Recent Advances in Device Simulation Using Standard Transport Models

G. BACCARANI, M. RUDAN, M. LORENZINI and C. SALA

In this paper we address a number of issues related with device simulation in 3-D, and point out a few deficiencies which still prevent 3-D device simulation to be widely accepted as a standard tool for device design and optimization in an engineering environment. More specifically, such deficiencies have to do with structure definition and mesh generation, as well as with the computational burden which is typically associated with discretization meshes featuring 50,000 nodes or more. Next, we address the problem of validating advanced physical models in 3-D by means of our device simulator HFIELDS-3D and use a flash-EEPROM cell manufactured at ST-Microelectronics as a test vehicle for the validation of hot-carrier injection into the floating gate and Fowler-Nordheim tunneling across the gate oxide.

Keywords: Semiconductor-Device Simulation, Flash-EEPROM Cell, Hot-Carrier Injection, Fowler-Nordheim Tunneling, Interpoly Capacitance, Electron-Energy Distribution Function.

1. INTRODUCTION

Numerical simulation of silicon devices has reached a mature development stage. Two-dimensional simulation codes are increasingly being used both for a better understanding of device performance and for design purposes; furthermore, commercial tools featuring a friendly user interface and sophisticated simulation environments are currently made available to interested users. Three-dimensional simulation, instead, is still in an evolutionary phase, since a number of unsolved practical and theoretical problems still exist, which make it difficult to achieve an accurate prediction of the device performance. Such problems range from structure definition to mesh generation; also, the size of the computational problem is often too large to be properly handled in an engineering design environment. In addition, some physical models turn out to be very hard to quantitatively predict.

In this paper, we illustrate results achieved in the simulation of a flash-EEPROM cell designed for a 4 Megabit memory chip and, in doing so, we point out problems and inefficiencies arising from the incomplete availability of the necessary software tools. More specifically, we discuss the problem of structure definition and mesh generation and, from the simulation of the programming and erasing characteristics of the cell, address the problem of validating the physical models which are being used for the above functions, i.e. hot-electron injection into the floating gate, and Fowler-Nordheim tunneling across the gate oxide, respectively.
2. STRUCTURE DEFINITION AND MESH GENERATION IN 3-D

A major problem with structure definition in 3-D is the lack of reliable 3-D process-modeling tools. Generally speaking, such tools have a lesser predicting capability than device simulators, and this is especially true in 3-D, where process modeling is still in an early development stage. As a result, the device morphology is currently inferred from TEM cross sections, which involves a rather cumbersome procedure.

Also, the impurity concentration is inferred from 1-D and 2-D process simulations empirically combined to get the 3-D profiles. However, due to the difficulty of measuring experimental profiles in 3-D, the only feedback comes from the comparison between simulated and measured device characteristics. Quite often, a reverse-engineering process is needed, and the impurity profiles are adjusted in order to fit the turn-on characteristics.

Under such circumstances, the validation of sophisticated physical models, such as hot-electron injection into the gate oxide, turns out to be rather difficult, as the uncertainty in the device morphology and impurity concentration may lead to inaccurate electric fields.

Another area of importance is mesh generation. The spatial discretization of the semiconductor device equations has a substantial influence not only on the accuracy of the numerical results, but also on the convergence behavior. An appropriate spatial discretization must fulfill a number of requirements [1], including adaption to irregular geometries, accurate approximation of physical quantities, efficient allocation of grid points and suitable element form-factors.

From the experience gained in the past, it has now become clear that no unique algorithm or discretization element can fulfill all the requirements which are needed in order to generate meshes suitable for different tasks, such as process and device modeling, simulation of micro-electro-mechanical sensors and actuators and so on. More recently, a new approach based on a modular implementation of different algorithms has been devised [2]. The latter allows new basic elements and properties to be incorporated within the code, without loss of generality. These modules can be selected depending on the grid requirements for the specific application. The potential of the above approach raises the hope for a new generation of better and more reliable mesh generators.

