Study of Warm-Electron Injection in Double-Gate SONOS by Full-Band Monte Carlo Simulation

Gino Giusi
Dipartimento di Elettronica, Informatica e Sistemistica, Università della Calabria
Giuseppe Iannaccone
Dipartimento di Ingegneria dell'Informazione: Elettronica, Informatica, Telecomunicazioni, Università di Pisa
Mohamed Mohamed
University of Illinois at Urbana–Champaign, Urbana
Umberto Ravaioli
University of Illinois at Urbana–Champaign, Urbana

G. Giusi, G. Iannaccone, M. Mohamed, U. Ravaioli, *Study of Warm-Electron Injection in Double-Gate SONOS by Full-Band Monte Carlo Simulation*, IEEE Electron Device Letters, **29**, n. 11, pp. 1242 -1244, (2008)

Study of Warm-Electron Injection in Double-Gate SONOS by Full-Band Monte Carlo Simulation

Gino Giusi, Giuseppe Iannaccone, Mohamed Mohamed, and Umberto Ravaioli

***Abstract*—In this letter, we investigate *warm*-electron injection in a double-gate SONOS memory by means of 2-D full-band Monte Carlo simulations of the Boltzmann transport equation. Electrons are accelerated in the channel by a drain-to-source voltage V_{DS} smaller than 3 V, so that programming occurs via electrons tunneling through a potential barrier whose height has been effectively reduced by the accumulated kinetic energy. Particle energy distribution at the semiconductor/oxide interface is studied for different bias conditions and different positions along the channel. The gate current is calculated with a continuum-based postprocessing method as a function of the particle distribution obtained from Monte Carlo simulation. Simulation results show that the gate current increases by several orders of magnitude with increasing drain bias, and warm-electron injection can be an interesting option for programming when short-channel effects prohibit the application of larger drain bias.**

***Index Terms*—FinFET memory, nonvolatile memory, SONOS.**

I. Introduction

RECENTLY, multigate MOSFET architectures proposed to reduce short-channel effects have been investigated for nonvolatile memory applications also [1], [2]. Multigate architectures keep short-channel effects under control for reading bias, but in the case of channel hot-electron programming, the maximum applicable drain-to-source voltage is limited by punchthrough. In particular, for aggressively scaled devices, V_{DS} cannot be larger than 3.15 V corresponding to the silicon oxide–silicon potential barrier in the conduction band [1], [3]. This means that electrons cannot acquire in the channel sufficient kinetic energy to overcome the potential barrier represented by the gate dielectric, i.e., are not sufficiently "hot." However, experiments show that gate injection for V_{DS} smaller than 3 V is much more efficient than in the case of Fowler–Nordheim programming, meaning that a "warm-electron tunneling" mechanism can represent a reasonable option for nonvolatile memory programming [1], [3].

In this letter, we aim to investigate the warm-electron tunneling regime and to evaluate its efficiency in SONOS programming. Because data information is stored in the ONO stack through gate tunneling, accurate modeling of the gate current is extremely important for evaluating device performance. In such short devices, the transport problem can only be accurately modeled by solving the Boltzmann transport equation (BTE). However, the gate current is several orders of magnitude smaller than the drain current, and its calculation poses a tremendous challenge to particle-based methods. Attempts to solve the BTE for the gate current problem were made [4], [5]. An energy transport model and a Monte Carlo approach were successfully applied to gate current calculations for the case of hot-carrier injection [6]–[9].

In this letter, we use the Monte Carlo approach to calculate charge distribution and the electrostatic potential, and we compute the transmission coefficient as a function of energy using the WKB approximation. Size quantization and barrier lowering are neglected. We are interested in particular in the contribution to charge injection due to electrons whose energy is lower than the barrier height (i.e., tunneling electrons).

