Università di Pisa

Low voltage hot-carrier programming of ultra-scaled SOI Finflash memories

J. Razafindramora
CEA LETI-MINATEC, Grenoble
Luca Perniola
CEA LETI-MINATEC, Grenoble
C. Jahan
CEA LETI-MINATEC, Grenoble
P. Scheiblin
CEA LETI-MINATEC, Grenoble
M. Gély
CEA LETI-MINATEC, Grenoble
C. Vizioz
CEA LETI-MINATEC, Grenoble
C. Carabasse
CEA LETI-MINATEC, Grenoble
F. Boulanger
CEA LETI-MINATEC, Grenoble
Barbara De Salvo
CEA LETI-MINATEC, Grenoble
S. Deleonibus
CEA LETI-MINATEC, Grenoble
S. Lombardo
CNR-IMM, Catania
C. Bongiorno
CNR-IMM, Catania
Giuseppe Iannaccone
Dipartimento di Ingegneria dell'Informazione: Elettronica, Informatica, Telecomunicazioni, Università di Pisa

J. J. Razafindramora, L. Perniola, C. Jahan, P. Scheiblin, M. Gély, C. Vizioz, C. Carabasse, F. Boulanger, B. De Salvo; S. Deleonibus, S. Lombardo, C. Bongiorno, G. Iannaccone, *Low voltage hot-carrier programming of ultra-scaled SOI Finflash memories*, Proceeding of ESSDERC, pp. C4L-B3-1-4, Munich, Germany, 11-13 September 2007.

Low voltage hot-carrier programming of ultra-scaled SOI Finflash memories

J. Razafindramora, L. Perniola, C. Jahan, P. Scheiblin, M. Gély, C. Vizioz, C. Carabasse, F. Boulanger, B. De Salvo and S. Deleonibus
CEA LETI-MINATEC, 17 rue des Martyrs
38054 Grenoble Cedex 09, France
carine.jahan@cea.fr

S. Lombardo, C. Bongiorno
CNR-IMM, Catania - Italy

G. Iannaccone
Università di Pisa, Pisa, Italy

***Abstract* — In this paper, we present a deep investigation of ultra-scaled Finflash memories, fabricated on Silicon on Insulator (SOI) substrate, with Silicon NanoCrystal (Si-NC) or nitride layers acting as storage nodes. Electrical characteristics of devices with channel length (L_G) as short as 30nm, and fin width (W_{FIN}) as narrow as 10nm are shown. Effective Channel Hot Electron (CHE) writing with sub-3.2V drain biases (i.e. ΔV_{TH}=3V at $V_G/V_D/t_{stress}$=9V/2.5V/100µs), as well as Hot Hole Injection (HHI) erasing with sub-4.5V drain biases are demonstrated. Finally, fully three dimensional Monte Carlo simulations, coupled with an original semi-analytical approach, allow us to give a qualitative explanation of the obtained experimental data.**

I. Introduction

It is widely believed that the scaling of Flash memories down to the 32nm technological node and beyond will face major issues, due to the high electric fields required for the programming and erasing operations and the stringent leakage requirements for long term charge storage [1].
In this context, new transistor architectures such as tri-gate Finflash memory devices [2] coupled with the discrete storage node approaches (i.e. nitride storage layer [3] or Silicon Nanocrystals, Si-NC, [4]) offer the possibility of scaled gate dielectrics, implying scaled operating voltages, along with short channel effect immunity and higher sensing current drivability.

II. Device Fabrication

A schema of the Finflash structure fabricated in this work is shown in Fig.1, where the critical dimensions of the device are also reported.
The fabrication of our Finflash devices is based upon a standard Finfet process flow [5]. E-beam lithography and resist trimming are used to pattern both the fin and the gate. Sidewall oxidation is carried out to round fin corners and decrease fin width.

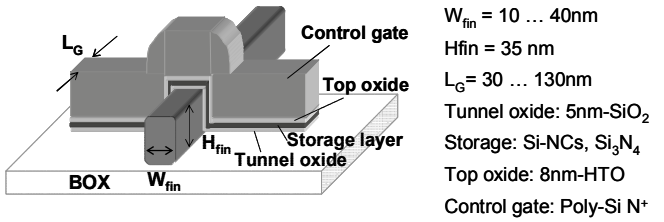

Fig.1: Schema of the Finflash memory cells.