3. PHYSICAL MODELS

Within the context of standard simulation tools, some physical effects turn out to be very difficult to model; this is especially true for physical mechanisms characterized by an energy threshold, such as impact ionization and hot-carrier injection into the oxide. In order to accurately predict such effects, the knowledge of the energy-distribution function of both types of carriers is a necessary prerequisite. This involves the solution of the Boltzmann Transport Equation (BTE) which is typically done using the Monte Carlo technique. If such information is not available, a relationship must be established between impact ionization and electric field or carrier temperature, if the latter is available.

Within homogeneous materials, and under a uniform electric field, there is a one-to-one correspondence between electric field, carrier temperature and energy distribution function. In deep submicron devices, however, the device length is comparable with the energy-relaxation distance: hence the energy distribution is never in equilibrium with the local field. Hence, any impact-ionization model based on the local value of the electric field turns out to be grossly inadequate and typically leads to an overestimation of the impact-ionization effect. Likewise, hot-carrier injection into the oxide requires a non-local model to account for the dynamics of high-energy electrons.

Better results can be achieved using the hydrodynamic transport model. Within this model, a simplified energy distribution function is associated with the carrier temperature. The occurrence of the investigated effect is then inferred from the population of the tail of the distribution function. Such an approach is obviously not general enough to warrant its applicability to a variety of devices under different operating
conditions. However, for a given class of devices, we expect the above assumption to be reasonably correct.

4. COMPARISON WITH EXPERIMENTS

In this section we are going to show simulation results of a flash-EEPROM cell which is a staked-gate device designed for the SGS-Thomson's 4 Megabit flash chip. This cell is intrinsically a three-dimensional device due to the morphology of the field-oxide region, shown in the TEM Fig. 1 micrograph of figure 1, which represents a cross section normal to the current flow. For this reason, a two-dimensional simulation of this device cannot predict the cell behavior, and reasonable results can be achieved only after a tuning of the effective device width and after an experimental determination of the coupling coefficient between the control gate and the floating gate. The latter, however, is a function of the specific operating conditions. Hence, an accurate prediction of the device performance in 2-D is very difficult to achieve.

Using a 3-D simulator, all capacitive couplings between the floating gate and the accessible contacts are automatically taken into account, and the scaling of the device current with respect to channel width can be investigated. Moreover, the space distribution of the generation rate due to impact ionization and band-to-band tunneling around the drain and source diffusions, in programming or erasing conditions, can be carefully studied. Therefore, the simulations are expected to predict the actual device behavior. The device cross-section parallel to the current Fig. 2 flow is sketched in figure 2.

To investigate such effects, cells with different widths were considered. It is worth noting that, in order to minimize process modifications which would introduce additional uncertainties in the geometry, the fabrication of cells with different channel widths has been attained by modifying only one process mask,
namely the one controlling the size of the active area. As a consequence, the dimension of the floating gate was kept constant, thus yielding a variable wing dimension. This means that the floating and control-gate areas are almost the same in cells having different widths, resulting in an unchanged interpoly capacitance, i.e. the capacitance between the control gate and the floating gate. In order to optimize the trade-off between accuracy and computational workload, significant effort has been devoted to the generation of a suitable structure for the simulation. Rather than attempting to reproduce the shape of the bird’s beak by a process simulator, the morphology of the cell was extracted from the TEM micrograph in figure 1 to act as a template. A commercial solid modeller and mesh generator were used to generate the two-dimensional grid, which was then replicated in the third dimension. As usual, symmetry conditions were taken advantage of by considering only half of the EEPROM cell.

The fine-tuning of the cell geometry and doping profiles for the source, drain, and channel regions was pursued by reverse-engineering of the transfer and output characteristics of the cell for three different channel widths. So doing, a validation of the 3-D structure thus generated was achieved. The agreement between simulations and measurements shown in figures 3 and 4 is acceptable, in spite of a slight deviation in threshold voltage and saturation behavior. It is worth mentioning that no physical parameter was adjusted to fit the data. The former deviation can be ascribed to inaccuracies in the surface-channel doping and to an insufficient mesh refinement in the same region, while the deviations in the output current could be due to a limitation of the surface-mobility model in the high doping region. In any case, local changes of the threshold voltage on wafer are of the same order of magnitude.

