Physical Understanding of Program Injection and Consumption in Ultra-Scaled SiN Split-Gate Memories

L. Masoero, V. Della Marca, G. Molas, M. Gély, O. Cueto, P. Colonna, A. de Luca, P. Brianceau, C. Charpin, D. Lafond, et al.

Physical understanding of program injection and consumption in ultra-scaled SiN Split-Gate memories

L. Masoero, G. Molas, V. Della Marca[+], M. Gély, O. Cueto, J. P. Colonna, A. De Luca, P. Brianceau, C. Charpin, D. Lafond, V. Delaye, F. Aussenac, C. Carabasse, S. Pauliac, C. Comboroure, P. Boivin[+], G. Ghibaudo*, S. Deleonibus, B. De Salvo

CEA, LETI, MINATEC Campus, 17 rue des Martyrs, 38054 GRENOBLE Cedex 9, France, lia.masoero@cea.fr
(+) STMicroelectronics, Rousset, France (*) IMEP-LAHC CNRS, Grenoble, France

***Abstract*— In this work, a detailed study of the physical mechanisms governing the Source Side Injection programming in ultra-scaled (down to 20nm) SiN split-gate memories is presented. Experimental measurements coupled to static and dynamic TCAD simulations are shown. In particular, we claim that adjusting the select gate voltage in moderate inversion allows for the optimization of the compromise between high electron injection and limited consumption. Then, we show that scaling the dimensions of the select gate can induce a higher consumption, while scaling the memory gate leads to lower programming energy (<1nJ) due to higher injection efficiency, suitable for low power applications.**

I. Introduction

In the last years, the demand for highly reliable, low cost and low power embedded memories has strongly increased driven by industrial and automotive products. One attractive solution [1-2] consists in the split-gate charge trap memories that combine the advantages of discrete storage layer (robustness to SILC, scalability…) and those of the split-gate memory architectures (low power, small circuitry, high speed…). In [3] we presented the impact of the memory gate scaling on the split-gate memory window. In this work, we focus on the understanding of the physical mechanisms beside the Source Side Injection (SSI) operation and the role of the select gate on the program injection. Then, we investigate the impact of the select gate and the memory gate scaling on the injection efficiency and programming current consumption.

A. Devices under test - In our samples, a 6nm LPCVD Si_3N_4 charge trapping layer is embedded between a 5nm tunnel oxide and an 8nm HTO control dielectric. Electron beam lithography was used to define select gates (L_{SG}) down to 40nm and the channel widths (W) down to 100nm. Thanks to the overlap of the memory gate over the select gate we achieved electrical memory gate length (L_{MG}) down to 20nm (Fig.1).

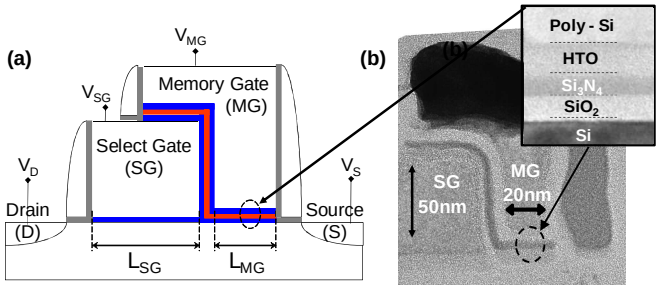

Figure 1. (a) Schematic and (b) TEM cross section of the SiN split-gate memory studied in this work.

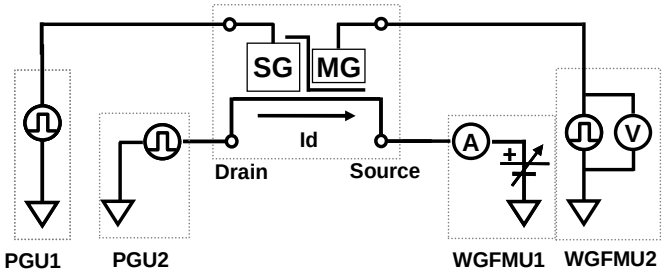

Figure 2. Experimental setup used to measure the current comsuption during the Source Side Injection (SSI) programming operation.

