Embedded non–volatile memory study with surface potential based model

D. Garetto[§◇], A. Zaka[†‡], V. Quenette[†], D. Rideau[†], E. Dornel[△], O. Saxod[†], W. F. Clark[§]
M. Minondo[†], C. Tavernier[†], Q. Rafhay[‡], R. Clerc[‡], A. Schmid[◇], Y. Leblebici[◇], and H. Jaouen[†]

[§] IBM France, 850, rue Jean Monnet, Crolles, France
[◇] Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[‡] IMEP, MINATEC, 3 parvis Louis Néel, Grenoble, France
[†] ST Microelectronics, 850, rue Jean Monnet, Crolles, France
[△]IBM France, Crolles, France – Altran, Echirolles, France
Corresponding authors: david.garetto@fr.ibm.com, alban.zaka@st.com

ABSTRACT

Coupling coefficients calculation is known to be a critical issue in embedded Non-Volatile Memory (eNVM) compact modeling [1]. In this paper we have implemented the charge balance method within the Brew's Charge Sheet Model equation, determining the floating gate potential of the cell and deriving the NVM coupling coefficients. The results have been compared with TCAD simulations and demonstrate that short channel effects, overlap capacitances and velocity saturation dominate over the intrinsic behavior of the cell in scaled devices. We have studied the transient behavior of the eNVM, reproducing the V_{TH} shift due to charge injection and demonstrating the capability of our model to simulate the full electrical behavior of the device. Moreover we have included the charge balance equation in the VerilogA implementation of PSP MOSFET model and proved with SPICE simulations the suitability of our approach to compact modeling.

Keywords: embedded non–volatile memory, surface potential compact model, short channel effects, saturation velocity, overlap capacitance

1 INTRODUCTION

Embedded non–volatile memory (eNVM) compact models present in literature are generally well suitable for reproducing the behavior of a single flash cell in read conditions, but often they don't consider all the physical effects that can influence the performance of the device at the sub–micron scale. In order to enable IP designers implementing support circuits and analyzing the performance of the device (Program/Erase speed, endurance, data retention, reliability, etc), a more physical model of the cell is required. Among the eNVM compact models, both the capacitive network model (CNM) [1] and the charge balance model (CBM) [2] rely on the floating gate (FG) potential V_{F} calculation.

In CNM, the FG potential V_{F} is calculated from capacitive coupling coefficients between FG and the other terminals of the device, defined as $\alpha_{\mathrm{I}} = \frac{\partial V_F}{\partial V_I}$ (with I indicating each terminal of the device). This approach is often used in industry for its simplicity and computational efficiency. However, the experimental estimation of these coefficients is difficult and inaccurate, since it is usually performed from indirect DC measurements [3]. Compact models based on CNM suffer from the absence of a bias-dependent model for α_{I}, whose values strongly depend on the stored charge and on the applied biases, as demonstrated in [4] and in this work. For these reasons, the accurate modeling of aggressively scaled cells, such as embedded NVM, is compromised.

More recently, CBM models have been proposed, where V_{F} is found solving the charge balance equation at the FG node using an iterative bisectional algorithm [2]. This compact derivation of CBM that relies on BSIMv3 as core MOSFET model has been used by to study the bias and geometry dependence of α_{I} [4]. However, some problems can be identified in this approach: the BSIMv3 model generates unphysical discontinuities in the bias–dependency behavior of α_{I}, the complexity of the drain / source - gate overlap region is not taken into account and the impact of short channel effects (SCE) on charges and capacitances is not accurately reproduced.

In this paper we studied the behavior of the coupling coefficients with TCAD simulations. We implemented CBM in a semi–analytical charge sheet model (CSM) based on Brew's equation. Our model is accounting for SCE, overlap capacitances and saturation velocity effects on currents but also on charges [5]. In Section II, we present the flash structure studied by means of TCAD simulations. The semi–analytical NVM–CSM model is described and validated in Section III. Finally, in Section IV, a compact implementation of NVM–CSM based on PSP is shown.