After fin patterning and boron channel implantation, gate stack deposition is performed, i.e. the 5-nm thermal SiO_2, the storage layer made of Si-NC (directly deposited by LPCVD or obtained by Silicon Rich Oxide annealing) or 6nm-thick LPCVD Si_3N_4, the blocking dielectric (8nm-thick HTO) and, finally, the 100nm N^+ Poly-Si control gate. After the gate etching, nitride spacers are deposited and etched. Raised Source/Drain are epitaxially grown in order to decrease the series resistance. After the completion of source/drain implantation, the flow is terminated by standard Back-End-Of-Line.
TEM images of the Finflash devices are reported in Fig.2, demonstrating fin widths W_{FIN} and gate lengths L_G down to 10nm and 30nm, respectively.

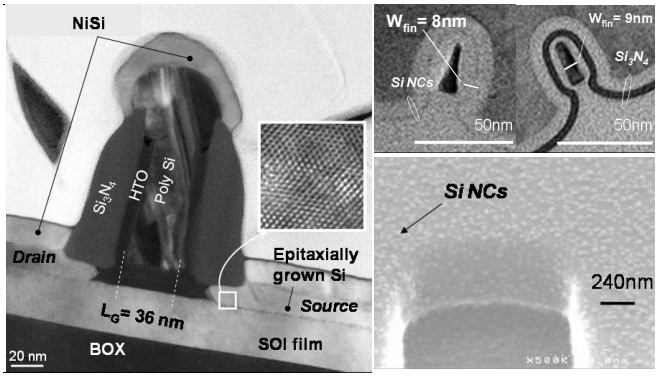

Fig. 2: TEM views of Finflash devices with different storage nodes.

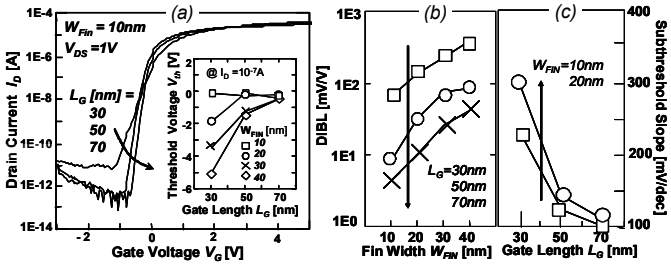

Fig.3: (a) I_D-V_G of Si-NCs Finflash (in the virgin state) with W_{FIN}=10nm and different L_G. Inset: V_{TH} versus L_G, for devices with different W_{FIN}.(b) DIBL versus W_{FIN} , for devices with different L_G. (c) Subthreshold slope versus L_G, for devices with different W_{FIN}.

III. FINFET CHARACTERISTICS

Fig.3(a) shows the transfer characteristics (I_D-V_G) of virgin Si-NC Finflash cells with W_{FIN}=10nm and different gate lengths. The enhanced electrostatic control of the gate over the channel at very small fin widths clearly appears. In particular, in the Inset, we can see that the threshold voltage V_{TH} roll-off disappears in narrow fins. Figs.3(b),(c) show that in the smallest devices (W_{FIN}/L_G=10/30nm), 220 mV/dec Subthreshold Slope ($SS@V_D$=1V) and 0.7V/V Drain Induced Barrier Lowering (DIBL) are achieved.

IV. MEMORY CHARACTERISTICS

Si-NC FinFlash Results - Ultra-scaled Si-NC Finflash devices are first studied in NOR configuration (i.e. Channel Hot Electron writing & Fowler-Nordheim erasing). Fig.4 and Fig.5 show that, in scaled devices (W_{FIN}/L_G=10/30nm and W_{FIN}/L_G=20/30nm, respectively), CHE yields large programming window with low V_D biases (lower than the Si/SiO_2 conduction band difference, *i.e.* ~3.2V). In particular, ΔV_{TH} ~3V can be achieved when V_D=2.5V, V_G=9V, and t_{stress}=100µs. We can also observe that Fowler-Nordheim erasing can be achieved in Silicon Nanocrystal FinFlash devices even with a 5nm-thick tunnel oxide (nevertheless, a saturation of the erase V_{th} occurs in the smallest device, Fig.4(c)).

