Evaluation of program, erase and retention times of Flash memories with very thin gate dielectric

Giuseppe Iannaccone
Dipartimento di Ingegneria dell'Informazione: Elettronica, Informatica, Telecomunicazioni,
Università di Pisa

G. Iannaccone, Evaluation of program, erase and retention times of Flash memories with very thin gate dielectric, IEEE Nanotechnology 2002, p. 247, Washington DC 2002.

Evaluation of program, erase and retention times of Flash memories with very thin gate dielectric

G. Iannaccone

Dipartimento di Ingegneria dell'Informazione, Università degli studi di Pisa,
Via Diotisalvi 2, I-56122 Pisa, Italy
e-mail: g.iannaccone@iet.unipi.it

Abstract — **We have developed a code for the simulation of the complete program and erase process of a Flash EEPROM via direct of Fowler-Nordheim tunneling from the channel. The code is based on the solution of the Poisson-Schrödinger equation in one dimension, and on the computation of the tunneling current from the detailed solution of the barrier transmission problem. We are able to compute the program, erase and retention time of a flash memory structure with a given gate stack, and therefore to design an optimized layer structure with respect to the trade-off between program and retention times.**

I. INTRODUZIONE

Numerical simulation of the program and erase operations of Flash EEPROMs would enable device physicist and engineers to design the gate stack in order to optimize the trade-off between fast write-erase times and long retention times. Such a tool would allow us to evaluate the properties of more complex gate stacks with alternative dielectrics, and to explore the possibility of multi-layered tunnel barriers [1]. In this paper, we focus on thin-oxide Flash EEPROMs, in which the floating gate is charged through Fowler-Nordheim (FN) or direct tunneling. In such devices, the tunnel oxide should be thinner than in present day commercial Flash EEPROMs (8-10 nm), in order to obtain write times and operating voltages comparable to flash memories based on Channel Hot Electrons (CHE) charging [2]; on the other hand, thin oxides considerably increase leakage to the channel, so that even non-defective cells may have data retention times smaller than ten years, which is the present standard requirement.

We have developed a code for the simulation of the complete time-dependent process of charging and discharging of the floating gate in one dimension, with a fully quantum-mechanical approach, including tunneling through the barrier and quantum confinement in the channel. Then, we have used the code to perform an investigation of the process of charging and discharging of the floating gate as a function of oxide thickness, device structure, and applied voltages.

II. APPROACH

In order to describe the modeling approach using a meaningful example, let us consider a Flash EEPROM with the following structure: substrate doping $N_A = 10^{16}\,\mathrm{cm}^{-3}$, tunnel oxide t_{ox}, n-type polysilicon floating gate ($N_D = 10^{19}\,\mathrm{cm}^{-3}$), triple Oxide-Nitride-Oxide (ONO) layer (5 nm SiO_2, 10 nm Si_3N_4, 5 nm SiO_2 oxide), and n^+ poly control gate ($N_D = 10^{20}\,\mathrm{cm}^{-3}$).

We solve the coupled Poisson and Schrödinger equations in the vertical direction with the boundary conditions imposed by the applied gate voltage V_{GS}. Three separate Fermi levels are used for the substrate (E_{FS} = 0 eV), the floating gate (E_{FFG} = 0 eV), and the control gate $E_{FFG} = E_{FS} - qV_{GS}$. The Fermi-level in the floating gate E_{FFG} depends on the charge stored per unit area Q_{FG}: we will actually fix the value of E_{FFG}, and by varying it in a broad range we will find the relationship among E_{FFG}, Q_{FG}, and the current density J_T.

The quantum model (based on the solution of the Schrödinger equation with mean field approximation) is used only for electrons at the Si-SiO_2 interface (on the substrate side for the program operation, on the side of the floating gate for the erase operation. If x is the direction perpendicular to the channel, the electron density $n(x)$ is given by

$$
\begin{aligned}
n(x) = {} & \frac{2kTm_t}{\pi\hbar^2}\sum_i \left|\Psi_{il}(x)\right|^2 \ln\left[1+\exp\left(\frac{E_F - E_{il}}{kT}\right)\right] \\
& + \frac{4kT\sqrt{m_l m_t}}{\pi\hbar^2}\sum_i \left|\Psi_{it}(x)\right|^2 \ln\left[1+\exp\left(\frac{E_F - E_{it}}{kT}\right)\right]
\end{aligned}, \qquad (1)
$$

where m_l and m_t are the longitudinal and transverse effective masses in silicon, k is the Boltzmann constant, T is the temperature, Ψ_{ik} and E_{ik} ($k=l,t$) are the i-th eigenfunction and eigenvalue of the Schrödinger equation

0-7803-7538-6/02/$17.00©2002 IEEE

$$-\frac{\hbar^2}{2m_k}\nabla^2\Psi_{ik}(x)+E_C(x)\Psi_{ik}(x)=E_{il}\Psi_{ik}(x) \qquad (2)$$

Electron densities in other regions and hole densities are computed with the semiclassical approximation [3]. The Poisson and Schrödinger equations are then solved self-consistently with an iterative under-relaxation algorithm.

