# Precise EOT regrowth extraction enabling performance analysis of Low Temperature Extension First devices

Jessy Micout, Quentin Rafhay, Xavier Garros, Mikael Cassé, Jean Coignus, Luca Pasini, Cao-Minh Vincent Lu, Nils Rambal, Claire Fenouillet-Beranger, Laurent Brunet, et al.

▶ **To cite this version:**

Jessy Micout, Quentin Rafhay, Xavier Garros, Mikael Cassé, Jean Coignus, et al.. Precise EOT regrowth extraction enabling performance analysis of Low Temperature Extension First devices. 2017 ESSDERC - 47th European Solid-State Device Research Conference, Sep 2017, Leuven, Belgium. pp.144-147, 10.1109/ESSDERC.2017.8066612 . cea-01525266

**HAL Id: cea-01525266**
**https://cea.hal.science/cea-01525266**

Submitted on 19 May 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Precise EOT regrowth extraction enabling performance analysis of Low Temperature Extension First devices

J. Micout[1,2], Q. Rafhay[1], X. Garros[2], M. Cassé[2], J. Coignus[2], L. Pasini[1,2,3],C.-M. V. Lu[2], N. Rambal[2], C. Fenouillet-Beranger[2], L. Brunet[2], G. Romano[3], R. Gassilloud[2], P. Batude[2], M. Vinet[2] and G. Ghibaudo[1]

[1]IMEP-LAHC, MINATEC/INPG, Univ. Grenoble Alpes, F-38016, Grenoble, France
[2]CEA, Leti, MINATEC Campus, Univ. Grenoble Alpes, F-38054, Grenoble, France
[3]STMicroelectronics, 850 rue Jean Monnet, F-38926, Crolles Cedex, France
jessy.micout@cea.fr

Abstract— **3D sequential integration requires top FETs processing with a low thermal budget (500°C). The analysis of the origin of the performance difference between Low Temperature (LT) MOSFET and high temperature standard process must take into account a potential EOT modification for short gate lengths. In this work, the difficulty of precise EOT extraction for scaled devices is observed by CV measurements and an alternative methodology using IV measurements is proposed. This methodology has been applied to an extension first integration, and the extraction accuracy is high enough to conclude to an EOT regrowth for the low temperature nFETs only. Thus, the origin of performance degradation between LT and HT, previously attributed to larger access resistance, highlights also a detrimental role of gate stack instability. The origin of this variation is attributed to oxygen ingress, through the thin extension first liner which should be suppressed by minor process optimizations.**

## I. INTRODUCTION

3D sequential integration is an alternative approach to conventional scaling. It consists in fabricating stacked layers of devices, sequentially one after the other, and enables to obtain the highest density of contact between the stacked layers (e.g. $10^8$/mm² for a 14nm node technology) [1]. To develop this integration, the thermal budget of the top transistor has to be reduced down to 500°C in order to preserve the underlying devices. The main challenge in thermal budget reduction concerns the thermal dopant activation usually made at temperatures around 1050°C. Low Temperature (LT) activation using Solid Phase Epitaxy Recrystallization (SPER) at 600°C has been shown to lead to device performance close from state of the art devices [2], as seen in **Figure 1**.

In this previous study, the performance degradation has been attributed to access resistance degradation, assuming a constant EOT with scaled gate lengths. Indeed EOT has been only extracted on large device in order to have a sufficient signal.

However an EOT variation for the shortest length devices could be observed and this regrowth might be different for the high temperature and low temperature process. As a reminder, a 1 Å difference between HT POR and LT splits is expected to leads to 10% performance degradation. This potential EOT regrowth might extend the original conclusion obtained on the performance degradation of LT process, presented in our previous work [2].

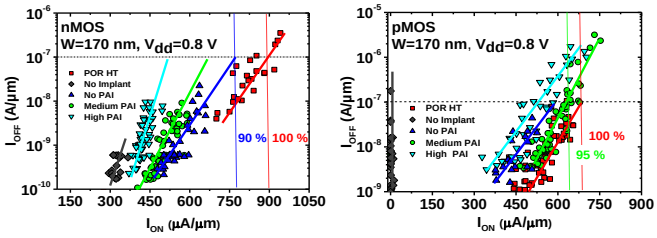



**Figure 1:** $I_{on}$ $I_{off}$ trade-off for low temperature extension first devices compared to high temperature process of reference [2]

In this in-depth investigation, the eventual EOT regrowth has been evaluated using both CV and IV measurements. The first section describes the device fabrication process; the second and third section presents the CV and IV results and the last section details the interpretation using models of gate current tunneling.