II. Monte Carlo Simulation

Simulations have been performed with the Monte Carlo solver MoCa [10], purposely modified to include the simulation of the gate current with a continuum-based method (as opposed to the particle-based method used to compute particle distributions and transport in the channel). We believe that a full-band 3-D Monte Carlo solver, including short-range particle–particle interaction via the P3M method, already implemented in MoCa [11] would provide a natural and accurate means of computing electron distributions in the channel. In this letter, we want to examine the concept of warm-electron injection, and we prefer to adopt an approximate approach to reduce the computational complexity of the problem by using a 2-D full-band version of MoCa, including all the relevant scattering mechanisms: Electron–electron interaction is approximately taken into account by self-consistently solving the Poisson equation on a fine grid, without explicitly introducing electron–electron scattering mechanisms such as those proposed in [12]. In the proposed method, the gate current is calculated with a postprocess approach by extracting the particle distribution in position and energy from the Monte Carlo solver. Let us refer to the double-gate (DG) SONOS structure shown in Fig. 1, where y is the direction of tunneling. Our model assumes that the total carrier energy (E) and the transversal momentum (k_x, k_z) are conserved during tunneling and that the dispersion relation in

Manuscript received July 21, 2008. First published October 10, 2008; current version published October 22, 2008. This work was supported in part by the FinFlash IST-NMP Project, by FIRB Project RBIP06YSJJ, and by the RHESSA Project of the Minister of Foreign Affairs. The review of this letter was arranged by Editor X. Zhou.

G. Giusi is with the Dipartimento di Elettronica, Informatica e Sistemistica, University of Calabria, 87036 Arcavacata di Rende, Italy (e-mail: ggiusi@deis.unical.it).

G. Iannaccone is with the Dipartimento di Ingegneria dell'Informazione, Elettronica, Informatica, Telecomunicazioni, Università di Pisa, 56126 Pisa, Italy (e-mail: g.iannaccone@iet.unipi.it).

M. Mohamed and U. Ravaioli are with the University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA (e-mail: mohamed@uiuc.edu; ravaioli@uiuc.edu).

Digital Object Identifier 10.1109/LED.2008.2004784

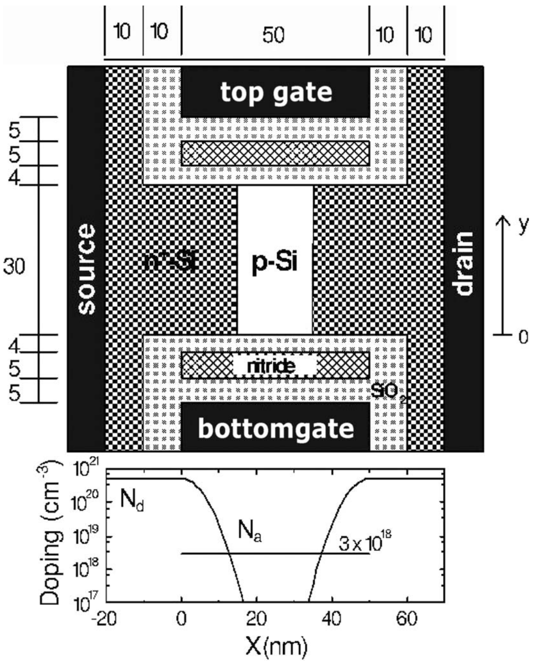

Fig. 1. Simulated structure is a DG SONOS memory with a 4-nm tunnel oxide, a 5-nm nitride oxide, a 5-nm top oxide, a 30-nm fin width, and a 50-nm gate length. The acceptor fin doping is 3×10^{18} cm^{-3}, while the source/drain doping extends under the gate for 15 nm from each side. y is the direction of tunneling (perpendicular to the Si$-$SiO$_2$ interface). The interfaces are at $y = 0$ and $y = 30$ nm.