B. Experimental setup

In order to quantify the programming consumption, the source current was measured during the programming operation using the dynamic technique proposed in [4]. The developed setup (Fig 2) uses two waveform generators combined with two WGFMUs (Waveform Generator and Fast Measurement Units) integrated in an Agilent B1500A semiconductor device analyzer. The setup was verified comparing the $I_S(V_{SG})$ transfer characteristics (at low V_S) with the average current consumed during a programming pulse for various select gate voltages. The good matching between the current measured in continuous and dynamic mode (Fig. 3) demonstrates the validity of our setup. Moreover it proves the capability of the select transistor to control the current even when high bias voltages are applied to the source and memory gate electrodes.

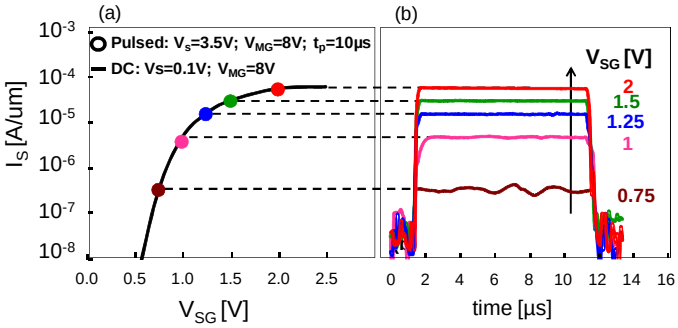

Figure 3. Comparison between (a) $I_S(V_{SG})$ transfer characteristic and (b) channel consumption current as a function of the select gate voltage (V_{SG}) during a 10µs programming pulse. In dynamic mode, each point corresponds to the average current measured during the pulse.

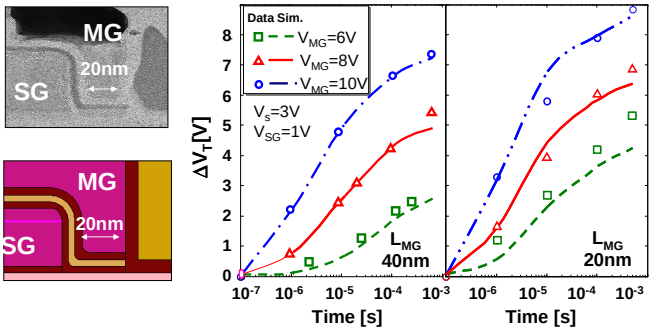

Figure 4. Left: (up) TEM image of the tested device with a 20nm memory gate length; (down) corresponding TCAD simulated structure. Right: Measured and simulated programming characteristics using the Fiegna model [5].

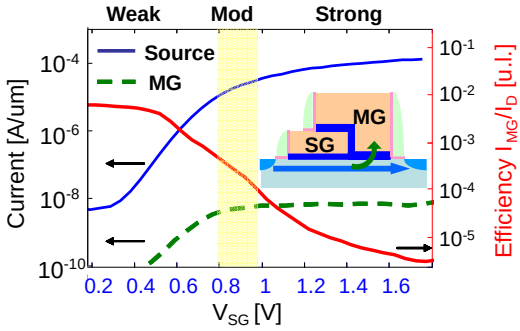

Figure 5. Measured source current (I_D); memory gate current I_{MG}; and computed injection efficiency (I_{MG}/I_D) for a split-gate with a dummy memory stack composed by a 5nm SiO_2 (L_{SG}=200nm; L_{MG}=100nm).