2 TCAD SIMULATIONS

The structure studied in this paper is a single Flash cell with terminals D (drain), S (source), B (bulk), C (control gate) and F (floating gate) (Figure 1). The device can be modeled in two independent parts: a stacked oxide/nitride/oxide (ONO) capacitor C_{CF}, placed between the terminals C and F, and an equivalent MOS transistor. For every cell length, we extracted the value

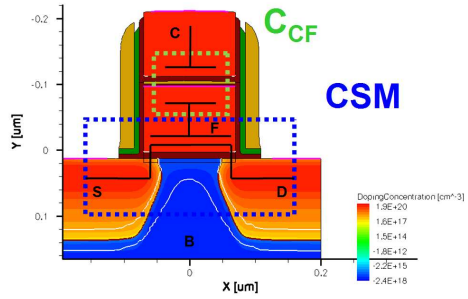


Figure 1: TCAD geometry of the flash cell and equivalent circuit for semi-analytical model.

of C_{CF} from TCAD simulations and considered it constant with respect to the applied biases (e.g. C_{CF}=83.1 aF for L=0.14 μm). This assumption is justified by the negligible impact of polydepletion effects on C_{CF}, for high values of FG and CG doping.

We performed 2D TCAD simulations with a commercial simulation package [6]. TCAD process calibration has been done on the 65 nm node manufacturing process flow based on measurements (not shown here) for various gate lengths, using the most advanced dopant / defect pair diffusion models [7]. We performed a series of TCAD simulations varying the length of the device in Figure 1 from 0.14μm to 1μm to analyze the impact of SCE and overlap capacitances. In this way, we could easily discriminate between the intrinsic behavior of the device, prominent for long cells, and the extrinsic effects dominating at sub–micron scale.

We used these simulations both to validate our NVM–CSM model and to demonstrate the extrinsic effects on the capacitive coupling coefficients $\alpha_{\rm I}$. The coupling coefficients can be calculated by deriving the FG potential with respect to the applied biases. We determined the evolution of $\alpha_{\rm I}$ with respect to $V_{\rm D}$ and $V_{\rm C}$ biases, varying from 0 to 9 V. We maintained $V_{\rm B}$ and $V_{\rm S}$ grounded, but the same analysis can be performed for the extraction of $\alpha_{\rm S}$ and $\alpha_{\rm B}$, respectively varying the biases $V_{\rm S}$ and $V_{\rm B}$ (not shown). Figure 2 shows the bias dependence of $\alpha_{\rm C}$ and $\alpha_{\rm D}$ coefficients. Significant difference are observed between the long and the short device. In the former, $\alpha_{\rm C}$ reaches a maximum value near the intrinsic threshold voltage of the MOS and then abruptly decreases. In the latter, the variation of $\alpha_{\rm C}$ is smoother, the relative amplitude is reduced and the coefficient exhibits a monotonous decrease vs. $V_{\rm C}$. The behavior of $\alpha_{\rm D}$ is similar to the one of $\alpha_{\rm C}$, when passing from subthreshold to overthreshold, but the magnitude of the variation is lower.

3 NVM CHARGE BALANCE MODEL

Using the charge balance approach, we implemented an iterative method to determine $V_{\rm F}$. The charge balance

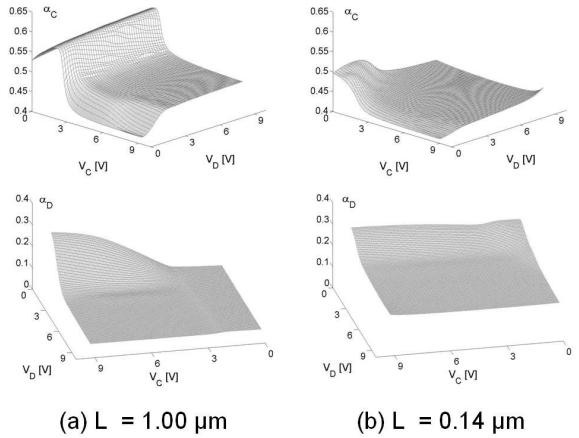


Figure 2: 2D TCAD simulations showing large differences in gate coupling coefficient $\alpha_{\rm C}$ and drain coupling coefficient $\alpha_{\rm D}$, vs. $V_{\rm C}$ and $V_{\rm D}$ ($V_{\rm S} = V_{\rm B} = 0$), in (a) long and (b) short cells.

