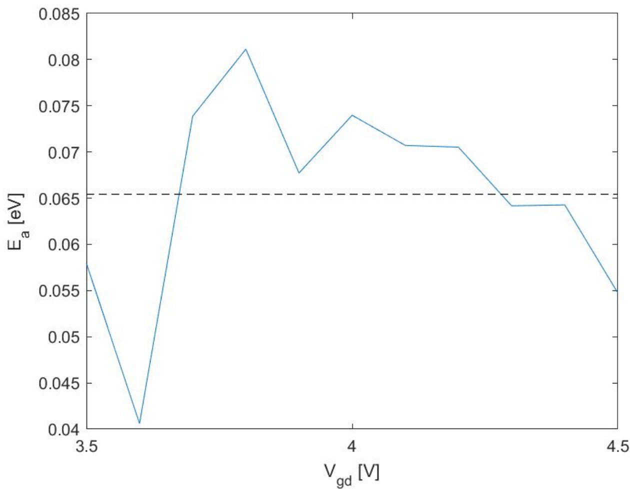

Fig. 7. Activation energy E_{a} against gate-drain voltage V_{gd}. The dotted line plots the average value of the activation energy in the plotted range.

In equation 1, there is still one constant not yet calculated; the pre-exponential factor A. Like the activation energy, this value can be calculated from Fig. 5, where $ln(A)$ is equal to the y-intercept of the fitted curve. However, unlike the activation energy, the factor A is dependent on the gate voltage.

The pre-exponential factor A has been plotted against gate-drain voltage V_{gd} in Fig. 8. From the plot, it is clear that there is a linear relationship between this factor and the gate-drain voltage. By extension, that means that there is a relation between the rate of reaction k and the gate-drain voltage V_{gd} applied to the MOS capacitor. For the voltage range in which SILC is present, the rate of reaction k increases as the gate-drain voltage increases.

Within literature, the pre-exponential factor A is referred to as the attempt frequency of the reaction. Since a clear

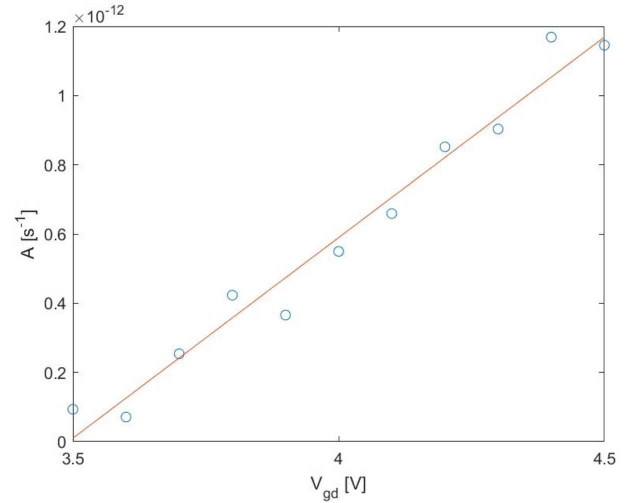

Fig. 8. Pre-exponential factor A plotted against the gate-drain voltage V_{gd} for the voltage range $3.5 - 4.5$ V.

linear relationship is found between the gate-drain voltage V_{gd} and the factor A, the data seems to suggest that the recovery behaviour is not constant over the whole I-V curve. This is in turn, would indicate that at different gate-drain voltages, different traps are being annealed. However, the near-constant activation energy seems to contradict this observation, suggesting that throughout the gate-drain voltage range of $3.5 - 4.5$ V the same traps are being annealed.

It is also possible to calculate the activation energy using a fixed pre-exponential factor A, allowing the activation energy to fluctuate as the gate-drain voltage varies. It was calculated that the activation energy varies between $0.01 - 0.06$ eV as V_{gd} varies between $3.5 - 4.5$ V. Such a strong fluctuation seems to suggest that using the current parameter set (A and E_{a}) is insufficient to fit the measured data. As a result, it is possible to conclude that the Arrhenius model, though useful, is not enough to completely describe the annealing process over time.