## II. DEVICE FABRICATION

The FDSOI devices are fabricated on 6nm Si and $SiGe_{27\%}$ thick channels with 20nm Buried Oxide, followed by $HfO_2$/ TiN/Poly-Si gate stack. Modifications appear afterwards, at the beginning of the junction formation. More specifically, **Figure 2** shows the difference between the HT POR and the extension first ($X^{1st}$) flow. The Low Temperature devices have been processed using the $X^{1st}$ flow (**Figure 2b)**). This process aims at localizing dopants at the gate edge as this low temperature process allows no dopant diffusion. Indeed, instead of depositing and etching a 6 nm Offset Spacer as for the POR HT case, implantations are carried out directly at the channel entrance through a 3 nm thin nitride liner. After the implantation, a nitride deposition is performed in order to reach the final POR spacer thickness. This thin and implanted (thus damaged) liner can be oxidized and thus could act as an oxygen reservoir very close to the gate edge.

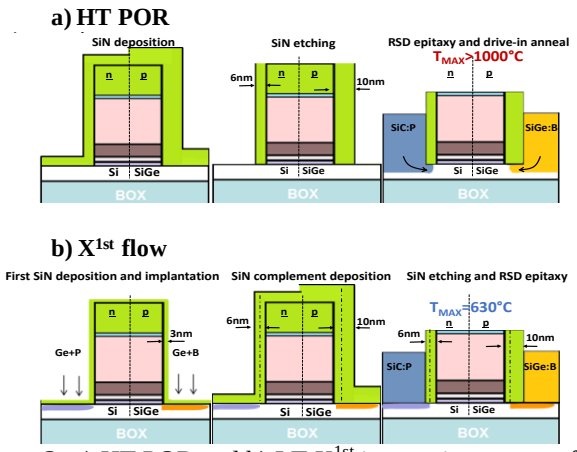



**Figure 2:** a) HT POR and b) LT $X^{1st}$ integration process flow.

It is therefore possible that this process flow could lead to an EOT regrowth for short devices ($L_g$<100nm), by oxygen ingress. To understand if the thin liner and/or the implantation can lead the EOT to increase, three technological splits will be compared, i.e.:

- The HT POR: standard FDSOI integration
- The LT Implant: $X^{1st}$ integration with implanted liner
- The LT No Implant: $X^{1st}$ integration without implanted liner

## III. EOT EXTRACTION BY C-V MEASUREMENTS

The EOT absolute value, extracted by CV measurements at 90 KHz, is obtained for all the gate length by fitting the curve between the measured capacitance and the simulated one, using [3].The differentiation of CV values for two close gate lengths enable to remove parasitic capacitance and to remove gate length uncertainties compared to the mask dimension. Moreover, an interdigitated transistor structure enables to measure the capacitance for the shortest gate lengths. **Figure 3** shows the C-V and the absolute EOT which is deduced for both n&p MOS. The differentiation has been carried out for the following couples of gate lengths: $\Delta L_1$ = 150 nm – 60 nm, and $\Delta L_2$ = 26 nm – 24 nm.

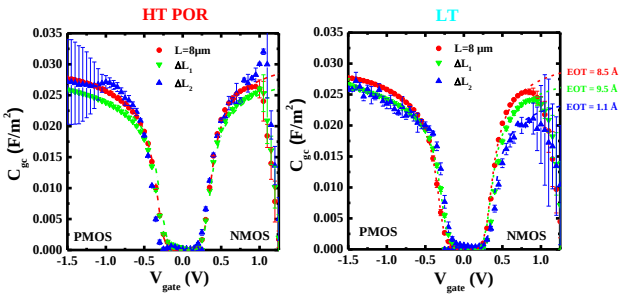



**Figure 3:** C-V graph and the deduced absolute EOT values

While for long channel the dispersion of the data is suitable for a correct extraction of EOT, the dispersion is too large for short channels. For example, in **Figure 3**, for the nMOS POR HT split, ΔEOT = 1 Å ± 1 Å, and for the nMOS LT No implant split, ΔEOT = 2 Å ± 1 Å. A conclusion about an EOT variation is therefore impossible using CV measurement. In addition, for pMOS HT POR, the fit for the shortest gate length is impossible due to the large dispersion.

The problem of the large dispersion could be explained by 1/ the linear dependency of the oxide capacitance with the EOT and by 2/ a too large gate leakage.

This variation is thus not satisfying to fully conclude about an EOT regrowth due to the LT process flow, and another method is proposed in this study.

