J-6-1

Numerical Simulation of the Read Disturb Behavior on the ONO Scaling Margin in SONOS Flash Memory

C. H. Lee[1,2], C.W. Wu[1], S. W. Lin[1], T. H. Yeh[1], S. H. Gu[1], K.F. Chen[1], Y. J. Chen[1], J. Y. Hsieh[1], I. J. Huang[1], N. K. Zous[1],
T. T. Han[1], M. S. Chen[1], W. P. Lu[1], K. C. Chen[1], Tahui Wang[1,2], and C. Y. Lu[1]

[1]Macronix International Co., Ltd, No. 16, Li-Hsin Road, Science-Based Industrial Park, Hsin-Chu, Taiwan
[2]Dept. of Electronics Engineering, National Chiao-Tung University, Hsin-Chu, Taiwan
Phone: +886-3-5786688-78088 E-mail: nkzou@mxic.com.tw, shku@mxic.com.tw, twang@cc.nctu.edu.tw

1. Introduction

SONOS type devices have received great attention for their scalability and simple process for NAND applications [1,2]. Even while the program and erase speeds could be significantly improved by using a high-k blocking layer with a metal gate [3], some concerns still exist such as program saturation [4] and read disturb [5]. Previously, we have proposed a model to emulate the program transient of such devices [4]. Here, we extend the study to a lower field region, which will allow us to further understand the read disturb behavior. Moreover, the process margins can be given according to its program speed/read disturb constraints. Experiments are exercised on a p^+-gate SONOS with O/N/O thickness of 8/6/3 nm and W/L= 0.065/0.1 μm.

2. Experiments and Results

Read disturb model

In our previous study [4], the program transient has been well simulated with a trap depth of 1.65eV as shown in Fig.1. From the schematic energy band diagram in Fig.2 (a), FN tunneling is the dominant injection mechanism (Eq.(1) of Table I) due to a larger gate bias during programming operation. Contrarily, read operation, which uses a lower gate bias, may also affect a cell V_T once a longer stress time is applied and modified FN tunneling probably is the main mechanism to drive this disturbance, as depicted in Fig.2 (b). The corresponding equation in this region is listed in (3) of Table 1. Except the saturation behavior, in Fig.3, the simulated curves (solid line) fit well for the Si-data from V_g=7V to 9V with a trap depth of 1.65eV. This non-saturated phenomenon suggests that the electrons may be trapped in a deeper state of the nitride layer. To confirm this, a negative gate acceleration test, which proposed in [6], is used to compare the retention characteristics under different programming V_g. The data, which utilizes the substrate hot electron (SHE) injection technique, is plotted in Fig.4 (a) for reference. The program V_T is 4V and V_g=-7V is applied to accelerate the charge loss. Apparently, a cell programmed by V_g=9V shows the best retentivity while the SHE injected one is the worst. These results suggest that the trap depth in the nitride layer may be affected by the injected electron energy and a deeper trap state is favor for V_g ranges from 7V to 9V. Repeated experiments were executed on other cells and comparison of ΔV_T at retention time of 2000 seconds is shown in Fig.4 (b). All cells show similar result, which suggests that our observation is a common behavior. In addition, the trap energy, which was used in the simulation, was adjusted according to Fig.4 (a) and reference [7]. Finally, the disturb curve of Fig.3 is well simulated up to 10^4 seconds (dash line).

Read bias in NAND array operation

Typically, the initial V_T of a NAND cell is around zero volt and the schematic cell V_T distributions during NAND array operation is shown in Fig.5. During read operation, in order to achieve sufficient sensing current at around 1μA, a vertical surface field ~ 2.6MV/cm is necessary. This value can be obtained by using the drain current equation at linear region [8]and is consisted with [9]. This read criteria should be considered when a cell at erase state is being read and also for those programmed cells in the same bit line string, as enough read current should pass through the string. The final $V_{pass,R}$ is thus determined by considering the required read current and effective oxide thickness (EOT). Additionally, the width of the cell V_T distribution (w) should also be taken care to gain enough read margin during production. Here, we set the read margin =0.5V. Read retention characteristics measured at different $V_{pass,R}$ are shown in Fig.6. The lifetime is defined when the threshold shift of a disturbed cell reaches 2V. To meet a 10 years lifetime requirement, $V_{pass,R}$ should be lower than 5.44V (=$V^*_{pass,R}$). Fig.7 shows the relationship between the standard deviation of the cell V_T distribution and $V_{pass,R}$ for 1M bits read. Under the limitation of $V^*_{pass,R}$, the width of V_T distribution (w) must be less than 1V. This narrow cell V_T distribution will degrade the program performance [10]. If we loosen the distribution to 1.5V, $V_{pass,R}$ should be as high as 5.7V to obtain enough read current. In this case, our device will suffer a read disturb failure.