3.1 Extrapolation

Because both the activation energy and the pre-exponential factor in equation 1 have been determined, it is also possible to extrapolate what value the rate of reaction would have at lower temperatures, such as 25 °C or 50 °C. Because the rate of reaction is also dependent on the gate voltage, it is possible to plot the rate of reaction k against the gate voltage V_{g}.

Fig. 9 plots the rate of reaction for 5 different temperatures, two of which are extrapolated data points. Unsurprisingly, the lower temperatures show a lower rate of reaction than the higher temperatures. However, the rate of reaction of the extrapolated temperatures does not differ too much from the measured data when the gate voltage is around 4.0 V. The extrapolated rates of reaction differ more from the measured rates of reaction as the gate voltage either decreases to 3.5 V or increases to 4.5 V.

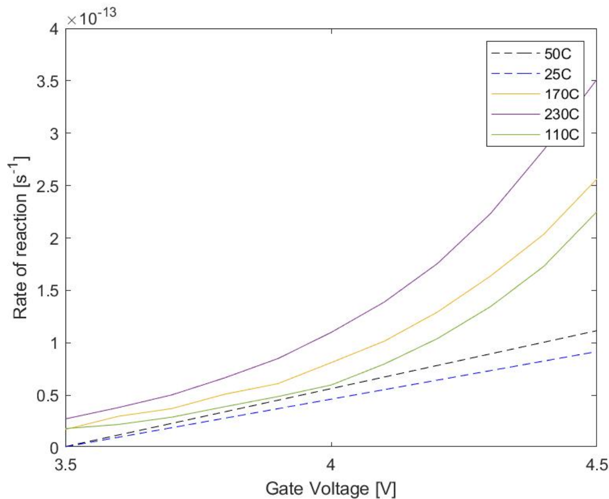

Fig. 9. Rate of reaction plotted against gate voltage for different temperatures. Dotted lines illustrate extrapolated data.

Now it is possible to predict from the extrapolated rates of reaction in Fig. 9 how SILC decreases over time at lower temperatures. With this, it is assumed that at lower temperatures the annealing process follows a similar exponential relationship with respect to time, as observed in Fig. 5 and 4.

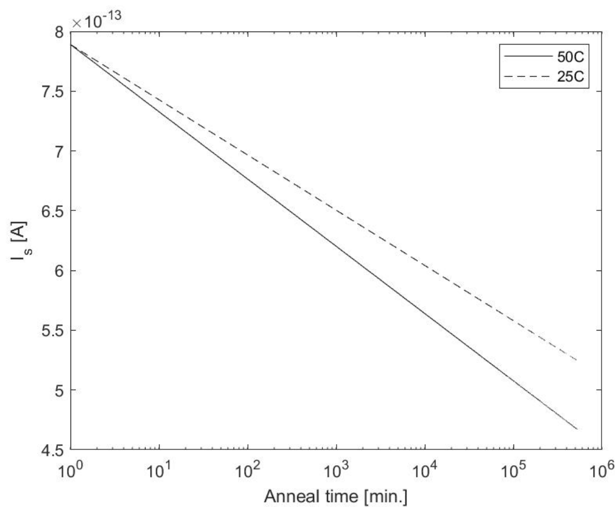

Fig. 10. Predicted SILC recovery plotted against anneal time. The initial value at $t = 0$ is taken from earlier measurements. Gate-drain voltage $V_{gd} = 4V$.

Fig. 10 plots the calculated SILC against time at a gate voltage of $V_{\mathrm{g}} = 4\ V$ for 25 $°C$ and 50 $°C$. Similar to Fig. 5 and 4, the current shows clear exponential behaviour. Fig. 10 extrapolates SILC for an anneal time of 1 year, to gain some insight into the long term behaviour of the annealing process. From this figure, it is also possible to calculate the time constant of the annealing process at these low temperatures to see whether self-healing is realistic in a typical setting (T=25-50°C). The time constant is the time it takes for the SILC measured to decay to 36.8% of its initial value. Calculating the time constant for $T_a = 50°C$ yields a time constant of $\tau = 1000$ years. As a result, it is unrealistic for significant annealing to occur at operating temperatures within a timeframe of 10 years.