## IV. EOT EXTRACTION BY I-V MEASUREMENTS

In the simple direct Fowler-Nordheim model, the gate current is linked to the EOT following equation (1):

$$I_g = W.L.A.e^{\left(-\frac{EOT}{B}\right)} \quad (1)$$

where A and B are constants depending on material parameters, but independent of the gate length L and the width W. Because of the exponential dependency, the measured gate current is expected to lead to a larger signal in case of EOT regrowth. It is however assumed here that any variation of the material parameters (like gate dielectric constant or effective masses) is interpreted as an EOT variation. The impact will however be the same at the performance level.

To ensure unbiased extraction, outlier data are at first removed by using Grubbs-Smirnov's statistical test [4].Then, to avoid the impact of short channel effects, the gate current is measured at low drain voltage ($V_d$ = 25 mV). In addition, to avoid trap-assisted current leakage, the gate current is taken at high $V_g$. Finally, as no gate resistance can distort the measurements as $I_g^{max}$<100 µA, the highest value of available gate voltage is taken, which is about 1.7 V. **Figure 4** shows the average gate current $I_g^{lin}$ versus the gate voltage $V_g$ measurements, for both HT POR and LT splits and for three gate lengths.

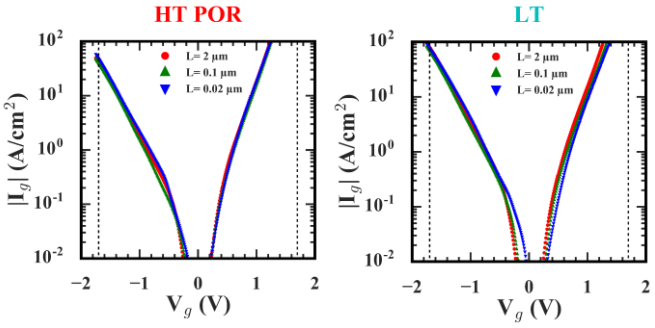



**Figure 4:** I-V graph

The comparison of the averaged $I_g^{lin}$ as a function of the gate length allows to identify if any degradation occurred. This degradation could be explained by 1/ an EOT variation, 2/ a difference between the mask and the effective channel length or 3/ a difference of the potential between the overlap and the channel, which will impact the gate current in the same manner as 2/. These three hypotheses will be studied in the following paragraphs.

At first, it has to be noted that with a lithography precision estimated around ±2 nm in the worst case scenario, the hypothesis 2/ would lead at maximum to a 20% degradation of $I_g^{lin}$ for the shortest gate length.

Then, assuming that there is no EOT regrowth and using the very simple direct tunneling model, 2/ and 3/would influence the gate current according to equation (2):

$$I_g(L) = W.(L+\Delta L).A.e^{\left(-\frac{EOT}{B}\right)} + I_{g,overlap} \quad (2)$$

Using a differential method with respect to the gate length, would lead to equation (3), which becomes independent of the gate length:

$$I_g^{new}\left(\frac{L_2+L_1}{2}\right) = \frac{\partial I_g}{\partial L} = \frac{I_g(L_2)-I_g(L_1)}{L_2-L_1} = W.A.e^{\left(-\frac{EOT}{B}\right)} \quad (3)$$

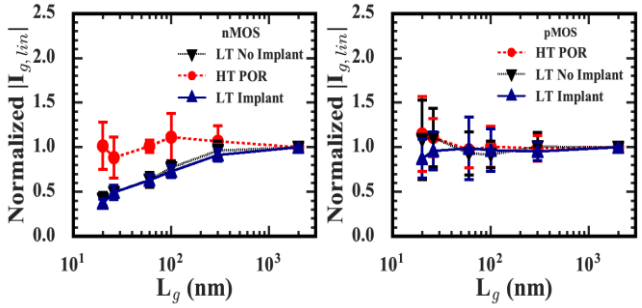
Therefore, if a monotonic gate current variation is caused by an extra gate length ΔL, applying the differential method would give an $I_g^{new}$ parameters independent of L.

On the contrary, if the gate current variation is cause by an EOT variation with L, the differential method would give $I_g^{new}$ values still varying with L, according to:

$$I_g^{new}\left(\frac{L_2+L_1}{2}\right) = W.A.e^{\left(-\frac{EOT(L)}{B}\right)}\left(1-\frac{L+\Delta L}{B}\frac{\partial EOT(L)}{\partial L}\right) \quad (4)$$

In that situation, EOT(L) could be extracted more simply from the logarithm of $I_g$ per unit area, if a value of B is known.