Si-NC Finflash devices can also be programmed in the NROM operating scheme (i.e. Channel Hot Electron writing & Hot Hole Injection erasing). The W/E dynamics are reported in Fig.6, with the programmed threshold voltages read either in the forward mode (V_{DS}=1V) or in the reverse mode (V_{SD}=1V) [6]. Indeed, we can clearly observe the asymmetry between the forward/reverse V_{th}s, clearly suggesting that even for such strongly scaled devices the charges injected at the drain do not spread over to the source. Moreover, it can be noticed that erasing by hot holes is effective at a very low drain bias (lower than the Si/SiO_2 valence band difference, *i.e.* ~4.5V).

Fig. 7 shows the virgin and written V_{th}s in forward and reverse mode versus Lg. We observe a write boost for the shortest devices and detached forward/reverse V_{th}s.

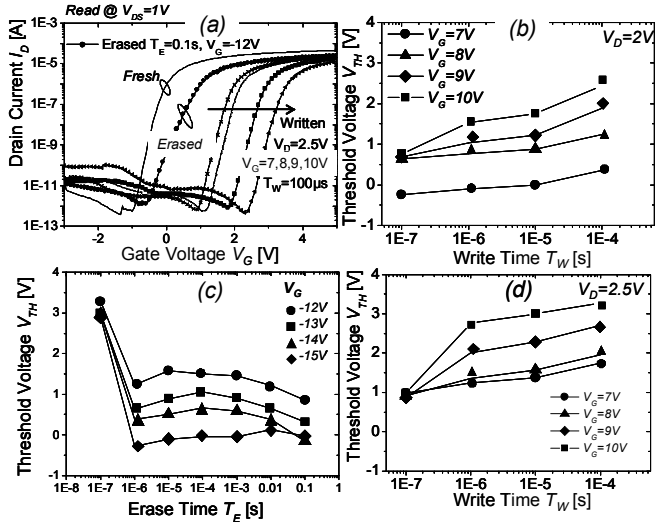

Fig.4: CHE/FN characteristics of Si-NC Finflash with W_{FIN}=10nm, L_G=30nm. (a) I_D-V_G in virgin, written (CHE) and erased (FN) states. (b,d) Write and Erase (c) dynamics.

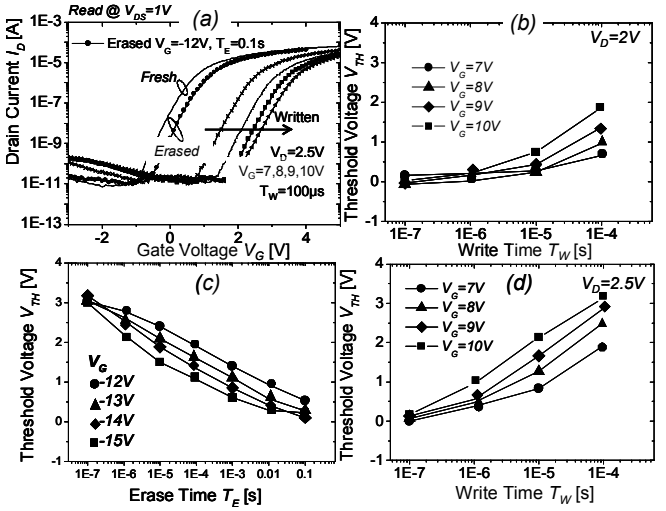

Fig.5: CHE/FN characteristics of Si-NC Finflash with W_{FIN}=20nm, L_G=30nm. (a) I_D-V_G in virgin, written (CHE) and erased (FN) states. (b,d) Write and Erase (c) dynamics.

