Scalability of split-gate charge trap memories down to 20nm for low-power embedded memories

L. Masoero, G. Molas, F. Brun, M. Gély, P. Colonna, V. Della Marca, O. Cueto, E. Nowak, A. de Luca, P. Brianceau, et al.

Scalability of split-gate charge trap memories down to 20nm for low-power embedded memories

L. Masoero, G. Molas, F. Brun, M. Gély, J. P. Colonna, V. Della Marca[+], O. Cueto E. Nowak, A. De Luca, P. Brianceau, C. Charpin, R. Kies, A. Toffoli, D. Lafond, V. Delaye, F. Aussenac, C. Carabasse, S. Pauliac, C. Comboroure, 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, split-gate charge trap memories with electrical gate length down to 20nm are presented for the 1st time. Silicon nanocristals (Si-ncs), or silicon nitride (Si_3N_4) and hybrid Si-nc/SiN based split-gate memories, with SiO_2 or Al_2O_3 control dielectrics, are compared in terms of program erase and retention. Then, the scalability of split-gate charge trap memories is studied, investigating the impact of gate length reduction on the memory window, retention and consumption. The results are analyzed by means of TCAD simulations.

Introduction

Due to the increasing demand for consumer, industrial and automotive products, highly reliable, and low integration cost embedded memories are more and more required. In this context, split-gate charge trap memories were proposed for microcontroller products, combining the advantage of a discrete storage layer (as robustness to SILC, scalability…) and of the split-gate configuration (as low power consumption, small circuitry, high speed…). Two main approaches exist in the literature, integrating either silicon nanocrystal (Si-nc) [1,2] or silicon nitride (Si_3N_4) [3] charge trapping layer. So far, no clear comparison of Si-nc and Si_3N_4 based split-gate memories was provided, and the scalability of the concept down to aggressive dimensions was not clearly demonstrated up to now. Moreover, no studies have been reported in the literature concerning the impact of high-k control dielectrics in split-gate charge-trap devices.

Results and discussion

1. Technological details

Split-gate charge trap memories were processed with a "memory last" configuration, meaning that the Memory Gate (MG) is deposited on the Select Gate (SG) electrode. Electron beam lithography was used to define control gates down to 40nm and the channel widths (W) down to 100 nm. The electrical memory gate length is controlled by the poly-Si layer overlapping the memory channel (Fig.1), allowing to achieve gate length down to 20nm (Fig.2). In the following, L_{GM} will refer to this electrical length. Various gate stacks were integrated with Si-nc (Sample A), Si_3N_4 (B), and hybrid Si-nc/SiN (C) charge trapping layer CTL [4]. Finally, Si-nc/SiN CTL was combined to HTO/Al_2O_3 control dielectrics to allow FN bottom erase (D). Technological details of the various samples are given in Table 1.

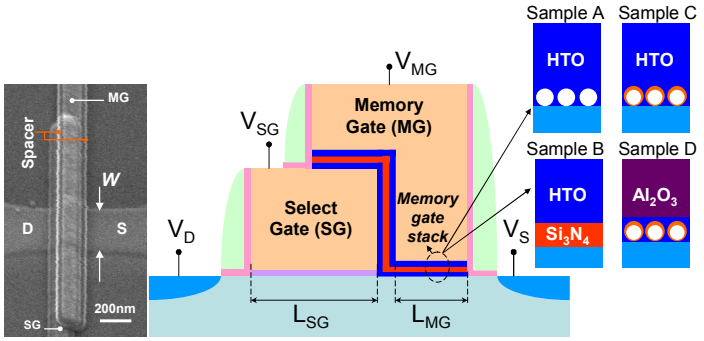

Figure 1. SEM plane view of channel, select gate and memory gate (**left**), and schematic cross section (**right**) of a typical split-gate memory processed in this work.

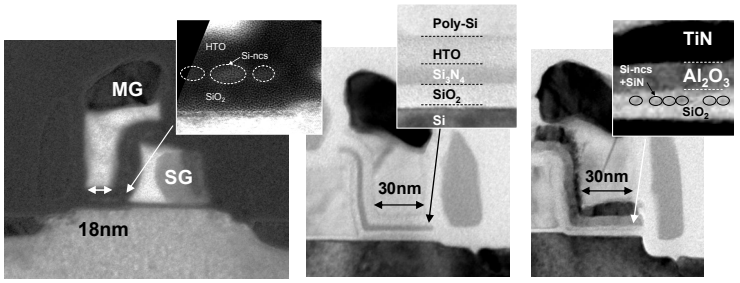

Figure 2. TEM images of split-gate memories for various memory gate stacks, **(a)** Si-nc, **(b)** Si_3N_4 **(c)** Si-nc/SiN charge trapping layers. (a) was obtained in Energy Filtered mode (selecting Si), (b) and (c) in High Resolution mode.