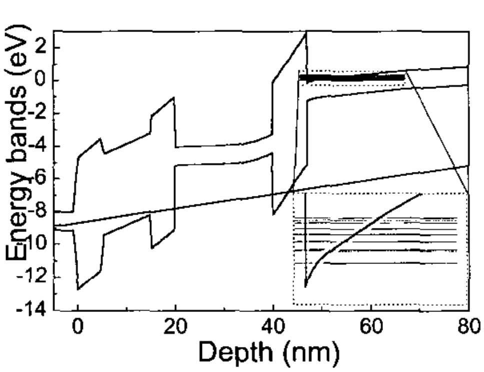

Fig. 1: Valence and conduction bands in the vertical direction for a device with t_{ox} = 10 nm and V_{GS} = 8 V. Eigenvalues in the two-fold (solid line) and the four-fold (dotted line) degenerate conduction band minima are shown in the inset

In Fig. 1 the valence and conduction bands in the vertical direction are plotted for a device with t_{ox} = 10 nm and applied gate voltage V_{GS} = 8 V, corresponding to a program operation. Eigenvalues for both the two-fold and the four-fold degenerate conduction band minima at the Si-SiO$_2$ interface are shown in the inset. Once the band profile is known, the density of the tunnel current can be readily obtained on the basis of Bardeen's tunneling Hamiltonian as:

$$J_T=\sum_i 2q\frac{T_{il}}{\tau_{il}}\frac{kTm_t}{\pi\hbar^2}\left\{\ln\left[\frac{1+(E_{il}-E_F)/kT}{1+(E_{il}-E_{FFG})/kT}\right]\right\} \\ +\sum_i 4q\frac{T_{it}}{\tau_{it}}\frac{kT\sqrt{m_l m_t}}{\pi\hbar^2}\left\{\ln\left[\frac{1+(E_{it}-E_F)/kT}{1+(E_{it}-E_{FFG})/kT}\right]\right\} \qquad ,(3)$$

where T_{id} is the tunneling probability of an electron with energy E_{id} and effective mass m_d ($d=l,t$), obtained by solving the Schrödinger equation in the barrier with the transfer matrix approach; τ_{id} ($d=l,t$) is the corresponding round trip time.

We repeat the calculation by varying the Fermi level in the floating gate E_{FFG} in a broad range, and compute the corresponding tunnel current density J_T and charge density Q_{FG} stored in the floating gate. In Fig. 2 we plot J_T versus -Q_{FG} for devices with oxide thickness t_{ox} ranging from 2 to 10 nm. All other parameters have the value already mentioned. The applied gate voltage is V_{GS} = 18 V, meaning that we are putting electrons in the floating gate, i.e., performing a "program" operation.

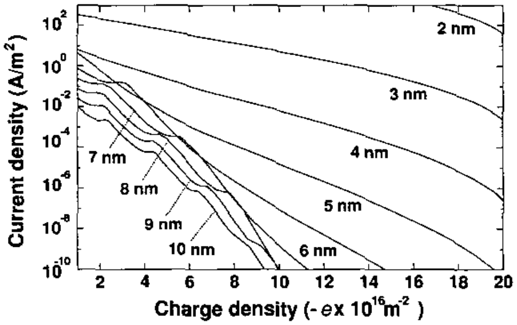

Fig. 2. Tunneling current density as a function of the charge density in the floating gate, for different values of oxide thickness

The mean descending behavior of the curves is due to the fact that as electrons enter the floating gate (Q_{FG} is negative), the electrostatic repulsion rises the potential barrier seen by electrons in the channel, and the tunneling current J_T consequently decreases. Therefore, a reliable estimate of write and retention times must rely upon the accurate simulation of the complete time evolution of potential profiles and charge densities.

For t_{ox} > 6 nm, J_T exhibits the well known oscillations associated to Fowler-Nordheim tunneling: indeed, for large values of $|Q_{FG}|$ the barrier becomes trapezoidal, and oscillations of J_T disappear. As can be seen, for barriers thinner than 6 nm, oscillations in J_T do not occur, since tunneling is always direct.

Starting from the continuity equation $J_T(Q_{FG}) = dQ_{FG}/dt$, we can obtain the time required to put a given charge Q_{FG} in the floating gate:

$$t(Q_{FG})=\int_{Q_{FG}(0)}^{Q_{FG}}\frac{1}{J_T(Q'_{FG})}dQ'_{FG} \qquad (4)$$

where $Q_{FG}(0)$ is the charge in the floating gate at time zero, i.e., with constant Fermi level in the whole device.