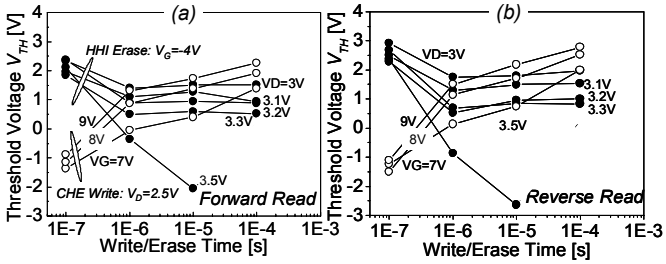

Fig.6: CHE/HHI characteristics of Si-NC Finflash with W_{FIN}=10nm, L_G=30nm. Programmed threshold voltages are read (a) in the forward mode (V_{DS}=1V) or (b) in the reverse mode (V_{SD}=1V).

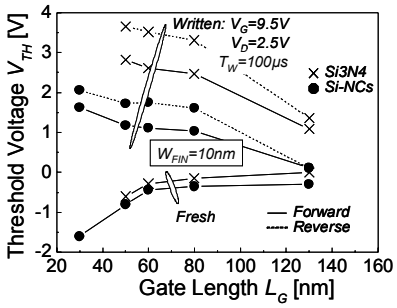

Fig.7: Virgin and written (by CHE) threshold voltages V_{TH}s (read in the forward and reverse mode) versus L_G of Si-NC and Si_3N_4 Finflash with W_{FIN}=10nm.

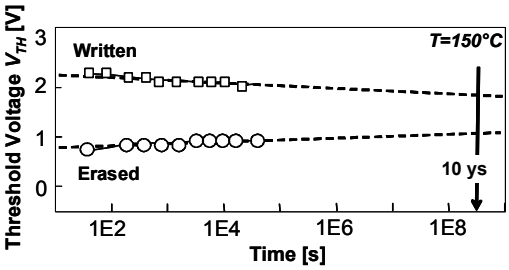

Fig.8: Data retention @ T=150°C of Si-NC Finflash with *W_{FIN}=20nm, L_G=30nm (CHE/FN Written/Erased).*

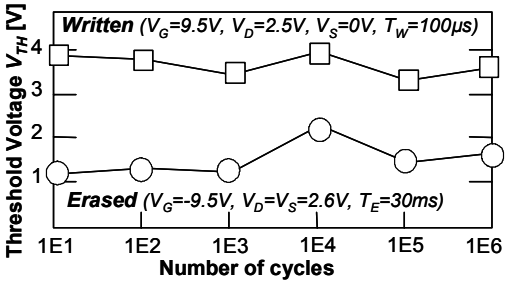

Fig.9: Endurance of Si-NC FinFlash with *W_{FIN}=20nm, L_G=30nm.*

Data-retention of Si-NC device with W_{FIN}/L_G=10/30nm is reported in Fig.8, showing small charge loss at high temperature (150°C). Good endurance (up to 1E6 cycles) of Si-NC device with W_{FIN}/L_G=20/30nm also appears in Fig.9. Nevertheless, it should be stated that a slight degradation of the I_D-V_G characteristics appeared after 1E5 cycles

Si_3N_4 FinFlash Results - Ultra-scaled nitride Finflash devices are studied in NROM configuration (i.e. Channel Hot Electron writing & Hot Hole Injection erasing), the Fowler-Nordheim erasing of charged nitride memories being not effective with 5nm-thick tunnel oxide.

As we previously observed in Si-NC devices, strongly scaled Si_3N_4 devices can be efficiently written with V_D biases lower than 3.2V and erased by HHI with V_D biases lower than 4.5V (Fig.9), while these low-voltage stresses are not effective for long devices.