Bottom oxide thickness effect

Device tuning is necessary in order to overcome the read disturb issue when $V_{pass,R}$ =5.7V is applied. Increasing the bottom oxide thickness is one of the strategies to extend the device's lifetime. However, thicker bottom oxide thickness will also degrade programming speed. By our simulation, Fig.8 illustrates the bottom oxide thickness effect on both program degradation and disturb improvement when the EOT of the device is fixed at 14.3nm. When the criteria of 4V window in 1ms program (20V) and 2V window in 10years disturb (5.7V) are set, the allowed bottom oxide thickness range can be determined. Fig.9 shows $V_{pass,R}$ should increase with an increase in the EOT. Based on the simulation, allowed T_{OX} range versus EOT is also drawn in Fig.9. One may deduce that as EOT is being scaling down, the allowed T_{OX} range will finally vanish due to the criteria of the read disturb.

3. Conclusions

The read disturb behaviors can be accurately simulated by our proposed model. No saturation phenomenon under read operation is observed when MFN tunneling is applied. We suspect that MFN tunneling leads to a deeper trap depth than FN tunneling. $V_{pass,R}$ is found to be affected both by the standard deviation of cell V_T distribution and the effective oxide thickness of the device. Optimized range of bottom oxide thickness which guarantees both programming speed and read disturb criteria is from 4.7nm to 5.1nm under the EOT is 14.3nm.

References

[1] M. H. White et al. *IEEE Circuits and Devices*, Vol. 16, p.22, 2000.

[2] Y. park et al., *IEDM*., p.29, 2006.

[3] C. H. Lee et al., *IEDM*., p.613, 2003.

[4] C. H. Lee et al., *NVSMW*., 2008

[5] A. Furnemont et al., NVSMW. p.94, 2007.

[6] C. C. Yeh et al., *IEDM*., p.931, 2002**.**

[7] T. Wang et al., IEDM., p.169, 2003

[8] S. M. Sze ,In *Physics of Semiconductor Devices* ,John Wiley, New York, 1981, p.440

[9] J. D. Choi et al., *IEDM*., p.767, 2000.

[10] T. Cho et al., *ISSCC*., p.28, 2001.

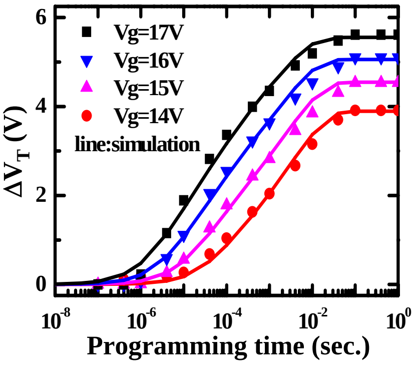

Fig.1 Experimental and simulated program transient at various program V_g. The symbols represent measured data and the lines are simulated results.

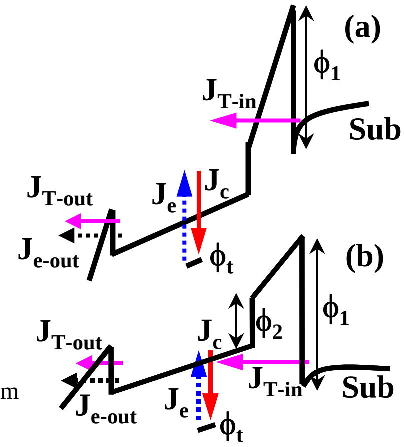

Fig.2 Schematic energy band diagram and current flows when **(a)** FN tunneling dominates; **(b)** Modified FN tunneling dominates.

Table 1 Equations for calculating tunneling current through bottom oxide or top oxide according to the electric field, **(1)** FN tunneling; **(2)** Direct tunneling; **(3)** Modified FN tunneling.

$$J_{FN} = A \cdot \exp\left(-\frac{B_1}{E_{OX}}\right) \quad \text{when } E_{OX} \geq \frac{\phi_1}{T_{OX}}, \text{where } A=\frac{q^3}{16\pi^2\hbar\phi_1}\cdot E_{OX}^2\,; B_1 = \frac{4}{3}\cdot\frac{\sqrt{2m_{OX}}}{q\hbar}\phi_1^{\ 3/2} \quad (1)$$

$$J_{DT} = A \cdot \exp\left(-\frac{B_2}{E_{OX}}\right) \quad \text{when } \frac{\phi_1-\phi_2}{T_{OX}} < E_{OX} < \frac{\phi_1}{T_{OX}}, \ \text{where } B_2 = B_1\cdot\left[1-\left(1-\frac{E_{OX}\cdot t_{OX}}{\phi_1}\right)^{3/2}\right] \quad (2)$$

$$J_{MFN} = J_{DT} \cdot \exp\left(-\frac{B_3}{E_{SiN}}\right) \text{when } E_{OX} \leq \frac{\phi_1-\phi_2}{T_{OX}}, \text{where } B_3 = \frac{4}{3}\cdot\frac{\sqrt{2m_{SiN}}}{q\hbar}\left(\phi_1-\phi_2-E_{OX}\cdot T_{OX}\right)^{3/2} \quad (3)$$

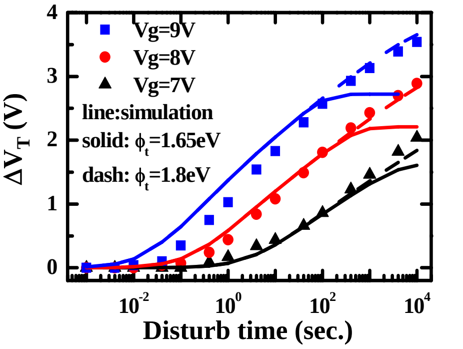

Fig.3 Measured (symbol) and simulated (line) disturb transient at various read V_g. Solid and dash lines are the results with ϕ_t=1.65eV and ϕ_t=1.8eV, respectively.