the oxide is parabolic with isotropic effective mass m_{ox}. The component of the kinetic energy contributing to tunneling is therefore

$$E_y = E_{\text{kin}} - \frac{\hbar^2}{2m_{\text{ox}}}\left(k_x^2 + k_z^2\right) \tag{1}$$

and the effective barrier height is $\phi_s = B - E_y$, where $B = 3.15$ eV is the barrier height of the Si$-$SiO$_2$ interface. We consider only particles that are at the Si$-$SiO$_2$ interface and have a positive velocity (v_y) in the tunneling direction. For each particle, we can calculate E_y from (1) and compute the quantity $\langle v_y n(x, y, E_y)\rangle$, where n is the electron density per unit area per unit E_y. References [13] and [14] have shown that with proper barrier parameters, the I–V characteristics of thin gate stacks can be reproduced with reasonable accuracy of several orders of magnitude without taking into account barrier lowering and with the WKB approximation. We therefore compute the transmission coefficient $T = T(E_y)$ as in [13]. The tunneling current density can be calculated using the formula

$$J_G(x, z) = q\int \langle v_y n(x, z, E_y)\rangle T(E_y)dE_y. \tag{2}$$

An alternative approach consists of assuming that only the total energy is conserved during tunneling, without any assumption on oxide bands. In this case, E_y can be calculated by

$$E_y = E_{\text{kin}}(k_x, k_y, k_z) - E_{\text{min}}(k_x, k_z) \tag{3}$$

where E_{min} is the minimum energy that can be obtained by conserving the transversal momenta k_x and k_z. In other words, $E_{\text{min}} = \min_{ky}\{E_{\text{kin}}(k_x, k_y, k_z)\}$, where k_y can vary in the whole Brillouin zone. This solution is identical to the previous one only if the band structure in the silicon and in the oxide are parabolic and if the masses along k_x and k_z are identical. The advantage of this second solution is that we do not have to take into account the band index and the particular conduction-band minimum at the price of some computational cost for obtaining E_{min}.

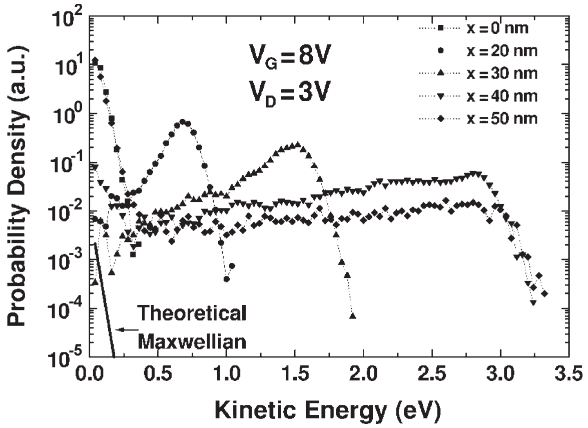

Fig. 2. Distribution of the kinetic energy E_{kin} for electrons at the silicon–oxide interface in various positions along the channel. The electron distribution obeys the Maxwell–Boltzmann law only at the source ($x = 0$ nm), and it progressively departs from an equilibrium distribution as the drain is approached.

III. Simulations

The simulated structure is an n-channel DG SONOS memory with a 50-nm channel length and a 4/5/5-nm ONO stack (Fig. 1). In Fig. 2, we show the distribution of the kinetic energy E_{kin} for electrons at the silicon–oxide interface in various positions along the channel. As can be seen, only at the source ($x = 0$ nm) electrons obey a Maxwell–Boltzmann distribution. The distribution progressively differs from a displaced equilibrium distribution when approaching the drain side, and it has a maximum at a kinetic energy value very close to the potential energy drop with respect to $x = 0$ nm, which would correspond to ballistic electron transport. In addition, the distribution becomes more and more asymmetric with respect to the maximum, and in the drain region, it is rather flat for energies up to the maximum, with a thermal equilibrium tail for larger kinetic energies corresponding to the lattice temperature. Since transport is partially ballistic, for a drain voltage V_D, a significant portion of electrons injected from the source have in the vicinity of the drain a kinetic energy eV_D so that they see a barrier toward the gate of approximate height $B - eV_D$. Such effective barrier lowering significantly increases local tunneling close to the drain. Fig. 3 shows the gate current density J_G as a function of the longitudinal position for different drain biases. Gate current density was calculated by (2), and the transmission coefficient was calculated by assuming parabolic oxide bands (1). No significant difference in the gate current has been observed by considering effective bands (3). As one can see, the drain voltage significantly increases gate tunneling, which also increases by several orders of magnitude between the source ($x = 0$ nm) and the drain ($x = 50$ nm). It is very important to underline that injection is also important for eV_D lower than