C. *TCAD simulations*

The experimental results were explained by TCAD dynamic simulations, using Fiegna's model [5] in the Synopsys suite able to compute the injected charge during programming. In Fiegna's model, the hot carrier injection current I_G is calculated as an integral along the semiconductor-insulator interface over the product of energy dependent normal to interface carrier velocity ($v_\perp$), carrier distribution energy (f), and carrier density of states (g), so that:

$$I_g = q\int P_{ins}\left(\int_{E_{B0}}^{\infty} v_\perp(\varepsilon)f(\varepsilon)g(\varepsilon)d\varepsilon\right)ds \quad (1)$$

The carrier distribution energy is approximated to the case of a parabolic and an isotropic band structure, and equilibrium between lattice and electrons leading to a simplified expression of the gate current:

$$I_g = q\frac{A}{3\chi}\int P_{ins} n \frac{F_{eff}^{3/2}}{\sqrt{E_B}} e^{\frac{\chi E_B^3}{F_{eff}^{3/2}}} ds \quad (2)$$

where A is a fitting parameter; χ is a constant of the high-energy distribution function; E_B is the Si-SiO_2 barrier energy; n the electron density; P_{ins} the probability that an electron does not scatter in the image potential well; and F is the effective electric field that replaces the local electric field to capture at first order the effects of the non-locality of hot electron injection [6].

The simulation parameters were calibrated by fitting the programming characteristics over different V_S/V_{MG} for two memories with respectively 20nm and 40nm memory gate lengths (Fig. 4). In our structures, 3V of programming V_S is sufficient to generate hot carriers due to the short L_{GM}. Note that the numerical simulations correctly reproduce our experimental results.

II. RESULTS AND DISCUSSION

The understanding of the physical mechanisms beside the SSI operation in split-gate memories is crucial for the interpretation of different aspects of the experimental results [7-11]. In this section we use simulations to understand the impact of the memory and select gate scaling on the measured programming efficiency and current consumption.

A. *Basics of SSI injection*

In split-gate charge trap memories, the electrons flow through the channel below the select gate to be successively injected toward the charge trapping layer. This happens because the source voltage induces a high electric field parallel to the interface that gives to the electrons the needed energy to pass over the Si/SiO_2 barrier. Then, the electrons are driven to the charge trapping layer by the attractive transversal electric field induced by a strong positive voltage applied on the memory gate.

We started with measuring a reference split-gate structure with a dummy memory transistor, composed by a 5nm oxide instead of the Oxide/Nitride/Oxide memory stack, in order to directly monitor the injected current through the tunnel oxide. We measured the source current (I_S); the memory gate current I_{MG}; and we computed the injection efficiency (I_{MG}/I_S) as a function of the select gate voltage for a high source and memory gate bias (V_S=3V; V_{MG}=3V). Fig 5 shows that the best choice for the select gate voltage during the programming operation is close to the threshold voltage, giving the best compromise between a low current consumption and a high current injection. Indeed, when the select transistor is in weak inversion, the gate and source currents increase nearly exponentially, on the contrary when the select transistor is in strong inversion, the injection current saturates and using a higher V_{SG} is inefficient.

The experimental results were figured out by the device simulations. The simulated channel potential during programming operation when the select transistor is in weak (V_{SG}=0.3V), moderate (V_{SG}=0.9V), and strong inversion (V_{SG}=2V), is reported in Fig.6-a. First it should be noted that most of the hot electrons are generated by the strong electric field created across the weak-controlled gap between select and memory gates. Indeed, in this thin region occurs the major voltage drop between the select and memory gates. The injection current (Eq.2) depends, in a first approximation, on the product (defined as *p2*) between a monotonic function of the local electric field F and the number of channel electrons n; this product (Fig.6-b), in agreement with the experimental results, is low at V_{SG}=0.3V and remains in the same order of magnitude between V_{SG}=0.9V and V_{SG}=2V. This can be explained by the fact that F and n have opposite behavior as the select gate bias increases.

When the select transistor operates in weak inversion, the injected current increases with V_{SG}, due to the increasing of the amount of electrons provided by the select transistor.

