# *Time-dependent analysis of low VDD program operation in double-gate SONOS memories by full-band Monte Carlo simulation*

**Gino Giusi**
Dipartimento di Elettronica Informatica e Sistemistica, Università degli Studi della Calabria
**Giuseppe Iannaccone**
Dipartimento di Ingegneria dell'Informazione: Elettronica, Informatica, Telecomunicazioni, Università di Pisa
**U.Ravaioli**
University of Illinois at Urbana-Champaign, Urbana, Illinois

G. Giusi, G. Iannaccone, U. Ravaioli, *Time-dependent analysis of low VDD program operation in double-gate SONOS memories by full-band Monte Carlo simulation*, Journal Applied Physics, **106**, pp.104506-1 – 104506-8 (2009).

# Time-dependent analysis of low $V_{DD}$ program operation in double-gate SONOS memories by full-band Monte Carlo simulation

G. Giusi,[1] G. Iannaccone,[2,a)] and U. Ravaioli[3]
[1]*DEIS, University of Calabria, Via P. Bucci 41C, I-87036 Arcavacata di Rende (CS), Italy*
[2]*DIIEIT, University of Pisa, Via Caruso 16, I-56126 Pisa, Italy*
[3]*University of Illinois at Urbana-Champaign, Urbana, Illinois 61801 USA*

(Received 15 July 2009; accepted 6 October 2009; published online 17 November 2009)

In this paper, we investigate warm electron injection (WEI) as a mechanism for NOR programming of double-gate SONOS memories through two dimensional (2D) full-band Monte Carlo simulations. WEI is characterized by an applied $V_{DS}$ smaller than 3.15 V, so that electrons cannot easily accumulate a kinetic energy larger than the height of the $Si/SiO_2$ barrier. We perform a time-dependent simulation of the program operation where the local gate current density is computed with a continuum-based method and is adiabatically separated from the 2D full Monte Carlo simulation used to obtain the electron distribution in the phase space. Trapping and detrapping from the nitride layer is taken into account by using a simplified Shockley–Read–Hall model. In this way, we are able to compute the time evolution of the charge stored in the nitride layer and of the threshold voltages corresponding to forward and reverse biases. We show that WEI is a viable option for NOR programming in order to reduce power supply and preserve reliability and complementary metal-oxide-semiconductor logic level compatibility. With the limitations of our adopted physical model, our results confirm the experimental observation showing that WEI provides a well localized trapped charge and offers interesting perspectives for multilevel and dual bit operation, even in devices with negligible short channel effects. © *2009 American Institute of Physics*. [doi:10.1063/1.3259409]

## I. INTRODUCTION

Multiple-gate and discrete-storage nonvolatile memories (NVMs) offer the combined advantages of improved retention due to the suppression of stress-induced leakage currents, improved short channel effects, and reduced intercell capacitive coupling. These aspects make them particularly promising for aggressive downscaling into the nanoscale regime and justify a significant research effort.[1–4] However, reliability concerns may limit the maximum longitudinal electric field and therefore the maximum applicable drain voltage $V_{DS}$ during NOR programming. Interestingly, experiments suggest that "warm electron injection," where electrons cannot accumulate kinetic energies higher than the $Si/SiO_2$ barrier height ($V_{DS}<3.15$ V), can represent a reasonable option for NVM programming.[4–6] In a previous work,[7] we studied the problem of warm electron injection in SONOS memories showing a strong dependence of the injected current on the drain voltage in the initial phase of the program operation, when the nitride layer is neutral. In this work, we largely extend the scope of Ref. 7 by investigating the time-dependent program operation in the warm electron injection regime and the dynamic trapping and detrapping of electrons in the silicon nitride layer. The simulation methodology helps us draw conclusions useful from a device design point of view.

Charge transport in nanoscale field-effect transistors biased in far from equilibrium conditions can be accurately modeled by solving the Boltzmann transport equation (BTE) through full-band Monte Carlo (MC) simulation. On the other hand, the tunneling gate current is several orders of magnitude smaller than the drain current, and the calculation of its profile along the channel poses a tremendous challenge to particle-based methods. Attempts to solve the BTE for the gate current problem have been made.[8,9] An energy transport model and a MC approach were successfully applied to gate current calculations for the case of hot carrier injection.[10–13]

In this work, we propose a simulation methodology to calculate the injected gate current density based on adiabatically decoupling the relatively slower process of gate injection from the faster process of electron transport in the metal oxide semiconductor field-effect transistor channel, which is modeled by two dimensional (2D) full-band MC simulation.[14] Trapping and detrapping occurring in the nitride layer are taken into account by using a simplified Shockley–Read–Hall (SRH) model.

Approaches to the simulation of the time evolution of the charge injected into the nitride layer are found in the literature. For example, in Ref. 15, only hot electrons are considered and all the injected charge is considered trapped. In Ref. 16, a trapping-detrapping model includes thermal excitation as the main discharge mechanism. Reference 17 proposes a method to accelerate the iterative MC procedure. In Ref. 18, the stored charge is evaluated as the difference between the injected charge and the charge emitted via the Poole–Frenkel effect.

In this work, we use the MC approach to calculate charge density and electrostatic potential in the device. The transmission coefficient for each point of the silicon/silicon

a)Electronic mail: g.iannaccone@iet.unipi.it.

 

oxide interface is calculated using the WKB approximation. Quantum confinement in the channel and barrier lowering are not considered. Differing from the cases mentioned above, we calculate gate injection contributions also for "warm" carriers, whose kinetic energy is lower than the barrier height and therefore tunnel through the gate oxide

By studying gate charge injection during the programming transient, we are able to gather insights into the evolution of the trapped charge. This information is particularly important for dual bit operation where physical charge localization is used to store more than 1 bit per cell.