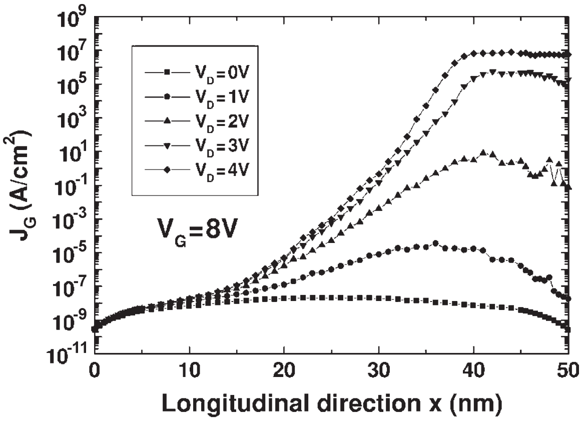

Fig. 3. Gate current density J_G as a function of the longitudinal position for different drain biases. The drain voltage significantly increases gate tunneling, which also increases by several orders of magnitude from the source ($x = 0$ nm) to the drain ($x = 50$ nm).

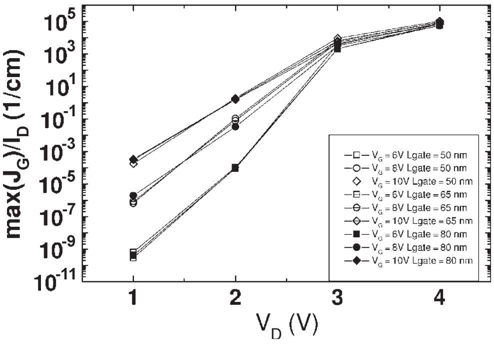

Fig. 4. Programming efficiency defined as the ratio of the maximum gate current density in the channel to the drain current. In the warm-electron injection regime ($V_D < 3$ V), the programming efficiency strongly increases with increasing V_{DS}, whereas it almost saturates for $V_D > 3$ V.

the barrier height (3.15 eV) so that gate programming can be obtained also by "warm electrons."

One should also keep in mind that gate current densities may change significantly in time due to nitride charging. In our calculations, this effect is not considered, and the nitride layer is assumed to be neutral, as it is in the initial phase of the program operation. Still, our approach should provide significant information on the injection mechanisms.

Fig. 4 shows the "programming efficiency" defined in this case as the ratio of the maximum gate current density along the channel to the drain current for different program bias conditions and for different gate lengths (it is an inverse length). In the warm-electron injection regime ($V_{\mathrm{DS}} < 3$ V), the programming efficiency exponentially increases with increasing V_{DS}, whereas it almost saturates for $V_{\mathrm{DS}} > 3$ V. This means that since short-channel effects in nanoscale SONOS memories pose an upper limit to the maximum applicable drain bias during programming, warm-electron injection can provide an interesting option for achieving fast programming speed, as confirmed by experimental results [1], [3], [15].

IV. Conclusion

In this letter, we have investigated through 2-D full-band Monte Carlo simulations the gate current injection in a DG SONOS memory programmed with warm electrons, i.e., with V_{DS} smaller than the silicon oxide–silicon barrier height. The particle energy distribution at the semiconductor/oxide interface is extracted at different bias conditions and different positions along the channel. Gate current is calculated during a postprocessing phase as a function of the particle distribution, neglecting size quantization. Simulation results show that injection is effective also for low drain biases because of the very strong dependence of gate current on V_{DS}. Warm-electron injection could be an interesting option for very short devices for which punchthrough does not allow one to apply larger V_{DS}.

References

[1] F. Hofmann, M. Specht, U. Dorda, R. Kömmling, L. Dreeskornfeld, J. Kretz, M. Städele, W. Rösner, and L. Risch, "NVM based on FinFET device structures," *Solid State Electron.*, vol. 49, no. 11, pp. 1799–1804, Nov. 2005.