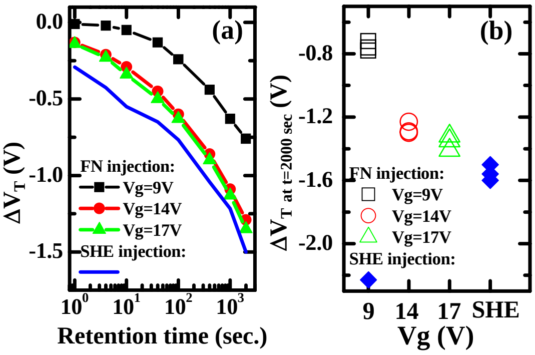

Fig.4 (a) Effect of program method on retention behavior. Program window=4V. Accelerated Vg during retention=-7V. (b) Comparison of ΔV_T at t=2000 sec. under four different conditions. The injected method by SHE shows the largest ΔV_T.

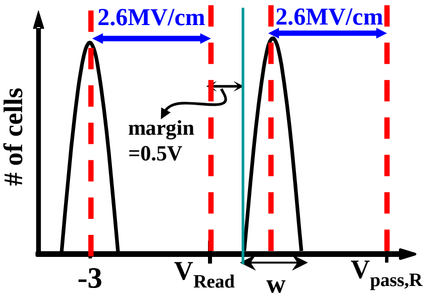

Fig.5 Schematic cell V_T distribution. Additional 2.6MV/cm with respect to programmed and erased V_T is needed for enough read current. "w" is the width of cell V_T distribution.

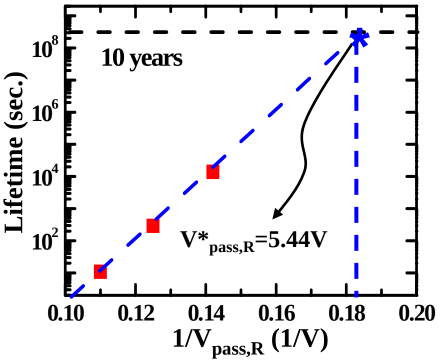

Fig.6 Read retention behavior. With ΔV_T=2V and a disturb lifetime of 10 years, $V_{pass,R}$ can be as high as 5.44V.

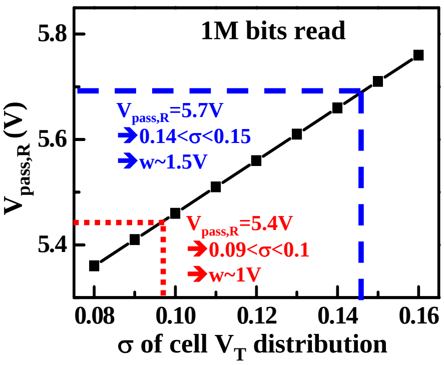

Fig.7 Relationship between $V_{pass,R}$ and standard deviation σ of cell V_T distribution. $V_{pass,R}$ is directly proportion to σ.

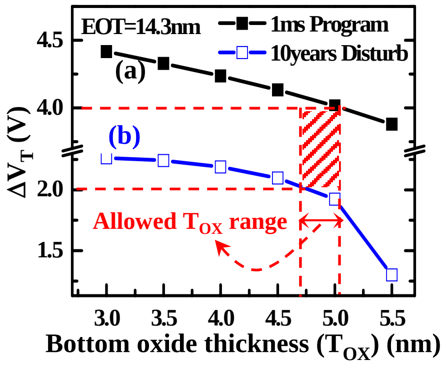

Fig.8 Simulated ΔV_T **(a)** at program time=1ms and **(b)** at disturb time=10years for different bottom oxide thicknesses (T_{OX}) but under fixed EOT. Program voltage=20V and disturb voltage=5.7V.

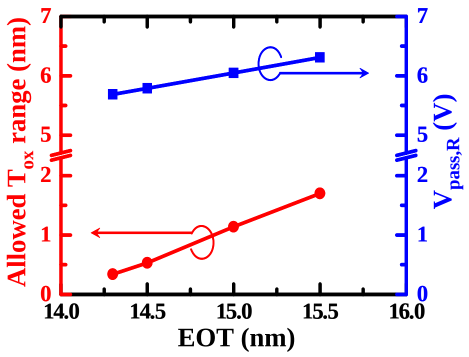

Fig.9 Allowed T_{OX} range versus EOT is drawn. Relationship between $V_{pass,R}$ and EOT is also shown.