In order to reproduce the writing and erasing characteristics of the cells, i.e. the curves representing the threshold-voltage shift as a function of time, it is necessary to determine the relation between the threshold-voltage shift $\Delta V_t$ and the floating-gate charge $Q_{fg}$ as evaluated by HFIELDS-3D. A typical assumption is that the stored charge shifts the threshold voltage according to the relation [3,4]

$$\Delta V_t = -\frac{Q_{fg}}{C_{pp}}$$

(1)

where $C_{pp}$ is the interpoly capacitance. For an unwritten cell it is $Q_{fg} = 0$, hence $\Delta V_t = 0$. During a writing operation, the building-up of negative charge within

![FIGURE 3 Simulated and experimental turn-on characteristics for the three EEPROM cells with different channel widths. $V_{ds} = 0.1$ V](image_url)
the floating gate progressively increases the threshold-voltage. During an erasing operation, instead, the threshold-voltage progressively decreases due to the extraction of electrons from the floating gate.

By simulating the transfer characteristics with different amounts of charge stored in the floating gate, the linear relation between $Q_{fg}$ and $\Delta V_t$ expressed by Eq. (1) is confirmed and the interpoly capacitance relative to half cell can be estimated as $1.8 \times 10^{-15}$ F for all channel widths, as shown in figure 5.

A non-linear scaling of the drain current with channel width $W$ is experimentally observed: the narrower cell carries a higher current density, as shown in figure 6 where the normalized drain current $I_d/W$ is...
reported against the control-gate voltage. This can be explained as follows. Considering for instance a transfer characteristic at $V_g = 0$, $V_b = 0$, $Q_{fg} = 0$, and using a simplified capacitive equivalent circuit for the cell, the floating-gate coupling to other contacts can be represented by the following linear equation \[ V_{fg} = \alpha_g V_{cg} + \alpha_d V_d \] \[ \text{Eq. (2)} \]

where $\alpha_g = C_{pp}/C_T$ is the coupling ratio between control and floating gates, $\alpha_d = C_d/C_T$ is the floating gate to drain coupling ratio, and $C_T$ is the total capacitance ($C_T = C_{pp} + C_d + C_s + C_b$). Eq. (2) simplifies to

$$V_{fg} = \alpha_g V_{cg} = \frac{C_{pp}}{C_T} V_{cg}$$ \[ \text{Eq. (3)} \]

due to the fact that $\alpha_d$ is typically one order of magnitude smaller than $\alpha_g$ and $V_d = 0.1$ V. Simulations show that increasing the channel width at constant $V_{cg}$ decreases the potential of the floating gate, leading to a higher threshold voltage. As expected, this means that, while $C_{pp}$ is unchanged, $C_T$ increases, due to the increase in the planar region of the oxide. The increase in the floating-gate potential at decreasing channel widths strongly affects the erasing behavior of the cell, as shown in section 6.

5. PROGRAMMING OPERATION

The cell under investigation is programmed by hot-electron injection at the drain end of the channel. Hot-electron injection requires a large lateral electric field in silicon for the carriers to attain a high enough energy to overcome the barrier. Once they are injected into the floating gate, they are confined by the potential barrier of approximately 3 eV at the polysilicon-oxide interface.

Figure 7 shows the threshold voltage shift vs programming time for different values of the drain voltage. As more and more negative charge is injected into the floating gate, the voltage of the latter decreases until it reaches the drain voltage. At this point, the field in the oxide reverses, and slows down further charge injection. This effect is revealed by the sudden slope change visible in the curves of figure 7. The information contained in the above measurements is converted to a curve relating the injection current to the floating-gate voltage, using the following expression:

$$I_{fg} = \frac{dQ_{fg}}{dt} = C_{pp} \frac{dV_{fg}}{dt}$$ \[ \text{Eq. (4)} \]
and

\[ V_{fg} = \alpha_g (V_{cg} - \Delta V_t) + \alpha_d V_d \]  

(5)

A typical example of the gate current vs floating-gate voltage is shown in figure 8. Clearly, the current drops off as soon as the floating gate voltage approaches the drain voltage, due to the reverse oxide field which opposes electron injection.