Simulation results show that injection is effective also for low drain bias due to the very strong dependence of the gate current on $V_{DS}$. With the limitations of our adopted physical model, it is shown that charge injection is well localized, offering interesting perspectives for dual bit and multibit operation even in devices with reduced short channel effects such as multigate devices. Warm electron injection could be very useful for increasing reliability, reducing supply voltage and, hence, power dissipation.

The remainder of the paper is organized as follows. In Sec. II, the physical model of gate tunneling and of trapping/detrapping in the nitride layer is described. In Sec. III, the simulation methodology to calculate the time- and space-dependent gate current density and stored charge density is presented. An application of the method on double-gate SONOS memories is shown in Sec. IV. Conclusions are drawn in Sec. V.

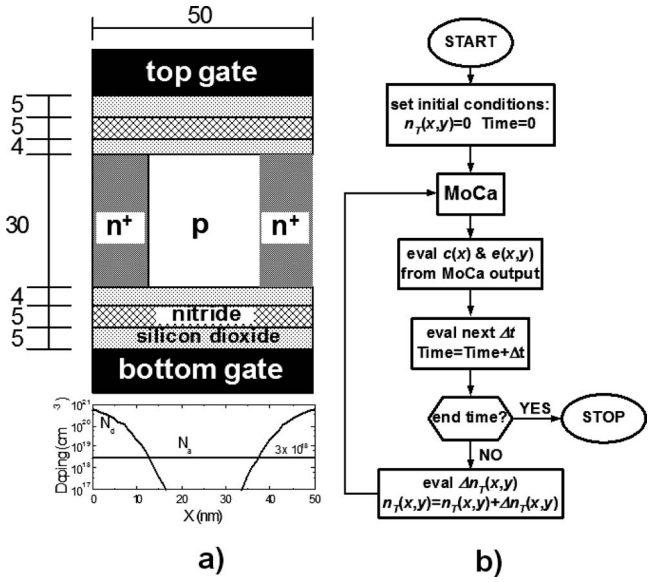


FIG. 1. (a) (Left) The simulated structure, an n-channel DG SONOS memory with a 50 nm channel length, and a 4/5/5 nm ONO stack. The acceptor fin doping is $3\times10^{18}$ cm$^{-3}$, while the source/drain doping extends under the gate for 15 nm from each side. $x$ is the channel direction, and $y$ is the direction of tunneling (perpendicular to the $Si/SiO_2$ interface). One interface is at $y=0$, $y>0$ is silicon, and $y<0$ is oxide. (b) (Right) Simulation flowchart. The electron energy distribution and potential are extracted by MC simulation and used to calculate the injected charge and capture and emission coefficients. The SRH equation rate [Eq. (8)] is solved with a convenient choice of an adaptive time step $\Delta t$. Next, the total charge into the nitride is updated with the trapped charge obtained from Eq. (8). The cycle restarts with the next MC simulation until the desired programming time is reached.

## II. PHYSICAL MODEL

In this section, the adopted physical model is presented. In Sec. II A, we discuss tunneling of electrons from the channel through the $Si/SiO_2$ interface, while in Sec. II B we discuss the physical model of charge trapping-detrapping in the silicon nitride layer.

### A. Physical model of electron tunneling

As stated in Sec. I, we consider not only "hot electrons" with a total kinetic energy higher than the $Si/SiO_2$ interface barrier height $B=3.15$ eV but also warm electrons with lower kinetic energies that provide the major contribution to the charging current for low $V_{DS}$ programming. The tunneling model has been introduced in Ref. 7, and we report it here in more detail for readers' convenience. Different from previous contributions in the literature, we need to compute the local tunneling current at each point of the $Si/SiO_2$ interface because injection is highly nonuniform in space and because charge trapping is localized. Our tunneling model is relatively simple: we assume that (i) total energy and transverse momentum are conserved during tunneling and that (ii) the dispersion relation in $SiO_2$ is parabolic with isotropic effective mass $m_{ox}$. We have verified in Ref. 7 that including the full dispersion relation of $SiO_2$ does not lead to relevant differences.

The considered structure is sketched in Fig. 1(a). In the diagram, $x$ is the channel direction and $y$ is the direction of tunneling (perpendicular to the $Si/SiO_2$ interface). In the following, we refer to the interface positioned at $y=0$, with $y>0$ for silicon and $y<0$ for oxide. $E$ is the total carrier energy, and $(k_x,k_z)$ is the transverse wave vector. If $E_{kin}$ is the kinetic energy of the particle at the interface on the silicon side at a given $x$, then the kinetic energy of the particle at the interface on the oxide side for the same $x$ $E_{ox\_kin}$ is

$$E_{ox\_kin}=E_{kin}-B. \tag{1}$$

The dispersion relation in the oxide is assumed to be

$$E_{ox\_kin}=\frac{\hbar^2}{2m_{ox}}(k_{ox\_x}^2+k_{ox\_y}^2+k_{ox\_z}^2), \tag{2}$$

where $k_{ox\_x}$, $k_{ox\_y}$, and $k_{ox\_z}$ are the components of the wave vector. We should note that in the case of tunneling, $E_{ox\_kin}$ is negative and $k_{ox\_y}$ is purely imaginary, while $k_{ox\_x}$ and $k_{ox\_z}$ are real because they are conserved during tunneling ($k_{ox\_x}=k_x$ and $k_{ox\_z}=k_z$). From Eq. (2), we obtain $k_{ox\_y}$, and the component of the kinetic energy in the oxide along the tunneling direction is

$$E_{ox\_y}=\frac{\hbar^2}{2m_{ox}}k_{ox\_y}^2=E_{ox\_kin}-\frac{\hbar^2}{2m_{ox}}(k_x^2+k_z^2). \tag{3}$$

The component of the kinetic energy contributing to tunneling, $E_y$, can be separated as

$$E_y=E_{kin}-\frac{\hbar^2}{2m_{ox}}(k_x^2+k_z^2), \tag{4}$$

so that the effective barrier height is identified as $\phi_s=B-E_y$.

