Electrostatic Effect of Localized Charge in Dual Bit Memory Cells with DiscreteTraps

L. Perniola(1, 2,*), S. Bernardini(3), G. Iannaccone(1,6), B. De Salvo (4), G. Ghiaudino (2), P. Masson(3), C. Gerardi(5)

(1)Dipartimento di Ingegneria dell’Informazione, Università degli Studi di Pisa, Via Caruso, 56122 Pisa, Italy, *perniola@enserg.fr
(2)IMEP-CNRS/INPG, Avenue de Martyrs 32, 38016 Grenoble, France
(3)L2MP-Polytech – IMT Technopôle de Château Gombert, 13451 Marseille Cedex 20 France
(4)CEA-LETI, Avenue de Martigny 16, 38054 Grenoble, France
(5)STMicroelectronics, Catania, Italy
(6)IEIIT-CNR, Via Caruso, 56122 Pisa, Italy.

The following article, presented at ESSDERC 2004 conference, illustrates the capability of ATLAS in the validation and the comprehension of complex effects of new promising devices like non-volatile discrete traps memory devices.

Abstract

In this paper the electrostatic impact of Channel Hot Electron (CHE) injection in discrete-trap memories is quantitatively addressed. The dual bit behavior of the transfer characteristic during forward and reverse read of a written cell is thoroughly analysed with the help of an analytical model. Such model allows, for the first time, to estimate the effective charged portion of the discrete storage layer, \(L_2\), and the quantity of electrons, \(Q\), injected in the trapping sites from the experimental parameters of the \(I_d-V_g\) characteristics, the reverse-forward threshold voltage shift \(\Delta V_{RF}\) and the total threshold voltage shift \(\Delta V_{tot}\). The viability of this model is confirmed with tests performed on nanocrystal memories, under different bias conditions. These results are confirmed with the help of a 2D drift-diffusion commercial code (ATLAS-SILVACO).

1. Introduction

Channel Hot Electron injection is widely used as a standard writing method for non-volatile discrete-trap memory products [1]-[2]. It provides the opportunity to localise the charges injected in a small region of the trapping medium, and two-bit operation is achieved through multilevel storage [3]. The basic principle on which two-bit operation resides, is common to NROM memories and nanocrystal memories [4]-[5]. It is possible to trap charges near one junction (drain or source) with a programming stress, and read them in the reverse mode, compared to the programming, enhancing the electrostatic effect of these charges on the conductivity of the active channel (see Figure 1).

It has been shown in the literature [1]-[3]-[6] that the threshold voltage during the forward read, \(V_{th-F}\), is lower than the threshold voltage during the reverse read, \(V_{th-R}\), when the cell is polarised in the saturation regime.

This is due to the strong two dimensional effects near the charged junction. If the injected charge, near the drain, is completely screened by the high \(V_{ds}\) applied in forward read (which induces a long pinchoff region), the \(I_d-V_g\) characteristic results very close to the characteristic of the fresh cell. In this case we have a low \(V_{th-F}\). On the other hand, during the reverse read the high \(V_{ds}\) applied is not able to screen the effect of electrons and the conductivity of the active channel is lowered by the “bottleneck” near the low-voltage contact. In such a case we have a high \(V_{th-R}\) [2]-[6].

Continued on page 2...
With a simple 1D approach (Gradual Channel Approximation) [6] it is not possible to simulate such an asymmetric behavior. In this paper, with the help of a simple quasi 2D analytical approach, we are able to quantify for the first time, both the length of the charged region, \( L_2 \), and the number of injected electrons per unit area in the trapping sites, \( Q \), from the two quantities available from experiments, the total threshold voltage shift, \( \Delta V_{th-tot} \) (i.e., the difference between the threshold voltage in the reverse read, \( V_{th-R} \), and the threshold voltage of the fresh cell, \( V_{th-fresh} \)) and the reverse-forward threshold voltage shift, \( \Delta V_{th-F} \) (i.e. the difference between \( V_{th-R} \) and \( V_{th-F} \)).

The viability of the model is tested with 2D numerical simulations of a commercial TCAD code (ATLAS-SILVACO).

Main results from the theory are widely tested, under different bias conditions, on nanocrystal memories fabricated by ST-Microelectronics [4]. The influence of the bulk potential, as well as that of the drain-to-source potential, are carefully assessed in this work.

2. Quasi-2D Analytical Model

With a simple 1D approach it is not possible to describe properly the 2D effects near the charged region, which are the first cause of the dual bit behavior. For the first time, to our knowledge, we present an analytical approach, which starts from quasi 2D considerations [7].