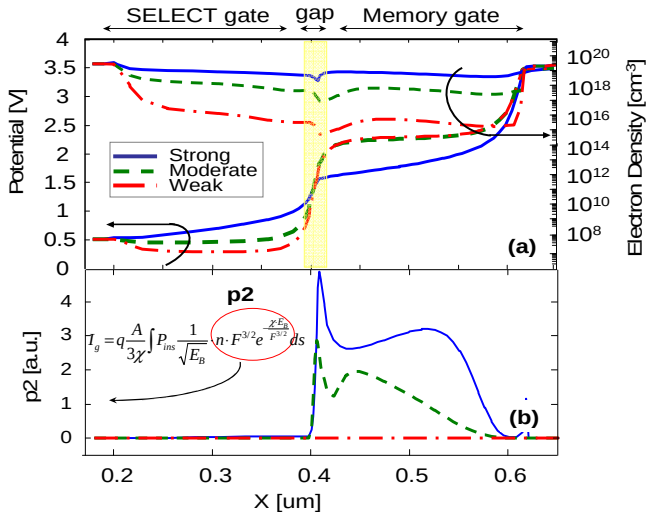

Figure 6. (a) Simulated potential profile and electron concentration at the Si/SiO_2 interface during programming operation. The bias conditions are: V_{MG}=3V V_S=3V V_{SG}=0.3V (weak) V_{SG}=0.9V (moderate) V_{SG}=2V (strong inversion). The device dimensions are L_{SG}=200nm and L_{MG}=200nm. (b) Corresponding simulated injected current (in arbitrary units), plotted along the memory channel.

On the other hand, in strong inversion, the injected current is limited by the reduction of the electric field, as V_{SG} increases. Indeed, in strong inversion, the select gate potential is disturbed by the memory gate, causing a lowering of the potential difference at the gap side that results in a lower electric field. Analog experiments have been done for the memory devices described in section I where we measured the consumed current at the source electrode during a programming pulse (V_{MG}=8V V_D=3V t=100us) and the memory gate threshold voltage shift. Fig.7 shows the memory window (ΔV_T) and current consumption I_S when the select gate is in weak, moderate and strong inversion, for a large memory gate length. The previous behavior (Fig.5) was found again, confirming that the optimal choice for V_{SG} is in a region strictly above the select gate threshold voltage, insuring the best compromise between a high programming window and a limited current consumption.

B. *Select Gate Scaling*

The effect of the select gate scaling on the programming current consumption was investigated by measuring the select gate threshold voltage lowering and the programming window for devices with a select gate length from 350nm down to 40nm.

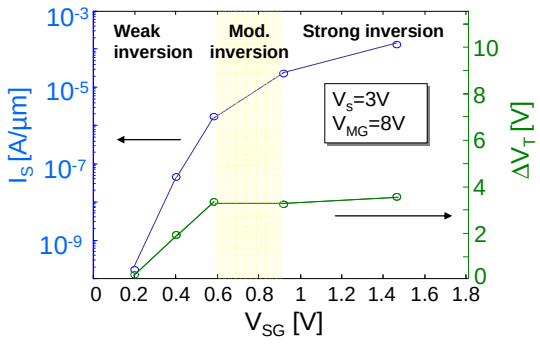

Figure 7. Measured channel current and programming window versus the normalised select gate voltage during a 10µs programming pulse with Vs=3V and V_{MG}=8V. L_{MG}=200nm

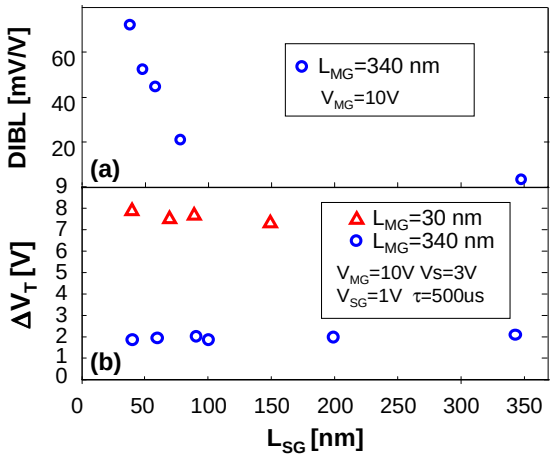

Figure 8. (a) Programming window as a function of the select gate length for devices with a long (L_{MG}=340nm) and short (L_{MG}=40nm) memory gate length. (b) DIBL due to select gate scaling