We consider for tunneling only particles that are at the $Si/SiO_2$ interface and have a positive velocity $v_y$ in the tunneling direction. During the MC solution of the BTE, for each particle, we can calculate $E_y$ from Eq. (4) and compute the distribution $n(x,E_y)$, where $n(x,E_y)dE_y$ is the density per unit volume of electrons in $y=0$ that have a component of the kinetic energy between $E_y$ and $E_y+dE_y$. The tunneling current density can be calculated using the formula

$$J_G(x) = q\int v_y n(x,E_y)T(x,E_y)dE_y. \quad (5)$$

References 19 and 20 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 $T(x,E_y)$, as in Ref. 19.

### B. Physical model of trapping-detrapping

SONOS memories operate by storing charge in localized states in the nitride layer of the oxide-nitride-oxide (ONO) stack. At present, no clear consensus exists about the nature and the distribution in energy and space of nitride traps. As the reader can imagine, this information is very important to model the stored charge, which in turn influences the threshold voltage shift and thus the "stored information." There is common consensus on the fact that nitride traps have an amphoteric behavior, which is a neutral state $D^0$ when they are filled with one electron, a negative state $D^-$ when they are filled with two electrons, and a positive state $D^+$ when they are filled with a hole. During capturing, traps are in state $D^0$ with an energy $E_{TB}$, while during emission, traps are in state $D^-$ with energy $E_{TA}$ [energies are positive and represent the distance to the bottom of the nitride conduction band (CB)]. In the "positive correlation energy model,"[21] $E_{TB} > E_{TA}$; that is, traps are deeper in energy when they are in the $D^0$ state than when they are in the $D^-$ state. On the other hand, in the "negative correlation energy model," $E_{TB} < E_{TA}$.[22] Here, we prefer to use the first approach. Nevertheless, we will show in Sec. IV that in our particular case the choice of the correct value of energy parameters practically does not influence the final result.

Typically, a trap profile is determined by experiments and simulation fitting,[21,23–30] as a uniform trap distribution in space and a constant trap energy between 0.8 and 1.4 eV below the nitride CB. Here, we make the same assumption considering a monoenergetic level $E_{TA}=1.0$ eV below the nitride CB for emission and $E_{TB}=2.0\pm0.4$ eV below the nitride CB for capture, as suggested in Refs. 21 and 24–26.

Generation-recombination in the nitride is governed by SRH generation-recombination.[31] Generally, traps can be filled or emptied by capture and emission processes of electrons and holes. We neglect hole contribution during programming because (a) in the gate there are few holes that can tunnel into the nitride, (b) the valence band shift (4.7 eV) is higher than the CB shift (3.15 eV) at the $Si/SiO_2$ interface.

The processes considered here are illustrated by Fig. 2. Traps in the nitride can be filled by channel electrons tunneling through the $Si/SiO_2$ barrier. These electrons, depending on their kinetic energy $E_y$ along the tunnel direction at the interface, can be trapped directly (process B in the figure) or indirectly by thermal recombination (process A+C). Detrapping is due to two processes:[23] thermal emission of electrons into the CB (process D) and trap-to-band tunneling of electrons directly into the gate CB (E). Other charge loss processes as band-to-trap tunneling, trap-to-trap tunneling, and Poole–Frenkel emission have been neglected, as in Refs. 21–23. We neglect also the redistribution of charge between nitride traps, which was found to be governed by Poole–Frenkel conduction[32] because this process is too slow with respect to the processes we have considered to be relevant during programming[23,32–34] and multiphonon assisted emission[35] because of the relatively low field in the nitride layer (see the discussion in Sec. IV).

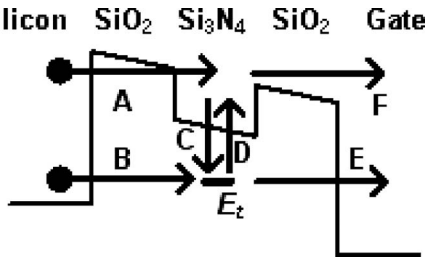


FIG. 2. Energy band diagram and mechanisms involved during program operation. (A) Electron injection from the silicon CB to the nitride CB. (B) Tunneling and capture from the silicon CB to the nitride traps. (C) Thermal recombination from the nitride CB to the nitride traps. (D) Thermal emission from the nitride traps to the nitride CB. (E) Trap-to-band tunneling from the nitride traps to the gate CB. (F) Electron injection from the nitride CB to the gate CB. Only A+C and E have been found to be not negligible in our simulation.

The nitride region is subdivided into spatial bins along the $x$ and $y$ directions. For each $(x,y)$ bin, the SRH equation is

$$\frac{dn_T(x,y)}{dt} = c(x)p_T(x,y) - e(x,y)n_T(x,y), \quad (6)$$

where $n_T$ is the concentration of occupied traps, $p_T$ is the concentration of free traps, $c(x)$ is the total capture rate, and $e(x,y)$ is the total emission rate. The total capture rate is due to the sum of the capture rates of processes A+C and B, while the total emission rate is due to the sum of the emission rates of processes D and E. The capture rate due to the thermal recombination (process A+C) can be calculated from the gate current density given by Eq. (5), $\sigma J_G(x)/q$, where $\sigma$ is the trap capture cross section. Figure 3 shows the CB diagram and the nitride trap energy at the bottom gate side for different programming times for the bias $V_{GS}=8$ V, $V_{DS}=2.8$ V. As can be seen, the capture trap energy level $E_{TD}$ $(2.0\pm0.4$ eV) remains always below the silicon CB edge so that direct trapping (process B) is not very significant in our case.