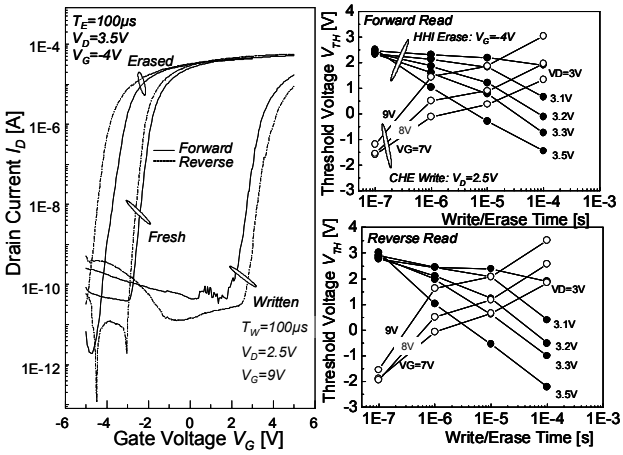

Fig.10: CHE/HHI characteristics of Nitride FinFlash with (W_{FIN}=10nm, L_G=30nm). Left: I_D-V_G characteristics. Right: W/E dynamics, with Vth read in forward and reverse mode.

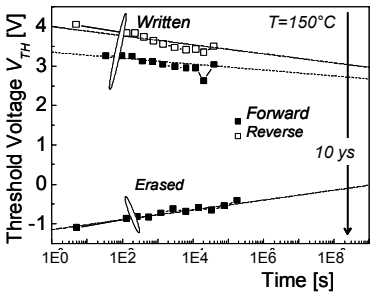

Fig.11: Data retention @ T=150°C of Nitride FinFlash with (W_{FIN}=10nm, L_G=30nm) (CHE/HHI Written/Erased).

Moreover, even in nitride devices with ultra reduced cell lengths, a good threshold voltage difference between the reverse and forward states appears.

In Fig.7 we can remark that the nitride storage layer gives rise to a larger programming window than the Si-NC storage layer, probably due to the higher trap density of amorphous nitride compared to crystalline Si-NCs.

Data-retention of *Si_3N_4* devices with W_{FIN}/L_G=10/30nm is reported in Fig.11, showing small charge loss at high temperature (150°C) and still detached forward and reverse threshold voltages after 10 years.

V. MODELING

The effective Channel Hot Electron writing with sub-3.2V drain bias achieved in our ultra-small devices can be qualitatively explained through Monte Carlo simulations [7] shown in Fig.12. From the results, we can consider that in a device with a 30nm-channel length the electrons acquire a large kinetic energy in the vicinity of the drain, even up to 2 eV.

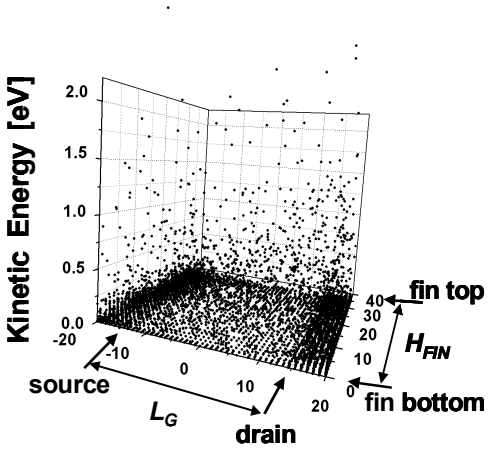

Fig. 12: 3D Monte Carlo simulation of CHE injection: snapshot of kinetic energy distribution of electrons in the channel for W_{FIN}/L_G=20/30nm at V_G/V_D= 10V/2.5V.

This means that electrons do not have sufficient energy to overcome the gate oxide barrier, but – since the gate oxide is quite thin (5nm in our devices) – they can still charge the storage layer via Fowler-Nordheim tunnelling. Indeed, FN tunneling probability is extremely sensitive to electron kinetic energy: the higher the kinetic energy, the smaller and the thinner the barrier to overcome. Programming with sub-3.2 V V_{DS} therefore requires both thin oxides and short channels, a condition fulfilled in the devices here considered. Obviously, the same phenomenon cannot take place in long devices (*i.e.* L_G>100nm), because the longitudinal electric field between the source and the drain is not sufficiently high to let electrons acquire enough kinetic energy in their characteristic mean free path (i.e. without scattering events). Thus long devices cannot be written with low V_D bias.