Fig.8-a shows that as the select gate dimensions scale, the memory window remains unchanged but the select gate threshold voltage decreases due to DIBL (drain induced barrier lowering). This parasitic effect causes, for a given $V_{SG,}$ an increase of the consumed current during program operation. For instance, as the select gate scales from 90nm to 40nm we measured during a pulse of 10µs with V_S=3V; V_{MG}=10V; V_{SG}=1V (corresponding to the same ΔV_T in the two devices: see Fig.8-b), a current consumption increase of about one decade. Indeed, the DIBL in devices with scaled L_{SG} is a consequence of the insufficient control of the channel potential by the select gate. At high applied source voltages this induces, similarly to the case of the strong inversion described above, an increasing of the consumed current and a lowering of the electric field that results in a lower injection efficiency ($\Delta V_T/I_S$). Therefore in ultra-scaled devices, optimizing the junction implantation is of great importance to control the consumption.

C. *Memory Gate Scaling*

The impact of the memory gate scaling on the current consumption has been investigated by studying the programming characteristics of devices with a 100nm select gate length and a memory gate length from 180nm down to 30nm. Fig.9 shows the programming window after a 500µs program pulse (V_{MG}=10V; V_S=3V; V_{SG}=1V) as a function of the memory gate length. With the shrinking of the memory dimensions the programming window strongly increases from 3V to 9V. This result has been explained by the means of TCAD simulations. In long devices the electric field in the memory channel shows two peaks (Fig.10-a), the first one is located in the gap, due to the difference between the memory gate and the select gate potentials; the second peak is created at the channel source junction. As the gate length is further reduced, the two peaks merge and the maximum of the electric field increases, leading to an enhanced injected charge in the nitride layer (Fig.10-b). This memory window enhancement in scaled devices can be used to reduce the programming consumption. To analyze this effect, we first measured for various memory gate lengths the programming characteristics and the current consumption as a function of the programming time (Fig. 11).

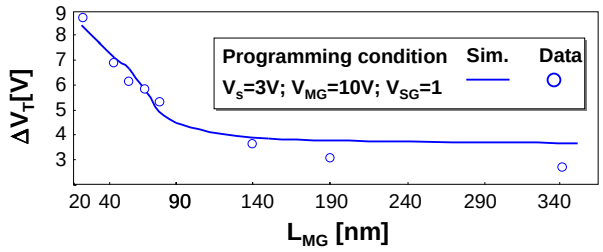

Figure 9. Measured and simulated programming windows as a function of the memory gate length.

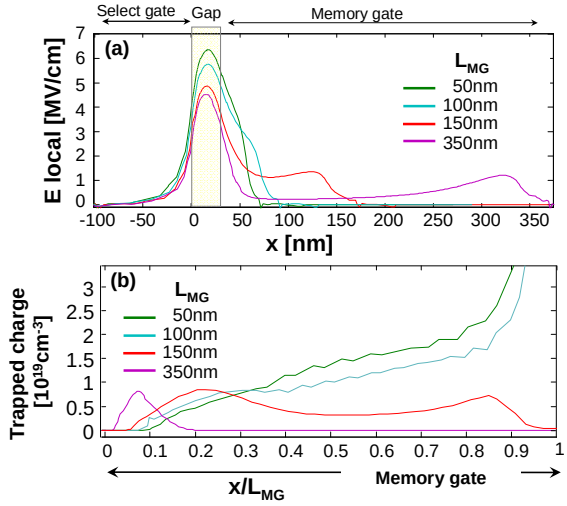

Figure 10. (a) Simulated local electric field in the channel during Source Side Injection programming operation for various gate lengths. (b) Trapped charge after a programming pulse (V_{MG}=10V V_S=3V V_{SG}=1 t=500us) as a function of the normalised memory gate length.

Then based on these graphs, we extrapolated the required programming time to reach a given programming window of 3.5V and the corresponding energy consumption. The consumed energy is calculated as the integral along the programming time of the channel current times the applied source voltage. In scaled devices the memory window is higher but the average current consumed during a programming pulse is nearly constant (Fig.12) as it only depends of V_{SG}.