The emission rate due to trap-to-band tunneling (process E) is given by $e_{TBT0}T_G(x,y)$, where $e_{TBT0}$ is the "attempt-to-escape frequency" and $T_G(x,y)$ is the transmission probabil-

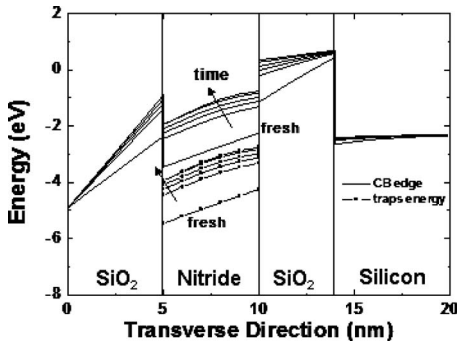


FIG. 3. CB diagram and trap energy at the bottom gate side for different programming times (fresh and $10^{-9}$, $10^{-8}$, $10^{-7}$, $10^{-6}$, and $10^{-5}$ s) for the biases $V_{GS}$=8 V and $V_{DS}$=2.8 V. The capture trap energy level $E_{TD}$ (2.0 eV) remains always below the silicon CB edge so that direct trapping (process B) is not important in our case, and the total capture rate is due only to thermal recombination.

ity of the barrier from the trap position to the control gate CB. The emission rate due to thermal emission (process D) can be written as[23]

$$e_{th} = \frac{\sigma v_{th} N_C(m_n^*)}{\exp[E_{TD}/kT]}, \tag{7}$$

where $N_C$ is the effective density of states in the nitride, which is a function of the nitride effective mass ($m_n^*$), and $v_{th}$ is the thermal velocity. Here, we assume $\sigma$ to be equal to $5\times10^{-13}$ cm$^2$,[21] so that $e_D \approx 5\times10^{-4}$ s$^{-1}$ (this rate is not a function of the trap position). As will be clear in Sec. IV, thermal emission rate is much lower with respect to the trap-to-band emission rate, which dominates the emission process.

A complete charge trapping-detrapping model should also include a transport model in the nitride.[36] This is a very complicated task, and many parameters such as the trap density $N_T$, the trap cross section, energy relaxation time for electrons, and electron mobility in silicon nitride have to be known accurately. As stated before, several different values can be found in the literature for these parameters, which in the end are all extracted through fitting with experiments: typical extracted values for $N_T$ are in the range $10^{19}-10^{20}$ cm$^{-3}$, and those for $\sigma$ are in the range $5\times10^{-12}-5\times10^{-13}$ cm$^2$.[21,24–27]

We believe that in our case we can make a reasonable simplification to strongly reduce the number of free parameters, assuming that all injected charge is uniformly trapped along $y$ in the silicon nitride layer. Moreover, we neglect a short-range Coulomb interaction between trapped charges and electrons and the hot electron lateral shift from the injection point during thermalization.[37]

Our assumption would not hold in general but it is based on the fact that in our case electrons are injected into the nitride layer with a relatively low kinetic energy and can lose much of it before reaching the high control oxide barrier and being reflected. The assumption of uniform trapping of injected electron in the layer thickness is acceptable for a very thin layer as in our case, and it allows us not to consider transport in the nitride layer explicitly. Actually, if the layer thickness is much smaller than the channel length, the details of the charge distribution along $y$ have a negligible effect on the electrostatics in the channel. Therefore, we can substitute $\sigma N_T$ with the inverse of the nitride layer thickness $t_N$ in Eq. (6) in order to obtain

$$\frac{dn_T(x,y)}{dt} = \frac{J_G(x)}{qt_N} - e_{TBT0} T_G(x,y) n_T(x,y). \tag{8}$$

In our trapping-detrapping model, now only two free parameters remain, $E_t$ and $e_{TBT0}$, which determine the behavior of the trap-to-band tunneling process.

In the pioneering work of Lundkvist[38] on charge loss in SONOS, the attempt-to-escape frequency was expressed as $e_{TBT0}=E_t/h$ (where $h$ is Planck's constant), and the transmission coefficient $T_G$ was calculated by assuming a rectangular barrier. The former assumption would have a physical basis only if the trap energy considered was taken with respect to the bottom of the CB in a potential well and not with respect to the CB out of the well, as in this case (for $E_t$ =1 eV, one would obtain $e_{TBT0}=2.4\times10^{14}$ s$^{-1}$). Several values for $e_{TBT0}$ can be found in literature in the range $10^{12}-10^{14}$ s$^{-1}$.[23,25–27,35] Here, we prefer to use the value extracted in Ref. 26 from the comparison between retention experiments and simulations ($e_{TBT0}=2\times10^{12}$ s$^{-1}$).

The assumption of a rectangular barrier in the control oxide is not very realistic for a large electric field in the oxide. For this reason, we compute $T_G(x,y)$ in detail for generic barrier shapes with the WKB approximation.

## III. SIMULATION METHOD

The problem of calculating the charge trapped in the nitride can be addressed by solving the time-dependent SRH rate equation (6) and (8), where capture and emission rates depend on the occupation density $n_T$. In order to perform a time-dependent simulation, we adiabatically decouple transport in the channel from electron tunneling into and from the nitride because the former is a faster mechanism. In addition, we assume that the tunneling current is a negligible fraction of the drain current, and we do not consider it explicitly when computing transport properties in the channel.