equation is expressed as:

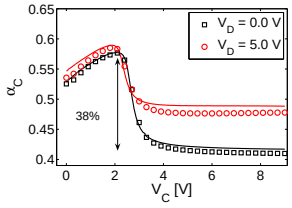
$$Q_G(V_F, V_S, V_D, V_B) + C_{CF}(V_F - V_C) = Q_0 = \text{const.} \quad (1)$$

where Q_G indicates the charge on the MOS gate, and Q_0 the total charge stored in FG, constant in DC conditions and only depending on the state of the cell. An iterative bisectional approach is required to solve the non–linear equation: at each iteration, the MOSFET charge Q_G is determined using a Matlab implementation of Charge Sheet Model (namely CSM [8]) based on Brews equation [9].

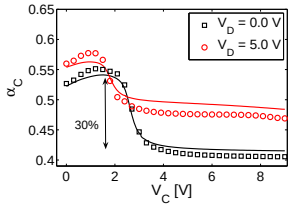
For our NVM–CSM simulations, we used an uniform doping profile with constant value $N_{Ch} = 10^{18}\text{cm}^{-3}$ together with an oxide thickness of 9.6 nm. We included an advanced 3-terminals-overlap model that takes into account the doping profile in the overlap region and thus is able to reproduce the dependency of the overlap capacitances with respect to the surface potential in the channel [10]. The fringing capacitances' contributions are calculated using an electric-field-lines approach [11]. SCE and reverse SCE have been modeled considering the reduction of bulk charge due to the sharing between the source and drain terminals [12]. We will clearly present that for submicron devices all these physical effects must be accurately accounted for.

In Figure 3 the predictions of our model have been compared with TCAD simulations for 3 different device lengths (0.14μm, 0.37μm and 1μm) and excellent agreement has been found. Such an agreement would be difficult to obtain without the advanced SCE model and the overlap model used in the calculation. Indeed, the SCE model used in NVM–CSM [12], based on Wu's charge sharing model [13], not only accurately reproduces the V_{TH} roll–off, but has also a strong impact on the charges in short devices. Similarly, overlap charges also significantly impact the total amount of charges in the shortest

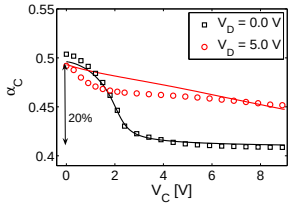
device and thus must be accounted for in the calculation. We performed a series of simulations progressively enabling our SCE and overlap capacitance models. In Figure 4(a), it can be seen that the intrinsic behavior of α_C, emerging when our SCE and overlap capacitance models are disabled, is characterized by a peak in proximity of V_{TH} and is identical to the behavior in long devices. When SCE are introduced, the peak is shifted to lower V_C (V_{TH} roll–off) and its amplitude decreases. Including overlap and fringing capacitances in the calculation, the curve is further flattened and shifted to lower values of α_C.



(a) L=1.00μm



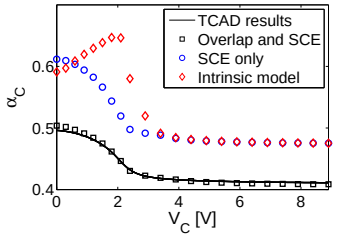
(b) L=0.375μm



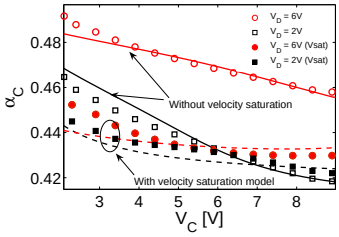
(c) L=0.14μm

Figure 3: TCAD (lines) and NVM-CSM (symbols) matching of the gate coupling coefficient α_C as a function of the control gate voltage V_C, for different drain biases V_D and different cell dimensions: (a) L=1μm, (b) L=0.375μm, (c) L=0.14μm.