	Sample A	Sample B	Sample C	Sample D
Tunnel Oxide	SiO_2 (5nm)	SiO_2 (5nm)	SiO_2 (5nm)	SiO_2 (4nm)
Charge trapping layer	Si-ncs (Φ~6nm)	SiN (6nm)	Si-ncs+SiN (3nm)	Si-ncs+SiN (3nm)
Control dielectrics	HTO (8nm)	HTO (10nm)	HTO (10nm)	HTO (3nm) Al_2O_3 (8nm)
Control Gate	Poly-Si	Poly-Si	Poly-Si	TiN

Table 1. Technological details of the studied split-gate charge trap memories.

2. Basics of split-gate charge trap memories

$I_D(V_{SG}, V_{MG})$ and $I_D(V_{MG}, V_{SG})$ transfer characteristics were measured for 20nm and 70nm gate length memory devices (Fig.3-4). On the 20nm memory, a slight V_T reduction appears as V_{SG} increases, due to a parasitic control of the memory channel by the select transistor. Despite this issue, the channel current can efficiently be controlled from ON to OFF state by biasing the select gate.

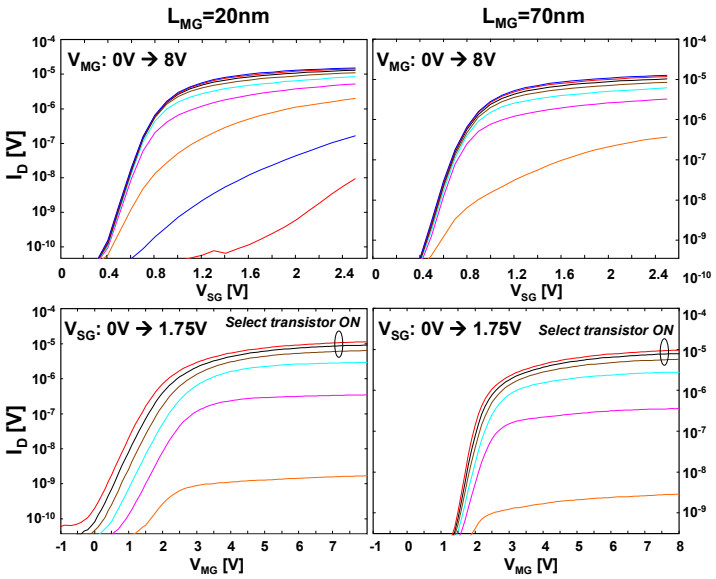

Figure 3. $I_D(V_{SG}, V_{MG})$ **(top)** and $I_D(V_{MG}, V_{SG})$ **(bottom)** characteristics for 20nm and 70nm memory gate lengths.

In Fig.4, the channel current is dynamically measured during a programming pulse of 10µs (V_S=3.5V) as a function of V_{SG}.
Thus the select gate bias allows to control the programming current, with typical values of 1-10µA when the select transistor operates in the subthreshold regime.

3. Impact of charge trapping layer

3.1 Program Erase

The split-gate memories are programmed using Source Side Injection [5], biasing both the memory gate and the source electrode at high voltages. The select gate potential is set in order to operate close to the threshold regime (I_S~10µA). Fig.5 shows the program characteristics of Si_3N_4, Si-ncs and hybrid Si-nc/SiN split-gate memories, for various programming V_{MG} and V_S and compares their memory windows.

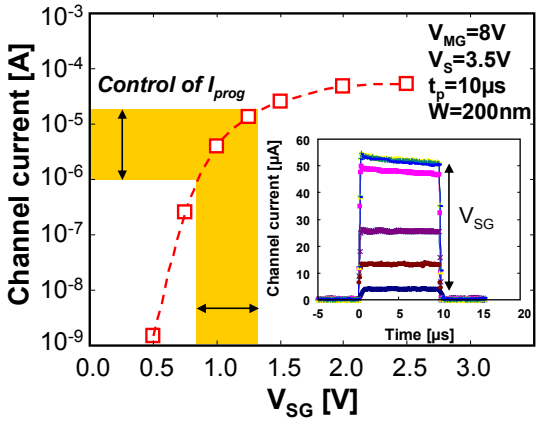

Figure 4. Channel consumption current as a function of the select gate voltage during SSI programming (10µs pulse). Each point corresponds to the average value measured during the pulse.