The simulation flowchart is shown in Fig. 1(b). Simulation time is broken into several time steps. Between two successive steps, electronic distribution and potential at the interface are computed with the 2D MC simulator, MoCa,[14] and they are used to calculate the new capture and emission rates. Now, the occupation factor, $f_T=n_T/N_T$, of traps at the generic position $(x,y)$ can be calculated by an explicit forward integration of Eq. (6) [(8)]:

$$f_T(t+\Delta t) = f_T(t) + \left(\frac{c}{c+e} - f_T(t)\right)[1-e^{-\Delta t/\tau}], \tag{9}$$

where $t$ is the time, $\Delta t$ is the time step, and $\tau=(c+e)^{-1}$ is the characteristic time of the charge-discharge process. Obviously, $c$ and $e$ are functions of $t$ and $\Delta t$, which has to be small enough so that $c$ and $e$ are practically constant during $\Delta t$.

In Eq. (9), it is apparent that trap occupation increases if $c\tau > f_T(t)$; otherwise, it decreases. During programming, trapping reduces the number and the average kinetic energy

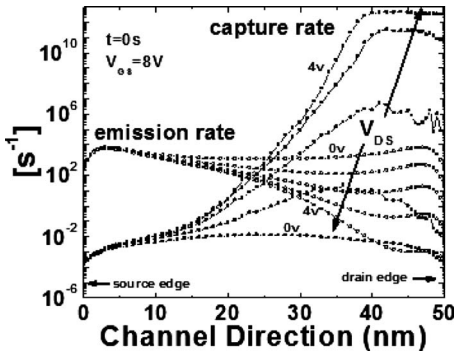


FIG. 4. Capture and emission rates at the top oxide/nitride interface along the channel and for different $V_{DS}$ values (0, 1, 2, 3, and 4 V) at the beginning of the injection ($t=0$ s). Charge trapping is possible only on the drain side where the capture rate is several orders of magnitude higher than the emission rate.

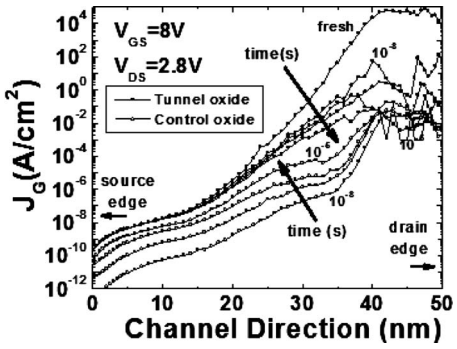


FIG. 5. Current density along the channel through the tunnel oxide (in) and through the control oxide (out) for increasing time during the program operation. As time progresses, capture and emission rates become comparable at the drain side and charge injection saturates.

of charge carriers so that $c$ reduces also. On the other hand, the emission rate $e$ increases with increased trap filling; therefore, at one point of the charge process, the condition $c>e$ is not true and the trapping process ends.

Because during each MC simulation step the electrostatic potential is considered constant, during the time step $\Delta t$ the occupation factor of all traps must have a negligible change. We choose an adaptive $\Delta t$ ten times smaller than the smallest value of $\tau$ for all interface points on the $(x,y)$ plane: $\Delta t=\min_{x,y}(\tau)/10$. We have found that this small value also ensures convergence of the forward time integration.

Each time the occupation factor is computed, the total charge in the nitride layer is updated, and the simulation continues with the next MC time step until the desired programming time is reached.

As discussed in Ref. 7, our choice of computing the particle distribution with a 2D MC simulator raises the issue of the correct evaluation of electron-electron interaction. We are aware of the fact that there are accurate approaches to naturally include a short-range particle-particle interaction in MC simulation with a three dimensional (3D) solver, as implemented, for instance, in the simulator developed by one of the authors of this work.[39] The cost of 3D solutions remains, however, prohibitive, and in order to reduce the computational complexity of the problem, we had to limit our approach to the 2D full-band version of MoCa, where we approximately take into account particle-particle interaction by self-consistently solving the Poisson equation on a relatively fine grid, without explicitly introducing electron-electron scattering mechanisms.[40]

## IV. SIMULATION RESULTS

To validate the proposed simulation method and the physical model, we used the same test structure considered in Ref. 7: an n-channel double gate (DG) SONOS memory with a 50 nm channel length and a 4/5/5 nm ONO stack. The acceptor fin doping is $3\times10^{18}$ cm$^{-3}$, and the source/drain doping extends under the gate for 15 nm on each side [Fig. 1(a)].

In Fig. 4, we plot local capture and emission rates at $t=0$ s for a fixed $V_{GS}=8$ V and for different $V_{DS}$. The capture rate has been calculated by assuming a capture cross section $\sigma$ of $5\times10^{-13}$ cm$^2$,[21] and the shown emission rate (which depends on $y$) is the one at the interface between nitride and the control oxide. As stated in Sec. II, the overall emission rate is much higher with respect to the calculated thermal emission rate so that the emission process is dominated by the trap-to-band tunneling. In Fig. 4, it is also clear that multiphonon emission has a negligible role because of the relative low nitride field ($<1$ MV/cm) and the relative low temperature ($T=298$ K). In fact, Ref. 35 shows that for a field of 1 MV/cm and $T=573$ K, $dn_T/dt$ can be at most $10^{14}$ cm$^{-3}$ s$^{-1}$ for traps close to the injecting interface. For a nitride trap density of $10^{19}$ cm$^{-3}$, the corresponding multiphonon emission rate would be $10^{-5}$ s$^{-1}$, which is much lower than the calculated trap-to-band emission rate.

It is apparent that rates are very sensitive to $V_{DS}$ and that carrier injection is localized at the drain side. Moving toward the source, the capture rate becomes smaller than the emission rate. But since the occupation factor of traps near the source is negligible, the emission current is practically negligible with respect to the capture current. Indeed, as shown in Fig. 7, trapping occurs also at the source, but the stored charge is several orders of magnitude lower than the charge stored at the drain side. The storing of charge at the source side continues as the programming time increases. Here, the limiting factor in charge trapping is the low capture rate and not the high emission rate because charge trapping at the source side does not show any saturation.