**a) Classical method**




**b) Differential method**

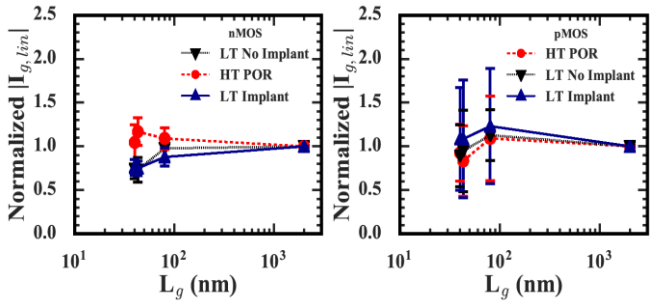



**Figure 5**: Relative $I_{g,lin}$ vs gate length, for W = 1 µm, |Vd| = 0.025 V and |Vg| = 1.7 V for both methods

**Figure 5a)** plots the $I_g^{lin}$ per unit area, normalized with respect to the $I_g^{lin}$ per unit area of the longest gate length, as a function of the gate length, for nMOS and pMOS devices of the HT POR and the LT processes. **Figure 5b)** plots the result of the differential method described by equation (3) and (4) and is therefore based on gate current values not divided by the length. The results of the differential methods are then normalized with respect to the value obtained for the longest gate length.

The following results are shown in **Figure 5a)**:

- For nMOS, the normalized $I_g^{lin}$ of the HT POR is relatively constant, which clearly indicates a stable EOT or a very limited ΔL. A significant degradation however occurs starting from 300 nm for the LT splits, and larger than a factor 2 for the shorter gate length of 20 nm.
- For pMOS, the variation of the normalized $I_g^{lin}$ is far less significant, and the variability in the value of $I_g^{lin}$ is much larger than in the nMOS case, as indicated by the errors bars (standard deviation of $I_g^{lin}$ around its average).

The use of the differential method shown in **Figure 5b)** is then used to clarify the origin of the previous results:

- As expected, the differential of $I_g^{lin}$ for nMOS devices of HT POR is relatively constant, which clearly indicates a stable EOT and very limited ΔL. For the LT splits however, the differential is not constant, which indicates a degradation of the EOT, and not an impact of the ΔL.
- The weakly varying values for pMOS confirm a weaker regrowth of the EOT. Note that the presence of a source of variability strongly impacts the gate current with this method.

These results therefore suggest that it is possible to extract a value of ΔEOT from the logarithm of the gate current per unit area, if a value of B is known:

$$\ln(I_g(L_{Long})) - \ln(I_g(L)) = \frac{\Delta EOT(L)}{B} \quad (5)$$

## V. COMPARISON BY MODELISATION

The value of B in equation (5) is highly dependent on the model of gate tunneling current considered. The simple direct tunneling model is convenient to separate the different contribution of the variation of $I_g^{lin}$, as highlighted in the previous section, but it suffers from too strong assumptions to be applied to the complex $SiO_2/HfO_2$ gate stack of the devices considered here. To obtain a better estimation of the B parameter, a more complete modeling of the gate current has hence been carried using the scattering matrix formalism, assuming a 3D electron gas in the semiconductor [5][6]. The $SiO_2$ thickness has been varied from 2 Å to 4 Å and the $HfO_2$ one from 19 Å to 21 Å. The value of the $SiO_2$ effective mass has been set to 0.5 $m_0$, while the one of $HfO_2$ has been fixed at 0.165 $m_0$. The tunneling currents obtained with this approach are shown in **Figure 6**.

A value of 1.2 Å has been extracted from these calculations for the B parameter. It will be used in the next section to deduce the resulting ΔEOT observed in LT nMOS devices.

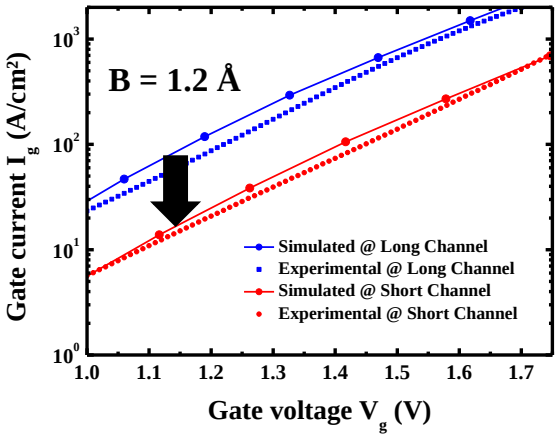



**Figure 6:** Comparison between simulated and experimental gate current for short and long gate lengths.

## VI. Analysis

The ΔEOT value, plotted in **Figure 7**, is hence deduced from equation (5). The value of B is obtained using the scattering matrix formalism.