The hot-electron injection model incorporated in HFIELDS-3D accounts to some extent for both the non-Maxwellian form of the electron-energy distribution and for the non-local nature of carrier heating. Fundamentally, it is based on the expression for the
energy distribution of hot electrons \( f(E) \) derived in [6] under some simplifying assumptions for electron transport (i.e., carriers belong to a single nonparabolic conduction band and their dominant energy-loss mechanism is the emission of optical phonons)

\[
f(E) = C \exp \left( -\frac{\chi E^3}{|F|^{1.5}} \right)
\]

where \( E \) is the electron energy, \( F \) represents the electric field, \( C \) is a weak function of \( F \), and \( \chi = \text{const} \). In order to account for non-local effects, in [7] the actual electric field \( F \) is replaced by an effective field \( F_e \) depending linearly on the electron temperature \( T_e \). In case of carrier heating (i.e., \( T_e > T_L \) where \( T_L \) is the lattice temperature), the hot-electron injection current density can be written as

\[
J_{inj} = -qnA \int_{E_B + \Delta E_B}^{\infty} E^{1.5} \exp \left( -\frac{\chi E^3}{F_e^{1.5}} \right) dE
\]

where \( E_B \) denotes the height of the barrier at the Si-SiO₂ interface; \( \Delta E_B \) is the increase in the barrier height due to the oxide-field reversal which may occur near the drain if the drain voltage exceeds the floating-gate voltage; \( n \) is the electron concentration, and the values of the constants \( A \) and \( \chi \) are determined by fitting experimental data. Eq. (7) is then multiplied by a damping term, taking into account the electron scattering in the image force potential well into the SiO₂ layer [8]. The model has been implemented as a post processor, due to the fact that the gate current is normally orders of magnitude lower than the other currents of interest. The above model was then improved and extended to holes [9] in analogy with a substrate current model based on carrier temperature [10].

Figure 9 shows a comparison between simulation and experiment. An obvious shortcoming of the model is that the simulated threshold-voltage shift is sensibly larger than the measured one at large values of time, since the former curve does not account for the sudden change in the Fig. 10 slope which characterizes the experimental data. This is corroborated even further by figure 10, where simulated and experimental current data are reported in a linear plot versus the floating-gate voltage. Figure 10 exhibits a smoother drop of the simulated current as the floating-gate voltage decreases.

Such a shortcoming has been overcome by assuming a modified functional form for the electron-distri-

![Figure 9](image_url)

**Figure 9** Experimental and simulated threshold-voltage shift against programming time for \( W = 0.7 \mu m, V_{cg} = 11 \text{ V} \) and \( V_d = 4.5 \text{ V} \)
bution function [11,12]. A linear combination of two exponentials has been posited

\[
f(E) = C \left[ \exp \left( -\frac{\chi_n E^3}{T_n^{1.5}} \right) + C_0 \exp \left( -\frac{\chi_p E^3}{T_p^{1.5}} \right) \right]
\]  

(8)

in order to overcome the assumption that the electrons belong to a single nonparabolic conduction band. The resulting gate current model has the decided advantage to fit correctly the observed shoulder in the measured electron gate current at the expense of a larger number of fitting parameters. This is shown in figure 11, where the experimental and simulated gate currents turn out to be in a much better agreement than before [11].

The computational workload involved in the simulation of the programming characteristics of realistic three-dimensional structures is rather heavy, primarily due to the size of the mesh and the transient nature of the effect under investigation. This seems to be one of the most important hurdles to the acceptance of three-dimensional device simulators in an engineering environment.

6. ERASING OPERATION

Cell erasing is performed by keeping the drain contact floating, whilst quickly ramping a moderate positive bias on the source contact and a large negative bias on the gate contact. In this situation, a large number of carriers tunnel through the Poly-Si-SiO\(_2\) potential barrier. For emission into a dielectric, using a free-electron gas model for polycrystalline silicon, and the WKB approximation for the tunneling probability, and taking the effect of the image force on the barrier into account, the following expression for the Fowler-Nordheim tunneling current density is obtained [13,14]

\[
J_{FN} = \frac{q^3 m_F^2}{8 \pi \hbar m^* \phi} \frac{1}{r^2(y)} \sin(\pi c k T) \exp \left( -\frac{4 \sqrt{2m^* \phi^4}}{3q \hbar F} v(y) \right)
\]  