As time progresses, capture and emission processes become comparable at the drain side and charge injection saturates, as can be seen in the plot of the tunneling currents through the two oxides as a function of the program time (Fig. 5). It can be seen in Fig. 6 that the maximum of the injected current density, for a fixed $V_{GS}=8$ V, decreases as $1/t$, with a behavior independent of $V_{DS}$. This is due to the fact that charge trapping near the drain, which imposes the maximum $J_G$, is essentially limited by the difference between the average energy of hot carriers at the drain (imposed by $V_{DS}$) and the effective barrier height (determined by the trapped charge). For this reason, if we use a larger $V_{DS}$ and start programming, we will reach a time at which the trapped charge reduces $J_G$ at the drain ($\sim J_{G\max}$) at the level obtained with a smaller $V_{DS}$ and uncharged nitride. From there on, the

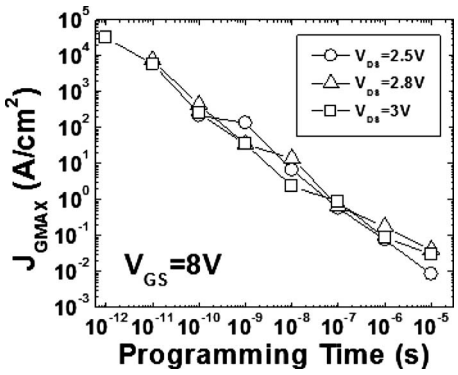


FIG. 6. The maximum of $J_G$ as a function of the programming time is quite independent of the drain bias and has roughly a $1/t$ behavior in a log-log scale. This behavior is attributed to a balance between a higher injection field and a lower trapping probability.

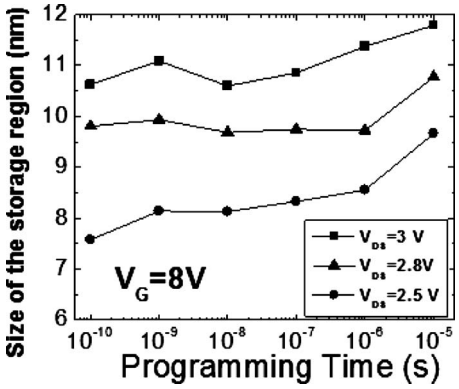


FIG. 8. Effective injection length as a function of the programming time for different values of $V_{DS}$. As one can see, the stored charge is localized in a region of a length of about 10 nm on the drain side, and it does not change significantly during programming operation but it increases slightly for higher $V_{DS}$ values.

time-dependent behavior of $J_{G\max}$ for the two values of $V_{DS}$ is identical since it is mainly determined only by the charging dynamics, which in turn is determined by $J_{G\max}$. Of course, even if $J_{G\max}$ has the same behavior as a function of time for different values of $V_{DS}$, more charge is trapped for the same total programming time for larger $V_{DS}$ because the starting $J_{G\max}$ was higher.

During programming, the trapped charge (Fig. 7) remains rather localized on the drain side because far from the drain tunneling currents are relatively very small (Fig. 5) and the variation in the trapped charge is negligible. Let us highlight that we neglected a short-range Coulomb interaction between charged traps and electron charges, which may produce a lower value of the trapped charge and a higher lateral spread.

We can define the effective size of the charge storage region as the ratio of the total stored charge per unit length to the peak charge density per unit area. This quantity is plotted in Fig. 8. As one can see, the stored charge is localized in a region of length of about 10 nm on the drain side, and it does not change significantly during programming operation but slightly increases for higher $V_{DS}$ values.

Figures 9 and 10 show the threshold voltage shift, calculated by a constant current gate voltage shift in a subthreshold region using Medici, for different gate and drain bias values. "Forward read" indicates that during the read phase, the drain and source contacts are the same as those used for programming. "Reverse read" indicates that during the read phase the drain and source contacts are exchanged with respect to the program. The read voltage is 0.4 V. The voltage window between the threshold voltages in reverse and forward read observed in Fig. 9 favors the possibility of using warm electron injection for dual bit cell operation. This is also confirmed by experiments on a structure similar to what we simulated.[6] Figure 9 also shows the threshold voltage shift obtained for $E_{TA}$=2.0 eV corresponding to the negative energy trap correlation model and the case of no emission for $V_{DS}$=2.8 V. Data for $E_{TA}$=2.0 eV and no emission are totally overlapped, showing that the trap-to-band transmission coefficient is very low for $E_{TA}$=2.0 eV. Moreover, data for $E_{TA}$=1.0 eV and $E_{TA}$=2.0 eV show that emission is not important in our case and that trapping saturation is due to the increased local threshold during injection.

Moreover, Fig. 10 (where the drain bias is held constant and the gate bias is changed) emphasizes that the programming time for a given required $V_T$ shift is reduced down to two orders of magnitude for an increase of 1 V in $V_{GS}$, and for the same programming time the $V_T$ shift increases by

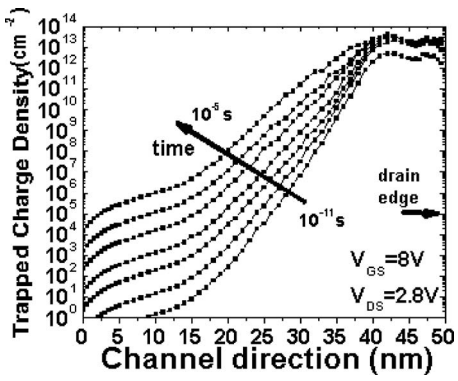


FIG. 7. Trapped charge density in the nitride layer along the channel as a function of the programming time. During programming, the trapped charge remains rather localized on the drain side because far from the drain tunneling currents are relatively very small.