[2] S. Lombardo, C. Gerardi, L. Breuil, C. Jahan, L. Perniola, G. Cina, D. Corso, E. Tripiciano, V. Ancarani, G. Iannaccone, G. Iacono, C. Bongiorno, C. Garozzo, P. Barbera, E. Nowak, R. Puglisi, G. A. Costa, C. Coccorese, M. Vecchio, E. Rimini, J. Van Houdt, B. De Salvo, and M. Melanotte, "Advantages of the FinFET architecture in SONOS and nanocrystal memory devices," in *IEDM Tech. Dig.*, Washington, DC, Dec. 2007, pp. 921–924.

[3] L. Perniola, E. Nowak, G. Iannaccone, P. Scheiblin, C. Jahan, G. Pananakakis, J. Razafindramora, B. De Salvo, S. Deleonibus, G. Reimbold, and F. Boulanger, "Physical model for NAND operation in SOI and body-tied nanocrystal FinFLASH memories," in *IEDM Tech. Dig.*, Washington, DC, Dec. 2007, pp. 943–946.

[4] Z. Han, C. Lin, N. Goldsman, I. Mayergoyz, S. Yu, and M. Stettler, "Gate leakage current simulation by Boltzmann transport equation and its dependence on the gate oxide thickness," in *Proc. SISPAD*, 1999, pp. 247–250.

[5] H. Lin and J. Peng, "An efficient physics-based gate current calculation by solving space-dependent Boltzmann transport equation," in *Proc. Univ./Gov./Ind. Microelectron. Symp.*, 1995, pp. 193–196.

[6] C. M. Yih, G. H. Lee, and S. Chung, "A consistent gate and substrate current model for submicron MOSFETs by considering energy transport," in *Proc. VLSI Technol., Syst., Appl.*, 1995, pp. 127–130.

[7] A. Harkar, R. W. Kelsall, and J. N. Ellis, "Monte Carlo study of the lateral distribution of gate current density along the channel of submicron LDD MOSFETs," *VLSI Des.*, vol. 13, no. 1–4, pp. 301–304, 2001.

[8] Y. Ohkura, C. Suzuki, H. Amakawa, and K. Nishi, "Analysis of gate currents through high-*k* dielectrics using a Monte Carlo device simulator," in *Proc. SISPAD*, 2003, pp. 67–70.

[9] K. Hasnat, C.-F. Yeap, S. Jallepalli, S. A. Hareland, W.-K. Shih, V. M. Agostinelli, A. F. Tasch, and C. M. Maziar, "Thermionic emission model of electron gate current in submicron NMOSFETs," *IEEE Trans. Electron Devices*, vol. 44, no. 1, pp. 129–138, Jan. 1997.

[10] G. A. Kathawala, B. Winstead, and U. Ravaioli, "Monte Carlo simulations of double-gate MOSFETs," *IEEE Trans. Electron Devices*, vol. 50, no. 12, pp. 2467–2473, Dec. 2003.

[11] C. J. Wordelman and U. Ravaioli, "Integration of a particle–particle–particle–mesh algorithm with the ensemble Monte Carlo method for the simulation of ultra-small semiconductor devices," *IEEE Trans. Electron Devices*, vol. 47, no. 2, pp. 410–416, Feb. 2000.

[12] M. V. Fischetti, S. E. Laux, and E. Crabbé, "Understanding hot-electron transport in silicon devices: Is there a shortcut?" *J. Appl. Phys.*, vol. 78, no. 2, pp. 1058–1087, Jul. 1995.

[13] M. Depas, B. Vermeire, P. W. Mertens, R. L. Van Meirhaeghe, and M. M. Heyns, "Determination of tunnelling parameters in ultra-thin oxide layer poly-Si/SiO$_2$/Si structures," *Solid State Electron.*, vol. 38, no. 8, pp. 1465–1471, 1995.

[14] P. Palestri *et al.*, "Comparison of modeling approaches for the capacitance–voltage and current–voltage characteristics of advanced gate stacks," *IEEE Trans. Electron Devices*, vol. 54, no. 1, pp. 106–114, Jan. 2007.

[15] L. Breuil *et al.*, "Nitride based FinFLASH memory device using multilevel hot carrier program/erase," in *Proc. Non Volatile Semiconductor Memory Workshop*, 2007, pp. 46–47.