(9)

where

\[
c = \frac{2 \sqrt{2m^* \phi}}{q \hbar F} f(y),
\]
\[ F \] is the electric field, \( m_{ox}^* \) is the effective mass of the electron in the forbidden gap of the oxide, \( \phi \) represents the Poly-Si-SiO\(_2\) barrier height, and
\[
y = \frac{1}{\phi} \sqrt{\frac{q^3 F}{4\pi \varepsilon_0 m_{ox}^* \phi}}
\]
is the normalized image-force barrier lowering \((y = \Delta \phi / \phi)\). Both \( \nu(y) \) and \( t(y) \) are slowly varying functions of \( y \); their expressions can be found in [13]. The other quantities have their usual meaning. A simplification is often made because \( f^2 \) varies only between 1.00 and 0.81 and the correction due to the multiplicative term \( \pi c k T / \sin(\pi c k T) \) is \( \approx 1 \) for the ordinary temperatures; the following expression is obtained
\[
J_{FN} = \frac{q^3 m F^2}{8\pi \hbar m_{ox}^* \phi} \exp\left( -\frac{4\sqrt{2m_{ox}^* \phi^3}}{3q\hbar F} \nu(y) \right).
\]

Only if the image-force correction is completely neglected, it is possible to derive from Eq. (10) the usual Lenzlinger and Snow expression
\[
J_{FN} = A_{FN} F^2 e^{-B_{FN} / F},
\]
where
\[
A_{FN} = \frac{q^3 m}{8\pi \hbar m_{ox}^* \phi}, \quad B_{FN} = \frac{4\sqrt{2m_{ox}^* \phi^3}}{3q\hbar}
\]
are constants, i.e., independent of the electric field. Assuming \( m_{ox}^* = 0.5 \ m \) [15], one obtains
\[
A_{FN} = 3.08 \times 10^{-6} \frac{A}{V^2}, \quad B_{FN} = 4.83 \times 10^7 \phi^{3/2} \ V/cm
\]
where \( \phi \) is expressed in eV. The tunneling current flows almost uniformly across the thin oxide, and no important edge effect was detected. For this reason, one can argue that the injection through the oxide depends on the channel width via the injection area and the coupling coefficient \( \alpha_g \). The former effect increases the tunnel current at increasing channel widths, whereas the latter causes a variation in the opposite direction. In figure 12, measured erasing curves at different channel widths are represented: the narrower the cell, the faster the discharging process.

Hence, the geometrical effect mentioned above has a strong impact on the erasing process. Note that the largest cell is markedly slower than the others. Coming now to the simulations, it was found that on a given structure it is possible to fit the erasing curve at different source voltages, as shown in figures 13, 14 and 15. However it is not possible to do so with a unique set of fitting parameters (i.e., with the same value of \( \phi \)) for all the curves at different channel widths, as shown in figure 16. Since there are no
FIGURE 12 Measured threshold-voltage shift vs erasing time for the three different channel widths. $V_{cg} = -8 \text{ V}$, $V_t = 4 \text{ V}$, $I_d = 0 \text{ A}$

FIGURE 13 Measured and simulated threshold-voltage shift vs erasing time for three different values of the source voltage. $W = 0.9 \text{ \textmu m}$, $V_{cg} = -8 \text{ V}$, $\Phi = 2.95 \text{ eV}$
FIGURE 14 Measured and simulated threshold-voltage shift vs erasing time for three different values of the source voltage. $W = 0.7 \, \mu m$, $V_{cg} = -8 \, \text{V}$, $\Phi = 2.83 \, \text{eV}$

FIGURE 15 Measured and simulated threshold-voltage shift vs erasing time for three different values of the source voltage. $W = 0.5 \, \mu m$, $V_{cg} = -8 \, \text{V}$, $\Phi = 2.83 \, \text{eV}$
physical reasons to believe that the barrier height is different in cells having different widths and, on the other hand, the values obtained for $\phi$ are different and lower than expected (i.e., it is necessary to enhance the field emission into the oxide), such a lack of agreement seems to indicate the existence of oxide thinning at the end of the bird’s beak of the cell for the smaller cells. This effect leads to a local enhancement of the tunneling field which, due to the exponential dependence of the tunneling current upon the field, may cause a strong enhancement of the erasing characteristics.