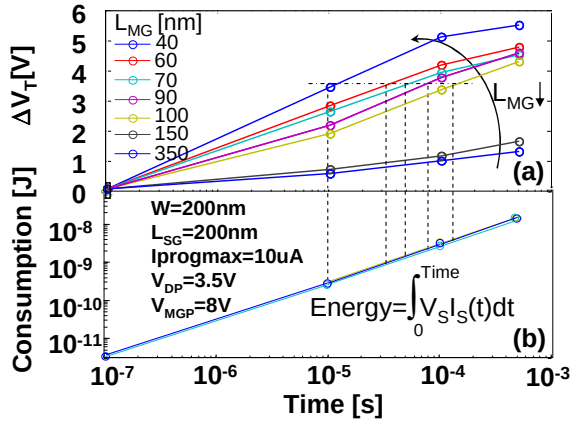

Figure 11. Measured programming characteristics (a) and measured consumed current (b) as a function of the programming time for various devices with different memory gate lengths.

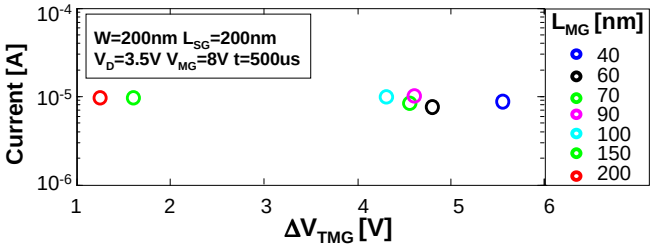

Figure 12. Avarage current measured during a programming pulse of 500us for devices with differente memory gate lenghts plots as a function of the relative programming window.

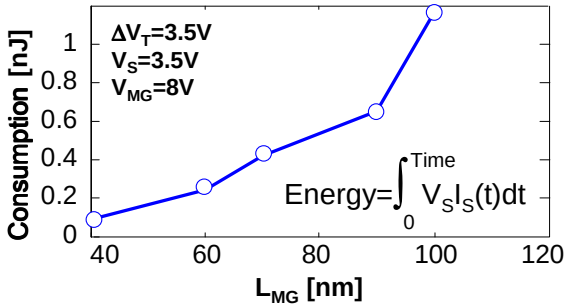

Figure 13. Measured current consumption during a programming pulse with V_S=3.5V; V_{MG}=8V to reached a programming windows of 3.5V as a function of the memory gate length.

The result shows an improvement of over 10 times of consumption energy when the memory length passes from 100nm to 40nm. In particular, for sub-90nm gate length devices, <1nJ of programming energy is reached, suitable for low power applications (fig.13).

III. Conclusion

Experiments on ultra-scaled (down to 20nm) SiN split-gate memories, coupled to static and dynamic TCAD simulations, allowed us to understand the physical mechanisms behind the Source Side Injection programming. In particular, we showed that by adjusting the select gate voltage in moderate inversion we can optimize the compromise between the high electron injection and the limited consumption. Then, we showed that scaling the dimensions of the select gate can induce a higher consumption, while scaling the memory gate leads to lower programming energy (<1nJ) due to higher injection efficiency, suitable for low power applications.

Acknowledgment

This work was done in the frame of CEA-LETI/ ST-Microelectronics bilateral collaboration and the CATRENE REFINED project.

[1] J. Yater, et al. , proc. of IMW 2011
[2] T. Tanaka et al., proc. of VLSI 2003
[3] L. Masoero, et al. proc of IEDM 2011
[4] V. Della Marca et al., proc. of ISDRS 2011
[5] C. Fiegna, E Sangiorgi, IEEE Trans. on Elec. Dev., 40, pp.619-627, 1993
[6] A. Zaka et al., proc. of ISDRS 2009
[7] L. Breuil et al., IEEE Trans. on Elec. Dev., 52, 2005
[8] P. Palestri et al., IEEE Trans. on Elec. Dev., 53, pp. 488-493, 2006
[9] K. Sridhar et al., proc of IPFA 2005
[10] Y.-H. Wang et al., IEEE Proc. Circuits Dev. Syst., Vol. 153, No. 2, 2006
[11] W. Stefanutti et al.; IEEE Trans. on Elec. Dev., 53, pp. 89-96, 2006