The analytical approach describes the behavior of the surface potential \( \Psi \) along the channel of a MOSFET with non-uniform threshold voltage. The distribution of trapped charge may be in general a complicate function of the spatial coordinates. We assume that the charge distribution along the channel can be described by a step function which represents an “effective” distribution. As a result, we can distinguish two regions: one close to the source with a length \( L_1 \), where storage nodes are not charged (uncharged region); and the second one close to the drain with a length \( L_2 \) which is uniformly charged by \( Q \) electrons per unit area. In Figure 2 it can be seen how the written memory is divided in the two regions. The trapped charge raises the flat band voltage, \( V_{fb-2} \) of the charged region \( L_2 \). In particular the flat band voltage difference between the charged and the uncharged region is:

\[
\Delta V_{fb} = + \frac{Q}{C_t} \cdot \frac{1}{C_2}
\]

Where \( q \) is the electron charge, \( C_t \) is the top oxide capacitance per unit area.

In the following model, we do not consider charges over the drain junction, but, as already noted by [6], from Figure 3 it appears that they do not influence \( I_{V2} \) characteristics.

The model solves separately, in the two regions, the differential equations (2) and (3). Gauss’ law is applied to a transversal slice of thickness dy along y and cross section equal to the average depletion width \( X_{dep} \) (i=1,2) (see Figure 2), and allows us to write the flux of the electric field through the faces perpendicular to y in terms of the surface potential as \( (X_{dep}, \eta) d\Psi/dy \). where \( \eta \) is a fitting parameter. During the gate voltage sweep, frequently the uncharged region goes in strong inversion, while the charged region is still pinched-off. The presence of both inversion charge \( Q_{inv} \) and depletion charge at inversion \( Q_{dep} \) are carefully considered in Equation (3):

\[
\varepsilon_{in} \frac{X_{dep1}}{\eta} \frac{d^2 \Psi_x}{dy^2} + \varepsilon_{in} \frac{V_x - V_{fb1} - \Psi_x}{t_{inv}} = Q_{dep1} + Q_{inv} \quad (2)
\]

\[
\varepsilon_{in} \frac{X_{dep2}}{\eta} \frac{d^2 \Psi_x}{dy^2} + \varepsilon_{in} \frac{V_x - V_{fb2} - \Psi_x}{t_{inv}} = qN_{ab}X_{dep2} \quad (3)
\]
where $\epsilon_s$ is the electric permittivity of silicon (oxide), $V_g$ is the gate voltage, $t_{ox}$ is the equivalent oxide thickness, $N_{sub}$ is the doping level of the bulk. If a bulk voltage, $V_b$, is applied, as sketched in Figure 2, the drain and source potentials must be raised of $V_b$, setting at zero the potential of the bulk. The space charge and the inversion charge have the following expression:

$$Q_{d} = \sqrt{2} q N_{inf} \varepsilon_s \cdot (\Psi_{L-1} - V_b)$$

$$Q_{inf} = \frac{KT}{q} mC_{inf} \ln(1 +$$

$$+ \frac{C_{drp}}{mC_{inf}} \exp \left( \frac{q (V_g - V_{drp}) - q U_{min}}{KT} \right) )$$

Where $i=1,2, \psi_{L-}^{(i)}$ is the surface potential in the case the influence of drain and source junctions is completely neglected, $U_{min}$ is the quasi-Fermi potential referred to the source of the uncharged equivalent transistor.

The solution to Equation (2) is a linear combination of hyperbolic sins:

$$\Psi_{L-1}(y) = (\Psi_{L-1}^{(i)} - \Psi_{L-1}^{(i)}) \sinh(\frac{y}{\lambda_i}) +$$

$$+ (\Psi_{L}^{(i)} - \Psi_{L-1}^{(i)}) \sinh(\frac{y - L_i}{\lambda_i})$$

where $\Psi_{L-}^{(i)}$ is the right-hand (left-hand) boundary condition shown in Figure 2. $\Psi_{L}^{(i)}$ is referred to the value of the potential in the bulk. The solution to Equation (3) is of similar form of Equation (6).

The parameter $\lambda_i = \sqrt{\frac{\varepsilon_s}{\eta \lambda_i t_{ox}}} \frac{X_{drp}( \varepsilon_s, \eta )}$ describes the short channel effects, as it provides a clue on how strong is the influence of the junction potential on $\Psi_S$ under the gate. The higher the value of $\lambda_i$, the smoother appears the $\Psi_S$ curve along the channel, the weaker is the control effect of the gate on the channel.

From the description of the surface potential, the drain current can be calculated. As in the subthreshold region the current is a diffusion dominated process, the following expression can be derived [8]:

$$I_d = \mu_{eff} \frac{W}{L} \varepsilon_s q N_{inf} \left( \frac{KT}{q} \right)^{2} \left( \frac{N_{inf}}{N_{sub}} \right)^{2} \frac{d\Psi_{min}^{(i)}}{dx}$$

where $\mu_{eff}$ is the effective mobility of electrons in the channel, $KT$ is the thermal energy, $\eta$ is the intrinsic electron concentration in silicon and $\Psi_{min}$ is the minimum of the surface potential in the channel.