In principle, it would be possible to use a unique set of values for $A_{FN}$ and $B_{FN}$ by introducing a smaller effective-oxide thickness. Once the parameters are fixed for the largest cell, they are expected to scale as $A_{FN}/\gamma^2$ and $B_{FN}/\gamma$ for the narrower cells, where $\gamma = t_{ox}/t_{eff} = F_{eff}/F_{ox}$ is determined by fitting the experimental curves. Of course, the above reasoning is based on the assumption that the oxide thinning affects more sensibly the narrower cells than the largest one, as it seems.

When considering the amount of oxide thinning, it turns out that the above discrepancies are fully explained by assuming a thinning of 0.5 nm, i.e. about one atomic layer. Thus, the extreme sensitivity of the tunneling current upon the oxide thickness makes it very hard to accurately predict the erasing behavior of the cell. Moreover, the uniformity of the process can hardly be controlled within the above dimensional limits, so that the simulation is within the spreading of the experimental results.

7. CONCLUSION

In this paper we examine some of the problems related with device simulation in 3-D, and address the major deficiencies which still prevent a wide acceptance of 3-D simulation in an engineering design environment for process and device optimization. More specifically, we identify structure definition, mesh generation and CPU time as the most important areas where a substantial effort is needed in order to make
3-D device simulation a useful tool for engineering purposes. Next, we address the problem of validating advanced physical models in 3-D and use a flash-EEPROM cell as a test vehicle for the validation of the basic mechanisms employed to perform the writing and erasing functions, i.e., hot-carrier injection into the floating gate and Fowler-Nordheim tunneling across the gate oxide. Three different channel widths are considered, namely 0.9, 0.7 and 0.5 μm and a non-linear scaling behavior of the drain current is detected and properly interpreted.

Acknowledgements
The authors acknowledge the contribution of Ann Concannon to the hot-carrier injection model, which was improved during her stay in Bologna within the framework of the Human Capital and Mobility Program supported by the EU.

References

Biographies
Giorgio Baccarani received his Dr. Ing. degree in Electrical Engineering in 1967 and his Dr. degree in Physics in 1969, both from the University of Bologna, Italy, where he is currently Full Professor of “Digital-System Electronics”. His research interests include numerical-device simulation, integrated-circuit design, analog and digital architectures for image processing, character and voice recognition, neural networks and fuzzy systems. He is author or coauthor of over 100 papers published on international scientific journals, and editor of two books. He is senior member of IEEE, and corresponding member of the Academy of Sciences in Bologna.

Massimo Rudan received a degree in Electrical Engineering in 1973 and a degree in Physics in 1976, both from the University of Bologna, Italy. He joined the Department of Electronics (DEIS) of the University of Bologna in 1975, where he investigated the physical properties of the MOS structures and the problems of analytical modelling of semiconductor devices. Since 1983 he has been working in a group involved in numerical analysis of semiconductor devices, acting as Task leader in a number of EEC-supported Projects in the area of CAD for VLSI. In 1986 he has been a visiting scientist, on a one-year assignment, at the IBM Thomas J. Watson Research Center at Yorktown Heights, NY. In 1990, he was appointed Full Professor of Microelectronics at the University of Bologna.

Martino Lorenzini received his degree in Electrical Engineering from the University of Bologna, Italy, in 1993. Currently, he is working towards his Ph.D. degree at the Department of Electronics (DEIS) of the same University in the field of the numerical simula-
tion of non-volatile memory cells. His research interests include the physics and modeling of semiconductor devices.

Cristiano Sala received his degree in Physics from the University of Milan, Italy, in 1986 and his Ph.D. degree from the Katholieke Universiteit Leuven, Belgium, in 1991. He is currently Senior TCAD Engineer at SGS-Thomson Microelectronics in Agrate Brianza, Milan. His current research interest include the simulation of semiconductor processes, with particular emphasis on high-energy, high-angle implantation and diffusion phenomena, and the three-dimensional simulation of non-volatile memory devices.
Submit your manuscripts at http://www.hindawi.com