Velocity saturation is another important phenomenon that occurs in short devices at high V_D and V_C. In this conditions, the cell is usually programmed and an accurate model of the α_I is required to guarantee a correct estimation of the injection current. As depicted in Figure 4(b), we demonstrated that also this parameter plays a major role on α_C bias dependency at the submicron scale. These NVM–CSM simulations on submicron devices confirm that extrinsic effects dominate over the intrinsic behavior of the cell, each playing an independent role on the coupling coefficient α_C. Figure 5 shows the dependency of V_F and drain current with respect to the static stored FG charge. The shift of the curves corresponds to the variation of the threshold voltage V_{TH} of the cell due to the presence of the stored charge. It can be noticed that, even in absence of stored charge, V_F is negative below inversion and changes sign at high V_C.



(a) Impact of extrinsic effects



(b) Impact of velocity saturation

Figure 4: (a) Impact of extrinsic effects on the coupling coefficient α_C for different control gate voltages V_C at $V_D = V_S = V_B = 0$; (b) impact of saturation velocity on α_C for different values of V_D. The lines are referred to TCAD simulations, the markers indicate NVM-CSM results.

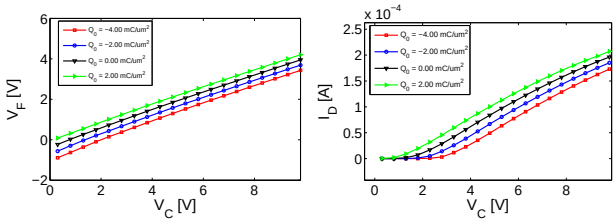


Figure 5: Influence of the charge Q_0 stored inside the floating gate on the FG potential V_F (top) and on α_C (bottom) for different control gate V_C biases (L=0.14μm).

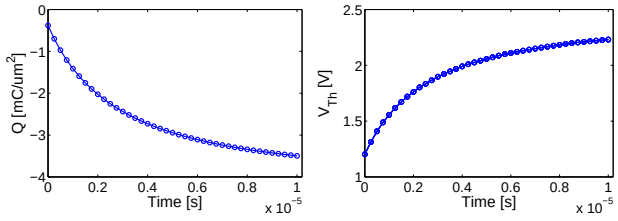


Figure 6: Variation of the stored charge Q and of the threshold voltage V_{Th} during programming by channel Hot-Electron injection simulated with NVM-CSM.

To demonstrate the capability of NVM–CSM of simu-

lating the complete behavior of a flash cell, we extended our research effort to transient simulations, implementing a charge injection model for programming [14] and Fowler - Nordheim tunneling for erasure. Figure 6 shows the dynamics of the cell during the programming phase ($V_{\mathrm{C}} = 9V$ and $V_{\mathrm{D}} = 4V$). As electrons are injected in the FG node, the stored charge decreases and the V_{TH} of the cell increases.

4 COMPACT MODEL IMPLEMENTATION

Now comes the question of a compact implementation of our approach, suitable for SPICE simulations. We used the latest Verilog–A release of the PSPv103 MOSFET model [15] and included an additional internal node for FG. Iteratively solving Eq.1, we performed the same procedure for the extraction of the coefficients and obtained the same intrinsic behavior as in TCAD and NVM–CSM for long devices. However, slight discrepancies (more pronounced in shorter devices) are present in proximity of the peak in α_{C}; this could be due to the fact that the PSP SCE model does not take into account the charge sharing [13] in the surface potential calculation. Further work along this line should be performed to refine the SCE compact model on the basis of TCAD, NVM–CSM and experimental results.

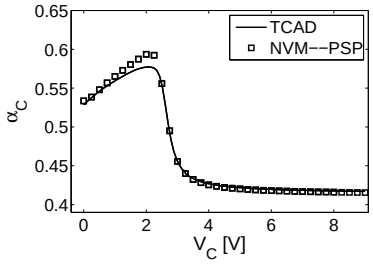


Figure 7: TCAD and NVM-PSP matching of the gate coupling coefficient α_{C} as a function of the control gate voltage V_{C}, for $V_{\mathrm{D}} = 0V$ and L=1μm.

