Program, Erase and Retention Times of Thin-oxide Flash-EEPROMs

G. IANNACCONE* and S. GENNAI

Dipartimento di Ingegneria dell’Informazione, Università degli studi di Pisa, Via Diotisalvi 2, I-56126 Pisa, Italy

We present an investigation of the process of charging and discharging of the floating gate of thin oxide Flash EEPROMs based on a fully quantum mechanical approach to transport in the vertical direction. Our approach allows us to compute program, erase, and retention times as a function of the gate stack structure, and applied voltages.

Keywords: Thin-oxide; Retention times; CHE; Schrödinger equation

I. INTRODUCTION

The purpose of this work is the simulation of program, erase, and retention times of thin-oxide Flash EEPROMs, in which the floating gate is charged through Fowler-Nordheim (FN) or direct tunneling. The use of thin oxides is required to have write times and operating voltages comparable to flash memories based on Channel Hot Electrons (CHE) charging [1]; 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 currently standard requirement.

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

II. APPROACH

Let us consider a Flash EEPROM with the following structure: substrate doping $N_A = 10^{18}$ cm$^{-3}$, tunnel oxide $t_{ox}$ n-type polysilicon floating gate ($N_D = 10^{19}$ cm$^{-3}$), triple O.N.O. layer (5 nm-thick oxide, 10 nm-thick nitride, 5 nm-thick oxide), and n$^+$ poly control gate ($N_D = 10^{20}$ 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.

*Corresponding author. Tel.: +39 050 568677, Fax: +39 050 568522, e-mail: g.iannaccone@iet.unipi.it
for the substrate \((E_{FS} = 0)\), the floating gate \((E_{FFG})\), and the control gate \((E_{FCG} = E_{FS} - qV_{GS})\). The Fermi-level in the floating gate \(E_{FFG}\) depends on the charge stored per unit area \(Q_{FG}\); we will actually assume the value of \(E_{FFG}\), and by varying it in a broad range we will find the relationship between \(E_{FFG}, Q_{FG}\), and the current density \(J_T\).

The Schrödinger equation is used only for electrons at the Si-SiO\(_2\) interface (in the substrate for the program operation, in the floating gate for the erase operation); therefore, the electron density \(n(x)\) is given by

\[
n(x) = \frac{2k_BTm_i}{\pi\hbar^2} \sum_i |\Psi_{ik}(x)|^2 \ln\left[1 + \exp\left(\frac{E_F - E_{ik}}{k_BT}\right)\right] + \frac{4k_BT\sqrt{m_im_t}}{\pi\hbar^2} \sum_i |\Psi_{ik}(x)|^2 \ln\left[1 + \exp\left(\frac{E_F - E_{it}}{k_BT}\right)\right],
\]

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

\[
-\frac{\hbar^2}{2m_k} \nabla^2 \Psi_{ik}(x) + E_c(x)\Psi_{ik}(x) = E_{ik}\Psi_{ik}(x).
\]

Electron densities in other regions and hole densities are computed with the semiclassical approximation [2]. The Poisson and Schrödinger equations are then solved self-consistently with an iterative under-relaxation algorithm.

In Figure 1 the valence and conduction bands in the vertical direction are plotted for a device with \(t_{ox} = 10\) nm and \(V_{GS} = 8\) 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 current density can be readily obtained on the basis of Bardeen’s tunneling Hamiltonian as:

\[
J_T = \sum_i \frac{2q T_{il} k_BT m_l}{\pi\hbar^2} \ln\left[1 + \frac{(E_{il} - E_F)/k_BT}{1 + (E_{il} - E_{FFG})/k_BT}\right] \ln\left[1 + \frac{(E_{it} - E_F)/k_BT}{1 + (E_{it} - E_{FFG})/k_BT}\right] + \sum_i 4q T_{it} k_BT(m_l m_t)^{0.5} \frac{\hbar^2}{T_{ik}} \ln\left[1 + \frac{(E_{ik} - E_F)/k_BT}{1 + (E_{ik} - E_{FFG})/k_BT}\right],
\]

where \(T_{ik}\) is the tunneling probability of an electron with energy \(E_{ik}\) and effective mass \(m_k(k = l, t)\) obtained by solving the Schrödinger equation in the barrier with the transfer matrix approach; \(\tau_{ik}\) is the corresponding round trip time.

We repeat the calculation sweeping 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 Figure 2 we plot the \(J_T\) versus \(-Q_{FG}\) for devices with oxide thickness \(t_{ox}\) ranging from 2 nm to 10 nm, in steps of 1 nm. All other parameters are those 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.
FIGURE 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 $J_T$ undergoes a steep decrease. 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 $t_{ox} \geq 6$ 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:

$$t(Q_{FG}) = \int_{Q_{FG}(0)}^{Q_{FG}} \frac{1}{J_T(Q_{FG})} dQ_{FG}, \quad (5)$$

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 perform the integral of Eq. (5) on the curves shown in Figure 2 in order to calculate the time required to obtain a given threshold voltage.

III. RESULTS AND DISCUSSION

In Figure 3, 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 $t_{ox}$ 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.

The discharging process of the floating gate can be simulated analogously. The only differences are the fact that the Schrodinger equation is solved in the floating gate, and that current direction is reversed. We assume that a memory initially programmed with a threshold voltage shift $\Delta V_T$, 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 $\Delta V_T$, 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.,

$$t_{erase}(Q_{FG}) = \int_{Q_{FG}}^{Q_{FG}(0)} \frac{1}{J_F(Q_{FG})} dQ_{FG}, \quad (6)$$

In Figure 4 the initial $\delta V_T$ is plotted as a function of $t_{erase}$ for several values of oxide thickness, and $V_{GS} = -18$ V. The erase time is slightly

FIGURE 3 Programmed shift of the threshold voltage as a function of the program time, for different values of oxide thickness, $V_{GS} = 18$ V.
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 behaviour: for \( t_{ox} \geq 6 \text{ nm} \) the erase time decreases exponentially with decreasing oxide thickness, while for \( t_{ox} < 6 \text{ nm} \) the decrease is 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. 6 by considering \( Q_{FG} = Q_{FG}(0) + 0.9[Q_{FG} - Q_{FG}(0)] \). In Figure 5 the initially programmed \( \Delta V_T \) is represented on the y-axis while the retention time is on the x-axis. Since \( V_{GS} = 0 \), the barrier is trapezoidal for all barrier thicknesses, therefore we observe a very regular behaviour as the oxide thickness is reduced. For an initial \( \Delta V_T = 4 \text{ V} \), data retention times decreases by 3–4 orders of magnitude per nanometer of \( t_{ox} \).

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 \( t_{ox} < 6 \text{ nm} \) the ten year requirement cannot be met even in principle, and that very thin oxides would imply unreasonably low retention times.

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 [3]. An accurate evaluation of such times requires a fully quantum-mechanical solution of the time-dependent charging-discharging process.

References