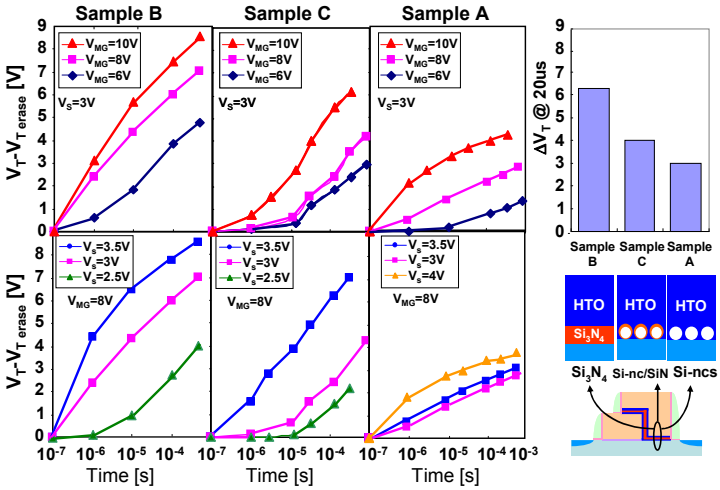

Figure 5. **Left**: Program characteristics in Source Side Injection mode of sample B (Si_3N_4), sample C (Si-nc/SiN), sample A (Si-nc) with 30nm memory electrical gate length, for various programming V_{MG} and V_S. **Right**: Memory windows of Si-nc, Si-nc/SiN and Si_3N_4 split-gate charge trap memories for same programming conditions (V_{GM}=10V, V_S=3V, t_w=20us).

Due to higher density of trapping sites, nitride memories exhibit a higher ΔV_T than Si-ncs. The hybrid Si-ncs/SiN layers offer a good way to enlarge the Si-ncs memory window [4].
Various erasing modes were used depending on the nature of the charge trapping layer: (1) nitride based memories are erased using Hot Hole Injection (HHI), (2) Si-nc memories with HTO control dielectrics are erased by Fowler-Nordheim (FN) injection through the top oxide, and (3) Si-nc memories with high-k control dielectrics are erased by FN injection through the bottom oxide. Fig.6 presents the corresponding erasing characteristics; HHI allows faster erasing speed but suffer from a higher current consumption. In FN mode, +16V and -16V gate voltages are respectively used to erase memory samples with HTO and Al_2O_3 control dielectrics.

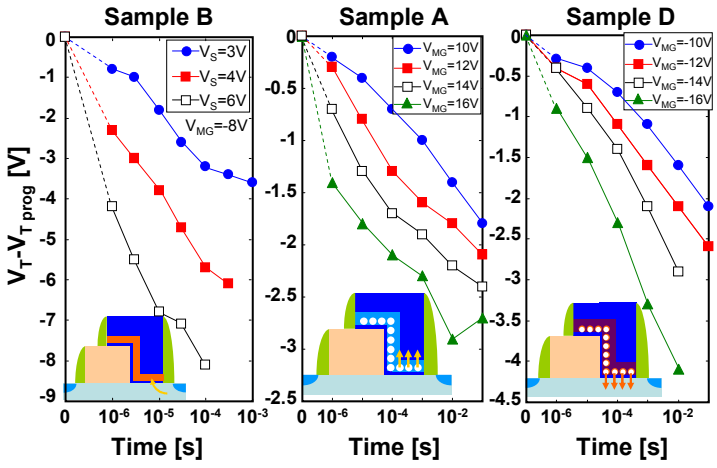

Figure 6. Erase characteristics of 30nm memories with various erasing mechanisms: Hot Hole Injection (sample B: Si_3N_4), FN through top dielectric (sample A: Si-ncs) and FN through bottom dielectric (sample D: Si-ncs with high-k top oxide).