3. Comparison Between the Analytical Model and ATLAS Results

From Equation (2), the importance of $\eta$ is apparent.

To determine the proper value of this parameter, we have used the commercial TCAD tool ATLAS from SILVACO. We have simulated a class of nanocrystal memories, fabricated from STMicroelectronics. Parameters of the memory cell are $L=0.28 \mu m$, $W=0.16 \mu m$, tunnelling oxide thickness $t_{tun}=5.5 \text{nm}$, control oxide thickness $t_{c}=8 \text{nm}$. The charged nanocrystals have been considered as a uniform charged oxide region, of thickness $t_{ch}=2 \text{nm}$, sandwiched between the tunnelling oxide and the control gate oxide.

![Graph showing comparison between ATLAS and analytical model](image-url)
As specified before, we simply assumed that the charge distribution in the oxide is a step function where the charged length can be varied and the density of charge has been chosen to fit different flat band voltages $V_f'$. In Figure 4(a) a sample comparison between ATLAS results and the analytical model for $L_2=L/4$ is shown. The parameter $\eta$ to fit $I_d-I_g$ numerical characteristics has shown a weaker dependence on the density of electrons trapped in dots, $Q$, than on the charged length $L_2$. This property is shown in Figure 4(b), where, provided one value of $L_2$, it is possible to fairly fit different values of $Q$, with one value of $\eta$. In the inset, on the other hand, the dependence of $\eta$ on $L_2$ is provided (strong dependence of $\eta$ on $L_2$).

Already the plot of Figure 4(a) shows that the forward and reverse bits are not clearly detached for $L_2=L/4$ ($L_2=70$ nm), and cannot be easily sensed as $V_{th,F}$ and $V_{th,R}$ are similar. From Figure 5, it is clear that one-bit operation appears for $L_2=70$ nm ($\Delta V_{th}=0$V). It is not required that the whole region of the channel is uniformly charged to have $V_{th}=V_{th,F}$.

Another interesting property is the charge density insensitivity of $\Delta V_{RF}$ once a saturation limit is achieved. In Figure 6 such a behavior is shown for $L_2=L/8$ ($L_2=35$ nm) and for $L_2=L/4$ ($L_2=70$ nm). In the case of $L_2=L/4$, for $Q>2 \times 10^{12}$ el./cm$^2$ the window between the forward and reverse threshold voltage is almost constant; while in the case of $L_2=L/8$, this property appears for $Q>5 \times 10^{12}$ el./cm$^2$. Therefore $\Delta V_{RF}$ can be used as an absolute indicator of the charged region dimension.

4. Comparison Between the Analytical Model and Experiments

Experiments performed on STMicroelectronics nanocrystal memory cells [4] confirm both the validity of the analytical model and suggest how to perform a reading procedure which allows to enhance the asymmetry between forward and reverse read.

From Figure 7, it is clear that raising $V_d$ enhances the electrostatic effect of trapped electrons, and decreases $V_{th,F}$ to the value of the fresh cell threshold voltage, $V_{th,Fresh}$. In experiments the maximum reading $V_d$ has been put at 1.5V. In Figure 8, the viability of this analytical model with different bulk voltages, $V_b$, has been tested. The most relevant conclusion of this work is presented in...
Figure 8. Analytical model versus experiments obtained from nanocrystal memories with $V_{ds}=1.5\text{V}$ and different bias condition for $V_g$.

Figure 9. As an example it is shown a contour plot of $\Delta V_{tot}(L_2, Q)$ and $\Delta V_{RF}(L_2, Q)$ that can be calculated from this model. From the two experimental parameters $\Delta V_{es}$ (blue line) and $\Delta V_{es}$ (red line) is possible to find in the contour plot the crossing point and then deduce the values of $L_2$ and the density of electrons $Q$ injected in the trapping medium.

5. Conclusions

A detailed model concerning Channel Hot Electron (CHE) electrostatic impact on forward/reverse reading has been provided. Both numerical simulations and experiments on nanocrystal memories, performed under different bias conditions confirm the viability of the model. At our knowledge, for the first time in literature, a contour plot is provided (Figure 9) where, from the experimental results at hand ($\Delta V_{es}$ and $\Delta V_{es}$), it is possible to assess the effective charged length, $L_2$ and the density of injected electrons, $Q$. Support from EU projects ADA-MANT (IST-2001-34234), SINANO (IST-506844) and from the CNR through the FIRB project is gratefully acknowledged.

References