Once we compute the relationship between Q_{FG} and the threshold voltage, we can integrate the curves shown in Fig. 3 (Eq. (4)) in order to calculate the time required to obtain a given threshold voltage.

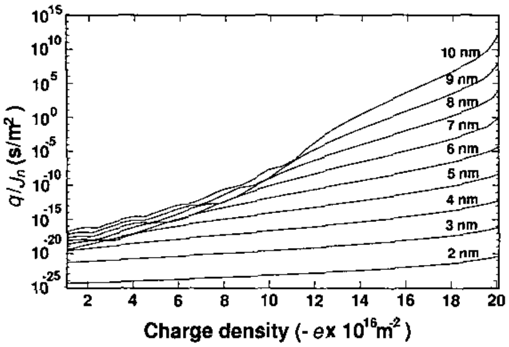

Fig. 3: Average time required to store an electron per unit area (q/J_T) as a function of the charge per unit area already stored in the floating gate, for different tunnel oxide thickness.

III. Results and discussion

In Fig. 4, we plot the threshold voltage shift, with respect to the threshold voltage with Q_{FG}= $Q_{FG}(0)$, versus the program time, for t_{ox} ranging from 2 to 10 nm. Again, oscillations for $t_{ox} \geq 6$ nm are associated to oscillations of the Fowler-Nordheim tunneling current. For smaller values of t_{ox} we observe a qualitatively different behavior, due to the occurrence of direct tunneling: oscillations disappear and the program time becomes much more sensitive to the oxide thickness.

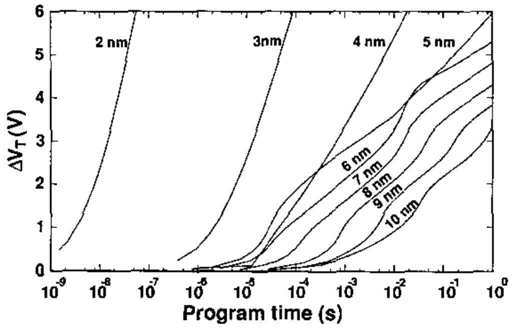

Fig. 4: Programmed shift of the threshold voltage as a function of the program time, for different values of oxide thickness, V_{GS} = 18 V.

The discharging process of the floating gate can be simulated analogously. The only differences are the fact that the Schrödinger equation is solved in the floating gate, and that current direction is reversed (Figs. 5 and 6). We assume that a memory initially programmed with a threshold voltage shift ΔV_T, corresponding to a stored charge Q_{FG}, is erased when enough charge is extracted to reduce the threshold voltage shift to 10% of the initial ΔV_T, i.e. when the residual charge in the floating gate is $Q_{FG} = Q_{FG}(0) + 0.1[Q_{FG} - Q_{FG}(0)]$, i.e.,

$$t_{erase}(Q_{FG}) = \int_{Q_{FG}}^{Q_{FG}} \frac{1}{J_T(Q'_{FG})} dQ'_{FG} \tag{5}$$

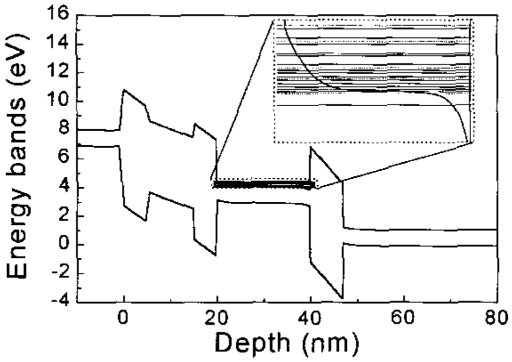

Fig. 5: Example of band profile during the erase operation with V_g = -8 V.

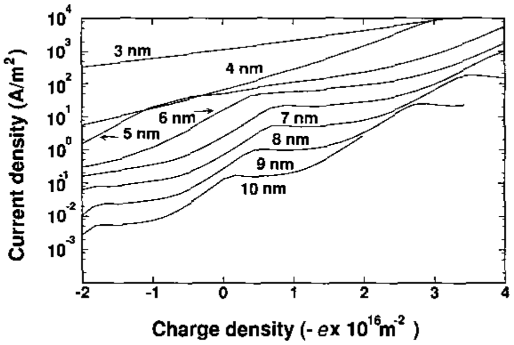

Fig. 6: Current density inthe erase operation as a function of the charge density store in the floating gate for VG=-18 V.