3.2 Retention

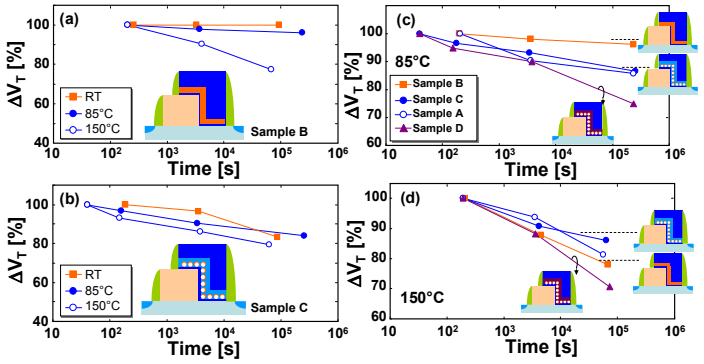

Figure 7. **(a, b)**: Retention characteristics for various temperature for Si_3N_4 (sample B) and Si-nc / SiN (sample C) charge trapping layers. **(c, d)**: Comparison of retention characteristics at 85°C and 150°C of split-gate charge trap memories with various gate stacks.

Fig.7 presents the retention characteristics. A high temperature activation is measured with a nitride CTL with a strong charge loss at 150°C (Fig.7-a), while Si-nc/SiN CTL exhibits a more stable behavior as the temperature is increased (Fig.7-b). Fig.7-c and Fig.7-d show the comparison of the memory samples at 85°C and 150°C. Up to 85°C, Si_3N_4 CTL offers the best retention performances, while for higher temperatures, an inversion of trend is observed as Si-nc memories present the smallest charge loss. For all the investigated temperatures, memories with high-k control dielectrics show faster charge decay, due to the thinner tunnel oxide and the lower barrier height of Al_2O_3 compared to SiO_2. Finally, Fig.8 presents the retention characteristics of a Si-nc/SiN memory at 150°C with a 20nm gate length, extrapolating a memory window of more than 2.5V after 10 years, demonstrating the scalability of this concept.

4. Impact of the memory gate scaling

In this section we investigate the impact of the memory gate scaling on the split-gate memory performances. Impact of gate length on the electron/hole mismatch was studied in [6]. Here we address the impact of scaling on the memory window and injected electron distribution. The experimental trends were explained by TCAD dynamic simulations, using Fiegna's model [7] in the Synopsys tool suite to compute the injected charge during programming. In order to analyze the trapped charge location, V_T shift after a 500µs program pulse (V_{GM}=10 V_S=3) was measured in forward (V_{DS}>0) and reverse (V_{DS}<0) modes.

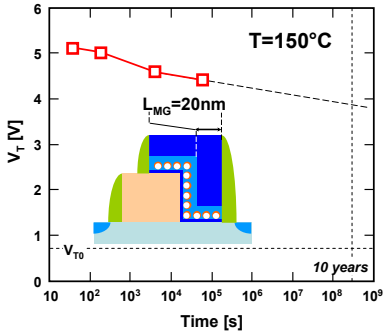

Figure 8. Retention characteristics at 150°C of a 20nm gate length memory device with Si-nc / SiN charge trapping layer (Sample C).

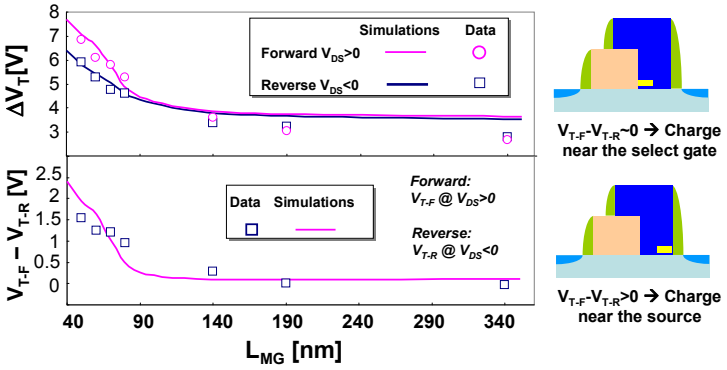

Figure 9. Measured and simulated $V_{T\text{-}F}$-$V_{T\text{-}R}$ as a function of the memory gate length (Sample B).

Fig.9-up shows that, as L_{GM} is reduced, the memory window strongly increases. Moreover, Fig.9-down shows that the difference between forward and reverse V_T is larger for scaled devices. This behaviour can be explained by means of TCAD simulation. Fig.10 shows the simulated parallel electric field in the channel during SSI operation. First, we observe that scaling L_{GM} leads to higher channel electric field, explaining the higher ΔV_T. Then, we can see that two peaks of electric field appear: one close to the select gate, the other near the source electrode. For large gate length devices (down to 150nm), the injection point is located on the side of the select gate. As the gate length is further reduced, the two peaks merge and the maximum electric field peak is shifted toward the source electrode, in agreement with [8]. This is confirmed by Fig.11 which shows that the maximum concentration of trapped charge gradually moves to the source electrode in short devices.