4 Conclusion

The recovery of SILC over time through thermal annealing has been observed at different temperatures. During annealing, the detrapping process shows clear exponential behaviour over time. Measurements at various temperatures resulted in an activation energy for the detrapping process. Furthermore, the rate of reaction for the detrapping process could be modelled using the Arrhenius equation.

References

[1] H.-T. Lue, P.-Y. Du, C.-P. Chen, W.-C. Chen, C.-C. Hsieh, Y.-H. Hsiao, Y.-H. Shih, and C.-Y. Lu, "Radically Extending the Cycling Endurance of Flash Memory (to $>$ 100M Cycles) by Using Built-in Thermal Annealing to Self-heal the Stress-induced Damage," *Hsinchu Science Park, Hsinchu, Taiwan.*, pp. 199–202, 2012.

[2] N. Mielke, H. P. Belgal, A. Fazio, Q. Meng, and N. Righos, "Recovery Effects in the Distributed Cycling of Flash Memories," *44th Annual International Reliability Physics Symposium*, pp. 29–35, 2006.

[3] J. Maserjian and N. Zamani, "Behavior of the Si/SiO_2 interface observed by Fowler-Nordheim tunneling," *Journal of Applied Physics*, vol. 53, pp. 559–567, 1982.

[4] P. Olivo, T. Nguyen, and B. Ricco, "High-Field-Induced Degradation in Ultra-Thin SO2 Films," *IEEE Transactions on electron devices*, vol. 35, pp. 2259–2267, 1988.

[5] G. G. et al., "Emerging Oxide Degradation Mechanisms: Stress-Induced Leakage Current (SILC) and Quasi-Breakdown," *Microelectronic Engineering*, vol. 49, pp. 41–50, 1999.

[6] R. Moazzami and C. Hu, "Stress-Induced Current in Thin Silicon Dioxide Films," *Department of Electrical Engineering and Computer Sciences, University of California, Berkeley*, 1992.

[7] F. Irrera and B. Riccò, "SILC Dynamics in MOS Structures Subject to Periodic Stress," *IEEE Transactions on Electron Devices*, vol. 49, pp. 1729–1735, 2002.

[8] V. Houtsma, "Gate Oxide Reliability of Poly-Si and Poly-SiGe CMOS devices," pp. 43–49, 1999.

[9] J. Frenkel, "On pre-breakdown phenomena in insulators and electronic semi-conductors," *American Physical Society*, October 1938.

[10] Cyferz, "Flash cell structure," Wikimedia Commons, July 2007.

[11] R. Bez, E. Camerlenghi, A. Modelli, and A. Visconti, "Introduction to flash memory," *Proceedings of the IEEE*, vol. 91, May 2003.

[12] S. Arrhenius, "Über die dissociationswärme und den einfluss der temperatur auf den dissociationsgrad der elektrolyte," *Zeitschrift für Physikalische Chemie*, vol. 4U, no. 1, pp. 96–116, 1889.

[13] J. Schmitz, A. van de Ven, B. de Wachter, J. Mees, and L. L. Cam, *Description of the MINOXG mask set*, Philips Electronics, July 2002.

[14] M. D. V. Huard and C. Parthasarathy, "NBTI degradation: From physical mechanisms to modelling," *Microelectronics Reliability*, vol. 46, no. 1, pp. 1–23, 2006.

[15] G. X. Duan, J. Hatchtel, X. Shen, E. X. Zhang, C. X. Zhang, B. R. Tuttle, D. M. Fleetwood, R. D. Schrimpf, R. A. Reed, J. Franco, D. Linten, J. Mitard, L. Witters, N. Collaert, M. Chisholm, and S. T. Pantelides, "Activation energies for oxide- and interface-trap charge generation due to negative-bias temperature stress of Si-capped SiGe-pmosfets," *IEEE Transactions on Device and Materials Reliability*, vol. 15, no. 3, pp. 352–358, September 2015.