We also investigate the relationship between the device channel length L_G and the localized pocket of charge after CHE, characterized by the two following parameters: L_{TRAP}, the length of the trapped charge region close to the drain contact, and Q_{TRAP}, the uniform trapped charge density. To this aim, the experimental data of forward and reverse programming windows of Fig.7 are analyzed through a compact model, detailed in [8]. The model, based on the Green's function theory, provides the impact of a uniformly distributed trapped charge on the potential distribution inside the SOI fin.

In Fig.13, we show the behavior of the surface potential Ψ_S (at the fin/tunnel oxide interface, along a longitudinal cut in the middle of the fin) for the devices with different channel lengths.

	Si-NCs			Si3N4	
L_G [nm]	Q_{TRAP} [x 1e12 cm^{-2}]	L_{TRAP} [nm]		Q_{TRAP} [x 1e12 cm^{-2}]	L_{TRAP} [nm]
50	~7	~24		~11.2	~25
60	~7.3	~27		~11	~30
80	~5.7	~44		~9.5	~45

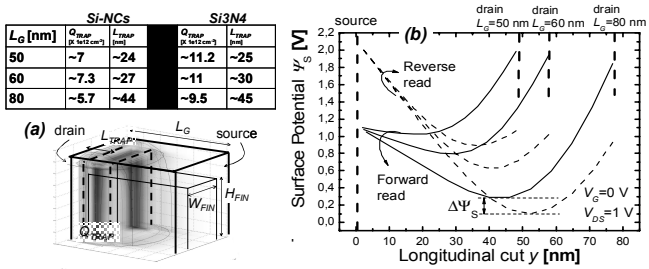

Fig. 13: Table of Q_{TRAP} and L_{TRAP} of Finflash structures with L_G=50, 60 and 80 nm, as extracted by applying the compact model of [8] to data of Fig.7. (a) Schema of the simplified structured used in the compact model [8]. (b) Surface potential (at the fin/tunnel oxide interface) behavior of the modeled devices.

We remark that the forward and reverse read provide two well detached Ψ_S minima, thanks to the localized pocket of trapped charge. According to the model, we can conclude that, in Finflash devices (with channel lengths from 50nm to 80nm) programmed by CHE with same voltages, the length of the pocket of the injected charges, L_{TRAP}, scales with the overall cell length L_G.

One can qualitatively explain the scaling of the charged region with L_G based on the following considerations. In fact, the kinetic energy of the most energetic electrons (say E_{hot}) increases smoothly from the source to the drain along the channel, up to a value close to $q.V_{DS}$. Thus, roughly, $E_{hot} \sim q.V_{DS}.y(x/L_G)$, where x is the longitudinal coordinate and $y(x/L_G)$ is a monotonic function (~0 at the source; ~1 at the drain). Since the trapped charge distribution f_Q is a function of the local gate current $J_{TUN}(x)$, which is in turn a function of E_{hot}, we have indeed that $f_Q=f_Q(x/L_G)$, i.e. the spatial extension of the stored charge distribution scales with L_G.

VI. CONCLUSION

Ultra-small SOI Finflash with Si-NC or nitride layers have shown to achieve functionalities in NOR and NROM operating schemes down to W_{FIN}/L_G=10/30nm. The effective CHE write and HHI erase at very low drain stress voltages in the sub-100nm gate length region have been demonstrated. 3D Monte Carlo simulations, coupled to a compact model, allow for a good comprehension of the physics related to the hot electron writing behaviour.

REFERENCES

[1] G. Atwood, *IEEE Trans. on Dev. & Mat. Rel., 07, 27, p.76, 2004.*

[2] S.H.Lee et al., *Tech. Dig. of IEDM 2006*, p.33.

[3] C.Friederich et al., *Tech. Dig. of IEDM 2006*, p.963.

[4] B.deSalvo et al., *Tech. Dig. IEDM 2003*, p. 597.

[5] C.Jahan et al, *VLSI Tech. Dig. 2005*, p. 112.

[6] B.Eitan et al, *IEEE El. Dev. Lett., Vol. 21, No. 11, 2000, p. 543.*

[7] G. A. Kathawala et al., *Tech. Dig. IEDM 2003, p. 683.*

[8] L.Perniola et al., *Proc. of SISPAD 2006, p.228.*