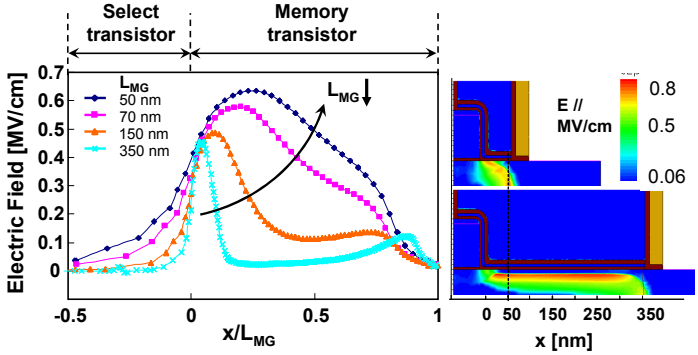

Figure 10. Simulated parallel electric field (Fiegna model) during Source Side Injection programming operation for various gate lengths.

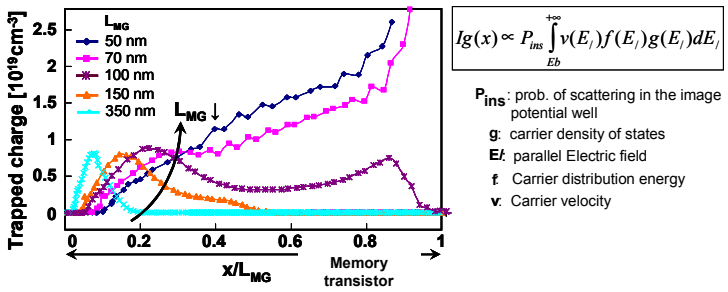

Figure 11. Simulated trapped charge in Si_3N_4 (Fiegna model) during Source Side Injection programming operation for various gate lengths.

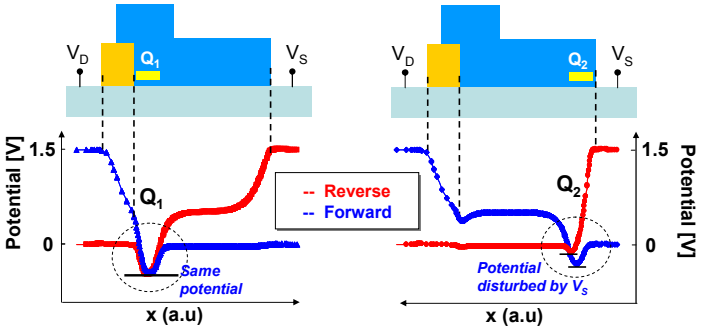

Figure 12. Simulated $V_{T\text{-}F}$-$V_{T\text{-}R}$ assuming the spread of a pocket of charge located close to the select gate (Q1) for a long device or to the source electrode (Q2) for a shorter device.

Fig.12 shows the channel potential during reading operation, and illustrates that the source partly screens the trapped charge above, leading to positive $V_{T\text{-}F}$-$V_{T\text{-}R}$ value [9]. On the contrary, $V_{T\text{-}F}$-$V_{T\text{-}R}$~0 when the charge is near the select transistor. Based on these considerations, we analyzed the retention characteristics. Fig.13 demonstrates that even after 10^5s, short devices still exhibit a larger ΔV_T. Then, the time evolution of $V_{T\text{-}F}$-$V_{T\text{-}R}$ was measured in Fig.14-Left, showing two behaviors: for long devices, the charge remains close to the select transistor as V_{TF}~$V_{T\text{-}R}$. For short devices the charge initially close to the source diffuses toward the select gate and $V_{T\text{-}F}$-$V_{T\text{-}R}$ decreases, as predicted by the TCAD simulations (Fig. 14-Right).

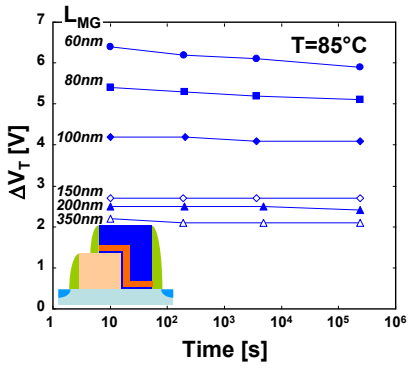

Figure 13. Impact of the memory gate length on the retention characteristics at 85°C (Sample B with Si_3N_4 CTL).