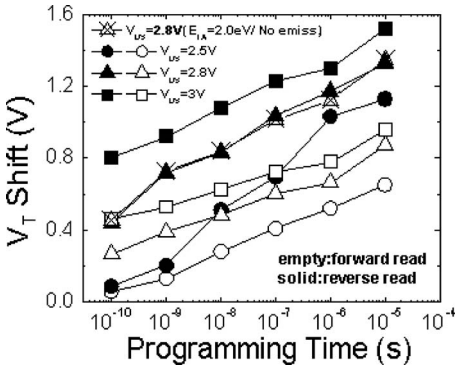


FIG. 9. Threshold voltage displacement in forward and reverse read (read voltage is 0.4 V) as function of the programming time for a fixed $V_{GS}$ =8 V and for different $V_{DS}$. It is evident that dual bit operation is possible also with $V_{DS}<3.15$ V. Also shown in the picture is the reverse threshold voltage shift for the case of a trap energy of 2.0 eV and for the case of no emission.

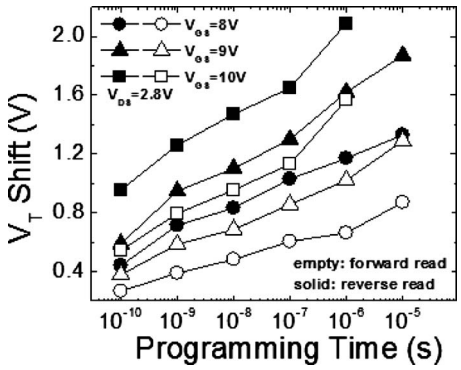


FIG. 10. Threshold voltage displacement in forward and reverse read as a function of the programming time for a fixed $V_{DS}$=2.8 V and for different $V_{GS}$. Multilevel operation seems to have a larger programming window with respect to multibit operation.

0.3–0.4 V for each volt of increase in $V_{GS}$. Also, this result is in agreement with experiments reported in Ref. 6. Moreover, by comparing Fig. 9 with Fig. 10, it seems that multilevel operation (the storing of different charge levels in the same physical position in the nitride layer obtained by changing the gate bias) has a larger programming window with respect to multibit operation (the storing of the same charge in different physical regions obtained by changing the drain bias). Let us stress the fact that the comparison between our simulations and Ref. 6 can only be qualitative because we are considering double-gate devices, while experiments in Ref. 6 regard trigate devices where injection at the corners (especially for high fields) can be dominant.[41–43]

## V. CONCLUSION

We have developed a simulation methodology based on a MC device simulator to investigate the warm electron injection programming regime ($V_{DS}<3.15$ V) of NOR double-gate SONOS based on the adiabatic separation of electron transport in the channel with respect to trapping/detrapping in the ONO layer. The gate current is calculated as a post-processing step of the MC simulation by a continuum-based method in conjunction with the particle-based method used to compute particle distributions and transport in the channel. Warm electron injection emerges as a viable option for NOR programming, preserving reliability, power dissipation, and complementary metal-oxide-semiconductor logic level compatibility at the cost of a slower programming time with respect to hot electron operation. We have also shown that the stored charge is well localized, providing a significant forward-reverse threshold voltage window both for multibit and for multilevel operation. This last aspect requires further investigation since in order to keep the computational resources required under control, we have adopted a model that does not consider transport in the nitride and short-range Coulomb interaction and can therefore underestimate charge spreading.

## ACKNOWLEDGMENTS

This work was supported by the FIRB Project No. RBIP06YSJJ and by the RIFLASH project of the Italian Ministry for Foreign Affairs.