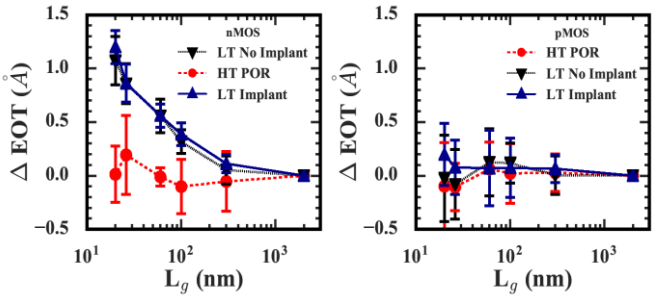



**Figure 7**: ΔEOT variation as function of the gate length, for W=1µm, |Vd|=0.025 V and |Vg|= 1.7 V

In the case of pMOS, no EOT regrowth is found for HT POR and LT with a high precision of 0.4 Å.

For nMOS, HT POR shows also no EOT regrowth, while for LT, a clear degradation of 1.2 Å ± 0.2 Å between short and long channel is observed. This degradation corresponds to a variation of 10% of the EOT and could thus lead to 10% $I_{on}$ degradation, as $I_{on}$ is inversely proportional to EOT.

As anticipated, this EOT regrowth could be attributed to an oxygen ingress coming from an oxidation of the thin liner. This ingress is expected to occur during the spacer complement deposition, which is made at 630°C during 2 hours. The thin liner oxidation is hence not linked to the implantation as the EOT regrowth is strictly the same of the LT splits, with or without implantation. The difference in EOT regrowth between n and p MOS LT splits can be explained by the fact that the latter is built on $SiGe_{27\%}$ channel and the oxidation kinetics in SiGe is lower than for Si [7].

To resolve this problem, a small reducing treatment could be applied before spacer complement deposition. A queue time reduction between the first liner deposition and the complement one can also be envisaged.

## VII. Conclusion

In conclusion, gate current measurement appears as an interesting source of information to see the EOT value evolution for short gate lengths, enabling a higher precision than a capacitance measurement. A precise evaluation of the B parameters (which could be considered as a characteristic tunneling length) is however needed. It has been estimated in this work using the scattering matrix formalism.

These electrical characterizations have shown that EOT regrowth could be observed between LT and HT splits for nMOS only, while from CV measurements the uncertainties is too high to fully conclude about an EOT regrowth. Thanks to this work, the origin of performance degradation between LT and HT, previously only attributed to larger access resistance [2], has been extended to include the detrimental role of gate stack instability.

To avoid EOT regrowth for low temperature process, it is suggested to use a small reducing treatment, which should allow to obtain greater performance.


## Acknowledgment

This work is partly funded by the French Public Authorities through NANO 2017 program and EQUIPEX FDSOI11, ST-LETI Alliance program and by LabEx Minos ANR-10-LABX-55-01.


## References


[1] P. Batude et al., "3DVLSI with CoolCube process: An alternative path to scaling," in 2015 Symposium on VLSI Technology (VLSI Technology), 2015, pp. T48–T49.

[2] L. Pasini et al., "High performance CMOS FDSOI devices activated at low temperature," in 2016 IEEE Symposium on VLSI Technology, 2016, pp. 1–2.

[3] K. Romanjek, F. Andrieu, T. Ernst, and G. Ghibaudo, "Characterization of the effective mobility by split C(V) technique in sub 0.1 µm Si and SiGe PMOSFETs," Solid-State Electron., vol. 49, no. 5, pp. 721–726, May 2005.

[4] C. Adam, Essential mathematics and statistics for forensic science, Wiley. 2010.

[5] J. Coignus, C. Leroux, R. Clerc, G. Ghibaudo, G. Reimbold, and F. Boulanger, "Experimental investigation of transport mechanisms through HfO2 gate stacks in nMOS transistors," in 2009 Proceedings of the European Solid State Device Research Conference, 2009, pp. 169–172.

[6] D. K. Ferry, Quantum mechanics : an introduction for device physicists and electrical engineers, CRC Press. 2001.

[7] J. Grabowski and R. B. Beck, "Oxidation kinetics of silicon strained by silicon germanium," J. Telecommun. Inf. Technol., vol. nr 3, pp. 30–32, 2007.