In Fig. 7 the initial ΔV_T is plotted as a function of t_{erase} for several values of oxide thickness, and $V_{GS} = -18$ V. The erase time is slightly reduced with increasing initial charge storage, since the presence of charge in the floating gate increases the field in the barrier and, as a consequence, the tunneling probability. It is worthy noticing that, also in the case of the erase operation, very thin barriers have a qualitatively different behavior: for $t_{ox} \geq 6$ nm the erase time decreases exponentially with decreasing oxide thickness, while for $t_{ox} < 6$ nm the decrease is much more accelerated.

To evaluate the data retention time, we assume that zero voltage is applied to the control gate, and that the information is reliably stored until less than 10% of the programming charge leaks out. Therefore, we still use Eq. (5) by considering Q_{FG} =Q_{FG} (0)+0.9[Q_{FG} - Q_{FG} (0)]. In Fig. 8 the initially programmed ΔV_T is represented on the *y*-axis while the retention time is on the *x*-axis. Since V_{GS} = 0V, the barrier is trapezoidal for any barrier thickness, therefore we observe a very regular behavior as the oxide thickness is reduced. For an initial $\Delta V_T = 4$ V, data retention times decreases by 3-4 orders of magnitude per nanometer of t_{ox}.

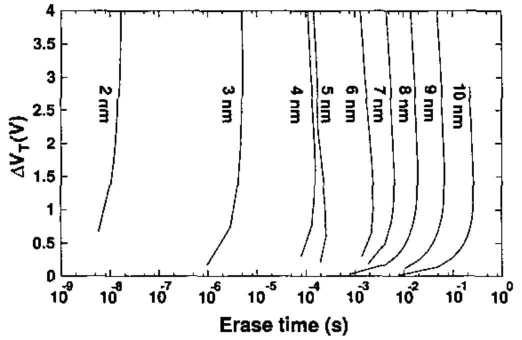

Fig. 7 : Erase times and associated initial ΔV_T for different oxide thickness and V_{GS}=-18 V. The memory is considered erased when the threshold voltage shift is reduced to 10% of the initial value.

The vertical dotted line represents ten years, which is the currently required data retention time for flash memories. It is important to point out that we are considering "perfect" oxides, while usually data retention time is limited by defects in the oxide, induced by high field stress during program-erase cycles, which would allow trap-assisted-tunneling in the oxide. The data retention time we have computed must be considered an upper limit. However, it is clear that if t_{ox} = 6 nm the ten year requirement can be met, for ideal oxides, only if ΔV_T is smaller than 8 V, if t_{ox} = 5 nm if ΔV_T < 3 V, if t_{ox} = 4 nm if ΔV_T < 1.2 V. Thinner oxides would imply unacceptably low retention times even in ideal conditions.

The developed code represents a useful tool for evaluating different structures for the gate stack, including alternative solutions that have been proposed to obtain a better tradeoff between program, erase and retention times[1][4]. An accurate evaluation of such times requires a fully quantum-mechanical solution of the time-dependent charging-discharging process.

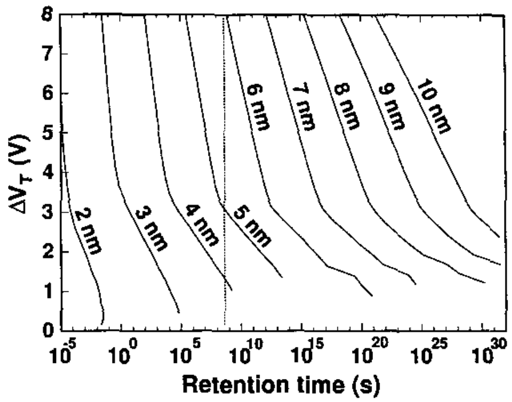

Fig. 8: Data retention times and associated initial ΔV_T for different oxide thickness. Data is considered lost when 10% of the charge injected during the program operation leaks out.

ACKNOWLEDGMENTS

Support from the NANOTCAD project (EU Contract IST-1999-10828) and from the Fondazione Cassa di Risparmio di Pisa is gratefully acknowledged.

REFERENCES

[1] K. K. Likharev, "Layered tunnel barriers for nonvolatile memory devices", *Appl. Phys. Lett.*, vol. 73, pp. 2137-2139, 1998.

[2] P. Pavan, R. Bez, P.Olivo, E. Zanoni, "Flash memory cells - an overview", *Proc. IEEE*, vol. 85, pp. 1248-1271, 1997.

[3] S. Sze, *Physics of semiconductor devices*, New York: Wiley and Sons, 2nd Ed., 1981.

[4] J. J. Welser, S. Tiwari, S. Rishton, K. Y. Lee, Y. Lee, "Room temperature operation of a quantum-dot flash memory", *IEEE Electron Device Letters*, vol. 18, pp. 278-280, 1997.