[1] M. H. White, Y. L. Yang, A. Purwar, and M. L. French, IEEE Trans. Compon., Packag. Manuf. Technol., Part A **20**, 190 (1997).
[2] M. H. White, D. A. Adams, and J. Bu, IEEE Circuits Devices Mag. **16**, 22 (2000).
[3] F. Hofmann, M. Specht, U. Dorda, R. Kommling, L. Dreeskornfeld, J. Kretz, M. Stadele, W. Rosner, and L. Risch, Solid-State Electron. **49**, 1799 (2005).
[4] 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, Tech. Dig. - Int. Electron Devices Meet. **2007**, 921.
[5] J. Razafmdramora, L. Perniola, C. Jahan, P. Scheiblin, M. Gely, C. Vizioz, C. Carabasse, F. Boulanger, B. De Salvo, S. Deleonibus, S. Lombardo, and C. Bongiorno, ESSDERC, 2007, pp. 414–417.
[6] L. Breuil, M. Rosmeulen, A. Cacciato, J. Loo, A. Furnémont, L. Haspeslagh, and J. Van Houdt, NVSMW, 2007 (unpublished), pp. 46 and 47.
[7] G. Giusi, G. Iannaccone, M. Mohamed, and U. Ravaioli, IEEE Electron Device Lett. **29**, 1242 (2008).
[8] Z. Han, C. Lin, N. Goldsman, I. Mayergoyz, S. Yu, and M. Stettler, Simulation of Semiconductor Processes and Devices (SISPAD), 1999 (unpublished), pp. 247–250.
[9] H. Lin and J. Peng, University/Government/Industry Microelectronics Symposium, 1995 (unpublished), pp. 193–196.
[10] C. M. Yih, G. H. Lee, and S. Chung, VLSI Technology, Systems, and Applications, 1995 (unpublished), pp. 127–130.
[11] A. Harkar, R. W. Kelsall, and J. N. Ellis, VLSI Des. **13**, 301 (2001).
[12] Y. Ohkura, C. Suzuki, H. Amakawa, and K. Nishi, Simulation of Semiconductor Processes and Devices (SISPAD), 2003 (unpublished), pp. 67–70.
[13] K. Hasnat, C.-F. Yeap, S. Jallepalli, S. A. Hareland, W.-K. Shih, V. M. Agostinelli, F. Al Tasch, and C. M. Maziar, IEEE Trans. Electron Devices **44**, 129 (1997).
[14] G. A. Kathawala, B. Winstead, and U. Ravaioli, IEEE Trans. Electron Devices **50**, 2467 (2003).
[15] R. Hagenbeck, S. Decker, P. Haibach, T. Mikolajick, G. Tempel, M. Isler, C. Jungemann, and B. Meinerzhagen, Simulation of Semiconductor Processes and Devices (SISPAD), 2006 (unpublished), pp. 322–325.
[16] Y. Song, Z. Xia, J. Yang, G. Du, J. Kang, R. Han, and X. Liu, ICSICT, 2006 (unpublished), pp. 772–774.
[17] G. Kathawala, T. Thurgate, Z. Liu, M. Kwan, M. Randolph, and Y. Sun, NVSMW, 2007 (unpublished), pp. 106–109.
[18] C. H. Lee, C. W. Wu, S. W. Lin, T. H. Yeh, S. H. Gu, K. F. Chen, Y. J. Chen, J. Y. Hsieh, I. J. Huang, N. K. Zous, T. T. Han, M. S. Chen, W. P. Lu, T. Wang, and C. Y. Lu, NVSMW, 2008 (unpublished), pp. 109–110.
[19] M. Depas, B. Vermeire, P. W. Mertens, R. L. Van Meirhaeghe, and M. M. Heyns, Solid-State Electron. **38**, 1465 (1995).
[20] P. Palestri, N. Barin, D. Brunel, C. Busseret, A. Campera, P. A. Childs, F. Driussi, C. Fiegna, G. Fiori, R. Gusmeroli, G. Iannaccone, M. Karner, H. Kosina, A. L. Lacaita, E. Langer, B. Majkusiak, C. Monzio Compagnoni, A. Poncet, E. Sangiorgi, L. Selmi, A. S. Spinelli, and J. Walczak, IEEE Trans. Electron Devices **54**, 106 (2007).
[21] Y. Yang and M. H. White, Solid-State Electron. **44**, 949 (2000).
[22] V. A. Gritsenko, Yu. N. Novikov, A. V. Shaposhnikov, H. Wong, and G. M. Zhidomirov, Phys. Solid State **45**, 2031 (2003).
[23] Y. Wang and M. H. White, Solid-State Electron. **49**, 97 (2005).
[24] T. H. Kim, J. S. Sim, J. D. Lee, H. C. Shin, and B. G. Park, Appl. Phys. Lett. **85**, 660 (2004).
[25] H. Aozasa, I. Fujiwara, A. Nakamura, and Y. Komatsu, Jpn. J. Appl. Phys., Part 1 **38**, 1441 (1999).
[26] A. Arreghini, N. Akil, F. Driussi, D. Esseni, L. Selmi, and M. J. Duuren, ESSDERC, 2007 (unpublished), pp. 406–409.
[27] S. H. Gu, C. W. Hsu, T. Wang, W. P. Lu, Y. H. J. Ku, and C. Y. Lu, IEEE Trans. Electron Devices **54**, 90 (2007).
[28] A. Campera, G. Iannaccone, and F. Crupi, IEEE Trans. Electron Devices **54**, 83 (2007).
[29] G. Iannaccone, F. Crupi, B. Neri, and S. Lombardo, IEEE Trans. Electron Devices **50**, 1363 (2003).

[30]G. Iannaccone, F. Crupi, B. Neri, and S. Lombardo, Appl. Phys. Lett. **77**, 2876 (2000).
[31]W. Shockley and W. T. Read, Phys. Rev. **87**, 835 (1952).
[32]A. Furnemont, M. Rosmeulen, J. Van Houdt, K. De Meyer, and H. Maes, ESSDERC, 2006 (unpublished), pp. 447–450.
[33]H. Pang, L. Pan, L. Sun, D. Wu, and J. Zhu, International Conference on Solid State Devices and Materials, Yokohama, 2006 (unpublished), pp. 988–989.
[34]A. Furnémont, M. Rosmeulen, K. Van der Zanden, J. Van Houdt, K. De Meyer, and H. Maes, IEEE Trans. Electron Devices **54**, 1351 (2007).
[35]M. Hermann and A. Schenk, J. Appl. Phys. **77**, 4522 (1995).
[36]E. Vianello, F. Driussi, P. Palestri, A. Arreghini, D. Esseni, L. Selmi, N. Akil, M. Van Duuren, and D. S. Golubovic, ESSDERC, 2008 (unpublished), pp. 107–110.
[37]D. Fuks, A. Kiv, Y. Roizin, M. Gutman, R. Avichail-Bibi, and T. Maximova, IEEE Trans. Electron Devices **53**, 304 (2006).
[38]L. Lundkvist, I. Lundstrom, and C. Svensson, Solid-State Electron. **16**, 811 (1973).
[39]C. J. Wordelman and U. Ravaioli, IEEE Trans. Electron Devices **47**, 410 (2000).
[40]M. V. Fischetti, S. E. Laux, and E. Crabbé, J. Appl. Phys. **78**, 1058 (1995).
[41]G. Fiori, G. Iannaccone, G. Molas, and B. De Salvo, Appl. Phys. Lett. **86**, 113502 (2005).
[42]G. Iannaccone, A. Trellakis, and U. Ravaioli, J. Appl. Phys. **84**, 5032 (1998).
[43]G. Fiori, G. Iannaccone, G. Molas, and B. De Salvo, IEEE Trans. Nanotechnol. **4**, 326 (2005).