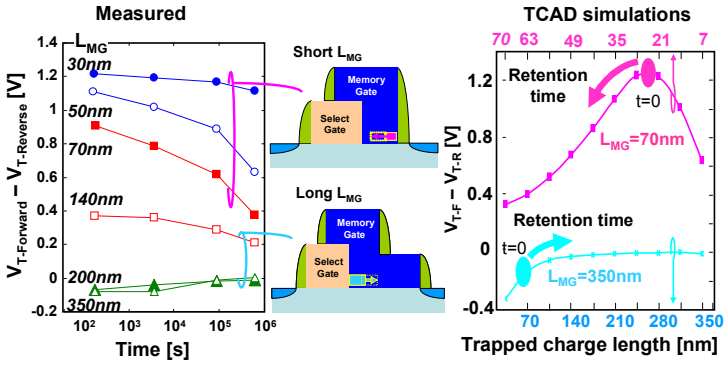

Figure 14. **Left:** Time evolution at RT of $V_{T\text{-}F}$-$V_{T\text{-}R}$ during retention for various memory gate lengths (Sample C). $V_{T\text{-}F}$-$V_{T\text{-}R}$ gives indications on the trapped charge location. V_T is measured at $|V_{DS}|$=1.5V. **Right:** Simulated $V_{T\text{-}F}$-$V_{T\text{-}R}$ assuming the spread of a pocket of charge located close to the select gate for a long device or to the source electrode for a shorter device.

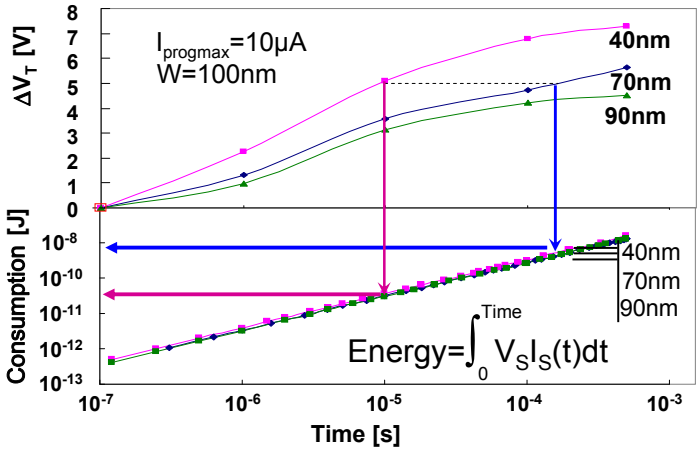

Figure 15. Simulated split-gate programming characteristics (Fiegna model) for various gate lengths, and corresponding programming energy consumption.

Finally, Fig.15 shows the simulated programming characteristics and programming energy consumption for various L_{GM}, Indeed, the gain in programming speed demonstrated previously in small devices allows a clear energy consumption reduction (~nJ), promising for contactless embedded applications.

Conclusions

Charge trap memories with Si-ncs or Si_3N_4 charge trapping layers were processed for the 1st time with electrical gate lengths down to 20nm. While Si_3N_4 exhibits higher memory window and better retention up to 85°C, Si-nc presents better retention at 150°C. Based on electrical measurements coupled with TCAD simulations, it was demonstrated that scaling the memory gate leads to an enlargement of the memory window. Moreover, the scaling also allows a reduction of the programming time and of the energy consumption. Indeed, these results make split-gate charge trap memories promising candidates for advanced technology nodes of embedded memory products.

Acknowledgments

This work was partly done in the frame of the bilateral collaboration between the CEA-LETI and ST-Microelectronics.

References

[1] J. Yater et al., proc. of Int. Mem. Workshop 2011, pp.79-82.
[2] G. Chindalore et al., 2008 Symp. on VLSI Tech., pp.136-137.
[3] T. Tanaka et al., 2003 Symp. on VLSI Circ., pp.211-212.
[4] G. Molas et al., proc. of IEDM 2007, pp.453-456.
[5] J. Van Houdt et al., IEEE Trans. on Elec. Dev., 42, pp.1314-1320, 1995.
[6] L. Breuil et al, Solid State Elec., 49, pp.1862-1866, 2005.
[7] C. Fiegna et al., IEEE Trans. on Elec. Dev., 38, pp. 603-610, 1991.
[8] Y. Tsuji et al., IEEE Trans. on Elec. Dev., 57, pp.466-473, 2010.
[9] L. Perniola et al., proc. of IEDM 2005, pp.857-860.