5 CONCLUSION

In this paper, we have shown the dependence of α_{I} coefficients on the applied biases, implementing a semi– analytical charge balance model for the study of embedded non–volatile memory device. We validated our model using calibrated TCAD simulations, extracted the intrinsic behavior of the coefficients and demonstrated the importance of extrinsic effects (SCE, overlap capacitances, velocity saturation) both in read and programming conditions. We refined our method with the implementation of a charge sharing model for SCE analysis and an advanced overlap model, improving the model proposed by [4]. We applied our model to transient simulations and demonstrated the capability of simulating the full electrical behavior of a flash cell. Finally we implemented a compact version of our semi–analytical model using PSP and CBM. However, in CBM the model computational efficiency is reduced, compromising its application to compact modeling. In this scenario, an alternative technique would consist in accurately modeling the bias–dependency of the coupling coefficients α_{I} using CBM. The accurate bias–dependence of the coefficients can then be used in a CNM–based compact model, to achieve both efficiency and accuracy, guaranteed respectively by CNM and CBM.

REFERENCES

[1] Kolodny *et al.*, "Analysis and modeling of fg eeprom cells," *Electron Devices, IEEE Transactions on*, vol. 33, no. 6, pp. 835–844, 1986.

[2] Larcher *et al.*, "A new compact dc model of fg memory cells without capacitive coupling coefficients," *Electron Devices, IEEE Transactions on*, vol. 49, no. 2, pp. 301–307, 2002.

[3] Wong *et al.*, "Analysis of the subthreshold slope and the linear transconductance techniques for the extraction of the capacitance coupling coefficients of fg devices," *Electron Device Letters, IEEE*, vol. 13, no. 11, pp. 566–568, 1992.

[4] Larcher *et al.*, "Bias and w/l dependence of capacitive coupling coefficients in fg memory cells," *Electron Devices, IEEE Transactions on*, vol. 48, no. 9, pp. 2081–2089, 2001.

[5] Galup-Montoro *et al.*, "A compact model of mosfet mismatch for circuit design," *Solid-State Circuits, IEEE Journal of*, vol. 40, no. 8, pp. 1649–1657, 2005.

[6] Synopsys, *Synopsys Sentaurus, release Z-2007.03, Sprocess and Sdevice simulators*, 2007.

[7] Schermer *et al.*, "On a computationally efficient approach to boron-interstitial clustering," pp. 342–345, 2007.

[8] Gilibert *et al.*, "Dc and ac mos modelling in presence of high gate leakage and experimental validation," *Solid State Electronics*, vol. 48, no. 4, pp. 597–608, 2004.

[9] Brews, "A charge-sheet model of the mosfet," *Solid-State Electron*, vol. 21, no. 2, pp. 345–355, 1978.

[10] Klein *et al.*, "Description of the bias dependent overlap capacitance at lddmosfets for circuit applications," in *Electron Devices Meeting, 1993. Technical Digest., International*, 1993, pp. 493–496.

[11] Goren *et al.*, "An interconnect-aware methodology for analog and mixed signal design, based on high bandwidth (over 40 ghz) on-chip transmission line approach," in *DATEC, 2002. Proceedings*, 2002, pp. 804–811.

[12] Quenette *et al.*, "Dynamic charge sharing modeling for surface potential based models," in *WCM*, 2009.

[13] Wu *et al.*, "An analytic and accurate model for the threshold voltage of short channel mosfets in vlsi," *Solid-state electronics*, vol. 27, no. 7, pp. 651–658, 1984.

[14] Gilibert *et al.*, "Characterization and modeling of gate-induced-drain-leakage," *IEICE Transactions on Electronics*, vol. 3, pp. 829–837, 2005.

[15] Gildenblat *et al.*, "Introduction to psp mosfet model," in *WCM*, 2005, pp. 19–24.